STATISTICS IN MEDICINE
Statist. Med. 2003; **22**:1365–1381 (DOI: 10.1002/sim.1501)

Multiple additive regression trees with application in epidemiology

Jerome H. Friedman[1,*,†] and Jacqueline J. Meulman[2,‡]

[1] *Department of Statistics, Stanford University, U.S.A.*
[2] *Data Theory Group, Leiden University, The Netherlands*

SUMMARY

Predicting future outcomes based on knowledge obtained from past observational data is a common application in a wide variety of areas of scientific research. In the present paper, prediction will be focused on various grades of cervical preneoplasia and neoplasia. Statistical tools used for prediction should of course possess predictive accuracy, and preferably meet secondary requirements such as speed, ease of use, and interpretability of the resulting predictive model. A new automated procedure based on an extension (called 'boosting') of regression and classification tree (CART) models is described. The resulting tool is a fast 'off-the-shelf' procedure for classification and regression that is competitive in accuracy with more customized approaches, while being fairly automatic to use (little tuning), and highly robust especially when applied to less than clean data. Additional tools are presented for interpreting and visualizing the results of such multiple additive regression tree (MART) models. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: predictive learning; regression trees; boosting; data mining; MART; cervical cancer

1. PREDICTIVE LEARNING

In the predictive learning problem one has a system consisting of a random 'output' or 'response' variable y and a set of random 'input', 'explanatory' or 'predictor' variables $\mathbf{x}=\{x_1,\ldots,x_n\}$. In the application that will be given below, the predictor variables are a collection of qualitative and quantitative features of cells taken from the cervix, and the response variable gives the histological diagnosis, determined by a biopsy, for each of the cases into five categories, ranging from mild dysplasia to invasive squamous cell carcinoma. The predictive learning problem defines a 'training' sample, $\{y_i,\mathbf{x}_i\}_1^N$ of known $(y,\mathbf{x})$-values, where $(y_i,\mathbf{x}_i)$ links the features of the ith case with the ith value of the diagnosis variable. The goal is then to use these data to obtain an estimate of the function that maps the values of the predictor variables $\mathbf{x}$, here the cell features, into the values of the response variable y,

* Correspondence to: Jerome H. Friedman, Department of Statistics, Stanford University, U.S.A.
† E-mail: jhf@stanford.edu
‡ E-mail: meulman@fsw.leidenuniv.nl

the diagnosis. This function can then be used to predict the diagnosis for future observations, where only $\mathbf{x}$ is known.

In general, we wish to obtain an estimate $\hat{F}(\mathbf{x})$, of the function $F^*(\mathbf{x})$ mapping $\mathbf{x}$ to y that minimizes the expected value of some specified loss function $L(y,F(\mathbf{x}))$ over the joint distribution of all $(y,\mathbf{x})$-values

$$\hat{F}(\mathbf{x}) \cong F^*(\mathbf{x}) = \arg\min_{F(\mathbf{x})} E_{y\mathbf{x}} L(y,F(\mathbf{x})) \tag{1}$$

In this paper, we shall focus on a particular approach to prediction, which is by the use of regression and classification tree (CART) models [1]. Formally, a *regression* tree is defined by a continuous numerical response variable, and a *classification* tree by a categorical, usually unordered response variable. In the multivariate analysis context, a regression tree can be compared to multiple regression, and a classification tree to discriminant analysis. In this paper, we will loosely talk of regression trees for both continuous and categorical response variables. A regression tree model $T(\mathbf{x};\{R_j\}_1^J)$ partitions the $\mathbf{x}$-space into J disjoint regions $\{R_j\}_{j=1}^J$ and predicts a separate constant value in each one

$$\mathbf{x} \in R_j \Rightarrow T(\mathbf{x};\{R_j\}_1^J) = \bar{y}_j$$

or equivalently

$$T(\mathbf{x};\{R_j\}_1^J) = \sum_{j=1}^{J} \bar{y}_j I(\mathbf{x} \in R_j) \tag{2}$$

Here $\bar{y}_j = \text{mean}_{\mathbf{x}_i \in R_j}(y_i)$ is the mean of the response y in each region R_j, so a tree predicts a constant value $\bar{y}_j$ within each region R_j. Regression trees are induced by top–down recursive splitting based on a least-squares fitting criterion. The parameters of this particular model are the regions $\{R_j\}_1^J$ of the partition, which in turn are defined by the identities of the predictor variables used for splitting and their corresponding split points.

2. USING TREES FOR DATA MINING

Regression trees have the following desirable properties for data mining, or the analysis of complicated data sets in general. The input variables can be mixtures of all types; numeric, ordinal, binary and categorical variables are all handled with equal ease. Trees directly handle missing values in a very elegant fashion based on 'surrogate' splitting [1]. No imputation schemes need to be employed. The predictive model is invariant under strictly monotone transformations of the input variables; performing a transformation on any variable $x_j \leftarrow g_j(x_j)$, where $g_j(x_j)$ is any monotone function, produces the same model. Thus, there is no issue of trying to find 'good' transformations beforehand. Trees are immune to the effects of extreme outliers among the predictor x-variables, thereby easing the data cleaning burden. If least-absolute-deviation loss, one of the loss functions implemented, is used then there is also complete immunity from outliers in the output response y, thereby providing total immunity to the effects of outliers. Unlike near-neighbour and kernel based methods, and support vector machines, trees are invariant to changing the relative scales of the predictor variables, so there is no need to experiment with various metrics.

Regression trees perform automatic variable subset selection. In many situations where there are a large number of predictor variables, only a few of them are actually relevant to prediction. The performance of many methods (neural networks, near-neighbour, kernel methods, support vector machines) can degrade dramatically when extra irrelevant predictors are included. The performance of tree based models are highly resistant to the inclusion of a large number of irrelevant variables, thereby making feature selection much less of an issue.

With all these advantageous properties, it is no surprise that fitting a regression or classification tree has become a very popular tool for predictive learning in data mining. However, trees have one major disadvantage, *inaccuracy*. Although they are sometimes competitive, usually tree based models do not achieve accuracy close to the best possible in any given application. Fortunately, a remedy became available through so-called 'boosting', and boosting tree based models almost always dramatically increases their accuracy. This is illustrated with the example in Section 6 as well as in Freund and Shapire [2], Friedman *et al.* [3], and Friedman [4]. As we shall see in the sequel, MART models combine boosting with regression trees as their primary ingredient, so they inherit nearly all of the advantages of tree based modelling, while overcoming inaccuracy, their primary disadvantage.

3. GRADIENT BOOSTING

Technically, boosting takes the estimate $\hat{F}(\mathbf{x})$ to be an 'additive' expansion of the form

$$\hat{F}(\mathbf{x}) = \sum_{m=0}^{M} \beta_m h(\mathbf{x}; \mathbf{a}_m) \tag{3}$$

where the functions $h(\mathbf{x}; \mathbf{a})$ ('base learner') are usually chosen to be simple functions of $\mathbf{x}$ with parameters $\mathbf{a} = \{a_1, a_2, \ldots\}$. The expansion coefficients $\{\beta_m\}_0^M$ and the parameters $\{\mathbf{a}_m\}_0^M$ are jointly fit to the training data in a forward 'stagewise' manner. One starts with an initial guess $F_0(\mathbf{x})$, and then for $m = 1, 2, \ldots, M$

$$(\beta_m, \mathbf{a}_m) = \arg\min_{\beta, \mathbf{a}} \sum_{i=1}^{N} L(y_i, F_{m-1}(\mathbf{x}_i) + \beta h(\mathbf{x}_i; \mathbf{a})) \tag{4}$$

and

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \beta_m h(\mathbf{x}; \mathbf{a}_m) \tag{5}$$

Here $L(y, F)$ is the loss criterion (1) chosen to define lack-of-fit.

Gradient boosting [4] approximately solves (4) for arbitrary (differentiable) loss functions $L(y, F(\mathbf{x}))$ with a two-step procedure. First, the function $h(\mathbf{x}; \mathbf{a})$ is fit by least-squares

$$\mathbf{a}_m = \arg\min_{\mathbf{a}, \rho} \sum_{i=1}^{N} [\tilde{y}_{im} - \rho h(\mathbf{x}_i; \mathbf{a})]^2 \tag{6}$$

to the current 'pseudo'-residuals

$$\tilde{y}_{im} = -\left[\frac{\partial L(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x}) = F_{m-1}(\mathbf{x})} \tag{7}$$

Then, given $h(\mathbf{x};\mathbf{a}_m)$, the optimal value of the coefficient β_m is determined by

$$\beta_m = \arg\min_{\beta} \sum_{i=1}^{N} L(y_i, F_{m-1}(\mathbf{x}_i) + \beta h(\mathbf{x}_i;\mathbf{a}_m)) \tag{8}$$

This strategy replaces a potentially difficult *function* optimization problem (4) by one based on least-squares (6), followed by a *single parameter* optimization (8) based on the general loss criterion L. Often used loss criteria are $L(y,F)=|y-F|$ (least-absolute-deviation loss) and $L(y,F)=(y-F)^2$ (squared-error loss) for regression ($y\in R$), and $I(y\neq F)$ for K-class classification $y\in\{c_1,\ldots,c_K\}$. The 'target' function $F^*(\mathbf{x})$ (1) is by definition the best predictor of $y\,|\,\mathbf{x}$ under the defined loss L. The data based estimate $\hat{F}(\mathbf{x})$ is used as a surrogate to predict y-values for future observations where only the values of $\mathbf{x}$ are known.

4. AN ALGORITHM FOR BOOSTING TREES

Multiple additive regression trees (MART) specializes the gradient boosting approach to the case where the base learner $h(\mathbf{x};\mathbf{a})$ is a J-terminal node regression or classification tree (2). MART employs an iterative algorithm, where now at *each* iteration m, a new regression tree $T_m\ (\mathbf{x};\{R_{jm}\}_1^J)$ is built to partition the $\mathbf{x}$-space into J disjoint regions $\{R_{jm}\}_{j=1}^J$ and predict a separate constant value in each one

$$T_m(\mathbf{x};\{R_{jm}\}_1^J) = \sum_{j=1}^{J} \bar{y}_{jm} I(\mathbf{x}\in R_{jm}) \tag{9}$$

Here, $\bar{y}_{jm} = \text{mean}_{\mathbf{x}_i\in R_{jm}}(\tilde{y}_{im})$ is the mean of (7) in each region R_{jm} induced at the mth iteration. The parameters of the base learner are the regions $\{R_{jm}\}_1^J$ of the partition, defined by that tree. With regression trees, (8) can be solved separately within each region R_{jm} defined by the corresponding terminal node j of the mth tree. Because the tree (9) predicts a constant value within each region R_{jm}, the solution to (8) reduces to a simple 'location' estimate based on the criterion L

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{\mathbf{x}_i\in R_{jm}} L(y_i, F_{m-1}(\mathbf{x}_i) + \gamma)$$

The current approximation $F_{m-1}(\mathbf{x})$ is then separately updated in each corresponding region

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu\gamma_{jm} I(\mathbf{x}\in R_{jm}) \tag{10}$$

The 'shrinkage' parameter $0<\nu\leqslant 1$ controls the learning rate of the procedure. Empirically [4], it was found that small values ($\nu\leqslant 0.1$) lead to much better results in terms of prediction error on future data.

This leads to the following MART algorithm for generalized boosting of regression trees:

1. $F_0(\mathbf{x}) = \arg\min_{\gamma}\sum_{i=1}^{N} L(y_i,\gamma)$
2. For $m=1$ to M do:
3. $\tilde{y}_{im} = -\left[\frac{\partial L(y_i,F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})},\ i=1,N$
4. $\{R_{jm}\}_1^J = J-\text{terminal node tree}(\{\tilde{y}_{im},\mathbf{x}_i\}_1^N)$

5. $\gamma_{jm} = \arg\min_{\gamma} \sum_{\mathbf{x}_i \in R_{jm}} L(y_i, F_{m-1}(\mathbf{x}_i) + \gamma)$
6. $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu\gamma_{jm} I(\mathbf{x} \in R_{jm})$
7. endFor.

Friedman [4] presented specific algorithms based on this template for several loss criteria including: least-squares, $L(y,F) = (y-F)^2$; least-absolute-deviation, $L(y,F) = |y-F|$; Huber-M, $L(y,F) = (y-F)^2 I(|y-F| \leqslant \delta) + 2\delta(|y-F| - \delta/2) I(|y-F| > \delta)$, and for classification, K-class multinomial negative log-likelihood.

For K-class classification, MART builds a separate model $\hat{F}_k(\mathbf{x})$ for the log-odds of each class relative to the other classes. Each of these models is a sum of trees

$$\hat{F}_k(\mathbf{x}) = \sum_{m=1}^{M} T_{km}(\mathbf{x}) \tag{11}$$

induced via algorithm 1. The models $\{\hat{F}_k(\mathbf{x})\}_{m=1}^{M}$ are coupled in that they are constrained so that at each $\mathbf{x}$ the corresponding probabilities are in the range $[0,1]$ and sum to one.

5. INTERPRETATION

Unfortunately, boosted trees forfeit one of the most important advantages of single tree models, that is, high interpretability. The information contained in the entire single tree model can be completely represented by a two-dimensional graphic (binary tree) that is easily interpreted. MART models, on the other hand, are based on linear combinations of many trees, and therefore lose this important feature. The results of a MART analysis must therefore be interpreted in a different way. Among the most important ingredients in any interpretation are identifying which variables are important for prediction, and understanding their joint effect on the response.

5.1. Relative importance of predictor variables

As noted above, all of the input predictor variables are seldom equally relevant for prediction. Often only a few of them have substantial influence on the response; the vast majority are irrelevant and could just as well have not been measured. It is often useful to learn the relative importance or contribution of each input variable in predicting the response.

For a single tree T, Breiman *et al.* [1] proposed a measure of (squared) relevance $I_j^2(T)$ for each predictor variable x_j, based on the number of times that variable was selected for splitting in the tree weighted by the squared improvement to the model as a result of each of those splits. This importance measure is easily generalized to additive tree expansions (3); it is simply averaged over the trees

$$I_j^2 = \frac{1}{M} \sum_{m=1}^{M} I_j^2(T_m) \tag{12}$$

Owing to the stabilizing effect of averaging, this measure (12) turns out to be more reliable than is its counterpart for a single tree. Since these measures are relative, it is customary to assign the largest a value of 100 and then scale the others accordingly.

For K-class classification MART builds a separate model (11) for each class. In this case (12) can be generalized to

$$I_{jk}^2 = \frac{1}{M}\sum_{m=1}^{M} I_j^2(T_{km}) \tag{13}$$

Here I_{jk} is the relevance of x_j in separating the class k observations from the other classes. The overall relevance of x_j is obtained by averaging over all of the classes

$$I_j^2 = \frac{1}{K}\sum_{k=1}^{K} I_{jk}^2 \tag{14}$$

The individual I_{jk} however can themselves be quite useful. One can summarize the $n \times K$ matrix of these relevance values in a variety of ways. Individual columns $I_{\cdot k}$ give the relative variable importances in separating class k. The individual rows $I_{j\cdot}$ reveal the influence of x_j in separating the respective classes. One can average the matrix elements (13) over chosen subsets of classes to determine variable relevance for that subset. Similarly, one can average over subsets of variables to obtain an idea of which classes the chosen variable subset is most influential in separating.

5.2. *Partial dependence plots*

After the most relevant variables have been identified, the next step is to attempt to understand the nature of the dependence of the approximation $\hat{F}(\mathbf{x})$ on their joint values. Visualization is one of the most powerful of such interpretational tools. Graphical representation of the value of $\hat{F}(\mathbf{x})$ as a function of its arguments provides a comprehensive summary of its dependence on the joint values of the input variables.

Unfortunately, such visualization is limited to low dimensional arguments. For more than two or three relevant variables, viewing functions of the corresponding higher dimensional arguments is more difficult. A useful alternative can sometimes be to view a collection of plots, each one of which shows the partial dependence of the approximation $\hat{F}(\mathbf{x})$ on a selected small subset of the input variables. Although such a collection can seldom provide a comprehensive depiction of the approximation, it can often produce helpful clues, especially when $\hat{F}(\mathbf{x})$ is dominated by low order interactions.

Consider a subset $\mathbf{z}_l$ of size $l < n$ of the input predictor variables $\mathbf{x}$

$$\mathbf{z}_l = \{z_{1l}, z_{2l}, \ldots, z_{ll}\} \subset \{x_1, x_2, \ldots, x_n\}$$

and let $\mathbf{z}_{\setminus l}$ be the complement set $\mathbf{z}_{\setminus l} \cup \mathbf{z}_l = \mathbf{x}$. The approximation $\hat{F}(\mathbf{x})$ will in principle depend on all of the input variables $\hat{F}(\mathbf{x}) = \hat{F}(\mathbf{z}_l, \mathbf{z}_{\setminus l})$. If one conditions on a specific set of joint values for $\mathbf{z}_{\setminus l}$, then $\hat{F}(\mathbf{x})$ can be regarded as a function of $\mathbf{z}_l$ conditioned on the chosen values for $\mathbf{z}_{\setminus l}$

$$\hat{F}_{\mathbf{z}_{\setminus l}}(\mathbf{z}_l) = \hat{F}(\mathbf{z}_l | \mathbf{z}_{\setminus l}) \tag{15}$$

The form of this function (15) will generally depend on the values chosen for $\mathbf{z}_{\setminus l}$. One can view (15) as a function of $\mathbf{z}_l$ characterized by 'parameters' $\mathbf{z}_{\setminus l}$.

If the form of (15) for the chosen subset $\mathbf{z}_l$ does not vary greatly for different values of $\mathbf{z}_{\setminus l}$, then the average or 'partial' dependence of $\hat{F}(\mathbf{x})$ on $\mathbf{z}_l$

$$\bar{F}_l(\mathbf{z}_l) = E_{z_{\setminus l}} \hat{F}_{z_{\setminus l}}(\mathbf{z}_l) \tag{16}$$

can serve as a useful description of the 'effect' of the chosen subset on $\hat{F}(\mathbf{x})$, where the expected value is over the joint distribution of $\mathbf{z}_{\setminus l}$ values. An estimate of $\bar{F}_l(\mathbf{z}_l)$ (16) can be rapidly computed for tree based models [4].

It is important to note that partial dependence functions defined in (16) represent the effect of the variable subset $\mathbf{z}_l$ on $\hat{F}(\mathbf{x})$ after accounting for (eliminating) the (average) effects of the other variables $\mathbf{z}_{\setminus l}$ on $\hat{F}(\mathbf{x})$. They are *not* the effect of $\mathbf{z}_l$ on $\hat{F}(\mathbf{x})$ *ignoring* the effects of $\mathbf{z}_{\setminus l}$. An analogy can be made with ordinary simple linear regression. The coefficient obtained by regressing y on each x_j separately measures the effect of x_j on y *ignoring* the other predictor variables. On the other hand, the coefficient of x_j obtained in a multiple regression on all of the predictors measures its effect after *accounting* for the effects of the other variables. Although the effects here are functions (16) rather than simply coefficients, the concept is the same. Since their locations are arbitrary, all partial dependence functions are centred to have zero mean over the data distribution.

For K-class classification, there are K separate models (11), one for each class. Each one is estimating the log-odds of its class. Thus each $F_k(\mathbf{x})$ is a monotone increasing function of its respective probability on a logarithmic scale. Partial dependence plots of each respective $\hat{F}_k(\mathbf{x})$ (11) on its most relevant predictors (13) can help reveal how the odds of realizing that class depend on the respective input variables.

6. APPLICATION TO THE HISTOLOGICAL DIAGNOSIS OF CERVICAL (PRE)NEOPLASIA

The data used in this application of MART were collected at the Leiden Cytology and Pathology Laboratory, and concern characteristics of cells obtained from patients with various grades of cervical preneoplasia and neoplasia. To obtain the samples, taken from the ectocervix as well as the endocervix, special sampling and preparation techniques were used [5]. The correct histological diagnosis was known by a subsequently taken biopsy. Previous analyses [6] concerned a subset of the data, containing, according to the histological diagnosis, 50 cases with mild dysplasia (histological group 1), 50 cases with moderate dysplasia (histological group 2), 50 cases with severe dysplasia (histological group 3), and 50 cases with carcinoma in situ (histological group 4). The number of cases with invasive squamous cell carcinoma (histological group 5), rather scarce in the Dutch population due to intense screening activities, could be enlarged to 42 by the co-operation of physicians in Indonesia, where cervical carcinoma is very frequent. For each of the 242 cases, seven qualitative features of the cells were determined. The features were rated by a pathologist on a scale ranging from 1 (normal) to 4 (very abnormal); so these seven variables are ordered categorical. The features under consideration are *nuclear shape*, *nuclear irregularity*, *chromatin pattern*, *chromatin*

distribution, *nucleolar irregularity*, *nucleus/nucleolus ratio and nucleus/cytoplasm ratio*. In addition, four quantitative features of each sample were established: *number of abnormal cells per fragment* (mean values), *total number of abnormal cells*, *number of mitoses*, and *number of nucleoli* (mean values).

Meulman *et al.* [6] inspected the mean values on each variable for each of the five groups in the histological diagnosis (the response variable). It turned out that apart from minor deviations, the group means were monotonically increasing with the severity of the grade of (pre)neoplasia, except for the feature *nucleus/cytoplasm ratio*. The mean values for the latter variable for each histological group were 2.34, 2.96, 3.22, 3.54 and 3.05, thus the mean value for group 5 (invasive squamous cell carcinoma) is smaller than the mean values for group 3 and group 4. Therefore, the variable *nucleus/cytoplasm ratio* was treated as categorical in the MART analysis, while all other variables were treated as numerical. Meulman *et al.* [6] analysed the data with canonical discriminant analysis (with two optimal linear combinations of the predictors), and obtained an average error rate of 0.37 for the test sets; the error rates in the five groups separately were 0.22, 0.34, 0.46, 0.48 and 0.33. A standard CART analysis (Breiman *et al.* [1], as implemented by Steinberg and Colla [7]) gave a tree with 10 regions (terminal nodes), and with an overall error rate of 0.42.

The MART analysis was repeated 30 times with different random samples in the test set to obtain a stable estimate for the error rate. The average error rate was 0.30 (with standard deviation 0.052). The tuning parameters (see Friedman [4]) were set at the default values, except for learn.rate = 0.10 (ν in (10)), sample.frac = 0.50 (because of the relatively small number of observations), and the number of regions in each tree was set to tree.size = 4. The results of one of these MART analyses, with error rate 0.292, will be discussed below in some detail (a parallel analysis with tree.size = 2, a model with no interactions among predictor variables, had error rate 0.333; tree.size = 3 gave 0.313, tree.size = 5 gave 0.333, and tree.size = 6 gave 0.333, so it was decided to choose a model with multiple additive trees of size 4).

In Figure 1 the error rates are displayed, from the total error rate at the top left (from which it is seen that class 1, mild dysplasia, has the largest error rate, while class 5, invasive squamous cell carcinoma, has the smallest error rate), to the separate error rates per class, with the error rate of class 5 at the bottom right. The plots per class also show details of the misclassification, which makes clear that in most cases, misclassification occurs into adjacent classes; for example, misclassification for class 4 (carcinoma *in situ*) occurs in equal proportions into class 3 (severe dysplasia) and class 5 (invasive squamous cell carcinoma).

6.1. Relative importance and contributions of predictor variables

Figure 2 displays the relative importance of the predictor variables, for all classes jointly (14) at the top left, and next for each of the classes separately (13). There is a clear differential effect. For example, variable 8 (*nucleus/cytoplasm ratio*) is by far the most important variable for separating class 1, while it is less important for class 5. Variable 7 (*nucleus/nucleolus ratio*) is the most important variable for class 5, while being (relatively) unimportant for the classes 2 and 3. The qualitative variables 1 to 4 (*nuclear shape*, *nuclear irregularity*, *chromatin pattern* and *chromatin distribution*) are much more important for the dysplasia classes than for the carcinoma classes.

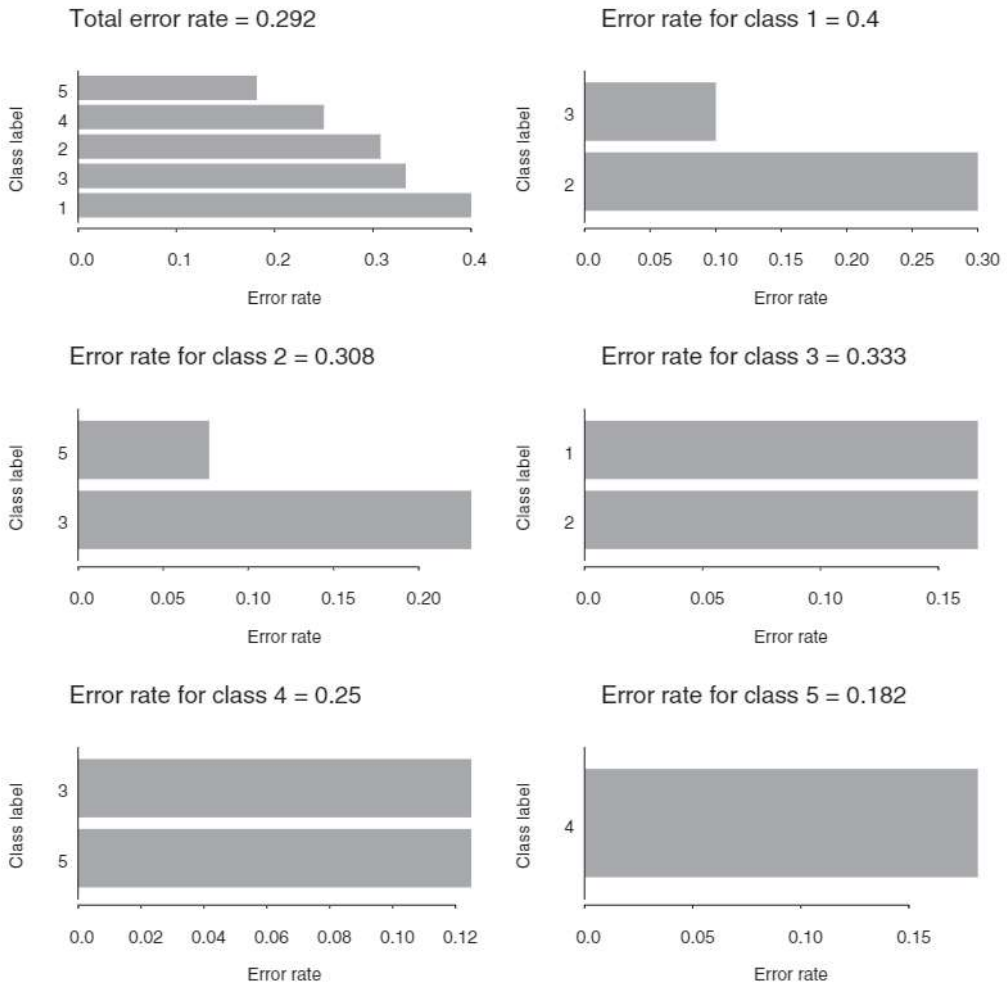

Figure 1. Total error rate, and error rates for each of the five histological classes of cervical (pre)neoplasia: 1 = mild dysplasia; 2 = moderate dysplasia; 3 = severe dysplasia; 4 = carcinoma *in situ*; 5 = invasive squamous cell carcinoma. The upper left bar plot compares error rates over the classes. The bar plots for each individual class compare the misclassification rates for that class into the other classes.

The latter becomes even more obvious from the 'contribution plots' for the variables 1, 2, 3 and 4 in Figure 3. The contribution plots answer the question, given a predictor, to which classes it contributes most with respect to their separation. Variable 1 (*nuclear shape*) contributes only to the separation of class 1, while variable 7, variable 11 (*chromatin pattern*), and variable 6 (*nucleolar irregularity*) do not contribute to class 1 at all. In contrast, the variables 5, 6 and 7 contribute most to separating class 5. Variable 8 (*chromatin pattern*), variable 9 (*number of abnormal cells per fragment*) and variable 10 (*total number of abnormal cells*) contribute to the separation of all classes.

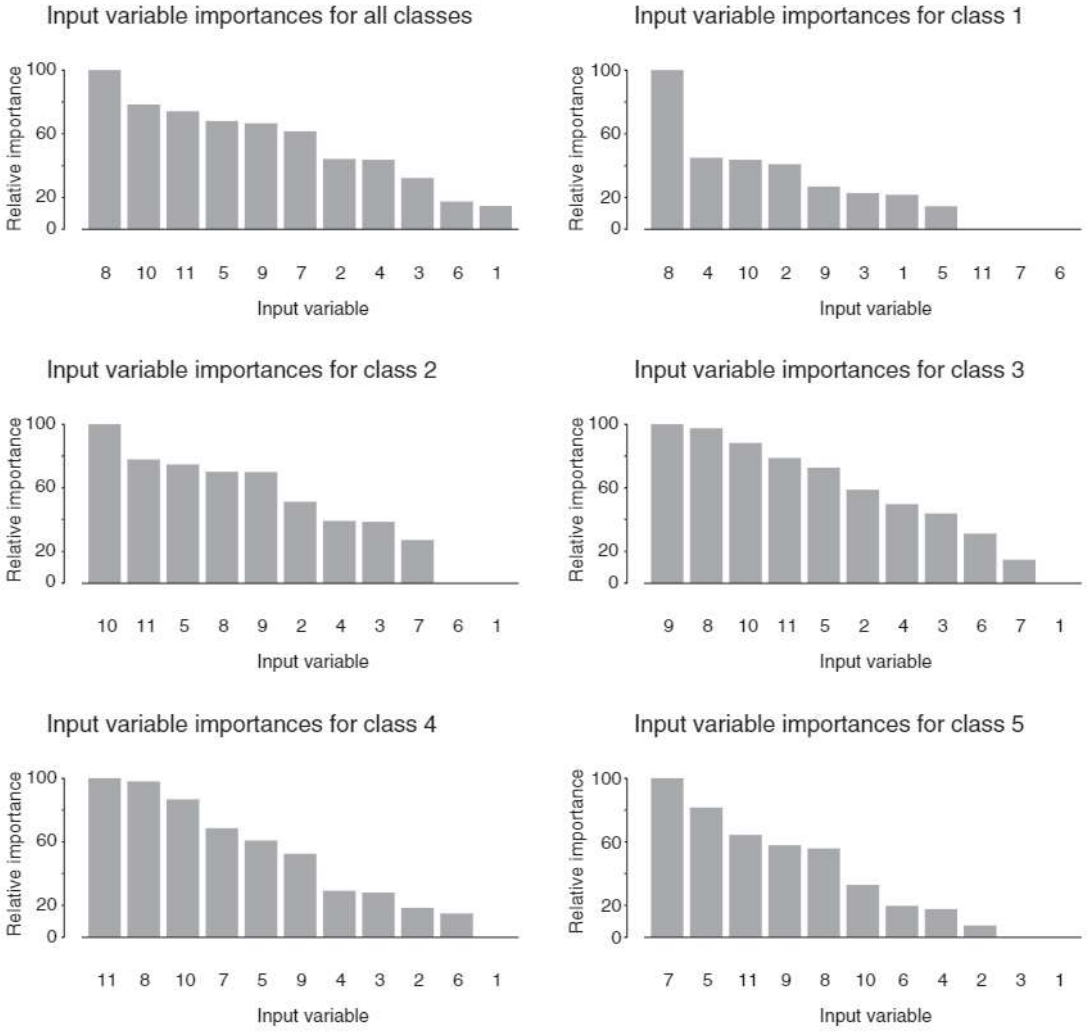

Figure 2. Ordering of the 11 predictors according to overall importance, and the importance of the predictors in separating each of the five given classes from the other four classes. Variable 8 = nucleus/cytoplasm ratio, 10 = total number of abnormal cells, 11 = number of mitoses, 5 = number of nucleoli, 9 = number of abnormal cells per fragment, 7 = nucleus/nucleolus ratio, 2 = nuclear irregularity, 4 = chromatin distribution, 3 = chromatin pattern, 6 = nucleolar irregularity, 1 = nuclear shape. Class 1 = mild dysplasia, 2 = moderate dysplasia, 3 = severe dysplasia, 4 = carcinoma *in situ*, 5 = invasive squamous cell carcinoma.

6.2. *Partial dependence plots*

Figure 4 displays the partial dependence of the response on variable 3 (left) and variable 7 (right) for each of the five classes. It is clear that there is a differential effect between the two different variables, and within the two variables for the different classes. The partial

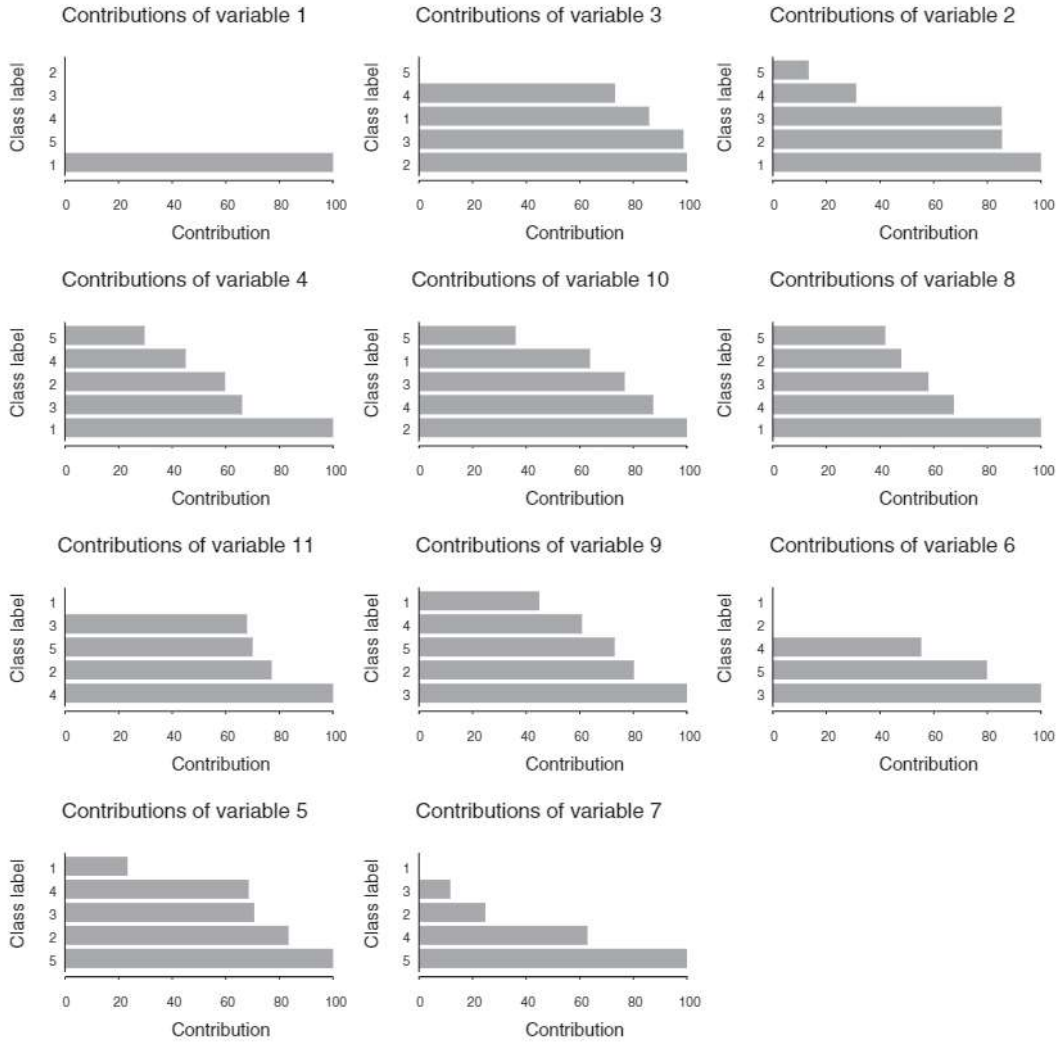

Figure 3. Different orderings of the five classes according to the contribution of each predictor variable to their separation. The 11 plots are ordered according to the contribution of each predictor to the separation of class 5. Variable 1 = nuclear shape, 3 = chromatin pattern, 2 = nuclear irregularity, 4 = chromatin distribution, 10 = total number of abnormal cells, 8 = nucleus/cytoplasm ratio, 11 = number of mitoses, 9 = number of abnormal cells per fragment, 6 = nucleolar irregularity, 5 = number of nucleoli, 7 = nucleus/nucleolus ratio.

dependence functions are step functions due to the discrete nature of these variables. Starting with variable 3 (*chromatin pattern*), we first notice that it has only three categories, since category 1 (normal) does not occur in the total sample. For class 1 (upper plot at the left), we see a sharp decrease going from rating 2 to the ratings 3 and 4 of *chromatin pattern*, and a more gradual decrease for the probability to be in class 2. For the classes 3 and 4 the

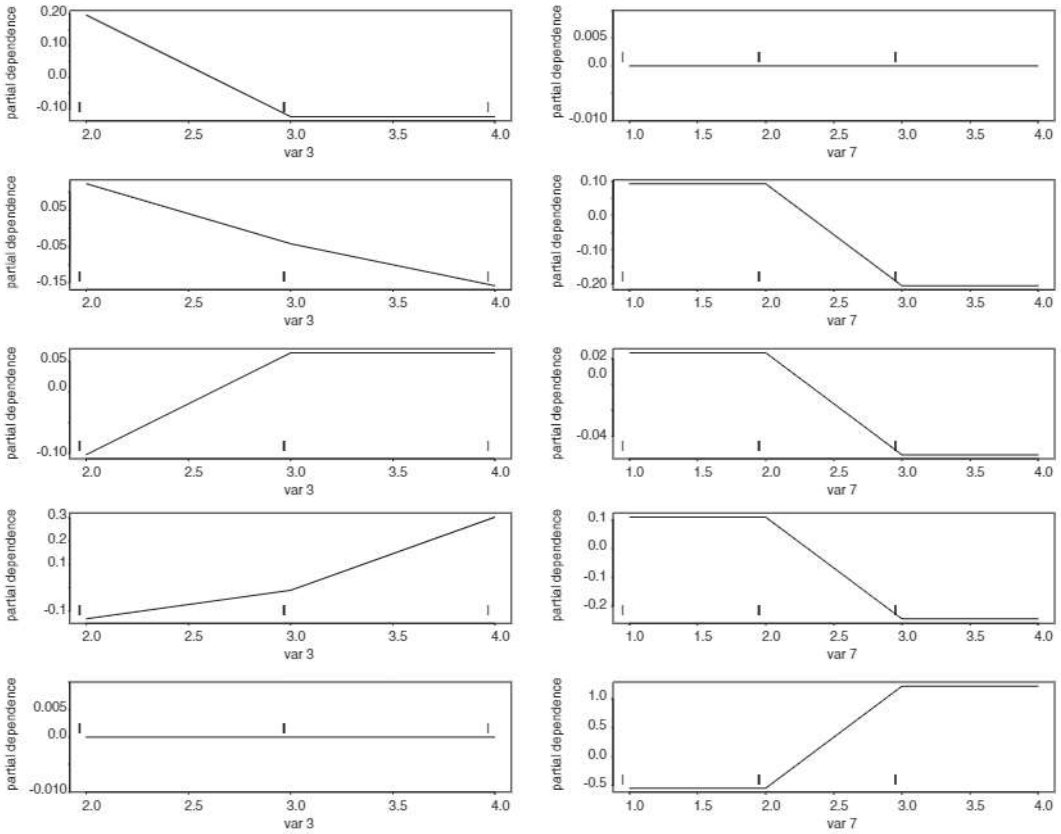

Figure 4. Partial dependence plots for the variables 3 (left panels) and 7 (right panels) for the classes 1 to 5 (from top to bottom). Variable 3 = chromatin pattern, 7 = nucleus/nucleolus ratio.

picture is reversed: going from rating 2 to 4, we see a sharp increase and a gradual increase, respectively. After accounting for the other variables, there is no influence of *chromatin pattern* on the prediction into class 5.

Inspecting variable 7 next, there is no partial dependence on *nucleus/nucleolus ratio* for the prediction into class 1. The step functions for the classes 2, 3 and 4 decrease going from rating 1 and 2, to 3 and 4, but for class 5, the partial dependence is increasing. Note that the scales on the vertical axes of these plots are different, and that the plot for class 5 signifies a much larger effect than the plot for class 3.

In Figure 5, the partial dependence plots are given for the four most important variables for class 2; these are the variables 10, 11, 5 and 8. Starting at the plot at the top left, we see that variable 10 (*total number of abnormal cells*) has a very skew distribution for class 2: while the maximum over all 242 cases is 1800, there are no counts higher than 385. There is a high peak in the partial dependence function around the value 75, while the function is completely flat after the value 250. For variable 11 (*number of mitosis*) we see a sharp decrease from the value 0 to the value 2, and the function becomes flat after that point. The function for

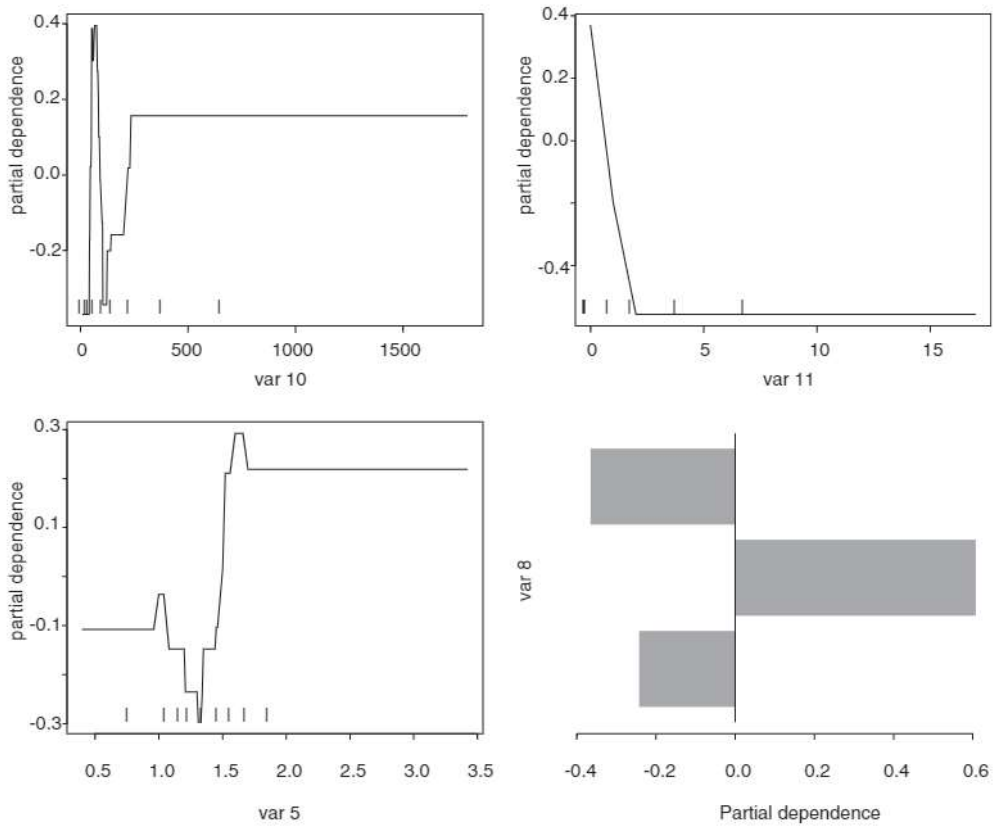

Figure 5. Partial dependence plots for the four most important variables for class 2 (moderate dysplasia). Variable 10 = total number of abnormal cells, 11 = number of mitoses; 5 = number of nucleoli, 8 = nucleus/cytoplasm ratio. The hash marks at the base of each plot locate the deciles of the data distribution on each variable Since variable 8 is categorical with three levels its partial dependence is displayed as a bar plot.

variable 5 starts flat, and increases to a peak in between the values 1.3 and 1.65. After that point, the function becomes constant again. The function for variable 8 (*nucleus/cytoplasm ratio*) is presented as a bar plot because the predictor has been treated as categorical. The probability to be in class 2 is clearly largest for the value 3 (and much less for both the values 2 and 4).

Figure 6 scrutinizes class 2 again, and looks at the different functions for variable 10 (*total number of abnormal cells*) according to the three different levels of variable 8 (*nucleus/cytoplasm ratio*). To obtain functions on a comparable scale on the horizontal axis (indicating the number of abnormal cells), the extreme values were cut off, using the 90 per cent quantiles. The plot for category 3 of *nucleus/cytoplasm ratio* is clearly different from the plots for category 2 (upper left panel) and category 4 (lower left panel), and can be compared to the overall plot for class 2 in the upper left panel of Figure 5.

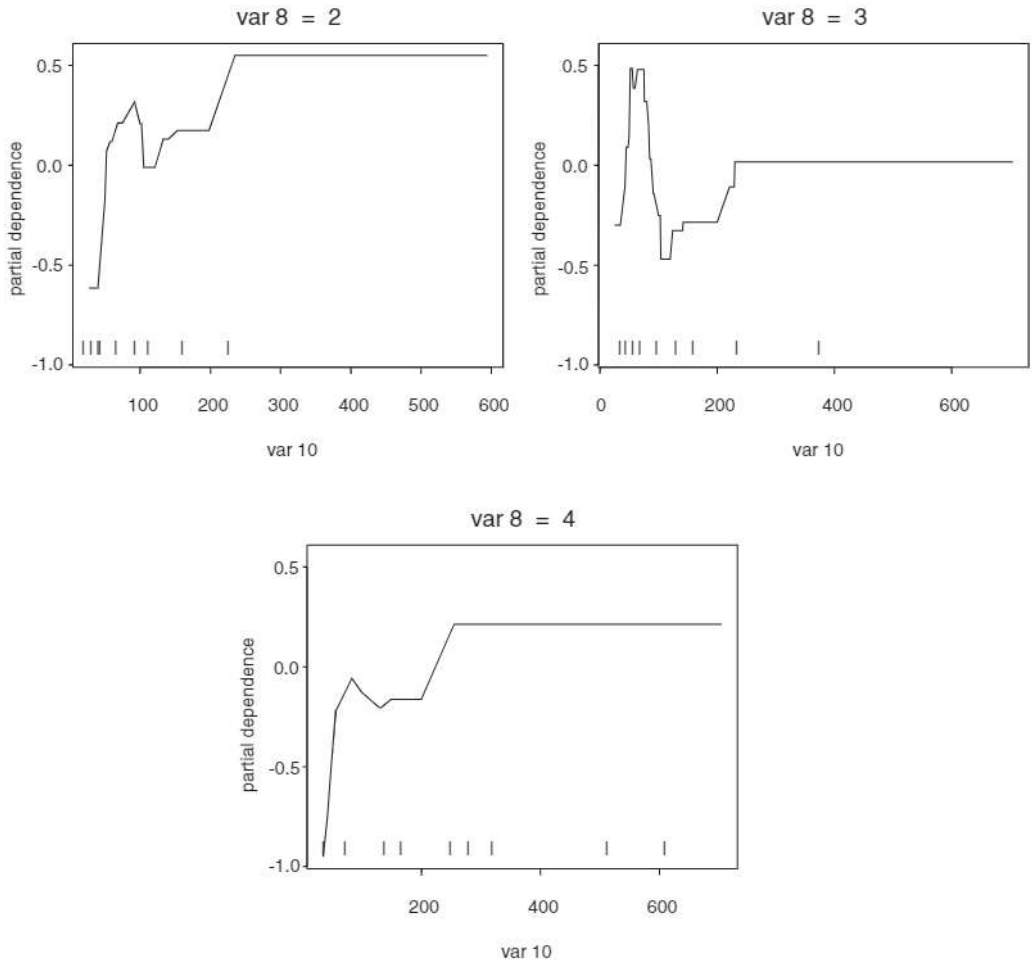

Figure 6. Joint partial dependence plots for variables 10 and 8 for class 2 (moderate dysplasia). From top left to bottom the partial dependence of variable 10 according to the three different levels of variable 8. Variable 10 = total number of abnormal cells, 8 = nucleus/cytoplasm ratio.

To continue to demonstrate the use of partial dependence plots, Figure 7 shows joint partial dependence plots for class 2 for its four most important (numerical) variables: *total number of abnormal cells*, *number of mitoses*, *number of nucleoli* and *number of abnormal cells per fragment*. Because they display the joint partial dependence on two variables in a single plot, the plots are now three-dimensional, showing the different interactions of the pairs of variables involved. The fact that there is also a differential effect across classes is demonstrated in Figure 8, where the interaction of variable 11 (*number of mitoses*) and variable 5 (*number of nucleoli*) is given for the classes 1, 3, 4 and 5. The plots differ considerably both from each other, as well as from the plot for class 2 in the bottom left panel in Figure 7.

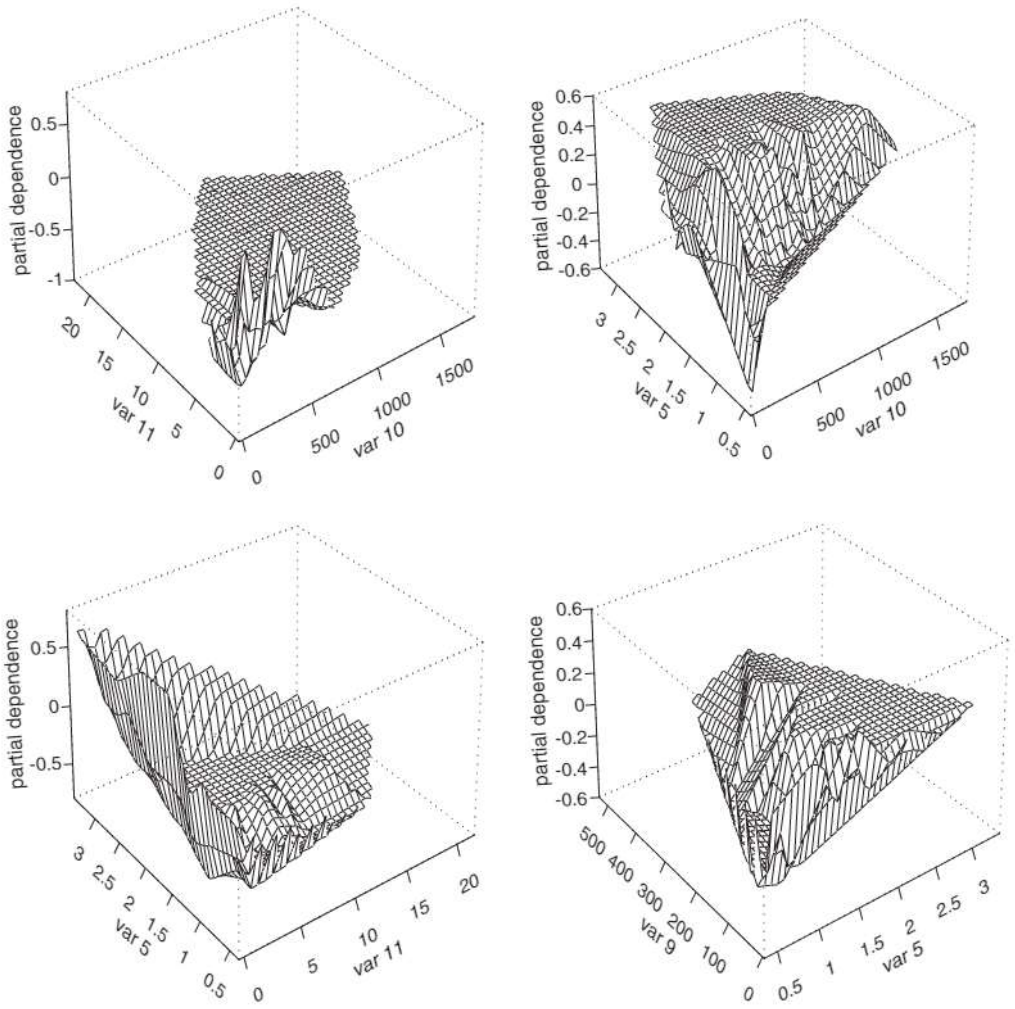

Figure 7. Joint partial dependence plots for the four most important (numerical) variables for class 2 (moderate dysplasia). Variable 10 = total number of abnormal cells, 11 = number of mitoses, 5 = number of nucleoli, 9 = number of abnormal cells per fragment.

7. CONCLUSION

In this paper we summarize the approach of using multiple additive regression trees (MART) to analyse complicated data sets, as in data mining. Complicated data sets usually contain mixtures of nominal, ordinal and numerical variables, and include missing values. Regression and classification trees are eminently fit to find structure among such variables with respect to the prediction of the response variable, and without having to define appropriate transformations of the predictor variables beforehand. MART combines regression and classification trees with a powerful extension, called boosting, that takes care of the major disadvantage of regression trees, that is, their inaccuracy.

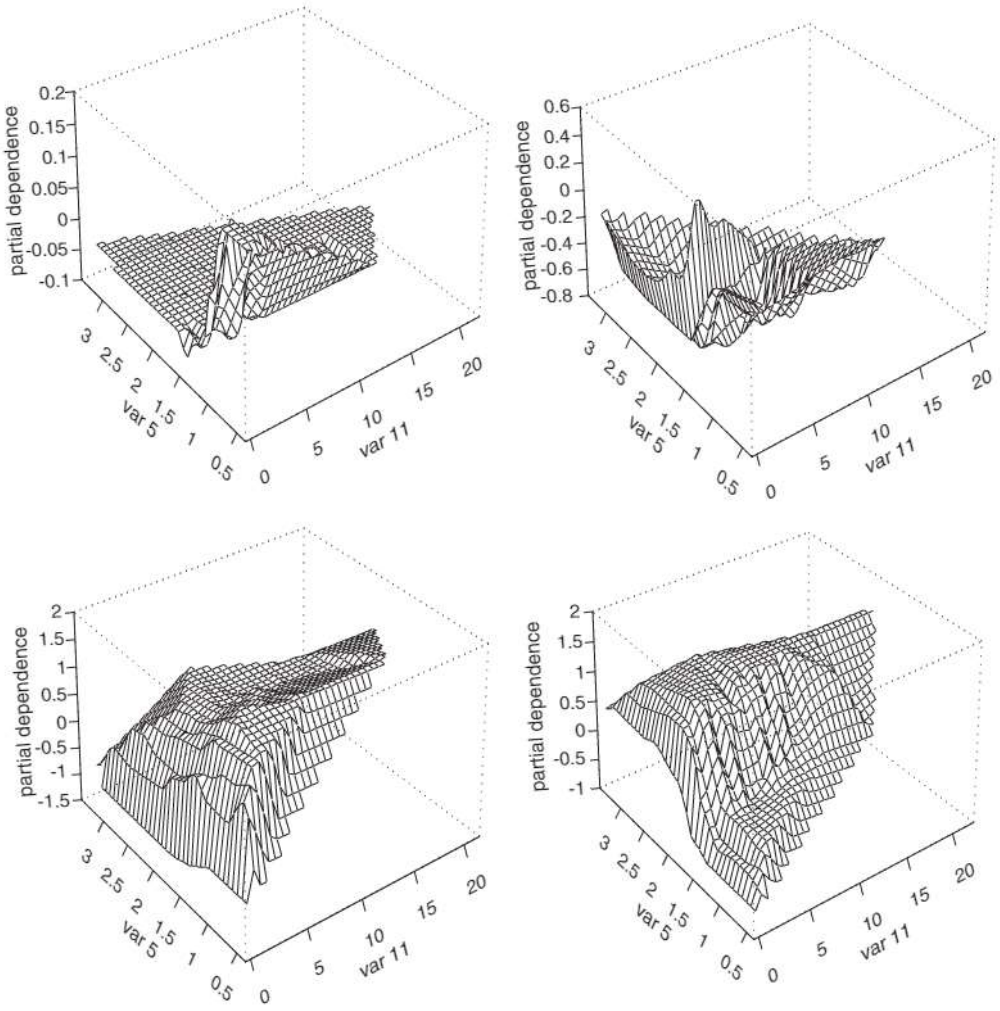

Figure 8. Joint partial dependence plots for variables 11 and 5 for (from the top left to the bottom right) the classes 1 (mild dysplasia), 3 (severe dysplasia), 4 (carcinoma *in situ*) and 5 (invasive squamous cell carcinoma). Variable 11 = number of mitoses, 5 = number of nucleoli.

The ingredients of the MART procedure have been described, and MART has been applied to a real life data set from epidemiology, concerning the prediction of cervical cancer and its precursor lesions, from a number of qualitative and quantitative variables. The prediction of the histologic diagnosis into five diagnostic categories (mild, moderate and severe dysplasia, carcinoma *in situ*, and invasive squamous cell carcinoma) turns out to be superior for MART compared to both a straightforward classification tree (using CART) and a canonical discriminant analysis. The differential importance of the predictor variables has been detailed by using powerful visual tools, displaying their influence on the prediction of the diagnosis.

REFERENCES

1. Breiman L, Friedman JH, Olshen R, Stone C. *Classification and Regression Trees*. Wadsworth: Pacific Grove, 1984.
2. Freund Y, Schapire R. Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, 1996; 148–156, Morgan Kaufmann: San Francisco, CA.
3. Friedman JH, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion). *Annals of Statistics* 2000; **28**:337–407.
4. Friedman JH. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 2001; **29**: 1189–1232.
5. Boon ME, Zeppa P, Ouwerkerk-Noordam E, Kok LP. Exploiting the tooth-pick effect of the cytobrush by plastic embedding of cervical samples. *Acta Cytologica* 1990; **35**:57–63.
6. Meulman JJ, Zeppa P, Boon ME, Rietveld WJ. Prediction of various grades of cervical preneoplasia and neoplasia on plastic embedded cytobrush samples: discriminant analysis with qualitative and quantitative predictors. *Analytical and Quantitative Cytology and Histology* 1992; **14**:60–72.
7. Steinberg D, Colla P. *CART – Classification and Regression Trees*. Salford Systems: San Diego, CA, 1997.

