Experimental work on Data Clustering using Enhanced Random K-Mode Algorithm

S. Sathappan [1], D.C. Tomar [2]
[1,2](Department of Computer Science & Engineering, Sathyabama University, Chennai, India)

ABSTRACT: *Clustering the uncertainty data is not an easy task but an essential task in data mining. The traditional algorithms like K-Means clustering, UK Means clustering, density based clustering etc, to cluster uncertain data are limited to using geometric distance based similarity measures and cannot capture the difference between uncertain data with their distributions. Such methods cannot handle uncertain objects that are geometrically indistinguishable, such as products with the same mean but very different variances in customer ratings. Because of its complexity, the clustering takes high execution time resulting in high computational cost. In this proposed method is Enhanced Random K-Mode algorithm which is also called as ERK-Mode to cluster the uncertainty data. The K-mode concept classifies the dataset and separates as certain and uncertain data from the whole dataset. Again enhanced random K-Mode is used to cluster the uncertainty data. The Weather data values are taken in to the account for experiments. The experiment shows that the proposed algorithm is very efficient with fast execution time and low complexity.*

Keywords - *uncertainty clustering, weather dataset, random k-mode, probability density function.*

I. INTRODUCTION

In the real world, data mining applications are affected by data's uncertainty. All clustering algorithms aim of dividing the collection all data objects into subsets or similar clusters. A cluster is a collection of objects which are „similar" between them and are „dissimilar" to the objects belonging to other clusters and a clustering algorithm aims to find a natural structure or relationship in an unlabeled data set. Due to uncertainty into account during the computations, designing of data mining technique has become critical. For measuring the applications like weather station monitors, weather conditions, hardware techniques in the real time, the uncertainty is measured through elements like temperature, precipitation amount, humidity etc. The Uncertain data can be clustered by probability density function equation:

$$PDF(z_{l,1}, z_{l,2} \dots z_{l,m}) = \int_1^m f_l(z_{l,m})dx$$

For example, the mode of an object's probability density function can be used as typical point. However, it gives outstanding clustering results than traditional methods. K-mode algorithm uses values as the highest number of occurrence as a cluster head. K-means supports only for numerical but K-mode supports both Numerical and categorical dataset. Major contributions in the real time application for this paper: Weather dataset has been taken with parameters like temperature, humidity and are analyzed. The previous studies on clustering uncertain data are largely various extensions of the traditional clustering algorithms considered for certain data. Here the object in certain dataset is considered as a single point and distribution concerning the object itself is not considered in traditional clustering algorithms. New algorithm Enhanced Random K-mode has been proposed to cluster the uncertain data efficiently.

II. PROBLEM STATEMENT

Clustering the certain data is a normal process but clustering the uncertainty data is not an easy task. 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 randomized k-mode algorithm is used to improve the accuracy of the clusters. In k-mode clustering, number of cluster depends on dataset value which results in time consumption.

III. PROPOSED METHOD

Here we consider weather dataset values consist of different parameters like temperature, humidity etc. The dataset is given as input to the K-mode process where the dataset can be classified into two clusters as

certain and uncertain data. Uncertainty data's are given as input for the Enhance random K-mode algorithm to cluster these data into different categories to find out uncertain values for each parameter to evaluate its efficiency. Then, we compare the output of the two algorithms to find the best one to predict uncertain values in climatic or weather conditions.

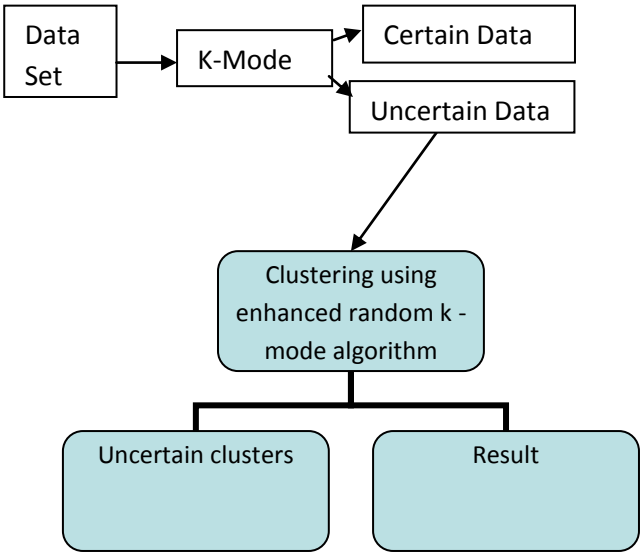

Fig. 1: Proposed diagram for algorithm

Consider the Weather Dataset as D which has the set of objects X, X consists of set of attributes 'X' each attribute has different domain values 'Dom' which is represented as 'x' these are described below

$D=\{X_1, X2 \ldots X_n\}$

Where n is the number of objects or Data Sets

$Xi = [x_{i1}, x_{i2} \ldots x_{im}]$

Where m is number of categorical attributes

$Dom(X_j) = \{x_j^{(1)}, x_j^{(2)} \ldots, x_j^{(pj)}\}$

Where *pj* is the number of category values of attribute Xj.

Algorithm Steps

Phase I:

a) Select K initial modes, one for each of the cluster.
b) Allocate data object to the cluster whose mode is nearest to it according to parameters like temperature, humidity etc.
c) Compute new modes of all clusters.
d) Repeat step 2 to 3 until no data object has changed cluster membership.

First, choose the cluster head by probability density function and then select the object nearby the cluster head. Now allot the object to cluster which is near to the object. Similarly compute for remaining modes of all the clusters. Repeat the above steps until the data object doesn't change the cluster membership.

Phase II:

Pseudo code:

Input: Uncertain dataset from phase I

Output: uncertain clusters

1. Associate each data point to the most similar mode.
2. For each mode m for each non mode data o pick random value m and o and compute the total cost of the randomness.
3. Select the random value with the lowest cost.

4. Repeat steps 2 to 3 until the clusters are not changed.

In the second phase, uncertain data is taken as input from phase I and random k-mode algorithm is applied to get efficient clusters without lot of computations.

IV. DATASET DESCRIPTION

Weather dataset is taken and various readings for the past 4 years are taken from 2012 to 2015 consisting of temperature and humidity values. The dataset includes reading time, temperature, and speed of fan and some humidity details. The fundamental data includes the data relating to humid and temperature values taken at 100 samples per second. Uncertain data are grouped into different clusters using phase I followed by Phase II of above algorithm. Month wise data is also taken and analyzed.

4.1 TEMPERATURE AND HUMIDITY VALUES

The following example shows the different temperature and humidity values taken in the corresponding year.

2012 1 13.007 1 25.9 1 20.7 1 21 1 2013 1 11.03 1 21.96 1 17.87 1 17 1 2013 1 14.01 1 27.9 1 18.9 1 18 1 2014
x L y L m L n L u L x L y L m L n L u L
x L y L m L n L u L

x - year
m - minimum humidity(%)
n - maximum humidity (%)
u - No of Uncertainty
L - separator

In the above example, the values represent the various temperature and humidity values which are taken in the corresponding years. Each reading is separated by a separator.

4.2 DATASET SETUP

The dataset is collected from metrological department for the year 2010 to 2015to predict the weather values in the year 2016 regarding climate changes due to uncertain temperature and humidity values. We use dot net framework to implement the system. The dataset is collected month wise day wise for the past years.

V. RESULTS

TABLE I. Weather Data Taken from different file

Year	Minimum humidity	Maximum humidity	Mode Humidity	Uncertain Value
2012	13.006	25.997	20.786	21
2013	11.003	21.967	17.876	18
2014	14.001	27.998	18.997	18
2015	13.889	16.893	14.988	19

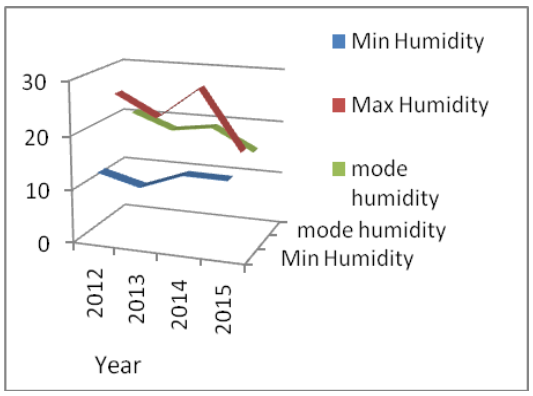

Fig. 2: Humidity values at different time

In the above graph, x-axis represents Year Wise Prediction values. By analyzing the graph we can conclude that the uncertain values depends mode values and the number of uncertain data decreases when the mode value decreases.

TABLE II. Weather Data temperature for year 2015

Month	Temperature
January	49
February	50
March	51
April	50
May	52
June	50
July	48
August	51
September	54
October	51
November	53
December	49

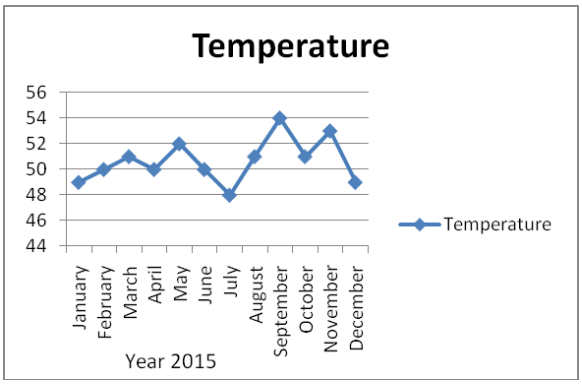

Fig. 3: Data taken from same Data set file

In the above figure, the x-axis represents months for year 2015. By the above graph we can analyze that the temperature for the year 2016 can be predicted.

VI. CONCLUSION

The traditional algorithms were focused for neither categorical nor numerical data, but our proposed method is suitable for all kinds of data. The above experiments are proved our enhanced k-Mode algorithm

effectively with Gas Sensor's numerical values. The K-Mode algorithm and Probability density calculations only used and so the complexity is also reduced. Moreover the computational cost is very low. The accuracy in producing the resulting clusters is good.

VII. ACKNOWLEDGEMENTS

The authors thank Dr. S. Sridhar, Professor and Dean, Cognitive & Central Computing, R.V. College of Engineering, Bangalore, India for guiding to present the ideas in the form of a paper by giving a good shape.

REFERENCES

[1] W. K. Ngai, B. Kao, R. Cheng, M. Chau, S. D. Lee,D. W. Cheung, and K. Y. Yip, "Metric and trigonometric pruning for clustering of uncertain data in 2D geometric space", Information Systems,36 (2), 2011, 476–497.
[2] B. Kao, S. D. Lee, F. K. F. Lee, D. W. L. Cheung and W. S. Ho, "Clustering uncertain data using Voronoi diagrams and R-tree index", IEEE TKDE, 22(9) ,2010, 1219–1233.
[3] T. Velmurugan, "Efficiency of K-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points", IJCTA, 2012.
[4] S. Anjani and M. Wangjari, "Clustering of uncertain data object using improved K-Means algorithm", IJARCSSE, 2013.
[5] B. Kao Sau D. L. Foris, K. F. L, David, W. Cheung and W. S. Ho," Clustering Uncertain Data using Voronoi Diagrams and R-Tree Index", IEEE, 2010.