Chapter 6
A Study on Efficient Clustering Techniques Involved in Dealing With Diverse Attribute Data

Pragathi Penikalapati
Vellore Institute of Technology, India

A. Nagaraja Rao
Vellore Institute of Technology, India

ABSTRACT

The compatibility issues among the characteristics of data involving numerical as well as categorical attributes (mixed) laid many challenges in pattern recognition field. Clustering is often used to group identical elements and to find structures out of data. However, clustering categorical data poses some notable challenges. Particularly clustering diversified (mixed) data constitute bigger challenges because of its range of attributes. Computations on such data are merely too complex to match the scales of numerical and categorical values due to its ranges and conversions. This chapter is intended to cover literature clustering algorithms in the context of mixed attribute unlabelled data. Further, this chapter will cover the types and state of the art methodologies that help in separating data by satisfying inter and intracluster similarity. This chapter further identifies challenges and Future research directions of state-of-the-art clustering algorithms with notable research gaps.

INTRODUCTION

In this web 3.0 era, information is accelerating from the Big data and with the applicability of IoT devices. Each day, information is generated from the use of Social sites to wearable devices and IOT's. For instance, millions of google searchers, thousands of YouTube video uploads, data from service-based applications (like medical, transport, logistics, education, shopping sites, etc.), tweets, comments are being generated. With this information, researchers are trying to extract the patterns that help in analyzing and understand-

DOI: 10.4018/978-1-7998-1839-7.ch006

ing data. However, this raw data cannot be analyzed using any algorithm. Often in real-time situations, data is not available with any appropriate classifications and pre-defined labels. Consequently, there is a need to develop certain models of machine learning capable of precise classification of the new data, based on certain similarities in features. This process can be accomplished by ‘Clustering’, an algorithm of unsupervised learning. In machine learning () as well as data mining, this analysis of clustering is a very important technique. The objective of clustering analysis is to segregate an ensemble of undefined objects into diverse clusters in such a way that the data objects of a specific cluster are either different or similar to the data objects of another cluster. The applications of cluster analysis are numerous including the categorization of customers, setting market targets, analysis of social networks, bioinformatics, and analysis of scientific data (Han & Kamber, 2000). Segmentation of a specific dataset into a homogeneous collection is performed by an optimization model of partitioning depicted by a cost function, in such a way that there is a similarity among the observations inside a cluster, while dissimilarity among the observations of other clusters. Input: An unlabeled training set with attribute values *D*={*observations, i*=1,…,*N*} with N objects described by d attributes where *observations*= {*Attribute*_1, *Attribute*_2,…, *Attribute_d*} ÎR^d K depicts the total number of initial clusters. Output: A set of K clusters C_1, C_2,…, C_k. The variations in size, shape, and density in the resultant clusters largely depend on the number of clusters K and the processes of clustering adopted. The prime characteristic feature of a good clustering is in its intense compactness, which means that the intra-cluster observations should be as proximate as can be possible, and isolation which means that the inter-cluster variations in observations should be as scattered as can be possible.

As illustrated in Figure 1, it can be noted that clusters C1 and C3 are different in shape but very compact, while C2 and C3 are comparatively not so compact. Certain observations are found to be secluded from the cluster’s core. Such secluded observations could be the representations of the outlier and noise in the resultant clusters and may not be causing a negative influence on the comprehension towards the close of the process (Ben Salem, Naouali, & Chtourou, 2018).

d1 is the representation of the inter-distance existing between the observations associated with diverse clusters requiring to be maximized. This results in the procurement of isolated clusters. Similarly, d2 depicts the intra-distance existing between the observations pertaining to the same cluster requiring to be minimized. This results in the procurement of compact clusters.

Accuracy in finding the clusters and the high scalability are the major constraints to be considered in the design of new approaches of clustering, more specifically in the context of larger datasets. In general, there are certain major issues requiring to be addressed in the process of clustering such as discovering suitable measure (distance) of similarity for the optimization of objective function, implementation of competent iterative steps for the discovery of the most precise clusters and the derivation of an appropriate description for the contextualization of the elements in every individual clusters and according permission for the extraction of patterns (Ben Salem et al., 2018). An objective function’s optimization can be represented as

$$Obj\_fun = \sum_{i=1}^{n} \left| Observation_j . \mathrm{Centroid}_j \right|$$

where Centroid$_j$ is the center of the jth cluster, *Observation*$_j$ is the ith object selected from Observations and ||.|| is a distance metric.