Research on Association Rules Mining Base on Positive and Negative Items of FP-tree

Chunwei Chen[1, a], Dandong Wang[2, a]

[1]School of Hangzhou Dianzi University, Hangzhou 310000, China

Keywords: Data Mining; Association Rule; FP-tree Positive & Negative items; Apriori

Abstract. This paper analyzes the two kinds of classic traditional association rules algorithm Apriori and FP – tree which those are advantages and disadvantages .In order to solve the Apriori algorithm for mining association rules requiring multiple scanning database and generate a large number of candidate frequent sets, base on the inherited FP -tree algorithm the advantages of scanning second times of database project on the basis of combining the positive and negative association rules mining algorithm, not only can solve real problems in negative association rules mining, can also delete the contradictory relationship, effectively improve the efficiency of the algorithm .The final process of a set of real transaction of mining experiments show that the algorithm to improve the quality and efficiency of mining rules, and to avoid invalid mode rules.

Introduction

Association rule mining is unearth hidden behind the huge amounts of data between data items, people can be found useful value from the connection between the data. It is used to denote dependencies between two items, if two items transaction between certain associated, so one item transactions can be as a prerequisite at a certain probability inference another item. Its purpose is to find items from a large amount of data transaction between interesting association. The theoretical research of it by original frequent patterns to closed model, incremental mining pattern mining, interest measure mining, data stream mining, and other form of association rule mining. Its application from the original basket analysis to classification association rules, knowledge extraction and recommendation, association analysis of traditional Chinese medicine drugs, protein structure analysis, software bug mining, equipment fault diagnosis, the traffic accident model analysis, etc.

Representation of the traditional association rules is $A \Rightarrow B$, originally composed of state Richard armitage grawal and others first proposed in 1993 [1], and in 1994 put forward a fast algorithm of Apriori [2], then many scholars mainly on the efficiency of the association rule algorithm and performance improvement methods are proposed. Wu Xingdong, [3] proposed A model of PR as $A \Rightarrow B$ type associated with an important complement, three forms to the negative association rules ($A \Rightarrow \neg B$, $\neg A \Rightarrow B$, such $A \Rightarrow \neg B$) made a research and analysis. Professor Han Jiawei in 2000 put forward another classic algorithm FP - tree algorithm [4], and solved the Apriori repeatedly scanning database and produce the problem such as a large number of candidate sets. In this paper, the author raises a kind of positive and negative association rules and FP - tree algorithm, can not only inherits the advantages of FP - tree algorithm, and mined frequent item focus is at the same time, the negative association rules, can remove conflicting rules, so as to improve the efficiency and practicability of the algorithm.

Research and analysis of the traditional association rules algorithm

Apriori and FP – tree.

Apriori and FP - tree is two kinds of classic algorithm of association rules. Apriori algorithm is proposed by R.Agrawal and algorithm mining rules are divided into two steps: (1) find all frequent set.(2) Find out all the strong association rules.

The algorithm's disadvantage is the need to repeatedly scan the database, so as to increase algorithm spent reading and writing operation time, resulting in mining time costs rise, the cost with

the increase of data storage is a geometric series rise. Then use the algorithm generated a large number of candidate frequent item set, resulting in poor performance of the algorithm, and even lead to lack of memory. Using the Weka software experiment proved the shortcomings of this algorithm, with a huge amount of data the data set (UCI open-source database retai) analysis of the association rule, caused by insufficient memory space and the algorithm of running crash .

Apriori algorithm cannot meet the needs of modern large data outbreak. The association rules of the researchers put forward a lot of improved algorithm. FP Tree algorithm on data sets for data mining the main idea is to first creating a tree (frequent pattern tree, then map the relationship between database transaction data and data to the frequent pattern tree, created the frequent pattern tree traversal, get the final rule. Mainly divided into three steps: (1)structure FP tree (2) mining frequent item set (3) found Association rules.

The algorithm significantly improves the memory space and time efficiency, using a recursive divide and rule strategy, frequent items mining. To scan the transaction database for database compression for a frequent pattern tree (FP Tree or tree, reserved item set associative information; the compressed data into a set of predicated database (a special type of projection database) [5], each premise database associated with a frequent item to excavate the database of the premise.

Positive & Negative Items Of FP-tree.

But in real life, there are often negative related items caused by misleading Association, the replacement of the problem of false positive correlation relationship. The negative association rule is the complement of the general association rules. Mining negative association rules will be related to many non frequent item set, in order to effectively and produce positive and negative association rules, mining model of Xindong Wu, are a kind of PR, discovering both positive and negative association rules; Xiangjun Dong, et al. gives the multilayer minimum support degree model MLMs at the same time, the minimum reliability and correlation coefficient of positive and negative association rules mining.

In order to solve the positive and negative association rules generated a large number of frequent item sets and the performance of the algorithm problem, a mechanism is proposed for mining positive and negative association rules is improved FP Tree, its characteristics: advantages (1) Inheritance of FP Tree, do not need to scan database repeatedly, do not produce candidate item set; (2) the negative items as similar item transaction inserts a tree is constructed , not expanding the original database, with compressed data structure to store the transaction database of relevant information, different nodes can be shared prefix path; (3) doesn't generate conditional pattern base, do not need to a large number of conditions pattern tree resources waste memory and time overhead structure.

Mining analysis of FP-tree improved algorithm based on positive and negative correlation

Related concepts.

Set I ={ i_1, i_2, i_3,... , i_n } is a collection of items, and the elements in the formula are called items. D is a set of transaction database transactions, where each transaction is a set of T for the $T \subseteq I$. Each transaction has a unique identifier, denoted TID, X, $\Rightarrow$Y is a collection of I items. Rule x y in D is denoted as (support), transaction set contains A and B the number of transaction and all transactions than, denoted as support (x$\cup$ y). That is

Support(X$\Rightarrow$Y) $=\frac{|T:X\cup Y\subseteq T,T\in D|}{|D|}$（3.1）

Rule x$\Rightarrow$ y in D confidence degree is denoted by (confidence, denoted by C) refers to contain a and B, the number of transactions and contains a number of transaction ratio.

Confidence (X$\Rightarrow$Y) = $\frac{|T:X\cup Y\subseteq T,T\in D|}{|T:X\subseteq T,T\in D|}$（3.2）

Theorem 1 $X\subset I, Y\subset I, A\cap B=\emptyset$ is up to:

Support($\neg$X)=1-Support(X)（3.3）

Support(X$\cup\neg$Y)=Support(X)-Support(X$\cup$Y)（3.4）

Theorem 1 $X\subset I, Y\subset I, A\cap B=\emptyset$ is up to:

Confidence(X⇒ ¬Y)=1-Confidence(X⇒Y)（3.5）

Confidence(¬X⇒Y)=$\frac{Support(Y)-Support(X\cap Y)}{1-Support(X)}$（3.6）

Confidence(¬X⇒ ¬Y)=1-Confidence(¬X⇒Y)（3.7）

The procedure of constructing positive and negative association rules FP-tree.

Input: the entire transaction database R, according to the characteristics of the user or the minimum support of experts developed I: FP-tree.

Step1: scanning database one time to produce items all set S, the set is a set and the negative set, produces all the support, and in descending order.

Step2: the second scan database, on a per - transaction Di perform operation: the root node of a tree root=null, if Di contains S in all the di all control sequences S affairs frequent item set [p|P], executive insert action.

Step3: It will be the support of the children of root plus 1, while the P will be generated by the new [p|P]: to continue to perform the insertion operation. Otherwise, the P as the children of root link up; the support of P is set to 1, a new [p|P]Root=p; continue to insert the operation.

Step4: generated FP-tree at this time, the FP-tree implementation of the mining action, resulting in sub FP-tree, generating frequent pattern set.

Final output: frequent pattern sets.

Among them, the meaning of [p|P]: P support for the highest level of a channel, P said the remaining frequent items.

Insert operation, in fact, can be such an understanding of if (root children =p).

Algorithm applied to real transaction data set.

We provide a small business database to illustrate this problem, small transaction data S(T1(i1,i2,i3),T2(i2,i4),T3(i2,i4),T4(i1,i2,i4),T5(i1,i3),T6(i2,i3),T7(i1,i3),T8(i1,i2,i3,i5),T9(i1,i2,i3)).The support statistics of the project of scanning database .

({i1:6,i2:7,i3:6,i4:2,i5:2},{-i1:3,-i2:2,-i3:3,-i4:7,-i5:7})

According to the characteristics of the transaction database, we set up the minimum support number 5, delete the non frequent positive, negative items get{i1:6,i2:7,i3:6,-i4:7,-i5:7}. According to the above construction method, to create a positive and negative items of the FP-tree, as the database of 9 transaction entries into the operation, the structure of the FP-tree as shown Fig. 1:

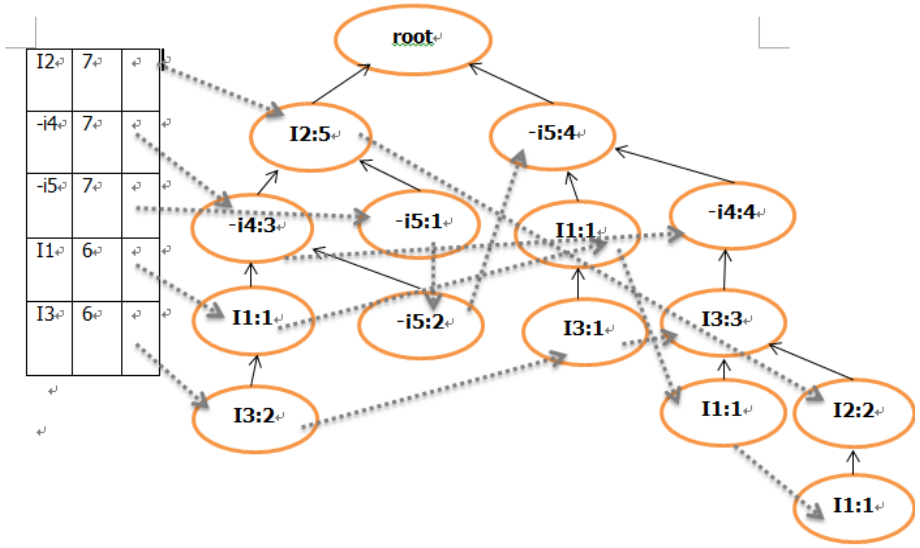

Fig. 1 Construction tree with improved positive and negative items

For positive and negative project FP-tree construction is completed, now frequent 2- items, 3- item set for mining. From here we can see that each item to expand the model, only use the current item conditions for frequent projects and its conditional frequent suffix item, and the item has nothing to do with the current item.

Improved algorithm validation.

Due to small transaction database is very difficult to clear the time-consuming effect comparison algorithms.in order to have a better result, we (UCI open source data from the selected a relatively large source data (retail), the two algorithms run in time, we can map (Fig. 2).

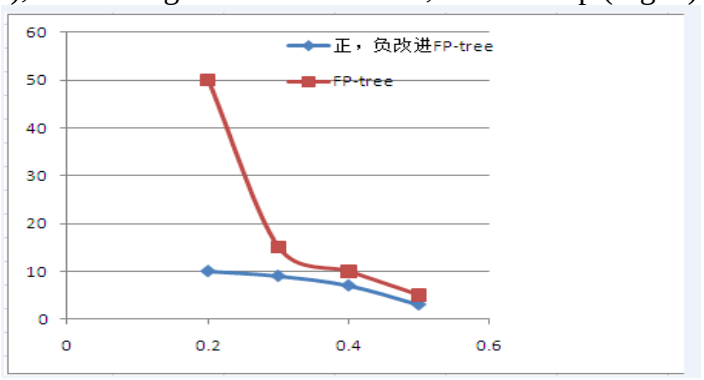

Fig. 2 Comparison of algorithm time consuming performance

Relatively less frequent project due to small transaction database is very difficult to clear the time-consuming effect comparison algorithms, this paper from the UCI open source data) selected a relatively large source data (retail), more frequent project by two algorithms run in time, from that shown in Figure 2 conclusion, when the support parameter is smaller, there is a need to explore more prominent the performance of the proposed algorithm is compared, with the support parameter is larger, need to dig, the performance of algorithm converge.

Table 1 Comparison of the number of mining

algorithm	Positive association rules	Negative association rule
FP-tree	46	0
Positive & Negative Items Of FP-tree	31	9

Summary

Firstly, the advantages and disadvantages of the two kinds of classical algorithm of association rules of Apriori and FP-tree are analyzed, and according to practical data of the experimental test and display, combined with the reality of the existence of the negative influence of association rule mining, this paper leads to a positive and negative based on improved FP-tree mining, the mining algorithm inherits the FP-tree does not need to repeatedly scanning the database, expanding the original database does not generate candidate set, under the condition of considering the positive items and negative items, so that the algorithm is closer to reality, remove the conflicting rules, which saves time and space, and finally the contrast experiment for the improved algorithm, clearly shows that the improved algorithm is efficient higher in actual mining.

References

[1]. Agrawal R, Imielinski T, Swami A N. Mining Association Rules Between Sets of Items in LargeDatabases, SIGMOD Conference[J]. Acm Sigmod Record, 1993, 22:207--216.

[2]. Agrawal R, Imieli&#, Ski T, et al. Mining Assocation Rules between Sets of Items in Large Databases[J]. Proc.conf.on Management of Data, 1993, 22:207--216.

[3]. Wu X, Zhang C, Zhang S. Mining Both Positive and Negative Association Rules.[J]. Proceedings of Int.conf.on Machine Learning, 2002.

[4]. Han J, Pei J, Yin Y, et al. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach[J]. Data Mining & Knowledge Discovery, 2004, 8(1):53-87.

[5]. Wu X, Zhang C, Zhang S. Efficient mining of both positive and negative association rules[J]. Acm Transactions on Information Systems, 2004, 22(3):381-405.