ELSEVIER

Computers in Biology and Medicine 37 (2007) 637–641

Computers in Biology and Medicine

www.intl.elsevierhealth.com/journals/cobm

A tree-based decision rule for identifying profile groups of cases without predefined classes: application in diffuse large B-cell lymphomas

Elias Zintzaras[a,*], Maria Bai[b], Christos Douligeris[c], Axel Kowald[d], Panayiotis Kanavaros[e]

[a]*Department of Biomathematics, University of Thessaly School of Medicine, Larissa, Greece*
[b]*Department of Pathology, University of Ioannina School of Medicine, Ioannina, Greece*
[c]*Department of Informatics, University of Piraeus, Piraeus, Greece*
[d]*Max Planck Institute for Molecular Genetics, Berlin, Germany*
[e]*Department of Anatomy-Histology-Embryology, University of Ioannina School of Medicine, Ioannina, Greece*

Received 28 July 2005; received in revised form 6 May 2006; accepted 5 June 2006

Abstract

In this paper, we examined the utility of a forward growing classification tree as a supplement to cluster analysis for deriving a decision rule for the identification of profile groups when the cases do not belong to predefined classes. The technique was applied for the identification of low and high proliferation profile groups of diffuse large B-cell lymphomas according to the immunohistochemical expression levels of proliferation proteins. In a forward growing classification tree method, the size of the tree is controlled by the improvement (threshold value) in the apparent misclassification rate after each split. The classes used in the tree were defined using k-means clustering. The decision rule consisted of the splitting points of the split variables used. The methodology was applied to the histology data from 79 cases of diffuse large B-cell lymphomas. Ten classes of individual cases were derived from k-means clustering. Then, a classification tree with a threshold of 2% was used to derive the decision rule. Branches at the left side of the tree consisted of individuals with a low proliferation profile and branches at the right side of the tree consisted of cases with a high proliferation profile. The classification tree, as a supplement method, not only identified but also provided decision rules for identifying profile groups. Finally, it also allowed for exploration of the data structure.

Keywords: Classification tree; Decision rule; Cluster analysis; B-cell lymphomas; Proteins; Proliferation

1. Introduction

In biomedical research, a commonly used multivariate method to identify groups of cases when the cases do not belong to predefined classes is cluster analysis. However, cluster analysis does not provide a decision rule to identify groups with certain profiles. A classification tree method in combination with cluster analysis may assist in the derivation of a rule for identifying profile groups and in the further exploration of the data structure.

For example, currently, the identification of low and high proliferation profile groups of cases of diffuse large B-cell lymphomas based on a set of immunohistochemical expression levels of proliferation-associated proteins (variables) taken from each individual case is carried out using cluster analysis [1]. Nevertheless, it is of great clinical importance to obtain a decision rule that identifies groups with certain profiles, i.e. high and low values of the associated variables, based on certain threshold values of the measured variables.

In this study, we propose a decision rule methodology to identify groups of cases with high and low profiles of measured variables for each case, based on the forward growing classification tree method [2], which is a variant of the method proposed by Breiman et al. [3]. In the forward growing method the size of the tree is controlled by the improvement in the apparent misclassification rate (AMR) after a split [2], and not by an upward pruning of the maximum tree [3]. In addition, the proposed technique can be useful for the exploration of the data structure. The classification tree method allows groups of individual cases to be determined and at the same time the

* Corresponding author. Tel./fax: +30 2140 565270.
E-mail address: zintza@med.uth.gr (E. Zintzaras).

0010-4825/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2006.06.001

importance and the diagnostic value of the measured variables to be precisely evaluated.

Since there were no predefined classes of the cases, which is a definite requirement for the application of the classification tree, the conventional technique of k-means clustering was first applied to the individual cases. The derived clusters were declared as classes of individuals, i.e. we derived a seed for the classification tree analysis. Then, the splitting variables and the corresponding splitting points were used to define the decision rule.

Classification trees have been used in various medical applications: for improving accuracy of diagnosis, for prognostic applications, for examining predictors of survival and for analyzing prospective epidemiologic studies [4].

The purpose of this paper is to examine the utility of the forward growing classification tree as a supplement to cluster analysis for the identification of profile groups. As an example, the proposed technique was used to identify the immunohistochemical expression levels of the proliferation-associated proteins Ki67, cyclin A and cyclin B1 used for the identification of groups with certain proliferation profiles [1,5,6]. The proliferation profile of diffuse large B-cell lymphomas as a distinct malignant tumor type [7] has not been extensively investigated so far by multivariate methods. The identification of distinct groups of diffuse large B-cell lymphomas with respect to the proliferation profile is important since an increased proliferation was reported to be associated with aggressive clinical behavior in these tumors [8,9].

2. Methods

2.1. Histological data

We used the forward growing classification tree method, as proposed by Zintzaras et al. [2], in combination with a conventional clustering method, namely the k-means, to derive the decision rule for classifying 79 cases of diffuse large B-cell lymphomas [7]. The classification was based on the immunohistochemical expression values of the proliferation-associated proteins (variables) Ki67, cyclin A and cyclin B1 [5].

2.2. Growing the classification tree

In the classification tree method, the individual cases are sorted for each measured variable (x_i). Then the variable with the splitting point which best discriminates between the classes that the cases belong, is chosen. Afterwards, the initial set of cases is split into two subsets and the two subsets are partitioned independently. The above process is repeated recursively. The algorithm can be presented as a tree structure. If t is the node representing the initial set of cases then the algorithm splits t into two subsets (sub-nodes) t_L and t_R in such a way that the sub-nodes are purer (a node is maximally pure when it contains cases from only one class) than the parent node t. Then a portion p_L of individual cases in t goes to t_L and a portion p_R goes to t_R. If the splitting variable is x_k ($k=1-3$) and the splitting point is s, then t_L contains all the individuals which have values of x_k less than s and t_R contains the remainder. The goodness of split is defined by the increase in purity $\mathrm{DI}(s)$ due to split s: $\mathrm{DI}(s) = I(t) - p_L I(t_L) - p_R I(t_R)$, where $I(t) = 1 - \Sigma_j p^2(j/t)$ and $p(j/t)$ is the proportion of individuals of class j in node t [3].

After each split, classes are assigned to the new nodes using the majority rule. The growth of the tree is determined by the improvement (threshold value) in the AMR (the proportion of misclassified individuals to their classes using resubstitution) after each split [2].

2.3. Decision rule

Based on the splitting variables and the corresponding splitting points, we can derive a decision rule for determining groups of cases according to the degree of proliferation of diffuse large B-cell lymphomas.

2.4. Class specification

Prior to the classification tree analysis, a clustering method [10], namely the k-means clustering, was applied to the data in order to identify clusters of individuals. K-means clustering starts with k random clusters and then moves cases between those clusters to: (i) minimize variability within clusters and (ii) maximize variability between clusters. Then these clusters were declared as classes, i.e. a seed, applicable for the classification tree analysis.

The classification tree was applied to these declared classes for various thresholds of improvement in AMR. Then, the split point values at the branch nodes were recorded providing a decision tool for characterizing groups of diffuse large B-cell lymphomas with low and high proliferation profiles. The terminal nodes consisted of groups of individual cases with a certain proliferation profile of diffuse large B-cell lymphomas.

2.5. Computer implementation

The classification tree algorithm was implemented as a C program with a Windows interface. The program generates a graphical output (the resulting tree), which is displayed on the screen or exported as a postscript file. The software is available by the corresponding author upon request.

3. Results

The k-means clustering for $k = 10$ produced 10 distinct clusters of unequal size and each cluster was defined as a class of individual cases (Fig. 1). Then, based on these classes the classification tree method was applied at a threshold value of 1%. The tree produced 15 terminal nodes with a misclassification rate $R = 0.013$, i.e. only one case was misclassified (Fig. 1). Each terminal node corresponded to a group of individual cases characterized by a combination of protein expressions. In this tree there are terminal nodes with one or two cases (nodes: 16, 11, 28, 27, 20) which are very close to large nodes. Therefore, we may produce a tree with a higher threshold value,