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

Ensemble Gene Selection Versus Single Gene Selection: Which Is Better?

Randall Wald, Taghi M. Khoshgoftaar, David J. Dittman
Florida Atlantic University
rdwald@gmail.com, khoshgof@fau.edu, ddittman@fau.edu

Abstract

One of the major challenges in bioinformatics is selecting the appropriate genes for a given problem, and moreover, choosing the best gene selection technique for this task. Many such techniques have been developed, each with its own characteristics and complexities. Recently, some works have addressed this by introducing ensemble gene selection, which is the process of performing multiple runs of gene selection and aggregating the results into a single final list. The question is, will ensemble gene selection improve the results over those obtained when using single gene selection techniques (e.g., filter-based gene selection techniques on their own without any ensemble approach)? We compare how five filter-based feature (gene) selection techniques work with and without a data diversity ensemble approach (using a single feature selection technique on multiple sampled datasets created from an original one) when used for building models to label cancerous cells (or predict cancer treatment response) based on gene expression levels. Eleven bioinformatics (gene microarray) datasets are employed, along with four feature subset sizes and five learners. Our results show that the techniques Fold Change Ratio and Information Gain will produce better classification results when an ensemble approach is applied, while Probability Ratio and Signal-to-Noise will, in general, perform better without the ensemble approach. For the Area Under the ROC (Receiver Operating Characteristics) Curve ranker, the classification results are similar with or without the ensemble approach. This is, to our knowledge, the first paper to comprehensively examine the difference between the ensemble and single approaches for gene selection in the biomedical and bioinformatics domains.

Introduction

Gene expression profiles are an important tool in discovering patterns in the biomedical and bioinformatics domains. However, one significant challenge with these datasets is high dimensionality, the problem of having a very large number of features (genes) for each sample or instance. One method for combating this issue is through the use of gene selection techniques.

Gene selection is a specific application of feature selection, a data pre-processing technique from the field of data mining. The goal of gene selection is to choose an optimum subset of the full gene set and only use the gene subset in subsequent analysis. This is achieved by identifying the irrelevant and redundant genes and removing them from consideration, leaving only the optimum subset. The benefits of gene selection include reduced computation time, potentially improved classification, and a set of genes which can be further examined using laboratory techniques to determine if they are potentially of interest. Many gene selection techniques exist, which provides many options for practitioners but can also make it difficult to choose the appropriate technique for a given situation.

One option to improve the results of some gene selection techniques is to apply an ensemble approach to the process of gene selection. Ensemble gene selection is the process of performing multiple runs of gene selection and then aggregating those results in to a single feature subset. There are a number of benefits of ensemble feature selection including: more stable feature lists (the chosen features are more likely to remain valid when changes to the data occur) and comparable or superior classification results compared to the results from the single gene selection approach (e.g., using a single gene selection technique alone without an ensemble).

Despite the increased focus on ensemble gene selection in recent years, there is still the question of whether or not a gene selection technique will be improved through the use of the ensemble approach. This is an important fact to consider due to the increased computation time associated with ensemble techniques (multiple runs of gene selection vs. a single run of gene selection). Our paper is a thorough study of the effects of the ensemble approach on five gene (feature) selection techniques, when used to identify cancerous cells (or predict patient response to cancer treatment) on eleven biomedical datasets (specifically, gene microarray datasets). All five techniques were tested using both the ensemble and single gene selection approaches. We also used four feature subset sizes and five classifiers (learners) to build the models. Our results show that Information Gain and Fold Change Ratio favor the ensemble gene selection approach and that the Signal-to-Noise and Probability Ratio techniques favor the single gene selection approach. The final technique Area Under the ROC curve does, in general,

favor single selection but in three of the five learners ensemble gene selection outperforms single gene selection 50% of the time; thus, the proper choice will depend on the specifics of the experiment.

This paper is organized as follows. The Related Works section contains some background information regarding our topic. The Methodology section contains the process of our experiments. The Results section describes what we observed during our experiments. Lastly, the Conclusion section presents our findings and future work.

Related Works

Due to the high dimensionality (large number of genes) of DNA microarray datasets, dimensionality reduction techniques are a necessary preprocessing step. A study performed by Inza et al. (Inza et al. 2004) found that classification performed on reduced feature subsets derived from the original DNA microarray datasets outperformed classification using the whole feature set in a majority of cases, and that feature selection drastically reduced computation time. However, the large degree of high dimensionality causes a number of feature selection techniques to be infeasible in terms of computational time (Somorjai, Dolenko, and Baumgartner 2003). Therefore, a majority of the work on feature selection has been using univariate feature rankers.

However, univariate feature ranking techniques are very well suited for work in bioinformatics. There are a number of reasons why these techniques are ideal for this problem, including: the output of the techniques (a ranked list of features) is intuitive and easy to understand; the ranking of genes makes it easy for researchers to further validate the results through laboratory techniques; and the relatively small computational time when compared to other types of feature selection techniques (such as filter-based subset evaluation, wrapper approaches, etc.) (Saeys, Inza, and Larraaga 2007).

The use of ensembles for feature selection is a relative new concept. Originally, ensembles were used to develop models for decision making. It has been shown that these ensemble learners are competitive with other learners and in some cases are superior even in the biomedical domain (Dittman et al. 2011). Now, there have been studies in applying the ensemble concept to the process of feature selection (Haury, Gestraud, and Vert 2011). Current research has shown that not only do models built with feature subsets created using ensemble methods have comparable (or better) classification performance (when compared to models built using a single feature selection method), but the feature subsets themselves are more robust and can be appropriately applied to other data from the same problem (Awada et al. 2012).

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.

Table 1: Details of the Datasets

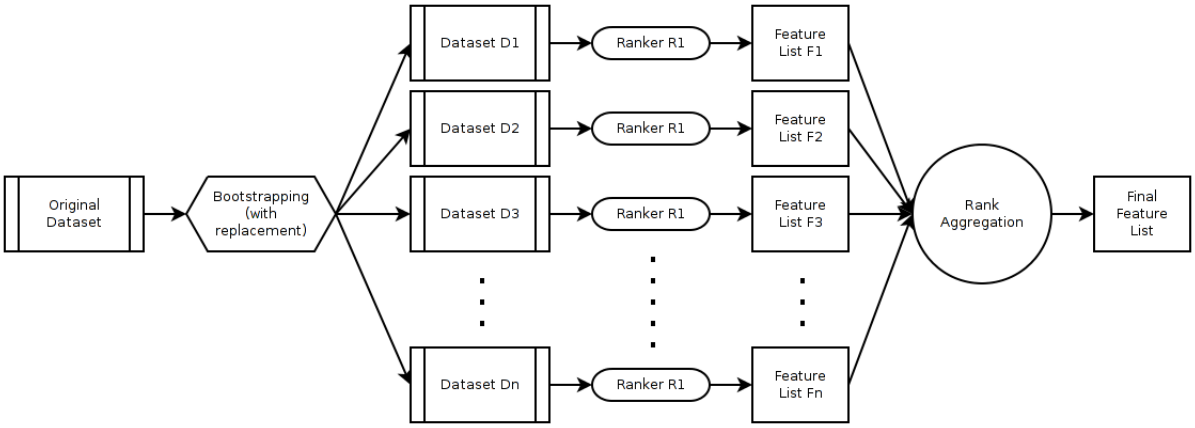
Name	Total # of Instances	# of Attributes	Average AUC
colon	62	2001	0.79413
ovarian_mat	66	6001	0.78958
prostate	136	12601	0.78225
Brain_Tumor	90	27680	0.72096
lungcancer-ontario	39	2881	0.71968
ECML_Pancreas	90	27680	0.67226
mulligan-r-pd	126	22284	0.65265
breast-cancer	97	24482	0.60085
mulligan-r-nr	169	22284	0.59308
DLBCL-NIH	240	7400	0.58527
CNS	60	7130	0.51893

As some of 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, in the case of the mulligan-r-pd and mulligan-r-nr datasets, 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. The final column, Average AUC (the Area under the ROC Curve, where the ROC Curve itself is a plot of True Positive Rate versus False Positive Rate, and thus the AUC shows how the model balances these two values), refers to the classification performance on these datasets when building models without feature selection. This is used to show that in addition to having many thousands of features, these datasets are notable for being difficult to model (such that models do not perform well), which suggests that they may also be difficult to select good features from. The values were calculated using a set of six different classification models: 5 Nearest Neighbor, Multilayer Perceptron, Naïve Bayes, Support Vector Machine, C4.5 D, and C4.5 N using the Weka data mining tool set (Witten and Frank 2011) and using five-fold cross-validation (see the Cross-Validation section for the details on cross-validation). The 5 Nearest Neighbor, Multilayer Perceptron, Naïve Bayes, and Support Vector Machine are decribed in the Classification section of this paper. C4.5 is a decision tree learner which uses information gain to select features for splitting the tree at each level. C4.5 D refers to using the C4.5 learner with the default values and C4.5 N refers to using a C4.5 variant with Laplace smoothing and no pruning. The reason for using these difficult datasets is because with the more difficult datasets, it is necessary to use techniques such as feature selection in order to improve the results of classification.

Gene Selection Techniques

In this paper we use five univariate or filter-based feature (gene) selection techniques: Area Under the ROC Curve (ROC) (Dittman et al. 2011), Probability Ratio (PR) (Forman 2003), Fold Change Ratio (FCR) (Jeffery, Higgins, and Culhane 2006), Signal-to-Noise (S2N), and Information Gain (IG) (Hall and Holmes 2003). The filter approach uses only the raw dataset to decide which features are to be used

Figure 1: Ensemble: Data Diversity



to create the best classifier. Since no classifier is used, filter methods must rely on statistical measures. Filters can either be rankers or subset evaluators, depending on whether they examine features one at a time or in groups. The reason we chose filter based gene ranking techniques is because of the relatively low computation time compared to other techniques (filter-based subset evaluation, wrapper, etc.) and the output of these techniques, a ranked lists of genes, is quite intuitive and can be easily validated through various laboratory techniques.

The five techniques can be placed into three categories: Threshold-Based Feature Selection (TBFS), First Order Statistics (FOS) based feature selection, and Information Gain. In TBFS, each attribute is evaluated against the class, independent of all other features in the dataset. After normalizing each attribute to have a range between 0 and 1, simple classifiers are built for each threshold value $t \in [0, 1]$ according to two different classification rules (whether instances with values above the threshold are considered positive or negative class examples). The normalized values are treated as posterior probabilities: however, no real classifiers are being built. The ROC and PR techniques fall under this category. ROC begins with the Receiver Operating Characteristic curve, which is a graphical representation of the trade off between the rate of detection (true positive rate) and false alarms (false positive rate) for a particular gene across all thresholds. In order to get a single metric we calculate the area under the curve with the larger the value, between zero and one, the more predictive power the gene has. The PR metric is a simple metric which the ratio of the true positive rate and the false positive rate.

FOS techniques are ranking techniques which focus on first order statistics such as mean and standard deviation to calculate the predictive power of these genes. S2N and FCR are both members of this family of techniques. S2N is a simple calculation of the ratio of the difference between the mean of the gene's value in the positive class and the mean of its value in the negative class to the sum of the standard deviations of the values in the positive and negative classes. FCR is the ratio of the mean of the gene's value in the positive class to the mean of the its value in the negative class

The last technique is IG, which is a commonly used technique for ranking features or genes. IG determines the significance of a feature based on the amount by which the entropy of the class decreases when considering that feature.

All of the above feature ranking techniques, with or without ensemble gene selection (described in the next section), produce a ranked list of features from best to worst. However, for building a classification model, a specific subset of features must be chosen. In this paper, we chose subset sizes of 10, 25, 50, and 100. This represents a wide range of possible subsets while significantly reducing the number of features compared with the original datasets.

Ensemble Gene Selection

As the goal of our work is to determine if ensemble approaches will improve the classification results of feature rankers, we must perform both single and ensemble gene selection. Single gene selection refers to applying the gene selection technique with no ensemble approach and using the single run for finding the gene subset. Ensemble gene selection uses multiple runs of gene selection whose results are then aggregated into a single result which is used to find the gene subset.

In this work we use the data diversity approach of ensemble feature selection. The data diversity approach (Figure 1) applies a single feature selection technique on multiple sampled datasets created from an original dataset. The sampled datasets are created through the use of bootstrapping with replacement (choosing random instances to create a new dataset where after each choice the instance is not removed from consideration in subsequent choices). This step creates diversity within the data being used for feature selection. The resulting feature lists are aggregated into a single final list (Awada et al. 2012). The reason why we chose data diversity over other ensemble approaches is that it only uses a single feature selection technique instead of an ensemble of different feature selection techniques. This allows us to

directly compare the difference between the feature selection techniques when used with data diversity and when used as a single run of feature selection, without the additional bias of combining different feature selection techniques with each other. In this paper we use each of the five gene selection techniques separately on fifty sampled datasets. While smaller numbers of iterations may be appropriate, we chose fifty iterations so that the number of iterations is sufficiently large enough to not be a mitigating factor with the results. The resulting ranked gene lists are aggregated through the commonly used aggregation technique: mean aggregation. Mean aggregation takes the mean rank of the gene across the lists as the final ranks and ranks the features based on this aggregated rank. In addition to the computational simplicity of the technique, mean aggregation has shown that it can outperform techniques which are designed with bioinformatics in mind (Wald, Khoshgoftaar, and Dittman 2012).

Classification

We used five different classifiers (learners) to create inductive models from the features (genes) chosen by both the single gene selection techniques and the ensemble gene selection techniques. These models are used to evaluate the predictive power of the genes chosen by applying them to a set of learners with varied properties. The five learners work as follows: 5 Nearest Neighbor (5-NN) (Witten and Frank 2011) classifies instances by finding the five closest instances to the test instance and comparing the total weight of the instances from each class (using 1/Distance as the weighting factor). Multilayer Perceptron (MLP) (Haykin 1998) builds an artificial neural network with three nodes in its single hidden layer, with 10% of the data being held aside for validating when to stop the backpropagation procedure. Naive Bayes (Witten and Frank 2011) uses Bayes' Theorem to determine the posterior probability of membership in a given class based on the values of the various features, assuming that all of the features are independent of one another. Support Vector Machines (SVM) (Witten and Frank 2011) find a maximal-margin hyperplane which cuts through the space of instances (such that instances on one side are in one class and the other side are in the other class), choosing the plane which preserves the greatest distance between each of the classes. For this study, we set SVM's complexity parameter c to 5.0 and its $buildLogisticModels$ parameter to "true" to provide proper probability estimates. Logistic Regression (Dittman et al. 2011) is a statistical technique that builds a logistic regression model to decide the class membership of future instances. All five learners use the built-in implementations in the Weka machine learning toolkit (Hall and Holmes 2003), using the default parameter values unless noted in the preceding descriptions.

Cross-Validation

Cross-validation refers to a technique used to allow for the training and testing of inductive models without resorting to using the same dataset. The process of cross-validation is that the dataset will be split as evenly as possible into a predetermined number of subsets or folds. The models (including feature ranking) are then built on the first $n - 1$ folds where n is the total number of folds. The model is then tested on the final fold and the results are collected. The final step is to change which fold is the testing fold and repeat the training and testing process until each fold has been the test fold exactly once. 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 lucky or unlucky split. It should be noted that the feature ranking process was performed inside the cross-validation step: that is, for each run and each "training set" within the cross-validation procedure, both the data diversity algorithm and single feature selection were performed for all five rankers, and ten different ranked lists (for each of five rankers, one list aggregated from the data diversity lists and one taken from the single feature selection) were created. These were then used in conjunction with the four different feature subset sizes and five learners to build models which were tested on the test fold. In total we built (11 datasets $\times$ 4 runs $\times$ 5-fold cross-validation $\times$ 5 gene rankers $\times$ (50 iterations for the ensemble gene selection and 1 run for single single selection) = 56,100 ranked feature lists (not counting the lists created through aggregation). In terms of inductive models we built (11 datasets $\times$ 4 runs $\times$ 5-fold cross-validation $\times$ 5 gene rankers $\times$ 2 gene selection approaches (with or without ensemble) $\times$ 4 feature subset sizes $\times$ 5 learners) = 8,800 models.

Results

This section contains the results from our experiment comparing the classification performance of five filter-based gene selection techniques with and without the use of an ensemble approach for feature selection. The performance was compared by evaluating both ensemble feature selection and single feature selection on eleven biomedical and bioinformatics datasets with the ensemble approach using fifty iterations. Tables 2 through 6 contain the results of our experiment, with each table representing one of the five learners. Each entry is the average AUC (Area Under the ROC Curve) value across the eleven datasets when holding static the learner, gene selection technique, feature subset size, and whether or not we use an ensemble approach (labeled Ensemble in the tables) or not (labeled Single in the tables). The top performing value for each selection technique using the same subset size is in boldface.

When looking across all of the learners, we find that each ranker will clearly prefer either the ensemble approach or single feature selection. The rankers IG and FCR both are improved by the application of the ensemble techniques. ROC, PR and S2N will all favor the single feature selection approach rather than the ensemble approach.

Upon looking at each of the individual learners we find that there are some exceptions to the above trends. With the ROC ranker, for three of the learners (MLP, Naïve Bayes, and SVM) the ensemble technique will outperform the single feature selection approach in 50% of the cases. This allows us to state that the decision of whether or not to apply the ensemble approach is left to the practitioner. In PR, the single feature selection approach will either outperform the ensemble approach or match it except in Naïve Bayes

Table 2: Average Classification Results of the 5 Gene Selection Techniques Using 5-NN

Subset	ROC		PR		S2N		FCR		IG	
Size	Ensemble	Single	Ensemble	Single	Ensemble	Single	Ensemble	Single	Ensemble	Single
10	0.73791	**0.74500**	**0.71820**	0.71723	0.72398	**0.74481**	**0.65119**	0.63631	**0.74263**	0.73069
25	0.74676	**0.75559**	**0.72235**	0.72101	0.72865	**0.75609**	**0.66857**	0.66125	**0.76466**	0.75177
50	**0.76672**	0.76161	0.72015	**0.72089**	0.74016	**0.75454**	0.65784	**0.67841**	**0.76649**	0.76357
100	0.75872	**0.76504**	0.72970	**0.73402**	0.74284	**0.75600**	**0.68517**	0.67015	0.76037	**0.76561**

Table 3: Average Classification Results of the 5 Gene Selection Techniques Logistic Regression

Subset	ROC		PR		S2N		FCR		IG	
Size	Ensemble	Single	Ensemble	Single	Ensemble	Single	Ensemble	Single	Ensemble	Single
10	0.74362	**0.74470**	0.71074	**0.71930**	0.71380	**0.74907**	**0.65354**	0.63420	**0.73252**	0.71730
25	**0.71080**	0.70292	0.69736	**0.72280**	**0.71209**	0.70055	**0.66011**	0.65565	**0.72591**	0.71598
50	0.69947	**0.70259**	0.67930	**0.71477**	0.69675	**0.70463**	0.66999	**0.67289**	**0.69573**	0.68960
100	0.68938	**0.69405**	0.66863	**0.69204**	0.69663	**0.70243**	**0.67860**	0.67548	**0.68994**	0.68828

Table 4: Average Classification Results of the 5 Gene Selection Techniques Using MLP

Subset	ROC		PR		S2N		FCR		IG	
Size	Ensemble	Single	Ensemble	Single	Ensemble	Single	Ensemble	Single	Ensemble	Single
10	0.75155	**0.75277**	0.73795	**0.74561**	0.73597	**0.75396**	**0.66495**	0.63405	**0.75838**	0.74888
25	**0.76375**	0.75789	0.73856	**0.73989**	0.75211	**0.75671**	0.66641	**0.66857**	**0.76995**	0.76815
50	**0.76318**	0.75539	0.73792	**0.74387**	0.75546	**0.76138**	0.67267	**0.67273**	**0.76803**	0.76507
100	0.75884	**0.75967**	0.73904	**0.74030**	0.75649	**0.76500**	0.67270	**0.68656**	**0.77038**	0.76642

Table 5: Average Classification Results of the 5 Gene Selection Techniques Using Naïve Bayes

Subset	ROC		PR		S2N		FCR		IG	
Size	Ensemble	Single	Ensemble	Single	Ensemble	Single	Ensemble	Single	Ensemble	Single
10	0.74580	**0.74925**	**0.72684**	0.72544	0.72708	**0.73911**	**0.63501**	0.61468	**0.74199**	0.70902
25	0.75467	**0.76117**	**0.71793**	0.71710	0.72630	**0.74275**	**0.64517**	0.64119	**0.76445**	0.74106
50	**0.74798**	0.74772	0.69945	**0.70162**	0.71832	**0.73715**	0.64181	**0.64916**	0.75280	**0.75684**
100	**0.74397**	0.74097	**0.69107**	0.68814	0.69662	**0.71355**	**0.64821**	0.64118	0.74307	**0.74767**

Table 6: Average Classification Results of the 5 Gene Selection Techniques Using SVM

Subset	ROC		PR		S2N		FCR		IG	
Size	Ensemble	Single	Ensemble	Single	Ensemble	Single	Ensemble	Single	Ensemble	Single
10	0.75857	**0.76112**	0.74400	**0.74806**	0.74234	**0.76281**	**0.66398**	0.63333	**0.76802**	0.75511
25	**0.77170**	0.76465	0.74617	**0.74631**	0.75914	**0.76229**	**0.69239**	0.67812	**0.78027**	0.77298
50	**0.76756**	0.76031	**0.74487**	0.73629	0.75843	**0.76470**	0.69352	**0.70455**	**0.77393**	0.76411
100	0.75664	**0.75783**	**0.74305**	0.74162	0.75635	**0.75846**	0.69936	**0.71081**	**0.76296**	0.76019

in which the ensemble approach does outperform the single feature selection approach. S2N only has a single case where the ensemble approach outperforms the single approach (using Logistic Regression and a feature subset size of 25). FCR favor the ensemble approach in all learners except for MLP in which it favors the single feature selection approach. Finally, IG only has two cases in which it performs better when using the single feature selection approach (using a feature subset size of 50 and 100 when using Naïve Bayes).

Conclusion

Gene selection has become a necessary step for working with high dimensional bioinformatics or biomedical datasets. The decision of which gene selection technique to implement is an important but daunting task. In addition to choosing which gene selection technique to apply, one must also choose whether or not to apply an ensemble approach (multiple runs of gene selection which are then aggregated into a single list) along with the gene selection technique. The question is: will the ensemble approach improve the classification performance of the gene selection technique, or will the gene selection technique achieve better classification performance without the addition of the ensemble approach? This paper is a study of the classification performance of five gene selection techniques where we applied the techniques both with an ensemble approach (data diversity where a single feature selection technique is applied toward a number of sampled datasets derived from a single dataset, then the resulting lists are aggregated into a single list) as well as with single gene selection (a single single run of gene selection) to determine how the ensemble approach will affect the classification performance. In the process of our experiments we used eleven biomedical (gene microarray) datasets from the field of cancer research as well as five learners and four feature subset sizes. In terms of the ensemble approach we use fifty iterations of gene selection.

Our results show that each of the gene selection techniques react to the inclusion of an ensemble approach differently. The Information Gain and Fold Change Ratio techniques, in a majority of cases, will be improved by the inclusion of the ensemble approach. The Signal-to-Noise technique and the Probability Ratio techniques in a majority of cases are negatively effected by the use of the ensemble approach. The final technique, Area Under the ROC Curve, does in general favor the single feature selection approach but for three of the five learners the ensemble approach will outperform the single feature selection approach in 50% of the cases. This allows us to recommend that for Information Gain and Fold Change Ratio, the ensemble approach should be used, while for Signal-to-Noise and Probability Ratio, the single feature selection approach will work better. As for Area Under the ROC Curve we believe that the choice is relatively minor in terms of classification and the decision falls to the practitioner. While there are exceptions to the above trends we feel reasonably confident in our recommendations.

Future work in this area will focus on the addition of more datasets. The inclusion of more datasets will allow us to further test the trends discovered by our experiments. Additionally, by including more datasets with a specific purpose (i.e. patient response prediction) we can test these trends with a more focused goal.

References

Awada, W.; Khoshgoftaar, T.; Dittman, D.; Wald, R.; and Napolitano, A. 2012. A review of the stability of feature selection techniques for bioinformatics data. In *2012 IEEE 13th International Conference on Information Reuse and Integration (IRI)*, 356–363.

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

Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research* 3:1289–1305.

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.

Haury, A.-C.; Gestraud, P.; and Vert, J.-P. 2011. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE* 6(12):e28210.

Haykin, S. 1998. *Neural Networks: A Comprehensive Foundation 2nd edition*. Prentice Hall.

Inza, I.; Larraaga, P.; Blanco, R.; and Cerrolaza, A. J. 2004. Filter versus wrapper gene selection approaches in dna microarray domains. *Artificial Intelligence in Medicine* 31(2):91–103.

Jeffery, I.; Higgins, D.; and Culhane, A. 2006. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics* 7(1):359.

Saeys, Y.; Inza, I.; and Larraaga, P. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19):2507–2517.

Somorjai, R.; Dolenko, B.; and Baumgartner, R. 2003. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics* 19(12):1484–1491.

Wald, R.; Khoshgoftaar, T. M.; and Dittman, D. 2012. Mean aggregation versus robust rank aggregation for ensemble gene selection. In *2012 11th International Conference on Machine Learning and Applications (ICMLA)*, volume 1, 63–69.

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