Hybrid Variable-Length Spider Monkey Optimization with Good-Point Set Initialization for Data Clustering

Athraa Qays Obaid, Maytham Alabbas*
Department of Computer Science, College of Computer Science and Information Technology, University of Basrah, Basrah, Iraq
E-mail: itpg.athraa.qays@uobasrah.edu.iq, ma@uobasrah.edu.iq
*Corresponding author

Keywords: data clustering, spider monkey optimization, k-means clustering, variable-length spider monkey optimization, good-point set

Received: May 21, 2023

Data clustering refers to the process of grouping similar data points based on patterns or characteristics. It finds applications in image analysis, pattern recognition, and data mining. The k-means algorithm is commonly used for this purpose, but it has two main limitations. Firstly, it necessitates the user to explicitly specify the number of clusters. Secondly, it is highly sensitive to the initial selection of cluster centroids. To overcome these limitations, this study presents a novel approach that utilizes a variable-length spider monkey optimization algorithm (VLSMO) with a hybrid measure to determine the optimal number of clusters and initial centroids. Experimental results obtained from real-life datasets demonstrate that VLSMO outperforms the standard k-means algorithm and other techniques in terms of accuracy and clustering capacity.

Povzetek: Prispevek opisuje novo metodo za združevanje podatkov z uporabo optimirnega algoritma VLSMO, ki odpravlja omejitve predhodnih algoritma, izboljšuje natančnost in zmogljivost združevanja.

1 Introduction

Data clustering is one of the most important data mining approaches, which involves partitioning data instances into smaller groups, where each group comprises objects that are similar to each other but distinct from those in other clusters. Clusters are defined by a center point and a proximity metric that measures the similarity or dissimilarity of the candidate data points. Clustering analysis has as its primary objective the creation of clusters consisting of the highest density of similar points and the most distant clusters of different points. Clustering cannot be performed manually due to the large volume of data. Instead, specialized computing techniques are used. Nonetheless, clustering differs from classification since most data is unlabeled, implicitly performing classification. Therefore, it is considered unsupervised learning.

The practice of clustering is prevalent throughout industries, as it allows related objects to be grouped. For example, a clustering approach can be used in marketing to identify consumers with similar purchasing habits. In educational settings, it can be useful in analyzing students' academic achievement by grouping those with similar study habits. In addition, clustering can also be utilized in diverse applications, including the segmentation of images, the detection of outliers, the detection of tumors, and the detection of fraud. Furthermore, clustering is a powerful method of uncovering hidden patterns within a dataset. Despite its wide range of applications, clustering poses a number of challenges [1].

Many clustering algorithms have been developed, including K-means, density-based spatial clustering of applications with noise (DBSCAN), expectation maximization (EM), and hierarchical agglomerative clustering (HAC). The K-means algorithm is widely recognized as the most commonly employed clustering algorithm. The K-means algorithm is popular in scientific research and industrial applications because it is simple, fast convergence, and scalable. Nevertheless, k-means clustering, which involves randomly distributing starting points during center initialization, frequently results in local optimal clustering outcomes that could result in inaccurate categorization due to instability. It has a number of limitations, including the necessity of specifying the number of clusters and its sensitivity to initial center points. In order to overcome the limitations of this algorithm, a globally optimized approach must be adopted [2]. Several techniques have been proposed in the literature to overcome these limitations, including the elbow method, the gap statistic, and the canopy method. The appropriate number of clusters (k) can be determined using these techniques. Furthermore, various algorithms can be used to identify the initial centroids, including the Forgy method, random partition method, and k-means++ algorithm. As far as the authors are aware, there are limited techniques available to determine both parameters simultaneously based on using optimization techniques like genetic algorithms (GA), artificial bee colony (ABC), and particle swarm optimization (PSO).

The current work aims to progress in this area by presenting a modified version of the spider monkey