Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease

Yanwei Xing, Jie Wang and Zhihong Zhao
Guanganmen Hospita
Chinese Academy of Medical Sciences
Beijing 100053, China
xingyanwei12345@163.com

Yonghong Gao
Dongzhimen hospital
Beijing university of Chinese medicine
Beijing 100700, China

Abstract

The prediction of survival of Coronary Heart Disease (CHD) has been a challenging research problem for medical society. The goal of this paper is to develop data mining algorithms for predicting survival of CHD patients based on 1000 cases .We carry out a clinical observation and a 6-month follow up to include 1000 CHD cases. The survival information of each case is obtained via follow up. Based on the data, we employed three popular data mining algorithms to develop the prediction models using the 502 cases. We also used 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes. The results indicated that the SVM is the best predictor with **92.1％** accuracy on the holdout sample artificial neural networks came out to be the second with91.0％ accuracy and the decision tress models came out to be the worst of the three with 89.6% accuracy. The comparative study of multiple prediction models for survival of CHD patients along with a 10-fold cross-validation provided us with an insight into the relative prediction ability of different data.

1. Introduction

Coronary heart disease (CHD) is one of leading causes of morbidity and mortality in China [1]. Each year, about 1 million have heart attacks and 1 million die of CHD-related causes. Furthermore, it is on the rise and has become a true pandemic that respects no borders.

Although CHD research is generally clinical or biological in nature, data driven statistical research is becoming a common complement. In medical domains where data and statistics driven research is successfully applied, new and novel research directions are identified for further clinical and biological research. The work of data driven study in Reference [2] found statistical evidence to tie the gene to the disease, and now researchers are looking for biological and clinical evidence to support his theory.

Predicting the outcome of a disease is one of the most interesting and challenging tasks in which to develop data mining applications. Survival analyses is a field in medical prognosis that deals with application of various methods to historic data in order to predict the survival of a particular patient suffering from a disease over a particular time period. With the increased use of computers powered with automated tools, storage and retrieval of large volumes of medical data are being collected and are being made available to the medical research community who has been interested in developing prediction models for survivability. As a result, new research avenues such as knowledge discovery in databases (KDD), which includes data mining techniques, has become a popular research tool for medical researchers who seek to identify and exploit patterns and relationships among large number of variables, and be able to predict the outcome of a disease using the historical cases stored within datasets [3].

It is the combination of the serious effects of CHD, the promising results of prior related research, the potential benefits of the research outcomes and the desire to further understand the nature of CHD that provided the motivation for this research effort. In this paper, we report on our research project where we developed models that predict the survivability of diagnosed cases for CHD. One of the salient features of this research effort is the authenticity and the large volume of data processed in developing these survivability prediction models. Based on surveyed data, we used three different types of classification models: Support vector machine (SVM), artificial neural network (ANN) and decision tree along with a 10-fold cross-validation technique to compare the accuracy of these classification models.

The amount of data being collected and stored in databases (both in medical and in other fields) has increased dramatically due to the advancements in software capabilities and hardware tools that enabled the automated data collection (along with the decreasing trend of hardware and software cost) in the last decade. As a result, traditional data analysis techniques have become inadequate for processing such volumes of data, and new techniques have been developed. A major area of development is called KDD. KDD encompasses variety of statistical analysis, pattern recognition and machine learning techniques. In essence, KDD is a formal process whereby the steps of understanding the domain, understanding the data, data preparation, gathering and formulating knowledge from pattern extraction, and ‘‘post-processing of the knowledge’’ are employed to exploit the knowledge from large amount of

0-7695-3038-9/07 $25.00 © 2007 IEEE
DOI 10.1109/ICCIT.2007.204

recorded data [3]. The step of gathering and formulating knowledge from data using pattern extraction methods is commonly referred to as data mining [4]. Applications of data mining have already been proven to provide benefits to many areas of medicine, including diagnosis, prognosis and treatment

The remainder of this paper is organized as follows. Section 2 provides the reader with the background information onCHD. In Section 3, the prediction results of all three algorithms are presented. In Section 4, we explained in detail the data and data processing, prediction methods and the 10-fold cross-validation algorithm.In Section 5, the conclusion of the research is given.

2. Survival analysis

Survival analysis is a field in medical prognosis that deals with application of various methods to estimate the survival of a particular patient suffering from a disease over a particular time period. ‘‘Survival’’ is generally defined as a patient remaining alive for a specified period of time after the diagnosis of disease. Traditionally conventional statistical techniques such as Kaplan-Meier test and Cox-Propositional hazard models [5] were used for modeling survival. These techniques are conditional probability based models that provide us with a probability estimate of survival. With advances in the field of knowledge discovery and data mining a new stream of methods have came into existence. These methods are proved to be more powerful as compared to traditional statistical methods [6].

Many research projects define survival as a period of 10 years or longer. Survival estimates developed using such a definition of survival may not accurately reflect the current state of treatment and the probability of survival . Recent improvements in early detection and treatment have increased the expectations of survival . As a result, for the purposes of this research effort, we have defined ‘‘survival’’ as any incidence of CHD where the person is still living after sixty months from the date of diagnosis.

3. Three data mining prediction models

We used three different types of classification models: Support vector machine, artificial neural networks, decision trees. These models were selected for inclusions in this study due to their popularity in the recently published literature as well as their better than average performance in our preliminary comparative studies. What follows is a short description of these classification model types and their specific implementations for this research

3.1 Support vector machine

The SVM is a state-of-the-art maximum margin classification algorithm rooted in statistical learning theory [7-8]. SVM performs classification tasks by maximizing the margin separating both classes while minimizing the classification errors. We used sequential minimal optimization algorithm to train the SVM here. Figure 1 shows the topology of SVM.

3.2 Artificial neural networks

Artificial neural networks (ANNs) are commonly known as biologically inspired, highly sophisticated analytical techniques, capable of modeling extremely complex non-linear functions. Formally defined, ANNs are analytic techniques modeled after the processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so-called learning from existing data[9]. We used a popular ANN architecture called multi-layer perceptron (MLP) with back-propagation (a supervised learning algorithm). The MLP is known to be a powerful function approximator for prediction and classification problems. It is arguably the most commonly used and well-studied ANN architecture. Our experimental runs also proved the notion that for this type of classification problems MLP performs better than other ANN architectures such as radial basis function (RBF), recurrent neural network (RNN), and self-organizing map (SOM). In fact, Hornik et al. [10] empirically showed that given the right size and the structure, MLP is capable of learning arbitrarily complex nonlinear functions to arbitrary accuracy levels. The MLP is essentially the collection of nonlinear neurons (a.k.a. perceptrons) organized and connected to each other in a feedforward multi-layer structure.

3.3 Decision trees

Decision trees are powerful classification algorithms that are becoming increasingly more popular with the growth of data mining in the field of information systems. Popular decision tree algorithms include Quinlan’s ID3, C4.5, C5 [11], and Breiman et al.’s CART [12]. As the name implies, this technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy. In doing so, they use mathematical algorithms (e.g.,information gain, Gini index, and Chi-squared test) to identify a variable and corresponding threshold for the variable that splits the input observation into two or more subgroups. This step is repeated at each leaf node until the complete tree is constructed.The objective of the splitting algorithm is to find a variable-threshold pair that maximizes the homogeneity (order) of the resulting two or more subgroups of samples. The most commonly used mathematical algorithm for splitting includes Entropy based information gain (used in ID3, C4.5, C5), Gini index (used in CART), and Chi-squared test (used in CHAID). Based on the favorable prediction results we have obtained from the preliminary runs, in this study we chose to use C5 algorithm as our decision tree method, which is an improved version of C4.5 and ID3 algorithms [11].

4. Survival Data concerning on CHD

The study population consisted of 1000 consecutive patients who underwent coronary angiography for known coronary atherosclerosis at Anzhen hospital, capital University of Medical Sciences, Beijing from August 2005 to December 2005.

4.1 Inclusion and exclusion criteria

Diagnostic criteria for ACS and SAP are according to the American college of Cardiology/American Heart Association (ACC/AHA) guideline for the management of patients with unstable angina and non-ST-segment elevation myocardial infarction in 2002, the ACC/AHA guidelines for the management of patients with acute myocardial infarction in 1999, and the ACC/AHA/ACP-ASIM guidelines for the management of patients with chronic stable angina in 1999. Patient with pregnancy, acute infection, serious hepatopathy or nephropathy, severe heart failure, malignant tumor, rheumatism, cerebral stroke, trauma, or a history of surgery during the previous month were excluded.

4.2 Basic information of the data

Table 1 is the variables included in the survey.

Variables	Type
TNF	Continuous
IL6	Continuous
IL8	Continuous
HICRP	Continuous
MPO1	Continuous
TNI2	Continuous
Sex	Category
Age	Discrete
Smoke	Category
Hypertension	Category
Diabetes	Category
Survival	Category

We used a binary categorical survival variable, which was calculated from the variables in the raw dataset, to represent the survivability where survival is represented with a value of ‘‘1’’ and non-survival is represented with ‘‘0’’. Among 1000 CHD cases, the death is 202 cases and the survival is 798 cases.

5. Results

In this paper, we employed three hackneyed performance measures: accuracy, sensitivity and specificity. A distinguished confusion matrix is obtained to calculate the three measures. Confusion matrix is a matrix representation of the classification results. the upper left cell denotes the number of samples classifies as true while they were true (i.e., true positives), and lower right cell denotes the number of samples classified as false while they were actually false (i.e., true false). The other two cells (lower left cell and upper right cell) denote the number of samples misclassified. Specifically, the lower left cell denoting the number of samples classified as false while they actually were true (i.e., false negatives), and the upper right cell denoting the number of samples classified as true while they actually were false (i.e., false positives).Once the confusion matrixes were constructed, the accuracy, sensitivity and specificity are easily calculated as: sensitivity = TP/(TP + FN); specificity = TN/(TN + FP). Accuracy = (TP + TN)/(TP + FP + TN + FN); where TP, TN, FP and FN denotes true positives, true negatives, false positives and false negatives, respectively.

10-fold cross validation is used here to minimize the bias produced by random sampling of the training and test data samples. Extensive tests on numerous datasets, with different learning strategies, have shown that 10 is about the right number of folds to get the best estimate of error, and there is also some theoretical evidence that backs this up [13].

Every model was evaluated based on the three measures discussed above (classification accuracy, sensitivity and specificity). The results were achieved using average value of 10 fold cross-validation for each algorithm. We found that the MLP achieved classification accuracy of 91.0%with a sensitivity of 91.73% and a specificity of 88.12%. The C5 achieved a classification accuracy of 89.6% with a sensitivity of 90.98% and a specificity of 84.16%.However, The SVM preformed the best of three models evaluated. It achieved a classification accuracy of 92.1%with a sensitivity of 92.87% and a specificity of 89.11%. Table 2 shows the complete set of results in a tabular format. The detailed prediction results of the validation datasets are presented in form of confusion matrixes.

6. Conclusion

In this paper, we report on a research effort where we developed several prediction models for CHD survivability. Specifically, we used three popular data mining methods。We survey a data of 1000 CHD cases with 11 attributes. In this research, we defined survival as any incidence of CHD where person is still alive after 6 months from the date of diagnosis. We used a binary categorical survival variable, which was calculated from the variables in the raw dataset, to represent the survivability where survival is represented with a value of ‘‘1’’ and non-survival is represented with ‘‘0’’. In order to measure the unbiased prediction accuracy of the three methods, we used a 10-fold cross-validation procedure. That is, we divided the dataset into 10 mutually exclusive partitions (a.k.a. folds) using a stratified sampling technique. Then, we used 9 of 10 folds for training and the 10th for the testing. We repeated this process for 10 times so that each and every data point would be used as part of the training and testing datasets. The accuracy measure for the model is calculated by averaging the 10 models performance numbers. We repeated this process

for each of the three prediction models. This provided us with a less biased prediction performance measures to compare the tree models. The aggregated results indicated that the SVM performed the best with a classification accuracy of 92.1%, the ANN model (with multi layered perceptron architecture) came out to be second best with a classification accuracy of 91.0%, and the decision tree model came out to be the worst with a classification accuracy of 89.6%. The results showed here make clinical application more accessible, which will provide great advance in healing CHD.

Acknowledgement

The work has been supported by the National Basic Research Program of China (973 Program) under grant No. (2003CB517103)

Reference

1. World Health Organization. World Health statistics Annual, Geneva, Switzerland: World Health Organization (2006)
2. Ritter M. Gene tied to manic-depression. Newspaper article in Tulsa World June 16, 2003: D8.
3. Lavrac N. Selected techniques for data mining in medicine. Artif Intell Med 1999;16:3—23.
4. Cios KJ, Moore GW. Uniqueness of medical data mining. Artif Intell Med 2002;26:1—24.
5. Cox DR. Analysis of survival data. London: Chapman & Hall; 1984.
6. Ohno-Machado L. Modeling medical prognosis: survival analysis techniques. J Biomed Inform 2001;34:428—39.
7 .Vapnik, K. : Statistical learning theory. Wiley, New York (1998)
8. Graf, A., Wichmann. F., Bulthoff .H., etc. : Classification of faces in man and machine. Neural Computation, 18 (2006) 143-165.
[9] Haykin S. Neural networks: a comprehensive foundation.
New Jersey: Prentice Hall; 1998.
[10] Hornik K, Stinchcombe M, White H. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward network. Neural Netw 1990;3:359—66.
[11] Quinlan J. C4.5: programs for machine learning. San Mateo,CA: Morgan Kaufmann; 1993.
[12] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Monterey, CA: Wadsworth & Brooks/ Cole Advanced Books & Software; 1984.
[13]Witten, I.H., FrankMichalewicz, E. Z. : Data Mining: Practical machine learning tools and techniques 2nd ed. Morgan Kaufmann, San Francisco (2005)

Table 2. Confusion matrix shows the classification of the
cases in the test dataset. In confusion matrix, the columns denote the actual cases and the rows denote the predicted. Sensitivity = TP/(TP + FN); Specificity = TN/(TN + FP). Accuracy = (TP + TN)/(TP + FP + TN + FN);

Algorithm	TP	FN	Sensitivity	Specificity	Accuracy
	TF	TN			
SVM	**741**	57	**92.87%**	**89.11%**	**92.1%**
	22	**180**			
MLP	732	87	91.73%	88.12%	91.0%
	24	178			
C5	726	87	90.98%	84.16%	89.6%
	32	170			

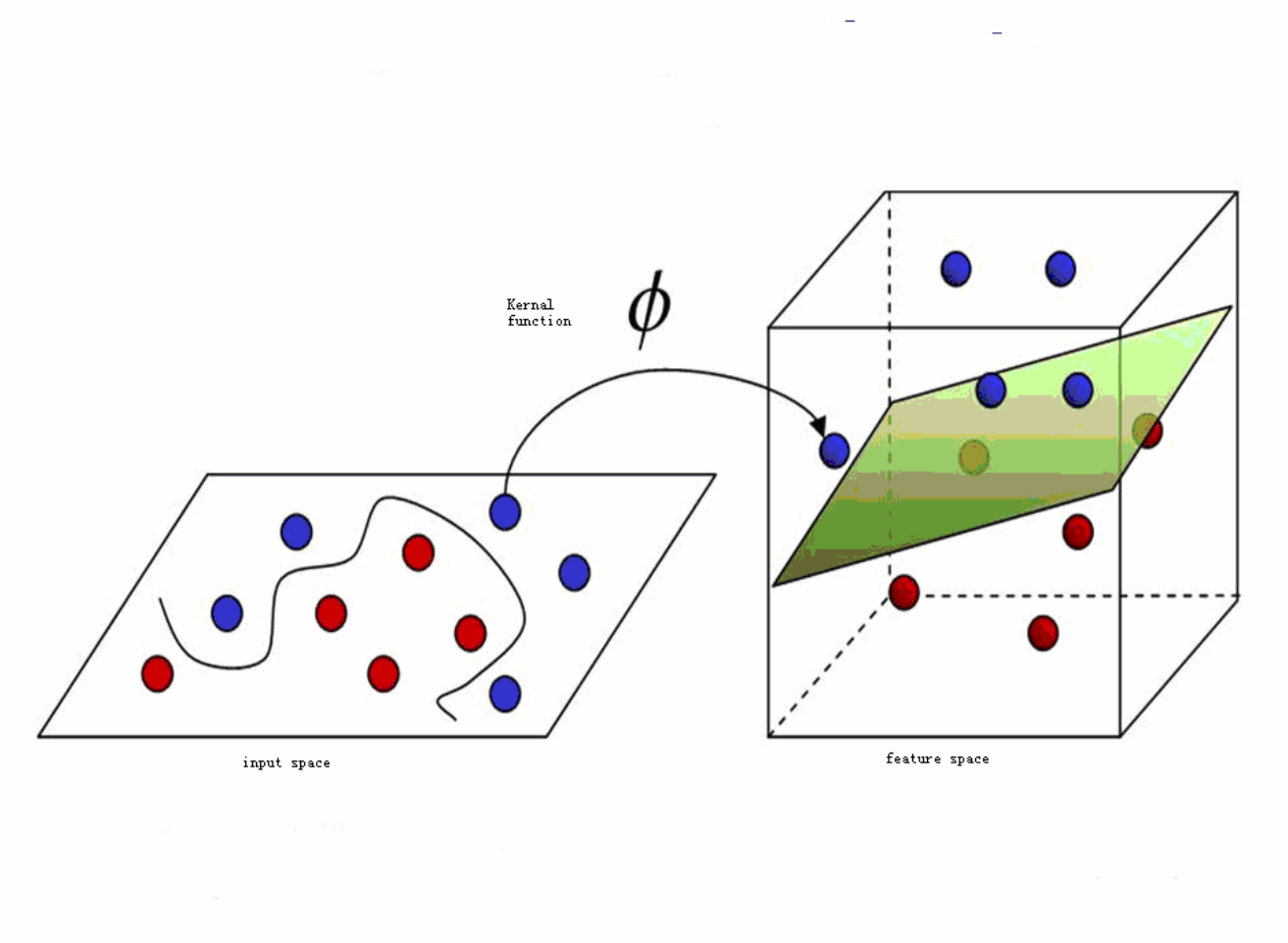

Figure 1. The topology of SVM