Improved Genetic Algorithm Based on Simulated Annealing and Quantum Computing Strategy for Mining Association Rules

Dongsheng Liu
College of Computer Science & Information Engineering
Zhejiang Gongshang University, Hangzhou 310018, China
Email: lds1118@zjgsu.edu.cn

Abstract—**Association rules mining is an important content in data mining. It can discover the relations of different attributes by analyzing and disposing data which is in database. This paper proposes a novel data mining algorithm to enhance the capability of exploring valuable information from databases with continuous values. The algorithm combines with quantum-inspired genetic algorithm and simulated annealing to find interesting association rules. The final best sets of membership functions in all the populations are then gathered together to be used for mining association rules. The experiment result demonstrates that the proposed approach could generate more association rules than other algorithms.**

Index Terms—**data mining, association rule, transaction data, quantum-inspired genetic algorithm, simulated annealing**

I. INTRODUCTION

Data mining is a form of knowledge discovery essential for solving problems in a specific domain. Individual data sets may be gathered and studied collectively for purposes other than those for which they were originally created. Data mining is most commonly used in attempts to induce association rules from transaction data [1]. It can discover the relations of different attributes by analyzing and disposing data which is in database. Most previous studies focused on binary-valued transaction data. Transaction data in real-world applications, however, usually consist of quantitative values. Designing a data mining algorithm able to deal with various types of data presents a challenge to workers in this research field.

GA can dispose large-scale data gather. It is widely applied in mining association rules since it is powerful search techniques in solving difficult problems. GA starts with some randomly selected gene combinations called chromosomes to form the first generation. Each individual chromosome in each generation corresponds to a solution in the problem domain. The fitness function is used to evaluate the quality of each chromosome such that the chromosomes with high quality will survive and from the next generation. By using the following three GA operators (selection, crossover, and mutation), a new generation is recombined to find the best solution. The process will be continuously repeated until obtaining the optimum solution, or a constant number of iterations have been performed. GA plays an important role in data mining technology, which is decided by its own characteristics and advantages [2]. However, traditional GA would come forth local convergence in search process.

Quantum-inspired genetic algorithm (QIGA) is based on the concept and principles of quantum computing such as qubits and superposition of states. By adopting qubit chromosome as a representation, QIGA can represent a linear superposition of solutions due to its probabilistic representation .However, the performance of QIGA is easy to be trapped in local optimal so as to be premature convergence. In other words, the qubit search with quantum mechanism and genetic search with evolution mechanism need to be well coordinated and the ability of exploration and exploitation need to be well balanced as well.

In this paper, parallel quantum-inspired genetic mining algorithm based on simulated annealing (SA) strategy, called PQGMA, is applied in extracting association rule for data mining. SA originations from the method of the statistical physics and first employed by Kirkpatric to solve the optimization problem. We chose PQGMA due to its simplicity and its capability as a powerful search mechanism. We take advantage of this principle in developing modified frameworks essentially using the GA at its core.

The rest of this paper is organized as follows. Related work is discussed in Section 2. The PQGMA is described in Section 3. Section 5 experiments are carried out to investigate the effectiveness of the HCSGA. Finally, section 6 concludes this paper and indicates some directions for future study.

II. RELATED WORKS

Recently, some research works have been done on discovering association rules for dealing continuous attributes. Associations allow capturing all possible rules

This study is supported by the Natural Science Foundation of Zhejiang Province (Y1090617, Z1091224); Zhejiang Science and Technology Plan Project (2009C13G2050020)

doi:10.4304/jsw.5.11.1243-1249

that explain the presence of some items according to the presence of other items in the same transaction.

In [3] illustrated fuzzy versions of confidence and support. Gyenesei presented two different methods for mining fuzzy continuous 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. The authors generate a concept relation dictionary and a classification tree from a random set of daily business reports database of text classes concerning retailing. An approach in [4] uses fuzzy association thesaurus and query expansion for information retrieval.

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 [5].M. Kaya proposed a GA-based clustering method to derive a predefined number of membership functions for getting a maximum profit [6].In [7] proposed a GA-based fuzzy data-mining method for extracting both association rules and membership functions from quantitative transactions.

Analysis of real-world data in data mining often necessitates simultaneous dealing with different types of variables, viz., categorical/symbolic data and numerical data. Nauck [8] has developed a learning algorithm that creates mixed fuzzy rules involving both categorical and numeric attributes. However, quantity is a very useful piece of information. Realizing the importance of quantity, people started to concentrate on quantitative attributes. The support for any particular value is likely to be low, while the support for intervals is much higher. Srikant and Agrawal used equi-depth partitioning to mine quantitative rules [9]. They separate intervals by their relative ordering and quantities equally. Hong et al. proposed an algorithm that integrates fuzzy set concepts and Apriori mining algorithm to find interesting fuzzy association rules from given transactional data [10]. In [11] employed these measures of fuzzy rules for function approximation and pattern classification problems. Gyenesei presented two different methods for mining fuzzy quantitative association rules, namely without normalization and with normalization. In [12] 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. In [13] presented a geometric-based algorithm, called BitOP, to perform clustering for numerical attributes. They showed that clustering is a possible solution to figure out meaningful regions and support the discovery of association rules. 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. The approach developed by Zhang extends the equi-depth partitioning with fuzzy terms [14]. However, it assumes fuzzy terms as predefined. However, the specified fuzzy linguistic terms in fuzzy association rules can be given only when the properties of the attributes are estimated. In real life, contents of columns may be unknown and meaningful intervals are usually not concise and crisp enough.

III. PQGMA ALGORITHM FOR EXTRACTING ASSOCIATION RULE

In this section, the proposed PQGMA algorithm for mining association rules is described. Association rules show attributes that occur frequently together in a given dataset. The following is a formal statement of the problem of mining association rules. An association rule is an expression $X \Rightarrow Y$, 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. $AS = \{I, T, Sup, Conf, X, Y,\}$ represents the association rules in data mining, in which $I = \{i_1, i_2, ..., i_k, ..., i_n\}$ be a set of literals, called items or attributes, and $T = \{t_1, t_2, ...t_k, ..., t_m\}$ be a set of transactions, where each transaction $t_k (1 \le k \le n)$ is a set of items such that $t_k \subseteq I$. Associated with each transaction is a unique identifier. A transaction T contains X , a set of some items in I , if $X \subseteq T$. An association rule is an implication of the form of $X \Rightarrow Y$, where $X \subset I, Y \subset I$,and $X \cap Y = \varnothing$. X is called antecedent and Y is called consequent of the rule. In general, a set of items is called an itemset.

Each itemset has an associated measure of statistical significance called support. The support of the rule is defined as the percentage of the records having both attributes X and Y ,namely $Sup(X \Rightarrow Y) = \frac{|\{D:X \cup Y \subseteq T, D \in T\}|}{|T|}$. The rule $X \Rightarrow Y$ has a measure of its strength called confidence. The confidence of the rule is defined as the percentage of the records having Y given that they also have X , namely $Conf(X \Rightarrow Y) = \frac{|\{D: X \cup Y \subseteq T, D \in T\}|}{|D: X \subseteq T, D \in T|}$.If support degree and confidence degree exceed each minimum value, then $X \Rightarrow Y$ can be regarded as significant association rules in T .

PQGMA starts with an initial population of randomly or heuristically generated individuals, and advances toward better individuals by applying genetic operators modeled on the genetic processes occurring in nature. The population undergoes evolution in a form of natural selection. During successive iterations, called generations, individuals in the population are rated for their adaptation as solutions, and on the basis of these evaluations. As a result, a new population of individuals is formed using a selection mechanism and specific genetic operators such as crossover and mutation. To form a new population, individuals are selected according to their fitness. Consequently, an evaluation or fitness function must be devised for each problem to be solved. Given a particular individual, a possible solution, the fitness function returns a single numerical fitness, which is supposed to be proportional to the utility or adaptation of the solution represented by that individual. The genetic algorithm uses crossover and mutation operators to generate the offspring of the existing population. Before genetic operators are applied, parents should be selected for evolution to the next generation. An individual is additionally selected to initialize offspring to be generated in the other half of the population, not in the part of the candidates as parents.

A. Chromosome

It is important to encode membership functions as string representation for PQGMA to be applied. A PQGMA requires a population of feasible solutions to be initialized and updated during the evolution process. Each individual within a population is a possible set of triangular membership functions for an item. The initial set of chromosomes in a population is randomly generated within some constraints for forming feasible membership functions.

B. Fitness Function

This paper adopts support and confidence for filtering rules. Then correlative degree is confirmed in rules which satisfy minimum support degree and minimum confidence degree. The fitness function is defined as follows：

$$f(X \Rightarrow Y) = exp(\frac{Sup(X \Rightarrow Y) + Conf(X \Rightarrow Y)}{2})$$

C. Selection

Selection is the transmission of personal information from the parent individual to the offspring individuals. We used “Stochastic Universal Sampling” [15] as our selection method. A form of stochastic universal sampling is implemented by obtaining a cumulative sum of the fitness vector. Thus, only one random number is generated, all the others used being equally spaced from that point. The index of the individuals selected is determined by comparing the generated numbers with the cumulative sum vector. The probability of an individual being selected is then given by

$$P_s(x_i) = f(x_i) / \sum_{i=1}^{PopSize} f(x_i)$$

where $f(x_i)$ is the fitness of individual x_i and $P_s(x_i)$ is the probability of that individual being selected. $PopSize$ is the size of the population.

D. Crossover

The crossover operator usually operates on two individuals/parents, to produce two children. Common crossover methods that are readily used for binary gene include one-point crossover, two-point crossover, and uniform crossover. We are adopted uniform crossover strategy, which exchanges can occur at any position on the chromosome [16]. For example, if we have two parents V_1 (1101001) and V_2 (1011111), then to produce two offspring ($V_1^{'}$ and $V_2^{'}$) .We go through each of the seven genes of one of the parent. For each gene, there is a mutation probability p_c (e.g. p_c =0.9) that this gene will be exchanged with the other parent. As such, if an exchange happens at position 2 and 4, then the two offspring produced will be 1110101 and 1001011.

After crossover operation, computing the fitness value $f(V_1^{'})$and $f(V_2^{'})$,if min $\{1, exp(-(f(V^{'}) - f(V) / T_k)\}$>random[0,1], it accepts the two offspring($V_1^{'}$ and $V_2^{'}$). T_k is the evolution temperature of the kth time.

E. Mutation

Mutation operator works on each individual offspring. The mutation operator helps prevent early convergence of the QIGA by changing characteristics of chromosomes in the population [17]. In PQGMA, a quantum gate $U(\theta)$ is employed to update a qubit individual as a variation operator.

$$\begin{pmatrix} \alpha_i^{'} \\ \beta_i^{'} \end{pmatrix} = U(\theta) \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \cos(\theta_i) & -\sin(\theta_i) \\ \sin(\theta_i) & \cos(\theta_i) \end{pmatrix} \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix}$$

F. The solution process of PQGMA

The solution process of PQGMA as follows:

Step 1: Randomly generate $PopSize$ binary strings (chromosomes) for the initial population $Pop(k)$, each for an item; each individual in a population represents a possible set of membership functions for that items, Initialize the maximum evolution iteration $MaxGen$, the number of items m , the number of transaction data of n , the minimum support γ , a minimum confidence τ ,evolution algebra $k=0$;

Step 2: Evaluate the fitness value of each chromosome in each population

Step 3: Execute selection operation for $Pop(k)$ so as to generate $Pop(k)^{'}$;

Step 4: Execute crossover operation $Pop(k)^{'}$ so as to generate $Pop(k)^{''}$;

Step 5: Update $Pop(k)^{''}$ to generate $Pop(k+1)$ using quantum rotation gates;

Step 6: $k = k+1$ .if $k \geq MaxGen$, output the set of association rules with its associated set of membership functions; otherwise, go back to step 2.

Ⅳ. SIMULATION EXPERIMENT AND RESULT ANALYSIS

In order to demonstrate the application of the proposed algorithm, experiments have been carried out. We adjusted the parameters of the HCSGA by experiments, and finally selected the following combination of the parameters: The initial population size $PopSize$ was set at 100, the maximum number of generations $MaxGen = 500$, the mutation probabilities $P_m = 0.15$.

In this section, we present our simulation results for the comparison of PQGMA with fuzzy sets, QIGA and GA. Fig. 1-3 show the runtime required by each of the three algorithms to find the required ten, fifteen and fifty association rules. As can be seen easily from Fig. 1-3, PQGMA outperforms other algorithms, as runtime is concerned. Fig. 4 is the curves that show the number of association rules found according to three different methods for different values of minimum support. In this experiment, the minimum confidence was set to 38% and the number of association rules is ten for both GA and PQGMA approaches as well as for fuzzy sets and the number of fuzzy sets is also ten. Here, another important point is that the difference between the curves of PQGMA and GA increases in favor of PQGMA. This is quite consistent with our intuition since a large number of transaction sets will make quantities of an item in different transactions easily scattered in difference sets. It can be observed from Fig. 4 that the number of rules decreased along with the increase of the minimum support values. Besides, the curve with a large minimum confidence value was smoother than those with a small value.

Fig. 5 investigates the number of association rules for different values of minimum confidence. It can be observed from Fig. 5 that the number of association rules decreased along with the increase of the minimum confidence values. Fig. 6 shows the number of association rules for three algorithms with different generations. It can be observed from Fig. 6 that the proposed approach could generate more association rules than other algorithms. As the generation continues, further improvement is found in average population fitness as demonstrated in Fig. 7. The offspring are generated by the average of the parents which have higher fitness. As the generation progresses further, the increment of fitness falls down a little because most of the chromosomes are already converged to the high fitness. From Fig.7, we can see that PQGMA which is based on evolution strategy can not only conquer getting in state of local convergence, but also accelerate search speed.

The precision for four mining algorithms are shown in Fig. 8-10 .The recall for four mining algorithms are shown in Fig. 11-12. Analyzing the precision and recall show in Fig. 8-12, we see that on average, the PQGMA algorithm obtain a higher accuracy value than other three mining algorithms.

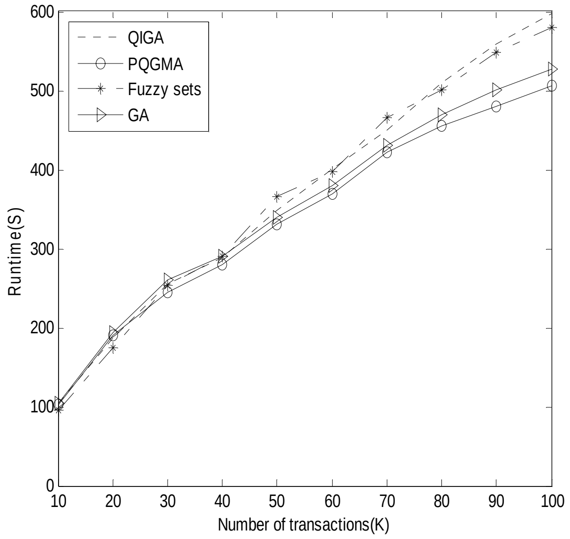

Fig.1. The runtime for four mining algorithms to find 10 association rules

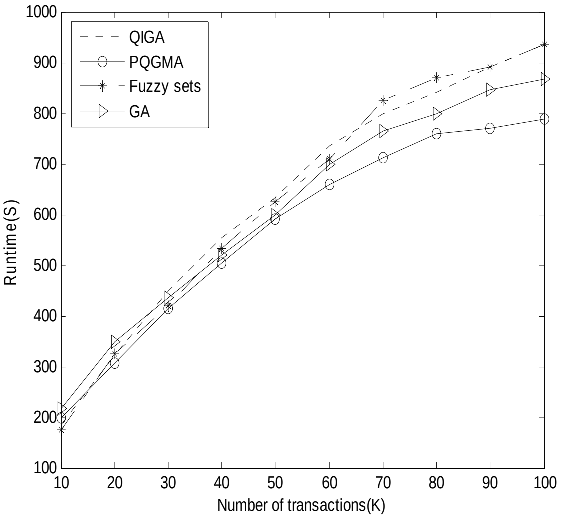

Fig.2. The runtime for four mining algorithms to find 15 association rules

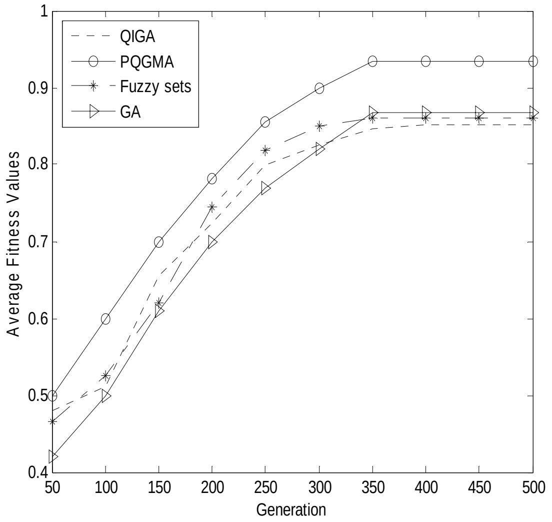

Fig.3. The runtime for four mining algorithms to find 50 association rules

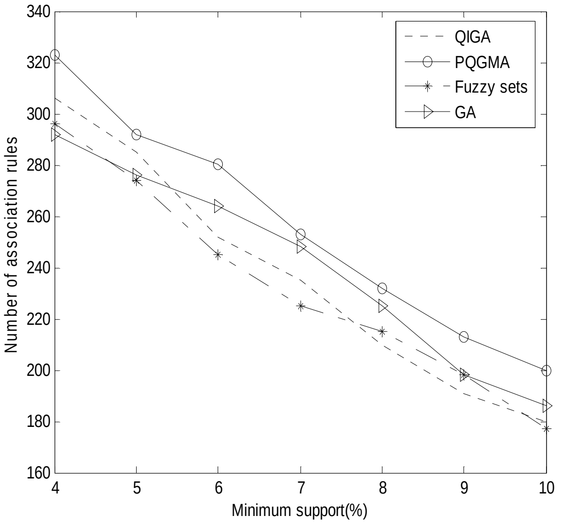

Fig.4. Number of association rules for different minimum support values.

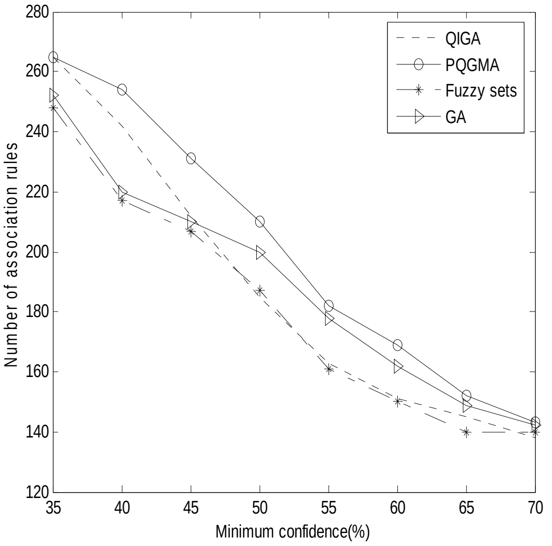

Fig.5. Number of association rules for different minimum confidence values.

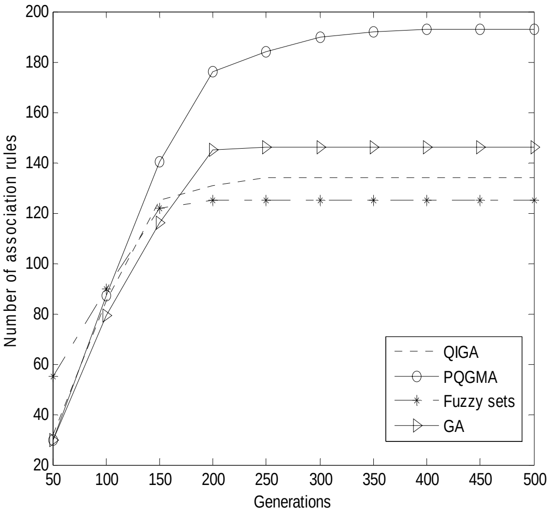

Fig.6. Number of association rules for different generations

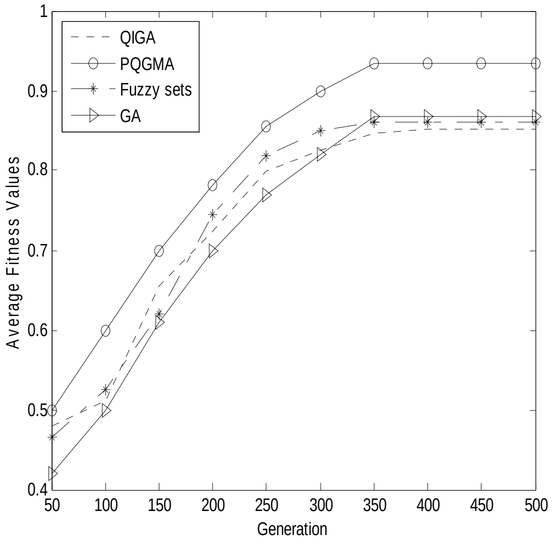

Fig.7. Number of association rules for different generations

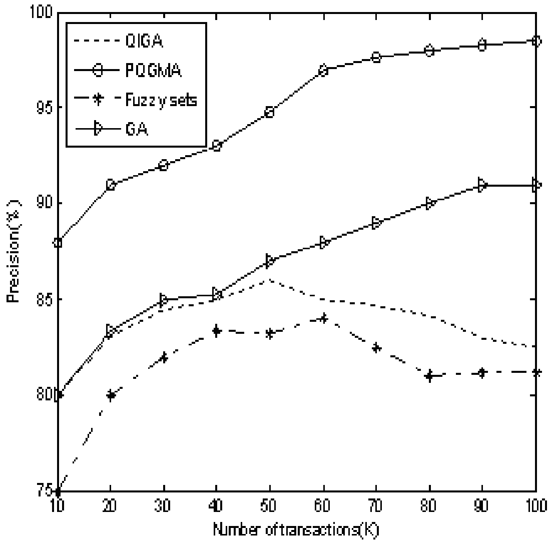

Fig.8. The precision for four mining algorithms to find 10 association rules

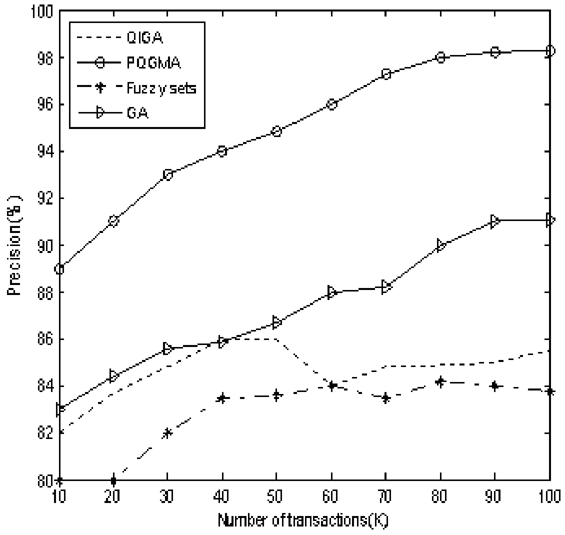

Fig.9. The precision for four mining algorithms to find 15 association rules

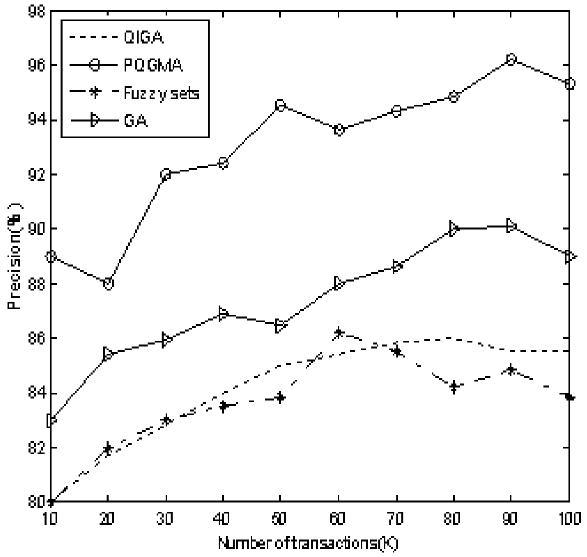

Fig.10. The precision for four mining algorithms to find 50 association rules

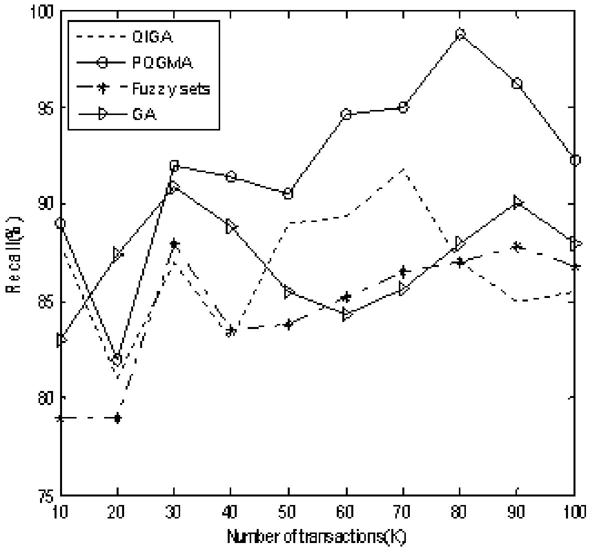

Fig.11. The Recall for four mining algorithms to find 15 association rules

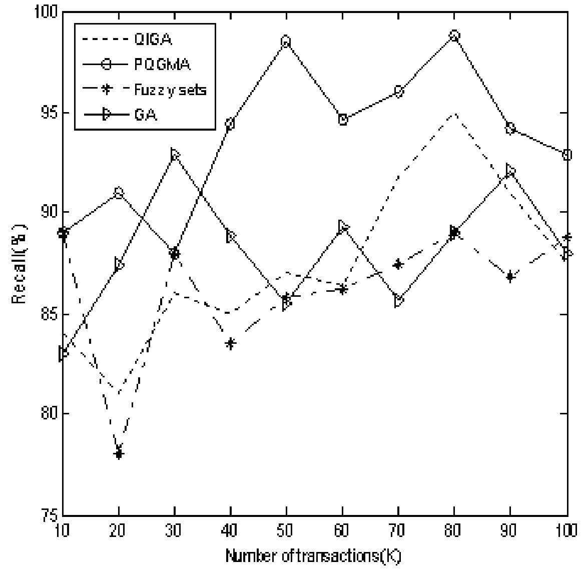

Fig.12. The Recall for four mining algorithms to find 50 association rules

Ⅴ.CONCLUSIONS

Data mining is most commonly used in attempts to induce association rules from transaction data [18]. In this paper, we have presented PQGMA algorithm to apply in mining association rules. From the experiment result, we can concluded that One of the most important advantages of PQGMA-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 PQGMA-based approach is larger than those obtained by applying other methods.

ACKNOWLEDGMENT

Dongsheng Liu wish to thank the Natural Science Foundation of Zhejiang Province (Y1090617, Z1091224); Zhejiang Science and Technology Plan Project (2009C13G2050020) and Modern business centre of Zhejiang Gongshang University for support.

REFERENCES

[1] D. A. Chiang, L. R. Chow, and Y. F. Wang. Mining time series data by a fuzzy linguistic summary system[J] .Fuzzy Sets System, 2000, pp: 419–432.

[2] L.Wang, Intelligent Optimization Algorithm with Application[C], Tsinghua University & Springer Press, Beijing, 2001.

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

[4] 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 International Conf. on Fuzzy Systems, 2001, pp. 2:724 – 2:727.

[5] I.W. Flockart,N.J.Radcliffe. GA-miner: parallel Data Mining with Hierarchical genetic algorithm, EPCC-AIKMS-GA-Miner-Report 1.0, University of Edimburgh, 1995.

[6] M. Kaya and R. Alhajj, Genetic algorithm based framework for mining fuzzy association rules[J], Elsevier, 2004.

[7] T. P. Hong, C. H. Chen, Y. L. Wu and Y. C. Le, "Mining membership functions and fuzzy association rules", The 2003 Joint Conference on AI, Fuzzy System, and Grey System, 2003.

[8] D. Nauck, "Using symbolic data in neuro-fuzzy classification," in Proc. NAFIPS 99, New York, June 1999, pp. 536–540.

[9] R. Srikant, R. Agrawal, Mining quantitative association rules in large relational tables, in: Procedure. ACM SIGMOD International Conference Management of Data, 1996, pp. 1–12.

[10] T.P. Hong, C.S. Kuo, S.C. Chi,A fuzzy data mining algorithm for quantitative values, in: Procedure International Conference Knowledge Based Intelligent Information Engineering Systems, 1999, pp. 480–483.

[11] R.J. Miller, Y. Yang, Association rules over interval data, in: Proc. ACM SIGMOD International Conference Management of Data,1997, pp. 452–461.

[12] A. Gyenesei, Mining weighted association rules for fuzzy quantitative items, TUCS Technical Report No: 346, May 2000.

[13] W.H. Au, K.C.C. Chan, An effective algorithm for discovering fuzzy rules in relational databases, in:

Procedure IEEE International Conference Fuzzy Systems, 1998, pp. 1314–1319.

[14] W. Zhang, Mining fuzzy quantitative association rules, Proc. IEEE International Conference Tools Artificial Intelligent, 1999, 99–102.

[15] Baker, J. E. Reducing bias and inefficiency in the selection algorithm, Proceeding ICGA 2, Lawrence Erlbuam Associates, Publishers, 1987. pp:14-21

[16] H. Ishibuchi, T. Nakashima, T.Yamamoto, Fuzzy association rules for handling continuous attributes, in: Proc. IEEE ISIE, 2001, pp. 118–121.

[17] C. Zhang and S. Zhang. Association Rule Mining: models and algorithms[C]. Springer, Sydney, Australia, 2002, pp. 238-251.

[18] J. McCarthy, Phenomenal data mining, association for computing machinery, Communications of the ACM ,2000,43 (8): 75－80.

Dongsheng Liu received his PHD degree in the school of information engineering, in 2008, Zhejiang Gongshang University (ZJGSU), China. He is currently an associate Professor in the school of information engineering at ZJGSU. His research interests include data mining, electronic commerce and wireless network.