Expert Systems with Applications 42 (2015) 4239–4252

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Parallel alternatives for evolutionary multi-objective optimization in unsupervised feature selection

Dragi Kimovski [a], Julio Ortega [b], Andrés Ortiz [c,*], Raúl Baños [d]

[a] University of Information, Science & Technology, Ohrid, Macedonia
[b] Dept. Computer Architecture and Technology, CITIC, University of Granada, Spain
[c] Dept. Communications Engineering, University of Malaga, Spain
[d] Dept. Business Administration and Management, Catholic University of Murcia, Spain

A R T I C L E I N F O

Article history:
Available online 4 February 2015

Keywords:
Feature selection
High-dimensional data
Multi-objective clustering
Parallel evolutionary algorithms
Speedup models
Unsupervised classification

A B S T R A C T

Many machine learning and pattern recognition applications require reducing dimensionality to improve learning accuracy while irrelevant inputs are removed. This way, feature selection has become an important issue on these researching areas. Nevertheless, as in past years the number of patterns and, more specifically, the number of features to be selected have grown very fast, parallel processing constitutes an important tool to reach efficient approaches that make possible to tackle complex problems within reasonable computing times. In this paper we propose parallel multi-objective optimization approaches to cope with high-dimensional feature selection problems. Several parallel multi-objective evolutionary alternatives are proposed, and experimentally evaluated by using some synthetic and BCI (Brain–Computer Interface) benchmarks. The experimental results show that the cooperation of parallel evolving subpopulations provides improvements in the solution quality and computing time speedups depending on the parallel alternative and data profile.

1. Introduction

Many relevant applications imply high-dimensional pattern classification or modelling tasks where feature selection techniques must be applied to reduce the dimensionality, and therefore, to remove redundant, noisy-dominated, or irrelevant features. In particular, dimensionality reduction is very important when the number of features in the input pattern is higher than the number of available patterns. Thus, feature selection is important for increasing the learning accuracy and result comprehensibility.

An interesting review on feature selection techniques used in bioinformatics is provided in Saeys, Inza, and Larrañaga (2007) along with analyses and references of feature selection in bioinformatics applications such as sequence analysis, microarray analysis, and mass spectra analysis. For example, one of the problems in sequence analysis is the identification of relevant motifs, by relating them with levels of gene expression through regression models where feature selection is useful to improve the model fitting. In the prediction of protein function from sequence, feature selection techniques can be useful for determining relevant amino acid subsets. Dimension reduction in the input patterns has been also applied to Electroencephalogram (EEG) classification for recognizing epileptiform patterns (Acir & Güzelis, 2004). Precisely, EEG classification has to cope with Lotte, Congedo, Lcuyer, Lamarche, and Arnaldi (2007) (1) the presence of noise or outliers in the features (as EEG signals have a low signal-to-noise ratio); (2) the need to represent time information in the features (as the brain patterns are usually related to changes in time in the EEG signals); (3) the non-stationarity of EEG signals, that may change quickly over time or experiments; and (4) the low number of patterns (EEGs) available for training (as the experimental work required to register the EEGs for different events is time consuming). As the solution to these problems usually implies the increase of the dimensionality of the feature vectors, the classification of EEG signals, for example in BCI applications, has to be accomplished from relatively few feature vectors of very high dimensionality. This circumstance determines the so called curse-of-dimensionality problem, as the number of patterns needed to properly define the different classes increases very fast with the dimension of the feature vectors)from five to ten times as many training samples per class as the dimension (Raudys & Jain, 2014).

This way, feature selection will reduce the dimension of the input patterns to be classified and thus it makes possible to: (1)

* Corresponding author.
E-mail addresses: dragi.kimovski@uist.edu.mk (D. Kimovski), jortega@ugr.es (J. Ortega), aortiz@ic.uma.es (A. Ortiz), rbanos@ucam.edu (R. Baños).

http://dx.doi.org/10.1016/j.eswa.2015.01.061
0957-4174/

decrease the computational complexity of the procedure, (2) remove irrelevant/redundant features that would make the learning of the classifier more difficult, and (3) avoid the curse of dimensionality in problems with many features and a low number of available data to be classified. Nevertheless, as the size of the search space depends exponentially on the number of possible features, an exhaustive search of the best set is not feasible, even for a modest number of features. Procedures based on branch-and-bound, simulated annealing, or evolutionary algorithms have been proposed. Moreover, parallel processing is as an interesting alternative to take advantage of high performance computer architectures for feature selection in high-dimensional cases.

This paper deals with parallel processing in feature selection, considered as a multi-objective optimization problem. In papers such as Emmanouilidis, Hunter, and MacIntyre (2000), Kim, Street, and Menczer (2002), Morita, Sabourin, Bortolozzi, and Suen (2003), Oliveira, Sabourin, Bortolozzi, and Suen (2003), Handl and Knowles (2006), Mierswa and Wurst (2006) and Huang, Buckley, and Kechadi (2010), feature selection for either supervised or unsupervised classification problems has been recently approached as a multi-objective optimization problem. Indeed, in Mierswa and Wurst (2006), it is shown that feature selection in unsupervised learning problems is inherently a multi-objective problem. With respect to the use of parallel processing for feature selection, related papers are Garcia, Hall, Goldgof, and Kramer (2004), De Souza, Matwin, and Japkowitz (2006), Guillén et al. (2009), Zhao, Zhang, Cox, Duling, and Sarle (2013) and Sun (1991). In Sun (1991), a MapReduce model is used to obtain a parallel implementation of a feature selection procedure based on mutual information to evaluate the statistical dependence between variables. Parallel feature selection based on a forward–backward algorithm applied to a k-nearest neighbours clustering method, to separate genuine and non-genuine images in stenography problems, is shown in Guillén et al. (2009). In Zhao et al. (2013), a large-scale feature selection algorithm based on the abilities of the features to explain the data variance is proposed. Random features selection implemented in parallel is compared with a wrapper method in Garcia et al. (2004), where Support Vector Machines (SVMs) are used as the supervised learning algorithm for classification. Finally, De Souza et al. (2006) compare parallel implementations of different feature selection procedures previously proposed (including a genetic algorithm) with the FortalFS algorithm presented in the paper. A description of the essential characteristics to generate efficient parallel procedures from sequential ones is also provided by De Souza et al. (2006). Hence, these previous papers have not considered feature selection from a parallel multi-objective approach as we propose in this paper. The reason to use a multi-objective formulation of the feature selection problem is that the performance of a classifier is usually expressed not only by its accuracy for a given set of patterns but also by other measures that quantify properties such as the generalization capability. In this way, a multi-objective formulation can be considered a straightforward approach for feature selection. Along with the previous justifications given for using parallel processing and multi-objective optimization, the reasons for using unsupervised classification should be also provided. In this case, it has been taken into account that in many classification problems, the patterns are not labelled. Moreover, unsupervised learning is mandatory whenever the number of classes is unknown or to extract unknown relations among the features, thus being also a suitable approach to select the best set of features.

Once the relevance of feature selection and the reasons for a parallel multi-objective unsupervised learning approach to this problem have been summarized in this introduction, Section 2 deals with the formulation of feature selection as an unsupervised multi-objective optimization problem, while Sections 3 and 4 are devoted to its parallel implementation. More specifically, Section 3 reviews the parallel processing issues of evolutionary multi-objective optimization including references to previous works and Section 4 describes the parallel procedures we propose for unsupervised feature selection. The results obtained from the experiments performed on different benchmarks are presented and discussed in Section 5. Finally, the conclusions are given in Section 6.

2. Multiobjective optimization in unsupervised feature selection

Approaches for dimensionality reduction can be classified into two main alternatives, (1) feature space transformation through linear or non-linear transformations, and (2) the selection of a subset of features. In this paper, we consider feature selection, which can be defined as the search of a set of features which optimizes a cost function that evaluates the utility of these features according to the classifier performance once it has been trained with patterns whose components are the selected features. These are the so called *wrapper* approaches for feature selection.

As it has been previously stated in Section 1, the formulation of feature selection as a multi-objective optimization problem has been proposed by authors such as Emmanouilidis et al. (2000), Kim et al. (2002), Morita et al. (2003), Oliveira et al. (2003), Handl and Knowles (2006), Mierswa and Wurst (2006) and Huang et al. (2010). A multi-objective optimization problem can be defined as the problem of finding a vector of decision variables $x \in \mathbb{R}^n$, $x = [x_1, x_2, \ldots, x_n]$, that satisfies a restriction set, $g(x) \leqslant 0$, $h(x) = 0$, and optimizes a function vector $f(x)$, whose scalar values $(f_1(x), f2(x), \ldots, f_m(x))$ represent the objectives to be optimized. These objectives are usually in conflict, and the concept of optimum must be redefined in this context. Thus, instead of providing only one optimal solution, the procedures applied to these multi-objective optimization problems should obtain a set of non-dominated solutions, known as Pareto optimal solutions (Goldberg, 1989), from which a decision maker will choose the most convenient solution according to the circumstances. These Pareto optimal solutions are optimal in the sense that in the corresponding hyper-area known as *Pareto front*, there is not any solution worse than any of the other ones when all the objectives are taken into account.

The formulation of feature selection as a multiobjective optimization problem could provide some advantages that depend on whether the classification procedure is either supervised or unsupervised (Handl & Knowles, 2006). In supervised classification procedures, the goal is usually the maximization of the classifier performance while the number of features is minimized as larger feature sets could produce overfitting and low generalization problems. Thus, a multi-objective optimization approach that takes into account the classifier performance and the number of features adequately allows a satisfactory formulation of this goal.

The situation in unsupervised classification problems is different. In this case, the utility should be determined from a definition of clustering quality without having knowledge about the corresponding labels, or even the number of clusters. Frequently, the clustering quality measures use ratios between intra-cluster compactness measures and inter-cluster separation ones. Nevertheless, the distances between points tend to similar values as dimensionality increases and these quality solutions are biased towards lower dimension solutions (Handl & Knowles, 2006). This way, as the applied validation techniques usually present a dimensionality-bias to either smaller or larger cardinality feature sets, a