Hyperspectral Image Classification Using Relevance Vector Machines

Begüm Demir, *Student Member, IEEE*, and Sarp Ertürk, *Member, IEEE*

***Abstract*—This letter presents a hyperspectral image classification method based on relevance vector machines (RVMs). Support vector machine (SVM)-based approaches have been recently proposed for hyperspectral image classification and have raised important interest. In this letter, it is genuinely proposed to use an RVM-based approach for the classification of hyperspectral images. It is shown that approximately the same classification accuracy is obtained using RVM-based classification, with a significantly smaller relevance vector rate and, therefore, much faster testing time, compared with SVM-based classification. This feature makes the RVM-based hyperspectral classification approach more suitable for applications that require low complexity and, possibly, real-time classification.**

***Index Terms*—Classification, hyperspectral images, relevance vector machines (RVMs), support vector machines (SVMs).**

I. Introduction

SUPPORT vector machine (SVM)-based approaches [1]–[5] have been recently proposed for regression and classification tasks in multispectral [6], [7] and hyperspectral [8]–[12] images. For example, in [8], SVM classifiers have been applied to hyperspectral Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) data. The effectiveness of SVM-based hyperspectral image classification has been addressed in [9]. An SVM classification system that estimates the SVM parameters in a fully automatic way is presented in [10]. Different kernel-based approaches and their properties have been analyzed for hyperspectral image classification in [11]. In [12], a smoothing preprocessing step is introduced before SVM classification, to take spatial context into consideration and improve the classification rate. In [13] and [14], it has been proposed to use SVM-based regression for density estimation of class conditional probabilities for maximum *a posteriori* classification.

Relevance vector machine (RVM)-based regression and classification have been proposed in [15]–[17]. The advantages of the RVM over the SVM are probabilistic predictions, automatic estimations of parameters, and the possibility of choosing arbitrary kernel functions [15], [16]. Most importantly, RVM classification results in fewer relevance vectors (RVs) compared with the number of support vectors (SVs) obtained in the SVM classification. Hence, classification can be carried out much faster with the RVM compared to the SVM. For example, in [17], the RVM has been used for the detection of microcalcification clusters in digital mammograms, and it has been shown that the RVM classifier is much more suitable for real-time processing and reduces the computational complexity compared to SVM-based classification, while maintaining similar detection accuracy.

It is proposed in this letter to utilize the RVM for classification of hyperspectral images. It is shown that the RVM-based classification approach can provide similar classification accuracy (AC) as the SVM-based classification, with a significantly reduced number of RVs. This feature makes the RVM-based hyperspectral classification approach more suitable for applications that require low complexity and, possibly, real-time classification.

II. RVM Classification

Supervised learning techniques make use of a training set that consists of a set of sample input vectors $\{\mathbf{x}_n\}_{n=1}^N$ together with the corresponding targets $\{t_n\}_{n=1}^N$. The targets are basically real values in regression tasks or class labels in classification problems. It is typically desired to learn a model of the dependency of the targets on the inputs from the training set, so that accurate predictions of t can be made for previously unseen values of $\mathbf{x}$. Commonly, these predications can be based on some function $y(\mathbf{x})$ defined over the input space in the form of

$$y(\mathbf{x};\mathbf{w}) = \sum_{i=1}^{M} w_i \phi_i(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}) \tag{1}$$

as a linearly weighted sum of M (generally nonlinear and fixed) basis functions $\boldsymbol{\varphi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_M(\mathbf{x}))^T$. Although this model is linear in the parameters (or weights) $\mathbf{w} = (w_1, w_2, \ldots, w_M)^T$, it can still be highly flexible as the size of the basis set M can be effectively large.

Learning is basically the process of inferring the function or, equivalently, the parameters of the function $y(\mathbf{x})$. In this context, it is desired to estimate reasonable values for the parameters (or weights) $\mathbf{w} = (w_1, w_2, \ldots, w_M)^T$. Given a set of N corresponding training pairs $\{\mathbf{x}_n, t_n\}_{n=1}^N$, the objective is to find values for the weights $\mathbf{w} = (w_1, w_2, \ldots, w_M)^T$, such that $y(\mathbf{x})$ generalizes well enough to new data, yet only a few elements of $\mathbf{w}$ are nonzero [15]. Having only a few nonzero weights facilitates a sparse representation with the advantage of providing fast implementation.

Manuscript received January 11, 2007; revised April 2, 2007. This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under the "Hyperspectral Classification, Segmentation and Recognition" project.

The authors are with the Kocaeli University Laboratory of Image and Signal Processing (KULIS), University of Kocaeli, 41040 Kocaeli, Turkey.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2007.903069

The SVM [1]–[5] provides a successful approach to supervised learning by making predictions based on a function in the form of

$$y(\mathbf{x};\mathbf{w}) = \sum_{i=1}^{N} w_i K(\mathbf{x},\mathbf{x}_i) + w_0 \tag{2}$$

where w_i shows the model weights, and $K(\cdot,\cdot)$ is a kernel function effectively defining one basis function for each sample in the training set. The key feature of the SVM classification is that, its target function attempts to minimize a measure of error on the training set, while simultaneously maximizing the margin between the two classes that are implicitly defined in the feature space by the kernel K [2]. This process results in a sparse model that depends only on a subset of kernel functions, namely, those associated with training examples that lie either on the margin or on the wrong side, and the corresponding training examples are referred to as "SVs." SVM is quite popular in supervised learning applications and has been recently applied for regression and classification of multispectral [6], [7] as well as hyperspectral images [8]–[12], and therefore, the reader is referred to these references to avoid rephrasing the basics of SVM. Although the SVM classification provides successful results, a number of significant and practical disadvantages are identified as [15], [16] follows.

- Although SVMs are relatively sparse, the number of SVs typically grows linearly with the size of the training set, and therefore, SVMs make unnecessarily liberal use of basis functions.
- Predictions are not probabilistic, and therefore, SVM is not suitable for classification tasks in which posterior probabilities of class membership are necessary.
- In SVM, it is required to estimate the error/margin tradeoff parameter C, which generally entails a cross-validation procedure which can be a waste of data as well as computation.
- In SVM, the kernel function must satisfy Mercer's condition; hence, it must be a continuous symmetric kernel of a positive integral operator.

The RVM has been introduced by Tipping [15], [16] as a Bayesian treatment alternative to the SVM that does not suffer from the aforementioned limitations. The RVM introduces a prior over the model weights governed by a set of hyperparameters, in a probabilistic framework. One hyperparameter is associated with each weight, and the most probable values are iteratively estimated from the training data. The most compelling feature of the RVM is that it typically utilizes significantly fewer kernel functions compared to the SVM, while providing a similar performance.

For two-class classification, any target can be classified into two classes such that $t_n \in \{0,1\}$. A Bernoulli distribution can be adopted for $p(t|\mathbf{x})$ in the probabilistic framework because only two values (0 and 1) are possible. The logistic sigmoid link function $\sigma(y) = 1/(1+e^{-y})$ is applied to $y(\mathbf{x})$ to link random and systematic components, and generalize the linear model. Following the definition of the Bernoulli distribution, the likelihood is written as

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} \sigma\left\{y(x_n;\mathbf{w})\right\}^{t_n} \left[1-\sigma\left\{y(x_n;\mathbf{w})\right\}\right]^{1-t_n} \tag{3}$$

for the targets $t_n \in \{0,1\}$.

The likelihood is complemented by a prior over the parameters (weights) in the form of

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{N} \frac{\sqrt{\alpha_i}}{\sqrt{2\pi}} \exp\left(-\frac{\alpha_i w_i^2}{2}\right) \tag{4}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)^T$ shows the hyperparameters introduced to control the strength of the prior over its associated weight. Hence, the prior is Gaussian, but conditioned on $\boldsymbol{\alpha}$.

For a certain $\boldsymbol{\alpha}$ value, the posterior weight distribution conditioned on the data can be obtained using Bayes' rule, i.e.,

$$p(\mathbf{w}|\mathbf{t},\boldsymbol{\alpha}) = \frac{p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})}{p(\mathbf{t}|\boldsymbol{\alpha})} \tag{5}$$

where $p(\mathbf{t}|\mathbf{w})$ is the likelihood, $p(\mathbf{w}|\boldsymbol{\alpha})$ is the prior, and $p(\mathbf{t}|\boldsymbol{\alpha})$ is referred to as evidence.

The weights cannot be analytically obtained, and therefore, a Laplacian approximation procedure [18] is used.

1) Since $p(\mathbf{w}|\mathbf{t},\boldsymbol{\alpha})$ is linearly proportional to $p(\mathbf{t}|\mathbf{w}) \times p(\mathbf{w}|\boldsymbol{\alpha})$, it is possible to aim to find the maximum of

$$\log\left\{p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})\right\} = \sum_{n=1}^{N}\left[t_n \log y_n + (1-t_n)\log(1-y_n)\right] - \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w} \tag{6}$$

for the most probable weights $\mathbf{w}_{\mathrm{MP}}$, with $y_n = \sigma\{y(\mathbf{x}_n;\mathbf{w})\}$ and $\mathbf{A} = \mathrm{diag}(\alpha_0, \alpha_1, \ldots, \alpha_N)$ being composed of the current values of $\boldsymbol{\alpha}$. This is a penalized logistic log-likelihood function and requires iterative maximization. The iteratively reweighed least-squares algorithm [15], [19] can be used to find $\mathbf{w}_{\mathrm{MP}}$.

2) The logistic log-likelihood function can be differentiated twice to obtain the Hessian in the form of

$$\nabla_{\mathbf{w}}\nabla_{\mathbf{w}} \log p(\mathbf{w}|\mathbf{t},\boldsymbol{\alpha})\big|_{\mathbf{w}_{\mathrm{MP}}} = -(\boldsymbol{\Phi}^T\mathbf{B}\boldsymbol{\Phi} + \mathbf{A}) \tag{7}$$

where $\mathbf{B} = \mathrm{diag}(\beta_1, \beta_2, \ldots, \beta_N)$ is a diagonal matrix with $\beta_n = \sigma\{y(\mathbf{x}_n;\mathbf{w}_{\mathrm{MP}})\}[1-\sigma\{y(\mathbf{x}_n;\mathbf{w}_{\mathrm{MP}})\}]$, and $\boldsymbol{\Phi}$ is the 'design' matrix with $\boldsymbol{\Phi}_{nm} = K(x_n, x_{m-1})$ and $\boldsymbol{\Phi}_{n1} = 1$. This result is then negated and inverted to give the covariance $\boldsymbol{\Sigma}$, as shown as follows, for a Gaussian approximation to the posterior over weights centered at $\mathbf{w}_{\mathrm{MP}}$:

$$\boldsymbol{\Sigma} = (\boldsymbol{\Phi}^T\mathbf{B}\boldsymbol{\Phi} + \mathbf{A})^{-1}. \tag{8}$$

In this way, the classification problem is locally linearized around $\mathbf{w}_{\mathrm{MP}}$ in an effective way with

$$\mathbf{w}_{\mathrm{MP}} = \boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{B}\hat{\mathbf{t}} \tag{9}$$

$$\hat{\mathbf{t}} = \boldsymbol{\Phi}\mathbf{w}_{\mathrm{MP}} + \mathbf{B}^{-1}(\mathbf{t}-\mathbf{y}). \tag{10}$$

TABLE I
NUMBER OF TRAINING AND TEST SAMPLES

Class	Training	Test
C1-Corn-no till	742	692
C2-Corn-min till	442	392
C3-Grass/Pasture	260	237
C4-Grass/Trees	389	358
C5-Hay-windrowed	236	253
C6-Soybean-no till	487	481
C7-Soybean-min till	1245	1223
C8-Soybean-clean till	305	309
C9-Woods	651	643
Total	4757	4588

These equations are basically equivalent to the solution of a generalized least-squares problem. After obtaining $\mathbf{w}_{\mathrm{MP}}$, the hyperparameters α_i are updated using $\alpha_i^{\mathrm{new}} = \lambda_i / w_i^2$, where w_i is the ith posterior mean weight, and λ_i is defined as $\lambda_i = 1 - \alpha_i \Sigma_{ii}$, where Σ_{ii} is the ith diagonal element of the covariance, and can be regarded as a measure of how well determined each parameter w_i is by the data. During the optimization process, many α_i will have large values, and thus, the corresponding model weights are pruned out, realizing sparsity. The optimization process typically continues until the maximum change in α_i values is below a certain threshold or the maximum number of iterations is reached.

III. EXPERIMENTAL RESULTS

RVM and SVM classification methods have been applied to a sample hyperspectral image which is taken over northwest Indiana's Indian Pine test site in June 1992 [20], as the ground truth classification result of this image is already available. The data consist of 145 $\times$ 145 pixels with 220 bands. The number of spectral bands is initially reduced to 200 by removing bands, covering water absorption as well as noisy bands. The original ground truth has actually 16 classes, but some classes have a very small number of elements, and therefore, nine classes that have the highest number of elements have been selected and used to generate 4757 training samples and 4588 test samples, which are shown in Table I.

The most popular kernels used in SVM and RVM are the linear, polynomial, and radial basis function (RBF) kernels. The linear kernel typically shows a lower performance and is therefore not employed in the provided results. Note that γ serves as an inner product coefficient for the polynomial kernel, whereas it determines the RBF width in the case of the RBF kernel.

Linear kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j. \tag{11}$$

Polynomial kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i \cdot \mathbf{x}_j)^d. \tag{12}$$

RBF kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right). \tag{13}$$

TABLE II
CLASSIFICATION AC AND NUMBER OF SVS FOR SVM

Method	Kernel type	Kernel Parameter		C	AC	SV
		γ	D			
SVM	RBF	0.1	-	1000	90.88	3195
SVM	RBF	1	-	65	91.95	3259
SVM	RBF	2	-	40	92.67	3393
SVM	RBF	2	-	1000	92.56	3207
SVM	Poly.	1	7	60	90.03	2066

TABLE III
CLASSIFICATION AC AND NUMBER OF RVS FOR RVM

Method	Kernel type	Kernel parameter		AC	RV
		γ	d		
RVM	RBF	0.1	-	89.41	414
RVM	RBF	0.5	-	90.30	541
RVM	RBF	1	-	90.14	514
RVM	RBF	2	-	90.32	592
RVM	RBF	3	-	90.03	509
RVM	Poly.	1	7	89.39	412

The SVM is intrinsically biclass [21], and extending it to multiclass problems is an ongoing research issue [22], [23]. Because it is computationally more expensive to directly solve multiclass problems [3], it is common to combine several binary SVM classifiers for this purpose. In the one-against-all method [4], each class is trained against the remaining $K-1$ classes, where K is the total number of classes. The one-against-all method has to test K binary decision functions to predict a sample data point. In the one-against-one method [5], $K(K-1)/2$ binary classifiers are trained and $K(K-1)/2$ binary tests are required to make a final decision. Each outcome gives one vote to the winning class. The class with the most votes is selected as the final result. In this letter, we have applied the second approach for SVM as well as RVM because it typically provides faster training. Although RVM is theoretically not limited to binary classifiers, this is of little use in practice, since the size of the Hessian matrix (used while maximizing the likelihood and updating the weights) grows with the number of classes [24].

Table II shows results for the SVM-based classification (with LIBSVM that uses sequential minimal optimization [25]), and Table III shows results for the RVM-based classification of the hyperspectral test image. It is seen from these results, that for similar classification AC, RVM requires a significantly less number of RVs; hence, the classification time is considerably reduced. Comparing the maximum classification AC, it is seen that RVM provides a slightly lower (about 2%) maximum classification AC compared to SVM, and the main reason is possibly that the RVM classifier uses a significantly lower number of RVs compared with the number of SVs used in SVM, resulting in a more sparse representation, which is analogous to results presented in the literature. Note that cross validation is used to obtain the parameters for SVM; also, results are given for a variety of different parameter combinations. The same parameters are used for RVM to evaluate its performance.

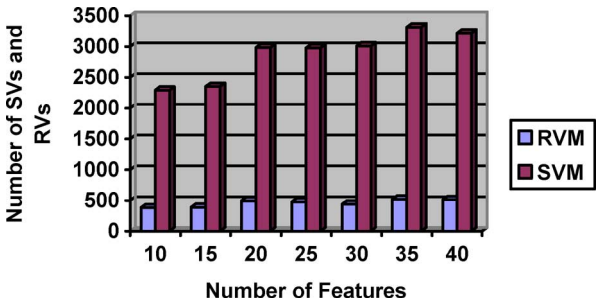

Fig. 1. Number of SVs for SVM classification and number of RVs for RVM classification with respect to the number of features.

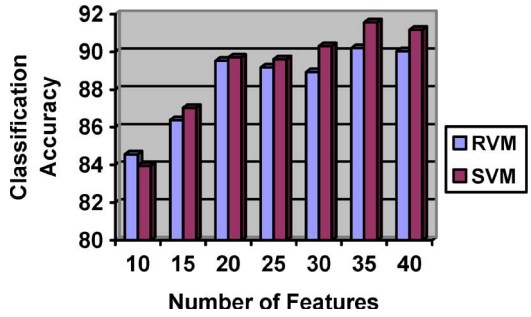

Fig. 2. SVM and RVM classification ACs with respect to the number of features.

To compare the performance of RVM and SVM classifications with different number of features, the steepest ascent [9], [26] algorithm is used to reduce the number of features (i.e., bands used in classification). Fig. 1 shows the number of SVs for the SVM classification and RVs for the RVM classification, and Fig. 2 shows the RVM and SVM classification ACs for different number of features. It is seen that the classification AC of RVM and SVM shows a similar behavior under feature reduction, whereas the RVM always results in a smaller number of RVs compared with the number of SVs obtained in the SVM. Note that $\gamma = 2$ and $C = 1000$ are used for these results.

For comparative evaluation, the one-norm SVM, as presented in [27], is also utilized for hyperspectral classification, as it is known to provide a rather simple and sparse classification method, and a classification AC of 78.81% is obtained. This value is significantly lower than the accuracies obtained with SVM and RVM.

Experimental results show that RVM is superior to SVM in terms of the number of kernel functions that needs to be used in the classification (i.e., testing) stage. Therefore, RVM is preferable to SVM in applications that require low complexity and, possibly, real-time classification with *a priori* training.

Although RVM is certainly preferable to SVM in terms of the time performance in the test (classification) phase, it has to be noted that the training time of RVM is longer than SVM because the update rules for the hyperparameters depend on computing the posterior weight covariance matrix. It is shown in [17] that the training time of RVM is about seven to eight times longer than that of SVM, whereas the testing time of RVM is about seven to eight times shorter than SVM. It is however noted in [16] that the increased training time of RVM is significantly offset by the lack of necessity to perform cross validation over nuisance parameters. It must also be noted that the training complexity of RVM in the training phase is due to the necessity of repeatedly computing and inverting the Hessian matrix, which for a set of N samples requires $\mathrm{O}(N^2)$ storage and $\mathrm{O}(N^3)$ computation. For large data sets, this makes training considerably slower than SVM [15].

TABLE IV
CLASSIFICATION RESULTS FOR DIFFERENT TRAINING DATA SIZE

Training Size	Test Size	SVM		RVM	
		AC	SV	AC	RV
590	4588	84.39	937	80.95	192
1320	4588	87.94	1510	84.45	254
2376	4588	90.56	2217	87.05	353
4757	4588	92.67	3393	90.32	592

Table IV shows the effect of the training set size on the SVM and RVM classifications for $\gamma = 2$ and $C = 40$. In this case, classification is carried out with five differently chosen training sets of the given size, and the median result is presented to avoid bias. It is seen that RVM always results in a significantly lower number of RVs compared with the number of SVs in SVM, at the cost of a slightly lower classification AC. Note that the number of SVs and RVs always shows the total number used in the one-against-one classifications with common vectors being separately counted.

IV. CONCLUSION

RVM-based hyperspectral image classification is presented in this letter. It is shown to provide similar classification AC, with a significantly smaller RV rate and, therefore, much faster testing time, compared with the SVM-based classification. Hence, the RVM classification is superior to the SVM classification in terms of sparsity. This makes the RVM-based hyperspectral classification approach more suitable for applications that require low complexity and, possibly, real-time classification. However, the classification AC yielded by RVM is less accurate than SVM, particularly for reduced training samples.

ACKNOWLEDGMENT

The authors would like to thank R. Johansson for his help with the RVM, D. Landgrebe for providing the AVIRIS data [20], C.-J. Lin for the LIBSVM software [25], and the anonymous reviewers whose comments have significantly improved this letter.

REFERENCES

[1] B. E. Boser, I. M. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. ACM Workshop Comput. Learn. Theory*, 1992, pp. 144–152.

[2] C. Burges, "A tutorial on support vector machines for pattern recognition," in *Proc. Data Mining and Knowl. Discovery*, U. Fayyad, Ed., 1998, pp. 1–43.

[3] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[4] L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Muller, E. Sackinger, P. Simard, and V. Vapnik, "Comparison of classifier methods: A case study in handwriting digit recognition," in *Proc. Int. Conf. Pattern Recog.*, 1994, pp. 77–87.

[5] S. Knerr, L. Personnaz, and G. Dreyfus, "Single-layer learning revisited: A stepwise procedure for building and training a neural network," in *Neurocomputing: Algorithms, Architectures and Applications*, J. Fogelman, Ed. New York: Springer-Verlag, 1990.

[6] C. Huang, L. S. Davis, and J. R. G. Townshend, "An assessment of support vector machines for land cover classification," *Int. J. Remote Sens.*, vol. 23, no. 4, pp. 725–749, Feb. 2002.

[7] F. Roli and G. Fumera, "Support vector machines for remote-sensing image classification," *Proc. SPIE*, vol. 4170, pp. 160–166, 2001.

[8] J. A. Gualtieri, S. R. Chettri, R. F. Cromp, and L. F. Johnson, "Support vector machine classifiers as applied to AVIRIS data," in *Proc. Summaries 8th JPL Airborne Earth Sci. Workshop*, 1999, pp. 217–227, JPL Pub. 99-17.

[9] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[10] Y. Bazi and F. Melgani, "Toward an optimal SVM classification system for hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3374–3385, Nov. 2006.

[11] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1352–1362, Jun. 2005.

[12] W. M. Lennon, G. Mercier, and L. Hubert-Moy, "Classification of hyperspectral images with nonlinear filtering and support vector machines," in *Proc. IGARSS*, 2002, vol. 3, pp. 1670–1672.

[13] A. A. Farag, R. M. Mohamed, and A. El-Baz, "A unified framework for MAP estimation in remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 7, pp. 1617–1634, Jul. 2005.

[14] P. Mantero, G. Moser, and S. B. Serpico, "Partially supervised classification of remote sensing images through SVM-based probability density estimation," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 559–570, Mar. 2005.

[15] M. E. Tipping, "The relevance vector machine," in *Advances in Neural Information Processing Systems*, vol. 12, S. A. Solla, T. K. Leen, and K.-R. Müller, Eds. Cambridge, MA: MIT Press, 2000.

[16] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[17] W. Liyang, Y. Yongyi, R. M. Nishikawa, M. N. Wernick, and A. Edwards, "Relevance vector machine for automatic detection of clustered microcalcifications," *IEEE Trans. Med. Imag.*, vol. 24, no. 10, pp. 1278–1285, Oct. 2005.

[18] D. J. C. MacKay, "The evidence framework applied to classification networks," *Neural Comput.*, vol. 4, no. 5, pp. 720–736, 1992.

[19] I. T. Nabney, "Efficient training of RBF networks for classification," in *Proc. 9th ICANN*, 1999, vol. 1, pp. 210–215.

[20] *AVIRIS NW Indiana's Indian Pines 1992 Data Set*. (original files), (ground truth). [Online]. Available: ftp://ftp.ecn.purdue.edu/biehl/MultiSpec/92AV3C

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[22] D. Anguita, S. Ridella, and D. Sterpi, "A new method for multiclass support vector machines," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jul. 2004, vol. 1, pp. 407–412.

[23] R. Rifkin and A. Klautau, "In defense of one-versus-all classification," *J. Mach. Learn. Res.*, vol. 5, pp. 101–141, 2004.

[24] R. Johansson and P. Nugues, "Sparse Bayesian classification of predicate arguments," in *Proc. 9th Conf. Comput. Natural Language Learn., 43rd Annu. Meeting Assoc. Comput. Linguistics*, Ann Arbor, MI, 2005, pp. 177–200.

[25] C.-C. Chang and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines*, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[26] S. Serpico and L. Bruzzone, "A new search algorithm for feature selection in hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1360–1367, Jul. 2001.

[27] F. Glen and O. Mangasarian, "A feature selection Newton method for support vector machine classification," *Comput. Optim. Appl.*, vol. 28, no. 2, pp. 185–202, Jul. 2004.

Pattern Recognition 44 (2011) 330–349

Contents lists available at ScienceDirect

Pattern Recognition

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

Mining data with random forests: A survey and results of new tests

A. Verikas [a,b,*], A. Gelzinis [b], M. Bacauskiene [b]

[a] Intelligent Systems Laboratory, Halmstad University, Box 823, S 301 18 Halmstad, Sweden
[b] Department of Electrical & Control Equipment, Kaunas University of Technology, Studentu 50, LT-51368, Kaunas, Lithuania

ARTICLE INFO

Article history:
Received 18 December 2009
Received in revised form
2 August 2010
Accepted 7 August 2010

Keywords:
Random forests
Variable importance
Variable selection
Classifier
Data proximity

ABSTRACT

Random forests (RF) has become a popular technique for classification, prediction, studying variable importance, variable selection, and outlier detection. There are numerous application examples of RF in a variety of fields. Several large scale comparisons including RF have been performed. There are numerous articles, where variable importance evaluations based on the variable importance measures available from RF are used for data exploration and understanding. Apart from the literature survey in RF area, this paper also presents results of new tests regarding variable rankings based on RF variable importance measures. We studied experimentally the consistency and generality of such rankings. Results of the studies indicate that there is no evidence supporting the belief in generality of such rankings. A high variance of variable importance evaluations was observed in the case of small number of trees and small data sets.

1. Introduction

Growing size of data sets increases the variety of problems characterized by a large number of variables. Nowadays, it is not uncommon that the number of variables N is larger than the number of observations M. Microarray gene expression data is a characteristic example, where $N \gg M$ most often. Traditional statistical techniques experience problems, when $N > M$. Therefore, machine learning-based techniques are usually applied in such cases.

Support vector machine (SVM) [1,2], multilayer perceptron (MLP) [3], relevance vector machine (RVM) [4], and various ensembling approaches [5] are probably the most popular machine learning techniques applied to create predictors. SVM and RVM make no assumptions about the data, are able to find the global minimum of the objective function, and can provide near optimal performance. Moreover, the complexity of these techniques depends on the number of support (relevance) vectors, but not on the dimensionality of the input space. However, predictors based on these techniques provide too little insight as to the importance of variables to the predictor derived. The transparency is very important in some application areas, such as medical decision support or quality control, for example.

By contrast, classification and regression trees [6,7] are known for their transparency. However, decision trees are rather sensitive to small perturbations in the learning set. It has been demonstrated that this problem can be mitigated by applying bagging [8,9]. Random forests proposed by Breiman [10] and studied by Biau et al. [11] is a combination of the random subspace method proposed by Ho [12] and bagging. RF have been used for large variety of tasks, including: identification of DNA-binding proteins [13], segmentation of video objects [14], classification of hyper-spectral data [15,16], prediction of the vegetation type occurrence in Belgian lowland valley based on spatially distributed measurements of environmental conditions [17,18], to predict distributions of bird and mammal species characteristic to the eastern slopes of the central Andes [19], Czech language modeling in the lecture recognition task [20], diagnosing Alzheimer's disease based on single photon emission computed tomography (SPECT) data [21], genetic polymorphisms identification [22], prediction of long disordered regions in protein sequences [23], classification of agricultural practices based on Landsat satellite imagery [24], classification of aerial images [25], analysis of phenolic antioxidants in chemistry [26], recognition of handwritten digits [27], categorizing time-depth profiles of diving vertebrates [28], and many others.

2. Weak learners and random forests

Let us assume that given is a set of training data $\mathcal{X}_t = \{(\mathbf{x}_m, y_m), m = 1, \ldots, M\}$, where $\mathbf{x}_m$ is an input observation and y_m is a predictor output. A weak learner can be created using the training set $\mathcal{X}_t$. A weak learner is a predictor $f(\mathbf{x}, \mathcal{X}_t)$ having a low bias and a high variance [30]. By randomly sampling from the

* Corresponding author at: Intelligent Systems Laboratory, Halmstad University, Box 823, S 301 18 Halmstad, Sweden. Tel.: +46 35 167140.
E-mail addresses: antanas.verikas@hh.se (A. Verikas), adas.gelzinis@ktu.lt (A. Gelzinis), marija.bacauskiene@ktu.lt (M. Bacauskiene).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.08.011

set $\mathcal{X}_t$, a collection of weak learners $f(\mathbf{x},\mathcal{X}_t,\boldsymbol{\theta}_k)$ can be created, with $f(\mathbf{x},\mathcal{X}_t,\boldsymbol{\theta}_k)$ being the kth weak learner and $\boldsymbol{\theta}_k$ is the random vector selecting data points for the kth weak learner. By applying bootstrap sampling to generate $\boldsymbol{\theta}_k$, for example, about two-thirds of the data points are used by each weak learner. About one-third of the observations are out of the bootstrap sample or out-of-bag (OOB). The $\boldsymbol{\theta}_k$ are independent and identically distributed, i.i.d.

It can be shown that combining i.i.d. randomized weak learners into a committee by averaging, leaves the bias approximately unchanged while reduces the variance by a factor of $\overline{\rho}$—the mean value of the correlation between the weak learners [10]. Thus, if correlation and bias of i.i.d. randomized weak learners are kept low, a big reduction in test set error can be obtained.

RF is a committee of weak learners for solving prediction (both classification and regression) problems. In RF, a decision tree, i.e. CART (classification and regression trees), is used as a weak learner. When solving classification problems, the RF prediction is the un-weighted majority of class votes. Fig. 1 presents a general architecture of RF, where B is the number of trees in RF and k_1, k_2, k_B, and k are class labels. As the number of trees in RF increases, the test set error rates converge to a limit, meaning that there is no over-fitting in large RFs [10]. Low bias and low correlation are essential for accuracy. To get low bias, trees are grown to maximum depth. To achieve low correlation, randomization is applied:

i. Each tree of RF is grown on a bootstrap sample of the training set.
ii. When growing a tree, at each node, n variables are randomly selected out of the N available.
iii. Usually, $n \ll N$. It is suggested starting with $n = \lfloor \log_2(N)+1 \rfloor$ or $n = \sqrt{N}$ and then decreasing and increasing n until the minimum error for the OOB data set is obtained. At each node, only one variable, providing the best split, is used out of the n selected.

In RF, n is the only parameter to be selected experimentally. RF can handle thousands of variables of different types with many missing values. For a tree grown on a bootstrap data, the OOB data can be used as a test set for that tree. As the number of trees increases, RF provides an OOB data-based unbiased estimate of the test set error. OOB data are also used to estimate importance of variables. These two estimates (test set error estimate and variable importance) are very useful byproducts of RF.

2.1. Variable importance

There are four variable importance measures implemented in the RF software code [29,30]. Two measures, based on the Gini index of node impurity and classification accuracy of OOB data, are usually used.

Given a node t and estimated class probabilities $p(k|t), k = 1,\ldots,Q$, the Gini index is defined as [6]

$$G(t) = 1 - \sum_{k=1}^{Q} p^2(k|t) \tag{1}$$

where Q is the number of classes.

To calculate the Gini index based measure, at each node the decrease in the Gini index is calculated for variable x_j used to make the split. The Gini index-based variable importance measure $\overline{\Delta}_j$ is then given by the average decrease in the Gini index in the forest, where the variable x_j is used to split a node.

The classification accuracy-based estimator of variable importance prevails in various studies. The measure computes the mean decrease in classification accuracy of the OOB data. Having bootstrap samples $b=1,\ldots,B$, the importance measure $\overline{D}_j$ for variable x_j is calculated as follows:

i. Set $b=1$ and find the OOB data points $\mathcal{L}_b^{oob}$.
ii. Classify $\mathcal{L}_b^{oob}$ using the tree T_b and count the number of correct classifications, R_b^{oob}.
iii. For variables x_j, $j=1,\ldots,N$:
   (a) permute the values of x_j in $\mathcal{L}_b^{oob}$, the permutation results into $\mathcal{L}_{b,j}^{oob}$;
   (b) use T_b to classify $\mathcal{L}_{b,j}^{oob}$ and count the number of correct classifications, $R_{b,j}^{oob}$.
iv. Repeat Steps i–iii for $b=2,\ldots,B$.
v. The importance measure $\overline{D}_j$ for variable x_j is then given by

$$\overline{D}_j = \frac{1}{B}\sum_{b=1}^{B}(R_b^{oob} - R_{b,j}^{oob}) \tag{2}$$

vi. Compute the standard deviation s_j of the decrease in correct classifications and a z-score:

$$z_j = \frac{\overline{D}_j}{s_j/\sqrt{B}} \tag{3}$$

vii. Assuming Gaussian distribution, covert z_j to a significance value.

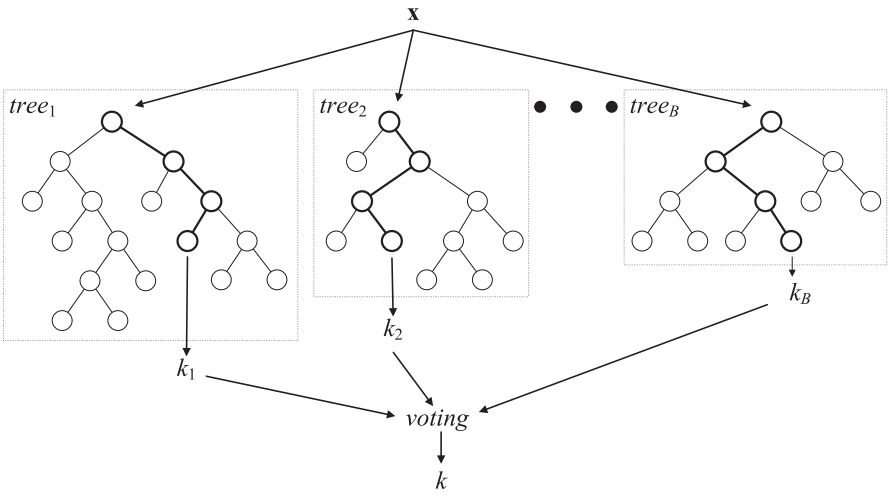

Fig. 1. A general architecture of a random forest.

2.2. Data proximity matrix

The proximity matrix available from RF is a very useful information source. To obtain the matrix, for each tree grown, the data are run down the tree. If two observations $\mathbf{x}_i$ and $\mathbf{x}_j$ occupy the same terminal node of the tree, $prox(i,j)$ is increased by one. When RF is grown, the proximities are divided by the number of trees in RF. Data proximities can be used to replace missing values, to find outliers and mislabeled data, to visualize data by applying the multidimensional scaling to the proximity matrix, for example.

3. Objectives of the study

Several researchers have found that RF error rates compare favorably to other predictors, including logistic regression (LR), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), k-NN, MLP, SVM, classification and regression trees (CART), naive Bayes [10,31–33], and other techniques. However, contra-examples can also be easily found [34], see Table 2. This is expected, since according to the No Free Lunch theorem [35], there is no single classifier model, which is best for all problems. In addition to a classification or regression model, RF also provides an estimate of variable importance for the model. Variable importance estimates are often used for data understanding. However, very few studies have been done to explore the adequacy of these estimates.

The objective of this work is twofold: to present a survey of RF applications and to study some aspects related to variable importance estimates available from RF. Four issues are considered in the survey, namely, prediction accuracy, exploitation of RF to study variable importance, exploitation of RF data proximity estimates, and modifications of RF. Large scale studies including RF we present in a separate section. Small scale studies are summarized in several tables.

Our studies regarding variable importance evaluations are focussed on the *consistency* and the *generality* of the variable importance estimates provided by the measures $\overline{D}_j$ and $\overline{\Delta}_j$. Consistency reflects the stability of feature rankings upon variations in parameters and procedure used to create RF. By generality we mean the effectiveness of a set of features corresponding to highest values of the measures $\overline{D}_j$ and $\overline{\Delta}_j$ in classifiers of various types, including RF.

We also study the complexity of several classification problems providing diverse results when solved using RF and SVM. To assess the complexity, several measures of decision boundary complexity studied in [36,37] are used. The aim of these tests was to get some insights into "suitability" of a problem at hand for RF-based classification.

4. Large scale studies including random forests

Meyer et al. investigated SVM performance and fulfilled a large scale comparison of 16 classification and nine regression techniques [38]. RF, SVM, MLP, and bagged ensembles of trees were among the techniques compared. Hyper-parameters of RF, SVM, and MLP were carefully selected. The classification benchmark was done on 21 data sets, while 12 data sets were used for the regression tests. Most of the data sets come from the UCI machine learning database. The performance was assessed using 10 times repeated 10-fold cross-validation. In the classification tasks, SVM always was ranked in the top 3, except for two data sets. However, SVM was outperformed in 10 out of 21 data sets. In the regression tasks, SVM was almost always in the top 3. However, SVM was outperformed in all occasions, except 2. The authors have found that MLP, RF, and projection pursuit regression are good alternatives to SVM regression and often yielded better performance than SVM.

Banfield et al. compared several decision tree ensemble creation techniques against standard bagging [39]. Boosting [40], random subspaces [12], three types of random forests (exploiting 1, 2, and $\lfloor \log_2(N)+1 \rfloor$ variables for node splitting) [10], and randomized C4.5 [9] have been used in the studies performed on 57 different data sets. For each technique, 1000 trees were used to make an ensemble. The classification accuracy was assessed using 10-fold and 5×2-fold [41] cross-validation. The statistical significance of differences in accuracy was assessed using the t-test on results of the 10-fold cross-validation and the F-test for the 5×2-fold cross-validation. In addition to these tests, the techniques were compared on the multiple data sets by comparing their average ranks [42]. The Friedman test [42] was applied to the ranks, to see if there is any statistically significant difference between the algorithms for the data sets. For 37 of the 57 data sets, none of the techniques showed statistically significant improvement over bagging. Boosting with 1000 trees and RF with $\lfloor \log_2(N)+1 \rfloor$ variables for node splitting appeared to the best techniques (most wins over bagging and never losing). Boosting with only 50 trees was less accurate. RF with only two variables showed the lowest average rank (the lower the rank the better the technique). The ranks of boosting with 1000 trees and RF with $\lfloor \log_2(N)+1 \rfloor$ were close to the lowest rank.

Diaz-Uriarte and Alvarez de Andres [43] studied the ability of RF to select a small set of genes retaining a high predictive accuracy. Gene selection was based on the OOB variable importance measure available from RF. The algorithm examined all forests resulting from iterative elimination of a fraction of genes (0.2 for example) used in the previous iteration. When tested on several microarray data sets with the number of genes ranging from 2000 to 9868, the algorithm was able to find very small sets of genes and retain predictive performance. The classification accuracy obtained from RF was comparable with that achieved by a linear SVM.

Inspired by work presented in [43], Statnikov et al. performed rigorous comparison of RF and SVM on 22 diagnostic and prognostic microarray data sets [34]. The data sets include 50–308 samples and 2000–24,188 variables. RFs made of 500, 1000, and 2000 trees were tested. The authors found that both on average and in the majority of the data sets SVM outperform RF, often by a large margin. The superiority of SVM was observed in both settings: without gene selection and with several gene selection techniques. Classifier performance was evaluated using receiver operating curves (ROC) and area under the receiver operating curves (AUC) for binary classification problems and the relative classifier information (RCI) [44] for multi-category tasks. RCI (entropy-based measure) evaluates how much the uncertainty in decision making is reduced by a classifier if compared to classifying based only on the a priori class probabilities. Both AUC and RCI are not sensitive to unbalanced distributions. The average performance of SVMs was 0.775 AUC and 0.860 RCI, while the average performance of RFs was 0.742 AUC and 0.803 RCI. The authors emphasize that the choice of RF parameters (the number of variables randomly selected at each node and the number of trees) creates large variation in RF performance. SVM was much less sensitive to the choice of hyper-parameters.

5. Small scale studies including random forests

5.1. Accuracy related studies

There are a large number of small scale studies aiming to compare the performance of several techniques on one or very few data sets. A large variety of application areas are explored.

Rigor of the comparisons varies greatly in different studies. We summarize results of these studies in two tables. Table 1 presents a survey of studies where random forests outperformed or performed on the same level as other techniques. Table 2 summarize studies where random forests were outperformed by other techniques.

Table 1
A survey of studies where random forests outperformed or performed on the same level as other techniques.

Area and techniques explored	Study details & results
Customer churn prediction (RF, SVM, MLP, DT) [45,46]	**Data**: from a Chinese bank, training and test sets each of 762 observations, two unbalanced classes, the smaller one contains about 5% of the data, 27 continuous and categorical variables, 12 removed as being irrelevant or with more than 30% of missing values. **Result**: RF $>$ all the other techniques in terms of accuracy and AUC. The statistical significance not tested
Customer churn prediction (RF, LR, SVM) [47,48]	**Data**: from a publishing company on subscription renewal, training and test sets of 45,000 observations, two classes, training set is balanced, the smaller class of the test contains 11% of the data, 32 variables both continuous and categorical [47]. **Result**: RF $>$ SVM and LR in terms of AUC (significant at 95% level) and accuracy
Credit card fraud detection (RF, SVM, kNN, LR, QDA, CART, NB) [49]	**Data**: two data sets from two banks of about 47,000 observations each, two classes, the smaller class contains about 30% of the data, 87 and 91 continuous as well as categorical variables, 70% of data for training and 30% for test. **Result**: RF $>$ all the other techniques
Discrimination of fish populations using parasites as biological tags (RF, MLP, LDA) [50]	**Data**: a set of 763 observations, $\approx$ evenly distributed in five classes—different regions in the North East Atlantic, 80% for training and 20% for test, 31 continuous variables. **Result**: RF $>$ MLP and LDA, in terms of average accuracy, precision (specificity), and recall (sensitivity) when assessed by 10-fold stratified cross-validation
Automatic e-mail filing into folders (multi-class problem) and spam e-mail filtering (two-class problem) (RF CART, SVM, NB) [51]	**Data**: multi-class: five users (data sets) with 545, 423, 888, 926, and 982 observations in 7, 6, 11, 19, and 6 very unbalanced classes, respectively; two-class: two data sets of 1099 and 2893 observations, with about 44% and 17% in the smaller class. IG technique [52] was used to select variables for text description, 256 continuous variables. RF of 10 trees, 9 variables to split a node. **Result**: RF $>$ all the other techniques, in terms of accuracy, precision, and recall when assessed by 10-fold stratifiedcross-validation
Aircraft engine fault diagnosis (RF, MLP, SVM) [53]	**Data**: a set of 19,635 observations in 1 normal and 6 classes of faults, 2805 points in each class, 11 continuous variables, seven binary classifiers one-against-all were used. **Result**: RF of 500 trees and 3 features to split a node $>$ SVM, and CART, in terms of accuracy, false positive, and false negative rates when assessed by 5-fold cross-validation, RF $\approx$ MLP
Bacterial species identification (RF, SVM, MLP) [54]	**Data**: a set of 3012 observations represented by 105 continuous variables and coming from 3 classes (961, 378, and 1673 observations) or 213 classes. **Result**: RF of 1000–4000 trees (assessed on OOB data) $>$ SVM and MLP in terms of AUC, sensitivity, and precision; performance of SVM and MLP was assessed by 10-fold stratified cross-validation
Recognition of face images (RF, SVM) [55].	**Data**: a set of 2414 images represented by 3584 or 3342 variables (gray values of pixels of resized images) and evenly distributed in 38 classes. Training and test sets were of the same size. **Result**: accuracy of RF $\gg$ SVM. Best RF consists of 500 trees and generated using 100 variables to split a node
Diagnosis of induction motor faults (RF, *k*-NN, SVM, CART) [56]	**Data**: a set of 180 training samples, 90 test samples evenly distributed in 9 classes (normal an 8 fault classes), 63 continuous variables. **Result**: RF $\gg$ than all the other techniques in terms of average accuracy. Best RF of 1200 trees, 1 variable to split a node
Classification of protein-localization patterns within florescent microscope images (RF, SVM, DT) [57]	**Data**: a set of 862 images approximately evenly distributed in 10 classes, 180 continuous variables. **Result**: average accuracy of RF $\gg$ than the other two techniques. Best were relatively small forests (56–86 trees) created using a large number of variables (43–164) to split a node, 20/80, 50/50, and 80/20% were training and test set proportions explored
Discrimination between acidic and alkaline enzymes (RF, SVM, MLP, NB, *k*-NN, DT, Bayes net, boosted ensemble, bagged ensemble) [58]	**Data**: a strictly screened two-class data set of 105 acidic enzymes and 111 alkaline enzymes, 60 continuous variables characterizing the secondary structure of amino acid compositions **Result**: RF $>$ all the other techniques in terms of accuracy, AUC, selectivity, and specificity, assessed by 10-fold cross-validation
Forest areas prediction (RF, CART, MLP) [59]	**Data**: a two-class (presence and absence of given species) balanced data set of 16,510 observations, 14 continuous variables. **Result**: RF $>$ MLP and CART in terms of AUC, when 2/3 of the data were used for learning and 1/3 for test. Best RF of 500 trees, 6 variables to split a node
Prediction of current and future tree distributions (MARS, CART, BT, RF consisting of 1000 trees) [60]	**Data**: a regression (predict the distribution of four tree species) data set of 9782 study cells characterized by 36 continuous variables. **Result**: RF and BT $>$ CART and MARS, assessed by Kappa statistics, which measures the level of agreement between the distribution of categories in map *A* and map *B*
Distinguishing between the chemical spaces of metabolites and non-metabolites (RF, CART, CPNN) [61]	**Data**: a two-class set of 6409 compounds, including 1811 metabolites and 4598 non-metabolites, 25 continuous variables. **Result**: RF of 1000 trees $>$ CART and CPNN in terms of average accuracy, assessed by 10-fold cross-validation for CART and on OOB data for RF
Categorization of cancer cases (RF, multinomial logit) [62]	**Data**: a four-class (4 categories of cancer diagnosis) data set, with 54%, 34%, 7%, and 5% class proportions, collected from 5608 subjects, 63 variables, many of which categorical. **Result**: RF of 500 trees $\gg$ multinomial logit, in terms of average accuracy assessed by 8-fold stratified cross-validation
Predicting the condition of vegetation across the state of Victoria of Australia (RF, CART, MTRT, BT) [63]	**Data**: a regression task, 16,967 observations of ecological and remote-sensed data, 40 continuous variables. **Result**: RF $>$ the other techniques. The difference in accuracy between RF and BT was not statistically significant according to the 10 times repeated 10-fold cross-validation and the corrected Friedman test [42]—test for the average ranks of the algorithms
Distinguishing between the diseased and non-diseased eyes (RF, BT, boosted C4.5 trees) [64]	**Data**: a three-class (normal and 2 classes of diseases) set of 254 (119+36+99) observations, 15-121 continuous variables given by Zernike or pseudo-Zernike polynomials. **Result**: RF $\approx$ the other techniques in terms of accuracy assessed by 10-fold cross-validation
Predicting partial defection by behaviorally loyal clients (RF, ARDNN, LR) [65]	**Data**: two-class data collected from 32,371 customers, the smaller class of 25%, 61 variables—several categorical, but mainly continuous. **Result**: RF of 5000 trees and 8 variables to split a node $\approx$ the other techniques in terms of accuracy and AUC assessed on a hold-out set of 50% of the data

DT: decision tree; NB: naive Bayes; CCNN: cascade correlation neural network; $>$: more accurate; $\approx$: approximately the same performance; NCS: nearest shrunken centroids [66]; MARS: multivariate adaptive regression splines; CPNN: counterpropagation neural network; MTRT: multi-target regression trees; ARDNN: automatic relevance determination neural network [67,68]; and BT: bagged trees.

Table 2
A survey of small scale studies where random forests were outperformed by other techniques.

Area and techniques explored	Study details & results
Recognition of five types of underwater plankton images (SVM, RF, C4.5, CCNN, BT) [69]	**Data**: a 5-class problem (unbalanced classes) with 1285 binary images and a 6-class problem (balanced classes) with 6000 binary images, 29 continuous variables characterizing shape; **Result**: SVM > all the other techniques when assessed by 10-fold cross validation and a paired *t*-test at the 95% confidence level (for the 6-class problem the difference between SVM and RF was not significant). BT and RF consist of 100 trees and the default number of variables to split a node
Face recognition using Haar features (RF, SVM) [70]	**Data**: a set of 464 images from 6 classes and a set of 40 images from 5 balanced classes, 25 continuous features. **Result**: SVM > RF of 10 trees, in terms of recognition accuracy, the way used to assess the accuracy not provided
Hyper-spectral remote sensing and geographic data classification (RF, ensemble of SVMs) [71]	**Data**: a set of 2019 observations from 7 data channels (4 landsat channels, 1 elevation, 1 slope, and 1 aspect), 50% for training, 10 ground-cover classes, continuous variables. **Result**: Ensembles of SVMs > RF in terms of average accuracy
Spam email detection (CBART, LR, SVM, CART, NB, MLP, RF ranging from 10 to 500 trees) [72]	**Data**: a two-class problem, three data sets of 2893 (spam 16.6%), 1099 (spam 43.8%), and 4601 (spam 39.4%) observations, continuous variables—256 in the first two sets and 57 in the third. **Result**: CBART > all the other techniques, in terms of average error rate, confirmed by 10-fold cross-validation. The difference between CBART and RF not significant. RF provided the highest AUC
Identification of strawberry cultivars from mass spectrometry data (RF, PDA, DPLS) [73]	**Data**: a set of 233 observations, 9 classes (21 observations in the smallest class and 30 in the largest), 231 continuous variables. **Result**: Both techniques > RF with 15 variables to split a node, in terms of average accuracy assessed by LOO
Categorization of near infrared spectra of red grape homogenates (RF, PDA, MARS) [74]	**Data**: a 3-class problem, a set of 284 spectra (original and transformed) made of 1024 sample wavelengths, 1024 continuous variables. **Result**: PDA and MARS > RF in terms of accuracy. The number of trees in RF, the number variables used to split a node, and the way to assess the accuracy not provided
Network intrusion detection (RF, ANFIS [75], LGP) [76]	**Data**: a 5-class problem, a set of 5092 (1000+500+3002+27+563) training and 6890 test observations, 41 variables (categorical and continuous). **Result**: ANFIS and LGP > RF generated by using 3 variables (a fixed number) to split a node, in terms of accuracy. The number of trees in RF not provided
The task of diagnosing scrapie (disease) in sheep (18 classifiers including LDA, NB, SVM, trees, BT, AdaBoost ensembles, RF) [77]	**Data**: a two-class data set of 3113 observations (83% in the scrapie class), 90% for training and 10% for test, 100 random splits, 125 binary variables. **Result**: Pruned J48 trees > RF and AdaBoost with unpruned J48 (significant difference at 95% confidence level). Ensembles, including RF, were of 50 trees, the number of variables used to split a node not provided
Rock glacier detection based on terrain analysis and multispectral remote sensing data (LR, GLM, GAM, LDA, stabilized LDA, penalized LDA, SVM, BT, and RF) [78]	**Data**: a two-class problem, a data set of 2071 observations (only 86 in the rock glacier class), 26 continuous variables. False-positive rates at sensitivity of 70% were assessed by 100-repeated 5-fold cross-validation and compared. **Result**: PLDA, GAM, GLM > RF of 200 trees, the number of variables used to split a node not provided. Data over-fitting by RF was emphasized. A larger number of trees, perhaps, could mitigate the problem
Customer loyalty prediction (RF, MLR, ARDNN) [79]	**Data**: a regression task, a set of 878 observations, 35 variables several of them categorical. **Result**: MLR with 4 variables > RF of 5000 trees and ARDNN (both using all 35 variables), in terms of determination coefficient R^2 and MSE, assessed by the leave-one-out technique. MLR with 35 variables < RF with 35 variables
Predicting the chromatographic retention of basic drugs (RF, CART, TreeBoost, PLS, GA-MLR) [80]	**Data**: a regression task, a set of 83 drugs, 1272 variables (molecular descriptors). **Result**: Optimal RF of 600 trees, 382 variables to split a node. GA-MLR > RF, in terms of determination coefficient R^2, assessed by LOO or by using the OOB set (for bagged and boosted models)

CBART: classification Bayesian additive regression trees; PDA: penalized discriminant analysis; PLS: partial least squares; DPLS: discriminant partial least squares; ANFIS: adaptive-network-based fuzzy inference system; LGP: linear genetic programming; GLM: generalized linear models; GAM: generalized additive models; MLR: multiple linear regression; GA-MLR: genetic algorithms-based multiple linear regression; and LOO: leave-one-out.

It is interesting to note that in some cases, RF outperformed all other techniques used for comparisons, including SVM, by a large margin. This was observed for problems characterized with both relatively large and small number of variables. The highest RF accuracy was sometimes achieved using the number of variables for node splitting, which deviated significantly, to both directions, from the approximate number, $\lfloor \log_2(N)+1 \rfloor$, suggested by Breiman.

We found fewer examples, where RF were outperformed by other techniques. However, the variety of techniques outperforming RF was large, including multiple linear regression. It is worth mentioning that some comparisons may be biased, in the way that when RF was applied for the first time in a specific application area, much effort was put to find the optimal forest size and the optimal number of variables used for node splitting, while selection of parameters for other techniques was not so diligent.

5.2. Studies related to variable importance

In spite of the fact that variable importance measures available from RF are widely used for data exploration, very few attempts were made to study behavior of the measures. Below, we briefly summarize the results of these studies.

Strobl et al. [81] studied the behavior of $\overline{\Delta}_j$, $\overline{D}_j$, and the variable selection frequency-based measure on a set of artificial data with one continuous and four categorical variables. The authors came to a conclusion that the measures may give misleading results when variables are of different types or the number of levels differ in different categorical variables. The Gini index-based creation of classification trees is seen as the underlying mechanism behind this deficiency. Gini index favors continuous variables or categorical variables with a large number of levels.

Reif et al. [82] applied $\overline{D}_j$ of RF to identify relevant features in sets of high-dimensional gene and protein data containing both categorical and continuous variables. A 100 data sets containing up to 1550 variables were generated for the studies. Random forests of 10,000 trees were grown. In contrast to the findings of Strobl et al. [81], the authors make a conclusion that random forests are robust to noise and are able to identify relevant variables of either type in high-dimensional data containing variables measured on multiple scales.

Archer and Kimes [83] studied both $\overline{\Delta}_j$ and $\overline{D}_j$ variable importance measures using RF consisting of 2500 trees and multivariate normal data generated by a linear regression model. The degree of correlation between variables and the degree of association of variables with the output were varied in the study. The authors came to a conclusion that random forests are attractive in settings with a large number of correlated variables. The authors indicate one drawback of the measure $\overline{D}_j$ in the case of small data sets–a rather low resolution, which is approximately equal to $3/M$, with M being the number of observations.

In most of the studies surveyed in this article, evaluations of variable importance measures are directly used for data exploration, understanding, and interpretation. In some cases, such interpretations can be a point of contention, since correlation between variable rankings obtained using RF created under different conditions can be rather low. For example, Okun and Priisalu [84] used RF for gene expression-based cancer classification and demonstrated that two forests, created using different number of variables to split a node, may exhibit similar accuracy on the same data set, but correlation between the lists of variables ranked according to RF variable importance can be weak. Two very small data sets with 2000 and 822 genes (variables), and Gini index-based variable ranking were used in the experiment. The number of trees in the RF was set to 500. Table 3 summarizes studies exploiting RF-based variable importance evaluations.

5.3. Exploiting data proximity

Wang et al. [95] and Yang et al. [96] proposed an interesting data classification approach based on the data proximity matrix. Classifier design and data classification proceeds as follows:

i. Having a set $\mathcal{X}^{M-1}=\{\mathbf{x}_1,\ldots,\mathbf{x}_{M-1}\}$ of training observations, create an RF.
ii. Given an unknown observation $\mathbf{x}_M$, run a set $\mathcal{X}^{M}=\{\mathbf{x}_1,\ldots,\mathbf{x}_{M-1},\mathbf{x}_M\}$ down the RF and create a proximity matrix $prox(i,j)$ of size $M\times M$.
iii. Initialize the set of prediction classes $\Gamma^{\varepsilon,\tau_M}=\emptyset$.
iv. For $j=1$ to Q (the number of classes) do:
  (a) assume that j is the label of $\mathbf{x}_M$;
  (b) calculate scores $\alpha_1,\alpha_2,\ldots,\alpha_{M-1},\alpha_M$ for the set $\mathcal{X}^M$ according to the following rule:

$$\alpha_i=\frac{\sum_{k=1}^{K} prox_{ik}^{-y_i}}{\sum_{k=1}^{K} prox_{ik}^{y_i}},\quad i=1,\ldots,M \tag{4}$$

  where y_i is the class label of the observation $\mathbf{x}_i$, $prox_{ik}^{-y_i}$ is the jth largest proximity between $\mathbf{x}_i$ and observations having different labels than y_i, $prox_{ik}^{y_i}$ is the jth largest proximity between $\mathbf{x}_i$ and observations having label y_i, and K is a user defined parameter;
  (c) calculate p_y (confidence value) for the assumption $\{\mathbf{x}_M \in \texttt{class}\ j\}$ according to the following equation:

$$p_y=\frac{|\{i=1,\ldots,M:\alpha_i>\alpha_M\}|+\tau_M|\{i=1,\ldots,M:\alpha_i=\alpha_M\}|}{M} \tag{5}$$

  where τ_M is a uniformly distributed random number in [0,1].
  (d) if $p_y>\varepsilon$ (significance level), $\Gamma^{\varepsilon,\tau_M}:=\Gamma^{\varepsilon,\tau_M}\cup j$.
v. Output the classification result, $\Gamma^{\varepsilon,\tau_M}$.

Thus a set of possible classes and confidence values are obtained as a result of the classification. The authors applied the classifier in the chronic gastritis diagnosis task, characterized by 54 binary symptoms, using a data set collected from 709 subjects [95]. The number of trees included into the forest ranged from 1000 to 10,000. A 10-fold cross-validation was used to assess the classification accuracy. The technique significantly outperformed SVM, ordinary RF, and the Bayesian classifier.

Wiseman et al., aiming to visually explore data characterizing benign and malignant thyroid tumors, applied multidimensional scaling (MDS) to the data collected in the RF proximity matrix [97]. The MDS plot showed good separation between the benign and malignant cases. Spearman correlation between the Gini index-based variable importance and the variable importance evaluated with the Mann-Whitney U test [98] for the top 35 variables (out of 57) was highly significant ($\rho=0.764, p<0.0001$).

Wu et al. [99] used RF to distinguish good chip delay test signatures from bad chip delay test signatures as well as to detect outliers in the post-silicon delay test data. The RF proximity matrix was used to detect outliers. Gislason et al. [100] applied RF for classification of remote sensing and geographic data, where outlier detection was also based on analysis of the RF proximity matrix.

5.4. Modifications

There were several attempts to modify RF, aiming to increase the prediction accuracy, to reduce the number of trees, or to make an online implementation. Gray and Fan [101] suggested using genetic algorithms (GA) for designing classification forests. Starting from randomly generated trees, genetic operations are applied to evolve a forest with a low number of trees. Linear combinations of two or three variables can be utilized for node splitting. Experimental tests on real data sets have shown that the forests evolved by GA are not as accurate as RF. Rodriguez et al. [102,103] proposed using rotation forests instead of RF. To create one of L base classifiers in a rotation forest, a feature set is randomly split into K (parameter of the technique) subsets, a bootstrap sample set is selected for each subset and PCA is applied. All PCs are retained and used as features to train the classifier. It was found that rotation forests are more accurate than RF and AdaBoost.

Tsymbal et al. [104] proposed using adaptive RF. Trees to be included into RF are dynamically selected for each observation. The proximity matrix is used to determine the trees to be included into RF. Voting and weighted voting aggregation schemes were explored. The local accuracy estimates are used to determine weights in the case of weighted voting. Both aggregation schemes showed rather similar performance and adaptive RFs were statistically significantly more accurate than the ordinary ones. Tripoliti et al. [105] used RF for diagnosing Alzheimer's disease based on fMRI data. In addition to majority voting, different weighted voting schemes were explored. Weighted voting based on a distance between the unknown and training observations provided the best performance.

Leshem and Ritov suggested using RF as a weak learner in the AdaBoost algorithm [106]. The technique was applied to a traffic flow prediction task. However, no comparisons were performed with an ordinary RF. Osman [107] proposed a technique for online RF design with incremental feature selection and demonstrated

Table 3
A survey of studies exploiting random forests-based variable importance evaluations.

Techniques applied	Task & results
$\overline{D}_j$ of RF applied to a set of 13 categorical variables [85]	**Task**: ranking the importance of drivers, vehicles, and environment characteristics on crash avoidance maneuvers of drivers; studied in a binary classification task—evasive actions or no evasive actions of drivers. **Result**: rankings are useful for data exploration in 3 types of collisions studied using three unbalanced data sets: rear-end collisions (10,867 observations), head-on collisions (1105), and angle collisions (5878)
Based on $\overline{D}_j$ of RF applied to a set of 19 continuous and categorical variables [86]	**Task**: discovering the most important factors for tooth loss/survival; studied in a binary classification task. **Result**: obtained variable importance rankings make sense from a clinical standpoint; data on 6463 teeth (from 355 subjects) were used, tooth loss observed for 17%
$\overline{\Delta}_j$ of RF, coefficients of MLR, and α of ARDNN (hyper-parameter for each input variable reflecting variable importance) [79]	**Task**: Variable importance in the customer loyalty prediction task, 35 continuous and categorical variables, see Table 2. **Result**: low correlation between variable importance values obtained by the different techniques: the correlation coefficient $\rho=0.0886$ between the MLR and RF, and $\rho=0.1205$ between the ARDNN and RF importance values
$\overline{\Delta}_j$, $\overline{D}_j$ of RF applied to a set of 25 continuous variables [61]	**Task**: variable importance for distinguishing between the chemical spaces of metabolites and non-metabolites, see Table 1. **Result**: "the importance of descriptors was to a large extent in accordance with the rules extracted by a single classification tree". Moderate agrement between $\overline{\Delta}_j$ and $\overline{D}_j$ was observed
$\overline{D}_j$ of RF, multinomial logit-based backward and forward variable selection [62]	**Task**: importance of 63 variables, many of which categorical, for categorizing cancer cases, see Table 1. **Result**: seven of the top-ten variables were the same in both lists; the rank of importance largely agreed with what could be expected based on experience and rationale considerations
$\overline{D}_j$ of RF, α of ARDNN [65]	**Task**: variable importance to predict partial defection by behaviorally loyal clients, 61 variables, several categorical, see Table 1. **Result**: the Spearman correlation coefficient $\rho=0.313$ between the rankings provided by the techniques. Six of the top-ten variables were the same in both lists
$\overline{D}_j$ of RF [87,88]	**Task**: variable selection in intrusion detection using a large balanced data set with 41 categorical and continuous variables. **Result**: elimination of 7 least important variables allowed improving the detection accuracy. Selection result highly depends on a training set
$\overline{\Delta}_j$ of RF [89]	**Task**: variable selection (from a set of 3629 wavelengthes) using a set of 641 observations for hierarchical binary classification in infrared spectroscopy. **Result**: similarity of wavelength regions selected by RF and regions selected by other authors using SVM and MLP was observed
$\overline{D}_j$ of RF [90]	**Task**: variable selection for classification of mass spectrometry profiles in a prostate cancer diagnosis task, 322 observations, thousands of variables. **Result**: iterative procedure to take into account relations between variables, at each iteration 30% of the most important variables were chosen
$\overline{\Delta}_j$ of RF [91]	**Task**: variable importance evaluations to study issues affecting customer retention and profitability, regression and binary classification tasks, 100,000 observations, 30 continuous and categorical variables. **Result**: variable importance evaluations were found being useful for exploring the data
$\overline{D}_j$ of RF [92]	**Task**: identification of the most important identity markers when studying income redistribution preferences across identity groups, 10 categorical variables, 13,024 observations. **Result**: variable importance evaluations were found being useful for exploring the data
$\overline{D}_j$ of RF, Fisher's Exact test [93]	**Task**: identification of a small number of risk-associated single nucleotide polymorphisms (SNPs) among large number of un-associated SNPs in complex disease models, two data sets of 1000 observations with 100 and 1000 ordinal variables. **Result**: when risk SNPs interact, $\overline{D}_j$ significantly outperformed Fisher's Exact test as an SNPs screening tool
$\overline{D}_j$ of RF [94]	**Task**: to select features from a set of 304 continuous variables for electroencephalogram classification, 98,160 observations. One-step feature selection was compared with recursive feature elimination. **Result**: classification accuracy obtained from the recursive approach was much higher than using the one-step procedure. The error rate decreased from 22.1%, when using all variables, to 12.3%, when using 24 recursively selected variables
$\overline{\Delta}_j$ of RF [84]	**Task**: a gene expression-based cancer detection task, a data set of 74 observations and 822 variables (genes) and a set of 62 observations and 2000 genes. **Result**: correlation between the variable importance rankings obtained from RF, created using different number of variables to split a node, can be weak

that the online RF attains approximately the same performance as the off-line counterpart.

6. New tests regarding variable importance measures

We used four public databases for these tests: *Waveform*, *Satimage*, *Thyroid*, and *Wisconsin Diagnostic Breast Cancer—WDBC* (http://archive.ics.uci.edu/ml/).

There are three classes of waves in the *Waveform* database [6] with equal number of instances in each class, where each class is generated from a combination of 2 of 3 "base" waves. All of 40 variables used include noise with mean 0 and variance 1. The later 19 variables are all noise variables with mean 0 and variance 1. The *Satimage* data set contains 6435 observations categorized into six classes and represented by 36 features. The task in *Thyroid* medical database is to decide whether the patient's thyroid has over-function, normal function, or under-function. It is a 21-dimensional, three-class database containing 7200 examples. The class frequencies are 5.1%, 92.6%, and 2.3%, respectively. There are 30 real-valued features in the two-class *WDBC* problem. The features are computed from a digitized image of a fine needle aspirate of a breast mass and describe characteristics of the cell nuclei present in the image. There are 569 instances, 357 benign and 212 malignant.

The classification accuracy presented in this section is the average accuracy computed from 50 trials. In each trial, a data set was randomly split into learning—70% and test—30% subsets. The

statistical significance of difference in accuracy was assessed by using a paired t-test at the 95% significance level.

We studied the *consistency* and *generality* of the variable importance evaluations provided by $\overline{D}_j$ and $\overline{\Delta}_j$. In a finite sample case, unimportant features usually degrade the performance of a classifier. However, this sensitivity is different for different classifiers. It is well known that degradation in performance often accompanies addition of new, unimportant, equally weighted features in k-NN classifiers. However, MLP and SVM, suffer from the 'curse of dimensionality' to a significantly less extent than k-NN. The dependency of the test set classification accuracy on the number of features used by SVM and k-NN classifiers to classify the Satimage data set, shown in Fig. 2, illustrates such behavior.

To create the plots, features were selected by the k-NN-based forward selection and by taking features corresponding to the highest values of the $\overline{D}_j$ and $\overline{\Delta}_j$ measures. Values of the measures were computed using RFs with the optimal number of trees and the optimal number of features used for node splitting. Selected features were then used in k-NN and SVM classifiers. SVM was trained with three different sets of features: selected by the k-NN, and according to the $\overline{D}_j$ and $\overline{\Delta}_j$ measures. Hyper-parameters of the SVM (regularization constant and width of the Gaussian kernel) were carefully selected by cross-validation. A mode can be clearly observed in the k-NN curve. However, there is no clear mode in the SVM plots. Therefore, we focused on k-NN and RF classifiers, when studying the generality of variable importance evaluations provided by the $\overline{D}_j$ and $\overline{\Delta}_j$.

6.1. Results of the consistency studies

We assess the consistency by computing the width of the confidence interval for the measures and by evaluating the average Spearman correlation coefficients, $\overline{\rho}_{D_j}$ and $\overline{\rho}_{\Delta_j}$, between rankings of variables obtained in different experiments controlled by various values of parameters governing the RF designing process.

6.1.1. Variance of the variable importance estimates

First, we studied the consistency using the artificial 40-dimensional Waveform data. Previous tests of various variable selection techniques on this data set have shown that correct identification of all the noise variables is not an easy task [108,109]. However, this was not the case for the $\overline{D}_j$ and $\overline{\Delta}_j$ rankings-based variable selection. Fig. 3 presents characteristic plots of average values of the $\overline{D}_j$ and $\overline{\Delta}_j$ measures for the variable set. The plots were obtained by varying the number of trees in the forest from 900 to 1100, using 5000 observations to create RF, and using $\lfloor \log_2(N)+1 \rfloor$ variables to split a node. The figure also provides the 95% confidence intervals for the measures, which are very narrow and barely seen in the plots.

One can easily notice a clear difference between the variable importance values computed for the pure noise and the relevant variables. Observe that variables $\langle 1,2,20,21 \rangle$ are also almost equivalent to the noise variables. The measure $\overline{D}_j$ provides a higher contrast between the pure noise and the relevant variables than $\overline{\Delta}_j$. The variance of the measures is very low. The variance of the measures remained low even when using a relatively small number of trees in RF and/or a small number of observations. Fig. 4 (*left*) presents $\overline{D}_j$ computed by varying the number of trees in RF from 900 to 1100 and using 500 observations, while Fig. 4 (*right*) plots $\overline{D}_j$ computed by varying the number of trees from 50 to 150 and using 500 observations. As can be seen, the width of the confidence intervals remains very small even for small number of trees and observations. The same behavior of the measures was observed for the other data sets.

Next, we studied variability of the variable importance values by varying the number of variables used to split a node when generating RF. Fig. 5 plots the average values of $\overline{D}_j$ along with the 95% confidence intervals computed for variables of the Satimage (*left*) and the Waveform (*right*) data sets when varying the number of variables used to split a node. The number of features was varied from 4 to 20 for the Satimage data set and from 6 to 25 for the Waveform data set. A 100 trees and 500 observations were used to create RFs.

As one can see, the confidence intervals are much wider and, for many variables of the Satimage data set, the difference in variable importance cannot be deemed as being statistically significant. The same variability degree of the measures was also observed for the other two data sets. Thus, variable importance rankings may, to a great extent, depend on the number of variables used to split a node when designing RF.

6.1.2. Correlation of the variable importance rankings

We studied correlation between rankings obtained by varying the number of trees in RF and the number of features used to split a node. Fig. 6 plots the average Spearman correlation coefficient

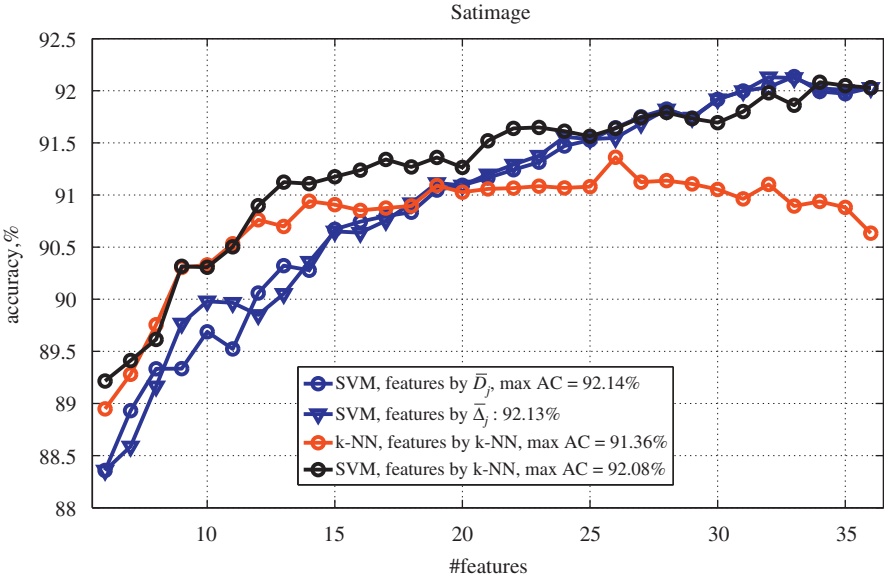

Fig. 2. The dependency of the test set classification accuracy of the SVM and k-NN classifiers on the number of variables used for the Satimage data set.

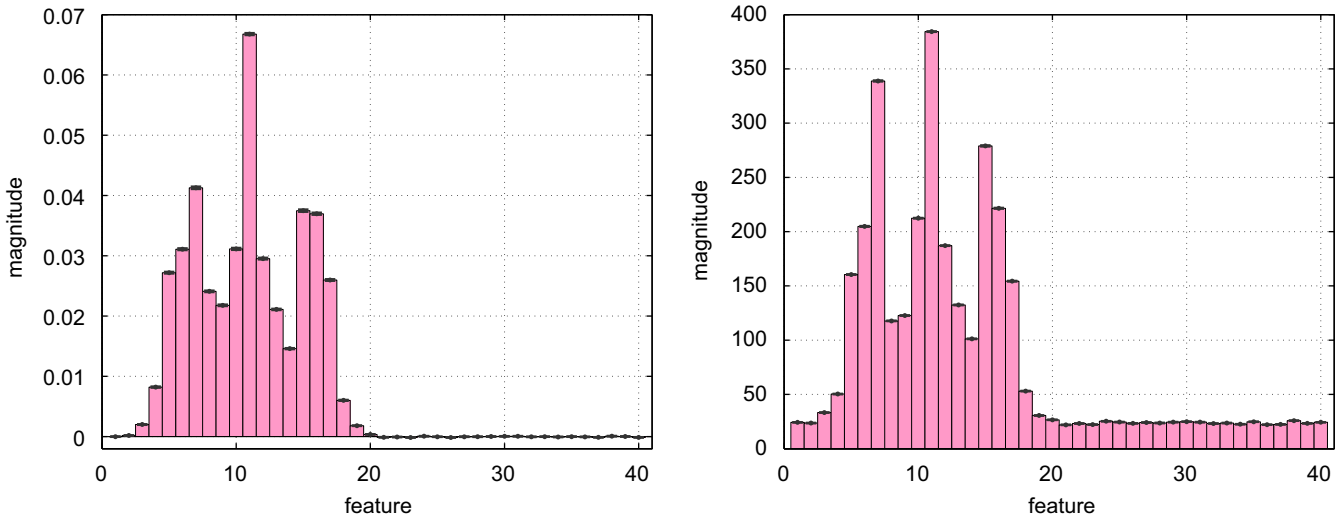

Fig. 3. The average values of $\overline{D}_j$ (*left*) and $\overline{A}_j$ (*right*) along with the 95% confidence intervals computed for variables of the Waveform data set using a large number of observations and a large number of trees.

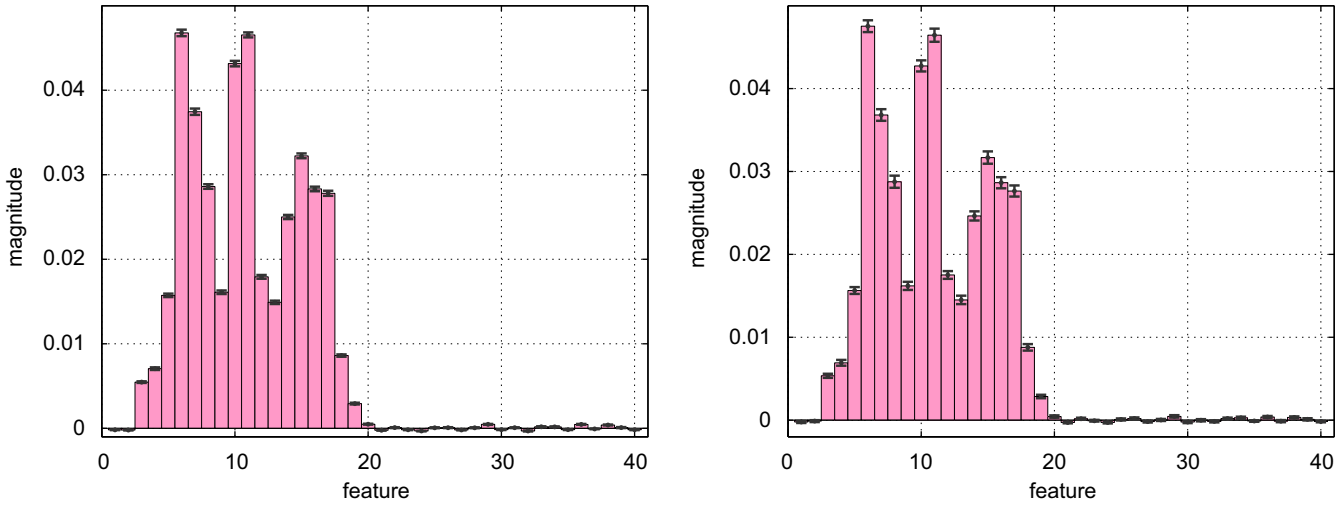

Fig. 4. The average value of $\overline{D}_j$ along with the 95% confidence intervals computed for variables of the waveform data set using a small number of observations (*left*) and a small number of both trees and observations (*right*).

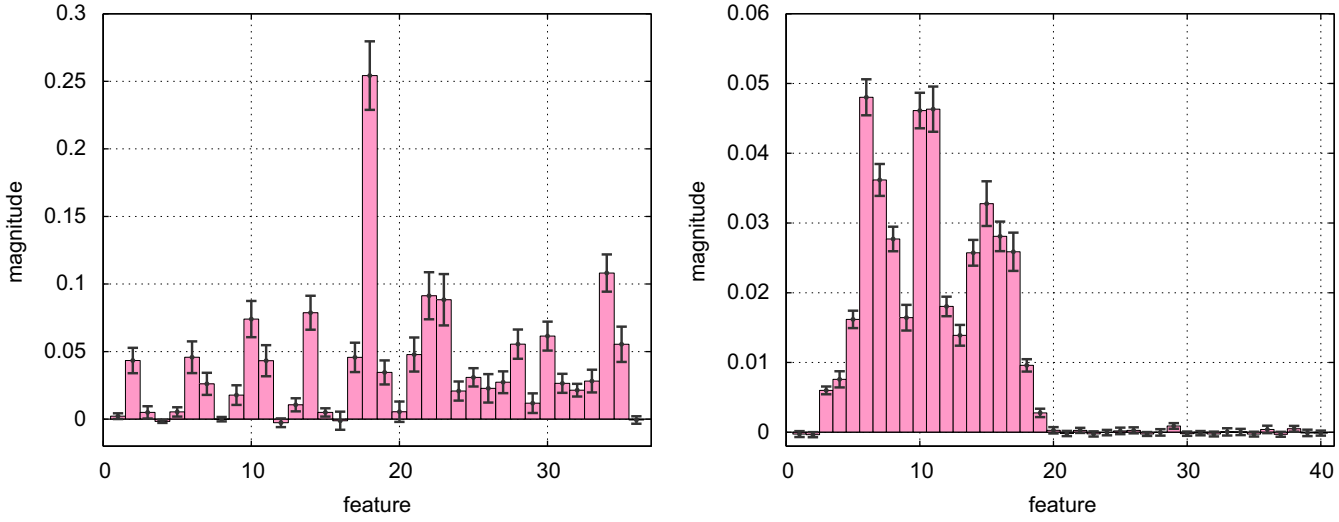

Fig. 5. The average values of $\overline{D}_j$ along with the 95% confidence intervals computed for variables of the Satimage (*left*) and the Waveform (*right*) data sets when varying the number of variables used to split a node.

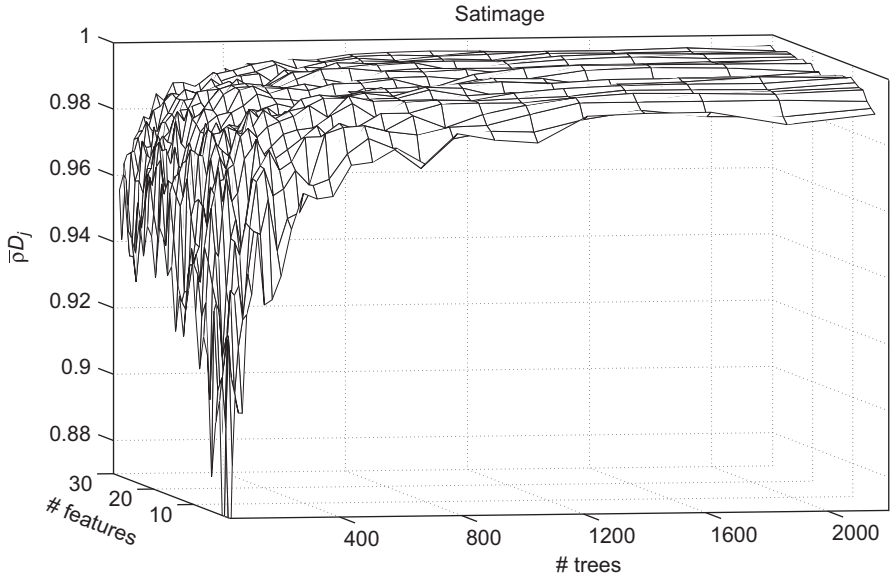

Fig. 6. The average Spearman correlation coefficient $\overline{\rho}_{D_j}$ as a function of the number of trees in RF and the number of features used to split a node.

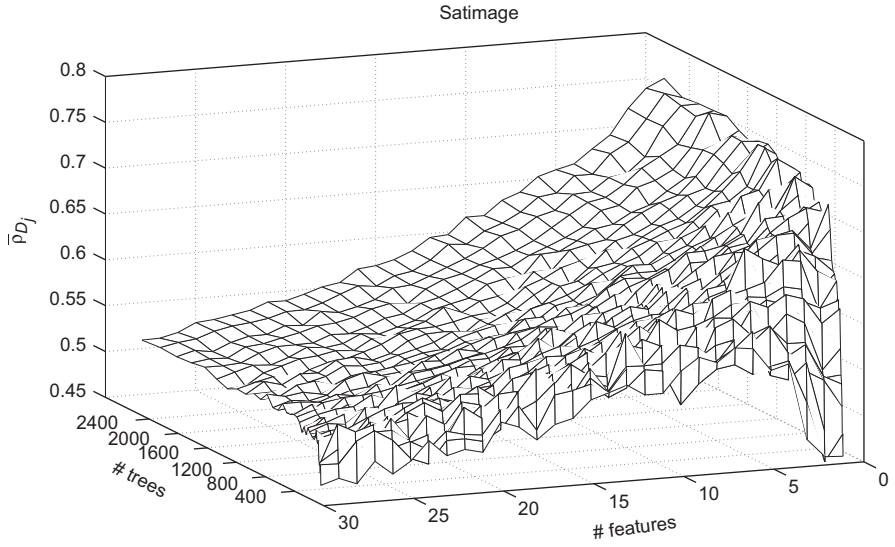

Fig. 7. The average Spearman correlation coefficient $\overline{\rho}_{D_j}$ for the Satimage data set as a function of the number of trees in RF and the number of variables used to split a node, when averaging over 20 different runs performed with the same number of trees and variables. A new training set of 500 observations was randomly selected for each run.

$\overline{\rho}_{D_j}$ computed for the Satimage data set as a function of the number of trees in RF and the number of variables used to split a node when generating the RF. The plot was created using all observations of the data set (6435) in the RF designing process. $\overline{\rho}_{D_j}$ values were obtained by averaging locally ρ_{D_j} values computed for a pair of variable lists generated by a pair of random forests with a different number of trees and/or variables.

As one can see from Fig. 6, for RF consisting of a large number of trees created using a large number of observations (6435), correlation between variable rankings is very high. However, the correlation decreases significantly in the case of RF consisting of a relatively small number of trees. For example, using RFs of 100 trees, which is a popular number if various applications of RF, a significant decrease of correlation is observed. To create the plot shown in Fig. 6, correlations were computed between for a pair of variable lists generated by a pair of random forests with a very similar number of trees and/or variables (adjacent numbers). One can expect lower correlations, when the numbers of variables used to split a node in two RF differ more. Relatively wide confidence intervals shown in Fig. 5 indicate this.

Fig. 7 explores the stability of D_j-based variable ranking from a slightly different prospective. The figure presents $\overline{\rho}_{D_j}$ for the Satimage data set as a function of the number of trees in RF and the number of variables used to split a node, when correlating rankings and averaging the correlations obtained in 20 different runs performed with the same number of trees and variables. For each run, 500 observations were randomly selected out of 6435 available. As can be seen from Fig. 7, a rather low correlation between rankings is obtained even for a large number of trees. The decrease in correlation, however, is mainly due to randomness in the data set. The plot presented in Fig. 8 substantiates the fact. The plot was created in the same way as that shown in Fig. 7, except that all observations from the WDBC data set were used in each run. Observe that there are only 569 observations in the

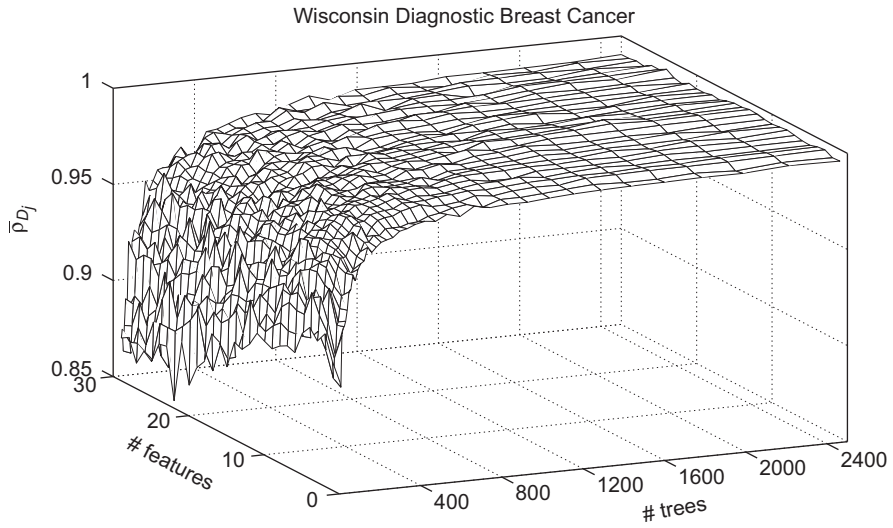

Fig. 8. The average Spearman correlation coefficient $\overline{\rho}_{D_j}$ for the WDBC data set as a function of the number of trees in RF and the number of features used to split a node, when averaging over 20 different runs performed with the same number of trees and features.

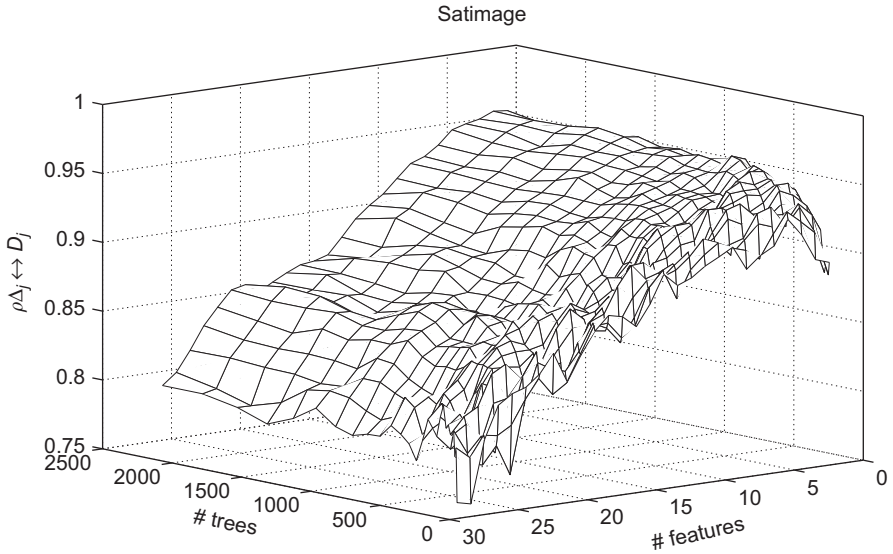

Fig. 9. The Spearmen correlation coefficient between variable rankings obtained by the two measures as a function of the number of trees in RF and the number of features used to split a node for the Satimage data set.

WDBC data set, but the correlation magnitude is much higher, especially for RF consisting of a large number of trees.

Plots presented in Figs. 9 and 10 explore correlation between variable rankings based on the $\overline{\Delta}_j$ and $\overline{D}_j$ variable importance measures. Fig. 9 plots the Spearmen correlation coefficient between variable rankings obtained by the two measures for the Satimage data set. Fig. 9 shows that the two measures provide very similar variable rankings, especially when the number of features used to split a node is small. The same pattern of correlations is also observed for the WDBC data set, see Fig. 10.

What observations can be made from the results of the consistency studies?

i. Variable importance rankings may, to a great extent, depend on the number of variables used to split a node when designing RF, especially when the number of trees in RF is small. This fact should not be forgotten when using variable importance evaluations for data exploration and understanding. In many applications of RF, the number of variables used to split a node is set to the default value given by $\lfloor \log_2(N)+1 \rfloor$ or $\sqrt{N}$ and not optimized. However, the optimal number of variables used to split a node may differ significantly from the default value. Thus, a rather different ranking of variables may be obtained when using the optimal number of variables to split a node instead of the default one.
ii. Variable rankings obtained for the same data set from two large RFs designed using a similar number of variables to split a node are very similar.
iii. For small data sets, variable importance evaluations obtained for unseen data can be rather different from those computed in the RF designing process.
iv. Both measures, $\overline{\Delta}_j$ and $\overline{D}_j$, provide similar variable rankings.

6.2. Results of the generality studies

The effectiveness of a set of features corresponding to highest values of the measures $\overline{D}_j$ and $\overline{\Delta}_j$ in SVM, *k*-NN, and RF classifiers was studied. Since variable rankings obtained using the two

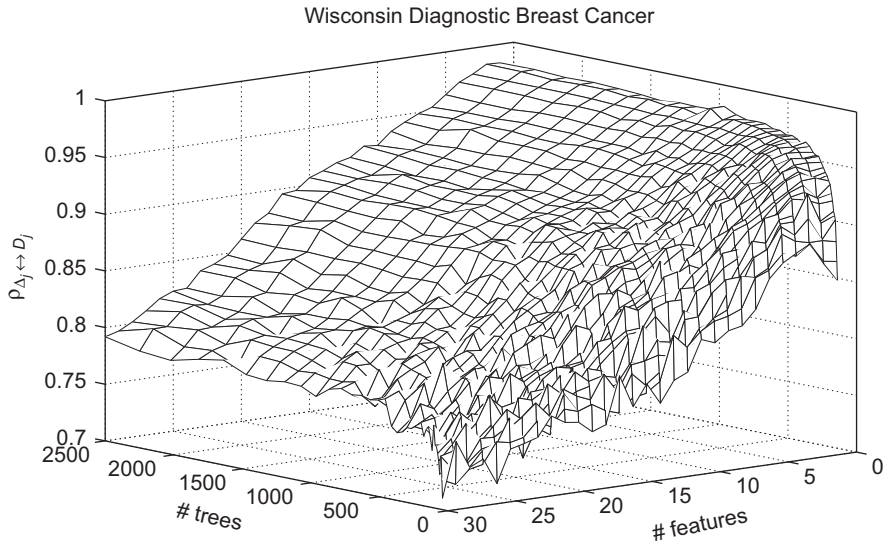

Fig. 10. The Spearmen correlation coefficient between variable rankings obtained by the two measures as a function of the number of trees in RF and the number of features used to split a node for the WDBC data set.

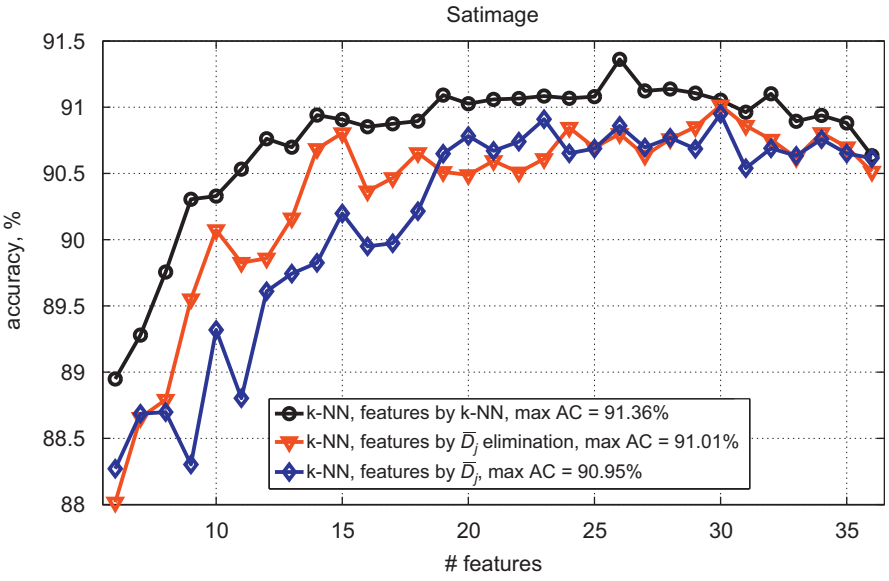

Fig. 11. The dependency of the test set data classification accuracy of the *k*-NN classifier on the number of features selected by the different techniques for the Satimage data set.

measures were highly correlated, we present here results only for the $\overline{D}_j$ measure. Fig. 11 presents the dependency of the test set data classification accuracy of the *k*-NN classifier on the number of features selected by different techniques for the Satimage data set. Three techniques have been applied for feature selection: *k*-NN-based forward selection (backward elimination provided similar results), selection by taking features corresponding to highest values of the $\overline{D}_j$ measure, and recursive feature elimination one-by-one based on $\overline{D}_j$. The maximum achieved accuracy (AC) is also provided in the figure. As can be seen from Fig. 11, the *k*-NN-based forward feature selection outperformed the other two techniques. The difference in accuracy was statistically significant for most of the feature sets. Similar results have also been obtained for the WDBC and Thyroid data sets, presented in Figs. 12 and 13, respectively. Variables selected by the *k*-NN technique seems to be more general than the variables selected according to the RF variable importance measure $\overline{D}_j$. As can be seen in Fig. 13, the accuracy of the *k*-NN classifier deteriorates significantly when switching from feature sets selected by the *k*-NN technique to feature sets selected according to the variable importance measure $\overline{D}_j$.

One can expect the *k*-NN-based variable selection to outperform the other variable selection techniques, since in the case of *k*-NN, a classifier of the same type is used for both variable selection and classification. However, plots presented in Figs. 14–17 demonstrate that features selected by the *k*-NN technique are also efficient in SVM and RF classifiers.

Figs. 14 and 15 present the dependency of the test set data classification accuracy of the SVM classifier on the number of features selected by the different techniques for the WDBC and Satimage data sets, respectively. In the figures, "features by SVM" refers to recursive forward feature selection based on SVM accuracy. As can be seen from Figs. 14 and 15, for most of the feature sets, SVM exploiting features selected using $\overline{D}_j$ performed

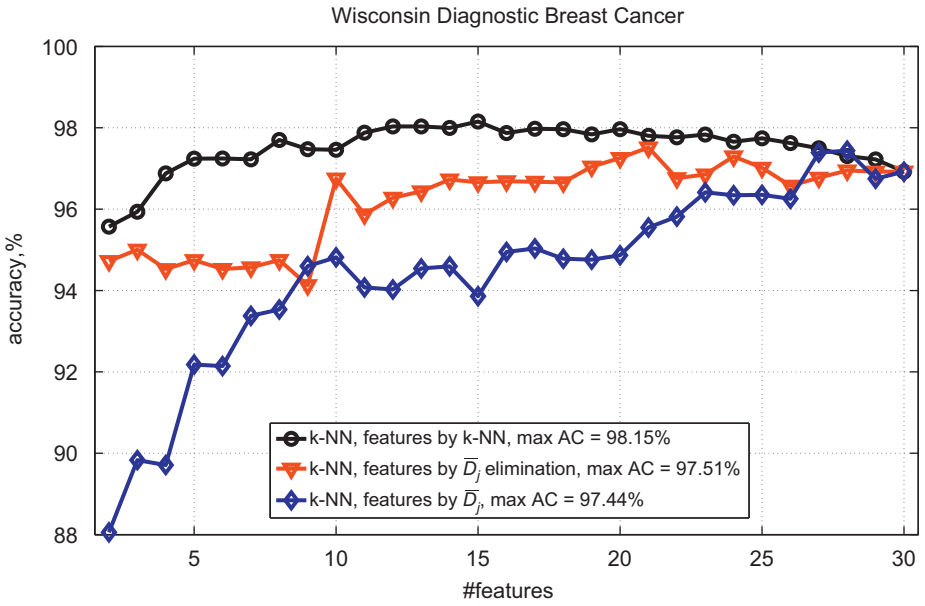

Fig. 12. The dependency of the test set data classification accuracy of the *k*-NN classifier on the number of features selected by the different techniques for the WDBC data set.

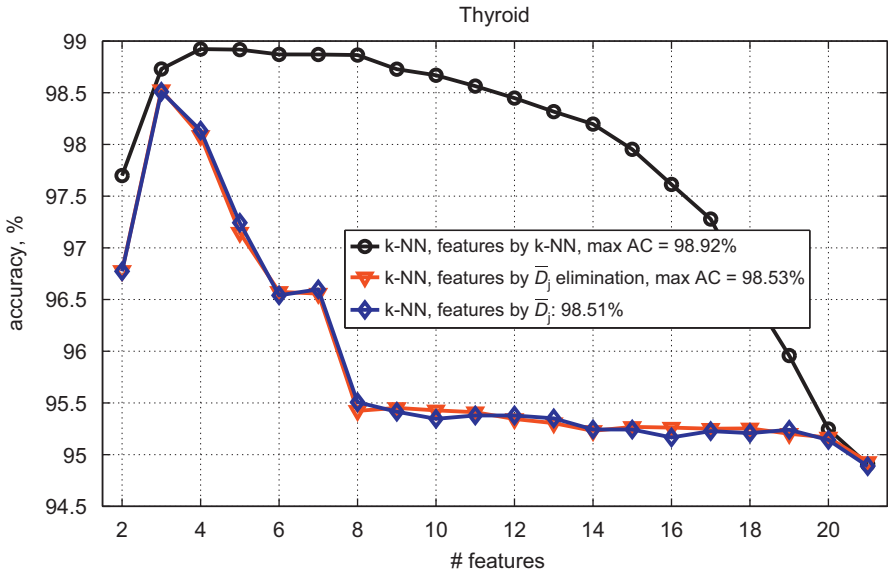

Fig. 13. The dependency of the test set data classification accuracy of the *k*-NN classifier on the number of features selected by the different techniques for the thyroid data set.

significantly worse than the *k*-NN and SVM features-based counterparts.

In Figs. 16 and 17, "features by RF" refers to recursive feature elimination one-by-one based on RF accuracy. As expected, this was the best approach to feature selection when using RF as a classifier. It is interesting to see that, for small feature sets, RF exploiting features selected by the *k*-NN technique was more accurate than the counterparts using features selected according to the $\overline{D}_j$ measure and the $\overline{D}_j$ measure-based recursive elimination. For many of the feature sets the difference in accuracy was statistically significant.

SVM exploiting feature selected by the SVM provided the highest classification accuracy for both Satimage and WDBC data sets. Thus, it is interesting to see correlation between the feature ranking obtained from the SVM-based feature selection and rankings produced by the $\overline{D}_j$ and $\overline{\Delta}_j$ measures. Fig. 18 presents the Spearman correlation coefficient of such rankings for the Satimage data set as a function of the number of trees in RF and the number of features used to split a tree node. Though statistically significant, the correlations are not high.

The following observations can be made from the generality studies:

i. There is no evidence supporting the claim of high generality of feature subsets selected based on the $\overline{D}_j$ and $\overline{\Delta}_j$ measures.
ii. Feature subsets selected using the *k*-NN classifier-based forward or backward feature selection exhibit higher generality and efficiency (in terms of classification accuracy) than feature subsets determined using the $\overline{D}_j$ and $\overline{\Delta}_j$ measures.

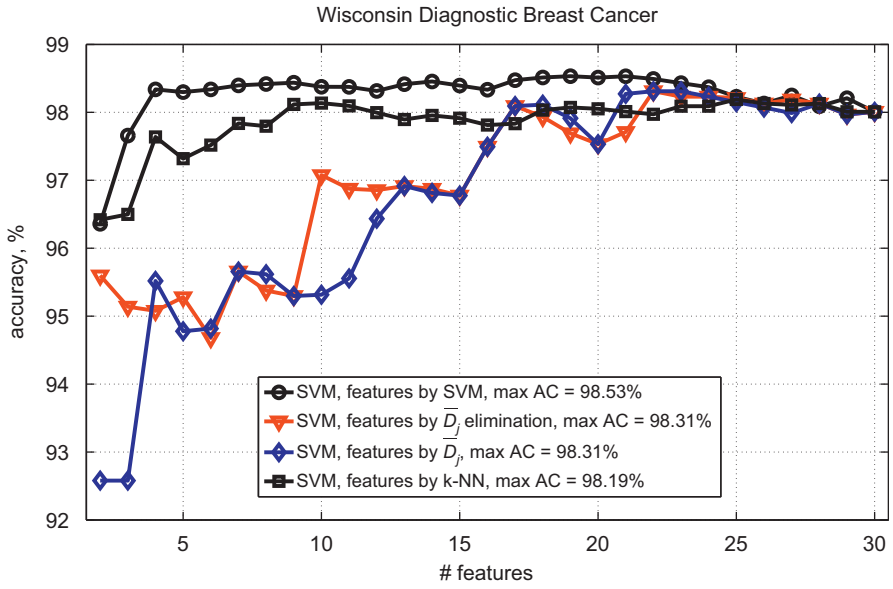

Fig. 14. The dependency of the test set data classification accuracy of the SVM classifier on the number of features selected by the different techniques for the WDBC data set.

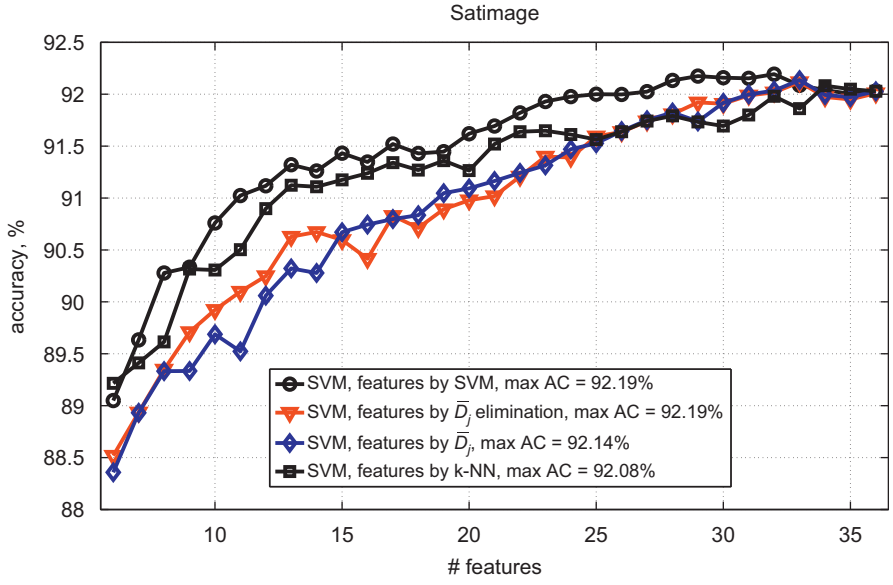

Fig. 15. The dependency of the test set data classification accuracy of the SVM classifier on the number of features selected by the different techniques for the Satimage data set.

iii. The RF accuracy-based recursive feature selection is capable of providing a significantly higher classification accuracy than feature selection based on the $\overline{D}_j$ and $\overline{\Delta}_j$ measures.
iv. In problems where RF is outperformed by other techniques (by SVM for example), correlation between rankings of features in feature subsets providing the highest classification accuracy (feature subsets determined by SVM accuracy-based recursive forward feature selection, for example) and rankings based on the $\overline{D}_j$ and $\overline{\Delta}_j$ measures can be rather low.

7. Tests concerning problem complexity

The aim of these studies is to get some insights into "suitability" of a problem at hand for RF-based classification. To assess the problem complexity, several measures studied in [36,37] are used in this work. The measures are listed in Table 4.

The measures F1, F2, F3, and F4 reflect the degree of overlap of individual feature values, while the measures L1 and L2 assess linear separability of classes. To compute values of the measures L1, L2, and L3, a linear classifier is build. An SVM with a linear kernel trained by the sequential minimal optimization algorithm [110] is used to build the linear classifier in this work. The measures N1, N2, and N3 express mixture identifiability and are attributed to the class of measures reflecting separability of classes [37]. To describe geometry, topology, and density of manifolds spanned by different classes, measures L3, N4, T1, and T2 are used.

Fifteen data sets have been used in these studies. Three of the data sets were used in our studies discussed in the previous subsections, while the other twelve were taken from the literature [111,112]. Information on the data sets used in the experiment is given in Table 5. The table also provides the average test set data classification accuracy we obtained in the experiment for the SVM

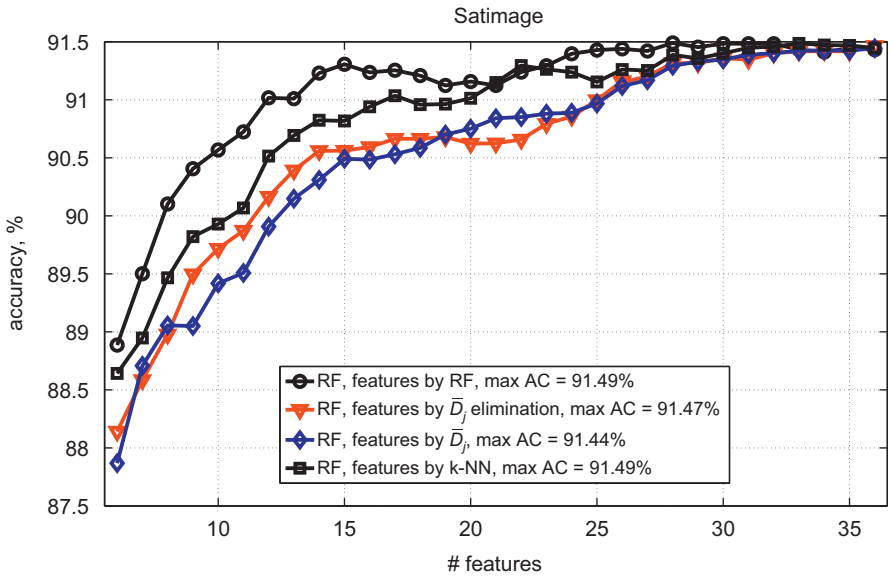

Fig. 16. The dependency of the test set data classification accuracy of the RF classifier on the number of features selected by the different techniques for the Satimage data set.

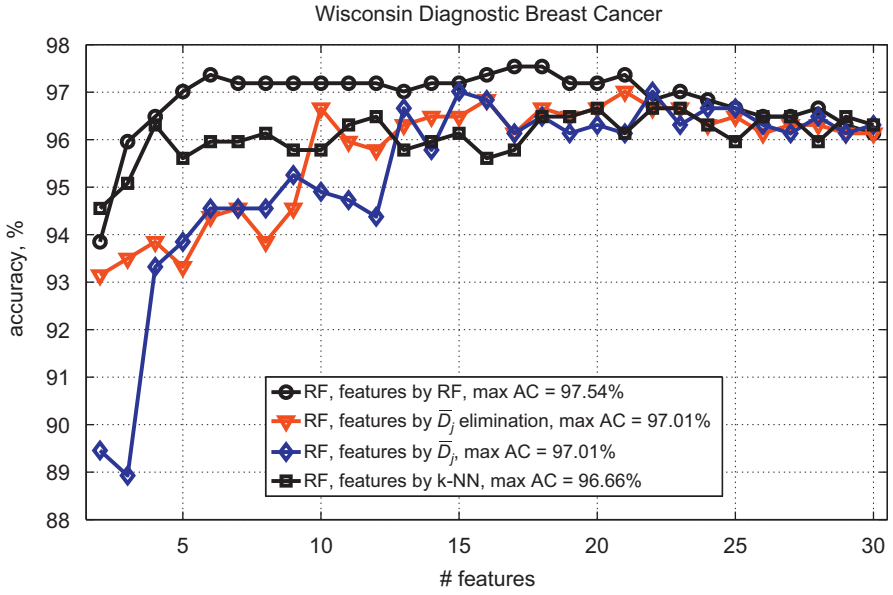

Fig. 17. The dependency of the test set data classification accuracy of the RF classifier on the number of features selected by the different techniques for the WDBC data set.

and RF classifiers and the normalized difference in accuracy d. The formalized difference is given by

$$d = \frac{A_{\mathrm{SVM}} - A_{\mathrm{RF}}}{100 - \overline{E}} \tag{6}$$

where A_{SVM} and A_{RF} are the SVM and RF test set classification accuracy, and $\overline{E}$ is the average of the SVM and RF classification error. Hyper-parameters of SVM (width of the Gaussian kernel and the regularization parameter) and RF were carefully selected when running the experiments. The RF accuracy was assessed using the OOB data. To assess the SVM accuracy 80% of data were randomly selected for training and 20% for validation. The experiment was repeated as many times as it was required to get 15,000 observations for validation in total. Then the average classification accuracy for these 15,000 observations was calculated.

We have chosen to compare RF and SVM, since SVM is often considered as being one of the most successful types of classifier models. Values of the 13 problem complexity measures computed for these data sets are given in Table 6. For multi-class problems, values of the measures were computed for pairwise binary classifications and averaged. In the table are also given values of the correlation coefficient R computed between the measures and the difference in accuracy d. Values shown in bold indicate correlation significant at the 90% confidence level. Thus, only four measures, F2, F3, N2, and T2 out of 13 exhibit statistically significant correlations with the difference in classification accuracy of the SVM and RF classifiers. The statistically significant negative correlation between the F3 and d indicates that for problems with a large maximum efficiency of individual features, one can expect obtaining a higher accuracy from RF than from SVM. The negative correlation of T2 and d should indicate that problems with a large number of observations per

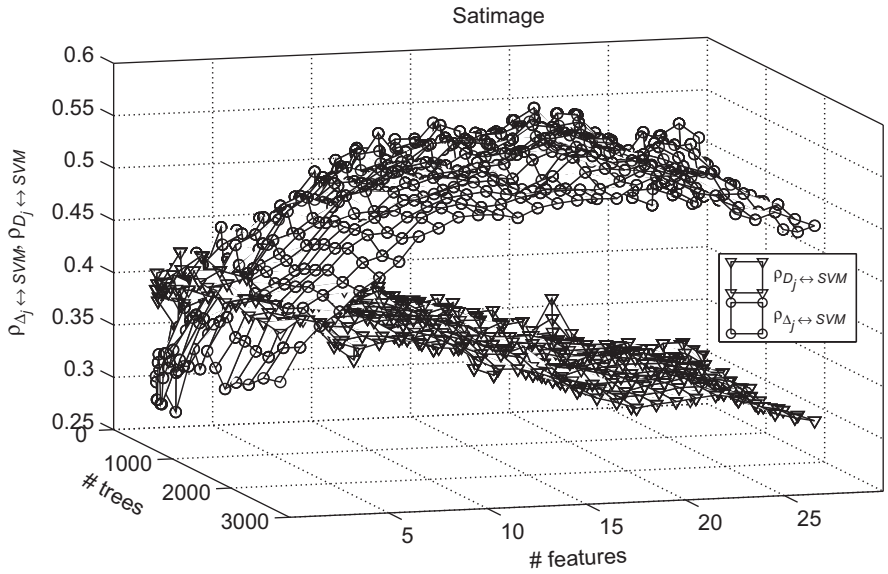

Fig. 18. The Spearman correlation coefficient between the feature ranking obtained from the SVM accuracy-based feature selection and rankings produced by the $\overline{D}_j$ and $\overline{\Delta}_j$ measures as a function of the number of trees in RF and the number of features used to split a tree node.

Table 4
A list of data complexity measures used.

F1:	Maximum Fisher's discriminant ratio
F2:	Volume of overlap region
F3:	Maximum (individual) feature efficiency
F4:	Collective feature efficiency (sum of individual feature efficiencies)
N1:	Fraction of points on class boundary
N2:	Ratio of average intra/inter class nearest neighbor (NN) distance
N3:	Leave-one-out error rate of 1-NN classifier
N4:	Nonlinearity of 1-NN classifier
L1:	Mean absolute training error of a linear classifier
L2:	Error rate of a linear classifier
L3:	Nonlinearity of a linear classifier
T1:	Fraction of points with associated adherence subsets retained
T2:	Average number of points per dimension

Table 5
Test set data classification accuracy for the SVM and RF classifiers along with the normalized difference in accuracy.

N#	Data set	# Classes	# Data	# Features	% SVM	% RF	Difference
1	Plankton [111]	5	8440	47	88.33	86.71	0.1299
2	Australian credit [112]	2	690	14	86.78	87.41	−0.0493
3	German credit [112]	2	1000	24	76.59	77.44	−0.0368
4	Pima diabetes [112]	2	768	8	77.26	76.68	0.0252
5	Vehicle silhouettes [112]	4	846	18	85.12	75.03	0.5063
6	Balance scale [112]	3	625	4	97.32	89.56	1.1828
7	Image segmentation [112]	7	2310	18	96.52	98.21	−0.6401
8	Optical digits [112]	10	5620	62	98.72	98.52	0.1480
9	WDBC original [38]	2	683	9	97.09	97.40	−0.1119
10	Heart disease [38]	2	296	13	82.21	82.22	−0.0009
11	Credit card [38]	2	653	15	87.20	87.90	−0.0562
12	Bupa liver [38]	2	345	6	72.27	74.17	−0.0710
13	Satimage	6	6435	36	92.12	91.49	0.0765
14	Thyroid	3	7200	21	98.16	99.66	−1.3768
15	WDBC	2	569	30	98.01	97.54	0.2102

dimension are more suitable for RF than SVM. However, the diverse results obtained for the 6th and 14th data sets suggest that one should be careful with such a conclusion. Analysis of relation between a linear combination of the four measures and the difference in accuracy d does not provide much more insights into the results.

8. Discussion

Fast training, the possibility of obtaining the generalization error estimate without splitting the data set into learning and validation subsets, variable importance evaluations available as a

Table 6
Values of the complexity measures computed for different data sets along with values of the correlation coefficient R between the measures and the difference in accuracy d.

Data set	F1	F2	F3	F4	N1	N2	N3	N4	L1	L2	L3	T1	T2
1	10.12	0.00	0.14	0.48	0.13	0.55	0.07	0.09	0.78	0.11	0.06	1.00	71.83
2	3.88	0.00	0.03	0.07	0.28	0.55	0.21	0.15	0.29	0.15	0.13	1.00	49.29
3	0.72	0.66	0.01	0.02	0.49	0.85	0.33	0.27	0.72	0.27	0.46	1.00	41.67
4	0.20	0.25	0.01	0.02	0.44	0.84	0.29	0.27	0.69	0.35	0.50	1.00	96.00
5	7.63	0.03	0.25	0.80	0.22	0.58	0.13	0.17	0.67	0.24	0.21	0.99	23.50
6	0.98	1.00	0.00	0.00	0.26	0.65	0.18	0.22	0.37	0.11	0.35	0.89	104.16
7	149.20	0.00	0.81	0.95	0.02	0.14	0.01	0.02	0.37	0.02	0.01	0.76	36.67
8	27.92	0.00	0.30	0.91	0.01	0.44	0.00	0.00	0.64	0.00	0.00	0.98	18.13
9	4.09	0.22	0.12	0.24	0.06	0.33	0.04	0.03	0.46	0.03	0.01	0.80	75.89
10	1.75	0.21	0.02	0.06	0.37	0.74	0.24	0.15	0.56	0.19	0.15	1.00	22.77
11	3.67	0.00	0.03	0.07	0.28	0.51	0.18	0.13	0.27	0.14	0.12	1.00	43.53
12	0.09	0.07	0.03	0.11	0.57	0.91	0.37	0.34	0.84	0.42	0.50	1.00	57.50
13	11.58	0.00	0.51	0.85	0.04	0.37	0.02	0.03	0.99	0.04	0.02	0.95	59.58
14	3.33	0.00	0.81	0.84	0.06	0.25	0.03	0.10	0.23	0.05	0.35	0.93	228.57
15	10.29	0.00	0.52	1.00	0.07	0.56	0.05	0.03	0.54	0.05	0.03	1.00	18.97
R	−0.305	**0.514**	**−0.573**	−0.272	0.193	**0.443**	0.204	0.204	0.270	0.160	−0.010	0.222	**−0.487**

byproduct of training, only one parameter to be tuned experimentally, make RF a very popular data mining technique.

8.1. Classification accuracy

Random forests have been applied in a variety of fields and their performance was compared to the performance of many other popular techniques. As one can expect, both the few large scale comparisons and numerous small scale studies demonstrate that the superiority of RF over or loss to the other techniques is very problem dependent. When studying the difference in accuracy between RF and SVM classifiers, statistically significant correlations were found between several problem complexity measures and the difference in accuracy. However, the correlations are not very strong and do not allow making strong statements.

8.2. Consistency of variable importance evaluations

Variable importance evaluations based on the variable importance measures available from RF are widely used for data exploration and understanding. Our experimental investigations have shown that both measures, $\overline{\Delta}_j$ and $\overline{D}_j$, provide similar variable rankings. The investigations have also shown that for large data sets, RF consisting of a large number of trees provides consistent variable importance evaluations, exhibiting low variance and high correlation when varying the number of trees in RF. However, variable importance rankings may, to a great extent, depend on the number of variables used to split a node when designing RF, especially when the number of trees in RF is small. This fact should not be forgotten when using variable importance evaluations for data exploration and understanding. In many applications of RF found in the surveyed literature, the number of variables used to split a node is set to a default value, given by $\lfloor \log_2(N)+1 \rfloor$ or $\sqrt{N}$, without any further optimization. However, the optimal number of variables used to split a node may differ significantly from the default value. Thus, a rather different ranking of variables may be obtained when using the optimal number of variables to split a node instead of the default one. On the other hand, one can expect obtaining very similar variable rankings from two large RFs designed using the default and the optimal number of variables for node splitting, respectively, when the difference between these numbers is small. One should also bear in mind that for small data sets, as it is often the case with various RF applications, variable importance evaluations obtained for unseen data can be rather different from those computed in the RF designing process.

8.3. Generality of variable importance evaluations

The results of experimental investigations concerning generality of variable importance evaluations indicate that there is no evidence supporting the belief in high generality of feature subsets selected based on the $\overline{D}_j$ and $\overline{\Delta}_j$ measures. In most of the cases, feature subsets selected using the k-NN classifier-based forward or backward feature selection exhibited higher generality and efficiency (in terms of classification accuracy) than feature subsets determined using the $\overline{D}_j$ and $\overline{\Delta}_j$ measures, especially for small feature subsets. Such superiority of the k-NN-based feature selection was observed even in the cases where RF was used as a base classifier. Thus, an unconditional belief in variable importance evaluations available from RF seems to be not justified. The experimental tests have also shown that the RF accuracy-based recursive feature selection is capable of providing a significantly higher classification accuracy than feature selection based on the $\overline{D}_j$ and $\overline{\Delta}_j$ measures.

When studying feature subsets related to problems where RF was outperformed by other techniques (by SVM for example), it was found out that correlation between rankings of features in feature subsets providing the highest classification accuracy (feature subsets determined by SVM accuracy-based recursive forward feature selection, for example) and rankings based on the $\overline{D}_j$ and $\overline{\Delta}_j$ measures can be rather low.

8.4. Designing issues

What issues are worth bearing in mind when designing and using RF for solving a problem at hand? First, to avoid over-fitting, the number of trees in RF should be sufficiently large. The optimal number of features used for node splitting in the RF designing process is to be selected experimentally, rather than set to the default value, given by $\lfloor \log_2(N)+1 \rfloor$ or $\sqrt{N}$. Selection of salient variables for the problem at hand should be based on advanced feature selection techniques, rather than on values of the $\overline{D}_j$ and $\overline{\Delta}_j$ measures, available from the RF designing process. The proximity matrix available from RF is a very useful information source and can be used for data exploration: to detect outliers and mislabeled data, to replace missing values, to visualize data by applying the multidimensional scaling to

the proximity matrix, for example. It is worth mentioning that noisy irrelevant features, which are not used in the final RF, do not have any affect on the proximity values.

Only one parameter to be tuned experimentally and the possibility of obtaining the generalization error estimate without splitting the data set into learning and validation subsets, make the designing process of RF much faster than those of MLP or SVM. Computations in one grown tree are also very fast, since only one comparison instruction is performed in each node of the tree. The computation time in RF grows linearly with the increasing number of trees in RF. Computations in different trees of RF can be easily parallelized, if required.

8.5. Challenges

Low correlation between RF trees and low RF bias are essential for RF accuracy. Thus, when aiming to improve the RF performance, RF designing techniques capable of reducing the average correlation between RF trees without increasing the RF bias is an interesting and challenging research direction. Dynamic selection of relevant trees to be included into RF of dynamic size can lead to RF, specific for each observation being analyzed and may help reducing the average correlation. The proximity matrix available from RF can be an important information source for designing data dependent RFs. Majority voting is the scheme applied in an ordinary RF to combine decisions of single trees. Schemes for weighted voting using dynamic, data dependent aggregation weights is another promising area of research aiming to improve the RF performance. The task of variable importance evaluations, especially when variables of different types are used to create RF, also needs more studies.

Bagging and random features are two types of randomness usually used in the RF designing process. It is interesting to explore other types of randomness, random Boolean combinations of features [10], for example. The size of bootstrap samples in bagging does not necessary to be the same as that of the original data set [11].

Acknowledgements

Useful suggestions from the referees are gratefully acknowledged. We acknowledge the support from the agency for International Science and Technology Development Programmes in Lithuania (COST Actions IC0602 and IC0806). The infrastructure for parallel and distributed computing, and e-services (LitGrid) was used in the studies.

References

[1] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[2] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, UK, 2004.
[3] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, Singapore, 2006.
[4] M.E. Tipping, Sparse bayesian learning and the relevance vector machine, Journal of Machine Learning Research 1 (2001) 211–244.
[5] A. Verikas, A. Lipnickas, K. Malmqvist, M. Bacauskiene, A. Gelzinis, Soft combination of neural classifiers: a comparative study, Pattern Recognition Letters 20 (1999) 429–444.
[6] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Chapman & Hall, 1993.
[7] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, Los Altos, CA, 1999 〈http://www.cs.waikato.ac.nz/ml/weka〉.
[8] L. Breiman, Bagging predictors, Machine Learning 24 (2) (1996) 123–140.
[9] T.G. Dietterich, An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, Machine Learning 40 (2) (2000) 139–157.
[10] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.
[11] G. Biau, L. Devroye, G. Lugosi, Consistency of random forests and other averaging classifiers, Journal of Machine Learning Research 9 (2008) 2015–2033.
[12] T.K. Ho, The random subspace method for constructing decision forests, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (8) (1998) 832–844.
[13] G. Nimrod, A. Szilagyi, C. Leslie, N. Ben-Tal, Identification of DNA-binding proteins using structural, electrostatic and evolutionary features, Journal of Molecular Biology 387 (4) (2009) 1040–1053.
[14] H.T. Chen, T.L. Liu, C.S. Fuh, Segmenting highly articulated video objects with weak-prior random forests, in: A. Leonardis, H. Bischof, A. Pinz (Eds.), ECCV 2006, Part IV, Lecture Notes in Computer Science, vol. 3954, Springer-Verlag, Berlin, Heidelberg2006, pp. 373–385.
[15] J. Ham, Y. Chen, M.M. Crawford, J. Ghosh, Investigation of the random forest framework for classification of hyperspectral data, IEEE Transactions on Geoscience and Remote Sensing 43 (3) (2005) 492–501.
[16] M.M. Crawford, J. Ham, Y. Chen, J. Ghosh, Random forests of binary hierarchical classifiers for analysis of hyperspectral data, in: 2003 IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data, IEEE, Greenbelt, MD, USA2004, pp. 337–345.
[17] J. Peters, B. De Baets, N.E.C. Verhoest, R. Samson, S. Degroeve, P. De Becker, W. Huybrechts, Random forests as a tool for ecohydrological distribution modelling, Ecological Modelling 207 (2007) 304–318.
[18] J. Peters, N.E.C. Verhoest, R. Samson, M. Van Meirvenne, L. Cockx, B. De Baets, Uncertainty propagation in vegetation distribution models based on ensemble classifiers, Ecological Modelling 220 (2009) 791–804.
[19] P.A. Hernandez, I. Franke, S.K. Herzog, V. Pacheco, L. Paniagua, H.L. Quintana, A. Soto, J.J. Swenson, C. Tovar, T.H. Valqui, J. Vargas, B.E. Young, Predicting species distributions in poorly-studied landscapes, Biodiversity and Conservation 17 (2008) 1353–1366.
[20] I. Oparin, O. Glembek, L. Burget, J. Cernocky, Morphological random forests for language modeling of inflectional languages, in: 2008 IEEE Workshop on Spoken Language Technology, SLT 2008, IEEE, Goa, India2008, pp. 189–192.
[21] J. Ramirez, J.M. Gorriz, R. Chaves, M. Lopez, D. Salas-Gonzalez, I. Alvarez, F. Segovia, SPECT image classification using random forests, Electronics Letters 45 (12) (2009) 604–605.
[22] A.G. Heidema, J.M.A. Boer, N. Nagelkerke, E.C.M. Mariman, D.L. van der A, E.J.M. Feskens, The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases, Accident Analysis and Prevention 7 (23) (2006) 1–15.
[23] P. Han, X. Zhang, R.S. Norton, Z.P. Feng, Large-scale prediction of long disordered regions in proteins using random forests, BMC Bioinformatics 10 (8) (2009) 1–9.
[24] J.D. Watts, R.L. Lawrence, P.R. Miller, C. Montagne, Monitoring of cropland practices for carbon sequestration purposes in north central montana by landsat remote sensing, Remote Sensing of Environment 113 (2009) 1843–1852.
[25] S.R. Joelsson, J.A. Benediktsson, J.R. Sveinsson, Feature selection for morphological feature extraction using random forests, in: Seventh Nordic Signal Processing Symposium, IEEE, New York, Reykjavik, Iceland2006, pp. 138–141.
[26] S. Gupta, S. Matthew, P.M. Abreu, J.A. de Sousa, QSAR analysis of phenolic antioxidants using MOLMAP descriptors of local properties, Bioorganic & Medicinal Chemistry 14 (2006) 1199–1206.
[27] S. Bernard, L. Heutte, S. Adam, Using random forests for handwritten digit recognition, ICDAR 2007: Ninth International Conference on Document Analysis and Recognition, vol. 1–2, Curitiba, Brazil2007, pp. 1043–1047.
[28] M. Thums, C.J.A. Bradshaw, M.A. Hindell, A validated approach for supervised dive classification in diving vertebrates, Journal of Experimental Marine Biology and Ecology 363 (2008) 75–83.
[29] A. Liaw, M. Wiener, Classification and regression by random forest, R News 2 (3) (2002) 18–22.
[30] L. Breiman, RFtools—for predicting and understanding data, Technical Report, Berkeley University, Berkeley, USA 〈http://oz.berkeley.edu/users/breiman/RandomForests/cc.papers.htm〉, 2004.
[31] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, B.P. Feuston, Random forest: a classification and regression tool for compound classification and QSAR modeling, Journal of Chemical Information and Computer Sciences 43 (6) (2003) 1947–1958.
[32] V. Svetnik, T. Wang, C. Tong, A. Liaw, R.P. Sheridan, Q.H. Song, Boosting: an ensemble learning tool for compound classification and QSAR modeling, Journal of Chemical Information and Modeling 45 (3) (2005) 786–799.
[33] A. Folleco, T.M. Khoshgoftaar, J. Van Hulse, L. Bullard, Software quality modeling: the impact of class noise on the random forest classifier, 2008 IEEE Congress on Evolutionary Computation, vol. 1–8, IEEE, Hong Kong, PR China2008, pp. 3853–3859.
[34] A. Statnikov, L. Wang, C.F. Aliferis, A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification, BMC Bioinformatics 9 (319) (2008) 1–10.
[35] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., John Wiley & Sons, New York, 2001.
[36] T.K. Ho, A data complexity analysis of comparative advantages of decision forest constructors, Pattern Analysis and Applications 5 (2002) 102–112.

[37] T.K. Ho, M. Basu, Complexity measures of supervised classification problems, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (3) (2002) 289–300.
[38] D. Meyer, F. Leisch, K. Hornik, The support vector machine under test, Neurocomputing 55 (1–2) (2003) 169–186.
[39] R.E. Banfield, L.O. Hall, K.W. Bowyer, W.P. Kegelmeyer, A comparison of decision tree ensemble creation techniques, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (1) (2007) 173–180.
[40] Y. Freund, R.E. Schapire, Experiments with a new boosting algorithm, in: Proceedings of the 13th International Conference on Machine Learning, 1996, pp. 148–156.
[41] E. Alpaydin, Combined 5×2 cv F test for comparing supervised classification learning algorithms, Neural Computation 11 (8) (1999) 1885–1892.
[42] J. Demsar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research 7 (2006) 1–30.
[43] R. Diaz-Uriarte, S. Alvarez de Andres, Gene selection and classification of microarray data using random forest, BMC Bioinformatics 7 (3) (2006) 1–13.
[44] V. Sindhwani, P. Bhattacharyya, S. Rakshit, Information theoretic feature crediting in multiclass support vector machines, in: R. Grossman, V. Kumar (Eds.), Proceedings of the First SIAM International Conference on Data Mining, SIAM, Chicago, USA2001, pp. 1–18.
[45] Y. Xie, X. Li, E.W.T. Ngai, W. Ying, Customer churn prediction using improved balanced random forests, Expert Systems with Applications 36 (2009) 5445–5449.
[46] W. Ying, X. Li, Y. Xie, E. Johnson, Preventing customer churn by using random forests modeling, in: IEEE International Conference on Information Reuse and Integration (IRI-2008), IEEE, Las Vegas, USA2008, pp. 429–434.
[47] K. Coussement, D. Van den Poel, Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques, Expert Systems with Applications 34 (2008) 313–327.
[48] K. Coussement, D. Van den Poel, Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers, Expert Systems with Applications 36 (2009) 6127–6134.
[49] C. Whitrow, D.J. Hand, P. Juszczak, D. Weston, N.M. Adams, Transaction aggregation as a strategy for credit card fraud detection, Data Mining and Knowledge Discovery 18 (1) (2009) 30–55.
[50] D. Perdiguero-Alonso, F.E. Montero, A. Kostadinova, J.A. Raga, J. Barrett, Random forests, a novel approach for discrimination of fish populations using parasites as biological tags, International Journal for Parasitology 38 (2008) 1425–1434.
[51] I. Koprinska, J. Poon, J. Clark, J. Chan, Learning to classify e-mail, Information Sciences 177 (2007) 2167–2187.
[52] C. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA, USA, 1999.
[53] W. Yan, Application of random forest to aircraft engine fault diagnosis, in: IMACS Multiconference on Computational Engineering in Systems Applications (CESA), IEEE, Beijing, PR China2006, pp. 468–475.
[54] B. Slabbinck, B. De Baets, P. Dawyndt, P. De Vos, Towards large-scale FAME-based bacterial species identification using machine learning techniques, Systematic and Applied Microbiology 32 (2009) 163–176.
[55] A.Z. Kouzani, Faceparts for recognition, in: TENCON 2006–2006 IEEE Region 10 Conference, IEEE, Hong Kong, PR China2006, pp. 1232–1235.
[56] B.S. Yang, X. Di, T. Han, Random forests classifier for machine fault diagnosis, Journal of Mechanical Science and Technology 22 (2008) 1716–1725.
[57] A.Z. Kouzani, Subcellular localisation of proteins in fluorescent microscope images using a random forest, 2008 IEEE International Joint Conference on Neural Networks, vol. 1–8, IEEE, Hong Kong, PR China2008, pp. 3926–3932.
[58] G. Zhang, H. Li, B. Fang, Discriminating acidic and alkaline enzymes using a random forest model with secondary structure amino acid composition, Process Biochemistry 44 (2009) 654–660.
[59] M.B. Garzon, R. Blazek, M. Neteler, R.S. de Dios, H.S. Ollero, C. Furlanello, Predicting habitat suitability with machine learning models: the potential area of *Pinus sylvestris* L. in the Iberian peninsula, Ecological Modelling 197 (2006) 383–393.
[60] A.M. Prasad, L.R. Iverson, A. Liaw, Newer classification and regression tree techniques: bagging and random forests for ecological prediction, Ecosystems 9 (2006) 181–199.
[61] S. Gupta, J. Aires-deSousa, Comparing the chemical spaces of metabolites and available chemicals: models of metabolite-likeness, Molecular Diversity 11 (1) (2007) 23–36.
[62] S. Tognazzo, B. Emanuela, F.A. Rita, G. Stefano, M. Daniele, S.C. Fiorella, Z. Paola, Probabilistic classifiers and automated cancer registration: an exploratory application, Journal of Biomedical Informatics 42 (2009) 1–10.
[63] D. Kocev, S. Dzeroski, M.D. White, G.R. Newell, P. Griffioen, Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition, Ecological Modelling 220 (2009) 1159–1168.
[64] K. Marsolo, M. Twa, M.A. Bullimore, S. Parthasarathy, The support vector machine under test, IEEE Transactions on Information Technology in Biomedicine 11 (2) (2007) 203–212.
[65] W. Buckinx, D. Van den Poel, Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting, European Journal of Operational Research 164 (2005) 252–268.
[66] R. Tibshirani, T. Hastie, B. Narasimhan, G. Chu, Class prediction by nearest shrunken centroids, with applications to DNA microarrays, Statistical Science 18 (1) (2003) 104–117.
[67] D.J. MacKay, Bayesian interpolation, Neural Computation 4 (1992) 415–447.
[68] D.J.C. MacKay, The evidence framework applied to classification networks, Neural Computation 4 (5) (1992) 720–736.
[69] T. Luo, K. Kramer, D.B. Goldgof, L.O. Hall, S. Samson, A. Remsen, T. Hopkins, Recognizing plankton images from the shadow image particle profiling evaluation recorder, IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics 34 (4) (2004) 1753–1762.
[70] V. Belle, T. Deselaers, S. Schiffer, Randomized trees for real-time one-step face detection and recognition, in: 19th International Conference on Pattern Recognition, vol. 1–6, Tampa, FL, USA, 2008, pp. 3547–3550.
[71] J.A. Benediktsson, J. Chanussot, M. Fauvel, Multiple classifier systems in remote sensing: from basics to recent developments, in: M. Haindl, J. Kittler, F. Roli (Eds.), Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 4471, Springer-Verlag, Berlin, Heidelberg2007, pp. 501–512.
[72] S. Abu-Nimeh, D. Nappa, X. Wang, S. Nair, Bayesian additive regression trees-based spam detection for enhanced email privacy, in: The Third International Conference on Availability, Reliability and Security, IEEE Computer Society, Barcelona, Spain2008, pp. 1044–1051.
[73] P.M. Granitto, F. Biasioli, E. Aprea, D. Mott, C. Furlanello, T.D. Mark, F. Gasperi, Rapid and non-destructive identification of strawberry cultivars by direct PTR-MS headspace analysis and data mining techniques, Sensors and Actuators B 121 (2007) 379–385.
[74] D. Donald, D. Coomans, Y. Everingham, D. Cozzolino, M. Gishen, T. Hancock, Adaptive wavelet modelling of a nested 3 factor experimental design in NIR chemometrics, Chemometrics and Intelligent Laboratory Systems 82 (2006) 122–129.
[75] J.R. Jang, ANFIS: adaptive-network-based fuzzy inference system, IEEE Transactions on Systems, Man, and Cybernetics 23 (1993) 665–685.
[76] A. Zainal, M.A. Maarof, S.M. Shamsuddin, A. Abraham, Ensemble of one-class classifiers for network intrusion detection system, in: Fourth International Symposium on Information Assurance and Security, IEEE Computer Society, Napoli, Italy2008, pp. 180–185.
[77] L.I. Kuncheva, V.J. del Rio Vilas, J.J. Rodriguez, Diagnosing scrapie in sheep: a classification experiment, Computers in Biology and Medicine 37 (2007) 1194–1202.
[78] A. Brenning, Benchmarking classifiers to optimally integrate terrain analysis and multispectral remote sensing in automatic rock glacier detection, Remote Sensing of Environment 113 (1) (2009) 239–247.
[79] W. Buckinx, G. Verstraeten, D. Van den Poel, Predicting customer loyalty using the internal transactional database, Expert Systems with Applications 32 (2007) 125–134.
[80] T. Hancock, R. Put, D. Coomans, Y.V. Heyden, Y. Everingham, A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies, Chemometrics and Intelligent Laboratory Systems 76 (2005) 185–196.
[81] C. Strobl, A.L. Boulesteix, A. Zeileis, T. Hothorn, Bias in random forest variable importance measures: illustrations, sources and a solution, BMC Bioinformatics 8 (25) (2007) 1–21.
[82] D.M. Reif, A.A. Motsinger, B.A. McKinney, J.E. Crowe, J.H. Moore, Feature selection using a random forests classifier for the integrated analysis of multiple data types, in: Proceedings of the 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, IEEE, Toronto, Canada2006, pp. 171–178.
[83] K.J. Archer, R.V. Kimes, Empirical characterization of random forest variable importance measures, Computational Statistics & Data Analysis 52 (4) (2008) 2249–2260.
[84] O. Okun, H. Priisalu, Random forest for gene expression based cancer classification: overlooked issues, in: J. Marti et al.(Ed.), IbPRIA 2007, Part II, Lecture Notes in Computer Science, vol. 4478, Springer-Verlag, Berlin, Heidelberg2007, pp. 483–490.
[85] R. Harb, X. Yan, E. Radwan, X. Su, Exploring precrash maneuvers using classification trees and random forests, Accident Analysis and Prevention 41 (2009) 98–107.
[86] J. Fan, M.E. Nunn, X. Su, Multivariate exponential survival trees and their application to tooth prognosis, Computational Statistics and Data Analysis 53 (2009) 1110–1121.
[87] J. Zhang, M. Zulkernine, A. Haque, Random-forests-based network intrusion detection systems, IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews 38 (5) (2008) 649–659.
[88] J. Zhang, M. Zulkernine, A hybrid network intrusion detection technique using random forests, in: Proceedings of the First International Conference on Availability, Reliability and Security (ARES'06), IEEE Computer Society, Vienna, Austria2006, pp. 262–269.
[89] B.H. Menze, W. Petrich, F.A. Hamprecht, Multivariate feature selection and hierarchical classification for infrared spectroscopy: serum-based detection of bovine spongiform encephalopathy, Analytical and Bioanalytical Chemistry 387 (5) (2007) 1801–1807.
[90] D. Donald, T. Hancock, D. Coomans, Y. Everingham, Bagged super wavelets reduction for boosted prostate cancer classification of seldi-tof mass spectral serum profiles, Chemometrics and Intelligent Laboratory Systems 82 (2006) 2–7.

[91] B. Lariviere, D. Van den Poel, Predicting customer retention and profitability by using random forests and regression forests techniques, Expert Systems with Applications 29 (2005) 472–484.
[92] L.C. Keely, C.M. Tan, Understanding preferences for income redistribution, Journal of Public Economics 92 (2008) 944–961.
[93] K.L. Lunetta, L.B. Hayward, J. Segal, P. Van Eerdewegh, Screening large-scale association study data: exploiting interactions using random forests, BMC Genetics 5 (32) (2004) 1–13.
[94] K.Q. Shen, C.J. Ong, X.P. Li, Z. Hui, E.P.V. Wilder-Smith, A feature selection method for multilevel mental fatigue EEG classification, IEEE Transactions on Biomedical Engineering 54 (7) (2007) 1231–1237.
[95] H. Wang, C. Lin, F. Yang, X. Hu, Hedged predictions for traditional Chinese chronic gastritis diagnosis with confidence machine, Computers in Biology and Medicine 39 (2009) 425–432.
[96] F. Yang, H.Z. Wang, H. Mi, A novel classification method of microarray with reliability and confidence, Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, vol. 1–7, IEEE, Kunming, PR China2008, pp. 1726–1733.
[97] S.M. Wiseman, A. Melck, H. Masoudi, F. Ghaidi, L. Goldstein, A. Gown, S.J.M. Jones, O.L. Griffith, Molecular phenotyping of thyroid tumors identifies a marker panel for differentiated thyroid cancer diagnosis, Annals of Surgical Oncology 15 (10) (2008) 2811–2826.
[98] G.W. Corder, D.I. Foreman, Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach, Wiley, New Jersey, 2009.
[99] S.H. Wu, B.N. Lee, L.C. Wang, M.S. Abadir, Statistical analysis and optimization of parametric delay test, IEEE International Test Conference, vol. 1–2, IEEE, Santa Clara, CA, USA2007, pp. 613–622.
[100] P.O. Gislason, J.A. Benediktsson, J.R. Sveinsson, Random forests for land cover classification, Pattern Recognition Letters 27 (2006) 294–300.
[101] J.B. Gray, G. Fan, Classification tree analysis using TARGET, Computational Statistics & Data Analysis 52 (2008) 1362–1372.
[102] J.J. Rodrguez, C.J. Alonso, Rotation-based ensembles, Current Topics in Artificial Intelligence, Lecture Notes in Computer Science, vol. 3040, Springer-Verlag2003, pp. 498–506.
[103] J.L. Rodrguez, L.I. Kuncheva, C.J. Alonso, Rotation forest: a new classifier ensemble method, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (10) (2006) 1619–1630.
[104] A. Tsymbal, M. Pechenizkiy, P. Cunningham, Dynamic integration with random forests, in: J. Fürnkranz, T. Scheffer, M. Spiliopoulou (Eds.), ECML 2006, Lecture Notes in Artificial Intelligence, vol. 4212, Springer-Verlag, Berlin, Heidelberg2006, pp. 801–808.
[105] E.E. Tripoliti, D.I. Fotiadis, M. Argyropoulou, An automated supervised method for the diagnosis of Alzheimer's disease based on fMRI data using weighted voting schemes, in: IEEE International Workshop on Imaging Systems and Techniques—IST 2008, IEEE, Chania, Greece2008, pp. 338–343.
[106] G. Leshem, Y. Ritov, Traffic flow prediction using adaboost algorithm with random forests as a weak learner, in: Proceedings of World Academy of Science, Engineering and Technology, vol. 19, Bangkok, Thailand, 2007, pp. 193–198.
[107] H.E. Osman, Online random forests based on CorrFS and CorrBE, in: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, Anchorage, AK, USA2008, pp. 863–869.
[108] A. Verikas, M. Bacauskiene, Feature selection with neural networks, Pattern Recognition Letters 23 (11) (2002) 1323–1335.
[109] A. Verikas, M. Bacauskiene, D. Valincius, A. Gelzinis, Predictor output sensitivity and feature similarity-based feature selection, Fuzzy Sets & Systems 159 (2008) 422–434.
[110] J.C. Platt, Fast training of support vector machines using sequential minimal optimization, in: B. Scholkopf, C.J.C. Burges, A.J. Smola (Eds.), Advances in Kernel Methods: Support Vector Learning, MIT Press, Cambridge, MA1998, pp. 185–208.
[111] T. Luo, K. Kramer, D.B. Goldgof, L.O. Hall, A.R.S. Samson, T. Hopkins, Active learning to recognize multiple types of plankton, Journal of Machine Learning Research 6 (2005) 589–613.
[112] S.J. Wang, A. Mathew, Y. Chen, L.F. Xi, L. Mab, J. Lee, Empirical analysis of support vector machine ensemble classifiers, Expert Systems with Applications 36 (2009) 6466–6476.

Antanas Verikas is currently holding a Professor position at both Halmstad University Sweden and Kaunas University of Technology, Lithuania. His research interests include image processing, pattern recognition, artificial neural networks, fuzzy logic, and visual media technology. He is a member of the International Pattern Recognition Society, European Neural Network Society, International Association of Science and Technology for Development, Swedish Society of Learning Systems, and a member of the IEEE.

Adas Gelzinis received the M.S. degree in Electrical Engineering from Kaunas University of Technology, Lithuania, in 1995. He received the Ph.D. degree in Computer Science from the same university, in 2000. He is a Senior Researcher in the Department of Electrical and Control Equipment at Kaunas University of Technology. His research interests include artificial neural networks, kernel methods, pattern recognition, signal and image processing, texture classification.

Marija Bacauskiene is a Senior Researcher in the Department of Electrical and Control Equipment at Kaunas University of Technology, Lithuania. Her research interests include artificial neural networks, image processing, pattern recognition, and fuzzy logic. She participated in various research projects and published numerous papers in these areas.