Effects of the Use of Boosting on Classification Performance of Imbalanced Bioinformatics Datasets

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

Abstract—**In the domain of bioinformatics, two common problems encountered when analyzing real-world datasets are class imbalance and high dimensionality. Boosting is a technique that can be used to improve classification performance, even in the presence of class imbalance. In addition, data sampling and feature selection are two important preprocessing techniques used to counter the adverse effects of both challenges collectively. In this study, we examine whether the inclusion of boosting along with joint deployment of feature selection and data sampling techniques affect the classification performance of inductive models. To this end, we used two approaches: filter-based feature selection followed by either data sampling (denoted as FS-DS) or a hybrid data sampling and boosting technique entitled RUSBoost (denoted as FRB) which integrates random undersampling within the boosting process. We conducted an extensive experimental study using six high dimensional and imbalanced bioinformatics datasets along with three learners and four feature subset sizes. Our results show that the improvement of classification performance due to boosting depends on the choice of learner used to build the model. We recommend FRB because it outperforms FS-DS for nearly all scenarios. Additionally, our ANOVA analysis shows that the FRB is statistically distinguishable from the FS-DS when using the LR learner. To our knowledge, this is the first study to investigate the effects of boosting along with combined feature selection and data sampling on classification performance of inductive models in the domain of bioinformatics.**

Keywords-**High dimensionality; feature subset size; class imbalance; data sampling; boosting; RUSBoost;**

I. Introduction

Class imbalance happens when one class has many more instances than the other class(es) and it is often found to be a problem in bioinformatics data that can lead to increase bias toward the majority class and suboptimal classification results. Boosting is a technique that can be used to tackle the challenge of class imbalance [16] and improve classification performance. AdaBoost [12] is the most common boosting algorithm that iteratively builds an ensemble of models to improve the performance of any weak classifier. Seiffert et al. [21] developed an innovative and effective hybrid data sampling and boosting algorithm called RUSBoost that is designed to improve the performance of models trained on imbalanced data. The RUSBoost algorithm integrates random undersampling within each iteration of boosting.

Another technique that can be used to counter the class imbalance problem is data sampling. It modifies the composition of the original data by removing instances from the majority class until the target class ratio is achieved or by adding instances to the minority class until the target class ratio is reached. In this study, we used random undersampling to achieve a class balance ratio of minority to majority of 50:50; this selection of random undersampling and class ratio were based on preliminary investigations supported by early research [1].

In addition to class imbalance, high dimensionality is a common problem found in bioinformatics datasets. High dimensionality occurs when there is an excessive number of features (genes) within the data. Feature selection [15] is a common preprocessing technique designed to alleviate the problem of high dimensionality by choosing an optimum subset of features from a very large set of features. These techniques attain this objective by identifying the redundant or irrelevant features and choosing only the most useful features for subsequent analysis. This reduction in the number of features not only improves the efficiency of building the models from the reduced feature set, but improves the performance of these models in most cases [27]. Feature ranking techniques have become a very practical tool in the analysis of bioinformatics datasets because these techniques require smaller computational demands and produce output that is intuitive to understand.

The main objective of this paper is to investigate whether there is a difference (in terms of classification performance) when we deploy boosting along with joint application of feature selection and data sampling in building effective classification models. Thus, we conducted an empirical study using six high dimensional and

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.68

imbalanced bioinformatics datasets, three learners, and four feature subset sizes, performing an in-depth analysis of the effect of the inclusion of boosting when combined with feature selection and data sampling on the classification models' performances. Our results show that the inclusion of boosting improved the classification performance for all but one scenario. Thus we recommend boosting for the increased classification performance. However, because for the 5-NN and Naïve Bayes learners the differences are not significant, boosting can be removed for improved computational costs though there will likely be a reduction in classification performance.

No previous work has considered such an extensive empirical study in the field of bioinformatics that investigate the effects of boosting along with the joint application of feature selection and data sampling on classification performance of the build models. In addition, our recommendations are the main contributions to the domain of bioinformatics by providing guidelines on techniques to use or to avoid by practitioners and researchers in the field.

The rest of the paper is organized as follows. Section II discusses previous research that relates to our empirical study. Section III describes data sampling approaches and the RUSBoost technique utilized in this work. Section IV outlines methods that we use to conduct our experiment, including datasets used in this work. Section V presents the results of our work along with discussions of the results. Lastly, Section VI presents our conclusions and topics for future work.

II. Related Works

Working with datasets from the field of bioinformatics can be a challenging task because of the two prevalent problems of high dimensionality and class imbalance. Much research has been conducted with respect to the problems of high dimensionality and class imbalance in isolation. For example, Batuwita et al. [3] indicated that class imbalance is a major challenge particularly in the field of bioinformatics. This challenge is due to the fact that many traditional classifiers assume that the classes are relatively equal in size [14]. Another study, by Sharma et al. [22], stated that feature selection not only aims to alleviate the common problems associated with high dimensionality, the redundant and/or irrelevant features, by choosing an optimum subset of the original features by identifying the most important features and removing the rest, but it can improve the generalization capability of classification models built using the selected features.

However, there has been little work on investigating both problems simultaneously. One such study stated that high-dimensional imbalanced datasets exhibit suboptimal classification performance [6]. A study by Al-Shahib et al. [2] examined the joint effects of feature selection and data-sampling techniques on the performance of three classifiers. They used a single protein sequence dataset, a wrapper-based feature selection technique, and a random undersampling technique. Their finding was that, in general, applying both techniques (feature selection followed by random undersampling) improves classification performance. Unlike our study, their work was limited because it considered only one dataset that is not high dimensional (433 features), three classifiers, and one wrapper-based feature selection technique. However, in this study, we used six high-dimensional imbalanced datasets and three learners.

Blagus and Lusa [5] studied the performance of two data-sampling techniques, including random undersampling, on three breast cancer gene expression datasets. One drawback of this work, however, was that only one dataset was imbalanced, whereas in our study, all six datasets are highly imbalanced. They found that, for all classifiers, random undersampling outperformed SMOTE, particularly when feature selection was performed prior to the application of data sampling. These significant procedural differences make the present work more comprehensive than their work and improve the generalizibility of our results.

Boosting is another technique that can be used to alleviate the problem of class imbalance [21] and to improve classification performance for the built models. Boosting has not received much attention in the context of learning from bioinformatics imbalanced data. Lee et al. [19] conducted a comparative study to assess the performance of a variety of classification methods (including boosting) on seven microarray datasets. They found that boosting improved the classification performance. Khoshgoftaar et al. [16] stated that while boosting attempts to tackle the class imbalance problem using adjusted weights of instances, the boosting alone may not be sufficient to address the problem for severely imbalanced datasets. Seiffert et al. [21] developed an innovative hybrid boosting method called RUSBoost that integrates data sampling (random undersampling) into each iteration of boosting before building the classification models. We use RUSBoost in our present study due to its effectiveness compared to other boosting techniques [21].

To our knowledge, this is the first extensive work in the domain of bioinformatics considering the classification performance of joint feature selection and data sampling along with boosting (denoted as FRB) and without boosting (denoted as FS-DS). We used six high-dimensional and imbalanced bioinformatics datasets using three learners and four feature subset sizes to investigate whether deploying boosting can improve the performance of a classifier. As a results of this research we provide recommendations that will help practitioners and researchers in the field decide which techniques to deploy or to avoid.

III. Data Sampling Approaches and RUSBoost

Data sampling is a common preprocessing technique used to alleviate the class imbalance problem. One approach to combine feature selection and data sampling (to tackle both challenges of high dimensionality and class imbalance simultaneously) is to apply feature selection (FS) prior to data sampling (DS) and using the sampled data along with the selected features (denoted as FS-DS) as training data [1]. In this study, we use random undersampling (RUS) as the sampling technique due to its relatively higher performance over other data sampling techniques [21].

Boosting [12] is an ensemble technique to improve the performance of weak classifiers for building inductive models iteratively. At each iteration, boosting builds a new hypothesis and evaluates error associated with the hypothesis. Then, it adjusts the weights of each instance by increasing the weights of misclassified instances while decreasing the weights of correctly classified instances. Finally, all hypotheses from each iteration are used in a weighted vote to classify a new instance.

In 2010, Seiffert et al. [21] developed an innovative hybrid boosting method called RUSBoost (See Figure 1) that simply integrates random undersampling into each iteration of the boosting algorithm before building the classification models. As one can see, random undersampling (discarding instances from the majority class in a random fashion to achieve a desired balance class ratio) is applied to achieve a more balanced post-sampling class ratio before building inductive models. In this work, we applied filter-based ranking feature selection technique followed by RUSBoost (denoted as FRB) not only to tackle both problems jointly but to determine the effects (if any) of boosting on classification performance of inductive models.

In this work, we used the 50:50 post-sampling class ratio of minority to majority to compare classification performance of inductive models built using both approaches FS-DS and FRB. This class ratio was chosen due to preliminary investigations showing that it is appropriate and because the goal of data sampling is to have a balanced dataset.

IV. Methodology

A. Datasets

Table I contains the list of all six datasets used in this study, along with citations discussing the datasets in details. This table includes different characteristics of each dataset, such as name, total number of minority-class instances, total number of instances, percentage of instances from the minority class, the number of features (genes), and the average AUC values for all datasets. This average AUC value is based on classification models built on raw data using no data sampling and no feature selection. To create these AUC scores, five-fold cross-validation was employed and the average performance from six classification learners was used: Naïve Bayes, Multilayer Perceptron, 5-Nearest Neighbors, Support Vector Machines, and two versions of C4.5 decision trees (one using default parameter values, one using Laplace smoothing and no pruning [29]). Note that the results from these classifiers were used only to determine the difficulty of the datasets and have no further bearing on the rest of the experiment in this study. All models were built using the Weka data mining toolkit [13]. For more information on the process of calculating these values please refer to [9].

All six datasets are DNA microarray datasets publicly available from a number of different bioinformatics and medical projects. We chose these datasets because they exhibit a variety of class imbalance levels (in the range of 8.89%–25.56%) and large number of genes (in the range of 4,027–54,614, with the majority of datasets having more than 10,000 genes). In addition, these datasets exhibit a large variety of high class imbalance levels and none of the datasets can be considered balanced, thus data sampling is appropriate.

B. Feature Ranking Techniques

In this work, we use only one ranker, Area Under the ROC Curve (ROC), from a family of rankers we refer to as "Threshold-Based Feature Selection" (TBFS) [28]. The feature ranker ROC was chosen due to preliminary investigations supported by early research [10] that this ranker is effective for bioinformatics datasets. We use feature ranking because filter- and wrapper-based subset selection techniques can be computationally prohibitive,

Fig. 1: RUSBoost combined with external feature selection

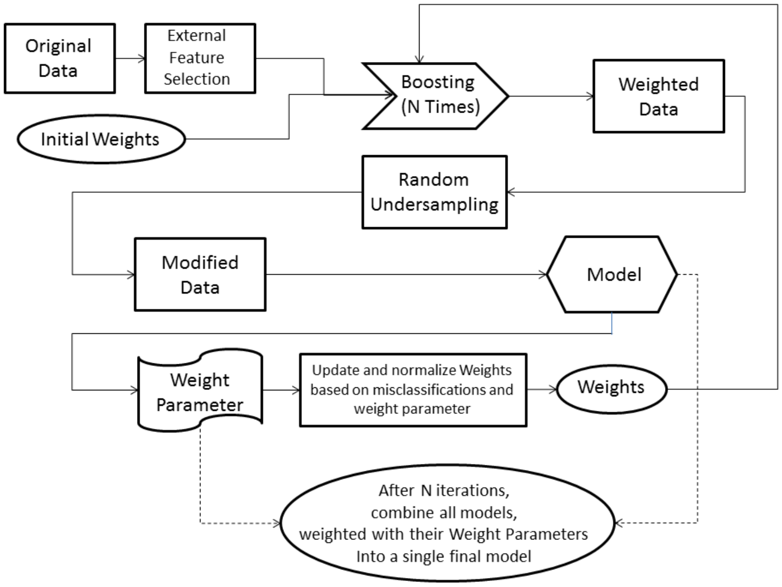
Original Data
External Feature Selection
Initial Weights
Boosting (N Times)
Weighted Data
Random Undersampling
Modified Data
Model
Weight Parameter
Update and normalize Weights based on misclassifications and weight parameter
Weights
After N iterations, combine all models, weighted with their Weight Parameters Into a single final model

TABLE I: Details of the Datasets

Name	# Minority Instances	Total # of Instances	% Minority Instances	# of Attributes	Average AUC
Brain Tumor [26]	23	90	25.56%	27680	0.7210
ECML Pancreas [23]	8	90	8.89%	27680	0.6723
Lung 50k [8]	70	400	17.50%	54614	0.8150
Ovarian MAT [6]	16	66	24.24%	6001	0.7896
Lymphoma [25]	23	96	23.96%	4027	0.8511
GSE1456 [20]	40	159	25.16%	12066	0.6108

in particular for datasets with high numbers of features (genes), which are very common in the field of bioinformatics.

Area Under the ROC Curve (ROC) [28], the area under the receiver operating characteristic (ROC) curve, is a measurement based on statistical decision theory. The ROC curve is used to measure and optimize the trade-off between True Positive Rate (TPR) versus False Positive Rate (FPR) across all decision thresholds. In this work, the ROC is used to determine the quality of a feature and to select the top features of each dataset to conduct our empirical study.

As the main objective of feature selection is to choose an optimum subset of features that can be used for subsequent analysis using classifiers to build inductive models, one must decide on how many of the features to use for this purpose. In this work, we decided on four feature subset sizes: 25, 50, 100, and 200. These four subset sizes are appropriate according to previous research [7].

C. Classifiers

In this paper, we use three classifiers: 5-Nearest Neighbor (5-NN), Logistic Regression (LR), and Naïve Bayes (NB). Because these are well-understood techniques, we provide only a brief discussion of these learners here. We refer interested reader to provided references for more information. All models in this paper were built using the Weka data mining toolkit [13] with default parameter values unless otherwise specified. Note that any changes to default parameter values were applied when experimentation showed an overall improvement of the classification performance [24].

5-Nearest Neighbor [30] is a lazy classifier that predicts the class of a new instance by finding the five training-set instances closest to the test instance and having them vote on the class. The “weight by 1/Distance” parameter was used for this voting process.

Logistic Regression [18] creates a simple logistic model of the data which predicts the class variable of new instances.

Naïve Bayes [30] is a Bayesian learner that attempts to approximate the posterior probability of an instance belonging to a particular class, given its values for the different features. By using Bayes's rule and making the naïve assumption of conditional independence, this can be computed based on the individual probabilities of each feature value given each class.

D. Cross-Validation and Performance Metric

Cross-validation [17] is a process that divides the original dataset into N approximately equal-size partitions (folds), builds the model using $(N-1)$ of these folds, then tests 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 fold for evaluating purposes. In this study, we used four runs of five-fold cross-validation to reduce any bias due to randomness. Additionally, we performed feature selection and random undersampling for each run and fold of every instance of 4 runs of 5-fold cross-validation, which results in 20 feature rankings per instance. We build (6 datasets x 1 feature ranking technique x 4 feature subset sizes x (1 iteration for FS-DS + 10 iterations for FRB) x 3 learners x 4 runs x 5-fold cross-validation) = 15,840 inductive models to evaluate the predictive power of the selected feature sets.

We use Area Under the ROC Curve [11] to evaluate all built classification models. The Area Under the ROC Curve plots the curve of True Positive Rate (TPR) versus False Positive Rate (FPR) across all decision thresholds; the area under the curve represents balanced levels that the model makes trade off between TPR and FPR. Because we use the Area Under the ROC Curve as both a classification performance metric and a ranker, as described in Section IV-A, anytime that confusion may arise, we use AUC as the performance metric and ROC as the feature ranker.

V. Results

Table II contains the average AUC values for every classification model constructed over four runs of five-fold cross-validation for all three learners and four feature subset sizes across all six high-dimensional and imbalanced microarray datasets. To improve readability, we present the best results of sampling approaches for each combination of learner and feature subset size in **boldface**. In addition, there are 12 (3 learners x 4 feature subset sizes) AUC values for each data sampling approach (FS-DS and FRB) and we examine the frequency with which each approach performs better than the other. Looking at these results, we observe that FRB outperforms FS-DS in 11 out of 12 scenarios (91.67%). The one exception is when using NB and 100 features. Based on these results we conclude that boosting can improve the classification performance of a model.

In order to more rigorously examine the differences between the approaches (FS-DS vs. FRB), we conducted an ANalysis Of VAriance (ANOVA) [4] to validate these results further and examine whether the difference in performance among the two approaches is statistically significant or not. Table III represents the ANOVA results for one factor (boosting) for each individual learner. 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 two approaches are statistically significant only for the classifier, LR. In addition, we performed a multiple comparison test with Tukey's Honestly Significant Difference (HSD) [4] criterion to find out which approach outperforms the other. Figure 2 shows the multiple comparison for both approaches for LR. In this figure, two approaches are statistically significantly different if their intervals are disjoint, otherwise they are not significantly different. This figure shows that FRB outperforms the other approach and it is statistically different from FS-DS.

Based on the experimental results along with the statistical analysis we find that the inclusion of boosting does improve classification performance for bioinformatics datasets in a majority of scenarios. However, for the learners 5-NN and NB, because the difference is not significant, we recommend that the inclusion of boosting be done if the computational resources are available. Otherwise, a practitioner can choose not to use boosting to reduce computational costs but the results may not be good.

VI. Conclusion

Class imbalance and high dimensionality are often found to be two major impediments when dealing with bioinformatics data. In this study, we examine the effects of boosting using two approaches (feature selection followed by data sampling versus feature selection followed by a hybrid approach of integrating random undersampling with boosting) on classification performance of bioinformatics datasets. Thus, we conducted a thorough

TABLE II: Classification Results - FS-DS vs. FRB

Learner	25		50		100		200	
	FS-DS	FRB	FS-DS	FRB	FS-DS	FRB	FS-DS	FRB
5-NN	0.87741	**0.88341**	0.88211	**0.88213**	0.88085	**0.88627**	0.88343	**0.89109**
LR	0.82265	**0.86040**	0.82038	**0.85075**	0.83008	**0.86410**	0.84252	**0.86911**
NB	0.87549	**0.87601**	0.87349	**0.87387**	**0.86798**	0.86436	0.86860	**0.87724**

TABLE III: ANOVA Results: FS-DS vs. FRB

Class Ratio	Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
5-NN	Boosting	0.0055	1	0.00547	0.39	0.5311
	Error	13.3628	958	0.01395		
	Total	13.3683	959			
LR	Boosting	0.2485	1	0.24854	11.09	0.0009
	Error	21.4778	958	0.02242		
	Total	21.7263	959			
NB	Boosting	0.0005	1	0.00053	0.04	0.8475
	Error	13.5982	958	0.01419		
	Total	13.5987	959			

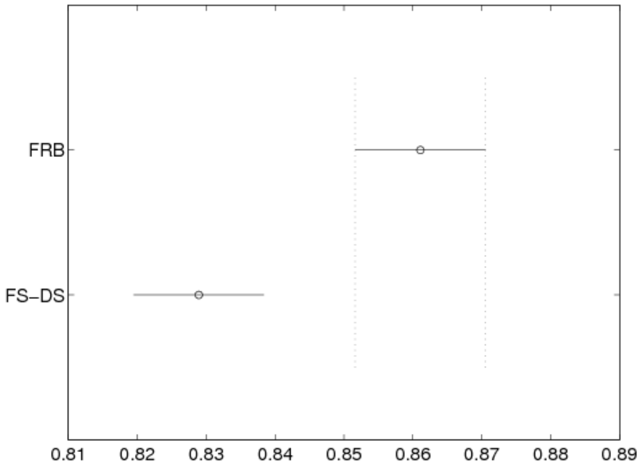

Fig. 2: Tukey's HSD Results: FS-DS vs. FRB - Logistic Regression

empirical study and compared both techniques using six high dimensional and imbalanced bioinformatics datasets, three learners, and four feature subset sizes.

To our knowledge, no previous work has performed such an extensive empirical study in the field of bioinformatics that investigates the effects of boosting along with the joint application of feature selection and data sampling on classification performance. Additionally, the main contributions of this study to the domain of bioinformatics are our recommendations that provide guidelines on techniques to use or to avoid by practitioners and researchers in the field.

Our results show that whether the inclusion of boosting can improve the performance of a classification model. Based on the results of this study, we recommend boosting (FRB) because it is the most frequent top performer. In addition, our results show that the boosting (FRB) is statistically significantly different from FS-DS when using the LR learner. Thus, this give us confidence to state that boosting (using FRB) is a good and safe choice to tackle both problems (class imbalance and high dimensionality) jointly and improve classification results. However, for NB and 5-NN boosting can be removed for a slight performance decrease if the computational resources are not available.

Future work will continue to include additional boosting algorithms, datasets, sampling techniques, feature rankers, and feature subset sizes. Another possibility is to include datasets from another application domain (health informatics) to investigate whether these results will generalize.

REFERENCES

[1] A. Abu Shanab, T. M. Khoshgoftaar, R. Wald, and A. Napolitano, "Impact of noise and data sampling on stability of feature ranking techniques for biological datasets," in *2012 IEEE International Conference on Information Reuse and Integration (IRI)*, August 2012, pp. 415–422.

[2] A. Al-Shahib, R. Breitling, and D. Gilbert, "Feature selection and the class imbalance problem in predicting protein function from sequence," *Applied Bioinformatics*, vol. 4, no. 3, pp. 195–203, 2005. [Online]. Available: http://www.ingentaconnect.com/content/adis/abi/2005/00000004/00000003/art00004

[3] R. Batuwita and V. Palade, "A new performance measure for class imbalance learning. application to bioinformatics problems," in *International Conference on Machine Learning and Applications*, Dec. 2009, pp. 545–550.

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

[5] R. Blagus and L. Lusa, "Evaluation of smote for high-dimensional class-imbalanced microarray data," in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, vol. 2, 2012, pp. 89–94.

[6] 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.

[7] 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.

[8] 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.

[9] D. J. Dittman, T. M. Khoshgoftaar, R. Wald, and A. Napolitano, "Simplifying the utilization of machine learning techniques for bioinformatics," in *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, 2013, pp. 396–403.

[10] ——, "Comparison of data sampling approaches for imbalanced bioinformatics data," in *FLAIRS Conference*, 2014.

[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] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proceedings of the 13th International Conference on Machine Learning*, 1996, pp. 148–156.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org.ezproxy.fau.edu/10.1145/1656274.1656278

[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] I. Inza, P. Larraaga, R. Blanco, and A. J. Cerrolaza, "Filter versus wrapper gene selection approaches in dna microarray domains," *Artificial Intelligence in Medicine*, vol. 31, no. 2, pp. 91–103, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0933365704000193

[16] T. M. Khoshgoftaar, D. J. Dittman, R. Wald, and A. Napolitano, "Contrasting undersampled boosting with internal and external feature selection for patient response datasets," in *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, vol. 2, Dec 2013, pp. 404–410.

[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. Le Cessie and J. C. V. Houwelingen, "Ridge estimators in logistic regression," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, pp. 191–201, 1992.

[19] J. W. Lee, J. B. Lee, M. Park, and S. H. Song, "An extensive comparison of recent classification tools applied to microarray data," *Computational Statistics and Data Analysis*, vol. 48, no. 4, pp. 869 – 885, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016794730400101X

[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] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Rusboost: A hybrid approach to alleviating class imbalance," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 40, no. 1, pp. 185–197, January 2010.

[22] A. Sharma, S. Imoto, and S. Miyano, "A filter based feature selection algorithm using null space of covariance matrix for dna microarray gene expression data," *Current Bioinformatics*, vol. 7, no. 3, pp. 289–294, 2012. [Online]. Available: http://www.ingentaconnect.com/content/ben/cbio/2012/00000007/00000003/art00007

[23] 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.

[24] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in *Proceedings of the 24th international conference on Machine learning*, ser. ICML '07. New York, NY, USA: ACM, 2007, pp. 935–942. [Online]. Available: http://doi.acm.org/10.1145/1273496.1273614

[25] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, and R. Wald, "Feature selection with high-dimensional imbalanced data." in *ICDM Workshops'09*, 2009, pp. 507–514.

[26] ——, "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.

[27] R. Wald, T. M. Khoshgoftaar, and D. J. Dittman, "Mean aggregation versus robust rank aggregation for ensemble gene selection," in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, vol. 1, Dec 2012, pp. 63–69.

[28] 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.

[29] G. M. Weiss and F. J. Provost, "Learning when training data are costly: the effect of class distribution on tree induction," *J. Artif. Intell. Res.(JAIR)*, vol. 19, pp. 315–354, 2003.

[30] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2nd edition, 2005.