Multi-Objective Genetic Algorithm for Robust Clustering with Unknown Number of Clusters

Amit Banerjee, The Pennsylvania State University, USA

ABSTRACT

In this paper, a multi-objective genetic algorithm for data clustering based on the robust fuzzy least trimmed squares estimator is presented. The proposed clustering methodology addresses two critical issues in unsupervised data clustering – the ability to produce meaningful partition in noisy data, and the requirement that the number of clusters be known a priori. The multi-objective genetic algorithm-driven clustering technique optimizes the number of clusters as well as cluster assignment, and cluster prototypes. A two-parameter, mapped, fixed point coding scheme is used to represent assignment of data into the true retained set and the noisy trimmed set, and the optimal number of clusters in the retained set. A three-objective criterion is also used as the minimization functional for the multi-objective genetic algorithm. Results on well-known data sets from literature suggest that the proposed methodology is superior to conventional fuzzy clustering algorithms that assume a known value for optimal number of clusters.

Keywords: Fuzzy C-Means, Genetic Algorithms, Multi-Objective Optimization, Optimal Clusters, Outliers, Robust Clustering

INTRODUCTION

Clustering procedures produce a partition of data into clusters, with similar entities grouped in a cluster and dissimilar entities in different clusters. The process of organizing data into groups is one of the most fundamental modes of understanding and learning (Jain & Dubes, 1998). The k-Means (MacQueen, 1967) class of clustering algorithms generates non-overlapping k-partitions of the data by minimizing the least squares (LS) residual within every cluster. Fuzzy c-Means or FCM as it is popularly known (Bezdek, 1981) is the fuzzy extension of k-Means and generates overlapping clusters where every entity is shared with varying degrees of membership across all the clusters. However, since both k-Means and FCM are based on LS residual minimization, they are extremely sensitive to outliers in data. There have been attempts at robustifying k-Means and FCM by incorporating low-breakdown regression estimators such as ones based on covariance and absolute deviations of residu-
als, and several high-breakdown estimators like the least median of squares (LMS), the least trimmed squares (LTS) estimator, among others. Another standing criticism of k-Means and FCM is the requirement that the number of clusters be known \textit{a priori}. In this paper, a novel robust clustering methodology with a multi-objective genetic algorithm is proposed. The technique uses FCM to find an optimal partition on a retained subset of the original noisy data, and optimizes the entities in the retained set and the number of clusters within the retained set using an elitist multi-objective genetic algorithm. The methodology presented in this paper is a significant extension of the work in Banerjee and Louis (2007), where a canonical genetic algorithm was used to implement an LTS estimator for robust clustering for a known number of clusters. The paper is structured as follows - in the next section, brief survey of related work in genetic algorithm-based robust clustering is presented, followed by a discussion of the feasible solution algorithm to implement the LTS estimator. The multi-objective criterion is subsequently introduced followed by the proposed multi-objective genetic algorithm-driven robust clustering methodology. Computational results on benchmark and artificial data sets and a summary bring the paper to a conclusion.

**GENETIC ALGORITHMS IN ROBUST CLUSTERING**

The clustering problem is often mapped to one of searching for an optimum partition from all possible and valid partitions such that some goodness of fit criterion is optimized. It can therefore be formally considered a particular class of NP-hard problem (Falkenauer, 1998) which has led to the development of efficient metaheuristics to provide near-optimal solutions in reasonable time. Evolutionary algorithms are a class of population-based optimization techniques that mimic the biological evolution process and include among others genetic algorithms (GA), genetic programming (GP), evolutionary strategies (ES), evolutionary programming (EP) and newer techniques such as differential evolution (DE) and memetic algorithms (MA). These methods solve the optimization problem by evolving a population of possible solutions by using a set of evolutionary operators that act on solutions based on their relative fitness. The methods differ in the way they represent potential solutions and the way the evolutionary operators are handled, which dictates their applicability to certain types of optimization problems. Selection, crossover and mutation are the most widely used evolutionary operators. The population evolves when the selection operator chooses a subset of the fittest solutions in the present generation to reproduce using recombination operators such as crossover and mutation. The evolved population on the average is fitter than the population it replaces.

The aim of evolutionary clustering algorithms is to evolve a population of sub-optimal partitions to a population that contains potential near-optimal partitions. Evolutionary clustering algorithms can be categorized into algorithms that cluster data when the number of clusters is known \textit{a-priori} and others where the number of clusters is considered variable. For a fixed number of clusters, a partition can be encoded in the population using the integer cluster-label representation which for a dataset of size \( n \) is \( n \)-bits long where the value of the \( i \)-th bit is the label of the cluster it belongs to (Krishna & Murty, 1999; Lu et al., 2004). Since the number of clusters is assumed known \( (k) \), the value each bit can take varies between 1 and \( k \). It can be easily seen that the cluster-label representation is naturally redundant i.e., the same partition can be represented by \( k! \) different genotypes by varying the labels of the cluster. Another encoding scheme is the prototype-based representation which includes the real-valued \( k \)-medoid representation and the real-valued \( k \)-centroid representation (Bandyopadhyay & Maulik, 2002; Merz & Zell, 2002; Kivijärvi, Fränti, & Nevalainen, 2003). In clustering terms, a medoid is a data object that can be assumed representative of the cluster and a centroid on the other hand is a cluster prototype whose parameters may not necessary coincide with any of the data.
Fractal Geometry and Computer Science
www.igi-global.com/chapter/fractal-geometry-computer-science/28386?camid=4v1a