Hepatitis-C Classification using Data Mining Techniques

Huda Yasin
Department of Computer Science,
University of Karachi,
University Road, Karachi,
75270, Pakistan

Tahseen A. Jilani
Department of Computer Science,
University of Karachi,
University Road, Karachi,
75270, Pakistan

Madiha Danish
Department of Anatomy, Liaquat
College of Medicine & Dentistry,
Karachi, 75290, Pakistan

ABSTRACT

In this paper, we scrutinize factors that dole out significantly to augmenting the risk of hepatitis-C virus. The dataset has been taken from the machine learning warehouse of University of California. It contains nineteen features along with a class feature having binary classification. There is a total of 15 binary attributes together with a class attribute and 5 continuous attributes. The dataset contains 155 records. In order to prevail over the missing values problem, data normalization techniques are applied. First, the dimension of the problem is trimmed down. Next binary logistic regression is applied to classify the cases by using qualitative and quantitative approaches for data reduction. The three stage procedure has produced more than 89% accurate classification. Our proposed approach has a low feature complexity with a good classification rate as it is working by using only 37% of the total fields.

General Terms

Data Mining

Keywords

Binary logistic regression analyses, data mining, hepatitis-C Virus (HCV), principle component analysis.

1. INTRODUCTION

The liver is a wedge shaped organ. It is located on the upper right side of the body beneath the rib cage. This organ, which is considered to be the largest, makes up 2 to 3 percent of the overall body’s mass. Unlike the heart or stomach, the liver doesn’t have only one function. According to hepatologists, this single organ has more than 140 functions **[1].** These consist of generating bile required for digestion, piling up of minerals and vitamins, assisting in blood clotting (vitamin K), deactivating toxic, producing amino acids to build strong muscles, regulating energy, sustaining hormonal balance and processing sedatives **[2]**. When an individual becomes affected with hepatitis virus, this virus can assault his liver and cause swelling and redness in it **[3]**.

1.1. Introduction to Hepatitis

Inflammation (itis) of the liver (hepar), burning or swelling of the liver cells refers to hepatitis. There are several sources of hepatitis which comprise viral infections A, B and C; however, it also consists of auto-immune hepatitis, fatty liver hepatitis, spirituous hepatitis and toxin induced hepatitis **[1] [4] [5]**.

On the worldwide stage, it is predicted that around 250 million people are affected by hepatitis C. Furthermore, it is estimated 400 million people are chronic haulers of hepatitis B **[1].**

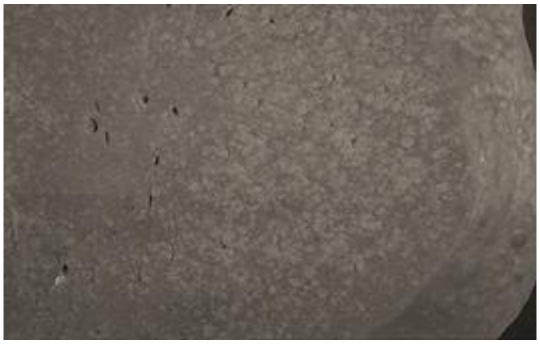

Figure 1: Showing cut section of liver

Hepatitis is a massive health issue. In fact, there is a high probability that you may be encountered with at least one or more people having this virus. Often people with hepatitis virus find it easy to live their lives without letting know others about it. This is because hepatitis may be infectious. By doing so, they thwart themselves either from facing the ignorant attitude or sympathy of others. However, people who know that they have infectious hepatitis need to take only few basic precautions to avoid passing the infection around **[5] [6] [7]**.

Generally, the hepatitis diagnosis is made by a routine blood testing or during a blood donation. Risk factors are blood transfusions, tatoos and piercing, drug abuse, hemodyalisis, health workers, sexual contact with hepatitis carrier **[8] [9]**.

There are so many scrutinizing factors for the diagnosis of hepatitis virus which makes the physician’s job difficult. A physician usually makes judgments by assessing the current test results of a patient and by referring to the previous judgments made on other patients with the similar condition.

The former method depends entirely on the physician’s knowledge, which depends on the physician’s experience to compare his/her current patient with his earlier patients. This job is not unproblematic because there are large numbers of factors that she has to evaluate. In this crucial step, she may require an accurate tool that lists her previous judgments on the patient having same (or similar) symptoms.

1.2. Hepatitis C

The World Health organization (WHO) estimates that 170 million people, i.e. 3% of the world’s population, are currently infected with the hepatitis-C virus (HCV) **[1]**. In majority of such cases, symptoms are not appeared in the infected people for many years, thus leaving them totally unaware of their condition. Liver damage is not caused by the virus itself but by the immune reaction of the body to the attack. This damage can be extremely serious, therefore resulting in liver failure and death of the patient. The indications of this type of hepatitis are normally less critical than hepatitis B. Hepatitis C spreads through

- Contaminated blood or blood products,
- Sexual contact,
- Contaminated intravenous needles.
- With some cases of Hepatitis C, no approach of transmission can be recognized.

The current treatment for HCV, according to the United Kingdom's clinical guidelines, is with a combination therapy of two drugs: Interferon-alpha and Ribavirin **[10]**. A chief factor in prescribing combination therapy is that both drugs generate side effects in most inhabitants. The cost of combination therapy is between £3000 and £12,000 per patient per year.

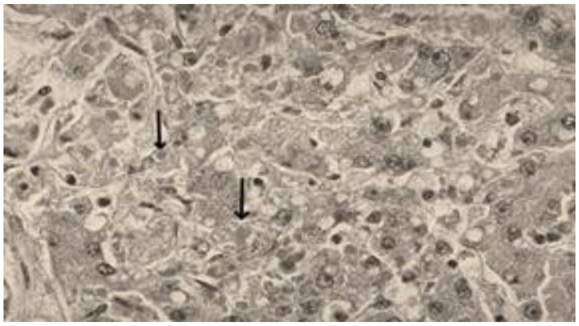

Figure 2: Microscopic Features-Hepatitis B

It is a general thinking that treating patients from pricey drugs with potentially severe side effects may be unsuitable unless there is a clear evident that patient has been infected from the virus. A liver biopsy is presently the only technique available to assess HCV activity. The biopsy involves removing a small core of tissue, which is approximately 15 mm in length and 2–3 mm in diameter. This core is then goes on in paraffin wax, cut into pieces along its length and stained.

At this level, a trained histopathologist will investigate the samples under a light microscope and use his/her practice, combined with a comprehensive definition, to evaluate the level of damage. The damage can usually be classified into two types and it is general to assign a numerical score relative to the level of damage for each type. One of the most widely used scoring methods is the Ishak system **[11]**, which can be summarized as:

- Inflammation: assigned a necroinflammatory2 (activity) score from 0 to18.
- Scarring: assigned a fibrosis3 (stage) score in the range 0–6.

Kedziora et al. **[12]** demonstrated that Phylogenetics trees and Hamming distances best reflect the differences between HCV populations present in the organisms of patients who responded positively and negatively to the applied therapy. Hodgson et al. **[4]** presented an automated system for the quantification of inflammatory cells in hepatitis-C-infected liver biopsies. The required features are extracted from color-corrected biopsy images at positions of interest, and are identified by adaptive thresholding and clump decomposition.

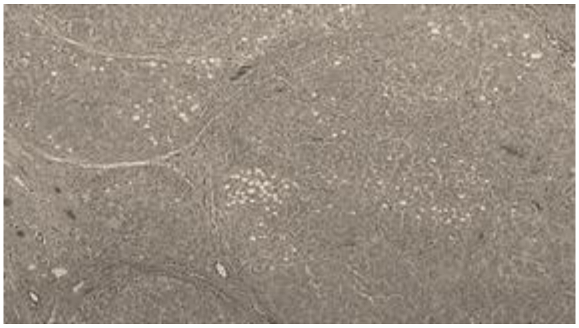

Figure 3: Microscopic Features-Hepatitis C

The experimental results show that this system can rank 15 test images, with varying degrees of inflammation, in strong agreement with five expert pathologists. Kemal Polat and Salih Güne **[8]** presented a novel method for diagnosis of hepatitis virus based on a hybrid method that uses feature selection (FS) and artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism. The obtained classification accuracy of their system was 92.59%. Also, sensitivity, and specificity values for the experimental results were 100% and 85%. Peng Guan, De-Sheng Huang and Bao-Sen Zhou **[13]** studied the application of artificial neural network (ANN) in forecasting the incidence of hepatitis A, using ARIMA model and compared the results with an ANN model. They concluded that ANN is superior to conventional methods in forecasting the incidence of hepatitis A which has an auto-regression phenomenon. Avendao **[14]** formulated a model to describe the dynamics of hepatitis C virus (HCV). They discussed the efficacy of the therapy methods for hepatitis C in terms of threshold parameter. Triumph of the treatment could possibly be forecasted from the early viral dynamics in the patients. For this purpose, they considered four populations: uninfected liver cells, infected liver cells, HCV and T cells. Moneim and Mosa **[15]** constructed a mathematical model to study the spread of HCV-subtype 4a amongst the Egyptian population. The relation between HCV-subtype 4a and the other subtypes has also been studied in the paper.

In this research paper, we have used logistic regression model to investigate factors that contribute significantly in the classification of HCV cases. The paper is organized as follows: In section 2, we have given a brief description regarding data. In section 3, we have set an introduction to data mining techniques. Section 4 discusses the models and methods involved in logistic regression. In section 5, we present experimental results of logistic regression. Section 6 concludes the paper with future studies.

2. DATA DESCRIPTION

The data available at UCI machine learning data repository contains 19 fields with one output field **[16]**. The output shows whether patients with hepatitis are alive or dead. The intention of the dataset is to forecast the presence or absence of hepatitis virus given the results of various medical tests carried out on a patient. The Hepatitis dataset contains 155 samples belonging to two different target classes. There are 19 features, 13 binary and 6 features with 6–8 discrete values. Out of total 155 cases, the class variable contains 32 cases that died due to hepatitis.

3. INTRODUCTION TO DATA MINING

Finding unrevealed information and useful patterns in a database is often referred to as data mining. The terms knowledge discovery, information retrieval, deductive learning and exploratory data analysis can be used in place of data mining. To accomplish different tasks, many different algorithms are involved in data mining. Usually data mining scopes are partitioned into predictive and descriptive areas with application specific changes pertaining to the requirements of the problems. Making prediction about data values by using previously known results from some other data is done by predictive model where identification of patterns in data is made by descriptive model **[17] [18]**.

Table 1: Description of attributes from UCCI machine learning dataset

S.No	Variable	Values
1	Class	Die, Alive
2	AGE	10, 20, 30, 40, 50, 60, 70, 80
3	SEX	Male, female
4	STEROID	No, Yes
5	ANTIVIRALS	No, Yes
6	FATIQUE	No, Yes
7	MALAISE	No, Yes
8	ANOREXIA	No, Yes
9	LIVER BIG	No, Yes
10	LIVER FIRM	No, Yes
11	PLEEN PALPABLE	No, Yes
12	SPIDERS	No, Yes
13	ASCITES	No, Yes
14	VARICES	No, Yes
15	BILIRUBIN	0.39, 0.80, 1.20, 2.0, 3.0, 4.0
16	ALK PHOSPHATE	33, 80, 120, 160, 200, 250
17	SGOT	13, 100, 200, 300, 400, 500
18	ALBUMIN	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
19	PROTIME	10, 20, … , 90
20	HISTOLOGY	No, Yes

3.1. Principal Component Analysis

Dimension of a large data set can be reduced by using principal component analysis which is considered as one of the most popular and useful statistical methods.

This method transforms the original data into new dimensions. The new variables are formed by taking linear combinations of the original variables of the form:

$$Z_1 = b_1'Y = b_{11}Y_1 + b_{12}Y_2 + \cdots + b_{1m}Y_m$$
$$Z_2 = b_2'Y = b_{21}Y_1 + b_{22}Y_2 + \cdots + b_{2m}Y_m$$
$$\ldots..$$
$$Z_p = b_p'Y = b_{p1}Y_1 + b_{p2}Y_2 + \cdots + b_{pm}Y_m$$

In matrix form, we can write Z=B.Y, where b_{11}, b_{12} ,…, b_{pp} are called the loading parameters. The new axes are adjusted such that they are orthogonal to each other with maximum information gain.

$$Var(Z_i) = b_i' \sum b_i \quad . \qquad i = 1.2. \ldots p$$
$$Cov(Z_i.Z_K) = b_i' \sum b_K \quad . \quad i = 1.2. \ldots p$$

Y_1 is the first principal component having the largest variance. As the direct computation of matrix B is not possible so in feature transformation, the first step is to determine the covariance matrix *U* which can be defined as **[19]**.

$$U_{m \times n} = \frac{1}{m-1}\left[\sum_{i=1}^{m}(Y_i - \bar{Y})'.(Y_i - \bar{Y})\right].$$

$$\text{where} \quad \bar{Y} = (\frac{1}{m})\sum_{i=1}^{m} X_i$$

The next step is to calculate the eigen values for the covariance matrix 'U'. Finally, a linear transformation is defined by n eigen vectors correspond to n eigen values from a m-dimensional space to n-dimensional space (n<m).

Table 2: Selection of attributes after data preparation and data reduction

Selected Variables	Variable Description
Steroid	Any groups of lipids with a specific 7 carbon atom ring system.
Antiviral	It's a drug that decreases or inhibits the effects of viral. Whereas Antiviral are agents that works against the viruses to decrease the effects or inhibits the effect of viruses.
Fatigue	A state of increased discomfort and decreased efficiency due to prolonged or excessive exertion (work). Loss of power or capacity to respond to stimulation (any signal).
Malaise	Felling of discomfort
Anorexia	Lack or loss of appetite for food
Liver Big	Liver increased in size or fatty Liver is called liver Big.
Liver Firm	Liver strong when liver as regular continuous structure which is not like hard and irregular called Liver firm.
Class	Die/ Live

Principal axes, which are also called eigen vectors, $E_1. E_2. \ldots .E_m$ correspond to eigen values $\lambda_1 + \lambda_2 + \cdots + \lambda_n$. Mostly, the first few principal components contain most of the information. Using the proportion of the Analysis of variances can inform us of how many principal components to be retained from the dataset.

4. REGRESSION MODEL

Regression allows forecasting future values on the basis of past values. The relationship's strength between two variables can be evaluated by bivariate model. The following equation gives the general form of linear regression model:

$$z = a_0 + a_1y_1 + \cdots + a_my_m + \epsilon$$

Here $\in$ represents the random number, X_1, X_2,…, Xm represent the input variables and are called regressors. a_0, a_1, a_2,…, a_m are the constants which are chosen to match the input samples using statistical estimation. Because the number of predictors is more than one, so it is sometimes referred to as 'multiple linear regression'- which means that this is a regression model in hyper-dimensional space **[19]**. The data values that are exceptions to the expected data are called outliers. Mostly, the preprocessing step of the data mining model building steps includes analyzing outliers and interventions.

A. *Logistic regression model*

Modeling the probability of the event occurs as a function of a linear set of predictors variable is referred to as logistic regression model **[20]**. The logistic regression model can be described as:

$$E\left(\frac{z}{x}\right) = \frac{e^T}{1+e^T} = \pi(X) = \frac{e^T}{1+e^T} \qquad \textbf{(1)}$$

Where, $\pi(x)$ represents the expected value of the response variable, natural logarithms base is e and T is:

$$T = p_0 + p_1X_1 + p_2X_2 + \cdots + p_hX_h$$

Where, p_j and X_j are coefficients and predictors respectively for *h* predictors $j = 1.2.\ldots.h$

B. *Testing hypothesis about the coefficients*

In order to determine whether a specific predictor is significant or not, a hypothesis test called Wald test is performed. This is defined as:

$$W_i = \left(\frac{\beta_i}{SE_{\beta_i}}\right)^2, \; i = 1.2.\ldots.h$$

where, SE refers to the standard error of the coefficient as estimated from the data.

C. *Assessing the goodness of fit of the model*

In a statistical model, how well a model fits an observation set is explained by the goodness of its fit **[19]**. By analyzing the residuals, a majority of the tests for goodness of fit of a model are carried out. However, for a binary (0-1) outcome variable, this approach is not beneficial. The likelihood function $1\left(\frac{p}{x}\right)$ is a parameters function $p = p_0.p_1.p_2.\ldots.p_m$ which expresses the observed data's probability **[21]**. The log-likelihood function can be written as:

$$1\left(\frac{p}{x}\right) = \sum_{i=1}^{n}[z_i \ln \pi(x_i) + (1-z_i)\ln(1-\pi(x_i))]$$

Where, z_i and $\pi(x_i)$ are the actual outcome and the predicted probability respectively of event occurring.

5. RESULTS

Initially, we started working with all the available variables taken from UCI machine learning data repository. The classification results obtained are up to 90.6%. Correlation exists among the variables thus it is infeasible to work with all the nineteen variables. As the data suffers from the curse of dimensionality problem, certain steps were necessary.

Table 3: Values in the equation

	B	S.E.	Wald	Sig	Exp(B)
Antiviral	0.947	0.626	2.293	0.13	2.579
Bilirubin	-0.177	0.837	0.045	0.833	0.838
SGOT	-0.006	0.006	1.145	0.283	0.994
Albumin	0.002	0.003	0.557	0.455	1.002
Protime	0.626	0.464	1.82	0.177	1.871
Histology	0.018	0.017	1.029	0.31	1.018
Steriod	3.142	1.451	4.688	0.03	23.142
Ascites	1.691	0.632	7.169	0.007	5.427
Liver-firm	2.01	0.775	6.731	0.009	0.134
Age	-0.073	0.024	9.174	0.002	0.93
Alk-Phos	-0.893	0.306	8.502	0.004	0.41

Before proceeding for model fitting, first we applied a data reduction technique to reduce the dimensions. After applying principal component analysis on the nineteen independent variables, we have found that the first seven principle components cover more than 98% of the total variability of the continuous data space.

We have observed that data mining using data reduction resulted in better values than performance indicators like mean square error and coefficient of determination. After data reduction, the seven independent variables are steroid, antiviral, fatigue, malaise, anorexia, liver big and liver firm. Table 3 presents the estimation the logistic regression model. This table gives the coefficients, standard error for coefficients, Wald statistics, and significance value for Wald statistic.

Table 4: Classification table

		Predicted		
		Die	Live	% correct
Observed	Die	18	13	58.1
	Live	<u>**3**</u>	120	97.6
	Overall %			**89.6**

Table 5 represents classification accuracy in percentage of our proposed method. The classification precision of other approaches which were used for the diagnosis of hepatitis is also represented in the table **4**.

A. *Test hypothesis about the coefficients*

Table 3 represents the calculated Wald statistic and its corresponding significance level to test the null hypothesis for possible rejection. The significant level of age is 0.002 which indicates its higher prevalence in the risk of hepatitis C virus.

The positive coefficient of albumin and histology reveals that the risk of hepatitis C virus augments with the augmenting value of these factors. Similarly, the negative coefficient of SGOT indicates that the risk of virus increases with the decreasing values of this factor.

B. Classification of cases

The classification of cases forecasted is represented in Table 4. From the table it is revealed that 18 patients who were infected with the hepatitis C virus were correctly envisaged by the model after their death. This implies that 58.1% of the patients appropriately classified who have died. In a similar way, 123 patients were predicted accurately to be alive i.e. 97.6% of the patients were classified correctly to be alive. The off-diagonal records illustrate the number of patients that were classified inaccurately, i.e. 13 patients who have died were classified incorrectly or we can referred it to as type-I error. Similarly, only 03 patients who are alive were classified incorrectly as dead.

Table 5: Classification accuracies obtained by using hepatitis diagnostic methods

Used method	Article author's	Classification accuracy (%)
RBF	Özyıldırım, Yıldırım, et al.	83.75
15NN, stand. Euclidean	Grudzinski	89
FSM without rotations	Adamczak	88.5
MLP with BP	Stern and Dobnikar	82.1
QDA, quadratic discriminant analysis	Stern and Dobnikar	85.8
ASI	Stern and Dobnikar	82
MLP + BP (Tooldiag)	Adamczak	77.4
LDA	Stern and Dobnikar	86.4
MLP	Özyıldırım, Yıldırım, et al.	74.37
RBF (Tooldiag)	Adamczak	79
1NN	Stern and Dobnikar	85.3
Naïve Bayes and semi-NB	Stern and Dobnikar	86.3
Fisher discriminant analysis	Stern and Dobnikar	84.5
LVQ	Stern and Dobnikar	83.2
GRNN	Özyıldırım, Yıldırım, et al.	80
ASR	Stern and Dobnikar	85
IncNet	Norbert Jankowski	86
CART (decision tree)	Stern and Dobnikar	82.7
LFC	Stern and Dobnikar	81.9
Proposed Method	Used in this study	**89.6**

From total number of 155 cases, the overall percent appropriately predicted appears reasonably superior at 89.6%.

6. CONCLUSION & FUTURE STUDIES

In this paper we have investigated factors which have higher prevalence of the risk of hepatitis C virus. In future, we will extend our research by using different techniques like outlier analysis and link analysis also referred to as association rule mining on large number of patients. Also, we aim to investigate other factors such as hepatic enzymes (ALT etc.), blood picture, urine analysis, serology for viral markers etc. Moreover, we will apply fuzzy learning models for further enhanced forecasting of hepatitis C virus.

7. ACKNOWLEDGMENTS

We want to give our special thanks to Mr. Muhammad Yasin Subhan for his encouragement and support with the evaluation.

REFERENCES

[1] WHO, Hepatitis C (Fact Sheet No. 164), World Health Organization, Geneva, 2000.

[2] WHO, Hepatitis C global prevalence (update), Weekly Epidemiological Record (World Health Organization), 74, 1999, pp. 421–428.

[3] Information regarding hepatitis C from the staff of Mayo Clinic; available at: http://www.mayoclinic.com/health/hepatitis-c/DS00097.

[4] Hodgson S., Harrison R. F., Cross S. S., An automated pattern recognition system for the quantification of inflammatory cells in hepatitis-C-infected liver biopsies, Image and Vision Computing 24, 2006, pp. 1025–1038.

[5] Moriishi K. and Y. Matsuura, "Mechanisms of hepatitis C virus infection", Antivir. Chem. Chemother 14, 2003, pp. 285–297.

[6] Fattovich G. and Schalm S.W., Hepatitis C and cirrhosis, in Hepatitis C, T.J. Liang, J.H. Hoofnagle, San Diego, Academic Press, pp. 241–264 eds, 2000.

[7] Lawrence S. P., "Advances in the treatment of hepatitis C", Advanced in International Medicine. 45, 2000, pp. 65–105.

[8] Polat K., Gunes S., "Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation", Digital Signal Processing 16, 2006, pp. 889–901.

[9] Polat, K., & Gunes, S., An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digital Signal Processing, 17(4), 2007, pp. 702–710.

[10] Booth J., O'Grady J. and Neuberger J. (2001), Clinical guidelines on the management of hepatitis C, Gut 1, 11–21.<www.kidshealth.org/parent/infections/bacterial_viral/hepatitis.html> Accessed 19.02.10.

[11] Ishak K., A. Baptista, L.B. Histological, et al., Histological grading and staging of chronic hepatitis, Journal of Hepatology, 22, 1995, pp. 696–699.

[12] Kedziora P., Figlerowicz M., Formanowicz P., Alejska M., Jackowiak P., Malinowska N., Fratczak A., Blazewicz J., and Figlerowicz M., "Computational Methods in Diagnostics of Chronic Hepatitis C", Bulletin of the Polish Academy of Sciences, Technical Sciences, 53 (3), 2005, pp.273-281.

[13] Peng Guan, De-Sheng Huang, Bao-Sen Zhou, "Forecasting model for the incidence of hepatitis A based on artificial neural network", China World Journal of Gastroenterol , 10 (24), 2004, pp. 3579-3582.

[14] Avendao R., Esteva L., Flores J. A., et al., A Mathematical Model for the Dynamics of Hepatitis C, Computational and Mathematical Methods in Medicine 4(2), 2002, pp.109- 118.

[15] Moneim I. A. and Mosa G. A., "Modeling the Hepatitis C with Different Types of Virus Genome", Computational and Mathematical Methods in Medicine, 7(1), 2006, pp. 3-13.

[16] Blake, C. L., & Merz, C. J. (1996). UCI reporsitory of machine learning databases. Available from: <http://www.ics.uci.edu./ ~mlearn/MLReporsitory.html>.

[17] Dunham M. H. and Sridhar S., Data Mining: Introductory and Advanced topics, Pearson Education, 2006

[18] Larose D. T., Data mining methods and models. John Wiley and sons, 2006.

[19] Kantardzic M., Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons 2003.

[20] Neter J., Waserman W. and Kutner, M., Applied Linear Statistical Models: Regression Analysis of Variance and Experimental Designs, McGraw Hill, 3rd edition, 1996.

[21] Hosmer D. and Lemeshow S., Applied Logistic Regression, John Wiley and Sons, 2nd edition, 2000.