Variable Selection Using Random Forests

Marco Sandri and Paola Zuccolotto[1]

Dipartimento Metodi Quantitativi,
Università di Brescia
c.da S. Chiara, 50 - 25122 Brescia, Italy
zuk@eco.unibs.it

Abstract. One of the main topic in the development of predictive models is the identification of variables which are predictors of a given outcome. Automated model selection methods, such as backward or forward stepwise regression, are classical solutions to this problem, but are generally based on strong assumptions about the functional form of the model or the distribution of residuals. In this paper an alternative selection method, based on the technique of Random Forests, is proposed in the context of classification, with an application to a real dataset.

1 Introduction

In many empirical analyses a crucial problem is the presence in the data of a set of variables not significatively contributing to explain the analyzed phenomenon, but capable to create a random noise which prevents from distinguishing the main effects and the relevant predictors. In this context proper methods are necessary in order to identify variables that are predictors of a given outcome. Many automatic variable selection techniques have been proposed in the literature, for example the backward or forward stepwise regression (see Miller (1984) and Hocking (1976)) or the recent stepwise bootstrap method of Austin and Tu (2004). These methods are for the most part based on assumptions about the functional form of the models or on the distribution of residuals. These hypothesis can be dangerously strong in presence of one or more of the following situations: (*i*) a large number of observed variables is available, (*ii*) collinearity is present, (*iii*) the data generating process is complex, (*iv*) the sample size is small with reference to all these conditions. Data analysis can be basically approached by two points of view: data modeling and algorithmic modeling (Breiman (2001b)). The former assumes that data are generated by a given stochastic model, while the latter treats data mechanism as unknown, a *black box* whose insides are complex and often partly unknowable. The aim of the present paper is to propose a variable selection method based on the algorithmic approach and to examine its performance on a particular dataset. In the mid-1980s two powerful new algorithms for fitting data were developed: neural nets and decision trees, and were applied in a wide range of fields, from physics, to medicine, to economics, even if in some applications (see e.g. Ennis et al. (1998)) their performance was poorer

than that of simpler models like linear logistic regression. The main shortcomings of these two methods were overfitting and instability, the latter with particular reference to decision trees.

While overfitting has been long discussed and many techniques are available to overcome the problem (stopping rules, cross-validation, pruning, ...), few has been made to handle instability, a problem occurring when there are many different models with similar predictive accuracy and a slight perturbation in the data or in the model construction can cause a skip from one model to another, close in terms of error, but distant in terms of the meaning (Breiman (1996a)). The proposal of Random Forests (Breiman (2001a)), a method for classification or regression based on the repeated growing of trees through the introduction of a random perturbation, tries to manage these situations averaging the outcome of a great number of models fitted to the same dataset. As a subproduct of this technique, the identification of variables which are important in a great number of models provides suggestions in terms of variable selection. The proposal of this paper is to use the technique of Random Forests (RF) as a tool for variable selection, and a procedure is introduced and evaluated on a real dataset. The paper is organized as follows: in section 2 the technique of RF is briefly recalled, confining the attention to the case of classification, in section 3 a variable selection method based on RF is proposed, the application to a real dataset is reported in section 4, conclusive remarks follow in section 5.

2 Random Forests

A population is partitioned into two or more groups, according to some qualitative feature. It follows that each individual in the population belongs to (only) one group. The information about the group is contained in the categorical variable Y, while relevant further information is collected in a set of exogenous variables $\mathbf{X}$, always known, which is assumed to somewhat affect Y. Given a random sample $S = \{(y_1, \mathbf{x}_1); \cdots ; (y_n, \mathbf{x}_n)\}$, several statistical techniques are available in order to determine an operative rule $h(\mathbf{x})$ called *classifier*, used to assign to one group an individual of the population, not contained in the sample, for which only the exogenous variables $\mathbf{x}_{n+1}$ are known. A *random classifier* $h(\mathbf{x}, \boldsymbol{\theta})$ is a classifier whose prediction about y depends, besides on the input vector $\mathbf{x}$, on a random vector $\boldsymbol{\theta}$ from a known distribution $\boldsymbol{\Theta}$. Given a i.i.d. sequence $\{\boldsymbol{\theta}_k\} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots , \boldsymbol{\theta}_k, \cdots\}$ of random vectors from a known distribution $\boldsymbol{\Theta}$, a Random Forest $\mathrm{RF}(\mathbf{x}, \{\boldsymbol{\theta}_k\})$ is itself a random classifier, consisting of a sequence of random classifiers $\{h(\mathbf{x}, \boldsymbol{\theta}_1), h(\mathbf{x}, \boldsymbol{\theta}_2), \cdots , h(\mathbf{x}, \boldsymbol{\theta}_k), \cdots\}$ each predicting a value for y at input $\mathbf{x}$. The RF prediction for y is expressed in terms of probability of Y assuming the value y, $Pr\{Y = y\}$. By definition a RF is composed by an infinite number of classifiers, but from an operational point of view the term is used

to indicate a finite set of classifiers $\{h(\mathbf{x}, \boldsymbol{\theta}_1), h(\mathbf{x}, \boldsymbol{\theta}_2), \cdots, h(\mathbf{x}, \boldsymbol{\theta}_k)\}$. The k-set's prediction for y corresponds to the prediction whose frequency exceeds a given threshold[1]. Asymptotic results have been derived in order to know the behavior of the set as the number of classifiers increases. Limiting laws and statistical features of RF have been developed by Breiman (2001a) and a detailed explanation can be found in Sandri and Zuccolotto (2004). The theory of RF is quite general and can be applied to several kinds of classifiers and randomizations: examples are already present in literature, for instance the bagging technique of Breiman (1996b) or the random split selection of Dietterich (2000). Moreover other well-known techniques, like bootstrap itself, although introduced in different contexts, can be led back to the RF framework. Nevertheless by now the methodology called Random Forests is used uniquely with reference to its original formulation, due to Breiman (2001a), which uses CART-structured (Classification And Regression Trees, Breiman et al. (19984)) classifiers. RF with randomly selected inputs are sequences of trees grown by selecting at random *at each node* a small group of F input variables to split on. This procedure is often used in tandem with *bagging* (Breiman (1996b)), that is with a random selection of a subsample of the original training set at each tree. The trees obtained in this way are a RF, that is a k-set of random classifiers $\{h(\mathbf{x}, \boldsymbol{\theta}_1), h(\mathbf{x}, \boldsymbol{\theta}_2), \cdots, h(\mathbf{x}, \boldsymbol{\theta}_k)\}$ where the vectors $\boldsymbol{\theta}_i$ denote the randomization injected by the subsample drawing and by the selection of the F variables at each node.

2.1 Variable importance measures

The main drawback of using a set of random classifiers lies in its explanatory power: predictions are the outcome of a *black box* where it is impossible to distinguish the contribution of the single predictors. With RF this problem is even more crucial, because the method performs very well especially in presence of a small number of informative predictors hidden among a great number of noise variables. To overcome this weakness the following four measures of variable importance are available in order to identify the informative predictors and exclude the others (Breiman (2002)):

- **Measure 1**: at each tree of the RF all the values of the h-th variable are randomly permuted and new classifications are obtained with this new dataset, over only those individuals who have not contributed to the growing of the tree. At the end a new misclassification error rate $\hat{e}_h$ is

[1] In the standard case, the k-set's prediction for y corresponds to the most voted prediction, but a generalization is needed, as sometimes real datasets are characterized by extremely unbalanced class frequencies, so that the prediction rule of the RF has to be changed to other than majority votes. The optimal cutoff value can be determined for example with the usual method based on the joint maximization of sensitivity and specificity.

then computed and compared with $\hat{e}$. The $M1$ measure for h-th variable is given by

$$M1_h = \max\{0; \hat{e}_h - \hat{e}\}.$$

- **Measure 2**: for an individual $(y, \mathbf{x})$ the margin function $mg(y, \mathbf{x})$ is defined as a measure of the extent to which the proportion of correct classifications exceeds the proportion of the most voted incorrect classifications. If at each tree all the values of the h-th variable are randomly permuted, new margins $mg_h(y, \mathbf{x})$ can be calculated over only those trees which have not been grown with that subject. The $M2$ measure of importance is given by the average lowering of the margin across all cases:

$$M2_h = \max\{0; av_S[mg(y, \mathbf{x}) - mg_h(y, \mathbf{x})]\}.$$

- **Measure 3**: in the framework just described for $M2$, the $M3$ measure is given by the difference between the number of lowered and raised margins:

$$M3_h = \max\{0; \#[mg(y, \mathbf{x}) < mg_h(y, \mathbf{x})] - \#[mg(y, \mathbf{x}) \geq mg_h(y, \mathbf{x})]\}.$$

- **Measure 4**: at each node z in every tree only a small number of variables is randomly chosen to split on, relying on some splitting criterion given by a heterogeneity index such as the Gini index or the Shannon entropy. Let $d(h, z)$ be the decrease in the heterogeneity index allowed by variable $\mathbf{X}_h$ at node z, then $\mathbf{X}_h$ is used to split at node z if $d(h, z) > d(w, z)$ for all variables $\mathbf{X}_w$ randomly chosen at node z. The $M4$ measure is calculated as the sum of all decreases in the RF due to h-th variable, divided by the number of trees:

$$M4_h = \frac{1}{k} \sum_z [d(h, z) I(h, z)]$$

where $I(h, z)$ is the indicator function that is equal to 1 if h-th variable is used to split at node z and 0 otherwise.

3 Variable selection using Random Forests

In this paper the possible use of RF as a method for variable selection is emphasized, relying on the above mentioned four importance measures. A selection procedure can be defined, observing that the exogenous variables described by the four measures can be considered as points in a four-dimensional space, with the following steps: (1) calculate a four-dimensional centroid with coordinates given by an average (or a median) of the four measures; (2) calculate the distance of each point-variable from the centroid and arrange the calculated distances in non-increasing order; (3) select the variable whose distance from the centroid exceeds a given threshold, for example the average distance. This simple method is often quite effective, because the noise variables represented in the four-dimensional space tend to cluster together

in a unique group and the predictors appear like outliers. A refinement of this proposal, which provides a useful graphical representation, can be proposed observing that the four measures are often correlated and this allows a dimensional reduction of the space where the variables are defined. With a simple Principal Component Analysis (PCA) the first two factors can be selected and a scatterplot of the variables can be represented in the two-dimensional factorial space, where the cluster of noise variables and the "outliers" can be recognized. The above described procedure based on the calculation of the distances from an average centroid can be applied also in this context and helps deciding which points have to be effectively considered outliers[2]. Simulation studies show that these methods very favorably compare with a forward stepwise logistic regression, even when the real data generating mechanism is a logistic one. Their major advantage lies in a sensibly smaller number of wrongly identified predictors. The main problem of these methods consists in the definition of the threshold between predictive and not predictive variables. To help deciding if this threshold exists and where it could be placed, a useful graphical representation could be a sort of scree-plot of the distances from the centroid, where the actual existence of two groups of variables, and the positioning of the threshold between them, can be easily recognized.

4 Case study

A prospective study was conducted from January 1995 to December 1998 by the First Department of General Surgery (Ospedale Maggiore di Borgo Trento, Verona, Italy) in patients affected by acute peptic ulcer who underwent endoscopic examination and were treated with a particular injection therapy. The aims of the study were to identify risk factors for recurrence of hemorrhage, as early prediction and treatment of rebleeding would improve the overall outcome of the therapy. The dataset consists of 499 cases, observed according to 32 exogenous variables related to patient history (gender, age, bleeding at home or during hospitalization, previous peptic ulcer disease, previous gastrointestinal hemorrhage, intake of nonsteoridal anti-inflammatory drugs, intake of anticoagulant drugs, associated diseases, recent - within 30 days - or past - more than 30 days - surgical operations), to the magnitude of bleeding (symptoms: haematemesis, coffee-ground vomit, melena, anemia; systolic blood pressure, heart rate, hypovolemic shock, hematocrit

[2] In this case the distance function can take into consideration the importance of the two factors and a weight can be introduced given, for example, by the fraction of total variance accounted for by each factor or by the correspondent eigenvalue. Actually this procedure could be redundant, as the space rotation implied by the PCA, already involves a overdispersion of points along the more informative dimensions.

and hemoglobin level, units of blood transfused), to endoscopic state (number, size, location of peptic ulcers, Forrest classification, presence of gastritis or duodenitis). The values of all the variables are classified into categories according to medical suggestions. We think that the use of the raw data could allow a more detailed analysis. The results were presented in a paper (Guglielmi et al. (2002)) where a logistic regression with variables selected relying both on statistical evidences and on medical experience was able to provide a (in-sample) 24% misclassification error with sensitivity and specificity equal to 76%[3]. In this paper two logistic regressions are fitted to the same data, with variables selected respectively by a AIC stepwise procedure[4] (Model A) and by our RF-based method (Model B).

The AIC stepwise variable selection method identifies nine relevant predictors[5], while using the RF procedure, eight predictors are selected[6]. In both cases the resulting predictors has been judged reasonable on the basis of medical experience. In the left part of Figure 1 the scree plot of variable distances from the centroid are represented for three approaches (the basic method in the four-dimensional space, the refinement based on the PCA of the four measures with Euclidean distance or weighted Euclidean distance), while the right part of Figure 1 shows the two-dimensional scatterplot of the variables in the first two principal components space, with a virtual line separating the outlier variables selected as predictors.

In order to evaluate the performance of the two models, a cross-validation study has been carried out with validation sets of size 125 (25% of the sample) and $r = 1000$ repeated data splittings. The estimated probabilities of the two models are used to classify a patient being or not at risk of rebleeding, according to a cutoff point determined by minimizing the absolute difference between sensitivity and specificity in each validation set. Results are reported in Table 1, where also the corresponding in-sample statistics are shown.

The two models exhibit a substantially equal goodness-of-fit and also have a high agreement rate (in the in-sample analysis 91.58% of the individuals is classified in the same class by the two models). However it has to be noticed that Model B, built with the RF variable selection, has a reduced number of

[3] The predictors included in the model were: associated diseases/liver cirrhosis (`livcir`), recent surgical operations (`recsurg`), systolic blood pressure (`sbp`), symptoms/haematemesis (`hematem`), ulcer size (`size`), ulcer location (`location(2)`), Forrest class (`Forrest`).

[4] Coherently with our previous simulative studies, a forward selection is used. Anyway, the backward option was experimented: it leads to a less parsimonious model with substantially the same predictive performance.

[5] Forrest class, systolic blood pressure, ulcer size, recent surgical operations, ulcer location, units of blood transfused (`uobt`), age (`age`), symptoms/haematemesis, intake of anticoagulant drugs (`anticoag`).

[6] Systolic blood pressure, Forrest class, hypovolemic shock (`shock`), recent surgical operations, age, ulcer size, symptoms/haematemesis and-or melena (`symptoms`), ulcer location.

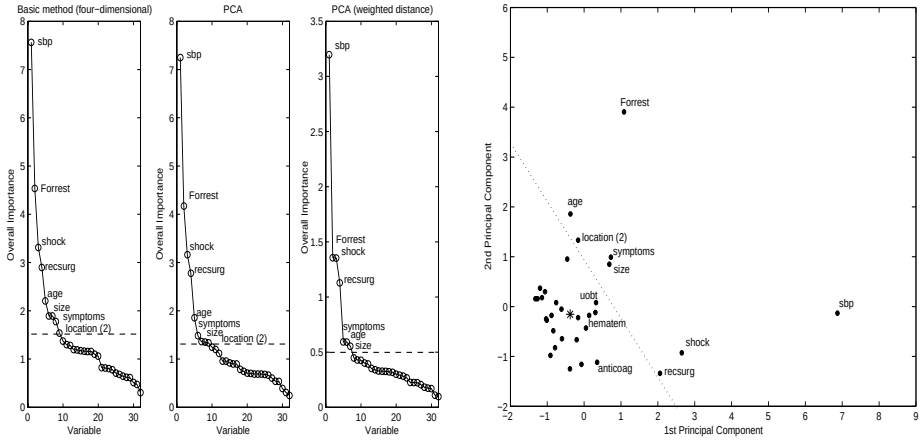

Fig. 1. Left: Scree plots of variable distances from centroid (thresholds: average distances); Right: Scatterplot of the variables in the first two principal components space.

Table 1. *Misclassification error, Cohen k, sensitivity, specificity, cutoff value of the two logistic models (In- and out-of-sample analyses)*

	Misclassification error	Cohen k	Sensitivity	Specificity	Cutoff
	Model A/Model B - In-sample analysis				
Model A	23.05%	0.3482	77.03%	76.94%	0.1783
Model B	22.65%	0.3556	77.03%	77.41%	0.1642
	Model A - Out-of-sample analysis				
25th percentile	21.60%	0.2224	55.55%	73.15%	0.1629
Median	24.00%	0.2751	63.63%	76.19%	0.1742
75th percentile	26.40%	0.3320	70.65%	79.25%	0.1841
	Model B - Out-of-sample analysis				
25th percentile	22.40%	0.2138	55.55%	71.96%	0.1558
Median	24.80%	0.2686	63.63%	75.23%	0.1658
75th percentile	27.20%	0.3196	72.22%	78.57%	0.1771

predictors, coherently with the simulation results assessing a good capability of the method in the false variables selection rate.

5 Concluding remarks

In this paper a variable selection method based on Breiman's Random Forests is proposed and applied to a real dataset of patients affected by acute peptic ulcers, in order to identify risk factors for recurrence of hemorrhage. The main advantage of selecting relevant variables through an algorithmic modeling technique is the independence from any assumptions on the relationships among variables and on the distribution of errors. After having selected the

predictors, a model could be developed with some given hypothesis, and this outlines Random Forests as a technique for preliminary analysis and variable selection and not only for classification or regression, which are its main purposes. The results on real data confirm what expected on the basis of simulation studies: the RF-based variable selection identifies a smaller number of relevant predictors and allows the construction of a more parsimonious model but with predictive performance similar to the logistic model selected by the AIC stepwise procedure. Further research is currently exploring the advantages deriving from the combination of measures coming from model-based prediction methods and algorithmic modeling techniques. Moreover simulation studies have highlighted the presence of a bias effect in a commonly used algorithmic variable importance measure. An adjustment strategy is under development (Sandri and Zuccolotto (2006)).

References

AUSTIN, P. and TU, J. (2004): Bootstrap methods for developing predictive models. *The American Statistician, 58, 131–137.*

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. and STONE, C.J. (1984): *Classification and Regression Trees.* Chapman & Hall, London.

BREIMAN, L. (1996a): The heuristic of instability in model selection. *Annals of Statistics, 24, 2350–2383.*

BREIMAN, L. (1996b): Bagging predictions. *Machine Learning, 24, 123–140.*

BREIMAN, L. (2001a): Random Forests. *Machine Learning, 45, 5–32.*

BREIMAN, L. (2001b): Statistical modeling: the two cultures. *Statistical Science, 16, 199–231.*

BREIMAN, L. (2002): Manual on setting up, using, and understanding Random Forests v3.1. *Technical Report*, `http://oz.berkeley.edu/users/breiman`.

DIETTERICH, T. (2000): An experimental comparison of three methods for construction ensembles of decision trees: bagging, boosting and randomization. *Machine Learning, 40, 139–157.*

ENNIS, M., HINTON, G., NAYLOR, D., REVOW, M. and TIBSHIRANI, R. (1998): A comparison of statistical learning methods on the gusto database. *Statistics in Medicine, 17, 2501–2508.*

GUGLIELMI, A., RUZZENENTE, A., SANDRI, M., KIND, R., LOMBARDO, F., RODELLA, L., CATALANO, F., DE MANZONI, G. and CORDIANO, C. (2002): Risk assessment and prediction of rebleeding in bleeding gastroduodenal ulcer. *Endoscopy, 34, 771–779.*

HOCKING, R.R. (1976): The analysis and selection of variables in linear regression. *Biometrics, 42, 1–49.*

MILLER, A.J. (1984): Selection of subsets of regression variables. *Journal of the Royal Statistical Society, Series A, 147, 389–425.*

SANDRI, M. and ZUCCOLOTTO, P. (2004): Classification with Random Forests: the theoretical framework. *Rapporto di Ricerca del Dipartimento Metodi Quantitativi, Università degli Studi di Brescia, 235.*

SANDRI, M. and ZUCCOLOTTO, P. (2006): Analysis of a bias effect on a tree-based variable importance measure. Evaluation of an empirical adjustment strategy. *Manuscript.*