ASSESSMENT OF UNCERTAIN DATA CLUSTERING USING A MODIFIED K-NEAREST NEIGHBOR METHOD

Kim S.Y. & Riccardi P.M.
Department of Orthopedic Rehabilitation, Incheon Medical University, Incheon, South Korea
Department of Musculoskeletal Medicine, Naples Clinical Academy, Naples, Italy

ABSTRACT

Clustering is the process of making the group of abstract objects into classes of similar objects. A cluster of data objects can be treated as one group. Clustering of uncertain data has been well recognized as an important issue. This research paper proposes clustering of uncertain datasets, where similarities can be measured between datasets according to the characteristics. First, the user is registered with the server and the server verifies the user's details with the database. After verification user sends the uncertain data to the server. The server uses KLL divergence mechanism for classifying discrete and continuous case data and computes the similarity of the data. Finally, apply K-NN algorithm to compute the distance between the nearest nodes and cluster the data. This method provides efficient clustering of uncertain data compared to other clustering methods. In the proposed system, KL (Kullback-Leibler) divergence and KNN approach are used together to overcome the existing drawback and to produce an effective performance with less time complexity.

Keywords: *Clustering, Uncertain Data, Probability mass function, Probability density function, density estimation*.

I. INTRODUCTION

Cluster analysis, first partition the set of data into groups based on data similarity and then assigns the label to the groups. It has application in various areas of computer science such as machine learning, data compression, data mining, or pattern recognition. Depending on the application we want to cluster such diverse objects as text documents, probability distributions, etc. different objects and different applications also require different notions of dissimilarity of objects. As a consequence, there are numerous different formulations of clustering. First step towards understanding clustering problems with nonmetric or metric dissimilarity measures, like Kullback-Leibler divergence. 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 centre. 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 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).For sensor measurements may be imprecise at a certain degree due to the presence of various noisy factors (e.g., signal noise, instrumental errors, and wireless transmission) at a certain degree due to the presence of various noisy factors (e.g., signal noise, instrumental errors, and wireless transmission).

Clustering Analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing. Clustering can also help marketers discover distinct groups in their customer basis. And they can characterize their customer groups based on purchasing patterns. In field of biology it can be used to derive plant and animal taxonomies, categorize genes with similar functionality and gain insight into structures inherent in populations. Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according house type, value, and geographic location.

As a data mining function, Cluster Analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster. To overcome this issue our proposed system introduces K-nearest-neighbour algorithm to calculate the nearest neighbour. The K-nearest-neighbour (KNN) algorithm measures the distance between a query scenario and a set of scenarios in the data set. Here, distance is calculated for both continuous and discrete cases by using Probability mass function. Then, nearest neighbour is calculated by applying KNN approach. Thus, our proposed overcomes the existing drawback and produce effective result.

II. LITERATURE REVIEW

The K-Nearest Neighbor algorithm (k-NN) is a non-parametric method for classifying objects based on closest training examples in the feature space. K-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor. Dimension reduction – both feature selection and sample selection. Dimension reduction is of particular importance with k-NN as it has a big impact on computational performance and accuracy [30].

High Calculation Complexity: To find out the k nearest neighbor samples, all the similarities between the training samples must be calculated. When the number of training samples is less, the KNN classifier is no longer optimal, but if the training set contains a huge number of samples, the KNN classifier needs more time to calculate the similarities. This Algorithm can be solved in 3 ways: reducing the dimensions of the feature space; using smaller data sets; using improved algorithm which can accelerate.

Dependency on the training set: The classifier is generated only with the training samples and it does not use any additional data. This makes the algorithm to depend on the training set excessively; it needs recalculation even if there is a small change on training set.

No Weight Difference between Samples*:* All the training samples are treated equally; there is no difference between the samples with small number of data and huge number of data. So it doesn't match the actual phenomenon where the samples have uneven distribution commonly.

Anil K.Jain provided a brief overview of clustering, summarize well known clustering methods, discuss the major challenges and key issues in designing clustering algorithms, and point out some of the emerging and useful research directions, including semi-supervised clustering, ensemble clustering, simultaneous feature selection during data clustering, and large scale data clustering. Clustering is in the eye of the beholder, so indeed data clustering must involve the user or application needs.

Yang et al., proposed an efficient data clustering algorithm. It is well known that K-Means (KM) algorithm is one of the most popular clustering techniques because it is unproblematic to implement and work rapidly in most situations. But the sensitivity of KM algorithm to initialization makes it effortlessly trapped in local optima. K-Harmonic Means (KHM) clustering resolves the problem of initialization faced by KM algorithm.

Chen Zhang et al., presented a new clustering method based on K-Means that have avoided alternative randomness of initial center. This approach focused on K-Means algorithm to the initial value of the dependence of K selected from the aspects of the algorithm is improved. First, the initial clustering number is N. Second, through the application of the sub-merger strategy the categories were combined.

III. PROPOSED METHODOLOGY

The proposed method includes K-nearest-neighbor (KNN) algorithm to measure the nearest neighbor in uncertain data. To examine the distance calculation, our system uses probability mass function. The KNN works based on matrix computation for both continuous and discrete cases. The proposed methodology system architecture design is as follow.

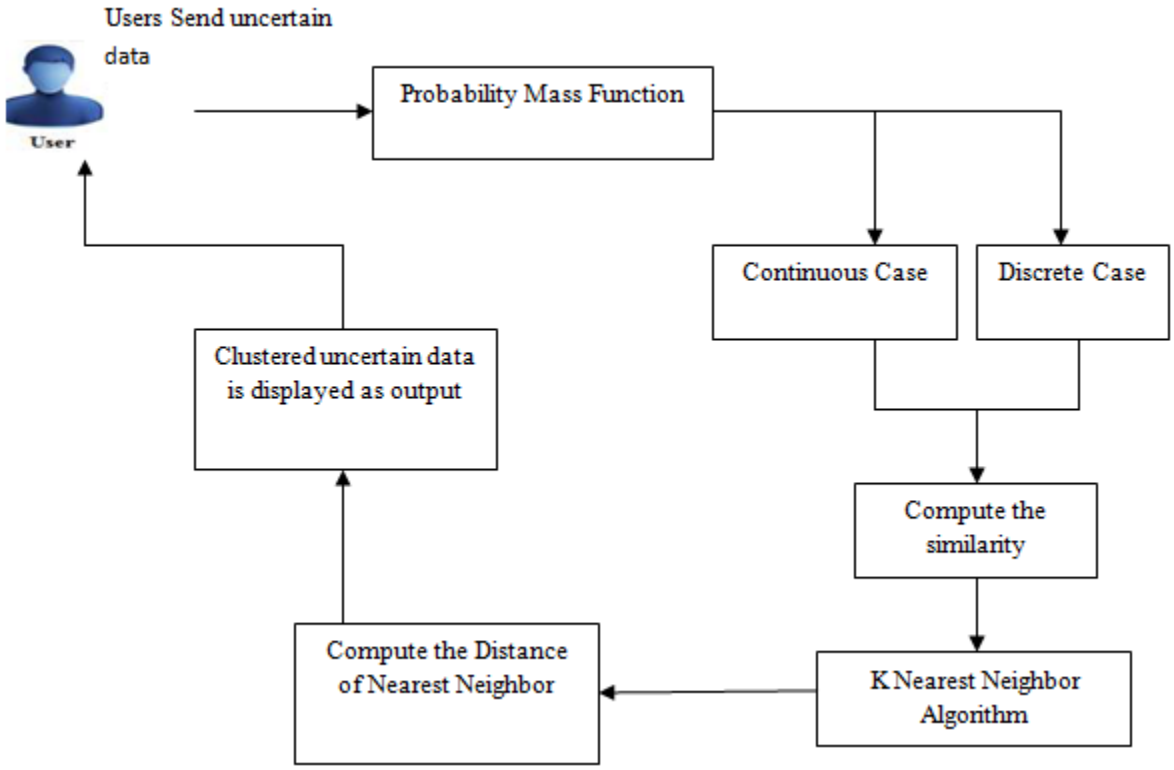

Figure 1. Proposed Methodology Architecture

User Authentication
It contains authentication details between client and server. First users make registration by entering the required details and sent to server. Server validates the user details and sent authentication details

KL Divergence Approach
This approach finds the similarity of uncertain data. This similarity is more helpful to cluster the data. Then cluster the uncertain data according to the similarity.

KNN Approach
After partitioning the datasets, the partitioned data sets are taken as input for KNN approach. This KNN approach finds the distance between every two node. Then if the distance is nearest than other node then that node is considered as nearest node.

Data Transmission to User
After clustering the dataset the server sends the dataset to requested user. Then the user receives the dataset and views. This process contains details about transmitting the data to user after clustering the data.

IV. ALGORITHM AND APPROACH

KL-divergence algorithm
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. It is important to note that the definition of KL divergence necessitates that for any x ϵ ID if P (x) > 0 then Q (x) > 0. To ensure that the KL divergence is defined between every pair of uncertain objects, we smooth the probability mass/density function of every uncertain object P so that it has a positive probability to take any possible value in the domain.

KNN approach

K-Nearest Neighbours (KNN) classification divides data into a test set and a training set. For each row of the test set, the K nearest (in Euclidean distance) training set objects are found, and the classification is determined by majority vote with ties broken at random. If there are ties for the Kth nearest vector, all candidates are included in the vote.

Suppose each sample in our data set has n attributes which we combine to form an n-dimensional vector: x = (x1, x2,..xn). These n attributes are considered to be the independent variables. Each sample also has another attribute, denoted by y (the dependent variable), whose value depends on the other n attributes x. We assume that y is a categoric variable, and there is a scalar function, f, which assigns a class, y = f(x) to every such vectors. We do not know anything about f (otherwise there is no need for data mining) except that we assume that it is smooth in some sense. We suppose that a set of T such vectors are given together with their corresponding classes:x(i), y(i) for i = 1, 2, . . . , T. This set is referred to as the training set.

V. CONCLUSION

A new mechanism is proposed in this paper for clustering uncertain data. First the user is registered with server and the server verifies the user's details with the database. After verification user send the uncertain data to the server. The server uses KLL divergence mechanism for classifying discrete and continuous case data and computes the similarity of the data. Finally apply K-NN algorithm to compute the distance between the nearest nodes and cluster the data. This proposed method provides efficient clustering of uncertain data compared to other clustering methods.

REFERENCES

1. *C. Fernandes, A.M. Mora, J.J. Merelo, V. Ramos and J.L.J. Laredo, "KohonAnts: A Self-Organizing Ant Algorithm for Clustering and Pattern Classification", http://arxiv.org/abs/0803.2695v1, 2008.*
2. *J. Han and M. Kamber. "Data Mining: Concepts and Techniques", Morgan Kaufmann, San Francisco, 2001.*
3. *J. Handl, J. Knowles and M. Dorigo, "Strategies for the increased robustness of ant-based clustering, Lecture Notes in Computer Science, Vol. 2977, 2004, pp. 90-104.*
4. *Lewis-Beck, Michael S. Data Analysis: an Introduction, Sage Publications Inc, ISBN 0-8039-5772-6, 1995*
5. *A.K. Jain, M.N. Murty and P.J. Flynn, "Data Clustering: a review", ACM Computing Surveys, Vol. 31, 1999, pp. 264-323.*
6. *Kiran Kumar Reddi, Integrating Fuzzy C-Means Clustering Technique with K-Means Clustering Technique for CBIR, International Journal of Computers and Distributed Systems, Vol. 3, Issue III, Aug-Sep 2013.*
7. *M.N. Murty, A.K. Jain, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264–323.*
8. *Aristidis Likas, Nikos Vlassis, JakobJ. Verbeek – "The global k-means clustering algorithm" 2003- 451 461- Pattern Recognition Society. Published by Elsevier Science Ltd. served. PII: S0031-3203(02)*
9. *Haung, A similarity measure for text document clustering. In proceeding of the sixth New Zealand computer research student conference, page 49-56, 2008.*
10. *Cao,F.Liang, J.Li, D.Bai. "A dissimilarity measure for k-mode clustering algorithm" , knowledge based system,26(1):120-127, 2012..*