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
Email addresses: enrica.santucci@gmail.com, didaci@diee.unica.it, fumera@diee.unica.it, roli@diee.unica.it
URL: http://pralab.diee.unica.it

Abstract

Randomization 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 a given learning algorithm. In this work we argue that randomization techniques can be defined also by directly manipulating the parameters of a given classifier, i.e., by sampling their values from a given probability distribution. A classifier ensemble can thus be built without first defining a procedure for manipulating the training data or the learning algorithm and the training the individual classifiers. The key issue is to define a suitable parameter distribution for a given base classifier. This provides also a different perspective on existing techniques: the resulting classifier parameters can be seen as if they were drawn from the distribution implicitly defined by a given randomization technique that explicitly manipulates the training data or the learning algorithm. Accordingly, if this distribution is known or can be approximated, one could re-implement existing techniques by randomly sampling the classifier parameters from the corresponding distribution. 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 some well-known classifiers (nearest-mean, linear and quadratic discriminant), and empirically show that, using the approximated distribution, our approach provides ensembles very similar to Bagging. We also give a first example of the definition of a novel randomization technique based on our approach.

Keywords: Multiple classifier systems, Ensemble construction techniques, Randomization, Bagging

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. The main randomization techniques are Bagging [5], the Random Subspace Method [3], Random Forests [4], and the more recent Rotation Forests [6]. 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, in Bagging each base classifier is trained on a bootstrap replicate of the original training set; in the Random Subspace Method, each classifier is trained using 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 component of the expected misclassification probability of a base classifier. Accordingly, such techniques are particularly effective for *unstable* classifiers, i.e., classifiers that exhibit large changes in their output as a consequence of small changes in the training set. Decision trees and neural networks are well-known examples of unstable classifiers, as opposed, e.g., to the nearest neighbor classifier [5]. It is worth noting that randomization techniques are *parallel* ensemble construction techniques, as opposed to another state-of-the-art approach, boosting, which is a *sequential* one [7].

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 the base classifier at hand. Accordingly, we propose that new randomization techniques can be obtained by directly defining a *suitable* probability distribution on the parameters of a given classifier; a classifier ensemble can therefore be built simply by randomly sampling the parameter values of its members, without actually manipulating the training data nor running any learning algorithm.

Our approach also provides a different perspective on existing randomization techniques. If their underlying parameter distribution is known, or can be approximated, one could build a classifier ensemble as described above, without running the corresponding procedure for manipulating the training data or modifying (and anyway running) the chosen learning algorithm.

As mentioned above, the key issue of our approach is to define a suitable probability distribution on the parameters of a given base classifier, i.e., capable of providing an advantageous trade-off, in terms of the ensemble performance, between accuracy and diversity of the resulting classifiers. Since this is not a straightforward task, and to our knowledge no previous work investigated the distribution of classifier parameters induced by randomization techniques, in this paper we make a first step by starting from the analysis of the distribution induced by one of the most popular techniques, Bagging. To this aim, we consider base classifiers and learning algorithms whose distribution induced by Bagging can be analytically approximated: two linear classifiers (nearest mean and linear discriminant analysis), and the quadratic classifier. To evaluate the accuracy of our approximation, we compare the empirical parameter distribution produced by Bagging with the one obtained by our model, on artificial and real-world data sets. Based on these results, we also give a first example of a new randomization technique that can be defined based on our approach, by modifying the distribution induced by Bagging on the same classifiers above.

The rest of this paper is structured as follows. In Sect. 2 we provide the necessary details about Bagging. Our approach is then presented in Sect. 3, where the three base classifiers mentioned above are also described. In Sect. 4 we develop an analytical model of the parameter distribution induced by Bagging on such classifiers. Experimental results are reported in Sect. 5. We conclude this paper by discussing limitations and possible extension of our work, in Sect. 6.

2. Background

The notation used in this paper is summarized in Table **??**. Throughout the paper we shall use...

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 $T = \{(x_i, y_i)\}_{i=1}^n$, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, and a learning algorithm $\mathcal{L}$, a randomization technique independently learns N different base classifiers $h_j(\cdot; \theta_j) : \mathcal{X} \rightarrow \mathcal{Y}$, $j = 1, \ldots, N$, by repeatedly calling $\mathcal{L}$, where $\theta_1, \ldots, \theta_N$ are independent and identically distributed (i.i.d.) instances of some random variable Θ.

In practice, the above idea can be implemented by introducing some randomness to the training process of the individual classifiers, i.e., either in the training data or in 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 component of the expected error (mean squared error) of a given regression algorithm [5]. In the same work it has also been extended to classification algorithms. According the above formalization, in the case of Bagging the random variable Θ is associated to the bootstrap sampling procedure; accordingly, 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, which is an acronym for "bootstrap aggregating"). Therefore, 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.

As the ensemble size N increases, its output approaches the asymptotic Bagging prediction. If majority voting is used, for an input instance $\mathbf{x}$ the asymptotic Bagging prediction is the class label y^* defined as:

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

Several authors (e.g., [5, 8, 9]) 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 [5], Bagging is known to be effective especially for unstable classifiers. Well-known examples of unstable classifiers are decision trees and neural networks. In particular, several authors have observed that Bagging tends to reduce the variance component of the expected loss (usually, the misclassification probability) of a given base classifier [10, 11]. Other explanations of Bagging have also been proposed; for instance, in [12] it has been argued that Bagging equalizes the influence of training instances, and thus reduces the influence of outliers; this is due to the fact that every instance in T has a probability approximately equal to 0.632 of appearing in a bootstrap replicate, and thus each outlier is present on average only in 63% of bootstrap replicates.

A thorough analysis of the stabilizing effect of Bagging has been carried out in [8, 13] for two widely used linear classifiers: the Linear Discriminant and the Nearest Mean Classifier. In these works it was pointed out that the degree of instability of a given classifier depends also on the training set size: the smaller the training set, the higher the instability, which in turn leads to a worse classification performance. In particular, the above linear classifiers turned out to exhibit a very unstable behavior (and thus a maximum of the generalization error) for critical values of the training set size n around the number of features d, and Bagging was capable of improving their performance only under this condition.

With regard to our approach to the development of new randomization techniques, we point out that useful insights on the definition of suitable parameter distributions could be provided by the study of the distributions induced by existing techniques, like Bagging.

3. A parameter randomization approach for ensemble construction

Consider a given classification algorithm, e.g., a linear classifier with discriminant function $\mathbf{w}^\top \cdot \mathbf{x} + w_0$ implemented as the linear discriminant classifier (LDC), or a neural network trained with the back-propagation algorithm. Let ψ denote the parameters that are set by its learning algorithm $\mathcal{L}$, e.g., the coefficients of a linear classifier (in this case, $\psi = (\mathbf{w}, w_0)$), or the connection weights of a neural network.

Consider now any given randomization technique R (e.g., Bagging), defined by some manipulation procedure of the training set or of $\mathcal{L}$. The classifiers of an ensemble of size N obtained using R can be denoted as $h_1(\mathbf{x}; \psi(\theta_1))$, $\ldots$, $h_N(\mathbf{x}; \psi(\theta_N))$, where θ_j, $j = 1, \ldots, N$, denote N i.i.d. realizations of the random variable Θ associated to R, and the $\psi(\theta_j)$'s denote the parameters of the corresponding classifiers, where we explicitly point out their dependence on the θ_j's. For instance, if Bagging is applied to a linear classifier, $\psi(\theta_j)$ denotes the coefficients $(\mathbf{w}_j, w_{0,j})$ obtained by the chosen learning algorithm (e.g., LDC) on a bootstrap replicate T_j^* of the original training set. In the above setting, the parameters $\psi(\theta_j)$ can be seen as i.i.d. realizations of a random variable $\Psi = \Psi(\Theta)$, whose distribution is implicitly defined by the considered randomization technique R and learning algorithm $\mathcal{L}$, and is parametrized by the training set at hand. Accordingly, we write such a distribution as $\mathbb{P}_{\mathrm{R},\mathcal{L}}[\Psi]$. Note that also the realizations $\psi(\theta_i)$ are i.i.d. because they are functions of i.i.d. realizations of Θ.

A first consequence of the above formalization is that it enables an alternative procedure for constructing a classifier ensemble using a *known* technique R, in the case when the distribution $\mathbb{P}_{\mathrm{R},\mathcal{L}}[\Psi]$ is known or can be approximated. The traditional procedure [4] is to run $\mathcal{L}$ for N times on (possibly perturbed versions of) the training set (e.g., in the case of Bagging, on bootstrap replicates of the original training set), or on a modified version of $\mathcal{L}$, according to R. The alternative procedure is to independently draw N i.i.d. realizations $\psi_1, \ldots, \psi_N$ of the classifier parameters by sampling from $\mathbb{P}_{\mathrm{R},\mathcal{L}}[\Psi]$. In this case, no manipulation of the training set is actually carried out, as well as no run of the learning algorithm. For instance, if one wants to build an ensemble of LDCs using Bagging, the distribution $\mathbb{P}_{\mathrm{R},\mathcal{L}}[\Psi]$ should be modelled based on the analysis of the coefficients $\psi = (\mathbf{w}, w_0)$ of a LDC trained on bootstrap replicates of a given training set, and on the training data at hand (a detailed example will be given in Sect. 4). A possible advantage of this approach is a lower processing time for ensemble construction.

More interestingly, the above formalization suggests a different approach for developing *novel* randomization techniques, alternative to the definition of a given procedure for manipulating the training data or modifying $\mathcal{L}$, and then to run $\mathcal{L}$ to obtain each ensemble member. This approach consists in directly defining a suitable distribution $\mathbb{P}_{\mathcal{L}}[\Psi]$, not related to any actual procedure R for manipulating the training data or the learning algorithm. one can then build the ensemble members by sampling from such a distribution the corresponding parameter values, without carrying out any training procedure. As pointed out in section1, this approach translates the requirement of defining an effective procedure (in terms of ensemble performance) for manipulating the training data or the learning algorithm, into defining a suitable distribution $\mathbb{P}_{\mathcal{L}}[\Psi]$. This task is not straightforward, for several reasons.

One reason is that different distributions should be defined for base classifiers characterized by different parameters (e.g., decision trees and neural networks). This does not happen for several techniques based on the traditional approach, instead, like Bagging and the Random Subspace Method, which consist in manipulating only the training data, and can therefore be applied to any base classifier. On the other hand, a technique like the Random Forest is specific to decision trees. Another reason is that understanding the *joint* effect of the parameter values of a given classifier (e.g., the connection weights of a neural network) on its performance, and on the performance of an *ensemble* of such classifiers, can be very difficult.

Accordingly, in this work we chose to investigate our approach starting from the analysis of the parameter distribution induced on a given classifier by a randomization technique defined according to the traditional approach, i.e., by an explicit manipulation procedure of training data or of the learning algorithm. The reason is that the results of such an analysis could provide useful insights on the characteristics of a suitable distribution should exhibit, as mentioned at the end of Sect. 2. To this aim, we focus on the popular Bagging technique, and consider three base classifiers and the corresponding learning algorithms for which an analytical approximation of the corresponding parameter distributions is feasible. The chosen classifiers are summarized in Sect. 4.1.

4. Analytical study of the parameter distribution of "bagged" classifiers

In this section we present our analytical study of the parameter distribution of the training sets obtained by Bagging.

With no loss of generality, we consider a two-class problem, with class labels $\mathcal{Y} \in \{+1, -1\}$. The results can be easily extended to multi-class problems. For the sake of simplicity, we assume that the training set T contains the same number n of instances from both classes: $T = \{(\mathbf{x}_i, +1)\}_{i=1}^{n} \bigcup \{(\mathbf{x}_i, -1)\}_{i=n+1}^{2n}$. The analytical results are very similar when the above assumption does not hold. We also make the usual assumption that training instances are i.i.d. To enable analytical derivations, we consider Gaussian class-conditional distributions. We then show in Sect. 4.6 how our results can be generalized to the case of non-Gaussian, and even unknown distributions.

In the following we shall use Greek letters to denote the parameters of the probability distributions, 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 T (*e.g.*, the sample mean of class +1), we shall denote by $\mathbf{a}^*(j)$, $j = 1, \ldots, N$, its bootstrap replicates, and with $\mathbf{a}^*$ the corresponding random variable.

We denote by $\mathbf{m}_1 = (m_{1,1}, \ldots, m_{1,d})^\top$ and by $\mathbf{S}_1$ the maximum likelihood estimates of the mean μ_1 and covariance matrix Σ_1 of class +1, *i.e.*:

$$\mathbf{m}_1 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \qquad \mathbf{S}_1 = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m}_1)(\mathbf{x}_i - \mathbf{m}_1)^\top \ . \tag{2}$$

We denote the analogous quantities for class -1 by $\mathbf{m}_2$ and $\mathbf{S}_2$.

In our derivations four random variables play the main role: the sample mean $\mathbf{m}_k^*$ and the sample covariance matrix $\mathbf{S}_k^*$ of the bootstrap replicates of T, $k = 1, 2$.

Since the sample mean of n i.i.d. instances drawn from a multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$ follows the distribution $\mathcal{N}(\mu, \frac{1}{n}\Sigma)$, according to the results in [14] we can approximate the distribution of $\mathbf{m}_k^*$, $k = 1, 2$, with two independent multivariate Gaussian distributions:

$$\mathcal{N}\left(\mu_k, \frac{1}{n}\Sigma_k\right), \ k = 1, 2. \tag{3}$$

Note that the above approximation is valid even when the data distribution is non-Gaussian, provided that n large enough, in virtue of the Central Limit Theorem (CLT). At this purpose, there exists a *heuristic rule* according to which for $n \geq 30$ the application of the CLT is well justified. In some cases, a smaller value of n is enough, as we shall show in Sect. 5.1.

Beside $\mathbf{m}_k^*$, the sample covariance matrix $\mathbf{S}_k^*$ is a random variable, too. However, it is not easy to analytically solve the problem by considering both mean and covariance matrix as random variables because, as a consequence of Eq. (2), we should consider two dependent random variables $\mathbf{m}_k^*$ and $\mathbf{S}_k^*$ and the derivation of the parameter distributions would become very difficult (especially when we will compute the inverse of the covariance matrix).

To further simplify our analysis, we shall approximate the sample covariance matrices of *any* bootstrap replicate $T^*(j)$ with the corresponding (constant) covariance matrix of the data distribution:

$$\mathbf{S}_k^*(j) \simeq \Sigma_k, \ j = 1, \ldots, N \ . \tag{4}$$

We shall evaluate by numerical simulations the accuracy of this approximation in Sect. 5. Accordingly, in our analysis only the $\mathbf{m}_k^*$'s are random variables.

Based on the above assumptions and results, in the following subsections we derive the parameter distributions of the classifiers mentioned above, and summarized in Sect. 4.1, under the assumption of Gaussian class-conditional distributions, and under different forms of the covariance matrices Σ_1 and Σ_2:

- Case 1: identical covariance matrices, proportional to the identity matrix: $\Sigma_1 = \Sigma_2 = \sigma^2\mathbf{I}$.
- Case 2: identical covariance matrices, proportional to the identity matrix but with different diagonal values: $\Sigma_1 = \Sigma_2 = \Sigma = \vec{\sigma}^2\mathbf{I}$, where $\vec{\sigma} = (\sigma_1^2, \ldots, \sigma_d^2)$.
- Case 3: identical covariance matrices having a general form: $\Sigma_1 = \Sigma_2 = \Sigma$.
- Case 4: diagonal covariance matrices, different from each other: $\Sigma_1 = \vec{\sigma}_1^2 I$, $\Sigma_2 = \vec{\sigma}_2^2 I$ such that $\vec{\sigma}_1 \neq \vec{\sigma}_2$, $\vec{\sigma}_1^2 = (\sigma_{1,1}^2, \ldots, \sigma_{1,d}^2)$, $\vec{\sigma}_2^2 = (\sigma_{2,1}^2, \ldots, \sigma_{2,d}^2)$.

We finally consider in Sect. 4.6 the most general case of non-Gaussian or unknown data distribution.

4.1. Base classifiers

Here we summarize the three base classifiers considered in this work considering their "ideal" discriminant function, *i.e.*, written in terms of the true parameters of the underlying data distribution. For the sake of simplicity we only consider the case of identical class priors. All such classifiers provide the optimal discriminant function (either asymptotically with respect to training set size, or in the ideal case when the data distribution is known) when the class-conditional distributions are Gaussian, under different forms of the covariance matrices of the two classes. This fact makes it easier to analytically derive the parameter distributions of the corresponding "bagged" classifiers.

Nearest-mean classifier (NMC). This is a linear classifier whose discriminant function is defined as:

$$g(\mathbf{x}) = \mathbf{w}^\top \cdot (\mathbf{x} - \mathbf{x}_0) \ , \tag{5}$$

where

$$\mathbf{w} = \mu_1 - \mu_2, \quad \mathbf{x}_0 = \frac{1}{2}(\mu_1 + \mu_2) \ . \tag{6}$$

This is the optimal classifier, if the class-conditional distributions are Gaussian and $\Sigma_1 = \Sigma_2 = \sigma^2\mathbf{I}$.

Linear Discriminant Classifier (LDC). The LDC is another well-known linear classifier. Its discriminant function is equal to (5), where:

$$\mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_2), \quad \mathbf{x}_0 = \frac{1}{2}(\mu_1 + \mu_2) \ , \tag{7}$$

and $\Sigma = \frac{1}{2}(\Sigma_1 + \Sigma_2)$. The LDC is the optimal classifier, if the class-conditional distributions are Gaussian with identical covariance matrices (of any form): $\Sigma_1 = \Sigma_2 = \Sigma$.

Quadratic Discriminant Classifier (QDC). This classifier produces a quadratic discriminant function:

$$g(\mathbf{x}) = \mathbf{x}^\top \mathbf{W} \mathbf{x} + \mathbf{w}^\top \mathbf{x} + w_0 \ , \tag{8}$$

where

$$\begin{aligned} \mathbf{W} &= \frac{1}{2}(\Sigma_2^{-1} - \Sigma_1^{-1}) \ , \\ \mathbf{w} &= (\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2) \ , \\ w_0 &= \frac{1}{2}(\mu_2^\top \Sigma_2^{-1}\mu_2 - \mu_1^\top \Sigma_1^{-1}\mu_1) \ . \end{aligned} \tag{9}$$

The QDC is the optimal classifier, when the class-conditional distributions are Gaussian, without constraints on the form of covariance matrices.

4.2. Case 1: identical covariance matrices proportional to the identity matrix

In this section we assume

$$\Sigma_1 = \Sigma_2 = \sigma^2\mathbf{I} = \Sigma \ , \tag{10}$$

for some value of $\sigma \in \mathbb{R}$, where Σ denotes the common covariance matrix.

4.2.1. Parameter distribution of the "bagged" classifiers

Due to our approximation (4), under assumption (10) the LDC and QDC classifiers coincide with the NMC (see Sect. 4.1). Their discriminant function is given by Eq. (5), where $\mathbf{w} = (w_1, \ldots, w_d)^\top = \mu_1 - \mu_2$, and $\mathbf{x}_0 = (x_{01}, \ldots, x_{0d})^\top = \frac{1}{2}(\mu_1 + \mu_2)$. Such a function is a hyperplane orthogonal to the line joining μ_1 and μ_2, and it is independent on Σ. A single classifier is therefore described by means of $d+1$ independent parameters, *i.e.* $\mathbf{w}$ (a d-dimensional vector) and $w_0 = \mathbf{w}^\top \cdot \mathbf{x}_0$ (a scalar value). Consequently, the parameter vector is $\Psi = (\mathbf{w}, w_0) \in \mathbb{R}^{d+1}$.

According to approximation (4), also the "bagged" QDC and LDC coincide with the "bagged" NMC, which is defined by $\mathbf{w}^* = \mathbf{m}_1^* - \mathbf{m}_2^*$ and $\mathbf{x}_0^* = \frac{1}{2}(\mathbf{m}_1^* + \mathbf{m}_2^*)$, where both quantities are independent on Σ.

Our goal is to derive the distribution of the corresponding parameter vector $\Psi^* = (\mathbf{w}^*, w_0^*) \in \mathbb{R}^{d+1}$. However, whereas the distribution of $\mathbf{w}^*$ is Gaussian, the one of $w_0^* = (\mathbf{w}^*)^\top \cdot \mathbf{x}_0^* = \frac{1}{2}((\mathbf{m}_1^*)^\top \cdot \mathbf{m}_1^* - (\mathbf{m}_2^*)^\top \cdot \mathbf{m}_2^*)$ is not, and involves non-central Chi-Squared distributions which are more difficult to treat. For this reason we consider the following, redundant parameter vector:

$$\Psi^* = (\mathbf{w}^*, \mathbf{x}_0^*) = (w_1^*, \ldots, w_d^*, x_{01}^*, \ldots, x_{0d}^*) \in \mathbb{R}^{2d} , \tag{11}$$

since also the distribution of $\mathbf{x}_0^*$ is Gaussian. From the above discussion, it follows that the distribution of Ψ^* can be approximated by a Gaussian:

$$\mathcal{N}(\xi, \Sigma_\xi) , \tag{12}$$

where the expected value $\xi \in \mathbb{R}^{2d}$ and the $2d \times 2d$ covariance matrix Σ_ξ are given by

$$\xi = (w_1, \ldots, w_d, x_{01}, \ldots, x_{0d}), \quad \Sigma_\xi = \begin{pmatrix} \Sigma_{\mathbf{w}^*} & \Sigma_{\mathbf{w}^*,\mathbf{x}_0^*} \\ \Sigma_{\mathbf{w}^*,\mathbf{x}_0^*} & \Sigma_{\mathbf{x}_0^*} \end{pmatrix} , \tag{13}$$

and $\Sigma_{\mathbf{w}^*,\mathbf{x}_0^*}$ is a $d \times d$ matrix whose components are the covariances among all the $\mathbf{w}^*$ and $\mathbf{x}_0^*$ components:

$$\Sigma_{\mathbf{w}^*,\mathbf{x}_0^*} = \{\mathrm{cov}(w_i^*, x_{0j}^*)\}_{i,j=1,\ldots,d} . \tag{14}$$

According to assumption (10) we have:

$$\Sigma_{\mathbf{w}^*} = \frac{2\sigma^2}{n}\mathbf{I}_d, \quad \Sigma_{\mathbf{x}_0^*} = \frac{\sigma^2}{2n}\mathbf{I}_d, \quad \Sigma_{\mathbf{w}^*,\mathbf{x}_0^*} = \mathbf{0}_{d\times d} \ . \tag{15}$$

Note that the above results follow from the following properties: *i)* $\mathbf{m}_1^*$ and $\mathbf{m}_2^*$ are independent random variables; *ii)* the components of the random vectors $\mathbf{m}_1^*$ and $\mathbf{m}_2^*$ are independent on each other, since the features are uncorrelated according to assumption (10); *iii)* the Normal distribution belongs to the *Lévy alpha-stable distribution family* [18], *i.e.* linear combination of independent Normal variables is a Normal variable. We also point out that the off-diagonal terms of the sub-matrix $\Sigma_{\mathbf{w}^*,\mathbf{x}_0^*}$ are equal to zero because the random variables x_{0i}^* and w_j^* are independent for $i \neq j$. The situation is different for the diagonal terms $\mathrm{cov}(x_{0i}^*, w_i^*)$ because the random variables $x_{0i}^* = (\mu_{1,i}^* - \mu_{2,i}^*)/2$ and $w_i^* = \mu_{1,i}^* - \mu_{2,i}^*$ are not independent; the only exception is when both classes have the same number of training instances, in which case also the latter terms are null.[1]

Finally, we point out that, although the discriminant function of a single NMC does not depend on Σ, the covariance matrix Σ_ξ of the corresponding parameter distribution, given by Eq. (12), does depend on Σ.

4.2.2. Confidence regions for the distribution parameters

In this section we derive the confidence regions for the parameters of the distribution derived above. We assume to deal with a finite number n of instances. In this case the estimation of the "distance" between the true and the estimated statistic is given by the confidence regions involving the Student's t-distribution [19] for the one-dimensional case, and the Hotelling's T-squared distribution [20] (which is a generalization of the former) used for multivariate tests. We consider this kind of distributions in place of the standard confidence intervals because generally we do not know the covariance matrix of the data

[1] Denoting by n_1 and n_2 the number of training instances of the two classes, the diagonal terms of $\Sigma_{\mathbf{w}^*}$, $\Sigma_{\mathbf{x}_0^*}$, $\Sigma_{\mathbf{w}^*,\mathbf{x}_0^*}$ become respectively $\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$, $\frac{\sigma^2}{4}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$, $\frac{\sigma^2}{2}\left(\frac{1}{n_1}-\frac{1}{n_2}\right)$.

and we use its maximum likelihood estimate (see Eq. (2)). In particular, in the one-dimensional case ($d = 1$) the considered classifiers are defined only by the parameter $x_0 = (\mu_1 + \mu_2)/2$. In this case the confidence interval is given by:

$$\frac{m_1 + m_2}{2} \pm t_{2n-1}^{(1-\alpha)} \cdot \frac{s}{\sqrt{2n}} \ , \tag{16}$$

where $t_{2n-1}^{(1-\alpha)}$ is the $(1-\alpha)$-th percentile of the Student's t-distribution for $2n-1$ degrees of freedom, m_1 and m_2 are the sample means, and s is the estimated standard deviation of the data.

In the more general case of $d > 1$, the set of hypotheses we have to test is:

$$\begin{cases} H_0 : \ \xi = \xi^{(0)} \\ H_1 : \xi \neq \xi^{(0)} \quad \text{for some } i \end{cases} \tag{17}$$

where $\xi^{(0)} = (\xi_1^{(0)}, \xi_2^{(0)}, \ldots, \xi_{2d}^{(0)})$ is a known vector.

According to the T^2 Hotelling test (which generalizes the Student's t-test discussed above), the hypothesis H_0 is accepted with probability $1 - \alpha$, if:

$$Pr\{T_{\Psi^*}^2 < T_{2d}^2(n-1, \alpha)\} = 1 - \alpha \ , \tag{18}$$

where $T_{\Psi^*}^2 = (\mathbf{m}_{\Psi^*} - \xi^{(0)})^T \mathbf{S}_{\Psi^*}^{-1} (\mathbf{m}_{\Psi^*} - \xi^{(0)})$ (with $\mathbf{m}_{\Psi^*}$ and $\mathbf{S}_{\Psi^*}$ estimated values of ξ and Σ_ξ respectively) and $T_{2d}^2(n-1, \alpha)$ is the α-th percentile of the $2d$-dimensional T^2 Hotelling distribution with $n-1$ degrees of freedom. Eq. (18) represents the *confidence ellipsoid* centered in $\xi^{(0)}$. Obviously, the hypothesis H_0 is refused (with the same probability), if $T_{\Psi^*}^2 > T_{2d}^2(n-1, \alpha)$.

4.3. Case 2: identical, diagonal covariance matrices

We now discuss the case in which the covariance matrices of the classes are identical and proportional to the identity matrix, but with different diagonal values, *i.e.*:

$$\Sigma_1 = \Sigma_2 = \Sigma = \vec{\sigma}^2 \mathbf{I} \ , \tag{19}$$

where $\vec{\sigma} = (\sigma_1^2, \ldots, \sigma_d^2)$.

The LDC is the optimal classifier under the above assumption, and coincides with the QDC. We analyze first these two classifiers. Their decision function is the one of Eq. (5), where $\mathbf{x}_0$ and $\mathbf{w}$ are given by Eq. (7), with $\mathbf{w}$ depending on the matrix Σ. Accordingly, the resulting discriminant function is again a hyperplane, but it is not orthogonal to the line joining μ_1 and μ_2.

In order to derive the distribution of the parameters $\Psi^* = (\mathbf{w}^*, \mathbf{x}_0^*)$ of the "bagged" classifier, we recall the following well-known property. If $X \sim \mathcal{N}(\mu, \Sigma)$ is a p-dimensional random variable with a multivariate Normal distribution, and A and b are respectively a non-singular matrix and a vector of proper size, then also $Y = AX + b$ has a multivariate Normal distribution, such that $Y \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$. This implies that Ψ^* has a multivariate Normal distribution $\mathcal{N}(\xi, \Sigma_\xi)$, where:

$$\xi = \left[\frac{\mu_{1,1} - \mu_{2,1}}{\sigma_1^2}, \ldots, \frac{\mu_{1,d} - \mu_{2,d}}{\sigma_d^2}, \frac{\mu_{1,1} + \mu_{2,1}}{2}, \ldots, \frac{\mu_{1,d} + \mu_{2,d}}{2}\right], \tag{20}$$

and Σ_ξ has the same structure as in Eq. (13), where:

$$\Sigma_{\mathbf{w}^*} = \frac{2}{\vec{\sigma}^2 n}\mathbf{I}_d, \quad \Sigma_{\mathbf{x}_0^*} = \frac{\vec{\sigma}^2}{2n}\mathbf{I}_d, \quad \Sigma_{\mathbf{w}^*, \mathbf{x}_0^*} = \mathbf{0}_{d \times d} \ . \tag{21}$$

Note that also in this case $\Sigma_{\mathbf{w}^*, \mathbf{x}_0^*}$ is the null matrix only if the classes exhibit identical prior probabilities, otherwise it is a diagonal matrix.[2]

Consider now the NMC, which is suboptimal under assumption (19). In this case, the parameter distribution of the "bagged" NMC turns out to be the one derived in Sect. 4.2.1, given by Eqs. (13) and (15), where $\sigma^2 = \left(\frac{1}{d}\sum_{i=1}^{d} \sigma_i^2\right)$.

Consider finally the confidence regions for the distribution parameters (20) and (21). We obtain results similar to the ones discussed in Sect. 4.2.2 where $\mathbf{m}_{\Psi^*}$ and $\mathbf{S}_{\Psi^*}$, in the $T^2_{\Psi^*}$ formula, are the estimated values of ξ and Σ_ξ respectively, given by Eqs. (20) and (21).

[2] For two-class problems with n_1 and n_2 instances of the two classes in the training set, the diagonal terms of $\Sigma_{\mathbf{w}^*}, \Sigma_{\mathbf{x}_0^*}$ and $\Sigma_{\mathbf{w}^*, \mathbf{x}_0^*}$ become respectively $\frac{1}{\sigma_i^2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$, $\frac{\sigma_i^2}{4}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$, for $i = 1, \ldots, d$, and $\frac{1}{2}\left(\frac{1}{n_1} - \frac{1}{n_2}\right)$.

4.4. Case 3: identical covariance matrices

Here we discuss the case of identical covariance matrices having a general form:

$$\Sigma_1 = \Sigma_2 = \Sigma \ . \tag{22}$$

In this case the features are correlated, which does not allow us to analytically derive all the elements of the covariance matrix Σ_ξ of the parameter distribution of the "bagged" classifiers. Nevertheless, by performing an appropriate rotation of the feature space, we obtain the diagonal matrix $A^{-1}\Sigma A$ (where A is the eigenvector matrix) whose elements are the eigenvalues $\lambda_1, \ldots, \lambda_d$ of Σ. This leads us to the case already discussed in Sect. 4.3. The distribution of the parameter Ψ^*, for the different classifiers considered, is therefore the one derived in Sect. 4.3, where σ_i^2 is replaced by λ_i, $i = 1, \ldots, d$, and $\mathbf{w}$ and $\mathbf{x}_0$ refer to the values computed in the rotated feature space. Similarly, the same results presented in Sect. 4.2.2 hold for the corresponding confidence regions.

4.5. Case 4: different, diagonal covariance matrices

Here we present the results when the covariance matrices of the classes are different and have a diagonal form:

$$\Sigma_1 = \vec{\sigma}_1^2 \mathbf{I}, \quad \Sigma_2 = \vec{\sigma}_2^2 \mathbf{I}, \quad \vec{\sigma}_1 \ \neq \ \vec{\sigma}_2 \ , \tag{23}$$

where $\vec{\sigma}_k^2 = (\sigma_{k,1}^2, \ldots, \sigma_{k,d}^2)$, $k = 1, 2$.

4.5.1. Parameter distribution of the "bagged" classifiers

The QDC is the optimal classifier under assumption (23). Its decision function is given by Eqs. (8) and (9). Due to assumption (4), the quantity $\mathbf{W} = \frac{1}{2}(\Sigma_2^{-1} - \Sigma_1^{-1})$ is a constant term. Accordingly, the parameter of the "bagged" QDC classifier whose distribution we have to derive is $\Psi^* = (\mathbf{w}^*, w_0^*)$. First, according to Eq. (9) we have:

$$\mathbf{w}^* = \Sigma_1^{-1}\mathbf{m}_1^* - \Sigma_2^{-1}\mathbf{m}_2^* \ . \tag{24}$$

It is easy to see that $\mathbf{w}^*$ approximately follows a multivariate Normal distribution:

$$\mathcal{N}\left(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2, \frac{1}{n}\Sigma_1^{-1} + \frac{1}{n}\Sigma_2^{-1}\right) \ . \tag{25}$$

Next, to derive the distribution of w_0^* it is convenient to multiply it by the number of training instances of each class, n, and to rewrite the resulting quantity (see Eq. (9)) as $w_{0,1}^* - w_{0,2}^*$, where:

$$w_{0,k}^* = n(\mathbf{m}_k^*)^\top \Sigma_k^{-1}(\mathbf{m}_k^*) = n\sum_{i=1}^{d}\left(\frac{m_{k,i}^*}{\sigma_{k,i}}\right)^2 \ , \ k = 1, 2. \tag{26}$$

Consider now that $\{\mathbf{m}_{k,i}^*\}_{i=1,\dots,d}$, $k = 1, 2$, are independent random variables, and their distribution is approximately Gaussian; this implies:

$$\frac{\sqrt{n}\mathbf{m}_{k,i}^*}{\sigma_{k,i}} \ \sim \ \mathcal{N}(\frac{\sqrt{n}\mu_{k,i}}{\sigma_{k,i}}, 1), \quad i = 1, \dots, d, \quad k = 1, 2 \ . \tag{27}$$

It follows that the random variables $w_{0,k}^*$ in Eq. (26) approximately follow noncentral Chi Squared distributions with d degrees of freedom [21]:

$$w_{0,k}^* \ \sim \ \chi^2(d, \rho_k), \quad k = 1, 2 \ , \tag{28}$$

where ρ_k is given by:[3]

$$\rho_k = n\mu_k^\top \Sigma_k^{-1}\mu_k = n\sum_{i=1}^{d}\left(\frac{\mu_{k,i}}{\sigma_{k,i}}\right)^2 \ . \tag{29}$$

We point out that the components of the random variable $\mathbf{w}^* = (w_1^*, \dots, w_d^*)$ are independent on each other (due to assumption (23)), but they are not independent on $w_{0,1}^*$ and $w_{0,2}^*$. For instance, the covariance between the first component of $\mathbf{w}^*$ and $w_{0,1}^*$, calculated under the assumption of independent features, is $\mathrm{cov}\left(w_1^*, w_{0,1}^*\right) = \frac{2\mu_{1,1}}{\sigma_{1,1}^2}$. In the same way one obtains the covariance between the other components of $\mathbf{w}^*$ and $w_{0,1}^*$ or $w_{0,2}^*$.

Consider finally the "bagged" LDC and NMC, which are suboptimal under assumption (23). Their parameter distribution is the one we obtained in Sects. 4.2–4.4, where $\Sigma = \frac{1}{2}(\Sigma_1 + \Sigma_2)$, depending on the form of Σ.

[3] For the sake of simplicity we do not consider the case when the number of instances from the two classes is different. In this case the results are very similar to the case of identical number of instances, as we have already shown in previous sections for NMC and LDC.

4.5.2. Confidence regions for the distribution parameters

The confidence region for the multivariate Gaussian random variable $\mathbf{w}^*$ given by Eq. (24) can be computed by means of the T^2 Hotelling test, as previously discussed. The set of hypotheses that has to be tested is indeed:

$$\begin{cases} H_0 : \ \omega_{1i} = \omega_{0i} & \forall i = 1, \ldots, d, \\ H_1 : \omega_{1i} \neq \omega_{0i} & \text{for some } i, \end{cases} \tag{30}$$

where $\omega_0 = (\omega_{01}, \ldots, \omega_{0d})$ is a known vector. We accept the hypothesis H_0 with probability $1 - \alpha$, if:

$$Pr\{T^2_{\mathbf{w}^*} < T^2_d(n-1, \alpha)\} = 1 - \alpha \ , \tag{31}$$

where $T^2_{\mathbf{w}^*} = (\mathbf{m}_{\mathbf{w}^*} - \omega_0)^\top \mathbf{S}^{-1}_{\mathbf{w}^*}(\mathbf{m}_{\mathbf{w}^*} - \omega_0)$, whereas $\mathbf{m}_{\mathbf{w}^*}$ and $\mathbf{S}_{\mathbf{w}^*}$ are the estimates of the distribution parameters of $\mathbf{w}^*$ in Eq. (25). The derivation of the confidence region for the non-central Chi-Squared variables $w^*_{0,1}$ and $w^*_{0,2}$ is more difficult [15], and we omit it for the sake of simplicity.

4.6. General case: non-Gaussian or unknown data distribution

Up to now we derived the distribution of the parameters of "bagged" classifiers under the assumption that the data has a multivariate Gaussian distribution with known class-conditional means μ_k and covariance matrices Σ_k. In practice one has no access to the true values of μ_k and Σ_k. Nevertheless, in the case of Gaussian data distribution, all the above results still hold by further approximating the distribution (3) of the random variable $\mathbf{m}^*_k$ by the following Gaussian distribution, in which the sample means $\mathbf{m}_k$ and covariance matrices $\mathbf{S}_k$ (estimated from T, see Eq. (2)) are used in place of μ_k and Σ_k:[4] *i.e.*:

$$\mathcal{N}\left(\mathbf{m}_k, \frac{1}{n}\mathbf{S}_k\right) , \ k = 1, 2 \ . \tag{32}$$

[4] Although in a boostrap replicate T^* of size $2n$ the number of samples from each class can be different from n, Eq. (32) is a good approximation for a sufficiently large value of n, thanks to the CLT.

Moreover, thanks to the CLT the distribution of $\mathbf{m}_k^*$ is approximated with good accuracy by Eq. (32) even if the underlying data distribution is non-Gaussian, provided that the sample size n is sufficiently large as already mentioned above (say, $n > 30$, although in practice even a small value of n is enough, as we will show in Sect. 5.1). In particular, this allows the above results to be exploited also in the practical case of unknown data distribution.

According to our approach and to the above results, we are now in the position of presenting the procedure for constructing an ensemble of N NMC, LDC or QDC classifiers, using our approach to simulate Bagging, in a practical setting with unknown data distribution. Given the decision functions of such classifiers in Eqs. (5)–(9), one has to independently sample N instances of their parameters, *i.e.*, $\Psi^*(j) = (\mathbf{w}^*(j), \mathbf{x}_0^*(j))$ for NMC and LDC, and $\Psi^*(j) = (\mathbf{w}^*(j), w_0^*(j))$ for the QDC, with $j = 1, \ldots, N$.

The corresponding distributions depend on μ_k and Σ_k, that we approximate with $\mathbf{m}_k$ and $\mathbf{S}_k$, $k = 1, 2$. Note that, generally, the sample covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ are different and non-diagonal. For each base classifier, the distributions are the following ones:

NMC : the distribution of $\Psi^* = (\mathbf{w}^*, \mathbf{x}_0^*)$ is approximated by a multivariate Gaussian $\mathcal{N}(\xi, \Sigma_\xi)$ as in Sect. 4.2; the values of its mean ξ and covariance matrix Σ_ξ are given by Eqs. (13) and (15), where the scalar σ^2 is approximated by the mean value of the diagonal elements of $\frac{1}{2}(\mathbf{S}_1 + \mathbf{S}_2)$.

LDC : according to Sect. 4.4, the distribution of $\Psi^* = (\mathbf{w}^*, \mathbf{x}_0^*)$ is a multivariate Gaussian $\mathcal{N}(\xi, \Sigma_\xi)$, and the values of ξ and Σ_ξ are given respectively by Eqs. (20) and (21). In practice, the vector $\vec{\sigma}^2$ is approximated by the diagonal of the matrix $\mathbf{S} = \frac{1}{2}(\mathbf{S}_1 + \mathbf{S}_2)$. If needed, we first rotate the feature space such that the resulting $\mathbf{S}$ is diagonal (its elements are the eigenvalues of the original matrix $\mathbf{S}$, as explained in Sect. 4.4).

QDC : in Sect. 4.5 we derived the distribution of the parameter $\Psi^* = (\mathbf{w}^*, w_0^*)$. In practice, an alternative and easier way to obtain the parameters of the

Table 1: Characteristics of the two-class data sets used in our experiments. The number of instances in each class are shown between brackets.

Data set	**Instances**	**Features** (d)
German	1000 (700+300)	24
Pima	768 (500+268)	8
Breast Cancer	699 (458+241)	9
Blood	748 (570+178)	4
Ilpd	583 (416+167)	9
Bands	365 (135+230)	19
Ionosphere	351 (225+126)	34
Heart	267 (55+212)	44
Uncorrelated Gaussian	1000 (500+500)	10
Correlated Gaussian	400 (200+200)	30

"bagged" QDC according to our approach is to sample only the values of the random variables $\mathbf{m}_1^*$ and $\mathbf{m}_2^*$, whose distributions are approximated by Eq. (32), and to plug them into Eq. (9), in which the covariance matrices Σ_k are approximated by the corresponding $\mathbf{S}_k$.

5. Experiments

To evaluate the proposed randomization approach we carry out experiments on ten different data sets, using the three base classifiers NMC, LDC and QDC. Our first aim is to verify whether and to what extent the distribution of the parameters of classifiers obtained by Bagging can be approximated by the Normal distributions we derived in Sect. 4. Secondly, we compare the classification performance of Bagging and of classifier ensembles obtained by our approach using the same classifier parameter distributions derived for Bagging. Finally, we show an example of a "synthetic" randomization technique obtained by modifying the classifier parameter distributions we derived for Bagging.

The main characteristics of the data sets are reported in Table 1. We used two artificial datasets whose distribution is known, and eight real-world data sets from the UCI repository.[5] Both artificial data sets exhibit Gaussian class-conditional distributions; in 'Uncorrelated Gaussian' they are identical and proportional to the identity matrix, which is the setting investigated in Sect. 4.2 (Eq. (10)), whereas in 'Correlated Gaussian'[6] they are identical to a diagonal matrix such that the variance of the second feature is equal to 40 and the other ones are equal to 1. The 'Correlated Gaussian' data set is also rotated for the first two features using a rotation matrix $\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$.

We randomly subdivided each data set, using stratified sampling, into a training set made up of 80% of the instances and a test set containing the remaining instances. To evaluate the effect of training set size on our approach, we considered four different training sets of increasing size, $2n^{(1)} < 2n^{(2)} < 2n^{(3)} < 2n^{(4)}$ where $2n^{(4)}$ corresponds to the original training set; we then set the smallest size $2n^{(1)}$ equal to the number of features d, which corresponds to the "instability region" where Bagging was found to particularly effective for the considered classifiers in [13] (see Sect. 2); we then set the intermediate sizes $2n^{(2)}$ and $2n^{(3)}$ equally divided between $2n^{(1)}$ and $2n^{(4)}$. We built the training sets of size lower than $2n^{(4)}$ by a stratified sampling from the original training set. For each base classifier and training set size we built two ensembles of $N = 31$ classifiers: one using Bagging, and one using our approach and the parameter distribution idicated in Sect. 4.6. We repeated the above procedure for ten times. All the results will be reported in terms of averages and standard deviations over the ten runs of the experiments.

5.1. Verification of the Gaussianity of the classifier parameters obtained by Bagging

To evaluate the accuracy of the approximation of the distribution of classifier parameters we derived in Sect. 4, we focused on two data sets: the artificial

[5] `http://archive.ics.uci.edu/ml`

[6] This data set has been defined in [13].

Uncorrelated Gaussian data set, and the real-world Breast Cancer data sets. We remind the reader that the former exhibits Gaussian class-conditional covariance matrices, whereas the distribution of the latter is unknown and its features are correlated.

In particular, we evaluated whether the distributions of the vectors $\mathbf{w}^*$ and $\mathbf{x}_0^*$ obtained by Bagging using the NMC and LDC are well approximated by the derived multivariate Normal distributions, using the well known *Jarque-Bera* gaussianity test [16]. It is commonly used for verifying if the input comes from a Normal distribution with unknown parameters, corresponding to the null hypothesis. In particular, when the *p-value* [17] is smaller than 0.05, the test rejects the null hypothesis at the default 5% significance level (which means that the distribution is not Gaussian), otherwise the null hypothesis is accepted (*i.e.*, the distribution is considered Gaussian).

We performed the test for each $\mathbf{w}^*$ and $\mathbf{x}_0^*$ component separately because, if a random vector follows a Gaussian distribution, its individual components are Gaussian random variables too. We also performed the test for two training set sizes: comparison in the instability region (we chose $2n^{(1)} = 10$ for both data sets, since the number of features is 10 for Uncorrelated Gaussian and 9 for Breast Cancer) and for a larger training set size (corresponding to $2n^{(2)} = 300$ and $2n^{(3)} = 300$, respectively).

Let us note that we did not perform the test for the QDC because in this case a different approach was used (as explained in Sec. 4.6). Indeed, it is obvious that we can not get a Gaussian bootstrap replicate distribution by sampling from random variables $\mathbf{m}_1^*$, $\mathbf{m}_2^*$ and plugging the obtained values into Eq. (9).

We show the results in Tables 2, 3 and 4. Tables 2 and 3 show the results for the Uncorrelated Gaussian and Breast Cancer data set, respectively, obtained using NMC as the base classifier, where $\bar{\mathbf{x}}_0^{(s)}$ and $\sigma^2_{\bar{\mathbf{x}}_0^{(s)}}$ denote mean and covariance of each $\mathbf{x}_0$ component obtained by simulating Bagging with our approach (columns 1 and 2), and $\bar{\mathbf{x}}_0^*$ and $\sigma^2_{\bar{\mathbf{x}}_0^*}$ denote the same quantities for the components of $\mathbf{x}_0$ obtained by the original Bagging (columns 3 and 4). We point out that such means and variances were computed over 310 values, given by 31 classifiers × 10 runs of the experiments. Similarly, $\bar{\mathbf{w}}^*$ and $\sigma^2_{\bar{\mathbf{w}}^*}$ denote the same quantities for the components of $\mathbf{w}$ obtained by the simulated (columns 6 and 7) and the original Bagging (columns 8 and 9). In columns 5 and 10 we show the *p-value* related to each $\mathbf{x}_0$ and $\mathbf{w}$ component, respectively.

According to the test, the random variables $\mathbf{x}_0^*$ and $\mathbf{w}^*$ obtained from Bagging follow Gaussian distributions for both data sets, and for both training set sizes. Indeed, the p-value is always greater than the default value 0.05 except for sporadic cases in the instability region. In particular, the p-value increases in both cases as the training set size increases, which is in agreement with the CLT.

We point out that, for the two data sets above, in the instability region the training set size $2n^{(1)} = 10$ is lower than 30, which, according to the heuristic rule mentioned in Sect. 4, is the minimum value which justifies the application of the CLT. This fact provides evidence that, as we mentioned in Sect. 4, the distribution of the classifier parameters obtained by Bagging can be well approximated by a Gaussian also for a training set size lower than 30, even if the original data distribution is not Gaussian.

We finally point out that the average parameter values obtained by the simulated Bagging are very close to the ones of the original Bagging.

In Table 4, we show the results obtained using the LDC as base classifier. Since the vector $\mathbf{x}_0$ is identical to the one of the NMC classifier, we omit it. For this classifier it was not possible to compute the parameter $\mathbf{w}^*$ for a training set size equal to $2n^{(1)}$, since in the instability region the covariance matrix of the data is ill-conditioned. On the other hand, for a training set size $2n = 300$, the random variable $\mathbf{w}^*$ (as well as $\mathbf{x}_0$, see above) is well approximated by a Gaussian distribution for both data sets, as in the case of the NMC.

Accordingly, Bagging can be effectively simulated by our approach for the NMC and LDC classifiers using the distribution of classifier parameters we derived in Sect. 4, also in the practical cases of data sets with unknown distribution.

5.2. Performance comparison

In this section we further compare the original Bagging with its implementation based on our approach (that we implemented according to Sect. 4.6), in terms of their classification performance. For each base classifier and training set size we report the error rate of Bagging, $\mathrm{E_b}$, and the relative difference between it and the error rate $\mathrm{E_s}$ of our approach (where 's' stands for "simulated Bagging"), given by $\frac{E_\mathrm{b}-E_\mathrm{s}}{E_\mathrm{b}}$. The results are reported in Tables 5–7, respectively for the NMC, LDC and QDC. We report only the average values over the ten runs of our experiments, because the variance was alwys very small: its maximum value (over all classifiers and data sets) was about 0.06, but most of its values were of the order of 10^{-3}. In some cases (denoted by "–" in the tables) it was not possible to use the LDC and QDC, due to their singular covariance matrices. We also summarize the relative differences between the error rates in Fig. 5.2.

Tables 5 and 6 show that our approach provided a classification performance very close to Bagging, when the NMC and LDC were used as the base classifier. The only exception is the LDC on Breast Cancer and Blood datasets: in both cases our approach attained a significantly greater error rate than Bagging, in the first dataset for the training set size equal to $2n^{(1)}$, and in the second one for training set sizes equal to $2n^{(3)}$ and $2n^{(4)}$.

When the QDC was used as the base classifier, Table 7 shows a similar trend, with a few more exceptions: our approach attained a significantly greater error rate than Bagging on Breast Cancer (for a training set size of $2n^{(2)}$), Blood (for all training set sizes) and Correlated Gaussian (for the $2n^{(3)}$ and $2n^{(4)}$ training set sizes).

The reason of the lower performance attained by our approach when using the LDC and QDC can be twofold. Firstly, if the covariance matrix of the data is ill-conditioned, the accuracy of the approximation of the classifier parameter distribution decreases, as it depends on the inverse of the covariance matrix. Secondly, our assumption that the covariance matrix of bootstrap replicates is constant and equal to the one of the original training set (see Eq. (4)) may be not accurate enough in some cases.

Table 2: NMC base classifier, Uncorrelated Gaussian data set. Comparison between mean value and variance of the $\mathbf{x}_0^*$ and $\mathbf{w}^*$ components of the classifier parameter obtained by our approach (columns 1-4) and by Bagging (columns 6-9), for training set sizes $2n^{(1)} = 10$ and $2n^{(2)} = 300$. The p-value for the Normality test (see text) is shown for all the elements of the above vectors obtained by Bagging: a value higher than 0.05 means that the corresponding random variable has a Normal distribution, at the default 5% significance level.

$2n = 10$									
Our approach		**Bagging**			**Our approach**		**Bagging**		
$\bar{\mathbf{x}}_0^{(s)}$	$\sigma^2_{\bar{\mathbf{x}}_0^{(s)}}$	$\bar{\mathbf{x}}_0^*$	$\sigma^2_{\bar{\mathbf{x}}_0^*}$	**p-value**	$\bar{\mathbf{w}}^{(s)}$	$\sigma^2_{\bar{\mathbf{w}}^{(s)}}$	$\bar{\mathbf{w}}^*$	$\sigma^2_{\bar{\mathbf{w}}^*}$	**p-value**
0.4403	0.0931	0.4400	0.0990	0.1469	0.5878	0.3842	0.6489	0.3935	0.3455
0.3857	0.0892	0.3373	0.0990	0.0266	-0.7716	0.4037	-0.6758	0.3935	0.4728
0.7065	0.0823	0.6572	0.0929	0.0834	0.4011	0.3615	0.3153	0.3331	0.0733
0.6414	0.1020	0.6813	0.0868	0.0187	0.1434	0.4264	0.0212	0.4065	0.5000
0.4841	0.1027	0.5468	0.1074	0.5000	-0.4147	0.3768	-0.2477	0.4113	0.0015
0.8863	0.1102	0.8477	0.0848	0.5000	0.4958	0.3852	0.3079	0.3650	0.5000
0.4627	0.0948	0.4830	0.0976	0.5000	-0.3146	0.3392	-0.2643	0.4209	0.2741
0.5203	0.1109	0.4764	0.1172	0.5000	0.4820	0.3945	0.4082	0.4561	0.2204
0.7066	0.1143	0.7031	0.0962	0.0867	-0.2706	0.3986	-0.4562	0.3594	0.5000
0.7450	0.0765	0.7418	0.0803	0.2839	-0.0418	0.3623	-0.0799	0.3311	0.5000
$2n = 300$									
Our approach		**Bagging**			**Our approach**		**Bagging**		
$\bar{\mathbf{x}}_0^{(s)}$	$\sigma^2_{\bar{\mathbf{x}}_0^{(s)}}$	$\bar{\mathbf{x}}_0^*$	$\sigma^2_{\bar{\mathbf{x}}_0^*}$	**p-value**	$\bar{\mathbf{w}}^{(s)}$	$\sigma^2_{\bar{\mathbf{w}}^{(s)}}$	$\bar{\mathbf{w}}^*$	$\sigma^2_{\bar{\mathbf{w}}^*}$	**p-value**
0.4644	0.0069	0.4702	0.0056	0.5000	0.5435	0.0222	0.5315	0.0174	0.5000
0.3980	0.0063	0.4011	0.0056	0.5000	-0.6715	0.0252	-0.6746	0.0209	0.1695
0.6647	0.0058	0.6597	0.0052	0.3223	0.3376	0.0201	0.3591	0.0209	0.5000
0.6613	0.0051	0.6582	0.0062	0.5000	0.0868	0.0231	0.0724	0.0207	0.5000
0.5131	0.0053	0.5037	0.0070	0.3422	-0.3748	0.0244	-0.3831	0.0214	0.5000
0.8822	0.0053	0.8794	0.0057	0.1412	0.4639	0.0221	0.4495	0.0241	0.5000
0.4401	0.0053	0.4305	0.0047	0.1244	-0.3001	0.0235	-0.3033	0.0244	0.3213
0.4592	0.0070	0.4649	0.0057	0.5000	0.4066	0.0202	0.4100	0.0235	0.2833
0.7219	0.0054	0.7212	0.0057	0.5000	-0.3173	0.0276	-0.3455	0.0243	0.5000
0.7084	0.0054	0.7079	0.0048	0.4572	-0.0302	0.0206	0.0038	0.0215	0.4309

5.3. Synthesizing new randomization techniques: an example

The above experiments showed that Bagging can be also implemented according to our approach, using the considered base classifiers, thanks to the

Table 3: NMC base classifier, Breast Cancer data set. See caption of Table 2 for the details.

$2n = 10$									
Our approach		**Bagging**			**Our approach**		**Bagging**		
$\bar{\mathbf{x}}_0^{(s)}$	$\sigma^2_{\bar{\mathbf{x}}_0^{(s)}}$	$\bar{\mathbf{x}}_0^*$	$\sigma^2_{\bar{\mathbf{x}}_0^*}$	**p-value**	$\bar{\mathbf{w}}^{(s)}$	$\sigma^2_{\bar{\mathbf{w}}^{(s)}}$	$\bar{\mathbf{w}}^*$	$\sigma^2_{\bar{\mathbf{w}}^*}$	**p-value**
0.5173	0.0046	0.5028	0.0036	0.0769	-0.4322	0.0166	-0.4295	0.0143	0.1455
0.3976	0.0034	0.3973	0.0036	0.0607	-0.5166	0.0149	-0.5266	0.0143	0.0014
0.3925	0.0035	0.4013	0.0030	0.5000	-0.5028	0.0135	-0.5095	0.0162	0.0874
0.3381	0.0042	0.3445	0.0050	0.3038	-0.3950	0.0169	-0.4313	0.0222	0.5000
0.3659	0.0035	0.3638	0.0033	0.0217	-0.2958	0.0125	-0.3086	0.0131	0.0124
0.4575	0.0051	0.4572	0.0054	0.0814	-0.6217	0.0206	-0.6151	0.0208	0.0011
0.3992	0.0041	0.4057	0.0031	0.2081	-0.3760	0.0155	-0.4008	0.0127	0.5000
0.3297	0.0048	0.3674	0.0055	0.0662	-0.4672	0.0229	-0.4702	0.0232	0.5000
0.1789	0.0035	0.1882	0.0034	0.0010	-0.1596	0.0140	-0.1542	0.0134	0.0121
$2n = 300$									
Our approach		**Bagging**			**Our approach**		**Bagging**		
$\bar{\mathbf{x}}_0^{(s)}$	$\sigma^2_{\bar{\mathbf{x}}_0^{(s)}}(10^{-3})$	$\bar{\mathbf{x}}_0^*$	$\sigma^2_{\bar{\mathbf{x}}_0^*}(10^{-3})$	**p-value**	$\bar{\mathbf{w}}^{(s)}$	$\sigma^2_{\bar{\mathbf{w}}^{(s)}}(10^{-3})$	$\bar{\mathbf{w}}^*$	$\sigma^2_{\bar{\mathbf{w}}^*}(10^{-3})$	**p-value**
0.5095	0.1989	0.5061	0.2030	0.5000	-0.4291	0.7989	-0.4219	0.7421	0.5000
0.3940	0.2077	0.3944	0.2030	0.3722	-0.5236	0.8155	-0.5226	0.7421	0.4137
0.4003	0.2051	0.3999	0.1650	0.4124	-0.5106	0.7703	-0.5091	0.7163	0.0411
0.3469	0.2295	0.3467	0.2569	0.5000	-0.4188	0.8968	-0.4147	0.9767	0.1936
0.3698	0.1949	0.3708	0.1782	0.5000	-0.3192	0.7778	-0.3165	0.6256	0.2785
0.4513	0.2242	0.4508	0.2821	0.4785	-0.6207	0.8866	-0.6180	0.9973	0.4089
0.4043	0.1995	0.4035	0.1798	0.1910	-0.3879	0.7424	-0.3869	0.6274	0.4889
0.3572	0.2034	0.3563	0.3057	0.4836	-0.4524	0.8493	-0.4563	1.1133	0.5000
0.1808	0.1935	0.1810	0.1453	0.5000	-0.1527	0.7895	-0.1524	0.6306	0.5000

accuracy of the approximation of the corresponding classifier parameter distribution that we derived in Sect. 4.

Based on this result, we further investigated a possible implementation of a "synthetic" randomization technique obtained by directly defining the distribution of the classifier parameters. As explained in Sect. 3, in our first attempt we started by modifying the distribution obtained by Bagging. To this aim, we modified the covariance matrix Σ_ξ of Eq. (15) into a new covariance matrix defined as $\Sigma_\xi^{'} = n\Sigma_\xi$, *i.e.*:

$$\Sigma_{\mathbf{x}_0^*} = \frac{\Sigma}{2}, \quad \Sigma_{\mathbf{w}^*} = 2\Sigma, \quad \Sigma_{\mathbf{x}_0^*,\mathbf{w}^*} = \mathbf{0}_{d\times d} \ . \tag{33}$$

Table 4: LDC base classifier, Uncorrelated Gaussian and Breast Cancer datasets. Comparison between the mean value and variance of the $\mathbf{w}^*$ component of the classifier parameter obtained by our approach and by Bagging (columns 1–2 and 3–4, respectively), for a training set size $2n = 300$, and p-value for the Normality test (see caption of Table 2 for the details).

Uncorrelated Gaussian ($2n = 300$)				
Our approach		**Bagging**		
$\bar{\mathbf{w}}^{(s)}$	$\sigma^2_{\bar{\mathbf{w}}^{(s)}}$	$\bar{\mathbf{w}}^*$	$\sigma^2_{\bar{\mathbf{w}}^*}$	**p-value**
0.5672	0.0207	0.5795	0.0276	0.5000
-0.6781	0.0300	-0.6618	0.0276	0.5000
0.4049	0.0391	0.3656	0.0360	0.5000
0.0070	0.0262	0.0092	0.0340	0.5000
-0.4258	0.0249	-0.3602	0.0266	0.5000
0.4622	0.0188	0.4743	0.0400	0.5000
-0.4187	0.0330	-0.4022	0.0363	0.5000
0.4642	0.0186	0.4688	0.0346	0.5000
-0.3295	0.0374	-0.3102	0.0467	0.0665
-0.0178	0.0211	-0.0264	0.0273	0.5000
Breast Cancer ($2n = 300$)				
Our approach		**Bagging**		
$\bar{\mathbf{w}}^{(s)}$	$\sigma^2_{\bar{\mathbf{w}}^{(s)}}$	$\bar{\mathbf{w}}^*$	$\sigma^2_{\bar{\mathbf{w}}^*}$	**p-value**
-9.9882	10.8395	-8.7553	9.0166	0.5000
-5.7018	10.8395	-5.4385	9.0166	0.0208
-4.2521	9.5173	-3.4918	6.9907	0.5000
-1.7440	4.5151	-1.4015	4.9878	0.0277
-2.2664	7.1830	-2.2900	8.9879	0.5000
-11.3144	6.1746	-11.6187	7.0698	0.0164
-3.6291	5.9336	-4.4045	7.0778	0.0910
-4.3743	3.8106	-3.8320	4.8822	0.5000
1.2484	7.9784	0.3680	8.1049	0.5000

Intuitively, we expect that increasing the variance of the parameter distribution results in a greater diversity among the individual classifiers, and thus in their lower average accuracy. In other words, this should shift the accuracy-diversity trade-off in favour of a higher diversity. Accordingly, for comparison we also applied the *Nice Bagging* technique proposed in [8] for discarding “poor” individual classifiers: it simply works by adding to the ensemble only classifiers exhibiting a lower error rate than the one of a base classifier trained on the

Table 5: NMC base classifier: error rates of Bagging for the different training set sizes.

	Training set size			
Dataset	$2n^{(1)}$	$2n^{(2)}$	$2n^{(3)}$	$2n^{(4)}$
Uncorrelated Gaussian	0.346	0.303	0.295	0.294
German	0.458	0.362	0.368	0.369
Pima	0.440	0.346	0.340	0.344
Breast	0.044	0.041	0.041	0.041
Blood	0.372	0.327	0.333	0.327
Correlated Gaussian	0.290	0.285	0.296	0.300
Ilpd	0.476	0.444	0.450	0.446
Bands	0.415	0.411	0.415	0.414
Ionosphere	0.225	0.259	0.259	0.259
Heart	0.370	0.352	0.346	0.343

original training set, where the error rate is estimated on the same training set.

We carried out these experiments using only the NMC as base classifier. For reference, we compare the classification performance of our approach, with and without using the Nice Bagging technique, with the performance of the original Nice Bagging and Bagging, respectively. The results are reported in Table 8 and Fig. 5.3.

It can be seen that our “synthetic” randomization technique attained a performance which is “reasonable”, in the sense that it is close to the one of “real” randomization techniques like Bagging and Nice Bagging; in particular, it even outperformed such techniques on Breast Cancer (for all training set sizes), Blood and Ilpd (for training set sizes $2n^{(3)}$ and $2n^{(4)}$), and Heart (except for the smallest training set size).

This is a first evidence of the viability of our alternative randomization approach, which motivates further investigations on criteria for defining suitable distributions of classifier parameters.

6. Discussion and conclusions

We proposed a novel approach to randomization-based techniques for classifier ensemble construction. It is based on modeling the probability distribution

Table 6: LDC base classifier: error rates of Bagging for the different training set sizes.

	Training set size			
Dataset	$2n^{(1)}$	$2n^{(2)}$	$2n^{(3)}$	$2n^{(4)}$
Uncorrelated Gaussian	0.398	0.271	0.263	0.264
German	0.452	0.306	0.276	0.233
Pima	0.323	0.244	0.231	0.234
Breast	0.088	0.035	0.037	0.038
Correlated Gaussian	0.169	0.071	0.066	0.070
Ilpd	0.471	0.399	0.315	0.305
Heart	0.428	0.333	0.287	0.269
Blood	0.431	0.355	0.207	0.207
Bands	0.441	0.382	0.341	0.337

Table 7: QDC base classifier: error rates of Bagging for the different training set sizes.

	Training set size			
Dataset	$2n^{(1)}$	$2n^{(2)}$	$2n^{(3)}$	$2n^{(4)}$
Uncorrelated Gaussian	0.500	0.363	0.326	0.289
Pima	0.381	0.321	0.275	0.287
Breast	–	0.072	0.064	0.045
Blood	–	0.381	0.375	0.257
Correlated Gaussian	0.499	0.391	0.175	0.145
Ilpd	0.438	0.420	0.443	0.426
Heart	0.515	0.206	0.204	0.211

of the parameters of a given base classifier induced by the use of a given randomization technique, for a given learning algorithm and training set. The classifiers of the ensemble can then be obtained by directly and independently sampling their parameter values from such a distribution, instead of actually manipulating the training data and running the chosen learning algorithm for each of them.

On the one hand, our approach can be applied to existing randomization techniques only if the induced parameter distribution can be analytically derived or can be at least approximated, which can be difficult from some techniques, base classifiers and learning algorithms. In this case, the main practical

Table 8: Error rates of Bagging and Nice Bagging, using NMC as the base classifier.

		Training set size			
Dataset	**Method**	$2n^{(1)}$	$2n^{(2)}$	$2n^{(3)}$	$2n^{(4)}$
UncGaussian	Bagging	0.346	0.289	0.283	0.280
	Nice Bagging	0.230	0.306	0.320	0.321
German	Bagging	0.403	0.366	0.367	0.366
	Nice Bagging	0.353	0.351	0.353	0.350
Pima	Bagging	0.382	0.350	0.343	0.340
	Nice Bagging	0.418	0.333	0.346	0.340
Breast	Bagging	0.044	0.039	0.039	0.040
	Nice Bagging	0.0393	0.0364	0.0364	0.0364
Blood	Bagging	0.349	0.340	0.343	0.340
	Nice Bagging	0.349	0.348	0.334	0.334
CorrGaussian	Bagging	0.315	0.308	0.308	0.306
	Nice Bagging	0.230	0.306	0.320	0.321
Ilpd	Bagging	0.451	0.442	0.449	0.454
	Nice Bagging	0.458	0.465	0.445	0.448
Bands	Bagging	0.408	0.390	0.392	0.392
	Nice Bagging	0.437	0.429	0.427	0.427
Ionosphere	Bagging	0.276	0.272	0.282	0.269
	Nice Bagging	0.193	0.230	0.206	0.207
Heart	Bagging	0.341	0.337	0.337	0.335
	Nice Bagging	0.382	0.370	0.365	0.359

advantage is the reduction of processing cost in the ensemble construction stage.

On the other hand, this approach may also enable the definition of novel, "synthetic" randomization techniques, by explicitly defining a suitable parameter distribution for a given base classifier, without deriving it from an actual data manipulation procedure. The crucial point in this case is to define a distribution that can provide a good trade-off between accuracy and diversity of the resulting classifiers, in terms of the corresponding ensemble performance. To this aim, a useful starting point is the analysis of the parameter distribution induced by existing randomization techniques. In this work we made a first step in this direction: first, we analytically derived the parameter distribution induced by Bagging on three well-known classifiers; second, we modified the derived distribution (only for the NMC, as a preliminary attempt) to obtain a different, “synthetic” randomization technique.

Our results have shown that the parameter distribution induced by Bagging on the considered classifiers can be approximated by a Gaussian, parametrized by the statistics estimated from the original training set. The accuracy of such an approximation increases as the training set size n increases, and is already good for $n \simeq 30$. This is witnessed by the corresponding ensemble performance, which turned out to be very close to the one of Bagging on several benchmark data sets, especially when the base classifier is the NMC.

The accuracy of our approximation could be further increased by deriving also the distribution of the sample covariance matrix of bootstrap replicates of the training set, that in our derivations was assumed to be a constant value.

Our analytical results are limited to a single randomization technique (Bagging) and to three classifiers (NMC, LDC and QDC). Deriving the parameter distribution for other techniques or classifiers (e.g., Bagging applied to neural networks, or the Random Forests technique) can be more difficult, but is an obvious and very interesting follow-up of this work. Another necessary and interesting follow-up is to investigate criteria for defining novel randomization techniques through the direct definition of the corresponding parameter distribution.

Acknowledgement

This work has been partly supported by the project ”Computational quantum structures at the service of pattern recognition: modeling uncertainty” [CRP-59872] funded by Regione Autonoma della Sardegna, L.R. 7/2007, Bando 2012.

References

[1] L.I. Kuncheva, Combining Pattern Classifiers, Second Edition, John Wiley & Sons, Inc, Hoboken, NJ, USA, 2014.

[2] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, Chapman & Hall/CRC, 2012.

[3] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Patt. Anal. Mach. Intell. 20 (1998) 832–844.

[4] L. Breiman, Random Forests, Machine Learning. 45 (2001) 5–32.

[5] L. Breiman, Bagging Predictors, Machine Learning. 24 (1996) 123–140.

[6] J.J. Rodriguez, L.I. Kuncheva, C.J. Alonso, Rotation Forest: A New Classifier Ensemble Method, IEEE Trans. Patt. Anal. Mach. Intell. 28 (2006) 1619–1630.

[7] Y. Freund, R.E. Schapire, Experiments with a New Boosting Algorithm, In Int. Conf. on Machine Learning (1996), pp. 148–156.

[8] M. Skurichina, R.P.W. Duin, Bagging for linear classifiers, Pattern Recognition. 31 (1998) 909–930.

[9] G. Fumera, F. Roli, A. Serrau, A Theoretical Analysis of Bagging as a Linear Combination of Classifiers, IEEE Trans. Pattern Anal. Machine Intell. 30, 1293–1299.

[10] R. Tibshirani, Bias, Variance and Prediction Error for Classification Rules, (1996).

[11] D.H. Wolpert, W.G. Macready, An Efficient Method To Estimate Bagging's Generalization Error, Machine Learning 35 (1999) 41–55.

[12] Y. Grandvalet, Bagging Equalizes Influence, Machine Learning 55 (2004) 251–270.

[13] M. Skurichina, R.P.W. Duin, Bagging, Boosting and the Random Subspace Method for Linear Classifiers, Patt. Anal. Appl. 5 (2002) 121–135.

[14] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall/CRC, 1993.

[15] J.T. Kent, T.J. Hainsworth, Confidence intervals for the noncentral chi-squared distribution, J. of Stat. Planning and Inference. 46 (1995) 147–159.

[16] C.M. Jarque, A.K. Bera, A Test for Normality of Observations and Regression Residuals, Int. Stat. Rev. 55 (1987) 163.

[17] R.A. Fisher, Statistical Methods for Research Workers, (1925), Oliver & Boyd, Edinburgh.

[18] B. Mandelbrot, The Pareto-Levy Law and the Distribution of Income, International Economic Review 1 (1960) 79.

[19] R.V. Hogg and A. T. Craig: Introduction to Mathematical Statistics. The Macmillan Company, New York (1978).

[20] H. Hotelling, The Generalization of Student??s Ratio, in: Breakthroughs in Statistics, Springer, 1992, pp. 54–65.

[21] S. Kotz, N. Balakrishnan, N.L. Johnson, Continuous Multivariate Distributions, John Wiley & Sons, 2005.

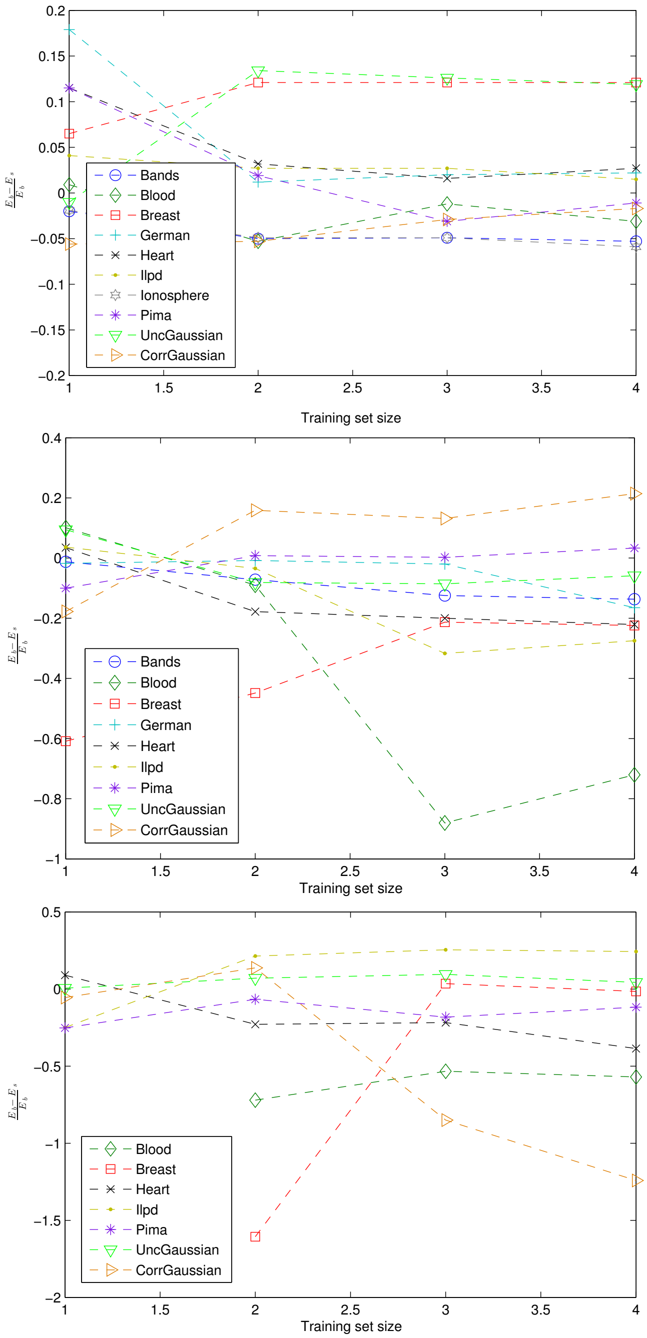

Figure 1: Relative difference between the error rate of Bagging and of our approach that simulates Bagging, for the different training set sizes, on all data sets. Base classifiers: NMC (top), LDC (middle), QDC (bottom).

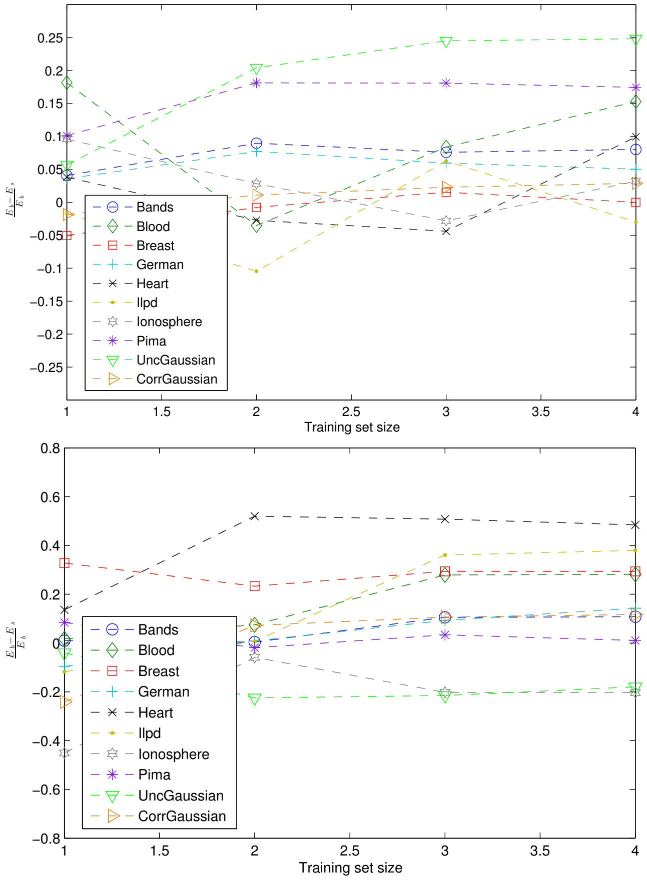

Figure 2: Top: relative difference between the error rate of Bagging and of our "synthetic" randomization technique described in Sect. 5.3, using NMC as the base classifier. Bottom: the same comparison between the "Nice" versions of both techniques.