Clustering Process for Mixed Dataset Using Shortest Path Non Parameterised Technique

Dr. V. Kavitha
Department of MCA, Hindusthan College of Arts and Science,
Coimbatore, India,
kavithahicas@gmail.com

R. Annamalai Saravanan
Department of Computer Applications, Nehru College of Arts and Science,
Coimbatore, India
malai24003@gmail.com

Abstract**—**Clustering in mixed dataset is a dynamic research focus in data mining concepts. The predictable clustering algorithm related to be more supportive to only one kind of attribute not for the mixed data type. Hence, the traditional clustering techniques processed with mixed attributes either by converting the numerical data type to categorical type or categorical type to numerical data type. But, utmost of the clustering processes are improved by converting numerical attributes. This progression of grouping ends up with two boundaries, the earlier limitation is that conveying numerical values to all types of categorical data is simply difficult. On the other hand the later drawback lies in the parameterized clustering which needs number of clusters as response for grouping the datasets. To succeed over the limitations the clustering technique is organised by incorporating shortest path and non-parameterized clustering. The proposed work of Shortest path non parameterised Clustering technique, the input parameter (number of clusters) is discovered spontaneously and the data objects of the cluster are grouped that are at the shortest distance.

Keywords*-Clustering, Mixed data type, Similarity Measure*

*****

I. INTRODUCTION

Clustering is the process where similar data objects are stimulated into clusters. Generally, the dataset is mingled with the kind of both categorical and numerical data types. Clustering concept is not able to function in the type of mixed data type properly. Hence the mixed data type needs to separate the individual data types. With the support of split function, the mixed data type is separated as categorical and numerical data for finding the similarity. Then, the active clustering algorithm is invented that produces the sub groups established on the similarity. Hence, similarity measure is utilized to estimate the similar data objects in the clusters. Herewith, the similarity plays a vital role for shaping the cluster quality. Figure 1 illustrates the process of clustering techniques with mixed data types. Generally data set is the combination of numerical and categorical data type that is separated with individual data sets using any effective clustering technique.

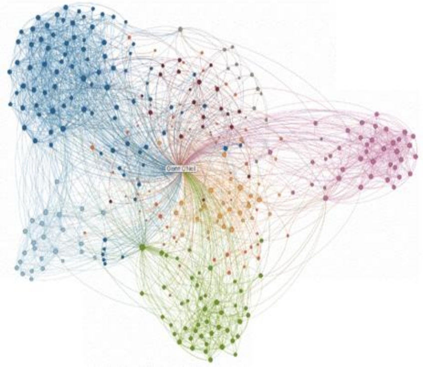

Figure 1 : Clustering process with various mixed data type objects

Normally the mixed type of data attribute is very huge and high dimensional. The incredible evolution in the real world applications like bio informatics, medical informatics, geo informatics, banking, networks and information retrieval produces massive data that constitutes mixed type attributes. Now a days extracting the hidden information from these dataset is the major obstacle. Hence, the clustering process needs to divide the mixed data set into various meaningful groups based on the similarity measure like the objects in the same group has high degree of similarity.

Figure 2 Shortest path non parameterised Clustering Technique with Mixed Data type provides the clear idea about the clustering process which progressing with the help of processing automated cluster points and find out the cluster shortest path. The essential core of the research work is the dataset and the clustering techniques that are adopted to enhance the efficiency and effectiveness of the research performance. The mixed data set is applied in this research work. As already mention that the mixed dataset is referred as the dataset that embraces both the numeric and categorical data. To handle the mixed dataset and to obtain the unknown knowledge from the extensive data the clustering technique is adopted. The clustering technique is used for grouping the mixed nature of data objects in such a way that the objects of the same clusters are similar to each other whereas the data objects belonging to various clusters are dissimilar to each other. Generally, the clustering process initiates with the user input of number of clusters as input parameter which slows down the clustering performance. To enrich the grouping

performance, the research work proceeds with Non-parameterized clustering**.** This is referred as the process of clustering performed without any user intervention.

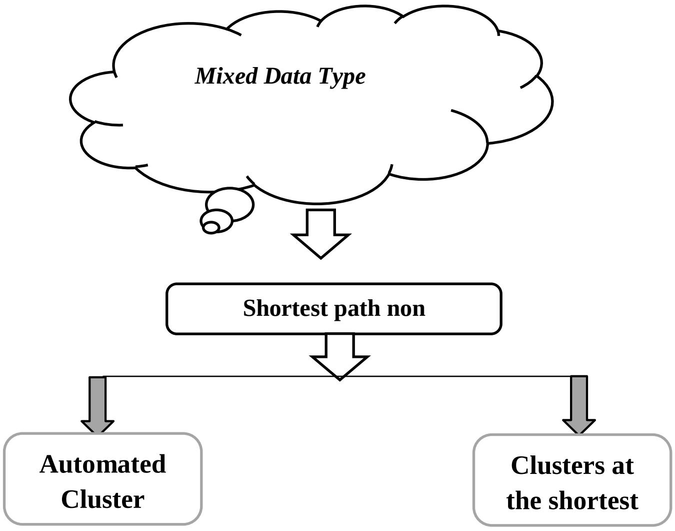

Figure 2 Shortest path non parameterised Clustering Technique with Mixed Data Type

In Clustering the mixed data objects are grouped based on the distance of each and every data objects. When the data objects that are at the minimal distance are grouped into one single cluster then it is referred as shortest-path clustering and it is accomplished in this researchworkto improve the effectiveness.

II. LITERATURE SURVEY

(Yihong Dong and Yueting.Zhuang, 2004) stated that new clustering algorithm which is an association of fuzzy and hierarchical clustering. With the assistance of partitioning clustering method the clusters are divided into numerous sub clusters. That sub clusters are constructed with the help of linked fuzzy degree. A cut graph of linked fuzzy degree is used to connect the fuzzy graph components for managing the high dimensional dataset and also it is used to perform the cluster with arbitrary shape of the clusters. This algorithm is handled with the mixed nature of numeric attributes and categorical attributes. The outcome of investigational analysis with the mixed nature of data set are proved and encouraged with the arbitrary size and shape. The resulting performance is demonstrated with the web log files that are used to determine the user access patterns efficiently. Moreover, this algorithm produces best quality clusters than established clustering algorithms and it performs the large and high dimensional databases.

(Pedro Pereira Rodrigues, et al, 2008) stated the innovative clustering algorithm named as ODAC- Online Divisive Agglomerative Clustering. Basically, ODAC clustering technique is based on the traditional hierarchical clustering technique. Hierarchical clustering technique s results in tree like clusters in which small clusters of data objects that are discovered to be strongly same to each data objects are nested together within the huge clusters that enclose minimum similar data objects.

Hierarchical clustering technique is broadly divided into agglomerative and divisive. The output of a hierarchical clustering technique is produced in the form of dendrogram. It is usually symbolized in the tree with numeric levels associated to its branches. The numeric value indicates the similarity levels of the cluster formation. At any level of similarity a line is draw to the similarity axis. In this manner, the tree's each branches are developed with the sub tree rooted at each branch. All the data objects are assembled in the single cluster when the level is in lower level.

Mostly, hierarchical agglomerative clustering technique operates on the approach of stored matrix. Hence, agglomerative hierarchical clustering groups N data objects by utilizing the stored matrix approach. Practically, the hierarchical clustering technique is belongs to the linkage methods namely single linkage method, complete linkage method and group average method.

This hierarchical clustering algorithm continuously monitors a hierarchical tree structure matrix of clusters that progresses with data with the help of top down strategy. The splitting principle is a dissimilarity measure which is based on correlation. The ODAC clustering algorithm also uses a merging process that adopted in response of changing the existing clusters diameter. This clustering algorithm is designed to develop the fast flow of data streams at high rate. The main advantage of this clustering algorithm is it does not depend on the number of instances in the data set, memory usage and update time. Moreover, the time and memory required to process an example decreases whenever the cluster structure expands. Experimental results on artificial and real data assess the processing qualities of the system, suggesting a competitive performance on clustering streaming time series, exploring also its ability to deal with concept drift.

III.PROPOSED CLUSTERING TECHNIQUE OF SHORTEST PATH NON PARAMETERISED ALGORITHM FOR MIXED DATA SET

Clustering in mixed dataset object is measured as a dynamic research in modern years. Most of the background learning related to clustering states that the clustering process is much suitable for numerical data objects rather than categorical data objects. Naturally, the mixed data is a mixture of both categorical and numerical. This inadequacy appears the necessity for the enrichment in the grouping process. Thereby, the proposed shortest path non parameterised algorithm framework is focused and the grouping process is calculated to handle the mixed- type dataset powerfully. The proposed design of the research work is illustrated in figure 3. The framework of the proposed work is in the form of two folds. When the primary fold is nominated to automate the grouping process by identifying the number of clusters spontaneously and the secondary fold deals with clustering the similar objects that are at the shortest distance.

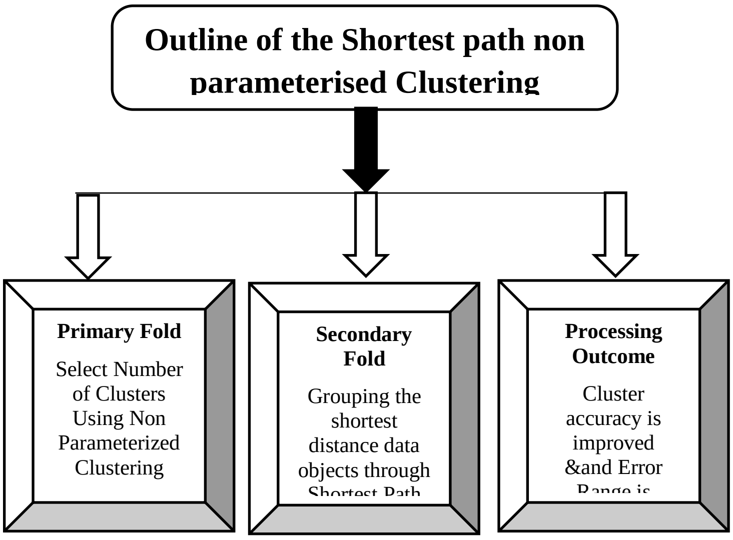
Outline of the Shortest path non parameterised Clustering
Primary Fold
Select Number of Clusters Using Non Parameterized Clustering
Secondary Fold
Grouping the shortest distance data objects through Shortest Path
Processing Outcome
Cluster accuracy is improved &and Error Range is

Figure 3 Outline Design of the Shortest path non parameterised Clustering

Processing outcome of the shortest path non parameterised clustering algorithm
enrich the clustering performance of both the accuracy and computation speed of the clustering process.

Algorithm for Shortest path non parameterised clustering

Input: Mixed Dataset

Prcess 1: Assign the labels to n points in d dimensions such that A Ɛ R ^{nxd}

Process 2: Determine the cluster path between each pair of points min subject to

Process 3: Constrain the total area between every pair of points using the weight function $\sum_{i<j}$

Determining the number of clusters through the primary fold Non Parameterized grouping the next step is to discover the shortest cluster path for grouping. The shortest cluster path is derived through the convex objective function through the following algorithm procedure.

IV.PERFORMANCE FACTORS AND EXPERIMENTAL SETUP

Recently Medical, science, engineering and geographical fields generates more volume of mixed dataset. Naturally the mixed dataset is not belongs to any single category. It belongs to numerical and also categorical data type. The experimental analysis is performed on the iris dataset, mush room data set and adult dataset of mixed nature of data sets. The performance factor of cluster quality of accuracy and error range.

The quality of grouping process in mixed nature of data type is experimented on high dimensional iris dataset, adult data set and mushroom databases. These datasets were acquired from the UCI archive datasets. The mixed nature of iris dataset contains four attributes and the number of instances is 250. Whereas the adult dataset enfolds fourteen attributes and size of the data base is 52,780. Accordingly, the mush room dataset encloses twenty two attributes and volume of the data base is 9174.

The objective of this shortest path non parameterised clustering is to automate clusters by deriving the number of cluster automatically without any user intervention. Also, from the clusters by grouping the mixed data object that are in the shortest distance. **Results and Discussions**

Clustering of huge high-dimensional and mixed nature of data sets is a wide spread obstacles and challenging issues. Many real time scientific applications generates huge amount of high dimensional and mixed kind of dataset. The various performance measures on these datasets are as follows.

Cluster Accuracy

Cluster quality is measured in the form of cluster accuracy through the mixed data type. The nature of cluster quality is enriched by clustering technique which is measured among the parameterized clustering technique and non-parameterized clustering technique. The table 1 gives a detail explanation about the cluster quality performance factor among the parameterized and non parameterized clustering techniques.

TABLE 1
Comparison of Cluster quality in Traditional Parameterized & Non- Parameterized Clustering Techniques

Data Set	Traditional Parameterized Clustering Technique (%)	Shortest Path Non-Parameterized Clustering Technique (%)
Iris Dataset	72.34	*89.53*
Adult Dataset	71.12	*90.23*
Mushroom Dataset	69.29	*72.56*

Table 1 shows the well performance of the proposed grouping technique of shortest path non-parameterized clustering technique towards the existing and traditional grouping technique of parameterized clustering technique through the cluster quality performance factor. These two techniques are compared with the support of three kinds of databases namely iris dataset, adult dataset and mushroom dataset. Among the three datasets the proposed algorithm obtains higher quality in the Iris dataset than the other mixed data sets. The reason for the performance variation is the number of attributes increased the performance of clustering technique is reduced.

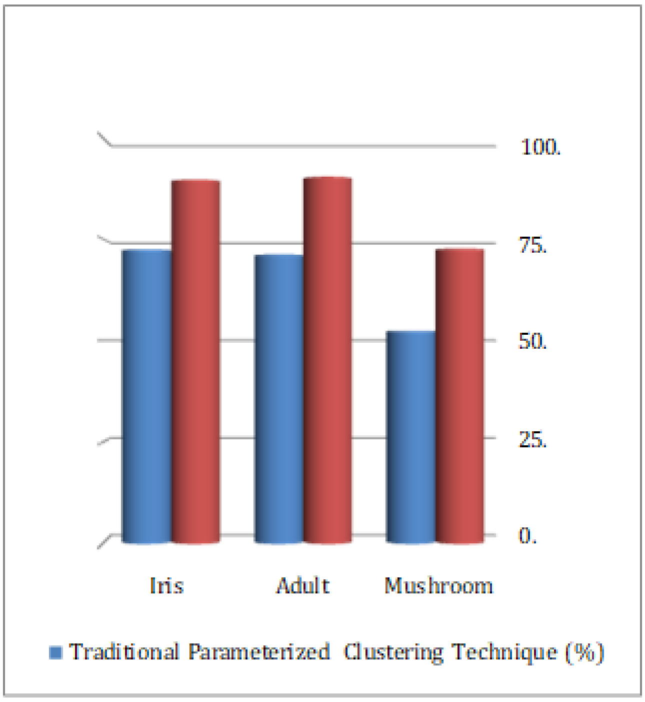

Figure 4
Cluster Quality of various Mixed Datasets

The Figure 4 depicts the comparison between the existing and the innovative new clustering technique of shortest path non parameterised clustering in terms of cluster quality. The quality of the datasets such as iris dataset, adult dataset and mushroom dataset are shown in the form of cluster quality in percentage. The x-axis of the figure shows the various mixed datasets of iris, adult and mushroom data set and y-axis of the figure shows the cluster accuracy in percentage that varies from 0 to 100%.

Error range

Clustering performance factor of error range is practically measured through the proposed non parameterized shortest pat clustering technique of the research. Table 2 provides a proper description about the performance factor of error range among tradition clustering technique with specified number of input parameters and shortest path non parameterised clustering technique, which clearly obtain that the least error range of the proposed clustering technique. Moreover theses clustering techniques are experimented with the mixed nature of data sets namely iris, adult and mush room data set.

TABLE 2

Comparison of error range-Parameterized Clustering vs Shortest Path Non-Parameterized Clustering

Data Set	Traditional Parameterized Clustering (PM) Technique (0-1 range)	Shortest Path Non-Parameterized Clustering (SPNP) Technique (0-1 range)
Iris Dataset	0.9	0.5
Adult Dataset	0.9	0.7
Mushroom Dataset	1	0.9

Figure 5 depicts the comparison among the existing and proposed clustering techniques in terms of the clustering performance factor error range. The x-axis of the figure shows the various mixed nature of datasets and y-axis of the figure shows the clustering performance factor of error range from 0-1. The least performance factor of error range is considered in the proposed clustering technique compared with the existing and traditional clustering system.

The Error range is minimal in the proposed clustering technique of shortest path non parameterised clustering Ultimately, the clusters are formulated from iris dataset, the existing clustering technique obtain the Error range is 0.9 to generate the clusters whereas the proposed clustering technique obtain the Error range is 0.5. Error range is reduced in the proposed clustering technique as 0.4. The minimal Error range is obtained by the proposed non-parameterized shortest path clustering technique than the existing clustering technique.

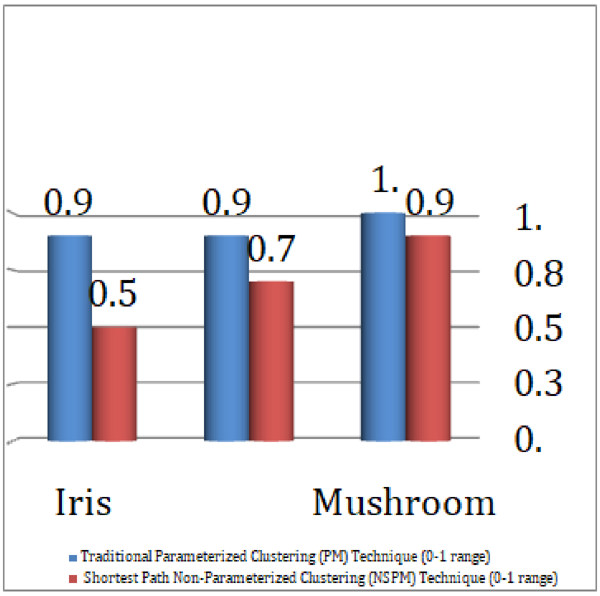

Figure 5
Error Range of mixed various Datasets

Likely, the clusters are devised from adult dataset, the existing clustering technique obtains maximal Error range of 0.9 and the proposed clustering technique obtains minimal Error range of 0.7. Comparing above two clustering techniques the performance factor of Error range is minimized in the proposed clustering technique as 0.2. The minimum Error range is obtained through the proposed non-parameterized shortest path clustering technique than the existing clustering technique.

Similarly, the clusters are formulated from mushroom dataset, the existing clustering technique acquires maximum Error range of 1 and the proposed clustering technique obtains minimal Error range of 0.9. The minimum Error range is acquired through the proposed Shortest Path non-parameterized shortest path clustering technique than existing parameterized clustering technique. Hence, the Error range deviations conforms that the optimal performance with the assistance of three kinds of datasets.

V.CONCLUSION

Clustering in mixed dataset is a dynamic research focus in data mining concepts. The traditional clustering techniques processed with mixed attributes either by converting the numerical data type to categorical type or categorical type to numerical data type. But, utmost of the clustering processes are improved by converting numerical attributes. This progression of grouping ends up with two boundaries, the earlier limitation is that conveying numerical values to all types of categorical data is simply difficult. On the other hand the later drawback lies in the parameterized clustering which needs number of clusters as response for grouping the datasets. To succeed over the limitations the clustering technique is organised by incorporating shortest path and non-parameterized clustering. The proposed work of Shortest path non parameterised Clustering technique, the input parameter (number of clusters) is discovered spontaneously and the data objects of the cluster are grouped that are at the shortest distance.This clustering technique is mainly concentrated in moving the process of clustering to automated clustering and grouping the objects by detecting the shortest cluster path. The research findings promise the optimal performance of the proposed clustering methodology than the existing clustering technique. Although, the upgradation improves the grouping process to the next level, the system has to be improved by considering the inadequacy in managing the cluster data objects that belongs to two or more clusters.

REFERENCES

[1]. Pantelis n.Karamolegkos, Charalampos Z.Patrikakis Nikolaos D.Doulamis Panagiotis, “An Evaluation Study of Clustering Algorithms in the Scope of user Communities Assessment” Computers $ Mathematics with Applications, Elsevier, Vol No 58, issue no 8, October 2009, Pages 1498 - 1519.

[2]. Man Abdel - Maksoud, Mohammed Elmogy, Rashid Al-Awadi, “Brain Tumor Segmentation Based on a Hybrid Clustering Technique”, Egyptian Informatics Journal, Vol No 16, Issue no 1, March 2005, Pages 1 - 81.

[3]. Madjid Khalilian, Norwati Mustapha, Data Stream Clustering: Challenges and Issue, Proceedings of the International Multi conference of Engineers and Computer Scientists 2010 Vol No1, IMECS 2010,March 17-19 2010.

[4]. Maryam Mousavi1 , Azuraliza Abu Bakar, and Mohammadmahdi Vakilian, “Data Stream Clustering Algorithms: A Review”, International Journal of Advance Soft Computer Applications Vol o 7, Issue No 3, November 2015, ISSN 2074-8523.

[5]. Jose R. Fernandez,” A Framework and Algorithm for Data Stream Cluster Analysis”, International Journal of Advanced Computer Science and Applications, Vol No 2, Issue No11, Pages 87, 2011.

[6]. Twinkle B Ankleshwaria, Twinkle B Ankleshwaria, Mining Data Streams: A Survey, International Journal of Advance Research in Computer Science and Management Studies, Vol No 2, Issue No 2, Feb 2014, ISSN: 2321-778.

[7]. Amineh Amini, Teh Ying Wah, “Density Micro-Clustering Algorithms on Data Streams: A Review”, Proceedings of the International MultiConference of Engineers and Computer Scientists 2011 Vol No 1, IMCES 2011, March 16-18, 2011.

[8]. Silva, J. A., Faria, E. R., Barros, R. C., Hruschka, E. R., de Carvalho, A. C. P. L. F., and Gama, J, “Data stream clustering: A survey”, ACM Computing Surveys, Vol No 46, Issue No1, Article 13, October 2013, Pages 31.

[9]. DoniaAugustine, “A Survey on Density based Micro-clustering Algorithms for Data Stream Clustering”, International Journal of Advanced Research in Computer Science and Software Engineering Research, Vol No 7, Issue No 1, January 2017.

[10]. Dure Supriya Suresh, Prof. Wadne Vinod, “Survey Paper on Clustering Data Streams Based on Shared Density between Micro-Clusters”, International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395 -0056, Vol No 04 ,Issue No 01, January 2017.

[11]. Amini A, Wah TY, Saboohi H, “On density-based data streams clustering algorithms: A survey”,Journal of Computer Science and Technology, Pages 116–141, January 2014, DOI 10.1007/s11390-013-1416-3.

[12]. Safal V Bhosale, “A Survey: Outlier Detection in Streaming Data Using Clustering Approache”, International Journal of Computer Science and Information Technologies, Vol No 5, 2014, 6050-6053 ISSN 0975 - 9646.

[13]. Prashant V. Desai, Vilas S. Gaikawad, “Novel approach for data stream clustering through micro-clusters shared Density”,International Journal of Computer Sciences and Engineering Volume-5, Issue-1 E-ISSN: 2347-2693.

[14]. M.S.B.PhridviRaj, C.V.GuruRao, “Data Mining - Past, Present and Future - A Typical Survey on Data Streams”, Elsevier Procedia Technology”, Vol No 12, 2014, Pages 255 - 263.

[15]. Yisroel Mirsky, Bracha Shapira, Lior Rokach, and Yuval Elovici, “pcStream: A Stream Clustering Algorithm for Dynamically Detecting and Managing Temporal Contexts”, Springer International Publishing Switzerland 2015, PAKDD 2015, Part II, LNAI 9078, pp. 119–133, 2015. DOI: 10.1007/978-3-319-18032-8_10.

[16]. Shufeng Gong, Yanfeng Zhang, Ge Yu1, “Clustering Stream Data by Exploring the Evolution of Density Mountain”, PVLDB, 11(4) 2017. DOI: 10.1145/3164135.3164136.