Integrated Genetic-Fuzzy Approach for Mining Quantitative Association Rules

Shaikh Nikhat Fatma Shaikh
Department Of Computer , Mumbai University ,
Pillai's Institute Of Information Technology, New Panvel

Jagdish W Bakal
Phd,Mumbai University
Shivajirao S Jondhale College Of Engineering,
Dombivli (E)

Madhu Nashipudimath
Department of Information Technology Pillai's Institute of Information Technology,
New Panvel

ABSTRACT
Data mining of association rules from items in transaction databases has been studied extensively in recent years. However these algorithms deal with only transactions with binary values whereas transactions with quantitative values are more commonly seen in real-world applications. As to fuzzy data mining, many approaches have also been proposed for mining fuzzy association rules. Most of the previous approaches, however, set a single minimum support threshold for all the items or itemsets and identify the relationships among transactions. In real applications, different items may have different criteria to judge their importance and quantitative data may exist. Thus the fuzzy data mining approaches are divided into two types, namely single-minimum-support fuzzy-mining (SSFM) and multiple-minimum-support fuzzy-mining (MSFM) problems. These algorithms integrates fuzzy set concepts and the apriori mining algorithm to find fuzzy association rules in given transaction data sets.

Keywords
k-means Clustering, data mining, fuzzy set, genetic algorithm, Fuzzy Association Rules, Quantitative transactions

1. INTRODUCTION
At present, more and more databases containing large quantities of data are available. These industrial, medical, financial and other databases make an invaluable resource of useful knowledge. The task of extraction of useful knowledge from databases is challenged by the techniques called data-mining techniques . One of the widely used data-mining techniques is association rules mining. Data Mining is commonly used in attempts to induce association rules from transaction data. An association rule is an expression X→Y, where X is a set of items and Y is a single item. It means in the set of transactions, if all the items in X exist in a transaction, then Y is also in the transaction with a high probability. Association rules mining identifies associations (patterns or relations) among database attributes and their values. It is a pattern-discovery technique which does not serve to solve classification problems (it does not classify samples into some target classes) nor prediction problems (it does not predict the development of the attribute values). Association rules mining generally searches for any associations among any attributes present in the database. An example of association rule can be as follows 'if a customer buys a toothbrush, then he also probably buys toothpaste (in the same transaction)', the rule can be written as: {toothbrush} → {toothpaste}.
Transaction data in real-world applications, however, usually consist of quantitative values. For example, assume whenever customers in a supermarket buy bread and butter, they will also buy milk. From the transactions kept in the supermarkets, an association rule such as "Bread and Butter →Milk" will be mined out. Most previous studies focused on binary valued transaction data. Transaction data in real-world applications, however, usually consist of quantitative values. Many sophisticated data-mining approaches have also been proposed to deal with various types of data presents a challenge to workers in this research field.
In the remaining paper in section 2 we take an over view of Association Rule Mining . In section 3 we take the review of how fuzzy mining is done and how it actually works. In section 4 we have the drawbacks of fuzzy mining. In section 5 we have Integrated fuzzy genetic approaches and in section 6 we make a conclusion.

2. ASSOCIATION RULE MINING
Data mining techniques have been used in different fields to discover interesting information from databases in recent years. Association rules mining is one of the popular data mining techniques since it was proposed by Agrawal et al. [15]. Many business enterprises accumulate large quantities of data from their day-to-day operations. For example, huge amounts of customer purchase data are collected daily at the checkout counters of grocery stores. Table 1 illustrates an example of such data, commonly known as market basket transactions. Each row in this table corresponds to a transaction, which contains a unique identifier labeled TID and a set of items bought by a given customer. Retailers are interested in analyzing the data to learn about the purchasing behavior of their customers. Such valuable information can be used to support a variety of business-related applications such as marketing promotions, inventory management, and customer relationship management.
This methodology is known as association analysis, which is useful for discovering interesting relationships hidden in large data sets. The uncovered relationships can be represented in the form of association rules or sets of frequent items. For example, the following rule can be extracted from the data set shown in Table 1:
{Diapers} → {Beer}.

Table 1. An example of market basket transactions.

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

The rule suggests that a strong relationship exists between the sale of diapers and beer because many customers who buy diapers also buy beer. Retailers can use this type of rules to help them identify new opportunities for cross selling their products to the customers.
Besides market basket data, association analysis is also applicable to other application domains such as bioinformatics, medical diagnosis, Web mining, and scientific data analysis. In the analysis of Earth science data, for example, the association patterns may reveal interesting connections among the ocean, land, and atmospheric processes. Such information may help Earth scientists develop a better understanding of how the different elements of the Earth system interact with each other. Even though the techniques presented here are generally applicable to a wider variety of data sets, for illustrative purposes, our discussion will focus mainly on market basket data. There are two key issues that need to be addressed when applying association analysis to market basket data. First, discovering patterns from a large transaction data set can be computationally expensive. Second, some of the discovered patterns are potentially spurious because they may happen simply by chance.

3. FUZZY MINING

For many applications, an association rule may be more interesting if it reveals relationship among some useful concepts, such as "high income", "new car", and "frequent customer". These concepts are often imprecise or uncertain. Interesting concepts are defined using fuzzy terms and interpreted based on fuzzy set. We refer association rules involving fuzzy terms as *fuzzy quantitative association rules*.

For example, Age = young and income = high →Risk Level = medium High is a fuzzy quantitative association rule, where "young", "high" and "medium High" are fuzzy terms.

This is a process of fuzzifying numerical numbers into linguistic terms, which is often used to reduce information overload in hurnan decision making process. The numerical salary, for example, may be perceived in linguistic terms as high, average and low. One way of determining membership functions of these linguistic terms is by expert opinion or by people's perception.
Fuzzy association rules use linguistic variables. These linguistic variables define the value of a variable both qualitatively, by defining a symbol for a fuzzy set, and quantitatively, by defining the meaning of the fuzzy set.

Table 2. An example of fuzzy dataset.

ID	Age	Degree	Salary
E1	Adult	M.Tech.	30000 (High)
E2	Old	B.A.	18000 (Normal)
E3	Young	B.Tech.	28000 (High)
E4	Adult	M.C.A.	10000 (Low)

Table 2 contains a sample fuzzy dataset. We can determine the value of the attribute i_k of the j^{th} record by using the convention $t_j[i_k]$. For example, if we want to determine the value of salary of third record, we will write t_3[Salary] and obtain the value 28000. In Table 1, the attribute salary has been denoted using the fuzzy set Salary = {high, normal, low}, dividing the salary interval into low, normal and high. For the interval (Rs. 10000 to Rs. 30000) we have normal salary, for (Rs. 10,000 and below) we have low salary and for (Rs.30,000 and above)we have high salary.
Basically, the fuzzy mining algorithms first uses membership functions to transform each quantitative value into a fuzzy set in linguistic terms. The algorithm then calculated the scalar cardinality of each linguistic term on all the transaction data. The mining process based on fuzzy counts was then performed to find fuzzy association rules.

4. DRAWBACKS OF FUZZY MINING

In fuzzy data mining, the first thing that it needs to be done is to define appropriate membership functions because they have a critical influence on the final mining results. Generally speaking, membership functions are defined by experts. That is the best approach, absolutely. However, experts may not always do this since the customers' favorites change all the time. Most of the previous fuzzy data mining algorithms thus assume the membership functions are already known. Of course, it is not suitable when we try to apply it to real applications. The algorithms that can derive both the appropriate membership functions and fuzzy rules automatically are thus developed. . Most of these fuzzy data mining algorithms assume the membership functions are already known. The developing of mining algorithms that can mine both appropriate membership functions and fuzzy association rules automatically is thus an important task. There are two reasons for the drawback of fuzzy mining. The first one is that companies may not always ask experts to define the appropriate membership function because it needs to spend lots of money and time. The second reason is that the favourite things of customers change all the time. Some mechanisms are thus needed to adapt the membership functions to these changes automatically.

5. INTEGRATED GENETIC FUZZY APPROACHES

A survey of several algorithms that can mine both appropriate membership functions and fuzzy association rules were made. We divide them into two different genetic-fuzzy data mining problems according to the utilized approaches and two types of problems in fuzzy data mining, namely Integrated Genetic-Fuzzy approach for items with Single Minimum Supports

(IGFSMS) , Integrated Genetic-Fuzzy approaches for items with Multiple Minimum Supports (IGFMMS).
In the integrated genetic-fuzzy approaches, they encoded all membership functions of all items (attributes) into a chromosome (also called an individual). The genetic algorithms are then used to derive a set of appropriate membership functions according to the designed fitness function. Finally, the best set of membership functions are then used to mine fuzzy association rules.

5.1 Integrated Genetic-Fuzzy approaches for items with Single Minimum Support

In IGFSMS problem, many approaches have been published [2], Hong et al. proposed a genetic-fuzzy data-mining algorithm for extracting both association rules and membership functions from quantitative transactions [2]. They propose a GA-based framework for searching membership functions suitable for mining problems and then use the final best set of membership functions to mine association rules. The proposed framework is shown in Figure 1.

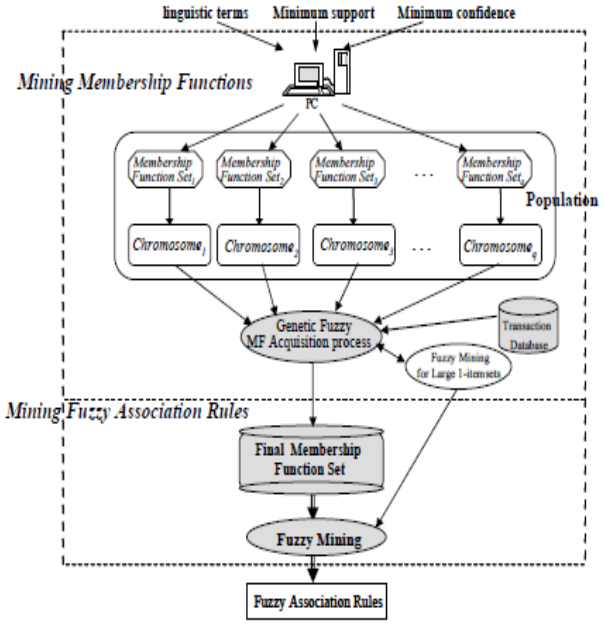

Fig 1: GA-based framework for searching membership functions

The proposed framework consists of two phases, namely mining membership function and mining fuzzy association rules phases. In the first phase, the proposed framework maintains a population of sets of membership functions, and uses the genetic algorithm to automatically derive the resulting one. It first transforms each set of membership functions into a fixed-length string. The chromosome was then evaluated by the number of large 1-itemsets and the suitability of membership functions. The fitness value of a chromosome *Cq* is then defined as:

$$f(C_q) = \frac{|L_{1q}|}{\text{Suitability}(C_q)}$$

where $|L1|$ is the number of large 1-itemsets obtained by using the set of membership functions in *Cq*. The suitability measure was used to reduce the occurrence of bad types of membership functions. The two bad types of membership function are shown in Figure 2, where the first one is too redundant, and the second one is too separate.

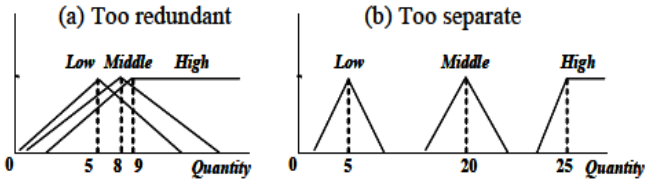

Fig 2:The two bad types of membership functions

Two factors, overlap factor and coverage factor, are used to avoid those shapes. The overlap factor is designed for avoiding the first bad case (too redundant), and the coverage factor is for the second one (too separate). After fitness evaluation, it then chooses appropriate chromosomes for mating, gradually creating good offspring membership function sets. The offspring membership function sets then undergo recursive evolution until a good set of membership functions has been obtained. In the second phase, the final best membership functions are gathered to mine fuzzy association rules. The fuzzy mining algorithm is adopted to achieve this purpose.[2]

The calculation for large 1-itemsets, however, would still take a lot of time, especially when the database to be scanned could not totally fed into main memory. An enhanced approach, called the cluster-based fuzzy-genetic mining algorithm were thus proposed to speed up the evaluation process and keep nearly the same quality of solutions as that in [2]. The proposed approach also maintains a population of sets of membership functions, and uses the genetic algorithm to automatically derive the resulting one. Before fitness evaluation, the clustering technique was first used to cluster chromosomes. In other words, it used the *k*-means clustering approach to gather similar chromosomes into groups. The two factors, overlap factor and coverage factor, were then used as attributes for clustering. For example, assume there have ten chromosomes with its coverage and overlap factor and shown in Table 3, where the column "suitability" represents the pair (coverage factor, overlap factor). The *k*-means clustering approach is executed to divide the ten chromosomes into *k* clusters. In this example, assume the parameter *k* is set at 3. The three clusters found are shown in Table 3. The representative chromosomes in the three clusters are *C5* (4.37, 0.33), *C4* (4.66, 0) and *C9* (4.09, 8.33).

Table 3. The coverage and the overlap factor of ten chromosomes

C_q	*Suitability*	C_q	*Suitability*
C_1	(4, 0)	C_6	(4.5, 0)
C_2	(4.24, 0.5)	C_7	(4.45, 0)
C_3	(4.37, 0)	C_8	(4.37, 0.53)
C_4	(4.66, 0)	C_9	(4.09, 8.33)
C_5	(4.37, 0.33)	C_{10}	(4.87, 0)

Table 4. The three clusters found from the ten chromosomes

$Cluster_i$	*Chromosome*	*Representative chromosome*
$Cluster_1$	C_1, C_2, C_5, C_8	C_5
$Cluster_2$	C_3, C_4, C_6, C_7, C_{10}	C_4
$Cluster_3$	C_9	C_9

All the chromosomes in a cluster use the number of large 1-itemsets derived from the representative chromosome in the cluster and their own suitability of membership functions to calculate their fitness values. Since the number for scanning a database decreases, the evaluation cost can thus be reduced. In this example, the representative chromosomes are chromosomes *C4*, *C5*, *C9* and it only needs to calculate the number of large 1-itemsets three times. The evaluation results are utilized to choose appropriate chromosomes for mating in the next generation. The offspring membership function sets then undergo recursive evolution until a good set of membership functions has been obtained. Finally, the derived membership functions are used to mine fuzzy association rules.

Besides, Kaya et al. also proposed several genetic-fuzzy data mining approaches to derive membership functions and fuzzy association rules. In [14] the proposed approach tries to derive membership functions, which can get a maximum profit within an interval of user specified minimum support values and then using the derived membership functions to mine fuzzy association rules. The concept of their idea is shown in Figure 3.

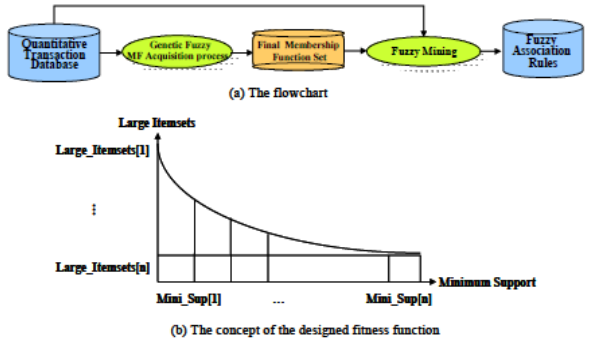

Figure 3: The concept of Kaya's approach

As shown in Figure 3(a), they first used genetic-fuzzy mining approach to derived membership functions form the given quantity transaction database. The final membership functions are then used to mining fuzzy association rules. The concept of maximizing the large itemsets of the given minimum support interval is used to design the fitness function as shown in Figure 3(b). They also extended the approach to mine fuzzy weighted association rules. Furthermore, since the fitness functions are not easily to be defined for optimizing problems, the multi-objective genetic algorithm has thus been developed for this situation. In other words, there always needs more than one criterion to do things well in real applications. A set of solutions, namely non-dominate points (also called Pareto-Optimal Surface) is derived and given to user instead the best one solution by genetic algorithm. Kaya et al. thus proposed a multi-objective genetic algorithm based approaches for mining membership functions, which can generate interesting fuzzy association rules. Three objective functions, namely strongness , interestingness and comprehensibility, are used to find the Pareto-Optimal Surface.

5.2 Integrated Genetic-Fuzzy approaches for Items with Multiple Minimum Supports

In above subsection, we can know that lots of researches are focused on integrated genetic-fuzzy approaches for items with single minimum support. However, different items may have different criteria to judge their importance. For example, if there have some items and they are expensive, then people usually may not buy it. It is easy to know that the support value of that item is low. Even people may not buy it, manager may also interest on those produces because of their high profit. In such cases, the IGFSMS approaches may not suitable for this problem. In this subsection, we then introduced another genetic-fuzzy data mining approach which extends the approach proposed in [2] to solve this problem. The proposed algorithm combines the clustering, fuzzy and genetic concepts to derive minimum support values and membership functions for items. The final minimum support values and membership functions are then used to mine fuzzy association rules. The genetic-fuzzy mining framework is shown in Figure 4. It is also the first proposed approach for achieving the purpose.

As shown in Figure 4, it can be divided into two phases. The first phase searches for suitable minimum support values and membership functions of items and the second phase uses the final best set of minimum support values and membership functions to mine fuzzy association rules. The proposed framework maintains a population of sets of minimum support values and membership functions, and uses the genetic algorithm to automatically derive the resulting one. It first uses the *k*-means clustering approach to gather similar items into groups. All items in the same cluster are considered to have similar characteristics and are assigned similar values when a population is initialized. The values (or initialization information) include an appropriate number of linguistic terms for each item, its reasonable membership functions, and a range of its possible minimum support values. It then generates and encodes each set of minimum support values and membership functions into a fixed-length string according to the initialization information.

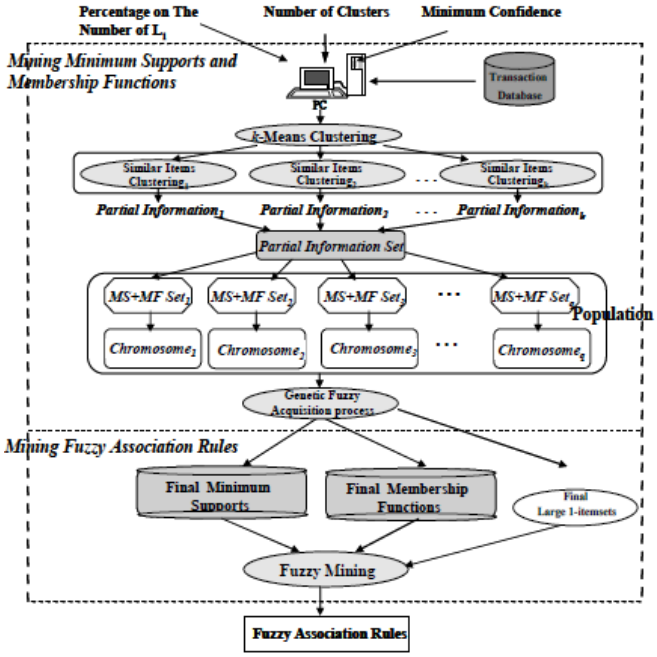

Figure 4: The proposed genetic-fuzzy mining framework for items with multiple minimum supports

In this approach, the minimum support values of the items may be different. It is hard to assign the values. As an alternative, the values can be determined according to the required number of rules. It is, however, very time-consuming to obtain the rules for each chromosome. Usually, a larger number of 1-itemsets will result in a larger number of all itemsets with a higher probability, which will thus usually imply more interesting association rules. The evaluation by 1-itemsets is faster than that by all itemsets or interesting association rules. Using the number of large 1-itemsets can thus achieve a trade-off between execution time and rule interestingness [2].

A criterion should thus be specified to reflect the user preference on the derived knowledge. In this approach, the required number of large 1-itemsets (*RNL*) is proposed for this purpose. It is the number of linguistic large 1-itemsets that a user wants to get from an item. It can be defined as the number of linguistic terms of an item multiplied by the predefined percentage which reflects users' preference on the number of large 1-itemsets. It is used to reflect the closeness degree between the number of derived large 1-itemsets and the required number of large 1-itemset. For example, assume there are three linguistic terms for an item and the predefined percentage *p* is set at 80%. The *RNL* value is then set as $\lfloor 3*0.8 \rfloor$, which is 2. *RNL* is thus used in the fitness function to evaluate the goodness of a chromosome. For example, assume there are three linguistic terms for an item and the predefined percentage p is set at 80%. The *RNL* value is then set as $\lfloor 3*0.8 \rfloor$, which is 2. The fitness function is then composed of the suitability of membership functions and the closeness to the *RNL* value. The minimum support values and membership functions can thus be derived by GA and are then used to mine fuzzy association rules by a fuzzy mining approach for multiple minimum supports such as the one in[13].

6. CONCLUSION

Since association rules was proposed Agrawal et al., lots of researchers devoted themselves to it, no matter what they just improve the existing algorithms, create new ideas or combine the existing one with other techniques. In fact, association rules mining approaches become an important topic. In this article, we thus focus on genetic-fuzzy data mining techniques and lots of literatures are introduced.

7. REFERENCES

[1] Chun-Hao Chen, Tzung-Pei Hong, "Cluster-Based Evaluation in Fuzzy-Genetic Data Mining" ,IEEE transactions on fuzzy systems, Vol. 16, No. 1, February 2008 249, pp. 249-262.

[2] T. P. Hong, C. H. Chen, Y. L. Wu, and Y. C. Lee, "A GA-based fuzzy mining approach to achieve a trade-off between number of rules and suitability of membership functions," *Soft Computing*, vol. 10, no. 11, pp. 1091–1101, 2006.

[3] Hung-Pin Chiu, Yi-Tsung Tang, "A Cluster-Based Mining Approach for Mining Fuzzy Association Rules in Two Databases", *Electronic Commerce Studies*, Vol. 4, No.1, Spring 2006, Page 57-74.

[4] Tzung-Pei Hong, Chan-Sheng Kuo, Sheng-Chai Chi, "Trade-Off Between computation time and number of rules for fuzzy mining from quantitative data", *International Journal of Uncertainty, Fuzziness and Knowledge-Based systems*, Vol. 9, No. 5, 2001, page 587- 604.

[5] M. Sulaiman Khan, Maybin Muyeba, Frans Coenen, "Fuzzy Weighted Association Rule Mining with Weighted Support and Confidence Framework", *The University of Liverpool, Department of Computer Science, Liverpool, UK.*

[6] H. Ishibuchi and T. Yamamoto, "Rule weight specification in fuzzy rule-based classification systems," *IEEE Trans. on Fuzzy Systems*, Vol. 13, No. 4, pp. 428-435, August 2005.

[7] Miguel Delgado, Nicolás Marín, Daniel Sánchez, and María-Amparo Vila," Fuzzy Association Rules: General Model and Applications",*IEEE transactions on fuzzy systems*, vol. 11, no. 2, April 2003

[8] Tzung-Pei Hong, Li-Huei Tseng and Been-Chian Chien,"Learning Fuzzy Rules from Incomplete Quantitative Data by Rough Sets",

[9] H. J. Zimmermann, "Fuzzy set theory and its applications", *Kluwer Academic Publisher*, Boston, 1991.

[10] Tzung-Pei Hong, Ming-Jer Chiang and Shyue-Liang Wang "Mining from Quantitative Data with Linguistic Minimum Supports and Confidences",2002 IEEE Proceedings.

[11] S. Yue, E. Tsang, D. Yeung, and D. Shi, "Mining fuzzy association rules with weighted items," in *Proc. IEEE Int. Conf. on Systems, Man and Cybernetics*, 2000, pp. 1906–1911.

[12] M. Kaya, R. Alhajj , *"Genetic algorithm based framework for mining fuzzy association rules",* 2004 Elsevier B.V.

[13] Y. C. Lee, T. P. Hong and W. Y. Lin, "Mining fuzzy association rules with multiple minimum supports using maximum constraints", *Lecture Notes in Computer Science*, Vol. 3214, pp. 1283-1290, 2004.

[14] M. Kaya and R. Alhajj, "A clustering algorithm with genetically optimized membership functions for fuzzy association rules mining," *The IEEE International Conference on Fuzzy Systems*, pp. 881-886, 2003.

[15] R. Agrawal, T. Imielinksi and A. Swami, "Mining association rules between sets of items in large database," *The 1993 ACM SIGMOD Conference*, Washington DC, USA, 1993.