Two-Class with Oversampling Versus One-Class Classification for Microarray Datasets

Beatriz Pérez-Sánchez(✉), Oscar Fontenla-Romero,
and Noelia Sánchez-Maroño

Department of Computer Science, Faculty of Informatics, University of A Coruña,
Campus de Elviña s/n, 15071 A Coruña, Spain
{bperezs,ofontenla,nsanchez}@udc.es

Abstract. Microarray datasets are a challenge for classical computational techniques because of the large dimensionality of their feature space front to a reduced number of samples, besides they usually present unbalanced classes. Thanks to this unbalanced situation, in a previous research, the superiority of one-class classification for handling microarray datasets was proved. This paper presents a new study that tries to improve the behavior of the traditional techniques, specifically Support Vector Machines, by considering oversampling techniques. The experimental results achieved demonstrate that despite inclusion of these methods the performance of classical classifiers still remains below one-class approach.

1 Introduction

Microarray datasets are commonly used for cancer diagnosis distinguishing two approaches: binary and multiple classes. Firstly, the binary approach tries to differentiate patients with cancer from healthy persons and, on the other hand, the multiple classes approach tries to distinguish different variants of the same type of cancer. This paper is focused on the first approach and, since unhealthy patients are less common, these datasets are usually unbalanced. The intrinsic characteristics of microarray datasets – large dimensionality of the feature space (usually several thousand of genes) and small number of samples available (often less than a hundred) – restrict the application of classical learning machine techniques. To date, two-class classification methods are mainly used, being Support Vector Machines (SVMs) among the most notable classifiers for this task. However, in the context of microarray classification some authors proposed to use a one-class classification (OCC) for classifying microarrays due to its ability to deal with unbalanced and noisy data [1]. In OCC only instances from one of the classes are available or considered. They are known as *target* objects whereas the other are the *outlier* ones. Using OCC, models are constructed from objects belonging to only one class distribution and are robust when handling inherent data difficulties. In a previous work [2], we compared the behavior of two-class (specifically, SVM) versus OCC over microarray datasets whilst analyzing the effect of feature selection (FS). This experimental study proved the superiority

A.E.P. Villa et al. (Eds.): ICANN 2016, Part II, LNCS 9887, pp. 398–405, 2016.
DOI: 10.1007/978-3-319-44781-0_47

of the one-class approach achieving both a fine performance and a good trade-off between evaluation measures. However, a criticism to this work is that the success of SVM was limited because of the imbalanced problem that could be partially solved by sampling techniques [3]. Therefore, in this paper we present the results of a study where some of these sampling techniques are applied to improve the SVM behavior for classifying the microarray datasets denoting that, even so, OCC is superior.

This paper is structured as follows. In Sect. 2 a brief introduction about sampling techniques is given and the oversampling methods used in this experimental study are introduced. In Sect. 3 the conditions for experimental study are established. In Sect. 4 we compare the behavior of one-class classifiers and two-class methods with sampling techniques for classifying different benchmark microarray datasets, also the results are discussed. Finally, Sect. 5 is devoted to conclusions.

2 Sampling Techniques

From literature, we can find different methods to face imbalanced datasets. Among them, the most commonly employed ones are: oversampling minority class, undersampling majority class, ensemble methods, cost-sensitive learning or asymmetric classification [4]. Undersampling and oversampling are the simplest approaches. The former consists on randomly select a portion of instances from majority class whereas the latter randomly duplicates samples belonging to the minority class. Taking into account that microarray datasets enclose a reduced number of samples, undersampling does not seem a viable alternative as, it may lead to a loss of useful information. Thus, for this preliminary experimental study we focus on oversampling techniques to overcome the limitations associated to unbalanced sets. Specifically we have selected three widely applied algorithms to deal with imbalance distributions:

1. *Resampling* consists on random duplication of instances belonging to the minority class [5].
2. *Synthetic Minority Oversampling Technique* (SMOTE) algorithm generates synthetic or artificial samples by means of the nearest neighbor rule, interpolating new instances instead of duplicating them as in the case of the resampling method [6]. SMOTE does not consider the distribution of minority classes and latent noises in dataset when it generates synthetic examples. To overcome this limitation, Modified SMOTE (MSMOTE) algorithm [7] categorizes the instances belonging to the minority class into three groups according to the label of their nearest neighbors: noise (all of them belong to other classes), safe (when all neighbors belong to the minority class) otherwise, it is considered as border. Then MSMOTE chooses one of the k-nearest neighbor for safe samples and the nearest neighbor for border ones whereas in the case of noise samples the algorithm does nothing.

3. *Critical SMOTE* (CSMOTE) algorithm [4] is an improved version of the MSMOTE method that follows the idea of generating artificial samples employing only a subset of the minority class. In a first phase this algorithm extracts from the class two subsets of patterns: edge and border samples. This categorization is based on the method proposed in [8]. Edge samples define the boundary of the class and they are enough to represent the original dataset when all classes in the dataset are separated. Border samples are carefully picked in the overlapping region between adjacent classes so as to obtain the best decision surface possible. After this categorization, new patters are generated following MSMOTE. For each border sample CSMOTE randomly chooses one of the nearest neighbors whilst for each edge samples the nearest neighbor is picked.

3 Experimental Setup

The aim is to check the suitability of oversampling techniques to improve two-class classification on microarray datasets. These results are compared to those reached by one-class approach. Two of the most up-to-date classifiers are selected: SVMs for two-class classification [9] and Support Vector Data Description (SVDD) [10] as one-class classifier. It is worth mentioning that the OCC is addressed by using both minority and majority class as target concept and oversampling is not applied in any case because it is unnecessary. Next, we establish certain considerations which have been taken into account in the experimental study.

- In order to obtain statistically significant results, 30 simulations were run with the cross-validation technique to tune the parameters of each method, specifically the width parameter in the radial basis function kernel for SVDD and the kernel function (linear, radial basis and polynomial) for SVM.
- For the implementation of classifiers two different toolboxs for Matlab was used. The data description toolbox, DDtools library [11], for SVDD and the Statistics and Machine Learning toolbox for SVM.
- Similarly to our previous study [2], we have applied feature selection methods as a preprocessing step with the aim of discarding irrelevant features/genes while retaining the relevant ones. All these techniques are available in the well-known Weka tool [12], except for mRMR filter, whose implementation is available for *Matlab*.
- To evaluate the goodness of the selected set of genes in terms of accuracy of the classifier it is necessary to have an independent test set with data which have not been seen by neither the feature selection method nor the classifier. The selected data sets come originally distributed into training and test sets, so the training set was employed to perform the feature selection process and posterior classification while the test set was used to evaluate the appropriateness of the selection and the posterior classification.
- For the sake of fair comparison, only the training set is oversampled when using SVM, whereas the test dataset remains the same.

- Finally, a statistical study was conducted to determine whether the results are statistically different. First at all, the normality conditions of each distribution are checked by means of Kolmogorow Smirnov test. As in any case, normal conditions are verified then the non parametric Kruskal-Wallis test was applied.

Datasets, FS methods and evaluation measures employed for experimental study are briefly introduced below.

Datasets characteristics. Breast and Prostate datasets are widely applied due to two main properties: (1) come originally separated in training and test and (2) present more imbalance in the test set. Both datasets are available for download at [13,14]. Table 1 provides for train and test sets the number of attributes (# Atts.), examples (# Ex.) and the percentage of examples for majority (% Ma) and minority (% Min) classes. The last column corresponds to imbalance ratio (IR), a value of 1 indicates balance whereas a large value denotes a high imbalance. As can be seen in Table 1 both datasets present more imbalance in the test set specially in the case of Prostate dataset. *Dataset shift* problem [15] occurs when the joint distribution of inputs and outputs is different between training and test stages, hampering the classification process that may lead to poor performance results. This problem may be caused by different situations, such in Prostate dataset where the test set was extracted from a different experiment. Accordingly, this dataset raises a challenge for machine learning methods. For this reason some classifiers, whose features are selected according to the training set, assign all samples to the majority class.

Table 1. Description of the train and test binary datasets.

		Train				Test			
Dataset	# Atts.	# Ex	% Min	% Maj	IR	# Ex	% Min	% Maj	IR
Breast	24.481	78	43,59	56,41	1,29	19	36,84	63,16	1,71
Prostate	12.600	102	49,02	50,98	1,04	34	26,47	73,53	2,78

FS methods. Seven classical FS methods widely used in this field are selected: Correlation-based FS (CFS) [16], Fast Correlation-Based Filter (FCBF) [17], INTERACT algorithm [18], Information Gain (IG) [19], ReliefF [20], minimum Redundancy Maximum Relevance (mRMR) [21] and Support Vector Machine based on Recursive Feature Elimination (SVM-RFE) [22]. All of them, with the exception of the last one, correspond to the filter methods that rely on the general characteristics of the training data to select feature independent of any predictor. The three first CFS, FCBF and INTERACT return a subset of features. Thus, from the original 24,481 attributes of Breast dataset 130, 99 and 102 are selected respectively. While in the case of Prostate, 89, 77 and 73 are chosen from the 12,600 initial features. An ordered ranking of the features is obtained by the four last (IG, ReliefF, mRMR and SVM-RFE). For simplicity we introduce the performance keeping the top 10 and top 50 features. Finally, SVM-RFE is the most famous embedded method to specifically deal with gene selection for cancer classification. This method iteratively trains a SVM classifier with the current set of features and basing on its internal parameters the least important are removing.

Evaluation measures. For a binary classification problem, accuracy indicates how well the system predicts both categories. However accuracy is inappropriate when the prior class probabilities are very different since it does not consider misclassification costs and therefore, it is sensitive to class skews and it is biased in favor of the majority class. Then, alternative measures should be considered. The true positive rate (recall or sensitivity) is the percentage of correctly classified positive instances (e.g. the rate of cancer patients who are correctly identified as having cancer). The true negative rate (specificity) is the percentage of correctly classified negative examples (e.g. the rate of healthy patients who are correctly classify as not having cancer). The ideal predictor should be 100 % specific and 100 % sensitive. Regarding OCC, it should be mentioned that sensitivity and specificity measures are always calculated considering as negative the healthy samples and as positive the cancer ones.

4 Experimental Results

In this section the results achieved in the Breast and Prostate datasets are introduced. Table 2 shows the results obtained by SVM and SVDD classifiers, specifically Accuracy (Acc), Sensitivity (Se) and Specificity (Sp) are used to assess their performance. In the case of SVDD we introduce the results reached by using both classes (majority and minority) as the target concept in training process. Regarding SVM we include the results obtained by using resampling, SMOTE and CSMOTE as oversampling techniques. Each column represents one of the three performance measures while rows indicate the FS methods, the last row provides the results when no FS method is applied. To facilitate the analysis of the results, best values (statistically speaking) of each performance measures for each dataset are marked in bold.

Firstly, we focus on SVM with oversampling methods. At first glance, it seems that the behavior of the SVM is similar independently of the oversampling technique. An ideal predictor should be 100 % sensitive and 100 % specific but Table 2 shows that SVM tends towards one of the classes. Comparing to the original results (without oversampling) introduced in [2], it can be seen that the inclusion of oversampling methods lead to particular performance improvements without an outstanding enhancement in the trade-off between Se and Sp.

Regarding OCC, SVDD overcomes the results obtained by SVM showing important differences. In order to know if such differences are significant a statistical study was conducted. As it was previously commented, for each performance measure, FS method and dataset the best values are marked in bold face.

Table 2. Results for SVM (with oversampling techniques) and SVDD classifiers on Breast and Prostate datasets.

	FS method	*Acc*					*Se*					*Sp*				
		SVM			SVDD		SVM			SVDD		SVM			SVDD	
		SVM^a	SVM^b	SVM^c	Min	Maj	SVM^a	SVM^b	SVM^c	Min	Maj	SVM^a	SVM^b	SVM^c	Min	Maj
Breast	CFS	0.52	0.58	0.57	**0.62**	**0.65**	0.32	0.25	0.24	**0.49**	**0.56**	0.64	**0.77**	**0.76**	**0.83**	**0.80**
	FCBF	0.63	0.58	0.58	**0.70**	**0.70**	0.09	0.16	0.14	**0.72**	**0.73**	**0.94**	0.83	0.83	0.67	0.65
	INT	0.58	0.57	0.59	**0.71**	**0.71**	0.13	0.16	0.17	**0.74**	**0.74**	**0.84**	**0.80**	**0.83**	0.66	0.67
	IG-10	0.54	0.54	0.53	**0.67**	**0.67**	0.39	0.31	0.32	**0.63**	**0.63**	**0.63**	**0.67**	**0.65**	**0.74**	**0.74**
	IG-50	0.52	0.54	0.56	**0.74**	**0.73**	0.35	0.26	0.26	**0.80**	**0.79**	0.62	**0.71**	**0.73**	0.64	0.62
	RelieF-10	0.48	0.49	0.51	**0.79**	**0.79**	0.52	0.47	0.49	**0.75**	**0.75**	0.46	0.51	0.52	**0.86**	**0.86**
	RelieF-50	0.49	0.51	0.55	**0.74**	**0.73**	0.54	0.37	0.39	**0.69**	**0.67**	0.46	0.59	0.64	**0.84**	**0.83**
	SVM-RFE-10	0.50	0.52	0.52	**0.89**	**0.89**	0.57	0.49	0.51	**0.90**	**0.89**	0.47	0.53	0.53	**0.87**	**0.87**
	SVM-RFE-50	0.42	0.48	0.49	**0.84**	**0.84**	0.57	0.55	0.53	**0.84**	**0.84**	0.33	0.44	0.46	**0.84**	**0.83**
	mRMR-10	0.49	0.49	0.49	**0.76**	**0.74**	0.46	0.47	0.46	**0.78**	**0.76**	0.50	0.51	0.51	**0.73**	**0.72**
	mRMR-50	0.53	0.56	0.56	**0.76**	**0.76**	0.41	0.25	0.26	**0.81**	**0.80**	0.61	**0.74**	**0.73**	**0.68**	**0.68**
	no FS	0.47	0.55	0.56	**0.63**	**0.62**	**0.57**	0.27	0.28	0.46	0.47	0.42	0.72	0.72	**0.91**	**0.92**
Prostate	CFS	0.59	0.59	0.58	**0.97**	**0.97**	0.29	0.29	0.26	**0.97**	**0.97**	0.69	0.69	0.69	**1.00**	**1.00**
	FCBF	0.62	0.67	0.63	**0.92**	**0.92**	0.16	0.26	0.18	**0.90**	**0.89**	0.78	0.82	0.79	**0.99**	**0.99**
	INT	0.65	0.66	0.66	**0.96**	**0.96**	0.13	0.14	0.14	**0.94**	**0.94**	0.84	0.85	0.85	**1.00**	**1.00**
	IG-10	0.60	0.57	0.59	**0.94**	**0.95**	0.32	0.28	0.27	**0.92**	**0.94**	0.71	0.68	0.69	**0.98**	**0.97**
	IG-50	0.64	0.65	0.65	**0.99**	**0.99**	0.19	0.20	0.16	**0.99**	**0.98**	0.80	0.81	0.83	**1.00**	**0.99**
	RelieF-10	0.61	0.61	0.59	**0.93**	**0.93**	0.25	0.25	0.23	**0.91**	**0.91**	0.74	0.74	0.73	**1.00**	**1.00**
	RelieF-50	0.68	0.68	0.69	**0.96**	**0.96**	0.13	0.12	0.14	**0.94**	**0.94**	0.88	0.88	0.89	**0.99**	**0.99**
	SVM-RFE-10	0.64	0.62	0.61	**0.87**	**0.87**	0.27	0.24	0.23	**0.86**	**0.86**	0.77	0.76	0.75	**0.91**	**0.90**
	SVM-RFE-50	0.63	0.62	0.64	**0.96**	**0.96**	0.22	0.20	0.23	**0.95**	**0.95**	0.78	0.77	0.79	**1.00**	**1.00**
	mRMR-10	0.62	0.63	0.62	**0.94**	**0.96**	0.18	0.21	0.21	**0.93**	**0.95**	0.77	0.78	0.77	**0.97**	**0.99**
	mRMR-50	0.63	0.62	0.62	**0.94**	**0.94**	0.25	0.23	0.23	**0.91**	**0.91**	0.77	0.76	0.76	**1.00**	**1.00**
	no FS	0.61	0.59	0.62	**0.93**	**0.88**	0.13	0.26	0.28	**0.91**	**0.84**	0.78	0.71	0.75	**1.00**	**1.00**

[a] SVM corresponds to SVM with resampling technique.

[b] SVM corresponds to SVM with SMOTE technique.

[c] SVM corresponds to SVM with CSMOTE technique.

Only for Breast set, SVM obtains (in some cases) a higher value in the *Sp* measure, however in all cases SVDD achieves the best value of *Acc* and *Se* and also balanced values for *Se* and *Sp*. Finally two issues should be pointed out. On one hand, FS not only may lead to better performance results, specially in the case of Breast (for instance, see the differences between SVM-RFE-10 and the last row for this dataset) but also to significantly reduce the computational and time requirements. On the other hand, as it was previously remarked SVDD allows using minority or majority class as the target class in the training process and both exhibit a good performance. Even when the provided results are not statistically distinct, SVDD can remain the best results depending on the specific application. Since the aim of this work was to compare SVM and SVDD, there is no statistically study to compare the application or not of FS methods. However, considering FS or not, and either the minority or majority class, SVDD achieves the best performance results.

5 Conclusions

Imbalanced datasets are very common in real world for example for the diagnosis of a disease as cancer, becoming an important challenge for machine learning field. In this context, the classifiers tend towards the majority class achieving poor performance results. In a previous work we compare the results obtained by one and two class classifiers, SVDD and SVM respectively, on two microarray datasets. SVDD significantly overcame the SVM achieving a fine global performance. In this paper we include oversampling techniques to avoid the effects associated with imbalanced distributions and improve the performance of the SVM classifiers. Despite our initial idea the experimental results show that such modification does not enhance significantly the behavior of the SVM that still remains below SVDD. It is possible that this fact is caused by the peculiarities of the selected datasets. For this reason, we have in mind to extend this study including more imbalanced datasets (with higher IR) and more complex oversampling techniques to ensure the supremacy shown by the OCC in this preliminary study.

Acknowledgments. This work has been supported in part by the Secretaría de Estado de Investigación of the Spanish Government (Grant TIN2015-65069-C2-1-R), and by the Xunta de Galicia (Grant GRC2014/035) with the European Union FEDER funds.

References

1. Krawczyk, B.: Combining one-class support vector machines for microarray classification. In: Federated Conference on Computer Science and Information Systems (FedCSIS 2013), pp. 83–89 (2013)
2. Pérez-Sánchez, B., Fontenla-Romero, O., Sánchez-Maroño, N.: One-class classification for microarray datasets with feature selection. In: Iliadis, L., Jayne, C. (eds.) EANN 2015. CCIS, vol. 517, pp. 325–334. Springer, Heidelberg (2015). doi:10.1007/978-3-319-23983-5_30

3. Akbani, R., Kwek, S.S., Japkowicz, N.: Applying support vector machines to imbalanced datasets. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) ECML 2004. LNCS (LNAI), vol. 3201, pp. 39–50. Springer, Heidelberg (2004)
4. Nanni, L., Fantozzi, C., Lazzarini, N.: Coupling different methods for overcoming the class imbalance problem. Neurocomputing **158**, 48–61 (2015)
5. Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans. Syst. Man, Cybern. B, Cybern. **SMC–2**(3), 408–421 (1972)
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Oversampling Technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
7. Hu, S., Liang, Y., Ma, L., He, Y.: MSMOTE: improving classification performance when training data is imbalanced. In: 2nd International Workshop on Computer Science and Engineering (IWCSE 2009), vol. 2, pp. 13–17 (2009)
8. Li, Y., Maguire, L.: Selecting critical patterns based on local geometrical and statistical information. IEEE Trans. Pattern Anal. Mach. Intell **33**(6), 1189–1201 (2011)
9. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
10. Tax, D.M.J., Duin, R.P.W.: Support vector data description. Mach. Learn. **54**, 45–66 (2004)
11. Tax, D.M.J.: DDtools, the data description toolbox for matlab, Delft University of Technology (2005)
12. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
13. Kent Ridge Bio-Medical Dataset. http://datam.i2r.a-star.edu.sg/datasets/krbd. Accessed Feb 2016
14. Microarray Cancers, Plymouth University. http://www.tech.plym.ac.uk/spmc/links/bioinformatics/microarray/microarray_cancers.html. Accessed Feb 2016
15. Moreno-Torres, J.G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N.V., Herrera, F.: A unifying view on dataset shift in classification. Pattern Recogn. **45**(1), 521–530 (2012)
16. Hall, M.: Correlation-Based Feature Selection for Machine Learning. Ph.D. Thesis (1999)
17. Yu, L., Liu, H.: Feature selection for high-dimensional data: a fast correlation-based filter solution. In: 20th International Conference on Machine Learning (ICML 2003), pp. 856–863 (2003)
18. Zhao, Z., Liu, H.: Searching for interacting features. In: 20th International Joint Conference on Artifical Intelligence (IJCAI 2007), pp. 1156–1161 (2007)
19. Hall, M., Smith, L.: Practical feature subset selection for machine learning. In: 21st Australasian Computer Science Conference (ACSC 1998), pp. 181–191 (1998)
20. Kononenko, I.: Estimating attributes: analysis and extensions of RELIEF. In: Bergadano, F., De Raedt, L. (eds.) ECML-94. LNCS, vol. 784, pp. 171–182. Springer, Heidelberg (1994)
21. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell **27**, 1226–1238 (2005)
22. Guyon, I., Weston, J., Barnhill, S., Vapnik, V., Cristianini, N.: Gene selection for cancer classification using support vector machines. Mach. Learn. **46**(1–3), 389–422 (2002)