Neurocomputing 135 (2014) 13–20

Contents lists available at ScienceDirect

Neurocomputing

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

Data classification using an ensemble of filters

V. Bolón-Canedo *, 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 22 October 2012
Received in revised form
15 February 2013
Accepted 20 March 2013
Available online 4 January 2014

Keywords:
Ensemble learning
Feature selection
Classification
Microarray data

ABSTRACT

Ensemble learning has been the focus of much attention, based on the assumption that combining the output of multiple experts is better than the output of any single expert. Many methods have been proposed of which bagging and boosting were the most popular. In this research, the idea of ensembling is adapted for feature selection. We propose an ensemble of filters for classification, aimed at achieving a good classification performance together with a reduction in the input dimensionality. With this approach, we try to overcome the problem of selecting an appropriate method for each problem at hand, as it is overly dependent on the characteristics of the datasets. The adequacy of using an ensemble of filters rather than a single filter was demonstrated on synthetic and real data, paving the way for its final application over a challenging scenario such as DNA microarray classification.

1. Introduction

Classically, machine learning methods have used a single learning model to solve a given problem. However, the technique of using multiple prediction models for solving the same problem, known as *ensemble learning*, has proven its effectiveness over the last few years [1]. The idea builds on the assumption that combining the output of multiple experts is better than the output of any single expert. Typically, ensemble learning has been applied to classification, where the most popular methods are bagging [2] and boosting [3] due to their theoretical performance guarantees and robust experimental results. However, it can be also thought as a means of improving other machine learning disciplines such as *feature selection*.

Feature selection is a vital part of the preprocessing stage in machine learning. It consists of identifying and removing irrelevant and redundant features from the training data, so that the learning algorithm focuses only on those aspects of the training data useful for analysis and future prediction [4]. This reduction in the input dimensionality involves, most of the time, an improvement in the performance. Among the different feature selection methods available, this research will be based on the filter approach, because it allows for reducing the dimensionality of the data without compromising the time and memory requirements of machine learning algorithms.

Feature subset selection was employed as a useful technique for creating diversity in classification ensembles. In this case, diversity was incorporated as an objective in the search for obtaining the best collection of feature subsets. While traditional feature selection algorithms have as goal to find the best subset for both the learning task and the selected inductive learning algorithms, the aim of this ensembled feature selection was additionally finding a set of feature subsets that promote disagreement among the base classifiers [5]. Ho [6] has shown that simple random selection of feature subsets may be an effective technique for ensemble feature selection because the lack of accuracy in the ensemble members is compensated by their diversity. Optiz [7] describes an ensemble feature selection technique for neural networks called *Genetic Ensemble Feature Selection* and another ensemble method for decision trees is called *Stochastic Attribute Selection Committees* [8]. More recently, Aly et al. [9] proposed several novel variations to the basic feature subset ensembles present in the literature, trying to improve their results. Finally, in [10] a large-scale analysis of ensemble feature selection was conducted to show their adequacy over biomarker selection.

However, our idea of ensemble feature selection is a little different. Real life datasets come in diverse flavors and sizes, and so their nature imposes several substantial restrictions for both learning models and feature selection algorithms [11]. Datasets may be very large in samples and number of features, and also there might be problems with redundant, noisy, multivariate and non-linear scenarios. Thus, most existing methods alone are not capable of confronting these problems, and something like “the best feature selection method” simply does not exist in general, making it difficult for users to select one method over another. In order to make a correct choice, a user not only needs to know the domain well and the characteristics of each dataset,

* Corresponding author.
E-mail addresses: vbolon@udc.es (V. Bolón-Canedo), nsanchez@udc.es (N. Sánchez-Maroño), ciamparo@udc.es (A. Alonso-Betanzos).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.03.067

but also is expected to understand technical details of available algorithms [12].

So, the idea is to use an ensemble of filters to induce diversity, instead of a single method. In [13] this idea was introduced by proposing an ensemble which combined filters and classifiers obtaining a classification prediction for each of them and deciding on a final result by simple voting. In this paper, further research in ensembles of filters is presented, exploring new ensembles that could improve performance. The objective is again to introduce diversity and increase the stability of the feature selection process, since it takes advantage of the strengths of the single selectors and overcomes their weak points. A total of five configurations for the ensemble of filters are proposed in this research, which are tested using four different classifiers. Experimental validation of the methodology on synthetic data shows the adequacy of the proposed ensembles, paving the way to their application on real and DNA microarray data obtaining high level performance results.

2. The proposed filter ensemble approaches

When dealing with ensemble feature selection, a typical practice is to use different features for each of the base classifiers, i.e. the whole set of features is distributed into the instances of the classifier, which implies almost exhausting the features. However, with the ensemble proposed herein, not all the features have to be necessarily employed. The idea of this research consists of applying several filters based on different metrics so as to have a diverse set of selections. This diversity is the key to this approach since, for a specific dataset, employing one or another filter varies the selected subset of features and, consequently, the performance result obtained by a machine learning algorithm. By using this ensemble of filters, the user is released from the task of choosing an adequate filter for each scenario, because this approach obtains acceptable results independent of the characteristics of the data.

Among the broad suite of filters available in the literature, five filters were selected according to a previous study [13], all of them based on different metrics. Thus, the proposed ensemble will be formed by the filters Correlation-based Feature Selection (CFS) [14], Consistency-based Filter [15], INTERACT [16], Information Gain (IG) [17] and ReliefF [18]. Two distinct general approaches are proposed: Ensemble1 and Ensemble2 (see Fig. 1). The main difference between them are that the former uses several filters and classifies once for each filter, as an integration method for the outputs of the classifier is necessary, whilst the later uses several filters, combines the different subsets returned by each filter, and finally obtains a classification output for this unique subset of features.

2.1. Ensemble1

Within this approach (see Fig. 1a), each one of the F filters selects a subset of features and this subset is used for training a given classifier. Therefore, there will be as many outputs as filters were employed in the ensemble (F). Due to the different metric the filters are based on, they select different sets of features leading to classifier outputs that could be contradictory, so an integration method becomes necessary. Note that in each execution F filters and only one classifier are used, but the classifier is trained F times (once for each filter). More details can be found in [13,19] and the pseudo-code is shown in Algorithm 1. Different variants of this philosophy will be implemented regarding the combination of the F outputs. Two different methods are considered, producing two implementations of Ensemble1. The first uses the well-known simple voting (E1-sv), where for a particular instance, each classifier votes for a class and the class with the greatest number of votes is considered the output class. The second implementation (E1-cp) stores the probability with which an instance has been assigned to a class. The class with the highest cumulative probability is considered the output class.

Algorithm 1. Pseudo-code for *Ensemble1*.

1. F:=number of filters
   (a) for each f from 1 to F
      i. select attributes A using filter f
      ii. build classifier C_f with the selected attributes A
      iii. obtain prediction P_f from classifier
   (b) apply a combination method over predictions $P_1 \dots P_f$
   (c) obtain prediction P

Instead of using the same classifier for all five filters, one may think that there are classifiers more suitable for certain feature selection methods. In fact, in [20] states that CFS, Consistency-based, INTERACT and InfoGain select a small number of relevant features, whilst ReliefF is very effective at removing redundancy. On the other hand, IB1 [21] and SVM [22] deteriorate their performance when irrelevant features are present whereas naive Bayes [23] is robust with respect to irrelevant features but deteriorates with redundant ones. In this situation, the authors propose to try an ensemble which uses naive Bayes together with ReliefF and IB1 with the remaining filters (E1-ni) and another which uses again naive Bayes together with ReliefF and SVM with the remaining filters (E1-ns). Both these configurations can be seen in Fig. 2.

2.2. Ensemble2

This approach consists of combining the subsets selected by each one of the F filters obtaining only one subset of features. This method has the advantage of not requiring a combiner method in order to obtain the class prediction. On the contrary, it needs a method to combine the features returned by each F filter, as can be seen in Fig. 1b. In previous works, strategies such as the union [19]

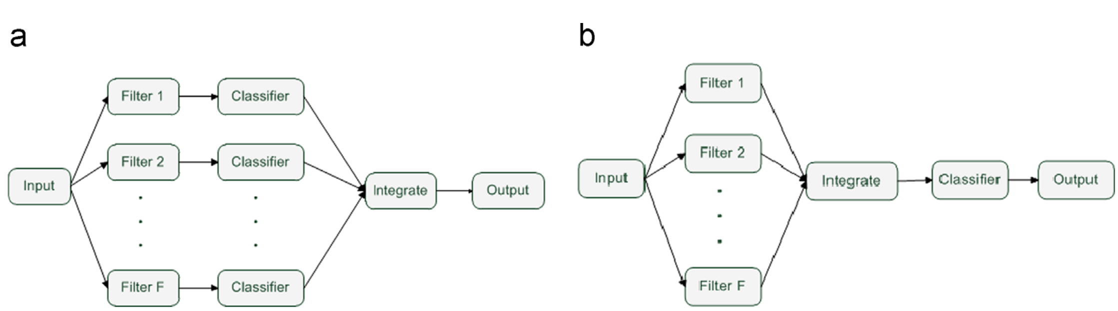

Fig. 1. Implementations of the ensemble. (a) Ensemble1 and (b) Ensemble2.