ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Data complexity measures for analyzing the effect of SMOTE over microarrays

L. Morán-Fernández and V. Bolón-Canedo and A. Alonso-Betanzos *

Laboratory for Research and Development in Artificial Intelligence (LIDIA),
Computer Science Dept., University of A Coruña, 15071 A Coruña, Spain

Abstract. Microarray classification is a challenging issue for machine learning researchers mainly due to the fact that there is a mismatch between gene dimension and sample size. Besides, this type of data have other properties that can complicate the classification task, such as class imbalance. A common approach to deal with the problem of imbalanced datasets is the use of a preprocessing step trying to cope with this imbalance. In this work we analyze the usefulness of the data complexity measures in order to evaluate the behavior of the SMOTE algorithm before and after applying feature gene selection.

1 Introduction

During the last two decades, advances in molecular genetics technologies —such as DNA microarrays— have allowed the expression levels of thousands of genes to be measured simultaneously, stimulating a new line of research both in bioinformatics and in machine learning (ML). Microarray technology is used to collect information from tissue and cell samples regarding gene expression differences that could be useful for diagnosis diseases, as they enable distinct kinds or subtypes of tumors to be classified according to expression patterns (profiles). The classification of this type of data has been viewed as a particular challenge for ML researchers mainly because of their extremely high dimensionality (from 2000 to 25000 features) in contrast to small samples sizes (often fewer than 100 patients) [1]. The existence of many fields relative to few samples means that false positives findings due to chance are very likely (in terms of both identifying relevant genes and building predictive models). Moreover, several studies have demonstrated that most of the genes measured in a DNA microarray experiment do not actually contribute to efficient sample classification. To avoid this "curse of dimensionality", feature selection (FS) —defined as the process of identifying and removing irrelevant features from the training data— is advisable so as to identify the specific genes that enhance classification accuracy.

Apart from the obvious problem of having an extremely high dimensionality, microarray datasets present other properties than can complicate the classification task, as the imbalance of the data. The class imbalance problem occurs when a dataset is dominated by a major class or classes which have significantly more instances than the other rare/minority classes in the data. For example, in the domain at hand, the cancer class tends to be rarer than the non-cancer class because usually there are more healthy

*This research has been financially supported in part by the Spanish Ministerio de Economía y Competitividad (research projects TIN 2012-37954 and TIN2015-65069-C2-1-R), by European Union FEDER funds and by the Consellería de Industria of the Xunta de Galicia (research project GRC2014/035). V. Bolón-Canedo acknowledges Xunta de Galicia postdoctoral funding (ED481B 2014/164-0).

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

patients in a real situation. However, it is important for practitioners to predict and prevent the appearance of cancer. In these cases, standard classifier learning algorithms have a bias toward the classes with a greater number of samples, since the rules that correctly predict those samples are positively weighted in favor of the accuracy metric, whereas specific rules that predict instances from the minority class are usually ignored (treated as noise), because more general rules are preferred. Therefore, minority class samples are more often misclassified than those from the other classes.

In contrast, it is widely acknowledged that the prediction capacities of classifiers greatly depend on the characteristics of the data as well. Data complexity analysis is a relatively recent proposal by Ho and Basu [2] to identify data particularities which imply some difficulty for the classification task. Several studies have explored the use of data complexity analysis to characterize data and to relate data characteristics to classifier performance over microarray data [3, 4].

The aim of this work is to show that the data complexity measures are adequate to analyze the effect of the preprocessing in unbalanced data for classification. Specifically, we will consider the "Synthetic Minority Oversampling Technique" (SMOTE) [5] over five microarray datasets before and after applying FS, obtaining promising results.

2 Oversampling approach: the SMOTE algorithm

The traditional preprocessing techniques used to overcome the class imbalance problem are undersampling and oversampling methods. Undersampling is a technique which creates a subset of the original datasets by eliminating samples. It aims to attain the same number of samples of the majority class as in the minority class. As we mentioned above about microarray data, a very small number of samples are available in the whole dataset. Consequently, elimination of observations is not a good option for this type of datasets. In contrast, oversampling methods create a superset of the original dataset by replicating some instances or creating new instances from existing ones. Specifically, in this research we have chosen a popular oversampling method, the SMOTE algorithm.

When applying SMOTE, the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of oversampling required, neighbors from the k-nearest neighbors (k-NN) are randomly chosen. Synthetic samples are generated as follows: (1) take the difference between the feature vector (sample) under consideration and its nearest neighbor, (2) multiply this difference by a random number between 0 and 1, and (3) add it to the feature vector consideration. This causes the selection of a random point along the line segment between two specific features. This approach effectively forces the decision region of the minority class to become more general. In short, its main idea is to form new minority class samples by interpolating between several minority class samples that lie together. Thus, the overfitting problem is avoided and causes the decision boundaries for the minority class to spread further into the majority class space.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

3 Data complexity measures

To analyze the theoretical complexity of the microarray datasets chosen for this research, we have used some of the measures proposed in [2]. As these measures have been designed for two-class problems, we have converted the original multiclass problem to many instances of two-class problems by using the "one-versus-rest" strategy.

- *F1*: Maximum Fisher's discriminant ratio. This measure, which computes the maximum discriminative power of each feature, is defined as:

$$f = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

  where μ_1, μ_2 are the means and σ_1^2 and σ_2^2 are the variances of the two classes in that feature dimension. We compute f for each feature and take the maximum as the F1 measure. Higher values indicate simpler classification problems.

- *F3*: Maximum (individual) feature efficiency. In a procedure that progressively removes unambiguous points falling outside the overlapping region in each chosen dimension, the efficiency of each feature is defined as the fraction of all remaining points separable by that feature. To represent the contribution of the most useful feature, we use maximum feature efficiency as measure F3.

- *L1*: Minimized sum of the error distance by linear programming. This measure evaluates to what extent the training data is linearly separable. It returns the sum of the differences between a linear classifier predicted value and the actual class value. Unlike Ho and Basu, we use SVM with a linear kernel. A zero value for L1 indicates that the problem is linearly separable.

- *N1*: Fraction of points on the class boundary. This measure constructs a class-blind minimum spanning tree over the entire dataset, counting the number of points incident to an edge going across the two classes. This index thus reflects the fraction of such points over all points in the dataset. High values indicates smaller separation in distributions and a more difficult classification task.

4 Experimental results

The experiments were carried out on five microarray datasets. Table 1 profiles the main characteristics of these datasets in terms of number of features, number of samples, number of classes and imbalance ratio (IR). The imbalance ratio is defined as the number of samples in the majority class divided by the number of samples in the minority class, where a high score indicates a highly unbalanced dataset.

As classifier we have chosen k-NN [8]. It was executed using the Weka tool [9], using $k = 1$. A 5-fold cross validation is performed. In order to reduce the number of features, *Correlation-based Feature Selection* (CFS) [10] is used in this work. This simple multivariate filtering algorithm ranks feature subsets according to a correlation-based heuristic evaluation function. The evaluation function is biased towards subsets

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

Table 1: Characteristics of five microarray datasets.

Dataset	# Classes	# Features	# Samples	IR	Download
CLL-SUB-111	3	11340	111	4.63	[6]
Leukemia-1	3	5327	72	4.22	[7]
BrainTumor-2	4	10367	50	2.14	[7]
BrainTumor-1	5	5920	90	15.00	[7]
LungCancer	5	12600	203	23.16	[7]

containing features that are highly correlated with the class and uncorrelated with each other. Irrelevant features with low correlation with the class are ignored. Redundant features are screened out as they would be highly correlated with one or more of the remaining features. The acceptance of a feature depends on the extent to which it predicts classes in areas of the instance space not already predicted by other features.

Table 2 shows the true positives rate for the four approaches: (1) classification before applying FS and oversampling (only classifier), (2) classification after applying FS (CFS), (3) classification after applying oversampling (SMOTE) and (4) classification after applying oversampling and FS (CFS+SMOTE).

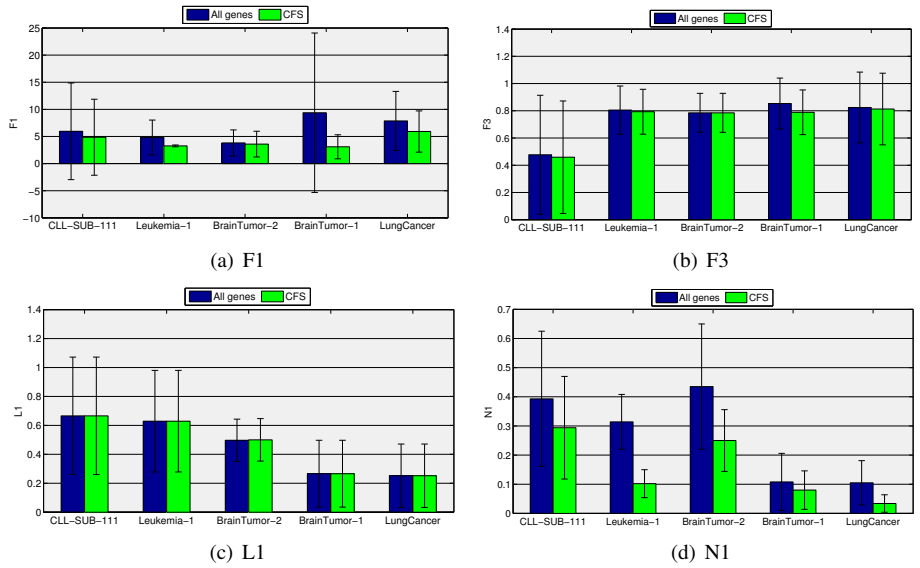
Table 2: True positives rate before and after applying feature selection and oversampling on five multiclass microarray datasets.

		Distribution (%)	Only classifier	CFS	SMOTE	CFS+SMOTE
CLL-SUB-111	Class-0	9.91	100.00	**100.00**	100.00	100.00
	Class-1	44.14	47.04	**56.87**	41.67	53.84
	Class-2	45.95	68.49	**85.15**	58.37	72.90
Leukemia-1	Class-0	52.78	100.00	**96.67**	86.28	95.32
	Class-1	12.50	66.67	**100.00**	80.00	100.00
	Class-2	34.72	64.93	**100.00**	81.67	100.00
BrainTumor-2	Class-0	28.00	63.33	96.67	68.33	**89.33**
	Class-1	14.00	20.00	50.00	89.33	**93.33**
	Class-2	28.00	60.33	86.67	63.33	**100.00**
	Class-3	30.00	54.00	64.00	88.33	**75.33**
BrainTumor-1	Class-0	66.67	96.79	94.33	94.13	**96.93**
	Class-1	11.11	83.33	100.00	100.00	**93.33**
	Class-2	11.11	83.33	66.67	93.33	**75.00**
	Class-3	4.44	100.00	100.00	100.00	**100.00**
	Class-4	6.67	40.00	40.00	40.00	**60.00**
LungCancer	Class-0	68.47	95.63	95.31	92.76	**93.28**
	Class-1	8.38	68.67	75.33	90.00	**85.00**
	Class-2	10.34	84.00	87.00	92.00	**100.00**
	Class-3	9.85	93.33	100.00	100.00	**100.00**
	Class-4	2.96	70.00	60.00	80.00	**83.33**

The first conclusion that can be drawn is that classification results are better after applying FS (both before and after applying the SMOTE method). In order to shed light on this issue, we applied the data complexity measures to the microarrays datasets after

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

redundant and/or irrelevant genes were removed. Figure 1 illustrates the behavior of these measures with FS (CFS) and without FS (All genes), where the value of the complexity measure is the average of the binary sub-problems into which each multiclass dataset was decomposed. Standard deviations are also shown. Whilst F3 and L1 maintained their values, and F1 slighltly diminishes, the N1 measure decreased considerably. Higher N1 values indicate smaller separation in distributions and a more difficult classification task, so it seems logical to think that due to the nature of the measure (its nearest neighbor base definition), that lower values after applying the CFS filter would improve classification performance.



(a) F1 (b) F3

(c) L1 (d) N1

Fig. 1: Data complexity measures with feature selection (CFS) and without feature selection (All genes) on five microarray datasets.

Trying to explain in detail the effect of SMOTE on the minority classes, Table 3 briefly summarizes the results obtained by the four data complexity measures for the five microarray datasets chosen. Indicated for each data property (overlap between classes, non-linearity and closeness to class boundary) are the corresponding classes and the related complexity measures. As can be seen in Table 2, the minority classes of CLL-SUB-111 and Leukemia-1 achieved a true positives rate of 100% (even before applying oversampling) whilst the results for the majority classes are not as good (specially for CLL-SUB-111). This is happening because of the complexity problems that the classes with more samples present, and that is the reason why SMOTE algorithm cannot help to improve the classification performance on these datasets. In contrast, the effect of the SMOTE algorithm is remarkable on the minority class of BrainTumor-2 (class 1) —which does not show any complexity problem— improving its true positives rate from 20% to 93.33% after applying SMOTE along with the CFS filter. The analysis followed for these datasets can be extended to BrainTumor-1 and LungCancer.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

Table 3: Theoretical complexity of the five microarray datasets.

	Overlap (F1, F3)	Non-linearity (L1)	Boundary Class (N1)
CLL-SUB-111	1, 2	1, 2	1, 2
Leukemia-1		0	
BrainTumor-2	2, 3	2,3	
BrainTumor-1	0, 4	0, 4	
LungCancer	0	0	

5 Conclusions

In this work we have analyzed a common problem in microarray data, the so-called class imbalance problem, using an oversampling method which is a reference in this area, the SMOTE algorithm, before and after applying feature selection.

We have observed that the imbalance ratio is not enough to predict the adequate performance of the classifier. As an alternative approach, we have computed several data complexity measures over the imbalanced datasets in order to support the application or not of an oversampling method. In view of the experimental results over five microarray datasets, we recommend to analyze the theoretical complexity of the datasets —through the data complexity measures— before applying the SMOTE algorithm. The rationale for this relies in the fact that the classifier is more affected by the complexity of the data itself than by the imbalance problem. We also suggest the use of a feature selection method before applying SMOTE. As future research, we plan to extend this study to other oversampling methods.

References

[1] Putri W Novianti, Victor L Jong, Kit CB Roes, and Marinus JC Eijkemans. Factors affecting the accuracy of a class prediction model in gene expression data. *BMC bioinformatics*, 16(1):199, 2015.

[2] T. K. Ho and M. Basu. *Data complexity in pattern recognition*. Springer, 2006.

[3] A. C. Lorena, I. G. Costa, N. Spolaôr, and M. C. de Souto. Analysis of complexity indices for classification problems: cancer gene expression data. *Neurocomputing*, 75(1):33–42, 2012.

[4] O. Okun and H. Priisalu. Dataset complexity in gene expression based cancer classification using ensembles of k-nearest neighbors. *Artificial intelligence in medicine*, 45(2):151–162, 2009.

[5] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, pages 321–357, 2002.

[6] Arizona State University. Feature selection datasets. http://featureselection.asu.edu/datasets.php [Last access: October 2015].

[7] A. Statnikov, C.Aliferis, and I. Tsardinos. Gems: Gene expression model selector. http://www.gems-system.org/ [Last access: October 2015].

[8] D. W. Aha, D.Kibler, and M. K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.

[9] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[10] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.