Universidade do Minho
Escola de Engenharia

Eva Margarida Correia Duarte

Técnicas de Mineração de Dados para Suporte à Decisão no Planeamento de Horários em Empresas de Transportes Públicos

Novembro de 2008

Universidade do Minho
Escola de Engenharia

Eva Margarida Correia Duarte

Técnicas de Mineração de Dados para Suporte à Decisão no Planeamento de Horários em Empresas de Transportes Públicos

Mestrado Integrado em Informática

Trabalho efectuado sob a orientação do
Prof. Orlando Manuel de Oliveira Belo
co-orientação do
Prof. João Pedro Carvalho Leal Mendes Moreira

Novembro de 2008

Técnicas de Mineração de Dados para suporte à decisão no Planeamento de Horários em Empresas de Transportes Públicos

Eva Margarida Correia Duarte

Dissertação apresentada à Universidade do Minho para obtenção do grau de Mestre em Informática, elaborada sob orientação do Professor Doutor Orlando Manuel de Oliveira Belo.

2008

Agradecimentos

Ao professor Orlando Belo, por me ter orientado ao longo de todo este projecto.
Ao professor João Moreira, por me ter incentivado a realizar este trabalho e por toda a ajuda e sugestões fornecidas.
Ao professor Jorge Freire, administrador da STCP, que tornou possível a realização deste trabalho.
Ao Engenheiro José Miguel Magalhães, ao Engenheiro Carlos Abreu e ao Pedro Gonçalves, da STCP, pela ajuda e sugestões que me deram.
Ao Jorge, por estar sempre a meu lado e por me apoiar e incentivar nos bons e nos maus momentos.
Aos meus pais e irmãos que estiveram a meu lado durante todo o meu percurso de vida académica que agora termina.

Resumo

Técnicas de Mineração de Dados para suporte à decisão no Planeamento de Horários em Empresas de Transportes Públicos

A fiabilidade dos sistemas de transportes públicos de passageiros é uma das maiores preocupações tanto dos passageiros como das próprias empresas que fornecem o serviço. Os avanços tecnológicos ocorridos nas últimas décadas permitiram que as empresas de transportes públicos armazenassem grandes quantidades de informação acerca das viagens realizadas. Isto possibilita que essa informação seja analisada posteriormente, podendo-se assim identificar erros de planeamento e padrões de comportamento que podem ser utilizados para fornecer uma melhoria do serviço no futuro. Este trabalho foi feito tendo como alvo de estudo a STCP, uma empresa de transportes públicos de passageiros, que pretende melhorar o desempenho do seu sistema no que toca ao cumprimento dos horários, de forma a aumentar a satisfação do cliente e diminuir os prejuízos da empresa decorrentes dos sucessivos incumprimentos. As técnicas de mineração de dados podem ser uma ferramenta poderosa para extrair informação útil para as empresas de transportes públicos, no sentido de melhorar o cumprimento dos horários. Foi feito um estudo de técnicas de mineração de dados que poderiam ser utilizadas para extrair esta informação, tendo-se optado por utilizar Árvores de Decisão, Descoberta de CARs (*Class Association Rules*) e Classificação Baseada em Associação. Realizaram-se diversas experiências com diferentes conjuntos de dados utilizando estas técnicas, analisando as potencialidades e os pontos fracos de cada uma delas. Após a análise de todas as técnicas, conclui-se que a técnica de Classificação Baseada em Associação é a mais adequada para o problema em causa, tendo em conta os objectivos que se pretendiam atingir. Em comparação com as outras técnicas estudadas, CBA demonstrou ter melhor desempenho no conjunto de características mais importantes: esta técnica fornece regras com elevado factor de confiança, tem boa capacidade de detecção de erros sistemáticos e boa interpretabilidade.

Abstract

Data Mining Techinques for Schedule Planning Decision Support in Public Transport Companies

Reliability of public transport systems is one of the major concerns of both passengers and public transport companies. The technological evolution that has occurred in the last decades made possible to companies the storage of large amounts of information about the trips. This information can be used in later analysis, making possible the identification of planning errors and behaviour patterns, which can be used to supply a better service in the future.

The case study used in this work was STCP, a public transport company which intends to improve its service in terms of schedule reliability, in order to increase clients satisfaction and to decrease damages caused by successive schedule deviations. Data mining techniques can be a powerful tool to extract useful information to public transport companies in order to improve its schedule reliability. In this study, we used Decision trees, Class Association Rules and Classification Based in Association. Various experiments with different data sets were done using these techniques in order to analyse the strong and weak points of each one. After the analysis of all the techniques, the results suggested that Classification Based in Association is the most adequate data mining technique to use in this case, concerning to business objectives. Compared with the other techniques, CBA has the best results in the most important characteristics analyzed: it has a good performance in detecting systematic schedule deviations, it gives high confidence rules and it has good interpretability.

Índice

Lista de Abreviaturas

A

AML – Amial

C

CAR – Class Association Rules

CBA – Classification Based in Assocations

CBA-RG – Classification Based in Associations – Rule Generator

CF – Confidence Factor

CMAR – Classification Based on Multiple Association Rules

CMP – Campanhã

CPAR – Classification Based on Predictive Association Rules

CQ – Castelo do Queijo

CRISP-DM – Cross-Industry Standard Process for Data Mining

CVP – Critical Value Pruning

D

DW - Data Warehouse

E

EBP – Error Based Pruning

G

GARC – Gain based Association Rule Classification

H

HSJ – Hospital de São João

K

KDD – Knowledge Discovery in Databases

M

MCCP – Minimal Cost-Complexity Pruning

MEP – Minimum Error Pruning

P

PEP – Pessimistic Error Pruning

R

RAEP – Rotunda AEP

REP – Reduced Error Pruning

S

SR – São Roque

Índice de Figuras

Índice de Tabelas

Capítulo 1

Introdução

1.1 Transportes públicos rodoviários

A fiabilidade dos sistemas de transportes públicos de passageiros é uma das maiores preocupações tanto dos passageiros como das próprias empresas que fornecem o serviço. Tempos de espera demasiado elevados, chegadas atrasadas ou adiantadas aos destinos e conexões perdidas podem induzir nos passageiros sentimentos de insatisfação perante o sistema (Liu and Sinha 2007). Para além dos problemas causados pela perda de clientes, a falta de fiabilidade do serviço pode forçar as empresas de transportes a activar recursos adicionais numa tentativa de fazer cumprir os horários e satisfazer a procura por parte dos clientes, o que resulta num aumento dos custos para a empresa (Strathman, Dueker et al. 1999).

A fiabilidade do serviço prestado pelas empresas de transportes públicos é um assunto que tem vindo a ser estudado por diversos investigadores (Carey 1994; Bates, Polak et al. 2001; Rietveld, Bruinsma et al. 2001; Chen, Skabardonis et al. 2003). Existem vários indicadores que podem ser utilizados para medir a fiabilidade do serviço, sendo que o mais utilizado é a aderência ao horário (Strathman, Dueker et al. 1999). No caso particular das empresas de transporte rodoviário, a aderência ao horário mede a probabilidade de um autocarro estar no local correcto à hora correcta, de acordo com o horário previsto. Habitualmente, as empresas de transportes públicos rodoviários informam os clientes acerca dos horários previstos para cada percurso, não apenas para o início e fim de viagem, mas também para algumas paragens intermédias ao longo

do percurso, normalmente denominadas por *pontos de horário.* É prática habitual as empresas considerarem que uma viagem está dentro do horário previsto se o autocarro chega a um ponto de horário não mais do que 1 minuto adiantado ou 5 minutos atrasado (Strathman and Hopper 1993; Strathman, Dueker et al. 1999). Quando os autocarros operam de forma consistente segundo esta janela temporal, os utilizadores podem programar a sua chegada à paragem de forma a minimizar o tempo de espera, com a confiança de que o autocarro não terá ainda partido e que o seu tempo de espera não será demasiado elevado.

Todavia, a pontualidade das chegadas e das partidas em cada paragem é muitas vezes difícil de manter, uma vez que existem muitos factores que podem afectar o cumprimento dos horários previstos. Estes factores são divididos por Liu e Sinha (2007) (Liu and Sinha 2007) nos seguintes grupos: características de tráfico, como os níveis de congestionamento diário; características do percurso, como a localização das paragens, o número de intersecções e o tamanho do percurso; características dos passageiros, como o número de entradas e saídas em cada paragem e a variabilidade do volume de passageiros; características operacionais, tais como o sistema de construção dos horários, o sistema de bilhética utilizado ou ainda a experiência e comportamento do condutor.

Uma vez que a fiabilidade do serviço é um indicador fundamental tanto para empresas como passageiros, as empresas recorrem a diversos métodos para tentar reduzir os efeitos dos factores mencionados anteriormente, tais como (Carey 1998; El-Geneidy, Horning et al. 2007): fazer compassos de espera em algumas paragens de forma a que o horário seja cumprido nos pontos de horário quando o veículo circula adiantado; modificações no percurso (tamanho, localização das paragens, entre outros); implementação de controlo na hora a que as viagens estão a decorrer, através da introdução de veículos adicionais quando necessário ou de indicações dadas aos motoristas para abrandar a marcha, acelerar ou desviar a sua rota; alocação de tempo adicional para a realização das viagens, por forma a compensar eventuais atrasos.

Tal como referido por Rietveld *et al.* (2001) (Rietveld, Bruinsma et al. 2001), a fiabilidade dos transportes públicos está intimamente ligada com a construção de horários. Quando constroem novos horários, ou reajustam os já existentes, os planeadores tentam melhorar a fiabilidade do serviço ou, no mínimo, manter os níveis já existentes. A construção de horários para sistemas de transportes públicos é um assunto que tem vindo a ser estudado por diversos investigadores. Ceder (1986) propôs diversos métodos para construir horários de autocarros utilizando informação relativa ao número de passageiros (Ceder 1987); Palma *et al.* (2000) propuseram soluções para horários “óptimos” numa determinada linha onde os utilizadores diferem relativamente às horas a que preferem viajar, associando custos ao facto de estes não viajaram à hora que desejariam (Palma and Lindsey 2001); Ceder *et al.* (2000) analisaram o problema de gerar horários para uma

dada rede de forma a maximizar a sincronização e descreveram um algoritmo para resolver um problema em tempo polinomial (Ceder, Golany et al. 2001).

A análise das viagens (tanto na hora em que estas se realizam como posteriormente) é uma tarefa que permite identificar as falhas nos horários que se encontram em vigor e fazer os necessários reajustamentos. Em particular, a análise da pontualidade não apenas no início e fim da viagem, mas também em cada um dos pontos de horário pode conduzir a consideráveis melhoramentos na qualidade do serviço. Isto porque, atrasos ou chegadas antes do tempo no início do percurso podem contribuir para um fraco desempenho ao longo de toda a viagem, especialmente se estratégias de controlo e espera nos pontos de horários não forem utilizadas (Strathman and Hopper 1993). Para além disso, o custo para o utilizador de uma viagem perdida pode ser muito elevado, especialmente em percursos com baixas frequências, o que realça a importância da pontualidade nos pontos de horário.

Durante muito tempo, esse tipo de análise apenas era possível através das informações dadas pelos motoristas ou pelos passageiros. No entanto, a evolução de algumas tecnologias como o GPS permitiu às empresas de transportes públicos criarem sistemas de controlo dos seus veículos. Isto possibilita o armazenamento de informação detalhada acerca das viagens realizadas, nomeadamente a hora a que começou e terminou a viagem, os locais por onde passou e respectiva hora, entradas e saídas de passageiros, o motorista que conduzia o veículo, entre outros. Vários estudos demonstram que a utilização destes dados no momento ou para posterior análise pode contribuir para uma melhoria do desempenho do sistema (Strathman and Hopper 1993; Ding and Chien 2001; Strathman, Kimpel et al. 2002). Isto porque, o facto de se ter conhecimento acerca da localização de cada veículo a cada momento permite que no decorrer das viagens sejam tomadas medidas no sentido de prevenir eventuais desvios do horário previsto ou compensar eventuais atrasos. Para além disso, uma análise posterior destes dados permite identificar erros de planeamento e padrões de comportamento que podem ser utilizados para fornecer uma melhoria do serviço no futuro.

1.2 Motivação e Objectivos

Os avanços tecnológicos ocorridos nas últimas décadas permitiram que as empresas de transportes públicos armazenassem grandes quantidades de informação acerca das viagens realizadas. Estes dados estão geralmente armazenados em sistemas de *data warehousing* (DW), repositórios de dados especificamente orientados para a análise de dados e suporte à decisão. Estes sistemas reúnem informação proveniente de diversas fontes de dados, efectuando a sua limpeza, tratamento e armazenamento. A informação fica assim disponível de forma organizada e consolidada, podendo ser utilizada pelos analistas para extrair conhecimento útil para a empresa.

Normalmente, os sistemas de DW possuem ferramentas próprias para a extracção e visualização da informação guardada. No entanto, estas aplicações podem não ser suficientes para satisfazer as necessidades dos analistas, sendo que muitas vezes outras ferramentas são construídas à medida das suas necessidades.

Tal como já foi referido anteriormente, a qualidade do serviço oferecido pelas empresas de transportes públicos é muitas vezes medida através da aderência aos horários, ou seja, a probabilidade de um autocarro estar no local previsto à hora prevista. Torna-se por isso fundamental uma análise comparativa entre os horários pré-estabelecidos e os tempos de viagem reais, tentando perceber em que circunstâncias os horários não são cumpridos e os motivos pelos quais isso acontece. Isto irá permitir à empresa a aplicação de medidas que possibilitem um melhor cumprimento dos horários. Este tipo de análise poderia ser feita utilizando estatística descritiva, no entanto, esta não permite a extracção de informação mais complexa, como a detecção automática de erros sistemáticos e padrões de comportamento. Para obtermos este tipo de informação, é necessário por isso recorrer a outras técnicas de análise de dados, sendo a mineração de dados (*data mining*) uma alternativa possível.

A mineração de dados é um processo que consiste na aplicação de análise de dados e algoritmos de descoberta que, usualmente, produzem padrões ou modelos acerca dos dados (Fayyad and Uthurusamy 1996). Redes neuronais, árvores de decisão, regras de associação e algoritmos genéticos são apenas alguns exemplos de técnicas que podem ser utilizadas para extrair esse conhecimento. No entanto, antes de serem aplicadas estas técnicas, os dados devem passar por várias fases de pré-processamento como limpeza, selecção e transformação. A mineração dos dados propriamente dita pode por isso ser considerada apenas como uma fase de um processo maior denominado *Knowledge Discovery in Databases* (KDD) – em português, Descoberta de Conhecimento em Bases de Dados – que pode ser definido como o processo não trivial de identificação de padrões válidos, desconhecidos, potencialmente úteis e compreensíveis, nos dados (Fayyad and Uthurusamy 1996).

Este trabalho foi feito com base em dados provenientes de uma empresa de transportes públicos que tem como principal objectivo melhorar o desempenho do seu sistema no que toca ao cumprimento dos horários, de forma a aumentar a satisfação do cliente e diminuir os prejuízos da empresa decorrentes dos sucessivos incumprimentos. A empresa pretende por isso detectar situações sistemáticas em que o horário previsto não é cumprido e identificar as condições em estas ocorrem. Para atingir esse fim, foram delineados os seguintes objectivos para este trabalho:

- Estudar técnicas de mineração de dados e desenvolver modelos que sejam capazes de fornecer o tipo de informação pretendida;
- Analisar a qualidade da informação fornecida pelos modelos desenvolvidos, verificando se esta vai de encontro aos objectivos da empresa;

- Efectuar uma comparação das técnicas estudadas, analisando os prós e contras de cada uma, para que no futuro se possa proceder ao desenvolvimento de uma aplicação informática que integre uma ou mais destas técnicas.

1.3 Estrutura do Documento

Além do presente capítulo, onde se apresentam as linhas gerais do trabalho que foi realizado, este documento encontra-se estruturado como se descreve em seguida. No Capítulo 2 é descrito o caso de estudo que é utilizado neste trabalho. É ainda descrita a metodologia utilizada para resolver o problema em causa, bem como as técnicas de mineração de dados e aplicações informáticas utilizadas para a realização deste trabalho. No Capítulo 3 é descrito o pré-processamento dos dados, que inclui tarefas de compreensão e preparação dos dados. É feita a análise dos dados disponíveis, a avaliação da sua qualidade, limpeza e construção do conjunto de dados final através da junção de tabelas e derivação de novos atributos. De seguida, no Capítulo 4, são apresentados os testes efectuados na fase de modelação e os resultados obtidos, para cada uma das técnicas de mineração de dados testadas. Por mim, no capítulo 5, são apresentadas as conclusões do trabalho, através de uma análise comparativa dos modelos obtidos, a avaliação do trabalho efectuado e perspectivas para trabalho futuro.

Capítulo 2

Análise de uma Linha de Transportes Rodoviários

2.1 O alvo de estudo

A empresa que foi alvo de estudo neste trabalho foi a STCP, Sociedade de Transportes Colectivos do Porto, SA. A STCP é o maior operador de transporte público urbano de passageiros do Grande Porto, desenvolvendo a sua actividade num cenário misto: monopólio legal do modo rodoviário no Porto e concorrência com os demais operadores fora dos limites da cidade. A empresa serve cerca de 1,3 milhões de habitantes, distribuídos por 6 concelhos, através das 94 linhas que compõem os 496 quilómetros de rede.

Habitualmente, a análise do desempenho do sistema e do cumprimento dos horários é feita por linha. Por essa razão, neste trabalho optou-se por utilizar os dados referentes a apenas uma linha, mas sem comprometer a generalidade do problema, ou seja, de forma a que as técnicas utilizadas possam ser aplicadas qualquer que seja a linha escolhida. Uma *linha* é um conjunto fixo de ligações e paragens na rede que é servida por um conjunto de veículos de acordo com um horário pré-estabelecido (Liu and Sinha 2007). A linha escolhida para este trabalho foi a 205, que faz a ligação entre a estação da Campanhã e o Castelo do Queijo. Esta linha foi escolhida por ser a linha da STCP que transporta maior número de passageiros por dia (cerca de 15000 nos dias úteis), possuindo um percurso bastante longo e um elevado número de viagens por dia. O percurso efectuado pelos autocarros que circulam nesta linha tem aproximadamente 18km e está

representado no mapa da Figura 1. Esta linha possui 6 pontos de horário, que se podem observar na Figura 1 assinalados a cor amarela: Campanhã, São Roque, Hospital de São João, Amial, Norteshopping (Rotunda AEP) e Castelo do Queijo.

Figura 1: Mapa da linha 205

Tal como outras empresas de transportes públicos, a STCP efectua o controlo das viagens no momento em que estas se realizam, através de uma sala de controlo instalada na sede da empresa. Nesta sala trabalham vários operadores que controlam a posição dos autocarros a cada momento, tendo a possibilidade de comunicar com os motoristas via rádio ou até mesmo enviar mensagens de texto directamente para o autocarro, para que estas possam ser visualizadas pelos passageiros. Estes operadores podem assim dar indicações aos motoristas, como, por exemplo, abrandar ou acelerar a marcha, para que o horário previsto seja cumprido tanto quanto possível. No entanto, uma política de controlo no momento que não é adoptada nesta empresa é a de fazer compassos de espera nos pontos de horário. Ou seja, se um autocarro está numa paragem que é ponto de horário e os passageiros já se encontram todos dentro do autocarro, este abandona a paragem independentemente de estar a partir antes da hora prevista ou não. Isto faz com que os autocarros circulem muitas vezes adiantados durante grande parte do percurso. No entanto, a situação inversa também acontece, existindo muitas situações em que os autocarros acumulam

sucessivos atrasos, comprometendo até as viagens seguintes. Este tipo de situações pode ser observada na Figura 2[1].

Na Figura 2 estão representadas as viagens previstas e realizadas na linha 205 nas primeiras horas da manhã do dia 10/03/2008. As linhas a tracejado representam o horário previsto, sendo possível observar o local onde deveria estar cada autocarro a cada momento. As linhas coloridas representam os dados reais. Neste gráfico é possível observar todo o tipo de situações: autocarros que circulam adiantados, como o caso da viagem V2 no sentido Ida, entre as 7h e as 8h da manhã; autocarros que circulam atrasados, como a viagem V3 entre as 7h e as 8h da manhã no sentido Volta, que acaba por acumular atrasos fazendo com que a viagem seguinte parta atrasada; autocarros que se mantêm aproximadamente dentro do horário previsto, como o caso da viagem V2 no sentido IDA entre as 6h e as 7h da manhã. Como se pode observar, a análise deste tipo de gráficos é um processo moroso que não permite tirar automaticamente conclusões acerca das situações em que os desvios acontecem.

Tal como já foi referido anteriormente, neste trabalho pretendia-se fazer o estudo de ferramentas que permitissem detectar as situações em que ocorrem desvios sistemáticos do horário previsto. Para isso, foi feita uma análise da aderência ao horário nas paragens que são pontos de horário, classificando o desvio relativamente ao horário previsto e procurando depois encontrar padrões de comportamento que induzissem à ocorrência desses desvios, utilizando para isso técnicas de mineração de dados. A classificação dos desvios foi feita consoante a diferença (em minutos) entre a hora prevista de passagem na paragem e a hora a que o autocarro de facto chegou à paragem, sendo explicada com mais detalhe na secção 3.2.3 deste trabalho.

[1] A Figura 2 foi obtida a partir do Sistema de Apoio à Exploração de Informação (SAEI) da STCP.

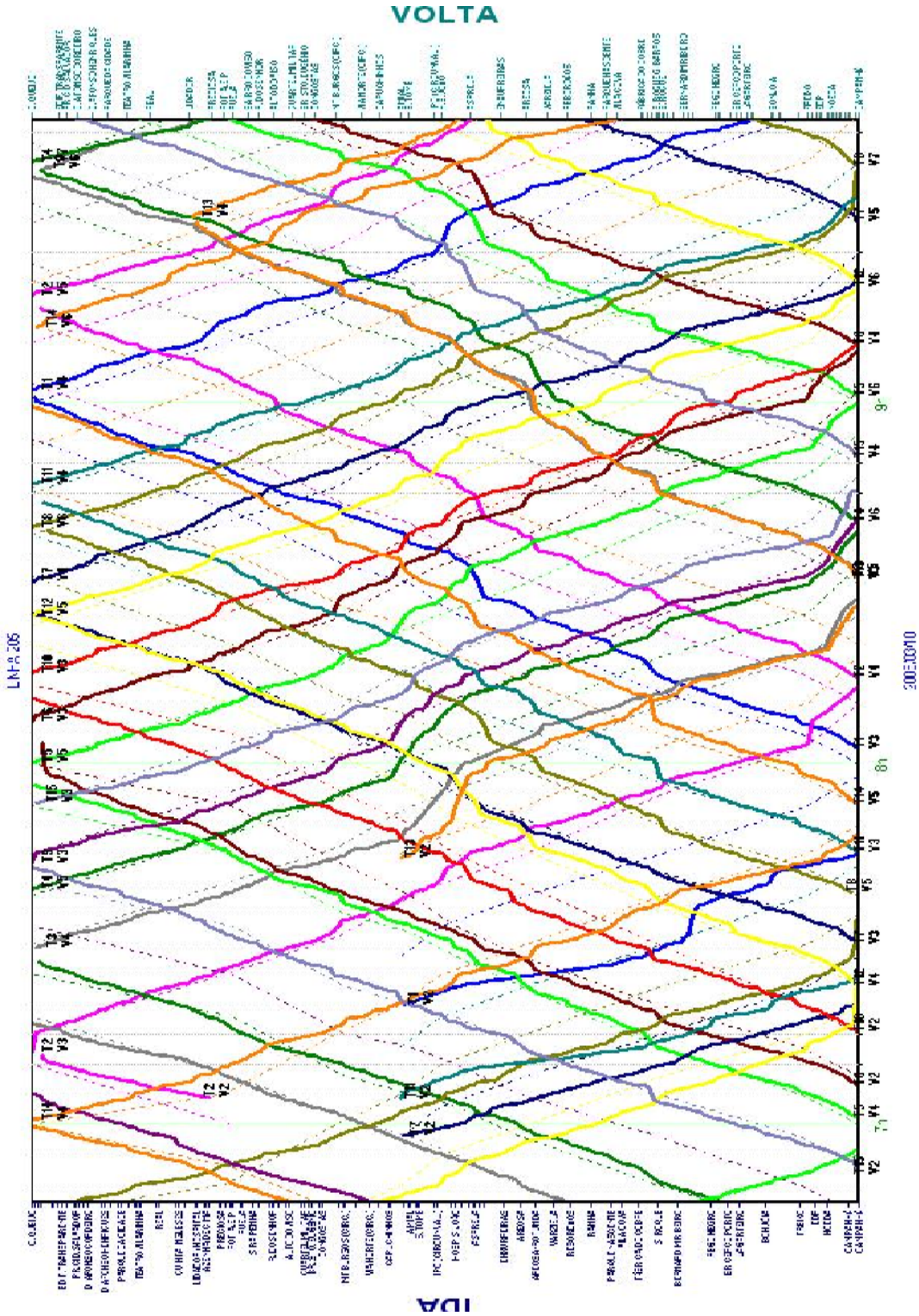

Figura 2: Diagrama espaço-tempo das viagens na linha nas primeiras horas da manhã do dia 10/03/2008

2.2 O processo de extracção de conhecimento e a metodologia utilizada

A utilização de técnicas de mineração de dados para extracção de conhecimento a partir de bases de dados é considerada por Fayyad *et* al. (1996) como sendo apenas uma fase de um processo maior designado por *Descoberta de Conhecimento em Bases de Dados* (KDD – Knowledge Discovery in Databases). A mineração de dados consiste na aplicação de algoritmos específicos para a extracção de padrões nos dados. No entanto, a execução de tarefas como selecção, preparação e limpeza dos dados, incorporando conhecimento prévio apropriado e a avaliação adequada dos resultados, são essenciais para se garantir que informação útil possa ser extraída a partir dos dados (Fayyad, Piatetsky-Shapiro et al. 1996; Fayyad and Uthurusamy 1996). Estas actividades consomem muitas vezes mais tempo do que a mineração de dados propriamente dita, tendo uma grande influência no resultado final do processo (Feelders, Daniels et al. 2000).

Apercebendo-se de que existia a necessidade de uma metodologia que orientasse todo o processo de KDD, as empresas DaimlerChrysler[2], SPSS[3], e NCR[4] desenvolveram em conjunto a CRISP-DM[5] (CRoss-Industry Standard Process for Data Mining) (Chapman, Clinton et al. 2000). A CRISP-DM é uma metodologia não proprietária e disponível gratuitamente que, dado o seu carácter generalista, pode ser aplicada na maioria dos projectos de KDD. Por essa razão, esta será a metodologia utilizada neste trabalho. No entanto, tal como observado pelos próprios autores, "(…) *a descrição das fases e tarefas como passos seguidos numa ordem específica é uma sequência idealizada de eventos. Na prática, muitas das tarefas podem ser realizadas numa ordem diferente e pode ser necessário voltar aos passos anteriores e repetir determinadas acções."* (Chapman, Clinton et al. 2000). Assim, poderá haver situações em que a metodologia

[2] http://www.daimler.com/

[3] http://www.spss.com/

[4] http://www.ncr.com/

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

não será seguida rigorosamente, mas será sempre a base para a realização do processo de KDD neste trabalho.

Na metodologia CRISP-DM, o processo de KDD é dividido em 6 fases principais: compreensão do negócio, compreensão dos dados, preparação dos dados, modelação, avaliação e desenvolvimento.

Figura 3: Metodologia CRISP-DM (baseado em (Chapman, Clinton et al. 2000))

A Figura 3 apresenta o processo de KDD segundo a metodologia CRISP-DM, através da representação das diferentes fases do processo e as ligações entre elas:

- Compreensão do negócio (*Business Understandig*): compreensão dos objectivos do projecto numa perspectiva de negócio, convertendo depois esse conhecimento para um problema de mineração de dados e estabelecendo as linhas gerais que vão ser seguidas para a execução do projecto.
- Compreensão dos dados (*Data Understanding*): consiste na recolha e exploração dos dados, no sentido de identificar as suas principais características e problemas de qualidade.
- Preparação dos dados (*Data Preparation*): esta fase cobre todas as actividades que constituem a construção do conjunto de dados final, tal como a integração de dados provenientes de diferentes tabelas, derivação de novos atributos e limpeza dos dados.
- Modelação (*Modeling*): consiste na selecção e aplicação de várias técnicas de mineração de dados, tentando encontrar os parâmetros para cada um dos modelos que melhor se adequam ao problema e aos dados em causa.

- Avaliação (*Evaluation*): consiste na avaliação dos modelos desenvolvidos, tentando perceber se estes realmente permitem atingir todos os objectivos de negócio.
- Desenvolvimento (*Deployment*): depois de definidos os modelos é necessário que a informação fornecida pelos mesmos seja organizada e apresentada de forma a que o cliente possa utilizá-la. Dependendo da natureza do projecto e dos requisitos de negócio, a fase de desenvolvimento pode passar simplesmente pela geração de um relatório ou ser um processo mais complexo como a implementação de um processo de mineração de dados repetitivo.

A execução de cada uma destas fases é descrita ao longo deste documento. A fase de compreensão do negócio está implícita neste capítulo, através da descrição dos principais objectivos de negócio e descrição das técnicas e metodologias que irão ser utilizadas. A compreensão e preparação dos dados é descrita no Capítulo 3. A fase de modelação é descrita no Capítulo 4 e a fase de avaliação é descrita no Capítulo 5, integrada nas conclusões do trabalho. A fase de desenvolvimento é a única que não é incluída neste trabalho e que se pretende que venha a ser realizada como trabalho futuro.

Como referido, a fase de mineração de dados propriamente dita é o passo mais importante num processo de KDD. Os principais objectivos da mineração de dados são a *previsão* e a *descrição*. A previsão envolve a utilização de atributos da base de dados para prever os valores de outros atributos e a descrição envolve a descoberta de padrões nos dados (Fayyad and Uthurusamy 1996). No entanto, estes objectivos não são completamente distintos porque pode haver situações em que estes se interligam, como por exemplo na utilização de modelos de previsão para fazer descrições dos dados devido à sua boa interpretabilidade.

Para atingir estes objectivos podem ser utilizadas diversos métodos tais como *classificação*, *regressão, clustering, associação*, entre outras. Neste trabalho foram utilizados métodos de classificação e associação. A classificação é uma técnica que procura encontrar propriedades comuns entre as instâncias que constituem o conjunto de dados, classificando-as em diferentes classes, de acordo com o modelo de classificação utilizado (Chen, Han et al. 1996); A associação também chamada de *agrupamento de afinidades*) é uma técnica de mineração de dados que procura encontrar afinidades entre diversos objectos através da análise dos objectos que frequentemente aparecem juntos e criando assim regras de associação entre eles (Michael and Gordon 2004). Estes métodos podem ser utilizados recorrendo a diversas técnicas de mineração de dados, consoante os dados que se estão a analisar e os objectivos que se pretendem atingir. As técnicas que irão ser utilizadas e os motivos para a sua utilização são descritos na secção 2.3.

Para a fase de construção dos modelos foram utilizadas duas aplicações informáticas: *Weka*[6] e *CBA 1.0*[7]. *Weka* (Witten and Frank 2005) é uma colecção de algoritmos de aprendizagem de máquina para tarefas de mineração de dados, implementados em linguagem *Java*. Possui ferramentas de pré-processamento dos dados, classificação, regressão e visualização dos dados. *CBA* (Bing Liu, Wynne Hsu et al. 1998) é uma ferramenta de mineração de dados desenvolvida na Escola de Computação da Universidade Nacional de Singapura que permite executar tarefas de classificação e previsão, mineração de regras de associação, entre outras. A ferramenta CBA é disponibilizada sem limitações na versão comercial ou solicitando a versão completa para efeitos académicos. O contacto com os autores no sentido de utilizar a versão completa foi efectuado, mas não houve qualquer resposta, razão pela qual foi utilizada a versão *demo* que possui algumas limitações. Para os efeitos deste trabalho, as limitações fundamentais são o facto da versão utilizada limitar o tamanho dos conjuntos de teste a 50000 registos e não possibilitar efectuar testes utilizando a técnica de *10-fold-cross validation*.

2.3 Técnicas de Mineração de Dados Aplicadas

2.3.1 Classificação - Árvores de Decisão

As árvores de decisão são uma ferramenta muito popular utilizada para tarefas de classificação e previsão. Na opinião de muitos investigadores, a popularidade desta técnica deve-se sobretudo à sua simplicidade e transparência (Maimon 2007). As árvores de decisão podem ser utilizadas para classificar um objecto como membro de uma determinada classe ou prever um valor numérico, baseado nos atributos que o caracterizam (classificação).

Uma árvore de decisão é constituída por vários nodos, ligados por ramos. Um nodo com ramos descendentes é chamado de *nodo interno* ou *nodo de teste,* enquanto que os nodos sem ramos descendentes são designados por *folhas*. Cada nodo interno divide o espaço das instâncias

[6] http://www.cs.waikato.ac.nz/ml/weka/

[7] http://www.comp.nus.edu.sg/~dm2/

em dois ou mais sub-espaços, de acordo com um teste em função dos valores dos atributos. O resultado do teste efectuado num nodo interno define qual será o próximo nodo a ser visitado. Quando um nodo não tem mais nós descendentes, significa que chegamos a uma folha. Uma folha pode estar associada a uma classe ou valor numérico, consoante estamos a lidar com variáveis objectivo nominais ou numéricas.

A classificação das instâncias é feita percorrendo a árvore da raiz até às folhas, de acordo com os resultados dos testes efectuados em cada nodo.

As árvores de decisão construídas a partir da aprendizagem de um determinado conjunto de dados são muitas vezes complexas e opacas, o que dificulta muito a sua interpretação por parte dos especialistas (a complexidade de uma árvore de decisão é normalmente medida como o número de nodos da árvore (LA Breslow 1997)). Por essa razão, é muitas vezes necessário aplicar métodos de simplificação da árvore. O ideal é que esses métodos sejam capazes de simplificar a árvore, aumentando assim a sua capacidade de interpretação, mas tentando ao mesmo tempo manter ou até aumentar a sua precisão. Existem vários estudos que comprovam o sucesso de diversos métodos de simplificação de árvores de decisão, demonstrando que é possível aumentar a eficácia do modelo em termos de previsão recorrendo a estas técnicas (LA Breslow 1997). Todavia, o método utilizado e a "quantidade" de simplificação a ser aplicada continua a ser objecto de debate, uma vez que a aplicação de técnicas de simplificação pode também diminuir a eficácia do modelo. Schaffer (Schaffer 1993) defende que o sucesso de muitos dos métodos de simplificação já demonstrados se deve não só ao método aplicado mas também ao conjunto de dados ao qual foi aplicado. Isto significa que as estratégias de simplificação são tanto mais úteis quanto mais apropriadas forem para o problema que se está a tratar. É por isso importante que seja conhecida informação adicional acerca do domínio e contexto do problema (não apenas relativamente ao conjunto de dados) para que possam ser estimados os níveis de ruído e complexidade do problema, por forma a escolher a técnica e o grau de simplificação que vai ser aplicado. Hoje em dia, o leque de técnicas de simplificação de árvores de decisão é bastante variado (LA Breslow 1997; Murthy 1998): técnicas de poda da árvore, que consistem na remoção de algumas sub-árvores "desnecessárias"; restrições ao número mínimo de nodos por folha; restrição ao nível dos dados, nomeadamente a redução do número de atributos que descrevem os objectos; transformação das árvores em estruturas alternativas, nomeadamente num conjunto de regras de produção.

Um dos métodos mais utilizados na simplificação de árvores de decisão é a técnica de *poda da árvore* (*pruning*). Basicamente, os algoritmos de *pruning* recebem uma árvore T e dão como resultado uma árvore T', que resulta de remoção de uma ou mais árvores de T , substituindo essas sub-árvores por folhas (LA Breslow 1997). Existem diversos algoritmos de *pruning*, nomeadamente *Minimal Cost-Complexity Pruning* (MCCP) (Breiman, Friedman et al. 1984),

Reduced Error Pruning (REP) (Quinlan 1987), *Minimum Error Pruning* (MEP) (Niblett and Bratko 1987), *Critical Value Pruning* (CVP) , *Pessimistic Error Pruning* (PEP) (Quinlan 1987), *Error Based Pruning* (EBP) (Quinlan 1993), entre outros. Vários estudos foram já efectuados no sentido de avaliar as potencialidades de cada um destes métodos e efectuar comparações entre eles. Em (LA Breslow 1997), Breslow *et al.* fazem um sumário das conclusões dos estudos de Mingers (Mingers 1989) e Esposito (Floriana Esposito 1997). O método REP é destacado pelo seu bom desempenho pois produz árvores muito próximas do tamanho "correcto" e a sua precisão é boa. Eles destacam ainda a maior eficácia da previsão dos métodos PEP e EBP relativamente aos outros métodos, deixando contudo a salvaguarda de que ambas as técnicas têm tendência para construir árvores de maiores dimensões. No entanto, em (Lawrence O. Hall 2002) , Lawrence O.Hall *et al.* afirmam que estes resultados para o método EBP são apenas válidos quando é utilizado o parâmetro por defeito para o factor de confiança, ou seja, CF = 25. No algoritmo EBP, o factor de confiança (CF) é o parâmetro que controla a técnica de *pruning*: se CF = 100 significa que não irá haver poda da árvore; quanto mais baixo for o valor de CF maior será a poda (Lawrence O. Hall 2002). Isto significa que se reduzirmos o valor de CF poderemos obter árvores de menores dimensões.

As árvores de decisão possuem determinadas características que as distinguem de outros métodos de classificação e previsão e que fazem com que estas sejam uma ferramenta tão popular para realizar este tipo de tarefas (Landgrebe 1991; Murthy 1998; Maimon 2007):

- A estrutura hierárquica das árvores de decisão possibilita um melhor uso dos atributos que caracterizam a instância e um aumento da eficiência em termos computacionais;
- As árvores de decisão permitem classificar instâncias através de uma sequência de testes simples cuja semântica é intuitiva para os utilizadores; não é necessário ser um conhecedor de técnicas de mineração de dados para compreender e utilizar uma árvore de decisão;
- As árvores de decisão podem ser representadas graficamente como estruturas hierárquicas, o que torna mais fácil a sua interpretação;
- As árvores de decisão podem ser facilmente transformadas em regras de produção.

Estas características vão de encontro aos objectivos do modelo que se pretendia obter neste trabalho, razão pela qual esta técnica foi escolhida para integrar o conjunto de técnicas testadas.

Existem muitos algoritmos que podem ser utilizados para gerar árvores de decisão, No entanto, o algoritmo C4.5, proposto por Quinlan em 1993 (Quinlan 1993), é um dos algoritmos de referência nesta área, razão pela qual este será o algoritmo utilizado para a construção de árvores de decisão neste trabalho.

2.3.2 Associação – Descoberta de CARs

A descoberta de regras de associação é uma técnica de mineração de dados que identifica relações interessantes entre as variáveis em grandes bases de dados. Seja $I = \{i_1, i_2, \dots, i_m\}$ um conjunto de atributos (itens), seja D um conjunto de transacções em que cada transacção T é um conjunto de itens tal que $T \subseteq I$. Uma regra de associação é uma implicação da forma $X \Longrightarrow Y$ tal que $X \subset I, Y \subset I \; e \; X \cap Y =$ (Rakesh Agrawal and Srikant 1994).

As investigações nesta área foram impulsionadas essencialmente pela crescente necessidade das grandes empresas de retalho conseguirem obter informação útil a partir dos dados das suas vendas (Agrawal 1993; Rakesh Agrawal and Srikant 1994; Craig Silverstein 2004). Este tipo de análise é conhecido como *market basket analysis* (análise de cestos de compras), em que o objectivo é descobrir padrões de compra, procurando encontrar os produtos que habitualmente são comprados ao mesmo tempo (Craig Silverstein 2004). O exemplo de uma regra de associação é a regra $Fraldas \wedge Copos \Longrightarrow Cerveja$, que indica que num cesto de compras onde existem fraldas e copos existe uma grande probabilidade de existir também cerveja. A análise deste tipo de regras permite que as empresas tomem medidas no sentido de aumentar as suas vendas, nomeadamente na elaboração de campanhas de marketing, na disposição dos artigos nas lojas, na elaboração de catálogos, entre outros.

No entanto, a análise de cestos de compras não é a única área em que podem ser aplicadas técnicas de descoberta de regras de associação. Aplicações como a descoberta de padrões em genes (S.Hanash 2003), a classificação automática de e-mails com base no seu conteúdo (Itskevitch 2001) ou a extracção de informação útil e relevante em grandes conjuntos de textos (Hany Mahgoub 2007), são apenas alguns exemplos. Neste trabalho, a descoberta de regras de associação é feita no sentido de descobrir regras que associem o comportamento dos autocarros com as condições em que esses comportamentos ocorrem.

No entanto, a mineração de regras de associação na sua forma "tradicional" é de pouca utilidade no contexto do problema em estudo. Isto porque, aplicando esta técnica, iriam ser obtidas regras cujo lado direito da implicação poderia ser qualquer subconjunto do nosso conjunto de atributos. Ora, aquilo que se pretende é obter regras cujo lado direito da implicação tenha apenas um elemento e que esse elemento seja um tipo de atraso.

Em 1998, Liu *et al.* propuseram um novo tipo de regras de associação: *regras de associação para classificação* (CARs – do inglês *class association rules*). Estas são regras de associação cujo lado direito da implicação é restrito ao atributo de classificação. Seja D um conjunto de dados, I o conjunto de todos os atributos em D e Y o conjunto de identificadores da classe. Uma regra de associação para classificação (CAR) é uma implicação na forma $X \Longrightarrow y$, em que $X \subseteq I$ e $y \in Y$. Uma regra $X \Rightarrow Y$ é satisfeita no conjunto D com confiança c se c% dos casos em D que contêm X estão etiquetados com a classe y. A regra $X \Rightarrow Y$ tem suporte s em D se s% dos casos em D que

contêm X estão etiquetados com a classe y. Para possibilitar a descoberta de CARs, Liu *et al.* (Bing Liu, Wynne Hsu et al. 1998) adaptaram o algoritmo de descoberta de regras de associação *Apriori*, proposto em 1994 por Srikant *et al* (Rakesh Agrawal and Srikant 1994). O objectivo é gerar o conjunto completo de CARs que satisfazem os valores de mínimo suporte e confiança mínima propostos pelo utilizador. Esta adaptação dá ainda a possibilidade de efectuar *pruning* das regras geradas através do método *Pessimistic Error Pruning* utilizado também na geração de regras do algoritmo C4.5 (Quinlan 1993).

Uma vez que o algoritmo pode gerar um elevado número de regras, é importante definir critérios para avaliar a qualidade e interesse das regras geradas. Os critérios utilizados podem ser muito variados (Hussain, Liu et al. 2000): confiança, suporte, senso comum, aplicabilidade, novidade, entre outros. Neste trabalho, optou-se por medir a qualidade e interesse das regras geradas tendo em conta sobretudo os valores de confiança.

Os motivos para a inclusão desta técnica nas experiências deste trabalho prendem-se com a sua flexibilidade em obter regras com a qualidade desejada pelo utilizador. Isto porque o utilizador pode escolher quaisquer valores de suporte e confiança, permitindo assim à partida a selecção das regras de maior interesse.

2.3.3 Classificação Baseada em Associação

A descoberta de regras de classificação e a descoberta de regras de associação são duas técnicas de mineração de dados muito utilizadas. A primeira procura descobrir um pequeno conjunto de regras na base de dados para formar um classificador preciso (Quinlan 1986) enquanto a segunda procura todas as regras na base de dados que satisfazem as condições de mínimo suporte e confiança (Rakesh Agrawal and Srikant 1994). Em 1998, *Liu et al.* (Bing Liu, Wynne Hsu et al. 1998) propuseram uma nova técnica denominada de *classificação associativa* que tem como objectivo integrar associação e classificação para construir classificadores mais precisos. Eles propuseram o algoritmo CBA (*Classification Based in Associations*) para implementar essa técnica e demonstraram que este consegue muitas vezes produzir classificadores mais precisos do que aqueles construídos a partir do algoritmo C4.5. Isto porque o conjunto de regras de associação que satisfazem um mínimo suporte e confiança constituem o conjunto de todas as regras que contêm informação importante, o que faz com que esta técnica tenha um grande potencial para reflectir a verdadeira estrutura dos dados (Wang, Zhou et al. 2000).

O algoritmo CBA utiliza por isso uma técnica denominada de *classificação associativa* e é composto por duas partes: a descoberta de regras de associação para classificação (CARs) e a construção de um classificador a partir dessas regras. A primeira parte desse algoritmo é descrita

na secção 2.3.2 e será utilizada também neste trabalho de forma independente. A segunda parte do algoritmo consiste na construção de um classificador usando as CARs obtidas na primeira fase do algoritmo.

A construção do classificador baseia-se no conceito de precedência entre duas regras. Dadas duas regras r_i e r_j, r_i precede r_j ($r_i \succ r_j$) se

1. a confiança de r_ié maior do que a confiança de r_j, ou
2. a confiança de r_ié igual à confiança de r_j mas o suporte de r_ié maior do que o suporte de r_j, ou
3. ambas as regras têm os mesmo valores de confiança e suporte mas r_i é gerada antes de r_j.

Esta fase do algoritmo consiste em escolher um conjunto de regras de elevada precedência que classifique todas as instâncias do conjunto de treino. O classificador obtido é constituído por uma lista ordenada de regras e uma *classe por defeito*. Na classificação de uma determinada instância, percorre-se o classificador pela ordem gerada pelo algoritmo em busca de uma regra que satisfaça a instância: a primeira regra encontrada classifica a instância; caso não seja encontrada nenhuma regra que satisfaça a instância, esta é classificada pela classe por defeito. O classificador produzido pelo algoritmo CBA satisfaz uma condição que garante a qualidade das regras incluídas: *cada instância de treino é coberta pela regra com maior precedência relativamente a todas as regras que podem cobrir essa instância* (Bing Liu, Wynne Hsu et al. 1998).

Uma vez que a primeira fase do algoritmo CBA consiste na descoberta de CARs, é necessário definir os valores de mínimo suporte (*minsup*) *e* confiança (*minconf*) que o algoritmo irá utilizar. A escolha de *minsup* é a mais complexa porque este tem um grande efeito na qualidade do classificador produzido: se o valor for muito elevado, as regras que não satisfazem *minsup* mas que poderiam ter elevada confiança não são incluídas no classificador; por outro lado, regras com um suporte muito baixo significa que são pouco comuns, podendo por isso representar *outliers* (Bing Liu, Wynne Hsu et al. 1998; McGarry 2005).

Esta técnica foi escolhida para este trabalho porque a integração das técnicas de associação e classificação faz com que o classificador seja constituído por um conjunto de regras de elevada qualidade que constitui uma descrição global do conjunto de dados.

Para além do algoritmo CBA, existem agora outros algoritmos que constroem classificadores associativos, como o caso do CMAR (Han and Pei 2001), CPAR (Yin and Han 2003), CorClass (Zimmermann and De Raedt 2004) e GARC (Chen, Liu et al. 2006). Existem também alguns trabalhos que identificaram certas “fraquezas” do algoritmo CBA e propuseram novas versões deste algoritmo com alguns melhoramentos (Janssens, Wets et al. 2003, Liu, 2000 #101). Apesar

disto, neste trabalho iremos utilizar o algoritmo CBA na sua versão original, por ser o algoritmo de referência nesta área.

Capítulo 3

Pré-Processamento dos Dados

3.1 Compreensão dos Dados

3.1.1 Recolha inicial dos Dados

Para a realização deste trabalho foi utilizada informação proveniente do *data warehouse* (DW) da STCP e ainda informação proveniente do Departamento de Operações, correspondente aos horários previstos.

A informação seleccionada do DW deu origem a duas tabelas:

- *Viagens,* que contém informação acerca das viagens realizadas, em que cada registo corresponde a uma viagem. Nesta tabela podemos encontrar informação como hora de início e hora de fim (prevista e real), motorista, serviço, kilómetros percorridos, entre outros;
- *DW_REGESTATISTICA,* que contém todos os eventos de detecções de paragem com os respectivos dados estatísticos associados, tais como carga, hora, tipo de dia, entre outros.

Estes dados correspondem a viagens realizadas entre 1 de Janeiro de 2007 e 31 de Outubro de 2007.

O motor de bases de dados utilizado para armazenar estas tabelas e efectuar tarefas de pré-processamento dos dados foi o *SQL Server 2005.*

3.1.2 Descrição dos Dados

As secções que se seguem descrevem as tabelas VIAGENS e DW_REGESTATISTICA, enumerando cada um dos atributos, a respectiva descrição e tipos de dados.

A) TABELA VIAGENS

Esta tabela contém informação sobre inícios e fins de viagens entre 1 de Janeiro de 2007 e 31 de Outubro de 2007.

Atributo	Descrição	Tipo de Dados
LINHA	Identificador da linha	nvarchar(20)
DIA_FV	o dia em que a viagem ocorreu	datetime
DTHR_VIAGEM	data e hora da viagem	datetime
TURNO	o turno do da viagem	nvarchar(40)
SERVIÇO	Identificador do Serviço	navarchar(40)
NR_VIAGEM	Número da Viagem	int
SENTIDO	Ida ou volta ('I' ou 'V')	nvarchar(1)
P_DTHR_VGMINI	data e hora de início da viagem previstas	datetime
P_DTHR_VGMFIM	data e hora de fim da viagem previstas	datetime
R_DTHR_VGMINI	data e hora de início da viagem reais	datetime
R_DTHR_VGMFIM	data e hora de início da viagem reais	datetime
P_COD_PRG_INI	paragem de início prevista	nvarchar(6)
P_COD_PRG_FIM	paragem de fim prevista	nvarchar(6)
R_KMS	quilómetros percorridos durante a viagem	nvarchar(10)
P_KMS	quilómetros previstos para serem percorridos durante a viagem	nvarchar(20)
R_TIPOVGM	tipo da viagem que pode ser completa, incompleta ou com erros	nvarchar(Max)

Tabela 1: Atributos da tabela Viagens

B) TABELA DW_REGESTATISTICA

Esta tabela contém informação sobre todos os eventos de detecções de paragem com os respectivos dados estatísticos associados, para viagens entre 1 de Janeiro de 2007 e 31 de Outubro de 2007.

Atributo	Descrição	Tipo de Dados
NR_VEICULO	Identificador do Veículo	Numeric(4,0)
TIMESTAMP	hora de entrada na informação no SAEI	Datetime
DATA_TRAMA	hora de passagem na paragem detectada pelo GPS	Datetime
NR_MAT	código do motorista	Numeric(5,0)
SERVICO_KM	Número de kilómetros percorridos durante o serviço	Numeric(6,1)
NR_SERVICO	Número do Serviço	Numeric(4,0)
EPOCA	Época do Ano	Numeric(2,0)
DIA_TIPO	tipo de dia a que se refere (feriado, etc…)	Numeric(2,0)
NR_LINHA	Identificador da linha	Nvarchar(4)
NR_TURNO	Número do turno	Numeric(2,0)
NR_VIAGEM	Número da Viagem	Numeric(2,0)
TIPO_VIAGEM	Tipo de Viagem	Numeric(1,0)
NR_ORDEM_PARAGEM	Número de Ordem de Paragem	Numeric(2,0)
NR_VAR_PERC	Número de variáveis do percurso	Numeric(2,0)
SENTIDO	Sentido	Numeric(1,0)
NR_PTPARAGEM	código da paragem	Numeric(4,0)
ID_PARAGEM_STCP	identificador da paragem	Nvarchar(6)
CARGA	carga do veículo no momento da paragem	Numeric(3,0)
NUM_PASSAGEIROS_ENTRADA	Número de Passageiros que entraram na paragem	Numeric(3,0)
NUM_PASSAGEIROS_SAIDA	Número de passageiros que saíram na paragem	Numeric(3,0)
RDATE	Data da viagem	Datetime
TEMPO_PARADO	Tempo que o autocarro esteve parado na paragem	Numeric(5,0)
PARAGEM_PREV	Hora de paragem prevista	datetime

Tabela 2: Atributos da tabela DW_REGESTATISTICA

O atributo DIA_TIPO é representando como um inteiro, em que cada número corresponde a um tipo de dia, segundo o seguinte mapeamento:

- DIA_TIPO = 1 → Dia útil
- DIA_TIPO = 2 → Sábado
- DIA_TIPO = 3 → Domingo/Feriado
- DIA_TIPO = 4 → Especial

- DIA_TIPO = 5 → Ponte
- DIA_TIPO = 6 → Férias Escolares

3.1.3 Exploração e análise da qualidade dos dados

Por serem tabelas obtidas a partir de um DW*, o*s dados que constam nas tabelas descritas anteriormente foram obtidos como resultado de um pré-processamento. Por essa razão, muitos dos erros que a informação não filtrada poderia conter foram já eliminados. No entanto, podiam ainda prevalecer diversos tipos de erros. Uma vez que o objecto de análise deste trabalho eram apenas as viagens referentes à linha 205, a análise limitou-se aos dados que dizem respeito a essa linha.

Através de uma análise aos atributos das tabelas, foi possível verificar de imediato que não existe um atributo que identifique univocamente uma viagem, o que faz com que não exista uma ligação directa entre as viagens que constam na tabela VIAGENS e os eventos de detecção de paragens que se encontram na tabela DW_REGESTATISTICA. Esta limitação teve consequências ao nível do método utilizado para fazer a junção das tabelas. Uma vez que neste trabalho foi utilizada informação proveniente das duas tabelas, foi necessário encontrar condições alternativas que possibilitem fazer essa junção.

Outra observação importante tem a ver com a inexistência da hora de passagem prevista em cada paragem. Apesar de existir um atributo para este efeito na tabela DW_REGESTATISTICA, estes valores não estavam preenchidos em nenhum registo. Isto fez com que não fosse possível calcular directamente o desvio existente entre o horário previsto e a hora real de passagem; para obter o desvio foi necessário derivar primeiro o atributo relativo à hora prevista de passagem, através da junção da informação proveniente das tabelas do *data warehouse* com as tabelas de horários para a linha 205.

As secções seguintes apresentam uma análise mais pormenorizada de alguns atributos que são fundamentais para o tipo de análise que se deseja fazer neste trabalho.

A) HORA DE ÍNICIO E FIM DE VIAGEM

Os atributos que indicam a hora de início e fim da viagem (prevista e real) encontram-se na tabela VIAGENS e são os seguintes: P_DTHR_VGM_INI, P_DTHR_VGM_FIM, R_DTHR_VGM_INI, R_DTHR_VGM_FIM. Através da observação visual dos registos da tabela verificou-se que existem muitos registos em que estes atributos têm valor nulo. Verificou-se ainda que, na generalidade dos casos, quando um dos atributos relativos à hora prevista tem valor nulo, o outro também tem, o mesmo acontecendo com os atributos relativos à hora real.

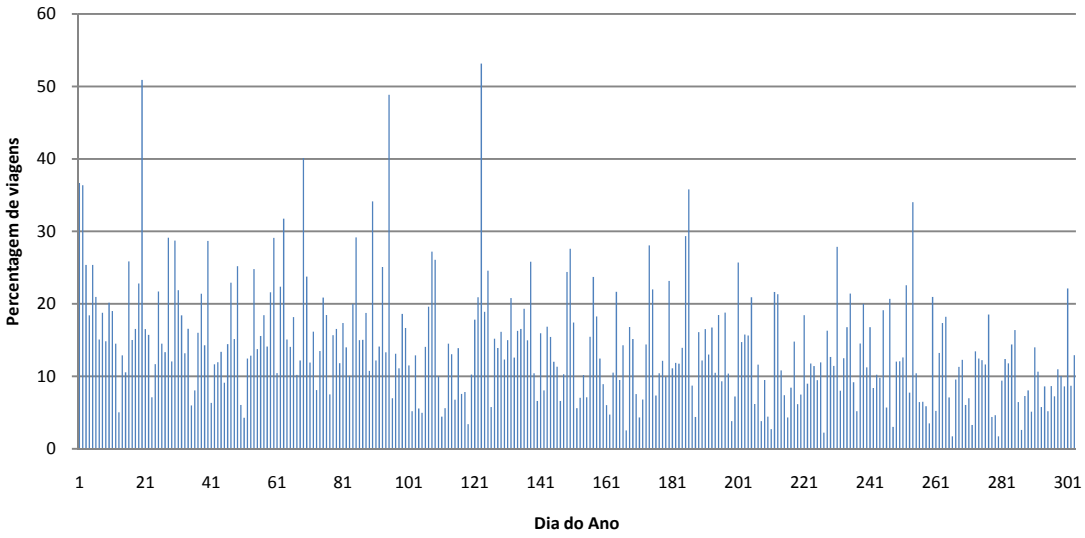

Figura 4: Gráfico relativo à percentagem de viagens com valores nulos nos atributos que indicam a hora de início e fim da viagem, em função do dia do ano

Estatística	Valor
Média	14,23
Desvio Padrão	7,96
Mínimo	1,72
Máximo	53,15

Tabela 3: Tabela de estatísticas amostrais relativas aos dados do gráfico da Figura 4

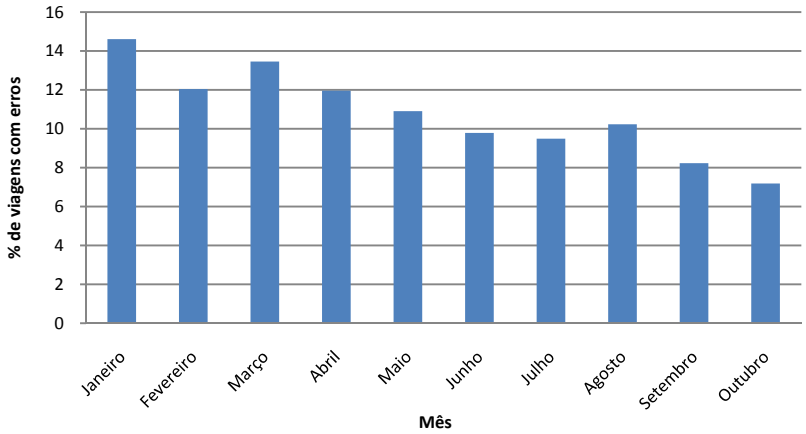

Figura 5: Percentagem de viagens com erros na hora de início e fim da viagem, em cada mês

No gráfico da Figura 4 é possível observar a percentagem de viagens com valores nulos nos atributos que indicam a hora de início e fim da viagem, por dia do ano. Através da análise do gráfico, verificou-se que a percentagem destes erros não é constante ao longo do ano e tem uma distribuição aparentemente aleatória. No entanto, foi possível notar um ligeiro decréscimo ao longo do ano. Este decréscimo é confirmado no gráfico da Figura 5 onde se apresenta a percentagem de viagens com erros em cada mês do ano. Janeiro é o mês que atinge o valor mais alto, 14.606, sendo que os valores vão diminuindo gradualmente até Outubro, atingindo 7.185. Na Tabela 3 apresentam-se ainda os valores da Média, Desvio Padrão, Mínimo e Máximo relativo a estes dados.

A ocorrência deste tipo de erros pode invalidar a análise destas viagens uma vez que os valores presentes nestes atributos são fundamentais para calcular a duração da viagem (prevista e real), calcular atrasos, etc. Uma vez que os eventos de início e fim de viagem reais são introduzidos pelos motoristas, uma possível explicação para este fenómeno são as falhas humanas. Outra possível explicação é o facto de muitas vezes serem realizadas viagens que não estavam planeadas, o que explica a nulidade dos valores previstos para o início e fim de viagem.

B) PARAGEM DE INÍCIO E FIM DA VIAGEM (PREVISTA E REAL)

Os atributos que dizem respeito às paragens de início e fim da viagem (prevista e real) encontrava-se na tabela VIAGENS e são os seguintes: P_COD_PRG_INI, P_COD_PRG_FIM, R_COD_PRG_INI, R_COD_PRG_FIM. Através de uma análise visual desta tabela, verificou-se que existe uma quantidade significativa de registos em que estes atributos surgem com valor nulo. Em particular, observou-se que, na generalidade dos casos em que os valores reais de início e fim de viagem se encontram a nulo, os valores reais para a paragem de início e de fim também se encontram a nulo. Existiam ainda diversas situações em que o valor real de paragem inicial ou final era nulo, o que pode dever-se a falhas na detecção das paragens pelo sistema GPS.

No gráfico da Figura 6 é possível observar a percentagem de viagens com valores nulos nos atributos referentes às paragens de início e fim da viagem, em função do dia do ano. Tal como no caso dos registos relativos ao início e fim de viagem, observou-se que a ocorrência destes erros tem uma distribuição aparentemente aleatória. Comparando este gráfico com o gráfico da Figura 4, verificou-se que os dois são muito semelhantes, o que confirma a ligação entre os valores nulos das paragens e de hora de início e fim de viagem (ver ponto B) desta secção). Na Tabela 4 apresentam-se ainda os valores da Média, Desvio Padrão, Mínimo e Máximo relativo a estes dados.

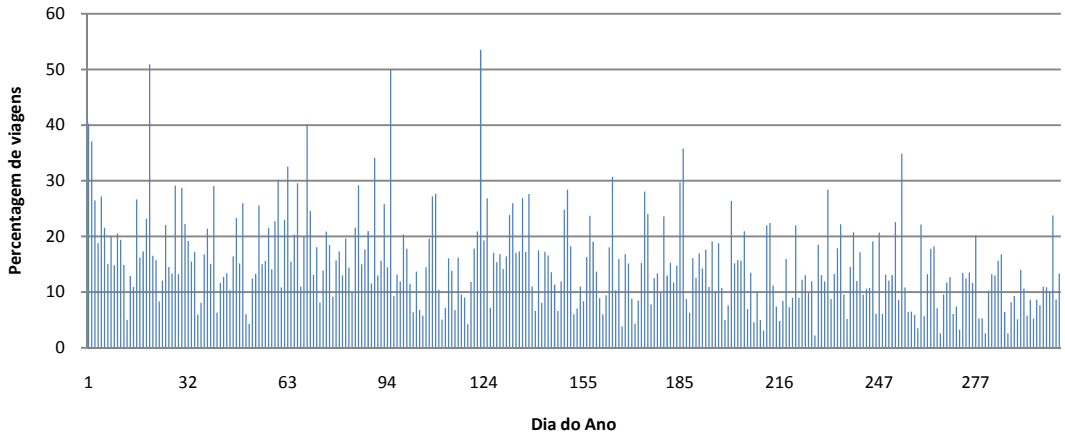

Figura 6: Gráfico relativo à Percentagem de viagens com valores nulos nos atributos referentes às paragens de início e fim da viagem, em função do dia do ano

Estatística	Valor
Média	15,1419
Desvio Padrão	8,151942
Mínimo	2,222222
Máximo	53,54331

Tabela 4: Tabela de estatísticas amostrais relativas aos dados do gráfico da Figura 6

C) PARAGENS INTERMÉDIAS

Os eventos de detecção de paragens iniciais, finais e intermédias de uma viagem estão registados na tabela DW_REGESTATISTICA. Uma vez que neste trabalho foram apenas utilizadas as paragens que são pontos de horário, foi feita uma análise do número de registos que existem relativamente a cada uma dessas paragens. Apesar de algumas das viagens não passarem em todas as paragens referidas, este número é pouco significativo, pelo que se esperava que não existisse uma diferença significativa no número de registos relativos a cada uma das paragens.

Através da análise do gráfico da Figura 7 verificou-se que os resultados iam de encontro aquilo que era esperado. A diminuição observada no número de registos relativos às paragens da Campanhã (CMP) nos meses de verão (7 e 8) deve-se provavelmente ao facto de que, nestes meses, existe uma grande parte das viagens que não passa nestas paragens, ao contrário do que acontece ao longo do resto do ano. Observa-se ainda uma diferença significativa no número de registos nas paragens em Castelo do Queijo (CQ) relativamente às outras paragens, o que não era esperado. No entanto, este fenómeno poderá ser explicado pela existência de obras na zona

ou problemas nos detectores GPS daquelas paragens, o que poderia fazer com que muitas vezes estas não fossem detectadas. Na análise deste gráfico é ainda de realçar o facto de que, apesar de o número de registos para cada paragem não ser constante em todos os meses, a relação entre o número de registos para cada paragem em cada um dos meses mantém-se na maioria dos meses. Por exemplo, em todos os meses a paragem que tem maior número de registos é HSJ, seguida de SR e AML. Esta diferença entre o número de registos por paragem é também evidente no gráfico da Figura 8 onde se apresenta o número de registos por paragem para o ano inteiro.

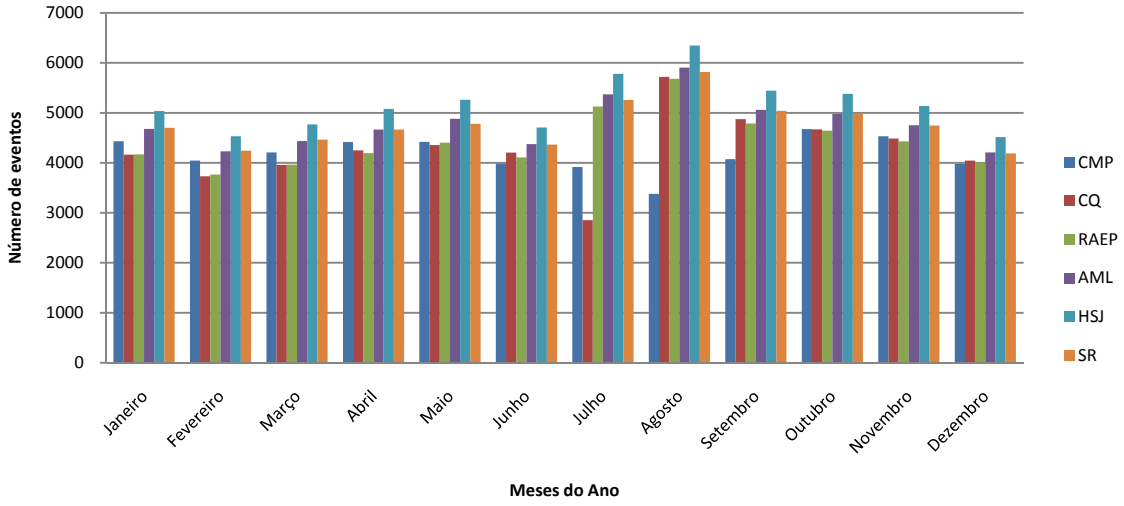

Figura 7: Gráfico com a representação do número de registos por paragem em função do mês

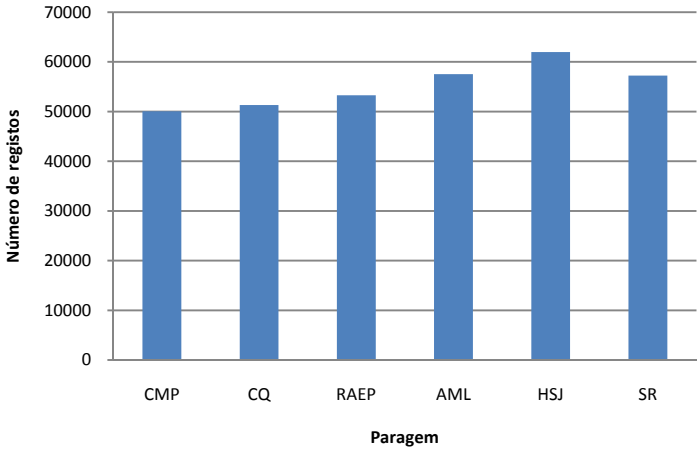

Figura 8: Número de registos por paragem durante o ano de 2007 para a linha 205

D) NÚMERO DE PASSAGEIROS

Apesar de existirem atributos nas tabelas destinados a albergar informação acerca do número de passageiros (NUM_PASSAGEIROS_ENTRADA e NUM_PASSAGEIROS_ SAIDA), estes atributos têm o valor 0 na grande maioria dos registos. Verificou-se que, para o atributo NUM_PASSAGEIROS_ENTRADA, apenas 0,42% dos registos têm valor diferente de 0. Para o atributo NUM_PASSAGEIROS_SAIDA este valor é de 0,49%.

E) CARGA

O atributo CARGA, presente na tabela DW_REGESTATISTICA poderia ser usado como indicador da variação do número de passageiros no autocarro. No entanto, analisando visualmente a tabela verificou-se que existe um número muito elevado de registos onde o atributo Carga tem o valor 0. De facto, esta afirmação é verificada pelo gráfico da Figura 9.

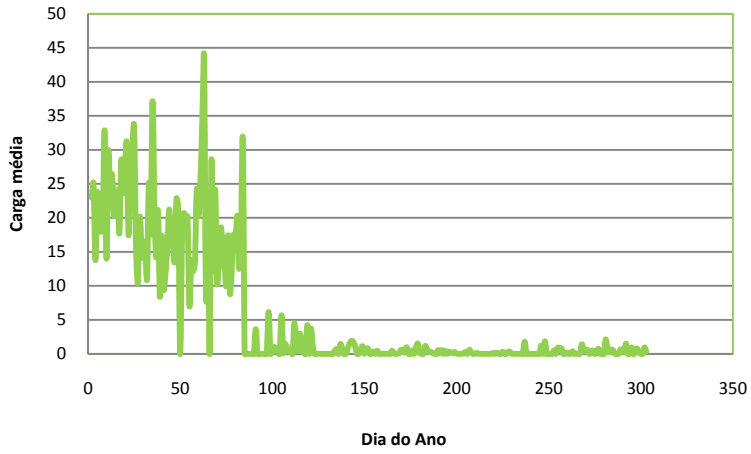

Figura 9: Carga média por dia do ano

Da análise do gráfico da Figura 9 concluiu-se que a média dos valores de Carga por dia do ano não é constante. Nos primeiros 84 dias do ano (até 24 de Março) observam-se valores médios de carga dispersos, que variam entre 0 e 63. No gráfico da Figura 10 é possível observar com mais detalhe a distribuição desses dados. Depois do dia 84, observa-se um período de 36 dias em que os valores médios de carga oscilam entre 0 e 7. A partir do dia 120, os valores médios de carga são praticamente constantes, em valores muito próximos de zero. Observa-se por isso uma grande inconsistência nos valores médios de Carga por dia do ano, tendo-se concluindo assim estes dados podem não ser fiáveis.

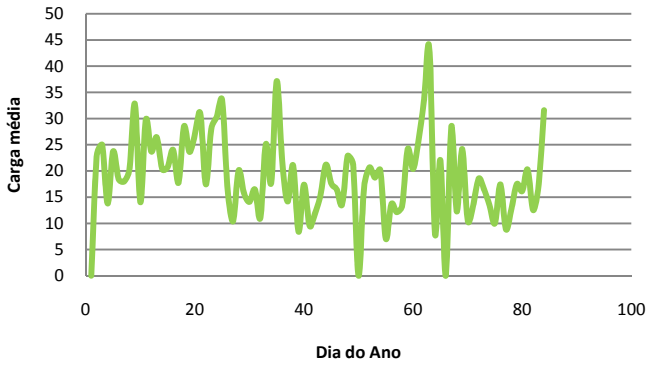

Figura 10: Carga média por dia do ano (até 24 de Março)

F) NÚMERO DE REGISTOS POR DIA

O número de registos por dia esperava-se que fosse aproximadamente constante ao longo de todo o ano, consoante o tipo de dia. A excepção seria feita apenas para os meses de verão (entre 16 de Julho e de Setembro de 2007), uma vez que os horários são alterados e a oferta é diferente do resto do ano. A análise do número de registos por dia foi feita com o objectivo de perceber se existiam diferenças significativas nesses valores ao longo do ano, que podiam indicar a existência de erros.

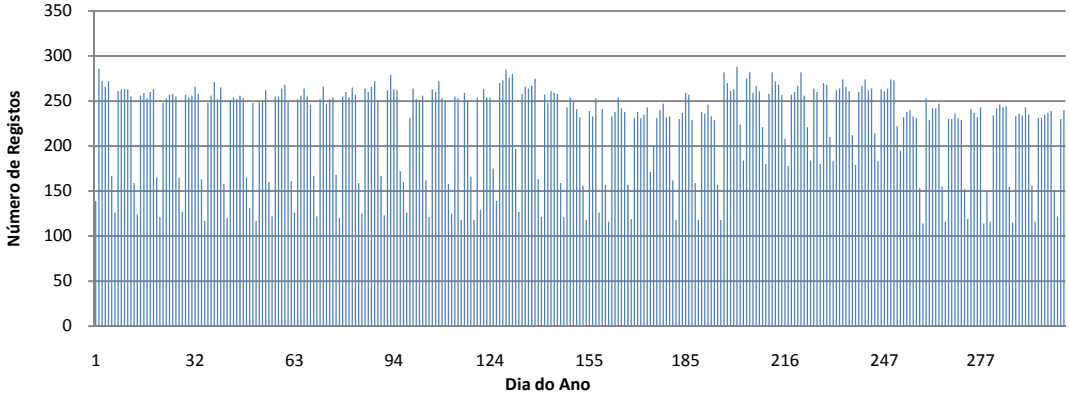

Figura 11: Número de registos por dia do ano na tabela VIAGENS

O gráfico da Figura 11 representa o número de registos por dia do ano na tabela VIAGENS. Observa-se que não existe uma variação significativa do número de registos por dia, relativos aos dias úteis (os valores mais altos do gráfico). Os valores mais baixos que ocorrem com

periodicidade regular correspondem aos fins-de-semana, onde a oferta é bastante inferior à dos dias úteis. Também estes valores não sofrem variações significativas ao longo do ano, com excepção dos meses de Verão, nos quais existe um acréscimo visível. Após a consulta dos horários em vigor no período de análise, verificou-se que existiu um aumento da oferta aos fins-de-semana, pelo que estes valores são normais.

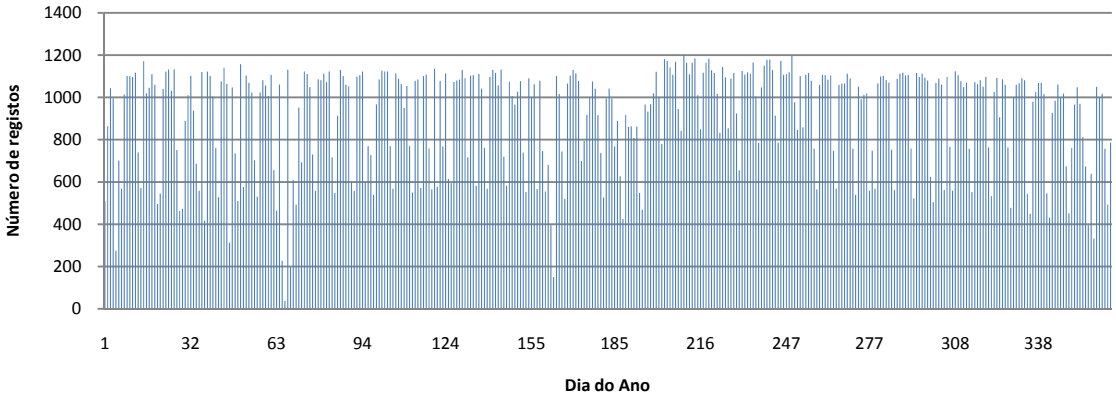

Figura 12: Número de registos por dia do ano das paragens que representam pontos de horário (tabela DW_REGESTATISTICA)

No gráfico da Figura 12 é possível observar que o número de registos na tabela DW_REGESTATISTICA tem um comportamento menos constante por comparação com a tabela VIAGENS. Em particular, existem duas zonas do gráfico, perto dos dias 63 (início de Março) e 155 (início de Junho) em que se observa uma diminuição significativa do número de registos. Apesar disto, uma análise global do gráfico permite confirmar aquilo que se esperava: o número de registos para cada tipo de dia não tem uma variação significativa ao longo do ano, com excepção dos sábados e domingos nos meses de Verão onde a oferta aumentou, tal como já foi referido anteriormente.

G) ÉPOCA DO ANO

Como se pode observar na Tabela 2, existe um atributo na tabela DW_REGESTATISTICA, o atributo EPOCA, que corresponde à época do ano. No entanto, após a análise da tabela verificou-se que o atributo tem o mesmo valor em todos os registos, o que faz com que na prática não exista qualquer distinção nos registos relativamente à época do ano.

3.2 Preparação dos dados

Na fase de preparação dos dados, começou-se por fazer a selecção dos dados que iam ser utilizados, descrevendo os atributos que deveriam constar no conjunto de dados final (3.2.1). De seguida, passou-se à tarefa de integração dos dados, ou seja, junção da informação proveniente das diversas tabelas (3.2.2). Só depois de realizar esta tarefa é que foi possível efectuar a derivação dos novos atributos, pois estes resultam da combinação de atributos provenientes de tabelas diferentes (3.2.3).

3.2.1 Selecção dos dados

Como já foi referido anteriormente, os dados utilizados foram apenas aqueles que se referem à linha 205. Relativamente à tabela DW_REGESTATISTICA, a informação seleccionada foi apenas a que diz respeito às paragens que são pontos de horário nesta linha. Foram ainda excluídos os registos que correspondem aos tipos de dia Especial, Ponte e Férias Escolares. De notar que, neste caso, as férias escolares não se referem ao período de Verão, mas sim aos períodos de férias de Natal, Carnaval e Páscoa (dias úteis). Apesar de nestes dias vigorarem os horários referentes aos dias úteis durante o ano curricular, o facto de não haver aulas para a maioria das crianças ou o facto de existir um grande número de pessoas que não trabalha porque faz ponte, vai produzir alterações no trânsito. Por representarem situações “anormais”, estes dias não devem ser utilizados para avaliar se o horário está ou não a ser cumprido, por isso optou-se por excluir estes dados da análise. Pelas mesmas razões, foi também excluído o período que se refere à semana da queima das fitas de 2007 do Porto, ou seja, o período entre os dias 04 de Maio de 2007 e 12 de Maio de 2007. A junção da informação referente às três tabelas referidas anteriormente deu origem a uma tabela com informação mais completa relativa a cada passagem de um autocarro nos pontos de horário. Os atributos que constam nesta tabela (a qual designaremos por *tabela_1*) foram depois utilizados para derivar outros atributos e construir os conjuntos finais que iam ser utilizados.

A selecção dos atributos a utilizar através de métodos matemáticos não foi um tópico abordado neste estudo. Em vez disso, optou-se por escolher o conjunto de atributos que, na visão dos planeadores de horários da STCP, seria o mais adequado para as análises que se pretendem efectuar. Nesse conjunto foram incluídos atributos que dessem indicações relativas ao número de passageiros, como CARGA, NUM_PASSAGEIROS_ENTRADA e NUM_PASSAGEIROS_SAIDA. No entanto, depois das análises feitas a estes atributos na secção 3.1.3, optou-se por não os incluir, pois a sua informação é pouco fiável e poderia, eventualmente, perturbar os resultados de forma negativa.

Assim, pretendia-se que o conjunto de dados fosse constituído pelos seguintes atributos:

- Identificador da Paragem;
- Sentido;
- Tipo de Dia;
- Dia da Semana;
- Época do Ano (Ano Lectivo ou Férias escolares);
- Dia do Ano;
- Hora do dia (em minutos);
- Tipo de Hora (Hora de Ponta da Manhã, Nocturna, etc);
- Tipo de Desvio (Atrasado, Adiantado ou Pontual);

Nem todos os atributos foram usados em todos os testes. Isto aconteceu devido às especificidades de cada algoritmo, sendo que alguns funcionam melhor (ou só funcionam) com determinados tipos de dados. No entanto, este foi o conjunto de dados base a partir do qual serão construídos os conjuntos de dados específicos a cada tarefa modelação, bastando para isso remover os atributos não utilizados e filtrar os dados relativos ao período de análise desejado. Este conjunto dá-nos a possibilidade de identificar temporal e espacialmente as condições em que os desvios relativamente ao horário previsto acontecem.

3.2.2 Conciliação dos dados

A tarefa de conciliação dos dados consiste na junção de informação proveniente de diversas tabelas. Como já foi referido anteriormente, foi necessário criar uma tabela intermédia, a partir da qual foram derivados novos atributos e construído o conjunto de dados final. Esta tabela resultou da junção das tabelas VIAGENS, DW_REGESTATISTICA e das tabelas referentes aos horários da linha 205. O processo de junção das tabelas não irá ser explicado em detalhe. São apenas apresentados os principais passos que foram seguidos:

1) Definição das condições de junção das tabelas VIAGENS e DW_REGESTATISTICA, tendo em conta que a tabela resultante deverá conter um registo por cada evento de detecção de paragem presente na tabela DW_REGESTATISTICA;
2) Efectuar a junção das tabelas e acrescentar o atributo HORA_NA_PRG_PREVISTA à tabela resultante;
3) Definir as condições de junção das tabelas de horários da linha 205 com a tabela obtida no passo 2), tendo em conta o período em que os horários estiveram em vigor e o tipo de dia a que se referem (Dias Úteis, Sábados ou Domingos);

4) Efectuar a junção das tabelas de horários da linha 205 com a tabela obtida no passo 2), preenchendo assim o campo HORA_NA_PRG_PREVISTA para todas as paragens que são pontos de horário.

Os atributos da tabela resultante são os que constam na Tabela 5. Uma vez que a tabela VIAGENS possui apenas informação até 31 de Outubro de 2007, a tabela resultante apresenta apenas registos até essa data.

Atributo	**Descrição**
DATA	Data da viagem
INICIO PREVISTO	início da viagem previsto
INICIO REAL	início da viagem real
FIM PREVISTO	fim da viagem previsto
FIM REAL	fim da viagem real
PARAGEM INICIO	Paragem de início da viagem
PARAGEM FIM	Paragem de fim da viagem
ID_PARAGEM	Identificador da paragem
HORA_NA_PRG_REAL	Hora a que o veículo chegou à paragem
HORA_NA_PRG_PREVISTA	Hora a que o veículo deveria ter chegado à paragem
LINHA	Linha
SENTIDO	Sentido da viagem
SERVIÇO	Serviço no qual está inserida a viagem
TIPO DE DIA	Tipo de dia em que ocorre a viagem

Tabela 5: Atributos da *tabela_1*.

3.2.3 Construção e Limpeza dos dados

A etapa de construção e limpeza dos dados inclui tarefas de preparação dos dados como a produção de atributos derivados, a transformação de valores de atributos já existentes e a eliminação de registos que possam perturbar a qualidade dos resultados.

A derivação de novos atributos e a transformação dos atributos já existentes foi feita com o auxílio de diversas funções, algumas já pré-definidas no *SQL Server 2005* e outras que foram propositadamente construídas para esse efeito.

A) IDENTIFICADOR DA PARAGEM

Uma vez que em muitas zonas existe mais do que uma paragem, a representação das paragens não é uniforme, ou seja, para as paragens em Campanhã (por exemplo), podem existir representações como CMP1, CMP2, 1CMP, entre outras. Assim, a representação das paragens foi uniformizada para os seguintes valores:

- Campanhã - CMP;
- São Roque – SR;
- Hospital de São João – HSJ;
- Amial – AML;
- Rotunda AEP – RAEP;
- Castelo do Queijo – CQ;

B) SENTIDO E TIPO DE DIA

O atributo TIPO_DE_DIA, apesar de já constar na *tabela_1*, passou a ser representado por uma *String*, segundo o mapeamento definido na secção 3.1.2, para ser de mais fácil interpretação. O mesmo aconteceu com o atributo SENTIDO, que passou também a ser representado por uma *String,* podendo assumir os valores IDA ou VOLTA.

C) ANO LECTIVO

Como foi verificado na fase de exploração dos dados 3.1.3, o valor do atributo EPOCA na tabela DW_REGESTATISTICA é sempre o mesmo. Sendo assim, com o auxílio de uma função especialmente definida para esse efeito, foi derivado o atributo ANO_LECTIVO com base na data do evento. O atributo é do tipo booleano, sendo que o seu valor é *true* caso a data coincida com o ano lectivo e *false* caso a data se refira a um período de férias escolares. Os limites utilizados para definir o ano lectivo e as férias escolares foram os utilizados pela STCP para o funcionamento do horário de Verão (férias escolares entre 16 de Julho de 2007 e 09 de Setembro de 2007, inclusive).

D) HORA E TIPO DE HORA

O atributo TIPO_DE_HORA indica o período do dia a que uma viagem foi realizada, tendo em conta a classificação adoptada pela STCP.

Dias Úteis

Ponta da Manhã	07:30-10:00
Hora Normal da Manhã	10:00-12:00
Meio-dia	12:00-14:00
Hora Normal da Tarde	14:00-17:00

Ponta da Tarde	17:00-20:00
Nocturno	20:00-01:00
Madrugada	01:00-06:00
Sábados, Domingos e Feriados	
Manhã	06:00-14:00
Tarde	14:00-21:00
Nocturno	21:00-01:00
Madrugada	01:00-06:00

Este tipo de classificação foi utilizado para a definição dos horários em cada linha, fazendo variar a oferta de viagens consoante o período do dia. No entanto, é necessário notar que esta é apenas uma classificação genérica que não é utilizada de forma rígida e que pode inclusive variar consoante a linha a que se refere. Existe ainda outra representação para a hora do dia, que corresponde ao atributo HORA que é representado pelos minutos do dia. Este atributo permite-nos saber de forma mais detalhada a hora a que se realizou a viagem.

E) TIPO DE DESVIO

O atributo TIPO_DE_DESVIO foi construído com base na diferença em minutos entre a hora de passagem prevista numa paragem e a hora a que o autocarro de facto passou, consoante a classificação definida na secção 2.1. É por isso importante analisar estatisticamente os valores dessas diferenças, para verificar a existência de erros que devem ser eliminados.

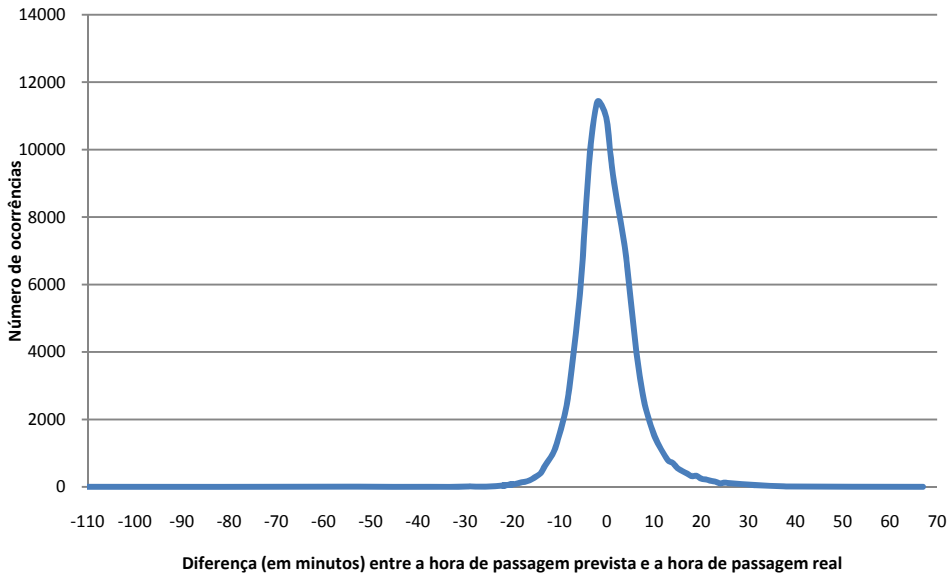

Figura 13: Distribuição dos desvios em relação à hora de passagem na paragem prevista (em minutos)

Estatística	Valor
Média	0,157
Desvio Padrão	6,487
Mínimo	-110
Máximo	67

Tabela 6: Tabela de estatísticas amostrais relativas aos dados do gráfico da Figura 13.

Após a análise do gráfico da Figura 13 e da Tabela 6 verifica-se claramente que existe uma vasta gama de valores que não têm expressão a nível estatístico por terem um número de ocorrências muito reduzido. Para além disso, estes valores podem corresponder a ruído e erros nos dados, pelo que devem ser removidos. Optou-se então por eliminar estes registos, adoptando o critério de *eliminar todos os registos cujo número de ocorrências seja inferior ao percentil 0,01% ou superior ao percentil 99,99%.* Foram por isso eliminados todos os registos com desvio inferior a -17 e superior 23.

É política habitual considerar-se que um autocarro circula dentro do horário, ou seja, Pontual, se chega a um ponto de horário não mais do que 1 minuto adiantado ou 5 minutos atrasado (Strathman and Hopper 1993; Strathman, Dueker et al. 1999). Por isso, existiam à partida 3 tipos de desvio: Pontual, Atrasado e Adiantado. No entanto, esta classificação pode ser insuficiente pois não permite fazer uma distinção entre as situações mais graves e aquelas que merecem menos atenção. Por exemplo, no caso de autocarros que circulam adiantados, merece muita mais atenção uma situação em que estes circulem sistematicamente 7 minutos adiantados do que situações em que o adiantamento é de apenas 2 minutos. Recorreu-se por isso ao auxílio dos valores de percentil associados a cada desvio para definir os intervalos que iam ser utilizados. O gráfico da Figura 14 representa o desvio (em minutos) associado a cada percentil.

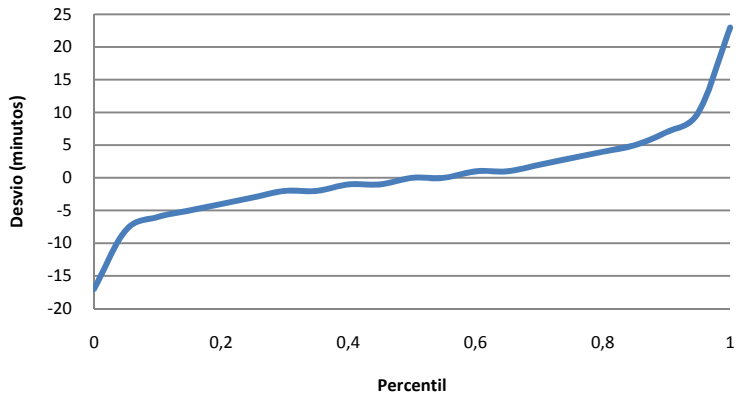

Figura 14: Desvio (em minutos) associado ao Percentil

Após a análise do gráfico e tendo em conta que os limites para o tipo de desvio *Pontual* já estavam definidos e correspondem aos percentis 40% e 85%, optou-se por manter apenas uma classe para os autocarros que chegam atrasados e definir uma classe que corresponde a desvios classificados como *Muito Adiantado*, utilizando o percentil 10%, que corresponde a -6 minutos de desvio. Sendo assim, foram obtidas as seguintes classes:

- Muito Adiantado: se o desvio é menor do que -6;
- Adiantado: se o desvio é maior ou igual a -6 e menor do que -1;
- Pontual: se o desvio é maior ou igual a -1 e menor ou igual a 5;
- Atrasado: se o desvio é maior do que 5;

O gráfico da Figura 15 apresenta o número de instâncias classificadas por cada um dos tipos de desvio. Como se pode observar, a classe que contém um maior número de valores é a classe Pontual, seguida da classe Adiantado.

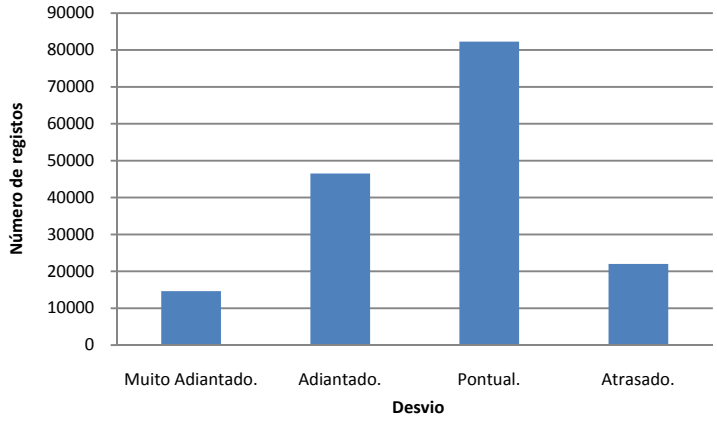

Figura 15: Número de instâncias classificadas por cada um dos tipos de desvio

F) OUTROS ATRIBUTOS

Os atributos DIA_SEMANA e DIA_ANO foram derivados com o auxílio da função *datepart* do *SQL Server 2005* a partir do atributo que representa a data da viagem. Uma vez que a função *datepart* retorna um número inteiro para cada dia da semana, este atributo foi depois convertido em *String* e os números substituídos pelas expressões que identificam cada um dos dias da semana, para ser mais fácil a sua interpretação.

3.2.4 Descrição do conjunto de dados resultante

O conjunto de dados resultante é constituído por 165451 instâncias. Os atributos que o compõem e os valores que podem assumir estão apresentados na Tabela 7:

Identificador da Paragem	**Valores Possíveis**
Sentido	IDA ou VOLTA
Tipo de Dia	SABADO, DIA ÚTIL, DOMINGO/FERIADO
Dia da Semana	SEGUNDA-FEIRA, TERÇA-FEIRA, QUARTA-FEIRA, QUINTA-FEIRA, SEXTA-FEIRA, SÁBADO, DOMINGO
Época do Ano	ANO LECTIVO ou FÉRIAS ESCOLARES
Dia do Ano	Números inteiros entre 1 e 303
Hora do dia	Números inteiros entre 1 e 1440, representando a hora do dia em minutos
Tipo de Hora	MANHÃ, PONTA DA MANHÃ, HORA NORMAL DA MANHÃ, MEIO-DIA, TARDE, HORA NORMAL DA TARDE, PONTA DA TARDE, NOCTURNO, MADRUGADA
Tipo de Desvio	MUITO ADIANTADO, ADIANTADO, PONTUAL, ATRASADO

Tabela 7: Conjunto de Dados resultante

Capítulo 3

Modelação

4.1 Modelação de Árvores de Decisão

4.1.1 Desenho dos testes

Uma vez que neste trabalho se pretendiam avaliar métodos que fossem capazes de fornecer informação útil independentemente do conjunto de dados escolhido pelo utilizador, foram efectuados testes em seis conjuntos de dados de diferentes dimensões (Tabela 8).

Nome	Descrição	Número de Instâncias
Out_14_20	Viagens realizadas no mês de Outubro entre as 14h e as 20h	9194
Ago_Out_8_12	Viagens realizadas entre 15 de Agosto e 15 de Outubro entre as 8 e as 12h	15856
Junho	Viagens realizadas durante o mês de Junho	17437
Jan_Mar	Viagens realizadas nos meses de Janeiro, Fevereiro e Março	41010
Ago_Out	Viagens realizadas entre 15 de Agosto e 15 de Outubro	44727
Mai_Out	Viagens realizadas entre Maio e Outubro de 2007	111901

Tabela 8: Conjuntos de dados utilizados nas experiências com árvores de decisão

O objectivo era avaliar a capacidade do método de lidar com conjuntos de dados de diferentes dimensões. Para isso, foram testadas duas técnicas de *pruning* de forma a avaliar aquela que melhor se adequava ao problema em causa tendo em conta os conjuntos de dados utilizados. As técnicas utilizadas foram *Reduced Error Pruning* (REP) e *Error Based Pruning* (EBP), tendo sido escolhidas pelos seus bons desempenhos já destacados na secção 2.3.1. Para a técnica de EBP, foram ainda testados diferentes valores para o factor de confiança (25, 1, 0.1 e 0.01) pois o desempenho deste algoritmo está directamente relacionado com o valor escolhido para CF (Lawrence O. Hall 2002). O número mínimo de instâncias por folha usado foi 50. Todos os outros parâmetros utilizados serão os parâmetros utilizados por defeito pelo algoritmo.

A análise e comparação dos testes efectuados foram feitas com base nos valores de tamanho da árvore, número de folhas e percentagem de instâncias correctamente classificadas. Como já foi mencionado anteriormente, aquilo que se pretendia não era a obtenção de modelos que fossem capazes de classificar instâncias desconhecidas, mas sim um modelo que fosse capaz de descrever o conjunto de dados de uma forma que seja de fácil interpretação para o utilizador e tão precisa quanto possível. Por essa razão, a avaliação do número de instâncias correctamente classificadas foi feita utilizando o conjunto de treino, através da técnica de *10 fold cross-validation*.

4.1.2 Resultados e Discussão

Os testes foram efectuados segundo o planeamento descrito na secção anterior. Os resultados podem ser observados nos histogramas das Figura 16 e Figura 17.

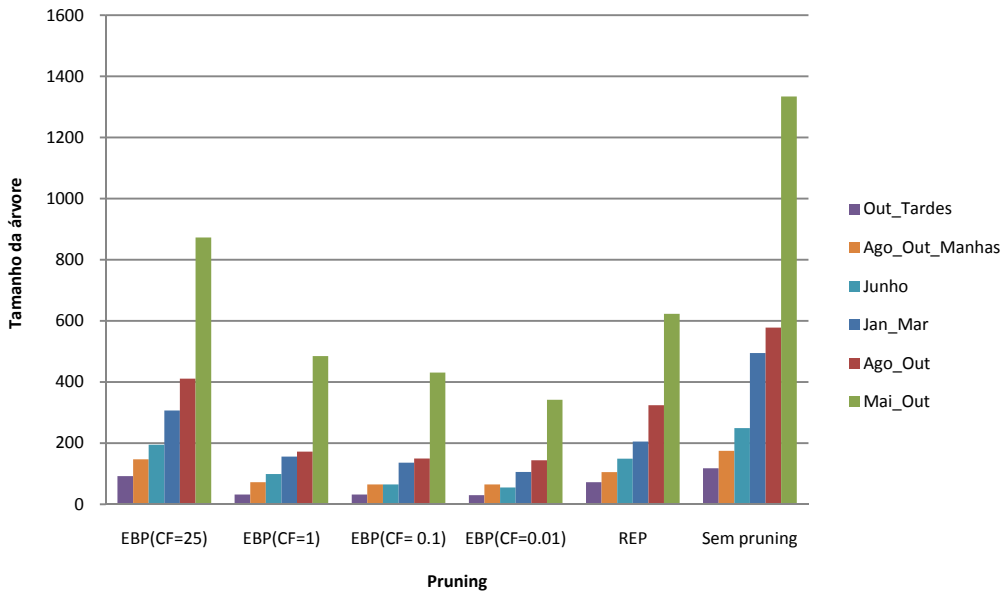

Figura 16: Tamanho da árvore gerada em cada um dos testes

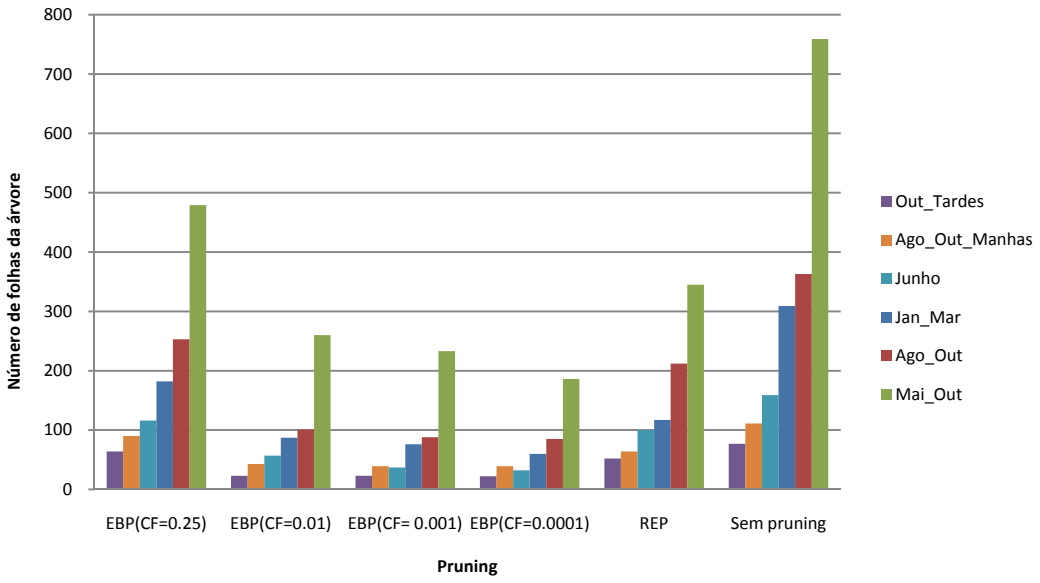

Figura 17: Número de folhas da árvore gerada em cada um dos testes

Como se pode observar nos gráficos da Figura 16 e Figura 17, tanto o tamanho como o número de folhas da árvore crescem com o aumento do tamanho do conjunto de dados. Observa-se ainda que, no caso do algoritmo EBP, a diminuição do valor de CF resulta na diminuição do tamanho e do número de folhas da árvore, tal como esperado. Comparando com o algoritmo REP, este só obtém árvores de menor tamanho quando utilizamos CF = 25 no algoritmo EBP. Em todos os outros casos, o algoritmo EBP obtém árvores de menores dimensões.

A diminuição do tamanho da árvore aumenta muito a interpretabilidade da mesma. No entanto, no caso dos conjuntos de dados maiores, mesmo utilizando técnicas para reduzir o tamanho da árvore, estas continuam a ter dimensões que tornam a sua interpretação uma tarefa difícil. Para além disso, é importante que a simplificação da árvore não implique uma redução significativa da capacidade de previsão pois, nesse caso, esta pode não compensar a perda de precisão. O gráfico da Figura 18 apresenta a percentagem de instâncias correctamente classificadas em cada um dos testes.

No gráfico da Figura 18 observa-se que os melhores resultados em termos de previsão se referem ao algoritmo EBP utilizando CF = 25 ou à não utilização de qualquer técnica de *pruning*. No entanto, a não utilização de nenhuma técnica de *pruning* conduz a um maior número de regras e consequentemente, a árvores mais complexas. Relativamente à técnica de *pruning* EBP, à medida que se reduz o valor de CF, a percentagem de instâncias correctamente classificadas também diminui. No pior caso, o conjunto de dados AGO_OUT, temos uma diminuição de 3,12%. Verifica-

se ainda que a precisão do modelo não está directamente relacionada com o tamanho do conjunto de dados. Isto porque estes estão ordenados por ordem crescente, da esquerda para a direita, e verifica-se que o conjunto de dados JAN_MAR é aquele que obtém melhores resultados.

Um aspecto negativo que pode ser também observado neste histograma tem a ver com os fracos resultados, em termos de percentagem, de instâncias classificadas correctamente. A árvore que tem melhor desempenho neste aspecto classifica correctamente apenas 62,73% das instâncias.

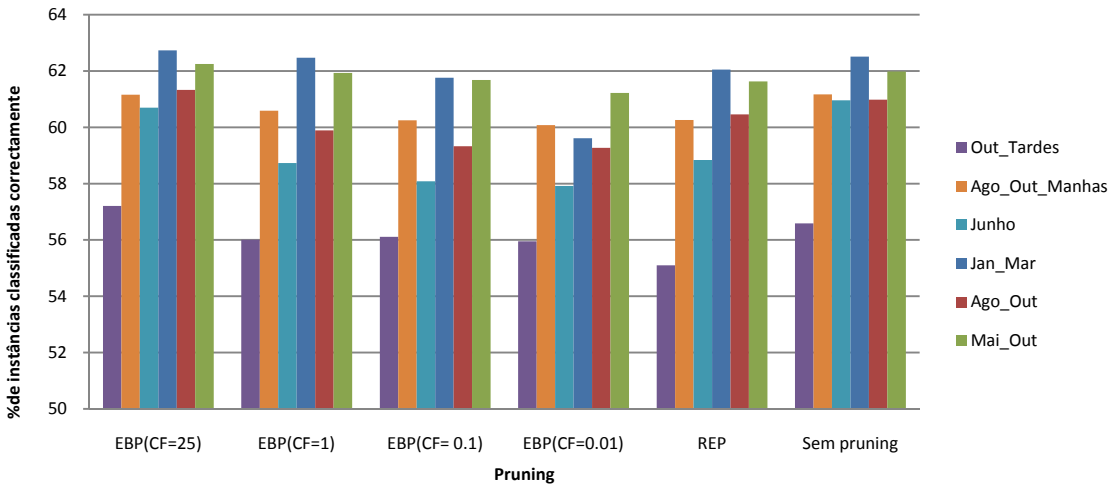

Figura 18: Percentagem de instâncias classificadas correctamente em cada um dos testes

Pode-se assim concluir que, tendo em conta os objectivos que se pretendem atingir com a criação do modelo, o mais adequado será utilizar o método de *Error Based Pruning* (EBP) para fazer a poda da árvore. A utilização de CF = 1% é uma boa opção pois permite obter as árvores de dimensões mais reduzidas, sem sacrificar substancialmente a precisão. No entanto, para determinados conjuntos de dados pode ser necessário diminuir o valor de CF para que seja obtida uma árvore de menores dimensões, sendo assim mais fácil de interpretar.

Para demonstrar o tipo de informação que pode ser fornecida por uma árvore de decisão, apresenta-se na Figura 19 a árvore de decisão construída a partir do conjunto Out_Tardes, utilizando a técnica de *pruning* EBP com CF = 1%.

A percentagem de instâncias classificadas correctamente na totalidade da árvore é de 55,06%. Existem por isso muitas folhas da árvore cuja precisão é reduzida, mas ainda assim, a informação fornecida por estas é útil pois sabemos que a classe representada na folha ocorre mais vezes do que todas as outras classes. Para além disso, a árvore de decisão permite-nos ter uma visão global do comportamento das viagens neste período.

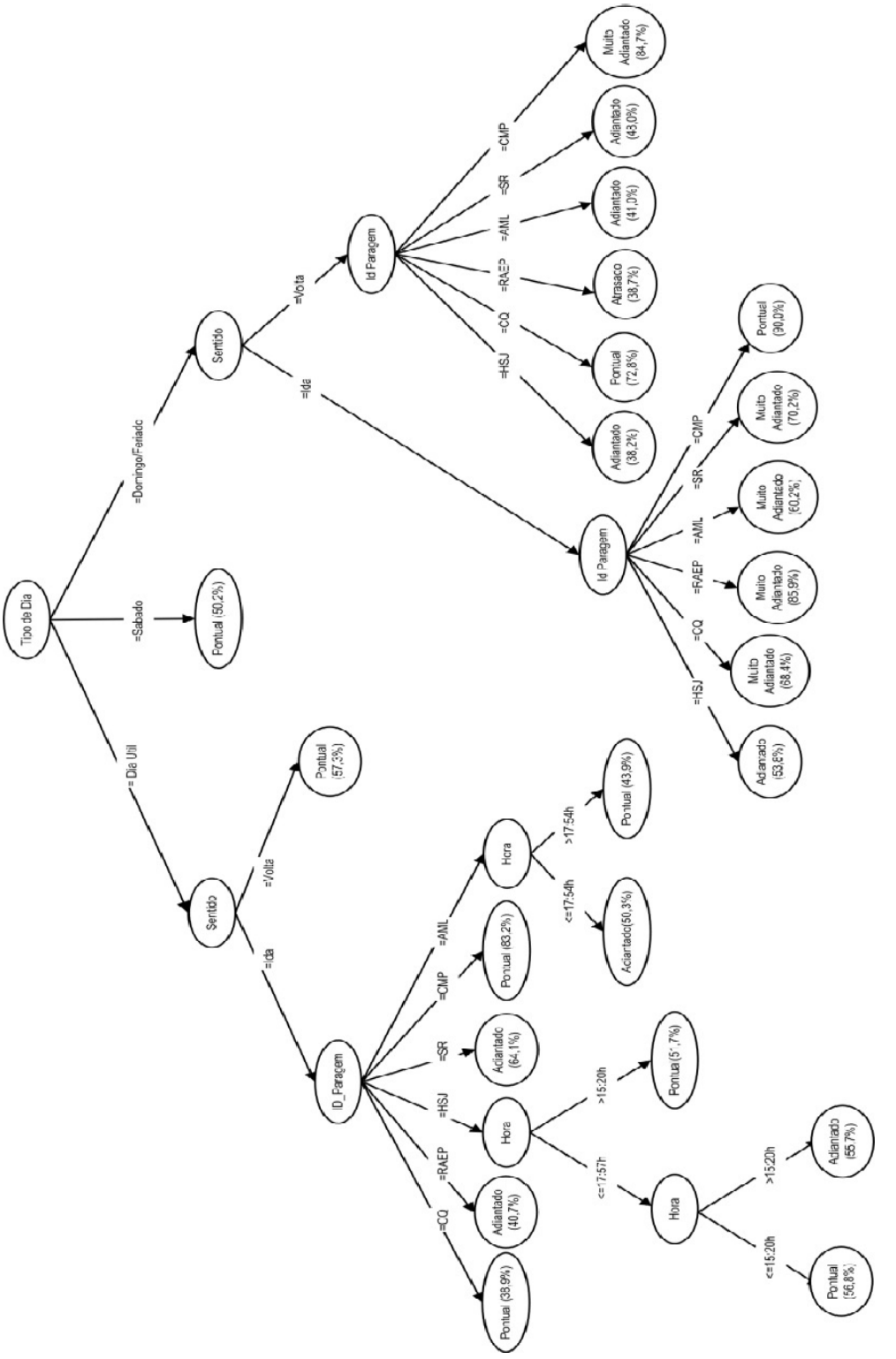

Figura 19: Árvore de decisão obtida a partir do conjunto de dados Out_Tardes,, utilizando a técnica de *pruning* EBP com CF = 1%

Depois de uma primeira análise da árvore, uma das folhas que poderia chamar mais à atenção é a folha com o caminho

```
Tipo de Dia = Domingo/Feriado, Sentido = Volta, Id Paragem = CMP,
```

cuja classificação é Muito Adiantado em 84,7% dos casos. No entanto, na prática, este não é um caso que mereça muita atenção. Isto porque o problema de as viagens estarem adiantadas é mais grave nas paragens intermédias ou na primeira paragem, pois pode fazer com que o utilizador perca o autocarro. Neste caso, no sentido Volta, as viagens terminam na paragem CMP, o que significa apenas que o autocarro vai terminar a viagem mais cedo.

Este caso é apenas um exemplo que demonstra a importância da análise da árvore ser feita por especialistas do domínio: é importante que todos os aspectos do problema e a informação empírica que estes especialistas possuem sejam tidos em conta para que a informação fornecida pelo modelo seja verdadeiramente útil.

4.2 Descoberta de CARs

4.2.1 Desenho dos testes

A ferramenta *Weka* possui uma implementação do algoritmo *Apriori* original, tendo ainda a opção de utilizar o algoritmo para obter CARs, através da adaptação proposta em 1998 por Liu *et al.* (Bing Liu, Wynne Hsu et al. 1998). Após algumas experiências realizadas com esta ferramenta, verificou-se que esta não realiza *pruning* após a elaboração das regras. Isto faz com que seja gerado um grande número de regras redundantes com mesma informação e abrangendo o mesmo número de casos. Desta forma, e apesar da sua limitação no que toca ao número de registos que podem ser utilizados, optou-se por realizar as experiências recorrendo à ferramenta *CBA*. Então, foram feitas experiências com vários conjuntos de dados e para diferentes valores de mínimo suporte e confiança mínima. O objectivo era analisar o número e a qualidade das regras obtidas fazendo variar estes valores, bem como o tamanho da amostra. Os conjuntos de dados que foram utilizados são os que se apresentam na Tabela 9. O conjunto de dados Mai_Out não foi utilizado devido à limitação do número de registos na ferramenta CBA.

Nome	Descrição	Número de Instâncias
Out_Tardes	Viagens realizadas no mês de Outubro entre as 14h e as 20h	9194
Ago_Out_Manhãs	Viagens realizadas entre 15 de Agosto e 15 de Outubro entre as 8 e as 12h	15830
Junho	Viagens realizadas durante o mês de Junho	17437
Jan_Mar	Viagens realizadas nos meses de Janeiro, Fevereiro e Março	41010
Ago_Out	Viagens realizadas entre 15 de Agosto e 15 de Outubro	44727

Tabela 9: Conjuntos de dados utilizados nas experiências de associação

Os valores testados para suporte mínimo e confiança mínima foram:

Suporte mínimo: 5%, 2%, 1%, 0,5%, 0,1% e 0,05%;

Confiança mínima: 50%, 75% e 90%.

Tendo em conta que este algoritmo só trabalha com atributos nominais, não foi possível utilizar-se o atributo Hora representado pelos minutos do dia (tal como se tinha utilizado nas árvores). Sendo assim, optou-se por utilizar a sua representação nominal, construída com base nos critérios utilizados pela STCP, ou seja, o atributo TIPO DE HORA.

Relativamente ao atributo DIA DO ANO, este também foi representado por um número inteiro o que significa que tem de ser transformado num atributo discreto para ser utilizado. O método utilizado para fazer essa transformação foi o método proposto por Fayyad e Irani em (Fayyad and Irani 1993), método este que é também utilizado no algoritmo CBA (Bing Liu, Wynne Hsu et al. 1998). Os atributos utilizados foram os seguintes: identificador da paragem, tipo de dia, dia da semana, ano lectivo, tipo de hora, dia do ano, tipo de atraso da chegada (classe que queremos classificar).

4.2.2 Resultados e Discussão

Os resultados dos testes efectuados relativamente ao número de regras geradas estão apresentados nos histogramas das Figuras Figura 20, Figura 21 e Figura 22.

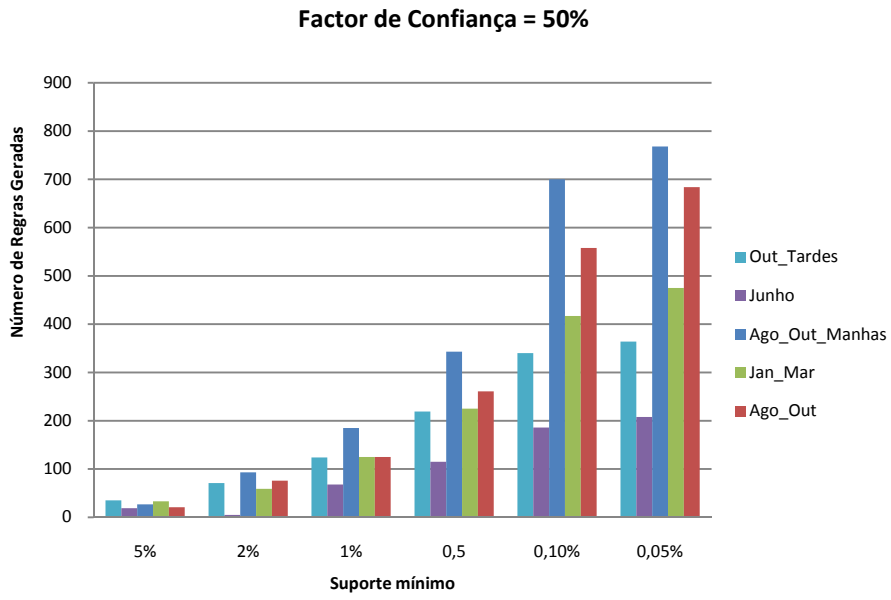

Figura 20: Número de regras geradas por conjunto de dados e suporte mínimo, com CF = 50%

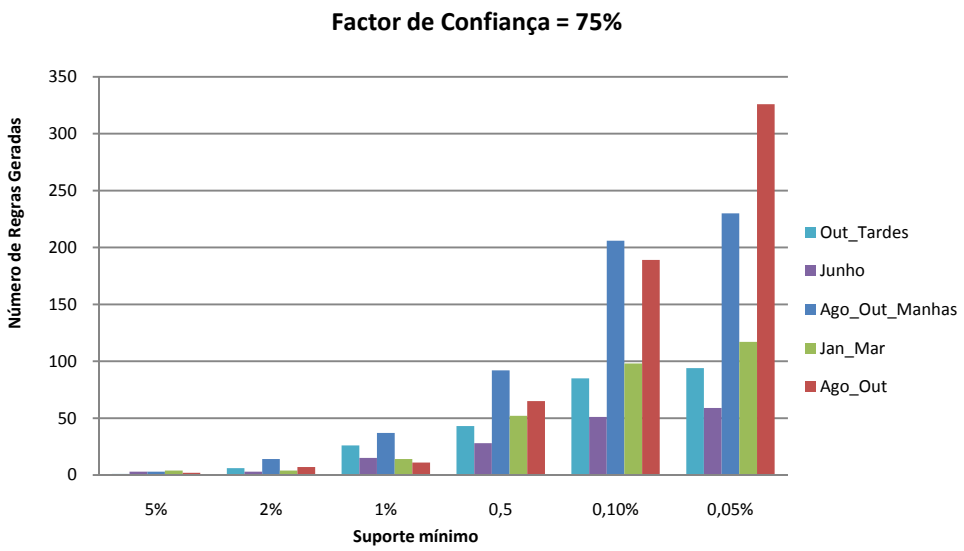

Figura 21: Número de regras geradas por conjunto de dados e suporte mínimo, com CF = 75%

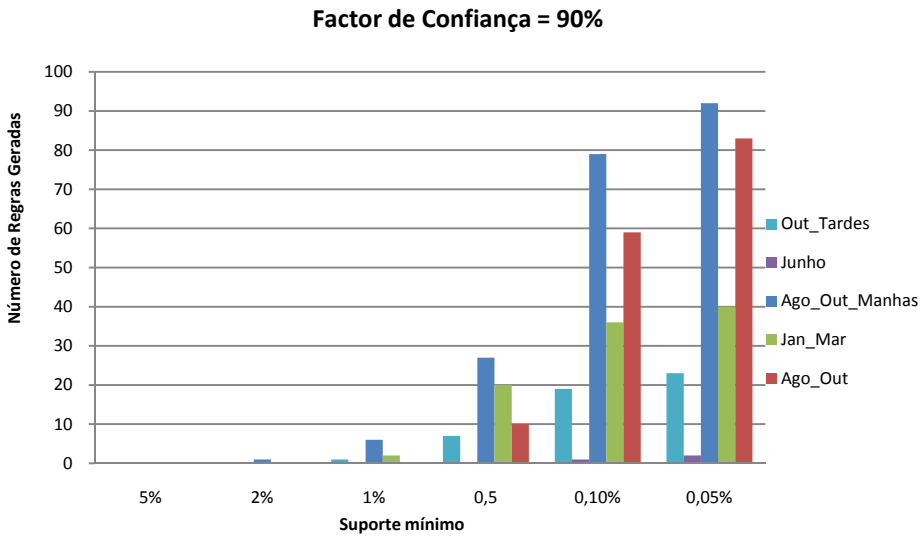

Figura 22: Número de regras geradas por conjunto de dados e suporte mínimo, com CF = 90%

Observa-se ainda que para valores de suporte de 2% e 5% o número de regras geradas é muito reduzido, nos casos em que CF = 75% e CF = 90%, sendo mesmo 0 em muitos dos testes. No gráfico da Figura 22, em que CF = 90%, é ainda de notar o número de regras geradas pelo conjunto Junho: mesmo para os valores de suporte mais baixo, o número de regras máximo gerado é 2.

Verifica-se ainda que o número de regras geradas para os mesmos valores de suporte e confiança não está directamente relacionado com o tamanho do conjunto de dados. Este facto é evidente quando se analisam os conjntos Jan_Mar (41010 instâncias) e Ago_Out_Manhas (15586 instâncias): na grande maioria das vezes, o número de regras geradas pelo segundo conjunto é substancialmente maior do que o número de regras geradas pelo primeiro.

Da análise dos gráficos é possível concluir que a utilização de um factor de confiança baixo (50%) pode conduzir a um elevado número de regras em todos os conjuntos de dados, especialmente para valores de suporte mínimo também baixos. A utilização deste factor de confiança pode não ser a mais adequada pois pode haver um grande número de regras geradas que não têm interesse para o problema. Isto porque uma regra com factor de confiança baixo, 50% por exemplo, caracteriza casos cujo comportamento é muito pouco consistente, ou seja, apesar de 50% das vezes a regra se verificar, existem outras 50% de vezes em que não se verifica, assumindo um comportamento que é desconhecido para o utilizador. Em contrapartida, a utilização de um factor de confiança elevado, como 90%, pode conduzir a um número muito reduzido de regras, caso o suporte mínimo não seja suficientemente pequeno.

Como é evidente pela análise dos gráficos, a escolha do valor utilizado para o suporte mínimo influencia muito o número de regras geradas. Esta escolha deve por isso ter em conta o tamanho do conjunto de dados e o detalhe que se pretende para as regras geradas: quanto maior for o detalhe que se pretende, menor deverá ser o suporte mínimo. As regras seguintes foram obtidas do conjunto Ago_Out, com os valores de 0.1 e 0.05 para suporte mínimo:

```
tipo_de_dia = SABADO
 sentido = IDA
 id_paragem = RAEP
 ->     class = Muito_Adiantado
 (confiança: 84.969%, suporte: 0.910%)

ano_lectivo = TRUE
hora = Meio-dia
dia_da_semana = QUINTA-FEIRA
tipo_de_dia = DIA_UTIL
sentido = IDA
id_paragem = SR
->      class = Adiantado
(confiança: 80.556%, suporte: 0.065%)
```

Verifica-se que ambas as regras têm elevada confiança, mas enquanto a primeira se refere a todas as viagens de um determinado tipo de dia, a segunda especifica o tipo de hora a que a viagem acontece e acrescenta ainda o detalhe de esta apenas se verificar durante o ano lectivo.

Como seria de esperar, verificou-se que os atributos TIPO DE DIA e DIA DA SEMANA podem conduzir a um elevado número de regras com a mesma informação. Isto porque, aos sábados estes atributos assumem o mesmo valor, o que faz com que seja obtida mais do que uma regra com o mesmo significado, tal como ilustrado no exemplo seguinte:

```
tipo_de_dia = SABADO
sentido = IDA
id_paragem = RAEP
->     class = Muito_Adiantado
(confiança: 91.371% suporte: 1.137%)

dia_da_semana = SABADO
sentido = IDA
id_paragem = RAEP
->     class = Muito_Adiantado
(confiança: 91.371% suporte: 1.137%)
```

A análise destes casos poderia levar à eliminação de um destes atributos das experiências. No entanto, ambos os atributos podem revelar-se importantes para detectar comportamentos que ocorrem sistematicamente. No caso do Dia da Semana, este pode ser utilizado para detectar fenómenos que ocorrem apenas em determinados dias da semana. Exemplos deste tipo de fenómenos são alterações no comportamento da linha devido a feiras que ocorrem nas localidades e que podem condicionar o trânsito, ou a diminuição do movimento junto a uma escola onde não há aulas à tarde num determinado dia da semana. Embora este tipo de fenómenos não conduza a alterações de horário pois não é possível fazer um horário diferente para cada dia da semana, é importante que estes sejam conhecidos para que a empresa possa tomar medidas no sentido de diminuir o efeito dos mesmos. Estas medidas podem passar por prevenir os motoristas de que poderão ter de alterar a sua marcha normal por forma a cumprir o horário naqueles dias ou ter preparados veículos de reforço que podem ser necessários para cobrir os atrasos nesses dias. Nos testes efectuados foram encontradas várias regras que se referiam especificamente a um dia da semana, por exemplo:

```
ano_lectivo = TRUE
hora = Meio-dia
dia_da_semana = QUARTA-FEIRA
```

```
sentido = IDA
id_paragem = SR
->     class = Adiantado
(confiança:83.784%, suporte: 0.069%)

ano_lectivo = FALSE
hora = Meio-dia
dia_da_semana = QUARTA-FEIRA
sentido = IDA
id_paragem = SR
->     class = Pontual
(confiança: 96.552%, suporte: 0.063%)


ano_lectivo = TRUE
hora = Meio-dia
dia_da_semana = SEXTA-FEIRA
sentido = IDA
id_paragem = SR
->     class = Adiantado
(confiança: 82.143%, suporte: 0.051%)
```

Relativamente ao atributo Tipo de Dia, a sua permanência também é importante pois pode ser importante para detectar fenómenos que ocorrem apenas nos dias úteis ou aos domingos e feriados (não apenas domingos). Exemplos desses fenómenos são aqueles descritos nas regras que se apresentam em seguida:

```
hora = Madrugada
tipo_de_dia = DIA_UTIL
sentido = IDA
id_paragem = AML
->     class = Muito_Adiantado
(confiança: 75.000%, suporte: 0.732%)

hora = Madrugada
tipo_de_dia = DIA_UTIL
sentido = IDA
id_paragem = CQ
->     class = Muito_Adiantado
(confiança: 61.017%, suporte: 0.351%)
```

É por isso importante que ambos os atributos sejam mantidos.

Existem ainda outras situações em que ocorre a replicação de regras que representam a mesma informação. Apesar de ser utilizada a técnica de *pruning* para fazer o pós-processamento das regras, esta é muitas vezes insuficiente para eliminar todas as redundâncias. Exemplos desse tipo de situação são os que se apresentam nas regras seguintes:

```
hora = Tarde,
tipo_de_dia = DOMINGO/FERIADO,
sentido = VOLTA,
id_paragem = CMP
   ->      class = Muito_Adiantado
   (confiança: 95.833% suporte: 0.750%)

ano_lectivo = TRUE,
hora = Tarde,
tipo_de_dia = DOMINGO/FERIADO,
sentido = VOLTA,
id_paragem = CMP
 ->  class = Muito_Adiantado
 (confiança: 95.833% suporte: 0.750%)

 ano_lectivo = FALSE
 hora = Hora_Normal_da_Manha
 ->  class = Pontual
 (confiança: 56.822% suporte: 2.374%)

 ano_lectivo = FALSE
 hora = Hora_Normal_da_Manha
 tipo_de_dia = DIA_UTIL
 ->  class = Pontual
 (confiança: 56.822% suporte: 2.374%)
```

Analisando as primeiras duas regras, verifica-se que a segunda regra é desnecessária uma vez que o suporte e a confiança em ambas as regras é igual, mas a segunda contém mais uma condição (ano lectivo = true). Isto acontece porque no conjunto de dados utilizado o valor do atributo é sempre o mesmo, o que faz com que a existência ou não desta condição seja irrelevante. Relativamente às duas regras seguintes, o problema é semelhante, sendo que nestas a condição irrelevante é tipo_de_dia = dia útil. Neste caso, isto acontece porque a condição Hora = Hora_Normal_Manha tem implícita a informação de que a regra só se aplica nos dias úteis porque

este tipo de hora só está definido para os dias úteis. Isto faz com que a existência ou não da condição Tipo de Dia = Dia útil seja irrelevante.

É possível assim concluir que o facto de muitos dos atributos estarem relacionados faz com que seja gerado um grande número de regras que não trazem informação relevante.

Em resumo, as conclusões que se podem tirar acerca dos resultados obtidos são:

- O suporte utilizado deve ser escolhido em função do detalhe que se pretende obter nas regras; deve ter-se em conta que suportes de 1% ou mais podem conduzir a um número muito reduzido de regras, especialmente no caso de elevados factores de confiança; suportes de 0.05% ou inferiores podem não fazer sentido pois significam um número mínimo de registos por regra muito reduzido;
- O factor de confiança influencia a quantidade e a qualidade das regras obtidas, devendo por isso ser escolhido em função dos objectivos que se pretendem atingir, tendo em conta que quanto menor for o factor de confiança, maior será o número de regras obtidas;
- Apesar de ser utilizada a técnica de *pruning*, o facto de muitos dos atributos estarem relacionados faz com que seja gerado um grande número de regras redundantes.

4.3 Construção de um Classificador Associativo

4.3.1 Desenho dos testes

Foram feitas experiências com vários conjuntos de dados e para diferentes valores de mínimo suporte e confiança mínima. O objectivo era analisar a quantidade e a qualidade das regras obtidas fazendo variar estes valores, bem como o tamanho da amostra.

Os conjuntos de dados que foram ser utilizados são os mesmos utilizados nas experiências de mineração de CARs (Tabela 9).

Valores entre 1% e 2% são muito utilizados como suporte mínimo para o algoritmo CBA (Han and Pei 2001; Mutter, Hall et al. 2004; Hu and Li 2005; Thabtah, Cowling et al. 2006). No entanto, para o caso que se estava a estudar pode fazer sentido utilizar valores inferiores a 1% por forma a obter regras que descrevam fenómenos menos frequentes mas com elevados factores de confiança. Sendo assim, os valores testados para suporte mínimo serão:

2%, 1%, 0,5%, 0,1% e 0,05%.

Relativamente ao valor para a confiança mínima, este tem menor impacto na qualidade do classificador (Thabtah, Cowling et al. 2006), desde que não seja demasiado elevado. Foram por isso testados os valores de 50% (Han and Pei 2001; Mutter, Hall et al. 2004; Hu and Li 2005) e 30% (Thabtah, Cowling et al. 2004). Tendo em conta que este algoritmo só trabalha com atributos

nominais, foram utilizados os mesmos atributos que se utilizaram na descoberta de CARs (Secção 4.2), procedendo à transformação do atributo Dia do Ano num atributo discreto segundo o método proposto por Fayyad *et al* em (Fayyad and Irani 1993). Os atributos utilizados são então os seguintes: Identificador da paragem, Tipo de Dia, Dia da Semana, Ano Lectivo, Tipo de Hora, Dia do Ano e Tipo de Atraso da Chegada (classe que queremos classificar).

4.3.2 Resultados e Discussão

Os resultados dos testes efectuados relativos à precisão e ao número de regras geradas pelo modelo estão representados nas Figuras 23, 24, 25 e 26.

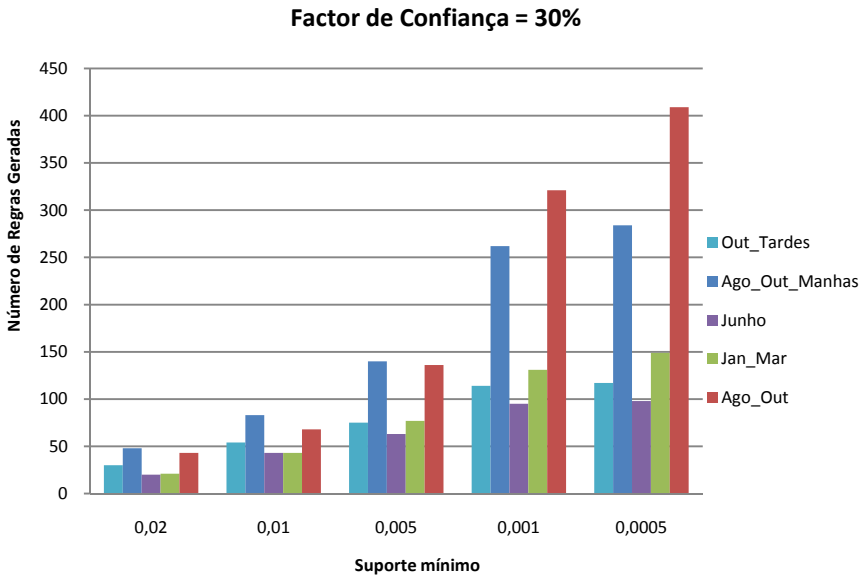

Figura 23: Número de regras do classificador obtido com CF = 30%, para cada um dos conjuntos de teste

Nos gráficos das Figuras 23 e 24 é possível observar que o número de regras obtidas pelo classificador é maior para valores de suporte mais baixos, o que já seria de esperar. No entanto, em algumas situações, esse crescimento é muito reduzido. É o caso dos conjuntos Out_Tardes e Junho, quando passamos de suporte 0,1% para 0,05%. Isto pode dever-se ao facto da diminuição de suporte não resultar num aumento significativo do número de regras com elevada confiança. Em ambos os gráficos, o conjunto Ago_Out distingue-se por ser aquele que obtém maior número de regras, para a grande maioria dos valores de suporte mínimo e confiança testados. Um dos motivos para este resultado pode ser o facto de este ser o maior de todos os conjuntos de dados

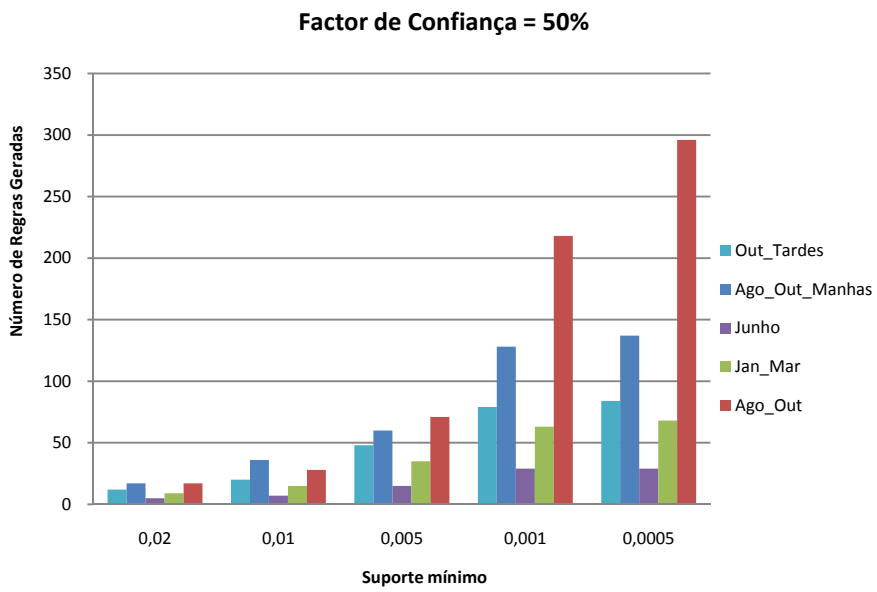

Figura 24: Número de regras do classificador obtido com CF= 50%, para cada um dos conjuntos de teste

utilizados. No entanto, este factor não é suficiente pois, se o tamanho do conjunto de dados estivesse directamente relacionado com o número de regras geradas, não teríamos o conjunto Ago_Out_Manhas (15856 instâncias) a gerar um número muito maior de regras do que o conjunto Junho (17437 instâncias), no caso em que CF = 50% (Figura 24). Para além disso, também o conjunto Jan_Mar (41011 instâncias), apesar de ser substancialmente maior do que os conjuntos Junho e Ago_Out_Manhas, gera na maioria dos testes um número de regras inferior a estes conjuntos. Sendo assim, apesar de podermos considerar que o tamanho do conjunto de dados possa ter alguma influência no conjunto de regras geradas, a variedade e complexidade da informação contida em cada um deles pode também exercer uma grande influência.

No caso dos conjuntos Ago_Out e Ago_Out_Manhas, o elevado número de regras obtido pode dever-se ao facto de ambos se referirem a um período de tempo que engloba a época de Verão e o Ano Lectivo. Nestas duas épocas, para além dos horários e a oferta de viagens serem distintos, o volume de trânsito é também muito diferente de uma época para a outra, o que faz com que possamos ter comportamentos muito variados, aumentando por isso o número de regras geradas.

A análise do número de regras geradas permite-nos analisar a interpretabilidade do modelo. No entanto, é necessário também avaliar a sua precisão, para que possamos ter uma perspectiva global da qualidade da descrição dos dados feita pelo modelo. Os gráficos das figuras Figura 25 e Figura 26 apresentam a percentagem de instâncias correctamente classificadas para cada um dos factores de confiança testados.

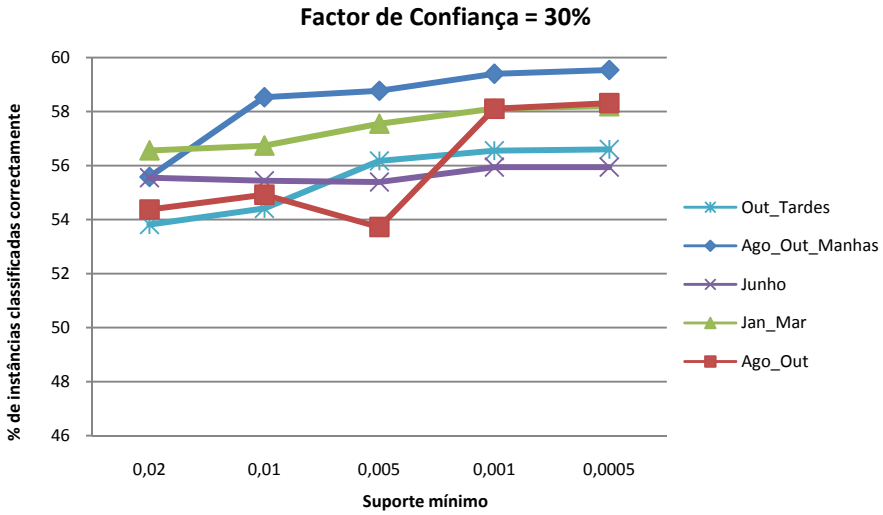

Figura 25:Percentagem de instâncias classificadas correctamente pelo classificador obtido com CF = 30%, para cada um dos conjuntos de teste

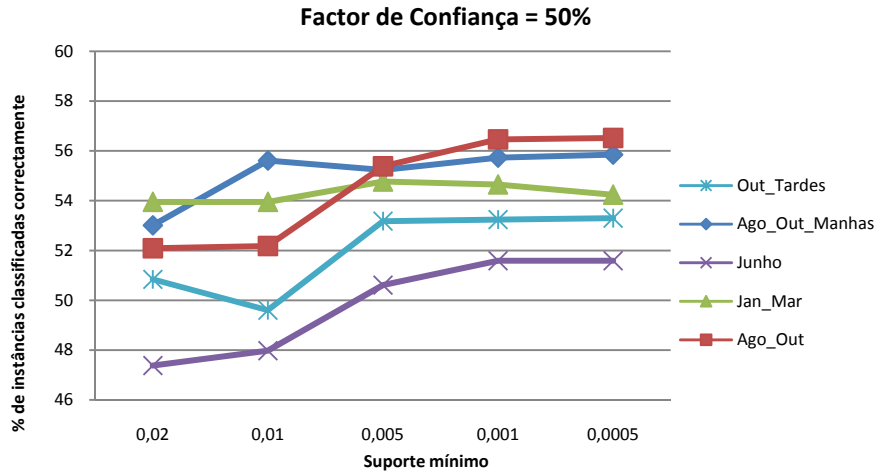

Figura 26: Percentagem de instâncias classificadas correctamente pelo classificador obtido com CF = 50%, para cada um dos conjuntos de teste

Nos gráficos figuras Figura 25 e Figura 26 observa-se claramente que, na grande maioria dos casos, os resultados são melhores quando se utiliza CF = 30%. A excepção é feita apenas para o conjunto de dados Ago_Out, com suportes 0,5%, onde a precisão do modelo é ligeiramente superior no caso de CF = 50%. No entanto, em todos os outros casos a precisão é melhor com CF = 30%. O caso onde essa diferença é mais notória é o conjunto Junho, onde se chega a atingir uma diferença de 8,17% com suporte 2%. Uma possível explicação para este fenómeno é a existência de um elevado número de regras com factores de confiança baixos e que por isso influenciam a qualidade do classificador obtido.

Tal como seria de esperar, o suporte mínimo utilizado também exerce influência sobre a precisão do modelo. Na maioria dos casos, a diminuição do suporte resulta num aumento da precisão, sendo os melhores resultados obtidos quando o suporte é 0,05%. O único caso em que isto não acontece é no conjunto Jan_Mar, quando CF = 50%, que obtém o melhor valor com suporte de 0,5%..

Apesar de ser possível melhorar a precisão do modelo através de alterações no factor de confiança ou suporte mínimo, os resultados obtidos são muito pouco satisfatórios para um modelo classificador. O melhor resultado, 59,54%, é alcançado pelo conjunto Ago_Out_Manhas com CF = 30% e suporte de 0,05%. Este resultado, embora não fosse desejável, já seria de esperar, tendo em conta que os comportamentos descritos nos conjuntos de dados podem ser muito irregulares em determinadas situações. No entanto, isto não invalida a utilização deste tipo de classificador para obter uma descrição geral do conjunto de dados, até porque muitas das regras geradas podem ter um elevado factor de confiança.

Fazendo uma análise global dos gráficos referentes ao número de regras e precisão do modelo (Figura 23, Figura 24, Figura 25 e Figura 26), podemos afirmar que, apesar de poder conduzir a um número de regras mais elevado, a utilização de CF = 30% é mais adequada, tendo em conta que o aumento da precisão pode ser muito significativo em determinados casos. Relativamente ao suporte mínimo, é aconselhável que seja utilizado um valor inferior a 1% pois obtém melhores resultados na precisão do modelo. Na maioria dos casos, a utilização de um suporte inferior a 0,5% não resulta num aumento significativo da precisão, aumentando apenas o número de regras. Por essa razão, a utilização de um suporte inferior a 0,5% pode ser prejudicial, aumentando a complexidade do modelo sem aumentar a precisão. Para além disso, para conjuntos de dados pequenos não faz sentido utilizar um suporte muito baixo pois cada regra poderá corresponder a um número muito reduzido de casos, o que pode não ter interesse para o problema.

Para demonstrar o tipo de informação que pode ser fornecida por este modelo, apresentam-se em seguida as primeiras 20 regras geradas a partir do conjunto Out_tardes, com 0,5% de suporte e 30% de confiança. O classificador é composto por 75 regras e é apresentado na sua totalidade no Anexo I.

Como se pode observar pela análise da Tabela 10, o classificador gera um grande número de regras com confiança elevada. No entanto, existem conjuntos de regras que representam situações que ocorrem em condições semelhantes: é o caso dos conjuntos {1,6} e {11, 17}. No primeiro conjunto de regras é descrita a situação do autocarro chegar muito adiantado à paragem da rotunda AEP, no sentido Ida, sendo que a única diferença entre as condições é que na regra 1 se refere apenas aos domingos e na regra 6 refere-se aos domingos e feriados. No segundo conjunto de regras é descrita a situação do autocarro chegar adiantado à paragem SR nas horas

normais da tarde, sendo que a regra 11 se refere a esta situação apenas às segundas-feiras enquanto a regra 17 se refere a esta situação em todos os dias úteis (uma vez que a classificação Hora_Normal_da_Tarde existe apenas para os dias úteis). Estas situações não representam replicação de informação nas regras porque elas não representam exactamente as mesmas condições nem têm os mesmos valores de suporte e confiança. No entanto, a semelhança da informação fornecida demonstra a necessidade de uma análise cuidada de todas as regras para que se possa tirar o melhor proveito da informação extraída. Neste caso, as regras 1 e 17 deveriam merecer mais atenção por representarem situações mais abrangentes. A informação fornecida por estas regras poderia inclusive conduzir a alterações nos horários desta linha nesses períodos.

1	dia_da_semana = DOMINGO, sentido = IDA, id_paragem = RAEP -> class = Muito_Adiantado (confiança: 91.667% , suporte: 0.598%)
2	dia_da_semana = QUARTA-FEIRA, sentido = IDA, id_paragem = CMP -> class = Pontual (confiança: 90.244% , suporte: 1.207%)
3	tipo_de_dia = DOMINGO/FERIADO, sentido = IDA, id_paragem = CMP -> class = Pontual (confiança:90.000% , suporte: 0.783%)
4	hora = Hora_Normal_da_Tarde, sentido = IDA, id_paragem = CMP -> class = Pontual (confiança: 88.532% , suporte: 2.099%)
5	hora = Tarde, sentido = IDA, id_paragem = CMP -> class = Pontual (confiança: 88.356% , suporte: 1.403%)
6	tipo_de_dia = DOMINGO/FERIADO, sentido = IDA, id_paragem = RAEP -> class = Muito_Adiantado (confiança: 85.897% , suporte: 0.729%)
7	dia_da_semana = DOMINGO, sentido = VOLTA, id_paragem = CMP -> class = Muito_Adiantado (confiança: 85.714% , suporte: 0.522%)
8	dia_da_semana = QUARTA-FEIRA, sentido = VOLTA, id_paragem = RAEP -> class = Pontual (confiança: 85.246% , suporte: 1.131%)
9	tipo_de_dia = DOMINGO/FERIADO, sentido = VOLTA, id_paragem = CMP -> class = Muito_Adiantado (confiança: 84.722% , suporte: 0.663%)
10	sentido = IDA, id_paragem = CMP -> class = Pontual (confiança: 84.225% , suporte: 6.852%)

11	hora = Hora_Normal_da_Tarde, dia_da_semana = SEGUNDA-FEIRA, sentido = IDA, id_paragem = SR -> class = Adiantado (confiança: 83.636% , suporte: 0.500%)
12	hora = Ponta_da_Tarde, dia_da_semana = QUARTA-FEIRA, sentido = VOLTA, id_paragem = CQ -> class = Pontual (confiança: 82.432% , suporte: 0.663%)
13	dia_da_semana = QUARTA-FEIRA, sentido = VOLTA, id_paragem = CQ -> class = Pontual (confiança: 80.769% , suporte: 1.142%)
14	hora = Ponta_da_Tarde, dia_da_semana = SEGUNDA-FEIRA, sentido = VOLTA, id_paragem = CQ -> class = Pontual (confiança: 80.000% , suporte: 0.827%)
15	dia_da_semana = SEGUNDA-FEIRA, sentido = VOLTA, id_paragem = CQ -> class = Pontual (confiança: 78.528% , suporte: 1.392%)
16	hora = Hora_Normal_da_Tarde, sentido = VOLTA, id_paragem = RAEP -> class = Pontual (confiança: 77.725% , suporte: 1.784%)
17	hora = Hora_Normal_da_Tarde, sentido = IDA, id_paragem = SR -> class = Adiantado (confiança: 77.626% , suporte: 1.849%)
18	dia_da_semana = SEGUNDA-FEIRA, sentido = VOLTA, id_paragem = RAEP -> class = Pontual (confiança: 77.333% , suporte: 1.262%)
19	hora = Ponta_da_Tarde, sentido = VOLTA, id_paragem = CQ -> class = Pontual (confiança: 76.839% , suporte: 3.067%)
20	hora = Ponta_da_Tarde, dia_da_semana = QUARTA-FEIRA, sentido = IDA, id_paragem = SR -> class = Adiantado (confiança: 75.342% , suporte: 0.598%)

Tabela 10: Primeiras 20 regras do classificador gerado a partir do conjunto Out_tardes com 0,5% de suporte e 30% de confiança

Capítulo 5

Conclusões e Trabalho Futuro

5.1 Análise crítica dos Resultados

Através dos resultados obtidos nos testes efectuados, verificou-se que as técnicas utilizadas são capazes de detectar e caracterizar a ocorrência de situações sistemáticas relativamente aos desvios entre o horário previsto e a hora real de passagem. No entanto, na análise da informação obtida é importante ter em conta todos os aspectos do problema e a forma como foram construídos os modelos, para que se possa tirar o melhor partido da informação devolvida.

Um aspecto importante é o facto das diferenças registadas, ou seja, os desvios, serem calculados tendo em conta a hora a que o autocarro chega à paragem ou, mais precisamente, a hora a que o GPS detecta a chegada do autocarro à paragem. Isto faz com que não seja possível determinar com exactidão se o tipo de desvio registado à chegada (Muito Adiantado, Adiantado, Pontual, etc.) é o mesmo tipo de desvio registado quando o autocarro parte. Por exemplo, um autocarro que chega a uma determinada paragem atrasado 4 minutos tem um desvio classificado como Pontual, mas se estiver parado mais de um minuto, o desvio relativamente à hora de partida prevista será classificado como Atrasado. Esta incerteza relativamente ao tempo que o autocarro fica parado numa certa paragem pode perturbar a análise da informação fornecida pelo modelo. É por isso importante que o conhecimento adquirido pelos planeadores de horário através da sua

experiência seja sempre utilizado pois este pode fornecer informações adicionais que podem ser decisivas. Por exemplo, se existe uma situação que ocorre sempre na mesma paragem e que é classificada como Adiantada (desvio entre -1 e -6), é importante que o analista tenha pelo menos uma ideia do fluxo de passageiros e do tempo que o autocarro poderá estar parado naquela paragem. Isto porque, numa avaliação precipitada da situação poderíamos considerar necessário o reajustamento do horário para que o autocarro não parta antes da hora. No entanto, o conhecimento e a experiência do planeador de horários poderá ajudar a perceber se o autocarro irá de facto partir adiantado ou se, nos casos em que existem muitos passageiros e o autocarro fica muito tempo parado e ele na realidade partirá a horas.

Outro aspecto importante é a quantidade de desvios que são classificados como *Pontual*, que representam aproximadamente 50% de todos os desvios calculados. Poderia ser questionado o facto de estas situações serem incluídas neste tipo de análise, visto que o objectivo principal é detectar situações passíveis de serem corrigidas. No entanto, o conjunto de desvios que se encontram nesta classe é constituído por uma janela temporal muito grande, podendo por isso representar uma grande variedade de situações. Por um lado este é um bom indicador pois demonstra que a grande maioria das viagens chega às paragens "pontualmente", tendo em conta esta classificação. Todavia, estar-se perante uma situação de um desvio é sistemático de 3 ou 4 minutos é mais grave do que um desvio sistemático (à chegada) de -1 ou 0 minutos. Por essa razão, tendo em conta os objectivos que se pretendem atingir neste trabalho, poderia ser mais adequado termos considerado outra janela temporal para os autocarros que circulam "pontualmente".

Ainda assim, pode ser útil para o utilizador ter conhecimento das situações que são classificadas como *Pontuais*, tendo em conta a classificação actual. Em primeiro lugar, porque pode evitar que sejam reajustados horários que não precisam ou não devem ser reajustados. Em segundo lugar porque uma situação classificada como *Pontual* pode significar que o autocarro vai partir atrasado, nos casos em que o tempo gasto nas paragens para entrada e saída de passageiros é mais elevado. Mais uma vez se denota a importância da experiência que os analistas possuem relativamente a esses aspectos. Por exemplo, uma situação classificada como *Pontual* no Hospital de São João poderá ser alvo de atenção pois é sabido que este é um local onde entram e saem muitos passageiros podendo fazer com que o autocarro parta atrasado e não seja capaz de recuperar esse atraso.

5.2 Comparação entre os métodos

Uma comparação entre os diferentes métodos, realçando os pontos fortes e fracos de cada um deles é fundamental para se decidir aquele que é mais adequado para ir de encontro aos

objectivos do negócio. Serão comparados os métodos sob um conjunto de critérios, sendo que alguns deles só podem ser aplicados aos métodos de classificação: precisão do classificador, interpretabilidade, qualidade/interesse das regras geradas, facilidade de classificação de uma nova instância, capacidade para detectar erros sistemáticos e versatilidade relativamente aos tipos de dados utilizados.

A comparação e análise da precisão do classificador é importante porque fornece uma ideia geral da qualidade da descrição que o modelo irá fazer dos dados e das regras que irão ser obtidas. Ou seja, um modelo que classifica 60% das instâncias correctamente poderá ter muitas regras (folhas) com elevado factor de confiança (percentagem de instâncias classificadas correctamente), mas também terá certamente um elevado número de regras/folhas em que esses valores são muito baixos. Ambos os métodos estudados que constroem um classificador (árvores de decisão e CBA) não conseguem obter precisões elevadas: nas árvores de decisão variam entre 62,25% e 55,1% e no algoritmo CBA variam entre 59,54% e 53,3%. Na generalidade dos casos, os resultados são melhores nas árvores de decisão, o que contraria os resultados obtidos em (Bing Liu, Wynne Hsu et al. 1998), que demonstram que a precisão do algoritmo CBA é geralmente melhor do que a do algoritmo C4.5.

A interpretabilidade dos modelos obtidos é um aspecto fundamental para os objectivos de negócio pois determina a facilidade com que o utilizador (planeador de horários) irá extrair informação a partir do modelo. Se a informação for fornecida em “grandes quantidades”, apresentada de forma confusa e desorganizada, o utilizador não será capaz de extrair qualquer informação útil para a empresa. A complexidade de uma árvore de decisão pode ser medida pelo número de folhas da árvore (LA Breslow 1997), sendo que podemos utilizar um critério semelhante para os outros algoritmos, medindo a sua complexidade através do número de regras geradas. Avaliando segundo este critério, as árvores de decisão conseguem produzir modelos mais fáceis de interpretar uma vez que tendem a produzir um menor número de regras do que o algoritmo CBA (Hu and Li 2005), (Mutter, Hall et al. 2004), e por sua vez, o algoritmo CBA produz menos regras do que o algoritmo CBA-RG. Para além disso, a representação hierárquica das árvores faz com que seja possível ter uma visão global do conjunto de dados de uma forma simples e intuitiva. Isto já não acontece no classificador associativo e nas CARs uma vez que a representação na forma de uma lista de regras dificulta a compreensão geral do conjunto, especialmente se o número de regras for elevado.

A qualidade das regras geradas é também um aspecto fundamental nos modelos obtidos. O método utilizado na construção de árvores de decisão divide sucessivamente o conjunto de dados fazendo com que exista apenas um caminho (regra) possível para cada instância (Quinlan 1993). Pelo contrário, a técnica utilizada no algoritmo CBA e na produção de CARs procura encontrar o conjunto completo de regras de classificação através da procura exaustiva de relacionamentos

entre os atributos (Bing Liu, Wynne Hsu et al. 1998), o que possibilita que se encontrem regras mais interessantes e com elevado factor de confiança. O algoritmo CBA-RG é aquele que poderá produzir maior número de regras com elevado factor de confiança, uma vez que estas apenas passam pela tarefa de *pruning* que procura eliminar a maioria das redundâncias. No entanto, este passo é muitas vezes insuficiente para eliminar todas as regras que não representam nenhuma informação interessante ou adicional relativamente às outras, o que faz com que o conjunto de CARs seja constituído ao mesmo tempo por um grande número de regras interessantes e um grande número de regras redundantes.

Associado à qualidade das regras geradas está a capacidade de cada um dos modelos de detectar erros sistemáticos. Pelos motivos que se explicaram anteriormente, a mineração de CARs é a técnica que têm melhor desempenho neste aspecto.

Outro aspecto importante na avaliação dos métodos de classificação é a facilidade de classificação de uma nova instância. Os métodos classificadores utilizados neste trabalho permitem ao utilizador monitorizar o comportamento dos autocarros em qualquer situação. Por exemplo, se o utilizador colocar a questão *qual é o comportamento das viagens nos dias úteis à hora de ponta?*, tanto um classificador baseado em árvores de decisão como um classificador baseado em associação poderia responder à questão. Nas árvores de decisão o processo de classificação é bastante simples uma vez que cada condição vai dividindo sucessivamente o espaço de resultados possíveis em dois ou mais conjuntos. Todavia, o mesmo não acontece num classificador associativo no qual é necessário percorrer todas as regras a partir do início até encontrar aquela que classifica a instância em questão. Se o classificador for composto por muitas regras este pode ser um processo moroso e cansativo.

O último aspecto que irá ser analisado para comparar os três métodos prende-se com a versatilidade na utilização de diferentes tipos de dados. As técnicas de mineração de CARs e construção de um classificador associativo não permitem a utilização de valores numéricos. Isto faz com que atributos deste tipo tenham de ser transformados em atributos discretos podendo perder-se precisão nas regras geradas. O mesmo já não acontece nas árvores de decisão onde a utilização de valores numéricos nos atributos é possível.

Os aspectos até aqui abordados para comparar as técnicas de mineração de dados utilizadas neste trabalho estão apresentados de forma resumida na Tabela 11.

Apesar de todos os aspectos analisados serem importantes, é preciso ter em conta quais são as características mais importantes que o modelo deve ter para que os objectivos de negócio sejam cumpridos. Tendo em conta que se pretende que o modelo final seja capaz de fornecer (de forma simples) informação que permita detectar desvios sistemáticos do horário previsto, é fundamental que este tenha uma boa capacidade de fornecer regras interessantes e de detectar erros sistemáticos. Neste aspecto, as técnicas baseadas em associação têm um desempenho

	Árvores de decisão	**CBA**	**CARs**
Precisão do Classificador	Não tem elevada precisão mas consegue obter melhores resultados do que o CBA.	Razoável, mas inferior às árvores de decisão.	Não se aplica.
Interpretabilidade	Boa. A mais intuitiva de todas as técnicas. A árvore torna-se mais difícil de interpretar com o aumento do tamanho.	Razoável. As regras são fáceis de interpretar mas é difícil ter uma visão global do conjunto, principalmente se o número de regras for elevado.	Razoável. As regras são fáceis de interpretar mas o grande número de regras redundantes perturba a análise do conjunto.
Qualidade/Interesse das regras geradas	Razoável. Como a precisão não é muito elevada existem muito poucas folhas cuja percentagem de instâncias correctamente classificadas é elevada.	Boa. Apresenta um grande número de regras com elevado factor de confiança.	Razoável. De entre todas as técnicas é a que gera maior número de regras com elevado factor de confiança. No entanto, gera muitas regras sem interesse por serem redundantes.
Facilidade de classificação de uma nova instância	Boa. Simples e intuitiva.	Razoável. Se o número de regras for muito elevado pode obrigar a percorrer uma grande parte do classificador para classificar uma dada instância.	Não se aplica.
Capacidade de detectar erros sistemáticos	Razoável.	Boa.	Boa. A melhor de todas as técnicas porque gera o maior número de regras com elevado factor de confiança.
Versatilidade nos tipos de dados	Boa.	Razoável. Não permite a utilização de valores numéricos.	Razoável. Não permite a utilização de valores numéricos.

Tabela 11: Tabela comparativa entre as técnicas de mineração utilizadas

muito melhor do que as árvores de decisão. A grande vantagem de utilizar a mineração de CARs é que esta permite ao utilizador seleccionar apenas as regras mais "interessantes" através da definição dos valores de confiança e suporte adequados. O mesmo já não acontece na construção do classificador em que é necessário utilizar valores de suporte e confiança baixos para que o classificador seja correctamente construído, fazendo assim com que sejam fornecidas regras que podem ter pouco interesse para o utilizador. No entanto, na mineração de CARs também são geradas muitas regras que não trazem qualquer informação útil ao utilizador pois a mesma informação pode ser fornecida por várias regras. Por outro lado, a construção de um classificador associativo, apesar de poder omitir muitas das regras que são geradas na produção de CARs, dá-nos a garantia de que cada instância de treino é coberta pela regra com maior precedência relativamente a todas as regras que poderiam cobrir o caso (Bing Liu, Wynne Hsu et al. 1998). Para além disso, o classificador associativo é composto apenas por um subconjunto de regras que constituem as CARs, o que torna a sua interpretação mais simples. Estes argumentos fazem-nos concluir que a utilização de um classificador associativo poderá ser a solução mais adequada tendo em conta os objectivos de negócio

5.3 Avaliação e Trabalho Futuro

Durante a realização deste trabalho, surgiram diversas situações que dificultaram a sua realização e que condicionaram a forma como este foi conduzido. O facto de a informação utilizada ter sido fornecida em diferentes tabelas que não tinham atributos que identificassem univocamente cada um dos registos, fez com que fosse necessário encontrar um processo de junção das tabelas mais complexo. O facto de este processo envolver a utilização de mais do que um atributo, sendo necessário que os atributos utilizados na junção não tivessem valores nulos, fez com que fosse perdido um grande número de registos que não preenchia as condições da junção. Para além disso, o facto de não existir um atributo que identificasse a hora prevista de passagem em cada paragem obrigou à utilização das diversas tabelas de horários que estiveram em vigor no ano de 2007 para calcular esse valor, complicando ainda mais o processo de construção do conjunto de dados final.

Existem ainda situações em que poderíamos ter optado por outras opções e que poderiam ter melhorado as condições em que as experiências ocorreram e os resultados obtidos. A utilização de outras janelas temporais para a definição dos intervalos que classificam os desvios, particularmente no caso *Pontual*, poderia enriquecer os resultados obtidos, diminuindo o número de casos classificados como pontuais e aumentando as possibilidades de descobrir um maior número de situações interessantes. A utilização de mais do que uma linha para realizar as experiências poderia também ter dado origem a diferentes resultados. Uma vez que linha 205 é

uma linha em que o número de autocarros que circulam por hora é muito elevado, poderia ser interessante efectuar a análise de uma linha em que a frequência dos autocarros é menor. Para além disso, nas experiências efectuadas para mineração de CARs poderia ter-se explorado mais a fase de pós-processamento das regras geradas, tentando assim obter um conjunto de regras mais pequeno mas igualmente rico em regras interessantes. Existem diversos estudos feitos neste sentido que poderiam ter sido utilizados: em (Liu, Hsu et al. 1997), Liu *et al.* propõem um algoritmo que selecciona as regras mais interessantes através da sua comparação com "impressões gerais" especificadas pelo utilizador numa linguagem criada para o efeito; em (Liu and Hsu 1996), os mesmos autores propõem uma técnica que permite ao utilizador comparar as regras geradas com o seu conhecimento e as suas hipóteses; em (Baesens 2000), Baesens *et al.* destacam a necessidade do pós-processamento de regras de associação e fazem uma revisão de alguns métodos que podem ser utilizados para esse objectivo.

Apesar de ainda existirem muitos aspectos que podem ser melhorados, perante os resultados obtidos, podemos concluir que a integração dos modelos estudados numa aplicação informática pode ser útil para as empresas de transportes. Em particular, a utilização de métodos de classificação baseada em associação parece ser a alternativa mais interessante.

Bibliografia

Agrawal, R. a. I., Tomasz and Swami, Arun N. (1993). Mining Association Rules between Sets of Items in Large Databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C.

Baesens, B., Viaene, S. and Vanthienen (2000). Post-processing of association rules. In The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2000), Boston (MA), U.S.A.

Bates, J., J. Polak, et al. (2001). "The valuation of reliability for personal travel." Transportation Research Part E **37**(2-3): 191-229.

Bing Liu, Wynne Hsu, et al. (1998). "Integrating Classification and Association Rule Mining." In Proceedings of KDD'98: 80-86.

Breiman, L., J. H. Friedman, et al. (1984). "Classification and Regression Trees " Wadsworth,Belmont, CA.

Carey, M. (1994). "Reliability of interconnected scheduled services." European Journal of Operational Research **79**(1): 51–72.

Carey, M. (1998). "Optimizing scheduled times, allowing for behavioural response." Transportation Research Part B **32**(5): 329-342.

Ceder, A. (1987). "Methods for creating bus timetables." Transportation research. Part A: general **21**(1): 59-83.

Ceder, A., B. Golany, et al. (2001). "Creating bus timetables with maximal synchronization." Transportation Research Part A **35**(10): 913-928.

Chapman, P., J. Clinton, et al. (2000). CRISP-DM 1.0: Step-by-step data mining guide. CRISP-DM consortium.

Chen, C., A. Skabardonis, et al. (2003). "Travel-Time Reliability as a Measure of Service." Transportation Research Record **1855**: 74-79.

Chen, G., H. Liu, et al. (2006). "A new approach to classification based on association rule mining." Decision Support Systems **42**(2): 674-689.

Chen, M. S., J. Han, et al. (1996). "Data mining: an overview from a database perspective." IEEE Transactions on Knowledge and Data Engineering **8**(6): 866-883.

Craig Silverstein , S. B. a. R. M. (2004). "Beyond Market Baskets: Generalizing Association Rules to Dependence Rules " Data Mining and Knowledge Discovery **2** (1): 39-68.

Ding, Y. and S. I. Chien (2001). "Improving Transit Service Quality and Headway Regularity with Real-Time Control." Transportation Research Record **1760**: 161-170.

El-Geneidy, A., J. Horning, et al. (2007). Using Archived ITS Data to Improve Transit Performance and Management. St. Paul, Minnesota, University of Minnesota.

Fayyad, U., G. Piatetsky-Shapiro, et al. (1996). "Knowledge Discovery and Data Mining: Towards a Unifying Framework." Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR: 82-88.

Fayyad, U. and R. Uthurusamy (1996). "Data mining and knowledge discovery in databases." Communications of the ACM **39**(11): 24-26.

Fayyad, U. M. and K. B. Irani (1993). "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning." International Joint Conference on Artificial Inteligence **13**: 1022-1022.

Feelders, A., H. Daniels, et al. (2000). "Methodological and practical aspects of data mining." Information & Management **37**(5): 271-281.

Floriana Esposito, D. M., Giovanni Semeraro (1997). "A comparative analysis of methods for pruning decision trees " IEEE Transactions on Pattern Analysis and Machine Intelligence **19**: 476-491.

Han, W. L. J. and J. Pei (2001). "CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules." Proc. of IEEE-ICDM: 369-376.

Hany Mahgoub, D. R., Nabil Ismail and Fawzy Torkey (2007). "A Text Mining Technique Using Association Rules Extraction." International Journal of computational intelligence **4**(1): 21-28.

Hu, H. and J. Li (2005). "Using association rules to make rule-based classifiers robust." Proceedings of the 16th Australasian database conference: 47-54.

Hussain, F., H. Liu, et al. (2000). Relative measure for mining interesting rules. Principles of Data Mining and Knowledge Discovery 4th European Conference, Lyon, France.

Itskevitch, J. (2001). Automatic hierarchical e-mail classification using association rules, Simon Fraser University. **Master of Science**.

Janssens, D., G. Wets, et al. (2003). Integrating Classification and Association Rules by proposing adaptations to the CBA Algorithm. Proceedings of the 10th International Conference on Recent Advances in Retailing and Services Science, Portland, Oregon (USA).

LA Breslow, D. A. (1997). "Simplifying decision trees: a survey." Knowledge Engineering Review **12**(1): 1-40.

Landgrebe, S. R. S. a. D. (1991). "A survey of decision tree classifier methodology." IEEE transactions on systems, man, and cybernetics **21**(3): 660-674.

Lawrence O. Hall, R. C., Kevin W. Bowyer and Robert Banfield1 (2002). Error-Based Pruning of Decision Trees Grown on Very Large Data Sets Can Work. International Conference on Tools for Artificial Intelligence.

Liu, B. and W. Hsu (1996). Post-Analysis of Learned Rules. Proc. of the Thirteenth National Conference on Artificial Intelligence (AAAI '96), Portland, Oregon.

Liu, B., W. Hsu, et al. (1997). Using general impressions to analyze discovered classification rules. 3rd Int. Conf. Knowledge Discovery & Data Mining, 1997 Newport Beach, California.

Liu, R. and S. Sinha (2007). Modelling urban bus service and passenger reliability. Proc. of the International Symposium on Transportation Network Reliability, The Hague, Netherlands.

Maimon, L. R. a. O. (2007). Data Mining with Decision Trees - Theory and Applications, World Scientific Publishing Company.

McGarry, K. E. N. (2005). "A survey of interestingness measures for knowledge discovery." The Knowledge Engineering Review **20**(1): 39-61.

Michael, J. A. B. and S. L. Gordon (2004). Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, John Wiley & Sons.

Mingers, J. (1989). "An Empirical Comparison of Pruning Methods for Decision Tree Induction " Machine Learning **4**(2): 227-243.

Murthy, S. K. (1998). "Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey." Data Mining and Knowledge Discovery **2** (4): 345-389.

Mutter, S., M. Hall, et al. (2004). "Using classification to evaluate the output of confidence-based association rule mining." Australian Conference on Artificial Intelligence, Cairns, Australia, Springer: 538–549.

Niblett, T. and I. Bratko (1987). Learning decision rules in noisy domains. Proceedings of Expert Systems' 86, The 6Th Annual Technical Conference on Research and development in expert systems III, Cambridge University Press.

Palma, A. and R. Lindsey (2001). "Optimal timetables for public transportation." Transportation Research Part B **35**(8): 789-813.

Quinlan, J. R. (1986). "Induction of Decision Trees." Machine Learning **1**(1): 81-106.

Quinlan, J. R. (1987). "Simplifying decision trees." Int. J. Man-Mach. Stud. **27**(3): 221-234.

Quinlan, J. R. (1993). C4.5: programs for machine learning, Morgan Kaufmann Publishers Inc.

Rakesh Agrawal and R. Srikant (1994). Fast Algorithms for Mining Association Rules. Proc. 20th Int. Conf. Very Large Data Bases, Santiago, Chile.

Rietveld, P., F. R. Bruinsma, et al. (2001). "Coping with unreliability in public transport chains: A case study for Netherlands." Transportation Research Part A **35**(6): 539-559.

S.Hanash, C. C. a. (2003). "Mining gene expression databases for association rules " Bioinformatics **19**(1): 79–86.

Schaffer, C. (1993). "Overfitting Avoidance as Bias." Machine Learning **10**(2): 153-178.

Strathman, J. G., K. J. Dueker, et al. (1999). "Automated Bus Dispatching, Operations Control, and Service Reliability: Baseline Analysis." Transportation Research Record **1666**: 28-36.

Strathman, J. G. and J. R. Hopper (1993). "Empirical analysis of bus transit on-time performance." Transportation research. Part A, Policy and practice **27**(2): 93-100.

Strathman, J. G., T. J. Kimpel, et al. (2002). "Evaluation of transit operations: data applications of Tri-Met's automated Bus Dispatching System." Transportation **29**(3): 321-345.

Thabtah, F., P. Cowling, et al. (2006). "Improving rule sorting, predictive accuracy and training time in associative classification." Expert Systems with Applications **31**(2): 414-426.

Thabtah, F., P. Cowling, et al. (2004). MCLA: Multi-label Classification Learning Algorithm. ACIT' 2004. Mentouri University of Constantine, Algeria.

Wang, K., S. Zhou, et al. (2000). Growing decision trees on support-less association rules, ACM New York, NY, USA.

Witten, I. H. and E. Frank (2005). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco, Morgan Kaufmann.

Yin, X. and J. Han (2003). CPAR: Classification based on Predictive Association Rules. Proceedings of the Third SIAM International Conference on Data Mining San Francisco, California (USA).

Zimmermann, A. and L. De Raedt (2004). "CorClass: Correlated Association Rule Mining for Classification." Lecture Notes in Computer Science **3245**: 60-72.

Anexos

I. Classificador construído utilizando a técnica CBA

O classificador apresentado na foi gerado a partir do conjunto Out_tardes com 0,5% de suporte e 30% de confiança.

1	dia_da_semana = DOMINGO, sentido = IDA, id_paragem = RAEP -> class = Muito_Adiantado (confiança: 91.667% , suporte: 0.598%)
2	dia_da_semana = QUARTA-FEIRA, sentido = IDA, id_paragem = CMP -> class = Pontual (confiança: 90.244% , suporte: 1.207%)
3	tipo_de_dia = DOMINGO/FERIADO, sentido = IDA, id_paragem = CMP -> class = Pontual (confiança:90.000% , suporte: 0.783%)
4	hora = Hora_Normal_da_Tarde, sentido = IDA, id_paragem = CMP -> class = Pontual (confiança: 88.532% , suporte: 2.099%)
5	hora = Tarde, sentido = IDA, id_paragem = CMP -> class = Pontual (confiança: 88.356% , suporte: 1.403%)
6	tipo_de_dia = DOMINGO/FERIADO, sentido = IDA, id_paragem = RAEP -> class = Muito_Adiantado (confiança: 85.897% , suporte: 0.729%)
7	dia_da_semana = DOMINGO, sentido = VOLTA, id_paragem = CMP -> class = Muito_Adiantado (confiança: 85.714% , suporte: 0.522%)
8	dia_da_semana = QUARTA-FEIRA, sentido = VOLTA, id_paragem = RAEP -> class = Pontual (confiança: 85.246% , suporte: 1.131%)
9	tipo_de_dia = DOMINGO/FERIADO, sentido = VOLTA, id_paragem = CMP -> class = Muito_Adiantado (confiança: 84.722% , suporte: 0.663%)
10	sentido = IDA, id_paragem = CMP -> class = Pontual (confiança: 84.225% , suporte: 6.852%)
11	hora = Hora_Normal_da_Tarde, dia_da_semana = SEGUNDA-FEIRA, sentido = IDA, id_paragem = SR -> class = Adiantado (confiança: 83.636% , suporte: 0.500%)
12	hora = Ponta_da_Tarde, dia_da_semana = QUARTA-FEIRA, sentido = VOLTA, id_paragem = CQ -> class = Pontual (confiança: 82.432% , suporte: 0.663%)

13	dia_da_semana = QUARTA-FEIRA, sentido = VOLTA, id_paragem = CQ -> class = Pontual (confiança: 80.769% , suporte: 1.142%)
14	hora = Ponta_da_Tarde, dia_da_semana = SEGUNDA-FEIRA, sentido = VOLTA, id_paragem = CQ -> class = Pontual (confiança: 80.000% , suporte: 0.827%)
15	dia_da_semana = SEGUNDA-FEIRA, sentido = VOLTA, id_paragem = CQ -> class = Pontual (confiança: 78.528% , suporte: 1.392%)
16	hora = Hora_Normal_da_Tarde, sentido = VOLTA, id_paragem = RAEP -> class = Pontual (confiança: 77.725% , suporte: 1.784%)
17	hora = Hora_Normal_da_Tarde, sentido = IDA, id_paragem = SR -> class = Adiantado (confiança: 77.626% , suporte: 1.849%)
18	dia_da_semana = SEGUNDA-FEIRA, sentido = VOLTA, id_paragem = RAEP -> class = Pontual (confiança: 77.333% , suporte: 1.262%)
19	hora = Ponta_da_Tarde, sentido = VOLTA, id_paragem = CQ -> class = Pontual (confiança: 76.839% , suporte: 3.067%)
20	hora = Ponta_da_Tarde, dia_da_semana = QUARTA-FEIRA, sentido = IDA, id_paragem = SR -> class = Adiantado (confiança: 75.342% , suporte: 0.598%)
21	tipo_de_dia = DIA_UTIL, sentido = VOLTA, id_paragem = CQ -> class = Pontual (7.113% 74.618% 654 488 5.308%)
22	sentido = VOLTA, id_paragem = CQ -> class = Pontual (8.832% 74.261% 812 603 6.559%)
23	hora = Hora_Normal_da_Tarde, dia_da_semana = QUARTA-FEIRA, id_paragem = CMP -> class = Pontual (0.968% 73.034% 89 65 0.707%)
24	tipo_de_dia = DIA_UTIL, sentido = VOLTA, id_paragem = RAEP -> class = Pontual (6.646% 72.831% 611 445 4.840%)
25	hora = Ponta_da_Tarde, dia_da_semana = QUARTA-FEIRA, id_paragem = CMP -> class = Pontual (1.621% 71.812% 149 107 1.164%)

26	dia_da_semana = SEGUNDA-FEIRA, sentido = IDA, id_paragem = SR -> class = Adiantado (1.631% 70.000% 150 105 1.142%)
27	hora = Ponta_da_Tarde, dia_da_semana = SEXTA-FEIRA, id_paragem = CQ -> class = Pontual (0.979% 70.000% 90 63 0.685%)
28	tipo_de_dia = DOMINGO/FERIADO, sentido = IDA, id_paragem = SR -> class = Adiantado (0.914% 69.048% 84 58 0.631%)
29	tipo_de_dia = DOMINGO/FERIADO, sentido = IDA, id_paragem = CQ -> class = Muito_Adiantado (0.859% 68.354% 79 54 0.587%)
30	hora = Hora_Normal_da_Tarde, dia_da_semana = SEXTA-FEIRA, sentido = VOLTA -> class = Pontual (2.186% 68.159% 201 137 1.490%)
31	hora = Ponta_da_Tarde, dia_da_semana = QUARTA-FEIRA, sentido = VOLTA -> class = Pontual (4.851% 67.489% 446 301 3.274%)
32	DIA_ANO = (-inf-274_5], hora = Ponta_da_Tarde, sentido = VOLTA -> class = Pontual (1.262% 67.241% 116 78 0.848%)
33	hora = Hora_Normal_da_Tarde, id_paragem = CMP -> class = Pontual (4.818% 66.817% 443 296 3.219%)
34	sentido = VOLTA, id_paragem = RAEP -> class = Pontual (8.201% 65.915% 754 497 5.406%)
35	dia_da_semana = QUINTA-FEIRA, sentido = IDA, id_paragem = SR -> class = Adiantado (1.240% 65.789% 114 75 0.816%)
36	dia_da_semana = SEXTA-FEIRA, tipo_de_dia = DIA_UTIL, id_paragem = CQ -> class = Pontual (1.849% 65.294% 170 111 1.207%)
37	hora = Ponta_da_Tarde, dia_da_semana=SEGUNDA-FEIRA,sentido=VOLTA, id_paragem=HSJ -> class = Pontual (1.001% 65.217% 92 60 0.653%)
38	dia_da_semana = QUARTA-FEIRA, sentido = VOLTA -> class = Pontual (8.625% 64.943% 793 515 5.601%)

39	tipo_de_dia = SABADO, id_paragem = CMP -> class = Pontual (1.490% 64.964% 137 89 0.968%)
40	hora = Tarde, sentido = IDA, id_paragem = RAEP -> class = Muito_Adiantado (1.610% 64.865% 148 96 1.044%)
41	tipo_de_dia = DIA_UTIL, sentido = IDA, id_paragem = SR -> class = Adiantado (6.689% 64.065% 615 394 4.285%)
42	hora = Hora_Normal_da_Tarde, dia_da_semana = QUINTA-FEIRA, id_paragem = CQ -> class = Pontual (0.903% 63.855% 83 53 0.576%)
43	hora = Hora_Normal_da_Tarde, sentido = VOLTA, id_paragem = AML -> class = Pontual (2.817% 63.707% 259 165 1.795%)
44	hora = Ponta_da_Tarde, dia_da_semana = SEGUNDA-FEIRA, id_paragem = CQ -> class = Pontual (1.762% 63.580% 162 103 1.120%)
45	sentido = IDA, id_paragem = SR -> class = Adiantado (8.375% 63.117% 770 486 5.286%)
46	DIA_ANO = (-inf-274_5], sentido = VOLTA -> class = Pontual (2.186% 63.184% 201 127 1.381%)
47	hora = Hora_Normal_da_Tarde, dia_da_semana = QUINTA-FEIRA, sentido = VOLTA -> class = Pontual (2.752% 62.846% 253 159 1.729%)
48	hora = Ponta_da_Tarde, id_paragem = CMP -> class = Pontual (7.581% 62.697% 697 437 4.753%)
49	hora = Hora_Normal_da_Tarde, dia_da_semana = SEGUNDA-FEIRA, sentido = VOLTA -> class = Pontual (3.687% 61.947% 339 210 2.284%)
50	hora = Ponta_da_Tarde, dia_da_semana = SEGUNDA-FEIRA, sentido = VOLTA -> class = Pontual (6.015% 61.302% 553 339 3.687%)

51	hora = Ponta_da_Tarde, id_paragem = CQ -> class = Pontual (6.994% 61.275% 643 394 4.285%)
52	dia_da_semana = QUARTA-FEIRA, id_paragem = CQ -> class = Pontual (2.708% 61.044% 249 152 1.653%)
53	dia_da_semana = DOMINGO, id_paragem = SR -> class = Adiantado (1.327% 60.656% 122 74 0.805%)
54	tipo_de_dia = DOMINGO/FERIADO, sentido = IDA, id_paragem = AML -> class = Muito_Adiantado (0.903% 60.241% 83 50 0.544%)
55	hora = Hora_Normal_da_Tarde, sentido = VOLTA -> class = Pontual (15.249% 60.128% 1402 843 9.169%)
56	tipo_de_dia = SABADO, id_paragem = CQ -> class = Pontual (1.555% 60.140% 143 86 0.935%)
57	dia_da_semana = QUINTA-FEIRA, id_paragem = CQ -> class = Pontual (2.404% 59.729% 221 132 1.436%)
58	tipo_de_dia = DOMINGO/FERIADO, id_paragem = SR -> class = Adiantado (1.729% 59.119% 159 94 1.022%)
59	DIA_ANO = (-inf-274_5], hora = Ponta_da_Tarde -> class = Pontual (2.197% 58.911% 202 119 1.294%)
60	tipo_de_dia = DIA_UTIL, sentido = VOLTA, id_paragem = AML -> class = Pontual (7.342% 58.815% 675 397 4.318%)
61	hora = Nocturno, id_paragem = AML -> class = Pontual (1.077% 58.586% 99 58 0.631%)
62	hora = Ponta_da_Tarde, dia_da_semana = SEGUNDA-FEIRA, id_paragem = HSJ -> class = Pontual (1.751% 57.764% 161 93 1.012%)
63	sentido = VOLTA, id_paragem = AML -> class = Pontual (8.821% 57.460% 811 466 5.069%)

64	tipo_de_dia = DIA_UTIL, sentido = VOLTA -> class = Pontual (42.974% 57.302% 3951 2264 24.625%)
65	dia_da_semana = DOMINGO, id_paragem = HSJ -> class = Adiantado (1.273% 57.265% 117 67 0.729%)
66	tipo_de_dia = DOMINGO/FERIADO, id_paragem = HSJ -> class = Adiantado (1.642% 56.954% 151 86 0.935%)
67	hora = Ponta_da_Tarde, dia_da_semana = QUARTA-FEIRA -> class = Pontual (9.115% 55.489% 838 465 5.058%)
68	hora = Ponta_da_Tarde, dia_da_semana = SEXTA-FEIRA, sentido = IDA -> class = Pontual (2.556% 55.319% 235 130 1.414%)
69	sentido = VOLTA -> class = Pontual (52.284% 54.941% 4807 2641 28.725%)
70	DIA_ANO = (-inf-274_5] -> class = Pontual (4.079% 54.933% 375 206 2.241%)
71	tipo_de_dia = SABADO, id_paragem = HSJ -> class = Pontual (1.512% 54.676% 139 76 0.827%)
72	hora = Nocturno, id_paragem = HSJ -> class = Pontual (1.077% 54.545% 99 54 0.587%)
73	dia_da_semana = SEGUNDA-FEIRA, id_paragem = HSJ -> class = Pontual (3.307% 52.303% 304 159 1.729%)
74	hora = Ponta_da_Tarde -> class = Pontual (44.007% 51.656% 4046 2090 22.732%)
75	hora = Hora_Normal_da_Tarde, sentido = IDA, id_paragem = HSJ -> class = Adiantado (2.317% 51.643% 213 110 1.196%)

Tabela 12: Classificador construído utilizando o algoritmo CBA, a partir do conjunto Out_tardes com 0,5% de suporte e 30% de confiança