Using Random Forests for Data Mining and Drowsy Driver Classification Using FOT Data

Cristofer Englund[1], Jordanka Kovaceva[2],
Magdalena Lindman[2], and John-Fredrik Grönvall[2]

[1] Viktoria Institute
Lindholmspiren 3A
S-417 56 Gothenburg, Sweden
cristofer.englund@viktoria.se
[2] Volvo Car Coorporation
S-405 31 Gothenburg, Sweden
{jkovace1,mlindman,jgronval}@volvocars.com

Abstract. Data mining techniques based on Random forests are explored to gain knowledge about data in a Field Operational Test (FOT) database. We compare the performance of a Random forest, a Support Vector Machine and a Neural network used to separate drowsy from alert drivers. 25 variables from the FOT data was utilized to train the models. It is experimentally shown that the Random forest outperforms the other methods while separating drowsy from alert drivers. It is also shown how the Random forest can be used for variable selection to find a subset of the variables that improves the classification accuracy. Furthermore it is shown that the data proximity matrix estimated from the Random forest trained using these variables can be used to improve both classification accuracy, outlier detection and data visualization.

Keywords: Data mining, Random Forest, Drowsy Driver Detection, Proximity, Outlier detection, Variable selection, Field operational test.

1 Introduction

Data mining is used to extract implicit, interesting, and previously unknown information that may be hidden within large data sets [1,2,3]. Both regression and classification methods are used in data mining and methods involving, Neural Networks [4], Support Vector Machines (SVM) [5], k-NN [6], Classification and Regression Trees (CART) [7] and Self-Organizing Maps [8] are amongst the most popular.

This paper presents investigations on how some of these data mining methods can be applied to a recently established Field Operational Test (FOT) database. The database consists of vehicle Controller Area Network (CAN) data and video material from cameras in the vehicle both facing the road and the driver from 100 cars operated during normal conditions. The establishment of the FOT database

R. Meersman et al. (Eds.): OTM 2012, Part II, LNCS 7566, pp. 752–762, 2012.

is part of the European 7th Framework Program within the euroFOT (European Large-Scale Field Operational Test on In-Vehicle Systems)[1] project. One of the objectives of the euroFOT project is to assess the capabilities and performance of intelligent vehicle systems (IVS) and to observe the behaviour of the driver while interacting with those systems.

Typically Driver Drowsiness Warning (DDW) systems monitor a number of alertness indicators and establishes a driver profile of each driver. As the behaviour deviates from this profile, the system will warn the driver. The DDW technology is designed to alert tired and distracted drivers. This function steps in at 65 km/h and monitors the car's progress between the lane markers and warns the driver if his or her driving pattern changes in a random or uncontrolled way. Generally, the data used in such systems are for example changes in steering angle [9,10]. In [11] drowsy measures are divided into three categories; Subjective measures; physiological measures; vehicle-based measures. Subjective measures involves self-assessment by the driver. The driver is asked to rate their sleepiness on some scale. The most common ones are the 7-point Standford Sleepiness Scale (SSS) or the 9-point Karolinska Sleepiness Scale (KSS). Physiological measures involves monitoring brain activity, heart rate or eye movements. In particular, electroencephalographic (EEG) and eyeblink (eye closure) duration are amongst the most common measures. Vehicle-based measures used in DDW systems are typically speed variation, standard deviation of lane position and steering wheel movement [10,11].

In this work we utilize the information in the FOT database to create a Random Forest (RF) that can be used to separate drowsy from alert driver behaviour. We also compare the classification performance of a RF, a SVM and a NN used to separate drowsy from alert drivers. Furthermore, based on the RF, data proximity is estimated and used to find outliers and to visualize the high dimensional data.

1.1 Random Forests

Recently RF, that are formed by an ensemble of CARTs, was proposed by Breiman [12]. RF are used for both classification and regression and they incorporate methods to estimate variable importance and data proximity [13,14]. Given a data set $\mathbf{X}$, each tree in the RF is trained using a bootstrap sample from the data. Bootstrapping implies creating a vector χ_k that describes the samples to be drawn from the training data with replacemen. For each tree, the data not used for training, the out of bag (OOB) data, can be used to test the generalization performance (OOB error) and it may also be used to estimate variable importance. On average each data point is in the OOB set about two-thirds of the time.

For each tree k the random vectors that selects the bootstrapping data χ_k and the variables used for splitting Θ_k are used to create a predictor $f(\mathbf{x}, \chi_k, \Theta_k)$. Θ_k consists of a number of independent random variables m between 1 and M and by

[1] `http://www.eurofot-ip.eu/`

varying the number of variables used different generalization performance may be achieved, however, starting the search from $m = \lfloor \log_2(M) + 1 \rfloor$ or $m = \sqrt{M}$ is often suggested [13,14]. As more trees are added to the RF the generalization error will converge to a limiting value and thus there is no risk of over-fitting in RF as the number of trees grows large [13].

Figure 1 shows a general design of a RF. The classifiers $f(\mathbf{x}, \chi_k, \Theta_k)$ outputs a class, given the input $\mathbf{x}$, and the most popular class among the trees is the output of the RF. A description of the data used to train the RF in this paper is found in Section 2.

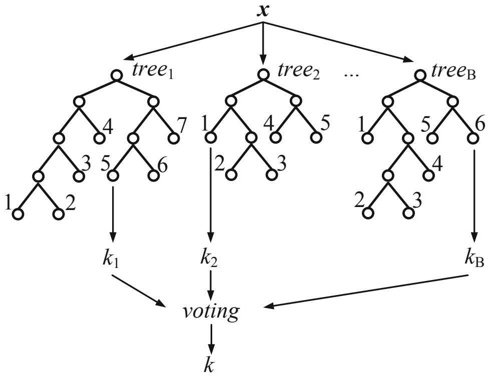

Fig. 1. General design of RF

1.2 Variable Importance

After training a RF, variable importance may be assessed. Variable importance is estimated by monitoring the OOB generalization error while permuting the observed variable whereas the others are left unchanged, after testing all variables, the variable that increases the error the least is the least important and is eliminated. This procedure is repeated until only one variable is left. There are at least three other methods for estimating the variable importance from RF [14,15].

1.3 Proximity

Proximity is a valuable measure and may be used for outlier detection, the detection of mislabeled data, to replace missing values, and as input to visualization methods such as Multi Dimensional Scaling [16] or t-Distributed Stochastic Neighbor Embedding (t-SNE)[17].

In [14] the data proximity matrix is estimated as the RF is generated. For each tree grown, the data is run down the tree and if two samples x_i and x_j activate the same leaf node, prox(i,j) is increased by one. When the RF is generated, the proximities are divided by the number grown trees.

2 Data Description

The FOT data contains more than 300 variables, e.g. wiper speed, brake pedal position, engine torque and cruise control activation signal. The data is stored with 10 Hz sampling frequency during the trip. Data from 100 cars has been collected under normal driving conditions during 18 months. Seven variables, that enable monitoring of drivers' actions are used in this work. Below is a description of the variables:

y — mDIMONWarning is a vehicle warning signal generated from a system designed to alert tired and distracted drivers.
x_1 — mInCarTemp is the temperature inside coupé of the car.
x_2 — mVehicleSpeed is the vehicle speed.
x_3 — mAnyLaneOffset is the shortest distance to any of the the lane markers min(d_l,d_r) where d_l and d_r are the distances to the left and right lane respectively.
x_4 — mSteeringAngle is the steering angle.
x_5 — mLateralAcc is the lateral acceleration of the vehicle.
x_6 — mYawRate is the yaw rate of the vehicle.
x_7 — mAccelPedalPos is the acceleration pedal position.

To be able to capture the dynamics of the drivers the standard deviation of $x_2 - x_7$ during three different time frames was estimated. Time frames of 30s, 60s, and 120s, showed the best initial performance. These variables are denoted D300, D600, D1200 respectively. In total, there are 25 input data for training the RF.

3 Experimental Investigations

The RF is trained using the data described in Section 2. In order to demonstrate the performance of the RF the results are compared with two well known data mining methods described shortly.

3.1 Experimental Setup

The data samples in this work are collected from the FOT data base and have been carefully annotated by watching video recordings of the drivers in order to confirm the state of the driver. The annotation procedure involved classifying driver behaviour at warnings in groups of, distracted or drowsy drivers or driving in a way that caused the system to produce a warning. The final training data consists of 15 trips where the driver was drowsy and additionally 6 trips where the driver was alert during the whole trip, all trips are validated (annotated) by an expert.

From the annotated data, the training data was selected according to the following procedure:

1. Locate time index for the drowsy driver warnings generated by the internal vehicle system.
2. Select 20 samples with 5 seconds interval around (before and after) the drowsy detection warning. Assign these samples $y = 1$.
3. Select 100 samples uniformly distributed during an interval of 15-5 minutes before the drowsy warning. Assign these samples $y = 0$.
4. In files without any drowsy warnings 100 samples was selected uniformly distributed during the whole trip. Assign these samples $y = 0$.

The data was divided into two parts, train and validation (75%) and test (25%).

3.2 Drowsy Driver Classification

We compare the results from the RF with a SVM and a NN , since they are amongst the most popular, and the most powerful data mining methods available. The train and validation part of the data was used to perform ten-fold cross validation in order to find the parameters of the SVM and NN models. The RF consists of 100 trees and m was set to $\sqrt{M}$.

The SVM is constructed with a Radial basis-based kernel and the best parameters was found to be $\sigma = 0.05$ and $C = 1$. The NN is a Multilayer perceptron (MLP) with three nodes in the hidden layer. A Hyperbolic tangent sigmoid and a Logarithmic sigmoid transfer function was used for the for the hidden and output layers respectively. The Bayesian regularization learning rule was used for training the NN. As can be seen, the RF has significantly higher accuracy than the SVM and the NN.

A comparison between the generalization performance, mean and standard deviation of the ten-fold cross validation test error, between the RF, a SVM and a NN is shown in Table 1. The training and validation data is used for training the models using ten-fold cross validation and the test data is used for testing the generalization performance (cross validation test error).

Table 1. Mean ($\overline{m}$) and standard deviation of mean ($\sigma_{\overline{m}}$) of the ten-fold cross validation test error of the RF, a SVM and a NN trained using all 25 variables

Model	$\overline{m}$	$\sigma_{\overline{m}}$
Random forest	0.017	0.033
SVM	0.0238	0.1463
NN	0.0297	0.1498

3.3 Variable Selection

The variable importance measure described in Section 1.2 was applied in order to select only the most important variables. The first variable that is removed, the least significant, is assigned one credit score. For each variable being removed the assigned credit score is incremented by one. Thus the last variable, the most significant, is given the highest score. According to Figure 2, on average the best performance in this application is achieved by removing 20 of the 25 variables. The variable selection procedure was repeated 10 times and the mean results can be found in Table 2. It was found that the most important variables are the ones incorporating dynamics (variation) of the driver, thus, the standard deviation of the signals from a 30, 60 and 120 seconds window. The current in-vehicle temperature and vehicle speed (x_1 and x_2) are the two signals that has the highest score among the signals describing one time instance. These two signals have some inertness built in, thus they are less noisy and therefore they are ranked higher than the signals $x_3 - x_7$. The signals $x_3 - x_7$ along with $mAccelerationPedalPos_{D300}$ are the least significant variables.

Table 2. Mean score of variable importance estimation

Rank	Score	Variable name	Rank	Score	Variable name
1	23.4	$mLateralAcc_{D300}$	14	12.0	$mYawRate_{D600}$
2	19.7	$mVehicleSpeed_{D1200}$	15	9.1	$mAnyLaneOffset_{D1200}$
3	19.0	$mAccelPedalPos_{D1200}$	16	8.7	$mYawRate_{D300}$
4	18.7	$mSteeringAngle_{D600}$	17	8.0	$mLateralAcc_{D600}$
5	18.5	$mVehicleSpeed_{D600}$	18	7.7	$mAnyLaneOffset_{D300}$
6	18.3	$mSteeringAngle_{D1200}$	19	7.0	$mAccelPedalPos_{D600}$
7	18.0	$mVehicleSpeed$	20	6.9	$mYawRate$
8	17.7	$mInCarTemp$	21	5.2	$mSteeringAngle$
9	17.3	$mYawRate_{D1200}$	22	4.6	$mAccelPedalPos_{D300}$
10	14.8	$mAnyLaneOffset_{D600}$	23	3.6	$mAccelPedalPos$
11	12.8	$mSteeringAngle_{D300}$	24	2.5	$mAnyLaneOffset$
12	12.8	$mLateralAcc_{D1200}$	25	1.1	$mLateralAcc$
13	12.6	$mVehicleSpeed_{D300}$			

Table 3 shows the comparison between the RF, SVM and NN modes used to separate the drowsy from alert drivers trained using the variables with the highest score in Table 2. Using ten-fold cross validation the same model parameters as when all variables was available gave the best generalization performance for the SVM and the NN. As can be seen, the RF outperforms both the SVM- and the NN-based classifiers.

3.4 Visualization

t-SNE[17] is a fast and effective visualization method that allows 2D visualization of high dimensional data. It takes as input a pairwise similarity matrix and is able to capture both local structure while also apprehending global structures, that may form clusters in several dimensions in the high dimensional data. A RF may be used to create a proximity matrix from the data, see Section 1.3, which can be

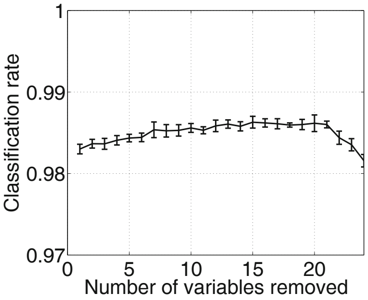

Fig. 2. Mean RF classification rate for the variable importance estimation

Table 3. Mean ($\overline{m}$) and standard deviation of mean ($\sigma_{\overline{m}}$) of the ten-fold cross validation test error of the RF, a SVM and a NN trained using the five variables with the highest score in Table 2

Model	$\overline{m}$	$\sigma_{\overline{m}}$
Random forest	0.0138	0.0010
SVM	0.0271	0.1524
NN	0.0259	0.1466

used as input for the t-SNE method. The left part of Figure 3 visualizes the data using t-SNE and the right part visualizes the data using Principal Component Analysis (PCA). For this we calculate eigenvalues λ_i ($\lambda_1 > \lambda_2 > ... > \lambda_m$) and the associated eigenvectors $\mathbf{u}_i$ of the covariance matrix $\mathbf{C} = \frac{1}{N}\sum_{j=1}^{N} \mathbf{x}_j \mathbf{x}_j^T$, where N is the number of samples.

We project the $N \times m$ matrix $\mathbf{X}$ of the vectors $\mathbf{x}$ onto the first P eigenvectors $\mathbf{u}_k$, $\mathbf{A} = \mathbf{XU}$. By comparing the 2D plots, see Figure 3, we clearly see that the proximity-based t-SNE approach provides better separation capability than PCA between the drowsy and alert drivers. For both plots in Figure 3 drowsy drivers, indicated by red '$\diamond$' (the samples from around to the drowsy detection warning), are grouped together, however it is difficult to separate them from the alert drivers for the data projected onto the eigenvectors of the covariance matrix belonging to the 1st and 2nd largest eigenvalues (right part of Figure 3). The data indicated by green '$*$' are the drivers that are alert during the whole trip, and the data indicated by the blue '$+$' are the data from the drowsy driver 15-5 minutes before the warning.

The left part of Figure 4 visualizes the data proximities using t-SNE and the right part of Figure 4 visualizes using PCA. The variables used are the five with the highest score in Table 2. t-SNE generates more distinct clusters when using only the five variables with the highest importance score, compared to when all variables are used to estimate the data proximity. Applying the PCA-analysis to the five variables with highest score does not improve the separation of alert and drowsy drivers, see right part of Figure 4.

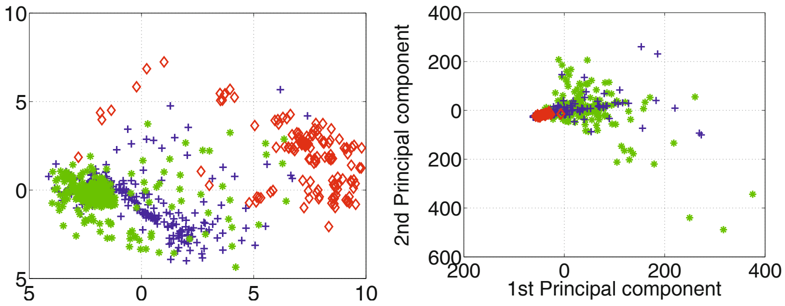

Fig. 3. *Left:* Visualizing using t-SNE generated from the proximity matrix generated from the RF trained using all available variables. *Right:* Visualizing the data using first and second Principal component estimated from all available training data.

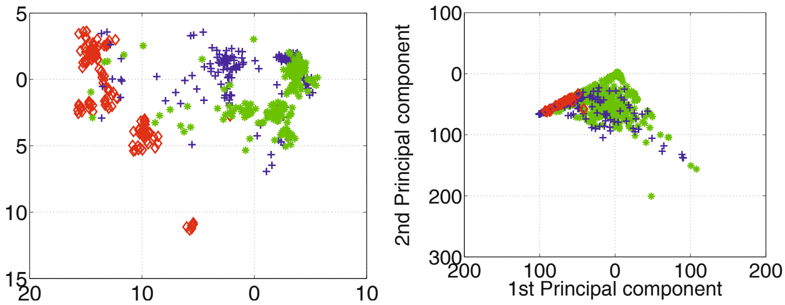

Fig. 4. *Left:* Visualizing using t-SNE generated from the proximity matrix generated from the RF trained using the five variables with the highest score in Table 2. *Right:* Visualizing the data using first and second Principal component estimated from the five variables with the highest score in Table 2.

3.5 Outlier Detection

Outliers can be seen as extreme values, appearing at the outskirts of the data set. The proximity matrix may be used to estimate an outlier value ρ of the data, see Equation 1. Figure 5 shows the outlier value [14] ρ.

$$\rho_n = 1/\sum_k (prox(x_n, x_k), n \neq k \tag{1}$$

Proximity indicate how similar two samples are, therefore, if one sample differ from the others a high ρ is obtained. By comparing the data in the left, and right parts of Figure 5 and Figure 6 we see that samples that are assigned a high ρ in Figure 5 are also appearing as outliers in Figure 6.

Figure 6 visualizes the data using t-SNE estimated from the proximity matrix generated by the RF trained using all variables (left) and trained using the five

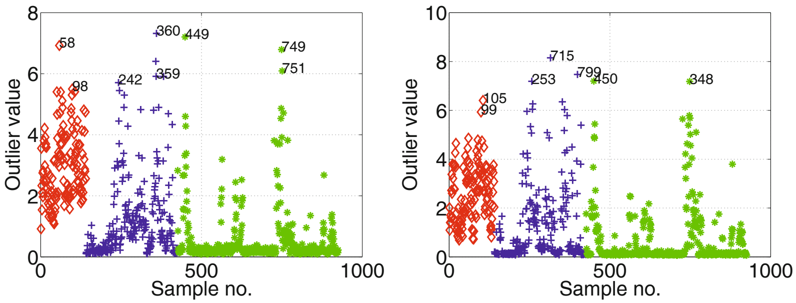

Fig. 5. *Left:* Plot of outlier value for all training data using all variable available. *Right:* Plot of outlier value for all training data using the five variables with the highest score in Table 2.

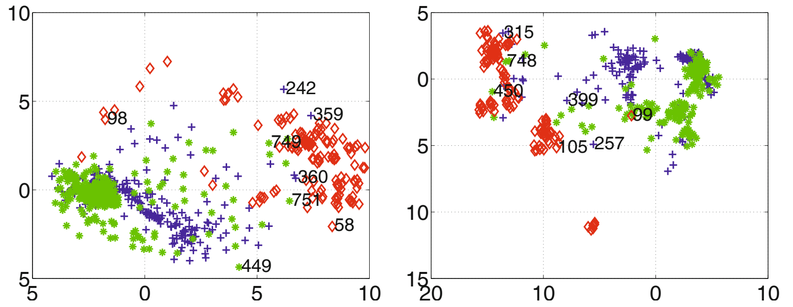

Fig. 6. *Left:* Visualizing using t-SNE that use the proximity matrix from the RF trained using all available variables. *Right:* Visualizing using t-SNE that use a proximity matrix from the RF trained using the five variables with the highest score in Table 2.

with the highest score in Table 2 (right). The samples are numbered to be able to track them. The samples indicated by red diamonds are the drowsy drivers. As can be seen, in Figure 6, we find samples indicated by blue '+' (possibly mislabeled samples) and green '$*$' (possibly alert drivers that behave like drowsy drivers) that mix with the samples from drowsy drivers. Additionally, one of the drowsy driver samples, indicated by red '$\diamond$' appears close to the center of the alert drivers - further investigation of these samples can be made if necessary.

4 Conclusion and Future Work

We have shown how a RF may be used to gain knowledge about data from a recently established FOT data base. The RF is used to classify the state of a driver as being either alert or drowsy. For training, two classes of driver behaviour was available from 21 trips. The generalization test error showed that the method was capable of separating the two classes with very few misclassifications. It is further shown that the RF outperforms both a SVM, and a MLP NN.

While comparing the performance between different algorithms the No Free Lunch Theorem [18] is often refered to, saying there can be no algorithm that is always better than another one. In [19] for example, the authors compared RF and SVM and it was found that the performance of the algorithms was highly problem dependent—for some problems RF was significantly better while for others SVM significantly outperformed RF. Several data complexity measures aiming to find out for which data types RF is more suitable than SVM and vice versa was also applied however the relations was found to be very unclear [19]. Nevertheless, in this work we conclude that the classification problem at hand is well suited for RF.

A variable selection method based on the RF was applied and by using only the five variables with the highest importance score the classification rate for separating drowsy from alert driver behaviour was improved.

By utilizing the data proximity matrix generated from the RF it is also possible to visualize the data. By using t-SNE, significantly better visual separation performance was achieved than traditional PCA. Both using all available variables and using only the five variables with the highest importance score. We also show how an outlier measure, based on the data proximity matrix, can be used to find samples that do not appear as expected, e.g. possibly mislabeled data.

To fully understand the potential of this method it needs to be validated using more data. Moreover, further development of the method could incorporate an iterative process where new drowsy drivers found by the proposed method may be added to the training and validation data. Thus, the new model will incorporate more knowledge about drowsy drivers. This procedure could be repeated until the false positives rate is sufficiently low.

References

1. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann Publishers Inc., San Francisco (2005)
2. Shneiderman, B.: Inventing Discovery Tools: Combining Information Visualization with Data Mining. In: Abe, N., Khardon, R., Zeugmann, T. (eds.) ALT 2001. LNCS (LNAI), vol. 2225, p. 58. Springer, Heidelberg (2001)
3. Zhu, D.: A hybrid approach for efficient ensambles. Decision Support Systems 48, 480–487 (2010)
4. Bishop, C.: Pattern Recognition and Machine Learning. Springer, Singapore (2006)
5. Vapnik, V.: Statistical Learning Theory. Whiley, New York (1998)
6. Devroye, L., Gyorfi, L., Krzyzak, A., Lugosi, G.: On the strong universal consistency of nearest neighbor regression function estimates. Annals of Statistics 22, 1371–1385 (1994)
7. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Wadsworth and Brooks, Monterey (1984)

8. Kohonen, T.: Self-Organizing Maps. Springer, Berlin (1995) (Second Extended Edition 1997)
9. Lesemann, M.: Testing and evaluation methods for ict-based safety systems, deliverable D1.1: State of the art and evalue scope. Technical report, eValue project (2008), http://www.evalue-project.eu/pdf/evalue-080402-d11-v14-final.pdf
10. Kircher, A.: Vehicle control and drowsiness. VTI Meddelande 922A, Swedish National Road Transport Resesarch Institute, Linköping (2002)
11. Liu, C.C., Hosking, S.G., Lenné, M.G.: Predicting driver drowsiness using vehicle measures: Recent insights and future challenges. Journal of Safety Research 40, 239–245 (2009)
12. Breiman, L.: Bagging predictors. Machine Learning 24, 123–140 (1996)
13. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
14. Breiman, L., Cutler, A.: RFtools—for predicting and understanding data, Technical Report. Berkeley University, Berkeley, USA (2004)
15. Breiman, L.: Manual on setting up, using, and understanding random forests v3.1. Berkeley University, Berkeley (2002)
16. Kruskal, J., Wish, M.: Multidimensional scaling. Quantitative applications in the social sciences. Sage Publications (1978)
17. van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-SNE. Journal of Machine Learning Research 9, 2579–2605 (2008)
18. Wolpert, D.H., Macready, W.G.: No free lunch theorems for search. Technical Report SFI-TR-05-010, Santa Fe Institute (1995)
19. Verikas, A., Gelzinis, A., Bacauskiene, M.: Mining data with random forests: A survey and results of new tests. Pattern Recognition 44, 330–349 (2011)