CHARACTER BASED WEIGHTED SUPPORT THRESHOLD ALGORITHM USING MULTI CRITERIA DECISION MAKING TECHNIQUE

Dr.T.Christopher,

Assistant Professor, Government Arts College (Autonomous),
KARUR - 639 005, Tamilnadu, INDIA.
chris.hodcs@gmail.com

Abstract - An association rule technique generally used to generate frequent itemsets from databases and generates association rules by considering each item in the datasets. However, the values of items are different in many aspects in a number of real applications, such as retail marketing, network log, etc. The difference between items makes a strong impact on the decision making in these applications. Therefore, traditional Association Rule Mining(ARM) cannot meet the demands arising from these applications. In this paper a new approach is introduced for computing profit weight of an item and generating frequent itemsets by minimum support threshold. The profit or the importance of the items in the itemsets is computed, based on the item subjective measures of characteristic through the proposed Global Profit Weight (GPW) algorithm using multi criteria decision making technique to improve the quality of output

Keywords: Association Rule mining, Profit Based Pattern, Subjective Measures, Global Support Weight

1. INTRODUCTION

Many previous works focused on the market basket (binary) association rules, which is in the form of "The transactions show that there are many customers who purchase product A will purchase the product B". Here all the items are treated uniformly. Association Rule Mining (ARM) identifies frequent itemsets from databases and generates association rules by considering each item in equal value. But, items are actually different in many aspects in a number of real applications, such as retail marketing, network, etc. The difference between items makes a strong impact on the decision making in these applications. Therefore, traditional ARM cannot meet the demands arising from these applications. By considering the different values of individual items as significance, profit/value based mining focuses on identifying the itemsets with high profit/ utilities. Unlike the support threshold defined by [1][2] in traditional association rule mining algorithm, profit based support measures can be applied to the non-binary numerical data associated with items in a transaction, allowing for a more insightful analysis of the impact of itemsets in terms of stock, cost or profit.

Discovery of efficient association rules has been found useful in many applications. However, without fully considering the importance and significance of items and transactions, it is noted that some rules which are discovered might have expired from users' point of view. Value based measures play an important role in data mining, regardless of the kind of patterns being mined. These measures are intended for selecting and ranking patterns according to their potential interest to the user.

In retail market analysis the product characteristics(set of criteria) such as damage of the product, offers provided by the product, the quality of the product or easiness to sell the product, brand or trade-mark of the product etc., may be interesting factors. The list of user interested factors is called profit support measure. The usefulness of the profit support measure improves the quality of the mining results. A user may not be interested in frequent itemsets alone that do not generate significant interestingness.

II. MULTI CRITERIA DECISION METHODS

Many multi criteria decision methods (MCDM) have been developed to assist data mining for analyzing and solving multiple criteria decision problems. The discipline is divided into two major sub-areas, namely, multi criteria decision mechanism and multi objective optimization. The primary goal of these methods is to find the best solution that is shaped by the preferences of the data mining for both quantitative and qualitative objectives.

The multi criteria decision methods[3][4][5] are developed to assist data mining , in either ranking a known set of alternatives for a problem or making a choice among this set while considering the conflicting criteria. The preferences of the data mining are elicited either before or during the evaluation of the alternatives and the criteria. The alternatives are compared with each other based on how they perform relative to each criterion. Similarly, some methods require comparison of the criteria to determine the relative importance of each criterion. Multi criteria decision methods will then utilize this information to assign ranks to the alternatives. The alternative with the highest rank is selected as the best compromise solution. The multi objective optimization methods are developed to generate a set of non-inferior solutions to problems that are modeled as optimization models consisting of multiple objectives. The preference information of the data mining is then incorporated to identify the best compromise solution from the non-inferior region of objective space.

Analytic Hierarchy Process (AHP) is one of multi criteria decision making [6][7] method to derive ratio scales from paired comparisons. The input can be obtained from actual measurement such as price, weight etc., or from subjective opinion such as satisfaction feelings and preference. The analytic hierarchy process enables decision makers to structure decisions hierarchically with the overall goal of the decision at the top of the model, strategic objectives in the higher levels, evaluation criteria in the middle levels, and alternative choices at the bottom. The AHP provides a structured framework for setting priorities on each level of the hierarchy using pair-wise comparisons, a process of evaluating each pair of decision factors [8][9] at a given level on the model for their relative importance with respect to their parent. The consistency of the judgments is tracked using the mathematical analysis behind the analytic hierarchy process to validate the decision process. In cases where inconsistency is above ten percent it is recommended that the criteria and judgments be revisited [10][11]. Decision makers are then able to create a model of their priorities where the weight of the decision is distributed from the goal downwards. If a user increases the weight of a criterion, the alternatives that performed well on that criterion will always get higher scores. This sensitivity analysis is portrayed extremely valuable for testing the impact of changing priorities on alternative business decision choices.

This process uses pair wise comparisons and then computes the weighting factors through evaluation of a set of criteria elements. The decision maker starts by laying out the overall hierarchy of the decision. This hierarchy reveals the factors to be considered as well as the various alternatives in the decision. Then, a number of pair-wise comparisons are done, which result in the determination of factor weights and factor evaluations. The process has been used to assist numerous corporate and government decision makers.

In this process the problems are decomposed into a hierarchy of criteria and alternatives. An important part of the process is accomplished by the three steps. The steps are stating the objective, defining criteria, and picking the alternatives. This information is then arranged in a hierarchical tree and synthesized to determine relative rankings of alternatives. Both qualitative and quantitative criteria can be compared using informed judgments to derive weights and priorities

For simplicity, considering that there are 4 Criteria: C_1, C_2, C_3, and C_4 they form a pair-wise comparison matrix A, where the number in the i^{th} row and j^{th} column gives the relative importance of C_i as compared with C_j. By using a 1–9 scale, with a_{ij} = 1 if the two criteria's are equal in importance, a_{ij} = 3 if C_i is weakly more important than C_j, a_{ij} = 5 if C_i is strongly more important than C_j, a_{ij} = 7 if C_i is very strongly more important than C_j, a_{ij} = 9 if C_i is absolutely more important than C_j, a_{ij} = 1/3 if C_j is weakly more important than C_i

The prioritization of items is measured using pair-wise comparison in 1-9 scale in the following method.

1- equally preferred
2- Equally to moderately preferred
3- Moderately preferred
4- Moderately to strongly preferred
5- Strongly preferred
6- Strongly to very strongly preferred
7- Very strongly preferred
8- Very strongly preferred to extremely strongly preferred
9- Extremely preferred

III .THEORETICAL MODEL Of GLOBAL PROFIT WEIGHT

In this section, the global profit weight and global profit support values are determined through an empirical illustration. Let I = {i_1, i_2, …….i_m} be a set of literals, called items. Let D = {T_1, T_2, T_3…T_n} be a set of n transactions, where for each transaction T ε D, T $\subseteq$I. A set of Items X $\subseteq$I is called an itemset. Transaction T contains X if X $\subseteq$T. Each itemset X is associated with a set of transactions T_x = {T ε D|T $\subseteq$X}, the transaction containing X.

Definition 1: The total profit of an item, denoted TP_i is dependent on the value of attribute associated with frequency of the product (Q_i) and the marginal profit of the product (UP_i) for each item in the database.

$$TP_i = Q_i * UP_i$$

……. (1)

Definition 2: The product weight of an item, denoted Pw_i calculates the ratio of total profit (TP_i) of each item and the sum of total profit of all items i.e

$$TP = \sum_{i=1}^{n} (Q_i \; * UP_i) = \sum_{i=1}^{n} TP_i$$

……. (2)

$$Pw_i = {TP_i}/{TP}$$

……. (3)

For example, if the product A has sold 4 units and the profit per unit is 6.50 then the total profit (TP_i) is TP_1 = 4 * 6.50 = 26. The total profit value of all items is represented by the sum of the total profit value, which is used to compute the product weight (Pw_i) of an item.

Definition 3: A criteria ratio of criteria is, denoted CR_i, and is the multiplication value of the preference value of each criterion. The preference values are evaluated through the pair wise comparison using 1 – 9 scaling technique. i = (1, 2, 3 … m), where m is number of criteria factor.

$$CR_i = \prod_{i=1, j=1}^{m} C_{ij}$$

……. (4)

Definition 4: Priority ratio of a criteria denoted ZC_i, is obtained from the normalized criteria ratio with respect to the objective. Let the Criteria C= {C_1, C_2, C_3 … C_m}, where "m" is number of criteria factor

The priority ratio of criteria is defined as

$$ZC_i = \frac{\left(\prod_{i=1, j=1}^{m} C_{ij} \right)^{1/m}}{\sum_{i=1}^{m} \left(\prod_{i=1, j=1}^{m} C_{ij} \right)^{1/m}} = \frac{(CR_i)^{1/m}}{\sum_{i=1}^{m} (CR_i)^{1/m}}$$

……. (5)

Definition 5: A product criteria ratio denoted PCR_i, is the value associated with the pair-wise comparison of the products using 1-9 scaling techniques. Let P is a product set, P= {P_1, P_2… P_n}. Where n is the number of products/items. P_{ij} is obtained from the pair-wise comparison between the products with respect to the criteria factor.
The product criteria ratio is defined as

$$PCR_i = \prod_{i=1, j=1}^{n} P_{ij}$$

……. (4.1)……. (6)

Definition 6: Priority ratio of product (level 2 hierarchy) with respect to each criteria is denoted ZR_i, is obtained from the normalized product criteria ratio. The ZP_i value indicates that the product weight with respect to the criteria. Where n is the number of products/items.

$$ZP_i = \frac{\left(\prod_{i=1, j=1}^{n} P_{ij} \right)^{1/n}}{\sum_{i=1}^{n} \left(\prod_{i=1, j=1}^{n} P_{ij} \right)^{1/n}} = \frac{(PCR_i)^{1/n}}{\sum_{i=1}^{n} (PCR_i)^{1/n}}$$

……. (7)

The same procedure is used to compute the priority ratio of the product with respect to all other criteria.

Let P is a product set, P= {P_1, P2, … P_n}. Suppose there are several comparison matrices at level 2. These comparison matrices are made for each choice, with respect to each factor. The ZP_i value indicates that the product weight with respect to the criteria of a product

Definition 7: Factor weight of an item is denoted as Fw_i, is computed using the priority ratio of objective (defined in definition (4)) and priority ratio of criteria (defined in definition (6)).

The value the priority ratio with respect to each criteria is assigned as each column of (m X m) matrix and the matrix is multiplied with priority ratio of level-1 hierarchy objective. Where m is the number of criteria and n is number of items. The factor weight is defined as,

$$Fw_i = \sum_{i=1,j=1}^{i=n,j=m} ZP_{ij} * ZC_{j1}$$

……. (8)

$Fw_1 = ZP_{11}*ZC_{11} + ZP_{12}*ZC_{21} + ZP_{13}*ZC_{31} \ldots\ldots ZP_{1m}*ZC_{m1}$
$Fw_2 = ZP_{21}*ZC_{11} + ZP_{22}*ZC_{21} + ZP_{23}*ZC_{31} \ldots\ldots ZP_{2m}*ZC_{m1}$
$Fw_3 = ZP_{31}*ZC_{11} + ZP_{32}*ZC_{21} + ZP_{33}*ZC_{31} \ldots\ldots ZP_{3m}*ZC_{m1}$

o o o o o o o o

o o o o o o o o

$Fw_n = ZP_{n1}*ZC_{11} + ZP_{n2}*ZC_{21} + ZP_{n3}*ZC_{31} \ldots\ldots ZP_{nm}*ZC_{m1}$

……. (9)

Definition 8: Global profit weight of an item is denoted as GPW_i, is the sum of the product weight (Pw_i) and factor weight (Fw_i) of each item. It is a unique weight of every product, which denotes the importance of product in the highest weight order. The global profit weight is defined as:

$$GPW_i = Fw_i + Pw_i$$

……. (9)

$$GPW_i = \sum_{i=1,j=1}^{i=n,j=m} ZP_{ij} * ZC_{j1} + \frac{TP_i}{TP}$$

……. (10)

Definition 9: Global Support Weight of an item is denoted as GSW_i, is derived from the frequency or occurrences of the items (F_i) in all the transaction in the database and global profit weight (GPW_i). For example, if the item B is occurred 13 times in the transaction and the global profit weight is 0.2585 then the global support weight of an item B is

$GSW_1 = 13 * 0.2585 = 3.3605$.

The global support weight is defined as:

$$GSW_i = F_i * GPW_i$$

……. (11)

By this mean, the global support weight of an item can be defined as the product of global profit weight of the item and the frequency of occurrences of the item in the transaction.

An item is called profitable, if the global support weight of an item is greater or equal to the minimum support threshold. An association rule X → Y is called an interesting pattern if X U Y is a large itemsets and the pattern is greater than or equal to global profit weight threshold.

Proposed Global Profit Weight (GPW) algorithm

The algorithm for global profit weight has the following inputs and outputs.

Inputs: A database D with the transaction T, Product (P), Quantity (Q) of item in each transaction and unit price (UP) of the item.

Outputs: A set of global profit weight of each item.

```
Procedure GlobalProfitWeight (P, Q, UP)
        Begin       GTP ← ∅;
                         For each Pi ≤ n do
                TPi ← Qi * UPI ;
                 GTP ← GTP + TPi
        End
        For each Pi ≤ n do
                PWi ← TPi / GTPi ;
        End
                For each  i ≤ n do
        ZCi ← ((∏CRi)^1/m) / (∑(∏CRi)^1/m) ;
                End
        For each i ≤ m do
    ZPi ← ((∏PCRi)^1/n) / (∑(∏PCRi)^1/n) ;
        End
        For each i ≤ n
            FWi ← ∅;
            For each x ≤ r
                For each y ≤ k
        FWix ← FWix  + (ZPxy * ZCyi) ;
                End
            End
        End
                For each Pi ≤ n do
                        GPWi ← FWi + PWi ;
                End
        End
```

IV. FREQUENT ITEMSETS WITH USING SUPPORT THRESHOLD

The different kinds of data like scientific data, medical data and marketing data are very important in the real time applications. Finding frequent itemsets is computationally the most expensive step in association rule discovery and therefore it has attracted significant research attention [12][13]. In this context, the pruning or filtering the unwanted or items which are not interest to the user's point of the view, in terms of profit or value in the transaction, is very important process to find frequent itemsets. The pruning is done based on the support measures. The items which are greater than or equal to the given minimum support threshold is defined as frequent items. Optimized measures can be used to prune uninteresting patterns during the data mining process to narrow the search space and thus improve the mining efficiency. For example, in a certain application, user could be more interested in the rules that contain "sales" rather than "customer". So, subjective interestingness measures can play an important role in knowledge discovery.

V. RESULTS AND DISCUSSIONS

The GPW algorithm for sort out the highly interested items has been implemented and the results have been compared with the computation method of support value for binary attributes, quantitative attributes, weighted support for binary attributes and weighted support of quantitative attributes. The results were significant in terms of finding the high priority, value based items. The GPW algorithm has been designed and implemented using JAVA language under windows operating system. The Pentium IV 3.06GHz PC system used to conduct the experiments.

The performance of the proposed model of value based association rule mining was ascertained based on the support value or the importance of each item in the database. The results of the experiment, that is comparison of support of binary and quantitative attribute, weighted support of binary and quantitative attribute and proposed global support weight are discussed.

Support value comparisons and result discussions

The performance of the proposed model of computing the value based weight was ascertained based the criteria, frequency, number of item sold, and marginal profit of each item in the database. The results of the experiments, which are comparison of support of binary and quantitative attribute, weighted support of binary and quantitative attribute and global support weight are given in the Table 1.

Support value (Binary Attribute)

The traditional support value (Binary attribute) computations are considered only the items occurred in the transaction. That is the support value of a pattern depends on the number of occurrences of the item in the market basket not the quantity purchased. For example the 1-itemset {C} has 50% support if the sale of item C occurred in 10 of the 20 transactions, the item {J} has 65% support if the sale of item J occurred in 13 of the 20 transactions.

Support value (Quantitative Attribute)

Simply choosing the occurrences of itemset does not reflect the impact of any factor except the frequency of the items. Therefore, numerical value can provide with an item of each transaction. Each value in the transaction dataset indicates the quantity sold of an item. The 1-itemset {J} has a higher support than the 1-itemset {C} if it occurs in three more transactions than C. Suppose, the total quantity of item C sold is 58, while that of item J is 17, so in fact C is sold more frequently. While considering the quantity of product sold, the product C has more support than the product J.

Weighted Support (Binary and Quantitative Attribute)

The weight as proposed in [14] its itemset profit w, where $0 \leq w \leq 1$, indicates the importance of the itemset. Any value can be assigned, within the bound 0 to 1, to any item. For example, if the weight of the itemset X is 0.95, it tells an itemset is important in the set of transactions. The weight of 0.1 indicates a less important set. Both the occurrences of an item and the important ratio (weights) factors are to be considered to compute the weighted support of the rule. Weights can provide the users with a convenient way to indicate the importance of the attributes. The weighted support of a binary attribute considered only the presence of the item (frequency) in the transactions whereas the quantitative attribute consider the total quantity sold in the transactions.

The weighted support of an item X is

$$wSup(X) = \left(\sum_{i_J \in (X)} w_j \right) \left(Sup(X) \right) \quad \text{…….. (12)}$$

Where the weights of the items $\{i_1, i_2, i_3, \ldots\ldots i_n\}$ are $\{w_1, w_2, \ldots\ldots w_n\}$ respectively.

Table 1 Product details

Products	No. of Occurrences	No. of Item Sold (Quantitative)	Weight of Item	Unit Profit
A	915	9512	0.65	6.5
B	950	9632	1	12
C	756	6448	0.4	7
D	1108	9608	0.2	5
E	488	808	0.6	2
F	910	1480	0.35	15.5
G	686	792	0.8	13.5
H	560	1002	5	25
I	474	782	0.7	5
J	914	2112	0.3	15

Comparisons of Global Support Weight

The Table 1 describes the items or products present in the real life database of D2K.T5.M10. It consists of the name of the products, number of occurrences of the products in the database, total number of quantity of a products sold and the weight of each item. And the columns of the Table 2 shows the support value of binary and quantitative attribute, weighted support of binary and quantitative attribute and global support weight. The importance or priority/rank of a product based on the support value is given in the square bracket.

Table 2 Comparisons of the Global Support Weight threshold with the existing Support threshold

Products	Binary Attribute %	Quantitative Attribute	Weighted Support (Binary)	Weighted Support Quantitative)	Global Support Weight
A	40 [5]	25 [4]	0.26 [5]	0.81 [4]	1.1296 [7]
B	35 [7]	38 [3]	0.35 [2]	1.90 [2]	1.4756 [6]
C	50 [2]	58 [1]	0.20 [7]	1.16 [3]	1.6850 [3]
D	30 [9]	49 [2]	0.06 [10]	0.49 [5]	0.9192 [9]
E	50 [2]	11 [7]	0.30 [4]	0.33 [8]	2.5330 [2]
F	25 [10]	05 [10]	0.09 [9]	0.09 [10]	0.8885 [10]
G	40 [5]	09 [9]	0.32 [3]	0.36 [6]	3.7208 [1]
H	45 [4]	14 [6]	2.25 [1]	3.50 [1]	1.5399 [5]
I	35 [7]	10 [8]	0.25 [6]	0.35 [7]	0.9422 [8]
J	65 [1]	17 [5]	0.195 [8]	0.255 [9]	1.6211 [4]

It is observed from the above table, that the global support value of the product {G} has highest priority (Profit/Weight) based on subjective measures, over other existing support method. Even though the product G has assigned more weight in the weighted support method, it does not have the highest priority value because number of item sold is less. But in the retail market some items are very expensive, consequently they are not purchased so often, but the profit gained those expensive items are as important as other frequently bought items to the retailer. As shown in the Table 2, the product {G} has highest Profit, and can believe that this product must have first priority than other products.

It is also noticed that, depending on the subjective measures assigned to the item criteria, the product {E} has the second highest profitable global support value in GPW algorithm. In practice, the quantity sold alone may not express the semantics of applications, because the user's interest may be related to other factors. That is, the sales person may not interested the items which are sold more with less profit. This product has occurred 488 times in the transaction and the total number of products sold is 808. It is clearly understood that this product is more profitable in both support (binary attribute) value and Global Support weight techniques.

The Table 2 also indicates that the product {F} has lowest priority in almost all the support computation method, because is has less importance in all aspects.

The product {D} has the 9th priority/rank value in the GPW method, whereas this item has higher preference in the support (quantitative value) method. According to the profit based subjective measures, and the direct profit (15.5) gained form the product, the sales person may not focus this product because of less profit, even though the product sold very often.

It is concluded that the Table 2 shows, the best improvements occur in the Global Support Weight for selection of profitable\Quality items. The items H and B have almost high weighted support value in binary and quantitative weighted support method with respect to the direct profit of the items but the global support value of these items are varies considerably. Depending on the subjective measure assigned to the item criteria, it is possible that the GPW method might produce the valuable and quality items as shown in the above Table 2. The results were significant in terms of finding the value based itemset. So it is found that the GPW performance is better than other existing methods.

The proposed method produces highly profitable frequent items based on the criteria and also gives better accuracy/quality in terms of profit than other existing methods.

VI. CONCLUSION

In this paper character based weighted support threshold algorithm known as Global Profit Weight algorithm and a mathematical model to find the global profit weight for prioritize the item in the transaction which is based on the product characteristics using multi criteria decision mechanism is devised. The method of computation of the global profit weight is discussed in detail using retail market basket analysis as a real life application. The performance of the global support weight method has been compared with other existing methods.

REFERENCES

[1]. Agrawal.R and Srikant.R "Fast algorithms for mining association rules". In Proceedings the International Conference of Very Large Data bases, Santiago, Chile, September 1994, pages 487-499.

[2]. Aggarwal.C, "Towards long pattern generation in dense databases" In Proceeding of the ACM SIGKDD Explorations, January 2001, Vol. 3(1) pages 20-26.

[3]. Koch and Tom, "A Pilot Study on transplant Eligibility Criteria", Pediatric Nursing, Mar.-Apr. 1997, Vol. 23(2), pages160-62.

[4]. Hemaida.R.S. and Kalb.E, "Using the analytic hierarchy process for the selection of first-year family practice residents", Research and Perspectives on Healthcare, 2001, Vol. 79(1), pages 11-15.

[5]. Lenca.P, MeyerP, Vaillant.B, and Lallice.S, "A multi criteria decision aid for interestingness measure selection", Technical Report LUSSI-TR-2004-01-EN, GET/ENST, Bretagne, France, May 2004, pages 1-24.

[6]. Kang, Moonsig, "PAHAP: A Pairwise Aggregated Hierarchical Analysis of Ration-Scale Preferences", Decision Sciences, July 1994, pages 607-24.

[7]. Ishizaka.A and Lusti.M., "An Expert Module to Improve the Consistency of an AHP Matrix", The 16th Triennial Conference of the International Federation of Operational Research Societies, Edinburgh, 2002, pages 112-119.

[8]. Hahn.E.D, "Better decisions come from a results-based approach", Marketing News, Academic Research Library, September 2004, Vol. 36(19), pages 24-25

[9]. Zhang.H, Padmanabhan.B and Tuzhilin.A, "On the discovery of significant statistical quantitative rules", In Proceeding of the 10th International Conference on Knowledge Discovery and Data Mining, Seattle, USA, August 2004, pages 374-383.

[10]. Wang.K, Zhou.S and He.Y, "Growing decision trees on support-less association rules", In Proceedings of International Conference on Knowledge Discovery and Data Mining, August 2000, pages 265-269.

[11]. Goh and Chon, "AHP for Robot Selection", Journal of Manufacturing Systems, January 1997, pages 381-87.

[12]. Zheng.Z, Kohavi.R and Mason.L, "Real world performance of association rule algorithms", In Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining, ACM SIGKDD'01, San Francisco, California, August 2001, pages 401-406.

[13]. Zaki.M.J, "Scalable algorithms for association mining", In IEEE Transaction of Knowledge and Data Engineering, 2000, Vol. 12(2), pages 372–390.

[14]. Yang.J, Wang.W and Yu.P, "Efficient Mining of Weighted Association Rules (WAR)", In Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data mining, 2000, pages 270-274