ISSN No:-2456-2165

A Coherent Approach towards Clustering Uncertain Data

Sylvia S. Nair[1], Prasanna Lohe[2]
[1] Department of Computer Science and Engineering, J.D. College of Engineering and Management, Nagpur, India
[2] Department of Computer Science and Engineering, J.D College of Engineering and Management, Nagpur, India

Abstract:- Data mining consists of a very vital task that is clustering, where not only the predictable data but also the clustering of uncertain data can be seen. In the previous papers, to cluster set of uncertain data some basic techniques that were used like DBSCAN, k-means and other methods which bank on distances between the object in a geometrical way. The methods that were used in previous papers cannot tackle with such uncertain data that has high computational complexity, it's like when a product has alike mean but varies in customer's grading. Thankfully, the probability distribution method, which is one of the most significant part in uncertain data, and will help in finding efficient ways to overcome the previous methodology used. In this paper we can comprehensively distinguish the data in continuous as well as in discrete form. In the previous papers, Kullback-Leibler Divergence has also been used to calculate the difference between data in continuous as well as discrete form, respectively. Using the KL-Divergence gave appropriate results but had a costly impression on the project. Therefore, in this proposed paper Skew Divergence which measures the alikeness of uncertain objects is put to use with Fuzzy c-means making the work of clustering uncertain data feasible but the more improved version of clustering is being used i.e. fast gauss transformation which is more flexible than the previous methods of clustering.

Keywords:- *Uncertainty of Data, Clustering, Fuzzy C-Means, Skew Divergence, Fast Gauss Transformation.*

I. INTRODUCTION

In past few years, a variety of methods came into existence and was put to use to experiment things on data mining when it comes to clustering. Uncertain data was one of the topics that had put the researches into exploring data mining more deeply to know about it because nobody had ever put much focus on such kind of objects.

These uncertain objects or data are the ones that cannot be tackled using the regular methods of clustering. For example, k-means is one of the techniques used in the previous papers but k-means itself has computational problems because of which it uses many heuristic functions to reduce the problem of complex computation. These uncertain data or objects are mostly seen in large databases and warehouses where data from various sources are taken.

Uncertain data are basically the error that makes it deviate from the desired output. Data is continuously growing especially on web. For example, there may be a possibility of uncertain address of a customer in an enterprise data set or uncertain weather with varying temperature or uncertain temperature captured by an old senor network giving wrong readings. There may be possibility of errors while capturing weather information because of environment instability which makes instability in database as well more likely known as uncertain data. Since there are varieties of users so every end-user will have its own perception of a particular product. Variety of end-user have different scrutiny on a particular product, now this does not mean uncertainty, basically when these end-users have different perception on its features then that may cause uncertainty, like some customers may be fascinated with the latest software in that laptop some may be happy to have more RAM, but there may be some end-user who find the new software confusing because its new for them. So every data that are captured in the database come from different sources that may contain data certain or uncertain data. If a certain data is important then uncertain data is no less, it is as important as certain data as it may contain or give us some vital information.

Next example, where a device detects the temperature, moisture in air, direction of the wind, possibility of rain or heavy storm these kinds of data are all uncertain data. This instability of weather information is gathered forming a cluster. The data of an entire month is gathered and is analysed to form a cluster of uncertain data.

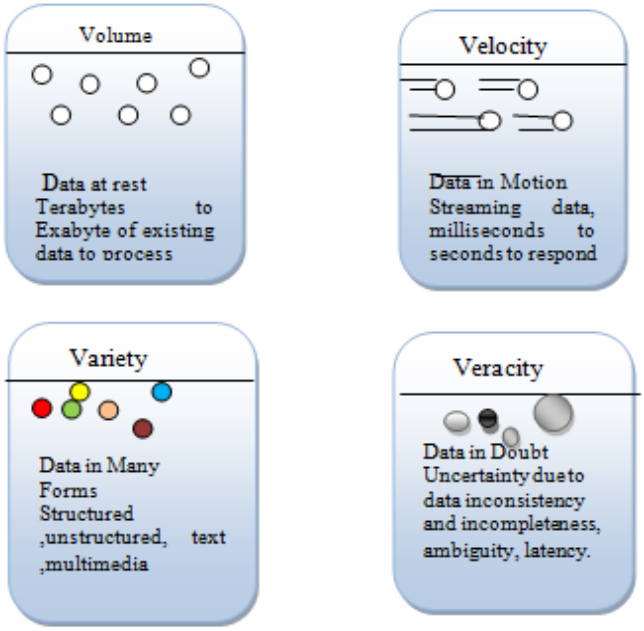

Fig 1

ISSN No:-2456-2165

In figure 1, it states that data can be of any kind like a data in rest mode or a moving data or different kinds of data lie structured or unstructured or data that are doubted if that really is a data or just some unnecessary object. The uncertainty of data can come in any type.

II. LITERATURE SURVEY

Lately, various researches have tried to work on clustering. Clustering similar objects or data was somehow easy but it came to clustering uncertain data things got complicated.

Anil K. Jain [1] had put forward K-means method for the clustering uncertain data. He says that K-means had been used since 50 years and is still widely used. The clustering of uncertain data requires indexing and vornoi clustering as K-means works with indexing that helps in increasing the efficiency whereas, vornoi clustering in filtering the set of clusters formed as an output.

Aakanksha Dixit [2] proposed the method of UK-mean in clustering uncertain data. According to her paper UK-mean algorithm tends to be more efficient that the K-mean algorithm as the UK-mean calculates Expected Distance (ED) between the objects in every cluster as it works on probability function (pdf) that makes the UK-mean more reliable than K-mean and to overcome its drawbacks UK-mean was introduced.[3] rainfalls or hot climate have intense effect on urban and rural areas, the climate may sometimes have dry air or moisture because of which the precipitation in the atmosphere varies that can give uncertain temperature data sets, therefore, UK-Means helps in calculating the results.

In [4] Ajit Patil, M.D. Ingle proposed that KL-Divergence calculates the distributed similarity amongst the data objects. Although there are many techniques that are used to cluster uncertain data but not all techniques can help in finding the results after clustering uncertain data as some may need geometrical expression and not every data set might work on it, therefore, KL-Divergence was introduced but it also has a major drawback of being expensive that makes it inefficient. [5] KL-Divergence here calculates the similarity of the objects. Objects in the data set may be different from one another like some may be discrete values and some may be continuous depending upon that the data is modeled. Nevertheless, this naïve approach may be quite expensive.

In [6] this paper proposed DBSCAN that overcame all the problems that occurred during using the above mentioned methods. Other methods single values to find or calculate distance between two similar objects in the data set this becomes a naïve approach towards calculating the distances amongst the objects and to deal with those values single handedly made it have a major drawback for using it.

Donghua Pan; Lilel Zhao [7] proposed the method of DBSCAN with CIR (Core Influence Rate). This CIR-DBSCAN works on models based on the distance calculated between the objects in data set that extends the methods used in simple DBSCAN. There was one more method used for clustering uncertain data FDBSCAN which when compared with CIR-DBSCAN gave poor results which made CIR-DBSCAN more efficient.

H-P. Kriegel ; M. Pfeifle[8] has proposed OPTICS that had helped many end users to get an overview of the large dataset. When the OPTICS are used it considers the single values that has been calculated. Also when two fuzzy objects are being calculated FOPTICS are being used that gave more effective results but again it would handle only single values. [9] Unlike OPTICS that works on single values the FOPTICS also works on single values but using different and simpler approach, therefore, making this method slow and naïve.

[10] An Efficient Methodology for Clustering Uncertain Data Based on Similarity Measure, a paper that works on clustering uncertain data using skew divergence also shows the comparison between KL divergence and skew divergence where skew divergence turned out to be better than KL divergence.

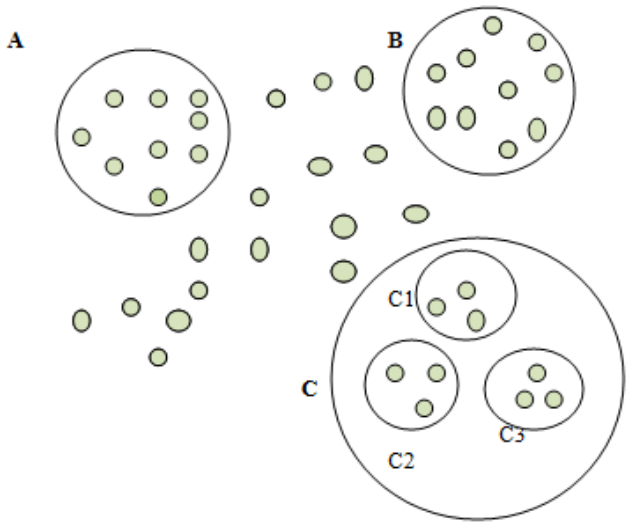

Fig 2

III. PROPOSED WORK

The proposed work shows the amalgamation of different methods with one another like fuzzy cognitive maps along with skew divergence that produces efficient results. Since fuzzy cognitive maps has the capability to check objects from various clusters if they can be a part of other clusters or not also it lets those object to be held by other clusters.

Fuzzy cognitive maps also provide a pass towards the core of the cluster depending upon the distance of the object and the core of the cluster. If the object is near to the core of the cluster then it will be given pass according to its distance. Therefore, fuzzy cognitive maps are used more than other methods in various concepts.

Here we have used the concept of FCM, basic clustering and a divergence technique known as skew where all these concepts are used to look for the similarity of objects in clusters. Also this FCM method permits some objects to belong to one or more clustering sets. This flexibility makes this algorithm effective and efficient. Therefore, it allows the objects of clusters to have a membership access depending upon its distance from the center part of a cluster. Which clearly means that addition the membership of all object are same which is determined by

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}{}^{m} \;\; ||x_i - c_j||^2$$

Where, m is could be any real number greater than 1, N is the number of data, C is the number of cluster, u_{ij} is the degree of membership of the objects, x_i is the i^{th} degree of the membership and c_j is the dimensional cluster center.

The FCM algorithm starts with choosing the number of clusters to initialize u_{ij}, once u_{ij} is initialized the minimum u_{ij} is tried to achieve using:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}}$$

Now, let's discuss about skew divergence where the distance between data point and cluster centre is calculated using:

SD (q,r) = D (r ||∝ q + (1 - ∝) r)
= D (r || z)
$= r * \frac{\log r}{\log z}$

And,
z = $\propto + q[(1-\propto)r]$

Thus using the above formula skew divergence can be easily calculated making the output to generate quickly but there is one method that over comes the previous problem of time consumption that technique is Fast Gauss Transformation where it clusters uncertain data and generates the output in no time making this method easy and quick for clustering.

IV. EXPERIMENTAL RESULT

In this paper the method used for clustering uncertain data is Fast Gauss Transformation. This is the fastest algorithm used in clustering the unpredictable data. As the previous methodology states different use of clustering techniques where each technique has a unique way to cluster data and all those methods give appropriate output but in some specific conditions. Those conditions sometimes work and sometimes doesn't. Not every methodology can work on unpredictable data some might take time to give output but some may just waste time without generating any output. Therefore, this paper introduces a new way of clustering uncertain data that takes less time and also is effective and reliable. This Gauss transformation has been proven to be the best of all the previous methodology. The formula for Gauss Transformation is:

$$G(y_j) = \sum_{i=0}^{n} q_i \; e^{-||y_j - x_i||^2}/h^2$$

Where, q_i is the weight coefficient, x_i is the source point of the Gaussians whereas h is the bandwidth parameter of Gaussians. By using this formula and its algorithm the fast gauss transformation has prove to be the fastest of all the previous methodologies. The table below shows the result when fast gauss transformation being compared with previous method of clustering uncertain data.

Serial no.	No .of cluster	FCM skew time	FCM fast gauss time
1	2	44,238	31,986
2	3	30,456	18,443
3	4	23,213	11,223
4	5	32,643	20,111

Table 1:- Time Taken to Cluster Uncertain Data

The graph below shows the actual comparison of FCM with skew divergence and FCM with fast gauss transformation where the clustering is done as per the centroids taken while calculating the PDF and PMF of the given or provided data, stating that fast gauss transformation is much better than skew divergence.

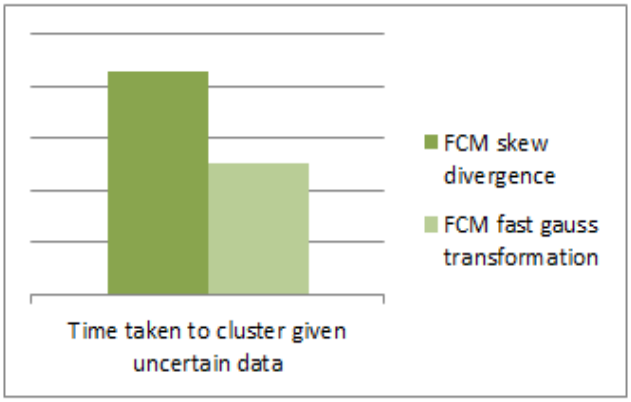

Fig 3:- Bar Graph Showing More and Less Time Consumption

V. CONCLUSION

Since I have gone through various methods of clustering uncertain data in different papers by which the conclusion has been made that the methods used in this paper is more efficient and convenient. The technique used in this paper is effortless also the desired output will be seen in no time.

ACKNOWLEDGMENT

I wish to acknowledge, Prof. Prasanna Lohe (Computer Science Engineering Department, JD college of Engineering & Management, Nagpur) for doing great job in helping and guiding me well, where I have learned a lot about this project through his amazing knowledge about technology and with his support this project has become a huge success.

REFERENCES

[1]. Alexander Topchy, BehrouzMinael-Bidgoli, Anil K Jain, William F Punch, proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. 1, 272-275,2004.
[2]. Clustering Uncertain Data using UK-Means, Aakanksha Dixit (aakankshadixit@csitdurg.in), CSIT, Durg. AbbishekMisal (abhishekmisal@csitdurg.in), CSIT, Durg.
[3]. Temperature influences on intense UK hourly precipitation and dependency on large-scale circulation. S Blenkinshop, SC Chan, EJ Kendon, NM Roberts, HJ Fowler. Environmental Research Letters 10 (5), 054021,2015.
[4]. Clustering on Uncertain Data using KullbackLeibler Divergence Measurement based on Probability Distribution. JSCOE, Department of Computer Engineering, Pune University, Hadpsar Pune. Accepted 27 July 2015, Available online 28 July 2015, Vol.5, No.4 (Aug 2015)
[5]. Bin Jian and Jian Pei Simon Fraser University, Burnaby, Yufei Tao, Chinese University of Hong Kong, Hong Kong, Xuemin Lin, The University of New South Wales, Sydney and East China Normal University, China.
[6]. Density-based clustering of uncertain data, Hans-Peter Krigel, Martin Pfeifle. Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 672-677, 2005,
[7]. Uncertain data cluster based on DBSCAN, Donghua Pan and Lilei Zhao, Dalian University of Technology, Liaoning, 116024, China.
[8]. Hierarchical density-based clustering of uncertain data, H-P Kriegel, Martin Pfeifle, Fifth IEEE International Conference on Data Mining (ICDM'05), 4 p.p., 2005.
[9]. Schneider, Johannes; Vlachos, Michail (2013). "Fast parameter less density-based clustering via random projections". 22nd ACM International Conference on Information and Knowledge Management (CIKM).
[10]. An Efficient Methodology for Clustering Uncertain Data Based on Similarity Measure by Manisha Padole and Sonali Bodkhe, department of computer science and engineering, volume 18, issue , Ver. V (Jul.-Aug. 2016).