Applied Soft Computing 13 (2013) 89–108

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

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

A generalized automatic clustering algorithm in a multiobjective framework

Sriparna Saha [a,*], Sanghamitra Bandyopadhyay [b]

[a] Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
[b] Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India

ARTICLE INFO

Article history:
Received 30 December 2011
Received in revised form 3 July 2012
Accepted 2 August 2012
Available online 6 September 2012

Keywords:
Clustering
Multiobjective optimization (MOO)
Symmetry
Relative neighborhood graph
Multi-center
Automatic determination of number of clusters

ABSTRACT

In this paper a new multiobjective (MO) clustering technique (GenClustMOO) is proposed which can automatically partition the data into an appropriate number of clusters. Each cluster is divided into several small hyperspherical subclusters and the centers of all these small sub-clusters are encoded in a string to represent the whole clustering. For assigning points to different clusters, these local sub-clusters are considered individually. For the purpose of objective function evaluation, these sub-clusters are merged appropriately to form a variable number of global clusters. Three objective functions, one reflecting the total compactness of the partitioning based on the Euclidean distance, the other reflecting the total symmetry of the clusters, and the last reflecting the cluster connectedness, are considered here. These are optimized simultaneously using AMOSA, a newly developed simulated annealing based multiobjective optimization method, in order to detect the appropriate number of clusters as well as the appropriate partitioning. The symmetry present in a partitioning is measured using a newly developed point symmetry based distance. Connectedness present in a partitioning is measured using the relative neighborhood graph concept. Since AMOSA, as well as any other MO optimization technique, provides a set of Pareto-optimal solutions, a new method is also developed to determine a single solution from this set. Thus the proposed GenClustMOO is able to detect the appropriate number of clusters and the appropriate partitioning from data sets having either well-separated clusters of any shape or symmetrical clusters with or without overlaps. The effectiveness of the proposed GenClustMOO in comparison with another recent multiobjective clustering technique (MOCK), a single objective genetic algorithm based automatic clustering technique (VGAPS-clustering), *K*-means and single linkage clustering techniques is comprehensively demonstrated for nineteen artificial and seven real-life data sets of varying complexities. In a part of the experiment the effectiveness of AMOSA as the underlying optimization technique in GenClustMOO is also demonstrated in comparison to another evolutionary MO algorithm, PESA2.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Clustering [1,2] is a popular unsupervised pattern classification technique which partitions the input space into K regions based on some similarity/dissimilarity metric where the value of K may or may not be known *a priori*. The aim of any clustering technique is to evolve a partition matrix $U(X)$ of the given data set X (consisting of, say, n patterns, $X = \{x_1, x_2, \ldots, x_n\}$) such that

$$\sum_{j=1}^{n} u_{kj} \geq 1 \qquad \text{for} \quad k = 1, \ldots, K,$$

$$\sum_{k=1}^{K} u_{kj} = 1 \qquad \text{for} \quad j = 1, \ldots, n, \quad \text{and}$$

$$\sum_{k=1}^{K} \sum_{j=1}^{n} u_{kj} = n.$$

The partition matrix $U(X)$ of size $K \times n$ may be represented as $U = [u_{kj}]$, $k = 1, \ldots, K$ and $j = 1, \ldots, n$, where u_{kj} is the membership of pattern x_j to cluster C_k. In crisp partitioning $u_{kj} = 1$ if $x_j \in C_k$, otherwise $u_{kj} = 0$. Elements of U are real numbers in the interval (0, 1).

Determining the appropriate number of clusters from a given data set is an important consideration in clustering. For this

* Corresponding author. Tel.: +91 8809559190.
E-mail addresses: sriparna.saha@gmail.com, sriparna@iitp.ac.in (S. Saha), sanghami@isical.ac.in (S. Bandyopadhyay).

1568-4946/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2012.08.005

purpose, and also to validate the obtained partitioning, several cluster validity indices have been proposed in the literature. The measure of validity of the clusters should be such that it will be able to impose an ordering of the clusters in terms of their goodness. The classical approach of determining the number of clusters is to apply a given clustering algorithm for a range of K values and to evaluate a certain validity function of the resulting partitioning in each case. The partitioning exhibiting the optimal validity is chosen as the true partitioning. In [3] a genetic clustering technique is proposed which uses a cluster validity index as the objective function. Authors have experimented with several different cluster validity indices. In [4] authors have experimented with three clustering algorithms, hard K-Means, single linkage, and a simulated annealing (SA) based technique, in conjunction with four cluster validity indices, namely Davies–Bouldin index, Dunn's index, Calinski–Harabasz index, and a recently developed index I. In [5] a new cluster validity index named CS-index is developed which is able to detect clusters of different densities. In addition, authors also propose a modified K-means algorithm that can assign more cluster centers to areas with low densities of data than the conventional K-means algorithm does. Some fuzzy logic based cluster validity indices are proposed in [6]. A partitional clustering method based on graph theory and a clustering tendency index are proposed in [7]. The number of clusters and the partition that best fits the data set, are selected according to the optimal cluster tendency index value. In [8], authors have presented an analysis of design principles implicitly used in defining cluster validity indices and reviewed a variety of existing cluster validity indices in the light of these principles. After that authors proposed some remedies to overcome the limitations of the existing indices. Based on these remedies six new cluster validity indices are proposed.

The method of using cluster validity indices for searching the optimal number of cluster number depends on the selected clustering algorithm, whose performance may depend on several factors including the initial values, algorithm's parameters, optimization approach and assumptions regarding the cluster distributions. Similarly, most of the validity measures usually assume a certain geometrical structure in the cluster shapes. But if several different cluster structures exist in the same data set, these have often been found to fail.

The global optimum of these validity functions correspond to the most "valid" solutions. Thus Genetic Algorithms (GAs) have been applied to optimize the validity functions to determine the appropriate number of clusters and the appropriate partitioning of a data set simultaneously [3,9,10]. Simple GA (SGA) [11] or its variants are used as the genetic clustering techniques in [3,9,10]. In [12], a function called Weighted Sum Validity Function (WSVF), which is a weighted sum of the several normalized validity functions, is used for optimization along with a Hybrid Niching Genetic Algorithm (HNGA) to automatically evolve the proper number of clusters from a given data set. Within this HNGA, a niching method is developed to prevent premature convergence by preserving both the diversity of the population with respect to the number of clusters encoded in the individuals and the diversity of the subpopulation with the same number of clusters during the search. In [13], a variable string length GA (VGA) based clustering method (named VGAPS-clustering) is proposed which uses a newly developed point symmetry based distance [14] for assignment of points to different clusters and optimizes a newly developed point symmetry (PS) based cluster validity index, *Sym*-index [15,13]. Use of the PS-distance enables the proposed VGAPS-clustering to evolve the clusters of any shape and size as long as they possess the symmetry property.

1.1. Relevance of multiobjective optimization for clustering

Clustering is considered to be a difficult task as no unambiguous partitioning of the data exists for many data sets. Most of the existing clustering techniques are based on only one criterion which reflects a single measure of goodness of a partitioning. However, a single cluster quality measure is seldom equally applicable for different kinds of data sets with different characteristics. Hence, it may become necessary to simultaneously optimize several cluster quality measures that can capture the different data characteristics. In order to achieve this the problem of clustering a data set has been posed as one of multiobjective optimization in literature. In [16], a multiobjective clustering technique called MOCK is developed which outperforms several single-objective clustering algorithms, a modern ensemble technique, and two other methods of model selection. Although the objectives of [16] are very useful, it can only handle clusters either having hyperspherical shape or "connected" but well-separated structures. It fails for datasets having overlapping clusters which do not contain any hyperspherical shape. Moreover MOCK uses locus-based adjacency representation proposed in [17]. Thus when the number of data points is too large the string length becomes high too and convergence becomes slow.

In this paper we have developed a new multiobjective clustering technique with encoding of cluster centers instead of data points. The technique can detect the appropriate number of clusters and the appropriate partitioning from data sets with many different types of cluster structures. A newly developed simulated annealing based multiobjective optimization technique, AMOSA, is used as the underlying optimization strategy. The concept of "multiple centers" corresponding to each cluster is used in this article. Each cluster is divided into several non-overlapping small hyperspherical sub-clusters and the centers of these sub-clusters are encoded in a string to represent a particular cluster. Three cluster validity indices are optimized simultaneously using the search capability of AMOSA. One of these cluster validity indices reflects the total compactness of a particular partitioning, another represents the total symmetry present in a particular partitioning and the last one measures, in a novel way, the degree of "connectedness" of a particular partitioning.

Any multiobjective optimization technique generates a large number of non-dominated solutions on its final Pareto optimal front. Each of these solutions provides a way of partitioning the particular data set. All these solutions are equally important from the algorithmic point of view, but sometimes the user wants a single solution. Thus in this article we have also developed a new semi-supervised method to identify a single best solution from the set of final Pareto-optimal solutions. The superiority of the proposed *GenClustMOO* in comparison with MOCK, a recently proposed MO clustering technique, a single objective genetic clustering technique VGAPS-clustering [13], K-means clustering technique and single linkage clustering techniques, is shown for nineteen artificial data sets (including most of the data sets used in [16]) and seven real-life data sets of varying complexities. In a part of the experiment, the effectiveness of AMOSA as the underlying optimization technique in *GenClustMOO* is also demonstrated in comparison to another evolutionary MO algorithm, PESA2. In a part of the paper we have also experimented with a second criterion of selecting a single solution from the final Pareto optimal set.

2. The SA based MOO algorithm: AMOSA

Archived multiobjective simulated annealing (AMOSA) [18] is an efficient MO version of the simulated annealing (SA) algorithm. MOO is applied when dealing with the real-world problems where

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات