Acute Coronary Syndrome Prediction Using Data Mining Techniques- An Application

Tahseen A. Jilani, Huda Yasin, Madiha Yasin, Cemal Ardil

***Abstract*—**In this paper we use data mining techniques to investigate factors that contribute significantly to enhancing the risk of acute coronary syndrome. We assume that the dependent variable is diagnosis – with dichotomous values showing presence or absence of disease. We have applied binary regression to the factors affecting the dependent variable. The data set has been taken from two different cardiac hospitals of Karachi, Pakistan. We have total sixteen variables out of which one is assumed dependent and other 15 are independent variables. For better performance of the regression model in predicting acute coronary syndrome, data reduction techniques like principle component analysis is applied. Based on results of data reduction, we have considered only 14 out of sixteen factors.

***Keywords*—**Acute coronary syndrome (ACS), binary logistic regression analyses, myocardial ischemia (MI), principle component analysis, unstable angina (U.A.).

I. INTRODUCTION

ONE of the major causes of death worldwide is the cardiovascular disease (CVD). Acute coronary syndrome (ACS) is considered as one of the most common form of heart syndrome. The term ACS is used to cover any clinical symptom's group compatible with acute myocardial ischemia. The chest pain occurs due to the insufficient blood supply to the heart muscle is called 'Acute Myocardial Infarction' that results from coronary artery disease (also called coronary heart disease) [1]. The rupture of an atherosclerotic plague is the cause of acute coronary syndrome [2].

Acute coronary syndrome (ACS) includes three acute manifestations of ischemic heart disease [3]:

1) Unstable angina (UA)
2) Non-ST elevation (MI)
3) Sudden cardiac death

ECG changes include ST segment depression or T wave flattening. In unstable angina cardiac enzymes are not elevated while in non-ST elevation MI (NSTEMI) cardiac enzymes become elevated. In cardiac hospital, most of the patients have diagnosed ACS which may be a sign of left ventricular dysfunction during pain [3].

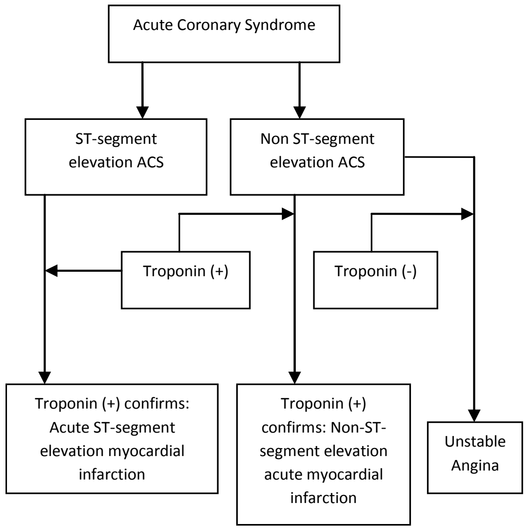

Fig. 1 Classification of Acute Coronary Syndrome [4]

In Pakistan, the prevalence of ACS is increasing rapidly. For example, 414 patients were admitted in National Institute of Cardiovascular Diseases in September 2000 with 71.25% males. Around 72.92% of the patients were in the fifth decade of life. The most common presentation was the acute coronary syndrome (ACS), present in 39.8% of the patients. Similarly, a total of 446 patients were admitted in September 2005. Now, males were 63%. Of these, 71.29% were in the fifth, sixth, and seventh decades of life. The patients admitted with acute coronary syndromes (ACS) were around 43.04% see [5]. Thus, there is a need of exploration of those factors responsible to enhancing the risk of ACS for reducing the prevalence of this syndrome.

Data for this research were collected from two different cardiac hospitals, Karachi in the year 2008. There were 319 observations in the data set. The data set comprises one dependent variable (Diagnosis) and sixteen independent variables as given in Table I. In the data set, there were 104 patients without ACS and 215 patients with ACS. The data set is highly volatile and noisy due to the diversity of patients' history, physical, mental, social and economical classes. Even

Dr. T. A. Jilani is Assistant Professor in the Department of Computer Science, University of Karachi, Pakistan (phone: +92-333-3040963; e-mail: tahseenjilani@uok.edu.pk).
H. Yasin, is undergraduate student (software engineering) in the Department of Computer Science, University of Karachi, Pakistan.
Dr. M. Yasin is demonstrator in the Department of Anatomy, Liaquat College of Medicine and Dentistry, Karachi, Pakistan.
Cemal Ardil is with the Azerbaijan National Academy of Aviation, Baku, Azerbaijan.

TABLE I
DESCRIPTION OF DEPENDENT AND INDEPENDENT VARIABLES

Variable Label	Variable Name	Level	Value Label
1. Dependent variable			
Diagnosis	ACS	0	Absent
		1	Present
2. Independent variables			
a. Categorical Variables			
Gender	Patient's gender	0	Male
		1	Female
Smoke	Smoking	0	Non-smoker
		1	Smoker
Hyp	Hypertension	0	No
		1	Yes
FamHistory	Family History	0	No
		1	Yes
DM	Diabetics Miletus	0	No
		1	Yes
SK	Streptokinase	0	No
		1	Yes
b. Continuous Variables			
Age	Patient's age		
FBS	Fasting blood sugar		
RBS	Random blood sugar		
BP (sys)	Blood Pressure Systolic		
BP (dias)	Blood Pressure Diastolic		
Chol	Cholesterol		
Hb	Hemoglobin		
RR	Respiratory Rate		
HR	Heart Rate		
PR	Pulse Rate		

then we have tried to efficiently preprocess the data set using data mining preprocessing techniques, and obtained good experimental results.

In this research paper we have used logistic regression model to investigate factors that contribute significantly to enhancing the risk of ACS. For analyzing this problem, we observe whether a person have or does not have ACS. The paper is organized as follows. In section II we have given a brief introduction to data mining techniques. Section III discusses the models and methods involved in logistic regression. In section IV, we present experimental results of logistic regression. Section V concludes the paper with future studies. In the following, we have given some literature review about the applications of data mining and intelligent systems in acute coronary syndrome (ACS).

Lavesson et al [6] applied several data mining techniques to predict the severity of an ACS based on electrocardiograms. Only two classes unstable Angina (UA) and Myocardial Infarction (MI) were assumed as values of dependent variable. Based on 28 features, they evaluated different types of features selection techniques and applied supervised neural network for prediction model. McCullough et al. [7] used neural networks to examine 13 features with single target output that indicate symptoms of ACS with separate models for males and females. They also compared the results with receiver OC curves. Kostakis et al [8] investigated the patterns in cardiovascular risk factors with their matched controls. They discussed the application of OLAP-specific procedures in order to explore hidden pathways associated with risk factors among patients and controls. Rao et al. [9] proposed a probabilistic framework for Re liable Extraction and Meaningful Inference from Non-structured Data (REMIND) that integrates the structured and unstructured clinical data in patient records to automatically create high-quality structured clinical data. REMIND also performs inference with data from multiple sources and to enforce consistency between different medical conclusions drawn from the data -- via a probabilistic reasoning framework. Scott et al [10] discussed the measurements and quality checking of care in health care patients especially, acute coronary syndromes. Massad et al. [11] reviewed the current state of the art of logic applications in medical diagnosis. Tamil et al [12] reviewed feature extraction and classification method for bio-signal processing which concentrates on electrocardiogram (ECG) signal processing. They in depth discussed the discrete wavelet transform for feature extraction and neuro-fuzzy logic for classification. Quteishat and Lim [13] discussed the intelligent data mining techniques like min-max neural networks to medical diagnosis. They choose real medical records from suspected ACS patients is collected and used for experimentation.

II. 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 the 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 [14].

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

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

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.

$$\mathrm{Var}(Z_i) = b_i'\Sigma\, b_i \qquad , \quad i = 1,2,...,p$$
$$\mathrm{Cov}(Z_i, Z_k) = b_i'\Sigma\, b_k \ , \quad i = 1,2,...,p$$

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 [15]:

$$U_{mxn} = \frac{1}{m-1}\left[\sum_{i=1}^{m}\left(Y_i - \overline{Y}\right)^t . \left(Y_i - \overline{Y}\right)\right],$$

where $\overline{Y} = \left(\frac{1}{m}\right)\sum_{i=1}^{m} X_i$

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). Principal axes are also called eigen vectors $E_1, E_2, ..., E_m$ corresponds to eigen values $\lambda_1 + \lambda_2 + ... + \lambda_n$. Mostly, the first few principal components contain most of the information. Using Analysis of variances' proportion tells how many principal components to be retained from the dataset [15].

III. 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 regression [14]. The following equation gives the general form of linear regression model:

$$z = a_0 + a_1 y_1 + ... + a_m y_m + \in$$

Here $\in$ represents the random number, m represents 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. Because the number of predictors is more than one so it is sometimes referred to as 'multiple linear regression' that is a regression model in hyper-dimensional space [14]. The data values that are exceptions to the expected data are called outliers. Mostly, the preprocessing step of the data mining model building steps included analysis of the outliers and interventions.

A. Logistic regression model

Modeling the probability of the event occurs as a function of linear set of predictors variable is referred as logistic regression model [15]. The logistic regression model can be described as:

$$E(Z/x) = \frac{e^T}{1+e^T} = \pi(x) = \frac{e^T}{1+e^T} \qquad (1)$$

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

$$T = \rho_0 + \rho_1 X_1 + + \rho_2 X_2 + ... + \rho_h X_h$$

Where, ρ_j and X_j are coefficients and predictors respectively for *h* predictors $j = 1,2,...,h$.

B. Testing hypothesis about the coefficients

In order to determine whether a specific predictor is significance or not, a hypothesis test is performed which is called Wald test see [16]. It is defined as:

$$W_i = \left(\frac{\mu_i}{S.E_{\beta_i}}\right)^2, \ i = 1,2,...,h$$

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

C. Partial correlation

Partial correlation between each of the independent variables and dependent variable can be obtained with range from -1 to +1. Sample partial correlation coefficient estimates the measure of linear relationship between any two variables leaving the effects of the remaining variables [17]. Partial correlation can be defined by the given equation:

$$P_{corr} = \pm\sqrt{\frac{\text{Wald Statistics} - 2\,\text{df}}{-2\,\text{Loglikelihood}_{(0)}}}$$

where df represents degree of freedom and $-2\,\text{Loglikelihood}$ of a base model with no variable or a base model which contains the intercept only.

D. Assessing the goodness of fit of the model

In a statistical model, how well a model fits an observation set is explained by goodness of fit [16]. By analyzing the residuals, majority of the tests for goodness of fit of a model are carried out; although for binary (0-1) outcome variable, this approach is not good [17]. The likelihood function $l(\rho / x)$ is a parameters function $\rho = \rho_0, \rho_1, \rho_2, ...\rho_m$ which expresses the observed data probability [16]. The log-likelihood function can be written as:

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

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

IV. RESULTS

As the data have the problem of curse of dimensionality, therefore, before proceeding for model fitting, first we have applied some data reduction technique to reduce the dimensions. After applying principal component analysis on the ten independent numeric variables, we have found that the first eight principle components cover more than 98% of the total variability of the continuous data space. Respiratory rate and hemoglobin have small eigen values and thus their influence is minimal on the information contents of the data set. We have observed that the data mining using data reduction resulted in better values of the performance indicators like mean square error and coefficient of determination. After data reduction, the fourteen independent variables are age, gender, smoke, hypertension, family history, diabetics mellitus, fasting blood sugar, random blood sugar, cholesterol, streptokinase, blood pressure (systolic), blood pressure (diastolic), heart rate and pulse rate.

Table-II presents the estimation the logistic regression model. This table gives the coefficients, standard error for coefficients, Wald statistics, and significance value for Wald statistic.

A. Test hypothesis about the coefficients

Table II represents the calculated Wald statistics and its corresponding significance level to test the null hypothesis for possible rejection. The significance level of smoking is 0 which indicates its higher prevalence in the risk of ACS. The positive coefficient of BPs, HR and RBS reveals that the risk of ACS increases with the increasing value of these factors. Similarly, the negative coefficient of BPd and PR indicates that the risk of disease increases with the decreasing values if these factors.

B. Classification of cases

Table-III represents the classification of cases predicted. Results show that 43 individual not having ACS were correctly predicted by the model which indicates that 37.1% of the individuals correctly classified without ACS. In the same way, 203 individuals were correctly predicted to have ACS i.e., 88.7% of the individuals were classified correctly with ACS. The off-diagonal entries show the number of individual that were incorrectly classified i.e.73 individuals not having ACS were classified incorrectly or we can referred it to as type-I error. In the same way, 23 individuals having ACS were incorrectly classified as not having ACS. Of 319 cases, 69.9% of the cases were correctly classified. Although, based on analysis, the false positive cases that is, those who have no ACS and they are predicted as having ACS is not very serious case. The most significant issue arises from true negative cases that are those who are diseased and predicted as non-diseased. Based on this discussion, if we focus ourselves at true negative cases then the error rate in the study reduces to 7.21% with prediction accuracy of 92.79%. Due to many internal and external reasons, like war against terrorism and internal financial crises, Karachi city has become the hub for the whole people migrating from other parts of the country. Therefore patient history diversity; physical, mental health, social status etc are variables of high impact on the acute coronary syndrome analysis. Thus, for such a noisy and volatile dataset, a model accuracy of 92.79% may be appreciated.

TABLE-II
VARIABLES IN THE EQUATION

	B	S.E.	Wald	Sig	Exp(B)
Gender	0.561	0.302	3.445	0.063	1.753
Age	-0.003	0.011	0.049	0.824	0.997
BPs	0.01	0.008	1.676	0.195	1.01
BPd	-0.016	0.012	1.92	0.166	0.984
HR	0.003	0.005	0.455	0.5	1.003
Smoking	1.078	0.296	13.233	0	2.938
Hyp	0.275	0.391	0.493	0.483	1.316
PR	-0.012	0.008	2.075	0.15	0.988
Fam-Hist	-0.1	0.262	0.146	0.703	0.905
Diabetics	-0.481	0.577	0.694	0.405	0.618
FBS	-0.003	0.004	0.776	0.378	0.997
RBS	0.001	0.002	0.104	0.747	1.001
SK	-0.546	0.335	2.661	0.103	0.579
Cholesterol	0.007	0.007	1.18	0.277	1.007
Constant	-0.053	1.764	0.001	0.976	0.949

TABLE III
CLASSIFICATION TABLE

		Predicted		
		Non-Diseased	Diseased	% correct
Observed	Non-diseased	43	73	37.1
	Diseased	**23**	180	88.7
	Overall %			69.9

V. CONCLUSION AND FUTURE STUDIES

In this paper we have investigated factors which have higher prevalence of the risk of acute coronary syndrome. We observed that in comparison with other factors, smoking is the most significant factor. In future, we will extend this paper to obtain further improved results using outlier analysis and link analysis (association rule mining). We aim to investigate the effects of diet, environmental, social and fluctuations on acute coronary syndrome. Also, we will apply fuzzy learning models for further improved prediction of acute coronary syndrome.

REFERENCES

[1] American Heart Association - acute coronary syndrome. Available: http://www.americanheart.org.

[2] M. A. Chisholm-Burns, B. G. Wells, T. L. Schwinghammer, P. M. Malone, J. M. Kolesar, J. C. Rotschafer and J. T. Dipiro, Pharmacotherapy Principles & Practice, McGraw Hill 2007, chapter 5.

[3] M. I. Danish, Medical Diagnosis and Management, 5th edition.

[4] S.A. Spinler, Pharmacotherapy Self-Assessment Program-acute coronary syndrome, 5th Edition.

[5] S. F. Kazim, A. Itrat, N. W. Butt and M. Ishaq, "Comparison of cardiovascular disease patterns in two data sets of patients admitted at a Tertiary Care Public Hospital in Karachi five years apart", Pak J Med Sci 2009, vol. 25, no.1, pp. 55-60.

[6] N. Lavesson, A. Halling, M. Freitag, J. Odeberg, H. Odeberg, P. Davidsson (2009), "Classifying the severity of an Acute Coronary Syndrome by Mining Patient Data", 25th Annual Workshop of the Swedish Artificial Intelligence Society, Linköping University Electronic Press, ISSN 1650-3686.

[7] C. L. McCullough, A. J. Novobilski, F. M. Fesmire (2007), "Use of Neural Networks to Predict Adverse Outcomes from Acute Coronary Syndrome for Male and Female Patients", 6th International Conference on Machine Learning and Applications (ICMLA), 13-15, December. Cincinnati, Ohio, USA.

[8] H. Kostakis , B. Boutsinas, D. B. Panagiotakos and L. D. Kounis (2008), **"**A Computational Algorithm for the Risk Assessment of Developing Acute Coronary Syndromes, Using Online Analytical Process Methodology Source", International Journal of Knowledge Engineering and Soft Data Paradigms, Pages 85-99.

[9] R. B. Rao, S. Krishnan and R. S. Niculescu (2006), Data mining for Improved Cardiac Care**, A**CM SIGKDD Explorations Newsletter, 8(1), pp. 3 – 10.

[10] I. A. Scott, C. P. Denaro, J. L. Flores, C. J. Bennett, A. C. Hickey and A. M. Mudge (2002), Quality of care of patients hospitalized with acute coronary syndromes, Royal Australasian College of Physicians, Australia.

[11] Massad E., Ortega N. R.S., L. C Barros and C. J. Struchiner (2008), "...and Beyond: Fuzzy Logic in Medical Diagnosis", Fuzzy Logic in Action: Applications in Epidemiology and Beyond, Studies in Fuzziness and Soft Computing, vol, 232/2008. Springer-Verlag.

[12] E. B. M. Tamil, N. H. Kamarudin, R. Salleh and A. M. Tamil(2008), A Review on Feature Extraction & Classification Techniques for Biosignal Processing (Part I: Electrocardiogram), 4th Kuala Lumpur International Conference on Biomedical Engineering (BIOMED), 25–28 June 2008 Kuala Lumpur, Malaysia, pp. 107-112. IFMBE Proceedings, Springer-Verlag

[13] A. Quteishat , C. P. Lim(2008), "Application of the Fuzzy Min-Max Neural Networks to Medical Diagnosis", Lecture Notes In Artificial Intelligence, vol. 5179.

[14] Proceedings of the 12th international conference on Knowledge-Based Intelligent Information and Engineering Systems, Part III, pp. 548 – 555, Springer-Verlag

[15] M. H. Dunham and S. Sridhar, Data Mining: Introductory and Advanced topics, Pearson Education 2006, chapter 1, chapter 3, chapter 4.

[16] M. Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons 2003, chapter 5.

[17] D. T. Larose, Data mining methods and models. John Wiley and sons, 2006, chapter 4.