International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Clustering on Uncertain Data using Kullback Leibler Divergence Measurement based on Probability Distribution

Ajit Patil†* and M.D. Ingle†

†JSCOE, Department of Computer Engineering, Pune University, Hadpsar Pune.

Accepted 27 July 2015, Available online 28 July 2015, **Vol.5, No.4 (Aug 2015)**

Abstract

Cluster analysis is one of the important data analysis methods and is a very complex task. It is the art of a detecting group of similar objects in large data sets without requiring specified groups by means of explicit features or knowledge of data. Clustering on uncertain data is a most difficult task in both modeling similarity between uncertain data objects and developing efficient computational method. The most of the previous method for a clustering uncertain data extends partitioning clustering algorithms and Density based clustering algorithms. These methods are based on geometrical distance between two uncertain data objects. Such method not capable to handle uncertain objects, which are cannot distinguishable by using geometric characteristics and Distribution related to object itself is not considered. Probability distribution is a most important characteristic of uncertain object is not taking into account during measuring the similarity between two uncertain objects. The very popular technique Kullback-Leibler divergence used to measures the distribution similarity between two uncertain data objects. Integrates the effectiveness of KL divergence into both partition and density based clustering algorithms to properly cluster uncertain data. Calculation of KL-Divergence is very costly to solve this problem by using popular technique kernel density estimation and employ the fast Gauss transform method to further speed up the computation to decrease execution time

Keywords: *Uncertain data, Clustering, Fast-Gauss transformation, probabilistic distribution, KL-divergence.*

1. Introduction

Clustering is a fundamental, primary and initial step in data analysis. Clustering is called as an unsupervised classification of patterns into numbers of groups. Clustering methods are required in various areas such as mining skin lesion images, machine learning situations, pattern analysis and many other fields. It is difficult to analyze this vast amount of data by hand since data are massive and complex. Thus the goal is to obtain valuable knowledge by using advanced data analysis mechanisms that exploit the computing power available today. Clustering of a data is a process of organizing similar data objects into groups. A clustering algorithm divides a data set into numbers of group such that the similarity within a group is larger than among groups. In the domain of data mining, several clustering algorithms are proved for their clustering quality. Data uncertainty brings new challenges to clustering (Bin jiang, Jian Pei, 2013) since clustering uncertain data demands a measurement of similarity between uncertain data objects.

Uncertainty is generally caused by limited perception or understanding 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 real space to the pattern space.

Clustering is a task of partitioning a set of objects into a several set of meaningful subclasses is called cluster. Clustering is called as unsupervised classification i.e. there is not available any predefined classes. A good clustering technique produces cluster with high quality. In which the similarity in intra-cluster is high and inter-cluster similarity is low.

Uncertainty in the today's data comes with new challenges into the clustering of uncertain data. The

*Corresponding author: **Ajit Patil** is a M.E. Scholar and **M.D. Ingle** is working as Professor

most of previous studies in the field of a clustering uncertain data are based on improvement to the traditional clustering algorithms which are particularly designed for certain data. Any data object within certain data set is assumed as single point, the distribution of an uncertain object is not considered in traditional (previous) clustering algorithms. Thus, the existing studies are extended to the previous algorithms to cluster uncertain data. These existing methods are restricted to using geometrical distance based similarity measurement, and these earlier approaches cannot capture the difference between uncertain objects in the form 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 determine clusters with different sizes and shapes. The core idea of density based clustering is to cluster the objects based on density i.e. number of neighborhood objects exceed some threshold value. The earlier density based clustering algorithms are fails to discover clusters which are exist within cluster and is also does not work on varied density.

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

2. Literature Survey

The problem of clustering uncertain data has been studied in the recent years and find out feasible solutions on this problem. But the most of the earlier methods for clustering uncertain data are 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 depends geometric properties of data object. The geometric properties are used to measure similarity between object and they cannot consider the distribution similarity. There are three mostly used methods.

1. Partition based Clustering approaches.
2. Density based clustering approaches.
3. Possible world clustering approaches.

2.1 Partition based Clustering Approach

The partition based clustering approaches for clustering uncertain data extend the mostly used algorithm k-means method by using expected distance between uncertain objects to measure the similarity. It is required to first calculate expected distance between each pair of an object and select cluster representative iteratively. In Partition based algorithm have restriction on the user to specifying 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 a sample from an uncertain data set. Each of the possible worlds contains an instance from each object. Clustering process is applied individually on each possible world and the final clustering is performed by combining the clustering results of all possible worlds into a single global clustering. A sampled possible world of a data object cannot take into account the distribution of a data object, so 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 stable and meaningful clustering result at object level. It is computational infeasible because of there are 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 at the same region. But these objects have 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 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 minpts. It fails discovers clusters with varying density and it cannot discovers cluster which exist within cluster.

Partition based, possible world and density based clustering techniques for clustering uncertain data depends on the geometric characteristics of data object and they only focus on 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 a most important characteristic of uncertain object. Two uncertain objects may overlap but they have different distribution.

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 discover 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.

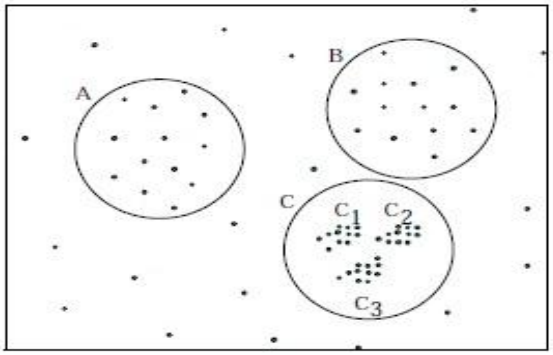

Fig 1: Cluster with different density and cluster within cluster

3. Related Work

Kullback Leibler divergence work on probability distribution of objects. So that it is required to first calculate probability distribution of each 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 on the quality of clustering results and should be reflected in the similarity. In order to capture 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 discrete domain with a finite number of values, the Kullback-Leibler divergence between p and q is defined below.

If domain is discrete with 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 distribution difference between uncertain objects by using probability distribution of each 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 a discrete domain with finite number values. The KL divergence between f and g

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

2. In 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 a 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. An 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*-center 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 made 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 each object as the 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 an 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 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 dataset. First calculates Kullback Leibler divergence between objects i.e. distribution distances. Then DBSCAN algorithm works based on KL divergence.

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

1. Select each unvisited point P from dataset.
2. Retrieve all points (neighborpts) density reachable from P with respect to Eps (distance/radius) and minpts according to the Kullback Leibler Divergence.
3. If P is core point a cluster is formed.
4. Expand cluster until all neighbor points in cluster are processed.
5. If P is a border point, no points are reachable from P and DBSCAN visits the next point of the dataset.
6. Continue the process until all of the points have been processed and no point can be included into any cluster.

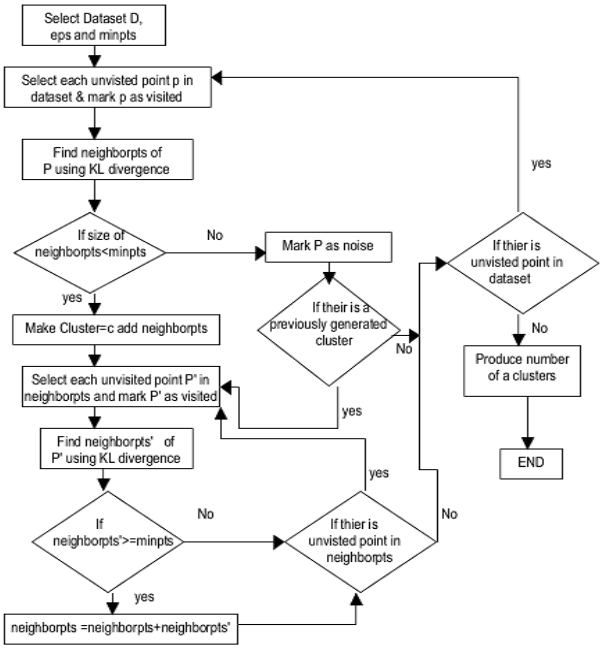

Fig 2: Flow Chart of DBSCAN algorithm

4.2 Randomized K-Mediod Algorithm

The randomized k-medoids method, instead of finding the optimal non representative object for swapping with representative object, 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 swapping framework. At 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 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 non representative object is randomly selected. Randomly selected non representative object is replaced with 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 representatives 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 tehnique 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 a two phases, first is a building phase and second is a swapping phase.

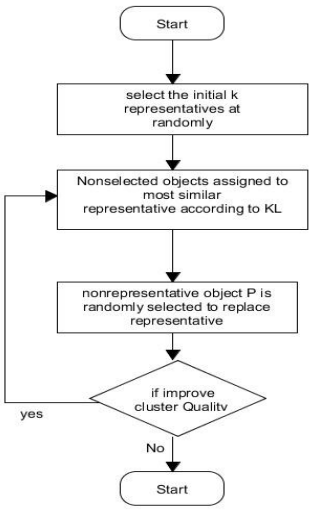

Fig 3: Flow chart of Randomized K-mediod Algorithm

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

Algorithm of 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

Swapping Phase: In the swapping phase the uncertain *k*-medoid method recursively improves the clustering quality by swapping a no representative data objects with the representative objects to which it is assigned.

Algorithm of swapping phase

Step1: Swapping the representative with the non representative data.

Step2: Repeat step 2, 3 and 4 of 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 affect the result and 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 (radius)
Minpts=minimum number of a points
D= {O1, O2, O3,-,--On}
P= {P1, P2, P3, P4, P5, P6, P7}
P= {Select input, select object, find, formation of cluster, expansion of cluster, check, produce cluster}

1. Select input(D, Eps, Minpts)
2. Select(each unvisited data point in D)
   Until (unvisited-list) = = null)
3. Find (neighborpts of P)
   Find neighbor points of P using KL divergence.
4. formation of cluster (Ci)
   if(neighborpts >= minpts) then Cluster Ci is formed.
   Other than noise will be considered and go to step P2
5. Expansion of Cluster (Ci)
   Select each unvisited point P' in neighborpts and apply process P2 to find neighborpts'.
   If (neighborpts'>= minpts) then
   Neighborpts= neighborpts + neighborpts'.
   Until (neighborpts=unvisited)
6. check unvisited(D)
   if (unvisited-list !=null) then go to P2.
7. Produce cluster.

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

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 used data set movie rating, the movie rating dataset is extracted from URL: - http://www.imdb.com/.

Table 1: Dataset information

Number of items	280067
Number of users	4300
Number of Records	116931

6.1 Experimental setup

1) Hardware

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

2) Software

Java technology is used to implement Distance and partition based clustering algorithm with distribution. In this paper Netbean IDE and MySQL server is 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).

mid	uid	rating
44310	1	7
39033	2	9
39033	3	7
44145	4	9
44145	6	7
44145	7	10
44014	8	8
44014	9	9
45362	10	8
44014	11	8
45362	12	8
44014	13	7
44014	15	10
44014	16	8
44014	17	8
44014	18	8
44014	19	10
44014	20	7
44014	21	7
44014	22	7
44014	23	10
44014	24	8
44014	25	7
44014	26	7
44014	27	6
44014	28	7

Fig 4: Movie rating Dataset

The below diagram shows calculation of Kullback Leibler Divergence as well as clustering result.

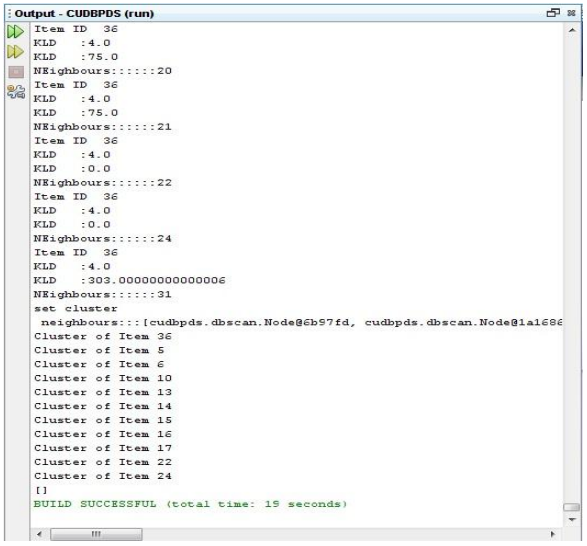

Fig 5: Uncertain DBSCAN with KL divergence.

Table 2: Clustering result DBSCAN algorithm with KL

	DBSCAN- KL
Numbers of record	280067
Eps (distance)/minpts	8/30
Run times	19 sec
Number of cluster	20

Conclusions

In this paper shows clustering of an 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.

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. M.D. Ingle 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 across the Internet for maintaining valuable document and resource on the Internet.

References

Banerjee, S. Merugu, I.S. Dhillon, and J. Ghosh,(2005)Clustering with Bregman Divergences, J. Machine Learning Research, vol. 6, pp. 1705-1749

Bin jiang, Jian Pei(2013) Clustering Uncertain data based on probability distrinbution similarity, IEEE.

N.N. Dalvi and D. Suciu,(2007) Reducing Uk-Means to k-Means Management of Probabilistic Data: Foundations and Challenges, Proc. ACM.

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu,(1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,Proc. Second Intl Conf. Knowledge Discovery and Data Mining (KDD).

J. Han and M. Kamber,(200) Data Mining: Concepts and Techniques,Elsevier,

R. Jampani, F. Xu, M. Wu, L.L. Perez, C.M. Jermaine, and P.J. Haas(2008) , Mcdb: A Monte Carlo Approach to Managing Uncertain Data, Proc. ACM.

B. Kao, S.D.Lee, D.W.Cheung, W.-S. Ho, and K.F. Chan,(2008) Clustering Uncertain Data Using Voronoi Diagrams,. Proc. IEEE Intl Conf. Data Mining (ICDM).

H.-P. Kriegel and M. Pfeifle, (2005) Density-Based Clustering of Uncertain Data,Proc. ACM SIGKDD Intl Conf. Knowledge Discovery in Data Mining (KDD).

H.-P. Kriegel and M. Pfeifle,(2005) Hierarchical Density-Based Clustering of Uncertain Data, Proc. IEEE Intl Conf. Data Mining (ICDM).

S. Kullback and R.A. Leibler, (1998) On Information and Sufficiency,The Annals of Math. Statistics.

S.D. Lee, B. Kao, and R. Cheng, (2007) Reducing Uk-Means to k-Means, Proc. IEEE Intl Conf. Data Mining Workshops (ICDM).

W.K. Ngai, B. Kao, C.K. Chui, R. Cheng, M. Chau, and K.Y. Yip, (2006) Efficient Clustering of Uncertain Data,,Proc. Sixth Intl Conf. Data Mining (ICDM).

B.W. Silverman (1986) Density Estimation for Statistics and Data Analysis,Chapman and Hall.

P.B. Volk, F. Rosenthal, M. Hahmann, D. Habich, and W. Lehner, (2009) Clustering Uncertain Data with Possible Worlds,, Proc. IEEE Intl Conf. Data Eng. (ICDE).

J. Xu and W.B. Croft, (1999) Cluster-Based Language Models for Distributed Retrieval, Proc. 22nd Ann. Intl ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR).

M. Ankerst, M.M. Breunig, H.-P. Kriegel, and J. Sander (1999) Optics: Ordering Points to Identify the Clustering Structure,Proc. ACM SIGMOD Intl Conf.

J. Pei, B. Jiang, X. Lin, and Y. Yuan,(2007) Probabilistic Skylines on Uncertain Data, Proc. 33rd Intl Conf. Very Large Data Bases (VLDB).

Michael Chau, Reynold Cheng,(2011) Uncertain Data Mining: An Example in Clustering Location Data, IEEE

Zhong Su, Qiang Yang, (2001) Correlation based Documnet clustering using Web Logs IEEE

C. Yang, R. Duraiswami, N.A. Gumerov, and L.S. Davis (2003), Improved Fast Gauss Transform and Efficient Kernel Density Estimation,Proc. IEEE Intl Conf. Computer Vision (ICCV).