SciMedCentral

JSM Mathematics and Statistics

Research Article

L_1 Penalized logistic Regression Feature Preselection for Classification Tree: Application to the Diagnostic of Severe Imported Malaria Based on Heterogeneous Data

Luca Talenti[1], Margaux Luck[2], Nicolas Argy[3], Sandrine Houzé[3], Anastasia Yartseva[2] and Cecilia Damon[2]*

[1]*Ecole Centrale Paris, Grande Voie des Vignes, France*
[2]*Institut HyperCube, 51 Esplanade du Général de Gaulle, France*
[3]*Laboratoire de Parasitologie – Mycologie / CNRPalu, Hopital Bichat-Claude Bernard, France*

***Corresponding author**
Cecilia Damon, Institut HyperCube, 51 Esplanade du Général de Gaulle, 92907 Puteaux, France, Tel: 33-6-27-88-42-39; Email: cecilia.damon@institut-hypercube.org

Submitted: 20 October 2015

Accepted: 02 December 2015

Published: 04 December 2015

Keywords
- Decision trees
- Logistic regression
- Imported malaria
- Diagnosis

Abstract

The actual criteria for the classification of the different forms of imported malaria are complex and do not take into account the heterogeneity of the individual profiles. Multivariate classification methods using explanatory and predictive models are necessary to characterize groups with a high risk of developing severe forms of imported malaria. We investigate two standard approaches implementing two different strategies: L_1 logistic regression that models a single global solution, which is a linear combination of a subset of the input features, and classification trees that models multiple local solutions corresponding to discriminate sub regions of the feature space. As an alternative to pruning, which limits the complexity of the decision tree by removing unstable branches once the model is built, we explore an original approach known as L_1 LR-Tree, which combines the two previous strategies. This combined method constrains the dimension of the initial set of features before fitting the decision tree. Through the comparison of these methods, we aim to study the influence of biological and epidemiological factors dismissed in the current clinical/biological picture used for the diagnosis of severe forms of imported malaria based on a heterogeneous population of 353 patients. This method builds sparse and stable models that can significantly differentiate the patients with moderate imported malaria from those severely affected, and the severely affected from the critically affected (p-values≤0.01 for recall and specificity scores). Moreover, it outperforms all the other methods in terms of accuracy by achieving a 70% correct classification rate through a leave-one-out evaluation. The combined models identify new risk profiles that may be useful in order to improve the diagnosis and patient treatment.

ABBREVIATIONS

GB: White Blood Cells; L_1LR: L_1 penalized Logistic Regression.

INTRODUCTION

Despite the decrease in malaria cases in endemic areas since 2000 [1], the increasing number of travelers between endemic regions and Western countries favors the entry of malaria strains (called imported malaria) in areas that would not otherwise be affected. Out of all European countries, metropolitan France is considered to have the highest risk of exposure. The mortality rate of imported malaria is strongly correlated to severe malaria form and with delays in access to health care. The World Health Organization (WHO) has drawn up a concrete definition of the different clinical and biological criteria for severe malaria in order to speed up the diagnosis and health care of patients that require urgent and intensive care units [1]. However, the clinical/biological picture that infers this diagnosis is based on multiple criteria, complex and does not take into account epidemiological information. Indeed, in contrast to endemic regions in Africa, the populations of patients with imported malaria are heterogeneous and composed of first generation migrants (born in endemic regions and residing in France), second generation migrants (born in France to immigrant parents) and African or European travelers (adults or children with a different history of malaria and genetic background).

Risk factors for developing severe malaria that are currently not included in the clinical/biological picture of the diagnosis have been investigated in the context of heterogeneous patient populations with classical univariate statistical methods [2]. However, these methods have revealed their limits indistinguishing risk factors in subgroups of patients as they can only assess the statistical association between each factor and the

severe form of malaria independently of each other and without identifying a predictive relationship. The use of explanatory multivariate classification tool share essential and still under researched to efficiently characterizing groups with a high risk of developing severe imported malaria or effectively reducing the mortality of *Plasmodium falciparum* infection in France based on multi-source data [3].

Therefore, we explored and compared two off-the-shelf multivariate predictive methods based on two model-learning strategies: L_1 logistic regression and decision trees [4,5]. Both methods attempt to capture the relationship between binary response variables and a set of heterogeneous explanatory variables.

The choice of the L_1 regularization for logistic regression aims to reduce the risk of over fitting induced by potential co-linearity and the combinatorial exploration of all two-way possible interactions [6,7]. L_1 logistic regression uses linear combinations of explanatory variables to learn a single decision boundary and build an easily understandable linear model. It selects a subset of discriminant features and assesses the predictive contribution of each of them in the model. However, it only considers linear interactions between features and their global variation related to the binary outcome. Moreover, it does not take into account missing data.

The decision tree approach is non-parametric, non-linear and particularly helpful when exploring which feature subspaces are predictive of a class of subjects [8,9]. It learns multiple decision boundaries parallel to feature axes and builds easy-to-interpret models under the form of a set of if-then-else decision rules. It also handles data as complex as missing values, numerical and categorical data, multi colinear variables, outliers and local relationships among variables. However, decision trees could generate over-complex and locally optimal solutions increasing the over fitting risk and unstable decision trees due to small variations in the data.

Pruning is generally applied to avoid the over fitting phenomenon. Recent studies have tried to combine logistic regression and decision trees, particularly by applying a logistic regression model at the leaves of the decision trees in order to smooth the final model as an alternative to the standard pruning [10].Other methods, such as Random Forest and Gradient Boosting, also called ensemble methods, have been proposed to avoid over-complex solutions and unstable decision trees [5]. We do not investigate these methods since they do not allow building explanatory models that are easily interpretable and that could be used in practice for diagnosing imported malaria. Another popular and efficient way to produce more accurate and stable classifiers is the feature selection method [11].

In this study, we proposed and tested an original and innovative approach, known as L_1LR-Tree, that combines the two previous strategies. The objective of this novel approach is to fit more robust and simple decision tree learners by applying first a feature selection resulting from the L_1 logistic regression model.

In the materials and methods section, we first present the dataset and the two comparative experiments of severity forms of imported malaria, and then, we briefly explain: the three classification methods, the model selection and the evaluation methodologies. In section Results, we describe the different performance results and the learned models. Finally, in section Conclusions, we conclude on the benefit of the combined method L_1LR -Tree and the perspectives derived from this work.

MATERIALS AND METHODS

A) Dataset

The French National Reference Center of Malaria (FNRCM[1]) monitors imported malaria for epidemiological purposes through a national network of correspondents in hospital centers. In a prospective manner, demographic (age, sex, ethnic origin, medical history, history of malaria, chemoprophylaxis taken), epidemiological (native country, country of residence, visited area), clinical (disease history, severity criteria, patient management, treatment), biological (severity criteria, biochemical parameters, hematological parameters, diagnostic tools, serological status) and transcriptomic (parasite genome) data have been collected by a center at the day of the diagnosis (D0) in a secured database. All the blood samples used for the parasitological diagnosis were sent to the FNRCM laboratory. The objective of this monitoring, conducted from 2010 to 2013, is to identify higher-risk groups for the development of severe malaria. The patients included in the study are infected with *P. falciparum* species as confirmed by the FNRCM and they consists of: first-generation migrants (born in malaria-endemic regions and living in Metropolitan France), second–generation migrants (born in Metropolitan France from first-generation migrants parents and living in Metropolitan France), and travelers or expatriates (born and living in non-malaria-endemic regions). For each patient we dispose of whole blood sample available in sufficient quantity at recovery day (D0). Patients had no preventive malaria treatment within 30 days preceding the diagnosis as checked by the presence of anti malarial drugs in plasma at D0. Informed consent was not required since the sampling procedures and the examination of biological samples are part of the French national recommendations for the care and monitoring of malaria.

For this study, we removed from our analyses the variables that are directly used to infer the target, such as organ or metabolic dysfunctions and blood smear measures, all of which serve to determine the severity of the malarial infection. The choice of the variables used in this study come either from risk factors of severe imported malaria identified in previous studies (age, ethnic origin, health care delay, visited an endemic area, etc.) [2,12,13] or variables influenced by severe malaria development as platelet count, malarial history and serological status [14-16].

We also distinguished two subgroups of patients among those affected by severe malaria according to the existence of neurological and multi-organ clinical dysfunctions. The first one is called serious imported malaria and the second one is called critical imported malaria because this last form of the disease has a high probability of being fatal.

Finally, the studied dataset is composed of 353 patients diagnosed with three severity levels of imported malaria:

[1] http://www.cnrpalu-france.org/

moderate, serious or critical. For each patient we have a total of 39 features that providevarious informationof different sources (Table 1). None of them are confounders with a good repartition of the variables distribution within the different classes.

We define two experiments, each one comparing two subject groups:

I. The first experiment includes the whole dataset by comparing subjects having moderate imported malaria to those having a severe form of the disease (i.e. serious and critical). 353 subjects are included in this experiment with 202 patients having a moderate form and 151 having a severe form. The objective of this experiment is to identify risk factors predictive of severe malaria in a heterogeneous population of patients.

II. The second experiment compares subjects suffering from serious imported malaria to those displaying a critical form of the disease. 151 subjects are included in this experiment, 88 displaying the serious form and 63 displaying the critical form. The objective of this experiment is to characterize and to validate the relevance of these two subgroups among patients suffering from severe malaria.

B) Methods

In the following, we first briefly describe the two standard methods: L_1 logistic regression and decision trees, and the combined method L_1LR -Tree. We then present the methodology used for model selection and finally the procedure for assessing and comparing the different models.

B.1) L_1 logistic regression: For the classification step, we used an L_1 regularized logistic regression[2] (i.e. L_1LR), modeling the class membership probability as a linear combination of explanatory variables [6,7]. Standard logistic regression (i.e. non-penalized) estimates a binary decision function by assuming that the log it can be modeled as a linear function of features:

$$\log\left(\frac{p_\beta(\mathbf{x_i})}{1-p_\beta(\mathbf{x_i})}\right)=\eta_\beta(\mathbf{x_i}) \quad with,$$

$$\eta_\beta(\mathbf{x_i})=\beta_0+\sum_{j=1}^{q}\mathbf{x}_{i,j}^t\beta_j,\ p_\beta(\mathbf{x_i})=\mathbb{P}_\beta\left(Y=1\,|\,\mathbf{x_i}\right).$$

$Y\in\{0,1\}$ is the binary target, $X=\{x_1,\cdots,x_q\}\in R^n$ are the explicative variables, β_0 is the intercept and $\boldsymbol{\beta}\in R^q$ is the regressor vector.

The L_1 regularization parameter is introduced in the model to shrink the estimates of the regression coefficients towards zero and set some of them to zero relative to the maximum likelihood estimates:

$$\hat{\boldsymbol{\beta}}=\operatorname{argmin}_{(\beta_0,\boldsymbol{\beta})\in R^{q+1}}F_\lambda(\beta_0,\boldsymbol{\beta})=\operatorname{argmin}_{(\beta_0,\boldsymbol{\beta})\in R^{q+1}}-l(\beta_0,\boldsymbol{\beta})+\lambda\|\boldsymbol{\beta}\|_1$$

where $l\left(\beta_0,\boldsymbol{\beta}\right)$ is the log-likelihood function:

$$l\left(\beta_0,\boldsymbol{\beta}\right)=\frac{1}{N}\sum_{i=1}^{N}y_i\left(\beta_0+\mathbf{x}_i^t\boldsymbol{\beta}\right)-\log\left(1+e^{\beta_0+\mathbf{x}_i^t\boldsymbol{\beta}}\right)$$

Note that genomic data of the parasite have not been included in L_1LR method as it requires not-empty features and many subjects have missing values for these features. Several solutions are proposed to impute missing data such as replacing the missing values using the mean, median or k-nearest neighbor along the feature. However, they all represent biased estimators and our attempt to complete missing data with the median has deteriorated the performances of the models.

B. 2) Decision Trees: Decision tree analysis[3] successively splits the dataset into increasingly homogeneous subsets by binary recursive partitioning of multidimensional covariate space until it is stratified to meet a splitting criterion [4, 7].

The splitting criterion used is $SS_T-\left(SS_L+SS_R\right)$, where $SS_T=\sum_i\left(y_i-\bar{y}\right)^2$ is the sum of squares for the node and SS_R, SS_L are the sums of squares for the right and left children respectively. This is equivalent to choosing a split that maximizes the between-groups sum of squares in an analysis of variance. We constrained a minimum number of 10 observations in the leaf nodes in order to avoid over-fitting.

Note that the *rpart* implementation of the decision trees deals with missing data, which avoids the removal of the observations having at least one missing value and allows the consideration of the parasite's genomic data.

B. 3) L_1LR -Tree: The combined model consists of two steps:

1. Select a subset of features by fitting an $\mathbf{L_1}$ logistic regression.
2. Build a decision tree on the selected features.

The purpose of this combined approach is the same as the

[2]Computed in Rwith the package *glmnet*, https://cran.r-project.org/web/packages/glmnet/glmnet.pdf

Table 1: List of input variables by data type.

Input variables	
Data Types	**Variable name & description**
Demographic	Age, Sex, Caucasian (dichotomous), African (dichotomous), Chemoprophylaxis taken (dichotomous)
Epidemiological	Vis West Africa (Visit in West Africa, dichotomous), Vis Central Africa (Visit in Central Africa, dichotomous), Vis Other (Visit in another endemic country, dichotomous), Res France or other non-endemic country (Resident in France or in another non-endemic country, dichotomous)
Clinical	ATCD (history of the disease), Delay 2(days from 1st symptoms to recovery), Immunodependency (dichotomous)
Biological	GB (White Blood Cells count), Platelets (platelets count), Serology, Serological Interpretation, Titration
Transcriptomic	A1, A2, A3, B1, B2, C1, C2, BC1, BC2, Var1, Var2csa, var3 (parasite genome, i.e. expression of *var* genes)

[3]Computed in R with the package rpart. https://cran.r-project.org/web/packages/rpart/rpart.pdf

pruning tree approach: to limit the appearance of over-complex solutions and to lower the over-fitting risk. However, the penalization of these two approaches occurs in two different ways. Instead of pruning the learned decision tree to limit its size, the **L_1LR**- Tree approach prior constrains the model by reducing the dimension of input features.

B. 4) Model selection: For both methods, we applied a model selection to limit the complexity of the solution with a K-fold cross-validation. We chose $K = \left\lfloor \frac{n}{10} \right\rfloor$ to both ensure a bias-variance trade-off of the test error estimates and a sufficient representation of the two groups within the test sets [17]. In each experimental dataset, we kept the original proportion of the two classes within each fold and we used the same sample distribution for all methods.

For $\mathbf{L_1}$ logistic regression, we optimized the penalization coefficient λ so that we captured two levels of model complexity.

Therefore, we selected two values of λ: λ_{min} such that λ_{min} - L_1 LR is the best model minimizing the K-folds cross-validation mean squared error and λ_{1se} such that λ_{1se} - L_1 LR corresponds to the simplest model which is no more than one standard error worse than the best model according to the one standard error rule [18].

For decision trees, we applied a cost-complexity minimization to limit the size of the tree and we called this simplified tree, the pruned tree T_α :

$$T_\alpha = \operatorname{argmin}_T R_\alpha(T),$$
$$R_\alpha(T) = R(T) + \alpha |T|$$

The $\alpha \in [0, \infty[$ parameter defines the cost of adding another split to the model. This parameter is optimized within the K-fold cross-validation.

B.5) Evaluation: As the output is dichotomous and the two regression methods, **L_1LR** and decision trees, estimate the membership probability $\hat{y} = \hat{p}(\mathbf{x_i}) = \mathbb{P}(Y = 1 | \mathbf{x_i})$, we defined the following decision function to classify our samples:

$$\text{class}(\hat{y}) = \begin{cases} 0 & \text{if} \quad \hat{y} \le 0.4 \\ 1 & \text{if} \quad \hat{y} > 0.4 \end{cases}$$

We fixed the threshold of the decision function to 0.4with respect to the distribution of the two classes in both experiments (i.e. 0.4 corresponds to the proportion of the symptomatic patients) in order to avoid biased models due to imbalanced classes. Hence, the misclassification of the predominant class 0 is more strongly penalized.

We checked and compared the predictive power of our constructed classifiers, based on the L_1LR, the decision trees and the L_1LR -tree methods, through three performance indicators:

Recall: $\frac{\text{True Positives}}{\text{Positives}}$

Specificity: $\frac{\text{True Negatives}}{\text{Negatives}}$

Accuracy: $\frac{\text{Recall} + \text{Specificity}}{2}$

The Recall (resp. Specificity) score aims to quantify the overall rate of samples correctly classified for the second (resp. first) class. They give two complementary information about the quality of classification performances of the different methods. Indeed, from a medical point of view, we aim to discriminate and well classify the two groups of patients and not only the predominant class.

We also assessed the statistical significance of the recall and specificity scores with a binomial test with a probability of success of 0.5 and a confidence interval level of 95%. We defined three significance levels:*$p\text{-value} \le 5\%$, **$p\text{-value} \le 1\%$, ***$p\text{-value} \le 0.1\%$.

These performance indicators are computed with the leave-one-out validation method, which consists in training our models on a train set composed of all the subjects except one kept for the test set and the model evaluation. This approach classifies each patient one at a time in order to generate stable learning models (highly correlated). This choice is due to the high patient heterogeneity. We thus obtained for each method 353 models and we compared the set of features selected by these methods. A measure of their stability is given by the frequency of the input variables selected by the different methods over the 353 models generated by the leave-one-out validation. For the combined method, we were able to visualize an example of a stable decision tree giving precise information about the definition of the subgroups of patients (the threshold or range of values for quantitative variables and the category for qualitative variables).

RESULTS

In this part, we first compared the results between the standard methods, i.e. L_1LR and classification trees, and their sparse form, λ_{min} - L_1LR *vs* λ_{1se} - L_1LR and Tree *vs* Prune for both experiments: moderate *vs* severe malaria and serious *vs* critical. Then, we selected the best forms of each standard method in term of performance scores and sparsity and we combined them to build a L_1LR-Tree model. Below, we reported the results of this combined model for both experiments.

A) L_1 logistic regression- and classification trees-based models

A.1) Performance scores: Classification tree-based models (i.e. Tree and Prune) have a better accuracy than L_1-LR-based models (i.e. λ_{1se} - L_1LR and λ_{min} - L_1LR) for both experiments (Figures 1,2). The tree method outperforms the other methods with an accuracy score of 68% (resp.70%) to discriminate moderate and severe (resp. serious and critical) imported malaria. It is also the only method to have both highly significant recall and specificity scores ($pvalue \le 0.001$) for the two experiments.

Indeed, L_1LR -based models tend to well-classify the second class of the experiments, corresponding to the least represented composed of the patients with severe (resp. critical) imported malaria in the first (resp. second) experiment with significant

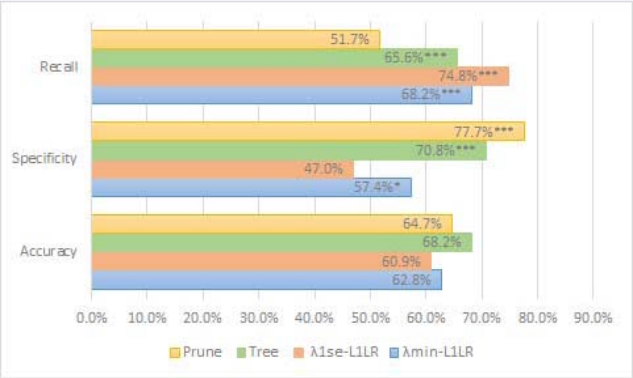

Figure 1 Moderate malaria form *vs* severe form (1st experiment): comparison of the 4 models, Tree, Prune, λ_{min} - L_1LR and λ_{1se} - L_1LR, through the three leave-one-out performance scores: recall, specificity and accuracy. For each score-bar, the color refers to the model; the length indicates the score's value (rounded percent) with its significance level for the recall and specificity scores.

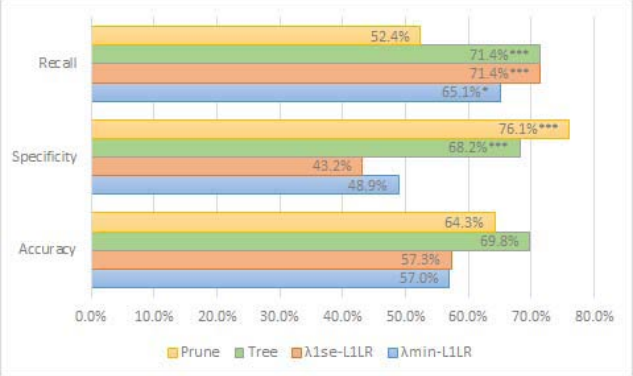

Figure 2 Serious malaria form *vs* critical form (2nd experiment): comparison of the 4 models, Tree, Prune, λ_{min} - L_1LR and λ_{1se} - L_1LR, through the three leave-one-out performance scores: recall, specificity and accuracy. For each score-bar, the color refers to the model; the length indicates the score's value (rounded percent) with its significance level for the recall and specificity scores.

recall scores while they tend to fail to classify the first class.

Conversely, prune models well-classify the first class of the experiments, corresponding to the most represented composed of the patients with moderate (resp. serious) imported malaria in the first (resp. second) experiment with the best significant specificity scores while they fail to classify the second class.

We can assume that the Tree method is more robust to unbalanced groups of samples and therefore extract discriminant decision rules efficiently generalizable to predict both classes.

A.2) Selected variables: As expected, the simpler models, namely λ_{1se} - L_1LR and Prune, include less features than the standard models, namely λ_{min} - L_1LR and Tree. For both first experiment and particularly the first one, the Tree-based models are on average sparser but less stable than the L_1LR-based models (Figures 3,4). Indeed, they capture on average less features but some of them are selected only a few times corresponding probably to locally optimal solutions.

A common pattern of selected stable features (i.e. selected almost systematically over the leave-one-out models) for all the methods is composed of White blood cells count (GB), platelets count, serological status and titration variables for the first experiment. We observed the same common pattern of selected stable features plus the age for the second experiment. Note that the serological status is a discrete feature deriving from the titration values and so they are considered as similar features.

Here, we focused on the Tree method, since it is the most powerful approach and it gives meaningful information on the models through the learned decision rules. Indeed, these latter characterize the different discriminant sub regions of the feature space specific to subgroups of subjects.

In addition to the common pattern, the three models capture the following stable features: the immune depression and sex variables for the first experiment, and the log-transformed of the expression of a subgroup of *var* genes and visit in West Africa variables for the second experiment.

Some of these results confirmed the observations of previous studies about the potential interactions between *Plasmodium falciparum* during severe malaria and the hematological components of the inflammatory response like GB [19] and platelets count [14,20] on the one hand, and on the other, the immunological protection, by previous exposure to the parasite, on the development of the different severity forms of imported malaria. The immunological protection may be represented by the serological status, visit in West Africa, malaria ATCD and membership of African ethnie [21, 15].

Furthermore, being older is a well-known risk factor to develop "very severe malaria" [22]. In [22], a statistical relationship has been reported between visiting West Africa, especially Gambia, and the risk of fatal malaria.

Concerning the impact of gender on the discrimination between moderate and severe malaria, no statistical relation has been proven between gender and malaria severity. Nevertheless, one study showed that women are more susceptible to cerebral complications than men [12]. Concerning the expression of group A *var* gene, some studies have highlighted the role of the *var* gene family in cerebral malaria [23].

B) Combined L_1LR -tree-based models

To effectively penalize the Tree method with a prior L_1LR -based feature selection step, we used the $\lambda_{1se} - L_1LR$ method.

B.1) Performance scores: The combined $\lambda_{1se} - L_1LR$ -tree method achieves similar or higher performances than the Tree ones, except for the recall score of the first experiment which is 2% inferior (Figures 5,6).

B.2) Selected variables: As the combined method builds the decision tree based on the features selected with λ_{1se} - L_1LR, it efficiently reduces the set of input features. The set of stable variables selected by the combined models corresponds to the previous observed common patterns for both experiments: GB, platelets count and serological status/titration (resp. plus age) variables for the first (resp. second) experiment (Figures 7,8).

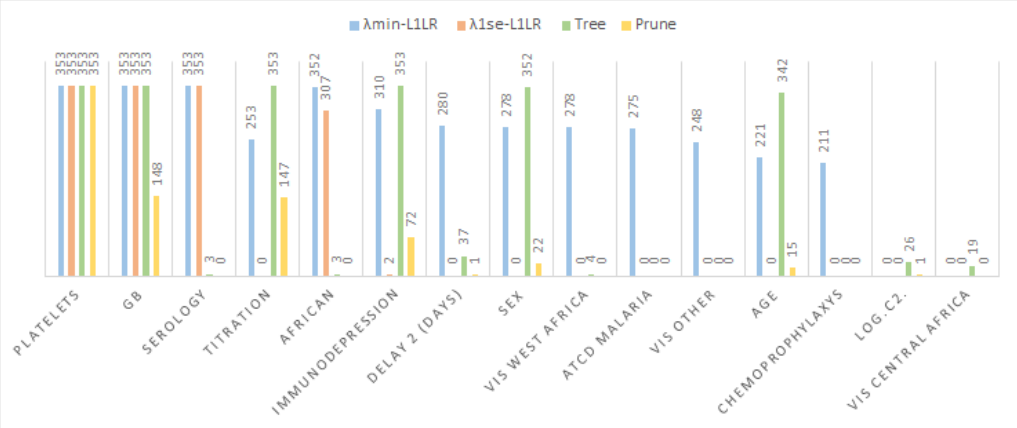

Figure 3 Moderate malaria form *vs* severe form (1st experiment): frequency of the input variables selected by the four methods (Tree, Prune, λ_{min} - L_1LR and λ_{1se} - L_1LR), over the 353 models generated by the leave-one-out validation. We only represented features selected at least 10 times.

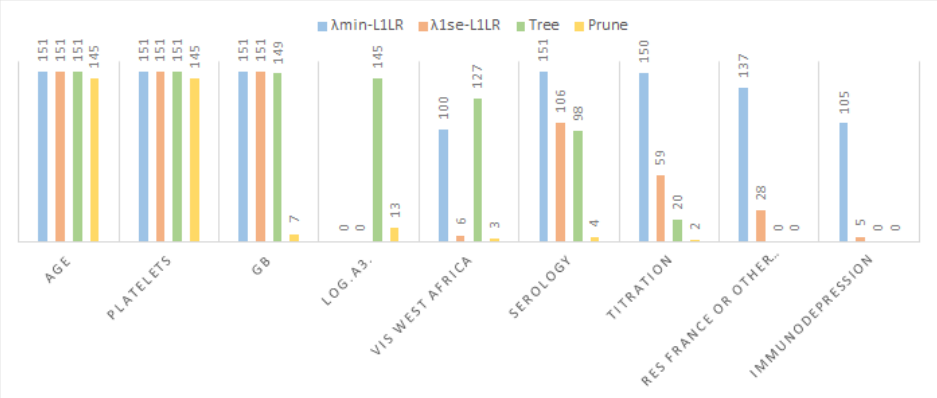

Figure 4 Serious malaria form *vs* critical form (2nd experiment): frequency of the input variables selected by the four methods (Tree, Prune, λ_{min} - L_1LR and λ_{1se} - L_1LR), over the 353 models generated by the leave-one-out validation. We only represented features selected at least 10 times.

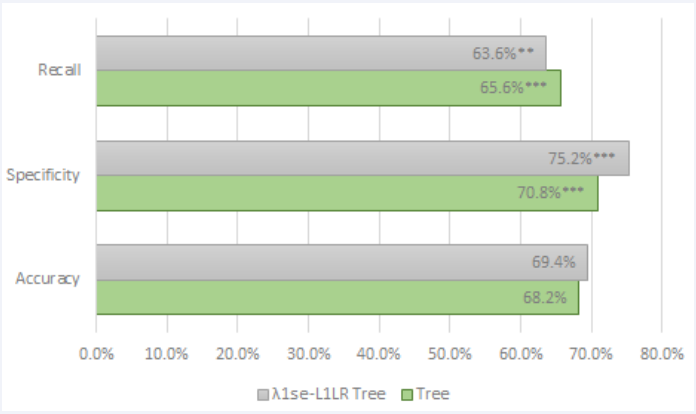

Figure 5 Moderate malaria form *vs* severe form (1st experiment): comparison of the Tree and λ_{min} - L_1LR Tree models through the three leave-one-out performance scores: recall, specificity and accuracy. For each score-bar, the color refers to the model; the length indicates the score's value (rounded percent) with its significance level for the recall and specificity scores.

Note that for the second experiment the 151 combined models have selected either serology or titration leading to a total frequency of 103 for both variables.

Therefore, the combined method led to sparser, more stable and discriminant (in terms of accuracy performances) models than those achieved by the Tree method.

Figures 9 and 10 show examples of stable $\lambda_{1se} - L_1LR$ Tree

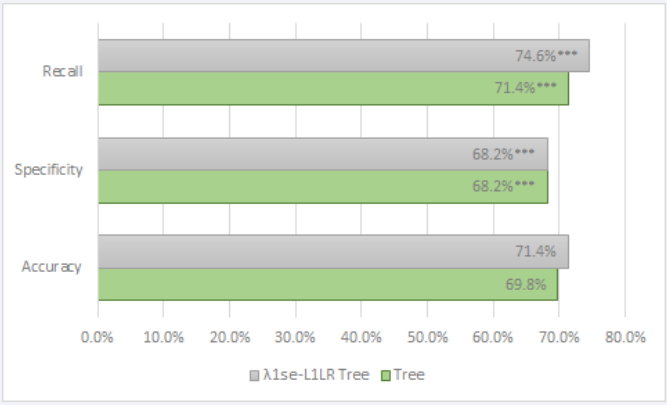

Figure 6 Serious malaria form *vs* critical form (2nd experiment): comparison of the Tree and λ_{min} - L_1LR Tree models, through the three leave-one-out performance scores: recall, specificity and accuracy. For each score-bar, the color refers to the model; the length indicates the score's value (rounded percent) with its significance level for the recall and specificity scores.

Figure 7 Moderate malaria form *vs* severe form (1st experiment): frequency of the input variables selected by the three methods (λ_{min} - L_1LR , Tree and λ_{1se} - L_1LR Tree), over the 353 models generated by the leave-one-out validation. We only represented features selected at least 10 times.

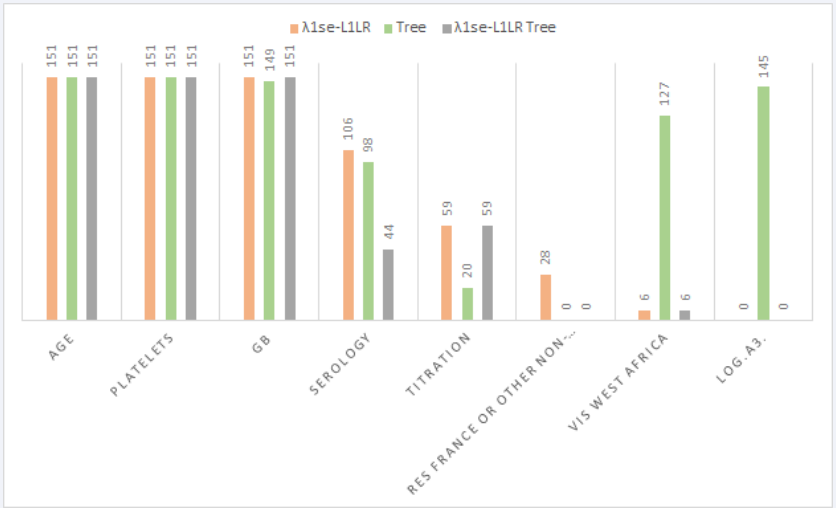

Figure 8 Serious malaria form *vs* critical form (2nd experiment): frequency of the input variables selected by the three methods (λ_{min} - L_1LR , Tree and λ_{1se} - L_1LR Tree), over the 151 models generated by the leave-one-out validation. We only represented features selected at least 10 times.

models for each experiment. We can deduce from this model the feature space sub regions predictive of the different forms of imported malaria. It seems that patients with previous exposure of malaria and a platelets count superior to 46 000 G/l are more frequently affected by the moderate form of malaria contrary to "naïve" patients (i.e. no previous malarial history) with severe thrombocytopenia (< 46000g/l) that suffer from the more severe forms of malaria. In addition, "naïve" patients with extreme ages (<16 and >44 years old) and severe thrombocytopenia (< 46000G/l) are most at risk of developing critical malarial forms.

CONCLUSION AND DISCUSSION

Among the standard approaches, i.e. L_1 logistic regression and decision trees, only the Tree method efficiently well-classify the two classes of patients for both experiments. However, the Tree models are not sparse and stable enough to provide locally optimal solutions reflecting the intrinsic heterogeneity of the studied dataset.

The pruning method drastically simplifies the Tree models while leading to poor non-significant recall scores.

This phenomenon could be explained by the fact that pruning tends to eliminate unstable branches, corresponding to variables with a great variance on threshold values and positions across cross-validation trees.

Therefore, a pre-selection of the input features can be a good alternative solution to pruning in order to constrain the complexity and to increase the robustness to small data variations of the decision trees by removing variables with local unstable phenomena in the studied population.

Our new method, called $\lambda_{1se} - L_1LR$ -tree, significantly discriminates the two classes for both experiments and outperforms all the other methods in terms of accuracies. Moreover, it efficiently leads to sparser and stable models than the Tree ones.

We can conclude that our combined method is a relevant sparse tree-based method for classification problems.

Concerning the diagnosis of the severe forms of imported malaria, the combine method correctly classifies around 70% of the patients for both studied experiments. Hence, the sub classification of the severe imported malaria in serious and critical classes is valid. Moreover, the final combined models include both hematological, immunological and demographic factors identified in some studies as being related to severe forms of imported malaria. These models propose more than a set of predictive variables, which are platelets, GB, age and previous malarial history represented by the serological status. They also provide precise insights about the representative subgroups of patients displaying moderate versus severe and serious versus

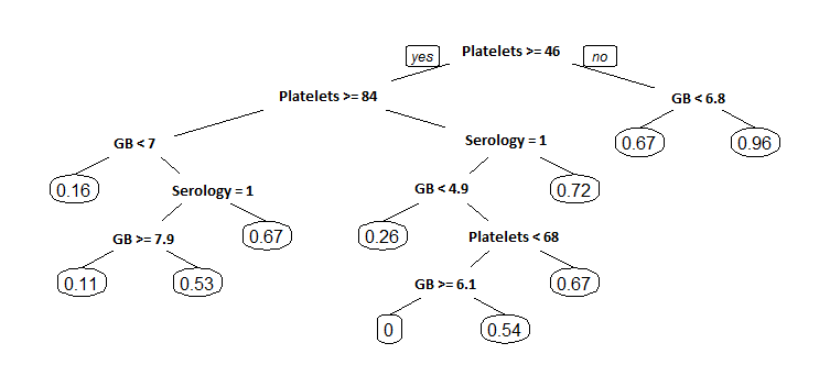

Figure 9 Moderate malaria form *vs* severe form (1st experiment): example of a stable $\lambda_{1se} - L_1LR$ -tree model. The leaves indicate the membership probability of the class severe malaria form.

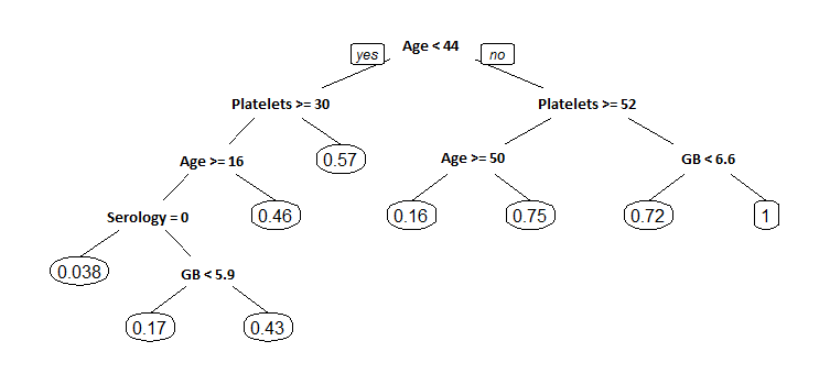

Figure 10 Serious malaria form *vs* critical form (2nd experiment): example of a stable $\lambda_{1se} - L_1LR$ -tree model. The leaves indicate the membership probability of the class critical malaria form.

critical malarial forms by fixing thresholds and range of values for quantitative variables and categories for qualitative variables. Moreover, the identified biological and clinical factors may be quickly available during the diagnostic process. Concerning the platelets, it is the first time that a direct implication in the development of severe forms of imported malaria is assessed with fixed thresholds. As the role of the platelets is not well known, many investigations are useful for understanding their interaction mechanisms with *Plasmodium falciparum.* The **s**erological status could not be currently quickly monitored, which limits its use for a rapid diagnosis, but it may be sometimes indirectly deduced by other easy-to-evaluate factors such as malarial history and the ethnicity.

The combined models did not capture some local phenomena in a stable way, probably due to their low representation in the dataset. This may explain a part of the cases where patients were misclassified. A solution would be to expand the sample size, while ensuring the diversity of the surveyed population, in order to increase the statistical reliability of these phenomena. A part of the classification error may also result from a bias in the definition of the classes. Indeed, as explained in the introduction, the diagnosis of severe imported malaria is multi-criteria, complex and does not take into account the heterogeneity of the individual profiles. The integration of our results in the clinical-biological picture may improve the current classification**.** Moreover, we are currently considering the use of clinical variables as outcomes to provide new insights on the diagnostic criteria.

It is also important to mention that the use of the $\lambda_{1se} - L_1LR$ as a feature selection step prior to fitting the decision tree, may be challenged to overcome the limitations of the L_1LR method (linear interactions, no missing data, etc.). In future work, it would be interesting to investigate other L_1 penalized approaches and to assess the comparison of the different methods on simulated data.

REFERENCES

1. World malaria report. World Health Organization. 2014.
2. Seringe E, Thellier M, Fontanet A, Legros F, Bouchaud O, Ancelle T, et al. Severe imported Plasmodium falciparum malaria, France, 1996-2003. Emerg Infect Dis. 2011; 17: 807-813.
3. Kajungu DK, Selemani M, Masanja I, Baraka A, Njozi M, Khatib R, et al. Using classification tree modelling to investigate drug prescription practices at health facilities in rural Tanzania. Malar J. 2012; 11: 311.
4. Perlich C, Foster Provost, Jeffrey S. Simonoff. Tree induction vs. logistic regression: A learning-curve analysis. The Journal of Machine Learning Research, 2003; 4: 211–255.
5. Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. Journal of clinical epidemiology. 2013; 66: 398–407.
6. David W. Hosmer, Stanley Lemeshow, Rodney X. Sturdivant. Applied logistic regression. Wiley Series in Probability and Mathematical Statistics. 2013.
7. Park SY, Liu Y. Robust penalized logistic regression with truncated loss functions. Can J Stat. 2011; 39: 300-323.
8. Breiman L et al. Classification and Regression trees. Chapman and Hall/CRC, 1984.
9. Therneau TM, Elizabeth J. Atkinson. An Introduction to Recursive Partitioning Using the RPART Routines.Technical report cran r-project. 2015.
10. Landwehr N. Logistic model trees. Machine Learning. 59: 161–205.
11. Guyon I, Elisseeff, A. An introduction to variable and feature selection. The Journal of Machine Learning Research. 2003; 3: 1157–1182.
12. Mühlberger N, Jelinek T, Behrens RH, Gjørup I, Coulaud JP, Clerinx J, et al . Age as a risk factor for severe manifestations and fatal outcome of falciparum malaria in European patients: observations from TropNetEurop and SIMPID Surveillance Data. Clinical Infections Diseases. 2003; 36: 990-995.
13. Phillips A, Bassett P, Zeki S, Newman S, Pasvol G. Risk factors for severe disease in adults with falciparum malaria. Clin Infect Dis. 2009; 48: 871-878.
14. Lampah DA, Yeo TW, Malloy M, Kenangalem E, Douglas NM, Ronaldo D, et al. Severe malarial thrombocytopenia: a risk factor for mortality in Papua, Indonesia. J Infect Dis. 2015; 211: 623-634.
15. Bouchaud O, Cot M, Kony S, Durand R, Schiemann R, Ralaimazava P, et al. Do African immigrants living in France have long-term malarial immunity? Am J Trop Med Hyg. 2005; 72: 21-25.
16. Pistone T, Diallo A, Mechain M, Receveur MC, Malvy D. Epidemiology of imported malaria give support to the hypothesis of 'long-term' semi-immunity to malaria in sub-Saharan African migrants living in France. Travel Med Infect Dis. 2014; 12: 48-53.
17. Kohavi, R. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. 14th International Joint Conference on Artificial Intelligence. 1995; 2: 1137-1143.
18. Friedman,JH. The Element of Statistical Learning.2nd Edition.Springer Series in Statistics. 2009.
19. Berens-Riha N, Kroidl I, Schunk M, Alberer M, Beissner M, Pritsch M, et al. Evidence for significant influence of host immunity on changes in differential blood count during malaria. Malar J. 2014; 13: 155.
20. Morrell CN, Aggrey AA, Chapman LM, Modjeski KL. Emerging roles for platelets as immune and inflammatory cells. Blood. 2014; 123: 2759-2767.
21. Muwonge H, Kikomeko S, Sembajjwe LF, Seguya A, Namugwanya C. How Reliable Are Hematological Parameters in Predicting Uncomplicated Plasmodium falciparum Malaria in an Endemic Region?. ISRN TropicalMedicine, 2013; 2013: 1-9
22. Checkley AM, Smith A, Smith V, Blaze M, Bradley D, Chiodini PL, et al. Risk factors for mortality from imported falciparum malaria in the United Kingdom over 20 years: an observational study. BMJ. 2012; 344: e2116.
23. Argy N, Houzé S. Paludisme grave : de la physiopathologie aux nouveautés thérapeutiques. Journal des Anti-infectieux. 2014; 16: 13-17.

Cite this article

Talenti L, Luck M, Argy N, Houzé S, Yartseva A, et al. (2015) L_1 Penalized logistic Regression Feature Preselection for Classification Tree: Application to the Diagnostic of Severe Imported Malaria Based on Heterogeneous Data. JSM Math Stat 2(2): 1012.