Distribution rules with numeric attributes of interest *

Alípio M. Jorge[1], Paulo J. Azevedo[2] and Fernando Pereira[1]

[1] LIACC, Faculty of Economics, University of Porto, Portugal amjorge@liacc.up.pt
[2] Departamento de Informática, University of Minho, Portugal pja@di.uminho.pt

Abstract. In this paper we introduce distribution rules, a kind of association rules with a distribution on the consequent. Distribution rules are related to quantitative association rules but can be seen as a more fundamental concept, useful for learning distributions. We formalize the main concepts and indicate applications to tasks such as frequent pattern discovery, sub group discovery and forecasting. An efficient algorithm for the generation of distribution rules is described. We also provide interest measures, visualization techniques and evaluation.

1 Introduction

Learning and discovering probability distributions is an important and difficult problem in statistics, machine learning and data mining [12]. Machine learning has focused particularly on learning conditional probabilities of one target variable y (either numerical or categorical) with respect to a set of input variables X. However, the output of a learning algorithm is typically reduced to associating the most adequate value of y to each combination of values of the variables in X. This is the case in regression, classification and association discovery. Learning whole distributions goes beyond point estimation. In this paper, we approach the problems of discovering and presenting important conditional distributions of a target variable with respect to a set of input variables. Our approach is based on association rule discovery [1].

Association rules (AR) are highly legible chunks of knowledge that can be discovered from data. On top of that, the process for generating association rules is efficient enough to deal with very large databases, and the intended result is very well defined and free of heuristics. Although devised mainly for descriptive purposes, AR can also be useful in classification [14], clustering [10], regression [17], recommendation and subgroup discovery [11].

Typically, algorithms for the discovery of AR deal with categoric attributes only. Srikant [19] proposed a specific approach for the discretization of numerical attributes bearing in mind the descriptive aim of AR. In predictive tasks such as regression, [17] or classification [14] the independent numeric variables can be discretized using the supervised discretization MDL based algorithm [8].

* Supported by POCI/TRA/61001/2004 Triana Project (Fundação Ciência e Tecnologia), FEDER e Programa de Financiamento Plurianual de Unidades de I & D.

Avoiding pre-discretization, Fukuda *et al.* [9] proposed an algorithm for handling pairs of numeric attributes on the LHS of association rules. Aumann and Lindell [2] introduced *Quantitative Association Rules (QAR)*, where a frequent itemset (on the LHS of the rule) is associated with a statistical summary of a numeric attribute of interest(on the RHS). Numeric attributes appearing on the LHS are pre-discretized. Other authors have meanwhile improved some aspects of the original QAR [20, 21], in a different direction from the work proposed here.

To learn and discover distributions we propose *distribution rules* (DR). These associate a frequent itemset with an empirical distribution of a numeric attribute of interest without any loss of information. Distribution rules can be used in descriptive data mining tasks with the advantage of avoiding pre-discretization of the numeric variable of interest. We provide an efficient algorithm that discovers distribution rules and describe how to filter interesting rules, using the statistical distribution of Kolmogorov-Smirnov. Distribution rules can be easily visualized as frequency polygons and viewed by a domain expert or data analyst. Besides, DRs can also potentially be used in a predictive setting, and are not fundamentally limited to numeric properties of interest.

2 Distribution Rules

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

In this paper we will assume y is a numeric variable. Nevertheless, the concept of distribution rules is extended for categorical attributes as well. The attributes on the antecedent are either categorical or are discretized as in [8].

Example: Suppose we have clinical data describing habits of patients and their level of cholesterol. The distribution rule $smoke \wedge young \rightarrow chol = \{180/2, 193/4, 205/3, 230/1\}$ represents the information that, of the young smokers on the data set, 2 have a *cholesterol* of 180, 4 of 193, 3 of 205 and 1 of 230. This information can be represented graphically, for example, as a frequency polygon. The attribute *chol* is the property of interest.$\diamond$

Given a dataset S, the task of *distribution rule discovery* consists in finding all the DR $A \rightarrow y = D_{y|A}$, where A has a support above a determined mininum σ_{min} and $D_{y|A}$ is statistically significantly different (w.r.t. a pre-defined threshold) from the default distribution $D_{y|\emptyset}$. The default distribution is the one obtained with all the values of y for the whole dataset.

2.1 Presentation and visualization

Although distribution rules can be output as text, the length of the empirical distribution is normally too long to be readable in practice. Since the consequent

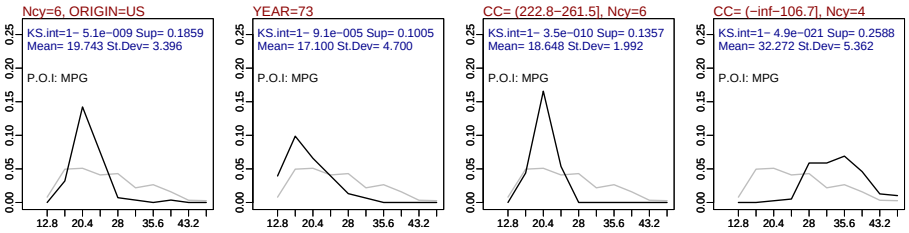

Fig. 1. Graphical representation of one distribution rule for the dataset auto-mpg

of one distribution rule is an empirical distribution, it can be represented as a frequency polygon. In Figure 1 we can see 4 rules obtained from the dataset auto-mpg [16]. The antecedent of each rule (e.g., the leftmost) is displayed as the main title. Some selected measures of the distribution and the name of the property of interest (P.O.I.: MPG) are shown within the plot. The x axis is the domain of the P.O.I. and the y axis the estimated probability density. The polygon is drawn by binning the domain of the P.O.I. into a given number of intervals (default 10) with equal width w. For each interval I, the pair x, y is plotted. The value of x is the lower limit of the interval and $y = freq_I/(freq_r * w)$, where $freq_I$ is the number of values in I and $freq_r$ is the total number of values of the P.O.I. covered by the rule.

The distribution for the set of cases that satisfy the condition is shown in black, and the default distribution for the whole population is shown in grey. For the distribution rule shown in Figure 1 we can see that cars with 6 cylinders built on the US tend to make less miles per galon than the whole set of cars. For those cars, the values of MPG are very concentrated around 20. Nevertheless, we can see that there are some economic cars in this group because of the right tail of the black curve. The interest of this rule is shown as KS.int, the complement to 1 of the Kolmogorov-Smirnov test p-value as explained in the following section.

Alternatively, the empirical distribution could be represented by a parametric distribution curve (e.g., Normal), or a boxplot. In this paper we adopted the frequency polygon, since it does not require any assumption about the distribution of the P.O.I. and it minimizes the loss of information regarding the distribution.

2.2 Measuring the interest of DRs

The interest of a discovered pattern can be measured according to objective and subjective criteria [18]. In the case of association rules, objective interest measures typically try to assess how much the observed frequency of the consequent of the rule, under the conditions imposed by the antecedent, deviate from the frequency that would be expected assuming that antecedent and consequent were independent. This is the case of measures such as *lift* (a.k.a. *interest*), *leverage* or *conviction*[5]. The χ^2 statistical test has also been extensively used for testing the statistical independence between the antecedent and consequent of association rules [15].

In the case of distribution rules, objective interest can be measured by assessing the difference between the distribution of the consequent and a reference distribution. This, in principle, is the distribution of the whole population. The difference between two empirical distributions can be assessed through a statistical goodness of fit test, such as *Kolmogorov-Smirnov* [7].

Definition. Given a set of transactions DS, a property of interest y in DS, and a distribution rule $A \rightarrow D_{y|A}$ obtained from DS, the *KS-interest* of that rule is $1 - p$ where p is the p-value of the Kolmogorov Smirnov test for the two empirical distributions $D_{y|A}$ and $D_{y|\emptyset}$.$\diamond$

Given this notion of the interest of a distribution rule, we can filter a set of DR's by selecting the ones with KS-interest above a pre-defined threshold. This threshold can be intuitively set by a data analyst since it has a clear statistical meaning. Although other notions of interest can also be defined using other statistical tests, we will for now focus on the use of the KS test.

3 Using DRs

Distribution rules can be used in descriptive pattern discovery tasks, although they can also be adopted in predictive tasks as well. One immediate advantage of their use in these situations is that it is not required to previously discretize the attribute y.

One way to handle distribution rules is by working with them as regular association rules. In Table 1 we can see a textual representation of one rule discovered for the dataset *Determinants of Wages from the 1985 Current Population Survey in the United States*, a.k.a. *Wages*, also used in [2]. While the antecedent of each of these rules is a frequent itemset, the consequent is a raw distribution. Although the rules can be represented in this textual form, they are internally stored using compact data structures. To present the rule, it is more effective to graphically visualize them, or to summarize them, as for example in [2].

Having obtained a set of distribution rules, these can be presented, sorted and filtered in many different ways. In this paper, we propose one particular graphical multi-plot presentation (Figure 2). In each plot, the default distribution is used as a term of comparison and appears in grey. The rules shown are a subset of the 35 rules produced for the dataset *Wages*, obtained with a minimal antecedent support of 0.1 and a min KS-interest of 0.95. We have also applied an improvement filter, as suggested in [4], on the KS-interest. In this case, $improvement(A \rightarrow B)$ can be defined as min({KS-interest$(A \rightarrow B)$−KS-

Table 1. A distribution rule produced for the dataset *Wages*, with min-sup=0.1, min-KS.int=$1 - 0.05$ and a minimal KS improvement of 0.01

```
Sup=0.118 KS.int=1-0.0085 Mean=10.982 St.Dev=6.333
EDUCATION=(12.5-15.5] & SOUTH=0 & RACE=3
  ->  WAGE={ 3.98/1,4.0/1,4.17/1,4.5/1,4.55/1,4.84/1,5.0/1,5.62/1,5.65/1,5.8/1,6.0/1,6.25/4,7.14/1,7.5/1,7.67/1,7.7/1,7.96/1,
          8.0/2,8.4/1,8.56/1,8.63/1,8.75/1,8.9/1,9.22/1,9.63/1,9.75/1,9.86/1,10.0/3,10.25/1,10.5/1,10.53/1,10.58/1,10.61/1,
          11.11/1,11.25/2,12.0/1,12.47/1,12.5/4,13.07/1,13.75/1,13.98/1,14.29/1,15.0/1,16.0/1,16.14/1,16.42/1,17.25/1,17.86/1,
          18.5/1,21.25/1,22.5/1,26.0/1,44.5/1 }
```

interest$(A_s \rightarrow B) \mid A_s \subseteq A\}$. The minimal KS-interest improvement used in these experiments was 0.01.

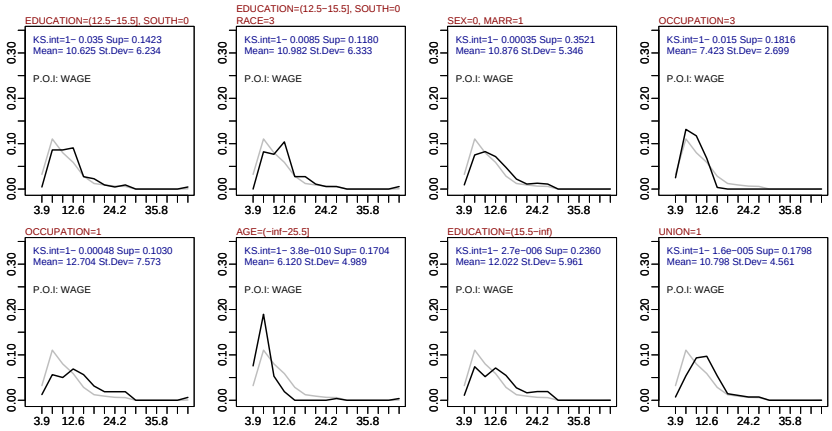

Fig. 2. Multi plot of a set of 8 distribution rules. Each is plotted against the default distribution

The 35 rules can be visualized in mosaic plots of $n \times m$. We show a mosaic of 2×4 with 8 rules selected from the 35. The selection can be done visually by paging the 35 rules in mosaics of $n \times m$. We will refer to the rules on Figure 2 by number from 1 to 8, reading from left to right and from top to bottom. Rules 1 and 2 describe people with 13 to 15 years of education which are not from the South. Their wages distribution is significantly different from the whole population and visibly better although concentrated on the same interval. Rule 2 is a refinement of rule 1 with higher interest. It seems that people with race=3 (white) in the conditions stated before have a slightly but significantly better situation. Rule 3 describes married males, and rules 4 and 5 show that occupation 1 has a wider and higher range of income than occupation 3. Rule 6 shows the impact of age, and rule 8 the positive effect of holding a union membership. Rule 7 indicates that people with higher education have higher wages.

Used in this setting, distribution rules are selected by the Kolmogorov-Smirnov statistic. Improvement enables the elimination of non informative sub rules. The visualization of the distributions gives a broader picture of the subset of data covered by each rule. With these parameters (min KS-int=0.95 and min improvement=0.01) we get very few rules.

Distribution rules can be naturally applied to the data minig task of subgroup discovery [13] both for numeric and categorical properties of interest. An interesting subgroup corresponds to a KS-interesting distribution rule.

Distribution rules can also potentially be used in predictive tasks such as regression as in [17] or probability density estimation as in [6]. In this paper we have focused on the fundamental concepts and on the processes of generating, filtering and presenting the rules.

Input: *minsup*, KS-int $= 1 - \alpha$, DB

Rules $= \emptyset$;
First database scan (count items)
Build DI = {items of the form $y = v_i$ belonging to property of interest};
Build AI = {antecedent items with count $\geq$ *minsup*} ;
Second database scan (bitmaps mounting)
Mount coverage bitmap for each item in AI and DI;
Compute $D_{y|\emptyset}$ using DI bitcounting;
foreach *transaction* $t \in DB$ **do**
    Set correspondent bit in each item (in AI and DI) occurring in t;
    Count 2-itemsets occurring in t;
end

`/* (Expansion phase)` `*/`

foreach *frequent item* $i \in$ AI **do**
    Compute $D_{y|i}$ from *bitmap*(i) and *bitmaps*(DI);
    if $KS(D_{y|\emptyset}, D_{y|i}) < \alpha$ **then** $Rules = Rules \cup \{i \rightarrow D_{y|i}\}$;
    foreach *frequent item* $i' > i$ *(> refers to items ordering)* **do**
        $a = \{i, i'\}$
        bitmap(a) = *bitmap*$(i) \oplus$ *bitmap*(i');
        if *support*$(a) \geq$ *minsup* **then**
            Compute $D_{y|a}$ from *bitmap*(a) and *bitmaps*(DI);
            if $KS(D_{y|\emptyset}, D_a) < \alpha$ **then** $Rules = Rules \cup \{a \rightarrow D_{y|a}\}$;
            $Rules = Rules \cup$ `Expansion`$(a, i', D_{y|\emptyset}, \alpha)$;
        end
    end
end

Output: *Rules*

Algorithm 1: CAREN-DR Depth First Distribution rules derivation

4 Rule Generation

A set of distribution rules can be obtained from a given database by computing all the frequent itemsets a not involving the property of interest y. For each frequent itemset a we compute the associated distribution $D_{y|a}$. Counting operations are efficiently implemented through the use of bitmaps.

4.1 Algorithm and computational complexity

The algorithm CAREN-DR works by finding frequent itemsets and, simultaneously, their associated p.o.i. distributions. For each antecedent item, a bitmap that represents its coverage is built. Antecedents are formed by depth first expansion. When an item is added to the antecedent to build a new itemset, a new bitmap is calculated (through bit-intersection) and its support can be counted through a bitcounting operation. To help in unfrequent itemset pruning during itemset expansion, the algorithm builds a flat matrix with 2-itemsets counts. Thus, expensive bitcounting operations can be avoided if subsets of the candidate itemsets are not frequent.

Input: (itemset,lastitem,$D_{y|\emptyset}$,α)
$R = \emptyset$;
foreach $i \in$ AI , $i >$ *lastitem (> refers to items ordering)* **do**
  if $\forall\, a \in itemset \quad support(\{a, i\}) \geq minsup$ **then**
    $new = itemset \cup \{i\}$;
    $bitmap(new) = bitmap(itemset) \oplus bitmap(i)$;
    $support(new) = bitcount(new)$;
    if $support(new) \geq minsup$ **then**
      Compute $D_{y|new}$ from $bitmap(new)$ and $bitmaps$(DI);
      if $KS(D_{y|\emptyset}, D_{new}) < \alpha$ **then** $R = R \cup \{new \rightarrow D_{new}\}$
      $R = R \cup$ `Expansion`$(new, i, D_{y|\emptyset}, \alpha)$;
    end
  end
end
return R

Function `Expansion`(*itemset,lastitem,*$D_{y|\emptyset}$,α)

For an efficient rule's consequent calculation, each distribution item (the numeric values associated with the p.o.i.) also keeps a bitmap. Deriving a new distribution requires intersection operations between the bitmap of the antecedent itemset and the bitmaps of the distribution items. The algorithm extracts significant rules by performing a Kolmogorov-Smirnov test between each new rule ($D_{y|a}$) and the *a priori* distribution ($D_{y|\emptyset}$).

The algorithm receives as input a minsup for antecedent filtering and an α that is used to set the minimal KS-interest threshold in $1-\alpha$. It can also receive an improvement threshold value. The theoretical complexity of the method is dominated by the complexity of finding frequent itemsets, which is known to be linear on the number of cases. Bitmap operations for the P.O.I. distribution update are also linear on the number of cases.

CAREN-DR algorithm conceptually resembles OPUS-IR algorithm [20], since it also uses a depth-first approach. In fact, with minimal adjustments, our proposal can be easily modified to work in a top-N rules search mode. However, we use bitmaps to represent itemset coverage and to calculate p.o.i. distributions. Our implementation of CAREN-DR is part of the java-based association rule discovery engine CAREN [3]. In relation to the QAR proposal of Aumann and Lindell, our algorithm does not require an extra database scan to compute the distributions associated to each rule. Furthermore our method outputs whole distributions and defines the interest of a rule in terms of comparison of distributions rather than the comparison of means.

5 Evaluation

In this section we show how our algorithm CAREN-DR performs on 4 different datasets described in Table 2. The algorithm has been run with different values of minimal support for a minimal KS-interest of 0.95 and with the improvement

Table 2. Description of the datasets used to measure the computation time (upper table). The column #Distinct indicates the number of distinct values of the property of interest (p.o.i.). Times in seconds and number of rules generated for the datasets for different minimal supports (lower table).

Dataset	#Attr	#Records	p.o.i.	#Distinct
mpg	7	398	MPG	129
housing	13	506	MEDV	211
abalone	8	4177	RINGS	28
cal. houses	9	20640	mhousevalue	3842

	Time in seconds				Number of rules generated			
min.sup	MPG	Housing	Abalone	Cal. Houses	MPG	Housing	Abalone	Cal. Houses
0.3	3.202	6.811	12.425	36.993	4	98	4	6
0.2	3.220	7.107	12.313	43.936	15	310	4	22
0.1	3.629	8.790	12.281	56.326	67	1490	41	74
0.05	5.089	13.894	12.790	72.084	240	5848	516	197
0.01	5.643	48.962	15.095	153.867	1369	51185	3158	1867

switch turned off. We can see that the algorithm scales up quite well with the number of examples and the value of minimal support. Table 2 (bottom) shows the times in seconds spent on a Pentium IV, 1.6GHz and 1GB RAM. These times include writing the rules to a csv file (one of the possible output modes). Table 2 (bottom) also shows the number of rules produced per run. We stress, however, that by turning improvement on, the number of rules falls dramatically.

These experiments show that the algorithm is capable of generating a very large number of distribution rules (and writing them as text to disk) in a very reasonable time (51185 rules for *Housing* in 49 seconds). In the case of the dataset *Cal. Houses*, CAREN-DR processes the 20640 cases in 2.5 minutes. Additional experiments with this dataset show that the time spent by CAREN-DR grows practically linearly as the number of examples rises from 5000 to 20000. In another set of experiments we observe that the time spent by the KS-test is also linear w.r.t the size of the distributions.

In our approach, the number of different values of the property of interest is also a source of complexity. However, typical numerical attributes tend to have low numerical precision, thus low variety of values. In the event of having to deal with a high precision attribute, we can round the values to a reasonable number of significant digits. Experimentally, we observe that the p-value of the KS test is robust to rounding to 3 significant digits.

5.1 An artificial dataset

In order to test the ability of the KS test to identify interesting rules, we have generated an artificial dataset with 1000 cases. The values of the attributes were chosen so that specific interesting distribution rules should appear. Thus, we have randomly initially generated the values for the p.o.i. y from a N(0,1) distribution. After that we have randomly assigned values of r or s to the categorical attributes a, b, and c. Whenever a and b had the values r we added to the value of y an extra random value from a distribution N(-2,1). The algorithm CAREN-DR produced the distribution rules shown in Figure 3 with a minimal support of 0.1, and a

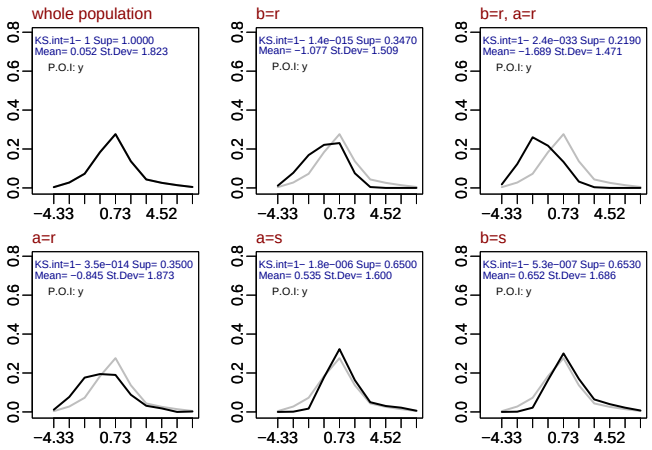

Fig. 3. Distribution rules for the artificial dataset

minimal KS-interest of 0.95. The minimal improvement on KS-interest used was 0.001. If the improvement filter is turned off, some redundant rules appear.

As we can see, only 5 rules have been identified, apart from the *a priori* rule. The condition $a = r \,\&\, b = r$ appears as expected, but also its items separately and their complements. The attribute c does not appear since the distribution of $y|c$ is similar to the distribution of $y|\emptyset$.

6 Case Study

We have applied distribution rules to the analysis of the main causes of delays in trip time duration for buses in a urban centre. This is a real dataset with about 8000 cases describing trips of a specific bus line. The dataset has 16 attributes plus the property of interest TripTime. The numeric attributes of the antecedent have been previously discretized using an implementation of the algorithm of Fayyad and Irani [8]. Since this discretization approach requires a class attribute, it is done with respect to a discretized version of the P.O.I, as in [17]. Afterwards, the P.O.I. is used in its continuous version. We obtain 36 relevant rules with support above 0.05 and KS-interest above 1-1E-05. Improvement is 0.0001.

In Figure 4 we can see a selection of the rules. Most rules have only one condition on the antecedent due to the effect of improvement filtering. We can see for example the difference in the distribution of the time a bus takes to make its route in March (Month=3) and in August (Month=8). Holydays also have a positive impact on trip time (plot 6) . The last two plots show the difference between Sundays and Fridays. The other attributes that appear on Figure 4 are Start, which is the starting time of the bus trip in seconds, DayOfYear, which is the number of days passed since 1st of January of the year being studied, and TypeOfDay, which can have values normal, or bank holiday.

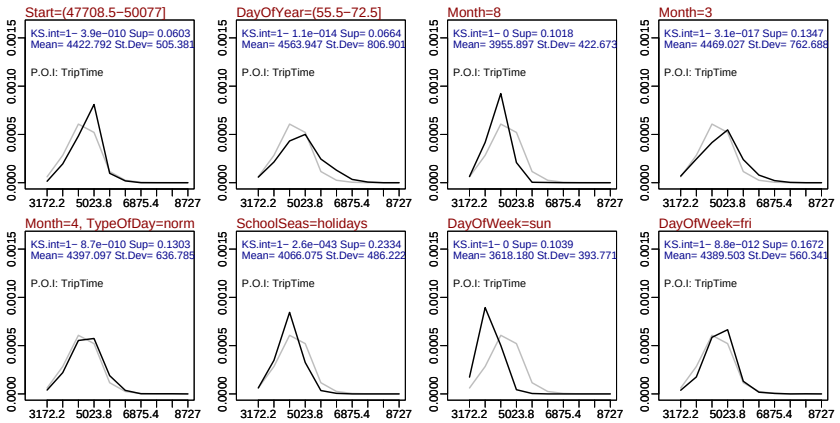

Fig. 4. Distribution rules for the buses dataset

This type of rules are being used to attempt to reduce the costs with personnel, since unpredicted delays often force the bus company management to pay for extra labour time. This way, distribution rules can be used both to give managers indications about the most relevant causes of delay and also enable to predict the probability that TripTime will be higher than a certain threshold.

7 Related Work

Distribution rules are mainly related to learning probability distributions [12], subgroup discovery [13] and quantitative association rules (QAR).

Aumann and Lindell's work on QAR uses a z-test to identify rules significance. As already pointed by Webb [20], z-test is inappropriate for small samples. The OPUS-IR authors propose the use of the standard t-test to decide on rules significance since the t-test tends to the z-test as the number of degrees of freedom increases. However, both z-test and t-test assume normality which in pratice cannot be guaranted. In this sense, using the KS approach is an advantage since no further distribution assumptions need to be considered.

Aumann and Lindell [2] propose an elaborated mechanism to identify and filter all the significante basic rules and sub-rules. They propose a notion of basic rule and an algorithm to find all significante "sub-rules" and "super-rules". The algorithm works as a post-mining step and builds a lattice of frequent sets to identify when a rule is basic or a sub-rule. The notion of super-rule is related to our notion of improvement. We also filter super-rules that do not bring about an improvement in the property of interest (in our case the KS-interest). However, applying improvement filtering does not require any sofisticated algorithm with lattice traversel. In fact, improvement filtering is computed on the fly, along rule derivation.

The QAR authors also present a mechanism to derive rules with more than one p.o.i. in the consequent. In pratice, it seems interesting to analyse several numerical properties in parallel. QAR has this feature as a post-processing step. We include this feature in the CAREN-DR engine during rule derivation. Thus, it is only required to specify the different properties to derive rules for.

Association rules have been used in subgroup discovery. APRIORI-SD [11] uses association rules to discover interesting subgroups with categorical properties of interest. Our approach enables the discovery of subgroups with numeric and categoric properties of interest. In this paper we employ the KS test to handle numeric properties.

8 Conclusion

We have introduced the concept of Distribution Rules as a generalization of association rules. We provide the basic concepts, such as the general form, support and objective interest of distribution rules. We also describe how to visualize distribution rules. DRs are particulary interesting when there is a numerical property of interest, although the concept can be extended to categorical properties as well. With classical association rules we would have to pre-discretize the numerical attribute of interest. With quantitative association rules, we would reduce the set of values in the consequent to a summary given by the mean or median. In the case of distribution rules, we keep the whole set of values of the property of interest and use these in graphical representations or post processing. Distribution rules can be presented as text or graphically and can be used in tasks of descriptive and predictive knowledge discovery.

References

1. R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.
2. Y. Aumann and Y. Lindell. A statistical theory for quantitative association rules. *Journal of Intelligent Information Systems*, 2003.
3. P. J. Azevedo and A. M. Jorge. The class project. http://www.niaad.liacc.up.pt/~amjorge/Projectos/Class/, 2006.
4. R. J. Bayardo, R. Agrawal, and D. Gunopulos. Constraint-based rule mining in large, dense databases. In *ICDE*, pages 188–197. IEEE Computer Society, 1999.
5. S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In J. Peckham, editor, *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, pages 255–264, Tucson, Arizona, 13–15 June 1997.
6. M. Carney, P. Cunningham, J. Dowling, and C. Lee. Predicting probability distributions for surf height using an ensemble of mixture density networks. In *Proceedings of the 22 International Conference on Machine Learning, ICML' 05, Bonn, Germany*, 2005.

7. W. J. Conover. *Practical Nonparametric Statistics - Third Edition*. John Wiley & Sons, New York, 1999.
8. U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *IJCAI*, pages 1022–1029, 1993.
9. T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Data mining using two-dimensional optimized association rules: scheme, algorithms, and visualization. In *SIGMOD '96: Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, pages 13–23, New York, NY, USA, 1996. ACM Press.
10. E.-H. Han, G. Karypis, V. Kumar, and B. Mobasher. Clustering based on association rule hypergraphs. In *Research Issues on Data Mining and Knowledge Discovery*, 1997.
11. B. Kavsek, N. Lavrac, and V. Jovanoski. Apriori-sd: Adapting association rule learning to subgroup discovery. In M. R. Berthold, editor, *Advances in Intelligent Data Analysis V*, volume 2810 of *Lecture Notes in Computer Science*, pages 230 – 241, Berlin Heidelberg, 2003. Springer-Verlag.
12. M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. E. Schapire, and L. Sellie. On the learnability of discrete distributions. In *STOC '94: Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 273–282, New York, NY, USA, 1994. ACM Press.
13. W. Klősgen. Explora: A multipattern and multistrategy discovery assistant. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA, 1996.
14. B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *KDD '98: Proceedings of the fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 80–86, New York, NY, USA, 1998. ACM Press.
15. B. Liu, W. Hsu, and Y. Ma. Pruning and summarizing the discovered associations. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 125–134, New York, NY, USA, 1999. ACM Press.
16. C. J. Merz and P. Murphy. Uci repository of machine learning database. http://www.cs.uci.edu/∼mlearn, 1996.
17. A. Ozgur, P.-N. Tan, and V. Kumar. Rba: An integrated framework for regression based on association rules. In M. W. Berry, U. Dayal, C. Kamath, and D. B. Skillicorn, editors, *SDM '04: Proceedings of the Fourth SIAM International Conference on Data Mining, Lake Buena Vista, Florida, USA*, 2004.
18. A. Silberschatz and A. Tuzhilin. On subjective measure of interestingness in knowledge discovery. In *KDD '95: Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, pages 275–281. AAAi Press, 1995.
19. R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. In *SIGMOD '96: Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, pages 1–12, New York, NY, USA, 1996. ACM Press.
20. G. I. Webb. Discovering associations with numeric variables. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 383–388, New York, NY, USA, 2001. ACM Press.
21. H. Zhang, B. Padmanabhan, and A. Tuzhilin. On the discovery of significant statistical quantitative rules. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 374–383, New York, NY, USA, 2004. ACM Press.