International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-2 Number-2 Issue-4 June-2012

Web Usage Mining Clustering Using Hybrid FCM with GA

Karunesh Gupta[1], Manish Shrivastava[2]
[1]M. Tech (IT), LNCT, Bhopal, India
[2]Department of Computer Science Engineering, LNCT, Bhopal, India

Abstract

The most widely used clustering algorithm implementing the fuzzy philosophy is Fuzzy C-Means (FCM) .In this paper, we have proposed a new Hybrid FCM with Genetic Algorithm (GA), we get an improved FCM algorithm which has not only the global search capability of GA but also the local search capability of FCM, and hence can better solve the clustering problem. An improved version of this hybrid clustering algorithm is, therefore, proposed to reduce the number of iterations. The proposed algorithm is tested with MSNBC.com web access logs collected from UCI (University of California Irvine) dataset repository.

Keywords

Fuzzy C-Means, Genetic Algorithm, Fuzzy clustering

1. Introduction

Fuzzy C-Means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method was developed by Dunn [1] in 1973 and improved by Bezdek [2] in 1981.FCM is a method of solving the problem of clustering. It has such merits as simple designing, universal application and easy transformation to optimization problem so as to be solved by classic nonlinear projecting theory. GA operates on a population of potential solutions applying the principle of survival of the fittest to produce (hopefully) better and better approximations to a solution [3]. At each generation, a new set of approximations is created by the process of selecting individuals according to their level of fitness in the problem domain and breeding them together using operators borrowed from natural genetics. This process leads to the evolution of populations of individuals that are better suited to their environment than the individuals that they were created from, just as in natural adaptation .GA, as an optimization algorithm, contains three operators: selection, crossover and mutation. Having a balanced global and local search capability due to its crossover and mutation operation, it doesn't get stuck in local optimums easily. When combining GA with FCM, we get a new algorithm which has not only the global search capability of GA but also the local search capability of FCM, and hence can better solve the clustering problem. An improved version of this hybrid clustering [4, 5] algorithm is, therefore, proposed in this paper to improve the clustering results.

The remainder of this paper is organized as follows. The proposed clustering by Hybrid FCM with GA is described in the section 3. The section 4 illustrates experimental setup of the proposed approach. This section also gives performance evaluation with the existing algorithms. Finally, the section 5 concludes the paper.

2. Literature Review

The FCM algorithm evolved from Hard C-means (HCM) has become one of the basic, important, and popular methods that can be used to solve uncertainties accompanying real-life situations. According to [6], more data require more time for calculation; to improve the clustering efficiency, we usually use sampling for database shrinkage. [7] Propose a local search technique that searches for the optimum by virtue of *FCM* iterative *hill-climbing.* The clustering results are vulnerable to the randomly selected cluster centroid. Besides, the algorithm will suffer from the problem of the local optimum during the optimal clustering. To address this problem, scholars use different algorithms to find the cluster centroid most close to the global optimum and then proceed with FCM procedures, so as to avoid the problem of the local optimum. Genetic Algorithm is widely used as the optimization method in evolutionary computation based on inheritance and natural selection. As the multi-point concept is applied to search for data space and the mutation rule is random but not fixed, we may reduce the probability of getting trapped on a false peak. [8] suggest that a traditional FCM algorithm is always converged to its local optimum instead of a global optimum.

3. Proposed Approach

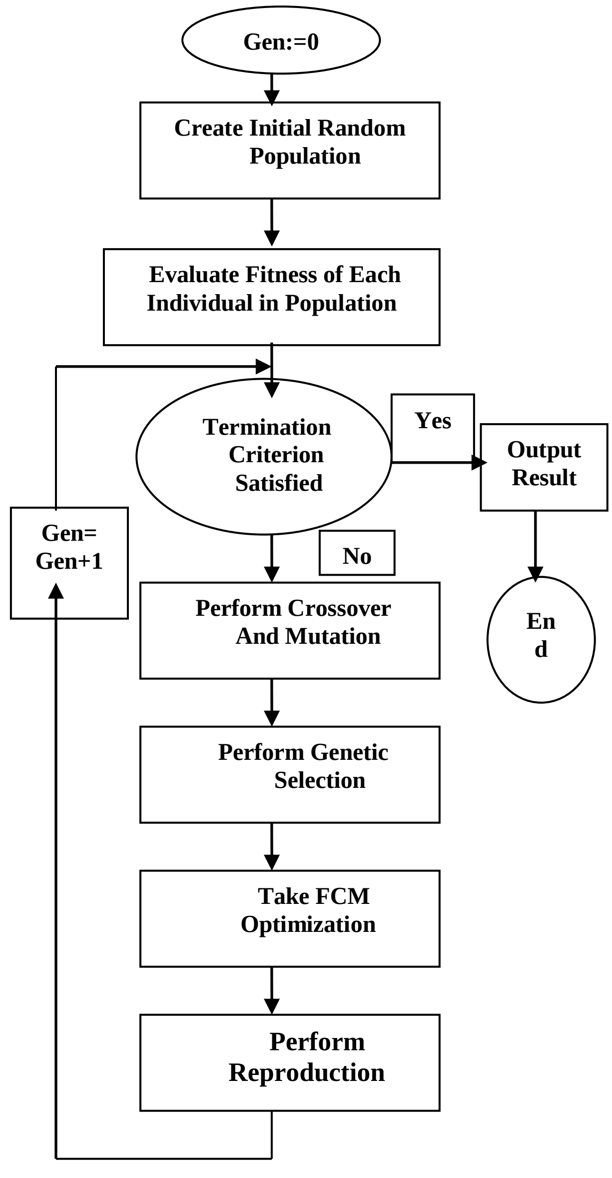
Hybrid FCM with GA, we get an improved FCM algorithm which has not only the global search capability of GA but also the local search capability of FCM, and hence can better solve the clustering problem. An improved version of this hybrid clustering algorithm is, therefore, proposed to improve the clustering results. First, the individual stochastic match crossover in GA is changed to proportional selection individual stochastic match crossover. Then the individual reproduction after proportional selection in GA is canceled. The results are both an increase in the colony the sequence of crossover and mutation in GA is changed to parallel and a new population is chosen from its parent generation colony and the post-genetic-operation offspring colonies, which is composed of the same number of excellent individuals and retains the same scale. As a result, the colony diversity will be preserved and the robustness and efficiency of the algorithm will be improved. FCM optimization is applied after each generation of genetic operation, which reduces the number of iteration.

Here mention step of algorithm
Step 1: Input web log data X;
Step 2: Chromosome is coded in real number and initialize population X (i), i = 0 at random;
Step 3: Compute the fitness of each individual in the current instant;
The fitness function of algorithm is determined by F(x).

$$F(X) = \begin{cases} (\overline{\alpha} + 2\beta) - \alpha i, & \alpha i < \overline{\alpha} + 2\beta \\ 0, & \alpha i \geq \overline{\alpha} + 2\beta \end{cases}$$

i=1, 2,…………………………….., N

where $\overline{\alpha}$ and β are the respective average value and standard error of the objective function value αi.
Step 4: Check the termination conditions. If the termination conditions are satisfied, then turn to step 5, otherwise, turn to step 6;
Step 5: Decode the chromosome and calculate the optimal clustering centers and fuzzy partition matrixes. And set the optimal clustering partition according to maximum membership principle and output the results.
Step 6: Perform the parallel crossover and mutation operation on population X(i), then we can get population Y(i), Z(i) respectively;
Step 7: Carry out the genetic selection on the instant composed of population X(i), Y(i), Z(i) and population W(i) is got;
Step 8: Apply the FCM optimization on population W (i) and generate the next generation X (i +1). Then turn to step 3.



a) Flow chart of Hybrid FCM with GA

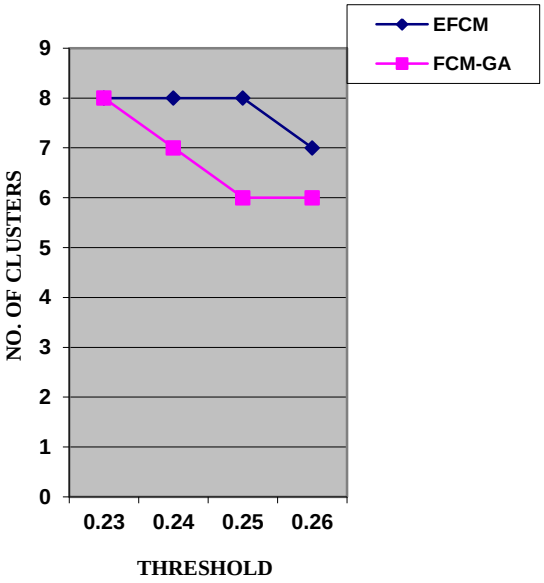
4. Performance and Comparative Studies

In the comparative tests described in this section the threshold value is used to compare the Hybrid FCM-GA with Entropy based FCM (EFCM) [8, 9, 10, and

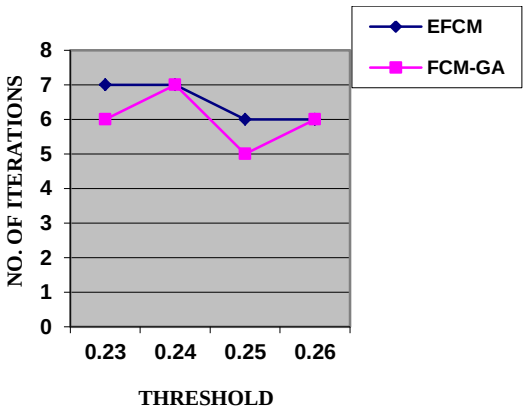
11] algorithm. Hybrid FCM-GA is applied number of times from different groups to obtain the number of the optimum clustering centers.

FCM Based on Entropy Weighting

EFCM uses the information entropy to determine the number of cluster centers, which reduces the algorithm's dependence on the initial cluster centers and improves clustering performance greatly. In the clustering process, we simultaneously combine with the ideas of merger. Then, introduces the weighting parameter into the fuzzy clustering, which makes the clustering center position is closer to the actual position. In the meantime, the introduction of weighting factor makes the algorithm deal with certain noisy data problems, which broadens the scope of application of the algorithm and reduces the error. In the clustering process, the algorithm combines the merger of ideas, which can make the algorithm get any classes shape, and reduce the algorithm's dependence on the data sets.



A) Between threshold and no. of clusters.



B) Between threshold and no. of iterations

5. Conclusion

In this paper we have presented a Hybrid FCM with GA which proved to be a powerful extension of the famous Fuzzy C-Means algorithm. Essentially the Hybrid FCM with GA algorithm is a learning algorithm which can mine databases to build an optimized and efficient clustering results .If we combined FCM and GA together, a good clustering result is likely to be achieved. But if the two methods are simply put together, we can't achieve the optimal clustering results. Hence, the crossover, selection and mutation parts of the genetic algorithm are improved, which strengthens the global search capability. At the same time, this improved algorithm performs FCM optimization immediately after each generation of genetic operation, which reduces the number of iteration. The result of the experiment shows that the proposed algorithm can significantly improve the clustering results.

References

[1] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters'' Journal of Cybernetics, Vol.3, pp.32-57, 1973.

[2] Bezdek, J. C., "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, NY, 1981.

[3] Hu Yusuo,Cheb Zonghai., "Clustering analysis based on hybrid GA" [J].Pattern Recognition and Artificial Intelligence, 2001.

[4] R.Alcalá,J.M.Benítez,J.Casillas,O.Cordón, and R.Pérez. "Fuzzy control of HVAC systems optimized by genetic algorithms". Applied Intelligence, vol.18, pp.155-177, 2003.

[5] Odukoya, O.H, Aderounmu, G.A. And Adagunodo, E.R, "An Improved Data Clustering

Algorithm for Mining Web Documents" 978-1-4244-5392-4, 2010.

[6] Mishra, N., Oblinger, D. & Pitt, L. 2001. "Sublinera time Approximate clustering". Proc. 12th Ann. ACM-SIAM Symp. Discrete Algorithms (SODA), 2001.

[7] Cai, L., Huang, Y. & Chen, J. 2006. "A genetic-based fuzzy clsutering algorithm for fault diagnosis in satellite attitude determination system". Paper presented at the International Conference on Intelligent Systems Design and Applications,2006.

[8] Peng, H., Xu, L. & Jiang, Y. 2006. "Improved genetic FCM algorithm for color image segmentation". Paper presented at the International Conference on Signal Processing,2006.

[9] J. Vellingiri and S. Chenthur Pandian, "Fuzzy Possibilistic C-Means Algorithm for Clustering on Web Usage Mining to Predict the User Behavior**",** European Journal of Scientific Research ISSN 1450-216X Vol.58 No.2, pp.222-230 (2011).

[10] K.Suresh, R.Madana Mohana, A.RamaMohan Reddy**,** "Improved FCM algorithm for Clustering on Web Usage Mining", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 1, January 2011.

[11] Xuan SU ,Xiaoye WANG ,Zhuo WANG ,Yingyuan XIAZO, "An New Fuzzy Clustering Algorithm Based On Entropy Weighting", Journal of Computational Information Systems,pp.-3319-3326,2010.

Karunesh Gupta received the B.E in Computer Science and engineering from MPCT in 2005, Gwalior, Madhya Pradesh, pursuing M.Tech in Information Technology from LNCT, Bhopal. He has 3 years of experience in Software Company as a software engineer. His areas of interest include Data Mining, Web Mining and Database Systems.