A machine learning method for incomplete and imbalanced medical data

Issam SALMAN[1] and Jiří VOMLEL[2]

[1]*Department of Software Engineering*
Faculty of Nuclear Sciences and Physical Engineering
Czech Technical University in Prague
Trojanova 13, 12000, Prague, Czech Republic
issam.salman@fjfi.cvut.cz

[2]*Institute of Information Theory and Automation*
Academy of Science of the Czech Republic
Pod vodárenskou věží 4, 18208, Prague, Czech Republic
vomlel@utia.cas.cz

Abstract

Our research reported in this paper is twofold. In the first part of the paper we use standard statistical methods to analyze medical records of patients suffering myocardial infarction from the third world Syria and a developed country - the Czech Republic. One of our goals is to find whether there are statistically significant differences between the two countries. In the second part of the paper we present an idea how to deal with incomplete and imbalanced data for tree-augmented naive Bayesian (TAN). All results presented in this paper are based on a real data about 603 patients from a hospital in the Czech Republic and about 184 patients from two hospitals in Syria.

Keywords: Machine Learning, Data analysis, Bayesian networks, Missing data, Imbalanced data, Acute Myocardial Infarction.

1 Introduction

Acute myocardial infarction (AMI) is commonly known as a heart attack. A heart attack occurs when an artery leading to the heart becomes completely blocked and the heart doesn't get enough blood or oxygen. Without oxygen, cells in that area of the heart die. AMI is responsible for more than a half of deaths in most countries worldwide. Its treatment has a significant socioeconomic impact.

One of the main objectives of our research is to design, analyze, and verify a predictive model of hospital mortality based on clinical data about patients. A model that predicts well the mortality can be used, for example, for the evaluation of the medical care in different hospitals. The evaluation based on mere mortality would not be fair to hospitals that treat often complicated cases. It seems better to measure the quality of the health care using the difference between predicted and observed mortality.

A related work was published by [1]. The authors analyze the mortality data in U.S. hospitals using the logistic regression model. Other work was published by [2]. The authors compare different machine learning methods using a real medical data from a hospital.

2 Data

Our dataset contains data about 787 patients characterized by 24 variables. 603 patients of them are from the Czech Republic [2] and 184 are from Syria. The attributes are listed in the Table 1. Most of the attributes are real valued, four attributes are nominal. Only a subset of attributes was measured for the Syrian patients.

Most records contain missing values, i.e., for most patients only some attribute values are available. The thirty days mortality is recorded for all patients. In the Czech Republic the results of blood tests are reported in millimoles per liter of blood. In Syria some of the measurements are reported in milligrams per liter and some in millimoles per liter. We standartize all measurements to the millimoles per liter scale.

We will note $\mathbf{U} = \{X_1, X_2, \ldots, X_m\}$ for a discrete domain, where $X_i, i \in \{1, 2, \ldots, m\}$ is a discrete attribute and take on values from a finite set, denoted by $Val(X_i)$. We use capital letters such as X, Y, Z for attribute names, and lower-case letters such as x,y,z to denote specific values taken by those variables. Sets of variables are denoted by boldface capital letters such as $\mathbf{X}$,$\mathbf{Y}$,$\mathbf{Z}$ and assignments of values to the variables in these sets are denoted by boldface lowercase letters $\mathbf{x}$,$\mathbf{y}$,$\mathbf{z}$. A classified discrete domain is a discrete domain where one of the attributes is distinguished as class. We will use $\mathbf{U}_C = \{A_1, A_2, \ldots, A_n, C\}$ for a classified discrete domain. A dataset $D = \{\mathbf{u}_1, \ldots, \mathbf{u}_N\}$ of instances of $\mathbf{U}_C$, where each $\mathbf{u}_i, i \in \{1, \ldots, N\}$ is a tuple of the form $(a_i^1, \ldots, a_i^n, c_i)$ where $a_i^1 \in Val(A_1), \ldots, a_i^n \in Val(A_n)$ and $c_i \in Val(C)$. Also we note that the class is always known, and a missing value in the dataset is denoted by NA.

Table 1: Attributes

Attribute	**Code**	**type**	**value range in data**	**Country**
Age	AGE	real	[23, 94]	SYR, CZ
Height	HT	real	[145, 205]	CZ
Weight	WT	real	[35, 150]	CZ
Body Mass Index	BMI	real	[16.65, 48.98]	CZ
Gender	SEX	nominal	{male, female}	SYR, CZ
Nationality	NAT	nominal	{Czech, Syrian}	SYR, CZ
STEMI Location	STEMI	nominal	{inferior, anterior, lateral}	SYR, CZ
Hospital	Hospital	nominal	{CZ, SYR1, SYR2}	SYR, CZ
Kalium	K	real	[2.25, 7.07]	CZ
Urea	UR	real	[1.6, 61]	SYR, CZ
Kreatinin	KREA	real	[17, 525]	SYR, CZ
Uric acid	KM	real	[97, 935]	SYR, CZ
Albumin	ALB	real	[16, 60]	SYR, CZ
HDL Cholesterol	HDLC	real	[0.38, 2.92]	SYR, CZ
Cholesterol	CH	real	[1.8, 9.9]	SYR, CZ
Triacylglycerol	TAG	real	[0.31, 11.9]	SYR, CZ
LDL Cholesterol	LDLC	real	[0.261, 7.79]	SYR, CZ
Glucose	GLU	real	[2.77, 25.7]	SYR, CZ
C-reactive protein	CRP	real	[0.3, 359]	SYR, CZ
Cystatin C	CYSC	real	[0.2, 5.22]	SYR, CZ
N-terminal prohormone of brain natriuretic peptide	NTBNP	real	[22.2, 35000]	CZ
Troponin	TRPT	real	[0, 25]	CZ
Glomerular filtration rate (based on MDRD)	GFMD	real	[0.13, 7.31]	CZ
Glomerular filtration rate (based on Cystatin C)	GFCD	real	[0.09, 7.17]	CZ

3 Preliminary Statistical Analysis

For a preliminary statistical analysis [3] we selected a subset of attributes that are highly correlated with the class [5] and present in both groups, namely, we considered these variables: age, nationality, gender, STEMI location, and the class mortality.

The STEMI location encoded by 1 denotes a STEMI.inf, 2 denotes a STEMI.ant, and 3 denotes a STEMI.lat. The nationality is encoded by a binary variable, where 0 means Czech and 1 means Syrian. The Gender is encoded by a binary variable where 0 denotes a man, while 1 stands for

a female. The mortality is also encoded as a binary variable, where 0 means that the patient survived 30 days, while 1 means that he/she did not.

Already from Figure 1, where the histogram of the age values is presented, we can see that from patients that didn't survive a high percentage are young patients from Syria.

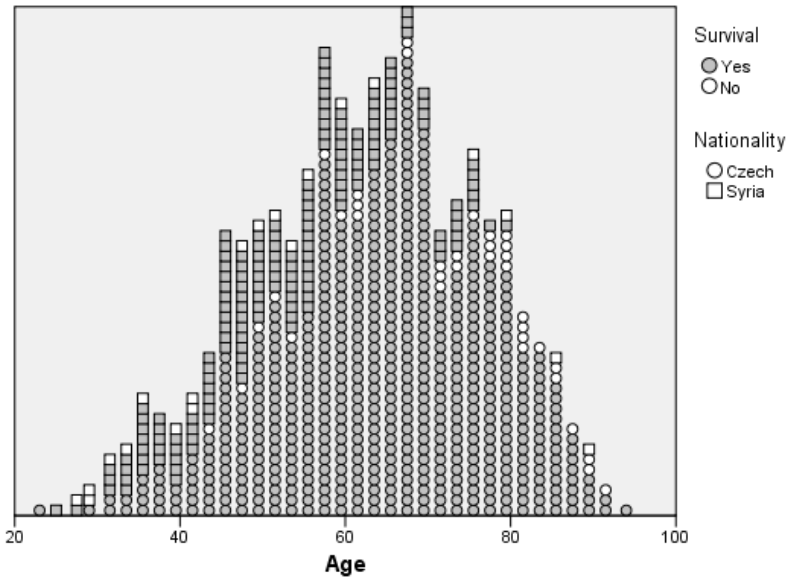

Figure 1: Histogram of the age values

Table 2: The Chi-Square Test of conditional independence

		gender	STEMI loc.	mortality	nationality
age	value	**.174**	-.010	.048	**-.381**
	sign.	**.0001**	.775	.181	**.0001**
gender	value		.022	.068	**.92**
	sign.		.53	.057	**.01**
STEMI loc.	value			-.026	-.036
	sign.			0.46	.312
mortality	value				**.089**
	sign.				**0.013**

The standard chi-square test of conditional independence between two variables reveals (see Table 2) that there is a significant dependence (at the level 0.05) between the mortality and nationality, the gender and nationality, also there are a significant dependencies between the gender and age, the mortality and gender – the patients from Syria have the lowest probability to survive, also they are younger and there is higher percentage of woman.

Finally, we learned the logistic regression model, that describes the relationship between the considered independent variables and the mortality as the dependent variable. We have got:

$$\begin{aligned} \text{logit } P(C=1|A=a) &= \beta_0 + \beta_1 a_1 + \ldots + \beta_4 a_4 \\ &= -0.034 + 0.001 \cdot a_1 + 0.027 \cdot a_2 - 0.007 \cdot a_3 + 0.065 \cdot a_4 \end{aligned}$$

where a_1: age, a_2: gender, a_3: STEMI loc, and a_4: nationality. Variables age and nationality appeared to be statistically significant for mortality prediction.

From the preliminary statistical analysis we can conclude that:

- In Syria the mortality from AIM is significantly higher than in the Czech Republic – 87.3% Syrian patients survive, while 94.7% patients from the Czech Republic survive.
- The age of patients in Syria is lower in average (the average difference is 13 years) and there is a higher prevalence of women among the patients with AIM in Syria than in the Czech Republic.
- The STEMI location is related to the mortality.

4 Machine Learning Methods

The preliminary statistical analysis studied mostly the pairwise relations only. Since the explanatory variables may combine their influence and the influence of a variable may be mediated by another variable it is worth of studying the relations of variables alltogether. Our data are incomplete and imbalanced. We will present an idea for dealing with that type of data using tree-augmented naive Bayesian (TAN).

4.1 Bayesian networks

A Bayesian network [6] is an annotated directed acyclic graph that encodes a mass probability distribution over a set of random variables $\mathbf{U}$. Formally, a Bayesian network for $\mathbf{U}$ is a pair $B = \langle G, \Theta \rangle$. The first component, G, is a directed acyclic graph whose vertices correspond to the random variables $\mathbf{U} = \{X_1, X_2, \ldots, X_m\}$, and whose edges represent direct dependencies between the variables. The graph G encodes independence assumptions: each variable X_i is independent of its non-descendants given its parents in G. The second component of the pair, namely Θ, represents the set of parameters that quantifies the network. It contains the parameter $\theta_{x_i|\Pi_{x_i}} = f(x_i|\Pi_{x_i})$ for each possible value x_i of X_i and Π_{x_i} of Π_{X_i}, where Π_{X_i} denotes the set of parents of X_i in G. Accordingly, a Bayesian network B defines a unique joint probability distribution over $\mathbf{U}$ given by:

$$f(X_1 = x_1, \ldots, X_m = x_m) \quad = \quad \prod_{i=1}^{m} f(X_i = x_i | \Pi_{X_i} = \Pi_{x_i}) \quad = \quad \prod_{i=1}^{m} \theta_{x_i|\Pi_{x_i}}$$

for each Π_{X_i} which is a parent of X_i.

4.2 Learning with Trees

A directed acyclic graph on $\{X_1, X_2, \ldots, X_n\}$ is a tree if Π_{X_i} contains exactly one parent for all X_i, except for one variable that has no parents (this variable is referred to as the root). A tree network can be described by identifying the parent of each variable [7]. A function $\pi : \{1, \ldots, n\} \to \{0, \ldots, n\}$ is said to define a tree over $X_1, X_2, \ldots, X_n$ if there is exactly one i such that $\pi(i) = 0$ (namely the root of the tree), and there is no sequence $i_1, \ldots, i_k$ such that $\pi(i_j) = i_{j+1}$ for $i \leq j < k$ and $\pi(i_k) = i_1$ (i.e., no cycles). Such a function defines a tree network where $\Pi_{X_i} = \{X_{\pi(i)}\}$ if $\pi(i) > 0$ and $\Pi X_i = \emptyset$ if $\pi(i) = 0$.

4.3 Learning Maximum Likelihood TAN

Let $\{A_1, A_2, \ldots, A_n\}$ be a set of attribute variables and C be the class variable. We say that B (Bayesian network) is a TAN model if $\Pi_C = \emptyset$ and there is a function that defines a tree over $\{A_1, A_2, \ldots, A_n\}$. The optimization problem consists on finding a tree defining function π over $\{A_1, A_2, \ldots, A_n\}$ such that the log likelihood is maximized [8] $LL(B_T|D) = \sum_{\mathbf{u} \in D} \log f(\mathbf{u})$. To learn the maximum likelihood TAN we should use the following equation to compute the parameters [8], $\theta_{a_i,\Pi_{a_i}} = \frac{N_{a_i,\Pi_{a_i}}(a_i,\Pi_{a_i})}{N_{\Pi_{a_i}}(\Pi_{a_i})}$ where $N_{a_i,\Pi_{a_i}}(a_i, \Pi_{a_i})$ stands for the number of times that attribute i has value a_i and its parents have values Π_{a_i} in the dataset. Similarly, $N_{\Pi_{a_i}}(\Pi_{a_i})$ is the number of times that the parents of attribute A_i have values Π_{a_i} in the dataset.

5 Learning TAN from incomplete data

Missing data are a very common problem which is important to consider in a many data mining applications, and machine learning or pattern recognition applications. Some variables may not be observable (i.e. hidden) even for training instances. Now more and more datasets are available, and most of them are incomplete. Therefore, we want to find a way to build a new model from an incomplete dataset. Normally, to learn the maximum likelihood TAN structure [8], we need a complete data, such that all

instances $\mathbf{u}_i, i \in \{1,\ldots,N\}$ from $\mathbf{U}_C$ are complete and don't have any missing value. In case the data are incomplete and there is an instance which has a missing value, we will not use the whole instance in TAN structure learning i.e. not use the other known values from that instance in TAN structure learning. Note that the class is always known, and a missing value in the dataset is denoted by NA. Our goal is to learn a tree-augmented naive Bayesian (TAN) from incomplete data. Some previous work by [13] propose maximizing conditional likelihood for BN parameter learning. They apply their method to MCAR (Missing Completely At Random) incomplete data by using available case analysis in order to find the best TAN classifier. In other work by [9] also deals with TAN classifiers and expectation-maximization (EM) principle for partially unlabeled data. In their work, only the variable corresponding to the class can have missing. Also, other work by [10] deals with TAN based on the EM principle, where they have proposed an adaptation of the learning process of Tree Augmented Naive Bayes classifier from incomplete data. In their work, any variable can have missing values in the dataset. The TAN algorithm can be adapted to learn from incomplete datasets, such that most available data will be used in TAN structure learning. The procedure is shown in Algorithm 1, where the Conditional Mutual Information "CMI" is defined as:

$$I(X,Y|Z) = \sum_{\mathbf{x,y,z}} f(\mathbf{x,y,z}) \log \frac{f(\mathbf{z})f(\mathbf{x,y,z})}{f(\mathbf{x,z})f(\mathbf{y,z})}$$

where the sum is only over $\mathbf{x,y,z}$ such that $f(\mathbf{x,z}) > 0$ and $f(\mathbf{y,z}) > 0$.

Algorithm 1 TAN For Incomplete Data

1: **procedure** CMI($A_i, A_j, C\}$) ▷ // Conditional Mutual Information
2: $\overline{D} = \{\overline{\mathbf{u}}_1,\ldots,\overline{\mathbf{u}}_N\}, \overline{\mathbf{u}}_m = (a_i, a_j, c), m \in \{1,\ldots,N\}$, such that $\mathbf{u}_m = (a_1,\ldots,a_n,c) \in D$
3: **Foreach** $\overline{\mathbf{u}}_m \in \overline{D}$
4: **If**$(a_i == NA | a_j == NA)$
5: Delete $\overline{\mathbf{u}}_m$ from $\overline{D}$
6: **endfor**
7: Compute $I_p = I(A_i, A_j|C)$ from $\overline{D}$
8: **return** I_p
9: **Endprocedure**
10: Read $D = \{\mathbf{u}_1,\ldots,\mathbf{u}_N\}, \mathbf{u}_m = (a_1,\ldots,a_n,c), m \in \{1,\ldots,N\}$
11: var:
12: n the number of attribute variables A;
13: $\mathbb{I}_p[n][n]$ the WeightMatrix;
14: UG the UndirectedGraph;
15: UT the UndirectedTree;
16: T the DirectedTree;
17: TAN the DirectedGraph;
18: **Foreach** $A_i, i \in \{1,\ldots,n-1\}$
19: **Foreach** $A_j, j \in \{2,\ldots,n\}$
20: $I_{p_ij} = CMI(A_i, A_j, C)$
21: $\mathbb{I}_p[i][j] = I_{p_ij}$
22: $\mathbb{I}_p[j][i] = I_{p_ij}$
23: **EndForeach**
24: **EndForeach**
25: G = ConstructUndirectedGraph($\mathbb{I}_p[i][j]$)
26: UT = MaximumWeightedSpanningTree(G);
27: T = MakeDirected(UT);
28: TAN = AddClass(T);

In Algorithm 1, on line 25 we build a complete undirected graph in which the vertices are the attributes $A_1,\ldots,A_n$. Annotate the weight of an edge connecting A_i to $A_j, i \neq j$ by $I_{p_ij} = I(A_i, A_j|C)$ One line 26 we build a subgraph from G, without any cycles and with the maximum possible total edge weight. On line 27 we transform the resulting undirected tree to a directed one by choosing a root variable and setting the direction of all edges to be outward from it. On line 28 we add the class C to the graph as a node and add edges from C to all other nodes in the graph

The idea behind Algorithm 1 is that we believe if we use more data then the estimates of conditional mutual information are more reliable.

6 Imbalanced Data

In case of imbalanced data the classifiers are more sensitive to detecting the majority class and less sensitive to the minority class. Thus, if we don't take care of the issue, the classification output will be biased, in many cases resulting in always predicting the majority class. Many methods have been proposed in the past few years to deal with imbalanced data. In our research the mortality rate of patients with myocardial infarction refers to the percentage of patients who have not survived more than 30 days, where the results are 89% of patients survive and 11% of patients do not survive, therefore the data are quite imbalanced. One of the most common and simplest strategies to handle imbalanced data is to under-sample the majority class [11, 12]. While different techniques have been proposed in the past, they did not bring any improvement with respect to simply selecting samples at random. So, for this analysis we propose the following steps:

- Let M be the number of samples for the majority class, and N be the number of samples for the minority class, and M be L times greater than N.
- Divide the instances which have majority class into L distinct clusters.
- Train L predictors, where each predictor is trained on only one of the distinct clusters, but on all of the data from the rare class. To be clear, the data from the minority class are used in the training of all L predictors.
- Use model averaging for the L learned predictors as your final predictor. i.e (in our case we will compute a conditional mutual information between each pair of attributes $(A_i, A_j), i, j \in 1, 2, \ldots, n, i \neq j$ given the class L times for each pair, in each time will use only one of the distinct clusters and all data from the minority class, then we will use the average of conditional mutual information for each pair to compute a weight matrix).

After integrating this step into the Algorithm 1, we will have a TAN algorithm which deals with an incomplete and imbalance data 2:

Algorithm 2 TAN for incomplete and imbalance data

1: var
2: M The number of samples for the majority class
3: N The number of samples for the minority class
4: D_T All instances of the majority class, $D_T \subset D$
5: D_F All instances of the minority class, $D_F \subset D$
6: integer division $L = M/N$
7: Divide D_T to L parts, $D_{T_k}, k \in \{1, \ldots, L\}$
8: **Foreach** D_{T_k}
9: $D_k = D_{T_k} \cup D_F$
10: **EndForeach**
11: Compute WeightMatrix $\mathbb{I}_{p_k}[n][n]$ foreach D_k
12: $\hat{\mathbb{I}}_p[n][n]$ = the average of $\mathbb{I}_{p_k}[n][n], k \in 1, \ldots, L$ ▷ // $\hat{\mathbb{I}}_p$ is the WeightMatrix which wwill be used in Algorithm 1
13: Continue from line 26 in Algorithm 1 using $\hat{\mathbb{I}}_p$

7 Results

For each data record classified by a classifier there are four possible classification results. Either the classifier got a positive example labeled as positive (in our data the positive example is the patient survived) or it made a mistake and marked it as negative. Conversely, a negative example may have been mislabeled as a positive one, or correctly marked as negative. Our results are summarized in Figure 2 using the ROC curves. We use the 10 fold cross validation as the model evaluation method. The ROC curve shows how the classifier can sacrifice the true positive rate (TP rate: number of positive examples, labeled as such over total positives) for the false positive rate(FP rate: number of negative examples, labeled as positive over total negatives) (1-specificity) by plotting the TP rate to the FP rate. In other words, it shows how many correct positive classifications can be gained as you allow for more and more false positives by changing the threshold.

In Figure 2 we compare our results with normal TAN ([8]) and SMOTE algorithm ([4]) for TAN. Algorithm 2 has achieved the highest area under the ROC curve (AUC) with 0.82. The results of Algorithm 1 (ROC = 0.77) is better than the normal TAN algorithm (ROC = 0.62). But SMOTE algorithm with TAN (ROC = 0.802) is better than Algorithm 1.

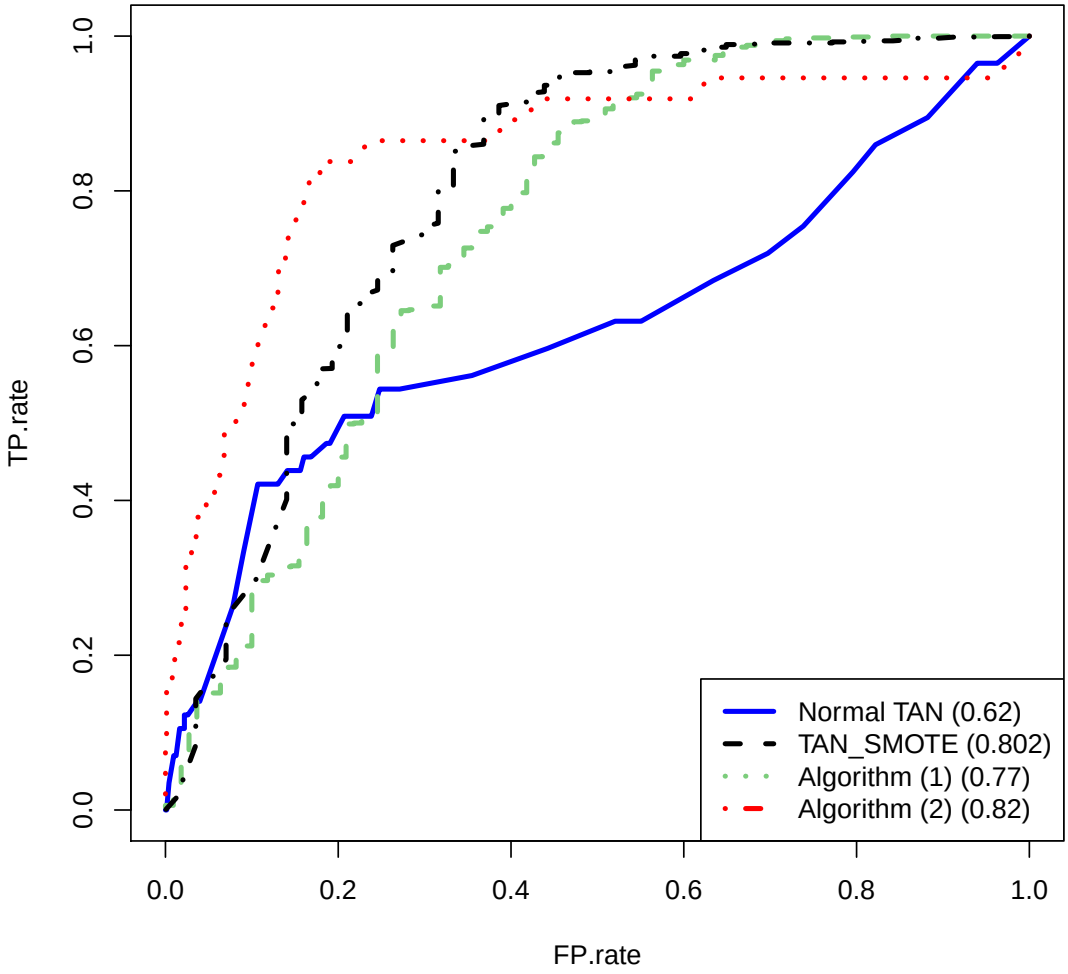

Figure 2: ROCs (TAN , TAN_SMOTI , Algorithm(1) , Algorithm(2))

8 Conclusions

First, we used medical data on patients with AIM for preliminary statistical analysis. We found a significant difference between Syrian patients and Czech patients. Second, Bayesian networks are a tool of choice for reasoning in uncertainty, with incomplete data. However, often, Bayesian network structural learning only deals with complete data. We have proposed here an adaptation of the learning process of the Tree Augmented Naive Bayes classifier from incomplete and imbalanced datasets. This methods have been successfully tested on our dataset. We have seen that our Algorithm 2 performed better than normal TAN and TAN-SOMTE.

Acknowledgement
This work was supported by the Czech Science Foundation through projects 16-12010S, and the student grant CTU SGS16/253/OHK3/3T/14.

References

[1] H. M. Krumholz, S.-L. T. Normand, D. H. Galusha, J. A. Mattera, A. S. Rich, Y. Wang and Y. Wang, *Risk-Adjustment Models for AMI and HF 30-Day Mortality, Methodology*, Harvard Medical School,

Department of Health Care Policy, (2007).

[2] J. Vomlel and H. Kružík and P. Tůma and J. Přeček, and M. Hutyra, *Machine Learning Methods for Mortality Prediction in Patients with ST Elevation Myocardial Infarction*, In the Proceedings of The Nineth Workshop on Uncertainty Processing WUPES'12, Czech Republic, 204–213, (2012).

[3] L. Wasserman. *All of Statistics*, Springer-Verlag New York, (2004).

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *Synthetic Minority Over-sampling Technique*, Journal of Artificial Intelligence Research, Volume 11, Issue 16, 321–357, (2002).

[5] M. Hall and E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and H. Witten, *The WEKA Data Mining Software: an Update*, In 'ACM SIGKDD Exploration ACM SIGKDD Explorations', Volume 11, Issue 1. (2009), 10–18.

[6] F. V. Jensen, *An Introduction to Bayesian Networks*, Springer, (1996).

[7] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, (1988).

[8] N. Friedman, D. Geiger, and M. Goldszmidt, *Bayesian network classifiers*, Machine Learning Journal, Volume 29, Issue 2. (1997). 131–163.

[9] I. Cohen and F. Cozman and N. Sebe and M. C. Cirelo and T. S. Huang, *Semi-supervised learning of classifiers: theory, algorithms and their application to human-computer interaction*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 26, Issue 12, 1553–1568, (2004).

[10] O. C. H. Francois and P. Leray, *Learning the Tree Augmented Naive Bayes Classifier from incomplete datasets*,

Third European Workshop on Probabilistic Graphical Models, 91–98, (2006).

[11] R. Laza, R. Pavon, M. Reboiro-Jato and F. Fdez-Riverola R. Laza and et al, Evaluating the effect of unbalanced data in biomedical document classification, Journal of Integrative Bioinformatics, Volume 16, Issue 3, pp. 177, (2011).

[12] M. M. Rahman and D. N. Davis, Addressing the Class Imbalance Problem in Medical Datasets, International Journal of Machine Learning and Computing, volume 3, Issue 2, 224-228,(2013)

[13] R. Greiner and W. Zhou, Structural extension to logistic regression, Eighteenth Annual National Conference on Artificial Intelligence (AAI02), 167–173, (2002).