Mining Association Rules with respect to Support and Anti-support-Experimental Results

R. Słowiński[1,2], I. Szczęch[1], M. Urbanowicz, S. Greco[3]

1 Inst. of Computing Science, Poznań University of Technology,60-965 Poznań, Poland
{Roman.Slowinski, Izabela.Szczech} @cs.put.poznan.pl
2 Inst. for Systems Research, Polish Academy of Sciences, 01–447 Warsaw, Poland
3 Faculty of Economics, University of Catania, Corso Italia, 55, 95129 Catania, Italy
salgreco@mbox.unicit.it

Abstract. Evaluating the interestingness of rules or trees is a challenging problem of knowledge discovery and data mining. In recent studies, the use of two interestingness measures at the same time was prevailing. Mining of Pareto-optimal borders according to support and confidence, or support and anti-support are examples of that approach. Here, we consider induction of "*if..., then...*" association rules with a fixed conclusion. We investigate ways to limit the set of rules non–dominated wrt support and confidence or support and anti-support, to a subset of truly interesting rules. Analytically, and through experiments, we show that both of the considered sets can be easily reduced by using the valuable semantics of confirmation measures.

Keywords: Association rules, Induction, Support, Anti–support, Confirmation, Confidence, Pareto–optimal border.

1 Introduction

In data mining and knowledge discovery, the discovered knowledge patterns are often expressed in a form of "*if..., then...*" rules. They are consequence relations representing correlation, association, causation etc. between independent and dependent attributes. In order to increase the relevance and utility of selected rules and, thus, also limit the size of the resulting rule set, quantitative measures, also known as interestingness measures, have been proposed and studied (e.g. confidence, support, gain, conviction, lift). Among widely studied interestingness measures, there is, moreover, a group of Bayesian confirmation measures, which quantify the degree to which a piece of evidence built of the independent attributes provides "evidence for or against" the hypothesis built of the dependent attributes [4]. Another approach to evaluation of generated rules concentrates on the use of two different interestingness measures. In this paper,

we show a way to limit the set of rules generated with respect to pairs of measures: support–confidence and support–anti–support, by filtering out the rules for which the premise does not confirm the conclusion. This proposition is based on imposing the confirmation perspective on the analyzed two–dimensional evaluations.

The paper is organized as follows. In section 2, there are preliminaries on rules and their quantitative description. In section 3, we investigate the idea and advantages of mining only rules with positive confirmation from Pareto–optimal border with respect to support and confidence. Section 4 concentrates on the proposal of limiting the set of rules generated with respect to support and anti–support. Theoretical considerations are supported by experimental results. The paper ends with conclusions.

2 Preliminaries

Since discovering rules from data is the domain of inductive reasoning, its starting point is a sample of larger reality often given in a form of a data table. Formally, a *data table* is a pair $S = (U, A)$, where U is a nonempty finite set of objects called *universe*, and A is a nonempty finite set of *attributes* such that $a : U \rightarrow V_a$ for every $a \in A$. The set V_a is a domain of a. A *rule* induced from S is denoted by $\phi \rightarrow \psi$ (read as "*if* ϕ, *then* ψ"). It consists of antecedent ϕ and consequent ψ, called *premise* and *conclusion*, respectively. In this paper, similarly to [2], we consider evaluation of rules with the same conclusion.

2.1 Partial Preorder on Rules in terms of Two Measures

Let us denote by $\preceq_{qt}$ a partial preorder given by a dominance relation on a set X of rules in terms of any two different interestingness measures q and t, i.e. for all $r_1, r_2 \in X r_1 \preceq_{qt} r_2$ if $r_1 \preceq_q r_2$ and $r_1 \preceq_t r_2$. Recall that a *partial preorder* on a set X is a binary relation R on X that is reflexive and transitive. The partial preorder $\preceq_{qt}$ can be decomposed into its asymmetric part $\prec_{qt}$ and its symmetric part $\sim_{gt}$ in the following manner: given a set of rules X and two rules $r_1, r_2 \in X, r_1 \prec_{qt} r_2$ if and only if $q(r_1) \leq q(r_2) \wedge t(r_1) < t(r_2)$, or $q(r_1) \leq q(r_2) \wedge t(r_1) < t(r_2)$, moreover, $r_1 \sim_{qt} r_2$ if and only if $q(r_1) = q(r_2) \wedge t(r_1) = t(r_2)$. If for a rule $r \in X$ there does not exist any rule $r' \in X$, such that $r \prec_{qt} r'$ then r is said to be non–dominated (i.e. Pareto–optimal) wrt interestingness measures q and t. A set of all non–dominated rules wrt q and t is also referred to as an *q–t Pareto–optimal border*.

2.2 Monotonicity of a Function in its Argument

Let x be an element of a set of rules X and let $g(x)$ be a real function associated with this set, such that $g : X \rightarrow \mathbf{R}$. Assuming an ordering relation $\succ$ in X, function g is said to be monotone (resp. anti–monotone) in x, if for any $x, y \in X$, relation $x \succ y$ implies that $g(x) \geq g(y)$ (resp. $g(x) \leq g(y)$).

2.3 Support, Confidence and Anti–support Measures of Rules

Among measures very commonly associated with rules induced from information table S, there are *support* and *confidence*. The *support* of condition ϕ, denoted as $sup(\phi)$, is equal to the number of objects in U having property ϕ. The support of rule $\phi \rightarrow \psi$, denoted as $sup(\phi \rightarrow \psi)$, is the number of objects in U having property ϕ and ψ.

The *confidence* of a rule (also called *certainty*), denoted as $conf(\phi \rightarrow \psi)$, is defined as: $conf(\phi \rightarrow \psi) = \frac{sup(\phi \rightarrow \psi)}{sup(\phi)}$, $sup(\phi) > 0$.

Anti–support of a rule, denoted as $anti - sup(\phi \rightarrow \psi)$, is equal to the number of objects in U having the property ϕ but not having the property ψ. Thus, anti–support is the number of counter–examples, i.e. objects for which the premise ϕ evaluates to true but which fall into a class different than ψ. Note that anti–support can also be regarded as $sup(\phi \rightarrow \neg\psi)$.

2.4 Bayesian Confirmation Measures

Bayesian confirmation measures constitute a group of interestingness measures that quantify the degree to which a premise ϕ provides "support for or against" a conclusion ψ [4]. Under the "closed world assumption" adopted in inductive reasoning, and because U is a finite set, a confirmation measure denoted by $c(\phi \rightarrow \psi)$ is required to satisfy the following definition:

$$c(\phi \rightarrow \psi) = \begin{cases} > 0 \text{ if } conf(\psi \rightarrow \phi) > sup(\psi)/|U|, \\ = 0 \text{ if } conf(\psi \rightarrow \phi) = sup(\psi)/|U|, \\ < 0 \text{ if } conf(\psi \rightarrow \phi) < sup(\psi)/|U|. \end{cases} \quad (1)$$

For the confirmation measures a desired property of monotonicity (M) was proposed in [5]. This monotonicity property says that, given an information system S, a confirmation measure is a function non–decreasing wrt

$sup(\phi \rightarrow \psi)$and $sup(\neg\phi \rightarrow \neg\psi)$, and non–increasing wrt $sup(\neg\phi \rightarrow \psi)$ and $sup(\phi \rightarrow \neg\psi)$. Among confirmation measures that have property (M) there is e.g. confirmation measure f [4] defined as:
$f(\phi \rightarrow \psi) = \frac{conf(\psi\rightarrow\phi)-conf(\neg\psi\rightarrow\phi)}{conf(\psi\rightarrow\phi)+conf(\neg\psi\rightarrow\phi)}$.

2.5 A Brief Description of a Dataset and Experiments

For the purpose of these experiments we used a dataset *adult* [7]. The number of analyzed instances reached 32 561. They were described by 9 nominal attributes differing in domain sizes. Missing values were substituted by the most frequently appearing one. Two experiments were conducted: one generating rules wrt support and confidence, and the second one generating rules according to support and anti-support. Both of them proceeded in a two step Apriori-like framework:

- firstly, all conjunctions of elementary conditions (i.e. itemsets) that exceeded the minimum rule support threshold (i.e. frequent itemsets) were found;
- secondly, those frequent itemsets were used to generate association rules having either confidence or anti-support not smaller than the user's defined threshold.

The detailed description as well as the efficiency comparison of the applied algorithms (based on [1,6]) can be found in [9]. Throughout the experiment, the value of support was expressed as a relative value between 0 and 1. During the frequent itemset generation phase, only itemsets that exceeded 0.15 support threshold were approved. No confidence nor anti-support thresholds were applied in order to show the complete Pareto-optimal border exceeding the support threshold.

3 Support–Confidence Pareto–optimal Border

Bayardo and Agrawal [2] proposed evaluation of the set of rules in terms of two popular interestingness measures being rule support and confidence. They have proved that for a class of rules with fixed conclusion, the support–confidence Pareto–optimal border includes optimal rules according to several different interestingness measures, such as gain, lift, conviction, etc. Thus, by solving an optimized rule mining problem wrt rule support and confidence one can identify a set of rules containing most interesting (optimal) rules according to several interestingness measures. However, despite those valuable features of the support–confidence

Pareto–optimal border, one cannot, in general, claim that the set of dominated rules is without interest. It can be e.g. due to the fact that in order to cover the analyzed concept (decision class) one has to use both dominated and non–dominated rules. Of course, a user can set some thresholds both on rule support and confidence, but still taking under the consideration both dominated and non–dominated rules can result in a large, difficult to analyze set of rules. Hence, we propose a way to limit the set of the analyzed rules by using the valuable semantic of confirmation measures.

3.1 The Confirmation Perspective on the Support–Confidence Evaluations

The advantages of semantic utility of confirmation measures in general over confidence have been widely studied in [3,5]. Thus, we find it valuable to impose the confirmation perspective on the analyzed support–confidence evaluations and limit in this way the set of rules to be analyzed. It has been analytically proved in [3] that for a fixed value of rule support, confidence is monotone wrt any confirmation measure having the desired property of monotonicity (M) proposed in [5].

Let us observe that according to definition (1) of $c(\phi \rightarrow \psi)$, we have:

$$conf(\phi \rightarrow \psi) > 0 \Leftrightarrow conf(\phi \rightarrow \psi) > \frac{sup(\psi)}{|U|} \tag{2}$$

Since, we limit our consideration to rules with the same conclusion, then $|U|$ and $sup(\psi)$ should be regarded as constant values. Thus, (2) shows that rules laying under a constant, expressing what percentage of the whole dataset is taken by the considered class ψ, are characterized by negative values of confirmation (see Fig. 1). For those rules ψ is satisfied less frequently when ϕ is satisfied rather than generically.

It is also interesting to investigate a more general condition $c(\phi \rightarrow \psi) \geq k$, $k \geq 0$, for some specific confirmation measures. In the following, we consider confirmation measure $f(\phi \rightarrow \psi)$.

Theorem 1. (See proof in [8])

$$f(\phi \rightarrow \psi) \geq k \Leftrightarrow conf(\phi \rightarrow \psi) \geq \frac{sup(\psi)(k+1)}{|U|-k(|U|-2sup(\psi))} \tag{3}$$

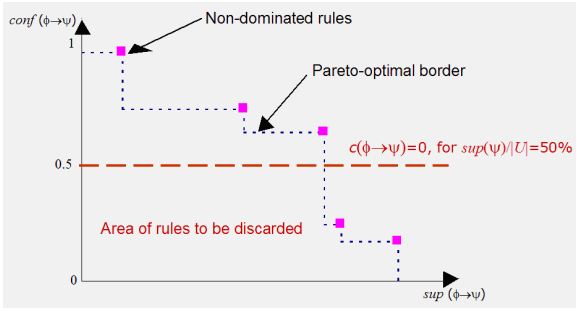

Fig. 1. An example of a constant line representing $c(\phi \rightarrow \psi) = 0$ in a support–confidence space. Rules laying under it should be discarded from further analysis

3.2 Experiments with Rule Induction with respect to Support and Confidence

On Fig. 2 we show association rules generated, according to mentioned thresholds for the conclusion: workclass='Private'. This class contains information about people working in a private sector. Rules are presented in a support–confidence space.

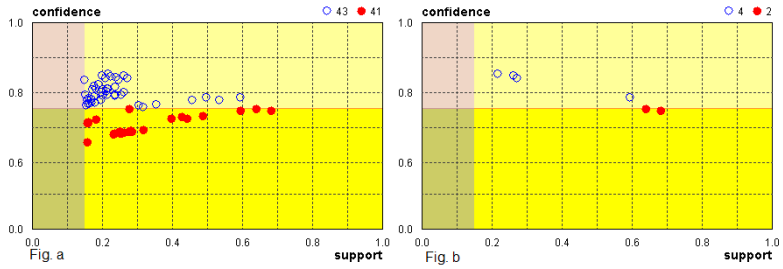

Fig. 2. Rules generated for a conclusion workclass='Private' with positive (empty circles) and non–positive confirmation measure value (solid circles) in a support–confidence space. Fig. a – all generated rules, Fig. b – the Pareto-optimal border only.

This experiment makes it evident that in practice even rules with high value of confidence (exceeding even 0.7) can be found useless as their premise disconfirms the conclusion (those rules are marked by solid circles). It is therefore clear, that the semantic scale of the confidence measure is not enough and that confirmation measures are very much needed. Sometimes even rules from the Pareto–optimal border need to be discarded from further analysis as their value of confirmation is non–positive. On Fig. 2 a constant line was placed separating the rules with positive confirmation (situated above the line) from those with non–positive confirmation (situated below the line). Fig. 2 visualizes result (2) and says how big (in comparison to the whole dataset) is the considered class of

rules for the analyzed conclusion workclass='Private'. Illustrations for other classes can be found in [8,9]. By imposing the confirmation perspective, the number of rules to be analyzed by the domain expert can be significantly reduced. For the conclusion being worklass='Private', 41 out of 84 rules had to be discarded for disconfirming the conclusion. Tab. 1 shows results for other conclusions that we have considered.

Table 1. Information about the percentage of rules with non-positive confirmation in the set of all generated rules for different conclusions.

Considered conclusion	**No. of all rules**	**No. of all rules with non–positive confirm.**	**Reduction percentage**
workclass='Private'	84	41	49%
sex=Male	85	24	28%
income<=50kUSD	87	43	49%

Table 2. Information about the percentage of rules with non-positive confirmation laying on the support–confidence Pareto–optimal border for different conclusions.

Considered conclusion	**No. of all rules on Pareto border**	**No. of all rules with non–positive confirm.**	**Reduction percentage**
workclass='Private'	6	2	33%
sex=Male	6	1	17%
income<=50kUSD	5	1	20%

Tab. 2 shows how many rules with non–positive confirmation laid on the support–confidence Pareto–optimal border for different considered conclusions. Even Pareto–optimal borders, i.e. objectively the best sets of rules, contain rules that are misleading. In some cases, the support–confidence Pareto–optimal border could be reduced by even 33%, like for the conclusion workclass='Private'.

4 Support–Anti–support Pareto–optimal Border

Presentation of association rules in dimensions of rule support and anti–support was proposed in [3]. The idea of combining those two dimensions came from a critical remark towards support–confidence Pareto–optimal border. In [3], it was proved that a rule maximizing a confirmation measure satisfying the property (M) is on the support–confidence Pareto–optimal border only if a specific condition is satisfied. Thus, in general, not all rules maximizing such a measure are on the support–confidence Pareto–optimal border. However, due to valuable semantics of confirmation measures, mining all rules that maximize confirmation measures

with (M), became an interesting problem. The solution is support–anti–support Pareto–optimal border. It was proved in [3] that the best rule according to any of confirmation measures with (M) must reside on the support–anti–support Pareto–optimal border. Moreover, it was pointed out in [3] that the Pareto–optimal border of support–anti–support contains the support–confidence Pareto–optimal border. Despite all good characteristics of the support–anti–support Pareto–optimal border, one can still remain interested in the set of dominated rules. Thus, analyzing whether one can limit the set of rules, by imposing a confirmation perspective on the spport–anti–support evaluations, is interesting.

4.1 The Confirmation Perspective on the Support–Anti–support Evaluations

It has been analytically proved in [3] that for a fixed value of rule support, any confirmation measure $c(\phi \rightarrow \psi)$ having the desired property of monotonicity (M) is anti–monotone (i.e. non–decreasing) wrt anti–support. Let us observe that a simple transformation of definition (1) leads to the following result:

$$c(\phi \rightarrow \psi) \geq 0 \Leftrightarrow anti - sup(\phi \rightarrow \psi) \leq sup(\phi \rightarrow \psi) \left[\frac{|U|}{sup(\psi)} - 1 \right] \qquad (4)$$

Having limited our consideration to rules with the same conclusion, $|U|$and $sup(\psi)$ should be regarded as constant values. Thus, the result (4) shows that a simple linear function bounds rules that are characterized by positive values of confirmation from those with non–positive confirmation values (see Fig. 3).

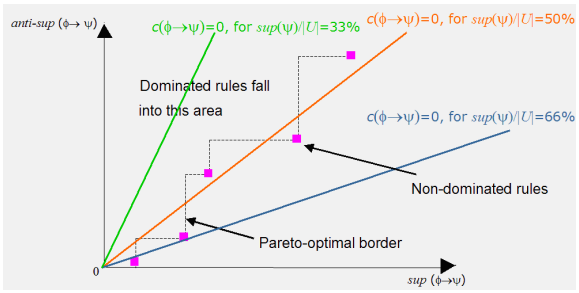

Fig. 3. Three examples of linear functions representing $c(\phi \rightarrow \psi) = 0$ in a support–anti–support space. Lines were drawn according to a set of rules for conclusions different in cardinality. Rules laying above them should be discarded from further analysis.

It is also interesting to investigate a more general condition $c(\phi \rightarrow \psi) \geq k, k \geq 0$. Let us consider again $f(\phi \rightarrow \psi)$.

Theorem 2. (See proof in [8])

$$f(\phi \rightarrow \psi) \geq k \Leftrightarrow anti - sup(\phi \rightarrow \psi) \leq sup(\phi \rightarrow \psi)(U - sup(\psi))\frac{1-k}{(1+k)sup(\psi)} \quad (5)$$

4.2 Experiments with Rule Induction with respect to Support and Anti–support

On Fig. 4, we show association rules generated, according to mentioned threshold, for the conclusion: workclass='Private'.

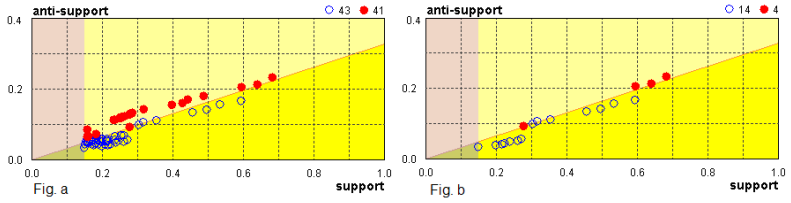

Fig. 4. Rules generated for a conclusion workclass='Private' with positive (empty circles) and non–positive (solid circles) confirmation measure value in a support–anti–support space. Fig. a – all generated rules, Fig. b – the Pareto-optimal border only.

This experiment makes it clear, that despite the valuable properties of support–anti–support Pareto–optimal border, it is necessary to take under consideration also the information brought by the sign of the confirmation measures. Within the Pareto-optimal set presented on Fig. 4, 22% of rules need to be discarded as their value of confirmation is non–positive. On Fig. 4, a linear function was placed separating the rules with positive confirmation (situated under the line) from those with non–positive confirmation. Fig. 4 visualizes result (4). Tab. 3 presents the percentage of rules to be discarded from the support-anti-support Pareto-optimal border. In the conducted experiment the set of rules to be analyzed could be reduced by e.g. about 22% (workclass='Private').

Table 3. Information about the percentage of rules with non-positive confirmation laying on the support–anti–support Pareto–optimal border for different conclusions.

Considered conclusion	No. of all rules on Pareto border	No. of all rules with non–positive confirm.	Reduction percentage
workclass='Private'	18	4	22%
sex=Male	8	3	38%
income<=50kUSD	15	4	27%

5 Conclusions

In this paper, we investigated rules induced for a fixed conclusion and evaluated in spaces of support–confidence and support–anti–support. The Pareto–optimal borders of those spaces have some valuable features. However, these worthy features, do not assure that the number of induced rules would not exceed the human user capabilities to analyze them. Inspired by the strength of the semantics of confirmation measures, we show that it is reasonable to limit the set of rules by eliminating those that are characterized by non–positive or small values of confirmation. We have shown analytically that a simple constant line imposed on the support–confidence space bounds the rules with positive values of confirmation measure from those with non–positive confirmation values. This is a very practical result allowing to limit the set of analyzed rules only to those with positive confirmation values, without actually calculating the value of a particular confirmation measure for each of the induced rules. Analogous analysis has been conducted for rules in support–anti–support space. We have shown that a simple linear function separates the rules with positive and non–positive values of confirmation. Again, this is an easy approach to limit the set of analyzed rules. Experimental results show how big the reduction can be.

References

1. R. Agrawal and T.Imielinski. Mining associations between sets of items in massive databases. *Proc.of ACM-SIGMOD Int'l Conf. on Management of Data*, 1993.
2. R.J. Bayardo and R.Agrawal. Mining the most interesting rules. *Proc. of 5th ACM-SIGKDD Int'l Conf. on Knowledge Disc.and Data Mining*, pages 145–154, 1999.
3. I. Brzezińska, S.Greco, and R.Słowiński. Mining pareto-optimal rules with respect to support and anti-support. *Eng. Applic.of Artif. Intelligence,to appear.*
4. B. Fitelson. *Studies in Bayesian Confirmation Theory.* PhD thesis, University of Wisconsin, Madison, 2001.
5. S. Greco, Z.Pawlak, and R.Słowiński. Can bayesian confirmation measures be useful for rough set decision rules? *Eng. Applic.of Artif. Intelligence*, 17:345–361, 2004.
6. J. Han, J.Pei, and Y.Yin. Mining frequent patterns without candidate generation. *Proc. ACM SIGMOD Conference on Management of Data*, pages 1–12, 2000.
7. R.J. Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. *Proc. of the 2nd Int'l Conf. on Knowledge Discovery and Data Mining*, 1996.
8. R. Słowiński, I.Szczech, M.Urbanowicz, and S.Greco. Experiments with induction of assoc. rules wrt support,anti-support. Technical Report RA-018/06, II-PP, 2006.
9. M. Urbanowicz. *Induction of association rules with given support, confidence and confirmation.* PhD thesis, Poznań University of Technology, 2007.