INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 6 (2017) PAGES 2524 - 2530
RECEIVED : 25.10.2017 PUBLISHED : 01.12.2017
Dec 01, 2017

CLUSTERING IN HEAVY-TAILED DATA DISTRIBUTIONS BY USING FUZZY SEMANTIC CLUSTERING

Gurram.Bobby Gowtham
B.Tech - Computer science And Engineering
VIT University, Vellore, India.

Abstract:

A distributed frameworks cluster is a gathering of machines that are for all intents and purposes or topographically isolated and that cooperate to give a similar services or application to customers. It is conceivable that huge numbers of the services you keep running in your system today are a piece of a distributed system cluster. Cluster is a vital information mining procedure which expects to isolate the information objects into important gatherings called as groups. It is the way toward gathering objects into bunches to such an extent that items from a similar group are comparable and objects from various groups is unique. In information mining, information bunching has been examined for long time utilizing diverse calculations and ordinary patterns are proposed for better results around tailed data. The fuzzy semantic strategy is look at to group the overwhelming followed information by utilizing some technique for remove. An appraisal think about is introduced in view of time and exactness. In this method proposed here is evaluated to other relational clustering schemes using various propinquity matrices as input. Simulations demonstrate the scheme to be very effectual.

Keywords: Clustering, Heavy-Tailed Distribution, Fuzzy semantic Algorithm.

1. INTRODUCTION

The data clustering, in the framework of databases, that refers to the capability a number of servers or instances to bond to solitary database. An instance is an anthology of memory and procedures that interrelates with a database, Which is the no of substantial files that essentially store data. The main aim of data clustering is that grouping or assign data points which have similar properties or values that makes to the user convenient to retrieve. Normally clustering problems can divided into 2 categories: 1) Hard clustering (or) Crisp Set Theory(CST) clustering. 2) Soft Clustering (or) Fuzzy Set Theory (FST).

In Hard Clustering there is only one cluster for the Data point. This Crisp Set Theory is the clusters exposed more clusters are disjoined. But in soft clustering data point may have more than one cluster with some higher chances. The clustering techniques where applied in many industries which by using various data like Archaeology, Anthropology, Economics, Geography, criminology, anthropology, sociology, Remote sensing and Medicine. For the modern investigates on the topic of data clustering by using fuzzy clustering to take account of this article.

Zadeh was developed the Fuzzy Set Theory(FST) in 1965 in that investigate work particularly for the meadow of modeling and ambiguity[1]. Velmurugan analyzes about the performance of k-

Means and Fuzzy C-Means and author conclude the process calculation time of k-Means algorithm is not as much of than the FCM algorithm for the certain application[2]. This highlights the significance of data clustering as a key technique of data mining and pattern recognition, knowledge discovery and statistics. From the diversity of clustering algorithms, the Partitioned algorithms have the benefit of being able to integrate facts about the global shape or size of clusters by using suitable prototypes and distance measures in the point function. Another important step to find and measure the distance of clusters this may find the similarity of clusters and elements are calculated. This will give influence the clusters shape , clusters some data point may very nearby one to another based on the distance some time to far away according other data point. Where the distinction of clustering may use symmetric or asymmetric distance.

Nidhi Grover [3] was present detailed study about fuzzy clustering algorithms. the author briefly explained advantages and difficulties various fuzzy clustering algorithm like , Fuzzy Possibilistic C-Means Algorithm (FPCM), Possibilistic C-Means Algorithm (PCM),Fuzzy C-Means Algorithm (FCM), and Possibilistic Fuzzy C Means Algorithm (PFCM). Anjana Gosain et al [4] was presented broad study on performance and analysis of the various clustering algorithm and they are conclude Density Oriented Fuzzy C-Means(DOFCM) worked and produce best result and identify outlier first and then give best centroids contrasted to all other algorithm.

As of late record bunching has been widely researched. With the fast development of the volumes of content information, archive grouping can help in sorting out the accumulation, accordingly encouraging future route and inquiry. Archive bunching is helpful in numerous data recovery errands, for example, record perusing, association and survey of recovery comes about, age of progressive systems of reports in web indexes and so forth. Record bunching is a subset of the bigger field of information grouping, which obtains ideas from the fields of Information Retrieval (IR), Natural Language Processing (NLP), and Machine Learning (ML), among others. Dissimilar to archive order, no named records are given in grouping; thus bunching is otherwise called unsupervised learning. In spite of the fact that there are different customary grouping methods, yet they can't be connected for bunching content information because of the essential properties of content databases: volume of the information database, high dimensionality of list of capabilities, inadequacy in record vector, perplexing and uncertain semantics, and loud information.

In web where display heavy tails data when distribution of large no of files, that to gather with no of files requested by web users ,no of files are transmitted to the network, duration of file transmission and storing data into servers. where the distribution of data may contain multimedia element such as audio, video, graphics and rich text. This situation where files are transmitted over the network that transmission duration appear to heavy - tailed distribution. This impact might be clarified by the measure of the documents themselves, as they too show substantial tails. And the conveyance of documents that are accessible on an arrangement of servers are overwhelming – followed. The substantial followed property of transmission times is probably going to be caused mostly by the dissemination of accessible documents, as opposed to by the idea of client requirement. In this paper present detailed investigate of heavy tailed data distribution and clustering based on fuzzy semantic clustering and Euclidean distance[7].

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 6 (2017) PAGES 2524 - 2530
RECEIVED : 25.10.2017 PUBLISHED : 01.12.2017
Dec 01, 2017

2. INITIAL CONCEPTS

1. Heavy Tailed Data Distribution

Heavy tailed Data distribution are probability distribution which tails were not exponentially enclosed, that is very heavy tails compare to exponential distribution. In numerous applications it is the correct tail of the distribution that is of intrigue, yet a distribution may have an overwhelming left tail, or the two tails might be substantial[8].

THEOREM 1

The distribution of random variable R distribution function F is supposed to heavy (right) tail if an instant produce of F, $M_f(t)$ is finite for all t>0.

$$\int_{-\infty}^{\infty} e^{tr}\, dF(R) = \infty \text{ for all } t > 0$$

The random variable R with heavy tailed distribution have huge values with restricted probabilities ensuring in no outliers. Getting sampling from heavy distribution results in normally tiny values with some huge valued samples. In the heavy tailed input with simulation will take long time to attain stable state also the variance can be huge.

$$| \bar{x}_n - \mu | \approx cn^{\frac{1}{\alpha}-1}$$

Where C is constant.

THEOREM 2

The distribution of a random variable X with distribution function F is said to have a long right tail if for all $t > 0$,

$$\lim_{x\to\infty} \Pr[X > x + t \mid X > x] = 1,$$

This will be the instinctive for a long - tailed distributed amount that if the long tailed amount surpasses some high level, the probability towards 1 that may exceed any other higher level.

3. EARLY FUZZY CLUSTERING ALGORITHMS

Clustering is the method that means huge amount of dataset is divided into some tiny groups or segment called clusters using some similarity measures, that the similarity between any two or more substance in the same group is more that between two objects in two different groups. Recent days fuzzy clustering play important role of data extraction, pattern reorganization and model identification. In soft clustering fuzzy dataset object can be in the right place one or more clusters.

Fuzzy c -means is the one of the most broadly used technique in the fuzzy clustering algorithms. Originally it was proposed by Dunn[5] and it has modified no of occasion by different authors.

$$j(_m) = \sum_{i=1}^{N}\sum_{j=1}^{C} Uij \;\; |xi - cj|^2$$

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 6 (2017) PAGES 2524 - 2530
RECEIVED : 25.10.2017 PUBLISHED : 01.12.2017
Dec 01, 2017

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center.

Fuzzy c-means have advantage that produce good result and data point belong to one or more cluster objects. But is take too much computational time and also Euclidean distance measures can unevenly weight essential factors.

Fuzzy k-means is another fuzzy based algorithm most widely used n recent days. It is simplest unsupervised clustering algorithm proposed by MacQueen [6] this simple procedure and classify through certain no of clusters. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as bary centers of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

4. HEAVY TAILED DATA CLUSTERING

Heavy tailed data fuzzy semantic clustering used two distance methods Euclidean and Minkowski Distance. where the results are compare with their precision.

Euclidean distance :

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_p-y_p)^2}=\sqrt{(x-y)^t(x-y)}$$

Minkowski Distance:

$$d(x,y)=\left[\sum_{i=1}^{p}|x_i-y_i|^m\right]^{1/m}$$

Fuzzy Semantic Clustering Algorithm (Clustering *N* Objects):

1. Initiate with n clusters, All holding at least a single object and N x N symmetric matrix of distances (or resemblances) D={d ik}

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 6 (2017) PAGES 2524 - 2530
RECEIVED : 25.10.2017 PUBLISHED : 01.12.2017
Dec 01, 2017

2. Find the distance matrix for the adjacent(most related) couple of clusters. Let the distance between "most related" clusters C and V be d_{cv}.

3. Combine clusters C and V. Label the newly formed clusters(CV).Update the accesses in the distance matrix.

(a) Removing the rows and columns matching to clusters C and V and

(b) addition a row and column charitable the distance between cluster (CV) and the residual clusters.

4. Repeat Steps 2 and 3 a total of *n-1* times. (All substance will be in a single cluster after the algorithm terminates.) proof the individuality of clusters that are combined and the levels (distances) at which the combines take place.

There are 3 linkage schemes. The main distinction amid these schemes are the distances between *(CV)* and some other cluster *W*.

(I) Single Linkage:

$$d_{(CV)W} = \min \{d_{CW}, d_{VW}\}$$

(II) Complete Linkage:

$$d(CV)w = \max\{d_{CW}, d_{VW}\}$$

(III) Average Linkage:

$$d(CV) = \frac{\sum_i \sum_k d_{ik}}{n_{(CV)} n_w}$$

where d_{ik} is the distance between object *i* in the cluster *(CV)* and object *k* in the cluster *W*, and n_{CV} and n_W are the number of items in clusters *(CV)* and *W*, respectively.

The simulation results shown are shown Table I and Figure 1and Figure 2 .Figure 1 illustrate time variations of fuzzy semantic clustering scheme with Euclidean and Minkowski Distance with same distance.

Evolution index	Distance	Mean(time in Second)	Mean(precision)
Xie-Beni	Euclidean	0.306	0.00016528%
	Minkowski Distance	0.318	2.256968e-09

Table I. Precision values.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 6 (2017) PAGES 2524 - 2530
RECEIVED : 25.10.2017 PUBLISHED : 01.12.2017
Dec 01, 2017

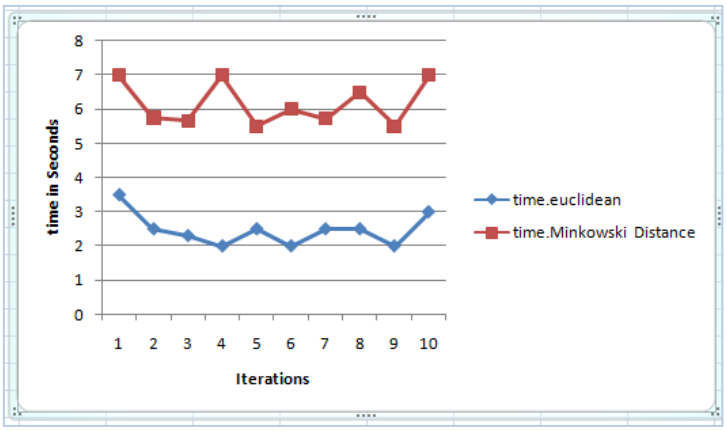

Fig 1. The time of Fuzzy Semantic Clustering method with Euclidean and Minkowski distances.

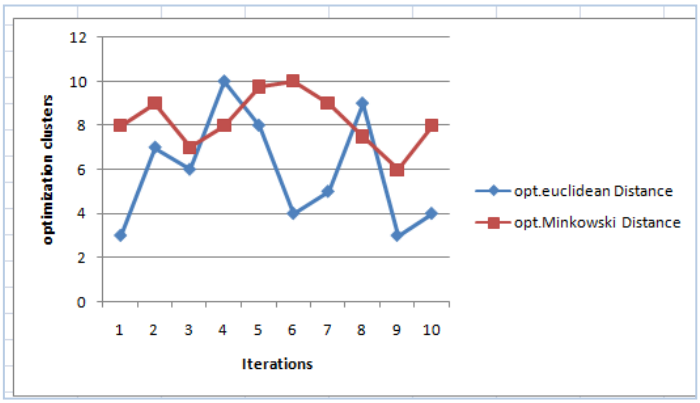

Fig 1. The time of Fuzzy Semantic Clustering method with Euclidean and Minkowski distances.

In table I shows the simulation test for 600 data with fuzzy semantic clustering method and two distances for contrast in terms and precision of clusters with Cauchy distribution and normal distribution.

DISCUSSION

In this result discussion compare with previous data clustering methods and fuzzy semantic clustering algorithm for data distribution. In the Fuzzy c-means and Cauchy distribution with using two distances namely Euclidean and Manhattan distances. Where the previous schemes Euclidean distance is very less than time with Minkowski distance. The precision value also greater than all other previous distances. So this approach working good for heavy tailed data distribution.

CONCLUSION

This paper presented new approach fuzzy semantic clustering for heavy tailed data distribution and this approach present detailed distribution of data clustering and distance. This method are considered to heavy tailed data clustering based on the Euclidean distance and Minkowski distance. our Minkowski distance approach take less time compare all other distance for clustering label. finally this research area potential topic in future research.

REFERENCES

[1]. L.A. Zadeh, "Fuzzy sets and systems", J. Fox (Ed.), System Theory, Polytechnic Press, Brooklyn, NY (1965), pp. 29-39.

[2]. T. Velmurgan, "Austria Performance Comparison Between K-means and Fuzzy C- means", Wulfenia Journal Using Arbitrary Data Points, vol. 19, pp. 1-8, 2012.

[3] N. Grover, "A Study of Various Fuzzy Clustering Algorithms", International Journal of Engineering Research, vol.3(3), pp. 177-181, 2014.

[4]. Anjana Gosain and Prabhjot KaurA," density oriented fuzzy C-means clustering algorithm for recognising original cluster shapes from noisy data" , International Journal of Innovative Computing and Applications ,Volume 3, Issue 2 ,DOI: 10.1504/IJICA.2011.039591.

[5]. J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics, vol. 3, pp. 32-57, 1973.

[6] MacQueen JB. Some methods of classification and analysis of multivariate observations, *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, 1967, p. 281-297.

[7].S. Mahmoud Taheri and A. Mohammadpour ,"On Fuzzy Clustering for Heavy-Tailed Data" 2017 5th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS 7-9 March, Qazvin Islamic Azad University, Tehran, Iran.

[8] In probability theory, heavy-tailed distributions,"https://en.wikipedia.org/wiki/Heavy-tailed_distribution.

[10] J. Ilow, and D.Hatzinakos," Analytic Alpha-Stable Noise Modeling in a Poisson Field of Interferers or Scatterers", IEEE Transactions on Signal Processing, vol. 46(6), pp. 1601-1611, 1998.

[11] J. P Nolan, "Stable Distributions-Models for Heavy Tailed Data," Birkhauser, Boston 2015.

[12] L. B. Klebanov," Heavy Tailed Distribution", Charles University, Prague, 2003.

[13] M. B. Ferraro, and P. Giordani, " A Toolbox for Fuzzy Clustering Using the R Programming Language, "Fuzzy Sets and Systems,vol. 279, pp. 1– 16, 2015.

[14] M. Fajardo, A. McBratney, and Whelan, "Fuzzy Clustering of Vis–NIR Spectra for the Objective Recognition of Soil Morphological Horizons in Soil Profiles", Geoderma, vol. 263, pp. 244–253, 2016