Universidade do Minho
Escola de Engenharia

Joana Pinto Fernandes

Descoberta de Padrões em Sistemas de Dados
Uma Aplicação à Área da Prescrição de Medicamentos

Outubro de 2009

Descoberta de Padrões em Sistemas de Dados
Uma Aplicação à Área da Prescrição de Medicamentos
Joana Pinto Fernandes
UMinho | 2009

Universidade do Minho
Escola de Engenharia

Joana Pinto Fernandes

Descoberta de Padrões em Sistemas de Dados
Uma Aplicação à Área da Prescrição de Medicamentos

Mestrado em Informática

Trabalho efectuado sob a orientação do
Professor Doutor Orlando Manuel de Oliveira Belo

Outubro de 2009

À minha Família.

Agradecimentos

Agradeço em primeiro lugar ao meu orientador Professor Orlando Belo, por toda a dedicação e ajuda que me disponibilizou na realização desta dissertação. Em todos os momentos que necessitei de algum tipo de ajuda e orientação, foi a pessoa que sempre me ensinou como dar o passo seguinte e me fez crescer em termos técnicos e humanos. Apesar de por vezes achar os seus conselhos demasiados cautelosos, no fundo sempre os aceitei devido à sua vasta experiência e enorme respeito e confiança que tenho pela sua pessoa. Muito obrigado pelos seus conselhos e pela sua paciência demonstrada em diversos assuntos.

Agradeço também ao Nuno Alexandre Carvalho pela sua total disponibilidade relativamente a todos os problemas eoncontrados na utilização do servidor. Todos os pequenos problemas encontrados relativamente ao uso do servidor, foram todos resolvidos quase de imediato com a ajuda do Nuno Carvalho, mesmo os que aconteceram em dias menos próprios de trabalho, como ao fim de semana.

Sem a facultação dos dados por parte da Administração Regional de Saúde do Norte, era impossível a realização desta dissertação, consequentemente, queria fazer um agradecimento especial à ARSN.

Queria também agradecer à minha família por todas as alturas que me proporcionaram as condições de trabalho que necessitava. Queria agradecer em especial à minha mãe por todos os momentos disponíveis e por todo o encorajamento que me deu. Mesmo nos momentos mais complicados, ela deu-me todo o seu apoio e palavras que me fizeram acreditar. Por ser a pessoa que mais admiro e respeito, as suas palavras e sinceridade, fizeram-me sempre olhar em frente e nunca para trás.

Por fim queria agradecer à pessoa que mais me acompanhou e ajudou neste processo, Tiago Martins. Queria agradecer por todos os momentos em que me fez ver a realidade, fosse ela dura ou não. Por todos os momentos de descontracção que me proporcionou e por toda a paciência que teve comigo. É graças principalmente a ele, que encarei esta fase da minha vida sempre com a confiança que era capaz e a sorrir. Muito obrigado!

Resumo

Descoberta de Padrões em Sistemas de Dados

A vasta quantidade de dados armazenada pelas empresas e a necessidade da sua consequente análise, com o objectivo de retirar informação útil que possa ajudar na tomada de decisões, tem ajudado ao grande impulso do estudo e utilização das técnicas de mineração de dados no seio das organizações. A aplicação das técnicas de mineração de dados possibilita a extracção de informação útil, anteriormente, desconhecida, a partir de grandes quantidades de dados, através da descoberta de tendências, de padrões e de possíveis anomalias nos dados. A descoberta de padrões nos dados proporciona às empresas um acréscimo da taxa de crescimento dos seus negócios e serviços, através da tomada de decisões, suportadas pela informação descoberta. A descoberta de padrões pode ser aplicada a qualquer tipo de dados, como por exemplo, dados relativos à prescrição de medicamentos.

Por ano, em Portugal, existem milhões de pessoas que recorrem aos centros de saúde independentemente dos seus problemas de saúde. Devido a isso, existem dias ou mesmo meses mais assoberbados que outros, e neste caso uma boa gestão de recursos de saúde, permite um bom funcionamento desses centros, e a respectiva satisfação e confiança por parte dos utentes. No entanto, as decisões tomadas com esse objectivo, têm de ser suportadas por algum tipo de informação. Essa informação encontra-se "escondida", por exemplo, nos dados guardados, relativos à prescrição de medicamentos e só através de análises pormenorizadas se consegue obter a informação desejada. A aplicação das ténicas de mineração de dados permite a descoberta de tendências e padrões nos dados, que através de técnicas de análise convencionais, seriam impossíveis de descobrir. Neste trabalho, além da abordagem genérica às principais técnicas de descoberta de padrões, aplicou-se as respectivas técnicas a um caso de estudo real. As técnicas abordadas foram a associação, a segmentação e a classificação, dando-se particular ênfase aos algoritmos mais usuais. Após o tratamento de dados necessário ter sido realizado, este trabalho focou-se no objectivo final, a descoberta de padrões de prescrições de medicamentos através da aplicação de técnicas de mineração de dados. Tendo em consideração alguns centros de saúde dos distritos de Aveiro, Braga, Bragança, Porto, Viana do Castelo e Vila Real, foram realizados 3 estudos em separado, um para cada técnica explorada. Por fim, deve-se referir que, a aplicação das técnicas a um caso de estudo real só foi possível porque a Associação Regional de Saúde do Norte disponibilizou os dados alvo, com o objectivo de se descobrir eventuais padrões de prescrição de medicamento e possíveis tendências de prescrições de medicamentos, relativos a utentes, médicos e laboratórios prescritos.

Palavras-chave: ARSN; Prescrição de Medicamentos; Descoberta de Padrões; *Data Warehouse*; Técnicas de Mineração de Dados; Associação; Segmentação; Classificação.

Abstract

Finding Patterns in Data Systems

The huge amount of data stored by companies and the need for analyzing it, with the aim of obtain useful information for decision making, have helped the huge impulse of the study and use of data mining techniques among the organizations. Data mining techniques can be used to extract useful information, by finding patterns, trends and anomalies in large databases. Discovering the hidden patterns, provide useful information that can be used by the organizations, to make crucial business decisions before competition does. Data mining is not specific to any industry, therefore, it can be applied to data related with drugs prescription.

For year, in Portugal, there are millions of persons who seek for medical treatment regardless their health problems. Sometimes the number of persons in a day or month increases, and in this situations a proper resource management allows a better health care systems, and consequent the confidence and satisfaction of the patients. However, all the decisions with that purposes must be supported by some kind of information. That information is somehow "hidden", for example, in the data stored related to drug prescription, and only through detailed analysis it's possible to obtain the desired information. The use of data mining techniques allows the discovery of hidden patterns and unexpected trends in the data, that with the use of conventional techniques it would be impossible to extract that kind of information. In this work, in addiction to the generic approach of the main techniques for finding patterns, it was applied those same techniques to a real study case. The techniques discussed were association, clustering and classification, giving particular emphasis to the most common algorithms. After doing the pre-processing needed, this work focused on the aim goal, the discovery of patterns in drug prescription data by applying data mining techniques. Taking in consideration some health centers in the districts of Aveiro, Braga, Bragança, Porto, Viana do Castelo e Vila Real, 3 studies were conducted separately, for each one of the techniques exploited. Finally, it should be noted that this work was only possible due to the data provided from Associação Regional de Saúde do Norte, in order to identity any possible patterns and trends of drug prescription, related to the patients and the doctors prescription.

Keywords: ARSN; Drug Prescription; Finding Patterns; Data Warehouse; Data Mining Techniques; Association; Clustering; Classification.

Índice

Introdução **1**

1.1 Contextualização 1
1.2 Descoberta do Conhecimento através da Mineração de Dados 2
1.3 Motivação 8
1.4 Organização da Dissertação 9

Modelos e Padrões **11**

2.1 Descoberta de Padrões 13
2.1.1 Padrões Frequentes 15
2.1.2 Padrões Frequentes Sequenciais 17
2.1.3 Episódios Frequentes 21
2.2 Grau de Interesse de um Padrão 24
2.3 Dos Padrões aos Modelos 26
2.4 Regras de Associação 27
2.4.1 Medidas de Interesse 29
2.4.2 Algoritmos de Descoberta de Regras 32
2.4.3 Casos de Estudo 35
2.5 Segmentação 37
2.5.1 Medidas de Similaridade 38
2.5.2 Técnicas da Segmentação 41
2.5.3 Casos de Estudo 48
2.6 Classificação 50
2.6.1 Pré Tratamento dos Dados 50
2.6.2 Técnicas de Classificação 52
2.6.3 Avaliação do Desempenho de um Modelo 64
2.6.4 Casos de Estudo 66
2.7 Análise Geral das Técnicas 68
2.7.1 Associação 69
2.7.2 Segmentação 70
2.7.3 Classificação 71

Descoberta de Padrões **75**

3.1 Um Caso de Estudo 75
3.2 O Modelo de Dados 77
3.3 Análise Exploratória dos Dados 82

3.4 Análise Crítica da Qualidade dos Dados 92
3.5 Selecção e Limpeza dos Dados 94
3.6 Construção, Integração e Formatação dos Dados 95
3.7 Análise Exploratória Complementar 98
3.8 Aplicação das Técnicas de Mineração de Dados ao Modelo de Dados Final 99
3.8.1 Associação 100
3.8.2 Segmentação 108
3.8.3 Classificação 118
3.9 Análise Geral da Aplicação das Técnicas de Mineração de Dados 132
3.10 Software Utilizado 133

Conclusões e Trabalho Futuro 135
4.1 Conclusões 135
4.2 Trabalho Futuro 141

Bibliografia 143

Referências WWW 155

Anexos I 157

Índice de Figuras

Figura 1 – Processo da descoberta do conhecimento em bases de dados [Fayyad *et al.* 1996b]..... 4

Figura 2 – Estrutura típica de uma arquitectura necessara à aplicação de técnicas de mineração de dados (Han & Kamber 2006). 7

Figura 3 – Exemplo do funcionamento do algortimo *Apriori*. 17

Figura 4 – Exemplos de diferentes tipos de episódios. 22

Figura 5 – Divisão dos tipos de algoritmos. 33

Figura 6 – Segmentação dos dados.....37

Figura 7 – Cálculo das distâncias, Euclidiana e Manhattan, entre dois objectos [Han e Kamber 2006]. 38

Figura 8 – Exemplo de funcionamento da medida MND [Jain *et al*. 1999]. 41

Figura 9 – Funcionamento das duas sub técnicas hierárquicas. 42

Figura 10 – Exemplo da estrutura de uma árvore de decisão [Han & Kamber 2006]. 52

Figura 11 – Exemplo da estrutura de uma rede neural, em que (a) representa um grafo acíclico e (b) representa a tabela das probabilidade condicionais para a variável cancro dos pulmões [Han & Kamber 2006]. 61

Figura 12 – Percentagem da prescrição por distritos 82

Figura 13 – Percentagem de prescrições dos centros de saúde estudados. 83

Figura 14 – Nº de prescrições de distritos por 1000 utentes. 84

Figura 15 – Nº de prescrições por distritos e por grupo etário relativamente ao sexo feminino. 84

Figura 16 – Nº de prescrições por distritos e por grupo etário relativamente ao sexo masculino. 85

Figura 17 – Percentagem do número de prescrições por 1000 utente, tendo em conta os eixos de análise do sexo e faixa etária do utente. 86

Figura 18 – Percentagem do número de prescrições dos meses do ano. 87

Figura 19 – Número de prescrições de medicamentos por 1000 utentes, segundo o grupo etário. 87

Figura 20 – Percentagem de prescrições de medicamentos genéricos e não genéricos.....89

Figura 21 – Percentagem de prescrições de medicamentos genéricos por grupos etários. 89

Figura 22 – Nº de prescrições de medicamentos genericos e não genericos por 1000 utentes, segundo os distritos estudados. 90

Figura 23 – Esquema do *data mart* implementado. 98

Figura 24 – Percentagem do número de prescrições por 1000 utentes, tendo em conta os eixos de análise das estações do ano e dos dias da semana. 99

Figura 25 – Exemplo de uma estrutura FP-tree, em que (a) é a lista de transacções ordenada descendentemente pelo suporte de cada elemento e (b) representa alguns passos da construção da árvore e a árvore resultante. 101

Figura 26 – Sub árvores dos caminhos dos prefixos de *e* e *d*. 101

Figura 27 – Exemplo do funcionamento do algoritmo *FP-growth* através da *FP-tree* com um suporte minímo de 2 [www5].....102

Figura 29 – Representação gráfico dos valores médios dos centróides para os atributos em estudo.....114

Índice de Tabelas

Tabela 1 – Exemplo de Compras por Clientes. 15

Tabela 2 – Exemplos de produtos comprados por clientes por dia. 18

Tabela 3 – Sequência de compras de cada cliente.....18

Tabela 4 – Sequências de suporte igual ou superior a 2. 19

Tabela 5 – Sequências de compras transformadas. 20

Tabela 6 – Avaliação dos algoritmos de regras de associação [Zheng *et al*. 2001, Goethals 2003, Han & Kamber 2006, Hipp *et al*. 2000]. 70

Tabela 7 – Avaliação dos algoritmos de segmentação [Andritsos 2002, Han & Kamber 2006]. 73

Tabela 8 – Avaliação dos algoritmos da classificação [Han & Kamber 2006, Boullé 2007, Friedman *et al*. 1997, Cunningham & Delary 2007, Orponen 1994]. 74

Tabela 9 – Locais de atendimento relativos ao estudo em causa. 78

Tabela 10 – Descrição do modelo de dados inicial. 79

Tabela 11 – Descrição do modelo de dados inicial (continuação). 80

Tabela 12 – Descrição do modelo de dados inicial (continuação). 81

Tabela 13 – Descrição do modelo de dados inicial (continuação). 82

Tabela 14 – TOP 5 das unidades de saúde com taxa minima e maxima de prescrições por 1000 utentes. 83

Tabela 15 – Estatísticas descritivas do número de prescrições tendo em conta diversos factores. 86

Tabela 16 – Estatísticas descritivas tendo em conta o estudo dos medicamentos e laboratórios. 88

Tabela 17 – Estatísticas descritivas do top 10 dos médicos, tendo em conta o diferente número de laboratórios prescritos. 90

Tabela 18 – Análise da comparticipação dos 10 utentes com maior número dez prescrições.91

Tabela 19 – Conjuntos frequentes descobertos pela aplicação do *FP-growth* à lista de transacções da Figura 25 (a). 102

Tabela 20 – Regras de Associação entre os medicamentos prescritos. 105

Tabela 21 – Regras de Associação entre os médicos e os laboratórios prescritos. 107

Tabela 22 – Gama de valores dos atributos em estudo na aplicação do algoritmo *k-means*. 110

Tabela 23 – Resultados da avaliação do *k-means* para $k = 2$ até ao $k = 8$. 112

Tabela 24 – Resultados da avaliação do *k-means* para $k = 9$ até ao $k = 14$ 112

Tabela 25 – Resultados do segmento 0 ao segmento 6. 113

Tabela 26 – Resultados do segmento 7 ao segmento 12. 113

Tabela 27 – Parâmetros de qualidade de cada árvore gerada e o número mínimo de folhas respectivo. 122

Tabela 28 – Listagem dos príncipios activos dos utentes da Tabela 18. 162

Lista de Siglas e Acrónimos

AGNES	Aglomerative Nesting
AIM	Autorização de Introdução no Mercado
API	Application Program Interface
ARSN	Associação Regional de Saúde do Norte
BFS	Breadth-First Search
BIRCH	Balanced Iterative Reducing and Clustering
CART	Classification and Regression Trees
CLARA	Clustering Large Applications
CLARANS	Clustering Large Applications based upon Randomized Search
CLS	Concept Learning System
COD	Cause of Death
CURE	Clustering Using Representatives
DB	Davies Bouldin
DBSCAN	Density Based Spatial Clustering Applications with Noise
DIANA	Divisive Analysis
DFS	Depth-First Search
EM	Expectation-Maximization
FP	Frequent Pattern
ICPC	Classificação Internacional de Cuidados Primários
ID3	Iterative Dichotomiser
INFARMED	Instituto Nacional da Farmácia e do Medicamento
JDBC	Java Database Connectivity
KDD	Knowledge Discovery in Databases
MI	Mutual Information
MND	Mutual Neighbor
NN	Nearest Neighbor
OLAP	Online Analytical Processing
PAM	Partitioning Around Medoids
ROCK	Robust Clustering Algorithm for Categorical Data

SAM	Sistema de Apoio ao Médico
SOM	Self-Organizing feature Maps
SSE	Soma dos Quadrados dos Erros
STING	Statistical Information Grid-based Method
STR	Survival Time Recode
TID	Transaction Identifier
USF	Unidades de Saúde Familiares
VST	Vital Status Recode
XML	Extensible Markup Language

Capítulo 1

Introdução

1.1 Contextualização

Numa sociedade em constante crescimento, e numa era considerada como sendo a era da informação, a vantagem competitiva de uma empresa é alcançada de acordo com a habilidade de adquirir e manusear a maior quantidade de informação útil dos dados existentes. Os recorrentes avanços na tecnologia dos sistemas de informação possibilitam estratégias mais fáceis e avançadas na recolha dos dados. Essa crescente recolha de dados fez com que as empresas acrescentassem um novo processo no seu variadíssimo rol de actividades: a exploração selectiva de dados. Isto é, o estudo dos hábitos comportamentais das pessoas, possibilita às empresas a tomada de decisão, com base na possível descoberta dos seus comportamentos futuros, com o objectivo final de melhorar o serviço prestado aos seus clientes e assim ganhar vantagem competitiva. Uma das formas de realizar tal processo é através da descoberta de padrões de comportamente dos dados relativos aos seus clientes.

Os avanços na tecnologia dos sistemas de informação possibilitaram maneiras mais fáceis e avançadas na recolha e análise desses dados [Fayyad *et al*. 1996a], que são maioritariamente guardados em sistemas de base de dados para uma maior facilidade de análise [Sumathi & Sivanandam 2006]. O processo de análise dos dados consiste, essencialmente, na descoberta de tendências, padrões ou possíveis anomalias existentes nos dados armazenados, que inicialmente poderão não ser visíveis através de técnicas de análise ditas convencionais [Lloyd-Williams *et al*. 1995], impossibilitando assim uma análise pormenorizada dos dados. A necessidade de exploração dessas tendências e padrões é feita através de mecanismos de mineração de dados, que permitem a extracção de informação

útil, anteriormente desconhecida, através da aplicação de diversos algoritmos a grandes quantidades de dados [Sumathi & Sivanandam 2006]. Consequentemente, a aplicação de técnicas de mineração de dados aos sistemas operacionais das empresas, permite-lhes um aumento de informação útil, que lhes irá possibilitar a tomada de decisões de maneira mais segura e favorável.

1.2 Descoberta do Conhecimento através da Mineração de Dados

As técnicas de análise de dados convencionais são normalmente usadas na análise da regressão e segmentação, na análise multidimensional e de séries temporais, em modelos estocásticos e em diversos conjuntos de variadas análises estatísticas [Sumathi & Sivanandam 2006]. No entanto, essas técnicas, são essencialmente, orientadas à extracção quantitativa e estatística das características dos dados, e como tal, apresentam algumas limitações inerentes. Isto é, por exemplo, as análises estatísticas podem determinar a correlação entre duas variáveis existentes nos dados, no entanto não permitem caracterizar as dependências a um nível abstracto, nem apresentar razões para a existência dessas dependências. Uma análise estatística pode determinar a tendência central e a variância de dados factores, bem como utilizar uma análise de regressão para ajustar uma curva a um conjunto de pontos. Contudo, mais uma vez, esta análise apresenta limitações, pois não permite obter uma descrição qualitativa das regularidades e determinar os factores de dependência que não se encontram explícitos nos dados [Sumathi & Sivanandam 2006]. Mas, a principal limitação é a análise de elevadas quantidades de registos em tempo real, algo que cada vez mais se torna um requisito no mundo empresarial. Essas limitações, entre outras, são consideradas um grande entrave às análises estatísticas mais usuais, e devido a essas limitações emergiu uma nova área de pesquisa, conhecida como a mineração de dados (*Data Mining*).

A mineração de dados, termo inicialmente apenas abordado por estatísticos, só mais tarde começou a ganhar alguma popularidade na área de base de dados [Fayyad *et al*. 1996b]. Apesar de à primeira impressão, essas duas áreas serem bastantes similares, ao pormenor, constata-se que existem diferenças notórias entre elas, tendo em consideração as limitações e abrangência do resultado final de ambas. Essa similaridade entre as duas áreas, provocou alguma confusão, tendo até, em alguns casos, causado algum tipo de antipatia. O aparecimento de uma nova técnica, através de novas pessoas que tinham a pretensão de resolver problemas que até então caiam no domínio dos estatísticos, gerou alguma preocupação aos estatísticos. Essa preocupação foi mais notável devido ao nome sonante

desta nova técnica, que causou interesse e curiosidade no mundo, e também, devido ao facto desta técnica apresentar uma relevância particular em relação às preocupações comerciais [Hand 1999]. Uma das grandes diferenças entre as duas técnicas reflecte-se no objectivo final de cada uma delas. Isto é, o objectivo central da mineração de dados é a descoberta do conhecimento, ao contrário da estatística, que concentra-se na melhor maneira de colectar os dados com o intuito de conseguir responder mais acertadamente a uma questão específica. A mineração de dados assume, essencialmente, que os dados já foram colectados anteriormente, e a sua preocupação cinge-se apenas, na descoberta de informação escondida, mas valiosa, nos dados. De uma forma geral, a mineração de dados é considerada com um processo exploratório dos dados, enquanto que a estatística é considerada como um processo confirmatório [Hand 1999].

A mineração de dados é também encarada como sendo uma passo específico da descoberta do conhecimento em bases de dados - *Knowledge Discovery in Databases* (KDD) -, consistindo essencialmente, na aplicação de algoritmos específicos com vista à descoberta de informação [Fayyad *et al*. 1996a, Fayyad *et al*. 1996b]. O processo KDD, abordado pela primeira vez na primeira conferência de KDD em 1989 (Piatetsky-Shapiro 1991), é referenciado como sendo o processo global, não trivial, da descoberta de padrões novos, válidos, potencialmente úteis e de boa compreensão dos dados [Fayyad *et al*. 1996b]. Em que os dados são um conjunto de factos (casos de uma base de dados, por exemplo), e os padrões são uma expressão numa linguagem, que descrevem um subconjunto dos dados ou um modelo aplicável ao subconjunto. Neste contexto, a extracção de um padrão também pode ser designada como o ajuste de um modelo aos dados e/ou a descoberta de uma estrutura que define os dados. De uma forma geral, consiste em fazer uma descrição de alto nível de um conjunto de dados. O termo processo implica o facto de o KDD abranger um conjunto de passos como a preparação dos dados, descoberta de padrões, avaliação do conhecimento e refinamento. E, por fim, ao ser não trivial, significa que alguma pesquisa ou inferência é necessária, isto é, o processo KDD não é feito através de uma computação directa, como o cálculo da média de um conjunto de valores [Fayyad *et al*. 1996b]. O processo KDD é um processo interactivo e iterativo, que envolve a realização de um conjunto de passos, com inúmeras decisões necessárias por parte do utilizador. De seguida serão descritos todos os passos a ter em consideração na realização deste processo [Fayyad *et al*. 1996a, Fayyad *et al*. 1996b, Santos & Azevedo 2005]:

1. Selecção dos dados: após a aprendizagem e compreensão do domínio da aplicação, é feita a selecção e recolha de um conjunto de dados ou focalização da presquisa em amostras de dados, de acordo com alguns critérios.

2. Limpeza e pré-processamento dos dados: este passo inclui as operações básicas, como a remoção de ruído e erros dos dados e decisão da melhor estratégia a tomar, na existência de valores em falta e determinação da informação relevante.

3. Transformação dos dados: consiste, basicamente, em descobrir características úteis que representem os dados, dependendo do objectivo final. Através de métodos de redução e transformação da dimensionalidade, o número efectivo de variáveis em consideração pode diminuir, ou representações invariáveis dos dados podem ser encontradas; nesta fase, torna-se evidente o uso de um *data warehouse*, pois nestas estruturas, os dados são não voláteis, classificados por assunto e de natureza histórica, tendendo, assim, a tornarem-se grandes repositórios de dados, extremamente organizados.

4. Mineração de Dados: esta etapa pode ser subdividida em três outras; numa fase inicial, é necessário escolher qual a melhor técnica de mineração de dados a aplicar, de acordo com os objectivos finais deste processo; após a escolha da técnica a utilizar é preciso escolher os algoritmos existentes relativamente à técnica em causa; neste caso é necessário ter especial atenção aos dados em estudo e ao resultado expectável; a última fase consiste na descoberta de padrões de interesse numa determinada forma representativa ou num conjunto de tais representações, tal como, construção de regras ou árvores de classificação, equações de regressão, segmentos, entre outras.

5. Interpretação dos resultados obtidos: inclui a visualização dos padrões extraídos e a sua interpretação, a remoção dos padrões irrelevantes e redundantes e a tradução dos padrões considerados úteis em informação compreensível aos utilizadores; caso seja necessário, existe ainda a possibilidade de retornar para qualquer uma das fases anteriores; após a obtenção dos resultados desejados, apenas falta a incorporação do conhecimento descoberto no desempenho do sistema, através de tomadas de decisão, ou simplesmente, documentar e guardar os resultados obtidos, de maneira, a se conseguir resolver potenciais e futuros conflitos.

A componente mineração de dados de um processo KDD envolve repetidas iterações de aplicações de métodos particulares da mineração de dados [Fayyad *et al*. 1996a]. Isto acontece, devido à necessidade de ajustar os parâmetros dos algoritmos, de maneira a se conseguir o melhor resultado possível, tendo em conta o método e os algoritmos utilizados na aplicação da mineração de dados. Apesar da mineração de dados ser apenas um passo específico de todo o processo KDD, ela tem recebido muita atenção e assumido particular importância quando comparada com o processo global em si. No entanto, é preciso referir, que nunca se deve esquecer todo o processo global, pois só através da execução de todos os

passos, se consegue obter um resultado final credível e com algum impacto prático, disponibilizando conhecimento útil. Na Figura 1 estão apresentados de forma sucinta os principais passos do processo de descoberta do conhecimento em base de dados.

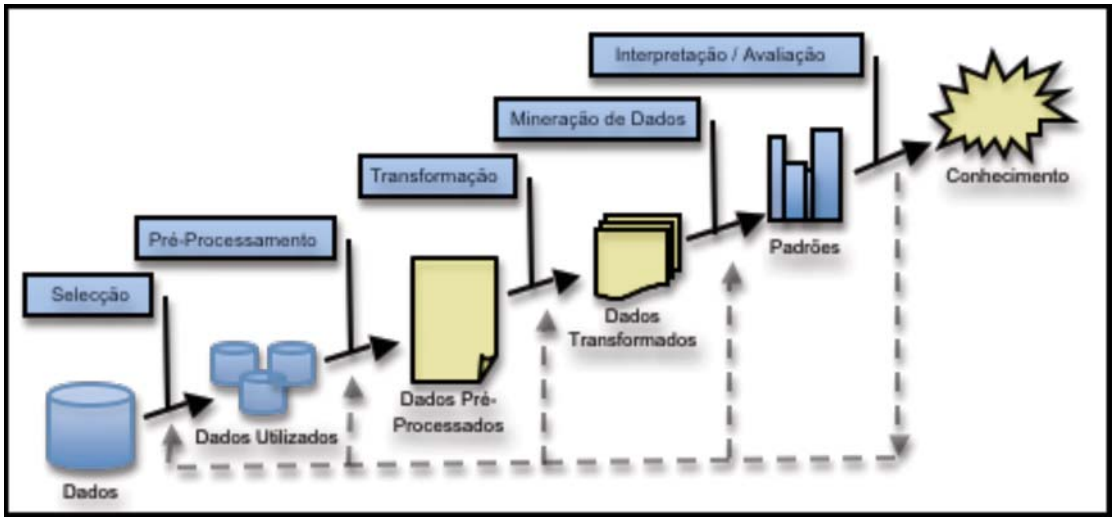


Figura 1 – Processo da descoberta do conhecimento em bases de dados [Fayyad *et al.* 1996b].

Como foi visto anteriormente, as capacidades da mineração de dados são enormes e possibilitam a obtenção de informação bastante valiosa. A aplicação de técnicas de mineração de dados, consiste em ajustar modelos a um conjunto elevado de dados, ou descobrir padrões existentes nesses mesmos dados [Fayyad *et al*. 1996a, Fayyad *et al*. 1996]. No caso de ajuste de modelos é necessário haver um conhecimento prévio do comportamento dos dados, de maneira a ser possível a criação de um modelo, que permita a previsão do comportamento futuro dos dados, com alguma fiabilidade, ou a obtenção de uma descrição bastante rigorosa dos dados, com o intuito de obter nova informação. No entanto, a qualidade e utilidade de um modelo é bastante subjectiva, dependendo dos objectivos e resultados esperados pelo utilizador [Fayyad *et al*. 1996a, Fayyad *et al*. 1996b, [www1]]. Na descoberta de padrões em sistema de dados é possível obter-se um número elevado de padrões, que por vezes poderão ser bastante óbvios (logo desinteressantes) ou mesmo não ser "reais", uma vez que apenas reflectem uma estrutura especifica ou relação, em vez de reflectir o comportamento geral do conjunto de dados [Hand 1999]. Em ambos os casos, a qualidade da informação obtida depende do tipo de informação necessária ao utilizador em causa (apenas padrões que constatem nova informação são considerados interessantes), mas também de alguns outros parâmetros que avaliam essa qualidade (percentagem de previsão correcta de uma classe, ou suporte e confiança de uma regra). No capítulo seguinte será abordado com algum pormenor os possíveis métodos de avaliação de um modelo, dando mas ênfase à descoberta de padrões.

Em suma, a mineração de dados é um campo bastante interdisciplinar que aglomera um conjunto de técnicas estatísticas, de visualização, de base de dados, de aprendizagem de algoritmos, de reconhecimento de padrões e de redes neurais. A aplicação de tal conceito, tem permitido a transformação de grandes volumes de informação em conhecimento válido e útil, através da descoberta de verdades fundamentais em dados aparentemente aleatórios [Sumathi & Sivanandam 2006].

Uma das maiores dificuldades para as empresas na aplicação de técnicas de mineração de dados, é identificar quais os dados que se apresentam em condições favoráveis para a prática da mineração de dados. A mineração de dados necessita de uma fonte de dados consistente, separada, integrada e limpa. De acordo com esses requisitos, um *data warehouse* é o sistema de dados ideal para fornecer a informação para a aplicação de técnicas de mineração de dados. Tal pode-se justificar pelas seguintes razões:

- A consistência e a qualidade dos dados é um pré-requisito para a mineração, de maneira a assegurar a precisão dos modelos de previsão. Os *data wareshouses* são povoados com dados limpos e consistentes [Connolly & Begg 1998]. Desta forma, a mineração pode ser concentrada apenas na mineração dos dados em si, e não na limpeza e na integração dos dados de forma exclusiva [Inmon 1996].
- A obtenção de dados de diversas fontes para a mineração, possibilita a descoberta de tantas inter-relações quanto possíveis, que consequentemente remetem para um conhecimento mais elevado [Connolly & Begg 1998] – com sabemos, um *data warehouse* pode conter dados provenientes de diversas fontes operacionais.
- Um *data warehouse* assegura a existência dos dados de uma forma detalhada e sumarizada, através das suas capacidades [Connolly & Begg 1998]. A apresentação dos dados de forma detalhada é necessária, quando o objectivo da mineração de dados é examinar os dados na sua forma mais granular, e possivelmente, retirar inúmeros padrões escondidos, de grande relevância. A sumarização dos dados assegura que não seja necessário repetir a realização de análises (anteriormente realizadas por outra pessoa) no início de cada processo de exploração. Isto permite salvaguardar o trabalho excessivo em análises repetitivas, mesmo por parte de outros utilizadores [Inmon 1996].
- Os resultados da aplicação de técnicas de mineração de dados são bastantes úteis, apenas se existir uma maneira de investigar os padrões descobertos [Connolly & Begg 1998]. Isso só é possível através da análise histórica dos dados, que permite a compreensão sazonal do comportamento dos dados [Inmon 1996]. Os *data warehouses* são responsáveis por guardar informação histórica dos dados, permitindo

assim detectar tendências e padrões de comportamento ao longo do tempo, através da aplicação da mineração de dados.

Assim, num contexto geral, um *data warehouse* é um repositório de dados, especialmente orientado para suporte à decisão e que armazena dados provenientes de diferentes fontes operacionais, dando a possibilidade às organizações de tomarem decisões de acordo com a informação existente [Connolly & Begg 1998]. É construído sob uma perspectiva de armazenamento de informações a longo prazo, isto é, um *data warehouse* mantém a informação histórica à medida que existem alterações nas bases de dados.

Na Figura 2 podemos observar a estrutura típica de uma arquitectura necessária à aplicação de técnicas de mineração de dados. Neste caso, existem 2 fontes de dados e dados provenientes de fontes externas (e.g. documentos), que após um tratamento específico serão integrados no *data warehouse*. Os dados existentes no *data warehouse* podem ser acedidos de forma automática por ferramentas de análise e exploração de dados, como a mineração de dados, ou podem ser modelados em estruturas OLAP (*Online Analytical Processing*), e só depois acedidos por ferramentas de mineração de dados.

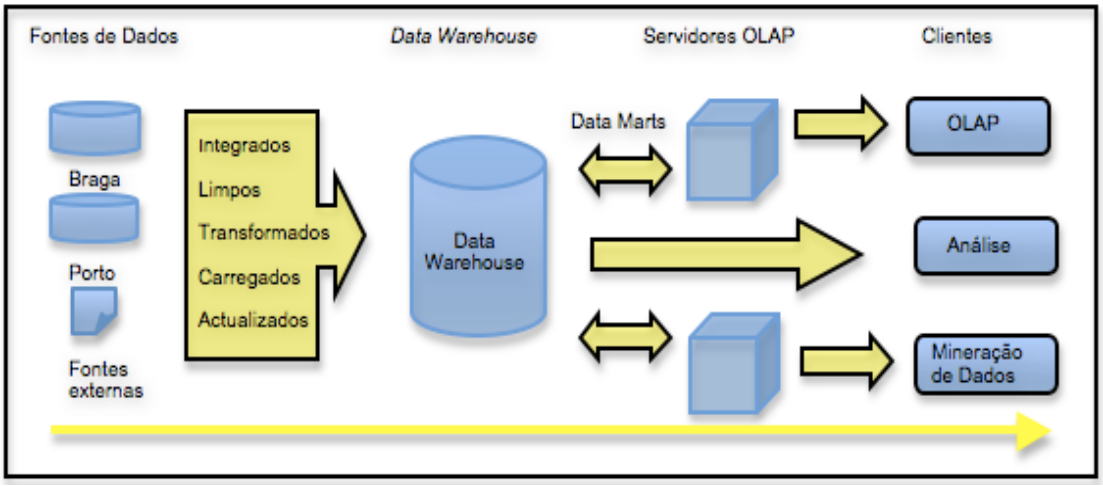


Figura 2 – Estrutura típica de uma arquitectura necessara à aplicação de técnicas de mineração de dados (Han & Kamber 2006).

Tal como os *data warehouses*, que facilitam o sucesso da aplicação da mineração de dados, os hipercubos também facilitam a utilização da mineração de dados. No entanto, existem algumas diferenças entre aplicação da mineração de dados a estruturas de *data warehouse*, e entre a aplicação a estruturas OLAP. Por exemplo, as dimensões de um hipercubo proporcionam a análise através de diversas prespectivas materializadas no cubo e as medidas são os valores que substanciam essas análises. Mais concretamente, significa que,

relativamente à análise dos dados de uma organização, através de um hipercubo é possível analisar os dados referentes a um departamento específico separadamente, enquanto que um *data warehouse* é referente aos dados da organização em si. Concluindo, os dados de um hipercubo apresentam um nível elevado de sumarização, quando em comparação com os dados existentes no *data warehouse*. Como tal, a mineração de dados aplicada a estruturas OLAP, pode ser considerada como uma análise secundária e mais pormenorizada dos dados, enquanto que a aplicação a sistemas de *data warehousing* pode ser considerada como a primeira análise, devido à natureza mais elementar encontrada nos dados nos *data warehouses* [Inmon 1996].

1.3 Motivação

A globalização dos serviços de prescrição electrónica, possibilitam hoje o armazenamento de elevadas quantidades de dados relativas às características das consultas realizadas. A introdução do registo electrónico na prática clínica permite aumentar a eficiência organizacional e generalizar o acesso e a troca de informação, bem como a sua recolha e análise. Os estudos existentes nesta área sugerem que estes sistemas aumentam a segurança dos utentes e diminuem os gastos, tantos hospitalares, como de todos os outros sistemas de saúde associados [Tomé *et al*. 2008].

A redução dos erros de registos, sejam eles devidos a ilegibilidade ou mesmo falta de dados, é uma das maiores vantagens da introdução destes sistemas na prática clínica. No entanto, as suas vantagens vão bastante além da redução dos erros de registo, sendo ainda visíveis em diversas áreas [Tomé *et al*. 2008]:

- **Nas consultas.** Os registos electrónicos para além de reduzirem os erros de registos, diminuem o tempo perdido com procura de registos passados, eliminam a necessidade de criar novas fichas e registos por perda dos anteriores, aumentam a fiabilidade dos registos (completos e legíveis), permitem o cruzamento de vários tipos de informação clínica e de base de dados que originam, se for o caso, alertas em tempo real relativas a possíveis erros de dosagem ou potenciais reacções alérgicas ou tóxicas, o que origina melhores resultados terapêuticos.
- **Na qualidade dos serviços prestados.** Relativamente a este área, a prescrição electrónica aboliu a falsificação de receitas, possibilitou a avaliação de parâmetros de qualidade e transmissão de informações directas entre outras instituições e mesmo farmácias, que por sua vez estimula a melhoria da qualidade dos serviços.

- **Nas instituições.** Devido à falta de necessidade de criar novas fichas e registos por perda dos anteriores e à maior fiabilidade dos registos, a prescrição electrónica possibilita a diminuição global dos custos de prescrição e dos custos dos exames complementares de diagnóstico. Possibilita ainda a diminuição de custos em processos judiciais, pois a informação encontra-se melhor disponibilizada, e consequentemente aumenta a produtividade e a eficiência. Isto significa, que existe um retorno rápido do investimento feito na instalação destes sistemas.

Existem já alguns estudos realizados relativamente à prescrição electrónica em Portugal, mas, além de maioritariamente dos estudos serem focados ao centro de Portugal e à região de Lisboa, são também, a maior parte deles, relativos à prescrição de antibióticos. Relativamente ao norte do país os estudos são também bastante limitados, e também aqui os estudos realizados são relativos à prescrições de antibióticos [Carvalho 2008]. Consequentemente, não se sabe se o padrão de prescrições varia consoante o sexo e idade do utente, tempo ou local da prescrição.

Através da aplicação SAM (Sistema de Apoio ao Médico) instalada nos centros de saúde da região norte, é possível monitorizar a prescrição realizada pelos médicos nas consultas aos utentes dos centros de saúde, e suas extensões e unidades de saúde familiares associadas, possibilitando assim uma análise efectiva da actividade clínica realizada. Consequentemente, a ausência de estudos nesta área (principalmente relativas à região norte) e a existência do sistema SAM, justifica a realização de um estudo, que tem como objectivo principal descobrir padrões de prescrições de medicamentos, existentes na região norte, com base em eixos de análise temporais, locais e características dos utentes. Este estudo, é ainda complementado com a descoberta de possíveis associações entre as prescrições de medicamentos de respectivos laboratórios, por determinados médicos.

1.4 Organização da Dissertação

Além do presente capítulo, esta dissertação está organizada em mais três, nomeadamente:

- O capítulo 2 é relativo à revisão bibliográfica realizada para esta dissertação. É abordado os diferentes tipos de padrões existentes e é explicado três técnicas mais usuais na descoberta de padrões: a associação, a segmentação e a classificação. No final, é apresentado para cada uma das técnicas, as diferenças essenciais entre alguns dos diferentes algoritmos existentes. A grande extensão deste capítulo remete para a necessidade de estudo extensivo que foi necessária para a realização desta

dissertação, como tal, o seu conteúdo baseia-se essencialmente em informação retirada e assimilada de diversas referências bibliográficas.

- No capítulo 3 é apresentado um caso prático de estudo real, focalizado na descoberta de padrões, num conjunto de dados, respectivo à prescrição de medicamentos. São realizados 3 estudos diferentes para cada uma das técnicas abordadas no capítulo 2, com objectivos finais diferentes, mas complementares. Este capítulo termina com uma pequena análise dos resultados obtidos para cada uma das técnicas e possíveis modificações dos parâmetros.

- No capítulo 4 são mencionadas as conclusões da realização desta dissertação e referenciados alguns possíveis e futuros estudos.

A metodologia seguida na realização desta dissertação, cobriu todos os passos essenciais e necessários na aplicação das técnicas de mineração de dados a um determinado caso de estudo real. Cada passo foi cuidadosamente detalhado e referenciado, com o objectivo de se obter o melhor resultado possível e que fosse de encontro com as expectativas da organização responsável pela facultação dos dados.

Capítulo 2

Modelos e Padrões

A mineração de dados é um termo que engloba uma grande diversidade de técnicas, que permite inferir características interessantes a partir de um grande volume de dados. No entanto, a informação retirada através dessas técnicas podem aparecer em diferentes formatos. Consequentemente, é necessário identificar quais as questões que se pretendem descobrir e, a partir daí, escolher a técnica que melhor responde à pergunta em causa [Laxman 2006]. De uma maneira geral, estas respostas são estruturas, que diferem de acordo com os objectivos que se pretendem, com a aplicação de técnicas de mineração de dados. Estes objectivos, são usualmente, denominados como de tarefas da mineração de dados (*tasks of data mining*), e que, por sua vez, podem ser divididas em 5 grupos gerais [Hand *et al*. 2001, Laxman 2006].

1. Análise Exploratória dos Dados. Como o próprio nome indica, o objectivo é explorar os dados através de ferramentas interactivas. Normalmente, esta análise consiste na visualização gráfica dos dados, o que no entanto, torna bastante difícil a análise quando é feita para dimensões superiores a 3.

2. Modelos Descritivos. O objectivo de um modelo descritivo é descrever os dados. A ideia básica é agrupar os registos similares através de medidas de similaridade entre registos.

3. Modelos de Previsão. Nos modelos de previsão, o objectivo é prever o valor de um atributo, em função dos outros. A principal distinção entre os modelos de previsão e descritivos, é que os modelos de previsão tem como objectivo um único atributo (o de previsão), enquanto que os descritivos não são apenas focados num atributo central.

4. Descoberta de Padrões e Regras. As tarefas especificadas acima centram-se na construção de modelos. No entanto, existem outras aplicações da mineração de dados que se centram na descoberta de padrões. Um desses exemplos, é a descoberta de potenciais fraudes com base nos padrões de comportamento dos dados.

5. Recuperação por conteúdo. O objectivo aqui é descobrir padrões similares no conjunto de dados, com base num padrão de interesse definido pelo utilizador. Esta tarefa é mais usada para conjuntos de dados de texto ou imagens.

Como é possível constatar pelo parágrafo acima, as diferentes estruturas resultantes da aplicação de técnicas de mineração de dados, cingem-se a duas estruturas globais, padrões e modelos [Hand *et al*. 2001, Fayyad *et al*. 1996a, Fayyad *et al*. 1996b, Laxman 2006]. Um modelo é uma representação global e abstracta dos dados, isto é, são basicamente um conjunto de parâmetros que, na íntegra, possibilitam a extracção de informação útil, com base na análise do sistema de dados [Hand *et al*. 2001, Laxman 2006]. Dando um exemplo no negócio do retalho, através da criação de modelos, é possível associar um certo produto a um conjunto de pessoas com certas características, ou prever qual o conjunto de pessoas que poderão optar pela compra desse produto.

Um exemplo de um modelo linear simples é $Y = aX + b$, em que Y é uma função linear de X. Na estatística convencional, um modelo é linear se for uma função linear dos parâmetros (a e b), mas no contexto de um modelo de mineração de dados, só faz sentido considerar a linearidade como sendo uma função entre as variáveis de interesse (Y e X) e não entre os parâmetros [Hand *et al*. 2001]. Consequentemente, num modelo, é necessário estimar os valores apropriados para os parâmetros, de maneira a maximizar ou minimizar o resultado da função, para melhor se ajustar o modelo aos dados em estudo. Os modelos podem ser classificados em duas categorias principais, modelos de previsão e modelos descritivos [Hand *et al*. 2001]. Os modelos de previsão são modelos que tem como base a previsão da variável de resposta, Y, em função da variável exploratória, X. Deste modo, é possível construir um modelo para prever a probabilidade de uma pessoa se encontrar engripada ou não (por exemplo), baseado nos seus sintomas e no estudo do comportamento de outros utentes existentes na base de dados. O registo de um dado utente i no passado, pode ser representado por $\{(x(i), y(i))\}$. Em que $y(i)$ é resultado da classe (doente ou não doente) do utente i, e $x(i)$ é o vector $x = (x_1(i),...,x_p(i))$ que representa os sintomas do utente i. O modelo suporta previsões do tipo $y = (x_i,...,x_p;\alpha)$ em que y é a previsão do modelo e α representa os parâmetros da estrutura do modelo. Quando o Y é representado por valores numéricos, a técnica usada para estimar Y em função das p-dimensões de X, é a

regressão. Por sua vez, quando o Y é representado por um valor discreto, o processo de aprendizagem para prever Y em função das p-dimensões de X, é denominado de classificação. Ambas as técnicas podem ser consideradas como problemas de aproximação de funções, em que o objectivo é através do estudo das variáveis p-dimensionais de X, descobrir Y [Hand *et al*. 2001]. Os modelos descritivos são, basicamente, modelos que tem como objectivo produzir uma descrição dos dados. Neste contexto existem duas possibilidades. Se os dados disponibilizados já forem os finais, então nenhuma dedução terá relevância e, consequentemente, o objectivo é meramente a simplificação da descrição. No entanto, se os dados fornecidos pertencerem apenas a uma parte do total dos dados, ou se apresentarem algum erro de medida (isto quer dizer quando forem novamente recolhidos apresentarão diferentes valores), o objectivo principal será obter uma dedução sobre os resultados, se o modelo estruturado é considerado bom, por exemplo [Hand *et al*. 2001].

Em contraste com os modelos, os padrões são estruturas locais que proporcionam afirmações específicas sobre uma pequena parte dos dados, ou espaço em que os dados possam ocorrer [Laxman 2006, Hand *et al*. 2001]. Numa pesquisa pela base de dados pode-se, por vezes, encontrar registos que apresentem o mesmo tipo de comportamento [Hand *et al*. 2001], esse conjunto de registos é caracterizado como sendo um padrão de comportamento. Por exemplo, através da pesquisa numa base de dados de compras de livros, constata-se que quem compra o livro "Equador", do autor Miguel Sousa Tavares, muitas vezes também compra o segundo romance, "Rio das Flores", escrito pelo mesmo autor, isso representa um padrão de comportamento por parte de algumas pessoas.

Apesar de a diferença entre modelos e padrões ser bastante importante no contexto da comparação e categorização dos algoritmos de mineração de dados, facilmente se constata que por vezes existem casos em que a categorização de uma estrutura, em padrão ou em modelo, pode não ser muito clara [Laxman 2006, Hand *et al*. 2001]. Contudo, este trabalho não é direccionado nessa vertente, mas sim para a descoberta de padrões. Na secção seguinte será feita uma descrição mais abrangente sobre a descoberta de padrões e sobre os diversos tipos de padrões frequentes, bem como quais os algoritmos que permitem a sua descoberta.

2.1 Descoberta de Padrões

Ao contrário dos outros objectivos especificados no inicio deste capítulo - como a classificação (modelos de previsão), a segmentação (modelos descritivos) e a comparação de padrões (recuperação por conteúdo) -, em que as suas origens provêm de diversas áreas,

como a aprendizagem de algoritmos e o reconhecimento de padrões, as origens da descoberta de padrões provém da própria mineração de dados em si. Neste contexto, a descoberta de padrões, através da sua natureza exploratória e não supervisionada, é considerado o ponto fulcral da mineração de dados [Laxman 2006]. Como foi referido anteriormente, padrões são estruturas locais existentes nos dados, cujo objectivo principal objectivo é descobrir todos os padrões de interesse que poderão existir num dado conjunto de dados. Um conceito bastante importante na mineração de dados e referente à descoberta de padrões, é a descoberta de padrões frequentes. Padrões frequentes, são padrões que ocorrem com bastante regularidade nos dados. É no estudo, do desenvolvimento de algoritmos que permitem a descoberta desse tipo de padrões e na formulação de estruturas de padrões úteis, que se tem centrado as pesquisas na área da mineração de dados. A descoberta de padrões frequentes é considerada bastante importante, pois através dessa descoberta é possível descobrir importantes regras, que permitem deduzir alguns comportamentos habituais, de bastante interesse, nos dados [Laxman 2006].

Uma regra consiste num par de afirmações ou preposições booleanas (verdadeiro ou falso) acerca de acontecimentos no mundo. Esse par é composto por uma preposição do lado esquerdo (o antecedente ou condição) e por uma preposição do lado direito (consequente) [Laxman 2006, Hand *et al*. 2001]. Uma regra obriga a quando o lado esquerdo é verdadeiro, então o lado direito também é considerado verdadeiro (exemplo: "Se está a nevar, logo está frio"). Mas numa regra probabilística o conceito é um pouco modificado juntando a noção de probabilidade à preposição do lado direito. Mais concretamente, o lado direito é considerado verdadeiro com uma probabilidade de p, quando ocorre a veracidade do lado esquerdo [Hand *et al*. 2001]. As regras assumem um grande papel na representação da informação nos algoritmos de aprendizagem e na inteligência artificial. Na classificação, as árvores de decisão, podem ser referenciadas como um caso específico do uso da aprendizagem de um conjunto de regras, com o intuito de criar um modelo de previsão. Nesta técnica, as condições dos nodos existentes, ao longo dos ramos de cada folha, podem ser consideradas como um conjunto de afirmações que representam o lado esquerdo de uma regra, e em que o lado direito da regra é a classe atribuída à folha [Hand *et al*. 2001]. As regras são usualmente apresentadas em formato discreto, isto é, ambas as afirmações, do lado esquerdo e do lado direito, são booleanas. Consequentemente, as regras são particularmente boas para a modelação de variáveis discretas, pois esses tipos de variáveis são facilmente mapeadas em valores booleanos. Mas mesmo assim, é possível expandir a utilização de regras para valores contínuos, através da discretização desses valores em intervalos (exemplo: "Se $a > 30.5$ então $b < 2$"). É através da discretização de valores que as árvores de classificação manipulam a existência de valores contínuos [Hand *et al*. 2001]. Existem ainda duas técnicas em que a sua base recai na descoberta de padrões, são elas a associação

e a segmentação. Ambas as três técnicas serão exploradas ao pormenor neste capítulo. Mas para se compreender o funcionamento destas técnicas, é necessário, primeiro perceber o funcionamento da descoberta de padrões frequentes. Como tal, será feita uma primeira abordagem à noção de padrões frequentes e será descrito o algoritmo base na descoberta de padrões frequentes. Mas ao longo do estudo da descoberta de padrões frequentes, surgiu a necessidade de especificar melhor o tipo de padrões que se procura. Consequentemente a descoberta de padrões frequentes foi ramificada em dois tipos de classes de padrões principais, padrões sequenciais e episódios frequentes. Cada uma das classes de padrões será abordada bem como as extensões feitas ao algoritmo inicial da descoberta de padrões frequentes.

2.1.1 Padrões Frequentes

Como foi mencionado na secção anterior, a existência de padrões frequentes acontece quando variáveis de um conjunto em estudo ocorrem frequentemente juntas. O funcionamento da descoberta de padrões frequentes é bastante simples: atribuindo um suporte de limite mínimo s, o objectivo é encontrar todos os conjuntos de padrões frequentes que respeitem s e assim sucessivamente [Hand *et al.* 2001]. Aplicando estes conceito à Tabela 1, e atribuindo um $s = 2$, verificamos que os conjuntos frequentes são {X}, {Y}, {Z} e {XY}, pois são os únicos que respeitam o suporte mínimo.

Clientes/Produto	X	Y	Z
Cliente 1	1	0	1
Cliente 2	1	1	0
Cliente 3	1	1	0
Cliente 4	0	1	1
Cliente 5	1	1	0

Tabela 1 – Exemplo de Compras por Clientes.

Através da descoberta destes conjuntos frequentes é possível construir a regra $X \Rightarrow Y$ com uma confiança de $\frac{3}{4} = 0.75$. A confiança é uma medida de avaliação que determina a precisão de uma regra de associação, sendo obtida através do número de casos que cobrem a regra sobre o número de casos que cobrem o antecedente. Para este caso temos $conf(X \Rightarrow Y) = \frac{s(X \Rightarrow Y)}{s(X)}$ [Hand *et al.* 2001].

Como foi referido anteriormente, a descoberta de padrões está bastante associada à criação de regras, isto quer dizer que a descoberta de padrões é feita através dos algoritmos de descoberta de regras de associação. A descoberta das regras de associação é feita através da descoberta de todas as regras que preencham os critérios mínimos dos valores, previamente especificados, do suporte e da medida de avaliação em causa. Constata-se então, que essa descoberta é bastante simples se todos os padrões frequentes tiverem sido encontrados. Uma maneira trivial de encontrar padrões frequentes é calcular o suporte de todos os padrões [Hand *et al*. 2001], mas obviamente isto seria um processo bastante demoroso, pois quando falamos de uma base de dados de grandes dimensões, é possível encontrar um vasto número de padrões existentes e, no final, chegar à conclusão que a descoberta de maior parte deles foi inútil. Então, a principal observação que se deve ter em conta na descoberta de padrões frequentes, é que um conjunto X de variáveis pode ser frequente somente se os subconjuntos de X forem frequentes. Esta observação subentende que não é necessário descobrir o suporte de nenhum conjunto X, se houver algum subconjunto de X que não respeite o valor do suporte mínimo [Hand *et al*. 2001]. Consequentemente, é possível encontrar todos os conjuntos frequentes através da aplicação de um algoritmo que, inicialmente, comece por procurar os conjuntos frequentes com uma variável. Assumindo a realização deste passo inicial, é necessário de seguida criar conjuntos candidatos de duas variáveis, {A,B}, tal que {A} e {B} sejam frequentes. Depois de criados todos os candidatos o algoritmo repete-se, isto é, para cada um dos conjuntos candidatos é necessário descobrir quais são os verdadeiramente frequentes. Tendo sido encontrado todos os conjuntos frequentes de duas variáveis, o passo seguinte é criar conjuntos com três variáveis e proceder-se à descoberta dos conjuntos realmente frequentes de três variáveis e assim sucessivamente. Este algoritmo é conhecido por *Apriori* [Agrawal *et al*. 1993], e foi o primeiro algoritmo a ser apresentado com o intuito de gerar todas as regras de associação significativas entre os registos de uma bases de dados. Este algoritmo foi apresentado em 1993, na conferência internacional de gestão de dados [Agrawal *et al*. 1993]. Na Figura 3, é apresentado um exemplo do funcionamento deste algoritmo com base no registo de compras de quatro clientes, em que o suporte mínimo imposto foi 2. Este exemplo serve para mostrar como são os candidatos formados e de que maneira o suporte de cada candidato é gerado. Por exemplo, os candidatos da iteração I_2 foram gerados através da união dos termos frequentes da iteração I_1_1. De notar, que a iteração I_4 nunca seria gerada pois um subconjunto do termo {1,2,4,5} é {1,2,4}, e como é possível observar na iteração I_3_1 esse termo não é considerado um termo frequente, logo {1,2,4,5} também não é considerado um termo frequente. Este exemplo ressalva bem a ideia que um termo só poderá ser frequente se todos os seus subconjuntos forem frequentes, e essa é a ideia base deste algoritmo. Mas para saber se um termo candidato é frequente é necessário saber o seu suporte. O cálculo do suporte de um termo é basicamente um *count* das vezes que esse termo aparece nos

registos da base de dados. Por exemplo, para o termo {1,4} (iteração I_2_1) o suporte é 2, pois os únicos clientes que fizeram essa compra em conjunto foram os clientes 2 e 4. O processo de gerar progressivamente conjuntos de termos de tamanho superior é continuado até a um ponto em que não exista mais nenhum termo frequente, e isso marca o final do processo da descoberta de padrões frequentes [Agrawal *et al*. 1993].

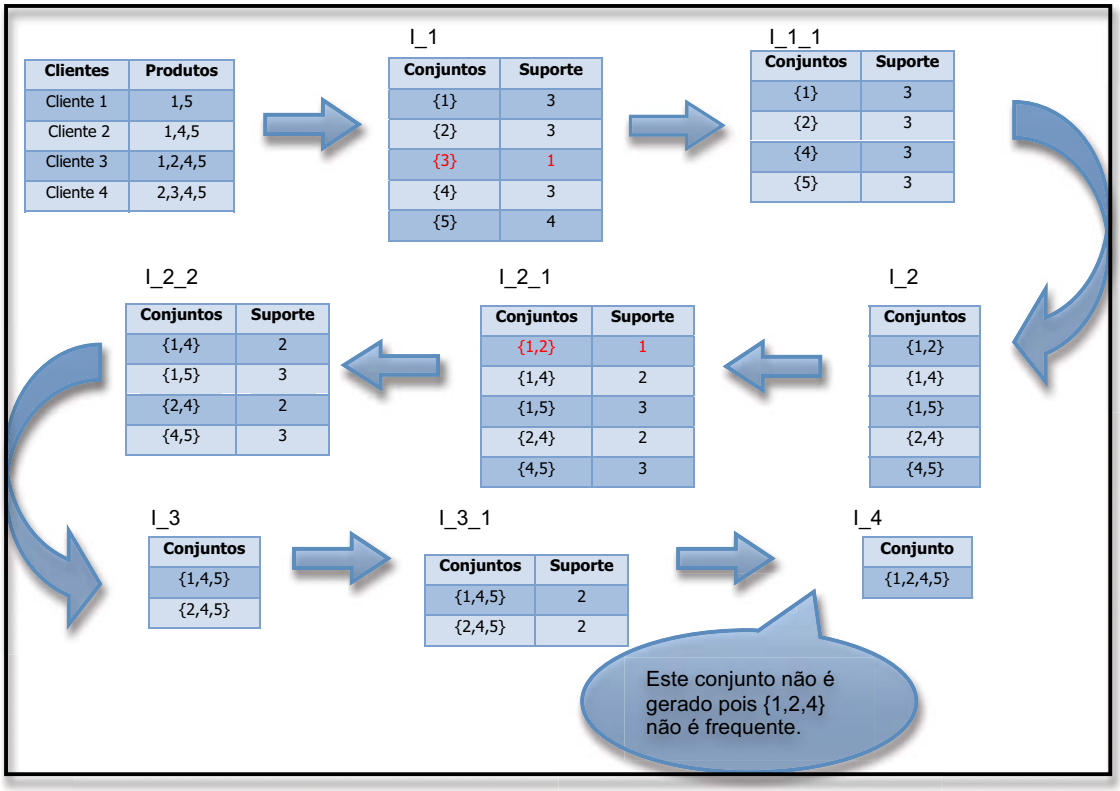


Figura 3 – Exemplo do funcionamento do algortimo *Apriori*.

2.1.2 Padrões Frequentes Sequenciais

A primeira abordagem de padrões sequenciais foi feita em 1995 por *Agrawal* e *Srikant* [Agrawal & Srikant 1995], mostrando que por vezes existem padrões que acontecem sequencialmente, e a sua descoberta pode ser uma mais valia na avaliação de comportamentos [Agrawal & Srikant 1995]. Um exemplo de um padrão sequencial pode ser a compra de DVDs de séries televisivas. Isto significa que, normalmente, uma pessoa quando se interessa por uma determinada série, tem tendência a comprar primeiro a temporada 1, depois a temporada 2, e assim sucessivamente. Estas compras não precisam de ser necessariamente consecutivas, pode haver compras de outras séries diferentes no entretanto [Agrawal & Srikant 1995]. A questão essencial deste tipo de padrões é a compra sequencial das temporadas da mesma série (por exemplo), pois uma pessoa muito dificilmente começa pela última temporada para chegar à primeira temporada.

O algoritmo utilizado na descoberta deste tipo de padrões é uma extensão ao *Apriori* [Agrawal *et al*. 1993], e proposto pelo mesmo *Agrawa*l e *Srikant* [Agrawal & Srikant 1995]. Neste caso a base de dados deixa de lado o conceito de um conjunto de registos desordenados, e aborda o conceito de uma base de dados ordenada, associando a cada compra uma etiqueta temporal e o respectivo cliente. Na realidade, existe uma base de dados com sequências de compras, em que cada sequência é uma lista de compras, ordenada por tempo. Para exemplificar este tipo de padrões é usado o exemplo proposto na secção de padrões frequentes, mas, agora, foi acrescentado um registo temporal para se ter a noção temporal da sequência das compras.

Compras/Clientes	Dia	Produtos
Cliente 1	1 de Maio	X
	5 de Maio	Y
Cliente 2	2 de Maio	W,O
	3 de Maio	X
	5 de Maio	T,Z,P
Cliente 3	29 de Abril	X,M,P
Cliente 4	29 de Abril	X
	3 de Maio	T,P
	6 de Maio	Y
Cliente 5	2 de Maio	Y

Tabela 2 – Exemplos de produtos comprados por clientes por dia.

Como é possível observar na Tabela 2, existe um conjunto de compras associadas a cada cliente e identificadas pelo dia, num espaço de tempo entre 29 de Abril a 6 de Maio. Através destes registos, criou-se uma tabela que apresenta a sequência das compras por cada cliente (Tabela 3). Uma sequência s de um conjunto de elementos é representada por $\langle s_1 s_2 \ldots s_n \rangle$, em que s_j é um conjunto de elementos. Como s é um conjunto de n-elementos, então é denominada por uma n-sequência. Por exemplo, o cliente 1 tem uma 2-sequência de compras.

Clientes	Sequência de Compras Original
Cliente 1	<(X),(Y)>
Cliente 2	<(W,O),(X),(T,Z,P)>
Cliente 3	<(X,O,P)>
Cliente 4	<(X),(T,P),(Y)>
Cliente 5	<(Y)>

Tabela 3 – Sequência de compras de cada cliente.

Um padrão sequencial só pode ser tido em conta se for considerado como largo e máximo [Hand *et al*. 2001]. Uma sequência é considerada larga se apresentar um suporte igual ou superior ao suporte mínimo indicado, neste caso, foi atribuído um suporte mínimo de 2. Para ser máxima, uma sequência larga, não pode estar contida noutra sequência larga, por exemplo, a sequência larga $<(T,P)>$, do cliente 4 está contida na sequência larga $<(X),(T,P)>$, do cliente 2. Consequentemente, a sequência $<(T,P)>$, apesar de ter um suporte igual ao suporte mínimo, não pode ser considerado um padrão sequencial, pois está contido na sequência $<(X),(T,P)>$ Com base nestes dois critérios, apenas sobram dois padrões sequenciais, que são $<(X),(Y)>$ e $<(X),(T,P)>$ (Tabela 4), pois são os únicos que além de respeitar o suporte mínimo também são considerado máximo. Todos os outros elementos da Tabela 4, coloridos a vermelho, apesar de respeitarem o suporte mínimo, não são considerados máximos, logo não são considerados padrões sequenciais.

Sequências	Suporte
<(X)>	4
<(Y)>	3
<(T)>	2
<(P)>	3
<(T,P)>	2
<(X),(Y)>	2
<(X),(T)>	2
<(X),(P)>	2
<(X),(T,P)>	2

Tabela 4 – Sequências de suporte igual ou superior a 2.

De uma maneira geral, o mecanismo da descoberta de padrões sequenciais começa com a descoberta de todos os possíveis conjuntos com suporte mínimo (fase de transformação). Nesta fase é possível usar o algoritmo *Apriori* [Agrawal *et al*. 1993], descrito na secção anterior, apenas com uma pequena alteração no cálculo do suporte. No algoritmo descrito anteriormente o suporte é definido pelo número de compras em que o conjunto de produtos aparece, em todas as compras existentes. Neste caso, o suporte é o número de clientes que compraram a sequência de produtos em pelo menos uma das suas compras [Agrawal & Srikant 1995]. Uma vez descobertas todas as sequências de produtos consideradas largas, uma espécie de nova base de dados é criada, em que cada compra é substituída pelo conjunto de sequências largas contidas nessas compras (Tabela 5). Estas sequências são consideradas as sequências candidatas.

Clientes	Sequência de Compras Transformada
Cliente 1	<{(X)},{(Y)}>
Cliente 2	<{(X)},{(T),(P),(T,P)}>
Cliente 3	<{(X)},{(P)}>
Cliente 4	<{(X)},{(T),(P),(T,P)}, {(Y)}>
Cliente 5	<{(Y)}>

Tabela 5 – Sequências de compras transformadas.

Após a fase de transformação, o passo seguinte é designado por fase sequencial. Nesta fase, cada passo é iniciado com um conjunto de sequências largas consideradas como sementes. Estas sementes são usadas para gerar potenciais sequências novas e maiores, denominadas de sequências candidatas. A essas sequências candidatas é calculado o seu suporte e as que forem consideradas largas serão consideradas as sementes para o próximo passo. De notar, que as sementes do primeiro passo são as sequências de tamanho 1 que respeitam o suporte mínimo [Agrawal & Srikant 1995]. Para esta fase foram apresentadas duas famílias de algoritmos por *Agrawal* e *Srikant* [Agrawal & Srikant 1995]: o "*Count-all Algorithm*" e o "*Count-Some Algorithms*". O "*Count-all Algorithm*" conta, incialmente, todas as sequências largas e depois retira todas as sequências consideradas não máximas num passo de pós-processamento. Este algoritmo é também baseado no ideia do *Apriori* para a contagem de padrões frequentes. Num primeiro passo todos as sequências largas de tamanho 1 são descobertas e através dessas são calculadas as sequências de tamanho 2, combinando, de todas as maneiras, as sequências largas de tamanho 1 existentes. De seguida o processo é repetido e por aí adiante [Agrawal & Srikant 1995]. O algoritmo "*Count-Some Algorithms*" funciona de uma maneira mais inteligente, pois explora a ideia que um padrão sequencial só pode ser considerado se for máximo. Como a pesquisa é feita só para sequências máximas, é possível evitar contar as sequências que se encontram contidas em sequências maiores. Para isso é necessário verificar primeiro as sequências de maior dimensão. Deste modo, este algoritmo tem uma fase avançada que consiste em descobrir todas as sequências frequentes de um determinado tamanho, e só depois é descoberto as sequências frequentes que sobram [Agrawal & Srikant 1995]. Naturalmente este algoritmo apresenta uma contra partida. Se foram descobertas muitas sequências de grande dimensão que não respeitem o suporte mínimo, o tempo ganho na exploração da restrição de maximização de sequências, pode ser subvertido pelo tempo perdido na descoberta de sequências que não apresentem o suporte mínimo [Hand *et al*. 2001].

Estes algoritmos para a descoberta de padrões sequenciais são bastante eficientes e são usados na maioria das aplicações temporais da mineração de dados. Apesar disso foram feitas extensões a estes algoritmos [Shintani & Kitsuregawa 1998, Zaki 1998, Lin & Lee 2003], mas o desempenho destes novos algoritmos é inferior quando os dados apresentam

longas sequências com um suporte igual ao suporte mínimo, ou quando o suporte mínimo definido é baixo. Uma maneira de reverter este problema, é não só procurar por sequências consideradas largas mas também por sequências consideradas fechadas. Uma sequência é considerada frequente e fechada se respeitar o suporte mínimo e caso esteja contida noutra sequência considerada frequente, essa sequência (a não contida) tem que apresentar um suporte inferior à sequência considerada fechada [Pasquier *et al*. 1999]. Esta ideia foi introduzida em 1999 por Pasquier *et al*, [Pasquier *et al*. 1999] e entretanto foram propostas técnicas para a descoberta de padrões sequenciais fechados [Yan *et al*. 2003, Wang & Han 2004].

2.1.3 Episódios Frequentes

Na descoberta de episódios frequentes, os dados são fornecidos no formato de uma sequência longa de eventos, em que o objectivo é a descoberta de padrões temporais (episódios) que ocorrem com alguma frequência ao longo da sequência fornecida [Mannila *et al*. 1997]. Esta estrutura foi originalmente aplicada na análise da chegada de alarmes numa rede de telecomunicações [Mannila *et al*. 1997]. Existem várias tipos de alarmes que são accionados por diferentes acontecimentos numa rede de telecomunicações. É esses acontecimentos que esta estrutura procura encontrar. A descoberta de episódios frequentes tem utilidade noutro tipo de conjuntos de dados, como nos *logs* de navegação *Web* [Mannila *et al*. 1997], e no registo de vendas do Wal-Mart [Atallah *et al*. 2004], por exemplo.

Uma sequência de eventos é representada por $\langle (E_1,t_1),(E_2,t_2),...,(E_n,t_n)\rangle$, em que n é o número de eventos na sequência de dados. Em cada evento (E_i,t_i), E_i representa o tipo de evento e t_i representa o tempo da ocorrência do evento. Um exemplo de uma sequência de 10 eventos é representada na Equação (1) [Laxman 2006]:

$$\langle (A,1),(B,3),(B,7),(C,8),(D,10),(D,11),(A,13),(C,14),(B,15),(D,18)\rangle \quad (1)$$

Um episódio, é considerado um conjunto de grupos frequentes de tipos de eventos, que respeitam algumas regras relativas à ordem da sua ocorrência. Estes grupos de tipos de eventos são considerados padrões interessantes, que podem disponibilizar informação útil no que toca a correlações entre os tipos de eventos. [Laxman 2006]. Existem duas extensões aos episódios normais, os episódios em paralelo ou em série. Quando a sequência de eventos de um episódio se encontra ordenada cronologicamente, este episódio é considerado um episódio em série. Caso os eventos de um episódio não apresentem nenhuma ordem

concreta, então o episódio é considerado um episódio em paralelo [Mannila *et al*. 1997]. Os episódios podem ser descritos como gráficos acíclicos:

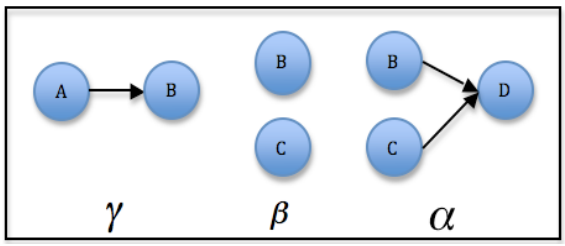


Figura 4 – Exemplos de diferentes tipos de episódios.

Na Figura 4 é demonstrado cada tipo de episódio. O grafo γ é um episódio em série pois o evento B só acontece se o evento A tiver acontecido antes. As setas servem para enfatizar a ordem da sequência. O grafo β demonstra um episódio em paralelo que diz que o evento B e o evento C ocorrem sem uma ordem especifica. Por fim, α é um episódio, mas que não é considerado nem em série, nem em paralelo, pois para ser em série, o evento C e o evento B não poderiam ser paralelos. Por outro lado, α não pode ser considerado um episódio em paralelo pois existe a noção de ordem sequencial, pois o evento D só poderá ocorrer se entretanto os eventos B e C tiverem ocorrido. Apesar de existirem estes três tipos de episódios, os estudos concentram-se mais na pesquisa de episódios em paralelos ou em série [Mannila *et al*. 1997]. Diz-se que um episódio ocorre numa sequência de eventos, se existirem eventos na sequência que ocorrem exactamente pela mesma ordem da existente no episódio. Por exemplo, na sequência (Equação (1)) os eventos (A,1), (B,3) e (D,10) constituem uma ocorrência do episódio em série $(A \rightarrow B \rightarrow D)$, enquanto que os eventos (D,11), (A,13) e (B,15) não constituem, pois para este tipo de episódio em série ocorrer, o B precisa de vir primeiro que o D. Ambos estes conjuntos de eventos são ocorrências válidas do episódio em paralelo (ABD), pois neste tipo de episódios não existem restrições quanto à ordem de ocorrência dos eventos [Laxman 2006].

Nos padrões sequenciais foi abordado a noção de uma sequência contida noutra. Uma noção similar para os episódios frequentes é a noção de sub episódios. Um episódio β é considerado um sub episódio de um episódio α, se todos os eventos de β ocorrem em α, e se todas as relações entre os eventos ocorridos em β estejam presentes em α. Sintacticamente esta relação é representado por $\beta \preceq \alpha$ [Mannila *et al*. 1997]. Por exemplo $(A \rightarrow D)$ é um sub episódio de tamanho 2 do episódio em série $(A \rightarrow B \rightarrow D)$, enquanto que $(D \rightarrow A)$ não é seu sub episódio. No caso de episódios em paralelo, esta restrição de

ordem não é imposta, por isso todos subconjuntos dos eventos de um episódio são considerados sub episódios.

Na descoberta de episódios frequentes, é imposto a ideia base do algoritmo Apriori em relação à frequência, isto é, a frequência de um episódio será sempre menor do que a dos seus sub episódios [Mannila et al. 1997]. Esta regra assegura que um episódio de tamanho n é considerado um episódio frequente candidato, apenas se todos os seus sub episódios de tamanho $n-1$ forem frequentes [Laxman 2006]. *Mannila et al.* [Mannila et al. 1997] definiu a frequência de um episódio como sendo o número de vezes que esse episódio ocorre numa janela de tamanho fixo. O tamanho da janela é pré-definido e consiste na criação de intervalos de tempo na sequência de eventos original [Hand et al 2001]. Aplicando este conceito à Equação (1) e definindo uma janela de tamanho 10, temos no total duas janelas. A primeira é $(\langle(A,1),(B,3),(B,7),(C,8)\rangle 1,10)$, em que 1 é o inicio da janela e o 10 é o fim, sendo o último evento de cada janela, o evento $t-1$ do especificado como final. O que significa que para a primeira janela o evento limite é aquele que apresentar um $t=9$. Para a segunda janela obtemos um conjunto de seis eventos, e neste caso o último evento terá de ser $t<19$, $(\langle(D,10),(D,11),(A,13),(C,14),(B,15),(D,18)\rangle 10,20)$.

Dada uma sequência de eventos, um tamanho de janela, e um limite de frequência, o passo seguinte é a descoberta de episódios frequentes numa sequência de eventos. O processo da descoberta de episódios frequentes é feito através de um algoritmo do estilo do *Apriori* que começa com a descoberta de episódio frequentes de tamanho 1. Estes episódios descobertos, num segundo passo, são combinados de maneira a formar episódios candidatos de tamanho 2, em que posteriormente é calculado a sua frequência, e consequentemente são extraídos os episódios frequentes de tamanho 2. Este processo continua até todos os episódios frequentes de todos os tamanhos serem descobertos [Mannila *et al*. 1997]. Tal como no *Apriori*, um episódio só é considerado um episódio candidato se todos os seus sub episódios forem considerados frequentes [Laxman 2006]. Mas como vimos anteriormente existem dois tipos de episódios, os episódios em paralelo ou em série. Apesar de o processo inicial de ambos ser igual (começam com o mesmo episódio inicial), o algoritmo que gera os episódios candidatos difere para cada tipo diferente de episódios [Mannila *et al*. 1997]. A contagem de episódios em paralelo é praticamente linear. Como os episódios em paralelos são praticamente considerados como um conjunto de elementos, então a contagem do número de janelas em que esse episódio ocorre não difere muito do cálculo do suporte de um conjunto de produtos de uma lista de compras por clientes [Mannila *et al*. 1997]. A contagem de episódios em série envolve outros conceitos. Isto acontece, porque ao contrário dos episódios em paralelo, é necessário a utilização de autómatos de estado finitos para reconhecer episódios em série [Mannila *et al*. 1997]. Mais especificamente, um determinado

autómato de estado de tamanho n pode ser usado para reconhecer as ocorrências de um episódio em série de tamanho n. Por exemplo, para o episódio $(A \rightarrow B \rightarrow D)$, tem de existir um autómato de estado de tamanho 3 que transita para o seu primeiro estado quando ocorre um evento do tipo A e que depois espera por um evento do tipo B para transitar para o estado seguinte, e assim sucessivamente. Quando este autómato transita para o estado final, o episódio é reconhecido na sequência de eventos e a sua frequência é incrementada [Laxman 2006]. Para cada episódio em que é necessário calcular a sua frequência, tem de existir um autómato associado.

A estrutura descrita até agora depende do cálculo da frequência de episódios por janela. No entanto, também em *Mannila et al.* [Mannila *et al*. 1997] é proposto outra alternativa baseada no conceito "*minimal ocurrences*" (ocorrências mínimas) de um episódio. Outras abordagens ao cálculo da frequência foram abordadas por *Casas-Garriga* [Casas-Garriga 2003] e posteriormente por *Laxman et al*. [*Laxman et al*. 2004, *Laxman et al*. 2005], em que estas últimas apresentavam já uma grande vantagem no tempo de execução, apesar da complexidade da contagem do tamanho da janela ser o mesmo do algoritmo exposto por *Mannila et al*. [Mannila *et al*. 1997].

2.2 Grau de Interesse de um Padrão

A mineração de dados tem o potencial de gerar milhões e milhões de padrões através da análise dos dados. Consequentemente, surge a questão se todos os padrões são interessantes. Obviamente, a resposta é não. Apesar dos inúmeros padrões descobertos apenas uma pequena parte é considerada como útil para o utilizador em causa [Silberschatz & Tuzhilin 1996]. Partindo deste princípio surgem outras três importantes questões, como é que se avalia o grau de interesse de um padrão? Pode a mineração de dados gerar todos os padrões interessantes? Pode a mineração de dados gerar apenas os padrões com interesse?

Um padrão é considerado interessante se for:

- facilmente compreensível para os humanos;
- válido para casos novos ou testados com um mínimo grau de certeza;
- potencialmente úteis;
- inesperado.

Um padrão é também considerado interessante se validar uma hipótese que o utilizador procura confirmar [Silberschatz & Tuzhilin 1996]. Mais concretamente um padrão interessante representa conhecimento.

Existem medidas que permitem avaliar o grau de interesse de um padrão. Essas medidas são baseadas na estrutura da descoberta de padrões e à estatística ligada a essas medidas [Silberschatz & Tuzhilin 1996]. Uma medida de extrema importância para as regras de associação é o suporte de uma regra - esta medida já foi abordada ao longo do capítulo, mas será explicada mais concretamente na secção das regras de associação. Mas existem mais possíveis medidas que avaliam o interesse de uma regra, são elas a confiança, *lift*, teste do χ^2, entre outras, que também serão abordadas na secção das regras de associação.

Apesar de as medidas de interesse terem como objectivo identificar padrões considerados interessantes, estas medidas são insuficientes, a não ser que sejam combinadas com outras medidas que reflictam as verdadeiras necessidades e interesses do utilizador em causa. Por exemplo, padrões que descrevam as características das compras de clientes é de extrema importância para a pessoa responsável pelo *marketing*, mas poderá ser uma descoberta inútil quando, ao estudar a mesma base de dados, o objectivo é obter padrões sobre o desempenho dos empregados [Hand *et al*. 2001]. Além disso, existem padrões que apresentem valores consideráveis em relação às medidas de interesse, mas que por vezes não representam nenhuma informação nova e como tal a descoberta desses padrões é considerada inútil [Hand *et al*. 2001]. Como tal, a subjectividade das medidas de interesse é baseada de acordo com a informação necessária para o utilizador. Estas medidas permitem descobrir padrões interessantes se esses padrões forem considerados inesperados ("*unexpectedness*" [Silberschatz & Tuzhilin 1996]) e oferecerem informação estratégica para a tomada de decisões por parte doo utilizador ("actionability" [Silberschatz & Tuzhilin 1996]).

A segunda pergunta refere-se à completude dos algoritmos de mineração de dados. É irreal e ineficiente afirmar que os sistemas de mineração se dados conseguem sempre gerar todos os padrões possíveis. Para isso é necessário que as restrições e as medidas de interesse referidas pelo utilizador sejam focadas na descoberta de padrões desejados [Hand *et al*. 2001], de maneira a encontrar o maior número possível de padrões úteis ao utilizador. As regras de associação são um perfeito exemplo em que o uso das restrições e medidas de interesse conseguem assegurar a completude da mineração.

Por fim a terceira questão é considerada um problema de optimização na mineração de dados. Naturalmente, é altamente desejado que os sistemas de mineração de dados gerem apenas os padrões considerados interessantes, porque isso permitiria aos utilizadores e

mesmo aos sistemas de mineração de dados, não terem de fazer sucessivas passagens pelos padrões gerados com o intuito de descobrirem os que são realmente importantes. Tem havido bastantes progressos nessa direcção, mas mesmo assim a optimização de algoritmos de mineração de dados continua a ser um dos grandes temas de pesquisa de mineração de dados [Hand *et al*. 2001].

Medidas de interesse de padrões são essenciais para a descoberta de padrões úteis para os utilizadores. Mais precisamente, estas medidas podem ser usadas como guia e restrições no processo da descoberta, melhorando a eficiência da procura, através da filtração dos conjuntos de elementos dos padrões que não satisfaçam determinadas restrições de interesse [Hand *et al*. 2001].

2.3 Dos Padrões aos Modelos

Como se constatou pelas secções anteriores, a descoberta de padrões consiste, basicamente, na aplicação de algoritmos que permitem a extracção de padrões dos dados [Kamath & Musick 1998], através da combinação de técnicas estatísticas e tecnologias de bases de dados. A aplicação de técnicas de mineração de dados permite, por exemplo, ajudar na descoberta de potenciais localizações de recursos naturais, alertar as pessoas de uma eventual catástrofe ecológica, prever possíveis surtos de doenças infecciosas através da análise de registos médicos, ou potenciar aplicações de marketing mais eficazes através da análise das transacções de clientes, entre outras possíveis ajudas em situações reais [Laxman & Sastry 2006].

A mineração de dados envolve ajustar modelos a determinados padrões descobertos nos dados [Fayyad *et al*. 1996b]. As técnicas mais comuns, na mineração de dados, que permitem ajustar modelos a padrões descobertos são a associação [Hand *et al*. 2001], a classificação e a segmentação [www2]. Embora a classificação e a segmentação sejam consideradas técnicas similares, a sua aplicação aos dados em estudo é bastante diferente [www2]. Nesta secção será abordado uma pequena descrição de cada uma das técnicas. Nas secções seguintes será abordado uma descrição mais abrangente de cada técnica e dos seus algoritmos mais usuais, será também exposto aplicações destas técnicas em casos de estudo reais, para uma melhor compreensão da dimensão destas técnicas.

- **Associação.** Esta técnica é normalmente usada no processo da descoberta de relações entre diferentes atributos existente nos dados de uma base de dados. Estes atributos tanto podem ser considerados booleanos como quantitativos. O objectivo

das regras de associação é encontrar a essência das casualidades entre o valores dos diferentes atributos [Aggarwal & Yu 1999]. No contexto das compras de clientes, exemplo que temos vindo a utilizar, dá-se uma grande importância em como a compra de um produto pode afectar na compra de outro produto. As regras de associação permitem a descoberta dessas relações com um grau de precisão bastante grande, através da descoberta de todas as regras numa base de dados que satisfaçam determinadas restrições mínimas [Agrawal and Srikant 1994]. Esta técnica pode ser generalizada para o uso na classificação [Liu *et al*. 1998, Hand *et al*. 2001] e na segmentação [Hand *et al*. 2001] de dados de grande dimensão.

- **Segmentação.** Os problemas de segmentação consistem em dividir os dados em grupos de registos que apresentem características similares [Berkhin 2002]. Dependendo da consequente aplicação, cada um dos segmentos pode ser tratado de maneira diferente. Por exemplo, em bases de dados de imagens e vídeos, a segmentação pode ser usada para detectar padrões espaciais interessantes e características através de histogramas coloridos, textura, entre outros. Enquanto que em aplicação de seguros (por exemplo), as diferentes partições podem representar os diferentes segmentos demográficos da população, a que a cada um dos segmentos estão associadas diferentes características de risco, que devem ser analisados separadamente [Aggarwal & Yu 1999].
- **Classificação.** A classificação é considerada uma técnica bastante similar à segmentação, mas enquanto que a classificação é considerado uma técnica de aprendizagem supervisionada, a segmentação é considerada uma técnica de aprendizagem não supervisionada. Na classificação, o conjunto de treino é usado para modelar a relação existente entre as características dos atributos e a classe. Este modelo é criado com o intuito de prever a classe de uma nova instância, através das características dos atributos que identificam a nova instância [www3] [Aggarwal & Yu 1999]. Considerando um caso da aplicação de crédito e considerando um conjunto de dados de treino, em que os diferentes registos representam as características dos valores correspondentes ao comportamento bancário da população, o objectivo é descobrir a referente classe, risco de crédito ou não, a que um novo caso pertence.

2.4 Regras de Associação

O tópico das regras de associação foi introduzido em 1993 [Agrawal *et al*. 1993]. Seja $I = I_1, I_2, ..., I_n$ um conjunto de atributos binários chamados elementos e T um conjunto de transacções de elementos. Cada transacção t é representada como sendo um vector binário (tuplo), em que $t[k] = 1$ se na transacção t existir o elemento I_k, caso contrário, $t[k] = 0$.

Existe apenas um tuplo na base de dados para cada transacção [Agrawal *et al*. 1993]. Uma regra de associação representa uma implicação na forma $X \Rightarrow I_j$, em que X é um conjunto de alguns elementos de I, e I_j é um elemento em I que não está presente em X. A regra $X \Rightarrow I_j$ é satisfazível num conjunto de transacções T com uma confiança de $0 \leq c \leq 1$, se pelo menos c% das transacções em T que satisfaçam X também satisfaçam I_j. Dada uma transacção T, o interesse das regras de associação é gerarem todas as regras que satisfaçam duas formas de restrições [Agrawal *et al*. 1993]:

- **Restrições sintácticas.** Estas restrições envolvem restrições em elementos que poderão aparecer em regras. Por exemplo, o utilizador pode estar interessado apenas em regras que contenham o especifico elemento I_x como consequente, ou regras que contenham o especifico elemento I_y como antecedente. Combinações das restrições exemplificadas acima também são possíveis, isto é, o utilizador pode querer todas as regras que contenham determinados elementos de X como consequente, ou regras que contenham determinados elementos de outro conjunto de elementos Y, como antecedentes.
- **Restrições de suporte.** Estas restrições referem-se ao número de transacções em T que suportam uma regra. O suporte de uma regra é definido como sendo a fracção de transacções T ,que satisfaçam a união de elementos no consequente de uma regra com o seu antecedente. O suporte de uma regra não deve ser confundido com a confiança. Enquanto que a confiança é uma medida de força da regra, o suporte corresponde à significância estatística de uma regra. Para além da significância estatística, outra motivação para o uso das restrições de suporte, provém do facto de que apenas regras que apresentem um suporte superior a um limite mínimo são consideradas interessantes, do ponto de vista do mundo real. Se o suporte de uma regra não for suficiente grande, significa que essa regra não é considerada, ou que simplesmente é considerada menos preferível.

O processo de criação de regras de associação pode-se dividir em duas fases principais, que representam os dois problemas principais na criação de regras. Numa primeira fase, é necessário gerar todas as combinações de elementos que apresentam um suporte superior ao suporte mínimo especificado. Essas combinações são denotadas como sendo combinações largas, enquanto que as que não apresentem suporte mínimo são consideradas como sendo pequenos conjuntos de elementos. Todas as combinação são geradas tendo em conta as restrições sintácticas de maneira a restringir as combinações geradas [Agrawal *et al*. 1993]. A segunda fase, consiste na construção, propriamente dita, das regras. Para um dado conjunto de elementos largos, $Y = I_1 I_2 .. I_k, k \geq 2$, são geradas todas as regras que usem os elementos

existentes no conjunto $I_1,I_2,...,I_k$. O antecedente de cada uma destas regras será um subconjunto X de Y, tal que X tem $k-1$ elementos, e o consequente será o elemento $Y-X$. Para gerar a regra $X \Rightarrow I_j \mid c$ (na qual c representa o factor de confiança da regra), em que $X = I_1 I_2 ..I_{j-1} I_{j+1} ..I_k$, é necessário calcular a confiança da regra, através da divisão do suporte de Y pelo suporte de X. Se o resultado for maior que c, então a regra é satisfeita com o factor de confiança de c, caso contrário a regra é descartada. É preciso ter em atenção que se o conjunto de elementos Y é largo, então todos os subconjuntos de Y também são considerados largos, e como tal é necessário saber todos os seus suportes na primeira fase, descrita acima. Outra condição essencial é que todas as regras derivadas de Y têm de satisfazer as restrições de suporte, pois Y satisfaz as restrições de suporte e Y é a união de elementos provenientes, do consequente e do antecedente, de todas as regras existentes [Agrawal *et al*. 1993]. Depois de encontrados todos os conjuntos de elementos considerados largos, a segunda fase é processada de maneira trivial. Logo, como se constata o problema encontra-se na resolução da primeira fase. Nesta fase, é necessário ter atenção às medidas de interesse e ao algoritmo utilizado para descobrir o conjunto de elementos largos. De seguida serão abordadas as possíveis medidas de interesse mais utilizadas e os algoritmos principais na descoberta de regras de associação.

2.4.1 Medidas de Interesse

Sendo D uma base de dados de tamanho N, r uma regra do formato $A \Rightarrow B$, em que A e B são conjuntos de elementos. Esta será a notação utilizada em todas as abordagens das 8 medidas de interesse feitas nesta secção [www4].

1. Suporte. A noção de suporte foi introduzida em 1993 por *Agrawal et al*. [*Agrawal et al*. 1993]. O suporte é usado como uma medida de significância de um conjunto de elementos (Equação (2)). Esta medida concentra-se na contagem de transacções existentes na base de dados que contenham a regra.

$$\sup(A) = P(A) \tag{2}$$

O problema desta medida acontece, por exemplo, quando os dados apresentam conjuntos de elementos infrequentes, mas que a inclusão desses conjuntos nas regras, poderiam produzir regras bastante importantes em termos de interesse.

2. Confiança. A confiança é uma medida que também foi abordada em 1993 [Agrawal *et al*. 1993]. A confiança é considerada como uma medida que transmite a força de uma regra. É definida como a probabilidade de o consequente de uma regra se observar numa transacção, se o antecedente de uma regra também se observar nessa transacção (Equação (3)). A confiança de uma regra aceita valores entre 0 e 1 [Azevedo & Alípio 2007].

$$conf(A \Rightarrow B) = \frac{\sup(A \Rightarrow B)}{\sup(A)} \tag{3}$$

Esta medida é bastante sensível à frequência do consequente da regra. O que quer dizer, que se o consequente de uma regra apresentar um suporte elevado, então irá produzir uma confiança de valor elevado, mesmo que não existia tal associação entre os elementos.

3. *Lift*: A medida *lift*, que originalmente foi denominada como interesse, foi apresentada em 1997 [Brin *et al*. 1997]. Esta medida mede a quantidade de vezes que A e B ocorrem juntos, em relação ao que seria esperado e se são estatisticamente independentes (Equação (4)). Quanto o valor de *lift* for 1, indica que A e B co-ocorrem na base de dados como seria esperado, sob o ponto de vista da independência. Valores superiores a 1 indicam que A e B estão associados [Hashler & Hornik 2008].

$$lift(A \Rightarrow B) = \frac{conf(A \Rightarrow B)}{\sup(B)} \tag{4}$$

Esta medida não apresenta o problema exposto para o suporte. Conjuntos de elementos com baixa frequência, que por vezes ocorrem juntos algumas vezes (ou mesmo só uma) produzem um valor *lift* enorme.

4. *Conviction*. Introduzida em 1997 [Brin *et al*. 1997], esta medida foi proposta com o intuito de colmatar os problemas das medidas confiança e *lift*. Ao contrario da medida *lift*, esta medida é sensível no que toca ao sentido de uma regra, isto é $conv(A \Rightarrow B) \neq conv(B \Rightarrow A)$. O principal objectivo desta medida é medir o grau de implicação de uma regra (Equação (5)). A medida *conviction* é infinita para implicações lógicas, e é 1 se A e B forem independentes. Esta medida toma valores entre $0.5,...,1,...\infty$ [Azevedo & Alípio 2007]. Tal como a *lift*, regras com valores de *conviction* mais afastados de 1 indicam melhores regras.

$$conv(A \Rightarrow B) = \frac{1 - \sup(B)}{1 - conf(A \Rightarrow B)} \quad (5)$$

Esta medida captura, intuitivamente, a noção de implicação de regras [Brin *et al*. 1997]. Logo, a regra $A \Rightarrow B$ pode ser reescrita como $\neg(A \wedge \neg B)$. Ao contrário da confiança, ambos o suporte do antecedente e do consequente de uma regra são, considerados nesta medida [Azevedo & Alípio 2007].

5. *Leverage*. Apresentada por *Piatetsky-Shapiro* [Piatetsky-Shapiro 1991], esta medida tem como objectivo medir a diferença entre A e B aparecerem juntos no conjunto de dados e o que seria esperado se ambos fosses estatisticamente dependentes (Equação (6)). O valor possível para esta medida vai de $[-0.25,0.25]$ [Azevedo & Alípio 2007].

$$leve(A \Rightarrow B) = \sup(A \Rightarrow B) - \sup(A) \times \sup(B) \quad (6)$$

Esta medida pode sofrer também do mesmo problema exposto no suporte, pois também na utilização desta medida existe um *leverage* mínimo, que significa uma restrição na criação de regras.

6. χ^2. Esta é a única medida capaz de medir a independência estatística entre o antecedente e o consequente de uma regra (Equação (7)). O teste estatístico pode ser usado como uma medida no processo de criação de regras.

$$\chi^2(A \Rightarrow B) = N \times \sum_{X \in (A, \neg A), Y \in (C, \neg C)} \frac{(\sup(X \Rightarrow Y) - \sup(X).\sup(Y))^2}{\sup(X) \times \sup(Y)} \quad (7)$$

Esta medida não informa sobre a força da correlação entre o antecedente e o consequente, apenas informa sobre a independência de uma regra, o que significa que esta medida é considerada impraticável para problemas de *rank* [Azevedo & Alípio 2007].

7. *Jaccard*. Esta medida, ao contrário de indicar a ausência de independência estatística entre A e B, mede o grau de sobreposição entre os casos cobertos por A e B (Equação (8)). Esta medida apresenta valores ente $[0,1]$ e avalia a distância entre o

antecedente e o consequente, pela fracção entre os casos cobertos por ambos e os casos cobertos por um só (A e B).

$$jacc(A \Rightarrow B) = \frac{\sup(A \Rightarrow B)}{\sup(A) + \sup(B) - \sup(A \Rightarrow B)} \tag{8}$$

Valores próximos de 1 indicam que A e B tendem a cobrir os mesmos casos [Azevedo & Alípio 2007].

8. *Mutual Information* (MI). Esta medida é baseada na noção de entropia, e mede a redução de incerteza no consequente quando se toma conhecimento do antecedente (Equação (9)).

$$MI(A \Rightarrow C) = \frac{\sum_i \sum_j \sup(A_i \Rightarrow B_j) \times \log\left(\frac{\sup(A_i \Rightarrow B_j)}{\sup(A_i) \times \sup(B_j)}\right)}{\min\left(\sum_i -\sup(A_i) \times \log(\sup(A_i)), \sum_j -\sup(B_j) \times \log(\sup(B_j))\right)} \tag{9}$$

Tal que $A_i \in \{A, \neg A\}$ e $B_j \in \{B, \neg B\}$. Esta medida pode tomar valores entre $[0,1]$ [Azevedo & Alípio 2007].

De notar que as medidas *lift*, *leverage*, χ^2, *jaccard* e *MI* são consideradas simétricas, enquanto que a confiança e a *conviction* são assimétricas [Azevedo & Alípio 2007].

2.4.2 Algoritmos de Descoberta de Regras

Existem duas categorias principais nas quais se podem encaixar os algoritmos de descoberta de regras de acordo com o seu funcionamento. Estes algoritmos podem ser considerados como algoritmos de procura em profundidade (*Depth-First Search* (DFS)) ou algoritmos de procura em largura (*Breadth-First Search* (BFS)). Para cada um dos casos será explorado dois tipos de algoritmos, os de contagem e os de intersecção (Figura 5) [Hipp *et al*. 2000].

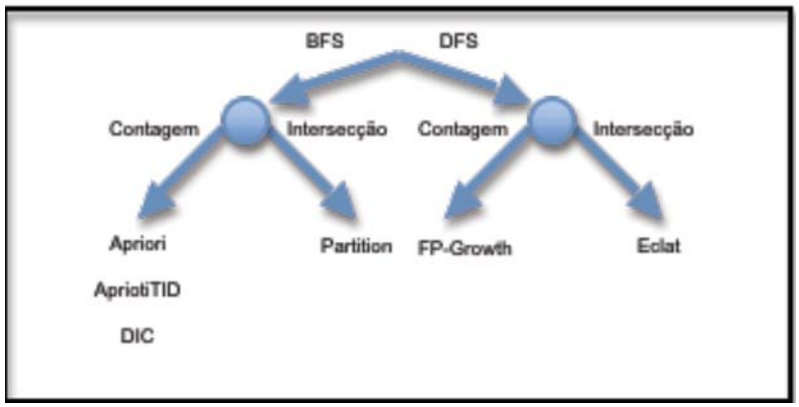


Figura 5 – Divisão dos tipos de algoritmos.

1. **BFS e Contagem de Ocorrências.** O algoritmo mais popular deste tipo é o *Apriori* [Agrawal *et al*. 1993], em que foi introduzido também o conceito de que se um conjunto de elementos é frequente, então todos os elementos desse conjunto são obrigatoriamente frequentes. Este algoritmo, adicionalmente, retira todos os candidatos que contenham subconjuntos infrequentes antes de contar os seus suportes. Esta optimização é possível devido ao funcionamento BFS, que assegura que o valor do suporte de todos os subconjuntos já seja conhecido em avanço. O *Apriori* conta todos os candidatos de cardinalidade k numa só passagem pela base de dados. A fase critica deste algoritmo é procurar os candidatos em cada uma das transacções. Com o intuito de subverter este problema, foi introduzido em 1994 [Agrawal and Srikant 1994], uma estrutura *hashtree* [Nguyen & Abiteboul 2001]. Os elementos em cada transacção são usados para descer na *hashtree*, e quando se chega a uma folha, é encontrado o conjunto de candidatos que têm o mesmo prefixo contido na transacção. De seguida, esses candidatos são procurados na transacção que foi, previamente, codificada como um *bitmap* (de maneira a tornar os testes mais rápidos), em que cada *bit* do *bitmap* corresponde a um elemento [Agrawal and Srikant 1994]. Em caso de sucesso, o contador do candidato da árvore é incrementado [Hipp *et al*. 2000]. Uma extensão ao algoritmo básico do *Apriori* é o *AprioriTID* [Agrawal and Srikant 1994]. A grande diferença do algoritmo principal para a sua extensão, consiste no formato inicial dos dados, isto é, em vez da pesquisa ser feita na base de dados em bruto, o *AprioriTID* representa cada transacção pelos candidatos existentes nessa transacção. Nesse mesmo artigo foi ainda apresentado mais dois algoritmos, em que um deles denominado por *AprioriHybrid*, e é uma combinação do *Apriori* com o *AprioriTID* [Hipp *et al*. 2000]. Outra variação do *Apriori,* que obteve um reconhecimento bastante satisfatório, é o DIC [Brin *et al*. 1997]. Este algoritmo suaviza a separação entre a fase de contagem e a geração dos candidatos [Hipp *et al*. 2000]. Sempre que um candidato atinge o suporte mínimo, isso pode até acontecer quando não se tiver visto todas as transacções existentes, o algoritmo começa a gerar candidatos adicionais, com base nesse candidato. Para esse efeito é criado uma

árvore de prefixos. Em contraste, com as *hashtree*, a cada nodo da árvore é atribuído apenas uma candidato frequente respectivamente. Isso quer dizer que cada vez que se chega a um nodo pode-se ter a certeza que o conjunto de elementos associado a nodo está contido na transacção. Outra vantagem é a diminuição do tempo de passagens pela base de dados [Brin *et al*. 1997].

2. BFS e Intersecções TID-List. O algoritmo *Partition* [Savasere *et al*. 1995] é um algoritmo do género do *Apriori*, que usa intersecções de listas de transacções para determinar o valor do suporte. Como foi descrito, o *Apriori* determina o valor do suporte de todos os $k-1$ candidatos, antes de calcular o suporte dos k candidatos. O problema é que o funcionamento dos algoritmos anteriores precisam de várias passagens pela base de dados para descobrir todas as regras, e ainda por cima, esse número de passagens nem é possível ser determinado [Savasere *et al*. 1995]. Obviamente, este problema cresce rapidamente para além das limitações de I/O das máquinas comuns. Para superar este problema, o *Partition* divide a base de dados em partições de tamanho inferior, que são tratadas independentemente. O tamanho destas partições é escolhido com o intuito de todas as listas de transacções caberem na memória principal. Depois de determinado os conjuntos de elementos frequentes para cada partição da base de dados, uma passagem extra é necessária para assegurar que todas os conjuntos frequentes locais são também globais [Hipp *et al*. 2000]. Isto permite, que no máximo haja apenas 2 passagens pela base de dados [Savasere *et al*. 1995].

3. DFS e Contagem de Ocorrências. A contagem de ocorrências assume conjuntos de candidatos de tamanho razoáveis, em que, por cada conjunto de candidatos é necessário uma passagem pela base de dados. Quando se usa um algoritmo DFS, os conjuntos de candidatos consistem apenas nos conjuntos de elementos de um dos nodos da árvore. Obviamente, percorrer a base de dados para cada nodo resulta numa elevada computação. Consequentemente, a simples combinação da técnica DFS à contagem de ocorrência é uma solução impraticável [Hipp *et al*. 2000]. Contudo, em 2000 foi apresentado um novo algoritmo baseado na técnica DFS com contagem de ocorrência [Han *et al*. 2000], que veio tirar partido destas duas técnicas juntas, o *FP-growth*. Num passo de pré-processamento, este algoritmo representa uma estrutura original de árvores de padrões frequentes (*FP-tree*), que evita custos adicionais em *scans* à base de dados [Han *et al*. 2000]. A geração da *FP-tree* é feita através da contagem de ocorrências em combinação com a técnica DFS. Em contraste com as abordagens dos anteriores algoritmos DFS, o *FP-growth* não segue os nós da árvore na procura, mas descende directamente para uma parte dos conjuntos de elementos no espaço de procura. Para assegurar que a estrutura da árvore seja compacta e informativa, apenas elementos frequentes de tamanho 1 serão nodos na árvore. Os nodos da

árvore são arranjados em ordem descendente de suporte, de maneira a que os nodos que ocorrem mais frequentemente tenham maiores hipóteses de dividir nodos, do que os que apresentam menor suporte. Uma árvore deste tipo, encontra-se bastante compactada, logo apresenta uma magnitude inferior à da base de dados original, o que permite trabalhar com um conjunto de dados inferior [Han *et al*. 2000]. Num segundo passo, o *FP-growth* usa a *FP-tree* gerada para calcular o valor do suporte de todos os conjuntos de elementos frequentes [Savasere *et al*. 1995].

4. **DFS e Intersecções TID-LIst.** Em 1997 é introduzido o algoritmo *Eclat* [Zaki *et al*. 1997], que combina a técnica DFS com a intersecção de listas de transacções. Quando se usa a técnica DFS, não é necessário dividir a base de dados em partições, como acontecia no algoritmo *Partition* da técnica BFS. Apesar do *Partition* fazer apenas duas passagens pela base de dados, à medida que o número de partições aumenta, o número de conjuntos de elementos frequentes locais também aumenta. Por essa razão o *Partition* pode gastar muito tempo na realização de computação redundante [Zaki *et al*. 1997]. O *Eclat* implementa uma optimização chamada "*fast intersections*". Isto é, sempre que se intersecta duas listas de transacções, o resultado só interessa se a lista de transacções final apresentar uma cardinalidade superior ao suporte mínimo. Por outras palavras, deve-se descartar uma intersecção sempre que se aperceba que ela não chegará ao suporte mínimo. O *Eclat* originalmente gera apenas conjuntos de elementos frequentes de tamanho igual ou superior a 3 (Hipp *et al*. 2000).

2.4.3 Casos de Estudo

Como foi mencionado acima, o estudo das regras de associação em dados de transacções comerciais tem sido estudado exaustivamente [Agrawal & Srikant 1994, Savarese *et al*. 1995, Srikant & Agrawal 1995, Agrawal *et al*. 1996, Han & Fu 1995, Srikant & Agrawal 1996]. No estudo de 1995, realizado por *Srikant* e *Agrawal* [Srikant & Agrawal 1995], é usado uma base de dados de transacções de grande dimensão, em que cada transacção consiste num conjunto de elementos. Neste artigo é apresentado dois algoritmos que permitem descobrir associações entre os diferentes produtos de qualquer hierarquia. Isto é, um casaco é considerado uma roupa para sair, que por sua vez é considerado uma peça de roupa no geral. Através desta hierarquia é possível descobrir que pessoas que compram roupa para sair tendem a comprar sapatos, e por sua vez, ainda mais especifico, pessoas que compram casacos tem tendência para comprar sapatos [Srikant & Agrawal 1995]. São abordadas comparações entre os dois algoritmos apresentados e entre o algoritmo considerado "básico", e comprova-se que estes dois algoritmos são 2 a 5 vezes mais rápidos, neste tipo de estudo

(quando aplicado a um caso de estudo real apresentou uma melhoria superior a 100 vezes na execução temporal). É abordado também a confiança e o suporte das regras, e é apresentado uma nova medida que descarta 40% a 60% das regras redundantes existentes no conjunto de dados real. O estudo realizado por *Han* e *Yu* [Han & Yu 1996] é mais uma abordagem ao tipo de estudo mencionado em cima. Outro exemplo do estudo de transacções foi abordado por *Brijs et al*. [Brijs *et al*. 1999]. Neste abordagem foi proposto um modelo microeconómico para selecção de produtos com base no uso de conjuntos de elementos frequentes, obtidos através de regras de associação. Mais especificamente, foi integrado a noção de conjuntos frequentes com alguns parâmetros microeconómicos importantes, que são frequentemente usados pelos retalhistas para suporte de decisão no processo de selecção de produtos. São usados dados provenientes de transacções de vendas de uma loja de conveniência [Brijs *et al*. 1999].

Por vezes, uma base de dados apresenta valores quantitativos e qualitativos, neste caso é necessário proceder-se a um certo mapeamento e discretização de valores. Essa é a base principal em que assenta o estudo de *Srikant* e *Agrawal* [Srikant & Agrawal 1996]. O estudo é aplicado a 500.000 registos existentes numa base de dados, em que 5 são considerados numéricos (salário mensal, limite de crédito, saldo corrente, balanço anual e interesse anual) e 2 discretos (emprego e estado civil) [Srikant & Agrawal 1996].

Apesar de grande parte dos estudos, realizados no âmbito da descoberta de regras de associação, serem sobre transacções realizadas por clientes, existem outras diversas áreas que tiram partido desta técnica de mineração de dados, como os estudos na área da saúde, mais propriamente na análise de dados de expressão de genes [Creighton & Hanash 2003, Becquet *et al*. 2002]. O objectivo destes estudos é tentar determinar como a expressão de um gene em particular pode afectar a expressão de outros genes. Outro objectivo destes estudos é tentar determinar quais os genes existentes em células doentes ou saudáveis [Creighton & Hanash 2003]. Através destes estudos foram descobertas várias associações entre determinados genes, em que maior parte aparentam ter uma grande significância biológica [Creighton & Hanash 2003, Becquet *et al*. 2002]. *Creighton* e *Hanash* [Creighton & Hanash 2003] têm como trabalho futuro a aplicação de técnicas de mineração de dados, como a associação, a um conjunto de dados sobre o cancro da mama, de maneira a encontrarem regras que relacionem resultados clínicos com certos padrões existentes na associação de genes.

2.5 Segmentação

A análise de *clusters* (segmentos) é a organização de um conjunto de padrões dentro de segmentos, baseados na similaridade. Mais concretamente, padrões referentes a um determinado segmento são mais similares entre eles, do que a outro padrão referente a outro segmento diferente [Jain *et al*. 1999]. Um exemplo da segmentação é retratado na Figura 6. Em que os padrões de entrada são mostrados na Figura 6 (a), e os segmentos criados são mostrados na Figura 6 (b). Aos pontos pertencentes ao mesmo segmento, é dado a mesma designação.

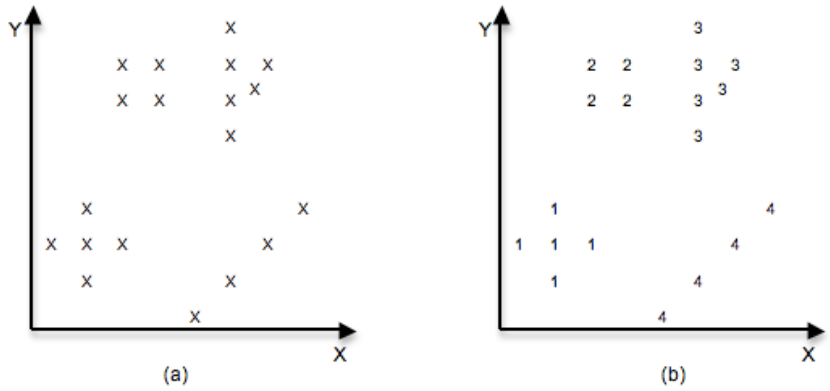


Figura 6 – Segmentação dos dados.

Apesar da segmentação e da classificação serem consideradas técnicas bastantes similares, a sua aplicação aos dados em estudo é muito diferenciada [www2]. É importante compreender a diferença entre estas duas técnicas. Na classificação (uma técnica supervisionada) é fornecido um conjunto de padrões previamente classificados, em que o objectivo é classificar uma nova instância (caso) que ainda não está classificada. Neste caso, os padrões já classificados são usados como aprendizagem das classes para servirem de base para a classificação de novos casos. Por sua vez, na segmentação, o problema é agrupar um conjunto de padrões não classificados em segmentos significativos. Na segmentação também é atribuído uma classificação a cada padrão, mas essa classificação é atribuída apenas através dos dados [Jain *et al*. 1999].

Para melhor compreensão da explicação apresentada, é necessário apresentar algumas notações que serão usadas ao longo desta secção. Seja $\mathrm{X} = (x_1,...,x_d)$ um padrão, que consiste, basicamente, num vector de dimensão (espaço do padrão) d, em que cada x_i é um atributo. Um conjunto de padrões é denotado por $\aleph = \{\mathrm{X}_1,...,\mathrm{X}_n\}$. O padrão i em $\aleph$ é

representado por $\mathrm{X}_i = (x_{i,1},...,x_{i,d})$. Em muitos casos, um conjunto de padrões é visto como uma matriz de $n \times d$ [Jain *et al*. 1999]. Um padrão pode medir, ou um objecto físico (como uma cadeira) ou uma noção abstracta (estilo de escrita). Como foi constatado no parágrafo acima, os padrões são representados, normalmente, como vectores multidimensionais, em que cada dimensão é um atributo apenas. Estes atributos podem ser quantitativos ou qualitativos. Por exemplo, se usarmos o peso e a cor como atributos, uma possível representação pode ser (20, laranja), em que laranja é a cor do objecto, e 20 é o seu peso. Dentro dos atributos quantitativos, os valores podem ser contínuos (e.g. peso), discretos (e.g. número de computadores) ou podem tomar a forma de intervalos de valores (e.g. duração de eventos). Os atributos qualitativos, podem ser nominais (e.g. cor) ou ordinais (e.g. quente ou frio). É muito importante ter a noção de que tipo de valores poderão existir, pois uma simples transformação dos atributos existentes pode significar uma melhoria significativa nos resultados da segmentação. Uma boa representação de um padrão pode produzir um segmento simples e de fácil compreensão, enquanto que uma má representação de um padrão pode resultar num segmento complexo, em que poderá ser difícil ou mesmo impossível, discernir alguma informação sobre a sua estrutura real [Jain *et al*. 1999].

2.5.1 Medidas de Similaridade

Como a noção de similaridade é fundamental para a definição de um segmento, uma medida de similaridade entre dois padrões, traçada a partir da mesma característica espacial, é essencial para a maioria dos procedimentos da segmentação. Devido à variedade de tipo de atributo, é necessário escolher a medida de distância com alguma cautela [Jain *et al*. 1999]. A escolha da medida influencia a forma dos segmentos, pois um elemento poderá estar perto de outro de acordo com uma medida, e mais longe de acordo com outra medida (Figura 7) [Han & Kamber 2006].

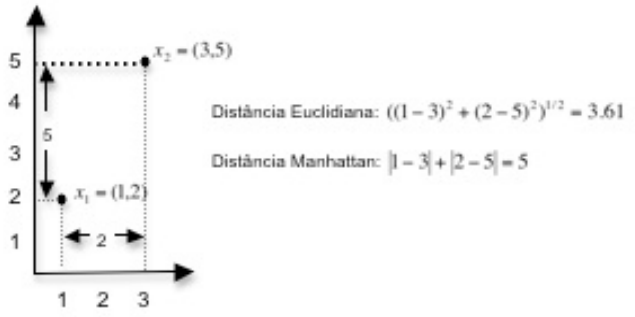


Figura 7 – Cálculo das distâncias, Euclidiana e Manhattan, entre dois objectos [Han e Kamber 2006].

De seguida irão ser apresentadas as medidas mais comuns na utilização de algoritmos de segmentação.

1. *Euclidean*: É considerada a medida mais comum no estudo de atributos contínuos. Esta medida é normalmente usada para calcular a proximidade de dois objectos num espaço de duas ou três dimensão. É também denominada de norma 2 (Equação (10)) [Jain *et al.* 1999].

$$d_2(\mathrm{X}_i,\mathrm{X}_j) = \left(\sum_{k=1}^{d}\left|x_{i,k} - x_{j,k}\right|^2\right)^{1/2} = \left\|\mathrm{X}_i - \mathrm{X}_j\right\|_2 \tag{10}$$

A medida euclidiana é uma medida que apresenta bons resultados, quando o conjunto de dados tem segmentos compactos e isolados. Mas apresenta desvantagens para correlações negativas.

2. *Manhattan*: Esta medida é semelhante à medida euclidiana, só que em norma 1 (Equação (11)).

$$d_1(\mathrm{X}_i,\mathrm{X}_j) = \left(\sum_{k=1}^{d}\left|x_{i,k} - x_{j,k}\right|^1\right)^{1} = \left\|\mathrm{X}_i - \mathrm{X}_j\right\|_1 \tag{11}$$

Ambas estas distâncias satisfazem as exigências matemáticas de uma função de distância (Han & Kamber 2006):

- $d(i,j) \geq 0$, a distância nunca toma valores negativos,
- $d(i,i) = 0$, a distância de um objecto a ele próprio é zero,
- $d(i,j) = d(j,i)$, a distância é uma função simétrica,
- $d(i,j) \leq d(i,h) + d(h,j)$, a distância de um objecto i para um objecto j no espaço, é sempre menor ou igual à soma das distâncias desses objectos a outro objecto h (Desigualdade triangular).

3. *Mahalanobis*: As medidas mencionadas acima assumem a independência dos atributos. Por sua vez, esta medida é usada em algoritmos que não assumam tal propriedade

[Berkhin 2002]. Sendo S um exemplo de uma matriz de covariâncias, dos padrões, a medida de *mahalanobis* é definida pela seguinte Equação (12).

$$d_M(\mathrm{X}_i,\mathrm{X}_j) = \sqrt{(\mathrm{X}_i - \mathrm{X}_j)^T S^{-1}(\mathrm{X}_i - \mathrm{X}_j)} \tag{12}$$

Esta medida atribui diferentes pesos a diferentes atributos, com base na sua variância e na correlação linear entre os atributos. A grande diferença entre a medida *euclidean* e a *mahalanobis*, é que esta última tem em conta a correlação do conjunto de dados e não depende da escala de medições [Berkhin 2002].

4. *Tanimoto*: Ao contrário das três medidas descritas acima, esta medida é usada para atributos qualitativos. Assumindo atributos binários com valores α, $\beta = \pm$, seja $d_{\alpha\beta}$ o número de atributo que apresente resultados α em X_i e β em X_j. Então os índices R (*Rand*) e J (*Jaccard*), também conhecidos como *Tanimoto* são representados na Equação **(**13).

$$R(\mathrm{X}_i,\mathrm{X}_j) = \frac{(d_{++} + d_{--})}{(d_{++} + d_{+-} + d_{-+} + d_{--})} \qquad R(X_i X_j) = \frac{(d_{++})}{(d_{++} + d_{+-} + d_{-+})} \tag{13}$$

De notar que *jaccard* trata os valores positivos e negativos assimetricamente, o que torna esta medida, a medida de escolha para dados transaccionais, em que + significa que um elemento encontra-se presente na transacção [Berkhin 2002].

5. *Mutual Neighbor* (MND): Existem algumas medidas em que a distância entre dois pontos é calculada tendo em conta o efeito dos pontos vizinhos circundantes [Jain *et al*. 1999]. A similaridade entre dois pontos X_i e X_j, dado este contexto é expresso pela Equação (14).

$$s(\mathrm{X}_i,\mathrm{X}_j) = f(\mathrm{X}_i,\mathrm{X}_j,\wp) \tag{14}$$

Em que $\wp$ é o contexto (o conjunto dos pontos circundantes). Uma medida definida através de um contexto é a MND, que é dada pela Equação (15).

$$MND(\mathrm{X}_i,\mathrm{X}_j) = NN(\mathrm{X}_i,\mathrm{X}_j) + NN(\mathrm{X}_j,\mathrm{X}_i) \tag{15}$$

Em que $NN(\mathrm{X}_i,\mathrm{X}_j)$ é o número de vizinhos de X_j, de acordo com X_i.

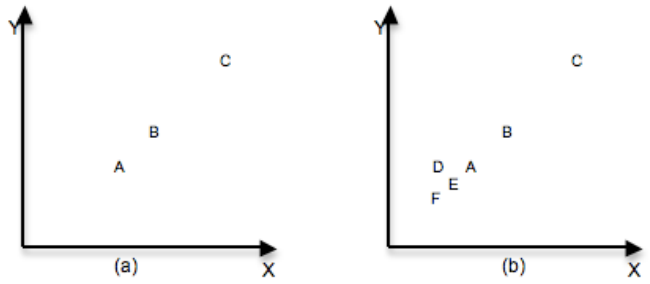


Figura 8 – Exemplo de funcionamento da medida MND [Jain *et al*. 1999].

Na Figura 8 estão dois gráficos que servirão de ajuda para exemplificar como funciona o cálculo da distância entre dois pontos através desta medida. Na gráfico (a), o vizinho mais próximo de A é B, e o vizinho mais próximo de B é A. Logo $NN(A,B)=NN(B,A)=1$, logo $MND(A,B)=2$. No entanto $NN(B,C)=1$, mas $NN(C,B)=2$, logo $MND(B,C)=3$. O gráfico (b) foi obtido através do gráfico (a), adicionando 3 novos pontos. Na análise deste gráfico, o $MND(B,C)=3$, mas agora o $MND(A,B)=5$. O valor do $MND(A,B)$ aumentou através da introdução de 3 novos pontos, mesmo que os pontos A e B não tenham sofrido qualquer alteração de localização. Agora o valor de $NN(B,A)=4$. Através desta explicação, constata-se facilmente que esta medida não satisfaz a desigualdade triangular. Apesar disso, esta medida tem sido aplicada em algoritmos de segmentação [Jain *et al*. 1999].

As medidas apresentadas são apenas algumas das utilizadas, existindo um vasto número de variedade de medidas e extensões das mesmas. Como foi dito anteriormente, cada medida pode apresentar valores diferentes para calcular a mesma distância entre dois pontos, como tal é bastante importante ter a perfeita noção de qual a medida mais adequada para o estudo em causa.

2.5.2 Técnicas da Segmentação

Existe um extenso número de algoritmos para a técnica segmentação. As duas técnicas mais tradicionais são a segmentação hierárquica e a segmentação particionada. A segmentação hierárquica pode ser subdividida em aglomerativa [Jain & Dubes 1988] ou divísel [Kaufman & Rousseeuw 1990]. Apesar destas duas técnicas serem consideradas as mais tradicionais e abrangentes, existe ainda um vasto número de outras técnicas, sendo três delas mais sonantes. Estas três técnicas, foram abordadas com o intuito de combater as restrições das

técnicas hierárquicas e particionadas [Berkhin 2002]. São técnicas mais especificas, baseadas em estruturas e casos específicos. Nesta secção serão exploradas as 5 diferentes técnicas, dando mais importância às duas primeiras.

1. ***Hierarchical*****:** As técnicas hierárquicas têm como base a construção de uma árvore de sucessivos segmentos, usando segmentos previamente existentes. Esta técnica pode ser subdividida numa técnica aglomerativa ou divisível, dependendo de como a decomposição hierárquica é feita. A qualidade da aplicação da uma técnica puramente hierárquica sofre com a impossibilidade de realizar ajustes sempre que uma divisão ou uma junção de segmentos é executada. Isto significa que, se numa fase posterior, uma divisão ou junção de um segmento em particular for considerado uma má escolha, é impossível reverter esse passo de maneira a corrigi-lo [Han & Kamber 2006].

- ***Aglomerative*****:** Esta técnica consiste em atribuir a cada ponto existente um segmento, para nos passos seguintes juntar estes segmentos atómicos em segmentos cada vez maiores, até que se obtenha um único segmento ou então até certas condições de paragem serem satisfeitas. Grande parte dos algoritmos hierárquicos opta por utilizar este tipo de técnica, em que apenas diferem na definição de similaridade entre os segmentos [Han & Kamber 2006, Berkhin 2002].
- ***Divisive*****:** O funcionamento desta técnica é o contrário da técnica aglomerativa. Inicialmente começa-se com um único segmento inteiro e, a cada passo, procede-se à divisão em segmentos de tamanho inferiores, até cada ponto representar um segmento, ou até ser satisfeito as condições de paragem, como o número mínimo de segmentos, ou o diâmetro de cada segmento estar dentro de uma determinado limite dado.

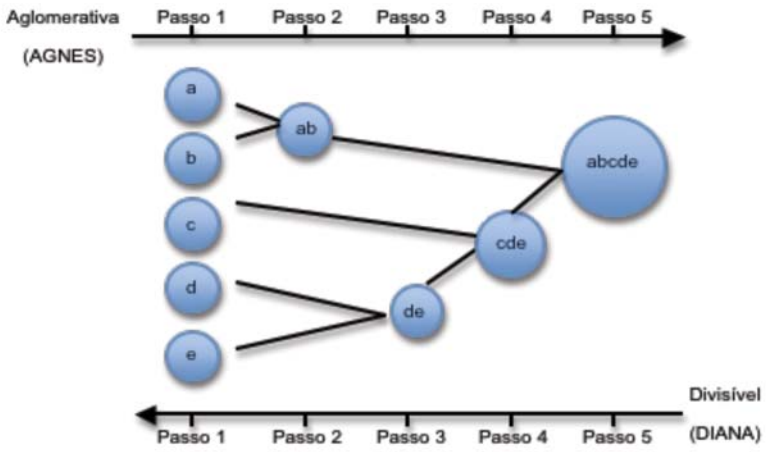


Figura 9 – Funcionamento das duas sub técnicas hierárquicas.

Na Figura 9, é apresentado um exemplo da aplicação de AGNES [Kaufman & Rousseeuw 1990] (*AGlomerative NESting*), um algoritmo da técnica hierárquica aglomerativa, e de DIANA [Kaufman & Rousseeuw 1990] (*DIvisive ANAlysis*), um algoritmo da técnica hierárquica divisível, a um conjunto de dados de 5 elementos $\{a,b,c,d,e\}$. Inicialmente, o algoritmo AGNES coloca cada elemento num segmento distinto. De seguida os segmentos são aglomerados passo por passo, de acordo com o critério definido. Por exemplo, os segmentos dos respectivos elementos a e b podem ser aglomerados num só segmento, se a distância euclidiana entre os dois elementos for menor do que qualquer outra distância entre outros dois elementos. É uma abordagem de ligação única, em que cada segmento é representado por todos os elementos do segmento, e a similaridade entre dois segmentos é medida através da similaridade do mais próximo par de pontos pertencentes a outros segmentos. O processo de aglomeração de segmentos é repetido, eventualmente, até todos os elementos estarem num único segmento [Han & Kamber 2006]. O algoritmo DIANA, parte do ponto inicial em que existe um único segmento com todos os objectos. Este segmento inicial é dividido de acordo com alguns critérios, como a distância euclidiana máxima entre os elementos vizinhos do segmento. Este processo é repetido até, eventualmente, cada segmento conter apenas um elemento [Han & Kamber 2006]. Em ambos os casos foi usada a expressão eventualmente, quando referido à paragem do algoritmo, pois para cada um dos algoritmos pode-se estipular o número desejado de segmentos [Han & Kamber 2006], e aí o critério de paragem pode mudar.

Os algoritmos hierárquicos são bastante mais versáteis que os particionais. Por exemplo, os algoritmos aglomerativos apresentam bons resultados quando a aplicados a conjuntos de dados contendo segmentos não isotrópicos, enquanto que o *k-means* (algoritmo particional) apresenta bons resultados em conjuntos de dados, que apresentem as mesmas propriedades, isto é que sejam isotrópicos [Nagy 1968]. Mas por outro lado, os algoritmos particionais apresentam melhores resultados, quando falamos na complexidade e do tempo de execução [Jain *et al*. 1999]. De maneira a tirar proveito das características dos dois tipos de técnicas, foi explorado um algoritmo híbrido em 1980 [Murty & Krishna 1980]. Entretanto, em 1996 foi apresentado o BIRCH [Zhang et al. 1996] (Balanced Iterative Reducing and Clustering), que tem como objectivo integrar técnicas hierárquicas com outras técnicas de segmentação, resultando numa segmentação de múltiplas fases [Han & Kamber 2006]. Os algoritmos hierárquicos tendem a ser bastante sensíveis à existência de *outliers* ou ruído nos dados, devido à medida da distância entre dois segmentos apresentar dois extremos (aglomerativa – distância mínima, divisível - distância máxima) [Han & Kamber 2006]. O algoritmo CURE [Guha et al. 1998] (*Clustering Using REpresentatives*) veio melhorar essa deficiências nas técnicas hierárquicas, pois este algoritmo é capaz de encontrar segmentos de

diferentes tamanhos e formas, e é insensível a *outliers*. Outro exemplo deste tipo de algoritmos é o ROCK [Guha et al. 2000] (*Robust Clustering Algorithm for Categorical Data*), mas este tem como objectivo trabalhar com atributos qualitativos [Berkhin 2002].

2. *Partitional*: Dada uma base de dados de n elementos ou tuplos de dados, os algoritmos particionais particionam os dados em k partições, em que cada partição representa um segmento e $k \leq n$. Isto é, estes algoritmos classificam os dados em k segmentos, em que cada segmento contém pelo menos um elemento e cada elemento deve pertencer a um único segmento [Han & Kamber 2006]. Os algoritmos particionais apresentam uma complexidade temporal, bastante boa, pois, normalmente os k segmentos são encontrados logo numa única iteração [Steinbach *et al*. 2000]. A vantagem deste método é diminuir a confusão, por exemplo, caso se tenha 25 segmentos, a análise dos resultados poderá ser um pouco confusa, sendo possível ultrapassar essa situação com a escolha de apenas 10 segmentos finais (escolha do k). Mas por outro lado, o facto de se escolher o número de segmentos, pode levar a uma degradação no resultado final [Jain *et al*. 1999], pois pode haver outro k que apresente um resultado mais distinto e observável ao analista [www1]. Esta desvantagem, provoca na prática, diversas execuções do algoritmo para diferentes k, e o melhor resultado encontrado de todas as execuções feitas, é considerado o resultado final da aplicação da técnica de segmentação [Jain *et al*. 1999]. Consequentemente, apesar dos algoritmos particionais conseguirem executar em apenas uma iteração, é necessário executá-los por algumas vezes, de maneira a encontrar o melhor resultado que se pode obter. Os algoritmos particionais mais comuns são o *k-means* e o *k-medoids*, que serão abordados de seguida.

- ***K-Means*:** O algoritmo *k-means* recebe um k, e particiona um conjunto de n elementos em k segmentos, o que leva a um similaridade entre segmentos baixa, mas uma similaridade alta entre os elementos existentes nos segmentos [Han & Kamber 2006]. A similaridade de um segmentos é calculada através do valor médio dos elementos num segmento, que pode ser visto como o centróide do segmento. Inicialmente, o *k-means* selecciona k elementos, que representarão os centróides de cada segmento. Para cada um dos elementos restantes, é associado um segmento, com base na distância entre o elemento em causa e o centróide do segmento. A medida por defeito utilizada no cálculo da distância entre elementos é medida euclidiana (Equação (10)). De seguida é calculado o novo centróide para todos os segmentos existentes. Este processo é repetido até que a função objectiva convirja [Berkhin 2002, Han & Kamber 2006]. A função, normalmente utilizada é a soma dos quadrados dos erros [Berkhin 2002], que é dada pela Equação **(**16).

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \quad \textbf{(}16\text{)}$$

Em que p é o ponto no espaço representando um dado elemento, e m_i é a média do segmento C_i. Por outras palavras, para cada elemento de cada segmento, a distância entre o elemento e o centro do segmento é calculada ao quadrado, e de seguida as distâncias são somadas. Este critério tem como objectivo fazer os k segmentos mais compactos e o mais separados possíveis [Han & Kamber 2006]. Este algoritmo apresenta algumas desvantagens, como é óbvio. O facto de só poder ser aplicado assim que for calculado a média de todos os segmentos é uma desvantagem, pois quando temos atributos qualitativos nos dados, o cálculo da média torna-se bastante complicado. A necessidade de se especificar o k poderá ser também uma desvantagem, como já foi abordado. O *k-means* é uma algoritmo pouco aconselhado para a descoberta de segmentos com formato pouco compacto. Além disso, é bastante sensível a *outliers* e a ruído nos dados, pois mesmo existindo um pequeno número de dados desse género, é possível influenciarem em muito o cálculo do centróide de um segmento [Han & Kamber 2006, Berkhin 2002]. O *k-modes* é uma extensão ao *k-means* que possibilita o uso de atributos qualitativos, substituindo a média dos segmentos por modos, usando novas medidas de desigualdade que lidam com atributos qualitativos. É um algoritmo baseado na frequência, para actualizar os modos dos segmentos [Han & Kamber 2006].

- ***K-Medoids*:** Como foi mencionado, o *k-means* é bastante sensível à existência de *outliers*, pois um elementos com um valor extremamente elevado pode distorcer, substancialmente, a distribuição dos dados. Este efeito é particularmente acentuado devido ao uso da soma dos quadrados dos erros (Equação **(**16)). O *k-medoids* ganha vantagem ao *k-means* nessa situação. O *k-medoids* em vez de pegar no valor médio do elemento num segmento como ponto de referência, pega no elemento em si como representação do segmento, designado por *medoid*. Isto é feito para todos os segmentos existentes. Cada elemento que sobra é associado ao segmento que apresenta o elemento representativo mais similar ao elemento que sobra [Han & Kamber 2006]. A partição é feita com o principio de minimizar a soma das similaridades entre cada elemento e o respectivo *medoid* (ponto de referência do segmento), isto é, a função objectivo é a soma do erro absoluto definida pela Equação (17).

$$E = \sum_{j=1}^{k} \sum_{p \in C_j} |p - o_j| \quad (17)$$

Em que p é o ponto no espaço representando um dado elemento no segmento C_j, e o_j é o elemento representante do segmento C_j. Em geral, o algoritmo itera, eventualmente, até que cada elemento de referência seja o *medoid*, isto é, o cálculo do novo *medoid* seja igual ao anterior. Este algoritmo é mais robusto que o *k-means* na presença de ruído e de *outliers*, devido à escolha do *medoid* ser influenciada pela localização de maior predominância de elementos num segmento [Berkhin 2002]. No entanto, o seu processamento é mais complexo e em ambos os casos é necessário especificar o número de segmentos, k. Duas versões do *k-medoid* é o algoritmo PAM [Kaufman & Rousseeuw 1990] (*Partitioning Around Medoids*) e o CLARA [Kaufman & Rousseeuw 1990] (*Clustering LARge Applications*). O algoritmo PAM, após uma selecção aleatória dos k elementos representativos, tenta iterativamente optimizar a escolha desses mesmos elementos representativos, através da análise de todos os possíveis pares. Cada par é constituído pelo elemento representativo e por outro elemento que não o seja [Han & Kamber 2006]. O PAM apresenta resultados bastantes satisfatórios para um conjunto de dados pequeno, mas não é tão satisfatório para um conjunto de dados de larga dimensão. Com o intuito de obter bom resultados na aplicação a conjuntos de dados de grande dimensão, foi apresentado o algoritmo CLARA [Han & Kamber 2006]. Progressos adicionais estão associados a *Ng & Han*, que introduziram o algoritmo CLARANS [Ng & Han 2002] (*Clustering Large Applications based upon RANdomized Search*). Basicamente, o processamento do CLARANS pode ser visto como uma pesquisa através de um grafo, em que cada nodo é uma potencial solução (um conjunto de k *medoids*). Este algoritmo tem sido experimentado e tem mostrado ser mais eficaz do que o PAM e o CLARA. O CLARANS também é capaz de detectar *outliers*, no entanto apresenta uma complexidade um bocado elevada [Han & Kamber 2006].

3. *Density based*: A maior parte dos algoritmos particionais são baseados na distância entre elementos. Esses algoritmos conseguem facilmente encontrar segmentos de formas esféricas, mas apresentam bastantes problemas na descoberta de segmentos que apresentem formas irregulares. Essas dificuldades foram ultrapassadas através da criação de algoritmos baseados na densidade. A ideia principal deste tipo de algoritmos, é o crescimento do segmento em causa, até à densidade da vizinhança (número de elementos) exceder um limite estabelecido. Isto é, para cada elemento de uma dado segmento, a vizinhança desse

elemento, dado o raio, tem de conter um número mínimo de elementos. Esta abordagem é vantajosa na existência de *outliers* e permite descobrir segmentos que apresentem formas irregulares. O DBSCAN [Ester *et al*. 1996] (*Density Based Spatial Clustering of Applications with Noise*) e suas extensões, são algoritmos criados baseados na noção de densidade, que proporcionam o crescimento dos segmentos com base na análise da densidade [Han & Kamber 2006].

4. *Grid Based*: Este tipo de algoritmos quantificam os elementos numa partição espacial, distribuídos pelo um número finito de células que formam uma estrutura em rede [Berkhin 2002]. Todas as operações de segmentos são realizadas numa estrutura em rede. A principal vantagem desta abordagem é a rapidez no tempo de processamento, que é normalmente independente do número de elementos nos dados, e dependente do número de células em cada dimensão da partição espacial [Han & Kamber 2006]. O algoritmo STING [Wang *et al*. 1997] (*STatistical INformation Grid-based method*) é um algoritmo típico baseado numa estrutura em rede, que colecciona informações estatísticas numa rede de células [Han & Kamber 2006].

5. *Model based*: Os algoritmos baseados na construção de modelos admitem a hipótese de um modelo para cada segmento existente, e escolhem o melhor ajuste dos dados ao modelo determinado [Berkhin 2002]. Este tipo de algoritmos baseiam-se, frequentemente, na hipótese de que os dados são gerados por uma mistura de distribuições probabilísticas, que reflectem a distribuição espacial dos elementos [Han & Kamber 2006]. Três exemplos deste tipo de algoritmos são: O EM [Dempster *et al*. 1977] (Expectation-Maximization) que é uma extensão do *k-means*, o COBWEB [Fisher 1987] é um algoritmo de aprendizagem conceptual que funciona através da análise das probabilidades, e por fim o SOM [Kohonen 1982] (*Self-Organizing feature Maps*) que é uma aproximação das redes neurais à segmentação [Han & Kamber 2006].

A escolha de um algoritmo de segmentação depende do tipo de dados disponível, bem como do objectivo da utilização da técnica de segmentação. No entanto existem algoritmos que abordam mais do que uma ideia, dos algoritmos explorados acima, o que por vezes leva a uma indecisão na sua utilização. Mas pelo contrário, pode existir aplicações que requerem a aplicação de mais de uma das técnicas de segmentação [Han & Kamber 2006].

Por algumas vezes foi abordado o tema *outliers* neste trabalho. Um *outlier* é um elemento que apresenta um comportamento completamente diferente do existente, no resto dos dados. Um *outlier* pode ser causado por um erro de inserção, por exemplo atribuir a uma pessoa de 29 anos a idade 299. Ou simplesmente, pode ser causado por uma variação de

dados normal, por exemplo, o salário de um chefe, pode muito bem ser um *outlier* quando em causa de estudo se tem os salários dos empregados de uma grande empresa [Han & Kamber 2006]. Muitos dos algoritmos de mineração de dados, tentam diminuir a influência dos *outliers*, ou por vezes, procede-se mesmo à sua eliminação. No entanto, isto pode resultar na perda de informação relevante que se encontra escondida, como no caso de detecção de fraudes, em que um *outlier* pode indicar uma actividade fraudulenta [Han & Kamber 2006]. A aplicação de algoritmos de segmentação facilita a descoberta de *outliers*, pois cada elemento que se encontra bastante disperso dos restantes elementos dentro dos segmentos, pode claramente ser considerado um *oulier*.

2.5.3 Casos de Estudo

A segmentação é usada em inúmeras estudos de diagnósticos médicos, na detecção de fraudes, na detecção de anomalias, na segurança de redes [Bianco *et a*l. 2005], entre outros exemplos [Berkhin 2002]. Através do uso de técnicas de segmentação é possível fazer distinção de grupos, que possibilita a extracção de conclusões acerca dos dados, de maneira a prevenir acontecimentos considerados ilegais. Um exemplo do uso da segmentação para prevenção de acontecimentos, é o estudo de 2008, intitulado "*Internal Fraud Risk Reduction: Results of a Data Mining Case Study*" [Jans *et al*. 2008]. Como o próprio nome indica, o estudo em causa, não tem como objectivo indicar se uma acção é fraudulenta ou não, mas sim reduzir o risco de fraudes, através da análise descritiva dos dados. Neste estudo foram usados dados de uma empresa, que se encontra classificada dentro do top 20 financeiro europeu. Através da aplicação da segmentação foram extraídos quatro grupos, que após uma observação atenciosa, foram classificados em casos fraudulentos, casos de evasão de procedimentos, erros ou enganos e *outliers*. A observação fraudulenta faz parte da detecção de fraudes, enquanto que as observações dos casos de evasão de procedimentos podem ser criados por erros ou podem ser indícios de fraude, e assim, é possível a sua prevenção [Jans *et al*. 2008]. O artigo "*Comprehensive Survey of Data Mining-based Fraud Detection Research*" [Phua *et al*. 2005], faz uma abordagem geral de vários estudos, referentes à detecção de fraudes para diversas áreas.

Um das áreas bastante explorada pela segmentação é a área da saúde. Um estudo bastante recente, explorou a propagação do vírus dengue numa população aborígene na Tailândia [Mammen *et al*. 2008]. Este estudo tem um grande impacto neste tipo de população, principalmente, quando se fala de um vírus que é o principal causador de doenças na população aborígene em todo o mundo. Este estudo baseou-se na análise da propagação do vírus dentro de um contexto temporal e geográfico. Ao contrário de um estudo realizado em

2005 [Beckett *et al*. 2005], em que o objectivo final era o mesmo, este estudo só foi feito em relação a crianças, pois os adultos podem apresentar uma maior imunidade ao vírus dengue. Outra grande diferença deste estudo para o anterior envolve a colheita geográfica, isto é, enquanto que em 2005 a recolha foi feita a uma área urbana num hospital, para o estudo de 2008, o estudo foi realizado numa área rural, com base numa escola. Foram recolhidos dados de 100 casas, em que crianças febris (segmento positivo) ou não febris (segmento negativo) foram acompanhadas durante dois picos de propagação do vírus dengue. Este estudo permite, num futuro próximo, explorar através da segmentação, como é que a imunidade do hospedeiro e os aspectos comportamentais podem ter impacto na transmissão do vírus, e assim criar estratégias de prevenção [Mammen *et al*. 2008].

Cada vez mais é debatido e explorado, a origem e possíveis aparecimentos do cancro. Apesar de já ser possível prevenir numa fase mais prematura, é necessário continuar as pesquisas nesta área para mais soluções se poder ter, para uma doença que ainda mata milhões de pessoas por ano. Um dos cancros bastante comum nas mulheres é o cancro da mama. Como tal, diversos estudos tem sido avançados nessa área, e para isso as técnicas de mineração de dados são bastante necessárias. Um estudo realizado em 2006 [Carey *et al*. 2006] abordou o aparecimento do cancro da mama em mulheres, de acordo com a sua idade e a sua raça. Para este estudo foram usados os registos de 24 municípios da zona este e central da Carolina do Norte, e foram analisadas mulheres entre os 20 e os 74 anos de raça afro-americana e branca. A estratégica da amostra foi equilibrar os registos dos 4 grupos (mulheres afro americanas de idade inferior a 50 anos, mulheres afro americanas com idade superior a 50 anos, mulheres não afro americanas de idade inferior a 50 anos e mulheres não afro americanas com idade superior a 50 anos), para se poder realizar comparações estatísticas entre os 4 grupos. Este estudo concluiu que o aparecimento do cancro da mama é mais predominante nas mulheres afro americanas de idade inferior a 50 anos. No entanto, é bastante importante avaliar o aparecimento do cancro da mama para as diferentes raças, com base na fase do tumor ou no historial familiar de aparecimentos de tumores [Carey *et al*. 2006]. Na área da saúde, a análise de padrões de expressões genéticas, através da técnica da segmentação, são também bastante usuais [Ben-Dor *et al*. 1998, Bezerra *et al*. 2005].

Outra abordagem, de bastante relevo, na aplicação da técnica de segmentação, é a detecção de crimes ou terrorismo. No artigo de *Phua et al*. [Phua *et al.* 2005] é mencionado a possível detecção de terroristas com base na aplicação de técnicas de mineração de dados, tais como a segmentação. Entretanto em 2006, foi apresentado um estudo com base na detecção de padrões de crimes [Nath 2006]. Este estudo é feito a partir da análise de diferentes tipos de crimes, agrupando-os em diferentes segmentos, isto é, cada crime cometido é associado a uma ou mais pessoas que o cometeram. Essa pessoa é estudada de acordo com a sua idade,

raça, idade da vítima, arma utilizada, entre outros. Logo, para cada tipo de crime, são associadas certas características humanas que poderão ser semelhantes, ou não, a pessoas causadores de um crime no futuro. De notar, que em nada este tipo de estudo serve para assumir o culpado de um crime, e muito menos serve para substituir os investigadores dos crimes. No entanto, este estudo serve para uma possível ajuda aos investigadores, apresentando possíveis características do possível infractor. Num estudo futuro, o objectivo é prever os locais mais usuais de ocorrência de crimes (este tipo de estudo já foi aplicado pelo departamento policial de Richmond[1]), para uma maior eficiência da utilização de recursos policiais [Nath 2006].

2.6 Classificação

A classificação é uma técnica bastante diferente da segmentação. Apesar de serem similares no aspecto de também criar segmentos de registos denominados por classes, a análise da classificação necessita que o analista conheça antes a maneira como essas classes estão definidas. isto é, o objectivo não é explorar os dados de maneira a descobrir segmentos, mas sim decidir como as novas instâncias deverão ser classificadas [www2]. Num problema de classificação, os dados de entrada são categorizados como sendo o conjunto de treino, que contém um conjunto de registos que por sua vez, cada registo contem múltiplos atributos. A cada exemplo do conjunto de treino é categorizado uma classe, que pode ser categórica ou quantitativa. Caso a classe seja quantitativa, então o problema da classificação é referenciado como sendo um problema de regressão linear. O objectivo principal da classificação é através do conjunto de treino criar um modelo de classificação, e através desse modelo prever os valores da classe do conjunto de teste [Aggarwal & Yu 1999]. A classificação é considerado uma técnica supervisionada, pois ao contrário de uma técnica não supervisionada (segmentação), em que a categorização de uma classe de cada registo do conjunto de testes é previamente fornecida [Han & Kamber 2006].

2.6.1 Pré Tratamento dos Dados

Para uma melhor precisão, eficiência e escalabilidade do processo de classificação, é necessário proceder-se a um pré tratamentos dos dados, como à diminuição do ruído, à transformação de valores e à eliminação de atributos irrelevantes na análise.

[1] http://www.spss.com/dirvideo/richmond.htm?source=govtlawen&zone=rtsidebar

- **Limpeza dos dados:** Este pré tratamento refere-se à eliminação ou redução do ruído (através da aplicação de técnicas de suavização do valor de previsão da classe, por exemplo) ou tratamento de valores em falta (através da substituição do valor em falta, pelo valor mais comum do atributo em causa, ou do mais provável de acordo com estatísticas, por exemplo). Apesar de maior parte dos algoritmos de classificação apresentarem mecanismos de tratamento de nulos e ruído, o processamento deste passo ajuda na redução da confusão durante o processo de aprendizagem do algoritmo [Han & Kamber 2006]. Este pré processamento dos dados, pode, por vezes, ajudar ainda mais na redução dos valores indesejáveis, e resultar num modelo de maior utilidade e precisão.
- **Transformação de valores:** A transformação de valores abrange a normalização de valores. Esta normalização transforma os valores dos atributos, em valores que se situem dentro de um escala de pequena variação de valores. Esta técnica é bastante utilizada quando se fala no atributo salário, por exemplo, em que temos valores bastantes elevados, e através desta técnica podemos transformar um valor 73.600 num valor entre 0 e 1 [Han & Kamber 2006]. Esta normalização permite um aumento da velocidade da fase de aprendizagem. A transformação de valores pode também ser referenciada pela descretização de valores, isto é, quando se tem em conta o atributo temperatura do ar, é possível ter uma grande quantidade de valores diferentes. Esta variedade de valores, pode levar a uma análise menos concisa e a um elevado número de operações, principal quando se fala em atributos contínuos. De maneira a subverter este problema, pode-se generalizar o valor da temperatura, em baixa, média ou alta. De igual forma, para atributos discretos, como a morada, pode-se generalizar em cidades, e assim obter uma menor variedade de valores. Na transformação de valores é possível, também, a construção de novos atributos através de outros atributos do conjunto de dados, de maneira a obter-se um melhor resultado no processo da mineração de dados [Han & Kamber 2006].
- **Análise da relevância dos atributos:** É muito normal no conjunto de dados de entrada existirem atributos irrelevantes para o estudo em causa. Através da análise de correlação entre dois atributos, é possível constatar que um deles pode ser descartado. Isto só poderá acontecer se a correlação entre eles for considerada bastante forte, caso contrário, significa que o comportamento de um atributo em causa, não influencia em nada o comportamento do outro. A própria base de dados pode conter atributos irrelevantes. Nesta situação procede-se a uma selecção de atributos, de maneira a reduzir o conjunto de atributos, de tal maneira que a probabilidade da distribuição das classes dos dados seja o mais parecida possível com a distribuição original obtida, usando todos os atributos. Por

fim, a análise da relevância, através da análise da correlação e da selecção de atributos, pode detectar atributos que em nada interessam para o estudo em causa, mas que por sua vez, podem causar uma redução da velocidade do processo de aprendizagem [Han & Kamber 2006].

2.6.2 Técnicas de Classificação

Uma técnica da classificação é uma aproximação sistemática para a construção de modelos de classificação a partir de um conjunto de dados de entrada [Tan *et al*. 2006]. Neste secção serão abordadas quatro técnicas para a construção de modelos: *Decision Trees*, *Bayesian Classifiers*, *Lazy Learnes* e *Neural Networks*. Cada técnica utiliza um algoritmo de aprendizagem, com o intuito de identificar o modelo que melhor ajusta a relação entre o conjunto de atributos, e a classe referente aos dados de entrada [Tan *et al*. 2006, Han & Kamber 2006]. O modelo gerado tem como objectivo prever correctamente a classe de um novo caso.

1. ***Decision Trees*:** A ideia das árvores de decisão consiste em particionar os dados até que cada partição contenha maioritariamente exemplos da mesma classe [Aggarwal & Yu 1999]. Uma árvore de decisão é um fluxograma com uma estrutura igual a uma árvore, em que cada nodo interno é um teste a um atributo, cada ramo representa o resultado do teste e cada folha representa a classe final [Tan *et al*. 2006, Han & Kamber 2006]. Na Figura 10 mostra o exemplo da estrutura de uma árvore de decisão, em que classe de previsão é se compra ou não computador. Se a pessoa em causa for jovem e for estudante, então provavelmente comprará computador, mas por sua vez, se for idosa e possuir uma pensão baixa, provavelmente, não deverá comprar computador.

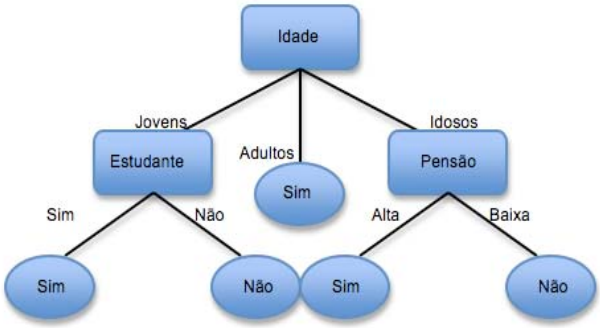


Figura 10 – Exemplo da estrutura de uma árvore de decisão [Han & Kamber 2006].

Cada nodo da árvore contém um ponto de divisão, que usa um critério de condição de como os dados deverão ser particionados. O desempenho da árvore de decisão depende bastante da escolha do ponto (atributo) de divisão [Aggarwal & Yu 1999]. A escolha desse ponto de divisão será abordado mais abaixo [Aggarwal & Yu 1999].

Esta técnica é considerada uma técnica bastante usual e popular, sendo a técnica mais utilizada quando se fala na previsão através da classificação. Isto deve-se à vantagem das árvores de decisão poderem trabalhar com grandes dimensões de dados, e a sua forma em árvore é bastante intuitiva, o que torna a fácil a sua compreensão [Han & Kamber 2006]. Em geral, as árvores de decisão apresentam bastantes bons resultados em relação à previsão de modelos, no entanto, por vezes, essa precisão só é obtida através de um pré tratamento manual dos dados. Apesar do processo de classificação de árvores de decisão ser considerado um processo fácil, é necessário ter em conta duas questões bastantes importantes, de maneira a que se consiga obter um bom modelo de previsão [Tan *et al*. 2006]. A primeira questão consiste na partição do conjunto de treino. Em cada passo recursivo do processo do crescimento da árvore é necessário escolher um atributo como teste de condição que divida os registos em subconjuntos de menor tamanho. Essa divisão é feita com base no tipo de atributos (contínuos, discretos e discretos e binários), e com base numa medida que avalia qual a melhor divisão. A segunda questão consiste no critério de paragem de divisão de uma árvore. Isto é, é necessário uma condição de paragem no processo de crescimento de uma árvore. Neste caso existem duas opções, continuar a expandir a árvore até todos os registos restantes pertencerem à mesma classe, ou terem valores de atributos idênticos. Apesar de ambas as condições serem suficientes para a paragem de crescimento da árvore, outros critérios podem ser impostos de maneira a que o processo termine antes [Tan *et al*. 2006]. As vantagens de uma terminação antecipada serão abordadas quando for mencionado o problema de *overfitting* das árvores de decisão. Como foi mencionado é necessário ter em conta o tipo de atributos a que se vai ajustar o modelo, pois a divisão difere para cada um dos tipos:

- **Atributos discretos:** Nos atributos discretos, para cada valor que o atributo pode tomar, é formado um novo ramo [Han & Kamber 2006]. Por exemplo, na Figura 10, os registos do atributo idade só podem tomar valores de jovens, adultos ou idosos, como tal, são representados os três tipos de possíveis registos na árvore.
- **Atributos contínuos:** No caso dos atributos contínuos, sendo A um novo caso, e N o nodo correspondente ao teste em causa, tem-se que esse que N só pode ter dois possíveis resultados, ou $A \geq split\_point$ ou $A < split\_point$, respectivamente. Em que *split_point* é o ponto escolhido para a partição, obtido

através do processo de escolha do atributo de divisão [Han & Kamber 2006]. Imaginado na Figura 10, que em vez dos ramos referentes à pensão serem definidos por alta e baixa, fossem definidos por valores, como por exemplo, $Pensão < 500$ e $Pensão \geq 500$, obter-se-ia um divisão de atributos contínuos. Neste caso se $A \geq 500$, então haveria maior probabilidade dessa pessoa comprar um computador, caso contrário, se $A < 500$, então haveria maior probabilidade de não comprar computador.

- **Atributos discretos e binários:** Neste tipo de atributos, o teste do nodo N é da forma $A \in S_A$, em que S_A é o subconjunto da divisão para A, obtida também, através do processo de escolha do atributo de divisão [Han & Kamber 2006]. Isto é, imaginando na Figura 10, que o nodo principal fosse $S_A = \{jovens, adultos\}$, os ramos só poderiam ser sim ou não (valores binários).

Uma medida de selecção de um atributo é uma heurística para a selecção do critério de divisão, que melhor separa um conjunto de treino de dados, em classes individuais. Numa forma geral, o melhor critério de divisão, é aquele que resultar em resultados mais próximos dos registos, numa partição, pertencentes à mesma classe [Han & Kamber 2006]. Uma medida de selecção de atributos, apresenta uma classificação (resultado) para cada atributo pertencente ao conjunto de treino. O atributo que apresentar melhor resultado para a medida em causa, é escolhido como o atributo a ser dividido [Tan *et al*. 2006]. As medidas mais comuns para este efeito são a *information gain* (ou entropia), *gain ratio* e *gini index*. Seja D, a partição de dados, um conjunto de treino de tuplos já classificados e N o nodo que representa os tuplos da partição D. E assumindo que a classe pode ter m valores distintos que definem m classes distintas, $C_i = (1,...,m)$. Em que $C_{i,D}$ é o conjunto de tuplos da classe C_i em D, e $|C_{i,D}|$ e $|D|$ representam o número de tuplos existentes em $C_{i,D}$ e D, respectivamente.

- ***Information Gain*:** Esta medida é usada no algoritmo ID3 [Quinlan 1986] (*Iterative Dichotomiser*) como medida de selecção de atributos. É baseada na teoria da informação de *Shanon*, responsável pelo estudo do conteúdo de informação nas mensagens. O atributo com a maior valor de *information gain* é escolhido como o atributo de divisão para o nodo N. Esta abordagem minimiza o número esperado de testes necessários para classificar um dado tuplo, e garante a geração de uma árvore simples [Han & Kamber 2006]. A informação necessária para classificar um tuplo em D é dada pela Equação (18).

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{18}$$

Em que p_i é a probabilidade de um tuplo pertencer à classe C_i. É usada uma função logarítmica de base 2, devido à informação ser codificada em bits.

Agora, supondo que uma partição é realizada nos tuplos em D num dado atributo A que contém v valores distinctos $\{a_1,a_2,...,a_v\}$, então o atributo A pode ser usado para dividir o tuplo D em v partições ou sub conjuntos $\{D_1,D_2,...,D_v\}$, em que D_j contém os tuplos em D consequentes pela partição sobre o atributo D. Neste caso, a informação é calculada pela Equação (19).

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{D} \times Info(D_j) \tag{19}$$

$Info_A(D)$ é a informação esperada necessária para classificar um tuplo de D baseado numa partição pelo atributo A. Quanto menor for a informação esperada necessária, maior é a considerada a pureza das partições. A *Information gain* é definida pela diferença entre a informação necessária original (baseada na proporção de classes) e a informação necessária (obtida após a partição de A) (Equação (20)).

$$Gain(A) = Info(D) - Info_A(D) \tag{20}$$

Por outras palavras, a Equação (20), indica qual será o ganho caso a partição seja realizada em A. O atributo que apresentar um maior ganho (*Gain*) é o escolhido como o atributo de divisão do nodo N.

- ***Gain ratio*:** A medida *information gain* é preferível quando os atributos apresentam uma vasta gama de valores. Por exemplo, considerando um atributo que apresenta um identificador único numa base de dados, a divisão desse atributo, iria resultar num número de partições iguais ao número de registos, em que cada partição conteria apenas um tuplo. Devido à informação ser pura, o resultado de *information gain* para este atributo seria 0, o que resultaria numa ganho máximo de informação. No entanto, uma partição deste género é completamente inútil na classificação. O algoritmo C4.5 [Quinlan 1993], um

sucessor do ID3, possiblita a escolha de uma extensão da medida *information gain*, denominada de *gain ratio*, que tem como objectivo diminuir o erro de tendência central [Han & Kamber 2006]. Nesta extensão é aplicada uma espécie de normalização do ganho da informação, usando um valor de informação de divisão definido analogamente pela medida *Info* (Equação (21)).

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \tag{21}$$

Este valor representa a potencial informação gerada através da divisão do conjunto de treino D, em v partições, correspondentes aos v resultados de um teste ao atributo A. De notar, que para cada resultado de v, é considerado o número de tuplos que apresentam esse resultado, em relação ao número total de tuplos em D (Equação (21)) [Han & Kamber 2006]. Através da Equação (20) e da Equação (21), obtemos a Equação (22), que representa a maneira como é calculado o *gain ratio*.

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \tag{22}$$

O atributo que apresentar um valor maior é seleccionado como o atributo de divisão. De notar que à medida que *SplitInfo* se aproxima de 0, o *gain ratio* torna-se instável. De maneira a ultrapassar este problema, uma restrição é adicionada, em que o ganho da informação do teste seleccionado tem de ser maior que a média do ganho de todos os testes examinados [Han & Kamber 2006].

- ***Gini Índex*:** Esta medida é usada no algoritmo CART [Breiman *et al*. 1984] (*Classification and Regression Trees*), e tem como objectivo, escolher o atributo que maximiza a redução da impureza. Para isso é necessário calcular primeiro a impureza de D, através da Equação (23).

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^{\,2} \tag{23}$$

Em que p_i é a probabilidade de um tuplo em D pertencer à classe C_i. A soma é calculada tendo como base as m classes. De seguida é necessário calcular a

impureza para cada atributo possível. Este cálculo é feito através de uma divisão binária para cada atributo. Considerando, inicialmente, o caso em que o atributo A é um valor discreto que tem v valores distintos que ocorrem em D. Para determinar a melhor divisão binária em A, é necessário examinar todos os subconjuntos que podem ser formados usando valores conhecidos de A. Cada subconjunto S_A pode ser considerado como um teste binário para o atributo A, na forma $A \in S_A$. Dado um tuplo, o teste é satisfeito, se o valor de A estiver entre os valores de S_A. Um exemplo deste tipo foi mencionado na altura em que se falou dos atributos discretos e binários. Quando acontece este caso, a medida *gini índex* é calculada através da uma soma ponderada da impureza de cada partição resultante (Han & Kamber 2006). Por exemplo, se uma divisão binária em A , particionar D em D_1 e D_2, o valor de $Gini(D)$ é dado pela Equação (24).

$$Gini(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \tag{24}$$

Para cada atributo, cada uma das possíveis divisões são consideradas. Para um valor discreto, o subconjunto que apresentar o menor valor para esse atributo, é seleccionado como sendo o subconjunto dividido. Para valores contínuos, cada ponto de divisão possível deve ser considerado, em que o ponto médio entre cada par de valores adjacentes é dado como um possível ponto de divisão. O ponto que apresente um menor valor *Gini índex*, para um dado atributo é considerado o ponto de divisão desse atributo (Han & Kamber 2006). De notar que para um possível ponto de divisão de A, D_1 é o conjunto de tuplos em D que satisfazem $A \leq split\_point$, e D_2 é o conjunto de tuplos em D que satisfazem $A > split\_point$.

Após o cálculo da impureza para cada possível atributo de divisão, de seguida é necessário calcular a diferença entre a impureza da partição de dados D, para a impureza de cada um dos atributos de divisão possível, através da Equação (25).

$$\Delta Gini(A) = Gini(D) - Gini_A(D) \tag{25}$$

O atributo que apresentar uma variação maior é considerado o atributo de divisão, pois é o que apresenta um menor grau de impureza.

Quando uma árvore de decisão é construída, muitos dos ramos irão reflectir anomalias no conjunto de treino devido ao ruído ou *outliers*. E quanto mais profundos forem os nós das folhas, menos exemplos têm a cobri-los, o que provoca uma significância estatística baixa. Por outras palavras, um modelo bem ajustado deve ter um valor de erro de treino baixo, bem como um valor de erro de generalização baixo. É importante salientar isto, pois um modelo que ajuste o conjunto de treinos bastante bem, pode ter um erro de generalização mais fraco do que um modelo que apresente um erro de treino elevado [Tan *et al*. 2006]. Esta situação é denominada de *overfitting* do modelo. A maneira de lidar com o problema de *overfitting* é a aplicação da técnica *pre-pruning* ou *post-pruning* à árvore em causa. Na abordagem da técnica *pre-pruning*, o algoritmo responsável pelo crescimento da árvore, apresenta uma condição de paragem de crescimento, que é activada antes que a árvore se ajuste perfeitamente aos dados do conjunto de treino. Essa condição de paragem, impõe que o crescimento de uma folha pare, assim que se observe que o ganho da medida de impurezas seja menor do que um certo limite mínimo imposto. A vantagem desta técnica previne a geração de sub árvores extremamente complexas, e que se sobre ajustem ao conjunto de treino. No entanto, a escolha do limite mínimo para a condição de paragem é considerado difícil [Tan *et al*. 2006]. No *post-pruning*, inicialmente, a árvore cresce para o seu tamanho máximo. Após o processo de crescimento da árvore, é feito o *pruning* da árvore das folhas à raíz. Esta "apara" da árvore é feita através da substituição de uma sub árvore, por uma nova folha em que a sua classe é definida pela classe mais frequente nessa sub árvore [Tan *et al*. 2006, Han & Kamber 2006]. Este processo termina quando mais nenhuma melhoria é observada. A técnica *post-pruning* tem tendência a produzir melhores resultados que a técnica *pre-pruning*, pois as decisão de *pruning* são feitas com base na árvore final obtida, enquanto que através da aplicação de *pre-pruning* a árvore pode ter uma terminação prematura do seu crescimento. No entanto, na aplicação de post-pruning são feitos cálculos computacionais adicionais, pois para ser aplicado o *pruning* é necessário, primeiro, a árvore chegar ao seu tamanho máximo [Tan *et al*. 2006].

Entretanto, mais recentemente foi criado uma extensão comercial do C4.5, o C5.0[2], que é considerado significamente mais rápido, produz árvores de decisão consideravelmente mais pequenas mas com a mesma precisão, permite a redução do ruído, é mais eficiente no uso da memória, entre outras vantagens.

2. *Naive Bayes Classifiers*: Os classificadores *bayesianos* são classificadores estatísticos. Estes classificadores têm como objectivo calcular a probabilidade de um novo caso pertencer a uma classe respectiva [Han & Kamber 2006]. Estudos de comparação de

[2] Para mais informações sobre o C5.0, consultar http://www.rulequest.com/see5-comparison.html

algoritmos de classificação levaram ao aparecimento de um classificador *bayesiano* simples, conhecido por *Naive Bayes* e que era comparável, em termos de desempenho, às árvores de decisão e às redes neuronais. O termo *Bayes* provém das regras de *Bayes*, criadas por *Thomas Bayes*, e o *Naive* provém da característica que esta técnica assume a independência dos atributos. Isto é, o naive bayes assume que o valor de um atributo numa dada classe é independente dos valores dos outros atributos. Este pressuposto faz aumentar a simplicidade da computação desta técnica, mas mesmo apesar desta simplicidade, esta técnica apresenta bons resultados quando testada em conjuntos de dados de grande dimensão [Witten & Frank 2005]. No entanto, como se constata, quando surge a necessidade de fazer previsões com base em atributos dependentes, esta técnica não apresenta nenhuma utilidade. Logo, para a representação da dependência entre os subconjuntos de atributos é usado a técnica Redes *Bayesianas*, que são modelos gráficos também usados na classificação [Han & Kamber 2006].

Seja D um conjunto de treino de tuplos já classificados respectivamente. Em que cada tuplo é representado por um vector de atributos n-dimensional, $A_1,A_2,...,A_n$, e supondo que existem m classes, $C_1,C_2,...,C_m$. Dado um tuplo X, o classificador irá prever qual a classe a que X pertence, através do cálculo da probabilidade de X para todas as classes existentes, e a que obter um maior valor é considerada a classe do tuplo X. Isto é, o classificador *naive bayes*, prevê que um tuplo X pertence a uma classe C_i, se e só se $P(C_i \mid X) > P(C_j \mid X)$, para $i \leq j \leq m$ e $j \neq i$. Deste modo $P(C_i \mid X)$ é maximizado, e essa classe é chamada de hipótese máxima *posteriori*. Como $P(X)$ é constante para todas as classes, só é necessário maximizar $P(X \mid C_i)P(C_i)$. No entanto, se as probabilidades das classes não forem conhecidas, então, é normalmente assumido que as classes são igualmente prováveis, isto é, $P(C_1) = P(C_2) = ... = P(C_m)$, e consequente só seria necessário maximizar $P(X \mid C_i)$. Caso contrário é mesmo necessário maximizar $P(X \mid C_i)P(C_i)$. A probabilidade de cada classe pode ser estimada por $P(C_i) = \frac{|C_{i,D}|}{|D|}$, em que $|C_{i,D}|$ é o número de tuplos de treino da classe C_i, existentes no conjunto de treino D. Através destas noções, os classificadores *bayesianos*, usam o Teorema de *Bayes*, representado na Equação (26).

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \tag{26}$$

No entanto, a Equação (26) só pode ser utilizada se for assumido que o conjunto de treino apenas possui um atributo. Pois para um conjunto de treino de diferentes atributos, o cálculo

de $P(X \mid C_i)$, seria computacionalmente dispendioso. De maneira a reduzir essa computação, os classificadores *bayesianos* assumem que os valores dos atributos são incondicionalmente independentes uns dos outros, dada a classe do tuplo (Han & Kamber 2006). Através desta assumpção é possível calcular a probabilidade da classe C_i gerar o tuplo X, através da Equação (27).

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) = P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times ... \times P(x_k \mid C_i) \tag{27}$$

Em que x_k refere-se ao valor do atributo A_k, para o tuplo X. Desta forma, facilmente se consegue estimar a probabilidade de $P(x_1 \mid C_i)$, $P(x_2 \mid C_i)$,..., $P(x_k \mid C_i)$, dos tuplos de treino. Para cada atributo é necessário ter em atenção o seu tipo, isto é, se é discreto ou contínuo. Para isso de seguida é mencionado as diferenças a ter em causa para um dos diferentes tipos:

- **Atributos discretos:** Se A_k é discreto, então $P(x_k \mid C_i)$ é o número de tuplos da classe C_i em D, tal que o valor x_k pertença a A, a dividir pelo número de tuplos da classe C_i em D, $|C_{i,D}|$ [Witten & Frank 2005].
- **Atributos contínuos:** Caso A_k seja contínuo, é necessário executar mais alguns cálculos. Para um atributo contínuo assume-se, normalmente, uma distribuição gaussiana com uma média μ, e com um desvio padrão de σ, definida pela Equação (28).

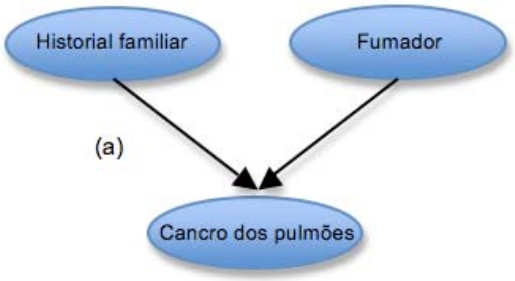
$$g(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{28}$$

  Logo, $P(x_k \mid C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i})$. No entanto antes de se calcular a distribuição gaussiana é necessário calcular a média μ_{C_i}, e o desvio padrão σ_{C_i}, dos valores dos atributos A_k para os tuplos de treino da classe C_i, e só depois é calculada a Equação (27), de maneira a estimar $P(x_k \mid C_i)$ [Witten & Frank 2005].

Concluindo, para se prever a classe de X, $P(X \mid C_i)P(C_i)$ é calculado para cada classe C_i. Assim, o classificador prevê que a classe do tuplo X é a classe C_i, se e só se $P(X \mid C_i)P(C_i) > P(X \mid C_j)P(C_j)$, para $i \le j \le m$ e $j \neq i$. Por outras palavras, a classe prevista é a classe C_i para o qual $P(X \mid C_i)P(C_i)$ seja máximo [Han & Kamber 2006].

Vários estudos empíricos deste classificador, em comparação com as árvores de decisão e as redes neurais, indicaram que o classificador *naive bayes* é o que apresenta a taxa mínima de erro. Na prática, este classificador, não apresenta sempre os melhores resultados, devido à restrição de apenas classificar atributos independentes. Nesse contexto, as redes *bayesianas*, são redes probabilísticas, que permitem, dado a classe de um tuplo, a independência incondicional entre os valores dos atributos [Han & Kamber 2006]. Isto é, aplica o mesmo conceito que o *naive bayes* (conceito de probabilidades), permitindo a independência entre os atributos. As redes *bayesianas* (Figura 11 (a)) pertencem à família dos modelos de gráficos probabilísticos, mais concretamente aos grafos acíclicos directos. Estas estruturas de grafos são usadas para representar o conhecimento acerca de um domínio específico e são definidas por dois conjuntos, o conjuntos dos nodos (vértices) e o conjunto de arestas directas. Os nodos representam variáveis aleatórias e são desenhadas como círculos associados a cada nome das variáveis. As arestas representam as dependências directas entre as variáveis e são desenhadas como setas entre os nodos [Ben-Gal 2007]. Mais concretamente, uma aresta do nodo x_i para o nodo x_j representa uma dependência estatística entre as respectivas variáveis. Neste caso, o nodo x_i é denominado de pai de x_j, e logicamente, x_j é denominado de filho de x_i.

Além da de um gráfico acíclico directo, as redes *bayesianas* também utilizam uma tabela de probabilidades condicionais (Figura 11 (b)), para cada variável. Nessa tabela é listado a probabilidade local que um nodo filho assume para cada um dos valores possíveis, isto é, para cada combinação de valores dos seus pais [Ben-Gal 2007].



(b)

	HF,S	HF,~S	~HF,S	~HF,S
CP	0.8	0.5	0.7	0.1
~CP	0.2	0.5	0.3	0.9

Figura 11 – Exemplo da estrutura de uma rede neural, em que (a) representa um grafo acíclico e (b) representa a tabela das probabilidade condicionais para a variável cancro dos pulmões [Han & Kamber 2006].

Uma rede *bayesiana* B, é um grafo acíclico que representa a dependência entre variáveis aleatórias V. A rede é definida por um par $B = \langle G, \Theta \rangle$, tal que G representa o grafo acíclico

directo em que os nodos $X_1,X_2,...,X_n$ representam variáveis aleatórias, e em que as arestas representam as dependências directas entre essas variáveis. O Θ denota o conjunto de parâmetros da rede. Este conjunto contém os parâmetros $\theta_{x_i|\pi_i} = P(x_i \mid \pi_i)$ para cada realização de x_i em X_i, condicionado por π_i, o conjunto dos pais de X_i em G. Por consequência, B define uma distribuição conjunta através de V, pela Equação (29).

$$P(x_1,...,x_n) = \prod_{i=1}^{n} P(x_i \mid \pi_i) \tag{29}$$

Em que $P(x_1,...,x_n)$ determina de forma unívoca, a distribuição conjunta para os valores de X [Ben-Gal 2007, Han & Kamber 2006].

3. ***Lazy Learners*****:** A técnica mais conhecida do conjunto de técnicas desta categoria é o *K-Nearest Neighbors* (*K-NN*). De uma maneira básica, a técnica dos k vizinhos mais próximos consiste em classificar uma nova instância com base no cálculo da distância dos k vizinhos mais próximos. Isto é, depois de se calcular os k vizinhos mais próximos, a classe predominante nesse conjunto será a classe atribuída à nova instancia [Hand *et al*. 2001]. De notar que a distância é definida em termos de um dado espaço p-dimensional. A partir desta pequena introdução surgem logo duas questões, qual o valor de k e qual a métrica utilizada para o cálculo da distância.

A forma mais básica desta técnica, assume um $k = 1$, mas isso cria um classificador bastante instável, e normalmente as previsões são mais consistentes quando $k > 1$, pois a nova instância a classificar pode ser mais similar a outra classe, do que à classe do vizinho mais próximo [Cover & Hart 1967, Hand *et al.* 2001]. No entanto, o aumento do k significa que poderão existir casos no conjunto de treino, que poderão ser incluídos no cálculo dos vizinhos mais próximos, mas que, devido à escolha de um k grande, não estão, necessariamente, perto do novo caso. Isto significa que o objectivo desta técnica de estimar a classe de uma nova instância com base nos vizinhos mais próximos, pode não ser funcional na prática. Isto é, como a associação da classe a uma nova instância é feita através da classe predominante dos vizinhos mais próximos, se a classe predominante corresponder ao vizinhos mais distantes, dentro dos k vizinhos mais próximos, a previsão da classe pode não ser considerada a mais correcta. Apesar deste problema, não existe nenhuma forma teórica de como calcular o melhor k, pois depende da estrutura particular do conjunto de dados. No entanto, a melhor maneira de calcular o k, é através de tentativas de vários valores, e o que apresentar melhores resultados, é escolhido como o k final do estudo [Hand *et al.* 2001]. Após a escolha do k, também é necessário escolher uma medida para calcular a distância

entre os diferentes casos existentes e o novo. Normalmente a métrica utilizada para atributos contínuos é a medida euclidiana (Equação (10)), no entanto, esta medida assume que todos os atributos apresentam igual importância [Witten & Frank 2005]. Este problema pode ser ultrapassado atribuindo diferentes pesos aos atributos [Hand *et al.* 2001] no cálculo da distância. Para atributos discretos, a distância é zero quando são idênticos, e um no caso contrário. Isto é, se o atributo em causa for referente a cores, a distância entre verde e verde será 0, mas a distância entre verde e azul será 1 [Witten & Frank 2005].

Esta técnica apresenta inúmeras propriedades atractivas, é fácil de programar e nenhuma optimização ou treino é necessário. Apresenta resultados bastante bons para espaços dimensionais pequenos. No entanto, é uma técnica que, computacionalmente, é bastante pesada, pois numa fase inicial, este tipo de algoritmos necessitam de calcular todas as distâncias entre as instâncias de treino e a nova instância, e de seguida é necessário guardar todas essas distância, o que leva a uma utilização excessiva de memória [Hand *et al*. 2001].

4. *Neural Networks*: As redes neuronais surgiram como uma importante técnica da classificação. Vastos estudos na classificação neuronal estabeleceram que as redes neuronais são um promissora alternativa às técnicas convencionais da classificação. A vantagem desta técnica provém de quatro aspectos. Em primeiro lugar, as redes neuronais são técnicas que conseguem facilmente se ajustar aos dados sem que nenhuma especificação funcional ou espacial do modelo em causa, seja feita. Em segundo lugar consegue aproximar qualquer função, com uma precisão arbitrária, sem ter que procurar uma relação funcional entre o conjunto de treino e os atributos do novo caso. Em terceiro lugar, as redes neuronais são modelos não lineares, o que as torna flexíveis na modelação de relações complexas do mundo real. E por fim, as redes neuronais conseguem estimar as probabilidades posteriores, que fornecem as bases para a formação de regras de classificação, e realizar análises estatísticas [Zhang 2000]. De uma maneira geral, uma rede neuronal é uma estrutura de dados construída através de uma rede de neurónios, que são funções, que pegam em um ou mais valores e retornam como saída a classe que deve ser atribuída à nova instância. Durante a fase de treino, os dados são alimentados à rede neuronal um por um, e as funções dos neurónios são modificadas com base nas taxas dos erros dos resultados de saída. É necessário a execução de múltiplos passos sobre os dados, de maneira a ser possível prever a classe correcta dos tuplos de entrada. Este tipo de técnica envolve tempos de treino bastante longos, por isso, a utilização desta técnica só é possível para a aplicações em que o longo tempo de espera seja uma opção [Aggrawal & Yu 1999, Han & Kamber 2006]. Apesar da desvantagem do tempo necessário de execução de uma rede neuronal, esta técnica também foi criticada pela sua fraca interpretação aos olhos humanos, o que levou, inicialmente, a um menor interesse desta técnica [Han & Kamber 2006]. No entanto, as redes

neuronais são bastante tolerantes ao ruído nos dados, e adaptam-se muito bem a valores contínuos de entrada e saída, ao contrário da maior parte dos algoritmos das árvores de decisão [Han & Kamber 2006].

Existem diferentes tipos de redes neuronais e algoritmos de redes neuronais. O algoritmo mais comum desta técnica é o *backpropagation* [Rumelhart *et al*. 1986], que ganhou reputação em meados do século 19 [Han & Kamber 2006].

2.6.3 Avaliação do Desempenho de um Modelo

É, normalmente, útil medir o desempenho de um modelo no conjunto de treino, devido a tal medida fornecer uma estimativa imparcial da taxa de erro. A precisão ou a taxa de erro, medida através do conjunto de treino, pode também ser comparada com o desempenho relativo dos diferentes classificadores no mesmo domínio. No entanto, para isso acontecer, é necessário conhecer as respectivas classes dos registos de teste [Tan *et al*. 2006]. De seguida são apresentadas quatro métodos que são usados para a classificação do desempenho de um classificador.

1. *Holdout Method*: Nesta abordagem, os dados em causa são aleatoriamente repartidos em dois conjuntos independentes, o conjunto de treino e o conjunto de teste. Normalmente, o conjunto de treino é dois terços do conjunto de dados inicial, e o um terço que sobra é considerado o conjunto de teste [Han & Kamber 2006, Tan *et al*. 2006]. O conjunto de treino é responsável pela criação do modelo, em que a precisão é determinada pelo o conjunto de teste. Esta abordagem apresenta bastantes limitações. Primeiro, menos exemplos classificados estão disponíveis para treino, pois alguns dos registos estão reservados para teste, no entnto, o modelo induzido pode não ser tão bom, caso todos os exemplos classificados sejam usados para treino. Segundo, o modelo pode ser bastante dependente da composição dos conjuntos de treino e teste, isto é, quanto maior for o tamanho do conjunto de treino, menos confiável é a precisão estimada do conjunto de testes. Por outro lado, quanto menor for o tamanho do conjunto de treino, maior é a variância do modelo. Finalmente, devido aos conjuntos de treino e teste serem subconjunto dos dados iniciais, uma classe que esteja bastante representada num subconjunto, está pouco representada noutro, e vice versa [Tan *et al*. 2006].

2. *Random Subsampling*: Este método é uma variação do *holdout*, em que neste caso é possível repetir, os passos executados no método *holdout*, diversas vezes de maneira a melhorar a estimativa do desempenho do classificador [Han & Kamber 2006, Tan *et al*.

2006]. No entanto, este método também apresenta algumas desvantagens do *holdout*, pois também, neste método, não é utilizado a totalidade dos dados iniciais. Este método não apresenta nenhum controlo no número de vezes que cada registo é usado para treino e teste, consequentemente, alguns registos podem ser usados para treino mais vezes que outros registos [Tan *et al*. 2006].

3. *Cross Validation*: Uma alternativa ao método anterior é o *cross validation* (validação cruzada). Nesta abordagem cada registo é usado o mesmo número de vezes para treino e apenas uma vez para teste. Inicialmente os dados são divididos em k subconjuntos iguais. De seguida, numa primeira iteração, uma das partições é escolhida para teste, enquanto que as restantes são usadas para treino. Numa segunda iteração é escolhido outra partição para testes e as restantes (a contar com a partição de teste na 1ª iteração) são usadas como treino. Este processo é repetido k vezes (iterações), de maneira a que cada partição seja usada para teste apenas uma vez. Ao contrários dos dois métodos explicados anteriormente, aqui, cada partição é usada o mesmo número de vezes para treino e apenas uma vez para teste. O cálculo da precisão de um modelo, é o número total de classificações correctas nas k iterações, dividido pelo número total de tuplos dos dados iniciais. Um caso especial deste método utiliza o tamanho do conjunto de dados como sendo o k. Este caso é denominado de *leave-one-out*, que significa que um conjunto de teste apenas contém um registo. Esta abordagem apresenta vantagens ao utilizar o maior número de dados possíveis para treino. No entanto, o problema de repetir N (tamanho do conjunto de dados inicial) vezes este processo, leva a uma desvantagem computacional excessiva [Han & Kamber 2006, Tan *et al*. 2006].

4. *Bootstrap*: Os métodos já abordados assumem que os registos de treino são recolhidos aleatoriamente sem substituição, consequentemente, não existem registos duplicados nos conjuntos de treino e teste. Neste método, os registos de treino são recolhidos aleatoriamente com substituição. Isto quer dizer, que cada vez que um tuplo é seleccionado, é igualmente provável que seja seleccionado outra vez, e seja de novo adicionado ao conjunto de treino. Se os dados originais tiverem N registos (tuplos), então o conjunto de treino ou amostra *bootstrap* de tamanho N, contém, aproximadamente, 63,2% dos registos dos dados originais. Esta aproximação é feita através da probabilidade de um registo ser escolhido por uma recolha aleatório ser calculada por $1-\left(1-\frac{1}{N}\right)^N$. Quando o N é suficiente grande, a probabilidade aproxima-se de $1-e^{-1}=0.632$. Os tuplos dos dados que não pertencerem ao conjunto de treino são usados no conjunto de testes. Existem vários métodos deste género, o mais usado é o *.632 bootstrap* [Han & Kamber 2006, Tan *et al*. 2006].

2.6.4 Casos de Estudo

Os modelos de classificação, são usados para prever a classe de registos desconhecidos. Este tipo de modelos funcionam como uma caixa preta que automaticamente atribui uma classe, assim que é dado como entrada, um conjunto de atributos de um registo desconhecido [Tan *et al*. 2006]. Esta técnica apresenta bastantes funcionalidades quando aplicada a casos do mundo real. Por exemplo, através da análise de dados bancários, é possível descobrir quais as aplicações de empréstimo que apresentam um risco de crédito para o banco, ou são consideradas créditos seguras. Outro exemplo, este na área da saúde, um médico analisa os dados referentes ao cancro da mama de maneira a prever qual o tratamento especifico que a paciente deverá receber de acordo com a sua fase em que se encontra o cancro.

As áreas exploradas na segmentação, também são exploradas na classificação, dependo do tipo de análise que queremos. Como foi visto, os casos de estudo relativos à aplicação de técnicas de segmentação tinham o objectivo de analisar acontecimentos. Um dos casos foi a análise do cancro do mama, em que foi observado uma maior tendência do aparecimento deste tipo de cancro nas mulheres afro-americanas com idades inferiores a 50. Na mesma área é possível, através de técnicas de classificação, detectar certos acontecimentos, como a existência de cancro ou não, numa pessoa, ou mesmo prever a taxa de sobrevivência dependendo de vários factores. Em 2004 [Delen *et al*. 2004], foi realizado um estudo inicial, com o objectivo da previsão de sobrevivência de mulheres afectadas pelo cancro da mama. No entanto mais recentemente, apareceu outro estudo baseado no mesmo contexto, que vinha colmatar algumas falhas apontadas no estudo de 2004. Esse estudo foi levado a cabo por *Bellaachia* e *Guven* [Bellaachia & Guven 2006], e teve em atenção dois novos campos. No estudo realizado anterior (2004), existiam dois grupos, os sobreviventes e os não sobreviventes, que dependiam do tempo de sobrevivência recorde. Isto é, os registo em que o campo "*Survival Time Recode*" (STR) era superior a 60 meses, eram considerados como sendo sobreviventes, caso contrário, não sobreviventes. No entanto o número de casos não sobreviventes, não era propriamente o número de casos não vivos "*Vital Status Recode*" (VST), e daí ter sido inserido um novo campo para esta hipótese. Outro caso considerado neste novo caso, foi a causa de morte das pessoas relacionadas no estudo, isto é, nem todas as pessoas morreram devido à existência de cancro, mas de outros factores alheios a este problema. Devido a isso o campo "*Cause of Death*" (COD), também foi tido em conta para este novo estudo. Os dados utilizados para ambos os estudos são os mesmos e foram provenientes do "SEER *Cancer Incidence Public-Use Database*" para os anos de 1993 a 2000, e eram compostos por 433,272 registos. A grande diferença destes estudos, foi um animador

aumento da percentagem da taxa de sobrevivência, de 41.7% para 76.8%, e de certa maneira uma maior realidade sobre este assunto, pois foi abordado tendo em conta mais factores externos. Apesar de o cancro da mama ter sido o mais explorado neste trabalho, o estudo através da utilização de técnicas de classificação abrange todos os tipos de cancro existentes [Wu *et al*. 2003], bem como outros problemas de saúde, cujo o objectivo é a previsão de algo, como um diagnóstico [Zelic *et al*. 1997] ou tipo de tratamento mais recomendado.

No entanto o uso da classificação é também bastante necessário na previsão de outros acontecimentos, para além da saúde. Existem alguns estudos na base da educação, com o objectivo de prever o sucesso escolar de alunos, e consequentemente aplicar novas metodologias com o intuito de combater o insucesso escolar. Um estudo realizado por *Merceron* e *Yacef* [Merceron & Yacef 2005], analisou uma ferramenta escolar (*Logic-ITA* [Yacef 2005]), responsável pelo armazenamento de todos os registos dos trabalhos realizados por alunos e professores da universidade de *Sidney,* desde 2001. Com base nos registos existentes nessa ferramenta, já tinham sido abordadas novas metodologias na maneira de ensino, que resultou numa melhoria do número de alunos aprovados de 2003 para 2004. No entanto, o objectivo principal deste estudo foi prever quais os alunos que poderiam estar em risco de reprovar, e caso isso fosse provável, fornecer-lhes um maior apoio. Outro estudo baseado nesta ideia foi realizado por *El-Halees* [El-Halees 2008] na universidade islâmica de Gaza. Em 2007 [Romero & Ventura 2007] foi apresentado uma abordagem geral, dos estudos realizados através de técnicas de mineração de dados, em termos educacionais.

Outro abordagem bastante importante, principalmente nos dias de hoje, foi realizada em 2001 por *Schultz et al*. [Schultz et al. 2001], sobre a detecção de executáveis maliciosos. Este estudo foi realizado tendo como base o sistema operativo da *Microsoft*, com um total de 4.266 programas, sendo 3.256 considerados maliciosos (5% eram *trojans* e 95% vírus) e 1.001 considerados limpos. Foi criado uma ferramenta baseada no algoritmo *naive bayes* para a realização deste estudo e foi conseguido uma percentagem de previsão correcta de 97.76%. Os anti-vírus são normalmente actualizados uma vez por mês, nesse mesmo período de tempo, 240 a 300 novos executáveis maliciosos são criados. Esta nova ferramenta permite a detecção de 216 a 270 destes novos programas maliciosos, enquanto que os métodos tradicionais só conseguem detectar cerca de 87 a 109 desses novos executáveis. O anti-vírus utilizado nestes testes foi o *MacAfee's* [Schultz *et al*. 2001]. Este estudo, foi como uma extensão ao estudo realizado por investigadores da IBM[3], em 1996 através da técnica de

[3] http://www.research.ibm.com/antivirus/SciPapers/Tesauro/NeuralNets.html.

redes neuronais. As redes neuronais são usadas em inúmeros aplicações de negócios. *Vellido et al.* [Vellido et al. 1999] apresentaram em 1999, um estudo global da aplicação de redes neuronais na área dos negócios, entre os anos de 1992 e 1998.

A partir do 11 de Setembro de 2001, tem havido um maior controlo sobre o terrorismo no mundo. A classificação assimila as propriedades comuns entre os diferentes crimes, e organiza-os em diferentes classes. Isso permite identificar a fonte de *e-mails* com base nos padrões linguísticos do remetente e características estruturais. Esta técnica permite ainda prever padrões de crimes, e reduzir o tempo de procura dos causadores dos crimes [Chen *et al*. 2004]. Todas estas vantagens possíveis através da classificação, foram abordadas através da criação de uma nova ferramenta em 2004 [Chen *et al*. 2004], com base na experiência do projecto *Coplink*[4], um projecto levado a cabo por investigadores da universidade de Arizona com os departamentos policiais de *Tucson* e *Phoenix*. Os testes realizados com esta nova ferramenta foram realizados sobre a base de dados de *Tucson* que continha um total de 1.3 milhões de suspeitos e que apresenta dados desde o ano de 1970.

Muitas das aplicações da classificação são complementadas com o estudo da segmentação, exemplo disso são alguns dos estudos [Romero & Ventura 2007, El-Halees 2008, Chen *et al*. 2004] mencionados acima.

2.7 Análise Geral das Técnicas

A escolha de uma técnica de mineração de dados depende do objectivo que se pretende. É impossível avaliar uma técnica pelo seu desempenho, pois a utilização de cada técnica depende do objectivo final que se pretende obter. No entanto, entre cada técnica é possível avaliar os algoritmos conforme a características dos dados, e as características de funcionamento do próprio algoritmo. Mas antes de se proceder a uma comparação de algoritmos é necessário saber quais os parâmetros a ter em conta, e para diferentes técnicas e diferentes algoritmos, existem diferentes parâmetros a ter em conta. Por isso, para cada técnica será apresentado as características de cada algoritmo, de acordo com determinados parâmetros a ter em conta.

[4] http://ai.bpa.arizona.edu/coplink.

2.7.1 Associação

A associação é usada com o objectivo de se descobrir as relações existentes entre os dados, isto é, encontrar a natureza das casualidades entre os valores dos diferentes atributos [Aggarwal & Yu 1999]. Como foi visto na secção 2.4.2, os algoritmos para a descoberta de regras de associação podem ser subdivididos em quatro tipos, em que cada tipo apresenta vantagens e desvantagens relativamente a cada um. A comparação de um algoritmo para a descoberta de regras é avaliado de acordo:

- **Escalabilidade:** Habilidade de manipular um conjunto de dados em termos de quantidade.
- **Padrões frequentes:** Habilidade de trabalhar com conjuntos de elementos frequentes, se grandes (longos) ou pequenos.
- **Scans da base de dados:** Quantidade de passagens necessárias pela base de dados, para obtenção das regras.
- **Desempenho:** Custos computacionais envolvidos na geração de regras de associação.

A Tabela 6, é apenas uma tabela que pode servir de guia conforme as características que o estudo apresenta. Isto é, apesar do *Apriori* apresentar uma complexidade elevada e executar bastantes passagens pela base de dados, em comparação com o *Eclat*, quando a base de dados em estudo apresenta padrões frequentes substancialmente pequenos, o *Apriori* apresenta resultados bastante melhores do que o *Eclat* [Han & Kamber 2006] e que os outros algoritmos [Zheng *et al*. 2001]. No entanto, tendo em conta padrões frequentes densos (longos), o algoritmo *Partition* apresenta melhores resultados que o *Eclat*, mesmo tendo uma complexidade mais elevada [Goethals 2003]. O *Eclat* e o *FP-Growth* apresentam bastantes similaridades no seu funcionamento, mas a principal diferença entre ambos é a sua complexidade computacional, isto é, apesar de a complexidade de ambos ser baixa, o *Eclat* consegue obter uma complexidade bem mais baixa do que o *FP-Growth* [Goethals 2003].

Associação		Escalabilidade	Padrões Frequentes	Scans	Desempenho
	BFS e contagem de ocorrências				
	Apriori	Grande/Pequens	Pequenos	Muitos	Elevado
	BFD e intersecções TID-List				
	Eclat	Grande	Grandes	Poucos	Baixo
	DFS e contagem de ocorrências				
	Partition	Grande	Pequenos/Grandes	Poucos	Elevada
	DFS e interseções TID-List				
	FP-Growth	Grande	Pequenos/Grandes	Poucos	Baixo

Tabela 6 – Avaliação dos algoritmos de regras de associação [Zheng *et al*. 2001, Goethals 2003, Han & Kamber 2006, Hipp *et al*. 2000].

2.7.2 Segmentação

Como foi mencionado na secção 2.5, a segmentação é uma técnica responsável pela geração de modelos descritivos, em que o objectivo é agrupar os registos em segmentos similares, de maneira a ser possível a análise do comportamento da respectiva base de dados [Aggarwal & Yu 1999]. Nessa secção, foram abordados 5 técnicas, no entanto a avaliação dos algoritmos de segmentação será feita com base em apenas quatro: *Hierarchical*, *Partitional*, *Density based* e *Grid based*. Para a avaliação de algoritmos de segmentação é necessário ter em conta [Andritsos 2002, Han & Kamber 2006]:

- **Parâmetros de Entrada:** É necessário ter em conta os parâmetros de entrada dos algoritmos de segmentação, pois alguns poderão ser bastantes sensíveis em relação à escolha desses parâmetros.
- **Escalabilidade:** Habilidade de manipular um conjunto de dados em termos de quantidade.
- **Dimensionalidade dos dados:** Habilidade de trabalhar com dados que apresentem uma grande dimensão, isto é, que consigam trabalhar com dados que apresentem um número superior a 10 atributos (dimensões).
- **Forma dos segmentos:** A forma normalmente corresponde ao tipo de segmentos que um algoritmo consegue encontrar.
- **Habilidade de trabalhar com ruído nos dados:** Como o próprio nome indica, é necessário ter em conta os algoritmos que conseguem produzir bons resultados mesmo na existência de *outliers*.
- **Desempenho:** Custos computacionais envolvidos na geração de modelos de segmentação.

A Tabela 7, apresenta as principais diferenças entre os algoritmos de segmentação. Mais uma vez, é referido, que esta tabela não serve para escolher qual o melhor algoritmo, mas sim para ajudar na escolha do utilizador, consoante as características da base de dados em estudo. No entanto, facilmente se constata que o CLARANS é melhor que o PAM e o CLARA, porque abrange um maior número de opções existentes no mundo real, isto é, apresenta uma melhor computação computacional e normalmente apresenta melhores segmentos [Andritsos 2002]. A única desvantagem do CLARANS é a necessidade de mais um parâmetro de entrada, o que por vezes, a escolha destes parâmetros torna-se bastante difícil, pois pode não corresponder à melhor escolha.

2.7.3 Classificação

O objectivo desta técnica, passa pela previsão de classes, com base na aprendizagem do comportamento dos dados, aprendizagem supervisionada [Aggarwal & Yu 1999]. Cada uma das técnicas apresenta as respectivas vantagens e desvantagens, de acordo com os seguintes parâmetros [Han & Kamber 2006]:

- **Precisão:** Precisão final do modelo final gerado, na previsão da classe para novas instâncias.
- **Escalabilidade:** Habilidade de manipular um conjunto de dados em termos de quantidade.
- **Habilidade de trabalhar com ruído nos dados:** Como o próprio nome indica, é necessário ter em conta os algoritmos que conseguem produzir bons resultados mesmo na existência de *outliers*.
- **Compreensão dos resultados:** A compreensão é subjectiva, mas mesmo assim existem resultados, que apresentam uma noção geral de mais ou menos compreensíveis, ao mais comum dos humanos.
- **Desempenho:** Custos computacionais envolvidos na geração de modelos de classificação.

Na classificação, os algoritmos de árvores de decisão são bastantes similares, em que a única mudança entre o ID3, CART e C5.0, consiste na medida de escolha do atributo de corte e na escolha do método de aplicação de *pruning* [Han & Kamber 2006]. A escolha da melhor medida para a selecção do atributo de corte não é fácil, pois cada uma apresenta vantagens e desvantagens relativamente às outras. Tal como na escolha da melhor medida, a escolha do melhor do método de aplicação de *pruning*, se *prepruning* ou *post-pruning*, é complicado, como se viu na secção 2.6.2.

Em relação aos classificadores *bayesianos*, não existe muito a noção de qual é melhor, pois ambos existem com diferentes propostos. Isto é, a aplicação do *naive bayes* é proposta na aplicação de atributos independentes, enquanto que a aplicação do *bayesian belief networks* é como uma técnica contemplar ao *naive bayes*, que apresenta o mesmo funcionamento, mas permite a utilização sobre atributos dependentes, no entanto apresenta um complexidade maior. Em suma, a Tabela 8, pode ajudar na descoberta de um modelo de classificação através da escolha de uma técnica geral, de acordo com o tipo de dados. Se possível, optar por aplicar mais do que uma técnica ao mesmo conjunto de dados e, a partir das técnicas de avaliação de um modelo, descobrir a técnica que melhor se adapta aos dados em estudo.

Segmentação		Parâmetros de Entrada	Escalabilidade	Dimensionalidade	Forma	Ruído	Desempenho
	Hierarchical						
	BIRCH	Factor de ramificação, diâmetro máximo	Grande	Pequena	Esférica	Sim	$\Theta(n)$
	CURE	Número de segmentos, número de segmentos representativos	Média	Pequena	Arbitrária	Sim	$\Theta(n^2 \log n)$
	Partitional						
	K-MEANS	Número de segmentos	Grande/Pequena	Grande	Esférica	Não	$\Theta(lkn)$
	CLARA	Número de segmentos	Média	Pequena	Esférica	Não	$\Theta(ks^2 + k(n-k))$
	PAM	Número de segmentos	Pequena	Grande	Esférica	Não	$\Theta(lk(n-k)^2)$
	CLARANS	Número de segmentos, número máximo de vizinhos	Grande	Grande	Esférica	Não	$\Theta(kn^2)$
	Density based						
	DBSCAN	Raio dos segmentos, número mínimo de pontos nos segmentos	Grande	Grande	Arbitrária	Sim	$\Theta(n \log n)$
	DENCLUE	Raio dos segmentos, número mínimo de objectos	Grande	Grande	Arbitrária	Sim	$\Theta(n \log n)$
	Grid based						
	STING	Número de células no nível mais baixo, número de objectos numa célula.	Grande	Grande	Limites verticais e horizontais	Sim	$\Theta(n)$
	CLIQUE	Tamanho da rede, número mínimo de objectos numa célula	Grande	Grande	Arbitrária	Sim	$\Theta(n)$

n = número de objectos, k = número de segmentos, s = tamanho da base de dados, l =número de iterações

Tabela 7 – Avaliação dos algoritmos de segmentação [Andritsos 2002, Han & Kamber 2006].

Classificação		Precisão	Escalabilidade	Ruído	Compreensão	Desempenho
	Decision Trees					
	ID3	Boa	Pequena	Não	Fácil	$\Theta(n \log n)$
	CART	Boa	Grande	Não	Fácil	$\Theta(n \log n)$
	C5.0	Melhor que ID3 e CART	Grande	Não	Fácil	$\Theta(n \log n)$
	SPRINT	Boa	Grande	Não	Fácil	$\Theta(n \log n)$
	Naive Bayes Classifiers					
	Naive Bayes	Bastante Boa	Grande	Sim	Fácil	$\Theta(\log n)$
	Bayesian Belief Networks (TAN)	Bastante Boa	Grande	Sim	Fácil	$\Theta(n^2 \log n)$
	Lazy Learners					
	K-NN	Boa	Pequena	Sim	Fácil	$\Theta(n^2)$
	Neural Networks					
	Backpropagation	Bastante Boa	Grande	Sim	Difícil	$\Theta(n^3)$

n = número de objectos

Tabela 8 – Avaliação dos algoritmos da classificação [Han & Kamber 2006, Boullé 2007, Friedman *et al*. 1997, Cunningham & Delary 2007, Orponen 1994].

Capítulo 3

Descoberta de Padrões

3.1 Um Caso de Estudo

A informação, apresenta hoje uma fonte de conhecimento bastante valiosa que contribui para o acréscimo da qualidade de vida das pessoas, através da tomada de decisões suportadas por um conjunto de informação recolhida e analisada ao pormenor [Tomé *et al*. 2008]. Consequentemente, a introdução do registo electrónico na prática clínica, abrangendo todos os passos existentes de uma consulta, pode servir de apoio à decisão e à monitorização dos serviços prestados, quer aos utentes, quer aos clínicos ou mesmo aos gestores. Apesar de existir um número relativamente baixo sobre os estudos efectuados sobre a avaliação da eficácia destes sistemas, os resultados já obtidos sugerem um aumento da segurança dos pacientes e uma redução de gastos, tanto hospitalares, como dos centros de saúde, ou outras entidades relacionadas.

Os registos electrónicos em saúde tem apresentado uma evolução bastante satisfatória, nomeadamente no Reino Unido, onde aproximadamente 100% dos médicos, a nível dos cuidados de saúde primários, utilizam esses serviços. No entanto, apesar de em Portugal essa ainda não ser a realidade, várias acções têm sido implementadas com esse intuito, em contraste com os Estados Unidos da América, onde a taxa de utilização ronda apenas os 15%. A utilização destes sistemas em Portugal pode ser feita através de três sistemas existentes o SAM, o MedicineOne e o VITACARE que permitem a prescrição electrónica [Tomé *et al*. 2008]:

- **SAM (Sistema de Apoio ao Médico).** Este sistema foi inicialmente desenvolvido tendo em conta questões económicas e não clínicas. No entanto, apresenta um grande potencial de melhoria, e comparativamente à prescrição manual é considerado uma mais-valia.
- **MedicineOne.** O MedicineOne foi concebido tendo já em conta a prestação de apoio clínico ao prescritor, sem que as preocupações financeiras fossem descuradas.
- **VITACARE.** O VITACARE, tem como objectivo principal a preocupação pelo utente. Isto indica que todos os registos efectuados, independentemente dos "factores externos" (como o médico), estão associados ao utente.

A aplicação SAM foi o projecto piloto instalado em Portugal, inicialmente, em cinco centros de saúde, em Julho de 2000. A partir do ano de 2002, a exploração desta nova aplicação por parte dos médicos é já uma realidade a nível nacional. Esta aplicação permite efectuar o registo completo do motivo da consulta e o respectivo diagnóstico do utente em causa, possibilitando ainda o registo dos medicamentos prescritos e dos meios complementares utilizados na análise clínica do utente. Consequentemente, é possível identificar os medicamentos prescritos e as respectivas quantidades em cada consulta, de cada utente, desagregada por idade e sexo. A cada prescrição é associado o código de embalagem de um medicamento, o que possibilita a extracção de informação útil, como a substância activa, a dosagem do respectivo medicamento, a quantidade da embalagem em causa ou a forma terapêutica [Carvalho 2008]. Para o sistema SAM, a base de dados de medicamentos é da responsabilidade da INFARMED, sendo actualizada de forma automática em todos os centros de saúde abrangidos pela aplicação. No entanto, essas actualizações nem sempre são funcionais, excepto quando um medicamento é retirado do mercado, em que o sistema não permite a prescrição desse medicamento e alerta para esse facto. O problema funcional existente, assenta na não distinção entre os medicamentos com *Autorização de Introdução no Mercado* (AIM) e os realmente comercializados, o que quer dizer que qualquer medicamento existente na base de dados pode ser prescrito mesmo não sendo comercializado. Em relação aos medicamentos que não se encontrem registados na base de dados, como os dietéticos, podem ser prescritos em texto livre, seleccionando a opção outras prescrições [Tomé *et al*. 2008]. Em Fevereiro de 2008, a utilização da aplicação SAM já era possível em todos os centros de saúde do Alentejo e Norte, e também em todos os centros de saúde das sub-regiões de Aveiro, Castelo-Branco, Guarda e Santarém [Carvalho 2008]. A informatização dos registos de prescrição, possibilita a existência de estudos aprofundados sobre a análise de tendências e evolução da prescrição a nível nacional. A aplicação de técnicas de mineração de dados, permite descobrir diversas associações entre a prescrição de medicamentos e os outros factores existentes numa prescrição.

É na descoberta do conhecimento para além da prescrição electrónica, através da aplicação de técnicas de mineração de dados, que se concentra este trabalho. Os principais objectivos serão descobrir possíveis associações entre prescrições de medicamentos, isto é, descobrir se a prescrição de determinados medicamentos está implicitamente associado à prescrição de outros determinados medicamentos. Encontrar padrões de prescrições relacionadas com o sexo, idade, localidade do utente, marcas temporais (por exemplo, fins de semana, feriados.), laboratórios, entre outros. Descobrir, se, realmente, a prescrição por parte dos médicos é feita tendo alguma associação com os laboratórios dos medicamentos prescritos, uma suspeita existente nas direcções de saúde existentes a nível nacional. Existe ainda a suspeita de, por vezes, algumas prescrições de medicamentos, de instituições de tratamento de saúde, serem feitas em nome dos utentes que apresentem maiores valores de comparticipação. Isso resulta num elevado número de medicamentos prescritos associados a esses utentes em causa e numa ficha clínica não tão viável. O objectivo deste estudo concentra-se também nesta questão, na afirmação de tal realidade ou não.

Existe um inúmero de possíveis estudos e análises a dados referentes à prescrição de medicamentos, no entanto, em Portugal, a realização desse tipo de estudos é bastante reduzida [Tomé *et al*. 2008, Carvalho 2008]. Tendo em conta este défice, este trabalho permitirá a obtenção de novos resultados acerca da prescrição electrónica, ajudando os serviços de saúde na sua organização e tomada de decisões. Conciliando técnicas de mineração de dados, como a associação, a classificação e a segmentação, com dados fornecidos pela *Administração Regional de Saúde do Norte* (ARSN) sobre a prescrição de medicamentos, é possível atingir o objectivo final e principal deste trabalho, a descoberta de padrões em sistemas de dados.

3.2 O Modelo de Dados

Todos os dados utilizados neste trabalho foram disponibilizados pela ARSN e foram obtidos a partir da prescrição de medicamentos na região Norte. Os dados fornecidos são referentes ao ano de 2008, desde o dia 1 de Janeiro de 2008 até ao dia 31 de Dezembro de 2008, num total de 21301073 registos. No entanto, o estudo foi limitado a um conjunto menor de dados, tendo em conta os centros de saúde e extensões de algumas localidades, referenciadas como possiveís representativas de toda a população, no que diz respeito a características de desenvolvimento do concelho em que estão inseridas, por exemplo, habilitações e nível socioeconómico. De seguida estão apresentados na Tabela 9, todos os centros de saúde seleccionados, associados ao respectivo distrito e tipologia urbano/rural, para este estudo.

Distrito	Centros de Saúde	Tipologia urbano/rural
Braga	Fafe, Vila Verde e Terras de Bouro	4, 4 e 5
Bragança	Macedo de Cavaleiros e Vinhais	3 e 5
Aveiro	Vale Cambra	4
Porto	Paranhos/Vale Formoso, Castelo da Maia, Barão do Corvo e Lousada	1, 2, 2 e 4
Viana do Castelo	Darque e Paredes de Coura	2 e 5
Vila Real	Chaves 1 e Stª Marta de Penaguião	5 e 2

Tabela 9 – Locais de atendimento relativos ao estudo em causa.

Os registos relativos aos centros de saúde e extensões, mencionados na Tabela 9 perfazem um total de 2261955 registos, em que cada registo é referente a cada medicamento prescrito por um médico, associado a uma receita, referente a um determinado utente, e a um respectivo local de prescrição. A cada medicamento, pode ser associado as seguintes características:

- código e descrição do medicamento;
- código e descrição da embalagem referente ao respectivo medicamento;
- código e descrição da via de administração do medicamento e embalagem;
- código e descrição do laboratório responsável pelo medicamento;
- código e descrição dos princípios activos do medicamento;
- código e descrição da catalogação de genérico ou não genérico;
- código e descrição do grupo homogéneo a que pertence o medicamento;
- código do grupo farmacoterapêutico;
- preço de venda ao público.

Cada prescrição é correspondente a um utente. Para cada utente é possível descobrir as seguintes características:

- código do utente (cifrado);
- sexo do utente;
- idade do utente;
- código de freguesia de habitação (cifrado).

Por fim, para cada medicamento prescrito, é possível retirar a seguinte informação:

- código da receita;

- código do respectivo medicamento e utente, que através desses campos é possível retirar informação respectivo a cada um;
- código da localidade onde foi elaborada a consulta (cifrada);
- data, hora e minuto da prescrição;
- código do médico prescritor (cifrado) e código de especialidade respectivo;
- informação se o médico em causa é o médico de família do utente em causa,
- tipo de consulta ocorrida;
- código do programa de saúde;
- número de episódio (SAM);
- ICPC2s envolvidos no episódio – Diagnósticos até 10;
- comparticipação do medicamento referente ao utente;
- quantidade de medicamento por registo (1).

Os dados foram inicialmente integrados numa única tabela e, a partir daí, foi realizado um reconhecimento técnico dos dados. As Tabela 10, Tabela 11, Tabela 12 e Tabela 13 apresentam os dados existentes e as respectivas características.

Campo	Descrição	Tipo de Dados	Domínio/Restrição	Nulo	Cardinalidade
cod_local	Código de localidade	Integer	Número de 1 a 7 dígitos (0-9)	Não	55
cod_dia	Data da prescrição	Integer	Número de 8 digitos (0-9)	Não	366
cod_hora	Hora da prescrição	Varchar(2)	String de tamanho 2	Sim	25
cod_minuto	Mês da prescrição	Varchar(2)	String de tamanho 2	Sim	60
cod_idade	Idade do utente	Integer	Número de 1 a 3 dígitos (0-9)	Não	106
cod_utente	Código do utente	Integer	Número de 1 a 9 digitos (0-9)	Não	213631
cod_sexo	Sexo do utente	Varchar(2)	String de tamanho 2	Não	2
cod_freg_habita	Freguesia de habitação do utente	Varchar(6)	String de tamanho 6	Sim	961
cod_med_fam	Médico de Família do Utente	Varchar(1)	String de tamanho 1	Não	3
cod_receita	Código da receita	Integer	Número de 7 digitos (0-9)	Não	926956
prog_saude	Programa de saúde	Integer	Número de 1 digito (0-6)	Não	7
tipo_cons	Tipo de consulta	Integer	Número de 1 a 3 digitos (0-9)	Não	39

Tabela 10 – Descrição do modelo de dados inicial.

Campo	Descrição	Tipo de Dados	Domínio/Restrição	Nulo	Cardinalidade
cod_espec	Código da especialidade da consulta	Integer	Número de 1 a 2 digitos (0-9)	Não	8
medico	Médico prescritor	Integer	Número de 1 a 6 digitos (0-9)	Não	291
cod_n1_medicamento	Código do medicamento	Integer	Número de 1 a 5 digitos (0-9)	Não	5317
des_n1_medicamento	Descrição do medicamento	Varchar(110)	String de tamanho 110	Não	4146
cod_n2_medicamento	Código da embalagem	Integer	Número de 1 a 5 digitos (0-9)	Não	8021
des_n2_medicamento	Descrição da embalagem	Varchar(110)	String de tamanho 110	Não	499
cod_n1_apresentacao	Código do método de administração do medicamento	Integer	Número de 1 a 5 digitos (0-9)	Não	17
des_n1_apresentacao	Descrição do método de administração do medicamento	Varchar(100)	String de tamanho 100	Sim	13
cod_n2_apresentacao	Código do método de administração da embalagem	Integer	Número de 1 a 5 digitos (0-9)	Não	128
des_n2_apresentacao	Código do método de administração da embalagem	Varchar(50)	String de tamanho 50	Sim	17
cod_laboratorio	Código do laboratório	Integer	Número de 1 a 4 digitos (0-9)	Não	319
des_laboratorio	Descrição do laboratório	Varchar(100)	String de tamanho 100	Não	316
cod_principios_activos	Código do principio activo	Integer	Número de 1 a 6 digitos (0-9)	Sim	1099
des_principios_activos	Descrição do principio activo	Varchar(250)	String de tamanho 250	Sim	1099
cod_generico	Código de indicação se o medicamento é genérico ou não	Integer	Número de 1 digito (0-9)	Sim	3
des_generico	Descrição de indicação se o medicamento é genérico ou não	Varchar(6)	String de tamanho 6	Sim	3

Tabela 11 – Descrição do modelo de dados inicial (continuação).

Campo	Descrição	Tipo de Dados	Domínio/Restrição	Nulo	Cardinalidade
cod_grupo_homogeneo	Código do grupo homogéneo	Varchar(6)	String de tamanho 6	Sim	489
des_grupo_homogeneo	Descrição do grupo homogéneo	Varchar(110)	String de tamanho 110	Sim	489
cod_n5_cft	Código do grupo farmacoterapêutico	Varchar(10)	String de tamanho 10	Sim	213
num_episodio	Número de episódio (SAM)	Integer	Número de 6 digitos (0-9)	Sim	198217
Ic1	Diagnóstico 1 do ICPC2s envolvidos no episódio	Varchar(6)	String de tamanho 6	Sim	978
Ic2	Diagnóstico 2 do ICPC2s envolvidos no episódio	Varchar(6)	String de tamanho 6	Sim	922
Ic3	Diagnóstico 3 do ICPC2s envolvidos no episódio	Varchar(6)	String de tamanho 6	Sim	866
Ic4	Diagnóstico 4 do ICPC2s envolvidos no episódio	Varchar(6)	String de tamanho 6	Sim	765
Ic5	Diagnóstico 5 do ICPC2s envolvidos no episódio	Varchar(6)	String de tamanho 6	Sim	695
Ic6	Diagnóstico 6 do ICPC2s envolvidos no episódio	Varchar(6)	String de tamanho 6	Sim	608
Ic7	Diagnóstico 7 do ICPC2s envolvidos no episódio	Varchar(6)	String de tamanho 6	Sim	521
Ic8	Diagnóstico 8 do ICPC2s envolvidos no episódio	Varchar(6)	String de tamanho 6	Sim	412
Ic9	Diagnóstico 9 do ICPC2s envolvidos no episódio	Varchar(6)	String de tamanho 6	Sim	327
Ic10	Diagnóstico 10 do ICPC2s envolvidos no episódio	Varchar(6)	String de tamanho 6	Sim	245
preço_pvp	Preço de venda ao público	Decimal(5,2)	Número de 5 digitos com 2 casas decimais (0-9)	Sim	3404

Tabela 12 – Descrição do modelo de dados inicial (continuação).

Campo	Descrição	Tipo de Dados	Domínio/Restrição	Nulo	Cardinalidade
comparticipacao	Comparticipação da respectiva prescrição relativa ao medicamento e ao utente	Decimal(5,2)	Número de 5 digitos com 2 casas decimais (0-9)	Sim	3482
qtd_medicam_presc	Quantidade de medicamentos prescritos por registo (1)	Integer	Número de 1 digito (0-9)	Sim	1

Tabela 13 – Descrição do modelo de dados inicial (continuação).

3.3 Análise Exploratória dos Dados

Nos centros de saúde em questão (Tabela 9) foram prescritas 2261955 embalagens de medicamentos, no período entre 1 de Janeiro de 2008 e 31 de Dezembro de 2008, para um total de 303419 utentes. A taxa de prescrição para esses distritos foi aproximadamente de 10588 prescrições por 1000 utentes. É perceptível pela Figura 12 que o Porto foi o distrito que obteve o maior número de prescrições associadas e que Bragança, por sua vez, apresentou o menor número de prescrições.

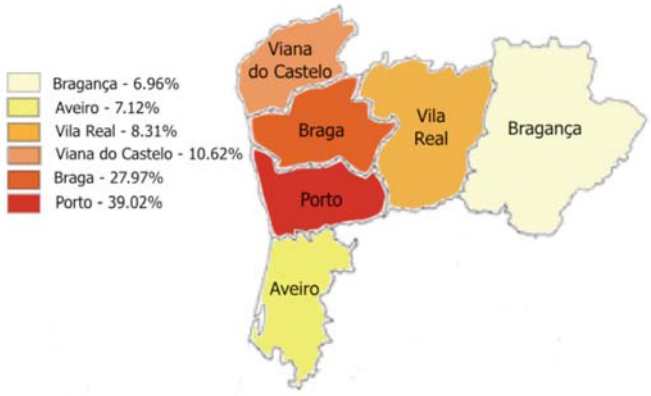


Figura 12 – Percentagem da prescrição por distritos

Relativamente aos centros de saúde, o de Fafe, no distrito de Braga, tendo em consideração as suas extensões e unidades de saúde familiares (USF), foi o que apresentou uma maior percentagem de prescrições (14.29%), enquanto que o centro de saúde de Stª Marta de Penaguião, em Vila Real, foi o que apresentou uma menor percentagem de prescrições (2.47%) (Figura 13). Por sua vez, a unidade de saúde familiar IRIS, associada ao centro de

saúde Castelo da Maia, no distrito do Porto foi a que apresentou a percentagem mínima de prescrições por 1000 utentes, enquanto, que a extensão de Agrochão do centro de saúde de Vinhais, no distrito de Bragança foi a unidade de saúde que apresentou a maior taxa de prescrição de medicamentos (Tabela 14).

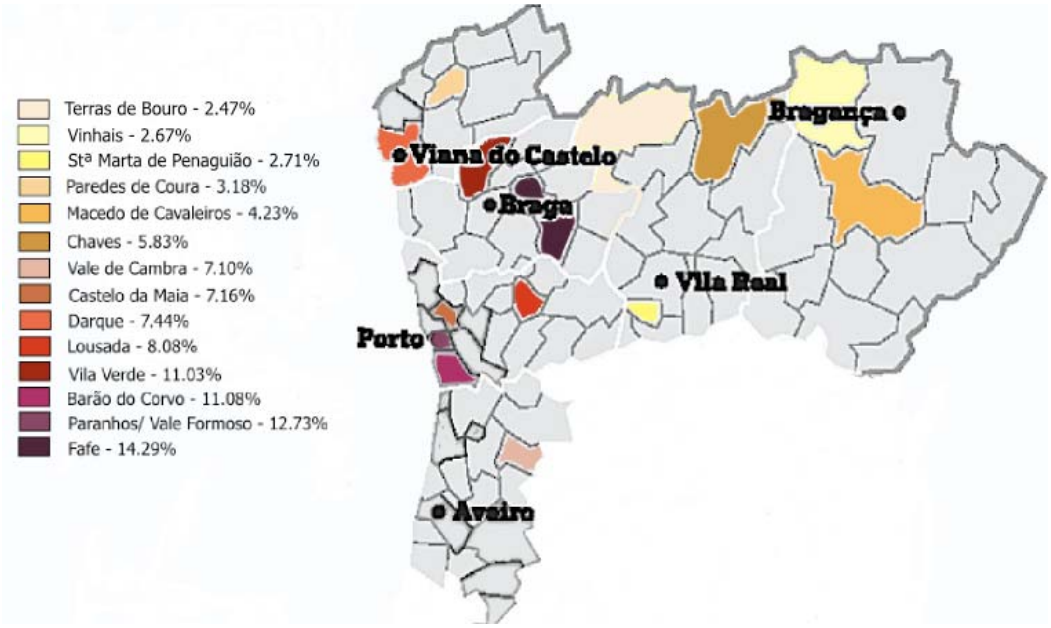


Figura 13 – Percentagem de prescrições dos centros de saúde estudados.

Taxa Mínima			
Distrito	**Centros de Saúde**	**Extensões/USF**	**Taxa de Prescrição**
Porto	Castelo da Maia	USF IRIS	2722 / 1000 utentes
Braga	Vila Verde	Ext. Cervães	6148 / 1000 utentes
Braga	Fafe	Fafe (sede)	7419 / 1000 utentes
Porto	Lousada	Ext. Meinedo	7838 / 1000 utentes
Braga	Vila Verde	Ext. Prado	8505 / 1000 utentes
Taxa Máxima			
Bragança	Vinhais	Ext. Agrochão	26929 / 1000 utentes
Bragança	Vinhais	Ext. Ervedosa	14064 / 1000 utentes
Aveiro	Vale Cambra	Ext. Junqueira	13776 / 1000 utentes
Aveiro	Vale Cambra	Ext. Macieira de Cambra	13570 / 1000 utentes
Bragança	Vinhais	Ext. Rebordelo	13229 / 1000 utentes

Tabela 14 – TOP 5 das unidades de saúde com taxa minima e maxima de prescrições por 1000 utentes.

Apesar de Aveiro ser dos distritos com o menor número de prescrições (Figura 12), é o distrito com a maior taxa de prescrições por 1000 utentes. Isto significa, que existe um grande número de prescrições comparativamento com o número de utentes existente nesse distrito. Isso é constatável através da Figura 14.

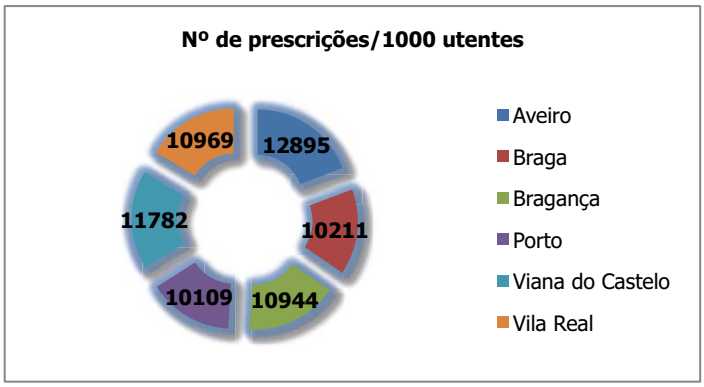
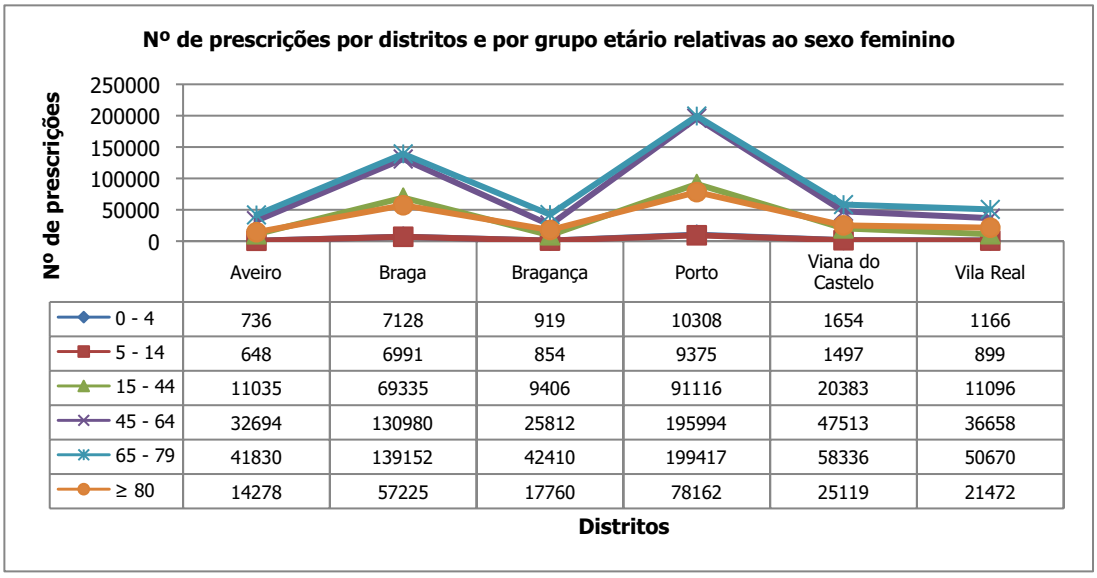


Figura 14 – Nº de prescrições de distritos por 1000 utentes.

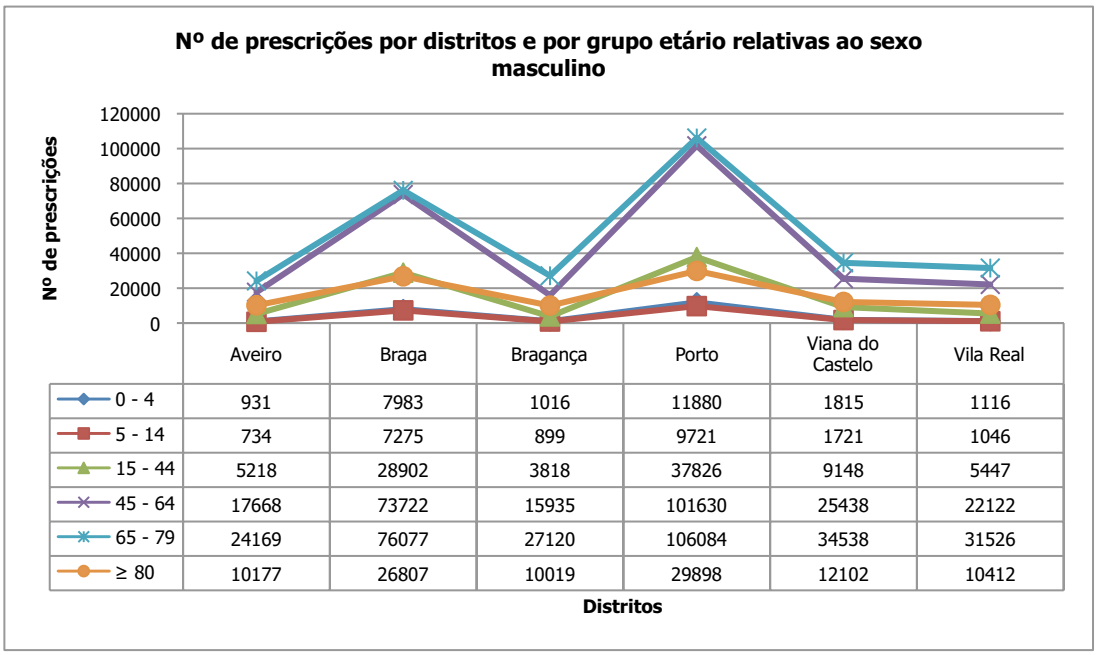
A Figura 15 e Figura 16 são referentes ao número de prescrições, tendo em conta o distrito onde foi realizada a consulta e o grupo etário do utente em causa. A única diferença de análise, remete para o facto de a Figura 15 ser relativa aos utentes do sexo feminino, e a Figura 16 aos utentes do sexo masculino. Pela comparação entre as duas figuras, é perceptível o maior número de prescrições por parte do sexo feminino, e em ambos os casos, os distritos que apresentam maior prescrições são o Porto e Braga. Para os restantes distritos os valores apresentam alguma uniformidade, não sendo notável tanta discrepância como nesses dois distritos.



	Aveiro	Braga	Bragança	Porto	Viana do Castelo	Vila Real
0 - 4	736	7128	919	10308	1654	1166
5 - 14	648	6991	854	9375	1497	899
15 - 44	11035	69335	9406	91116	20383	11096
45 - 64	32694	130980	25812	195994	47513	36658
65 - 79	41830	139152	42410	199417	58336	50670
≥ 80	14278	57225	17760	78162	25119	21472

Figura 15 – Nº de prescrições por distritos e por grupo etário relativamente ao sexo feminino.

Mais concretamente, relativamente aos grupos etários, os grupos dos 45 aos 64 e dos 65 aos 79, são os que apresentam uma maior quantidade de prescrições, e os grupos dos 0 aos 4 e dos 5 aos 14 apresentam valores bastantes semelhantes, e só nestes casos o número de prescrições do sexo masculino é superior ao número de prescrições do sexo feminino em ambas as figuras.



	Aveiro	Braga	Bragança	Porto	Viana do Castelo	Vila Real
0 - 4	931	7983	1016	11880	1815	1116
5 - 14	734	7275	899	9721	1721	1046
15 - 44	5218	28902	3818	37826	9148	5447
45 - 64	17668	73722	15935	101630	25438	22122
65 - 79	24169	76077	27120	106084	34538	31526
≥ 80	10177	26807	10019	29898	12102	10412

Figura 16 – Nº de prescrições por distritos e por grupo etário relativamente ao sexo masculino.

Após uma análise relativa à prescrição demográfica de medicamentos, é apresentado de seguida (Tabela 15) uma outra análise tendo agora em conta outros parâmetros que poderão influenciar bastante a prescrição de medicamentos, tais como a faixa etária, sexo do utente e os meses do ano. A prescrição de medicamentos foi mais frequente no sexo feminino (7856 prescrições/1000 utentes), do que no sexo masculino (6810 prescrições/1000 utentes). Na Figura 17 (gráfico da esquerda) é possível constatar a distribuição percentual segundo o sexo do utente. Em termos percentuais, os grupos etários dos 0 aos 4 anos e dos 5 aos 14 anos, foram os que apresentaram um menor número de prescrições, apenas 2%, enquanto que o grupo etário dos 65 aos 79 apresentou a maior taxa percentual de prescrições, 37% (Figura 17 (gráfico da direita)). No entanto, pela Tabela 15, é perceptível, que apesar de o grupo etário dos 65 aos 79 anos apresentar um maior número de prescrições, o grupo etário de idade superior a 80 anos, é o que apresenta a taxa mais elevada de prescrições por 1000 utentes (Figura 19). O que significa que a classe dos utentes, de idade igual ou superior a 80 anos, recorrem com mais frequência aos serviços de saúde disponíveis, o que normalmente

se constata na vida real, devido ao aparecimento de um maior número de problemas de saúde com o avanço da idade.

	Informação	Nº de prescrições	Nº de utentes	Prescrição / 1000 utentes
Sexo	Feminino	1470015	129766	11328 /1000 utentes
	Masculino	791940	83868	9443 / 1000 utentes
Grupo Etário	0 - 4	46652	11146	4186 / 1000 utentes
	5 - 14	41644	14186	2936 / 1000 utentes
	15 - 44	302730	64870	4667 / 1000 utentes
	45 - 64	726166	66554	10911/ 1000 utentes
	65 - 79	831329	45862	18198 / 1000 utentes
	≥ 80	313431	15822	19810 / 1000 utentes
Meses	Janeiro	216085	213631	1011 / 1000 utentes
	Fevereiro	185616	213631	869 / 1000 utentes
	Março	173528	213631	812 / 1000 utentes
	Abril	198338	213631	928 / 1000 utentes
	Maio	180285	213631	844 / 1000 utentes
	Junho	177397	213631	830 / 1000 utentes
	Julho	184997	213631	866 / 1000 utentes
	Agosto	141834	213631	664 / 1000 utentes
	Setembro	223250	213631	1045 / 1000 utentes
	Outubro	226472	213631	1060 / 1000 utentes
	Novembro	185063	213631	866 / 1000 utentes
	Dezembro	169090	213631	791 / 1000 utentes

Tabela 15 – Estatísticas descritivas do número de prescrições tendo em conta diversos factores.

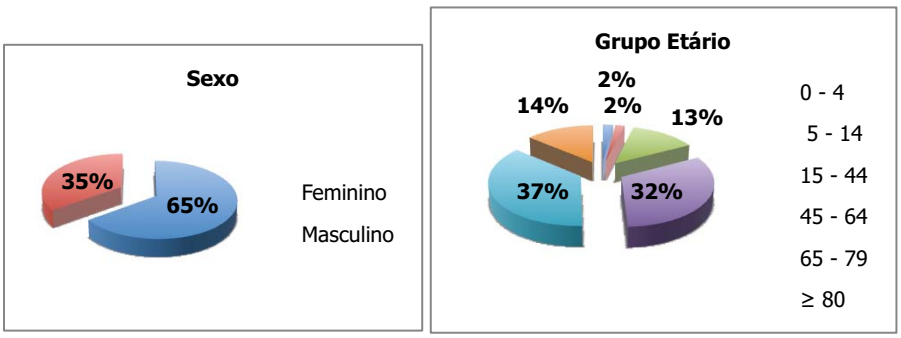


Figura 17 – Percentagem do número de prescrições por 1000 utente, tendo em conta os eixos de análise do sexo e faixa etária do utente.

Na Figura 18 está representado o gráfico com os valores percentuais das prescrições ocorridas em cada mês. Como é possível observar pela mesma figura e também através da

Tabela 15, os meses que apresentam uma maior número de prescrições, são os meses de Janeiro, Setembro e Outubro. O que é algo expectável, muito em parte devido à afluência das pessoas aos centros de saúde, com o intuito de se preservarem contra a gripe que normalmente ataca nos meses de Setembro e Outubro. No entanto, em Dezembro existe um decréscimo do número de prescrições, provavelmente devido a ser um período festivo e de férias para algumas pessoas, o que diminuiu a probabilidade de contágio de doenças e ser aproveitado como um tempo de descanso e relaxamento.

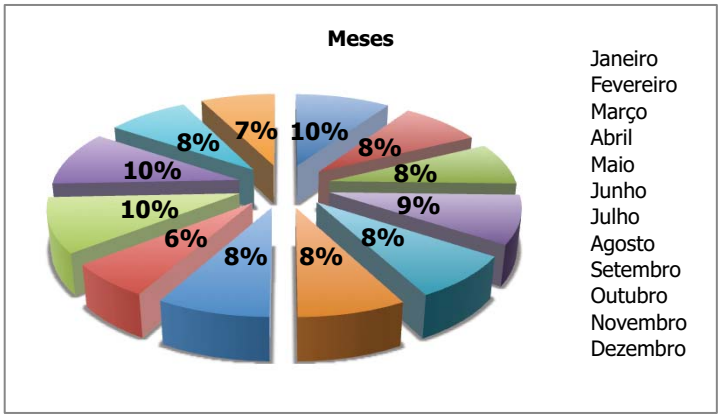


Figura 18 – Percentagem do número de prescrições dos meses do ano.

Pelo gráfico da Figura 19 é possível constatar-se a predominância do sexo feminino relativamente ao número de prescrições por 1000 utentes, com excepção nos dois primeiros grupos etários. Como foi mencionado anteriormente, apesar de o maior número de prescrições se suceder nos grupos etários dos 45 aos 64 e dos 65 aos 79, o grupo etário de idade superior ou igual a 80 anos, é o que apresenta uma maior taxa de prescrição de medicamentos por 1000 utentes. O grupo etário que apresenta uma menor taxa de prescrições por 1000 utentes é o dos 5 aos 14 anos.

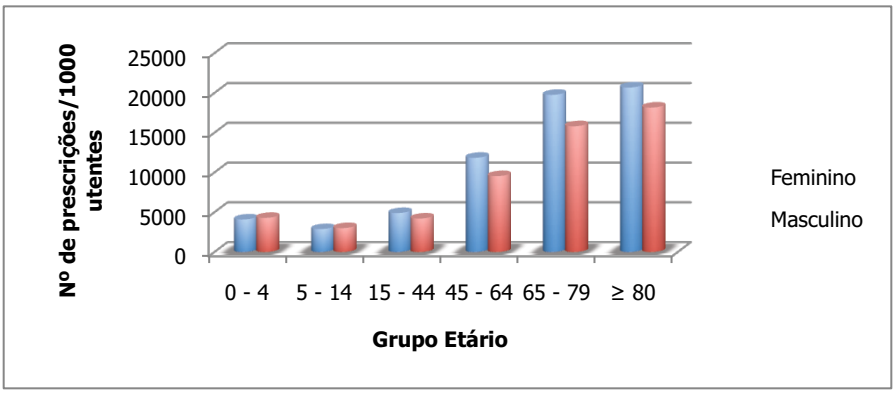


Figura 19 – Número de prescrições de medicamentos por 1000 utentes, segundo o grupo etário.

De seguida será realizada uma análise mais direccionada aos medicamentos prescritos, bem como aos laboratórios existentes, de maneira, a se ter uma noção quais são as classes predominantes para cada um dos casos. Na Tabela 16, são apresentadas algumas estatísticas discritivas tendo em consideração a prescrição de genéricos, os medicamentos mais prescritos e os laboratórios predominantes nas prescrições.

	Informação	Nº de prescrições	Nº de utentes	Prescrição / 1000 utentes
Genérico	Sim	388278	213631	1818 presc /1000 utentes
	Não	1873620	213631	8771 presc / 1000 utentes
Top 6 Medicamentos	Ben-U-Ron	43922	213631	206 presc / 1000 utentes
	Tromalyt	27237	213631	128 presc / 1000 utentes
	Risidon	25145	213631	118 presc / 1000 utentes
	Diamicron	22433	213631	105 presc / 1000 utentes
	Lorenin	21998	213631	103 presc / 1000 utentes
	Lasix	21402	213631	101 presc / 1000 utentes
	Outros	2099818	213631	9829 presc / 1000 utentes
Top 6 Laboratórios	Sanofi Aventis	121032	213631	567 presc / 1000 utentes
	Lab. Pfizer	88036	213631	412 presc / 1000 utentes
	Servier	86335	213631	404 presc / 1000 utentes
	Neo-F armacêutica	84410	213631	395 presc / 1000 utentes
	Merck	75602	213631	354 presc / 1000 utentes
	Novartis Farma	67827	213631	317 presc / 1000 utentes
	Outros	1738713	213631	8139 presc / 1000 utentes

Tabela 16 – Estatísticas descritivas tendo em conta o estudo dos medicamentos e laboratórios.

Na Figura 20, é mostrado o gráfico referente à percentagem de prescrições de medicamentos genéricos e não genéricos. Em conclusão, com a Tabela 16, é facilmente destacável que a prescrição de genéricos é algo ainda não muito usual, apesar de todos os esforços que têm havido na propaganda de tal.

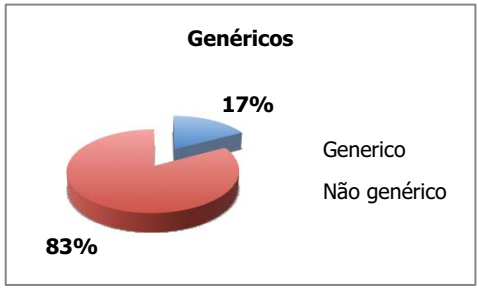


Figura 20 – Percentagem de prescrições de medicamentos genéricos e não genéricos.

Na Figura 21 está apresentada uma análise à percentagem de prescrições de genéricos tendo em consideração os grupos etários existentes. Nos grupos etários superiore (45 - 64, 65 - 79 e ≥ 80) é notório um aumento da percentagem de prescrições de genéricos comparativamente com os grupos etários inferiores (0 - 4, 5 - 14 e 15 - 64). Isso poderá dever-se, de alguma maneira, ainda ao cepticismo das pessoas em relação aos genéricos. Isto é, apesar de serem mais baratos, poderá haver ainda alguma desconfiança na sua segurança e eficácia, o que leva as pessoas a apostarem nos medicamentos não genéricos quando se trata de crianças, mesmo estes tendo um custo mais elevado.

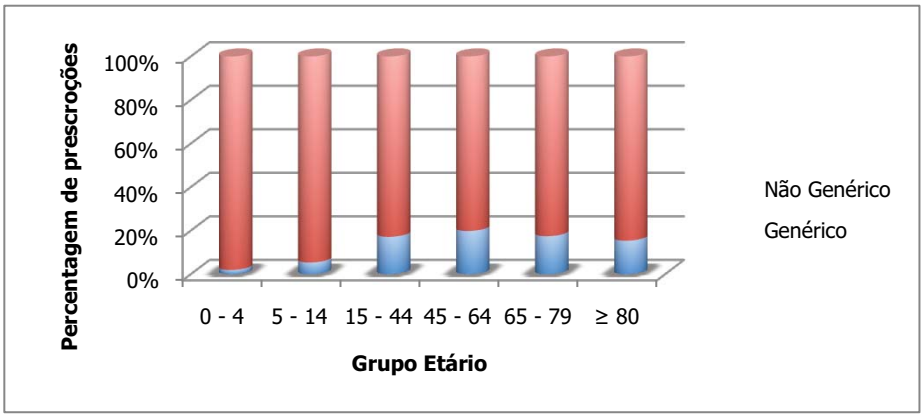


Figura 21 – Percentagem de prescrições de medicamentos genéricos por grupos etários.

Pelo gráfico apresentado na Figura 22 é possível observar que a prescrição de medicamentos genéricos é bastante inferior à prescrição de medicamentos não genéricos, em todos os distritos estudados, não havendo nenhum distrito em que a prescrição de genéricos seja já algo notável. Praticamente em todos os distritos, as prescrições de medicamentos genéricos e não genéricos, por 1000 utentes, são idênticas. No entanto, denota-se um ligeiro aumento de prescrições de medicamentos genéricos, por 1000 utentes, em Aveiro, Viana do Castelo e Vila Real comparativamente com os outros distritos, apesar de Aveiro, também ser o distrito com mais prescrições de medicamentos não genéricos por 1000 utentes.

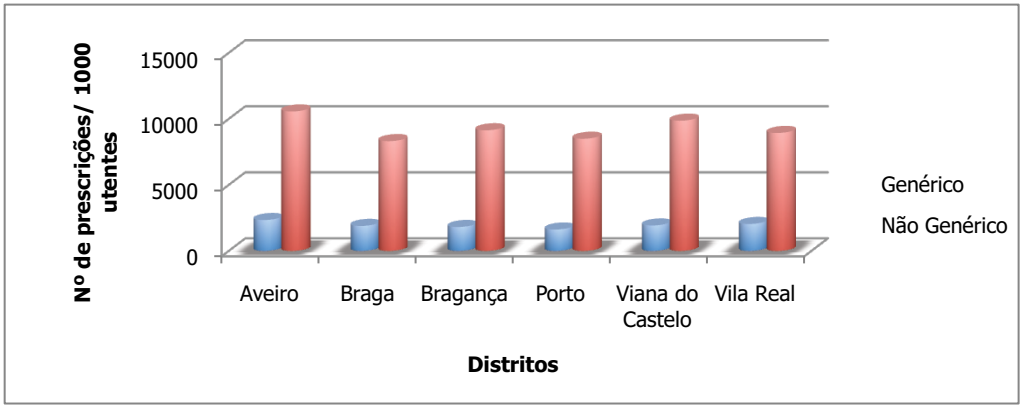


Figura 22 - Nº de prescrições de medicamentos genericos e não genericos por 1000 utentes, segundo os distritos estudados.

Na Tabela 17 são apresentadas algumas estatísticas relativas ao top 10 de médicos mais prescritores. À informação do número de prescrições dos médicos, está associado o número dos diferentes laboratórios prescritos, complementando-se, na última coluna, com a informação de quantos laboratórios são usados em cada 1000 prescrições feitas pelos médicos. Após a observação dos dados, pode-se concluir que, em principio, não existe nenhuma associação vísivel entre a prescrição dos médicos e os laboratórios. Isto é, apesar de existir uma predominância de alguns laboratórios nas prescrições gerais, não é notável que os médicos estejam ligados exclusivamente a determinados laboratórios, o que remeteria para possíveis acordos entre laboratórios e médicos. Este assunto será abordado com mais atenção, na aplicação de técnicas de mineração de dados, com o intuito de descobrir se realmente não existe nenhuma associação vísivel entre os médicos e as suas prescrições, ou se por outro lado, poderá existir alguma associação.

	Informação	**Nº de prescrições**	**Nº de laboratórios**	**Laboratórios / 1000 prescrições**
Top 10 Médicos	777470	20474	182	8.89 labs / 1000 prescrições
	773470	17646	214	12.12 labs / 1000 prescrições
	140121	16620	203	12.21 labs / 1000 prescrições
	797421	16586	199	12.00 labs / 1000 prescrições
	218970	16169	201	12.43 labs / 1000 prescrições
	683028	15859	204	12.86 labs / 1000 prescrições
	327279	15833	184	11.62 labs / 1000 prescrições
	500377	15315	192	12.54 labs / 1000 prescrições
	713470	14744	199	13.50 labs / 1000 prescrições
	303270	14664	192	13.09 labs / 1000 prescrições

Tabela 17 – Estatísticas descritivas do top 10 dos médicos, tendo em conta o diferente número de laboratórios prescritos.

A comparticipação de cada utente difere consoante as suas condições de saúde e profissionais. Consequentemente, dentro de um agregado familiar poderá existir uma pessoa com uma comparticipação mais elevada do que o resto da família, o que por vezes remete para um "abuso" de utilização dessa pessoa na prescrição de medicamentos. Mais concretamente, por vezes, os medicamentos prescritos a um agregado familiar, são pedidos no nome do familiar que apresenta uma maior comparticipação. Este problema remete para uma ficha clínica não muito fiável, pois associado à mesma pessoa poderão estar um sem número de medicamentos de princípios activos diferentes, e por vezes, esses medicamentos poderão ser completamente incompatíveis. Na Tabela 18 estão apresentadas algumas características dos 10 utentes com mais prescrições associadas, de maneira, a se poder descobrir se realmente existem ou não casos de prescrições associadas a um único familiar. Esta tabela é complementada com a idade e sexo do utente, e também com as questões económicas, como o preço total que pagou, associado à comparticipação que teve no total. Normalmente, pessoas que apresentam uma elevada percentagem de comparticipação, poderão ser pessoas com doenças crónicas, nesse caso, a existência de diversos princípios activos associados a uma única pessoa, pode ser considerado normal.

Utente	Idade	Sexo	Nº presc.	Preço de Venda	Comp.	% comp.	Princípios Activos
799710208	74	F	211	4387.73	3182.16	73%	60
736825200	83	M	204	3051.88	1102.12	36%	28
44967785	69	M	194	2845.36	1749.93	62%	34
526825203	34	F	186	853.79	425.95	50%	17
140316210	72	F	180	3206.70	2471.40	77%	74
851917244	57	F	164	2316.40	1338.11	58%	27
998056243	78	F	163	1181.55	765.97	65%	32
534689281	75	F	162	3354.61	2515.14	75%	32
446815219	66	F	158	3599.07	2810.46	78%	103
732642781	79	F	152	2176.27	1207.64	55%	18

Presc. = prescrições e Comp.= comparticipação

Tabela 18 – Análise da comparticipação dos 10 utentes com maior número dez prescrições.

Mas para uma melhor análise destes 10 utentes, é necessário descobrir quais os princípios activos associados a cada um dos utentes. A informação relativa ao número de princípios activos associados a cada utente encontra-se na última coluna da Tabela 18. No entanto, para uma análise mais pormenorizada (lista dos princípios activos de cada utente), é necessário recorrer à Tabela 28 (Anexos I) para que seja possível tirar algumas conclusões acerca deste possível problema.

As análises estatísticas realizadas, já demonstram algumas características importantes dos dados e permitem tirar algumas conclusões acerca deles. No entanto, o estudo dos dados permite também aprofundar o conhecimentos da qualidade existente nos dados, e consequentemente afirmar quais os campos que possivelmente poderão ser desnecessários no estudo ou que se encontram inutilizáveis.

3.4 Análise Crítica da Qualidade dos Dados

Numa análise geral, a qualidade dos dados em causa é considerada muito boa, no entanto existem alguns problemas em alguns dos seus atributos. O problema mais notório encontra-se nos dados dos ICPC2s, em que em todos eles existe uma predominância de existência de valores nulos. A possível utilidade destes campos, permitiria análises bastantes interessantes, como por exemplo, consoante os sintomas do utente, quais seriam os medicamentos prescritos. No entanto, devido à imensa falta de registos, relativos a estes campos, foi completamente impossível fazer qualquer tipo de análise deste género. Esta situação acontece, devido à não inserção dos sintomas do utente no sistema, aquando a realização da consulta. Também foi notória a existência de alguns registos em que o campo hora se encontrava a *Null*, mas mesmo assim foi possível a utilização deste campo no estudo. No entanto, a sua utilização remeteu para a necessidade de tomada de decisões, isto é, numa fase seguinte, foi criado um novo campo denominado de período, que indicava se a consulta foi dada num período diurno ou nocturno. Como o povoamento deste novo campo provinha da informação contida no campo hora, foi necessário definir estratégias alternativas, para quando o campo se encontrava a nulo. Consequentemente, ao campo período, apenas nesta situação, foi atríbuido o período predominante no restantes registos (período diurno).

Relativamente aos laboratórios, havia um laboratório em específico que no ficheiro original de dados fornecido, após a designação desse laboratório encontrava-se um '\n' em vez de um '\t'. Isso significa, que os campos seguintes nesses registos ficarão a nulos, pois os campos após o '\n', são considerados como um novo registo, apesar de na realidade não o serem. Por esse motivo, é que nas Tabela 10, Tabela 11, Tabela 12 e Tabela 13, os campos seguintes ao atributo *des_laboratorio* apresentam todos valores nulos, caso contrário, apenas alguns apresentariam (por exemplo os ICPC2s). A solução para este problema foi retirar todos esses registos, uma vez que correspondiam a um laboratório, pouco utilizado e sem significado relevante, estatisticamente.

Por fim, foram encontrados 3 registos de uma possível inserção incorrecta, pois associado a esses registos encontra-se um utente cuja idade aparece como sendo -31 anos. Como, obviamente, este é um caso completamente impossível, a inclusão destes registos no estudo em causa, não forneceria qualquer informação válida, logo, estes registos foram retirados do estudo.

Todos estes casos são aqueles que foram analisados com mais atenção relativamente à qualidade, no entanto, existem ainda algumas situações que poderiam ser revistas. Por exemplo, existem medicamentos cujo o nome são iguais, que apesar de serem realmente diferentes, foram inseridos com o mesmo nome. Isto quer dizer, que numa pesquisa de medicamentos, para um dado número de códigos obtem-se um menor número de descrições. Isto pode levar, desnecessariamente, a uma análise confusa e compreensão incorrecta num primeiro contacto com os dados. Este problema também foi encontrado relativamente aos laboratórios, existem laboratórios com designações iguais e códigos diferentes. Mas, mais uma vez, são na realidade laboratórios diferentes (focados para diferentes tipos de medicamentos), apesar de a entidade superior ser a mesma. Relativamente a estes casos, poderia ser dada mais alguma atenção na modifcação das designações, pois iria permitir uma análise mais intuitiva e sem possíveis enganos. Outra possível modificação, seria associar não a idade do utente nas prescrições, mas sim a sua data de nascimento. Isto quer dizer, que quando se está a fazer o estudo das prescrições, por vezes aparece o mesmo utente mas associado a idades diferentes. Seria bem mais fácil e intuitivo, se associado aos utentes estivesse a sua data de nascimento e não a sua idade. Houve duas situações que geraram alguma confusão na criação de um sistema de *data warehousing*, numa fase mais adiante, ambas pelo mesmo motivo. Associado ao mesmo cliente, poderá existir duas freguseias de habitação, caso o utente em causa tivesse mudado de casa. O mesmo é vísivel relativamente ao preços dos medicamentos. O formato dos dados fornecido, criou alguns problemas neste caso. Optou-se então, na criação do sistema de *data warehousing*, pela criação de chaves de substituição, o que significa que o mesmo cliente pode estar associado a duas chaves de substituições diferentes, em que diferem apenas na localidade ou a idade. No entanto, para pesquisas relativas aos clientes, foi usado os seus códigos de raíz, pois são esses realmente os clientes existentes. O mesmo aconteceu com os medicamentos, relativamente aos diferentes preços.

Numa análise posterior, relativamente à aplicação da associação ao modelo de dados final, foi encontrado uma situação no mínimo estranha. Ao mesmo código de receita estão associados dois ou mais pacientes e/ou dois ou mais médicos. O normal seria uma relação de 1 para um entre os códigos de receita e as entidades médico e paciente, o que em alguns registos isso não acontece. Este problema nos dados deve-se, provavelmete, a uma falha no

sistema, podendo ter sido causada por erro humano ou não. Para este caso, optou-se por encarar esta situação como "normal", não se procedendo a nenhum tratamento.

3.5 Selecção e Limpeza dos Dados

Após a análise estatística dos dados, do conhecimento da qualidade dos mesmos e tendo em conta o objectivo da mineração de dados, foram seleccionados apenas alguns campos dos existentes no modelo de dados inicial. Relativamente a uma análise temporal foi retirado o campo minuto, pois seria um campo de pouca utilidade, pois o minuto não é propriamente um eixo de análise interessante para o estudo em causa. Neste estudo, o eixo de análise temporal que terá mais importância, será provavelmente o mês, uma vez que já irá permitir obter bastante informação relativamete aos padrões de prescrição, consoante a variação dos meses (por exemplo, em Setembro existe uma grande procura da vacina contra a gripe). Devido à elevada quantidade de valores nulos e consequente impossibilidade de retirar qualquer tipo de informação relativamente aos campos do ICPC2s, optou-se por retirar esses campos, consequentemente, ao todo, 10 campos foram retirados tendo em conta os diagnósticos de 1 até 10. O campo *num_episodio*, está associado ao número de episódio do sistema de apoio ao médico (SAM). Como através da informação fornecida, por este campo, não é possível tirar qualquer conclusão relativamente aos padrões de prescrições, optou-se pela não inclusão deste campo no estudo. Concluindo, dos 45 campos iniciais, foram retirados um total de 12 campos, ficando-se com um total de 33 campos no final.

Como foi mencionado acima (secção 3.4), foi encontrado um problema num dos laboratórios, em que após a designação do laboratório encontrava-se um '\n' em vez de um '\t'. O laboratório em causa é o "BYK GULDEN LOMBERG" com o código 557. Todos os registos que apresentavam estas características foram retirados do conjunto de dados final utilizado, sendo necessário também retirar todos os registos, que apesar de pertencerem ao registo anterior, devido ao '\n' foram dados como um novo registo. Isto resultou numa remoção de um total de 114 registos. Como se constata, o facto de se retirar todos os registos relativos a este laboratório não apresenta nenhuma modificação notável no conjunto de dados final. Logo, esta decisão não traz alterações de relevância no resultado final da análise.

Por vezes, a prescrição de medicamentos engloba prescrições de material como ligaduras, bandas elásticas, fraldas, entre outros. Apesar de não serem propriamente medicamentos, a sua prescrição aparece em conjunto com o resto das prescrições de medicamentos. No entanto, a informação proveniente da análise deste tipo de registos não é considerada de grande importância para a ARSN, como tal, a inserção destes registos no estudo não

apresenta nenhuma sobre-valia ao estudo em geral. Consequentemente, foi realizada a remoção completa destes registos, cuja descrição dos princípios activos era igual a uma *string* vazia. No total, foram removidos 23280 registos, um número que é aproximadamente 0.1% do total dos registos descritos inicialmente.

Relativamente ao problema mencionada na secção 3.4, em que existia um utente cuja a sua idade estava indicada como sendo -31, a opção de tratamento dos dados, neste caso, recaiu na eliminação destes 3 registos, pois qualquer modifcação poderia não estar de acordo com a realidade. Logo, de maneira a não distorcer a realidade e de se obter uns resultados mais transparentes e correctos, a remoção destes 3 registos foi a decisão mais acertada.

Após a limpeza dos dados ter sido realizada, o número total de registos obtidos, e que servirão de estudo para este trabalho, foi de 2238615 registos. No entanto, para uma melhor compreensão do modelo de dados, para uma exploração dos dados mais intuitiva e lógica e para a possibilidade de um refinamento que ajudaria na descoberta de mais padrões e associações, foi criado um sistema de *data warehousing*, com todos esses registos.

3.6 Construção, Integração e Formatação dos Dados

Após a limpeza e selecção dos dados a utilizar foi implementado uma estrutura que melhor respondesse às necessidades impostas pela a utilização de técnicas de mineração de dados e às necessidades de variados eixos de análise. Consequentemente, o objectivo final foi criar um sistema de *data warehousing*, com novos eixos de análise. Entretanto, numa fase anterior ao da criação do sistema de *data warehousing*, foi criado uma área de retenção. Nessa área de retenção foi criada uma tabela idêntica à existente inicial, mas que, além de já não apresentar os campos removidos, foi povoada com os registos finais. A criação desta área de retenção foi feita com o intuito de preservar o modelo de dados inicial intacto. Só após este procedimento, foi implementada a estrutura adequada para a utilização final dos dados.

O processo de povoamento do *data warehouse*, é considerado como sendo um processo realizado em três fases. A primeira fase consiste em passar os dados do sistema operacional para a área de rentenção (processo de extracção dos dados) em que os dados podem ser sujeitos a diversas transformações, definições de chaves estrangeiras e chaves de substituições (processo de transformação dos dados). Após a transformação dos dados, é possível passar-se à passagem dos dados da área de rentenção para o *data warehouse* (processo de integração dos dados). O sistema de *data warehousing* implementado integra no seu *data mart* um esquema em estrela envolvendo 6 dimensões de análise. As dimensões

criadas correspondem à dimensão tempo, laboratório, médico, período, medicamento e utente. Para cada uma das dimensões será, de seguida, indicado os seus campos e o método do povoamento realizado.

- **DimTempo.** Esta dimensão serve de marca temporal relativamente à prescrição de medicamentos, sendo composta por campos que servem de marcadores temporais. É possível analisar as prescrições por data, ano, mês, dia, hora, dia da semana, fim de semana, feriado e estação. O povoamento desta dimensão foi realizado através de um cursor e de 4 funções auxiliares criadas para os campos, dia da semana, fim de semana, estação do ano e feriado. A função auxiliar *Dia_Semana()* recebe o dia da semana em inglês e converte-o para português, através desse resultado e da função *Fim_Semana()*, o campo fim de semana é preenchido. A função *Feriado()*, dado uma data devolve sim ou não, caso o dia em causa seja feriado ou não, por fim, a função *Estacao()* tem como *input* uma data e devolve a estação do ano em que essa data se insere. Para os restantes campos foram usados apenas funções já existentes na biblioteca do *MySQL*, aquando a realização do cursor. A chave primária desta dimensão é um número sequencial gerado pelo sistema.
- **DimPeriodo.** Apesar de estar numa dimensão à parte da dimensão tempo, esta dimensão é também um eixo de análise temporal, que permite analisar se uma prescrição foi realizada num período diurno (das 8h às 20h) ou nocturno (das 20h às 8h). Esta dimensão é composta por dois campos, sendo um o código do período (chave primária) e outro a descrição. A chave primária desta tabela é também um campo gerado sequencialmente, no entanto, a descrição do período é povoada manualmente com apenas dois registos, nocturno e diurno.
- **DimMedico.** A dimensão médico é composta apenas por um campo já existente no modelo de dados inicial. Apesar de existir apenas um campo associado à entidade médico, como é considerado um eixo de análise demasiadamente importante, foi considerado como sendo uma dimensão. O povoamento desta tabela foi automático, através do comando "*INSERT INTO ... SELECT DISTINCT*".
- **DimLaboratorio.** Como o próprio nome indica, esta dimensão possibilita a análise das prescrições tendo em conta os laboratórios existentes. É constituída por dois campos, o código e a descrição do laboratório. Neste caso, a chave primária é o próprio código do laboratório já existente no modelo de dados inicial, consequentemente, o povoamento desta tabela também foi feito automaticamente através do comando "*INSERT INTO ... SELECT DISTINCT*".
- **DimUtente.** Na dimensão utente é guardada todo o tipo de informação que está, directamente, relacionada com o utente. Além dos campos referentes aos utentes, do modelo de dados inicial, foi acrescentado mais dois campos. O campo *cod_utente_id*

que serve como chave primária da tabela, devido à possível existência de utentes associados a mais do que uma idade ou a localidade (o que remetia para uma violação da chave primária, caso essa chave fosse o código do utente original). Este campo é um número sequencial, gerado automaticamente. O outro campo gerado é o campo relativo ao grupo etário em que o utente se encontra, isto é, foram criados 6 grupos etárias diferentes, que coincidem com os grupos etários utilizados nas análises realizadas na secção 3.3. Para este campo foi criada outra função auxiliar, *Grupo_Etario()*, que dado uma idade devolve o respectivo grupo etário a que o respectivo utente se encontra. Este campo foi povoado, bem como todos os outros campos existentes nesta dimensão, através da execução de um cursor implementado, que por sua vez, para cada registo chama a função *Grupo_Etario()* para a inserção do valor correcto no campo correspondente.

- **DimMedicamento.** Esta dimensão é responsável por toda a informação directamente relacionada com os medicamentos. Todos os campos existentes nesta tabela são provenientes do modelo de dados inicial, exceptuando a chave primária. Neste caso foi necessário a criação deste campo, pelo mesmo motivo que na dimensão utente, pois o código de um medicamento pode estar associado a diversos códigos de embalagens ou mesmo a diferentes preços. Este campo, é também um número sequencial gerado automaticamente. Como todos os campos, já existiam anteriormente, o povoamento desta tabela foi realizado automaticamente através do comando "*INSERT INTO ... SELECT DISTINCT*".

A tabela de factos é constituída por um total de 16 eixos de análise, sendo 9 dos eixos considerados dimensões degeneradas, e por 1 medida que representa o número total medicamentos receitados por registo. O grão do *data mart* é definido pelos medicamentos prescritos por um médico, a um dado utente num respectivo local e data, período e hora. Esta tabela é representativa das características existentes numa prescrição, como o tipo de consulta, o local de consulta, a comparticipação do medicamento tendo em conta o utente a que é prescrito o medicamento, entre outros. Também nesta tabela foram criados novos campos, o centro de saúde associado à prescrição, o distrito em que foi realizado essa prescrição e uma chave de substituição que servirá como o campo identificador da tabela. Para o povoamento do campo centro de saúde foi criado uma função auxiliar *CSaude()*, que dado um local, devolve o respectivo centro de saúde associado. Para o campo distrito, a função auxiliar *Distrito()* funciona exactamente da mesma maneira, no entanto, neste caso, devolve o distrito em causa. Estas funções foram executadas, para cada registo existente, através da execução de um cursor que foi implementado, que chama ambas as funções e atribui o valor de saída aos respectivos campos. Mais uma vez, a chave primária, é um

número sequencial gerado automaticamente aquando a inserção dos registos. Na Figura 23, pode-se observar o esquema final do *data mart* referido.

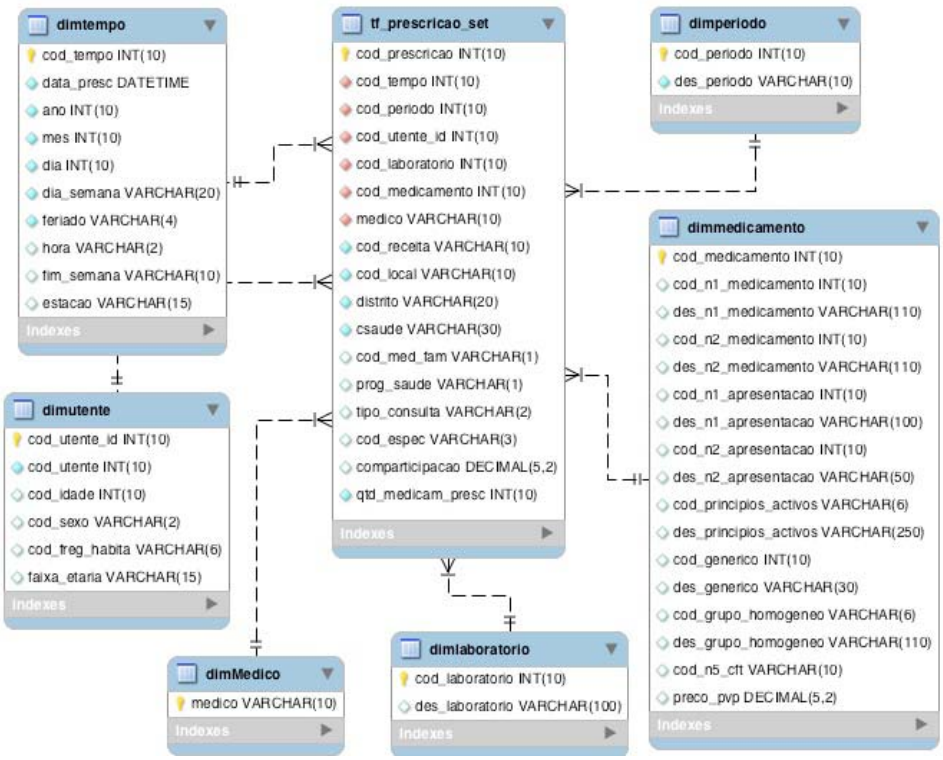


Figura 23 – Esquema do *data mart* implementado.

3.7 Análise Exploratória Complementar

O modelo de dados final apresenta uma gama mais vasta de possíveis atributos de estudos, em que alguns apresentam um peso relevante nos resultados finais. Mais concretamente, os atributos relativos às estações do ano e aos dias da semana, foram usadas na classificação e na segmentação e através deles obteve-se informação relevante. Consequentemente, será feita uma breve análise estatística, tendo em conta estes dois atributos. Na Figura 24, são apresentados dois gráficos correspondentes à análise do número de prescrições por 1000 utentes, relativas às estações do ano (gráfico da esquerda) e dias da semana (gráfico da direita). Como é perceptível, o Outono é a estação que apresenta uma maior taxa de prescrições, o que é considerado normal, devido ao aparecimento, por exemplo, da gripe sazonal e de outros factores. Relativamente aos dias da semana, a segunda-feira e a terça-feira são o que apresentam uma maior taxa do número de prescrições, como também seria de esperar. No entanto, é realmente a segunda-feira que apresenta o maior número de

prescrições (2274 prescrições por 1000 utentes à segunda e 2242 prescrições por 1000 utentes à terça).

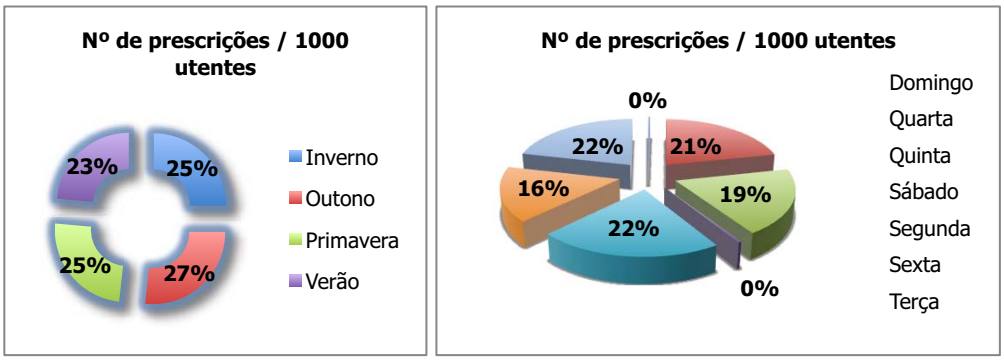


Figura 24 – Percentagem do número de prescrições por 1000 utentes, tendo em conta os eixos de análise das estações do ano e dos dias da semana.

Através deste gráfico (Figura 24 (gráfico da direita)) é possível comprovar que realmente existe um maior número de consultas à segunda-feira, e por sua vez, é na sexta-feira que existe um menor número de consultas durante a semana. Esta teoria é normalmente afirmada pelos médicos e realmente comprova-se isso através desta análise estatística. Apesar de, no gráfico relativo aos dias da semana, a percentagem relativamente ao sábado e domingo ser 0%, em ambos os casos, existem prescrições associadas. No entanto, são demasiado pequenas comparadas com as da semana, o sábado apresenta uma taxa de 40 prescrições por 1000 utentes e o domingo apresenta uma taxa de 32 prescrições por 1000 utentes. Relativamente às análises realizadas anteriormente, e às modificações executadas na obtenção do modelo final, a proporção é semelhante, como tal não se considera relevante a sua repetição.

3.8 Aplicação das Técnicas de Mineração de Dados ao Modelo de Dados Final

Após a análise dos dados, da selecção de dados e atributos e após a implementação do sistema de *data warehousing*, a aplicação de técnicas de mineração de dados aos dados é praticamento automática. Para a aplicação das técnicas ao modelo de dados final foi utilizada a ferramenta *RapidMiner*. Uma grande vantagem na utilizadae desta ferramenta é a possibilidade de conexão directa com uma base de dados, independentemente da versão do sistema de gestão de base de dados. Neste trabalho foi utilizado a versão 5.1 do *MySql*, o que, por exemplo, para o *SPSS Clementine* impossibilita a ligação directa à base de dados. Mas, como o *RapidMiner* é uma ferramenta programada em *java*, basta, através do conector

jdbc e definição das restantes características (*hostname*, *porta*, *user*, *password*, etc.) criar a respectiva ligação. Consequentemente, em todos os estudos realizados para cada uma das técnicas estudadas, os dados vieram quase directamente do *data warehouse*, no entanto, nenhuma modificação foi realizada no *data warehouse* desenvolvido. Ao modelo de dados final vão ser aplicados as três técnicas descritas anteriormente no capítulo 2: a associação, a classificação e a segmentação. Cada uma das técnicas vai ser descrita em separado, e para cada uma delas será realizada uma descrição do algoritmo utilizado, bem como dos parâmetros relevantes.

3.8.1 Associação

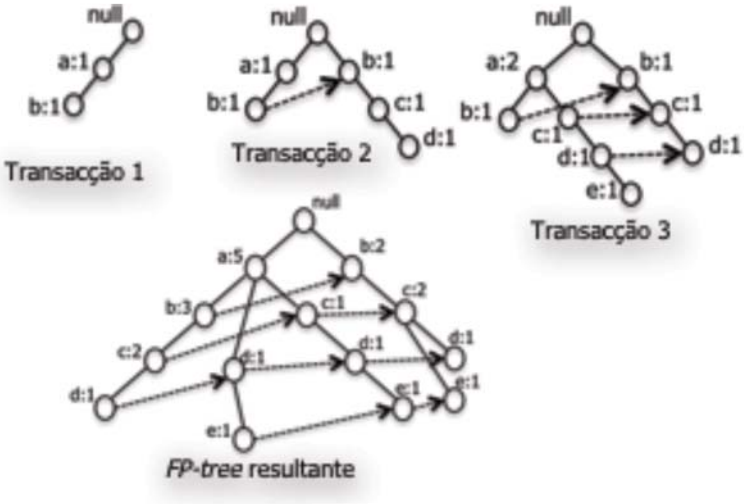
Na aplicação da técnica de associação, ao modelo de dados final, foi utilizado o algoritmo *FP-growth* (*frequent pattern growth*). A implementação e consequente funcionamento do *FP-growth* originou melhorias notáveis comparativamente ao funcionamento do *Apriori* como também em relação aos algoritmos implementados em fases posteriores ao *Apriori* [Wu *et al*. 2008]. A eficiência deste algoritmo é alcançada através da aplicação de 3 métodos [Han *et al*. 2000]:

- Uma base de dados de grande dimensão é compactada numa estrutura bastante menor e mais condensada, que evita custos adicionais nos *scans* à base de dados.
- A estrutura *FP-tree* adopta um método de padrão de crescimento fragmentado que previne custos adicionais na criação de elevados conjuntos de candidatos.
- O método de partição com base no algoritmo *divide-and-conquer* reduz drasticamente o espaço de procura.

Uma das vantagens do *FP-growth* consiste no facto de só ser necessário duas passagens pela base de dados. Na primeira passagem, todos os conjuntos frequentes e o seu respectivo suporte são calculados e guardados pela ordem descendente de suporte em cada transacção. Na segunda passagem, os conjuntos de cada transacção são juntos num prefixo da árvore e são contados os conjuntos (nodos) comuns em diferentes transacções. Cada novo nodo é associado a um elemento e, depois, é realizada a contagem para esse nodo. Os nodos comuns são ligados por um apontador chamado *node-link*. Devido ao facto, de os elementos estarem distribuídos descendentemente pelo valor do suporte, os nodos perto da raíz do prefixo da árvore são os mais comuns nas transacções. Na Figura 25 (a) é apresentado a lista de transacções, já ordenada decrescentemente pelo suporte de cada elemento, da respectiva *FP-tree* criada em (b).

Transacção	Elementos
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{b,c,e}

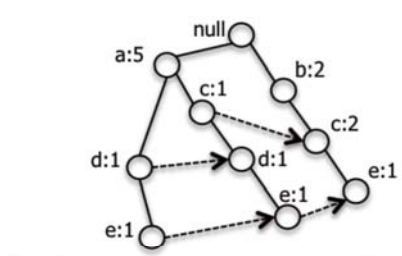
(a)



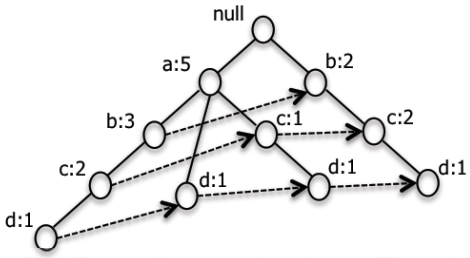
(b)

Figura 25 – Exemplo de uma estrutura FP-tree, em que (a) é a lista de transacções ordenada descendentemente pelo suporte de cada elemento e (b) representa alguns passos da construção da árvore e a árvore resultante.

A partir da *FP-tree*, o algoritmo *FP-growth* começa pelas folhas (os elementos menos frequentes) e extrai os conjuntos frequentes de elementos que contém o elemento em causa e depois funciona recursivamente para os outros elementos [Wu *et al*. 2008]. Mais concretamente, numa primeira fase é construída as sub árvores dos caminhos de cada um dos prefixos. Na Figura 26 é apresentado as sub árvores dos caminhos dos prefixos *e* e *d*.



Sub árvore dos caminhos dos prefixos *e*



Sub árvore dos caminhos dos prefixos *d*

Figura 26 – Sub árvores dos caminhos dos prefixos de *e* e *d*.

De seguida, através da abordagem do algoritmo *divide and conquer*, cada sub árvore dos caminhos dos prefixos é processada recursivamente, de maneira a extrair os conjuntos frequentes. A geração dos conjuntos frequentes é feita a partir dos prefixos menos frequentes para os mais frequentes, isto é, o algoritmo *FP-growth* começa nos elementos das folhas até à raíz. Na Figura 27, é apresentado o exemplo do funcionamento do *FP-growth* através da *FP-tree* para alguns dos conjuntos frequentes de *e*. Inicialmente a sub árvore, dos

caminhos dos prefixos de *e* (Figura 27 (a)), é transforma na *FP-tree* condicional de *e* (Figura 27 (b)), através da actualização do número do suporte de cada elemento. Tendo em consideração um suporte mínimo de 2, constata-se que de (a) para (b), o elemento *b* foi retirado da árvore, pois não respeitava o suporte mínimo. Isso é possível verificar através da Figura 25 (b), na *FP-tree* resultante, em que uma das contagens de *b* é relativa à transacção 2 e a outra à transacção 7, em que depois se subdividem em ramos diferentes, em que num deles o elemento *e* não aparece. Isto significa, que quando se desenha a *FP-tree* condicional de *e*, não se pode ter em conta a transacção 2 e, consequentemente, o suporte de *be* e *ce* baixa. No entanto, como o elemento *c* encontra-se ligado por um *node-link* a outro elemento *c*, o suporte de *ce* é igual a 2, logo *ce* é frequente. Após a obtenção da *FP-tree* condicional o processo é aplicado recursivamente, até se obter todos os conjuntos frequentes relativos a *e*. Na Figura 27 (c) e (d) é representado a sub árvore dos caminhos dos prefixos de *de* (que por (b) observa-se que é um conjunto frequente) e a *FP-tree* condicional de *de*, respectivamente. Por essas duas imagens é observável que o conjunto *ade* é frequente, no entanto o *acde* não é frequente, pois o elemento *cde* também não é frequente, para um suporte mínimo de 2.

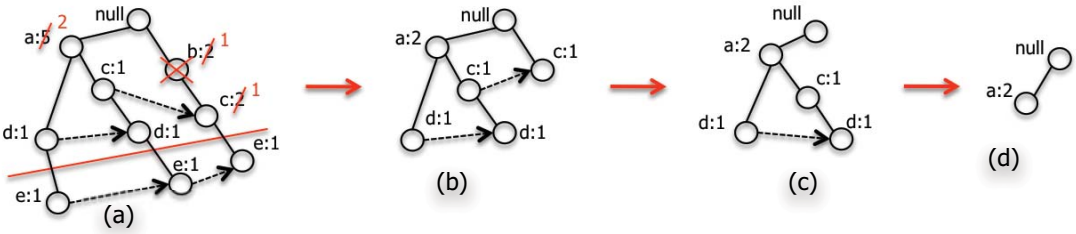


Figura 27 – Exemplo do funcionamento do algoritmo *FP-growth* através da *FP-tree* com um suporte minímo de 2 [www5].

Após a aplicação do algoritmo FP-Growth, ao exemplo mencionado na Figura 25 (a), foram descobertos os conjuntos frequentes descritos na Tabela 19.

Sufixos	Conjuntos Frequentes
{e}	{e},{d,e},{a,d,e},{c,e},{a,e}
{d}	{d},{c,d},{b,c,d},{a,c,d},{b,d},{b,c,d},{a,d}
{c}	{c},{b,c},{a,b,c},{a,c}
{b}	{b},{a,b}
{a}	{a}

Tabela 19 – Conjuntos frequentes descobertos pela aplicação do *FP-growth* à lista de transacções da Figura 25 (a).

Florian Verhein [www5] é responsável por uma breve e clara apresentação relativa à criação da estrutura *FP-tree*, e pelo funcionamento do *FP-growth* na descoberta de elementos frequentes através da *FP-tree*.

A escolha deste algoritmo provém, no facto, de que comparativamente com o *Apriori*, o *FP-growth* apresenta baixos custos computacionais através da estrutura *FP-tree*, devido a ser apenas necessário dois *scans* à base de dados, e por sua vez é um algoritmo confiável na descoberta de padrões frequentes sejam eles pequenos ou grandes (Tabela 6).

Pré-processamento

A aplicação do algoritmo *FP-growth* ao modelo de dados final, envolveu a necessidade de aplicação de alguns passos de pré-processamento. De seguida é apresentado o código XML usado para o pré-processamento dos dados e aplicação da técnica da associação.

```
<operator name="Root" class="Process" expanded="yes">
      <operator name="DatabaseExampleSource" class="DatabaseExampleSource">
          <parameter key="work_on_database"  value="false"/>
          <parameter key="database_system"   value="MySQL"/>
          <parameter key="database_url" value="jdbc:mysql://localhost:3306/joana"/>
          <parameter key="username"   value="joana"/>
          <parameter key="password"   value="cB1pwz1nlIQ="/>
          <parameter key="query"      value="select       receita       as       TID,
GROUP_CONCAT(medicamento  ORDER  BY  1  SEPARATOR  “,”)  as  ITEM  FROM  fpg_medi  WHERE
distrito=”Porto” GROUP BY receita;"/>
          <parameter key="id_attribute"      value="TID"/>
          <parameter key="datamanagement"    value="int_array"/>
      </operator>
      <operator name="Split" class="Split">
          <parameter key="attributes" value="ITEM"/>
          <parameter key="apply_to_special_features" value="true"/>
          <parameter key="split_pattern"     value=","/>
          <parameter key="split_mode" value="unordered_split"/>
      </operator>
      <operator name="FPGrowth" class="FPGrowth">
          <parameter key="keep_example_set"  value="true"/>
          <parameter key="find_min_number_of_itemsets"      value="false"/>
          <parameter key="min_number_of_itemsets"    value="100"/>
          <parameter key="min_support"       value="0.0010"/>
          <parameter key="max_items"  value="-1"/>
      </operator>
      <operator name="AssociationRuleGenerator" class="AssociationRuleGenerator">
          <parameter key="keep_frequent_item_sets"   value="true"/>
          <parameter key="criterion"  value="confidence"/>
          <parameter key="min_confidence"    value="0.3"/>
          <parameter key="min_criterion_value"       value="0.8"/>
          <parameter key="gain_theta" value="2.0"/>
```

```
        <parameter key="laplace_k"  value="1.0"/>
    </operator>
</operator>
```

Para a aplicação do *FP-growth* é necessário que os dados estejam agregados pelo atributo de transacção. Neste caso, o atributo de transacção foi o *cod_receita*, pois o principal objectivo é descobrir associações entre os medicamentos prescritos, isto é, quais os medicamentos que são normalmente prescritos em conjunto. Para este problema a solução encontrada passou por duas fases. Numa primeira fase, os dados em estudo (as receitas e os medicamentos) foram importados para uma nova tabela (*fpg_medi*), criada com o intuito de facilitar o processo e de não ser necessário trabalhar com o modelo de dados final de aproximadamente 300 Mb. Numa segunda fase, o objectivo foi agrupar os medicamentos prescritos por cada receita através da execução de um comando SQL que utiliza a função *GROUP_CONCAT()* do *MySql*: "*SELECT receita as TID, GROUP_CONCAT(medicamento ORDER BY 1 SEPARATOR ',') as ITEM FROM fpg_medi GROUP BY receita*". Este comando agrupa os medicamentos, por receita, separados por uma vírgula e foi executado através do operador "*DatabaseExampleSource*".

Pelo o código XML, apresentado anteriormente, é possível verificar-se a utilização do operador "*Split*", que serviu para transformar os dados agrupados para o formato binomial. Neste operador é mencionado qual o atributo a transformar (*ITEM*) e qual o separador ",". O parâmetro *split_mode=unordered_split,* é o principal parâmetro responsável em transformar os dados num formato binomial. Após o pré-processamento dos dados é aplicado o operador "*FPGrowth*" e de seguida o operador "*AssociationRuleGenerator*", que transforma os conjuntos frequentes em regras, com base num parâmetro de confiança escolhido. Em ambos estes operadores, são mencionadas medidas de avaliação já apresentadas na secção da associação, como tal não faz sentido descrevê-las novamente.

Avaliação dos Resultados Otidos

Na aplicação do operador "*FPGrowth*" é necessário escolher o suporte (Equação (2)) mínimo que foi definido em 0.001. Através deste operador foram gerados todos os conjuntos frequentes que respeitem o suporte mínimo, e de seguida foram criadas regras através do operador "*AssociationRuleGenerator*" com o parâmetro *criterion=confidence* e com uma confiança (Equação (3)) mínima de 0.3. Foram escolhidos estes valores para os critérios de avaliação, devido ao elevado número de registos e da quantidade de possíveis medicamentos. Foi também realizado o estudo tendo em conta o valor do parâmetro *lift* (Equação (4)) e serão apresentados também as regras que apresentem um *lift* superior a 50. Isto indica que na regra $A \Rightarrow B$, o aparecimento do elemento A, encontra-se associado ao

aparecimento do elemento B. De notar que é uma medida simétrica, ao contrário da confiança. O objectivo da aplicação deste operador era, inicialmente, descobrir a associação entre a prescrição dos medicamentos para todos os distritos em conjunto, no entanto, a memória disponível não o permitiu. Consequentemente, aplicou-se o conjunto de operadores, mencionados no código XML, a cada distrito separadamente, sempre com os mesmos parâmetros. Na Tabela 20 estão apresentadas as regras que respeitam os parâmetros mínimos definidos, divididas pelos diferentes distritos.

Distritos	Regras	Sup.	Conf.	Lift
Aveiro	Adalgur N ⇒ Arthtotec	0.001	0.180	67.92
	Arthtotec ⇒ Adalgur N	0.001	0.385	67.92
	Vigantol ⇒ Cebiolon	0.002	0.747	444.1
	Cebiolon ⇒ Vigantol	0.002	0.896	444.1
Braga	Ben-U-Ron ⇒ Brufen suspensão	0.001	0.240	55.15
	Brufen suspensão ⇒ Ben-U-Ron	0.001	0.247	55.15
	Atrovent Unidose ⇒ Ventilan	0.002	0.404	94.22
	Vigantol ⇒ Cebiolon	0.001	0.430	248.5
	Ventilan ⇒ Atrovent Unidose	0.002	0.452	94.22
	Cebiolon ⇒ Vigantol	0.001	0.817	248.5
Bragança	Vigantol ⇒ Cebiolon	0.001	0.750	436.5
	Cebiolon ⇒ Vigantol	0.001	0.857	436.5
Porto	Ben-U-Ron ⇒ Brufen suspensão	0.002	0.240	54.26
	Brufen suspensão ⇒ Ben-U-Ron	0.002	0.359	54,26
	Atrovent Unidose ⇒ Ventilan	0.002	0.378	73.99
	Profenid ⇒ Relmus	0.001	0.386	69.72
	Relmus ⇒ Profenid	0.001	0.181	69.72
	Ventilan ⇒ Atrovent Unidose	0.002	0.458	73.99
	Relmus ⇒ Voltaren	0.003	0.582	113.0
	Vigantol ⇒ Cebiolon	0.003	0.594	169.6
	Voltaren ⇒ Relmus	0.003	0.626	113.0
	Cebiolon ⇒ Vigantol	0.003	0.756	169.6
Viana do Castelo	Atrovent Unidose ⇒ Ventilan	0.001	0.269	104.2
	Relmus ⇒ Profenid	0.001	0.332	188.5
	Ventilan ⇒ Atrovent Unidose	0.001	0.423	104.2
	Relmus ⇒ Voltaren	0.002	0.507	172.7
	Voltaren ⇒ Relmus	0.002	0.572	172.7
	Profenid ⇒ Relmus	0.001	0.624	188.5
Vila Real	Atrovent Unidose ⇒ Ventilan	0.002	0.430	105.8
	Ventilan ⇒ Atrovent Unidose	0.002	0.470	105.8
	Vigantol ⇒ Cebiolon	0.002	0.803	302.7
	Cebiolon ⇒ Vigantol	0.002	0.808	302.7

Tabela 20 – Regras de Associação entre os medicamentos prescritos.

Pela Tabela 20, constata-te grandes associações entre alguns medicamentos, através do alto valor de *lift*. Isto significa que, por exemplo, para o distrito de Aveiro, a prescrição do medicamento *Cebiolon* é normalmente prescrito em conjunto com o medicamento *Vigantol*. Isto é, as prescrições que contém o medicamento *Cebiolon* tendem a conter o medicamento *Vigantol*, mais vezes do que prescrições do medicamento *Vigantol* que não contém o medicamento *Cebiolon*. A associação entre os medicamentos é bastante visível, pois, maior parte das regras apresentadas estão associadas a um valor de *lift* elevado, mesmo com uma confiança baixa, como é o caso da regra *Adalgur N* $\Rightarrow$ *Arthtotec*. Esta regra tem uma confiança de apenas 0.182, o que significa que apenas 18% das prescrições que contém o medicamento *Adalgur N* também contém o medicamento *Arthtotec*, no entanto, 30% das prescrições que contém o medicamento *Arthtotec* também contém o medicamento *Adalgur N*. Apesar de serem duas regras bastantes idênticas, no caso da confiança, é medida a associação entre cada uma de acordo com a ordem que elas aparecem nas prescrições. Consequentemente, nesta análise, o parâmetro *lift*, é considerado um melhor parâmetro de avaliação de regras para a prescrição dos medicamentos, pois a ordem de prescrição dos medicamentos em nada modifica a associação da sua prescrição. Councluindo, a medida *lift* é uma medida simétrica, que mede a co-ocorrência de dois ou mais medicamentos, independentemente da sua ordem, e é isso que está em causa neste estudo inicial.

Relativamente ao estudo da associação ao modelo de dados final, foi realizado ainda outro estudo, mas desta feita para descobrir possíveis associações entre os médicos prescritores e os laboratórios prescritos. A única alteração neste estudo, foi a criação de outra tabela, mas desta vez com os atributos relativos ao código das receitas, dos médicos e dos códigos dos laboratórios. Numa fase posterior foi necessário incluir outro operador "*Split*", com o intuito de agrupar os laboratórios por receita, enquanto que o inicial "Split" foi utilizado para agrupar os médicos por receita. Como neste caso, a gama de valores dos códigos dos laboratórios é bastante inferior à gama dos medicamentos, foi possível gerar regras de associação para todos os distritos em conjunto. No entanto, também foi experimentado gerar para cada distrito, mas os resultados não foram tão satisfatórios (muito poucas associações entre médicos e laboratórios, pois o número de médicos era bastante inferior ao número de laboratórios existentes). Neste caso a confiança mínima foi reduzida para 0.2, pois o facto de 20% das prescrições de um médico conterem um determinado laboratório, é já considerado algo razoavelmente interessante.

Regras	Sup.	Conf.	Lift
Med_819674 ⇒ Sanofi Aventis	0.001	0.200	1.635
Med_219975 ⇒ Sanofi Aventis	0.001	0.200	1.638
Med_458534 ⇒ Sanofi Aventis	0.001	0.203	1.661
Med_13734 ⇒ Sanofi Aventis	0.001	0.203	1.662
Med_989878 ⇒ Sanofi Aventis	0.001	0.207	1.690
Med_366277 ⇒ Sanofi Aventis	0.001	0.207	1.694
Med_6370 ⇒ Sanofi Aventis	0.001	0.208	1.697
Med_162131 ⇒ Sanofi Aventis	0.001	0.208	1.701
Med_77941 ⇒ Sanofi Aventis	0.001	0.209	1.704
Med_812629 ⇒ Sanofi Aventis	0.001	0.211	1.720
Med_576377 ⇒ Sanofi Aventis	0.002	0.212	1.729
Med_989879 ⇒ Lab. Pfizer	0.002	0.212	2.348
Med_140121 ⇒ Sanofi Aventis	0.002	0.214	1.749
Med_528379 ⇒ Sanofi Aventis	0.001	0.214	1.751
Med_777470 ⇒ Sanofi Aventis	0.002	0.216	1.764
Med_750475 ⇒ Sanofi Aventis	0.001	0.216	1.768
Med_915872 ⇒ Sanofi Aventis	0.001	0.217	1.770
Med_987832 ⇒ Sanofi Aventis	0.001	0.217	1.773
Med_940838 ⇒ Sanofi Aventis	0.001	0.218	1.781
Med_610021 ⇒ Sanofi Aventis	0.001	0.218	1.782
Med_149172 ⇒ Sanofi Aventis	0.001	0.218	1.783
Med_276976 ⇒ Sanofi Aventis	0.002	0.219	1.791
Med_412552 ⇒ Merck Genéricos	0.002	0.222	5.733
Med_181172 ⇒ Sanofi Aventis	0.001	0.222	1.818
Med_412552 ⇒ Sanofi Aventis	0.002	0.223	1.824
Med_675076 ⇒ Sanofi Aventis	0.002	0.225	1.837
Med_550371 ⇒ Sanofi Aventis	0.002	0.227	1.851
Med_462576 ⇒ Sanofi Aventis	0.001	0.227	1.856
Med_986877 ⇒ Sanofi Aventis	0.001	0.228	1.861
Med_683028 ⇒ Sanofi Aventis	0.002	0.242	1.975
Med_257973 ⇒ Sanofi Aventis	0.001	0.242	1.979
Med_250921 ⇒ Generis	0.001	0.252	6.141
Med_623024 ⇒ Sanofi Aventis	0.001	0.255	2.084
Med_735424 ⇒ Sanofi Aventis	0.001	0.265	2.167
Med_245979 ⇒ Sanofi Aventis	0.001	0.266	2.170
Med_179175 ⇒ Sanofi Aventis	0.001	0.283	2.312
Med_592328 ⇒ Tetrafarma	0.002	0.311	22.09

Tabela 21 – Regras de Associação entre os médicos e os laboratórios prescritos.

Na Tabela 21, são apresentadas as regras de associação relativamente aos laboratórios dos medicamentos prescritos pelos médicos. Nesta caso é possível constatar que para quase

todas as regras, o valor de *lift* é bastante próximo de 1, o que significa que a ocurrência do antecedente da regra, não tem praticamente efeito na ocurrência do consequente da regra. Como neste caso, os diferentes laboratórios podem ser prescritos por diversos médicos, a medida *lift* não confere um conhecimento útil. No entanto, a confiança, neste caso, apresenta um medida com muito mais significância, uma vez, que o objectivo é descobrir se existe alguma associação entre a prescrição dos médicos e os laboratórios prescritos. Isto é, se existem médicos cujas as suas prescrições são tendenciosas para algum laboratório. Tendo em conta uma confiança superior a 20%, obteve-s algumas regras, em que, em praticamente todas, evidenciam a preferência de determinados médicos pelos laboratórios *Sanofi Aventis*. Apenas quatro regras apresentaram a tendência de médicos por outros laboratórios, sendo duas das regras referentes a laboratórios de medicamentos genéricos. Como a percentagem de laboratórios genéricos é bastante limitada, essas duas regras poderão não ser consideradas como sendo algo relevante. No entanto, a última regra, principalmente, apresenta um padrão fora do normal, pois é o único médico cuja as suas prescrições estão associadas ao laboratório TretaFarma, e a sua confiança (30%) é superior a todas as outras regras.

3.8.2 Segmentação

Neste caso, o algoritmo utilizado foi o *k-means*, que é um dos algoritmos mais utilizados na mineração de dados. Devido ao facto do algoritmo apresentar condições para uma elevada dimensionalidade e escalabilidade, e de o modelo de dados final não apresentar ruído, a opção da utilização do *k-means* é bastante viável (Tabela 7). De uma maneira geral, o *k-means* é inicializado através da escolha de k pontos, de um conjunto de vectores d-dimensionais, como sendo os k segmentos iniciais, representados por centróides. A escolha desses centróides é feita com base numa amostra aleatória retirada do conjunto de dados, definindo-os como sendo a solução da segmentação de pequenos subconjuntos de elementos do conjunto de dados. Numa segunda fase, é feita uma associação dos restantes dados a cada um dos centróides, que resulta na partição dos dados [Wu *et al*. 2008]. Por fim, é realizado iterativamente, uma recolocação dos centróides, até que a convergência seja alcançada ou sejam impostos critérios de paragem. Uma descrição mais detalhada é descrita na secção 2.5.2, na parte dos algoritmos particionais. Apesar de ser um algoritmo com bastante visibilidade, é também bastante susceptível à escolha de maus centroides iniciais. Isto significa, que para o mesmo conjunto de dados, caso os centróides sejam mal atribuídos, o algoritmo pode convergir, mas mesmo assim apresentar maus resultados, quando, comparativamente, à convergência com a atribuição de bons centróides. Consequentemente, poderá ser necessário executar o algoritmo mais do que uma vez, até se obter o melhor

resultado possível [Wu *et al*. 2008]. Consequentemente, na aplicação deste algoritmo ao modelo de dados, foi tido em conta diversas execuções do algoritmo, através do parâmetro *max_runs*.

Neste caso, os atributos escolhidos foram o atributo cod_prescricao como chave identificadora, e os atributos *distrito*, *des_generico*, *des_periodo*, *grupo_etario*, *estacao*, *cod_sexo*, *mes* e *dia_semana*, como atributos regulares. Sendo criado o atributo *cluster* como *label* do modelo de dados

Pré-processamento

Para a aplicação do *k-means* ao modelo de dados, foi necessário realizar alguns procedimentos de pré-processamento, tanto para ser possível trabalhar com o tipo de dados em causa, mas também para se obter os melhores resultados possíveis. O *k-means* é um algoritmo que não possibilita a utilização directa de atributos nominais, como tal, foi necessário proceder-se à passagem dos atributos nominais para atributos numéricos. O *RapidMiner* apresenta um operador que automaticamente faz esse procedimento, "*Nominal2Numerica*l". Este operador consiste, basicamente, em transformar os valores nominais em valores numéricos, através de uma mapeamento directo, isto é, para o caso do atributo mês, o mês de Janeiro será convertido no número 1, o mês de Fevereiro no número 2 e assim sucessivamente.

Recentemente, foi realizado um estudo [Visalakshi *et al*. 2009] sobre o impacto da normalização na aplicação do algoritmo *k-means*, em que se provou que, normalmente, através da normalização dos dados, se obtém melhores resultados. Consequentemente, neste estudo, foi aplicado a normalização dos dados através do operador "*Normalization*". O estilo de normalização aplicada foi a transformada de Z (Equação (30)), que representa uma distribuição normal com média 0 e variância 1, $N(0,1)$.

$$Z_i = \frac{X_i - \mu}{\sqrt{\sigma^2}} \tag{30}$$

Em que X é a variável aleatória em estudo com média μ e desvio padrão σ (Equação (31)).

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1} \tag{31}$$

Na Tabela 22, é apresentado a gama de valores para cada um dos atributos regulares usados no estudo relativo à aplicação do algoritmo *k-means*, após a aplicação da normalização.

Atributos	Gama de Valores
distrito	[-0.882,2.167]
des_generico	[-0.458,2.183]
des_periodo	[-0.157,6.356]
grupo_etario	[-1.488,3.201]
cod_sexo	[-0.733,1.365]
Estacão	[-1.342;1.305]
mes	[-1.571,1.584]
dia_semana	[-1.353,2.758]

Tabela 22 – Gama de valores dos atributos em estudo na aplicação do algoritmo *k-means*.

Após a normalização dos atributos em causa, foi aplicado o algoritmo ao modelo de dados. No entanto, como foi mencionado acima, o *k-means* apresenta algumas desvantagens, em que uma delas é a escolha dos k centróides inicias. Para se tentar encontrar o melhor resultado, a opção de *max_runs* foi colocada a 50, isto quer dizer que o algoritmo nesta condições será aplicado 50 vezes ao modelo de dados. Isto permite descobrir melhores resultados, pois em cada vez que o algoritmo é corrido os centróides inicias são mudados (isto significa que a variável *use_local_random_seed* se encontra a falso), e isso quer dizer, que com outros centróides aleatórios se poderá obter melhores resultados. Mas a incrementação do número de *runs* do algoritmo poderá não ser muito eficaz, caso o número de operações de optimização seja pequeno (poderá não permitir a convergência do algoritmo). Isto é, o *k-means* tem duas opções de paragem, ou converge ou é utilizado um número máximo de passos de optimização. Neste caso, o *max_optimization_steps* foi colocado a 500, o que significa que o algoritmo ou converge, ou após 500 iterações o algoritmo pára. A segunda grande desvantagem do *k-means*, é a escolha acertada do k, isto é, a escolha do número de segmentos finais pode declarar a qualidade do resultado final obtido, como tal, é necessário uma escolha ponderada do factor k. O objectivo do *k-means* é minimizar a distância intra-segmentos e maximizar a distância inter-segmentos, e a medida que melhor define estes dois objectivos é a medida *Davies Bouldin* (DB) [Davies & Bouldin 1979] (Equação (32)). Esta medida tem em consideração a separação relativa dos segmentos mais separados. O objectivo do DB é minimizar o seu valor, pois valores baixos correspondem a segmentos mais compactos e cujo os centróides estão mais distantes dos outros [Mierswa 2009].

$$w_{DB} = \frac{1}{k}\sum_{q=1}^{k}\max_{q,r\neq q}\left\{\frac{s_q + s_r}{d(C_q,C_r)}\right\} \tag{32}$$

Na Equação (32), o s_q e s_r são a distancia média intra-segmentos para os segmentos C_q e C_r, respectivamente. No entanto, esta medida funciona de uma maneira diferente no *RapidMiner*, isto porque, o *RapidMiner* apenas consegue realizar optimizações, através da maximização de funções. Logo, com o intuito de transformar um problema de minimização num problema de maximização, simplesmente, é necessário multiplicar a função original por -1 e em vez disso maximizá-la. Resumindo, nesta caso, é necessário maximizar o valor de DB (valor negativo) obtido pelo operador "*ClusterCentroidEvaluator*". Como acréscimo de informação, foram usadas, também, mais duas medidas de caracterização de qualidade dos segmentos, a soma dos quadrados dos erros (SSE), que calcula o erro de cada ponto, isto é, utiliza o valor da distância euclidiana ao centróide mais próximo e de seguida calcula a soma total dos quadrados dos erros (Equação (33)). O objectivo é minimizar este valor, uma vez que significa que os centróides dos segmentos representam melhor os pontos existentes nesse segmento (menor distância entre intra-segmentos) [Tan *et al*. 2006].

$$SSE = \sum_{i=1}^{k}\sum_{x\in C_i} dist(c_i,x)^2 \tag{33}$$

Em que c_i é a média do centróide do segmento. Esta medida é bastante sensível ao k, pois à medida que o k aumenta, o SSE diminui, devido a isso é que se utiliza também a medida DB como avaliação dos segmentos. Esta medida é calculada através do operador "*ItemDistributionEvaluator*". Por fim, a última medida utilizada foi a média da distância intra-segmento, isto é, para cada segmento é calculada a média da distância dos centróides aos pontos do segmento, e de seguida, é feita a média dessas médias, calculadas anteriormente. De uma maneira geral, esta medida apresenta as mesmas qualidades que a SSE, visto que a SSE é calculada através desta medida (c_i), como tal o objectivo é também minimizar o valor obtido. No entanto, também esta medida é implementada como um problema de maximização, consequentemente, o objectivo final é maximizar o seu valor, que é obtido através do operador "*ClusterCentroidEvaluator".*

Observando a Tabela 23 e a Tabela 24, constata-se que o valor do SSE e da média são valores que vão decrescendo à medida que o k aumenta, daí serem medidas sensíveis ao k. Observando os valores da medida DB, constata-se que o k que maximiza este valor é o

$k = 13$, consequentemente será o k escolhido e como tal o modelo final resultante será composto por 13 segmentos.

Medida	k=2	k=3	k=4	k=5	k=6	k=7	k=8
SSE	0.501	0.478	0.356	0.261	0.200	0.178	0.143
Avg.	-6.515	-5.561	-4.832	-4.420	-4.110	-3.822	-3.573
DB	-1.967	-1.488	-1.411	-1.423	-1.616	-1.440	-1.421

Tabela 23 – Resultados da avaliação do *k-means* para $k = 2$ até ao $k = 8$.

Medida	k=9	k=10	k=11	k=12	k=13	k=14
SSE	0.133	0.120	0.102	0.092	0.083	0.079
Avg.	-3.315	-3.147	-2.984	-2.849	-2.766	-2.665
DB	-1.406	-1.401	-1.371	-1.368	-1.347	-1.380

Tabela 24 – Resultados da avaliação do *k-means* para $k = 9$ até ao $k = 14$.

De seguida é apresentado o código XML do processo de criação dos 13 segmentos com todos os operadores e respectivos parâmetros utilizados para tal.

```
<operator name="Root" class="Process" expanded="yes">
      <operator name="DatabaseExampleSource" class="DatabaseExampleSource">
          <parameter key="work_on_database"  value="false"/>
          <parameter key="database_system"   value="MySQL"/>
          <parameter key="database_url" value="jdbc:mysql://localhost:3306/dw_tese"/>
          <parameter key="username"   value="joana"/>
          <parameter key="password"   value="N2VcmJDbe8M="/>
          <parameter key="query"      value="select
tf.cod_prescricao,tf.distrito,dm.des_generico,dp.des_periodo,du.grupo_etario,du.cod_se
xo,dt.estacao,dt.mes,dt.dia_semana  from  tf_prescricao_set  as  tf  &#10;inner  join
dimmedicamento as dm on tf.cod_medicamento=dm.cod_medicamento inner join dimperiodo as
dp   on   tf.cod_periodo=dp.cod_periodo   inner   join   dimutente   as   du&#10;on
tf.cod_utente_id=du.cod_utente_id     inner     join     dimtempo     as     dt     on
tf.cod_tempo=dt.cod_tempo;"/>
          <parameter key="id_attribute"      value="cod_prescricao"/>
          <parameter key="datamanagement"    value="double_array"/>
      </operator>
      <operator name="Nominal2Numerical" class="Nominal2Numerical">
          <parameter key="return_preprocessing_model"       value="true"/>
          <parameter key="create_view"       value="true"/>
      </operator>
      <operator name="Normalization" class="Normalization">
          <parameter key="return_preprocessing_model"       value="true"/>
          <parameter key="create_view"       value="true"/>
          <parameter key="method"     value="Z-Transformation"/>
          <parameter key="min"        value="0.0"/>
          <parameter key="max"        value="1.0"/>
```

```
    </operator>
    <operator name="KMeans" class="KMeans">
        <parameter key="keep_example_set"  value="true"/>
        <parameter key="add_cluster_attribute"     value="true"/>
        <parameter key="k"  value="13"/>
        <parameter key="max_runs"   value="50"/>
        <parameter key="max_optimization_steps"    value="500"/>
        <parameter key="use_local_random_seed"     value="false"/>
        <parameter key="local_random_seed" value="1"/>
    </operator>
    <operator name="ItemDistributionEvaluator" class="ItemDistributionEvaluator">
        <parameter key="measure"    value="SumOfSquares"/>
    </operator>
    <operator name="ClusterCentroidEvaluator" class="ClusterCentroidEvaluator">
        <parameter key="main_criterion"    value="Davies Bouldin"/>
    </operator>
</operator>
```

Avaliação dos Resultados Obtidos

Como se pode ver pelo código XML, os resultados foram obtidos através do operador *k-means*, em que os parâmetros definidos foram *k=13*, *max_runs=50*, *max_optimization_steps=500* e *add_cluster_attribute=true*. Os 13 segmentos gerados apresentam diferentes números de instâncias e a qualidade entre os segmentos difere, isto é, a distancia média intra-segmentos difere de segmento para segmento. Pela Tabela 25 e pela Tabela 25, é possível verificar, que o segmento com mais instâncias associadas é o segmento 1, e o que o segmento 10 é o mais compacto, consequentemente é o que melhor representa as instâncias a si associadas.

	seg. 0	**seg. 1**	**seg. 2**	**seg. 3**	**seg. 4**	**seg. 5**	**seg. 6**
Nº de instâncias	54080 (2.42%)	239437 (10.70%)	98500 (4.40%)	146178 (6.53%)	225137 (10.06%)	198902 (8.89%)	175061 (7.82%)
Avg.	-6.328	-2.568	-4.656	-1.709	-1.736	-1.845	-3.273

Tabela 25 – Resultados do segmento 0 ao segmento 6.

	seg. 7	**seg. 8**	**seg. 9**	**seg. 10**	**seg. 11**	**seg. 12**	**Total**
Nº de instâncias	167043 (7.46%)	200179 (8.94%)	208888 (9.33%)	148582 (6.64%)	196730 (8.79%)	179898 (8.02%)	2238615 (100%)
Avg.	-3.088	-1.768	-2.394	-1.399	-4.386	-4.192	-2.766

Tabela 26 – Resultados do segmento 7 ao segmento 12.

Cada segmento obtido apresenta as suas próprias características, consoante as instâncias que o representam. Consequentemente, a cada segmento é atribuído uma média dos valores

de cada atributo que melhor designam esse segmento. A Figura 28 descreve o que foi dito, mais propriamente, é possível verificar que para cada segmento criado é associado o valor médio de cada atributo.

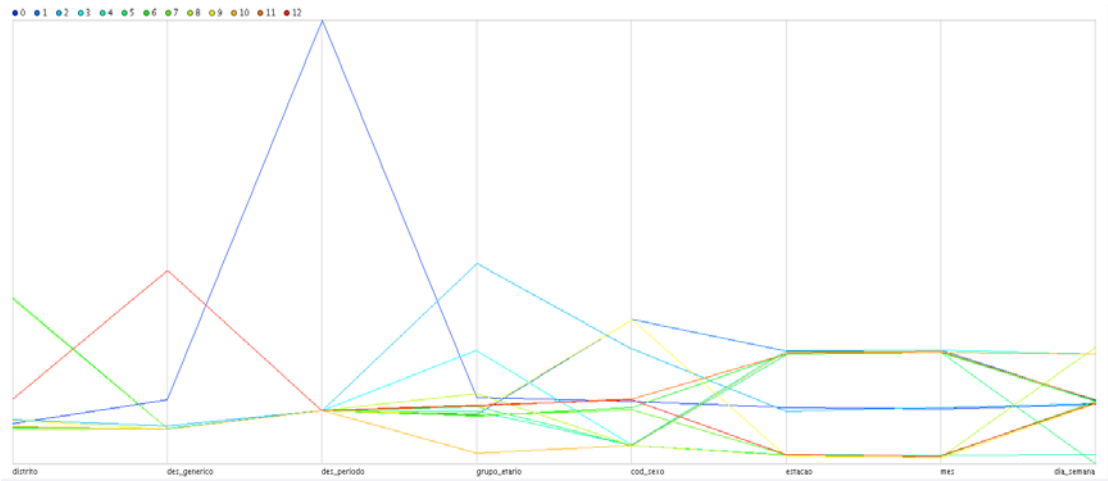


Figura 28 – Representação gráfico dos valores médios dos centróides para os atributos em estudo.

Foi usado também o operador "*DatabaseExampleSetWriter*", mas para isso foi necessário retirar os operadores "*ItemDistributionEvaluator*" e "*ClusterCentroidEvaluator*", pois para este operador é preciso dar como *input* o resultado final. O operador "DatabaseExampleSetWriter" foi utilizado com o intuito de importar o resultado obtido, através da aplicação do algoritmo *k-means*, para uma tabela criada com os respectivos campos em estudo. Desta maneira, através da nova tabela criada e do operador "*DatabaseExampleSource*", foi possível aplicar o operador "*W-J48*" ao resultado obtido pela segmentação, e assim obter um conjunto de regras que melhor definem os segmentos finais obtidos. Este operador funciona da mesma maneira que o algoritmo C4.5 das árvores de decisão e o seu resultado final é apresentado num formato não gráfico, consequentemente utiliza menos memória. Nesta segunda fase da segmentação é executado um algoritmo de classificação, e todo o processo associado é idêntico ao usado na secção 3.8.3, como tal, os parâmetros de qualidade usados serão mencionados e explicados na devida secção. Os resultados finais foram obtidos através do seguinte XML.

```
<operator name="Root" class="Process" expanded="yes">
      <operator name="DatabaseExampleSource" class="DatabaseExampleSource">
          <parameterkey="database_url" value="jdbc:mysql://localhost:3306/clust"/>
          <parameter key="username"   value="joana"/>
          <parameter key="password"   value="N2VcmJDbe8M="/>
          <parameter key="table_name" value="distri"/>
          <parameter key="label_attribute"    value="cluster"/>
          <parameter key="id_attribute"       value="cod_prescricao"/>
```

```
          <parameter key="datamanagement"    value="double_array"/>
      </operator>
      <operator name="XValidation" class="XValidation" expanded="yes">
          <parameter key="sampling_type"     value="suffled sampling"/>
          <operator name="W-J48" class="W-J48">
              <parameter key="C" value="0.5"/>
              <parameter key="M" value="500"/>
          </operator>
          <operator name="OperatorChain" class="OperatorChain" expanded="yes">
              <operator name="ModelApplier" class="ModelApplier">
                  <list key="application_parameters">
                  </list>
              <operator                                  name="ClassificationPerformance"
class="ClassificationPerformance">
                  <parameter key="accuracy" value="true"/>
                  <parameter key="classification_error" value="true"/>
                  <parameter key="weighted_mean_recall" value="true"/>
                  <parameter key="weighted_mean_precision" value="true"/>
                  <list key="class_weights">
                  </list>
              </operator>
          </operator>
      </operator>
  </operator>
```

O modelo final, resultante da aplicação do algortimo "*W-J48*", caracteriza os dados com bastante qualidade e certeza, visto que a sua percentagem de previsão é de 99.75%, consequentemente a taxa de erro é 0.25%. Foi também avaliado o parâmetro *precision* (99.65%) e *recall* (99.68%), que calculando pela Equação (36), obteve-se um *f-measure* de 99.66%, o que mais uma vez comprova que foi obtida uma boa previsão do modelo. Analisando as regras obtidas através da aplicação da classificação ao conjunto de segmentos obtidos através da segmentação, pode-se concluir alguns padrões de prescrição.

Relativamente à prescrição de genéricos não existem padrões muito vincados, no entanto, destaca-se que são as pessoas de idade superior a 15 anos que recorrem mais a este tipo de medicamentos. Tendo em consideração as prescrições dos medicamentos não genéricos, é possível assumir algumas características importantes dos dados, nomeadamente relativamente às estações do ano, sexo do utente e dias da semana. Isto é, nas estações da Primavera e Inverno, as prescrições dos utentes do sexo feminino apresentam características interessantes e que diferem bastante das características de prescrições dos utentes do sexo masculino. Nos distritos de Braga, Porto e Bragança os padrões de prescrições das mulheres nas estações de Primavera e Inverno são semelhantes. No entanto, nos dias da semana de segunda e quarta, dos meses Janeiro, Fevereiro, Março, Abril, Maio e Junho, existe uma predominância de mulheres de idade superior a 15 anos, sendo que as de idade inferior, mais propriamente dos 0 aos 4 anos, apresentam um aumento de prescrições durante a estação

do Inverno (Janeiro, Fevereiro e Março), é à quarta-feira. Nos restantes dias da semana, à sexta existe uma maior afluência dos utentes de idade superior aos 15 anos, e na terça e quinta, durante os meses de Janeiro, Fevereiro, Março, Abril e Maio é também predominante as prescrições a utentes femininos de idade superior aos 65 anos, sendo mais notório as consultas de utentes entre os 15 e os 79 anos no mês de Junho. Relativamente aos utentes de idade entre os 0 e os 4 anos, as suas prescrições são mais notórias nos meses de Janeiro, Fevereiro e Março. Ainda nos distritos de Braga, Porto e Bragança, no mês de Dezembro existe um maior número de prescrições relacionadas com os utentes do sexo feminino dos 0 aos 64 anos. Por sua vez, nos distritos de Viana, Aveiro e Vila Real, mais uma vez se destaca a distinção entre os dias da semana. Mais concretamente, as prescrições às quartas-feiras destacam-se nos distritos de Aveiro e Vila Real, comparativamente ao distrito de Viana. Às terças e sextas, o maior número de prescrições recai no grupo etário superior a 45 anos, por sua vez, os utentes dos 15 aos 44 anos recorrem aos centros de saúde maioritariamente às sextas. Por fim, à quinta-feira o número de prescrições é elevado nos distritos de Aveiro e Vila Real, sendo que no distrito de Viana, é predominante os utentes dos 15 aos 79 anos. Tendo em conta as segundas, terças e quintas, o distrito de Viana destaca-se no número de prescrições de utentes entre os 0 e os 14 anos.

Para as mesmas características mencionadas acima (prescrição de medicamentos não genéricos e estações do ano de Inverno e Primavera), os padrões relativos ao sexo masculino não apresentam tanta diversidade, o que significa que existe um comportamento mais uniforme por parte do sexo masculino. Nos distrito de Viana, Braga, Bragança e Porto, e nos meses de Janeiro, Fevereiro, Março, Abril, Maio e Junho, as prescrições relativas ao sexo masculino foram bastantes superiores nos dias de semana úteis para os utentes de idade igual ou superior aos 45 anos. Relativamente aos utentes de idade inferior aos 45 anos, mais propriamente nos meses de Janeiro, Fevereiro e Março, houve uma maior adesão aos serviçoes de saúde por parte dos utentes entre os 15 e os 44 anos. No mês de Dezembro, predominam as prescrições de utentes de idade superior a 45 anos. Nos restantes distritos, Aveiro e Vila Real, as prescrições apresentaram maiores contornos nos meses de Janeiro, Fevereiro, Março, Abril, Maio e Junho para os utentes masculinos de idade superior a 15 anos.

Ainda relativamente à prescrições de medicamentos não genéricos, mas tendo em conta as estações do ano de Verão e Outono, os padrões de prescrições do sexo feminino, diferem essencialmente de acordo com os dias da semana. Às segundas e quartas, os utentes do sexo feminino de idade igual ou superior a 15 anos, recorrem ao serviço de saúde em maior número nos distritos de Bragança, Braga e Porto, enquanto que os restantes distritos, mais propriamente Viana, apresenta a particularidade do número de prescrições ser mais elevado

às quartas-feiras, para a mesma faixa etária. Os utentes dos 0 aos 15 anos recorrem aos serviços de saúde mais na época do Verão, sendo que uma parte satisfatória dos utentes dos 0 aos 4 anos recorrem aos serviços de saúde do Porto no Outono. Nos restantes dias da semana, pode-se dividir a população em dois grupos etários, os de idade superior a 65 anos e os de idade entre os 0 e os 64 anos. Relativamente aos utentes femininos de idade superior a 65 anos, as suas prescrições são mais representativas nos distritos de Porto, Braga, Bragança, no entanto, à terça e à sexta, os centros de saúde e extensões de Viana do Castelo constatam um acréscimo de utentes dos 65 aos 79 anos. Relativamente aos utentes dos 0 aos 65 anos, nos distritos de Braga, Porto e Bragança predominam os utentes dos 15 aos 64 anos, e os utentes dos 0 aos 15 anos recorrem com mais frequência aos serviços de saúde na estação do Verão, tendo em conta os mesmos distritos. No distrito de Viana, existe um acréscimo de prescrições à terça e à sexta-feira dos utentes feminino dos 45 aos 64 anos, no entanto, comparativamente com o distrito de Viana do Castelo, os distritos de Vila Real e Aveiro em conjunto recolhem um maior número de prescrições.

De acordo com as características de prescrições de medicamentos não genéricos e das estações do ano de Verão e Outono, os utentes do sexo masculino, de idade igual ou superior a 45 anos, são predominantes nos distritos de Viana do Castelo, Braga, Bragança e Porto. Relativamente aos mesmos distritos, mas tendo em conta os utentes de idade inferior a 45 anos, na estação de Verão a prescrição é mais elevada, no entanto, relativamente ao Outono existe uma predominância de utentes do sexo masculino, de idade compreendida entre os 15 e os 44 anos. Nos distritos de Aveiro e Vila Real é notório o maior número de prescrições dos utentes de idade igual ou superior a 15 anos.

Concluindo, pode-se assumir que os distritos de Braga, Bragança e Porto apresentam normalmente as mesmas características de prescrições, bem como os distritos de Aveiro, Vila Real e Viana do Castelo que também apresentam características semelhantes entre eles. É perceptível que as mulheres recorrem com mais frequência aos centros de saúde às segundas e às quartas feiras. Por sua vez, os homens não apresentam padrões diferentes relativamente aos dias da semana. Em Dezembro, é normal os serviços de saúde atenderem pessoas mais adultas e idosas comparativamente com as crianças, enquanto que no Verão existe um acréscimo de prescrições a utentes dos 0 aos 15 anos. Associados aos distritos de Braga, Porto e Bragança, estão maioritariamente utentes de idade superior aos 15 anos. Nos distritos de Viana, Vila Real e Aveiro, o número de prescrições de utentes masculinos de idade inferior a 15 anos é bastante baixa, no entanto, nos restantes distritos os números já são relevantes, mas comparativamente com os utentes de idade superior a 15 são muito poucas. Relativamente ao sexo feminino, as prescrições de utentes de idade inferior a 15 anos não apresenta grande relevância em nenhum dos distritos.

3.8.3 Classificação

Na descoberta de padrões através da classificação, foi utilizado árvores de decisões, devido à fácil interpretação do modelo final apresentado e da possível manipulação de uma elevada quantidade de dados (Tabela 8). Tendo em conta a escolha de árvores de decisão, a escolha do algoritmo recaiu no C5.0. No entanto, devido a este algoritmo não vir implementado na ferramenta *RapidMiner*, o algoritmo utilizado foi o C4.5 [Quinlan 1993], que comparativamente, consome mais recursos computacionais e pode apresentar uma árvore um pouco mais complexa. O C4.5 é descendente do CLS (Concept Learning System) e do ID3 [Quinlan 1979], e difere na possibilidade de construir conjuntos de regras que facilitam a compreensão do modelo final, comparativamente com as árvores. O funcionamento do C4.5 é baseado no algoritmo *divide-and-conquer*, que é aplicado ao conjunto de casos iniciais. Na aplicação recursiva do algoritmo, poderão surgir duas situações [Wu et al. 2008]:

- Se todos os casos em causa pertenceram à mesma classe, ou se o conjunto de casos for menor que o número de casos mínimos para dividir uma árvore, então, essa árvore é considerada como sendo uma folha, e é-lhe atribuída a classe mais frequente dos casos em causa.
- Caso contrário, é realizado um teste a todos os atributos, de maneira a descobrir qual o que apresenta maior ganho no caso de uma partição. A partir daí, o atributo que maximizar o ganho é o escolhido como atributo de divisão, que consequentemente irá resultar em novos sub conjuntos, que irão formar novas sub árvores a que lhes será aplicado o algoritmo recursivamente.

Neste caso, a escolha do atributo de divisão é realizada através da heurística *information gain* (Equação (26)). Na secção 2.6.2 é apresentado um exemplo de uma árvore de decisão.

Pré-processamento

Para a aplicação de técnicas de classificação ao modelo de dados não foi realizado nenhum passo de pré-processamento, isto porque, na fase de construção do sistema de *data warehousing*, já foi realizado a descretização dos dados e o tratamento de valores nulos. A única diferença relativamente ao modelo de dados finais foi a escolha dos atributos, os mesmo escolhidos na segmentação com o acréscimo do atributo classe, princípio activo do medicamento. Também nesta análise, devido à memória existente, não foi possível obter os resultados desejados, que seria descobrir os padrões de prescrição, tendo em conta os

diversos medicamentos prescritos, em todos os distritos. Esta impossibilidade deveu-se à gama de valores existente de diferentes medicamentos ser bastante elevada, o que significa que as possíveis classes de previsão da árvore de decisão, também apresentariam uma gama de diferentes valores bastante elevada. E devido a esse problema, não foi possível calcular a qualidade da previsão do modelo, devido ao excesso de memória necessário para o operador "*ModelApplier*". Para contornar este problema, foi gerado diferentes árvores, com os seus respectivos parâmetros de qualidade, cada uma correspondente a cada distrito, mas tendo em conta o atributo dos princípios activos como classe. Esta decisão (princípios activos por distrito) foi tomada principalmente devido à falta de memória, mas também, porque, como cada princípio activo pode estar associado a diferentes medicamentos, de diferentes laboratórios, assim diminui-se o número de possíveis classes, e obtém-se uma análise semelhante à desejada inicialmente.

No cálculo da qualidade de previsão das árvores geradas foram tidos em conta 3 parâmetros [Witten & Frank 2005]. Todos os parâmetros de qualidade utilizados abordam as noções de falsos positivos (false positive), falsos negativos (false negative), acertos (true positive) e erros (true negative). Um acerto é quando os tuplos positivos são correctamente classificados, um erro é quando um tuplo é negativo e é calculado como tal, um falso positivo são os tuplos negativos que foram classificados incorrectamente (como positivos), e um falso negativo são os tuplos positivos que foram classificados incorrectamente (como negativos). Encara-se como negativo e positivo, a classe designada por sim ou não, como por exemplo na Figura 10. A medida de qualidade de previsão de uma árvore de classificação é denominada por *accuracy*, que mede a percentagem de acerto do modelo e é dada pela Equação (34).

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{34}$$

Outra das medidas utilizadas foi a taxa de erro da previsão, dado pela Equação (35), e que é o inverso da *accuracy*. O que quer dizer, que a soma da *accuracy* com a soma da taxa do erro deve dar a percentagem de 100%.

$$error\_rate = \frac{FP + FN}{TP + FP + FN + TN} \tag{35}$$

Por fim, a última medida de avaliação gerada foi a *f-measure*, que foi calculada através dos valores da *precision* e *recall*, fornecidos pela aplicação do operador

"*ClassificationPerformance*" ao modelo final. Consequentemente, para calcular o valor da medida *f-measure*, foi necessário a utilização da Equação (36).

$$f\_measure = \frac{2 \times precision \times recall}{recall + precision} \tag{36}$$

Esta medida varia entre 0 e 1 e valores mais altos sugerem melhores previsões. De notar que ambas as medidas mencionadas acima (*accuracy* e *error rate*) foram calculadas automaticamente através do operador "*ClassificationPerformance*".

Foram apresentados na secção 2.6.3, quatro métodos de avaliação do desempenho de um modelo. Para este estudo em particular foi usado o método *cross validation* (validação cruzada), através do operador "*XValidation*", com um total de 10 iterações (*number_of_validation=10*) e através de amostras aleatórias (*sampling_type=shuffled sampling*). De seguida é apresentado o código XML utilizado na realização deste estudo, contudo, é preciso referir que houveram pequenas modificações para cada distrito nomeadamente na *query* SQL e no número mínimo de instâncias por folha. Essas modificações serão apresentadas e explicadas na altura de avaliação de cada árvore, correspondente a cada distrito (Tabela 27).

```
<operator name="Root" class="Process" expanded="yes">
      <operator name="DatabaseExampleSource" class="DatabaseExampleSource">
          <parameter key="work_on_database"  value="false"/>
          <parameter key="database_system"   value="MySQL"/>
          <parameter key="database_url" value="jdbc:mysql://localhost:3306/joana"/>
          <parameter key="username"   value="joana"/>
          <parameter key="password"   value="cB1pwz1nlIQ="/>
          <parameter key="query"      value="select
cod_presc,distrito,princactiv,generico,periodo,mes,estacao,semana,sexo,idade      from
c45_pricactiv where distrito='Braganca';"/>
          <parameter key="id_attribute"      value="cod_presc"/>
          <parameter key="datamanagement"    value="int_array"/>
      </operator>
<operator name="XValidation" class="XValidation" expanded="yes">
          <parameter key="keep_example_set"  value="false"/>
          <parameter key="create_complete_model"     value="false"/>
          <parameter key="average_performances_only" value="true"/>
          <parameter key="leave_one_out"     value="false"/>
          <parameter key="number_of_validations"     value="10"/>
          <parameter key="sampling_type"     value="shuffled sampling"/>
          <parameter key="local_random_seed" value="-1"/>
          <operator name="W-J48" class="W-J48">
              <parameter key="keep_example_set"      value="false"/>
              <parameter key="U"      value="false"/>
              <parameter key="C"      value="0.5"/>
```

```
            <parameter key="M"      value="500.0"/>
        </operator>
         <operator name="OperatorChain" class="OperatorChain" expanded="yes">
            <operator name="ModelApplier" class="ModelApplier">
                <parameter key="keep_model"        value="false"/>
                <list key="application_parameters">
                </list>
                <parameter key="create_view"       value="false"/>
            </operator>
            <operator                                 name="ClassificationPerformance"
class="ClassificationPerformance">
                <parameter key="keep_example_set"  value="false"/>
                <parameter key="main_criterion"    value="first"/>
                <parameter key="accuracy"  value="true"/>
                <parameter key="classification_error"     value="true"/>
                <parameter key="weighted_mean_recall"     value="true"/>
                <parameter key="weighted_mean_precision"  value="true"/>
                <list key="class_weights">
                </list>
            </operator>
         </operator>
      </operator>
</operator>
```

Relativamente ao código XML, nos operadores "*W-J48*" e "*ClassificationPerformance*", só foram apresentados os parâmetros cujo houve uma alteração aos seus valores por defeito.

Avaliação dos Resultados Obtidos

Como foi mencionado acima, foi gerado uma árvore por cada distrito, e calculado os seus respectivos parâmetros de qualidade, tento em consideração os princípios activos como classe de previsão. Nesta secção, serão apresentados os parâmetros de qualidade obtidos para cada uma das árvores geradas e será feita uma descrição de cada uma. Assim sendo, na Tabela 27, é possível observar os valores obtidos para cada uma das árvores e o respectivo número mínimo de folhas (*M*) utilizado para cada uma das árvores criadas. De resto, todos os parâmetros são idênticos para todas as árvores. Como por exemplo, o parâmetro *C=0.5*, indica a confiança mínima para ser executado o *prunning* à árvore.

Medida	Aveiro	Braga	Bragança	Porto	Viana do Castelo	Vila Real
Nº min de folhas	500	1000	500	1000	500	500
Accuracy	0.0580	0.0754	0.0660	0.0649	0.0658	0.0678
Error rate	0.9412	0.9246	0.9344	0.9351	0.9342	0.9322
Recall	0.0057	0.0042	0.0036	0.0031	0.0044	0.0045
Precision	0.0016	0.0025	0.0016	0.0014	0.0021	0.0024
F-Measure	0.0024	0.0024	0.0022	0.0019	0.0028	0.0031

Tabela 27 – Parâmetros de qualidade de cada árvore gerada e o número mínimo de folhas respectivo.

Como é possível verificar na Tabela 27, a qualidade de previsão dos princípios activos para todos os distritos é bastante baixa. Isto deve-se ao facto de existirem acima de 1000 classes, o que remete para uma gama de possíveis classes bastante elevada. No entanto, mesmo assim, é possível obter-se uma ideia de quais são os tipos de medicamentos mais prescritos por distritos. Será de seguida feito uma descrição para cada uma das árvores geradas, ordenadas, relativamente ao distrito, pela Tabela 27.

1. **Aveiro:** No distrito de Aveiro, as prescrições de medicamentos genéricos diferem essencialmente no sexo do utente. Relativamente ao sexo feminino, aos utentes de idade superior a 45 anos é associado o consumo de medicamentos, cujo o princípio activo é a sinvastatina (reduzir os riscos de doenças cardiovasculares). Também no caso dos utentes do sexo masculino, de idade superior a 15 anos, a prescrição de medicamentos deste tipo é predominante. Para os utentes do sexo feminino de idade entre os 0 e os 4 anos, é normalmente receitado expectorantes (ambroxol), e para os utentes de idade compreendida entre os 5 e os 14 anos, são normalmente receitados medicamentos antialérgicos (cetirizina). Por fim, ainda no sexo feminino, aos utentes dos 15 aos 44 anos, estão associados consumos de medicamentos diferentes, consoantes a estações do ano. Isto é, na altura do Verão e Inverno, são normalmente receitados antidepressivos (fluoxetina), e na altura da Primavera e Outono a prescrição de ansiolíticos (alprazolam) é elevada. Relativamente ao sexo masculino, os utentes dos 5 aos 14 anos consomem, em maiores quantidades, fármacos do grupo dos anti-inflamatórios (ibuprofeno), enquanto que os utentes do grupo etário dos 0 aos 5 anos estão associados ao maior consumo de antibacterianos, mais propriamente penicilina (amoxicilina e ácido clavulânico). Para os medicamentos genéricos, o factor de divisão principal recai no grupo etário dos utentes. Para os utentes de idade compreendida entre 65 e 79 anos, do sexo feminino, é predominante a receita de analgésicos e antipiréticos (ácido acetilsalicílico), com excepção dos meses de Setembro e Outubro, em que o maior número de prescrições recai na vacina da gripe, e no mês de Março, em que é receitado na maioria analgésicos, associados ao princípio activo paracetamol. Relativamente ao sexo masculino, a prescrição é bastante semelhante à prescrição do sexo feminino, diferenciando apenas no

mês de Março, que também é receitado com mais frequência analgésicos e antipiréticos (ácido acetilsalicílico). A prescrição de medicamentos aos utentes de idade entre os 45 e os 64 anos, difere, principalmente, relativamente ao sexo do utentes. Para os utentes do sexo feminino, nos meses de Janeiro, Maio, Julho, Agosto e Setembro, é normalmente receitado medicamentos usados no tratamento das doenças endócrinas (levotiroxina sódica), nos meses de Fevereiro, Abril e Junho, a prescrição recai em ansiolíticos (bromazepam) e nos meses de Março e Dezembro, a prescrição de medicamentos associados ao paracetamol é mais elevada. E como seria de esperar, nos meses de Setembro e Outubro a vacina contra a gripe é o medicamentos mais receitado nesses dois meses. Para os utentes do sexo masculino, existe uma predominância na prescrição de medicamentos usados no tratamento das doenças endócrinas (levotiroxina sódica), nos meses de Janeiro, Fevereiro, Abril, Maio, Agosto, Novembro e Dezembro. No mês de Março a prescrição de antidiabéticos (gliclazida) é mais acentuada comparativamente com os outros medicamentos, o que também é visível nos meses de Junho e Julho, mas neste caso, apesar de o objectivo final ser o mesmo, os princípios activos dos medicamentos diferem, sendo neste caso a metformina. Por fim, nos meses de Setembro e Outubro, é receitado com mais predominância vacinas contra a gripe. Relativamente aos utentes dos 0 aos 4 anos, no sexo masculino é mais receitado medicamentos para acréscimo da vitamina D3 no sangue (colecalciferol), e no sexo feminino a vacina contra a meningite (vacina pneumocócica conjugada) é o "medicamento" mais prescrito. Para os utentes de idade compreendida entre os 5 e os 14 anos a prescrição de medicamentos recai essencialmente em anti-helmínticos (Flubendazol). Relativamente aos utentes dos 15 aos 44 anos, existe uma pequena diferença de prescrições relativamente aos meses do ano. Nos meses de Janeiro, Abril, Maio, Julho, Agosto, Novembro e Dezembro são receitado maioritariamente anticoncepcionais (etinilestradiol e gestodeno), nos meses de Fevereiro, Março, Junho e Setembro a prescrição de analgésicos (paracetamol) é superior, e no mês de Outubro a vacina contra a gripe, é uma vez mais bastante receitada. Por fim, nos utentes de idade superior a 80 anos, nos meses de Janeiro, Março, Abril, Junho, Novembro e Dezembro as prescrições dos medicamentos para o tratamento das doenças endócrinas (levotiroxina sódica) são mais elevadas e nos meses de Setembro e Outubro a vacina da gripe é predominante. Ainda no grupo etário dos utentes superiores a 80 anos, no mês de Fevereiro, para os utentes do sexo feminino é receitado maioritariamente medicamentos associados ao princípio activo levotiroxina sódica, enquanto que aos utentes do sexo masculino é normalmente receitado medicamentos para o tratamento de insuficiência cardíaca (furosemida). Finalmente, nos meses de Maio e Julho, para o sexo feminino é predominante a prescrição de medicamentos para a insuficiência cardíaca, e medicamentos para o tratamento de doenças endocrónicas para o sexo masculino.

2. Braga: Comparativamente com o distrito de Aveiro, Braga apresenta uma prescrição de medicamentos mais uniforme consoante as idades, os meses e o sexo dos utentes. Na prescrição de medicamentos genéricos, aos utentes superiores a 45 anos é normalmente receitado medicamentos de prevenção de doenças cardiovasculares (sinvastatina). Aos utentes dos 15 aos 45 anos, do sexo feminino a prescrição de ansiolíticos é elevada, e nos homens, a prescrição de medicamentos para a prevenção de doenças cardiovasculares é ainda bastante elevada. Por fim, a prescrição de medicamentos contendo penicilina é superior aos outros medicamentos prescritos, para os utentes entre os 0 e os 15 anos. Na prescrição de medicamentos não genéricos, a prescrição de paracetamol está bastante presente nos utentes dos 0 aos 4 anos, com excepção à terça-feira para os utentes do sexo masculino, a que são receitados, maioritariamente, medicamentos para a vitamina D3 (colecalciferol). Também para os utentes dos 5 aos 14 anos, o paracetamol é o tipo de medicamento mais receitado em todos os meses do ano, com excepção para os meses de Junho e Agosto, em que são mais receitados medicamentos anti-helmínticos (albendazol), e no mês de Outubro em que a prescrição da vacina contra a gripe apresenta proporções de prescrições elevadas. Relativamente aos utentes do grupo etários dos 15 aos 44 anos, para o sexo masculino, mais uma vez os analgésicos (paracetamol) são os mais receitados em todos os meses, com excepção dos meses relativos à prescrição da vacina da gripe, Setembro e Outubro. No caso dos utentes do sexo feminino, a prescrição de anticoncepcionais é elevada em todos os dias da semana, sendo, no fim de semana os analgésicos o tipo de medicamentos mais receitados. Aos utentes de idade entre os 45 e os 64 anos, do sexo feminino, é maioritariamente receitado analgésicos (paracetamol), excepto nos meses de Setembro e Outubro em que a prescrição de vacinas contra a gripe dispara. No sexo masculino, também é perceptível o aumento da vacina da gripe nos meses de Setembro e Outubro, no entanto, nos meses de Fevereiro, Março, Abril, Junho, Agosto, Novembro e Dezembro a prescrição de analgésicos e antipiréticos (ácido acetilsalicílico) é elevada, e finalmente, nos meses de Maio e Julho o número de prescrições de antidiabéticos (metformina) superior às restantes rescrições. A prescrição de analgésicos e antipiréticos, é predominante nos utentes de idade compreendida entre os 65 e os 79 anos do sexo masculino, com excepção dos meses de Setembro e Outubro em que a vacina da gripe ultrapassa o número de prescrições comparativamente com os outros medicamentos. Para o sexo feminino, a prescrição de analgésicos e antipiréticos está também acente em alguns meses, como Fevereiro, Abril, Junho, Julho, Agosto e Dezembro. Nos restantes meses, as prescrições são divididas entre a vacina contra a gripe nos meses de Setembro, Outubro e Novembro, e diuréticos (indapamida) nos meses de Janeiro, Março e Maio. Relativamente aos utentes de idade superior a 80 anos, as prescrições são também maioritariamente de analgésicos e antipiréticos, com excepção dos meses de Setembro e Outubro, em que a maioria das

prescrições recai na vacina contra a gripe, e nos meses de Abril, Junho e Julho, é receitado maioritariamente medicamentos para combater a insuficiência cardíaca (furosemida).

3. Bragança: No distrito de Bragança, relativamente à prescrição de medicamentos genéricos, aos utentes dos 0 aos 4 anos é normalmente receitado anti-infecciosos (cefixima), enquanto que aos utentes dos 5 aos 14 anos é normalmente receitado anti-inflamatórios (ibuprofeno). Nos utentes de idade dos 15 aos 44 anos a prescrição de ansiolíticos (alprazolam) é elevada, e nos utentes entre os 45 e os 79 anos, a prescrição de medicamentos é essencialmente associada a medicamentos de redução de doenças cardiovasculares (sinvastatina). Para os utentes de idade superior a 80 anos, ainda no sexo feminino, os medicamentos para o tratamento da úlcera duodenal (omeprazol) são bastante prescritos. A prescrição de medicamentos para o tratamento da úlcera duodenal também é predominante nos utentes do sexo masculino de idade superior a 80 anos, mas também nos utentes de idade compreendida entre os 15 e os 44 anos. Relativamente aos utentes do grupo etário dos 44 aos 79 anos, a prescrição de medicamentos para a redução de doenças cardiovasculares é elevada. Para os utentes entre os 5 e os 14 anos, é maioritariamente receitado antibióticos associados ao princípio activo azitromicina, e por fim, para os utentes dos 0 aos 4 anos a prescrição de anti-inflamatórios é predominante. Na prescrição dos medicamentos não genéricos, também no distrito de Bragança, a prescrição do paracetamol é maioritária nos utentes dos 0 aos 4 anos. Nos utentes do sexo masculino, de idade compreendida entre os 5 e os 14 anos, a maioria das prescrições são relativas a anti-inflamatórios, e no sexo feminino, a maioria das prescrições recai em anti-helmínticos (flubendazol). Relativamente aos utentes do grupo etário dos 15 aos 44 anos, o paracetamol é o tipo de medicamento mais prescrito, no entanto, nos meses de Setembro e Outubro, as prescrições para a vacina contra a gripe são predominantes comparativamente com as prescrições de outros medicamentos. Ainda relativo aos utentes dos 15 aos 44 anos, no mês de Julho, o medicamento mais prescrito está associado a anti-inflamatórios, e no mês de Março à prescrição de ansiolíticos (mexazolam). Nos grupos etários superiores, a prescrição de medicamentos é mais variada, comparativamente com o distrito de Braga. Os utentes dos 44 aos 65 anos, nos meses de Janeiro, Abril e Agosto, estão associados a prescrições de analgésicos, enquanto que nos meses de Setembro e Outubro, as prescrições relativas à vacina da gripe aumentam. No mês de Fevereiro, as prescrições dividem-se entre analgésicos e antidiabéticos, o que é visível também no mês de Dezembro, no entanto, os analgésicos estão associados ao sexo feminino, e os antidiabéticos ao sexo masculino. Os antidiabéticos são também muito receitados em Maio, e em conjunto com analgésicos e anti-inflamatórios (piroxicam) no mês de Julho. No mês de Junho, a prescrição de anti-inflamatórios não hormonais, associados ao produto etoricoxib, são prescritos em maioria, e no mês de Março a prescrição de anti-inflamatórios não esteróide (diclofenac) e de ansiolíticos (lorazepam) é

mais elevada, comparativamente com os outros medicamentos prescritos. Por fim, no mês de Julho os medicamentos prescritos em maioria dividem-se entre analgésicos e anti-inflamatórios (piroxicam) e entre antidiabéticos (gliclazida), e no mês de Novembro os medicamentos mais prescritos são anti-inflamatórios (ibuprofeno) e analgésicos (paracetamol). Relativamente aos utentes dos 65 aos 79 anos, a prescrição de medicamentos também varia bastante consoante os meses em causa. Nos meses de Março e Maio os analgésicos e antipiréticos (ácido acetilsalicílico) são o tipo de medicamentos mais receitados. Nos meses de Setembro e Outubro, mais uma vez, a prescrição das vacinas contra a gripe destacam-se como "medicamentos" mais prescritos. No mês de Dezembro o paracetamol é tipo de medicamento mais receitado, sendo também no meses de Julho e Fevereiro em conjunto com os analgésicos e antipiréticos associados à substância ácido acetilsalicílico, e no mês de Agosto em conjunto com diuréticos (indapamida e furosemida). No mês de Junho os medicamentos associados à indapamida são os mais prescritos, e nos restantes meses, em Abril, a maioria das prescrições divide-se entre anti-inflamatórios não esteróides (diclofenac), antidiabéticos (gliclazida) e entre analgésicos e antipiréticos (ácido acetilsalicílico), e em Novembro, entre antidiabéticos (gliclazida) e entre analgésicos e antipiréticos (ácido acetilsalicílico). As prescrições relativas a utentes de idade superior a 80 anos são mais uniformes, e variam essencialmente entre o paracetamol e ácidos acetilsalicílico. Nos meses de Janeiro, Março, Abril, Maio, Junho, Agosto e Novembro a prescrição de medicamentos para o sexo feminino, recai essencialmente nos analgésicos (paracetamol), e para o sexo masculino a maioria das prescrições estão associadas a analgésicos e antipiréticos (ácido acetilsalicílico). Para os meses de Setembro e Outubro a prescrição de vacinas contra a gripe é predominante, e nos meses de Julho e Dezembro os medicamentos contendo a substância ácido acetilsalicílico são os mais prescritos.

4. Porto: Na prescrição de genéricos, o atributo sexo é o que apresenta maior relevância, e de seguida o atributo grupo etário. Assim sendo, relativamente ao sexo feminino, as prescrições dos utentes dos 0 aos 14 anos, são maioritariamente antibactireanos (amoxicilina e ácido clavulânico), e as prescrições dos utentes dos 15 aos 44 anos dividem-se em dois tipos de medicamentos principais, os ansiolíticos (alprazolam) e os anti-inflamatórios não esteróides (nimesulida). Os medicamentos para a redução de doenças cardiovasculares (sinvastatina) são os mais prescritos para o grupo etário dos 45 aos 79 anos, sendo ainda dos mais prescritos para os utentes de idade superior a 80 anos, principalmente nos meses de Fevereiro, Abril, Maio, Junho, Junho, Agosto, Outubro, Novembro e Dezembro. Nos restantes meses, Janeiro, Março e Setembro os medicamentos mais prescritos são para o tratamento da úlcera duodenal (omeprazol). Para o sexo masculino, as prescrições de genéricos não diferem muito, comparativamente com as prescrições do sexo feminino. Relativamente aos utentes dos 0 aos 4 a prescrição de medicamentos é igual ao sexo

feminino, isto é, existe uma maioria de prescrições de antibacterianos. Aos utentes dos 15 ao 79 anos, é maioritariamente prescrito medicamentos para a redução de doenças cardiovasculares, o que também acontece nos utentes de idade superior a 80 anos, exceptuando nas estações da Primavera e Outono, em que é maioritariamente prescrito medicamentos para o tratamento de úlceras duodenais. Na prescrição de medicamentos não genéricos, para os utentes dos 0 aos 4 anos, o paracetamol é receitado na maioria dos meses, com excepção de Julho e Agosto, em que é mais prescrito a vacina pneumocócia conjugada. Os analgésicos (paracetamol) são também bastante receitados aos utentes de idade compreendida entre os 5 e os 14 anos, no entanto, no meses de Junho, Agosto, Setembro e Outubro a maioria das prescrições de medicamentos recai em anti-helmínticos benzimidazólicos (albendazol). Como se tem observado, os analgésicos são bastantes prescritos na região do Porto, consequentemente, também no grupo etário dos 15 aos 44 anos se verifica uma hegemonia da prescrição de medicamentos associados ao paracetamol. No caso dos utentes do sexo masculino, de idade entre os 15 e os 44 anos, os analgésicos são receitados em praticamente todos os meses, com excepção dos meses coincidentes com a prescrição da vacina contra a gripe, Setembro e Outubro e no mês de Julho, em que existe uma predominância de anti-inflamatórios não-esteróides nas prescrições. Para o sexo feminino, os padrões de prescrição não diferem em nada, com excepção no mês de Julho em que em vez de serem prescritos mais anti-inflamatório não-esteróides, são prescritos em maioria analgésicos e anticoncepcionais. Relativamente aos utentes dos 45 aos 64 anos do sexo feminino, as prescrições são, também, maioritariamente analgésicos, com excepção, também nos meses de Setembro e Outubro, em que o número de prescrições da vacina contra a gripe atinge a maioria das prescrições, e do mês de Junho, em que os analgésicos são o tipo de medicamentos mais prescritos, em conjunto com os anti-inflamatórios não-esteróides. Para o sexo masculino, as prescrições diferem das do sexo feminino, para os utentes dos 45 aos 64 anos. Neste caso, o tipo de medicamentos mais prescritos são os antidiabéticos, nomeadamente nos meses de Março, Junho, Novembro e Dezembro, e também nos meses de Janeiro, Julho e Agosto em conjunto com os analgésicos e antipiréticos (metformina). Por sua vez, os analgésicos e antipiréticos são prescritos em maioria e sozinhos nos meses de Março e Maio. Em comparação com os utentes do sexo feminino, nos meses de Setembro e Outubro, as vacinas contra a gripe são também o maior motivo das prescrições. Para os utentes de idade compreendida entre os 65 e os 79 anos do sexo feminino, as prescrições são maioritariamente relativas a analgésicos, em que apenas nos meses de Março e Julho, os analgésicos são mais prescritos em conjuntos com os ansiolíticos (lorazepam) e com os anti-inflamatórios não-esteróides (diclofenac), respectivamente. Como seria de esperar, nos meses de Setembro e Outubro, a prescrição de vacinas contra a gripe é a principal causa das prescrições. No sexo masculino, aos utentes de idade entre compreendida entre os 65 e os 79, é também, maioritariamente receitado

analgésicos e antipiréticos, no entanto, associados à substância ácido acetilsalicílico. Também neste caso, nos meses de Setembro e Outubro, as prescrições da vacina contra a gripe crescem exponencialmente. As prescrições dos utentes masculinos de idade superior a 80 anos, apresentam exactamente, os mesmo padrões, comparativamente com os utentes masculinos dos 65 aos 79 anos. No caso, dos utentes de idade superior aos 80 anos, mas do sexo feminino, a maioria das prescrições é referente à prescrição de analgésicos, nos meses de Maio, Julho, Agosto, Novembro e Dezembro. Nos meses de Janeiro, Fevereiro, Março e Junho a prescrição de ansilíticos é superior aos outros tipos de medicamentos, e no mês de Março a prescrições de analgésicos e antipiréticos é predominante, em relação à prescrição dos outros medicamentos nesse mesmo mês. Por fim, mais uma vez, a prescrição da vacina contra a gripe destaca-se nas prescrições relativas ao mês de Setembro e Outubro.

5. Viana do Castelo: A prescrição de medicamentos genéricos é bastante similar tendo em conta ambos os sexos. Para o sexo masculino, nos utentes dos 0 aos 14 anos, os antibacterianos são o tipo de medicamentos mais prescritos, enquanto que, para utentes de idade superior a 15 os medicamentos mais receitados são os de prevenção de doenças cardiovasculares. Este tipo de medicamentos é predominante também nos utentes de idade superior a 45 anos no sexo feminino. Aos utentes do sexo feminino, dos 15 aos 44 anos, a maioria das prescrições recai em ansiolíticos (alprazolam), enquanto que aos utentes dos 5 aos 14 é maioritariamente prescrito antbacterianos. Por fim, aos utentes femininos dos 0 aos 4 anos, a maioria das prescrições são relativas a antipsicóticos (risperidona). Tendo em conta as prescrições de medicamentos não genéricos, aos utentes dos 0 aos 4 anos de idade é maioritariamente prescrito analgésicos durante os dias da semana, no entanto, existe também um acréscimo de prescrições da vacina pneumocócica conjugada. Relativamente aos utentes do 5 aos 14 anos, as prescrições de medicamento também se encontram divididas por dias da semana, consequentemente, à segunda, terça e quarta é normalmente mais receitado anti-helmínticos benzimidazólicos, à quinta e sexta a maioria das prescrições são relativas a anti-inflamatórios não esteróides, e ao sábado a prescrição de antiasmáticos é mais elevada. Para os utentes de idade compreendida entre os 15 e os 44 anos do sexo feminino, as prescrições de anticoncepcionais predominam em quase todos os meses, com excepção do mês de Outubro, que apresenta uma maioria de prescrições da vacina contra a gripe, e nos meses de Novembro e Dezembro em que a prescrição de analgésicos é superior aos outros medicamentos prescritos. Para os mesmos utentes, mas de sexo masculino, a prescrição de medicamentos difere consoante os meses. Consequentemente, nos meses de Janeiro, Fevereiro e Abril a maioria das prescrições recai em analgésicos (paracetamol). Os anti-inflamatórios não-esteróides (ibuprofeno) são normalmente mais prescritos nos meses de Março e Dezembro. Nos meses de Junho e Novembro, existe uma maioria de prescrições de antidiabéticos, enquanto que nos meses de Setembro e Outubro as vacinas contra a gripe

são o "medicamento" mais prescrito. No mês de Julho a prescrição de insulina para diabetes (perindropil) é o tipo de medicamentos mais prescrito, enquanto que no mês de Agosto a maioria das prescrições recai em anti-epilépticos e anticonvulsivantes (ácido valpróico). E por fim, no mês de Dezembro a prescrição de medicamentos utilizados na hipertensão, insuficiência cardíaca crónica e angina pectoris (bisoprolol), foi superior aos restantes medicamentos prescritos. Relativamente aos utentes de idade compreendida entre os 45 e os 64 anos, do sexo feminino, o paracetamol é tipo de medicamento mais receitado, com excecpção da vacina contra a gripe nos meses de Setembro e Outubro. Para os utentes do sexo masculino, o tipo de medicamento mais prescrito são os antidiabéticos (metformina), exceptuando nos meses relacionados com a prescrição da vacina contra a gripe, em Setembro e Outubro, e no mês de Julho em que a maioria das prescrições estão associadas a anti-inflamatórios não-esteróides. No caso dos utentes de idade entre os 65 e os 79, a prescrição de medicamentos varia bem mais consoante os diferentes meses, comparativamente com os utentes dos 45 aos 64 anos. Relativamente aos do sexo feminino, nos meses de Janeiro, Fevereiro e Março, o número de prescrições de analgésicos é superior aos restantes, e nos meses de Setembro e Outubro a prescrição da vacina contra a gripe é mais uma vez superior ao resto das prescrições. Os anti-inflamatórios não-esteróides são prescritos em maioria nos meses de Novembro e Dezembro, e em conjunto com antidiabéticos e diuréticos (indapamida) nos meses de Junho e Agosto, respectivamente. Os antidiabéticos são ainda prescritos em maioria no mês de Julho e em conjunto com os analgésicos no mês de Maio. Por fim, no mês de Março, os analgésicos e os diuréticos são prescritos em maior quantidade do que qualquer outro tipo de medicamento. Para o mesmo grupo etário, mas do sexo masculino, a prescrição da vacina contra a gripe é prescrita nos meses e Setembro e Outubro em maioria, sendo que nos meses de Janeiro, Julho e Agosto os antidiabéticos são bastante prescritos. Nos meses de Março, Junho, Novembro e Dezembro, os anti-inflamatórios não-esteróides apresentam uma predominância nas prescrição desses meses. Nos restantes meses, existe uma conjugação dos anti-inflamatórios não-esteróides com os antidiabéticos, relativamente ao tipo de medicamentos mais prescritos. Por fim, os utentes de idade superior a 80 anos estão associados maioritariamente associados a diuréticos, no entanto, nos meses de Junho e Dezembro, o número de prescrições é repartido por digitálicos (digoxina) e por anti-inflamatórios não-esteróides, respectivamente. No mês de Novembro existe uma maioria de prescrições de anti-inflamatórios não-esteróides, enquanto que nos meses de Setembro e Outubro a vacina contra a gripe associado ao maior número de prescrições desses dois meses. Por fim, no mês de Fevereiro, aos utentes do sexo feminino de idade superior a 80 anos, são essencialmente prescritos analgésicos, enquanto que aos utentes do sexo masculino, são essencialmente prescritos diuréticos.

6. Vila Real: No distrito de Vila Real a prescrição de medicamentos genéricos varia mais comparativamente com os outros distrito, seja relativamente ao sexo do utente ou à idade. Para os utentes do sexo masculino, de idade entre os 0 e os 4 anos, é receitado em maioria expectorantes mucolíticos (ambroxol), para os utentes do grupo etário seguinte, dos 5 aos 14 anos, são maioritariamente receitados antipsicóticos atípicos (risperidona). Os medicamentos de tratamento de úlceras pépticas (lansoprazol) são normalmente receitados a utentes dos 15 aos 44 anos. Por fim, os medicamentos de redução dos riscos de doenças cardiovasculares e os medicamentos usados na retenção urinária (tansulosina), são os medicamentos mais receitados aos grupos etários dos 45 aos 79 anos e superior aos 80 anos, respectivamente. Para os utentes do sexo feminino, dos 0 aos 4 anos, os anti-histamínicos são o tipo de medicamentos mais receitados, enquanto que aos utentes dos 5 aos 14 anos, são normalmente mais receitados antibacterianos. A prescrição de medicamentos genéricos aos utentes femininos dos 15 aos 44 anos, difere consoante as estações do ano. Consequentemente, no Inverno e no Verão são prescritos um maior número de ansiolíticos, e nas estações do Outono e Primavera, os medicamentos para tratamento de úlceras pépticas estão em maioria. Por fim, aos utentes de idade superior a 45 anos do sexo feminino, são maioritariamente receitados medicamentos para a redução dos riscos de doenças cardiovasculares. Na prescrição dos medicamentos não genéricos, aos utentes dos 0 aos 4 anos, a prescrição de vitamina C (ácido ascórbico) é bastante usual, nomeadamente, nos dias da semana de terça, sexta e sábado. À segunda existe uma predominância do número de prescrições de vitaminda D3, sendo que à quarta os analgésicos apresentam o maior número de prescrições. A prescrição de anti-helmínticos benzimidazólicos é predominante nas prescrições da quinta-feira. Relativamente aos utentes dos 5 aos 14 anos, os do sexo feminino estão essencialmente associados à prescrição de paracetamol, enquanto que os do sexo masculino, estão associados à prescrição de anti-helmínticos benzimidazólicos. O paracetamol está bastante presente nas prescrições dos utentes dos 15 aos 44 anos, nomeadamente, nos meses de Janeiro, Fevereiro, Março, Maio, Julho, Agosto, Novembro e Dezembro. Nos restantes meses, a prescrição da vacina da gripe é associada aos meses de Setembro e Outubro, e por fim os anti-inflamatórios não-esteróides (ibuprofeno) são prescritos em maioria, nos meses de Março e Junho. Os anti-inflamatórios não-esteróides, nomeadamente os contendo a substância ácido acetilsalicílico, apresentam a maioria das prescrições dos utentes masculinos, dos 45 aos 64 anos, com excepção de medicamentos de controle do ácido úrico (alopurinol) no mês de Março, e das vacinas contra a gripe, nos meses de Setembro e Outubro. Para o sexo feminino, a prescrição de analgésicos contendo a substância paracetamol, apresenta-se em maioria para os meses de Janeiro, Fevereiro, Março, Maio, Junho, Julho, Novembro e Dezembro, sendo que nos meses de Abril e Junho, a maioria das prescrições é repartida com os anti-inflamatórios não-esteróides. Nos restantes meses, Setembro e Outubro, mais uma vez se verifica a hegemonia

das prescrições da vacina contra a gripe. Relativamente às prescrições de medicamentes aos utentes do sexo masculino, de idade compreendida entre os 65 e os 79 anos, os anti-inflamatórios não-esteróides são predominantes em quase todos os meses, com excepção dos meses de Setembro, Outubro e Novembro, em que maioria das prescrições está associada à vacina contra a gripe. Também no sexo feminino, os anti-inflamatórios não-esteróides (ácido acetilsalicílico) apresentam uma parte substancial das prescrições, no entanto, só nos meses de Julho e Agosto é que este tipo de medicamentos são maioritariamente prescritos. Isto quer dizer, que em alguns meses, os anti-inflamatórios não-esteróides, são prescritos em associação com outro tipo de medicamentos. No mês de Janeiro, Março e Abril, são prescritos em associação com diuréticos, que também são prescritos maioritariamente nos meses de Fevereiro e Maio. No mês de Junho as prescrições são essencialmente relativas também a anti-inflamatórios não-esteróides, mas neste caso esses anti-inflamatórios dividem-se entre a substância ácido acetilsalicílico e diclofenac. No mês de Dezembro as prescrições dividem-se entre os anti-inflamatórios não-esteróides e os analgésicos. Finalmente nos meses de Setembro, Outubro e Novembro, a vacina contra a gripe é responsável pela maior parte das prescrições, no entanto, no mês de Novembro existe também um grande número de prescrições de medicamentos contendo a substância ácido acetilsalicílico. Por fim, os utentes de idade superior aos 80 anos, apresentam um número de prescrições de vacinas contra a gripe bastante superior ao resto dos medicamentos, nos meses de Setembro, Outubro e Novembro. Nos restantes meses, os anti-inflamatórios não-esteróides são prescritos em quase maioria dos meses, com excepção nos meses de Fevereiro, Abril e Maio. No mês de Fevereiro e Abril, os anti-inflamatórios não-esteróides são associados aos utentes do sexo feminino, enquanto que ao sexo masculino estão associados os diuréticos e os anti-anginosos (trimetazidina). Finalmente, no mês de Maio, relativamente ao sexo feminino, o maior número de prescrições é relativa aos anti-anginosos, e associado ao sexo masculino encontram-se, mais uma vez, os anti-inflamatórios não-esteróides.

Após a análise da prescrição de medicamentos para cada um dos distritos, é possível destacar alguns padrões existentes para todos os distritos. Relativamente à prescrição de medicamentos de genéricos, independentemente do sexo do utente, aos utentes de idade superior 65 anos, é maioritariamente receitado medicamentos associados à substância sinvastatina. Para os utentes de idade inferior a 15 anos, a prescrição de antibacterianos e anti-inflamatórios apresenta uma grande percentagem de prescrições neste grupo etário. Relativamente à prescrição de medicamentos não genéricos, é indiscutível a prescrição da vacina contra a gripe nos meses de Setembro e Outubro, para utentes de idade superior a 15 anos. Destaca-se ainda que as prescrições de medicamentos contendo a substância paracetamol, estão normalmente associadas ao sexo feminino, sendo que as prescrições de

medicamentos contendo a substância ácido acetilsalicílico são maioritárias nos utentes do sexo masculino. Associado às mulheres do grupo etário dos 15 aos 44 anos, existe também um grande número de prescrições de anticoncepcionais. Nos utentes dos 0 aos 4 anos, a prescrição da vacina pneumocócia conjugada, é associado a diversos distritos, como sendo a prescrição mais realizada naquele grupo etário, em conjunto com os analgésicos (paracetamol) e com os antibactireanos (amoxicilina e ácido clavulânico). Nos utentes dos 5 aos 14 anos, a prescrição de analgésicos (paracetamol), de anti-helmínticos benzimidazólicos (albendazol) e de anti-inflamatórios não-esteróides (ibuprofeno), estão em maioria comparativamente com os restantes medicamentos prescritos. Algo que é possível constatar é a maior uniformidade da prescrição de medicamentos no distrito do Porto, Braga e também Aveiro, comparativamente com os restantes distritos de Bragança, Viana do Castelo e Vila Real.

3.9 Análise Geral da Aplicação das Técnicas de Mineração de Dados

De acordo com os objectivos propostos pela ARSN, foram elaborados possíveis estudos com a finalidade de tirar partido da máxima funcionalidade de cada uma das técnicas. Ao modelo de dados final foram aplicadas três técnicas de minerção de dados, a associação, a segmentação e a classificação, sendo que cada aplicação de cada diferente técnica é relativa a um estudo diferente, no entanto todos eles se enquadram na perspectiva da descoberta de padrões.

No caso da associação, a aplicação do algoritmo *FP-growth* serviu essencialmente para descobrir regras de associações entre os dados, condicionadas pela escolha dos atributos escolhidos no estudo. Pelo o facto de se ter lidado com uma quantidade de dados bastante significativa, foi necessário utilizar um suporte mínimo relativamente baixo, de maneira a se conseguir gerar determinadas regras. As regras apresentadas apresentam alguma informação relevante, contudo, seria bastante interessante extender este estudo ao conjunto de dados inicial, os 22 milhões de registos. O algoritmo utilizado na aplicação da segmentação foi o *k-means*. O objectivo da aplicação deste algoritmo ao modelo de dados consistiu na descoberta de padrões de prescrições nos dados. Neste caso, os resultados foram obtidos tendo em consideração as possíveis limitações deste algoritmo como a escolha do *k* e a escolha dos *k* centróides iniciais. Consequentemente, os segmentos finais obtidos, foram os que melhor caracterizam o comportamente dos dados, e no final podê-se concluir algumas considerações interessantes, relativamente ao objectivo do estudo em causa. Na classificação, a escolha do algoritmo recaiu obrigatoriamente no C4.5, dentro da escolha de árvores de classificação, devido às limitações da ferramenta de mineração de dados escolhida. Para este estudo o

objectivo final consistiu na descoberta de padrões de prescrição de medicamentos. Isto é, consoante as características, sejam elas temporais ou humanas, qual o tipo de medicamento prescrito que mais se adequa a tais características.

Apesar de terem sido elaborados estudos diferentes para cada uma das técnicas, é constatável, que a associação foi a que concluiu melhores resultados tendo em conta o objectivo final desejado. Como também é evidente, que apesar de se poder tirar conclusões acerca do resultado devolvido pela aplicação da classificação, não se pode afirmar essas conclusões com determinado grau de certeza, uma vez que a percentagem de qualidade do modelo é demasiadamente baixa. No caso da segmentação não se pode determinar a qualidade do modelo, uma vez, que a segmentação é uma técnica exploratória, em que o seu objectivo não é prever classes, mas sim explorar os dados. Nesse aspecto, foi possível obter algumas conclusões acerca dos dados, no entanto, mais uma vez, não é possível afirmar tais resultados com elevado grau de certeza.

Concluindo, após apresentar os objectivos de cada uma das técnicas neste estudo, e avaliar os resultados obtidos, é possível afirmar, que através da aplicação da associação ao modelo de dados em causa, consegue-se obter resultados bastantes satisfatórios. Como tal, seria interessante explorar mais possíveis estudos acerca dos dados em causa, através da técnica da associação e outros possíveis algortimos, como o *Apriori*, ou o Eclat que apesar de apresentar características semelhantes ao FP-growth, é um algoritmo baseado em DFS e intersecção de lista de transacões.

3.10 Software Utilizado

Para este trabalho foram utilizadas, maioritariamente, ferramentas *open-source*, com excepção do editor de texto e a ferramenta usada na exploração gráfica. Relativamente às ferramentas *open-source*, foi utilizado para a gestão de base de dados o *MySQL* versão 5.1, bem como todas as ferramentas gráficas associadas, como *MySQL Query Browser* (realização de *queries* e pesquisa sobre os dados), *MySQL Administrator* (administração da base de dados e execução de *backups*) e *MySQL Workbench* (desenho físico do modelo). A opção da utilização das ferramentas disponibilizadas pelo *MySQL* recaiu, essencialmente, no facto de serem gratuitas e apesar disso apresentar emuma resposta bastante boa em diversas situações. O único problema encontrado neste sistema foi o excessivo tempo de execução de cursores. Para a aplicação das técnicas de mineração de dados foi utilizada a ferramenta *RapidMiner* versão 4.5, devido a ser *open-source* e comparativamente com o *weka* apresentar muitas mais soluções de pré-processamento e não só. Esta ferramenta é

totalmente implementada em *java*, e uma das suas principais vantagens é mesmo o número de operadores implementados, o que permite uma grande variedade de possíveis soluções para os diversos problemas que poderão ocorrer. No entanto, esta ferramenta apresenta uma grande desvantagem relativamente ao *weka* (por exemplo), o consumo excessivo de memória, por parte de alguns operadores, o que se torna impraticável em máquinas mais usuais, obrigando a diferentes abordagens dos problemas.

As duas únicas ferramentas utilizadas sem serem open-source, foram as ferramentas *Office* da *Microsoft*, mais concretamente o *Microsoft Word* e *Excel* 2008. O *Microsoft Word* foi utilizado como editor de texto, enquanto que o *Microsoft Excel* foi utilizado para a exploração gráfica realizada aquando a análise exploratória.

Capítulo 4

Conclusões e Trabalho Futuro

4.1 Conclusões

A prescrição electrónica é cada vez mais uma realidade na sociedade de hoje, o que provoca o armazenamento de elevadas quantidades de dados, que futuramente poderão ser exploradas e analisadas. A informatização dos serviços de saúde, através da aplicação de técnicas de mineração de dados, permite a obtenção de informação bastante valiosa e útil para a tomada de decisão. A descoberta de padrões de prescrição permite uma gestão melhorada de recursos, tanto humanos como técnicos, o que permite uma redução significativa dos gastos em recursos excessivos. A análise dos dados relativos à prescrição de medicamentos possibilita, ainda, a possível descoberta de irregularidades nas prescrições dos médicos e possíveis associações entre laboratórios.

Neste trabalho foram descritos e analisados os diferentes tipos de padrões – por exemplo, frequentes, frequentes sequenciais e episódios frequentes. A descoberta de padrões é essencialmente realizada através de três técnicas de mineração de dados: a associação, a segmentação e a classificação. Para cada uma das técnicas mencionadas foi realizada uma revisão bibliográfica, tendo em consideração a possível ajuda na descoberta de padrões. A revisão bibliográfica consistiu num estudo pormenorizado dos algoritmos mais utilizados para cada uma das técnicas, sendo no final, apresentado diferentes casos de estudo.

A aplicação da associação a um modelo de dados, permite descobrir possíveis relações existentes nos dados, consoante os parâmetros mínimos de avaliação impostos. Alguns desses parâmetros, denominados como medidas de avaliação de regras, são apresentados, e

explorados o seu significado na avaliação de uma regra. No entanto, as medidas de avaliações de regras em nada servem sem a inicial descoberta de regras de associação. Consequentemente, relativamente à técnica da associação, foram descritos quatro tipos de possíveis funcionamento de algoritmos na descoberta de regras de associação (*Breadth-First Search* e contagem de ocorrências, *Breadth-First Search* e intersecções TID-*List*, *Depth-First Search* e contagem de ocorrências, e por fim *Depth-First Search* e intersecções TID-*List*). De seguida, foram apresentados alguns casos de estudo realizados através da associação. A segunda técnica revista foi a segmentação. Esta técnica permite descobrir padrões entre os dados, através do seu agrupamento em segmentos que contenham as mesmas características, ou seja, as mesmas similaridades. A similaridade entre os dados pode ser calculada através de diversas medidas, como a medida euclidiana. Esta e outras medidas foram apresentadas na secção da segmentação. Existem duas técnicas principais na segmentação, a *partitional* e a *hierarchical* que foram abordadas neste trabalho. No entanto, foram também abordadas mais três técnicas, baseadas em estruturas e casos mais específicos: *density based*, *grid based* e *model based*. Também, após a revisão bibliográfica da segmentação, foram apresentados diferentes casos de estudo, que através da aplicação de algoritmos da segmentação permitiram fornecer informação bastante útil e interessante. Por fim, na revisão bibliográfica relativa à classificação, foram apresentados, inicialmente, métodos de pré-processamento de dados, que se deve ter em conta antes da aplicação dos algoritmos ao modelo de dados. Para esta técnica, foram analisadas quatro diferentes técnicas, *decision trees*, *bayesian classifiers*, *lazy learnes* e *neural networks*. A utilização de cada uma delas remete para um modelo final resultante que se ajusta ao dados. No entanto, é necessário avaliar a qualidade de ajuste aos dados, isto é, se o modelo final consegue transpor com realidade e segurança as características existentes nos dados. Consequentemente, foram abordados quatro métodos de avaliação de modelos, que permitem calcular o desempenho de previsão de um modelo relativo ao conjunto de treino. Os quatro métodos apresentados foram *houldout method*, *random subsampling*, *cross validation* e *bootstrap*. No final da secção relativa à classificação, foram mencionados e analisados alguns casos de estudo, tendo em conta a aplicação de diferentes técnicas da classificação.

A revisão bibliográfica foi terminada com uma pequena revisão para cada uma das técnicas exploradas, que teve como objectivo final, essencialmente, ajudar na escolha do melhor algoritmo, consoante as características dos dados, o objectivo final e os custos computacionais necessários. Numa segunda fase deste trabalho, foi realizado a análise sobre um modelo de dados através da aplicação das técnicas mencionadas acima. Ao modelo de dados inicial, fornecido pela Associação Regional da Saúde do Norte (ARNS), referente às prescrições de medicamentos durante o ano de 2008, foram aplicados diversos passos de

pré-processamento aos dados. Numa fase inicial, foi delimitado o conjunto de dados, tendo em conta a análise de apenas alguns centros de saúde de alguns distritos, e foi realizada uma descrição dos dados, mais concretamente, a definição do seu significado, da sua cardinalidade, etc. Com base nesse subconjunto, foi realizado uma análise exploratória, com o objectivo de descobrir possíveis relações entre os dados e ter uma ideia superficial da qualidade dos dados em termos técnicos e informacionais. No final da análise exploratória, chegou-se à conclusão de que os dados apresentavam uma qualidade consideravelmente boa, uma vez que os dados apresentam-se bastante completos, existindo apenas alguns atributos com alguns valores nulos que impossibilitavam qualquer tipo de análise. Logo, para cada um desses casos foram tomadas decisões de preservação da qualidade do modelo final, quer fossem elas de eliminação de registos ou não inclusão de determinados atributos na análise final. Após a selecção e limpeza de dados foi realizada a construção, integração e formatação de dados, que consistiu, essencialmente, na construção de um sistema de *data warehousing* através de processos de povoamento e criação de novos novos eixos de análise. Para algumas dimensões foram criados novos atributos, tendo em conta os interesses de análise da ARSN e discretização dos dados, como foram os casos dos atributos mês e estação do ano, entre outros. A criação do sistema de *data warehousing*, permitiu implementar uma estrutura que melhor respondesse às necessidades impostas pela a utilização de técnicas de mineração de dados, e às necessidades de variados eixos de análise.

Na fase seguinte à implementação do sistema de *data warehousing*, foram aplicadas à estrutura criada as três técnicas expostas na revisão bibliográfica apresentada neste trabalho. A aplicação de cada técnica ao modelo de dados final, foi abordada em separado. No entanto, a abordagem ao problema foi realizada de forma idêntica nas três. Isto é, para cada uma das aplicações, foi mencionado e explicado o funcionamento do algoritmo a utilizar, os parâmetros de avaliação dos resultados finais, bem como os passos de pré-processamento, caso necessário. Todos esses processamentos, foram executados através dos operadores existentes no *RapidMiner*, uma ferramenta *open-source* totalmente implementada em *java*.

Após a obtenção e explicação dos resultados finais, foi apresentada uma breve conclusão acerca dos resultados obtidos. De notar que, para cada uma das técnicas utilizadas, a escolha dos atributos em estudo variou consoante as necessidades do trabalho em causa. No caso da associação, o algoritmo utilizado foi o *FP-growth*, devido às suas boas características para manusear grandes quantidades de dados a baixos custos computacionais. Para este algoritmo foi apresentada uma explicação mais pormenorizada sobre o seu funcionamento e, só depois, foram descritos os parâmetros de qualidade e os passos de pré-processamento. Para a técnica de associação foram elaborados dois estudos. O primeiro teve como objectivo descobrir possíveis associações entre os medicamentos prescritos, e o segundo, descobrir se

existe alguma associação entre as prescrições dos médicos e os laboratórios prescritos. Para o primeiro estudo, foi necessário descobrir as associações dos medicamentos para cada distrito em separado, devido à memória existente ser insuficiente para trabalhar os dados todos em conjunto. Em suma, foram obtidas regras de associação, tendo em conta o parâmetro *lift*, para cada um dos distritos. No segundo estudo, foram geradas regras de associação entre os médicos prescritores e os laboratórios prescritos, tendo em conta o parâmetro de avaliação confiança. Neste caso foram geradas as regras para todos os distritos em conjunto, e foram obtidas algumas regras que talvez sejam merecedoras de alguma atenção.

Para o caso da segmentação foi utilizado o algoritmo *k-means*. Como este algoritmo já tinha sido, previamente, explicado de uma forma clara, nesta fase apenas foi feita uma descrição sucinta acerca do *k-means*. No caso da segmentação foram usados determinados passos de pré-processamento, como a passagem dos dados de nominais para numéricos e de seguida fo realizada a normalização dos valores. A utilização do algoritmo *k-means* está associado a diversos parâmetros, que conforme os seus valores poderão indicar um modelo de melhor ou pior qualidade. Consequentemente, foi necessário avaliar alguns parâmetros que indicam a qualidade dos segmentos finais gerados e foi apresentada uma breve descrição de cada um e do seu significado na avaliação dos segmentos finais. Após a aplicação do algoritmo *k-means*, foi aplicado o algoritmo C4.5 ao conjunto de segmentos resultantes, com o intuito de se obter um modelo de análise mais facilitado. Isto é, a análise dos resultados através de uma árvore de decisão é mais compreensível do que a análise registo a registo, de acordo com os diferentes segmentos. No final de todo o processo, foi possível retirar algumas conclusões acerca de alguns padrões de prescrição. Mais concretamente, foi possível verificar uma diferença de padrões entre os utentes do sexo masculino e os utentes do sexo feminino, e também entre os diferentes grupos etários. É também notável um conjunto de características idênticas nos padrões entre os distritos de Braga, Porto e Bragança, e entre os distritos de Viana do Castelo, Aveiro e Vila Real.

Por fim, a aplicação da classificação ao modelo de dados foi realizada através do algoritmo C4.5. Como também neste caso o funcionamento das árvores de decisão tinham sido previamente explicados, apenas foi realizado uma breve descrição do seu funcionamento e especificada qual a medida de selecção de atributo neste estudo. O objectivo da classificação neste estudo foi descobrir possíveis padrões de prescrição de tipos de medicamentos, isto é, descobrir, por exemplo, quais os tipos de medicamentos que são normalmente prescritos a utentes com idades superiores a 80 anos. Neste caso, devido ao pré-processamento realizado anteriormente, não foi necessário aplicar qualquer tipo de pré-processamento aos dados. Logo, apenas foi aplicado o algoritmo C4.5 com o método de avaliação 10 *cross validation*,

para cada um dos distritos em separado, mais uma vez devido ao excesso de memória necessária. Apesar dos resultados de *accuracy* apresentados serem bastante baixos, foi possível retirar já bastante informação relativa aos padrões de prescrições de medicamentos, para cada um dos distritos. No entanto, existem diversos medicamentos de laboratórios diferentes associados ao mesmo princípio activo, como tal o estudo não foi relativo aos medicamentos em si, mas sim às substâncias associadas. Com isso conseguiu-se obter resultados um pouco mais concretos e acertados. Concluindo, foram descobertos alguns padrões de prescrições de medicamentos, relativamente aos distritos, sexo dos utentes, grupo etário, entre outros. Por exemplo, a prescrição nos distritos de Braga e Porto é bastante mais uniforme do que nos restantes distritos. A receita de medicamentos de paracetamol é bastante elevada para quase todas as faixas etárias, mas mais concretamente no sexo feminino e a prescrição de medicamentos de prevenção de doenças cardiovasculares, é bastante comum nos utentes de idade superior a 65 anos e em especial nos homens. A vacina contra a gripe é quase sempre associada ao maior número de prescrições nos meses de Setembro e Outubro para os utentes de idade superior a 45 anos.

A escolha de cada algoritmo, utilizado na aplicação de cada uma das técnicas ao modelo de dados, é explicada na respectiva secção de cada uma das aplicações das técnicas no capítulo 3. No entanto, no capítulo 2 ficou claro algumas limitações dos algoritmos utilizados em cada uma das técnicas. Para o caso da associação, foi utilizado o algoritmo *FP-growth*, mas seria bastante interessante, observar os resultados obtidos através da aplicação de algoritmos do tipo BFS com contagem de ocorrências, mais concretamente, o DIC (uma extensão do *Apriori*). Para o caso da aplicação da associação, os parâmetros de qualidade de regras foram escolhidos consoante os estudos em causa. Por exemplo, no primeiro estudo, caso se escolhe-se a confiança como parâmetro de qualidade das regras, estaria-se a ignorar algumas regras interessantes que demonstram a associação entre medicamentos diferentes. O mesmo se pode dizer relativamente ao segundo estudo e o parâmetro de qualidade *lift*, em que se esse fosse mesmo o parâmetro de qualidade, impondo o *lift* a 50 (como no primeiro estudo), não apareceriam regras nenhumas. Concluindo, os parâmetros de avaliação de regras devem ser escolhidos consoante o objectivo final do estudo realizado. O suporte mínimo para ambos os casos foi imposto a 0.001, devido à elevada quantidade de dados, no entanto, diminuindo ainda mais esse limite, provavelmente o número de regras aumentaria, mas também os custos computacionais.

A segmentação é uma técnica de aprendizagem não supervisionada, e consequentemente o seu objectivo é realizar uma análise dos dados. Neste caso, foi aplicado o algoritmo *k-means* que é um algoritmo particional, em que é necessário a escolha do número de segmentos. A escolha deste factor é preponderante na qualidade dos segmentos finais, conciliado com a

escolha aleatória dos *k* centroides iniciais. De maneira, a se conseguir os melhores resultados possíveis foi imposto um número de 500 de iterações máximas de passos de optimização, com o objectivo de se atingir a convergência do algoritmo, e um número de 50 execuções do algoritmo, com o intuito de se conseguir obter os melhores centróides iniciais possíveis. Obviamente com o aumento destes parâmetros obtêm-se melhores valores para o parâmetro *Davies Bouldin*, no entanto, o tempo de processamente dispara. Através desse parâmetro (DB), foi calculado qual o melhor *k* para o conjunto de dados em estudo. Outro dado interessante, que permitiu a obtenção de resultados bastantes melhores do parâmetro DB, foi a normalização dos dados, o que significa, que diminuindo a gama de valores dos dados consegue-se obter segmentos mais compactos e distantes uns dos outros. Existem dois casos que seriam interessantes explorar no caso da segmentação, a aplicação de outro algoritmo particional (os hierárquicos de certeza que resultariam em resultados piores devido à impossibilidade de recolocação de centróides, apesar de apresentarem uma complexidade bastate inferior) como o *k-medoids*, que não apresenta a sensibilidade do k-means relativamente a outliers, e a utiliazação de outra medida de similaridade, em vez da euclideana. Apesar de apresentar algumas características relativamente ao *k-means*, não acho que grandes diferentes seriam apresentadas, devido à inexistência de outliers nos dados em estudo. Relativamente à medida de similaridade, de certeza que os resultados seriam diferentes, pois como se viu, a medida entre dois pontos difere consoante a medida de similaridade em estudo, consequentemente, seria interessante realizar os mesmos estudos com outras medidas e avaliar os resultados finais.

Como foi mencionado, o uso de árvores de decisão na aplicação da classificação ao modelo de dados deveu-se, essencialmente, pela faclidade de análise dos resultados obtidos. No entanto, a excessiva complexidade deste tipo de algoritmos impossibilitou realizar o estudo para todos os distritos em conjunto. Consequentemente, teve que se separar o estudos por distritos, e mesmo assim foi necessário ter atenção ao número mínimo de folhas devido à memória existente. Essa limitação obviamente condicionou os resultados finais, que poderiam ser melhores caso o número mínimo de folhas pudesse ser aumentado. No entanto, seria necessário ter atenção à árvore resultante para não se obter uma árvore sobre-ajustada. Outra grande limitação na aplicação desta técnica, foi a realização do *pre-prunning* em vez de *post-prunning*, também devido à memoria existente. Caso fosse possível, a realização do *post-prunning*, proporcionaria, claramente, resultados um pouco melhores, mas mesmo assim, desconfio que a *accuracy* das árvores não atingiria valores superiores a 25%, muito devido às diferentes características existentes nos dados para os diferentes distritos. Neste caso, a normalização não foi possível, devido ao algoritmo utilizado só permitir que a classe seja um atributo nominal, como tal não se pode afirmar que a normalização fosse uma solução ao problema da baixa qualidade de previsão. No caso da aplicação desta técnica,

seria bastante interessante realizar o mesmo estudo mas aplicando o algoritmo *naive bayes*, e comparar os resultados com os obtidos através do C4.5. Supostamente os resultados obtidos seriam melhores no caso do *naive bayes*, pelo menos é um algoritmo reconhecido por isso comparativamente com o C4.5, no entanto a análise dos resultados necessitaria de bastante mais tempo e atenção. Será certamente um estudo a ter bastante em conta numa análise futura.

4.2 Trabalho Futuro

Apesar dos resultados obtidos através dos três estudos realizados já indicarem tendências interessantes e, em alguns casos, fornecerem mesmo informação útil na gestão de recursos dos centros de saúde, pensa-se, nesta altura, que a realização de um estudo mais aprofundado, e com mais algum tempo, permitiria retirar bastante mais informação. Por exemplo, relativamente à classificação, poderia ser realizado um outro estudo, tendo em conta todos os distritos, mas mais pormenorizado, em que o número de casos por folha fosse menos limitado, mas tendo sempre em conta o sobre ajustamento dos dados. Seria interessante também fazer um estudo com todos os distritos, mas fazendo a sua separação por grupos etários e analisar as prescrições por parte de cada um desses grupos. Evidentemente que, nesse estudo, ter-se-ia uma maior informação relativamente aos medicamentos prescritos naquele grupo etário, consoante cada um dos distritos. Seguindo a mesma ideia, seria possível realizar estudos mais aprofundados, tendo como base alguns dos atributos existentes, consoante a necessidade de informação necessitada. Por exemplo, poderia-se saber quais os padrões de prescrição, mais concretos, relativamente a determinados meses ou estações, entre outros possíveis eixos de análise. Caso fosse possível, seria bastante interessante também, realizar um estudo tendo em conta os dados dos ICPC2s envolvidos no diagnóstico. Talvez estes dados relativos aos ano de 2009 já se apresentem mais completos, de tal maneira, que seja possível descobrir quais os medicamentos prescrito consoante as queixas dos utentes.

Apesar dos estudos já realizados e da possibilidade de aplicação de diferentes algoritmos aos mesmos, facilmente se constata a existência de diversos outros estudos possíveis. Por exemplo, focar o estudo na análise da prescrição de antibióticos, entre outras características, iria originar diferentes estudos, mais orientados a determinados assuntos, que poderiam fornecer informação mais útil consoante as necessidades de informação da ARSN.

Para finalizar, as técnicas de mineração de dados foram, apenas, aplicadas a 10% dos dados facultados pela ARSN. Consequentemente, seria necessário e interessante, no futuro aplicar

as mesmas técnicas ao conjunto de dados incial, de aproximadamente 22 milhões, pois obviamente os resultados obtidos seriam diferentes e poderiam apresentar uma quantidade de padrões mais elevada e mais interessantes. No entanto, este estudo apenas é possível através da disponibilização de mais recursos computacionais, nomeadamente de máquinas com memória superior a 16 Gb.

As propostas dos estudos apresentados são apenas alguns na imensidão de possíveis estudos existentes com o conjunto de dados facultados, que deverão depender essencialmente das necessidades de informação existentes, relativamente à prescrição de medicamentos. No entanto, é preciso analisar cada uma dessas necessidades e, definir as melhores estratégias, de maneira a ir de encontro com os requisitos impostos. A aplicação da mineração de dados consiste em aplicar um conjunto de passos ao modelo de dados, com o objectivo de se descobrir possíveis padrões e tendências, proporcionando os resultados desejados às organizações em causa.

Bibliografia

[Agrawal et al. 1993] Rakesh Agrawal, Tomasz Imielinski e Arun Swami: "Mining Association Rules between Sets of Items in Large Databases". SIGMOD '93: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, vol. 12, 207--216, 1993.

[Agrawal & Srikant 1994] Rakesh Agrawal e Ramakrishnan Srikant: "Fast Algorithms for Mining Association Rules". Readings in database systems (3rd ed.), Morgan Kaufmann Publishers Inc., 580--592, 1994.

[Agrawal & Srikant 1995] Rakesh Agrawal e Ramakrishnan Srikant: "Mining Sequential Patterns". Proceedings of the Eleventh International Conference on Data Engineering, 3--14, 1995.

[Agrawal *et al*. 1996] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant e Hannu Toivonen e Inkeri A. Verkamo: "Fast discovery of association rules". Advances in Knowledge Discovery and Data Mining, 307--328, 1996.

[Aggarwal & Yu 1999] Charu C. Aggarwal e Philip S. Yu: "Data Mining Techniques for Associations, Clustering and Classification". PAKDD '99: Proceedings of the Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining, 13--23, 1999.

[Andritsos 2002] Periklis Andritsos: "Data Clustering Techniques". University of Toronto, Dep. of Computer Science (Qualifying Oral Exam), 2002.

[Atallah *et al*. 2004] Mikhail Atallah, Robert Gwadera e Wojciech Szpankowski: "Detection of Significant Sets of Episodes in Event Sequences". ICDM '04: Proceedings of the Fourth IEEE International Conference on Data Mining, 3--10, 2004.

[Azevedo & Alípio 2007] Paulo J. Azevedo e Alípio M. Jorge: "Comparing Rule Measures for Predictive Association Rules". ECML '07: Proceedings of the 18th European conference on Machine Learning, 510--517, 2007.

[Beckett *et al*. 2005] Charmagne G. Beckett, Herman Kosasih, Indra Faisal, Nurhayati, Ratina Tan, Susana Widjaja, Erlin Listiyaningsih, Chairin Ma'Roef, Suharyono Wuryadi, Michael J. Bangs, Tatang K. Samsi, Djoko Yuwono, Curtis G. Hayes e Kevin R. Porter: "Early Detection of Dengue Infections using Cluster Sampling around Index Cases". American Journal of Tropical Medicine and Hygiene, vol. 72, 777--782, 2005.

[Becquet *et al*. 2002] Céline Becquet, Sylvain Blachon, Baptiste Jeudy, Jean-Francois Boulicaut e Olivier Gandrillon: "Strong Association Rule Mining for Large Scale Gene Expression Data Analysis: a Case Study on Human SAGE Data". Genome Biology, vol. 3, 1--16, 2002.

[Bellaachia & Guven 2006] Abdelghani Bellaachia e Erhan Guven: "Predicting Breast Cancer Survivability Using Data Mining Techniques". IEEE Swarm Intelligence Symposium, 112--114, 2006.

[Ben-Dor *et al*. 1998] Amir Ben-Dor, Ron Shamir e Zohar Yakhini: "Clustering Gene Expression Patterns". Proceedings of the third annual international conference on Computational molecular biology, 1998.

[Ben-Gal 2007] Irad E. Ben-Gal: "Bayesian Networks". In F. Ruggeri, R. Kenett, and F. Faltin, editors, Encyclopedia of statistics in quality and reliability, John Wiley & Sons, 2007.

[Berkhin 2002] Pavel Berkhin: "Survey of Clustering Data Mining Techniques". Accrue Software, 2002.

[Bezerra *et al*. 2005] George B. Bezerra, Geraldo M. A. Cançado, Marcelo Menossi, Leandro N. De Castro e Fernando J. Von Zuben: "Recent Advantages in Gene Expression Data Clustering: a Case Study with Comparative Results". Genetics and Molecular Research, vol. 4, 514--524, 2005.

[Bianco *et al*. 2005] Andrea Bianco, Gianluca Mardente, Marco Mellia, Maurizio Munafò e Luca Muscariello: "Web Uses Session Characterization via Clustering Techniques". IEEE/ACM Transactions on Networking (TON), vol.17, 405--416, 2005.

[Boullé 2007] Marc Boullé: "Compression-Based Averaging of Selective Naive Bayes Classifiers". The Journal of Machine Learning Research, vol. 8, 1659--1685, 2007.

[Breiman *et al*. 1984] Leo Breiman, J. H. Friedman, R. A. Olshen e C. J. Stone: "Classification and Regression Trees", Wadsworth, 1984.

[Brijs *et al*. 1999] Tom Brijs, Gilbert Swinnen, Koen Vanhoof e Geert Wets: "Using Association Rules for Product Assortment Decisions: a Case Study". KDD '99: Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 254--260,1999.

[Brin *et al*. 1997] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman e Shalom Tsur: "Dynamic Itemset Counting and Implication Rules for Market Basket Data". ACM SIGMOD Record, vol. 26, 255--264, 1997.

[Carey *et al*. 2006] Lisa A. Carey, Charles M. Perou, Chad A. Livasy, Lynn G. Dressler, David Cowan, Kathleen Conway, Gamze Karaca, Melissa A. Troester, Chiu Kit Tse, Sharon Edmiston, Sandra L. Deming, Joseph Geradts, Maggie C. U. Cheang, Torsten O. Nielsen, Patricia G. Moorman, H. Shelton Earp, Robert C. Millikan: "Race, Breast Cancer Subtypes, and Survival in the Carolina Breast Cancer Study". Journal of the American Medical Association, vol. 295, 2492--2502, 2006.

[Carvalho 2008] Cristina Maria Carvalho: "Prescrição de Antibióticos nos Centros de Saúde da Região de Saúde do Norte: Padrão e Variabilidade Geográfica". Dissertação de Mestrado pela Faculdade de Medicina – Universidade do Porto, 2008.

[Casas-Garriga 2003] Gemma Casas-Garriga: "Discovering Unbounded Episodes in Sequential Data". PKDD '03: Proceedings of the Seventh European Conference Principles and Practice of Knowledge Discovery in Databases, 83--94, 2003.

[Chen *et al*. 2004] Hsinchun Chen, Wingyan Chung, Jennifer Jie Xu, Gang Wang, Yi Qin e Michael Chau: "Crime Data Mining: A General Framework and Some Examples", Computer, vol. 39, 50--56, 2004.

[Connolly & Begg 1998] Thomas M. Connolly e Carolyn Begg: "Database Systems: A Practical Approach to Design, Implementation and Management, 2nd Ed.". Addison-Wesley Longman Publishing Co., Inc., 1998.

[Cover & Hart 1967] Thomas M. Cover e Peter E. Hart: "Nearest Neighbor Pattern Classification". IEEE Transactions on Information Theory, vol. 13, 21--27, 1967.

[Creighton & Hanash 2003] Chad Creighton e Samir Hanash: "Mining Gene Expression Databases for Association Rules". Bioinformatics, vol. 19, 79--86, 2003.

[Cunningham & Delany 2007] Pádraig Cunningham e Sarah Jane Delany: "k-Nearest Neighbor Classifiers". Technical Report, UCD School of Computers Science and Informatics, 2007.

[Davies & Bouldin 1979] David L. Davies e Donald W. Bouldin: "A Cluster Separation Measure". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 2, 224--227, 1979.

[Delen *et al*. 2004] Dursun Delen, Glenn Walker e Amit Kadam: "Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods". Artificial Intelligence in Medicine, vol. 34, 113--127, 2004.

[Dempster *et al*. 1977] A. Dempster, N. Laird e D. Rubin: "Maximum likelihood from incomplete data via the EM algorithm". Journal of the Royal Statistical Society, vol. 39, 1--38, 1977.

[El-Halees 2008] Alaa El-Halees: "Mining Students Data to Analyze Learning Behavior: A Case Study", ACIT'2008: International Arab Conference on Information Technology, 2008.

[Ester *et al*. 1996] Martin Ester, Hans-Peter Kriegel, Jörg Sander e Xiaowei Xu: "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". International Conference on Knowledge Discovery and Data Mining, 226--231,1996.

[Fayyad *et al*. 1996a] Usama Fayyad, Gregory Piatetsky-Shapiro e Padhraic Smyth: "The KDD Process for Extracting Useful Knowledge from Volumes of Data". Communications of the ACM, vol. 39, 27--34, 1996.

[Fayyad *et al*. 1996b] Usama Fayyad, Gregory Piatetsky-Shapiro e Padhraic Smyth: "From Data Mining to Knowledge Discovery in Databases". AI Magazine, vol. 13, 37--54, 1996.

[Fisher 1987] Douglas H. Fisher: "Improving Inference through Conceptual Clustering". *AAAI'87:* Proceedings of the Sixth National Conference on Artificial Intelligence, 461--465, 1987.

[Friedman *et al*. 1997] Nir Friedman, Dan Geiger e Moisés Goldszmidt: "Bayesian Network Classifiers", Machine Learning, vol. 29, 131--163, 1997.

[Goethals 2003] Bart Goethals, "Survey on Frequent Pattern Mining". Technical report, Helsinki Institute for Information Technology, 2003.

[Guha *et al*. 1998] Sudipto Guha, Rajeev Rastogi e Kyuseok Shim: "CURE: An Efficient Clustering Algorithm for Large Databases". Proceedings of ACM SIGMOD International Conference on Management of Data, 73--84,1998.

[Guha *et al*. 2000] Sudipto Guha, Rajeev Rastogi e Kyuseok Shim: "ROCK: a Robust Clustering Algorithm for Categorical Attributes". Information Systems, vol. 25, 345--366, 2000.

[Han & Fu 1995] Jiawei Han e Yongjian Fu: "Discovery of Multiple-Level Association Rules from Large Databases". VLDB '95: Proceedings of the 21th International Conference on Very Large Databases, 420—431, 1995

[Han *et al*. 2000] Jiawei Han, Jian Pei e Yiwen Yin: "Mining Frequent Patterns without Candidate Generation": SIGMOD '00: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Cata, 1--12, 2000.

[Han & Kamber 2006] Jiawei Han e Micheline Kamber: "Data mining: Concepts and Techniques (2nd Ed.)". Morgan Kaufmann, 2006.

[Hand 1999] David J. Hand: "Statistics and Data Mining: Intersecting Disciplines". SIGKDD Explorations, vol. 1, 16--19, 1999.

[Hand *et al*. 2001] David J. Hand, Padhraic Smyth e Heikki Mannila: "Principles of Data Mining". MIT Press, 2001.

[Hahsler & Hornik 2007] Michael Hahsler e Kurt Hornik: "New Probabilistic Interest Measures for Association Rules". Intelligent Data Analysis, vol. 11, 437--455, 2007.

[Hipp *et al*. 2000] Jochen Hipp, Ulrich Güntzer e Gholamreza Nakhaeizadeh: "Algorithms for Association Rule Mining – A General Survey and Comparison". ACM SIGKDD Explorations Newsletter, vol. 2, 58--64, 2000.

[Inmon 1996] William H. Inmon: "The Data Warehouse and Data Mining". Communications of the ACM, vol. 39, 49--50, 1996.

[Jain & Dubes 1988] Anil Kumar Jain e Richard D. Dubes: "Algorithms for Clustering Data". Prentice-Hall Advanced Reference Series, 1988.

[Jain *et al*. 1999] Anil Kumar Jain, M. Narasimha Murty e Patrick Joseph Flynn: "Data Clustering: a Review". ACM Computing Surveys (CSUR), vol. 31, 264--323, 1999.

[Jans *et al*. 2008] Mieke Jans, Nadine Lybaert e Koen Vanhoof: "Internal Fraud Risk Reduction: Results of a Data Mining Case Study". 2th International Conference On Enterprise Information Systems, 161--166, 2008.

[Kamath & Musick 1998] Chandrika Kamath e Ron Musick: "Scalable Pattern Recognition for Large-Scale Scientific Data Mining". Center for Applied Scientific Computing Lawrence Livermore National Laboratory, 1998.

[Kaufman & Rosseeuw 1990] Leonard Kaufman and Peter J. Rousseeuw: "Finding Groups in Data: An Introduction to Cluster Analysis". Wiley, 1990.

[Kohonen 1982] Teuvo Kohonen: "Self-Organized Formation of Topologically Correct Feature Maps". Biological Cybernetics, vol. 43, 56--69, 1982.

[Laxman *et al*. 2004] Srivatsan Laxman, P. S. Sastry e K.P. Unnikrishnan: "Fast algorithms for frequent episode discovery in event sequences". Proceedings of the 3rd Workshop on Mining Temporal and Sequential Data, 2004.

[Laxman *et al*. 2004] Srivatsan Laxman, P. S. Sastry e K.P. Unnikrishnan: "Discovering frequent episodes and learning Hidden Markov Models: A formal connection". IEEE Transactions on Knowledge and Data Engineering, vol. 17, 1505--1517, 2005.

[Laxman 2006] Srivatsan Laxman: "Discovering frequent episodes: Fast algorithms, connections with HMMs and generalizations". PhD, Faculty of Engineering - Indian Institute of Science Bangalore, 2006.

[Laxman & Sastry 2006] Srivatsan Laxman e P. S. Sastry: "A Survey of Temporal Data Mining". Academy Proceedings in Engineering Sciences, 2006.

[Lin & Lee 2003] Ming-Yen Lin e Suh-Yin Lee: "Improving the efficiency of interactive sequential pattern mining by incremental pattern discovery". HICSS '03: Proceedings of the 36th Annual Hawaii International Conference on System Sciences,vol. 2, 6--9, 2003.

[Liu *et al*. 1998] Bing Liu, Wynne Hsu e Yiming Ma: "Integrating Classification and Association Rule Mining". KDD-98: Knowledge Discovery in Databases, 1998.

[Lloyd-Williams *et al*. 1995] Lloyd-Williams, M., Jenkins, J., Howden-Leach, H., Mathur, R., Cook, I. e Morris, C.: "Knowledge Discovery in an Infertility Database Using Artificial Neural Networks". IEE Colloquium on Knowledge Discovery in an Infertility Database Using Artificial Neural Networks, vol. 21, 1--3, 1995.

[Mannen et al. 2008] Mammen P. Mammen Jr., Chusak Pimgate, Constantianus J. M. Koenraadt, Alan L. Rothman, Jared Aldstadt, Ananda Nisalak, Richard G. Jarman, James W. Jones5, Anon Srikiatkhachorn, Charity Ann Ypil-Butac, Arthur Getis, Suwich Thammapalo, Amy C. Morrison, Daniel H. Libraty, Sharone Green, Thomas W. Scott: "Spatial and Temporal Clustering of Dengue Virus Transmission in Thai Villages", PLoS Med, vol. 5, 1605--1616, 2008.

[Mannila *et al*. 1997] Heikki Mannila, Hannu Toivonen e A. Inkeri Verkamo: "Discovery of Frequent Episodes in Event Sequences". Data Mining and Knowledge Discovery, vol. 1, 259--289, 1997.

[Merceron & Yacef 2005] Agathe Merceron e Kalina Yacef: "Educational Data Mining: a Case Study", 12th International Conference on Artificial Intelligence in Education, 467--474, 2005.

[Mierswa 2009] Ingo Mierswa: "Non-Convex and Multi-Objective Optimization in Data Mining", Dissertação de doutoramento pela Universidade de Dortmund, 2009.

[Murty & Krishna 1980] M. Narasimba Murty e G. Krishna: "A Computationally Efficient Technique for Data-Clustering". Pattern Recognition, vol. 12, 153--158, 1980.

[Nagy 1968] George Nagy: "State of the Art in Pattern Recognition". Proceedings of the IEEE, vol. 56, 836--862, 1968.

[Nath 2006] Shyam Varan Nath: "Crime Pattern Detection Using Data Mining". Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology, 41--44, 2006.

[Ng & Han 2002] Raymond T. Ng e Jiawei Han: "CLARANS: A Method for Clustering Objects for Spatial Data Mining". IEEE Transactions on Knowledge and Data Engineering, vol. 14, 1003--1016, 2002

[Nguyen & Abiteboul 2001] Benjamin Nguyen & Serge Abiteboul: ""A Hash-Tree based Algorithm for Subset Detection: Analysis and Experiments". Verso Internal Report, 2001.

[Orponen 1994] Pekka Orponen: "Computational Complexity of Neural Networks: A Survey". Nordic Journal of Computing, vol. 1, 94--110, 1994.

[Pasquier *et al*. 1999] Nicolas Pasquier, Yves Bastide, Rafik Taouil e Lotfi Lakhal: "Discovering Frequent Closed Itemsets for Association Rules". ICDT '99: Proceedings of the 7th International Conference on Database Theory, 398--416,1999.

[Phua *et al*. 2005] Clifton Phua, Vincent Lee, Kate Smith e Ross Gayler: "A Comprehensive Survey of Data Mining-based Fraud Detection Research", 2007.

[Piatetsky-Shapiro 1990] Gregory Piatetsky-Shapiro: "Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop". AI Magazine, vol. 11, 68--70, 1990.

[Piatetsky-Shapiro 1991] Gregory Piatetsky-Shapiro: "Discovery, Analysis, and Presentation of Strong Rules". Knowledge Discovery in Databases, 229-.248, 1991.

[Quinlan 1986] J. Ross Quinlan: "Induction of Decision Trees". Machine Learning, vol. 1 , 81--106,1986.

[Quinlan 1993] J. Ross Quinlan: "C4.5: Programs for Machine Learning". Morgan Kaufmann Publishers Inc., 1993.

[Romero & Ventura 2007] Cristobal Romero e Sebastian Ventura: "Educational data mining: A survey from 1995 to 2005". Expert Systems with Applications: An International Journal, vol. 33, 135--146, 2007.

[Rumelhart et al. 1986] David E. Rumelhart, Geoffrey E. Hilton e Ronald J. Williams: "Learning Internal Representations by Error Propagation". In D. E. Rumelhart and J. L. McClelland, editors, Parallel Distributed Processing. MIT Press, 1986.

[Santos e Azevedo 2005] Manuel Filipe Santos e Carla Sousa Azevedo: "Descoberta do Conhecimento em Bases de Dados". FCA – Editora de Informática, Lda., 2005.

[Savasere *et al.* 1995] Ashok Savasere, Edward Omiecinski e Shamkant B. Navathe: "An efficient Algorithm for Mining Association Rules in Large Databases". VLDB '95: Proceedings of the 21th International Conference on Very Large Data Bases, 432--444, 1995.

[Schultz et al. 2001] Matthew G. Schultz, Eleazar Eskin, Erez Zadok e Salvatore J. Stolfo: "Data Mining Methods for Detection of New Malicious Executables". Proceedings of the 2001 IEEE Symposium on Security and Privacy, 38--49, 2001.

[Silberschatz & Tuzhilin 1996] Avi Silberschatz e Alexander Tuzhilin: "What Makes Patterns Interesting in Knowledge Discovery Systems". IEEE Transactions on Knowledge and Data Engineering, vol. 8, 970--974, 1996.

[Shintani & Kitsuregawa 1998] Takahiko Shintani e Masaru Kitsuregawa: "Mining Algorithms for Sequential Patterns in Parallel: Hash Based Approach". PAKDD '98: Proceedings of the Second Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining, 283--294, 1998.

[Srikant & Agrawal 1995] Ramakrishnan Srikant e Rakesh Agrawal: "Mining Generalized Association Rules". Future Generation Computer Systems, vol. 13, 161--180, 1995.

[Srikant & Agrawal 1996] Ramakrishnan Srikant e Rakesh Agrawal: "Mining Quantitative Association Rules in Large Relational Tables". ACM SIGMOD Record, vol. 25, 1--12, 1996**.**

[Steinbach *et al.* 2000] Michael Steinbach, George Karypis e Vipin Kumar: "A Comparison of Document Clustering Techniques". KDD Workshop on Text Mining, 2000.

[Sumathi & Sivanandam 2006] Sumathi, S. e Sivanandam, S.N.: "Introduction to Data Mining and its Applications (Studies in Computational Intelligence (SCI))". Springer-Verlag Berlin Heidelberg, 2006.

[Tan *et al*. 2006] Pang-Ning Tan, Michael Steinbach e Vipin Kumar: "Introduction to Data Mining (1st Ed.)". Addison-Wesley Longman Publishing Co., Inc, 2006.

[Tomé *et al*. 2008] André Tomé, Paula Broeiro e António Faria-Vaz: "Os sistemas de prescrição electrónica". Associação Portuguesa dos Médicos de Clínica Geral, vol. 24, 632--640, 2008.

[Vellido *et al*. 1999] Alfredo Vellido, P. J. G. Lisboa e J. Vaughan: "Neural networks in business: a survey of applications (1992–1998)". Expert Systems with Applications, vol. 17, 51--70, 1999.

[Visalakshi *et al*. 2009] N. Karthikeyani Visalakshi e K. Thangavek: "Impact of Normalization in Distributed K-Means Clustering". International Journal of Soft Computing, vol. 4, 168--172, 2009.

[Wang & Han 2004] Jianyong Wang e Jiawei Han: "BIDE: Efficient Mining of Frequent Closed Sequences". ICDE '04: Proceedings of the 20th International Conference on Data Engineering, 79--90, 2004.

[Wang *et al*. 1997] Wei Wang, Jiong Yang e Richard R. Muntz: "STING: A Statistical Information Grid Approach to Spatial Data Mining". VLDB '97: Proceedings of the 23rd International Conference on Very Large Data Bases, 186--195, 1997.

[Witten & Frank 2005] Ian H. Witten & Eibe Frank: "Data Mining: Practical Machine Learning Tools and Techniques (2nd Ed.)". Morgan Kaufmann, 2005.

[Wu et al. 2003] Baolin Wu, Tom Abbott, David Fishman, Walter McMurray, Gil Mor, Kathryn Stone, David Ward, Kenneth Williams e Hongyu Zhao: "Comparison of Statistical Methods for Classification of Ovarian Cancer using Mass Spectrometry Data". Bioinformatics, vol. 19, 1636--1643, 2003.

[Wu *et al*. 2008] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand e Dan Steinberg: "Top 10 Algorithms in Data Mining". Knowledge and Information Systems, vol. 14, 1--37, 2008.

[Yacef 2005] Yacef, K., "The Logic-ITA in the Classroom: A Medium Scale Experiment". International Journal on Artificial Intelligence in Education, vol. 15, 41--60, 2005.

[Yan *et al*. 2003] Xifeng Yan, Jiawei Han e Ramin Afshar: "CloSpan: Mining Closed Sequential Patterns in Large Datasets". In Proceedings of the 2003 International SIAM Conference on Data Mining (SDM03), 166--177, 2003.

[Zaki et al. 1997] Mohammed Javeed Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara e Wei Li: "New Algorithms for Fast Discovery of Association Rules". University of Rochester, 1997.

[Zaki 1998] Mohammed Javeed Zaki: "Efficient enumeration of frequent sequences". CIKM '98: Proceedings of the seventh International Conference on Information and Knowledge Management, 68--75, 1998.

[Zelic *et al*. 1997] Igor Zelic, Igor Kononenko, Nada Lavrac e Vanja Vuga: "Induction of Decision Trees and Bayesian Classification Applied to Diagnosis of Sport Injuries". Journal of Medical Systems, vol. 21, 429--444, 1997.

[Zhang 2000] Guoqiang Peter Zhang: "Neural Networks for Classification: A Survey". IEEE Transactions on Systems, Man, and Cybernetics (Part C: Applications and Reviews), vol.30, 451--462, 2000.

[Zhang *et al*. 1996] Tian Zhang, Raghu Ramakrishnan e Miron Livny: "BIRCH: An Efficient Data Clustering Method for Very Large Databases". ACM SIGMOD Conference, 103--114, 1996.

[Zheng *et al*. 2001] Zijian Zheng, Ron Kohavi e Llew Mason: "Real World Performance of Association Rule Algorithms". KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining,401--406, 2001.

Referências WWW

[www1] http://www.thearling.com/text/dmtechniques/dmtechniques.htm

Alex Berson, Stephen Smith e Kurt Thearling: "An Overview of Data Mining Techniques". Este *site* apresenta uma pequena descrição de alguns algoritmos da mineração de dados e exemplos relacionados com a aplicação desses algoritmos, acompanhado com uma introdução histórica sobre a mineração de dados.

[www2] http://www.taborcommunications.com/dsstar/00/0704/101861.html

Ed Colet: "DSstar: Clustering and Classification: Data Mining Approaches". Neste site é descrito duas das técnicas de descoberta de padrões, a segmentação e classificação. É também abordado as principais diferenças entre ambas as técnicas.

[www3] http://www.psy.gla.ac.uk/~steve/pr/edel.html

Herb Edelstein: "Data Mining: Exploiting the Hidden Trends in Your Data". Este site apresenta um introdução às técnicas de mineração de dados, focalizando-se também, na relação e diferença entre a mineração de dados e *OLAP*.

[www4] http://michael.hahsler.net/research/association_rules/measures.html

Michael Hahsler: "A Comparison of Commonly Used Interest Measures for Association Rules", 2009. Neste site é descrito algumas das medidas de interesse mais utilizadas relativamente a regras de associação e conjuntos de elementos.

[www5] http://www.florian.verhein.com/teaching/2008-01-09/fp-growth-presentation_v1%20(handout).pdf

Florian Verhein: "Frequent Pattern Growth (FP-Growth) Algorithm - An Introduction", 2008. Neste site pode ser descarregado um apresentação que descreve o funcionamento do algoritmo *FP-Growth* de uma maneira sucinta e clara.

Anexos I

Princípios Activos dos utentes da Tabela 15									
799710208	**736825200**	**44967785**	**526825203**	**140316210**	**851917244**	**998056243**	**534689281**	**446815219**	**732642781**
Acetilcisteína	Ácido acetilsalicílico	Acarbose	Acemetacina	Acarbose	Acenocumarol	Acamprosato	Acetilcisteína	Acetilsalicilato de lisina	Acenocumarol
Acetilsalicilato de lisina	Ácido fólico	Ácido acetilsalicílico	Clotrimazol	Acenocumarol	Ácido mefenâmico	Acarbose	Alprazolam	Aciclovir	Ácido alendrónico + Colecalciferol
Ácido fusídico	Ácido valpróico	Amoxicilina + Ácido clavulânico	Desloratadina	Ácido zoledrónico	Betametasona + Clotrimazol + Gentamicina	Ácido acetilsalicílico	Amiodarona	Ácido acetilsalicílico	Bendazac
Ácido niflúmico	Alopurinol	Baclofeno	Dosulepina	Amlodipina	Bromazepam	Ácido fólico	Bisoprolol	Ácido alendrónico + Colecalciferol	Bioflavonóides
Ambroxol	Amiodarona	Bisacodilo	Etofenamato	Amoxicilina + Ácido clavulânico	Ciprofloxacina	Ácido fusídico	Calcitonina de salmão	Ácido zoledrónico	Bromazepam
Amorolfina	Amissulprida	Citrato de sódio + Laurilsulfoacetato de sódio	Etoricoxib	Azitromicina	Clotrimazol	Amitriptilina	Candesartan + Hidroclorotiazida	Albendazol	Diclofenac
Beta-histina	Amoxicilina + Ácido clavulânico	Clorazepato dipotássico	Flurazepam	Betametasona	Dimetindeno + Fenilefrina	Azitromicina	Citicolina	Alopurinol	Furosemida
Bioflavonóides	Bromazepam	Clotrimazol	Lansoprazol	Bezafibrato	Diosmina	Butilescopolamina + Paracetamol	Cloxazolam	Alprazolam	Glibenclamida

Carbonato de cálcio + Lactogluconato de cálcio	Cianocobalamina	Digoxina	Meloxicam	Bioflavonóides	Domperidona	Carbamazepina	Diclofenac	Amoxicilina	Irbesartan + Hidroclorotiazida
Carbonato de di-hidróxido de alumínio e sódio + Dimeticone	Citicolina	Dimeticone	Metoclopramida	Bisoprolol	Enoxaparina sódica	Cianocobalamina + Piridoxina + Tiamina	Diltiazem	Amoxicilina + Ácido clavulânico	Levotiroxina sódica
Celecoxib	Diflucortolona + Isoconazol	Espironolactona	Mexazolam	Bromazepam	Escitalopram	Clonixina	Escitalopram	Aspartato de magnésio	Metformina
Cetirizina	Furosemida	Flucloxacilina	Nimesulida	Bromexina	Hidroclorotiazida + Amilorida	Clotrimazol	Esomeprazol	Atenolol	Omeprazol
Cetoprofeno	Gabapentina	Furosemida	Oxazepam	Captopril	Ibuprofeno	Cocarboxilase	Etofenamato	Azitromicina	Paracetamol
Citrato de sódio + Laurilsulfoacetato de sódio	Glicerol	Glibenclamida + Metformina	Oxcarbazepina	Celecoxib	Levocetirizina	Diazepam	Flucloxacilina	Baclofeno	Piroxicam
Clobazam	Hesperidina + Ruscus aculeatus + Ácido ascórbico	Gliclazida	Risperidona	Cetirizina	Midazolam	Diclofenac	Ginkgo biloba	Bioflavonóides	Rosuvastatina
Clopidogrel	Levomepromazina	Idebenona	Tramadol + Paracetamol	Cianocobalamina	Nimesulida	Espironolactona	Guaifenesina + Salbutamol	Bisoprolol	Vacina contra a gripe
Cloreto de magnésio	Macrogol + Bicarbonato de potássio + Bicarbonato de sódio + Cloreto de sódio	Lansoprazol	Zolpidem	Citicolina	Omeprazol	Flurazepam	Irbesartan + Hidroclorotiazida	Bromazepam	Verapamilo
Cloreto de tróspio	Pantoprazol	Levodopa + Carbidopa	-	Clonixina	Paracetamol	Furosemida	Levocetirizina	Brometo de otilónio	-
Diclofenac	Paracetamol	Lisados polibacterianos	-	Cloreto de magnésio	Paracetamol + Pseudoefedrina	Gabapentina	Levotiroxina sódica	Budesonida	-
Esomeprazol	Perindopril	Loprazolam	-	Cloreto de tróspio	Prednisolona + Neomicina + Sulfacetamida	Haloperidol	Loperamida	Candesartan	-
Etofenamato	Piroxicam	Metformina	-	Cocarboxilase	Prednisona	Ibuprofeno	Lorazepam	Cefadroxil	-
Flupirtina	Proteínosuccinilato de ferro	Nateglinida	-	Diazepam	Sinvastatina	Insulina isofânica	Metolazona	Ceftriaxona	-
Fluticasona + Salmeterol	Sertaconazol	Nifedipina	-	Diclofenac	Terbinafina	Iodopovidona	Mometasona	Cetirizina	-
Hidroxizina	Solifenacina	Nimesulida	-	Digoxina	Trepibutona	Lorazepam	Naproxeno	Ciamemazina	-
Hipericão	Terbinafina	Nimodipina	-	Enalapril	Triamcinolona	Metamizol magnésico	Nimesulida	Ciclopirox	-
Ibuprofeno	Tramadol	Picetoprofeno	-	Esomeprazol	Trimebutina	Oxazepam	Nitrofurantoína	Ciprofloxacina	-
Imidapril	Vacina contra a gripe	Pidolato de magnésio	-	Etoricoxib	Vacina contra a gripe	Oxcarbazepina	Paracetamol	Claritromicina	-

Indometacina	Vacina pneumocócica poliosídica	Prednisolona + Neomicina + Polimixina B	-	Fenofibrato	-	Pancreatina	Paracetamol + Codeína	Clobazam	-
Levocetirizina	-	Saccharomyces boulardii	-	Flucloxacilina	-	Saccharomyces boulardii	Paroxetina	Clopidogrel	-
Loperamida	-	Selegilina	-	Fluconazol	-	Sulfato ferroso + Ácido fólico	Sinvastatina	Cloranfenicol	-
Lorazepam	-	Sinvastatina	-	Flunarizina	-	Tramadol	Vacina contra a gripe	Clorazepato dipotássico	-
Macrogol	-	Sobrerol	-	Fluoxetina	-	-	Varfarina	Cloreto de magnésio	-
Messalazina	-	Verapamilo	-	Furosemida	-	-	-	Cloreto de tróspio	-
Metformina	-	-	-	Glibenclamida	-	-	-	Cloropromazina	-
Mexazolam	-	-	-	Gliclazida	-	-	-	Clotrimazol	-
Mometasona	-	-	-	Glucosamina	-	-	-	Cloxazolam	-
Montelucaste	-	-	-	Hidroclorotiazida + Amilorida	-	-	-	Desloratadina	-
Naproxeno	-	-	-	Hidrocortisona	-	-	-	Diazepam	-
Nimesulida	-	-	-	Ibuprofeno	-	-	-	Diclofenac	-
Nitrofurantoína	-	-	-	Indapamida	-	-	-	Digoxina	-
Paracetamol	-	-	-	Lansoprazol	-	-	-	Domperidona	-
Picetoprofeno	-	-	-	Levocetirizina	-	-	-	Dosulepina	-
Piroxicam	-	-	-	Levodopa + Carbidopa	-	-	-	Ebastina	-
Prednisolona	-	-	-	Levodopa + Carbidopa + Entacapona	-	-	-	Enoxaparina sódica	-
Proglumetacina	-	-	-	Lisinopril	-	-	-	Esomeprazol	-
Ranitidina	-	-	-	Lorazepam	-	-	-	Etofenamato	-
Rupatadina	-	-	-	Losartan + Hidroclorotiazida	-	-	-	Fenofibrato	-
Saccharomyces boulardii	-	-	-	Metformina	-	-	-	Fenticonazol	-
Salbutamol	-	-	-	Mexazolam	-	-	-	Flucloxacilina	-
Sertralina	-	-	-	Mirtazapina	-	-	-	Fluconazol	-

Sinvastatina	-	-	-	Mometasona	-	-	-	Fluoxetina	-
Sulfametoxazol + Trimetoprim	-	-	-	Nabumetona	-	-	-	Fluticasona + Salmeterol	-
Sulpirida	-	-	-	Nimesulida	-	-	-	Formoterol	-
Tioconazol	-	-	-	Norfloxacina	-	-	-	Fosfato tricálcico + Colecalciferol	-
Tramadol	-	-	-	Omeprazol	-	-	-	Furosemida	-
Trazodona	-	-	-	Oxazepam	-	-	-	Glucosamina	-
Vacina contra a gripe	-	-	-	Paracetamol	-	-	-	Hidroclorotiazida + Amilorida	-
Venlafaxina	-	-	-	Paroxetina	-	-	-	Ibuprofeno	-
Zolmitriptano	-	-	-	Ranitidina	-	-	-	Insulina isofânica	-
-	-	-	-	Ropinirol	-	-	-	Irbesartan + Hidroclorotiazida	-
-	-	-	-	Sinvastatina	-	-	-	Lansoprazol	-
-	-	-	-	Sucralfato	-	-	-	Latanoprost	-
-	-	-	-	Sulfato ferroso + Ácido fólico	-	-	-	Levocetirizina	-
-	-	-	-	Terbinafina	-	-	-	Levotiroxina sódica	-
-	-	-	-	Tramadol	-	-	-	Lisinopril	-
-	-	-	-	Tri-hexifenidilo	-	-	-	Loflazepato de etilo	-
-	-	-	-	Trimetazidina	-	-	-	Loratadina	-
-	-	-	-	Valeriana	-	-	-	Lorazepam	-
-	-	-	-	Valsartan	-	-	-	Losartan + Hidroclorotiazida	-
-	-	-	-	Valsartan + Hidroclorotiazida	-	-	-	Maprotilina	-
-	-	-	-	Varfarina	-	-	-	Mebendazol	-
-	-	-	-	Zolmitriptano	-	-	-	Metformina	-
-	-	-	-	Zolpidem	-	-	-	Metoclopramida	-
-	-	-	-	-	-	-	-	Mirtazapina	-
-	-	-	-	-	-	-	-	Nebivolol	-

-	-	-	-	-	-	-	-	Nifedipina	-
-	-	-	-	-	-	-	-	Nimesulida	-
-	-	-	-	-	-	-	-	Nitrofurantoína	-
-	-	-	-	-	-	-	-	Nitroglicerina	-
-	-	-	-	-	-	-	-	Omeprazol	-
-	-	-	-	-	-	-	-	Oxazepam	-
-	-	-	-	-	-	-	-	Oxcarbazepina	-
-	-	-	-	-	-	-	-	Pantoprazol	-
-	-	-	-	-	-	-	-	Paracetamol	-
-	-	-	-	-	-	-	-	Pravastatina	-
-	-	-	-	-	-	-	-	Propranolol	-
-	-	-	-	-	-	-	-	Quetiapina	-
-	-	-	-	-	-	-	-	Ramipril	-
-	-	-	-	-	-	-	-	Risperidona	-
-	-	-	-	-	-	-	-	Rosuvastatina	-
-	-	-	-	-	-	-	-	Rupatadina	-
-	-	-	-	-	-	-	-	Saccharomyces boulardii	-
-	-	-	-	-	-	-	-	Sinvastatina	-
-	-	-	-	-	-	-	-	Sinvastatina + Ezetimiba	-
-	-	-	-	-	-	-	-	Sulfametoxazol + Trimetoprim	-
-	-	-	-	-	-	-	-	Sulfato ferroso + Ácido fólico	-
-	-	-	-	-	-	-	-	Trazodona	-
-	-	-	-	-	-	-	-	Valaciclovir	-
-	-	-	-	-	-	-	-	Valproato semisódico	-
-	-	-	-	-	-	-	-	Valsartan + Hidroclorotiazida	-
-	-	-	-	-	-	-	-	Varfarina	-

-	-	-	-	-	-	-	-	Zolmitriptano	-
-	-	-	-	-	-	-	-	Zolpidem	-

Tabela 28 – Listagem dos príncipios activos dos utentes da Tabela 18.