Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference

Selecting the Appropriate Ensemble Learning Approach for Balanced Bioinformatics Data

David J. Dittman, Taghi M. Khoshgoftaar, Amri Napolitano
Florida Atlantic University
ddittman@fau.edu, khoshgof@fau.edu, amrifau@gmail.com

Abstract

Ensemble learning (process of combining multiple models into a single decision) is an effective tool for improving the classification performance of inductive models. While ideal for domains like bioinformatics with many challenging datasets, many ensemble methods, such as Bagging and Boosting, do not take into account the high-dimensionality (large number of features per instance) that is commonly found in bioinformatics datasets. This work seeks to observe the effects of two relatively new ensemble learning methods (Select-Bagging and Select-Boosting: the Bagging and Boosting approaches with feature selection implemented within each iteration of their algorithms) on a series of seven balanced (greater than a 43.50% minority class distribution) bioinformatics datasets. Additionally, we included the results when no ensemble approach is implemented (denoted as No-Ensemble) so that we can observe the full effects of ensemble learning. In order to test the three approaches we use three feature rankers, four feature subset sizes, and two classifiers. The results show that Select-Bagging is the top performing ensemble approach and statistical analysis confirms that Select-Bagging is significantly better than No-Ensemble and better (though not significantly) than Select-Boosting. Our recommendation is that Select-Bagging is an excellent choice for improving classification performance for bioinformatics datasets. To our knowledge, this work is the first empirical study focused exclusively on balanced bioinformatics datasets that investigated the effects of ensemble learning and utilizes Select-Bagging and introduces Select-Boosting.

Introduction

As a result of complications, such as high-dimensionality (having a large number of features (genes) per instance) and difficult to learn class boundaries, that are common in bioinformatics datasets, researchers have utilized techniques from domains such as data mining and machine learning to assist in the effective analysis of said data. Applications of this partnership may include reducing the computational costs of analysis, the removal of irrelevant or redundant features, and the building of inductive models for use in analyzing future data.

One powerful tool from the domain of data mining that has potential benefit for bioinformatics analysis is ensemble learning. Ensemble learning seeks to improve performance by combining the power of multiple models into a single decision. Potential benefits of ensemble learning in addition to improved performance include: reduced bias towards any particular class and reduced risk of overfitting. Additionally, ensemble learning approaches are versatile due to their ability to incorporate a variety of classifiers and data preprocessing techniques into their algorithms (Khoshgoftaar *et al.* 2013).

Two popular ensemble learning techniques are Bagging and Boosting. Bagging takes a random sample of instances with replacement from the training dataset so that it creates a new dataset made up from instances of the training dataset. This process is repeated multiple times and the classifiers are trained using the new datasets and the final result is made of a majority vote of the trained classifiers. Boosting begins with the training dataset and gives an initial identical weight to each instance. Upon the training and testing of the classifier built, the misclassified instances are given more weight and the correctly classified instances are given less weight. These new weights are used to directly give more weight to the instances in the new training data (Boosting by reweighting) or they are used in a weighted sampling with replacement process which creates a new training dataset where the misclassified instances are more likely to show in the new training dataset than the correctly classified ones (Boosting by resampling). The process repeats using the new training dataset and the overall process is repeated a predetermined number of times. The final decision is a weighted majority vote of all the trained classifiers. However, both the Bagging and Boosting algorithms do not take into account the inherent high-dimensionality commonly found in bioinformatics dataset or any data pre-processing techniques such as feature selection to combat said high-dimensionality.

Our work focuses on the application of two relatively new ensemble approaches, Select-Bagging and Select-Boosting (the Bagging and Boosting algorithms with feature selection incorporated into each iteration of their respective algorithms), on balanced (no dataset has less than a 43.50% minority class distribution) bioinformatics datasets. We test these two approaches using a series of seven balanced bioinformatics datasets, three feature rankers, four subset sizes,

and two classifiers. Additionally, to better observe the absolute effect of ensemble learning, we also observed the results when no ensemble approach is applied (denoted as No-Ensemble in this work). The results show that Select-Bagging is the top performing approach for both classifiers, achieving the highest classification performance in a majority of the scenarios. Alternatively, we find that not including an ensemble approach generally results in the lowest performance. Statistical analysis confirms that Select-Bagging is significantly better than No-Ensemble and better, but not significantly better, than Select-Boosting for both classifiers. Therefore, we recommend the use of Select-Bagging, and by extension the use of ensemble learning approaches, for improving the classification performance for models built from bioinformatics datasets. To our knowledge, this is the first empirical study which focuses on balanced datasets and utilizes Select-Bagging and introduces Select-Boosting.

The rest of the paper is organized as follows. The Related Works section contains previous research which relates to our experiment. The Ensemble Learning section introduces the specifics of the two ensemble learning approaches used in our work. The Methodology section outlines the methodology of our experiment. The Results section presents the results of our work. Lastly, the Conclusion section presents our conclusions and topics for future work.

Related Works

Ensemble learning is the process of combining decisions of multiple classification models into a single final result (Koziol *et al.* 2009). The main objective of ensemble methods is not only improving overall classification performance (Dietterich 2000) but also more accurate generalization capability in classifying unseen instances (Yang *et al.* 2010). There are two key factors that affect ensemble method performance: the accuracy and the diversity of the base classifiers (Dietterich 2000). In this study, our focus is on the two most popular ensemble techniques: Bagging (Breiman 1996) and Boosting (Freund & Schapire 1996).

Several scholars have investigated both Bagging and Boosting in their works. For example, Nagi et al. (Nagi & Bhattacharyya 2013) conducted an empirical study using nine high-dimensional cancer datasets and three classifiers. The researchers proposed a new ensemble method and compared class-specific accuracy of their method versus each single classifier as well as Bagging and Boosting. Another work by Tan et al. (Tan & Gilbert 2003) used seven cancer gene expression datasets along with the C4.5 decision tree classifier, and two ensemble methods: Bagging and Boosting with decision trees as the classifier. Chen (Chen 2014) conducted an experiment using eight microarray datasets and one feature selection technique, Relief-F.

In 2014, our research group introduced the Select-Bagging ensemble method (Dittman *et al.* 2014). In this work we observed how Select-Bagging performed compared to when no ensemble approach is applied. We used a single classifier and two feature selection techniques in our case study. Our results showed that Select-Bagging performs significantly better than when no ensemble approach is applied.

However, there are a number of shortcomings found within these studies. Nagi et al. (Nagi & Bhattacharyya 2013) did not employ feature selection and chose their three learners based on classification results from a series of datasets which are not high dimensional and are not representative of bioinformatics datasets. Tan et al. (Tan & Gilbert 2003) applied feature selection but it was deployed outside 10-fold cross-validation. In addition, they applied one run of 10-fold cross-validation to some datasets but not all and they chose different feature subset sizes for different datasets. Chen (Chen 2014) performed feature selection outside the ensemble methods (it causes overfitting of the classification models) and they did not provide any information on the class distribution of those datasets or how many features were selected for their experiment. As a result of these shortcomings, the results provided may be called into question. Even our own work can be considered preliminary, as it only discusses Select-Bagging for ensemble approaches.

Contrary to these studies, our current work addresses each of these concerns. We are comparing different ensemble approaches in addition to no ensemble. We also used seven high-dimensional and balanced datasets, three feature rankers from three different families of feature selection methods along with four feature subset sizes, and two classifiers using four runs of five-fold cross validation. In addition, we performed feature selection within each run and each fold of the cross-validation process (as well as within each iteration of the ensemble approaches) to avoid overfitting of the built classification models. Lastly, all our results are validated by statistical analysis.

Select-Bagging and Select-Boosting

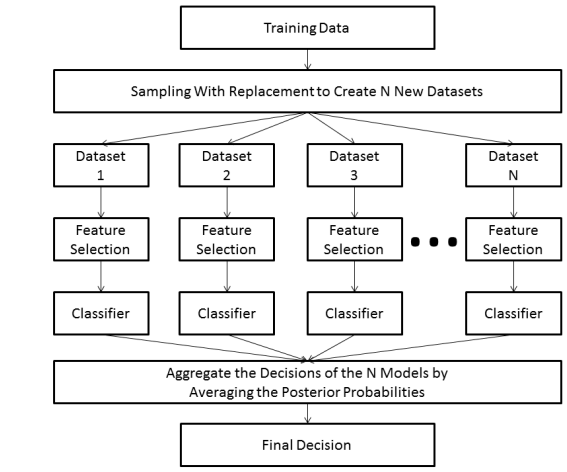
In this work we utilize two different ensemble learning approaches: Select-Bagging and Select-Boosting. Both of these techniques incorporate the feature selection process into their respective algorithms. This is an important distinction because both Bagging and Boosting (when using Boosting by resampling discussed later in this section) creates new training datasets with each iteration of their algorithms. Therefore, any feature selection performed before the ensemble approach will not be as valid with the new training datasets. Despite this, studies have performed the feature selection process either before the ensemble methods (Chen 2014) or even before the cross-validation process (if one is applied) (Tan & Gilbert 2003). As a result we developed Select-Bagging and Select-Boosting to apply the feature selection process on each new training dataset generated by their algorithms. However, as both Bagging and Boosting are well known ensemble learning techniques, we will focus on the particulars for the Select-Boosting and Select-Bagging processes. Both the Select-Bagging and Select-Boosting processes were implemented by our research group in the WEKA data mining toolset (Witten & Frank 2011). Each ensemble approach uses 10 iterations. For more information on the basic techniques please refer to (Breiman 1996) for Bagging and (Freund & Schapire 1996) for Boosting.

Select-Bagging (Dittman *et al.* 2014) (see Figure 1a) incorporates feature selection into the process of Bagging by

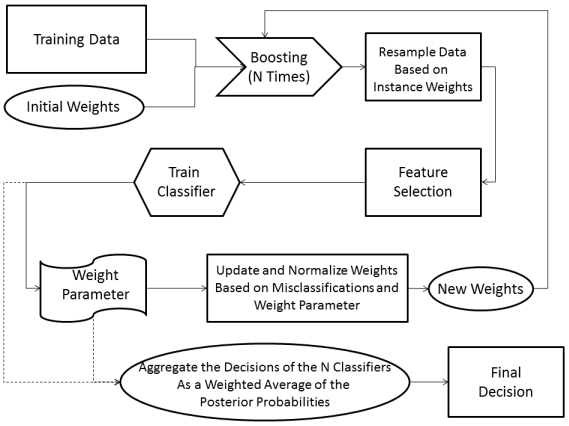
performing feature selection after the sampling with replacement for every iteration of the Bagging algorithm. After feature selection, a classifier is trained and the process is repeated the predetermined number of times. The final decision for Select-Bagging like with the Bagging algorithm is decided by taking the average of the posterior probabilities of the membership of the instance for the positive class from the collection of classifiers and using that average to make the final decision.

Select-Boosting (see Figure 1b), like Select-Bagging, incorporates the process of feature selection into the algorithm after the new training datasets are generated. However, in order to vary the training datasets for the feature selection process we used the Boosting by resampling option of the AdaBoost algorithm implemented in the WEKA data mining toolset. Boosting by resampling (activated by the "useResampling" option in WEKA being set to true), as opposed to Boosting by reweighting, resamples the training data based on the instance weights generated by that iteration. It should be noted, that the first iteration resamples based on the initial weights. As a result, a new training dataset is generated that is the same size as the original training dataset, with instances with high weights occurring more frequently than those with low weights. It is through this overrepresentation of the high weight instances and under representation of the low weight instances that the new weights are reflected. After feature selection is performed, a classifier is trained and is given a weight parameter and the process repeats for the predetermined number of iterations. The final decision of the Select-Boosting algorithm is the same as that of the AdaBoost algorithm: a weighted average of the posterior probabilities.

However, to truly observe the effects of ensemble learning, we need to compare the classification performance of Select-Bagging and Select-Boosting to that of a model built from feature selection on the training dataset followed by training a classifier (No-Ensemble). If the classification performance of the ensemble approaches are decisively better than that of No-Ensemble, then the implementation of the ensemble approaches are worth the effort.



(a) Select-Bagging



(b) Select-Boosting

Figure 1: Ensemble Approaches

Methodology

Datasets

Table 1 contains the list of datasets used in our experiment along with their characteristics. The datasets are all DNA microarray datasets acquired from a number of different real world bioinformatics, genetics, and medical projects. As the gene selection techniques used in this paper require that there be only two classes, we can only use datasets with two classes (in particular, either cancerous/noncancerous or relapse/no relapse following cancer treatment). The datasets in Table 1 show a large variety of different characteristics, such as number of total instances (samples or patients) and number of features. We chose these datasets because they have a variety of different levels of class imbalance but are all relatively balanced, as the smallest minority percentage is 42.50%.

Gene Selection Techniques and Feature Subset Size

We chose three forms of filter-based gene selection: a commonly used feature ranker, Information Gain; Threshold-Based Feature Selection (TBFS) used in conjunction with the Area Under the Receiver Operating Characteristic (ROC) Curve metric; and a first-order statistics based feature selection technique called Signal-to-Noise Ratio. For all three feature rankers we used four feature subset sizes: 25, 50, 100, and 200. These sizes were chosen because based on previous research, they are reasonable numbers of features (Khoshgoftaar *et al.* 2012).

Information Gain (Hall & Holmes 2003) is one of the simplest and fastest feature ranking techniques, and is thus popular in bioinformatics where high dimensionality makes some of the more complex techniques infeasible. Information Gain determines the significance of a feature based on the amount by which the entropy of the class decreases when considering that feature. Area Under the ROC Curve is a TBFS technique which treats feature values as ersatz posterior probabilities and classifies instances based on these probabilities, allowing us to use performance metrics as

Name	# Minority Instances	Total # of Instances	% Minority Instances	# of Attributes
DLBCL	23	47	48.94%	4027
Prostate	59	136	43.38%	12601
Breast Cancer	46	97	47.42%	24482
DLBCL NIH	102	240	42.50%	7400
BCancer50k	200	400	50.00%	54614
Spira2007	90	192	46.88%	22216
SotiriouMatrixData-Grade	45	99	45.45%	7651

Table 1: Details of the Datasets

filter-based feature selection techniques. The TBFS technique which uses Area Under the ROC Curve as its performance metric has been shown to be a strong ranker. For details on TBFS and the Area Under the ROC Curve metric please refer to (Abu Shanab *et al.* 2012). Signal-to-Noise Ratio is a measure of how well a feature separate the two classes. The ratio is defined as the difference between the mean value of that feature for the positive class instances and the mean value of that feature for the negative class instances over the sum of the standard deviation of that feature for the positive class and the standard deviation of that feature for the negative class. The larger the Signal-to-Noise Ratio, the more relevant a feature is to the dataset (Khoshgoftaar *et al.* 2012).

Classification, Cross-Validation, and Performance Metric

We used two different classifiers to create inductive models using the sampled data and the chosen features (genes). 5 Nearest Neighbor (k-nearest neighbors classifier with a k of five; denoted as 5-NN in this work) and Logistic Regression (LR), implemented using the WEKA toolkit (Witten & Frank 2011) using default values unless otherwise noted. Due to space limitations (and because these two classifiers are commonly used) we will not go into the details of these techniques. However it should be noted that for 5-NN the choice of a k of five and the weight by distance parameter being set to "Weight by $\frac{1}{distance}$" was chosen based on preliminary research. For more information on these learners, please refer to (Witten & Frank 2011).

Cross-validation refers to a technique used to allow for the training and testing of inductive models without resorting to using the same dataset. In this paper we use five-fold cross-validation. Additionally, we perform four runs of the five-fold cross validation so as to reduce any bias due to a chance split. However, it should be noted that the process of Select-Bagging, Select-Boosting, and No-Ensemble (including feature selection) is performed on every training dataset generated by the four runs of five-fold cross-validation. Therefore, we train 200 classifiers with feature selection for both Select-Bagging and Select-Boosting, as well as 20 classifiers with feature selection for No-Ensemble for every iteration of four runs of five-fold cross validation. The classification performance of each model is evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC) (Abu Shanab *et al.* 2012). Mathematically, this is the same metric as described above in the Gene Selection Technique and Feature Subset Size section, but there is a major distinction: for gene selection, we use an ersatz posterior probability to calculate the metric, but when used for evaluating classification models, the actual posterior probability from the model is used. To reduce confusion, we use AUC when referring to the performance metric. It should be noted that each datasets are separate from each other and none of the experiments combine any of the datasets though we do present the average classification results across the results from each dataset to to highlight more general trends.

Results

In this work, we seek to determine whether Select-Bagging or Select-Boosting is better suited for bioinformatics datasets using a series of seven balanced bioinformatics datasets. Additionally, we compare the two approaches to classification models built with no ensemble approach (No-Ensemble). In order to test these three approaches, we use three feature rankers, two classifiers, and four subset sizes. Table 2 contains the results of our experiment. Each entry in the table is the average AUC value across all seven datasets for every combination of ensemble approach, classifier, feature ranker, and feature subset size. The best ensemble approach for each combination of classifier, feature ranker, and feature subset size will be in **boldface** and the worst performing approach in *italics*.

Looking at 5-NN (top portion of Table 2), we see that Select-Bagging is the top performing ensemble approach for 12 out of 12 scenarios. It should also be noted that the approach of No-Ensemble is the worst performing approach for the ROC feature ranker. In the case of Information Gain and Signal-to-Noise, No-Ensemble is the worst performing approach in 50% of the scenarios. Specifically, No-Ensemble is the worst performing approach for subset sizes 25 and 200 with Information Gain and subset sizes 50 and 200 with Signal-to-Noise.

For Logistic Regression (bottom portion of Table 2), we see that unlike 5-NN, Select-Bagging is the top performing ensemble approach in 9 out of 12 scenarios. In terms of the three exceptions, Select-Boosting is the top performing approach, followed by Select-Bagging. Once again we see that No-Ensemble is the most frequent worst performing approach for the classifier for all 12 scenarios for Logistic Regression.

In order to further validate the results in our classification experiments, we performed two one-factor ANalysis Of VAriance (ANOVA) tests (Berenson, Goldstein, & Levine

Classifier	Subset Size	Information Gain			Area Under the ROC Curve			Signal-to-Noise		
		No-Ensemble	S.Boosting	S.Bagging	No-Ensemble	S.Boosting	S.Bagging	No-Ensemble	S.Boosting	S.Bagging
5-NN	25	*0.82121*	0.82375	**0.84697**	*0.80937*	0.82276	**0.83415**	*0.81288*	0.82486	**0.82842**
	50	0.82789	*0.82569*	**0.84477**	*0.81275*	0.82319	**0.82822**	0.81767	*0.81341*	**0.82512**
	100	0.83130	*0.83110*	**0.84523**	*0.81859*	0.82791	**0.83168**	*0.82446*	0.82518	**0.83114**
	200	*0.82857*	0.82939	**0.84125**	*0.81538*	0.82907	**0.83397**	0.82805	*0.82256*	**0.83175**
LR	25	*0.79156*	0.80923	**0.82071**	*0.79248*	0.81611	**0.83131**	*0.79787*	0.81141	**0.82803**
	50	*0.75411*	0.79538	**0.81517**	*0.76495*	0.80321	**0.82003**	*0.75487*	0.80897	**0.81702**
	100	*0.73920*	0.79938	**0.81375**	*0.75136*	0.80988	**0.81659**	*0.74286*	**0.81199**	0.80987
	200	*0.73927*	**0.81779**	0.81197	*0.75384*	0.82442	**0.82668**	*0.73874*	**0.82238**	0.80826

Table 2: Classification Results - Ensemble Approaches

Classifier	Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
5-NN	Approach	0.188	2	0.09407	3.84	0.0215
	Error	123.374	5037	0.02449		
	Total	123.563	5039			
LR	Approach	3.37	2	1.68492	64.39	2.43E-28
	Error	131.798	5037	0.02617		
	Total	135.168	5039			

Table 3: ANOVA Results: Ensemble Approaches

1983) (one for each classifier) with the choice of ensemble learning approach being the factor, across the seven datasets to determine if it has any statistically significant effect on the AUC levels. The results of the ANOVA tests (as seen in Table 3) show that the choice of ensemble approach is a significant factor for both classifiers. This is indicated by the Prob>F value being less than 0.05. Additionally, we performed a multiple comparison test using Tukey's Honestly Significant Difference (HSD) test (Berenson, Goldstein, & Levine 1983). Figure 2 contains the results of the Tukey's HSD tests. The results show that for 5-Nearest Neighbor, Select-Bagging is significantly better than No-Ensemble and better than Select-Boosting (but not to a statistically significant degree). In terms of Logistic Regression, both Select-Boosting and Select-Bagging are significantly better than No-Ensemble but are not significantly better than each other, with Select-Bagging being the top performing approach. Thus, we can recommend that the inclusion of an ensemble approach is beneficial and that we recommend using Select-Bagging as it is always significantly better than No-Ensemble and better than Select-Boosting for both classifiers.

Conclusion

Ensemble learning combines the power of multiple models into a single decision. Benefits from ensemble learning can include reduced overfitting and increased classification performance which makes ensemble learning a potential useful tool for bioinformatics. However, many ensemble approaches, such as Bagging and Boosting, do not take into account the inherent high-dimensionality found in these datasets. Thus we developed two new ensemble approaches, Select-Bagging and Select-Boosting, which incorporate the feature selection process into each iteration of their algorithms. In this work, we seek to determine whether Select-Bagging or Select-Boosting is best suited for bioinformatics datasets. Additionally, we include the results of the same experiments but with no ensemble approach applied in order to determine if the utilization of the ensemble learning approach is beneficial. We test the techniques using a series of seven balanced bioinformatics datasets along with three feature rankers, two classifiers, and four subset sizes.

Our results show that Select-Bagging is the most frequent top performing ensemble approach for both classifiers. Of the possible 24 scenarios, only three do not have Select-Bagging as the top performing approach, with Select-Boosting being the top performing approach for those scenarios. Additionally, the most frequent worst performing approach is No-Ensemble producing the worst performance for 20 out of the 24 scenarios. Statistical analysis shows that Select-Bagging is significantly better than No-Ensemble and better (though not significantly better) than Select-Boosting for both classifiers. Thus, it is our recommendation that Select-Bagging significantly improves the classification performance for balanced bioinformatics datasets. Future work may include the inclusion of more datasets, especially those for more specific purposes (tumor classification, patient response prediction, etc), to see if our findings remain valid.

References

Abu Shanab, A.; Khoshgoftaar, T. M.; Wald, R.; and Napolitano, A. 2012. 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)*, 415–422.

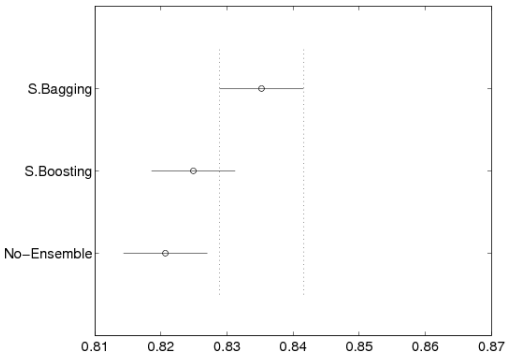
Berenson, M. L.; Goldstein, M.; and Levine, D. 1983. *Intermediate Statistical Methods and Applications: A Computer Package Approach 2nd Edition*. Prentice Hall.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24:123–140.

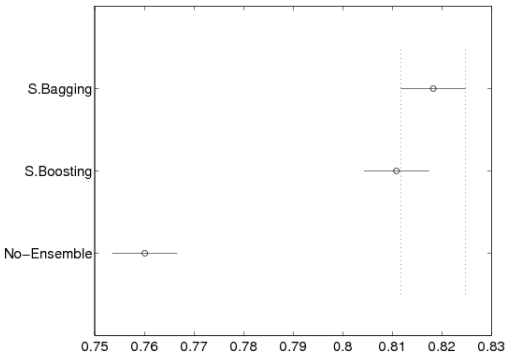
Chen, T. 2014. A selective ensemble classification method on microarray data. *Journal of Chemical & Pharmaceutical Research* 6(6):28602866.

Dietterich, T. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 1–15.

Dittman, D. J.; Khoshgoftaar, T. M.; Napolitano, A.; and Fazelpour, A. 2014. Select-bagging: Effectively combining



(a) 5-NN



(b) LR

Figure 2: Tukey HSD Results: Ensemble Approaches For Each Classifier

gene selection and bagging for balanced bioinformatics data. In *Bioinformatics and Bioengineering (BIBE), 2014 IEEE International Conference on*, 413–419.

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

Hall, M. A., and Holmes, G. 2003. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering* 15(6):392–398.

Khoshgoftaar, T. M.; Dittman, D. J.; Wald, R.; and Fazelpour, A. 2012. 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*, 151–157. ICMLA.

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

Koziol, J. A.; Feng, A. C.; Jia, Z.; Wang, Y.; Goodison, S.; McClelland, M.; and Mercola, D. 2009. The wisdom of the commons: ensemble tree classifiers for prostate cancer prognosis. *Bioinformatics* 25(1):54–60.

Nagi, S., and Bhattacharyya, D. 2013. Classification of microarray cancer data using ensemble approach. *Network Modeling Analysis in Health Informatics and Bioinformatics* 2(3):159–173.

Tan, A. C., and Gilbert, D. 2003. Ensemble machine learning on gene expression data for cancer classification. *Applied bioinformatics* 2(3 Suppl):S7583.

Witten, I. H., and Frank, E. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 3rd edition.

Yang, P.; Hwa Yang, Y.; B Zhou, B.; and Y Zomaya, A. 2010. A review of ensemble methods in bioinformatics. *Current Bioinformatics* 5(4):296–308.