Global Journals LaTeX JournalKaleidoscope™

Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Diagnosis Of Heart Disease Using Datamining Algorithm

Asha Rajkumar M.phil (Computer Science)[1] and G.Sophia Reena (HOD of BCA Department)[2]

[1] Bharathiar university

Received: 5 February 2010 Accepted: 1 March 2010 Published: 15 March 2010

Abstract

The diagnosis of heart disease is a significant and tedious task in medicine. The healthcare industry gathers enormous amounts of heart disease data that regrettably, are not ?mined? to determine concealed information for effective decision making by healthcare practitioners. The term Heart disease encompasses the diverse diseases that affect the heart. Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The reduction of blood and oxygen supply to the heart leads to heart disease. In this paper the data classification is based on supervised machine learning algorithms which result in accuracy, time taken to build the algorithm. Tanagra tool is used to classify the data and the data is evaluated using 10-fold cross validation and the results are compared.

Index terms— Naive Bayes, k-nn, Decision List, Tanagra tool

1 INTRODUCTION

he term heart disease applies to a number of illnesses that affect the circulatory system, which consists of heart and blood vessels. It is intended to deal only with the condition commonly called "Heart Attack" and the factors, which lead to such condition. Cardiomyopathy and Cardiovascular disease are some categories of heart diseases.The term -cardio vascular disease? includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Cardiovascular disease (CVD) results in severe illness, disability, and death. Narrowing of the coronary arteries results in the reduction of blood and oxygen supply to the heart and leads to the Coronary heart disease (CHD). Myocardial infarctions, generally known as a heart attacks, and angina pectoris, or chest pain are encompassed in the CHD. A sudden blockage of a coronary artery, generally due to a blood clot results in a heart attack. Chest pains arise when the blood received by the heart muscles is inadequate. High blood pressure, coronary artery disease, valvular heart disease, stroke, or rheumatic fever/rheumatic heart disease are the various forms of cardiovascular disease. ? high blood pressure ? diabetes ? smoking ? high cholesterol ? family history of heart attacks at ages younger than 60 years, one or more previous heart attacks, male gender ? obesity ? Postmenopausal women are at higher risk than premenopausal women. This is thought to be due to loss of the protective effects of the hormone estrogen at menopause. It was previously treated by hormone supplements (hormone replacement therapy, or HRT).

T However, research findings have changed our thinking on HRT; long-term HRT is no longer recommended for most women. ? Use of cocaine and similar stimulants.

2 III. EXTRACTION OF HEART DISEASE

3 DATAWAREHOUSE

The heart disease data warehouse contains the screening the data of heart patients. Initially, the data warehouse is preprocessed to make the mining process more efficient. In this paper Tanagra tool is used to compare the performance accuracy of data mining algorithms for diagnosis of heart disease dataset. The pre-processed data warehouse is then classified using Tanagra tool. The feature selection in the tool describes the attribute status of the data present in the heart disease. Using supervised machine learning algorithm such as Naive Bayes, k-nn and Decision list and the result are compared.Tanagra is a collection of machine learning algorithms for data mining tasks. The algorithms can be applied directly to a dataset. Tanagra contains tools for data classification, statistics, clustering, supervised learning, meta-supervised learning and visualization. It is also well suited for developing new machine learning schemes. This paper concentrates on functional algorithms like Naive Bayes, k-nn, and Decision list.

IV.

4 TANAGRA

Tanagra is a data mining suite build around graphical user interface. Tanagra is particularly strong in statistics, offering a wide range of uni-and multivariate parametric and nonparametric tests. Equally impressive is its list of feature selection techniques. Together with a compilation of standard machine learning techniques, it also includes correspondence analysis, principal component analysis, and the partial least squares methods. Tanagra is more powerful, it contains some supervised learning but also other paradigms such as clustering, supervised learning, meta supervised learning, feature selection, data visualization supervised learning assessment, statistics, feature selection and construction algorithms.The main purpose of Tanagra project is to give researchers and students an easy-to-use data mining software, conforming to the present norms of the software development in this domain , and allowing to analyze either real or synthetic data. Tanagra can be considered as a pedagogical tool for learning programming techniques. Tanagra is a wide set of data sources, direct access to data warehouses and databases, data cleansing, interactive utilization.

5 1) Classification

The basic classification is based on supervised algorithms. Algorithms are applicable for the input data. Classification is done to know the exactly how data is being classified. Meta-supervised learning is also supported which shows the list of machine learning algorithms. Tanagra includes support for arcing, boosting, and bagging classifiers. These algorithms in general operate on a classification algorithm and run it multiple times manipulating algorithm parameters or input data weight to increase the accuracy of the classifier. Two learning performance evaluators are included with Tanagra. The first simply splits a dataset into training and test data, while the second performs cross-validation using folds. Evaluation is usually described by accuracy, error, precision and recall rates. A standard confusion matrix is also displayed, for quick inspection of how well a classifier works.

6 2) Manifold machine learning algorithm

The main motivation for different supervised machine learning algorithms is accuracy improvement. Different algorithms use different rule for generalizing different representations of the knowledge. Therefore, they tend to error on different parts of the instance space. The combined use of different algorithms could lead to the correction of the individual uncorrelated errors. as a result the error rate and time taken to develop the algorithm is compared with different algorithm.

7 3) Algorithm selection

Algorithm is selected by evaluating each supervised machine learning algorithms by using supervised learning assessment (10-fold cross-validation) on the training set and selects the best one for application on the test set. Although this method is simple, it has been found to be highly effective and comparable to other methods. Several methods are proposed for machine learning domain. The overall cross validation performance of each algorithm is evaluated.

The selection of algorithms is based on their performance, but not around the test dataset itself, and also comprising the predictions of the classification models on the test instance. Training data are produced by recording the predictions of each algorithm, using the full training data both for training and for testing. Performance is determined by running 10-fold cross-validations and averaging the evaluations for each training dataset. Several approaches have been proposed for the characterization of learning domain. the performance of each algorithm on the data attribute is recorded. The algorithms are ranked according to their performance of the error rate.

8 4) Manuscript details

This paper deals with Naive Bayes, K-nn, Decision List algorithm . Experimental setup is discussed using 700 data and the results are compared . The performance analysis is done among these algorithms based on the accuracy and time taken to build the model.

9 V. ALGORITHM USED A Bayes classifier is a simple probabilistic classifier based on applying

Bayes theorem with strong (naive) independence assumptions. In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood. The advantage of using naive bayes is that one can work with the naive Bayes model without using any Bayesian methods.Naive Bayes classifiers have works well in many complex real-world situations. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.A decision list has only two possible outputs, yes or no (or alternately 1 or 0) on any input. a decision list is a question in some formal system with a yes-or-no answer, depending on the values of some input parameters . A method for solving a decision problem given in the form of an algorithm is called a decision procedure for that problem. The field of computational complexity categorizes decidable decision problem. Research in computability theory has typically focused on decision problems These inputs can be natural numbers, also other values of some other kind, such as strings of a formal language. Using some encoding, such as Godel numberings, the strings can be encoded as natural numbers. Thus, a decision problem informally phrased in terms of a formal language is also equivalent to a set of natural numbers. To keep the formal definition simple, it is phrased in terms of subsets of the natural numbers. Formally, a decision problem is a subset of the natural numbers. The corresponding informal problem is that of deciding whether a given number is in the set. Decision problems can be ordered according to many-one reducibility and related feasible reductions such as Polynomial-time reductions. Every decision problem can be converted into the function problem of computing the characteristic function of the set associated to the decision problem. If this function is computable then the associated decision problem is decidable. However, this reduction is more liberal than the standard reduction used in computational complexity (sometimes called polynomial-time many-one reduction). The k-nearest neighbor's algorithm (k-NN) is a method for classifying objects based on closest training data in the feature space. k-NN is a type of instance-based learning. The function is only approximated locally and all computation is deferred until classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms. The same method can be used for regression, by simply assigning the property value for the object to be the average of the values of its k nearest neighbors. It can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones The neighbors are taken from a set of objects for which the correct classification (or, in the case of regression, the value of the property) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. The k-nearest neighbor algorithm is sensitive to the local structure of the data . Nearest neighbor rules in effect compute the decision boundary in an implicit manner. It is also possible to compute the decision boundary itself explicitly, and to do so in an efficient manner so that the computational complexity is a function of the boundary complexity. The best choice of k depends upon the data; generally, larger values of k reduce the effect of noise on the classification, but make boundaries between classes less distinct. A good k can be selected by various heuristic techniques, for example, cross-validation. The special case where the class is predicted to be the class of the closest training sample (i.e. when k = 1) is called the nearest neighbor algorithm.The accuracy of the k-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance. Much research effort has been put into selecting or scaling features to improve classification. In binary (two class) classification problems, it is helpful to choose k to be an odd number as this avoids tied votes.Using an appropriate nearest neighbor search algorithm makes k-NN computationally tractable even for large data sets. The nearest neighbor algorithm has some strong consistency results. As the amount of data approaches infinity, the algorithm is guaranteed to yield an error rate no worse than twice the Bayes error rate (the minimum achievable error rate given the distribution of the data. k-nearest neighbor is guaranteed to approach the Bayes error rate, for some value of k (where k increases as a function of the number of data points). Various improvements to k-nearest neighbor methods are possible by using proximity graphs.The training data set consists of 3000 instances with 14 different attributes. The instances in the dataset are representing the results of different types of testing to predict the accuracy of heart disease. The performance of the classifiers is evaluated and their results are analyzed.In general, tenfold cross validation has been proved to be statistically good enough in evaluating the performance of the classifier. The 10-fold cross validation was performed to predict the accuracy of heart disease. The purpose of running multiple crossvalidations is to obtain more reliable estimates of the risk measures. The basic crisis is to predict the accuracy of heart disease that can be stated as follows: particular dataset of heart disease with its appropriate attributes. The main aim is to get a accuracy by classifying algorithms.

10 VI. EXPERIMENTAL SETUP

The data mining method used to build the model is classification. The data analysis is processed using Tanagra data mining tool for exploratory data analysis, machine learning and statistical learning algorithms. The training data set consists of 3000 instances with 14 different attributes. The instances in the dataset are representing the results of different types of testing to predict the accuracy of heart disease. The performance of the classifiers is evaluated and their results are analysed. The results of comparison are based on 10 ten-fold cross-validations. According to the attributes the dataset is divided into two parts that is 70% of the data are used for training and 30% are used for testing.

11 1) Description of dataset

The dataset contains the following attributes: 1) id: patient identification number 2) age: age in year, 3) sex: sex (1 = male; 0 = female), 4) painloc: chest pain location (1 = substernal; 0 = otherwise), 5) painexer (1 = provoked by exertion; 0 = otherwise), 6)

12 VII. CONCLUSION

Data mining in health care management is unlike the other fields owing to the fact that the data present are heterogeneous and that certain ethical, legal, and social constraints apply to private medical information. Health care related data are voluminous in nature and they arrive from diverse sources all of them not entirely appropriate in structure or quality. These days, the exploitation of knowledge and experience of numerous specialists and clinical screening data of patients gathered in a database during the diagnosis procedure, has been widely recognized. This paper deals with the results in the field of data classification obtained with Naive Bayes algorithm, Decision list algorithm and k-nn algorithm, and on the whole performance made known Naive Bayes Algorithm when tested on heart disease datasets. Naive Bayes algorithm is the best compact time for processing dataset and shows better performance in accuracy prediction. The time taken to run the data for result is fast when compared to other algorithms. It shows the enhanced performance according to its attribute. Attributes are fully classified by this algorithm and it gives 52.33% of accurate result. Based on the experimental results the classification accuracy is found to be better using Naive Bayes algorithm compare to other algorithms. From the above results Naive Bayes algorithm plays a key role in shaping improved classification accuracy of a dataset VIII.

1

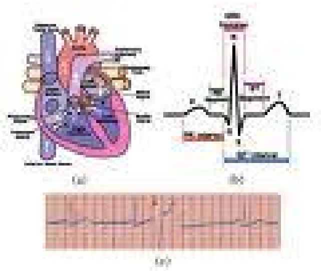

Figure 1: 1)

111

Figure 2: About 1 -?Fig. 1 1)

chart that naive bayes algorithm shows the superior performance compared to other algorithms.

Table Ii. Performance Study Of Accuracy

Algorithm	Values	Recall
		1-
used		precisi
Naïve	Left vent	0.4828
		0.4853
Bayes	hypertrophy	
Decision List	normal St-t-abnormal Left vent hypertrophy normal	0.5705
		0.4753
		0.0000
		1.0000
		0.4897
		0.4855
		0.5705
		0.4688

St-t-abnormal Left vent relrest (1 = relieved after rest; 0 = otherwise), 0.0000 1.0000 KNN 0.4552 0.5479 hy

9) chol: serum cholestoral 10) famhist: family history of coronary artery disease
(1 = yes; 0 = no)
11) restecg: resting electrocardiographic results
? Value 0: normal ? Value 1: having ST-T wave abnormality (T wave inversions and/or ST
? elevation or depression of > 0.05 mV)
? Value 2: showing probable or definite left ventricular hypertrophy ? by Estes' criteria 12) ekgmo (month o

? naive bayes,
? k-nn,
? decision list

Figure 3:

[Knowl. Dat. Eng ()] , *Knowl. Dat. Eng* 1996. 8 (6) p. .

[Chen and Greiner ()] ‘Comparing Bayesian Network Classifiers’. J Chen , R Greiner . *Proc. of UAI-99*, (of UAI-99) 1999. p. .

[Chen et al.] *Data Mining: An Overview from Database Perspective*, M Chen , J Han , P S Yu . Trans: IEEE.

[Guru et al. (2007)] ‘Decision Support System for Heart Disease Diagnosis Using Neural Network’. Niti Guru , Anil Dahiya , Navin Rajpal . *Delhi Business Review* January -June 2007. 8 (1) .

[Agrawal and Srikant (1994)] ‘Fast algorithms for mining association rules in large databases’. R Agrawal , R Srikant . *Proceedings of the 20th International Conference on Very Large Data Bases*, (the 20th International Conference on Very Large Data BasesSantiago, Chile) August 29-September 1994.

[Hospitalization for Heart Attack, Stroke, or Congestive Heart Failure among Persons with Diabetes Special report ()] ‘Hospitalization for Heart Attack, Stroke, or Congestive Heart Failure among Persons with Diabetes’. *Special report*, (New Mexico) 2001 -2003.

[Ordonez ()] ‘Improving Heart Disease Prediction Using Constrained Association Rules’. Carlos Ordonez . *Seminar Presentation atUniversity of Tokyo*, 2004.

[Franck Le Duff et al.] *Predicting Survival Causes After Out of Hospital Cardiac Arrest using 0%*, Cristian Franck Le Duff , Marc Munteanb , Philippe Cuggiaa , Mabob .

[Quinlan ()] J Quinlan . *C4.5: Programs for Machine Learning*, (San Mateo) 1993. Morgan Kaufmann.