2015 IEEE 16th International Conference on Information Reuse and Integration

Alterations to the Bootstrapping Process Within Random Forest: A Case Study on Imbalanced Bioinformatics Data

Taghi M. Khoshgoftaar, Alireza Fazelpour, David J. Dittman, Amri Napolitano
Florida Atlantic University, Boca Raton, FL 33431
Email: {khoshgof@fau.edu, alifazelpo@fau.edu, ddittman@fau.edu, amrifau@gmail.com}

***Abstract*—Class imbalance is a significant challenge that practitioners in the field of bioinformatics are faced with on a daily basis. It is a phenomenon that occurs when number of instances of one class is much greater than number of instances of the other class(es) and it has adverse effects on the performance of classification models built on this skewed data. Random Forest as a robust classifier has been utilized effectively to deal with challenging characteristics of imbalanced bioinformatics datasets. In this study, we seek the answer to the question, do alterations to the bootstrapping process within Random Forest improve its classification performance? Thus, we performed an experimental study using Random Forest with four bootstrapping approaches, including two new novel bootstrapping approaches, across 15 imbalanced bioinformatics datasets. Our results demonstrate that two of the bootstrapping approaches, including one of our proposed approaches, outperform other approaches; however, this difference is statistically insignificant. We conclude that Random Forest is a robust classifier, able to handle the challenge of class imbalance, and can be slightly improved by altering bootstrapping process. To the best of our knowledge, no previous work has studied the effects of multiple bootstrapping processes on the performance of Random Forest in the domain of bioinformatics. In addition, we proposed and implemented the two innovative bootstrapping approaches evaluated in this paper.**

***Keywords*-Class imbalance; Random Forest (RF); bootstrap; bioinformatics**

I. Introduction

Technologies, such as microarrays, sequencing, and mass spectrometry, allow reserachers in the medical field to study biological phenomena of interest, which requires computational resources to analyze as the technology generates a large amount of data [21] with challenging characteristics, such as class imbalance.

Class imbalance is a common challenge frequently found in bioinformatics datasets and it occurs where one class (the majority class) has many more instances than the other class (the minority class) [14]. The class imbalance problem affects a variety of bioinformatics applications, such as medical diagnosis, disease research, and patient treatment planning. Unfortunately, for many of these applications, the class of interest is the minority class, which is the more problematic one because many conventional supervised classification algorithms are designed to maximize overall accuracy without considering the significance of each class within the training dataset. Luckily, various machine-learning tools, such as supervised classification algorithms, can assist practitioners to transform the data into valuable knowledge.

Throughout the literature there have been a variety of supervised machine-learning algorithms utilized to build a classification model and then using the generated model as a valuable tool to predict the class of future instances. However, due to the challenging characteristics of bioinformatics data, such as class imbalance, small dataset size, and high dimensionality, the process of choosing the most appropriate classifier for the problem under consideration is a very hard task.

Among these learners, Random Forest has been deployed for analyzing DNA microarray datasets effectively. For example, one study found that Random Forest is effective in distinguishing between cancerous and non-cancerous cells [6]. Another study utilized Random Forest using molecular signature to create well-defined classes [24] for clinical purposes. Diaz-Uriarte et al. [7] stated that Random Forest will be in a "standard tool-box" of techniques for microarray analysis. In addition, Random Forest provides practitioners in the field with both prediction accuracy and model interpretability [22]. Our data-mining research team has been conducted a few studies that have been shown Random Forest with 100 trees and 200 top selected features is an optimal model for bioinformatics research [28] and it is an effective and robust classifier [8].

However. all these studies utilized Random Forest with its standard bootstrapping approach, which draws instances using sampling with replacement regardless of their class labels. In this study, we seek to answer the question of whether different bootstrapping approaches, to generate bootstrap datasets, for training unpruned decision trees can further improve the performance of Random Forest for imbalanced bioinformatics data? To this end, we conducted an extensive experimental study using 15 skewed datasets, three feature rankers, four feature subset sizes, and four variants of Random Forest. These variants include standard Random Forest with the standard bootstrapping approach as a baseline benchmark along with the other three bootstrapping approaches (See Section III): Balanced Random Forest (BRF), Class Ratio Preserved Random Forest (CRPRF), and

978-1-4673-6656-4/15 $31.00 © 2015 IEEE
DOI 10.1109/IRI.2015.59

IEEE computer society

Class Ratio Preserved Random Forest - Unmodified Minority (CRPRF-UM). The last two approaches were proposed and implemented by our research group for this study.

Our results indicates that the approach of Class Ratio Preserved with Random Forest (CRPRF) outperforms the other three approaches in 5 out of 12 scenarios (42%). The Balanced Random Forest (BRF) approach follows closely the CRPRF in 4 out of 12 scenarios (33%). In addition, both approaches achieve the highest average Area Under the receiver operating characteristic Curve (AUC) values with 200 feature subset size using Information Gain and area under the Receiver Operating Characteristic curve (ROC) feature rankers, respectively. However, the difference among four classification approaches is statistically insignificant. This lead us to conclude that Random forest is a robust classifier in handling the class imbalance challenge that very often found in bioinformatics datasets, but minor alterations in bootstrapping process (specifically, CRPRF and BRF) can improve the classification results though not significantly.

To the best of our knowledge, no previous work has performed such an extensive experimental study in the domain of bioinformatics to examine the effects of bootstrapping step within Random Forest on the classification performance in the context of imbalanced and high-dimensional datasets. In addition, our major contributions are introducing two new alterations to bootstrapping approaches, Class Ratio Preserved (CRP) and Class Ratio Preserved - Unaltered Minority (CRP-UM), and testing for the first time Balanced Random Forest [4] in the bioinformatics domain.

The remainder of this paper is organized as follows: Section II presents discussions of previous research that are relevant to our work. In section III, Random Forest along with four bootstrapping approaches used in this study, are discussed. Section IV outlines methods used to conduct our empirical study, including datasets, feature ranking techniques along with feature subset sizes, cross-validation process and performance metric. In section V, we present our results with discussions of our findings. Finally, in section VI, we present our conclusions and potential avenues of future study.

II. Related Works

Supervised classification models are the most widely used machine-learning tools utilized by practitioners in the bioinformatics field to build inductive models for further analysis. There are a number of different classifiers each with their own strengths and weaknesses that can be deployed to handle the class imbalance problem and to enhance the classification performance. Among them is Random Forest that is known to be an effective and robust classifier [8], which not only can deal with multiple challenging characteristics (skewed data, small data size, and high dimensionality) commonly found in bioinformatics data, but it also provides both prediction accuracy and model interpretability [18] [22].

In one study, Diaz-Uriarte and Alvares de Andres [7] performed an experiment using Random Forest applied to ten different microarray datasets. They found that Random Forest is competitive in terms of classification results compared to other learners. In addition, they recommended Random Forest due to its ease of use, interpretation, and effectiveness as a standard classification tool for microarray data. In another study, Dittman et al. [9] conducted an empirical study using fifteen patient response DNA microarray datasets. Based on the classification results, they recommended using Random Forest with 100 trees along with a feature selection technique to reduce the feature set to around 1000 features in order to maximize the performance of the classification model. Their results also showed that the choice of feature selection technique is not statistically significant as long as feature selection is performed in some fashion. Wald et al. [28] concluded that the optimal classification model for bioinformatics datasets is Random Forest with 100 trees using the top 200 selected genes, and the optimal model performs well regardless of the choice of feature ranker.

Contrary to our present work, all these studies have utilized Random Forest with its standard bootstrapping approach that draws instances (regardless of their class labels) from the entire original training dataset using sampling with replacement of size N (total number of instances of the original dataset). A limited number of studies have examined a variety of bootstrapping approaches to draw instances either from each class separately or exclusively from one class. For example, Chen et al. [4] proposed a variant of random forest, called Balanced Random Forest (BRF), which draws a bootstrap dataset with the same number of instances from the minority class and the majority class independently. This creates a fully balanced (minority to majority of 50:50) bootstrap dataset to address the problem of class imbalance. While the datasets that Chen et al. used in their study were imbalanced, they were not high dimensional bioinformatics data.

III. Classification Approaches

Classifiers are the most widely used machine-learning tools utilized by practitioners in the bioinformatics field to build inductive models for further analysis. In this study, we chose Random Forest because previous studies indicate that it is a robust and an optimal classification model for imbalanced bioinformatics data [8] [28].

Random Forest (RF) algorithm was proposed by Leo Breimann [3] in 2001. It is a popular and powerful ensemble approach that combines multiple classification tress, each built using bootstrap datasets (sampling with replacement from the original dataset). Then, it constructs a set of unpruned decision trees on those bootstrap datasets and uses majority voting to make the final decision [1]. The RF,

like many classifiers, suffers from the curse of class imbalance and may perform poorly when trained on extremely imbalanced bioinformatics datasets. Thus, throughout the literature, several variants of RF were developed to alleviate the class imbalance challenge. In this study, we focus on the bootstrapping algorithms used to generate the derived bootstrap datasets to manage class imbalance more effectively and improve the classification performance as well.

The standard RF relies on the original bootstrap approach that chooses each bootstrap instance regardless its class label (minority or majority) from the entire original dataset [16]. Therefore, the derived datasets may contain few instances from the minority class and classification models built using these datasets perform poorly with respect to the minority class. To address this problem, Chen et al. [4] proposed a variant of this classification algorithm called Balanced Random Forest (BRF), which applies the bootstrapping approach to each class separately. We use the principal of the BRF to apply bootstrap approach to both the majority and minority classes creating fully balanced (minority to majority class ratio of 50:50) bootstrap datasets. In this case, the number of instances of the minority and the majority class in the bootstrap datasets is the same.

In addition, we proposed and implemented two novel bootstrapping methods that preserve the original class ratio for the generated bootstrap datasets. The first approach, Class Ratio Preserved RF (CRPRF), applies bootstrapping method to each class separately. It chooses the number of instances from each class so the class distribution of the bootstrap dataset is the same as the class ratio of the original data.The other approach, Class Ratio Preserved RF - Unmodified Minority (CRPRF-UM), bootstraps only the majority class (to the same size of the original training majority class) and utilize the entire minority class's instances from the original training dataset to form each of the bootstrap datasets.

IV. Methodology

In this section, we discuss various aspects of our experiments, including datasets, feature selection techniques, feature subset sizes, and cross-validation.

A. Datasets

The list of all 15 imbalanced datasets used in our experiments along with their characteristics is presented in Table I. The datasets are either tumor classification or patient response data publicly available through a number of different real-world bioinformatics, genetics, and medical projects. For more information on these datasets, we refer interested readers to the provided citations within Table I. For each dataset we show its name, total number of minority-class instances, total number of instances, percentage of instances from the minority class (in the range of 8.89%–25.56%), and the number of features or genes (in the range of 6,001–27,680). We chose these datasets because of their levels of class imbalance.

B. Feature Selection Technique and Gene Subset Size

In this experimental work, we use three forms of filter-based feature rankers: Information Gain, Area Under the ROC Curve, and Signal-to-Noise ratio. We use rankers because filter- and wrapper-based subset selection techniques can be computationally prohibitive with high-dimensional bioinformatics datasets. We describe each feature ranker briefly and interested readers may consult the provided references for more information.

Information Gain (IG) [12] is a well-known feature ranker [29] from the family of "Commonly-used" feature selection technique that measures the number of bits of information gained when predicting the class of an instance by knowing the feature's value. The importance of a feature is determined by how much the entropy of the class decreases when considered with that feature.

Area under the Receiver Operating Characteristic curve is a "Threshold-Based Feature Selection" (TBFS) feature ranker [10] used in conjunction with the performance metric of AUC. TBFS treats feature values as ersatz posterior probabilities and classifies instances based on these probabilities, allowing us to use classifier performance metrics as filter-based feature selection techniques. The TBFS technique which uses ROC aims to measure and optimize the balance between True Positive Rate versus False Positive Rate across all thresholds of decision boundaries. The larger the area, the more relevant the feature is.

Signal-to-Noise ratio (S2N) [30] comes from the family of "First-Order Statistics-based" (FOS) [15] feature rankers. It represents how well a feature distinguishes instances of two classes. It is defined as the ratio of means difference to sum of standard deviations of both classes. The more relevant features have the larger S2N ratios [5].

Because feature selection techniques aim to choose an optimum subset of features that can be used for subsequent analysis, one must decide on how many features to pick for this task. In this work, we decided on four feature subset sizes: 25, 50, 100, and 200. Previous research [8] showed that these four feature subset sizes are appropriate.

C. Cross-Validation and Performance Metric

Cross-validation [17] is the process of dividing the original dataset into N approximately equal-size partitions (folds), building the model using $(N-1)$ of these folds, then testing the built model using the Nth fold. This process is repeated N times so that each fold is used $(N-1)$ times to build the models and used only once to test the built model. The advantage of N-fold cross-validation over random sub-sampling is that all instances are used for both training and testing, and each instance is used only once per run for evaluating purposes. In this study, we used four

Table I: Details of the Datasets

Name	# Minority Instances	Total # of Instances	% Minority Instances	# of Attributes
Brain Tumor [27]	23	90	25.56%	27,680
ECML Pancreas [26]	8	90	8.89%	27,680
GSE1456 [20]	40	159	25.16%	12,066
GSE20271 [25]	26	178	14.61%	22,284
GSE25055 [13]	57	306	18.63%	22,284
GSE25065 [13]	42	182	23.08%	22,284
GSE3494-GPL96-ER [19]	34	247	13.77%	22,284
GSE3494-GPL96-Grade [19]	54	249	21.69%	22,284
GSE3494-GPL97-ER [19]	34	247	13.77%	22,646
GSE3494-GPL97-Grade [19]	54	249	21.69%	22,646
Lung 50k [10]	70	400	17.50%	54,614
Ovarian MAT [5]	16	66	24.24%	6,001
Raponi 2007 No SD [23]	10	54	18.52%	22,284
Raponi 2007 R+SD [23]	14	58	24.14%	22,284
Watanabe 2006 [31]	11	46	23.91%	12,626

runs of five-fold cross-validation to reduce any bias due to randomness. In addition, we performed feature selection for each bootstrap dataset generated to be trained on an unpruned decision tree for a total of 100 times for each Random Forest. We build (15 datasets $\times$ 100 iterations $\times$ 3 feature rankers $\times$ 4 feature subset sizes $\times$ 4 versions of Random Forest $\times$ 4 runs $\times$ 5-fold cross-validation) = 1,440,000 inductive models to evaluate the effectiveness of these classification approaches.

We use the Area Under the Receiver Operating Characteristic Curve [11] to assess the performance of all classification models. The curve plots the True Positive Rate (TPR) versus the False Positive Rate (FPR) across all decision boundaries. The area under the curve represents the quality of the model. It should be noted that the AUC described here is different from the feature selection technique mentioned in Section IV-B. To prevent any confusion, we use the notation AUC for the classification metric and ROC for the feature ranking technique.

V. Results

Table II contains the average AUC values for every classification model constructed over four runs of five-fold cross-validation across all 15 Imbalanced and high-dimensional DNA microarray datasets. Each entry in the table presents the average AUC values for every combination of four classifiers, three feature rankers, and four feature subset sizes across all 15 datasets. To improve readability, we present the best results for each combination of feature ranker and feature subset size in **boldface**. In addition, there are four (because of four versions of Random Forest) AUC values for each combination of feature ranker and feature subset size and we examine the frequency of being top-performer as well as best performer in terms of absolute average AUC values for each scenario.

Looking at these results in terms of frequency of being top-performing approach, we observe that the CRPRF approach outperforms others in 5 out of 12 (42%) scenarios followed closely by BRF in 4 out of 12 (33%) scenarios. In addition, in terms of absolute average AUC values, the CRPRF and BRF approaches achieve the highest AUC values of 0.772874 and 0.766341, respectively. The CRPRF achieves its highest AUC value with IG feature ranker and feature subset size of 200 while the BRF achieves its highest AUC value with ROC feature ranker and 200 features.

To further validate our results investigating the effectiveness of the various classification approaches used in our study, we conducted an ANalysis Of VAriance (ANOVA) [2] to find out whether the difference in performance among four learning approaches (RF, BRF, CRPRF, and CRPRF-UM) is statistically significant or not. Table III represents the ANOVA results for one factor (learner or classification approach) for all datasets. We chose a significance level of 5% for this ANOVA analysis; thus a "Prob>F" score of less than 0.05 is considered to be statistically significant. The results show that the differences between the four learning approaches are statistically insignificant. Thus, we did not perform a multiple comparison test with Tukey's Honestly Significant Difference (HSD) [2] criterion.

VI. Conclusions

Class imbalance is a significant challenge that often found in many bioinformatics datasets and it occurs when one class has many more instances than the other class(es). Luckily, tools from the domain of data mining and machine learning, such as supervised classification models can be utilized not only to alleviate this problem but to improve the classification performance as well. Random Forest is a robust classifier that has been used in the domain of bioinformatics to address challenging characteristics of these data. However, almost every study has used the standard bootstrapping approach of the Random Forest. In this study, we are motivated to seek whether alterations to the bootstrapping approach within the Random Forest can further improve the classification performance.

Table II: Classification Results

Ranker	Learner	Feature Subset Size			
		25	50	100	200
IG	RF	0.748491	0.756540	0.765690	0.769406
	BRF	**0.762792**	0.759041	0.769006	0.767672
	CRPRF	0.757391	**0.760080**	**0.769761**	**0.772874**
	CRPRF-UM	0.755778	0.757724	0.767035	0.767684
ROC	RF	0.763840	0.758951	0.761067	0.763919
	BRF	0.764146	0.761636	0.760079	**0.766341**
	CRPRF	0.763003	0.765016	**0.763345**	0.765296
	CRPRF-UM	**0.766174**	**0.765949**	0.763004	0.762512
S2N	RF	0.741024	0.744559	**0.755026**	0.756217
	BRF	**0.750047**	**0.752412**	0.749623	0.758495
	CRPRF	0.749769	0.746073	0.753240	**0.762879**
	CRPRF-UM	0.744035	0.748616	0.752416	0.762600

Table III: ANOVA Results

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Approach	0.028	3	0.00927	0.18	0.9089
Error	734.032	14396	0.05099		
Total	734.059	14399			

To accomplish this, we performed an extensive empirical study using four different approaches with the bootstrapping step of Random Forest, including two novel approaches proposed and implemented by our research team, three feature rankers from three different families of feature selection techniques, four feature subset sizes, and 15 imbalanced bioinformatics datasets. Our results indicate that two modifications to the bootstrapping step in Random Forest, Class Ratio Preserved Random Forest and Balanced Random Forest, outperforms other bootstrapping approaches, including the standard approach originally developed within the Random Forest classifier. However, these results are statistically insignificant. Thus, based on these results, we conclude that Random Forest is not only a robust classifier in handling the problem of class imbalance, but alterations in bootstrapping approach can slightly improve the classification performance as well.

To the best of our knowledge, this study is the first to examine the effects of alterations of bootstrapping approaches on the performance of Random Forest within the context of imbalanced bioinformatics datasets. In addition, our main contribution is introducing two innovative bootstrapping approaches, Class Ratio Preserved - Random Forest (CRPRF) and Class Ratio Preserved - Random Forest with Unmodified Minority (CRPRF-UM). Both of these bootstrapping approaches were developed and implemented by our research team for this study. Future work can include datasets from other health informatics related areas to investigate whether these results would generalize.

Acknowledgement

The authors gratefully acknowledge partial support by the National Science Foundation, under grant number CNS-1427536. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] D. Amaratunga, J. Cabrera, and Y.-S. Lee, "Enriched random forests," *Bioinformatics*, vol. 24, no. 18, pp. 2010–2014, 2008.

[2] M. L. Berenson, M. Goldstein, and D. Levine, *Intermediate Statistical Methods and Applications: A Computer Package Approach 2nd Edition*. Prentice Hall, 1983.

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[4] C. Chen, A. Liaw, and L. Breiman, "Using random forest to leran imbalanced data," Department of Statistics, University of California - Berkeley, East Lansing, Michigan, Tech. Rep., 2006.

[5] X. Chen and M. Wasikowski, "Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems," in *Proceedings of the 14th ACM SIGKDD Int'l Conf. Knowldege Discovery and Data Mining (KDD '08)*. New York, NY: ACM, 2008, pp. 124–132.

[6] A. Cutler and J. R. Stevens, "[23] random forests for microarrays," in *DNA Microarrays, Part B: Databases and Statistics*, ser. Methods in Enzymology, A. Kimmel and B. Oliver, Eds. Academic Press, 2006, vol. 411, pp. 422–432. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S007668790611023X

[7] R. Diaz-Uriarte and S. Alvarez de Andres, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, pp. 1–13, 2006.

[8] D. J. Dittman, T. M. Khoshgoftaar, R. Wald, and A. Napolitano, "Random forest: A reliable tool for patient response prediction," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM) Workshops*. BIBM, 2011, pp. 289–296.

[9] ——, "Maximizing classification performance for patient response datasets," in *Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference on*, Nov 2013, pp. 454–462.

[10] D. J. Dittman, T. M. Khoshgoftaar, R. Wald, and J. Van Hulse, "Comparative analysis of dna microarray data through the use of feature selection techniques," in *Proceedings of the Ninth IEEE International Conference on Machine Learning and Applications (ICMLA)*. ICMLA, 2010, pp. 147–152.

[11] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861 – 874, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016786550500303X

[12] M. A. Hall and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 6, pp. 392–398, November/December 2003.

[13] C. Hatzis, L. Pusztai, V. Valero, and et al., "A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer," *JAMA*, vol. 305, no. 18, pp. 1873–1881, 2011. [Online]. Available: +http://dx.doi.org/10.1001/jama.2011.593

[14] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sept. 2009.

[15] T. M. Khoshgoftaar, D. J. Dittman, R. Wald, and A. Fazelpour, "First order statistics based feature selection: A diverse and powerful family of feature selection techniques," in *Proceedings of the Eleventh International Conference on Machine Learning and Applications (ICMLA): Health Informatics Workshop*. ICMLA, 2012, pp. 151–157.

[16] T. M. Khoshgoftaar, D. J. Dittman, R. Wald, and W. Awada, "A review of ensemble classification for dna microarrays data," in *Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference on*. IEEE, 2013, pp. 381–389.

[17] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, vol. 14, no. 2, 1995, pp. 1137–1145.

[18] S. Loscalzo, L. Yu, and C. Ding, "Consensus group stable feature selection," in *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2009, pp. 567–576.

[19] L. D. Miller, J. Smeds, J. George, V. B. Vega, L. Vergara, A. Ploner, Y. Pawitan, P. Hall, S. Klaar, E. T. Liu, and J. Bergh, "An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 38, pp. 13 550–13 555, 2005. [Online]. Available: http://www.pnas.org/content/102/38/13550.abstract

[20] Y. Pawitan, J. Bjohle, L. Amler, A.-L. Borg, S. Egyhazi, P. Hall, X. Han, L. Holmberg, F. Huang, S. Klaar, E. Liu, L. Miller, H. Nordgren, A. Ploner, K. Sandelin, P. Shaw, J. Smeds, L. Skoog, S. Wedren, and J. Bergh, "Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts," *Breast Cancer Research*, vol. 7, no. 6, pp. R953–R964, 2005. [Online]. Available: http://breast-cancer-research.com/content/7/6/R953

[21] D. Popovic, A. Sifrim, C. Moschopoulos, Y. Moreau, and B. De Moor, "A hybrid approach to feature ranking for microarray data classification," in *Engineering Applications of Neural Networks*. Springer, 2013, pp. 241–248.

[22] Y. Qi, "Random forest for bioinformatics," in *Ensemble Machine Learning*. Springer, 2012, pp. 307–323.

[23] M. Raponi, J.-L. Harousseau, J. E. Lancet, B. Lwenberg, R. Stone, Y. Zhang, W. Rackoff, Y. Wang, and D. Atkins, "Identification of molecular predictors of response in a study of tipifarnib treatment in relapsed and refractory acute myelogenous leukemia," *Clinical Cancer Research*, vol. 13, no. 7, pp. 2254–2260, 2007. [Online]. Available: http://clincancerres.aacrjournals.org/content/13/7/2254.abstract

[24] T. Shi, D. Seligson, A. Belldegrun, A. Palotie, and S. Horvath, "Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma," *Modern Pathology*, vol. 18, pp. 547–557, 2005. [Online]. Available: http://www.nature.com/modpathol/journal/v18/n4/abs/3800322a.html

[25] A. Tabchy, V. Valero, T. Vidaurre, A. Lluch, H. Gomez, M. Martin, Y. Qi, L. J. Barajas-Figueroa, E. Souchon, C. Coutant, F. D. Doimi, N. K. Ibrahim, Y. Gong, G. N. Hortobagyi, K. R. Hess, W. F. Symmans, and L. Pusztai, "Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide chemotherapy response predictor in a multicenter randomized trial in breast cancer," *Clinical Cancer Research*, vol. 16, no. 21, pp. 5351–5361, 2010. [Online]. Available: http://clincancerres.aacrjournals.org/content/16/21/5351.abstract

[26] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, and R. Wald, "Feature selection with high-dimensional imbalanced data," in *IEEE International Conference on Data Mining Workshops, 2009. ICDMW '09.*, December 2009, pp. 507–514.

[27] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, and R. Wald, "A comparative evaluation of feature ranking methods for high dimensional bioinformatics data," in *Proceedings of the IEEE International Conference on Information Reuse and Integration - IRI'11*, 2011, pp. 315–320.

[28] R. Wald, T. M. Khoshgoftaar, D. J. Dittman, and A. Napolitano, "Random forest with 200 selected features: An optimal model for bioinformatics research," in *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, vol. 1, Dec 2013, pp. 154–160.

[29] H. Wang, T. M. Khoshgoftaar, and J. Van Hulse, "A comparative study of threshold-based feature selection techniques," in *Granular Computing (GrC), 2010 IEEE International Conference on*, 2010, pp. 499–504.

[30] M. Wasikowski and X. wen Chen, "Combating the small sample class imbalance problem using feature selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 1388–1400, 2010.

[31] T. Watanabe, Y. Komuro, T. Kiyomatsu, T. Kanazawa, Y. Kazama, J. Tanaka, T. Tanaka, Y. Yamamoto, M. Shirane, T. Muto, and H. Nagawa, "Prediction of sensitivity of rectal cancer cells in response to preoperative radiotherapy by dna microarray analysis of gene expression profiles," *Cancer Research*, vol. 66, no. 7, pp. 3370–3374, 2006. [Online]. Available: http://cancerres.aacrjournals.org/content/66/7/3370.abstract