Statistical Approach for Data Mining to find the Frequent Item Sets

N.Venkateshwarlu
Assistant Professor, Department of Computer Science and Engineering
Prasad Engineering College, Jangaon, Andhra Pradesh, India.

K.Vinay Kumar Reddy
Assistant Professor, Department of Computer Science and Engineering
Prasad Engineering College, Jangaon, Andhra Pradesh, India.

Abstract- Association rule induction is a powerful data mining method. It is used to analyze the regularities in data trends by finding the frequent itemset and association between items or set of items.There is a great deal of overlap between data mining and statistics.In fact most of the techniques used in data mining can be in a Statistical frame work.In this paper an algorithm can be proposed for the purpose of finding Frequent Item sets.This algorithm is to capable to generate the frequent data items more close to the real life situations as it consider the Strength of Presence of each items implicitly.

Keywords – Frequent item set, Association rule, Data mining, Apriori algorithm, Frequency.

I. INTRODUCTION

In recent days, the nature of database applications and automated data collection tools lead to tremendous amounts of data stored in operational databases, data warehouses, and other information repositories. These large amounts of data are worthless unless they become knowledge or information using the concept of Knowledge Discovery Databases (KDD) [1, 2]. Data mining is a concept used hugely now days for the purpose of data explosion and knowledge discovery from the large dataset. Further, the concept of Data mining has recently attracted considerable attention from database practitioners and researchers because of its applicability in many areas such as decision support, market strategy and financial forecasts.

One of the common data mining techniques, Association rule induction is a powerful method used to find regularities in data trends [3]. It has two different phases; first is to find the frequent itemsets and the second is to find the rules from these item sets. An association rule expresses an association between items or sets of items. Finding the above association rule is valuable for cross – marketing analysis, attached mailing applications, store layout, and customer segmentation based on buying patterns etc.

The algorithms for mining frequent patterns can be classified into two approaches, namely, candidate generation and pattern-growth. The representative algorithm of the first approach is the Apriori algorithm and Apriori like algorithms [4]. Apriori algorithm traverses the Boolean search space in a pure breadth first manner and finds support information by explicitly generating and counting each node. Another approach is FP-growth algorithm [5]. FP-growth algorithm uses prefix-tree structure to mine frequent itemsets without generating candidates and scans thedatabase only twice. Some other approach like Dynamic Itemset Counting (DIC) [6] algorithm works on the principle of early candidate generation. Recently many other algorithms were developed for mining frequent patterns [7, 8, 9]. But most of them are an enhancement of Apriori Algorithm or FP-growth algorithm and use the technique for mining frequent itemsets in binary search space.

II. OVERVIEW OF APRIORI ALGORITHM

Apriori is the first algorithm that pioneered the use of support-based pruning to exponential growth of candidate itemsets [4] with systematic control. The data is assumed to be represented as binary dataset, where each row corresponds to a transaction and each column corresponds to an item. If any item belongs to a particular transaction, then its corresponding entry is equal to 1 otherwise, it is denoted as 0.

The Apriori principle states, If an itemset is frequent, then all of its subsets must also be frequent. The algorithm is based on the above principle and it iterates over two phases, the phase of candidate generation and the phase of verification, at each level.
Since, Apriori is a level-wise algorithm and it generates frequent itemsets one level at-a-time, from itemsets of level-1 to the longest frequent itemsets. At each level, new candidate itemsets are created by using frequent itemsets discovered at the previous level. At each level, the transaction database is scanned once to determine the actual support count of every candidate itemset. Also, Pruning in Apriori algorithm is essential for the removal of unwanted itemsets and is of prime importance in binary search space.

III.PROPOSED STATISTICAL APPROACH ALGORITHM FOR FINDING FREQUENT ITEM SETS

3.1.PROPOSED ALGORITHM.

Let I= {i1, i2, i3,…, in} be a set of n binary attributes called *items*. Let D= {t1, t2, t3,…., tn} be a set of transactions called the *database*. Each transaction in D has a unique transaction ID and contains a subset of the items in I. To illustrate the concepts, we use a small example from the supermarket domain. The set of items is I={milk,bread,butter,beer} and a small database containing the items (1 codes presence and 0 absence of an item in a transaction) is shown in the following table.

Example database with 4 items and 5 transactions

Transaction ID	milk	bread	butter	beer
1	1	1	0	0
2	0	0	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

An example rule for the supermarket could be {butter, bread} => {milk} meaning that if butter and bread are bought, customers also buy milk.
In this proposed algorithm, we find the set of frequent items by using the factor of each item frequency and also measure maximum number of the frequency. After considering the difference of maximum frequency to the individual frequency for referring the frequent items from the set of items in a particular respected transaction ID. For Finding Frequent data items by using this approach, there is no necessity of calculating minimum support as well as measuring of any minimum confidence.

3.2. THE ALGORITHM

The main function *SA_Frequent_Itemset* is used to generate the frequent itemsets at each level from the binary search space using the function *SA_Frequent_Itemset.*
//k is level identification number, C_k will store all possible itemsets at level k and L_k will store the frequent itemsets at level, F is a Frequency value and MF is a Maximum Frequency.
Function SA_Frequent_Itemset ()
***Initialize**: =1, n=Number of Items, C1=all the level 1 itemsets;*
Compute the dynamic support on each itemset of C1 and use the static support determine L1;
L1 :={ All Frequent itemsets of Level 1};
F :={ Frequency value each and every item in the given Transaction ID};
MF :={ Maximum Frequency count derived from Frequency Entry};
If L1 [MF] =0 V 1 then
Frequency item entry is frequent item set and moved to next iteration;
k:=k+1;
End
Answer:=$U_k L_k$
End Function

3.3 .PROPOSED ALGORITHM EXAMPLE

Let the set I= {I1, I2, I3, I4, I5} represent in the Transaction ID 100 as showed as follows.

TID	**List of Items**
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

From the above table, items are repeated many number of times. So we calculate the term frequency as number of times the item is repeated, then choose the maximum frequency number and also find the difference of each item frequency. If the difference is either 0 or 1, then the items are defined as frequent items, Otherwise the items are removed from list of items. The following table describes the first L1 as follows.

Items	Frequency(F)	D=MF-F
I1	6	1
I2	7	0
I3	6	1
I4	2	4
I5	2	4

(a)List of L1

Here MF referred as Maximum Frequency number in the Frequency Column entry. By using the above table, we also find the next L2 as follows.

Items	Frequency(F)	D=MF-F
{I1,I2}	4	0
{I2,I3}	4	0

(b)List of L2

By using the above table, we also find the next L3 as follows.

Items	Frequency(F)	D=MF-F
{I1,I2,I3}	2	0

(c)List of L3

Since, after the iteration equal to the number of products in the dataset. The algorithm stops. Hence the frequent itemsets obtained for the entire dataset is,

$L := L_1 \cup L_2 \cup L_3$

:={{I1},{I2},{I3},{I4},{I5},{I1,I2},{I2,I3},{I1,I2,I3}}

IV.CONCLUSION

In real life data analysis, predicting the frequent itemset with quantitative measure is a major concern, which can be the first step towards mining the association rules with broader objectives i.e. rules with quantitative measures of items. The proposed *SA Frequent Itemset* algorithm in this paper focuses upon mining association rules from binary dataset, with the consideration of said quantitative measure and focuses on Frequencies.
Further studies will concentrate on testing the algorithm on real environment in order to prove its practical utility and to extending the algorithm for binary dataset related to hierarchical items

REFERENCES

[1] Piatestsky-Shapiro G., "Knowledge discovery in databases", *AAI/MIT Press*, 1991
[2] Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R., "Advances in knowledge discovery and data mining", *AAAI Press*, 1996.
[3] Agrawal R, Imielinski T, Swami A., "Mining association rules between sets of items in large databases", *Intl. Conf. on Management of Data (Proc. ACM SIGMOD)*, Washington, DC, 1993, pp. 207-216.
[4] R. Agrawal, R. Srikant, "Fast algorithms for mining association rules", *Proceedings of the 20th Very Large DataBases Conference (VLDB'94)*, Santiago, Chile, 1994, pp. 487– 499.
[5] J. Han, J. Pei, Y. Yin, "Mining frequent patterns without candidate generation", *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, 2000, pp. 1–12.
[6] http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#autoId5.
[7] http://www.stat.tamu.edu/~eparzen/future.pdf