Chaotic Tornadogenesis Optimization Algorithm for Data Clustering Problems

Ravi Kumar Saidala, Acharya Nagarjuna University, Guntur, India
Nagaraju Devarakonda, LBR College of Engineering, Malavaram, India

ABSTRACT

This article describes how clustering is an attractive and major task in data mining in which particular set of objects are grouped according to their similarities based on some criteria. Among the numerous algorithms, k-Means is the best and efficient in address clustering problems. Any expert system is said to be good, only if it returns the optimal data clusters. The challenge of optimal clustering lies in finding the optimal number of clusters and identifying all the data groups correctly which is a NP-hard problem. Recently a new optimization algorithm TOA was developed to address these problems. However, the standard TOA is too often trapped at the local optima and premature convergence. To overcome this, this article proposes CTOA. The main objective of embedding chaotic maps into standard TOA is to compute and automatically adapt the internal parameters. The proposed CTOA is first benchmarked on standard mathematical functions and later applied to 10 data clustering problems. The obtained graphical and statistical results along with comparisons illustrate the capabilities of CTOA regarding accuracy and robustness.

KEYWORDS

Algorithms, Chaotic Maps, Cognitive Algorithms, Data Clustering Problem, Nature Inspired Meta-heuristic optimization, TOA

1. INTRODUCTION

Investigations and developments are made in recent decades to find optimal solutions for large and dynamic problems using nature-inspired algorithms (Chakraborty, Amrita, & Kumar Kar, 2017). This can be achieved by avoiding inefficient enumerating process. Many researchers have developed numerous optimization algorithms by looking into the nature, looking into the biology and tried to model some of the impressive and intellectual mechanisms (Pedrycz, & Witold, 2010) into new algorithms for different engineering applications (Bozorg-Haddad, & Omid, 2017). Population based Meta-heuristic optimization algorithms have numerous engineering applications (Patnaik, Srikanta, Yang, & Nakamatsu, 2017). These kinds are popular in optimizing complex functions. Finding optimal structures of data is a challenging task in data mining (Özbakır, Lale, & Turna, 2017; Jain, 2010; Han, & Kamber, 2001). Cluster analysis is an unsupervised technique and one of the best way to find the structures of data. This clustering process discovers the natural grouping of data points according to the similarity of measured intrinsic characteristics (Gagliardi, & Francesco, 2012). For an instance, in the k-Means partitional clustering, the similarity function that should be minimize centroid distance to obtain good clusters. i.e. minimization of sum of squared Euclidean distance of objects from respective cluster means that is shown as fallows.

DOI: 10.4018/IJSSCI.2018010104

Copyright © 2018, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
\[ d_{\text{min}} = \sum_{j=1}^{K} \sum_{Z_i \in C_j} Z_i - \mu_j^2 \]  

(1)

Where \( \mu_j \) is the mean of \( C_j \)

The problem of optimal data clustering is formalized as follows. Let consider a dataset \( D_T \), where \( T \) denotes the total number of objects and \( N \) be the dataset dimensions (variables) that are considered. Let the input N-dimensional dataset.

\[ D = \{ X_1, X_2, \ldots, X_n \} \]  

(2)

Each data object is then an \( N \)-dimensional vector:

\[ D_T = \begin{bmatrix} X_1, & X_2, & \ldots, & X_n \end{bmatrix} \]  

(3)

Assume dataset \( D_T \) with \( N \) dimensions is partitioned into \( C_i \) clusters, where \( i = 1 \) to \( k \), so that for each \( D_T \), The similarity measured by distance \( d \) then,

<table>
<thead>
<tr>
<th>Table 1. Notations of Uni and Multimodal benchmark functions Where V_no = 30, fmin = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Function name</strong></td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>( F_1(x) = \sum_{i=1}^{n} X_i^2 )</td>
</tr>
<tr>
<td>( F_2(x) = \sum_{i=1}^{n}</td>
</tr>
<tr>
<td>( F_3(x) = \sum_{i=1}^{n} \left( \sum_{j=1}^{i} x_j \right)^2 )</td>
</tr>
<tr>
<td>( F_4(x) = \max_{1 \leq i \leq n} { x_i } )</td>
</tr>
<tr>
<td>( F_5(x) = \sum_{i=1}^{n-1} \left[ 100 \left( x_{i+1} - x_i^2 \right)^2 + (x_i - 1)^2 \right] )</td>
</tr>
<tr>
<td>( F_6(x) = \sum_{i=1}^{n} \left(</td>
</tr>
<tr>
<td>( F_7(x) = \sum_{i=1}^{n} i x_i^4 + \text{random}[0,1] )</td>
</tr>
</tbody>
</table>
T-Spanner Problem: Genetic Algorithms for the T-Spanner Problem
www.igi-global.com/chapter/t-spanner-problem/180937?camid=4v1a

Agent-Based Middleware for Advanced Communication Services in a Ubiquitous Computing Environment
www.igi-global.com/chapter/agent-based-middleware-advanced-communication/64606?camid=4v1a