Research Inventy: International Journal Of Engineering And Science
Vol.4, Issue 11 (November2014), PP 51-57
Issn (e): 2278-4721, Issn (p):2319-6483, www.researchinventy.com

Web Based Fuzzy Clustering Analysis

[1]A.M. Sote, [2]Dr. S.R. Pande
[1]*A Department of Computer Science, ACS College Arvi India.*
[2]*Associate Professor and Head Department of Computer Science, SSESA's Science College,Nagpur,India*

ABSTRACT – *World wide web is a huge repository of information and there is a tremendous increase in the volume of information daily. The numbers of users are also increasing day by day. To reduce users browsing time lot of research is taken place. Clustering plays an important role in a broad range of applications like Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others. Clustering is the grouping of similar instances or objects. The key factor for clustering is some sort of measure that can determine whether two objects are similar or dissimilar. Cluster analysis is a technique for deriving natural groups present in the data. Fuzzy clustering uses membership degrees to assign data objects to clusters in order to handle uncertain data that shares properties of different clusters. Fuzzy clustering is an appropriate method since it separates the objects that are definite members of a cluster from the objects that are only possible members of a cluster. In this paper we focus on comparing and analyzing different fuzzy clustering algorithm on web data set.*
Keywords - *Fuzzy clustering, FCM, GG, GK, K-means, K-mediod.*

I. INTRODUCTION

Last two decades have witnessed an exponential growth of the Internet. This is mainly due to the great number of web applications available online and to the increasing number of their users. This has generated huge quantities of data related to the users' interactions with websites. This valuable information is stored by servers in user access log files or in web sites pages. In this respect, a number of research studies using web mining techniques have been carried out to analyze the interests and the profiles of web users so as to identify and recommend appropriate services [1], [2], [3]. This identification, also termed "profiling" is applied in several areas such as criminology, ecommerce, education, etc. In criminology, for example, detecting terrorists and racist groups is of utmost importance.

In [4] and [5], two approaches based on social networks and on the exploration of the web in general were proposed to identify terrorists and racist groups and to analyze their behaviour and profiles. Another computer profiling application concerns the identification of the personal interests of website users. A tool set that exploits neural networks and self-organizing maps (SOM) to identify customers' Internet browsing patterns is described in [6]. For their parts, [3] identified the potential customers of an online bookstore through web content mining whereas [7] provided a methodology combining a fuzzy K-means algorithm and neural networks to study Chilean bank client's behaviour.

As far as is education is concerned, the possibility of tracking user's behaviour in e-learning environments creates new possibilities for system architects, pedagogical and instructional designers to create and deliver learning contents [8]. Taking into account previous studies on profiling web users, it can be noticed that most of these works are based on analyzing access log files stored by servers and user's transaction records. However, access to these files is not always possible in all sites and as such cannot be easily extracted. To make up for this problem, the identification of profiles is based on texts available in web forums, blogs or social networks. To identify a user profile, we have to extract web messages, analyze them and detect texts written by our specific profile.In the present paper, focus is on presenting part of our research work, which is to clustering web data using different types clustering algorithms with validity measures. The rest of the paper is structured as follows. Section II explain five different types of clustering algorithms, Section III gives seven different types of validity measures applied on data, in section IV we explain experimental results and finally in section V we conclude the papers.

II. CLUSTERING ALGORITHM

The importance of clustering to Web mining, specifically in the domains of Web Content and Web Usage mining, make Web clustering an interesting topic of research. This includes clustering of Web documents, snippets and access logs. Usually the Web involves overlapping clusters. So a crisp usage of metrics is better replaced by fuzzy sets which can reflect, in a more natural manner, the degree of belongingness/membership to a cluster.

In this section we explain five different types of clustering algorithms which can be implemented on our web data as practical works. II.I K-means and K-medoid algorithms : The k-means and k-medoid algorithms are hard partitioning methods and they are simple and popular, though them results are not always reliable and these algorithms have numerical problems as well. The k-means and k-medoidalgorithms allocates each data point to one of *c* clusters to minimize the within-cluster sum of squares:

$$\sum_{i=1}^{c}\sum_{k\epsilon Ai} ||\mathrm{Xk}-\mathrm{Vi}||2 \qquad (1)$$

Where *Ai* is a set of objects (data points) in the *i*-th cluster and v_i is the mean for that points over cluster *i*. In k-means clustering v_i is called cluster prototypes.

$$v_i = \frac{\sum_{k=1}^{N_i} X_k}{N_i}, X_k \in A_i \qquad (2)$$

Where N_i is the number of objects in A_i .

In k-medoid clustering the cluster centers are the nearest objects to the mean of data in one cluster *V*= {vi ϵ *X*| $1 \le i \le c$} .

II.II Fuzzy C-means algorithm : The fuzzy c-means algorithm (FCM) can be seen as the fuzzified version of the k-means algorithm and is based on the minimization of an objective function called *c-means functional*:

$$J(\mathrm{X};\mathrm{U},\mathrm{V}) = \sum_{i=1}^{c}\sum_{K=1}^{N}(\mu_{ik})^{m}\,\|Xk-Vi\|_A^2 \qquad (3)$$

Where V = [v1;v2*; : : : ;*vc],v*i* ϵR^n is a vector of *cluster prototypes* (centers), which have to be determined, $D_{ikA}^2 = \|Xk-Vi\|_A^2 = (Xk-Vi)^T A(Xk-Vi)$is a squared inner-product distance norm, and the *N*×*c* matrix U=[*μik*] represents the fuzzy partitions, where *μik* denotes the membership degree that the i^{th} data point belongs to the k^{th} cluster. Its conditions are given by:

$$\mu_{ij}\epsilon[0,1],\forall i,k,\sum_{k=1}^{c}\mu ik = 1,\forall i,0<\sum_{i=1}^{N}\mu ik < N,\forall K \qquad (4)$$

FCM algorithm can find only clusters with the same shape and size because the distance norm *A* is not adaptive and it is often Euclidean norm (spherical clusters). The solution can be given by Lagrange multiplier method.

II.III The Gustafson- Kessel algorithm : Gustafson-Kessel algorithm (GK) is the extended version of the standard fuzzy c-means algorithm by employing an adaptive distance norm, in order to detect clusters of different geometrical shapes in one data set[9]. Each cluster has its own norm-inducing matrix A*i*. The objective function of GK algorithm is defined by

$$J(X;U,V,A) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^{m}D_{ikA_i}^2 \qquad (5)$$

The objective function cannot be directly minimized with respect to A*i*, since it is linear in A*i*. This means that *J* can be made as small as desired by simply making A*i* less positive definite. To obtain a feasible solution, A*i* must be constrained in some way. The usual way of accomplishing this is to constrain the determinant of A*i*. Allowing the matrix A*i* to vary with its determinant fixed corresponds to optimizing the cluster's shape while its volume remains constant

$$||A_i|| = \rho_i \quad , \quad \rho>0 \tag{6}$$

Where ρ_i is fixed for each cluster. Using the Lagrange multiplier method, the following expression for A_i is obtained :

$$A_i = [\ \rho_i det(F_i)\ ]^{1/n} F_i^{-1} \tag{7}$$

Where F_i is the *fuzzy covariance matrix* of the i^{th}cluster defined by –

$$Fi = \frac{\sum_{k=1}^{N}(\mu_{ik})^m (X_k - V_i)(X_k - V_i)^T}{\sum_{k=1}^{N}(\mu_{ik})^m} \tag{8}$$

GK algorithm can find clusters with different shape but with the same size.

II.IV The Gath–Geva algorithm : Gath-Geva algorithm (GG) is based on the fuzzy maximum likelihood estimation (FMLE) and it is able to detect clusters of varying shapes, sizes and densities[10].

$$D_{ik}(x_k,v_i) = \frac{\sqrt{detF_{wi}}}{\alpha_i}\left(\frac{1}{2}\left(X_k - V_i^{(l)}\right)^T F_{wi}^{-1}\left(X_k - V_i^{(l)}\right)\right) \tag{9}$$

The cluster covariance matrix is used in conjunction with an "exponential" distance, and the clusters are not constrained in volume.

$$Fwi = \frac{\sum_{k=1}^{N}(\mu_{ik})^w (X_k - V_i)(X_k - V_i)^T}{\sum_{k=1}^{N}(\mu_{ik})^w} \quad , \quad 1 \le i \le c \tag{10}$$

However, this algorithm is less robust in the sense that it needs a good initialization, since due to the exponential distance norm, it converges to a near local optimum. Using these five algorithms we can use seven different validity measures which can be explained in our next section validation.

III. VALIDATION

Validation of Cluster refers to the problem whether a given fuzzy partition fits to the data all[12]. The clustering algorithm always tries to find the best fit for a fixed number of clusters and the parameterized cluster shapes. However this does not mean that even the best fit is meaningful at all. Either the number of clusters might be wrong or the cluster shapes might not correspond to the groups in the data, if the data can be grouped in a meaningful way at all. Two main approaches to determining the appropriate number of clusters in data can be distinguished:

Starting with a sufficiently large number of clusters, and successively reducing this number by merging clusters that are similar (compatible) with respect to some predefined criteria. This approach is called compatible cluster merging [13].

Clustering data for different values of *c*, and using *validity measures* to assess the goodness of the obtained partitions. We used several indexes in our experiments and they are.

III.I Partition Coefficient (PC): It measures the amount of "overlapping" between cluster. It is defined by Bezdek[12] as follows :

$$PC(c) = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^2 \tag{11}$$

Where μ_{ij} is the membership of data point *j* in cluster *i*.

III.II Classification Entropy (CE): It measures the fuzziness of the cluster partition only, which is similar to the Partition Coefficient .

$$CE(c) = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}\ log(\mu_{ij}) \qquad (12)$$

III.III Partition Index (SC): It is the ratio of compactness and separation of the clusters. It is a sum of individual cluster validity measures normalised through division by the fuzzy cardinality of each cluster[14]

$$SC(c) = \sum_{i=1}^{c}\frac{\sum_{j=1}^{N}(\mu_{ij})^{m}\|x_j - v_i\|^2}{N_i\sum_{k=1}^{c}\|v_k - v_i\|^2} \qquad (13)$$

III.IV Separation Index (S): On the contrary of partition (SC),the separation index uses a minimum distance separation for partition validity[14].

$$S(c) = \frac{\sum_{i=1}^{c}\ \sum_{j=1}^{N}(\mu_{ij})^{2}\|x_j - v_i\|^{2}}{N min_{i,k}\|v_k - v_i\|^2} \qquad (14)$$

III.V Xie and Beni's Index (XB) : It aims to quantify the ratio of the total variation within clusters and the separation of cluster[15].

$$XB(c) = \frac{\sum_{i=1}^{c}\ \sum_{j=1}^{N}(\mu_{ij})^{m}\|x_j - v_i\|^{2}}{N min_{i,j}\|x_j - v_i\|^2} \qquad (15)$$

III.VI Dunn's Index (DI): This is originally proposed to use at the identification of "compact and well separated clusters". So the result of the clustering has to be recalculated as it was a hard partition algorithm.

$$DI(c) = min_{i\in c}\left\{min_{j\in c, i\neq j}\left\{\frac{min_{x\in C_i, y\in C_j}d(x,y)}{max_{k\in c}\{max_{x,y\in C}d(x,y)\}}\right\}\right\} \qquad (16)$$

III.VI Alternative Dunn Index (ADI) : The aim of modifying the original Dunn's index was that the calculation becomes more simple, when the dissimilarity function between two clusters ($min_{x}\in_{Ci,y}\in_{Cj}d(x,y)$) is rated in value from beneath by the train gle-non equality ;

$$d(x,y) \geq |d(y,v_j) - d(x,v_j)| \qquad (17)$$

Where v_j is the cluster center of the j[th]cluster.

$$ADI(c) = min_{i\in c}\left\{min_{j\in c, i\neq j}\left\{\frac{min_{x\in C_i, y\in C_j}|d(y,v_j) - d(x,v_j)|}{max_{k\in c}\{max_{x,y\in C}d(x,y)\}}\right\}\right\} \qquad (18)$$

We can use this seven validity measures in our experiments with five different clustering algorithms which can be explained in next section.

IV. EXPERIMENTAL RESULT

The objective of cluster analysis is the classification of objects according to similarities among them, and organizing of data into groups. Clustering techniques are among the unsupervised methods, they do not use prior class identifiers. The main potential of clustering is to detect the underlying structure in data not only for classification and pattern recognition but for model reduction and optimization.

The validity measures mentioned in Section III. Using the partitioning methods can be easily compared. In our experiment we use 300 data sets for clustering which is shown in Fig. 1,Fig. 2 Fig.3, Fig.4 and Fig.5 so the index-values are better determined at each type of clustering. We use MATLAB software with Fuzzy clustering and data Analysis Toolbox for implementation with the validity measures PC, CE, SC, S, XB, DI, ADI with different algorithms namely K-means, K-medoid, FCM, GK, GG results are collected and compared in Table 1. First of all it must be mentioned, that all these five algorithms use random initialization, so different runs issue in different partition results, i.e. values of the validation measures. On the other hand the results hardly depend from the structure of the data and no validity index is perfect by itself for a clustering problem. Several experiment and evaluation are needed that are not the proposition of this work.

Table 1. The numerical values of validity measures

	PC	CE	SC	S	XB	DI	ADI
K-means	1	Nan	0.095	0.0002	40.75	0.0152	0.0002
K-medoid	1	Nan	0.3454	0.0005	Inf	0.0048	0.0041
FCM	0.8076	0.2679	0.9791	0.0008	20.5631	0.0185	0.0126
GK	0.8516	0.2853	0.8892	0.0009	35.3245	0.0075	0.0201
GG	0.9729	0.0285	1.9431	0.004	6.2987	0.016	0.0097

In Table 1, PC and CE are not applicable for K-means and K-medoid, while they are hard clustering methods. But that is the reason for the bestresults in S, DI (and ADI), which are useful to validate crisp and well separated clusters. The Xie and Beni's index is infinity. On the score of the values of the two "most popular and used" indexes for fuzzy clustering (Partition Coefficient and Xie and Beni's Index) the Gath-Geva clustering has the very best results for this data set.

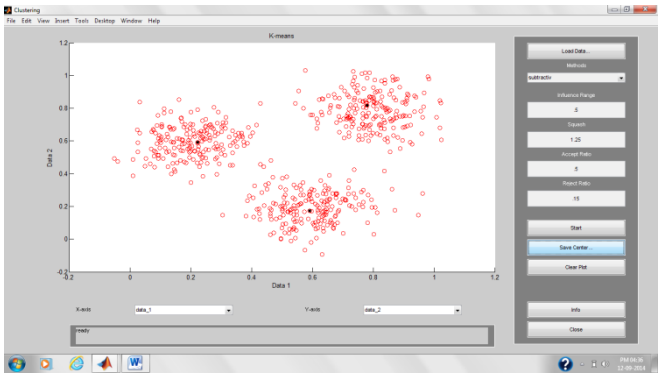

Fig.1. Result of k-means Algorithm

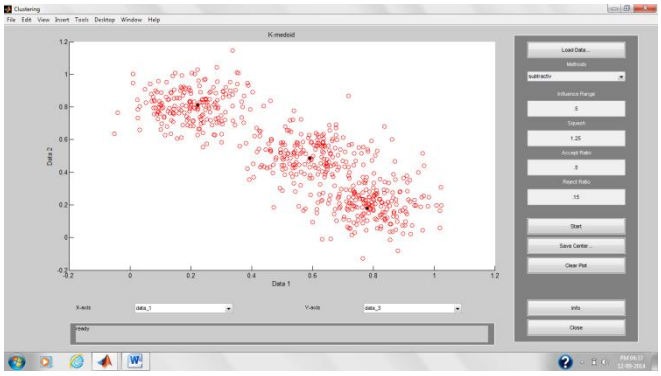

Fig.2. Result of k-medoid Algorithm

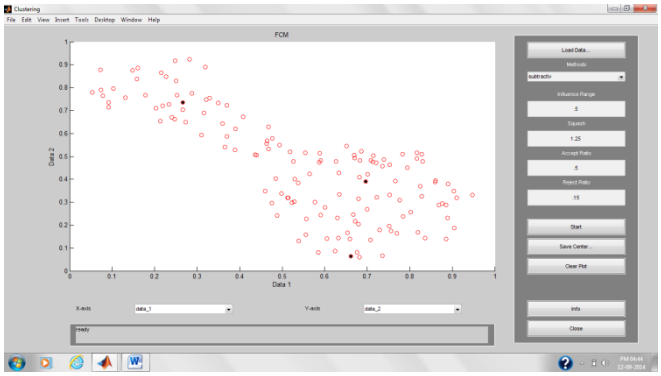

Fig.3. Result of FCM Algorithm

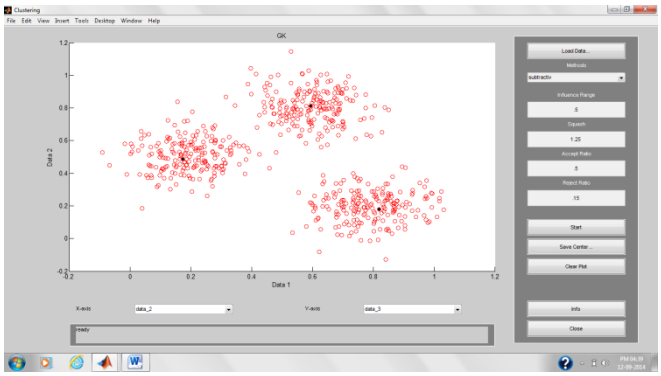

Fig.4. Result of GK Algorithm

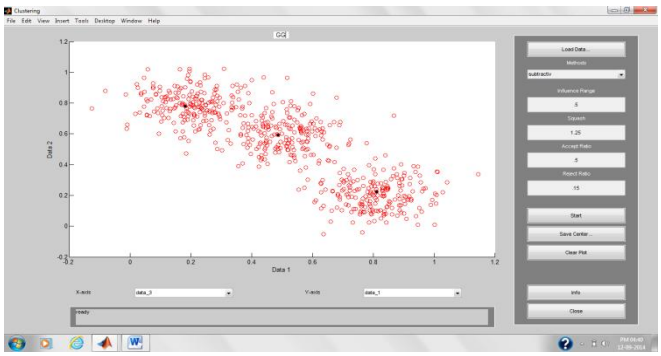

FIG.5.Result of GG Algorithm

Fig. 1 and Fig. 2 shows that hard clustering methods i.e. K-means and K-mediod which is also can find a good solution for the clustering problem, when it is compared with the figures of fuzzy clustering algorithms. On the contrary in Fig. 1 and Fig. 2 can show a typical example for the initialization problem of hard clustering. This caused the differences between the validity index values in Table 1.Fig.3, Fig.4,Fig.5 shows that fuzzy clustering methods i.e. FCM, Gustafson-Kesel, Gath-Geva algorithm respectively. This is shows better performance in clustering problem. All the seven validity measures are implemented with these hard and soft algorithms which can be fitted and secure on web data.

V. CONCLUSION

Clustering of numerical data forms the basis of many classification and system modelling algorithms. The purpose of clustering is to identify natural groupings of data from a large data set to produce a concise representation of a system's behaviour. In this paper, we have presented to clustering the web data based on fuzzy clustering algorithms such as FCM, Gustafson-Kessel algorithm, Gath-Geva algorithm with different validity measures such as PC, CE, SC, S, XB, DI, ADI.Results have proven that these algorithms seem to be very best results for this web data set shows in five figures. We were comparing the clustering methods in which we get secure result which is very reliable with these validity measures. Soft clustering algorithms shows better performance than hard clustering algorithms. The method is experimented and evaluated are found it is better method for clustering than the existing methods.

REFERENCES

[1] K. K. Chen , P. H. Chou, P. H. Li, M. J. Wu, Integrating web mining and neural network for personalized e-commerce automatic service, Expert System with applications, Vol.(37): 2898-2910, 2010

[2] Y. C. Yang. Web user behavioral profiling for user identification. Decision Support Systems, Vol.(49): 261–271.

[3] I. C. Yeh, C. H. Lien, T. M. Ting, C. H. Liu, Applications of web mining for marketing of online bookstore. Expert System with applications, Vol.(36) :11249-11256, 2009

[4] M. Chau, J. Wu , Mining communities and their relationships in blogs: a study of online hate group. Int. J. Human-Computer Studies, pp.57- 70, 2007

[5] H. Chen, W. Chung, J. Qin, E. Reid, M. Sageman, G. Weimann, Uncovering the Dark Web: A Case Study of Jihad on the Web. Journal of the American Society for Information Science and Technology, Vol.(59), Issue 8, pp: 1347–1359, 2008

[6] X. Zhang, J. Edwards, J. Harding , Personalised online sales using web usage data mining. Computers in Industry, 2007, Vol.(58): 772–782.

[7] S. Arayaa, M. Silvab, R. Weberc A methodology for web usage mining and its application to target group identification Fuzzy Sets and Systems 148 (2004) 139–152.

[8] J. M. Carbo, J. Minguillon , E. Mort , User navigational behavior in elearning virtual environments. IEEE/WIC/ACM International Conference on Web Intelligence, 2005

[9] D. Gustafson, W. Kessel, Fuzzy clustering with fuzzy covariance matrix, Proceedings of the IEEE CDC, San Diego (1979) pp. 761–766.

[10] I. Gath, A. Geva, Unsupervised optimal fuzzy clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 7 (1989) pp. 773–781.

[11] M. Setnes, Supervised fuzzy clustering for rule extraction, Proceedings of FUZZIEEE' 99, Seoul, Korea, (1999) pp. 1270–1274.

[12] B. Balasko,J.Abonyyi and B.Feil Fuzzy Clustering and Data analysis Toolbox

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

[14] A.M.Bensaid, L.O. Hall, J.C. Bezdek, L.P. Clarke, M.L. Silbiger, J.A.Arrington, and R.F. Murtagh. Validity-guided (Re)Clustering with applications to image segmentation. IEEE Transactions on Fuzzy Systems, 4:112-123, 1996.

[15] X.L. Xie and G.A. Beni. Validity measures for fuzzy clustering. IEEE Trans. PAMI, 3(8): 841-846,1991.