# HANDLING UNCERTAINTY AND CLUSTERING IN UNCERTAIN DATA BASED ON KL DIVERGENCE TECHNIQUE

Reshma MR[1]
[1]MTech Student Computer Science Department,
KMEA Engineering College, Kerala, India
[1]reshemmar@gmail.com

Suchismita Sahoo[2]
[2]Asst.Professor Computer Science Department,
KMEA Engineering College, Kerala, India
[2]suchismita.sh@gmail.com

**Abstract - Data uncertainty is an inherent property in various applications due to reasons such as outdated sources or imprecise measurement. Data mining problems are significantly influenced by the uncertainty in these underlying data. Clustering is one of the most comprehensively studied problems in the uncertain data mining literature. Techniques have been designed for clustering uncertain data based on the traditional partitioning clustering methods like k-means and density-based clustering methods like DBSCAN to uncertain data, they rely on geometric distances between objects. Such methods cannot handle uncertain objects that are geometrically indistinguishable. In the proposed system we are using probability distributions, which are essential characteristics of uncertain objects, and are considered in measuring similarity between uncertain objects. The well-known Kullback-Leibler divergence is used to measure similarity between uncertain objects and integrate it into partitioning and density-based clustering methods to cluster uncertain objects. The proposed system aims to handle the tolerance/uncertainty of uncertain data so we are using FCM clustering algorithm for tolerance/uncertainty and check the validity of the clusters, which are obtained after modelling uncertain objects using the KL divergence.**

**Keywords – Clustering, Uncertain data, Cluster analysis, Cluster validity**

## I. INTRODUCTION

Managing uncertain data has been studied ever since the eighties last century from the database society. With the emergence of many recent important and novel applications involving uncertain data, there has been a great deal of research attention dedicated to this field. The applications include data cleaning, data integration, information extraction, sensor networks, economic decision making, market surveillance, trend prediction, moving object management, etc. Uncertainty is inherent in such applications due to various factors such as data randomness and incompleteness, limitation of equipment, and delay or loss in data transfer. Clusterization has been well studied in data mining research. However, only a few studies on data mining or data clustering for uncertain data have been reported. Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. Clustering problem is about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters For example, consider a retail database records containing items purchased by customers.

A clustering procedure could group the customers in such a way that customers with similar buying patterns are in the same cluster. Thus, the main concern in the clustering process is to reveal the organization of patterns into "sensible" groups, which allow us to discover similarities and differences, as well as to derive useful conclusions about them. This idea is applicable in many fields, such as life sciences, medical sciences and engineering. Clustering may be found under different names in different contexts, such as unsupervised learning (in pattern recognition), numerical taxonomy (in biology, ecology), typology (in social sciences) and partition (in graph theory) [1]

In the clustering process, there are no predefined classes and no examples that would show what kind of desirable relations should be valid among the data that is why it is perceived as an unsupervised process [2]. On the other hand, classification is a procedure of assigning a data item to a predefined set of categories.[3] Clustering produces initial categories in which values of a data set are classified during the classification process.

By the way, each data is generally handled as a point in a pattern space and analyzed in clustering procedures. However; there are many cases that data have errors, ranges or missing value of attributes. Typical example of uncertain data is as follows:

*Missing value of attributes:* When we handle data from the questionnaires, there are several questions which are not answered. This causes the missing value of attributes.

These uncertain data are represented by other form instead of a point [4], [5]. Even now, handling uncertain data have been actively studied in many researches [4]–[7]. These methods can not only handle uncertain data but also obtain high quality results by considering data with uncertainty. This means handling uncertain data is one of the significant problems in the field of data mining.

In this paper, we propose cluster validity measures for data with tolerance in order to apply these measures in case for uncertain data. Also, we are measuring uncertain data using the Kullback-Leibler divergence (KL divergence) and later we integrate these into existing data clustering methods so that the clustering results could be effective. Next, we use the concept of tolerance and fuzzy c-means clustering for data with tolerance (FCMT). After that, we use cluster validity measures for data with tolerance to show the effectiveness of proposed measures.

*A. EXISTING WORKS*

The previous studies on clustering uncertain data are largely various extensions of the traditional clustering algorithms designed for certain data. As an object in a certain data set is a single point, the distribution regarding the object itself is not considered in traditional clustering algorithms. Thus, the studies that extended traditional algorithms to cluster uncertain data are limited to using geometric distance-based similarity measures, and cannot capture the difference between uncertain objects with different distributions.

Specifically, three principal categories exist in literature, namely partitioning clustering approaches [17], [18], [19], density-based clustering approaches [20], [21], and possible world approaches [22]. The first two are along the line of the categorization of clustering methods for certain data [23], the possible world approaches are specific for uncertain data following the popular possible world semantics for uncertain data [24], [25], [26].

As these approaches only explore the geometric properties of data objects and focus on instances of uncertain objects, they do not consider the similarity between uncertain objects in terms of distributions.

*Partitioning clustering approaches* [17], [18], [14] extend the k-means method with the use of the expected distance to measure the similarity between two uncertain objects. The expected distance between an object P and a cluster center c (which is a certain point) is ED(P, c) $=\int pfp(x)$dist( x ,c )dx, where fP denotes the probability density function of P and the distance measure dist is the square of Euclidean distance. In [14], it is proved that ED(P, c) is equal to the dist between the center (i.e., the mean) P.c of P and c plus the variance of P. That is,

$$ED(P,c) = dist(P.c, c) + V ar(P) \qquad (1)$$

Accordingly, P can be assigned to the cluster center $\text{argmin}_c \{ED(P, c)\} = \text{argmin}_c\{dist(P.c, c)\}$. Thus, only the centers of objects are taken into account in these uncertain versions of the k-means method. In our case, as every object has the same center, the expected distance-based approaches cannot distinguish the two sets of objects having different distributions.

*Density-based clustering approaches* [20], [21] extend the DBSCAN method [27] and the OPTICS method [28] in a probabilistic way. The basic idea behind the algorithms does not change—objects in geometrically dense regions are grouped together as clusters and clusters are separated by sparse regions. However, in our case, objects heavily overlap. There is no clear sparse regions to separate object into clusters. Therefore, the density-based approaches cannot work well.

*Possible world approaches* [22] follow the possible world semantics [24], [25], [26]. A set of possible worlds are sampled from an uncertain data set. Each possible world consists of an instance from each object. Clustering is conducted individually on each possible world and the final clustering is obtained by aggregating the clustering results on all possible worlds into a single global

clustering. The goal is to minimize the sum of the difference between the global clustering and the clustering of every possible world. Clearly, a sampled possible world does not consider the distribution of a data object since a possible world only contains one instance from each object. The clustering results from different possible worlds can be drastically different. The most probable clusters calculated using possible worlds may still carry a very low probability. Thus, the possible world approaches often cannot provide a stable and meaningful clustering result at the object level, not to mention that it is computationally infeasible due to the exponential number of possible worlds.

*B. Motivation*

In the previous methods they extend traditional partitioning clustering methods like k-means and density-based clustering methods like DBSCAN to uncertain data, thus it rely on geometric distances between objects. Such methods cannot handle uncertain objects that are geometrically indistinguishable, such as products with the same mean but very different variances in customer ratings. Similarly the case when we handle a set of data, data contains inherent uncertainty e.g., errors, ranges or some missing value of attributes. The concept of tolerance has been proposed from the viewpoint of handling such uncertain data.

## II.PROPOSED SYSTEM MODEL

In the proposed system while compared to the existing system in which they use geometric distance to measure similarity of uncertain data we are using probability distributions. We cluster uncertain objects according to the similarity between their probability distributions. In uncertain data similarity measurement between two probability distributions can be measured by the Kullback-Leibler divergence.

Next we demonstrate the effectiveness of KL divergence in partitioning , density-based clustering method and Fuzzy c- means (FCM) to cluster uncertain objects. The first two clustering methods are used in the existing system were they rely on the geometric distances between objects in uncertain data. Here we are showing the effectiveness of using KL divergence as the measure for the similarity of uncertain data by integrating it with both partitioning and density based methods.

In the proposed system it also aims to handle the uncertainty (tolerance) of uncertain data. By using the concept of tolerance, data with uncertainty is defined. In addition clustering algorithm for data with uncertainty (tolerance) is introduced in FCM clustering algorithm and to check the validity of the clusters, which are obtained after modelling uncertain objects using the KL divergence. We are using cluster validity measures for data with uncertainty namely the Xie-Beni's index and the Fukuyama- Sugeno's Index.

*A. Uncertain Objects and KL Divergence*

This section first models uncertain objects as random variables in probability distributions. We consider the discrete probability distributions and show the evaluation of the corresponding probability mass and density functions. Then, we recall the definition of KL divergence, and formalize the distribution similarity between two uncertain objects using KL divergence

*B. Modeling Uncertain Objects and Probability Distributions*

If the domain is discrete (e.g., categorical) with a finite or countably infinite number of values, the object is a discrete random variable and its probability distribution is described by a probability mass function (pmf). Otherwise, if the domain is continuous with a continuous range of values, the object is a continuous random variable and its probability distribution is described by a probability density function (pdf). For example, the domain of the ratings of cameras is a discrete set {1, 2, 3, 4, 5}, and the domain of temperature is continuous real numbers.

By overloading the notation, for an uncertain object P, we still use P to denote the corresponding random variable, the probability mass/density function, and the sample. For discrete domains, the probability mass function of an uncertain object can be directly estimated by normalizing the number of observations against the size of the sample. Formally, the pmf of object P is given as follows.

$$P(x) = \frac{|\{p \in P | p = x\}|}{|P|} \quad (2)$$

where p ∈ P is an observation of P and |.|is the cardinality of a set.

*C. KL Divergence*

In general, KL divergence between two probability distributions is defined as follows:

Kullback-Leibler Divergence : Let f and g be two probability mass functions in a discrete domain ID with a finite or countably infinite number of

values. The Kullback-Leibler divergence between f and g is given as:

$$D(f\|g) = \sum_{x \in ID} f(x)\log(f(x)/(g(x)) \quad (3)$$

KL divergence is defined only in the case where for any x in domain ID if f(x) > 0 then g(x) > 0. By convention, $0 \log \frac{0}{p} = 0$ for any $p \neq 0$ and the base of log is 2. KL divergence is not symmetric in general, that is, D (f ||g) ≠ D (g||f).

*D. Using KL Divergence as Similarity*

It is natural to quantify the similarity between two uncertain objects by KL divergence. Given two uncertain objects P and Q and their corresponding probability distributions, D (P || Q) evaluates the relative uncertainty of Q given the distribution of P. In fact, from (3) we have

$$D(P \| Q) = E\left[\log \frac{p}{q}\right] \quad (4)$$

which is the expected log-likelihood ratio of the two distributions and tells how similar they are. The KL divergence is always nonnegative, and satisfies Gibbs' inequality. That is, D (P || Q) ≥ 0 with equality only if P = Q. Therefore, the smaller the KL divergence,the more similar the two uncertain objects.

In the discrete case, it is straightforward to evaluate (3) to calculate the KL divergence between two uncertain objects P and Q from their probability mass functions calculated as (2).

## III.CLUSTERING ALGORITHMS

As the previous geometric distance-based clustering methods for uncertain data mainly fall into two categories, partitioning and density-based approaches, in this section, we present the clustering methods using KL divergence to cluster uncertain objects in these two categories. In this Section , we present the uncertain k-medoids method which belong to popular partitioning clustering method by using KL divergence .Then in the next Section we presents the uncertain DBSCAN method which integrates KL divergence into the framework of a typical density-based clustering method DBSCAN. We describe the algorithms of the methods and how they use KL divergence as the similarity measure.

*A. Partitioning Clustering Methods*

A partitioning clustering method organizes a set of n uncertain object$\mathbb{D}$ into k clusters $\mathbb{C}$ 1; . . . ;$\mathbb{C}$ k, such that $\mathbb{C}$ i $\subseteq \mathbb{D}$ (1 ≤ i ≤ k), $\mathbb{C}$ i ≠ ∅ ; $U_{i=1}^{k}$ $\mathbb{C}$ i = $\mathbb{D}$ , and $\mathbb{C}$ i ∩ $\mathbb{C}$ j =∅ ; for any i ≠ j. We use Ci to denote the representative of cluster $\mathbb{C}$ i. Using KL divergence as similarity, a partitioning clustering method tries to partition objects into k clusters and chooses the best k representatives, one for each cluster, to minimize the total KL divergence as below

$$TKL = \sum_{i=1}^{k} \sum_{P \in \mathbb{C}i} D(P\|Ci) \quad (5)$$

For an object P in cluster $\mathbb{C}$ i (1 ≤ i ≤ k), the KL divergence D(P || Ci) between P and the representative Ci measures the extra information required to construct P given Ci. Therefore, $\sum_{P \in \mathbb{C}i} D(P\|Ci)$ captures the total extra information required to construct the whole cluster $\mathbb{C}$ i using its representative Ci. Summing over all k clusters, the total KL divergence thus measures the quality of the partitioning clustering. The smaller the value of TKL, the better the clustering.

*B. Density-Based Clustering Methods*

Unlike partitioning methods which organize similar objects into the same partitions to discover clusters, density-based clustering methods regard clusters as dense regions of objects that are separated by regions of low density.

DBSCAN [27] is the first and most representative density-based clustering method developed for certain data. To demonstrate density-based clustering methods based on distribution similarity, we develop the uncertain DBSCAN method which integrates KL divergence into DBSCAN. Different to the FDBSCAN method [20] which is based on geometric distances and finds dense regions in the original geometric space, the uncertain DBSCAN method transforms objects into a different space where the distribution differences are revealed.

The uncertain DBSCAN method finds dense regions through core objects whose $\epsilon$ neighbourhood contains at least $\mu$ objects. Formally, P is a core object, if

$$|\{Q \in \mathbb{D} | D(Q\|P) \leq \epsilon\}| \geq \mu \quad (6)$$

An object Q is said to be direct density-reachable from an object P if D(Q || P) ≤ ε and P is a core object.

Initially, every core object forms a cluster. Two clusters are merged together if a core object of one cluster is density reachable from a core object of the other cluster. A noncore object is assigned to the closest core object if it is direct density reachable from this core object. The algorithm iteratively examines objects in the data set until no new object can be added to any cluster.

The quality of the clustering obtained by the uncertain DBSCAN method depends on the parameters ε and μ .

The complexity of the uncertain DBSCAN method is $O(n^2E)$ where n is the number of uncertain objects and E is the cost of evaluating the KL divergence of two objects.[35]

## IV. CLUSTERING ALGORITHMS FOR DATA WITH TOLERANCE

This section discusses a concept of tolerance which handles data with uncertainty. By using this concept, data with uncertainty is defined as data with tolerance.

First, we introduce a concept of tolerance which handles data with uncertainty. Second, we introduce new clustering algorithms for data with tolerance, FCM clustering for data with tolerance (FCMT). After that, we propose cluster validity measures for data with tolerance.

### *A. Concept of Tolerance*

When we handle a set of data, we regard each data as one point. However, there are many cases that data have errors, ranges or missing value of attributes. Typical examples of data with uncertainty are as follows:

*Errors:* A round-off error, a truncation error, and a reading error margin are often caused.

*Ranges:* When we handle an apple as color data, we allocate an apple from the real space to the three dimensional pattern space which is constructed by RGB colors. The color of apple is not unique but with some ranges.

*Missing value of attributes:* When we handle data from the questionnaires, there are several questions which are not answered. This causes the missing value of attributes.

These data with uncertainty are represented by other form (e.g., an interval a probability density function ) instead of a point. Even now, handling data with uncertainty have been actively studied in many researches [9]. This means handling data with uncertainty is one of the significant problems in the field of data mining.

From the viewpoint of handling data with uncertainty, a concept of tolerance and clustering algorithms have been proposed [14]. In these algorithms, data with uncertainty are defined as data with tolerance.

First, we define tolerance and tolerance vector. A tolerance $\kappa k = (\kappa k1, \ldots, \kappa kp)^T$ means the admissible range of each data. A set of tolerance vector is defined as $E = \{\varepsilon 1, \ldots, \varepsilon n\}$ in which $\varepsilon k$, ($k = 1, \ldots, n$) is a tolerance vector. $\varepsilon 1, \ldots, \varepsilon n$ are vectors of real $p$-dimensional space $\Re p$. A tolerance vector $\varepsilon k \in \Re p$ is the vector with real components $\varepsilon k1, \ldots, \varepsilon kp$, that is, $\varepsilon k = (\varepsilon k1, \ldots, \varepsilon kp)T \in \Re p$. A tolerance vector is the vector within the range of tolerance. In the conventional studies, a data is represented as $xk$. On the other hand, data with tolerance is represented as $xk + \varepsilon k$ by using this concept.

A constraint for tolerance vector is as follows:

$$(\varepsilon kj)2 \le (\kappa kj)2 , (\kappa kj \ge 0), \ \forall k, j. \quad (7)$$

From these formulation, data with uncertainty is handled as data with tolerance [35].

### *B. Fuzzy c-Means Clustering for Data with Tolerance*

Next, we introduce standard fuzzy $c$-means clustering for data with tolerance (sFCMT).

The objective function of sFCMT $Jst$ (U,E, V ) is as follows:

$$J_{st}(U,E,V) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ki})^m d_{ki} \quad (8)$$

The constraint for $u_{ki}$ is as follows:

$$Uf = \{ (u_{ki}) : u_{ki} \in [0,1], \sum_{i=1}^{c} u_{ki} = 1, \forall k \} \quad (9)$$

The dissimilarity for clustering is generally the squared L2-norm between each data and a cluster center is as follows :

$$dki = ||xk - vi||^2 = \sum_{j=1}^{P}(x_{kj} - v_{ij})^2 \quad (10)$$

## V. CLUSTER VALIDITY MEASURES FOR DATA WITH TOLERANCE

The cluster validity measures for data with tolerance. we propose cluster validity measures for data with tolerance by introducing the concept of tolerance into conventional ones the Xie-Beni's index, the Fukuyama- Sugeno's index [15][35].

### *A. Xie-Beni's Index for Data with Tolerance*

The Xie-Beni's index for data with tolerance is represented as Tolerance Xie-Beni's index (TXB).The formula for TXB is as follows:

$$TXB = \frac{\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ki})^m \left\|x_k + \epsilon_k - v_i\right\|^2}{n \max_{i,j}\left\|v_i - v_j\right\|^2} \quad (11)$$

### *B. Fukuyama-Sugeno's Index for Data with Tolerance*

The Fukuyama-Sugeno's index for data with tolerance represented as Tolerance Fukuyama-Sugeno's (TFS).The formula for TFS is as follows:

$$TFS = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ki})^m \left\{\left\|x_k + \epsilon_k - v_i\right\|^2 - \left\|v_i - \tilde{v}\right\|^2\right\} \quad (12)$$

where, ˜v is the centroid of a set of data.

$$\tilde{v} = \frac{1}{n}\sum_{k=1}^{n}(x_k + \epsilon_k)$$

## VI. EXPERIMENTAL RESULTS

The experiments were conducted on real data sets, to evaluate the effectiveness of KL divergence as a similarity measure for clustering uncertain data and the efficiency of the clusters were evaluated using the cluster validity measures.

The real data obtained was a weather data set from the National Center for Atmospheric Research data archive [33].The data set extracted from weather data set consists of 1000 records. Each record has five dimensions, temperature, precipitation, humidity, wind speed and direction. The clustering of records are conducted with different clustering algorithms namely partitioning method, Density based method and FCM clustering. The outputs obtained shows four clusters along with the default cluster including the whole record. The four clusters are based on the temperature [4.1°C], precipitation [6.3°C], wind speed [7km/h] and the direction [W].

The parameters used for the system evaluation are time and space complexity, Precision and recall and cluster validity measures. The validity of clusters are measured using the Xie-Beni's index, and the Fukuyama-Sugeno's index as we are using the FCM algorithm and the result of both shows that the FCM clustering for data with uncertainty shows more effectiveness while comparing with the other algorithm based on the partitioning clustering method. The same cannot be conducted using density based method as the values there obtained are density values.

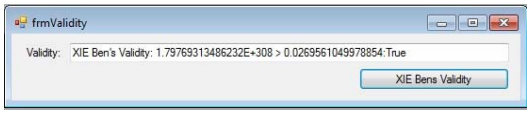


Table 6.1 The result of validity checking using the Xie Ben's index

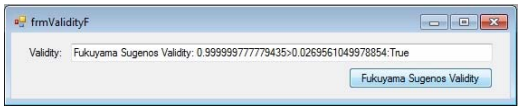


Table 6.2 The result of validity checking using the Fukuyama-Sugeno's index

The complexity measures of the algorithms are evaluated on the basis of its run time and the space utilized in the memory. The precision and recall for the algorithms are also evaluated; the density based method cannot be evaluated using the precision and recall as it contains the density values.

The Comparison table of time complexity obtained using partition method, FCM and density based method is shown in Table 6.3. Similarly the comparison chart of time complexity obtained using partition method, FCM and density based method is shown in Figure 6.1.

| Time complexity | Start(milli sec /count) | End(milli-sec /count) | Run(milli sec /count) |
|---|---|---|---|
| Partition method | 635131465349191151 | 635131465542012180 | 192821029 |
| FCM | 635131465134378865 | 635131465322879646 | 188500781 |
| Density-based method | 635131465582574500 | 635131465784156030 | 201581530 |

Table 6.3 The Comparison table of time complexity

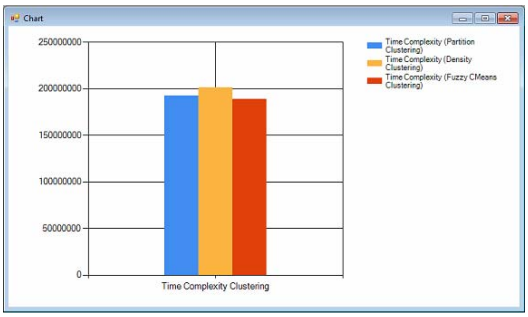


Fig 6.1 Chart for time complexity of clustering algorithms

The Comparison table of space complexity obtained using partition method, FCM and density based method is shown in Table 6.4 . Similarly the comparison chart of space complexity obtained using partition method, FCM and density based method is shown in Figure 6.2

| Space complexity | Start(milli sec /count) | End(milli-sec /count) | Run(milli sec /count) |
|---|---|---|---|
| Partition method | 1002 | 820 | 182 |
| FCM | 1002 | 820 | 182 |
| Density-based method | 1002 | 820 | 182 |

Table 6.4 The Comparison table of Space complexity

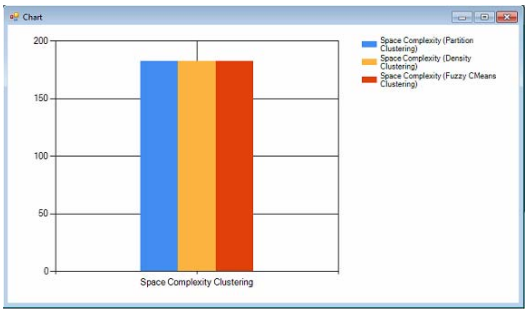


Fig 6.2 Chart for space complexity of clustering algorithms

The Comparison table of precision obtained using partition method and FCM is shown in Table 6.5. Similarly the comparison chart of precision obtained using partition method and FCM is shown in refer Figure 6.3

| Precision (%) | Partition method | FCM |
|---|---|---|
| | 0.249343206839091 | 0.249343206839091 |

Table 6.5 The Comparison table of Precision

The Comparison table of Recall obtained using partition method and FCM is shown in Table 6.6 Similarly the comparison chart of Recall obtained using partition method and FCM is shown in refer Figure 6.3

| Recall (%) | Partition method | FCM |
|---|---|---|
| | 0.5 | 0.5 |

Table 6.6 The Comparison table of Recall

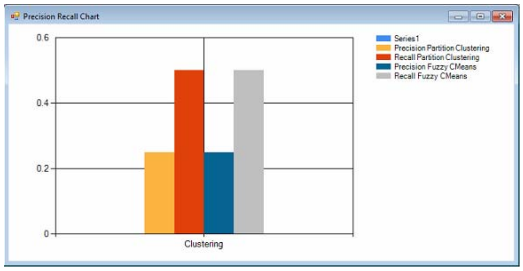


Fig 6.3 Chart for precision and recall of partition clustering and FCM

## VII. CONCLUSIONS

In this paper, we introduced the concept of tolerance used to define the uncertainty and clustering algorithms for data with uncertainty (tolerance) in order to handle uncertain data. Also cluster validity measures for data with uncertainty (tolerance). These proposed measures are quite different from conventional ones from the viewpoint of evaluating clustering results with uncertain data. Furthermore, it have verified the effectiveness of proposed measures with a real data set. Also it explore clustering uncertain data based on the similarity between their distributions. It used the Kullback-Leibler divergence as the similarity measurement. It also integrates KL divergence into the partitioning and density-based clustering methods to demonstrate the effectiveness of clustering using KL divergence. The experimental evaluation is conducted for partitioning clustering , density based method and FCM using the parameters namely time complexity , space complexity . the precision and recall and the cluster validity measures Xie-Beni's index and Fukuyamo-Sugeno's index. While comparing the time complexity and Validity measures, the experimental results proves that FCM performs better when comparing to partitioning clustering and density - based method. Also note that the results for space complexity are same for partitioning clustering, density based clustering and FCM whereas the precision and recall values of partitioning clustering and density based clustering are same.

## VIII. FUTURE WORK

In the future, besides clustering, similarity is also the fundamental significance to many other applications, such as nearest neighbour search, the study of those problems on uncertain data based on distribution similarity can be evaluated further. Similarly the discussion of cluster validity measures for data with tolerance has to be done from the viewpoint of evaluating classification results. Therefore, more comparisons between our proposed methods and conventional ones with other real data sets which include the missing value of attributes are quite important problems. The method of determining appropriate value of tolerance is also important. When these topics are developed, the proposed methods will be more effective and useful for uncertain data.

## ACKNOWLEDGMENT

I would like to express my gratitude to all those who gave me the possibility to complete this paper. I would like to thank the DEPARTMENT OF COMPUTER SCIENCE ENGINEERING, KMEA Engineering College for giving me support. I am bound to my parents for their stimulating support and encouragement which helped me in all the time of writing this paper. Particular thanks for A. Neela Madheswari, Associate Professor, Dept of CSE, KMEA for her suggestions on improving the paper at every step.

## REFERENCES

[1] Guha, S., Rastogi, R., and Shim K. (1998). CURE: An Efficient Clustering Algorithm for Large Databases. In *Proceedings of the ACM SIGMOD Conference*.

[2] Fayyad, M.U., Piatesky-Shapiro, G., Smuth P., Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI Press.

[3] Berry, M.J.A. and Linoff, G. (1996). *Data Mining Techniques For Marketing, Sales and Customer Support*. John Wiley & Sons, Inc., USA.

[4] R. J. Hathaway, J. C. Bezdek, Fuzzy c-means clustering of incomplete data, *IEEE Transactions on Systems, Man, and Cybernetics, Part B : Cybernetics*, Vol. 31, No. 5, pp. 735–744, 2001.

[5] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, K. Y. Yip, Efficient clustering of uncertain data, *Proc. of the Sixth International Conference on Data Mining* (ICDM2006), pp. 436–445, 2006.

[6] K. Honda, H. Ichihashi, Linear fuzzy clustering techniques with missing values and their application to local principal component analysis, *IEEE Transactions on Fuzzy Systems*, Vol. 12, No. 2, pp. 183–193, 2004.

[7] B. Kao, S. D. Lee, D. W. Cheung, W. Ho, K. F. Chan, Clustering uncertain data using voronoi diagrams, *Proc. of Eighth IEEE International Conference on Data Mining* (ICDM2008), pp. 333–342.

[8] Y. Endo, R. Murata, H. Haruyama, S. Miyamoto, Fuzzy c-means for data with tolerance, *Proc. of International Symposium on Nonlinear Theory and Its Applications* (Nolta2005), pp. 345–348, 2005.

[9] R. Murata, Y. Endo, H. Haruyama, S. Miyamoto, On fuzzy c-means for data with tolerance, *Journal of Advanced Computational Intelligence and Intelligent Informatics* (JACIII), Vol. 10, No. 5, pp. 673–681, 2006.

[10] Y. Hasegawa, Y. Endo, Y. Hamasuna, S. Miyamoto, Fuzzy c-means for data with tolerance defined as hyper-rectangle, *Proc. of Modeling Decisions for Artificial Intelligence* (MDAI2007), pp. 237–248, 2007

[11] Y. Hamasuna, Y. Endo, Y. Hasegawa, S. Miyamoto, Two clustering algorithms for data with tolerance based on hard c-means, *Proc. Of 2007 IEEE International Conference on Fuzzy Systems* (FUZZ-IEEE 2007), pp. 688–691, 2007.

[12] J. C. Dunn, Well separated clusters and optimal fuzzy partitions, *Journal of Cybernetics*, Vol. 4, pp. 95–104, 1974.

[13] I. Gath, A. B. Geva, Unsupervised optimal fuzzy clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 7, pp. 773–780, 1989.

[14] X. L. Xie, G. Beni, A validity measure for fuzzy clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13, No. 8, pp. 841–847, 1991.

[15] Y. Fukuyama, M. Sugeno, A new method of choosing the number of clusters for the fuzzy c-means method, *Proc. of fifth Fuzzy Syst. Symp.*, pp. 247–250, 1989 (in Japanese).

[16] D. L. Davies, D. W. Bouldin, A cluster separation measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.1, No.2, pp. 224–227, 1979.

[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] 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.

[19] S.D. Lee, B. Kao, and R. Cheng, "Reducing Uk-Means to k- Means," Proc. IEEE Int'l Conf. Data Mining Workshops (ICDM), 2007.

[20] 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.

[21] H.-P. Kriegel and M. Pfeifle,"Hierarchical Density-Based Clustering of Uncertain Data," Proc. IEEE Int'l Conf. Data Mining (ICDM), 2005.

[22] 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.

[23] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Elsevier, 2000.

[24] N.N. Dalvi and D. Suciu, "Management of Probabilistic Data: Foundations and Challenges," Proc. ACM SIGMOD-SIGACTSIGART Symp. Principles of Database Systems (PODS), 2007.

[25] A.D.Sarma, O. Benjelloun, A.Y. Halevy, and J. Widom, "Working Models for Uncertain Data," Proc. Int'l Conf. Data Eng. (ICDE), 2006.

[26] R. Jampani, F. Xu, M. Wu, L.L. Perez, C.M. Jermaine, and P.J. Haas, "Mcdb: A Monte Carlo Approach to Managing Uncertain Data," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2008.

[27] 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.

[28] M. Ankerst, M.M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: Ordering Points to Identify the Clustering Structure," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 1999.

[29] S. Kullback and R.A. Leibler, "On Information and Sufficiency," The Annals of Math. Statistics, vol. 22, pp. 79-86, 1951.

[30] S.P. Lloyd, "Least Squares Quantization in PCM," IEEE Trans. Information Theory, vol. IT-28, no. 2, pp. 129-137, Mar. 1982.

[31] J.B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," Proc. Fifth Berkeley Symp. Math. Statistics and Probability, 1967.

[32] L. Kaufman and P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, 1990.

[33] http://rda.ucar.edu/data sets/ds512.0/ : weather data set from the National Center or Atmospheric Research data archive

[34] Bin Jiang, Jian Pei, Senior Member, IEEE, Yufei Tao, Member, IEEE, and Xuemin Lin, Senior Member, IEEE "Clustering Uncertain Data Based on Probability Distribution Similarity" IEEE Transactions on knowledge and data engineering, vol. 25, no. 4, APRIL 2013
[35] Yukihiro Hamasuna, Yasunori Endo, and Sadaaki Miyamoto *Member, IEEE* "Cluster Validity Measures for Data with Tolerance" WCCI 2010 IEEE World Congress on Computational Intelligence July 2010

## AUTHORS PROFILE

**Reshma M R** received the B.Tech in Computer Science and Engineering from School of Engineering, (CUSAT) Cochin University of Science and Technology , Kerala, India during 2011. She is pursuing her M.Tech. in Computer Science and Engineering from KMEA Engineering College, MG University, Kerala, India. Her research interests are in Data Mining.

**Suchismita Sahoo** received her B.Tech in Computer Science and Engineering from ITER Bhubaneswar, India during 2009 and her M.Tech. during 2011 in Information Technology from USIT, GGSIP University , New Delhi ,India She is currently a Associate Professor of computer science and engineering at KMEA Engineering College, MG University, India.