Finding Interesting Contexts for Explaining Deviations in Bus Trip Duration Using Distribution Rules

Alípio M. Jorge[1], João Mendes-Moreira[2], Jorge Freire de Sousa[3], Carlos Soares[4], and Paulo J. Azevedo[5]

[1] LIAAD-INESC TEC DCC-FCUP, Universidade do Porto
[2] LIAAD-INESC TEC, DEI-FEUP, Universidade do Porto
[3] UGEI-INESC TEC, DEGI-FEUP, Universidade do Porto
[4] INESC TEC, FEP, Universidade do Porto
[5] Haslab-INESC TEC, Universidade do Minho

Abstract. In this paper we study the deviation of bus trip duration and its causes. Deviations are obtained by comparing scheduled times against actual trip duration and are either delays or early arrivals. We use distribution rules, a kind of association rules that may have continuous distributions on the consequent. Distribution rules allow the systematic identification of particular conditions, which we call contexts, under which the distribution of trip time deviations differs significantly from the overall deviation distribution. After identifying specific causes of delay the bus company operational managers can make adjustments to the timetables increasing punctuality without disrupting the service.

Keywords: Bus trip duration deviations, distribution rules.

1 Introduction

In the last two/three decades, passenger transport companies have made important investments in information systems, such as Automatic Vehicle Location (AVL), automatic passenger counting, automated ticketing and payment, multimodal traveler information systems, operational planning and control software and data warehouse technology, among others. As a consequence of this effort in Advanced Public Transportation Systems, passenger transport companies have been able to collect massive amounts of data. However, as in other areas of activity, the data collected are not being used as much as they could be in supporting public transport companies to accomplish their mission, despite the potential of both data and existing knowledge.

The planning of public transport companies is a complex task. It has as major goal: the achievement of the adequate offer of trips using the resources at a minimal cost. The two main resources are drivers and buses. The uncertainty of trip duration [14] must be taken into account in the definition of schedules, in order to obtain an adequate offer of trips at minimal costs. The problem is that

J. Hollmén, F. Klawonn, and A. Tucker (Eds.): IDA 2012, LNCS 7619, pp. 139–149, 2012.

it is not known how much uncertainty should be assumed because it is difficult to evaluate the trade-off between the operational costs (due to the amount of resources used) and client satisfaction. Additionally, the metrics used to measure client satisfaction vary according to the type of routes. For example, in highly frequent urban routes with a five minutes headway (where headway is the time gap between two consecutive vehicles) client satisfaction assessment is based on the stability of the headway. On low frequent suburban routes (e.g. 60 minutes headway) a metric based on the deviation from departure time is preferred.

In this paper we present a study on how to take advantage of stored AVL data in order to promote adjustments to existing timetables. For that we use the data mining technique of distribution rules [9]. We intend to detect systematic deviations between the actual and the scheduled trip duration and identify the causes of such deviations. Such tool can be integrated in a decision support tool for timetable adjustments [13].

We start by describing distribution rules, the data mining technique that we use to study trip duration deviation. We then present the datasets used and their preparation. Next, we provide details and results of the data mining step. Results are discussed both from a data mining and an operational point of view.

2 Distribution Rules

Distribution rules (DR) [9] are constructs similar to association rules (AR), but having a distribution of an attribute of interest A on the consequent. Whereas the antecedent of a DR is similar to the one of an AR, its consequent is a distribution of A under the conditions stated in the antecedent. In the DR setting, all the antecedents Ant which correspond to an interesting distribution $D_{A|Ant}$ (read as "the distribution of A given Ant") are found. In this case, interesting means that the distribution of A under Ant is significantly different from the distribution of A without any constraints (*a priori*).

Definition 1. *A distribution rule (DR) is a rule of the form* $Ant \rightarrow A = D_{A|Ant}$*, where Ant is a set of items as in a classical association rule, A is a property of interest (the target attribute), and* $D_{A|Ant}$ *is an empirical distribution of A for the cases where Ant is observed. This attribute A can be numerical or categorical.* $D_{A|Ant}$ *is a set of pairs* $A_j/freq(A_j)$ *where* A_j *is one particular value of A occurring in the sample and* $freq(A_j)$ *is the frequency of* A_j *for the cases where Ant is observed.*

In Figure 1 we can see one distribution rule derived from "Auto MPG", a data set with descriptions of cars where the property of interest (P.O.I.) is their fuel consumption in miles per galon (MPG) [6]. The antecedent is shown above the chart. The darker curve shows the density of $MPG|Ant$, the grey curve shows the density of MPG overall. We can also see some measures characterizing the rule: KS-interest, a measure of interest of the rule given by $1 - p_{KS}$, where p_{KS} is the *p-value* of the Kolmogorov Smirnov test; the support (Sup) of the

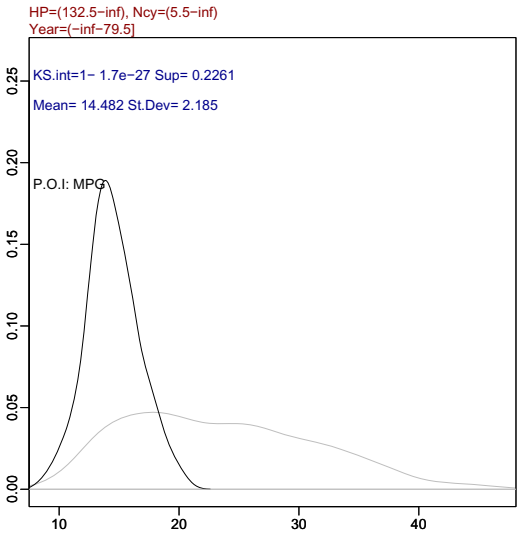

Fig. 1. Distribution rule for the "Auto MPG" data set

antecedent; the mean of $MPG|Ant$ and its standard deviation. The represented densities are estimated using kernel density estimation [8].

Given a dataset, the task of distribution rule discovery consists in finding all the DR $Ant \rightarrow A = D_{A|Ant}$, where Ant has a support above a determined minimum σ_{min} and $D_{A|Ant}$ is statistically significantly different (w.r.t. a predefined level of significance) from the default distribution $D_{A|\emptyset}$. The default distribution is the one obtained with all the values of A for the whole dataset or a distribution obtained from a holdout data set. To compare the distributions we use *Kolmogorov-Smirnov* (KS) [4], a statistical goodness of fit test. The value of the KS statistic is calculated by maximizing $|F_s(x) - F(x)|$, where $F(x)$ is the empirical cumulative distribution function for the whole domain of A and $F_s(x)$ is the cumulative distribution function for the cases covered by Ant.

3 Trip Time Deviation

The ultimate aim of this work is to improve the quality of urban bus service. One important aspect is adjusting schedules to operational conditions and vice versa. For that, we need to know what are the factors, or combination of factors, that are associated with deviations in trip duration. That can be done using descriptive data mining techniques [7]. The extracted patterns can then be used to help operational managers taking measures. In this paper we use the case of urban line 205 of STCP, the main Oporto bus operator.

Our property of interest is "trip duration deviation", named *Deviation* in this paper. This is defined as the difference, in seconds, between the time the bus actually takes from departure point to destination and the published scheduled

duration. Its a priori distribution corresponds to the whole population. Our operational problem of finding relevant factors of deviation will be translated, as a data mining problem, into discovering contexts that are associated with statistically significant changes in the distribution of the variable *Deviation*. Below, we formally define the notion of context.

Definition 2. *A context is a combination of conditions that can be observed at a given moment and influences operation. In logical terms, and in this work, a context is a conjunction of logical literals* $Cond_1 \wedge \ldots \wedge Cond_n$.

The descriptive data mining technique of distribution rule discovery presented in section 2 is able to find relevant contexts, given one property of interest. This discovery problem is also related to the problem of subgroup discovery [10][11]. What we do is: given a dataset with the description of trips (operational conditions and deviation), we obtain all interesting distribution rules. Each rule covers a subgroup of trips and associates one particular context with one particular distribution of the variable Deviation which is sufficiently different from its a priori distribution. To discover the rules we use the program Caren [1].

4 Data Preparation

From the data collected by the company, we analyse two different periods. The first period spans from January to September 2007 (dataset1 with 14169 records) and the second from November 2007 to March 2008 (dataset2 with 9203 records). The attributes of the datasets are described in Table 1. Each dataset line describes one bus trip from departure to destination. Along with static descriptors we have the actual time taken in that journey. From the published schedule we obtain the scheduled duration. The difference between actual duration and scheduled duration gives us the deviation.

We start by analyzing trip duration. The two boxplots in Figure 2 show that we have a very high number of outliers corresponding to extreme durations, mostly for the first period. However, our experience indicates that most of these outliers are caused by operational errors. In particular, bus drivers must press a button at the end of each trip so that arrival time is recorded. However, this operation may fail for different reasons. As a consequence, the recorded arrival time will be the next arrival time (or the one after that) which multiplies trip duration. Simple observation suggests a cut above 6000, for dataset1, and a cut above 5000 for dataset2. With this data cleaning operation we reduce the possibility of artificial deviations.

The attribute StartTime, originally in seconds elapsed on that day, has been discretized to 24 values ($\{0, 1, \ldots, 23\}$, where the value of k represents the interval between hour k and $k + 1$. This hourly discretization aims to capture the effects of different times of the day on trip time deviation.

Table 1. Attributes of the datasets, including derived attributes

Attribute	Description
Date	Date of trip
DepartureS	Departure time (in seconds)
Departure	Departure time discretized in 24 values (by hour)
Model	Vehicle model
Driver	Driver number. Zero means driver shift
DayOfYear	1st January is 1, 1st Feb is 32, etc.
WeekDay	From Monday to Sunday
DayType	Normal days, holidays, etc.
Duration	Duration of trip in seconds
SchDeparture	Scheduled time of departure
SchArrival	Scheduled time of arrival
SchDuration	Scheduled trip duration (seconds)
Deviation	Duration-SchDuration (seconds)

5 Discovering Interesting Contexts

We now look at the variable *Deviation* and use distribution rules to see how its distribution varies with the context. Our aim is to find combinations of conditions (contexts) that are associated with tendencies for positive and negative deviations, respectively delays and early arrivals.

From dataset1, using a level of significance of 0.05 on the Kolmogorov-Smirnov test and a minimum support of 2%, we obtained 30 rules. To avoid the generation of redundant rules we have also used a statistical significance filter. A rule $Ant_l \rightarrow D_l$ is added to the set of discovered rules only if there is no immediately simpler rule $Ant_s \rightarrow D_s$, such that Ant_s has exactly one item less than Ant_l and the distributions D_l and D_s are not significantly different. This latter test is performed using the same level of significance used for rule generation. We will examine some of the rules in more detail. We will use dataset2 to check the findings in dataset1. From dataset2, under the same conditions, we have obtained 25 rules.

5.1 Delays

In Fig. 3 we can see four of the top rules wrt positive deviation from dataset1. Positive deviation is associated with situations of frequent delays.

An emergent context is departure time between 2 and 3 P.M (Departure=14). Other contexts involving departure time around and immediately after lunch time also appear. This is an interesting information which implies that perhaps more care should be taken on that period, either by changing behavior (in order to observe defined schedules) or by adjusting schedules to operational conditions. The second rule shown indicates that a top cause of delay is the combination of driver's shifting (Driver=0 indicates change of driver) with the use of a particular

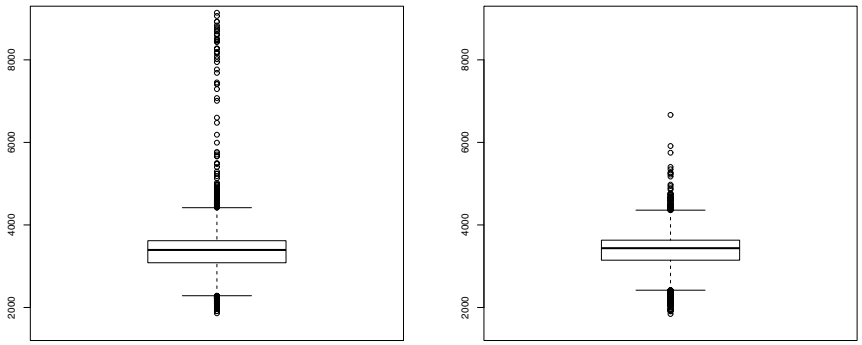

Fig. 2. Boxplots for trip duration in dataset1 and dataset2, respectively

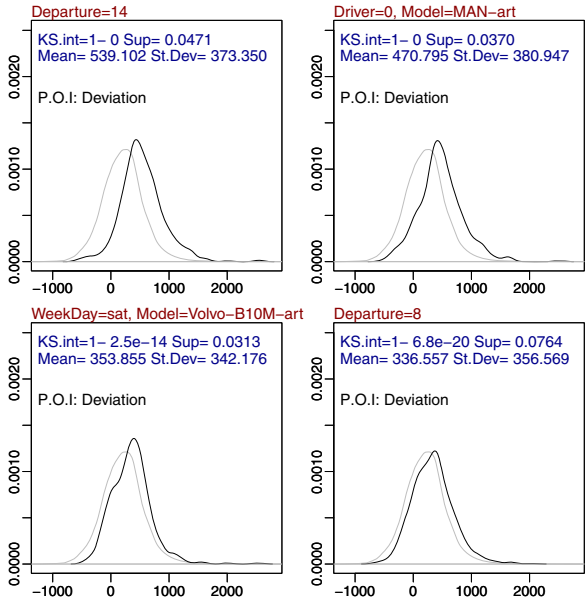

Fig. 3. Four top distribution rules for dataset1. The lighter line represents the a priori distribution of *Deviation*. The darker line represents the distribution in the context described by the condition or conditions above the box.

model of articulated vehicles. This draws the attention for the possibility of inefficiency in the process of transferring one vehicle from one driver to another. On the other hand, articulated vehicles may have difficulties in complying to the schedule. This may be caused by difficult manouvres in face of abusive street parking or narrow streets. The third context shown is Saturdays with another model of articulated vehicles. The fact that Saturdays are associated with delays

must also be studied by operational management. We can see that in these three contexts the distribution of trip duration deviation is clearly shifted to the right, with means over 300 seconds (5 minutes). Another interesting context for positive deviation is related to departure time between 8 and 9 A.M. These top rules cover from 3% to 7.6% of the trips.

The rules obtained from dataset2 confirm the importance of departure time. Early afternoon hours have a tendency for positive deviations. These are top rules here too. Driver change is also a cause for delay. The association between articulated vehicles and positive deviations is not observed. However, there is an association between the non-articulated model "MAN-3s" and negative deviations. A simple explanation for not having the model "MAN-art" as an interesting context here is that in this case this model represents more than 90% of the vehicles, whereas in dataset1 it is about 56%. Thus, the context "MAN-art" does not have a significantly different distribution of the whole set of trips.

5.2 Early Arrivals

From the same set of rules (dataset1) we obtain contexts that are related to negative deviations, in particular deviation distributions that are to the left of the a priori distribution (Fig. 4). Negative deviation is associated with early arrivals. In this case we observe that buses leaving after 7 P.M and before 8 tend to arrive earlier than in general. Moreover, there is a large majority of cases when these buses arrive before scheduled time. This is clear by observing the

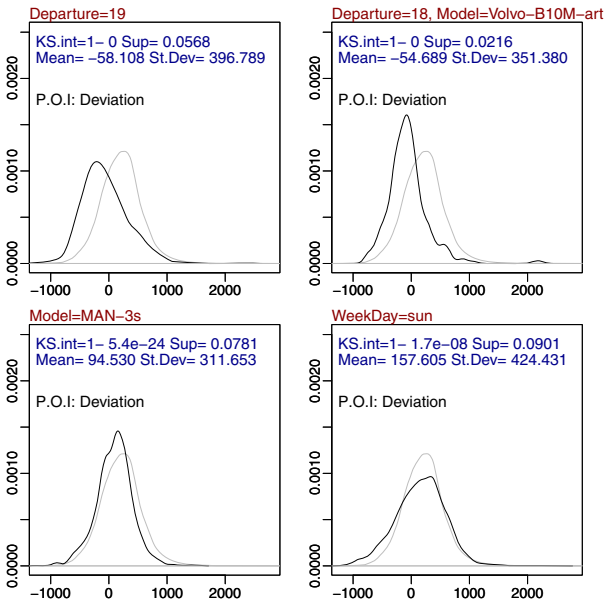

Fig. 4. Four top distribution rules with negative deviation for dataset1

resulting distribution for that context. Buses leaving beween 5 and 7 P.M. also have this tendency, but not as much. The use of vehicle model "Man-3s" is also associated with decreased delays. It is interesting to observe this association with this particular model, which is not articulated.

Another interesting context is related to departure time between 10 and 11 P.M. In this case there is a very good behavior with a high concentration around scheduled times (not shown in the Figure), i.e., practically no deviation is observed. This concentration effect is not surprising for this time of the day. The same behavior would also be expected for Sundays and holidays. Instead, however, Sundays appear as a day of many early arrivals, which can be very inconvenient for passengers. Similarly to Sundays, holidays have a large tendency for early arrivals. In fact, we can observe that early arrivals are also a tendency for departure between 8 and 9 P.M. but not for departures after 9 P.M.

With the rules obtained from dataset2 we confirm that early arrivals are associated with the early evening period (starting at 5 P.M. but mostly between 7 and 9 P.M.). As we have said above we can associate a non-articulated model with a lower tendency to delay. Regarding days of week and type of day, we also have Sunday as an interesting context.

6 Discussion

The use of distribution rules shows some advantages with respect to simple regression. With DRs we are able to find interesting contexts, or subgroups of trips, and look at the distribution of values instead of only having one or two parameters (typically mean and standard deviation). The discovery strategy based on the Kolmogorov-Smirnov test does actually look for interesting distributions. These may be shifted to the right, indicating a delay tendency, shifted to the left, indicating an early arrival tendency. By visually observing the discovered interesting distributions, we find particular situations of interest, as it was the case of Sundays and departures after 10 P.M. Moreover, we can distinguish between these two contexts which would, at a first sight, be similar. This exploratory visual inspection of relevant contexts provides the operational manager with hints for service improvement. The information thus obtained can be used to ameliorate the schedules. Next, we discuss some situations based on the results presented in Fig. 3 and 4.

In low frequency routes the planning should encourage timely departure times. For this type of routes, slack times (the time gap between the end of one trip and the beginning of the next for the same vehicle) should be used in order to accommodate trip time variation. The amount of variance assumed by the schedule should be a given percentage of the sample trip time distribution given by DR. It defines the trade-off between operational costs and client satisfaction. This choice is done by the planner. We use the right bottom plot of Fig. 4 to exemplify. The headway of this route on Sundays goes from 20 to 30 minutes (low frequency). We observe that a meaningful percentage of the trips have deviations lower than 0. This means that there is a high probability that the

bus will pass, at least in the last stops of the route, before the scheduled passing time. Considering the low frequency of this route on Sundays, this may imply that some clients lose the bus causing dissatisfaction. It would be advisable to reduce the scheduled trip time in this case or to communicate with drivers.

Situations like the one presented in the right bottom plot of Fig. 3 should be analyzed according to the slack time used. That is, if the slack time does not cover the majority of cases where the trip exceeded the scheduled trip duration, the slack time should be increased. On the other hand, if the planned slack time covers all the cases, it could be reduced in order to avoid waste of resources. Alternatively, bus driver scheduling could be optimized by including shifts at the end of trips that are expected to have higher slacks, in order to maximize driving time. The amount of delays covered by the slack time is, once more, something that must be defined by the planner.

There are other situations that should also be taken into account, such as the identification of problems with a particular vehicle model that should be considered during vehicle rostering; or measures to reduce delays due to the occurrence of driver's shifting during the trip.

6.1 Using DR to Determine Slack Time

The obtained distribution rules can be used to adjust slack times (ST) in order to optimize resources. Given a context Con we find the rule $Ant \rightarrow D$ whose conditions best apply to it. Then, we can determine the minimal duration t_{min} of ST that fails to cover at most a given percentage p_{uncov} of the trips. In this case, covering means that the slack time is sufficient to avoid delay. The value of t_{min} is the solution of the equation below, where fd_{Ctxt} is the density function given by the DR which applies to the context $Ctxt$.

$$t \text{ such as } p_{uncov} = \int_{-\infty}^{t} fd_{Ctxt}dx \tag{1}$$

If the value of t_{min} is negative then there is space to reduce trip time for the particular context considered. The rule that applies to a context is the one with the largest antecedent made true by the context. Ties are resolved by preferring rules with higher support.

6.2 Related Work

This paper discusses a tool that can be used to: (1) detect deviations between actual and scheduled trip duration and its causes; (2) support the definition of scheduled trip durations, slack times or frequencies.

As far as we know, the only existing study that addresses the first objective [5] uses classification based on association [12]. The authors discretize deviation into four classes. All discovered association rules have discretized deviation as consequent. This approach does not give useful information in order to address the second objective, as DR do (as described in Sect. 5).

Other works address only the second objective [3,15]. Using an economic perspective, [3] defines as objective function, the cost expressed in terms of scheduled trip durations, lateness and earliness unit costs. Using this approach it is possible to define the optimal scheduled trip durations and slack times (time between trips) for given ratios between the unit cost of scheduled trip durations and both the lateness and earliness unit costs. Another contribution of this work is the inclusion in the model of the effect of relaxation when the slack time is larger. I.e., it is known that when the schedule is tight, the actual trip duration is shorter than when it is large. Carey calls it the behavioral response. What Carey shows is that the timetable definition should be neither too tight, to avoid delays in departures, nor too large, to avoid behavioral inefficiency.

Under certain conditions, slack times can be optimized [15]. Using this approach, the shorter the slack time is, the shorter the scheduled headway is. The function to optimize defines the passengers' expected waiting time in terms of the scheduled headway and the variance of the delay. By using the function defined in [2] for the passengers' arrival at the bus stops, it is possible to adapt the solution to problems with large headways. These are analytic simplified general models. We can use distribution rules to model and estimate waiting time based on collected data. Moreover, this estimation can be done with respect to well defined emerging contexts instead of all the trips.

Our approach differs from these two [3,15] by leaving to the planner the decision on how to deal with the trade-off between operational costs and clients' satisfaction. Moreover, it is not made any assumption about the distribution of the data. The only assumption is about the sample. It is assumed that the sample is representative of the population.

7 Conclusion and Future Work

In this paper we have exploited distribution rule discovery to study deviations in bus trip duration. We have used real data collected from an urban bus company. Distribution rules allowed us to discover operational contexts that are, with high significance, statistically associated with relevant deviations in trip duration. Discovered rules can also be used to support the adjustment of timetables and schedules (e.g. redefining slack times).

Our next steps are to integrate these findings with bus operation in an ongoing collaboration with the bus company. Some of the discovered relevant contexts are already known by operational managers. Some others may yield suggestions that are not practical, due to the complexity of the processes. Some may even go against legally established principles. However, this work has shown that schedules must be brought closer to operational conditions. Moreover we can identify paths for addressing the problem of trip duration deviation. In particular, we will incorporate this tool in a decision support system for timetable adjustments [13]. The application of this approach to the tens of lines of the bus company can have a great impact in the reduction of inefficiency and the increase of client satisfaction.

Acknowledgements. This work is part-funded by the ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness), by the Portuguese Funds through the FCT (Portuguese Foundation for Science and Technology) within project FCOMP - 01-0124-FEDER-022701.

References

1. Azevedo, P.J.: CAREN - class project association rule engine (2008), http://www.di.uminho.pt/~pja/class/caren.html
2. Bowman, L.A., Turnquist, M.A.: Service frequency, schedule reliability and passenger wait times at transit stops. Transportation Research Part A 15, 465–471 (1981)
3. Carey, M.: Optimizing scheduled times, allowing for behavioural response. Transportation Research Part B 32(5), 329–342 (1998)
4. Conover, W.J.: Practical Nonparametric Statistics, 3rd edn. John Wiley & Sons, New York (1999)
5. Duarte, E., Mendes-Moreira, J., Belo, O.: Exploração de técnicas de classificação associativa no planeamento de horários de transportes públicos (exploitation of associative classification techniques in the planning the schedules of public transports). In: 9 Conferência da Associação Portuguesa de Sistemas de Informação, Viseu - Portugal (2009)
6. Frank, A., Asuncion, A.: UCI machine learning repository (2010)
7. Hand, D., Manila, H., Smyth, P.: Principles of Data Mining. MIT Press (2001)
8. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning. Springer (August 2001)
9. Jorge, A.M., Azevedo, P.J., Pereira, F.: Distribution Rules with Numeric Attributes of Interest. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS (LNAI), vol. 4213, pp. 247–258. Springer, Heidelberg (2006)
10. Kavšek, B., Lavrač, N., Jovanoski, V.: APRIORI-SD: Adapting Association Rule Learning to Subgroup Discovery. In: Berthold, M., Lenz, H.-J., Bradley, E., Kruse, R., Borgelt, C. (eds.) IDA 2003. LNCS, vol. 2810, pp. 230–241. Springer, Heidelberg (2003)
11. Klősgen, W.: Explora: A multipattern and multistrategy discovery assistant. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park (1996)
12. Liu, B., Hsu, W., Ma, Y.: Integrating classification and association rule mining. In: KDD 1998: Proceedings of the fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 80–86. ACM Press, New York (1998)
13. Mendes-Moreira, J., Duarte, E., Belo, O.: A decision support system for timetable adjustments. In: 13th EURO Working Group on Transportation Meeting (EWGT 2009), Padua - Italy (2009)
14. Mendes-Moreira, J., Jorge, A., de Sousa, J.F., Soares, C.: Comparing state-of-the-art regression methods for long term travel time prediction. Intelligent Data Analysis 16(3), 427–449 (2012)
15. Zhao, J., Dessouky, M., Bukkapatnam, S.: Optimal slack time for schedule-based transit operations. Transportation Science 40(4), 529–539 (2006)