Computing and Informatics, Vol. 31, 2012, 331–347

EFFICIENT MINING OF FUZZY ASSOCIATION RULES FROM THE PRE-PROCESSED DATASET

Zahra Farzanyar, Mohammadreza Kangavari

Deprtment of Computer Engineering, Iran University Science & Technology (IUST)
Tehran, Iran
e-mail: {Z_farzanyar, kangavari}@iust.ac.ir

Communicated by Huajun Chen

Abstract. Association rule mining is an active data mining research area. Recent years have witnessed many efforts on discovering fuzzy associations. The key strength of fuzzy association rule mining is its completeness. This strength, however, comes with a major drawback to handle large datasets. It often produces a huge number of candidate itemsets. The huge number of candidate itemsets makes it ineffective for a data mining system to analyze them. In the end, it produces a huge number of fuzzy associations. This is particularly true for datasets whose attributes are highly correlated. The huge number of fuzzy associations makes it very difficult for a human user to analyze them. Existing research has shown that most of the discovered rules are actually redundant or insignificant. In this paper, we propose a novel technique to overcome these problems; we are preprocessing the data tuples by focusing on similar behaviour attributes and ontology. Finally, the efficiency and advantages of this algorithm have been proved by experimental results.

Keywords: Knowledge discovery, data mining, fuzzy association rule, linguistic terms, domain ontology

1 INTRODUCTION

Data Mining, also referred as Knowledge Discovery in Databases (KDD), is a process of finding new, interesting, previously unknown, potentially useful, and ultimately understandable patterns from very large volumes of data [1, 2]. The regularities or exceptions discovered from databases through data mining have enabled human decision makers to better make decisions in many different areas [1, 2]. One important topic in data mining research is concerned with the discovery of interesting association rules.

Associations allow capturing all possible rules that explain the presence of some items according to the presence of other items in the same transaction. An association rule is an implication of the form $X \Rightarrow Y$, where X and Y are sets of attributes. Association rules can be rated by a number of quality measures, among which support and confidence stand out as the two essential ones [3]. The basic problem of mining association rules is then to generate all association rules $X \Rightarrow Y$ that have support and confidence greater than user-specified thresholds.

The problem of mining association rules was first introduced by Agrawal et al. [4], for databases consisting of only categorical attributes. In most real life applications, databases contain many other attribute values besides 0 and 1 such as cardinal or ordinal attributes. Unfortunately, the definition of categorical association rules does not translate directly to the case of quantitative attributes. It is therefore necessary to provide a definition of association rules for the case of a database containing quantitative attributes. To handle databases with both categorical and quantitative attributes, a quantitative association rule mining method was proposed by Srikant and Agrawal [5]. The method finds association rules by partitioning the quantitative attribute domain and then transforming the problem into binary one. Apparently, whatever partitioning methods are applied, "sharp boundaries" remain a problem, which may lead to an inaccurate representation of semantics. As a remedy to the sharp boundary problem, the fuzzy set concept has recently been used more frequently in mining quantitative association rules.

The fuzzy set theory introduced by Zadeh [6] is better than the interval method because fuzzy sets provide a smooth transition between member and non-member of a set. In these methods, each of quantitative attributes is replaced by a few other attributes that partition the range of the original one using the fuzzy theory. A column belongs to each of these partitions in the table containing transactions. So, the size of each transaction increases which leads to rising of their scan time. In frequent itemsets generation step, the number of candidate itemsets is several times higher than when all the exiting attributes are binary forms. It finally causes that definite time to compute the frequency of the candidate itemsets is longer. In spite of these problems, large databases make computation of knowledge discovery more and more expensive.

The other problem emerging here is the huge number of final produced rules. The huge number of associations makes it very difficult, if not impossible, for a human user to analyze them in order to identify those interesting/useful ones. The questions are: (1) "Can we enhance fuzzy association rule mining and preserve the full power of fuzzy association rule mining (i.e., its completeness) without overwhelming the user?" (2) "If yes, how can this be done?"

To overcome these problems, in this paper we are preprocessing the data tuples by focusing on similar behaviour attributes and ontology. Ontology represents knowledge with the relationships between the generalization and the specialization of concepts; therefore, it provides an alternative knowledge source than domain experts. It is organized as a DAG (Directed Acyclic Graph) hierarchy. There are many growing, large scale and shared ontologies which have been developed and utilized in various ways for helping the automation of data mining. If using similar behaving attributes and domain knowledge, we can reduce the search space, so knowledge discovery can be improved effectively. The other difference of this algorithm efficiency compared to other algorithms is in that the set of produced rules is typically very small.

This paper is organized as follows. In Section 2, the related works on the mining fuzzy association rules, studies on the interestingness and clustering and using taxonomic hierarchies for the mining association rules are outlined. Section 3 investigates the definitions and methodology of mining fuzzy association rules and introduces a formula to testing strong dependence between attributes and reviews some definitions related to ontology and description logics. So, it describes a new algorithm of mining fuzzy association rules along with an example to apply the proposed method. Section 4 shows the experimental results. The last section concludes the paper.

2 RELATED WORKS

In recent years, some work has been done on the use of fuzzy sets in discovering association rules for quantitative attributes. Miller and Yang [9] applied Birch clustering to identify intervals and proposed a distance-based association rules mining process, which improves the semantics of the intervals. Hirota and Pedrycz [28] proposed a context sensitive fuzzy clustering method based on fuzzy C-means to construct rule based models. To solve the qualitative knowledge discovery problem, Au and Chan [10] applied fuzzy linguistic terms to relational databases with numerical and categorical attributes. Later, they proposed the F-APACS method [11] to discover fuzzy association rules. Yager [24] introduced fuzzy linguistic summaries on different attributes. Some work has been done automatically determining the number and intervals of the clusters. Fu et al. [26] proposed an automated method to find fuzzy sets for the mining of fuzzy association rules. Their method is based on CLARANS clustering algorithm [27]. M. Kaya [30] presented a novel automated clustering method based on multi-objective GA [29]. Hong et al. [12] proposed definitions for the support and confidence of fuzzy membership grades and designed an algorithm to find interesting fuzzy association rules. Ishibuchi et al. [21] and E. Hullermeier [7] illustrated fuzzy versions of confidence and support. Gyenesei [15, 22] presented two different methods for mining fuzzy quantitative association rules, namely *without normalization* and *with normalization*. The experiments of Gyenesei showed that the numbers of large itemsets and interesting rules found by the fuzzy method are larger than the discrete method defined by Srikant and Agrawal [5]. Hu and Chen [18], by the algorithm named FGBRMA, proposed to generate fuzzy association rules from a relational database. Au and Chan [19] developed a fuzzy technique, called FARM II. FARM II is able to handle both relational and transactional data. Chen [13] proposed the fuzziness based fuzzy taxonomies. Some other researchers investigated the mining of weighted association rules. Some approaches by Cai, Fu et al. [14], Gyenesei [15], Shu, Tsang et al. [16], Lee [17] etc. have already been proposed, which are basically similar. Ya et al. [20] introduced an algorithm for mining fuzzy association rules by removing redundant fuzzy association rules.

In all the previous algorithms for fuzzy association rules mining, the objective is to find frequent fuzzy itemsets by expanding techniques presented for the binary form while the problems existing in the fuzzy form (introduced in Section 1) still remain.

3 FUZZY ASSOCIATION RULES AND PROPOSED ALGORITHM

Mining fuzzy association rules is the discovery of association rules using fuzzy set concepts such that the quantitative attributes can be handled. In this paper, we view each attribute as a linguistic variable, and the variables divided into various linguistic terms.

A linguistic variable is a variable whose terms are linguistic words or sentences in a natural language [31]. Triangular membership functions are used for each linguistic term defined in each quantitative attribute for simplicity. One way of determining membership functions of these linguistic terms is by expert opinion or by people's perception. Yet another way is by statistical methods [32]. Fuzzy clustering based on self-organized learning can also be used to generate membership functions [33].

In this study, we use the method presented in [18]. We claim higher degrees of correlation among attribute pairs, because of the existence of semantic links among them. Hence, in order to increase effectiveness and efficiency, we propose to directly incorporate knowledge about the existence of such links into the assessment process. Categorizing variables needs the existence of domain knowledge, such as the domain experts. However, domain experts are not always available during a data mining task. Domain ontology is another resource of domain knowledge which could specify semantic types for concepts. Therefore, we are making use of a large pre-existing concept hierarchy, which contains concepts from the data tuples.

3.1 Problem Statements

Let $I = \{I_1, I_2, \ldots, I_n\}$ be the items set where each I_j $(1 \leq j \leq n)$ is an attribute of the original dataset D. Each attribute I_j may have a binary, categorical or quantitative underlying domain Δ_j. Besides, each quantitative attribute I_j is associated with at least two fuzzy sets. If each quantitative attribute I_j is extended by its fuzzy set, we can get the extended attribute set I_f from I. Using the corresponding membership functions defined with each fuzzy set, the original dataset D is changed into a fuzzy dataset D_f. Given fuzzy dataset $D_f = \{t_1, t_2, \ldots, t_N\}$ with I_f, the discovered rules are of the form $A \Rightarrow B$, where t_i $(a \leq i \leq N)$ is a transaction in D_f, $A \subseteq I_f$, $B \subseteq I_f$, $A \bigcap B = \emptyset$ and $A \bigcup B$ do not contain any two items that are associated with the same attribute (for instance, will not contain "income low and income high"). Like Boolean association rules mining, A is called the antecedent of the rule and B is called the consequent of the rule. The standard approach to evaluate the significance of fuzzy association rules is to extend the definition of well-know support and confidence measures to fuzzy association rules [7]. The degree of support of the rule $A \Rightarrow B$ for the whole D_f is defined as:

$$D\,\mathrm{conf}(A \Rightarrow B) = \frac{\sum_{i=1}^{n} A(x) \otimes B(y)}{|D_f|} \tag{1}$$

and the degree of confidence is defined as:

$$D\,\mathrm{conf}(A \Rightarrow B) = \frac{\sum_{i=1}^{n} A(x) \otimes B(y)}{A(x)} \tag{2}$$

where $|D_f|$ is the total number of transactions in D_f, which is equal to N, the number of transactions in the quantitative database D. $A(x)$ and $B(y)$ denote the degree of membership of the elements x and y with respect to the fuzzy set A and B, respectively, $\otimes$ is a t-norm [8]. Based upon the notations of $D\,\mathrm{supp}$ and $D\,\mathrm{conf}$, a rule $A \Rightarrow B$ is interesting fuzzy association rule if

1. $D\,\mathrm{supp}(A \Rightarrow B) \geq \mathrm{Min\_supp}$;
2. $D\,\mathrm{conf}(A \Rightarrow B) \geq \mathrm{Min\_conf}$;

where Min_supp and Min_conf are the thresholds defined by users. Straightforward mining algorithm could be obtained easily. It is basically composed of two phases: generating all frequent itemsets from fuzzy dataset based on Apriori algorithm, then generating all rules from frequent itemsets and calculating confidence for each other. This straightforward algorithm is easy to understand and implement, but the disadvantages are also obvious (cf. Section 1).

The proposed way to build up this algorithm decreases the size of average transactions, data set and the number of candidate itemsets, leading to a shorter running-time along with improved performance. It also reduces the number of fuzzy association rules. In order to achieve this, we have preprocessed the dataset in the two phases. In the first phase, we fuse the similar behaviour attributes. For this purpose, we need a measure of dependency between two items. A widely used one has been introduced in [25]; it is chi-square and is based on support difference. In the second phase, we apply the ontology.

3.2 Chi-Square Test for Independence and Correlation

Chi-square test statistics (χ^2) is a widely used method for testing independence and/or correlation [25]. In our proposed technique, it is used for testing similar behaving between attributes. Essentially, χ^2 test is based on the comparison of observed frequencies with the corresponding expected frequencies. In other words,

χ^2 is used to test the significance of the deviation from the expected values. Let f_n be an observed frequency, and f be an expected frequency. The χ^2 value is defined as:

$$\chi^2 = \sum \frac{(f_0 - f)^2}{f} \tag{3}$$

χ^2 value of 0 implies the attributes are statistically independent. If it is higher than a certain threshold value, we reject the independence assumption.

3.3 Ontology

Nowadays, one of the most important and challenging problems in data mining is the definition of the prior knowledge; this can originate from the process or from the domain. This contextual information may help select the appropriate information, decrease the space of hypothesis, represent the output in a most comprehensible way and improve the process. Ontological foundation is a precondition for efficient automated usage of such information. Ontologies provide the formal framework for expressing knowledge about a domain and comprising semantic links among domain individuals that are described at the conceptual level as inter-concept relations or roles [19]. Therefore, it presents a better knowledge resource than domain experts. There are many large scale and shared ontologies which have been expanded and utilized in various manners for helping the automation of data mining. We can perceive the relation between ontologies and data mining in two manners:

From ontologies to data mining, we are incorporating knowledge in the process through ontologies use, i.e. how the experts comprehend and carry out the analysis tasks. Representative applications are intelligent assistants for discovering process, interpretation and validation of mined knowledge, ontologies for resource and service description and knowledge grids.

From data mining to ontologies, we are using the ontologies to represent the results of data mining. Therefore the analysis is done over these ontologies.

Our proposed algorithm precedes the first manner.

According to Corcho et al. [36] "a domain ontology can be extracted from special purpose encyclopedias, dictionaries, nomenclatures, taxonomies, handbooks, scientific special languages (say, chemical formulas), specialized KBs, and from experts." Ontology building (or ontology engineering) is a subfield of knowledge engineering that studies the methodologies for building ontologies. It studies the ontology development process for a special domain, the ontology life cycle, the methods and methodologies for building ontologies, and the tool suites and languages that support them [34, 35].

There are several mature methodologies that have been proposed to construct this process and thus to advance it. In computer science and artificial intelligence, ontology languages are formal languages used to construct ontologies. They do the encoding of knowledge about particular domains and often contain reasoning rules that support the processing of that knowledge. Ontology languages are usually declarative languages, and are commonly based on either first-order logic or on description logic. Domain ontology building for handling our problem is out of the scope of this study. We assume that a domain ontology matched with the dataset is built by ontology engineers.

3.4 Proposed Algorithm in Detail

In all the previous algorithms, the objective is to find frequent fuzzy itemsets by expanding techniques presented for the binary form while the problems existing in the fuzzy form remain. In our proposed algorithm, such procedure is not considered but the fuzzy association rule mining is taken under a novel look. In this method, two preprocessing are done on dataset. At first, the similar behaving attributes are found and fused. Such attributes have very close and sometimes equal degree of membership in the fuzzy transactions. Finding two similar behaviour attributes in one set and separate considering of each is time-consuming and leads to making many similar rules. In our proposed algorithm, we have used Chi-square statistic test to study the similar behaviour of attributes. We assume that quantitative attributes have their fuzzy linguistic terms.

At first we construct contingency tables for attributes existing in the I-set two by two. The rows and the columns of these tables are the fuzzy partitions that belong to each of such attributes. We obtain observed frequencies for all of the tables at the same time by scanning once the database, then for each of these tables, we obtain corresponding expected frequencies in case of accepting the independence assumption, and finally we calculate the χ value. The χ value obtained from the chi-square test for each of the contingency tables is compared to the chi-square table and the similar behaving attributes are determined.

For each of the similar behaving attributes, we consider two partitions which have the most frequency comparing to each other in the contingency table, we candidate them to be fused. After obtaining all the similar behaviour attributes, considering the common points between them we deduce the final similar behaving attributes. Then, we fuse the final similar behaving attributes on the primary dataset. We do so by multiplying degrees of membership of partitions related to similar behaving attributes candidate for fusing. Then we put the result in a new column and eliminate the columns related to these partitions. Therefore the size of each transaction is reduced.

Nevertheless, when we fuse similar behaving attributes of each transaction, the chances of finding similar transactions increases. During the writing, a tag is placed in front of every transaction to specify how many times that transaction exists in the dataset. While inserting the new transaction, the algorithm checks whether that transaction is already in the memory. If it is, it increases that transaction's counter by one. Otherwise, it inserts the transaction with count equal to one into the main memory.

So this technique reduces the average transaction length and also the dataset size significantly, so we can accumulate more transactions in the main memory.

These two reductions in transactions size and dataset size lead to reduction of scan time, which is very effective in increasing the performance of the proposed algorithm. The next phase of this algorithm concerns finding out the frequent itemsets in dataset. For this purpose, the frequency times of candidate itemsets in the new fuzzy dataset should be calculated. For generating candidate itemsets, this algorithm has used domain knowledge to guide the discovery process such that we can discover interesting rules.

The input domain ontology contains taxonomic relationships related to every concept, and the semantic relations among them. The other input is the dataset possessing transactions. Firstly, we classify the present attributes in dataset using higher-level concepts in relevant ontology and every class is given the same title as the relevant concept.

The semantic relations among these concepts in the ontology are also considered as the meta rules. These meta rules determine semantic relations among classes. For generating frequent itemsets, we only study the relations among the items related to a concept or the items related to concepts having semantic relations in the ontology. A great deal of candidate itemsets contain items related separate concepts, having no semantic relations in the ontology, so we omit these candidate itemsets as infrequent ones without spending time for counting their numbers in dataset.

In our algorithm the number of candidate fuzzy itemsets to be generated is much less than in the previous algorithms and it causes the considerable reduction of the time spent on computing the frequency of candidate itemsets, and therefore it leads to improving the performance of proposed algorithm.

The third phase of the algorithm is “making fuzzy association rules from obtained frequent itemsets”. In this phase, the produced rules inherit the semantic relations known as meta rules. Therefore, some of the produced rules also indicate the semantic relations that cause them to be more comprehensible. The following shows the details of the proposed algorithm.

Procedure Mining-Algorithm

Input A set of N transactions, each with n attribute, fuzzy linguistic terms for quantitative attributes, The user-specified minimum fuzzy support, The user-specified minimum fuzzy confidence, A domain Ontology.

Output

Phase I: Fuse similar behaving attributes;

Phase II: Generate Meta rules;

Phase III: Generate frequent fuzzy itemsets;

Phase IV: Make fuzzy association rules.

Method

```
Phase I:
    Step 1. Scan the database, and construct the contingency tables for each of two attributes.
    1-1. Compute the $\chi^2$ value.
    1-2. Compare obtained $\chi^2$ with chi-square table, if the obtained $\chi^2$ causes to reject independence assumption, announce these two attributes are similar behaving.
    Step 2. Deduction of final similar behaving attributes.
    Step 3. Fusing the similar behaving attributes in each transaction and forming a new column.
Phase II: Generate Meta rules;
    Step 1. Classify the attributes in database, using the existing concepts in ontology.
    Step 2. Entitle the obtained classes according to concepts.
    Step 3. Make the Meta rules according to the semantic relations among the obtained concepts of the ontology.
Phase III: Generate frequent fuzzy itemsets;
    Step 1. Find the frequent fuzzy 1-itemsets L1.
    Step 2. Find the frequent fuzzy 2-itemsets L2.
    2-1. Produce the candidate 2-itemsets C2 from 1- itemsets L1 in case that items of every itemsets belong to a concept or the concepts possessing semantic relations in the ontology.
    2-2. Compute the support degree for each itemset in 2-itemsets by Equation (1).
    2-3. Delete the 2-itemsets of which support is less than the user-specified minimum fuzzy support.
    2-4. Insert the remaining 2-itemsets into the frequent itemsets L2.
    Do
    Step 3. Find the frequent fuzzy $k$-itemsets L$k$.
    3-1. Produce the candidate $k$-itemsets C$k$ from frequent $(k-1)$ itemsets L$k-1$.
    3-2. Compute the support degree for each itemset in $k$-itemsets by Equation (1).
    3-3. Delete the $k$-itemsets of which support is less than the user-specified minimum fuzzy support.
    3-4. Insert the remaining $k$-itemsets into the frequent itemsets L$k$.
    Until no more frequent fuzzy itemset is produced.
Phase IV: Make fuzzy association rules with the found frequent itemsets.
    Step 1. Compute the confidence degree for each rule by Equation (2).
```

Step 2. Remove the rules with the confidence degree less than the user-specified minimum fuzzy confidence.

Step 3. Output the remaining fuzzy association rules with their support and confidence degrees and write main rules and Meta rules related to each of produced rules.

After finishing the algorithm, we would have all the fuzzy association rules resulting from this new fuzzy dataset. The rules obtained from the proposed algorithm are an exact and semantic summary of all the rules made in previous algorithms. We have used main rules for them to maintain exactness of the obtained rules. The main rules show the behavior of similar behaving attributes. In making all the rules we have considered the main rules as the basic ones, and also, we have made the other rules upon them so that they would not break the meta rules. The support degree of these rules is obtained from the related frequencies of contingency tables.

The user can now obtain a complete picture of the domain without being overwhelmed by a huge number of rules. From this summarized information, user can then find some interesting aspects to focus on. So complexity is reduced in phase III of the algorithm.

3.5 Complexity Analysis

Given a database containing N transactions such that each transaction is characterized by n attributes and each attribute is represented by m linguistic terms. The number of contingency tables for chi-square test should be calculated from the following formula:

$$c_n = \sum_{x=1}^{n-1} x = \frac{n(n-1)}{2}. \tag{4}$$

In each of the contingency tables complexity is $O(m^2N)$, so the complexity of the testing step of the attributes similar behaving is calculated as follows:

$$(m^2N)\left(\frac{n(n-1)}{2}\right) = (m^2N)(n^2) = O(m^2n^2N) \tag{5}$$

This complexity is desirable indeed.

3.6 An Example

This section shows an example to apply the second preprocessing. A part of the ontology is shown in Figure 1. For example; “person”, “paper”, “committee” and “conference” are higher-level concepts in the following ontology. These concepts have semantic relations together in the ontology considered as the meta rules. For example:

$$\text{Person} \xrightarrow{write} \text{Paper}$$

$$\text{Person} \xrightarrow{registers} \text{Conference}$$

We selected 10 000 transaction records as our testing synthetic dataset. The present attributes in related dataset concern to these higher-level concepts. For instance "age", "education", "skill", etc. concern to "person". Therefore, we put these attributes into "person" class. We classify the other attributes in the dataset using higher-level concepts in relevant ontology and every class has the same title as the relevant concept. The semantic relations among these classes in the ontology are considered as the meta rules.

The secondary phase of this algorithm concerns finding out the frequent itemsets in the dataset. For this purpose, the frequency times of candidate itemsets in the dataset should be calculated. We only study the candidate itemsets including items related to a concept, for instance the candidate itemsets including "age", "education" or "skill" that concern to "person" or the candidate itemsets including items related to concepts having semantic relations in the ontology.

In the following ontology, "Person" has semantic relations with "paper" and "conference" in the ontology, but it has no semantic relation with "committee". Then we only study the candidate itemsets including the items related to "person" and "paper" and also the candidate itemsets including the items related to "person" and "conference". So we omit candidate itemsets including items related to "person" and "committee" as infrequent ones without spending time for counting their numbers in dataset.

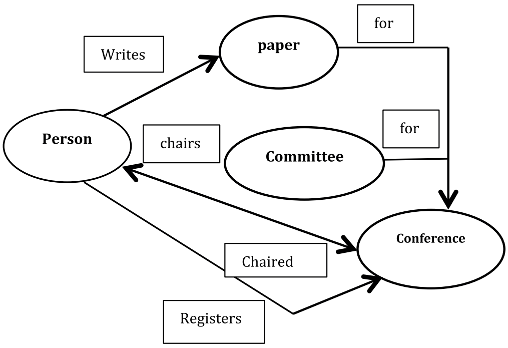


Fig. 1. A part of ontology

4 EXPERIMENTAL RESULTS

In order to evaluate the effectiveness of our new algorithm, we compare its performance with that of the other method. The difference between our new algorithm and the old one is that the latter one does not preprocess the data tuples by focusing on similar behaving attributes and ontology when generating fuzzy frequent itemsets.

We produced a synthetic dataset that described the use of domain ontology to evaluate our approach. It is also experimented on a real data set. The clinical dataset is from Australia and New Zealand Dialysis and Transplant Registry (ANZDATA) [37] which gathers 217 083 records of 19 220 kidney disease patients spanning 12.6 years. The ANZDATA is gathered by report form for every dialysis and transplant patient at yearly periods or real time entry for special incidents via the internet. The medical ontology base which is employed in this study is Unified Medical Language System (UMLS) [38]. It is a controlled compendium of many vocabularies which also provides a mapping structure between them. We applied the above algorithm as well as the old one to these datasets.

These two algorithms were implemented on PC with CPU 1 700-BANIAS, RAM 512 M, Windows XPPRO-TRAD, and Borland C++ 5.0.

At first, we show the effect of similar behaving attributes in decrease of the number of rules and of the execution time of the algorithm. The results of this experiment are shown in the following figures. In Figure 2, the experimental results show that the number of rules in our new algorithm is much less than that of the old algorithm, because with increasing number of attributes the number of the similar behaviour attributes is growing sharply.

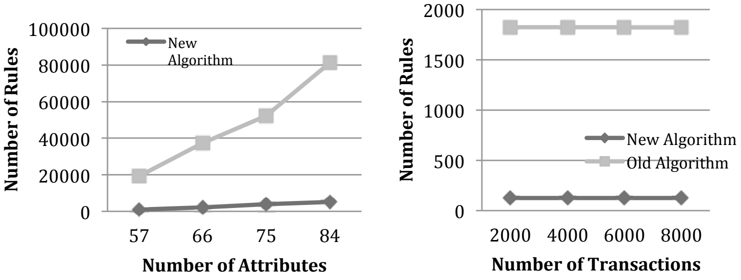


Fig. 2. Rules generated

From Figure 3, we can observe that fusing similar behaviour attributes causes decrease the execution time of the algorithm because it causes reduction of dataset size and of the number of candidate fuzzy itemsets.

Next we consider the algorithm in state of using ontology for generating candidate fuzzy itemsets after fusing similar behaving attributes in both datasets. In the following figures, we show the number of candidates generated and total time when the total number of attributes is increased. Increasing the number of attributes leads to increased number of candidates, most of which are created in the second pass. In our algorithm, the number of the candidate itemsets is growing slowly

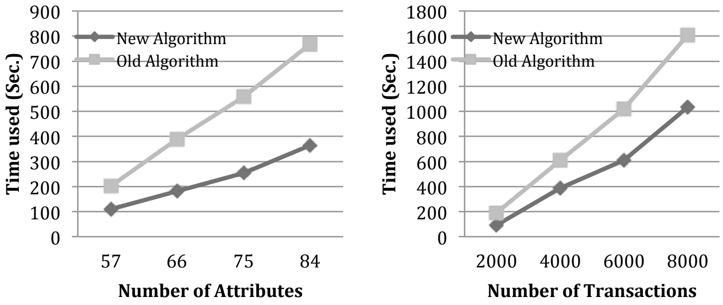


Fig. 3. Execution time

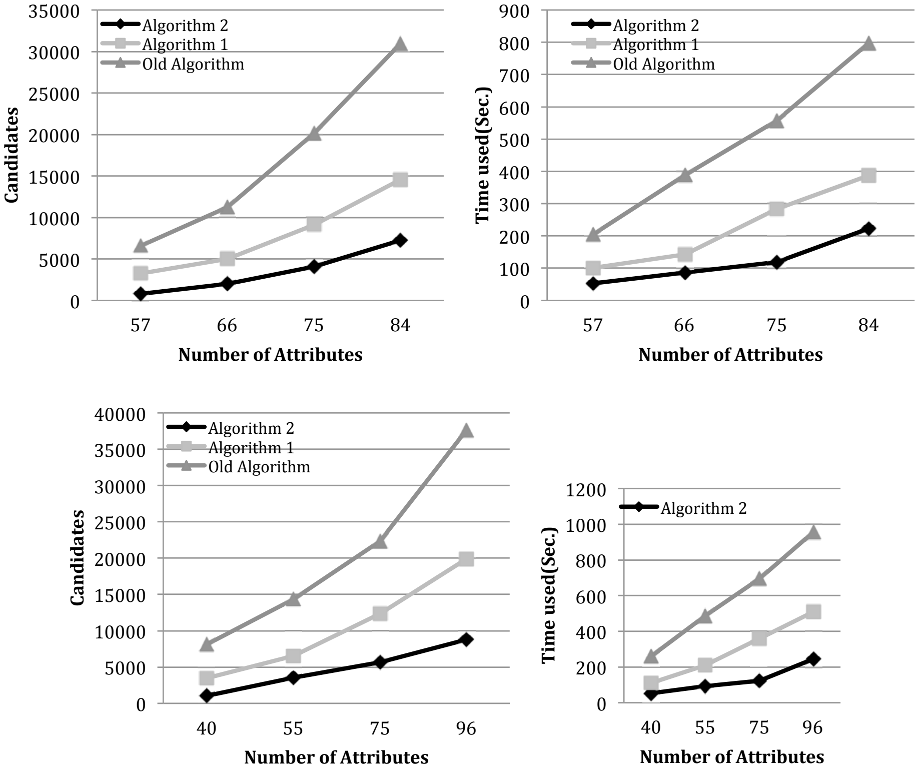


Fig. 4. Total running times and candidates depending on the number of attributes a) for synthetic dataset b) for clinical dataset

in comparison with the old algorithm. A correlation can be observed between the number of candidates and the execution time for this algorithm. The experimental results show that the running time of our new algorithm is much less than that of the old algorithm. The running time of the algorithms is determined by the number of passes and execution time of each pass. For our algorithm, the number of passes is less the old algorithm, because it removes many candidate itemsets. With decreasing the number of the candidate itemsets, the size of the largest frequent itemset reduces, which is the number of passes the algorithm has to perform. Furthermore, the execution time of each passes is less than the old algorithm, because of reduction in the number of calculating operations.

5 CONCLUSION

Association rule mining is an active data mining research area. Fuzzy association rules described by the natural language are well suited for thinking of human subjects. Thus, fuzzy association rules will be helpful to increase the flexibility for the users in making any decisions or designing the fuzzy systems. If mining procedure also produces a huge number of rules, it will be of limited use because a human user does not have the ability to analyze these rules. However, if such a huge number of rules do exist in the data, it will not be appropriate to arbitrarily discard any of them or to generate only a small subset of them. It is much more desi.rable if we can summarize them. This paper proposes such a technique. In this way, the fuzzy association rules can be manually inspected by a human user without too much effort. In our proposed method, the size of average transactions and original dataset is reduced effectively by the recognition and fusion of similar behaving attributes and mining is performed on the reduced dataset that produces a much smaller but richer set of fuzzy association rules which has been approved by experimental results.

REFERENCES

[1] Cai, Y.—Cercone, N.—Han, J.: Attribute-Oriented Induction in Relational Databases. In G. Piatetsky-Shapiro and W. J. Frawley (Eds.), Knowledge Discovery in Databases, AAAI/MIT Press 1991, pp. 213–228.

[2] Fayyad, U.—Piatetsky-Shapiro, G.—Smyth, P.: Knowledge Discovery and Data Mining: Towards a Unifying Framework. In KDD-96 Conference Proceedings, AAAI Press 1996.

[3] Han, J. W.—Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco 2001.

[4] Agrawal, R.—Imielinski, T.—Swami, A.: Mining Association Rules Between Sets of Items in Large Databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, May 1993, pp. 207–216.

[5] Srikant, R.—Agrawal, R.: Mining Quantitative Association Rules in Large Relational Tables. Proc. of ACM-SIGMOD, Montreal, Canada 1996.

[6] Zadeh, L. A.: Fuzzy Sets. Inform. and Control, Vol. 8, 1965, pp. 338–353.

[7] Hullermeier, E.—Beringer, J.: Mining Implication-Based Fuzzy Association Rules in Databases. In: B. Bouchon-Meunier, L. Foulloy, and R. R. Yager (Eds.): Intelligent Systems for Information Processing: From Representation to Applications, Elsevier 2003.

[8] Ruan, D.—Kerre, E. E.: Fuzzy Implication Operators and Generalized Fuzzy Method of Cases. Fuzzy Sets and systems, Vol. 54, 1993, No. 1, pp. 23–38.

[9] Miller, R. J.—Yang, Y.: Association Rules over Interval Data. In: Proc. ACM SIGMOD Internat. Conf. Management of Data 1997, pp. 452–461.

[10] Chan, K. C. C.—Au, W. H.: An Effective Algorithm for Mining Interesting Quantitative Association Rules. In Proc. of the 12th ACM Symp. on Applied Computing, San Jose, CA, Feb. 1997, pp. 88–90.

[11] Chan, K. C. C.—Au, W. H.: Mining Fuzzy Association Rules. In Proc. of the 6th ACM Int'l Conf. on Information and Knowledge Management, Las Vegas, Nevada, Nov. 1997, pp. 209–215.

[12] Hong, T. P.—Kuo, C. S.—Chi, S. C.: Mining Association Rules from Quantitative Data. Intell. Data Anal., Vol. 3, 1999, pp 363–376.

[13] Chen, G. Q.—Wei, Q.: Fuzzy Association Rules and the Extended Mining Algorithms. Information Sciences, Vol. 147, pp. 201–228.

[14] Cai, C. H.—Fu, W. C.—Cheng, C. H.—Kwong, W. W.: Mining Association Rules with Weighted Items. In: Proc. IDEAS 1998, pp. 68–77.

[15] Gyenesei, A.: Mining Weighted Association Rules for Fuzzy Quantitative Items. TUCS Technical Report No: 346, May 2000.

[16] Yue, S.—Tsang, E.—Yeung, D.—Shi, D.: Mining Fuzzy Association Rules with Weighted Items. Proc. IEEE Internat. Conf. Systems Man Cybernet., 2000, pp. 1906–1911.

[17] Lee, K. M.: Mining Generalized Fuzzy Quantitative Association Rules With Fuzzy Generalization Hierarchies. 0-7803-7078-3/01/10.00, 2001, IEEE.

[18] Hu, Y. C.—Chen, R. Sh.—Tzeng, G. H.: Discovering Fuzzy Association Rules Using Fuzzy Partition Methods. Knowledge-Based Systems, Vol. 16, 2003, pp. 137–147.

[19] Au, W. H.—Chan, K. C. C.: Mining Fuzzy Association Rules in a Bank-Account Database. IEEE Transactions on Fuzzy Systems, Vol. 11, 2003, No. 2.

[20] Gao, Y.—Ma, J.—Ma, L.: A New Algorithm for Mining Fuzzy Association Rules. Proc. IEEE Internat. Conf. Machine Learning and Cybernetics 2004.

[21] Ishibuchi, H.—Nakashima, T.—Yamamoto, T.: Fuzzy Association Rules for Handling Continuous Attributes. In: Proc. IEEE ISIE 2001.

[22] Gyenesei, A.: A Fuzzy Approach for Mining Quantitative Association Rules. TUCS Technical Report No: 336, March 2000.

[23] Zhang, W.: Mining Fuzzy Quantitative Association Rules. Proc. IEEE Internat. Conf. Tools Artif. Intell., 1999, pp. 99–102.

[24] Yager, R. R.: Fuzzy Summaries in Database Mining. Proc. Conf. Artif. Intell. Appl., 1995, pp. 265–269.

[25] Mills, F.: Statistical Methods. Pitman 1955.

[26] Fu, A. W. C.—Wong, M. H.—Sze, S. C.—Wong, W. C.—Wong, W. L.—Yu, W. K.: Finding Fuzzy Sets for the Mining of Association Rules for Numerical Attributes. In: Proc. IDEL, October 1998, pp. 263–268.

[27] Ng, R.—Han, J.: Efficient and Effective Clustering Methods for Spatial Data Mining. In: Proc. Internat. Conf. Very Large Databases 1994.

[28] Hirota, K.—Pedrycz, W.: Linguistic Data Mining and Fuzzy Modelling. Proc. IEEE Internat. Conf. Fuzzy Systems, Vol. 2, 1996, pp. 1448–1496.

[29] Kaya, M.—Alhajj, R.: Facilitating Fuzzy Association Rules Mining by Using Multi-Objective Genetic Algorithms for Automated Clustering. In: Proc. IEEE Internat. Conf. Data Mining, November 2003.

[30] Kaya, M.—Alhajj, R.: Genetic Algorithm Based Framework for Mining Fuzzy Association Rules, Fuzzy Sets and Systems. Elsevier 2004.

[31] Zadeh, L. A.: The Concept of a Linguistic Variable and Its Application to Approximate Reasoning. Information Science, Part 1. 1975.

[32] Civanlar, M. R.—Trussell, H. J.: Constructing Membership Functions Using Statistical Data. Fuzzy Sets and Systems, Vol. 18, 1986, pp. 1–14.

[33] Kohonen, R.: Self-Oganization and Associate Memory. Springer Verlag Berlin 1988.

[34] Gómez-Pérez, A.—Fernández-López, M.—Corcho, O.: Ontological Engineering: With Examples from the Areas of Knowledge Management, E-commerce and the Semantic Web. Springer Verlag 2004.

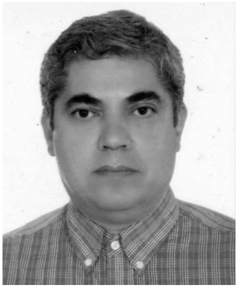
[35] De Nicola-Missikoff, M.—Navigli, R.: A Software Engineering Approach to Ontology Building. Information Systems, Vol. 34, 2009, pp. 258–275.

[36] Corcho, O.—Fernandez, M.—Gomez-Perez: Methodologies, Tools and Languages for Building Ontologies: Where is the Meeting Point? Data and Knowledge Engineering 46, 2003.

[37] `http://www.anzdata.org.au/`.

[38] `http://www.nlm.nih.gov/research/umls/`.

Zahra Farzanyar is currently a Ph. D. candidate in computer science from the Iran University of Science and Technology, Tehran, Iran, working under the supervision of Dr. Mohammadreza Knagavari. She is finalizing her Ph. D. thesis on "Developing Frequent Itemsets Mining Algorithms in Large scale Peer to Peer Environments with Data stream Theory". Her research interests are data stream mining and distributed and peer to peer data mining.



Mohammadreza Kangavari received the B. Sc. degree in mathematics and computer science form the Sharif University of Technology in 1982, the M. Sc. degree in computer science from Salford University in 1989, and the Ph. D. degree in computer science from the University of Manchester in 1994. He is currently a lecturer in the Computer Engineering Department, Iran University of Science and Technology, Tehran, Iran. His research interests include data mining, machine learning, and natural language processing.