Reducing the redundancy in association rule mining by frequent closed itemsets

Toshiyuki Suzuki

School of Knowledge Science,
Japan Advanced Institute of Science and Technology
March 2002

Keywords: Association rule mining, frequent closed itemsets, generator

Background and objectives

Knowledge Discovery in Database (KDD) —the rapidly growing interdisciplinary field that merges together database, statistics and machine learning—aims to extract useful and understandable knowledge from large volumes of data. Association rule mining, one branch of KDD, finds interesting associations and correlation relationships among a large set of items. The support defines the range of the rule, and the confidence defines the precision of the rule.

Association rule mining often produced too many rules and many of them are redundant in some sense. Our study is one effort to reduce the redundancy in association rule mining. Therefore we have four concrete objectives.

1. To investigate the problem of non-redundant association rules.
2. To try to formulate another form of non-redundant association rules.
3. To develop an algorithm that finds non-redundant association rules.
4. To try to improve the algorithm.

The framework of association rule mining

The common framework of association rule mining has two-step

1. To find frequent itemsets
2. To generate rules base on frequency itemsets

To find frequency itemsets

Most famous traditional algorithm is Apriori algorithm. This algorithm finds all frequent itemsets in database. But we do not use this algorithm. Because we use closed algorithm that fins frequent closed itemsets in database. This algorithm finds maximal set of itemsets. Therefore we can reduce the frequency items.

To generate rules

Previous work, two people defined definition of non-redundant association rules. Lakhal defined “minimal antecedent and maximal consequent with same support and same consequent”. If we generate the rules based on this definition, we use generator that is minimal itemsets to generate frequent closed itemsets and frequent closed itemsets, This definition indicates to generate only most informative rules. So we can deduce the other rules.

Our new definition based on Lakhal’s definition. We consider “minimal antecedent and maximal consequent in strong” if the rules satisfies the min_supp, i.e. minimum support threshold and min_conf i.e. minimum confidence threshold. Our definition changes the comparison space. This framework of association rule mining sets min_supp, so we consider no care rules support. If we want to know high support rules, we can set the high min_supp. Therefore we do not care each rule’s support. So we can compare in strong space.

The other form definition of non-redundant association rule is “minimal antecedent and minimal consequent with same support and same consequent” given by Zaki. If we generate rules based on this definition, they use frequent closed itemsets. This definition indicates small rules that are low antecedent and low consequent, so we can easily to understand rules. Therefore we can compare rules in generated rules. And so, we later selectively derive other rules of interest.

But Zaki’s algorithm is complicated because they first produce all candidate rules, after that they compare which is minimal. When we consider what does “minimal” mean, we discover the feature of minimal. It is generator of minimal itemsets, because generator produces frequent closed itemsets. So we use generators to generate rules based on Zaki’s definition, and we can generate rules straightforward to use by using generators.

Conclusion

We investigate the problem of redundant association rule. And so we try to formulate the other definition of non-redundant association rules. And we do experiments with some datasets and the above algorithm. We found that our new definition is effective in most cases. And we develop the efficient algorithm previous definition given by Zaki.

Finally we cannot say when or which time to use which algorithm. But we investigate all definition of previous and our non-redundant association rules, so if we analyze some datasets, our research help people who want to discover new knowledge by association rule mining.