International Journal on Emerging Technologies **2**(2): 51-55(2011)

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

Efficient Clustering Algorithm to Discover User Pattern Applying on Weblog Data

Ravindra Mangal* and Akash Saxena**

**Department of Physics, Maharaja Ganga Singh University*
***Deepshikha College of Technical Education, Jaipur, (RJ)*

(Received 9 August, 2011, Accepted 14 September, 2011)

ABSTRACT : WWW has becomes today not only an accessible and searchable information source but also one of the most important communication channels. One of the key steps in Knowledge Discovery in Databases is to create a suitable target data set for the data mining. Clustering of data is a method by which large sets of data is grouped into clusters of smaller sets of similar data.K-means clustering algorithm suffers from two major shortcomings, right value of clusters (*k*) are initially unknown and effective selections of initial seed are also difficult. In this dissertation, the efficient algorithm is proposed which overcomes initial seed problem and also the validation of cluster problem. The comparison is performed between proposed fuzzy clustering algorithm, fuzzy C-means and K-mean clustering algorithm on web log dataset to test its accuracy and efficiency.

Keywords: K-means, Initial seed, Validation, fuzzy clustering, Efficiency.

I. INTRODUCTION

Classification is a basic human conceptual activity.Children learns very early in their lives to classify the objects in their environment and to associate the resulting classes with nouns in their languages. "Cluster analysis" is the generic name for a wide variety of procedures that can be used to create classification. These procedures empirically, form "clusters" or group of highly similar entities. Often similarity is assessed according to a distance measure. Clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics.

Many Clustering algorithms have been developed for finding useful patterns in datasets, in which Fuzzy C-means is the simplest and easiest to implement. Fuzzy C-means is the most celebrated and widely used clustering technique. Despite of its popularity for general clustering K-means suffers from three major shortcomings: it scales poorly computationally, the number of clusters K has to be supplied by the user and the performance of K-means depends on the initial guess of partition. Many of the variants of K-means how to self generate the clusters without providing the number of clusters is not much explored at present, which is a major drawback of K-means clustering Algorithm. Another well used approach is Expectation Maximization algorithm (EM) and fuzzy K-means (FKEM). The Expectation Maximization algorithm is the most frequently used technique for estimating class conditional probability density functions (PDF) in both univariate and multivariate cases and in another technique, fuzzy logic applied with expectation maximization and gives optimal clusters and provides better results and Soft computing is a collection of new techniques in artificial intelligence, which exploit the tolerance for imprecision, uncertainty and partial truth to achieve tractability, robustness and low solution cost.

This paper proposed an algorithm that eliminates the drawback of K-means clustering algorithm. Validation of clustering is also a challenging task which improves the efficiency of clustering so this algorithm will also find the generated clusters are valid or not.

II. PREVIOUS WORK

A lot of algorithms have been developed that suit specific domains. K-means clustering algorithm is a simple and fast approach to classify the data sets. Initially we have large number of seeds. After those samples are assigned to each cluster based on its distance from the seed (centroid). The centroid is computed for each set and the data points are reassigned. The algorithm runs until it converges or until desired number of cluster is obtained.

Several variations and improvements of original K-means algorithm have been done means algorithm of Macqueen is widely used for its simplicity; Forgy algorithm shows convergence to a local minimum. Here convergence depends on the initial clustering and no guarantee for optimal clustering. Fuzzy logic is based on human reason of approximation. Fuzzy logic is used in problems where the results can be approximate rather than exact. Hence, the principles of fuzzy logic suit well to clustering problems. Clustering problems measure some kind of closeness between similar objects. Fuzzy logic has been widely used in various fields to provide flexibility to classical algorithm. Fuzzy C-means is the earlier well known approach to classify the data using fuzzy. 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 incorporates fuzzy set's concepts of partial membership and forms overlapping clusters to support it. It assigns membership value to the data items for the clusters within a range of 0 to 1. The

algorithm needs a fuzzification parameter *m* in the range [1, *n*] which determines the degree of fuzziness in the clusters. When *m* reaches the value of 1 the algorithm works like a crisp partitioning algorithm and for larger values of m the overlapping of clusters is tend to be more. The algorithm calculates the membership value μ as:

$$\mu_j(x_i) = \frac{\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{p}\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}$$

where $\mu_j(x_i)$ is the membership of x_i in j^{th} cluster

d_{ji} is the distance of x_i in C_j

p is the number of specified cluster

m is fuzzification parameter

And new cluster centers are calculated with the fuzzy membership values as:

$$C_j = \frac{\sum_i [\mu_j(x_i)]^m x_i}{\sum_i [\mu_j(x_i)]^m}$$

The main restriction is that the sum of membership values of a data point xi in all the clusters must be one and this tends to give high membership values for the outlier points, so the algorithm has difficulty in handling outlier points. Secondly, the membership of a data point in a cluster depends directly on its membership values in other cluster centers and this sometimes happens to produce unrealistic results.

In fuzzy C-means method a point will have partial membership in all the clusters. The third limitation of the algorithm is that due to the influence (partial membership) of all the data members, the cluster centers tend to move towards the center of all the data points. The fourth constraint of the algorithm is its inability to calculate the membership value if the distance of a data point is zero.

Fuzzy C-means algorithm is a most popular fuzzy clustering algorithm. Many approaches have been proposed to improve the performance of the algorithm.

1. C-means with Modified Distance Function: Frank Klawonn and Annette Keller have proposed a modified C-means algorithm with new distance function which is based on dot product instead of the conventional Euclidean distance. This method aims at identifying clusters with new shapes. With this modified C-means membership function, the fuzzy clustering algorithm can form clusters into their natural shapes.

2. Modified C-means for MRI Segmentation: Lei Jiang and Wenhui Yang presented a new approach for robust segmentation of Magnetic Resonance images (MRI) that have been corrupted by intensity in homogeneities and noise. The algorithm is formulated by modifying the objective function of the standard fuzzy C means (FCM) method to compensate for intensity in homogeneities.

3. Adaptive Fuzzy Clustering

   The adaptive fuzzy clustering algorithm is a modified version of the C-means clustering and it is proposed by Krisnapuram and Keller. The adaptive fuzzy clustering algorithm is efficient in handling data with outlier points. In comparison with C-means algorithm, it gives only very low membership for outlier points .Since the sum of distances of points in all the clusters involves in membership calculation, this method tends to produce very less membership values when the number of clusters and points increases and this is the main limitation of it.

4. In reformed fuzzy C-means a neighbourhood influence parameter γ at each pixel is calculated. The probabilistic constraint is removed by equating sum of membership function in a cluster to n.

$$C\sum_{n=1}(u,k) = n$$

Fuzzy K-means improves the basic K-means version with fuzzy. This method is divided into steps. In this method K-means is performed at the first step and fuzzy maximum likelihood is estimated at the second step. EM is an iterative algorithm that is used in problems where data is incomplete or missing. It is widely used in computer vision, speech processing and pattern recognition. The Expectation Maximization algorithm is the most frequently used technique for estimating class conditional probability density functions (PDF) in both univariate and multivariate cases. EM starts with an initial estimate for the missing variables and iterates to find the maximum likelihood (ML) for these variables. Maximum likelihood methods estimate the parameters by values that maximize the sample's probability for an event. EM algorithm is an iterative technique with four major steps. The first step is initializing the hidden variables. The second step estimates the unobserved variables with respect to known variables. In the third step we compute the maximum likelihood for unobserved data and then finally check for the stop condition. The frequent references to log-likelihoods in connection with maximum likelihood estimation are a computational short-cut, rather than a theoretical necessity. This is because maximum likelihood estimates involve the multiplication of many small probabilities which, because of rounding errors, may get truncated to zero. Let X = (x_1, …., x_n) be a random vector and $\{f_x(x/\theta) : \theta \in \Theta\}$ a statistical model parameterized by θ = (θ_1, …., θ_k), the parameter vector in the parameter space Θ. The likelihood function is a map $L : \Theta \to R$ given by

$$L(\theta/x) = f_x(x/\theta)$$

In other words, the likelihood function is functionally the same in form as a probability density function. However, the emphasis is changed from the *x* to the θ. The *pdf* is a

function of the *x*'s while holding the parameters θ's constant; *L* is a function of the parameters θ's, while holding the *x*'s constant. When there is no confusion, $L(\theta/x)$ is abbreviated to be $L(\theta)$.

The parameter vector $\hat{\theta}$ such that $L(\hat{\theta}) >= L(\theta)$ for all $\theta \in \Theta$ is called a maximum likelihood estimate, or MLE, of θ.

Many of the density functions are exponential in nature; it is therefore easier to compute the MLE of a likelihood function L by finding the maximum of the natural log of *L*, known as the log-likelihood function:

$$l(\theta/x) = \ln[L(\theta/x)]$$

III. APPROACH USED

This proposed clustering approach self generates the right number of cluster and minimizes the sum of square error after that find the generated clusters are valid or not using log likelihood function. After that test to efficiency and performance of this proposed algorithm apply on web log data.

Steps of Proposed work

Step 1: Resource extraction is the process of retrieving the desired web log data files from the web server. These access log files contain information in CERN (Common Log Format).

In this phase of our work:

1. Extract web log data file from web server which contains some common fields:

- User's IP address.
- Access date and time.
- Request method (GET or POST).
- URL of the page accessed.
- Transfer protocol (HTTP 1.0, HTTP 1.1,).
- Success of return code.
- Number of bytes transmitted.

2. Give this web log data file as an input to web log parsing tool: Web Log Explorer.

Web Log Explorer takes your log file, parse it and build reports by grouping or filtering the extracted data. Then we explore page view request in resulting table and export this report in CSV File (Excel).

3. Select task relevant data from excel file. This task relevant data contain 2 fields:

- Web pages.
- Frequency of web page.

Step 2: K-mean algorithm work on only numerical data so mapping is performed on web pages. And assign every web page to unique numerical value.

Step 3: After the Step 2, proposed clustering algorithms are performed. This clustering algorithm is applied on Data set of the bases of feature vector and fuzzy parameter. Proposed algorithm is completed in two stages: in the first stage fuzzy parameter is generated from the feature vector and according to that fuzzy parameter, number of cluster and their centroid is generated in the second stage according the clusters similarity they merge into a one cluster.

Step 4: After the Step 3, cluster validation phase is performed which will find the number of generated clusters are valid or not.

Proposed Algorithm

Step 1: Initialization of Object in a Gaussian Dataset.

(*a*) $X = \{X_1, X_2, \ldots X_n\}$ be a set of data objects.

(*b*) $(P_1, P_2 \ldots P_j)$ is the nature of data.

(c) Domain of nature is defined as:

$$\text{DOM}(P_j) = (P_1, P_2 \ldots P_j)$$

K_{th} object is $X_k (1 \le k \le n)$

$X_k = (X_{k1}, X_{k2} \ldots\ldots X_{kp})$

$X_{kj} \in \text{DOM}(P_j) (1 \le j \le P)$

Step 2: Finding dissimilar center of cluster.

(*a*) Initialize a generating function $M(X_{kj}, X_{lj})$

Here *D* is the dissimilar centre of cluster.

Here X_l and X_k are two different parameterized nature objects.

$D(X_k, X_l)$ = For $J = 1$ to P

If (X_k, X_l) then

$D(X_k, X_l) = 0$

Else

$D(X_k, X_1) = C_{\text{new}}$

Step 3: Initialization of Fuzzy parameter.

(*a*) Set nature of objects $S(1, \infty)$

F_m is a fuzzy Parameter *i.e.* $F_m \varepsilon S$

Initialize *V* is the Value of Max Cluster.

If F_m = V then

Reset $F_m \varepsilon (1, \infty)$

(*b*) Find the membership Function on the basis of fuzzy parameter on the given data and find the centroid of cluster.

Let *n* is a number of objects and *C* is a number of Clusters.

X_k is the K_{th} object

α_i is the ith center of cluster

β_{ik} is the membership value.

$U = \beta_{ik}$ and $V = \alpha_i$

$$\gamma = \frac{D(X_k, \alpha_i)}{D(X_k, \alpha_j)}$$

$$\beta_{ik} = (\gamma)\frac{1}{2m-1}$$

E_m is the estimation function; it is derived on the basis of value of β_{ik} and the value of α_i.

$$E_m(U,V) = \int_{k=1}^{n}\int_{i=1}^{c} (\beta_{ik})^{\frac{Fm}{2}} D(X_k, \alpha_i)$$

$$\int_{i=1}^{C} \beta_{ik} = \frac{M}{2}$$

Step 4: Find the Validity of Cluster

Set count = 0

For *i* = 1 to *n* do

For *j* = 1 to *n* do

If [ln($L(F_m'/X')$] = = ln[$L(F_m$/X)]

Then

Show "Cluster is Valid"

Else

Cluster is invalid

Count = count+1;

Show "count"

Exit

IV. RESULT ANALYSIS

Data Set

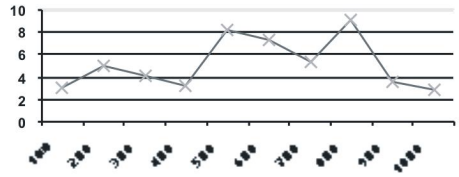

Fig. 1. Performance of K-means Algorithm.

Above graph and table shows the number of clusters developed when random number of data sets was taken. If results are taken theoretically and practically it is found that the results are different. It has been assume that whatever cluster formed is correct since this algorithm is inefficient to find the correct cluster and defomalization of cluster dataset take place. The points denoted on the graph shows fractional data set.

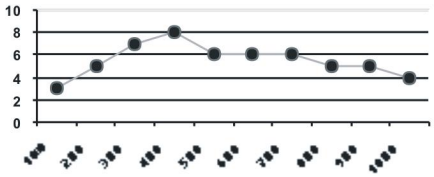

Fig. 2. Performance of fuzzy C-mean algorithm.

Table 1: Data Set and Cluster for K-means Algorithm.

Data set	100	200	300	400	500	600	700	800	900	1000
Cluster	3	5	7	8	6	6	6	5	5	4

Table 2: Data Set and Cluster for fuzzy C-mean Algorithm.

Data set	100	200	300	400	500	600	700	800	900	1000
Cluster	3	5	4.1	2	8.3	7.4	5.4	3.1	3.6	2.9

Above graph and table shows the performance of fuzzy C-means algorithm on random number of data sets. When membership value reached maximum so performance will be constant after that it will be degraded.

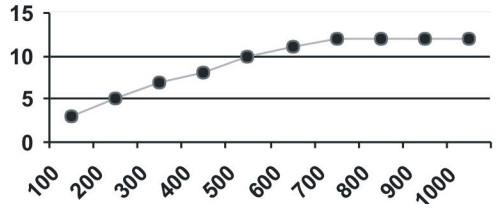

Fig. 3. Performance of Proposed algorithm.

Table 3: Data Set and Cluster for proposed Algorithm.

Data set	100	200	300	400	500	600	700	800	900	1000
Cluster	3	5	8	10	12	12	12	12	12	12

Above graph and table shows the performance of proposed algorithm by number of clusters developed when random number of data sets was taken. This algorithm generates all the clusters in non fractional format which is good symbol for efficiency in clustering.

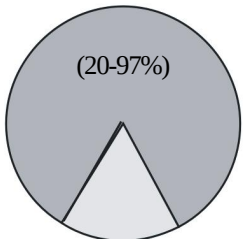

Fig. 4. Validation of Proposed algorithm.

This pie chart represents the validation of clusters in J-means algorithm; Where Cyan color shows the 90-97% of validity of generated clusters.

V. CONCLUSION

The new algorithm of clustering is proposed which efficiently overcome the major drawbacks viz. right number of cluster and initial seed (center point) problem. Proposed efficient clustering algorithm is based on two specific factors, fuzzy parameter which initially selects the random

value from the feature vector and decide the number of cluster. Second is, specific factor which merge the clusters according to the similarity. The proposed algorithm is efficiently applied on realistic web log data to mine number of web pages frequently access by the client.

REFRENCES

[1] A Modified Fuzzy K-means Clustering using Expectation Maximization Sara Nasser, Rawan Alkhaldi, Gregory Vert Department of Computer Science and Engineering, 171, University of Nevada Reno, Reno NV 89557, USA.

[2] Data Mining: Concepts and Techniques, Jiawei Han and Micheline Kamber Intelligent Database Systems Research Lab, School of Computing Science Simon Fraser University, Canada.

[3] On Clustering Validation Techniques: Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis Department of Informatics, *Athens University of Economics & Business, Patision* **76**, 10434, Athens, Greece (Hellas).

[4] An efficient algorithm to scale up k-medoid based algorithm for large databases, Maria camila N. Barioni, Humberto L. Razente, Agma J.M. Traina, Caetano Traina Jr. Computer Sciene department-ICMC/USP, Caixa postal 668-13560-970-sao carlos-sp-Brazil.

[5] Lei Jiang and Wenhui Yang, "A Modified Fuzzy C-Means Algorithm.for Segmentation of Magnetic Resonance Images" *Proc. VIIth Digital Image Computing: Techniques and Applications*, pp. 225-231, 10-12 Dec. (2003), Sydney.

[6] Ji He,Man Lan, Chew-Lim Tan, Sam-Yuan Sung, Hwee-Boon Low, "Initialization of Cluster refinement algorithms: a review and comparative study", Proceeding of International Joint Conference on Neural Networks.Budapest, (2004).

[7] E. Forgy. Cluster analysis of multivariate data: efficiency vs. interpretability of classifications. In WNAR meetings, Univ of Calif Riverside, number 768, (1965).

[8] E. Cox, Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration, Elsevier, 2005.

[9] A Modified Fuzzy C-Means Algorithm for Natural Data Exploration: Binu Thomas, Raju G., and Sonam Wangmo

[10] Frank Klawonn and Annette Keller, "Fuzzy Clustering Based on Modified Distance Measures". http://citeseer.istpsu.edu/fuzzy_clustering_62.

[11] J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, *University of California Press*, **1**: 281-297.

[12] "http://en.wikipedia.org/wiki/Expectationmaximization_algorithm.

[13] The expectation maximization algorithm, A short tutorial, Sean borman, July 05 (2008).

[14] Nong Ye: A handbook of Data Mining.

[15] Survey of Clustering Data Mining Techniques Pavel Berkhin Accrue Software, Inc.

[16] Data Clustering: A Review A.K. Jain Michigan State University M.N. Murty, P.J. Flynn Indian Institute of Science and the Ohio State University

[17] Machine Learning: Algorithms and Applications: Quang Nhat Nguyen, Faculty of Computer Science Free University of Bozen-Bolzano.

[18] Image segmentation using reformed fuzzy clustering technique: B Sowmya, B Sheelarani, CSI communication June (2009).