Machine Learning Algorithms for Predicting Severe Crises of Sickle Cell Disease

Clara Allayous
Département de Biologie, Université des Antilles et de la guyane
Stéphan Clémençon
MODALX - Univesité Paris X & Unité Méta@risk-INRA& LPMA-UMR CNRS 7599
Bassirou Diagne
MAPMO-UMR CNRS 6628 - Université d'Orléans
Richard Emilion
MAPMO-UMR CNRS 6628 - Université d'Orléans
Thérèse Marianne
Département de Biologie, Université des Antilles et de la guyane

July 25, 2008

Abstract

It is the main purpose of this paper to demonstrate how recent advances in predictive machine learning van be used for quantifying the risk of an acute splenic sequestration crisis (ASSC), a serious symptom of sickle cell disease (SCD). Precisely, based on a training data set, the goal is here to learn how to predict (estimate) the level of severity of the crisis, represented as a binary random variable Y ("mild" vs "severe"), from collected medical measurement $X = (X^1, ..., X^D)$. in practice, the prediction rule is obtained by thresholding a certain diagnostic test statistic $s(X)$ at a certain cutoff point t: patients with a test score above the selected cutoff point are predicted as "prone to severe crisis", otherwise as "prone to a mild crisis". Recently, new machine learning techniques have emerged for constructing a "good predictor", when the ability to distinguish between the two populations is measured either by the standard prediction error or else by the AUC criterion, a measure of quality extensively used for evaluating diagnostic tests in a wide variety of applications. In this article we show how machine learning methods such as boosting or ranking can be used for accurately evaluating the risk of a severe ASSC, while avoiding the question of modeling the underlying distribution of the response Y conditioned upon the explanatory observation X, in contrast to the logistic regression approach.

Keywords and phrases: sickle cell disease, ranking, ROC curve, AUC criterion, boosting, decision trees, bootstrap.
AMS 2000 **Mathematics Subject Classification:** 60G70, 60J10, 60K20.

1 Introduction

Sickle cell disease is an inherited disease cells, characterized by pain episodes, anemia (shortage of red blood cells), serious infections and damages to vital organs. The symptoms of sickle cell disease are caused by abnormal hemoglobin, the main protein inside red blood cells that carries oxygen from the lungs and takes it to every part of the body. Normally, red blood cells are round and flexible and flow easily through blood vessels. But when struck down by sickle cell disease, the abnormal hemoglobin causes red blood cells to become stiff and, under the microscope, may look like a C-shaped farm tool called a "sickle". These stiffer red blood cells can get stick in tiny blood vessels, cutting off the blood supply to nearby tissues. This is what causes pain (usually termed a "sickle cell pain episode" or "crisis" for short) and sometimes organ damage. Sickle-shaped red blood cells also die and break down more quickly than normal red blood cells, resulting in anemia. There ar several common forms of sickle cell disease, called SS (individual inherit one sickle cell gene from each parent), SC (the child inherits one sickle cell gene and one gene for another abnormal type of hemoglobin called "C") and S-beta thalassemia (the child inherits one sickle cell gene and one gene for beta thalassemia, another inherited anemia). The effects of sickle cell disease vary greatly from one person to another. Some affected children/adults are usually healthy, while others are frequently hospitalized. In order to describe the important clinical variability among such a crisis, medical experts distinguish between two symptom classes: mild and severe.

In this paper, we tackle the problem of explaining membership to a certain symptom class based on several medical measurements from a predictive perspective. There area wide variety of ways one can go trying to determine a good prediction. One may seek the opinions of medical experts for instance or else use a training data set, consisting of previously observed cases for which both binary response and predictor variables have been collected at the Caribbean Center for Sickle Cell Disease over the last ten years (refer to http://drepano.org for further details) and we investigate the performance of recent machine learning procedures such as boosting and bagging (on decisions trees on interpretability grounds) for building predictive models. As a first go, the quality of obtained prediction functions is measured by the (expected) prediction error in a standard fashion. However, most commonly used learning procedures output prediction rules based on comparing a certain test statistic to an adequate cutoff value and a standard approach for statistical evaluation of such "diagnostic tests" involves computing summary measures based on the so-termed Receiver Operator Characteristic curve (ROC curve, in abbreviated form), such as the AUC criterion (AUC standing for "area under the ROC curve"). One may refer to [13, 16, 15, 11] for instance, see also [14] for a recent account oriented to medical applications. Hence, the predictive power of diagnostic tests output by th machine learning algorithms considered in this article is also evaluated in terms of ROC curve. In this respect, the approach developed in [8] based on pairwise classification (see also [6,7]) for refining standard classification methods, is shown to yield significant improvements.

The paper is organized as follows. A precise formulation of the medical diagnostic prob-

lem considered in this paper is given in section 2 from the predictive machine learning angle, together with a brief description of the algorithms and data used in our experiments. And in section 3 the results of the various learning procedures we considered are compared in respect of their accuracy and interpretability.

2 Background-Data & Methods

In this section, we first set out the notation and recall certain key notions arising from prediction (machine) learning. The learning algorithms candidates considered in this article for building a good diagnostic test function, competitive in predicting sever acute splenic sequestration crises, are next briefly reviewed. Then, the data used for learning how to predict the level of severity of the next ASSC are datailed.

2.1 Formulation of the prediction problem - the bipartite setup

In the bipartite setup, the goal is to predict a binary output random variable Y, taking its values in $\{-1, +1\}$ say, from the observation of a set of explanatory variables $X = (X_1, ..., X_d)$. in the present application, Y represents the level of severity of the ASSC, mild and severe being respectively coded as -1 and $+1$, and X the collection of medical measurements described in Table?? for predicting Y. Using a training sample $\{(x_i, y_i)\}_{1 \leq i \leq n}$ of n cases for which the values of both the explanatory and the output variables have been jointly observed, the matter is to build a prediction rule C that maps all points $x = (x_1, ..., x_d)$ in the space $\mathcal{X}$ of all values of the input random vector to a point $C(x)$ in the space of response values, namely $\{-1, +1\}$ with a predictive risk

$$R(C) = \mathbb{P}(Y \neq C(X)))$$

as low as possible (the probability in (1) being taken over the joint distribution of (X, Y)).

2.2 The database

We have performed the above algorithms on a data set of 42 children described by 15 characteristics. These data were collected during 10 years at the "Centre Caribéen de la drépanocytose (Pointe-à-Titre, Guadeloupe, French West indies)", the structure of this data table is displayed below in table 1.

Variables	Description
GR	number of basic red globules
HT	basic ht
LEU	leucocyte number
NEU	neutrophyle number
PLT	number of plates of blood
HYGRO	hygrometry
HB	rate of basic hemoglobin
AGE	age of children
LIVER	liver
SPLEEN	spleen
DELTALEU	variation of leucocytes
DELTANEU	variation of neutrophiles
DELTAHB	variation of the hemoglobin lvel
DELTAPLT	variation of plates of blood
PERIODYEAR	period of the crisis in the year
CLASSE	CLASSE equal 1 when HG greater than 4 and -1 otherwise

Table 1: Description the input variables.

As it was observed that any ASSC is mainly characterized by a decrease of hemoglobin level, we have decide that crisis with hemoglobin level less 4g/dl are severe and the others are mild.
We didn't include HB as predictor in analyzes because our definition of ASSC was based on HG level. Also GR, DELTAHB were correlated with HT (respectively $t = 12.36, p < 2.2e-16; t = 9.45, p = 2.2e-16$), and DELTALEU, DELTANEU were strongly positively correlated with NEU (respectively $t = 38.8, p = 2.2e-16; t = 7.38, p < 2.2e-16$), in the total sample of 132 subjects. Therefore we eliminated: HB, GR, DELTAHB, DELTALEU, DELTANEU, DELTASPLEEN.
We have conducted the above learning algorithms and compared their performance in order to choose the variables that indicated a good prediction of the ASSC. We have used the **R** software where all these methods are available in some statistical packages, but we have implemented the ranktree, ranking boosting and ranking bagging.
The methods used here belong to the family of supervised learning ones, which are suitable for classification, ranking and predictions tasks. Supervised learning methods require a data set $\mathbf{S}$ of the form $\mathbf{S} = \{(\mathbf{x_i}, \mathbf{y_i})\}$ where the x_i are the medical measurements observed on the patients and also the age variable while y_i is the class label corresponding to the mild or severity symptom. This set of classified instances is used to learn the model and is therefore called the training set. A similar data set, called the test set, is used for evaluating the performance of the methods.
The classification model assigns each feature vector x_i of the test set to a class c_i, which is then compared to the true class y_i of the instance. If $c_i = y_i$, the instance is considered correctly classified, otherwise it counts as a misclassification. But in the case of ranking, the goal is to rank the positive instances higher than negative instance instead of simply

classifying them.

2.3 Methods

2.3.1 ADABOOST

Adaboost, a general discriminative learning algorithm invented by Freund and Schapire (1997) can be described as follows:
$F_0 = 0$
for $t = 1, ..., T$
$w_i^t = w_i^t \exp(-y_i F_{t-1}(x_i))$
Get h_t from weak learner
$\alpha_t = \frac{1}{2}(\frac{\sum_{i:h_t(x_i)=1, y_i=1} w_i^t}{\sum_{i:h_t(x_i)=1, y_i=-1} w_i^t}$
$F_{t+1} = F_t + \alpha_t h_t$

Figure 1. The Adaboost algorithm.

The basic idea of Adaboost is to repeatedly apply a simple learning algorithm called the weak or the base learner, to different weighting of the same training est. In its simplest form, Adaboost is intended for binary prediction problems where the training set consists of pairs $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n), x_i$ corresponds to the feature of an example, and $y_i \in \{-1, 1\}$ is the binary label to be predicted. A weighting of the training examples is an assignment of a non-negative real value w_i to each example (x_i, y_i).
On iteration t of the boosting process, the weak learner is applied to the training set with a set of weights $w_1^t, ..., w_n^t$ and produces a prediction rule h_t that maps x to $\{0, 1\}$. The requirement on the weak learner is for $h_t(x)$ to have a small but significant correlation with the example labels y when measured using the current weighting of the examples. After the rule h_t is generated, the examples weights are changed so that the weak predictions $h_t(x)$ and the label y are uncorrelated. The weak learner is then called with the new weights over the training examples, and the process repeats. Finally, all of the weak prediction rules are combined into a single strong rule using a weighted majority vote. One Can prove that if the rules generated in the iterations are all slightly correlated with the label, then the strong rule will have a very high correlation with the label, in order words, it will predict the label very accurately.
The whole process can be seen as a variational method in which an approximation $F(x)$ is repeatedly changed by adding to it small corrections given by the weak prediction functions. The strong prediction rule learned by Adaboost is denoted by $sign(F(x))$.

2.3.2 Bagging

Let's consider a learning sample of $\mathbf{S}$ consist of data $\{(x_i, y_i), i = 1, ..., N\}$, where $y \in \{1, ..., J\}$ denotes the class variable. Bootstrap aggregating or bagging (Breiman 1996) works by learning multiple classifiers on bootstrap samples. For each trial $t = 1, ..., T$, a

sample of size N is selected randomly with replacement from the original learning set $\mathbf{S}$. Classifiers $\phi_t(x, \mathbf{S}_t^{(B)})$ based on the bootstrap training set $\{\mathbf{S}_t^{(B)}\}$ is formed. The bagged classifier ϕ_B is then constructed by combining ϕ_t with simple voting.
Now we specifies the bagging algorithm as follows
1. A sequence of bootstrap samples $\{\mathbf{S}_t^{(B)}\}$ is drawn from the learning sample $\mathbf{S}$. Classifier $\phi_t = \phi$ generated by using ,$\forall t \leq T$, for a fixed number T.
2. Let N_j be the number of times that an instance x is classified as class j through $\phi_t = \phi(x, \mathbb{S}_t^{(B)}$. That is

$$N_j = \sum_{t=1}^{T} I[\phi(x, \mathbb{S}_t^{(B)}) = j]$$

for $j \in \{1, ..., J\}$, where I stands for the indicator function. 3. Bagged predictor is created by simple voting. That is

$$\phi_B(x) = \arg\max_{j=1}^{J} N_j.$$

2.3.3 Boosting Tree Ranking

The existing methods for growing decision trees typically use splitting criteria based on error/accuracy or discrimination. In this method we use AUC splitting criterion.

2.3.4 Logistic Regression

Denote the gold standards by a random variable Y and the other medical feature by $X_1, X_2, ..., X_n$. Let $Y = 1$ when a patient has a very severe crisis and $Y = 0$ when he has a mild one. The logistic model is of the form

$$\log \frac{Pr(Y = 1/X)}{1 - Pr(Y = 1/X)} = \alpha + \beta X,$$

where the random vector X consists of $X_1, X_2, ..., X_n$ an their interaction terms. The stepwise selection procedure starts from a null model. At each step, it adds a variable with the most significant score statistics among those not in the model, ten sequentially removes the variable with the least score statistic among those in the model whose score statistics are not significant. The process terminates if no further variable can be added to the model or if the variable just entered into the model is the only variable removed in the subsequent elimination. Here, the score statistic measures the significance of the effect of a variable.

2.3.5 Performance measurement: ROC curve

Receiving Operator Characteristic (ROC) curve (Zhou et al.) is a graphical representation used to assess the discriminatory ability of a dichotomous classifier by showing the tradeoffs between sensitivity and specificity for every possible cut off. Sensitivity is calculated by dividing the number of true positives (TP) by the number of all positives,

which is equal to the sum of the true positives and the false negative (FN). Specificity is calculated by dividing the number of the negative (TN) by the number of all negatives which equals the sum of the true negatives and the false positive (FP).

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

The ROC curve plot $1 - specificity$ on the X-axis and $sensitivity$ on the Y-axis.
A good classifier has its ROC curve climbing rapidly towards upper left hand corner of the graph. This can also be quantified by measuring the area under the curve. The closer the area is to 1, the better the classifier is while the closer the area is to 0.5, the worse the classifier is. We also use the ROC curve to compare the performance of these learning algorithms.

2.4 Results

Adaboost
With adaboost, the bootstrap error rate in the model rises to 9%. This was evaluated by doing 100 runs, each base learner receives different training set of n instances. With this sampling called bootstrap replication (Efron and Tibsirani (1993)), it has observed that, on average 36.8% of training instances are not used for building each tree but as test set, and then averaging over the test set errors. We choose the four most important variables which are HT, NEU, DELTAPLT, HYGRO by ranking the variables according to their frequency weighted by the position of the node.

Bagging

The results mostly agree with the bagging: Two similarity first features with adaboost are found to be very important. The bootstrap error rate in the model rises to 6%. This was evaluated by doing 100 runs. Bagging gives HT, HYGRO, PLT, and NEU as being the fours must important variables (see graph 1).

Boosting Tree Ranking ranking bagging

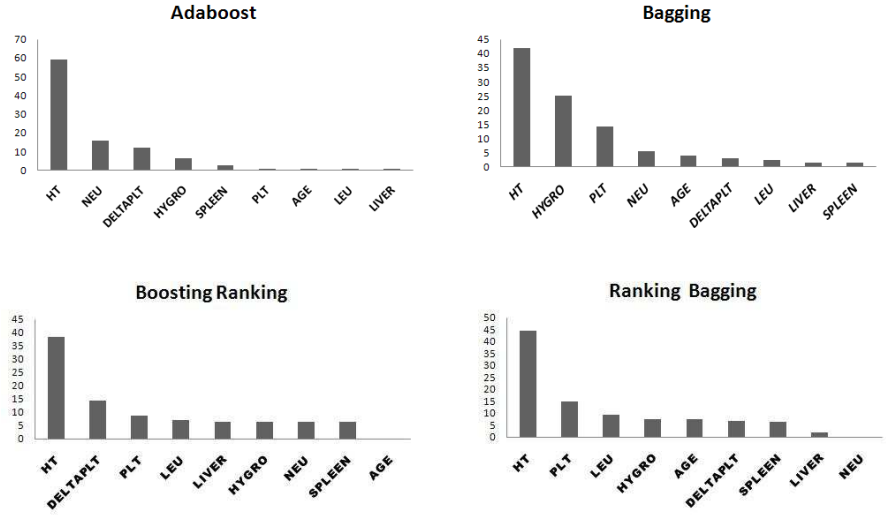
The existing methods for growing decision trees typically use splitting criteria based on error/accuracy or discrimination. In this method we use an AUC splitting criterion. This algorithm build a binary tree-structured scoring function. This method mimics the idea recursive approximation procedure of the optimal ROC curve.
Boosting ranking and ranking bagging use this tree as base learner and the same construction with adaboost and bagging respectively to obtain the finals methods.

Logistic Regression We use the likelihood ratio to test the significance of predictor variables included in the model. The likelihood ratio is given in the following equation:

$$D = -2\ln[l(B)].$$

This statistic, D, is known as the deviance. First, we want to determine the deviance of the model without any of the predictor variables (i.e with the intercept only), and compare this value with that of the model consisting of difference combinations of variables. As we add variables, we can evaluate the p-value of the deviance, which tests for the significance of that particular combination of predictor variables. A low p-value justifies the rejection of the null hypothesis, which is that all of he beta coefficients are equal to zero (i.e the all of the predictor variables are independent of the response variable). The rejection of the null hypothesis means that the variables included in the model are significant.
Backwards regression is the method we will use to find the overall best model. That is, we will run a model in R software containing all of the predictor variables and then analyze the p-value of the predictor variables to determine which of them should be removed. We will also note the deviance of the model, along with its associated p-value.
To refine the model, we remove several of the factors exhibiting the highest p-value the reevaluate the model. We repeat the removal process until we get to the point where the remaining predictor variables are all significant at the level $\alpha = 0.05$ level HT, PLT are the remaining risk variables.
Using the this model, we calculated the predictive error rate of logistic regression on the data set and obtained 11.36%. This was evaluated by doing 100 bootstraps and then averaging over the test set errors. The results presented show that all these methods

Figure 1: Variables Importance



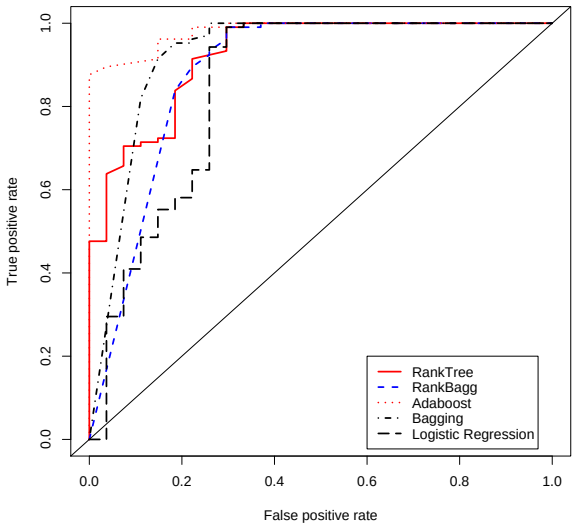
discriminate very well the severe and mild crisis. This is an indiction that there are clear distinction between the two class in terms of the 9 attributes that were used. The results of adaboost and bagging coincide in terms of what the most important variables are. The first four variables according to adaboost and bagging are HT, NEU, DELTAPLT, HYGRO. Also the the three first variables chosen by bagging ranking are ranked between the four most important variables according to boosting tree ranking.

2.5 ROC curves

The ROC curves constructed when we use these methods show that these methods improve on the default hypothesis (which would correspond to the diagonal line $TP = FP$). Moreover these curves confirm that Adaboost and RankTree give the best prediction.

Methods	AUC
Adaboost	0.92%
RankTree	0.90 %
RankBagg	0.88 %
Bagging	0.87 %
Logistic regression	0.80%

Table 2: Area under a ROC curve (AUC).



Conclusion

In this paper, we have explored Adaboost, Ranktree, Rankbagg, Bagging and logistic regression methods for diagnostic of severity of ASSC, a serious symptom in cycle cell

disease. The five methods were evaluated on diagnostic problem at a fixed level of haemoglobin to classify patient between mild and severe symptoms. The first interesting observation is the high accuracy of all methods. In particular the Adaboost and Ranktree achieved respectively 92% and 90% AUC, thus providing a very good model of the diagnostic process. This indicated that we can choose $HT, NEU, DELTAPLT$ and $HYGRO$ medical diagnostic to predict severity of ASSC.

References

[1] A Liaw and M. Wiener, Classification and regression by randomForest. Rnews, 2/3:1 8-22, Decembre 2002

[2] Breiman, L.,Friedman, J., Olshen, C. (1984). Classification and Regression Trees. Wadsworth, Belmont, California.

[3] Breiman,L. (1996) Bagging predictors. Machine learning, 24, 123-140

[4] Breiman,L. (2001) RandomForests. Machine learning, 45, 5-32

[5] Breiman, L. (2003). How to Use Survival Forests. Technical Report, Department of Statistics, University of California, Berkeley, URL:www.stat.berkeley.edu/breiman.

[6] Clémençon S., Lugosi G. and Vayatis N. (2005). Ranking and scoring using empirical risk minimization. In Proc. of COLT Bertino, Italy, June 27-30, 2005. Lecture Notes in computer Science 3559 Springer,1-15.

[7] Clémençon S., Lugosi G. and Vayatis N. (2006). From Classification to Ranking: a Statistical View. In Proc. of the 29th Annual Conference of the German Classification Society, GfKl 2005, 'Studies in Classification, Data Analysis and Knowledge Organization' series, Vol. 30. Springer-Verlag.

[8] Clémençon S., Lugosi G. and Vayatis N. (2007). U-processes and nonparametric scoring. To appear in Annals of Statistics, available at https://hal.archives-ouvertes.fr/hal-0020087.

[9] Efron, B. and Tibishirani, R.J 1993. An introduction to the bootstrap. London: Chapman & Hall.

[10] Freund, Y. an Schapire, R. E. (1997). A decision-theoritic generalization of online learning and an application to boosting. Journal of Computer and System Sciences, 55 (1). 119-139.

[11] Goddard, M.J. and Hinberg I. (1990). Receiver Operator Charcteristic curves and non-normal data: an empirical study. Statistics in Medicine, 9, 213-238.

[12] Hastie, T., R. Tibishirani, and J. Friedman (2003). The Elements of Statistical Learning. New York: Springer.

[13] Hsieh, F.and Turnbull B.W. (1996). Nonparametric Methods for Evaluating Diagnostic Tests. Statistica Sinica, 6, 47-62.

[14] Pepe, M.S. (2003). The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford: Oxford University Press.

[15] Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. Science, 240, 1285-1293.

[16] Swets, J.A. and Pickett R.M.(1982). Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. Academic Press, New York.

[17] Zhou X.H., Obuchowski N. and Obuchowski D. (2002). Statistical Methods in Diagnostic Medicine. New York: Wiley & Sons: 2002.