Available online at http://www.journalijdr.com

IJDR

International Journal of
DEVELOPMENT RESEARCH

ISSN: 2230-9926

International Journal of Development Research
Vol. 5, Issue, 02, pp. 3545-3549, February, 2015

Full Length Research Article

APPLICATION OF DECISION TREE AS A DATA MINING TOOL TO PREDICT BP SYSTOLIC DIASTOLIC

***Dr. Zeki S. Tywofik and Saif mohammed Ali**

Department of Computer, Dijlah University College (DUC),Baghdad, Iraq

ARTICLE INFO

Article History:
Received 17th November, 2014
Received in revised form
27th December, 2014
Accepted 16th January, 2015
Published online 27th February, 2015

Key words:
Decision Tree Modeling,
Healthcare Data Mining Applications,
Decision Rule Induction,
Data set.

ABSTRACT

Hemoglobin A1c is the most parameters for the monitoring of metabolic control of patients with diabetes mellitus. The aim of this study is to determine the reference rang of glycosylated hemoglobin (Hb A1c%) in an Iraqi population (males and females) and predict Bp systolic diastolic by using demonstrates the application of decision tree, as data mining tool, in the health care system. Data mining has the capability for classification, prediction, estimation, and pattern recognition by using health databases. Blood samples were collected from 100 healthy subjects (50 females and 50 females) are ranged between (20-75) years old as dataset. The reference value of HbA1c% was (5.34 + 0.67) % in female and (5.67 + 0.73) % in males. The present study found a strong relation between HbA1c % and systolic diastolic blood pressure in males whereas the relation in females no significant.

INTRODUCATION

Data mining can be defined as the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Alternatively, it can be defined as the process of data selection and exploration and building models using vast data stores to uncover previously unknown patterns. Data mining is not new—it has been used intensively and extensively by financial institutions, for credit scoring and fraud detection; marketers, for direct marketing and cross-selling or up-selling; retailers, for market segmentation and store layout; and manufacturers, for quality control and maintenance scheduling. In healthcare, data mining is becoming increasingly popular, if not increasingly essential. Several factors have motivated the use of data mining applications in healthcare. The existence of medical insurance fraud and abuse, for example, has led many healthcare insurers to attempt to reduce their losses by using data mining tools to help them find and track offenders.3 Fraud detection using data mining applications is prevalent in the commercial world, for example, in the detection of fraudulent credit card transactions.

Recently, there have been reports of successful data mining applications in healthcare fraud and abuse detection. Another factor is that the huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining can improve decision-making by discovering patterns and trends in large amounts of complex data.4 Such analysis has become increasingly essential as financial pressures have heightened the need for healthcare organizations to make decisions based on the analysis of clinical and financial data.

Insights gained from data mining can influence cost, revenue, and operating efficiency while maintaining a high level of care. 5 Healthcare organizations that perform data mining are better positioned to meet their long-term needs, Benko giving an illustration of a healthcare data mining application; and finally, highlighting the limitations of data mining and offering some future directions. The collection of information biological of medical are about patients lay in the filed bioinformatics. The field of bioinformatics relies heavily on work by experts in statistical methods and pattern recognition. Researchers come to bioinformatics from many fields, including mathematics, computer science, and linguistics. Unfortunately, biology is a science of the specific as well as the general. Bioinformatics is full of pitfalls for those who look for patterns and make predictions (Data mining Tool)

****Corresponding author: Dr. Zeki S. Tywofik***
Department of Computer, Dijlah University College (DUC),Baghdad, Iraq

without a complete understanding of where biological data comes from and what it means. Glycated or glycosylated hemoglobin A1c, HABa1C, A1C, or Hb1c, formed through the non enzymatic binding of circulating glucose to hemoglobin. Higher levels of glucose in blood contribute to more binging and consequents higher levels of glycosylated hemoglobin Milley (2000). This paper describes the use of decision tree and rule induction in data mining applications and use this tool to make decision tree from given dataset and put rules for predication Bp systolic diastolic (Milley, 2000; Salwa and Emad Abdul-Rehman, 2010 and Rafalski, 2002).

Previews Worked

Alina Van, Valerie C. Gay, Paul J. Kennedy (2007), Understanding Risk Factors in Cardiac Rehabilitation Patients with Random Forests and Decision Trees

Cardiac rehabilitation is a well-recognised non-pharmacological intervention recommended for the prevention of cardiovascular disease. Numerous studies have produced large amounts of data to examine the above aspects in patient groups. In this paper, datasets collected for over a 10 year period by one Australian hospital are analysed using decision trees to derive prediction rules for the outcome of phase II cardiac rehabilitation. Analysis includes prediction of the outcome of the cardiac rehabilitation program in terms of three groups of cardiovascular risk factors: physiological, psychosocial and performance risk factors. Random forests are used for feature selection to make the models compact and interpretable. Balanced sampling is used to deal with heavily imbalanced class distribution. Experimental results show that the outcome of phase II cardiac rehabilitation in terms of physiological, psychosocial and performance risk factor can be predicted based on initial readings of cholesterol level and hypertension, level achieved in six minute walk test, and Hospital Anxiety and Depression Score (HADS) anxiety score and HADS depression score respectively. This will allow for identifying high risk patient groups and developing personalised cardiac rehabilitation programs for those patients to increase their chances of success and minimize their risk of failure

Shital Shah*, Andrew Kusiak*, and Bradley Dixon (2003), Data Mining in Predicting Survival of Kidney Dialysis Patients -Invariant object approach

The number of patients on hemodialysis due to end stage kidney disease is increasing. The median survival for these patients is only about 3 years and the cost of providing care is high. Finding ways to improve patient outcomes and reduce the cost of dialysis is a challenging task. Dialysis care is complex and multiple factors may influence patient survival. More than 50 parameters may be monitored while providing a kidney dialysis treatment. Understanding the collective role of these parameters in determining outcomes for an individual patient and administering individualized treatments is of importance. Individual patient survival may depend on a complex interrelationship between multiple demographic and clinical variables, medications, and medical interventions. In this research, a data mining approach is used to elicit knowledge about the interaction between these variables and patient survival. Two different data mining algorithms are employed for extracting knowledge in the form of decision rules. Data mining is performed on the individual visits of the *"most invariant"* patients as they form *"signatures"* for their decision categories. The concepts introduced in this research have been applied and tested using a data collected at four dialysis sites. The computational results are reported.

Dr.K.P.Kaliyamurthie1, D. Parameswari (2011), Performance of Decision trees for Assessment of the Risk factors of Heart Disease

Coronary heart disease refers to the failure of coronary circulation to supply adequate circulation to cardiac muscle and surrounding tissue. The events myocardial infarction (MI), percutaneous coronary intervention (PCI), and coronary artery bypass graft surgery (CABG) were investigated The risk factors investigated were: 1) before the event: a) no modifiable—age, sex, and family history for premature CHD, b) modifiable—smoking before the event, history of hypertension, and history of diabetes; and 2) after the event: modifiable— smoking after the event, systolic blood pressure, diastolic blood pressure, total cholesterol, high- density lipoprotein, low-density lipoprotein, triglycerides, and glucose.. Data-mining analysis was carried out using the C5 decision tree algorithm for the aforementioned three events using five different splitting criteria. C4.5 is a widely-used free data mining tool that is descended from an earlier system called ID3 and is followed in turn by C5.0. It embodies new algorithms for generating rule sets, and the improvement is dramatic in accuracy, speed and memory.

Healthcare Data Mining Applications

There is vast potential for data mining applications in. Generally, these can be grouped as the evaluation of treatment effectiveness; management of healthcare; customer relationship management; and detection of fraud and abuse. More specialized medical data mining, such as predictive medicine and analysis of DNA micro-arrays, lies outside the scope of this paper.

Treatment effectiveness

Data mining applications can be developed to evaluate the effectiveness of medical treatments. By comparing and contrasting causes, symptoms, and courses of treatments, data mining can deliver an analysis of which courses of action prove effective.2 For example, the outcomes of patient groups treated with different drug regimens for the same disease or condition can be compared to determine which treatments work best and are most cost-effective (Cios and Moore, 2002). Along this line, United HealthCare has mined its treatment record data to explore ways to cut costs and deliver better medicine.15 It also has developed clinical profiles to give physicians information about their practice patterns and to compare these with those of other physicians and peer-reviewed industry standards. Similarly, data mining can help identify successful standardized treatments for specific diseases. In 1999, Florida Hospital launched the clinical best practices initiative with the goal of developing a standard path of care across all campuses, clinicians, and patient admissions (http://www.angoss.com/). A good account of data mining applications at Florida Hospital also can be found in Gillespie and Veletsos. Other data mining applications related to treatments include associating the various side-effects of treatment, collating common symptoms to aid diagnosis, determining the most effective drug compounds for treating sub populations that respond differently from the mainstream population to certain drugs, and determining proactive steps that can reduce the risk of affliction.

Healthcare management

To aid healthcare management, data mining applications can be developed to better identify and track chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims. For example, to develop better diagnosis and treatment protocols, the Arkansas Data Network looks at readmission and resource utilization and compares its data with current scientific literature to determine the best treatment options, thus using evidence to support medical care. Also, the Group Health Cooperative stratifies its patient populations by demographic characteristics and medical conditions to determine which groups use the most resources, enabling it to develop programs to help educate these populations and prevent or manage their conditions.1 Group Health Cooperative has been involved in several data mining efforts to give better healthcare at lower costs. In the Seton Medical Center, data mining is used to decrease patient length-of-stay, avoid clinical complications, develop best practices, improve patient outcomes, and provide information to physicians—all to maintain and improve the quality of healthcare (Paddison, 2000 and Schuerenberg, 2003). In this paper use Decision tree to predict which Bp systolic diastolic. Data mining can be used to analyze massive volume of data and statistics to search for patterns that might indicate Bp systolic diastolic and what relation with HbA1c and WE. Kgm (Ludwig et al., 2000).

Decision Tree Modeling

Decision trees are generated from training data (see table 1) in a top-down for specific direction. The initial state of a decision tree is the root node the is assigned to the outlook. The node is split into two classes which are Female (F) and Male (M). In process net node for each female and male split into according Hb A1c which include levels each level value range of Hb A1c and then attribute Age Yrs alkse have levels each level include value of range finally there is called leave nodes. The entropy good measure for splits of decision tree. A measure used from Information the entropy of a dataset can be considered to be how disordered it is. It has been shown that entropy is related to information, in the sense that the higher the entropy, or uncertainty, of some data, then the more information is required in order to completely describe that data. In building a decision tree, we aim to decrease the entropy of the dataset until we reach leaf nodes at which point the subset that we are left with is pure, or has zero entropy and represents instances all of one class (all instances have the same value for the target attribute). The measure the entropy of a dataset, S, with respect to one attribute, in this case the target attribute, with the following calculation:

$$Entropy(s) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

Data Set

The data gathered from the were collected from 100 healthy patients subjects 2010, and included records of 100 patients.

Database

The database management system used in the study was the Microsoft SQL Server 2000. This system was used for two reasons; the software used in analysis was compatible and efficient to use with the database management system, and the data to be analyzed was maintained in the database prior to the study.

The Data Mining Process

The data exploration and presentation process consisted of various steps. These steps were data preparation, data selection and transformation, data mining and presentation.

Data Preparation

In these steps, the data that was maintained in different tables was joined in a single table. The patients database of female (F) and male (M) are collected from peoples and select the data that more effect Bp systolic diastolic such as Age Year, weight in kilograms, Height in centimeters witch relation between them and Hb A1C put it in range. The other attributes are less effect see table 1. The process errors in the data were corrected.

Implementation

There is a strong correlation between the data obtained from collected blood samples for 100 healthy subjects include females and males. The dataset contains Sex, Age Yrs, WE. Kgm, Higfht Cm., Hb A1c and Bp systolic diastolic see Table 1.

Table 1. Dataset training

Sex	Age Yrs	WE. Kgm	Higfht Cm.	Hb A1c	Bp systolic diastolic (class)
F	21-32	55-66	170-162	4.0-4.8%	110/70
F	32-39	50-64	170-162	4.8-5.1%	120/70
F	34-40	64-74	162-175	5.1-5.5%	120/80
F	47-64	60-85	153-165	5.6-6.2 %	130/80
F	36-67	70-85	152-171	5.8-6.6%	130/90
F	66	79	165	63%	150-90
M	25-38	64-75	183-176	4.0-5.3%	120/80
M	34-35	70-74	170-173	4.3-5.4%	130/80
M	25-32	109-129	174-180	5.6-6.1%	120/90
M	36-37	30-37	167-181	5.0-6.9%	130/90
M	35-57	4-95	174-174	5.6-5.9%	150/90

Decision Rule Induction

Decision rules, in disjunctive normal form (DNF), may be induced from training data in a bottom-up specific-to-general style, or in a top-down general-to-specific style, as in decision tree building see Figure 1.

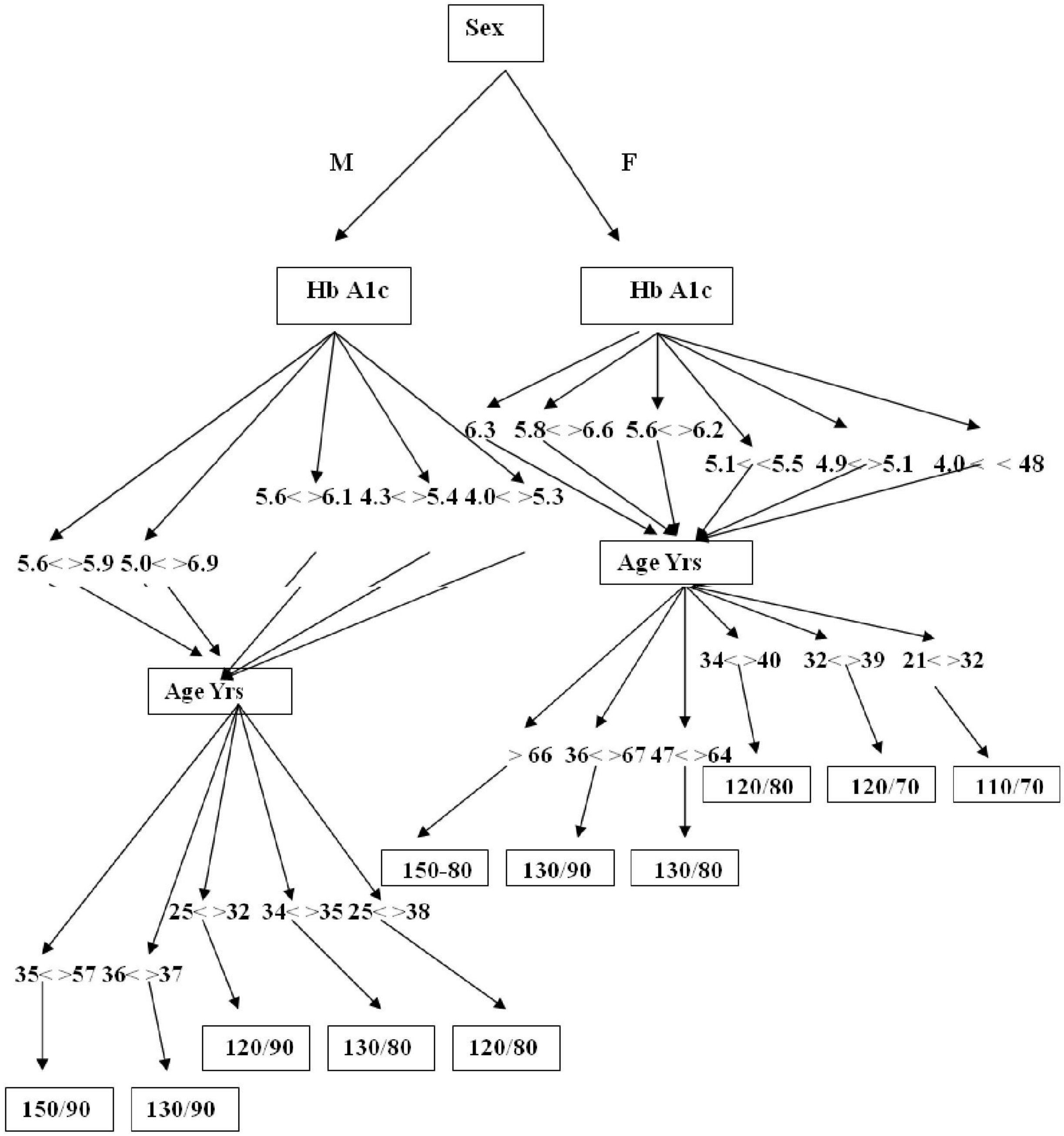

Figure 1. Decision Tree (classifier) to classify the types of Bp systolic diastolic

This section will highlight methodologies dealing with bottom-up specific-to-general approaches to rule induction. The initial state of a decision rule solution is indeed the collection of all individual instances or examples in a training data set, each of which may be thought of as a highly specialized decision rule. Most decision rule modeling systems employ a search process to evolve this set of highly specific and individual instances to more general rules. This search process is iterative, and usually terminates when rules can no longer be generalized, or some other alternate stopping criteria satisfied. As in the case of decision tree building, noise in the data may lead to over fitted decision rules, and various pruning mechanisms have been developed to deal with over fitted decision rule solutions. In the above tree have four leaf nodes. In decision tree, each leaf node represent a rule then have the following rule corresponding to the four give.

Rule induction methods attempt to find a compact \covering" rule set that completely partitions the examples into their correct classes. The covering set is found by heuristically searching for a single \best" rule that covers cases for only one class. Having found a \best" conjunctive rule for a Bp systolic diastolic, the rule is added to the rule set, and the cases satisfying it are removed from further consideration. The process is repeated until no cases remain to be covered.

Rules

Rule11 IF F and 4.0 < Hb A1c < 48% and 21< Age Yrs >32 Then Bp systolic diastolic is 110/70.

Rule2 IF F and 4.9< Hb A1c >5.1% and 32< Age Yrs >39 Then Bp systolic diastolic is 120/70.

Rule3 IF F and 5.1< Hb A1c <5.5% and 34< Age Yrs >40 Then Bp systolic diastolic is 120/80.

Rule4 IF F and 5.6< Hb A1c >6.2% and 47< Age Yrs >64 Then Bp systolic diastolic is 130/80.

Rule5 IF F and 5.8< Hb A1c >6.6% and 36< Age Yrs >67 Then Bp systolic diastolic is 130/90.

Rule6 IF F Hb A1c >6.3% and Age Yrs >66 Then Bp systolic diastolic is 150/90.

Rule7 IF M and 4.0< Hb A1c >5.3% and 25< Age Yrs >38 Then Bp systolic diastolic is 120/80.

Rule8 IF M and 4.3< Hb A1c >5.4 % and 34< Age Yrs >35 Then Bp systolic diastolic is 130/80.

Rule9 IF M and 5.6< Hb A1c >6.1 % and 25< Age Yrs >32 Then Bp systolic diastolic is 120/90.

Rule10 IF M and 5.0< Hb A1c >6.9% and 36< Age Yrs >37Then Bp systolic diastolic is 130/90.

Rule11 IF M and 5.6< Hb A1c >5.9 % and 35< Age Yrs >57 Then Bp systolic diastolic is 150/90.

RESULTS

From implementation of data mining Decision tree to the training set data it appear that.

1. Design Decision tree from dataset seen in Table 1.
2. Design rules to predict the diseases Bp systolic diastolic.
3. The decision tree appears Bp systolic diastolic depend on Hb A1c, Age yrs and Weight.
4. The decision tree appears in increasing Age yrs and Weight reason to increase A Bp systolic diastolic.
5. The decision tree good way to predict the Bp systolic diastolic depending on data in training.
6. The decision tree appear strong relation between Age Yrs, WE. Kgm, Hb A1c and Bp systolic diastolic.

Conclusion

The rise in attention and focus on decision support solutions using data mining techniques has refueled a big interest in classification, particularly symbolic techniques. This paper has attempted to provide the reader with the key issues of decision tree and decision rule modeling techniques. There is strong data as training data se obtained from collected from 100 healthy subjects (50 females and 50 females). The Data Mining was a success to solve the problem. Applied the concepts decision tree to real data and gained a working knowledge of data mining techniques. In this designed rules from this rules can predict the type of Hb A1c effect on the Bp systolic diastolic and then effect on the patients. The decision tree find relation between Bp systolic diastolic, Age Yrs and Bp systolic diastolic. Therefore fundamental concepts of extracting knowledge from data should be a goal for discovering important information.

REFERENCES

Milley, A. 2000. Healthcare and data mining. *Health Management Technology*, 21(8), 44-47.

Salwa H.N., Emad Abdul-Rehman, 2010. "Determining the reference Rang Values of Glycosylated Hemoglobin (Hb1c) by Immunoiturbid Assayin Iraqi Population" *Journal of* College of science.

Rafalski, E. 2002. Using data mining and data repository methods to identify marketing opportunities in healthcare. *Journal of Consumer Marketing*, 19(7), 607-613..

Cios, K.J. & Moore, G.W. 2002. Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, 26(1), 1-24.

Knowledge Seeker Data Mining Tool. “Angoss Home Page.” http://www.angoss.com/

Paddison, N. 2000. Index predicts individual service use. *Health Management Technology*, 21(2), 14-17.

Schuerenberg, B.K. 2003. An information excavation. *Health Data Management*, 11(6), 80-82.

Ludwig, L., Flies, D. and Wilson, A. 2000. “Data Mining Techniques Applied to the Relationship of Latitude and the Lifespan of Aspen Trees” http://epoxy.mrs.umn. edu ~ludwigl/datamining/research.pdf

Ministry of health

****