International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 3, Issue 8, August 2014 **ISSN 2319 - 4847**

Clustering Multi-Attribute Uncertain Data Using Jenson-Shannon Divergence

Mr. V.V.Kulkarni[1], Prof V.V. Bag[2]

[1]PG.Student M.E (CSE) N. K. Orchid College of Engineering and Technology, Solapur-413002

[2]Associate Professor N. K. Orchid College of Engineering and Technology, Solapur-413002

ABSTRACTS

Clustering uncertain data is one of the essential tasks in mining uncertain data. Uncertain data contains the notion of probabilityand it is typically found in the area of sensor networks, weather data, customer rating data etc. The earlier methods for clustering uncertain data also use probability distribution as a similarity measure to cluster uncertain objects.In this paper, uncertain object in discrete domain is modeled, where uncertain object is treated as a discrete random variable. The probability distribution of uncertain object is calculated based on probability mass function. The Jenson-Shannon divergence is used to measure the similarity between two uncertain objects.The partitioning and density based clustering approaches are used to evaluate the performance of Jenson-Shannon Divergence. Experiments are performed to verify the effectiveness and efficiency of model developed and results are at par with the existing approaches.

Keywords: Clustering, Discrete Domain, Multi-Attribute Data Uncertain Data.

1. INTRODUCTION

Clustering has been studied for years in data mining, machine learning, pattern recognition, bioinformatics, recommendation systems and some other fields. The process of grouping a set of physical or abstract objects into classes of similarobjects is called clustering. A cluster is a collection of data objects that are similarto one another within the same cluster and are dissimilarto the objects in other clusters [7] [11]. In computer science, uncertain data is the notion of data that contains specific uncertainty. When representing such data in a database, it contains some indication of the probability of the occurrence of various values. Uncertain data exist in Temperature data, sensor data, and marketing data. Clustering uncertain data has been well recognized as an important issue [2][3][4][5].The problem of clustering multi-attribute uncertain data according to their probability distribution happens in many scenarios. Customer reviews on e-commerce website are considered as useful information for recommendation system. Customer reviews contains various information like customer personal information, product information and reviews information based on text data containing customer opinion about the product and a ratings given by a customer. In general, a particular rating on a particular product by customer gives the customer satisfaction to that product.

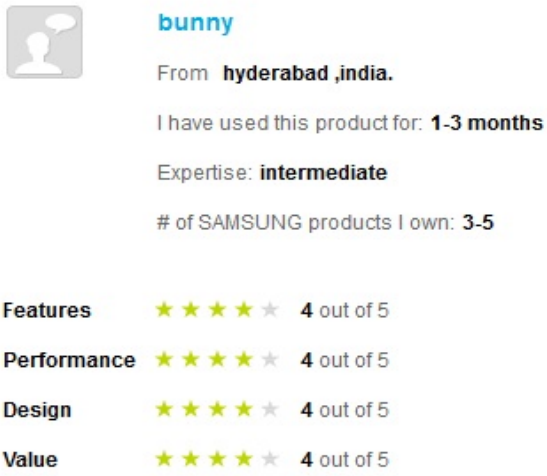

Figure 1: Sample of Customer ratings from e-commerce website

For example, the customer ratings on different mobiles, on different aspects of mobiles like features, performance, design, value etc. The figure 1 shows the samples of rating from an e-commerce website. Customer satisfaction to particular mobile can be considered as an uncertain object. Also two mobiles have same mean score, are substantially different if their score variances are different. To cluster the different mobiles based on their ratings on different aspect of each

mobile, it is needed to consider the similarity of both the mobiles on each aspect [12]. As another example, in recommendation system for generating list of good restaurants according to the consumer preferences. Consumer gives preferences or ratings on different services like overall rating, food rating, service rating etc. Consumer preferences to a particular restaurant can be considered as an uncertain object. Also, it is needed to consider the preferences on different aspects if also mean of two restaurants preferences is same, but their variances are different. Data uncertainty brings new challenges to clustering, since clustering uncertain data demands a measurement of similarity between uncertain data objects[8][9][10].

2. LITERATURE REVIEW

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, Density based clustering approaches, clustering with KL divergence. Third approach considers the probability distribution as similarity measure for clustering uncertain object and uses KL divergence [1].

2.1 Partitioning Clustering Approaches

Given D, a data set of n objects, and k, the number of clusters to form, a partitioning algorithm organizes the objects into kpartitions, where each partition represents a cluster. The clusters are formed to optimize an objective partitioning criterion, such as a dissimilarity function based on distance, so that the objects within a cluster are similar, whereas the objects of different clusters are dissimilar [7].Partitioning clustering approaches extend the k-means method with the use of the expected distance to measure the similarity between two uncertain objects. The expected distance (1) between an object P and a cluster centre c (which is a certain point) is

$$ED(P,c) = \int_p fp(x)dist(x,c)dx \qquad (1)$$

Where fp denotes the probability density function of P and the distance measure dist is the square of Euclidean distance.UK-means basically follows the well-known K-Means algorithm except that it uses expected distance when determining which cluster an object should be assigned to [3]. The second algorithm uses the idea of min max distance pruning in UK-means with the objective of reducing the number of expected distance calculations.UK-means starts by randomly selecting k points as cluster representatives. Each object o_i is then assigned to the cluster whose representative Pj has the smallest expected distance from o_i $(ED(o_i, p_j))$ among all clusters. After the assignment, cluster representatives are recomputed as the mean of the centres of mass of the assigned objects [4].

2.2 Density Based Clustering Approaches

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density- based clustering algorithm. Density-based clustering methods have been developed to discover clusters with arbitrary shape. These typically regard clusters as dense regions of objects in the data space that are separated by regions of low density. DBSCAN grows clusters according to a density-based connectivity analysis. The algorithm grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise. It defines a cluster as a maximal set of density-connected points [7]. In many modern applications ranges e.g. clustering of moving objects or sensor database, only uncertain data is available. For instance, in the area of mobile services the objects continuously change their position, so that exact position information is often not available. Fuzzy distance measure is used for measuring similarity between two fuzzy objects [5]. In real applications there is often no sharp boundary between clusters so that fuzzy clustering is useful for such data. Membership degrees between one and zero are used in fuzzy clustering instead of crisp assignments of data to cluster. The algorithm FDBSCAN is based on an enhanced version of core object definition and the core object probability of an object O indicates the likelihood that O is core object. The probability definition of core object differs from traditional approach where the similarity between fuzzy objects is measured by their distance expectation values. [6]

2.3 Clustering Based on KL-Divergence

Clustering uncertain objects according to the similarity between their probability distributions occurs in many scenarios. In information theory, the similarity between two distributions can be measured by the Kullback-Leibler divergence (KL divergence for short, also known as information entropy or relative entropy). The distribution difference cannot be captured by geometric distances. Uncertain objects are considered as random variables with certain distributions and both the discrete case and the continuous cases are considered. An uncertain object is considered as a random variable following a probability distribution in domain D. Uncertain objects can have any discrete or continuous distribution. In the discrete case, the domain has a finite number of values, for e.g., the rating of a camera can only take a value in {1, 2, 3, 4, and 5}. In the continuous case, the domain is a continuous range of values, for e.g., the temperatures recorded in a

weather station are continuous real numbers. Kullback-Leibler divergence is used to measure the similarity between two distributions [1].

2.3.1 KL-Divergence

In the discrete case, let f and g be two probability mass functions in a discrete domain D with a finite or countable infinite number of values. The Kullback-Leibler divergence (KL divergence for short) between f and g is

$$(\|) = \sum_{x \in D} f(x) \log(f(x)/g(x \quad (2)$$

3. METHODOLOGY

The sources of Multi-attribute uncertain data are customer ratings on mobiles on different aspects of mobiles like features, performance, design, value etc. In such a scenario we need to consider the probability of occurrences of rating on each mobile on each aspects of a mobile. So to cluster mobiles based on ratings on different aspects we need to consider the probability distribution as a similarity measure.

3.1 Uncertain Objects and Probability Distribution

Uncertain object is considered as a random variable following a probability distribution in a domain D. Uncertain object is modeled only in discrete case. If the domain is discrete 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. For example, the ratings of mobile are a discrete set {1, 2, 3, 4, and 5}.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 (Probability mass function) of objects P is

$$P(x) = \frac{|(p \in P | p = x)|}{|P|} \quad (3)$$

Where $p \in P$ is an observation of P and $|.|$ is the cardinality of a set.

3.2 Jenson-Shannon Divergence

In probability theory and statistics, the Jensen–Shannon divergence is a popular method of measuring the similarity between two probability distributions. It is also known as information radius or total divergence to the average. It is based on the Kullback–Leibler divergence, with some notable (and useful) differences including that it is symmetric and it is always a finite value. The Jensen–Shannon divergence (JSD) is a symmetrized and smoothed version of the Kullback–Leibler divergence. It is defined by

$$JSD(P\|Q) = \frac{1}{2} D(P\|M) + \frac{1}{2} D(Q\|M) \quad (4)$$

Where $M = \frac{1}{2}(P + Q)$

3.3 Clustering Algorithms

Clustering Multi-attribute uncertain data falls into two categories, partitioning clustering approaches and density based approaches. In this section, we present the clustering methods using JS divergence to cluster Multi-attribute uncertain objects in these two categories. Section 3.3.1 contains details of partitioning clustering approaches and section 3.3.2 contains details of density based approaches .Also this section describe the algorithms for the methods and how they the measure the similarity between two uncertain objects.

3.3.1 Partitioning Clustering Approaches

A Partitioning clustering method organizes a set of n uncertain objects O into k clusters $C_1,\ldots C_k$, such that $C_i \subseteq O \ (1 \leq i \leq k), C_i = \Phi$, $\bigcup_{i=1}^{k} C_i = O$, and $C_i \cap C_j$ for any $i \neq j$.Using JS 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 divergence. Using divergence as a 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 divergence. In partitioning clustering approaches we adopt the K-Medoids method to demonstrate the performance. In section (a) describes the different phases in randomized uncertain K-Medoids algorithm.

(a) Randomized Uncertain K-Medoids Algorithm

The randomized Uncertain K-Medoids method contains two phases, building phase and swapping phase.

Building Phase: In the building phase, the method obtains an initial clustering by selecting k representatives randomly one after another. At the beginning the building phase is simplified by selecting the initial k representatives as random.

Non-selected objects are assigned to the most similar representative according to JS divergence. After assigning the non-selected objects to nearest representative, the total current error is calculated for this assignment i.e. current error.

Swapping Phase: In the swapping phase, the method iteratively improves the clustering by swapping a non-representative object with the representative to which it is assigned. Then, in the swapping phase, we iteratively replace representative by non-representatives objects. In each iteration, a non-representative object P is randomly selected to replace the representative C to which P is assigned. To determine whether P is a good replacement of c, we examine the two cases in swapping phase.

- If P' currently belongs to C, when C is replaced by P, we will assign P' to P or one of another k-1 existing representatives, to which P' is the most similar.
- If P' currently belongs to a representative C' other than C, and $JSD(P'\|P) < JSD(P'\|C'), P'$ is reassigned to P.

After all non-representative objects are examined; the total decrease of the total divergence by swapping P and C is recorded. In swapping phase non-representatives are swapped with the representative, the total error is calculated for this assignment i.e. changed error. Current error and changed error are compared for replacing the non-representative with representative. In section (1) contains the algorithm for the randomized uncertain K-Medoids clustering method.

(1) Randomized Uncertain K-Medoids Algorithm

Algorithm:

A Randomized Uncertain K-Medoids Algorithm for Partitioning Clustering Approach

Input:

- k: The number of Clusters
- D: A Dataset Containing n list of objects (items)
- Features: list of features
- Method: JS(Jenson-Shannon Divergence)

Output:

A set of k clusters.

Method:

1. Randomly choose the k objects in D as the initial representative objects;
2. repeat
3. Assign each remaining object to the cluster with the nearest representative object based on current method; calculate the total current error for this assignment.
4. Randomly select a non-representative object O_{random}.
5. Compute the total changed error based on current method for swapping the representative object O_j with O_{random}.
6. If current error $\geq$ changed error, then replace O_j with O_{random}, to form the new set of representative object.
7. Until the largest decrease in error and which can improve the current clustering.
8. End.

3.3.2 Density Based Clustering Approaches

DBCAN (Density-Based Spatial Clustering of Applications with Noise) is a density- based clustering algorithm. Density-based clustering methods have been developed to discover clusters with arbitrary shape. These typically regard clusters as dense regions of objects in the data space that are separated by regions of low density. DBSCAN grows clusters according to a density-based connectivity analysis. In section (a) describes the different terminology used in uncertain DBSCAN algorithm.

(a) Uncertain DBSCAN Algorithm

The Uncertain DBSCAN method finds dense regions through core objects whose Ɛ-neighbourhood contains at least μ objects. Formally P is a core object, if

$$|\{Q \in O | JSD(Q\|P) \leq \}| \geq \mu \qquad (5)$$

An object Q is said to be directly density-reachable from an object P if $JSD(Q\|P) \leq$ and P is a core object. Every core object forms a cluster and a non-core 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 dataset until no new objects can be added to any cluster. The quality of clustering for Uncertain DBSCAN algorithm depends upon the parameters Ɛ and μ. In section (1) algorithm for uncertain DBSCAN clustering is described.

(1) Uncertain DBSCAN Algorithm

Algorithm:

A Uncertain DBSCAN Algorithm for Density Based Clustering Approach

Input:

- Ɛ: The neighborhood within a radius Ɛ of a given object is called the Ɛ-neighborhood of the object.
- μ : minimum number of points
- D: A Dataset Containing n list of objects (items)

- Features: list of features
- Method: JS(Jenson-Shannon Divergence)

Output:
A set of clusters.
Method:

1. Check the Ɛ-neighborhood of object, if contains the µ minimum number of points, then the object is called as a core object.
2. Create Cluster: If object is core object then create cluster with this core object.
3. Expand Cluster: Evaluate the density-reachable objects from this core object and add the objects to current cluster if they are density reachable from this core object.
4. Merge Cluster: Two clusters are merged together if core object of one cluster is density-reachable from a core object of the other cluster.
5. Iteratively examine the objects in the dataset until no new object can be added to any cluster.

4.DATASET AND RESULTS

4.1 Synthetic Dataset

The dataset is generated only in discrete domain. The user ratings are collected on different mobiles like Samsung, Micromax, Nokia, and Lava etc from different websites like Samsung website, 91mobiles.com. The ratings on five different attributes or features of mobile like Features, Performance, Design, Value, Overall ratings are collected. The ratings are in the range of 1 to 5 and the total 100 mobiles are considered in evaluation i.e. 100 uncertain objects are used to evaluate the performance.The figure 2 shows the part of synthetic dataset used.

mobile	username	Features	Performance	Design	Value	Overall Rating
galaxy core	Desparado	5	5	5	5	5
galaxy core	Gujjuboy	4	4	5	4	4
galaxy core	Core	5	5	4	5	5
galaxy core	Sara	4	4	5	5	4
galaxy core	RaviD	2	4	4	4	3
galaxy core	Champ	4	3	4	3	3
galaxy core	new core	3	3	3	2	3
galaxy core	vicky	4	1	3	2	2
galaxy core	reddy	5	5	5	5	5
galaxy core	Tangu	3	4	4	4	3
galaxy core	bablu	2	1	4	2	1
galaxy core	rohit	4	4	5	5	4
galaxy core	iv_dlh	3	3	3	2	3
galaxy core	surya	4	3	5	4	4
galaxy core	par_b	5	3	4	4	1
galaxy core	Core	4	5	5	5	5
galaxy core	Sudipta Das	3	1	4	1	1
galaxy core	karthik	5	4	5	4	4
galaxy core	believes_in_t	3	2	4	3	3
galaxy core	djkrutesh	5	5	5	5	5
galaxy core	Nicci	3	3	4	3	3
galaxy core	Pratik	5	5	5	3	5

Figure 2: Part of synthetic dataset

4.2 Real Dataset

To perform experiments and analyze the results of our approach, a dataset on Restaurant and Consumer dataset from UCI Machine Repository. The dataset consist of total 130 restaurants and also contains different context information about the consumer and demographic information about the different restaurants. We used only the rating provided by the consumers on different restaurants. It contains ratings of consumer for different restaurants on different aspects like Overall rating, Food rating, Service rating etc. The ratings are in the range of (0, 1 and 2). The results are analyzed based on Randomized Uncertain K-Medoids algorithm and Uncertain DBSCAN algorithm, forJS-Divergence.

4.3 Results

Clustering quality can be verified by internal validation techniques and external validation techniques. Internal validation techniques use internal criteria for validation and external validation techniques uses the external criteria for validation. External validation techniques usesdifferent measures like precision, recall and F-Measure and which are applicable in our case.In our case as no ground truth clusters are available, so we generated the expected cluster or ground truth clusters by comparing the different features of a mobile and similarity between different features. In statistical analysis the

F-Measure is a measure of a test's accuracy. It considers both the precisionP and the recallR of the test to compute the score. The F-Measure score can be interpreted as a weighted average of the precision and recall, where an F-Measure score reaches its best value at 1 and worst score at 0. Let G denote the ground truth clustering, C is the clustering obtained by a clustering method. Two objects are called a pair if they appear in the same cluster in a clustering. We define
TP true positive, the set of common pairs of objects in both G and C;
FP false positive, the set of pairs of objects in C but not G;
FN false negative, the number of pairs of objects in G but not C
Then, the precision, recall and F-Measure of a clustering C are calculated as formula (6), (7), (8).

$$Precision(C) = |TP| / (|TP| + |FP|) \quad (6)$$

$$Recall(C) = |TP| / (|TP| + |FN|) \quad (7)$$

$$F\text{-}Measure = (2 \times Precision \times Recall) / (Precision + Recall) \quad (8)$$

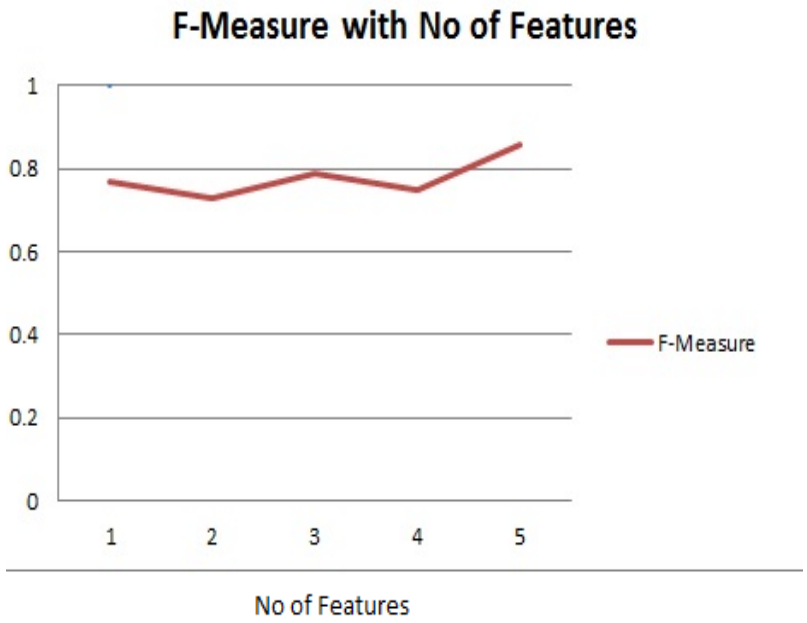

Figure 3: Comparison for Randomized Uncertain K-Medoids, F-Measure against the Number of Features

Table 1 contains the number of features and F-Measure values for randomized uncertain K-Medoids algorithm.

Table1: Comparison for Randomized Uncertain K-Medoids based on F-Measure values

No of Features	F-Measure
1	0.77
2	0.73
3	0.79
4	0.75
5	0.86

The clusters are formed based on similarity between one or more features. The performance of both the algorithms are evaluated based on number of features. The total five features are used for evaluations which are Features, Performance, Design, Value, and Overall Rating. The output of clustering algorithm varies according to number of features. Figure.3. shows comparison on F-Measure for Randomized K-Medoids algorithm for JS-Divergence against the number of features. Experiments are performed for number of cluster k=3.The performance of clustering is analyzed in both cases by increasing the number of features. For randomized uncertain K-Medoids method values of F-Measure are also depends upon the user input i.e. number of clusters.

Table 2: Comparison for Uncertain DBSCAN based on F-Measure values

No of Features	F-Measure
1	0.66
2	0.64
3	0.68
4	0.63
5	0.69

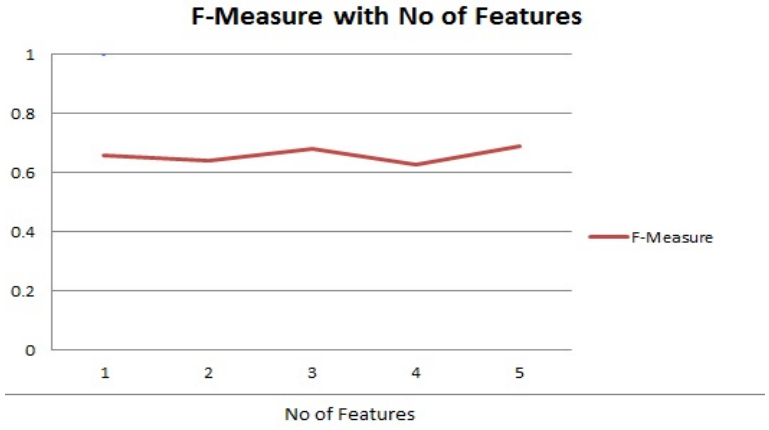

Figure 4: Comparison for Uncertain DBSCAN, F-Measure against the Number of Features

Table 2 contains the number of features and F-Measure values for uncertain DBSCAN algorithm.Figure.4. shows comparison on F-Measure for Uncertain DBSCAN algorithm for JS-Divergence against the number of features. F-Measure is measured against the number of clusters. The performance of Uncertain DBSCAN algorithm depends upon the two input parametersƐ-neighbourhood and µ-minimum number of objects. Experiments are performed for µ=3 and, variations in the values of Ɛ. The performance of clustering is analysed in both cases by increasing the number of features.

4. CONCLUSION

In this paper we explore clustering multi-attribute uncertain data based on probability distribution as a similarity measure. The results are analyzed based on Jenson-Shannon Divergence as a similarity measure. The data considered only in the discrete case, like multi-attribute data like customer ratings on different mobiles on different features of mobile. Jenson-Shannon Divergence is symmetric which measures the similarity of both the uncertain objects. Weintegrated Jenson-Shannon divergence in the partitioning clustering and density based clustering approaches.The contribution of this paper is to analyze the results of Jenson-Shannon Divergence as a similarity measure in discrete case and evaluation of the effectiveness of probability distribution as a similarity measure in real time datasets. In future, we will study problems related to evaluation policies related to clustering Multi-Attribute uncertain data and also variations in different kinds of ratings data.

REFERENCES

[1] Bin Jiang, Jian Pei, Yufei Tao and Xuemin Lin "Clustering Uncertain Data Based On Probability Distribution Similarity". IEEE Transaction On Knowledge and Data Engineering, 2013.
[2] J.Pei, B.Jiang, X.Lin and Y.Yuan "Probabilistic skylines on uncertain data". In VLDB, 2007.
[3] WangKayNgai, Ben Kao, ChunKitChui, Reynolds Cheng, Michael Chau, KevinY.Yip "Efficient Clustering Of Uncertain Data". In ICDM, 2005.
[4] B. Kao, S. D. Lee, D. W. Cheung, W.-S. Ho and K. F. Chan. Clustering uncertain data using voronoi diagrams. In ICDM, 2008.
[5] Hans-Peter Kriegel, Martin Pfeifle "Density Based Clustering of Uncertain Data". In KDD 2005.
[6] H.P.Kriegel and M. Pfeifle. Hierarchical density-based clustering of uncertain data. In ICDM, 2005.
[7] Jiawei Han, Micheline Kamber "Data Mining Concepts and Technique".
[8] A.Banerjee, S.Mergu, I.S.Dhillion, and J. Ghosh "Clustering Using Bregman Divergences". Journal of Machine Learning Research, 2003.
[9] R.Cheng, D. V. Kalashnikov, and S. Prabhakar. Evaluating probabilistic queries over imprecise data. In SIGMOD, 2003.
[10] N. N. Dalvi and D. Suciu. Management of probabilistic data: foundations and challenges. In PODS, 2007.
[11] A.K.JAIN Michigan State University, M.N.MURTHY Indian Institute of Science and P.J.FLYNN The Ohio State University "Data Clustering: A Review".
[12] For Dataset- Samsung website- http://www.samsung.com/in/consumer/mobile-phone/mobile-phone/smartphone/-reviews. And www.91mobiles.com.
[13] Real Dataset- Restaurant and Consumer dataset from UCI Machine Repository- https://archive.ics.uci.edu/ml/datasets/Restaurant+%26+consumer+data