Pattern Recognition 69 (2017) 1–13

Contents lists available at ScienceDirect

Pattern Recognition

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

A parameter randomization approach for constructing classifier ensembles

Enrica Santucci*, Luca Didaci, Giorgio Fumera, Fabio Roli

Dept. of Electrical and Electronic Eng., University of Cagliari Piazza d'Armi, 09123 Cagliari, Italy

ARTICLE INFO

Article history:
Received 23 June 2016
Revised 3 February 2017
Accepted 26 March 2017
Available online 29 March 2017

Keywords:
Multiple classifier systems
Ensemble construction techniques
Randomization
Bagging

ABSTRACT

Randomization-based techniques for classifier ensemble construction, like Bagging and Random Forests, are well known and widely used. They consist of independently training the ensemble members on random perturbations of the training data or random changes of the learning algorithm. We argue that randomization techniques can be defined also by directly manipulating the parameters of the base classifier, i.e., by sampling their values from a given probability distribution. A classifier ensemble can thus be built without manipulating the training data or the learning algorithm, and then running the learning algorithm to obtain the individual classifiers. The key issue is to define a suitable parameter distribution for a given base classifier. This also allows one to re-implement existing randomization techniques by sampling the classifier parameters from the distribution implicitly defined by such techniques, if it is known or can be approximated, instead of explicitly manipulating the training data and running the learning algorithm. In this work we provide a first investigation of our approach, starting from an existing randomization technique (Bagging): we analytically approximate the parameter distribution for three well-known classifiers (nearest-mean, linear and quadratic discriminant), and empirically show that it generates ensembles very similar to Bagging. We also give a first example of the definition of a novel randomization technique based on our approach.

1. Introduction

Ensembles methods have become a state-of-the-art approach for classifier design [1,2]. Among them, ensemble construction techniques based on randomization are well-known and widely used, e.g., Bagging [6], Random Subspace Method [3], Random Forests [4], and the more recent Rotation Forests [7]. Randomization techniques have been formalized in [4] as independently learning several individual classifiers using a given learning algorithm, after randomly manipulating the training data or the learning algorithm itself. For instance, Bagging and Random Subspace Method consist in learning each individual classifier respectively on a bootstrap replicate of the original training set, and on a random subset of the original features; Random Forests (ensembles of decision trees) combine the bootstrap sampling of the original training set with a random selection of the attribute of each node, among the most discriminative ones.

The main effect of randomization techniques, and in particular Bagging, is generally believed to be the reduction of the variance of the loss function of a base classifier. Accordingly, they are effective especially for *unstable* classifiers, i.e., classifiers that exhibit large changes in their output as a consequence of small changes in the training set, like decision trees and neural networks, as opposed, e.g., to the nearest neighbor classifier [6]. It is worth noting that randomization techniques operate in *parallel*, contrary to another state-of-the-art approach, boosting, which is a *sequential* ensemble construction technique [8].

In this work we propose a new approach for defining randomization techniques, inspired by the fact that existing ones can be seen as implicitly inducing a probability distribution on the parameters of a base classifier. Accordingly, we propose that new randomization techniques can be obtained by directly defining a *suitable* parameter distribution for a given classifier, as a function of the training set at hand; an ensemble can therefore be built by directly sampling the parameter values of its members from such a distribution, without actually manipulating the available training data nor running the learning algorithm. In this way, an ensemble can be obtained even without having access to the training set, but having access only to a pre-trained classifier. Some information about the training set, such as mean and covariance matrix,

* Corresponding author.
E-mail addresses: enrica.santucci@gmail.com (E. Santucci), didaci@diee.unica.it (L. Didaci), fumera@diee.unica.it (G. Fumera), roli@diee.unica.it (F. Roli).
URL: http://pralab.diee.unica.it (F. Roli)

http://dx.doi.org/10.1016/j.patcog.2017.03.031
0031-3203/

Table 1
Summary of the notation used in this paper.

Symbol	Meaning
$\mathcal{X}$, $\mathcal{Y}$	Feature space and class label set
$(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$	Feature vector and label of the i-th instance
T	Training set
$\mathcal{L}$	Learning algorithm
$h : \mathcal{X} \mapsto \mathcal{Y}$	Individual classifier
R	Randomization technique
Θ_R	Random variable associated to R
$\Psi(\Theta_R)$	Random variable of the classifier parameters associated to R
μ, Σ	True mean and covariance matrix
$\mathbf{m}, \mathbf{S}$	Sample mean and covariance matrix

is enough to apply our method, and it could be obtained from a pre-trained classifier.

Our approach also allows a different implementation of existing randomization techniques. If the distribution induced by a given technique on the parameters of a given base classifier is known or can be approximated, one could build an ensemble as described above, instead of running the corresponding procedure and then the learning algorithm.

As mentioned above, the key issue of our approach is to define a suitable parameter distribution for a given base classifier, i.e., capable of providing a trade-off between accuracy and diversity of the resulting classifiers which is advantageous in terms of ensemble performance. To our knowledge no previous work investigated the distribution of classifier parameters induced by randomization techniques, which is not a straightforward problem. To take a first step in this direction, in this work we start from the analysis and modeling of the distribution induced by one of the most popular techniques, Bagging, on base classifiers that can be dealt with analytically: the nearest mean, linear discriminant, and quadratic discriminant classifiers. We then assess the accuracy of our model by comparing the corresponding, empirical parameter distribution with the one produced by Bagging. The results of our analysis, that have to be extended in future work to other base classifiers and randomization techniques, are aimed at obtaining insights on the parameter distributions induced by existing randomization techniques, and thus hints and guidelines for the definition of *novel* techniques based on our approach. We give a first example of the definition of a new randomization technique, starting from our model of the distribution induced by Bagging on the classifiers mentioned above.

The rest of this paper is structured as follows. In Section 2 we summarize the main relevant concepts about randomization techniques and Bagging. We then present our approach and describe the considered base classifiers in Section 3. In Section 4 we model the parameter distribution induced by Bagging on such classifiers. In Section 5 we empirically evaluate the accuracy of our model, and give an example of the definition of new randomization techniques based on our approach. In Section 6 we discuss limitations and extensions of our work.

2. Background

The notation used in this paper is summarized in Table 1. We shall use Greek letters to denote probability distribution parameters, and Roman letters for other quantities, including estimated distribution parameters (statistics); vectors in Roman letters will be written in bold. For a given statistic $\mathbf{a}$ estimated from a training set we shall denote by $\mathbf{a}^*(j)$ its j-th bootstrap replicate, and with $\mathbf{a}^*$ the corresponding random variable.

Randomization techniques for ensemble construction can be formalized as follows [4]. Given a feature space $\mathcal{X} \subseteq \mathbb{R}^d$, a set of class labels $\mathcal{Y}$, a training set $T = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, a base classifier and its learning algorithm $\mathcal{L}$, a randomization technique R independently learns N different classifiers $h_j(\cdot\,; \theta_j)$, $j = 1, \ldots, N$, by repeatedly calling $\mathcal{L}$, where $\theta_1, \ldots, \theta_N$ are independent and identically distributed (i.i.d.) realizations of some random variable Θ_R. In practice, the above idea can be implemented by introducing some randomness into the training process of the individual classifiers, by manipulating either the training data or the learning algorithm, or both.

As an example, we focus here on the popular Bagging technique. It has been originally devised for regression tasks, with the aim of reducing the variance of the expected error (mean squared error) of a given regression algorithm, and has been extended to classification algorithms [6]. According to the above formalization, the corresponding random variable Θ_R is associated with the bootstrap sampling procedure: its values correspond to the possible bootstrap replicates T^* of the original training set T of size n, obtained by randomly drawing with replacement n instances from it (hence the name "Bagging", which is an acronym for "bootstrap aggregating"). Each base classifier h_j, $j = 1, \ldots, N$, is learned on a bootstrap replicate T_j^*, and can be also denoted as $h_j(\cdot; T_j^*)$. The ensemble prediction is usually obtained by majority voting. For base classifiers that output a real-valued score, simple averaging can also be used [6].

As the ensemble size N increases, its output approaches the asymptotic Bagging prediction, which, when majority voting is used, is defined as:

$$y^* = \arg\max_{y \in \mathcal{Y}} \mathbb{P}[h(\mathbf{x}; T^*) = y] \ . \tag{1}$$

Several authors (e.g., [6,9,10]) have shown that ensembles of 10 to 25 "bagged" classifiers attain a performance very similar to the one of larger ensembles, and thus of the asymptotic Bagging. This is a useful, practical guideline to attain a trade-off between computational (both space and time) complexity and classification performance.

Since [6], Bagging is known to be effective especially for unstable classifiers like decision trees and neural networks. In particular, it mainly works by reducing the variance component of the loss function (usually, the misclassification probability) of a given base classifier [11,12]. Other explanations have also been proposed; for instance, in [13] it has been argued that Bagging equalizes the influence of training instances, and thus reduces the effect of outliers; this is due to the fact that every instance in T has a probability of about 0.632 of appearing in a bootstrap replicate, and thus each outlier is present on average only in 63% of them.

A thorough analysis of the stabilizing effect of Bagging has been carried out in [9,14] for the Linear Discriminant and the Nearest Mean classifiers. Their degree of instability was found to depend also on the training set size n: the smaller the training set, the higher the instability, which in turn worsens classification performance. In particular, the above classifiers turned out to very unstable (thus exhibiting a maximum of the generalization error) for critical values of n around the number of features d, and Bagging was capable of improving their performance only under this condition.

In Section 4 we shall analyze and model the parameter distribution induced by Bagging on some base classifiers, including the ones considered in [9,14], as a first step toward the development of novel randomization techniques based on the definition of suitable parameter distributions.

3. A parameter randomization approach for ensemble construction

Consider a given classification algorithm, e.g., a parametric linear classifier with discriminant function $\mathbf{w}^\top \cdot \mathbf{x} + w_0$ implemented