Artificial Intelligence in Medicine 61 (2014) 35–44

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Multi-test decision tree and its application to microarray data classification

Marcin Czajkowski[a,*], Marek Grześ[b], Marek Kretowski[a]

[a] *Faculty of Computer Science, Bialystok University of Technology, Wiejska 45a, 15-351 Bialystok, Poland*
[b] *School of Computer Science, University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada*

ARTICLE INFO

Article history:
Received 24 June 2013
Received in revised form 11 January 2014
Accepted 30 January 2014

Keywords:
Decision trees
Univariate tests
Underfitting
Gene expression data

ABSTRACT

Objective: The desirable property of tools used to investigate biological data is easy to understand models and predictive decisions. Decision trees are particularly promising in this regard due to their comprehensible nature that resembles the hierarchical process of human decision making. However, existing algorithms for learning decision trees have tendency to underfit gene expression data. The main aim of this work is to improve the performance and stability of decision trees with only a small increase in their complexity.
Methods: We propose a multi-test decision tree (MTDT); our main contribution is the application of several univariate tests in each non-terminal node of the decision tree. We also search for alternative, lower-ranked features in order to obtain more stable and reliable predictions.
Results: Experimental validation was performed on several real-life gene expression datasets. Comparison results with eight classifiers show that MTDT has a statistically significantly higher accuracy than popular decision tree classifiers, and it was highly competitive with ensemble learning algorithms. The proposed solution managed to outperform its baseline algorithm on 14 datasets by an average 6%. A study performed on one of the datasets showed that the discovered genes used in the MTDT classification model are supported by biological evidence in the literature.
Conclusion: This paper introduces a new type of decision tree which is more suitable for solving biological problems. MTDTs are relatively easy to analyze and much more powerful in modeling high dimensional microarray data than their popular counterparts.

1. Introduction

Decision trees [1,2] are one of the most popular classification techniques in data mining and machine learning. Due to their comprehensible nature, they are particularly useful when the aim of modeling is to understand the underlying processes of the environment. Decision trees are also useful when the data do not satisfy the rigorous assumptions required by more traditional methods [3]. Tree-based classifiers can be successfully applied to solving biological problems [4–6]. Popular techniques for microarray data involve decision tree ensembles like random forest [7] and boosted decision trees [8]. However, existing attempts to apply decision trees to classification using gene expression data showed that single tree algorithms are not sufficient for inducing competitive classifiers [9,10].

In this paper, we tackle the problem of improving the performance of decision trees on gene expression data, with the constraint of preserving simplicity of decision trees. Standard techniques for improving the performance of classification algorithms, e.g., ensemble methods, do not satisfy this constraint when applied to decision trees because resulting classifiers become complex and almost impossible to understand [11,12]. We propose a multi-test approach to decision trees in which several univariate tests can be used to create a single splitting rule in every non-terminal node of the classification tree. We also search for alternative, lower-ranked features in order to obtain more stable and reliable predictions.

1.1. *Gene expression data analysis*

Cells represent basic organizational units of all living organisms. Each cell contains instructions for the creation of proteins and the regulation of processes in a living body. This collection of instructions is contained in the DNA. Each protein has a corresponding gene which can be seen as a recipe for how to create a given protein. If the gene is expressed, a corresponding protein will be produced [13]. A significant step in genomic research was the ability to monitor the expression level of genes in living cells. Specifically,

[*] Corresponding author. Tel.: +48 85 746 91 63; fax: +48 85 746 90 57.
E-mail address: m.czajkowski@pb.edu.pl (M. Czajkowski).

0933-3657/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.artmed.2014.01.005

cDNA microarray and high-density oligonucleotide chips allow the expression level of thousands of genes to be monitored simultaneously [14]. The outcome of these diagnostic tests is known as gene expression (or microarray) data.

Microarray data allows for numerous analyses of living organisms. The application of a mathematical apparatus and computations tools is indispensable here, since gene expression observations are represented by high dimensional feature vectors. The important questions are what kind of outcomes can be expected and what kind of questions can be answered using these tools. The answer comes from two fundamental approaches to mathematical modeling, which are equally important in the case of gene expression data. Scientific modeling attempts to understand the true model that is behind the data generated according to that model. In the case of gene expression data, it is concerned with problems of causal relationships between, for example, genes, or genes and proteins. Technological modeling has different aims. Here, the purpose is to build a model from past data that would be good at predicting future data regardless of whether the model is close to reality or not [15]. Discriminant analysis is an example of this kind of modeling in a general sense. It has also been widely used in post-genome cancer research studies [16,17].

Gene expression data poses many research challenges, and is not limited to research areas that are concerned with living organisms. This kind of data is also extremely challenging for computational tools and mathematical modeling [18]. Each observation is described by a high dimensional feature vector with a number of features that reach into the thousands, but the number of observations is rarely higher than 100. Therefore, this kind of data requires new computational tools to extract significant and meaningful rules, and some feature selection should be taken into account. Providing a group of most relevant genes may significantly improve classification performance [19].

1.2. Decision trees

Decision trees (also known as classification trees) represent one of the main techniques for discriminant analysis in data mining and knowledge discovery. They predict the class membership (dependent variable) of an instance using its measurements of predictor variables.

The most popular algorithms for decision tree induction are based on top-down greedy search [20]. First, the test attribute (and the threshold in the case of continuous attributes) is decided for the root node. Instances are split through the tree from the root node to a leaf node, which provides classification of a given instance. At each non-terminal node through which the instance passes, one (or more) attribute of the instance is tested and the instance is moved down to the branch that corresponds to an outcome of the test. The process is recursively repeated for each branch. When to stop partitioning and create a leaf node is still one of the major problems in the area.

Classification trees have many advantages that make them applicable in various scenarios, particularly when the data does not satisfy the rigorous assumptions required by more traditional methods. In this paper, the following facts are significant:

- learning of decision trees is fast, even with huge data sets, due to greedy search;
- classification is very fast, flexible, and allows for straightforward approaches to the problem of missing values;
- decision trees are easy to understand and analyze, as they reflect a hierarchical way of human decision making. They are thus the opposite of the ‘black-box’ approaches where model parameters are not understandable.

This introduction applies to cases in which tests in internal nodes of trees are based on one attribute. There are also algorithms which apply multivariate tests [21,22] based mostly on linear combination splits. Decision trees that allow the testing of multiple features at the node are potentially smaller than those limited to single univariate splits. Additionally, when only one attribute at each node is tested, it may cause replication of specific subtrees in the decision tree [23]. In effect, some features may be tested more than once in the decision tree. However, trees with simple tests are still desirable because experts can understand them. This fact is explicitly emphasized in the related literature. Brodley and Utgoff [24] say: “*A small tree with simple tests is most appealing because a human can understand it. There is a tradeoff to consider in allowing multivariate tests: using only univariate tests may result in large trees that are difficult to understand, whereas the addition of multivariate tests may result in trees with fewer nodes, but the test nodes may be more difficult to understand*”. Our focus is therefore on univariate trees, since they are a ‘white-box’ technique, which makes them particularly interesting for scientific modeling. It is easy to find explanation for decisions of univariate classification trees.

1.3. Background and motivation

As stated in the previous section, decision trees with univariate splits are convenient. They are much easier to understand than trees with multivariate splits, and it is much easier to learn from the data. However, traditional algorithms, for example, C4.5 [25] or CART [26], fail to produce decision trees with high classification accuracy of gene expression data. Our previous work with various univariate decision tree algorithms showed that these algorithms produce considerably small trees that perfectly classify the training data but fail to classify unseen instances [10]. Only a small number of attributes is used in such trees, and their model complexity is low (high bias). Therefore, they underfit the training data [2]. Producing bigger trees using standard algorithms such as C4.5 does not solve the problem in the case of gene expression data because small trees often classify the training data perfectly [10]. This indicates that the issue of split complexity could be advocated here, since not much can be gained from bigger univariate decision trees with this kind of data. This line of research is pursued in our paper.

We are motivated by the fact that univariate decision tree induction represents a white-box approach and improvements of such algorithms have considerable potential for genomic research and scientific modeling of underlying processes. Thus, our goal is to improve the classification accuracy of decision trees and imply more informative analysis of microarray data in a way that will make them still easy to understand. Decision trees with multivariate splits or bagging/boosting methods often outperform existing univariate algorithms on gene expression data [9,27,28]. However, those approaches generate complex rules that from a medical point of view are more difficult to understand and analyze. Our goal is to increase the complexity of univariate decision trees to the extent that makes them easy to understand and more competitive in terms of classification accuracy. We believe that the use of individual univariate splits may cause the classifier to underfit the learning data, since it leads to trees that are not robust enough and do not take information about other most relevant attributes into account. Our novel technique uses several univariate tests in each internal node to avoid these problems. As multi-test nodes are based on univariate tests, trees learned with this approach will be much easier to analyze than trees with classical multivariate splits.

In this paragraph, we attempt to justify why our approach is suitable for gene expression data and why this may lead to high classification accuracy. Gene expression data is characterized by a very high ratio of features to observations, which poses serious