

Clinical Data Mining and Knowledge Discovery

V. Prasathkumar¹, B. Sandhiya²

Assistant Professor

Department of Information Technology

Sri Shakthi Institute of Engineering and Technology, Coimbatore

Corresponding authors' email id: *prasathkumar@siet.ac.in¹, sandhiya@siet.ac.in²*

Abstract

Applying data mining in the medical field is a very challenging undertaking due to the idiosyncrasies of the medical profession. By introducing status and progress of data mining and analyzing characteristics of medical data, a mathematical model of medical data mining based on computation intelligence such as data preprocessing, statistical techniques, artificial neural network, decision tree and manifold learning have been introduced. This paper focusing on data mining on clinical datasets, examining of high risk population with hypertension, cancer and child obesity etc., the main objective of data mining in clinical data warehouse had been an appropriate and sufficiently sensitive method to analyze the outcomes of effective treatments and reduce costs. Medical diagnosis is extremely important but complicated task that should be performed accurately and efficiently. In contrast, the slight difference could change the balance between life or death. Towards this direction, in this study a context-aware approach is proposed, aiming to provide medical supervisors with a series of applications and personalized services targeted to exploit the multi parameter analysis for the assessment of medical –related risk factors targeting in the reduction of diseases.

Keywords: *Data Mining, Clinical data, Regression Analysis, Artificial Neural Network, Decision Tree*

INTRODUCTION

Data mining is the analysis of observational datasets to find unsuspected relationships and summarize data in novel ways that are both understandable and useful pattern. Data mining also can discover valuable and hidden knowledge from databases. In the clinical mining the discovered knowledge can be used to improve the quality of life by reduce the number of adverse drug effect, and to suggest less expensive therapeutically equivalent alternatives.

The following are some of the important areas of interests where data mining can be tremendous use in health care management.

- Data modeling for health care applications
- Forecasting treatment costs and demand of resources
- Anticipating patient's future behaviors with their history

The major steps in KDD process is data preparation, data modeling and post-processing and model evaluation.

OBJECTIVES

The objectives of this paper are the following:

1. To enumerate current uses and highlight the importance of data mining in medicine and public health,
2. To find data mining techniques used in other fields that may also be applied in the health sector.
3. To identify issues and challenges in data mining as applied to the medical practice.

DATA MINING TECHNIQUES IN HEALTH CARE

Different data mining techniques are used depends on the applications of mining and its accuracy.

A) Regression analysis

Regression modeling represents a powerful and elegant method for estimating the value of a continuous target variable. The analysis used to model the relationship between one or more independent or predictor values and dependent or response variables. In the present context of data mining the predictor variables or attributes of interest describing the topple which are known as values. The response variable is what we want to predict.

A regression task begins with a dataset in which the target values are known. In the

present investigation the dataset values of 'of patients treated in each group and in each mode of treatment.

The straight line regression analysis involves a responsible variable 'y' and a single predictor variable 'x'. It is the simplest form of regression and models 'y' as linear regression function of 'x':

$$Y=b+wx$$

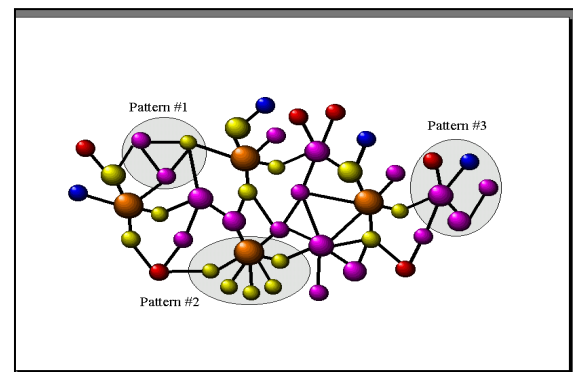
Where the variance of y is assumed to be constant, and b and w are regression coefficients specifying the y-intercept and slope of the line respectively. the regression coefficient 'w' and 'b' can also be thought as weights, so that we can write $Y=w_0+ w_1 x$

B) Neural Networks

An artificial neural network (ANN), usually called neural network (NN), is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural

networks are non-linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs or to find patterns in data.

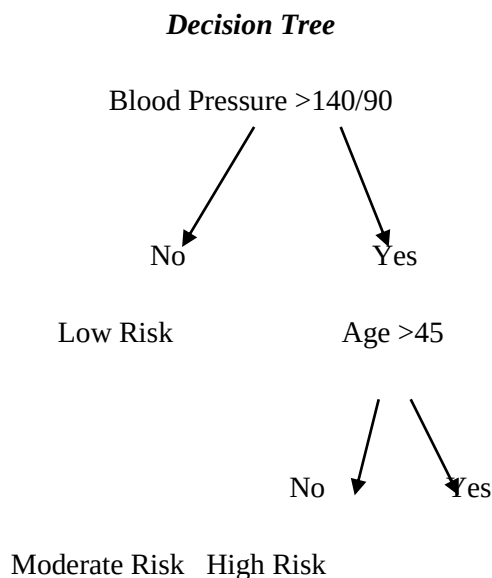
The human brain has many disadvantages over modern parallel computers such as massive parallelism, learning ability, generalization ability, adaptiveness, fault tolerance, low energy consumption, distributed representation and distributed computation. ANN is based on biological neural networks in the human brain as shown in figure:



C) Decision Tree

Decision trees are powerful classification algorithms that are popular in information systems. Widely used decision tree algorithms include Iterative Dichotomiser 3 (ID3), assistant algorithm, c4.5, c5 and CART. This technique performs separate observation in branches recursively to construct a tree for prediction. Splitting algorithms-i.e information gain, Gini

Index and Chi-squared test are used to identify a variable and its corresponding threshold and split the input observation into two or more subgroups. then the steps are repeated until a complete tree is built. The reason for the splitting algorithm is to find the variable threshold pair(s) which maximizes the order of the samples. Figure 2 describes a structure of a decision tree.



D) Rule set classifiers

Complex decision trees can be difficult to understand, for instance information about one class is usually distributed throughout the tree. C4.5 introduced an alternative formalism consisting of a list of rules of the form “if A and B and C and ---then class X”, where rules for each class are grouped together.

E) Fuzzy Systems

Stochastic back propagation algorithm is used for the construction of fuzzy based neural network. The steps involved are as follows: First, initialize weights of the connections with random values. Second for each unit compute net input value, output value and error rate. Third to handle uncertainty for each node, certainty measure for each node is calculated. The level of the certainty is computed using the following conditions.

- a) if $0.8 < c_1$ then there exists very high certainty
- b) if $0.6 < c_0.8$, then there exists high certainty
- c) if $0.4 < c_0.6$, then there exists average certainty
- d) if $0.1 < c_0.4$, then there exists less certainty
- e) if $c_0.1$, then there exists vary less certainty

The network constructed consists of 3 layers namely an input layer, a hidden layer and an output layer

DISEASES AND MINING TECHNIQUES

A) Hypertension:

Hypertension, or high blood pressure, is dangerous because it can lead to strokes, heart attacks, heart failure, or kidney disease and many more disease ailments. High blood pressure is often called the “silent killer” because it has no symptoms and can go undetected for years. The goal of hypertension treatment is to lower high blood pressure and protect important body organs, like the brain, heart, and kidneys from damage.

Treatment for hypertension has been associated with reductions in stroke (average of 35–40%), heart attack (20–25%), and heart failure (more than 50%) according to literature review. Hypertension is widely considered to be one of the most important risk factors for these diseases and is strongly associated with death from stroke, congestive heart failure and coronary heart disease. The high prevalence of the disease and rapidly growing number of older people suggest that hypertension is an enormous public health problem and yet detection and treatment remain relatively low.

High blood pressure is classified as:

- Normal blood pressure: less than 120/80 mmHg
- Hypertension: greater than 140/90 mmHg

- Prehypertension: 120–139/80–89 mmHg
- Stage 1 Hypertension: 140–159/90–99 mmHg
- Stage 2 Hypertension: 160 or greater/100 or greater mmHg

All patients with blood pressure readings greater than 120/80 should be encouraged to make lifestyle modifications, such as eating a healthier diet, quitting smoking, and undergo more physical exercise. Treatment with medication is recommended to lower blood pressure to less than 140/90 mmHg. For patients who have diabetes or chronic kidney disease, the recommended blood pressure is less than 130/80 mmHg.

The study reveals that results on hypertension with urban population around 30% and rural 25%. And the age factor prevail that >18 years with 5%, and 25% with the age group of 40 years. One solution can be obtained for classify the people and age group and gender regression analysis can be used. The age group can be classified into two age groups as young and old. The young group is denoted as ‘Y’ and the old age group is denoted as ‘O’. The prediction is denoted as ‘P’. The young group includes the age group of (15-25, 25-34 and 35-44) and the

old group includes the age group of (35-44,45-54 and 55-64) in both young and old age group includes the '35-44' as common age group. It is the upper limit for young and lower limit for old group.

The equation that used to calculate the prediction of both young and old age groups is given below.

$$P(Y) = [p(1) + p(2) + p(3)] \text{ ----- (1)}$$

$$P(O) = [p(3) + p(4) + p(5)] \text{ ----- (2)}$$

B) Cancer:

Cancer imaging is a vast field focused on non-invasive measurement for cancer detection, diagnosis and grading, and to monitor response to therapy by measuring tumor size and detecting regrowth. Having started out as mainly anatomical measurement for anomaly detection as an indicator of malignancy, for instance using ultrasound and magnetic imaging(MRI), cancer imaging has developed into multiple modalities of physiological imaging including spectroscopy, with Magnetic Resonance Spectroscopy(MRS) and spatially localized Chemical Shift Images(CSI),with Positron-Emission Tomography(PET) and Single-Photon Emission CT(SPECT) clinically widely used to measure glucose metabolism and therefore find regions with exceptionally high energetic rates ,which are

characteristics of tumor growth. An accessible review of machine learning applications to detection and diagnosis of disease may be found.

A recent development is the use of specific functional imaging for tumor delineation, that is to say, to measure the contour during surgical planning and to identify tumor striations that may not be apparent clinically even during re-section. Initial efforts in this direction applied blind signal separation to single-voxel spectra to separate infiltrating tumor mass compared with necrotic tissue.

However, tumors are not homogeneous, often mixing different types of tissue and sometimes with different grades of malignancy. One way to resolve this mixture is to segment the tumor composition, which is called no sological imaging. a related and very important clinical area is follow-up imaging to determine tumor progression and to differentiate recurrent tumor growth from treatment induced tissue changes.

K-means clustering algorithm can be used to analyze cervical cancer patients and found that clustering found better predictive results than existing medical opinion. They found a set of interesting attributes that could be used by doctors as

additional support on whether or not to recommend a biopsy for a patient suspected of having the cervical cancer.

C) Childhood Obesity

Obesity can be defined as an excessive accumulation of body fat. Prediction means indication in advance based on observation, experience, or scientific reason. A childhood obesity prediction system, therefore, is a system to identify young children who are at risk of excessive body fat (obesity) by predicting from their early growth records. Using data mining, childhood obesity predictions are possible. We present three data mining techniques that have been used for childhood obesity prediction systems.

For the Artificial Neural Network (ANN) childhood obesity prediction system, the first stage is data about pregnancy course; morphological changes of the mother, and her psychological attitude were collected and observed. In the second stage, observations of newborns took place. In the third stage, children from 1 to 12 months of age were monitored along with their nutrition and vaccinations. Observation was repeated when they were 3 years old and again they were at the entrance of the grammar school, with periodic health examinations by specialists. The parents also took part in

questionnaires about their health and their economic and social status. The reason for such observation is to identify risk factors of childhood obesity and children's tendency towards obesity.

The first risk factor is the family itself: history of obesity in the family, parental disease, the mother's weight during pregnancy, family nutrition style, education and level of physical activity. The second factor is lifestyle: good and bad habits, handling in early years of life, nutrition, physical activities, illness and personal character of the child. The third factor is the social environment: family size, income and status. These risk factors are used as inputs in the ANN childhood obesity prediction system.

Naïve Bayes Classifiers – In this system, data were collected from children during the first 2 years of life. Using Naïve Bayes classifiers, the collected data were employed to predict whether the children belonged to the lean or obese class. Several attributes were chosen for the training process: gender, adjusted Standard Deviation Scores (SDS) birth weight; adjusted SDS weight gain from birth to 6 weeks, adjusted SDS length at 6 weeks, adjusted SDS height at 2 years, length of gestation, adjusted SDS BMI at 2 years, BMI at 8 months, and whether the child

was obese or lean at 3 years old. To evaluate prediction success rate, four parameters were studied:

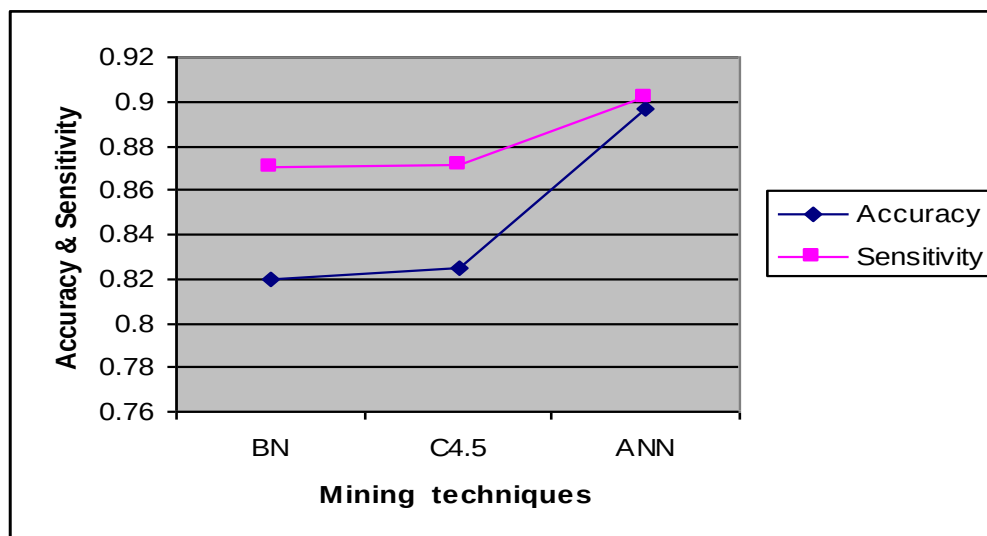
Sensitivity, specificity, positive predictive value and negative predictive value. The most important parameters are the sensitivity value, which refers to the number of correctly predicted obese cases compared to total obese cases. A higher sensitivity value means that the prediction success rates, tests were done when the children were 6 weeks, 8 months, and 2 years old.

Decision Tree – Decision tree can be used for the weight management counseling system for obese children. The system contains two **modules**: a Knowledge Management Module and Case Based Reasoning Module. The Knowledge Management Module has three **components**: a data repository, a knowledge miner and a case based rule structure. In the Knowledge miner component, the decision tree technique is used to build trees from a database of cases.

The decision tree technique is also used in the Case Based Reasoning Module to find possible solutions to a new child case.

COMPARISON OF MODELS

Every model is evaluated based on the two measures- accuracy and sensitivity. Bayesian model achieved classification accuracy of 0.82 with a sensitivity of 0.87. The Decision trees (C4.5) achieved a classification accuracy of 0.825 with a sensitivity of 0.8717. However, the Artificial Neural Network Model (ANN) performed the best of all three models evaluated. It is achieved with a classification accuracy of 0.897 with a sensitivity of 0.9017. Since syndrome is a combination of symptoms. So syndrome is identified by human brain and neural network is considered as best modeler of human brain.



Comparison of three models

CONCLUSION

By overviewed the problem of summarizing different algorithms for predicting combinations of several target attributes. Continuous data can also be used instead of just categorical data. Text mining also can be used to mine the vast amount of unstructured data available in healthcare databases. Further work will be mainly focus on carrying out depth analysis of manifold learning results, and play emphasis on not only potential transformation probability but also quantify relationship in low-dimensional embedding between different diseases. The aim of all these work is to apply theories and methods of information process and data mining to disease risk factors discovery and health early warning , and achievements will definitely promote invaluable development of public health.

REFERENCES

1. D.Barlow, Anxiety and its Disorders: the Nature and Treatment of Anxiety and Panic, 2nd ed.New York:Guilford press,2002,pp.1-36
2. P.Sajda,"Machine learning for detection and diagnosis of disease", annu.Rev.Biomed.Eng.,Vol.8,pp.537-565,2006.
3. Cheng, T.H., Wei, C.P., Tseng, V.S. (2006) Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches. Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06). Wong, W.K

4. A.K.Jain,J.Mao, K.M. Mohiuddin,
“Artificial neural network : a
tutorial”,IEEE, Michigan state
University, 1996,pp.31-44.

5. W.Peng, H.Zhou, J.Chen, “An
implementation of ID3 decision
tree learning algorithm”, Project of
Comp9417, Machine Learning,
university of New South Wales,
Sydney, unpublished.

6. E.Le Doux , “Brain mechanisms of
emotion and emotional learning
current opinions in neurobiology”,
Curr.Opin.Neurobiol.,vol.2, no.2,
pp.191-197,1992.

7. Cao, X., Maloney, K.B. and
Brusic, V. (2008). Data mining of
cancer vaccine trials: a bird's-eye
view. Immunome Research, 4:7.
DOI:10.1186/1745-7580-4-7