ONTOLOGY IN ASSOCIATION RULES PRE-PROCESSING AND POST-PROCESSING

Inhauma Neves Ferraz
Computer Science Department – Universidade Federal Fluminense
Rua Passo da Pátria 156 Bloco E Sala 326 – Niterói 24210-240 Brazil

Ana Cristina Bicharra Garcia
Computer Science Department – Universidade Federal Fluminense
Rua Passo da Pátria 156 Bloco E Sala 326 – Niterói 24210-240 Brazil

ABSTRACT

Data mining has emerged to address the problem of transforming data into useful knowledge. Although most data mining techniques, such as Association Rules, substantially reduce the search space, oftentimes one finds that the solution obtained surpasses the human ability to handle the resulting information. Furthermore, a good part of the information in repositories may be wrongfully dismissed due to the mining methods' inability to grasp the relationships between stored data from world knowledge that makes it possible to discover new valuable results, as well as eliminate irrelevant ones. This paper studies domain ontology as an instrument to enhance the mining results of Association Rules, which also acts to reduce the number of generated association rules. The adopted model is based on generalization and specialization processes in which the rules are filtered by metrics based on the coverage and confidence indicators.

KEYWORDS

Data Mining, Association Rules, Ontology, Preprocessing, Post processing, Pruning.

1. INTRODUCTION

A knowledge of the world, normally formalized by means of Ontologies, is a source of data mining enhancement, since the aim is to obtain knowledge, not only data, from the information repository.
Apparently, knowledge of the world is exacerbating the problem of inflated mined rules instead of contributing to its solution. However, the result obtained is much better as it allows the processes of mass cardinality reduction through pruning patterns and associations to utilize semantic criteria. The processes that are strictly syntactic reduce cardinality, but also prune what should not be pruned and do not prune what should be pruned. This does not occur when one uses knowledge of the world as a pruning mechanism. Without the use of world knowledge, of ontologies, this knowledge would be dismissed or would need to be filtered by the analyst's intuition.
The simplest way of utilizing subjacent knowledge is to group attribute instances into value ranges. By grouping purchasable items by price range (cheap, normal price, expensive), individuals by height (short, medium, tall), by productivity (low, medium, high, excellent), one is able guide findings more effectively than when trying to do the same thing with isolated instances.
Comprehensive variables, such as value ranges, are called dependent, since whenever a specific (or determinant) variable is known, the comprehensive (or dependent) variable is also known. Each type of relationship has its representation and resolution dependency mechanisms.
Research of this nature can be found for the study of "is-a" relationships with taxonomies (Srikant, 1995) but utilizing world knowledge in the pre-processing phase and filtering by interestingness.
Our null hypothesis states that our suggested model does not execute the filtering of a set of mined association rules by maintaining precision and reducing recall. Filtering is done by rule generalization and specialization process. Aggregating variables give origin to more general or comprehensive rules. These rules are compared to the mined rules through metrical structures based on objective measurements (coverage and confidence) that indicate if the new rule may substitute a set of original rules (generalization) or if it does not add information, in which case it must be unconsidered (rule specialization). The comprehensive rules derive from domain ontology.
The comparison of results obtained by semantic filtering and syntactic filtering can be performed quantitatively and qualitatively. The quantitative comparison is performed by the amount of pruned rules.

The refutation of the null hypothesis comes from recall reduction obtained by filtering. The qualitative comparison has to be made in light of domain ontology. Syntactic proposals of association rule pruning are not good enough because they prune what should not be pruned and refrain from pruning what should be pruned. The refutation of the null hypothesis is due to precision maintenance, as filtering does not prune what should not be pruned.

2. SEMPRUNE

This Section describes the model adopted for considering knowledge of the world in the post-processing of association rules. The SemPrune model is an alternative to the usual consideration of incorporating world knowledge in the pre-processing phase of rules mining. Semantic filtering as well as semantic enhancement are the basis of our pos-processing model of association rules and the enhancement of results in the post-processing phase will be described, and the model will be introduced below.

As in any pos-processing method SemPrune starts with the results of association rules (AR) data mining technique. Dependences between items in each AR are identified and heuristic rules are applied to semantic filter the irrelevant.

Let D be a relationship belonging to a multidimensional database. Consider x and y as two attributes of D. Let the dependence of x and y be given by f: Domain(x) $\Rightarrow$ Domain(y). Consider y_i an element of Domain y, and x_{ij} an element of Domain x, so that $f(x_{ij}) = y_i$. Consider that $X_{ij} = \{x = x_{ij}\}$ and $Y_i = \{y = y_i\}$ two sets of conditions defined over attributes x and y of D, respectively. If there is a dependence relationship between attributes x and y, or between sets X_{ij} and Y_i, if one instance satisfies the condition of set X_{ij} , then this instance also satisfies the condition of set Y_i.

Filtering will be performed by generalization and specialization processes. In generalizations, we value the synthesizing ability of discovered relationships, whereas in specializations, we value the rules' discriminatory ability. The choice between generalization (preferential) and specialization is made initially using the measurement of coverage interest, which is the traditional measurement to ascertain the generality of a rule. In this case, the general rule is the one that does not bring any added value and can be pruned. We begin defining two relevance indicators.

Definition *(CRg). Let D be a multidimensional database. Let* R_i *be an association rule in the form of expression* $Y_i \wedge A \Rightarrow B$*, and* $S_i = \{r_{ij} \mid j = 1..n\}$ *the set of corresponding rules in the form of expression* $X_{ij} \wedge A \Rightarrow B$*, obtained from D. The value of the measure CRg for* R_i *and* S_i *is given by:*

The measure CRg is based on the Coverage interest measure (Lavrac, 1999) and can be interpreted as the conditional probability that an instance could satisfy the antecedent of one of the more specific rules, given that the instance satisfies the antecedent of the more general rule. The value of the Coverage of a rule is given by the support of the antecedent of this rule.

Definição *(CRm). Let D be a multidimensional database. Let* R_i *be as association rule in the form of the expression* $Y_i \wedge A \Rightarrow B$*, and* $S_i = \{r_{ij} \mid j = 1..n\}$ *the set of corresponding rules in the form of the expression* $X_{ij} \wedge A \Rightarrow B$*, obtained from D. The value of the Confidence for* R_i *and* S_i *is given by:*

$$CRm(R_i, S_i) = \frac{\sup\left(A \cap Y_i \cap B \cap \left(\overline{\cup X_{ij} \cap A}\right)\right)}{\sup\left(A \cap Y_i \cap \left(\overline{\cup X_{ij} \cap A}\right)\right)} \qquad (2.1)$$

If the more general rules continues to be valid with this restriction – that is, if the value of the CRm measure is above the minimum valued specified by the user – this means that the behavior of the population is uniform, seeing as the behavior described by the more general rule is valid for whatever values of the attributes that determine the dependent attribute. This measure varies between 0 and 1. For trivial substitution process, you can write:

$$CRm(R_i, S_i) = \frac{\sup(R_i) - \sum_{j=1..n} \sup(r_{ij})}{\frac{\sup(R_i)}{conf(R_i)} - \sum_{j=1..n} \frac{\sup(r_{ij})}{conf(r_{ij})}} \qquad (2.2)$$

This measure also is based on the Confidence measure and the same formula is used with the inclusion of the restriction of the instances considered in the calculation of the Confidence of the more general rule, taking into consideration only those not covered by the more specific singular rules, of high support.

The semantic filtering algorithm of the set of mined associations rules with the Database enhanced with added items is done in two stages. In the first stage, the trivial rules, such as those that present in the antecedent dependent as well as determinant items or items that are determinant in the antecedent and items dependent in the consequent (and so on), are promptly eliminated. En suite, subsets of rules that present the same consequent are generated by the algorithm described by Domingues (2004). For each of

these subsets, a set *G* of more general rules is generated. In a second stage, each general rule generated is analyzed to verify its generalization capacity.

The second stage begins by generating the set *E* of more specific rules that are redundant with each general rule. Based on the more general rule, and the corresponding set of more specific rules, the value of measure *CRg* is calculated. If the value of the *CRg* measure is greater or equal to the minimum value specified by the user (*CRgMin*), the more specific rules are eliminated by way of a generalization process. If it is not, the value for the *CRm* measure is calculated. If this value is greater or equal to the minimum value specified by the user, once again the process of rule generalization is applied. If the value of the *CRm* measure is lower than the minimum value specified by the user, the process that is executed is the rule specialization with the elimination of the more general rule. In fact CRg and CRm are the relevance indexes used in filtering process described in SEMPRUNE. When CRg or CRm reach the lower limit of CRgMin or CRmMin, respectively, the more general rule is preferred to be fired. Otherwise the general rule is eliminated, i. e., it is considered irrelevant.

The post-processing enhancement of results is conducted by an algorithm that receives the following as entrance parameters: a set *R* of association rules, a set *RD* of dependency relations among the attributes, a mapping *DetAttr* that relates the attributes that possess dependencies, and a mapping *DepAttr* that stores those attributes that are dependent. The algorithm returns an enhanced set of association rules.

For each rule, a list, which is initially empty, is created with the determinant attributes present in the rule. The rule's items are scanned. If the attribute of the current item *i* is an attribute with a dependent, the algorithm will recover all the items *idep* that are dependent on *i*. A specific function returns the attributes with which you calculate the value of *idep*. If any determinant attribute of *idep* is not present in the rule, nothing is done. Otherwise, with this set, you calculate the value *val* of the dependent attribute. A new enhancing, temporary, rule is generated to substitute all the determinant attributes by the attribute with the value *val*.

Each more general enhancing rule should substitute a number of specific rules by way of a generalization process. If this is possible, there is simultaneously a semantic enhancement of the set of mined association rules as well as reduction in the cardinality of the set of rules. Nevertheless, if the population of the Database is not balanced, that is, if the results to be added do not have a reasonably uniform distribution, then a process of specialization discards the aggregating rule that was semantically generated. This analysis is conducted in the second stage of the semantic filtering algorithm described in the previous section.

The model used in the study is composed of a number of modules. The descriptions of the algorithms in these models were condensed as much as possible to adjust to the specifications of the Conference, but the text in full is available upon request.

The proposed model, SemPrune, is illustrated in figure 2. The user starts the process selecting the set of mined association rules and the domain ontology. After configuring the filter's parameters SemPrune pos processing begins. The Enhancer module is responsible to include each association rule information from the domain concerning the relationships among the items in each itemset. For each generic association rule the respective specific rules is selected. The Relevance Analyzer is responsible for CRg and CRm calculation. The Pruner finalizes the process by ordering and selecting the most relevant association rules according CRg and CRm.

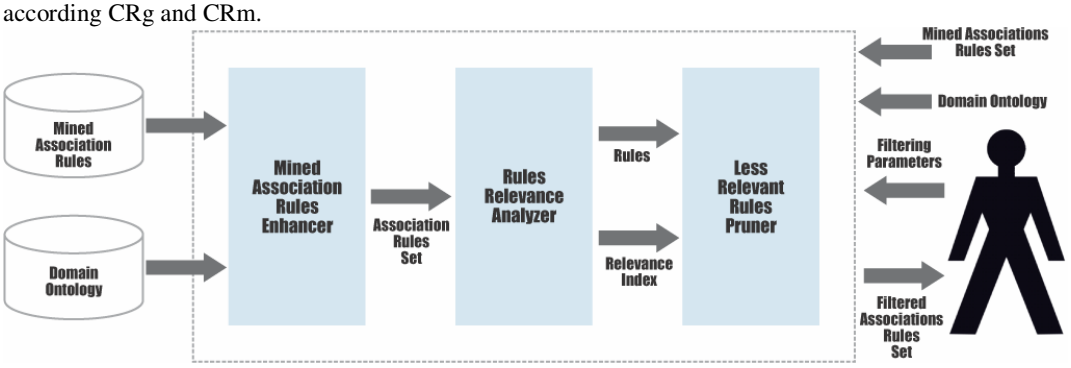

Figure 1 The SemPrune model

3. EXPERIMENTAL EVALUATION

To compare the results from the semantic enhancements made in either the pré-processing or post-processing phase, we use the Databases from the usual Internet repositories and introduce dependent attributes, like range and aggregator, which is common in Data Mining. In each case, the rules enhancements were compared using knowledge of the world during pre-processing and post-processing. The association rules mining was done using the Apriori algorithm (Agrawal, 1993). The knowledge of the world utilized dependencies among attributes that resulted from relationships of the "is-a" type. We chose to test the Adult and Labor Databases originating from the UCI Machine Learning Repository (Merz,1998). The STULONG Database was obtained from ECML/PKDD 2004 Discovery Challenge (ECML/PKDD, 2004). The support values adopted were 4% for the Labor Database and 2% for the other two databases. The confidence value adopted in all the minings was 70%. The maximum size specified for the frequent itemsets was 3 for the Stulong Database and 5 for the other two databases. We initially show the comparison between the quantitative results of the semantic filtering that was done only in the post-processing and done in both the pre-processing and post-processing. There were 2 aggregator attributes for the Labor Database and three for the other two databases.

Table 1. Ontology used only in the post-processing

Database	# Items	# Mined Rules	# Filtered Rules	% Filtered
Adult	14	36.998	28.386	76,72%
Labor	16	181.229	159.330	87,92%
Stulong	19	1.798	1.652	91,88%

Table 2. Ontology used in the pre-processing and in the post-processing

Database	# Items	# Mined Rules	# Filtered Rules	% Filtered
Adult	17	129.349	89.881	69,49%
Labor	18	333.745	235.775	70,65%
Stulong	22	2.764	2.249	81,37%

The difference in the computation costs of the proposed model can be observed in tables 1 and 2.

For the qualitative comparison of data mining post-processing between syntactical methods (pruning using Conviction, Specificity, Lift and Novelty) and semantic methods (pruning using generalization/ specialization) performance tables can be employed. The qualitative comparison must be conducted in light of the domain ontology through the observation of an analyst who can attest to the conformity of the results obtained by filtering with the domain ontology.

In the performance tables, similar to the confusion matrixes, it can be said that the pruning of any rule can be viewed in light of two aspects: it should (S) or should not (SN) have been pruned and it was (W) or was not (WN) pruned.

Experiments were conducted in the Databases described with the said four objective measures using four values of each one of the said measures as a cutting point.

Unequivocally, syntactic measures, by varying the cutting points, are able to prune a much larger number of rules that the semantic method in this study. However, they eliminate rules that should not have been eliminated and leave behind rules that should have been eliminated. Tables 3 and 4 show the qualitative aspect for the sample Databases. The results obtained from the SemPrune model were 18,63% for Stulong, 3,21% for Labor and 30,51% for Adult Databases.

Table 3. Qualitative comparison of Conviction and Specificity syntactic filters

Metric	Conviction				Specificity			
Database	1.1	1.2	1.3	1.4	0,95	0,97	0,98	0,99
Stulong	1.30%	1.48%	1.63%	1.63%	1,05%	1,37%	1,74%	1,74%
Labor	3.71%	3.73%	3.74%	3.75%	0,19%	0,48%	1,16%	1,36%
Adult	13.67%	15.04%	16.42%	17.57%	2,41%	5,93%	10,60%	18,98%

Table 4. Qualitative comparison of Lift and Novelty syntactic filters

Metric	Lift				Novelty			
Database	1	1,1	1,2	1,3	0	0,1	0,2	0,3
Stulong	0,00%	0,65%	1,81%	1,95%	0,00%	2,93%	2,97%	2,97%
Labor	0,00%	0,07%	1,07%	1,15%	0,00%	5,04%	5,05%	5,05%
Adult	0,00%	7,01%	9,54%	11,09%	0,00%	47,28%	47,50%	47,50%

The comparison of results obtained with the introduction of aggregator attributes in the pre-processing and post-processing clearly indicates the superiority of the second option. Filtered rules are the rules that were not pruned. The greatest insight obtained using knowledge of the world was achieved in exchange of the expansion of a set of rules that was already too large.
The ideal scenario is one where the fraction of filtered rules is as little as possible. It has been observed that rules enhancement in post-processing prunes a smaller fraction of rules than enhancement in the pre-processing. But since the number of rules that enters the filter is substantially lower, the cardinality of the mined rules set is reduced without creating spurious rules and with increasing insight for analysts being a clear advantage.

4. CONCLUSION

This paper described a study of using a new pos processing technique to filter association rules in order to reduce recall while maintaining precision. Related work can be found in (Srikant, 1995) and Bur, 2006). In the first the filtering was made by frequency (degree of interestingness) and in the latter the associations rules enhancement was made during pre processing step and only for the "is-a" type of relationships.
Recall reduction is already efficiently obtained by syntactic filtering measures that use rules' objective interest measures. Nevertheless, the simple fact that the number of competing objective measures runs in several dozens is already an indication that none of them are absolute. When analyzing the pruned rules by these various measures it becomes clear that the degree of superposition is low and, therefore, a loss in precision is inevitable.
Our suggestion, which uses knowledge of the world obtained from domain ontology, may enhance the semantics of the rules set obtained and substantially reduce its cardinality. Furthermore, semantic rules pruning does not eliminate relevant rules nor does it fail to eliminate redundant rules.
The results of the experiments obtained with the public Databases showed that the suggested model fully met the desired goals. As a side effect, the integration passage of the pre-processing ontology to the post-processing significantly reduces computational costs, as shown in the number of rules to treat (tables 1 and 2).
The study also contemplates the application of the proposed model to other types of relationships between domain items, such as the "part-of" items.

REFERENCES

Agrawal, R., Imielinski, T., Swami, A., 1993 Mining Association Rules between Sets of Items in *Large Databases in Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*. Washington DC, P. Buneman and S. Jajodia, Eds. ACM Press, 207–216.

Agrawal, R.; Srikant, R., 1994 Fast algorithms for mining association rules. In: Bocca, J. B.; Jarke, M.; Zaniolo, C. (Ed.). *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*. Morgan Kaufmann, 1994. p. 487–499. ISBN 1-55860-153-8. Santiago, Chile. Available at: <citeseer.ist.psu.edu/article/agrawal94fast.html>.

Domingues, M., Rezende, S., 2004. *Descrição de um algoritmo para generalização de regras de associação*. [S.l.]. ICMC/USP Technical Report- Number 228. São Carlos - Brazil.

ECML/PKDD 2004 *Discovery Challenge Homepage*. Available at: [http://lisp.vse.cz/challenge/ecmlpkdd2004/].

Lavrac, N.; Flach, P.; Zupan, B., 1999 *Rule evaluation measures: A unifying view*. Dzeroski, S., Flach, P. (Eds.) ILP-99, LNAI 1634, pp. 174-185. Springer-Verlag Berlin Heidelberg

Merz, C., and Murphy. P., 1998 *UCI repository of machine learning databases*. University of California, Irvine, Department of Information and Computer Sciences. Available at: [http://www.ics.uci.edu/mlearn/MLRepository.html].

Srikant, R.; Agrawal, R., 1995. Mining generalized association rules. *Proc. of the 21st Int'l Conference on Very Large Databases*, p. 407–419, Zurich, Switzerland.

Bürkle, P. Um Método de Pós-processamento de Regras de Associação com Base nas Relações de Dependência entre os Atributos. M. Sc. Dissertation — Instituto de Computação, UFF, Niterói, Brazil, 2006.