Indian Journal of Science and Technology, *Vol 8(8), 707–714, April 2015*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Provide a New Approach for Mining Fuzzy Association Rules using Apriori Algorithm

Heydar Jafarzadeh[1], Rouhollah Rahmati Torkashvand[2], Chamran Asgari[2*] and Amir Amiry[3]

[1]Department of Computer Engineering, Science and Research Branch,
Islamic Azad University, Mehran, Iran; Heydar.Jafarzadeh@gmail.com
[2]Department of Computer Engineering, Payame Noor University, Iran;
rahmati.r42@gmail.com, asgari.chamran@gmail.com
[3]Department of Computer Engineering, Islamic Azad University, Malayer, Iran; aamiry57@yahoo.com

Abstract

Association rules mining is one of the most popular data mining models. Minimum-support is used in association rules mining algorithms, like Apriori, FP-Growth, Eclat and etc. One problem Apriori algorithm and other algorithms in the field association rules mining, this is user must determine the threshold minimum-support. Suppose that the user wants to apply Apriori algorithm on a database with millions of transactions, definitely user can not have the necessary knowledge about all the transactions in the database, and therefore would not be able to determine an appropriate threshold. In this paper, using averaging techniques, we propose a method in which Apriori algorithm would specify the minimum support in a fully automated manner. Our goal in this paper improved algorithm Apriori, to achieve it, initially will try to use fuzzy logic to distribute data in different clusters, and then we try to introduce the user the most appropriate threshold automatically. The results show that this approach causes the any rule which can be interesting will not be lost and also any rule that is useless cannot be extracted. The simulation results on a real example show that our approach works better than the classic algorithms.

Keywords: Apriori Algorithm, Association Rules, Data Mining, Fuzzy Logic & C-Means Clustering, Frequent Patterns, Support

1. Introduction

Data mining is a logical process used for finding relevant data in large data sets[1]. For have this information, we try to find frequent patterns in the given data set. Patterns that are interesting and certain enough according to the user's measures are called knowledge[1]. The output of a program that discovers such useful patterns is called discovered knowledge. Data mining is a sub process of Knowledge Discovery in Databases in which the different available data sources are analyzed using various data mining algorithms[2]. Given there might be interesting associations among the data, we need automated and efficient tools to find and organize these associations. Therefore, many of the data mining tools are presented with various analysis techniques[3]. Association Rules Mining is one of the most important tasks used in data mining, which can be applied in different domains[4] in the field of data mining and association rules, many researchers have tried to obtain more favorable rules, to achieve this goal, they have used different strategies[5]. In the field of frequent pattern mining and association rules mining, the main three approaches have been developed that are Apriori approach and FP-Growth approach and Eclat algorithm[6]. We believe that the field of mining frequent patterns and association rules mining is still a research area has raised interest among researchers because the researchers are working to provide effective and efficient methods.

**Author for correspondence*

1.1 Problem Statement

A problem of classical association rules is that not every kind of data can be used for mining. Rules can only be derived from data containing binary data, where an item either exists in a transaction or it does not exist[7]. These types of calculations are in the crisp sets category, where expression is strictly one item is available in a transaction or not. But in front of crisp sets, fuzzy sets are given, Where different degrees of membership is defined, So that the degree of membership of a particular item can be part of various transactions. This idea cover crisp sets weaknesses[8]. There are also other challenges related to the Apriori algorithm and other algorithms in the field of Association rules mining are that these algorithms are based on the assumption that users can specify the threshold: minimum-support. Since, there is, in fact, no appropriate index based on which we can define an appropriate threshold as the standard for the measurement of the minimum-support It is impossible that users give a suitable minimum-support for a database to be mined if the users are without knowledge concerning the database. Our aim in this paper is that transactions that are stored in the database as the crisp enter to a fuzzy environment of the so-called fuzzification. The paper is organized as follows: section 2 contains the general model for association rules and an overview of related approaches to the discovery of association rules. We present an efficient fuzzy association rules mining algorithm based on our proposed approach in section 3, then in section 4 we implement a case study based on the proposed approach. Evaluate the effectiveness of the proposed approach experimentally perform in Section 5. Finally, we Discussion Conclusions and recommendations in section 6 and 7.

1.2 Research Background

Transactional database contains a lot of items, so there must be effective tools be able to review this item. Therefore, many data mining tools have been developed that allow a great variety of analysis techniques, mostly derived from classical statistics[9]. Since its introduction in[10], the technique of association rules mining has received great interest by the data mining community and a lot of research has been done resulting in the development of many different algorithms. Association rules is especially useful for conducting a market basket analysis where transaction data can be analyzed. Regularities in data of a supermarket for example can be found in this way. An association rules could be "If a customer buys bread and milk, he will mostly buy butter as well". This information is very useful for business because promotion actions can be designed accordingly[11].

2. Principles of Association Rules

Mining association rules from the database principles that all researchers who are active in this area, according to these principles, try to improve the algorithms. The following will describe the principles.

2.1 Association Rules

Mining association rules is one of the most important tasks in data mining area[2]. A simple association rule can be represented as: Bread→Cheese [support=0.1, confidence=0.8]. Simply put, this rule states that there is an association between buying bread and cheese with the support major indicating that bread and cheese come together in 10% of the transactions and the confidence major stating that cheese has taken part in transactions where bread is also present. In this case, 80% of the transactions included cheese with bread also present in those transactions. With this rule, we can assume that in the future, those who buy bread are most likely to also buy cheese during those transactions. Such information can help the retailers explore opportunities for cross-selling. We represent the matter of association rules as follows: $I=\{i_1,i_2,\ldots,i_m\}$ is aitemset and $T=\{t_1,t_2,\ldots,t_n\}$ is a set of transactions, each of which contains items from I. Therefore, transaction t_i contains a set of items where $t_i \subseteq I$. Association rules is a concept in the form of $X \rightarrow Y$ where $X \subset I$ and $Y \subset I$ as well as $X \cap Y = \varphi$ where X and Y are called itemsets.

2.2 Frequent Itemset

An itemset whose support is greater than or equal to a minimum-support threshold is known as frequent itemsets. In many situations, we only care about association rules involving sets of items that appear frequently in baskets. Then find the association rules only involving a high-support set of items i.e., $\{X_1,X_2,\ldots,X_n,Y\}$ must appear in at least a certain percent of the baskets, called the support threshold[12]. Support and confidence are the two most important quality measures for evaluating the interestingness of a rule.

Support: Support is measured by:

$$\sup\left(X \to Y\right) = \frac{X \cup Y}{N} \qquad (1)$$

Where, *A* and *B* are different items in the database and *X* is the total of the items in the database. This rule mines all the transactions where *A* and *B* are present and then compares it with the minimum-support specified by the user. And only chooses and lists those transactions in which support is equal to or bigger than the minimum-support and the rest are eliminated as uninteresting rules[6].

Confidence: Then the confidence is applied to the new list as follows:

$$conf\left(X \Rightarrow Y\right) = P\left(Y \middle| X\right) = \frac{\sup\left(X \to Y\right)}{\sup\left(X\right)} \qquad (2)$$

This is a very important measure to determine whether a rule is interesting or not. It looks at all transactions which contain a certain item or itemset defined by the antecedent of the rule. Then, it computes the percentage of the transactions also including all the items contained in the consequent[6].

3. The Main Algorithms in the Field of Mining Association Rules

Different ways by researchers to discover frequent patterns and generate the association rules from data stored in the data basis presented[6]. But the fact is that among these algorithms only three approaches are that with different ways to generate association rules, these include: (1) Apriori approach, (2) FP-Growth approach and (3) Eclat approach. Other algorithms and approaches in this area only have tried that improve the performance of the three algorithms.

3.1 Apriori Algorithm

Apriori algorithmin 1994 by Aggarwal and Srikant first algorithm to generate association rules was proposed. The main advantage of this algorithm simply uses it to generate association rules. Apriori algorithm works in two stages to produce association rules.

- Frequent Pattern Mining

Most complex process in production association rule is discovery process frequent patterns, because doing so requires repeated visits of the database. Due to the features of Apriori a pattern is frequent at level that which has been frequent in the level k-1 and it support is greater than or equal to the threshold[13]. Apriori algorithm for finding frequent itemsets the beginning generated of candidate itemset.

- Generating Association Rules

After complex process of finding frequent patterns, Begins the process generated association rules. Association rules can be generated only from the of frequent itemset that to do this used from measure confidence and the rule will be accepted as an association rule that their confidence greater than or equal to the threshold[13].

Apriori algorithm to generate association rules faced with two main challenges: (1) Repeated visits from database that will be slowing speed of algorithm, (2) Generate candidate items that creates a lot of over head and increases the complexity of the space.

3.2 FP-Growth Algorithm

FP-Growth algorithm was presented in 2000 by Hanet al. this algorithm over come the two main challenges Apriori algorithm. These algorithm visits only two-time the database and to produce a frequent itemset does not generate any candidate itemset. The FP-Growth algorithms use a depth-first search method and on the first visit the database compressed and saved all the data in the form of a tree structure called FP-Tree[14]. This algorithm to accelerate the process of discovering frequent patterns using of dividing large items into smaller units. FP-Growth algorithm performs the process of discovering frequent patterns in two stages.

- Pre-processing Data

On the first visit the database all items that are extracted, first compared with a threshold and items that can not satisfy this threshold are deleted the very beginning[8]. Used from remaining itemset to compression in to the FP-Tree and form the nodes of this tree. For each node in FP-Tree three values are stored they are: (1) name, (2) Support and (3) Link to connect the current node with neighboring nodes.

- FP-Tree

FP-Tree will be created in a structure descending such that items with maximum support make above and root node in the tree and Items with the least support make its leaves. FP-Growth algorithm with recursive traversal of the tree discovering the frequent patterns. This process is repeated until no new frequent item to be produced. The FP-Growth algorithm is two defects: First, the implementation and application of this approach is more difficult than the Apriori algorithm and second, these algorithms are not flexible to change the items in the database and if you add a new data into the database the FP-Tree must be built and loaded again[15].

3.3 Eclat Algorithm

There are two data formats for data storage: (1) horizontal Data format (TID, Items) and (2) Vertical data format (Items, TID). Both algorithms presented above use of the horizontal format while the Eclat algorithm use of the vertical format[16]. Eclat algorithm performs discovery of frequent patterns so that to generate the candidate itemset first visit the database and Items can be extracted with a list of the number of transactions in which they are present. This algorithm visit the databases only once, this work covers the overhead caused by the production of all subsets of a transaction and check them again to calculate the support[17].

4. The Proposed Approach

By examining the three algorithms presented above and all the algorithms have been trying to improve these algorithms, this result is obtained that this approach assumes that the user have to determine the threshold and when the algorithm runs, from user will be asked that determine the threshold. The question that comes to mind here is that how a user can have the required knowledge of all transactions in the database to be able to use that knowledge determines the threshold. Since transactional databases containing millions of items sufficient knowledge of these items is difficult or impossible for a user, therefore a user can not definitely determine an appropriate threshold.

In this article we will attempt to the transfer of crisp data to a fuzzy environment[19] and feeding the Apriori algorithm with fuzzy data improve the flexibility of this algorithm and provide a mechanism that algorithm to determine the threshold automatically. Our goal is to become the Apriori algorithm to a fully automated algorithm that there is no dependence on the user. We try also extracting rules, stronger and more interesting than the algorithms presented above.

4.1 Fuzzification and Mining Fuzzy Association Rules

As explained above, we want to put the fuzzy data as input to the Apriori algorithm. To transfer data to fuzzy environment in this paper from clustering approach is used that from proposed algorithms in the field of clustering we chose the popular clustering algorithm Fuzzy C_means[20,21]. The algorithm whit gets the crisp data sets and the number of clusters of user distributes the data in between clusters.

Fuzzy C-Means algorithm output is used as input for Apriori algorithm. Our approach to determine fuzzy support and confidence works as follows:

- Calculate Support in a Fuzzy Environment

As you can see in formula (3) on the first visit to the database to produce the candidate 1_items, calculate the total degree of membership for an item and consider the result as a fuzzy support[18].

$$fuzzysum = \sum_{i=1}^{n} \mu(x) \tag{3}$$

After generating the candidate 1_items to produce candidate 2_items that for calculate the fuzzy support 2_items the degree of membership of both the item to compare and choose between them minimum. Finally, total minimum obtained consider as the fuzzy support of two items. The process to generate a set of candidate 3_items and above also takes a similar form (See formula (4)). This process is repeated until no new candidate itemset to be generated.

$$fuzzy\sup(A \to B) = \sum_{i=1}^{n} \min(f_A(x), f_B(x)) \tag{4}$$

- Calculate Confidence in a Fuzzy Environment

As mentioned above Apriori algorithm produces the two-step association rules which is the second and final stage generated association rules. To get past this stage we must calculate the fuzzy confidence for frequent patterns obtained from the previous step. Formula (5) shows how to calculate the fuzzy confidence. It is clear that for the calculation fuzzy confidences of two items first calculate the fuzzy support of two items and then divided the total the minimum of the first item. The output of this relationship is compared with a threshold and rules that are greater than or equal to the threshold accepted as association rules[18].

$$fuzzyconf(A \Rightarrow B) = \frac{\sum[fuzzy\sup(A \to B)]}{\sum(\min(A))} \tag{5}$$

4.2 Calculating the Fuzzy Minimum-Support using Proposed Approach

We mean for the algorithm to automatically introduce an appropriate minimum-support to the user. When experts want to compare one set of data with an index and such index does not exist to be compared to and to classify the data bigger (better, stronger…) or smaller (worse, weaker…) than this index, they resort to statistical techniques and try to define an index based on formulas such as averaging, standard deviation, median, mode, variance, etc. so they can compare their data to this index and extract the information needed.

Since, there is, in fact, no appropriate index based on which we can define an appropriate threshold as the standard for the measurement of the minimum-support, we will also use statistical techniques. Of the techniques present in this field, we've chosen averaging technique to be able to define a suitable minimum-support to the extraction of frequent patterns is a way that this minimum-support is neither so low that it generates a large body of irrelevant patterns nor so high that we lose interesting patterns. After the use of fuzzy c-means clustering approach, crisp data were transferred into a fuzzy environment, we use the formula (6) to get minimum-support threshold and then to introduce users.

$$fuzzy\,\min\sup\left(A,B\right)=\frac{\sum\left[sum\left(f_x\left(A\right),f_x\left(B\right)\right)\right]}{\left|T\right|} \quad (6)$$

5. The Simulation Results of our Approach on a Real Example

In this section we describe all stages of the mining association rules using the proposed approach for automatically determining the minimum support and use Fuzzy C-Means clustering algorithm. To do this we use a database of online food stores, table 1 shows the number of items purchased from the online shop.

MATLAB provides the users the possibilities that do fuzzification using Fuzzy C-Means clustering algorithm and obtain statistical information from the data which are suitable for the analysis of data. Table 2 shows some of the statistical information of the data used in this paper. In this study, we define three clusters for each item that each item in this cluster is distributed Based on the amount buy.

Table 1. The number of products purchased by customers

TID	Drink	Meal	Canned Fish	Canned Beans	Eggs	...
1	21	33	4	13	23	...
2	2	21	34	10	55	...
3	13	11	8	0	63	...
4	5	6	19	14	36	...
5	56	83	32	34	27	...
6	11	23	43	17	54	...
7	9	17	65	0	3	...
8	55	34	15	11	48	...
...	...	...	...	...	...	...

Table 2. Statistical information for sale of products

Eggs	Canned Beans	Canned Fish	Meal	Drink	
0	2	0	2	0	**Minimum**
7	6	9	9	9	**Maximum**
0.6	2.44	0.4	3.4	1.2	**Center1**
1	4	3.6	7.6	4.9	**Center2**
6.4	5.5	9	8.5	8.6	**Center3**

Graphical representation of the membership functions helps users to analyze data. Figure 1 shows the membership functions obtained in this paper.

Given that the in this paper three clusters are defined for each item, table 3 shows sample of the data and clusters.

Min-sup obtained by the our approach was equal to 30%, the itemset are compared with 30% and support items that are greater than or equal 30% known as the frequent patterns and then generate the association rules. Table 4 shows the characterization.

The ultimate output of simulations of the our approach is showed in Table 5 (Min-sup= 30%, Min-conf= 50%) Show understand able user of the output is as follows:

If Meal = Low Then Drink = Low, Confidence=60%.

If Drink = Low and Drink=Medium Then Meal = High, Confidence= 97%.

If Drink = Medium and Meal = High and Drink = Low, Confidence= 99%.

6. Discussion

For this study we used a same database for the three algorithms Apriori, FP-Growth, Eclat and our approach. Simulation results show that Performance of our approach to find rules is better and more accurate compared with three other algorithms. Figure 2 shows the comparison.

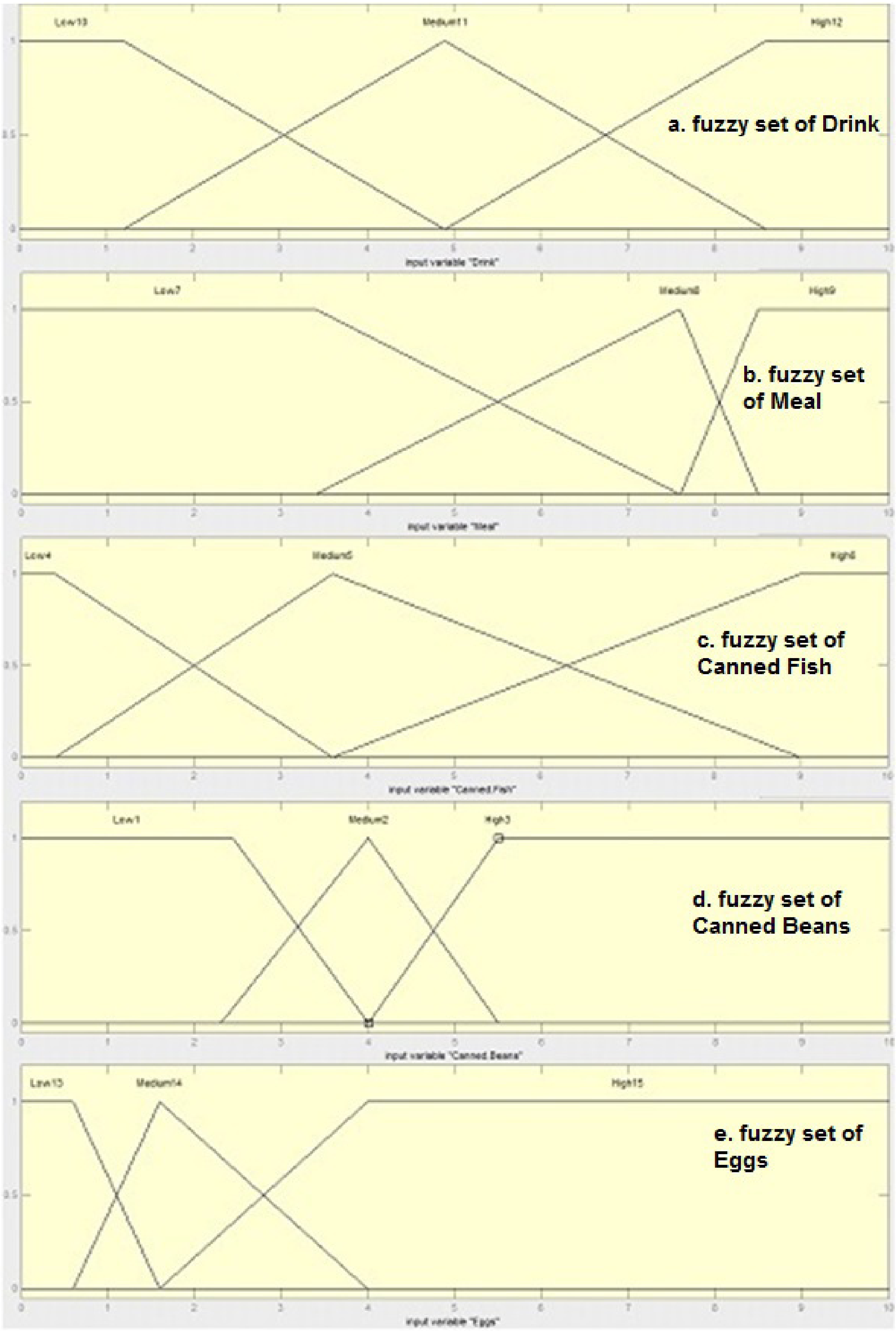

Figure 1. Membership functions obtained by Fuzzy C-Means Algorithm.

Table 3. The output of the FCM

Canned Fish.Medium	Canned Fish.Low	Meal.High	Meal.Medium	Meal.Low	Drink.High	Drink.Medium	Drink.Low
0.75	0.21	0.09	0.91	0	0.022	0.16	0.840
0.023	0.34	0.02	0.79	0.19	0.115	0.878	0.118
027	0.54	0.08	0.02	0.90	0.07	0.80	0.13
0	1	0.79	0	0.21	0.147	0.4	0.453
.	0.9	1	0	0	060	0.25	0.15
0.18	0.66	0.87	0.12	0.1	0.18	0.11	071
0.5	0.5	0.12	0.08	0.8	0.06	0.17	0.77
0.90	0.088	0.04	0.11	0.85	0.30	0.065	0.635

Table 4. Generate association rules

Confidence= Sup (Antecedent, Consequent)/ Sup (Antecedent)	Sup (Antecedent)	Sup (Antecedent, Consequent)	Antecedent→Consequent
23%	90%	21%	Drink.Low→Drink.Medium
15%	67%	10%	Meal.Low→Drink.Low
17%	66%	11%	Meal.Low→Drink.Medium
64%	80%	51%	Meal.Medium→Meal.High
79%	50%	40%	Drink.Low,Deink.Medium→Meal.High
51%	79%	40%	Drink.Low,Meal.High→Drink.Medium
98%	40%	39%	Drink.Low,Meal.High→Meal.Meaium
78%	50%	39%	Meal.Medium,Meal.High→Drink.Low

Table 5. The output of the our approach

Conf (Antecedent → Consequent)	Sup (Antecedent)	Sup (Antecedent, Consequent)	AR
60%	40%	23%	Meal.Low→Drink.Low
49%	51%	27%	Meal.Medium→Drink.Low
48%	40%	19%	Meal.Low→Drink.Medium
60%	39%	22%	Drink.High→Meal.Low
49%	40%	22%	Meal.Low→Drink.High
89%	48%	43%	Drink.High→Meal.High
48%	81%	43%	Meal.High→Drink.High
97%	88%	85%	Drink.Low,Drink.Medium→Meal.High
99%	86%	85%	Drink.Medium,Meal.High→Drink.Low

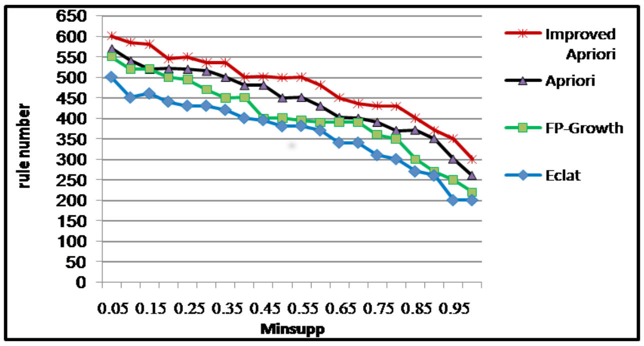

Figure 2. Performance of our approach compared to algorithms Apriori, FP-Growth and Eclat

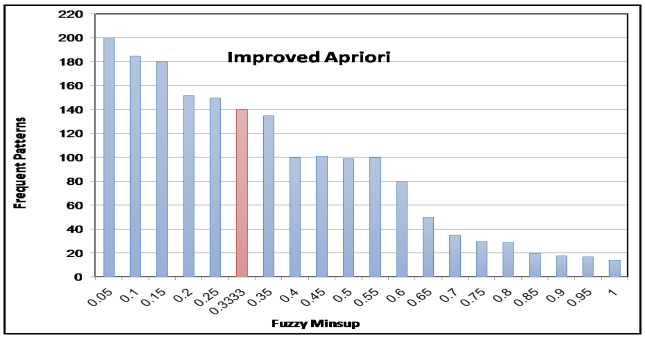

Figure 3. Comparison of determining the automatically minimum support and minimum-support determined by the user.

Automatically determining the threshold for the minimum-support by the algorithm will be generating all frequent itemsets the appropriate. Figure 3 shows a determined of automated support threshold and other supports specified by the user.

7. Conclusions

In this paper we deal with the important issue of automated determining the threshold for the minimum-support by the algorithm itself, that this was ignored by all the algorithms presented in the field of association rules. In this paper we used a database similar to the comparison approach proposed and the three basic algorithms Apriori, FP-Growth and Eclat. Comparative analysis of the results showed our approach has better performance than other algorithms.

8. References

1. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Magazine. 1996; 17(3):37–54.
2. Kabir S, Ripon S, Rahman M, Rahman T. Knowledge-based data mining using semantic web. International Conference on Applied Computing, Computer Science and Computer Engineering. 2014; 7:113–9.
3. Hong TP, Tung Y-F, Wang SH-L, Wu YL, Wu M-T. A multi-level ant-colony mining algorithm for membership functions. Inform Sci. 2012; 182(1):3–14.
4. Pears R, Koh SY, Dobbie G, Yeap W. Weighted association rule mining via a graph based connectivity model. Inform Sci. 2013; 218(1):61–84.
5. Yun U, Lee G, Ryu K. Mining maximal frequent patterns by considering weight conditions over data streams. Knowl Base Syst. 2014; 55:49–65.
6. Han J, Cheng H, Xin D, Yan X. Frequent pattern mining: current status and future directions. Data Min Knowl Discov. 2007; 15:55–88.
7. Kamsu-Foguem B, Rigal F, Mauget F. Mining association rules for the quality improvement of the production process. Expert Syst Appl. 2013; 40(4):1034–45.
8. Kaya M, Alhajj R. Genetic algorithm based framework for mining fuzzy association rules. Fuzzy Set Syst. 2005; 152(3):587–601.
9. Kardan A, Ebrahimi M. A novel approach to hybrid recommendation systems based on association rules mining for content recommendation in asynchronous discussion groups. Inform Sci. 2013; 219(10):93–110.
10. Imielienskin T, Swami A, Agrawal R. Mining association rules between set of items in large databases. Management of Data; 1993.
11. Petelin B, Kononenko I, Malacic V, Kukar M. Multi-level association rules and directed graphs for spatial data analysis. Expert Syst Appl. 2013; 40(12):4957–70.
12. Yoon Y, Lee G. Two scalable algorithms for associative text classification. Inform Process Manag. 2013; 49 (2):484–96.
13. Mannila H, Srikant R, Toivonen H, Inkeri A, Agrawal R. Fast discovery of association rules. Advances in Knowledge Discovery and Data Mining; 1996. p. 307–28.
14. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. Proceeding of the ACM-SIGMOD International Conference on Management of Data; 2000. p. 1–12.
15. Gruca A. Improvement of FP-growth algorithm for mining description-oriented rules. Advances in Intelligent Systems and Computing. 2014; 242:183–92.
16. Thabtah F-A, Cowling P-I. A greedy classification algorithm based on association rule. Applied Soft Computing. 2007; 7(3):1102–11.
17. Zaki M-J, Scalable algorithms for association mining. IEEE Transactions on Knowledge and Data Engineering; 2002. p. 372–90.
18. Lee Y-Ch, Hong T-P, Lin W-Y. Mining Fuzzy association rules with multiple minimum supports using maximum constraints. Knowledge-Based Intelligent Information and Engineering Systems. 2004; 3214:1283–90.
19. Zadeh L-A. Some reflections on the anniversary of fuzzy sets and systems. Fuzzy Set Syst. 1999; 100(1):1–3.
20. Bezdek J-C, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm. Comput Geosci. 1984; 10(2–3):191–203.
21. Shihab AI, Burger P. The analysis of cardiac velocity MR images using fuzzy clustering. Proceeding of SPIE Medical Imaging Physiology and Function from Multidimensional Images. 1998; 3:176–83.