Ms.Suchita S.Mesakar, Prof.M.S.Chaudhari / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 3, Issue 3, May-Jun 2013, pp.701-704

A Design of Fuzzy Approach for Data Clustering

Ms.Suchita S.Mesakar*, Prof.M.S.Chaudhari**
*(IV Sem M.Tech, Department of Computer Science and Engg., BCCE, Nagpur)
** (Asstt.Professor, Department of Computer Science and Engg., BCCE, Nagpur)

ABSTRACT

Data mining is the process of discovering patterns in large datasets. The objective of data mining is to extract information from a data set and transform it into an understandable structure for further use. Clustering in data mining is used to discover distribution patterns in the underlying data. Clustering aims to group data into clusters based on similarity/dissimilarity measures. Clustering algorithms uses the distance metric based similarity measure in order to partition the database such that data points in the same partition are more similar than points in different partitions. The clustering approaches differ in various aspects like flat or hierarchical structure, crisp or soft (fuzzy) cluster assignments. In this paper, the fuzzy based clustering approach is used to cluster the data.

Keywords:-Clustering, Data Mining, Fuzzy Clustering, Hard clustering, Soft clustering,

1. Introduction

Data mining is the process of extraction of knowledge or meaningful information from large dataset or huge databases. Data clustering is fundamental technique in data mining. Clustering aims to group objects into subsets in such a fashion that similar objects are grouped together, while different objects belong to different groups. Clustering is used in many areas which include machine learning, pattern recognition together with data mining, document retrieval, image segmentation. The learning methods are classified as unsupervised and supervised. In unsupervised learning for given set of patterns, a collection of clusters is to be discovered and additional patterns are assigned to correct cluster. In supervised learning set of classes (clusters) are given, new pattern (point) are assigned to proper cluster, and are labeled with label of its cluster. Clustering is unsupervised learning method because it deals with finding a structure in a collection of unlabeled data and no class values are given which can decide the priori grouping of data. Clustering uses the distance measures to find the similarity or dissimilarity between the objects. The common distance measures which are used for clustering are Euclidean, Manhattan, Minkowski and Mahalanobis distances. Many clustering algorithms are present, but selection of particular clustering algorithm depends upon the type of data. Clustering can be applied to numerical data and categorical data. The numerical data consist of numeric attributes, such as age, cost such attribute values can be ordered in specific manner and the properties can be used to apply the distance measures. The categorical data has different structure than the numerical data. The categorical data consists of non numeric attributes. The example of categorical attributes are color={orange,blue,white}.The categorical data do not have specific ordering as numeric data, so distance measures cannot be directly applied to categorical attributes. Clustering strategies can be classified as hard clustering and fuzzy clustering. In the hard clustering approach each object of the dataset belongs to only one cluster. In fuzzy clustering each object can belong to more than one clusters depending upon the degree of membership associated with it.

The major clustering methods are classified as

- Hierarchical Algorithms
- Partitional Algorithms
- Density Based Algorithms
- Grid Based Algorithms

Hierarchical clustering algorithms organize data into hierarchical structure. It starts with each case in a separate cluster and then combines the clusters step by step, reducing the number of clusters at each step until only one cluster is left. Hierarchical algorithms are classified into agglomerative methods and divisive methods. Agglomerative method is step-by-step clustering of objects and groups to larger groups and divisive method is step-by-step splitting of the whole set of objects into the smaller subsets and individual objects.

Partitional Clustering is simply division of the set of data objects into disjoint clusters.

Density-based algorithms identify clusters as dense regions of objects in the data space separated by regions of low density.

Grid based methods first divide space into grids, and then performs clustering on the grids. The main advantage of Grid based method is its

fast processing time which depends on number of grids in each dimension in quantized space. The density-based partitioning methods work best with numerical attributes, and grid-based methods work with attributes of different types.

2. Review of Different Clustering Algorithms

In clustering, the most widely used clustering algorithm is fuzzy clustering. Zadeh in 1965 proposed the fuzzy set theory & it gave an idea of uncertainty of belonging which was described by a membership function. The use of fuzzy set provides imprecise class membership function .The basic idea of fuzzy clustering is non-unique partition of the data into clusters. The objects are assigned membership values for each of the clusters and fuzzy clustering algorithm allow the clusters to grow accordingly.

Many fuzzy clustering algorithms are present in literature. In many situations fuzzy clustering is more natural the hard clustering. Objects are not forced to belong to one cluster as in hard clustering; rather they are assigned membership degrees between 0 and 1 indicating their partial membership. Fuzzy clustering algorithms are broadly classified into two groups: i) Classical and ii) Shape-based [6]. There exist many classical fuzzy clustering algorithms in the literature, among the most popular and widely used being: i) Fuzzy cmeans (FCM) [2], ii) Suppressed fuzzy c-means (SFCM) [7], iii) Possibilistic c-means (PCM) [4], and (iv) Gustafson-Kessel (GK) [8], while the shape-based fuzzy clustering algorithms include: i) Circular shape- based [9], ii) Elliptical shape-based [10], and (iii) Generic shape-based techniques [11].

The detailed review of classical fuzzy clustering algorithms is as below.

Fuzzy c means clustering method was developed by Dunn in 1973[1] and improved by Bezdek in 1981[2].The FCM employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one. After each iteration membership and cluster centers are updated according to the formula [3].

Krishnapuram and Keller [4] proposed a new clustering model named Possibilistic c-Means (PCM). The possibilistic C-means algorithm (PCM) was proposed to address the drawbacks associated with the memberships used in algorithms such as the fuzzy C-means (FCM). The approach differs from the existing clustering methods. In this the resulting partition of the data can be interpreted as a possibilistic partition, and the membership values can be interpreted as degrees of possibility of the points belonging to the classes. An appropriate objective function whose minimum will characterize a good possibilistic partition of the data is constructed, and the membership and prototype update equations are derived from necessary conditions for minimization of the criterion function.

To overcome difficulties of the PCM, Pal [5] defines a clustering technique that combines the features of both Fuzzy a Possibilistic c-means called Fuzzy Possibilistic c-Means (FPCM). Membership and Typicality's are very significant for the accurate characteristic of data substructure in clustering difficulty [3].An objective function in the FPCM depends on both membership and typicality. FPCM generates memberships and possibilities at the same time, together with the usual point prototypes or cluster center for each cluster.

The Gustafson-Kessel (GK) algorithm [8] is a powerful clustering technique that has been used in various image processing, classification and system identification applications. Gustafson and Kessel extended the standard fuzzy cmeans algorithm by employing an adaptive distance norm, in order to detect clusters of different geometrical shapes in one data set.

3. Proposed Work

In this paper, a novel fuzzy clustering approach is used for clustering of data.

The input to the algorithm will be the dataset containing numerical data. The algorithm begins by converting each value into fuzzy set depending upon the membership function. After the fuzzy regions are decided the maximum count value for each region is found. The fuzzy itemset will be decided from the maximum count values. The candidate set is decided according to which the clusters are decided. The term cluster matrix and data cluster matrix is build and finally the values are added to the best cluster.

The proposed approach can be given in following steps

Algorithm

Input:

2	1	1	0	0	0
1	1	0	0	0	0
1	0	2	0	0	0
0	0	0	3	0	2
0	0	0	11	1	1
0	1	0	4	0	0
0	0	0	8	1	2
3	0	1	0	0	0
0	1	0	3	0	0
0	0	0	8	2	1

Fig.1. Input values

Step 1. Transform each term frequency into fuzzy set.

Step 2. For all fuzzy regions, calculate the value of count.

	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H
	1.67	1.33	1	2	1	1	2	1	1	0	0	0	0	0	0	0	0	0
	2	1	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	2	1	1	0	0	0	1.67	1.33	1	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	1.33	1.67	1	0	0	0	1.67	1.33	1
	0	0	0	0	0	0	0	0	0	1	1	2	2	1	1	2	1	1
	0	0	0	2	1	1	0	0	0	1	2	1	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	1	1.43	1.57	2	1	1	1.67	1.33	1
	1.33	1.67	1.67	0	0	0	2	1	1	0	0	0	0	0	0	0	0	0
	0	0	0	2	1	1	0	0	0	1.33	1.67	1	0	0	0	0	0	0
	0	0	0	0.00	0	0	0	0	0	1	1.43	1.57	1.67	1.33	1	2	1	1
Count	7	5	4	8	4	4	5.67	3.33	3	6.66	9.2	8.14	5.67	3.33	3	7.34	4.66	4

Fig.2. Calculation of Fuzzy Regions

Step3. Find the region of each term with maximum count.

Step 4. Find fuzzy frequent itemset L1

Count	Support Values (Count/Total No.)
7	7.00/10 = 70 %
8	8.00/10 = 80 %
5.67	5.67/10 = 57 %
9.2	9.20/10 = 92 %
5.67	5.67/10 = 57 %
7.34	7.34/10 = 73 %

Fig.3. Fuzzy Frequent Itemset

Step 5. Generate the candidate set.

Step 6. Candidate cluster is generated based on the fuzzy frequent itemsets.

Step 7. Build p x k term cluster matrix G = [gmax-R]

Terms / clusters	c^{-1} (1)	c^{-1} (2)	c^{-1} (3)	c^{-1} (4)	c^{-1} (5)	c^{-1} (6)	c^{-1} (1,2)	c^{-1} (3,4)	c^{-1} (5,4)
1.Low	1.00	0.52	0.71	0.00	0.00	0.00	1.19	0.00	0.00
2. Low	0.50	1.00	0.25	0.50	0.00	0.00	0.42	0.00	0.00
3. Low	1.00	0.35	1.00	0.00	0.00	0.00	1.67	0.00	0.00
4. Mid	0.00	0.40	0.00	1.00	0.42	0.60	0.00	1.00	0.70
5. Low	0.00	0.00	0.00	1.00	1.00	1.00	0.00	1.67	1.67
6. Low	0.00	0.00	0.00	1.00	0.77	1.00	0.00	1.67	1.29

Fig.4.Term Cluster Matrix

Step 8. Build n x k data cluster matrix V.

1.Low	2.Low	3.Low	4.mid	5.Low	6.Low
1.67	2.00	2.00	0.00	0.00	0.00
2.00	2.00	0.00	0.00	0.00	0.00
2.00	0.00	1.67	0.00	0.00	0.00
0.00	0.00	0.00	1.67	0.00	1.67
0.00	0.00	0.00	1.00	2.00	2.00
0.00	2.00	0.00	2.00	0.00	0.00
0.00	0.00	0.00	1.43	2.00	1.67
1.33	0.00	2.00	0.00	0.00	0.00
0.00	2.00	0.00	1.67	0.00	0.00
0.00	0.00	0.00	1.43	1.67	2.00

Fig.5.Data Cluster Matrix

Step 9. Assign a data to a best target cluster.

	c^{-1} (1)	c^{-1} (2)	c^{-1} (3)	c^{-1} (4)	c^{-1} (5)	c^{-1} (6)	c^{-1} (1,2)	c^{-1} (3,4)	c^{-1} (5,4)
1	**4.67**	3.58	3.69	1.00	0.00	0.00	6.15	0.00	0.00
2	3.00	**3.05**	1.93	1.00	0.00	0.00	3.21	0.00	0.00
3	**3.67**	1.64	3.10	0.00	0.00	0.00	5.16	0.00	0.00
4	0.00	0.67	0.00	**3.34**	1.99	2.67	0.00	4.46	3.32
5	0.00	0.40	0.00	**5.00**	3.96	4.60	0.00	7.67	6.61
6	1.00	2.80	0.50	**3.00**	0.84	1.20	0.83	2.00	1.40
7	0.00	0.57	0.00	**5.10**	3.89	4.53	0.00	7.55	6.48
8	**3.33**	1.40	2.95	0.00	0.00	0.00	4.92	0.00	0.00
9	1.00	**2.67**	0.50	**2.67**	0.70	1.00	0.83	1.67	1.17
10	0.00	0.57	0.00	**5.10**	3.81	4.53	0.00	7.55	6.36

Fig.6.Targeted Clusters

4. Conclusion

Fuzzy clustering is a powerful unsupervised method for the analysis of data and construction of models. The overview of the most frequently used fuzzy clustering algorithms is discussed in this paper and a novel fuzzy based clustering approach is proposed to cluster the data. The approach starts by deciding the fuzzy regions for given data and fuzzy frequent itemset from the fuzzy region.Finally, the term cluster matrix and data cluster matrix is build and the values are added to the best cluster.

References

[1] J. C. Dunn (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", *Journal of Cybernetics 3: 32-57*

[2] J. C. Bezdek (1981): "Pattern Recognition with Fuzzy Objective Function Algoritms", Plenum Press, New York.

[3] R.Suganya, R.Shanthi," Fuzzy C-Means Algorithm-A Review", *International Journal of Scientific and Research Publications, Volume 2, Issue 11, November 2012.*

[4] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Transactions on Fuzzy Systems, vol. 1, no. 2, pp. 98–110, 1993.*

[5] Pal N.R, Pal K, Keller J.M. and Bezdek J.C, "A Possibilistic Fuzzy c-Means Clustering Algorithm", *IEEE Transactions on Fuzzy Systems, Vol. 13, No. 4, Pp. 517–530, 2005.*

[6] Hoppner, F., et al., Fuzzy Cluster Analysis: methods for classification, data analysis, and image recognition. 1999, New York: John Wiley & Sons, Ltd.

[7] Fan, J.L., Zhen, W.Z., and Xie, W.X.,Suppressed fuzzy c-means clusteringalgorithm. Pattern Recognition Letters, 2003.24: pp. 1607- 1612.

[8] Gustafson, D.E. and Kessel, W.C.,Fuzzy clustering with a fuzzy covariance matrix. *In Proceedings of IEEE Conference on Decision Control. 1979. pp. 761-766.*

[9] Man, Y. and Gath, I., Detection and separationof ring shaped clusters using fuzzy clustering,*IEEE Transaction on Pattern Analysis and Machine Intelligence, 1994.16(8): pp. 855-861*.

[10] Gath, I. and Hoory, D., Fuzzy clustering of elliptic ring-shaped clusters. PatternRecognition Letters, 1995.16(7): pp. 727-741.

[11] Ameer Ali, M., Dooley, L.S., and Karmakar,G.C.,Object based segmentation using fuzzy clustering, *in IEEE International Conference on Acoustics, Speech, and Signal Processing.2006.*