Analysis of complexity indices for classification problems: Cancer gene expression data

Ana C. Lorena a,1, Ivan G. Costa b,1, Newton Spolaôr a, Marcilio C.P. de Souto b,*

a Centro de Matemática, Computação e Cognição, Universidade Federal do ABC, Brazil
b Centro de Informática, Universidade Federal de Pernambuco, Brazil

1. Introduction

1.1. Motivation

Gene expression profiling studies of human diseases, such as cancer, have as main aim the identification of causal molecular mechanisms, as well as helping to improve diagnosis [3,41]. Indeed, technologies for measuring the gene expression of complete cell genomes, like microarray, have paved the way towards personalized medicine [41]. In other words, diagnoses of diseases can be based on molecular level information of individual patients, which can enhance the accuracy of diagnoses in relation to classical methods.

In the previous context, supervised machine learning (ML) methods have been successfully applied for performing gene expression-based cancer diagnosis [39]. However, cancer gene expression data sets exhibit certain characteristics that could make the classification task hard. For instance, such data present a very large number of attributes (genes) relative to the number of examples (patients) [21,39].

A great deal of research in supervised ML has focused on the development of algorithms able to create competitive classifiers with respect to generalization ability and computational time. Classification using ML techniques consists of inducing a function \( f(x) \) from a known training data set composed of \( n \) pairs \((x_i, y_i)\), where \( x_i \) is an input data and \( y_i \) corresponds to its class [28]. The induced function (classifier) should be able to predict the class of new data for which the classification is unknown, performing the desired discrimination.

As stated in [20], in several cases drawbacks in the classifier performance could arise not because of ML algorithms, but due to characteristics intrinsic to the data. In such a context, data set complexity analysis is a recent area of research. One of the aim of this research area, also known as meta-analysis of supervised ML algorithms [22], is to characterize the intrinsic complexity of a data set and find relationships (correlations) with the accuracy of the classifiers created.

The analysis presented in this paper is an extension of our previous works in [5,8,25], where we performed an investigation on the difficulty in classifying cancer gene expression data. Such a task was accomplished, mainly, using some of the complexity indices proposed in [20]. These indices measure statistics of data geometry, topology and shape of the classification boundary.

More specifically, as already discussed, microarray data are often very sparse, showing a high number of genes (features) and a low number of patients (examples). Moreover, they are noisy and can present a class unbalance: for a given data set, some types of tumors/tissues (classes) have fewer examples compared...
to others. All these could lead to difficulties in gene expression data classification.

The general framework employed in experiments for gene expression data classification often involves the following steps [35]: (1) initial pre-processing for discarding, for example, missing values, (2) feature (gene) selection for dimension reduction, and (3) induction of classification models.

Differently from our previous work in [5,8], in which Step 2 (feature selection) was not considered, in this paper we adopt all the three steps. In our work in [25], we studied the correlation of the classification errors of classifiers generated with the data sets resulting after the application of different feature selection procedures. In contrast to that work, here we perform a deeper investigation, from the point of view of the complexity indices, on the impact of the feature selection procedure in the resulting data sets.

With respect to the complexity indices, besides the ones already analyzed in our previous works, in the current paper we introduce three other indices: class balance (normalized class entropy), ratio of the principal component dimensionality to the number of instances, and the ratio of the principal component dimensionality to the real dimensionality. These complexity indices have as objective to measure two important characteristics of gene expression data: class unbalance and feature correlation.

Regarding Step 3, we apply only linear support vector machines (SVMs) in the induction of the classification models. Our choice is motivated by the experimental results presented in [5,8,24] that showed that most cancer gene expression data sets analyzed were linearly separable. Moreover, SVM consistently outperformed methods as k-NN, naive Bayes and logistic regression [5].

In summary, our basic goal is to investigate the capability of the complexity indices to explain the difficulty in the classification of cancer gene expression data, considering the popular experimental framework usually employed in the analysis of such data. This will be accomplished mainly by analyzing the correlation of the classification error rates of the classifiers generated to the values yielded by the complexity indices.

The remainder of the paper is organized as follows. Section 1.2 describes related work. Section 2 introduces some background on gene expression data analysis. The materials and methods employed in the experiments are described in Section 3. The experiments performed, along with their results, are presented and discussed in Section 4. Finally, Section 5 presents some concluding remarks.

1.2. Related work

In terms of computational experiments, as discussed in the previous section, the analysis presented here is an extension of our work in [5,8,25]. Apart from our own work, the study mostly related to ours is the one in [30]. However, in contrast to the analysis presented in this paper, whose aim is to present an extensive study of the complexity of different data sets, the main purpose in [30] is the proposal of a scheme to build multi-classifiers employing the data set complexity measures as guide.

In terms of meta-analysis of supervised ML algorithms for pattern recognition, by using a methodology that relates the classifier’s behavior to the complexity indices, the authors in [20,27,23] investigated the domain of competence of a set of popular classifiers. Based on the results from the experiments performed with different data sets from the UCI, they found that the simplest classifiers — the nearest neighbor and the linear classifier — have extreme behavior in that they mostly behave either as the best approach for certain kinds of problems or as the worst approach for other types of problems.

Still in the context of meta-analysis of supervised ML algorithms, in [22] the authors presented analyses regarding issues such as the discovery of similarities among classification algorithms, and among data sets. One of the differences of the work in [22] to the ones in [20,27,23] is the set of indices used to characterize the data sets. For instance, the measures used in [22] are basically statistical and information theoretic descriptors (e.g., percentage of symbolic attributes and normalized class entropy), whereas in [20,27,23] the focus is on measures that capture the data geometry.

More recently, based on the works in [22,10], the authors in [11], employed meta-learning techniques to the problem of algorithm recommendation for gene expression data classification.

2. Gene expression data

Genes are linear sequences of nucleotides along a segment of a DNA molecule that provide the coded instructions for synthesizing RNA molecules [1]. RNA molecules are often translated into proteins: the main building blocks of all organisms. This whole process is called gene expression.

The expression level of a gene can be regarded as an estimate of the amount of proteins it produces in a given period. Different technologies can be used to measure the expression levels of genes. One of the most important representatives is the microarray technology, which allows the measurement of the expression level of thousands of genes simultaneously [31]. By employing this technology, different kinds of biological experiments can be performed.

One can apply, for example, microarray to perform an essay whose goal is to compare gene expression levels in different types of tissues (e.g., normal and tumor tissues). Next, the data obtained from this experiment could be employed to aid the diagnosis of diseases, through the classification of distinct kinds or subtypes of tumors according to their expression patterns (profiles) [3,13,14,40,44]. It is also possible to design experiments whose aim is the identification of genes that are mostly related to a certain disease, which could be then target for future medicines and genetic therapies (e.g., the work in [18]).

Our work is mainly concerned with data regarding cancer diagnosis. Cancer diagnosis, in general, relies on a variety of microscopic and immunologic tissue tests. The presence of tumor samples with atypical morphologies can often make such a task harder [32]. Furthermore, some cancer tissues from different kinds of tumors (or subtypes) can present low differentiation, which can make the laboratory identification based only on morphology and immunophenotyping complex. To minimize this problem, one can use microarray to design a biological experiment with the aim of characterizing the molecular variations among tissues, by monitoring gene expression profiles in a genomic scale [13,18,31,44].

3. Material and methods

3.1. Data sets

In our investigation, we use 23 microarray data sets. They are a subset of a set of benchmark microarray data introduced in [7]. As Table 1 illustrates, such data sets present different characteristics for aspects like: type of microarray chip (Chip); number of data items (n); number of classes (c); distribution of data within the classes (Dist. Classes), where we have the mean (mean),

\[ \text{mean} \]
standard deviation (std), minimum (min) and maximum (max) numbers of examples per class; dimensionality (d); and the ratio between the number of data items and the dimensionality of data (n/d).

Regarding the second column in Table 1, microarray technology is usually available as two different types of platforms, single- or double-channel microarrays (e.g., Affymetrix) or double-channel microarrays (e.g., cDNA) [31,35]. Other microarrays technologies are also based on either single or double channel methods. Since the data sets analyzed here are constrained to those collected with cDNA and Affymetrix microarrays, we use the terms cDNA and Affymetrix to denote double or single-channel arrays, respectively.

Since many of the complexity indices are, by definition, only suitable for binary classification problems [20], for multi-class classification data sets, we employ a one-against-all decomposition. That is, from a data set with c class, we build c binary data sets. In each one of these data sets, one class is considered positive (+1) and the remaining classes are considered negative (−1).

The use of a one-against-all decomposition strategy has also been indicated in studies analyzing the problem of multi-class gene expression data classification with support vector machines as often leading to overall best classification performance [40]. After the application of the decomposition procedure, we have a total of 80 binary data sets.

For many of the previous data sets there is an unbalance in the distribution of data items per class. One can see this clearly by looking at the standard deviation values of the number of examples per class. For instance, there are very large standard deviation values in data sets such as Bhattacharjee, Gordon and Yeoh-V1, indicating that they are highly unbalanced regarding the distribution of examples per class. The data sets Armstrong-V2, Alizadeh-V1 and Bhattacharjee, on the other hand, are very balanced.

Still with respect to the data sets, from column "n/d" we can observe the data sparsity induced by the high dimensionality of the data sets. On average this ratio is of 0.06 with a standard deviation of 0.04 for all data sets. This clearly indicates that the ratio between the number of examples and genes is low, that is, there are usually much more genes than data items (patients) available. This is a characteristic of most cancer gene expression data sets, since obtaining patients’ samples is much more laborious and expensive than measuring a high number of gene expression values.

Hereafter, we will use the following notation. Let X be an n by d matrix representing a gene expression data set, where xij denotes the expression value of sample i and feature (gene) j, x is a d-dimensional vector with the expression values of sample (patient) i and x is an n-dimensional vector representing the expression values of feature (gene) j. Moreover, y is an n-dimensional vector, where yi ∈ {1,...,c} corresponds to the class (or cancer type) of patient i and c is the total number of classes (cancer types) in the data set.

3.2. Feature selection procedure

In terms of feature selection, one of the main aspects is the use of statistics to select features (or genes) which best discriminate the classes of patients. We inspect in this work a simple and well-known feature selection procedure based on the Fisher linear discriminant analysis [13]. This measure is basically the ratio of their between-class to within-class sums of squares.

For feature (gene) j this measure is defined as follows:

$$\text{BSS/WWSS}(j) = \frac{\sum_{i \in C_1} \sum_{j} (y_i - \bar{y}_{C_1})^2}{\sum_{i \in C_1} \sum_{j} (y_i - x_j)^2}$$

where \(x_j\) represents the average level of expression of gene \(j\) among all samples, \(x_{ij}\) represents the average level of expression of gene \(j\) among the samples belonging to class \(k\) and \(I\) is an indicator function.

In the previous context, before selecting the features, we first rank them according to their respective BSS/WWSS values: the larger the BSS/WWSS value for a feature, the better is its discrimination power.

Finally, it is important to point out that, in spite its simplicity and feature independence assumption, the results of the experiments in [13] showed the BSS/WWSS presented good results in the context of cancer gene expression data.

3.3. Classification method: support vector machines

Support vector machines (SVMs) are learning algorithms based on the Statistical Learning Theory [6]. Given a training data set composed of \(n\) instances from classes \((-1,+1)\), the algorithm searches for a hyperplane \(\mathbf{w} \cdot \mathbf{x} + b = 0\) able to separate the instances from these two classes, where \(\mathbf{w}\) is a weight vector orthogonal to the hyperplane and \(b\) is an offset term.

According to the principles of the Statistical Learning Theory, the optimal hyperplane is the one that maximizes the margin of separation between the classes. This hyperplane can be obtained by solving the following optimization task:

$$\text{minimize} : \quad \|\mathbf{w}\|^2 + \frac{C}{n} \sum_{i=1}^{n} \xi_i$$

s.t. : \(y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i\), \(\xi_i \geq 0\), \(i = 1, \ldots, n\)

where \(C\) is a regularization parameter that imposes a trade-off between training error and generalization and the \(\xi_i\) are slack variables. The constraints are imposed in order to ensure that few training samples should be within the margins or misclassified, allowing the technique to deal with noisy data. The number of training errors and instances within the margins allowed is controlled by the minimization of the term \(\sum_{i=1}^{n} \xi_i\).

The previous definition can be generalized to non-linearly separable classification problems by the use of non-linear Kernel functions. However, as explained in the Introduction, since the

<table>
<thead>
<tr>
<th>Data set</th>
<th>Chip</th>
<th>n</th>
<th>c</th>
<th>Dist. classes</th>
<th>d</th>
<th>n/d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std</td>
<td>Min.</td>
<td>Max.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Armstrong-V2</td>
<td>Affy</td>
<td>72</td>
<td>3</td>
<td>24.0 4.0</td>
<td>28</td>
<td>2194.03</td>
</tr>
<tr>
<td>Bhattacharjee</td>
<td>Affy</td>
<td>203</td>
<td>5</td>
<td>40.6 55.3</td>
<td>139</td>
<td>1543.03</td>
</tr>
<tr>
<td>Dyrskjot</td>
<td>Affy</td>
<td>40</td>
<td>3</td>
<td>13.3 5.9</td>
<td>20</td>
<td>1203.03</td>
</tr>
<tr>
<td>Golub-V2</td>
<td>Affy</td>
<td>72</td>
<td>3</td>
<td>24.0 14.5</td>
<td>38</td>
<td>1877.04</td>
</tr>
<tr>
<td>Gordon</td>
<td>Affy</td>
<td>181</td>
<td>2</td>
<td>90.5 84.1</td>
<td>150</td>
<td>1626.11</td>
</tr>
<tr>
<td>Lallo</td>
<td>Affy</td>
<td>37</td>
<td>2</td>
<td>18.5 14.8</td>
<td>29</td>
<td>2202.02</td>
</tr>
<tr>
<td>Nutt-V1</td>
<td>Affy</td>
<td>50</td>
<td>4</td>
<td>12.5 3.7</td>
<td>15</td>
<td>1377.04</td>
</tr>
<tr>
<td>Pomeroy-V1</td>
<td>Affy</td>
<td>34</td>
<td>2</td>
<td>17.0 11.3</td>
<td>25</td>
<td>857.04</td>
</tr>
<tr>
<td>Ramaeky</td>
<td>Affy</td>
<td>59</td>
<td>5</td>
<td>14.6 6.0</td>
<td>30</td>
<td>1363.04</td>
</tr>
<tr>
<td>Shipp</td>
<td>Affy</td>
<td>77</td>
<td>2</td>
<td>38.5 27.6</td>
<td>58</td>
<td>798.10</td>
</tr>
<tr>
<td>Su</td>
<td>Affy</td>
<td>174</td>
<td>10</td>
<td>17.4 9.3</td>
<td>28</td>
<td>1571.04</td>
</tr>
<tr>
<td>West</td>
<td>Affy</td>
<td>49</td>
<td>2</td>
<td>24.5 11.3</td>
<td>25</td>
<td>1198.04</td>
</tr>
<tr>
<td>Yeoh-V1</td>
<td>Affy</td>
<td>248</td>
<td>2</td>
<td>124.0 114.6</td>
<td>205</td>
<td>2526.10</td>
</tr>
<tr>
<td>Alizadeh-V1</td>
<td>cDNA</td>
<td>42</td>
<td>2</td>
<td>21.0 0.0</td>
<td>21</td>
<td>1095.04</td>
</tr>
<tr>
<td>Alizadeh-V2</td>
<td>cDNA</td>
<td>62</td>
<td>3</td>
<td>20.7 18.5</td>
<td>42</td>
<td>2093.03</td>
</tr>
<tr>
<td>Bittner</td>
<td>cDNA</td>
<td>38</td>
<td>2</td>
<td>19.0 0.0</td>
<td>19</td>
<td>2201.02</td>
</tr>
<tr>
<td>Broadel</td>
<td>cDNA</td>
<td>50</td>
<td>3</td>
<td>16.7 13.2</td>
<td>31</td>
<td>1739.03</td>
</tr>
<tr>
<td>Garber</td>
<td>cDNA</td>
<td>66</td>
<td>4</td>
<td>16.5 16.7</td>
<td>40</td>
<td>4553.01</td>
</tr>
<tr>
<td>Khan</td>
<td>cDNA</td>
<td>83</td>
<td>4</td>
<td>20.8 7.9</td>
<td>11</td>
<td>1069.08</td>
</tr>
<tr>
<td>Lapoint-V2</td>
<td>cDNA</td>
<td>110</td>
<td>4</td>
<td>27.5 14.8</td>
<td>41</td>
<td>2406.04</td>
</tr>
<tr>
<td>Liang</td>
<td>cDNA</td>
<td>37</td>
<td>3</td>
<td>12.3 13.7</td>
<td>3</td>
<td>1411.03</td>
</tr>
<tr>
<td>Risinger</td>
<td>cDNA</td>
<td>42</td>
<td>4</td>
<td>10.5 7.0</td>
<td>19</td>
<td>1771.02</td>
</tr>
<tr>
<td>Tomlins-V1</td>
<td>cDNA</td>
<td>104</td>
<td>5</td>
<td>20.8 8.7</td>
<td>32</td>
<td>2315.04</td>
</tr>
</tbody>
</table>
data sets analyzed are mostly linearly separable, in this work we employ only linear SVMs. More specifically, we apply the Sequential Minimal Optimization (SMO) algorithm implemented in the Weka tool [42] for the induction of the SVMs. Furthermore, since we observed low variations of results with respect to different values of the parameter C for the data used in this current work, we decided to use its default value.

3.4. Complexity measures of supervised classification tasks

Following the description in [20], we present the complexity indices, which will be analyzed in this work, by grouping them according to which aspect in the data they should focus: (1) measures of overlap, (2) measures of linear separability, and (3) measures of topology. Note that indices that measure the same property could display some degree of correlation.

3.4.1. Measure of overlap

Fisher’s discrimination ratio (F1) returns the Fisher statistics of the feature with the largest contribution to class discrimination:

\[
F_1 = \frac{\max_{j} \left( \frac{\bar{x}^{(k)}_j - \bar{x}^{(c)}_j}{\sigma^{(k)}_j + \sigma^{(c)}_j} \right)^2}{\max_{j} \left( \frac{\bar{x}^{(k)}_j - \bar{x}^{(c)}_j}{\sigma^{(k)}_j + \sigma^{(c)}_j} \right)^2}
\]

where \( \bar{x}^{(k)}_j \) is the mean of feature \( j \) in class \( k \) and \( \sigma^{(k)}_j \) is the variance of the class \( k \) for feature \( j \). F1 has positive values, where higher values indicates simpler classification problems.

Length of the overlapping region (F2) measures the length of the overlap between the distributions of values in distinct classes. For each feature, we measure the area of overlap between classes and normalize it by the total length of the distribution of both classes. Let \( \text{max}(x_j) \) and \( \text{min}(x_j) \) be the maximum and minimum value of feature \( j \) at class \( k \).

\[
F_2 = \frac{1}{N} \sum_{x} \max(0, \min(\text{max}(x_j), \text{max}(x_j)) - \min(\text{min}(x_j), \text{min}(x_j)))
\]

For such an index 0 indicates no overlap between classes, whereas positive values indicates overlap, that is, more complex classification data sets.

3.4.2. Measures of class separability

Linear separability indices (L1 and L2) evaluate if the classes are linearly separable. In order to calculate these indices, a linear programming method for finding the optimal linear classifier is employed only linear SVMs. More specifically, we apply the Sequen-

Principal component dimensionality/sample ratio (T2): To measure the principal component dimensionality of a data set, we apply principal component analysis to the expression matrix \( X \). We define the principal component dimensionality, \( d \), as the number of principal components (PCs) explaining 95% of the data variation. The final estimate is

\[
T_2 = \log(d/n)
\]

Principal component dimensionality/dimensionality ratio (T3): This index measures the proportion of dimension reduction when using only the PCs explaining 95% of the data variability:

\[
T_3 = d/d
\]

Classes balance (C1) measures the classes balance in a data set by estimating the normalized entropy of the class size distribution [22], according to

\[
C_1 = - \frac{1}{\log(c)} \sum_{k=1}^{c} \frac{n_k}{n} \log \frac{n_k}{n}
\]

where \( n_k \) is the number of samples in class \( k \) and \( c \) is the number of classes. This index will have value 0 if one class has all samples and value 1 if classes have same number of samples. As often ML techniques, because of overfitting, have problems in dealing with unbalanced data sets, lower values of C1 could be seen as indicating more complex data sets.

3.4.4. Discussion

In [20], the authors present an index similar to our F2. Our version of the index, F2, was introduced in [5]. The basic difference between our version and the one in [20] is that the latter, instead of a sum (length of the overlap), computes the product (volume of the overlap) of the normalized lengths of the overlapping ranges. A side effect of employing the product is that the value of the measure decreases drastically as dimensionality increases—F2 would be the product of values in the range [0,1]. Thus, as the data we deal with present a very high dimension, we opted to use the sum.

As already mentioned, for calculating indices T2 and T3 we use principal component analysis (PCA). PCA can be used to transform a number of possibly correlated variables (features) into a smaller number of uncorrelated variables called principal components (PCs) [12]. The first PC accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

Indeed, the motivation for the definition of T2 and T3 is the fact that groups of genes participating in the same biological process, also called expression modules [33], usually have the same expression pattern over a particular biological condition. For example, the current state of art in marker selection is to use biological networks, such as pathways or protein interaction [4], to extract correlated and functionally related set genes as markers. Therefore, one can also expect genes (or features) of cancer data sets to be highly correlated.

The main motivation for the index C1 is the frequent class size unbalance present in cancer gene expression data sets [2] (see Table 1). More specifically, like in [22], we use the very well-known normalized class entropy equation.

Finally, we would like also to point out that, as in [20], the complexity measures are computed for each problem using all available instances. There are no data source models for any of the problems. We also assume that attributes that present missing values have the values replaced by the respective mean value of the

3.4.3. Measures of topology

Dimensionality/samples ratio (T1) simply measures the log of the ratio between the number of features and the number of samples:

\[
T_1 = \log(d/n)
\]

Principal component dimensionality/sample ratio (T2): To measure the principal component dimensionality of a data set, we apply principal component analysis to the expression matrix \( X \). We define the principal component dimensionality, \( d \), as the number of principal components (PCs) explaining 95% of the data variation. The final estimate is

\[
T_2 = \log(d/n)
\]

Principal component dimensionality/dimensionality ratio (T3): This index measures the proportion of dimension reduction when using only the PCs explaining 95% of the data variability:

\[
T_3 = d/d
\]

Classes balance (C1) measures the classes balance in a data set by estimating the normalized entropy of the class size distribution [22], according to

\[
C_1 = - \frac{1}{\log(c)} \sum_{k=1}^{c} \frac{n_k}{n} \log \frac{n_k}{n}
\]

where \( n_k \) is the number of samples in class \( k \) and \( c \) is the number of classes. This index will have value 0 if one class has all samples and value 1 if classes have same number of samples. As often ML techniques, because of overfitting, have problems in dealing with unbalanced data sets, lower values of C1 could be seen as indicating more complex data sets.

3.4.4. Discussion

In [20], the authors present an index similar to our F2. Our version of the index, F2, was introduced in [5]. The basic difference between our version and the one in [20] is that the latter, instead of a sum (length of the overlap), computes the product (volume of the overlap) of the normalized lengths of the overlapping ranges. A side effect of employing the product is that the value of the measure decreases drastically as dimensionality increases—F2 would be the product of values in the range [0,1]. Thus, as the data we deal with present a very high dimension, we opted to use the sum.

As already mentioned, for calculating indices T2 and T3 we use principal component analysis (PCA). PCA can be used to transform a number of possibly correlated variables (features) into a smaller number of uncorrelated variables called principal components (PCs) [12]. The first PC accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

Indeed, the motivation for the definition of T2 and T3 is the fact that groups of genes participating in the same biological process, also called expression modules [33], usually have the same expression pattern over a particular biological condition. For example, the current state of art in marker selection is to use biological networks, such as pathways or protein interaction [4], to extract correlated and functionally related set genes as markers. Therefore, one can also expect genes (or features) of cancer data sets to be highly correlated.

The main motivation for the index C1 is the frequent class size unbalance present in cancer gene expression data sets [2] (see Table 1). More specifically, like in [22], we use the very well-known normalized class entropy equation.

Finally, we would like also to point out that, as in [20], the complexity measures are computed for each problem using all available instances. There are no data source models for any of the problems. We also assume that attributes that present missing values have the values replaced by the respective mean value of the
attribute. Hence, we constrain our claims to be for the apparent complexity of the underlying problems as presented in the given data sets.

4. Experiments and results

As discussed in the Introduction, we adopt a three step experimental framework [35]:

1. Initial pre-processing for discarding, for example, missing values and scaling the values of the features.
2. Feature (gene) selection for dimension reduction.
3. Induction of classification models.

For the initial pre-processing, we apply the same strategy as in [7]. That is, we discard genes that showed more than 10% of missing values. The remaining missing values are substituted by the average expression values of the genes through all samples. We also transform the feature values so that they lie within similar ranges (scale).

As already discussed, given that most of the complexity indices are only able to be employed in classification problems with two classes, we applied a one-against-all decomposition for each data set. This yielded 80 decomposed data sets, for which the classification results are reported.

The feature selection method employed is the BSS/WSS, described in Section 3.2. As base line, we also conducted a random selection of genes for each data set. Due to the stochastic nature of such a procedure, we performed this selection 10 times for each combination of data set and number of genes.

More specifically, for each data set we performed feature selection using a nested cross-validation procedure, selecting for each training data set the top 500, 250, 100, 50, 25 and 10 genes. In the end, we keep the best test error and the number of selected genes for each data set. We induce classification models through the use of linear SVMs (see Section 3.3). As the data sets present a very small number of instances, all error rates are estimated through a leave-one-out procedure.

In the following, we discuss the results of the complexity indices considered. Then, we analyze further the ability of some of these indices to capture the class structure in the data sets.

4.1. Complexity indices results

We computed first the classification complexity measures for all the original data sets in Table 1. This step was then repeated for the case where we applied feature selection procedures (BSS/WSS and random selection). More specifically, for the latter, we calculate the complexity indices for those data sets with best performances after the gene selection.

Fig. 1 presents histograms of the distributions of the values of the complexity indices for the data sets (a) with all features, and with a subset of the features, selected by (b) BSS/WSS or (c) random selection. It is important to point out that, since the feature selection procedure do not make changes in the numbers of examples per class, $C_{1}$ has the same distribution for all data sets.

Having a closer look at the indices and their distributions, we can identify some interesting properties of the data sets (see Fig. 1).

- Taking into account that higher values of $F_{1}$ indicates simpler classification problems, in some cases the use of BSS/WSS led to a decrease in the complexity of the classification problem.

![Fig. 1. Histogram—classification complexity values.](image-url)
One should observe that, since both F1 and BSS/WSS are based on the same principles (Fisher discriminant analysis), a higher correlation in this context would indeed be expected.

- According to the distribution of values for F2, there is less overlap between classes for data sets resulting after BSS/WSS feature selection, when compared to the one found for the original data. This was also expected, since the BSS/WSS principle lies on choosing features that maximize the separation of the classes, which, in its turn, will tend to minimize the measure of overlap.

- For N1, N2 and N3, the original data sets led to higher values, whereas BSS/WSS data sets tend to present lower values for such measures. This indicates a simplification of data structure after feature selection. As these indices can be seen as metric based (Euclidean distance), the results obtained reinforce that high dimensionality pose problems to data classification. More specifically, for N1, for example, higher values indicate smaller separation in the distributions of the classes. This is consistent with the remarks made for F1 and F2 in the previous two items.

- According to the values of L1 and L2, basically, all data sets considered are linearly separable. That is, most of the values calculated for L1 and L2 were zero or very close to it. In the case of L1, values close to 0 also indicate that examples incorrectly classified by a linear classifier are near the class boundary.

- By definition, the values of T1 are higher for the classifier errors when compared to those with feature selection. This holds also true for the case of T2. Moreover, in the presence of redundant features, as T2 is based on PCA, the values for such an index should be much smaller than those for T1.

- From the values for T3, we can see that the BSS/WSS procedure led in some cases to a number of features close to that which indeed capture most of the variability of the data. In contrast, this behavior cannot be observed when all features are used. In fact, for all features, the results of T3 are mainly close to 0, that is, the total number of features is much larger than the number PCs that represent most of the variability of the data after a PCA.

- In the case of C1, 16 out of the 80 data sets have C1 lower than 0.93, which indicates that the largest class has at least twice as many samples than the smallest class. This is in accordance to Table 1, where the sizes of the largest and smallest classes shows large differences.

From the previous results, we can see that (1) the BSS/WSS feature selection procedure seem to be effective in reducing data set complexity (values of the indices F1, F2, N1, N2, N3); and (2) the distribution of values for some indices reflect the consequences of the curse of dimensionality, a problem well-known to be related to cancer gene expression data sets (values of indices T1, T2, T3) [13,26].

For a further investigation, we drew a scatter plot of the classifier errors versus the values of the complexity measures calculated. Fig. 2 show scatter plots of the indices (x-axis) versus classification error (y-axis) for the classification and feature selection methods employed. Each point corresponds to a binary data set.

Besides the complexity indices discussed in Section 3.4, we also included the number of samples n and the dimension of the data sets d. In order to analyze if the classification error rate of a given method is related to a specific complexity measure, we calculated their correlation coefficient and performed a t-test [37]. These numbers are shown in the left/right upper corner of each scatter plot. Those in italic and bold indicate correlations that are statistically significant (t-test with p-value < 0.05).

First, we consider the indices measuring the dimensionality and sparsity of data: n, d, T1, T2 and T3.

The number of samples n has a significant negative correlation with the error rate in all cases. Since statistically higher numbers of samples leads to lower classification errors, this was somewhat expected. Also, in accordance to [16], these results evidence that the main source of problem in microarray data sets can be the small number of samples they have.

The dimension d of the data sets did not significantly relate to the error rates of the classifiers induced. When we consider T1, which is the ratio of d to n, there is a significant correlation for the case in which all genes are used. For T2, there is a significant correlation when all genes and BSS/WSS are used. In this context, T2 can be considered more precise than T1 in capturing the relationship between the dimensionality of the data sets and their number of samples, since we discard redundant dimensions via PCA. As a consequence, the correlation of the values of T2 with the classifier errors are higher in the scenarios where feature selection was performed. This indicates the usefulness of this new complexity index in measuring data sparsity.

There are also significant correlations with the error rates for the indices: F1, F2, N1, N2, N3 and C1. F1 is negatively correlated to the errors, since lower values of F1 indicate more complex data sets, which, in its turn, will tend to produce classifiers with higher error rates. F2 is significantly related to the error rates of the classifiers generated after feature selection only. Since BSS/WSS minimizes the overlap between the classes, the relation between T2 and the classification error is highlighted.

As expected from their definition, we also see that higher values for N1, N2 and N3 implied in less accurate classifiers. The correlations of N1 and N3 with the error rate is increased for the data sets after feature selection. As already mentioned, such indices can be regarded as metric based (Euclidean distance). Thus, intrinsically, they would have more difficulties in capturing characteristics of high dimension data sets, as those using all genes.

It is important to point that, though the definition of N1, N2 and N3 are similar, N2 displayed lower correlation than N1 and N3. A possible reason for this, as observed in [8], is the fact that N2 is based on the ratio of the average of intra-/inter-class nearest neighbors, whereas N1 and N3 are sensitive to which (intra or inter) neighbor (sample) is closer to a point. Thus, as N2 deal with averages, it could be more susceptible to the data sparsity induced by the high dimensionality of the data sets.

Concerning C1, we can verify a positive correlation between error rate and class balance. This indicates that classifier induced tend to be bias towards classifying the largest class. This has already been observed in gene expression data [2] and indicates the usefulness of this index in measuring classification complexity.

For some of the indices (L1, L2 and T3), we see no significant correlation to the classifier errors. This is the case for both L1 and L2. As defined in [20], the linear classifier used to calculate L1 and L2 is built using all the available data. Here it is important to point out that data sets with a small number of instances and high dimensionality, like the ones used in this paper, often lead to sparsity, introducing another layer of difficulty via a lack of restrictions on the generalization rules for the classifier. Thus, in our specific context, the values near zero of L1 and L2 reflect not that all data sets are linearly separable but that an overfitting occurred. Therefore, these measures, such as they have been defined, do not seem appropriate to describe our type of data sets.

It is interesting to notice that the random selection of genes was generally successful in generating gene lists with good predictive performance. Fig. 3 presents a histogram of the error rates of the SVM classifiers with and without gene selection,
where the notation used is the same as the one in Fig. 2. A table with all error rates is not shown due to the large number of data sets and, consequently, classification results. Also, we can clearly see that BSS/WSS tend to produce classifiers with the lowest error rates.

BSS/WSS was also more effective in reducing the dimensionality of the data sets, as indicated by Fig. 4, which presents a histogram of the numbers of genes selected in the feature selection experiments. Random selection, on the other hand, selected larger numbers of genes. Therefore, in order to obtain a good predictive performance, random selection has to take more genes into account. We can also say that when more genes are considered, it is relatively easy to find different subsets with good predictive performance. The relatively good results of a random selection of genes makes clearer the existence of multiple lists of genes that provide good predictive results in cancer diagnosis. This is in accordance to observations in [15,16,19,34,38].
4.2. Class structure in the data

In this section, we investigate further the ability of some of the indices in capturing the intrinsic complexity of cancer gene expression data. To accomplish this, we first added noise to the original data sets, by performing permutations in the labels of the samples in a rate of 20% and 40%, respectively. By doing so, we intend to modify the existing class structure of the data. This modification is expected to make the resulting classification problems more complex than the original ones. Hereafter, we will refer to the sets in this collection as random noise data sets.

Once we have built the random noise data sets, we compute the values of complexity indices for them and compare to those obtained for the original data. The main goal is to observe whether the complexity indices could, indeed, capture modifications (noise) made to the class structure of the gene expression data sets investigated.

The error rates of the distinct random noise data sets are depicted in Fig. 5. As expected, the increase of the classifier error rates was proportional to the amount of noise added to the data sets (20% and 40%, respectively).

We also analyzed if, by comparing the distribution values of the complexity indices, we could identify a threshold which could separate the original and random noise data sets—Fig. 6. We only show indices that had highlighted results in the previous analysis and that present distinct distributions for the original and random data sets. For example, by definition, there are some complexity indices which will exhibit the same distribution for either type of data sets (e.g., T1, T2 and T3).

As Fig. 6 illustrates, indices F1, N1, N2 and N3 were effective in assigning the original and random data sets collections to opposite sides of the ranges. We can assess this by finding a threshold value for each index, which can best separate random from original data sets.

Once we establish a kind of threshold, to have an idea on how well the original and random data is separated, we count how many data sets are in the wrong value ranges. For example, in the case of N1, considering a threshold of 0.275, we can observe that most of the original data sets are below this threshold, whereas most of the random data sets (20% random labels) are above the threshold.

More specifically, only four (out of the 80 decomposition) of the original data sets presented a value higher than 0.275 and two random data sets (20% random labels) had value lower than 0.275. When we consider the random data sets with 40% of noise, the threshold is 0.325. In this context, there is only one misplaced data set, which is one of the original data sets. That is, the random and original label were almost perfectly distinguished by the complexity index N1.

We also measured the percentage of data sets that were wrongly assigned by the threshold method for the complexity indices F1, F2, N1, N2 and N3. The percentage of misplaced data sets for indices F1, F2, N1, N2 and N3 were, respectively, 31.25%, 42.50%, 7.5%, 25.0% and 6.25%, when the original data is compared with 20% random labels data. Comparing the original data sets and noise data sets with 40% random labels, the percentage of misplaced data sets were, respectively, 8.75%, 23.75%, 1.25%, 2.5% and 1.25%. That is, N1 and N3 were the indices more successful in discriminating original from random data sets. N2 and F1 also displayed some discriminative power, whereas F2 had a rather modest performance.
5. Final remarks

The previous results show that some of the characteristics of cancer gene expression data do influence the classification performance of the classifiers built. The most important aspects are related to data sparsity and class imbalance. In this context, here, we proposed three new complexity indices, two of which were successful in capturing these aspects. We also showed that, when a dimensionality reduction is accomplished by a proper feature selection, the impact of these characteristics tends to be decreased.

As will be discussed next, the results in this paper are in accordance with our previous findings and complement them [5, 8, 25].

For example, in [5] we showed that N1 and T1 were correlated with the classification test error rates for different classifiers generated, independently of the ML algorithm used—in that work, besides linear SVMs, we performed experiments with other simple classifiers (e.g., k-NN and Naive Bayes). More specifically, the correlation with T1 pointed out that the sparsity of the data is a major aspect for indicating the complexity of gene expression-based classification. These results were also verified in this current work.

In [8] we confirmed that some of these indices are indeed good descriptors of the difficulty involved in using these data sets for classification, in particular N1 and N3. There, we generated noisy versions of the data sets and verified how the index values varied when data structure is changed. In the present paper a similar procedure was adopted, but including also a higher level of noise. Here, besides N1 and N3, N2 and F1 also displayed some discriminative power.

In [25] we performed experiments with T1 that showed that feature selection reduced the influence of data sparsity. Such a behavior was also observed in this work—but, here, besides T1, we analyzed the other complexity indices.

In the context of gene expression, feature redundancy is a relevant aspect, which is not approached by the classical complexity indices proposed in [20]. We proposed here two indices for measuring data sparsity which aimed to capture the non-redundant dimension (T2 and T3). However, none of these two indices displayed an improvement in capturing the data sparsity as compared to T1. The use of other methodologies for measuring feature redundancy, such as the one in [43] is an important future aspect.

A challenging topic for further research is the definition of complexity measures, which take, at the same time, both data preparation (e.g., feature selection and treatment of missing values) and classification into account. Other important future direction is to employ complexity indices, in particular those related to data sparsity and class balance, for building a method for recommending the number of samples necessary for a robust gene selection and classification results [16]. Another interesting direction is the use of the complexity indices as meta-attributes for meta-learning. In such a context, for instance, we could build a classifier to suggest the most appropriate classification/feature selection method for a given data set [9, 29].

Acknowledgment

The authors would like to thank the Brazilian research agencies CNPq, CAPES, FACEPE and UFABC for the financial support provided.

References


