Hybrid Clustering using Elitist Teaching Learning-Based Optimization: An Improved Hybrid Approach of TLBO

D.P. Kanungo, Department of Computer Science Engineering and Information Technology, Veer Surendra Sai University of Technology (VSSUT), Odisha, India
Janmenjoy Nayak, Department of Computer Science Engineering and Information Technology, Veer Surendra Sai University of Technology (VSSUT), Odisha, India
Bighnaraj Naik, Department of Computer Science Engineering and Information Technology, Veer Surendra Sai University of Technology (VSSUT), Odisha, India
H.S. Behera, Department of Computer Science Engineering and Information Technology, Veer Surendra Sai University of Technology (VSSUT), Odisha, India

ABSTRACT

Data clustering is a key field of research in the pattern recognition arena. Although clustering is an unsupervised learning technique, numerous efforts have been made in both hard and soft clustering. In hard clustering, K-means is the most popular method and is being used in diversified application areas. In this paper, an effort has been made with a recently developed population based metaheuristic called Elitist based teaching learning based optimization (ETLBO) for data clustering. The ETLBO has been hybridized with K-means algorithm (ETLBO-K-means) to get the optimal cluster centers and effective fitness values. The performance of the proposed method has been compared with other techniques by considering standard benchmark real life datasets as well as some synthetic datasets. Simulation and comparison results demonstrate the effectiveness and efficiency of the proposed method.

KEYWORDS
Clustering, Elitist TLBO, GA, K-Means, PSO, TLBO

INTRODUCTION

The main aim of any clustering technique is to divide n no. of patterns into some of the clusters in a search space. Literature survey (Jain et. al, 1999; Hruschka et. al., 2009; Hasan and Ramakrishnan, 2011) reveals that different clustering techniques have been developed in recent years. Some descent types of clustering methods those have been frequently used by the researchers are: a) Partition based,
b) Hierarchy based, c) Density based, d) Grid based. Basically, the clustering algorithms are based on some criterion to assess the quality of cluster partition. In particular, depending on the parameters, the algorithms receive the inputs as the no. of clusters, cluster density, etc. and find the good partition. K-means (MacQueen, 1967) survey is a partition based hard clustering method. It is a center based algorithm where the Euclidian distance from each point to the center is computed and the aim is to calculate the nearest cluster center. In recent days, several attempts (Arbelaitz et. al., 2013; Roy et. al., 2012; Vanfleteren et. al, 2013; Michie et. al, 2013; Arumugam et. al., 2011; Zhang et. al., 2013; Parmar and Pandi, 2015; Ward et. al., 2015; Wang et. al., 2015; Nguyen et. al., 2015; Cosby et. al., 2015; Boersma et. al., 2014; Jose et. al., 2014; Chen et. al., 2014; Bose et. al., 2013) have been made by using this method and hybridizing it with various optimization techniques to solve diversified problems. Basically, K-means algorithm is more famous than some other clustering algorithms due its simple execution steps with a minimal parameter settings and ability to run quickly for large datasets. After consideration of a no. of successful applications of K-means, some researchers found some limitations (Arthur and Vassilvitskii, 2007; Singh et. al., 2011; Komali et. al., 2015; Acharya et. al., 2015; Liang et. al., 2015) about the algorithm which need to be addressed. As the algorithm is based on some initial selection of random cluster centers, so after the completion of each iteration, it produces different results. Although, the algorithm is able to minimize the intra cluster variance, but it is ineffective in producing global minimum variance. In a dense clustering environment, the data set is divided into k no. of clusters. For each data, the closest center point is computed based on the distance calculation procedure. But for a cluster having none of the assigning data points, it will result in an empty cluster. Also, for the data set having some outliers, the algorithm is unable to calculate an accurate cluster centroid as it originally would have been. Apart from these problems, the biggest deal is to find the global solution as K-means is highly sensitive to local optima due to the fixed unchangeable cluster centers after some point of calculations of the algorithm. However, a no. of improved versions (Cheung, 2003; Charalampidis, 2005; Şerban and Moldovan, 2006; Zhang et. al., 2008; Celebi, 2011; Tarpey, 2012; Ma et. al., 2015; Shao et. al., 2015; Duwairi and Abu-Rahmeh, 2015) and hybridization algorithm of K-means (Ahmadyfard and Modares, 2008; Wu, 2008; Kao and Lee, 2009; Schreiber, 1991; Gibou amd Fedkiw, 2005; Li et. al., 2015; Aparna and Nair, 2014; Pan et. al., 2014) have been developed. Somehow, these improved algorithms performed well for some particular applications, still they failed to obtain a global solution which is our area of focus.

In recent days, many evolutionary algorithms have been developed and successfully used in various engineering applications (Dey et. al., 2013; Nayak et. al., 2014; Ashour et. al., 2015; Behera et. al., 2015; Naik et. al., 2015a; Gorripotu et. al., 2015; Naik et. al., 2015b; Sahu et. al., 2015; Nayak et. al., 2015a; Sekhar et. al., 2016; Nayak et. al., 2015b; Kanungo et. al., 2015; Shuaib et. al., 2015). Amongst them, Teaching Learning based optimization (TLBO) is a recent population based metaheuristic developed by Rao (Rao et al., 2011; Rao and Kalyankar, 2013). It is influenced by the teaching and learning mechanism of a teacher and learners respectively. In a classroom environment, the teaching process of a teacher makes an impression on the quality of learners and reflects in the examination process. A good teacher always tries to give his/her best in the class and creates an environment for equal level learning to all the learners in the class. But, practically, it is not possible to maintain the same level for all the learners and henceforth, the mean of the learners can be considered as a good solution.

Elitist TLBO (Rao and Patel, 2014) is an improvised version of TLBO having capacity to perform better than TLBO. The term elitist itself indicates for the best; i.e. in the algorithmic implementation, the best within all good solutions will be considered from the population. The elitist term will not only help to choose the best solution, but also it will replace the redundant or duplicate solutions (if any in the population), as in the TLBO the mean is considered from the population of learners. The details about ETLBO have been described in the next section. Although TLBO algorithm is quite new in the optimization community, but still researchers have used this algorithm in a variant of application areas of science and engineering.
Movie Analytics for Effective Recommendation System using Pig with Hadoop
Arushi Jain and Vishal Bhatnagar (2016). *International Journal of Rough Sets and Data Analysis* (pp. 82-100).
www.igi-global.com/article/movie-analytics-for-effective-recommendation-system-using-pig-with-hadoop/150466?camid=4v1a