ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 8, October 2013

Coupled Shortest Fuzzy C-Means Clustering Algorithm (CS-FCM) In Mixed Dataset

Mrs. M.Punithavalli[1] Mr. A.S.Naveen Kumar[2]

Director & Head, Sri Ramakrishna Engineering College, Tamil Nadu, India[1]

Ph.D. scholar in the Dept. of Computer Science, Bharathiar University, Tamil Nadu, India[2]

ABSTRACT: Nowadays Clustering in mixed dataset is a dynamic research topic in data mining concepts. Most of the clustering process is based on numerical attributes. That processes are not suitable for mixed dataset. The nature of mixed dataset is the combination of numeric and categorical data type. Hence, the proposed technique required more efficiency to handle the mixed data set. This paper proposes a hybrid clustering technique CS-FCM that combines the concepts of hierarchical and fuzzy clustering for mixed dataset. Hence, in this paper, the proposed technique can handle the mixed data set easily. Moreover, it is coined for automated and effective clusters. This proposed technique is experimented with three type's data sets. Obviously, the experimental result shows that, they are inventive and have the capability to discover the number of clusters automatically from the mixed-dataset.

K*eywords*: Mixed Dataset, Hierarchical Clustering, Fuzzy Clustering, Fuzzy-C-Means clustering

III. I. INTRODUCTION

Ever growing data in almost of entire fields can provide foremost and significant data such as mixed data type. The nature of mixed data is the combination of categorical and numerical data sets. Traditional data mining techniques are suitable for categorical dataset or numerical dataset. To handle the mixed dataset required the efficient mining technique. The knowledge discovery mechanisms are pertained for extracting hidden knowledge and useful information embedded in the data. The knowledge discovery system employs a vital role of machine learning mechanisms to discover the relationships between the tuples. Generally, clustering and classification are two most knowledge discovery techniques that are directed to extract fruitful knowledge from the large database. Classification is based on supervised learning techniques, which is suitable for predefined datasets. On the other hand the clustering is based on unsupervised learning techniques, which is suitable for distinct dataset such as mixed data set. Hence, this paper is concentrated the clustering technique to discover the efficient cluster groups. Mostly, the clustering technique requires the input parameter referred number of clusters to be described by the user which results in parameterized clustering. The drawback of this parameterized clustering are takes more time to process, produce improper clusters and poor cluster quality. To succeed over the above said limitations this paper concentrated with the integration of non parameterized clustering and Fuzzy C-Means techniques. The association of above said techniques involved to improve the cluster quality and to speed up the clustering process.

II. LITERATURE REVIEW

CACTUS is a categorical based clustering technique. The foremost unique feature of this clustering technique is based on agglomerative hierarchical clustering that utilizes the concepts of strong connections, support and similarities to cluster categorical data. The attributes are strongly connected their support exceeds value with the assumption of attributes independence. This technique is extended with more number of attributes. The clustering process is referred as a region of attributes that are strongly connected with the property of pair wise connection, and its support exceeds the attribute independence assumption which is expected. Computational complexity is the major issue of this technique and moreover, the computational cost is very high.[1]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 8, October 2013

Fuzzy Hierarchical Clustering technique is the incorporation of hierarchical clustering and fuzzy logic. Dataset was divided into various sub groups using the clustering technique, fuzzy graph of sub groups was constructed by examining the linked fuzzy degree between the sub clusters. A cut graph is used to connect the components of the fuzzy graph. This fuzzy graph logic is integrated to the traditional hierarchical clustering technique which is used to discover clusters with cluster arbitrary shape and size from high dimensional data set. And it is concluded that the experimental result shows that the better performance of the fuzzy hierarchical clustering technique than the traditional clustering techniques.

III. PROPOSED METHODOLOGY

All Clustering is an unsupervised learning technique where no pre defined classes are assigned. The key requirement for clustering is the similarity measure that exists between the patterns or instances. The main goal of clustering is to group n patterns into c desired clusters, such that the data points within clusters are more similar than across clusters. Moreover, the presence of mixed datasets requires the necessity of embedding specialised techniques with the traditional clustering. Thereby, the specialised approach called Fuzzy is coupled with clustering that ensures the degree of similarity and membership of the cluster in an efficient manner. The major difference between traditional clustering and fuzzy clustering is that the former approach generates partition wherein each pattern in each partition belongs to one and only one cluster. In the later case, each pattern is associated with every cluster using the membership function.

In this paper Fuzzy C-Means clustering is coupled with shortest path algorithm for handling the mixed type dataset. FCM is considered as the fuzzified form of the K-Means algorithm in which a piece of data fits into two or more clusters. This approach produces an optimal c partition by reducing the weighted sum of square error objective function $J_{fcm.}$

$$J_{fcm}(V, U, X) = \sum_{i=1}^{c} \sum_{i=1}^{n} u_{ij}$$

The CS-FCM is an enhanced method that is elevated to cluster the data of varying shape, size and density. The unique contribution of this approach is achieved by embedding a method called Fuzzy Maximum Likelihood Estimates (FMLE) that calculates norms and weights of the data for efficient clustering performance. The Maximum Likelihood data is derived through Co-variance matrix and conjunction function. The method chooses the data points based on two factors. The former is related with finding the maximum similarities or weights through fuzzy learning that brings out the optimal data points for clustering. The later is the exponential distance metric used for cluster formation. The formation of the cluster is acquired by coupling shortest path algorithm with Fuzzy C-Means. Furthermore, Low-rank approximation method is used to increase the quality of the cluster and also to detect the quantization of error. Therefore, this method is merely used for finding the appropriate data points without any data reduction instead remarkable data are selected and grouped with the help of Fuzzy C-Means clustering. Indeed the proposed method is a unique initiative for sweeping the quality of the cluster.

In the given fuzzy set x the membership function and the likelihood is defined as

$$L(\dot{x}, x) = P(x, \dot{x}) = \int x \, \mu_{\dot{x}}(x) g(x, \dot{x}) \quad \text{.....} \quad 1.1$$

The above equation denotes the probability of a fuzzy set. The variance of the set is denoted by

$$\mu_{\tilde{x}}(x) = \prod_{i=1}^{n} \mu_{\tilde{x}_i}(x_i) \quad \text{.....} \quad 1.2$$

The shortest Path is calculated using the Euclidean distance measures which outreaches with the shortest distance between two points.

IV. RESULTS AND DISCUSSION

The clustering progression is measured through the investigational results and discussions. The Clustering quality process in mixed data is investigated on high dimensional iris dataset, adult dataset and mushroom dataset. These datasets were taken from the UCI archive repository (http://www.sgi.com/tech/mlc/db). The above said datasets

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 8, October 2013

descriptions are explained in the following. Table 1 illustrates about the dataset description with various datasets, number of instances and number of attributes.

TABLE 1

Dataset Description

Data Set	Number of Attributes	Number of Instances
Iris Dataset	Four	150
Adult Dataset	Fourteen	48842
Mushroom Dataset	Twenty Two	8124

The performance of Coupled shortest Fuzzy C-Means Clustering Algorithm (CS-FCM) is measured with the evaluation factor namely cluster accuracy for proving the elevated cluster quality. Herewith, the outcoming performance factors are compared uniquely through the proposed technique of CS-FCM and Non-Parameterized Shortest Path algorithm are sketched below in Figures. The proposed CS-FCM is compared with the existing Non-Parameterized Shortest Path Clustering algorithm. The performance factor of cluster accuracy is measured among the above two techniques that derives the cluster quality. Table 2 explains about the comparison among the proposed and existing clustering techniques. These comparisons are illustrated in the figure 1.

TABLE 2
Accuracy comparison of Existing NSPM towards CS-FCM

Data Set	Non-Parameterized Shortest Path Clustering (NSPM)	Coupled Shortest Fuzzy C-Means Clustering (CS-FCM)
Iris	93.53	**95.34**
Adult	90.23	**92.29**
Mushroom	88.56	**90.98**

Figure 1shows the accuracy of proposed and existing clustering techniques that derives the cluster quality of iris, Adult and mushroom datasets. The CS-FCM gains 95.34%, in iris dataset, 92.29 % in Adult dataset and 90.98 in mushroom dataset. The accuracy percentage is decreased by dataset to dataset because of the number of attributes are varied in each dataset. It has accomplished that the Coupled Shortest Fuzzy C-Means Clustering(CS-FCM) technique adopts higher accuracy compared with existing Non-Parameterized Shortest Path Clustering (NSPM) technique.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 8, October 2013

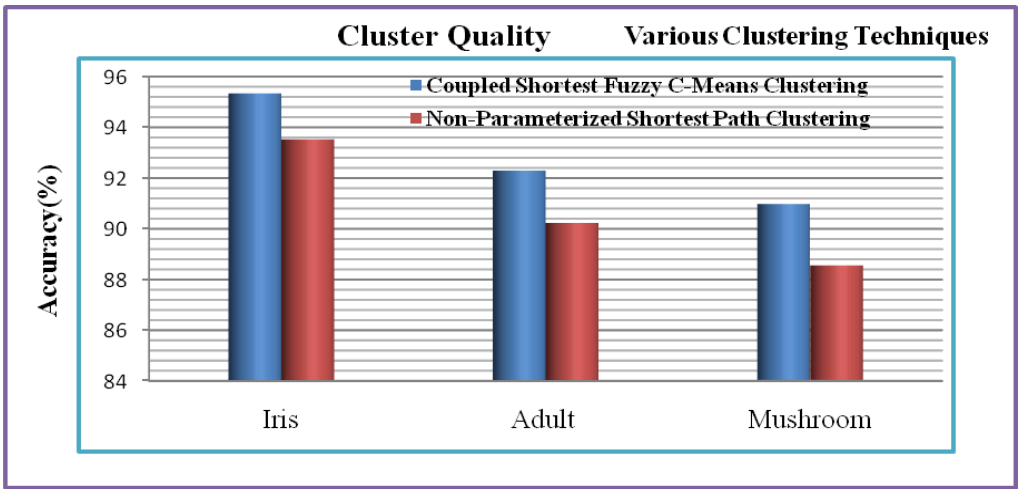

Figure 1. Comparing Proposed and Existing Clustering Techniques

V. CONCLUSION

Nowadays Clustering in mixed dataset is a dynamic research topic in data mining concepts. Most of the clustering process is based on numerical attributes. That processes are not suitable for mixed dataset. The nature of mixed dataset is the combination of numeric and categorical data type. Hence, the proposed technique required more efficiency to handle the mixed data set. This paper proposes a hybrid clustering technique CS-FCM that combines the concepts of hierarchical and fuzzy clustering for mixed dataset. Hence, in this paper, the proposed technique can handle the mixed data set easily. Moreover, it is coined for automated and effective clusters. This proposed technique is experimented with three types data sets namely iris dataset, Adult dataset and mushroom dataset. Obviously, the experimental result shows that, they are inventive and have the capability to discover the number of clusters automatically from the mixed-dataset. Moreover, the cluster quality improves in the CS-FCM than the NSPM clustering technique.

REFERENCES

[1]Amir Ahmad, Lipika Dey, A k-mean clustering algorithm for mixed numeric and categorical data, Elsevier Data and Knowledge Engineering. April 2007.

[2] Yihong Dong, Yueting.Zhuang, Fuzzy Hierarchical Clustering Algorithm Facing Large Databases, Proceedings of the 5th World Congress on Intelligent Control and Automation, June 15-19,2004, Hangzhou, P.R. China.

[3]A. Colubi. Statistical inference about the means of fuzzy random variables: Applications to the analysis of fuzzy- and real-valued data Fuzzy Sets and Systems,160(3):344–356, 2009.

[4] E. C˘ome, L. Oukhellou, T. Denoeux, and P. Aknin. Learning from partially supervised data using mixture models and belief functions. Pattern Recognition, 42(3) : 334– 348, 2009.

[5] R. Coppi. Management of uncertainty in statistical reasoning: The case of regression.International Journal of Approximate Reasoning, 47(3):284–305, 2008.

[6] R. Coppi, M. A. Gil, and H. A. L. Kiers. The fuzzy approach to statistical analysis. Computational Statistics & Data Analysis, 51(1):1–14, 2006.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, B39:1–38, 1977.

[8] T. Denoeux. Maximum likelihood from evidential data: an extension of the EM algorithm. In C. Borgelt et al., editor, Combining soft computing and statistical methods in data analysis (Proceedings of SMPS 2010), Advances in Intelligent and Soft Computing, pages 181– 188, Oviedeo, Spain, 2010. Springer.

[9] T. Denoeux, M. Masson, and P.-A. H´ebert. Nonparametric rank-based statistics and significance tests for fuzzy data. Fuzzy Sets and Systems, 153:1–28, 2005.

[10] T. Denoeux and M.-H. Masson. Principal component analysis of fuzzy data using autoassociative neural networks. IEEE Transactions on Fuzzy Systems,12(3):336–349, 2004.

[11] D. Dubois, W. Ostasiewicz, and H. Prade. Fuzzy sets: History and basic notions.In D. Dubois and H. Prade, editors, Fundamentals of Fuzzy sets, pages 21–124.Kluwer Academic Publishers, Boston, 2000.

[12] D. Dubois and H. Prade. Possibility Theory: An approach to computerized pro-cessing of uncertainty. Plenum Press, New-York, 1988.

[13] P. D'Urso and P. Giordani. A weighted fuzzy c-means clustering model for fuzzy data. Computational Statistics and Data Analysis, 50(6):1496–1523, 2006.

[14] J. Gebhardt, M. A. Gil, and R. Kruse. Fuzzy set-theoretic methods in statistics.In R. Slowinski, editor, Fuzzy sets in decision analysis, operations research and statistics, pages 311–347. Kluwer Academic Publishers, Boston, 1998.

[15] M. A. Gil and M. R. Casals. An operative extension of the likelihood ratio test from fuzzy data. Statistical Papers, 29:191–203, 1988.