A Literature Study On Traditional Clustering Algorithms For Uncertain Data

Abstract

Numerous traditional Clustering algorithms for uncertain data have been proposed in the literature such as k-medoid, global kernel k-means, k-mode, u-rule, uk-means algorithm, Uncertainty-Lineage database, Fuzzy c-means algorithm. In 2003, the traditional partitioning clustering algorithm was also modified by Chau, M *et al* to perform the uncertain data clustering. They presented the UK-means algorithm as a case study and illustrate how the proposed algorithm was applied. With the increasing complexity of real-world data brought by advanced sensor devices, they believed that uncertain data mining was an important and significant research area. The purpose of this paper is to present a literature study as foundation work for doing further research on traditional clustering algorithms for uncertain data, as part of PhD work.

Key words *: clustering algorithms, uncertain data, traditional partitioning*

Introduction

Aynur Dayanik, Craig G. Nevill-Manning (2004) have proposed a clustering approach by exploiting the relational structure of biological data to help with the next step: to enhance understanding of the data by combining techniques from information retrieval with those from bioinformatics. By computing over a network of sequence-structure-literature relationships it is possible to infer clusters of related articles, sequences and structures. They describe the general framework and its application to several biological domains. In the same year they proposed a uncertain version of the *k*-anonymity model which is related to the well known deterministic model of *k*-anonymity for the problem of privacy-preserving data mining. Most privacy transformations use some form of data perturbation or representational ambiguity in order to reduce the risk of identification. The final results from privacy transformation methods often require the underlying applications to be modified in order to work with the new representation of the data. While the results of privacy-transformation methods are a natural form of uncertain data, the two problems have generally been studied independently. The uncertain version of the *k*-anonymity model has the additional feature of introducing greater uncertainty for the adversary over an

equivalent deterministic model. As specific instantiations of this approach, we test the effectiveness of the privacy transformation on the problems of query estimation and classification, and show that the technique retains greater accuracy than other *k*-anonymity models.

Experimental study

M. Steinbach, G. Karypis, V. Kumar (2000) have presented the results of an experimental study of some common document clustering techniques: agglomerative hierarchical clustering and K-means. They used both a "standard" K-means algorithm and a "bisecting" K-means algorithm. Their results indicate that the bisecting K-means technique is better than the standard K-means approach and (somewhat surprisingly) as good or better than the hierarchical approaches that they tested. Y. Zhao and G. Karypis (2004) have proposed a method that evaluates the performance of different criterion functions in the context of partitioning clustering algorithms for document datasets. Their study involves a total of seven different criterion functions, three of which are introduced in their method and four that have been proposed in the past. They presented a comprehensive experimental evaluation involving 15 different datasets, as well as an analysis of the characteristics of the various criterion functions and their effect on the clusters they produce. Their experimental results show that there are a set of criterion functions that consistently outperform the rest, and that some of the newly proposed criterion functions lead to the best overall results. Their theoretical analysis shows that the relative performance of the criterion functions depends on (i) the degree to which they can correctly operate when the clusters are of different tightness, and (ii) the degree to which they can lead to reasonably balanced clusters.

New weighted dissimilarity measure for K-Modes

S.Aranganayagi and K.Thangavel (2009) proposed a new weighted dissimilarity measure for K-Modes, based on the ratio of frequency of attribute values in the cluster and in the data set. The new weighted measure is experimented with the data sets obtained from the UCI data repository. The results are compared with K-Modes and K-representative, which show that the new measure generates clusters with high purity. Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon and Joydeep Ghosh (2005) have proposed and analyzed parametric hard and soft clustering algorithms

based on a large class of distortion functions known as Bregman divergences. The proposed algorithms unify centroid-based parametric clustering approaches, such as classical kmeans, the Linde-Buzo-Gray (LBG) algorithm and information-theoretic clustering, which arise by special choices of the Bregman divergence. The algorithms maintain the simplicity and scalability of the classical k-means algorithm, while generalizing the method to a large class of clustering loss functions. This is achieved by first posing the hard clustering problem in terms of minimizing the loss in Bregman information, a quantity motivated by rate distortion theory, and then deriving an iterative algorithm that monotonically decreases this loss. In addition, they show that there is a bijection between regular exponential families and a large class of Bregman divergences, that they call regular Bregman divergences. This result enables the development of an alternative interpretation of an efficient EM scheme for learning mixtures of exponential family distributions, and leads to a simple soft clustering algorithm for regular Bregman divergences.

Hierarchical density-based clustering algorithm

Hans-Peter Kriegal and Martin Pfeifle (2005) have proposed hierarchical density-based clustering algorithm OPTICS (Ordering Points To Identify the Clustering Structure) has proven to help the user to get an overview over large data sets. When using OPTICS for analyzing uncertain data which naturally occur in many emerging application areas, e.g. location based services, or sensor databases, the similarity between uncertain objects has to be expressed by one numerical distance value. Based on such single-valued distance functions OPTICS, like other standard data mining algorithms, can work without any changes. They proposed to express the similarity between two fuzzy objects by distance probability functions which assign a probability value to each possible distance value. Contrary to the traditional approach, they do not extract aggregated values from the fuzzy distance functions but enhance OPTICS so that it can exploit the full information provided by these functions. The resulting algorithm FOPTICS helps the user to get an overview over a large set of fuzzy objects.

Probabilistic formulations of frequent item sets

Thomas Bernecker, Hans-Peter Kriegel, Matthias Renz, Florian Verhein and Andreas Zuefle (2009) have proposed new probabilistic formulations of frequent item sets

based on possible world semantics in Uncertain transactional databases. The consideration of existential uncertainty of item(sets), indicating the probability that an item(set) occurs in a transaction, makes traditional techniques inapplicable. In this probabilistic context, an itemset X is called frequent if the probability that X occurs in atleast minSup transactions is above a given threshold r . In consideration of the probabilistic formulations, they present a framework which is able to solve the Probabilistic Frequent Itemset Mining (PFIM) problem efficiently.

Two-layer approach for modeling and indexing technique

Das Sarma A., Benjelloun O., Halevy A. and Widom J. (2006) have proposed a two-layer approach for modeling and querying the uncertain data: an underlying logical model that is complete, and one or more *working models* that are easier to understand, visualize, and query, but may lose some information. They explore the space of incomplete working models, place several of them in a strict hierarchy based on expressive power, and study their closure properties. They describe how the two-layer approach is being used in our prototype DBMS for uncertain data, and they identify a number of interesting open problems to fully realize the approach. Samir N. Ajani and Prof. Mangesh Wanjari (2013) have proposed indexing technique for clustering the uncertain dataset. Probability Density Functions (PDF) is used to represent uncertain data objects. It has been observed that if the clustering algorithm is combined with indexing method then the clustering of uncertain data objects can be done very easily. They proposed a plan in which a K-means algorithm is used with Voronoi Diagram and indexing method to increase the performance of K-Means algorithm. In future this proposed plan can be implemented to prove the increased performance of K-Means algorithm. Voronoi diagram is an important technique for answering nearest-neighbor queries for spatial databases. To improve the performance of k-Means, this algorithm is combined with Voronoi diagram. They also studied how the Voronoi diagram can be used on uncertain data, which are inherent in scientific and business applications.

Quadratic penalty-vector regularization and fuzzy c-means algorithm

Y.Endo , Yasunori Endo , Yasushi Hasegawa , Yukihiro, Hamasuna and Yuchi Kanzawa (2011) have proposed clustering algorithm for uncertain data using quadratic penalty-vector regularization and fuzzy c-means algorithm. This method

provides optimization and helps in obtaining an optimal solution to handle uncertainty appropriately. The data uncertainties have been represented as interval ranges for which many clustering algorithms are constructed, but the lack of guidelines in selecting available distances in individual cases has made selection difficult and raised the need for ways to calculate dissimilarity between uncertain data without introducing a nearest-neighbor or other distance. They used tolerance concept which represents uncertain data as a point with a tolerance vector, not as an interval, while this is convenient for handling uncertain data, tolerance-vector constraints make mathematical development difficult. So , they attempted to remove the tolerance-vector constraints using quadratic penalty-vector regularization. The effectiveness was verified in numerical examples and derived appropriate penalty coefficient Wk corresponding to data distribution for individual algorithms.

Approximation by Single Gaussian and CURE

Lurong Xiao and Edward Hung (2007) have proposed an efficient method Approximation by Single Gaussian (ASG) to calculate the expected distance by a function of the means and variances of samples of uncertain objects. In the tasks such as clustering or nearest-neighbor queries, expected distance is often used as a distance measurement among uncertain data objects. Traditional database systems store uncertain objects using their expected (average) location in the data space. Distances can be calculated easily from the expected locations, but it poorly approximates the real expected distance values. Recent research work calculates the expected distance by calculating the weighted average of the pair-wise distances among samples of two uncertain objects. However the pair-wise distance calculations take much longer time than the former method. Theoretical and experimental studies show that ASG has both advantages of the latter method's high accuracy and the former method's fast execution time. They suggested that ASG plays an important role in reducing computational costs significantly in query processing and various data mining tasks such as clustering and outlier detection. S. Guha, R. Rastogi, and K. Shim (1998) have proposed a new clustering algorithm called CURE (Clustering using Representatives) that is more robust to outliers, and identifies clusters having non-spherical shapes and wide variances in size. CURE achieves this by representing each cluster by a certain fixed number of points that are generated by selecting well scattered points from the

cluster and then shrinking them toward the center of the cluster by a specified fraction. Having more than one representative point per cluster allows CURE to adjust well to the geometry of non-spherical shapes and the shrinking helps to dampen the effects of outliers. To handle large databases, CURE employs a combination of random sampling and partitioning. A random sample drawn from the data set is first partitioned and each partition is partially clustered. The partial clusters are then clustered in a second pass to yield the desired clusters. Their experimental results confirm that the quality of clusters produced by CURE is much better than those found by existing algorithms. Furthermore, they demonstrate that random sampling and partitioning enable CURE to not only outperform existing algorithms but also to scale well for large databases without sacrificing clustering quality.

Robust hierarchical clustering algorithm and t^1K-means algorithm

S. Guha, R. Rastogi, and K. Shim (2000) have proposed a new robust hierarchical clustering algorithm ROCK (Robust Clustering using linKs) for data with boolean and categorical attributes. They show that traditional clustering algorithms that use distances between points for clustering are not appropriate for boolean and categorical attributes. Instead, they propose a novel concept of links to measure the similarity/proximity between a pair of data points. Their methods naturally extend to non-metric similarity measures that are relevant in situations where a domain expert/similarity table is the only source of knowledge. In addition to presenting detailed complexity results for ROCK, they conducted an experimental study with real-life as well as synthetic data sets to demonstrate the effectiveness of their techniques. For data with categorical attributes, their findings indicate that ROCK not only generates better quality clusters than traditional algorithms, but it also exhibits good scalability properties. Michael Chau *et al* (2006) have implemented the t^1K-means algorithm, which aims at improving the accuracy of clustering by considering the uncertainty associated with data. Although in their paper, they only presented clustering algorithms for uncertain data with uniform distribution, the model was generalized to other distribution (e.g., by using sampling techniques). They also suggested that their concept of using expected distance was applied to other clustering approaches (such as nearest neighbor clustering and self-organizing maps) and other data mining techniques (such as data classification).

Extension of the density-based algorithm

Volk Habich, Clemens Utzny, Ralf Dittmann and Wolfgang Lehner (2007) have proposed an error-aware extension of the density-based algorithm DBSCAN. They applied this method especially in the field of photomask and semiconductor development. In this area, data, e.g. sensor data, has to be collected and analyzed for each process in order to ensure process quality. Furthermore, they presented some quality measures which could be utilized for further interpretation of the determined clustering results. With this new cluster algorithm, we can ensure that masks are classified into the correct cluster with respect to the measurement errors, thus ensuring a more likely correlation between the masks. Jianqiafig Yang and Steve Gunn (2007) have proposed a approach of input uncertainty classification. The approach have developed a technique which extends the support vector classification (SVC) by incorporating input uncertainties. Kernel functions was used to generalize that proposed technique to non-linear models and the resulting optimization problem was a second order cone program with a unique solution. Hae-Sang Park and Chi-Hyuck Jun (2008) proposed a new algorithm for K-medoid algorithm. It is a classical partitioning method to cluster the data. A partitioning clustering method organizes a set of uncertain data into K number of clusters. Their proposed algorithm calculates the distance matrix once and uses it for finding new medoids at every iterative step. To evaluate the proposed algorithm, they use some real and artificial data sets and compare with the results of other algorithms in terms of the adjusted Rand index. Experimental results show that this algorithm takes a significantly reduced time in computation with comparable performance against the partitioning around medoids.

Uncertainty-Lineage Database

Benjelloun et al (2006) proposed Uncertainty-Lineage Database (ULDB) which uses database with both uncertainty and lineage for clustering uncertain data. In data management applications, the influence of data source is an important factor that should be accounted. Zhang J. and Mani I. (2003) have described an application of simple kNN approach to a novel classification problem with an unbalanced class distribution. They empirically study the effects of under-sampling on the k nearest neighbour (kNN) approach. Their experimental results show that the kNN method is sensitive to number of negative examples selected and the random selection of

negative examples works better than other selection methods. Chau, M., Cheng, R., and Kao, B. (2005) proposed a framework new research direction in uncertain data mining. They proposed that when data mining is performed on uncertain data, data uncertainty has to be considered in order to obtain high quality data mining results. They presented the UK-means clustering algorithm as an example to illustrate how the traditional K-means algorithm can be modified to handle data uncertainty in data mining.

Traditional SVM classifier

Jinbo Bi and Tong Zhang (2005) have adapted the traditional SVM classifier to the uncertain data. They presented a general statistical framework to tackle the problem of noisy data. Based on the statistical reasoning, they proposed a formulation of support vector classification, which allows uncertainty in input data. They derived an intuitive geometric interpretation of the proposed formulation, and develop algorithms to efficiently solve it. Empirical results were included to show that the newly formed method was superior to the standard SVM for problems with noisy input. After this work, researchers tried to modify the various classification algorithms for uncertain data.Lamis Hawarah, Ana Simonet, and Michael Simonet (2006) dealt with the problem of missing values in decision trees during classification. Their approach is derived from the ordered attribute trees method, proposed by Lobo and Numao in 2000, which builds a decision tree for each attribute and uses these trees to fill the missing attribute values. Their method takes into account the dependence between attributes by using Mutual Information. The result of the classification process is a probability distribution instead of a single class. They explain their approach, we then present tests performed of our approach on several real databases and we compare them with those given by Lobo's method and Quinlan's method. They also measure the quality of our classification results. Finally, they calculate the complexity of our approach and we discuss some perspectives. Zufle A., Emrich T., Schmid K. A., Mamoulis N., Zimek A. and Renz M. (2014) have proposed a framework that targets the problem of computing meaningful clusterings from uncertain data sets, based on possible-worlds semantics; when applied on an uncertain dataset, it computes a set of representative clusterings, each of which has a probabilistic guarantee not to exceed some maximum distance to the ground truth clustering, i.e., the clustering of the actual

(but unknown) data. This framework can be combined with any existing clustering algorithm and it is the first to provide quality guarantees about its result. In addition, their experimental evaluation shows that their representative clusterings have a much smaller deviation from the ground truth clustering than existing approaches, thus reducing the effect of uncertainty.

Data model

Barbara, D., Garcia-Molina, H. and Porter D. (1992) have developed a data model that includes probabilities associated with the values of the attributes. The notion of missing probabilities is introduced for partially specified probability distributions. This model offers a richer descriptive language allowing the database to more accurately reflect the uncertain real world. Probabilistic analogs to the basic relational operators are defined and their correctness is studied. A set of operators that have no counterpart in conventional relational systems is presented. P. Langley, Wayne. lba, and K. Thompson (1992) presented an average-case analysis of the Bayesian classifier, a simple induction algorithm that fares remarkably well on many learning tasks. Their analysis assumes a monotone conjunctive target concept, and independent, noise-free Boolean attributes. They calculate the probability that the algorithm will induce an arbitrary pair of concept descriptions and then use this to compute the probability of correct classification over the instance space. The analysis takes into account the number of training instances, the number of attributes, the distribution of these attributes, and the level of class noise. They also explore the behavioral implications of the analysis by presenting predicted learning curves for artificial domains, and give experimental results on these domains as a check on our reasoning. Tian Zhang, Raghu Ramakrishnan and Miron Livny (1996) have proposed a data clustering method named BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), and demonstrates that it is especially suitable for very large databases. BIRCH incrementally and dynamically clusters incoming multi-dimensional metric data points to try to produce the best quality clustering with the available resources (i. e., available memory and time constraints). BIRCH can typically find a good clustering with a single scan of the data, and improve the quality further with a few additional scans. BIRCH is also the first clustering algorithm proposed in the database area to handle “noise)’ (data points that are not part of the underlying pattern) effectively.

Hybrid fuzzy clustering method

Hesam Izakian, Ajith Abraham, Vaclav Snasel(2009) have proposed a hybrid fuzzy clustering method based on FCM and fuzzy PSO (FPSO) is proposed which make use of the merits of both algorithms. Fuzzy c-means (FCM) algorithm is one of the most popular fuzzy clustering techniques because it is efficient, straightforward, and easy to implement. However FCM is sensitive to initialization and is easily trapped in local optima. Particle swarm optimization (PSO) is a stochastic global optimization tool which is used in many optimization problems. Experimental results show that their proposed method is efficient and can reveal encouraging results. Charu C. Aggarwal (2009) has discussed the problem of projected clustering of uncertain data streams. The use of uncertainty is especially important in the high dimensional scenario, because the sparsity property of high dimensional data is aggravated by the uncertainty. The uncertainty information is important for not only the determination of the assignment of data points to clusters, but also that of the valid projections across which the data is naturally clustered. The problem is especially challenging in the case where the data is not available on disk and arrives in the form of a fast stream. In such cases, the one-pass constraint in data stream computation poses special challenges to the algorithmic sophistication required for incorporating uncertainty information into the high dimensional computations. They have shown that the projected clustering problem can be effectively solved in the context of uncertain data streams.

Generalization of the k-median problem

Ackermann M.R., Blomer J. and Sohler C. (2008) have studied a generalization of the k-median problem with respect to an arbitrary dissimilarity measure D. Given a finite set P of size n, our goal is to find a set C of size k such that the sum of errors is minimized. Their algorithm requires time $n^{2O(mk \log(mk/\varepsilon))}$, where m is a constant that depends only on ε and D. Using this characterization, they obtain the first linear time $(1 + \varepsilon)$ approximation algorithms for the k-median problem in an arbitrary metric space with bounded doubling dimension, for the KL (Kullback-Leibler) divergence (relative entropy), for the Itakura-Saito divergence, for Mahalanobis distances, and for some special cases of Bregman divergences.Moreover, they obtain previously known results for the Euclidean k-median problem and the Euclidean k-means problem in a simplified manner. Charu C. Aggarwal and Philip S. Yu (2009) have discussed a

survey of uncertain data mining and management applications. They explored the various models utilized for uncertain data representation. In the field of uncertain data management, they examined traditional database management methods such as join processing, query processing, selectivity estimation, OLAP queries, and indexing. In the field of uncertain data mining, they examined traditional mining problems such as frequent pattern mining, outlier detection, classification, and clustering. We discuss different methodologies to process and mine uncertain data in a variety of forms. Achtert E., Kriegel H.P., Reichert L., Schubert E., Wojdanowski, R. and Zimek A. (2010) demonstrated a visualization tool based on a unification of outlier scores that allows to compare and evaluate outlier scores visually even for high dimensional data.

Clustering data objects with location uncertainty

Ngai W. K., Kao B., Cheng R., Chau M., Lee S. D., Cheung D. W. and Yip K. Y. (2011) have studied the problem of clustering data objects with location uncertainty. They proposed pruning methods that are based on metric properties (Met) and trigonometry (Tri) to speed up UK-means for clustering data with uncertainty . They studied two pruning methods namely pre-computation (PC) method which promotes reuse of pre-computed values and the cluster shift (CS) method that utilizes calculations made during clustering. In some cases the methods can prune more than 99.9% expected distance calculations. In their model, a data object is represented by an uncertainty region over which a probability density function (pdf) is defined. Romero C., Ventura S. and Espejo P.G. (2008) have developed a specific mining tool for making the configuration and execution of data mining techniques easier for instructors. They have used real data from seven Moodle courses with Cordoba University students. They have also applied discretization and rebalance preprocessing techniques on the original numerical data in order to verify if better classifier models are obtained. They also compare different data mining methods and techniques for classifying students based on their Moodle usage data and the final marks obtained in their respective courses. Finally, they propose that a classifier model appropriate for educational use has to be both accurate and comprehensible for instructors in order to be of use for decision making. Ben Kao ,Sau Dan Lee ,David W. Cheung, Wai-Shing Ho and K. F. Chan (2010) have presented the clustering algorithm based on Voronoi diagrams to reduce the number of expected distance calculation that is the drawback of

the UK-means clustering algorithm. Their techniques were analytically proven to be more effective than the basic bounding-box-based technique previously known in the literature. They introduced an R-tree index to organize the uncertain objects so as to reduce pruning overheads. They conducted experiments to evaluate the effectiveness of their novel techniques. Finally, they showed that their techniques were additive and, when used in combination, significantly outperform previously known methods.

Improved k-means algorithm

Shamir N Ajani (2013) proposed the improved k-means algorithm to handle the uncertainty data. Using the synthetic dataset, it clusters the data first to find the mean value and then applying to the nearer mean value in the cluster but it takes the large computation time and its variation depends on the initial k-value. Hence the indexing techniques are applied to the k-means algorithm and then the cluster generation time is significantly reduced. Sonia Manhas, Rajeev Vashisht, Parvinder S. Sandhu and Nirvair Neeru (2013) proposed a method that focuses on the assignment of cluster centroid selection so as to improve the clustering performance by K-Means clustering algorithm. Some of the problems associated with current clustering algorithms are that they do not address all the requirements adequately, and need high time complexity when dealing with a large number of dimensions and large data sets. This method uses Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance to assign for cluster centroid. Experimental result suggests that the proposed approach results in better clustering result when compared to the conventional technique.

U-relations

Antova L., Jansen T., Koch C. and Olteanu D. (2008) have introduced U-relations, a succinct and purely relational representation system for uncertain databases. U-relations support attribute-level uncertainty using vertical partitioning. If we consider positive relational algebra extended by an operation for computing possible answers, a query on the logical level can be translated into, and evaluated as, a single relational algebra query on the U-relational representation. The translation scheme essentially preserves the size of the query in terms of number of operations and, in particular, number of joins. Standard techniques employed in off-the-shelf relational database management systems are effective for optimizing and processing queries on U-

relations. In their experiments they show that query evaluation on U-relations scales to large amounts of data with high degrees of uncertainty. Reynold Cheng ,Yuni Xia, Sunil Prabhakar, Rahul Shah and Jeffrey Scott Vitter (2004) have proposed a method called Probabilistic Threshold Queries (PTQ) for clustering. They address the efficient computation of these types of queries. develop two index structures and associated algorithms to efficiently answer PTQs. The first index scheme is based on the idea of augmenting uncertainty information to an R-tree. They establish the difficulty of this problem by mapping one-dimensional intervals to a two-dimensional space, and show that the problem of interval indexing with probabilities is significantly harder than interval indexing which is considered a well-studied problem. To overcome the limitations of this R-tree based structure, they apply a technique we call variance-based clustering, where data points with similar degrees of uncertainty are clustered together. They used an extensive index structure can answer the queries for various kinds of uncertainty pdfs, in an almost optimal sense. They conduct experiments to validate the superior performance of both indexing schemes. Dhillon I.S., Mallela S. and Kumar R. (2003) proposed a new information theoretic divisive algorithm for feature/word clustering and apply it to text classification. Feature clustering is a powerful alternative to feature selection for reducing the dimensionality of text data. In comparison to the previously proposed agglomerative strategies divisive algorithm is much faster and achieves comparable or higher classification accuracies. They have shown that feature clustering is an effective technique for building smaller class models in hierarchical classification.

Kernel k-means

Grigorios F Tzortzis and Aristidis C Likas (2009) proposed the kernel k-means to handle the non-linearly separable cluster data and independent of initial k-value but it does not support the large dataset. The fast global kernel k-means algorithm supports the large dataset. William. W. Cohen (1995) have proposed evaluation of the rule learning algorithm IREP (Inceremental Reduced Error Pruning) on a large and diverse collection of benchmark problems. They show that while IREP is extremely efficient, it frequently gives higher error rates. They proposed a number of modifications to the IREP algorithm and found a new algorithm called RIPPERk rule. It reduces error rates and scales nearly linearly with the number of training examples, and can efficiently

process noisy datasets containing hundreds of thousands of examples. RIPPERk rule spilt the training datasets into growing set and pruning set. It uses the sequential covering approach to extract rules from the data. The algorithm extracts the rules one class at a time for a dataset. Let ($y1, y2, \ldots, yn$) be the ordered classes according to their frequencies, where $y1$ is the least frequent class and yn is the most frequent class. During the kth iteration, instances that belong to yi are labeled as positive examples, while those that belong to other classes are labeled as negative examples. A rule is desirable if it covers most of the positive examples and none of the negative examples.

Probabilistic query evaluation

Cheng R., Kalashnikov D. and Prabhakar S. (2003) have discussed probabilistic query evaluation based upon uncertain data. A classification of queries is made based upon the nature of the result set. For each class, they developed algorithms for computing probabilistic answers. They address the important issue of measuring the quality of the answers to these queries, and provide algorithms for efficiently pulling data from relevant sensors or moving objects in order to improve the quality of the executing queries. Extensive experiments are performed to examine the effectiveness of several data update policies. Shehroz S Khan and Dr. Shri Kant (2007) have proposed an approach to compute initial modes for K-mode clustering algorithm to cluster categorical data sets. Here, they utilize the idea of Evidence Accumulation for combining the results of multiple clusterings. Initially, n F − dimensional data is decomposed into a large number of compact clusters; the K-modes algorithm performs this decomposition, with several clusterings obtained by N random initializations of the Kmodes algorithm. The modes thus obtained from every run of random initializations are stored in a Mode-Pool, PN. The objective is to investigate the contribution of those data objects/patterns that are less vulnerable to the choice of random selection of modes and to choose the most diverse set of modes from the available Mode-Pool that can be utilized as initial modes for the K-mode clustering algorithm. Experimentally they found that by this method we get initial modes that are very similar to the actual/desired modes and gives consistent and better clustering results with less variance of clustering error than the traditional method of choosing random modes.

Achtert E., Goldhofer S., Kriegel H.P., Schubert E. and Zimek A. (2012) have provided a tool to visually support the assessment of clustering results in comparing multiple clusterings. Along the way, the suitability of a couple of clustering comparison measures can be judged in different scenarios. This tool provides best results when comparing with any evaluation metric which breaks down the available information to a single number. A lot of evaluation metrics are around, that are not always concordant nor easily interpretable in judging the agreement of a pair of clusterings.

New algorithm called DAGer

Dan Olteanu and Sebastiaan J. van Schaik (2012) have proposed a new algorithm called DAGer to cluster uncertain data. DAGger can work on arbitrarily correlated data and can compute both exact and approximate clusterings with error guarantees. They demonstrated DAGger using a real-world scenario in which partial discharge data from UK Power Networks is clustered to predict asset failure in the energy network. Jampani R., Xu F., Wu M., Perez L.L., Jermaine C.M. and Haas P.J. (2008) have proposed a new approach to handling enterprise-data uncertainty, embodied in a prototype system called MCDB approach , a system for managing uncertain data that is based on a Monte Carlo approach. MCDB represents uncertainty via "VG functions," which are used to pseudo randomly generate realized values for uncertain attributes. VG(Variable Generation) functions can be parameterized on the results of SQL queries over "parameter tables" that are stored in the database, facilitating what-if analyses. By storing parameters, and not probabilities, and by estimating, rather than exactly computing, the probability distribution over possible query answers, MCDB avoids many of the limitations of prior systems. For example, MCDB can easily handle arbitrary joint probability distributions over discrete or continuous attributes, arbitrarily complex SQL queries, and arbitrary functionals of the query-result distribution such as means, variances. In order to achieve good performance, MCDB uses novel query processing techniques, executing a query plan exactly once, but over "tuple bundles" instead of ordinary tuples. Experiments indicate that their enhanced functionality can be obtained with acceptable overheads relative to traditional systems.

Mining problem of clustering

Cormode G. and McGregor A. (2008) have studied the core mining problem of clustering on uncertain data, and defined appropriate natural generalizations of standard clustering optimization criteria. Two variations arise, depending on whether a data point is automatically associated with its optimal center, or whether it must be assigned to a fixed cluster no matter where it is actually located. In case of uncertain versions of k-means and k-median, they show reductions to their corresponding weighted versions on data with no uncertainties. These are simple in the unassigned case, but require some care for the assigned version. Their most interesting results are for uncertain k-center, which generalizes both traditional k-center and k-median objectives. They show a variety of bi-criteria approximation algorithms. Swagatam Das, Ajith Abraham, Amit Konar (2008) have proposed an application of DE(Differential Evolution) to the automatic clustering of large unlabeled data sets. In contrast to most of the existing clustering techniques, the proposed algorithm requires no prior knowledge of the data to be classified. Rather, it determines the optimal number of partitions of the data "on the run." Superiority of the new method is demonstrated by comparing it with two recently developed partitional clustering techniques and one popular hierarchical clustering algorithm. The partitional clustering algorithms are based on two powerful well-known optimization algorithms, namely the genetic algorithm and the particle swarm optimization. An interesting real-world application of the proposed method to automatic segmentation of images is also reported.

Building an R-tree index

S.Thirunavukkarasu and Dr.K.P.Kaliyamurthie (2014) have proposed method of building an R-tree index in order to organize uncertain objects. This can effectively reduce overheads pertaining to pruning. They discussed the process of clustering uncertain objects and the usage of PDFs (Probability Density Functions) to describe their locations is considered. they demonstrated UK-means algorithm is inefficient for clustering uncertain data. Biao Qin and Yuni Xia (2010) used a new rule based on classification and prediction technique for classifying the uncertain data. This algorithm introduces new measures for generating, pruning, and optimizing the rules. But it is not efficiently pruning the data. Based on the new measures, the optimal splitting attribute and splitting value can be identified and used for classification and

prediction. Their proposed u-Rule algorithm can process uncertainty in both numerical and categorical data. Their experimental results show that u-Rule has excellent performance even when data is highly uncertain.

Dynamic cost-sensitive fussy clustering approach

B. Kao, R. Cheng, M. Chau, S. D. Lee, D. W. Cheung and K. Y. Yip (2015) have proposed a dynamic cost-sensitive fussy clustering approach for uncertain data based on the genetic algorithm (GADCSFA). They gave definition of dynamic cost and adjacent interval, and the uncertain attributes are disposed as the interval number. Secondly, we give the method of fuzzy c-means clustering based on the interval data, and the interval numbers of fussy clustering solution and cost space are coded by its centre and radius. At last, the dynamic fussy clustering approach for uncertain data based on the genetic algorithm is structured, which uses the genetic algorithm to search the optimal clustering centre and cost by the hybridization, the mutation and selection. The experiments show that, compared to the other fussy clustering algorithm for uncertain data, GADCSFA has higher classification accuracy and performance, and the total expenditure is lower. Guru, D. S. and Nagendraswam, H. S. (2006) have proposed a novel similarity measure for estimating the degree of similarity between two symbolic patterns, the features of which are of interval type is proposed. A method for clustering data patterns based on the mutual similarity value (MSV) and the concept of k-mutual nearest neighbourhood is explored. The concept of mutual nearest neighbourhood exploits the mutual closeness possessed by the patterns for clustering thereby providing the naturalistic proximity characteristics of the patterns. Experiments on various datasets have been conducted in order to study the efficacy of the proposed methodology. Francesco Gullo, Giovanni Ponti, Andrea Tagarelli (2013) have proposed a novel formulation to the problem of clustering uncertain objects based on the minimization of the variance of the mixture models that represent the clusters being discovered. Analytical properties about the computation of variance for cluster mixture models are derived and exploited by a partitional clustering algorithm, called *MMVar*(Minimizing the variance). This algorithm achieves high efficiency since it does not need to employ any distance measure between uncertain objects. Experiments have shown that MMVar is scalable and outperforms state-of-

the-art algorithms in terms of efficiency, while achieving better average performance in terms of accuracy.

Uncertain object in discrete domain

Kulkarni V.V and Bag V.V (2014) have modeled uncertain object in discrete domain , where uncertain object is treated as a discrete random variable. The Jenson-Shannon divergence is used to measure the similarity between two uncertain objects and integrate it into partitioning and density based clustering approaches. Experiments are performed to verify the effectiveness and efficiency of model developed and results are at par with existing approaches. Joshua Zhexue Huang and Aranganayagi S (2015) used simple matching dissimilarity measure for categorical objects using k-mode algorithm with a heuristic approach which allows the use of the *k*-mode paradigm to obtain a cluster with strong intra-similarity, and to efficiently cluster large categorical data sets. It compares the two words and then finds the similarity and dissimilarity measures. The main aim of our paper is to derive rigorously the updating formula of the *k*-mode clustering algorithm with the new dissimilarity measure.

Clustering algorithm named CUDAP

Ping Jin , Shichao Qu , Yu Zong and Xin Li (2014) have proposed a novel clustering algorithm named CUDAP (Clustering algorithm for Uncertain Data based on Approximate backbone). In CUDAP, the steps are (1) make M times random sampling on the original uncertain data set D^m to generate M sampled data sets DS={Ds1,Ds2,…,DsM}; (2) capture the M local optimal clustering results P={C1,C2,…,CM} from DS by running UK-Medoids algorithm on each sample data set Ds_i, i=1,…M; (3) design a greedy search algorithm to find out the approximate backbone(APB) from P; (4) run UK-Medoids again on the original uncertain data set D^m guided by new initialization which was generated from APB. Experimental results on synthetic and real world data sets demonstrate the superiority of the proposed approach in terms of clustering quality measures. Ajit Patil and M.D. Ingle (2015) have proposed a 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. Lee S.D., Kao B. and Cheng R. (2007) have proposed a novel method for computing the EDs efficiently. It only works for a certain form of distance function. They developed an optimization to the UK-means algorithm, which generalizes the k-means algorithm to handle objects whose locations are uncertain. The location of each object is described by a probability density function (pdf). The UK-means algorithm needs to compute expected distances (EDs) between each object and the cluster representatives. The evaluation of ED from first principles is very costly operation, because the pdf 's are different and arbitrary. But UK-means needs to evaluate a lot of EDs. This is a major performance burden of the algorithm. They derived a formula for evaluating EDs efficiently. This tremendously reduces the execution time of UK-means, as demonstrated by our preliminary experiments. They illustrated that this optimized formula effectively reduces the UK-means problem to the traditional clustering algorithm addressed by the k-means algorithm.

Clustering algorithm named CUDAP

Le Li Zhiwen et al (2011) describes the automatic classification technique by soft classifier for the classification of uncertain data which appears in databases such as sensor ,location biometrics information databases with uncertainties. This data is generally imprecise in nature. This soft classifier technique is based on fuzzy c-means method with a fuzzy distance function to classify uncertain objects. The advantage of this method is that it works very well in uncertain data objects but not in certain data objects. Bin Wang, Gang Xiao, Hao Yu and Xiaochun Yang (2011) have proposed outlier detection method which is useful in many real time applications for clustering the uncertain data. They focused on distance-based outlier detection on uncertain data, in which each data is affiliated with a certain confidence value. They proposed a new definition of outlier on uncertain data. Based on the properties they discovered, both dynamic programming approach (DPA) and grid-based pruning approach(GPA) are used for detecting outliers on uncertain data efficiently. Detailed analysis and thorough experimental results demonstrate the efficiency and scalability of their method.

K-means algorithm

Velmurugan (2012) proposed a K-means algorithm to cluster the data. Here the given set of data is grouped into K number of disjoint clusters, where the value of K is to be fixed in advance. The algorithm consists of two separate phases: the first phase is to define K initial centroids, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to the nearest centroid. Generally Euclidean distance is used as the measure to determine the distance between data and the centroids. Then the centroids are recalculated and clustering is done with these new centroids. The process is repeated till the clusters are not changed . K-means method is not that much efficient to cluster the uncertain data, that is the main disadvantage.

Miss Pragati Pandey, Miss Prateeksha Pandey and Mrs. Minu Choudhary (2012) have presented uncertain data mining and management applications. They explored the various models utilized for uncertain data representation in the field of uncertain data management. They discussed different methodologies to process and mine uncertain data in a variety of forms. They examined traditional database management methods such as join processing, query processing, selectivity estimation, OLAP queries, and indexing for processing uncertain data. Samir N. Ajani, Prof. Mangesh Wanjari (2013) describes the Naive approach. It is for finding constraint frequent pattern from uncertain data and to find all the frequent patterns first approach , and checks these frequent patterns against the user constraints as a post-processing step – to filter out the patterns that do not satisfy the constraints.

Probability distribution similarity

Bin Jiang and Jian Pei (2013) proposed a new method for clustering uncertain data based on their probability distribution similarity. The previous methods extend traditional partitioning clustering methods like K-means, UK means and density-based clustering methods to uncertain data, thus rely on geometric distances between data. Probability distributions, are essential for characteristics of uncertain objects. Here systematically model uncertain objects in both continuous and discrete domains, where an uncertain object is modeled as a continuous and discrete random variable, respectively. Then use the well-known Kullback-Leibler (KL) divergence to measure similarity between uncertain objects in both the continuous and discrete cases, and integrate it into K-medoid method to cluster uncertain data. Compared to the

traditional clustering methods, K-Medoid clustering algorithm based on KL divergence similarity is more efficient.

Construction of wavelet decompositions

Y Zhao, C Aggarwal and P Yu (2010) have proposed a new method for the construction of wavelet decompositions of uncertain data on both the temporal and probabilistic aspects of the data, and designed a strategy for optimizing the relative effect of both components. Uncertain representations of data sets require significantly more space, and it is therefore even more important to construct compressed representations for such cases. They used a hierarchical optimization technique in order to construct the most effective partitioning for our wavelet representation. They explored two different schemes which optimize the uncertainty in the resulting representation. They show that the incorporation of uncertainty into the design of the wavelet representations significantly improves the compression rate of the representation. They presented experimental results illustrating the effectiveness of their approach.

Neural network method

Jiaqi Ge, and Yuni Xia (2010) have proposed a neural network method for classifying uncertain data (UNN). They extended the conventional neural networks classifier so that it was taken not only certain data but also uncertain probability distribution as the input. They started with designing uncertain perceptron in linear classification, and analyze how neurons use the new activation function to process data distribution as inputs. They illustrated how perceptron generates classification principles upon the knowledge learned from uncertain training data. They also constructed a multilayer neural network as a general classifier, and proposed an optimization technique to accelerate the training process.

Montecarlo database system

Volk P.B., Rosenthal F., Hahmann M., Habich D. and Lehner W. (2009) have proposed montecarlo database system (MCDB) approach was developed for uncertain data management based on the possible world scenario to overcome the problem of performance degradation in clustering uncertain data. As this methodology shows

significant performance and scalability enhancement, they adopt this method for the field of mining on uncertain data. They introduced a clustering methodology for uncertain data and illustrated current issues with this approach within the field of clustering uncertain data. Manisha Padole and Prof. Sonali Bodkhe (2016) have proposed the concept of FCM with Skew Divergence for clustering the uncertain dataset. The proposed Skew Divergence measure is quite different from conventional ones from the viewpoint of evaluating clustering results with uncertain data. They used Skew-Divergence as the similarity measure instead of Euclidean Distance which is the conventional method to calculate the distance between cluster center and data point. It provides better clustering results than the conventional methods. Bhumika Ingale and Priyanka Fulare (2014) have discussed various algorithms for clustering random data. They proposed hybrid algorithm for which gives better results for random data. A hybrid algorithm which combines hierarachical clustering with partitioning method was discussed.

Intuitive probabilistic approach

D.N. Campo, G. Stegmayera, D.H. Milonea (2016) have proposed a new index based on an intuitive probabilistic approach that is applicable to overlapped clusters. Given that recently there has been a remarkable increase in the analysis of data with naturally overlapped clusters, this new index allows to comparing clustering algorithms correctly. After presenting the new index, experiments with artificial and real datasets are shown and analyzed. Results over a real social network are also presented and discussed. The results indicate that the new index can correctly measure the similarity between two partitions of the dataset when there are different levels of overlap in the analyzed clusters.

MinMax k-Means algorithm

Grigorios Tzortzis and Aristidis Likas (2014) have proposed the MinMax k-Means algorithm, a method that assigns weights to the clusters relative to their variance and optimizes a weighted version of the k-Means objective. Weights are learned together with the cluster assignments, through an iterative procedure. The proposed weighting scheme limits the emergence of large variance clusters and allows high quality

solutions to be systematically uncovered, irrespective of the initialization. Experiments verify the effectiveness of our approach and its robustness over bad initializations, as it compares favorably to both k-Means and other methods from the literature that consider the k-Means initialization problem. Mrs. S. Vydehi and Dr. M. Punithavalli (2016) have proposed a method that combines Kullback-Leibler & Shennon Entropy (KLSE) in which the KL divergence method measures the similarity between objects using probability distribution and Shennon Entropy measures the uncertainty in a random variable. Herewith, in the dissimilarity between objects and inter-cluster distance are also considered to improve the cluster quality. Raymond T. Ng and Jiawei Han (2002) have proposed a new clustering method called CLARANS (Clustering Large Application Based upon RANdomized Search) , whose aim is to identify spatial structures that may be present in the data. Experimental results indicate that, when compared with existing clustering methods, CLARANS is very efficient and effective. CLARANS can handle not only point objects, but also polygon objects efficiently. One of the methods considered, called the IR-approximation, is very efficient in clustering convex and nonconvex polygon objects. Third, building on top of CLARANS, the paper develops two spatial data mining algorithms that aim to discover relationships between spatial and non spatial attributes. Both algorithms can discover knowledge that is difficult to find with existing spatial data mining algorithms. Yuni Xia and Bowei Xi (2007) have extended traditional conceptual clustering algorithm to explicitly handle uncertainty in data values. They proposed new total utility (TU) index for measuring the quality of the clustering. They developed improved algorithms for efficiently clustering uncertain categorical data, based on the COBWEB conceptual clustering algorithm. Experimental results using real datasets demonstrated how these algorithms and new TU measure can effectively improve the performance of clustering through the use of internal probabilistic information.

Based on genetic algorithm

C. Y. Liu (2013) have proposed a method on cost-sensitive clustering for uncertain data based on genetic algorithm (CSUDC). First, give the cost-sensitive learning for uncertain data. Use the interval to dispose continuous and discrete attribute of

uncertain data, so the traditional clustering method can cope with uncertain data. Second, a cost-sensitive clustering method for uncertain data is presented. Adopt the real encoding for the clustering data in genetic algorithm. The optimal cluster centers are searched by the selection, the crossover and mutation. The experimental results show, compared to the rest of the several common clustering method for uncertain data, CSUDC has higher accuracy of classification, and takes the total low cost in the clustering process. In recent years, naive bayes and neural network were modified to handle the uncertain data. Jiangtao Ren, Sati Dan Lee, Xianlu Chen. Ben Kao, Reynold Cheng and David Cheung (2009) have proposed a naive Bayes classification algorithm for uncertain data with a pdf. They addressed the problem of extending traditional naive Bayes model to the classification of uncertain data. They have extended the kernel density estimation method to handle uncertain data. For particular kernel functions and probability distributions, the double integral was analytically evaluated to give a closed-form formula, allowing an efficient formula-based algorithm. Extensive experiments on several UCI datasets showed that the uncertain naive Bayes model considering the full pdf information of uncertain data was produced classifiers with higher accuracy than the traditional model using the mean as the representative value of uncertain data. Time complexity analysis and performance analysis based on experiments showed that the formula-based approach has great advantages over the sample-based approach.

Conclusions

Numerous traditional Clustering algorithms for uncertain data have been proposed in the literature such as k-medoid, global kernel k-means, k-mode, u-rule, uk-means algorithm, Uncertainty-Lineage database, Fuzzy c-means algorithm. Many authors presented the UK-means algorithm as a case study and illustrate how the proposed algorithm was applied. With the increasing complexity of real-world data brought by advanced sensor devices, they believed that uncertain data mining was an important and significant research area. The purpose of this paper is to present a literature study as foundation work for doing further research on traditional clustering algorithms for uncertain data, as part of PhD thesis work.

Future study about the clustering algorithms with uncertain data

There are scopes for further research in this direction. We are working on finding out efficient methods to initialize the modes. In future, we will be able to expand the work by using a hypothesis as keeping aggregate of clusters as mode for clustering uncertain data in a more optimal manner. The long term goal of this approach is to provide new algorithms that use combination of B Trees or B+ trees for improving the efficiency of clustering results. We intend to apply this model in a wider variety of applications for different real time datasets.

REFERENCES

[1] Achtert E., Goldhofer S., Kriegel H.P., Schubert E. and Zimek A. (2012) , "Evaluation of clusterings – metrics and visual support", In Proc. ICDE, pp. 1285-1288.

[2] Achtert E., Kriegel H.P., Reichert L., Schubert E., Wojdanowski, R. and Zimek A. (2010), "Visual evaluation of outlier detection models", in Proc. DASFAA.

[3] Ackermann M.R., Blomer J. and Sohler C. (2008), "Clustering for Metric and Non-Metric Distance Measures", Proc. Ann. ACM-SIAM Symp. Discrete Algorithms (SODA).

[4] Aggarwal C.C. (2008), "On Unifying Privacy and Uncertain Data Models," Proc. 24th IEEE Int'l Conf. Data Eng. (ICDE).

[5] Aggarwal C.C. (2007), "On Density Based Transformations for Uncertain Data Mining", Proc. 23rd IEEE Int'l Conf. Data Eng. (ICDE).

[6] V. Vapnik. , "The Nature of Statistical Learning Theory.", Springer Verlag, 1995.

[7] Reynold Cheng ,Yuni Xia, Sunil Prabhakar, Rahul Shah and Jeffrey Scott Vitter (2004), "Efficient Indexing Methods for Probabilistic Threshold Queries over Uncertain Data", Proceedings of the 30th VLDB Conference, Toronto, Canada.

[8] Aggarwal C.C. and Yu P.S. (2008), "A Framework for Clustering Uncertain Data Streams", Proc. 24th IEEE Int'l Conf. Data Eng. (ICDE).

[9] Aggarwal C.C. and Yu P.S. (2008), "Outlier Detection with Uncertain Data", Proc. SIAM Int'l Conf. Data Mining (SDM).

[10] Antova L., Jansen T., Koch C. and Olteanu D. (2008) , "Fast and Simple Relational Processing of Uncertain Data", Proc. 24th IEEE Int'l Conf. Data Eng. (ICDE).

[11] Aranganayagi S.and Thangavel K. (2009), "Improved K-Modes for Categorical Clustering Using Weighted Dissimilarity Measure", International Journal of Information and Mathematical Sciences.

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

[13] Ben Kao ,Sau Dan Lee ,David W. Cheung, Wai-Shing Ho and K. F. Chan (2010), "Clustering Uncertain Data using Voronoi Diagrams", IEEE.

[14] Benjelloun O., Das Sarma A., Halevy A. and Widom J. (2006), "ULDBs: Databases with Uncertainty and Lineage", Proc. 32nd Int'l Conf. Very Large Data Bases (VLDB).

[15] Bernecker T., Kriegel H.P., Renz M., Verhein F. and Zufle A. (2009), "Probabilistic frequent itemset mining in uncertain databases", in Proc. KDD.

[16] Jinbo Bi and Zhang T. 2005, "Support vector classification with input data uncertainty", in Advances in Neural Information Processing Systems 17, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, pp. 161-168.

[17] Bin Jiang G, Jian Pei, Yufei Tao and Xuemin Lin (2013), "Clustering Uncertain Data Based on Probability Distribution Similarity" IEEE.

[18] Bin Wang, Gang Xiao, Hao Yu and Xiaochun Yang (2011), "Distance-Based Outlier Detection on Uncertain Data", IEEE Eleventh International Conference on Computer and Information Technology.

[19] *Burdick D., Deshpande P., Jayram T., Ramakrishnan R. and Vaithyanathan S. (2005), "OLAP Over Uncertain and Imprecise Data," in VLDB Conference Proceedings.*

[20] *Samir N. Ajani, Prof. Mangesh Wanjari (2013) , "An Approach for clustering uncertain data objects: A Survey", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Vol. 2, Issue 6.*

[21] *Charu C. Aggarwal (2009), "On High-Dimensional Projected Clustering of Uncertain Data Streams", Data Engineering, 2009. ICDE'09.*

[22] *Cheng R., Kalashnikov D. and Prabhakar S (2004), "Querying Imprecise Data in Moving Object Environments," IEEE Transactions on Knowledge and Data Engineering, Vol. 16,No. 9, pp. 1112-1127.*

[23] *C. Jebari and H. Ounelli, "Genre categorization of web pages." in ICDMW 2007, pp. 455464.*

[24] *Cheng R., Chau M., Garofalakis M. and Yu J. X. (2010), "Guest editors' introduction: Special section on mining large uncertain and probabilistic databases," IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 9, pp. 1201.*

[25] *Cormode G. and McGregor A. (2008), "Approximation Algorithms for Clustering Uncertain Data", Proc. Symp. Principles of Database Systems (PODS), M. Lenzerini and D. Lembo, eds., pp. 191-200.*

[26] *Das Sarma A., Benjelloun O., Halevy A. and Widom J. (2006), "Working Models for Uncertain Data," in ICDE Conference Proceedings(ICDE).*

[27] *Aggarwal C. and Yu P. S. (2009), "A survey of uncertain data algorithms and applications", IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 5, pp. 609-623.*

[28] *Chau, M., Cheng, R., and Kao, B. (2005) , "Uncertain Data Mining: A New Research Direction." in Proceedings of the Workshop on the Sciences of the Artificial, Hualien, Taiwan, December 7-8th.*

[29] *Dhillon I.S., Mallela S. and Kumar R. (2003), "A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification", J. Machine Learning Research, Vol.3, pp.1265-1287.*

[30] *Dunham H. and Sridhar S. (2006), Introduction to Data mining, Pearson Education.*

[31] *Giordani P. and Kiers H. A. L. (2006), "A comparison of three methods for principal component analysis of fuzzy interval data," Computational Statistics and Data Analysis, Vol. 51, No. 1, pp. 379-397.*

[32] *Hae-Sang Park & Chi-Hyuck Jun (2008), "A simple and fast algorithm for K-medoids clustering", Elsevier.*

[33] *Hans-Peter Kriegal and Martin Pfeifle (2005), "Density-based clustering of uncertain data", Proceeding of the eleventh ACM SIGKDD International conference on Knowledge discovery in data mining , pp.672 .*

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

[35] *Kao B., Lee S. D., Lee F. K. F., Cheung D. W. L., & Ho W. S. (2010), "Clustering uncertain data using Voronoi diagrams and R-tree index", IEEE TKDE, Vol. 22, No. 9, pp. 1219-1233.*

[36] *Hans-Peter Kriegal and Martin Pfeifle (2005), "Hierarchical Density-Based Clustering of Uncertain Data," Proc. IEEE Int'l Conf. Data Mining (ICDM).*

[37] *Le Li Zhiwen, Yul Zijian and Fengl Xiaohang Zhangl (2011), "Automatic Classification of Uncertain Data by Soft Classifier", Proceedings of the 2011 International Conference on machine Learning and Cybernetics, GuiJin.*

[38] *Lee S.D., Kao B. and Cheng R. (2007), "Reducing UK-Means to KMeans", Proc. First Workshop Data Mining of Uncertain Data (DUNE), in Conjunction with the Seventh IEEE Int'l Conf. Data Mining (ICDM).*

[39] *Lurong Xiao and Edward Hung (2007), "An Efficient Distance Calculation Method for Uncertain Objects", Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007).*

[40] *Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu (1996) " A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" , International Conference on Knowledge Discovery and Data Mining.*

[41] *Ngai W. K., Kao B., Cheng R., Chau M., Lee S. D., Cheung D. W. and Yip K. Y. (2011), "Metric and trigonometric pruning for clustering of uncertain data in 2D geometric space", Information Systems, Vol. 36, No. 2, pp. 476–497.*

[42] *Miss Pragati Pandey, Miss Prateeksha Pandey and Mrs. Minu Choudhary (2012), "Uncertain Data Algorithms and Applications", International Journal of Advanced Research in Computer Science and Software Engineering , Vol.2 , Issue 7,pp:274-280.*

[43] *Reynold Cheng, Xike Xie, Man Lung Yiu, Jinchuan Chen and Liwen Sun (2010), " UV-Diagram: A Voronoi Diagram for Uncertain Data", ICDE Conference IEEE.*

[44] *Samir Anjani H and Mangesh Wanjari (2013), "Clustering of uncertain data object using improved K-Means algorithm", IJARCSSE.*

[45] *Sujatha S. and Shanthi Sona A. (2013), " New Fast K-Means Clustering Algorithm using Modified Centroid Selection Method", IJERT.*

[46] *Velmurugan T. (2012), "Efficiency of K-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points", IJCTA.*

[47] *Volk Habich, Clemens Utzny, Ralf Dittmann and Wolfgang Lehner (2007), "Error-Aware Density-Based Clustering of Imprecise Measurement Values", Seventh IEEE International Conference on Data Mining Workshops, ICDM Workshops, IEEE.*

[48] *Volk P.B., Rosenthal F., Hahmann M., Habich D. and Lehner W. (2009), "Clustering Uncertain Data with Possible Worlds", Proc. IEEE Int'l Conf. Data Eng. (ICDE).*

[49] *Xu J. and Croft W.B. (1999), "Cluster-Based Language Models for Distributed Retrieval," Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR).*

[50] *Zufle A., Emrich T., Schmid K. A., Mamoulis N., Zimek A. and Renz M. (2014), " Representative clustering of uncertain data", in Proc. KDD, pp: 243–252.*

[51] *S.Thirunavukkarasu, Dr.K.P.Kaliyamurthie, "Clustering Uncertain Data using Voronoi Diagrams and R-Tree Index", International Journal of Innovative Research in Computer and Communication Engineering, Vol.2,issue.3,pp.3520-3525,2014.*

[52] *Jiangtao Ren, Sati Dan Lee, Xianlu Chen. Ben Kao, Reynold Cheng and David Cheung, "Naive Bayes Classification of Uncertain Data, "Ninth IEEE International Conference on Data Mining,pp.944-949,2009.*

[53] *Jiaqi Ge, and Yuni Xia, "UNN: A Neural Network for uncertain data classification", In Proceedings of PAKDD, vol.1, pp.449-460, 2010.*

[54] *Y.Endo , Yasunori Endo , Yasushi Hasegawa , Yukihiro, Hamasuna and Yuchi Kanzawa , " Fuzzy c-Means Clustering for Uncertain Data Using Quadratic Penalty-Vector Regularization ", In Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.15 , pp.76-82, No.1, 2011.*

[55] *Thuraisingham B., "A Primer for Understanding and Applying Data Mining", IT Professional, pp: 28-31, 2000.*

[56] *J. P. Bigus, ˜Data Mining with Neural Networks", McGraw-Hill, 1996.*

[57] *Usama M.Fayyad , " Data Mining and Knowledge Discovery in Databases: Implications from scientific databases", in Proc. of the 9th Int. Conf. on Scientific and Statistical Database Management, Olympia, Washington, USA. pp: 2-11. 1997.*

[58] *K. Yacoben and L. Carmichael, "Applying the Knowledge Discovery in Databases (KDD) Process to Fermilab Accelerator Machine Data", Fermi National Accelerator Laboratory, 1997.*

[59] *Frawley, W., Piatetsky-Shapiro, G., and Matheus, C. "Knowledge Discovery in Databases: An Overview". AI Magazine, pp: 57-70, 1992.*

[60] *Osmar R. Zaiane. "Chapter I: Introduction to Data Mining", CMPUT690 Principles of Knowledge Discovery in Databases, 1999.*

[61] *S. P. Lloyd. "Lease square quantization in pcm", IEEE Transactions on Information Theory, Vol: 28, No: 2, pp. 129-136, 1982.*

[62] *J. B. MacQueen, "Some Method for Classification and Analysis of Multivariate Observations", Proc. of Berkeley Symp. on Mathematical Statistics and Prob., Berkeley, U. of California Press, Vol: 1, pp: 281-297. 1967.*

[63] *M. Ester, H. Kriegel, J. Sander, and X. Xu, "A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", Proc. of KDD, 1996.*

[64] *J. Shi, and J. Malik, "Normalized Cuts and Image Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol: 22, No: 8, 2000.*

[65] *I. K. Ravichandra Rao, "Data Mining and Clustering Techniques". DRTC Workshop on Semantic Web. Bangalore, 8th – 10th December, 2003.*

[66] *Aynur Dayanik, Craig G. Nevill-Manning, "Clustering in Relational Biological Data", ICML-2004 Workshop on Statistical Relational Learning and Connections to Other Fields, pp: 42-47, 2004.*

[67] *Pham, D.T. and Afify, A.A. "Clustering techniques and their applications in engineering", Proceedings-Institution of Mechanical Engineers Part C Journal of Mechanical Engineering Science, Vol: 221; No: 11, pp: 1445-1460, 2007.*

[68] *M. Steinbach, G. Karypis, V. Kumar, "A Comparison of Document Clustering Techniques", Text Mining Workshop, KDD, 2000.*

[69] *Y. Zhao and G. Karypis, "Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering", Machine Learning, Vol: 55, No: 3, pp. 311-331, 2004.*

[70] *Hesam Izakian, Ajith Abraham, Vaclav Snasel, "Fuzzy Clustering Using Hybrid Fuzzy c-means and Fuzzy Particle Swarm Optimization", World Congress on Nature and Biologically Inspired Computing (NaBIC 2009), India, IEEE Press, pp. 1690-1694, 2009.*

[71] *Swagatam Das, Ajith Abraham, Amit Konar, "Automatic Clustering Using an Improved Differential Evolution Algorithm", IEEE Transactions on Systems, Man, and Cybernetics—Part. A: Systems And Humans, Vol. 38, No. 1, 2008.*

[72] *Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From Data Mining to Knowledge Discovery in Databases", AI Magazine, Vol. 17, pp. 37-54, 1996.*

[73] *Fayyad U.. Piatetsky-Shapiro G., Smyth P., "From data mining to knowledge discovery: an overview, Advances in knowledge discovery and data mining", American Association for Artificial Intelligence, Menlo Park, CA, AAAI/MIT Press, 1996, pp: 1-36.*

[74] *Romero C., Ventura S., Espejo P.G., and Hervas C., "Data Mining Algorithms to Classify Students", proceedings of the 1st Int'l conference on educational data mining. Canada. 2008, pp: 8-17.*

[75] *Zhang J., Mani I., "kNN Approach to Unbalanced Data Distributions: A Case Study involving Information Extraction", In Proceedings of The Twentieth International Conference on Machine Learning (ICML-2003), Workshop on Learning from Imbalanced Data Sets II, August 21, 2003.*

[76] *Charu C. Aggarwal and Philip S. Yu, "A Survey of Uncertain Data Algorithms and Applications", IEEE transactions on knowledge and data engineering. Vol. 21, No. 5 MAY 2009.*

[77] *Barbara, D., Garcia-Molina, H. and Porter D., "The Management of Probabilistic Data," IEEE Transactions on Knowledge and Data Engineering,vol.5, issue.5, pp:487-502,1992.*

[78] *Cheng, R.. Kalashnikov, D., and Prabhakar S., "Evaluating Probabilistic Queries over Imprecise Data," Proceedings of the ACM SIGMOD International Conference on Management of Data, June 2003.*

[79] *S. Guha, R. Rastogi, and K. Shim (1998), "CURE: An Efficient Clustering Algorithm for Large Databases", in ACM SIGMOD Conference.*

[80] *S. Guha, R. Rastogi, and K. Shim(2000), " ROCK: A Robust Clustering Algorithm for Categorical Attributes", in Information Systems, vol.25 , issue.5, pp: 345-366.*

[81] *A. Jain, and R. Dubes, "Algorithms for Clustering Data", Prentice Hall, New Jersey, 1998.*

[82] *L. Kaufman, P. Rousseeuw.,"Finding Groups in Data: An Introduction to Cluster Analysis", Wiley Interscience, 1990.*

[83] *J. R. Quinlan, C4. 5: Programs for Machine Learning. Morgan Kaufman Publishers, 1993.*

[84] *William. W. Cohen, "Fast effective rule induction," in Proc. of the 12th Intl. Conf on Machine Learning, 1995, pp. 115-123.*

[85] *Shehroz S Khan and Dr. Shri Kant, " Computation of Initial Modes for K-modes Clustering Algorithm using Evidence Accumulation," in Proceedings of the 20th international joint conference on Artifical intelligence(IJCAI), pp. 2784-2789, 2007.*

[86] *P. Langley, Wayne. lba, and K. Thompson, "An analysis of bayesian classifiers," in National Conf. on .4rtigicial Intelligence, 1992, pp. 223-228.*

[87] *R. Andrews, J. Diederich, and A. Tickle, "A survey and critique of techniques for extracting rules from trained artificial neural networks," Knowledge Based Systems, vol. 8, no. 6, pp. 373— 389, 1995.*

[88] *Lamis Hawarah, Ana Simonet, and Michael Simonet, "Dealing with missing values in a probabilistic decision tree during classification," in The Second International Workshop on Mining Complex Data, 2006, pp. 325-329.*

[89] *B. Kao, R. Cheng, M. Chau, S. D. Lee, D. W. Cheung and K. Y. Yip (2015), "Metric and trigonometric pruning for clustering of uncertain data in 2D geometric space", International Journal of Database Theory and Application Vol.8, No.2 , pp. 267-274.*

[90] D.N. Campo, G. Stegmayera, D.H. Milonea (2016) ," A new index for clustering validation with overlapped clusters" , Expert Systems with Applications , Vol.64, pp:549-556.

[91] Guru, D. S. and Nagendraswam, H. S. (2006), "Clustering of interval-valued symbolic patterns based on mutual similarity value and the concept of -mutual nearest neighbourhood", ACCV (2), pp. 234–243.

[92] Ajit Patil and M.D. Ingle (2015) ,"Clustering on Uncertain Data using Kullback Leibler Divergence Measurement based on Probability Distribution", International Journal of Current Engineering and Technology, Vol.5, No.4 , pp. 2614-2620.

[93] Kulkarni V.V and Bag V.V (2014), "Clustering Multi-Attribute Uncertain Data using Probability Distribution", International Journal of Computer Applications, Vol.102, No.5, pp:28-32.

[94] Manisha Padole , Prof. Sonali Bodkhe (2016), "An Efficient Methodology for Clustering Uncertain Data Based on Similarity Measure", IOSR Journal of Computer Engineering (IOSR-JCE), Vol.18, Issue 4, pp:12-16.

[95] Mrs. S. Vydehi and Dr. M. Punithavalli (2016), "Clustering Uncertain Gene Dataset using KLSE (Kullback-Leibler & Shennon Entropy) to improve Cluster Quality", International Journal of Applied Engineering Research, Vol.11, Number 4. pp:2693-2696.

[96] C. Y. Liu (2013) ,"Cost-Sensitive Clustering for Uncertain Data Based on Genetic Algorithm", International Journal of Applied Mathematics and Statistics ,Vol.40, Issue 10.

[97] Francesco Gullo, Giovanni Ponti, Andrea Tagarelli (2013), "Minimizing the variance of cluster mixture models for clustering uncertain objects", Statistical Analysis and Data Mining , Vol.6 , Issue 2, pp:116-135.

[98] Dan Olteanu and Sebastiaan J. van Schaik (2012), "DAGger: Clustering Correlated Uncertain Data", Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining.

[99] Grigorios Tzortzis and Aristidis Likas (2014), "The MinMax k-Means clustering algorithm", Pattern Recognition , pp: 2505–2516.

[100] Bhumika Ingale, Priyanka Fulare (2014),"Review of Algorithms for Clustering Random Data", International Journal of Computer Science and Information Technologies, Vol. 5 , Issue 3, pp:4444-4445.

[101] Raymond T. Ng and Jiawei Han (2002) "CLARANS: A Method for Clustering Objects for Spatial Data Mining" IEEE Trans on knowledge and data engineering, Vol. 14, No.5, pp.1003-1015.

[102] Y Zhao, C Aggarwal and P Yu (2010), "On Wavelet Decomposition of Uncertain Time Series Data Sets", Proceedings of the 19th ACM international conference, pp:129-138.

[103] Ping Jin , Shichao Qu , Yu Zong and Xin Li (2014),"CUDAP: A Novel Clustering Algorithm for Uncertain Data Based on Approximate Backbone", JOURNAL OF SOFTWARE, VOL. 9, NO. 3 , pp:732 - 737.

[104] Yuni Xia and Bowei Xi (2007), "Conceptual Clustering Categorical Data with Uncertainty", 19th IEEE International Conference on Tools with Artificial Intelligence , pp:329 -336.

[105] J. Han and M. Kamber. Data Mining Concepts And Techniques. Morgan Kaufmann, San Francisco, 2ed edition, 2006.

[106] Tian Zhang, Raghu Ramakrishnan and Miron Livny (1996), " BIRCH: An Efficient Data Clustering Method for Very Large Databases", ACM Sigmod Record, pp:103 -114.

[107] B. S. Duran and P. L. Odell.(1974)," Cluster Analysis: a Survey.", Springer-Verlag, NewYork.

[108] E. Diday and J. C. Simon. (1976),"Clustering anaslysis.", In K. S. Fu, editor, Digital Pattern Recognition, pages 47–94. Springer-Verlag, Secaucus, NJ.

[109] R. Michalski, R. E. Stepp, and E. Diday(1981)" A recent advance in data analysis: clustering objects into classes characterized by conjunctive concepts.", In L. Kanal and A. Rosenfeld, editors, Progress in Partial Recognition, volume 1, pages 33–56.North-Holland Publishing Co., The Netherlands.

[110] A. K. Jain, M. N. Murty, and P. J. Flynn.(1999)," Data clustering: a review.", ACM Computing Surveys, 31(3):264–323.