Pattern Recognition 47 (2014) 2481–2489

Contents lists available at ScienceDirect

Pattern Recognition

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

A framework for cost-based feature selection

V. Bolón-Canedo*, I. Porto-Díaz, N. Sánchez-Maroño, A. Alonso-Betanzos

Laboratory for Research and Development in Artificial Intelligence (LIDIA), Computer Science Department, University of A Coruña, 15071 A Coruña, Spain

ARTICLE INFO

Article history:
Received 19 July 2012
Received in revised form
15 November 2013
Accepted 21 January 2014
Available online 28 January 2014

Keywords:
Cost-based feature selection
Machine learning
Filter methods

ABSTRACT

Over the last few years, the dimensionality of datasets involved in data mining applications has increased dramatically. In this situation, feature selection becomes indispensable as it allows for dimensionality reduction and relevance detection. The research proposed in this paper broadens the scope of feature selection by taking into consideration not only the relevance of the features but also their associated costs. A new general framework is proposed, which consists of adding a new term to the evaluation function of a filter feature selection method so that the cost is taken into account. Although the proposed methodology could be applied to any feature selection filter, in this paper the approach is applied to two representative filter methods: Correlation-based Feature Selection (CFS) and Minimal-Redundancy-Maximal-Relevance (mRMR), as an example of use. The behavior of the proposed framework is tested on 17 heterogeneous classification datasets, employing a Support Vector Machine (SVM) as a classifier. The results of the experimental study show that the approach is sound and that it allows the user to reduce the cost without compromising the classification error.

1. Introduction

The proliferation of high-dimensional data has become a trend in the last few years. Datasets with a dimensionality over the tens of thousands are constantly appearing in applications such as medical image and text retrieval or genetic data. In fact, analyzing the dimensionality of the datasets posted in the UCI Machine Learning Repository [1] in the last decades, one can observe that in the 1980s, the maximum dimensionality of the data is about 100, increasing to more than 1500 in the 1990s; and finally in the 2000s, it further increases to about 3 million [2].

The high-dimensionality of data has an important impact in learning algorithms, since they degrade their performance when a number of irrelevant and redundant features are present. In fact, this phenomenon is known as the *curse of dimensionality* [3], because unnecessary features increase the size of the search space and make generalization more difficult. For overcoming this major obstacle in machine learning, researchers usually employ dimensionality reduction techniques. In this manner, the set of features required for describing the problem is reduced, most of the times along with an improvement in the performance of the models. *Feature selection* is arguably the most famous dimensionality reduction technique. It consists of detecting the relevant features and discarding the irrelevant ones. Its goal is to obtain a subset of features that describe properly the given problem with a minimum degradation in performance [4], with the implicit benefits of improving data and model understanding and the reduction in the need for data storage. With this technique, the original features are maintained, contrary to what usually happens in other techniques such as feature extraction, where the generated dataset is represented by a newly generated set of features, different than the original.

Feature selection methods can be divided into wrappers, filters and embedded methods [4]. While wrapper models involve optimizing a predictor as a part of the selection process, filter models rely on the general characteristics of the training data to select features with independence of any predictor. The embedded methods generally use machine learning models for classification, and then an optimal subset or ranking of features is built by the classifier algorithm. Wrappers and embedded methods tend to obtain better performances but at the expense of being very time consuming and having the risk of overfitting when the sample size is small. On the other hand, filters are faster and, therefore, more suitable for large datasets. They are also easier to implement and scale up better than wrapper and embedded methods. As a matter of fact, filters can be used as a pre-processing step before applying other more complex feature selection methods. For all these reasons, filters will be the focus of this work.

There is a broad suite of filter methods, based on different metrics, but the most common approaches are to find either a subset of features that maximizes a given metric or either an ordered ranking of the features based on this metric. Two of the

* Corresponding author at: Department of Computer Science, Facultade de Informática, Campus de Elviña s/n, University of A Coruña, 15071 A Coruña, Spain. Tel. +34 981 167 000x1305; fax: +34 981 167 160.

E-mail addresses: vbolon@udc.es (V. Bolón-Canedo), iporto@udc.es (I. Porto-Díaz), nsanchez@udc.es (N. Sánchez-Maroño), ciamparo@udc.es (A. Alonso-Betanzos).

0031-3203/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2014.01.008

most popular filter metrics for classification problems are correlation and mutual information, although other common filter metrics include error probability, probabilistic distance, entropy or consistency [5].

There are some situations where a user is not only interested in maximizing the merit of a subset of features, but also in reducing costs that may be associated to features. For example, for medical diagnosis, symptoms observed with the naked eye are costless, but each diagnostic value extracted by a clinical test is associated with its own cost and risk. In other fields, such as image analysis, the computational expense of features refers to the time and space complexities of the feature acquisition process [6]. This is a critical issue, specifically in real-time applications, where the computational time required to deal with one or another feature is crucial, and also in the medical domain, where it is important to save economic costs and to also improve the comfort of a patient by preventing risky or unpleasant clinical tests (variables that can be also treated as costs).

The goal of this research is to obtain a trade-off between a filter metric and the cost associated to the selected features, in order to select relevant features with a low associated cost. A general framework to be applied together with the filter approach is introduced. In this manner, any filter metric can be modified to have into account the cost associated to the input features. In this paper, and for the sake of brevity, two implementations of this framework will be presented as an example of use, choosing two representative and widely used filters: Correlation-based Feature Selection (CFS) and Minimal-Redundancy-Maximal-Relevance (mRMR). The results obtained with these two filters are promising, showing that the approach is sound.

The rest of the paper is organized as follows: Section 2 summarizes previous research on the subject; Section 3 describes the proposed method in detail; Sections 4 and 5 describe the experimental study performed and the results obtained, respectively; and finally, Section 6 presents the conclusions and the future work.

2. Background

Feature selection has been an active and effective tool in numerous fields such as DNA microarray analysis [7,8], intrusion detection [9,10], medical diagnosis [11] or text categorization [12]. New feature selection methods are constantly appearing, however, the great majority of them only focus on removing irrelevant and redundant features but not on the costs for obtaining the input features.

The cost associated to a feature can be related to different concepts. For example, in medical diagnosis, a pattern consists of observable symptoms (such as age and sex) along with the results of some diagnostic tests. Contrary to observable symptoms, which have no cost, diagnostic tests have associated costs and risks. For example, an invasive exploratory surgery is much more expensive and risky than a blood test [13]. Another example of the risk of extracting a feature can be found in [14], where for evaluating the merits of beef cattle as meat producers is necessary to carry out zoometry on living animals.

On the other hand, the cost can also be related to computational issues. In the medical imaging field, extracting a feature from a medical image can have a high computational cost. For example, in the texture analysis technique known as co-occurrence features [15], the computational cost for extracting each feature is not the same, which implies different computational times. In other cases, such as real-time applications, the space complexity is negligible, but the time complexity is very important [6].

As one may notice, features with an associated cost can be found in many real-life applications. However, this has not been the focus of much attention for machine learning researchers. As mentioned in Section 1, the purpose of this research is to propose a general framework to the problem of cost-based feature selection, trying to balance the correlation of the features with the class and their cost. There have been similar attempts to balance the contribution of different terms in other areas. For instance, in classification, Friedman et al. [16] included a regularization term to the traditional Linear Discriminant Analysis (LDA). The left side term of their cost function evaluates the error and the right side term would be the regularization one, which is weighted with λ. This provides a framework in which, according to the λ value, different regularized solutions can be obtained. Related to feature extraction, in [17] a criterion is proposed to select kernel parameters based on maximizing between-class scattering and minimizing within-class scattering. Applied to face recognition, Wright et al. [18] proposed a general classification framework to study feature extraction and robustness to occlusion via obtaining a sparse representation. Instead of measuring the correlation between a feature and the class, this method evaluates the representation error. However, our objective is completely different, as it is to provide a framework for feature selection where features with an inherent cost could be dealt with.

Despite the previous attempts in classification and feature extraction, to the best knowledge of the authors, there are only a few attempts to deal with this issue in feature selection. In the early 1990s, Feddema et al. [6] were developing methodologies for the automatic selection of image features to be used by a robot. For this selection process, they employed a weighted criterion that took into account the computational expense of features, i.e. the time and space complexities of the feature extraction process. Several years later, Yang and Honavar [13] proposed a genetic algorithm to perform feature subset selection where the fitness function combined two criteria: the accuracy of the classification function realized by the neural network and the cost of performing the classification (defined by the cost of measuring the value of a particular feature needed for classification, the risk involved, etc.). A similar approach was presented in [19], where a genetic algorithm is used for feature selection and parameters optimization for a support vector machine. In this case, classification accuracy, the number of selected features and the feature cost were the three criteria used to design the fitness function. Another proposal can be found in [20] by presenting a hybrid method for feature subset selection based on ant colony optimization and artificial neural networks. The heuristic that enables ants to select features is the inverse of the cost parameter.

The methods found in the literature that deal with cost associated to the features, which were described above, have the disadvantage of being computationally expensive by having interaction with a classifier, which prevents their use in large databases, a trending topic in the past few years [21]. However, the general framework proposed in this paper is applied together with the filter model, which is known to have a low computational cost and be independent of any classifier. By being fast and with a good generalization ability, filters using this cost-based feature selection framework will be suitable for application to databases with a great number of input features like microarray DNA data.

In light of the above, the novelty of our paper lies in that there does not exist too much research in cost-based feature selection methods. As a matter of fact, no cost methods can be found in the most popular machine learning and data mining tools. For instance, in Weka [22] we can only find some methods that address the problem of cost associated to the instances (not to the features), and they were incorporated in the latest release. RapidMiner [23] does in fact include some methods that take cost into account, but they are quite simple. One of them selects the attributes that have a cost