Information Sciences 326 (2016) 102–118

Contents lists available at ScienceDirect

Information Sciences

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

High-dimensional feature selection via feature grouping: A Variable Neighborhood Search approach

Miguel García-Torres [a,*], Francisco Gómez-Vela [a], Belén Melián-Batista [b], J. Marcos Moreno-Vega [b]

[a] Área de Lenguajes y Sistemas Informáticos, Universidad Pablo de Olavide, Ctra de Utrera, km. 1, Sevilla 41013, Spain
[b] Dpto. de Ingeniería Informática y de Sistemas, Universidad de La Laguna, La Laguna 38271, Spain

ARTICLE INFO

Article history:
Received 30 October 2014
Revised 15 July 2015
Accepted 17 July 2015
Available online 26 July 2015

Keywords:
Feature selection
High dimensionality
Metaheuristic
Feature grouping

ABSTRACT

In recent years, advances in technology have led to increasingly high-dimensional datasets. This increase of dimensionality along with the presence of irrelevant and redundant features make the feature selection process challenging with respect to efficiency and effectiveness. In this context, approximate algorithms are typically applied since they provide good solutions in a reasonable time. On the other hand, feature grouping has arisen as a powerful approach to reduce dimensionality in high-dimensional data. Recently, some authors have focused their attention on developing methods that combine feature grouping and feature selection to improve the model. In this paper, we propose a feature selection strategy that utilizes feature grouping to increase the effectiveness of the search. As feature selection strategy, we propose a Variable Neighborhood Search (VNS) metaheuristic. Then, we propose to group the input space into subsets of features by using the concept of Markov blankets. To the best of our knowledge, this is the first time in which the Markov blanket is used for grouping features. We test the performance of VNS by conducting experiments on several high-dimensional datasets from two different domains: microarray and text mining. We compare VNS with popular and competitive techniques. Results show that VNS is a competitive strategy capable of finding a small size of features with similar predictive power than that obtained with other algorithms used in this study.

1. Introduction

Feature selection for classification has become an increasingly important research area within machine learning and pattern recognition [42,44,45,59] due to rapid advances in data collection and storage technologies. These advances have allowed organizations from science and industry to create large, high-dimensional, complex and heterogeneous datasets that represent a new challenge to the existing methods in the feature selection field.

In high-dimensional spaces, in addition to the curse of dimensionality, the learning task suffers from the fact that usually not all the features have the same discriminative power. Moreover, as the number of dimensions becomes larger, not only the complexity of the datasets increases, but also the number of non informative features with respect to the class concept may increase, because of irrelevancy and redundancy [71]. In this context, feature selection plays a critical role for removing such

* Corresponding author. Tel.: +34 954977366.
E-mail addresses: mgarciat@upo.es (M. García-Torres), fgomez@upo.es (F. Gómez-Vela), mbmelian@ull.es (B. Melián-Batista), jmmoreno@ull.es (J.M. Moreno-Vega).

http://dx.doi.org/10.1016/j.ins.2015.07.041
0020-0255/

features and may yield some of the following benefits: (i) reduction in the cost of acquisition of the data, (ii) improvement of the comprehensibility of the final classification model, (iii) a faster induction of the final classification model and (iv) an improvement in classification accuracy.

Classically, feature selection in classification tasks is defined as the process that seeks the minimal size of relevant features such that the classification error is optimized. A relevant feature is neither irrelevant nor redundant to the target concept; an irrelevant feature does not affect the target concept in any way, and a redundant feature does not add anything new to the target concept [13]. In order to identify the optimal subset of relevant features, different criteria have been proposed to evaluate the goodness of feature subsets. Feature subset selection strategies are essentially divided into wrapper, filter and embedded methods [9,23]. Wrappers use the learner as a black box to score the subsets of features according to their predictive power. Thus, the quality of feature subsets for classification is defined with respect to the induction algorithms. The main advantage is that they include the interaction between feature subset and model selection, and have the ability to take into account feature dependencies. However, they have a higher risk of overfitting than filters and are computationally expensive [9]. Filter approaches select subsets of features as a preprocessing step, and so assess each subset according to intrinsic properties of the data. The advantages of these methods are that they are computationally fast, so that they easily scale to high-dimensional datasets [57]. A disadvantage is that they ignore the interaction with the classifier, which may lead to worse classification performance. Since they are independent of the learning algorithm, feature selection needs to be performed only once for a given training dataset. In contrast to the filter and wrapper techniques, the embedded methods cannot separate the learning and the feature selection since the structure of the class of functions under consideration play a crucial role. In this approach, the search for an optimal subset of features is done during the induction of the classifier. The main advantage of these methods is that they combine the interaction with the classification model such as the wrapper methods. However, they are far less computationally intensive than wrappers [28,58,75].

Selecting the most relevant features is usually suboptimal for building the model due to redundancy [23]. Despite the recent achievements carried out in the field of feature selection, feature relevance and redundancy are still two challenging issues in the field. Researchers firstly focused on identifying relevant features [3,5,8,31,36,73]. Then they also focused on redundancy [18,52,53,69,77], especially in high dimensional data [19]. Furthermore, the number of possible feature subsets grows exponentially with the number of features and many problems related to feature selection have been shown to be $\mathcal{NP}-hard$ [6]. For all these reasons, finding the optimal subset is usually intractable [35] even for a moderate number of features d. Therefore, approximate algorithms are typically applied since they provide satisfactory solutions in a reasonable time (see, for example, [20,26,41]). Even if the obtained solution is suboptimal and there is no guarantee of the distance between such solution and the optimal one, in general, they provide satisfactory solutions in a reasonable computational time.

The idea of feature clustering or feature grouping is a powerful approach for reducing the dimensionality. Moreover, the grouping of features is highly beneficial in learning with high-dimensional data. It can reduce the variance of the estimator [60], improve the stability of feature selection [32], and also helps to reduce the complexity of the model. As far as we know, it has been applied to text mining [2,50,54,64] and microarray [4,15,43,72,74] domains since the late 90s. For finding feature groups, some approaches use learning algorithms like self-organizing map [62], K-means [74], or a reminiscent of it [17], logistic regression [16], etc. Other techniques make use of information–theory measures [37,65], graph theory [65], kernel density estimation [76], and regularization techniques [68].

Approaches based on regularization techniques are worth mentioning. They are important embedded methods that attract increasing attention due to their good performance. These methods introduce additional constraints into the objective function. In effect, the model fits the data by minimizing the coefficients. Hence, features with coefficients that are close to 0 are then eliminated [49]. Some representative methods based on regularization techniques are: (a) the Lasso Regularization [67] based on l_1-norm, (b) Adaptive Lasso [78], which was proposed to improve the performance of the Lasso proposal, (c) Bridge regularization [29] and (d) Elastic net regularization [79] that is a mixture of bridge regularization (see [66] for more details).

Recently, some works focus their efforts on grouping correlated features. This approach produces feature selection results in the form of a set of feature groups, each consisting of features relevant to the class but highly correlated to each other, instead of the traditional form of a single subset of features. The main motivation of this approach lies on the key observation that in high-dimensional data, relevant features are highly correlated so that we can generate groups of correlated features that are resistant to the variations of the sample size. Such set of predictive feature groups, not only generalize well, but also provides additional informative group structure for expert domains to further investigate. Recently, two group-based feature selection frameworks were proposed to improve the robustness by identifying groups of correlated features [47,76]. In [76] the authors introduce the idea of Dense Feature Groups (DFG) based on Kernel Density Estimation (KDE). KDE is a popular non-parametric method for estimating probabilistic density functions and it is applied to estimate the density of the features. Therefore, DFG is composed of features which are close to the same density peak. This framework is motivated by two main observations in the sample space. Firstly, the dense core regions (peak regions), measured by probabilistic density estimation, are stable with respect to sampling of the dimensions (samples). Secondly, the features near the core region are highly correlated to each other, and thus should have similar relevance scores w.r.t. some class labels, assuming that the class labels are locally consistent. Under this framework, an algorithm named DRAGS (Dense Relevant Attribute Group Selector) was proposed, which finds a number of dense feature groups and evaluates the relevance of each group based on the average relevance of features in each group. A novel framework, called CGS (Consensus Group Stable Feature Selection) was proposed in [47]. This proposal, identifies consensus feature groups by subsampling training samples. In order to do this, the proposed approach approximates intrinsic feature groups by a set of