Chapter 31

Data Clustering Using Sine Cosine Algorithm:
Data Clustering Using SCA

Vijay Kumar
Thapar University, India

Dinesh Kumar
GJUS&T, India

ABSTRACT

The clustering techniques suffer from cluster centers initialization and local optima problems. In this chapter, the new metaheuristic algorithm, Sine Cosine Algorithm (SCA), is used as a search method to solve these problems. The SCA explores the search space of given dataset to find out the near-optimal cluster centers. The center based encoding scheme is used to evolve the cluster centers. The proposed SCA-based clustering technique is evaluated on four real-life datasets. The performance of SCA-based clustering is compared with recently developed clustering techniques. The experimental results reveal that SCA-based clustering gives better values in terms of cluster quality measures.

INTRODUCTION

Clustering is a technique that distributes unlabeled data into number of groups based upon some similarity measure that minimizes within-group variability, known as intra-cluster distance and maximizes the between-group variability, known as inter-cluster distance. It has been applied in various fields of science and engineering such as pattern recognition, image segmentation, web mining, remote sensing, etc. (Everitt, 1993, Abraham et al., 2008, Yamany et al 2015a). Data clustering algorithms is classified into two main categories: hierarchical and partitional (Leung et al, 2000). The former clustering construct a tree structure output which represent the nested grouping of the datapoints of a dataset (Frigui & Krishnapuram, 1999). However, it has some drawbacks such as it may fail to separate overlapping clusters and computationally expensive (Jain et al, 1999).

DOI: 10.4018/978-1-5225-2229-4.ch031

Whereas the latter one divides the dataset into a number of groups based upon certain clustering criteria. The clustering criterion directly affects the nature of clusters formed. The advantages of partitional clustering algorithms are the disadvantages of the hierarchical algorithms and vice versa. The partitional clustering is widely used in pattern recognition than hierarchical clustering as reported in literature (Xu & Wunsch, 2009, Yamany 2015b, Gaber et al 2016). The partitional clustering algorithm is the main focus of this chapter. The well-known partitional clustering technique is K-Means. The performance of K-Means clustering algorithm is affected from two problems such as the number of clusters specified and cluster center initialization. To resolve these clustering problems, metaheuristic techniques are genetic algorithms (GA), ant colony optimization (ACO), particle swarm optimization (PSO), gravitational search algorithm (GSA).

Recently, Mirjalili (2016) developed a novel metaheuristic algorithm namely sine cosine algorithm (SCA), which utilizes mathematical functions of sine and cosine. SCA is preferred over other metaheuristic techniques as it is easy to implement and fast convergence speed. Due to these advantages of SCA, it is employed on data clustering. The novel approach is proposed in which SCA is able to produce the optimal cluster centers.

This chapter aims to explore the applicability of sine cosine approach to the development of clustering technique. It includes a general overview of data clustering with emphasis on recently developed metaheuristic-based clustering techniques followed by proposed sine cosine based clustering technique. The performance evaluation has been done on the real-life datasets.

BACKGROUND

This section describes the related concepts of cluster analysis and related works on metaheuristics-based data clustering techniques.

Cluster Analysis

The partitional clustering technique is defined as follows. Let a dataset X consists of N data points, $X = \left\{x_1, x_2, ..., x_N\right\}$. Each data point is described by D features, where $x_j = \left(x_{j1}, x_{j2}, ..., x_{jD}\right)$ is a vector represent the j^{th} data point and x_{ji} represent the i^{th} feature of x_j. The main aim of clustering technique is to partition the dataset into a number of clusters (say K) $\left\{C_1, C_2, ..., C_K\right\}$ based on some similarity/dissimilarity measure. The value of K may or may not be known a prior. The partition matrix is represented as $U = [u_{kl}]$, $k = 1, 2, ..., K$ and $l = 1, 2, ..., N$, where u_{kl} is the membership of data point x_j to cluster C_k (Abraham et al., 2008). For the hard partitioning of the dataset, the following condition must be satisfied (Xu and Wunsch, 2009).

$$u_{kl} = \begin{cases} 1 & if \ x_l \in C_k \\ 0 & if \ x_l \notin C_k \end{cases} \tag{1}$$

Related Content