Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 08, August -2018

CLUSTERING ON UNCERTAIN DATA BASED PROBABILITY DISTRIBUTION SIMILARITY

[1]Prof.C.M.Jadhav, [2]Vanashri S.Shinde

[1](Head of department, Bharat Ratna Indira Gandhi College of Engineering, BIGCE, Solapur, India)
[2] (Bharat Ratna Indira Gandhi College of Engineering, BIGCE , Solapur, India)

Abstract — *Clustering is important task in data mining. The main purpose of clustering is grouping the same object data in a huge dataset and finding similarities between the objects. Clustering on unsure data is a most difficult task in both modeling similarity between unsure data objects and producing efficient computational method. Clustering uncertain data problems have been solved by using many different new data mining techniques and various algorithms. Techniques have recently been suitable for clustering uncertain data based upon the traditional dividing clustering methods like k-means and density-based clustering methods like DBSCAN to unsure data, they will determined by geometric distances between objects. Computing the similarity between the data objects will be based upon a similarity distance measure and further clustered with occurrence based clustering or hierarchical clustering methods. Such methods cannot handle uncertain items that are geometrically no difference. In the proposed system we could using probability that are essential characteristics of uncertain objects, and are considered in measuring likeness between uncertain objects. The very popular technique Kullback-Leibler divergence used to procedures the distribution similarity between two uncertain data items. First the probability division method for model unsure data object then there after measure the similarity between data objects using distance metrics, then finally best clustering methods such as partition clustering, density clustering.*

Keywords- Clustering, Clustering uncertain data, density based clustering, partition clustering, KL-divergence

1. INTRODUCTION:

Clustering is one of important task in data mining to group the similar information or data. Every clustering algorithms aim of dividing the collection all data objects into subsets or similar clusters. A cluster is a collection of objects which are 'similar' between them and are 'dissimilar' to the objects which belongs to other clusters. Clustering of the data is classified in three ways: Dividing clustering approaches Density-based clustering approaches and possible world approaches. The main characteristics of uncertain data are, they change continuously, we cannot predict their pattern the accurate position of uncertain objects is not known and they are geometrically indistinguishable. Because of these reason it is quite difficult to cluster the unsure data by using the traditional clustering methods. Clustering is a job of partitioning a set of objects into a several group of meaningful subclasses is called cluster. Clustering is called as unsupervised category i. e. there isn't basically any predefined classes. A good clustering approach produces cluster with high quality. In which the similarity in intracluster is high and inter-cluster likeness is low. A great real life example is weather station monitors weather conditions it provides various measurements like temperature, humidity, wind flow, and direction. The daily weather record differs from day to day, which is often modeled as an uncertain object represented by a distribution over the space formed by several measurements. Essentially, we have to group the uncertain objects matching to their distributions.

The figure shows it is not possible to obtain the clusters C3, C2, C1, A, and {W|M|N} simultaneously by using one global density parameter value. A global density-based clustering approach would discover only of the clusters C1, C2, C3, A, and B. If uses global density to discover the cluster C1, C2, C3 then objects from {W|M|N} and A are recognized as noise.

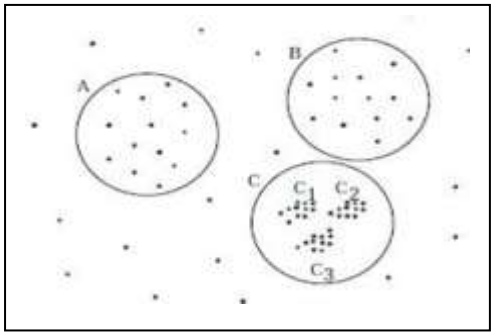

Figure: cluster with different density and Cluster within Cluster

2. RELATED STUDY:

Clustering uncertain data are mainly various extensions of the standard clustering algorithms made suitable for certain data. As an in a certain data set is target a one point, the distribution about the object itself is not considered in traditional clustering algorithms. Thus extended traditional algorithms to cluster unsure data are limited to using geometric distance-based likeness measures, and cannot get the difference between unsure objects with different allocation. Techniques including the density based clustering algorithm in [8] and hierarchical clustering algorithm in [9] are useful for working together with a specific application such as clustering or classification of information objects.

Various algorithms have been proposed for the clustering of uncertain data. Researchers are always attempting to increase the performance and efficiency of the grouped data. Doctor T. Velmurugan.[1] proposed a K-means algorithm to group the data. Here the given set of data is grouped into {E|T|P} number of disjoint groupings, in which the value of {E is usually to be T is usually to be |P is usually to be fixed in progress. The algorithm contains two separate phases: the first phase is to determine K initial centroids, one for each and every cluster. The next phase is to take each point belonging to the given data arranged and associate it to the nearest centroid. Generally Euclidean distance can be used as the measure to determine the distance between data and the centroids. After that the centroids are recalculated and clustering is completed with these new centroids. The process is repeated until the clusters are not changed. K-means method is not very much efficient to cluster the uncertain data, that is the key drawback.

Samir N. Ajani[2] proposes an better K-means algorithm to group the uncertain data effectively called UK means (Uncertain K means) clustering protocol. UK-means basically follows the well-known K-means algorithm other than that it uses Predicted distance (ED) calculation rather than using Euclidian distance. At first, k arbitrary point's c1... ck are chosen as the cluster representatives. After that, UK-means repeats the pursuing steps before the end result converges. First, for each and every data di, Expected Distance (di, cj) is computed for all centroids and data. Data di is then assigned to cluster cj that minimizes the Predicted Distance. Here the calculation of Expected distance requires numerically integrating functions, it is difficult to determine. The Expected distance calculations is one of the key problem in UK-means clustering. The efficiency of UK-means clustering is improved if the ED calculation is reduced.

Bin Jiang and Jian Pei.[6] proposed a new method for clustering uncertain data based on their possibility distribution similarity. The earlier methods extend traditional dividing clustering methods like K-means, UK means and density-based clustering methods to unsure data, thus rely on geometric distances between data. Probability distributions, which are essential characteristics of unsure objects. Here systematically model uncertain objects in both continuous and discrete domain names, where an uncertain thing is modelled as a continuous and discrete unique variable, respectively. Then use the well-known Kullback-Leibler (KL) divergence to measure likeness between uncertain objects in both the continuous and discrete cases, and combine it into K-medoid method to cluster uncertain data. Compared to the traditional clustering methods, K-Medoid clustering algorithm based on KL divergence similarity is more efficient.
Kriegel and Pfeifle [09] proposed the FDBSCAN formula which is a probabilistic extension of the deterministic DBSCAN algorithm [10] for clustering certain data. As DBSCAN is extended to a hierarchical density-based clustering method referenced to as OPTICS [3], Kriegel and Pfeifle [22] developed a probabilistic version of OPTICS called FOPTICS for clustering uncertain data items. FOPTICS outputs a hierarchical order in which data objects, rather than the decided clustering membership for each and every objects are clustered.

Volk et al. [36] followed the possible world semantics [1], [15], [8], [31] using Carlo sampling [17]. This approach finds the clustering of {a collection sampled possible sides using existing clustering methods for certain data. After that, the final clustering is aggregated from those group clustering.

3. METHODOLOGY:

Clustering is a primary data mining task. Clustering certain data has been considered for a long time in machine learning, data mining, pattern identity, Bioinformatics and other areas. Data uncertainty brings new challenges to clustering, since clustering uncertain data difficulty in the measurement of similarity between uncertain data objects. The majority of studies clustering uncertain data used distance-based similarity procedures and few theoretical studies considered using divergences to gauge the similarity between objects.
From this paper, we consider uncertain objects as random variables with certain distributions. We consider both the discrete case and the continuous case. In the continuous case, the domain is {a constant selection of values, for example, the temperatures recorded in a weather station are continuous real numbers.

3. 1 Uncertain Objects and Probability

We consider an uncertain object as a random variable following a probability distribution in a domain ID. We all consider both the individually distinct and continuous cases. If perhaps the domain is individually distinct (e. g., categorical) with a finite or countably infinite quantity of values, the object is an individually distinct random variable as

well as possibility distribution is described by a probability mass function (pmf for short). Normally, if the domain is continuous with a constant| variety of values, the subject is a consistent unique variable and its possibility distribution is described with a probability density function.

3. 2 KL Divergence

KL divergence measures how two distributions are different. It can be used to solution distribution difference between unsure objects by using possibility distribution of each and every object. The Kullback Leibler divergence is called as distance between two distributions.

1) In Discrete case, let f and g be two probability mass functions (pmf) in an individually distinct domain with finite amount values. The KL trick between f and g
2. In continuous reason, let f and g be two probability occurrence functions in a constant domain. The KL method between f nag G is

Calculate the likeness between two uncertain subjects by using Kullback Leibler divergence between their possibility distributions.

3. 3 Clustering Methods

Clustering uncertain subjects by considering the division of each object as the first class resident. Uncertain objects can have any discrete or constant distribution. We show that distribution dissimilarities between unsure objects cannot be decided by the sooner methods which are based on geometric distances. To solution the distribution difference between an uncertain objects by using recognized technique Kullback Leibler divergence. Demonstrate the effectiveness of KL method in both partitioning and density-based clustering methods.

3. 3. 1 DBSCAN Protocol

1. Select each unvisited point P from dataset.
2. Retrieve all items neighbours density reachable from P with respect to Eps (distance/radius) and minpts Kullback Leibler method.
3. If P is core point a group is formed.
4. Increase cluster until all neighbours points in cluster are processed.
5. If G is a border point, no points are obtainable from P and DBSCAN visits the next point of the dataset. Continue the process until all of the items have been processed without point can be included into any cluster.

3. 3. 2 Randomized K-Mediod Algorithm

The randomized k-medoids method, rather than finding the optimal non-representative object for swapping with representative subject, randomly selects a non-representative object for swapping if the clustering quality can be improved. The randomized k-medoids method works same in building and replacing framework. At the start, the building phase is conducted by selecting the initial k representatives article at random. Remaining thing i. e. not chosen objects are assigned to the most similar agent object according to KL divergence. Then perform replacing phase, in the replacing phase, recursively replace associates object by no- agent objects.

4. Hierarchical clustering techniques

DBSCAN is expanded to a hierarchical density-based clustering method referred to as OPTICS [36] by Kriegel. A great effective (deterministic) density established hierarchical clustering algorithm is OPTICS [36]. In this article, the core idea in OPTICS is quite similar to DBSCAN and it is depending on the principle of reachability distance between data points. While the method in DBSCAN identifies a large-scale density variable which can be used as a threshold in order to define reachability. It ensures the DBSCAN algorithm can be used for different values with this ordering, then a regular result is obtained.

4. 1 Partitioning Clustering

Dividing clustering K-means & K-medoids are two partioning methods. K-means algorithm be able to cluster the data. This technique is referred to as the UK-means algorithm. Ngai et al. [5] proposed the UK-means method extends the k-means method. The UK-means technique procedures the distance between an uncertain object and the cluster center (which is a certain point) by the expected distance. Lee et al. [7] showed that the UK-means method can be reduced to the k-means method on certain data points. In UK-means, an object is designated to the cluster in whose representative has the littlest expected distance to the object.

4. 4. 2 Density-Based Clustering Methods

As opposite to partitioning methods which coordinate similar objects into the same partitions to find out groupings, density-based clustering methods respectively clusters as dense parts of objects that are separated by regions of low density. DBSCAN [10] is the first and most consultant density-based clustering method developed for certain data. To demonstrate density-based clustering methods based on distribution likeness, we develop the unsure DBSCAN method which combine has a build-in} KL divergence into DBSCAN. Dissimilar to the FDBSCAN method [21] which is will be based upon geometric ranges and finds dense parts in the first geometric space.

4. CONCLUSION:

In this paper shows clustering of an uncertain data objects by considering distribution similarity. First of all calculates KL divergence as similarity measurement. KL method used to measure division differences. Integrated KL method into the density based clustering algorithm DBSCAN, displaying the effectiveness of KL divergence. This paper studies, the broad areas of probability based distribution likeness measurement associated with the field along with the key representational issues in uncertain data management. This represents both the dividing and density-based clustering methods with better clustering quality when using KL Leibler and other divergence as similarity than using distance metric.

5. 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.

[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.

