Classification Approach based on Association Rules mining for Unbalanced data: Application to In-hospital Maternal Mortality in Senegal and Mali

Cheikh Ndour[1,2] & **Simplice Dossou Gbété**[2] & **Aliou Diop**[1] & **Alexandre Dumont**[3]

[1] *University Gaston Berger, LERSTAD, Saint-Louis, Senegal*

[2] *University Pau et des pays de l'Ardour, UMR CNRS 5142, Pau, France*

[3] *Research Institute for Development, UMR IRD 216, Paris, France*

1 Introduction

This work deals with the supervised classification when the response variable is binary and its class distribution is unbalanced. In such situation, standard methods such as logistic regression [6], classification tree, discriminant analysis, etc [5] do not make it possible to build an efficient classification function. They tends to focus on the most prevalent class and to ignore the other one. To overcome this short-coming of these methods that provide classifiers with low sensibility, we tackled the classification problem here through an approach based on the association rules learning. Association rules is used when dealing with large database for unsupervised discovery of local-patterns that express hidden and potential valuable relationships between input variables [4]. In considering association rules from a supervised statistical learning point of view, a relevant set of weak classifiers is obtained from which one derives a classification function that performs well. More over this approach has the advantage of allowing the identification of the patterns that are well correlated with the target class.

2 Background: basic definitions and notations

Association rules learning is a well known method in the area of data-mining. **Agrawal & al. (1993)** introduced a methodology based on the concept of strong rules for the mining of association rules from transaction data.
Let m be an integer > 1 and $1:m$ is the set of all integers from 1 to m. Let's denote $(A_h)_{h=1:m}$ a set of m attributes that describe the elements of a population Ω, each attribute A_h being evaluated on a non numerical scale made of q_h levels $\left(a_{h,j(h)}\right)_{j(h)=1:q_h}$. A transaction is a sample unit $t \in \Omega$.

Definition 1. An item is a binary variable $A_{h,j(h)}$ such that $A_{h,j(h)} = 1$ if and only if $A_h = a_{h,j(h)}$

Definition 2. An itemset is a collection of items $\left(A_{h,j(h)}\right)_{h\in I}$ where $I \subset 1:m$. The length of the itemset $\left(A_{h,j(h)}\right)_{h\in I}$ is equal to the size of the set $I \subset 1:m$ and then a k-itemset is an itemset of length k.

Definition 3. Two itemsets $\left(A_{h,j(h)}\right)_{h\in I}$ and $\left(A_{h,j(h)}\right)_{h\in J}$ are disjoint itemsets if I and J are disjoint subsets of $1:m$. (i.e $I \cap J = \emptyset$)
They are nested itemsets if I and J are nested subsets of $1:m$. (i.e $I \subset J$)

Definition 4. Let's consider two disjoint itemsets $X = \left(A_{h,j(h)}\right)_{h\in I}$ and $Y = \left(A_{h,j(h)}\right)_{h\in j}$. An association rule is an implication of the form $X \to Y$ meaning that the probabilities $\Pr\left\{\left[\prod_{h\in I\cup J} A_{h,j(h)} = 1\right]\right\}$ and $\Pr\left\{\left[\prod_{h\in J} A_{h,j(h)} = 1\right] \mid \left[\prod_{h\in I} A_{h,j(h)} = 1\right]\right\}$ are significant.

Definition 5. Let's consider an association rule $X \to Y$ where $X = \left(A_{h,j(h)}\right)_{h\in I}$ and $Y = \left(A_{h,j(h)}\right)_{h\in J}$. The probability $\Pr\left\{\left[\prod_{h\in I\cup J} A_{h,j(h)} = 1\right]\right\}$ is called the support of the association rule and the conditional probability $\Pr\left\{\left[\prod_{h\in J} A_{h,j(h)} = 1\right] \mid \left[\prod_{h\in I} A_{h,j(h)} = 1\right]\right\}$ is its confidence.

Apriori is one of the most widely implemented association rules mining algorithms that pioneered the use of support-based pruning to systematically control the exponentially growth of candidate rules.

3 Classification by association rules

In the following we will consider the association rules whose consequence is reduced to the item $[Y = 1]$, indicating the modality of the target response variable Y and antecedent is an event based on an itemset only covariate.

Definition 6. Let Y be the classification target class indicator. A class association rule is an association rule of the form $X \to Y$ where $X = \left(A_{h,j(h)}\right)_{h\in I}$ is an itemset disjoint with Y.

Let's consider the classification function ϕ defined as $\phi(t) = \prod_{h\in I} A_{h,j(h)}(t)$ where $t \in \Omega$ is a transaction and $X = \left(A_{h,j(h)}\right)_{h\in I}$ is a given itemset. We have

- $TPR(X,Y) = \Pr([\phi_X = 1] \mid [Y = 1])$ (true-positive rates)
- $FPR(X,Y) = \Pr([\phi_X = 1]^c \mid [Y = 1]^c)$ (false-positive rates)

Definition 7. Let's consider an itemset $X = \left(A_{h,j(h)}\right)_{h\in I}$ where $I \subset 1:m$ and denote Y the indicator of the target class of a binary outcome. The relative risk of Y given the itemset X is the following probabilities ratio

$$\mathrm{RR}(X,Y) = \frac{\Pr([Y = 1] \mid [\phi_X = 1])}{\Pr([Y = 1] \mid [\phi_X = 1]^c)}$$

The itemset $X = \left(A_{h,j(h)}\right)_{h\in I}$ is a risk pattern for Y if the relative risk $RR(X,Y)$ exceeds a given threshold $\tau > 1$.

Processing data with the apriori algorithm usually produces a huge number of association rules, certainly more than it is necessary to build a classification function that is efficient and easy to implement. It is suitable to bring out some basic principles which could help to pruning association rules that generate very weak classification functions. To this end one will pay attention to the subset of rules whose the risk patterns are nested.

Definition 8. Let $U = \left(A_{h,j(h)}\right)_{h\in I}$ and $U' = \left(A_{h,l(h)}\right)_{h\in J}$ be two itemsets such that $I \subset J$. The itemset U' is redundant if the classification function generated by the class association $U \to Y$ has better performance measures.

Therefore the true-positive rate and the true-negative rate are sorted in the opposite way for the classification functions generated by two risk patterns if one of them is nested in the second one. This provides a criterium for pruning redundant risk pattern. Moreover on can state:

Proposition 1. *Let $X = \left(A_{h,j(h)}\right)_{h\in I}$ and $X' = \left(A_{h,j(h)}\right)_{h\in J}$ be two itemsets such that $I \subset J$ and both association rules $X \to Y$ and $X' \to Y$ are valid. If* $\Pr\{[\phi_X = 1],[Y = 1]\} = \Pr\{[\phi_{X'} = 1],[Y = 1]\}$ *then*

1. $\Pr([\phi_X = 1]^c \mid [Y = 1]^c) \leq \Pr([\phi_{X'} = 1]^c \mid [Y = 1]^c)$
2. $RR(X,Y) \leq RR(X',Y)$

It comes from the statement above that in case of equality of the true-positives rates of the classification functions generated by two nested risk patterns, not only the sparsest has the smallest false-positives rate but it has also the smallest relative risk. Therefore one can perform a statistical hypothesis testing where the null hypothesis $\Pr([\phi_U = 1], [Y = 1]) = \Pr([\phi_{U'} = 1], [Y = 1])$ is considered against its opposite and discard the risk pattern U if the null hypothesis is accepted for some U'.

Proposition 2. *Let* $X = \left(A_{h,j(h)}\right)_{h\in I}$ *and* $X' = \left(A_{h,j(h)}\right)_{h\in J}$ *be two itemsets such that* $I \subset J$ *and both association rules* $X \to Y$ *and* $X' \to Y$ *are valid. If* $\Pr\{[\phi_X = 1]^c, [Y = 1]^c\} = \Pr\{[\phi_{X'} = 1]^c, [Y = 1]^c\}$ *then*

1. $\Pr([\phi_X = 1] \mid [Y = 1]) \geq \Pr([\phi_{X'} = 1] \mid [Y = 1])$
2. $RR(X,Y) \geq RR(X',Y)$

This property has been pointed out first in Jiuyong Li & al. as the antimonotonic property [5]. Besides the statement that the relative risks of two nested patterns are in the same decreasing order as their sizes when their false-positives rates are equal, we can perform a statistical hypothesis testing where the null hypothesis $\Pr([\phi_{U'} = 1], [Y = 1]^c) = \Pr([\phi_U = 1], [Y = 1]^c)$ is considered against its opposite and discard the risk pattern U' and all the patterns generated by U (containing U) that are nested in U' if the null hypothesis is accepted for some U.

Proposition 3. *Let* $X = \left(A_{h,j(h)}\right)_{h\in I}$ *and* $X' = \left(A_{h,j(h)}\right)_{h\in J}$ *be two itemsets such that* $I \subset J$. *If* $\Pr\{[\phi_X = 1]\} = \Pr\{[\phi_{X'} = 1]\}$ *then the both following equalities holds:*

1. $\Pr\{[\phi_X = 1], [Y = 1]\} = \Pr\{[\phi_{X'} = 1], [Y = 1]\}$
2. $\Pr\{[\phi_X = 1], [Y = 1]^c\} = \Pr\{[\phi_{X'} = 1], [Y = 1]^c\}$

We have from the two first propositions above that the classification functions generated by the rules $U \to Y$ and $U \to Y$ have the same performance. In this case, the best classification function is the one with fewer parameters (the shortest). Therefore we can perform a statistical hypothesis testing where the null hypothesis $\Pr([\phi_U = 1]) = \Pr([\phi_{U'} = 1])$ is considered against its opposite and discard the risk pattern U' if the null hypothesis is accepted for some U.The comparison of nested risk patterns for pruning needs to carry out statistical hypothesis testing.

We perform a statistical hypothesis testing where the null hypothesis $H_0 = \Pr([Y\Phi_U = 1]) - \Pr([Y\Phi_W = 1])$ is considered against its opposite. We will consider the random variable defined by $Z = \mathbb{1}_{\{[Y\Phi_U=1]\}} - \mathbb{1}_{\{[Y\Phi_W=1]\}}$. One considers the hypothesis testing defined as follows: rejection of H_0 if $Z > 0$ with probability 1 or if $Z = 0$ with probability α.

After the pruning phase the statistical learning procedure is performed by using a validation dataset. The family $\mathscr{F}(Y)$ of the class association rules produced at the pruning stage is screened in order to select representative risk patterns. The first step in this stage consists in updating estimation of the relative risk for each pattern. After for each record in the target class identify all the risk patterns that describe the record and select as candidate one of the risk patterns whose relative risk value is maximum and was not already retained. Then a classification function is stated by combining these class association rules in such a way that an observation is classified as positive if it fits at least one risk pattern, and a negative one.

The final stage of the procedure deals with the assessment of the classification function and is performed on a test dataset. The performance of the classifier was assessed using the sensitivity and the specificity statistics.

4 Application to maternal mortality in Senegal and Mali

4.1 Classification using association rule approach

The data under consideration in this application has been gathered by using a randomized and controlled trial (trial QUARITE). The hospital is the unit of randomization and intervention while the patient admitted for childbirth is the unit of analysis. Only the patients' data are analysed in this paper. 89518 patients of the available sample are described by 25 variables split in three groups: a first group of seven variables describing the state of the patient status before the current pregnancy, a second group of eleven variables dealing with the progress of the pregnancy and a third group of seven variables describing the course of delivery. The binary target variable takes the value 1 if the patient died before being enable to leave the hospital (617 patients) and 0 otherwise (88901).

Classification rules (classifier) were obtained by setting algorithm's parameters as follows: maximum length of risk pattern: 3 or 4; threshold for local support: 9%, 10% or 15%; ratio of confidence by frequency of death: 3, 4 or 5. Combining these parameters results in eighteen classifiers. The best of classifier is selected from this set of classifier by examining the variation of the sensitivity with respect to the specificity (curve ROC).

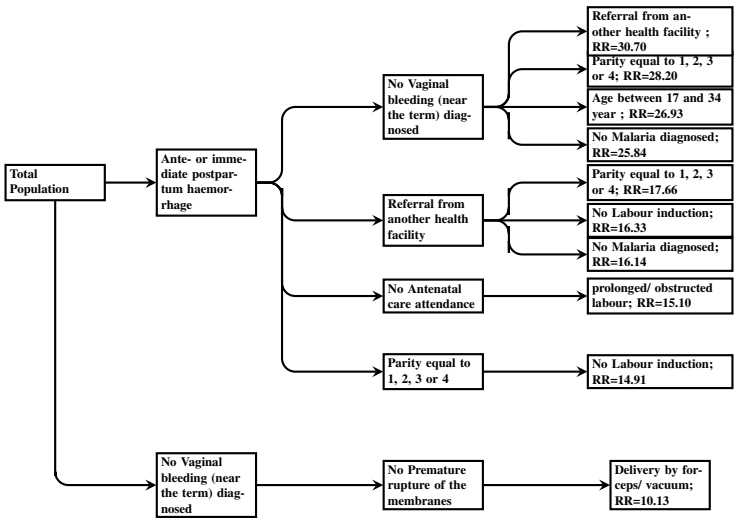

Figure 1: first part of the tree representation of rules mined

4.2 Comparison with alternatives methods

Bagging is a common way to improve a classification task on the basis of a logistic regression model or a classification tree. It consists in combining classifiers learned on bootstrap balanced samples formed by using weighted re-sampling scheme [2].

According to Table 1, bagging decision tree is less efficient than the model of classification association rules. Only the bagging logistic regression model [4] is comparable to the model of classification by association rules. The advantage of classification association rules is that it is possible to determine risk patterns and to present them in the form of tree easy to apprehend practical decision making situation (Figure 2). While it is not possible with Bagging logistic regression model.

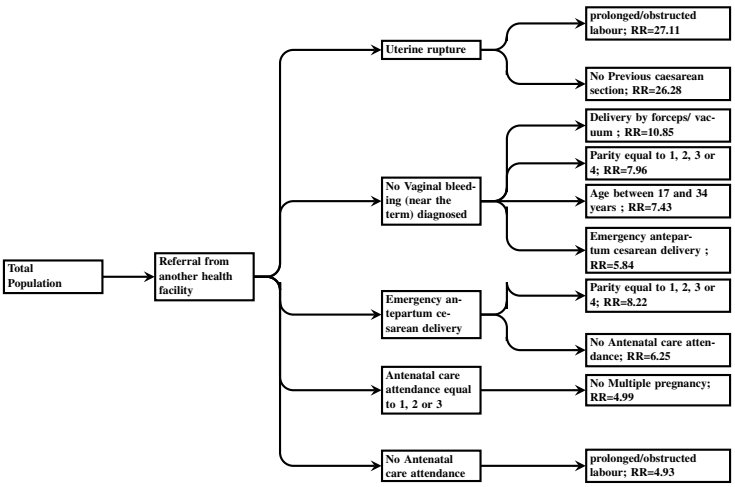

Figure 2: Second part of the tree representation of rules mined

Models		Observed				sensitivity	specificity	Global error
			Death	Not Death	Total			
A.R.M	Predicted	Death	164	5454	5618	0.824	0.816	0.184
		Not Death	35	24186	24221			
B.L.R	Predicted	Death	157	3023	3180	0.789	0.898	0.103
		Not Death	42	26617	26659			
B.D.T	Predicted	Death	149	7025	7174	0.749	0.763	0.237
		Not Death	50	22615	22665			

Table 1: Performances measures for two alternatives methods classification vs class association rule method: A.R.M: Association Rules Model; B.L.R: Bagging Logistic Regression ($\alpha = 0.5$); B.D.T: Bagging Decision Tree ($\alpha = 0.5$)

References

[1] Agrawal R., Srikant R. (1993). Fast algorithms for mining association rules. *VLDB-94.*

[2] Breiman L., (1996). Bagging predictors. *Machine Learning,* **24**(2),123 – 140.

[3] Hahsler M., Gruen B., Hornik K.(2005). arules: A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software,* **14**(15).

[4] Hualin W., Xiaogang S.(2010). Bagging probit models for unbalanced classification. *IGI Global,* **ch017***:*290 – 296.

[5] Li J., Ada W. F., Fahey P.(2009). Efficient discovery of risk patterns in medical data. *Artificial Intelligence in Medecine,* **45***:*77 – 89.

[6] Menardi G., Torelli N. (2010). Training and assessing classification rules with unbalanced data. *Working Paper Series,* **2**.

[7] Scarpa B., Torelli N. (2005) Selection the training set in classification problem with rare event. *Studies in Classification, Data Analysis, and Knowledge Organization,* 39 – 46