Intelligent Data Analysis 16 (2012) 25–47
DOI 10.3233/IDA-2011-0509
IOS Press

Optimal leverage association rules with numerical interval conditions

Alípio Mário Jorge[a,*] and Paulo J. Azevedo[b]
[a]*DCC-FCUP, Universidade do Porto, LIAAD, INESC Porto L.A., Porto, Portugal*
[b]*Departamento de Informática, Universidade do Minho, Braga, Portugal*

Abstract. In this paper we propose a framework for defining and discovering optimal association rules involving a numerical attribute A in the consequent. The consequent has the form of interval conditions ($A < x$, $A \geqslant x$ or $A \in I$ where I is an interval or a set of intervals of the form $[x_l, x_u)$). The optimality is with respect to leverage, one well known association rule interest measure. The generated rules are called Maximal Leverage Rules (MLR) and are generated from Distribution Rules. The principle for finding the MLR is related to the Kolmogorov-Smirnov goodness of fit statistical test. We propose different methods for MLR generation, taking into account leverage optimallity and readability. We theoretically demonstrate the optimality of the main exact methods, and measure the leverage loss of approximate methods. We show empirically that the discovery process is scalable.

Keywords: Numerical association rules, leverage, optimal association rules, distribution rules

1. Introduction

Association rules are an important technique for descriptive data mining. Especially, when the aim is to unveil properties of the data that can be potentially actionable. Fundamentally, association rules are concerned with sets of items or discrete attributes. The manipulation of numerical attributes in association rule discovery is not trivial. Numerical values cannot be directly treated as categorical at the cost of wasting generalization ability. On the other hand, pre-discretization forces early and sub-optimal decisions. The research literature has documented some non pre-discretization solutions for numerical values in the consequent of rules. For example, the target attribute can be summarized by showing its mean or median value, as it happens in the Quantitative Association Rules approach [2], or it can be graphically represented by its whole distribution, as it is the case of Distribution Rules [11]. Other approaches try to discover interval conditions on numerical attributes using association rule specific pre-discretization [25], clustering [6,14,18], rule templates [28], specific interest measures [12,13], bottom-up agglomeration of interval conditions [3], and genetic algorithms [24,27]. Some approaches are capable of discovering confidence optimal rules for fixed support and vice-versa [9,22] under provided rule templates. This paper describes an optimal approach for dealing with numerical attributes in the consequent of the rules. This approach avoids pre-discretization. The number of attributes in the antecedent is arbitrary. We exploit the path of determining, for each rule, optimal interval conditions on the numerical attribute in the consequent. All the other attributes in the antecedent are assumed

[*] Corresponding author. E-mail: amjorge@fc.up.pt.

to be categorical or previously discretized. It is not a concern of this paper the discretization of the variables in the antecedent. In particular, we look for conditions which are optimal under a well known notion of association rule interest: *leverage*, also known as Piatetsky-Shapiro measure [20]. Another related measure, *added value* [23], is also optimized. The rules formed with these optimal intervals are called “Maximal Leverage Numerical Rules”, or “Maximal Leverage Rules” for short. This work develops from our previous work on Distribution Rules [11]. There, a rule was objectively considered interesting if the numerical attribute in the consequent had a significantly different distribution, under the conditions in the antecedent, when compared with the overall distribution. This was measured using a Kolmogorov-Smirnov statistical test to determine whether the two distributions were likely to be the same or not. What we observe is that an interesting distribution rule under the Kolmogorov-Smirnov setting may become an interesting rule with interval conditions in the consequent under the optimal leverage setting. Moreover, we can assure that we do not overlook any such interval rule. What we propose is not a regression technique, but instead a technique for finding a well defined set of descriptive rules. On the other hand, in this paper we are not so much concerned with algorithmic issues of rule discovery but rather with theoretical interestingness properties. We use the previously existing implementation for discovering distribution rules from data. The interval rules with maximal leverage are obtained by post processing each of the resulting distribution rules. Our claims are theoretically justified and illustrated with examples on well known datasets. In this paper we start by presenting an overview of interval discovery for association rule mining. We then present Maximal Leverage Rules and describe how they can be obtained from distribution rules. We theoretically demonstrate the optimality of the proposed methods and then describe techniques for improving the readability of rules. In the following, completeness and complexity of the MLR generation are analyzed and some experiments are performed on regression data sets.

2. Related work

The simplest way to deal with numerical attributes in association rules is by employing some kind of pre-discretization, such as equal-widh or equal-depth. However, generic discretization techniques are not necessarily the most adequate to satisfy the aims of association rule discovery. When we discretize, we should know the impact in the resulting rules. We can also discretize so that optimal rules are found under well defined criteria.

Early discretization approaches for association rule discovery were focused on the aim of finding all rules with support and confidence above given values. Under this framework, Srikant and Agrawal [25] proposed a principled pre-discretization of numerical attributes. Since the aggregation of values makes the rules less precise but more general, the trade-off is settled by asking the user how much precision one is willing to lose. This parameter, which is user provided, determines the width of the smallest intervals. This discretization approach fits into the support-confidence association rule discovery framework. Miller and Yang [18] tried to overcome some limitations of Srikant and Agrawal’s proposal by taking into account the distribution of the values and their distance. Their proposal involved clustering the values of quantitative attributes. Their discovery framework was distance-based rather than based on support and confidence. Miller and Yang describe an algorithm for mining such rules but they do not provide any completeness property stating that this algorithm finds all interesting rules under their notion of interest. Lent et al. [14] also proposed clustering for generating rules of the form $X_1 \in Int_1 \wedge X_2 \in Int_2 \rightarrow Class = K$. Attributes X_1 and X_2 are numerical, Int_i are intervals and $Class$ is categorical. Their approach starts by binning the numerical attributes into equi-width intervals, after which association rules of the above form are generated for the base intervals (bins). Then, given a fixed K, rules are clustered together by merging neighboring intervals. The clustering is guided by a simplified Minimum Description Length measure, which simultaneously tries to minimize the number of rules and the number of cases that are not covered by any rule. The resulting algorithm is shown to be very efficient, which is in part justified by the very restricted language bias. Another factor contributing for efficiency is the heuristic search approach used to find the clusters.

Wang et al. [28] use a bottom-up interval merging algorithm for finding rules. Rule generation is governed by pre-defined templates which allow the use of quantitative and qualitative attributes. The merging of intervals is made on the basis of improving the pre-existing interestingness J*-measure*. Although experimental results with artificial data show interesting capabilities of the approach, no optimal solution is guaranteed. Besides, the use of templates implies some previous knowledge about the rules which are to be found.

In QAR-miner, by Cheung et al. [6], data attributes are assumed to be all numerical. The multi-dimensional space defined by the attributes is first discretized into equal sized base regions (hyper-rectangles defined by intervals on the attributes). Too small intervals are avoided by setting a minimum *density*. Density is the number of cases in the region (a hyper-rectangle) divided by the region's volume. A minimum volume threshold is also user-defined. Base regions are then merged through clustering. Rules are obtained by regarding the resulting regions as itemsets, and checking if their support and confidence are large enough. QAR-miner is also more efficient than Skrikant and Agrawal's approach, but we have no measure of what we are losing when we discretize. The guiding principle is that dense regions (intervals) are interesting. Although sharing some similarities with the distance based approach, Cheung's rules are characterized in terms of well known support and confidence. No completeness property of the algorithm is provided. Yiping Ke et al. [12,13] generate association rules with an arbitrary number of items, involving qualitative and quantitative attributes. Quantitative items have conditions on intervals. The key idea of their approach is the use of a specific interest measure *normalized mutual information* (*NMI*) which applies to both qualitative and quantitative items. They handle numeric attributes by pre-discretizing them with equi-depth intervals. Then, in the second step, they identify all pairs of attributes that are independent according to *NMI*. In the third step, itemsets are formed from combinations of these pairs of attributes, considering their possible values. In the case of numerical values, base intervals are considered, as well as combinations of base intervals up to a user provided maximal support. The algorithm is efficient and sound with respect to minimum support and confidence (all provided rules satisfy these criteria) but is not complete w.r.t. the same criteria (some rules may be missed). However, the method can be made complete with respect to the measure $interest$ (*lift*).

Ayubi et al. [3] devise an algorithm for discovering association rules involving items $X\Theta v$ where the operator Θ can be one of $=, \neq, >, \leqslant$. The procedure starts by finding frequent itemsets with operator $=$ (numerical and categorical attributes are treated equally without necessarily pre-processing). These are called simple itemsets. Then, it combines simple itemsets in order to include items with other operators. This is done by joining consecutive values of numerical attributes. One important aspect of this approach is that the database is visited only once for counting the supports of simple itemsets. The algorithm is efficiently implemented, produces compact rules when compared with boolean association rules, but is obviously not complete.

2.1. Optimal intervals

Fukuda et al. [9] contributed with an efficient algorithm that finds rules with the form $X \in Int \rightarrow Cons$, where $Cons$ is a fixed boolean condition. It finds rules which have enough support and are optimal in terms of confidence, and rules which have enough confidence and are optimal in terms of support. This is done by looking for optimal interval bounds using a non-trivial efficient procedure which is reduced to finding optimal slopes in support-confidence curves. The whole algorithm is relaxed in a number of ways for the sake of efficiency: the numerical attribute is first binned (which introduces some source of global sub-optimality); and the bins are determined using a sample of the whole data set if it is too large to fit in memory. Fukuda's proposal has the advantage of precisely characterizing the resulting rules. They also show that the method can be trivially extended to rules of the form $X \in Int \wedge Cond \rightarrow Cons$, where $Cond$ is a boolean condition. In their earlier paper [8], Fukuda et al. also propose rules involving two numerical attributes in the antecedent, thus defining rectangular regions. They show that looking for such rules is $O((N_A \times N_B)^{1.5})$, where N_A and N_B are the number of bins for the attributes involved. This is regarded as computationally demanding.

Rastogi and Shim [22] extended the work of Fukuda et al. [9] by allowing the rules to have disjunctions and an arbitrary number of quantitative and qualitative attributes. Rule search is restricted by the use of a template. A template is a rule where attribute names and values may be undetermined. Rule search and generation proceeds by filling in (instantiating) these free slots. Disjunctions enable conditions of the form $X_1 \in Int_1 \vee \ldots \vee X_k \in Int_k$. Although in general this problem is NP-hard, Rastogi and Shim propose a non-trivial search procedure which considers candidate rules as ordered by a combination of support and confidence $w_1 * supp + w_2 * conf$.

2.2. Evolutionary discretization

Interesting intervals can also be found using evolutionary approaches, which start with a population of itemsets with conditions of the form $X \in Int$. Then, appropriate operators for mutation and crossover are applied to improve the itemsets according to a pre-defined fitness function. Under this general setting, Mata et al. [27] propose GAR (Genetic Association Rules) for generating quantitative itemsets. The fitness function favours high support and itemset size, penalizing itemset coverage overlap (two itemsets covering the same tuples) and wide intervals. The proposal has no guaranteed result, but simple experiments with artificial datasets show some interesting capabilities of the approach.

Saleb-Aouissi et al. [24] have later noted that the GAR does not take rule confidence into account, since the evolutionary approach is only used for itemset generation. Their algorithm, QuantMiner, starts with a population of rules generated from user-provided templates. One important difference w.r.t. GAR is that the fitness function takes into account the *gain* of the rule, which is related to the rule's confidence. The fitness function also favors small intervals and penalizes low support.

3. Maximal leverage numerical rules

Our proposal for discovering association rules with interval conditions focuses on rules which have one numerical attribute in the consequent. This attribute is a specific property of interest (POI). In our approach, this POI is dynamically discretized according to the conditions in the antecedent of the rule. The discretization is optimal with respect to well defined criteria.

Given a dataset D, with all attributes categorical except one (A), which is numerical, we want interesting rules (or rules that can be potentially interesting in practice) of the form $Ant \rightarrow A \in I$, where Ant is a conjunction of conditions on the categorical attributes. I is an interval or union of intervals. For example I can have the form $[x_l, x_u)$. If such an interval is unbound on one of the sides the condition becomes $A < x$ or $A \geqslant x$. Let us also assume for now that we have some measure of interest on such rules. Under such a setting, the baseline approach discretizes attribute A and uses some existing association rule discovery engine. However, such a global pre-discretization does not guarantee optimal results, since the exact form of the interval may be different for each particular rule. Our aim is to avoid pre-discretization and the sub-optimal decisions that come with it. The interest measure we will use is *leverage*, also known as Piatetsky-Shapiro's measure [20].

$$\mathit{Lev}(r = A \rightarrow C) = \mathit{Supp}(r) - \mathit{Supp}(A).\mathit{Supp}(C)$$

Supp(*X*) is the support as usually defined [1], i.e., the relative frequence of X. Leverage takes values from the interval $[-0.25, 0.25]$ [26]. When antecedent and consequent are statistically independent it is zero. Therefore, for values close to zero, the rule is considered uninteresting. Commonly, negative values of leverage are not exploited either, since they indicate a negative association. Such rules, however, might be interesting for particular applications. In our case, we assume we are interested in rules with leverage above some positive threshold. Nevertheless, our results can be easily adapted for finding rules of minimal leverage. Although less popular, the *added value* (AV) interest measure [23] is very intuitive and will be useful in our analysis. It is the difference between the *a posteriori* and *a priori* values of confidence (*Conf*(.)).

$$AV(r = A \rightarrow C) = \mathit{Conf}(r) - \mathit{Supp}(C)$$

Below we show an example of a *maximal leverage numerical rule*, using the UCI dataset "Auto MPG" [17]. In the first line we give the value for the measures coverage (Cov), leverage (Lev), added value (AV) and confidence (Conf).

```
(Cov=0.226 Lev=0.148 AV=0.653 Conf=0.922)
HP=(132.5, inf) & Ncy=(5.5, inf) &
  Year=(-inf, 79.5]  ->  MPG < 18
```

The above rule states that cars with horse power (HP) above 132.5, more than 5 cylinders (Ncy) and made (Year)) before 1980 tend to have a miles per galon (MPG) value below 18 when compared to a generic car. In fact, the rule says that a generic car will only have such a bad performance with much lower probability (0.653 lower to be precise, as it is given by the added value AV). The condition $\mathit{MPG} < 18$ is the interval condition of the form $A < x_u$ or $A > x_l$ that maximizes leverage (and added value) for that sub group of cars. The confidence of the rule is 0.922. This characterizes the sub group in a very precise way. At the same time, tells us much about the relation between the features in the antecedent and the numerical property of interest. Note that the same rule but with the consequent $\mathit{MPG} < 20.3$ would have confidence equal to 1. However, its leverage would not be optimal. "Cov" is the coverage of the rule, which is simply the support of its antecedent. In this case, the antecedent of the rule applies to 22.6% of the examples in the data set.

The use of leverage as an interest measure is motivated by two reasons. First, leverage is one of the interest measures that assess the unexpectedness of the rule. The other reason is that leverage has mathematical properties that allow us to make clear statements. Other popular interest measures, such as *lift* (also known as interest), *conviction* [5] or χ^2 [15] will apparently require a different approach. Added value has a very intuitive reading but is a less popular measure. Our results have implications on both leverage and added value. Whenever we say "maximal leverage" or "optimal leverage" this also implies "maximal" or "optimal added value".

We should stress that such an optimal rule would be very unlikely to obtain if pre-discretization of the target variable had been used. One possibility would be to exhaustively consider all possible intervals.

However, this would be unfeasable, in general. Even less likely would be to find all optimal rules given a specific pre-discretization procedure.

This kind of rules can be potentially useful in practical applications where there is a numerical property of interest and the aim is to study its behavior. One example is travel time in urban public transports [11]. Managing schedules of buses and bus drivers is a very complex task that can be assisted through the use of predictive and descriptive data mining techniques [16]. A discovered maximal leverage rule can give a manager information about under which conditions (day of week, month, holiday season, driver, etc.) there are significant drifts in travel time. A maximal leverage rule indicates, for a given set of conditions, the intervals of travel time values that are more specific of those conditions. Such discovered knowledge can be potentially useful for bus schedule adjustment and for explaining deviations from the planned schedule.

3.1. Stating the problem

Let us now state our knowledge discovery problem of finding rules with maximal leverage having interval conditions on a numerical attribute in the consequent (maximal leverage rules, for short).

Given a dataset S with a numerical attribute A, a minimum coverage value $Cov_{\min}$ and a level of significance α, *Find* all the rules r of the form *Ant* $\rightarrow$ $A \in I$, where I is an interval or a union of intervals, such that I maximizes *Lev*(*r*) over *Ant*, *Cov*(*r*) $>$ $Cov_{\min}$ and *Ant* is a conjunction of conditions that defines an interesting subgroup with a level of significance α.

Cov(*r*) is the coverage of the rule. The condition on the interestingness of *Ant* will be verifyed by a goodness of fit test comparing the distribution of A for the cases which satisfy *Ant* with the a priori distribution of A. Our proposal is to use the Kolmogorov-Smirnov test.

4. Distribution rules

Distribution rules (DR) have been introduced in [11] as a kind of association rules with a numerical 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. As we will see, we can discover maximal leverage rules from distribution rules (DR) with advantage over a direct approach. 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*). By post-processing the interesting distributions, we can obtain optimally interesting intervals on A, as far as leverage is concerned.

Let us review what distribution rules are and how they can be derived from data [11].

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.*

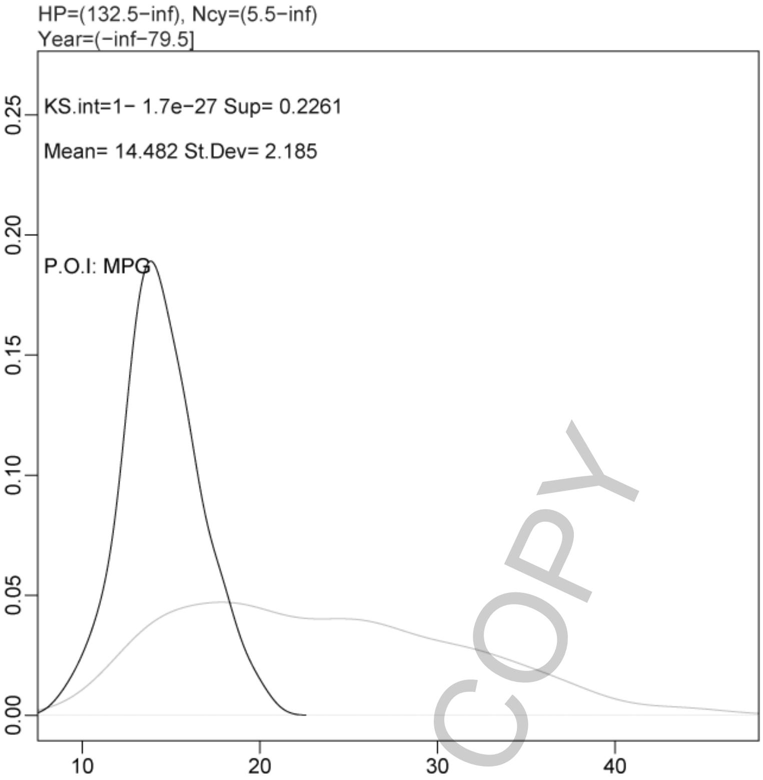

Fig. 1. Distribution rule for the "Auto MPG" data set.

In the context of this paper we will assume A is a numerical variable. However, the concept of distribution rules includes categorical attributes in the consequent as well. The attributes on the antecedent are either categorical or are discretized.

In Fig. 1 we can see one distribution rule derived from the "Auto MPG" data set. In fact this distribution rule is related to the same subgroup as the maximal leverage rule shown before. 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. In the chart we can also see some measures characterizing the rule: KS-interest, which is given by $1 - p_{KS}$, where p_{KS} is the *p value* of the Kolmogorov Smirnov test; the support (Sup) of the antecedent, the mean of *MPG*|*Ant* and its standard deviation. There is also an indication of what the property of interest (POI) is. The represented densities are estimated using kernel density estimation [10].

Given a dataset S, 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 mininum $\sigma_{\min}$ and $D_{A|Ant}$ is statistically significantly different (w.r.t. a pre-defined threshold) 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* [7], a statistical goodness of fit test. The value of the KS (Kolmogorov-Smirnov) 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*.

5. Finding maximal leverage rules

We now get back to our problem of finding rules with maximal leverage. We will see how such rules can be obtained from distribution rules. We start with a special case of conditions on intervals: inequalities. We show that, given a distribution rule *Ant* $\rightarrow D_{A|Ant}$, we can obtain a maximal leverage rule *Ant* $\rightarrow A \in I$, where I is one interval unbound on one of the sides. Such consequents can be written as $A < x_u$ or $A \geqslant x_l$. Then we turn to the general case where conjunctions of interval conditions are used.

5.1. Inequality consequents

In the process of determining whether a distribution rule is interesting or not we calculate the value of the KS (Kolmogorov-Smirnov) statistic. Although not interesting for the KS test, the point $x_{\max}$ where the maximum of $|F_s(x) - F(x)|$ occurs has a special meaning. In fact, if we have a rule $Ant \rightarrow A < x_{\max}$, the leverage of such rule is $(F_s(x_{\max}) - F(x_{\max})) \times$ *Supp*(*Ant*), in the case $F_s(x_{\max}) > F(x_{\max})$. If $F_s(x_{\max}) < F(x_{\max})$ then it will be more interesting to look at the rule *Ant* $\rightarrow A \geqslant x_{\max}$ with positive leverage $(F(x_{\max}) - F_s(x_{\max})) \times$ *Supp*(*Ant*). The case $F_s(x_{\max}) = F(x_{\max})$ does not occur, since we are assuming that a KS test was performed with success.

Since $|F_s(x_{\max}) - F(x_{\max})|$ is maximal, and *Supp*(*Ant*) is constant given the antecedent, then the rule with maximal added value, and hence leverage, with an inequality in the consequent, is one of the two above. If $F_s(x_{\max}) - F(x_{\max})$ is positive the optimal rule is *Ant* $\rightarrow A < x_{\max}$. If it is negative the optimal rule is *Ant* $\rightarrow A \geqslant x_{\max}$. Algorithm 1 describes the procedure for finding the optimal boundary of the inequality (threshold value). Theorem 1 formalizes the above line of reasoning.

Algorithm 1. Finding the optimal threshold for inequalities: method "inequality"

```
  Input: $D_s$: empirical distribution of the subgroup; $D$: empirical a priori distribution
  Output: $t$: threshold value
points = $\{x_1, x_2, \ldots, x_v\} \leftarrow values(D)$
for $i \in \{1, \ldots, v\}$ do
    $F(x_i) \leftarrow \sum_{p \in points < x_i} freq_D(p)$
    $F_s(x_i) \leftarrow \sum_{p \in points < x_i} freq_{D_s}(p)$
$t \leftarrow arg\max_{x \in points} |F_s(x) - F(x)|$
```

Theorem 1. *Let Ant* $\rightarrow D_{A|Ant}$ *be a distribution rule,* $F_s(A)$ *be the empirical probability distribution of A under the subgroup defined by Ant,* $F(A)$ *be the a priori empirical distribution function A. The rule of maximal leverage of the form Ant* $\rightarrow A < x_{\max}$ *or Ant* $\rightarrow A \geqslant x_{\max}$*, is the one where* $x_{\max} = arg\ min_{x \in Dom(A)} |F_s(x) > F(x)|$*. It has the first form* $(A < x_{\max})$ *if* $F_s(x_{\max}) > F(x_{\max})$ *and the second form* $(A \geqslant x_{\max})$ *otherwise. Its leverage is* $|F_s(x_{\max}) - F(x_{\max})| \times$ *Supp*(*Ant*)*. Dom*(*A*) *is the set of points observed in* D_A*.*

Proof. Assume $F_s(x_{\max}) > F(x_{\max})$. Then, by definition of leverage, the leverage of the rule with $A < x_{\max}$ is

$$(\mathit{Conf}(Ant \rightarrow A < x_{\max}) - \mathit{Supp}(A < x_{\max})) \times \mathit{Supp}(Ant) \qquad (1)$$

which is the same as $(F_s(x_{\max}) - F(x_{\max})) \times$ *Supp*(*Ant*). Since this is the maximal value of the difference, any other point will give a rule with lower leverage, which proves it for the first case.

If $F_s(x_{\max}) < F(x_{\max})$, then the leverage of the rule with $A \geqslant x_{\max}$ is $((1 - F_s(x_{\max})) - (1 - F(x_{\max})) \times$ *Supp*(*Ant*), which is the same as $(F(x_{\max}) - F_s(x_{\max})) \times$ *Supp*(*Ant*). Again, if the absolute difference is maximal for $x_{\max}$, then any other point will give a rule of this form with lower leverage. □

Example 1. *Consider the distribution rule in Fig. 1. If we take the empirical distributions $D_{MPG|Ant}$ and D_{MPG}, we can easily determine that the maximal value of $|F_s(x) - F(x)|$ is for $x = 18$. Therefore, the rule*

```
(Cov=0.226 Lev=0.148 AV=0.653 Conf=0.922)
HP=(132.5, inf) & Ncy=(5.5, inf) &
  Year=(-inf, 79.5]  ->  MPG < 18
```

is the maximal leverage rule with that antecedent and with an inequality in the consequent. We can observe that $x = 18$ is very close to the intersection point of the two estimated density curves. We will return to that later.

Example 2. *Some other maximal leverage rules with inequalities obtained from the "Auto MPG" dataset are*

```
(Cov=0.291 Lev=0.17 Conf=0.914)
Weight=(3422.5, inf) & ORIGIN=US &
Year=(-inf, 79.5]  ->  MPG < 19

(Cov=0.241 Lev=0.13 Conf=0.938)
Weight=(-inf, 2217]  ->  MPG >= 26
```

We can observe that we automatically obtain the sensible direction of the inequality. The first rule says that heavy old cars from the US tend to make less than 19 miles per galon. The second rule states that light cars tend to make 26 m.p.g or more. The rules have a very intuitive reading and clearly point a tendency. The rules also have high confidence.

5.2. The general case of intervals

Inequalities provide rules with a quite neat description of data. However, we must be open to the possibility of having rules with more general conditions of the form $Ant \to A \in I$, where I is an interval or a union of intervals. The intervals have the form $[x_l, x_u)$, i.e., they are closed on the left end and open on the right. This naturally subsumes inequalities.

Such general rules can also be found by studying the empirical distribution difference $F_s(x) - F(x)$. Unfortunately, they tend to be less readable than inequalities especially when I is a union of intervals. The existence of more than one interval is mostly caused by the natural instability of the empirical distribution. Albeit the reduction in readability, the leverage of an interval rule tends to be higher than its inequality rule counter-part, and it is at least as high.

Example 3. *Consider the distribution rule shown in Fig. 2. We can see that the density of MPG is higher until about 10, and then from about 13 to twenty something. If we try to transform this DR into a MLR with an inequality, we obtain*

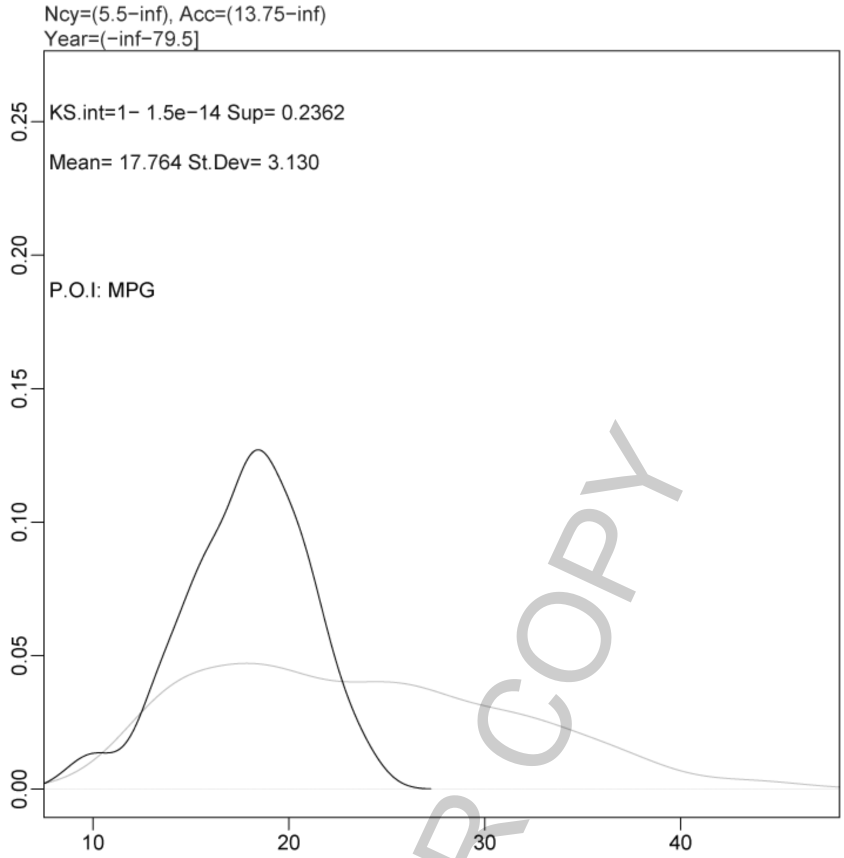

Fig. 2. Distribution rule with two interest areas in the property of interest.

```
(Cov=0.236 Lev=0.11 Conf=0.957)
Ncy=(5.5, inf) & Acc=(13.75, inf) &
Year=(-inf, 79.5]  ->  MPG<23
```

However, the following rule would have higher leverage.

```
(Cov=0.236 Lev=0.12 Conf=0.968)
Ncy=(5.5, inf) & Acc=(13.75, inf)
& Year=(-inf, 79.5]
-> MPG=[-Inf, 12) OR MPG=[13, 14.5)
 OR MPG=[15, 17.6) OR MPG=[18, 19.1)
 OR MPG=[19.2, 20.3) OR MPG=[20.5, 21.1)
 OR MPG=[21.5, 21.6) OR MPG=[22, 22.3)
 OR MPG=[22.5, 23) OR MPG=[23.9, 24)
```

Such rules can also be automatically obtained from distribution rules. However, readability is a problem. We may notice in the above example the existence of many small intervals, such as $[23.9, 24)$, which may be disposable at a possibly small cost. Other strategies may also increase readability at a small cost. We will look closer into this problem later on.

Let's now see how interval bounds are found. The process generalizes the one seen for finding inequality consequents. Given a distribution rule, we can compare its consequent distribution function

F_s against the *a priori* distribution F of the property of interest. In the case of inequalities, we need the *global* maximum of the difference $|F_s(x) - F(x)|$. Now we will look for the *local* maxima and minima of $F_s(x) - F(x)$. When this difference grows, leverage is increasing. When it stops growing (at a local maximum) we have found an interval right bound. When it starts growing again (at a local minimum) we have identified an interval left bound. All growing portions of the domain increase leverage. All decreasing portions reduce leverage. Algorithm 2 provides the outline for the procedure for finding the local minima and maxima of $F_s(x) - F(x)$, which become the lower and upper bounds of the intervals in the consequent. Theorem 1 proves the correctness of the algorithm.

Algorithm 2. Finding the optimal interval bounds: method "intervals"

```
  Input: $D_s$: distribution of the subgroup; $D$: a priori distribution
  Output: $LB$: set of interval lower bounds; $UB$ set of interval upper bounds
$points = \{x_1, x_2, \ldots, x_v\} \leftarrow values(D)$
for $i \in \{1, \ldots, v\}$ do
    $F(x_i) \leftarrow \sum_{p \in points < x_i} freq_D(p)$
    $F_s(x_i) \leftarrow \sum_{p \in points < x_i} freq_{D_s}(p)$
    $AV(x_i) \leftarrow F_s(x_i) - F(x_i)$
$UB = UpperBounds \leftarrow localmaxima(AV(x_i))$
$LB = LowerBounds \leftarrow localminima(AV(x_i))$
if $\min(UB \cup LB)$ is an upper bound then
    $LB \leftarrow \{-\infty\} \cup LB$
if $\max(UB \cup LB)$ is a lower bound then
    $UB \leftarrow \{+\infty\} \cup UB$
```

Theorem 2. *Let Ant* $\rightarrow D_{A|Ant}$ *be a distribution rule in the same conditions as in Theorem 1. The rule of maximal leverage of the form Ant* $\rightarrow A \in I$ *where* I *is an interval or a union of intervals of the form* $[x_l, x_u)$*, is the one whose bounds are obtained as described in Algorithm 2. The set of intervals are obtained by pairing the elements of* LB *with the ones of* UB *following their order. The leverage of the rule is obtained by adding, for each interval* $[x_l, x_u)$*,* $(F_s(x_u) - F_s(x_l)) - (F(x_u) - F(x_l))$.

Proof. Since each interval in I is limited by a local mininum and a local maximum of $F_s(x) - F(x)$, this implies that for any segment $[x_l, x_u)$ within the interval $F_s(x) - F(x)$ is monotonically increasing. Therefore $(F_s(x_u) - F_s(x_l)) - (F(x_u) - F(x_l))$ is always positive. As a consequence, if we take out any segment of the intervals, the overall leverage value reduces. Likewise, for any segment outside the given intervals $F_s(x) - F(x)$ is monotonically decreasing. Analougously, if we add to I any external segment the overall value of leverage will decrease. Which proves that we have found the rule of maximal leverage of the required form. □

Example 4. *Looking back at the maximal leverage rules with inequalities shown in Example 2, we now study the corresponding rules with generic intervals.*

```
(Cov=0.291 Lev=0.18 Conf=0.966)
Weight=(3422.5, inf) & ORIGIN=US
& Year=(-inf, 79.5]
->  MPG=[-Inf, 16.2) OR MPG=[16.5, 18.1)
 OR MPG=[18.2, 19) OR MPG=[19.2, 19.8)
 OR MPG=[20.5, 20.6)
```

```
(Cov=0.241 Lev=0.15 Conf=0.948)
Weight=(-inf, 2217]
->  MPG=[24.5, 25) OR MPG=[26, 26.4)
 OR MPG=[27, 27.2) OR MPG=[28, 28.1)
 OR MPG=[29, 29.9) OR MPG=[30, 30.7)
 OR MPG=[31, 31.3) OR MPG=[31.5, 31.6)
 OR MPG=[31.8, 32.2) OR MPG=[32.3, 32.4)
 OR MPG=[32.8, 32.9) OR MPG=[33, 34)
 OR MPG=[34.1, 36.4) OR MPG=[37, 43.4)
 OR MPG=[44, Inf)
```

Leverages are higher (0.18 against 0.17 in the first rule and 0.15 against 0.13 in the second). Confidence is higher for both interval rules as compared against inequality rules.

In the example above, again we observe the existence of too many "small" intervals which reduce the readability of the rules. Are they worthwhile keeping? What should our criteria be? That's what we will deal with in the next section.

6. Improving readability

As we have seen, we obtain the bounds of the intervals for the consequents of maximal leverage rules from the empirical distributions. Being samples, each distribution carries some instability which is observed when we study it closely. As a consequence, the empirical distributions have some "turbulence" areas and local optima of $F_s(x) - F(x)$ may abound in these regions. Since the interval bounds directly correspond to local optima, we will frequently have large collections of intervals as a consequent. The problem with such large collections is reduced readability.

Different strategies may be devised to improve readability. Always necessarily at the cost of reducing leverage, since we have proved how the optimal leverage is found. The simplest one is to use a tighter language bias, as it was the case with inequality rules. Another strategy tries to eliminate too small intervals, the ones which have little impact in the overall leverage. A third one tries to focus on the points of *Dom*(*A*) which are potencially more interesting for scrutiny. Finally, we can smooth the distribution curves to reduce turbulence by using density estimation.

As a proxy for improved readability and leverage loss, we will follow one rule about old large cars obtained with the "Auto MPG" data set. The optimal rule, without any readability concern is shown below. Its leverage is 0.155 and it has 5 conditions on the consequent.

```
(Cov=0.226 Lev=0.155 AV=0.687 Conf=0.989)
HP=(132.5, inf) & Ncy=(5.5, inf)
& Year=(-inf, 79.5]
->  MPG<17.6 OR MPG=[17.7, 18)
 OR MPG=[18.1, 18.6) OR MPG=[19.2, 19.8)
 OR MPG=[20.2, 20.3)
```

6.1. Using tighter language bias

We have already seen how this can be done with inequalities. Instead of using general intervals, we use less expressive but more readable conditions. The question is, given a tighter bias, how much do we lose in terms of leverage? Without any constraints on the underlying distributions, it is hard to say.

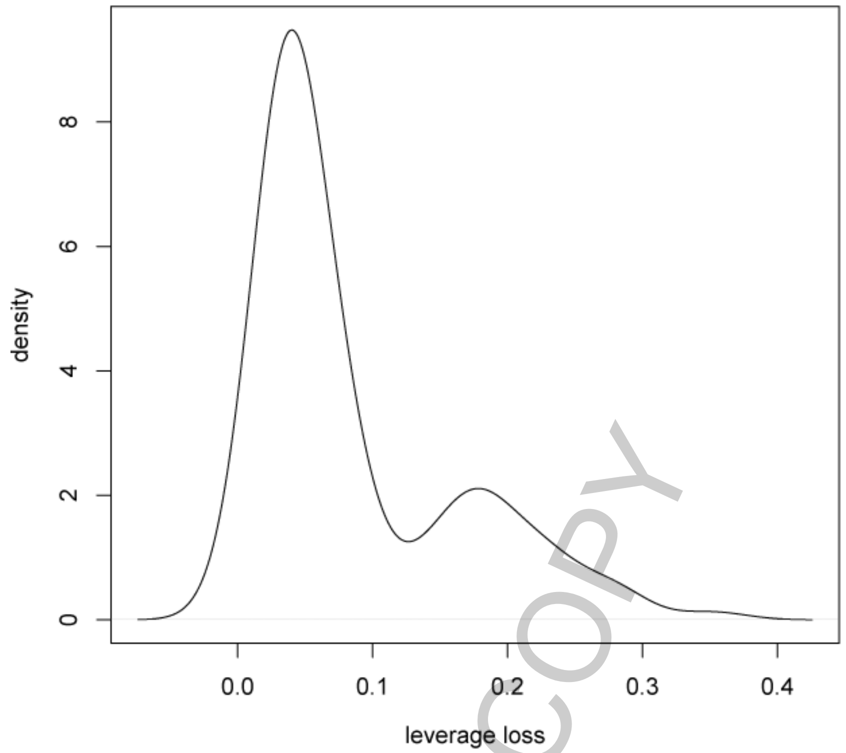

Fig. 3. Estimated density of the leverage loss from inequalities to intervals.

Empirically, we can observe the distribution of leverage loss from interval rules to inequality rules on particular discovery tasks. The rule for the same subgroup of old large cars is now as follows.

```
(Cov=0.226 Lev=0.148 AV=0.653 Conf=0.922)
HP=(132.5, inf) & Ncy=(5.5, inf)
& Year=(-inf, 79.5]
->  MPG < 18
```

We will also study leverage loss on the whole "Auto MPG" dataset. What we observe is that, for a minimal support of 0.2 and a KS significance (α) value of 0.05, the mean leverage loss is of 0.022, reaching a maximum of 0.078. Median is 0.012. This means that, in this data set, the cost of using a tighter bias is not high, but it can be relatively important for some particular subgroups (rule antecedents). Figure 3 gives a full picture of the leverage loss distribution. If we measure the reduction in the number of intervals, and we take this as a proxy for improved readability, the mean value for the "Auto MPG" dataset is of 88%. In other words, the single interval in the consequent of the inequality rule is 12% of the number of intervals for the optimal rules (on average).

6.2. Interval smoothing

Interval smoothing consists in identifying intervals that have little impact on overall leverage and remove them. Likewise, we can try to remove "negative" intervals, by joining/bridging two consecutive intervals with a small gap in between. Again, the gap (or negative interval) is removed if its removal has little impact. One direct strategy we have tried is to remove all gaps and intervals whose individual impact is below a (user provided) maximal impact threshold *max.imp*. Other strategies for measuring the impact of interval removal could be considered. Since the interval/gap removal decisions are independent of one another, the accumulated leverage loss for a given rule may reach *max.imp* × $n.int$, where *n.int* is the number of intervals in the consequent. In an extreme situation, this strategy may remove all the intervals, which is not desirable. In the case of the rule for old large cars we obtain

```
(Cov=0.226 Lev=0.147 AV=0.650 Conf=0.911)
HP=(132.5, inf) & Ncy=(5.5, inf)
& Year=(-inf, 79.5]
->  MPG<17.6
```

Empirically, we observe that the impact on leverage of removing intervals with a $max.imp$ of 0.02 on the MPG set of rules we have been using is relatively low, with a maximum of 0.062 and a mean of 0.017 (median is 0.011). On the other hand, the reduction in the number of intervals is significant (improved readability). This direct strategy reduced, in average and for this data set, the number of intervals in the consequent by 58%. Preceding interval removal by gap removal, degrades results both in terms of leverage loss (mean 0.029) and readability improvement (mean 32%). The gap removal strategy tries to delete each of the interval gaps, and accepts the removal of those which do not degrade leverage more than $max.imp$ (0.02). The example rule is shown below. By chance, this strategy found a rule with confidence 1.

```
(Cov=0.226 Lev=0.134 AV=0.593 Conf=1)
HP=(132.5, inf) & Ncy=(5.5, inf)
& Year=(-inf, 79.5]
->  MPG<17.6 OR MPG=[17.7, 20.3)
```

6.3. *Restricting the domain*

This strategy is based on restricting the pool of points considered as candidate interval bounds. In the optimal approach, the set of points taken as $Dom(A)$ is the set of all values in the dataset for the attribute A (the property of interest). By restricting the $Dom(A)$ for a given rule (subgroup) to the set of values of A observed in that subgroup, we get more readable rules. Again, necessarily, these are sub-optimal. Here is the observed example rule:

```
(Cov=0.226 Lev=0.152 AV=0.672 Conf=0.956)
HP=(132.5, inf) & Ncy=(5.5, inf)
& Year=(-inf, 79.5]
->  MPG<18 OR MPG=[18.1, 18.5)
 OR MPG=[19.2, 19.4)
```

An experiment with the "Auto MPG" data set indicates that this strategy can be effective in terms of improved readability. The mean reduction in the number of intervals is of 64%, which is better than interval smoothing for the same dataset. Mean leverage loss is slightly better (0.013) but its distribution has a long right tail, which means that it can have (with low probability) high values (maximum is 0.089).

6.4. Density estimation

In this case, readability is improved by using density estimations instead of empirical densities. So far, maximal leverage numerical rules have been determined on the basis of empirical distributions, as observed in the data set. Using kernel density estimation [10] we can smooth the behavior of $F(x))$ and find approximate maximal leverage rules with fewer interval conditions. The rules obtained using density estimation will be sub-optimal. However, they may also be more robust to the arrival of new observations. In the current implementation we have employed the R built-in function "density" with a gaussian kernel for density estimation.

The procedure for finding interval bounds with densities (which we will refer to as the MLR generation method "estimation") is equivalent to the one for finding optimal intervals for the empirical distributions. Given the two distributions D_s and D, we estimate the densities $d_s(x)$ and $d(x)$. The local optima of the distribution functions correspond to the points where the two density curves intersect. Therefore, the interval bounds are points from the domain of A (even if not observed in the data) where the difference $d_s(x) - d(x)$ changes sign.

```
**Algorithm 3**: Finding approximate interval bounds: method "estimation"
  **Input**: $D_s$: distribution of the subgroup; $D$: *a priori* distribution
  **Output**: $LB$: set of interval lower bounds; $UB$ set of interval upper bounds
$points = \{x_1, x_2, \ldots, x_v\} \leftarrow values(D)$
**for** $i \in \{1, \ldots, v\}$ **do**
    $d(x_i) \leftarrow$ kernel estimated density for $D$
    $F_s(x_i) \leftarrow$ kernel estimated density for $D_s$
    $sign(x_i) \leftarrow sign(d_s(x_i) - d(x_i))$
$UB = UpperBounds \leftarrow \{x_i, i \geqslant 1 | sign(x_i) = ``+'' \wedge sign(x_{i-1}) = ``-''\}$
$LB = LowerBounds \leftarrow \{x_i, i \geqslant 1 | sign(x_i) = ``-'' \wedge sign(x_{i-1}) = ``+''\}$
**if** $\min(UB \cup LB)$ is an upper bound **then**
    $LB \leftarrow \{-\infty\} \cup LB$
**if** $\max(UB \cup LB)$ is a lower bound **then**
    $UB \leftarrow \{+\infty\} \cup UB$
```

Let us see the result of one example. In Fig. 1 we can see that the density within the subgroup seems to be higher than the *a priori* density for for MPG values below approximately 18. The obtained rule, however, provides more detail that is not easily visible in Fig. 1. In particular the interval $MPG = [7.98, 8.01)$ seems redundant. Nevertheless, it is objectively important for increasing leverage. This interval appears in this rule because of some instability in the density curves.

```
(Cov=0.226 Lev=0.144 AV=0.636 Conf=0.956)
HP=(132.5, inf) & Ncy=(5.5, inf)
& Year=(-inf, 79.5]
->  MPG=[7.98, 8.01) OR MPG=[8.04, 18.31)
```

Experiments with the "Auto MPG" data set indicate that this estimation method greatly improves readability without sacrifycing too much leverage. On average, for the Auto data set, leverage loss is 0.02 and readability improves 85%. Moreover, as we will see in the experimental validation section of this paper, this method is very efficient.

7. Completeness

Our aim was to find all optimal rules with interval conditions on A in the consequent with significant interest given a level of significance α, and with coverage above given $Cov_{\min}$. The general process for discovering distribution rules identifies all the rules whose antecedent has support above a given $Sup_{\min}$. This will be the coverage of the maximal leverage rule. It is different from its support because the interval I will have to exclude some points. Therefore we satisfy the minimum coverage requirement for maximal leverage rules if we make $Sup_{\min} = Cov_{\min}$ in the DR discovery process. Hence, given the completeness of the maximal leverage rule discovery process is as complete as the process for discovering distribution rules. We should note that some filters can be used in the process of the discovery of distribution rules. These make the process faster but may cause incompleteness. However, this is empirically tolerated.

8. Complexity

As in the discovery of association rules, the time taken by the discovery of distribution rules is mainly affected by the number of cases N and the minimal support/coverage σ. Time spent grows linearly as N grows, and tends to grow exponentially as σ decreases. Other aspects are data set dependent and are related to the distribution of the cases [11]. The number of different numerical values V of the property of interest is also important, since, for each rule, we must compare two distributions involving at most V values for determining the value of the KS statistic and its p-value. This comparison is done by running through the sequence of values once and is therefore $O(V)$. Memory spent is also $O(V)$.

Obtaining maximal leverage rules from DRs implies going through the pair of distributions for each rule. Time is $O(V)$ either for finding local or global optima of $F_s(x) - F(x)$. For reducing the number I of intervals or gaps time is $O(I)$, if the specific leverage value for each interval is stored. If it is not stored, then the process is again at most $O(V)$. Density estimation adds an extra effort which is, at most, also $O(V)$.

9. Experiments

Our experiments have two aims. One is to study the scalability of the proposed algorithms with respect to the number of cases and the number of distinct values of the property of interest. The other is to perform a more general assessment of the impact of sub-optimal strategies on leverage. These experiments are performed on a number of regression datasets.

9.1. Running versions

Our current running version of the methods for generating MLR from data are implemented in java and R [21]. The distribution rule generation phase is implemented in java as the program *CAREN*, version 2.6 [4], also available as the R package *carenR*.[1] The MLR methods are currently prototyped in R, but will be integrated in *CAREN* in the future.

9.2. Data sets

The data sets used come from the UCI repository [17] and from the repository of Luís Torgo.[2] All the attributes except the property of interest are pre-discretized. The pre-discretization technique used is not of particular interest. Table 1 summarizes the data sets.

[1] http://www.liaad.up.pt/∼amjorge/Projectos/carenR/.

[2] http://www.liaad.up.pt/∼ltorgo/Regression/DataSets.html.

Table 1
Data sets used

Data set	#Expls	#POI Values	#Attr
Abalone	4177	28	9
Auto	398	129	9
Housing	506	229	14
Cal. Housing	20640	3842	9
Cart example	40768	40368	11

Table 2
Running times (in seconds) for the distribution rule generation phase and the maximal leverage rules generation phase, for different minimum coverages

$min.cov = 0.1$	$\#rules$	t_{DR}	t_{ineq}	t_{int}	t_{est}
abalone	41	0.87	0.30	1.09	1.86
auto	223	2.35	1.93	2.00	10.30
housing	709	2.51	9.35	11.66	33.50
cal-hou	72	14.96	19.26	318.36	11.44
cart	354	802.96	422.05	2922.13	119.46
$min.cov = 0.05$	$\#rules$	t_{DR}	t_{ineq}	t_{int}	t_{est}
abalone	479	1.51	3.73	22.31	22.32
auto	425	1.36	3.84	3.86	19.71
housing	1719	3.13	30.13	33.74	91.71
cal-hou	190	19.20	31.68	521.19	20.58
cart	1002	1041.74	864.72	6443.32	187.52
$min.cov = 0.01$	$\#rules$	t_{DR}	t_{ineq}	t_{int}	t_{est}
abalone	3054	3.50	72.79	95.95	185.56
auto	1561	1.52	5.30	5.24	26.36
housing	4731	5.43	173.36	179.16	396.44
cal-hou	1787	42.96	159.56	1398.15	129.44
cart	1002	1380.32	862.56	6424.88	187.86

9.3. Scalability

Here, we want to confirm the theoretical complexity analysis that says that MLR generation time grows linearly with the number of different POI values. Moreover, we confirm that the generation of distribution rules also grows linearly with the number of examples. The data sets used for these experiments have different sizes, ranging from small to medium size. We have generated rules for 3 values of minimal coverage, 0.1, 0.05 and 0.01. The three maximal leverage methods used were *inequalities*, *intervals* and *estimation*. Distribution rules were generated with an α (KS-test level of significance) of 0.05. Table 2 shows the execution times for the DR generation, and MLR generation, for each of the three methods. We can see that execution times are higher as the number of different POI values increases. The "Abalone" data set, for example, despite being about 10 times larger than "Housing" and "Auto", takes much less time. The number of rules generated also plays a role. However, it does not dominate, as we can see with "Cal. Housing". There, only 72 rules are generated and it is the second most time consuming data set.

Regarding the MLR methods, the most time consuming is the optimal one (intervals). Using inequalities is considerably faster, but the most stable runtimes are obtained with the estimation method. Although these data sets give an overall perspective of the computational effort taken, it is hard to devise trends due to too many varying factors. Therefore, we will engage in more focused experiments that help to clarify these trends.

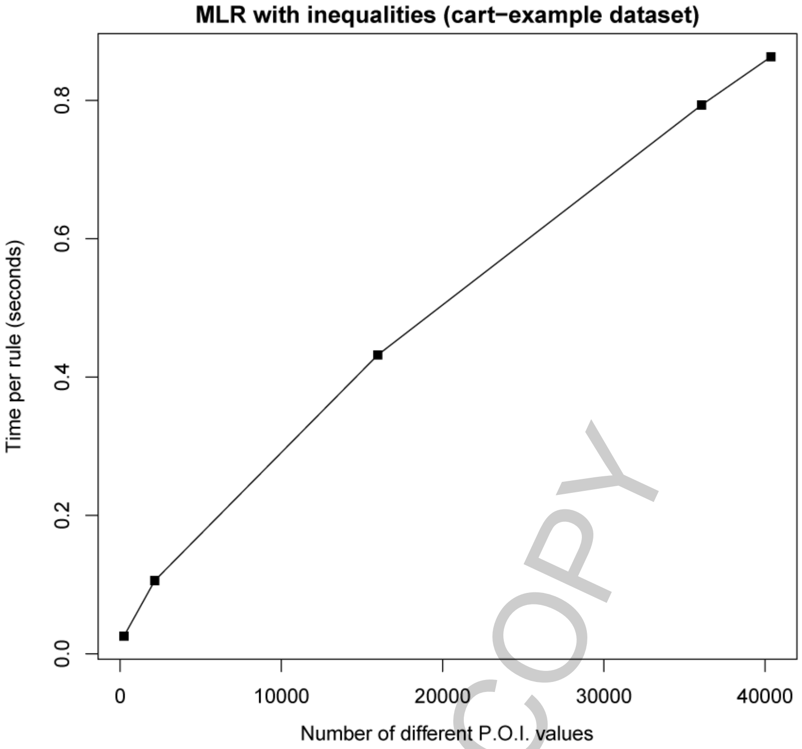

Fig. 4. Scalability of the inequality MLR method with the number of values of the POI (minimal support = 0.05, data set "cart-example").

To study the influence of the number of different POI values in the runtimes, we have generated 4 variants of the "cart-example" data set, by rounding the property of interest to 1, 2, 3 and 4 decimal places. This simple process produces different numbers of distinct values of the POI. As we can observe in Fig. 4, the computation time per rule grows linearly with the number of different values on the property of interest for the "inequality" MLR method. Figure 5 shows that the estimation MLR method is sublinear. This happens because the values of the density functions are estimated for a fixed number of points (512 in our experiments, which is the default parameter of R's density estimation function). In practice, this has an effect similar to reducing the number of different POI values considered as candidates for interval bounds. This particular experiment confirms our complexity analysis. The time spent for finding MLR grows at most linearly with the number of values. This implies that we can apply the method to other domains with a large number of different values in the POI.

We now confirm the linear scalability of distribution rule generation. We have randomly splitted "Cal. Housing", in 5 parts, and ran the DR generation on a growing sample of the data set. The results can be seen in Fig. 6.

9.4. Readability

In Section 6 we have already empirically measured the impact of different MLR generation methods on leverage and on readability. However, this has been done for one particular data set for illustrative purposes. In this section we extend the empirical assessment to other data sets and study in more depth the impact of minimal converage and different POI values on MLR readability. The experimental

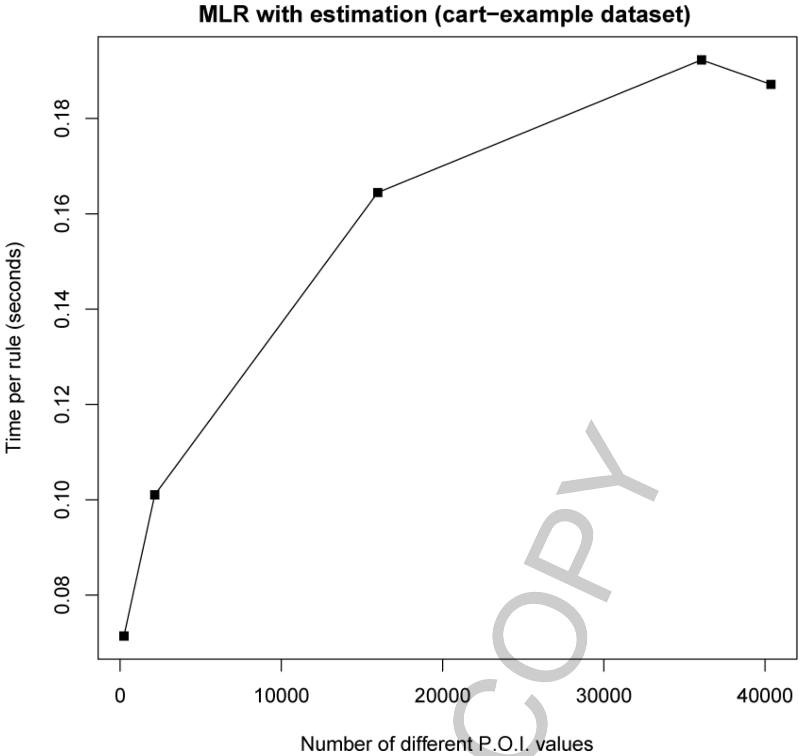

Fig. 5. Scalability of the estimation MLR method with the number of values of the POI (minimal support = 0.05, data set "cart-example").

measures used are "leverage loss" and "improved readability". The leverage loss of one rule r for a method m is the difference between the optimal leverage for rule r, as obtained by the method "interval", and the leverage of the corresponding rule obtained with method m. Improved readability of a rule r for a method m is 1 minus the quotient between the number of interval conditions in the consequent of r with method m and the number of interval conditions in the consequent of the corresponding optimal rule r. In Fig. 7 we can observe the relative performance of methods "inequality" and "estimation" on the 5 data sets. As we can see, no method is overall better. "Estimation" loses less leverage but has lower readability than "inequality". This is perfectly expected since the "inequality" method generates rules with 1 interval condition. Given these results, we can claim that the method "estimation" is a very good compromise. In particular, we can see the nearly extreme improvement for the data sets with many different POI values ("cal. housing" and "cart"). In fact, for these data sets we obtain values very close to 1. This happens because the optimal rule has tens of interval conditions. These experiments do not include the other presented techniques for improving readability (interval smoothing and domain restriction) due to poor computational performance of their current implementation.

10. Discussion

We have empirically confirmed the scalability of the optimal MLR generation methods proposed: "inequality" and "intervals". However, this second method is clearly computationally more expensive. The "estimation" method shows very good scalability features as the number of different POI values

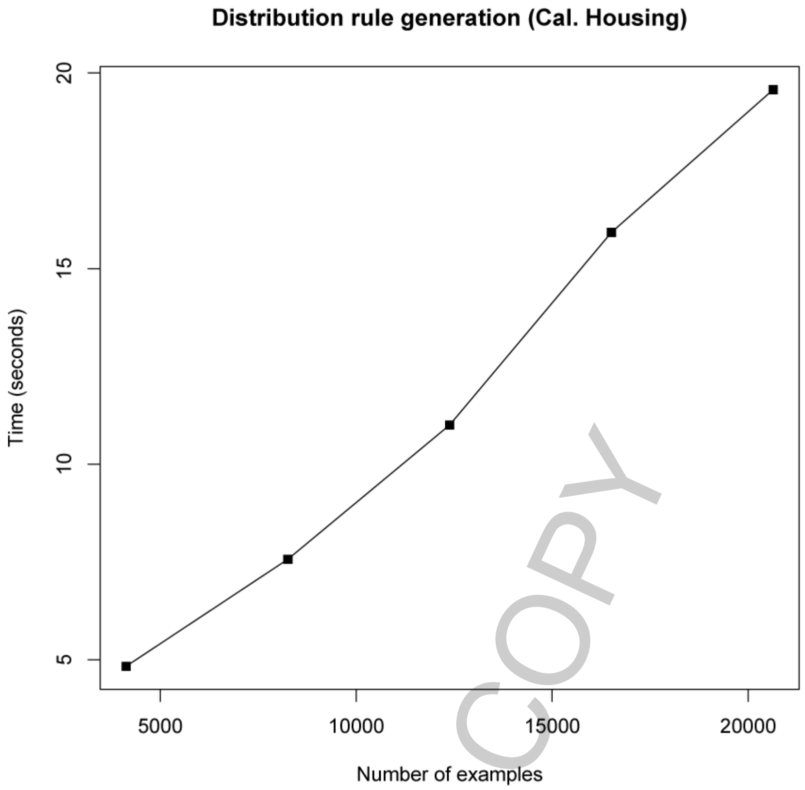

Fig. 6. Scalability of DR generation with the number of examples (minimal support=0.05, data set "Cal. Housing").

grows. We believe that the current implementation of the "intervals" method, despite essentially linear, can still be algorithmically optimized. The proposed interval smoothing technique, with interval and gap elimination, can certainly be much improved. The current proposal is a very simple one and relies on a user defined parameter $min.imp$. This parameter has a local impact on leverage, but the user has no control on the effect of the global leverage loss caused by interval and gap removal.

Another relevant issue regarding scalability and efficiency is the possibility of limiting the number of different values in the POI. Instead of the full set of values, we would use some sort of fine grained clustering. This would yield approximate results in terms of placing interval bounds. However, we have preliminary indications that these would be very good approximations. The main advantage of such an approximation is that computation time would be considerably limited since, as we have shown, MLR generation time grows linearly with the number of different values of the POI.

The interest measure "leverage" has convenient properties that enable the efficient algorithmic discovery of maximal leverage rules with dynamic discretization. In the case of leverage, the optimality criterion is easily defined. Although lift is probably the most intuitive interest measure for association rules, it does not seem to have a direct mathematical relation with the KS statistic. However, maximal lift rules can be discovered using either a post-processing step or by implementing directly an interval determination step within the discovery procedure. The problem, in the case of lift, is that maximal lift rules may have very low support (coverage). This is similar to what happens with optimal confidence rules which can only be found for a fixed value of support [9].

With respect to the state of the art, our proposal of Maximum Leverage Rules discovery is the only that optimizes *leverage* and *added value*. Most approaches presented in section 2 are heuristic. Some

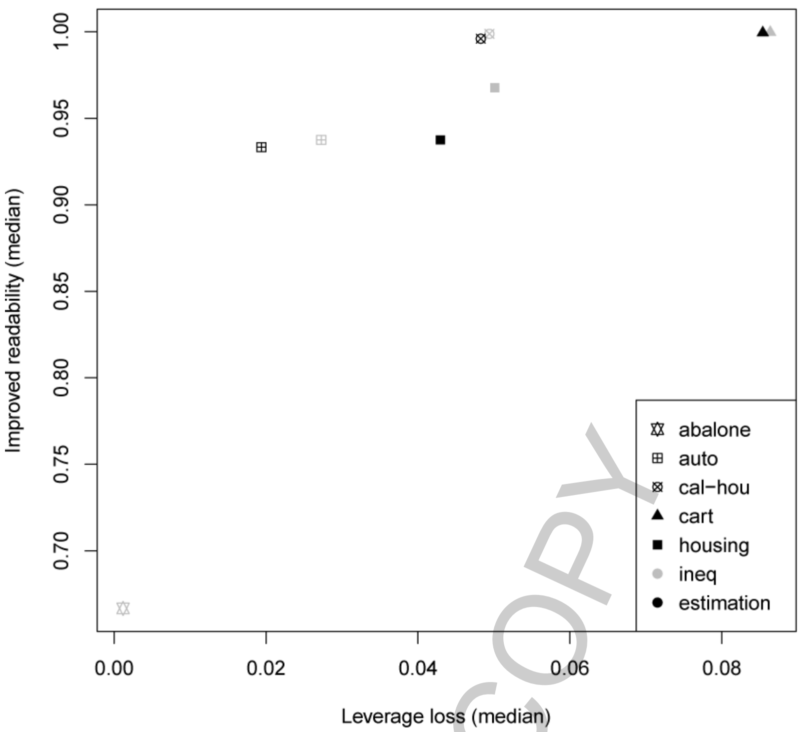

Fig. 7. Improved readability vs. leverage loss (min. coverage = 0.1). Grey points represent the inequality method and dark points the estimation method. Each dataset is represented by a different shape.

of these are able to discover rules in a general form, involving conditions on categorical and numerical attributes, but don't guarantee any optimal results. The approaches of Fukuda et al. [9] and [22] provide optimal rules under the support/confidence setting. In these cases, however, discovery is restricted to certain forms of rules. In the case of [9] only one interval condition is allowed in the antecedent. In the case of Rastogi and Shim [22] the user must supply rule templates to restrict search.

11. Conclusion

We have seen how to generate maximal (optimal) leverage rules with interval conditions in the consequent. Our proposal is highly scalable and avoids pre-discretization of the property of interest. We have proposed two basic methods: "intervals" and "inequalities" that dynamically find optimal interval bounds. We have proved that "intervals" is the optimal method for positive interval conditions in general. The method "inequalities" is optimal for a specific kind of interval conditions. A third method, "estimation", is sub-optimal but is a very good approximation with some advantages over the optimal method. These advantages are faster rule generation and more easily readable rules (with less interval conditions). The empirically observed leverage loss with respect to the optimal method is not important in general, as we have shown in our experiments (Fig. 7). To improve readability we have also proposed two other techniques: "interval smoothing" and "domain restriction". The first technique, when combined with the method "intervals" is worse than "estimation" both on leverage loss and improved readability and is therefore not interesting. The "domain restriction" technique is less lossy but yields more complex rules than “estimation”. We have proposed MLR approaches that take distribution rules as input. This enables modularity of the processes and allows MLR generation to immediatly benefit from DR concepts and filters. However, a more direct reimplementation would certainly result in a more efficient knowledge discovery program.

In terms of scalability, it is important to notice that MLR generation time (per rule) grows linearly with the number of examples and the number of different values of the property of interest. Our implemented versions are quite capable of dealing with tens of thousands of cases and tens of thousands of different POI values for a minimum coverage of 1%.

Future developments of the concept and algorithms for maximal leverage rules are as follows: most MLR methods are implemented in R and should be migrated to java. Algorithmic optimizations are still possible. Conceptually, we intend to parameterize the language bias so that we can find optimal rules with exactly or at most k conditions, for a given k. Another important research issue is the dynamic discretization of the attributes in the antecedent to avoid antecedent pre-discretization.

Acknowledgements

This work was partially supported by the FCT project MORWAQ (PTDC/EIA/68489/2006) and by Fundação Ciência e Tecnologia, FEDER e Programa de Financiamento Plurianual de Unidades de I & D. Special thanks to Brett Drury for giving some suggestions regarding the wording of two paragraphs.

References

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A.I. Verkamo, Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI/MIT Press, 1996.

[2] Y. Aumann and Y. Lindell, A statistical theory for quantitative association rules, *Journal of Intelligent Information Systems* **20**(3) (May 2003), 255–283.

[3] S. Ayubi, M.K. Muyeba, A. Baraani-Dastjerdi and J.A. Keane, An algorithm to mine general association rules from tabular data, *Inf Sci* **179**(20) (2009), 3520–3539.

[4] P.J. Azevedo, CAREN – class project association rule engine. http://www.di.uminho.pt/∼pja/class/caren.html, 2009.

[5] S. Brin, R. Motwani, J.D. Ullman and S. Tsur, Dynamic itemset counting and implication rules for market basket data. In Peckham [19], pages 255–264.

[6] D.W.-L. Cheung, L. Wang, S.-M. Yiu and B. Zhou, Density-based mining of quantitative association rules. In T. Terano, H. Liu, and A. L. P. Chen, editors, *PAKDD*, volume 1805 of *Lecture Notes in Computer Science*, Springer, 2000, pages 257–268.

[7] W.J. Conover, *Practical Nonparametric Statistics – Third Edition*, John Wiley & Sons, New York, 1999.

[8] 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.

[9] T. Fukuda, Y. Morimoto, S. Morishita and T. Tokuyama. Mining optimized association rules for numeric attributes, *J Comput Syst Sci* **58**(1) (1999), 1–12.

[10] T. Hastie, R. Tibshirani and J.H. Friedman, *The Elements of Statistical Learning*, Springer, August 2001.

[11] A.M. Jorge, P.J. Azevedo and F. Pereira, Distribution rules with numerical properties of interest. In *Proceedings of Principles of Data Mining and Knowledge Discovery* (*PKDD-06*), LNAI. Springer-Verlag, 2006.

[12] Y. Ke, J. Cheng and W. Ng, Mic framework: An information-theoretic approach to quantitative association rule mining, in: *ICDE*, L. Liu, A. Reuter, K.-Y. Whang and J. Zhang, eds, IEEE Computer Society, 2006, p. 112.

[13] Y. Ke, J. Cheng and W. Ng, An information-theoretic approach to quantitative association rule mining, *Knowl Inf Syst* **16**(2) (2008), 213–244.

[14] B. Lent, A.N. Swami and J. Widom, Clustering association rules. In W. A. Gray and P.-Å. Larson, editors, *ICDE*, pages 220–231. IEEE Computer Society, 1997.

[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*, New York, NY, USA, 1999, pages 125–134. ACM Press.

[16] J. Mendes-Moreira, C. Soares, A.M. Jorge and J.F. de Sousa, The effect of varying parameters and focusing on bus travel time prediction, in: *PAKDD*, T. Theeramunkong, B. Kijsirikul, N. Cercone and T.B. Ho, eds, volume 5476 of *Lecture Notes in Computer Science*, Springer, 2009, pp. 689–696.

[17] C.J. Merz and P. Murphy, UCI repository of machine learning database. http://www.cs.uci.edu/∼mlearn, 1996.

[18] R.J. Miller and Y. Yang, Association rules over interval data. In Peckham [19], pages 452–461.

[19] J. Peckham, editor. *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, 13–15 May 1997, Tucson, Arizona, USA*. ACM Press, 1997.

[20] G. Piatetsky-Shapiro, Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, AAAI/MIT Press, 1991, pages 229–248.

[21] R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-00-3.

[22] R. Rastogi and K. Shim, Mining optimized association rules with categorical and numeric attributes, *IEEE Trans Knowl Data Eng* **14**(1) (2002), 29–50.

[23] S. Sahar and Y. Mansour, An empirical evaluation of interest-level criteria. In *SPIE Conference on Data Mining and Knowledge Discovery, Orlando, FL*, 1999, pages 63–74.

[24] A. Salleb-Aouissi, C. Vrain and C. Nortet, Quantminer: A genetic algorithm for mining quantitative association rules, in: *IJCAI*, M.M. Veloso, ed., 2007, pp. 1035–1040.

[25] 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*, New York, NY, USA, 1996, pages 1–12. ACM Press.

[26] P.-N. Tan, V. Kumar and J. Srivastava. Selecting the right objective measure for association analysis, *Inf Syst* **29**(4) (2004), 293–313.

[27] J.M. Vázquez, J.L. Á. Macías and J.C.R. Santos, An evolutionary algorithm to discover numeric association rules. In *SAC*, ACM, 2002, pages 590–594.

[28] K. Wang, S.H.W. Tay and B. Liu, Interestingness-based interval merger for numeric association rules. In *KDD*, 1998, pages 121–128.