International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064

An Efficient Divergence and Distribution Based Similarity Measure for Clustering Of Uncertain Data

Geetha[1], Mary Shyla[2]

[1]Research Scholar, Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Bharathiar University, Coimbatore, Tamilnadu, India

[2]Assistant Professor, Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Bharathiar University, Coimbatore, Tamilnadu, India

Abstract: ***Data Mining is the extraction of hidden predictive information from large databases. Clustering is one of the popular data mining techniques. Clustering on uncertain data, one of the essential tasks in mining uncertain data, posts significant challenges on both modeling similarity between uncertain objects and developing efficient computational methods. The previous methods extend traditional partitioning clustering methods. Such methods cannot handle uncertain objects that are geometrically indistinguishable, such as products with the same mean but very different variances in customer ratings. Surprisingly, probability distributions, which are essential characteristics of uncertain objects, have not been considered in measuring similarity between uncertain objects. In Existing method to use the well-known Kullback-Leibler divergence to measure similarity between uncertain objects in both the continuous and discrete cases, and integrate it into partitioning and density-based clustering methods to cluster uncertain objects. It is very costly or even infeasible. The proposed work introduces the well-known Kernel skew divergence to measure similarity between uncertain objects in both the continuous and discrete cases. Measuring the cluster similarity with Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time and/or space and to further speed up the computation.***

Keywords: Clustering, uncertain data, Kernel skew Divergence and distribution

1. Introduction

Data mining is the process of extracting or mining knowledge from large amount of data. Data mining tools and techniques helps to predict business trends those can occur in near future such as Clustering, Classification, Association rule, Decision trees. As an important research direction in the field of data mining, clustering has drawn more and more attention to researchers in the data mining. Clustering is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. Clustering on uncertain data, one of the essential tasks in mining uncertain data, posts significant challenges on both modeling similarity between uncertain objects and developing efficient computational methods. It used to place data elements into related groups without advance knowledge of the group definitions..

1.1 Data Mining

Data mining is concerned with the nontrivial extraction of implicit, previously unknown, and potentially useful information from data (Frawley, Piatesky-Shapiro, and Matheus 1991). It is one of the steps in the process of knowledge discovery in databases(KDD) (Fayyad 1996; Fayyed, Piatetsky-Shapiro, and Smyth 1996a, 1996b, 1996c). And for this reason, data mining has been used interchangeably with KDD by many database researchers (Agrawal et al.1996: Han et al: 1996; Imielinski and Virmani 1995; Silberschatz, Stonebraker, and Ullman 1991,1996).

1.1.1 Foundation of Data Mining

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computer
- Data mining algorithms

The core components of data mining technology have been under development for decades, in research areas such as statistics, artificial intelligence, and machine learning. Today, the maturity of these techniques, coupled with high-performance relational database engines and broad integration efforts, make these technologies practical for current data warehouse environments.

1.1.2 Challenges in Data Mining

In four annual surveys of data miners, data mining practitioners consistently identify three key challenges that they face more than any others, specifically.

- Dirty data

- Explaining data mining to others, and
- Unavailability of data/ difficult access to data

1.2 Cluster Analysis and Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results.

Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties. The notion of a "cluster" cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms. There of course is a common denominator: a group of data objects. However, different researchers employ different cluster models, and for each of these cluster models again different algorithms can be given. The notion of a cluster, as found by different algorithms, varies significantly in its properties.

1.3 Problem Definition

Clustering on uncertain data, one of the essential tasks in mining uncertain data, posts significant challenges on both modeling similarity between uncertain objects and developing efficient computational methods. The previous methods extend traditional partitioning clustering methods. The Kullback-Leibler divergence to measure similarity between uncertain objects in both the continuous and discrete cases, and integrate it into partitioning and density-based clustering methods to cluster uncertain objects. Nevertheless, the implementation is very costly. Particularly, computing exact KL divergence in the continuous case is very costly or even infeasible. To tackle this problem, to estimate KL divergence in the continuous case by kernel density estimation and employ the fast Gauss transform technique. In this technique only measuring the similarity between the uncertain data not consider the similarity between the cluster it becomes the major problem in existing system, to overcome these problem we proposed a distribution similarity with Poisson distribution function between clusters and also improve the clustering accuracy of uncertain data object similarity by changing KL Divergence into Kernel skew divergence.

1.4 Objective of the Research

The main contribution of the research is to introduce a new Kernel Skew divergence based similarity measure used to measure the similarity between the uncertain data object that improves the clustering result. After measuring the similarity between the data object then find the probability result with Poisson distribution measure the uncertain data based on distribution similarity of the one clustered data to another clustered data.Finally cluster the uncertain data object with Distance-based clustering methods.

1.5 Contribution of the Research

The main contribution of the research is to introduce a new kernel skew divergence based similarity function that improves the clustering uncertain data object result .The key concept of the system is also measuring the similarity of the clustered data with Poisson distribution .It allow to measure the uncertain data based on distribution similarity of the one clustered data to another clustered data after the completion of the uncertain data object based similarity measure .It improves the clustering accuracy than the existing system .

2. Related Work

[Mihael Ankerst *et al.,* 1999] proposed to the density-based clustering's corresponding to a broad range of parameter settings. It is a versatile basis for both automatic and interactive cluster analysis. We show how to automatically and efficiently extract not only 'traditional' clustering information (e.g. representative points, arbitrary shaped clusters), but also the intrinsic clustering structure. For medium sized data sets, the cluster-ordering can be represented graphically and for very large data sets, we introduce an appropriate visualization technique. Both are suitable for interactive exploration of the intrinsic clustering structure offering additional insights into the distribution and correlation of the data.

[Hans-Peter Kriegel, Martin Pfeifle., 2005] The hierarchical density-based clustering algorithm OPTICS 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. In this paper, we propose 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, we 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. FOPTICS algorithm basically works like the OPTICS

algorithm. It always takes the first element from the seed list, add it to the result set, and carry out a range query.

[Wang Kay Ngai et al., 2006] proposed described the basic min-max-dist pruning method and showed that it was fairly effective in pruning expected distance computations. To further improve performance, we derived four bound-estimation methods. We conducted extensive experimental study evaluating those four pruning methods. Our results showed that Ucs and Lcs are very effective, especially when they work together. In some experiment setting, UcsLcs was a dozen times more effective than basic min-max-dist in terms of pruning effectiveness. Method UpreLpre , which is based on precomputation of anchor points' expected distances, also performed very well. The pre-computation overheads, however, made UpreLpre second-best to Ucs and Lcs . The four pruning methods are independent of each other and can be combined to achieve an even higher pruning effectiveness. Pruning is at its full-strength when all four are applied and if the pre-computation overhead could be discounted. A factor of 24 times more effective in pruning than min-max-dist was registered in some of the experiments.

[Marcel R. Ackermann *et al.*, 2008] proposed to understanding clustering problems with non-metric dissimilarity measures, like the Kullback-Leibler divergence. We consider a problem that is relatively well understood in the case of Euclidean and metric distances: k-median clustering. In k-median clustering we have a representative (sometimes called prototype) for each cluster. In the geometric version of the problem this is the cluster center. We are interested in minimizing the sum of error of the clustering, i.e. the error that is made by representing each input object by its corresponding representative. Since we allow non-metric dissimilarity measures, this version of k-median also captures other variants like the well-known Euclidean k-means clustering, where the goal is to minimize the sum of squared errors (with respect to Euclidean distance).

[Guadalupe J. Torres et al., 2008] The similarity measure that we proposed has experimentally demonstrated consistently similar results to popular measures of Euclidian distance (between cluster centroids) and Pearson correlation. The techniques can be extended to various real-world problems such as classification and clustering of malware, email analysis (finding social graph among the users based on email contents, for instance) in digital forensics. Since unsupervised clustering algorithms do not giveaccuracy; the proposed algorithm can be applied to find the best clustering algorithm for many real-life applications where clustering techniques are applied. The approach should enable users to experimentally compare various clustering algorithms and choose the one that best serves the problem.

[Thierry Denoeux, 2013] The proposed formalism combines aleatory uncertainty captured by a parametric statistical model with epistemic uncertainty induced by an imperfect observation process and represented by belief functions. Our method then seeks the value of the unknown parameter that maximizes a generalized likelihood criterion, which can be interpreted as a degree of agreement between the parametric model and the uncertain data. This is achieved using the evidential EM algorithm, which is a simple extension of the classical EM algorithm with proved convergence properties.

3. Methodology

3.1 Existing Method

Uncertain objects as random variables with certain distributions. Consider both the discrete case and the continuous case. In the discrete case, the domain has a finite number of values. In the continuous case, the domain is a continuous range of values, for example, the temperatures recorded in a weather station are continuous real numbers. Directly computing KL divergence between probability distributions can be very costly or even infeasible if the distributions are complex. Although KL divergence is meaningful, a significant challenge of clustering using KL divergence is how to evaluate KL divergence efficiently on many uncertain objects. First to study clustering uncertain data objects using KL divergence in a general setting, to make several contributions. After that develop a general framework of clustering uncertain objects considering the distribution as the first class citizen in both discrete and continuous cases. Uncertain objects can have any discrete or continuous distribution.

It shows that distribution differences cannot be captured by the previous methods based on geometric distances. KL divergence to measure the similarity between distributions, and demonstrate the effectiveness of KL divergence in both partitioning and density-based clustering methods. To tackle the challenge of evaluating the KL divergence in the continuous case, estimate KL divergence by kernel density estimation and apply the fast Gauss transform to boost the computation.

3.1.1 Kullback-Leibler Divergence based similarity measure

It is natural to quantify the similarity between two uncertain objects by KL divergence. Given two uncertain objects P and Q and their corresponding probability distributions, D(P||Q) evaluates the relative uncertainty of Q given the distribution of P. which is the expected log-likelihood ratio of the two distributions and tells how similar they are. The KL divergence is always nonnegative, and satisfies Gibbs' inequality, Kullback–Leibler divergence is a non-symmetric measure of the difference between two probability distributions P and Q. Specifically, the Kullback–Leibler divergence of Q from P, denoted $D_{KL}(P||Q)$, is a measure of the information lost when Q is used to approximate P: KL measures the expected number of extra bits required to code samples from P when using a code based on Q, rather than using a code based on P. Typically P represents the "true" distribution of data, observations, or a precisely calculated theoretical distribution. The measure Q typically represents a theory, model, description, or approximation of P.

$$D(P||Q) = E[\log P/Q]$$

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

In the continuous case, given the samples of P and Q, by the law of large numbers we have,

$$\left[\lim_{s\to\infty} 1/S\right]\left[\left(\sum_i (i=1)^{1}s \equiv \left[\log P(p_i)/Q(p_i)\right]\right)^{1}1 = D(P\|Q)\right]$$
$$\left[1/S\left(\sum_i (i=1)^{1}s \equiv \left[\log P(p_i)/Q(p_i)\right]\right)\right]^{1}1 = D(P\|Q)$$

3.1.2 Clustering methods

Geometric distance-based clustering methods for uncertain data mainly fall into two categories, partitioning and density-based approaches present the clustering methods using KL divergence to cluster uncertain objects in these two categories. In present the uncertain k-medoids method which extends a popular partitioning clustering method k-medoids [19] by using KL divergence. Develop a randomized k-medoids method based on the uncertain k-medoids method to reduce the time complexity. Presents the uncertain DBSCAN method which integrates KL divergence into the framework of a typical density-based clustering method DBSCAN the algorithms of the methods and how they use KL divergence as the similarity measure.

3.1.3 K-medoids used in research

K-means method represents each cluster by the mean of all objects in this cluster, while the k-medoids method uses an actual object in a cluster as its representative. In the context of uncertain data where objects are probability distributions, it is inefficient to compute the mean of probability density functions. K-medoids method avoids computing the means. For the sake of efficiency, we adopt the k-medoids method to demonstrate the performance of partitioning clustering methods using KL divergence to cluster uncertain objects. The uncertain k-medoids method consists of two phases, the building phase and the swapping phase.

3.1.4 The algorithm can be performed in two ways

In the building phase, the uncertain k-medoids method obtains an initial clustering by selecting k representatives one after another. The first representative C1 is the one which has the smallest sum of the KL divergence to all other objects in O.

$$C_1 1 = \arg\min_{\tau}(P \in O\{p\}) \sum_i (P^{1r} \in O\{p\}) \left[D(P^{1r}\|\right] P)$$

where P is the probability distribution , P' is the remaning data object that not belong to the objects in the P. $D(P^{1r}\|P)$ is smaller than the divergence between P' and any previously selected representatives P.

In the swapping phase, the uncertain k-medoids method iteratively improves the clustering by swapping a nonrepresentative object with the representation to which it is assigned. For a nonrepresentative object P, suppose it is assigned to cluster CC whose representative is C. Reassignment happens, the decrease of the total KL divergence by swapping P and C is recorded. After all nonrepresentative objects are examined; we obtain the total decrease of swapping P and C. Then, we select the object P_{max} which can make the largest decrease.

$$\left[P_i max = \arg\right]\min_{\tau}(P \in O\{c_i(1, c_i 2), \ldots c_i k\})\left[(DEC(\right] P))$$

Maximum number of the object in the dataset are selected based on the data objects that belongs to the decrease the probability (P) is denoted as DEC(P).

3.1.5 Partition based clustering

DBSCAN requires two parameters: ε(eps) and the minimum number of points required to form a cluster (minPts). It starts with an arbitrary starting point that has not been visited. This point's ε-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. Note that this point might later be found in a sufficiently sized ε-environment of a different point and hence be made part of a cluster. If a point is found to be a dense part of a cluster, its ε-neighborhood is also part of that cluster. Hence, all points that are found within the ε-neighborhood are added, as is their own ε-neighborhood when they are also dense. This process continues until the density-connected cluster is completely found. Then, a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.

3.1.6 DBSCAN in research

Unlike partitioning methods which organize similar objects into the same partitions to discover clusters, density-based clustering methods regard clusters as dense regions of objects that are separated by regions of low density. DBSCAN is the first and most representative density-based clustering method developed for certain data. To demonstrate density-based clustering methods based on distribution similarity, we develop the uncertain DBSCAN method which integrates KL divergence into DBSCAN.

DBSCAN method transforms objects into a different space where the distribution differences are revealed. The uncertain DBSCAN method finds dense regions through core objects whose "^{e} neighborhood contains at least $^{\mu}$ objects. Formally, P is a core object, if

$$|\{Q \in O|D(Q \mid H|P) \leq \varepsilon\}| \geq \mu$$

The uncertain k-medoids method, the randomized k-medoids method, and the uncertain DBSCAN method all require evaluation of KL divergences of many pairs of objects. As the number of uncertain objects and the sample size of each object increase, it is costly to evaluate a large amount of KL divergence expressions. Continuous case, the complexity of calculating the probability density functions is quadratic to the sample size of the uncertain object. Precomputation is not feasible since the domain is unaccountably infinite. No matter we evaluate KL divergences directly or evaluate the divergence differences the major cost is on evaluating the following expression:

$$\sum_{p\in P} log \sum_{q\in Q} \prod_{j=1}^{d} e^{\frac{-(p.D_j - q.D_j)^2}{2h_j^2}}$$

Improved fast Gauss transform to reduce the constant factor to asymptotically polynomial order. They adopt their improved Gauss transform to boost the efficiency of evaluating.

Step 1: Approximate the sample of C by clustering. Partition the N sample points of C into k clusters $S_1; \ldots ; S_k$ using the farthest point clustering algorithm.

Step 2: Choose parameter p for truncating Taylor expansion. Choose p sufficiently large such that the error estimate is less than the precision P. Here, the error at any point of P is bounded

$$N\left(\frac{2^p}{p!}\rho_1\rho_2 + e^{-p^2}\right)$$

Step 3: Compute the coefficients of the Taylor expansion. For each cluster S_i ($1 \leq i \leq k$), let c_i denote the center found by the farthest point clustering algorithm, compute the coefficients below

$$C_\alpha^i = \frac{2^{|\alpha|}}{\alpha!}\sum_{q\in S_i}\prod_{j=1}^{d} e^{\frac{-(p.D_j - q.D_j)^2}{2h^2 c_j}}\left(\frac{q-c_i}{h}\right)^\alpha$$

Step 4: Compute the sum of approximated Gaussian functions. For each point $s \in P$, find the clusters whose centers lie within the range $h\rho$. Then, the sum of Gaussian functions is evaluated

as, $\sum_{q\in C}\prod_{j=1}^{d} e^{\frac{-(pD_j - qD_j)^2}{2h^2 c_j}} = \frac{2^{|\alpha|}}{\alpha!}\sum_{q\in S_i}\prod_{j=1}^{d} e^{\frac{-(pD_j - qD_j)^2}{2h^2 c_j}}\left(\frac{q-c_i}{h}\right)^\alpha$

From this entire algorithm get the clustering results.

3.2 Proposed Method

KL divergence is a standard and well motivated distributional distance measure and α's role appears to be simply to guarantee that the skew divergence is always defined, two natural questions arise. First, does increasing α, thereby bringing the skew divergence closer to the KL divergence, always yield better results. The asymmetric skew divergence, on the other hand, simply smoothes one of the distributions by mixing it, to a degree determined by the parameter α, with the other distribution. Varying the value of α changed the performance of the skew divergence for both of our training sets. Performance curves are preserved across training sets, with the error rates rising and the minima shifting to the right for sparse. Again, the skew divergences as a family are also less affected by frequency filtering than the baseline, back-off**.**

3.2.1 Kernel Skew divergence and Poisson distribution

The highest value yielded the best performance and very small values resulted in the worst error rates, as one might expect; but the relationship between error rate and α for intermediate settings is not so simple. The skew divergence is essentially the relative entropy but with 'skewed' second argument. That is, the second argument σ is replaced by the convex combination $\alpha\rho+(1-\alpha)\sigma$, where α is a scalar ($0 < \alpha < 1$) which we call the skewing parameter to measure the uncertain data similarity . As one of its basic properties we will show that $S(\rho||\alpha\rho+(1-\alpha)\sigma)$ is no longer infinite but is bounded above by $-\log \alpha$, and we define the skew divergence as the skewed relative entropy divided by this factor $-\log \alpha$:

Skew divergence is defined as

$$\mathbf{SD_\alpha(\rho||\sigma) := \frac{1}{\log \alpha} S(\rho||\alpha\rho + (1-\alpha)\sigma).}$$

Improve the result of the similarity need some changes in the divergence based similarity function with skew, then add kernel entropy function that are most similar uncertain data objects are found in the dataset for clustering the data .It works by adding the k is a kernel if it can be interpreted as a scalar product on some other space. If we substitute $k(q, q')$ in the skew divergence function.

Now the skew divergence with kernel is changed as, the skew divergence as the skewed relative entropy divided by this factor $-\log \alpha$:

$$SD_\alpha(q||r) = \frac{1}{-\log \alpha} S(q||\alpha q + (1-\alpha)r)$$

$$KSD_\alpha(q||r) = \frac{1}{-\log \alpha} S(k(q,q')||\alpha k(q,q') + (1-\alpha)k(r,r'))$$

$$KS_{(q,q')} = \exp\left(\frac{-|q-q'|^2}{2\sigma^2}\right)$$

$$KS_{(r,r')} = \exp\left(\frac{-|r-r'|^2}{2\sigma^2}\right)$$

Steps in proposed system

Input : Uncertain data object O ,dataset D
Output : Best cluster uncertain distribution similarity
Step 1: Select uncertain data object O from the original dataset D
Step 2: Calculate the similarity that the skew divergence as the kernel skewed
relative entropy divided by this factor $-\log \alpha$,

$$KSD_\alpha(q||r) = \frac{1}{-\log \alpha} S(k(q,q')||\alpha k(q,q') + (1-\alpha)k(r,r'))$$

Step 3: Find best kernel skew divergence similarity results in the KSD
Step 4: After finding the Kernel skew divergence similarity function then performs the clustering using the clustering methods .
Step 5: Improve the clustering results find the distribution similarity using the poisson distribution. A discrete random variable X is said to have a poisson distribution with parameter $\lambda > 0$, if for $k = 0, 1, 2, ..$ the probability mass function of X is given by:

$$f(k,\lambda) = \Pr(X = K) = \frac{\lambda^K e^{-\lambda}}{k!}$$

Step 6: Finding the best distribution similarity improves the clustering results .

3.2.2 Poisson distribution

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume. The poisson distribution finds the uncertain data object similarity results from the kernel skew divergences result ,for improve the clustering result measure the distribution similarity that is the

similarity with highest probability values are considered as most similar data objects in the result. A discrete random variable X is said to have a poisson distribution with parameter $\lambda > 0$, if for $k = 0, 1, 2, \ldots$ the probability mass function of X is given by:

$$f(k,\lambda) = \Pr(X = K) = \frac{\lambda^K e^{-\lambda}}{k!}$$

Where e is the base of the natural algorithm ($e = 2.71828...$) and K! is the factorial of the k value that is the number of times the data object present in most similar in the Kernel Skew divergence based similarity results $\lambda = \lambda T$ when the number of events that is the number of times the similarity of uncertain data objects present in the dataset ,that is considered as important data for clustering the data at the interval that is difference one. The positive real number λ is equal to the expected number of the uncertain data objects results of x and also to its variance of the cluster distribution is defines as,

$$\lambda = E(X) = Var(X)$$

The Poisson distribution can be applied to systems with a large number of possible events in the uncertain data objects with each of which is rare in the data object .Based on these results finally cluster the data object using the clustering method in the existing work .The proposed system Clustering algorithm remains efficient in clustering large uncertain data objects and improves the accuracy of the system by measuring the similarity between the data objects and time complexity results are reduced based on the results.

4. Experimental Results

In this section measure the performance partitioning and density-based clustering methods have better clustering quality when using KL divergences as similarity than using geometric distances. Kernel Skew divergence (KSD) based similarity measure with distribution function. The results confirm that Kernel Skew divergence (KSD) based similarity measure with distribution function can naturally capture the distribution difference which geometric distance cannot capture. To boost the computation in the continuous case to battle the costly kernel density estimation, the fast Gauss transform can speed up the clustering a lot with an acceptable accuracy tradeoff. To scale on large data sets, the randomized k-medoids method equipped with the Gauss transform has linear complexity with respect to both the number of objects and the sample size per object. It can perform scalable clustering tasks on large data sets with moderate accuracy.

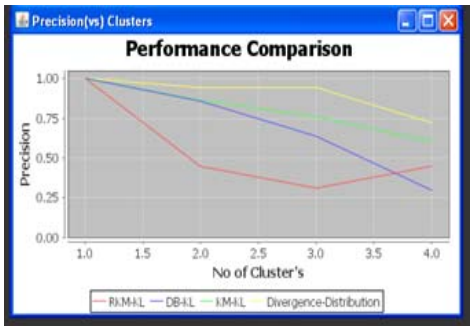

Figure 1: Precision comparison between no. of clustering

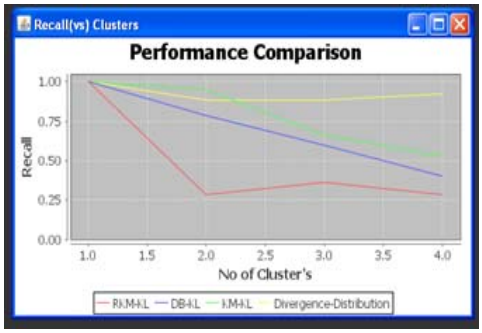

Figure 2: Recall comparison between no. of clustering

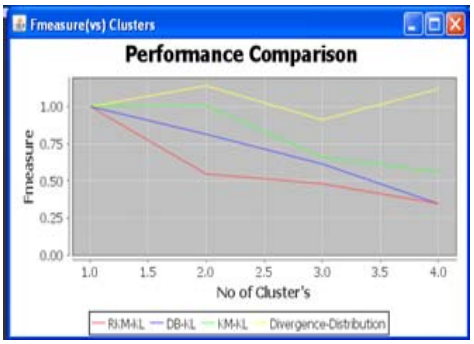

Figure 3: Fmeasure comparison between no. of clustering

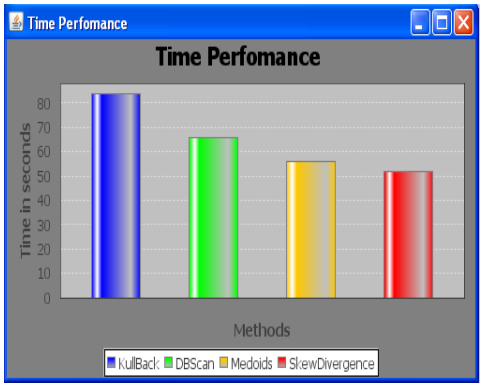

Figure 4: Time comparison between proposed and existing methods

5. Conclusion

In this research clustering uncertain data based on the similarity between their distributions. Using the kernel skew divergence as the similarity measurement between objects in both the continuous and discrete cases are measured. The integrated KSD divergence into the partitioning and density-based clustering methods to demonstrate the effectiveness of clustering results. To tackle the computational challenge in the continuous case, estimate KSD divergence by kernel density estimation and employ the Poisson distribution to find the probabilities of the distribution among data objects and then cluster the result further speed up the computation algorithm. Experimental results study the problems on uncertain data based on distribution similarity with clustering and it show better clustering results improves accuracy as well as time complexity is reduced.

6. Future Enhancement

Each and every distribution performs entirely different from each other, in future work can be applying different distribution uncertain data based on distribution similarity and it can be applied to any other data set with uncertain data.

References

[1] S. Abiteboul, P.C. Kanellakis, and G. Grahne, "On the Representation and Querying of Sets of Possible Worlds," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 1987.
[2] M.R. Ackermann, J. Blo¨mer, and C. Sohler, "Clustering for Metric and Non-Metric Distance Measures," Proc. Ann. ACM-SIAM Symp. Discrete Algorithms (SODA), 2008.
[3] M. Ankerst, M.M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: Ordering Points to Identify the Clustering Structure," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 1999.
[4] Banerjee, S. Merugu, I.S. Dhillon, and J. Ghosh, "Clustering with Bregman Divergences," J. Machine Learning Research, vol. 6, pp. 1705-1749, 2005.
[5] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," J. Machine Learning Research, vol. 3, pp. 993-1022, 2003.
[6] V. Cerny, "A Thermo dynamical Approach to the Travelling Salesman Problem: An Efficient Simulation Algorithm," J. Optimization Theory and Applications, vol. 45, pp. 41-51, 1985.
[7] R. Cheng, D.V. Kalashnikov, and S. Prabhakar, "Evaluating Probabilistic Queries over Imprecise Data," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2003.
[8] N.N. Dalvi and D. Suciu, "Management of Probabilistic Data: Foundations and Challenges," Proc. ACM SIGMOD-SIGACTSIGART Symp. Principles of Database Systems (PODS), 2007.
[9] I.S. Dhillon, S. Mallela, and R. Kumar, "A Divisive Information- Theoretic Feature Clustering Algorithm for Text Classification," J. Machine Learning Research, vol. 3, pp. 1265-1287, 2003.
[10] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," Proc. Second Int'l Conf. Knowledge Discovery and Data Mining (KDD), 1996.
[11] T. Feder and D.H. Greene, "Optimal Algorithms for Approximate Clustering," Proc. Ann. ACM Symp. Theory of Computing (STOC), 1988.
[12] T.F. Gonzalez, "Clustering to Minimize the Maximum Intercluster Distance," Theoretical Computer Science, vol. 38, pp. 293-306, 1985.
[13] L. Greengard and J. Strain, "The Fast Gauss Transform," SIAM J. Scientific and Statistical Computing, vol. 12, 1991.
[14] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Elsevier, 2000.
[15] T. Imielinski and W.L. Lipski Jr., "Incomplete Information in Relational Databases," J. ACM, vol. 31, 1984.
[16] A.K. Jain and R.C. Dubes, Algorithms for Clustering Data, Prentice Hall, 1988.
[17] R. Jampani, F. Xu, M. Wu, L.L. Perez, C.M. Jermaine, and P.J. Haas, "Mcdb: A Monte Carlo Approach to Managing Uncertain Data," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2008.
[18] Y. Tao, R. Cheng, X. Xiao, W.K. Ngai, B. Kao, and S. Prabhakar, "Indexing Multi-Dimensional Uncertain Data with Arbitrary Probability Density Functions," Proc. Int'l Conf. Very Large Data Bases (VLDB), 2005.
[19] P.B. Volk, F. Rosenthal, M. Hahmann, D. Habich, and W. Lehner, "Clustering Uncertain Data with Possible Worlds," Proc. IEEE Int'l Conf. Data Eng. (ICDE), 2009.
[20] J. Xu and W.B. Croft, "Cluster-Based Language Models for Distributed Retrieval," Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 1999.
[21] C. Yang, R. Duraiswami, N.A. Gumerov, and L.S. Davis, "Improved Fast Gauss Transform and Efficient Kernel Density Estimation," Proc. IEEE Int'l Conf. Computer Vision (ICCV), 2003.

Author Profile

Geetha Sarsavanan received the Bachelor's degree in Commerce with Computer Application from Periyar University in 2005. She received the Master's degree in Computer Application from Bharathiar University in 2010.

Mrs. E. Mary Shyla working as Assistant Professor and Placement Officer in Sri Ramakrishna College of Arts and Science for Women, Bharathiar University, Coimbatore, Tamilnadu. She has guided several PG and Research projects. She has presented her papers in International Conferences and has published papers in International Journals.