The Application Research of FCM Clustering Based on Genetic Algorithm in the Telephone User's Behavior

Hailiang Tang [1, a], LeiShi [2,b] and BinLiu [3,c]

[1] School of Information Science and Engineering, Shandong Normal University, JiNan, P.R.China, 250014

[2] ShanDong Provincial Academy of Educational Recruitment and Examination, JiNan, P.R.China, 250014

[3] School of Information Science and Engineering, QiLu Normal University, JiNan, P.R.China, 250014

[a]18766140935@qq.com, [b]sdnushilei@163.com, [c]15053188785@139.com

Keywords: FCM Clustering Algorithm; Genetic Algorithm; MATLAB; User Behavior Analysis

Abstract. Telecom user behavior analysis, is that in the case of gaining the basic consumption data of the users to disposal, count, and analyze the relevant data, and discover the law of the users consumption from it, and combine these laws with the telecom marketing strategies to find the problems in the current marketing campaign which will provide the basis for the design of the scientific decision-making, specific marketing and cross-selling program. Fuzzy C-Means Clustering Algorithm is a method of data mining, and also an algorithm of fuzzy set analysis theory which based on the soft partition. The FCM Clustering Algorithm is easily falling into local minimum. However, the Genetic Algorithm has ability of global optimization, so it is very reasonable to combine the genetic algorithm with FCM Clustering Algorithm into the Telecom user behavior analysis. This paper will combine the Genetic Algorithm with FCM Clustering Algorithm and use MATLAB to analyze the feasibility of the algorithm.

1. Introduction

With the continuous development of big data era, the data processing of the user behavior analysis has been increasingly valued by all walks of life. Basic data on user behavior is often disorderly and unsystematic, but at the same time it contains a huge social value and commercial value. It is a thorny issue faced by all walks of life that how to dig out the back value hidden in the basic data of the user behavior [1]. Major Chinese telecommunications companies have accumulated a lot of user behavior data, and urgently need to deal with the relevant data to get the user's interest points and better contribute to the society, promotion marketing, and scientific decision-making.

Data Mining is the birth of technology to handle the big data. It is applied to the practice continuously in recent years, and it also gets better results [2]. Telecommunications companies can use the relevant Data Mining technology to analyze the user behavior better and to lay out the foundation for the community service and promotion marketing.

Clustering Method is also known as unsupervised classification. As an important data processing technology in the Data Mining, Clustering Method is the process which bases on the similarity of things to distinguish and classify [3].

In Data Mining, the traditional clustering analysis (such as K-Means) is based on the proper data attribute partition, and a data attribute either-or hard classification. In real life, most object attribute don't have strict attribute distinction [4]. Fuzzy clustering analysis is to deal with this kind of object data set that doesn't have strict attribute distinction. Since the blur muster theory was proposed in 1965, fuzzy clustering analysis theory continuous development, and apply to practical application of all fields of work. So far, all walks of life both at home and abroad have been appearing many fuzzy clustering algorithms. Among these fuzzy clustering algorithms, Bezdek generalized fuzzy algorithm in 1981, and built the Fuzzy C-Means Clustering theory which based on the partition method, namely FCM Clustering Algorithm. It means clustering theory is based on the establishment of fuzzy

c-means clustering algorithm, namely, FCM Clustering Algorithm. Since then FCM Clustering Algorithm that based on the objective function was researched from various angles by more and more scholars, and also the most widely studied classic fuzzy clustering algorithm so far which based on the partition method [5].

In order to solve the problem that the FCM Clustering Algorithm is easily falling into the locally optimal solution in dealing with data set, we introduce the Genetic Algorithm. Genetic Algorithm is a global optimization algorithm, this algorithm is easy to operate, and can realize the effective searching for the global optimal solution, so adding the genetic algorithm into FCM Clustering Algorithm can overcome the shortcoming that FCM Clustering Algorithm is easily falling into the local optimal solution [6].

Because the data of telecom users behavior is the matrix form in nature. It is very easy to disposal by FCM Clustering Algorithm. Introducing the Genetic-FCM Algorithm into the telecom industry user behavior data analysis is benefit to serve the uses for the telecom industry, and also better to propose the scientific suggestion for the decision-making and marketing of the telecom industry [7].

2. Fuzzy C-Means Clustering Algorithm

Fuzzy C-Means Clustering Algorithm is a typical clustering algorithm based on both objective function is based on the Partition Method clustering algorithm, which is from the traditional the K-Means Algorithm, namely Hard C-Means Clustering Algorithm research evolved. Features FCM Clustering Algorithm is easy to understand, describe concise, practical, fast convergence and automatic classification and other advantages [8].

2.1 Introduction to the Fuzzy C-Means Clustering Algorithm

FCM Clustering Algorithm is a kind of algorithms that in fuzzy membership to determine each data point belong to some degree of clustering. It takes square-error and function as the objective function [9]. Assuming that the sample set is $X = \{x_1, x_2, \cdots, x_n\}$, if these data are divided into class C, correspondingly, it will have c cluster centers is C. Every sample j belong to some degree of membership of class "i" is u_{ji}. So, how to define the objective function of a FCM Clustering Algorithm (1-1) and constraints (1-2), as follows:

$$J_m = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} \left\| X_j - c_i \right\|^2 \qquad (1\text{-}1)$$

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1,2,\ldots,n \qquad (1\text{-}2)$$

The objective function of FCM Clustering Algorithm is calculated by the square-errors, where c is given to a number of categories, and $1 < c < n$; n is the data sample set in X number; u_{ij} is the degree of membership of class I in the j sample of the sample set X; m is the membership weighting exponent u_{ij}, it is also called fuzzy weighted index or smoothness index; c_i is the cluster center in the class "i"; $\left\| X_j - c_i \right\|^2$ is the error metric。

We need to explain here, the membership u_{ij} is generally set any real number between 0 and 1, and "$m \geq 1$".

We take the Lagrange multiplication to disposal the constraint condition and the objective function, and then we will obtain the objective function. As the following equation (1-3) below:

$$J = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} \left\| X_j - C_i \right\|^2 + \lambda_1\left(\sum_{i=1}^{c} u_{i1} - 1\right) + \cdots + \lambda_j\left(\sum_{i=1}^{c} u_{ij} - 1\right) + \cdots + \lambda_n\left(\sum_{i=1}^{c} u_{in} - 1\right) \qquad (1\text{-}3)$$

Then we have to take the derivative of the membership u_{ij} and the cluster center c_i, and take the constraint condition (1-2) into it, at last, we will draw the iterative formula (1-4) of the membership u_{ij} and the iterative formula (1-5) of the cluster centers c_i, as follows.

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\left\| x_j - c_i \right\|}{\left\| x_j - c_k \right\|}\right)^{\left(\frac{2}{m-1}\right)}} \qquad (1\text{-}4)$$

$$c_i = \frac{\sum_{j=1}^{n}(x_j u_{ij}^{m})}{\sum_{j=1}^{n} u_{ij}^{m}} \qquad (1\text{-}5)$$

We can draw the following conclusions by observing the iterative formula (1-4) of the membership u_{ij} and the iterative formula (1-5) of the cluster center "c_i".

When m = 1, FCM Clustering Algorithm will not have Blur Effect, it will metamorphose into HCM Clustering Algorithm, in other words, the traditional K-Means Algorithm;

When "$m \to 1$", FCM Clustering Algorithm will slowly lose Blur Effect, and gradually degenerate into HCM Clustering Algorithm.

When "$m \to +\infty$", FCM Clustering Algorithm will over blurred, FCM Clustering Algorithm will gradually lose divide effect [10].

2.2 The steps of FCM Clustering Algorithm

General steps of the FCM Clustering Algorithm:

（1）Standardization of original data, to form a standard sample data sets.

Because the FCM Clustering Algorithm is sensitive to noise, we should also process the related original data when we process the original data sets, Standardization of original data includes the characteristic standardization and dimension reduction.

（2）Initialization of relative parameters.

Determine the value of cluster class number c and the membership factor m, and the iterative time t, and the allowable deviationε.

（3）Initialize a membership u_{ij}

Initialization of the membership "u_{ij}", as follows (1-4).

（4）Computing the cluster center c_i according to the membership u_{ij}.

The computation formula of the cluster centerc_i, as follows (1-5).

（5）Calculation error $e = \sum_{i=1}^{c} \|X_j - c_i\|^2$;

If "$e < \varepsilon$", then the FCM Clustering Algorithm is end, and turn to step (7); otherwise, $t = t + 1$, and turn to step (6).

（6）According to the cluster centerc_i to calculate the membershipu_{ij}, go back to step (3), go and return in following a circle until to the stop of iteration t.

（7）According to the result of iteration to confirm the class of sample data, and show the final clustering result.

3. Genetic Algorithm

The concept of the Genetic Algorithm was first proposed by the Bagley J.D in the mid 1950s. Until 1960s Professor John Holland from Michigan University in America started to study and disposal systematically the mechanism of Genetic Algorithm, and put forward the complete theory and method on this basis. Genetic Algorithm is created by the thoughts of Darwin's Biological Evolution Theory and Mendel's Genetics Theory. It follows these mechanisms of "heredity", "variation", "survival of the fittest" and "superior bad discard ", and it is a global search algorithm [11]. With the continuous research and development of the Genetic Algorithm, it has been used widely in combinatorial optimization, artificial intelligence, neural networks, self-adaptive control and operational research and etcetera. It is one of the core technologies of the modern evolutionary computation method.

3.1 Introduction to the Genetic Algorithm

The Genetic Algorithm is based on the Biological Evolution. It is an artificial intelligence algorithm of self-organization and adaptation for biological evolution process and mechanism to solve problems [12]. Genetic Algorithm is to imitate an evolutionary process of artificial population in essence. With selection, hybridization, and variation to artificial population, and after several generations, at last, it can be top the best condition. The advantages of Genetic Algorithm are strong robustness, simple process, and easy to combine the other algorithms and so on. The mechanism of genetic algorithm is global parallel search, so it can actually be get global optimization [13].

The process of the Genetic Algorithm as follows:

The first step: Coding.

As the practical problem, firstly, we should look for the coding scheme that the solutions of problems are digitized. We generally use the real number to code because the real number encoding has the advantages of high precision, large search space and so on [14].

$$y = \frac{(a+x) \times (2^{16}-1)}{2a} \qquad (3\text{-}1)$$

X in the above formula can be decoded by the following formula:

$$x = -\frac{a \times (2^{16}-1) - 2axy}{(2^{16}-1)} \qquad (3\text{-}2)$$

The second step: Population Initialization.

Random initializing a population, the individuals in population are these digitized coding. For example, a population X was generated randomly, there are n individuals as the initial population, namely "$X = \{x_1, x_2, \cdots, x_n\}$".

The third step: the fitness evaluation.

Moderate evaluation to the individual in the population X, and fitness evaluation is better to the population evolution, the bigger of the individual fitness evaluation value the greater probability of the individual to be chosen.

Individual fitness evaluation function, as follows (3-3):

$$f_m = \frac{1}{1+F_m} \qquad (3\text{-}3)$$

"f_m" is the solution of the adaptive value, F_m is the function that have to take in.

Step four: Population Evolution.

That means let the individual in the population do the following operations such as selection, crossover, and mutation, and then produce the new generation of individuals. Selection also called the Reclaimed Operator, that is to say, using individual fitness evaluating function in the third step to select the better individual in the population. In general case, the bigger of the individual fitness evaluation value the greater probability of the individual to be chosen. In here, we take rotating disc type selection strategy, that is, we can get an individual every rotation.

Every individual's rotating disc type selection function in population X, as the following formula (3-4):

$$P_i = \frac{f_i}{\sum_{i=1}^{n} f_i} \qquad (3\text{-}4)$$

The total probability that all individuals are selected in the populations X, as the following equation (3-5):

$$P_x = \sum_{i=1}^{n} f_i \qquad (3\text{-}5)$$

Cross refers to an operation that the part encoding structure of two selected last individual exchange and recombine each other and regenerate a new-generation of individuals [15]. Cross requires that the population should be like in the nature does not destroy the excellent gene of the population gene and can produce the new-generation individuals that can adapt new environment.

Variation is to change some coding value of the individual coding in the population. It requires that the population should be like in the nature make the individuals of a population to produce several variations. Mutation operation is added in genetic algorithm that can improve the random hunting function of the genetic algorithm and prevent Genetic Algorithm from appearing the problem of premature convergence.

3.2 The general step of Genetic Algorithm

（1）The original data standardization, and form standard sample data set.

（2）The initialization of the coding and relevant parameters.

According to the formula (3-1) and (3-2) to encode, and set all kinds of genetic parameters, namely, the iteration number $t = 1$ and the biggest iteration number "t_{max}".

（3）The Population Initialization.

Randomly generated a population X, set the number of population X is n, as an initial individual, namely, "$X = \{x_1, x_2, \cdots, x_n\}$".

（4）The fitness evaluate of individual.

Using the fitness function (3-3) to calculate the fitness evaluate of individual in the population X.

（5）If individual match condition, and output match individual conditions, if the unsuitable individual number is 0 or $t = t_{max}$, ending the algorithm, if the unsuitable individual number is not 0, then carry out these unsuitable individual number into (6).

（6）Produce a new-generation of individuals. After individual selection, crossover and mutation, then there will produce a new-generation of individuals. So that $t = t + 1$, then return to the step (4).

4. GA-FCM Clustering Algorithm

A combination of FCM Clustering Algorithm and Genetic Algorithm, and using the powerful global searching ability of Genetic Algorithm, FCM Clustering Algorithm has the high convergence velocity and efficient local searching ability, and can do the fast and high-efficient clustering analysis to the data, and to get the best needed suboptimal clustering results. The GA-FCM Clustering Algorithm's basic idea is to use the FCM Clustering Algorithm to make every data-intensive data tend to each extreme point. By Genetic Algorithm it can break away from getting in local minimum during the convergence procedure. Repeat this operation until get the suboptimal clustering results [16-19].

4.1 The implementation procedure of the GA-FCM Clustering Algorithm

（1）The original data standardization, and form standard sample data.

Because the FCM Clustering Algorithm is sensitive to noise and required to the data, we should also process the related original data when we process the original data sets, Standardization of original data includes the characteristic standardization and dimension reduction.

（2）GA-FCM Clustering Algorithm parameters initialization.

Customize relative parameters of genetic algorithm include population size PS, crossover probability "P_c", mutation probability "P_m", the maximum iterations "t_{max}". Customize correlation parameter of FCM Clustering Algorithm include cluster classification number c, values of membership factor m;

（3）The initialization of the coding and population.

According to the formula (3-1) and (3-2) to encode, and randomly generated a population X, there are n individuals as the initial individual, namely "$X = \{x_1, x_2, \cdots, x_n\}$".

（4）The fitness evaluate of individual in the population.

According to the formula (4-4) to calculate fitness evaluate of individual in the population of GA-FAM Clustering Algorithm.

$$f_m = \frac{1}{1+J_m} \qquad (4\text{-}1)$$

J_m's computation formula, as follows (1-1).

（5）Produce a new-generation of individuals.

Using Genetic Algorithm to select, crossover and mutation for the individuals of the population, and generate a new-generation of individuals.

（6）If $t = t_{max}$, the Genetic Algorithm is end, and output the final data, and turn to the step (7); Otherwise, let "$t = t + 1$", and return back to the step (4).

（7）According to the global optimal solution to fuzzy divide entire data set[20].

According to the final data obtained by $t = t_{max}$, namely, the globally optimal solution of the GA-FCM Clustering Algorithm. According to the global optimal solution to fuzzy divide the entire data sets. Therefore, GA-FAM Clustering Algorithm execute is finished.

4.2The experimental analysis of the GA-FCM Clustering Algorithm

MATLAB (Matirx Laboratory) is a software which to use for the Matrix manipulation in the American Math Works company, and it can analysis algorithm performance and data analysis well [21-23]. This paper's experimental environment is MATLAB7.14, and the operating system is flagship version Windows 7 and CPU64.

Experimental data is the data of telecom user behavior analysis that supplied by the China Mobile. Here, the data we work with are the total consumption, call charges, message cost and GPRS fees of a month of China Mobile users.

In order to verify the stability and effectiveness of the algorithm, here we are going to analysis 2000 China Mobile user behavior data separately in FCM clustering algorithm and GA-FCM clustering algorithm. Some sample data as follow 4-1.

ID	Total Consumption （RMB）	Call Cost （RMB）	SMS Cost （RMB）	GPRS Cost （RMB）
1	46.4	12	8.4	10
2	38.3	15.6	5.7	5
3	69.8	34.7	10	15
4	75	40	10	20
5	46	15.8	5.2	15
6	78	24	10	20
7	45.1	18	5.1	15
8	58	12.4	3.1	10

Fig. 1：Data samples of user behavior

In the FCM Clustering Algorithm, we set the sample data set "$\mathbf{X} = \{x_1, x_2, \cdots, x_{2000}\}$", the initial cluster number "$\mathbf{c} = 3$", the membership factor "$\mathbf{m} = 3$", in the genetic algorithm, we set the population size "$\mathbf{PS} = 200$", crossover probability "$P_c = 0.75$", mutation probability "$P_m = 0.02$", the maximum number of iterations "$t_{max} = 50$". In the MATLAB, we use FCM Clustering Algorithm and GA-FCM Clustering Algorithm to disposal the data separately, as follow.

Table 1：Comparative FCM and GA-FCM in MATLAB.

	FCM	GA-FCM
The minimum of the objective function.	0.674-0.712	0.645-0.703
The average distance of class within.	0.952-1.002	0.852-0.901
Average number of iterations.	34	41

By 2000 sample data analysis in MATLAB, we can get the conclusion that the GA-FCM Clustering Algorithm is more advantage than the traditional FCM Clustering Algorithm in disposing the telecom user behavior analysis. Therefore, we can consider apply to this algorithm into the telecom user behavior analysis.

5. Summary

This paper introduces the basic principles and implementation FCM Clustering Algorithm and Genetic Algorithm, inadequate in dealing with telecommunications user behavior data for FCM Clustering Algorithm, the Genetic Algorithm into FCM Clustering Algorithm, Genetic Algorithm excellent global optimization FCM Clustering Algorithm ability and fast convergence capability, two different algorithms complement each other to improve the performance of the algorithm, making it easier to process the data telecommunications user behavior.

In order to verify the merits of the GA-FCM Clustering Algorithm, we collect relevant data on test simulation MATLAB, and the clustering results were analyzed, the results show, FCM Clustering Algorithm based on Genetic Algorithm alone when dealing with telecommunications user behavior data of FCM Clustering Algorithm than poly class of algorithm performance is much better.

References

[1] LI Si-nan, LI Ning, LI Zhan-huai. Multi-label Data Mining: A Survey [J]. Computer Science, 2013, 40(4), 14-20.

[2] HE Qing, LI Ning, LUO Wen-Juan, SHI Zhong-Zhi. A Survey of Machine Learning Algorithms for Big Data [J]. PR & AI, 2014, 27(4), 327-336.

[3] Wu Yu-hong. Discussion on the Method of Cluster Analysis [J]. Computer Science, 2012, 39(6A), 325-327.
[4] ZHOU Tao, LU Huiling. Clustering algorithm research advances on data mining [J]. Computer Engineering and Applications, 2012, 48(12), 100-111.
[5] PIAO Shang-Zhe, Chaomurilige, YU Jian. Cluster Validity Indexes for FCM Clustering Algorithm [J]. PR & AI, 2015, 28(5), 452-461.
[6] ZHU Wenjie, WU Nan, HU Xuegang. Improved cluster validity index for fuzzy clustering [J]. Computer Engineering and Applications, 2011, 47(5), 206-209.
[7] SONG Jiao, GE Lin-dong. Fuzzy cluster algorithm based on genetic algorithm and its application [J]. Computer Applications, 2008, 28(5), 1196-1199.
[8] WEN Zhong-wei, LIRong-jun. Fuzzy C-means clustering algorithm based on improved PSO [J]. Application Research of Computers, 2010, 27(7), 2520-2522.
[9] XIAO Man-sheng, WEN Zhi-cheng, ZHANG Ju-wu, WAN Xin-fan. An FCM clustering algorithm with improved membership function [J]. Control and Decision, 2015, 30(12), 2270-2274.
[10] Gao Cuifang, Wu Xiaojun. Improved Algorithm for Weighted Fuzzy Kernel Clustering Analysis [J]. Journal of Data Acquisition & Processing, 2010, 25(5), 631-636.
[11] WANG Fang, DAI Yong-shou, WANG Shao-shui. Modified chaos-genetic algorithm [J]. Computer Engineering and Applications, 2010, 46(6), 29-32.
[12] BIAN Xia, MI Liang, Development on genetic algorithm theory and its applications [J]. Application Research of Computers, 2010, 27(7), 2425-2429.
[13] MA Yong-jie, YUN Wen-xia, Research progress of genetic algorithm [J]. Application Research of Computers, 2012, 29(4), 1201-1206.
[14] ZHANG Chao-qun, ZHENG Jian-guo, QIAN Jie. Comparison of coding schemes for genetic algorithms [J]. Application Research of Computers, 2011, 28(3), 819-822.
[15] LI Shuquan, SUN Xue, SUN Dehui, BIAN Weipeng. Summary of crossover operator of genetic algorithm [J]. Computer Engineering and Applications, 2012, 48(1), 36-39.
[16] SHI Liang, ZOU Yi, YIN Yan, ZHUANG Zhen-quan. Fuzzy Cluster Technique Based on Active Evolution Based Genetic Algorithm [J]. MINI-MICRO SYSTEMS, 2005, 26(2), 204-208.
[17] GUAN Qing, DENG Zhaohong, WANG Shitong. Improved fuzzy C-means clustering algorithm [J]. Computer Engineering and Applications, 2011, 47(10), 27-29.
[18] Yang Cuiqiong, Jiang Hong, Yu Xiaolei. Analysis of an Improved Fuzzy C-Means Clustering [J]. Computer & Digital Engineering, 2010, 38(5), 1-3.
[19] CHEN Xiao-guo. Study on improvement of alterable weighted FCM clustering algorithm based on genetic algorithm [J]. Journal of Science of Teachers' College and University, 2011, 31(1), 12-15.
[20] ZHANG Yongku, YIN Lingxue, SUN Jinguang. Fuzzy clustering algorithm based on improved genetic algorithm [J]. CAAI Transactions on Intelligent Systems, 2015, 10(4), 627-635.
[21] ZHU Ran, LI Ji-ying. Contrast and Simulation of Several Clustering Centers of Optimized FCM Algorithms [J]. COMPUTER TECHNOLOGY AND DEVELOPMENT, 2015, 25(5),17-20.
[22] CHANG Li-yan, WANG Xue-fen. Based on MATLAB's Text Fuzzy Clustering Analysis and Application [J]. Computer Knowledge and Technology, 2012, 8(25), 5937-5942.
[23] Song Lihong. Matlab simulation design for K-means clustering [J]. Experimental Technology and Management, 2010, 27(10), 101-103.