Development of Fractional Genetic PSO Algorithm for Multi Objective Data Clustering

Aparna K., B. M. S. Institute of Technology, Bangalore, India
Mydhili K. Nair, M. S. Ramaiah Institute of Technology, Bangalore, India

ABSTRACT

Clustering is the task of finding natural partitioning within a data set such that data items within the same group are more similar than those within different groups. The performance of the traditional K-Means and Bisecting K-Means algorithm degrades as the dimensionality of the data increases. In order to find better clustering results, it is important to enhance the traditional algorithms by incorporating various constraints. Hence it is planned to develop a Multi-Objective Optimization (MOO) technique by including different objectives, like MSE, Stability measure, DB index, XB-index and sym-index. These five objectives will be used as fitness function for the proposed Fractional Genetic PSO algorithm (FGPSO) which is the hybrid optimization algorithm to do the clustering process. The performance of the proposed multi objective FGPSO algorithm will be evaluated based on clustering accuracy. Finally, the applicability of the proposed algorithm will be checked for some benchmark data sets available in the UCI machine learning repository.

KEYWORDS

DB Index, MSE, Multi-Objective Optimization, Partitional Clustering, Stability, Sym-Index, XB-Index

1. INTRODUCTION

Data clustering is the procedure of clustering together similar multi-dimensional data vectors. A comprehensive study and analysis of the different partitional clustering algorithms is given in (Aparna & Nair, 2015a). Clustering algorithms have been employed to a broad range of problems, together with exploratory data analysis, data mining (Evangelou et al., 2001), image segmentation (Lillesand et al., 1994) and mathematical programming (Andrews H.C, 1972), (Rao, 1971). Clustering techniques have been employed effectively to address the scalability problem of machine learning and data mining algorithms and also for developing optimized performance (Jain et al., 1999), (Quinlan, 1993), (Potgieter, 2002). Clustering replicates the statistical structure of the general collection of input patterns in the data and hence the subset of patterns has definite meanings (Roy & Sharma, 2010). The pattern can be symbolized mathematically by a vector in the multi-dimensional space.

Clustering algorithms can be clustered into two main classes of algorithms, namely supervised and unsupervised. The shortage of category information differentiates data clustering (unsupervised learning) from categorization or discriminant analysis (supervised learning). Clustering is the process of finding out different structures in data that are analytical in nature (Yip et al., 2004). No labelled data are accessible (Everitt et al., 2001), (Jain & Dubes, 1988) in unsupervised classification which is also called clustering. The objective of clustering is to divide a fixed unlabeled data set into a
fixed and separate set of “natural”, hidden data structures (Baraldi & Alpaydin, 2002), (Cherkassky & Mulier, 1988). For several learning domains, the characteristics that are potentially constructive are described manually. On the other hand, not all of these characteristics may be related. Selecting a subset of the original characteristics will frequently lead to improved presentation in such a case. Feature selection algorithms exploit some functionalities of predictive precision (Dy & Brodley, 2004) for supervised learning.

A lot of clustering algorithms have been proposed. One of the most famous hard clustering algorithms is K-Means which divides data objects into k clusters (Kanungo et al, 2002). Fuzzy algorithms can allocate data objects into multiple clusters. Fuzzy C-Means clustering is an efficient algorithm; moreover the arbitrary choice in initializing the centre points makes the iterative process in achieving local optimal solution without difficulty. In order to enhance the solution, many evolutionary algorithms such as Genetic Algorithm (GA) (Maulik & Bandyopadhyay, 2000), Simulated Annealing (SA) (Bandyopadhyay et al, 2001), Ant Colony Optimization (ACO) (Dai et al, 2009), and Particle Swarm Optimization (PSO) (Ghorpade & Metre, 2014) have been effectively used for the clustering. In addition, Multi-objective clustering is used to decompose a dataset into related groups, thereby maximizing multiple objectives. Multi-objective clustering can be looked out as a unique case of multi-objective optimization which plans to concurrently optimize multiple objectives under definite constraints.

2. LITERATURE SURVEY

For multi objective data clustering literature presents several theories. Now we assess some of the works associated to it: (Behera et al, 2011) have proposed that Canonical Variate analysis is by far more efficient and effective in reducing the dimensions of a high dimensional dataset. Clustering technique is applied to the reduced low dimensional data set using the modified K-Means clustering. In order to get better result, genetic algorithm is applied for the purpose of initializing the centroids of the Improved Hybridized K-Means Clustering Algorithm (IHKMCA) in their paper. As compared to other approaches, the work has shown effective and accurate results with less time consumption.

The Ant Colony Optimization based Clustering methodology (ACO-C) has been described by (Inkaya et al, 2015). This clustering methodology tackles quite a few disputes of the clustering problem with solution assessment, neighbourhood construction, and data set reduction. In this framework, they initially bring in two objective functions, namely adjusted compactness and relative separation. Each objective function assesses the clustering solution regarding the local features of the neighbourhoods. This permits to measure the quality of a broad range of clustering solutions without a priori data. ACO-C includes two pre-processing steps: neighbourhood construction and data set reduction. The former removes the local features of data points, while the latter was employed for scalability. The multi-objective assessment mechanism relative to the neighbourhoods improves the extraction of the arbitrary-shaped clusters containing density differences.

An interval weighted fuzzy C-Means clustering by genetically guided alternating optimization has been proposed by (Zhang et al, 2014). Interval number was considered for attribute weighting in the Weighted Fuzzy C-Means (WFCM) clustering, and it was demonstrated that the attained interval weighting was suitable from the point of view of geometric probability. In addition, a genetic heuristic approach for attribute weight searching was also considered to direct the alternating optimization (AO) of WFCM, and enhanced attribute weights in interval-constrained ranges and sensible data partitions were attained concurrently. The experimental results showed that the algorithm performed better. It exposes the interval weighted clustering as an optimization operator on the basis of the traditional
Deferred System's Design: Countering the Primacy of Reflective IS Development with Action-Based Information Systems

An Empirical Study on Patients’ Acceptance and Resistance Towards Electronic Health Record Sharing System: A Case Study of Hong Kong