Available online at www.sciencedirect.com

SCIENCE DIRECT®

European Journal of Operational Research 169 (2006) 477–489

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/ejor

Solving feature subset selection problem by a Parallel Scatter Search ☆

Félix García López, Miguel García Torres, Belén Melián Batista, José A. Moreno Pérez *, J. Marcos Moreno-Vega

Dpto. de Estadística, Investigación Operativa y Computación, Escuela Técnica Superior de Ingeniería Informática, Universidad de La Laguna, 38271 La Laguna, Spain

Available online 7 October 2004

Abstract

The aim of this paper is to develop a Parallel Scatter Search metaheuristic for solving the Feature Subset Selection Problem in classification. Given a set of instances characterized by several features, the classification problem consists of assigning a class to each instance. Feature Subset Selection Problem selects a relevant subset of features from the initial set in order to classify future instances. We propose two methods for combining solutions in the Scatter Search metaheuristic. These methods provide two sequential algorithms that are compared with a recent Genetic Algorithm and with a parallelization of the Scatter Search. This parallelization is obtained by running simultaneously the two combination methods. Parallel Scatter Search presents better performance than the sequential algorithms.

Keywords: Scatter Search; Feature subset selection; Parallelization; Metaheuristics

1. Introduction

In a classification problem, the goal is to classify instances that are characterized by a set of features. Then, the class to which each instance belongs is determined. In supervised machine learning, an induction algorithm is typically presented with a set of training instances (examples, cases), where each instance is defined by a vector of features and a class label. The task of the induction algorithm is to induce a classifier that will be used to classify future cases. The classifier is a

☆ This research has been partially supported by the projects TIC2002-04242-C03-01 (70% of which are FEDER founds) and TIC2002-10886E.

* Corresponding author. Tel.: +34 922 318186; fax: +34 922 318170.

E-mail addresses: jamoreno@ull.es (J.A. Moreno Pérez), fcgarcia@ull.es (F. García López), mgarciat@ull.es (M. García Torres), mbmelian@ull.es (B. Melián Batista), jmmoreno@ull.es (J.M. Moreno-Vega).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2004.08.010

mapping from the space of feature values to the set of class labels. Since, in practical applications, the set of features can be very large, in order to classify future instances, it is important to select a smaller subset of these features. As pointed in [10], this dimensionality reduction has several advantages: *a reduction in the cost of acquisition of the data*, *improvement of the comprehensibility of the final classification model*, *a faster induction of the final classification model and an improvement in classification accuracy*.

The feature subset selection problem consists of finding a subset of the original set of features, such that an induction algorithm using only these features is able to generate a classifier with the best performance. Selecting the optimal feature subset is an *NP*-hard optimization problem [12]. Therefore exact algorithms should not be used due to the complexity of the problem. For example, determining the optimal binary decision tree is an *NP*-hard problem [9]. Several heuristic algorithms have been proposed for solving the feature subset selection problem. One of the most widely used metaheuristics are the Genetic Algorithms. These algorithms have been proposed and analysed for the feature subset selection problem in [7,17,18,21]. The obtained results show that Genetic Algorithms are appropriate methods for this problem. We propose the use of another evolutive metaheuristic (Scatter Search [13]) to solve this problem and compare one of our proposed Scatter Search procedures with a recent Genetic Algorithms. We have not found any reference about the application of Scatter Search to the feature selection.

Two different approaches for selecting the subset of features can be considered: the wrapper and filter approaches. The filter approach selects the features using a preprocessing step that ignores the induction algorithm. The main disadvantage of this procedure is that it ignores the effect of the subset of features in the induction algorithm. Two filter-based algorithms are RELIEF [11] and FOCUS [3]. The first one assigns a weight to each feature according to its relevance for classifying. To do so it samples several examples randomly and compares the example with the two nearest examples of the same and opposite class. The latter algorithm examines all subsets of features by selecting the minimal subset of features that is sufficient to classify the examples.

In the wrapper approach, the induction algorithm selects the optimal subset of features by itself. Two well known wrapper approaches are forward subset selection (FSS) and backward subset selection (BSS) [5]. FSS starts with an empty subset of features and, at each step, it adds to the subset the feature that most improves the classification. This process is iterated until no improvement is possible. In BSS the initial subset consists of all the available features and, at each step, the worst feature is eliminated from the subset. As in FSS, this process is repeated until no improvement is possible. The Parallel Scatter Search proposed in this paper is based on the wrapper approach.

We consider three paradigms of learning: the *Instance-Based Learning* approach, the *Bayesian Learning procedures*, and the *Decision Tree* methods [14]. The first approach uses the nearest examples to predict the label of the instance, given a set of examples and an instance to be classified. In particular, we use the instance-based algorithm called IB1 [2], that classifies each instance with the label of the nearest example. For the purpose of classifying each instance with the label of the nearest example, IB1 considers all the features, although, in general, only a few of them are highly relevant. The Bayesian Learning algorithms use probability as an approach for classification. The Naive Bayes classifier consists in using Bayes theorem to estimate "a posteriori" probabilities of all possible classifications. For each instance, the classification with the highest "a posteriori" probability is chosen. Decision trees classify instances by testing the instance at each node it reaches. The procedure starts at the root and, at each node, moving down the trees branch according to the result of the test. Leaf nodes give the classification of all instances that reach the leaf. We use the top–down induction decision tree algorithm C4.5 devised by Quinlan [16]. The C4.5 algorithm is an improvement of the classical ID3 (Interactive Dichotomer 3) algorithm for constructing decision trees.