Expert Systems with Applications 39 (2012) 13046–13050

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

A novel approach to estimate proximity in a random forest: An exploratory study

C. Englund [a,*], A. Verikas [b,c]

[a] Viktoria Institute, Lindholmspiren 3A, S 417 56 Göteborg, Sweden
[b] Intelligent Systems Laboratory, Halmstad University, Box 823, S 301 18 Halmstad, Sweden
[c] Department of Electrical & Control Equipment, Kaunas University of Technology Studentu 50, LT-51368 Kaunas, Lithuania

ARTICLE INFO

Keywords:
Random forest
Proximity matrix
Support vector machine
Kernel matrix
Data mining

ABSTRACT

A data proximity matrix is an important information source in random forests (RF) based data mining, including data clustering, visualization, outlier detection, substitution of missing values, and finding mislabeled data samples. A novel approach to estimate proximity is proposed in this work. The approach is based on measuring distance between two terminal nodes in a decision tree. To assess the consistency (quality) of data proximity estimate, we suggest using the proximity matrix as a kernel matrix in a support vector machine (SVM), under the assumption that a matrix of higher quality leads to higher classification accuracy. It is experimentally shown that the proposed approach improves the proximity estimate, especially when RF is made of a small number of trees. It is also demonstrated that, for some tasks, an SVM exploiting the suggested proximity matrix based kernel, outperforms an SVM based on a standard radial basis function kernel and the standard proximity matrix based kernel.

1. Introduction

Random forests, introduced by Breiman (2001), is a general tool for solving classification and regression problems. RF have become a very popular data mining tool (Verikas, Gelzinis, & Macauskiene, 2011). Qualitative descriptive analysis of sensory data (Granitto, Gasperi, Biasioli, Trainotti, & Furlanello, 2007), classification of multi-source geographic data (Gislason, Benediktsson, & Sveinsson, 2006), characterization of driver behavior in association with crash avoidance maneuvers (Harb, Yan, Radwan, & Su, 2009), predicting customer retention (Larivière & den Poel, 2005), detection of Alzheimer's disease by analysis of single photon emission computed tomography (Ramírez et al., 2010), cardiac arrhythmia diagnosis (Özçift, 2011), appraisal of residential apartments (Antipov & Pokryshevskaya, 2012), machine fault diagnosis (Son, Niu, Yang, Hwang, & Kang, 2009) and analysis of pollutant emissions in a Kraft pulping process aiming to identify the most important variables (Mattieu & Heyen, 2010) are only a few RF application examples.

In addition to a classification or regression model, RF also provides a data proximity matrix (Breiman, 2001; Breiman & Cutler, 2004). The data proximity matrix is an important information source and can be used for a variety of data mining tasks, including data clustering, visualization of multidimensional data, outlier detection, substitution of missing values, and finding mislabeled data samples. The proximity matrix can also be used as a kernel matrix in SVM, as it is suggested in this work. Therefore, it is important to know how good the estimate of data proximity obtained from the RF software (Breiman & Cutler, 2004) is.

2. Background

2.1. Weak learners

A weak learner is a predictor that has low bias, however, the performance is usually not accurate due to high variance (Breiman & Cutler, 2004). Let us assume that given is a data set $\mathcal{X}_t = \{(\mathbf{x}_m, y_m),\ m = 1, \ldots, M\}$, where $\mathbf{x}_m$ is an input vector and y_m is the target output. A weak learner, a predictor $f(\mathbf{x}, \mathcal{X}_t)$, can be created using the set $\mathcal{X}_t$. By randomly sampling from the set $\mathcal{X}_t$, a collection of weak learners $f(\mathbf{x}, \mathcal{X}_t, \theta_k)$ can be created, where θ_k is the independent and identically-distributed, i.i.d., random vector selecting data points for the kth weak learner.

Combining i.i.d. randomized weak learners into an ensemble by averaging, leaves the bias approximately unchanged while reduces the variance by a factor of $\bar{\rho}$ equal to the mean value of the correlation between the weak learners (Breiman, 2001). This implies that by keeping correlation and bias of weak learners low, generalization performance may be improved.

2.2. Random forest

RF is an ensemble of weak learners, classification and regression trees (CART). When solving classification problems, the RF

* Corresponding author.
E-mail addresses: Cristofer.Englund@viktoria.se (C. Englund), antanas.verikas@hh.se (A. Verikas).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.05.094

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 more trees are added to RF, the generalization error converges to a limiting value, thus there is no over-fitting in large RFs (Breiman, 2001).

Low bias and low correlation are essential for accuracy. To get low bias, trees are grown to maximum depth without pruning. To achieve low correlation, randomization is applied:

i. Each tree of RF is grown on a bootstrap sample of the training set. By applying bootstrap sampling to generate θ_k, about two-thirds of the data points are used by each weak learner. About one-third of the data are out of the bootstrap sample or out-of-bag (OOB) and are used to test the generalization performance (OOB error).
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 (Breiman, 2001; Breiman & Cutler, 2004) starting with $n = \lfloor \log_2(N) + 1 \rfloor$ or $n = \sqrt{N}$ and then varying n until the minimum OOB error is obtained. At each node, only one variable, providing the best split, is used out of the n selected.

To obtain the proximity 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 (Breiman & Cutler, 2004).

3. Proposed technique

Proximity of two observations $\mathbf{x}_i$ and $\mathbf{x}_j$ estimated using only one tree attains a value of zero or one. This a very rough–binary–measure. As the number of trees in RF grows, the proximity estimate becomes more accurate due to averaging. A large number of trees is necessary to get stable estimates of data proximity (Liaw & Wiener, 2002). It is not uncommon, however, that random forests of rather few trees are used. A more elaborate estimate of data proximity is needed in such cases.

A novel approach to estimate data proximity in random forests is proposed in this work. The approach is based on measuring distance between two terminal nodes of a decision tree occupied by observations $\mathbf{x}_i$ and $\mathbf{x}_j$. We suggest assessing proximity of observations $\mathbf{x}_i$ and $\mathbf{x}_j$ according to the following equation:

$$p_{ij} = \frac{1}{K}\sum_{k}^{K} 1/(e^{w \cdot g_{ijk}}) \quad (1)$$

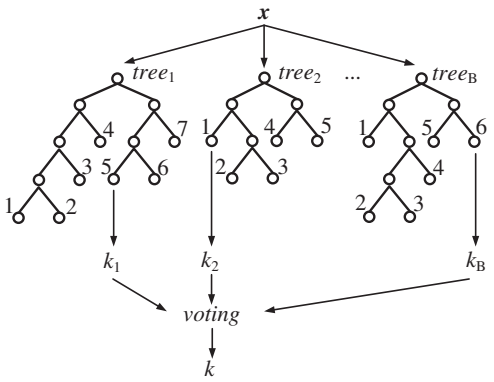

Fig. 1. A general architecture of a random forest.

Table 1
Data sets used in the experiments.

Data set	# dimensions	# samples	#positives	#negatives
Artificial data	2	200	100	100
Australian credit	14	690	307	383
Heart disease	13	270	120	150
German credit	24	1000	300	700

where k runs over the K trees, for which both $\mathbf{x}_i$ and $\mathbf{x}_j$ are among the OOB samples, w is a parameter, and g is the number of tree branches between the two terminal nodes occupied by $\mathbf{x}_i$ and $\mathbf{x}_j$. For example, $g = 3$ between the terminal nodes 1 and 3 of $tree_1$ in Fig. 1. If $\mathbf{x}_i$ and $\mathbf{x}_j$ occupy the same terminal node, then $g = 0$ and p_{ij} will be increased by one as in the original way to assess data proximity. The parameter w controls the influence of the distance between two terminal nodes occupied by the observations on the proximity values.

We propose a novel way of assessing the quality of data proximity measure. To assess the quality (consistency) of data proximity estimate, we suggest using the proximity matrix as a kernel matrix in a support vector machine (SVM). In the sequel, we call such type of kernel by proximity based kernel (PBK). We assume here that a data proximity matrix of higher quality leads to higher classification accuracy. It is worth observing that only features used by RF contribute to PBK. Noisy features that are not used by RF have no influence on PBK.

Thus, to assess the quality of a data proximity matrix, we train an SVM using the matrix as a kernel and evaluate the generalization error of the SVM. The lower the error, the higher is the quality of the matrix. In the same manner we find the optimal value of the parameter w—by varying w and using the proximity matrix in a PBK based SVM. Aiming to assess the quality of PBK, we also use an SVM based on the radial basis function (RBF) kernel and an SVM exploiting the standard proximity matrix (SPBK).

4. Experimental investigations

Four data sets, one artificial and three taken from the UCI[1] machine learning data base, have been used to test the proposed approach. Table 1 gives a summary of the data sets used in the experiments. All the sets concern a two-class (positive/negative) classification problem.

For each of the four data sets, we use RFs of 5, 10, and 100 trees. In all the experiments, we train SVM with three different kernels, RBF, PBK, and SPBK. The results presented here are obtained from a fivefold cross validation procedure. To enable condign comparison between the different SVMs, a search for the optimal SVM hyper-parameters is made. For the RBF SVM the width of the radial basis function γ and the cost parameter C are chosen experimentally. The parameter C controls the trade-off between the training error and the rigid margins allowing some misclassifications. For the PBK and SPBK SVMs the C and w parameters are chose experimentally.

4.1. Performance measure

To compare the results, we estimate the *True positives* (TP) rate, *True negatives* (TN) rate, *False positives* (FP) rate, and the *False negatives* (FN) rate. Thus, the correct classification rate CR in this work is estimated as:

$$CR = \frac{TP + TN + (100 - FP) + (100 - FN)}{4} \quad (2)$$

[1] http://archive.ics.uci.edu/ml/.

4.2. Artificial data

The artificial data are generated from two normal distributions with $\sigma = 0.6$, and centers at x_{c1}, $y_{c1} = (1,2)$ and x_{c2}, $y_{c2} = (-1.2, -0.5)$. The classes are perfectly balanced. Figs. 2–4 present the average generalization error obtained from the SVM based on the PBK and SPBK kernels created by RF of 5, 10, and 100 trees, respectively. Results obtained for the ordinary RBF kernel are also shown for the sake of comparison.

It is worth noting that the PBK kernel created by RF of 5 and 100 trees outperforms the ordinary RBF kernel. As expected, for small number of trees, the PBK kernel outperforms the SPBK one at some values of parameter w. Thus, the proposed technique improves the quality of data proximity estimation through selection of appropriate value of the parameter w.

4.3. Australian credit approval

The same pattern of behaviour of the three kernels was also observed for the *Australian credit approval* data set. The average generalization error obtained from the SVM based on the PBK and SPBK kernels created by RF of 5, 10, and 100 trees, respectively are shown in Figs. 5–7. Again, the PBK kernel outperformed the standard RBF kernel, see Figs. 6 and 7. For a small number of trees, the PBK kernel was significantly better than the SPBK one, see Fig. 5. As can be seen from Figs. 5 and 6, the classification accuracy obtained using the SPBK created of 5 trees is much lower than the accuracy achieved using information from 10 trees. Thus, for this data set, the RF consisting of five trees does not contain enough information to reflect data similarities well.

4.4. Heart disease

Figs. 8–10 present the average generalization error obtained for the *Heart disease* data set from the SVM based on the PBK and SPBK kernels created by RF of 5, 10, and 100 trees, respectively. For this data set, RBF kernel is significantly better than PBK and SPBK created using a small number of trees, see Figs. 8 and 9. However, when the number of trees is relatively large, the PBK kernel outperforms the RBF one, see Fig. 10. Thus, small forests do not contain enough information for assessing data proximities well. The PBK kernel outperforms the SPBK one in all the three experiments.

4.5. German Credit

The test results for the *German Credit* data set are shown in Figs. 11–13. For this data set, both the PBK- and the SPBK-based SVMs performed better than the RBF-based SVM for large as well as small

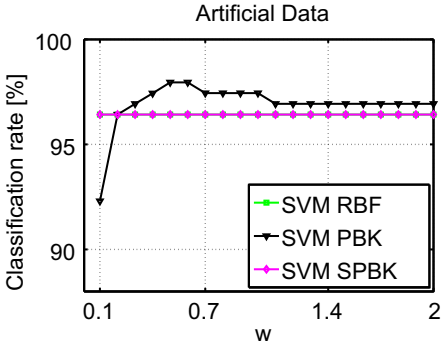

Fig. 2. The average generalization error obtained from the SVM based on the RBF kernel, and the PBK and SPBK kernels created by RF of 5 trees.

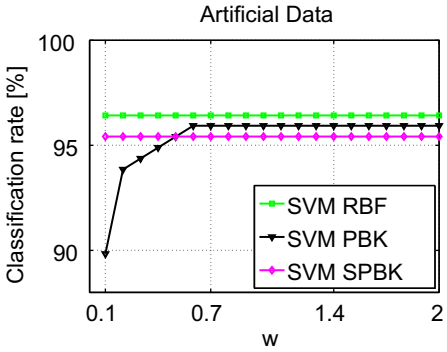

Fig. 3. The average generalization error obtained from the SVM based on the RBF kernel, and the PBK and SPBK kernels created by RF of 10 trees.

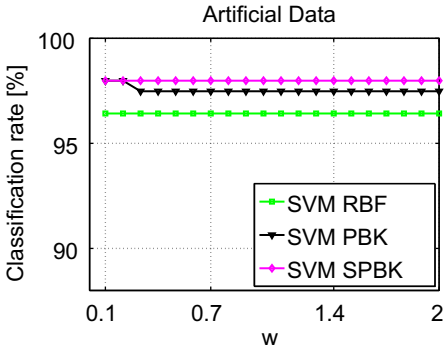

Fig. 4. The average generalization error obtained from the SVM based on the RBF kernel, and the PBK and SPBK kernels created by RF of 100 trees.

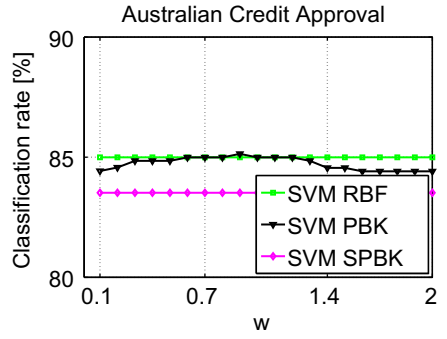

Fig. 5. The average generalization error obtained from the SVM based on the RBF kernel, and the PBK and SPBK kernels created by RF of 5 trees.

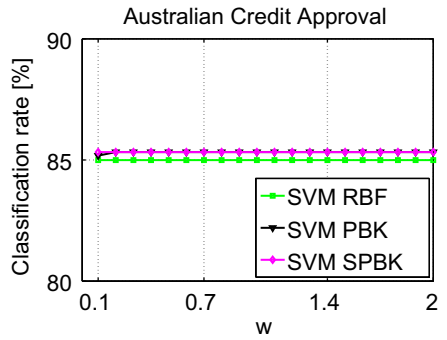

Fig. 6. The average generalization error obtained from the SVM based on the RBF kernel, and the PBK and SPBK kernels created by RF of 10 trees.

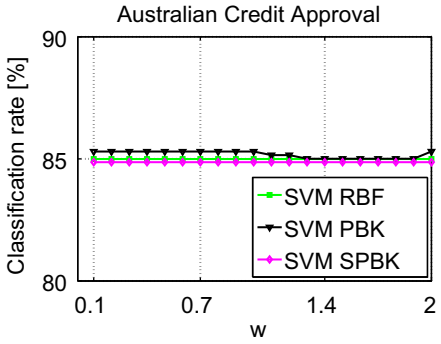

Fig. 7. The average generalization error obtained from the SVM based on the RBF kernel, and the PBK and SPBK kernels created by RF of 100 trees.

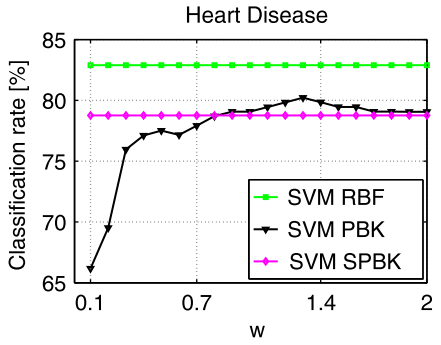

Fig. 8. The average generalization error obtained from the SVM based on the RBF kernel, and the PBK and SPBK kernels created by RF of 5 trees.

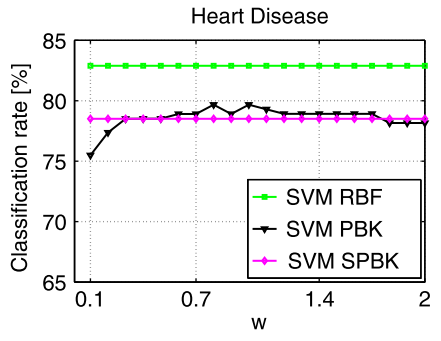

Fig. 9. The average generalization error obtained from the SVM based on the RBF kernel, and the PBK and SPBK kernels created by RF of 10 trees.

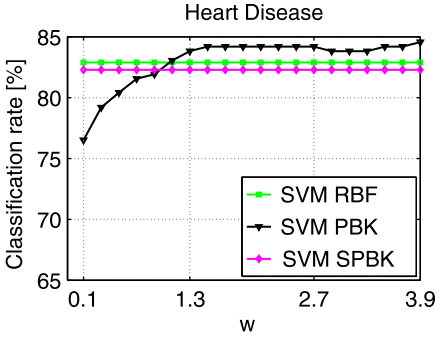

Fig. 10. The average generalization error obtained from the SVM based on the RBF kernel, and the PBK and SPBK kernels created by RF of 100trees.

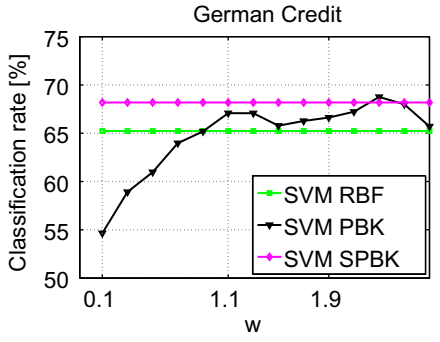

Fig. 11. The average generalization error obtained from the SVM based on the RBF kernel, and the PBK and SPBK kernels created by RF of 5 trees.

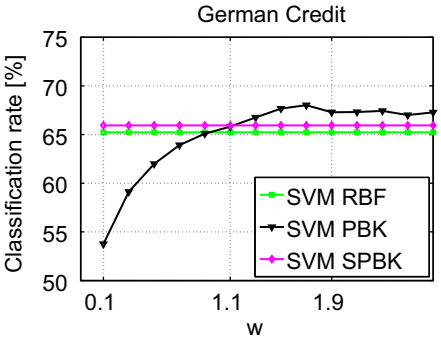

Fig. 12. The average generalization error obtained from the SVM based on the RBF kernel, and the PBK and SPBK kernels created by RF of 10 trees.

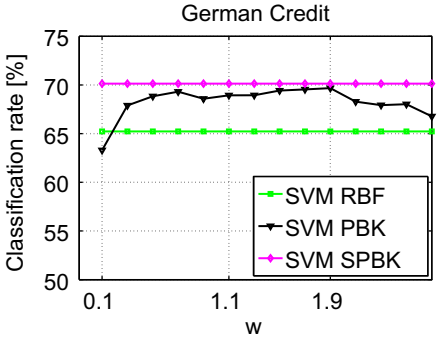

Fig. 13. The average generalization error obtained from the SVM based on the RBF kernel, and the PBK and SPBK kernels created by RF of 100 trees.

forests. As expected, for small number of trees, the PBK-based SVM performed better than the SPBK one, see Figs. 11 and 12.

5. Conclusions

We proposed a novel approach to data proximity estimation in random forests. The approach is based on measuring distance between two terminal nodes of a decision tree occupied by observations being compared. To assess the quality of data proximity estimate, we suggested using the proximity matrix as a kernel matrix in a support vector machine classifier.

It was demonstrated experimentally that the proposed technique improves the data proximity estimate, especially when random forests are made of a small number of trees. This behaviour was observed for all the four data sets used in the experiments.

It was also demonstrated that, for some tasks, an SVM exploiting the suggested proximity matrix based kernel, can outperform an SVM based on the standard radial basis function kernel. It is worth observing that only features used by a random forest contribute to the proximity matrix based kernel. Noisy features that are not used by the random forest have no influence on the kernel.

References

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Verikas, A., Gelzinis, A., & Macauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition, 44*, 330–349.

Granitto, P., Gasperi, F., Biasioli, F., Trainotti, E., & Furlanello, C. (2007). Modern data mining tools in descriptive sensory analysis: A case study with a random forest approach. *Food Quality and Preference, 18*, 681–689.

Gislason, P., Benediktsson, J., & Sveinsson, J. (2006). Random forest for land cover classification. *Pattern Recognition Letters, 27*, 294–300.

Harb, R., Yan, X., Radwan, E., & Su, X. (2009). Exploring prechrash maneuvers using classification trees and random forests. *Accident Analysis and Prevention, 41*, 98–107.

Larivière, B., & den Poel, D. V. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications, 29*(2), 472–484.

Ramírez, J., Górriz, J., Segovia, F., Chaves, R., Salas-Gonzalez, D., López, M., et al. (2010). Computer aided diagnosis system for the alzheimer's disease based on partial least squares and random forest SPECT image classification. *Neuroscience Letters, 472*, 99–103.

Özçift, A. (2011). Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. *Computers in Biology and Medicine, 41*, 265–271.

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications, 39*(2), 1772–1778.

Son, J.-D., Niu, G., Yang, B.-S., Hwang, D.-H., & Kang, D.-S. (2009). Development of smart sensors system for machine fault diagnosis. *Expert Systems with Applications, 36*(9), 11981–11991.

Mattieu, S., & Heyen, G. (2010). Performance monitoring of an industrial boiler: classification of relevant variables with random forests. *Computer Aided Chemical Engineering, 28*, 403–408.

Breiman, L., Cutler, A. (2004). RFtools—for predicting and understanding data, Technical Report, Berkeley University, Berkeley, USA (April 2004).

Liaw, A., & Wiener, M. (2002). Classification and regression by random forests. *R News, 2*(3), 18–22.