IJCAES
ISSN: 2231-4946
Volume IV, Special Issue, March 2014

International Journal of Computer Applications in Engineering Sciences
Special Issue on International Conference on Computing, Communication and Sensor Network (CCSN-2013)
www.caesjournals.org

A unique approach for Market Basket Analysis under the framework of Probability Based Association Rule Learning

Abhijit Sarkar[1], Apurba Paul[2], Anupam Mondal[3], Abhijit Ghosh[4]

Dept. of CSE, JIS College of Engineering, Kalyani, Nadia, West Bengal, India

[1]abhi41001@gmail.com, [2]apurba.saitech@gmail.com, [3]link.anupam@gmail.com, [4]abhijitghosh09@gmail.com

Abstract: - **In recent times, Data Mining plays one of the decisive role in business intelligence. Analyzing enormous amount of business transaction data is the order of the hour. Association rule learning is one of the major part in Data Mining that helps us to attain functional patterns understanding buying habits which can help in business decision making, increasing revenues, cutting cost etc. Apriori algorithm is one of the well-researched measure to generate association rules which are related set of data existing in transaction data. In this paper, we present an enhanced approach of market basket analysis under the framework of improved probability based Association Rule learning using the notion of Apriori Algorithm. This algorithm can be effectively used in any number of data in the field of continuous production, web usage mining, bioinformatics etc**.

Keywords: **Apriori algorithm, association rules, itemset, probability of occurrence, support count.**

I. INTRODUCTION

Data mining refers to the mining of new information in terms of patterns or rules from massive amount of data. Successful organizations view such huge databases as important pieces of the marketing infrastructure, as mining these databases gives useful patterns which is used for business decision making.

Data mining is aggravated by the decision support problem faced by most large retail organizations. So, it's always about transforming unprocessed data into business intelligence. In the field of Data mining, Association rule learning helps us to find patterns or rules present in raw data, which helps in the making of business decision. Apriori Algorithm can generate frequent itemsets which can effectively form association rules. Among the many possible association rules out of the frequent itemsets, we further filter them based on confidence threshold. The major drawback of the traditional approach is in the area of frequent scanning of huge database increasing the running time of the processing. In this paper, we have presented a modified probability based Apriori Algorithm for association rule learning to reduce the database scanning.

II. APRIORI ALGORITHM

1. k = 1
2. Find frequent set L_k from C_k of all candidate itemsets
3. Form C_{k+1} from L_k; k = k + 1
4. Repeat 2-3 until C_k is empty

Details about steps 2 and 3

Step 2: scan *D* and count each itemset in C_k, if it's greater than minSup, it is frequent
Step 3:

- For k=1, C_1 = all 1-itemsets.
- For k>1, generate C_k from L_{k-1} as follows:
  - *The join step*

  C_k = join of L_{k-1} with itself (itemset of size k)
  If both $\{a_1,\ldots,a_{k-2}, a_{k-1}\}$ & $\{a_1,\ldots, a_{k-2}, a_k\}$ are in L_{k-1}, then add $\{a_1,\ldots,a_{k-2}, a_{k-1}, a_k\}$ to C_k
  (We keep items **sorted**).
  - *The* prune *step*

  Remove $\{a_1,\ldots,a_{k-2}, a_{k-1}, a_k\}$ if it contains a non-frequent (k-1) subset.

III. ILLUSTRATION OF APRIORI ALGORITHM

Transaction	Item Present
T1	Milk, Bread, Butter ,Sugar
T2	Bread, Butter, Biscuit
T3	Butter, Sugar
T4	Bread, Butter, Sugar
T5	Bread, Sugar, Biscuit

Figure 1: Transaction database

Figure 1 shows the transaction database. It contains 5 different transactions and the items purchased in the respective transactions.

Minimum Support=2

Candidate 1 Itemset	Support Count
{Milk}	1
{Bread}	4
{Butter}	4
{Sugar}	4
{Biscuit}	2

C1

Compare with Minimum Support →

Frequent 1 Itemset	Support Count
{Bread}	4
{Butter}	4
{Sugar}	4
{Biscuit}	2

L1

Figure 2: Generation of candidate 1-itemset and frequent 1-itemset

Figure 2 shows the support count of different 1-itemsets and the selected 1-itemsets based on minimum support threshold value 2 after pruning.

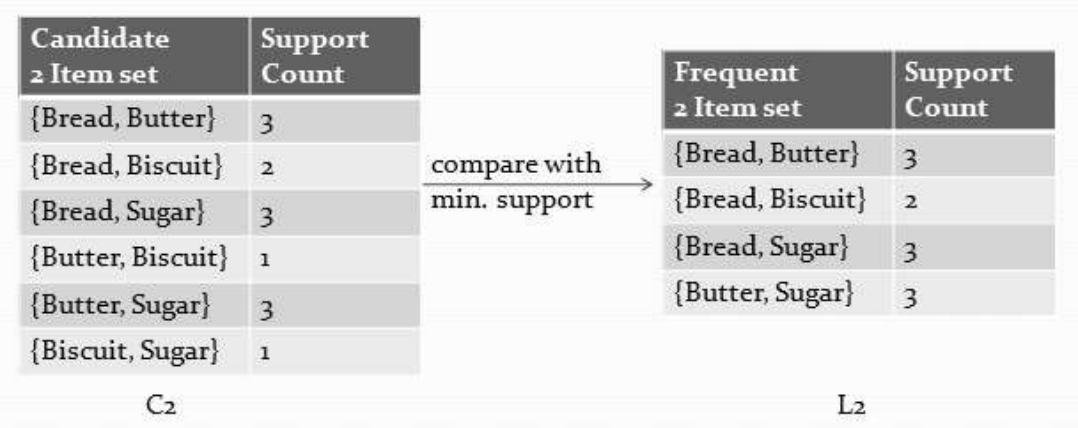

Figure 3: Generation of candidate 2-itemset and frequent 2-itemset

Figure 3 shows the support count of different candidate 2-itemsets from the frequent 1-itemset shown in Figure 2 and the selection of frequent 2-itemsets based on minimum support threshold value 2.

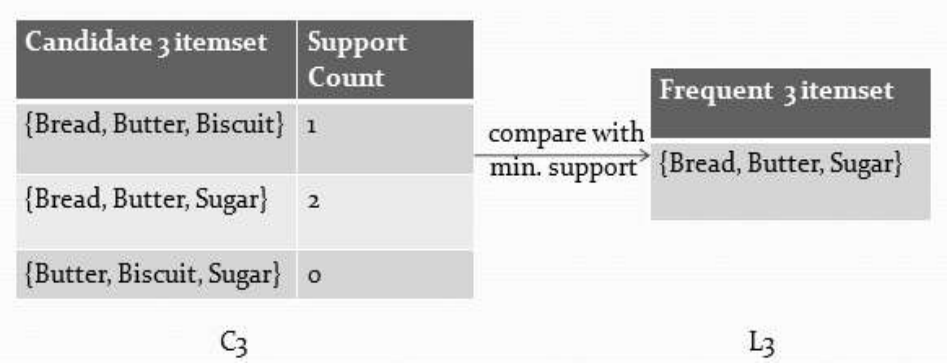

Figure 4: Generation of candidate 3-itemset and frequent 3-itemset

Figure 4 shows the support count of different candidate 3-itemsets and the selected frequent 3-itemset based on minimum support threshold value 2 after pruning and termination of frequent itemset generation. The final frequent itemset contains {Bread, Butter, Sugar}.

Association Rules	Confidence
R1: Bread^Butter->Sugar	Sup. Cnt.{Bread, Butter, Sugar}/Sup. Cnt.{ Bread, Butter}=2/3=67% SELECTED
R2: Bread^Sugar->Butter	Sup. Cnt.{Bread, Butter, Sugar}/Sup. Cnt.{ Bread, Sugar}=2/3=67% SELECTED
R3: Butter^Sugar->Bread	Sup. Cnt.{Bread, Butter, Sugar}/Sup. Cnt.{Butter, Sugar}=2/3=67% SELECTED
R4: Bread->Butter^Sugar	Sup. Cnt.{Bread, Butter, Sugar}/Sup. Cnt.{ Bread }=2/4=50%
R5: Butter-> Bread^Sugar	Sup. Cnt.{Bread, Butter, Sugar}/Sup. Cnt.{ Butter}=2/4=50%
R6: Sugar->Bread^Butter	Sup. Cnt.{Bread, Butter, Sugar}/Sup. Cnt.{ Sugar}=2/4=50%

Figure 5: List of selected association rules based on confidence threshold = 60%

Figure 5 shows the different possible association rules from the item existing in association rule set(frequent 3-itemset) shown in Figure 4 and selection of association rules R1, R2 & R3 among all the rules based on minimum confidence threshold = 60%.

IV. PROBABILITY BASED APRIORI ALGORITHM

Step 1: Transactions database, minimum support, and minimum confidence are taken as input for the processing. Two sets A= {all items}, RuleSet= {Ø} are initialized.

Step 2: Calculate the occurrence of each item in the transactional database & delete items having support count less than the minimum support count threshold in the set A.

Step 3: Select item having highest support count (≥ Minimum support) and insert it in RuleSet. In case, more than one item having same highest support count, select any arbitrary item.

Step 4: Take the combination of two items X & Y together such that X=last selected item to put into RuleSet and

Y ε {items belonging to set difference ({A}-{RuleSet}}. Calculate the occurrence probabilities of all X & Y together in
the transaction. Let, the highest probability value is P of item X & a particular Y (if Y is more than 1 item, consider any arbitrary Y).

Step 5: If P ≥ Minimum Support probability then select item Y & add item Y to RuleSet, delete item Y from set A & go to Step 4.

Step 6: If P< Minimum Support, then algorithm terminates & RuleSet items are returned and association rules are generated.

V. PROBABILITY BASED APRIORI ALGORITHM ILLUSTRATION & IMPLEMENTATION

Transaction	Item Present
T1	Milk, Bread, Butter ,Sugar
T2	Bread, Butter, Biscuit
T3	Butter, Sugar
T4	Bread, Butter, Sugar
T5	Bread, Sugar, Biscuit

Figure 6: Transaction database

Fig 6 shows the transaction database. It contains 5 different transactions and the items which have been purchased (same as Figure 1).

Minimum support threshold=2 (for this illustration).
Minimum support probability=Minimum support threshold/ number of transaction=2/5=0.4 (for this illustration)

Candidate 1 Itemset	Support Count
{Milk}	1 {Rejected}
{Bread}	4 {Selected}
{Butter}	4
{Sugar}	4
{Biscuit}	2

Figure 7: Finding out first item to be selected in Rule Set

Figure 7 shows the selection & inclusion of item "Bread" in the association RuleSet based on highest support count and rejection of item "Milk" from set A as support count is less than Minimum support threshold.
So, association RuleSet= {Bread} & set A= {Bread, Butter, Sugar, Biscuit}.

Rules (associating Selected Item Bread)	Probability of Occurrence
Bread → Butter	3/5=0.6≥Minimum Support(Selected)
Bread → Sugar	2/5=0.4
Bread → Biscuit	2/5=0.4

Figure 8: Finding out second item to be selected in RuleSet from the probability of occurrence values

Figure 8 shows the selection & inclusion of 2nd item "Butter" in the association RuleSet based on highest probability of occurrence (>=Minimum support probability)
So, current association RuleSet= {Bread, Butter}

Rules (associating Selected Item Butter)	Probability of Occurrence
Butter → Sugar	3/5=0.6 ≥Minimum Support (Selected)
Butter → Biscuit	1/5=0.2

Figure 9: Finding out third item to be selected in Rule Set from the probability of occurrence values

Figure 9 shows the selection & inclusion of 3rd item "Sugar" in the association rule set based on highest probability of occurrence (>=Minimum support probability)
So, association RuleSet= {Bread, Butter, Sugar}

Rules (associating Selected Item Butter)	Probability of Occurrence
Sugar → Biscuit	1/5=0.2 (Rejected)

Figure 10: Reaching termination of the processing

Figure 10 shows the termination of the processing, as rule's probability of occurrence is less than Minimum Support probability=0.4(for this illustration)

So, the final Association Rule Set after termination of the processing={Bread, Butter, Sugar} (for this illustration)

Association Rules	Confidence
R1: Bread^Butter->Sugar	Sup. Cnt.{Bread, Butter, Sugar}/Sup. Cnt.{ Bread, Butter}=2/3=67% SELECTED
R2: Bread^Sugar->Butter	Sup. Cnt.{Bread, Butter, Sugar}/Sup. Cnt.{ Bread, Sugar}=2/3=67% SELECTED
R3: Butter^Sugar->Bread	Sup. Cnt.{Bread, Butter, Sugar}/Sup. Cnt.{Butter, Sugar}=2/3=67% SELECTED
R4: Bread->Butter^Sugar	Sup. Cnt.{Bread, Butter, Sugar}/Sup. Cnt.{ Bread }=2/4=50%
R5: Butter-> Bread^Sugar	Sup. Cnt.{Bread, Butter, Sugar}/Sup. Cnt.{ Butter}=2/4=50%
R6: Sugar->Bread^Butter	Sup. Cnt.{Bread, Butter, Sugar}/Sup. Cnt.{ Sugar}=2/4=50%

Figure 11: Probability of occurrence based final selection and rejection of association rules

Figure 11 shows final selection and rejection of association rules containing different combination of items comprising final Association RuleSet based on minimum confidence threshold 60% (for this illustration).

So, in comparison to the conventional approach, the enhanced approach doesn't check the probability of occurrence of rules from any itemset whose size is greater than 2. While the traditional Apriori Algorithm is checking the support count of all possible candidate n-itemset based on frequent (n-1)-itemset (as shown in Figure 4, support count of candidate 3-itemsets are being checked), the modified probability based approach which have been applied for the same transactional database, is not at all checking the support count of those itemsets. But this approach still manages to give the same accurate outcomes of association rules. As a result, there will be fewer database scanning which improves the running time of the processing where we are dealing with huge amount of data.

VI. CONCLUSION

In the conventional Apriori algorithm one of the major inadequacies is that we have to access the database repetitively, which augment the running time of the algorithm. On account of this major drawback, we have developed the improved algorithm to minimize the database scanning while generating association rules, which gives us a good result. Always probability of occurrence of combination of any two items is checked among which one item have an explicit frequent occurrence, which reduces the time by a colossal factor. It increases the processing speed which is our primary target for developing a finer approach.

REFERENCES

[1] Han J, Kamber M, Pei J, Data Mining: Concepts and Techniques. Morgan Kaufmann Publisher.
[2] Jong S P, Ming S C, Philip S Y, An effective hash based algorithm for mining association rules, In Proceedings of the 2005 ACM SIGMOD International Conference On Management of Data.2005,24 (2): 175-186.
[3] Han, Pei, Y Yin and R Mao, Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery, 2004, 8:53-87.
[4] Kong Fang, Qian Xue-zhong, Research of improved apriori algorithm in mining association rules, Computer Engineering and Design, 2008, v29, n16, p4220- 4223.
[5] Li Qingzhong , Wang Haiyang, Yan Zhongmin, Efficient mining of association rules by reducing the number of passes over the database, Computer Science and Technology,2008,p 182-188.
[6] Nele Dexters, Paul W. Purdom, Dirk Van Gucht, A probability analysis for candidate-based frequent itemset algorithms, in Proceedings of the 2006 ACM Symposium on Applied Computing (SAC), Dijon, France, April 23-27, 2006.