Origin based Association Rule Mining using multiple MASP tree[☆]

Elsevier[1]

Radarweg 29, Amsterdam

Elsevier Inc[a,b], Global Customer Service[b,]*

[a] *1600 John F Kennedy Boulevard, Philadelphia*
[b] *360 Park Avenue South, New York*

Abstract

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases.

Keywords: data-mining, Association Rule Mining, frequent-itemset mining

1. Introduction

Association rule mining is a rule-based machine learning procedure to find interesting patterns in the transaction database based on individual and conditional frequencies. In the traditional approach, two steps are involved in generating rules. First, generate all frequent itemsets and pruned non-frequent ones and then in the second stage rules are derived from those frequent itemsets. An association rule e.g. bread, milk ⇒ butter in market basket analysis means if one purchase bread and milk together it is highly likely that they will also buy butter. Apart from market basket analysis, association rule mining is useful in intrusion detection, bioinformatics, and many other applications.

In 2014 Omer M. Soyal [1] proposed a new approach to extract mostly associated sequential patterns (MASPs) using less computational resources in terms

[☆] Fully documented templates are available in the elsarticle package on CTAN.

[*] Corresponding author
Email address: support@elsevier.com (Global Customer Service)
URL: www.elsevier.com (Elsevier Inc)

[1] Since 1880.

of time and memory while generating a long sequence of patterns that have the highest co-occurrence. This approach may produce different outcomes if we change the order of items in transactions. We propose an approach which is order independent. An association rule of the form A $\Rightarrow$ B must satisfy the threshold support and threshold confidence i.e. probability of occurrence of A and B together must surpass threshold support, and the probability of occurrence of B in transactions containing A must be greater than or equal to threshold confidence. It means, to calculate support and confidence, it is required to traverse complete transaction database. To generate all rules containing a particular item x it is reasonable to ignore all transactions(for calculating support and confidence) that come before the transaction in which that particular item appears for the first time. Embedding these two changes to the Omer M. Soyal [1] approach is the basis of our research.

2. Related works

In 1994 Rakesh Agrawal and Ramakrishnan Srikant published non-trivial algorithm(Apriori) [2] for finding association rules in large databases of sales transaction. Apriori algorithm produces association rules in two steps. First generates all frequent itemsets(prune non-frequent candidate itemsets) and then produce rules from those itemsets. This algorithm first find frequent itemsets of length 1 then frequent itemsets of length 2 using frequent itemsets of length 1 and so on till all frequent itemsets is not generated. This algorithm gave better result than the previously known fundamental algorithms AIS [3], SETM [4]. FP-Growth agorithm(2000) [5] also take two steps. Second step is same as apriori. FP-Growth do not generate candidate frequent itemsets. First it generates a tree(FP-Tree) and then find frequent itemsets. This algorithm is about an order of magnitude faster than the Apriori algorithm.

References

[1] O. M. Soysal, Association rule mining with mostly associated sequential patterns, Expert Systems with Applications 42 (5) (2015) 2582 – 2592. doi:http://dx.doi.org/10.1016/j.eswa.2014.10.049.

[2] R. S. Rakesh Agrawal, Fast algorithms for mining association rules, Proceedings of the 20th VLDB Conference Santiago, Chile.

[3] A. S. " R. Agrawal", "T. Imielinski", Mining association rules between sets of items in large databases, ACM SIGMOD Conference on Management of Data, Washington, D.C.

[4] A. S. "M. Houtsma", Set-oriented mining of association rules, Research Report RJ 9567, IBM Almaden Research Center, San Jose, California.

[5] Y. Y. "Jian Pei ", "Jiawei Han", Mining frequent patterns without candidate generation: A frequent-pattern tree approach, Data Mining and Knowledge Discovery.