Volume 7, Issue 1, January 2017 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Random Forest: A Review

Eesha Goel*, Er. Abhilasha
Computer Science & Engineering &GZSCCET Bhatinda,
Punjab, India

DOI: 10.23956/ijarcsse/V7I1/01113

Abstract— Ensemble is a data mining technique composed of number of individual classifiers to classify the data to generate new instances of data. Random Forest is the most popular ensemble technique of classification because of the presence of excellent features such as Variable importance measure, Out-of-bag error, Proximities etc. In this paper, the developments and improvements of Random Forest in the last 15 years are presented. This paper deals with the approach proposed by Brieman since 2001. This paper also presents the description of usage of Random Forest in various fields like Medicine, Agriculture, Astronomy, etc.

Keywords— Meta Random Forest, RelieF Random Forest, Dynamic Integration of Random Forest, Small Random Forest, Forest-RK, BAGA Algorithm, Dynamic Random Forest, CUDA.

I. INTRODUCTION

In the current century, numerous of classification problems are observed that consist of large amount of data. Most commonly used algorithms are failed to classify data with better accuracy. Therefore, Random forest is introduced for the classification and regression of large amount of data. Random forest is an ensemble Machine Learning Technique. Machine Leaning Techniques are one of the applications of Data Mining. Data Mining Techniques are classified into two groups that are described below:

- Descriptive Data Mining Techniques
- Predictive Data Mining Techniques.

Descriptive Data Mining Technique focuses on describing, summarizing and grouping the data into categories.

This technique is implemented using unsupervised machine learning techniques [Khulkarni and Sinha (2013)].

Predictive Data Mining Techniques focuses on analyzing the past data and predicts the conclusions for future. It is based on the classical model building process. The Predictive Model is based on the analysis of features of Predictor variables. One or more than one are considered as the Predictor variable. The output is generated in the form of function of predictor variable which is known as hypothesis. These hypothesis are tested for acceptance and rejection. Predictive Data Mining Techniques are implemented by using supervised machine learning techniques. Data set is divided into two sets namely Training sets and Test set. Leaning of model is implemented by using training set and test set is used to determine the accuracy of model [Khulkarni and Sinha (2013)].

Random Forest uses Decision Trees as base classifier. This ensemble learning method is used for classification and regression of data. An ensemble is consist of number of trained classifiers whose predictors are combined to classify new instances. Bagging proposed by Breiman and Boosting proposed by Freund and Schapire (1997) are the two main methods for ensemble learning method. Bauer & Kohavi, Drucker & Cortes; Brieman; Quinlan, presents that Bagging and Boosting are effective for decision trees. These techniques are used for re-sampling of data to generate different training sets for each classifiers with replacement [Fawagreh *et al.* (2014)].

Bagging technique works on the basis of bootstrap samples. Let the original dataset consist of N number of records and m be the number of individual classifiers generated as part of ensemble. Then, m number of training sets are generated of size N from original dataset by sampling with replacement. The classifiers obtained during Bagging are independent of each other.

In case of Boosting, weights are assigned to each sample from original dataset. The classifiers are generated sequentially. For example, if Ci is the classifier which is to be generated then, the weights of sample having classifiers of C are updated. In other words, the classifiers obtained during Boosting are dependent of each other.

Optiz and Maclin concentrates on Bagging and Boosting in order to generate disagreement among the classifiers by altering the each training set so that resulting classifiers will produce different predictors.

Dietterich proposed random split selection where at each node split is selected randomly from among the k number of best splits. Breiman generates new training set by randomizing the outputs from the original training set. Ho has proposed a random subspace method. This method is responsible for random selection of subset of features to grow each tree. Amit and Geman define and search large number of generic features for the best split at each node.

Tibshirani and Wolpert and Macready proposed the estimation of generalization error by using out-of-bag estimates as an ingredient. Breiman (2001) has introduced randomness during the construction of decision trees using Classification and Regression Technique (CART).

II. WHAT IS RANDOM FOREST?

A Random Forest is a classifier consisting of collection of tree-structured classifiers where independent random vectors are distributed identically and each tree cast a unit vote for the most popular class at input x.

A random vector is generated which is independent of the past random vectors with same distribution and a tree is generated by using the training test [Brieman (2001)]. For random forests, an upper bound is derived to obtain the generalization error in terms of two parameters that are given below:

- The accuracy of individual classifiers
- The dependency between the individual classifiers

The generalization of error for random forest includes two segments. These segments are defined below:

- The strength of the individual classifiers in the forest.
- The correlation between them in terms of raw margin function

In order to improve accuracy of random forest, the correlation should be minimized while retaining their strength. Brieman (2001) studied that forests consist of randomly selected inputs or combination of inputs at each node to grow tree. The class of procedures has desirable characteristics that are listed below:

- Accuracy is good and sometimes better.
- Relatively robust to outliers and noise.
- Faster than Bagging and Boosting.
- Simple and can easily be parallelized.

Brieman has proposed a randomization approach that works better with bagging or random space method. To generate each tree of random forest, following steps are followed that are described below:

- Training dataset consist of N number of records.
- Sampling of N number of records are performed randomly but with replacement.
- This sample of dataset is named as bootstrap sample.
- If this training set would consist of M number of input variables, m<<M number of inputs are selected randomly out of M and the best split on these m attributes is used to split the node.
- The value of m will remain constant during forest growing.
- The tree will be grow to the largest possible level.

There are two reasons for using Bagging approach. They are given below: [Brieman (1994)]

- It seems that the use of bagging along with the random features generates more accurate results.
- Bagging can be used to provide ongoing estimation of generalization error as well as the estimation of strength and correlation.

When the forest is trained to classify a new instance, this whole process is executed across all the trees included in the forest. Each tree of the forest has to provide a vote which is recorded as a classification for the new instance. The votes of all trees are combined together and the votes are counted, the class having maximum number of votes are declared as classification of new instance. This process is known as Forest RI process.

The description of process for building forest is given below:

- When bootstrap sample is built by sampling the data with replacement of each tree, then one-third of the instances are left out.
- The left out instances are known as OOB (Out of Bag) data.
- Each tree of the forest has its own OOB data which is used for the error estimation of individual trees known as OOB error estimation.
- Random forest also consist of in-built facilities for calculating variable importance and proximities.
- The proximities are used for removing and replacing missing values and outliers.

III. CURRENT RESEARCH WORK ON RANDOM FOREST

The classification and regression of data depends on the accuracy and performance of Random Forest. As Random forest is an ensemble technique, experiments are performed with its base classifier to improve its accuracy and performance.

In order to have a good ensemble model, base classifiers should be diverse and accurate. The improvements that are made in the Random Forest are explained below:

A. Meta Random Forestt

Meta learning techniques are applied to the random forest [Boinee *et al.* (2006)]. It is based on the concept that random forest is made as a base classifier. The performance of this model is tested and compared with the existing random forest algorithm. Meta Random Forest is generated by using bagging and boosting approach. In case of bagging, random forest is taken as a base classifier with the bagging approach and in case of boosting, random forest as a base classifier is implemented with the boosting approach. By comparing all approaches Bagged random forest had shown excellent results. The Boosting random forest fails due to the presence of deterioration in generalization performances. This is because the complexity of the classifier becomes complex and it is unable to improve its performance on datasets. For the improvement of performance in classification, extra computation effort is required in advance.

B. ReliefF Random Forest

Brieman has introduced Gini Index as a measure for attribute split in decision trees. Robnik and Sikonja experimented and observed that Gini index cannot able to detect strong conditional dependencies among the attributes. This is because it measures the impurity of the class value distribution before and after the split on evaluated attribute. The same behavior is observed from other measures such as Gain ratio, DKM, MDL, j-measure. This problem is solved by ReliefF [Robnik and Sikonja (2004)]. ReliefF in random forest was used to evaluate attributes in the pre-processing step and quality estimates are used as weight for selecting subsamples of attributes at each level of tree. This helps to decrease the correlation between the attributes while maintaining the strength.

C. Dynamic Integration of Random Forest

The original Random forest algorithm was using combination techniques and selection techniques that are listed below:

- Combination Technique known as Weighted Voting is used where each node will have weight which is proportional to the estimated generalization performance of the corresponding classification.
- Selection Technique known as Cross-Validation Majority (CVM) is used where the classifier with the highest cross-validation accuracy is selected.

These above approaches are static. They select only one model or the combination of models uniformly.

Dynamic integration of random forest was proposed by [Tsymbal *et al.* (2006)] considering each new instance. They use three approaches that are listed below:

- Dynamic Selection (DC)
- Dynamic Voting (DV)
- Dynamic Voting and Selection (DVS)

The process is followed as described below:

First of all, errors of each base classifier on each instance of the training set is evaluated by using cross-validation method. Then, K-nearest neighbor of each new instance is estimated. Lastly, the weighted nearest neighbor is used to evaluate the performance of base classifier.

- DS selects the classifier having least local error.
- In DV, the weight received by each base classifier is proportional to the estimation of local performance.
- In DVS, the base classifiers with the errors are discarded that falls under upper half of the error interval. Remaining classifiers are operated using DV approach.

D. Forest-RK

Previously, Brieman has introduced a RF method known as RI method. In this hyper -parameter i.e. k number of features are selected randomly at each node during the tree induction process. The value of k is set arbitrary without any theoretical or practical approach. Therefore, a method for arbitrary setting of value of k is introduced [Bernard et al. (2008)]. This method is known as Forest-RK. In this method, the value of k is not a hyper-parameter which means that k does not play an important role for growing accurate RF classifiers. Rather, this method provides at least statically accurate results as the accurate results are provided by Forest-RI method with default settings.

E. Small Random Forest

In order to develop the random forest, large number of trees are generated to make the model more stable and less prone to the prediction errors. But, due to the large number of trees it becomes difficult to interpret the forest. To resolve this problem [Zhang and Wang (2009)] two objectives were proposed to shrink the forest that are listed below:

- To maintain a similar level of accuracy for prediction.
- To reduce the number of trees to intercept easily.

Three measures are considered in order to determine the minimal size of the forest

- A tree will be removed if it has lesser impact on the accuracy of prediction. The process is as: first the prediction accuracy of the forest is determined. Second, for every tree T, the prediction accuracy is determined of the forest that excludes T trees. The difference from the original forest and the forest excluded T trees are calculated. These are least important and can be removed from the forest. This method is known as “by prediction”.
- Other remaining methods are on the basis of similarity of trees. It means that trees are removed from the forest that are similar to other trees in the forest.

Another method was introduced to reduce the number of trees which is known as McNemar Test [Latinne *et al.* (2001)]. In this, systematically minimum number of weekend classifiers are determined and are combined with the provided dataset. There is no need to select the best classifiers and overproduction of the classifiers.

F. BAGA Algorithm

Current approach for an ensemble classifier is that the individual classifiers are combined together to classify new instances. But, if the dataset is too large for ensemble classifiers to classify, then another alternative approach should be used. This is because if the additional learning technique is added with the ensemble classifier then, the computation overhead would be increased. Therefore, BAGA algorithm is introduced [Wu and Chen (2005)]. This algorithm is an expansion of "Overproduce and Choose" paradigm. BAGA algorithm will use combination of bagging and genetic algorithm in order to generate the component classifiers in proper execution time by considering the set of decision tree as an input. This algorithm can select the classifiers dynamically. First, the set of classifiers are created with decision trees and then genetic algorithm is applied to select the optimal set of classifiers for an ensemble.

G. Dynamic Construction of Random Forest

In previous method, there is a drawback that it has to select the classifiers in advance. But, a dynamic construction of random forest was proposed [Tripoli et al. (2010)]. In this method, dynamically the base classifiers are selected to form the forest. The construction of forest is performed by adding each tree. While adding the tree each time to form the forest the accuracy is recoded and an online fitting procedure is applied on the curve to express the variations in accuracy. When the difference between the curve of accuracy and online fitting procedure satisfy certain specific criteria then the procedure will be terminated.

H. Dynamic Random Forest

Current Random Forest has some drawbacks that are listed below:

- Number of trees in the forest has to be specified priori in order to increase the performance.
- Since, trees are added independently in the forest, some trees can also degrade the performance of forest.

To solve the above listed problems Dynamic Random Forest are created [Bernard *et al.* (2012)]. In this approach, the trees are grown by keeping previous sub-forest into account. By using this concept, only reliable trees are allowed to grow in the forest.

The process followed by the Dynamic Random Forest is described below:

- The initial tree was build according to the traditional approach of Random forest i.e. the use of bagging method.
- The weights of N number of training instances are updated using boosting method so that the weights of wrong classified trees are increased and the weights of right classified trees are decreased.
- With the use of second step the new tree is generated in the forest.
- The base classifier in this approach is subset of trees i.e. Random Forest rather than decision trees.

I. Compute Unified Device Architecture (CUDA) Random Forest

Traditional Random Forest algorithms are well suited for the parallel execution. By combining the feature of parallel random forest with GPU's the performance will be improved in better way. An experiment was performed by implementing the Random Forest algorithm at right situation due to the large set of data, they are able to outperform the two known CPU based algorithm known as LibRF and Weka with any loss in accuracy [Grahn *et al.* (2010)]. They have revised the Random Forest algorithm by parallelizing the traditional decision trees and execute them on Graphical Processing Unit (GPU).

IV. RANDOM FOREST ALGORITHM FOR ONLINE DATA

A. Steaming Random Forest Algorithm

It is a stream classification algorithm that uses same techniques of traditional random forest to build steaming decision trees [Abdulsalam *et al.* (2007)]. Decision trees are grown from a different block of data. The trees are grown with the use of Hoeffding bounds. This is used to decide the stoppage of growth of trees. The algorithm uses Gini Index to select the attribute for splitting. It considers two parameters that are listed below:

- The number of trees to be built.
- Tree Window: the number of records are used for the growth of trees.

The process followed by algorithm is described below:

- The value of each attribute whether it is a numeric or an ordinal variable is divided into fixed length intervals.
- Hoeffding and Gini index is applied when the frontier nodes of the current tree accumulate the defined parameter.
- When the Hoeffding is satisfied, then the frontier nodes of the current tree are converted into the internal node and best split point is provided by the Gini Index.
- If the number of records reached to frontier node exceeds from the defined threshold, then frontier node would be converted into leaf node only if the accumulated nodes are from similar class.

B. Dynamic Steaming Random Forest Algorithm

It is a self-adjusting steam classification algorithm as it has the ability to reflect the concept changes. This algorithm is just similar with the Steaming Random Forest Algorithm but, only one difference is that the value of tree window is not constant for all trees [Abdulsalam *et al.* (2008)]. When the number of trees contribute for building tree then, the classification error of current tree is computed. If the value of error exceeds the threshold value then, the building of tree will be stopped and move forward for building nest tree. But, if not then the process will be continued for building the current tree using half number of previous records before starting building next tree.

It provides the expected results when testing is performed on synthetic data with concept drift. The concept drift points were clearly detected, the defined parameters were adjusted automatically and the classification error of tree was approximately equal to the noise in the dataset.

C. Online Bagging and Boosting

Traditional popular ensemble learning methods bagging and boosting guarantees for strong experimental results. But, these algorithm require the whole training set to be available at once or even sometime it requires random access of available training dataset. To overcome this problem, online bagging and booting methods were proposed [Oza and Russell (2001)]. This method requires only one pass i.e. at once only where there is no need to store and reprocessing of data. These methods are used in the situations when the data is arriving continuously. They are also useful in the processing of large data sets by the use of secondary storage as batch processing algorithms need multiple passes of the training data set.

D. Extremely Random Forest

A new ensemble technique was introduced to build random forest for classification and regression [Geurts *et al.* (2006)]. It consists of attributes that are extremely randomized and the point from where the best splitting of node is performed. They can also be totally randomized that the output values of the sample become independent. The strength of the randomization can be adjusted according to the requirement of the specified problem and appropriate choice of parameter.

V. OTHER ADVANCEMENTS IN RANDOM FOREST

A. Semi-Supervised Random Forest

Random Forest does not require various binary classifiers for the evaluation of multi-class problem. But, it has some limitation that it needs large amount of labeled data to move to the best level of performance. Therefore, to overcome this problem Semi-Supervised Random Forest algorithm was proposed [Leistner *et al.* (2009)]. This algorithm processes both labeled and unlabeled training data. It is based on Deterministic Annealing. Using this approach, unlabeled data consist of labels that can effectively treated as additional optimization variables. The other advantage of using this algorithm is that by estimating the out-bag-error monitoring of unlabeled data is performed.

B. Rotation Forest

This method is proposed for generating the base classifiers ensembles on the basis of feature extraction [Rodriguze and Kuncheva (2006)]. The training set for base classifier is created by splitting the feature set into K subsets and Principal Component Analysis (PCA) is applied to each subset of features. All principal components are retained so that the variability information in the data could be preserved. Therefore, K axis of rotation is performed to develop the new features for base classifiers. This approach is used to maintain the accuracy and diversity within the ensemble.

C. Fuzzy Random Forest

The accuracy of classification can be increase by combining the individual classifiers to form Multiple Classifier Systems. Multiple Classifier Systems are based on a forest formed by fuzzy decision trees known as Fuzzy Random Forest. This method is proposed by Bonissone *et al.* (2010). This approach is a combination of:

- Robustness of Multiple Classifier Systems.
- Power of randomness to increase the diversity of decision trees.
- Flexibility of fuzzy logic and fuzzy sets for maintaining the management of imbalanced data.

In this approach there is no need to consider all the attributes of dataset to split the nodes. At each node, random subset of total set of attributes is available and then best is chosen from one of them to split the nodes. As all attributes are not needed to split but, in some situation excluded attribute might be useful by other split in the same tree.

VI. APPLICATIONS OF RANDOM FOREST

Online Learning and Tracking: Incremental Extremely Random Forest algorithm was introduced in 2009 [Wang *et al.* (2009)]. This algorithm is used for online learning classification and in video tracking problems. It deals with the small steaming labeled data. Whenever the examples are arrived at leaf node, Gini Index is calculated to determine the splitting node of the tree. This technique also increases the capability as compared with other co-training framework. The examples are stored in the memory so that they can be reuse again to perform the split test with small number and avoid the calculation of Hoeffding bounds for large number of attributes.

Analysis of Hyperspectral Data: it is a challenging task because of large input space in dimension. Therefore, the generated classifiers are unstable and have poor generalization. So, to improve the generalization error of the classifiers random forest of binary classifiers are introduced [Crawford *et al.* (2003)]. The new classifier includes Bagging of training data and adaptive random subspace feature selection with Binary Hierarchical Classifier (BHC) in such a way that number of features are selected are dependent on the size of training data.

Species Classification: Cutler et.al has compared the accuracy of random forest with other machine learning techniques. He conclude that random forest better as compared to others.

Cancer Classification: Klassen et.al experiment that random forest can precisely classify cancer. They concluded that with the use of various gene sets of data random forest performs better for microarray data on the basis of speed and accuracy.

Prediction of pathologic complete response in breast cancer: Hu has applied random forest algorithm for the prediction of pathologic complete response in breast cancer. He concluded that random forest is a best approach as it able identify important genes of biological significance using feature extraction scheme of random forest.

Astronomical object classification: Gao et.al has experimented on multi-wavelength classification. He had proved that random forest is best algorithm for astronomical object classification. This is because random forest is consist of various features such as classification, feature selection, feature weighting and detection of outliers.

Cause of Death prediction: Flaxman et.al has proposed a new method known as CCVA method with the random forest algorithm in order to predict the cause of death.

Traffic signs classification: Zaklouta et.al has used K-d trees along with the random forest in order to classify 43 different types of traffic signs by using HOG descriptors of different sizes and distance transforms.

Accurate classification of crops: Low et.al has utilized the combination of Support vector Machine (SVM) and random forest to classify the crops with better accuracy and to provide the spatial information on map uncertainty.

Land cover classification: Pal (2003) has experiment J48 and compare the performance of bagging and boosting with the random forest. He concluded that random forest performs better even in the presence of noise in the training data. He also conclude that AdaBoost is influenced by the presence of noise in the training data as compared with the bagging approach.

There are many more applications like Credit card fraud detection, banking fraud detection, Text Mining detection, etc.

VII. CONCLUSIONS

This paper presents a review on Random forest, current ongoing work on Random forest and the applications of Random forest. Random forest is an ensemble method which generates accurate results but, on the other hand it is a time consuming method too as compared with the other techniques.

REFERENCES

[1] Abdulsalam H, Skillicorn B, and Martin P, (2007): Streaming Random Forests, Proceedings of 11th International Database and Engineering Applications Symposium, Banff, Alta pp 225-232.

[2] Bernard S, Heutte L, Adam S, (2008) Forest-RK : A New Random Forest Induction Method, Proceedings of 4th International Conference on Intelligent Computing: Advanced Intelligent Computing Theories and Applications – with Aspects of Artificial Intelligence, Springer-Verlag

[3] Bernard S, Heutte L, and Adam S, (2012): Dynamic Random forests, Pattern Recognition Letters, 33, 1580-1586.

[4] Boinee P, Angelis A, Foresti G, (2006): Meta Random Forest, International Journal of Computational Intelligence 2.

[5] Bonissone P, Candenas J, Garrido M, Diaz R, (2010): A Fuzzy Random Forest: Fundamental for Design and Construction, Studies in Fuzziness and Soft Computing, Vol 249, 23-42.

[6] Bonissone P, Cadenas J, Garrido M, Diaz-Valladares R, (2010): A Fuzzy Random Forest, International Journal of Approximate Reasoning, 51, 729-747.

[7] Breiman L, Bagging Predictors, (1994): Technical report No 421.

[8] Brieman L, Random Forests, (2001): Machine Learning, 45, 5-32.

[9] Crawford M, Ham J, Chen Y, Ghosh J, (2003): Random Forests of Binary Hierarchical Classifiers for Analysis of Hyper-spectral Data, Advances in Techniques for Analysis of Remotely Sensed Data, 337-345, IEEE.

[10] Geurts P, Ernst D, Wehenkel L, (2006): Extremely Randomized Trees, Machine Learning, volume 63, 3-42.

[11] Grahn H, Lavesson N, Lapajne M, Slat D, A CUDA implementation of Random Forest – Early Results, Master Thesis Software Engineering, School of Computing, Blekinge Institute of Technology, Sweden.

[12] Latinne P, Debeir O, Decastecker C, (2001): Limiting the number of trees in Random Forest, MCS, UK.

[13] Leistner C, Saffari A, Santner J, Godec M, Bischof H, (2009): Semi-Supervised Random Forests, ICCV IEEE, Conference Proceedings, 506-513.

[14] Oza, Russell S, Online Bagging and Boosting, (2001): Proceedings of Artificial Intelligence and Statistics, 105-112.

[15] Pal M, (2003): Random Forests for Land Cover Classification, Proceedings of Geoscience and Remote Sensing Symposium, IEEE, 3510-3512.

[16] Rodriguze J, Kuncheva L, (2006) Rotation Forest: A New Classifier Ensemble Method, IEEE Transaction on Pattern Analysis and Machine intelligence, Vol 28, N0 10, 1619-1630.

[17] Tsymbal A, Pechenizkiy M, Cunningham P, (2006): Dynamic Integration with Random Forest, ECML, LNAI, 801-808, Springer-Verlag.

[18] Wu X, Chen Z, (2005): Toward Dynamic Ensemble: The BAGA Approach, Proceedings of the ACS/ IEEE International Conference on Computer Systems and Applications.

[19] Zhang H, Wang M, (2009): Search for the smallest Random Forest, Statistics and Its Interface Volume.2, pp 381-388.

[20] E Tripoli, D Fotiadis, G Manis, 2010: "Dynamic Construction of Random Forests: Evaluation using Biomedical Engineering Problems", IEEE.

[21] Vrushali Y Khulkarni and Pradeep K Sinha, 2013: Random Forest Classifiers: A Survey and Future Directions, Volume.36, pp 1144-1153, IJAC.

[22] Khaled Fawagreh, Mohamed Medhat Gaber and Eyad Elyan, 2014: Random forests: from early developments to recent advancements, Volume.2, pp-602-609, System Science and Control Engineering.

[23] Robnik M, Sikonja, (2004): Improving Random Forests, J F Boulicaut et al (eds): Machine Learning, ECML 2004 Proceedings, Springer, Berlin.

[24] Abdulsalam H, Skillicorn B, and Martin P, (2008): Classifying Evolving Data Streams Using Dynamic Streaming Random Forests, Database and Expert Systems Applications, pp 643-651.

[25] Aiping Wang, Guowei Wan, Zhiquan Cheng and Sikun Li, (2009): An incremental extremely random forest classifier for online learning and tracking, pp 1449-1452.

[26] Cutler, D. R., Edwards Jr., T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783–2792.

[27] Klassen, M., Cummings, M., & Saldana, G. (2008, April 9–11). Investigation of random forest performance with cancer microarray data. In T. Philip (Ed.), Proceedings of the ISCA 23rd International Conference on Computers and Their Applications, CATA 2008, Cancun, Mexico (pp. 64–69). Cary, NC: International Society for Computers and Their Applications.

[28] Hu, W. (2009). Identifying predictive markers of chemosensitivity of breast cancer with random forests. *Cancer*, *13*, 59–64.

[29] Gao, D., Zhang, Y.-X., & Zhao, Y.-H. (2009). Random forest algorithm for classification of multiwavelength data. Research in Astronomy and Astrophysics, 9(2), 14–39.

[30] Flaxman, A. D., Vahdatpour, A., Green, S., James, S. L., & Murray, C. J. L. (2011). Random forests for verbal autopsy analysis: Multisite validation study using clinical diagnostic gold standards. Population Health Metrics, 9(29), 1–11.

[31] Zaklouta, F., Stanciulescu, B., & Hamdoun, O. (2011). Traffic sign classification using kd trees and random forests. In The 2011 international joint conference on neural networks (IJCNN), San Jose, CA, USA (pp. 2151–2155). New York City, NY: IEEE.

[32] Löw, F., Schorcht, G., Michel, U., Dech, S., & Conrad, C. (2012, September 24). Per-field crop classification in irrigated agricultural regions in Middle Asia using random forest and Support vector machine ensemble. In SPIE remote sensing, Edinburgh, United Kingdom (pp. 85380R–85380R). Bellingham, WA: International Society for Optics and Photonics.

[33] Freund Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119–39.