Web Usage Data Clustering Using Improved Genetic Fuzzy C-Means Algorithm

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

Abstract

Web usage mining involves application of data mining techniques to discover usage patterns from the web data. Clustering is one of the important functions in web usage mining. Recent attempts have adapted the C-means clustering algorithm as well as genetic algorithms to find sets of clusters .In this paper; we have proposed a new framework to improve the web sessions' cluster quality from fuzzy c-means clustering using Improved Genetic Algorithm (GA). Initially a fuzzy c-means algorithm is used to cluster the user sessions. The refined initial starting condition allows the iterative algorithm to converge to a "better" local minimum. And in the second step, we have proposed a new GA based refinement algorithm to improve the cluster quality. The proposed algorithm is tested with web access logs collected from the UCI dataset repository.

Keywords

Web Usage Mining, Genetic Algorithm, Fuzzy C-Means.

1. Introduction

With the recent explosive growth of the amount of content on the Internet, it has become increasingly difficult for users to find and utilize information and for content providers to classify and catalog documents. Traditional web search engines often return hundreds or thousands of results for a search, which is time consuming for users to browse. In order to deal with these problems, researchers look toward automated methods of working with web documents so that they can be more easily browsed, organized, and cataloged with minimal human intervention.

Clustering and classification [1] have been useful and active areas of machine learning research that promise to help us cope with the problem of information overload on the Internet. With clustering the goal is to separate a given group of data item (the data set) into groups called clusters such that same cluster are similar to each other and dissimilar to the items in other clusters. In clustering methods no labeled examples are provided in advance for training (this is called unsupervised learning). Web mining methodologies [3] can generally be classified into one of three distinct categories: web usage mining, web structure mining, and web content mining a survey of techniques used in these areas.

In this paper, we are using FCM in first phase then we applying improved genetic algorithm. The first phase, we uses C-means 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. In second phase, the initial clusters centers are found using C-means algorithm. These give us centers that are widely spread within the data. GA takes these centers as it initial variables and iterates to find the local maxima. Hence, we get clusters that are distributed well using C-means and clusters that are compact using GA. In the subsequent section, we present the proposed architecture and experimentation results of the improved genetic FCM clustering comparing with entropy based FCM .Some conclusions are provided towards the end.

2. Web Usage Mining

In web usage mining the goal is to examine web page usage patterns in order to learn about a web system's users or the relationships between the documents. For example, the tool presented by Masseglia et al association rules from web access logs, which store the identity of pages accessed by users along with other information such as when the pages were accessed and by whom; these logs are the focus of the data mining effort, rather than the actual web pages themselves. Beeferman and Berger [2] described a process they developed which determines topics related to a user query using click-through logs and agglomerative clustering of bipartite graphs. The transaction-based method developed [2] creates links between pages that are frequently accessed together during the same session. Web usage mining [3] is useful for providing personalized web services, an area of web mining research that has lately become active. It promises to help tailor web services, such as web search engines, to the preferences of each individual user.

3. Proposed Approach

In this study, we propose a new algorithm to improve the C-means clustering in web usage data mining. The proposed algorithm consists of two steps. In the first step, to avoid local minima, we presented a simple and efficient method to select initial centroids. And the C-means algorithm is applied to cluster the data vectors. Then in the second step, Improved Genetic Algorithm (GA) is applied to refine the cluster to improve the quality of the clusters of users' sessions.

FCM algorithm

The C-means clustering algorithms are the simplest methods of clustering data. The C-means algorithm uses a set of unlabeled feature vectors and classifies them into C classes, where *C* is given by the user. From the set of feature vectors *C* of them are randomly selected as initial seeds. The feature vectors are assigned to the closest seeds depending on its distance from it. The mean of features belonging to a class is taken as the new center. The features are reassigned; this process is repeated until convergence.

To apply the C-means algorithm:

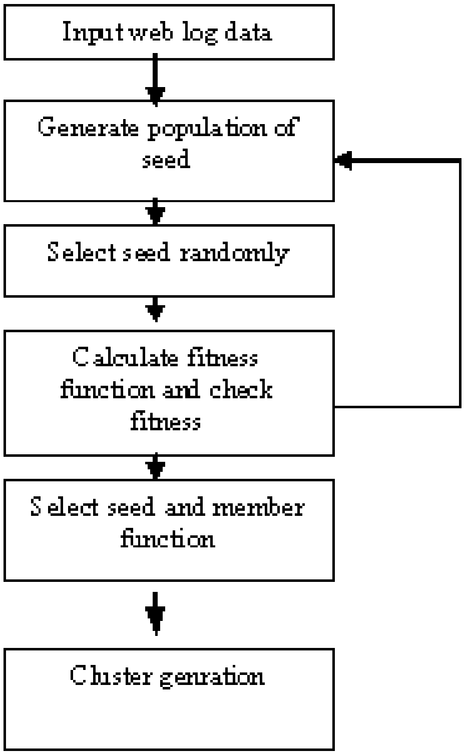
1. Choose C data points to initialize the clusters.
2. For each data point, find the nearest cluster center that is closest and assign that data point to the corresponding cluster.
3. Update the cluster centers in each cluster using the mean of the data points which are assigned to that cluster.
4. Repeat steps 2 and 3 until there are not more changes in the values of the means.

Improved Genetic FCM Algorithm

The cluster obtained from fuzzy C-means clustering [6] is considered as input to our refinement algorithm. Initially a random point is selected from each cluster; with this a chromosome is build. Like this an initial population with 10 chromosomes is build. For each chromosome the fitness value is calculated by using new fitness function. With this initial population, the genetic operators such as reproduction, crossover and mutation are applied to produce a new population. While applying crossover operator, the cluster points will get shuffled means that a point can move from one cluster to another. From this new population, we apply mutation process to find the optimized seed value to generate the final clusters. This process is repeated for N number of iterations. The fitness function of algorithm is determined by f(x).

$$F(x) = \{(\alpha + 2\beta) - \alpha i,\ \alpha i < \beta + 2\alpha \quad 0,\ \alpha i \geq \alpha i + 2\beta\}$$

$$i = 1, 2, \ldots\ldots\ldots\ldots\ldots\ldots\ldots, N$$


Input web log data
Generate population of seed
Select seed randomly
Calculate fitness function and check fitness
Select seed and member function
Cluster genration

a) Flow chart

4. Experiment Results

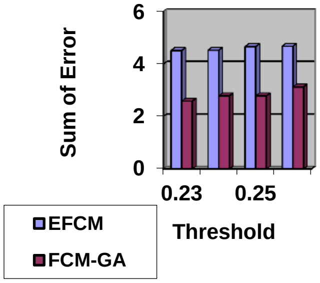
Improved Genetic FCM (FCM-GA) is applied number of times from different groups as the beginning, to obtain the number of the optimum clustering centers and the initial value of clustering centers .We compare the above final result with Entropy based FCM(EFCM) [7] by using sum of error and threshold. In the result of discriminated clustering, 90 percent are better than that using Entropy based FCM alone. The sum of error is defined as

$$SE = \sum_i^C (X_i - Y_i)$$

Where, *C* is the number of clusters that in this problem is equal to number of classes. X_i is the number of instances that belong to cluster *i*. and Y_i is the numbers of instances from cluster i that belong to a same class and also have maximum presence in cluster i.

Entropy based FCM:

EFCM [8] use information entropy to initialize the cluster centers to determine the number of cluster centers. it can be reduce some errors, and also can improve the algorithm introductions an weighting parameters .after that, combine with the merger of ideas, and divide the large chumps into small clusters. Then merge various small clusters according to the merger of the conditions, so that you can solve the irregular datasets clustering.



b) Comparison Graph

5. Conclusion

In this paper, an approach of the fuzzy clustering based on improved genetic algorithm is proposed, to a certain degree, which overcomes the defects that FCM is sensitive to initial value and it is easily able to be trapped in a local optimum. The practicability of this approach is analyzed in principle, and its practical effect is confirmed by experiment in technical. The experiment results show that the global distribution characteristic of the space clustering centers which are found during the process of the fuzzy clustering analysis by utilizing improved GA is suitably kept, so clustering effect is more rational.

References

[1] Athman Bouguettaya "On Line Clustering", IEEE Transaction on Knowledge and Data Engineering, Volume 8, No. 2, April 1996.

[2] D. Beeferman, A. Berger, Agglomerative clustering of search engine query log, KDD 2000.

[3] Qingtian Han, Xiaoyan Gao , Wenguo Wu, "Study on Web Mining Algorithm Based on Usage Mining", Computer- Aided Industrial Design and Conceptual Design, 2008.

[4] Yi Dong, Huiying Zhang, Linnan Jiao, "Research on Application of User Navigation Pattern Mining Recommendation", Intelligent Control and Automation, 2006. WCICA 2006, the Sixth World Congress, Volume 2.

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

[6] Mohanad Alata, Mohammad Molhim, and Abdullah Ramini, "Optimizing of Fuzzy C-Means Clustering Algorithm Using GA", World Academy of Science, Engineering and Technology 39, 2008.

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

[8] 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 Vol.58 No.2, pp.222-230 (2011).

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.