JMLR: Workshop and Conference Proceedings 11 (2010) 74–80 Workshop on Applications of Pattern Analysis

Closure-Based Confidence Boost in Association Rules

José L Balcázar JOSELUIS.BALCAZAR@UNICAN.ES
Departamento de Matemáticas, Estadística y Computación
Universidad de Cantabria, Santander, Spain

Editors: Tom Diethe, Nello Cristianini, John Shawe-Taylor

Abstract

We focus on association rule mining. It is well-known that naive miners end up often providing far too large amounts of mined associations to result actually useful in practice. Many proposals exist for selecting appropriate association rules, trying to measure their interest in various ways; most of these approaches are statistical in nature, or share their main traits with statistical notions.

Alternatively, some existing notions of redundancy among association rules allow for a logical-style characterization and lead to irredundant bases (axiomatizations) of absolutely minimum size. Here we follow up on a study of closure-based redundancy, which, in practice, leads to smaller bases than simpler alternative forms of redundancy, with the proviso that, in principle, they need to be complemented with an implicational basis.

One can push the intuition of redundancy further and gain a perspective of the interest of association rules in terms of their "novelty" with respect to other rules. An irredundant rule is so because its confidence is higher than what the rest of the rules would suggest; then, one can ask: how much higher? Among several variants, a recently proposed parameter, the confidence boost, succeeds in measuring a notion of novelty along these lines so that it fits better the needs of practical applications. However, that notion is based on plain redundancy, of relatively limited practical usefulness. Here we extend the confidence boost to closure-based redundancy, paying a small theoretical price to obtain several advantages in practical applications. We describe a rule-mining system implementing this contribution.

1. Introduction

Compared to more traditional Statistics, a Data Mining application is expected to find out semi-autonomously facts validated by the data, rather than validate on the data some specific hypotheses proposed externally. In the case of Association Rules, in essence, this amounts to enumerating all the rules that are not disproved by the data; in the presence of a support constraint, we even require that data "strongly disproves" an association before ruling it out. As there are exponentially growing quantities of potential associations, even relatively large datasets are unable to disprove most of them. Therefore, even somewhat demanding thresholds for the standard support and confidence parameters generate large numbers of rules with strong similarities among them, leading to intuitive redundancies.

We can see the problem of reducing the redundancy in an association miner's output as a study of "novelty". Indeed, novelty is, in an intuitive sense, a relative notion: it refers to facts that are, somehow, unexpected; hence, some expectation, lower than actually found,

must exist, and must be due to alternative facts or prediction mechanisms. Here we follow up a series of proposals to the effect that, as a minimum, each rule should be evaluated for novelty according to the rest of the rules mined, treated as "alternative" mechanism (see (Balcázar, 2010a) and the references there). Essentially, that approach proposes to measure novelty through the extent to which the confidence value is "robust", relative to those of related rules, as opposed to the absolute consideration of the single rule at hand.

As a preliminary filter, there are several essentially logical definitions of redundancy, patterned after similar intuitions in Propositional or First-Order Logic. This leads to minimum-size bases, like the Representative (or Essential) Rules (Aggarwal and Yu, 2001; Kryszkiewicz, 1998) for plain redundancy or the basis $\mathcal{B}^\star$ for closure-based redundancy, that works better in practice (as it works just on closed sets and spares the computation of minimal generators needed by the Representative Rules) but needs to be complemented with a basis for full implications. All these questions are surveyed in (Balcázar, 2010b).

In general, this still leaves too many rules as output. One can push the intuition of redundancy further. Intuitively, an irredundant rule is so because its confidence is higher than what the rest of the rules would suggest; then, one can ask: "how much higher?". If other rules suggest, say, a confidence of 0.8 (or 80%) for a rule, and the rule has actually a confidence of 0.81, the rule is indeed irredundant and brings in additional information, but its novelty, with respect to the rest of the rules, is not high; whereas, in case its confidence is actually 0.95, quite higher than the 0.8 expected, the fact can be considered novel, in that it states something really different from the rest of the information mined.

Several notions exist that attempt at measuring to what extent the confidence of the rule is substantially higher than that of related rules that would, intuitively, explain the same facts. If we are to require that related rules logically imply the rule at hand, the parameter obtained (confidence width) is not that bad, but still falls a bit short of working in practice. The choice is, then, to remain in a clean logical framework, or to attempt at finding better practical results by allowing a less logical, and more intuitive, notion of redundancy. We have started to explore this path for the Representative Rules in (Balcázar, 2010a), where the notion of confidence boost, related to plain redundancy and Representative Rules, is proposed and studied, with promising results.

Our contribution here is a new variant, the *closure-based confidence boost*, a somewhat sophisticated technical refinement of confidence boost that can be used to filter the rules in the $\mathcal{B}^\star$ basis, of better applicability than Representative Rules. We describe an open-source system that implements this notion and present an evaluation of the advantages of this system, both in a quantitative form and in a qualitative form by discussing the data mining process and its results on a dataset related to research in Machine Learning.

2. Closure-Based Confidence Boost

We will denote itemsets by capital letters from the end of the alphabet, and use juxtaposition to denote union, as in XY. The inclusion sign as in $X \subset Y$ denotes proper subset, whereas improper inclusion is denoted $X \subseteq Y$. For a given dataset $\mathcal{D}$, consisting of n transactions, support $s(X)$ and confidence $c(X \to Y)$ are defined as usual. We denote as $\overline{X}$ the closure of set X with respect to the given dataset $\mathcal{D}$: $\overline{X}$ is the largest set that includes X and has the same support as X in $\mathcal{D}$.

Clearly, if only confidence and support are considered, then rules $X \to Y$ and $X \to XY$ are equivalent, as are rules where some part of the left-hand side X is repeated in the right-hand side. We chose the convention that, in all our association rules $X \to Y$, $X \cap Y = \emptyset$.

Definition 1 *The* closure-based confidence boost *of a rule $X \to Y$ is $\overline{\beta}(X \to Y) =$*

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

The original notion of confidence boost in (Balcázar, 2010a) corresponds to the particular case where the closure operator is the identity function. Connections with lift and other similar notions are discussed there. This is the natural definition paralleling the confidence boost when the notion of reduncancy is closure-based: on one hand, the rules in the denominator may resort to the use of closures to make the rule at hand redundant, widening the options of redundancy; on the other hand, rules that are syntactically different from the rule at hand, but equivalent to it in closure-based redundancy, must be discarded, as they trivially entail it.

We give next an algorithm to filter rules according to their closure-based confidence boost.

Input: dataset $\mathcal{D}$; thresholds for support τ, for confidence c, and for confidence boost $b > 1$; rule $X \to Y$, with $c(X \to Y) \geq c$, and $s(XY) \geq \tau$
Output: boolean value indicating whether $\overline{\beta}(X \to Y) > b$
mine $\mathcal{D}$ for the basis $\mathcal{B}^\star$ at threshold c/b;
for *each rule $X' \to Y' \in \mathcal{B}^\star_{c/b}$, with $X' \subseteq \overline{X}$ and $Y \subseteq \overline{X'Y'}$* **do**
  if $\exists Z \subset \overline{X} - X'$ *such that* $\overline{X'Z} \subset \overline{X}$ *(with inequality) and* $c(X \to Y) \leq b \times c(X'Z \to Y)$ **then**
    return `False`
  end
  if $\exists A \in X'Y' - \overline{XY}$ *such that* $c(X \to Y) \leq b \times c(X \to AY)$ **then**
    return `False`
  end
end
otherwise: return `True`

A simple alternative algorithm that explores by brute force is better when we compute the boost of a single rule, but this algorithm is preferable when we are to filter the whole of $\mathcal{B}^\star_c$, as happens in most applications, provided that the intermediate basis $\mathcal{B}^\star_{c/b}$ is cached and not recomputed from scratch each time.

Theorem 2 *Let $X \to Y$ be a rule of confidence at least c. Then, this algorithm accepts it if and only if $\overline{\beta}(X \to Y) > b$.*

A main disadvantage often argued against confidence relates its unability to detect negative correlations. For instance, for a threshold of, say, 2/3, consider a representative rule $A \to B$ of confidence slightly beyond the threshold. If the actual frequency of B is say,

Conf.	1	1.05	1.1	1.15	1.2	1.25	1.3	1.35	1.4	1.45	1.5
70%	948	824	689	554	417	331	247	175	142	112	85
75%	639	541	444	356	266	212	161	112	97	76	56
80%	367	298	231	182	132	101	78	54	43	36	26

Table 1: Number of rules passing closure-based confidence boost bounds

4/5, then the correlation is, in fact, negative. This is one of the major criticisms that have 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 quite large; a good survey with many references is (Geng and Hamilton, 2006). But we have now an alternative: it is easy to prove that "negatively correlated" rules have always closure-based confidence boost of 1 or lower. The proof follows similar arguments as those used in (Balcázar, 2010a) to prove the analogous fact for plain confidence boost.

3. Preliminary Empirical Validation

The algorithm just described has been implemented in version 0.2.2 of the open source system `slatt` (available from `slatt.googlecode.com`); this system employs the free `apriori` implementation by Borgelt (Borgelt, 2003) to compute the closures, constructs the closure lattice, and offers algorithms to compute the GD basis of implications, the representative rules, and the $\mathcal{B}^\star$ basis; furthermore, as of version 0.2.1, the representative rules can be filtered through plain confidence boost (Balcázar, 2010a), and, as of version 0.2.2, the $\mathcal{B}^\star$ basis can be filtered at a closure-based confidence boost threshold. We have employed this system on real world datasets, with good results. Quantitatively, the figures that we obtain imply that large fractions of representative rules are somewhat uninteresting in that they fully lack any novelty, measured according to confidence boost.

However, all these quantitative arguments have a weakness: Are the actual rules passing the thresholds "the right ones"? An option is to involve "end-users" in the evaluation of the obtained association rules: persons that are extremely well-versed on the dataset at hand (see our discussion below). Here, instead, we go the other way around, and use a dataset for which the readers of this paper are expected to be reasonably knowledgeable: in the same vein as the evaluations in (Gallo et al., 2007), we employ the titles, topics, and abstracts of the reports submitted to the *e-prints* repository of the Pascal Network of Excellence along its early years of existence. This dataset, extracted from the repository by Professor Steve Gunn, was the object of a visualization challenge of the Pascal Network in 2006. Professor Gunn has also kindly furnished this author with a similar but much larger dataset, to which we plan to apply the same scheme in the near future.

The (mild) preprocessing consisted in removing punctuation and nonprintable characters, mapping all letters into lowercase, stripping off stop words as per the list from `www.textfixer.com`, and removing duplicate words from each of the transactions so obtained. This left 45185 total word occurrences chosen from a vocabulary of 8233 items. We checked the size of the closure space at supports of 10% (135 closures) and 5% (830 closures, still

somewhat small), and then at 1% (too large, as after a few minutes the program was still computing the closure lattice's edges—in fact, a later run showed that it consists of over 59713 closures). We settled for the far from trivial but manageable closure space consisting of 9620 closed itemsets obtained at 2% support. Then, we computed the $\mathcal{B}^\star$ basis at confidences 70% (1070 rules) and 80% (412 rules), and cut them down by filtering them at closure-based confidence boosts of 1, 1.05, 1.1, 1.15, 1.2, 1.25, 1.3, 1.35, 1.4, 1.45 and 1.5. All the runs were almost instantaneous. After looking at the resulting figures, the basis at confidence 75% (729 rules) was also computed and filtered at the same confidence boost thresholds. The figures obtained, given in Table 1, make it indeed possible to proceed to manual inspection of many of these options.

As a particular case, we chose to perform a manual examination of the 26 rules found at 2% support, 80% confidence, and 1.5 closure-based confidence boost, which revealed rules with little or no redundancy among themselves indeed, all of them semantically sensible, and a handful of them actually quite interesting (for this author). The whole process leading to these "nuggets" lasted less than two hours, *including all the preprocessing*, for a single person (the author) and quite limited computing power (an old Centrino Solo laptop). These rules are given in Table 3; let us insist here that the contents of this table is *not* a purportedly representative "selection" of the outcome of the mining process but *the whole of it*, at the indicated thresholds, which can be iteratively relaxed in order to obtain further information and, at the same time, avoiding a deluge of output rules. The predefined subjects of the e-prints Pascal server appearing in the table have been given in abbreviated form; Table 2 reports the abbreviations used for them.

At the same level of support, we have tested higher-confidence bases or other schemes with very good results. These comparisons will be reported in a larger version of this paper.

4. Discussion

Many sophisticated interestingness measures exist. We refer to (Geng and Hamilton, 2006) for an excellent survey of many options to relate supports of left-hand and right-hand sides of association rules to construct indicators of interestingness. Many of these only work on a single rule, with no reference to alternative rules with, say, smaller but otherwise arbitrary left-hand sides. Compared to this family of measures, confidence boost is finer as it can distinguish among many alternative antecedents to compare, at the price of being potentially more expensive to evaluate due to the search for smaller, arbitrary right-hand sides. A larger paper including the results here, currently in preparation (although a preliminary version is available from the author), will include a deeper discussion of related work, and will report

subject:B	Brain-Computer Interfaces
subject:I	Information Retrieval and Textual Information Access
subject:L	Learning/Statistics and Optimisation
subject:M	Machine Vision
subject:T	Theory and Algorithms

Table 2: Abbreviations of subjects for Table 3 below

conf.	supp %			
0.889	3.329	presents	⇒	paper
0.833	2.080	solve	⇒	problem
0.850	2.358	features selection	⇒	feature
0.818	2.497	graphs	⇒	subject:T
0.833	2.080	data second	⇒	subject:L
0.895	4.716	bound	⇒	subject:T
0.826	2.635	web	⇒	subject:I
0.941	2.219	art	⇒	state
0.907	5.409	documents	⇒	subject:I
0.882	2.080	approach method show	⇒	data
0.842	2.219	principal	⇒	component
0.914	4.438	document	⇒	subject:I
0.842	2.219	linear problem	⇒	subject:L
0.850	2.358	features subject:T	⇒	subject:L
0.842	2.219	methods subject:M	⇒	images
0.800	2.219	brain	⇒	subject:B
0.889	5.548	bounds	⇒	subject:T
0.818	2.497	data subject:M	⇒	subject:L
0.813	10.264	support	⇒	vector
0.818	2.497	more use	⇒	subject:L
0.919	4.716	object	⇒	subject:M
0.833	2.080	nonlinear subject:L	⇒	learning
0.813	3.606	variables	⇒	subject:T
0.810	2.358	kernel used	⇒	method
0.900	2.497	feature learning	⇒	subject:L
0.842	2.219	unlabeled	⇒	data

Table 3: The 26 rules at 2% support, 80% confidence, 1.5 boost

on further comparisons with several other approaches that we plan to perform in the near future (see (Jaroszewicz and Simovici, 2002), (Gallo et al., 2007), and their references).

As further work, we point out a couple of drawbacks of using closure-based confidence boost bounds, on which we are actively working at present. One is the need to choose yet another parameter for the mining process, besides confidence and support. In experiments, however, this problem does not seem to be that big; we tend to use a few "standard" values for confidence boost, like at 1.05 to prune really low novelty rules, at 1.2 to prune a bit more aggressively, and at (or near) 1.5 to reduce heavily redundancy at the potential price of killing all the output. These options tend to work well, and also make less critical the choice of the confidence threshold, that can be safely left at a somewhat low value (say, around 0.6 to 0.7), leaving to the boost parameter the task of reducing the output size.

Another shortcoming of this approach is that, sometimes, some of the full-confidence implications would be desirable indeed for inclusion in the output, given that working on the basis $\mathcal{B}^\star$ leaves them fully out. Partial progress is reported in (Balcázar et al., 2010a). We are also validating further our approach on the basis of end user advice; we are experimenting with e-learning datasets (Balcázar et al., 2010b), for which the teachers of the courses where the datasets originated are available for consultation.

Acknowledgments

This work has been partially supported by project TIN2007-66523 (FORMALISM) of Programa Nacional de Investigación, Ministerio de Ciencia e Innovación (MICINN), Spain, and by the Pascal-2 Network of the European Union. Dr. Cristina Tîrnăucă is entitled to the author's grateful thanks for discussions and proofreading of this paper.

References

Charu C. Aggarwal and Philip S. Yu. A new approach to online generation of association rules. *IEEE Trans. Knowl. Data Eng.*, 13(4):527–540, 2001.

José L. Balcázar. Objective novelty of association rules: Measuring the confidence boost. In Sadok Ben Yahia and Jean-Marc Petit, editors, *EGC*, volume RNTI-E-19 of *Revue des Nouvelles Technologies de l'Information*, pages 297–302. Cépaduès-Éditions, 2010a. ISBN 978-2-85428-922-0.

José L. Balcázar. Redundancy, deduction schemes, and minimum-size bases for association rules. Logical Methods in Computer Science. Available at: [http://personales.unican.es/balcazarjl], 2010b.

José L. Balcázar, Cristina Tîrnăucă, and Marta E. Zorrilla. Filtering association rules with negations on the basis of their confidence boost. To appear in KDIR 2010. Available at: [http://personales.unican.es/balcazarjl], 2010a.

José L. Balcázar, Cristina Tîrnăucă, and Marta E. Zorrilla. Mining educational data for patterns with negations and high confidence boost. To appear in Tamida 2010. Available at: [http://personales.unican.es/balcazarjl], 2010b.

Christian Borgelt. Efficient implementations of apriori and eclat. In Bart Goethals and Mohammed Javeed Zaki, editors, *FIMI*, volume 90 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2003.

Arianna Gallo, Tijl De Bie, and Nello Cristianini. Mini: Mining informative non-redundant itemsets. In Joost N. Kok, Jacek Koronacki, Ramon López de Mántaras, Stan Matwin, Dunja Mladenic, and Andrzej Skowron, editors, *PKDD*, volume 4702 of *Lecture Notes in Computer Science*, pages 438–445. Springer, 2007. ISBN 978-3-540-74975-2.

Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38(3), 2006.

Szymon Jaroszewicz and Dan A. Simovici. Pruning redundant association rules using maximum entropy principle. In Ming-Shan Cheng, Philip S. Yu, and Bing Liu, editors, *PAKDD*, volume 2336 of *Lecture Notes in Computer Science*, pages 135–147. Springer, 2002. ISBN 3-540-43704-5.

Marzena Kryszkiewicz. Representative association rules. In Xindong Wu, Kotagiri Ramamohanarao, and Kevin B. Korb, editors, *PAKDD*, volume 1394 of *LNCS*, pages 198–209. Springer, 1998. ISBN 3-540-64383-4.