# Web Usage Mining through Efficient Genetic Fuzzy C-Means

**Deepak Kumar Niware**
Research Scholar
Department of Computer Science & Engg.
TIT, Bhopal (INDIA)

**Setu Kumar Chaturvedi**
Professor
Department of Computer Science & Engg.
TIT, Bhopal (INDIA)

**Abstract**
In process of knowledge discovery from any web-log dataset, most widely and extensively used clustering algorithm for this purpose is Fuzzy c-means (FCM) algorithm because the data of web-log is unsupervised dataset. Due to sensitivity of FCM, it can be easily trapped in a local optimum, and it is also depends on initialization. In this paper we present use of Genetic algorithm in Fuzzy c-means algorithm to select initial center point for clustering in FCM. The purpose of this paper is to provide optimum initial solution for FCM with the help of genetic algorithm to reduce the error rate in pattern creation.

***Keywords:***
*Fuzzy C-means, Genetic Algorithm, Web log mining, Web usage mining, Web mining.*

## 1. Introduction

Database is used for keeping huge amount of data in a formatted manner, but data can also be in unformatted manner too, therefore it is suitable to apply data mining task for making intelligent business decisions. Web usage mining is a type of web mining which deals with the log files. It is also known as Web log mining. In application of web mining like Personalization, System Improvements, Modification of Web Site, Business Intelligence, Characterization of use etc. all can only be possible through web usage mining [6]. Clustering is one of the major data mining tasks and aims at grouping the data objects into meaningful classes (clusters) such that the similarity of objects within clusters is maximized, and the similarity of objects from different clusters is minimized [1]. Cluster can be viewed as subset of dataset, on the basis of these cluster, we can classify cluster technique as : *Hard (Crisp) clustering* methods are based on classical set theory, and require that an object either does or does not belong to a cluster. Hard clustering means partitioning the data into a specified number of mutually exclusive subsets. *Fuzzy clustering* methods, however, allow the objects to belong to several clusters simultaneously, with different degrees of membership. Objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial membership. *Fuzzy c-means clustering* involves two processes: the calculation of cluster centers and the assignment of points to these centers using a form of Euclidian distance. This process is repeated until the cluster centers stabilize. The algorithm is similar to k-means clustering in many ways but it assigns a membership value to the data items for the clusters within a range of 0 to 1. So it incorporates fuzzy set’s concepts of partial membership and forms overlapping clusters to support it [2]. A *genetic algorithm* (GA) is a search technique used in computing to find exact or approximate solutions to optimization and search problems. Genetic algorithms are a particular class of evolutionary algorithms (EA) that use techniques such as inheritance, mutation, selection, and crossover. In section 2 we shows some related work on Genetic algorithm and FCM, in section 3 we discuss the problem related with FCM, in section 4 overview of proposed method, in section 5 we present experiment setup and result, in last section 6 we shows the result and conclusion.

## 2. Related Work

In [3] propose a novel hybrid genetic algorithm (GA) that finds a globally optimal partition of a given data into a specified number of clusters. They hybridize GA with a classical gradient descent algorithm used in clustering viz., K-means algorithm. Hence, the name genetic K-means algorithm (GKA). They define K-means operator, one-step of K-means algorithm, and use it in GKA as a search operator instead of crossover. They also define a biased mutation operator specific to clustering called distance-based-mutation. Using finite Markov chain theory, and prove that the GKA converges to the global optimum. It is observed in the simulations that GKA converges to the best known optimum corresponding to the given data in concurrence with the convergence result. It is also observed that GKA searches faster than some of the other evolutionary algorithms used for clustering.
In [1] present a clustering algorithm based on Genetic k-means paradigm that works well for data with mixed numeric and categorical features. They worked to modified description of cluster center to overcome the numeric data only limitation of Genetic k-mean algorithm and provide a better characterization of clusters.
Pareto-based multi objective evolutionary algorithm rule mining method based on genetic algorithms is in [5]. Predictive accuracy, comprehensibility and interestingness are used as different objectives of the association rule

Manuscript received January 5, 2014
Manuscript revised January 20, 2014

mining problem. Specific mechanisms for mutations and crossover operators together with elitism have been designed to extract interesting rules from a transaction database.

## 3. Problem Statement

The process of web usage mining model falls into four sections as source data collection phase, data pretreatment phase, pattern mining phase and pattern analysis phase is shown in Fig-1[7].

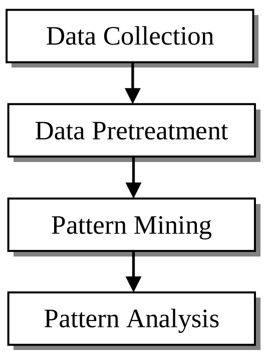



Figure -1

Pattern mining phase deals with making good clusters for the pattern analysis phase, each phase in web usage mining depend upon the previous phase for producing quality result. In this paper we are using Web log data which is huge and uncertain in nature. Due to the nature of web log data Fuzzy c-means algorithm which is inherited from k-means algorithm is used for clustering, because it is best suited for these types of data clustering. Pattern analysis is also depends on goodness of created cluster. In FCM, the cluster center which is chosen initially is not optimized solution. And pattern analysis is depended upon the cluster. The challenge is of better cluster center selection for the FCM. Because, if initial created cluster center is not optimized then rest cluster center will also not good. In this paper we proposed a Genetic Fuzzy c-mean algorithm, Genetic algorithm is used for the optimum solution for the cluster center in FCM.

## 4. Proposed Method

In this paper we proposed to combine two method Genetic algorithms which is used to local optimum solution. And other is Fuzzy c-means algorithms used for clustering in unsupervised data for knowledge discovery.

### 4.1 Genetic Algorithm

A **genetic algorithm (GA)** is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as Initialization, mutation, selection, and crossover [10].

#### 4.1.1 Initialization stage

The search space of all possible solutions is mapped onto a set of finite strings. Each string (called chromosomes) has a corresponding point in the search space. The algorithm starts with the initial solutions that are selected from a set of configurations in the search space called population using randomly generated solutions or by applying special algorithms. Each of the initial solutions (called an initial population) is evaluated using a user defined fitness function. A fitness function exists to numerically encode the performance of the chromosome.

#### 4.1.2 Selection stage

A set of individuals that have high scores in the fitness function is selected to reproduce itself. Such a selective process results in the best-performing chromosomes in the population to occupy an increasingly larger proportion of the population over time. From the selected set of individuals, some progeny is generated by applying different genetic operators (i.e. crossover, mutation).

#### 4.1.3 Crossover stage

One site crossover and two site crossover are the most common ones adopted. In most crossover operators, two strings are picked from the mating pool at random and some portions of the strings are exchanged between the strings. Crossover operation is done at string level by randomly selecting two strings for crossover operations. A one site crossover operator is performed by randomly choosing a crossing site along the string and by exchanging all bits on the right side of the crossing site as shown in Fig. 2.

| | |
|---|---|
| String1: 011\|**01100** | String1: 011\|*11001* |
| String2: 011\|*11001* | String2: 011\|**01100** |
| Before Crossover | After Crossover |

Figure 2: One site crossover operation

| | |
|---|---|
| String1: 011\|**011**\|00 | String1: 011\|*110*\|00 |
| String2: 011\|*110*\|01 | String2: 011\|**011**\|01 |
| Before Crossover | After Crossover |

Figure 3: Two-site crossover operation

In one site crossover, a crossover site is selected randomly (shown as vertical lines). The portion right of the selected site of these two strings is exchanged to form a new pair of strings. The new strings are thus a combination of the old strings. Two site crossover is a variation of the one site crossover, except that two crossover sites are chosen and the bits between the sites are exchanged as shown in Fig. 3. One site crossover is more suitable when string length is small while two site crossover is suitable for large strings. The underlying objective of crossover is to exchange information between strings to get a string that is possibly better than the parents.

### 4.1.4 Mutation stage

Mutation operates on a single chromosome: one element is chosen at random from the chain of symbols, and the bit string representation is changed with another one [11].

### 4.1.5 Termination

The terminating condition of algorithm can be controlled by the convergence degree of solution, and the inheritance can be controlled by the evolution algebra.

### 4.2 Fuzzy c-means Clustering

Fuzzy C-Mean (*FCM*) is an unsupervised clustering algorithm that has been applied to wide range of problems involving feature analysis, clustering and classifier design. One of the widely used clustering methods is the Fuzzy c-means (FCM) algorithm developed by Bezdek [9]. Fuzzy c-means partitions set of n objects $o = \{o_1, o_2, .., o_n\}$ in $R^d$ dimensional space into c $(1 < c < n)$ fuzzy clusters with $Z = \{z_1, z_2, ..., z_c\}$ cluster centers or centroids. The fuzzy clustering of objects is described by a fuzzy matrix $\mu$ with *n* rows and *c* columns in which *n* is the number of data objects and *c* is the number of clusters. $\mu_{ij}$, the element in the *i*th row and *j*th column in $\mu$, indicates the degree of association or membership function of the *i*th object with the *j*th cluster. The characters of $\mu$ are as follows [8]:

$$\mu_{ij} \in [0,1] \quad \forall i = 1,2, \dots, n; \ \forall j = 1,2, \dots, c \tag{1}$$

$$\sum_{j=1}^{c} \mu_{ij} = 1 \ \ \forall i = 1,2, \dots, n \tag{2}$$

$$0 < \sum_{i=1}^{n} \mu_{ij} < n \ \ \forall j = 1,2, \dots, c \tag{3}$$

The objective function of FCM algorithm is to minimize the Eq.(4):

$$J_m = \sum_{j=1}^{c} \sum_{i=1}^{n} \mu_{ij}^{m} d_{ij} \tag{4}$$

Where

$$d_{ij} = \left\| o_i - z_j \right\| \tag{5}$$

in which, m (m > 1) is a scalar termed the weighting exponent and controls the fuzziness of the resulting clusters and $d_{ij}$ is the Euclidian distance from object $o_i$ to the cluster center $z_j$. The $z_j$, centroid of the *j*th cluster, is obtained using Eq. (6).

$$z_j = \frac{\sum_{i=1}^{n} \mu_{ij}^{m} o_i}{\sum_{i=1}^{n} \mu_{ij}^{m}} \tag{6}$$

The FCM algorithm is iterative and can be stated as follows:

**Algorithm**: *Fuzzy c-means*

**Step 1.** Select m (m > 1); initialize the membership function values $\mu_{ij}$, *i* = 1, 2, ... , *n*; *j* = 1, 2, . . . , *c*.

**Step 2.** Compute the cluster centers $z_j$, *j* = 1, 2, ... , *c*, according to Eq. (6).

**Step 3.** Compute Euclidian distance $d_{ij}$, *i* = 1, 2,....... , *n*; *j* = 1, 2, ... , *c*.

**Step 4.** Update the membership function $\mu_{ij}$, *i* = 1, 2, .. , *n*; *j* = 1, 2, ... , *c* according to Eq. (7).

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{m-1}}} \tag{7}$$

**Step 5.** If not converged, go to step 2.

Several stopping rules can be used. One is to terminate the algorithm when the relative change in the centroid values becomes small or when the objective function, Eq. (4), cannot be minimized more. The FCM algorithm is sensitive to initial values and it is likely to fall into local optima.

## 5. Experiment and Result

Web log dataset used in this experiment is Microsoft Server log file, having 22 attributes in it. We have tested proposed method on two web log dataset having 259 KB and 559 KB size. We experimented on Pentium Dual Core 1.80 GHz and 1GB RAM with 160 GB HDD machine having Window XP Service Pack 3 with MATLAB Version 7.8. Result table and graphs are as follows.

| Method | Threshold | Error Rate | Time | Iteration |
|---|---|---|---|---|
| FCM | 0.1 | 50.391234 | 59.191740 | 2000 |
| FCM | 0.2 | 51.036754 | 62.939058 | 1000 |
| FCM | 0.3 | 51.682274 | 62.104490 | 667 |
| FCM | 0.4 | 52.327794 | 62.372756 | 500 |
| FCM | 0.5 | 52.973314 | 62.084189 | 400 |

| | | | | |
|---|---|---|---|---|
| **FCM** | 0.6 | 53.618834 | 62.155665 | 333 |
| **FCM** | 0.7 | 54.264354 | 61.654710 | 286 |
| **FCM** | 0.8 | 54.909874 | 60.510172 | 250 |
| **FCM** | 0.9 | 55.555394 | 60.662062 | 222 |
| **GFCM** | 0.1 | 19.317537 | 61.104803 | 2001 |
| **GFCM** | 0.2 | 19.539618 | 63.751756 | 1001 |
| **GFCM** | 0.3 | 19.761700 | 64.026603 | 668 |
| **GFCM** | 0.4 | 19.983781 | 63.583065 | 501 |
| **GFCM** | 0.5 | 20.205862 | 64.041465 | 401 |
| **GFCM** | 0.6 | 20.427943 | 63.856052 | 334 |
| **GFCM** | 0.7 | 20.650024 | 62.092531 | 287 |
| **GFCM** | 0.8 | 20.872105 | 63.207093 | 251 |
| **GFCM** | 0.9 | 21.094186 | 62.714621 | 223 |

**Table 1: Analysis Report of FCM and GFCM with weblog1 dataset**

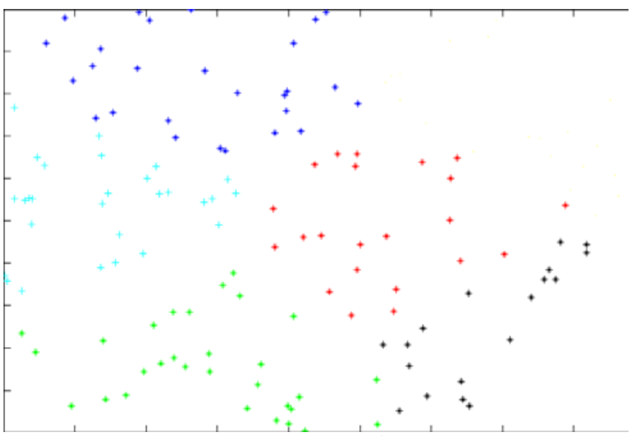


Figure 4: FCM Method of Dataset 1

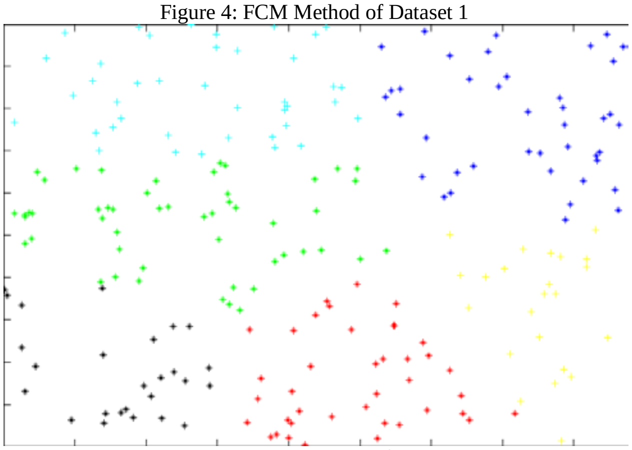
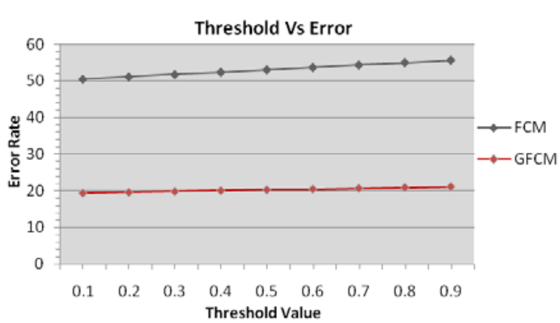
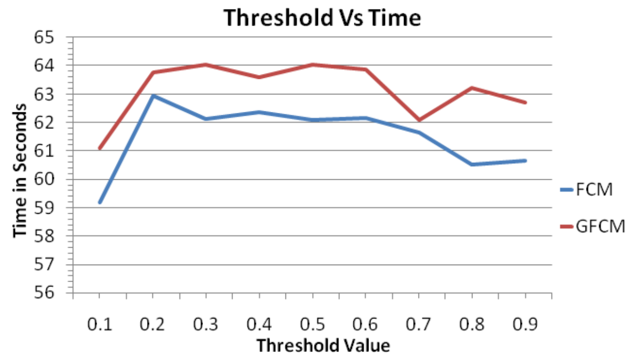


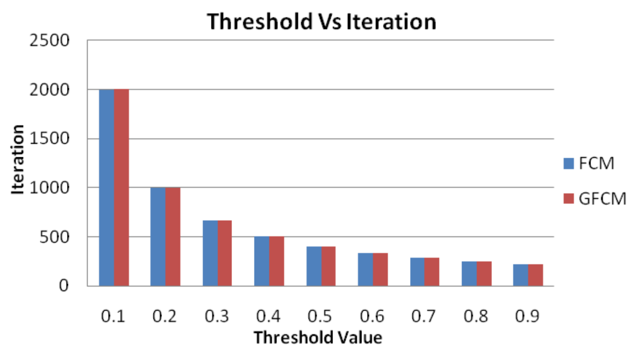
Figure 5: GFCM Method of Dataset 1




Graph 1.1: Threshold Vs Error of Dataset 1.




Graph 1.2: Threshold Vs Time of Dataset 1.




Graph 1.3: Threshold Vs Iteration of Dataset 1.

## 6. Conclustion

In the above shown graphs, in the FCM method as the value of threshold value increases the error rate also increases. In proposed GFCM method also error rate increased as the threshold value increases. But it is clearly shown in graph, if we comparing both methods that at same value of threshold value GFCM reduce the error rate more than 50% of the error rate of FCM. In graph of FCM, increases rate of error rate is high but in graph of GFCM, the increase rate of error rate is much low as compared to FCM. It shows that data loss in GFCM is low and the content of cluster is increased which is important parameter of evaluation. The other parameter of evaluation like time it is also shown in graph that time is varying because time is dependent upon CPU time other processes and time rate of increases and decrease is also not depend on threshold value. And in iteration parameter it is approximately same in both methods. Graphs, experiment and analysis concludes that GFCM is more efficient method for pattern recognition and cluster creation in web usage mining.

**References**


[1] Dharmendra K Roy , Lokesh K Sharma; "Genetic k-means clustering algorithm for mixed numeric and categorial datasets"; In proceeding of "International Journal of Artificial Intelligence & Applications (IJAIA)" Vol.1,No.2, page 23-28,2010.

[2] Raju G, Binu Thomas, Sonam Tobgay , Th. Shanta Kumar; “Fuzzy Clustering Methods in Data Mining-A comparative Case Analysis”; In “International Conference on Advanced Computer Theory and Engineering”, IEEE, page: 489-493, 2008.

[3] K. Krishna, M. Narasimha Murty; “Genetic K-Means Algorithm”; In “IEEE Transactions on systems, man, and cybernetics-part b: cybernetics” Vol. 29, No. 3, page: 433-439, JUNE 1999.

[4] Dharmendra K Roy, Lokesh K Sharma;"Genetic k-Means clustering algorithm for mixed numeric and categorical datasets"; In "International Journal of Artificial Intelligence & Applications"; Vol.1, No.2, page: 23-28, April 2010.

[5] Peter P. Wakabi-Waiswa, Venansius Baryamureeba; "Extraction of interesting association rules using genetic algorithms”; In "International Journal of Computing and ICT Research", Vol. 2 No. 1, Page: 26-33, June 2008.

[6] “Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan”; “SIGKDD Explorations 2000” ACM SIGKDD, Volume 1, Issue 2 - page 12-23, Jan 2000.

[7] Ya-Xiu Yu, Xin-Wei Wang; "Web Usage Mining Based on Fuzzy Clustering"; In "International Forum on Information Technology and Applications"; IEEE Computer Society, page: 268-271, 2009.

[8] H. Izakian, A. Abraham;"Fuzzy C-means and fuzzy swarm for fuzzy clustering problem"; In "Expert Systems with Applications 38(2011)"; Elsevier, page: 1835-1838, 2011.

[9] J. Bezdek, "Fuzzy Mathematics in Pattern Classification," Ithaca: Cornell University, 1973.

[10] http://en.wikipedia.org/wiki/Genetic_algorithm.

[11] Mu-Jung Huang, Hwa-Shan Huang, Mu-Yen Chen; "Constructing a personalized e-learning system based on genetic algorithm and case-based reasoning approach"; In "Expert Systems with Applications (2006)" Elsevier, 2006.

[12] T. Velmurugan, T. Santhanam; "Performance Evaluation of K-Means and Fuzzy C-Means Clustering Algorithms for Statistical Distributions of Input Data Points"; In "European Journal of Scientific Research" Vol.46 No.3, page: 320-330, 2010.