Information Fusion 24 (2015) 122–136

Contents lists available at ScienceDirect

Information Fusion

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

An ensemble approach of dual base learners for multi-class classification problems

M. Paz Sesmero *, Juan M. Alonso-Weber, German Gutierrez, Agapito Ledezma, Araceli Sanchis

Computer Science Department, Universidad Carlos III de Madrid, Avenida de la Universidad 30, Leganés 28911, Madrid, Spain

ARTICLE INFO

Article history:
Received 9 December 2013
Received in revised form 1 July 2014
Accepted 9 September 2014
Available online 22 September 2014

Keywords:
Ensemble of classifiers
Multi-class classification
Artificial Neural Networks
Feature Selection
Diversity

ABSTRACT

In this work, we formalise and evaluate an ensemble of classifiers that is designed for the resolution of multi-class problems. To achieve a good accuracy rate, the base learners are built with pairwise coupled binary and multi-class classifiers. Moreover, to reduce the computational cost of the ensemble and to improve its performance, these classifiers are trained using a specific attribute subset. This proposal offers the opportunity to capture the advantages provided by binary decomposition methods, by attribute partitioning methods, and by cooperative characteristics associated with a combination of redundant base learners. To analyse the quality of this architecture, its performance has been tested on different domains, and the results have been compared to other well-known classification methods. This experimental evaluation indicates that our model is, in most cases, as accurate as these methods, but it is much more efficient.

1. Introduction

A considerable amount of research in machine learning has been devoted to developing methods that automate the classification tasks. Despite the variety and number of models that have been proposed, the construction of a perfect classifier for any given task is far from achieved [1]. An alternative to improving the accuracy of individual models has appeared during the last decades in the form of classifier ensembles, which are considered one of the most promising areas of research in supervised learning [2].

A specific kind of problem that has been devoted fewer attention concerns the application of ensembles to multi-class problems. Moreover, the lack of efficient solutions grows when the input space has a high dimensionality.

Due to most of the classification systems have been designed for resolving dichotomous problems, the approach to multi-class classification usually consists in decomposing the multiclass problem into several binary sub-problems. Nevertheless, when the learning algorithm that is implicit in the classifiers is easily adaptable to multi-class problems, the binary decomposition might not be the best approach.

In this paper, we present the Binary-Complementary Ensemble (BCE), a homogeneous ensemble of classifiers that is designed to resolve multi-class problems in which the number of features that describe the examples is large. Given than in practical applications, training (space/time) complexity or the testing complexity can be factors as important as the accuracy; the main goal of the BCE architecture is improving the ensemble accuracy, especially in those problems with a high input dimensionality, while keeping the computational cost within reasonable bounds whenever possible.

The feasibility of the proposed ensemble has been empirically tested. This research makes a comprehensive analysis of the performance of the proposed ensemble on different domains, and the results are compared to other well-known classification methods.

This paper is organised as follows: Sections 2 and 3 provide a review of the literature on Classifier Ensembles and Feature Selection. Section 4 presents the architecture of BCE. Section 5 describes the data sets, the method and the measures used to evaluate BCE. Section 6 analyses the experimental results. Last, Section 7 presents concluding remarks and future work.

2. Ensemble of classifiers

An ensemble of classifiers is a set of classifiers whose individual decisions are combined to obtain a system that hopefully outperforms every one of its members [2]. To achieve this goal, the members of the ensemble, known as base learners or base classifiers, must be both accurate and diverse. A classifier is accurate if its

* Corresponding author. Tel.: +34 91 624 9111; fax: +34 91 624 9129.
E-mail addresses: msesmero@inf.uc3m.es (M.P. Sesmero), jmaw@ia.uc3m.es (J.M. Alonso-Weber), ggutierr@inf.uc3m.es (G. Gutierrez), ledezma@inf.uc3m.es (A. Ledezma), masm@inf.uc3m.es (A. Sanchis).

http://dx.doi.org/10.1016/j.inffus.2014.09.002
1566-2535/

classification error is lower than that obtained when the classes are assigned in a random way. Two classifiers are diverse if they make errors on different instances [2].

An important trend for these systems is the search for diversity [3]. Some of the research has been focused on *heterogeneous classifier ensembles*, where the base learners are generated with different learning algorithms, such as Artificial Neural Networks, decision trees, or nearest neighbour classifiers [4–6]. Another approach to achieve diversity is to *inject randomness into the learning algorithm*. For example, [7] show that training a series of Artificial Neural Networks (ANN) on the same training set but with different initial weights can provide a set of classifiers whose behaviour can be quite different. Another method based on this approach is *Randomization* [8]. This method generates decision tress [9] in which the criterion used to expand a node is randomly selected among the 20 best candidates.

Alternatively, diversity can be achieved by using different training data sets to build individual classifiers. Such data sets can be obtained in several ways [10]:

- *Resampling the training examples:* This approach includes two of the most widely known methods to construct ensembles of classifiers: *Bagging* [11] and *Boosting* [12]. *Bagging* builds multiple versions of the training set by applying random sampling with replacement. Each new data set has the same cardinality as the original training set, but some instances are repeated while others are omitted. *Boosting* also resamples the original data set with replacement. This last system is based on a sequential training scheme in which the data set used for building each member of the ensemble depends on the performance of the previously trained classifiers. Therefore, in *Boosting*, misclassified examples are chosen more frequently than correctly predicted examples.
- *Manipulating the input features:* Another way to achieve diversity between classifiers is the quantitative or qualitative modification of the set of features that is used to describe the instances. The quantitative modifications reduce this number by searching appropriate feature subsets. This reduction can be accomplished by random selection [13] or by applying different Feature Selection methods, such as genetic algorithms [14,15], heuristic search techniques [16], or wrapper models [17]. The qualitative modifications involve a change in the feature space. This group includes the methods of non-linear transformations proposed in [18].
- *Manipulating the targets:* A last way for generating diverse classifiers is the manipulation of the classes or categories of the training examples. These techniques are especially useful in multi-class problems. In this situation, the principal alternative is transforming the original problem into several binary sub-problems. These transformations can be accomplished in different ways: OAA – *One against All* – [19] (each classifier separates one class from the $(k-1)$ remaining classes), OAO – *One against One* – [20] (all classes are confronted pairwise), and PAQ – *P against Q* – [21] (each classifier separates a subset *P* of classes from a subset *Q*, where *P* and *Q* are disjoint).

A representative method of the PAQ approach is ECOC – *Error Correcting Output Codes* – [22]. In ECOC, for each classifier *i* of the ensemble, the class set $C = \{c_1, c_2, \dots c_k\}$, is randomly divided into two subsets, C_i^+ and C_i^-. Examples whose class is contained in C_i^+ are labelled as "1", and examples whose class is contained in C_i^- are labelled as "0". The training process delivers a set of binary classifiers that allow classifying new patterns by combining their outputs.

Another method based on PAQ decomposition is OAHO – *One Against Higher Order* – [23]. OAHO is based on a cascaded classifier architecture that ranks the *k* classes based on their number of training examples. The first classifier confronts the majority class (positive samples) against the remaining classes (negative samples). The successive classifiers repeat the same process, suppressing the previous majority class, i.e., taking the previous negative samples and confronting the next majority class against the remaining classes.

Most of the classification systems have been designed for dichotomous problems, and their extension to multi-class classification often leads to an increase in the computational cost or to a reduced system accuracy [24–26]. One way to address this difficulty is to divide the original problem into several binary sub-problems [23,27–32].

A drawback associated with some of the binary decomposition methods is that the mapping induced by the class recoding can provoke or increase the imbalance of the new classes [21]. Moreover, the dichotomous classifiers that integrate these models are trained only on partial knowledge and, in some of these architectures (OAA, OAHO), wrong decisions emitted by a binary classifier are not rectifiable [33]. In this scenario, the system accuracy depends mainly on the accuracy of its members but not on their diversity. Therefore, for certain problems, binary decomposition might not be the best approach.

3. Feature selection

A drawback when dealing with real-world problems is the dimensionality of the data and the computational cost of the classification models. In these situations it can be useful to perform a Dimensionality Reduction based on Feature Selection techniques.

Feature Selection (FS) [34,35] has been applied in literature pursuing the following aims: decreasing the computational cost; increasing the data understanding and data visualization; and reducing the curse of dimensionality. However, the main purpose of Feature Selection is to increase the model accuracy, applying the idea that using as much as possible input information does not imply a better performance. Therefore, the Feature Selection is the procedure of selecting just the relevant information avoiding irrelevant and redundant information, and therefore reducing the computational complexity of the learning task.

It is worth applying FS when: input variables are irrelevant, there is no correlation to the output to be predicted (classification, clustering, or regression); and when some input variables are related to others. Besides, FS can be applied for any prediction task (classification [36], regression [37], clustering [36]), or supervised and unsupervised learning [38].

In order to carry out a Feature Selection procedure to any prediction system, a selection criterion has to be carefully chosen to fix a suitable feature subset. Hence, the criterion can be based on information acquired just from the input and targets data itself, or based on the model accuracy. Based on these criteria, the literature [34,35] establishes a taxonomy for FS methods: Filter [39], Wrapper [40] and Embedded methods [34].

Due to its computational efficiency, in this work Feature Selection is carried out applying a Filter method as *Correlation-Feature Subset Selection* [41]. This is not applied on the whole feature set but on a selected feature subset obtained from the heuristic search known as Best First [42] ("greedy hill climbing augmented with a backtracking facility" [41]). This Feature Selection process is performed executing WEKA software [43].

4. Binary-complementary ensemble architecture

As was previously mentioned, the usual alternative for solving multi-class classification problems is the decomposition of the