Rules with Bounded Negations and the Coverage Inference Scheme

José L. Balcázar

Dept. LSI, UPC

C5, Campus Nord

08034 Barcelona, Spain

balqui@lsi.upc.es

http://www.lsi.upc.es/~balqui

Abstract

We develop an experimental study of the combination of the Coverage Inference Scheme proposed by Cristofor and Simovici with the bounded-negation association rules proposed by Fortes, Balcázar, and Morales. Our main contributions are:

1/ A fully independent experimental validation of the advantages of the Coverage Inference Scheme on two different datasets.

2/ An experimental validation of the interest of the bounded-negation association rules.

3/ An assessment of the applicability of the Coverage Inference Scheme to bounded-negation association rules.

Our results clearly indicate that this Inference Scheme can be advantageously used to summarize association rules as the authors suggest, and that its combination with bounded-negation association rules helps overcome the associated problem of handling a large number of output rules.

Research supported by the IST Programme of the EU under contract number IST-1999-14186 (ALCOM-FT), Spanish Government PB98-0937-C04 (FRESCO), and CIRIT 1997SGR-00366.

1. Introduction

The notions of frequent sets and association rules are currently a major topic in both the research area and the applications of Data Mining. Nearly all commercial data mining suites include algorithms to compute some form of frequent sets and association rules. Although finding them is a computationally intensive tasks, current algorithmics are able to offer a reasonable performance and provide the results, in most cases, within an affordable turnaround time.

Three associated research topics are: how to improve such algorithms in terms of performance, how to expand the domain of useful information that they can extract, and how to cope with the still too large, in most cases, set of obtained rules. We do not treat here at all the first topic. See [HK], [HGN] for good surveys of diverse algorithms, and [HP] and [BCGB] for alternative approaches.

Variations of the algorithms for frequent sets and association rules are able to yield other combinatorial concepts, potentially more useful and informative, at least for some practical cases; this includes the mining of sequential patterns in ordered transactions, the related problem of finding frequent episodes in long sequences, the search for generalized association rules in presence of taxonomic hierarchies, and the combination of the search with more general constraints. See [HK] and the references therein. A related cousin of the problem of finding association rules is that of eliciting perfect or approximate functional dependencies in databases. In the case of boolean attributes, an intermediate concept between standard association rules and functional dependencies was proposed in [FMB]: association rules with bounded negations, where some of the antecedents of the rule, and possibly the consequent as well, correspond to the *absence* of an item from a transaction, rather than to its *presence*. Richer, potentially more useful information could be found in this way; the main contribution of [FMB] was the possibility of constructing them along a standard Apriori-like exploration with no extra testing at all, against the input database, of the rules including negated attributes, so that efficiency of the algorithmics is maintained even though new, previously inaccessible rules are found.

A major strength of that algorithm was the possibility of controlling the number of negations allowed into the rules; this is important, as the possibility of using negations of the attributes opens the way to overwhelmingly large amounts of output rules, and it becomes important to keep up criteria to discard or organize them.

Actually, already in the standard problem of fully positive association rules, the mining of many a dataset at low levels of support and confidence is impractical, due to the fact that extremely large sets of rules are generated. Of course, a default position is thus to maintain a high threshold of one of them (most usually confidence); but this may be inappropriate in cases we do not want to impose such limitations. There are two additional approaches to cope with very large sets of generated rules. According to the first one, rules are pruned out after their obtention according to a number of criteria; proposals along this line can be found in [LHM], [CG], and the references there. We do not pursue that alley here.

According to the second approach, one tries to construct lossless summaries (frequently called covers) consisting of subsets of the mined rules from where *all* the mined rules can be recovered, possibly on account of the frequencies of all the frequent sets that are assumed to be available from a previous frequent-sets pass; thus, the information can be presented in a more careful way, combining well with exploratory analysis, but still being able to generate exactly the same set of rules that would be found at the chosen support and confidence. We describe next such a notion of cover, due to [CS], who applied it to a synthetic dataset and to the Mushroom database. We describe the results of employing this cover strategy on the databases Car (which is close to synthetic) and Contraceptive Method Choice, with real-world data coming from (a subset of) the 1987 National Indonesia Contraceptive Prevalence Survey; both are publicly available at the UCI repository [BM]. The results are in very reasonable correspondence with the ones obtained from Mushroom, even though we did not insist in producing the most informative cover. Note that our validation of the Covering Inference Scheme is fully independent in the sense that we purposely avoided the use of the same software, which is publicly available in the ARtool program [Cr].

Then we consider association rules with negations on these same datasets, and verify that indeed the number of association rules found grows pretty fast so that the bounded-negation technology from [FBM] is indeed necessary, but not sufficient; and then describe the results of combining the computation of covers with bounded-negation rules, obtaining much more reasonable results. We further discuss other potential progress avenues towards more useful concepts and algorithms along the same line.

2. Coverings

Our setting is fully standard in the texts about association rules. The input dataset is a collection of labelled transactions, each of them a subset of a fixed set of n items; one can think of the items as the headings of the columns in a tabular form in which the transactions would correspond to rows and each entry would indicate the boolean value corresponding to the presence or absence of the item in that transaction. We use capital letters early in the alphabet for items, A, B, $\ldots$ and capital letters from the end of the alphabet for itemsets, T, U, X, Y; these may be transactions too.

We use the conveniently overloaded notation $\sup(X)$ to denote both the set of all transactions T for which $X \subseteq T$, and the cardinality of the set. In all our statements including this notation, either the context will disambiguate whether we refer to a set or to a number (its cardinality), or will be a claim holding for both interpretations. We call $m = \sup(\emptyset)$, the number of transactions or size of the dataset. Usually all supports are scaled into percentages or probabilities by dividing by m or by $m/100$; this scaling is actually irrelevant for our purposes and we chose to speak of support directly as the number of transactions.

An association rule is a pair of itemsets, usually spelled as $X \to Y$; their intended meaning is that transactions containing X tend to contain Y too. Most (though not all) sources also require X and Y to be disjoint. We follow the standard support/confidence setting since optimal rules for other interest measures can be found on the optimal support/confidence border [BA]; besides, the original formulation of the Coverage Inference Scheme [CS] belonged there. The support of a rule $X \to Y$ is the support of the itemset $X \cup Y$ (or its scaling when all supports are considered scaled by m); the confidence of the rule is $\sup(X \cup Y)/\sup(X)$. The data mining process is assumed to start from user-specified thresholds for confidence, κ, and support, specified as a fraction σm of the size m of the database. The frequent-sets algorithms produce all rules whose numeric support and confidence run above the threshold; there starts our problem of organizing the (frequently too many) rules found.

There are a number of inference schemes that obtain rules whose support and confidence are high enough, from other rules. A part of [CS] is devoted to their description and comparison. We only work here with one contribution of [CS], which fares better in the comparisons: the Coverage Inference Scheme.

Proposition [CS]. Consider rules $X \to Y$ and $X' \to Y'$ such that $X' \cup Y' \subseteq X \cup Y$; if $\sup(X') \leq \sup(X)$ numerically, and $X \to Y$ has support and confidence above the thresholds, then $X' \to Y'$ does as well.

This can be readily proved from the definitions. Also, it is easy to check that the proof carries through to the rules with negated attributes using the framework of [FBM]. A rule covers another if the second can be inferred from the first according to this proposition; and likewise for sets of rules. A cover of a set of rules H is a minimal subset of it such that each rule of H is covered by a rule of the cover. In [CS], the coverage relation is analyzed and proved to have reasonable algebraic properties; and examples are given where, from a single rule, many other rules can be found through this scheme.

3. Experiments

The Car Evaluation Dataset [BR] in the UCI repository [BM] has seven categorical columns (six attributes and a class), which we preprocessed into a single boolean attribute per value of each categorical attribute, thus obtaining 27 boolean attributes. The dataset has 1728 instances, bijectively corresponding to each possible configuration of values of the six original attributes; thus, actually there is no statistical correlation at all among these attributes, and the only rules that might be meaningful are those that refer to the class, which can take four values (unacceptable, acceptable, good, and very-good). But, of course, such side knowledge was not used at all in our process here.

This dataset was mined for association rules at the low support level of 10%, and at confidence levels of 90%, 70%, 50%, 30%, and 10%, obtaining between 13 and 160 rules. However, mining for association rules with negations at 10% support and confidence thresholds produced, as could be expected, a fully unmanageable result of 2253890 rules, requiring over 141Mb of storage.

Applying the layered structure of bounded-negation association rules from [FBM], at the same set of thresholds for confidence and always at 10% of support, we obtained between 157 and 2347 rules with one single negated attribute and between 1507 and 21541 rules with two negated attributes. These figures, although still too high to explore manually, are promising in that applying the covering-based summarization could yield workable sizes.

Table 1 provides all the specific data for the experiments with the dataset Car. The first pair of columns refers to standard association rules, the second one to rules with at most one negated attribute, and the third pair to rules with at most two negated attributes. Rows correspond to confidence. In each pair, the larger (or equal) figure at the left of the slash is the number of raw rules that hold, and the number at the right of the slash is the number of covering rules.

	no neg	1 neg	2 neg
conf 90%	13 / 13	157 / 157	1507 / 1389
conf 70%	20 / 20	835 / 652	9263 / 5143
conf 50%	40 / 32	1295 / 860	14924 / 5996
conf 30%	131 / 50	1823 / 869	18204 / 6003
conf 10%	160 / 50	2347 / 869	21541 / 6003

Table 1. Results for Car

In all the cases we find that, as confidence goes below 50%, the number of raw rules found still grows with no apparent contention, yet the cover remains almost stable, and fully stable below confidence of 30%. This suggests that indeed the cover has found the rules (under the corresponding negation condition) that the dataset may offer. It is interesting to observe that the cover rules have generally high confidences, with only a few dropping below 60% and many still above 90%.

Note the interesting consistency that the step from one negation to two increased the the number of raw rules by a quite stable factor of about 10, and the number of cover rules by a smaller factor. Thus the case without negations confirms the interest of the approach of [CS], whereas the other two cases both confirm the usefulness of the bounded-negation rules of [FBM] and exhibit an excelent fit together of the two techniques.

As for the rules found, most of them exhibited as consequent a very frequent attribute such as "unacceptable car" (over 70% of the instances) or "not a good car" (about 96% of the instances). Given the scarcity of cars labeled as good or very good, the only way to know about them is by comparing "acceptable cars" with "not unacceptable cars", the latter corresponding to acceptable, good, or very good ones; this is only possible by using rules with negations. As tokens of rules with negations obtained, one finds that cars that are not very expensive, and seat 4 people, if they are still unacceptable then their safety is not high; or that a car labelled "acceptable" for five or more persons must not have low safety. Both rules hold at a low support close to the threshold but very high confidence.

The Contraceptive Method Choice dataset [[LLS]] has 1473 instances with 10 attributes each, one of them being the class. The data comes from a national survey in Indonesia, and consists of sociodemographical and personal information about families, including advances in studies of wife and husband, wife's age, number of children born by her, standard of living, occupations, religion, and contraception method: none, short term, or long term. Note that "some" contraception method can only be considered through a negated attribute.

In our preprocessing, the two numerical attributes (age and number of children of the wife) were crudely booleanized by simple comparison against the average. The 5 categorical attributes (including class, i.e. contraception method) were booleanized as in the previous case. Altogether we had 24 boolean attributes. We computed association rules at support 30%, for the same values of confidence as before, 90%, 70%, 50%, 30%, and 10%, obtaining between 37 and 179 rules.

However, as with the previous case (but not so dramatically), allowing negations in the rules produced 135209 of them, close to 7.5Mb of rules. At any rate, the bounded negations framework is also necessary here. As the confidence varied as indicated, we obtained between 611 and 2137 rules with one negated attribute and between 3665 and 11068 rules with two negated attributes. An amazing fact turned out to be that, in all three cases of no negations, one negation, and two negations, the same set of rules was obtained for a confidence threshold of 10% than for a confidence threshold of 30%. Thus we do not report anymore the results for the 10% case of the confidence. The resulting rules are shown in Table 2, according to the same layout as in Table 1.

	no neg	1 neg	2 neg
conf 90%	37 / 34	611 / 407	3665 / 1896
conf 70%	107 / 49	1470 / 472	8277 / 2082
conf 50%	148 / 49	1956 / 476	10620 / 2086
conf 30%	179 / 49	2137 / 476	11068 / 2086

Table 2. Results for CMC

The same phenomenon discussed for the previous dataset is readily apparent here, even more strikingly since the cover is almost stable already at the 70% confidence threshold. We see that bounding the negations to two and computing a cover provides a set of 2000 rules out of the 7.5Mb, which is laborious but feasible to explore by hand; and the column for one single negation is still better. Of course, the columns for no negations confirm the findings of [CS] with respect to the usefulness of the Coverage Inference Scheme.

Some examples of rules obtained where allowing negations became useful are: when the religion of the wife is Islam and she is not working, then the education of the husband is *not* low; or: when the education of the husband is high and the wife is younger than average, then her education is *not* low. Being "not low" is a property that the transactions themselves only enjoy in this "negation" setting, whereas, if we look just into the presence of items, these transactions will scatter into several values and fail to make up for a reasonably sized set of transactions that could be captured in a rule. This is also a way of validating the interest of rules with bounded negations.

4. Discussion

Our main interest in this work was to see whether the notion of cover would help reduce further the quantities of rules with negations already partially controlled through bounds on the number of negations. We consider satisfactory and promising the results obtained.

Let us discuss some limitations of this work. First, the datasets are not really large, and statistical noise is present to a still somewhat tolerable extent: mere visual inspection of the generated rules easily leads the eye to peculiar cases and discarding noisy effects is easy. This is not going to scale well, and better statistically well-grounded methods will be required to handle statistical noise on really large datasets. Additionally, two pretty naive preprocessing decisions were the individual booleanization for each possible value of each categorical attribute, as opposed to finer approaches like output coding, and the crude comparison against the average in the numeric attributes. The reason for our simplest process is that our intention was not truly to mine these databases but rather to test on these datasets the bounded negations and coverage approaches of previous works. In a real data mining process, of course these simplistic approaches should be considerably refined.

We developed additional schemes with the purpose of further simplifying the sets of rules. These correspond, intuitively, to the *Modus Tollens* rule of classical logic, and show that we can compute supports of rules with bounded negations from those without them. These rules are as follows:

a/ $\sup(X, \neg A \to B) = \sup(X \to B) - \sup(X, A \to B)$

b/ $\sup(X, A \to \neg B) = \sup(X \to A) - \sup(X, A \to B)$

c/ $\sup(X, \neg A \to \neg B) = \sup(X \to A) + \sup(X \to B) - \sup(X, A \to B) - \sup(X)$

d/ $\sup(X, \neg A, \neg B \to C) = \sup(X \to C) - \sup(X, A \to C) - \sup(X, B \to C) + \sup(X, A, B \to C)$

Similar formulas can be readily developed for the confidences of rules with one or two negations in terms of supports of rules without negations. This opens the possibility of further reducing the summaries of association rules by omitting some whose support and confidence can be computed from the remaining ones according to these rules. However, we do not develop this line further since, in our experiments, application of this approach gave absolutely no true reduction in the size of our obtained rule set. Thus, although sensible, this approach seems utterly inappropriate.

Our experiments also served as an independent confirmation of the experiments made in [CS], with the small caveat that one surprising effect, namely the reduction of the cover size for lower confidence values, was not observed in our experiments; the most likely reason for this slight divergence is our not insisting in constructing the so-called *most informative cover* advocated in [CS]. Rather, in our experiments, the cover size stabilized to a very robust point, to the extent that one is tempted to consider that some such stable cover (together with its consequences) is the true set of association rules that the dataset offers.

References

[BA] R. Bayardo, R. Agrawal. Mining the most interesting rules. In: Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, 145–154, 1999.

[BCGB] J. Baixeries, G. Casas-Garriga, J. L. Balcázar. A best-first strategy for finding frequent sets. In: Extraction et gestion des connaissances EGC 2002, D. Herin, D.A. Zighed eds. Montpellier, France, Jan 2002, 101–106.

[BM] C.L. Blake, C.J. Merz. UCI repository of machine learning databases. 1998. `http://www.ics.uci.edu/~mlearn/MLRepository.html`

[BR] M. Bohanec, V. Rajkovic. Knowledge acquisition and explanation for multi-attribute decision making. In: 8th Intl Workshop on Expert Systems and their Applications, Avignon, France, 59–78, 1988.

[CG] G. Casas-Garriga. A multi-step strategy to remove uninteresting association rules. Manuscript, may 2002.

[Cr] L. Cristofor. ARtool. `http://www.cs.umb.edu/~laur`

[CS] L. Cristofor, D. Simovici. Generating an informative cover for association rules. 2002, `http://www.cs.umb.edu/~laur`

[FMB] I. Fortes, J. L. Balcázar, R. Morales. Bounding negative information in frequent sets algorithms. Discovery Science 2001, LNAI 2226, KP Jantke and A Shinohara eds, Washington DC, USA, 50–58.

[HGN] J. Hipp, U. Güntzer, G. Nakhaeizadeh. Algorithms for association rule mining: A general survey and comparison. SIGKDD Explorations 2000, 58–64.

[HK] J.-W. Han, M. Kinber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001.

[HP] J.-W. Han, J. Pei. Mining frequent patterns by pattern-growth: methodology and implications. SIGKDD Explorations 2000, 30–36.

[LHM] B. Liu, W. Hsu, Y. Ma. Pruning and summarizing the discovered associations. In: Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, 125–134, 1999.

[LLS] T.-S. Lim, W.-Y. Loh, Y.-S. Shih. A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms. Machine Learning 40, 3 (2000), 203–228.