Interestingness and Pruning of Mined Patterns

Devavrat Shah[1] **Laks V. S. Lakshmanan**[1, 2] **Krithi Ramamritham**[1, 3]
S. Sudarshan[1]

[1] Indian Institute of Technology, Bombay
[2] Concordia University [3] Univ. of Massachusetts, Amherst
{devavrat,krithi,sudarsha}@cse.iitb.ernet.in
laks@math.iitb.ernet.in

Abstract

We study the following question: when can a mined pattern, which may be an association, a correlation, ratio rule, or any other, be regarded as interesting? Previous approaches to answering this question have been largely numeric. Specifically, we show that the presence of some rules may make others redundant, and therefore uninteresting. We articulate these principles and formalize them in the form of *pruning rules*. Pruning rules, when applied to a collection of mined patterns, can be used to eliminate redundant ones. As a concrete instance, we applied our pruning rules on association rules/positive association rules derived from a census database, and demonstrate that significant pruning results.

1 Introduction

Data mining can be described as the process of finding interesting patterns in large databases. A lot of work has focused on defining the notion of "interesting" patterns. Common approaches use statistical measures for finding interesting patterns. Patterns whose value with respect to a given measure exceeds a user-specified threshold are considered as interesting patterns. Several such measures have been proposed in the literature; some examples being the support-confidence measure [AIS93b], correlation measure [BMS97], ratio rules [KLKF98] and strongly collective patterns [AY98].

However, currently used techniques for defining interestingness have a major drawback in that along with the desired patterns they also generate redundant patterns. By this we mean that the same semantic information is captured by multiple patterns, and hence some of them can be pruned. This redundancy exists because each pattern is selected individually without taking into account the other patterns selected.

Example 1.1 To motivate the problem, we consider an example from census data. Each tuple corresponds to one person, and contains the attribute values presented in Table 1.[1] In this case, suppose we discover the following association rules:

1. `drives alone` $\wedge$ `born in US` $\rightarrow$ `not veteran`
   : confidence = 0.72
2. `drives alone` $\rightarrow$ `not veteran`
   : confidence = 0.67

These two rules have *similar* confidence, and the rules have the same r.h.s. part, namely person `not veteran`. The l.h.s. of the first rule is logically subsumed by the second rule and hence using first order logic, the second rule implies the first rule. We therefore say that the first rule is *redundant*, and only the second rule is interesting. □

Thus, we focus our attention on the issue of pruning out redundant patterns using the information gleaned from the other patterns generated. In this paper we describe our notion of "redundancy" in a rule set, and present strategies to prune out such redundant rules. Our notion of interestingness consists of adding the notion of redundancy on top of existing notions of interestingness. In other words, after interesting rules have been generated by any of the current measures, our techniques prune out those that are redundant.

In Section 2, we present the formal framework used in the presentation. We discuss pruning techniques for detecting redundant rules in Section 3.

[1] The census databases are those used in [BMS97] and [BMS98]. The division of attributes into cause/effect shown in the table is added by us.

No	Attribute	type	No	Attribute	type
0	`drives alone`	cause	1	`male or < 3 kids`	cause
2	`not veteran`	effect	3	`native land English`	cause
4	`not US citizen`	cause	5	`born in US`	cause
6	`married`	effect	7	`<= 40 years old`	cause
8	`male`	cause	9	`householder`	cause
10	`white`	cause	11	`Speaks Only English`	cause
12	`US citizen`	cause	13	`Speaks English well`	effect
14	`moved in past 5 years`	effect			

Table 1: Census Attributes

The results of our pruning techniques on rules generated from a real-life census databases are discussed in Section 4. Section 5 covers related work. Finally, in Section 6 we conclude, with directions for future work.

2 Rule Framework

In this section we present a cause-effect formalism for rules, which we use to motivate and present our pruning techniques. Although useful for motivation, the cause-effect framework is not essential; our pruning techniques can be applied, for e.g., on association rules. We assume the database consists of tuples, each with an associated set of attributes. Attributes can be divided into:

- *Cause Attributes* : These are the attributes that can occur as causes in the rules to be discovered. For example, in the census information, attribute `male` can be considered as a cause.
- *Effect Attributes* : These are the attributes which can occur as effects in the rules to be discovered. For example, in the census information attribute `married` can be considered as effect.

In general, cause and effect attributes may overlap, and in a limiting case, such as for association rules, all attributes may be considered to be both causes and effects.

Implication rules are rules of the form:

$$P_a\ (cause_1, cause_2, ..., cause_k) \longrightarrow P_o(\text{eff}_1, ..., \text{eff}_j) : [\texttt{Value(measures)}]$$

where, $cause_i$ is a *cause attribute*, eff_l is an *effect attribute* , and P_a, P_o are predicates.

Further, these rules must satisfy certain conditions which are dependent on the statistical measures used. For example, consider association rules in the market basket domain as defined in [AIS93b]. Such rules are of the form $itemset_1 \to itemset_2$. To understand such rules in our framework, we simply treat $itemset_i$ as a predicate requiring that all the attributes in the itemset be present. The conditions to be satisfied by association rules are:

1. Pr $[itemset_2|itemset_1] > Conf$, a user-define threshold.
2. *Support* ($itemset_1 \wedge itemset_2$) $> Supp$, a user-defined threshold. *Support* of an itemset in a given database is the fraction of the database containing that itemset.

One can define the notion of *strength* for implication rules based on their measure values. For instance, the strength of an association rule is the confidence of the rule, i.e. Pr $[itemset_2|itemset_1]$.

We outline below three classes of implication rules which we use in the rest of the paper.

- **Positive Rule :** A positive rule, denoted $A \xrightarrow{+} B$, where A are B are predicates on *cause* and *effect* attributes respectively, is an implication rule where the presence of the cause A is found to increase the probability of B's occurrence significantly. Formally, this means that for a given user-defined coefficient $P > 1$, $\Pr[B|A] > P * \Pr[B]$ should be satisfied. The strength of the rule is given by $\Pr[B|A]/\Pr[B]$).
- **Negative Rule :** A negative rule, denoted $A \xrightarrow{-} B$, where A are B are as before, is an implication rule where for a user-defined coefficient $N > 1$, $\Pr[B] > N * \Pr[B|A]$. For statistical significance, we also require that if A and B had been independent, they would be expected to occur often enough together; this is ensured by the additional constraint $\Pr[B] * \Pr[A] > thr$ where thr is a user defined threshold. The strength of the rule is given by $\Pr[B]/\Pr[B|A]$.

- **Subsumption Rule :** A subsumption rule, denoted by $A \xrightarrow{\subset} B$, is an implication rule that has a very high confidence. In this case, we permit implication rules where either both left hand side and right hand side predicates use only cause attributes, or both use only effect attributes. The motivation behind this classification is to see whether one cause/effect is subsumed by another cause/effect. Formally, for user-defined parameters $Conf_{sub}$ and $Supp_{sub}$, an implication rule as above is a subsumption rule if $\Pr[B|A] > Conf_{sub}$, with the relevant sets having greater support than $Supp_{sub}$.

  Subsumption rules can be understood as association rules between predicates on attributes, with the additional condition of high confidence.

One can extend this classification to different measures, for example the one presented in [AY98].

3 Pruning Techniques

At an abstract level we can describe the goal of pruning as minimizing the set of causes for a specific set of effects. Similarly, we would like to maximize the set of effects for a specific set of causes. We present several pruning techniques to achieve this.

Before we describe the pruning rules, we need the following definitions.

Redundant Rules : The rules that we prune away with our techniques are redundant rules.

Weak Rules : These are the rules that have been generated as valid rules using the statistical measure, but due to the presence of alternative causes, their validity may be questionable.

Strong Rules : Rules that are neither redundant nor weak are called strong rules.

We say that two rules are of *similar strength* if for a small pre-defined value $1 > \epsilon > 0$, $|strength(rule_1) - strength(rule_2)| < \epsilon$.

Pruning Rule 1: If there are two implications of the form $A \to C$ and $A \wedge B \to C$, and either both rules are positive or negative with similar strength, then $A \wedge B \to C$ is *redundant*.

Justification: This follows from first order logic.

Example 3.1 For the census database, say we discover two rules as :

1. `drives alone` $\wedge$ `born in US` $\to$ `not veteran` : confidence = 0.72
2. `drives alone` $\to$ `not veteran` : confidence = 0.67

With $\epsilon = 0.06$, pruning rule 1 implies that the first rule is redundant. □

Pruning Rule 2: If there are two implications, $A \longrightarrow C$, $B \longrightarrow C$, both either positive or negative rules with similar strength, then $B \longrightarrow C$ is redundant if, $B \xrightarrow{\subset} A$, but $A \xrightarrow{\subset} B$ is not true.

Justification : This rule handles the case when we have two implications for the same effect but the first implicant *subsumes* the second to a large extent. That is, whenever the second implicant is true, (in most cases) the first implicant is also true, and both rules imply the same effect. Hence it is justified to classify the second implicant as *redundant* on the given data.

Example 3.2 Suppose we discover the following two rules of similar strength :

1. `male` $\wedge$ `householder` $\to$ `married`
2. `(male or` $<$ `3 kids)` $\wedge$ `householder` $\to$ `married`

Note that (male or $<$ 3 kids) is a single attribute in the census database. It seems intuitive from the name of this attribute that the cause of the first rule should be subsumed by the cause of the second rule, although it is not a logical implication. In fact, the subsumption is discovered during the mining process. Subsumption does not hold in the opposite direction, and hence by pruning rule 2, the first rule is redundant. □

Pruning Rule 3: If there are two implications, $A \longrightarrow C$ and $B \longrightarrow C$, both either positive or negative rules with similar strength, then both of these rules are *weak* rules, if, $B \xrightarrow{\subset} A$, and $A \xrightarrow{\subset} B$.

Justification: Since we have two implications in which both implicants occur together very often (each subsumes the other) it is not clear which is the cause for the effect; it is also possible that they have a common cause which is also the cause of the observed effect. Hence these are *weak* implications.

Example 3.3 We discover the following two rules of similar strength as :

1. `native lang English` $\wedge$ `born in US` $\to$ `not veteran`
2. `native lang English` $\wedge$ `<= 40 years old` $\to$ `not veteran`

Both the causes were found to be subsumed by each other, and hence by pruning rule 3, we categorize them as *weak* rules. □

Pruning Rule 4: If $A \to C_1$ and $A \to C_1 \wedge C_2$, then $A \to C_1$ is redundant.

Justification: Effect $C_1 \wedge C_2$ is stronger than C_1 in logical sense. Hence the rule $A \to C_1$ is redundant.

Example 3.4 The following rules of similar strength are discovered :

1. `white` $\wedge$ `US citizen` $\rightarrow$ `speaks English well`
2. `white` $\wedge$ `US citizen` $\rightarrow$ `speaks English well` $\wedge$ `moved in past 5 years`

By pruning rule 4, the first rule is redundant. □

Pruning Rule 5: If $A \rightarrow C_1$ and $A \rightarrow C_2$, and further $C_1 \xrightarrow{\subset} C_2$, but not $C_2 \xrightarrow{\subset} C_1$, then $A \rightarrow C_2$ is redundant.
Justification: For a given cause, since the effect of the first rule is subsumed by the effect of the second, but not vice versa, the first rule is stronger in the logical sense. The second rule is actually implied by the first, and hence it is considered redundant.

Example 3.5 The following rules of similar strength are discovered :

1. `white` $\wedge$ `Speaks Only English` $\rightarrow$ `speaks English well`
2. `white` $\wedge$ `Speaks Only English` $\rightarrow$ `moved in past 5 years`

We further determine from the database that `moved in past 5 years` $\xrightarrow{\subset}$ `speaks English well` but not vice versa. By pruning rule 5, the first rule is redundant. □

These pruning rules applied together on the rules generated give a pruned set of rules. Application of these pruning rules should not be overlapping in the sense that if one rule is pruned once, it should not be considered to prune other rules. Pruning rules 1, 2 and 3 are used first to prune the rule space, while pruning rules 4 and 5 are applied on the remaining strong rules. The resulting set of rules is a minimal set of interesting rules given these pruning rules. Changing the order of application of these pruning rules may change the residual minimal rule set.

4 Experimental results

To see the effectiveness of our pruning rules, we implemented our strategies as a post-pass to rule generation. We tested our pruning rules on two real-life census datasets (these are similar to the ones used in [BMS97] and [BMS98]). We manually divided the attributes into cause and effect classes as shown in Table 1.

Due to lack of space we present results only for the following parameter values: $\epsilon = 0.06$, *support* $= 0.1$, $confidence = 0.5$, $N = 2$, $P = 2$, $confidence_{sub} = 0.9$.

Table 2 shows the results of sequentially applying the first 3 pruning rules in the order 1, 2, 3. Once a rule is pruned it is removed from the set, before applying the subsequent pruning rules. We do not show the effects of rules 4 and 5 – our experiments showed that although they did provide good pruning on the original rules, they almost always pruned only rules that were pruned away by the first three rules.

The first three rows deal with constrained rules where attributes are categorized as causes and effects. The last row corresponds to results for general association rules, i.e., without the cause effect semantics being considered. The results show that at the end of the application of pruning rules, many of the rules are pruned out. As shown in Table 2, the first row corresponds to general cause-effect implication rules rules (neither positive nor negative). One can see that 93 rules are pruned out as redundant rules from 125 rules, which is very effective pruning.

We note that almost all the pruning examples presented in the earlier sections were discovered during our experiments. As another example, in the case of census 2 we get the rules:

1. `male` $\wedge$ `Speaks Only English`$\rightarrow$ `speaks English well` : confidence $= 0.54$
2. `male` $\wedge$ `Islander or Indian` $\rightarrow$ `speaks English well` : confidence $= 0.56$
3. `Speaks Only English` $\xrightarrow{\subset}$ `Islander or Indian` : $confidence_{sub} = 0.92$

but *not*,
`Islander or Indian` $\xrightarrow{\subset}$ `Speaks Only English`
By pruning rule 2, the first one is redundant.

As we mentioned earlier, we could treat every attribute as both a cause and an effect, and with support/confidence based filtering, association rules result. The last row of Table 2 illustrates the results of applying our pruning techniques on association rules generated from the first census database.

5 Related Work

We believe that our work is the first to tackle the problem of succinctness in generated rulesets. We have used causality based arguments to prune the ruleset. Causality has been used in earlier work by Brin et. al. [BMS98]. They try to determine causal relationship between items using Bayesian learning techniques, which we do not address. However, they do not carry out any pruning of rules using knowledge of other related rules that are generated.

We would also like to contrast our work with traditional classification [J85]. The goal in classification is to classify the possible cause attributes into different classes based on the effect attribute one is interested in. Note that in the framework presented here, though we have the set of effects beforehand,

DB	Rule Type	Total Rules	Strong Rules	Pruned By Pruning Rule 1	Pruned By Pruning Rule 2	Pruned By Pruning Rule 3
census 1	Cause-effect	125	7	83	10	25
census 2	Positive	277	5	107	83	82
census 2	Cause-effect	893	14	343	226	310
census 1	Association	1131	101	390	86	554

Table 2: Results of Pruning Rules on Census Databases

we do not know on which effects we would like to classify. The pruning rules we apply here give us the minimal set of rules; however it does not reduce to classification because we prune among causes as well as effects. By contrast, the effects are fixed in classification.

In their work on generalized association rules, [SA95] introduce a notion of r-interestingness, which is based on confidences of rules on subclasses being sufficiently different from confidences of rules on superclasses in the hierarchy. This notion corresponds roughly to pruning rule 1 applied on class hierarchies. Chakrabarti et. al. [CSD98] define a notion of interestingness in temporal sequences of itemsets. They define a pattern as interesting if the correlation between the items of a pattern cannot be predicted given correlations of subsets, and correlations at earlier points in time. This measure of interestingness is orthogonal to our work, and does not consider global ruleset knowledge.

6 Conclusion and Future work

Existing statistical measures used for mining interesting rules generate a lot of redundant rules. We have proposed pruning strategies to eliminate these redundant rules. The pruning strategies are independent of specific mining measures, and our performance study indicates that the strategies work well.

An issue that we have not addressed here relates to efficient rule discovery. The pruning techniques that we have described here can perhaps be used in the rule-generation phase itself, instead of as a post-pass on the rule set.

In cases where causes/effects are quantitative or categorical in nature, we can extend our techniques by looking for alternative causes not only for the same effect value, but also for similar effect values. We can also prune based on similar values of causes when causes are categorical or quantitative. For example, suppose we have two rules:

1. `supplier` = `Tata` $\wedge$ `part-type` = `lathe` $\rightarrow$ `life` = `35 yrs`
2. `part-type` = `lathe` $\rightarrow$ `life` = `34 yrs`

then one can exploit the fact that 34 is similar to 35 to conclude that `supplier` = `Tata` is not a crucial factor, and hence the first rule is to be pruned.

We would also like to order causes for a given effect based on the magnitude of the effect. For example, given the rules

1. `supplier` = `ACME` $\rightarrow$ `life` = `2 yrs`
2. `supplier` = `Tata` $\rightarrow$ `life` = `1.5 yrs`

we can order the cause `supplier` = `ACME` as better than `supplier` = `Tata`, due to the longer life when supplier is ACME.

References

[AIS93a] R. Agrawal, T. Imielinski, A. Swamy. Database Mining: A Performance Perspective. In *IEEE Trans. on Knowl. and Data Engg.*, Vol. 5, No. 6, 1993, pp. 914-925.

[AIS93b] R. Agrawal, T. Imielinski, A. Swamy. Mining Association Rules between Sets of Items in Large Databases. In *SIGMOD Conf.* May 1993.

[AS94] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. In *VLDB Conf.*, Sep. 1994.

[AY98] C. C. Aggrawal and Philip S. Yu. A New Framework for Itemset Generation. In *PODS Symp.*, 1998.

[BMS97] Serge Brin, Rajeev Motwani and C Silverstein. Beyond Market Basket : Generalizing Association Rules. In *SIGMOD Conf.*, May 1997.

[BMS98] Serge Brin, Rajeev Motwani and C Silverstein. Scalable Techniques for Mining Causal Structures. In *VLDB Conf.*, 1998.

[CSD98] Soumen Chakrabarti, Sunita Sarawagi and Byron Dom. Mining Surprising patterns using temporal description length. In *VLDB Conf.*, 1998.

[J85] M. James. Classification Algorithms. Wiley, 1985.

[KLKF98] F. Korn, A. Labrinidis, Y. Kotidis and C. Faloutsos. Ratio Rules: A New Paradigm for Fast, Quantifiable Data Mining. In *VLDB Conf.*, 1998.

[NLHP98] Raymond T. Ng, Laks V. S. Lakshmanan, Jiawei Han, and Alex Pang. Exploratory Mining and Pruning Optimizations for Constrained Association Rules. In *SIGMOD Conf.*, May 1998.

[SA95] R. Srikant and R. Agrawal. Mining Generalized Association Rules. In *VLDB Conf.*, 1995.