International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:1, 2009

A Modified Fuzzy C-Means Algorithm for Natural Data Exploration

Binu Thomas, Raju G., and Sonam Wangmo

***Abstract*—**In Data mining, Fuzzy clustering algorithms have demonstrated advantage over crisp clustering algorithms in dealing with the challenges posed by large collections of vague and uncertain natural data. This paper reviews concept of fuzzy logic and fuzzy clustering. The classical fuzzy c-means algorithm is presented and its limitations are highlighted. Based on the study of the fuzzy c-means algorithm and its extensions, we propose a modification to the c-means algorithm to overcome the limitations of it in calculating the new cluster centers and in finding the membership values with natural data. The efficiency of the new modified method is demonstrated on real data collected for Bhutan's Gross National Happiness (GNH) program.

***Keywords*—**Adaptive fuzzy clustering, clustering, fuzzy logic, fuzzy clustering, c-means.

I. Introduction

IN Data Mining, Cluster analysis is a technique for breaking data down into related components in such a way that patterns and order becomes visible [5]. Clusters are natural groupings of data items based on similarity metrics or probability density models. Clustering algorithms maps a new data item into one of several known clusters. Membership of a data item in a cluster can be determined by measuring the distance from each cluster center to the data point [6]. In crisp clustering, the data item is added to a cluster for which this distance is a minimum. In fuzzy clustering techniques, a data item is given partial memberships in all the clusters within a range of membership values from zero to one. A cluster has a center of gravity which is basically the weighted average of the cluster.

The most popular fuzzy clustering technique is fuzzy c-means algorithm. This paper reviews the advantages and limitations of fuzzy c-means clustering. In order to address some its limitations we present a modified fuzzy c-means algorithm. The new algorithm is analyzed with a natural data set and found to give satisfactory performance.

The paper is organized as follows: Section II describes the basic notions of fuzzy logic. A brief introduction to the concepts in fuzzy clustering followed by a discussion on fuzzy c-means algorithm and its limitations are presented in section III. Many researchers have reported modified versions of fuzzy c-means algorithm. Three such extensions are described in section IV. The new algorithm is introduced in section V. This is followed by an illustration of application of the new algorithm on a natural data set and a comparative analysis of its performance. Finally, section VII concludes the paper.

Binu Thomas works with the Department of Computer Science, Royal University of Bhutan, Bhutan under the foreign deputation scheme of Government of India(e-mail binumarian@rediffmail.com).

G. Raju got his PhD in Computer science from Kerala University, Kerala, India. Now He is working as professor in Computer Science at SCMS Cochin,India (e-mail kurupgraju@rediffmail.com).

Sonam Wangmo is a lecturer in Computer Science at Royal University of Bhutan.

II. Fuzzy Logic

The modeling of imprecise and qualitative knowledge, as well as handling of uncertainty at various stages is possible through the use of fuzzy sets. Fuzzy logic is capable of supporting, to a reasonable extent, human type reasoning in natural form by allowing partial membership for data items in fuzzy subsets [2].

Fuzzy logic is logic of fuzzy sets; a Fuzzy set has, potentially, an infinite range of truth values between one and zero [10]. Propositions in fuzzy logic have a degree of truth, and membership in fuzzy sets can be fully inclusive, fully exclusive, or some degree in between [13].The fuzzy set is distinct from a crisp set that it allows the elements to have a degree of membership. The core of a fuzzy set is its membership function: a function which defines the relationship between a value in the sets domain and its degree of membership in the fuzzy set (1). The relationship is functional because it returns a single degree of membership for any value in the domain [11].

$$\mu = f(s,x) \quad (1)$$

Where,
μ : is he fuzzy membership value for the element
s : is the fuzzy set
x : is the value from the underlying domain.

Fuzzy sets provide a means of defining a series of overlapping concepts for a model variable since it represent degrees of membership. The values from the complete universe of discourse for a variable can have memberships in more than one fuzzy set.

III. Fuzzy Clustering

Integration of fuzzy logic with data mining techniques has become one of the key constituents of soft computing in handling the challenges posed by massive collections of natural data [1]. The central idea in fuzzy clustering is the non-unique partitioning of the data in a collection of clusters. The data points are assigned membership values for each of the clusters. The fuzzy clustering algorithms allow the clusters to grow into their natural shapes [15]. In some cases the membership value may be zero indicating that the data point is

not a member of the cluster under consideration. Many crisp clustering techniques have difficulties in handling extreme outliers but fuzzy clustering algorithms tend to give them very small membership degree in surrounding clusters [14].

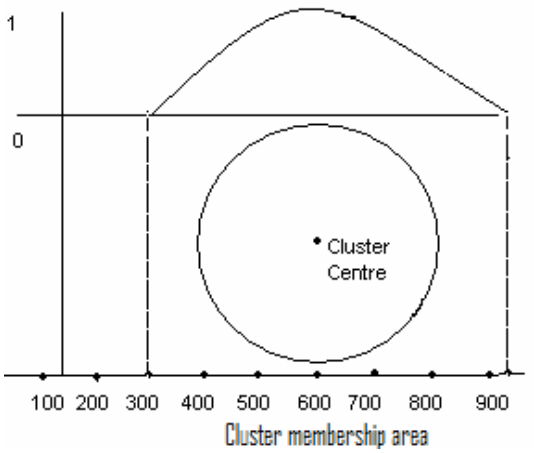

Fig. 1 Fuzzy membership in a cluster between zero and one

The non-zero membership values, with a maximum of one, show the degree to which the data point represents a cluster. As shown in Fig. 1, the points at the centre of the cluster have maximum membership values and the membership gradually decreases when we move away from the cluster centre. Thus fuzzy clustering provides a flexible and robust method for handling natural data with vagueness and uncertainty. In fuzzy clustering, each data point will have an associated degree of membership for each cluster. The membership value is in the range zero to one and indicates the strength of its association in that cluster.

A. C-Means Fuzzy Clustering Algorithm

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 with the formula,

$$\mu_j(x_i) = \frac{\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{p}\left(\frac{1}{d_{ki}}\right)^{\frac{1}{m-1}}} \quad (2)$$

where
$\mu_j(x_i)$: is the membership of x_i in the j^{th} cluster
d_{ji} : is the distance of x_i in cluster c_j
m : is the fuzzification parameter
p : is the number of specified clusters
d_{ki} : is the distance of xi in cluster C_k

Also the algorithm imposes a restriction which says the sum of memberships of a data point in all the clusters must be equal to one. This constrain is represented by expression 3.

$$\sum_{j=1}^{p} \mu_j(x_i) = 1 \quad (3)$$

The new cluster centers are calculated with the fuzzy membership values using (4).

$$c_j = \frac{\sum_i \left[\mu_j(x_i)\right]^m x_i}{\sum_i \left[\mu_j(x_i)\right]^m} \quad (4)$$

where
C_j : is the center of the j^{th} cluster
x_i : is the i^{th} data point
μ_j : the function which returns the membership
m : is the fuzzification parameter

This is a special form of weighted average. We modify the degree of fuzziness in x_i's current membership and multiply this by x_i. The product obtained is divided by the sum of the fuzzified membership. The c-means fuzzy clustering algorithm is given in Table I.

TABLE I
FUZZY C MEANS ALGORITHM

initialize p=number of clusters
initialize m=fuzzification parameter
initialize C_j (cluster centers)
Repeat
For i=1 to n :Update $\mu_j(x_i)$ applying (3)
For j=1 to p :Update C_i with(4)with current $\mu_j(x_i)$
Until C_j estimate stabilize

The first loop of the algorithm calculates membership values for the data points in clusters and the second loop recalculates the cluster centers using these membership values. When the cluster center stabilizes (when there is no change) the algorithm ends.

B. Limitations of the Algorithm

The fuzzy c-means approach to clustering suffers from several constrains that affect the performance [10]. The main drawback is from the restriction that the sum of membership values of a data point x_i in all the clusters must be one as in (4), 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 [10].The fourth constrain of the algorithm is its inability to calculate the membership value if the distance of a data point is zero(3).

IV. Some Common Modifications

Since fuzzy c-means algorithm is the most popular and widely used fuzzy clustering algorithm, many approaches have been proposed to improve the performance of the algorithm. Each of these modified methods proposes a new membership function for calculating the membership of data points in clusters. These new methods address the various limitations of the basic method. Three of them are discussed below:

A. 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[4].This method aims at identifying clusters with new shapes. In this method, they introduced a new membership function as given in expression (5).

$$\mu ik = \frac{1}{\sum_{j=1}^{c}\left(\frac{d^2(v_i, x_k)}{d^2(v_j, x_k)}\right)^{\frac{1}{m-1}}} \quad (5)$$

Here μ_{ik} is the membership of i^{th} data point in k^{th} cluster and c is the number of clusters. With this modified c-means membership function the fuzzy clustering algorithm can form clusters into their natural shapes.

B. 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 inhomogeneities and noise. The algorithm is formulated by modifying the objective function of the standard fuzzy C-means (FCM) method to compensate for intensity inhomogeneities[3]. Here the membership function is given as in expression (6)

$$\mu_{jk} = \frac{1}{\sum_{l=1}^{c}\left(\frac{\delta_{ik} + \gamma_k)}{\delta_{jl} + \gamma_l)}\right)^{\frac{1}{m-1}}} \quad (6)$$

In this expression, μ is the membership, δ denotes the distance and γ is defined as a term which represents the influence on a pixel by the neighbouring membership values.

C. 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 [10].The membership values in this method are calculated using Expression (7).

$$\mu_j(x_i) = \frac{n * \left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{p}\sum_{z=1}^{n}\left(\frac{1}{d_{kz}}\right)^{\frac{1}{m-1}}} \quad (7)$$

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[10].Since the sum of distances of points in all the clusters(7) involves in membership calculation this method tends to produce very less membership values when the number of clusters and points increase and this is the main limitation of it.

V. The New Fuzzy Clustering Method

In this section we propose a modification to the classical fuzzy c-means algorithm. First, we replaced the restriction imposed by exp (4) with a liberalized expression (8). That is, the sum of memberships in a cluster center must be n/2.

$$\sum_{i=1}^{n}\mu_j(x_i) = \frac{n}{2} \quad (8)$$

In c-mean the membership of a data point in a cluster depends directly on the sum of distances of the point in other cluster centers (2). Many limitations of the algorithm which affect the performance arise due this method [10]. Instead, if we consider the sum of distances of data members in a cluster for the calculation of memberships in that cluster, it might improve the performance of the algorithm. This leads to our second modification. The new membership function for i^{th} data point in j^{th} cluster is given below (9).

$$\mu_j(x_i) = \frac{n}{2} * \frac{\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}{\sum_{i=1}^{n}\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}} \quad (9)$$

K-means has demonstrated less sensitivity to initialization than the c-means algorithm [7]. A third improvement is to adopt the K-means method for initial segmentation. The results of segmentation are used as the initial centroids of our method. Membership values found with the (9) can be used in the same c-means (given in Table I) algorithm to produce better results.

VI. Illustration

Happiness and satisfaction are directly related with a community's ability to meet their basic needs and these are important factors in safeguarding their physical health. The unique concept of Bhutan's Gross National Happiness (GNH) depends on nine factors like health ecosystem, emotional well being etc.[16]. GNH regional chapter at Sherubtse College conducted a survey among 1311 villagers and the responses

were converted into numeric values. For the analysis of the new method we took the attributes income and health index as shown in Fig. 2. As we can see from the figure the response on health is converted into a numeric index on a ten point scale.

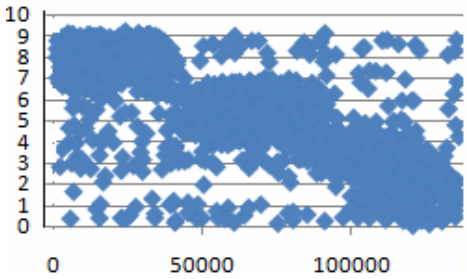

Fig. 2 The income(X axis) and health index(Y axis)

As we can see from the data set, in Bhutan the low income group maintains better health than high income group since they are self sufficient in many ways. Like any other natural data this data set also contains many outlier points which do not belong to any of the groups. If we apply c-means algorithm these points tend to get more membership values due to exp. (4).

To start the data analysis, first we applied k-means algorithm to find the initial three cluster centers. The algorithm ended with three cluster centers at C1(24243,6.7), C2(69794,5.1) and C3(11979.29,2.72). We applied these initial values in both c-means algorithm and the new method to analyze the data and the algorithms ended with centroids as given in Table II.

TABLE II
PERFORMANCE COMPARISON OF C-MEANS AND NEW METHOD

Centers	C-means		New Method	
	X	Y	X	Y
C1	24243.11	6.53	23464.1	7.3485
C2	69749.6	5.08	68707	5.71
C3	115979.3	2.83	112894.1	1.905

X values represent the income in Nulgtrum(Bhutan's currency). C1, C2 and C3 are the three final cluster centers.

From Fig. 2 and Table II, it can be seen that the final centroids of c-means method does not represent the actual centers of the clusters. This is due to the influence of outlier points. But the new method identifies the cluster centers in a better way by treating the outlier points in a different way.

VII. CONCLUSION

Fuzzy c-means algorithm, well known fuzzy clustering algorithm has several limitations in handling natural data with uncertainty and vagueness. In this paper we presented a modified version of fuzzy c-means algorithm. The new algorithm is applied on a natural data set and its performance is compared with that of classical fuzzy c-means algorithm. We found that the new method gives better performance in defining cluster centers. A detailed study with more data sets is necessary to ascertain the usefulness of the new method. Also a comparative analysis of the new algorithm with similar extensions of fuzzy c-means algorithm is to be carried out.

ACKNOWLEDGMENT

We would like to thank Mr. Nidup Gyelesten, Director, Gross National Happiness regional chapter, Sherubtse College, Bhutan for providing the data used for the analysis.

REFERENCES

[1] Sankar K. Pal, P. Mitra, "Data Mining in Soft Computing Framework: A Survey", IEEE transactions on neural networks, vol. 13, no. 1, January 2002.
[2] R. Cruse, C. Borgelt, "Fuzzy Data Analysis Challenges and Perspective". Available: http://citeseer.ist.psu.edu/ kruse99fuzzy.html
[3] 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.
[4] Frank Klawonn and Annette Keller, "Fuzzy Clustering Based on Modified Distance Measures", Available: http://citeseer.istpsu.edu/fuzzy_clustering_62
[5] W. H. Inmon, "The data warehouse and data mining", Commn. ACM, vol. 39, pp. 49–50, 1996.
[6] U. Fayyad and R. Uthurusamy, "Data mining and knowledge discovery in databases", Commn. ACM, vol. 39, pp. 24–27, 1996.
[7] Pavel Berkhin, "Survey of Clustering Data Mining Techniques", Available: http://citeseer.ist.psu.edu/berkhin02survey.html
[8] Chau, M., Cheng, R., and Kao, B, "Uncertain Data Mining: A New Research Direction", Available: www.business.hku.hk /~mchau/papers/UncertainDataMining_WSA.pdf
[9] Keith C.C, C. Wai-Ho Au, B. Choi, "Mining Fuzzy Rules in A Donor Database for Direct Marketing by A Charitable Organization", Proc of First IEEE International Conference on Cognitive Informatics, pp: 239 - 246, 2002
[10] E. Cox, Fuzzy Modeling And Genetic Algorithms For Data Mining And Exploration, Elsevier, 2005
[11] G. J Klir, T A. Folger, Fuzzy Sets, Uncertainty and Information, Prentice Hall,1988
[12] J Han, M Kamber, Data Mining Concepts and Techniques, Elsevier, 2003
[13] J. C. Bezdek, Fuzzy Mathematics in Pattern Classification, Ph.D. thesis, Center for Applied Mathematics, Cornell University, Ithica, N.Y., 1973.
[14] Carl G. Looney, "A Fuzzy Clustering and Fuzzy Merging Algorithm" Available: http://citeseer.ist.psu.edu/399498.html
[15] G. Raju, A. Singh, Th. Shanta Kumar, Binu Thomas, " Integration of Fuzzy Logic in Data Mining: A comparative Case Study", Proc. of International Conf. on Mathematics and Computer Science, Loyola College, Chennai, 128-136, 2008
[16] Sullen Donnelly, "How Bhutan Can Develop and Measure GNH", Available: www.bhutanstudies.org.bt/seminar/ 0402-gnh/GNH-papers-1st_18-20.pdf