

Objective Novelty of Association Rules: Measuring the Confidence Boost¹

José L Balcázar*

*Departamento de Matemáticas, Estadística y Computación
Universidad de Cantabria
Santander, Spain
joseluis.balcazar@unican.es

Résumé. On sait bien que la confiance des règles d'association n'est pas vraiment satisfaisant comme mesure d'intérêt. Nous proposons, au lieu de la substituer par des autres mesures (soit, en l'employant de façon conjointe a des autres mesures), évaluer la nouveauté de chaque règle par comparaison de sa confiance par rapport á des règles plus fortes qu'on trouve au même ensemble de données. C'est á dire, on considère un seuil "relative" de confiance au lieu du seuil absolute habituel. Cette idée se précise avec la magnitude du "confidence boost", mesurant l'increment relative de confiance prés des règles plus fortes. Nous prouvons que nôtre propose peut remplacer la "confidence width" et le blockage de règles employés a des publications précédentes.

1 Introduction

The outcome of a Data Mining project is expected to offer some degree of novelty. However, to formally study the novelty of Data Mining results is far from being a trivial task. Novelty refers to facts that are somehow unexpected, and therefore some expectation, lower than actually found, must exist, due to some alternative facts or prediction mechanisms.

Here we consider the specific task of Association Rule Mining, and follow up a recent proposal that, as a minimum, each rule should be evaluated for novelty according to the rest of the rules mined, treated as "alternative" mechanism [4], [5]. These works propose the *confidence width* as a measure of a relative form of objective novelty or surprisingness of each individual rule with respect to other rules that hold in the same dataset, and propose also to *block* some rules in case they do not bring in enough novelty with respect to a "blocker" itemset. Our contribution here is a new novelty notion, the *confidence boost*, similar in its syntactic definition to confidence width, but very different in its semantics ; its main trait is that it encompasses at once both the bound on the confidence width and the ability to detect that a rule would be blocked, so that the confidence boost bound embodies both of the bounds proposed in [5].

The notions of confidence width and rule blocking from [5] are similar to the "pruning" proposal from [14], in that the intuition is the same ; also our proposal here follows an analogous intuitive path. Major differences are that, in the proposals we discuss, a large portion of the pruning becomes unnecessary because we work on minimum-size bases (representative rules, proposed in [1], in [13], and in [16]) and, more importantly, that the pruning in [14]

is based on the χ^2 statistic, whereas we will look instead into the confidence thresholds that would make the rule “redundant”. Our notions are also similar to the notion of *improvement*, proposed in [6] (and also briefly discussed in [14]) but bear several key differences with it.

We will denote itemsets by capital letters from the end of the alphabet, and use juxtaposition to denote union, as in XY . For a given dataset \mathcal{D} , consisting of transactions, each of which is an itemset labeled with a unique transaction identifier, we can count the *support* $s(X)$ of an itemset X , which is the cardinality of the set of transactions that contain X . The *confidence* of a rule $X \rightarrow Y$ is $c(X \rightarrow Y) = s(XY)/s(X)$. In all association rules $X \rightarrow Y$ along this paper, we assume that $X \cap Y = \emptyset$; we do allow $X = \emptyset$.

2 Objective Novelty Measures

We propose to measure the novelty of each rule with respect to the rest of the outcome of the same data mining process. Among a number of notions of redundancy among association rules, there is a very precise and natural notion that allows for characterizing irredundant bases of minimum size (see [3]).

Lemma 1 *Let $X_0 \rightarrow Y_0$ and $X_1 \rightarrow Y_1$ be association rules. The following are equivalent :*

1. *In every dataset \mathcal{D} , $c(X_0 \rightarrow Y_0) \geq c(X_1 \rightarrow Y_1)$ and $s(X_0Y_0) \geq s(X_1Y_1)$.*
2. $X_1 \subseteq X_0 \subseteq X_0Y_0 \subseteq X_1Y_1$.

When these cases hold, we say that $X_1 \rightarrow Y_1$ makes $X_0 \rightarrow Y_0$ *redundant*, or also that $X_1 \rightarrow Y_1$ is *logically stronger* than $X_0 \rightarrow Y_0$. For a fixed confidence threshold, those rules that reach it, and are not made redundant by other rules also above the threshold, form the *representative* (or *essential*) *rule basis* for that confidence threshold [1], [13], [16].

2.1 Confidence Width

In [4], [5], the intuition of redundancy is pushed further in order to gain a perspective of novelty of association rules : an irredundant rule in the basis is so because its confidence is higher than what the rest of the basis would suggest ; then, one can ask : “how much higher ?”. This can be evaluated by means of the following definition. For the whole discussion, assume that a dataset \mathcal{D} and a support threshold τ have been fixed : all our rules are assumed to reach that support on \mathcal{D} .

Definition 1 *The confidence width of $X \rightarrow Y$ in \mathcal{D} is $w(X \rightarrow Y) =$*

$$= \frac{c(X \rightarrow Y)}{\max\{c(X' \rightarrow Y') \mid (X \rightarrow Y) \neq (X' \rightarrow Y'), X' \subseteq X, XY \subseteq X'Y'\}}$$

In case the denominator is null, the confidence width is infinite by convention.

To explain this notion, consider a rule $X \rightarrow Y$ of a given confidence, say $c(X \rightarrow Y) = c_0 \in [0, 1]$, and let’s see what happens as we mine the representative basis at a varying confidence threshold γ . If $c_0 < \gamma$, the rule at hand will not play any role at all, being of confidence too low for the threshold. At $\gamma = c_0$, the rule becomes part of the output of any standard

association mining process, but it could be that some other logically stronger rule appears at the same confidence. For instance, it could be that both rules $A \rightarrow B$ and $A \rightarrow BC$ have confidence c_0 : then $A \rightarrow B$ is redundant and will not belong to the basis for that confidence.

If no stronger rule appears, $X \rightarrow Y$ will belong to the representative basis for that threshold. Let's keep decreasing the threshold. At some lower confidence, a logically stronger rule may appear. If a logically stronger rule shows up early, at a confidence threshold γ very close to c_0 , then the rule $X \rightarrow Y$ is not novel : it is similar to the logically stronger one, and this shows in the fact that the interval of confidence thresholds where it is a representative rule is short. To the contrary, a stronger rule may take long to appear : in that case, only rules of much lower confidence entail $X \rightarrow Y$, so that the fact that it does reach confidence c_0 is novel in this sense. The interval of confidence thresholds where $X \rightarrow Y$ is a representative rule is large.

2.2 Blocking Rules

The main disadvantage often argued against confidence relates its inability to detect negative correlations. For instance, for a threshold of, say, $2/3$ (or around 66%), consider a representative rule $A \rightarrow B$ of confidence slightly beyond the threshold. It is going to be provided as interesting in the output, suggesting that transactions having A tend to have also B . However, in case the actual frequency of B is rather high, say, 80%, the correlation is in fact negative, since B appears *less* often among the transactions having A than in the whole dataset. In fact, this is the major criticism that has been made for confidence as a measure of “degree of implication”, and has motivated a large number of alternatives ; the literature about these notions is, in fact, quite large [9], [11], [12], [17]. A good survey with many references is [10].

Reference [5] suggests a notion of “rule blocking”. A specific motivating example that is described in that paper is as follows : mining for association rules at 5% support and 100% confidence the Adult training set from Irvine [2], 67 (out of 71) rules in the basis are of the form Husband + “something else” \rightarrow Male. The reason is that the rule Husband \rightarrow Male, that we would expect to hold, does not reach 100% confidence : indeed, tuple 7110 includes the items “Husband” and “Female” (instead of “Male”). This opens the door to many rules, intuitively uninformative, that enlarge a bit the left-hand side, enough to avoid tuple 7110 so as to reach confidence 100%. Further examples are given in the same paper. To approach this problem, [5] proposes a scheme that can be explained in an equivalent form as follows :

Definition 2 Given rule $X \rightarrow Y$, with $X \cap Y = \emptyset$, a proper subset $Z \subset X$ blocks $X \rightarrow Y$ at blocking threshold b if $(s(XY) - c(Z \rightarrow Y)s(X))/(c(Z \rightarrow Y)s(X)) \leq b$.

This definition compares the number of tuples having XY with the quantity that would be predicted from the confidence of the rule $Z \rightarrow Y$. Let $c(Z \rightarrow Y) = c$. If Y is distributed along the support of X at the same ratio as along the larger support of Z , we would expect $s(XY) \approx c \times s(X)$. In case the difference in the numerator is negative, it would mean that $s(XY)$ is even lower than what $Z \rightarrow Y$ would suggest. If it is positive but the quotient is low, $X \rightarrow Y$ still does not bring high enough confidence with respect to $Z \rightarrow Y$ to be considered : it remains blocked. But, if the quotient is larger, and this happens for all Z , then $X \rightarrow Y$ becomes novel since its confidence is higher enough than suggested by the rules of the form $Z \rightarrow Y$. It can be checked that the problems of the Adult dataset alluded to above get solved. Note, for later use, that the confidence width bound and the blocking bound are related in [5] as follows : if the confidence width bound is d , then the blocking bound is $b = d - 1$.

3 Confidence Boost

The definition of our main notion is very similar to the one given for confidence width ; but with a twist that, even though it is formally tiny, semantically changes it far enough so as to encompass the notion of blocking.

Definition 3 The confidence boost of $X \rightarrow Y$ is $\beta(X \rightarrow Y) =$

$$= \frac{c(X \rightarrow Y)}{\max\{c(X' \rightarrow Y') \mid (X \rightarrow Y) \neq (X' \rightarrow Y'), X' \subseteq X, Y \subseteq X'Y'\}}$$

Again, in case the denominator is null, the confidence boost is infinite by convention.

Under our assumption that $X \cap Y = \emptyset$ and $X' \cap Y' = \emptyset$, $Y \subseteq X'Y'$ is actually equivalent to $Y \subseteq Y'$. A low confidence boost corresponds to a low novelty. To see this, suppose that $\beta(X \rightarrow Y)$ is low, say $\beta(X \rightarrow Y) \leq b$, where b is just slightly larger than 1. According to the definition, there must be a *different* rule $X' \rightarrow Y'$, with $X' \subseteq X$ and $Y \subseteq X'Y'$, such that $\frac{c(X \rightarrow Y)}{c(X' \rightarrow Y')} \leq b$, or $c(X' \rightarrow Y') \geq c(X \rightarrow Y)/b$. This inequality says that the rule $X' \rightarrow Y'$, stating that transactions with X' tend to have Y' , has a confidence relatively high, not much lower than that of $X \rightarrow Y$; equivalently, the confidence of $X \rightarrow Y$ is not much higher (it could be lower) than that of $X' \rightarrow Y'$. But all transactions having X do have X' , and all transactions having $X'Y'$ have Y , so that the confidence found for $X \rightarrow Y$ is not novel, given that it does not give so much additional confidence over a rule that states as much or more, namely $X' \rightarrow Y'$. As a bare minimum, one should not consider rules with confidence boost 1 or less. Notice that this solves the objection against confidence that negative correlations go undetected : for instance, if the support of B is 80%, a rule $A \rightarrow B$ of confidence less than that would indeed yield a confidence boost below 1.

All rules filtered out of the output of an association miner due to low width will be filtered out as well due to low boost, if the same threshold is used, because, clearly, $\beta(X \rightarrow Y) \leq w(X \rightarrow Y)$. The same can be shown for blocking, provided that we take into account the difference in the meaning of the thresholds : if a rule $X \rightarrow Y$ is blocked by itemset Z at blocking threshold $b - 1$, then one can show $\frac{c(X \rightarrow Y)}{c(Z \rightarrow Y)} \leq b$, and get rather directly from there that $\beta(X \rightarrow Y) \leq b$. Hence, bounding the confidence boost at b ensures us that the rules that would have been filtered by either (or both) of the checks $w(X \rightarrow Y) \leq b$ or blocking at threshold $b - 1$ will be filtered as well by the confidence boost bound. In this sense, confidence boost embodies both low-novelty tests from [5], and with the same thresholds employed there.

3.1 Computing the Confidence Boost : Applications

In order to be of practical use, we need some deeper study of the confidence boost. As it currently stands, it makes no sense to traverse all the alternative rules to be taken into account for computing the maximum confidence in the denominator. The same sort of difficulty appears for confidence width and for blocking. A mild precomputation allows one to compute quite efficiently the width [4], but the same method does not seem to work for blocking or boost. In fact, the experiments reported in [5], as indicated there, resort to a weak approximation to blocking, the actual values being very expensive to compute. Here we develop a completely different alternative approach. Consider the following algorithm :

input : dataset \mathcal{D} ; thresholds for support s , confidence c , and confidence boost b ;
 rule $X \rightarrow Y$ with $c(X \rightarrow Y) \geq c$, $X \cap Y = \emptyset$
output : boolean value indicating whether $\beta(X \rightarrow Y) > b$
 mine \mathcal{D} for the representative rules \mathcal{R} at threshold c/b
 for each rule $X' \rightarrow Y' \in \mathcal{R}$ such that $X' \cap Y' = \emptyset$, $X' \subseteq X$ and $Y \subseteq X'Y'$:
 if $\exists Z \subset X - X'$ such that $c(X \rightarrow Y) \leq b \times c(X'Z \rightarrow Y)$: return `False`
 if $\exists A \in Y' - XY$ such that $c(X \rightarrow Y) \leq b \times c(X \rightarrow AY)$: return `False`
 otherwise : return `True`

Theorem 1 *Let $X \rightarrow Y$ be a rule of confidence at least c . Then, this algorithm accepts it if and only if $\beta(X \rightarrow Y) > b$.*

The comparisons are written in such a way as to avoid division by zero in the cases of infinite boost, such as $s(XAY) = 0$. Of course, the mining of \mathcal{R} is to be done just once and for all. For lack of space, descriptions of the empirical validation are deferred to a forthcoming paper. Among other tests, we have used this algorithm to compute the number of rules passing rather mild confidence boost thresholds of 1, 1.05, 1.1, 1.15, 1.2, 1.25, and 1.3 on three very different datasets, at two different thresholds of support, in the range 0.1% to 5%, and three different levels of confidence (70%, 80%, 90%). The whole computation took about 30 minutes on a low-range laptop. All three datasets consist of real world data, and are of very different characteristics : the training set of Adult from UCI, Retail from FIMI and a transactional form of fossil data from European paleontological sites. The results obtained were very satisfactory.

4 Conclusions and Further Work

The shortcomings of confidence thresholds discussed at the beginning of Subsection 2.2 have been often interpreted as an inadequacy of the notion of confidence. We propose here that a different alternative could be considered : possibly this sort of objections are not to be interpreted as the widespread consequence that “confidence is inappropriate” to filter and rank association rules, but that “an absolute threshold on confidence is inappropriate” to filter and rank association rules. Our message is that it could be complemented with *relative* confidence thresholds that assess the novelty of each rule by comparison with the confidence of logically (or intuitively) stronger rules. The identification of the appropriate measure is a clear research issue, to which we have contributed here via the notion of confidence boost : a reasonably natural definition that embodies both of two notions previously put forward for the same purpose. Of course, the use of confidence boost does not preclude a combination with lift or any other measure of intensity of implication ; to what extent these separate measures interact with confidence boost, and which ones perform best, is one among many open lines of future research.

Références

- [1] C C Aggarwal, P S Yu : A New Approach to Online Generation of Association Rules. *IEEE Transactions on Knowledge and Data Engineering*, 13 (2001), 527–540. (See also ICDE’98.)

- [2] A Asuncion, D J Newman : UCI Machine Learning Repository. Irvine, CA, University of California, School of Information and Computer Science, 2007.
- [3] J L Balcázar : Redundancy, Deduction Schemes, and Minimum-Size Bases for Association Rules ; submitted for publication.
- [4] J L Balcázar : Confidence Width : An Objective Measure for Association Rule Novelty. Workshop on Quality issues, measures of interestingness and evaluation of data mining models QIMIE'09 at PAKDD'09.
- [5] J L Balcázar : Two Measures of Objective Novelty in Association Rule Mining. Extended version of [4], to appear in LNCS (post-proceedings of the workshops of PAKDD'09).
- [6] R Bayardo, R Agrawal, D Gunopulos : Constraint-Based Rule Mining in Large, Dense Databases. ICDE'99, 188–197.
- [7] C Borgelt : Efficient Implementations of Apriori and Eclat. Workshop on Frequent Itemset Mining Implementations (2003) [borgelt.net].
- [8] Frequent Itemset Mining Implementations Repository [fimi.cs.helsinki.fi].
- [9] G C Garriga : Statistical Strategies for Pruning All the Uninteresting Association Rules. ECAI 2004, 430–434.
- [10] L Geng, H J Hamilton : Interestingness Measures for Data Mining : a Survey. *ACM Comp. Surveys* 38, 2006.
- [11] S Guillaume, F Guillet, J Philippé : “Improving the discovery of association rules with intensity of implication”, PKDD 1998, 318–327.
- [12] C Hébert, B Crémilleux : A Unified View of Objective Interestingness Measures. MLDM 2007, 533–547.
- [13] M Kryszkiewicz : Representative Association Rules. PAKDD'98, 198–209.
- [14] B Liu, W Hsu, Y Ma : Pruning and Summarizing the Discovered Associations. KDD'99, 125–134.
- [15] N Megiddo, R Srikant : Discovering Predictive Association Rules. KDD'98, 274–278
- [16] V Phan-Luong :The Representative Basis for Association Rules. ICDM'01, 639–640.
- [17] P-N Tan, V Kumar, J Srivastava : Selecting the Right Objective Measure for Association Analysis. *Inf. Syst.* 29(4) :293-313 (2004)
- [18] M Zaki : Mining Non-Redundant Association Rules. *Data Mining and Knowledge Discovery* 9 (2004), 223–248.

Summary

In association rule mining, it is well-known that merely imposing an absolute confidence threshold leads to certain shortcomings. Many alternative proposals have been suggested to overcome them. Here we propose, instead, to complement the process by filtering also the obtained rules according to their novelty, measured in a relative way with respect to the confidences of stronger rules from the same dataset. Our proposal, the confidence boost of a rule, encompasses two previous similar notions (confidence width and rule blocking) from previous works.