Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 10(1): 40-44, Mar. 2012 **Regular Paper**

Gene Algorithm of Crowd System of Data Mining

Jong-Min Park*, ***Member, KIICE***

Department of Cyber Security, Chosun College of Science & Technology, Gwangju 501-744, Korea

Abstract

Data mining, which is attracting public attention, is a process of drawing out knowledge from a large mass of data. The key technique in data mining is the ability to maximize the similarity in a group and minimize the similarity between groups. Since grouping in data mining deals with a large mass of data, it lessens the amount of time spent with the source data, and grouping techniques that shrink the quantity of the data form to which the algorithm is subjected are actively used. The current grouping algorithm is highly sensitive to static and reacts to local minima. The number of groups has to be stated depending on the initialization value. In this paper we propose a gene algorithm that automatically decides on the number of grouping algorithms. We will try to find the optimal group of the fittest function, and finally apply it to a data mining problem that deals with a large mass of data.

Index Terms: Algorithm, Crowd, Data mining, Gene

I. INTRODUCTION

Data mining is a process of mining out the useful and suggestive information from a large mass of data. Mining out the data means to pick out the meaningful pattern of information by allowing it to converge into a form of knowledge. The classification of the data is the same as grouping it. The key data source in data mining is divided into groups of data by maximizing the similarity in a given group and minimizing the similarity between groups. In order to be as efficient as possible in dealing with a large quantity of databases, the number of contacts with source data must be reduced. There is much interest in grouping techniques that shrink the quantity of the data an algorithm must process. Few studies have been systematically pursued on multimedia data mining despite the overwhelming amounts of multimedia data that have accumulated due to the development of computer capacity, storage technology, and the internet. In this paper, we propose a gene algorithm that is applied to grouping. Various functions will be tested to find the function that will lead to the best grouping by the gene algorithm [1, 2].

II. PRELIMINARIES

The gene algorithm is a model analogous to evolution of nature. That “nature selects the most suitable species or organism through survival” was a theory of Darwin. The concept of natural selection is the basis of the gene algorithm. The gene algorithm goes through such a process as detailed below.

First, an individual of a second echelon string that has a fixed length is generated to help deal with bits of information comfortably by encoding them, and an initial value must be assigned to the population. Second, the fidelity of all individuals in a hypothetical population should be calculated by manipulating such factors as hybridization fidelity,

Received 13 January 2012, **Revised** 12 February 2012, **Accepted** 27 February 2012
***Corresponding Author** E-mail: jmpark@cst.ac.kr

Open Access **http://dx.doi.org/10.6109/jicce.2012.10.1.040** print ISSN:2234-8255 online ISSN:2234-8883

comparison regeneration, and mutation in an individual. Lastly, the previous hypothetical population can be ignored and all of the above factors tested with the newly selected population. A gene algorithm, which is a method for acquiring the needed probability, can be extensively applied in a complex problem that is difficult to model and can even be lined up. In this study, the gene algorithm is applied to grouping. There is special emphasis on the fitness function that has a great impact on the ability of the gene algorithm. The fitness function is a measure of the fidelity of an individual. It could be called an object function because it plays a role in the environment that evaluates the probable solutions. The importance of a fitness function came out as it was used in various grouping evaluation functions. The concept of cohesion and separation was exercised in the grouping in a fitness function. Due to the two evaluation functions, the separation of the fitness function could be made by determining their grouping ability.

Some gene algorithms that are used in current data mining are the primitive gene algorithm, messy genetic algorithm, genetic programming, parallel genetic algorithm, GABIL, GA-IDE, GIGAR, and LONGEPRO. These are used in a wide range of data mining tasks, each with their unique characteristics.

A. NP-complete

The square root modulo *n* (SQROOT) problem is to find a square root of *a* modulo*n* for the given composite integer *n* and quadratic residue *a* modulo *n*. If the factors *p* and *q* are known, then the SQROOT problem can be solved in polynomial time. If the factors *p* and *q* are unknown, then the factoring problem of *n* is reduced to the SQROOT problem in polynomial time, and the factoring problem of *n* is NP−complete [3].

B. Gene Algorithm

Table 1. compares the gene algorithms.

Analysis content	Gene algorithm		
	Messy genetic algorithm	Parallel genetic algorithm	LONGEPRO
Last generation's incidental and analysis price of value	Good	Good	Good
Calculation Cost	Expensive	Expensive	Expensive
Optimization Guarantee	Impossible	Impossible	Impossible
Applying result	Easy	Easy	Easy
Encoding	Hard	Hard	Hard
Hybrid to neural net	Possible	Possible	Possible
Apply to optimize problem	Impossible	Impossible	Possible
Area of data handling	Narrow	Narrow	Wide

III. CLASSIFICATION USING THE GENE ALGORITHM

A. Classification Standard with the Gene Algorithm

This paper presents measures of each standard to classify to derails algorithm. Because the standard used performs a standard measure, placing stress on each algorithm's different performance side, user side, characteristics, and other information that draw each gene algorithm different several characteristics here, in the grouping process, the number of groups must be decided before the results are analyzed, and there is the problem that the performance of the algorithm is sensitive to the influence of the early group's center settings and to noise. Therefore, research using the technique or gene algorithm latest figure enemy and decide throng's number automatically is consisting. Also, research to solve the problem that can converge to a regional optimal solution using the gene algorithm is gone. The gene algorithm presented in this paper is useful for solving various combinations of optimization problems [4, 5]. It is a search algorithm that is probability enemy who allow fetters natural selection and principle of evolution. Specifically, is much used as a tool of impulse search and optimization, machine learning. This paper presents a measure for each standard to classify it into a detailed algorithm. The standard used here was extracted out from the characteristics of each gene algorithm. Each algorithm was set at a standard scale based on its efficiency, user interface, characteristics, and other information.

Split grouping must present the number of groups beforehand and the performance of the algorithm depends highly on the initial value and static status. These discomfort have motivated recent studies on topics such as the probability technique and studies that focus on deciding the number of groups automatically using a gene algorithm.

The gene algorithm used in this paper is known to be highly effective in solving various NP-complete optimally composed problems [4, 5]. It is a searching algorithm based on natural selection and evolutionary theory. It is especially useful for overall search and optimization and for machine learning equipment [6].

B. Population Initialization

The individual was expressed by a center value of each throng, and each individual was assigned a variableness length depending on the free value.

C. Fitness Function

The cohesion distance, an internal characteristic of a throng, and the approximate distance, which shows the external distance between throngs are calculated. The similarity [7, 8] value that takes into consideration the relationships and characteristics between throngs and use it as a fitness function is measured. The similarity is stated in the Equations below.

$$S_{ij} = \frac{1}{chD_{ij} \times AD_{ij}{}^2} \quad (1)$$

$$chD_{ij} = \frac{CD_{ij}}{\frac{CD_i + CD_j}{2}} \quad (2)$$

$$ADij = \frac{\sum_a^{ni} \sum_b^{nj} W_{ra} \times W_{rb} \times \|r_a - r_b\|^2}{n_i \times n_j} \quad (3)$$

$$CDij = \sum_a^{ni} \sum_b^{nj} W_{ra} \times W_{rb} \times \|r_a - r_b\|^2 \quad (4)$$

$$CDi = \frac{\sum_a^{ni} \sum_b^{nj} W_{ra} \times W_{rb} \times \|r_a - r_b\|^2}{2} \quad (5)$$

S: similarity, chD_{ij}: cohesion distance, CDi: connection distance, r_a, r_b: representative value vector, n_i: Number of location measures belonging to cattle group I, W_{ra} : individual number from the source data that representative value ra represents

D. Selection

A roulette wheel method and elitist model were used together.

E. Crossover

The genetic factors were not stated by their variableness length but with a fixed length. They were defined with no relevance to the positioning.

F. Mutation

A mutation operator and Gaussian function were used to change the specified genetic factor.

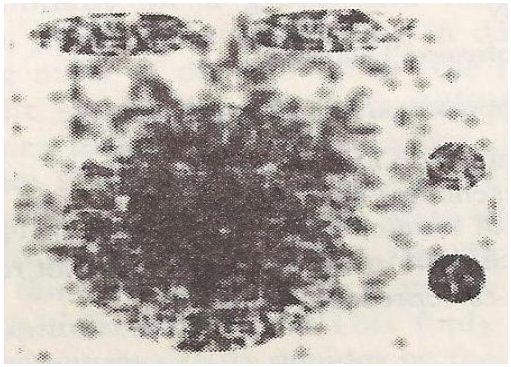
G. Analysis of the Gene Algorithm

The gene algorithm's input data type is basically based on a bit string. Also, the data size n is equivalent to 0 in a time complex. Although there is little room for scaling extensity, the concept of "very large (VL)" becomes possible by lining it up. Also, the optimal solution comes out as the final value making it possible to explain the conclusion and shows that a certain gene algorithm needs external training time although a typical gene algorithm does not need any. There are various kinds of factors such as crossover probability, mutation probability, dimension of population, with other selection, encryption, crossover, and mutation type making the use difficult. A parallel/analysis algorithm is possible and can be applied to the grounds of classification, estimation, and optimization [9-11].

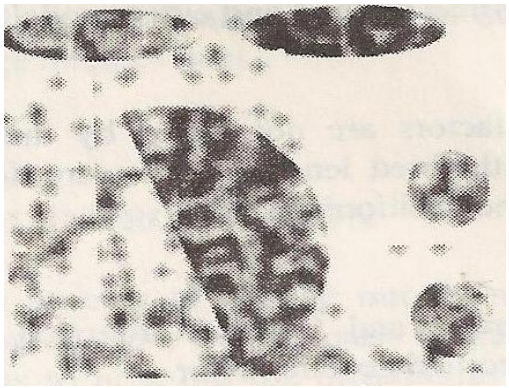
IV. EXPERIMENTS

This system uses C#, MSSQL 2000 in a Windows XP server environment, applying it to the student course guidance system. While the existing algorithm converges to a part value and it does not look for a suitable balance, the suggested algorithm not only finds the number of throngs automatically but also composes a comparatively accurate throng.

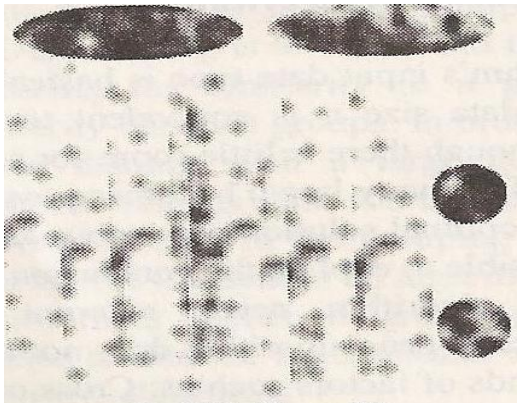
According to the results of the experiment with 2,000 data points and 5 throngs in 2-dimensional space, we conclude that the suggested algorithm finds good quality throngs amongst the data population clustered with static of all sizes. Fig. 1 shows two dimensional space data. Table 2 shows Performance of grouping techniques for the set of space data. Table 3 shows performance of grouping techniques for set of UCI data. The results are shown as follows.



(a) Data type 1



(b) Data type 2



(c) Data type 3

Fig. 1. Two-dimensional space data.

Table 2. Performance of grouping techniques for the set of space data (i)

	Hierarchic grouping		
	Shortest	Longest	Average
Data type 1	8.062	6.011	5.672
Data type 2	9.699	8.041	5.853
Data type 3	2.494	2.619	2.576

Table 3. Performance of grouping techniques for set of UCI data (Q)

	Hierarchical grouping		
	Shortest	Longest	Average
Australian	1.536	1.333	1.536
Diabetes	0.603	0.636	0.604
Heart	1.674	1.652	1.375
Iris	0.853	0.891	0.914
Soybean	3.091	3.091	3.091
Wine	0.986	0.711	0.984
Zoo	1.355	1.371	1.302

UCI: universal communications identifier.

After experimenting with the effectiveness of the algorithm on actual data, the (universal communications identifier) UCI machine learning repository, the algorithm showed a comparatively nice grouping performance. The performance evaluation function that evaluates the performance of the throng was measured by i in the following Equation. D_i represents the average distance among all data system in throng I and Q is the average of the D_i.

$$\mathrm{i} = \sum_{\mathrm{i}=1}^{\mathrm{k}} \frac{\mathrm{D_i}}{\mathrm{k}} \cdot \mathrm{D_i} = \frac{\sum_{\mathrm{a}=1}^{\mathrm{ni}} \sum_{\mathrm{b}=1}^{\mathrm{ni}} \sqrt{(\mathrm{x_a} - \mathrm{x_b})^2}}{\mathrm{n_i}(\mathrm{n_i} - 1)} \quad (6)$$

The most suitable algorithm was selected in each algorithm and the supporting evidence is all different. The data type 1, connection rule, and clustering algorithms that are focused on data mining are continuously improving their performance targeting towards a massive database, giving the means to evaluate the difference in the ability of each algorithm.

However, in the case of data type 2, it is impossible to compare and analyze the machine learning algorithm since its groundwork leans on the performance aspect. Therefore, algorithm type 2 is only referred to in the comparison of the gene algorithm regarding the fact that the last generation's final value represents the most suitable solution that comfortably expects no incidental explanation of a conclusion, the final value is easily used, the various data forms can be applied smoothly, it can handle a wide range of data, and that it can be applied to more optimized problems. Also, it can be integrated well with a neural net. The disadvantage is that it is difficult to encode the many questions in a fixed length gene, there is no guarantee of optimization, and it can be used only with small digits due to its high calculation cost. If some of these shortcomings are solved, we predict that gene algorithms will be widely used in the fields of data mining.

V. CONCLUSIONS

In this paper we propose and evaluate the idea that a more sufficient group can be found than the suggestion from a grouping algorithm that uses gene algorithm to automatically decide on the number of groups. Also, 8 gene algorithms were selected to be specifically analyzed. Since the environment of each algorithm and the purpose of mining differ greatly, it is impossible to experiment with them using the exact same standard and environment.

The sole purpose is to find and use the optimal algorithm in future systems. Therefore, the result of this study will help develop a foundation for data mining using the gene algorithm as a tool. It will also help fulfill the purpose of setting a suitable value as a guidepost in finding the right algorithm for data mining. In future studies, we plan to optimize the student course system and make the handling of the data more efficient by speeding up the calculation process.

REFERENCES

[1] M. J. A. Berry and G. Linoff, *Data mining techniques: for marketing, sales, and customer support*, New York: John Wiley & Sons; 1997.

[2] R. Agrawal and J. C. Shafer, "Parallel mining of association rules," *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 962-969, 1996.

[3] J. Katz, R. Ostrovsky, and M. Yung, "Efficient password authenticated key exchange using human memorable passwords," *Proceedings of the International Conference on the Theory and Application of Cryptographic Techniques: Advances in Cryptology*, Innsbruck, pp. 475-494, 2001.

[4] D. F. Specht, "Probabilistic neural networks," *Neural Networks*, vol. 3, no. 1, pp. 109-118, 1990.

[5] M. J. L. Orr, *Introduction to radial basis function networks*, Edinburgh: University of Edinburgh; 1996.

[6] A. Hinneburg and D. A. Keim, "An efficient approach to clustering in large multimedia databases with noise," *Proceeding of 4th International Conference of knowledge Discovery and Data Mining*, NewYork, pp. 58-65, 1998.

[7] J. D. Kelly and L. Davis, "Hybridizing the genetic algorithm and the k nearest neighbors classification algorithm," *Proceedings of the 4th International Conference on Genetic Algorithms*, San Diego, pp. 377-383, 1991.

[8] M. Ankerst, M. M. Breuning, H. P. Kriegel, and J. Sander, "OPTICS: ordering points to identify the clustering structure," *Proceedings of ACM SIGMOD International Conference on Management of Data*, Philadelphia, pp. 49-60, 1999.

[9] J. Bala, J. Huang, H. Vafaie, K. DeJong and H. Wechsler, "Hybrid learning using genetic algorithms and decision tree for pattern classification," *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, pp. 719-724, 1995.

[10] H. H. Lee, J. M. Park, and B. J. Cho, "Application of gene algorithm for the development of efficient clustering system," *The International Conference on Multimedia Technology and its Applications*, Uttar Pradesh,pp. 96-99, 2003.

[11] S. S. Anand, W. R. D. Patterson, J. G. Hughes, and D. A. Bell, "Discovering case knowledge using data mining," *Proceedings of the 2nd Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining*, Melbourne, pp. 25-35, 1998.

Jong-Min Park

He received the B.S. in Artificial Intelligence and Ph.D. degrees in the Dept. of Computer Engineering from Chosun University. In 2008 he joined the faculty of Chosun College of Science & Technology. His research interests are information security, biometrics, network security, information security, pattern recognition, and artificial intelligence.