Neurocomputing 75 (2012) 33–42

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

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

ARTICLE INFO

Available online 31 July 2011

Keywords:
Classification
Gene expression data
Complexity indices
Linear separability

ABSTRACT

Currently, cancer diagnosis at a molecular level has been made possible through the analysis of gene expression data. More specifically, one usually uses machine learning (ML) techniques to build, from cancer gene expression data, automatic diagnosis models (classifiers). Cancer gene expression data often present some characteristics that can have a negative impact in the generalization ability of the classifiers generated. Some of these properties are data sparsity and an unbalanced class distribution. We investigate the results of a set of indices able to extract the intrinsic complexity information from the data. Such measures can be used to analyze, among other things, which particular characteristics of cancer gene expression data mostly impact the prediction ability of support vector machine classifiers. In this context, we also show that, by applying a proper feature selection procedure to the data, one can reduce the influence of those characteristics in the error rates of the classifiers induced.

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 $f(\mathbf{x})$ from a known training data set composed of n pairs $(\mathbf{x}_i,y_i)$, where $\mathbf{x}_i$ 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

[*] Corresponding author.
E-mail addresses: ana.lorena@ufabc.edu.br (A.C. Lorena), igcf@cin.ufpe.br (I.G. Costa), newton.spolaor@ufabc.edu.br (N. Spolaôr), mcps@cin.ufpe.br (M.C. de Souto).

[1] The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.03.054

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.[2] 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**),

[2] http://algorithmics.molgen.mpg.de/Supplements/CompCancer/