Exploração de Técnicas de Classificação Associativa no Planeamento de Horários de Transportes Públicos

Eva Duarte[1], João Mendes Moreira[2,3], Orlando Belo[4].

1) Departamento de Informática, Escola de Engenharia, Universidade do Minho, Braga, Portugal
pg11964@alunos.uminho.pt
2) Departamento de Engenharia Informática, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal
jmoreira@fe.up.pt
3) LIAAD-INESC Porto L.A., Portugal
4) Departamento de Informática, Escola de Engenharia, Universidade do Minho, Braga, Portugal
obelo@di.uminho.pt

Resumo

Nos dias de hoje, a fiabilidade dos serviços dos sistemas de transportes públicos de passageiros é uma das nossas maiores preocupações, tanto ao nível dos passageiros como das próprias empresas que fornecem esses serviços. Os avanços tecnológicos ocorridos nas últimas décadas permitiram que as empresas de transportes públicos fossem colectando e armazenando enormes quantidades de informação acerca das viagens realizadas em repositórios de dados especializados. Agora, com base nessa informação conseguem identificar eventuais erros no planeamento dos seus serviços e identificarem padrões de comportamento que podem ser utilizados na melhoria dos seus serviços a médio e a longo prazo. Este trabalho, planeado e desenvolvido tendo como alvo de estudo a empresa de transportes públicos de passageiros STCP, pretendia melhorar o desempenho do seu sistema no que toca ao cumprimento dos horários. Pretendia-se estudar a viabilidade de aplicar o algoritmo CBA (*Classification Based in Association*) para detectar desvios sistemáticos do horário previsto. Com base nos resultados obtidos, pudemos concluir que este algoritmo permite atingir tais objectivos, sendo viável a sua integração numa aplicação informática com esse fim.

Palavras chave: Transportes públicos, Planeamento de Linhas, Mineração de dados, Classificação baseada em Associação.

1. Introdução

Desde a sua origem, 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). Tempos de espera demasiado elevados, chegadas atrasadas ou adiantadas aos destinos, e ligações perdidas podem induzir nos passageiros alguns 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). No caso particular das empresas de transportes rodoviários, a pontualidade das chegadas e das partidas em cada paragem é muitas vezes difícil de manter, uma vez que existem vários factores que podem afectar o cumprimento dos horários previstos, nomeadamente: os níveis de congestionamento diário, a localização das paragens, o número de passageiros, planeamento incorrecto dos horários, entre outros (Liu and Sinha 2007). A análise das viagens (tanto na hora em que estas se realizam como posteriormente) é uma tarefa que pode ser realizada no sentido de aumentar a fiabilidade dos transportes, pois 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 no fim de uma viagem, mas também em cada um dos pontos de horário, pode conduzir a consideráveis melhoramentos na qualidade do serviço prestado. 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 releva 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 próprios passageiros. No entanto, a evolução de algumas tecnologias como o GPS permitiram às empresas de transportes públicos criarem sistemas de controlo sobre os seus veículos. Isto permite fazer 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. Este trabalho tem como base a informação proveniente de uma empresa de transportes públicos - STCP, SA (Sociedade de Transportes Colectivos do Porto, SA) - 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 pretendia detectar situações sistemáticas em que o horário previsto não era cumprido e identificar as condições em que estas ocorrem. Para isso, foi estudada a viabilização da aplicação de uma técnica de mineração de dados, a *Classificação Baseada em Associação*, no sentido de atingir este objectivo.

2. Planeamento e Gestão de Horários de uma Linha de Transportes

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, 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. No entanto, os horários muitas vezes não são cumpridos, provocando descontentamento nos clientes e podendo causar prejuízos à empresa. 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.

Na maioria dos casos, 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 a 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 faz a ligação entre a parte oriental e a ocidental da cidade. 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, existindo 6 pontos de horário.

3. Planeamento de Horários baseado em Classificação Associativa

A descoberta de regras de classificação e a descoberta de regras de associação são duas técnicas de mineração de dados bastante utilizadas. Basicamente, a primeira procura descobrir um pequeno conjunto de regras na base de dados para formar um classificador preciso (Quinlan 1986), enquanto que 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 uma técnica denominada de *classificação associativa* e actua segundo duas vertentes: a descoberta de regras de associação para classificação (CARs) e a construção de um classificador a partir dessas regras. As CARs (do inglês *class association rules*) são regras de associação cujo lado direito da implicação é restrito ao atributo de classificação. 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. A segunda fase do algoritmo CBA 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 omissão*. 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. No contexto do actual problema, este tipo de técnica revela-se particularmente interessante pois reúne as vantagens da associação e da classificação: por um lado, a descoberta de CARs permite-nos encontrar relações interessantes entre as diferentes variáveis, tentando criar regras que expliquem os desvios do horário previsto; por outro lado, a construção de um classificador a partir dessas regras faz com que o conjunto de regras seja reduzido àquelas que têm mais relevância, obtendo-se ao mesmo tempo um classificador que descreve o conjunto de dados na sua totalidade.

4. Preparação e Enriquecimento dos Dados

Para a realização deste trabalho foi utilizada informação proveniente do *data warehouse* (DW) da STCP e ainda alguma outra informação proveniente do Departamento de Operações, correspondente aos horários previstos. Antes de aplicar o algoritmo CBA, os dados passaram por várias fases de pré-processamento que não serão explicadas em detalhe neste documento.

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 da STCP, seria o mais adequado para as análises que se pretendem efectuar, tendo em conta a viabilidade da sua utilização numa aplicação futura. 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 de efectuada uma breve análise dos dados disponíveis, optou-se por não os incluir, pois a sua informação era pouco fiável e poderia, eventualmente, perturbar os resultados de forma negativa.

Para classificar os desvios relativamente ao horário previsto foi criado o atributo nominal *Tipo de Desvio*. 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. 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 1 representa o desvio (em minutos) associado a cada percentil.

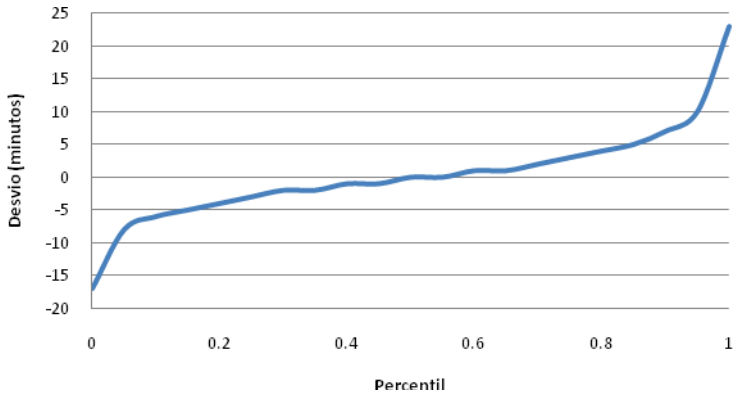

Figura 1: Desvio (em minutos) associado ao Percentil

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 2 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.

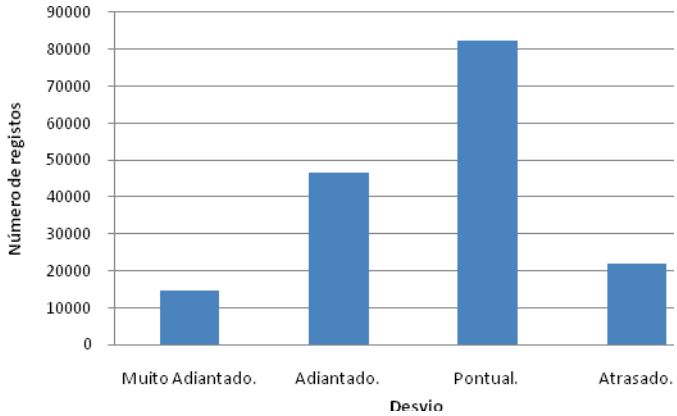

Figura 2: Número de instâncias classificadas por cada um dos tipos de desvio

No final da fase de pré-processamento, obtivemos um conjunto de dados constituído por 165451 instâncias. Estes dados correspondem a viagens realizadas entre 1 de Janeiro de 2007 e 31 de Outubro de 2007.

Cada entrada da tabela representa uma passagem (ou paragem) do autocarro num ponto de horário, sendo que o desvio relativamente ao horário previsto é calculado com base no momento em que o GPS detecta pela primeira vez o autocarro. Os atributos que o compõem e os valores que podem assumir estão apresentados na Tabela 1. 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 utilizados são os que se apresentam na Tabela 2.

5. Construção de um Classificador Associativo

5.1 Desenho dos testes

Para efectuar a construção dos modelos foi utilizada a ferramenta informática *CBA* (Bing Liu, Wynne Hsu et al. 1998). Esta 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.

Id. 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 1: Estrutura do conjunto de dados resultante

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*. 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%.

Nome	Descrição	Nº 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 2: Conjuntos de dados utilizados nas experiências de associação

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, procedendo à transformação do atributo Dia do Ano num atributo discreto segundo o método proposto por Fayyad *et al* em (Fayyad and Irani 1993).

5.2 Resultados Obtidos

Os resultados dos testes efectuados relativos à precisão e ao número de regras geradas pelo modelo estão representados nas Figuras 3, 5, 6 e 7.

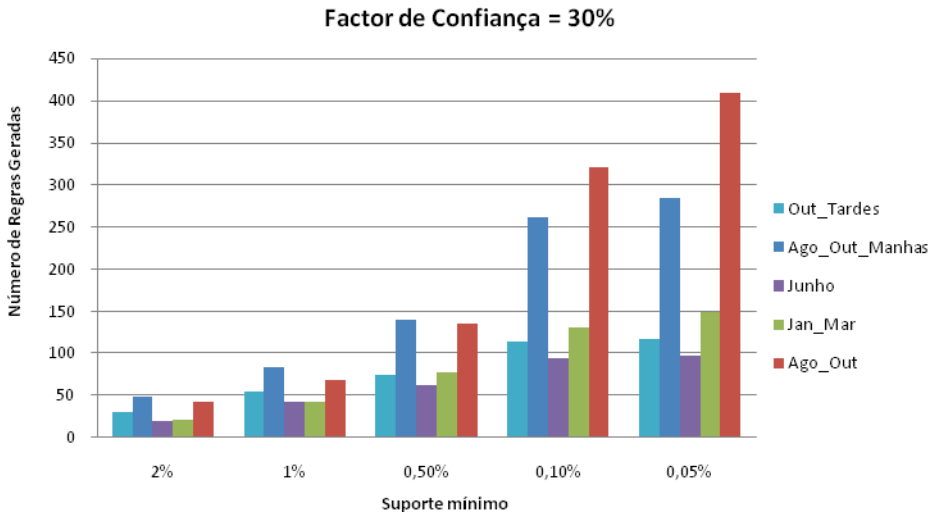

Figura 3: Número de regras do classificador obtido com CF = 30%, para cada um dos conjuntos de teste

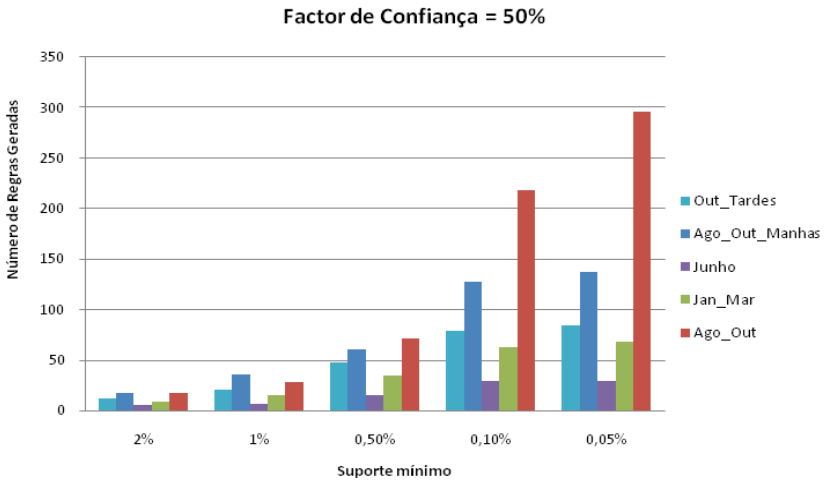

Figura 4: Número de regras do classificador obtido com CF= 50%, para cada um dos conjuntos de teste

Nos gráficos das Figuras Figura 3 e Figura 4 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 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 4). 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 apresentam a percentagem de instâncias correctamente classificadas para cada um dos factores de confiança testados.

Nos gráficos figuras Figura 5 e Figura 6 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.

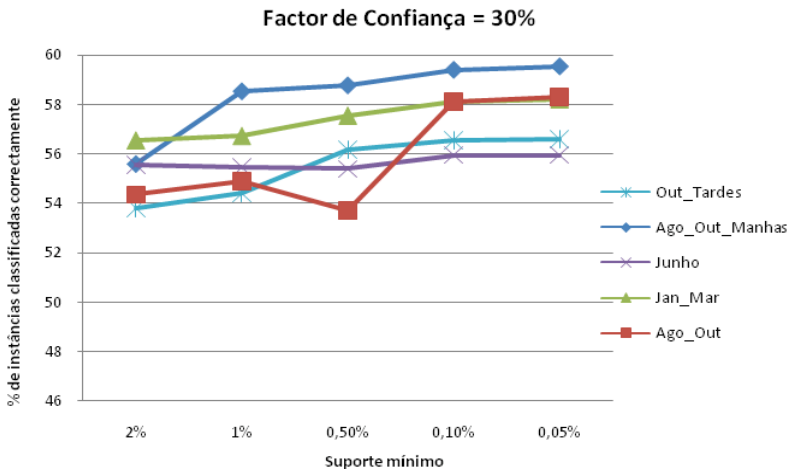

Figura 5: Percentagem de instâncias classificadas correctamente pelo classificador obtido com CF = 30%, para cada um dos conjuntos de teste

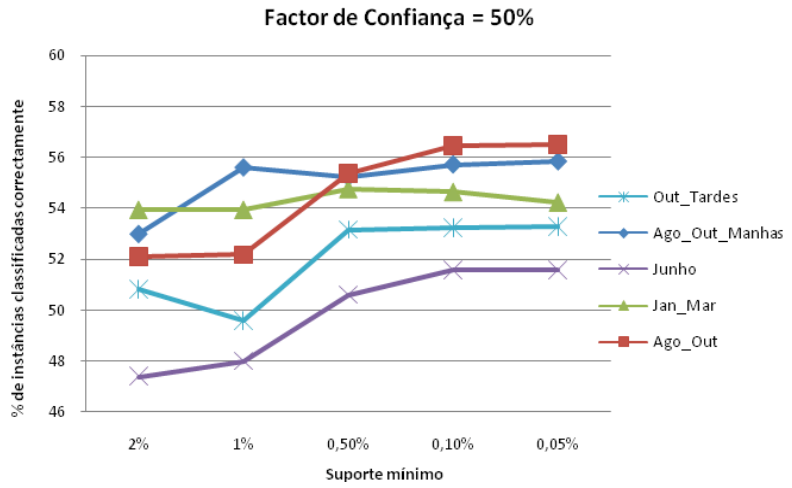

Figura 6: Percentagem de instâncias classificadas correctamente pelo classificador obtido com CF = 50%, para cada um dos conjuntos de teste

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 dos modelos (Figura 3, Figura 4, Figura 5 e Figura 6), 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 algumas regras do classificador obtido a partir do conjunto Out_tardes, com 0,5% de suporte e 30% de confiança. O classificador é composto na totalidade por 75 regras.

```
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%)
(...)
17  hora = Hora_Normal_da_Tarde, sentido = IDA, id_paragem = SR
 ->       class = Adiantado
(confiança: 77.626% , suporte: 1.849%
(...)
```

Figura 7: Excerto do classificador construído a partir do conjunto Out_tardes com 0,5% de suporte e 30% de confiança

Como se pode observar na Figura 7, as regras obtidas descrevem situações que ocorrem de forma sistemática e que podem conduzir ou não a desvios do horário previsto. Para o caso abordado neste estudo, as regras cujo desvio seja diferente de *Pontual* têm maior interesse, principalmente se tiverem factores de confiança elevados, como é o caso das regras 1 e 17. Isto porque estas regras descrevem situações em que os horários não são cumpridos com uma frequência muito elevada (91,667% e 77,626%), situação que pode ainda não ter sido detectada pelos planeadores e que pode conduzir a um reajustamento do horário.

5.3 Discussão dos Resultados

Através dos resultados obtidos, 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, 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 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 precisem ou não devam 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* num dado ponto de horário poderá ser alvo de atenção por ser um local onde entram e saem muitos passageiros podendo fazer com que o autocarro parta atrasado e não seja capaz de recuperar esse atraso.

6. Conclusões e Trabalho futuro

Como já foi referido, o principal objectivo deste trabalho era estudar a viabilidade do algoritmo CBA fornecer modelos de classificação que permitissem detectar erros sistemáticos do horário previsto em sistemas de transportes públicos. Depois de realizadas experiências com diferentes parâmetros e conjuntos de dados, concluiu-se que este modelo tem capacidade de devolver regras interessantes no sentido de se efectuarem reajustamentos nos horários.

Todavia, existem situações em que se poderia ter optado por outras opções e que poderiam ter melhorado os resultados obtidos, podendo agora servir como trabalho futuro. 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 a linha em

análise 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, podia ter-se optado por utilizar algoritmos de classificação baseada em associação mais recentes, como o CMAR (Han and Pei 2001), CPAR (Yin and Han 2003), CorClass (Zimmermann and De Raedt 2004) e GARC (Chen, Liu et al. 2006) ou novas versões do CBA com alguns melhoramentos (Janssens, Wets et al. 2003, Liu, 2000 #101).

Apesar de ainda existirem muitos aspectos que podem ser melhorados, perante os resultados obtidos, podemos concluir que a integração de modelos de classificação baseada em associação numa aplicação informática de análise de dados, aplicada aos transportes rodoviários, pode ser uma ferramenta muito útil para melhorar o desempenho destes sistemas.

7. Referências

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.

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

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.

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.

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.

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

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).

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.

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.

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

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.

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.

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.

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.