Chapter 15
Efficient Clustering Algorithms in Educational Data Mining

Anupama Chadha
Manav Rachna International Institute of Research and Studies, India

ABSTRACT

Higher education institutions are competing for excellence, and in this process, they are utilizing information technologies to gather relevant information for achieving academic excellence. The institutes are putting greater emphasis on meeting students' academic needs, enhancing the quality of service provided to students, providing better placements, course excellence, etc. The use of modern information technologies helps in storing huge data but requires the use of data mining technologies to extract useful information and knowledge from this data. Some of the knowledge achievable for higher education institutes through implementing several data mining techniques (classification, association learning, clustering, etc.) is the correlation between specialization and the chosen employment path, determining the subjects, courses, labs with high degree of difficulty, interesting subjects, courses, labs, facilities that might attract new students, etc. This chapter explores efficient clustering algorithms in educational data mining.

INTRODUCTION

Clustering is a technique of segregating the objects into partitions such that the objects in a group are more similar to each other than the objects in the other group. Clustering has its applications in variety of domains like health sector (Kaur, Harleen & Wasan, Krishan, Siri, 2006; Sharma, A. & Mansotra, V., 2014), E Commerce (Cheng, Yu & Ying, Xiong, 2009; Li, Mei & Feng, Cheng, 2010; Li, Yong-hong & Liu, Xiao-liang, 2010) etc. One of the areas where clustering is gaining boom is education (Hung, Jui-Long & Gao, Qingcheng, 2011; Hung, Jui-Long, Hsu, Yu-Chang & Rice, Kerry, 2012; Jin, Hanjun, Wu, Tianzhen, Liu, Zhiliang & Yan, Jianlin, 2009; Jing-miao, Zhang & Wei-xiao, Gao, 2008; Ma, Yiming, Liu, Bing, Wong, Kian, Ching, Yu, S., Philip & Lee, Ming, Shuik, 2000; Mei-lan, Chen, 2010; Ogor, 2007; Pal, 2012)

DOI: 10.4018/978-1-5225-3725-0.ch015

Many clustering algorithms have been proposed in the literature (Gan, Guojun, Ma, Chaoqun, & Wu, Jianhong, 2007; Jain, K., Anil & Dubes C. Richard, 1988; Wu, Xindong, Kumar, Vipin, Quinlan, Ross, J.,…Dan, 2007; Xu, Rui, & Wunsch, Donald, 2005). The clustering algorithms are broadly classified into two categories, Hierarchical and Partitional. K-Means is a famous partitonal clustering algorithm. Simplicity and speed in classification of massive data are two features which have made K-Means a very popular algorithm. However, K-Means has a major limitation -- the number of clusters, 'K', need to be pre-specified as an input to the algorithm. In absence of thorough domain knowledge, or for a new and unknown dataset, this advance estimation and specification of cluster number typically leads to "forced" clustering of data, and proper classification does not emerge. As K-Modes and K-prototype algorithms are variants of K-Means, they carry this demerit of K-Means.

In this chapter, new algorithms based on the K-Means, K-Modes and K-Prototype are presented which have advance features of intelligent data analysis and automatic generation of appropriate number of clusters. The clusters generated by the new algorithm are compared against results obtained with the original K-Means, K-Modes and K-Prototype. The practical application of these algorithms is discussed in the field of education to cluster the similar students based on their academic performance using datasets of different types and dimensions.

BACKGROUND

The work of some researchers to remove the limitation of giving the number of clusters required as an input in the K-Means for numerical data is discussed below:

Pelleg et al. (2000) presented XMeans algorithm as an extension of K Means which overcomes the limitation of inputting the value of K and performs faster than the original K Means. The algorithm starts with K as the lower bound of the given range and continues adding centroids until the upper bound is reached. During this process the centroid set that scores the best is recorded using the data structure kd-tree and is produced as output. The user has to input a range suggesting the lower and upper bound of K.

Tibshirani et al. (2000) used the technique of Gap Statistic that utilizes the output generated by any clustering algorithm to compare the change within cluster dispersion to that expected under an appropriate reference null distribution. This Gap method works well when the clusters are well separated.

Wagstaff, et al. (2001) utilized information about the problem domain in performing clustering. This information about the problem domain is used in specifying the constraints on the data set. While assigning the data points to a cluster it is ensured that none of the constraint is violated.

Cheung (2003) presented a new clustering technique named STep-wise Automatic Rival penalized (STAR) K-Means algorithm. This new algorithm is a generalization of the conventional K-Means algorithm but overcoming its two major limitations discussed earlier in this paper. The algorithm consists of two separate steps. The first step provides each cluster a center. The next step then adjust the units adaptively by a learning rule. This algorithm is computationally complex.

Leela et al. (2013) presented Y-Means algorithm based on K-Means algorithm. Initially, it runs K-Means algorithm on the data set and then follows the sequence of splitting, deleting and merging the clusters. The algorithm depends on K-Means algorithm to find the clusters initially.