ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI 28001100967420.

On the Independence of the Individual Predictions in Parallel Randomized Ensembles

Daniel Hernández-Lobato, Gonzalo Martínez-Muñoz and Alberto Suárez [*]

Escuela Politécnica Superior,
Universidad Autónoma de Madrid,
C/ Francisco Tomás y Valiente, 11, Madrid 28049 Spain
{daniel.hernandez, gonzalo.martinez, alberto.suarez}@uam.es

Abstract. In randomized parallel ensembles the class label predictions for a particular instance by different ensemble classifiers are independent random variables. Taking advantage of this independence we design a statistical test to identify instances near the decision borders, which are difficult to classify because of their proximity to these borders. For these instances, the performance of the ensemble is poor and approaches random guessing. The validity of this analysis and the usefulness of the proposed statistical test are illustrated in several real-world classification problems.

1 Introduction

Randomized parallel ensembles are composed of predictors built in independent applications of a randomized learning algorithm on a fixed set of labeled examples. By construction, the predictions of different ensemble members on a fixed test example are independent random variables, when conditioned to the training data. The independence of these predictions implies that the joint probability of error on a given instance factorizes as well. By contrast, when averaged over all instances, the errors of different predictors are not independent. These independence properties have been used in previous work to analyze the convergence of ensemble predictions [6], to estimate the prediction of the complete ensemble on the basis of a small subset of predictions [7] and to make inference on the prediction of an ensemble of infinite size [8]. The goal of the current investigation is to provide an empirical verification of the independence of the individual class predictions for a particular instance in parallel randomized ensembles (Section 2). We then take advantage of this independence to design a test that identifies data instances that are close to the decision borders (Section 3). For these instances, the predictions of the ensemble are close to random guessing. Therefore, they tend to concentrate most of the classification errors.

2 Predictions in Parallel Randomized Ensembles

The goal of supervised learning is to induce a predictor with good generalization properties from the set of labeled examples $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, y_i)) \in \mathcal{Z}\}_{i=1}^{N_{\text{train}}}$, where $\mathcal{Z} \equiv \mathcal{X} \times \mathcal{Y}$, $\mathcal{X}$ is the space of attributes and $\mathcal{Y}$ is the set of class labels. A

[*]The authors acknowledge financial support from the Spanish Dirección General de Investigación, project TIN2010-21575-C02-02.

predictor $h(\cdot) \in \mathcal{H}$ is a function $h : \mathcal{X} \to \mathcal{Y}$ that, given an unlabeled example $\mathbf{x} \in \mathcal{X}$, assigns a class label $h(\mathbf{x}) \in \mathcal{Y}$ to that example. Ensembles yield an aggregate decision by combining the outputs of a collection of such predictors. A common strategy to construct an ensemble is to use a randomized algorithm $\mathcal{L}_{\boldsymbol{\theta}} : \mathcal{Z} \to \mathcal{H}$ as a base learner. The learning algorithm $\mathcal{L}_{\boldsymbol{\theta}}$ generates a predictor $h(\cdot|\mathcal{D}, \boldsymbol{\theta}) \in \mathcal{H}$ when applied to some set of labeled instances $\mathcal{D} \in \mathcal{Z}$. The random variable $\boldsymbol{\theta}$ encodes the random decisions taken in the process of constructing the individual ensemble predictors [4]. By making T independent applications of the randomized learning algorithm $\mathcal{L}_{\boldsymbol{\theta}}$ on the available training data $\mathcal{D}_{\text{train}}$, one generates the parallel ensemble $\{h_t(\cdot) \equiv h_t(\cdot|\mathcal{D}_{\text{train}}, \boldsymbol{\theta}_t)\}_{t=1}^T$, where $\{\boldsymbol{\theta}_t\}_{t=1}^T$ are independent identically distributed random variables. Examples of these types of ensembles include bagging [3], random forest [4], extra trees [5], rotation forest [11] and class switching ensembles [9]. The nature and dimensionality of $\boldsymbol{\theta}$ depend on the particular base learning algorithm used to generate the individual predictors. In bagging $\boldsymbol{\theta}$ encodes the bootstrap samples used to built the individual predictors in the ensemble. These are obtained by sampling with replacement from the original training data. Therefore $\boldsymbol{\theta}$ is a vector of N_{resample} independent random integers, each of which takes values in the range $\{1, \ldots, N_{\text{train}}\}$ with equal probability. In standard implementations of bagging, $N_{\text{resample}} = N_{\text{train}}$, although other choices are possible and, in some cases, more effective [10]. In random forest $\boldsymbol{\theta}$ also includes a vector of independent random integers between 1 and K per internal node of the decision tree. The integer K is the number of attributes that are used to specify the decision at any given node [4]. In extra trees, additional variables indicate the choice of the threshold used to split an internal node in the decision tree [5].

Consider the predictors $h'(\cdot) \equiv h(\cdot|\mathcal{D}_{\text{train}}, \boldsymbol{\theta}')$ and $h''(\cdot) \equiv h(\cdot|\mathcal{D}_{\text{train}}, \boldsymbol{\theta}'')$, built in independent applications of the randomized learning algorithm $\mathcal{L}_{\boldsymbol{\theta}}$ on the available training data $\mathcal{D}_{\text{train}}$. Since $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}''$ are independent random variables, $h'(\cdot)$ and $h''(\cdot)$ are independent random functions. Hence, when conditioned to the training data, the corresponding predictions for a particular test instance $\mathbf{x}$ are also independent random variables

$$\mathcal{P}(h'(\mathbf{x}) = y', h''(\mathbf{x}) = y'') = \mathcal{P}(h'(\mathbf{x}) = y')\mathcal{P}(h''(\mathbf{x}) = y''), \tag{1}$$

where $y', y'' \in \mathcal{Y}$ are any pair of class labels. In consequence, their prediction errors on a fixed test instance $(\mathbf{x}, y)$ are also independent

$$\mathcal{P}(h'(\mathbf{x}) \neq y, h''(\mathbf{x}) \neq y) = \mathcal{P}(h'(\mathbf{x}) \neq y)\mathcal{P}(h''(\mathbf{x}) \neq y). \tag{2}$$

By contrast, the average prediction error will in general be dependent

$$\begin{aligned} &\mathbb{E}_{\mathbf{x},y}\left[\mathcal{P}(h'(\mathbf{x}) \neq y, h''(\mathbf{x}) \neq y)\right] = \mathbb{E}_{\mathbf{x},y}\left[\mathcal{P}(h'(\mathbf{x}) \neq y)\mathcal{P}(h''(\mathbf{x}) \neq y)\right] \\ &\neq \mathbb{E}_{\mathbf{x},y}\left[\mathcal{P}(h'(\mathbf{x}) \neq y)\right]\mathbb{E}_{\mathbf{x},y}\left[\mathcal{P}(h''(\mathbf{x}) \neq y)\right]. \end{aligned} \tag{3}$$

These relations are valid for any classification problem and for any parallel classification ensemble in which the individual classifiers are generated in independent applications of a randomized learning algorithm.

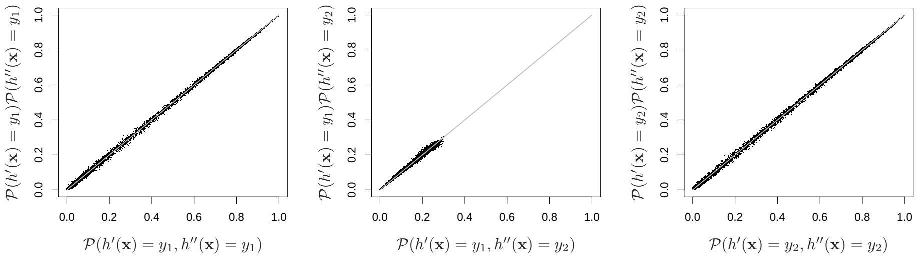

Fig. 1: Empirical estimates of the joint probability of class label predictions by two classifiers from a random forest ensemble in the classification problem *Breast Cancer*.

We now illustrate the independence of the predictions of the individual ensemble classifiers in the binary classification problem *Breast Cancer* from the UCI repository [2] using a random forest ensemble [4]. Similar results should be obtained for any prediction problem and for any parallel randomized ensemble. The experiments consist in generating 100 random partitions of each dataset into a training set and a test set of equal size. For each train and test partition, a random forest ensemble of $T = 1000$ trees is generated. The parameters used for the selection of the random splits in the internal nodes of the decision trees were set to their default values. The estimate of the joint prediction probability (left-hand side of (1)) is the fraction of counts that consecutive pairs of classifiers in the ensemble predict the specified pair of class labels. For the factorized form (right-hand side of (1)), we first estimate the probability of predicting a given class label using only the classifiers in the even positions of the ensemble. We then compute the corresponding estimates with the classifiers in the odd positions of the ensemble. The factorized estimator is the product of these two probabilities for each pair of class labels. Assuming that the independence hypothesis holds, the empirical estimate of the joint distribution and the factorized estimate should agree. Fig. 1 shows the 500 estimates obtained for each pair of class labels. The points in these plots correspond to the prediction probability on a given test instance for two random predictors using the joint estimator (horizontal axis) and the factorized estimator, which assumes independence (vertical axis). In all these plots, the points are aligned along the diagonal, within sample fluctuations. Therefore, both estimators of the joint prediction probability agree. This agreement illustrates the fact that the predictions of the individual ensemble classifiers are independent random variables.

In these experiments we also compare the estimates of the joint error probability error of two different classifiers using the joint estimator and the factorized estimator. The results are displayed in Fig. 2: The left plot compares the two estimates of the joint probability of error on individual test instances (2). The right plot presents the corresponding comparison for the average error (3). These graphs illustrate the fact that the prediction errors on individual test instances are independent. By contrast, the joint probability of the average prediction error does not factorize, which signals the presence of strong dependencies.

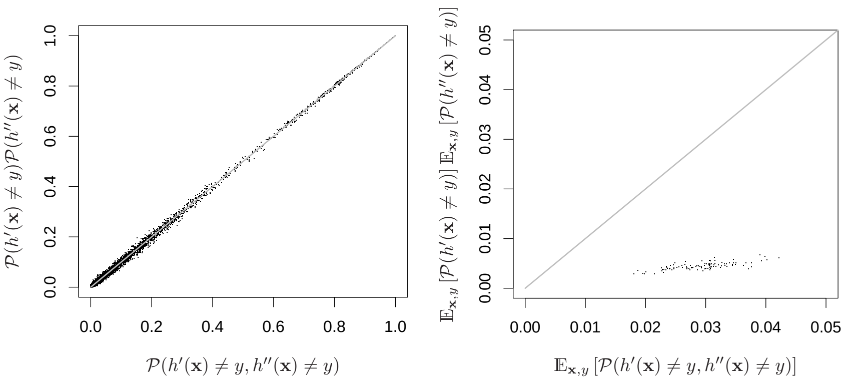

Fig. 2: Empirical estimates of the joint probability of individual (left plot) and average (right plot) prediction errors by two classifiers from a random forest ensemble in the classification problem *Breast Cancer*.

3 Identifying Examples that are Difficult to Classify

Consider a binary classification problem $\mathcal{Y} = \{y_1, y_2\}$. Assume that a parallel randomized ensemble composed of T predictors has been built. Let $\boldsymbol{T} = (T_1 \;\; T_2)$ be the random vector that encodes the ensemble predictions for a given instance $\mathbf{x}$, where T_1 (T_2) is the number of ensemble classifiers that predict class y_1 (y_2) and $T_1 + T_2 = T$. As a consequence of the independence of the predictions of different classifiers, this vector follows a binomial probability distribution

$$\mathcal{P}(\boldsymbol{T}|\boldsymbol{\pi}(\mathbf{x})) = \frac{T!}{T_1!T_2!}\pi_1(\mathbf{x})^{T_1}\pi_2(\mathbf{x})^{T_2}\,, \tag{4}$$

where the components of the random probability vector $\boldsymbol{\pi}(\mathbf{x}) = (\pi_1(\mathbf{x}) \; \pi_2(\mathbf{x}))$, $\pi_1(\mathbf{x})+\pi_2(\mathbf{x}) = 1$, are the probabilities that a classifier from the ensemble assigns to the data instance $\mathbf{x}$ the label y_1 and y_2, respectively. The values of these probabilities depend on the algorithm used to build the base learners, on the particular classification problem and on $\mathbf{x}$, the instance considered. Assuming that majority voting is used, the probability that an ensemble of size T assigns class label $y \in \mathcal{Y}$ to instance $\mathbf{x}$ is the sum of (4) over all the ensemble predictions in which that class receives more votes. In particular, for class y_1

$$\mathcal{P}(\hat{y}^T = y_1|T, \boldsymbol{\pi}(\mathbf{x})) = \sum_{\boldsymbol{T};T_1>T_2} \mathcal{P}(\boldsymbol{T}|\boldsymbol{\pi}(\mathbf{x})) = I_{\pi_1(\mathbf{x})}\left(\lfloor\frac{T}{2}\rfloor + 1, T - \lfloor\frac{T}{2}\rfloor\right)\,, \tag{5}$$

where $I_x(a,b)$ is the regularized incomplete beta function [1]. The classification boundary is defined by the set of examples $\mathbf{x}$ for which $\pi_1(\mathbf{x}) = \pi_2(\mathbf{x}) = 1/2$. For these examples the predictions given by the ensemble are equivalent to random guessing because $\mathcal{P}(\hat{y}^T = y_1|T, 1/2) = 1/2$, $\forall\, T > 0$. Therefore, the instances that are close to this decision boundary are more likely to be misclassified. Fig. 3 displays the prediction error rate for the test instances as a function of the

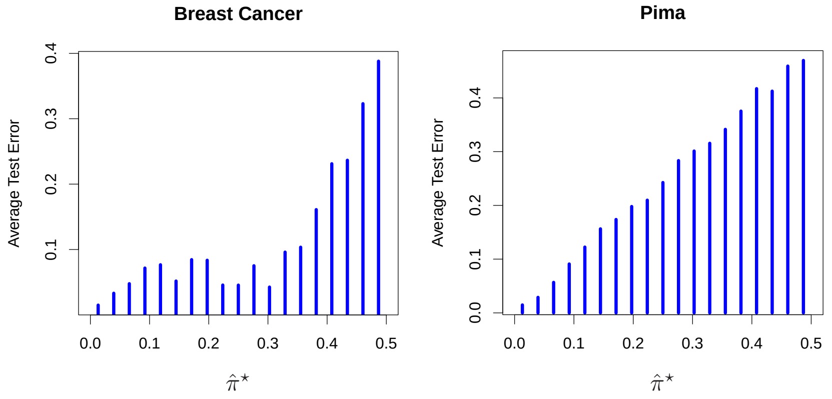

Fig. 3: Error rate as a function of $\hat{\pi}^\star = \min(\hat{\pi}_1, \hat{\pi}_2)$.

empirical estimates of the probability of the majority class $\hat{\pi}^\star = \min(\hat{\pi}_1, \hat{\pi}_2) \in [0, 1/2]$ ($\hat{\pi}_i = T_i/T, i = 1, 2$) for the *Breast Cancer* and *Pima* datasets, using a random forest ensemble. The results for other datasets and for other randomized parallel ensembles (e.g. random forests) are similar. The value $1/2 - \hat{\pi}^\star$ measures the distance to the classification boundary in the space of class votes. From these plots it is apparent that the test error tends to be larger for instances with higher values of $\hat{\pi}^\star$, which reflects the fact that the difficulty of classification increases with the proximity of the instances to the decision border. In fact, the error rates approaches 50% for $\hat{\pi}^\star \approx 1/2$. To identify the instances that are close to the decision boundary, which, for this reason, are difficult to classify, we design a binomial test based on the vector of predictions $\boldsymbol{T}$ for the test instance $\mathbf{x}$. The null hypothesis for this test is $\pi_1(\mathbf{x}) = \pi_2(\mathbf{x}) = 1/2$. The corresponding p-value is the probability of observing a vector of predictions more unlikely than the one actually observed ($\boldsymbol{T} = (T_1 \;\; T_2)$) assuming that the null hypothesis holds

$$\text{p-value} = 2 I_{1/2} \left(T - \min(T_1, T_2), 1 + \min(T_1, T_2) \right). \qquad (6)$$

When (6) is above 5%, $\mathbf{x}$ is identified as an example that is difficult to classify, in the sense that the classification by the ensemble will be close to random guessing (i.e. $\approx$ 50% chance of error). For $T = 1000$, this occurs when $\min(T_1, T_2) \geq 469$.

To assess the effectiveness of the test, we report the results of experiments in four binary classification problems from the UCI repository [2] (*Breast Cancer*, *Ionosphere*, *Sonar* and *Pima*), using random forests [4]. The experimental protocol described in the previous section is used to generate random train/test partitions of equal size and to build the random forest ensembles ($T = 1000$). Then, the instances whose p-value is above 5% are identified as being close to the decision borders and, in consequence, difficult to classify. The fraction of such instances are displayed in the second column of Table 1. Finally, the error rate on these instances (3rd column) is compared with the error rates on the remaining instances (4th column) and the global error rates (5th column). Analyzing the results in this table one sees that the prediction error on the set

Table 1: Properties of the test instances identified by the statistical test as being *potentially difficult to classify*. The results displayed are averages over different train/test partitions. The corresponding standard deviations are displayed after the ± symbols.

Dataset	**% difficult**	**Error difficult**	**Error rest**	**Error total**
Breast Cancer	0.6±0.4	46.6±37.4	2.7±0.7	3.0±0.7
Ionosphere	1.5±1.0	43.4±36.1	6.2±1.5	6.8±1.6
Pima	25.9±1.2	49.8±9.9	22.5±1.7	24.1±1.6
Sonar	9.5±3.0	47.5±16.6	16.9±4.7	19.9±4.5

of difficult instances is significantly larger than the error in the set of the remaining instances and in the whole test set. Furthermore, the error rates of the identified difficult instances are close to 50% (i.e. random guessing). The large values of the standard deviation for these rates are consistent with the expected distribution of the errors. These results illustrate the validity of the binomial test to identify difficult instances for classification.

In summary, we have illustrated the fact that in randomized parallel ensembles the predictions of different ensemble classifiers on a given instance are independent random variables. Taking advantage of this independence we have designed a statistical test to identify instances that are potentially difficult to classify. Experiments in several classification problems illustrate the validity of the analysis. The usefulness of this test in the design of robust boosting algorithms is currently under investigation.

References

[1] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover, New York, 1964.

[2] A. Asuncion and D.J. Newman. UCI machine learning repository, 2007. Available at http://www.ics.uci.edu/%7emlearn/MLRepository.html.

[3] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[4] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 36(1):3–42, 2006.

[6] L.K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.

[7] Daniel Hernández-Lobato, Gonzalo Martínez-Muñoz, and Alberto Suárez. Statistical instance-based pruning in ensembles of independent classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):364–369, 2009.

[8] Daniel Hernández-Lobato, Gonzalo Martńez-Muñoz, and Alberto Suárez. Inference on the prediction of ensembles of infinite size. *Pattern Recognition*, 44:1426–1434, 2011.

[9] Gonzalo Martínez-Muñoz and Alberto Suárez. Switching class labels to generate classification ensembles. *Pattern Recognition*, 38(10):1483–1494, 2005.

[10] Gonzalo Martínez-Muñoz and Alberto Suárez. Out-of-bag estimation of the optimal sample size in bagging. *Pattern Recognition*, 43(1):143 – 152, 2010.

[11] Juan J. Rodríguez, Ludmila I. Kuncheva, and Carlos J. Alonso. Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1619–1630, 2006.