A rules reduction algorithm based on significance measure

R. Abadía[1], M. Almiñana[2], L. F. Escudero[3], A. Pérez-Martín[2], A. Rabasa[2] & L. Santamaría[2]
[1]*Departamento de Ingeniería, Universidad Miguel Hernández, Spain*
[2]*Instituto Centro de Investigación Operativa (CIO), Universidad Miguel Hernández, Spain*
[3]*Departamento de Estadística e Investigación Operativa, Universidad Rey Juan Carlos, Spain*

Abstract

A lot of rules systems generated from decision trees (like CART, ID3, C4.5, etc) or from counting frequencies direct methods, usually provides non-significant or even contradictory rules. Most existing papers on the subject reach very important reductions over generated rules sets by searching and removing redundancies and conflicts and simplifying similarities between them.

In this paper we propose an algorithm (RBS: Reduction Based on Significance) for allocating a significance value to each rule in the system. This significance value may be used by experts to point out which of these rules must be considered preferable and to understand the exact correlation degree between different rule attributes. The significance is calculated from support and confidence parameters. For each rule, if its support is over a minimum level and its confidence is into a critical interval, its significance ratio is calculated by the algorithm. Thus, the rules space is divided according to these critical boundaries which are calculated by an incremental method. Finally, the significance function is defined in each of these intervals.

Like other rules reduction methods, our approach can also be applied to rules sets generated from decision trees or frequency counting algorithms, in an absolutely independent way and after the rules set was created. So, our RBS algorithm does not change the original accuracy of the rules.

The proposed method has been executed over three different data sets: two of them belong to UCI (University of California, Irvine) standard repository and the third is a real irrigation data set provided by the users. The validity of our reduction approach on the later data set is supervised and contrasted by experts. The computational experience provided in this paper supplies rules sets more reduced, ordered and easily understandable than the original ones.

Keywords: classification rules, reduction, significance measures, support, confidence, regions of significance.

WIT Transactions on Information and Communication Technologies, Vol 40, © 2008 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA080071

1 Introduction

Rules Systems generated from Counting Frequency Methods (like A Priori) or even from Decision Trees (like ID3, C4.5, etc.) are used in several contexts with excellent results. Although heuristics and pruning criteria are incorporated to these generating rules methods, often the provided Rules Sets are too large and disordered. It makes the expert's work (rules interpretation) really difficult and interpretation would become a really hard task. Most existing studies reach very important reductions over the rule set by locating and deleting redundant and conflictive rules, or by simplifying based on similarities found between them. In this paper we purpose a rules goodness measure, rule significance (*rs*) that is calculated from antecedent support and rule confidence values by significance domains. This measure is located into defined intervals so the expert can use it not only as a significance scaled indicator but also as a rule type label. In section 2 we present the formal problem definition, including problem domain, rule, support and confidence concepts. Functional principles of most important Rules Generation Systems are referred to, and most used goodness measures are described. The objectives of this study are specified and, finally, a detailed overview of algorithm is offered.

The computational experiments are described in section 3. It includes a brief description of used experimental data sets. The first of these data sets is a telemetric file of an irrigation network in Southern Spain. This network must be modelled from deduced rules in order to forecast their function in extreme demand situations. Thus, network schedulers know, in advance, critical situations such as burst pipes (caused by excessive pressure) or repeated non-attended user's demands. In order to give a major consistency to obtain reduction rules results, the algorithm is tested over two more data sets (mushrooms classification and animal classification) belonging to standard UCI [1] repository. Furthermore, we show three contingency tables used for experiments and their respective result tables containing reduction ratios reached and significance regions for each rule. Finally, we show an expert interpretation of reduced rule set in the irrigation described domain.

Conclusions derived from this study are given in section 4, where objectives reaching is examined and future research lines are proposed.

2 Problem description

2.1 Rules, support and confidence concepts

Let D be a data set containing one or more records t_i with $i = 1,...,N$ and a group of attributes or variables v_j, with $j = 1,...,m$.

D is a matrix where every row correspond to one record or instance of m attributes placed in columns:

$$D = (d_{ij})_{N \times m} \tag{1}$$

Let $RS_{A\to C}$ be the Rule Set where a rule r_k^{AC}, with $k = 1,...,n$, is defined by $A \to C$, being A the rule antecedent and C the rule consequent.

$$RS_{A\to C} = \left\{ r_1^{AC},...,r_n^{AC} \right\} \tag{2}$$

Rule antecedent A is defined as a tupla consisting of combinations of one or more attributes (maximum *m-1* attributes).

Rule consequent C is defined as a tupla with only one attribute called class attribute.

Without lost of generality, columns in matrix D can be re-ordered, such as *m-1* first columns corresponds to *m-1* attributes that could form the rule antecedent. We call X this subset. The m column that corresponds to consequent is called Y subset.

So, we define an antecedent as follows:

$$A =< v_{s_1},...,v_{s_P} > \tag{3}$$

With $v_{s_\bullet} \in X$ and $P \le m-1$.

The consequent is defined as follows:

$$C =< v_m > \tag{4}$$

An attribute domain is defined as a set of different possible values that can reach this attribute,

$$Dom(v_j) = \left\{ \bigcup_{i=1}^{N} d_{ij} \setminus d_{ij} \in v_j \right\} \tag{5}$$

The number of possible antecedents in a Rule Set is given by [[A]]:

$$[[A]] = \prod_{p=1}^{P} \left| Dom(v_{s_p}) \right| \tag{6}$$

Where $\left| Dom(v_{s_p}) \right|$ is the $Dom(v_{s_p})$ cardinality.

The number of possible consequents in a Rule Set is given by [[C]]:

$$[[C]] = \left| Dom(v_m) \right| \tag{7}$$

Where $\left| Dom(v_m) \right|$ is the $Dom(v_m)$ cardinality.

Antecedent rule support, $sup(A)$, is defined as follows:

$$sup(A) = \frac{[[A]]}{\left| RS_{A\to C} \right|} = \frac{[[A]]}{n} \tag{8}$$

Where $\left| RS_{A\to C} \right|$ is the Rule Set cardinality.

Rule confidence, $conf(A \to C)$, is defined as follows:

$$conf(A \to C) = \frac{[[A \to C]]}{[[A]]} \tag{9}$$

Where $[[A \to C]]$ is the number of rules with antecedent A and consequent C, and it is given by:

$$[[A \to C]] = \left(\prod_{p=1}^{P} \left| Dom(v_{s_p}) \right| \right) \cdot \left| Dom(v_m) \right| \tag{10}$$

Thus a rule has two properties: Region and Significance:

- Region: $r_k^{AC}.REG$ is the region of significance ($REG0$, $REG1$, $REG2$, $REG3$) and is calculated based on antecedent support and rule confidence boundaries.
- Significance: $r_k^{AC}.rs$ is the measure of rule significance and depends on region of significance.

2.2 Rules generation systems

Association Rules Systems shows data behaviour patterns from joint appearances of nominal instances of their attributes. The consequence of these rules is an attribute (or a set of them) over the data. The Counting Frequencies Method tries to find frequent item-sets by coverage, requiring a minimum support over the rules while they are built as explained by Hernández et al. [2]. Several algorithms focus on these kind of problems, but the most important one is the well known Apriori algorithm presented by Agrawal and Srikant [3]. There are a lot of variations and improvements over this algorithm for example, Apriori TID. Rough Set Theory is another approach that is especially suitable when the data set contains a high number of inconsistencies and lost values as Pawlak [4] and Li and Cercone [5] have showed. It drives to minimal rule sets, where redundant attributes are deleted and only the relevant ones are regarded. Tan and Gru [6] studied this situation.

Classification Rules (on which we focus in this work) can be considered as a particular case of Association Rules where the consequent attribute is unique (so called class attribute). The same class attribute is fixed for every rule set generation, and then all the previously described Association Rules generation methods are exactly suitable for Classification Rules generation too. Nevertheless, Decision Trees are considered to be more efficient and flexible methods in classification tasks because they include heuristics and pruning methods in the tree construction step. All rule sets, generated whether by Counting Frequency methods or Decision Trees methods, are susceptible to be reduced.

2.3 Rules goodness measures and existing reduction methods

In order to reduce a rules set, a rule goodness (or interest) measure becomes necessary. That is the correlation level existing between a given set of antecedent and consequent values. Agrawal et al. [7] focused on support value (rule support and antecedent support) as the most used indicator. Lately, this based on support measure evolved to consider support/confidence delimited spaces as Bayardo and Agrawal [8] has shown. The confidence value itself is considered to be a good precision measure of the rules. *Laplace* and *Conviction* are variations over confidence.

The *significance rule (sr)* measure, proposed in this work, is an interval-defined function of rule confidence and support antecedent over significance domains as we show in section 2.5.

Tan et al. [9] offers an exhaustive and comparative study of several measures into a really well referenced framework. In their study the central measures are: *Interest (I)* and *k-coefficient* and their respective variations.

Statistical independence between attributes is measured by *Interest, I* based on antecedent and consequent probabilities. Different variations of *I* are proposed by Tan et al. [9]: *Statistical Correction* to *I* is based on Bayes Theorem, *Certainty Factor*, *Collective Strength* (based on support measure only, but it considers negative correlations too), *Added Value* variation and *Piatesky-Shapiro*. Concordance degree between two attributes is given by *k-coefficient*. The most important variations to *k* are: *Entropy*, *J-Measure* and *Gini-Index* and they are based on probability distributions and the Entropy concept itself. Support and Confidence are frequently used to select the most important rules. A minimum Support and Confidence values are usually demanded by users.

In this paper we propose a new criteria to rule sets reduction based on selecting rules into different support/confidence regions defined by optimal boundaries of these measures which are data-dependent. Another regions approach is presented by Riquelme et al. [10] but the goal is deleting data examples from training data set. Several studies focus on over-specialization problems and reduction methods over redundancy and similarity concepts. Likewise some studies delete the conflicts (contradictory rules) and the inconsistencies (rules that do not obey data set prerequisites) from the final Rule Set.

2.4 Objectives

This study focuses on two main objectives:

- Obtaining reduced Rule Systems, ordered by significance: We define rule significance measure, *rs*, useful to filter the most significant rules and to reduce the final Rules System.
- Classifying rules by significance domains: *rs* is calculated using an intervals-function. Also, the *rs* value allocates each rule in its significance region.

2.5 Overview of the proposed reduction algorithm

We use a bi-dimensional space to allocate every rule on a given rule set. One coordinate is given by rule antecedent support (called support) and the other one is given by rule confidence. Thus, all rules are placed in a particular point in the space domain that must be divided in significance regions. So we need to establish the minimal and maximal limits of antecedent support and rule confidence. In the first step we calculate E_C and E_S as principal axis (Figure 1) that will be used in second step to establish the minimal (E_{Ci}, E_{Si}) and maximal (E_{Cs}, E_{Ss}) axis (Figures 2 and 3). In the third step for every rule, its significance region *REG* is assigned and its rule significance value *rs*, is calculated. In the fourth step the boundaries are adjusted to their possible limits in function of data

set, minimizing significance regions (Figure 4). Rules do not change of region (compare Figures 3 and 4). Finally, in step 5, rules in region 0 are removed and the other ones are re-ordered attending to the *rs* value.

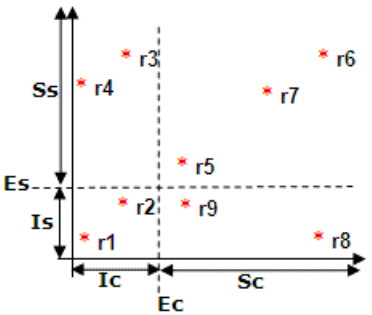

Figure 1: Example of rules distribution.

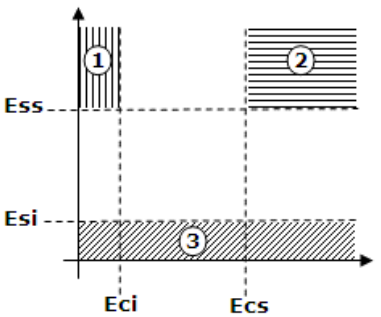

Figure 2: Example of significance regions division.

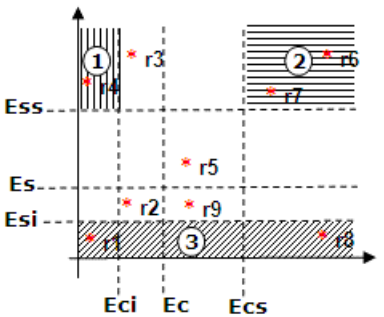

Figure 3: Example of boundary axis initializing.

Let $RS_{A\to C} = \{r_1^{AC},...,r_n^{AC}\}$ be the Rules Set, and $r_i^{AC} \in RS_{A\to C}$ is a rule $A \to C$ where A is the rule antecedent and C is the rule consequent, with $i = 1,...,n$

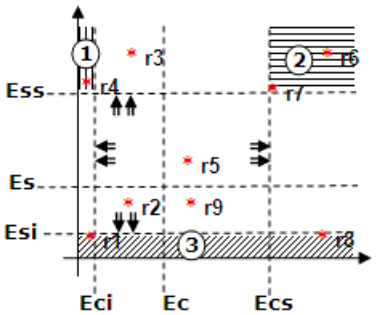

Figure 4: Example of boundary axis adjusting.

Every rule on Rules System belongs to only one of the defined exclusive regions.

3 Computational experiment

3.1 Data sets

Three data sets have been used to apply the reduction algorithm: Irrigation, Mushrooms and Zoo.
- Irrigation data set: This real-life data set consists of telemetric values of a tree-structured irrigation network in southern Spain. We use this data set to interpret the goodness of the proposed algorithm, with expert evaluation of filtered and classified rules. There are 6 categorical selected attributes and 14,843 rows.
- Mushrooms data set: provided by UCI standard repository. It has 6 categorical selected attributes describing mushroom morphology, and 8124 rows.
- Zoo data set: provided by UCI standard repository, with 6 categorical selected attributes (most of them binary attributes) and only 101 rows.

3.2 Empirical results

Defined contingency table for irrigation data set is shown in Table 1. The first row in contingency table shows the number of possible values for each attribute, the second and third rows are the attributes notation and interpretation. Next, every row, labelled as Ri, is a rule set consisting of attributes containing "1", as antecedent and the consequent placed in the last column.

Thus for example, in Table 1, R3 is the rule set containing ETO_D1 and ETO_D2 as the rule antecedents and Q_D as the rule consequent. For this rule set R3, a counting frequency method generates the rules shown in Table 2. The RBS algorithm fills in the columns *REG* and *rs* for every rule.

The RBS algorithm in the fifth step, removes all rules in REG0 and re-orders the rest.

For Mushrooms and Zoo data sets another two contingency tables (similar to Table 1) are generated, and every row on them generates rule set tables (similar to Table 2).

Table 1: Irrigation contingency table (extract).

[[Ci]]	4	5	5	5	5	7
Ci	C1	C2	C3	C4	C5	C6
	H_D	TMED_D1	TMED_D2	ETO_D1	ETO_D2	Q_D
R1	0	0	0	0	1	1
R2	0	0	0	1	0	1
R3	0	0	0	1	1	1
…	…	…	…	…	…	1
R31	1	1	1	1	1	1

Table 2: Rule Set R3, irrigation (extract).

ETO_D1	ETO_D2	Q_D	sop(A)	conf(A->C)	REG	rs
A	A	MB	0,09714	0,13431	REG0	0,00000
N	A	MB	0,13021	0,15675	REG0	0,00000
B	A	MB	0,03255	0,12698	REG3	0,00413
N	N	N	0,45521	0,30959	REG2	1,14093
N	N	o	0,45521	0,04143	REG1	-1,01886
B	N	o	0,03255	0,01389	REG3	0,00045
…	…	…	…	…	…	…

3.3 Irrigation rules. Results interpretation by expert

The antecedents and consequent attributes are shown in Table 3.

Table 3: Antecedents and consequent values on irrigation data set.

H_D	used turn for irrigation	T1, T2
TMED_D1	medium temp. on the previous day	MA, A, N, B, MB
TMED_D2	medium temp. two days before	MA, A, N, B, MB
ETO_D1	evapotransp. on the previous day	MA, A, N, B, MB
ETO_D2	evapotransp. two days before	MA, A, N, B, MB
Q_D (cons)	water consumption level	MA, A, N, B, MB, o, X
T1..T2: Turns 1st (morning) and 2nd (evening). MA: Very High; A: High; N: Medium; B: Low; MB: Very Low; o: missing; X: Error		

Evaluating final axis and *rs* values for every rules set, the most important expert interpretations about influence on water consumption are the following:

- Influence of evapotranspiration one day and two days before (R1, R2) over water consumption are very similar. If both values are considered (R3) the correlation level with consumption is analogous. We consider this is because evapotranspiration variations are minimal on time interval.

- Temperature one day and two days before (R4 and R8) have a very similar influence on water consumption. Considering both temperature values (R12) correlation does not increase.
- There are more significant rules correlating temperature with consumption than evapotranspiration with consumption. Normal levels of temperature implying normal levels of consumption is the most reliable rule. There are very few rules with high values of temperature, hence these rules have very low support. This data interpretation agree with farmer behaviour, because temperature is easily perceptible than evapotranspiration. That is the reason because water consumption in August is higher than in July while evapotranspiration is higher in July than in August.
- First irrigation turn has the most predictable behaviour, containing normal levels of water consumption the most reliable rule set (R16). There are very few instances of very high consumption in first turn. If temperature one day before is considered with turn (R24), correlation with consumption increases slightly.

4 Conclusions

Rule sets provided by classical rules generation algorithms are usually difficult to interpret. The RBS algorithm reduces and orders the rules into a given rule set when each of them is associated to a rule significance value and it is allocated into a specific significance region, thus it becomes more understandable to the expert. It allows to design and control remote systems safely. This method has been applied to different data sets and it has been interpreted by experts into an irrigation framework successfully, providing reasonable reductions over original rules sets and keeping the most important rules into their respective significance regions and removing only non-relevant rules. One of our papers, with several improvements over the RBS algorithm, is in progress.

References

[1] UCI Machine Learning; U.S., Univ. of California http://mlearn.ics.uci.edu/
[2] Hernández, J., Ramírez, M.J., Ferri, C., *Introducción a la Minería de Datos*, Pearson Prentice Hall, pp. 287–290, 2004.
[3] Agrawal, R., Srikant, R., *Fast Algorithms for Mining Association Rules*, IBM internal report, IBM Co., 1996.
[4] Pawlak, Z., Rough Sets and Intelligent Data Analysis, Elsevier. *Information Sciences* 147, pp. 1–12, 2002.
[5] Li, J., Cercone, N., *Introducing A Rule Importance Measure*, (Chap. 8), Transactions on Rough Sets V, Springer Berlin, pp. 167–189, 2006.
[6] Tan, S., Gu, J., An Efficient Rules Induction Algorithm for Rough Set Classification, *DS LNAI,* Springer-Verlag Berlin, pp. 330–337, 2004.
[7] Agrawal, R., Imielinski, T., Swami, A., Mining Association Rules Between Sets of Items in Large Databases, *ACM Sigmod Conf.*, pp. 207–216, 1993.

[8] Bayardo, R.,Agrawal, R., Mining the Most Interesting Rules, *Proc. 5th. ACM SIGKDD Internat. Conf. on Knowledge Discovery*, pp. 145–154, 1999.
[9] Tan, P.N., Kumar, V., Srivastava, J., Selecting Right Objective Measure for Association Analysis, *Information Systems*, Elsevier. 29, pp. 293–313, 2004.
[10] Riquelme, J.C., Aguilar-Ruiz, J.S., Toro, M., Finding Representative Patterns with Ordered Projections, *Pattern Recognition*, 36, Pergamon, pp. 1009–1018, 2003.