Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007)

Induction Tree methods to classify *M. tuberculosis* spoligotypes

Georges Valétudie
GRIMAAG - Université Antilles-Guyane,
BP 97110 Pointe-à-Pitre Guadeloupe
georges.valetudie@univ-ag.fr

***Abstract*— In this paper we compared and analyzed four graph induction methods to automatically classify spoligotypes. A spoligotype is a sequence of 43 binary values provided by a DNA analysis technique. This method is known to be useful and efficient to many supervised learning problems. We found it interesting to use these techniques especially for sequential data, in order to create a classifier based on one decision rule per class.**

I. INTRODUCTION

The direct Repeat (DR) locus of *Mycobacterium tuberculosis* is a suitable model to study molecular epidemiology and the evolutionary genetics of tuberculosis. This is achieved by a DNA analysis technique of genotyping called *spoligotyping* (*sp*acer *oligo*nucleotide *typing*) [1].
However, applied on a large scale, existing genetic methods analysis are very expensive, particularly for the third-world, mainly concerned by this problem. So, a real challenge would be to define new efficient, fast and low cost methods. Databases exist (see http://www.cdc.gov/ncidod/EID/vol8no11/02-0125-Table.htm), but their size increases constantly. So, our main goal in this paper is to investigate accurate data mining methods for a new application in such a context. We consider our approach as a sequences mining problem. Knowledge discovery from sequential data has been the objective of a lot of research over the last decade, particularly in the field of bioinformatics. But our sequences are small strings of 43 binary values, and a good challenge was to apply efficient supervised methods not specially adapted to sequential data. We've achieved this goal by applying Induction Tree methods, which have been successfully applied in different domains, as it provides suitable models. The aim of the work presented in this paper is to study a new application of this model for spoligotyping of *M. tuberculosis*, and to prove that classification could be achieved with spacers information. We want to show in witch way Decision Tree method is suitable for DNA researchers specialized on *Mycobacterium tuberculosis* complex analysis, as it provides simple, rapid and easily understandable decision rules. We achieved this by comparing some state-of-the-art methods in order to provide an accurate classifier for these data. In section 2, we present briefly the spoligotyping method. In section 3, we give a synthetic presentation of the model. In section 4, we present our experimental results and evaluate the classification accuracy of the obtained classifier. In section 5, we conclude and outline some future prospects.

II. THE SPOLIGOTYPING METHOD

Spoligotyping [1] is a DNA analysis technique of genotyping. It has been applied to characterize *Mycobacterium tuberculosis* complex strains and has gained increased international acceptance as a rapid first line discriminatory test. It is based on polymorphism in the chromosomal direct repeat locus, which represents a variable number of direct repetition (DVR). Each DVR is composed of a short repeated sequence and of a short non-repetitive one called *spacer*. A *spoligotype* is the result of spoligotyping method, representing the absence or presence of these spacers by a vector composed of 43 binary values (Fig.1).

This feature is of double interest in DNA data analysis of M. Tuberculosis : on one hand, it exploits a binary alphabet which simplifies the data-processing treatments and allows the black square/white square standardized presentation, very useful for an immediate visual analysis usually applied by the researchers. On the other hand, this presentation is composed of fixed length sequences, that solve the comparison problem of variable DNA sequences length, and allows supervised methods like Decision Tree, not specifically planned for sequence analysis.

III. THE MODEL

Among the different learning models, decision trees are particularly useful to produce rapid and easily interpretable decision rules, and this is the case for a majority of supervised learning problems. It consists in a succession of partitions on a population by optimizing a criterion which measures the quality of the obtained partition. This partition is accepted if it improves the value of the criterion employed, but is rejected otherwise. Many criteria exist, but they depend

Fig. 1. Example of spoligotype and its binary sequential representation

directly on the method used. They can be listed as following [1]:

- The *information gain* criterion, based on the *Shannon entropy* [11], with the formula :

$$\mathrm{I}(S_p) = \sum_{j=1}^{k} \frac{n_{.j}}{n} \left(-\sum_{i=1}^{m} \frac{n_{ij}}{n_{.j}} \log_2(\frac{n_{ij}}{n_{.j}}) \right) \quad (1)$$

  This criterion popularized by [3], allows us to measure the amount of information provided by the variable used to split.

- The *ratio gain* :

$$\mathrm{IGR}(S_p) = \frac{-\sum_{i=1}^{m} \frac{n_{i.}}{n} \log_2(\frac{n_{i.}}{n}) + \sum_{j=1}^{k} \frac{n_{.j}}{n} \sum_{i=1}^{m} \frac{n_{ij}}{n_{.j}} \log_2(\frac{n_{ij}}{n_{.j}})}{-\sum_{j=1}^{k} \frac{n_{.j}}{n} \log_2(\frac{n_{.j}}{n})} \quad (2)$$

  This criterion is also proposed by [3], but penalizes the multiple splits which bring out no more information.

- The *GINI gain*, a criterion popularized by [4] :

$$\mathrm{IG}(S_p) = \sum_{j=1}^{k} \frac{n_{.j}}{n} \sum_{i=1}^{m} \frac{n_{ij}}{n_{.j}} \left(1 - \frac{n_{ij}}{n.j} \right) \quad (3)$$

In fact, the criterion used for the tree construction is selected according to the corresponding method :

- the method based on C4.5 algorithm [12] is among the references. It proceeds by successive splits using the *gain ratio* criterion. The expansion is stopped when the variation gain is negative. This method applies a pruning step with an estimation of the real error on the training set. Pruning can be done on the tree or the decision rules set.

- CART method [4] is a reference also, and proceeds to the splits with the *index of Gini* criterion. It builds binary induction trees until the maximum size, i.e. when the impurity gain measured is lower than a fixed threshold (zero to obtain the finest partitions). Pruning is applied next, by estimation of the real error with a test sample. It should be noted that if the number of examples are not large enough to create this sample-test, it is possible to apply a cross-validation to evaluate the real error.

- ID3 method is also proposed by [3], and proceeds to the splits by using the *information gain* criterion. C4.5 is originally based on it, but differs by various extensions, particularly the pruning phase, not present in the classic formulations of this algorithm.

  SIPINA method [6] is particular, as it is a generalization of the existing methods. Its characteristic is to allow non-arborescent graphs construction with any of the criterion presented above. It proceeds by fusion or split, and stops the expansion as soon as the variation of measurement applied is negative. There is no post-pruning.

Each rule is extracted by just following a path in the tree, from the root to a leaf. Finally, our classifier is built with a combination of these different rules. So, each path in the resulting graph corresponds to a decision rule, according to the propositional logic formalism :

$$\textit{If} < \textit{condition} > \textit{then} \Rightarrow < \textit{conclusion} >$$

But the quality of the resulting classifier depends directly on rules one. Indeed, each rule is provided with some indicators in order to assess its accuracy. These indicators are all the more useful as one prefers to reject an assumption instead of keeping an unreliable conclusion. There is :

- the *number of examples*, used to evaluate the interest of the rules, and to define a hierarchy among them according to the number of examples concerned, particularly when different rules contribute to the same conclusion.

- the *sensitivity*, which supplements the preceding indicator, by evaluating the proportion of class examples concerned by the rule, in comparison with the number of examples in the class.

- the *accuracy*, to evaluate the proportion of correctly classified examples among those concerned by the rule. It supplements and refines the preceding indicators, but it is insensitive to the variation of the number of examples.

- the *implication level*, an index inspired by [10] which measures the degree of implication of the rule in the class definition, taking into account the premises. It is based on a statistical test, which compares the Hypothesis *H0* of an independence between conditions and conclusion, to a situation where the rule would randomly classify the individuals. A value close to 0 means independence, whereas a value close to 1 means a full implication.

- the *J-measure index*, proposed by [9] which indicates the interest of a rule, by measuring its ability to predict the class concerned, using the condition part of the rule. It

[1] S_p : current partition, m : number of classes, n : sample size, k : partition size, n_{ij} : number of class i instances in sub-group j, $n_{i.}$: number of class i instances, $n_{.j}$: size of the sub-group j.

indicates the degree of correlation between the *a priori* probability distribution on the class prediction *P*, and the *a posteriori* distribution when the conditions are known. The rule is all the more interesting as the value of the index is high, corresponding to a good predictive ability.

IV. Experimental results

As presented in [7] and refined and supplemented in [15], we use nine classes for the complex *M. Tuberculosis* : Afri, T, Beijing, EAI, Haarlem, LAM, CAS, X and Bovis. Our analysis continues these previous works with new data, and some families improved in quantity and quality. These classes were *a priori* defined by experts, highlighting some existing relations between spoligotypes that belonged to the same class (Fig. 2). They noticed that Beijing was particularly characterized by the absence of the spacers 1 to 34, T by the absence of spacers 33 to 36, EAI by the absence of 29 to 32 and 34 and presence of 33, LAM by the absence of 21 to 24 and 33 to 36, CAS by the absence of 4 to 7, X by the absence of 18 and 33 to 36, and Bovis by the absence of 39 to 43.

Although expert rules already exist, our objective to build trees is justified by three main reasons :

- to test the performances of the Expert rules,
- to try to simplify these rules composed sometimes with too many spacers for them to be easily used,

CLASS	DESCRIPTION
AFRI	*Spacers 8, 9 or 10 and 39 absent, 34 and 43 present.*
T	*One of the spacers 1-30 present, 33-36 absent, one of the spacers 21-24 present, 18 and 31 present, 9 or 10 present.*
BEIJING	*Spacers 1-34 absent.*
EAI	*Spacers 29-32 and 34 absent, 33 present, one of the spacers 1-28 present.*
HAARLEM	*Spacers 31 and 33-36 absent, 32 present, one of the spacers 1-30 present.*
LAM	*Spacers 21-24 and 33-36 absent, 18 and 31 present, one of the spacers 1-20, 25-30 present.*
X	*Spacers 31 present, 33-36 and 18 absent, sometimes related to 39-42.*
BOVIS	*Spacers 39-43 absent, 33-36 present.*
CAS	*Spacers 4-7 and 23-34 absent, 18 present.*

Fig. 2. Expert rules

TABLE I
DATA SET DAT1

Class	Size	Percentage
AFRI (1)	42	5 %
T (2)	226	29 %
BEIJING (3)	11	1 %
EAI (4)	99	13 %
HAARLEM (5)	94	12 %
LAM (6)	176	22 %
X (7)	52	7 %
BOVIS (8)	56	7 %
CAS (9)	29	4 %
Total	785	100 %

- if the success rate is low and/or not in agreement with the decision trees, perhaps it will be necessary to change the label for some examples and maybe discover new subclasses.

Our data file is composed according to the distribution presented in table I.

SIPINA for Windows software [5] was used to build the different graphs and their corresponding decision rules, with a selection of 4 induction algorithms : ID3, C4.5, CART, Sipina. For the CART method, we used a set composed with 75% of data for the learning phase. The 25% remaining used for a test-set during the pruning phase.

To choose the method, we proceed to a cross-validation procedure, which has the advantage to provide a reliable estimation of the real error and performance in generalization, directly from the training sample. More precisely, it's a stratified cross-validation, with 10 sub-groups as suggested by [14], in order to have a good estimation of the real error and confidence interval. The results are presented in table II. We note that the methods have similar average performances, but could be dissociated by the average number of rules. CART and C4.5 provide the weakest values, particularly CART for which the variation observed on average from one sub-group to another is weakest. More precisely, we can notice that the average quality of ID3 is slightly better than the other methods, but by doubling the average rules number, comparatively to CART. However, if CART provides the lowest average rules number, its performances are slightly lower than the others. In the end, C4.5 represents a good alternative as it provides good performances (identical to SIPINA, very close to ID3 and better than CART) and a reasonable average number of

TABLE II
STRATIFIED CROSS-VALIDATION WITH 10 SUB-GROUPS

Method	Average quality	Bias (%)	Average rules number	Bias
Id3	96 %	3	30.2	1.47
C4.5	95 %	3	19.9	1.76
Cart[Gini]	94 %	3	15.4	0.8
Sipina-[Fusbin]	95 %	3	30.2	1.47

rules (much lower than ID3 and SIPINA, and close to CART) in spite of the bias value.

Note that classification accuracy on the whole training data is not a way to assess the quality of a model (On this sole basis, methods as K-NN can easily obtain 100% accuracy rate) unlike the cross-validation procedure, which might be useful to choose a learning algorithm. But taking into account the cross-validation results, we could have a more precise idea about the methods selected through the quality of the built tree and by the analysis of the resubstitution error.

So we use a classification matrix in order to calculate the global success rate and error rate, and an average of these results for the cross-validation procedure. We use the number of levels in the tree to complete this comparison. Each method is evaluated through the induction tree. We analyze the quality of the partition by a confusion matrix, and we extract the indicators presented in table III. The better the induction tree is, the higher the global success rate (i.e. a weak error rate). This fact induce the presence of unclassified examples and few levels in the tree. We can notice a special column in this table corresponding to the undefined rate. It represents the situation where no class could be set to a leaf as at least two classes could be chosen. In this case, two solutions appears : to set it randomly or to keep it undefined (i.e. the class remains unspecified). This latter solution was adopted using a specification threshold : if at least 75% of the examples are not present in a leaf, we decide to keep it undefined. Results are presented in table III.

We could see that, in comparison with Sipina and ID3, CART and C4.5 methods produce a tree with few unclassified examples and the smaller number of levels, particularly for the CART method. If we consider the error rate, the tree built with CART is strongly penalized comparatively to ID3 and C4.5, which obtain both close results and the lowest rates. C4.5 method obtains the best global success rate.

Finally, three methods seem to be particularly interesting : ID3 with the second best success rate (94,27%) for a minimal error rate, CART with a good success rate and few unclassified examples, and C4.5 with the best success rate, an error rate close to the performances of the tree obtained with ID3, and the weakest unclassified examples rate.

ID3 induction tree generates many unclassified examples, and minimizes the error risk by this way. This is an important argument for us, as we prefer to reject an example wether than let it badly classified. But in our objective, the method should provide good results too. C4.5 induction tree give the best success rate for a weak error rate (close to ID3 error rate), with the weakest rate of unclassified examples. It means that the major part of the examples presented are correctly treated. If we consider these methods together, we could note a difference of 2.80% between the two unclassified examples rates with a 0.56% error rate variation, whereas the difference between the two success rates is changed to 2.16%. It means that a majority of the examples unclassified by ID3 induction tree are accepted and correctly classified by C4.5 induction tree. In addition, C4.5 induction tree is better than CART on all the criteria, except a second place for the tree level number. C4.5 method confirm to be a good choice as it gives better results on classification quality for the induced tree, best success rate, one of the weakest error rate, and a reasonable tree level number. This allows us to predict rather concise decision rules. This method is used for the next stage of the process, i.e. the extraction of decision rules from the tree and their evaluation (cf. table IV), with the following observations :

Spacer 34 is useful to separate AFRI and BOVIS class from the others. These two classes were easily separated by spacer 43 and more precisely by spacer 33.

Spacer 33 is useful to isolate the EAI class.

Spacer 36 is useful to isolate CAS and BEIJING classes, and they are easily separated by spacer 12.

Spacer 22, 24, 21 are useful to separate LAM class.

Spacer 31 and 32 are useful to isolate HAARLEM and X class, and they are easily separated by spacer 17.

For the T class, spacers 31 and 18 are useful to isolate them.

The following rules are presented according to the model defined before :

if spacers 43 and 34 are present, then Class = AFRI.

if spacers 18, 22 and 31 are present, and 33,34, 36 are absent, then Class = T.

if spacers 12 and 33 are absent, and 36, 39 are present, then Class = BEIJING.

TABLE III

METHODS PERFORMANCES

Method	Global success rate	Error rate	Undefined rate	Maximum tree level number
ID3	94.27	1.4	4.33	9
C4.5	96.43	2.04	1.53	8
Cart[gini]	94.11	4.04	1.85	6
Sipina-[Fushbin]	92.61	4.33	3.06	11

TABLE IV

C4.5 RULES PERFORMANCES

Class	Condition part size	Size	Sensitivity (%)	Accuracy (%)	Implication level(%)
Afri	2	32	88	97	100
T	5	163	97	97	99.94
Beijing	4	8	72.73	100	99.96
Eai	2	99	99	99	100
Haarlem	4	87	93	100	100
Lam	5	178	100	99	100
X	4	50	90	94	100
Bovis	3	57	100	98	100
Cas	4	29	97	97	100

TABLE V

PERFORMANCES BY TESTING THE LEARNING SET (RESUBSTITUTION)

Method	Global success rate	Error rate
ID3	96.94	3.06
C4.5	97.32	2.68
Cart[gini]	95.16	4.84
Sipina[Fushbin]	95.92	2.68

if spacer 33 is present and 34 absent, then Class = EAI.

if spacers 22 and 32 are present, 31, 33 are absent, then Class = HAARLEM.

if spacers 21, 22, 24, 34, and 36 are absent, then Class = LAM.

if spacers 12 and 36 are present, 33 and 34 absent, then Class = CAS.

if spacers 17 and 22 are present, 18 and 33 absent, then Class = X.

if spacers 33 and 34 are present, and 43 absent, then Class = BOVIS".

Comparatively to the Expert rules, the average simplification rate is 48% with C4.5 rules. The evaluation of these rules is presented in table IV with various criteria among those presented before.

According to these results, the rules are globally relevant with an implication rate and a precision generally close to 100 %.

In the next stage, we use the learning set as a test set to evaluate the *resubstitution* error. By this way, we can take into account the variability induced by this sample, and evaluate its incidence on the quality of the induction trees (cf. table V). C4.5 gives the weakest error rate estimation, but the best total success rate.

In the next stage, we estimate the quality of the classifier obtained with C4.5, by testing its ability to recognize 333 new examples of a test sample presented in table VI. Table VII shows the results. The global success rate is about 89% and

TABLE VI

TEST SET DAT2

Class	Size	Percentage
AFRI	22	7 %
T	99	30 %
BEIJING	7	2 %
EAI	46	14 %
HAARLEM	53	16 %
LAM	61	18 %
X	25	8 %
BOVIS	10	3 %
CAS	10	3 %
Total	333	100 %

TABLE VII

TEST PHASE RESULTS

Class	Correct classification		Misclassification	
	Size	Percentage	Size	Percentage
AFRI	17	6 %	5	14 %
T	89	30 %	10	27 %
BEIJING	0	0 %	7	19 %
EAI	39	13 %	7	19 %
HAARLEM	48	16 %	5	14 %
LAM	61	21 %	0	0 %
X	23	8 %	2	5 %
BOVIS	9	3 %	1	3 %
CAS	10	3 %	0	0 %
Total	296	100 %	37	100 %
Percentage	89 %		11 %	

confirm rather well the good quality of the classifier, although 11% of the examples remain misclassified, principally due to BEIJING and T class. To resolve these difficulties, we could take into account others analysis such as the possibility proposed by the Expert to use other types of data, the MIRU-VNTR [16]. Indeed, the former work presented by [17] underlined the interest of these markers in genetic epidemiology. They deserve to be exploited in our field.

Note that we obtain an unbiased estimator of the real error by using a new test set, but with a high variance, unless having a sample large enough (about 1000 examples according to [13]).

In order to have a better estimation measure of the decision tree rules quality, we made a comparative analysis with the Expert rules. If we consider at first, the observations concerning the number of premises used for the determination of a class, we can already note that the rules obtained with C4.5 method are more concise (reduction of the number of descriptors) and more accurate. So, we evaluate the ability of each method to recognize a file of 333 examples recently labeled and unexploited for the training phase. This file, called "Dat3", keeps the same distribution as "Dat2".

We present in two confusion matrices the results obtained by a test phase on a new datafile "Dat3", with the Expert rules (Fig. 3) and the rules extracted from C4.5 decision trees method (Fig. 4). The following notation is used :

Und : number of unlabeled examples, *Ind* : number of multi-labeled examples, *Txp* : success rate by class (sensitivity), *Txerr* : error rate, *Txind* : multi-labelled examples rate.

In fact, the Expert rules are penalized as they do not only cover the class examples concerned. We note a real improvement in the global success rate, from 59% with the expert rules, to 80% for the rules obtained with C4.5. By a finer analysis of these rules with the sensitivity measure (Txp), we could note a clear improvement in the quality results for some classes : LAM, X, BOVIS and CAS. The classes T, EAI and HARLEM provide similar results, and the BEIJING class performances are not significant. Only the AFRI class keeps better performances with the expert rules.

Txs = 59%

	1	2	3	4	5	6	7	8	9	Ind	Und	Txp(%)	Txer(%)	Txind(%)
1	20	0	0	0	0	0	0	0	0	2	0	90.91	0	9.09
2	0	49	0	0	4	0	0	0	0	0	14	49.49	9.09	27.27
3	0	0	0	0	0	0	0	0	0	7	0	0	0	100
4	0	0	0	45	0	0	0	0	0	1	0	97.83	0	2.17
5	0	0	0	0	50	0	0	0	0	3	0	94.34	0	5.66
6	0	0	0	0	0	26	0	0	0	35	0	42.62	0	57.38
7	0	0	0	0	0	0	2	0	0	23	0	8	0	92
8	0	0	0	0	0	0	0	5	0	4	1	50	0	40
9	0	0	0	0	0	0	0	0	0	10	0	0	0	100

Fig. 3. Confusion matrix with Expert rules

Txs = 80%

	1	2	3	4	5	6	7	8	9	Ind	Und	Txp(%)	Txer(%)	Txind(%)
1	13	1	0	0	0	0	0	2	1	0	5	59.09	18.18	0
2	0	58	0	0	0	1	1	0	0	0	39	58.59	2.02	0
3	0	0	1	0	0	1	0	0	0	0	5	14.29	14.29	0
4	0	0	0	44	0	0	0	0	2	0	0	95.65	4.35	0
5	0	0	0	0	48	1	0	0	0	0	4	90.57	1.89	0
6	0	0	0	0	0	61	0	0	0	0	0	100	0	0
7	0	0	0	0	0	0	22	0	0	1	2	88	0	4
8	0	0	0	0	0	0	0	10	0	0	0	100	0	0
9	0	0	0	0	0	0	0	0	9	0	1	90	0	0

Fig. 4. Confusion matrix with C4.5 decision rules

V. Conclusion and perspectives

In this paper, we compared and analyzed the contribution of four induction graph methods in the study of spoligotypes, data composed with 43 binary values representing the absence or presence of non-repetitive short sequences in DNA.

The induction tree is a very popular method, with a good efficiency and a great adaptability to a majority of supervised learning problems. It is particularly interesting for its interpretation simplicity, and the direct decision-making that it generates. The decision rules produced by C4.5, the method selected after a comparative study with three other methods (CART, ID3 and SIPINA), allows us to classify the spoligotypes with a good success rate, although some classes remain uneasy to classify because of their great diversity. A comparison with some Expert rules showed a real improvement of performances from a 59% success rate to 80%. The relevant and easily understandable rules automatically generated by a data-mining approach, are appreciated for their simplicity and their quick results. The financial aspect of such a fast and simple analysis is real, particularly when a good predictive accuracy is coupled with a small quantity of data.

However, there are some classes with misclassifications due to their great diversity and the presence of noise, or because of the necessity to create new classes or subclasses. Anyway, a new study was proposed to evaluate the contribution of other unexploited data by a data-mining approach of MIRUS [8].

This approach remains sensitive to overfitting, and for certain problems, it is also advisable to have a consequent population to reinforce the classifiers obtained, and to make them better suited to the generalization phase. These results encourage us to perform a largest scale analysis and to study the impact of pre-selection methods in the improvement of the quality of the corresponding classifier, and the possibility of cooperative methods such as ensemble tree methods, as they could produce better accuracy than a single one.

References

[1] J. Kamerbeek and L. Schouls and M. Van Agterveld and D. Van Soolingen and A. Kolk and S. Kuijper and A. Bunschoten and R. Shaw and M. Goyal and J. Van Embden, *Simultaneous detection and strain differentiation of mycobacterium tuberculosis for diagnosis and epidemiology*, J. Clin. Microbiol., 1997.

[2] C. E. Shannon and W. Weaver, *The mathematical theory of communication*, University of Illinois Press, 1949.

[3] J. R. Quinlan, *Induction Decision Tree*, Morgan Kaufmann, 1986.

[4] L. Breiman and J. H. Friedman and R. A. Olshen and C. J. Stone, *Classification and regression tree*, Chapman and Hall, 1984.

[5] R. Rakotomala and D. A. Zighed, *Graphes d'induction*, Paris Hermes, 2000.

[6] D. A. Zighed and J. P. Auray and G. Duru, *SIPINA : Méthode et logiciel*, Edition Alexandre Lacassagne, 1992.

[7] M. Sebban and I. Mokrousov and N. Rastogi and C. Sola, *A data-mining approach to spacer oligonucleotide typing of Mycobacterium tuberculosis*, Bioinformatics, No. 18, pp. 235-243, 2002.

[8] S. Ferdinand and G. Valétudie and C. Sola and N. Rastogi, *Data mining of Mycobacterium tuberculosis complex genotyping results using mycobacterial interspersed repetitive units validates the clonal structure of spoligotyping-defined families*, Research in Microbiology, 2004.

[9] R. M. Goodman and P. Smyth, *Information-theoretic rule induction*, European Conference on Artificial Intelligence, 1988.

[10] I. C. Lerman and G. Gras and H. Rostam, *Elaboration et évaluation d'un indice d'implication pour données binaires*, Mathématiques et Sciences Humaines, No. 5, vol. 74, pp. 5-35, 1981.

[11] C. E. Shannon and W. Weaver, *The mathematical theory of communication*, University of Illinois Press, 1949.

[12] J. R. Quinlan, *C4.5 : Program for Machine Learning*, Morgan Kaufmann, 1992.

[13] J. Catlett, *Megainduction : machine learning on very large databases*, University of Sydney, 1991.

[14] R. Kohavi, *A study of cross-validation and bootstrap for accuracy estimation and model selection*, International Joint conference on Artificial Intelligence - IJCIA'95, 1995.

[15] I. Filliol and R. Jeffrey Driscol and Dick Van Soolingen and N. Barry Kreiswirth and G. Valetudie and Dang Duc Anh and R. Barlow and D. Banerjee and P. J. Bifani and K. Brudey and A. Cataldi and R.C Cooksey and V. Cousins Debby and J.W. Dale and A. Dellagostin Odir and F. Drobnieski and Guildo Engelmann and S. Ferdinand and M. Gordon and C.M. Gutierrez and W.H. Haas and H. Heersma and G. Kallenius and E. Kassa-Kelembho and T. Koivula and H. Ly Minh and A. Makristathis and C. Mammina and G. Martin and P. Mostrom and I. Mokrousov and V. Narbonne and O. Narvskaya and A. Nastasi and Niobe-Eyangoh Sara Ngo and J.W. Pape and V. Rasolofo-Razanamparany and M. Ridell and M. L. Rossetti and Fritz Stauffer and N.P. Suffys and H. Takiff and J. Texier-Maugein and V. Vincent and J.H. De Waard and N. Rastogi and C. Sola , *Global distribution of Mycobacterium tuberculosis Spoligotype*, Emerging infectuous diseases, No. 11, vol. 8, 2003.

[16] R. Frothingham and W. A. Meeker-O'Connel, *Genetic diversity in the Mycobacterium tuberculosis complex based on variable numbers of tandem DNA repeats*, Microbiology, vol. 144, pp.1189-1196, 1998.

[17] P. Supply and S. Lesjean and A-L. Banuls and K. Kremer and D. Van Soolingen and M. Tibayrenc and C. Locht, *Minisatellite-based genotyping for the study of global epidemiology and population genetics of Mycobacterium tuberculosis* , indo French Sympsonium on Tuberculosis and AIDS, 2002.