2014 IEEE 14th International Conference on Bioinformatics and Bioengineering

Selecting The Appropriate Data Sampling Approach for Imbalanced and High-Dimensional Bioinformatics Datasets

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

***Abstract*—One of the more prevalent problems when working with bioinformatics datasets is class imbalance, when there are more instances in one class compared to the other class(es). This problem is made worse because frequently, the class of interest is also the minority class. A possible solution is data sampling, a powerful tool for combating class imbalance by adding or removing instances to make the dataset more balanced. In addition to the choice of including data sampling, one of the most important decisions when applying data sampling is what the final class ratio should be. Commonly, the final class ratio when data sampling is applied is 50:50, however it is an open question whether other ratios are more appropriate for certain imbalanced datasets (all datasets in this paper have 25.16% minority instances or less) where a 50:50 ratio requires extreme modification to the dataset. In this work we compare six different data sampling approaches (feature selection with the pairwise combinations of three data sampling techniques and two final class ratios) with feature selection without data sampling with the goal of determining if the inclusion of data sampling is beneficial and if so, what should be the final class ratio. In order to test the six data sampling approaches and feature selection alone thoroughly, we utilize seven imbalanced and high-dimensional datasets, three feature selection techniques, and six classifiers. Our results show that for a majority of scenarios, random undersampling along with either 35:65 or 50:50 is the best data sampling approach. Statistical analysis shows that there is no significant difference between the data sampling approaches. However, despite this, we still recommend using random undersampling along with 35:65 as the final class ratio. This is because of the frequency of both random undersampling and 35:65 being the most frequent top performing data sampling technique and class ratio respectively. Additionally, 35:65 will have fewer negative impacts than 50:50 (less data loss or overfitting, which makes it a better choice if all other factors are equal) and random undersampling is more computationally efficient than any other form of sampling, including "no sampling" (both by not requiring any internal calculations and by producing a reduced, easier-to-work-with dataset). To our knowledge, this is the most comprehensive work which focuses on the choice of the inclusion and implementation of data sampling with different final class ratios on bioinformatics datasets which exhibit such large levels of class imbalance.**

***Keywords*-Class Imbalance; Class Ratio; Data Sampling; DNA Microarray;**

I. Introduction

Class imbalance (when one class has more instances than the other class(es)) is a frequently encountered problem in bioinformatics which can lead to increased bias toward the majority class and an increased number of false classifications. Additionally, the minority class is frequently the class of interest, making these classification errors even more damaging. However, there is a powerful set of techniques designed to combat class imbalance: data sampling.

Data sampling uses the addition or removal of instances to transform the imbalanced dataset into a more balanced dataset. There are a number of different forms of data sampling that can be utilized including: randomly adding duplicates from the minority class, randomly removing instances from the majority class, or synthetically creating instances for the minority class based on the original ones.

In addition to the choice of technique, one of the most important decisions to make when applying data sampling is what the final class ratio should be. Commonly, researchers decide to make the sampled dataset to be perfectly balanced between the two classes (class ratio 50:50). However, for some cases of extreme class imbalance, making the classes perfectly balanced may not be appropriate: for undersampling approaches, this may lead to an excessive number of majority-class instances being removed from the dataset, while for oversampling approaches, this can create repetition within the minority class that will lead to overfitting.

In this work, we seek to determine if the inclusion of data sampling is recommended and if so, whether less-aggressive class ratios have the potential to be just as effective as the 50:50 ratio on imbalanced bioinformatics datasets. To accomplish this, we compare feature selection without data sampling with six different data sampling approaches: feature selection with one of six combinations of three data sampling techniques (random undersampling (RUS), random oversampling (ROS), and the Synthetic Minority Oversampling TEchnique (SMOTE) [1]) and two final class ratios (50:50 and 35:65). The 35:65 ratio was chosen because this study seeks to explore the potential of less-aggressive ratios in the bioinformatics application domain, not to quantify the precise ratio which will maximize performance and because 35:65 is an appropriate class ratio [12], [17]. These six data sampling approaches and feature selection alone were tested on a series of seven bioinformatics datasets which exhibit extreme degrees of class imbalance. In addition to the datasets and the data sampling techniques, we used a series of three feature selection techniques and six classifiers.

Our results show that in a majority of scenarios, RUS with 50:50 and RUS with 35:65 are the two top performing data sampling approaches. Additionally, we found that for a large majority of scenarios, an approach with data sampling outperforms feature selection alone, and at no point does feature selection alone become the top performing choice for any combination of feature ranker and classifier. Statistical analysis shows that there is no significant differences between the six data sampling approaches and feature selection alone. However, all other things being equal, 35:65 would be preferred over 50:50 due to its reduction of data loss (for random undersampling) or overfitting (for random oversampling and SMOTE); 50:50 has no inherent benefits for highly-imbalanced datasets (those which are naturally more imbalanced than 35:65 to start with). Among the three sampling techniques and the "no sampling" baseline approach, RUS would also be preferred because by reducing the dataset size, it reduces computational complexity of

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.61

modeling. Because of these facts, and because 35:65 is frequently the top performing class ratio, we recommend that RUS along with 35:65 be used for the data sampling approach on highly-imbalanced datasets. To our knowledge, this is the most comprehensive work regarding the choice of the inclusion of data sampling and the class ratio choice on severely imbalanced datasets in the domain of bioinformatics.

The remainder of this paper is organized as follows. Section II contains some related works to our topic. Section III outlines the case study used in this work. Section IV contains the results of our experiments. Lastly, Section V presents our conclusions and possible avenues for future work.

II. Related Works

Class imbalance is a frequent problem within bioinformatics datasets [3]. An example of class imbalance is demonstrated by Van Hulse et al. [22] who compared the correlations among nine rankers on five imbalanced datasets and a number of data sampling approaches (algorithms to improve the balance of datasets). Ramaswamy et al. [20] performed feature selection on a dataset where only 16% of the instances are in the class of interest. The presence of class imbalance has the potential to affect the classification performance of classifiers applied toward these imbalanced datasets. To ensure that we take these effects into account, the datasets in the present work all clearly exhibit class imbalance.

A possible reason why class imbalance tends to affect classification performance may be due to the fact that many classification algorithms assume that the classes will have an equal number of instances in the dataset [14]. This assumption can lead to some serious problems including increased bias against the minority class (which is frequently the class of interest) and an increased number of misclassifications [2]. One recommendation for combating some of these issues is applying data sampling methods. These work by either adding instances to the minority class (oversampling) or removing instances from the majority class (undersampling).

In 2005, Al-Shahib et al. [2] performed a study which utilized data sampling with multiple class ratios. The goal of the work was to see if the addition of a data sampling technique could improve the classification performance of protein function prediction. Their dataset consisted of 1,151 proteins (instances) from thirteen different functional groups with each of the protein samples being represented by a feature space of 433 features. In terms of data sampling they used random undersampling along with five different levels of undersampling ranging from no data sampling being applied to having the two classes become perfectly balanced. These classification models were evaluated using the Area Under the ROC Curve metric (AUC). Their findings were that the addition of the data sampling technique did improve classification performance, so as long as the data sampling was used to perfectly balance the two classes (a final class ratio of 50:50).

However, there are a number of differences between the present work and Al-Shahib et al. The first is that Al-Shahib et al. only uses a single dataset (even though it uses thirteen classification schemas), which raises the question of how applicable their results are to other datasets. The present work uses a collection of seven unique bioinformatics datasets from a variety of different genetics, medical, and biomedical studies. Next, all of the datasets in our work have many more features than the dataset in Al-Shahib et al. (between 12,066 and 54,614 features, compared to the 433 features in Al-Shahib et al.). Another major difference is that Al-Shahib et al. only uses a single data sampling technique (random undersampling) whereas this work uses three: random undersampling, random oversampling, and SMOTE. Additionally, only a single wrapper-based feature selection technique was used in Al-Shahib et al., while three filter-based feature selection techniques were used in this work. It should be noted that wrapper approaches such as those used by Al-Shahib et al. can become computationally infeasible on very high-dimensional datasets such as those used in the present work. Lastly, they introduce variability by bootstrapping (sampling with replacement) the test dataset twenty times which does not test the variability of the process of building the inductive model. The present work utilizes four runs of five-fold cross validation for each dataset and the entire model building process, including feature selection and data sampling, was repeated for every training dataset created.

In 2012, Blagus et al. [5] performed a study which included utilizing data sampling techniques on high dimensional bioinformatics datasets. In this work they used two sampling approaches: random undersampling and SMOTE. In their work they used a series of three breast cancer datasets each with two binary classification schemas (two distinct binary-class attributes) for a total of six datasets. The balance levels of the datasets ranged from 14% to 45%. The t-statistic was used for feature selection, with the features ranked based on their values and the top 40 features used for classification. The results of the work were that only the k-NN classifier seemed to benefit significantly from SMOTE, and this benefit was larger as the number of neighbors increased. However, for most of the other classifiers, it seemed that random undersampling was more useful than SMOTE.

There are a number of differences between the Blagus et al. work and the present work. The most important factor is that unlike Blagus et al., the present work only uses datasets which clearly exhibit an inherent class imbalance. Another difference is that while both papers apply cross-validation, Blagus et al. applies leave one out cross validation (the test dataset consists of only a single instance), whereas the present work uses four runs of five-fold cross validation. Also, Blagus et al. performed feature selection outside of the cross-validation process and then performed either no data sampling technique, SMOTE, or random undersampling. If a data sampling technique was applied, they created 50 new balanced datasets to train the models on. Additionally, this work uses an additional data sampling techniques when compared to the Blagus et al. work: random oversampling. Another major difference is that the Blagus et al. work only use a single filter-based feature selection technique, while this work uses a series of three filter-based feature selection techniques. Also, this work uses a series of seven different bioinformatics datasets (two of which were in Blagus et al.), compared to the three (each with two classification schemas) that Blagus et al. uses. Lastly, Blagus et al. only uses the balance level 50:50 in their microarray experiments. As a result, this work is a more comprehensive analysis for the topic of data sampling and class ratio selection.

III. Case Study

A. Data Sampling approaches

In this work, we use three different data sampling techniques which are applied after feature selection occurs: random undersampling, random oversampling, and Synthetic Minority Oversampling TEchnique or SMOTE [1], [11]. Random undersampling (RUS) seeks to create balance between the two classes by reducing the size of the majority class. This is accomplished by randomly removing instances from the majority class until the desired class ratio has been achieved. Alternatively, random oversampling (ROS) seeks to improve the class balance by increasing the size of the minority class. The increase is performed through randomly duplicating instances from the minority class until the desired class ratio is achieved.

SMOTE is another form of oversampling which seeks to improve the balance between the two classes through the increasing the size of the minority class. However, unlike random oversampling, SMOTE does not duplicate instances. Instead SMOTE creates new minority instances using the original ones as a basis. It starts with an instance from the minority class and looks at its nearest neighbor. Once the nearest neighbor has been identified, the differences between the two instances in terms of each feature is calculated. Finally a new instance is created by adding the product of the differences calculated and a random number between 0 and 1 to the original instance.

However, all three techniques have their downsides. Since random undersampling removes instances from the majority class, it is removing data from consideration for the model. Alternatively, random oversampling and to a lesser extent SMOTE run the risk of overfitting because of the addition of the duplicated or synthetic minority instances. As a result we decided to use two different final class ratios: 50:50 and 35:65. The two class ratios were chosen because 50:50 is the most common final class ratio for data sampling and previous research has shown that 35:65 is an appropriate final class ratio [12], [17].

In addition to the results with the inclusion of data sampling, we also present the results of feature selection without data sampling to determine if the inclusion of data sampling has some benefits. The goal of this paper was not to identify the absolute best solution: rather, we wanted to see if the inclusion of data sampling is beneficial and if we could achieve similar or even better classification performance while applying smaller changes to the data, which would enable us to create a model with less data loss or overfitting. In addition, the scope of this work (with the large collection of datasets, feature selection techniques, classification learners, and sampling techniques) did not permit us to consider a large collection of class ratios: the two we chose were sufficient to demonstrate the effectiveness of a less-aggressive class ratio.

B. Feature Selection Techniques

For all data sampling approaches, first step is applying feature selection techniques to reduce the feature set to a more manageable size. In this work we use three filter-based feature selection techniques: Information Gain (IG), Signal-to-Noise S2N, and Area under the ROC Curve (ROC). All three are filter-based feature selection techniques which have exhibited varying levels of stability and average to good classification performance. We feel this consideration is important as feature selection stability can affect the performance of the overall inductive model building process. According to previous research [9] we see that IG has average to below average stability; S2N is above average in terms of stability, and ROC is one of the most stable feature selection techniques. When using feature selection a final feature subset size must be chosen. In our work we use 25 features which, based on previous research, is a reasonable number of features [15].

Information Gain (IG) [13] 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. IG determines the significance of a feature based on the amount by which the entropy of the class decreases when considering that feature. S2N 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 difference between the standard deviation of that feature for the positive class and the standard deviation of that feature for the negative class. The larger the S2N ratio, the more relevant a feature is to the dataset [15].

ROC is Threshold-Based Feature Selection (TBFS) technique used in conjunction with the Area Under the Receiver Operating Characteristic (ROC) Curve metric. 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 ROC 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].

C. Classifiers

Six classifiers were considered in this work: 5-Nearest Neighbors (5-NN), Logistic Regression (LR), Multilayer Perceptron (MLP), Naïve Bayes (NB), Random Forest with 100 trees (RF100), and Support Vector Machines (SVM). All were implemented within the Weka data mining toolkit [24], with default parameters unless otherwise noted. We utilized the same parameter values for all models built in this work because the focus of this work was on comparing two different class ratios for sampling, not on tweaking the classification models, and using consistent parameters allowed us to make meaningful comparisons. Note that any changes to default parameter values were applied when experimentation showed an overall improvement of the classification performance [23]. Due to space limitations, all of the learners are described very briefly; please consult Witten et al. [24] for further information.

5-Nearest Neighbors is an instance-based classifier which, in order to classify a new instance, finds the five closest neighbors within the training set and takes a weighted vote (weighted by 1/Distance) of their class values. Logistic Regression builds a simple logistic regression model based on the training data. Multilayer Perceptron is a feed-forward artificial neural network learner wherein each node takes its value by finding the sigmoid of the weighted sum of its inputs; our implementation used a single hidden layer with three nodes, and held back 10% of the training data for validating when to stop the backpropagation-based training. Naive Bayes 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. RF100 is an ensemble classifier which builds a set of unpruned decision tree then uses majority voting on the resulting trees to perform prediction. Previous research [16] shows that the optimum number of trees is 100, so that is the number used in this study. Support Vector Machines attempts to find the hyperplane which best separates the two classes in feature-space; our version has its complexity constant set to 5.0 and its `buildLogisticModels` parameter set to `true`.

D. Cross Validation

Cross-validation [22] 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 both the feature selection and data sampling processes was performed inside the cross-validation step: that is, for each run and each "training set" within the cross-validation procedure, feature selection for subset size twenty-five was performed followed by data sampling to achieve a "training set" whose class ratio is either 35:65 or 50:50. The inductive models are trained using the reduced feature set and the sampled "training set".

E. Performance Metric

All classification performance is evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC) [10]. AUC seeks to find how the model balances false positives with false negatives at different decision thresholds. This is achieved by plotting the true positive rate (number of true positives/total number of positives) versus false positive rate (number of false positives/total number of negatives) over all of the decision thresholds. The area under the curve is used to describe the quality of the model. This metric was chosen because of the imbalanced nature of the datasets used in this study. It should be noted that this is distinct from the feature selection technique ROC discussed above.

F. Datasets

Table I 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 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 relapse/no relapse following cancer treatment). The datasets in Table I 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 are imbalanced enough to be able to utilize the 35:65 class ratio (the largest minority class distribution is 25.16% which is well below the 35% from the final class ratio).

The last column, Average AUC refers to the classification performance on these datasets when building models without feature selection or data sampling. The values were calculated using a set of six different classification models: 5-NN, MLP, NB, SVM, and two versions of a C4.5 decision tree (C4.5 D and C4.5 N). Descriptions of all of the learners are found in Section III-C with the exception of C4.5 D and C4.5 N. C4.5 D is the C4.5 decision tree classifier (implemented as J48 within WEKA) with the default parameter values. C4.5 N is the same classifier but with Laplace smoothing enabled and pruning disabled. Both are available using the WEKA Data Mining toolkit [24]. These results are used only to determine the difficulty of the datasets and have no further bearing on the rest of the experiment. The values show that while the datasets represent different levels of difficulty they are not trivial and therefore are excellent for comparing different techniques [7].

IV. RESULTS

In this study, we compare feature selection with no data sampling with six different data sampling approaches using a series of seven imbalanced, high-dimensional bioinformatics datasets, three feature selection techniques, and six classifiers. Table II contains the results of our experiment. Each number is the average AUC value across all of the datasets where the data sampling approach, class ratio when applicable, feature selection technique, and classifier are kept static. In terms of the results using data sampling, for every combination of data sampling technique, feature selection technique, and classifier the top performer between 50:50 and 35:65 is in **boldface**. Additionally, the top performer across all data sampling approaches and class ratios for each classifier will be underlined.

We begin with our comparison of feature selection with data sampling compared to feature selection without data sampling. The results show that for 85 of the 108 scenarios of data sampling with feature selection being compared to just feature selection, that the data sampling with feature selection shows better performance than just feature selection alone. Additionally, when we look at the top performer for each combination of classifier and feature selection technique, at no time does using feature selection alone outperform all possible approaches for data sampling.

Now that we see that data sampling does have a benefit, we must now look at the choice of final class ratio. When we look at the results using RUS we see that in twelve out of eighteen scenarios (66.7%), 35:65 outperforms 50:50. We see the same trend in terms of the individual feature selection techniques, with 35:65 outperforming 50:50 for 66.7% of the scenarios. In terms of the individual classifiers we see that for each classifier 35:65 will outperform 50:50 in at least 66.7% of the scenarios with the exception of 5-NN which has 50:50 outperforming 35:65 for all scenarios using RUS. However, for LR and RF100 35:65 outperforms 50:50 for all scenarios.

Looking at ROS, we see that for thirteen of the eighteen scenarios (72.2%) 35:65 outperforms 50:50. In terms of the individual feature selection techniques, we see that for each feature selection technique 35:65 will outperform 50:50 for at minimum 50% of the scenarios. In fact, the feature selection technique IG has 35:65 as the top performing class ratio for all scenarios. In terms of the individual classifiers we see that for each classifier 35:65 will outperform 50:50 in at least 50% of the scenarios for all scenarios

Table I
DETAILS OF THE DATASETS

Name	# Minority Instances	Total # of Instances	% Minority Instances	# of Attributes	Average AUC
GSE1456 [19]	40	159	25.16%	12066	0.6108
GSE20271 [21]	26	178	14.61%	22284	0.5867
GSE25055 [6]	57	306	18.63%	22284	0.6674
GSE25065 [6]	42	182	23.08%	22284	0.6384
GSE3494-GPL96-ER [18]	34	247	13.77%	22284	0.7688
GSE3494-GPL96-Grade [18]	54	249	21.69%	22284	0.8176
Lung 50k [8]	70	400	17.50%	54614	0.8150

Table II
CLASSIFICATION RESULTS: DATA SAMPLING APPROACHES

Feature Selection Technique	Classifier	Data Sampling approaches						
		None	RUS		ROS		SMOTE	
			35:65	50:50	35:65	50:50	35:65	50:50
IG	5NN	0.74699	0.76571	**0.77382**	**0.74887**	0.74226	**0.75668**	0.75407
	LR	0.73369	**0.72706**	0.69496	**0.73952**	0.73871	**0.73753**	0.73455
	MLP	0.76290	0.77882	**0.78020**	**0.75775**	0.74962	**0.75964**	0.75381
	NB	0.79395	**0.80644**	0.80289	**0.80585**	0.80510	**0.80262**	0.80082
	RF100	0.78543	**0.79975**	0.79491	**0.79365**	0.79241	**0.79764**	0.79515
	SVM	0.76941	**0.78481**	0.77905	**0.78100**	0.78019	**0.77947**	0.77452
ROC	5NN	0.75668	0.78226	**0.79442**	**0.77681**	0.77202	**0.78057**	0.77275
	LR	0.73371	**0.71295**	0.70715	0.74210	**0.74259**	**0.74026**	0.73637
	MLP	0.76516	**0.78739**	0.78636	0.76281	**0.76768**	**0.77162**	0.75595
	NB	0.80350	**0.80330**	0.80168	**0.80918**	0.80875	**0.81044**	0.81000
	RF100	0.79000	**0.80251**	0.80199	0.80353	**0.80645**	0.80553	**0.80563**
	SVM	0.76741	0.79095	**0.79281**	**0.79193**	0.79060	**0.79187**	0.78666
S2N	5NN	0.74728	0.77195	**0.77751**	**0.74671**	0.74088	**0.75544**	0.75269
	LR	0.74001	**0.71915**	0.69507	0.74437	**0.74551**	**0.74264**	0.73871
	MLP	0.77712	**0.78266**	0.78050	**0.77079**	0.75933	**0.77243**	0.75921
	NB	0.79979	0.79995	**0.80120**	0.80057	**0.80074**	**0.80040**	0.79937
	RF100	0.78680	**0.79270**	0.79021	**0.79424**	0.78935	**0.79124**	0.79027
	SVM	0.77347	**0.78267**	0.78172	**0.78363**	0.77955	**0.78065**	0.77736

Table III
ANOVA RESULTS: DATA SAMPLING APPROACHES

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Technique	0.083	6	0.01379	0.71	0.6386
Error	340.736	17633	0.01932		
Total	340.818	17639			

using ROS with the exception of LR. Additionally, for 5-NN and SVM 35:65 outperforms 50:50 for all feature selection techniques.

Lastly, when we look at SMOTE, we see that for seventeen out of eighteen scenarios (94.4%) 35:65 outperforms 50:50. In terms of the individual feature selection techniques and classifiers we see that 35:65 is the most frequent top performing class ratio. Additionally, we see that for the feature selection techniques IG and S2N as well as the classifiers LR, MLP, NB, and SVM that 35:65 outperforms 50:50 for all scenarios.

When we look across all of the factors, we see that 35:65 outperforms 50:50 for forty-two of the fifty-four (77.8%) scenarios. Additionally, there is no classifier or feature selection technique in which 35:65 is not the most frequent top performing class ratio across all three data sampling techniques. These facts allow us to state that 35:65 allows for better classification performance than 50:50.

In terms of the top performing data sampling approach to choose for each combination of feature selection technique and classifier, we see that RUS is the most frequent top performing data sampling technique with eleven of the possible eighteen (61.1%) combinations: five for 35:65 and six for 50:50. However, despite 50:50 having one more top performing approach count than 35:65, because 35:65 more frequently outperforms 50:50 within RUS and the lower data loss due to utilizing 35:65, we still recommend using 35:65 over 50:50 when using RUS.

In order to further validate the results in our classification experiments, we performed a one-factor ANalysis Of VAriance (ANOVA) test [4] with the choice of data sampling approach being the factor, across the seven datasets to determine if the choice of data sampling approach has any significant effect on the AUC levels. When we look at Table III we see that the choice of data sampling approach is not a significant factor. This is shown by the Prob>F score being above 0.05. While this indicates that the choice of data sampling approach is not significant, it is our recommendation to utilize RUS with 35:65 as the class ratio due to its frequency as the top performing data sampling approach, because the class ratio of 35:65 can reduce the amount of data loss or overfitting cause by sampling to a ratio of 50:50, and because

RUS will reduce the total amount of data and make further data mining procedures more computationally efficient.

V. Conclusion

Class imbalance is a prevalent problem found in bioinformatics datasets which can cause a number of complications for data analysis. A potential tool for combating class imbalance is data sampling. Commonly, the end results of data sampling is a perfectly balanced dataset (a class ratio of 50:50). However, the question remains: is the inclusion of data sampling beneficial and if so, is 50:50 the appropriate final class ratio for all scenarios, especially for datasets with extremely high levels of class imbalance? In this work, we compare feature selection with no data sampling with six different data sampling approaches: feature selection with one of six combinations of three data sampling techniques (RUS, ROS, and SMOTE) and two final class ratios (50:50 and 35:65). In order to test the six data sampling approaches along with feature selection alone we used a series of seven highly imbalanced datasets along with three feature selection techniques, and six classifiers.

Our results show that for a majority of scenarios RUS with 50:50 and RUS with 35:65 are the two most frequent top performing data sampling approaches. Additionally, we found that for a large majority of scenarios that an approach with data sampling outperforms the approach without data sampling and at no point does the feature selection alone become the top performing choice for any combination of feature ranker and classifier. We also performed a one factor ANOVA tests where the factor was the choice of of the data sampling approach. It was found that the choice of data sampling approach was not significant. Nonetheless, because RUS is the most frequent top performing data sampling technique (while being more computationally efficient than any other approach, including "no sampling") and given that the 35:65 class ratio can mitigate some of the data loss or overfitting of sampling with a 50:50 ratio, and because 35:65 did perform better than 50:50 more often than the reverse with all three data sampling techniques, we recommend using RUS with 35:65 as the class ratio for bioinformatics datasets which are highly imbalanced.

Future work in this area consists of including more datasets both within this application domain and in other domains to further confirm our results as well as other approaches towards combining data sampling and feature selection. In addition, works could consider a larger range of class ratios, although the lack of statistical significance found in this paper suggests that such variations might only have slight impacts on classification performance.

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

[7] D. J. Dittman, T. M. Khoshgoftaar, R. Wald, and A. Napolitano, "Maximizing classification performance for patient response datasets," in *Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference on*, Nov 2013, pp. 454–462.

[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 H. Wang, "Stability analysis of feature ranking techniques on biological datasets," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. BIBM, 2011, pp. 252–256.

[10] D. J. Dittman, T. M. Khoshgoftaar, R. Wald, and J. Hulse, "Feature selection algorithms for mining high dimensional dna microarray data," in *Handbook of Data Intensive Computing*. Springer New York, 2011, pp. 685–710.

[11] D. J. Dittman, T. M. Khoshgoftaar, R. Wald, and A. Napolitano, "Comparison of data sampling approaches for imbalanced bioinformatics data," in *Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference, FLAIRS 2014*, 2014, pp. 268–271.

[12] K. Gao, T. M. Khoshgoftaar, and A. Napolitano, "Impact of data sampling on stability of feature selection for software measurement data," in *Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on*, Nov 2011, pp. 1004–1011.

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

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

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

[16] T. M. Khoshgoftaar, M. Golawala, and J. Van Hulse, "An empirical study of learning from imbalanced data using random forest," in *19th IEEE International Conference on Tools with Artificial Intelligence, 2007. ICTAI 2007.*, vol. 2, October 2007, pp. 310–317.

[17] T. Khoshgoftaar, C. Seiffert, J. Van Hulse, A. Napolitano, and A. Folleco, "Learning with limited minority class data," in *Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on*, Dec 2007, pp. 348–353.

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

[19] 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

[20] S. Ramaswamy, K. N. Ross, E. S. Lander, and T. R. Golub, "A molecular signature of metastasis in primary solid tumors," *Nature Genetics*, vol. 33, pp. 49–54, 2003.

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

[22] 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.

[23] 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

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