Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 125 (2018) 780–789

Procedia Computer Science

www.elsevier.com/locate/procedia

6th International Conference on Smart Computing and Communications, ICSCC 2017, 7-8 December 2017, Kurukshetra, India

A Novel Heuristic for Evolutionary Clustering

Pranav Nerurkar*,[a], Archana Shirke[b], Madhav Chandane[c], Sunil Bhirud[d]

[a]*Dept. of Computer Engineering & IT, VJTI, Mumbai - 400019, India*
[b]*Dept. of Computer Engineering & IT, VJTI, Mumbai - 400019, India*
[c]*Dept. of Computer Engineering & IT, VJTI, Mumbai - 400019, India*
[d]*Dept. of Computer Engineering & IT, VJTI, Mumbai - 400019, India*

Abstract

Clustering is considered a challenging problem of data mining due to its unsupervised nature. The literature is inundated with algorithms and concepts related to determining the most suitable clustering structure in data. These techniques have a mathematical model of a cluster and attempt to obtain a result that shall represent this model as closely as possible. However as the problem of clustering is NP hard such strategies have disadvantages such as converging to local optima or suffering from the curse of dimensionality. In such scenario, meta heuristics could be more suitable strategies. Such techniques utilizes biologically inspired techniques such as swarm intelligence, evolution etc. to traverse the search space. Due to their inherent parallel nature, they are most robust towards converging to a local optima. The objective (cost) function used by such meta heuristics is responsible for guiding the agents of the swarm towards the best solution. Hence it should be designed to achieve trade-off between multiple objectives and constraints and at the same time produce relevant clustering. In this paper, a cost function is proposed (PSO-2) to produce compact well separated clusters by using the concept of intra-cluster and inter-cluster distances. Experiments have been performed on artificial benchmark data-sets where performance of the particle swarm optimizer using the proposed cost function is evaluated against other evolutionary and non evolutionary algorithms. The clustering structures produced by the methods have been evaluated using distance based and internal cluster validation metrics to demonstrate that the performance of PSO-2 is comparable to other techniques.

Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications

Keywords: Data clustering; Applications of evolutionary algorithms; Genetic algorithm

* Corresponding author. Tel.: +91-961-999-7797.
E-mail address: pranavn91@gmail.com

1877-0509
Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications
10.1016/j.procs.2017.12.100

1. Introduction

Clustering involves creating coarse grained descriptions of data. This is done with the main purpose of highlighting a set of autonomous regions in the data known as clusters. Clustering has found wide applicability in multiple fields ranging from web mining and information retrieval through customer segmentation and bio-informatics [1]. Regardless of the domain to which it is applied or the algorithm used, the goal of clustering is uniform i.e. to maximize homogeneity within cluster and heterogeneity between clusters [2]. Clustering algorithms produce partitions of data which may be hard (nodes have unique memberships to clusters), soft (nodes have memberships to multiple clusters) or fuzzy (nodes have varying degrees of memberships to every cluster). Thus, a hard partition of a dataset $X = x_1, x_2, x_3, x_4 ... x_n$ where x_i is a data point which stands for a n-dimensional feature vector, is a collection $C = C_1, C_2, ..., C_k$ of non null clusters such that $C_i \cap C_j = \phi$ for $i \neq j$. Relaxing the mutual dis-junction condition to allow $C_i \cap C_j \neq \phi$ for $i \neq j$ creates overlapping partitions [3].

Clustering is considered the most challenging task of machine learning due to the unavailability of spatial distribution of the data in terms of its clustering tendency [1] [2] [3]. Furthermore, the requirement of dealing with various types of attributes (binary, categorical, continuous, discrete), their conditions (complete or missing) and scales (nominal, ratio, interval, ordinal) has to be factored in while deciding the approach of clustering. In addition to this, the lack of information about the orientation of the clusters, their numbers, shapes, densities and volumes makes it difficult in selecting a particular clustering technique and in evaluating the results obtained by it [16] [17]. From the optimization perspective, Clustering is considered as a NP-hard grouping problem and the literature consists of a wide range of objective functions, approximation algorithms and heuristics for solving it [1].

Each approach has its own bias and comes with certain advantages and disadvantages to a given analysis or application scenario. Popular partition based clustering algorithms are K-means, PAM, CLARA, CLARANS. These algorithms have a objective function which is non-convex and hence the results might be locally optimal. Hierarchical algorithms too exist like AGNES, DIANA, BIRCH, ROCK and CHAMALEON etc. but these are computationally expensive. One inherent weakness of hierarchical clustering is that clusters formed in an iteration cannot be undone in subsequent iterations. DBSCAN, OPTICS are approaches that define clusters as dense regions separated by less dense regions. DBSCAN though is popular but the prior specification on ϵ and $MinPts$ makes it sensitive to parameter tuning. In real life experiments, various types of noise are introduced at different experimental stages during data collection, due to which the performance of these algorithms are less than ideal [1] [3]. To effectively handle noisy data, we need an algorithm that is able to overcome noise. These traditional algorithms are based on gradient based methods and are difficulty to extend to multi-objective problems as their basic design doesn't allow the consideration of multiple solutions [4] [5]. Population based methods such as evolutionary algorithms have an inherent advantage here [7] [8].

In the context of solving NP-hard optimization problems, evolutionary algorithms are considered to be particularly effective in obtaining near optimal solutions in reasonable time [9] [10]. Evolutionary algorithms are meta-heuristics based on optimization of a fitness function that guides the process of evolutionary search. Computational advantages are also present as the algorithms can be parallelized leading to increased coverage and possibly faster convergence [6]. However, evolutionary algorithms have their owns drawbacks as different features of these algorithms have to chosen in advance such as encoding scheme for the data points, choice of operators for crossover and mutation and a fitness (cost) function. Fitness functions are basically validity criteria and hence it is better to use multiple criteria in order to avoid the drawbacks of individual ones. One strategy for this is to assign weights to individual criteria and then optimize, but this approach only works when the multiple criteria are commensurable. In this paper, the fitness function proposed computes cost using a variation of intra-cluster distance and inter cluster distance. This variation allows computation to be performed in linear time. The conventional definitions of these distances requires $O(n^2)$ computations. Hence a reduction in computations is achieved leading to faster execution.

The rest of this paper is organized as follows. In Section II, an overview of the existing clustering algorithms used for identification of clusters in data is presented, these include both evolutionary and non evolutionary approaches. In