Addressing the Links Between Dimensionality and Data Characteristics in Gene-Expression Microarrays

J. Salvador Sánchez
Institute of New Imaging Technologies, Dept. 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'. Therefore the analysis (and interpretation) of these data sets has become a challenging problem. The hypothesis set out in this paper is that the curse of dimensionality is directly linked to other intrinsic data characteristics, such as class overlapping and class separability. To examine our hypothesis, here 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 exist meaningful relationships between dimensionality and some specific complexities that are inherent to data (especially, class separability and geometry of manifolds). Moreover, it is also discussed the behavior of three classifiers as a function of dimensionality and data complexities.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → *Life and medical sciences*; Bioinformatics;

KEYWORDS

Dimensionality, gene-expression microarray, data complexity measure, feature ranking

ACM Reference Format:
J. Salvador Sánchez and Vicente García. 2018. Addressing the Links Between Dimensionality and Data Characteristics in Gene-Expression Microarrays. In *LOPAL '18: International Conference on Learning and Optimization Algorithms: Theory and Applications, May 2–5, 2018, Rabat, Morocco.* ACM, New York, NY, USA, Article 1, 6 pages. https://doi.org/10.1145/3230905.3230909

1 INTRODUCTION

A major problem in many real-life applications refers to the 'curse of dimensionality' phenomenon, which indicates that the number of samples needed to estimate an arbitrary function with a given level of accuracy grows exponentially regarding the number of input variables (dimensionality) of the function [7]. A challenging example of this problem corresponds to gene-expression microarray data [5] where the number of genes (G) heavily 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, degrade the generalization ability of classifiers and hinder the understanding of the underlying relationships between the genes and the samples [9, 25]. Besides, overfitting is also a major issue in a high-dimensional, low-sample scenario [24].

Feature selection is the standard way to tackle this problem by choosing a small portion of informative variables for further analysis. In the specific context of microarray data, there exists a glaring need for dimensionality reduction not only because of the vast number of input variables, but also because many of them can be highly correlated with other variables. Many different algorithms have been proposed over the last years for feature (gene) selection using filter, wrapper, embedded and hybrid methods [2, 11, 15, 22].

A particularly popular strategy for feature selection over microarray data is 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 [10]. 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.

Bolón-Canedo et al. [6] 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. Several authors have investigated the possible connections between classifier performance and complexity of microarrays [3, 4, 19]. Lorena et al. [16] studied the complexity of several microarray data sets with and without dimensionality reduction using a support vector machine. Morán-Fernández et al. [17] demonstrated that there is a correlation between microarray data complexity and the classification error rates.

The critical question the present study intends to answer is how dimensionality and some intrinsic data characteristics are related. More specifically, this paper examines whether or not some data difficulty factors can be alleviated by dimensionality reduction and to what extent this affects the classification performance. Additionally, we propose a new index that allows to characterize the relationship between data set size requirements and dimensionality. To gain some insight into these questions, we analyze the tendency

LOPAL '18, May 2–5, 2018, Rabat, Morocco

ACM ISBN 978-1-4503-5304-5/18/05…$15.00
https://doi.org/10.1145/3230905.3230909

of several data complexity measures when varying the dimensionality of the feature space. For the experiments, we consider four public data sets of gene-expression microarrays.

2 QUANTIFICATION OF INTRINSIC DATA CHARACTERISTICS

The prediction performance of classifiers strongly depends on the particular characteristics of each data set. To analyze the theoretical complexity of each problem, several measures have been proposed in the literature [12, 13, 23] with the ultimate purpose of explaining the behavior of learning algorithms. The complexity measures that will be used in the experiments are often grouped into three categories according to the property of the data they focus on: (i) measures of overlap in feature values from different classes, (ii) measures of class separability, and (iii) measures of geometry, topology, and density of manifolds.

2.1 Measures of overlap in feature values from different classes

These measures focus on the effectiveness of an individual feature in separating the samples from different classes. To this end, they examine the range and spread of values in the data set within each class regarding to each feature, and check for overlaps among different classes.

- F1 (*maximum Fisher's discriminant ratio*): It computes how separated are two classes according to each individual feature.

$$\text{F1} = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{1}$$

  where μ_1, μ_2 are the means and σ_1^2, and σ_1^2 are the variances of the two classes in the feature.
- F2 (*volume of overlap region*): It computes, for each feature g_i, the length of the overlap range normalized by the length of the total range in which all values of both classes are distributed. Then, this measure can be defined as

$$\text{F2} = \prod_i \frac{\min(\max(g_i, c_1), \max(g_i, c_2)) - \max(\min(g_i, c_1), \min(g_i, c_2))}{(\max(\max(g_i, c_1), \max(g_i, c_2)) - \min(\min(g_i, c_1), \min(g_i, c_2))} \tag{2}$$

  where $i = 1, \ldots, G$ for a G-dimensional problem.

2.2 Measures of class separability

Linear separability refers to the maximum probability of correct classification when discriminating the pattern distribution with hyperplanes. In two-class problems, these measures evaluate to what extent the classes are separable by examining the existence and shape of the class boundaries.

- L1 (*minimized sum of error distance by linear programming*): This corresponds to the value of the objective function that tries to minimize a linear classifier obtained by a linear programming formulation. The method minimizes the sum of distances of error points to the separating hyperplane.
- L2 (*error rate of linear classifier by linear programming*): This measure is the error rate of the linear classifier defined for L1 on the training set.
- N2 (*ratio of average intra/inter-class nearest neighbor distance*): It compares the intra-class dispersion with the inter-class separability. Given a sample x_i, let $dist_{intra}(x_i)$ and $dist_{inter}(x_i)$ be the distance to its nearest neighbor from the same class and the distance to its nearest neighbor from the other class, respectively. Then, this measure can be computed as follows:

$$\text{N2} = \frac{\sum_{i=1}^{n} dist_{intra}(x_i)}{\sum_{i=1}^{n} dist_{inter}(x_i)} \tag{3}$$

- N3 (*error rate of the nearest neighbor classifier*): This is the error rate of the nearest neighbor classifier estimated by the leaving-one-out method. It indicates how close the samples of different classes are.

2.3 Measures of geometry, topology and density of manifolds

These measures are intended to describe the geometry or the shapes of the manifolds spanned by each class.

- T1 (ϵ-*neighborhoods*): This measure counts the number of balls needed to cover each class, being each ball centered at a sample and grown to the maximal size (in units of ϵ) before it reached a sample from the other class. Redundant balls lying completely in the interior of other balls are removed. This count is then normalized by the total number of samples.
- T2 (*average number of points per dimension*): It describes the density of spatial distributions of samples by computing the number of samples in the data set over the number of feature dimensions.
- L3 (*non-linearity of linear classifier by linear programming*): Given a data set, this method first generates a test by linear interpolation between randomly drawn pairs of points belonging to the same class. Then, the error rate of a linear classifier on such a test set is measured.
- N4 (*non-linearity of the nearest neighbor classifier*): Unlike the L3, here the error is calculated for the nearest neighbor classifier.

2.4 Relationship between data set size requirements and dimensionality

Although there is no strict guideline about what a sufficient data size is, the common wisdom is that the minimum number of samples needed to achieve good generalization should be around $n^* = 10 \times G \times c$, where c is the number of classes in a problem [18]. Taking this into account, the difference between the theoretical minimum sample size (n^*) and the current data set size (n) as a proportion of the theoretical minimum size is here proposed as a new index to characterize the relationship between data set size requirements and dimensionality:

$$I_{req} = \frac{n^* - n}{n^*} \tag{4}$$

This index indicates that the number of samples required grows exponentially with the dimensionality. For a perfectly modeled problem where $n^* = n$, $I_{req} = 0$. If $n^* > n$, $0 < I_{req} < 1$ and therefore, the data set is downsized with regard to dimensionality. On the other hand, if $n^* < n$, $I_{req} < 0$ suggesting that the data set is oversized.

3 DATABASES AND EXPERIMENTAL SET-UP

We conducted a pool of experiments on a collection of publicly available gene-expression microarray data sets taken from the Kent Ridge Biomedical Data Set Repository (http://datam.i2r.a-star.edu.sg/datasets/krbd). Table 1 summarizes the main characteristics of these data sets, reporting the number of genes (features), the number of samples, and the size of the positive and negative classes.

Table 1: Characteristics of the gene-expression microarray data sets

	#Genes	#Samples	Positive — Negative
Breast	24481	97	Relapse (46) — (51) Non-relapse
Colon	2000	62	Tumor (22) — (40) Normal
CNS	7129	60	Failure (39) — (21) Survivor
Prostate	12600	136	Tumor (77) — (59) Normal

For the present study, we varied the percentage of genes selected by the ReliefF algorithm from 5% to 100% with a step size of 5%. Bearing in mind that this paper aims to analyze how dimensionality might affect other data characteristics, not to find the best feature selection method, the experiments have been confined to the ReliefF algorithm.

3.1 The ReliefF algorithm

The basic idea of the ReliefF algorithm [21] lies on adjusting the weights of a vector $W = [w(1), w(2), \ldots, w(G)]$ to give more relevance to features that better discriminate the samples from neighbors of a 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, updating the values of the weight vector W as follows

$$w(f) = w(f) - \frac{\sum_{i=1}^{k} dist(f, x, h_i)}{t \cdot k} + \sum_{c \neq class(x)} \frac{P(c)}{1 - P(class(x))} \cdot \frac{\sum_{i=1}^{k} dist(f, x, m_i)}{t \cdot k} \quad (5)$$

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 to build the classifier with a presumably much smaller subset of features ($g \ll G$). Moreover, unlike other ranking methods such as those based on information theory (e.g., mutual information or information gain), the ReliefF algorithm considers the dependencies between genes [20].

4 RESULTS AND DISCUSSION

This section is divided into two blocks. Firstly, we will discuss the possible relationships between dimensionality and the data complexity measures introduced in Sections 2.1–2.3. The second part will be devoted to show the behavior of three classification models when varying the number of genes and how such a behavior was also related to the other intrinsic characteristics of data.

Before discussing the results related to each block, Figure 1 illustrates the plots of the index introduced in Section 2.4. As can be observed, the size of the experimental data sets with the original dimensionality is very far from the theoretical requirements to achieve high classification performance. However, by reducing the number of genes, the index I_{req} drops down and the probability of an accurate generalization increases.

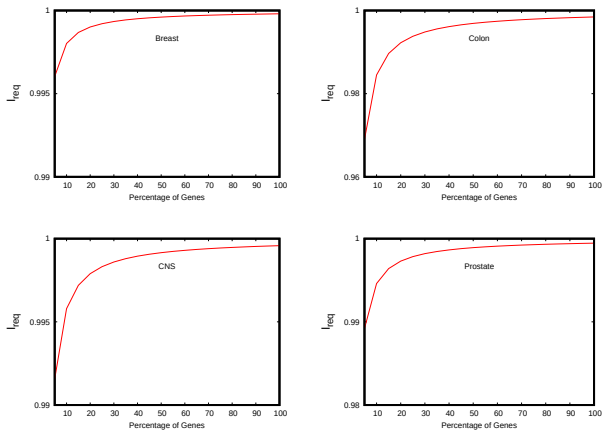

Figure 1: Plots of I_{req} when varying the percentage of genes

4.1 Dimensionality versus intrinsic data complexities

Figure 2 depicts the behavior of the data complexity measures as a function of the percentage of genes for each database. The plots of F1, F2 and T1 have not been included because the values of these measures kept constant across all dimensions; on the other hand, the values of L2 and L3 have not been plotted because they exhibit trends opposite to that of L1 in all cases.

The fact that F1 and F2 were constant across all dimensions indicates that there is no meaningful relation between the level of overlapping and dimensionality. F2 was very close to 0 for all cases, which suggests that the volume of overlap region was minimal.

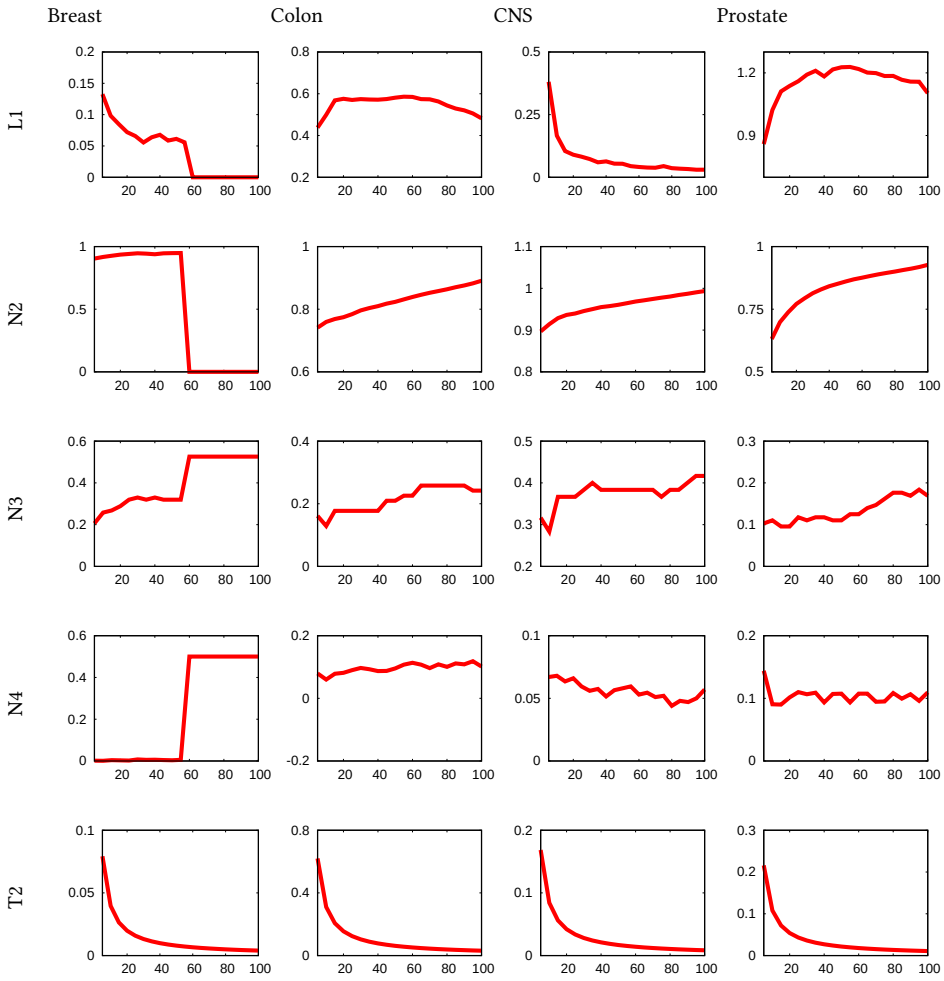

Figure 2: Plots of the data complexity measures when varying the percentage of genes

On the other hand, T1 was equal to 1 for all dimensions in all databases, which means that the class boundaries were not well defined irrespective of the dimensionality and therefore, it suggests that the ϵ-neighborhood is not related to dimensionality.

From the plots in Figure 2, it seems that N2, N3 and T2 correspond to the data complexity measures most directly related to dimensionality. However, it is also important to note that the Breast database with an extremely high dimensionality behaves differently from the rest of databases, probably because a 5% reduction of genes is not sufficient in this case to be compared with the others. Some comments should be drawn for a better understanding of the relationships between dimensionality and the other data complexities:

- L1 and N4 show that dimensionality reduction allows to increase the linear separability between classes.
- N2 indicates that dimensionality reduction alleviates the intra-class dispersion and increases the inter-class separability.
- N3 reflects that dimensionality reduction leads to classes more separated.
- As expected from the definition of this measure, T2 shows that the average number of samples per dimension decreases as dimensionality increases.

4.2 Classifier behavior as a function of dimensionality and other complexities

The classifiers included in this block of experiments were the nearest neighbor rule (1-NN), a support vector machine (SVM) using a linear kernel function with the sequential minimal optimization algorithm, the soft-margin constant $C = 1.0$ and a tolerance of 0.001, and the multi-layer perceptron (MLP) with the Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization method, two neurons in the hidden layer, a learning rate of 0.3 and 500 training epochs.

The five-fold cross-validation method was adopted for the design of this experiment because it appears to be one of the best estimators of performance compared to other strategies, such as bootstrap and re-substitution [1].

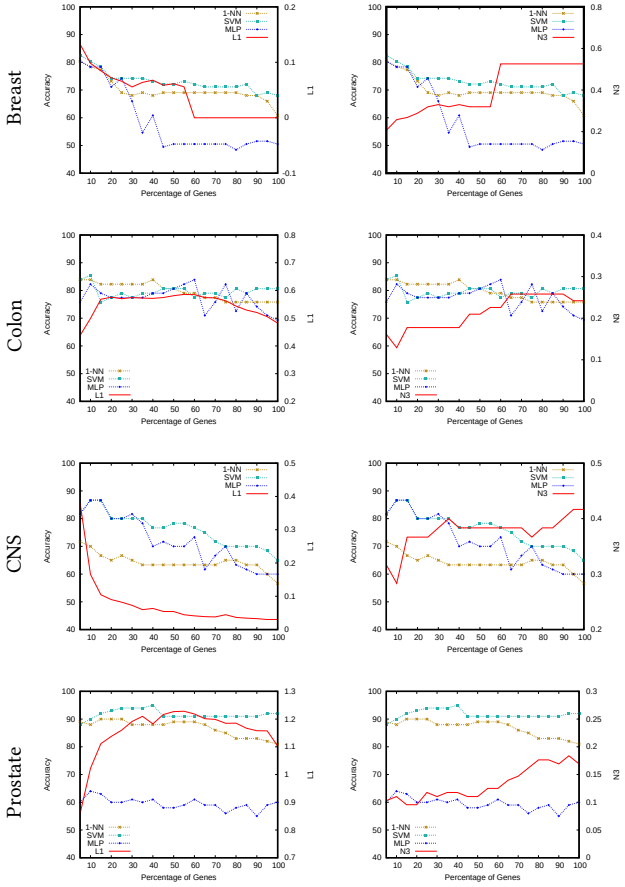

Figure 3: Plots of generalization accuracy vs. L1 and N3 when varying the percentage of genes

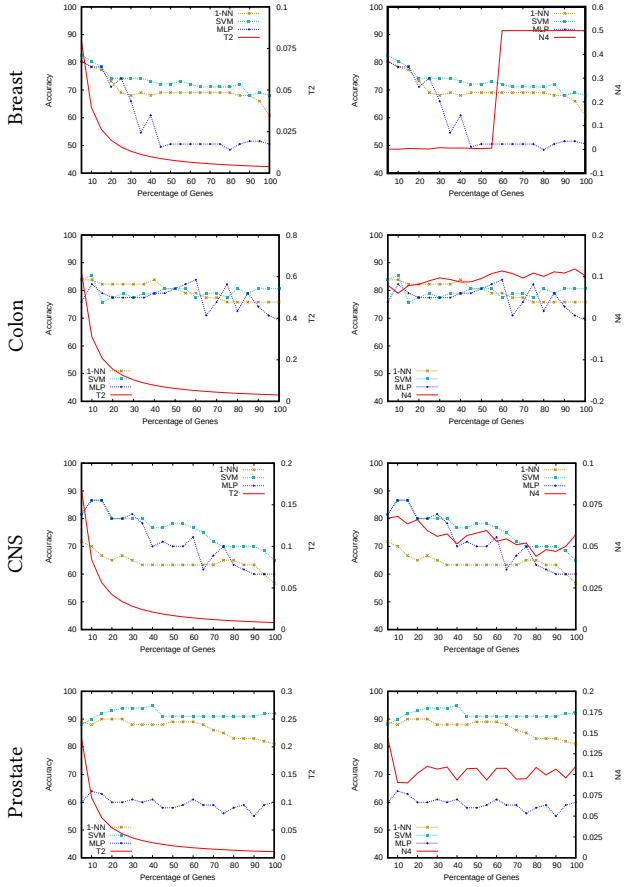

Figure 4: Plots of generalization accuracy vs. T2 and N4 when varying the percentage of genes

Figure 3 compares the accuracy rates of 1-NN, SVM and MLP with the values of L1 and N3 (two representative measures of class separability) as a function of the percentage of genes for each database. Analogously, Figure 4 depicts a comparison of the classification performance against the values of L3 and N4 (two measures of topology and geometry) when varying the percentage of genes.

It was found that all classifiers achieved the highest accuracy using a very low percentage of the top-ranked genes: 5% on Breast, and 10% on Colon and CNS. In the case of Prostate, the genes varied between 10% and 40% depending on the classifier. It is also interesting to remark that the SVM has shown superior performance in most cancer classification problems, probably because of its ability to deal with high-dimensional data and its robustness to noise [8, 14], and also because these data sets are linearly separable in the lowest dimensionalities [4].

Finally, Figure 5 shows the classification performance of 1-NN, SVM and MLP versus the values of the proposed index I_{req} when varying the percentage of genes. Examination of these figures confirms our initial hypothesis that there exist close relationships between dimensionality and data difficulty factors, which affect the performance of classifiers. In this sense, it appears that classification accuracy is related to both dimensionality and some data complexities or in other words, it seems that dimensionality reduction allows to alleviate those data difficulties and consequently, the classifiers can make better decisions.

5 CONCLUDING REMARKS

As one of the earliest works focusing on the relationships between dimensionality and several intrinsic data characteristics, this paper has analyzed the effect of dimensionality reduction on both some data complexities. As a result, two major contributions have been made.

First, we have shown that the curse of dimensionality phenomenon is related to other data difficulty factors, especially those concerning class separability. As a consequence of this, we have found that dimensionality reduction allows to reduce the effects of some other complexities on classification performance.

Second, we have introduced a new index, the I_{req}, which characterizes the link between data set size requirements and dimensionality. When applied to the experimental data sets of gene-expression

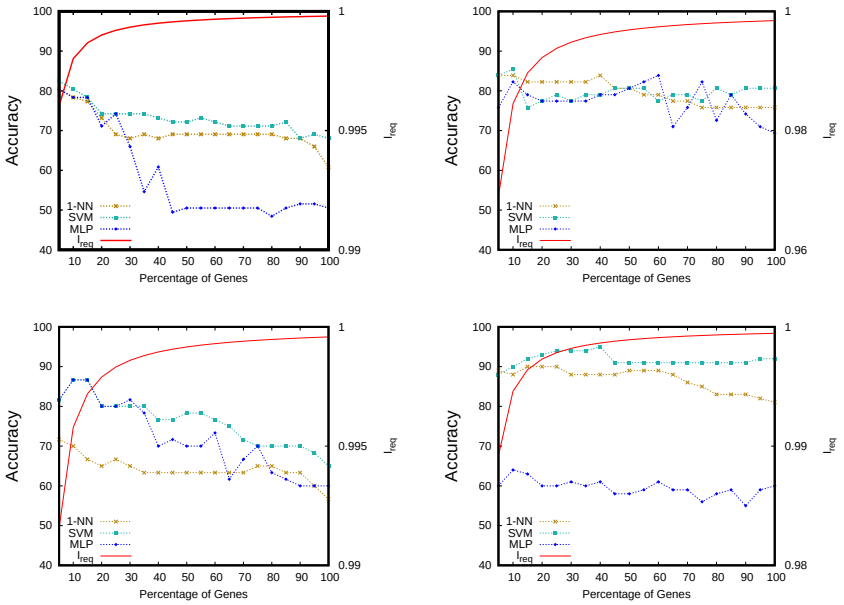

Figure 5: Plots of accuracy vs. I_{req} when varying the percentage of genes. From left to right, and from top to bottom: Breast, Colon, CNS, and Prostate

microarrays, we have observed that the number of samples required grows exponentially with the dimensionality.

ACKNOWLEDGMENTS

The work is supported by the Spanish Ministry of Economy under Grant No. TIN2013-46522-P, 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] E. Alpaydin. 2010. *Introduction to Machine Learning.* MIT Press, Cambridge, MA, USA.

[2] J.-C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed. 2016. Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection. *IEEE-ACM T. Comput. Biol. Bioinform.* 13, 5 (2016), 971–989.

[3] R. Baumgartner and R. L. Somorjai. 2006. Data complexity assessment in undersampled classification of high-dimensional biomedical data. *Pattern Recogn. Lett.* 27, 12 (2006), 1383–1389.

[4] V. Bolón-Canedo, L. Morán-Fernández, and A. Alonso-Betanzos. 2015. An insight on complexity measures and classification in microarray data. In *Proc. International Joint Conference on Neural Networks.* Killarney, Ireland, 1–8.

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

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

[7] L. Chen. 2009. Curse of Dimensionality. In *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu (Eds.). Springer, Boston, MA, 545–546.

[8] N. Cristianini and J. Shawe-Taylor. 2000. *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods.* Cambridge University Press, New York, NY, USA.

[9] E. R. Dougherty. 2001. Small sample issues for microarray-based classification. *Compar. Func. Genom.* 2, 1 (2001), 28–34.

[10] I. Guyon and A. Elisseeff. 2003. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* 3 (2003), 1157–1182.

[11] Z. M. Hira and D. F. Gillies. 2015. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Adv. Bioinformatics* 2015, ID 198363 (2015), 1–13.

[12] T.-K. Ho and M. Basu. 2002. Complexity measures of supervised classification problems. *IEEE T. Pattern Anal. Mach. Intell.* 24, 3 (2002), 289–300.

[13] T.-K. Ho, M. Basu, and M. H.-C. Law. 2006. Measures of Geometrical Complexity in Classification Problems. In *Data Complexity in Pattern Recognition*, Mitra Basu and Tin Kam Ho (Eds.). Springer, London, UK, 1–23.

[14] L. Huang, H.-H. Zhang, Z.-B. Zeng, and P. R. Bushel. 2013. Improved Sparse Multi-Class SVM and Its Application for Gene Selection in Cancer Classification. *Cancer Inform.* 12 (2013), 143–153.

[15] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe. 2012. A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis. *IEEE-ACM T. Comput. Biol. Bioinform.* 9, 4 (2012), 1106–1119.

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

[17] L. Morán-Fernández, V. Bolón-Canedo, and A. Alonso-Betanzos. 2017. Can classification performance be predicted by complexity measures? A study using microarray data. *Knowl. Inf. Syst.* 51, 3 (2017), 1067–1090.

[18] G. Nagy. 2004. Classifiers that improve with use. In *Proc. International Conference on Pattern Recognition and Multimedia.* Tokyo, Japan, 79–86.

[19] O. Okun and H. Priisalu. 2009. Dataset complexity in gene expression based cancer classification using ensembles of k-nearest neighbors. *Artif. Intell. Med.* 45, 2 (2009), 151–162.

[20] Y. Peng, W. Li, and Y. Liu. 2006. A Hybrid Approach for Biomarker Discovery from Microarray Gene Expression Data for Cancer Classification. *Cancer Inform.* 2 (2006), 301–311.

[21] M. Robnik-Šikonja and I. Kononenko. 2003. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Mach. Learn.* 53, 1–2 (2003), 23–69.

[22] Y. Saeys, I. Inza, and P. Larrañaga. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23, 19 (2007), 2507–2517.

[23] S. Singh. 2003. Multiresolution Estimates of Classification Complexity. *IEEE T. Pattern Anal. Mach. Intell.* 25, 12 (2003), 1534–1539.

[24] R. L. Somorjai, B. Dolenko, and R. Baumgartner. 2003. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics* 19, 12 (2003), 1484–1491.

[25] L. Wang, F. Chu, and W. Xie. 2007. Accurate cancer classification using expressions of very few genes. *IEEE-ACM T. Comput. Biol. Bioinform.* 4, 1 (2007), 40–53.