INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 3)
Available online at: www.ijariit.com

Clustering on uncertain data based on probability distribution similarity using Kullback-Leibler divergence

Vanashri Shrirang Shinde
vanashrishinde14@gmail.com
Bharat Ratna Indira Gandhi Collge of Engineering, Solapur, Maharashtra

C. M. Jadhav
chandaronly@gmail.com
Bharat Ratna Indira Gandhi Collge of Engineering, Solapur, Maharashtra

ABSTRACT

Cluster analysis is one in all the necessary knowledge analysis strategies and could be a terribly advanced task. It's the art of a sleuthing cluster of comparable objects in massive data sets while not requiring such teams by suggests that options or knowledge information bunch on unsure knowledge could be the toughest task in each modeling similarity between uncertain data objects. The foremost of the previous technique for a bunch unsure knowledge extends partitioning clustering algorithms and Density-based mostly clustering algorithms. These strategies are supported by the geometric distance between two unsure knowledge objects. Such technique unable to handle unsure objects, that are cannot distinguishable by victimization geometric characteristics and Distribution associated with object itself isn't thought about. Likelihood distribution could be the most vital characteristic of the unsure object isn't taking under consideration throughout measurement the similarity between 2 uncertain objects. The very fashionable technique Kullback-Leibler divergence won't to measures the distribution similarity between two unsure knowledge objects. Integrates the effectiveness of KL divergence into each partition and density based mostly bunch algorithms to properly cluster unsure knowledge. Calculation of KL-Divergence is extremely expensive to resolve this drawback by victimization well-liked technique kernel density estimation and use the quick Gauss remodel technique to any speed up the computation to decrease execution time.

Keywords— *Clustering, Clustering uncertain data, Density based clustering, Partition clustering, KL-divergence*

1. INTRODUCTION

Uncertainty is generally due to limited perception or knowledge of reality, limited observation equipment, limited resources to collect, store, transform, analyze, and understand data. Sensors used to collect data may be thermal, electromagnetic, chemical, mechanical, optical radiation or acoustic used in security, environment surveillance, manufacture systems. Ideal sensors are linear and sensitive, such that the output signal is linearly proportional to the value of the examined property. Practically due to changing environmental conditions ideal sensors outputs cannot be expected. Aggregation of data and granularity of data are also contributing to uncertainty in data. The data we handle have uncertainties in many cases (S.D. Lee, B. Kao, and R. Cheng, 2007)).One of the most general cases of uncertainty is the errors being introduced when the object is mapped from the actual space to the pattern space. Clustering is just a task of partitioning a set of objects into several groups of meaningful subclasses is called cluster. Clustering is called an unsupervised classification i.e. There is unavailable any predefined classes. A good clustering technique produces a cluster with high quality. In which the similarity in intra-cluster is high and inter-cluster similarity is low. Uncertainty in today's data comes with new challenges into the clustering of uncertain data.

The most of previous studies in the field of a clustering uncertain data are based on an improvement to the traditional clustering algorithms which are particularly designed for certain data. Any data object within certain data set is assumed a single point, the distribution of an uncertain object is not considered in traditional (previous) clustering algorithms. Thus, the existing studies are extended to the prior algorithms to cluster uncertain data. These existing methods are limited to using geometrical distance based similarity measurement, and these earlier approaches cannot capture the difference between uncertain objects in the shape of different distributions.

There are five different techniques available for clustering uncertain data. Partition based clustering method cluster the object based on the expected distance between objects. Density-based clustering algorithms are more important to find out clusters with different sizes and shapes. The core concept of density-based clustering would be to cluster the objects based on density i.e. number of neighbourhood objects exceed some threshold value. The earlier density-based clustering algorithms fail to discover clusters which exist within the cluster and is also doesn't work on varied density.

© 2019, www.IJARIIT.com All Rights Reserved

Clustering uncertain data by using Kullback Leibler divergence, (Bin Jiang, Jian Pei, 2013)) Clustering of uncertain data is known as an important issue in today's world. The issue of clustering uncertain data has been studied for many years and find out with this problem. The most of the earlier clustering algorithm for clustering uncertain data are an extended version of an existing clustering algorithm which are designed for clustering uncertain data. But extended existing algorithms to clustering uncertain data are limited because they depend on the geometric distance between uncertain object to measure similarity.

2. LITERATURE SURVEY

The problem of clustering uncertain data has been studied in recent years and find out feasible solutions on this problem. But most of the majority of the earlier methods for clustering uncertain data are an improvement to the existing clustering algorithms which are particularly designed by considering certain data. These improved existing algorithms of clustering uncertain data are limited because they depend on geometric properties of the data object. The geometric properties are used to measure the similarity between the object and they cannot consider the distribution similarity. There are three most used methods.

(a) Partition based Clustering approaches.
(b) Density-based clustering approaches.
(c) Possible world clustering approaches.

2.1 Partition based clustering approach

The partition-based clustering approaches for clustering uncertain data extends the mostly used algorithm k-means method by using the expected distance between uncertain objects to measure the similarity. It is required to first calculate the expected distance between each pair of an object and select cluster representative iteratively. In Partition based algorithm have restriction on the user to specify a number of cluster (k) before start clustering process. Partition based method cannot support the feature detection of an outliers and it forcefully add all objects into the clusters. Thus, only the centers or representative of objects are considered in these uncertain versions of the k-means method. If the two or more object has the same center then expected distance-based partition clustering approaches cannot differentiate the two sets of objects having different distributions.

2.2 Possible world approaches

Possible world approaches follow the possible world semantics. A set of possible worlds are taken as an example from an uncertain data set. Each of the possible worlds contains an instance from each object. The clustering process is applied individually on each possible world and the final clustering is completed by combining the clustering results of all possible worlds into a single global clustering. A sampled possible world of a data object cannot consider the distribution of a data object, therefore the possible world contains only one instance from each object. The clustering results obtained using different possible worlds it may be different. Thus, the possible world clustering methods often cannot provide a reliable and meaningful clustering result at the object level. It is computational infeasible because of |you can find an exponential number of possible worlds.

2.3 Density-based clustering approaches

Density-based clustering approaches for uncertain data are an improvement to the original DBSCAN (H.-P. Kriegel and M. Pfeifle, 2005) algorithm and the OPTICS algorithm (H.-P. Kriegel and M. Pfeifle, 2005). The core idea of these algorithms does not change, data objects in geometrically dense regions are grouped together as clusters and these clusters are separated by using sparse regions. However, objects are heavily overlapped and they are in the same region. But these objects have a different distribution. There are no clear sparse regions to separate objects into different clusters by considering distribution similarity. Therefore, the earlier density-based methods cannot work well.

In the (Bin Jiang, Jian Pei, 2013), proposed clustering uncertain data by using density-based clustering algorithm DBSCAN based on distribution similarity. Develop the uncertain DBSCAN algorithm (M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, 1996) which applies the effectiveness of Kullback Leibler divergence into the original DBSCAN algorithm to measure distribution similarity. But the original DBSCAN algorithm with Kullback Leibler divergence work on fixed input parameter values for eps and minds. It fails discovers clusters with varying density and it cannot discovers cluster which exists within the cluster.

Partition based, possible world and density-based clustering techniques for clustering uncertain data depend on the geometric characteristics of a data object and they just focus on an instance of an uncertain object, they cannot take into account the similarity between uncertain objects in terms of distribution. The distribution difference between object cannot be captured by geometric distances. The probability distribution is the most important characteristic of an uncertain object. Two uncertain objects may overlap but they have a different distribution.

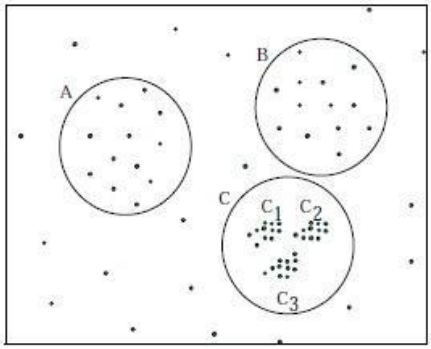

Fig. 1: Cluster with different density and cluster within a cluster

© 2019, www.IJARIIT.com All Rights Reserved

An important property of many real-data sets is that their exact cluster structure cannot be determined by global (single) density parameters. Very different densities (Minpts) may be needed to find out the cluster with high quality in different regions of data space. For example, in the data set showed in the below figure, it is not possible to obtain the clusters C3, C2, C1, A, and B simultaneously by using single global density parameter value. A global density-based clustering approach would detect only of the clusters C1, C2, C3, A, and B. If uses global density to discover the cluster C1, C2, C3 then objects from B and A are detected as noise.

3. RELATED WORK

Kullback Leibler divergence work on the probability distribution of objects. So that it is required to first calculate the probability distribution of each and every object. First Models the uncertain object as a random variable in both discrete and continuous probability distribution. The uncertainty is a very important feature in uncertain data objects, and the probability value will affect the quality of clustering results and should be reflected in the similarity. In order to capture the distribution difference between uncertain objects, we use KL divergence (Bin Jiang, Jian Pei, 2013)) to calculate the statistical difference between two objects in data. The KL divergence is a robust metric for measuring the difference between two data objects. Given p and q two distributions in the discrete domain with a finite number of values, the Kullback-Leibler divergence between p and q is defined below. If the domain is discrete with a finite number of values the object is a discrete random variable and its probability distribution is described by a probability mass function (pmf).

It is required to first calculate the probability mass function of all uncertain objects.

3.1 Kullback Leibler Divergence (S. Kullback and R.A. Leibler, (1998)): KL divergence measures how two distributions are different. It is used to measure the distribution difference between uncertain objects by using the probability distribution of each and every object. The Kullback Leibler divergence is called a distance between two distributions.

(a) In Discrete case, let f and g be two probability mass functions (pmf) in a discrete domain with finite number values. The KL divergence between f and g

$$D\,(f||g) = \sum_{x \in ID} f(x) \log f(x)/g(x)$$

(b) In the continuous case, let f and g be two probability density functions in a continuous domain. The KL divergence between f nag G is

$$D(f||g) = \int_{ID} \log\ f(x) \frac{f(x)}{g(x)}$$

Calculate the similarity between two uncertain objects by using Kullback Leibler divergence between their probability distributions.

3.2 Fast Gauss Transformation (C. Yang, R. Duraiswami, N.A. Gumerov, and L.S. Davis, 2003): The fast Gauss transform has proven to be a very efficient algorithm for solving many problems in applied mathematics and physics, and nonparametric statistics. All these problems require the evaluation of the discrete Gauss transform.

To break through this computational barrier, Greenland and Strain developed the fast Gauss transform, which requires O (M +N) operations, with a constant factor depending on the dimensionality d and the desired precision. The fast Gauss transform is an analysis based fast algorithm in the sense that it speeds up the computation by approximation of the Gaussian function to achieve the desired precision. The sources and targets can be placed on general positions. In contrast, the most popular fast Fourier transform requires the point to be on a regular mesh which is in general not available in the application| of statistics and pattern recognition. Implementation in two dimensions demonstrated the efficiency and effectiveness of the fast Gauss transform.

Despite its success in lower dimensional applications in mathematics and physics, the algorithm has not been used much in statistics, pattern recognition and machine learning where higher dimensions occur commonly. A multivariate Taylor expansion is applied to the improved fast Gauss transform which substantially reduces the number of the expansion terms in higher dimensions.

The k-centre algorithm is utilized to efficiently and adaptively subdivide the higher dimensional space according to the distribution of the points. A simpler and more accurate error estimate is reported, due to the simplification created by the new Taylor expansion and space subdivision schemes. The improved fast Gauss transform is capable of computing the Gauss transform in dimensions as high as tens which commonly occur in nonparametric statistics, pattern recognition. The behaviors of the algorithm in very high dimensional space (such as up to several hundred) will be studied and reported. The fast Gauss transform boosts the efficiency of our algorithms dramatically with only a small decrease of the clustering quality.

4. CLUSTERING ALGORITHM WITH KL

Develop a general framework of clustering uncertain objects by considering the distribution of every object as the first the very first class citizen. Uncertain objects can have any discrete or continuous distribution. We show that distribution differences between uncertain objects cannot be determined by the earlier methods which are based on geometric distances. To measure the distribution difference between uncertain objects by using well-known technique Kullback Leibler divergence. Demonstrate the effectiveness of KL divergence in both partitioning and density-based clustering methods.

To solve the challenge of evaluating the KL divergence in the continuous case, we calculate KL divergence by kernel density estimation and integrate the fast Gauss transform to speed up the evaluation process. We conducted experiments on real data sets to

© 2019, www.IJARIIT.com All Rights Reserved

show that clustering uncertain data by considering probability distribution is meaningful and clustering algorithm with Kullback Leibler divergence technique are efficient and scalable.

Before applying density based DBSCAN and partition based K-Mediod algorithm on a dataset. First calculates Kullback Leibler divergence between objects i.e. distribution distances. Then the DBSCAN algorithm works based on KL divergence.

4.1 DBSCAN Algorithm (H.-P. Kriegel and M. Pfeifle, 2005)

(a) Select each unvisited point P from the dataset.
(b) Retrieve all points (neighbourpts) density-reachable from P with respect to Eps radius and minpts according to the Kullback Leibler Divergence.
(c) If P is core point a cluster is formed.
(d) Expand cluster until all neighbor points in a cluster are processed.
(e) If P is a border point, no points are reachable from P and DBSCAN visits the next point of the dataset.
(f) Continue the process until all of the points have been processed and no point can be included in any cluster.

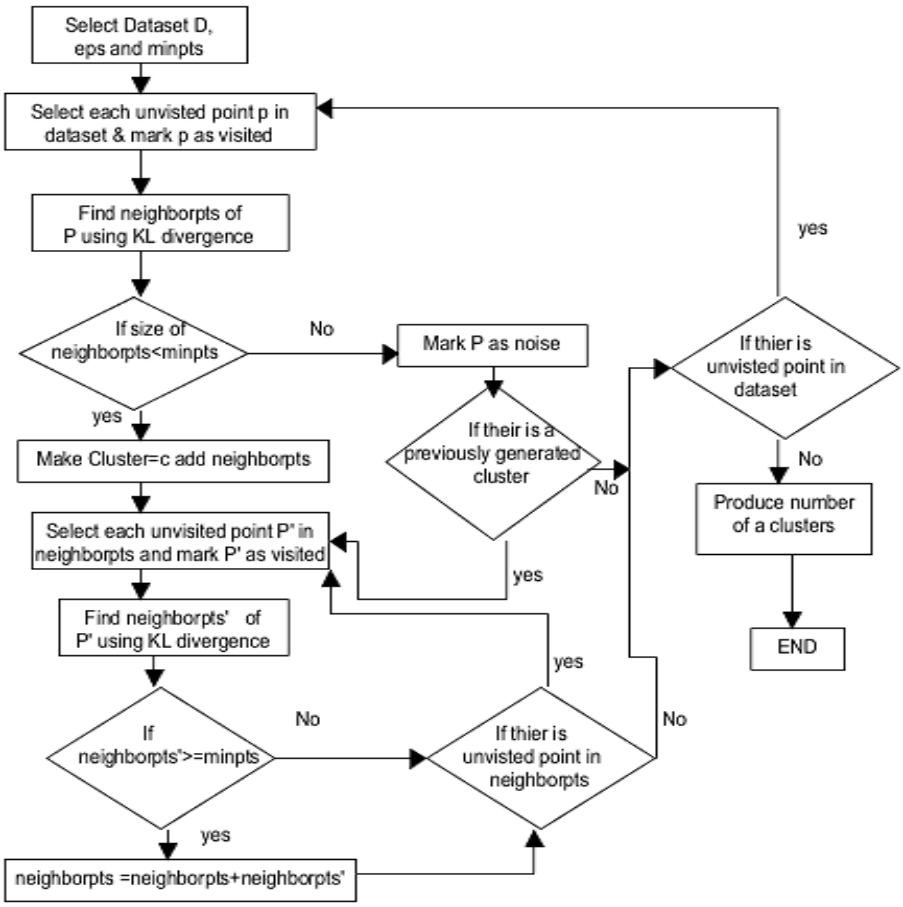

Fig. 2: Flow chart of the DBSCAN algorithm

4.2 Randomized K-Mediod algorithm

The randomized k-medoids method, instead of as opposed to finding the optimal non representative object for swapping with a representative object, randomly selects a low representative object for swapping if the clustering quality can be improved. The randomized k-medoids method works the same in building and swapping framework. In the beginning, the building phase is performed by selecting the initial k representatives' object at random. Remaining object i.e. non selected objects are assigned to the most similar representative object according to KL divergence. Then perform the swapping phase, in the swapping phase, recursively replace representatives object by non-representative objects. In each iteration instead of discovering the optimal non representative object for swapping in the uncertain k-medoids method, a low representative object is randomly selected. The randomly selected non representative object is replaced with the cluster representative object.

The randomized k-medoids method has time complexity O(r n E) where r is the number of iterations in the swapping phase and E is the complexity of determining the KL divergence of two objects. The cost of the building phase in the uncertain k-medoids method is removed because of the representative's objects are randomly initialized.

4.2.1 Applying KL divergence into K-medoid algorithm: K-medoid algorithm is a classical partitioning method to cluster the data. A partitioning clustering technique organizes a set of uncertain data objects into K number of clusters. Using KL divergence as similarity measurement, Partitioning clustering algorithms tries to organize data into K clusters and chooses the K representatives recursively, one for each cluster to minimize the total KL divergence. Here use K-medoid method to show the performance of clustering using KL divergence similarity. The K-medoid method works in two phases, first is a building phase and second is a swapping.

© 2019, www.IJARIIT.com All Rights Reserved

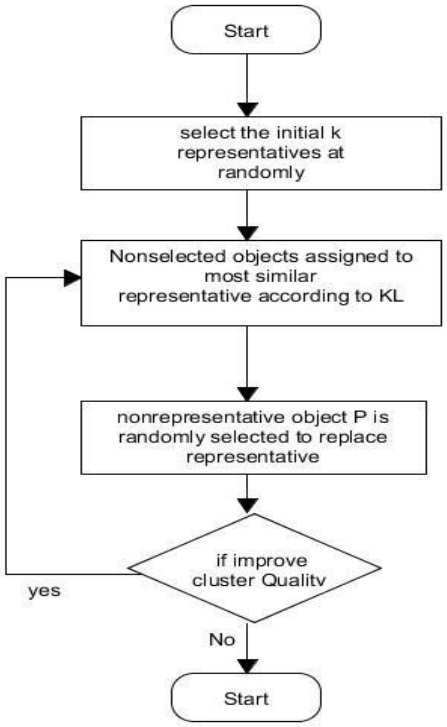

Fig. 3: Flow chart of Randomized K-mediod algorithm

4.3 Building phase

In the building phase, the K-medoid algorithm obtains an initial clustering by selecting initial medoids/representative objects randomly.

4.3.1 Algorithm of the building phase

Step 1: Randomly choose k number of data objects as the initial cluster.
Step 2: Calculate the KL divergence between each representative and the remaining data objects.
Step 3: Assign the non-representative to the medoid which has the smallest KL divergence with the medoid.
Step 4: Perform swapping phase: In the swapping phase, the uncertain k-medoid method recursively improves the clustering quality by swapping a number of representative data objects with the representative objects to which it is assigned.

4.3.2 Algorithm of swapping phase

Step 1: Swapping the representative with the non-representative data.
Step 2: Repeat step 2, 3 and 4 of the building phase until the clusters are not changed.

Clustering uncertain data based on their probability distribution similarity is very efficient clustering technique compare to existing methods. But in the building phase, the algorithm select initial representative objects randomly that impact on the quality of the resulting clusters and sometimes it generates unstable clusters which are meaningless. Also here the initial partition is based on the initial representative objects i.e. medoids and the initial partition affects the result and a total number of iterations. If the initial medoids are selected in an efficient way then it does not produce any empty clusters and also we can reduce the total number of iterations.

5. MATHEMATICAL MODEL

$S = \{I, O, P, F, S\}$
$I = \{D, eps, minpts\}$
Where

- D=dataset
- eps=epsilon distance
- minpts=minimum number of points

$D = \{d1, d2, d3, d4...dn\}$
$O = \{O1, O2...On\}$ Produce Number of Clusters.
$P = \{P\ 1, P\ 2, P\ 3, P\ 4, P\ 5, P\ 6, P\ 7\}$
P={select input, select, find, formation of cluster, expansion of cluster, check, produce cluster}
Step 1: select input(D, eps, minpts)
Step 2: select(each unvisited datapoint in D)

$until(unvisited - list\ !\ = null)$

© 2019, www.IJARIIT.com All Rights Reserved

Step 3: Find(neighborpts of P)

find neighbour points of P using KL Divergence. KL Divergence

i. In Discrete case

$$D\,(f||g) = \sum_{x \in ID} f(x) \log f(x)/g(x)$$

let f and g be two probability mass function.

ii. In Continuous case

$$D(f||g) = \int_{ID} \log\ f(x) \frac{f(x)}{g(x)}$$

let f and g be two probability density function.

Step 4: Formation of Cluster(Ci)

if (neighborpts ≥ minpts) then cluster Ci is formed
other than noise will be considered and goto step P2.

Step 5: Expansion of Cluster(Ci)

select each unvisited point P' in nighborpts and apply process p2 to neighborpts'. If (neighborpts' ≥ minpts) then
neighborpts = neighborpts + neighborpts'
until(*neighborpts = unvisited*)

Step 6: Produce cluster.

Check unvisited(D)
If (unvisited-list !=null) then go P2.

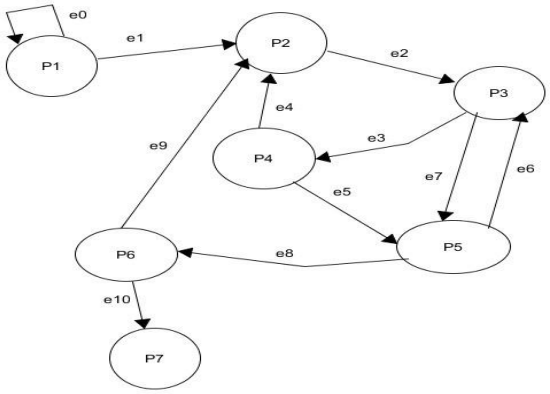

Fig. 4: State transition diagram

5.1 State Transition Diagram
S1=Select input,
S2=select object,
S3= find neighbor (i.e. similar data objects),
S4= formation of the cluster,
S5=expansion of the cluster,
S6=check unvisited objects,
S7=produce the numbers of cluster

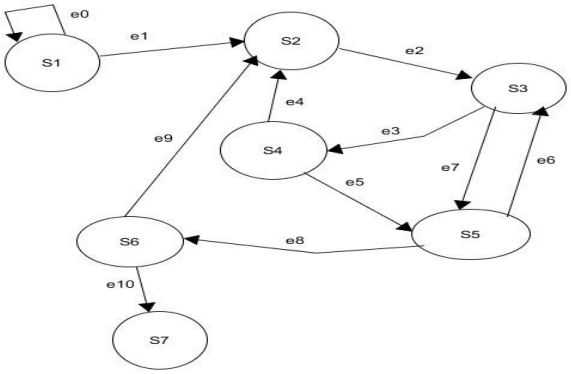

Fig. 5: State transition diagram

F = Fails to find out the cluster.
S= Return number of clusters and noise point.

© 2019, www.IJARIIT.com All Rights Reserved

6. EXPERIMENTAL RESULT

In this section first, present a Data set used then present Measurement of KL divergence. After that shows the comparison between original DBSCAN with KL and K-Mediod with KL divergence.

In our experiment, we have to pass two values for K and for specifying cluster range.

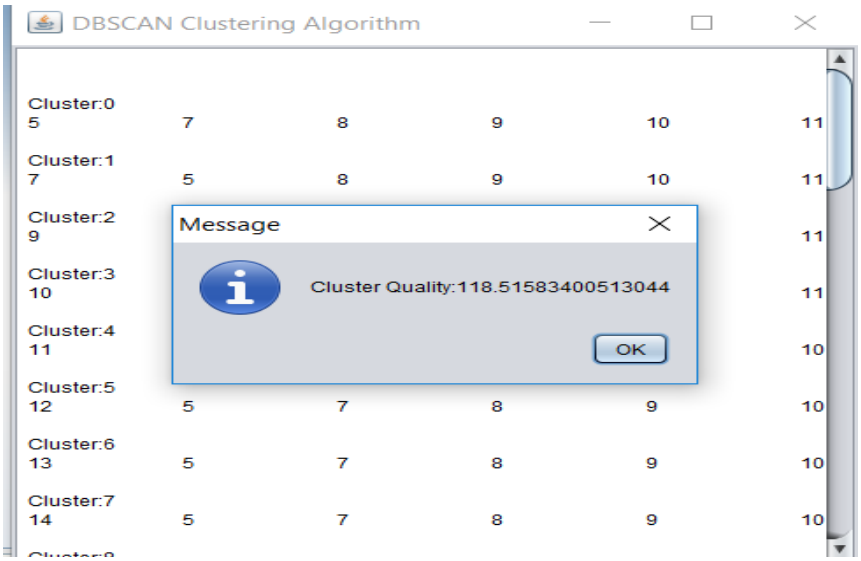

Fig. 6: Experimental result

6.1 Experimental setup

Hardware

Intel core 2 Duo Processor T7500,
HDD with 80GB,
RAM with 512MB was used

Software

Java technology is used to implement the Distance and partition based clustering algorithm with the distribution. In this paper Netbean IDE and MySQL serverare used. Apply these two algorithms on dataset. In the table show the comparison between DBSCAN-KL based on Computational time, input parameters and clustering result (Number of a cluster).

7. CONCLUSION

In this paper shows clustering of uncertain data objects by considering distribution similarity. First systematically calculates KL divergence as similarity measurement. KL divergence used to measure distribution differences. Integrated KL divergence into the density-based clustering algorithm DBSCAN, showing the effectiveness of KL divergence. The experimental result shows that KL divergence can capture distribution differences which cannot be captured by geometrical distance.

In the feature solve the issue selection of input parameters Eps and MinPts through some approach that can help determine these values. Also, it may happen that we are missing some core points which may cause loss of objects so this could also be solved. Improve the Computational time of Clustering.

8. ACKNOWLEDGEMENT

I wish to thank all the people who have directly or indirectly helped me in completing Paper work successfully. I express my gratitude towards my project guide Prof. C. M. Jadhav and also towards Head of Computer Engineering Department for their valuable suggestions and constant guideline during this paper work also acknowledge the research work done by all researchers in this field over the Internet for maintaining valuable document and resource on the Internet.

9. REFERENCES

[1] Dr T. Velmurugan "Efficiency of K-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points". IJCTA, 2012
[2] Samir Anjani and Prof. Mangesh Wangjari. "Clustering of uncertain data object using improved K-Means algorithm" IJARCSSE, 2013
[3] Bin Jiang, Jian Pei, Yufei Tao and Xuemin Lin. "Clustering Uncertain Data Based on Probability Distribution Similarity" IEEE, 2013
[4] L. Kaufman and P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, 1990.
[5] W.K. Ngai, B. Kao, C.K. Chui, R. Cheng, M. Chau, and K.Y. Yip, "Efficient Clustering of Uncertain Data," Proc. Sixth Int'l Conf. Data Mining (ICDM), 2006.
[6] B. Kao, S.D. Lee, D.W. Cheung, W.-S. Ho, and K.F. Chan, "Clustering Uncertain Data Using Voronoi Diagrams," Proc. IEEE Int'l Conf. Data Mining (ICDM), 2008.
[7] R. Cheng, D.V. Kalashnikov, and S. Prabhakar, "Evaluating Probabilistic Queries over Imprecise Data," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2003.

© 2019, www.IJARIIT.com All Rights Reserved

[8] H.-P. Kriegel and M. Pfeifle, "Density-Based Clustering of Uncertain Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining (KDD), 2005
[9] H.-P. Kriegel and M. Pfeifle, "Density-Based Clustering of Uncertain Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining (KDD), 2005.
[10] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," Proc. Second Int'l Conf. Knowledge Discovery and Data Mining (KDD), 1996
[11] C. C. Aggarwal and P. S. Yu, "Outlier detection with uncertain data," in Proceedings of the SIAM International Conference on Data Mining (SDM 2008), Atlanta, Georgia, USA, 2008, pp. 483–493.
[12] [L. Billard and E. Diday, Symbolic Data Analysis. Chichester, England: Wiley, 2006.
[13] L. A. Zadeh, "Fuzzy sets as a basis for a theory of possibility," Fuzzy Sets and Systems, vol. 1, pp. 3–28, 1978.
[14] J. Gebhardt, M. A. Gil, and R. Kruse, "Fuzzy set-theoretic methods in statistics," in Fuzzy sets in decision analysis, operations research and statistics, R. Slowinski, Ed. Boston: Kluwer Academic Publishers, 1998, pp. 311–347.
[15] S.D. Lee, B. Kao, and R. Cheng, "Reducing Uk-Means to kMeans," Proc. IEEE Int'l Conf. Data Mining Workshops (ICDM), 2007.
[16] G. Shafer, A mathematical theory of evidence. Princeton, N.J.: Princeton University Press, 1976. [28] P. Smets, "The combination of evidence in the Transferable Belief Model," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 5, pp. 447–458, 1990. [29] P. Smets and R. Kennes, "The Transferable Belief Model," Artificial Intelligence, vol. 66, pp. 191–243, 1994.
[17] W.K. Ngai, B. Kao, C.K. Chui, R. Cheng, M. Chau, and K.Y. Yip, "Efficient Clustering of Uncertain Data," Proc. Sixth Int'l Conf. Data Mining (ICDM), 2006.
[18] P. Smets and R. Kennes, "The Transferable Belief Model," Artificial Intelligence, vol. 66, pp. 191–243, 1994.
[19] A.D. Sarma, O. Benjelloun, A.Y. Halevy, and J. Widom, "Working Models for Uncertain Data," Proc. Int'l Conf. Data Eng. (ICDE), 2006.
[20] P.B. Volk, F. Rosenthal, M. Hahmann, D. Habich, and W. Lehner, "Clustering Uncertain Data with Possible Worlds," Proc. IEEE Int'l Conf. Data Eng. (ICDE), 2009.

© 2019, www.IJARIIT.com All Rights Reserved