Designing a Multi-Level Support Based Association Mining Algorithm

W. J. Samaraweera[*], S. Vasanthapriyan[*], Kavita S. Oza[**]

[*] Department of Computing and Information Systems, Sabaragamuwa University of Sri Lanka
[**] Department of Computer Sciences, Shivaji University, Kolhapur, India

Abstract- Finding of hidden and previously unknown information in large collection of data is the process of data mining. Mining association rules is a very important model in data mining. Using association rules different type of regularities and patterns can be identified. In most of the previous approaches a single minimum support threshold value is used for all the items or itemsets. But all the items in an itemset do not behave in the same way. Some appear very frequently and some very rarely. Therefore the support requirements should vary with different items. In this paper, a simple algorithm based on the Apriori approach is proposed to find the large-itemsets and association rules under arithmetic mean constraint and with multiple minimum supports to overcome the above mentioned problem. The proposed algorithm is easy and efficient and it saves time by focusing only on necessary associations.

Index Terms- apriori approach, association rule, data mining, mean constraint,multiple minimum supports.

I. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. [1]

Data mining can be recognized in many different terms such as; knowledge mining using data, knowledge extraction from data, pattern analysis, data dredging and data archaeology. But most frequently it is known as Knowledge Discovery from Data (KDD). Simply data mining is the process of knowledge discovery by analyzing the large volumes of data from various perspectives and then summarizing it into useful information so that people can use them in decision making. It has become an essential component in various fields of human life since it can be used to identify hidden patterns in a large data set. Data mining can be used on different kinds of data such as data warehouses, relational, transactional and advanced database systems, data streams, flat files, and the World Wide Web. The information and knowledge gained from data mining can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration. [3]

Data mining has several tasks such as association rule mining, classification and prediction, and clustering. [2] Association rule mining is the most common one among them [5] [6] [7] [8] [9] [10] [11] [14] [15] [17].And it has been used in many application domains. One of such is the business field where discovering relationships among items or sets of items in a transaction database in a supermarket or a shopping mall. When a set of transactions is given, association mining can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction. Technically association mining is used to discover elements that co-occur frequently within a dataset consisting of multiple independent selections of elements and to discover rules. An example for one of the applications in association mining is finding answers to questions such as "if a customer purchases product X, how likely is he/she to purchase product Y?" and "What products will a customer buy if he buys products A and B?"

Usually a supermarket or a shopping mall collects vast amount of data on sales, customer buying history, goods etc. Association mining help retailers in the supermarket/ shopping mall to identify the relationships among the goods purchased by customers in order to improve better customer satisfaction and retention.

Consider $I = \{I_1, I_2 \ldots I_m\}$, a set of items. Consider D as the task-relevant data, a set of database transactions where each transaction T is a set of items such that $T \subseteq I$. Consider X as a set of items. A transaction T is said to contain X if and only if $X \subseteq T$. An association rule is an implication of the form $X \rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \varphi$. Two measures, support and confidence, are evaluated to determine whether a rule should be kept. Let n be the number of transactions in T.

Support – The support of a rule, $X \rightarrow Y$, is the percentage of transactions in T that contains $X \cup Y$, and can be seen as an estimate of the probability, $P(X \cup Y)$. Support determines how frequent the rule is applicable in the transaction set T. The support of rule $X \rightarrow Y$ is computed as follows:

Support = count. $(X \cup Y)/n$

Confidence - The confidence of a rule, $X \rightarrow Y$, is the percentage of transactions in T that contain X also contain Y. It is the probability of conditional probability, $P(X \mid Y)$. The confidence of rule $X \rightarrow Y$ is computed as follows:

Confidence = count. $(X \cup Y)$/count.X

The support and the confidence of a rule must be larger than or equal to a user-specified minimum support and a minimum confidence respectively.

II. RESEARCH ELABORATIONS

Earlier work on association rules have showed minimum support (minsup) should be uniformly specified for all items in the database or for items within the same level. Therefore it

assumes that all items in the data set are of the same nature and have similar frequencies in the data.

But actually in real data different items may have different criteria to judge its importance. Since rare entities have useful knowledge patterns [13], it is not reasonable to have a single minimum support threshold value for all the items.

The support requirements should vary according to the item. For example, the minimum supports for cheaper items may be set higher than those for more expensive items. But if the minsup is set too high, rules that consider rare items will not be found. On the other hand if it is set very low, then those frequent items will be associated with one another in all possible ways and then there can happen a combinatorial explosion. That means it will provide too many rules. And many of them will be meaningless.

There are various algorithms that have been proposed to mine association rules.[14] [15] [17]AprioriAlgorithm uses single minimum support therefore it suffers from 'rare item problem'. [18]

Therefore, to extract frequent item sets involving rare items, an improved approach known as Multiple Support Apriori (MSApriori) has been proposed in [16].Multiple supports have been used in those algorithms. In MSApriori frequent item sets involving rare items are obtained by assigning minimum item support (MIS) value to each item. Then item sets has to satisfy the lowest MIS value among the respective items. The rules generated are then pruned based on confidence value.

A mining approach was proposed [12], which allowed the minimum support value of an itemset to be any function of the minimum support values of items contained in the itemset. This method is flexible when assigning the minimum supports to itemsets but the time complexity of this is very high because of its generality.

A solution for this was proposed [4], based on the Apriori approach to generate the large itemsets under the maximum constraints. Though this was a simple and efficient algorithm than with the minimum constraint, this algorithm generates very few association rules compared to the minimum constraint. There can be some important association rules which will be neglected by this approach.

Thus a new simple approach is proposed in this paper, which will not generate almost all the important association rules not creating a combinatorial explosion. It will consider about frequent items as well as rare items. It is basically based on Apriori Algorithm and it generates large itemsets under the arithmetic mean constraint.

III. The Proposed Mining Algorithm under the Arithmetic Mean Constraint

The multiple min-supports mining algorithm using mean constraints:

INPUT: A set of n transaction data T, a set of p items to be purchased, each item t_i with a minimum support value m_i, i = 1 to p, and a minimum confidence value (λ).

STEP 1: Calculate the count c_k of each item t_k, k=1 to p, as its occurrence number in the transactions; derive its support value s_{tk} as:

$s_{tk} = ck / n$ (1)

STEP 2: Check whether the support s_{tk} of each item t_k is larger than or equal to its predefined minimum support value m_{tk}. If t_k satisfies the above condition, put it in the set of large 1-itemsets (L1). That is:

$L1 = \{t_k | s_{tk} \geq m_{tk}, 1 \leq k \leq p\}$ (2)

STEP 3: Set r = 1, where r is used to keep the current number of items in an itemset.

STEP 4: Generate the candidate set C_{r+1} from L_r in a way similar to that in the Apriori algorithm [15] except that the supports of all the large r-itemsets comprising each candidate (r+1)-itemsetI_k must be larger than or equal to the arithmetic mean (denoted as mIk) of the minimum supports of items in these large r-itemsets.

STEP 5: Calculate the count c_{Ik} of each candidate (r+1)-itemsetI_k in C_{r+1}, as its occurrence number in the transactions; derive its support value s_{Ik} as:

$s_{Ik} = c_{Ik} / n$ (3)

STEP 6: Check whether the support s_{Ik} of each candidate (r +1)-itemsetI_k is larger than or equal to m_{Ik} (obtained in STEP 4). If I_k satisfies the above condition, put it in the set of large (r+1)-itemsets (L_{r+1}). That is:

$L_{r+1} = \{I_k | s_{Ik} \geq m_{Ik}, 1 \leq k \leq |C_{r+1}|\}$ (4)

STEP 7: IF $L_{r+1}$is null, do the next step; otherwise, set r = r+1 and repeat STEPs 4 to 7.

STEP 8: Construct the association rules for each large q-itemset$I_k$with items $\{I_{k1}, I_{k2},\ldots,I_{kq}\}$, $q \geq 2$, using the following sub steps:

(a) Form all possible association rules as follows:

$I_{k1} \ldots \Lambda\ I_{kj-1} \Lambda\ I_{kj+1} \Lambda \ldots \Lambda\ I_{kq} \rightarrow I_{kj}$, j=1 to q (5)

(b) Calculate the confidence values of all association rules using the formula:

$s_{Ik} / s_{Ik1} \ldots \Lambda\ I_{kj-1} \Lambda\ I_{kj+1} \Lambda \ldots \Lambda\ I_{kq}$ (6)

STEP 9: Output the rules with confidence values larger than or equal to the predefined confidence value λ.

OUTPUT: A set of association rules in the criterion of the arithmetic mean values of minimum supports.

IV. Results or Finding

A simple example is given to demonstrate the proposed algorithm and to show how the proposed algorithm can be used to generate association rules from a set of transactions with different minimum support values defined on different items.

Table 1 - Transaction Data

TID	Items
1	ABDG
2	BDE
3	ABCEF

4	BDEG
5	ABCEF
6	BEG
7	ACDE
8	BE
9	ABEF
10	ACDE

Table 2 - Pre-defined Minimum Support

Item	A	B	C	D	E	F	G
Min-Sup	0.4	0.7	0.3	0.7	0.6	0.2	0.4

STEP 1

Calculate the support value of each item.

Item	A	B	C	D	E	F	G
Support	0.6	0.8	0.4	0.5	0.9	0.3	0.3

STEP 2

Compare the support value of each item with the corresponding pre-defined minimum support values. Select the items which have higher support value than the corresponding minimum support values and put them in the large 1-itemsets.

L1 = {A, B, C, E, F}

STEP 3

Set r=1.

STEP 4

Generate the candidate set C2 from L1, and the supports of the two items in each itemset in C2 must be larger than or equal to the arithmetic mean of their predefined minimum support values.

Take the candidate 2-itemset {A, C} as an example. The supports of items A and C are 0.6 and 0.4 from STEP 1, and the arithmetic mean of their minimum support values is 0.35. Since both of the supports of these two items are larger than 0.35, the itemset {A, C} is put in the set of candidate 2-itemsets. All the candidate 2-itemsets generated in this way are found as:

C2 = {{A, B}, {A, C}, {A, E}, {A, F}, {B, E}, {C, F}}

STEP 5

Find the support of each candidate itemset in C2 using the transaction table.

2-itemset	{A, B}	{A, C}	{A, E}	{A, F}	{B, E}	{C, F}
Support	0.4	0.4	0.5	0.3	0.7	0.2

STEP 6

Compare the support value of each candidate 2-itemset with the arithmetic mean of the minimum support values of the items and put them in the set of large 2-itemsets L2.

L2 = {{A, C}, {A, E}, {A, F}, {B, E}}

STEP 7

Since L2 is not null, r is set at 2 and STEPs 4 to 7 are repeated. No candidate 3-itemset, C3, is generated and L3 is thus empty. The next step is then executed.

STEP 8

All association rules are formed for each large q-itemsets, $q \geq 2$.

1. A→C
2. C→A
3. A→E
4. E→A
5. A→F
6. F→A
7. B→E
8. E→B

STEP 9

Calculate the confidence values of the above association rules.

1. A→C = 0.67
2. C→A = 1
3. A→E = 0.83
4. E→A = 0.56
5. A→F = 0.5
6. F→A = 1
7. B→E = 0.875
8. E→B = 0.78

STEP 10

Compare the calculated confidence value for each of the rule with the pre-defined threshold confidence value. Select the rule which have a larger calculated confidence value.

1. C→A
2. F→A
3. B→E

V. Conclusion

In this paper, an efficient algorithm is developed for mining association rules with Multiple Minimum Supports based on Apriori Algorithm. A new approach has been introduced to find the large-itemsets and association rules under arithmetic mean constraint. The experiments have been done with other means such as geometric mean and harmonic mean also. But arithmetic mean gives the best result comparing to other two means. Arithmetic mean gives results covering both frequent items and rare items. And it will not give any unnecessary rules and create a combinatorial explosion.

The newly proposed algorithm is tested with real-life data sets using data from an outlet of Cargills (Ceylon) Food Company (Pvt.) Ltd. (http://www.cargillsceylon.com/Default.aspx) The data set has 38 items and 100 transactions. Each transaction has 2 – 5 items.

Eg:-

TID	Item
001	Cream Cracker, Cheese, Chocolate
002	Bread, Peanuts, Milk, Jam
003	Maggie, Cheese, Samahan, Cream Cracker
004	Jam, Soda, Potato Chips

It gave association rules regarding both frequent items and rare items comparing with using both minimum constraint and maximum constraint. And it does not give any unnecessary association rules also.
Eg:-

1. Bread → Butter
2. Maggie → Eggs
3. Cream Cracker → Cheese

The proposed mining algorithm using the arithmetic mean constraint finds less large itemsets and association rules than that using the minimum constraint. But it will eliminate the disadvantage of missing some rules when using maximum constraint. The proposed algorithm will focus more on rare items than using maximum constraint. It can find the large itemsets level by level without backtracking. Therefore it is more time-efficient than that with the minimum also.

REFERENCES

[1] KrishnaKumar, D.Amrita, N.SwathiPriya, Mining Association Rules between Sets of Items in Large Databases, International Journal of Science and Modern Engineering (IJISME), ISSN: 2319-6386,Volume-1, Issue-5, April 2013

[2] M. M. A. Baig, M.R. Pawar, S. F. Shazmeen, Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727Volume 10, Issue 6 (May. - Jun. 2013), PP 01-06

[3] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Second Edition, University of Illinois at Urbana-Champaign

[4] Y. C. Lee, T. P. Hong, W. Y. Lin, Mining Association Rules with Multiple Minimum Supports Using Maximum Constraints, Elsevier Science

[5] R. Agrawal, R. Srikant, Q, Vu, Mining association rules with item constraints,The Third International Conference on Knowledge Discovery in Databases and Data Mining, 1997, pp.67-73

[6] W.J. Frawley, G. Piatetsky-Shapiro,C.J. Matheus, Knowledge discovery indatabases: an overview, The AAAI Workshop on Knowledge Discovery inDatabases, 1991, pp.1-27

[7] T. Fukuda, Y. Morimoto, S. Morishita,T. Tokuyama, Mining optimizedassociation rules for numeric attributes, The ACM SIGACT-SIGMOD-SIGARTSymposium on Principles of Database Systems, 1996, pp.182-191

[8] H. Mannila, H. Toivonen, A.I. Verkamo, Efficient algorithm for discoveringassociation rules, The AAAI Workshop on Knowledge Discovery in Databases,1994, pp.181-192

[9] J.S. Park, M.S. Chen, P.S. Yu, Using a hash-based method with transactiontrimming for mining association rules, IEEE Transactions on Knowledge andData Engineering, 9(5)(1997) 812-825.

[10] R. Srikant, R. Agrawal, Mining generalized association rules, The Twenty-firstInternational Conference on Very Large Data Bases, Zurich, Switzerland, 1995,pp.407-419

[11] R. Srikant, R. Agrawal, Mining quantitative association rules in large relationaltables, ACM SIGMOD International Conference on Management of Data,Montreal, Canada, 1996, pp.1-12

[12] K. Wang, Y. H, J. Han, Mining frequent itemsets using support constraints, inProceedings of the 26th International Conference on Very Large Data Bases, 2000, pp. 43-52

[13] D. Rai, K. Verma, A.S. Thoke, MSApriori using Total Support Tree Data Structure,International Journal of Computer Applications (0975 – 8887) Volume 43– No.23, April 2012

[14] R.Agrawal, T.Imielinski, and A. Swami, "Mining association rules between sets of items in large databases." SIGMOD, 1993, pp. 207-216.

[15] R. Agrawaland R. Srikanth, "Fast algorithms for mining association rules." VLDB, 1994.

[16] Liu, W. Hsu, and Y. Ma, "Mining Association Rules with Multiple Minimum Supports." SIGKDD Explorations, 1999.

[17] J.Han, Y. Fu, "Discovery of multiple-level association rules from large database", in the twenty-first international conference on very large data bases, Zurich, Switzerland, 1995, pp. 420-431.W.-K. Chen, Linear Networks and Systems, Belmont, CA Wadsworth, 1993, pp. 123–135.

[18] Mannila, H. "Methods and Problems in Data Mining.",ICDT1997.

AUTHORS

First Author – W. J. Samaraweera Department of Computing and Information Systems, Sabaragamuwa University of Sri Lanka
Second Author – S. Vasanthapriyan, Department of Computing and Information Systems, Sabaragamuwa University of Sri Lanka
Third Author – Kavita S. Oza, Department of Computer Sciences, Shivaji University, Kolhapur, India