International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 10, October - 2013

Shortest Path Algorithm for Automated Clustering in Mixed Dataset

A.S.Naveen Kumar, Dr.M.Punithavalli
SNR & Sons College, Coimbatore, Tamilnadu.
Sri Ramakrishna Engineering College, Coimbatore, Tamilnadu.

Abstract

In recent years the efficient and automated clustering in mixed dataset (combination of categorical and numerical data) has raised the interest among numerous researchers of various fields. Automated clustering is generally referred as a process of finding the number of cluster sets automatically without any user intervention. Moreover, the grouping of clusters in automated clustering should choose the data objects that are both similar and are at the shortest distance. Hence, in this paper a Non-parameterized shortest path algorithm is coined for automated and efficient clustering. Finally, the results are elegant and have the ability to detect the number of clusters automatically from the mixed-dataset.

Keywords: Mixed dataset, Clustering, Non-Parameterized, shortest path algorithm.

1. Introduction

Ethically, the clustering algorithm handles three categories of datasets namely categorical, numerical and mixed datasets. The mixed dataset contains both numerical and categorical which are sparsely imposed in recent datasets. Handling these variant data types through a single algorithm is a crucial task. Since, the ability of the algorithm handles either numerical or categorical data. Moreover, the size of the datasets varies from one another. To gain knowledge from these dataset backgrounds, clustering is the key technique widely used to handle the complications. Usually, the clustering algorithm requires the input parameter termed "number of clusters" to be defined by the users which results in parameterized clustering. The limitations behind the parameterized clustering are i) Lack of performance ii) Time Complexity iii) in-efficient clustering. To prevail over the above said problems this paper outreach with a Non-parameterized technique with shortest path algorithm.

2. Related Work

In data mining applications, clustering is widely used to find patterns in datasets. Traditional clustering techniques mainly focus on a single type of attributes, either numerical or categorical attributes of datasets. In clustering process, a criterion function called similarity measure is used as one of the essential steps, i.e., in determining the candidate cluster number. The unique characteristics of categorical attributes are that the values of categorical attributes are not only discontinuous but also disordered while the values of numerical attributes are continuous in computing the distance between two values. Due to the difference of the characteristics between categorical and numerical attributes, similarity measures for categorical attributes or numerical attributes have focused on just their own characteristics, for example, entropy based similarity measure for categorical attributes and distance measure for numerical attributes.

Previously, most algorithms focus on numerical data in which geometric properties that inherent are exploited in order to define the data points naturally. The algorithms such as DBSCAN, BIRTH, CURE, CHAMELEON, and Wave cluster [2] [3] [4] [5] [6] [7] are adopted for clustering numerical data which are not appropriate for categorical attributes and mixed type attributes. Also, [8-14] paper proposes few algorithms for clustering categorical data but the algorithm for extended mixed-type attribute is unknown.

Some researchers [15] proposed K-modes and K-prototype algorithms that separate the categorical and mixed type attribute based on the total matches and weighted sum of Euclidean distance for numeric attributes. But, however the results may be biased due to improper calculation of weights. In [17] the distance measure is derived from the probabilistic model in which the distance between two clusters are equivalent to log-likelihood function as a result of merging. The Birth algorithm is also used for clustering. As a final on analyzing the background it is clear that the existing algorithms are capable of working efficiently to cluster either numerical or categorical data. Moreover, the user defined input (number of clusters) hinders the efficiency of clustering.

Therefore, a non-parameterized technique with shortest path algorithm is proposed to enhance the performance of the clustering process while handling the mixed type attribute dataset.

3. Proposed Methodology

Traditionally, the clustering process requires number of cluster points as input parameters to outreach with fruitful clusters. But, this input is insufficient to acquire the quality clusters. Therefore, the proposed work in this paper is engaged with automatically detecting the number of clusters [Non-parameterized input] for gaining the quality clusters. The Framework of the proposed Non-parameterized Shortest path algorithm is illustrated in the figure 1.

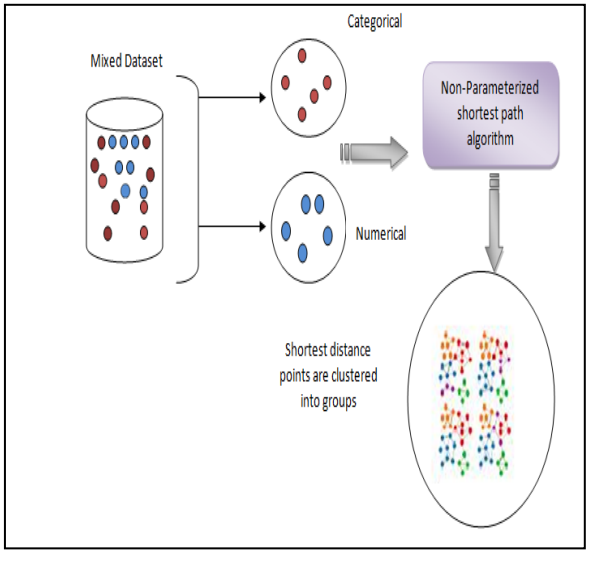

Figure 1 Work Procedure of NSPM

The foremost data mining technique of clustering is used to discover patterns in database. Traditional clustering techniques have been concentrated on a single type of attributes, either categorical attributes or numerical attributes of datasets. The nature of numerical attributes is continuous whereas the categorical attributes are discontinuous and disordered in estimating the distance among two data points. The clustering process is based on the similarity measure. Due to the various salient features of numerical and categorical attributes (Mixed datasets), the efficient distance metric is required. Hence, the shortest path distance method is used to measure the distance between the data points.

Herewith, the Non-Parameterized shortest path algorithm (NSPM) provides a new convex objective function for clustering mixed data attributes. The function is defined for every pair of data points within the cluster. The convex objective function is used to attain the optimal cluster points without manual input parameter. The parameter such as number of cluster is usually defined by the user to find the best clusters whereas, in NSPM the weight of the points are calculated and the number of the cluster is chosen by the algorithm automatically. Obviously the cluster point is chosen on the basis such that the point is both locally minimum and globally minimum.

Further, pairs are grouped on the account of distance where the points that are nearer to the centroid are grouped into clusters. Therefore, from the high-dimensional mixed dataset, the data that are nearer to the solution are grouped into categorical and numerical which further leads to meaningful cluster formation. Moreover, this method also reduces the dimensionality by grouping the data at the shortest distance. Hence, NSPM is responsible for both dimensionality reduction and efficient clustering performance. The work progression of the NSPM technique is,

Input: Data D $\in R^{nxp}$, Weight $WT_{i,j} > 0$, starting $\lambda >$
Output: Clusters
$D \rightarrow \alpha$
clusters = {{1},…..,{n}}
While clusters > 1 do
α, clusters $\rightarrow$ Solve L2(α, clusters, D, WT, λ)
$\lambda = \lambda + 1.5$
If cluster split is decided then clusters = {{1},…..,{n}}
End if
End while
Return with optimal λ values

The above NSPM technique is based on traditional hierarchical clustering algorithm which is initiated with the given data value(D), weight value(WT) and starting cluster event(ƛ). Generally, the hierarchical clustering technique is used to construct a tree shaped hierarchy structure (Dendrogram). This hierarchical clustering technique is divided into two types of approaches namely Divisive (Splitting Process) and Agglomerative (Merging Process). The NSPM technique is based on splitting process (Divisive approach). It is also referred as top down strategy, initiated with all data objects placed in one cluster and it divided the cluster into smaller clusters, until each data object forms a cluster on its own such as the diameter of each cluster id within a specific threshold value is reached. Diameter is termed as a maximum distance between the two data objects. Likewise, the above said technique is performed with a splitting process. Finally, the ƛ number of clusters is generated.

4. Results and Discussions

The clustering performance is measured through the experimental results and discussions. The quality of clustering 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 are explained in the following. Table 1 illustrates about the dataset description with various datasets, number of attributes and number of instances.

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 Non-Parameterized shortest path clustering is measured with the performance factor namely cluster accuracy for proving the elevated cluster quality. Herewith, the outcoming performance factors are compared uniquely through the proposed technique of Non-Parameterized Shortest path algorithm and sketched below in Figures. The proposed Non-Parameterized Shortest Path algorithm is compared with the existing Parameterized Clustering Technique. 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 2.

TABLE 2 Comparing Proposed and Existing Clustering Techniques

Data Set	Non-Parameterized Shortest Path	Parameterized Clustering Technique
Iris Dataset	93.53	89.34
Adult Dataset	90.23	85.56
Mushroom Dataset	88.56	82.29

FIGURE 2 Comparing Proposed and Existing Clustering Techniques

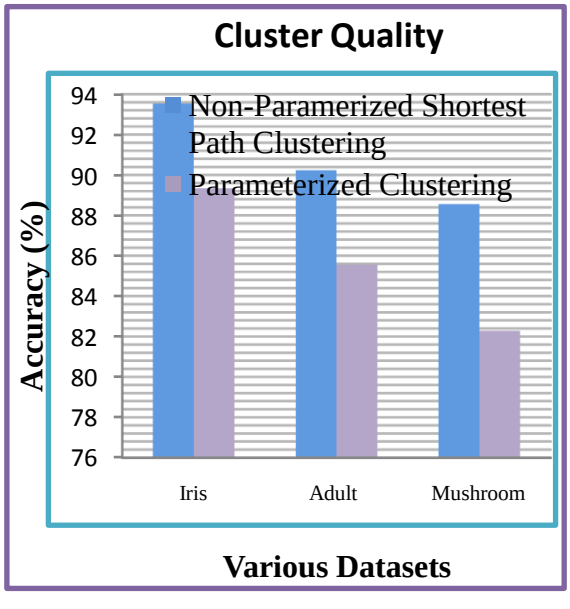

Figure 2 shows the accuracy of proposed and existing clustering techniques that derives the cluster quality of iris, Adult and mushroom datasets. The

NSPM gains 93.53%, in iris dataset, 90.23 % in Adult dataset and 88.56 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 concluded from the figure 2 that Non-Parameterized Shortest path technique adopts higher accuracy compared with existing techniques.

5. Conclusion

In recent years the efficient and automated clustering in mixed dataset (combination of categorical and numerical data) has raised the interest among numerous researchers of various fields. Automated clustering is generally referred as a process of finding the number of cluster sets automatically without any user intervention. Moreover, the grouping of clusters in automated clustering should choose the data objects that are both similar and are at the shortest distance. Hence, in this paper a Non-parameterized shortest path algorithm (NSPM) is coined for automated and efficient clustering. Obviously, the performance of NSPM clustering technique is measured through the cluster quality of accuracy. Cluster quality is improved in NSPM than the existing technique of parameterized clustering technique.

6. References

[1] Yi-Hong Chu, Jen-Wei Huang, Kun-Ya Chuang, DeNian Yang, "Density Conscious Subspace Clustering for High Dimensional Data" IEEE Transactions on Knowledge and Data Engineering. Vol 22, No 1, January 2010.

[2] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, E. Simoudis, J. Han, and U. Fayyad, Eds. AAAI Press, 1996, pp. 226–231.

[3] Y. Kim, W. Street, and F. Menczer, "Feature Selection in Unsupervised Learning via Evolutionary Search," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 365-369, 2000.

[4] M. Halkidi, Y. Batistakis, and M. Varzirgiannis, "On clustering validation techniques," Journal of Intelligent Information Systems, vol. 17, no. 2-3, pp. 107–145, 2001.

[5]. Pedro Pereira Rodriguess and Joao Pedro Pedroso, "Hierarchical Clustering of Time Series Data Streams," Sudipto Guha, Adam Meyerson, Nine Mishra and Rajeev Motiwani, "Clustering Data Streams: Theory and Practice", IEEE Transactions on Knowledge and Data Engineering. Vol. 15, no. 3, pp. 515-528, May/June 2003.

[6] Ashish Singhal, and Dale E Seborg, "Clustering Multivarriate Time Series Data," Journal of Chemometrics, vol. 19, pp. 427-438, Jan 2006.

[7] Sudipto Guha, Adam Meyerson, Nine Mishra and Rajeev Motiwani, "Clustering Data Streams: Theory and Practice", IEEE Transactions on Knowledge and Data Engineering. Vol. 15, no. 3, pp. 515-528, May/June 2003.