Universidade do Minho
Escola de Engenharia

Álvaro José Castro Moreira da Silva

Optimização da Cadeia de Abastecimento de um Hipermercado através da Utilização de Mineração de Dados

Dezembro, 2008

Universidade do Minho
Escola de Engenharia

Álvaro José Castro Moreira da Silva

Optimização da Cadeia de Abastecimento de um Hipermercado através da Utilização de Mineração de Dados

Tese de Mestrado em Informática

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

Dezembro, 2008

Universidade do Minho, ___/___/_____

Assinatura: _____

Computers have promised us a fountain of wisdom,
but delivered us a flood of data.
W. J. Frawley

Agradecimentos

Gostaria de agradecer a contribuição de algumas pessoas que, de diversas formas, me ajudaram no decorrer deste trabalho. A elas dedico-lhes esta dissertação expressando a minha total gratidão por toda a ajuda, motivação e apoios prestados.

Ao Doutor Orlando Belo, professor amigo já dos tempos da faculdade, que me tem orientado e aconselhado sempre que possível no decorrer desta investigação. Com ele e com os seus ensinamentos tenho crescido e passado bons momentos dentro e fora do seu gabinete. O seu papel de pedagogo, consultor e amigo permitiu-me dar este novo passo da minha carreira e a não desistir na recta final.

Ao meu pai, que me incentiva constantemente. Obrigado por todo o apoio demonstrado desde o início deste trabalho, motivando-me sempre para escrever melhor. Agradeço-lhe, acima de tudo, a educação que me tem dado, as ideias e princípios que me tem transmitido ao longo da vida e que me fazem crescer.

Aos meus irmãos pelos momentos de família. Se um dia os mencionei aqui, espero ter o prazer também de um dia ler o meu nome num trabalho feito por eles.

Aos meus amigos, que estão sempre perto, mesmo eu estando sempre longe. Este ano foi um ano especial, um ano difícil e isolado, que espero recompensar-vos no futuro.

Por fim, mas nunca em último, à minha princesa CI, que tem aguentado e suportado todos os meus momentos profissionais e académicos menos positivos, à sua coragem por me tentar animar diariamente e ao amor infinito que me tem dado.

Para a minha mãe, com muitas saudades...

Resumo

As organizações retalhistas e grossistas são duas das indústrias com maior dinamismo na actualidade. Devido à sua diversidade de processos, permitem gerar um volume gigantesto de dados operacionais diariamente. Do ponto de vista de um decisor de negócio é evidente que estes dados brutos, depois de "transformados" em informação pelos seus analistas, permitirão conhecer melhor o seu negócio e ganhar vantagem. Muitos dos gestores destas organizações na actualidade, infelizmente, para além de não possuirem sistemas analíticos apropriados de suporte à tomada de decisão, denominados por *Data Warehouses*, parte da informação do seu negócio é-lhes facultada pelos seus analistas com o apoio de ferramentas de extracção directa (ferramentas de *Business Intelligence*) dos seus dados brutos sobre os seus sistemas operacionais. Para além da informação directa recolhida a partir destes repositórios poder-se-á tentar induzir conhecimento válido e útil dos mesmos repositórios utilizando técnicas e algoritmos de Mineração de Dados (MD). Esta dissertação de mestrado surge no seguimento deste enquadramento e tem como objectivo principal o estudo de técnicas de MD para optimizar o processo de abastecimento de um hipermercado.

Abstract

Retailers and wholesalers are two of the most dynamic industries at the moment. Because of their processes diversity these organizations allow generating large volumes of operational data every day. To a business decision person it is clear that these raw data, after being "transformed" into information by their analysts, will increase their own business knowledge gaining advantage. Many of the recent managers of these organizations, unfortunately, besides not having analytical oriented systems, also known as Data Warehouses, appropriated to support their business decisions, the greatest part of their business information is supplied by their analysts using several tools (*Business Intelligence* tools) to do data extractions directly from their operational repositories. Besides these known information obtained from these repositories probably some valid and useful knowledge could be induced from the same repositories using Data Mining techniques and algorithms. Having this in mind, the current masters dissertation was proposed with the following objectives: study Data Mining techniques to improve one hypermarket supply chain process.

Índice

Índice de Figuras

Índice de Tabelas

Lista de acrónimos e abreviaturas

BI	–	Business Intelligence
DM	–	Data Mart
DW	–	Data Warehouse / Data Warehousing
ETL	–	Extract Transform and Load
KDD	–	Knowledge Discovery in Databases
OLAP	–	Online Analytical Processing
DCBD	–	Descoberta de Conhecimento em Bases de Dados
PMML	–	Predictive Modeling Markup Language
MBA	–	Market Basket Analysis
MD	–	Mineração de Dados
ROC	–	Receiver Operating Characteristic
UAC	–	Under Area Curve

Before everything else, getting ready is the secret of success.
Henry Ford

Capítulo 1

Introdução

A economia da actualidade vive uma instabilidade inegável. As empresas sobrevivem perante um clima de adversidade económica geral e todas elas são obrigadas a uma competitividade cada vez mais extrema, sobrevivendo aquelas cuja qualidade e serviços se evidenciem das demais. Neste contexto de competição extrema, a sobrevivência das empresas implica que todas elas adoptem estratégias que assegurem vantagem sustentável face aos seus principais competidores [Moura, 2006].

Neste sentido as empresas apenas conseguirão vantagens competitivas por via da diferenciação não só da qualidade mas também do serviço (valor) prestado aos clientes, operando com custos mais baixos (possíveis com melhorias de produtividade), ou ainda combinando ambos os factores. Cabena *et al* (1997) afirmam que as organizações são cada vez mais forçadas a reavaliar a forma como fazem os seus negócios e a encontrar novos caminhos que consigam responder às mudanças globais.

O mundo sofre imensas mudanças e mutações diárias a todos os níveis, sociais, políticas e económicas. Vive-se hoje numa era de grandes transformações e de inovação, de apostas, de grandes riscos e incertezas, que avançam a um ritmo exponencial, cujas fronteiras estão apenas limitadas pela nossa própria imaginação.

O conceito de globalização é representativo destas rápidas e exigentes mudanças e ao raio da sua acção atingindo, como o próprio nome sugere, o mundo inteiro. Para

Giddens (1999) a globalização, para além de económica, é também política, tecnológica e cultural sendo que é primordialmente "influenciada pelo progresso das comunicações".

Na actualidade há uma grande aposta e revolução nas áreas das tecnologias de inovação cuja aplicabilidade dentro de todo o tipo de organizações visa, sobretudo, o aumento e a manutenção da sua competitividade com a oferta de melhores produtos e serviços aos seus clientes. Nestas áreas tecnológicas, cujo mercado atinge também o limiar da hiper competitividade, as comunicações desempenham cada vez mais um papel preponderante nesta evolução. Segundo Moura (2006) as tecnologias de informação e da comunicação permitem contactos cada vez mais fáceis e baratos entre fornecedores e clientes, praticamente em tempo real, entre quaisquer lugares do mundo.

Duas indústrias afectadas por estas mudanças são, sem dúvida, as indústrias *retalhistas e grossistas*.

> Segundo o dicionário universal da língua portuguesa o conceito de retalho corresponde à "*parte de uma coisa que se retalhou, um pedaço, uma fracção ou um segmento*". Toda a indústria retalhista corresponde a uma organização responsável por vender produtos em pequenas quantidades, "a retalho", para consumo próprio.
>
> Por outro lado, o comércio *grossista* corresponde à venda de grandes volumes de produtos essencialmente destinados a elementos intermediários de uma cadeia de abastecimento tais como os próprios comerciantes retalhistas.

Hoje em dia qualquer tipo de retalhista, grossista e seus fornecedores estão presentes em multi canais de venda tais como catálogos, a *internet*, as lojas físicas, *call centers e kiosks.* O facto de haver uma oferta interminável de meios para a divulgação de novos produtos faz também com que as compras directas e/ou encomendas a fornecedores sejam mais facilitadas. Em estudos recentes relativos ao comportamento do consumidor

directo, está provado que a frequência de compras relativas a compradores multi canal é já superior à frequência de compradores num único local físico.

1.1 Motivação

Sabendo desta competitividade crescente nos mercados da actualidade e das mudanças tecnológicas claras, qualquer gestor terá de dar atenção a todo tipo de detalhe que lhe possibilitará tirar vantagem. Uma das formas que lhe permite alargar o seu conhecimento àcerca do seu próprio negócio é através da informação relativa às operações diárias do seu próprio negócio. Por exemplo, contextualizando na evolução histórica apresentada anteriormente, multi canais de venda permitirão guardar cada vez mais dados de diversas origens. Neste sentido, nos últimos anos, praticamente todas as organizações têm vindo a acumular volumes de dados, cada vez maiores, relativos a estas operações diárias necessárias para a movimentação do seu negócio: *as vendas directas a consumidores e as ordens de encomenda/compra a fornecedores e as respectivas recepções* são dois exemplos que permitem gerar diariamente imensos registos.

A título exemplificativo veja-se o seguinte cenário relativo a uma operação normal de um consumidor de um retalhista numa grande superfície física: *num determinado hipermercado, com uma grande gama de produtos, qualquer consumidor poderá seleccionar das prateleiras imensos produtos para o seu carrinho de compras. Os produtos da esquerda estão em promoção e os da direita, mesmo não estando em promoção, são os produtos de referência por serem necessidades básicas para a sua casa. Depois de encher o seu carrinho com alguns produtos, o consumidor dirigir-se-á a uma caixa para efectuar o respectivo pagamento, não se esquecendo de incluir ainda um dos produtos que está exposto mesmo ao lado da caixa de pagamento. Neste momento o pagamento poderá ser efectuado de diversas formas. O consumidor decide então pagar com o cartão de crédito uma parte e outra parte com dinheiro,*

apresentando também o cartão de descontos fornecido anteriormente pela mesma organização.

Analisando passo a passo este exemplo simples verifica-se que, em qualquer hipermercado, há uma razoável quantidade de produtos nas prateleiras. Esta existência e quantidade de produtos expostos não acontece ao acaso e sem qualquer estudo prévio. Existe, com certeza, um ou mais elementos do negócio responsáveis por tomar as melhores decisões relativas à selecção dos produtos para determinada loja, as quantidades e a sua disposição nas prateleiras e corredores (muito embora este processo seja completamente transparente ao olhos dos consumidores comuns).

Qualquer decisor perante este cenário prático irá imediatamente traçar um conjunto de pontos e interrogações a ter em linha de conta para as suas futuras decisões, por exemplo:

- *Quais os produtos que deverei ter na loja?*
- *De que forma serão dispostos estes produtos pelas próprias prateleiras?*
- *Em que quantidades deverão existir nas prateleiras?*
- *Qual a forma de obter os produtos? Serão produtos de marca própria ou fornecidos? Devo produzir estes produtos ou devo procurar um fornecedor que me forneça este produto?*
- *Quais os critérios para a selecção dos fornecedores?*

Para além da existência dos produtos e a sua disposição nas prateleiras, pode-se também reparar que, do lado esquerdo, existe um conjunto de produtos em promoção, mas do lado direito não. Muito embora as necessidades básicas residissem nos produtos do lado direito, este consumidor foi também induzido, por impulso, a encher o seu carrinho com alguns produtos que estavam em promoção.

Olhando também para este ponto, da mesma forma que foram colocadas as questões anteriores, os decisores podem formular as seguintes questões:

- *Que produtos devo colocar em promoção?*
- *Quando devo fazer uma promoção?*
- *Qual a minha política de preço?*
- *De que forma vou induzir os meus clientes a comprar o produto?*

Na parte final do exemplo, o consumidor poderá pagar de diversas formas, concretizando assim a sua compra. Diversas formas de pagamento são permitidas na mesma transacção, o que facilitou bastante o consumidor por não ter dinheiro suficiente naquela conta. Para além disso, este consumidor aceitou ser portador de um cartão de fidelização que lhe permite acumular alguns pontos, que após alguns meses de compras consegue abater no valor final da transacção. Daqui os decisores poderão colocar as seguintes questões:

- *Como devo fidelizar os meus clientes?*
- *Que formas de pagamento serão permitidas?*
- *Que tipo de vantagens terá o cartão de fidelização?*

Como se pode verificar, numa prespectiva organizacional, um simples exemplo pode representar um enorme conjunto de preocupações prévias para um decisor do negócio. As questões anteriormente colocadas são apenas um pequeno exemplo representativo das preocupações gerais de qualquer decisor às quais ele terá de obter resposta em tempo útil. No entanto, para conseguir dar essa resposta terá também de tomar um conjunto de decisões. Muito embora a situação apresentada seja bastante simplificada e representativa de um processo natural de compra de um consumidor, a quantidade de informação que se pode guardar e posteriormente extrair directamente é enorme. Desta forma os principais intervenientes do negócio de um retalhista poderão saber quais os

produtos mais comprados pelos seus consumidores e quais as formas de pagamento mais utilizadas. Possibilitará obter informação extra relativa ao próprio consumidor com a utilização de dados do cartão de fidelização que, em termos computacionais, ficam sempre associados à própria compra.

Figura 1 – Pirâmide demonstrativa do processo de decisão

Se diariamente qualquer decisor conseguir obter nova informação relativa ao seu negócio, então poderá também tomar as decisões mais acertadas e responder convenientemente a algumas das perguntas anteriormente exemplificadas. Neste contexto, a principal motivação para esta dissertação será o estudo de alternativas, com suporte computacional, que auxiliem qualquer decisor de um retalhista fictício na sua tomada de decisão relativamente a algumas questões pertinentes associadas ao processo de abastecimento, melhorando assim este processo.

Seguindo a pirâmide relativa ao processo de decisão exemplificada pela figura 1 o decisor terá como base dados operacionais guardados diariamente (*primeira camada da pirâmide*), *informação directa conhecida* transformada a partir dos dados (segunda camada da pirâmide) *e porventura desconhecida* dos seus repositórios de dados para a obtenção de novo conhecimento (terceira camada da pirâmide). Desta forma o decisor poderá melhorar o seu conhecimento para a tomada de decisão (topo da pirâmide) relativamente ao abastecimento dos seus armazéns e hipermercados.

1.2 Objectivos

Duas ideias chave orientaram este trabalho. Em primeiro lugar, a tentativa de consolidar conhecimentos funcionais na área de comércio grossista e retalhista focando o processo de compras de um cliente intermédio (retalhista) aos seus fornecedores. Em segundo lugar, compreender o processo da descoberta de conhecimento em grandes bases de dados utilizando dados provenientes desta interacção entre ambos os intervenientes.

No âmbito da descoberta de conhecimento em bases dados (*Knowledge Discovery in Databases*) serão aplicadas técnicas de Mineração de Dados (MD), conceitos apresentados posteriormente no capítulo 4. Serão utilizadas as técnicas de mineração para (a) indução de *Regras de Associação* numa tentativa de descobrir afinidades entre produtos comprados e recebidos; (b) criar um modelo de *Classificação*, utilizando *indução de árvores de decisão*, para previsão de atrasos e falhas nas quantidades de produtos entregues pelos fornecedores. Através da aplicação de ambas as técnicas será optimizado o processo de abastecimento.

Ainda para a componente prática, será seguida uma metodologia padrão denominada por *CRISP-DM (Cross-Industry Standard Process for Data Mining)*. Este tipo de metodologia, devidamente apresentada no capítulo 4, permite definir um conjunto de etapas *standard* na elaboração de um projecto de MD.

1.3 Estrutura da dissertação

A organização desta dissertação seguiu, em grande medida, a estrutura das actividades que se foram desenvolvendo ao longo do período de trabalho em questão. Assim, para além do presente capítulo, esta dissertação está organizada de acordo com o seguinte esquema:

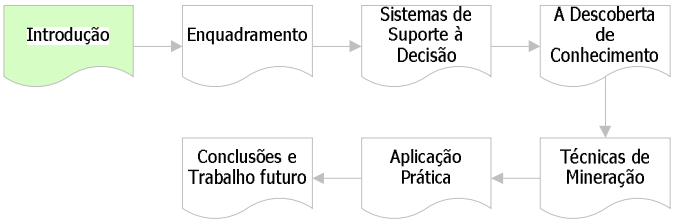


Figura 2 – Divisão dos capítulos do presente documento

Capítulo 2 – Enquadramento

Para melhor estruturar esta dissertação, decidiu-se por incluir na primeira parte do trabalho uma componente introdutória, representativa da problemática associada ao poder da compra e à diversidade interminável de oferta. Serão apresentados, de forma bastante simples (visto este não ser o foco principal deste trabalho), alguns conceitos teóricos associados à gestão da cadeia de abastecimento e, em particular, à função abastecimento.

Capítulo 3 – Sistemas de suporte à decisão

Num tentativa de melhorar o poder de decisão de um gestor do negócio será introduzida a evolução dos sistemas informáticos de suporte à decisão, realçando a mudança do paradigma relacional para o dimensional. Serão apresentados conceitos de *Data Warehousing, Business Intelligence e* também de análise multi dimensional *OLAP*.

Capítulo 4 – A descoberta de conhecimento

Na quarta parte deste trabalho, numa componente essencial para o estudo em causa, será feita uma revisão bibliográfica relativa à descoberta de conhecimento em grandes bases de dados - DCBD. No âmbito da descoberta de conhecimento, a MD será uma das etapas apresentadas e também aqui abordada.

Capítulo 5 – Técnicas de Mineração de Dados

A descoberta de Regras de Associação é uma das técnicas existentes na MD. Neste capítulo apresentar-se-ão os fundamentos teóricos associados a esta técnica assim como também alguns dos seus algoritmos mais estudados por toda a comunidade

científica. Como a componente prática incidirá no algoritmo mais tradicional *Apriori* e suas variantes para a geração das regras, será dado maior ênfase a este algoritmo. Para além das regras de associação será feita uma introdução à tarefa de Classificação estudando e analisando os respectivas técnicas e conceitos para a criação de um modelo. Será relevada a técnica de indução de árvores de decisão e revistos alguns dos seus algoritmos mais abordados na literatura.

Capítulo 6 – Aplicação prática

Para justificar os objectivos definidos para esta dissertação será apresentada uma componente prática sobre dados de compras e recepções relativos a um retalhista, mas devidamente mascarados por motivos de confidencialidade. Nesta componente aplicar-se-á o algoritmo para geração de regras *Apriori* e também o algoritmo baseado em CART para a construção de um modelo de *Árvores de Decisão* existentes na ferramenta *Oracle Data Mining (ODM) 10g*.

Capítulo 7 – Conclusões

Por fim, será feita uma revisão dos objectivos inicialmente propostos assim como uma análise global deste trabalho e considerações finais.

Complementarmente, nos apêndices, incluídos no final desta dissertação, apresentar-se-ão alguns pontos menos detalhados, todavia considerados importantes, estudados ao longo do trabalho.

We are entering an era of unprecedented choice. And that's a good thing.
Anderson Chris

Capítulo 2

Enquadramento funcional

No decorrer dos últimos anos as empresas têm evoluído no sentido de se tornarem físicas e digitais ao mesmo tempo [Carvalho *et al*, (S/D)]. Para Moura (2006) com o evidente desenvolvimento tecnológico, há agora novos modelos de fazer compras, em multi-canais, com vantagens para os compradores e vendedores. Actualmente pode-se comunicar livremente sob redes de informação cada vez mais rápidas e poderosas. Os jovens crescem hoje de uma forma bastante diferente, com acesso à informação em tempo real, sobre *bits e bytes* de dados, com lógica e padrões associados a novo saber, com imensa diversidade de opções. Vive-se numa época digital onde os consumidores são possuidores de mais e melhor informação acerca das ofertas dos diversos competidores e mais canais de compra [Hornick M. *et al*, 2007].

Para justificar esta evolução, à escala global, seleccionou-se um livro, em particular, cujo contexto e conteúdos poderão ser enquadrados na fase inicial deste trabalho. *The Long Tail* é um dos últimos livros de Anderson (2006) que apresenta a forma como no comércio retalhista e grossista moderno, cada vez mais dominado pela *Internet*, a oferta interminável de produtos implica também, cada vez mais, uma procura ilimitada dos mesmos, assim como uma maior quantidade de dados gerados nos sistemas operacionais das organizações.

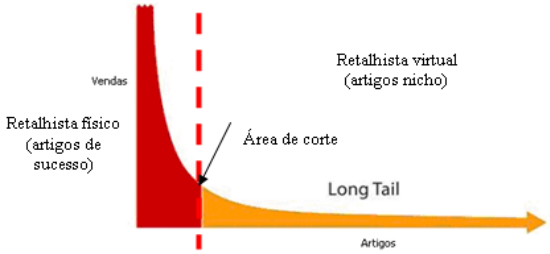


Figura 3 – Artigos nicho e artigos de grande visibilidade [Anderson (2006)]

A teoria que está por detrás deste livro é simples, perfeitamente válida e evidente no nosso dia a dia cada vez mais dominado pelas tecnologias de informação. O autor inicia a sua análise apresentando um paradoxo entre o resultado económico evidente dos grandes sucessos, do inglês *hits* (parte a vermelho no gráfico) e o valor resultante dos nichos de mercado representativo de artigos não considerados sucessos (parte a laranja). Este estudo orienta o leitor para dois cenários distintos limitados pela área de corte. O primeiro corresponde a compras efectuadas em estruturas físicas tais como em hipermercados e o segundo a compras efectuadas por outras vias: catálogos, televisão, *Internet*, entre outros.

Numa analogia curiosa, cuja visualização relembra os conceitos associados à Topologia da Mente de Freud, o autor desenha na mente do leitor um oceano representativo da nossa cultura e várias ilhas espalhadas que correspondem aos produtos mais populares e consequentemente que mais lucro permitirão gerar. A divisão entre as profundidades do mar e as ilhas é considerado o limiar de vendas necessárias para satisfazer os canais de distribuição. As ilhas são apenas pontas de grandes montanhas. Quando o custo da distribuição baixa, ou o nível das águas baixa, então é naturalmente provável que mais e mais produtos surjam à tona da água.

É evidente que a maioria dos retalhistas e grossistas actuais, possuidores de apenas estruturas físicas apenas se concentram na zona a vermelho. De facto a zona a vermelho oferece-lhes, à partida, uma garantia de venda desses artigos expostos nas prateleiras dado que são artigos cuja popularidade oferece intuitivamente garantia de compra.

As limitações físicas destas superfícies, no entanto, imagine-se um hipermercado, não permitem que a zona dedicada aos artigos de música seja uma zona ampla o suficiente para cobrir toda a quantidade de músicas procuradas. Para colmatar estas limitações existem então lojas físicas dedicadas apenas a artigos de música cujo espaço já permite colocar outro tipo de categorias, para outros gostos, nas suas prateleiras. Nestas superfícies específicas temos mais produtos e mais opções, mais oferta e consequentemente mais compras.

Muito embora este tipo de superfíces consiga cobrir segmentos mais específicos de consumidores, apresentando quantidades mais significativas de produtos, de facto nunca conseguirá expor nas suas prateleiras aquele produto com especificidades muito particulares (que se consideram nichos) para consumidores muito específicos. Visualizando o gráfico, com particular detalhe, é então evidente que a zona a laranja se extende à medida que mais produtos são colocados à venda. O conceito de loja virtual, cuja designação ficou anexada à terminologia inglesa *e-commerce*, não é um conceito novo e surgiu após a *Internet* ter "permitido" o comércio *online* em meados da década de noventa.

Uma loja virtual é uma loja onde uma determinada organização comunica directamente com os seus consumidores através de um computador. Por outro lado uma loja convencional, física, é restringida pela sua localização, o seu *layout*, pelos seus componentes operacionais (pessoal das vendas, as caixas de pagamento, prateleiras, etc.). Para além disso os consumidores necessitam de visitar fisicamente as lojas e somente quando as mesmas estão abertas. Esta é a grande vantagem das lojas virtuais – não são dependentes do tempo [Newman *et al* (2002)].

Voltando ao foco deste enquadramento, Anderson (2006) apresenta vários exemplos relativos às diferenças entre os artigos disponibilizados numa superfície física e noutra apenas virtual e os impactos de tais exposições nas vendas finais:

- No primeiro caso, num hipermercado de renome mundial, onde se vendem imensos artigos espalhados por toda a estrutura, existe uma parte física inteiramente dedicada à venda de artigos de música. Nesta superfície encontram-se expostos 4500 albums totalizando 25000.
- O retalhista virtual, por outro lado, vende todo o tipo de artigos de música, de nichos supostamente desconhecidos e também os grandes êxitos, que podem ser encomendados pelos seus clientes. Este vendedor virtual não possui uma estrutura física limitada, aparentemente, e por isso disponibiliza neste momento 1.5 milhões de músicas.

A pergunta que colocam ao autor no decorrer deste estudo tem a ver com o facto de, havendo uma quantidade interminável de produtos no segundo caso, se esses mesmos produtos seriam realmente vendáveis em cada quarto de ano. A resposta obtida para esta pergunta foi deveras surpreendente apontando para valores a rondar os 98% de certeza de que todos os artigos eram vendidos pelo menos uma vez em cada período de 3 meses. Se fosse gerado um gráfico para cada uma das lojas aqui apresentadas, seguramente a zona a laranja correspondente à loja virtual seria muito mais extensa. A conclusão que o autor retira daqui é clara e bastante pertinente. Em primeiro lugar há sempre consumidores para todo o tipo de artigos, mesmo sendo artigos de micro nichos. Olhando para zona laranja percebe-se que com o crescimento do número de artigos todos eles se vendem mesmo com valores reduzidos. Em segundo lugar, em termos financeiros, é evidente que à medida que a zona a laranja cresce, as vendas dos nichos crescem, e em consequência disso há um lucro bastante alto associado às vendas totais deste tipo de artigos.

Em termos informáticos este crescimento também tem um impacto enorme, bastante relevante neste estudo: os grandes retalhistas/grossistas não terão apenas estruturas físicas, mas também virtuais se quiserem competir em mercados e nichos mais específicos; torna-se também evidente que, no futuro, devido à facilidade evidente de apresentar as suas gamas de produtos, consequentemente também existirão mais

produtos, mais encomendas, mais vendas e maior volume de dados gerados diariamente directamente nas bases de dados operacionais. Consequentemente haverá uma necessidade de optimizar as cadeias de abastecimento.

2.1 Optimização da cadeia de abastecimento

Independentemente do tipo de estrutura onde o consumidor compre há a necessidade de abastecer sucessivamente as lojas físicas e/ou os armazéns para satisfazer a procura. Sendo o consumidor final o interveniente final de toda a gestão da cadeia, Moura (2006) sublinha que a optimização da função abastecimento é um dos desafios mais relevantes no seio de uma organização que permite tornar mais eficiente o serviço prestado ao cliente. A função abastecimento consiste "na obtenção de produtos e materiais de fornecedores internos e externos destinados à produção, ao consumo ou para revenda. Inclui todas as actividades necessárias para a disponibilização dos bens certos, no momento adequado e de forma económica". De entre estas actividades destacam-se a localização e selecção de fornecedores, a negociação de contratos, a emissão e acompanhamento de encomendas, entre outras. Este processo termina com a recepção da encomenda, nos prazos estipulados, nas quantidades certas e com a qualidade acordada entre ambas as partes [Moura, 2006].

A selecção de fornecedores é, de facto, uma das etapas mais importantes na gestão da cadeia de abastecimento. A falta de cuidado neste processo implica perdas financeiras para todo o tipo de organizações [Haery *et al*, 2008]. Depois de analisar vários trabalhos realizados por Dickson, Weber *et al*, De Boer *et al* e Morlacchi., Haery constatou que os estudos efectuados até então são limitadores, uma vez que apenas se focavam em variáveis de forma independente. Este aspecto não é, de todo, praticável numa realidade cada vez mais complexa e exigente.

A decisão pela selecção de fornecedores é uma tarefa complicada pelo facto de que existem imensos critérios a serem considerados em todo o processo de decisão. Para complicar mais este facto, a maioria dos fornecedores poderão ter diferentes

desempenhos relativamente a cada um desses critérios. Por exemplo, o fornecedor que tiver a melhor qualidade de entrega poderá não ter o melhor desempenho nos prazos de entrega. Por outro lado, o fornecedor que entregue sempre nos prazos estipulados poderá não conseguir responder aos pedidos de quantidade [Sucky E., S/D].

Carravilha & Oliveira (2000), seguindo a mesma esteira de pensamento, num estudo realizado sobre a temática da gestão da cadeia de abastecimento, constataram também que é importante atender a três factores fundamentais aquando do processo de tomada de decisão relativa a fornecedores: a instituição que irá fornecer os bens ou materiais (localização, gestão e estabilidade financeira); o produto que irá ser encomendado (qualidade e preço) e o serviço prestado por essa instituição (entrega atempada e quantidade adequada).

Por ser um dos processos mais importantes dentro da cadeia de abastecimento, a selecção de fornecedores é um desafio evidente, diário, para todos os responsáveis da organização. Todavia, não é fácil decidir pelo melhor fornecedor, ou pelas quantidades certas de encomenda, ou até mesmo pelos prazos certos, de forma a que todo este processo permita maximizar o lucro não impactando a satisfação final do consumidor.

O tema que se propõe para esta dissertação tem objectivos bem definidos e claros que visam optimizar os processos de abastecimento, utilizando sistemas computacionais de suporte à tomada de decisão, nomeadamente Técnicas de MD.

The quality of decisions is like the well-timed swoop of a falcon
which enables it to strike and destroy its victim
Sun Tzu

Capítulo 3

Sistemas de Suporte à Decisão

Numa breve e curiosa leitura efectuada relativamente ao poder da decisão e das dificuldades que essa tarefa acarreta, Girão I. *et al* (2000) definem o verbo decidir como sendo a capacidade de "*resolver, deliberar, emitir algum juízo, fazer escolhas. Portanto, implica pensar, julgar e agir, três componentes que caracterizam o processo de decisão*".

Como em tudo na vida tomar uma decisão não é uma tarefa fácil, porque na realidade implica bastante raciocínio e bom senso, para além de todo o risco associado. A decisão é uma actividade banal nos dias que correm e todos nós somos obrigados a decidir por diversos caminhos profissionais e pessoais ao longo da nossa vida. Numa vertente organizacional a tomada de decisão torna-se cada vez mais complexa e exigente devido à diversidade de opções e dos impactos que tais escolhas terão no seio da própria empresa.

Para apoiar tais decisões foram aparecendo ao longo dos anos sistemas computacionais capazes de auxiliar este e outros processos – denominados por sistemas de suporte à decisão (da terminologia inglesa *Decision Support Systems – DSS*). Alguns estudos relacionados com os sistemas de suporte tiveram início a meio da década de sessenta com um conjunto de investigadores a ter um papel relevante na construção de sistemas de suporte baseado em modelos. Num desses estudos Power *et al* (S/D), citando

diversos autores, afirmam que desde a década de cinquenta até meados da década de sessenta o conceito de suporte à decisão evolui devido aos trabalhos teóricos efectuados no departamento de tecnologia da Universidade de *Carnegie* e, posteriormente à década de sessenta, devido aos trabalhos mais técnicos em sistemas interactivos de computadores efectuados na mesma universidade.

De acordo com Sprague e Watson [cit. Power (S/D)], em meados de 1970 alguns jornais de negócio começaram então a publicar artigos relacionados com sistemas de decisão para gestão, sistemas de decisão para planeamento estratégico e também sistemas de suporte à decisão. Alguns nomes e trabalhos nesta área foram apresentados por Scott Morgan, Ferguson e Jones, entre outros nomes de referência no âmbito dos sistemas de suporte.

Para Scott Morton (1970), um sistema de suporte à decisão corresponde a um sistema computacional interactivo que auxilia os agentes de decisão a utilizar dados e modelos para resolver problemas não estruturados.

Para Keen & Scott Morton (1978), os sistemas de suporte agrupam os conhecimentos dos indivíduos com as capacidades computacionais de forma a melhorar a qualidade das decisões. São sistemas computacionais de suporte à gestão das decisões dos decisores que lidam com problemas semi-estruturados.

3.1 Sistemas Operacionais e Sistemas Analíticos

Ao longo da nossa formação académica, durante a escola e universidade, foram-nos apresentados conceitos associados às bases de dados tradicionais, com modelos E-R (entidade/relação) e a normalização nas suas formas normais. Durante estes anos aprendem-se as bases teórico-práticas para o desenvolvimento de sistemas de bases de dados cuja modelação é estritamente relacional. Os sistemas operacionais, denominados por sistemas OLTP (*On-line Transaction Processing*), têm no seu núcleo modelos

relacionais puros. Este tipo de sistemas, tal como o nome indica, aplicam-se em ambientes transaccionais onde a componente analítica não deveria ser misturada.

Ao contrário deste tipo de repositórios relacionais que existem em praticamente todas as organizações e cujo principal objectivo é o de efectuar operações diárias do próprio negócio, muitas organizações possuem já sistemas de suporte à decisão realmente analíticos, sistemas distintos, denominados também por sistemas de *Data Warehousing* e *Business Intelligence*.

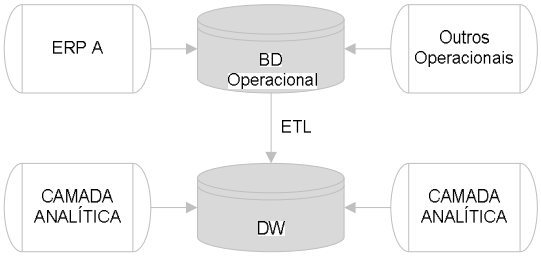


Figura 4 - Sistema Operacional vs Sistema Analítico

Para Inmon (1996) há várias razões evidentes para a existência desta separação entre sistemas operacionais de sistemas analíticos:

- Os dados que servem as necessidades operacionais são fisicamente diferentes daqueles que servem propósitos analíticos.
- As tecnologias suportadas por ambos são distintas.
- Os utilizadores finais operacionais são distintos dos analistas.
- As formas de processamento de ambos os sistemas são fundamentalmente diferentes.

Este tipo de repositórios analíticos, cujo volume poderá atingir valores inimagináveis e impossíveis de serem manipulados manualmente, foram criados a pensar num agente final de decisão. Para Berry *et al* (2000), durante o séc. XX, o volume de informação gerado e mantido por algumas empresas cresceu cerca de 100.000 vezes.

3.2 Os *Data Warehouses*

Na verdade, praticamente todas as empresas da actualidade possuem já um sistema capaz de armazenar informação operacional do seu negócio. Certamente, devido à dimensão do seu próprio negócio, não possuem e nem precisam de poderosos sistemas analíticos. Não obstante isto, é um facto que as grandes organizações da vanguarda mundial possuem uma quantidade imensa de dados, quer eles estejam em sistemas operacionais ou mesmo em grandes repositórios analíticos – denominados tradicionalmente por *Data Warehouses.*

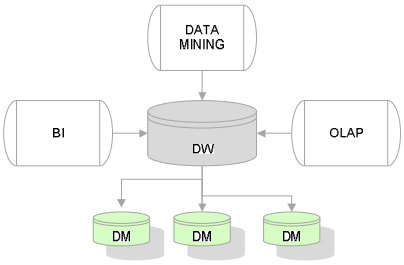


Figura 5 – *Data Warehouse* e Componentes Analíticas

Para Kimball R. (2002), percursor de grande parte dos conceitos associados a *Data Warehousing*, um DW corresponde a um repositório de dados, centralizado, que serve de suporte à tomada de decisão, composto por domínios distintos denominados por *Data Marts* (DM). Este tipo de sistemas analíticos seguem um paradigma de modelação dimensional relacional, bem distinto dos tradicionais sistemas cujas modelações são puramente relacionais. Um modelo dimensional pode ser representado de diversas formas, sendo as mais tradicionais as *estrela* e *floco de neve*, tal como está ilustrado nas duas imagens seguintes.

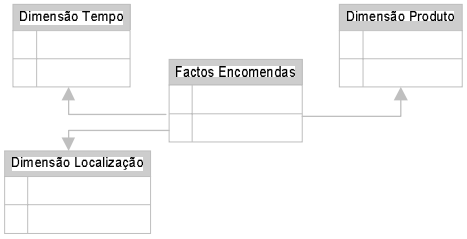


Figura 6 - Exemplo de dois tipos de Modelo Dimensional – Estrela

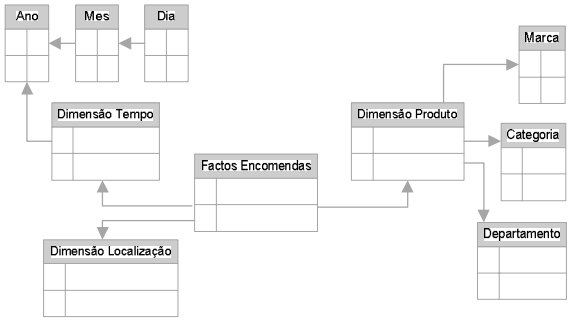


Figura 7 - Exemplo de dois tipos de Modelo Dimensional – Floco de Neve

Como são sistemas geralmente alimentados diariamente através de processos designados por ETL (*Extract Transform and Loading*), desenhados a pensar em analistas de negócio e nunca nos intervenientes operacionais do próprio negócio, os DW são sistemas que normalmente suportam gigantescos volumes de dados, estritamente focados nas análises dos dados. Os processos de ETL são um dos pontos críticos na concepção de um DW e podem ser caracterizados por três grandes fases: (i) a extracção de dados dos sistemas origem; (ii) a transformação dos mesmos; (iii) o carregamento para o repositório final. Dado que neste trabalho se pretendem aplicar técnicas de MD, devidamente analisadas no capítulo 5 e 6, pode-se desde já afirmar que, se esta etapa de ETL for devidamente efectuada, então o processo de pre-processamento para a tarefa de MD ficará também mais simplificado. Por outro lado, se este processo não estiver optimizado, então diversos problemas como ruído, valores inconsistentes e nulos poderão advir devido a um mau planeamento inicial do próprio DW.

Na literatura existem imensas abordagens para o conceito de DW. Para Inmon (1996), por exemplo, corresponde a uma colecção de dados orientados por assuntos, integrados, variáveis com o tempo e não voláteis, para apoiar o processo de tomada de decisão:

Orientado por assuntos:

- Organizado por assuntos importantes, tais como cliente, produto e vendas.
- Focado na modelação e análise de dados para a tomada de decisões, em vez de operações diárias e processamento de transacções.
- Fornece uma visão simples e concisa sobre questões de um tema particular através da exclusão de dados que não são importantes no suporte ao processo de decisão.

Variável Tempo:

- O horizonte de tempo para um DW é significativamente maior do que o de sistemas operacionais.
- Base de dados operacional: informação actual.
- Dados no DW: fornece informação numa perspectiva histórica (ex. últimos 5-10 anos).
- Cada estrutura chave no DW contém um elemento de tempo, explícita ou implcita, mas a chave de dados operacionais pode ou não conter um "elemento de tempo".

Integrado:

- Construído por integração de múltiplas e heterogéneas fontes de dados.
- Bases de dados relacionais, ficheiros simples, registos de transacções *on-line.*
- São aplicadas técnicas de limpeza de dados e integração de dados.

- É assegurada a consistência na convenção de nomes, codificação de estruturas, atributos de medidas, etc. entre diferentes fontes de dados.

Não Volátil:

- Um repositório fisicamente separado de dados transformados do ambiente operacional.
- Não ocorre actualização de dados operacional sobre a informação no DW.
- não requer mecanismos de processamento de transacções, recuperação e controlo de concorrência.
- Requer apenas duas operações de acesso a dados: carregamento inicial de dados e acesso a dados.

Neste momento, a título de curiosidade, a *Sybase*[1], numa notícia recente, estabeleceu um novo *record* do *Guinness World Record* permitindo que um DW de 1000 *terabytes* (1 *Petabyte*) seja suportado por um dos seus novos servidores. Este DW permitirá guardar 6 triliões de linhas de dados transaccionais e mais de 185 milhões de documentos com conteúdo pesquisável [www2].

3.3 A *Business Intelligence*

Em qualquer organização tomam-se imensas decisões diariamente. Este tipo de decisões, na maioria das vezes, são baseadas maioritariamente na experiência de quem as tomou ou então baseada em factos. Sabendo que a experiência e o conhecimento levam anos a aquirir então é necessário abordar a questão da decisão perante factos concretos, algo que possibilite a um decisor obter informação rapidamente [Ritacco M. & Carver A. (S/D)].

[1] A Sybase é uma das maiores empresa de software no Mundo focada nas áreas da Gestão da Informação, Desenvolvimento e Integração, Mobilidade e RFID, e Open Source.

Os DW's servem de repositórios centralizado de dados, tal como já foi apresentado, representando uma única versão da realidade de uma organização. No entanto, um repositório de dados é inútil se sobre ele não existir um mecanismo que permita extrair estes dados e apresentá-los sobre a forma de conhecimento a um agente final. Por esse motivo surgiu na década de 80 um novo conceito, geralmente associado a *Data Warehousing*, denominado por *Business Intelligence*. A BI é um termo que teve origem no *Gartner Group* e traduz-se num conjunto de etapas que permitem extrair dados dos repositórios empresariais, "transformá-los" em informação útil para a tomada de decisão e apresentá-los ao decisor final de uma forma simples e intuitiva. Em conjunto, as áreas de DW e BI permitem que os analistas e agentes finais de decisão antecipem tendências através da análise dos dados passados.

3.4 Os sistemas Multi Dimensionais OLAP – *Online Analytical Processing*

As ferramentas de BI permitem extrair informação de repositórios analíticos apresentando-a a analistas finais de negócio de forma intuitiva e organizada. Um dos principais desafios por detrás destas ferramentas tem a ver com o desempenho final obtido na apresentação de relatórios finais aos decisores de negócio. Neste sentido existem várias abordagens para a BI: (a) através da utilização de mapeamentos directos às bases de dados (ROLAP), extraindo informação directa, mas com piores desempenhos; (b) através de estruturas multi dimensionais previamente agregadas (MOLAP).

Na actualidade, em qualquer sistema de DW, a utilização de MOLAP é fundamental, mas não obrigatória. Kimball R. (2007) refere que a opção por utilizar MOLAP ou simplesmente um modelo dimensional relacional é uma decisão de negócio. A grande diferença será, com certeza, a melhoria substancial de desempenho na apresentação dos resultados finais devido às estruturas previamente agregadas.

A modelação multi dimensional permite que os dados sejam organizados em estruturas denominadas por cubos, onde os dados podem ser guardados e posteriormente visualizados sob diversos níveis de agregação. Estas estruturas, geralmente associadas ao conceito OLAP, neste caso MOLAP, têm como principal vantagem a rapidez de acesso aos dados, já que eles estão previamente consolidados sobre estruturas cúbicas.

Para Kimball R. (2007) o conceito de OLAP tem as seguintes propriedades**:**

- Muito melhor desempenho que um sistema relacional quando os cubos são desenhados correctamente, com menor necessidade de se proceder a complexas optimizações.
- Muitas mais capacidades analíticas do que o relacional.
- As ferramentas de *OLAP* são bastante superiores às ferramentas relacionais.
- A *OLAP* não sofre de problemas de limitações do próprio *SQL*.
- Quando há a necessidade de reconstruir um cubo é um processo demasiado lento.
- Os vendedores de soluções *OLAP* têm certas limitações que não existem no relacional, incluindo o número de membros nas dimensões, o número de valores distintos nos vários níveis da hierarquia e o tamanho global do cubo.

Não sendo este o foco principal deste estudo, apresentam-se de seguida algumas definições proferidas por alguns autores relativas a OLAP:

- Para Navega S. (2002) este conceito corresponde a uma categoria de *software* que permite aos analistas, gestores e executivos ter acesso à informação de forma bastante rápida, consistente e numa perspectiva multidimensional.
- Para o OLAP Council [www8] "OLAP (*On-line Analytical Processing)* corresponde a uma categoria de *software* que permite aos gestores, analistas e

executivos a análise dos dados, através de um acesso interactivo, consistente e rápido, de uma grande variedade de vistas possíveis".

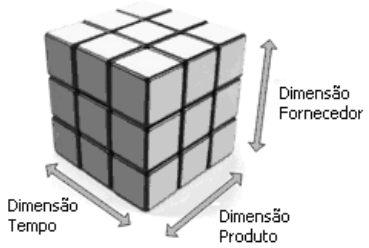


Figura 8 – Representação multi dimensional em forma de cubo

Ainda segundo Kimball R. (1996) pode-se navegar sobre estas estruturas de diversas formas, através de mecanismos de *drilling down, roll up, drilling across e drilling around*. Desta forma, utilizando as ferramentas de BI, será possível apresentar ao analista final do negócio uma única versão do seu negócio a vários níveis hierárquicos.

The key to success in business is to know something that nobody knows.
Aristotle Onassis

Capítulo 4

Descoberta de conhecimento

As tecnologias de informação têm um poder incrível nas mudanças de filosofias e pensamentos ao longo dos tempos. Antigamente os dados eram praticamente inexistentes do ponto de vista computacional e o conhecimento existente era fruto da experiência dos próprios trabalhadores ou então com suporte a qualquer outro tipo de sistema inadequado. Hoje em dia, as tecnologias permitem que todo o tipo de movimentações operacionais seja guardada numa qualquer base de dados, operacional ou analítica, e o conhecimento do próprio negócio seja melhorado com o auxílio a ferramentas que permitam extrair esses dados.

No capítulo anterior distinguiu-se brevemente o paradigma operacional do dimensional, focando os sistemas de *DW e BI*, passando pelos cubos multidimensionais de dados (MOLAP) cujo objectivo principal é a melhoria do desempenho através do armazenamento prévio dos dados em estruturas distintas, denominadas por cubos. Através de sistemas deste tipo as organizações têm uma versão centralizada e estratégica do seu negócio.

Para Brijs (2002) a vantagem competitiva não será conseguida utilizando apenas a informação conhecida nas bases de dados, mas sim pela capacidade de extrair, dos mesmos sistemas, o conhecimento desconhecido. Neste seguimento, o presente capítulo introduz novos conceitos associados com a descoberta de conhecimento em grandes bases de dados (DCBD) e também com a visualização dos resultados.

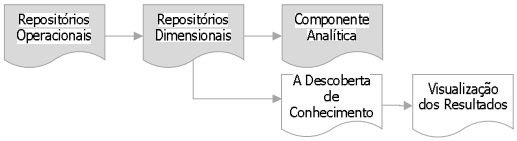


Figura 9 – Sequência das etapas apresentadas no capítulo 4

Os repositórios analíticos apresentados no capítulo anterior permitem obter conhecimento directo a partir dos dados brutos, ou por extracção directa de dados em tabelas ou sobre estruturas multi dimensionais utilizando ferramentas apropriadas de BI. No entanto, para Brijs (2002) e Piatetsky–Shapiro G. (1991), dentro dessas estruturas existe também uma oportunidade de descobrir novo conhecimento.

Para Piatetsky–Shapiro G. (1992), num artigo publicado posteriormente, estima-se que o aumento de informação no mundo duplique a cada 20 meses e, muito provavelmente, o número de bases de dados continue também a aumentar a um ritmo ainda mais rápido. Por este motivo, seguindo a mesma linha de pensamento de Anderson (2006) e devido à quantidade de informação cada vez maior (gerada sobre lojas físicas e virtuais), torna-se ainda mais evidente a necessidade de procurar padrões de conhecimento oculto.

Surge então o conceito de D*escoberta de Conhecimento em Bases de Dados* (da terminologia inglesa *KDD - Knowledge Discovery in Databases*). Para Piatetsky-Shapiro a DCBD corresponde *"à extracção não trivial de informação implícita, anteriormente desconhecida, potencialmente útil, existente nos dados"* ou então, segundo Fayyad U., Piatetsky-Shapiro G. & Smyth P. (1996), a "*um processo não trivial de informação de padrões*[2] *embutidos nos dados que sejam válidos, novos, potencialmente úteis e compreensíveis*".

[2] Um padrão interessante é considerado *conhecimento* e o resultado obtido pela execução de uma aplicação geradora de padrões é considerado *conhecimento descoberto* [Piatetsky-Shapiro *G.* (1992)].

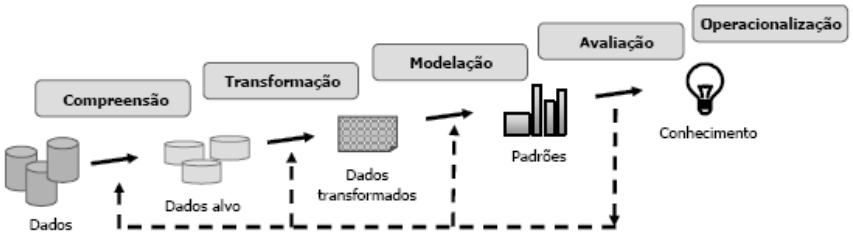


Figura 10 – Etapas da DCBD [adaptado de Piatetsky-Shapiro G. & Smyth P. (1996)]

Na figura anterior pode-se considerar que a descoberta de conhecimento dentro de bases de dados é um processo sequencial e interactivo que permite extrair padrões de conhecimento interessantes cujo critério de interesse é definido pelo próprio utilizador.

De acordo com Cabena *et al* (1997) o tempo e o esforço dispendidos nas etapas não é equitativo. Segundo os mesmos autores pode-se representar este esforço através da visualização do seguinte gráfico:

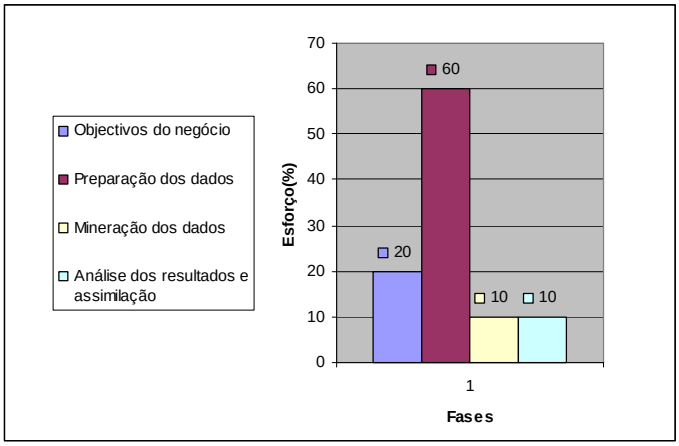


Figura 11 – Esforço dispendido nas etapas da DCBD [adaptado de Cabena *et al* (1997)]

4.1 Etapas gerais para a descoberta de conhecimento

Da literatura analisada existem bastantes estudos aplicados no comércio retalhista e grossista, sobre lojas físicas e virtuais, que relevam a importância da descoberta de

conhecimento dentro das bases de dados das organizações. Para além da bibliografia referida ao longo deste capítulo, salientam-se ainda os seguintes estudos: [Kusiak, 2007], [Brijs, 1999], [Brijs 2002], [Brijs, S/D], [Mladenic , S/D], [Borges, S/D] e [Kohavi, S/D].

Existem diversas etapas prévias que deverão ser aplicadas sobre um conjunto de dados brutos para permitir, sem quaisquer problemas, a descoberta de conhecimento dentro de bases de dados: alguns exemplos são o pré processamento (*extracção, limpeza e selecção*), a transformação dos dados, a MD e a interpretação dos resultados.

4.1.1 Pre processamento - Extracção, limpeza e selecção

Segundo Han *et al* (2001) as bases de dados do mundo real são altamente susceptíveis a ruído, a valores em falta, dados inconsistentes devido ao seu elevado volume de dados. De forma a atingir o máximo benefício da aplicação de um algoritmo de MD (conceito apresentado posteriormente) sobre um conjunto de dados (também denominado na terminologia inglesa por *dataset*), o pre processamento dos dados é necessário de forma a garantir integridade e fiabilidade dos dados. As tarefas básicas de pre processamento incluem a limpeza, a transformação, a integração, a redução e discretização dos dados [retirado de Larose 2005 cit. Kennedy 1998 e Pyle 1999].

Imagine-se um exemplo associado aos próprios processos de ETL para o carregamento de um domínio específico de um DW: um utilizador operacional efectua uma determinada acção no seu sistema operacional – considere-se um ERP[3]. Talvez por comodismo ou falta de atenção, não preenche todos os campos existentes para esse tipo de acção. Por exemplo, efectua o registo de um novo fornecedor sem indicar a cidade do mesmo. No final do dia, quando os processos de ETL forem executados, essa

[3] ERP – Enterprise Resource Planning – sigla correspondente a uma aplicação centralizada, operacional, que permite gerir toda a organização através da unificação de processos.

informação nunca virá para o DW, porque realmente nunca foi criada no sistema origem.

Ainda segundo Han *et al* (2001) existem vários mecanismos para preencher estes valores em falta: ignorando o tuplo, preenchendo manualmente o campo em falta, usando uma variável constante global, usando a média, usando o valor mais provável (através de regressão ou indução por árvores de decisão para calcular esses valores em falta, por exemplo).

Para além dos valores em falta, o ruído é outra preocupação. Considere-se, por exemplo uma determinada variável que poderá tomar imensos valores. Para ultrapassar este problema existem vários mecanismos que permitem "suavizar" os valores: através da discretização de valores (da terminologia inglesa *Binning),* através da segmentação *(da terminologia inglesa Clustering),* através da combinação de inspecções por computador e humanos e regressão.

Transformação e integração

Ainda na componente de transformação e integração os dados são também preparados para a aplicação de técnicas de extracção de conhecimento. Para Han *et al* (2001) existem também várias técnicas associadas:

- De igual forma através da "suavização" dos dados, através da discretização, da segmentação e regressão.
- Através da agregação, utilizando níveis estruturais mais elevados.
- Generalização dos dados, através da utilização das hierarquias. Por exemplo num atributo categórico a descrição do produto poderá ser generalizado para um nível mais alto como a categoria desse produto.
- Normalizando valores dentro de um determinado intervalo, por exemplo: 0.0 a 1.0.

- Criando novos atributos para enriquecimento do conjunto de dados e ajudar no processo de descoberta de conhecimento.

Redução dos dados

Um dos principais desafios relacionados com DW tem a ver com o facto de se trabalhar com volumes elevados espalhados pelos diversos domínios do próprio repositório. A técnica de redução dos dados aplica-se neste contexto, onde nem todas as variáveis serão necessárias para induzir apenas o conhecimento necessário. Com a redução dos dados obtêm-se sub conjuntos bem mais pequenos mantendo a integridade e fiabilidade do conjunto inicial. Para Han *et al* (2001) existem imensas técnicas para redução dos dados:

- Através da construção de um cubo de dados.
- Removendo dimensões e atributos irrelevantes.
- Fazendo a compressão dos dados através de mecanismos apropriados.
- Utilizando a discretização dos dados.

Selecção de atributos

A selecção dos atributos mais relevantes de um conjunto de dados é uma das etapas mais importantes na aplicação de um problema de Classificação e quando se tenta criar um modelo onde nem todas as variáveis são interessantes. Se para a criação de um modelo forem utilizadas todas as variáveis então mais processamento e memória serão necessários para a construção do mesmo, assim como possivelmente maior ruído e redundância. Na seguinte figura, sugerida por Liu H. (2003), apresenta-se um diagrama de fluxos simplificado e representativo deste processo:

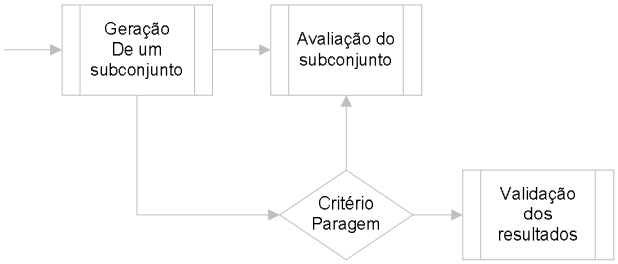


Figura 12 - Passos para a selecção de atributos

4.1.2 A Mineração de Dados

Para Barry *et al* (1997), com o crescimento exponencial dos dados armazenados, o desenvolvimento tecnológico e o aumento do conhecimento nas áreas de *machine learning*, os retalhistas e grossistas estão agora perante um cenário mais facilitado para a tomada de decisão utilizando a etapa de MD. Pode-se então considerar que a MD poderá ser um novo passo, um passo em frente, porventura acompanhando a BI, ou posterior, correspondendo à descoberta de conhecimento não evidente. Minerar dados é um processo exigente onde o grande desafio está relacionado com o imenso volume de dados e à incapacidade humana de manipular tais volumes sem o auxílio das próprias tecnologias de informação.

A motivação principal para a utilização da MD são a redução de custos, o aumento dos lucros, a descoberta de informação desconhecida e vantajosa, automatizar algumas tarefas humanas, identificar fraude e melhorar o serviço dos clientes. Segundo Hornick *et al* (2007) traduz-se numa vantagem competitiva não só para as indústrias retalhistas, mas para qualquer outro tipo de organizações. A MD corresponde a uma conjugação de diversas disciplinas, tal como representa a seguinte figura apresentada por Symeonidis *et al* (2005).

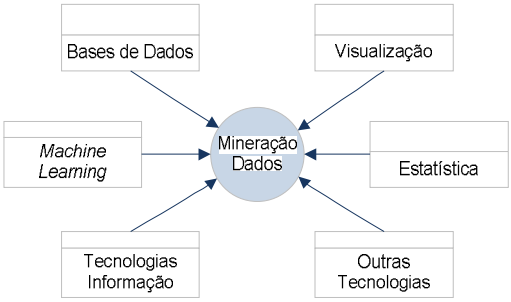


Figura 13 – Disciplinas associadas à MD [adaptado de Symeonidis *et al* (2005)]

Symeonidis *et al* (2005) apresentam também um quadro elucidativo da evolução dos sistemas existentes na década de sessenta até à presente decada, assim como as suas características.

Descrição (época)	Tecnologias
Armazenamento de dados (60's)	*Computadores, fitas, disquetes*
Gestão de dados (70's)	*DBMS, RDBMS*
Acesso aos dados (80's)	*RDBMS, SQL, ODBC*
DW's e SSD	*DW, OLAP*
MD	*Algoritmos avançados, multi processadores, grandes volumes de dados*

Quadro 1 – Evolução até à MD [adaptado de [Symeonidis *et al* (2005)]

Os mesmos autores, para além da evidenciarem novamente a incapacidade humana de manipular e interpretar quantidades enormes de dados, destacam dois outros factores que a caracterizam:

- Ao contrário dos *Sistemas de Suporte à Decisão* – SSD (da terminologia inglesa *Decision Suport Systems* - *DSS*), as técnicas da MD são *computer-driven* e, portanto, perfeitamente parametrizáveis.

- Permite a visualização e o entendimento de grandes conjuntos de dados de forma eficiente e simples.

Para descrever a MD e abordar as suas técnicas apresentam-se de seguida algumas definições proferidas por diversos autores:

- Para Piatetsky-Shapiro G. & Smyth P. (1996) a MD é um passo do processo de DCBD que consiste na análise dos dados e na aplicação de algoritmos capazes de extrair um conjunto de padrões ou modelos dos dados.
- Por sua vez, Witten *et al* (2005) definem a MD como sendo a descoberta de padrões nos dados cujo significado se traduz em alguma vantagem, usualmente económica.
- Berry e Linoff (1997) consideram que a MD corresponde ao processo de exploração e análise, de forma automática ou semi automática, de grandes quantidades de dados, com o objectivo da descoberta de novos padrões e regras que possuam algum significado.
- Para Cabena *et al* (1997) a MD é o processo de extracção de informação desconhecida, válida, que pode ser utilizada para a tomada de decisões cruciais para o negócio.
- Segundo Goebel *et al* a MD (1999) corresponde ao conjunto de técnicas que permitem descobrir informação estratégica escondida dentro de bases de dados de grandes dimensões.
- Por fim, para o Gartner Group cit. Larose (2005), a MD é o processo da descoberta de correlações com significado, padrões e tendências através da análise detalhada de grandes volumes de dados alojados em repositórios, utilizando tecnologias de reconhecimento de padrões e também técnicas estatísticas e matemáticas.

Para além da definição de MD, Navega S. (2002) define padrões como unidades de informação que se repetem, ou então como sequências de informação que dispõem de uma estrutura que se repete. O seu estudo apresenta um exemplo bastante simples, sobre uma sequência de letras, para indução de regras abstractas.Imagine-se a seguinte sequência original: *ABCXYABCZKABDKCABCTUABEWLABCWO.*

Para o autor a primeira etapa será tentar descobrir se existem padrões de letras que se repetem com alguma frequência. As sequências "AB" e "ABC" ocorrem com mais frequência do que as restantes. Após serem determinadas estas sequências verifica-se que elas segmentam o padrão original em diversas unidades independentes: *"ABCXY" "ABCZK" "ABDKC" "ABCTU" "ABEWL" "ABCWO"*.

No final fazem-se induções, que permitem gerar algumas regras genéricas indutivas e reduzir a informação original a algumas expressões simples: *"ABC??" "ABD??" "ABE??" e "AB???"*, onde o caractér '?' poderá representar qualquer letra.

Destas expressões genéricas, aproveitando a última sequência "AB???" pode-se inferir que sempre que encontramos AB na sequência original encontrar-se-ão mais três caracteres até completar o chamado *padrão*.

As letras representadas anteriormente, sendo abstractas, podem tomar diversos domínios. Pode-se, por exemplo, definir as letras como sendo produtos de uma loja. Neste caso quando o consumidor compra o produto A então também compra B. Outro exemplo seria: quando o navegador de um sítio na *Internet* visita a página "*A.html*" "então visita também a página "*B.html"*.

Para Piatetsky-Shapiro *et al* (1991) um padrão é considerado *interessante* se é facilmente compreensível por humanos, válido em dados novos ou dados de teste com algum grau de certeza, potencialmente útil ou se valida alguma hipótese que um utilizador necessita de confirmação.

As tarefas da MD

A MD pode ser dividida em duas vertentes: *predictiva e descritiva*. A vertente predictiva corresponde à determinação de um modelo capaz de prever um novo valor. Relativamente à vertente descritiva permite descrever um conjunto de dados de forma concisa apresentando características interessantes dos dados. O seguinte quadro enquadra as diferentes técnicas existentes de acordo com a sua vertente:

Predictiva	Descritiva
Classificação	*Segmentação*
Regressão	*Associação*
	Sumariação
	Modelação de Dependências
	Detecção de Alterações e Divergências

Quadro 2 - Enquadramento das tarefas nas vertentes *descritiva* e *predictiva*

Para Piatetsky-Shapiro G. & Smyth P. (1996) a *previsão* involve a utilização de variáveis e campos da base de dados para prever valores futuros, valores desconhecidos ou outras variáveis de interesse. Por outro lado *descrição* foca-se inteiramente na descoberta de padrões interpretáveis que descrevem os dados.

Standards para a MD

Existe, neste momento, um conjunto de *standards* estritamente relacionados com a tarefa de MD. Um *standard* corresponde a uma forma de fazer algo repetidamente, a um documento com especificações e critérios bem definidos que servirão como regra, guia ou uma boa definição. Neste caso específico, um quadro com os *standards* existentes poderão ser consultadas no quadro apresentado no apêndice 1. Para a presente dissertação consideram-se relevantes o *standard* CRISP-DM que permite definir um

conjunto de passos para elaborar um projecto de MD e também o PMML como um mecanismo bastante útil para guardar modelos sob a forma da XML.

CRISP-DM

O *CRISP-DM* (*Cross Industry Standard Process for Data Mining)* corresponde a um modelo processual criado por três organizações experientes (DaimlerChrysler (posteriormente Daimler-Benz), a SPSS (posteriormente ISL) e a NCR, com o objectivo único de uniformizar um conjunto de passos comuns na elaboração e execução de um projecto de MD. Esta metodologia tem como objectivo principal ser completamente transparente e neutra relativamente a quaisquer indústrias, aplicações e ferramentas. Por ser simples de perceber permite que qualquer projecto de MD seja elaborado de forma mais barata e rápida. Para além disso, segundo [www3], esta metodologia é não proprietária e tem sido validada com projectos de grande envergadura.

Para Hornick *et al* (2007) *CRISP-DM* corresponde a um processo bastante intuitivo e bastante reconhecido que começa pela definição do problema e objectivos, identificação dos dados para minerar e o acesso à qualidade dos dados. Segundo o mesmo autor, a disponibilidade de dados para mineração não implica que os dados estejam apropriados para serem minerados. Se os dados estãos “sujos” e contêm erros ou inconsistências então provavelmente terão de ser limpos primeiro.

A viabilidade de qualquer projecto deste tipo está inteiramente relacionado com a qualidade dos dados existentes. Tal como em qualquer outro projecto a fase inicial de análise e desenho é primordial para a obtenção de resultados satisfatórios no final. A seguinte figura representa as várias etapas do modelo CRISP-DM:

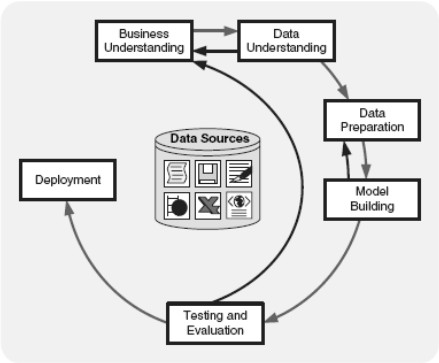


Figura 14 - O modelo CRISP-DM [Olson D. & Delen D. (2008)]

Esta metodologia segue seis passos *standard* que serão brevemente descritos [Chapman *et al* (2000)]:

Fase de conhecimento do negócio: tal como qualquer noutro projecto, a fase inicial é deveras importante. Nesta fase faz-se o levantamento de requisitos juntamente com os elementos do negócio e definem-se todos os objectivos a atingir. Segundo o modelo CRISP-DM existem dois tipos de objectivos: os objectivos definidos na linguagem associada ao próprio negócio e também os objectivos, mais técnicos, associados à MD.

Fase de percepção dos dados: depois de se ter percebido o problema e definidos os resultados pretendidos, é necessário saber que tipo de dados estarão disponiveis para dar seguimento a esta análise, assim como a qualidade dos mesmos.

Fase de preparação dos dados: sabendo da existência de dados brutos, esta fase é importante na medida em que aqui se faz a limpeza e tratamento dos dados inconsistentes, com valores a vazio. Posteriormente faz-se a separação destes dados em pequenos conjuntos, para permitir a análise dos mesmos.

Fase de criação do modelo*:* depois da existência de conjuntos de dados suficientemente aceitáveis prossegue-se com a fase de modelação onde são

escolhidas as técnicas de mineração e algoritmos, são ajustados todo e qualquer tipo de parâmetros para a execução dos próprios algoritmos.

***Fase de avaliação dos resultados**:* depois de criado um modelo temos de testá-lo para provar a sua eficácia relativamente aos objectivos definidos inicialmente.

Fase de Aplicação: após ter sido validada a eficácia do modelo será aplicado sobre dados reais.

Outras metodologias de processos

Para além da metodologia apresentada anteriormente existem outras que permitem elaborar passo a passo uma solução de MD. A metodologia SEMMA (*sample, explore, modify, model, assess*), por exemplo, criada pelo instituto SAS, permite que, seguindo um conjunto de etapas ligeiramente diferentes, se consigam atingir os mesmos objectivos finais propostos pelo modelo CRISP-DM.

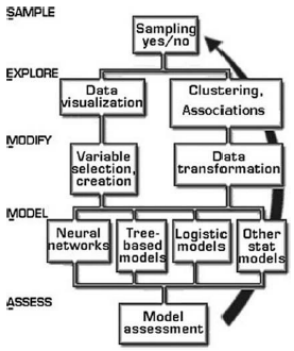


Figura 15 - O modelo SEMMA [Olson D. & Delen D. (2008)]

No entanto, segundo o sítio *KDNuggets*, num questionário efectuado em 2004, CRISP-DM ainda continua a ser a metodologia mais utilizada por parte de todos os investigadores na área (como se pode verificar na seguinte figura).

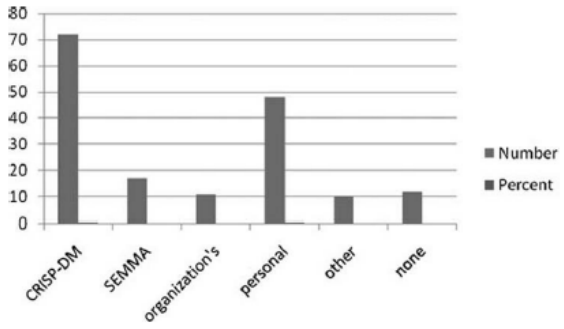


Figura 16 – Metodologias de processos mais utilizadas [KDNuggets, 2004]

PMML

Existe uma forma de representar e guardar modelos de MD sobre a forma de uma linguagem baseada em XML: a *PMML (Predictive Modeling Markup Language)*. A primeira versão desta linguagem apareceu em 1999 pelo *Data Mining Group*[4] e neste momento já existe a versão 3.2. Sabendo cada vez mais da existência de grandes redes e de múltiplos sistemas das organizações a PMML permite, de uma forma transparente, a troca de modelos entre aplicações distintas em ambientes distintos. Desta forma, uma aplicação qualquer depois de criar um modelo e guardá-lo no formato PMML, poderá enviar para outra aplicação remota para o aplicar directamente sobre os seus dados de teste.

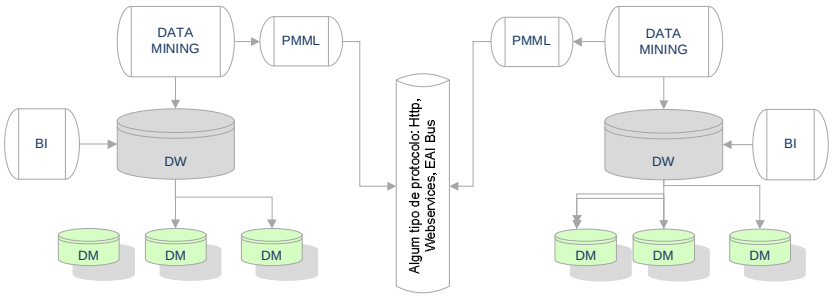


Figura 17 – Partilha de modelos em PMML entre aplicações distintas

[4] O DMG é um grupo independente que desenvolve um conjunto de standards para mineração de dados tal como a PMML.

Para Grossman (2004) um dos objectivos desta linguagem é criar um *interface standard* entre produtores de modelos, tal como sistemas estatísticos ou de MD e consumidores dos modelos, tal como motores de *scoring*, aplicações contendo modelos embebidos e outros sistemas operacionais. Basicamente a última versão da PMML permite gerar modelos dos seguintes tipos:

Tipos	
regressão polinomial	segmentação
regressão	associação
árvores de decisão	redes neuronais

Quadro 3 – Tipos gerados pela PMML

Ferramentas para MD

Existem imensas ferramentas comerciais e académicas que permitem aplicar técnicas da MD. No entanto algumas delas permitem utilizar um conjunto mais alargado de algoritmos do que outras:

Ferramentas	
Intelligent Miner da IBM1	*See5/C5.0*
Clementine da SPSS2	*R6*
DMMiner da DBMiner Technology	*Weka*
Enterprise Miner da SAS3	*Oracle Data Mining (ODM)*
Cubist e Magnum Opus da RuleQuest5	*Oracle Darwin*
S-Plus e Insightful Miner 2 da Insightful Corporation4	*Entre outras*

Quadro 4 – Algumas ferramentas para MD

Para o presente trabalho utilizou-se a ferramenta *ODM* por motivos estritamente profissionais. O ODM é um pedaço de *software* plenamente incorporado na base de dados *Oracle (desde a versão 9i).* Dado que a maioria dos *softwares* para MD necessitam de extrair dados de uma base de dados para ficheiros de texto e importá-los

de seguida, considerou-se interessante utilizar para o caso prático desta dissertação esta ferramenta. Esta solução é comercial e, portanto, qualquer implementação de uma solução de MD baseada no ODM tem este custo acrescido. Os modelos criados no ODM podem ser acedidos através da linguagem *PL/SQL* e também na linguagem *JAVA*, para além do *front end* existente para visualização gráfica de todos os objectos criados. O ODM já permite aplicar bastantes técnicas com os algoritmos apresentados no anexo 1.

If you have knowledge, let others light their candles at it.
Margaret Fuller

Capítulo 5

Técnicas de MD

Sabendo da competitividade extrema de todas as organizações actuais, a análise dos dados históricos permitirá, com certeza, melhorar as suas decisões no futuro. Por este motivo as organizações têm vindo a acumular todo o tipo de dados. Os repositórios de dados têm sofrido um crescimento bastante acelerado nos últimos anos de tal forma que hoje em dia, por existirem volumes tão gigantescos, tornou-se também praticamente impossível de serem manipulados convenientemente por qualquer ser humano sem o auxílio de ferramentas adequadas. Para Fayyad U., Piatetsky-Shapiro G. & Smyth P. (1996) as capacidades actuais para armazenamento de dados ultrapassa de longe a capacidade de analisar, sumarizar e extrair conhecimento desses dados. Seguindo o mesmo ponto de vista Brijs (2002) sublinha ainda que as organizações, por sua vez, estão também cada vez mais interessadas em estudar métodos para a descoberta de novo conhecimento. O crescimento do poder computacional da actualidade leva a que novas técnicas e mecanismos sejam desenvolvidos com o objectivo de analisar em tempo útil esses volumes gigantescos de dados.

Existem actualmente algumas técnicas de MD que permitem analisar estes dados, inferir conhecimento e representá-lo sob a forma de um modelo capaz de representar a realidade desses dados para posterior dedução de novo conhecimento a partir dele.

Figura 18 – Passos para o problema das regras de associação

A MD permite enquadrar as suas técnicas em dois grandes grupos: técnicas *supervisionadas e não supervisionadas.*Relativamente às técnicas *não supervisionadas* são técnicas que não possuem qualquer conhecimente à priori que possibilite a criação de qualquer modelo baseado em dados passados, por exemplo a técnica de geração de *Regras de Associação* ou *Segmentação*. Por outro lado o principal pressuposto por detrás das técnicas *supervisionadas* tem a ver com o facto de existir previamente um conjunto de dados devidamente classificado que possibilita a determinação de casos futuros desconhecidos baseados nestes casos passados: como exemplo a técnica de C*lassificação*.

Segundo Fayyad U., Piatetsky-Shapiro G. & Smyth P. (1996) a*s primeiras são* orientadas às verificações (onde o sistema verifica uma determinada hipótese) e as restantes orientadas para a descoberta (o sistema encontra novas regras e padrões de forma autónoma. Estas noções derivam das áreas de Inteligência Artificial e em particular da área científica de *Machine Learning*.

5.1 Regras de associação

Para tirar partido dos grandes volumes de dados Agrawal R., Imielinski, T., & Swami, A. (1993) apresentam técnicas de MD denominadas por Regras de Associação que permitem extrair conhecimento a partir de um elevado número de transacções sob a forma de regras. Desde então tem sido uma das metodologias mais aplicadas e estudadas na procura de *padrões* escondidos [Hipp *et al* (2000)]. Para entender o conceito de Regras de Associação existem algumas definições importantes a ter em linha de conta. Todas essas definições serão apresentadas de seguida, com alguns exemplos bastante simples e devidamente contextualizados no retalho.

Para Hipp *et al* (2000) uma regra de associação corresponde a uma expressão do tipo $X \Rightarrow Y$, onde X e Y são elementos de um determinado conjunto. Se imaginarmos uma base de dados qualquer D contendo um conjunto de transacções T, onde $T \in D$, então

X⇒Y significa que quando uma transacção contém X então provavelmente também conterá Y.

Uma regra poderá ter a seguinte representação onde o produto computador é o *antecedente* da regra e o disco o *consequente*:

Artigos recebidos | COMPUTADOR ⊳ DISCO
SE recepção = 'COMPUTADOR'
ENTÃO recepção = 'DISCO'

Quadro 5 – Exemplo de regra gerada

Se D corresponder a um conjunto de transacções onde cada transacção T é um conjunto de items tal que T⊆D. Como exemplo imagine-se um *dataset* relativo ao histórico de encomendas e/ou recepções contendo artigos encomendados a um determinado fornecedor.

Ordem Encomenda	**Artigos recebidos**
OE1	{computador, processador, memória}
OE2	{computador, memória}
OE3	{disco, memória}
OE4	{disco, monitor}
OEn	{m produtos}

Quadro 6 – Exemplo de artigos encomendados em conjunto

Um *itemset* X corresponde a um ou mais *items*. Diz-se que uma transacção contém o *itemset* X se X ⊆ T. Considere-se o *itemset* da OE1={computador, processador, memória}. O elemento "computador ⊆ {computador, processador, memória}".

5.1.1 Desafios associados à Mineração de Regras de Associação

No capítulo 2 deste trabalho apresentou-se um pequeno enquadramento baseado no livro de Anderson (2006) onde se salienta o papel das tecnologias de informação e a

facilidade de apresentar produtos aos consumidores através da *Internet*. O autor refere que a disponibilização ilimitada de produtos nas lojas virtuais permite não só vender mais como também vender produtos nicho inexistentes nas superfícies físicas (limitadas pela sua capacidade física).

O exemplo clássico que permite descobrir afinidades entre produtos corresponde à análise dos cestos de compras dos consumidores (tarefa denominada na terminologia inglesa por *Market Basket Analysis* - *MBA*). Sabendo que um determinado consumidor poderá comprar diversos produtos em conjunto, sobre superfícies físicas e/ou virtuais, o decisor poderá assim antecipar a procura abastecendo as suas lojas estrategicamente, promovendo a compra de um produto baseado nas compras do outro, ou até mesmo re-desenhando os seus próprios *layouts* físicos e virtuais baseado nas afinidades encontradas [Brijs *et al* (1999)].

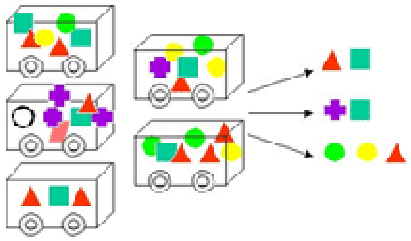


Figura 19 – Carrinhos de compras e a descoberta de afinidades

O enquadramento inicial apresentado no capítulo 2 não acontece por acaso. Sabendo da elevada quantidade de produtos disponibilizados nas estruturas das organizações actuais, existe uma probabilidade muito acentuada de existirem imensas transacções compostas apenas por alguns produtos muito específicos – dando origem a transacções muito esparsas. Lidar com tamanha esparsidade de dados é um dos desafios associados com a descoberta de regras de associações e à criação de novos algoritmos capazes de melhorar o desempenho do próprio processo. Para além da esparsidade das transacções, existem outros desafios, por exemplo, os resultados obtidos após a aplicação de algoritmos de pesquisa de regras poderão ser, tal como o próprio volume de dados inicial, elevados e impossíveis de serem manipulados facilmente por qualquer analista.

Para Zheng Z., Kohavi R. & Mason L. (2001) a geração de regras de associação pode gerar centenas de milhares de regras de associação. Em segundo lugar o próprio desempenho dos algoritmos associados a esta técnica torna-se pior quando aplicadas em repositórios mais volumosos. Por esse motivo, Agrawal, R., & Srikant, R. (1994) afirmam que os algoritmos terão de ser cada vez mais rápidos.

Relativamente ao primeiro caso, depois de aplicados os algoritmos para a descoberta de regras, podem-se obter volumes elevados de regras, mas muitos desses resultados não são interessantes. Por esse motivo existem medidas que permitem restringir o volume de regras geradas que permitem descobrir apenas todas as regras de associação que tenham *suporte* e *confiança* superiores a um suporte mínimo pré-definido (*minsup*) e confiança mínima (*minconf*) respectivamente [Agrawal, R., & Srikant, R. (1994)].

- A regra X⇒Y tem *suporte s* se em D s% das transacções de D contêm X U Y . O suporte representa a fracção de transacções em D que contêm ambos os items no antecedente e consequente da regra.
- A regra X⇒Y tem uma confiança de c% indica que num determinado conjunto de transacções D se X aparece então Y também aparece com confiança c%.

Designação	Fórmula
Suporte	$s = \frac{\bullet(\text{computador}, \text{processador}, \text{memória})}{\lvert D \rvert} = \frac{1}{5} = 0.2$
Confiança	$c(\text{computador} \rightarrow \text{processador}) = \frac{\bullet(\text{computador} \cup \text{processador})}{\bullet(\text{computador})} = \frac{X}{X} = X$

Quadro 7 – Suporte e confiança de uma regra

Os factores de suporte e confiança possibilitam "filtrar" o volume de regras geradas. Dessa forma o volume final dos resultados poderá ser cuidadosamente reduzido e, da mesma forma, melhorado o seu desempenho. Diz-se cuidadosamente na medida em que se se restringir o intervalo de regras a obter, usando as medidas apresentadas

anteriormente, então pode-se estar também a perder um conjunto de regras com elevado valor para o negócio.

5.1.2 Outras medidas de interesse para avaliação das regras geradas

Para além destas medidas, existem outras medidas propostas na literatura, mas não exaustivamente analisadas neste trabalho visto não ser esse o principal objectivo. Salientam-se, por exemplo: *Lift* (denominado também por *Interest*), *Conviction, Leverage, Coverage, Correlation.* O conceito de *Lift* foi introduzido por Brin S. (1997) e mede quantas vezes mais do que o esperado X e Y ocorrem em conjunto se forem estatisticamente independentes. De uma forma mais simplificada o *Lift* de uma regra de associação $X \Rightarrow Y$ indica o quanto mais frequente é *Y* quando X ocorre. O conceito de *Conviction* for introduzido por Brin S. (1997) compara a probabilidade de X ocorrer sem Y se ambos forem dependentes da actual frequência da ocorrência de X sem o Y. O conceito de *Leverage* foi introduzido por Piatetsky-Shapiro G. (1991) e mede a diferença entre X e Y aparecendo em conjunto e o que poderia ser esperado se X e Y fossem estatisticamente dependentes. O conceito de *Coverage* mostra que parte dos *itemsets* do consequente são cobertos por uma regra. Os seus valores variam entre [0,1]. O conceito de *Correlation* é um conceito estatístico que permite mostrar quando e como pares fortes de variáveis/*itemsets* estão relacionados [Sheikh L., Tanveer B., Hamdani S. (S/D)].

Para além destas medidas existem outras relevantes: *Laplace, Teste do* χ^2*, Jaccard, Cosine, Mutual Info, entre outras.* Para Hamdani *et al* (S/D) a combinação de algumas ou todas estas medidas possibilitariam realmente encontrar regras interessantes.

Designação	Fórmula
Lift	$\text{Lift}(X \Rightarrow Y) = \frac{P(X \wedge Y)}{P(X)P(Y)}$

Conviction	$conv(X \Rightarrow Y) = \frac{1 - s(Y)}{1 - conf(X \Rightarrow Y)}$
Leverage	$leverage\ (X \Rightarrow Y) = s(X \cup Y) - s(X) * s(Y)$
Coverage	$cov\,erage(X \Rightarrow Y) = sup\,orte(X)$

Quadro 8 – Outras medidas de interesse para avaliação de uma regra

5.1.3 Algoritmos para gerar regras de associação

Existe uma quantidade elevada de novos algoritmos ou variantes de alguns algoritmos conhecidos. Para Fayyad *et al* (1996) a escolha de um determinado algoritmo é uma arte e uma tarefa que compete a um analista fazer.

Segundo Savasere *et al* (1995) uma das particularidades da maioria dos algoritmos reside no facto de todos eles terem um fraco desempenho porque necessitam de percorrer várias vezes a base de dados. Todas estas passagens implicam que haja um grande esforço de *Input/Ouput*, especialmente sobre grandes repositórios. Os passos associados com a geração de regras de associação podem ser traduzidos na seguinte figura:

Figura 20 – Passos para o problema das regras de associação

a) Em primeiro lugar será feita a geração de todas as combinações de items;

b) Seguidamente a descoberta de conjuntos de items;

c) Finalmente a geração das regras de associação.

Na literatura existem bastantes estudos neste âmbito, sendo esta uma área bastante apreciada por toda a comunidade científica. Existem imensos algoritmos e algumas

variantes dos mesmos, mas por não ser o foco essencial deste trabalho colocam-se apenas alguns no quadro seguinte.

Algoritmo	Autores/Período	Algoritmo	Algoritmo
KID3	Piatestky-Shaphiro (1991)	*Partition*	Savasere A. (1995)
AIS	Agrawal *(1993)*	*Eclat*	Zaki (1997)
SETM	Houstman (1995)	*DIC*	Brin S. (1997)

Quadro 9 – Algoritmos para geração de regras de associação

5.1.4 Estudo relativo ao desempenho de alguns algoritmos

Segundo Houtsma (1995) a problemática do desempenho é um problema associado a muitos algoritmos cuja aplicação implique a mineração em grandes bases de dados. Neste momento existem bastantes algoritmos apresentados por diversos autores, capazes de gerar estas regras de associação. Para Savasere A. *et al* (1995) estes algoritmos variam essencialmente na forma como os *itemsets* candidatos são gerados e como os suportes para os *itemsets* candidatos são contados.

Este capítulo surge apenas para analisar os resultados de uma investigação efectuada por Zheng Z., Kohavi R. & Mason L. (2001), com dados reais e artificiais, fazendo a variação nas medidas de interesse estudadas no sub capítulo anterior. Existem bastantes estudos estritamente relacionados com o desempenho de alguns algoritmos para geração de regras. Geralmente para que estes estudos estejam coerentes com todos os que pretendam comparar com outros estudos, alguns *datasets standard* e outros artificiais são geralmente utilizados para o efeito. Por exemplo os *datasets IBM-Artificial, BMS-POS, BMS-WebView1 e BMS-WebView2* são de livre acesso e permitiram fazer a avaliação do desempenho. O seguinte quadro traduz os volumes testados.

	Transac-tions	Distinct Items	Maximum Transaction Size	Average Transaction Size
IBM-Artificial	100,000	870	29	10.1
BMS-POS	515,597	1,657	164	6.5
BMS-WebView-1	59,602	497	267	2.5
BMS-WebView-2	77,512	3,340	161	5.0

Quadro 10 – *Data Sets* utilizados por Zheng Z., Kohavi R. & Mason L. (2001)

Neste estudo compararam cinco algoritmos utilizando os *datasets* anteriores: *Apriori, FP-growth, Closet, Charm e Magnum-Opus*. Segundo este grupo de investigadores os resultados resultantes da aplicação dos mesmos ao *dataset* Artificial são bastante diferentes dos resultados quando aplicados aos outros três *datasets*. Este facto acontece porque o primeiro *dataset* apresenta muitas características distintas e não reais. Eles também afirmam que a escolha do algoritmo apenas interessa quando os níveis de suporte aplicados impliquem que a geração de regras ultrapasse as que realmente poderiam ser úteis. Neste estudo quando o suporte permitiu gerar 1 000 000 de regras então o *Apriori* conseguiu terminar em menos de 10 minutos. Quando a variaçao do suporte sofria ligeirissimas variações de 0.02% então o número de regras subiu de menos de 1 milhão para valores acima de um bilião, fazendo com que realmente para variações de suporte para além de um determinado intervalo a escolha do algoritmo seja irrelevante. De salientar que o estudo foi aplicado variando apenas o suporte e mantendo sempre a zero o valor da confiança.

5.2 Classificação

A Classificação é vista como sendo uma das tarefas mais benéficas e lucrativas da MD. Para Chen *et al* (1996) a *Classificação* é o processo que permite descobrir propriedades comuns entre um conjunto de objectos de um conjunto de dados e classificá-los em classes diferentes de acordo com um modelo de conhecimento.

Tal como se pode visualizar na figura seguinte, segundo Tan P., Steinbach M. & Kumar V. (2005), em qualquer problema deste tipo espera-se a entrada de um conjunto de dados. Cada linha desse conjunto de dados, também denominada por instância, é caracterizada por um tuplo (x,y) onde **x** corresponde a um conjunto de atributos e o **y** corresponde a uma classe ou atributo destino. Pretende-se com esta técnica criar modelos descritivos e/ou predictivos a partir de um conjunto de dados de entrada.

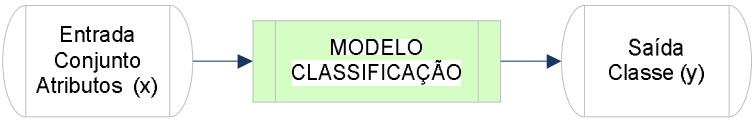


Figura 21 – Modelo de Classificação [Tan P., Steinbach M. & Kumar V. (2005)]

Segundo Han *et al* (2001) a Classificação de dados é um processo efectuado em dois passos:

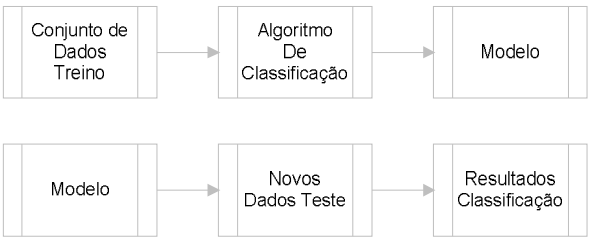


Figura 22 – Os dois passos para a Classificação

A construção de um modelo (também denominado por *classificador*) a partir de um conjunto de dados de treino cuja classe é previamente conhecida. Dado que esta classe para estes casos é já conhecida este tipo de técnica é *supervisionada* na medida em que para cada caso de treino fornecido é já conhecida a sua classe. O modelo construído é utilizado para fazer novas classificações[5]. Em primeiro lugar será analisada a eficiência do modelo, por exemplo aplicando um conjunto de casos de teste. Neste caso a eficiência do modelo corresponde à percentagem de casos de teste correctamente

[5] O processo de aplicar um determinado modelo a um conjunto de dados de teste é denominado por *scoring*.

classificados pelo modelo. Se este modelo for considerado aceitável então poderá ser utilizado noutros casos de testes cuja classe não seja conhecida. O primeiro passo que corresponde à construção de um modelo de Classificação é denominada por treino ou aprendizagem. Neste caso dados denominados por dados de treino ou de aprendizagem são utilizados. Posteriormente avalia-se este classificador através de um conjunto de teste.

Existem imensas formas para induzir um modelo de Classificação: através de árvores de decisão, classificadores baseados em regras, redes neuronais, SVM (*Support Vector Machines*), classificadores *Naíve Bayes*, entre outros. O presente estudo focará apenas a indução de árvores de decisão.

5.2.1 Classificação por indução de árvores de decisão

As árvores de decisão são classificadores bastantes simples de perceber e, talvez por esse motivo e pela sua simplicidade de visualização dos resultados, seja uma das metodologias mais populares e aplicadas na tarefa de Classificação. Este facto é reforçado em todas as leituras efectuadas.

Segundo Han *et al* (2001) uma árvore de decisão corresponde a uma estrutura em árvore onde cada nodo representa um teste num determinado atributo, cada braço representa um resultado do teste, cada nodo folha representa classes ou distribuições de classes. Uma árvore poderá ser binária, ou não, caso o número de classes tenha dois ou mais elementos.

Na seguinte figura apresenta-se uma árvore binária onde V1 corresponde ao nodo raiz, sendo este o primeiro nodo a ser considerado, e o nodo V2 que corresponde a um nodo não terminal. Este exemplo representa um problema de classificação binário onde as classes definidas para este caso são X e Y.

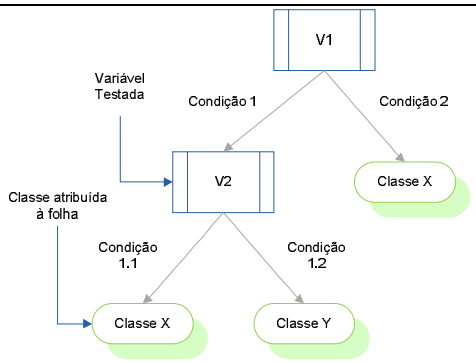


Figura 23 – Representação de uma árvore de decisão bidimensional

Quando o problema implicar um número maior de classes, *n-dimensional*, então a árvore poderá ter uma representação mais larga sugerida pela seguinte figura:

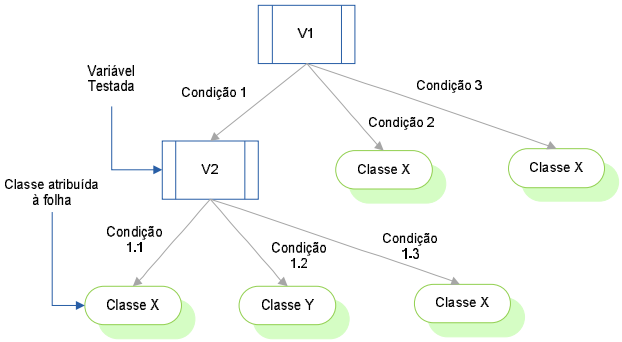


Figura 24 – Representação de uma árvore de decisão n-dimensional

Em ambas as árvores apresentadas a classificação para um novo caso é simples. Para tal bastará percorrer o caminho desde o nó raiz até à classe final. Durante este percurso várias regras de decisão vão sendo adquiridas.

5.2.2 Passos essenciais para a construção de uma árvore de decisão

Para Olson D. & Delen D. (2008) existem exponencialmente imensas árvores de decisão que poderão ser geradas a partir de um conjunto de atributos. Encontrar a melhor representação em árvore é praticamente impraticável devido ao tamanho exponencial do espaço de procura. No entanto, tal como será apresentado

posteriormente, existem bastantes algoritmos capazes de gerar árvores com bastante eficácia e em tempo útil.

O processo de aprendizagem de uma árvore de decisão é conhecido como *indução de árvores*. Existem imensos algoritmos que permitem construir árvores de decisão que se baseiam na estratégia de dividir para conquistar (da terminologia inglesa *divide and conquer*), ou seja, a cada passo que o algoritmo efectue, divide o problema em sub problemas de menor dimensão até que se encontre uma solução para cada um desses sub problemas. Os classificadores baseados em árvores de decisão procuram as melhores formas para efectuar essas divisões sucessivas.

Apesar de praticamente todos os classificadores baseados em árvores de decisão serem idênticos existem particularidades que os diferem. De entre essas particularidades destacam-se a escolha das características a utilizar em cada nó, a escolha, cálculo da partição do conjunto de treino, decidir quando o nó é folha, critério de selecção da classe. Para efectuar os particionamentos sucessivos da árvore e seleccionar os atributos nó existem apenas duas formas: ou através de heurísticas ou através de medidas estatísticas (por ex: o *ganho de informação*).

O *ganho de informação* (GI) de cada atributo permite avaliar a utilidade do mesmo para a Classificação. Segundo Han *et al* (2001) a medida GI é utilizada para seleccionar o atributo pela qual a árvore deverá ser partida. Desta forma, o atributo que possuir maior GI será o escolhido para o próximo passo. Da bibliografia consultada a *Entropia* e o *índice GINI* são os critérios mais mencionados e utilizados para efectuar estes particionamentos. Enquanto que a *Entropia* caracteriza a pureza dos dados, o índice Gini permite medir o grau de heterogeneidade dos mesmos. Para além destas medidas existem outras também relevantes, que não serão aplicadas na componente prática: por exemplo o *RELIEFF* proposto por Kononenko *et al* (1995) e (1997).

5.2.3 Algoritmos para geração de árvores de decisão

Existem imensos algoritmos que permitem gerar árvores de decisão. No decorrer deste sub-capítulo apresentar-se-á uma breve revisão bibliográfica sobre alguns deles assim como um enquadramento histórico. A seguinte tabela introduz alguns dos algoritmos mais referidos:

Algoritmo	Descrição	Autor / Período
CLS	*Concept Learning System*	[Hunt (1966)]
CART	*Classification And Regression Trees*	[Breiman (1984)]
ASSISTANT	*ASSISTANT*	[Kononenko *et al* (1984)]
ID3	*Iterative Dichotomiser 3*	[Quinlan (1986)]
GID3	*Generalized Iterative Dichotomiser 3*	[Cheng *et al* (1988)]
C4.5	*C4.5*	[Quinlan (1993)]
DBLearn	*DB Learn*	[Koonce (1997)].
SLIQ,	*SLIQ,*	[Mehta (1996);
RAINFOREST e	*RAINFOREST*	Ganti (1999);
SPRINT	*SPRINT*	Shafer (1996)]
1R	*One Rule*	[Holt (1993)]
CHAID	*Chi-Squared Automatic Interaction Detector*	[Kass (1980)]
QUEST	*Quick, Unbiased, Efficient Statistical Trees*	[Shih *et al* (1997)]

Quadro 11 – Alguns algoritmos para a geração de árvores de decisão

Os classificadores baseados nas árvores de decisão têm origem da década de cinquenta. Da bibliografia encontrada salientam-se o trabalho de Hunt *et al* (1966) com a apresentação do algoritmo CLS. Posteriormente o trabalho de Quinlan (1979) com o algoritmo ID3 que terá a sua última versão (ID3-IV) em Quinlan (1986). Posteriormente encontram-se diversas optimizações do ID3: o algoritmo ASSISTANT de Kononenko *et al* (1984), o GID3 de Cheng *et al* (1988), entre outros. Posterior ao ID3, relevante neste estudo, encontra-se o estudo de Breiman (1984) com a sugestão do algoritmo CART. Quinlan (1993) sugere também o algoritmo C4.5 que, neste momento, existe já na versão 5.

5.2.4 Desempenho dos modelos gerados

Depois de ser criado um modelo baseado em árvores de decisão, utilizando um algoritmo seleccionado, deve-se tentar perceber o quão eficiente é o classificador. Para este efeito deverá ser feita (a) uma avaliação do modelo sobre sub conjuntos de dados cuja classe é conhecida mas escondida para avaliação ou (b) sobre um sub conjunto de dados cuja classe não se conheça ainda. Segundo Rountree (1999) através da avaliação do modelo com casos cuja classe é desconhecida pode-se analisar a sua capacidade de *generalização*.

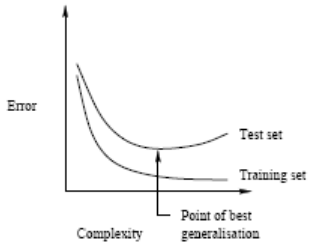


Figura 25 – Conceito de *Overfitting* [Rountree (1999)]

A figura anterior ilustra a capacidade de generalização do nosso classificador. Inicialmente, depois de induzida a árvore de decisão, dado que ainda não foi "treinada" para os casos de teste, a curva evidencia uma taxa de erro superior quando comparada com os valores do conjunto de treino – está *overfitted*. A solução para *overfitting* de uma árvore de decisão é fazer crescer a árvore ao máximo e depois fazer a poda sucessivamente de acordo com algum critério até o erro baixar. Ainda de salientar, segundo Breiman (1984), que a complexidade de uma árvore tem um efeito crucial na sua eficácia. Na maioria das vezes a complexidade de uma árvore é medida por uma das seguintes métricas: total de nodos, total de folhas, profundidade da árvore e o número de atributos utilizados. A complexidade é explicitamente controlada pelo critério de paragem e pelo método de poda empregues.

Existem várias abordagens que permitem avaliar e seleccionar classificadores: através da análise da *Taxa de Erro* e o *Ganho*, a matriz de confusão e a análise da curva ROC

(*Receiver Operating Characteristic*), curva *Lift* e UAC (*Under Area Curves*), para além de outros mecanismos. Existem diversos estudos que focam estas metodologias de forma mais detalhada. Da imensa bibliografia salientam-se os estudos de Fawcett (2003) e Provost (2001) relativos às curvas ROC e UAC. Quanto à metodologia *Lift* salienta-se o livro de Berry e Linoff (1999).

Metodologias para estimação de taxa de erro

A avaliação do desempenho de um classificador é uma tarefa de extrema importância. Um dos desafios associados a esta tarefa corresponde em estimar a taxa de erro de generalização que o mesmo pode atingir para novos casos. Da literatura consultada existem diversas possibilidades para avaliar estes erros: através de *Cross Validation* (*aka K-fold*), *Holdout test*, *BootStrapping*, entre outros. Blum (S/D) e Kohavi R. (1995) apresentam estudos acerca destas metodologias. Infelizmente, nem todas as ferramentas disponíveis no mercado e aqui referidas no sub capítulo 4.6 possibilitam aplicar estas técnicas.

Validação Cruzada *(da terminologia inglesa Cross Validation – aka K-Fold)*: A metodologia *K-Fold*, também designada por estimativa de rotação [Kohavi R., (1995)], é um método que consiste em dividir o conjunto de dados em K partições iguais. Apenas uma dessas partições corresponde inicialmente ao conjunto de teste e as restantes K-1 aos conjuntos de treino. Cada iteração utilizará um novo conjunto de testes e uma taxa de erro será calculada de etapa em etapa. No final de todos os passos a taxa de erro corresponderá à média das taxas de erro obtidas nas diferentes partições.

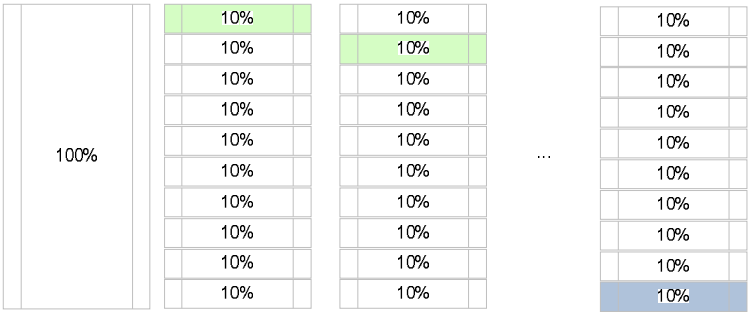


Figura 26 – Metodologia *Cross Validation*

***Holdout test**:* Este método, também denominado por estimativa por amostragem [Kohavi R. (1995)], permite partir o conjunto de dados em dois grupos t1 e t2, sendo um para conjunto de treino e outro para conjunto de teste (também denominado por *Holdout set*). Geralmente aplica-se a regra 2/3 para o primeiro caso e 1/3 para o segundo, respectivamente.

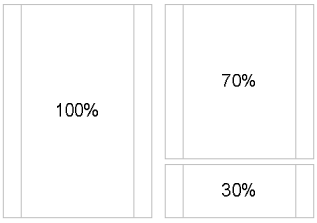


Figura 27 – Metodologia *Holdout test*

Matriz de confusão

Segundo Kohavi e Provost (1998) para se saber quão viável é um modelo criado procede-se a uma contagem dos casos de testes aplicados e cuja classificação foi efectuada correctamente. Estes resultados são então colocados num tabela denominada por matriz de confusão, sendo esta uma das técnicas mais utilizadas para os problemas de Classificação. A matriz de confusão permite avaliar os resultados de uma determinada Classificação fazendo o mapeamento entre os valores previstos por um modelo com os valores desejados.

		CLASSE VERDADEIRA P	CLASSE VERDADEIRA N
CLASSE HIPOTÉTICA	P	True Positive	False Positive
	N	False Negative	True Negative
TOTAIS			

Figura 28 - Representação de uma matriz de confusão bidimensional

Quando se lida com problemas de Classificação de problemas binários, onde a classe final poderá ter dois valores, algumas medidas terão de ser calculadas para permitir entender a matriz de confusão: número de verdadeiros positivos (TP), número de falsos positivos (FP), número de falsos negativos (FN) e número de verdadeiros negativos (TN). Como exemplo, considere-se o seguinte quadro para um problema idêntico:

Classe correcta	**Classe prevista**	**TP**	**FN**	**FP**	**TN**
-	+	0	0	1	0
-	+	0	0	1	0
-	-	0	0	0	1
-	-	0	0	0	1
-	-	0	0	0	1
-	-	0	0	0	1
-	+	0	0	1	0
-	-	0	0	0	1
+	+	1	0	0	0
-	+	0	0	1	0
+	-	0	1	0	0
+	-	0	1	0	0
	Totais.	1	2	4	5
		TP + FN = +		FP + TN = -	
		+ = 3		- = 9	

Quadro 12 – Cálculo das medidas necessárias para a matriz de confusão

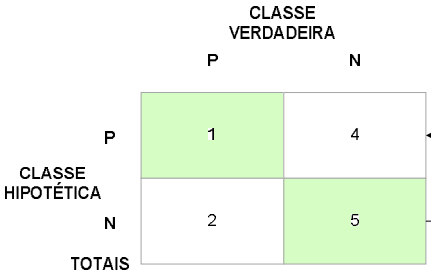


Figura 29 – Cálculo das medidas necessárias para a matriz de confusão

Analisando o quadro anterior, apesar de ser um exemplo extremamente simplificado, é facilmente perceptível que quando uma classe é negativa e foi classificada como positiva então corresponde a um falso positivo. Por outro lado quando uma classe é positiva mas classificada como negativa então corresponde a um falso negativo. Como consequência, analisando o quadro 9, constata-se que a primeira célula que contém o valor 1 indica o número de *verdadeiros positivos* para obter a classe P. Esta estatística indica que num caso o modelo previu correctamente a classe. Para a segunda célula o modelo previu 2 vezes que seria a classe N, mas a verdadeira classe seria P. A próxima célula indica que modelo previu 4 vezes erradamente a classe como sendo P. Por fim o último valor indica que o modelo conseguiu prever correctamente 5 casos como *verdadeiros negativos*. O que interessa saber realmente são os valores totais de verdadeiros positivos e verdadeiros negativos, que neste caso contabilizam o valor 6.

Uma matriz de confusão poderá também ter uma representação mais alargada caso o número de classes seja maior que dois, aumentando a complexidade da própria grelha. No entanto a lógica existente para a sua construção mantém-se. A seguinte figura apresenta uma matriz de confusão com 3 classes.

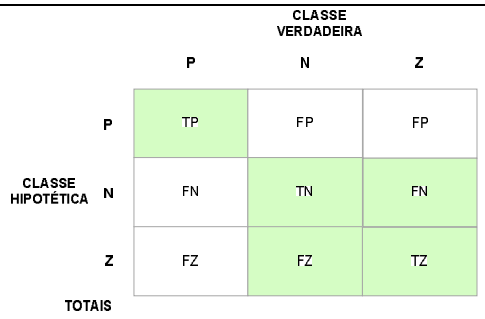


Figura 32 – Matriz de confusão com três classes, 3-dimensional

Segundo Fawcett T. (2003) existem métricas a ter em consideração na análise da matriz de confusão, cujas formulação se apresenta no quadro seguinte: *TP Rate*, *FP Rate (Recall)*, *Precisão e Eficácia*. Para Vuk M. e Curk T. (2006) estas medidas de precisão e eficácia servem geralmente para medir a qualidade de classificadores binários. Partindo dos valores cálculados anteriormente calculam-se as seguintes métricas:

Designação	Fórmula	Resultados
FP Rate	$\frac{FP}{N}$	$\frac{2}{5}$
TP Rate = Recall	$\frac{TP}{N}$	$\frac{1}{5}$
Precisão	$\frac{TP}{TP+FP}$	$\frac{1}{3}$
Eficácia	$\frac{TP+TN}{TP+FP}$	$\frac{2}{3}$
F-score	Precisão x *Recall*	$\frac{1}{3} \text{ x } \frac{2}{3}$

Quadro 13 – Métricas relacionadas com a matriz de confusão e resultados obtidos

ROC – *Receiver Operating Characteristics*

Segundo Egan (1975) as curvas ROC são medidas que servem para comparar classificadores, sendo estas muitas vezes utilizadas por sistemas de diagnóstico médico e de apoio à decisão (de acordo com Hanley e McNeil (1982) citado por Vuk M. e Curk T. (2006)).

Para Fawcett T. (2005) a análise ROC é uma das técnicas que permite organizar, visualizar e seleccionar classificadores baseada nos desempenhos dos mesmos. Esta análise permite analisar o limiar entre as taxas de acertos ou verdadeiros positivos (*TP Rate*) e taxas de erro ou falsos positivos (*FP Rate*). O mecanismo de construção do gráfico é relativamente simples. Em primeiro lugar coloca-se o valor de *FP Rate* no eixo X e o valor de *TP Rate* no eixo Y. O ponto (0,0) representa uma estratégia em que nenhuma classificação positiva é gerada ao contrário do ponto (1,1) que representa uma estratégia para gerar positivos. O ponto (0,1) representa uma classificação perfeita, ou seja todas as instâncias são bem classificadas.

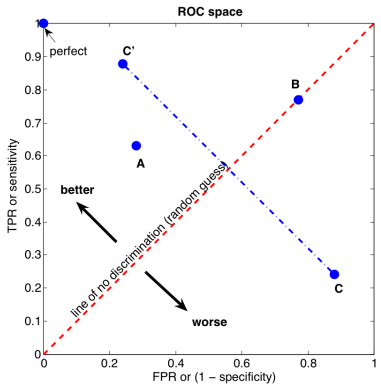


Figura 29 – Representação dos pontos no espaço ROC

Um ponto é considerado melhor do que outro ponto se este se encontra mais à direita e/ou mais à esquerda do que o outro ponto. Ou seja, visualizando a figura anterior é facilmente perceptível que o ponto C é melhor que os pontos A, B e C.

AUC – *Area Under Curve*

Para Fawcett T. (2005) para ser possível comparar classificadores é necessário reduzir a curva ROC a um valor escalar. Uma das formas de o fazer é calcular a área abaixo da curva ROC que geralmente é calculada unindo o respectivo ponto no espaço e os pontos (0,0) e (1,1).

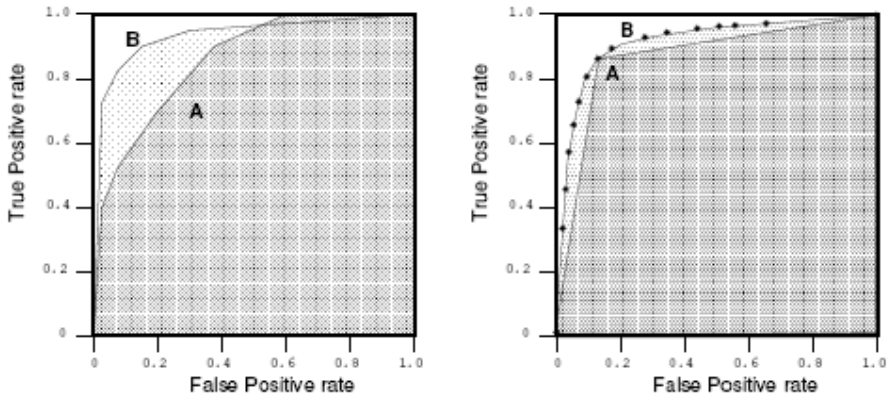


Figura 30 – Representação da UAC [Fawcett (2005)]

Relativamente ao primeiro caso apresentado na figura anterior a curva de A intersecta a curva de B. Apenas nesta altura existe uma ligeira vantagem de A sobre B. Quanto ao segundo caso é visível que o *classificador B* possui uma área maior, logo pode inferir-se também que tem um desempenho médio melhor que o *classificador A*. Sendo assim o classificador B seria o escolhido.

The key to success in business is to know something that nobody knows.
Aristotle Onassis

Capítulo 6

Aplicação Prática

Um dos grandes desafios de qualquer vendedor é entender o comportamento dos seus consumidores com o objectivo de: (a) antecipar a procura; (b) garantir a existência de novos produtos; (c) garantir *stock* suficiente para satisfazer toda a procura; (d) entregar atempadamente as encomendas efectuadas. Para Angerer (2005) todos os esforços realizados na tentativa de melhorar a cadeia de abastecimento são fúteis se, no final, o consumidor não conseguir comprar um produto por não se encontrar na prateleira de uma loja física. Com esta ideia em mente e o objectivo de aplicar a teoria apresentada nos capítulos anteriores, sugerem-se dois casos práticos utilizando dados de uma organização retalhista portuguesa, que visam optimizar a sua cadeia de abastecimento. Por motivos de confidencialidade não serão apresentados, no decorrer desta aplicação prática, quaisquer descrições relativas à organização.

Existem diversos estudos que visam optimizar a cadeia de abastecimento. O presente estudo tem como foco principal a aplicação de duas técnicas específicas da MD sobre dados reais: *Regras de Associação e Classificação*. Em primeiro lugar a descoberta de regras de associação incidirá sobre dados de ordens de encomenda de compradores intermédios da organização. Muito embora as técnicas de indução de regras de associação sejam tradicionalmente anexadas às compras de consumidores nas lojas, com as análises dos seus carrinhos de compra, neste caso específico analisam-se as recepções efectuadas nas lojas e/ou armazéns ou os envios realizados pelos fornecedores. Na

segunda componente, serão analisadas algumas variáveis que vão permitir fazer a classificação de uma encomenda, com determinadas características, que possibilitará prever possíveis falhas nos prazos de entregas e nas quantidades encomendadas. Uma das variáveis em análise será o próprio fornecedor, que perante determinadas quantidades de compra e prazos, não conseguirá responder adequadamente. Por este motivo, este trabalho também permitirá fazer a avaliação e a selecção de um fornecedor para uma nova encomenda, sabendo que a selecção do fornecedor mais adequado é a chave para a obtenção de níveis desejados de qualidade, níveis de suporte técnico, um nível de serviço atempado e a um preço certo [Darade *et al* (S/D)].

Para este caso prático, serão seguidos todos os passos de CRISP-DM, *standard* devidamente apresentado no sub capítulo 4.1.2.

6.1 Fase de conhecimento do negócio

A organização em causa dedica-se à comercialização de bens para consumo e permite a realização de compras sobre estruturas físicas e virtuais. As estruturas físicas desta organização denominam-se por unidades funcionais. Actualmente é constituída por, aproximadamente, 45 entrepostos e 1217 lojas instaladas sobre a Península Ibérica. No presente ano a organização aumentou substancialmente o seu número de estruturas, com cerca de 637 novas lojas abertas, sendo 24 virtuais. O total de fornecedores activos aproxima-se, neste momento, dos 20500, e o total de clientes aproxima-se já dos 3 milhões.

Os seguintes gráficos representam o número de novas estruturas físicas e virtuais abertas nos últimos anos.

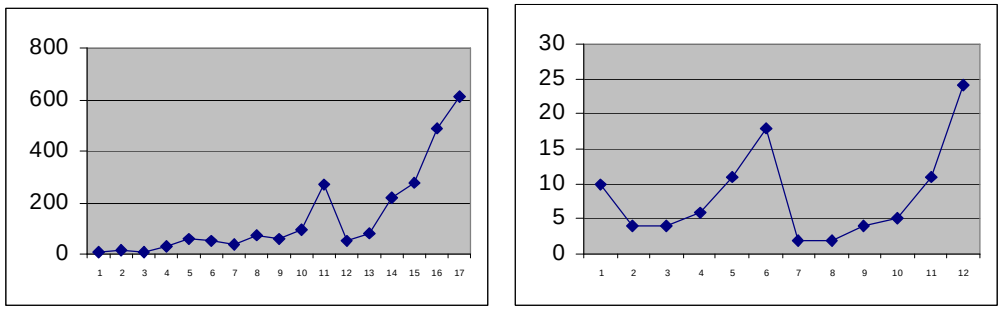


Figura 31 – Crescimento de lojas virtuais e físicas na organização

A organização efectua as suas encomendas a fornecedores *externos* e *internos*, sendo considerados externos aqueles fornecedores que permitem abastecer os entrepostos ou as lojas directamente através de ordens de encomenda especialmente feitas por compradores dos respectivos departamentos. Quanto às requisições internas permitem movimentar a mercadoria entre entidades da mesma organização.

Desta forma há três possibilidades para efectuar encomendas e respectivas recepções: (a) através de um mecanismo denominado por cross docking onde a mercadoria é encomendada a fornecedores externos para abastecimento de lojas, mas durante o seu percurso efectua uma passagem no entreposto, sem repor stock, até à loja final; (b) através de transferências directas entre armazéns e lojas; (c) finalmente através das encomendas directas a fornecedores para abastecimento de armazéns.

A seguinte figura exemplifica este processo apresentando os três mecanismos de *sourcing* [6]dos produtos da organização.

[6] O conceito de *sourcing* trata da identificação, avaliação, negociação e configuração de novos produtos e/ou fornecedores.

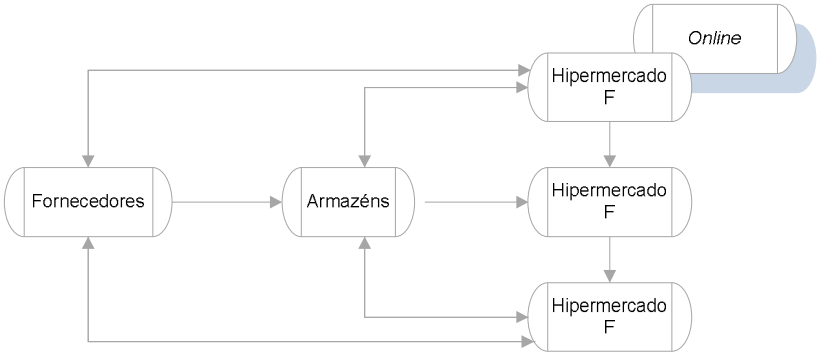


Figura 32 – Ciclo de compra e reaprovisionamento

6.2 Fase de percepção dos dados

A instituição em causa possui um DW, relativamente estável, mas com algumas falhas na qualidade dos seus dados. Este DW serve de repositório central e de suporte às equipas de negócio para as suas tomadas de decisão operacionais e estratégicas. Para se perceberem os dados armazenados iniciou-se este trabalho com uma análise de alguns DM, das suas dimensões e hierarquias, com particular interesse nos factos relativos às ordens de compra e respectivas recepções. Pretendeu-se assim descobrir quais as principais variáveis relevantes para o estudo em causa.

Para o primeiro caso, relativo à descoberta de afinidades entre produtos, havia duas aproximações válidas e perfeitamente lógicas: (a) por um lado, através da análise das vendas a consumidores. Com esta análise consegue-se descobrir afinidades entre os produtos comprados em conjunto possibilitando, posteriormente, optimizar a encomenda dos mesmos, reduzindo o número de fornecedores e, consequentemente, custos associados à distribuição. (b) por outro lado poder-se-ia tentar analisar as recepções nos armazéns e os produtos recepcionados em conjunto, para posteriormente proceder à encomenda dos mesmos, mas a menos fornecedores. Esta abordagem seria válida, por exemplo, quando o fornecedor entregasse produtos substitutos para colmatar um produto em falta. Optou-se inicialmente por analisar o DM de encomendas,

extraindo 6 meses de histórico (*64118 registos*) da sua tabela principal de ordens de compra para apenas uma categoria de produtos: *Informática.*

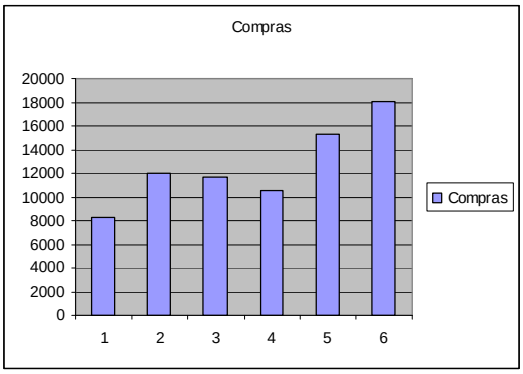


Figura 33 – Distribuição do volume de ordens de compra mensais

Relativamente à segunda parte do trabalho foram analisados os mesmos seis meses para a mesma categoria.

6.3 Fase de preparação dos dados

Na ferramenta utilizada, para se trabalharem os dados, é necessário fazer o armazenamento dos mesmos sobre tabelas, ao invés dos tradicionais ficheiros de texto necessários noutras ferramentas. Para a aplicação da técnica de regras de associação, o algoritmo interno do ODM necessita que os dados estejam numa tabela tipo *Nested*, tal como sugere a figura 34.

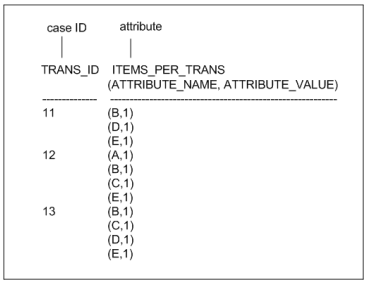


Figura 34 – Representação das transacções no ODM [retirado de (www8)]

Quanto ao segundo caso prático, efectuaram-se as etapas de pre-processamento sobre um conjunto de dados previamente definido e cujas variáveis foram extraídas de dimensões existentes no DW. Seguidamente criaram-se dois sub conjuntos de dados *t1 e t2* para permitir criar e aplicar o modelo induzido.

Na etapa de pre-processamento limparam-se instâncias desnecessárias, fez-se a discretização de alguns campos, a factorização de outros e também a resolução de valores nulos. Foi um trabalho complicado na medida em que os dados e variáveis se encontravam dispersos por diversos DM. Por outro lado algumas variáveis existiam apenas nos sistemas operacionais e foi necessário importar dados manualmente. No apêndice 2 encontram-se algumas variáveis seleccionadas para a tarefa de classificação.

6.4 Fase de criação dos modelos

Naturalmente, nem todos os artigos são comprados. Sendo assim, é perceptível que os produtos de cada transacção estejam bastante esparsos. Segundo o sítio da *Oracle*, o algoritmo *AprioriLike*, implementado pela ferramenta utilizada, está preparado para lidar com esta esparsidade de forma bastante eficiente. Iniciou-se então a criação do primeiro modelo para indução de regras sobre os volumes apresentados no seguinte quadro, seguindo a abordagem de Zheng Z., Kohavi R. & Mason L. (2001):

Transacções	Valores distintos	Máximo tamanho da transacção	Média de cada transacção
≈ 64000	2987	109	6

Quadro 14 – Total de registos utilizados para indução de regras

De forma a permitir diminuir o tempo para a geração de associações o algoritmo aguarda pela definição de uma variável extra, que permite limitar o tamanho da regra, para além das sugeridas por Agrawal *et al* (1993). Os valores iniciais foram sendo ajustados e aplicados, tendo como resultados o seguinte quadro:

Variável	Valor Inicial	Valor N
ASSO_MAX_RULE_LENGTH	∞	∞
ASSO_MIN_SUPPORT	0	80
ASSO_MIN_CONFIDENCE	0	90
Número de regras geradas	≈ 3 milhões	94

Quadro 15 – Valores iniciais de suporte e confiança para o APRIORI-*Like*

O caso prático de Classificação proposto permite fazer a previsão dos atrasos e falhas nas quantidades de uma determinada encomenda. Tal como referido anteriormente dividiu-se o conjunto inicial de dados em dois sub conjuntos distintos *t1* e *t2*. No primeiro sub conjunto, designado por conjunto de treino, as classes já são conhecidas. O outro caso, designado por conjunto de testes, será utilizado para aplicar o modelo criado a partir do conjunto de treino. Como se apresentou no sub capítulo 5.2.4 um dos pontos a ter em consideração na análise dos resultados tem a ver com as estimativas dos erros devolvidos pelos classificadores. A estimativa de erro é de extrema importância antes de se aplicar este modelo sobre dados não classificados. Sendo assim, partindo do primeiro sub conjunto de teste *t1*, divide-se novamente este conjunto em mais dois sub conjuntos menores *t11* e *t12*. A ferramenta permitiu efectuar esta divisão automaticamente através de um processo idêntico ao *Holdout Test* (apresentado no sub capítulo 5.2.4), onde se pôde definir a percentagem desejada para os casos de treino e para os casos de teste. Seguidamente, foram ajustados outros parâmetros tais como a escolha da medida de particionamento *Entropia* ou *índice Gini*, máxima profundidade da árvore, número mínimo de nodos, número mínimo de registos numa partição. O seguinte quadro apresenta esses ajustes.

Descrição	Descrição *Classificador 1 (Default)*	Descrição *Classificador 2*	Descrição *Classificador 3*
GINI/Entropia	Gini	Gini	Entropia
Máxima profundidade	7	8	8
Número mínimo de nodos	10	6	6
Percentagem Mínima de registos num nodo	0.05	0.1	0.1
Número mínimo de registos para uma partição	20	5	5
Percentagem Mínima de registos numa partição	0.1	0.1	0.1
	Seleccionado		

Quadro 16 – Valores iniciais para o *CART-Like*

A figura seguinte permite visualizar a estratégia de particionamento utilizada, ajustando as percentagens da partição para valores entre [60, 40] e [70, 30] (%).

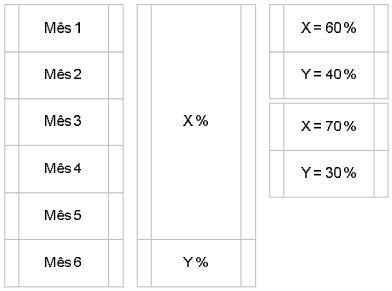


Figura 35 – Variação de percentagens para particionamentos t11 e t12

6.5 Fase de avaliação dos resultados e aplicação

No primeiro caso, depois de variados os parâmetros de *suporte* e *confiança*, induziram-se regras para o nível de granularidade mais baixo, o produto, e também para o nível da categoria. Quando se consideraram como domínio as categorias, ao invés do produto, os resultados foram melhores e obtiveram-se regras mais interessantes e com melhor

desempenho. Muito embora este processo de indução de regras seja um trabalho bastante iterativo, na procura de melhores resultados sucessivamente, os resultados obtidos vão de encontro aos objectivos propostos e permitem verificar a viabilidade deste trabalho. As regras induzidas têm a seguinte representação ao nível da categoria:

"Rule Id","If (condition)","Then (association)","Confidence (%)","Support (%)"

"90","05= 1 AND 03= 1","02= 1","82.0339","7.662269"

Estas regras sugerem que sempre que o fornecedor envia algum produto da categoria X também envia da categoria Y. Por exemplo sempre que o fornecedor envia produtos da sub categoria '05' e '03' então também envia da '02' com uma confiança 82.0339% e suporte 7.66%.

Seguidamente procedeu-se à indução de vários modelos de Classificação, onde se foram ajustando sucessivamente os parâmetros disponibilizados. Os resultados observados nas diversas iterações diferem, uma vez que, após cada iteração, foram-se tentando melhorar os resultados obtidos. Quando foi usada a *Entropia* como medida para particionamento, ao invés do *índice Gini*, visualizaram-se piores resultados e o desempenho também foi mais pobre. A figura seguinte apresenta a matriz de confusão encontrada para o modelo seleccionado para a *classe Atraso* com as suas respectivas ROC e UAC. De salientar que para o particionamento se optou por partir t1=60% e t2=40%.

	Outros	**SIM**
Outros	14297	2176
SIM	3563	5732

Quadro 17 – Matriz de confusão modelo para classe Atraso

A matriz de confusão indica que o modelo conseguiu classificar correctamente (14297+5732=20029) instâncias. Dos dados resultantes ainda se verifica que para a classe Atrasa=NAO o classificador conseguiu prever correctamente 87% dos casos e

para a classe SIM conseguiu prever 61%. Dos modelos analisados este apresenta uma *UAC » 0.872* maior e *Eficácia Média = 0.74*.

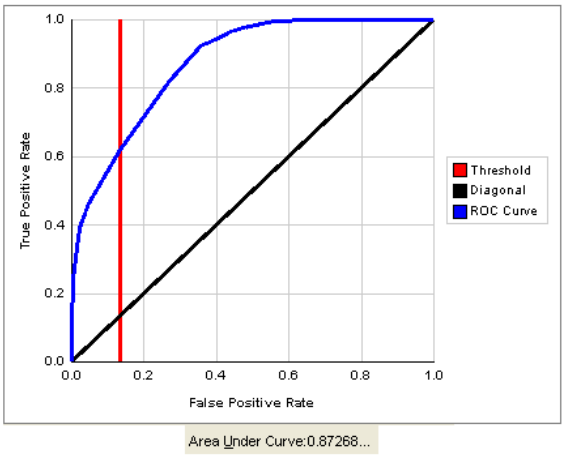


Figura 36 – ROC e UAC para a classe *Atraso_class*

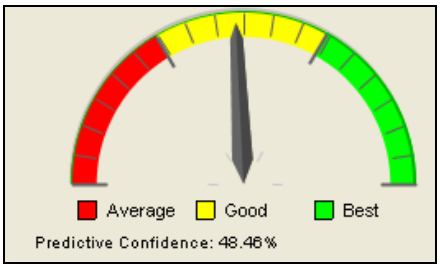


Figura 37 – Confiança do modelo final obtido para a classe *Atraso_class*

Os modelos finais obtidos para a classe *FalhaQty* foram relativamente satisfatórios, apresentando valores de confiança perto dos 45%. O modelo conseguiu classificar bem 55% para casos de *Falha* e 90% para casos de *NãoFalha. Pela* matriz de confusão verifica-se que o modelo conseguiu classificar correctamente 15072 instâncias. Um exemplo das regras geradas encontra-se no apêndice 3. Infelizmente os classificadores baseados em árvores, para estes casos, não são muito eficazes e, portanto, outras técnicas da MD permitirão fazê-lo com melhor qualidade. Sendo assim, também não garantem uma fiabilidade total para classificar novos casos.

What we think, or what we know, or what we believe is, in the end,
of little consequence. The only consequence is what we do.
John Ruskin

Capítulo 7

Conclusões

No início desta dissertação apresentou-se um exemplo bastante simples de uma rotina diária normal por parte de qualquer indivíduo que pretenda comprar algum produto. Deste exemplo diversas perguntas foram colocadas, o que permitiu abstrair daqui que para simples casos qualquer pessoa que decida tenha de ter imensas preocupações devido aos impactos de tais decisões. De seguida, fundamentado em Anderson (2006), apresentaram-se visões realistas sobre o poder interminável da oferta de produtos e da sua procura, referindo o volume crescente de dados que se geram diariamente nas superfícies físicas e virtuais. Enquadrou-se neste capítulo o processo de compras, realçando a importância cada vez mais estratégica deste processo dentro de uma organização. No capítulo seguinte introduziram-se os Sistemas de Suporte à Decisão. Estes sistemas computacionais permitem a qualquer decisor melhorar as suas decisões de negócio baseadas nos dados passados. Neste capítulo introduziram-se conceitos de *Data Warehousing, Business Intelligence e OLAP*, conceitos chave para o entendimento da etapa apresentada no capítulo 3 relativo à DCBD. Sabendo que um DW é um repositório histórico, então a quantidade de dados existente cresce diariamente atingindo volumes impraticáveis de ser manipulados por um ser humano sem o auxílio das tecnologias. Sendo assim há uma oportunidade não só para extrair directamente esses dados utilizando a BI, mas também para induzir/deduzir novo conhecimento a partir de sub conjuntos de dados provenientes do mesmo repositório. O capítulo 5 surge assim

para introduzir conceitos associados com a aplicação de técnicas de MD para a indução de modelos supervisionados e não supervisionados de conhecimento. Aqui foram analisadas com particular detalhe as técnicas de indução de *regras de associação* (supervisionadas) e também de *indução de árvores de decisão* para Classificação *(não supervisionadas).*

Relativamente às r*egras de associação,* considera-se que esta técnica é uma das técnicas mais utilizadas para recolher algum conhecimento inicial de grandes conjuntos de dados, tal como a técnica para segmentação que não foi estudada neste trabalho. Uma das aplicações mais conhecidas relacionadas com a mineração de regras, no retalho, corresponde à descoberta de afinidades entre produtos em transacções – denominada por *Market Basket Analysis.* Seguidamente introduziu-se a *Classificação* e as árvores de decisão. As árvores de decisão permitem induzir conhecimento, como o próprio nome indica, sob a forma de estruturas em árvore, permitindo uma visualização dos resultados clara e intuitiva do conhecimento adquirido.

Depois do estudo teórico, fundamentado em diversas leituras efectuadas, aplicaram-se ambas as técnicas de MD sobre um conjunto de dados de um retalhista no capítulo 6. A aplicação do algoritmo *Apriori-Like* para geração de regras de associação permitiu descobrir associações interessantes. A quantidade de regras geradas foi variando à medida que as medidas de *suporte, confiança e a variável extra máximo de regras* foram sendo afinadas, até que no final seleccionaram-se as apenas as regras de maior confiança para o volume de dados proposto. Relativamente ao desempenho obtido, o algoritmo teve um comportamento bastante aceitável dado o número reduzido de produtos. Quanto à aplicação do algoritmo *CART-Like* para Classificação conseguiu-se criar um modelo relativamente eficaz quando aplicado sobre casos reais não classificados.

Os objectivos propostos no início deste trabalho visavam optimizar a cadeia de abastecimento de um hipermercado. Através das técnicas apresentadas prova-se que a

partir do momento que se consegue analisar as afinidades entre produtos recebidos se pode optimizar o processo de compra através da análise criteriosa das regras resultantes. As classificações feitas pelos modelos de Classificação permitem prever possíveis falhas nos atrasos e quantidades, antecipando assim uma possível perda financeira. No caso prático apresentado, no entanto, o classificador não possui uma confiança elevada para ambas as classes.

7.1 Limitações desta investigação

A presente dissertação ficou limitada pelas condições impostas pela organização, confidencialidade, assim como as condições disponibilizadas sobre dados reais de produção e acessos a ambientes operacionais. Infelizmente, talvez por se tratar de uma área distinta da tradicional compra de consumidores em hipermercados, onde existem informação de clientes, a análise das compras organizacionais apresentam dificuldades acrescidas devido à inexistência e/ou inconsistência de informação relevante relativa a fornecedores dentro do DW e à necessidade de um conjunto vasto de variáveis para efectuar eficazmente este estudo. Por outro lado, o próprio DW possui imensas métricas que aparentemente são interessantes, visíveis pelas suas próprias descrições, mas que não são devidamente carregadas nos sistemas operacionais, invalidando a possibilidade de efectuar uma análise mais específica utilizando essas colunas.

7.2 Contribuições

As contribuições deste trabalho abrangem duas áreas distintas: uma área funcional (orientada para a indústria de comércio retalhista e grossista) e outra área mais técnica focada nas tecnologias de informação. Relativamente à tecnologias de informação, este trabalho permitiu: a) alargar os conhecimentos nas áreas de Suporte à Decisão com particular realçe para as áreas de *Data Warehousing* e *Business Intelligence*; b) introduzir a DCBD, aplicando técnicas de MD; c) criar um modelo de conhecimento supervisionado e outro não supervisionado, por indução de regras de associação e também por indução de árvores de decisão para classificação. Quanto à área funcional o

trabalho permitiu: a) descobrir afinidades entre produtos tendo como base as compras efectuadas pelos consumidores nas estruturas físicas e também virtuais; b) analisar afinidades de produtos recepcionados, sabendo que os fornecedores muitas vezes substituem certos produtos por outros similares, devido a falhas de *stock*; c) indução e dedução de novo conhecimento para previsão de atrasos e falhas nas encomendas.

7.3 Considerações finais e trabalho futuro

Esta dissertação representa o culminar de um processo pessoal e profissional, de aprendizagem. No decorrer deste trabalho consolidaram-se conhecimentos na área de Sistemas de Suporte à Decisão com particular interesse na área da DCBD utilizando a MD. Sendo este um projecto longo, complexo, aplicado num domínio muito específico do retalho, cada vez mais suportado pelas Tecnologias de Informação, foi possível alargar o conhecimento de negócio focado no processo de compras e assim obter uma visão mais alargada sobre o funcionamento de todo o processo de abastecimento de um hipermercado. Desta forma releva-se a importância, cada vez mais estratégica, da função compras e a relação de colaboração necessária entre fornecedores e compradores – processo fulcral nos dias que correm.

As duas componentes práticas podem também ser optimizadas em termos computacionais com a utilização de ferramentas com algoritmos mais recentes e com melhores desempenhos.

Ainda para este trabalho mais variáveis deveriam ter sido analisadas na tarefa de Classificação. No entanto, devido às dimensões do DW e às limitações já apresentadas, seleccionaram-se apenas um conjunto mínimo válido para o problema proposto. Relativamente às técnicas aplicadas, ficaram por descrever diversas técnicas, nomeadamente: a segmentação, as redes neuronais, as SVM (*Support Vector Machines*), os classificadores *Naíve Bayes e* outros. Este trabalho deverá ser realizado numa próxima etapa.

Referências bibliográficas

[1] Agrawal, R., Imielinski, T., & Swami, A. (1993). *Mining Association rules between sets of items in large databases.* In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data: Washington DC (USA), pp. 207-216.

[2] Agrawal, R., & Srikant, R. (1994). *Fast algorithms for mining association rules*. In: Proceedings of the 20th International Conference on Very Large Databases (VLDB), Santiago (Chile), pp. 487 - 499.

[3] Agrawal R., Gupta A. & Sarawagi S. (1997). *Modeling Multidimensional Databases.* 13th International Conference on Data Engineering. Birmingham, England.

[4] Anderson C. (2006). *The Long Tail: How endless choice is Creating Unlimited Demand*. Random House Business Books.

[5] Bayardo *et al* (1999). *Constraint-Based Rule Mining in Large, Dense Databases.* Proceedings of the 15th Int'l Conf. on Data Engineering, 1999, pp 188-197.

[6] Berry & Linoff (1997). *Data Mining Techniques – For Marketing, Sales and Customer Support.* John Willey e Sons, New York.

[7] Borges (S/D). *Toward a new supermarket layout: from industrial categories to one stop shopping organization throught a data mining approach*. Reims Management School, França.

[8] Brijs (2002). *Retail Market Basket Analysis – A quantitative modeling approach*. Dissertation at the Limburg University Center.

[9] Brijs *et al* (1999). *Using Association Rules for Product Assortment Decisions: A Case Study*. Limburg University Centre, Department of Applied Economic Sciences, Belgium.

[10] Brijs *et al* (S/D). *A Data Mining Framework for Optimal Product Selection in Convenience Stores*. Limburg University Center.

[11] Brin *et al* (1997). *Dynamic itemset counting and implication rules for market basket data*. SIGMOD Record 26(2), pp 255–264.

[12] Blum A.; Kalai A. & Langford J. (1999). *Beating the Hold-out – Bounds for K-fold and Progressive Cross-Validation – Computational Learing Theory.* In Proceedings of the International Conference on Computational Learning Theory, pp. 203-208.

[13] Chapman *et al* (2000). *CRISP-DM 1.0 – Step by Step data mining guide.* SPSS.

[14] Cabena *et al* (1997). *Discovering Data Mining – From Concept to Implementation*. Prentice- Hall.

[15] Egan J. (1975). *Signal Detection Theory and ROC analysis*. New York, Academic Press.

[16] Darade *et al* (S/D). *Supplier Evaluation for Part Procurement Using ISA.* Jawaharlal Nehru Engineering College, India.

[17] Davenport T. (2006). *Competing on Analytics*. Harvard Business Review.

[18] Fayyad U., Piatetsky-Shapiro G. & Smyth P. (1996). *From Data Mining to Knowledge Discovery in Databases.* American Association for Artificial Intelligence.

[19] Fawcett T. (2003). *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers.* HP Laboratories.

[20] Fawcett T. (2005). *An Introduction to ROC analysis.* Science Direct, Pattern Recognition Letters, pp 861-874.

[21] Grossman R. (2004). *Data Mining Standards, Services and Platforms.* University of Illinois, Chicago.

[22] Ganti *et al* (1999). *RainForest – A Framework for Fast Decision Tree Construction of Large Datasets*. Department of Cmputer Science, University of Wisconsin, Madison.

[23] Han *et al* (2001). *Data mining – Concepts and Techniques*. Morgan Kaufmann Publishers.

[24] Hartmann *et al* (2001). *Determining the Purchase Situation: Cornerstone of Supplier Relationship Management.* Competitive Paper submitted to the 17th Annual IMP Conference at the Norwegian School of Management BI, Oslo, Norway.

[25] Hornick *et al* (2007). *Java Data Mining / Strategy, Standard, and Practise – A practical Guide for Architecture, Design, and Implementation.* Morgan Kaufmann.

[26] Haery *et al* (2008). *Application of Association Rule Mining in Supplier Selection Criteria*. Proceedings of world academy of science, engineering and technology.

[27] Hipp *et al* (2000). *Algorithms for Association Rule Mining – A General Survey and Comparison*. SIG KDD.

[28] Hamdani *et al* (S/D). *Interesting Measures for Mining Association Rules.* FAST-NUCES, Lahore.

[29] Ray H. (2004). *Using CHAID for classification problems*. New Zealand Statistical Association 2004 Conference.

[30] Witten *et al* (2005). *Data Mining: Practical machine learning tools and techniques.* 2nd Edition, Morgan Kaufmann.

[31] Inmon W. (1996). *Building the Data Warehouse*. John Wiley & Sons.

[32] Girão I. *et al* (2000). *Teoria da decisão: Difícil decidir*. Brasil

[33] Kadav *et al* (S/D). *Data Mining Standards*. Indian Institute of Technology, Kanpur.

[34] Kohavi R. & Provost F. (1998). *On applied research in Machine Learning. Editorial for the special issue on applications of Machine Learning and the Knowledge Discovery Process*, Vol. 30, Nº 2/3.

[35] Kohavi R. (1995). *A Study of Cross-Validation and BootStrap for Accuracy Estimation and Model Selection*. Computer Science Department, Stanford University.

[36] Larose D. (2005). *Discovering Knowledge in Data – An introduction to Data Mining*. Wiley.

[37] Lee *et al* (2002). *An intelligent supplier management tool for benchmarking suppliers in outsources manufacturing*. Expert Systems with Applications, 22, pp. 213-224.

[38] Liu H. *et al* (2003). *The Handbook of Data Mining.* Lawrence Erlbaum Associates, Inc.Publishers.

[39] Shih *et al* (1997). *Split selection methods for classification trees*. Statistica Sinica, vol. 7, pp. 815-840.

[40] Kimball R. (1996). *The Data Warehouse ToolKit.* New York.

[41] Kimball R. (2007). *Dimensional Relational vs. OLAP: The Final Deployment Conundrum.* New York.

[42] Kononenko *et al* (1995). *Induction of decision trees using RELIEFF*. University of Ljubljana, Faculty of electrical engineering & computer science.

[43] Kononenko *et al* (1997). *Attribute selection for modeling.* University of Ljubljana, Faculty of electrical engineering & computer science.

[44] Kass G. (1980). *An Exploratory Technique for Investigating Large Quantities of Categorical Data*. Journal of Applied Statistics.

[45] Mladenic *et al* (S/D). *Exploratory analysis of Retail Sales of Billions of items.* Carnegie Mellon University, Pittsburgh.

[46] Morton M. (1978). *Some Perspectives on Computerized Management Decision.* IEEE "COMPSAC 77" Conference, Chicago.

[47] Mehta *et al* (1996). *SLIQ: A fast Scalable Classifier for Data Mining*. IBM Almaden Research Center, San Jose, CA.

[48] Moura B. (2006). *Logística – Conceitos e Tendências*. CentroAtlantico.pt, Lisboa.

[49] Newman A. & Cullen P. (2002). *Retailing: Environment & Operations.* Thomson Learning.

[50] Navega S. (2002). *Princípios Essenciais do Data Mining*. Intelliwise Research and Training, São Paulo.

[51] Oracle (2005). *Oracle Data Mining 10g Release 2 – Know More, Do More, Spend Less*. Oracle.

[52] Olson D. & Delen D. (2008). *Advanced Data Mining Techniques.* Springer.Verlag Berlin Heidelberg.

[53] Piatetsky-Shapiro G. (1991). *Knowledge Discovery in Databases.* A report on the IJCAI-89 Workshop, AI Magazine.

[54] Piatetsky-Shapiro G. (1992). *Discovery, analysis, and presentation of strong rules.* Knowledge Discovery in Databases, pp. 229-248.

[55] Ponniah P. (2001). *Data Warehousing Fundamentals – A comprehensive guide for IT professionals.* Wiley Student Edition.

[56] Pendse N. (2001). *What is OLAP?* OLAPreport.com.

[57] Power J. (2007). *A Brief History of Decision Support Systems.* DSSResources.com.

[58] Quinlan *et al* (1996). *Induction of decision trees*. Centre for Advanced Computing Sciences, Kluwer Academic Publishers, Boston.

[59] Rountree N. (1999). *Further Data Mining – Building Decision Trees.* N/A.

[60] Sheikh L., Tanveer B., Hamdani S. (S/D), *Interesting Rules for Mining Association Rules*. Lahore

[61] Smith *et al* (2005). *Oracle Database 10g Data Warehousing*. Elsevier Digital Press. UK.

[62] Silva M. (S/D). *Mineração de dados - Conceitos, Aplicações e Experimentos com Weka*. Universidade do Estado do Rio Grande do Norte, Instituto Nacional de Pesquisas Espaciais.

[63] Savasere A. *et al* (1995). *An efficient algorithm for Mining Association Rules in Large Databases.* Georgia Institute of Technology, Atlanta.

[64] Simon A. (1995). *Strategic Database Technology: Management for the year 2000*. Morgan Kaufmann Publishers Inc.

[65] Sucky E. (S/D). *A dynamic model for supplier selection.* Department of Supply Chain Management, Goethe University.

[66] Symeonidis *et al* (2005). *Agent Intelligence through Data Mining.* Springer, Verlag Berlin Heidelberg.

[67] Schmidt-Thieme L. (S/D). *Algorithmic Features of Eclat*. Computer based New Media Group, Institute of Computer Science, University of Freiburg, Germany.

[68] Shafer *et al* (1996). *SPRINT – A Scalable Parallel Classifier for Data Mining*. IBM Almaden Research Center, San Jose, CA.

[69] Tan P., Steinbach M. & Kumar V. (2005). *Introduction to Data Mining*. Addison-Wesley.

[70] Tan *et al* (2002). *Selecting the right interestingness measure for Association Patterns*. Department of Computer Science and Engineering, University of Minnesota.

[71] Vuk M. & Curk T. (2006). *ROC Curve, Lift Chart and Calibration Plot*. Department of Knowledge Technologies e University of Ljubljana, Faculty of Computer and Information Sciente, Slovenia.

[72] Williams S. (2003). *The business value of Business Intelligence*. Decision Path Consulting, Business Intelligence Journal.

[73] Tsur *et al* (1997). *Dynamic itemset counting and implication rules for market basket data*. In: Proceedings of the ACM SIGMOD Int'l Conf. on Management of Data.

[74] Zheng Z., Kohavi R. & Mason L. (2001). *Real World Performance of Association Rule Algorithms*. Blue Martini Software, KDD 2001.

[75] Zaki, M.J., Parthasarathy, S., Ogihara, M. & Li, W. (1997). *New algorithms for fast discovery of association rules*, Technical Report no.TR651.

Referências WWW

[www1] http://www.kdnuggets.com

[www2] http://www.sybase.com/guinness

[www3] http://www.crisp-dm.org

[www4] http://www.dssresources.com

[www5] http://www.olapreport.com

[www6] http://sourceforge.net/projects/pmml

[www7] http://www.twocrows.com/

[www8] http://www.olapcouncil.org

[www9] http://www.oracle.com

Apêndice 1

Standards para MD

Areas	*Standard*	Descrição
Processos	***CRISP-DM***	**Conjunto de etapas para desenvolver um projecto de MD**
XML	***PMML***	**Modelo para representação de dados estatísticos e de MD**
	CWM-DM	Modelo para metadados
API's	*SQL/MM, Java API (JSR-73), Microsoft OLE DB*	API para MD
Protocolos de transporte	*Data Space Protocol (DSTP)*	Protocolo utilizado para distribuir, pesquisar e retornar dados num determinado espaço de dados
Scoring Standard	*Predictive Scoring and Update protocol (PSUP)*	Avaliação *Online* e tempo real
Web	*XML para análise (XMLA)*	*Web Service interface*
	Semantic Web	Providencia uma *framework* para representar informação numa forma que permita ser processada pela máquina e pode ser usada para extrair conhecimento de sistemas de MD.
	Espaço de dados	Providencia uma estrutura para criar uma web de dados.
Grid	*Open Grid Standard Architecture*	Desenvolvida pela Globus, basei a-se numa arquitectura para organizações distribuídas virtuais

Quadro 18 – Alguns *standards* para MD segundo Kadav (S/D)

Apêndice 2

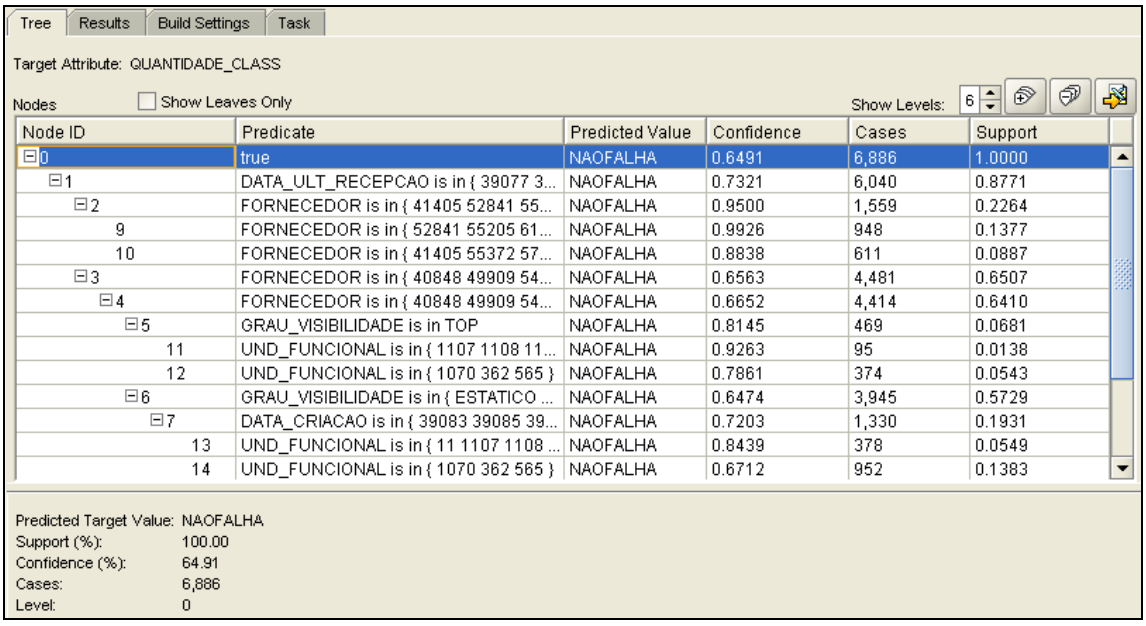
Algumas variáveis base seleccionadas

Variáveis	Descrição
ID	Identificador – Chave única – gerada
FORNECEDOR	Fornecedor da mercadoria
CATEGORIA	Categoria dos artigos
BUYER	Comprador
ARTIGO	Descrição do artigo
UND_FUNCIONAL	Código da Loja ou Armazém
DISTANCIA	Distancia entre entidades (*binned*)
TIPO_PRIORIDADE	Prioridade da encomenda efectuada {1,2}
QUANTIDADE_ENCOMENDA	Quantidade de encomenda (*binned*)
QUANTIDADE_RECEBIDA	Quantidade recebida real (*binned*)
NACIONAL	{S,N}
DESCR_PAIS	Descrição do País
FORNECEDOR_INTERNO	{S,N}
FORNECEDOR_P	{S,N}
GRAU_VISIBILIDADE	{TOP, SUPERDINAMICO, DINAMICO, REGULAR}
TRANSPORTE	Método de transporte
TP_MARCA	Tipo de marca do produto
COD_SECCAO	Descrição da Secção
COD_FAMILIA	Descrição da Família
COD_SUB_FAMILIA	Descrição da Sub Família
DIA_SEMANA_ENC	Dia da semana em que ocorreu a encomenda
DIA_SEMANA_REC	Dia da semana em que ocorreu a recepção
ATRASO_CLASS	{Atrasa; NaoAtrasa }
QUANTIDADE_CLASS	{Falha; NaoHaFalhas }

Quadro 19 – Esquema das variáveis e classes

Apêndice 3

Algumas regras geradas para a Classe Falha Quantidade



Tree | Results | Build Settings | Task

Target Attribute: QUANTIDADE_CLASS

Nodes ☐ Show Leaves Only Show Levels: 6

Node ID	Predicate	Predicted Value	Confidence	Cases	Support
⊟0	true	NAOFALHA	0.6491	6,886	1.0000
⊟1	DATA_ULT_RECEPCAO is in { 39077 3...	NAOFALHA	0.7321	6,040	0.8771
⊟2	FORNECEDOR is in { 41405 52841 55...	NAOFALHA	0.9500	1,559	0.2264
9	FORNECEDOR is in { 52841 55205 61...	NAOFALHA	0.9926	948	0.1377
10	FORNECEDOR is in { 41405 55372 57...	NAOFALHA	0.8838	611	0.0887
⊟3	FORNECEDOR is in { 40848 49909 54...	NAOFALHA	0.6563	4,481	0.6507
⊟4	FORNECEDOR is in { 40848 49909 54...	NAOFALHA	0.6652	4,414	0.6410
⊟5	GRAU_VISIBILIDADE is in TOP	NAOFALHA	0.8145	469	0.0681
11	UND_FUNCIONAL is in { 1107 1108 11...	NAOFALHA	0.9263	95	0.0138
12	UND_FUNCIONAL is in { 1070 362 565 }	NAOFALHA	0.7861	374	0.0543
⊟6	GRAU_VISIBILIDADE is in { ESTATICO ...	NAOFALHA	0.6474	3,945	0.5729
⊟7	DATA_CRIACAO is in { 39083 39085 39...	NAOFALHA	0.7203	1,330	0.1931
13	UND_FUNCIONAL is in { 11 1107 1108 ...	NAOFALHA	0.8439	378	0.0549
14	UND_FUNCIONAL is in { 1070 362 565 }	NAOFALHA	0.6712	952	0.1383

Predicted Target Value: NAOFALHA
Support (%): 100.00
Confidence (%): 64.91
Cases: 6,886
Level: 0

Quadro 20 – Regras de Classificação para prever valor NAOFALHA