ECOGA: Efficient Data Mining Approach for Fuzzy Association Rules

Wanneng Shu[1,2]
[1]State key lab of software engineering, Wuhan University
Wuhan 430072, China
[2]College of Computer Science, South-Central University for Nationalities
Wuhan 430074, China
Email:shuwanneng@yahoo.com.cn

Lixin Ding[1, 3,*]
[3] Computer School, Wuhan University, Wuhan, 430072, China
Email: lxding@whu.edu.cn

***Abstract*—Data mining is concerned with developing algorithms and computational tools and techniques to help people extract patterns from data. In this paper an efficient data mining approach, which is based on fuzzy set theory and clonal selection algorithm, is proposed. The main motivation is to benefit from the global search performed by this kind of algorithms. Experimental results show the number of fuzzy association rules obtained with the proposed method is larger than those obtained by applying other methods.**

***Index Terms*—data mining, association rules, fuzzy sets, efficient clonal optimizing genetic algorithm**

I. INTRODUCTION

In recent years, information growth has proceeded at an explosive rate. We are confronted with the paradox that more data useless information. This has opened a new challenge for computer science: the discovery of new and meaningful information. A new generation of techniques and tools are required to assist humans in intelligently analyzing voluminous data for pieces of useful knowledge.Data mining is concerned with developing algorithms and computational tools and techniques to help people extract patterns from data [1]. Finding the patterns by identifying the underlying rules and features in the data is done in an automatic way. Association rules are used to represent and identify dependencies between items in a database [2].These are an expression of the type $X \Rightarrow Y$ ($X \cap Y = \varnothing$), where both X and Y are defined as sets of attributes or items. It means that if all the items in X exist in a transaction then all the items in Y are also in the transaction with a high probability, and X and Y should not have a common item [3].

Most data mining algorithms are derived from machine learning, pattern recognition and statistics. These algorithms include classification, clustering and graphical models. Fuzzy sets are sets whose elements have degrees of membership. In classical set theory, the membership of elements in a set is assessed in binary terms according to a bivalent condition.Fuzzy set theory has been used more and more frequently in intelligent systems because of its simplicity and similarity to human reasoning [4]. Fuzzy rule-based systems have been successfully applied to various application areas such as control and classification [5]. While the main objective in the design of fuzzy rule-based systems has been the performance maximization, their comprehensibility has also been taken into account in some recent studies [6]. The automatic definition of fuzzy systems can be considered as an optimization or search process and nowadays evolutionary algorithms, particularly genetic algorithms (GA), are considered as the better known and used global search technique. GA try to simulate some aspects of Darwin's natural selection and have been used in several areas to solve problems considered intractable. For this reason, GA have been successfully applied to learn and to tune fuzzy systems in the last years [7].

Evolutionary algorithms (EAs) have the ability to process sets of solutions in parallel and explore big search spaces in reasonable time. There have been several efforts to make use of models based on GA in data mining [8]. In this paper we present an ECOGA (Efficient Clonal Optimizing Genetic Algorithm) method to derive the fuzzy sets from a set of given transactions, which combines the clonal selection mechanism of the immune system with the evolution equation of genetic algorithm. They can enhance exploratory capabilities, and keep a desirable balance between global search and local search, so as to accelerate the convergence speed.

The outline of the contribution is organized as follows. Related work is discussed in Section 2. Fuzzy quantitative association rules mining are defined in Section 3. The ECOGA process to find fuzzy sets and their membership functions is described in Section 4. Experiments are given in Section 5. Finally, section 6 concludes the paper and discusses some future research directions.

*Corresponding author

doi:10.4304/jsw.6.1.91-99

II. Related Works

The fuzzy set concept has recently been used more frequently in mining quantitative association rules [9], where the membership functions was assumed to be known in advance. In [10] the fuzzy set theory introduced provides an excellent means to model the "fuzzy" boundaries of linguistic terms by introducing gradual membership. In [11] use of fuzzy sets in discovering association rules for quantitative attributes e.g. However, in existing approaches fuzzy sets are either supplied by an expert or determined by applying an existing known clustering algorithm.

In [12] applied Birch clustering to identify intervals and proposed a distance-based association rules mining process, which improves the semantics of the intervals. In [13] introduced fuzzy linguistic summaries on different attributes. In [14] proposed a context sensitive fuzzy clustering method based on fuzzy C-means to construct rule-based models. However, the context-sensitive fuzzy C-means method cannot deal with the data consisting of both numerical and categorical attributes. In [15] applied fuzzy linguistic terms to relational databases with numerical and categorical attributes. Later, they proposed the F-APACS method to discover fuzzy association rules. They utilized adjacent difference analysis and fuzziness in finding the minimum support and confidence values instead of having them supplied by a user. They determine both positive and negative associations. In [16] proposed an algorithm that integrates fuzzy set concepts and Apriori mining algorithm to find interesting fuzzy association rules from given transactional data. In [17] proposed definitions for the support and confidence of fuzzy membership grades and designed a data mining approach based on fuzzy sets to find association rules with linguistic terms of human knowledge. In [18], a new rule representation model has been presented based on these concepts to perform a tuning of complex linguistic fuzzy models. An approach in [19] uses fuzzy association thesaurus and query expansion for information retrieval. Fuzzy composition operations like max-min, max-product and sum – product are used for constructing the thesaurus. Interactive query expansion shows the user, upon initial query, a ranked list of documents suggested by the system based on the fuzzy relation composition.

Since data mining works successfully to find valuable information within large datasets, it should be useful to improve the GA if each chromosome is considered as a transaction behavior. GA-miner is designed for running GA on large scale parallel data mining. M. Kaya proposed a GA-based clustering method to derive a predefined number of membership functions for getting a maximum profit [20].

III. Fuzzy Association Rules

Let $TI=\{T,I\}$ be a model of fuzzy association rules in data mining, in which $I=\{i_1,i_2,...,i_k,...,i_m\}$ represent all attributes (items) that appear in T, and $T=\{t_1,t_2,...t_k,...,t_n\}$ be a set of transactions, where each transaction $t_k(1\le k\le n)$ is a set of items such that $t_k\subseteq I$ .Each quantitative attribute i_k is associated with at least two fuzzy sets. Given a database of transactions T, its set of attributes I, and the fuzzy sets associated with quantitative attributes in I. Note that each transaction t_k contains values of some attributes from I and each quantitative attribute in I has two or more corresponding fuzzy sets. The target is to find out some interesting and potentially useful rules, i.e., fuzzy association rules with enough support and high confidence.

A fuzzy set is a pair (A,m) where A is a set and $m:A\to[0,1]$. For each $x\in A$, m(x) is called the grade of membership of x in (A,m). A fuzzy association rule is expressed as Rule R_q:

$$X \quad is \quad A \Rightarrow Y \quad is \quad B \quad (X\subset I, Y\subset I, X\cap Y=\varnothing)$$

Where $X=\{x_1,x_2,...,x_m\}$ and $Y=\{y_1,y_2,...,y_n\}$ are disjoint sets of attributes called itemsets; $A=\{f_1,f_2,...,f_m\}$ and $B=\{g_1,g_2,...,g_n\}$ contain the fuzzy sets associated with corresponding attributes in X and Y, respectively. f_i is the set of fuzzy sets related to attribute x_i and g_i is the set of fuzzy sets related to attributed y_i. For a rule to be interesting, it should have enough support and high confidence value, large than user specified thresholds.

In the area of data mining, two measures called confidence and support have often been used for evaluating association rules.The support of the fuzzy association rule is defined as the percentage of the records having both attributes X and Y ,namely $F\text{-}Sup(R_q)=\frac{|\{D:X\cup Y\subseteq T,D\in T\}|}{|T|}$,where $|T|$ is the number of transactions in database T ,and $|\{D:X\cup Y\subseteq T,D\in T\}|$ is the number of transactions that are compatible with both the antecedent X and the consequent Y . The support $F\text{-}Sup$ indicates the grade of the coverage by $R_q:X\Rightarrow Y$ .The confidence of the rule is defined as the percentage of the records having Y given that they also have X , namely $F\text{-}Conf(R_q)=\frac{|\{D:X\cup Y\subseteq T,D\in T\}|}{|D:X\subseteq T,D\in T|}$ .The confidence $F\text{-}Conf$ indicates the grade of the validity of $R_q:X\Rightarrow Y$ .If support and confidence exceed each minimum value, then $R_q:X\Rightarrow Y$ can be regarded as significant association rules in T .

IV. EFFICIENT OPTIMIZING CLONAL GENETIC ALGORITHMS

This section describes the characteristics of an ECOGA algorithm for fuzzy association rules. One of the most important steps in mining fuzzy association rules is to decide on the fuzzy sets according to which the values of each quantitative attribute are to be classified. In order to cope with these problems, we first concentrate on how fuzzy sets are determined automatically from the values of the given attributes. For this purpose, we have used an ECOGA-based clustering algorithm.

Clonal selection algorithm is very important for the immunology [21], and attracts much attention from the artificial intelligence researchers. In biological immune system, cloning means that a group of identical cells is generated from a single common ancestor and only antibodies with high affinity will be cloned to attack the pathogens [22]. Genetic algorithms are adaptive search methods that try to mimic the process of evolution. Genetic algorithms start with a population of solutions [23]. Every individual is assigned a fitness measure to indicate its quality. After generations, these survivors become similar to each other. With the random initial population, random crossover and mutation, different replications of a genetic algorithm may produce different solutions.

When using ECOGA to mine fuzzy association rules, there are three main concerns. The first is to define the internal representation. The internal representation defines the search space of an ECOGA algorithm. The second concern is to define the fitness measure. In our case, the fitness value is used as an optimization objective function for the algorithm. An ECOGA searches for individuals whose fitness is maximal. Thus, the aim of the ECOGA employed in this study is to maximize the number of all the large itemsets extracted by the adjusted membership functions. The third concern relates to the genetic operators. We use clonal, crossover, mutation and selection to build the individuals of the next generation. This paper presents an ECOGA algorithm that addresses these three concerns.

In ECOGA, cloning means that a group of identical cells is generated from a single common ancestor and only antibodies with high affinity will be cloned to attack the pathogens [24].Clone the antibody population, increases the searching scope and maintains the diversity of the population Antibodies are sorted by descending order in our algorithms. The clonal scale is proportional to fitness, namely, the large the antibody is, the more it is cloned.

The other individuals in the population are generated via crossover and mutation. Crossover generates new offspring by exchanging partial information from the parents. Crossover means exchanging substrings from pairs of chromosomes to form new pairs of chromosomes. Crossover involves combining pair of parents by randomly selecting a point at which pieces of parents are swapped. The probability of selection of parents depends on their fitness values. The single point crossover, which separates chromosomes into two substrings, and the double point crossover, which separates them into three substrings, is the most popular crossover methods. Mutation is the key mechanisms by which genetic algorithms explore the solution space. Mutation operators change the value of a gene to keep the solution's diversity. Mutation prevents the search process from falling into local maxima, but a mutation rate that is too high may cause great fluctuation, so the mutation rate is generally set at a low value.

The selection scheme involves survival competition and reproduction mating. The selection scheme controls the growth speed of "good" solutions. A number of different selection implementations have been proposed in the [25], such as roulette wheel selection, tournament selection, and linear normalization selection. Here linear normalization selection, which has a high selection pressure, has been implemented. Using the rank position rather than the actual fitness values avoids problems that occur when fitness values are very close to each other or when an extremely fit individual is present in the population. This selection technique pushes the population toward the solution in a reasonably fast manner, avoiding the risk of a single individual dominating the population in the space of one or two generations.

ECOGA extracts the clonal selection mechanism in biological immune system, and regards object function and constraints as antigen, and antibody as feasible solution to the problem. ECOGA algorithm is as follows.

Step 1: Produce $Size$ antibodies at random to form the initial population A_i , $i = 0$;

Step 2: Evaluate the fitness of every antibody;

Step 3: Clone A_i to generate the population $A_i^{'}$;

Step 4: Do crossover and mutation operations in $A_i^{'}$;

Step 5: Compute individual fitness after step 4. If the fitness of individual after step 4 is larger than the old one, then substitute the old one with it;

Step 6: Select k antibodies with large fitness in the population to clone, and clonal probability of antibody individual relates to its fitness;

Step 7: $i = i + 1$;Check if the end condition is satisfied or not. If it is satisfied, end the algorithm, otherwise turn to Step 2 to begin the next iteration.

V. EXPERIMENTS

In this section, we present some experiments that have been carried out to test the efficiency and effectiveness of the proposed approach. In all the experiments conducted in this study, the ECOGA process started with a population of 50 individuals and the maximum number of generations has been fixed at 300. We set the population size, crossover rate, and mutation rate of ECOGA search for our experiments. We use 100 organisms in the population and set the crossover at 0.75 and mutation rate at 0.15.

Experimental results and comparisons the F-APACS and GA-miner with our proposed algorithm. Figs. 1-7 are

the curves that show the number of fuzzy association rules found according to three different methods for different values of minimum support when the number of fuzzy sets from four to eight . Figs. 8-14 are the curves that show the number of fuzzy association rules found according to three different methods for different values of minimum confidence when the number of fuzzy sets from four to eight . Figs. 15-23 shows the execute time required by each of the three algorithms to find the required fuzzy sets from two to ten per quantitative attribute.

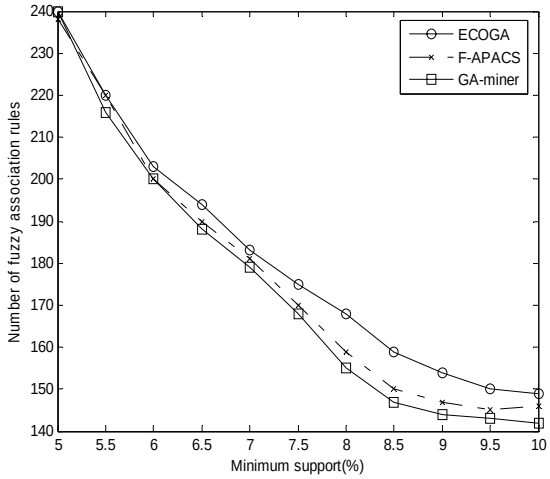

Fig. 1. Number of fuzzy association rules for different values of minimum support for four fuzzy sets

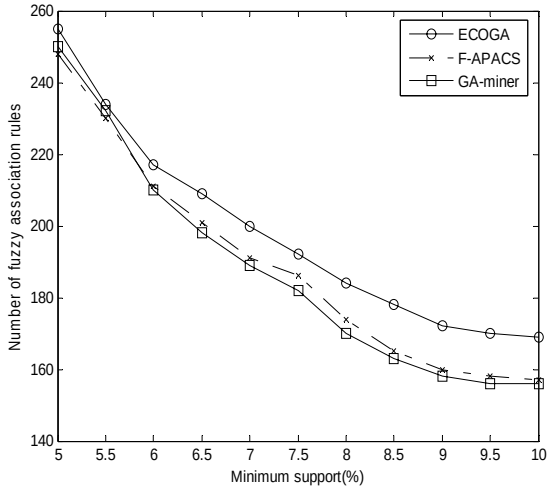

Fig. 2. Number of fuzzy association rules for different values of minimum support for five fuzzy sets

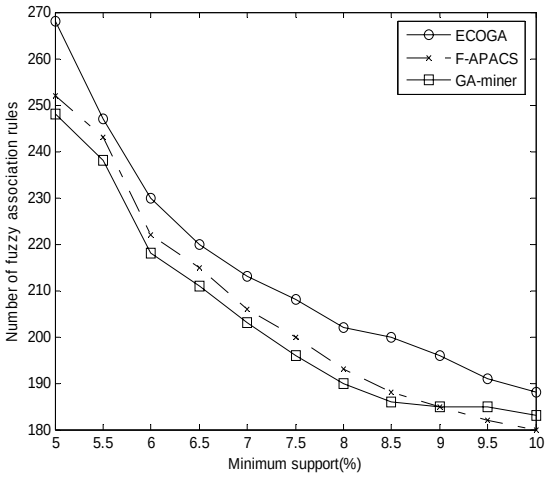

Fig. 3. Number of fuzzy association rules for different values of minimum support for six fuzzy sets

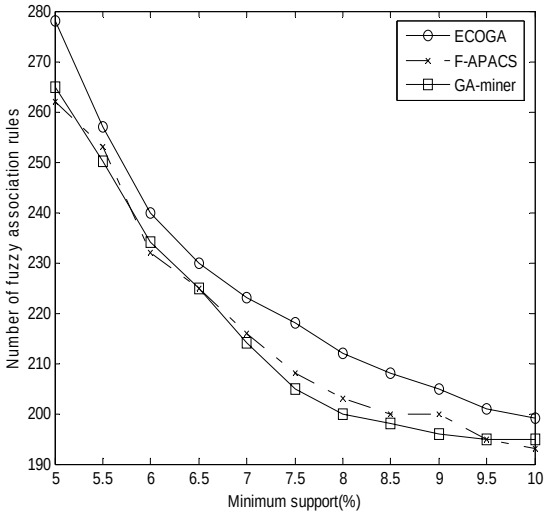

Fig. 4. Number of fuzzy association rules for different values of minimum support for seven fuzzy sets

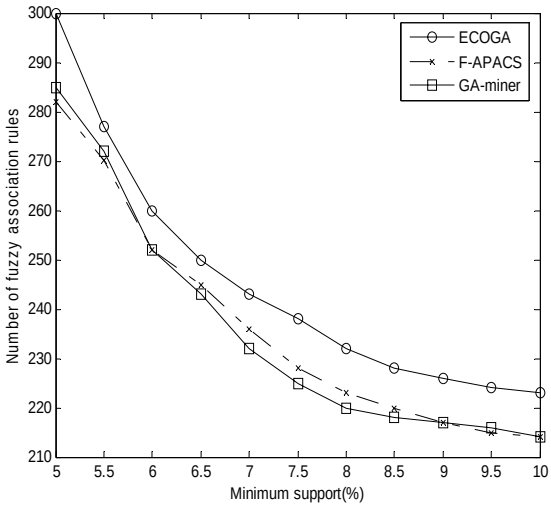

Fig. 5. Number of fuzzy association rules for different values of minimum support for eight fuzzy sets

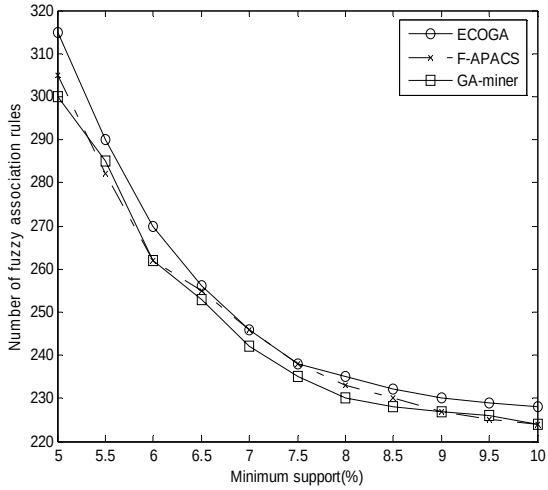

Fig. 6. Number of fuzzy association rules for different values of minimum support for nine fuzzy sets

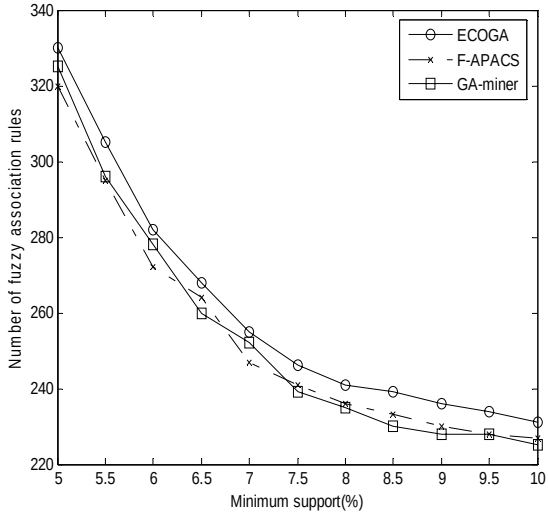

Fig. 7. Number of fuzzy association rules for different values of minimum support for ten fuzzy sets

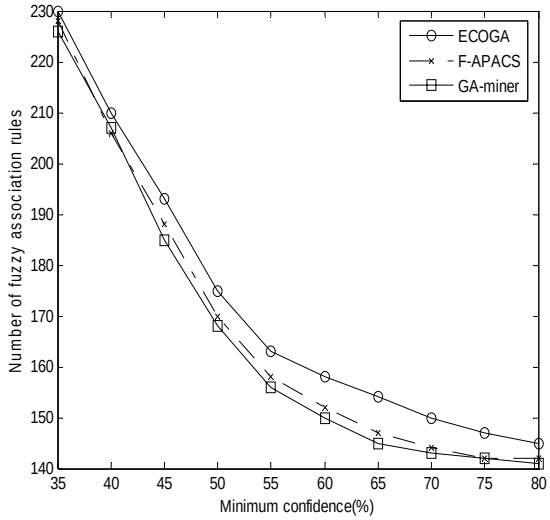

Fig. 8. Number of fuzzy association rules for different minimum confidence for four fuzzy sets

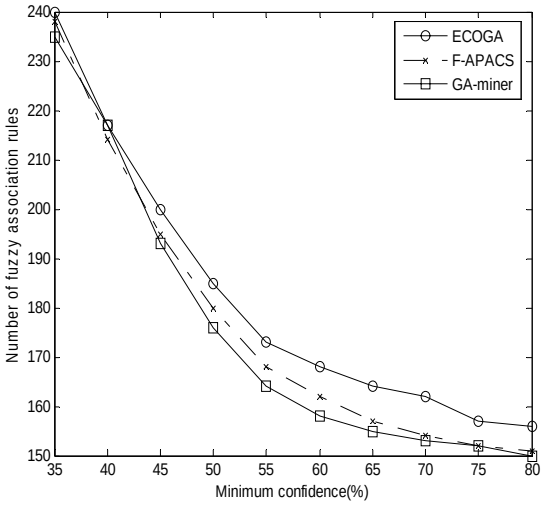

Fig. 9. Number of fuzzy association rules for different minimum confidence for five fuzzy sets

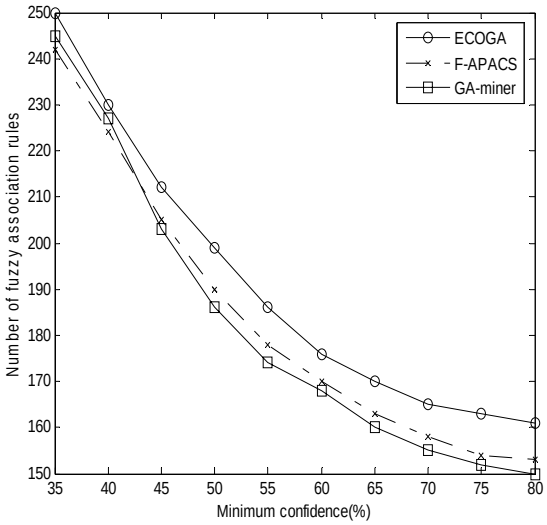

Fig. 10. Number of fuzzy association rules for different minimum confidence for six fuzzy sets

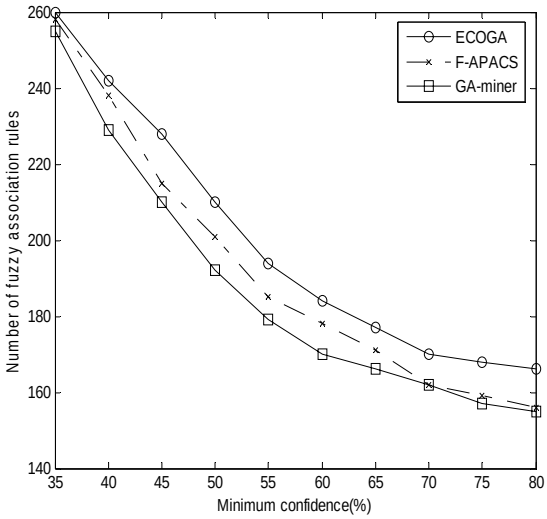

Fig. 11. Number of fuzzy association rules for different minimum confidence for seven fuzzy sets

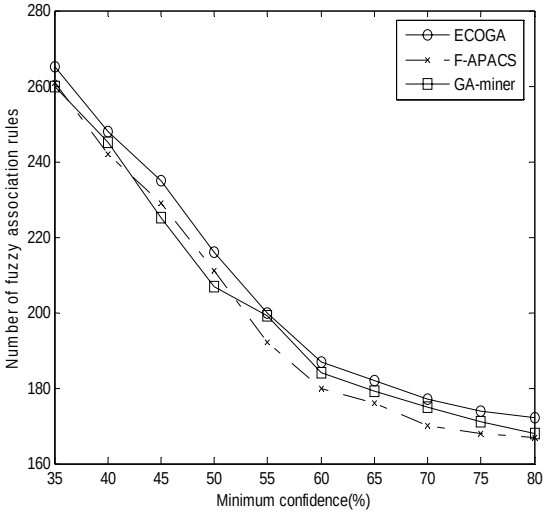

Fig. 12. Number of fuzzy association rules for different minimum confidence for eight fuzzy sets

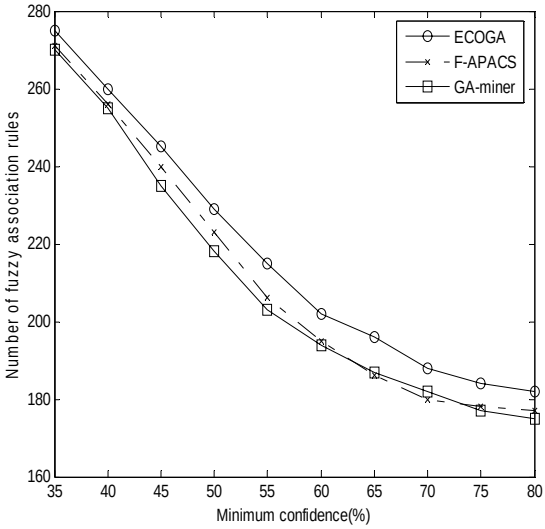

Fig. 13. Number of fuzzy association rules for different minimum confidence for nine fuzzy sets

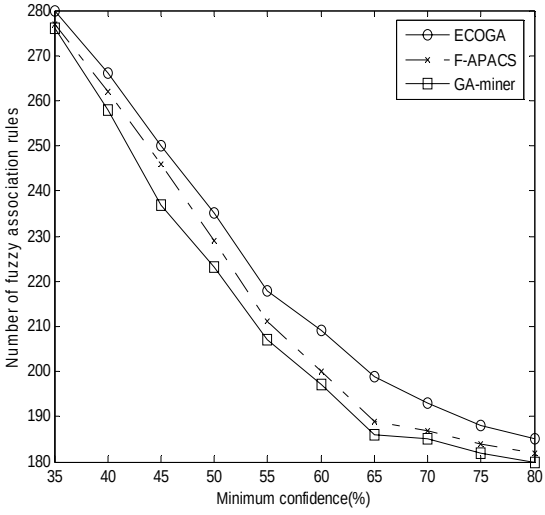

Fig. 14. Number of fuzzy association rules for different minimum confidence for nine fuzzy sets

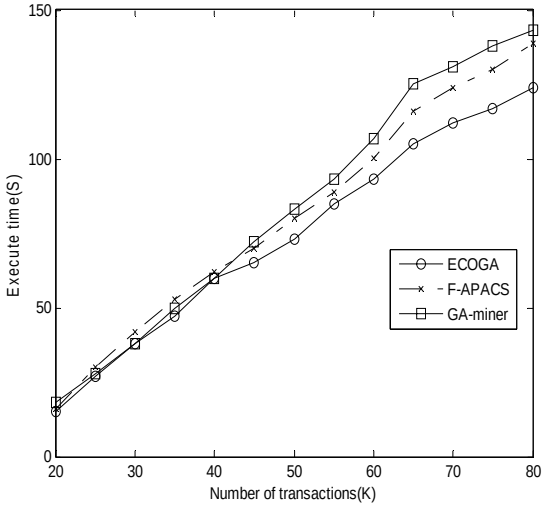

Fig. 15. The execute time required to find all large itemsets for two fuzzy sets for four fuzzy sets

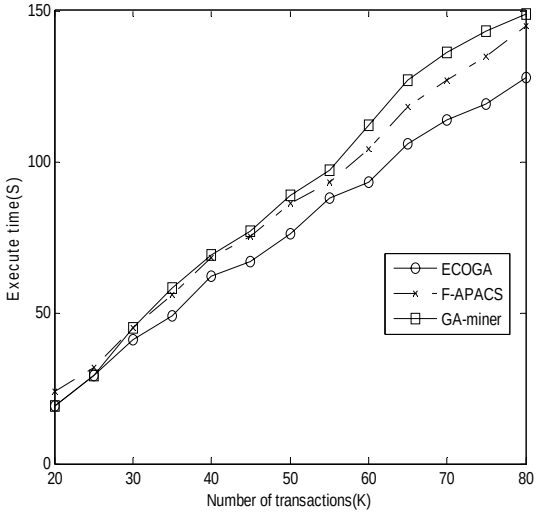

Fig. 16. The execute time required to find all large itemsets for three fuzzy sets

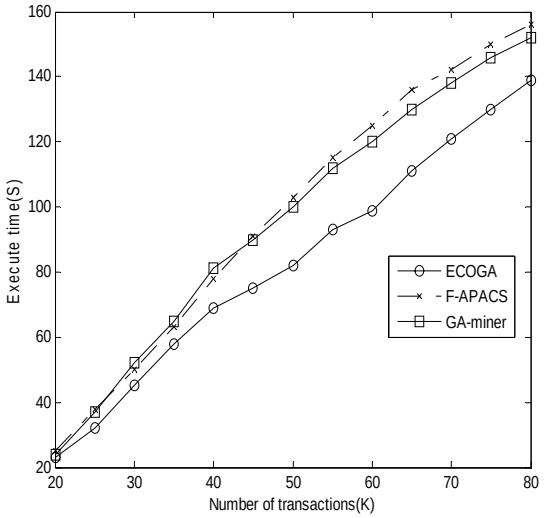

Fig. 17. The execute time required to find all large itemsets for four fuzzy sets

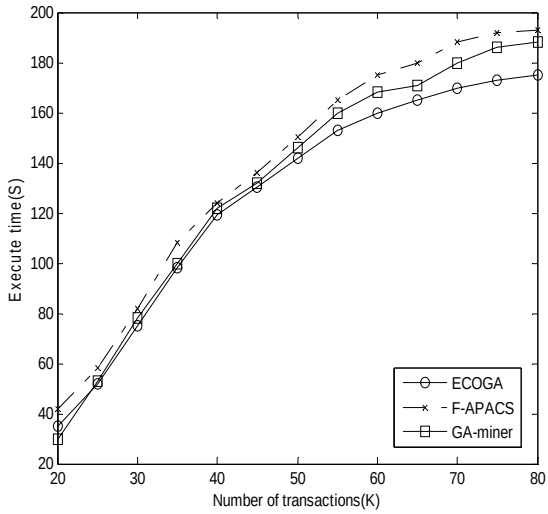

Fig. 18. The execute time required to find all large itemsets for five fuzzy sets

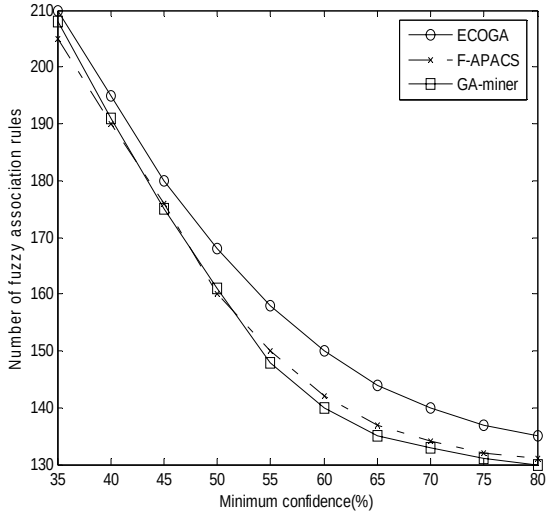

Fig. 19. The execute time required to find all large itemsets for six fuzzy sets

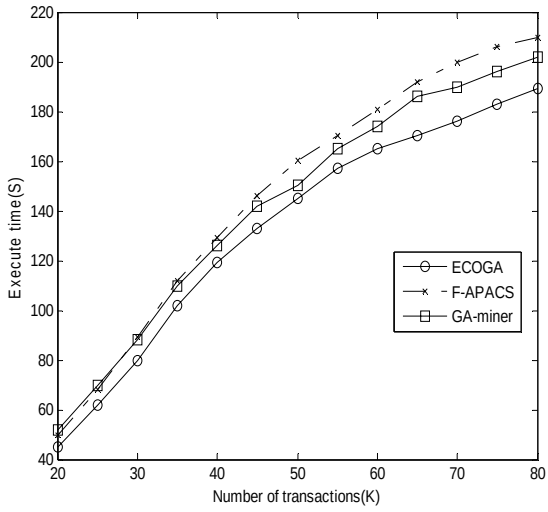

Fig. 20. The execute time required to find all large itemsets for seven fuzzy sets

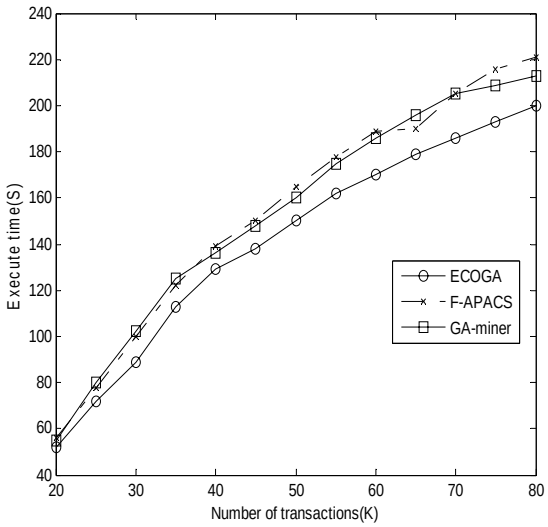

Fig. 21. The execute time required to find all large itemsets for eight fuzzy sets

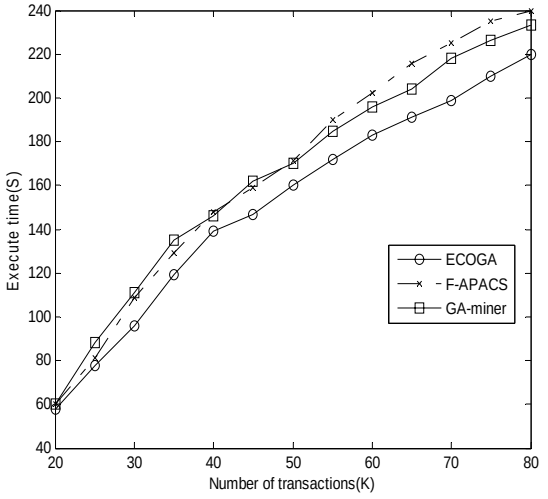

Fig. 22. The execute time required to find all large itemsets for nine fuzzy sets

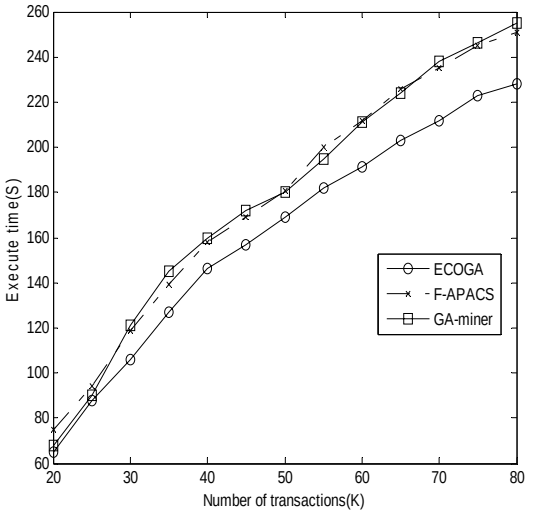

Fig. 23. The execute time required to find all large itemsets for ten fuzzy sets

From Figs. 1-7, we can conclude that the ECOGA yields the best results compared to the other examined approaches. As can be seen easily from Figs. 8-14, ECOGA outperforms other algorithms, as runtime is concerned. Figs. 15-23 demonstrate, using the proposed method results in more rules than the F-APACS and GA-miner algorithm. From the analysis and experimental result, it can be concluded that our method is superior to traditional F-APACS and GA-miner methods in data mining.

VI. Conclusions

Data mining is the process of discovering interesting and previously unknown patterns in data sets. In this paper, we put forwards an efficient clonal optimizing genetic algorithm, which is based on fuzzy set theory and genetic algorithms. One of the most important advantages of EOCGA-based method is that, as apart from other methods, our approach depends on a minimum support value given beforehand. So it is possible to obtain more appropriate solutions by changing the minimum support value. Also, the number of interesting rules obtained with the GA-based approach is larger than those obtained by applying other methods. In the future, we will investigate the possibility of reaching closer optimum by combining EOCGA with local search techniques.

Acknowledgments

This research work was supported by National Natural Science Foundation of China (Grant No. 60975050 and 60902053), the Research Fund for the Doctoral Program of Higher Education (Grant No.20070486081), the Fundamental Research Funds for the Central Universities(Grant No.6081014），the Natural Science Foundation of South-Central University for Nationalities (Grant No. YZY10004).

References

[1] D.Dubois, E.Hullermeier,H. Prade, A systematic approach to the assessment of fuzzy association rules, Data Mining and Knowledge Discovery, 2006,13 (2):167-192.

[2] M. Setnes, H. Roubos, GA-based modeling and classification: complexity and performance, IEEE Trans. Fuzzy Systems 2000, 8 (5):509–522.

[3] Y. Lee, T. Hong, T. Wang, Multi-level fuzzy mining with multiple minimum supports, Expert Systems Appl. 2008, 34 (1):459-468.

[4] M. Kaya, Multi-objective genetic algorithm based approaches for mining optimized fuzzy association rules, Soft Comput. 2006, 10 (7):578-586.

[5] E. Hullermeier, Y. Yi, In defense of fuzzy association analysis, IEEE Trans. Systems, Man, Cybernet.-Part B: Cybernet. 2007, 37 (4):1039-1043.

[6] L. Castillo, A. Gonzalez, P. Perez, Including a simplicity criterion in the selection of the best rule in a genetic fuzzy learning algorithm, Fuzzy Sets and Systems 2001,120 (2) :309–321.

[7] O. Cordón, F. Gomide, F. Herrera, F. Hoffmann, L. Magdalena, Ten years of genetic fuzzy systems: current framework and new trends, Fuzzy Sets and Systems ,2004,141 (1) :5–31.

[8] A. Jaszkiewicz, Genetic local search for multi-objective combinatorial optimization, European J. Oper. Res. 2002,137 (1):50-71.

[9] Y. Lee, T. Hong,W. Lin, Mining fuzzy association rules with multiple minimum supports using maximum constraints, in: KES, Lecture Notes in Computer Science, Springer, Berlin, 2004, Vol. 3214 ,pp. 1283–1290.

[10] T. Hong, Y. Lee, An overview of mining fuzzy association rules, in: H. Bustince, F. Herrera, J. Montero (Eds.), Studies in Fuzziness and Soft Computing, Springer, Berlin, Heidelberg, 2008, Vol. 220, pp. 397–410.

[11] T. Hong, C. Chen, Y. Wu, Y. Lee, A GA-based fuzzy mining approach to achieve a trade-off between number of rules and suitability of membership functions, Soft Comput. 2006, 10 (11): 1091–1101.

[12] Sankar K.Pal.Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions [J]. IEEE Transaction on neural networks, 2002, 13(5):1012-1024.

[13] H. Ishibuchi, T. Nakashima, effect of rule weights in fuzzy rule-based classification systems, IEEE Trans. Fuzzy Systems ,2001,9 (4) :506-515.

[14] W. Pedrycz, Fuzzy sets technology in knowledge discovery, Fuzzy Sets and Systems, 1998, pp.279–290.

[15] W.H.Au, K.C.C. Chan, An effective algorithm for discovering fuzzy rules in relational databases, in: Proc. IEEE Internat. Conf. Fuzzy Systems, 1998, pp. 1314–1319.

[16] Hisao Ishibuchi,Takashi Yamamoto. Fuzzy rule selection by multi-objective Genetic local search algorithm and rule evaluation measures in data mining [J].Fuzzy Sets and Systems, 2004, 141(2):59-88.

[17] Riyaz,s.,Selwyn P. Efficient genetic algorithm based data mining using feature selection with hausdorff distance[J], Information Technology and Management,2005,6(4):315-331

[18] R. Alcalá, J. Alcalá-Fdez, F. Herrera, A proposal for the genetic lateral tuning of linguistic fuzzy systems and its interaction with rule selection,IEEE Trans. Fuzzy Systems ,2007,15 (4) ;616–635.

[19] H.-M. Lee, S.-K. Lin, C.-W. Huang, Interactive query expansion based on fuzzy association thesaurus for web information retrieval, in: Proc. the 10th IEEE Internat. Conf. on Fuzzy Systems, 2001, pp. 2:724 – 2:727.

[20] M. Kaya, R. Alhajj, Genetic algorithm based framework for mining fuzzy association rules, Fuzzy Sets and Systems, 2005,152 (3):587-601.

[21] LI Yang-Yang, JIAO Li-Cheng. Quantum-Inspired Immune Clonal Algorithm for SAT Problem [J].Chinese Journal of Computers, 2007, 30(2): 176-183

[22] L.C.Jiao, L.Wang. A Novel Genetic Algorithm Based on Immunity [J].IEEE Transaction on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2000, 30(5):552-561.

[23] Zhong Wei-Cai, Liu Jing, Xue Ming-Zhi, Jiao Li-Cheng. A Multi-Agent Genetic Algorithm for Global Numerical Optimization. IEEE Trans.System,Man and Cybernetics—Part B. 34(2), (2004)1128-1141

[24] Jiao Licheng, Du Haifeng, “An Artificial Immune System: Pogress and Prospect”, ACTA ELECTRONICA SINICA, 31(10), (2003) 1540-1549

[25] Weisheng Dong,Guangming Shi,Li Zhang. Immune memory clonal selection algorithms for designing stack filters [J].Neurocomputing, 2007, 70(3):777-784.

Wanneng Shu (1981-), male. He is PHD Candidate in the State key lab of software engineering, Wuhan University, China. He is currently a lecturer in the College of Computer Science, South-Central University for Nationalities. His research interests include data mining, and intelligence computing.

Lixin Ding (1967-), male. He received his B.Sc. and M.Sc. degrees from the Department of Applied Mathematics, Hunan University, Changsha, China, in 1989 and 1992, respectively, and Ph.D. degree from the State Key Laboratory of Software Engineering (SKLSE), Wuhan University, Wuhan, China, in 1998. From 1998 to 2000, he was a Postdoctoral researcher at the Department of Armament Science and Technology, Naval University of Engineering, Wuhan, China. He is currently a professor at the SKLSE. His main research interests include evolutionary computation, intelligent information processing, statistical learning and cloud computing.