Decision Support Systems 51 (2011) 794–809

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Multi-objective design of hierarchical consensus functions for clustering ensembles via genetic programming

André L.V. Coelho [a,*], Everlândio Fernandes [a], Katti Faceli [b]

[a] Graduate Program in Applied Informatics, Center of Technological Sciences, University of Fortaleza, Av. Washington Soares, 1321/J30, 60811-905, Fortaleza, CE, Brazil
[b] Federal University of São Carlos, Sorocaba Campus, Rod. João Leme dos Santos, Km 110, Itinga, 18052-780, Sorocaba, SP, Brazil

ARTICLE INFO

Available online 1 February 2011

Keywords:
Cluster analysis
Clustering ensembles
Multi-objective clustering
Hierarchical fusion
Partition selection
Genetic programming

ABSTRACT

This paper investigates a genetic programming (GP) approach aimed at the multi-objective design of hierarchical consensus functions for clustering ensembles. By this means, data partitions obtained via different clustering techniques can be continuously refined (via selection and merging) by a population of fusion hierarchies having complementary validation indices as objective functions. To assess the potential of the novel framework in terms of efficiency and effectiveness, a series of systematic experiments, involving eleven variants of the proposed GP-based algorithm and a comparison with basic as well as advanced clustering methods (of which some are clustering ensembles and/or multi-objective in nature), have been conducted on a number of artificial, benchmark and bioinformatics datasets. Overall, the results corroborate the perspective that having fusion hierarchies operating on well-chosen subsets of data partitions is a fine strategy that may yield significant gains in terms of clustering robustness.

1. Introduction

Among the several data mining tasks that have been heavily investigated lie those involving the partitioning of data into groups via a process named clustering [29,41]. In a nutshell, the main goal of clustering is to find natural groupings of multidimensional data, called clusters, so that both the homogeneity within each cluster and the heterogeneity among different clusters are maximized. That is, instances (i.e., patterns, objects) that belong to the same cluster should be more similar to each other than the instances appearing in different clusters. Albeit simple to state, this is an ill-posed problem to pursue due mainly to its unsupervised nature and to the lack of a precise definition of what a cluster really is [5,34]. Actually, data can reveal clusters with different shapes and sizes, and the number of clusters depends very much on the resolution with which the data are effectively analyzed. As a consequence, different clustering algorithms have been proposed and new algorithms continue to emerge, each associated with different assumptions on data and different clustering validation criteria [12,28].

Even though conventional clustering algorithms have been successfully applied in a range of scenarios [41], the choice of the algorithm (and thus the clustering criterion) best suited to a given dataset is still a non-trivial task to realize. This is even more serious when dealing with the discovery of more than one underlying structure that can be present in the data. In fact, as Law et al. [34] have pointed out, "inability to detect clusters with diverse shapes and sizes is a fundamental limitation of every clustering algorithm irrespective of the clustering criterion (objective function) used." In order to overcome these problems, more advanced strategies for cluster analysis have been recently proposed and assessed in the literature.

Roughly speaking, these advanced strategies work by combining several clustering criteria in order to avoid the choice of one single criterion when nothing is known about the underlying structure present in the data and to guarantee the simultaneous retrieval of all the structures when more than one is available. These strategies encompass clustering ensemble algorithms [15,20], like those proposed by Strehl and Ghosh [38], Fern and Brodley [17], Topchy et al. [40], and Frossyniotis et al. [19]; multi-objective clustering methods [15,34], like MOCK (Multi-Objective Clustering with automatic K-determination) [24]; and multi-objective clustering ensemble models, like MOCLE (Multi-Objective CLustering Ensemble) [13,14].

It is worth noticing that all these alternatives still rely on the use of basic clustering algorithms. In fact, the traditional algorithms are very effective in uncovering the type of structure they were designed to find, if this structure is indeed present in the data. The main problem is that usually the real data are not "well-behaved" and nothing is known a-priori about the structures they hide to guide the choice of the best algorithm to adopt.

Typical clustering ensemble algorithms deal with the difficulty in the choice of the most suitable clustering algorithm while still focusing on the discovery of one single structure. The multi-objective approaches, in contrast, try to solve the two issues at once. As a matter of fact, all the aforementioned alternatives were constructed by

* Corresponding author. Tel.: +55 85 3477 3268; fax: +55 85 3477 3061.
E-mail addresses: acoelho@unifor.br (A.L.V. Coelho), everlandio@gmail.com (E. Fernandes), katti@ufscar.br (K. Faceli).

0167-9236/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2011.01.014

forming various combinations of partitions produced via traditional algorithms. However, none of them makes explicit which partitions to select and how these partitions are combined. Moreover, a single strategy is used to merge all the partitions, even in the multi-objective case, where the combinations are made progressively.

In this paper, we propose a cluster analysis framework that, like MOCLE, hybridizes the clustering ensemble and multi-objective clustering strategies. However, differently from the previously discussed alternatives, our aim is to synthesize the ensembles that will ultimately produce the partitions. In this framework, henceforth referred to as MCHPF (for "Multi-objective Clustering with Hierarchical Partition Fusions"), a Pareto-based version of genetic programming (GP) [33,36,43] is used to evolve a population of ensembles, taking into account complementary clustering validation measures. Each ensemble, in this case, is a hierarchy of consensus functions applied to a subset of base partitions, also produced with traditional clustering algorithms. That is, MCHPF automatically designs hierarchical consensus functions possibly considering different consensus functions already available in the literature.

By evolving the population in a multi-objective way, the proposed approach can find several structures present in the data, in accordance with more than one clustering criterion. Moreover, since the resulting ensembles are sets of individuals, we can easily observe which partitions were fused and how they were fused. Finally, several different strategies for combination can be used in one single hierarchical consensus function.

To evaluate our proposal, systematic experiments on artificial, benchmark and bioinformatics datasets with varying structures have been performed, comparing the quality of the partitions obtained by MCHPF with those delivered by traditional as well as other ensemble and multi-objective algorithms. Moreover, we assess the potentials of several distinct configurations of the conceptual components of the proposed framework so as to understand how the choice of these components may impact the levels of performance achieved.

The rest of the paper is organized as follows. In Section 2, we briefly revisit two of the fundamental limitations exhibited by traditional clustering algorithms and then comment upon some important clustering ensemble and multi-objective clustering approaches. Also in this section, the basic concepts and steps of GP are introduced. Section 3 is devoted to the detailed characterization of the proposed framework, providing the specification of its standard configuration as well as other ten variants. In this context, a contrast between MCHPF and MOCLE is provided, highlighting their distinctive conceptual aspects. In Section 4, we report on several simulation experiments we have conducted to evaluate the potential of the novel framework. After explaining how the experiments have been configured in terms of the datasets and clustering methods used, the control parameter settings, and the criterion adopted for measuring the quality of the resulting partitions, we present and discuss in a step-by-step manner the results obtained. Finally, Section 5 concludes the paper and provides remarks on future work.

2. Related work

The following subsections outline important aspects related, respectively, to basic clustering algorithms, clustering ensembles, multi-objective clustering, and genetic programming.

2.1. Limitations of traditional clustering algorithms

Traditional cluster analysis operates by seeking the structure present in the data that agrees most closely with some specified clustering criterion (that is, is in line with one of the possible cluster definitions). However, the practical application of such analysis suffers from two main drawbacks [19,28,34]. First, as there is no single cluster definition, it is very difficult to choose the clustering algorithm (optimizing only one clustering criterion) that is more appropriate to discover the underlying structure of the data at hand. Second, in several practical cases, the data themselves can present several underlying structures, each characterized by a different cluster definition or at a different refinement level.

To cope with the first problem, one alternative is to apply several clustering algorithms to the data and select the best partition generated according to a single validation measure. However, the validation measures employed to choose the best result(s) are usually biased towards one of the cluster definitions, since many of these measures are just more elaborated formulations of the available clustering criteria [12,25]. For instance, the intra-cluster variance, which is a measure that evaluates the compactness of clusters in a partition, has the same properties of the clustering criterion adopted by the *k*-means algorithm [41]. Another alternative to cope with the problem of clustering algorithm selection is to combine several partitions induced in consonance with different clustering criteria into a final solution with, hopefully, improved quality. Clustering ensembles [15,20], discussed in the following, are the simplest way to accomplish this.

2.2. Clustering ensembles

Instead of choosing the single best partition, clustering ensemble algorithms aim at finding a partition that represents the consensus among previously-obtained partitions. These algorithms usually consist of two steps: 1) the generation of a diverse set of base partitions; and 2) the application of a consensus function (fusion operator) to fuse such partitions into a consensus one [22,40]. The set of base partitions can be homogeneous (when generated by the same clustering algorithm) or heterogeneous (when different clustering algorithms are employed). Heterogeneous ensembles are the best suited for dealing with the problem of different cluster definitions and, therefore, were adopted in this work. With respect to the consensus function, the main existing methods are based on co-association, graph/hypergraph partitioning, mutual information or re-labeling [40].

One of the first and most popular approaches for clustering ensembles is that proposed by Strehl and Ghosh [38]. In their work, the authors formalize the clustering ensemble problem as a combinatorial optimization problem in terms of shared mutual information. In order to tackle the combinatorial complexity of the problem, they propose three algorithms (consensus functions): CSPA (Cluster-based Similarity Partitioning Algorithm); HGPA (Hyper-Graph Partitioning Algorithm); and MCLA (Meta-Clustering Algorithm). A supra-consensus (SC) function was also devised to select the best partition among the results of the earlier three functions.

The CSPA algorithm starts with the construction of a new similarity matrix over the objects according to the base partitions. Each entry of this matrix denotes the fraction of partitions in which two objects are assigned to the same cluster. This matrix is then employed to cluster the objects using any similarity-based clustering algorithm, delivering the consensus partition. In HGPA, the fusion is treated as a problem of partitioning a hypergraph whose hyperedges represent the clusters of the base partitions. The hypergraph is partitioned by cutting a minimal number of hyperedges. Finally, MCLA considers the fusion of partitions as a problem of finding the correspondence among the clusters of the base partitions.

By other means, Fern and Brodley devised the Hybrid Bipartite Graph Formulation (HBGF) clustering ensemble method, which is also based on graph partitioning [17]. First, HBGF constructs a bipartite graph from the set of base partitions, modeling their objects and clusters as vertices. Then, it partitions the graph using a familiar graph partitioning technique. The resulting division of the objects is deemed as the consensus partition.