ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

Clustering Data Using Fuzzy C-Means by Determining the Number of Clusters Using Gap Statistics

Dr.A.Joshi[1], Dr.V.Subedha[2], Vidhya.E[3], Vaishnavi.S[4]

Head, Dept. of IT, Panimalar Institute of Technology, Chennai, Tamil Nadu, India[1]

Head, Dept. of CSE, Panimalar Institute of Technology, Chennai, Tamil Nadu, India[2]

B.Tech Student, Dept. of IT, Panimalar Institute of Technology, Chennai, Tamil Nadu, India[3]

B.Tech Student, Dept of IT, Panimalar Institute of Technology, Chennai, Tamil Nadu, India[4]

ABSTRACT: Clustering is an unsupervised learning technique which is used to group samples of data based on their features and properties of instances. In any clustering algorithm determining the number of clusters is a significant task which needs to be efficient to group the data with relatively similar characteristics. In this paper we use a method Gap statistics algorithm to determine the number of clusters for a Fuzzy C-means clustering algorithm [2] to group the samples of data. In gap statistics method we calculate the error measure for each sample of data and evaluate it with a reference value and depending on the evaluation we obtain the optimal number of clusters which can be applied to the Fuzzy C-means clustering.

KEYWORDS: Fuzzy C-Means clustering, Gap Statistics, error measure, data characteristics, reference value.

I. INTRODUCTION

In today's scenario large amount of data is available that is to be processed to obtain some useful information and project these processed results as a source to certain applications. Data mining is a process which abstracts the useful data from large amount of available unprocessed datasets and then group these data into an useful information. Data mining is used in many different applications such as Image segmentation, speech detection, Configuration management, fraud detection etc. For grouping of the data based on the similarities and properties we use various clustering methods and algorithms. In any clustering algorithms determining an optimal number of clusters is a tedious task. We need to define an optimal method for determining the number of clusters for the efficient performance of the various clustering algorithms. This paper is divided into four sections. The first sections deals with Fuzzy C-means clustering. The second section tells gap statistics a method for determining the number of clusters for a given dataset. The third section deals with application of gap statistics method to Fuzzy C-means clustering algorithm in determining the number of clusters or k value and also the benefits of the gap statistics method over the other statistical approach in determining the number of clusters. The fourth section gives details about future work and conclusion.

II. LITERATURE SURVEY

A New Fuzzy Clustering Validity Index With a Median Factor for Centroid-Based Clustering-IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 23, NO. 3, JUNE 2015 Determining the number of clusters, which is usually approved by domain experts or evaluated by clustering validity indexes, is an important issue in clustering analysis.The performance of WLI and some existing clustering validity indexes are evaluated and compared by running the fuzzy c means algorithm for clustering various types of datasets , including artificial datasets , UCI datasets and images.

Identification of Bad Data of Power System Based improved GSA Judgment -Zhang Junfang , Ge Liang , Zhao Tong ,Tian Ming,Wu Junji IEEE/CAA JOURNAL OF Electricity Distribution , Jan 2015 An improved power

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

system security and stability of operation based on bad data detection using GSA(Gap Statistic Algorithm)data mining method, and is applied on bad data detection in power system.

SECTION I

Clustering is an unsupervised learning technique for grouping of data depending on the similarity of the characteristics of a sample data with other samples. In clustering we do not have predefined classes or groups. Depending on the features of the data we partition or group them.There are different types of clustering like partitioning, hierarchical, Density based, grid based, model based and constraint basedmethods. The Fuzzy C-means clustering algorithm used in this paper comes under the soft clustering algorithm.In soft clustering data elements can belong to more than one cluster, and are associated with the help of membership levels.These indicate the datapoints association with the particular cluster.Fuzzy C –means is used to assign the membership values to each datapoints in order to determine to which cluster it belongs to optimally.

The Fuzzy C-means algorithm was first proposed by James C.Bezdek[3] in 1981. Fuzzy C-means algorithm is a simple algorithm to group the similar objects together.In Fuzzy C-means clustering algorithm we are assigning the membership values to each datapoints based on their distance from the cluster centers.More the datapoint is closer to a particular cluster center more it has the probability of belonging to that particular cluster.Depending on the distance we are assigning the higher membership value for the datapoint which is nearest to it. This is the actual concept of Fuzzy C-means algorithm. The performance of the Fuzzy C-means algorithm can be improved if we determine the best optimal number of clusters for which the data needs to be partitioned. Several methods are there for determining the number of clusters out of them we are considering the gap statistic algorithm in determining the number of clusters which will be explained in the next section. The results of the Fuzzy C-means clustering algorithms[2] depends on the centroids of the clusters. If we get any new sample value we will be recalculating the centroids again to group the data efficiently.

The Fuzzy C-means clustering algorithm has the following steps:

Step1: Initially we need to randomly select the centres for the clusters for the given dataset.

Step 2: Now using the cluster centres from the above step to calculate the distance between each sample data and the centroid of the clusters.

Step3: Now comparing the distance measure obtained from the above step for each data sample, allot the data to the cluster to which it has nearby distance or minimum distance.

Step4: we will be recalculating the new cluster centre every time when a new data sample needs to be grouped. The formula for recalculating the cluster centre is :

$$v_{i=(\frac{1}{C_i})\sum_{j=1}^{C_i} X_i} \quad [2]$$

Ci – This represents no. of data points in a particular cluster.

Step5: whenever a new cluster centre is obtained recalculate the distance between the data sample and cluster centre again.

Step6: If no new sample data is present stop else repeat the process from Step3 till Step 5.

For Fuzzy C-means clustering we need to have a prior knowledge on the number of clusters and cluster centres. To determine the number of clusters we are using the gap statistics algorithm this is explained in detail in the following section.

SECTION II

Gap Statistic algorithm can be used in the data mining for determining the number of clusters for a given dataset thereby improving the performance of cluster analysis. There are two approaches local and global in determining the number of clusters. Gap statistics follow the global approach in determining the number of clusters. This gap statistics approach provides even a better way for the grouping of erroneous data.For determining the erroneous data we make use of an elbow curve method which takes the min elbow angle in the log(W(k))curve. This algorithm can be applied to any clustering techniques and distance function. In this method we calculate an error function logW(k) for each sample data and compare it with a reference value obtained. The best value of k is taken in such a way that the value logW(k) falls farthest below its expected reference curve. To use this gap statistics method first we need to select a valid or matching reference null distribution. There are two different ways in determining the reference distribution:

1. Generating the reference data by comparing each sample data over a set of observed values uniformly.

2. Generating the reference data using the important features of the data

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

In this paper we are considering the reference dataset using the first method because of its simplicity. After generating the reference dataset the following steps are considered to obtain the optimal k value which is the number of clusters that can be used for the clustering algorithms.

Step1: We will be clustering the data for a fixed number of clusters say k=1,2,3...n

Step2: calculate the W(k) measure for all values of k.

$$W(k) = \sum_{r=1}^{k} \frac{1}{2n_r} D_r \quad [6]$$

Here k is the number of clusters we are assuming nr is the size of the dataset Dr is the sum of distances for all the points in the cluster r taken in pairs.

$$D_{r=\sum_{i,i' \in C_r} d_{ii'}} \quad [6]$$

Dii' is the distance between the sample and cluster centre.

Step3: Now carryout the same process for the reference datasets let us assume the reference datasets as B. Now on clustering each of the reference datasets and calculating the Wb(k) for b=1,2,..B and k=1,2,...n .

Step4: Now the Gap statistics can be calculated based on the following formula.

$$Gap(k) = \frac{1}{B} \sum_b \log\left(W_{b^*}(k)\right) - \log\left(W(k)\right) \quad [6]$$

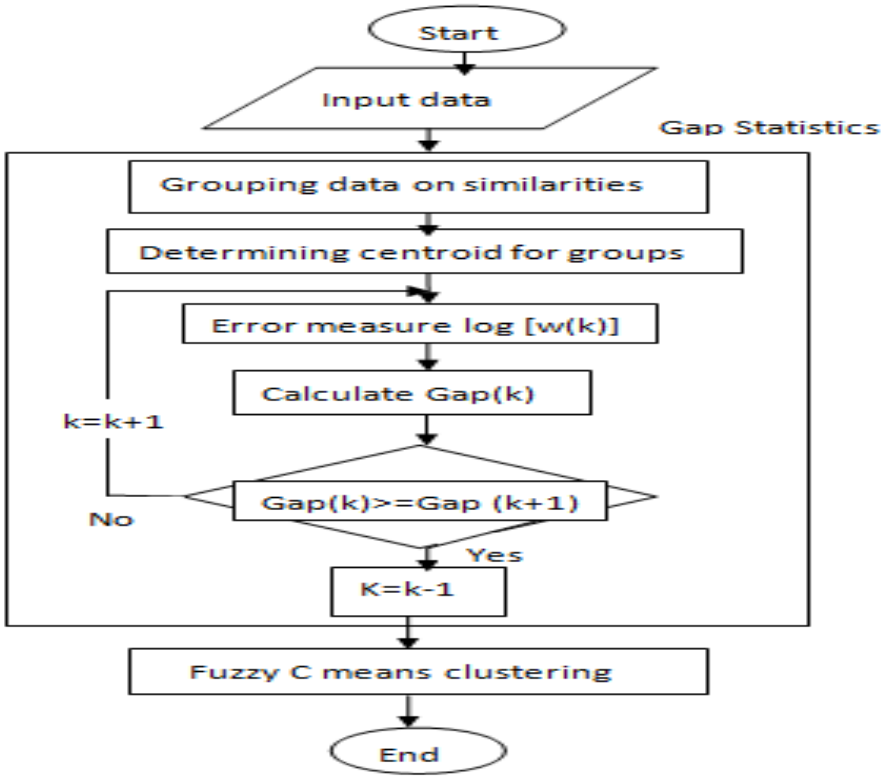
Step5: Repeat the process for each different values of k when Gap(k)>=Gap(k+1).

This value of k minus 1 will gives the optimum number of clusters that can be used for clustering the dataset.

The value of K which is obtained from the gap statistics method is the optimum number of clusters and this can be given as the predetermined number of clusters value for the Fuzzy C-means clustering algorithm.

SECTION III

This section provides the information regarding the applications of the Gap statistics method. The following flow diagram shows determining the number of clusters in Fuzzy C-means clustering algorithm using the Gap statistics method.



ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

First we are taking our dataset which needs to be clustered. Then using the gap statistics algorithm as defined in the previous section we are obtaining the number of clusters. This value can be used as the predefined number of clusters for the Fuzzy C-means clustering algorithm. The gap statistics method can also be used for grouping of the bad data. Because this method performs by observing the error measure for each sample with a reference value.Thus enabling in determining the level of erroneous data if present. When considering the other statistical methods like silhouette to determine the number of clusters, they need a maximum value of k that is the number of clusters value to obtain the optimal number of clusters. That is need to repeat the iteration for a larger number of times to obtain the optimal k value when compared to the Gap statistics method. Hence in other words the computational cost and time complexity is less in gap statistics method when compared to the other methods. One useful step in gap statistics methods is determining the reference datasets which can be used to compare against the sample data values. The gap method is found to have a better performance in identifying the well separated clusters.

SECTION IV

We have taken the computer features and price information as our dataset and used the gap statistics method to determine the number of clusters and clustering using the Fuzzy C-means algorithm. This method provided an optimal solution for finding the number of clusters and clustering of the relative data accordingly. In gap statistics all the data samples are assigned to any of the clusters without leaving any sample. We also achieved lesser time complexity and computational cost for large value of samples in the dataset. The following table compares the results for number of clusters obtained for Gap statistics and other different methods like Silhouette statistics[5] and Calinski and Harabasz (CH) methods.

	No Of Clusters		
Number of samples	Gap Statistics	Silhouette Statistics	CH
50	4	2	1
100	5	2	2
150	5	3	4
200	6	5	5

It is obvious from the above table the gap statistics provides better results than the other methods compared.

III. FUTURE WORK

The Gap statistics method can be used with other adaptive versions of Fuzzy C- means clustering algorithm for finding better solutions for elongated clusters. For taking the reference dataset in gap statistics we have used the first approach out of the two mentioned approaches. The other method of determining the reference dataset can be used for further efficient simulation of reference data.

REFERENCES

[1]Yun Liu Coll. of Commun. Eng., Jilin Univ.,Changchun, China Tao Hou ; Fu LiuElectronics Letters ,"Improving fuzzy c-means method for unbalanced dataset",IEEE Transaction on Clustering,Vol.51 , Issue: 23 pp.12-34 November 2015.

[2]Silva, L. Dept. of Inf. & Appl. Math., Fed. Univ. of Rio Grande do Norte, Natal Brazil Moura, R. ; Canuto, A.M.P.; Santiago, R.H.N.; Bedregal,B,"An Interval-Based Framework for Fuzzy Clustering Applications", IEEE Transactions on Fuzzy SystemsVol.23,Issue: 6 ,pp 25-34 November 2015.

[3]Chih-Hung Wu Dept. of Electr. Eng., Nat. Univ. of Kaohsiung,Kaohsiung,Taiwan Chen-Sen Ouyang; Li-WenChen; Li-Wei Lu,"A New Fuzzy Clustering Validity Index With a Median Factor for Centroid- BasedClustering", IEEE transactions on fuzzy systems, vol. 23, no. 3,pp 10-36 june 2015.

[4]Yizhang Jiang, Member, IEEE, Fu-Lai Chung, Member, IEEE, Shitong Wang,Zhaohong Deng, Senior Member, IEEE, Jun Wang, Member, IEEE, and Pengjiang Qian, Member, IEEE,"Collaborative Fuzzy Clustering From Multiple Weighted Views",IEEE transactions on cybernetics, vol. 45, no. 4,pp 22-28 april 2015.

[5]Pawan Lingras and Matt Triff,"Fuzzy and Crisp Recursive Profiling of Online Reviewers and Businesses",IEEE transactions on fuzzy systems, vol. 23, no. 4,pp 17-38 august 2015.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

[6]Zhang Junfang , Ge Liang , Zhao Tong ,Tian Ming,Wu Junji,"Identification Of Bad Data Of Power System Based Improved Gsa Judgment", IEEE/CAAJOURNAL OF Electricity Distribution ,vol.29,pp22-35 Jan 2015.
[7]I-Jen Chiang Grad. Inst. of Biomed. Inf., Taipei Med. Univ., Taipei,Taiwan Liu, C.C.-H. ; Yi-Hsin Tsai; Kumar, A.,"Discovering Latent Semantics in Web Documents Using Fuzzy Clustering", IEEE transactions on fuzzy systems,vol.23 ,issue: 6 ,pp 25-30 november 2015.
[8]Shao-Tung Chang Dept. of Math., Nat. Taiwan Normal Univ., Taipei, Taiwan Kang-Ping Lu; Miin-Shen Yang,"Fuzzy Change-Point Algorithms for Regression Models", IEEE Transactions onFuzzy Systems, Vol.23,Issue: 6 pp 12-25 November 2015.

BIOGRAPHY

Dr.A.Joshi is the Head Of The IT Department, College of Panimalar Institute Of Technology, Chennai. She is a Doctorate in Information Technology and has done Master in Mathematics.She has nearly 12 years of experience. Her research interest are Computer Network and Data Mining.

Dr.V.Subedha is the Head Of The CSE Department, College of Panimalar Institute Of Technology, Chennai. She is a Doctorate in Computer Science and Engineering .She has nearly 17 years of experience.Her research interest are Data Mining.

Vidhya.E is a student in Information Technology Department, College of Panimalar Institute Of Technology, Chennai. She is currently doing her final year in Bachelor of Technology. Her research interests is Data Mining.

Vaishnavi. S is a student in Information Technology Department, College of Panimalar Institute Of Technology, Chennai. She is currently doing her final year in Bachelor of Technology. Her research interests is Data Mining.