Feature Dimensionality vs. Distribution of Sample Types: A Preliminary Study on Gene-Expression Microarrays

J. Salvador Sánchez
Institute of New Imaging Technologies
Department of Computer Languages and Systems
Universitat Jaume I
Castelló de la Plana, Spain
sanchez@uji.es

Vicente García
División Multidisciplinaria de Ciudad Universitaria
Universidad Autónoma de Ciudad Juárez
Ciudad Juárez, Chihuahua, Mexico
vicente.jimenez@uacj.mx

***Abstract*—In gene-expression microarray data sets each sample is defined by hundreds or thousands of measurements. High-dimensionality data spaces have been reported as a significant obstacle to apply machine learning algorithms, owing to the associated phenomenon called 'curse of dimensionality'. The analysis and interpretation of these data sets have been defined as a very challenging problem. The hypothesis proposed in this paper is that there may exist some correlation between dimensionality and the types of samples (safe, borderline, rare and outlier). To examine our hypothesis, we have carried out a series of experiments over four gene-expression microarray databases because these data correspond to a typical example of the so-called 'curse of dimensionality' phenomenon. The results show that there indeed exist meaningful relationships between dimensionality and the proportion of each type of samples, demonstrating that the amount of safe samples increases and the total number of borderline samples decreases as dimensionality of the feature space decreases.**

***Index Terms*—Gene-expression microarray, feature dimensionality, sample types, feature ranking, classification**

I. Introduction

The 'curse of dimensionality' phenomenon (also known as the Hughes phenomenon) constitutes a challenging problem in many real-life applications. It refers to a situation in which the number of samples needed to estimate an arbitrary function with a given level of accuracy grows exponentially with respect to the number of input variables (dimensionality) of the function [1]. An illustrative example of this problem corresponds to gene-expression microarray data [2], [3] where the number of genes (G) massively exceeds the sample size (n): there are typically over tens of thousands of gene-expression levels and often less than 100 samples in the data set. This is a problem in itself because it may increase the complexity of classification/prediction, degrade the generalization ability of classifiers and hinder the understanding of the underlying relationships between the genes and the samples [4], [5]. Besides, overfitting is also a major issue in a high-dimensional, low-sample scenario [6].

Feature selection is the standard way to tackle this problem by choosing a subset of informative variables from the whole set of features for further analysis. In the specific context of microarray data, there exists an apparent need for dimensionality reduction not only because of the huge number of input variables, but also because many of them can be highly correlated with other variables. Throughout the last decades, many different feature (gene) selection algorithms have been proposed using filter, wrapper, embedded, ensembles and hybrid methods [7]–[11].

A particularly popular strategy for feature selection over microarray data refers to the use of gene ranking algorithms, which are filters that comprise some univariate scoring metric to quantify how much more statistically significant each gene is than the others [12]. These methods rank genes in decreasing order of the estimated scores under the assumption that the top-ranked genes correspond to the most informative (or differentially expressed) ones across different classes without redundancy.

The central question the present study intends to answer is how dimensionality of the feature space and some intrinsic data characteristics are related to each other. More specifically, this paper examines whether or not dimensionality reduction may alter the distribution of the different types of samples defined by several authors [13], [14]. To gain some insight into this question, we analyze the tendency of the amount of each type of samples when varying the dimensionality of the feature space. For the experiments, we consider four public data sets of gene-expression microarrays.

Over the past years, the potential links between feature dimensionality and several data complexities in microarrays have been a matter of concern for researchers. For instance, Baumgartner and Somorjai [15] used five real-life biomedical databases of increasing difficulty to show how the data complexity of a given classification problem can be related to the performance of regularized linear classifiers. Okun and Priisalu [16] explored the connections between data complexity and classification performance defined by low-variance and low-biased bolstered resubstitution error made by k-nearest neighbor classifiers. Souto et al. [17] computed different

measures characterizing the complexity of gene expression data sets for cancer diagnosis, and then investigated how those measures were related to the classification performances of support vector machines. Bolón-Canedo et al. [18] presented a review of a set of feature selection methods applied to DNA microarray data and analyzed the impact of class imbalance, class overlapping or data set shift on the classification results. Similarly, Sánchez and García [19] demonstrated that there exist meaningful relationships between dimensionality and class separability in gene-expression microarray data sets. Lorena et al. [20] studied the complexity of several microarray data sets with and without dimensionality reduction using a support vector machine. Seijo-Pardo et al. [21] proposed the use of three data complexity measures to automatically set a threshold value, which is then employed to obtain a subset of genes from the ordered ranking given by a ranker algorithm. Morán-Fernández et al. [22] demonstrated that there is some correlation between microarray data complexity and the classification error rates of a set of classifiers. Sun et al [23] proposed an ECOC algorithm to address the small sample size and class imbalance problems in multi-class microarray data sets by exploring data distributions based on data complexity theory.

Henceforth, the rest of the paper is organized as follows. Section II presents the types of samples according to the taxonomy proposed by Napierala and Stefanowski [14]. Section III provides the experimental set-up and the description of the databases used in our experiments. Next, the results are reported and discussed in Section IV. Finally, Section V summarizes the main conclusions and points out some directions for future research.

II. Types of Samples

Following the categorization proposed by several authors [13], [14], [24], two main types of samples should be distinguished: *safe* and *unsafe*. Safe samples refer to those located in homogeneous regions with data of a single class and are sufficiently separated from examples of other classes, whereas the rest of samples have to be considered as unsafe. The safe samples will be classified correctly by most models, but the classification of unsafe samples will usually be a very tricky task with a high error rate.

The general feature of unsafe samples is that they are placed close to examples from some other classes. However, this type of data can be further divided into three subgroups depending on their particular characteristics [14], [25]: *borderline*, *rare* and *outlier*. Borderline samples are located near the decision boundaries between classes. Rare samples are small groups of examples located far from the core of their class, creating small data chunks or subclusters. Finally, the outliers are single samples being surrounded by examples that belong to some other class.

A simple method to identify each type of samples is based on analyzing the local neighborhood of the examples. This can be performed either by searching for the k neighbors of a sample or by using some kernel function. Thus, one can guess that a safe sample x will be characterized by having a neighborhood with a majority of examples that belong to its same class; rare examples and outliers will be mainly surrounded by examples from different classes, whereas borderline samples will be surrounded by examples both from their same class and also from different classes.

Many authors often choose $k = 5$ because smaller values may poorly distinguish the nature of samples, while higher values would violate the assumption of the local neighborhood [14], [24]–[26]. Following this procedure, we can define the following cases:

- A sample x will be classified in the safe category if at least 4 out of the 5 nearest neighbors belong to the class of x.
- A sample x will be classified in the borderline category if 2–3 out of its 5 nearest neighbors belong to the class of x.
- A sample x will be classified in the rare category if only one nearest neighbor belongs to the class of x, and this has no more than one neighbor from its same class.
- A sample x will be classified in the outlier category if all its nearest neighbors are from the opposite class.

III. Databases and Experimental Protocol

We conducted a pool of experiments on a collection of publicly available gene-expression microarray data sets, which were taken from the Kent Ridge Biomedical Data Set Repository (http://datam.i2r.a-star.edu.sg/datasets/krbd). Table I reports the main characteristics of these databases, including the number of genes (features), the number of samples, and the size of each class (here designated as positive and negative).

TABLE I
Characteristics of the gene-expression microarray data sets

	#Genes	#Samples	#Positive	#Negative
Breast	24481	97	46	51
CNS	7129	60	21	39
Colon	2000	62	22	40
Prostate	12600	136	59	77

For the present study, we varied the percentage of genes from 5% to 100% with a step size of 5% by using the ReliefF algorithm, thus yielding 20 different subsets (each one with a percentage of the top-ranked features) for each database. The experiments have been circumscribed to the ReliefF algorithm because this paper aims to analyze how dimensionality of the feature space might affect the proportion of the different types of samples, not to find the best feature selection/ranking method.

A. The ReliefF Algorithm

The basic idea of the ReliefF algorithm [27] lies on adjusting the weights of a vector $W = [w(1), w(2), \ldots, w(G)]$ with the objective of giving more relevance to features that better discriminate the samples from neighbors of some different class.

It randomly picks out a sample x and searches for k nearest neighbors of the same class (hits, h_i) and k nearest neighbors from each of the different classes (misses, m_i). If x and h_i have different values on feature f, then the weight $w(f)$ is decreased because it is interpreted as a bad property of this feature. In contrast, if x and m_i have different values on the feature f, then $w(f)$ is increased. This process is repeated t times, and the values of the weight vector W are updated as follows:

$$w(f) = w(f) - \frac{\sum_{i=1}^{k} dist(f,x,h_i)}{t \cdot k} + \sum_{c \neq class(x)} \frac{P(c)}{1 - P(class(x))} \cdot \frac{\sum_{i=1}^{k} dist(f,x,m_i)}{t \cdot k} \quad (1)$$

where $P(c)$ is the prior probability of class c, $P(class(x))$ denotes the probability for the class of x, and $dist(f,x,m_i)$ represents the absolute distance between samples x and m_i in the feature f.

The algorithm assigns negative values to features that are completely irrelevant and the highest scores for the most informative features. In general, one will then select the g top-ranked features in order to build the classifier with a presumably much smaller subset of features ($g \ll G$). In addition, unlike other well-known ranking methods such as those based on information theory (e.g., mutual information or information gain), the ReliefF algorithm takes care of the dependencies between genes [28].

IV. Results and Discussion

This section is devoted to explore how the number of genes may have an effect on the amount of samples that belong to each type. As far as we know, there has been no systematic analysis on this problem; in fact, previous studies have focused on identifying the types of samples from the minority class in class imbalanced data sets and analyzing how the resampling techniques may alter the distribution/proportion of safe, borderline, rare and outlier samples [14], [24]–[26], [29], [30]

Bearing our purpose in mind, the experiments were as follows. First, we calculated the percentages of positive and negative samples from each type when varying the percentage of genes. Afterwards, we also run six classifiers of different nature over each subset of features: the 1-nearest neighbor (1-NN) rule with the Euclidean distance, a pruned C4.5 decision tree, a support vector machine (SVM) with a linear kernel using the sequential minimal optimization algorithm and a soft-margin $C = 1.0$, a normalized Gaussian radial basis function (RBF) neural network with the K-means clustering algorithm to provide the basis functions, the naive Bayes classifier (NBayes), and a multi-layer perceptron (MLP) with one hidden layer, a learning rate of 0.3 and 500 training epochs.

Fig. 1 shows the percentages of each positive sample type when varying the dimensionality of the feature space for each database. As can be seen, the percentage of safe samples in the positive class increases and the percentage of borderline positive samples decreases as dimensionality decreases. Although the percentages of rare and outlier samples are generally low, it was observed a very similar behavior to that of the borderline samples. This result could allow to gain some insight into the reasons why classification in lower dimensions is usually easier than in higher dimensions.

Analogously, Fig. 2 displays the percentages of the negative sample types when varying the dimensionality of the feature space for each database. In general, lines in these plots closely match the trend patterns recognized in the plots of Fig. 1, that is, the percentage of safe samples increases and the percentages of the different types of unsafe samples decrease as dimensionality decreases. Notwithstanding, for the safe and borderline samples, we observed an essential difference of behavior between the positive class and the negative class: while the percentages of safe positive samples were usually lower than those of the borderline positive samples, the percentages of safe negative samples always resulted much higher than those of the borderline negative samples. This behavior agrees with the expected one because the negative class corresponds to the majority class and therefore, the probability for a negative sample to be identified as safe is higher than the probability of being classified in some group of the unsafe samples.

Regarding the rare and outlier samples that belong to the negative class, we found that there was no substantial relationship between dimensionality of the feature space and the number of samples in both these types. Nevertheless, this fact should not become especially critical for a given classification problem because the amount of samples that belong to the rare and outlier types is minimal as compared to the total number of safe and borderline samples.

Plots in Fig. 3 correspond to the accuracy achieved by each classification model when applied to each of the 20 subsets. It is possible to observe that the accuracy of all classifiers tends to decrease as the amount of genes increases. A visual comparison between this figure and those of the sample types allows to demonstrate that there exists some significant link (positive correlation) between the dimensionality of the feature space and the distribution of sample types since the highest accuracies were achieved for the subsets with the largest number of safe samples and the smallest number of unsafe samples.

V. Concluding Remarks

As one of the earliest works on investigating the potential connections between feature dimensionality and sample types, this paper has to be viewed as a preliminary study of the effects of dimensionality reduction on the distribution of the different types of samples in a data set.

From the experiments carried out, we have observed that the proportions of safe, borderline, rare and outlier samples vary as the dimensionality of the feature space changes. More specifically, reduction in dimensionality generally leads to a

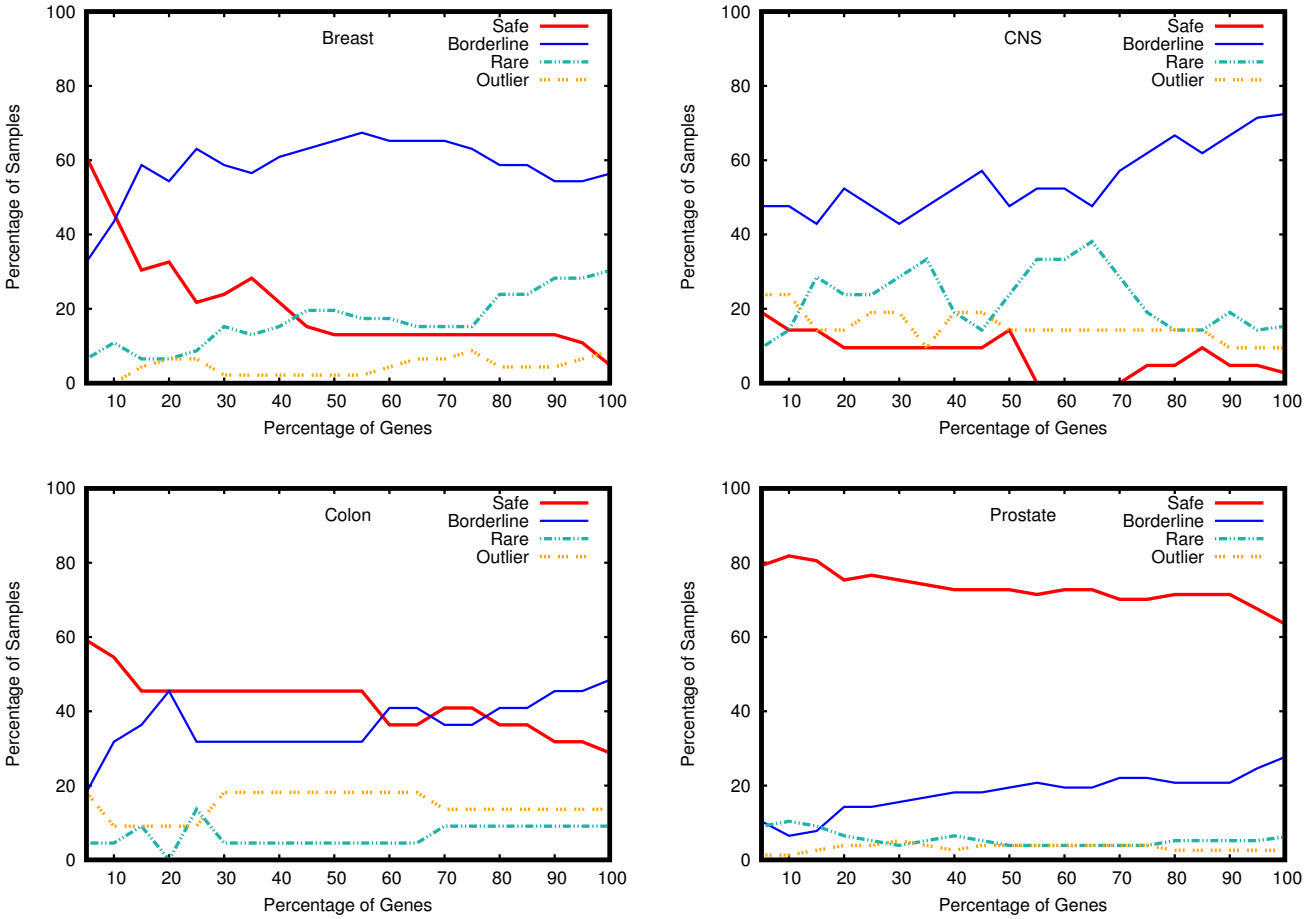

Fig. 1. Plots of the percentage of each type of positive samples when varying the percentage of genes

significant decrease in the amount of borderline samples and an increase in the number of safe samples. As showed in the experiments, this has a direct impact on the performance of classifiers because the classification of safe samples results much easier than the classification of any type of unsafe samples.

Through the characterization of databases by the distribution of their sample types, our hypothesis for further research is that it would be possible to define a meta-learning framework to choose the feature subset with the highest classification performance. Another direction for extending the present paper consists in the combined use of sample types and data complexity measures for the implementation of accurate preprocessing methods.

Acknowledgment

This research work has partially been supported by the Mexican PRODEP under Grant No. DSA/103.5/15/7004, the Generalitat Valenciana under Grant No. PROMETEOII/2014/062, and the Universitat Jaume I under Grant No. P1-1B2015-74.

References

[1] L. Chen, "Curse of dimensionality," in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds. Boston, MA: Springer, 2009, pp. 545–546.

[2] R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang, "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data," *Nat. Rev. Cancer*, vol. 8, no. 1, pp. 37–49, 2008.

[3] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Recent advances and emerging challenges of feature selection in the context of big data," *Knowl.-Based Syst.*, vol. 86, pp. 33–45, 2015.

[4] E. R. Dougherty, "Small sample issues for microarray-based classification," *Compar. Func. Genom.*, vol. 2, no. 1, pp. 28–34, 2001.

[5] L. Wang, F. Chu, and W. Xie, "Accurate cancer classification using expressions of very few genes," *IEEE-ACM T. Comput. Biol. Bioinform.*, vol. 4, no. 1, pp. 40–53, 2007.

[6] R. Somorjai, B. Dolenko, and R. Baumgartner, "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions," *Bioinformatics*, vol. 19, no. 12, pp. 1484–1491, 2003.

[7] I. Inza, P. Larrañaga, R. Blanco, and A. J. Cerrolaza, "Filter versus wrapper gene selection approaches in DNA microarray domains," *Artif. Intell. Med.*, vol. 31, no. 2, pp. 91–103, 2004.

[8] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[9] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe, "A survey on filter techniques for feature selection in gene expression microarray analysis,"

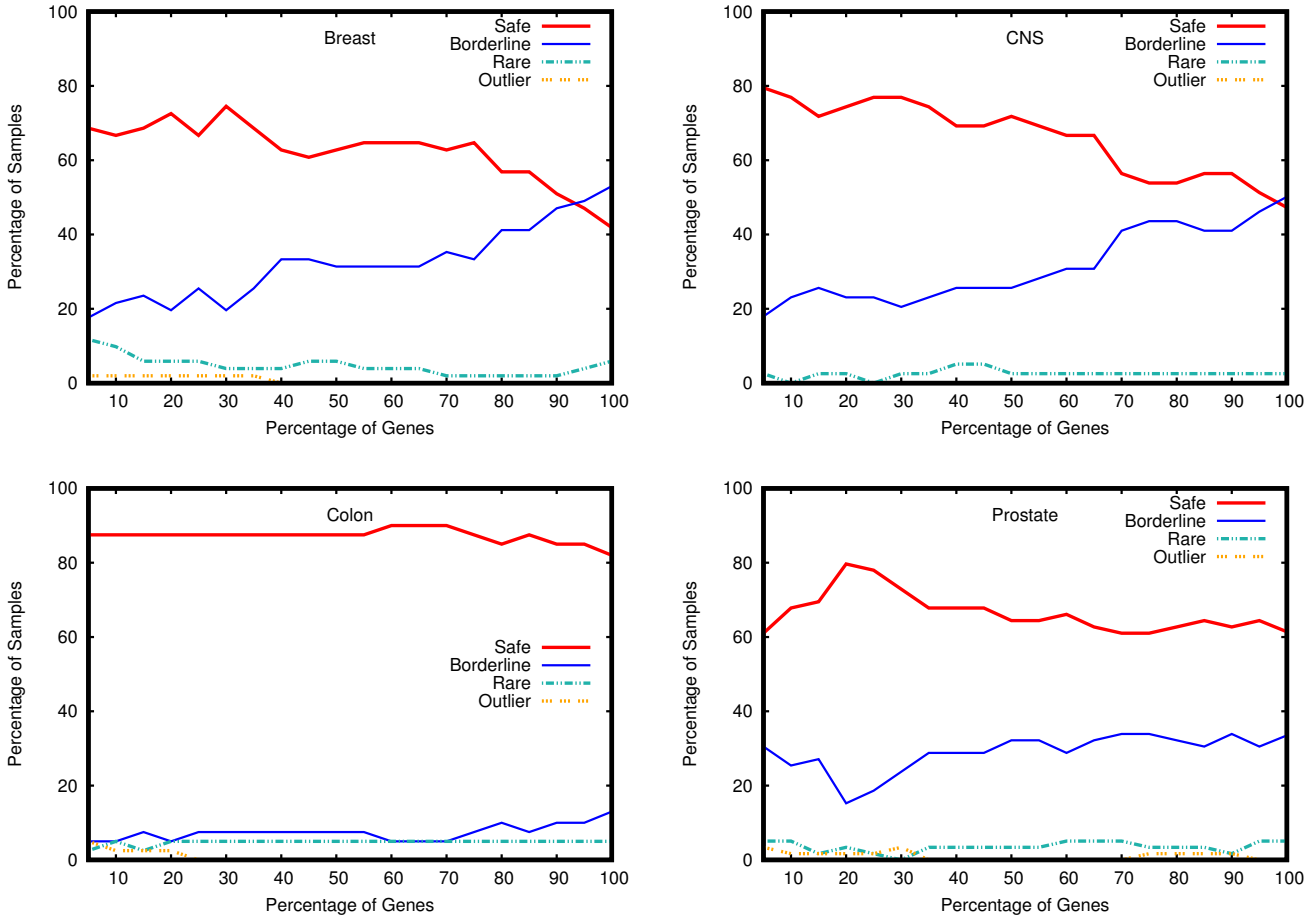

Fig. 2. Plots of the percentage of each type of negative samples when varying the percentage of genes

IEEE-ACM T. Comput. Biol. Bioinform., vol. 9, no. 4, pp. 1106–1119, 2012.

[10] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Adv. Bioinformatics*, vol. 2015, no. ID 198363, pp. 1–13, 2015.

[11] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection," *IEEE-ACM T. Comput. Biol. Bioinform.*, vol. 13, no. 5, pp. 971–989, 2016.

[12] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[13] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proc. 14th International Conference on Machine Learning*, Nashville, TN, 1997, pp. 179–186.

[14] K. Napierala and J. Stefanowski, "Types of minority class examples and their influence on learning classifiers from imbalanced data," *J. Intell. Inf. Syst.*, vol. 46, no. 3, pp. 563–597, 2016.

[15] R. Baumgartner and R. Somorjai, "Data complexity assessment in undersampled classification of high-dimensional biomedical data," *Pattern Recogn. Lett.*, vol. 27, no. 12, pp. 1383–1389, 2006.

[16] O. Okun and H. Priisalu, "Dataset complexity in gene expression based cancer classification using ensembles of k-nearest neighbors," *Artif. Intell. Med.*, vol. 45, no. 2, pp. 151–162, 2009.

[17] M. C. P. de Souto, A. C. Lorena, N. Spolaôr, and I. G. Costa, "Complexity measures of supervised classifications tasks: A case study for cancer gene expression data," in *Proc. International Joint Conference on Neural Networks*, Barcelona, Spain, 2010, pp. 1–7.

[18] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," *Inform. Sciences*, vol. 282, pp. 111–135, 2014.

[19] J. S. Sánchez and V. García, "Addressing the links between dimensionality and data characteristics in gene-expression microarrays," in *Proc. International Conference on Learning and Optimization Algorithms: Theory and Applications*, Rabat, Morocco, 2018, pp. 1–6.

[20] A. C. Lorena, I. G. Costa, N. Spolaor, and M. C. de Souto, "Analysis of complexity indices for classification problems: Cancer gene expression data," *Neurocomputing*, vol. 75, no. 1, pp. 33–42, 2012.

[21] B. Seijo-Pardo, V. Bolón-Canedo, and A. Alonso-Betanzos, "Using data complexity measures for thresholding in feature selection rankers," in *Advances in Artificial Intelligence*. Lecture Notes in Computer Science, Springer, 2016, vol. 9868, pp. 121–131.

[22] L. Morán-Fernández, V. Bolón-Canedo, and A. Alonso-Betanzos, "Can classification performance be predicted by complexity measures? A study using microarray data," *Knowl. Inf. Syst.*, vol. 51, no. 3, pp. 1067–1090, 2017.

[23] M. Sun, K. Liu, and Q. Hong, "An ECOC approach for microarray data classification based on minimizing feature related complexities," in *Proc. 10th International Symposium on Computational Intelligence and Design*, Hangzhou, China, 2017, pp. 300–303.

[24] J. A. Sáez, B. Krawczyk, and M. Woźniak, "Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets," *Pattern Recogn.*, vol. 57, pp. 164–178, 2016.

[25] B. Krawczyk, M. Woniak, and F. Herrera, "Weighted one-class classification for different types of minority class examples in imbalanced data," in *Proc. IEEE Symposium on Computational Intelligence and Data Mining*, Piscataway, NJ, 2014, pp. 337–344.

[26] P. Skryjomski and B. Krawczyk, "Influence of minority class instance types on SMOTE imbalanced data oversampling," in *Proc. 1st Interna-*

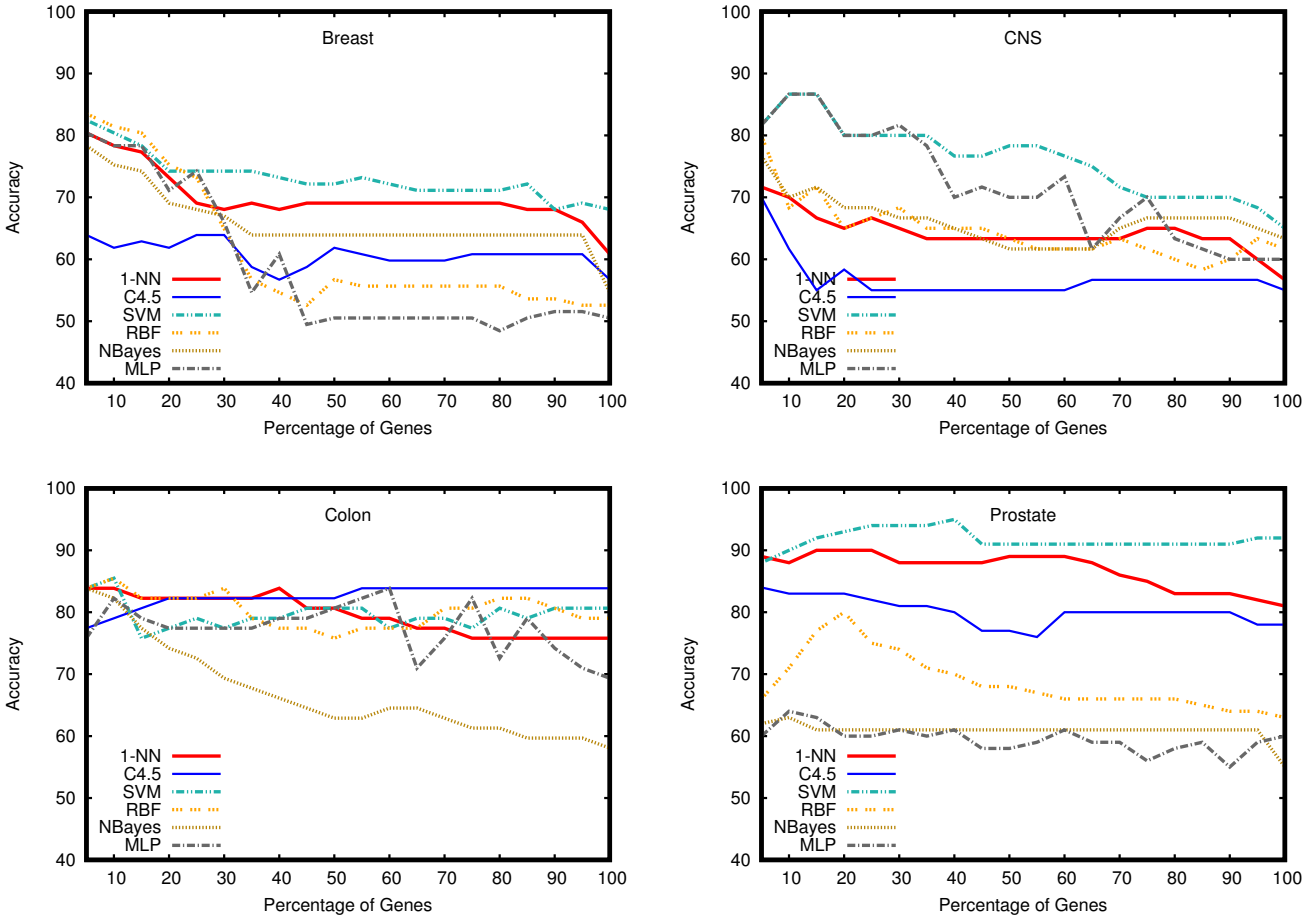

Fig. 3. Classification accuracies when varying the percentage of genes

tional Workshop on Learning with Imbalanced Domains: Theory and Applications, Skopje, Macedonia, 2017, vol. 74, pp. 7–21.

[27] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, no. 1–2, pp. 23–69, 2003.

[28] Y. Peng, W. Li, and Y. Liu, "A hybrid approach for biomarker discovery from microarray gene expression data for cancer classification," *Cancer Inform.*, vol. 2, pp. 301–311, 2006.

[29] S. Wang, L. L. Minku, and X. Yao, "Resampling-based ensemble methods for online class imbalance learning," *IEEE T. Knowl. Data En.*, vol. 27, no. 5, pp. 1356–1368, 2015.

[30] M. Lango and J. Stefanowski, "The usefulness of roughly balanced bagging for complex and high-dimensional imbalanced data," in *Proc. 4th International Workshop on New Frontiers in Mining Complex Patterns*. Porto, Portugal: Springer International Publishing, 2016, pp. 93–107.