Applied Soft Computing 29 (2015) 479–486

Contents lists available at ScienceDirect

Applied Soft Computing

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

Simultaneous feature selection and symmetry based clustering using multiobjective framework

Sriparna Saha[a,*], Rachamadugu Spandana[a], Asif Ekbal[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 13 May 2013
Received in revised form
17 November 2014
Accepted 6 December 2014
Available online 21 January 2015

Keywords:
Clustering
Multiobjective optimization (MOO)
Symmetry
Automatic feature selection
Automatic determination of number of clusters

ABSTRACT

In this paper a new framework based on multiobjective optimization (MOO), namely FeaClusMOO, is proposed which is capable of identifying the correct partitioning as well as the most relevant set of features from a data set. A newly developed multiobjective simulated annealing based optimization technique namely archived multiobjective simulated annealing (AMOSA) is used as the background strategy for optimization. Here features and cluster centers are encoded in the form of a string. As the objective functions, two internal cluster validity indices measuring the goodness of the obtained partitioning using Euclidean distance and point symmetry based distance, respectively, and a count on the number of features are utilized. These three objectives are optimized simultaneously using AMOSA in order to detect the appropriate subset of features, appropriate number of clusters as well as the appropriate partitioning. Points are allocated to different clusters using a point symmetry based distance. Mutation changes the feature combination as well as the set of cluster centers. Since AMOSA, like any other MOO technique, provides a set of solutions on the final Pareto front, a technique based on the concept of semi-supervised classification is developed to select a solution from the given set. The effectiveness of the proposed FeaClustMOO in comparison with other clustering techniques like its Euclidean distance based version where Euclidean distance is used for cluster assignment, a genetic algorithm based automatic clustering technique (VGAPS-clustering) using point symmetry based distance with all the features, *K*-means clustering technique with all features is shown for seven higher dimensional data sets obtained from real-life.

1. Introduction

Clustering [1,2] is a well-known technique in the field of unsupervised pattern classification which aims to divide the given data set into *K* number of partitions. Here the value of number of clusters may or may not be known *a priori*. The partitioning is formed based on some similarity or dissimilarity measurements. For clustering a data set usually all the available features of a given data set are used. Feature selection, or subset selection, is the method of reducing dimension in machine learning. It is important for different reasons: first total computation can be reduced if we can reduce the size of dimension. Secondly all the features may not be helpful to classify the data; some may be redundant and irrelevant from the classification point of view. Thus it is needed to determine the most relevant subset of features automatically. In order to address these problems, feature selection is needed both for unsupervised as well as supervised classification problems. There exists number of works that addressed the feature selection problem in a supervised setup [3–6]. But the literature shows that there are very few works which dealt with the problem of feature selection in an unsupervised machine learning framework. In case of unsupervised classification it is very difficult to measure the goodness of a particular feature.

In recent years some works have been reported to solve the unsupervised feature selection problem [7–12]. But most of these techniques pose the feature selection problem as a single objective optimization technique. They have mostly optimized a single cluster quality measure. In recent years, some new approaches have emerged which used multiobjective optimization (MOO) for solving the unsupervised feature selection problem. Morita et al. [13] developed an alternative multiobjective wrapper approach for solving the problem of unsupervised feature selection. They have used the *K*-means clustering technique as the underlying partitioning method and varied the number of clusters in a range. A genetic algorithm based multiobjective optimization technique, NSGA-II, [14] is used as the background optimization strategy and two objective functions are simultaneously optimized, namely

* Corresponding author. Tel.: +91 8809559190; fax: +91 8809559190.
E-mail addresses: sriparna@iitp.ac.in (S. Saha), rspandana@iitp.ac.in (R. Spandana), asif@iitp.ac.in (A. Ekbal).

http://dx.doi.org/10.1016/j.asoc.2014.12.009
1568-4946/

the number of features and the Davies–Bouldin-Index (DB-Index [15]). In 2002, Kim et al. [16] presented a multiobjective approach for wrapper based unsupervised feature selection. They have utilized a multiobjective approach named ELSA (Evolutionary Local Selection Algorithm [17]) as the background optimization strategy. The *K*-means algorithm is utilized as the underlying partitioning technique to determine a partitioning corresponding to a feature combination. ELSA is used to determine both the correct feature combination and the corresponding number of partitions. Four objective functions capturing different partitioning qualities: a function of number of features, the available number of clusters, intra cluster compactness and inter cluster separation, were used for optimization. The problem of feature selection coupled with clustering is also treated as a MOO problem in [18]. Here another evolutionary optimization technique, PESA-II, is used to develop a multiobjective feature selection technique. This technique utilizes Euclidean distance for assignment of points to different clusters and *K*-means as the underlying clustering technique. Thus it can only determine hyper spherical shaped equal sized clusters. In this paper we have posed the problem of feature selection for unsupervised classification as a MOO problem. An existing multiobjective simulated annealing based technique, AMOSA [19] is utilized as the background optimization strategy. Point symmetry based distance [20] is used in place of Euclidean distance for assignment of points to different clusters.

The proposed multiobjective clustering as well as feature selection technique, called FeaClusMOO technique, encodes number of features and number of cluster centers in the form of a string. Then points are assigned based on the features present in the string to different partitions using a symmetry based similarity measurement [20]. Three different objective functions are used for optimization. These are a symmetry based cluster validity index, *Sym*-index [21], Euclidean distance based cluster validity index, XB-index [22] and the number of features. The third objective function is used to balance the bias of the first two objective functions on dimensionality. Internal cluster validation techniques are based on some distance computations and thus those are biased towards lower dimensions [18]. In order to balance these bias we have used the third objective which will try to increase the number of features present in a data set. The final Pareto optimal front contains a set of solutions representing different feature combinations and cluster centers. The algorithm automatically identifies the proper partitioning with correct number of clusters and the effective feature combination from a data set. Results are shown for several higher dimensional real-life data sets. The performance of FeaClusMOO technique is compared with (a) FeaClusMOO using Euclidean distance in place of distance measurement based on the symmetry property for allocation of points to different partitions; (b) an automatic clustering technique utilizing the symmetry property, VGAPS-clustering, where all the features are utilized for distance computation and point symmetry based distance is used for cluster assignment, and (c) *K*-means clustering technique with all features and known number of clusters. In a part of the paper we have also compared three MOO techniques, simulated annealing based multiobjective optimization technique, AMOSA [19], particle swarm optimization based multiobjective optimization technique, AMOPSO [23] and genetic algorithm based multiobjective optimization technique, NSGA-II [14] as the background strategy for optimization of FeaClustMOO technique.

2. Proposed method of simultaneous feature selection and clustering

This section describes the newly proposed multiobjective optimization based feature selection and clustering technique, *FeaClusMOO*, in detail. Note that the proposed framework is a very general one. In order to optimize multiple objective functions, AMOSA [19] is utilized as the underlying optimization strategy. But any other optimization techniques based on genetic algorithm [24] or particle swarm optimization [25] could have also been utilized as the underlying optimization strategies.

Archived multiobjective simulated annealing (AMOSA) [19] is an effective multiobjective version of the simulated annealing (SA) algorithm. It has been shown in [19] that AMOSA performs better than some popular and existing MOO approaches specially for solving 3 or more objectives [19]. Inspired by this observation, in the current paper, it is used as the background optimization strategy.

In the commonly used genetic algorithm based MOO techniques (MOGA) or particle swarm optimization based MOO techniques (MOO-PSO), there was no positive probability involved in accepting the bad solutions. These MOGAs or MOO-PSOs are so designed that they simply discard the bad solutions; but in AMOSA during the execution process there are some positive probabilities in accepting the bad solutions. This is a very important characteristic existing in most of the single objective optimization techniques like genetic algorithm or simulated annealing which helps them to avoid getting stuck at local optima. Due to the presence of this characteristic, AMOSA has been widely used as the MOO technique for solving many real-life problems.

In this section we have described the general framework of simultaneous feature selection and clustering.

2.1. Representation of strings and initialization of archive

In *FeaClusMOO*, a state in AMOSA consists of two items: (a) a set of real numbers which represents the coordinates of the centers of the clusters hence the corresponding grouping of the data and (b) a set of binary values which denotes a feature combination. AMOSA is used to evolve the appropriate set of cluster centers and the appropriate set of feature combination. Suppose a particular string encodes the centers of K number of clusters and the total number of available features is F. Therefore, the length of the string will be $F+K\times F$. The K number of cluster centers are randomly chosen K points from the data set. Here feature combinations are some randomly chosen binary numbers. An example of a string is given below, where $K=3$ and $F=5$: $\langle c_1^1, c_1^2, c_1^3, \ldots, c_1^5, c_2^1, \ldots, c_2^5, c_3^1, c_3^2, \ldots, c_3^5, 11010\rangle$. This represents a partitioning having three cluster centers $\langle c_1^1, c_1^2, c_1^3, \ldots, c_1^5\rangle$, $\langle c_2^1, \ldots, c_2^5\rangle$, and $\langle c_3^1, c_3^2, \ldots, c_3^5\rangle$ and we have to consider only the features: first, second and fourth. These features are considered for cluster assignment and objective function calculation. Each string i in the archive initially contains K_i number of partitions, such that $K_i = (rand()\bmod(K^{max}-1))+2$. Here, $rand()$ is used to denote a function which returns an integer value, and K^{max} is assumed to be the higher limit of the value of partitions. The value of initially encoded partitions will vary between 2 and K^{max}. The procedure used for initializing the cluster centers is fully random. The strings of the archive are created randomly, i.e., for these solutions the K_i number of centers are selected randomly from the data set. Thereafter points are allocated to different clusters using the principles of well-known *K*-means algorithm. The features are also initialized randomly. The features encoded in a string are randomly initialized to either 0 or 1. Here, if the ith position of the feature vector of a given string is 0 then it represents that ith feature does not participate in the cluster assignment. Else, the value 1 indicates that the ith feature participates in the cluster assignment (Fig. 1).

2.2. Assignment of points

Cluster assignments are done based on a symmetry based similarity measurement [20], $d_{ps}(\bar{x}, \bar{c})$ as defined follows.