Scientific Journal Impact Factor (SJIF): 1.711
e-ISSN: 2349-9745
p-ISSN: 2393-8161
International Journal of Modern Trends in Engineering and Research
www.ijmter.com

Distribution Similarity based Data Partition and Nearest Neighbor Search on Uncertain Data

Mr. D. Parthipan[1], Dr. M. Moorthi [2]
Research Scholar, Kongu Arts and Science College, Erode, TN, India.
Ph.D, Assistant Professor, Kongu Arts and Science College, Erode, TN, India.

Abstract-Databases are build with the fixed number of fields and records. Uncertain database contains a different number of fields and records. Clustering techniques are used to group up the relevant records based on the similarity values. The similarity measures are designed to estimate the relationship between the transactions with fixed attributes. The uncertain data similarity is estimated using similarity measures with some modifications.

Clustering on uncertain data is one of the essential tasks in mining uncertain data. The existing methods extend traditional partitioning clustering methods like k-means and density-based clustering methods like DBSCAN to uncertain data. Such methods cannot handle uncertain objects. Probability distributions are essential characteristics of uncertain objects have not been considered in measuring similarity between uncertain objects.

The customer purchase transaction data is analyzed using uncertain data clustering scheme. The density based clustering mechanism is used for the uncertain data clustering process. This model produces results with minimum accuracy levels. The clustering technique is improved with distribution based similarity model for uncertain data. The nearest neighbor search technique is applied on the distribution based data environment. The system is designed using java as a front end and oracle as a back end.

I. Introduction

Clustering is the task of assigning a set of objects into groups so that the objects in the same cluster are more similar to each other than to those in other clusters. Clustering is a main task of explorative 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.

Uncertain data is the notion of data that contains specific uncertainty. Uncertain data is typically found in the area of sensor networks. When representing such data in a database, some indication of the probability of the various values. There are three main models of uncertain data in databases. In attribute uncertainty, each uncertain attribute in a tuple is subject to its own independent probability distribution. For example, if readings are taken of temperature and wind speed, each would be described by its own probability distribution, as knowing the reading for one measurement would not provide any information about the other.

The data clustering is carried out to partition the transaction sets. Transaction relationships are estimated for the clustering process. The similarity metrics are calculated to identify the transaction relationships. The transactions are assigned to relevant clusters based on the similarity value. Different types of similarity metric estimation mechanisms are used. The similarity metrics is also called as distance metrics.

II. Related Work

Clustering is a fundamental data mining task. Clustering certain data has been studied for years in data mining, machine learning, pattern recognition, bioinformatics, and some other fields. However, there is only preliminary research on clustering uncertain data. Data uncertainty brings new challenges to clustering, since clustering uncertain data demands a measurement of similarity between uncertain data objects. Ngai et al. [11] proposed the UK-means method which extends the k-means method. The UK-means method measures the distance between an uncertain object and the cluster center by the expected distance. Recently, Lee et al. [6] showed that the UK-means method can be reduced to the k-means method on certain data points.

Kriegel and Pfeifle [10] proposed the FDBSCAN algorithm which is a probabilistic extension of the deterministic DBSCAN algorithm for clustering certain data. As DBSCAN is extended to a hierarchical density-based clustering method referred to as OPTICS, Kriegel and Pfeifle [5] developed a probabilistic version of OPTICS called FOPTICS for clustering uncertain data objects. FOPTICS outputs a hierarchical order in which data objects, instead of the determined clustering membership for each object, are clustered.

Volk et al. followed the possible world semantics [8], [3] using Monte Carlo sampling [7]. This approach finds the clustering of a set of sampled possible worlds using existing clustering algorithms for certain data. Then, the final clustering is aggregated from those sample clusterings. The existing techniques on clustering uncertain data mainly focus on the geometric characteristics of objects, and do not take into account the probability distributions of objects. In this paper, we propose to use KL divergence as the similarity measure which can capture distribution difference between objects. To the best of our knowledge, this paper is the first work to study clustering uncertain objects using KL divergence.

We are aware that clustering distributions has appeared in the area of information retrieval when clustering documents. The major difference of our work is that we do not assume any knowledge on the types of distributions of uncertain objects. When clustering documents, each document is modeled as a multinomial distribution in the language model. For example, Xu and Croft discussed a k-means clustering method with KL divergence as the similarity measurement between multinomial distributions of documents. Assuming multinomial distributions, KL divergence can be computed using the number of occurrences of terms in documents. Blei et al. proposed a generative model approach—the Latent Dirichlet Allocation (LDA for short). LDA models each document and each topic as a multinomial distribution, where a document is generated by several topics. Dhillon et al. [9] used KL divergence to measure similarity between words to cluster words in documents in order to reduce the number of features in document classification. They developed a k-means like clustering algorithm and showed that the algorithm monotonically decreases the objective function as shown in [9], and minimizes the intracluster Jensen-Shannon divergence while maximizing intercluster Jensen-Shannon divergence. As their application is on text data, each word is a discrete random variable in the space of documents. Therefore, it is corresponding to the discrete case in our problem.

Banerjee et al. [4] theoretically analyzed the k-means like iterative relocation clustering algorithms based on Bregman divergences which is a general case of KL divergence. They summarized a generalized iterative relocation clustering framework for various similarity measures from the previous

work from an information theoretical viewpoint. They showed that finding the optimal clustering is equivalent to minimizing the loss function in Bregman information corresponding to the selected Bregman divergence used as the underlying similarity measure. In terms of efficiency, their algorithms have linear complexity in each iteration with respect to the number of objects. However, they did not provide methods for efficiently evaluating Bregman divergence nor calculating the mean of a set of distributions in a cluster. For uncertain objects in our problem which can have arbitrary discrete or continuous distributions, it is essential to solve the two problems in order to scale on large data sets, as we can see in our experiments.

Ackermann et al. [2] developed a probabilistic (1 + Є) approximation algorithm with linear time complexity for the k-medoids problem with respect to an arbitrary similarity measure, such as squared euclidean distance, KL divergence, Mahalanobis distance, etc., if the similarity measure allows the 1-medoid problem being approximated within a factor of (1 + Є) by solving it exactly on a random sample of constant size. They were motivated by the problem of compressing Java and C++ executable codes which are modeled based on a large number of probability distributions. They solved the problem by identifying a good set of representatives for these distributions to achieve compression which involves nonmetric similarity measures like KL divergence. The major contribution of their work is on developing a probabilistic approximation algorithm for the k-medoids problem.

The previous theoretical studies focused on the correctness of clustering using KL divergence [9], the correspondence of clustering using Bregman divergence in information theory [4] and the probabilistic approximation algorithm [2]. However, they did not provide methods for efficiently evaluating KL divergence in the clustering process, neither did they experimentally test the efficiency and scalability of their methods on large data sets. Different to them, our work aims at introducing distribution differences especially KL divergence as a similarity measure for clustering uncertain data. We integrate KL divergence into the framework of k-medoids and DBSCAN to demonstrate the performance of clustering uncertain data. More importantly, particular to the uncertain objects in our problem, we focus on efficient computation techniques for large data sets, and demonstrate the effectiveness, efficiency, and scalability of our methods on both synthetic and real data sets with thousands of objects, each of which has a sample of hundreds of observations.

III. Uncertain Data Clustering Scheme

Clustering uncertain data has been well recognized as an important issue. Generally, an uncertain data object can be represented by a probability distribution. The problem of clustering uncertain objects according to their probability distributions happens in many scenarios. Each camera may be scored by many users. Thus, the user satisfaction to a camera can be modeled as an uncertain object on the user score space. There are often a good number of cameras under a user study. A frequent analysis task is to cluster the digital cameras under study according to user satisfaction data. One challenge in this clustering task is that needed to consider the similarity between cameras not only in terms of their score values, but also their score distributions. One camera receiving high scores is different from one receiving low scores. At the same time, two cameras, though with the same mean score, are substantially different if their score variances are very different.

Similarity measurement between two probability distributions is not a new problem at all. In information theory, the similarity between two distributions can be measured by the Kullback-Leibler divergence. The distribution difference cannot be captured by geometric distances. The two objects have different

geometric locations [1]. Their probability density functions over the entire data space are different and the difference can be captured by KL divergence. In the geometric locations of the two objects are heavily overlapping, they have different distributions. The difference between their distributions can also be discovered by KL divergence, but cannot be captured by the existing methods.

The system considers uncertain objects as random variables with certain distributions. Both the discrete case and the continuous case. In the discrete case, the domain has a finite number of values, for example, the rating of a camera can only take a value in {1, 2, 3, 4, 5}. 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.

To the best of the knowledge, this system is the first to study clustering uncertain data objects using KL divergence in a general setting. A system builds 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. The distribution differences cannot be captured by the previous methods based on geometric distances.To tackle the challenge of evaluating the KL divergence in the continuous case, KL divergence is estimated by kernel density estimation and applies the fast Gauss transform to boost the computation.

IV. Issues on Uncertain Data Clustering

Clustering on uncertain data is one of the essential tasks in mining uncertain data. Partitional clustering is performed using the K-means clustering algorithm. The density based clustering is performed using DBSCAN technique. K-means and DBSCAN techniques are used for the certainty data environment. Probability distributions are used to measure similarity between uncertain objects. Kullback-Leibler divergence is used as the similarity measurement. KL divergence is integrated with the partitioning and density-based clustering methods. Distribution based similarity measures are integrated with K-means and DBSCAN techniques for uncertain data clustering process. The following drawbacks are identified in the existing system.

- Nearest neighbor search is not supported
- Clustering accuracy level is low
- Random centroid selection scheme
- Inter cluster distance factor is not considered

V. Data Partitioning and Nearest Neighbor Search on Uncertain Data

Clustering techniques are used to group up the customer data values. Probability distribution based similarity analysis is used for the uncertain data environment. Centroid optimization scheme is used to improve the clustering process. K-Nearest Neighbor search algorithm is used to fetch similar data values. The customer purchase transaction data is analyzed using uncertain data clustering scheme. Cluster centroid optimization scheme is used to improve the accuracy levels. K-means and DBSCAN clustering algorithms are integrated with the optimal centroid based analysis scheme. The nearest neighbor search technique is applied on the distribution based data environment. Fig 5.1 shows the cluster based nearest neighbor search.

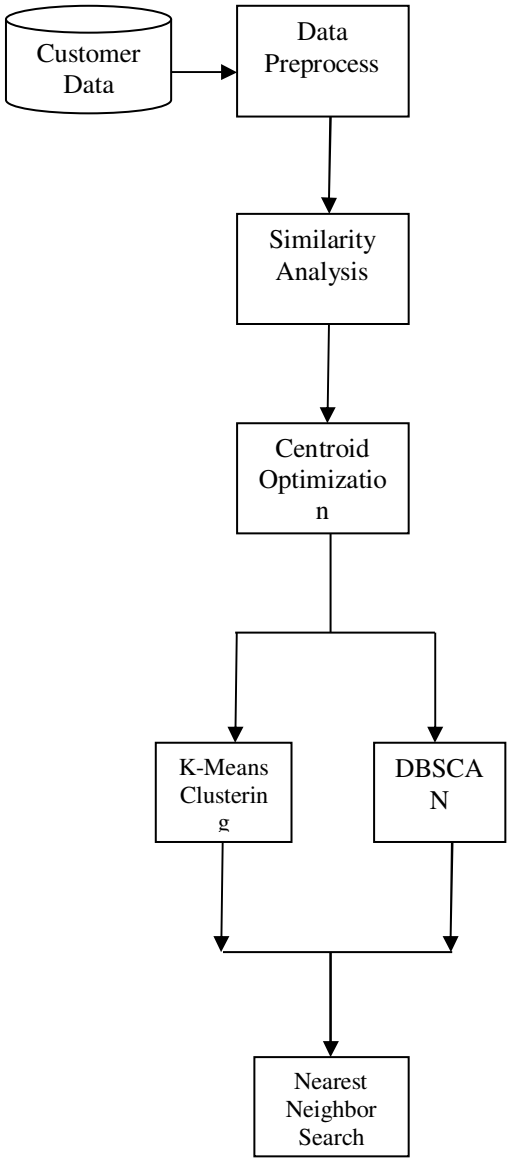

Figure 5.1. Cluster Based Nearest Neighbor Search

The system supports partitional and density based clustering process on uncertain data values. Cluster centroid optimization scheme is integrated in the K-means clustering algorithm. K-nearest neighbor search algorithm is used to search data on clusters. The system is divided in to five major modules. They are Data preprocess, Similarity analysis, Partitional clustering process, Density based clustering process and Nearest neighbor search. The data preprocess module is designed to construct transaction matrix. Similarity analysis module is used to estimate the relationship between the transactions. Partitional data clustering is performed with K-means and distribution based similarity analysis. Density based clustering is performed with DBSCAN and distribution similarity analysis. Nearest neighbor search is carried out on clustered data values.

5.1. Data Preprocess

Sales transactions are maintained in data files. Uncertainty analysis is carried out to verify product count in each record. Transaction matrix group ups the bill and product details. Product code and product name are maintained in product list. The customer transactions form is used to fetch the customer data values from the database. The uncertainty analysis form is used show the uncertainty level for the transactions. The customer transactions are converted into transaction matrix. The product code and product name details are provided in product details form.

5.2. Similarity Analysis

Similarity measures are used to fetch the transaction relationships. Geometric distance and (KL) divergence methods are used in the similarity analysis. Individual records values are compared with their weight values in geometric distances. KL divergence method uses the distribution ratio for the similarity analysis. The similarity analysis is used fetch the relationship between the transactions. The geometrix similarity is calculated for the selected transaction set. The Kullback Leibler divergence similarity models are used in the system. The similarity values are calculated and displayed in the same form.

5.3. Density Based Clustering Process

Density and distribution analysis is performed on transaction matrix values. Transaction matrix is updated with distribution values. DB Scan method is used for the clustering process. Centroid optimization scheme is used to improve the cluster accuracy levels. The density and distribution analysis is carried out to find out the transaction distribution levels. The DB Scan scheme is integrated with KL divergence similarity model. The clustering process is performed with cluster count collected from the user. The cluster results are produced in separate form. The Inter Cluster Distance (ICD) scheme is used to select optimal centroid values.

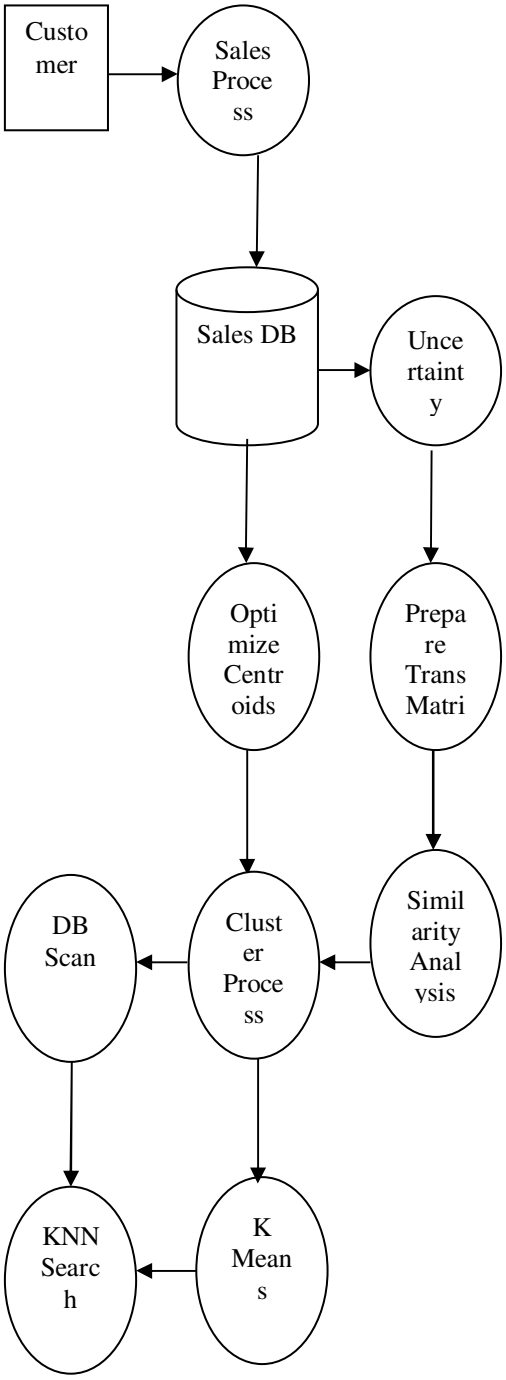

Figure 5.2. Uncertain Data Clustering Process

5.4. Partitional Clustering Process

Partitional clustering is performed using K-means clustering algorithm. KL divergence similarity is used with K-means clustering algorithm. Random Centroid selection scheme is replaced with optimal centroid selection model. Inter cluster distance analysis is applied in optimal centroid selection process. The K-means clustering scheme is used with KL divergence based similarity analysis process. The clustering process is performed with cluster count collected from the user. The system uses the Inter

Cluster Distance (ICD) based optimal centroid for clustering process. The clustering results are listed in separated form for the selected cluster name.

5.5. Nearest Neighbor Search

Nearest neighbor search process is carried out to identify relevant transactions. K-Nearest Neighbor (KNN) search technique is used in the system. KNN search process is applied on the cluster results. Top K items are fetched with reference to the similarity values. The K-nearest neighbor search mechanism is used in the system. The KNN searh operations are carried out on the clustered data values. The transactions are ranked with similarity values. The KL divergence similarity model is used to fetch the nearest neighbor transactions. The system performance is analyzed with different performance matrix and their results are represented in graphical form.

VI. Performance Analysis

The data clustering scheme is designed to partition the uncertain data values. The Density and distribution based clustering with Kullback-Leibler divergence similarity (DBKL) technique and K-Means clustering with Kullback-Leibler divergence similarity (KMKL) techniques are used in the clustering process. The clustering scheme is tuned to handle data uncertainty. Data distribution and density are analyzed using the DBScan method. The similarity analysis is performed with geographic relationships. The Kullback-Leibler divergence similarity is used to measure the relationship between the transactions. The K-Nearest Neighbor (KNN) search process is performed on the partitioned data values. The clustering system is implemented using the Java language and Oracle relational database environment. The system is tested with the Customer sales transaction data set. The F-measure, purity and separation index performance parameters are used to evaluate the cluster accuracy levels. K-Nearest Neighbor (KNN) search process is analyzed with accuracy levels.

6.1. Datasets

S. No:	Attribute Name	Description
1	Billno	Bill number
2	Dop	Date of purchase
3	Pname	Name of the product
4	Price	Unit price
5	Qty	Product quantity
6	Amount	Total amount

Table No: 6.1. Attribute details for Customer Sales Transaction Datasets

6.2. F-measure

The F-measure is a harmonic combination of the precision and recall values used in information retrieval. Each cluster obtained can be considered as the result of a query, whereas each preclassified set of transactions can be considered as the desired set of transactions for that query. Thus, the system can calculate the precision P(i, j) and recall R(I, j) of each cluster j for each class i.

If n_i is the number of members of the class i, n_j is the number of members of the cluster j, and n_{ij} is the number of members of the class i in the cluster j, then P(i, j) and R(i, j) can be defined as

$$P(i, j) = \frac{n_{ij}}{n_j}, \qquad (1)$$

$$R(i, j) = \frac{n_{ij}}{n_i} \qquad (2)$$

The corresponding F-measure F(i, j) is defined as

$$F(i, j) = \frac{2 * P(i, j) * R(i, j)}{P(i, j) + R(i, j)} \qquad (3)$$

Then, the F-measure of the whole clustering result is defined as

$$F = \sum_i \frac{n_i}{n} \max_j (F(i, j)), \qquad (4)$$

where n is the total number of transactions in the data set. In general, the larger the F-measure is, the better the clustering result is (2).

The F-measure analysis is performed to estimate the accuracy level for the clustering techniques. Precision and recall measures are calculated for the F-measure analysis process. The F-measure analysis is shown in figure 6.1. The analysis show that the K-Means clustering with Kullback-Leibler divergence similarity (KMKL) technique produces accuracy level 25% more than the Density and distribution based clustering with Kullback-Leibler divergence similarity (DBKL) technique.

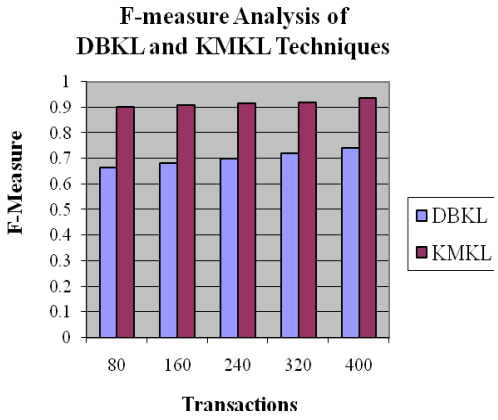

Figure 6.1: F-measure Analysis between DBKL and KMKL

6.3. Purity

The purity of a cluster represents the fraction of the cluster corresponding to the largest class of transactions assigned to that cluster; thus, the purity of the cluster j is defined as

$$Purity\ (j) = \frac{1}{n_j} \max_i (n_{ij}) \qquad (5)$$

The overall purity of the clustering result is a weighted sum of the purity values of the clusters as follows:

$$Purity = \sum_j \frac{n_j}{n} Purity\ (j) \qquad (6)$$

In general, the larger the purity value is, the better the clustering result is (6).

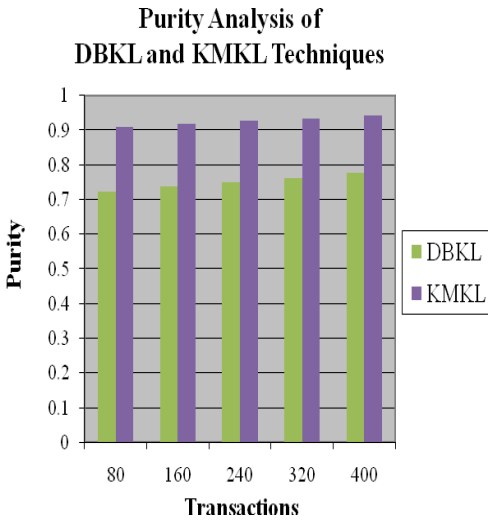

Figure 6.2: Purity Analysis between DBKL and KMKL

The purity measure is also used to evaluate the cluster accuracy levels. The purity analysis for DBKL and KMKL techniques are produced in figure 6.2. The analysis results show that the KMKL produces s accuracy level 20% than the DBKL technique.

6.4. Separation Index

Separation Index (SI) is another cluster validity measure that utilizes cluster centroids to measure the distance between clusters, as well as between points in a cluster to their respective cluster centroid. It is defined as the ratio of average within-cluster variance to the square of the minimum pairwise distance between clusters:

$$SI = \frac{\sum_{i=1}^{NC}\sum_{xj\in ci} dist(x_j, m_i)^2}{N_D \min_{1\le r,s\le N_c} dist(m_r, m_s)\}^2}$$

$$= \frac{\sum_{i=1}^{N_c}\sum_{xj\in ci} dist(x_j, m_i)^2}{N_D \cdot dist_{\min}^2}$$

where m_i is the centroid of cluster c_i, and distmin is the minimum pairwise distance between cluster centroids. Clustering solutions with more compact clusters and larger separation have lower Separation Index, thus higher values indicate better solutions. This index is more computationally efficient than other validity indices, such as Dunn's index, which is also used to validate clusters that are compact and well separated. In addition, it is less sensitive to noisy data.

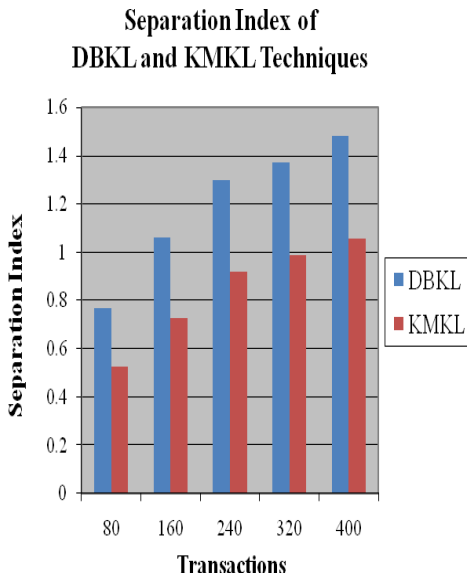

Figure 6.3: Separation Index Analysis between DBKL and KMKL

The separation index measure is used to analyze the cluster intervals and transaction interval values. The separation index analysis is shown in figure 6.3. The analysis result shows that the KMKL improves the cluster quality 20% than the DBKL technique.

6.5. Accuracy Level

The accuracy level analysis is performed for the K-Nearest Neighbor (KNN) search process. The KNN search process is performed on the partitioned data values. The accuracy level is measured with the ratio between the retrieved transactions and actual transactions for the query values. Different k values are used to measure the accuracy level. Density and distribution based clustering with Kullback-Leibler divergence similarity (DBKL) technique and K-Means clustering with Kullback-Leibler divergence similarity (KMKL) technique are integrated with the KNN search process. The accuracy level analysis is shown in figure 6.4. The KNN search with KMKL cluster model produces 20% accuracy level better than the KNN search with DBKL cluster model.

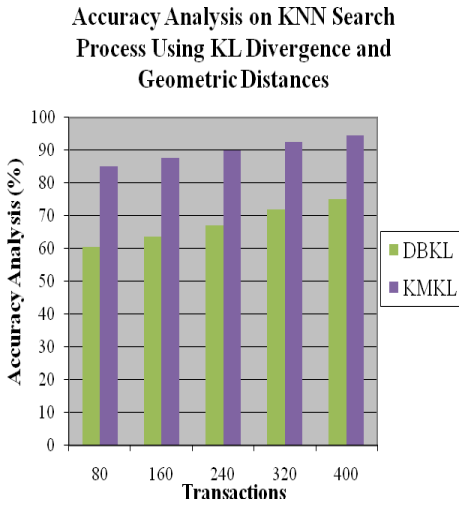

Figure 6.4: Accuracy Analysis between DBKL and KMKL

VII. Conclusion And Future Enhancement

Clustering techniques are used to group up the customer data values. Probability distribution based similarity analysis is used for the uncertain data environment. Centroid optimization scheme is used to improve the clustering process. K-Nearest Neighbor search (KNN) algorithm is used to fetch similar data values. The system supports uncertain data clustering with high scalability. Partitional and density based clustering operations are supported by the system. Optimal centroid model improves the cluster accuracy level. Nearest neighbor data search is handled by the system.

The customer transaction data clustering system is implemented to group up the sales transactions with relevance. Kullback Leibler divergence similarity is used with density and distribution analysis mechanism. The K-Means clustering scheme is used with optimal centroid estimation mechanism. The system also supports K-Nearest Neighbor (KNN) search process. The system can be enhanced with the following features.

- The system can be enhanced support uncertain data clustering under distributed environment.
- The system can be enhanced with privacy preservation models to support privacy preserved data mining operation on sensitive data values.
- The clustering scheme can be improved to support data stream based data partitioning process.

References

[1] Bin Jiang, Jian Pei, Yufei Tao and Xuemin Lin, "Clustering Uncertain Data Based on Probability Distribution Similarity", IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 4, April 2013

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

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

[4] A. Banerjee, S. Merugu, I.S. Dhillon, and J. Ghosh, "Clustering with Bregman Divergences," J. Machine Learning Research, vol. 6, pp. 1705-1749, 2005.

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

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

[7] 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.

[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] H.-P. Kriegel and M. Pfeifle, "Density-Based Clustering of Uncertain Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining (KDD), 2005.

[11] W.K. Ngai, B. Kao, C.K. Chui, R. Cheng, M. Chau, and K.Y. Yip, "Efficient Clustering of Uncertain Data," Proc. Sixth Int'l Conf. Data Mining (ICDM), 2006.

Mr. D. Parthipan received M.Sc(IT) degree from Anna University, Chennai, and doing M.Phil., Degree in Kongu Arts and Science College, Bharathiar University, Coimbatore, TN, India . He has presented paper in National Conference. E-mail ID: parthimail4u@gmail.com

Dr. M. Moorthi received MCA degree from Bharathiar University, Coimbatore, and M.Phil., Degree from Manonmaniam Sundaranar University and PhD Degree centre from Bharathiar University, Coimbatore, TN, India . He is currently the Assistant Professor in Kongu Arts and Science College, Erode, TN, India. He has 15 years of teaching and 10 years of research experience. He has guided eleven M.Phil students in the area of Computer Science. He has presented papers in National and International Conference and has published an article in National Journal. He is a member of ISCA and working as Associate Editor in Canadian Research & Development Center of Science and Cultures – Advances in Natural Science and Management Engineering – ISSN 1913-0341.His interests and expertise are in the area of Image Processing, Data Mining, Multimedia, Computer Graphics and Networks. E-mail ID: moorthi.bmka@gmail.com.

