Knowledge-Based Systems xxx (xxxx) xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

New margin-based subsampling iterative technique in modified random forests for classification☆

Wei Feng [a,b,*], Gabriel Dauphin [c], Wenjiang Huang [b], Yinghui Quan [d], Wenzhi Liao [e]

[a] School of Electronic Engineering, Xidian University, Shaanxi 710071, China
[b] Key laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China
[c] Laboratory of Information Processing and Transmission, L2TI, Institut Galilée, University Paris XIII, France
[d] Key Laboratory for Radar Signal Processing, Xidian University, Shaanxi 710071, China
[e] Department of Telecommunications and Information Processing, IMEC-TELIN-Ghent University, Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium

ARTICLE INFO

Article history:
Received 12 September 2018
Received in revised form 14 May 2019
Accepted 12 July 2019
Available online xxxx

Keywords:
Classification
Ensemble margin
Diversity
Random forests
Sub-sampling

ABSTRACT

Diversity within base classifiers has been recognized as an important characteristic of an ensemble classifier. Data and feature sampling are two popular methods of increasing such diversity. This is exemplified by Random Forests (RFs), known as a very effective classifier. However real-world data remain challenging due to several issues, such as multi-class imbalance, data redundancy, and class noise. Ensemble margin theory is a proven effective way to improve the performance of classification models. It can be used to detect the most important instances and thus help ensemble classifiers to avoid the negative effects of the class noise and class imbalance. To obtain accurate classification results, this paper proposes the Ensemble-Margin Based Random Forests (EMRFs) method, which combines RFs and a new subsampling iterative technique making use of computed ensemble margin values. As for comparative analysis, the learning techniques considered are: SVM, AdaBoost, RFs and the Subsample based Random Forests (SubRFs). The SubRFs uses Out-Of-Bag (OOB) estimation to optimize the training size. The effectiveness of EMRFs is demonstrated on both balanced and imbalanced datasets.

1. Introduction

Machine learning is widely used in solving real-world problems, including web page ranking [1], collaborative filtering [2], entity recognition [3], speech recognition [4] and remote sensing [5]. Data classification is a major research area in machine learning [6,7]. Its starting point is a set of observations described by features and associated with a class, thanks to some specific knowledge. Then, a classification model is developed, generally assuming that all these class-labeled features are drawn from a set of probability distributions, one for each class. Finally, this classification model displays the class membership of new instances [8].

Among the numerous learning approaches, we are concerned here with ensemble learning, which provides effective methods to develop accurate classification systems [9,10]. Accurate binary classifiers are obtained by aggregating weak classifiers which are only slightly better than a random guess [11]. And aggregation has also the capacity to increase classifiers' generalization ability [11]. Ensemble learning approaches could be addressed at the data level, the feature level or in how classifiers are combined.

We focus on Random Forests (RFs) [12] as their outstanding performance makes it receive more and more attention [13–15]. RFs are a set of decision trees, that are combined through a majority vote. Each decision tree is trained on a reduced number of samples having a reduced number of features, the latter could be the square root or the logarithm of the number of available features [16]. Both samples and features are drawn randomly from the training set. A practical asset is that it runs efficiently on large database handling thousand of input variables with reduced training time.

Ensemble diversity is a property of an ensemble of classifiers with respect to a set of data. It has been recognized as an important characteristic [10,17–19]. Diversity accounts for the amount of statistical independence of classifiers [20]. It accounts to how much incorrect decisions of some classifiers increase the variance and not the bias error of the ensemble classifier. In other words, ensemble learning is very effective, mainly because base

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.knosys.2019.07.016.

* Corresponding author at: Key laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China.
E-mail address: fw.enp@qq.com (W. Feng).

https://doi.org/10.1016/j.knosys.2019.07.016
0950-7051/

classifiers have different *biases* [8,11]. And we are addressing this challenging issue, yielding more diverse decision trees, thanks to new training procedures.

In [13], the process leading to new features specific to each randomly selected subset of samples is three-fold. First features are randomly partitioned into subsets generating a partitioned feature space. Each feature subspace is then reduced by applying Principal Component Analysis (PCA) on the randomly selected samples. Finally, all reduced feature subspaces are merged into a single reduced feature space. The collection of these selected samples and their modified features are used to train modified RFs. In [21], a Convolutional Neural Network (CNN) produces a nonlinear mapping of features with which RFs are trained. Both resulting classifiers have increased diversity and performance.

However, when challenged with some issues, such as class noise [22] or multi-class imbalance [23], it seems more appropriate to increase the diversity of an ensemble model at the data level and actually to use data sampling techniques [9]. The two following techniques presented in [24], exemplify how data sampling technique addresses the class imbalance. In a binary context of class imbalance, the Balanced Random Forests algorithm (BRF) trains each decision tree using an even number of samples drawn from each of the two class. The Weighted Random Forests algorithm (WRF) assigns to each class a weight modeling a misclassification cost being higher for the minority class. WRF has similar performances to BRF but reduced computational efficiency. Bernard et al. proposed a Dynamic Random Forest (DRF) [25]. DRF is an adaptive learning procedure where each new tree is expected to best complement the already learned trees. The adaptivity stems both from a weight-based resampled training data and from a randomly selected number of features taking into account a measure of feature information. However, in contrast to RFs, the trees in this method are no longer independent and the resulting classifier may lack class noise robustness.

Subsample based methods have proved to have higher ensemble diversity, and they have been used in many fields [26–29] with various subsampling ratios (i.e. the number of bootstrap samples to the number of available samples) and using two different sampling techniques: with or without replacement. On average, [30] notes that the behavior of an ensemble classifier for which samples are drawn with replacement is the same as when there is no replacement on the condition that the subsampling ratio is being modified according to a specific nonlinear mapping. The common choice is 50% without replacement, which is approximately equivalent to 100 % with replacement (as in bagging). Decreasing this subsampling ratio can reduce the bias and variance as shown by [31]. Addressing a very large database, [32] has achieved good performance with a subsampling ratio as small as 0.5% which indeed saves training time. Studying the impact of the subsampling ratio without replacement, [33] shows that depending on the dataset, increased generalization performance is sometimes obtained with a ratio of 20% or of up to 60% or even 80%. To select that ratio, they advise using the Out-Of-Bag (OOB) estimation base method, whose statistical properties have been studied in [34]. Unfortunately, when addressing imbalanced datasets, a bias towards the majority class may still occur.

Subsampling techniques can be improved using ensemble margin theory. The concept of margin was first proposed by Vapnik, who applied it to build Support Vector Machines (SVM) [35]. Ensemble margin consists in assigning to each sample a value named margin which models its importance. This has been used in class noise filtering [22], instance selection [36], [37], feature selection [38] and classifier design [39], [40], [41]. Taking into account such margin information helps to address several issues such as redundancy, class noise, and class imbalance. Ensemble margin theory can be used to detect the most important instances and thus help ensemble classifiers to avoid the negative effects of redundant or noisy samples.

The contribution of this paper is to propose a novel method named Ensemble-Margin based Random Forests method (EMRFs). Two improvements are done upon RFs: samples are selected according to their margin values, and ensemble diversity has been increased. In our previous work [19], the classification of imbalanced multi-class datasets was addressed by combining bagging with the computation of a margin. Although both methods are based on the margin value, there exist some big differences (1) improvements are applied to RFs instead of bagging as RFs is known to be more efficient (2) the technique addressing the imbalance issue is the Synthetic Majority Oversampling Technique (SMOTE) [42,43] instead of an undersampling technique (3) balanced datasets are here also addressed, whereas in [19] only imbalanced datasets were addressed. As for comparative analysis, the learning techniques used are: SVM [44], AdaBoost [45], RFs and the Subsample based Random Forests (SubRFs). The latter uses Out-Of-Bag (OOB) estimation to optimize the training size. In an experimental study, it is shown that the proposed algorithm is a clear enhancement of RFs and SubRFs, especially when applied to imbalanced data sets.

The rest of the paper is organized as follows. In Section 2, the ensemble margin theory is first presented, then an evaluation method of the instance importance based on ensemble margin is proposed. The EMRFs are proposed in Section 3. The comparative analysis showing the effectiveness of the new algorithm and a study on hyper-parameters are presented in Section 4 and discussed in Section 5. Finally, the conclusions are summarized in Section 6.

2. Evaluation of the data significance based on ensemble margin

2.1. Ensemble margin

Measuring the margin helps designing classifiers that are more robust to input perturbations and have better generalization properties. The margins can be defined into two main ways: sample margin and hypothesis margin [46]. The sample margin is defined as the distance between the feature vector and the decision boundary induced by a classifier. For example, SVM [35] aims to find the separating hyperplane with the sample margin. On the other hand, the hypothesis margin requires the existence of a distance measure on the hypothesis class, it measures how much the hypothesis can travel before it hits a feature vector without changing the way it labels any of the sample points. This definition requires a distance measure between classifiers [38], [46]. AdaBoost is an example of ensemble classifier using such a hypothesis margin [45].

When a large number of classifiers are available, margin can also be defined using the classifiers' predictions and such a metric is called ensemble margin. The ensemble margin considered in this paper is introduced by Shapire et al. [47] and redescribed here as Eq. (1). Let us define some notations before,

- $(\mathbf{x}, y)$ is an instance, with $\mathbf{x}$ as a vector with feature values and y as a label being one of the C class labels,
- $h_j(\mathbf{x})$ are the J labels predicted by the J available classifiers,
- $\pi \mapsto \delta(\pi)$ is a function mapping any predicate π into 1 or 0 depending on whether π is true or not,
- $v(\mathbf{x}, c) = \sum_{j=1}^{J} \delta(h_j(\mathbf{x}) = c)$ is the number of classifiers predicting the label c when the feature vector is $\mathbf{x}$,

This ensemble margin is defined as

$$\text{margin}(\mathbf{x}, y) = \frac{1}{J}\left(v(\mathbf{x}, y) - \max_{c \neq y} v(\mathbf{x}, c)\right) \tag{1}$$

Please cite this article as: W. Feng, G. Dauphin, W. Huang et al., New margin-based subsampling iterative technique in modified random forests for classification, Knowledge-Based Systems (2019), https://doi.org/10.1016/j.knosys.2019.07.016.

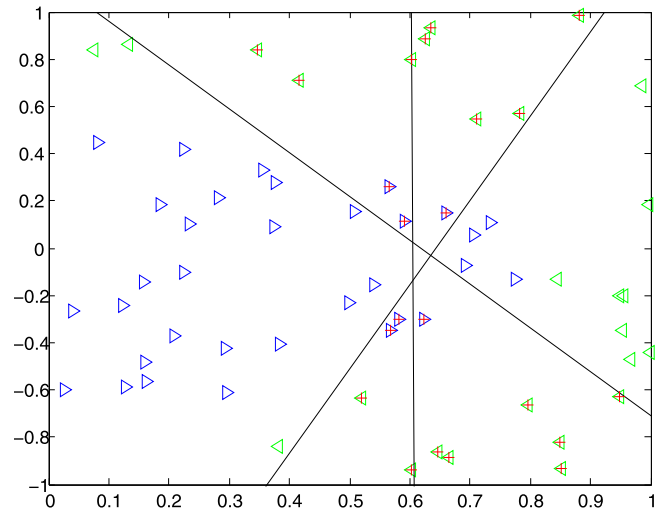


Fig. 1. Samples (◁ in green and ▷ in blue) are classified by three classifiers indicated by their decision boundaries (black straight lines). Samples on the left and on the right have a significance $W_i = 0$ whereas samples in between decision boundaries are indicated with a red plus + as they have a significance $W_i = 1$.

Note that a positive value of margin$(\mathbf{x}, y)$ indicates that this instance is correctly classified by the set of J classifiers when using a majority vote. A negative value indicates misclassification. When all base classifiers are unanimous, margin$(\mathbf{x}, y) = 1$ or margin$(\mathbf{x}, y) = -1$, margin$(\mathbf{x}, y)$ ranges from -1 to 1.

2.2. Assessing the significance of a sample

Ensemble margin is used here to assess the significance of a sample. The rationale is that for a given sample, when different classifiers do not agree on a label, the ensemble margin tends to be lower in magnitude and the sample is likely to be more useful in terms of classification. The significance of a sample $(\mathbf{x}, y)$ is defined as:

$$W(\mathbf{x}, y) = 1 - |\text{margin}(\mathbf{x}, y)| \tag{2}$$

The significance can be rewritten as:

$$W(\mathbf{x}, y) = 1 - \frac{1}{J}\left[v(\mathbf{x}, \tilde{h}_1(\mathbf{x})) - \min\left(v(\mathbf{x}, y), v(\mathbf{x}, \tilde{h}_2(\mathbf{x}))\right)\right] \tag{3}$$

where $\tilde{h}_1(\mathbf{x})$ and $\tilde{h}_2(\mathbf{x})$ are the top two labels in terms of number of classifiers:

$$\tilde{h}_1(\mathbf{x}) = \arg\max_c v(\mathbf{x}, c) \quad \text{and} \quad \tilde{h}_2(\mathbf{x}) = \arg\max_{c \neq \tilde{h}_1(\mathbf{x})} v(\mathbf{x}, c)$$

Assuming the number of classifiers predicting each label are listed and sorted in a decreasing order, $W(\mathbf{x}, y)$ can be thought as the normalized difference of the two top numbers of classifiers or as the normalized difference between the top number of classifiers and the number of classifiers selecting the true label, depending on whichever is the greatest value.

When there are only two classes, the significance can be rewritten as:

$$W(\mathbf{x}, y) = 1 - \frac{1}{J}\left[v(\mathbf{x}, \tilde{h}_1(\mathbf{x})) - v(\mathbf{x}, \tilde{h}_2(\mathbf{x}))\right] \tag{4}$$

Then $W(\mathbf{x}, y)$ can be thought of as the normalized difference of the two top numbers of classifiers.

Proof of Eq. (3) is in two steps.

- Let us first assume that the true label is the majority selected label, $\tilde{h}_1(\mathbf{x}) = y$, then

$$W(\mathbf{x}, y) = 1 - \left|v(\mathbf{x}, \tilde{h}_1(\mathbf{x})) - v(\mathbf{x}, \tilde{h}_2(\mathbf{x}))\right|$$

by definition of W in Eq. (2) and because $y = \tilde{h}_1(\mathbf{x})$

$$W(\mathbf{x}, y) = 1 - \left[v(\mathbf{x}, \tilde{h}_1(\mathbf{x})) - v(\mathbf{x}, \tilde{h}_2(\mathbf{x}))\right]$$

because $v(\mathbf{x}, \tilde{h}_2(\mathbf{x})) \leq v(\mathbf{x}, \tilde{h}_1(\mathbf{x}))$

which implies that the absolute value is not sign changing

$$W(\mathbf{x}, y) = 1 - \frac{1}{J}\left[v(\mathbf{x}, \tilde{h}_1(\mathbf{x})) - \min\left(v(\mathbf{x}, y), v(\mathbf{x}, \tilde{h}_2(\mathbf{x}))\right)\right]$$

because $v(\mathbf{x}, \tilde{h}_2(\mathbf{x})) \leq v(\mathbf{x}, \tilde{h}_1(\mathbf{x})) = v(\mathbf{x}, y)$

- If this assumption is not true, $\tilde{h}_1(\mathbf{x}) \neq y$, then

$$W(\mathbf{x}, y) = 1 - \left|v(\mathbf{x}, y) - v(\mathbf{x}, \tilde{h}_1(\mathbf{x}))\right|$$

by definition of W in Eq. (2) and because $\max_{c \neq y} v(\mathbf{x}, c)$

$= v(\mathbf{x}, \tilde{h}_1(\mathbf{x}))$

as $\tilde{h}_1(\mathbf{x}) \neq y$ and $v(\mathbf{x}, \tilde{h}_1(\mathbf{x})) \geq \max_c v(\mathbf{x}, c)$

$$W(\mathbf{x}, y) = 1 - \left[v(\mathbf{x}, \tilde{h}_1(\mathbf{x})) - v(\mathbf{x}, y)\right]$$

because $v(\mathbf{x}, \tilde{h}_1(\mathbf{x})) \geq v(\mathbf{x}, y)$

which implies that the absolute value is here sign changing

$$W(\mathbf{x}, y) = 1 - \frac{1}{J}\left[v(\mathbf{x}, \tilde{h}_1(\mathbf{x})) - \min\left(v(\mathbf{x}, y), v(\mathbf{x}, \tilde{h}_2(\mathbf{x}))\right)\right]$$

because $v(\mathbf{x}, y) \leq v(\mathbf{x}, \tilde{h}_2(\mathbf{x}))$

as the top value $v(\mathbf{x}, \tilde{h}_1(\mathbf{x}))$ is not available since $y \neq \tilde{h}_1(\mathbf{x})$

Proof of Eq. (4) is also in the same two steps.

- Let us first assume that $\tilde{h}_1(\mathbf{x}) = y$,
then $v(\mathbf{x}, \tilde{h}_2(\mathbf{x})) \leq v(\mathbf{x}, \tilde{h}_1(\mathbf{x})) = v(\mathbf{x}, y)$ and

$$\min\left(v(\mathbf{x}, \tilde{h}_2(\mathbf{x})), v(\mathbf{x}, y)\right) = v(\mathbf{x}, \tilde{h}_2(\mathbf{x}))$$

- If this assumption is not true, $\tilde{h}_1(\mathbf{x}) \neq y$,
then since there are only two classes, $\tilde{h}_2(\mathbf{x}) = y$ and

$$\min\left(v(\mathbf{x}, \tilde{h}_2(\mathbf{x})), v(\mathbf{x}, y)\right) = v(\mathbf{x}, \tilde{h}_2(\mathbf{x}))$$

2.3. Significance based data ordering

Let us consider a training set denoted as $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, where $\mathbf{x}_i$ is a vector with feature values and y_i is the value of the class label. The significance of a training sample $(\mathbf{x}_i, y_i)$ is assessed by:

$$W_i = W(\mathbf{x}_i, y_i) = 1 - \frac{1}{J}\left|v(\mathbf{x}_i, y_i) - \max_{c \neq y_i} v(\mathbf{x}_i, c)\right| \tag{5}$$

An example (Fig. 1) is used to show in a two-class setting, the relationship between the predicted labels of different samples by the three different classifiers and their significance values W_i. Samples are shown as green left triangles or blue right triangles depending on their labels. Three different decision boundaries are shown as black straight lines, they are related to the three different classifiers. Significant samples are indicated with a red plus. Indeed as indicated by Eq. (4), in this two-class setting with only three classifiers, W_i is a two-value metric: $W_i = 0$ when classifiers are unanimous and $W_i = 1$ when they are not.

Please cite this article as: W. Feng, G. Dauphin, W. Huang et al., New margin-based subsampling iterative technique in modified random forests for classification, Knowledge-Based Systems (2019), https://doi.org/10.1016/j.knosys.2019.07.016.

The significance metric is designed for applications where a large number of base classifiers are available, as this metric would then yield many intermediate values ranging from 0 to 1. This metric measures disagreement among classifiers and hence is able to indicate to what extent a sample is near a class decision boundary. Because samples near such boundaries are known to be important in classification, the main idea of the proposed method is to train new base classifiers that focus on significant samples, as measured by W_i. Technically the algorithm consists in reordering the training samples according to the decreasing order of W_i. The different base classifiers are trained on bootstrap sets using subsets of the reordered training dataset so as to preserve diversity.

3. Ensemble-margin based random forests (EMRFs)

The Ensemble-Margin based Random Forests (EMRFs) method is proposed to find a more accurate and diverse classifier. It has three main steps:

1. Computing the significance of training samples using the available base classifiers.
2. Constructing bootstrap subsets of the training set while focusing on significant samples and preserving diversity.
3. Training base classifiers on those subsets with random feature selection as in the traditional RFs.

The first step of EMRFs consists of training a robust ensemble classifier, namely the *random forests*, using the whole training set. The significance W_i of each training instance is then calculated using Eq. (5). And the training set is ordered according to the decreasing order of W_i, the resulting new training dataset is denoted as $S' = \{(\mathbf{x}'_1, y'_1), \ldots, (\mathbf{x}'_N, y'_N)\}$. Note that $i \mapsto W(\mathbf{x}'_i, y'_i)$ is a decreasing sequence.

In the second step, the algorithm aims to explore the most significant samples while preserving diversity. Let t be an index ranging from 1 to T. The evolving resampling rate a_t considered here is periodic and has over the first period, an arithmetic progression.

$$\Delta a = \frac{1}{T_a} \quad \text{and} \quad a_t = \frac{1 + \mathrm{mod}(t-1, T_a)}{T_a}$$

where $\mathrm{mod}(a, b)$ stands for the modulo operator, that is the remainder after division of one number by another; Δa is an EMRFs hyper-parameter ranging from 0 to 1 and constrained to be an integer inverse; T_a is the period, determined by Δa. Note that a_t is ranging from Δa to 1 and that Δa is the initial resampling rate: $a_1 = \Delta a$.

With a_t and as t ranges from 1 to T, we define T training subsets:

$$\tilde{S}'_t = \{(\mathbf{x}'_i, y'_i) | i \leq Na_t\}$$

T new bootstrap training sets containing $\lfloor Na_t \rfloor$ samples are drawn with replacement from these subsets:

$$S_t \sim \mathcal{B}_{\lfloor Na_t \rfloor}\left(\tilde{S}'_t\right)$$

Here are some examples to illustrate the definition of a_t and of the yielded subsets.

- If $T = 12$, $a = 0.1 \sim 1$ and $N = 100$, then the T resampling rates are: $a_1 = 0.1$, $a_2 = 0.2$, …, $a_{10} = 1$, $a_{11} = 0.1$, $a_{12} = 0.2$. The T subsets $\tilde{S}_t$ are the first 10, 20, …,100,10, 20 first samples of the reordered training set (Fig. 2(a)).
- If $T = 11$, $a = 0.25 \sim 1$ and $N = 100$, then the T resampling rates are: $a_1 = 0.25$, $a_2 = 0.5$, $a_3 = 0.75$, …, $a_{10} = 0.5$, $a_{11} = 0.75$. The T subsets $\tilde{S}_t$ are the first 25, 50, 75,100, …,50, 75 first samples of the reordered training set (Fig. 2(b)).
- If $a = 1$ and $T = 10$, and $N = 100$, then for all t, $\tilde{S}_t = S$ and S_t are bootstrap training sets built as in the classical RFs.

Fig. 3 presents the flowchart of the EMRFs. Finally, the results of a series of individual classifiers, generated by repeating the aforementioned steps several times, are fused according to majority vote rule.

Algorithm 1 Ensemble-Margin based Random Forests

Training phase
Input:
Training set $S = (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2,), \cdots, (\mathbf{x}_N, y_N)$;
Number of training instances N;
Number of classifiers T;
Ensemble creation algorithm ζ;
Initial resampling rate Δa;
Process:
The new ensemble classifier is set as containing no classifiers: $E \leftarrow \varnothing$
Construct an ensemble classifier with all training data $(\mathbf{x}_i, y_i) \in S$
Compute the significance of each sample $(\mathbf{x}_i, y_i)$ as measured by W_i.
Order the training samples in descending order of W_i, the ordered set is denoted as S'.
for t=1:T **do**
    Obtain a new training set S_t by performing a bootstrap on the first Na_t samples of S'.
    Train a decision tree $h_t = \zeta(S_t)$ using S_t.
    Add this new classifier: $E \leftarrow E \cup \zeta_t$
end for
Complete E with the majority vote as the pooling rule.
Output:
The ensemble classifier E.

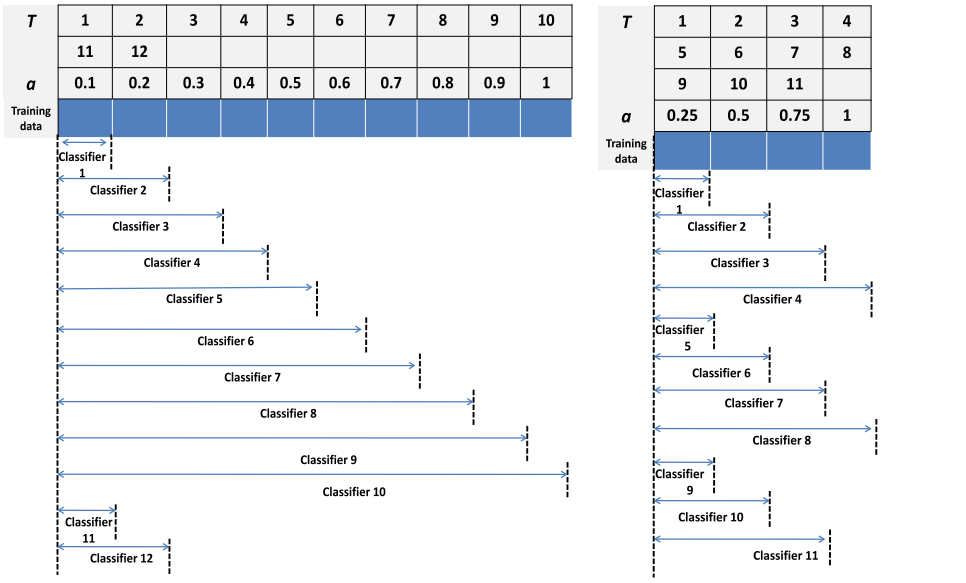
Prediction phase
Inputs:
The ensemble $E = \{h_t\}_{t=1}^{T}$;
A new sample $\mathbf{x}^*$.
Output:
Class label $y^* = \mathrm{argmax}_{\{1 \leq c \leq C\}} \sum_{t=1}^{T} \delta(\, h_t(\mathbf{x}^*) = c\,)$

4. Experiment

4.1. Experiment settings

In these experiments, RFs method is used to create an ensemble classifier involving Classification and Regression Trees (CART) [12] as base classifiers to get the margin values of the instances of the original data set. Evaluation of the performance of the proposed EMRFs is carried out using 5-fold cross-validation (CV). EMRFs are compared to SVM, AdaBoost, RFs and SubRFs which is a refined version of RFs, denoted as SubRFs [33,48]. Ensembles size T is set to 500, a number chosen to be higher than what is usually needed to deal with publicly available classification problems. The influence of the ensemble size for all the methods is further discussed in Section 4.6.1. The range of sampling parameter a is set to $0.1 \sim 1$ (i.e. $\Delta a = 0.1$). The influence of Δa for EMRFs is presented in Section 4.6.2. R-project package, "randomForest" is used to implement the proposed methodology and two reference schemes in the experiment. For all the RFs based methods, the number of features randomly sampled as candidates at each split is set to the square root of the number of features. Other parameters are kept to

Please cite this article as: W. Feng, G. Dauphin, W. Huang et al., New margin-based subsampling iterative technique in modified random forests for classification, Knowledge-Based Systems (2019), https://doi.org/10.1016/j.knosys.2019.07.016.



(a) Ensemble size $T = 12$, $a = 0.1 \sim 1$ (b) Ensemble size $T = 11$, $a = 0.25 \sim 1$

Fig. 2. Two examples to illustrate the process of the base classifier building along with the change of the parameter a. The training samples have been sorted in ascending order according to their margin values.

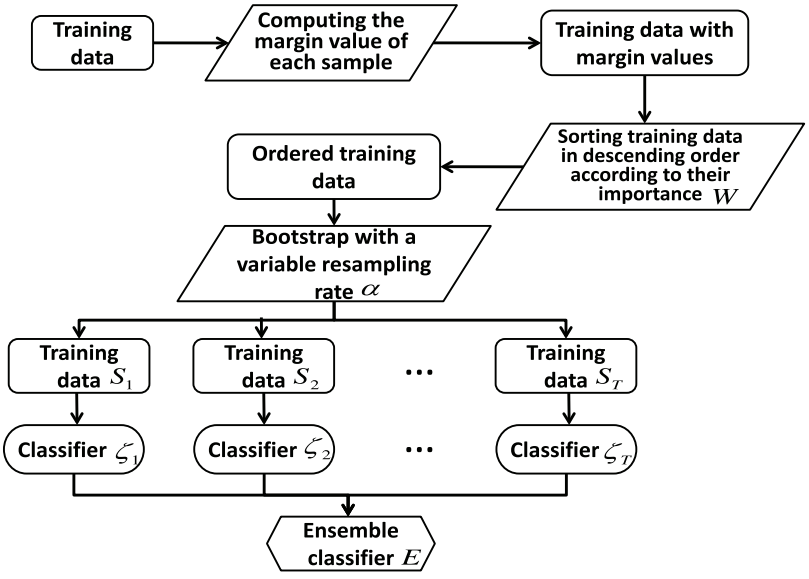


Fig. 3. Flowchart of ensemble-margin based random forests classification.

their default values in R-project package (https://cran.r-project.org/web/packages/randomForest/index.html). The SVM and AdaBoost methods are implemented using the "e1071" and "adabag" packages (https://cran.r-project.org/web/packages/e1071/index.html and https://cran.r-project.org/web/packages/adabag/). The kernel function used for SVM is the radial basis function. Both the gamma and the cost value used in the kernel are jointly selected so as optimize the average performance obtained with CV. The optimized values are searched respectively within $[2^{-8} : 2^{8}]$ and $[2^{1} : 2^{5}]$.

4.2. Evaluation methods

The performance measures adopted in the experiments are: *Overall Accuracy*, *Minimum Accuracy Per Class* and *(Kohavi-Wolpert) diversity* defined in [49] and denoted as KW.

- **Minimum Accuracy Per Class**, is the percentage of instances correctly classified in the class for which this percentage is the least. Let n_{ii} and n_{ij} represent the true prediction of the ith class and the false prediction of the ith class into jth class respectively. The Minimum Per Class Accuracy for class i can be defined as:

$$\text{Minimum Per Class Accuracy} = \min_i \frac{n_{ii}}{\sum_{j=1}^{C} n_{ij}} \tag{6}$$

where C stands for the number of classes. Note that $\frac{n_{ii}}{\sum_{j=1}^{C} n_{ij}}$ is called the Per Class Accuracy or also Recall. And [50] strongly recommends using this performance measure Recall to evaluate classification algorithms, especially when dealing with multi-class imbalance problems.

Please cite this article as: W. Feng, G. Dauphin, W. Huang et al., New margin-based subsampling iterative technique in modified random forests for classification, Knowledge-Based Systems (2019), https://doi.org/10.1016/j.knosys.2019.07.016.

Table 1
Descriptions of 15 UCI datasets and 8 high-dimensional microarray datasets.

	Datasets	No.of instances	Variables	Classes	IR
Low-dimensional datasets	Balance*	600	4	3	5.65
	Breast	680	9	2	1.92
	Breast-wdbc	460	30	2	1.67
	Clean	400	166	2	1.4
	Monk3	500	6	2	1.11
	Optdigit	2000	64	10	1.24
	Pendigit	4000	16	10	1.21
	Segment	2000	19	7	1.05
	Sonar	200	60	2	1.2
	Soybean*	560	35	15	4.84
	Statlog	2000	36	6	2.54
	UrbanLandCover*	300	147	7	5.55
	Waveform40	4000	21	3	1.03
	Wilt*	4000	5	2	17.69
	Wine	160	13	3	1.39
High-dimensional datasets	Alon	62	2000	2	1.82
	Christensen*	217	1413	3	5.95
	Golub	72	7129	2	1.88
	Gravier	168	2905	2	1.95
	Khan*	63	2308	4	2.88
	Pomeroy	60	7128	2	1.86
	Shipp*	77	7129	2	3.05
	Su	102	5565	4	1.22

- **Overall Accuracy** is a performance metric giving the same weight to each class overlooking the number of samples assigned to each class:

$$\text{Overall Accuracy} = \frac{\sum_{i=1}^{C} \text{Per Class Accuracy}_i}{C} \tag{7}$$

- **KW Diversity** presented in [49] and defined in [51] by the following equation:

$$KW = -\frac{1}{NT^2}\sum_{j=1}^{N} t(x_j)(T - t(x_j)) \tag{8}$$

where N is the number of samples in the training set, T is the number of classifiers, and $t(x_j)$ is the number of classifiers having predicted the correct label of x_j. A higher value of KW indicates a higher diversity. Note that this measure gives also the same weight to each class.

4.3. Datasets

The proposed algorithm is applied on 15 UCI datasets published in [52] and 8 high-dimensional microarray datasets (https://github.com/ramhiser/datamicroarray/tree/master/data). The chosen datasets include 12 multi-class and 11 binary data (Table 1). Those data sets deal with different machine learning issues in terms of sizes and features. Table 1 summarizes the properties of the selected datasets, including the number of classes, the number of attributes, the number of examples and the Imbalance Ratio, denoted as IR. IR is the ratio of the size of the most populated class to the size of the least populated class. The asterisk (*) indicates high IR in this dataset. 5-fold CV is done on each dataset. It is done during oversampling as [43] recommends especially for imbalanced datasets: for each fold, the new instances are generated using only instances from the corresponding training set. Note that this CV-technique is applied to multi-class datasets extending the binary context considered in [43]. Furthermore, to prevent the random oversampling method from overfitting, the SMOTE technique presented in [42] is adopted in the experiment on the more imbalanced datasets indicated with an asterisk, as a pre-processing task common to all learning techniques.

Table 2
Overall Accuracy of the SVM, AdaBoost, standard RFs, SubRFs and the proposed method EMRFs.

Datasets	SVM	AdaBoost	RFs	SubRFs	EMRFs
Balance	80.33(2.0)	78.57(5.0)	80.07(3.0)	78.70(4.0)	**81.33**(1.0)
Breast	96.44(3.0)	96.15(4.0)	96.85(2.0)	95.85(5.0)	**97.79**(1.0)
Breast-wdbc	96.87(2.0)	95.91(3.0)	95.22(4.0)	94.74(5.0)	**97.39**(1.0)
Clean	**90.05**(1.0)	84.15(2.0)	82.35(4.0)	76.85(5.0)	82.75(3.0)
Monk3	95.80(5.0)	97.24(2.0)	97.00(3.0)	96.92(4.0)	**98.60** (1.0)
Optdigit	96.67(4.0)	96.95(2.0)	96.72(3.0)	96.03(5.0)	**97.15**(1.0)
Pendigit	**99.30**(1.0)	98.19(3.0)	98.02(4.0)	97.47(5.0)	98.55(2.0)
Segment	90.47(5.0)	**97.69**(1.0)	96.80(3.0)	96.07(4.0)	97.25(2.0)
Sonar	83.00(2.0)	81.70(3.0)	79.40(4.0)	78.10(5.0)	**84.00**(1.0)
Soybean	92.50(2.0)	91.00(5.0)	92.39(3.0)	91.71(4.0)	**93.75**(1.0)
Statlog	90.50(2.0)	90.19(4.0)	90.26(3.0)	89.55(5.0)	**90.58**(1.0)
UrbanLandCover	75.93(5.0)	79.87(4.0)	83.33(3.0)	85.33(2.0)	**86.00**(1.0)
Waveform40	84.71(2.5)	83.95(5.0)	84.65(4.0)	84.71(2.5)	**85.75**(1.0)
Wilt	97.08(5.0)	98.13(2.0)	97.90(3.0)	97.57(4.0)	**98.23**(1.0)
Wine	97.88(2.0)	95.00(4.0)	96.12(3.0)	93.59(5.0)	**98.75**(1.0)
Average	91.17	90.98	91.14	90.21	**92.52**
Rank	2.90	3.27	3.27	4.30	1.27

Table 3
Minimum Accuracy Per Class of the SVM, AdaBoost, standard RFs, SubRFs and the proposed method EMRFs.

Datasets	SVM	AdaBoost	RFs	SubRFs	EMRFs
Balance	28.57(3.0)	41.66(2.0)	25.96(4.0)	21.26(5.0)	**44.94**(1.0)
Breast	95.54(2.0)	95.21(3.0)	95.18(4.0)	93.63(5.0)	**97.55**(1.0)
Breast-wdbc	93.52(3.0)	94.90(2.0)	92.89(4.0)	92.17(5.0)	**96.84**(1.0)
Clean	**87.28**(1.0)	81.67(3.0)	75.40(4.0)	66.25(5.0)	81.99(2.0)
Monk3	94.51(5.0)	96.64(2.0)	96.01(3.0)	95.29(4.0)	**97.91**(1.0)
Optdigit	93.35(4.0)	94.39(2.0)	93.54(3.0)	93.03(5.0)	**95.11**(1.0)
Pendigit	**98.45**(1.0)	96.04(3.0)	94.47(4.0)	93.65(5.0)	97.39(2.0)
Segment	79.69(5.0)	94.77(2.0)	93.07(3.0)	91.55(4.0)	**94.81**(1.0)
Sonar	75.34(3.0)	77.15(2.0)	72.46(4.0)	71.27(5.0)	**82.14**(1.0)
Soybean	79.80(2.0)	76.85(3.0)	75.42(5.0)	76.08(4.0)	**85.58**(1.0)
Statlog	74.81(2.0)	67.99(5.0)	71.57(3.0)	71.56(4.0)	**77.37**(1.0)
UrbanLandCover	57.10(5.0)	66.35(4.0)	69.24(3.0)	71.88(2.0)	**72.81**(1.0)
Waveform40	81.29(2.0)	80.96(3.0)	76.88(5.0)	79.50(4.0)	**83.72**(1.0)
Wilt	70.19(5.0)	**89.48**(1.0)	84.73(2.0)	84.47(3.0)	80.97(4.0)
Wine	96.85(2.0)	92.31(4.0)	93.05(3.0)	88.98(5.0)	**97.22**(1.0)
Average	80.42	83.09	80.66	79.37	**85.76**
Rank	3.00	2.73	3.60	4.33	1.33

4.4. Results

Table 2 shows the Overall Accuracy of the SVM, AdaBoost, RFs, SubRFs, and the proposed EMRFs. This table shows that the new scheme outperforms the reference methods. The best increases of the novel method respectively are up to **11%**, **6%**, **4%** and **6%**. Moreover, the proposed method is effective for not only the datasets of high quality but also the datasets of a complicated space distribution such as the imbalanced data *Balance, Soybean, UrbanLandCover* and *Wilt*. Although the SMOTE is performed to balance the class sizes of these difficult datasets in the preprocessing step, the operation has a high risk of producing artificial noise. The EMRFs always yield the best results, i.e. the proposed

Please cite this article as: W. Feng, G. Dauphin, W. Huang et al., New margin-based subsampling iterative technique in modified random forests for classification, Knowledge-Based Systems (2019), https://doi.org/10.1016/j.knosys.2019.07.016.

Table 4
Ensemble diversity of the AdaBoost, standard RFs, SubRFs, and the proposed method EMRFs.

Datasets	Diversity			
	AdaBoost	RFs	SubRFs	EMRFs
Balance	0.1586	0.1143	0.1115	**0.1691**
Breast	0.0618	0.0371	0.0514	**0.0639**
Breast-wdbc	0.0918	0.0535	0.0654	**0.1096**
Clean	0.2176	0.1836	0.1897	**0.2288**
Monk3	0.1393	0.0732	0.1113	**0.1520**
Optdigit	0.0232	0.0176	0.0202	**0.0283**
Pendigit	**0.0405**	0.0142	0.0164	0.0328
Segment	**0.0653**	0.0317	0.0367	0.0525
Sonar	0.2219	0.1964	0.2019	**0.2386**
Soybean	**0.0339**	0.0157	0.0193	0.0233
Statlog	**0.0873**	0.0379	0.0394	0.0591
UrbanLandCover	0.0607	0.0498	0.0566	**0.0675**
Waveform40	**0.1573**	0.1181	0.1208	0.1492
Wilt	0.0742	0.0324	0.0360	**0.1484**
Wine	0.1363	0.1047	0.1398	**0.1775**
Average	0.1046	0.0720	0.0811	**0.1134**

Table 5
Overall Accuracy and Minimum Accuracy Per Class of the proposed method EMRFs with the sampling range $a = 0.1 \sim 1$ and optimized value respectively.

Datasets	Overall Accuracy		Minimum Accuracy Per Class	
	$a = 0.1 \sim 1$	Optimized a	$a = 0.1 \sim 1$	Optimized a
Balance	81.33	**85.50**	44.94	**47.94**
Breast	97.79	**97.94**	97.55	97.55
Breast-wdbc	97.39	**98.04**	96.84	96.84
Clean	82.75	**87.25**	81.99	**85.24**
Monk3	98.60	**98.90**	97.91	**97.96**
Optdigit	97.15	**97.40**	95.11	**95.38**
Pendigit	98.55	**98.62**	97.39	**97.45**
Segment	97.25	**97.60**	94.81	**95.61**
Sonar	84.00	**88.50**	82.14	**84.50**
Soybean	93.75	93.75	85.58	**85.87**
Statlog	90.58	**91.22**	77.37	**79.94**
UrbanLandCover	86.00	**86.25**	72.81	**77.08**
Waveform40	85.75	**85.95**	83.72	**84.48**
Wilt	98.23	**98.52**	80.97	**84.74**
Wine	98.75	98.75	97.22	**98.44**
Average	92.52	**93.61**	85.76	**87.27**

method has better noise robustness as compared to SVM, AdaBoost, RFs and SubRFs, and is more suitable to deal with the imbalance problem.

The Minimum Accuracy Per Class of each approach is shown in Table 3. This table shows that the proposed method is effective in increasing the accuracy of classifying the most difficult class for most data sets when compared with the other schemes. Difficult class instances have typically low margin values and hence obtain more attention, whereas instances having high margin values are potentially redundant or noisy instances.

Table 4 shows the ensemble diversity of the AdaBoost, standard RFs, SubRFs, and EMRFs. EMRFs still yields the best result and significantly outperforms RFs and SubRFs. Indeed, as the base classifiers are built using training sets of different sizes and having different data distributions, the resulting ensemble classifier is expected to have increased ensemble diversity.

4.5. Statistical analysis

The Nemenyi statistical test presented in [53], is used here to assess, with high probability, that the measured performances are evidence of differences in performance among some of the tested learning techniques. This test is a post-hoc test as it is used several times to gather more information. It infers a possible significant difference between two techniques from the difference of their mean ranks in Overall Accuracy and in Minimum Accuracy Per Class. These mean ranks are shown in Tables 2 and 3. The null-hypothesis can be rejected and a claim can be assessed when the mean rank difference is greater than a threshold called the Critical Difference (CD) set in the BonferroniDunn:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{9}$$

where α, the significant level is here set to 0.05 and q_α is based on the Studentized range statistic divided by $\sqrt{2}$, [53]. N is the number of datasets, and k is the number of algorithms. Note that for this experiment, $CD = 1.442$. We use the plots described by [53] to present the graphical representation results of the Bonferroni-Dunn test in Fig. 4: the comparative performances of the tested learning techniques are shown both for the Overall Accuracy (above) and the Minimum Accuracy Per Class (below).

4.6. Influence of model parameters on classification performance

4.6.1. Influence of the ensemble size

In order to study the influence on random forests construction of the ensemble size, T (i.e. the total number of classifiers), we present in Figs. 5 and 6 the Overall Accuracy and Minimum Accuracy Per Class with respect to T ranging from 10 to 500. From both figures, we can see that the EMRFs accuracy curve is above the SVM, AdaBoost, RFs and SubRFs curves in most datasets. RFs method statistically has better behaviors than SubRFs. The SubRFs scheme has the risk of losing useful information. Note that for most databases and much in the same way as for AdaBoost, RFs, and SubRfs, EMRFs accuracy reaches, at some point, a horizontal asymptote. Therefore T should not be considered as hyper-parameter.

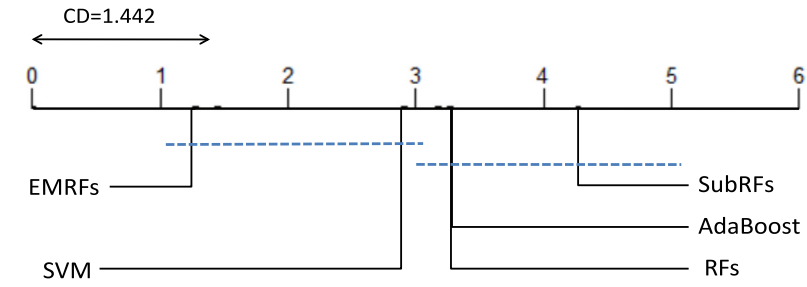
4.6.2. Influence of the initial resampling rate, Δa

This section aims to study the influence of the initial resampling rate Δa on the EMRFs classification performance. In this experiment, T is set to 500 and Δa ranges from $\frac{1}{40}$ to 1. Fig. 7 exhibits the two optimal initial resampling rates, Δa_o and Δa_m, yielding the best classification results on all the data sets according to respectively the Overall Accuracy and the Minimum Accuracy Per Class. The optimal Overall Accuracy and Minimum Accuracy Per Class are indicated above each black triangle. Table 5 presents for all the datasets, the Overall Accuracy and Minimum Accuracy Per Class achieved by EMRFs, using respectively $\Delta a = 0.1$, $\Delta a = \Delta a_m$ and $\Delta a = \Delta a_o$. Note that the values using $\Delta a = 0.1$ are those of the last column of Table 2, and the other values are those indicated in Fig. 7. From Fig. 7, it appears that for most datasets, Δa_o and Δa_m are similar. And from Table 5, accuracy is improved when Δa is tuned. The Overall Accuracy is increased up to 4% for the following datasets: *Balance*, *Clean* and *Sonar*. The Minimum Accuracy Per Class is increased by 4% for the *UrbanLandCover* dataset.

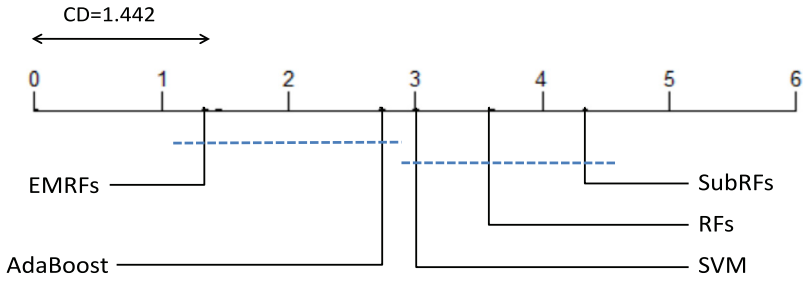
4.6.3. Performance of EMRFs on high-dimensional data

The previous experiment focuses on the accurate classification of the low-dimensional datasets. To further evaluate the effectiveness of the proposed algorithm, we analyze the performance of EMRFs on high-dimensional microarray datasets (see also Table 1). Ensemble methods cannot improve the performance of

Please cite this article as: W. Feng, G. Dauphin, W. Huang et al., New margin-based subsampling iterative technique in modified random forests for classification, Knowledge-Based Systems (2019), https://doi.org/10.1016/j.knosys.2019.07.016.



(a) Overall Accuracy



(b) Minimum Accuracy Per Class

Fig. 4. Comparison of all classifiers against each other with the Nemenyi test.

base classifiers on high-dimensional datasets when the variables are correlated [54]. The identification and removal of the relevant variables are very useful to reduce the class-overlap problem. Hence, those microarray datasets are preprocessed in the first step of our experiment. The variable importance is evaluated using the ideas by Breiman et al. [16]. Only the first 1000 important variables are kept and used for further analysis. The 5-fold CV is still used to estimate accuracy measures.

Table 6 presents the Overall Accuracy, Minimum Accuracy Per Class and ensemble diversity of the AdaBoost, standard RFs, SubRFs and the proposed EMRFs on high-dimensional microarray datasets. We do not consider SVM in our analyses. SVM does not perform well in the high-dimensional setting. In addition, although SMOTE has been adopted to alleviate the class imbalance problem, SVM is still very sensitive to the minority class instances. It can be observed in Table 6 that EMRFs still outperforms the reference ensemble methods. Although RFs has been reported as having good performance on high dimensional data, EMRFs is more efficient than RFs when data is preprocessed using feature importance filter and SMOTE. When compared with the Adaboost, RFs, and SubRF, the best increases of the EMRFs in Overall Accuracy are respectively about 3%, **7%** and **12%**. The best increases of the proposed method in Minimum Accuracy Per Class are respectively about **7%**, **40%** and **31%**. Moreover, the best results of the ensemble diversity are usually achieved with the proposed method. These positive results can be explained by the fact that it is suitable to design a random forest model by using the ensemble margin to define the significance of a sample to improve the classification accuracy of the high-dimensional datasets.

5. Discussion

1. When randomly selecting samples, many ensemble classifiers make use of classical subsampling techniques, resulting in a common data distribution shared by all base-classifier training sets. Such techniques have shown good accuracy when dealing with normal datasets. However, with imbalanced datasets, using a common data distribution entails the risk of losing important information and

Table 6
Overall Accuracy, Minimum Accuracy Per Class and diversity of the SVM, AdaBoost, standard RFs, SubRFs and the proposed EMRFs on high-dimensional microarray datasets.

		SVM	AdaBoost	RFs	SubRFs	EMRFs
Overall Accuracy	Alon	64.56	80.69(2.0)	80.25(3.0)	75.61(4.0)	**81.75**(1.0)
	Christensen	–	98.44(3.0)	**100.0**(1.0)	97.03(4.0)	99.72(2.0)
	Golub	65.31	96.25(2.0)	95.33(3.0)	85.50(4.0)	**97.58**(1.0)
	Gravier	66.09	**79.96**(1.0)	73.48(4.0)	73.80(3.0)	77.72(2.0)
	Khan	–	95.92(3.0)	98.58(2.0)	91.44(4.0)	**98.72**(1.0)
	Pomeroy	65.00	67.00(2.0)	60.83(4.0)	61.00(3.0)	**67.67**(1.0)
	Shipp	–	**94.22**(1.0)	92.35(3.0)	87.53(4.0)	93.31(2.0)
	Su	–	97.05(3.0)	97.95(2.0)	96.65(4.0)	**98.45**(1.0)
	Average	–	88.69	87.35	83.57	**89.37**
	Rank	–	2.13	2.75	3.75	1.38
Minimum Accuracy Per Class	Alon	–	70.82(3.0)	71.47(2.0)	62.39(4.0)	**71.90**(1.0)
	Christensen	–	90.67(4.0)	**100.0**(1.0)	99.33(2.0)	97.30(3.0)
	Golub	–	93.66(2.0)	89.78(3.0)	63.36(4.0)	**94.41**(1.0)
	Gravier	–	59.96(2.0)	42.55(3.0)	40.75(4.0)	**64.15**(1.0)
	Khan	–	89.78(3.0)	**96.07**(1.0)	90.00(2.0)	85.67(4.0)
	Pomeroy	–	44.12(2.0)	8.03(4.0)	30.75(3.0)	**48.45**(1.0)
	Shipp	–	**89.87**(1.0)	87.36(2.0)	85.28(4.0)	85.90(3.0)
	Su	–	92.43(4.0)	**94.79**(1.0)	94.12(2.0)	93.70(3.0)
	Average	–	78.91	73.76	70.75	**80.19**
	Rank	–	2.63	2.13	3.13	2.13
Diversity	Alon	–	0.2055	0.1909	0.1948	**0.2311**
	Christensen	–	0.0052	0.027	**0.1013**	0.0803
	Golub	–	0.1227	0.1692	0.1821	**0.1917**
	Gravier	–	0.2187	0.1948	0.1881	**0.2381**
	Khan	–	0.0672	0.1247	0.1297	**0.1451**
	Pomeroy	–	0.2345	0.2243	0.2182	**0.2452**
	Shipp	–	0.1381	0.1722	0.1784	**0.2096**
	Su	–	0.176	0.1891	0.1949	**0.2178**
	Average	–	0.1460	0.1615	0.1734	**0.1949**

Please cite this article as: W. Feng, G. Dauphin, W. Huang et al., New margin-based subsampling iterative technique in modified random forests for classification, Knowledge-Based Systems (2019), https://doi.org/10.1016/j.knosys.2019.07.016.

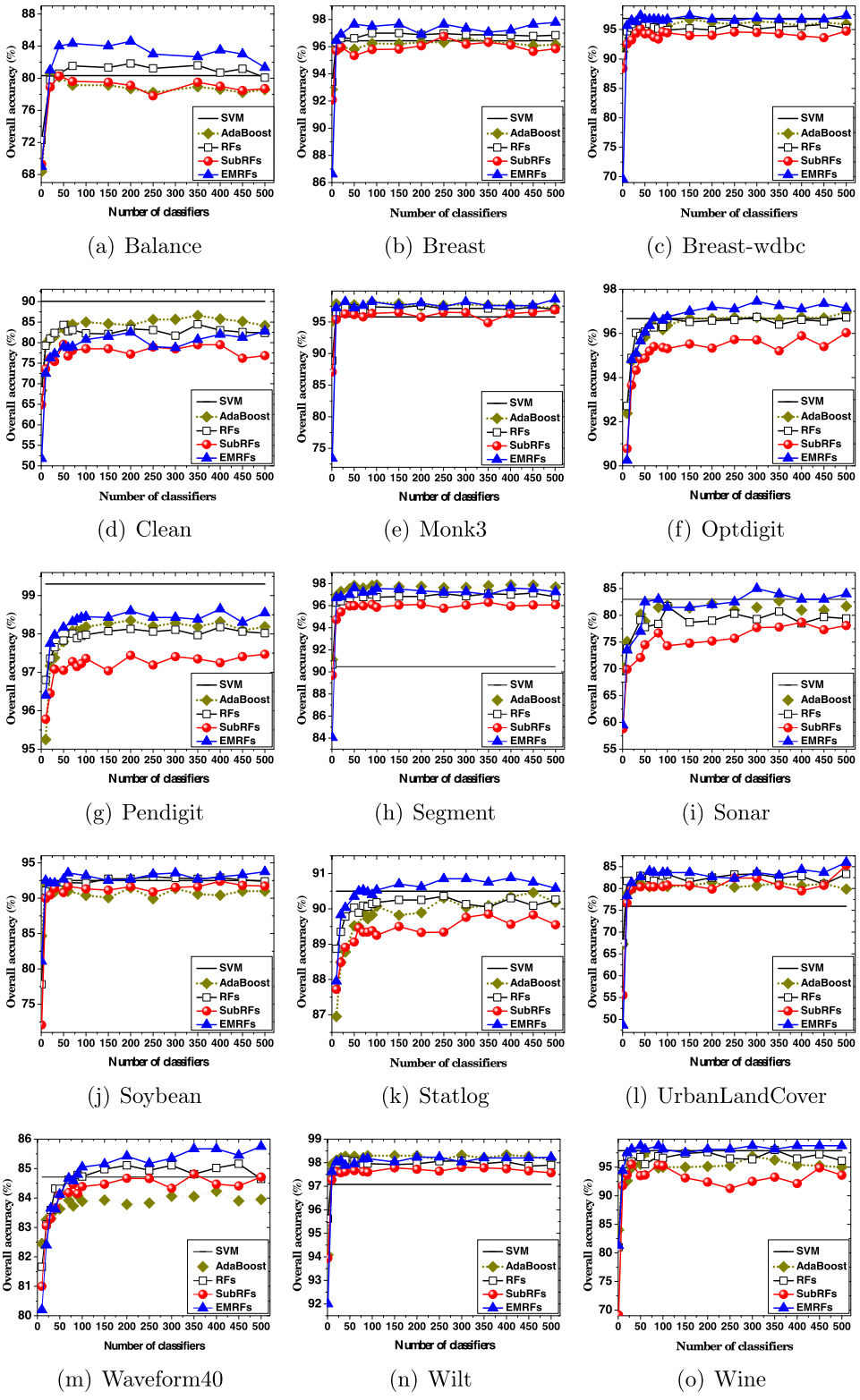


Fig. 5. Evolution of the Overall Accuracy according to the ensemble size, T.

failing to identify minority classes. The EMRFs subsampling technique gives priority to small-margin samples and to diversity, thereby ensuring that class boundary samples and minority class samples are more likely to be drawn.

2. When designing ensemble classifiers, the common practice is to set the size of the training set for each base classifier according to a trade-off between diversity and accuracy [20]. Namely increasing the size improves the accuracy whereas reducing the size improves the diversity. As opposed to RFs and SubRFs, the EMRFs base classifiers are trained with datasets of very different sizes, thanks to an iterative bootstrap technology. The resulting ensemble classifier yields both an improved diversity and accuracy.

Please cite this article as: W. Feng, G. Dauphin, W. Huang et al., New margin-based subsampling iterative technique in modified random forests for classification, Knowledge-Based Systems (2019), https://doi.org/10.1016/j.knosys.2019.07.016.

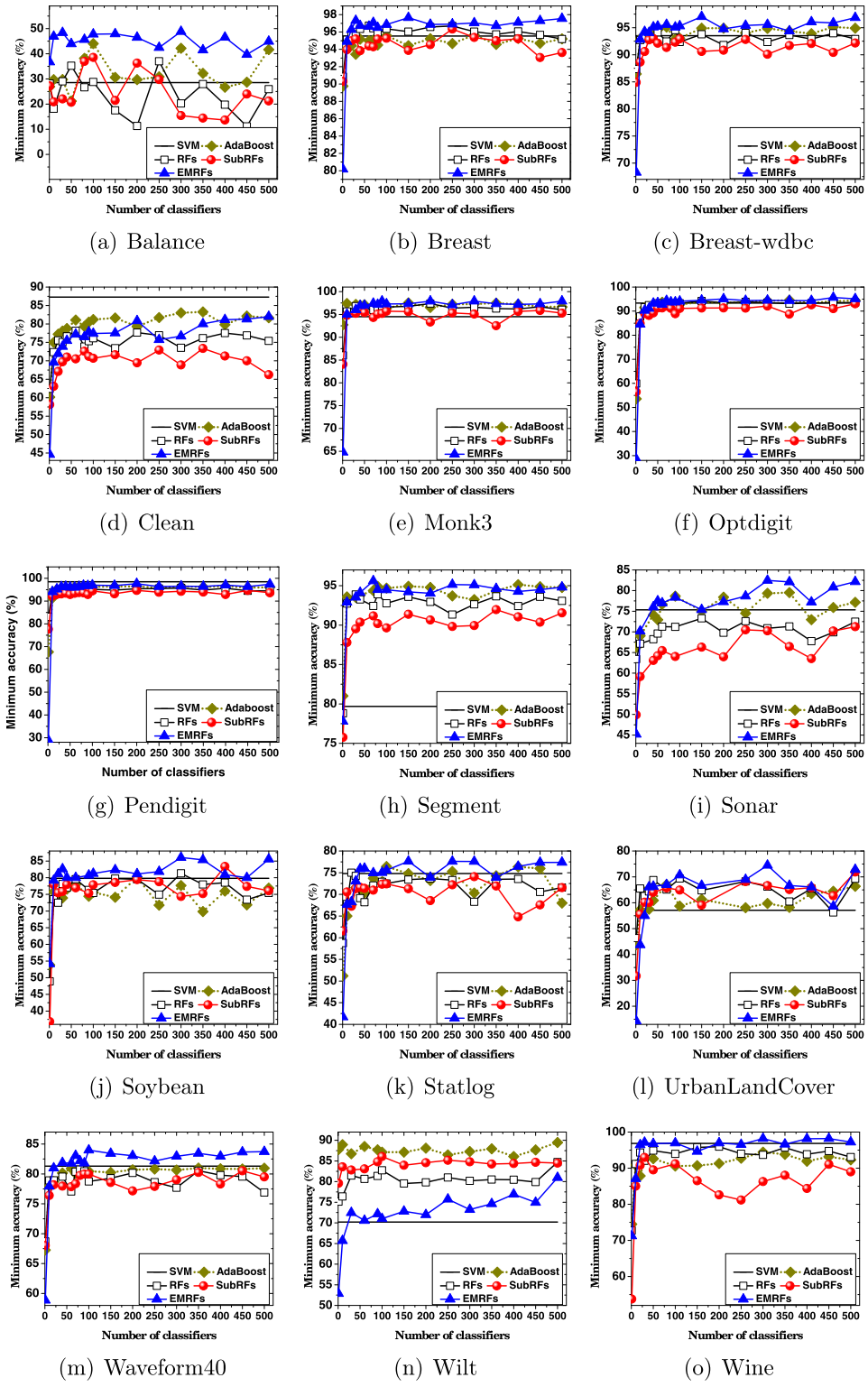


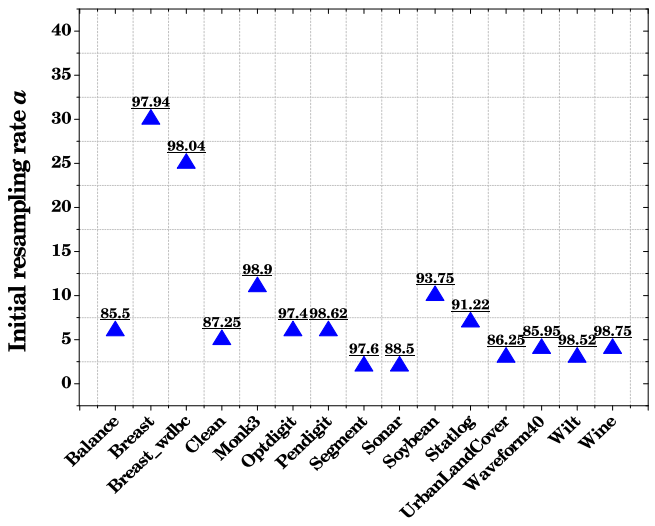
Fig. 6. Evolution of the Minimum Accuracy Per Class according to the ensemble size, T.

3. EMRFs bears some resemblance with boosting in that each sample does not receive the same interest. And boosting is known to be susceptible to class-label noise, see [55], an issue raised by [56] as weights are strongly label-dependent. On the other hand, Eqs. (3) and (4) show that significance has only small label-dependency. This provides EMRFs with more robustness to incorrect labeling.

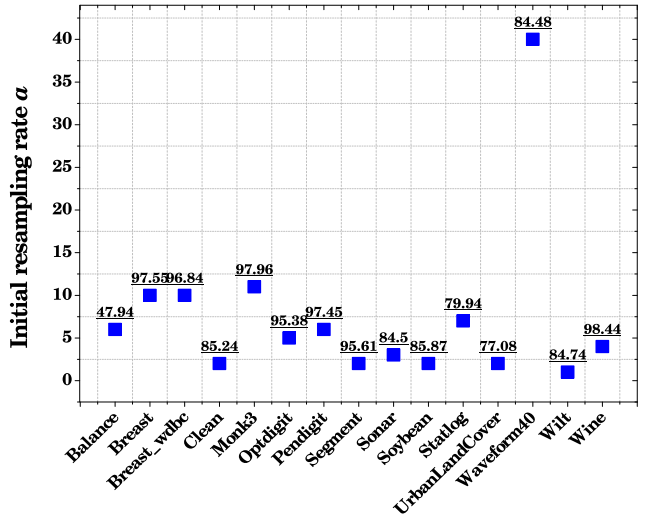
6. Conclusion

This paper proposes a novel ensemble-margin based random forests algorithm, named EMRFs. It assigns to each sample a measure of its significance by collecting information on an already trained random forests classifier. These assigned measures are used in an iterative bootstrapping technique to draw diverse sets,

Please cite this article as: W. Feng, G. Dauphin, W. Huang et al., New margin-based subsampling iterative technique in modified random forests for classification, Knowledge-Based Systems (2019), https://doi.org/10.1016/j.knosys.2019.07.016.



(a) Overall Accuracy



(b) Minimum Accuracy Per Class

Fig. 7. Optimal initial-resampling rate *a*, scaled in percentage, for all the data sets.

each of them being used to train a base classifier. It is this significance metric and this iterative bootstrapping technique that provide the resulting ensemble classifier with greater accuracy when dealing with more complex datasets and more robustness to class-label noise.

To evaluate the effectiveness of the proposed approach, SVM, AdaBoost, standard random forests and subsampled based forests are used in a comparative analysis. From this study, we have emphasized the superiority of the proposed method. The novel algorithm has three advantages: (1) the classification trees trained by focusing on the informative samples tend to be more accurate than those obtained by implementing the traditional bootstrap or the subsampling with optimized training size on original data (2) the ensemble classifier yields better performance with smaller ensemble size (3) the method is more effective for the classification of imbalanced data. The experimental results show that EMRFs has better performance in terms of Overall Accuracy, Minimum Accuracy Per Class and ensemble diversity. As future research, we plan to extend the margin-based ensemble framework by studying other iterative resampling rates.

Acknowledgments

This work is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA19080304), Hainan Provincial Key R&D Project of China (ZDYF2018073), National Natural Science Foundation of China (41601466, 41871339, 61461003), Wenzhi Liao is a postdoctoral fellow of the Research Foundation Flanders (FWOVlaanderen) and acknowledges its support.

References

[1] L. Jin, L. Feng, G. Liu, C. Wang, Personal web revisitation by context and content keywords with relevance feedback, IEEE Trans. Knowl. Data Eng. 29 (7) (2017) 1508–1521.
[2] T. Tran, D. Phung, S. Venkatesh, Collaborative filtering via sparse markov random fields, Inform. Sci. 369 (2016) 221–237.
[3] B. Bhasuran, G. Murugesan, S. Abdulkadhar, J. Natarajan, Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases, J. Biomed. Inform. 64 (2016) 1–9.
[4] A.E. Ferreira, D. Alarcao, Real-time blind source separation system with applications to distant speech recognition, Appl. Acoust. 113 (2016) 170–184.
[5] W. Feng, W. Bao, Weight-based rotation forest for hyperspectral image classification, IEEE Geosci. Remote Sens. Lett. 14 (11) (2017) 2167–2171.
[6] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, G. Zhang, Transfer learning using computational intelligence: A survey, Knowl.-Based Syst. 80 (2015) 14–23.
[7] J. Lu, X. Yang, G. Zhang, Support vector machine-based multi-source multi-attribute information integration for situation assessment, Expert Syst. Appl. 34 (2) (2008) 1333–1340.
[8] J. Abelln, C.J. Mantas, J.G. Castellano, S. Moral-Garca, Increasing diversity in random forest learning algorithm via imprecise probabilities, Expert Syst. Appl. 97 (2018) 228–243.
[9] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.
[10] B. Krawczyk, L.L. Minku, J. Gama, J. Stefanowski, M. Wozniak, Ensemble learning for data stream analysis: A survey, Inf. Fusion 37 (2017) 132–156.
[11] Z.H. Zhou, Ensemble Methods: Foundations and Algorithms, Chapman and Hall/CRC, 2012, p. 236.
[12] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Publisher: Wadsworth, 1984.
[13] J. Xia, S. Zhang, G. Cai, L. Li, Q. Pan, J. Yan, G. Ning, Adjusted weight voting algorithm for random forests in handling missing values, Pattern Recognit. 69 (2017) 52–60.
[14] R. Houborg, M.F. McCabe, A hybrid training approach for leaf area index estimation via cubist and random forests machine-learning, ISPRS J. Photogramm. Remote Sens. 135 (2018) 173–188.
[15] D. Turner, A. Lucieer, Z. Malenovsk, D. King, S.A. Robinson, Assessment of antarctic moss health from multi-sensor uas imagery with random forest modelling, Int. J. Appl. Earth Obs. Geoinf. 68 (2018) 168–179.
[16] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[17] Q. Dai, R. Ye, Z. Liu, Considering diversity and accuracy simultaneously for ensemble pruning, Appl. Soft Comput. 58 (2017) 75–91.
[18] H. Guo, H. Liu, R. Li, C. Wu, Y. Guo, M. Xu, Margin & diversity based ordering ensemble pruning, Neurocomputing 275 (2018) 237–246.
[19] W. Feng, W. Huang, J. Ren, Class imbalance ensemble learning based on the margin theory, Appl. Sci. 8 (815) (2018).
[20] L. Kuncheva, C. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Mach. Learn. 51 (2003) 181–207.
[21] J.I. Orlando, E.P.M. Fresno, M.B. Blaschko, An ensemble deep learning based approach for red lesion detection in fundus images, Comput. Methods Programs Biomed. 153 (2018) 115–127.
[22] W. Feng, S. Boukir, Class noise removal and correction for image classification using ensemble margin, in: 2015 IEEE International Conference on Image Processing (ICIP), 2015, pp. 4698–4702.
[23] H. Guo, Y. Li, S. Jennifer, M. Gu, Y. Huang, B. Gong, Learning from class-imbalanced data: Review of methods and applications, Expert Syst. Appl. 73 (2017) 220–239.
[24] C. Chen, A. Liaw, L. Breiman, Using Random Forest to Learn Imbalanced Data, Technical Report 666, Department of Statistics, University of California, Berkeley., 2004.
[25] S. Bernard, S. Adam, L. Heutte, Dynamic random forests, Pattern Recognit. Lett. 33 (12) (2012) 1580–1586.
[26] C. Zhang, X. Bian, P. Liu, X. Tan, Q. Fan, W. Liu, L. Lin, Subagging for the improvement of predictive stability of extreme learning machine for spectral quantitative analysis of complex samples, Chemometr. Intell. Lab. Syst. 161 (2017) 43–48.

[27] R.K.H. Galvao, M.C.U. Araujo, M.N. Martins, G.E. Jose, M.J.C. Pontes, E.C. Silva, T.C.B. Saldanha, An application of subagging for the improvement of prediction accuracy of multivariate calibration models, Chemometr. Intell. Lab. Syst. 81 (1) (2006) 60–67.
[28] G. Paleologo, A. Elisseeff, G. Antonini, Subagging for credit scoring models, European J. Oper. Res. 201 (2) (2010) 490–499.
[29] R. Genuer, J. Poggi, C. Tuleau-Malot, N. Villa-Vialaneix, Random forests for big data, Big Data Res. 9 (2017) 28–46.
[30] A. Buja, W. Stuetzle, Observations on bagging., Preprint. (2002).
[31] P. Buhlmann, Bagging, subagging and bragging for improving some prediction algorithms, in: M.G. Akritas, D.N. Politis (Eds.), Recent Advances and Trends in Nonparametric Statistics, JAI, Amsterdam, 2003, pp. 19–34.
[32] L. Breiman, Pasting small votes for classification in large databases and on-line, Mach. Learn. 36 (1) (1999) 85–103.
[33] G. Martinez-Munoz, A. Suarez, Out-of-bag estimation of the optimal sample size in bagging, Pattern Recognit. 43 (1) (2010) 143–152.
[34] L. Breiman, Out-Of-Bag Estimation, Technical Report, University of California, 1996.
[35] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[36] L. Li, A. Pratap, H.T. Lin, Y.S. Abu-Mostafa, Improving generalization by data categorization, in: Knowledge Discovery in Databases: PKDD 2005, Vol. 3721, Springer Berlin Heidelberg, 2005, pp. 157–168.
[37] E. Marchiori, Class conditional nearest neighbor for large margin instance selection, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2) (2010) 364–370.
[38] M. Alshawabkeh, Hypothesis Margin Based Weighting for Feature Selection Using Boosting: Theory, Algorithms and Applications (Ph.D. Thesis), Northeastern University, 2013.
[39] L.J. Li, B. Zou, Q.H. Hu, X.Q. Wu, D.R. Yu, Dynamic classifier ensemble using classification confidence, Neurocomputing 99 (2013) 581–591.
[40] C.H. Shen, H.X. Li, Boosting through optimization of margin distributions, Trans. Neur. Netw. 21 (4) (2010) 659–666.
[41] Z.X. Xie, Y. Xu, Q.H. Hu, P.F. Zhu, Margin distribution based bagging pruning, Neurocomputing 85 (2012) 11–19.
[42] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, J. Artif. Int. Res. 16 (1) (2002) 321–357.
[43] M.S. Santos, J.P. Soares, P.H. Abreu, H. Araujo, J. Santos, Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier], IEEE Comput. Intell. Mag. 13 (4) (2018) 59–76.
[44] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag New York, Inc., 1995.
[45] Y. Freund, R. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. System Sci. 55 (1) (1997) 119–139.
[46] K. Crammer, R. Gilad-bachrach, A. Navot, N. Tishby, Margin analysis of the lvq algorithm, in: Advances in Neural Information Processing Systems 2002, MIT press, 2002, pp. 462–469.
[47] R. Schapire, Y. Freund, P. Bartlett, W. Lee, Boosting the margin: a new explanation for the effectiveness of voting methods, Ann. Statist. 26 (5) (1998) 1651–1686.
[48] P. Probst, M. Wright, A. Boulesteix, Hyperparameters and tuning strategies for random forest, arXiv preprint arXiv:1804.03515 (2018).
[49] R. Kohavi, D. Wolpert, Bias plus variance decomposition for zero-one loss functions, in: 13th International Conference of Machine Learning, in: ICML'96, 1996, pp. 275–283.
[50] J.A. Sáez, B. Krawczyk, M. Woźniak, Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets, Pattern Recognit. 57 (2016) 164–178.
[51] M. Kapp, R. Sabourin, P. Maupin, An empirical study on diversity measures and margin theory for ensembles of classifiers, in: 10th International Conference on Information Fusion, 2007, pp. 1–8.
[52] A. Asuncion, D. Newman, UCI machine learning repository, 2007.
[53] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[54] L. Wei, J.J. Chen, Class-imbalanced classifiers for high-dimensional data, Brief. Bioinform. 14 (1) (2013) 13–26.
[55] C. Leistner, A. Saffari, P.M. Roth, H. Bischof, On robustness of on-line boosting - a competitive study, in: 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, 2009, pp. 1362–1369.
[56] R.E. Schapire, The strength of weak learnability, Mach. Learn. 5 (2) (1990) 197–227.

Please cite this article as: W. Feng, G. Dauphin, W. Huang et al., New margin-based subsampling iterative technique in modified random forests for classification, Knowledge-Based Systems (2019), https://doi.org/10.1016/j.knosys.2019.07.016.