Information Fusion 36 (2017) 26–51

Contents lists available at ScienceDirect

Information Fusion

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

NMC: nearest matrix classification – A new combination model for pruning One-vs-One ensembles by transforming the aggregation problem

Mikel Galar[a,e,*], Alberto Fernández[b], Edurne Barrenechea[a,e], Humberto Bustince[a,e], Francisco Herrera[c,d]

[a] Departamento de Automática y Computación, Universidad Pública de Navarra, 31006, Pamplona, Spain
[b] Department of Computer Science, University of Jaén, 23071, Jaén, Spain
[c] Department of Computer Science and Artificial Intelligence, University of Granada, 18071, Granada, Spain
[d] Faculty of Computing and Information Technology - North Jeddah, King Abdulaziz University, 21589, Jeddah, Saudi Arabia
[e] Institute of Smart Cities (ISC), Universidad Pública de Navarra, 31006, Pamplona, Spain

ARTICLE INFO

Article history:
Received 22 July 2015
Revised 15 February 2016
Accepted 3 November 2016
Available online 4 November 2016

Keywords:
Multi-class classification
One-vs-One
Decomposition strategies
Classifier selection
Ensemble pruning

ABSTRACT

The One-vs-One strategy is among the most used techniques to deal with multi-class problems in Machine Learning. This way, any binary classifier can be used to address the original problem, since one classifier is learned for each possible pair of classes. As in every ensemble method, classifier combination becomes a vital step in the classification process. Even though many combination models have been developed in the literature, none of them have dealt with the possibility of reducing the number of generated classifiers after the training phase, i.e., ensemble pruning, since every classifier is supposed to be necessary.

On this account, our objective in this paper is two-fold: (1) We propose a transformation of the aggregation step, which lead us to a new combination strategy where instances are classified on the basis of the similarities among score-matrices. (2) This fact allows us to introduce the possibility of reducing the number of binary classifiers without affecting the final accuracy. We will show that around 50% of classifiers can be removed (depending on the base learner and the specific problem) and that the confidence degrees obtained by these base classifiers have a strong influence on the improvement in the final accuracy.

A thorough experimental study is carried out in order to show the behavior of the proposed approach in comparison with the state-of-the-art combination models in the One-vs-One strategy. Different classifiers from various Machine Learning paradigms are considered as base classifiers and the results obtained are contrasted with the proper statistical analysis.

1. Introduction

Multi-class problems are present in many real-world applications, for example, the severity grading of diseases [1], fingerprint classification [2], the classification of micro-arrays [3] or people tracking [4] to name a few. Although the number of problems that can be viewed as multi-class ones is increasing, binary classifiers are much more studied in the literature. This is due to the fact that there are some classifier learning paradigms in which multi-class modeling is not straightforward. A well-known example of this situation is Support Vector Machine (SVM) [5].

One simple, yet effective way to address multi-class problems in these cases is by means of decomposition strategies [6]. In order to do so, multi-class problems are divided into easier-to-solve binary classification problems following the divide-and-conquer paradigm. As a result, a set of classifiers is learned, each one being responsible for a binary problem. In the testing phase, the outputs of all the classifiers for a given instance are aggregated to make the final decision [7]. Therefore, the difficulty in addressing the multi-class problem is shifted from the classifier itself to the combination stage.

Among decomposition strategies, the One-vs-One (OVO) [8] scheme stands out as one of the most popular techniques.

* Corresponding author. Fax: +34 948 168924.
E-mail addresses: mikel.galar@unavarra.es (M. Galar), alberto.fernandez@ujaen.es (A. Fernández), edurne.barrenechea@unavarra.es (E. Barrenechea), bustince@unavarra.es (H. Bustince), herrera@decsai.ugr.es (F. Herrera).

http://dx.doi.org/10.1016/j.inffus.2016.11.004
1566-2535/

Its usage to model multi-class problems with SVMs in very well-known software tools such as WEKA [9], LIBSVM [10] or KEEL [11], has made it prevalent in many applications. However, it should be mentioned that this strategy can be included in the broader framework of Error Correcting Output Codes (ECOC) [12,13] In OVO, the multi-class problem division is carried out in such a way that a new binary problem is generated for each possible pair of classes. This is why it is also known as *pairwise learning* [14]. Nevertheless, OVO is not only useful to deal with multi-class problems using classifiers without inherent multi-class support, but it also provides a better classification accuracy than addressing the problem directly using multi-class classifiers [15–19].

In the combination phase, the way in which the problem is divided has to be taken into account as a key factor. Several combination methods for the OVO strategy can be found in the literature [18], among which a voting strategy is the most intuitive one (each classifier votes for its predicted class and the most voted one is given as output). Nonetheless, more elaborated approaches have also been developed attending at the inherent difficulties in the OVO decomposition [20–22], although the same accuracy is achieved by simpler alternatives such as the Weighted Voting (WV) [14] or probability estimation methods [23]. An exhaustive empirical study on the combination methods for OVO can be found in [18], where the presence of non-competent classifiers in this strategy was stressed as a promising research line to improve previous combination models. Non-competent classifiers are those that have not been trained with instances from the class to which the example to be classified belongs to. Recent developments have shown that an effective handling of these classifiers allows one to improve the final classification accuracy rate [24,25].

In this paper our aim is to look at the aggregation phase from a different perspective, which may also take advantage of non-competent classifiers rather than avoiding them. Specifically, in our contribution we transform this aggregation by thinking of the outputs of the classifiers as new inputs to another classification problem, which is used to determine the final class labels of the dataset. This view is similar to Stacking [26], although neither a cross-validation procedure is followed (the same base classifier is used for all subproblems) nor a classifier is trained. Stacking and OVO together have been previously considered but with different purposes to ours, focusing on Stacking with cross-validation using different base classifiers [27] and making use of OVO as a Stacking method [28]. In our case, the main difference appears at the combination method. Once the outputs for every training instance are obtained (each one stored in a score-matrix), new instances are simply classified by the most similar score-matrices to that obtained for the new instance, that is, the k Nearest Neighbors (kNN) [29] classifier is applied over the score-matrices (neither requiring a cross-validation nor the usage of different types of base classifiers). This is why we named it as *Nearest Matrix Classification* (NMC).

We will show that by itself this strategy can be competitive and even superior to the state-of-the-art aggregations, although its behavior strongly depends on the underlying classifier and the quality of its confidence degrees. This fact together with the added storage necessity lead us to introduce prototype (in this case, score-matrix) selection methods [30]. This way, only those score-matrices coming from examples that are useful for the classification are maintained in the reference set for NMC classifier, reducing the storage necessity and improving the classification performance as a result of being more robust with respect to the different base classifiers.

More interestingly, this novel view allows us to introduce pruning techniques [31] into OVO, which have not been previously considered, since all classifiers are supposed to be necessary. Pruning techniques for ensembles aim at reducing the pool of classifiers, decreasing the storage necessity, improving performance and reducing testing times. Our new perspective on the combination phase turns the pruning (i.e., classifier selection) into a feature selection problem [32] for the kNN classifier. We will show that almost half of the classifiers in OVO can be safely removed for testing time (depending on the problem and the base classifier) and that if the appropriate confidence estimates are given by the underlying classifier, accuracy can also be boosted in some cases. In order to carry out the feature and instance selection, we consider the usage of a Genetic Algorithm (GA), which has been previously applied with success [33–35].

All these aspects are analyzed in a thorough experimental study, where twenty three real-world problems from the KEEL data-set repository[1] [11,36] are tested using several well-known classifiers from different Machine Learning paradigms as base learners, namely, SVMs [5], decision trees [37,38], instance-based learning [29], and decision lists [39]. Different evaluation criteria are considered to measure the performance, storage reduction and training times. The conclusions obtained are supported by the appropriate statistical tests as suggested in the literature [40,41]. In addition to NMC classifier, state-of-the-art combinations for OVO [18], including a novel Dynamic Classifier Selection (DCS) approach [24] are included in the empirical comparison.

The contributions of this paper are:

- A new combination strategy for OVO is proposed by transforming the aggregation problem.
- The possibility of carrying out pruning in OVO ensembles is introduced for the first time.
- An exhaustive experimental study showing the existence of redundant (non-necessary) classifiers in OVO is developed, which opens up new future research lines in the topic.

The rest of this paper is organized as follows. Section 2 recalls several concepts used in this work. Afterwards, Section 3 discusses other works related to our proposal. Next, Section 4 presents our NMC proposal to prune OVO ensembles. The set-up of the experimental framework is presented in Section 5, whereas the experimental analysis is carried out in Section 6. Finally, Section 7 concludes the paper and presents the future research lines.

2. Preliminaries

This section recalls the OVO scheme, including existing combinations. Afterwards, DTs and their application in OVO are explained.

2.1. The One-vs-One scheme

In the OVO strategy, a m-class problem is divided into $m(m-1)/2$ two-class problems (one for each possible pair of classes). Each binary classification sub-problem is addressed by a different classifier, which is built using training instances only from the two classes considered. This fact is what causes the *non-competence* problem [14,18,24,25] in testing phase.

An easy way of organizing the outputs of the base classifiers for an instance is by means of a score-matrix R, from which different combination models can be applied:

$$R = \begin{pmatrix} - & r_{12} & \cdots & r_{1m} \\ r_{21} & - & \cdots & r_{2m} \\ \vdots & & & \vdots \\ r_{m1} & r_{m2} & \cdots & - \end{pmatrix} \quad (1)$$

[1] http://www.keel.es/dataset.php.