Select-Bagging: Effectively Combining Gene Selection and Bagging for Balanced Bioinformatics Data

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

***Abstract*—Bioinformatics datasets have historically been difficult to work with. However, within machine learning, there is a potentially effective tool to combat such problems: ensemble learning. Ensemble learning generates a series of models and combines their results to make a single decision. This process has the benefit of utilizing the power of multiple models but the overhead of having to compute the multiple models. Thus, we must ask whether the benefits outweigh the detriments. In this study, we seek to determine if the ensemble learning technique Select-Bagging improves classification results over feature selection on the training dataset followed by classification (denoted as FS-Classifier in this work) on a series of balanced bioinformatics datasets. We test the two approaches with two filter-based feature rankers, four feature subset sizes and the Naïve Bayes classifier. Our results show that Select-Bagging clearly outperforms FS-Classifier for nearly all scenarios. Subsequent statistical analysis shows that the increase in performance generated by Select-Bagging is statistically significantly better than FS-Classifier. Therefore, we can state that the inclusion of Select-Bagging is beneficial to the classification performance of models built on high-dimensional and balanced bioinformatics datasets and should be implemented. To our knowledge this is the first study which looks at the effectiveness of bagging in conjunction with internal feature selection for balanced bioinformatics datasets.**

***Keywords*-high-dimensionality; ensemble; bagging; feature selection;**

I. Introduction

Bioinformatics datasets have a number of inherent problems, such as high-dimensionality (having a large number of features per instance), noisy data (errors in the datasets themselves), and difficult to learn class boundaries which are challenges for researchers to overcome. Commonly, techniques from the domain of machine learning have been utilized of multiple purposes including reducing the dimensionality of the datasets and building inductive models of making decision on new data. In particular there is a subset of techniques which can potentially powerful tool for bioinformatics: ensemble learning.

Ensemble classification is the process of combining the results of multiple related models into a single decision. There are a number of benefits associated with ensemble classification including, reduced bias, being less prone to overfitting, and improved classification performance. Additionally, ensemble classification is a very diverse area of study as many of the ensemble techniques can be utilized with any number of base classifiers and data pre-processing techniques.

One popular ensemble technique is bagging. The basic principle of bagging is to take a random sample of instances from the dataset with replacement so that the new dataset has the same size of the original training dataset. Steps can also be taken (as is done in this work) to preserve the class distribution of the original training dataset. A model is built from this new dataset and a decision is found. This process is repeated multiple times and the results of the models are aggregated into a single decision.

As a result of the inherent high-dimensionality of many bioinformatics datasets, it has become necessary to include dimensionality-reducing techniques such as feature (gene) selection, a process of choosing an optimum subset of features to build subsequent models. However, within the ensemble classification method, feature selection should be included in all iterations of the ensemble. In this paper we focus on a technique called Select-Bagging which incorporates feature selection into every iteration of the bagging algorithm. We test Select-Bagging against feature selection on the training dataset followed by classification (denoted as FS-Classifier) on a series of seven high-dimensional balanced (each class having a similar number of instances in the dataset)

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.66

bioinformatics datasets. Additionally we use two filter-based feature rankers, four feature subset sizes, and the Naïve Bayes classifier when performing our experiments.

The results show that Select-Bagging, on average, outperforms FS-Classifier for all feature ranking techniques and feature subset sizes. Additionally, when looking at the datasets individually, in a majority of cases Select-Bagging outperforms FS-Classifier for both feature rankers. Statistical analysis confirms that the increase in classification performance by Select-Bagging is a significant increase. Therefore, it is our recommendation to use Select-Bagging over FS-Classifier.

The rest of the paper is organized as follows. Section II contains a disscussion of previous research that relates to this study. Section III describes the process of both Select-Bagging and FS-Classifier. Section IV presents the specific process by which our experiments are conducted. Section V contains the results of our work along with a discussion of the results. Lastly, Section VI presents our conclusion and potential future avenues of research.

II. Related Work

Bioinformatics datasets frequently contain challenges such as high-dimensionality, noisy data, and difficult to learn class boundaries which make it necessary for advanced data-mining techniques to be implemented [9]. In particular, ensemble learning is an effective approach for increasing the classification performance for challenging datasets. These techniques (e.g., bagging) seek to alleviate these problems by incorporating the results over multiple classification models to reduce the potential for overfitting the training data [20]. This process allows researcher to increase the utility and more effectively use the data that they posses.

There have been a number of studies which have incorporated ensemble learning, including bagging, into their studies. Tan et al. [15] performed a study comparing ensemble methods to a single classifier C4.5 on a series of cancer classification gene microarray datasets and found that ensemble techniques show improved results for gene expression data. Chen et al [4] compared three different ensemble learning techniques on gene microarray data with preprocessing being performed by the Relief-F feature ranker prior ensemble implementation. They found that ensemble methods when combined with data pre-processing techniques show improved results. Nagi et al. [12] looked at bagging, boosting, and stacked generalization along with their own proposed method on a series of nine bioinformatics datasets. Results show that bagging outperforms both boosting and stack generalization and was competitive with their proposed method.

However, these studies are not without their downsides. Tan et al. [15] only uses one run of ten-fold cross-validation, instead of multiple runs of a smaller fold number cross-validation which reduces bias due to a chance split. Additionally, Tan et al. performs their dimensionality-reducing technique, discretization, before implementing cross-validation. This is a problem because the features chosen may not be as valid when the training dataset changes due to the process of cross-validation. This is further compounded when the datasets change again due to the ensemble learning techniques. Tan et al.'s datasets are not all balanced. One of their testing datasets is imbalanced (minority class distribution of 10.07%). This calls into question their results as that the chance of overfitting caused by relying on a small number of minority instances is increased with the multiple models. Chen et al. [4] only uses feature selection on two of the approaches being tested and uses them externally to the ensemble approaches. Additionally, they reported that their "experiments are repeated 30 times independantly" [4] but give no explanation as to how. Lastly, Chen et al. do not give any information on the class distribution of their datasets. Nagi et al. [12] does not perform any feature selection within their experiment. They, like Tan et al., also only perform one run of ten-fold cross-validation instead of multiple runs of a smaller fold cross-validation. Nagi et al., like Chen et al., do not give the class distribution of the training and testing datasets created from the original data used in their study. However, upon further investigation of the source of the data, the Kent Ridge Bio-Medical Dataset Repository (link can be found in the original article [12]), it was found that one of the original datasets is clearly imbalanced (minority class distribution of 17.13%) which becomes an issue when you do not take measures to counteract the imbalanced classes (e.g., data sampling). Lastly, all of the studies do not take measures to preserve the class distribution when performing ensemble techniques.

In contrast, our work addresses each of these issues. We perform four runs of five-fold cross-validation to reduce bias due to a chance split. The feature selection performed in this work is always performed upon the twenty training datasets generated by each iteration of four runs of five-fold cross-valaidation. Additionally, in the case of Select-Bagging, feature selection is performed on all new training datasets genereated from the sampling

with replacement process. Lastly, all of our datasets are binary class datasets which have relatively balanced class distributions with the minority class percentage ranging from 42.50% to 50.00%. We also take steps to preserve the class distribution of the original training dataset (see Section III for details). These factors give our work a unique view of the ensemble learning process and gives further credence to our findings.

III. Select-Bagging and Feature Selection on the Training Dataset Followed by Classification

Bagging was developed in 1996 [3] in order to improve the results of unstable single-run classifiers. The basic principle of bagging is to take a random sample of T instances from the dataset with replacement where T is the number of instances in the unsampled dataset. However in order to conserve the inherent class distribution ratio of the original training dataset, the T instances are split into P and N instances where P and N are the number of instances in the positive and negative classes found in the original training dataset. Using this new dataset, a model is built using a classifier and the results are recorded. This process repeats R times and the results of the R models are aggregated into a final decision.

Select-Bagging (see Figure 1) incorporates feature selection into the process of bagging by performing feature selection after the sampling with replacement for every iteration of the bagging algorithm. The reason why the feature selection is performed for every iteration is that a different dataset is used in every iteration and the feature chosen for one dataset is not guaranteed to be valid for another. Therefore, by incorporating the feature selection into each iteration we are not forced to rely on a set of features chosen for a different dataset. The Select-Bagging process is implemented in the Weka data mining toolset [8] using the default number of 10 iterations.

However, because ensemble techniques have a higher computational cost over single classifier learners, we need to compare the classification performance of Select-Bagging to that of a model built from feature selection on the training dataset followed by classification (FS-Classifier). If the classification performance of Select-Bagging is decisively better than that of FS-Classifier, then the implementation of Select-Bagging is worth the additional computational cost.

IV. Methodology

A. Datasets

Table I contains the list of all seven datasets used in this study, along with citations which contain the details of the datasets. Included in this table are the names, number of minority class instances, total number of instances, the % of instances in the minority class, the number of attributes, and the Average AUC value. This average AUC value is a measure of the inherent difficulty of building a model on the dataset. This value is based on classification models built on unaltered datasets 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 [18]). Note that the results from these classifiers were used only to determine how challenging it is to build a effective model on the dataset and have no further bearing on the rest of the experiment presented in this study. All models were built using the Weka data mining toolkit [8]. All seven datasets are DNA microarray datasets from a number of different bioinformatics and medical projects. We chose these datasets because they are all relatively balanced datasets (smallest minority class percentage is 42.50%) and none of the datasets can be considered trivial in terms of difficulty.

B. Feature Ranking Techniques

Feature selection has become a necessary step when analysing bioinformatics datasets. In this work we utilize two filter-based feature rankers: Area Under the Receiver Operating Characteristic Curve (ROC) and Signal-to-Noise (S2N). These techniques were chosen for their relatively good performance in terms of classification [6].

ROC is Threshold-Based Feature Selection (TBFS) technique. TBFS treats feature values as ersatz posterior probabilities and classifies instances based on these probabilities, allowing us to use performance metrics as filter-based feature selection techniques. The TBFS technique which uses Area Under the Receiver Operating Characteristic Curve as its performance metric has been shown to be a strong ranker. For details on TBFS and the ROC metric please refer to Abu Shanab et al. [1].

S2N is a measure of how well a feature separate the two classes. The metric is the ratio of the difference between the mean values of that feature form the positive

Fig. 1: Select-Bagging

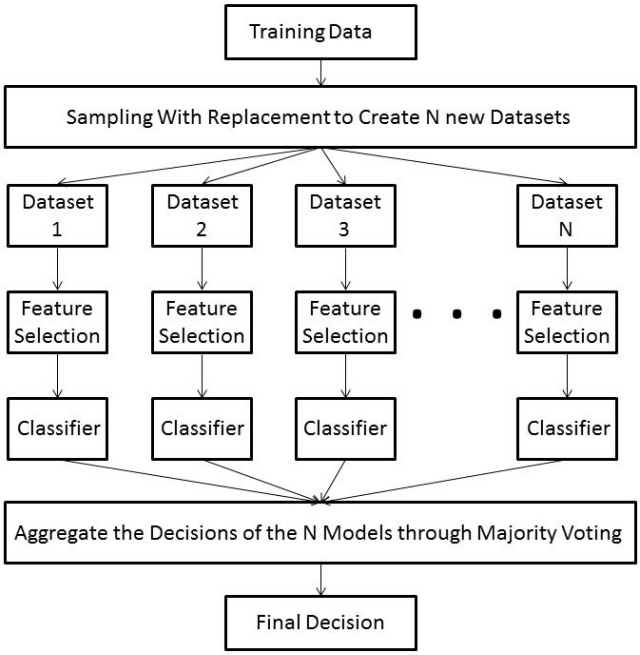


TABLE I: Details of the Datasets

Name	# Minority Instances	Total # of Instances	% Minority Instances	# of Attributes	Average AUC
DLBCL [16]	23	47	48.94%	4027	0.8675
Prostate [17]	59	136	43.38%	12601	0.7823
Breast Cancer [17]	46	97	47.42%	24482	0.6009
DLBCL NIH [17]	102	240	42.50%	7400	0.5853
BCancer50k [5]	200	400	50.00%	54614	0.8564
Spira2007 [14]	90	192	46.88%	22216	0.6661
SotiriouMatrixData-Grade [13]	45	99	45.45%	7651	0.6325

class instances and the negative class instances over the difference between the standard deviations of that feature for the positive class and the negative class instances. The larger the S2N ratio, the more relevant a feature is to the dataset [10].

C. Classifiers

In this paper, we use the Naïve Bayes (NB) classifier to build the inductive models. Because this is a well-understood technique, there is only a brief discussion of the learner here. Naïve Bayes [19] 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. All of the models in this paper were built using the Weka data mining toolkit [8] with the default parameter values.

D. Cross-Validation and Performance Metric

Cross-validation [11] 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 either FS-Classifier or the Select-Bagging

process for each of the twenty training datasets generated by every occurrence of 4 runs of 5-fold cross-validation, which results in 20 models per instance for FS-Classifier and 200 models for the Select-Bagging technique.

We use Area Under the ROC Curve [7] 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

In this study, we sought to observe how the inclusion of bagging affect classification performance on high-dimensional bioinformatics datasets. To conduct the study, we used a series of seven high-dimensional datasets along with two filter-based feature rankers, the Naïve Bayes classifier, and four feature subset sizes. Table II contains the results of our experiment. Each value in the table is the average AUC value across the seven datasets for each combination of feature selection technique, feature subset size, and choice to include bagging. The top performer between Select-Bagging and FS-Classifier is in **boldface** for every combination of feature ranker and subset size. It should be noted that in the tables FS-Classifier.

As we see from the results, Select-Bagging outperforms single feature selection for every scenario. In terms of differences in performance, the smallest difference occurs when using subset size 25 and the largest at subset size 50. This trend holds true for both feature rankers. When we observe the results from each dataset (not shown due to space considerations) we see that Select-Bagging outperforms single feature selection for 39 of the possible 56 (69.64%) scenarios across the seven datasets and both feature rankers. For the individual rankers Select-Bagging outperformed single feature selection 22 out of 28 scenarios (78.57%) for ROC and 17 out of 28 scenarios (60.71%) for S2N.

In order to further validate the results in our classification experiments, we performed a one-factor ANalysis Of VAriance (ANOVA) test [2] with the choice of data sampling approach being the factor, across the seven datasets to determine if the choice of including bagging has any significant effect on the AUC levels. When we look at Table III we see that the choice of bagging is a significant factor. This is shown by the Prob>F score being below 0.05. Additionally we performed a multiple comparison test using Tukey's Honestly Significant Difference (HSD) test [2]. Figure 2 shows that the two approaches yield significantly different performances and that Select-Bagging is the top performer. Thus, based on the results and the statistical analysis we can state that the inclusion of bagging is beneficial to classification performance on relatively balanced bioinformatics datasets.

VI. Conclusion

Ensemble learning has the potential to be a powerful tool for improving the performance of inductive models. However, one of the downsides to ensemble learning is the computational expense. The question remains, does the inclusion of ensemble learning improve the classification performance enough to implement it. This study focuses on the ensemble approach Select-Bagging and compares its classification performance to that of FS-Classifier. We test these two approaches using a series of seven balanced bioinformatics datasets, two feature rankers, and the Naïve Bayes classifier.

Our results show that the models which implement Select-Bagging outperform those which just use FS-Classifier. Looking across the datasets, for all subset sizes and feature rankers, Select-Bagging outperforms FS-Classifier. Additionally, when we observe the results for each dataset we see that, for a majority of scenarios, Select-Bagging outperforms FS-Classifier. To confirm our results we performed a one factor ANOVA test which shows that the choice of including Select-Bagging is significant. Tukey's HSD test show that the two approaches are disjoint and that Select-Bagging performs better. Therefore, we recommend that the inclusion of Select-Bagging is significantly beneficial to the classification performance of model built from balanced bioinformatics data.

Future work in this area consists of implementing other ensemble methods (e.g., boosting) to see if the same trends occur. In addition, by implementing these other ensemble methods we can determine which one is the most appropriate.

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.

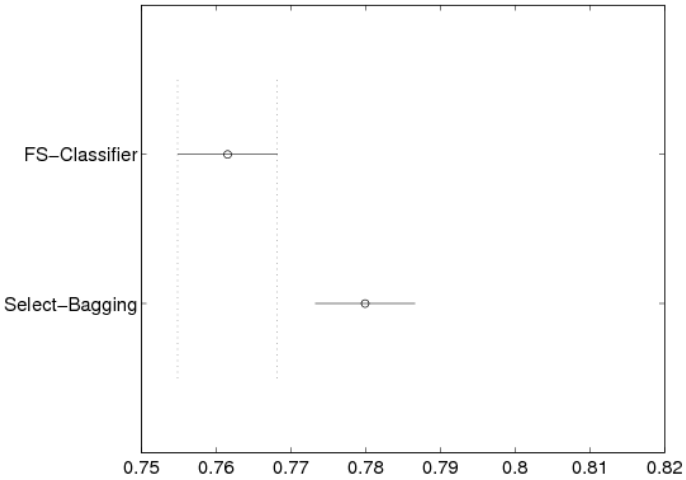
TABLE II: Classification Results - Feature Selection and Select-Bagging

Subset Size	ROC		S2N	
	FS-Classifier	Select-Bagging	FS-Classifier	Select-Bagging
25	0.78332	**0.79034**	0.77452	**0.78477**
50	0.75652	**0.78287**	0.76116	**0.78551**
100	0.75492	**0.77476**	0.75885	**0.77662**
200	0.74895	**0.77193**	0.75423	**0.77282**

TABLE III: ANOVA Results: FS-Classifier vs. Select-Bagging

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Include Bagging	0.1895	1	0.18948	7.28	0.007
Error	58.2199	2238	0.02601		
Total	58.4094	2239			

Fig. 2: Tukey's HSD Test: FS Only vs. Select-Bagging



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

[3] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996. [Online]. Available: http://dx.doi.org/10.1007/BF00058655

[4] T. Chen, "A selective ensemble classification method on microarray data." *Journal of Chemical & Pharmaceutical Research*, vol. 6, no. 6, p. 28602866, 2014.

[5] 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.

[6] 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.

[7] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861 – 874, 2006.

[8] 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.

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

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

[11] 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.

[12] S. Nagi and D. Bhattacharyya, "Classification of microarray cancer data using ensemble approach," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 2, no. 3, pp. 159–173, 2013.

[13] C. Sotiriou, S.-Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, P. Martiat, S. B. Fox, A. L. Harris, and E. T. Liu, "Breast cancer classification and prognosis based on gene expression profiles from a population-based study," *Proceedings of the National Academy of Sciences*, vol. 100, no. 18, pp. 10 393–10 398, 2003.

[14] A. Spira, J. E. Beane, V. Shah, K. Steiling, G. Liu, F. S. Schem-

bri, Gilman, Y.-M. Dumas, P. Calner, P. Sebastiani, S. Sridhar, J. Beamis, C. Lamb, T. Anderson, N. Gerry, J. Keane, M. E. Lenburg, and J. S. Brody, "Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer," *Nature Medicine*, vol. 13, pp. 361–366, 2007.

[15] A. C. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification," *Applied bioinformatics*, vol. 2, no. 3 Suppl, p. S7583, 2003.

[16] 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.

[17] ——, "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.

[18] 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.

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

[20] P. Yang, Y. Hwa Yang, B. B Zhou, and A. Y Zomaya, "A review of ensemble methods in bioinformatics," *Current Bioinformatics*, vol. 5, no. 4, pp. 296–308, 2010.