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

Comparison of Data Sampling Approaches for Imbalanced Bioinformatics Data

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

Abstract

Class imbalance is a frequent problem found in bioinformatics datasets. Unfortunately, the minority class is usually also the class of interest. One of the methods to improve this situation is data sampling. There are a number of different data sampling methods, each with their own strengths and weaknesses, which makes choosing one a difficult prospect. In our work we compare three data sampling techniques (Random Undersampling, Random Oversampling, and SMOTE) on six bioinformatics datasets with varying levels of class imbalance. Additionally, we apply two different classifiers to the problem (5-NN and SVM), and use feature selection to reduce our datasets to 25 features prior to applying sampling. Our results show that there is very little difference between the data sampling techniques, although Random Undersampling is the most frequent top performing data sampling technique for both of our classifiers. We also performed statistical analysis which confirms that there is no statistical difference between the techniques. Therefore, our recommendation is to use Random Undersampling when choosing a data sampling technique, because it is less computationally expensive to implement than SMOTE and it also reduces the size of the dataset, which will improve subsequent computational costs without sacrificing classification performance.

Introduction

Class imbalance is a problem that is prevalent among bioinformatics datasets and occurs when there is not an even distribution of instances between the classes. Additionally, in binary classification, it is frequently the minority class which is the class of interest. There are a number of problems associated with class imbalance, including: bias towards the majority class, reduced classification performance, and increased number of false negatives.

One of the potential ways of alleviating this issue is through data sampling. Data sampling transforms the dataset by either adding or removing instances in order to achieve a more balanced class ratio. There are a number of different forms that data sampling can take, including undersampling (removing instances of the majority class in either a directed or random manner) and oversampling (adding instances to the minority class in either a directed or random manner).

However, the variety of techniques raises the question of which technique to choose. In this work we compare the classification results when applying one of three data sampling techniques (Random Undersampling, Random Oversampling, and the Synthetic Minority Oversampling TEchnique or SMOTE) to bring the class ratio to 50:50. We test these techniques on a series of six high-dimensional bioinformatics datasets which contain various levels of class imbalance (although even the most balanced still has only 33.55% of instances in its minority class). Additionally, we also use two classifiers (5-Nearest Neighbors and Support Vector Machines) as well as the threshold-based feature selection technique Area Under the ROC Curve to reduce the number of features to twenty-five prior to applying the sampling techniques.

Our results show that there is little difference between the three data sampling techniques, although Random Undersampling is the most frequent top performer. In order to confirm our results we performed an ANalysis Of VAriance test as well as a multiple comparison test using Tukey's Honestly Significant Difference criterion, both of which confirm that there is no statistically significant difference among the three techniques. Due to the results gathered we recommend using Random Undersampling over Random Oversampling and SMOTE for the purposes of data sampling, because its reduced computational costs to implement (compared to SMOTE) and its end result of reducing the size of the dataset (compared to the oversampling techniques) both help to decrease the computational costs of the classification experiments while not reducing performance.

The rest of the paper is organized as follows. The Related Works section contains previous research which relates to our experiment. The Data Sampling section introduces the specifics of the three data sampling techniques used in our work. The Methodology section outline the methodology of our experiment. The Results section presents the results of our work. Lastly, the Conclusion section presents our conclusions and topics for future work.

Related Works

Perhaps the root of the trouble with class imbalance is in how classification algorithms are designed. A majority of classification algorithms assume that the classes involved will have an equal presence in the dataset (He & Garcia 2009). This assumption can lead to some serious problems. For example, feature selection with certain classifiers is known to focus on accuracy and so will focus on the majority class (Al-Shahib, Breitling, & Gilbert 2005). Some recommendations for combating some of these issues include applying data sampling methods. These methods work by either adding instances to the minority class (oversampling) or removing instances from the majority class (undersampling).

However, despite the frequency of imbalanced datasets, there has been little work on data sampling in the domain of bioinformatics, even if the work that does exists shows potential. In 2005, Al-Shahib et al. (Al-Shahib, Breitling, & Gilbert 2005) performed experiments to determine if the addition of a data sampling technique would improve classification results when singling out a single functional group from a set of thirteen. This study showed that applying data sampling to improve the class ratio to 50:50 (with or without feature selection) gave significantly better results than most other combinations of sampling and feature selection.

In 2012, Blagus et al. (Blagus & Lusa 2012) performed a study using data sampling on high-dimensional class-imbalanced DNA microarray data. They used two data sampling techniques, Random Undersampling and SMOTE, on a series of six datasets and a series of classifiers. Their results found that only the k-NN classifiers seem to benefit substantially from SMOTE and a number of the other classifiers seem to prefer Random Undersampling. One downside of this work however, was in the selection of the datasets. Some of the datasets chosen were not particularly imbalanced, with the minority class being as high as 45% of the instances. In these cases, data sampling will have little effect as the classes are fairly balanced to begin with.

Data Sampling

In this work we use three different data sampling techniques: Random Undersampling, Random Oversampling, and Synthetic Minority Oversampling TEchnique or SMOTE (Abu Shanab *et al.* 2012). 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 increasing the size of the minority class. However, unlike Random Oversampling, SMOTE does not duplicate instances. Instead SMOTE creates new minority instances which are interpolated between existing minority-class instances, thus creating a denser minority class. For all three sampling techniques, sampling was performed to create a 50:50 class ratio.

Methodology

Datasets

Table 1 contains the list of datasets used in our experiment along with their characteristics. The datasets are all DNA microarray datasets acquired from a number of different real world bioinformatics, genetics, and medical projects. As the gene selection technique used in this paper requires that there be only two classes, we can only use datasets with two classes (in particular, either cancerous/noncancerous or relapse/no relapse following cancer treatment). The datasets in Table 1 show a large variety of different characteristics such as number of total instances (samples or patients) and number of features. We chose these datasets because they have a variety of different levels of class imbalance but are not considered balanced, as the largest minority percentage is 33.55%.

Gene Selection Technique and Feature Subset Size

Based on previous research (Abu Shanab *et al.* 2012), feature selection was applied prior to data sampling, in order to select the top 25 features. We chose one form of feature selection, Threshold-Based Feature Selection (TBFS) used in conjunction 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. (Abu Shanab *et al.* 2012).

Classification, Cross-Validation, and Performance Metric

We used two different classifiers (learners) to create inductive models using the sampled data and the chosen features (genes). 5 Nearest Neighbor (5-NN) and Support Vector Machines (SVM), implemented using the WEKA toolkit (Witten & Frank 2011). Due to space limitations (and because these two classifiers are commonly used) we will not go into the details of these techniques. We do note that our implementation of SVM uses a complexity constant of 5.0 and the `buildLogisticModels` parameter set to `true`, instead of their default values in WEKA. For more information on these learners, please refer to (Witten & Frank 2011).

Cross-validation refers to a technique used to allow for the training and testing of inductive models without resorting to using the same dataset. In this paper we use five-fold cross-validation. Additionally, we perform four runs of the five-fold cross validation so as to reduce any bias due to a lucky or unlucky split. The classification performance of each model is evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC) (Abu Shanab *et al.*

Table 1: Details of the Datasets

Name	# Minority Instances	Total # of Instances	% Minority Instances	# of Attributes
Brain Tumor (Dittman *et al.* 2013b)	23	90	25.56%	27680
Chanrion 2008 (Dittman *et al.* 2013a)	52	155	33.55%	22657
GSE20271 (Tabchy *et al.* 2010)	26	178	14.61%	22284
GSE3494-GPL96-ER (Pittman *et al.* 2004)	34	247	13.77%	22284
Mulligan R-PD (Dittman *et al.* 2013a)	41	126	32.54%	22284
Ovarian MAT (Dittman *et al.* 2013b)	16	66	24.24%	6001

Table 2: Classification Results for the Three Data Sampling Methods

Dataset	5-NN				SVM			
	RUS	ROS	SMOTE	Range	RUS	ROS	SMOTE	Range
Brain Tumor	**0.88615**	*0.85475*	0.86423	0.03140	*0.91157*	0.92761	**0.92910**	0.01753
Chanrion 2008	*0.79379*	**0.79887**	0.79545	0.00508	**0.80959**	*0.80494*	0.80759	0.00465
GSE20271	**0.64423**	*0.58349*	0.59863	0.06074	**0.67583**	0.66453	*0.65059*	0.02524
GSE3494-GPL96-ER	**0.86873**	*0.84695*	0.86392	0.02178	**0.89342**	0.89065	*0.88848*	0.00494
Mulligan R-PD	0.65502	**0.65739**	*0.63783*	0.01956	0.65372	**0.65417**	*0.65364*	0.00053
Ovarian MAT	**0.93792**	*0.93125*	**0.93792**	0.00667	**0.92792**	*0.90833*	0.90958	0.01959

2012). Mathematically, this is the same metric as described above in the Gene Selection Technique and Feature Subset Size section, but there is a major distinction: for gene selection (denoted as ROC), we an ersatz posterior probability to calculate the metric, but when used for evaluating classification models (denoted as AUC), the actual posterior probability from the model is used.

Results

In this study, we compare the classification results of three data sampling techniques on a series of six high-dimensional bioinformatics datasets whose feature set has been reduced to twenty-five features using the filter-based feature selection technique ROC. We test the performance using two classifiers: 5-NN and SVM. Table 2 contains the results of our experiment. In each row (which is a combination of learner and dataset) the top performing data sampling technique is in **boldface** and the worst performing data sampling technique is in *italics*. The final column represents the difference between the top and worst performing data sampling techniques in order to show the range of values.

Looking at the results using 5-NN (The first section of Table 2), we see that for four of the datasets, Random Undersampling outperforms SMOTE and Random Oversampling. It should be noted however, that Ovarian MAT's Random Undersampling and SMOTE scores are effectively the same (though they are not identical, differing in the eighth decimal place). The remaining two datasets have Random Oversampling outperforming Random Undersampling and SMOTE. In terms of the worst performing data sampling technique, we see that Random Oversampling is the worst performer for four of the datasets. As for the remaining two datasets, Random Undersampling and SMOTE each have a dataset in which they are the worst performing data sampling technique.

When using SVM (the second section of Table 2), we see that like 5-NN, four of the six datasets show Random Undersampling outperforming Random Oversampling and SMOTE. However, unlike 5-NN, for the remaining two datasets Random Oversampling and SMOTE each have a dataset in which they are the top performing data sampling technique. When we look at the worst performing data sampling techniques, we see that SMOTE has three datasets in which it is the worst performing technique. Random Oversampling and Random Undersampling are the worst performing techniques for two and one datasets respectively.

When we look across the data sampling techniques we see that classification results are very similar between the three data sampling techniques. Looking at the "Range" columns in Table 2, we see that for most of the datasets the difference between the top performing data sampling technique and the worst performing technique are relatively small, although the differences in general are larger in 5-NN than SVM. In fact, for five of the dataset/learner combinations the difference between the best and worst performing techniques is ≤ 0.01 AUC. It should also be noted though that Random Undersampling is the most frequent top performing data sampling technique, by being the top performing technique for eight of the possible twelve combinations of learner and dataset.

In order to further validate the results in our classification experiments, we performed two sets of one-factor ANalysis Of VAriance (ANOVA) tests (Berenson, Goldstein, & Levine 1983) (one per classification learner) across the six datasets to determine if the choice of data sampling technique has any significant effect on the AUC levels. The table itself cannot be presented due to space limitations, but the results show that there are no significant differences between the three techniques for either classifier. This result is different from Blagus et al.'s (Blagus & Lusa 2012) findings that SMOTE is significantly better than RUS for k-NN classifiers. We believe this difference may come from the fact that our datasets (unlike Blagus et al.'s) are all clearly imbalanced.

Conclusion

Class imbalance is one of the most frequent problems when working with bioinformatics datasets because often the class of interest is the minority class and a number of classifiers seek to maximize accuracy which is biased towards the majority class. One method to alleviate this problem is data sampling, a process of adding or removing instances in order to improve the class ratio. In our work we compare the classification performance (using two classifiers, 5-NN and SVM) of three data sampling techniques, Random Undersampling, Random Oversampling, and SMOTE, on a series of six high-dimensional class-imbalanced bioinformatics datasets which have had their features reduced with the ROC feature ranking technique.

Our results show that there is little difference among the three data sampling techniques, although Random Undersampling frequently is the best performing of the three techniques. To confirm our results we include a one factor ANOVA test and a test of Tukey's HSD criterion (using the choice of data sampling technique as the factor being examined) and found that there is no statistically significant difference among the three techniques. Notably, this observation is contrary to previous research (Blagus & Lusa 2012). We believe this is due to our selection of dataset which were all clearly imbalanced dataset. Based on our results we recommend using Random Undersampling over Random Oversampling and SMOTE as the data sampling technique due to the smaller computational cost over SMOTE and the creation of a smaller dataset (compared to either oversampling technique), which will reduce computational costs in subsequent analysis with no sacrifice to classification performance.

Future work will consist of increasing the number of datasets used both generally and for focusing on a particular type of dataset (tumor classification, patient response prediction, etc.). Additionally, future work will look into other final class distributions beyond being perfectly balanced.

References

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

Al-Shahib, A.; Breitling, R.; and Gilbert, D. 2005. Feature selection and the class imbalance problem in predicting protein function from sequence. *Applied Bioinformatics* 4(3):195–203.

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

Blagus, R., and Lusa, L. 2012. Evaluation of smote for high-dimensional class-imbalanced microarray data. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume 2, 89–94.

Dittman, D. J.; Khoshgoftaar, T. M.; Wald, R.; and Napolitano, A. 2013a. Maximizing classification performance for patient response datasets. In *Proceedings of the 25th IEEE International Conference on Tools with Artifical Intelligence (ICTAI'13)*, 454–462. IEEE Computer Society.

Dittman, D.; Khoshgoftaar, T.; Wald, R.; and Napolitano, A. 2013b. Gene selection stability's dependence on dataset difficulty. In *Information Reuse and Integration (IRI), 2013 IEEE 14th International Conference on*, 341–348.

He, H., and Garcia, E. A. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 21(9):1263–1284.

Pittman, J.; Huang, E.; Dressman, H.; Horng, C.-F.; Cheng, S. H.; Tsou, M.-H.; Chen, C.-M.; Bild, A.; Iversen, E. S.; Huang, A. T.; Nevins, J. R.; and West, M. 2004. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences of the United States of America* 101(22):8431–8436.

Tabchy, A.; Valero, V.; Vidaurre, T.; Lluch, A.; Gomez, H.; Martin, M.; Qi, Y.; Barajas-Figueroa, L. J.; Souchon, E.; Coutant, C.; Doimi, F. D.; Ibrahim, N. K.; Gong, Y.; Hortobagyi, G. N.; Hess, K. R.; Symmans, W. F.; and Pusztai, L. 2010. Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide chemotherapy response predictor in a multicenter randomized trial in breast cancer. *Clinical Cancer Research* 16(21):5351–5361.

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