Pre-publication draft of a paper which appeared in Lecture Notes in Computer Science Vol. 3518: Proceedings of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD 2005), pages 71-80. Berlin: Springer

Pruning Derivative Partial Rules During Impact Rule Discovery

Shiying Huang and Geoffrey I. Webb

Faculty of Information Technology
Monash University
Melbourne VIC 3800, Austrailia
{Shiying.Huang, Geoff.Webb}@infotech.monash.edu.au

Abstract. Because exploratory rule discovery works with data that is only a sample of the phenomena to be investigated, some resulting rules may appear interesting only by chance. Techniques are developed for automatically discarding statistically insignificant exploratory rules that cannot survive a hypothesis with regard to its ancestors. We call such *insignificant* rules *derivative extended rules*. In this paper, we argue that there is another type of derivative exploratory rules, which is *derivative* with regard to their children. We also argue that considerable amount of such derivative partial rules can not be successfully removed using existing rule pruning techniques. We propose a new technique to address this problem. Experiments are done in impact rule discovery to evaluate the effect of this derivative partial rule filter. Results show that the inherent problem of too many resulting rules in exploratory rule discovery is alleviated.

Keywords

Exploratory rule discovery, impact rules, rule significance, derivative rules

1 Introduction

Exploratory rule discovery seeks to retrieve all implicit patterns and regularities that satisfy some user-defined set of constraints in a population, with respect to a set of available sample data. The best known such approach is *association rule discovery* [1]. Most approaches seeks rules $A \rightarrow C$ for which there is a correlation between the antecedent A and the consequent C. However, whenever one such rule is found, there is a risk that many derivative and potentially uninteresting rules $A' \rightarrow C'$ will also be found. These *derivative rules* are those for which there is a correlation between the antecedent and the consequent only by virtue of there being a correlation between A and C. For example, if A and C are correlated then for any term B that is unrelated to either A or C, AB will also turn out to be correlated with C.

Considerable research has been devoted to automatically identify and discard such derivative rules. The closed itemset techniques [12, 3, 16] can identify rules for which some elements can be removed without changing the support of the rule. Minimum improvement techniques [6] can to identify rules for which some elements can be removed without decreasing rule confidence. However, since exploratory rule discovery seeks to discover rules characterizing the features in a population, with respect to a given sample, rules may happen to be interesting simply due to sampling fluctuation. Statistical tests are also applied to assess whether there is evidence that no elements can be removed without significantly altering the status of the rule with respect to the population from which the sample data is drawn [11, 5, 9]. However, all these techniques relate only to identifying rules that are derivative due to the addition of irrelevant or unproductive elements.

There exists, however, another type of derivative rules that may also result in many rules that are likely to be of little interest to the user. For any rule $AB \rightarrow C$ which is not derivative from another rule and for which there is a correlation between the antecedent and the consequent, both A and B may each be correlated with C solely due to correlation between AB and C. In this case $A \rightarrow C$ and $B \rightarrow C$ will both be potentially uninteresting derivative rules that may be discovered by an exploratory rule discovery system.

The following example illustrates an occasion where such a potentially uninteresting rule may be generated.

Example 1. Suppose a retailer is trying to identify the groups of customers who is likely to buy some new products. After applying the impact rule discovery with the rule filters proposed by Huang and Webb [9, 10], two rules are identified as solutions:

$$District = A \rightarrow profit(coverage = 200, mean = 100)$$
$$District = A \ \& \ age < 50 \rightarrow profit(coverage = 100, mean = 200)$$

Although these two rules are both "significant" as is identified by the rule filter proposed by Huang and Webb [9], the first rule, which is an ancestor of the second one is misleading. Actually, no profit is produced by customers who belong to district A and are older than 50 years! The retailer's attention should more sensibly focus on the group of customers who are under age 50 in district A, instead of on all those in district A. Keeping the first rule in the resulting solutions may confuse the decision makers.

Impact rule discovery is a type of exploratory rule discovery that seeks rules for which the consequent is an undiscretized quantitative variable, referred to as the *target* and is described using its distribution. This paper investigates the identification of the second type of derivative rules in the context of impact rule discovery [9, 14].

The rest of this paper is organized like this: a brief introduction to exploratory rule discovery related concepts is presented in section 2. The definitions and notations of impact rule discovery is charaterized in section 3. Derivative impact rules are defined and relationship between different rules are clarified in section 4, together with the implementation of the derivative rule filter in section 3. Experimental results are evaluated in section 5, which is followed by our conclusions in section 6.

2 Exploratory Rule Discovery

Many machine learning systems discover a single model from the available data that is expected to maximize some objective function of interestingness on unknown future data. Predictions or classifications are done on the basis of this single model [15]. However, alternative models may exist that perform equally well. Thus, it is not always sensible to choose only one of the"best" models. Moreover the criteria for deciding whether a model is best or not also varies with the context of application. Exploratory rule discovery techniques overcome this problem by searching for multiple models which satisfy certain user-defined set of constraints and present all these models to the users to provide them with alternative choices. Greater flexibility is achieved in this way.

Exploratory rule discovery techniques [9] are classified into propositional rule discovery which seeks rules with qualitative attributes only and distributional-consequent rule discovery which seeks rules with undiscretized quantitative variables as consequent. Propositional rules are composed of Boolean conditions only. While the status or performance of the undiscretized quantitative attributes in distributional-consequent rules are described with their distributions. *Association rule discovery* [1], *contrast sets discovery* [5] and *correlation rule discovery* [8] are examples of propositional exploratory rule discovery, while *impact rule* [14] or *quantitative association rule discovery* [2], as is variously known, belongs to the class of distributional-consequent rule discovery. It is argued that distributional-consequent rules are able to provide better descriptions of the interrelationship between quantitative variables and qualitative attributes.

Considering the differences between propositional rule discovery and distributional-consequent rule discovery, there are inherent differences between the techniques for propositional and distributional-consequent rule pruning and optimizations. Researchers have devoted extensive efforts to develop rule pruning and optimization techniques. Reviews of such work can be found in many related works [9].

We define some key notions of exploratory rule discovery as follows:

1. For propositional rule discovery, a *record* is an element to which we apply Boolean predicates called conditions, while for distributional-consequent rule discovery, a record is a *pair* $< c, v >$, where c is the nonempty set of Boolean conditions, and v is a set of values for the quantitative variables in whose distribution the users are interested.
2. Rule r_1 is a parent of r_2 if the body of r_1 is a subset of the body of r_2. If the cardinality of the body of r_1 is smaller than that of r_2 by 1, then the second rule is referred to as a *direct parent* of the first rule, otherwise, it is a non-direct ancestor of the first rule.
3. We use the notion $coverset(A)$, where A is a conjunction of conditions, to represent the set of records that satisfy A. If a record x is in $coverset(A)$, we say that x is *covered* by A. If A is an $\emptyset$, $coverset(A)$ includes all the records in the database. $Coverage(A)$ is the number of records covered by A. $coverage(A) = |coverset(A)|$.

3 Impact Rule Discovery

We construct our impact rule discovery algorithm on the basis of OPUS [13] search algorithm, which enables successful discovery of the top k impact rules that satisfy a certain set of user-specified constraints.

We characterized the terminology of k-optimal impact rule discovery to be used in this paper as follows:

1. An impact rule takes the form of $A \rightarrow target$, where the target is describe by the following measures: *coverage*, *mean*, *variance*, *maximum*, *minimum*, *sum* and *impact*. This is an example of impact rules discovered by our algorithm:
$$Address = Brighton \ \& \ profession = programmer \rightarrow income$$
$$(coverage : 23\%, mean : 60000, variance : 4000, max : 75000,$$
$$min : 44000, sum : 1380000, impact : 3903.98)$$
2. $Impact$ is an interestingness measure suggested by Webb [14][1]: $impact(A \rightarrow target) = (mean(A \rightarrow target) - \overline{targ}) \times coverage(A))$.
3. An k-optimal impact rule discovery task is a 6-tuple:
$KOIRD(D, \mathcal{C}, \mathcal{T}, \mathcal{M}, \lambda, k)$.
   - D: is a nonempty set of records, which is called the database. A record is a pair $< c, v >, c \subseteq \mathcal{C}$ and v is a set of values for $\mathcal{T}$. D is an available sample from the global population $\mathcal{D}$.
   - $\mathcal{C}$: is a nonempty set of Boolean conditions, which are the set of available conditions for impact rule antecedents, which is generated from the given data in D.
   - $\mathcal{T}$: is a nonempty set of the variables in whose distribution we are interested.
   - $\mathcal{M}$: is a set of constraints. A constraint is a criteria which the resulting rules must satisfy.
   - λ: $\{X \rightarrow Y\} \times \{D\} \rightarrow \mathcal{R}$ is a function from rules and databases to values and defines an interestingness metric such that the greater the value of $\lambda(X \rightarrow Y, D)$ the greater the interestingness of this rule given the database.
   - k: is a user specified integer number denoting the number of rules in the ultimate set of solutions for this rule discovery task.

Pseudo code of the original algorithm for impact rule discovery is described in table 1. In this table, *current* is the set of conditions, whose supersets (children) are currently being explored. *Available* is the set of conditions that may be added to *current*. By adding the conditions in *available* to *current* one by one, the antecedent of the *current rule*: $New \rightarrow target$, is produced. *Rule_list* is an ordered list of the top-k interesting rules we have encountered by now.

The search space of this algorithm is illustrated in figure 1. Each node in this search space is connected with a potential impact rule, whose antecedent

[1] In this formula, *mean*$(A \rightarrow$ *target*$)$ denotes the mean of the *targets* covered by A, and *coverage*(A) is the number of the records covered by A.

```
Algorithm: OPUS_IR(Current, Available, 𝓜)

 1. SoFar := ∅
 2. FOR EACH P in Available
    2.1 New := Current ∪ P
    2.2 IF current rule New → target does not satisfy any of the prunable constraints in
        𝓜
           THEN go to step 2.
    2.4 ELSE IF current rule New → target satisfies all the nonprunable constraints in 𝓜
           Record New → target in the rule_list;
    2.5 OPUS_IR(New, SoFar, (M));
    2.6 SoFar := SoFar ∪ P
    2.7 END IF
 3. END FOR
```

Table 1. OPUS_IR

is composed of the conditions between the braces. By performing a depth-first search through such a search space, the algorithm is guarantee to access every nodes and generate all potential impact rules. Based on the OPUS structure, powerful search space pruning is facilitated [13], making it suitable for discovering impact rules in vary large, dense databases. The completeness of OPUS based algorithms is proved by Webb [13].

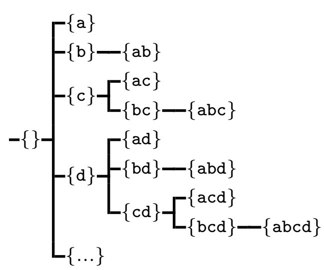

Fig. 1. fixed search space for OPUS_IR

4 Derivative Partial Impact Rules

Techniques for automatically discarding potentially uninteresting rules are extensively explored, examples are the constraint-based techniques, the non-redundant techniques and the techniques regarding the rule improvement and statistically significance. The first classes of techniques seek to identify whether a rule r fails to satisfy the constraints in $\mathcal{M}$. The second class of techniques assess whether the resulting rules are redundant or not by reference to the sample data. Example of non-redundant rule discovery techniques are the closed set related techniques and the trivial rule filter. Each assessment of whether r is desirable is not always free from the risk that the rule is not correct with respect to $\mathcal{D}$ due to the sampling fluctuation [15].

The third class was proposed to reduce the influence of sampling on resulting rules. Statistical tests have been utilized for discarding potentially uninteresting rules generated due to sampling fluctuation, both in context of propositional and distributional-consequent rule discovery. For propositional rule discovery, Brin et al. [8] proposed a pruning technique for removing insignificant correlation rules using a chi-square test; Liu et al. [11] also made use of the chi-square test to identify the significance of association rules with fixed consequents. Bay and Pazzani [5] applied a significance test to remove the insignificant contrast sets in STUCCO. Webb [15] sought to control the number of potentially uninteresting association rules which happen to be interesting due to the sampling by applying a Fisher exact test. For distributional consequent rule discovery, Aumann and Lindell [2] applied a standard z test to quantitative association rule discovery and Huang and Webb [9] also developed an insignificance filter in impact rule discovery whose efficiency is considerably improved by introducing several efficiency improving techniques for rule discovery in very large, dense databases.

However, the techniques mentioned above can only successfully remove a subset of derivative rules.

4.1 Relationship among Rules

As is argued in the introduction, there are derivative rules other than the derivative extended rules that the existing techniques cannot successfully remove. Even after both rules, $A \rightarrow target$ and $A\ \&B \rightarrow target$, have been identified as non-derivative extended rules, there is still a risk that either or both of them are potentially uninteresting. For example, if the target mean of $coverset(A\&\neg B)$ is not significantly higher than the target mean of $coverset(\neg A)$, it can be asserted that the notably high target mean for $coverset(A)$ derives solely from that of $coverset(A\&B)$, which is only a subset of $coverset(A)$. Such rules are defined as *derivative partial rules*, which are insignificant compared to fundamental rules which are their children.

The inter-relationships among different rules are explained in figure 2. In this figure, fundamental rules can also be regarded as non-derivative rules. Derivative extended rules are those referred to as insignificant rules in previous research. Unproductive rules are those exhibit no improvement in target mean comparing with their parent rules. Trivial rules are rules whose antecedents cover exactly the same records as one of their parent rules. As was proved by Huang and Webb [9], trivial rules are a subset of unproductive rules. while those that are productive, with respect to the sample, but fail the significance test are all classified as statistically unproductive.

4.2 Derivative Partial rules and Implementation

we first define derivative partial impact rules:

Definition 1. *A non-derivative extended impact rule, $A \rightarrow target$ is an derivative partial rule, iff there exists a condition x, not included in A, where the target*

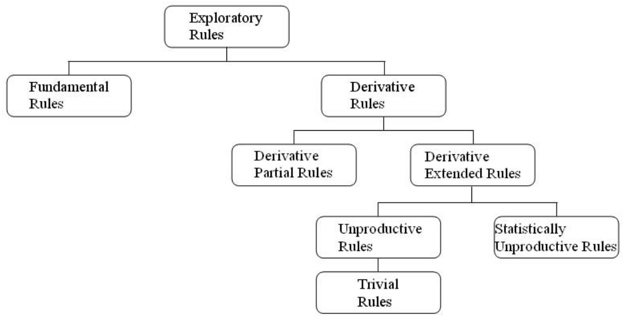

Fig. 2. Relationship of different rules

mean for $coverset(A) - coverset(A\&x)$ *is not higher than the target mean for* $coverset(\neg A)$ *at a user specified level of significance.*

$$InsigPartial(A \rightarrow target) = \exists x \in (\mathcal{C} - \{A\}),$$
$$TarMean(coverset(A\&\neg x)) \succ TarMean(coverset(\neg A))$$

Statistical Test Since by performing the exploratory rule discovery, we are aiming at discovering rules that characterize the features of the population with reference to sample data, hypothesis tests must be done to identify whether an impact rule is derivative or not. A t test is applied to assess whether a partial rule is derivative with regard to its children.

Implementation The new algorithm with derivative partial rule filter is provided in table 2. In this algorithm all the parent rules of the *current_rule* are stored in the *parent_rule_list* while checking whether *current_rule* is a derivative extended rule or not. After *current_rule* is identified as perspectively *fundamental* the derivative partial rule filter is then applied to check whether the parents are derivative with regard to *current_rule*. Derivative parent rules are deleted from the *rule_list*. Since all the parent rules of *current_rule* has already been explored before *current_rule* (please refer to the search space of OPUS_IR), every derivative rule is guaranteed to be removed.

5 Experimental Evaluations

We study the effectiveness of the algorithm in table 2 for the derivative partial rule filter by applying it to 10 large databases chosen from KDD archives [4] and UCI machine learning repository [7], in which many attributes are quantitative. Great differences exist among these databases with the smallest database in size having less than 300 records and the greatest having 2000 times as many records as that of the smallest. Number of attributes vary from only 9 to almost 90. Since complex interrelationships exist among the data, there is a strong necessity

```
Algorithm: OPUS_IR_Filter(Current, Available, M)

 1. SoFar := ∅
 2. FOR EACH P in Available
    2.1 New := Current ∪ P
    2.2 IF New satisfies all the prunable constraints in M except the nontrivial
        constraint THEN
      2.2.1 current_rule = New → target
      2.2.2 IF the mean of current_rule is significantly higher than all its direct parents
            THEN
        2.2.2.1 add the parent rules to the parent_rule_list
        2.2.2.2 IF the rule satisfies all the other non-prunable constraints in M.
                   THEN record Rule to the ordered rule_list
        2.2.2.3 END IF
        2.2.2.4 FOR EACH parent_rule in parent_rule_list
                   IF parent_rule is a derivative partial rule with regard to
                current_rule
                   THEN delete parent_rule from rule_list.
                   END IF
        2.2.2.5 END FOR
      2.2.3 OPUS_IR(New, SoFar, M)
      2.2.4 SoFar := SoFar ∪ P
      2.2.5 END IF
    2.3 END IF
 3. END FOR
```

Table 2. derivative rule Filter

for rule pruning. We choose a target attribute from among the quantitative attributes in each database, and discretize the rest using a 3-bin equal-frequency discretization. After discretization the numbers of available conditions turn out to be over 1500 for some of the databases. The significance level for the derivative rule filters is 0.05.

We did the experiments using following protocol. First, the program in table 1 is run using the insignificance filter proposed by Huang and Webb [9] to find the top 1000 significance rules from each database, with maximum number of conditions on rule antecedent set to 3, 4 and 5. Then, the algorithm with derivative partial rule filter in table 2 is executed to remove derivative partial rules from the resulting solutions. Results are organized in table 4. The numbers of fundamental rules found after both filters are applied are those before the slashes. Integers after the slashes are those found using the insignificance filter only. Decreases in resulting rules are also presented in percentage.

Here is an example of an impact rule which is discarded as derivative partial

$$Sex = M \rightarrow Shucked\_weight(coverage : 1528, mean : 0.432946,$$
$$variance : 0.049729, min : 0.0065, max : 1.351, sum : 661.542, impact : 112.428)$$

It is derivative regarding its parent rule:

$$Sex = M\&1.0295 <= Whole\_weight \rightarrow Shucked\_weight(coverage : 687,$$
$$mean : 0.619229, variance : 0.0284951, min : 0.315, max : 1.351,$$
$$sum : 425.411, impact : 178.525)$$

In this example, if an abalone is male but have a whole weight less than 1.0295 cannot have a very high shucked weight. The first rule is thus misleading!

database	**records**	**attributes**	**conditions**	**Target**
Abalone	4117	9	24	Shuckedweight
Heart	270	13	40	Max heart rate
Housing	506	14	49	MEDV
German credit	1000	20	77	Credit amount
Ipums.la.97	70187	61	1693	Total income
Ipums.la.98	74954	61	1610	Total income
Ipums.la.99	88443	61	1889	Total income
Ticdata2000	5822	86	771	Ave. income
Census income	199523	42	522	Wage per hour
Covtype	581012	55	131	Elevation

Table 3. Basic information of the databases

database	MNC=3		MNC=4		MNC=5	
Abalone	82/86	4.65%	127/138	7.97%	149/173	13.87%
Heart	43/57	24.56%	63/80	21.25%	81/100	19.0%
Housing	131/171	23.39%	168/255	34.12%	192/288	33.33%
German credit	152/197	22.84%	213/273	21.98%	222/295	24.75%
Ipums.la.97	949/1000	5.1%	867/1000	13.3%	809/1000	19.1%
Ipums.la.98	944/1000	5.6%	890/1000	11.0%	761/1000	23.9%
Ipums.la.99	959/1000	4.1%	930/1000	7.0%	896/1000	10.4%
Ticdata2000	803/1000	19.7%	739/1000	26.1%	674/1000	32.6%
Census income	894/1000	10.6%	776/1000	22.4%	744/1000	25.6%
Covtype	918/1000	8.2%	829/1000	17.1%	733/1000	26.7%

Table 4. Experimental results

From the experimental results in table 4, we make the following observation: When the number of maximum conditions on rule antecedent increases, generally, more derivative partial rules are produced by the impact rule discovery system. The greatest change for the numbers of resulting rules after the derivative partial rule filter is applied is as much as 34%. Even the database with a slightest change saw a decrease of over 4%. This justify the argument that there are considerable amount of derivative partial rules still exist in the resulting rules even after the derivative extended rule filter (insignificance filter) is applied. The derivative partial rules can be pruned using our proposed algorithm in reasonable period of time.

6 Conclusions

Exploratory rule discovery searches for multiple models within a set of given data to represent the underlying patterns or regularities. However, it often results in large numbers of rules. Sometimes, the resulting rules are too numerous for human to analysis. Research has investigated techniques for automatically discarding potentially uninteresting rules, thus reducing the number of rules and removing those that are unlikely to be of fundamental interest. One class of these techniques is to apply statistical tests to the resulting models, so as to alleviate the risk of accepting rules which appear to be interesting by reference to the given data which is only a sample, instead of the real world population. In this paper, we argued that there is a type of potentially uninteresting rules which existing techniques fail to remove. We call these rules derivative rules. A derivative rule filter is developed in a impact rule discovery system. Experiments showed a considerable decrease in the number of resulting rules.

References

1. R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*.
2. Y. Aumann and Y. Lindell. A statistical theory for quantitative association rules. In *Knowledge Discovery and Data Mining*, pages 261–270, 1999.
3. Yves Bastide, Nicolas Pasquier, Rafik Taouil, Gerd Stumme, and Lotfi Lakhal. Mining minimal non-redundant association rules using frequent closed itemsets. pages 972–986, 2000.
4. S. D. Bay. The uci kdd archive [http://kdd.ics.uci.edu], 1999.
5. S.D. Bay and M.J. Pazzani. Detecting group differences: Mining contrast sets. In *Data Mining and Knowledge Discovery*, pages 213–246, 2001.
6. Roberto J. Bayardo, Jr., Rakesh Agrawal, and Dimitrios Gunopulos. Constraint-based rule mining in large, dense databases. *Data Min. Knowl. Discov.*, 4(2-3):217–240, 2000.
7. C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.
8. S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In Joan Peckham, editor, *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*, pages 265–276. ACM Press, 1997.
9. Shiying Huang and Geoffrey I. Webb. Discarding insignificant rules during impact rule discovery in large database, 2004. Accepted for publication in SIAM Data Mining Conference 2005.
10. Shiying Huang and Geoffrey I. Webb. Efficiently identification of exploratory rules' significance, 2004.
11. B. Liu, W. Hsu, and Y. Ma. Pruning and summarizing the discovered associations. In *Knowledge Discovery and Data Mining*, pages 125–134, 1999.
12. Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Closed set based discovery of small covers for association rules. In *Proc. 15emes Journees Bases de Donnees Avancees, BDA*, pages 361–381, 1999.
13. G. I. Webb. OPUS: An efficient admissible algorithm for unordered search. *Journal of Artificial Intelligence Research*, 3:431–465, 1995.
14. G. I. Webb. Discovering associations with numeric variables. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 383–388. ACM Press, 2001.
15. G.I. Webb. Statistically sound exploratory rule discovery, 2004.
16. M. J. Zaki. Generating non-redundant association rules. In *Knowledge Discovery and Data Mining*, pages 34–43, 2000.