ISSN No. 2278-3091

Volume 2, No.2, March - April 2013
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://warse.org/pdfs/2013/ijatcse02222013.pdf

Clustering based integration of personal information using Weighted Fuzzy Local Information C-Means Algorithm

Capt. Dr. S Santhosh Baboo[1], Ms. P Shanmuga Priya[2]

[1] Associate Professor, P.G. & Research Department of Computer Science, D. G. Vaishnav College, Arumbakkam, Chennai, Tamil Nadu, India, *santhos2001@sify.com*
[2] Research Scholar, P.G. & Research Department of Computer Science, D. G. Vaishnav College, Arumbakkam, Chennai, Tamil Nadu, India, shanmusekar@gmail.com

ABSTRACT

This paper presents a variation of fuzzy c-means (FCM) algorithm that provides data clustering. The proposed algorithm incorporates the local spatial information in a novel fuzzy way. The new algorithm is called Weighted Fuzzy Local Information C-Means (WFLICM). WFLICM can overcome the disadvantages of the known fuzzy C-means algorithm and at the same time enhances the clustering performance. The major characteristic of WFLICM is the use of a fuzzy local similarity measure, aiming to guarantee noise insensitiveness and information detail preservation. Experiments performed on synthetic and real-world databases like ration card, passport and voter id show that WFLICM algorithm is effective and efficient, providing robustness to noisy data and faster retrieval of information.

Key words : Clustering, Fuzzy C-Means, Fuzzy factor Weighted Fuzzy Local Information C-Means (WFLICM)

1. INTRODUCTION

Data mining or knowledge discovery in databases (KDD) is the process of discovering useful knowledge from large amount of data stored in databases, data warehouses, or other information repositories[1]. Recently, a number of data mining applications and prototypes have been developed for a variety of domains including marketing, banking, finance, manufacturing, health care and other types of scientific fields [2, 3]. Clustering is the process of grouping a set of objects into classes of similar objects. In machine learning, clustering is an example of unsupervised learning. In general, there are two main clustering strategies: the hard clustering scheme and the fuzzy clustering scheme. The conventional hard clustering methods classify each point of the data set just to one cluster. As a consequence, the results are often very crisp, however, in many real situations, issues such as noise, inconsistent, irrelevant and incomplete data reduce the effectiveness of hard (crisp) clustering methods. Fuzzy set theory[4] has introduced the idea of partial membership, described by a membership function. Among the fuzzy clustering methods, fuzzy c-means (FCM) algorithm[5] is the most popular method used for image segmentation because it has robust characteristics for ambiguity and can retain much more information than hard segmentation methods[6]. This paper apply the fuzzy clustering technique to cluster the passport, ration card and voter id databases of Tamilnadu, by assigning weights to the attribute values and convert it into numeric data appropriate for clustering. The clustering is performed according to the query namewise, agewise, statewise or even yearwise. Although the conventional FCM algorithm works well on most noise-free data, it is very sensitive to noise and other artifacts.

2. LITERATURE REVIEW

An Introduction to data mining clustering techniques can be found in Han et al. (2001)[8]. General references about clustering included in Hartigan(1975)[12], Jain and Dubes(1988)[13], Kaufman and Rousseeuw(1990)[14], Everitt (1993)[15], Mirkin (1996)[16], Jain et al.(1999)[17], Fasulo(1999)[18] and Ghosh (2002)[19]. Clustering in data mining was brought to life by intense developments in information retrieval and text mining (Dhillon et al., 2001)[20]. Text mining research in general relies on a vector space model, first proposed by Salton (1971)[21] to model text documents as vectors in the feature space. Gibson et al. (1998)[22] presented a method that encoded datasets into a weighted graph structure where the individual attribute values correspond to weighted vertices[26].

The fuzzy C-means (FCM) clustering algorithm was first introduced by Dunn[7] and later extended by Bezdek [5]. Tolias and Panas [23] developed a fuzzy rule-based scheme called the ruled-based neighborhood enhancement system to impose spatial constraints by postprocessing the FCM clustering results. Noordam *et al.* [24] proposed a geometrically guided FCM (GG-FCM) algorithm, a semi-supervised FCM technique, where a geometrical condition determined is used by taking into account the local neighborhood of each pixel. Pham [25] modified the FCM objective function by including a spatial penalty on the membership functions. The penalty term leads to an iterative algorithm, which is very similar to the original FCM and allows the estimation of spatially smooth membership functions[27].

3. METHODOLOGY

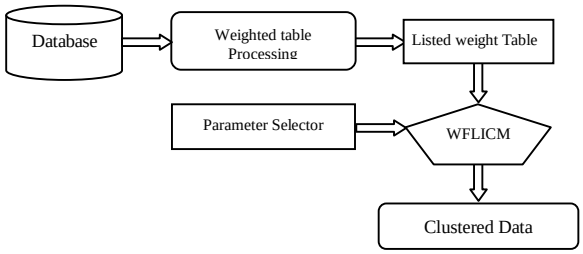

Figure 1: Architecture of the Weighted Fuzzy Local Information C Means

1. Weighted Table Processing block will fetch all the datasets from the database and assigns weights to all parameters in the database.
2. Listed weight table is the process of collecting all the weights assigned to the database in such a way that clustering process is done perfectly.
3. Clustering process should be done by means of WFLICM process on the basis of given parameters.
4. The Weighted Fuzzy local information c means algorithm is processed and clustering has been done for the given database with the given parameter.
5. The resultant output is the clustered data.

4. Weighted Fuzzy Local Information C-means (WFLICM) Clustering Algorithm

In this paper, a novel and robust FCM framework for clustering called Weighted Fuzzy Local Information C-means (WFLICM) clustering algorithm is proposed. Clustering makes the information retrieval about an individual from the database effortless.

4.1 Introducing the Fuzzy Factor G

The FCM method described in [27] yielded effective clustering but still have some disadvantages.

1. Although the introduction of local spatial information enhances their insensitiveness to noise upto some extent, they still lack enough robustness [9]–[11] to noise and outliers, especially in absence of prior knowledge of the noise.
2. The crucial parameter in the objective function is used to balance between robustness to noise and effectiveness of preserving the details of the information. Generally, its selection has to be made by experience or trial and error experiments.
3. They are all applied on the weighting factor assigned to the attributes, which has to be computed in advance. Details of the original information could be lost depending on the method used to generate the new cluster.

In order to overcome the above mentioned disadvantages a new factor in FCM objective function is needed. The new factor should have some special characteristics:

- to incorporate local spatial information in a fuzzy way in order to preserve robustness and noise insensitiveness;
- to control the influence of the neighborhood pixels depending on their distance from the central pixel;
- to use the original data avoiding preprocessing steps that could cause missing information;
- to be free of any parameter selection.

So, we introduce the novel fuzzy factor G_{ki} defined as

$$G_{ki} = \sum_{\substack{j \in N_i \\ i \neq j}} \frac{1}{d_{ij}+1} \left(1-u_{kj}\right)^m \left\| x_j - u_k \right\|^2 \qquad (1)$$

where the i^{th} pixel is the center of the local window (for example, 3X3), k is the reference cluster and the j^{th} pixel belongs in the set of the neighbors falling into a window around the i^{th} pixel n_i. $d_{i,j}$ is the spatial Euclidean distance between pixels i and j, u_{kj} is the degree of membership of the j^{th} pixel in the k^{th} cluster, m is the weighting exponent on each fuzzy membership, and u_k is the prototype of the center of cluster k. It is easy to see that the factor G_{ki} is completely free of using any parameter that controls the balance between the data noise and the data details. The control of this balance is automatically achieved by the definition of the fuzziness of each pixel. Also, by using d_{ij} the factor G_{ki} makes the influence of the pixels within the local window, to change flexibly according to their distance from the central pixel. Thus, more local spatial information can be used. It is worth indicating that the shape of the local window used in our experiments is square, but also, windows with other shapes such as diamond or circle can easily be adopted to the algorithm. As a whole, G_{ki} reflects the damping extent of the neighbors with the spatial distances from the central pixel. Moreover, there is no need of preprocessing steps to be applied to the algorithm, as it will be shown in the following. The important role of G_{ki} during the application of the algorithm will also be shown in the following subsection.

4.2 General Framework of WFLICM

By using the definition of G_{ki} the proposed robust FCM framework for data clustering, named Weighted Fuzzy Local Information C-Means (WFLICM) clustering algorithm, incorporates local spatial level information into its objective function, defined in terms of

$$J_m = \sum_{i=1}^{N} \sum_{k=1}^{c} \left[u_{ki}^m \left\| x_i - v_k \right\|^2 + G_{ki} \right] \qquad (2)$$

The two necessary conditions for J_m to be at its local minimal extreme, with respect to u_{ki} and u_k is obtained as follows:

$$u_{ki} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|x_i - v_k\|^2 + G_{ki}}{\|x_i - v_j\|^2 + G_{ji}} \right)^{\frac{1}{m-1}}} \quad (3)$$

$$v_k = \frac{\sum_{i=1}^{N} u_{ki}^m x_i}{\sum_{i=1}^{N} u_{ki}^m} \quad (4)$$

Thus, the WFLICM algorithm is given as follows

Step 1. Set the number c of the cluster prototypes, fuzzification parameter m and the stopping condition ε.
Step 2. Initialize randomly the fuzzy partition matrix.
Step 3. Set the loop counter $b = 0$.
Step 4. Calculate the cluster prototypes using (4).
Step 5. Compute membership values using (3).
Step 6. If Max $\{U^{(b)} - u^{(b-1)}\} < \varepsilon$ then stop, otherwise, go to step 4.

The measure used in the FLICM objective function (2) is still the Euclidean metric as in FCM, which is computationally simple. Moreover, differently from FCM, WFLICM is robust because of the introduction of the factor G_{ki}, which can be analyzed as follows:

The noise tolerance and outliers resistance property, completely relies on the definition of G_{ki}, as it is seen in (2). G_{ki} is automatically determined rather than artificially set, even in the absence of any prior noise knowledge. Two basic cases which describe the performance of the algorithm when outliers are present in the window will be presented in the following. As it will be shown the G_{ki} of the noise-corrupted pixels within a window will be kept with similar value to the central pixel ignoring the influence of the noise. The G_{ki} value will adaptively change in each iteration, converging to the central pixels value and thus preserving the insensitiveness to noise and outliers.

5. EXPERIMENTAL RESULTS

The performance of the proposed method by presenting numerical results and examples on various real dataset, with different types of noise and characteristics is shown in the table. The dataset consists of the passport, ration card and voter id databases of Tamilnadu, with approximately 6,000 records and the clustering process was performed namewise, district wise or yearwise with the proposed WFLICM algorithm. The information of an individual when queried is retrieved with less response time by integrating the information from all the databases due to the clustering technique by assigning weights to the attributes.

Table 1 : Clustered Dataset

Sl.No.	Character	Sex	Year	District	No. of Elements in a Cluster based on			
					Character	Sex	Year	District
1	A	Male	1994	Chennai	726	5211	3078	3590
2	G	Male	1995	Chennai	269	5211	2375	3590
3	S	Female	1996	Theni	1239	1445	8	1020
4	D	Male	1994	Madurai	378	5211	3078	345
5	M	Female	1995	Madurai	526	1445	2375	345
6	L	Female	1995	Chennai	140	1445	2375	3590
7	P	Male	1995	Madurai	527	5211	877	345

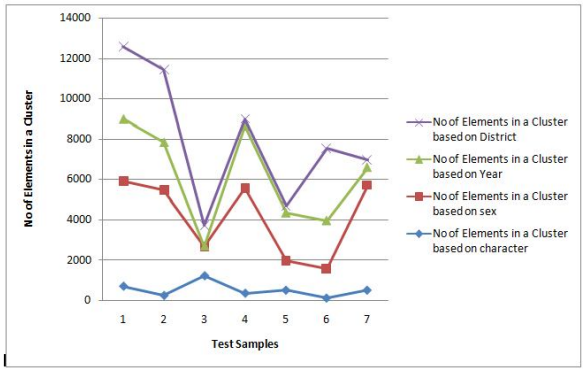

Figure 2 : Graph of Clustered Dataset of Table 1

The graph shows the representation of the number of elements present in each cluster in the test sample taken. The elements of the clusters are categorized based on Alphabet, Sex, Year, and District they belong. For example Test Sample 4 in which the condition to cluster the database are those persons whose names start with alphabet D, Sex Male, Year 1994, and belongs to district Madurai. The result which gives those Members whose names starts with 'D' as their first letter was about 378; Number of Males Present in our database was about 5211; No of persons born in the Year 1994 is 3078; and Number of Persons who are born in Madurai is 345; as per our database.

Table 2: Response Time of clustered Dataset

Methods	Processing Time (in Seconds	Response Time (in Seconds)						
		Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7
K-Means	1.5794	58.1956	12.9854	1.354	1.9863	1.7623	2.4987	1.5227
FCM	1.8576	56.7273	13.0187	0.7781	1.6649	1.3979	2.4781	1.2109
WFLICM	2.8232	56.5987	12.9659	0.6996	1.571	1.293	2.4702	0.96255

The table 2 obviously shows the results of K-Means, FCM and WFLICM. The processing time of K-means is less when compared to FCM which is less than WFLICM. Eventhough the processing time of WFLICM algorithm is more when compared to FCM and K-means the response time of the clusters is comparatively very less. For sample1 of table 1 the response time for WFLICM is less when compared to FCM which is less when compared to K-Means. The response time for sample2 of table 1 WFLICM is less when compared to FCM which is greater when compared to K-Means. Correspondingly the response time for all the rest of the samples of table 1 illustrates that WFLICM out performs the FCM which in turn is better than K-Means.

CONCLUSION

In this paper, a novel robust weighted fuzzy local information c-means (WFLICM) algorithm for data clustering was introduced. The proposed algorithm can detect the clusters of the data overcoming the disadvantages of the known FCM algorithms and their variants. This is achieved by incorporating local spatial information. The WFLICM introduces a new factor as a local (spatial domain) similarity measure which aims to guarantee robustness both to noise and outliers. Also, the algorithm is relatively independent of the type of the added noise, and as a consequence, in the absence of prior knowledge of the noise, WFLICM is the best choice for clustering. This is also enforced by the way that spatial information which is combined in the algorithm; the factor combines in a fuzzy manner the spatial level information, rendering the algorithm more robust to all kind of noises, as well as to outliers. Furthermore, all the other fuzzy c-means algorithms for clustering exploit, in their objective functions, a crucial parameter, which is used to balance the robustness and effectiveness of ignoring the added noise. This parameter is mainly determined empirically or using the trial-and-error method. The WFLICM is completely free of any parameter determination, while the balance between the noise and outlier details is automatically achieved by the fuzzy local constraints, enhancing concurrently the clustering performance. This is also enhanced, by the fact, that almost all the other methods perform the clustering on a precomputed data, while WFLICM is applied on the original data.

REFERENCES

[1]. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds., **Advances in Knowledge Discovery and Data Mining**, Menlo Park, CA: AAAI/MIT Press, 1996.

[2]. S. K. Pal and S.Mitra, **Neuro-Fuzz y Pattern Recognition: Methods in Soft Computing**. New York: Wiley, 1999.

[3]. Jiawei Han and Micheline Kamber, **Data Mining: concepts and techniques**, Morgan kaufman publishers, 2006.

[4]. L. Zadeh, "**Fuzzy sets**," Inf. Control, vol. 8, pp. 338–353, 1965.

[5]. J. Bezdek, **Pattern Recognition With Fuzzy Objective Function Algorithms**, New York: Plenum, 1981.

[6]. D. Pham, "**An adaptive fuzzy C-means algorithm for image segmentation in the presence of intensity inhomogeneities**," Pattern Recognition. Lett., vol. 20, pp. 57–68, 1999.

[7]. J. Dunn, "**A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters**," J. Cybern., vol. 3, pp. 32–57, 1974.

[8]. Han, J. M. Kamber and A.K.H Tung, **Spatial Clustering methods in Data Mining: A survey in Geographic Data Mining and Knowledge Discovery**, Miller, H. and J. Han(Eds). Taylor and Francis 2001.

[9]. R. Hathaway, J. Bezdek, and Y. Hu, "**Generalized fuzzy c-means clustering, strategies using L norm distance**," IEEE Trans. Fuzzy Syst., vol. 8, pp. 576–582, Oct. 2000.

[10]. K.Wu and M. Yang, "**Alternative c-means clustering algorithms,"** Pattern Recognition., vol. 35, no. 10, pp. 2267–2278, 2002.

[11]. J. Leski, "**Toward a robust fuzzy clustering**," Fuzzy Sets Syst., vol. 137, no. 2, pp. 215–233, 2003.

[12]. Hartigan J.A., **Clustering Algorithms**. 4th edition , John Wiley, new York, 1975.

[13]. Jain A.K and R.C. Dubes, **Algorithms for clustering data**, Prentice-Hall Inc., Englewood Cliffs, NJ., USA, 1988.

[14] Kaufman, L. and P.J. Rousseuw, **Finding Groups in Data: An Introduction to Cluster Analysis**, John Wiley and Sons, New York, 1990.

[15] Everitt B., **Cluster Analysis** 3rd Edition., Edward Arnold, London, UK, 1993.

[16] Mirkin B., **Mathematic Classification and Clustering** Kluwer Academic Publishers, USA, 1996.

[17] Jain, A.K., M.N. Murty and P.J. Flynn, **Data Clustering: A Review**. ACM Computing Surveys, Vol. 31: 264-323, 1999.

[18] Fasulo D., **An Analysis of recent work on clustering Algorithms** Technical report, University of Washington, 1999.

[19] Ghosh J., **Scalable Clustering Methods for Data Mining** In: Handbook of Data Mining, Nong, Y.(Ed.). Erlbaum, Lawrence, 2002.

[20] Dhillon I., J. Fan and Y. Guan, **Efficient Clustering of Very Large Document Collections** In: Data Mining for Scientific and Engineering Applications, Grossman R.Lt., C. Kamath, P. Kegelmeyer, V. Kumar and R.R. Namburu Eds Kluwer Academic Publishers, USA, 2001.

[21] Salton G., **The Smart Retreival System Experiment in Automatic Document Processing,** Prentice-Hall, Englewood Cliffs, New Jersy 1971.

[22] Gibson D., J. Kleinberg and P. Raghavan, 1998. **Clustering categorical data: An Approach based on dynamical systems**, International Conference on Very Large Databases, New York, USA., PP 311-322, 1998.

[23] Y. Tolias and S. Panas, "**Image segmentation by a fuzzy clustering algorithm using adaptive spatially constrained membership functions**," *IEEE Trans. Syst., Man, Cybern.*, vol. 28, no. 3, pp. 359–369, Mar. 1998.

[24] J. Noordam, W. van den Broek, and L. Buydens, "**Geometrically guided fuzzy C-means clustering for multivariate image segmentation**," in *Proc. Int. Conf. Pattern Recognition*, 2000, vol. 1, pp. 462–465.

[25] D. Pham, "**Fuzzy clustering with spatial constraints**," in *Proc. Int. Conf.Image Processing*, New York, 2002, vol. II, pp. 65–68.

[26] K. Premalatha and A.M Natarajan, "**A Literature Review ON Document Clustering**", Information Technology Journal 9(5): pp 993-1002, 2010.

[27] Stelios Krinidis and Vassilios Chatzis, "**A Robust Fuzzy Local Information C-Means Clustering Algorithm**" IEEE Transactions on Image Processing, Vol.19, No. 5, May 2010.