



Web Log Mining Based on Improved FCM Algorithm using Multi-objective Genetic Algorithm

Ranu Singhal¹, Nirupama Tiwari²

Shri Ram College of Engineering And Management Banmore^{1,2}
Gwalior INDIA

Abstract: Web log mining is important area of research in concern of user behavior, web cache improving and web security concern. For the purpose of mining various algorithms are used such as clustering, classification and rule mining technique in web log files. In this paper we discuss we log mining using improved fuzzy clustering algorithm using multi-objective genetic algorithm. For the optimization of cluster seed and controlling iteration of clustering technique we used multi-genetic algorithm. The proposed method minimized the error rate and generate accurate cluster for provided web log files. For the experimental process we used real time web log files and compared our technique with FCM algorithm. Our experimental evaluation result shows better clustering performance in compression of FCM algorithm.

Keywords: Web logs, clustering, FCM, MOGA

I. INTRODUCTION

Web log mining refers to behavior analysis of user and rating of website. Web log mining process implied in web usage mining concern. For the web log mining, log files of web passes through various state of data preprocessing such as data cleaning, data transformation and data validation. All these process implies the uncontrolled event in clustering process. 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 items (the data set) into groups called clusters such that items in the 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. As an important web data processing technology, clustering analysis has been widely applied to various fields. Because many of the real world objects or the things themselves are uncertain, people have to use fuzzy theory to study the objections or the things, and we called this study to fuzzy clustering analysis. 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 its 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. The rest of paper is organized as follows. In Section II discuss FCM algorithm. The Section III discusses proposed method IV experimental result followed by a conclusion in Section V.

II. FCM ALGORITHM AND RELATED WORK

Fuzzy clustering plays an important role in solving problems in the areas of pattern recognition and web log mining. A variety of fuzzy clustering methods have been proposed and most of them are based upon distance criteria [6]. One widely used algorithm is the fuzzy c-means (FCM) algorithm. It uses reciprocal distance to compute fuzzy weights. A more efficient algorithm is the new FCFM. It computes the cluster center using Gaussian weights, uses large initial prototypes, and adds processes of eliminating, clustering and merging. In the following sections we discuss and compare the FCM algorithm and FCFM algorithm. The fuzzy c-means (FCM) algorithm was introduced by J. C. Bezdek [2]. The idea of FCM is using the weights that minimize the total weighted mean-square error:

$$J(w_{qk}, z^{(k)}) = \sum_{(k=1,K)} \sum_{(k=1,K)} (w_{qk}) \|x^{(q)} - z^{(k)}\|^2 \quad (1)$$

$$w_{qk} = \frac{\sum_{(k=1,K)} (w_{qk})}{\sum_{(k=1,K)} (w_{qk})} = 1 \text{ for each } q$$
$$w_{qk} = (1/(D_{qk})^2)^{1/(p-1)} / \sum_{(k=1,K)} (1/(D_{qk})^2)^{1/(p-1)}, p > 1 \quad (2)$$

The FCM allows each feature vector to belong to every cluster with a fuzzy truth value (between 0 and 1), which is computed using Equation (4). The algorithm assigns a feature vector to a cluster according to the maximum weight of the feature vector over all clusters.

III. MODIFIED FCM ALGORITHM

Genetic algorithms are inspired by Darwin's theory about evolution. Solution to a problem solved by genetic algorithms is evolved. Algorithm is started with a set of solutions (represented by chromosomes or also called string) called population. Solutions from one population are taken and used to form a new population. This is motivated by a hope, that the new population will be better than the old one.

The process of k-means clustering and multi-objective function takes several process such as population, fitness function, mutation and crossover for propagation of clustering algorithm. Some steps are divided into six phase. Initializing population

The initial population is done by determining the length of chromosome with size $K \times d$. K is the number of chromosome a lot of d while d is the dimension of the cluster variables

Fitness function

K-Means clustering optimization with multi-objective genetic algorithm uses 2 objectives, i.e. minimizing error functions within each cluster (equation 1) and maximizing the centroid value between cluster (equation 2). The calculations used as follows.

$$\sigma_i^2 = \frac{1}{N} \sum_{j=1}^N \sigma_{ij}^2 \dots \dots \dots 3)$$

Where $i = 1, 2 \dots K$; K is the number of cluster

$$\sigma^2 \sum_{i=1}^K \sigma_i^2 \dots \dots \dots 4)$$

Whereas:

- σ_i^2 : Error in i -th cluster
- n_i : number of data in banyak data pada i -th cluster
- x_{ij} : Data in i -cluster, j -th variable
- z_{ij} : i -cluster average in j -variable
- V : Number of variable

The first function is to minimize error average in cluster which formulated as follow:

$$V(w) = \frac{1}{k} \sum_{i=1}^k \sigma_i^2, i = 1, 2 \dots k \quad (5)$$

Whereas:

- $V(w)$: Error in cluster
- k : Number of cluster

The second function is to maximize inter-cluster error which formulated as follow:

$$V(b) = \frac{1}{k} \sum_{i=1}^k \sum_{j=1}^v (z_{ij} - z_{\square j}) \quad 6$$

dengan:

- $V(b)$: Error in cluster
- z_{ij} : i -cluster average in variable
- $z_{\square}$: Grand mean of j -th variable

A. Fitness Function with Pareto Ranking Approach

Fitness function is calculated by using pareto ranking approach, where each individual datum is evaluated by the overall population based on the non-domination concept. Next, Pareto ranking approach is done by equations 3.

$$r_2(x, t) = 1 + nq(x, t) \dots \dots \dots 7)$$

Whereas:

- $r_2(x, t)$: x -th Completion rank in t -th iteration
- $nq(x, t)$: Solution number which dominate x completion in t -th iteration

B. Crossover

The crossover is one genetic operator in generating new chromosome (offspring). In this study, single point crossover is used with crossover probability which is calculated by using equation 4.

$$pz = \frac{F(s_z)}{\sum_{z=1}^Z F(s_z)} \dots \dots \dots 8)$$

Whereas:

pz : z -th completion selection probability
 $F(s_z)$: fitness value in s_z solution

E. Mutation

Another genetic operator is mutation. This process is exploits against the possibility of modifications to the existing results. The selection process of chromosome mutations, as well as the position of a gene to be transformed will be done in a random. The number of affected offspring is determined by the mutation probability as in equation 5.

$$pk = \frac{1.5 * d_{\max}(x_n) - d(x_n, c_k) + 0.5}{\sum_{k=1}^k (1.5 * d_{\max}(x_n) - d(x_n, c_k) + 0.5)} \dots \dots \dots 9)$$

Whereas:

p_k : Mutation probability in k -th cluster
 $d(x_n, c_k)$: jarak Euclidean distant between x_n data and c_k centre point of k -th cluster.
 $d_{\max}(x_n)$: $\max_k \{d(x_n, c_k)\}$.

If k -th group is empty then $d(x_n, c_k)$ is defined as 0.

F. Elitism

The random selection doesn't guarantee the non-dominated solutions will survive in the next generation. The real step for elitism in multi-objective genetic algorithm is doubling the non-dominated solution in population P_t , henceforth will be included in the population of P_{t+1} by selecting the non-dominated solution.

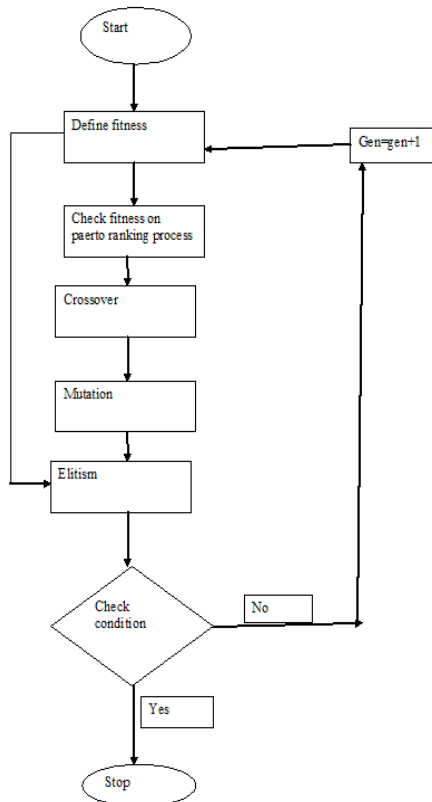


Figure 1 shows proposed model of modified working process.

IV. EXPERIMENTAL RESULT ANALYSIS

. For the evaluation of web log data used FCM and FCM_MOGA clustering algorithm implement in MATLAB 7.8.0 and found the similar pattern of cluster for ip, session, time and data. The cluster found in color group. The formation of cluster gives the information of valid and invalid cluster according to cluster valid index[13]. The clustering validity criteria are classified into internal, external, and relative. The clustering work focus on the relative association of month, grade and student relative criteria is used as the validity measure. The criteria widely accepted for partitioning a data set into a number of clusters are separation of the clusters, and their compactness. Thus these criteria are obviously good candidates for checking the validity of clustering results. The process of cluster validation defines a relative validity index, for assessing the quality of partitioning for each set of the input values. The proposal formalize clustering validity index based on clusters' compactness (in terms of cluster density), and clusters' separation (combining the distance between clusters and the inter-cluster density). The evaluation of clustering performance used some standard parameter such as number of valid cluster generation and number of cluster along with mean absolute error of clustering process. The mean absolute error process induced the error rate of clustering technique. The process of clustering used some set of data as number of instant as row in fashion of 1000, 2000, 3000 and 4000 thousand for small data to large size of data. To test the validation of cluster each cluster property assigned the color label of data the unlabeled cluster shows that invalid cluster in the process of cluster generation[7,8]. For validity of cluster and measurement of error used some standard formula given below. In clustering the mean absolute error (MAE) is a quantity used to measure how close real or predictions are to the eventual outcomes. The mean absolute error is given by

$$MAE = \frac{1}{n} \sum_{i=1}^n |f_i - y_i| = \frac{1}{n} \sum_{i=1}^n |e_i|. \quad (1)$$

As the name suggests, the mean absolute error is an average of the absolute errors $e_i = |f_i - y_i|$, where f_i is the prediction and y_i the true value.

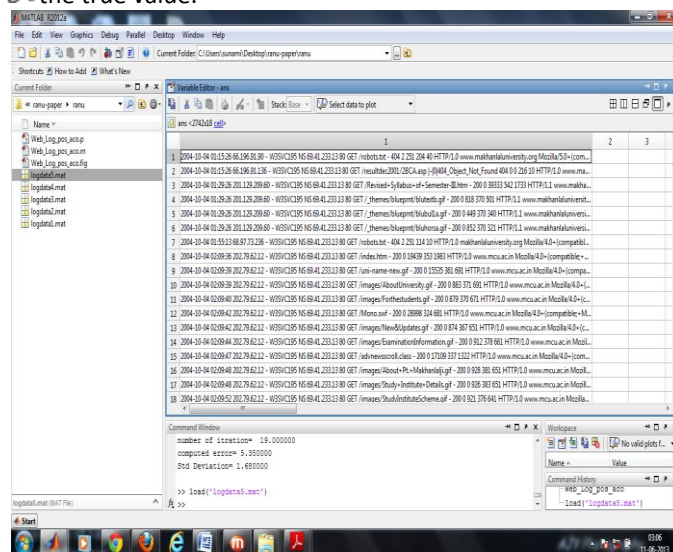


Figure 2 shows that web log data for clustering process

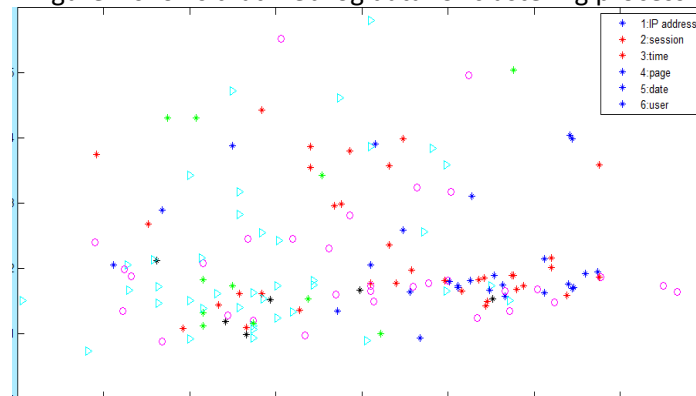


Figure 3 shows that web log cluster in color mode

Size of data	Number of cluster	Valid cluster	Error
Data1	8	6	4.607
Data2	7	6	7.890
Data3	8	8	2.304
Data4	5	4	10.34

Table 1 show that cluster generation of educational data and check number of cluster according to valid cluster

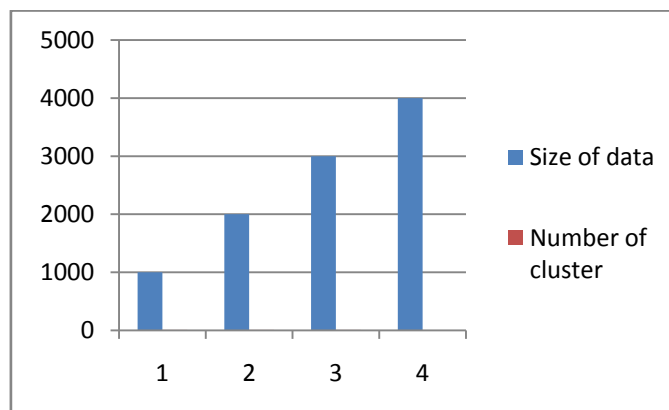


Figure 4 shows that comparative cluster generation according to data size

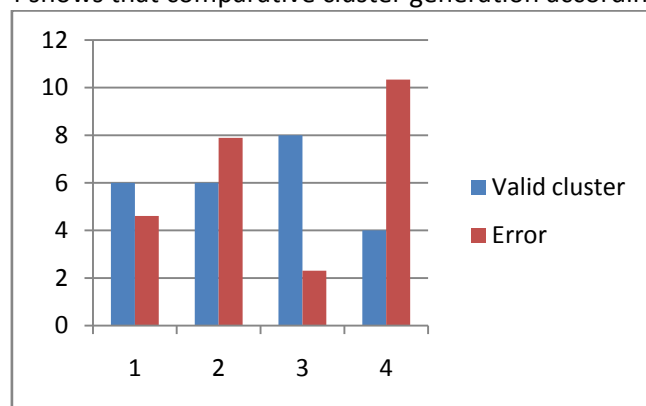


Figure 5 shows that comparative valid cluster generation and generation of error according to cluster size

V. CONCLUSION AND FUTURE WORK

In this paper we proposed a modified FCM algorithm using multi-objective genetic algorithm for web log clustering. 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.CAID/CD 2008. 9th International Conference on 22-25 Nov. 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 Journal of Computer Science Issues, Vol. 8, Issue 1, January 2011.