2011 21st International Conference on Systems Engineering

Heart disease Classification using Neural Network and Feature Selection

Anchana Khemphila
Software Systems Engineering Laboratory
Department of Mathematics and Computer Science
Faculty of Science, King Mongkut's Institute
of Technology Ladkrabang
Chalongkrung Rd., Ladkrabang, Bangkok 10520, Thailand.
s0067103@kmitl.ac.th

Veera Boonjing
Software Systems Engineering Laboratory
Department of Mathematics and Computer Science
Faculty of Science, King Mongkut's Institute
of Technology Ladkrabang
Chalongkrung Rd., Ladkrabang, Bangkok 10520, Thailand.
kbveera@kmitl.ac.th

***Abstract*—In this study, we introduces a classification approach using Multi-Layer Perceptron (MLP)with Back-Propagation learning algorithm and a feature selection algorithm along with biomedical test values to diagnose heart disease.Clinical diagnosis is done mostly by doctor's expertise and experience.But still cases are reported of wrong diagnosis and treatment.Patients are asked to take number of tests for diagnosis.In many cases,not all the tests contribute towards effective diagnosis of a disease.Our work is to classify the presence of heart disease with reduced number of attributes.Original,13 attributes are involved in classify the heart disease.We use Information Gain to determine the attributes which reduces the number of attributes which is need to be taken from patients.The Artificial neural networks is used to classify the diagnosis of patients.Thirteen attributes are reduced to 8 attributes.The accuracy differs between 13 features and 8 features in training data set is 1.1% and in the validation data set is 0.82%.**

Keywords: **Data mining , classification , Heart disease , Artificial neural networks , Feature Selection , Information Gain**

I. Introduction

Data mining is a crucial step in discovery of knowledge from large data sets. In recent years, Data mining has found its significant hold in every field including health care.Medical history data comprises of a number of tests essential to diagnose a particular disease [1].Clinical databases are elements of the domain where the procedure of data mining has develop into an inevitable aspect due to the gradual incline of medical and clinical research data. It is possible for the healthcare industries to gain advantage of Data mining by employing the same as an intelligent diagnostic tool. It is possible to acquire knowledge and information concerning a disease from the patient specific stored measurements as far as medical data is concerned. Therefore, data mining has developed into a vital domain in healthcare [2].It is possible to predict the efficiency of medical treatments by building the data mining applications. Data mining can deliver an assessment of which courses of action prove effective [3]by comparing and evaluating causes, symptoms, and courses of treatments. The real-life data mining applications are attractive since they provide data miners with varied set of problems, time and again. Working on heart disease patients databases is one kind of a real-life application. The detection of a disease from several factors or symptoms is a multi-layered problem and might lead to false assumptions frequently associated with erratic effects. Therefore it appears reasonable to try utilizing the knowledge and experience of several specialists collected in databases towards assisting the diagnosis process [4], [5].
The researchers in the medical field identify and predict the diseases besides proffering effective care for patients with the aid of data mining techniques. The data mining techniques have been utilized by a wide variety of works in the literature to diagnose various diseases including: Diabetes, Hepatitis, Cancer, Heart diseases and the like. Information associated with the disease, prevailing in the form of electronic clinical records, treatment information, gene expressions, images and more; were employed in all these works. In the recent past, the data mining techniques were utilized by several authors to present diagnosis approaches for diverse types of heart diseases [6], [7] and [8].Hence, more careful and efficient methods of cardiac diseases and periodic examination are of high importance.Our work is an attempt to introduce a classification approach using Multi-Layer Perceptron (MLP)with Back-Propagation learning algorithm and a feature selection using information gain with heart disease patients.
In the next section,we review the the data mining technique.Section 3 explains the data set used and discusses on reducing the number of attributes using Information Gain .Section 4 is result of experiment. Finally, section 5 contains concluding.

II. Data mining technique

Few works have been published Polat et al.(2007) have proposed a new decision making system based on combining of feature selection, fuzzy weighted preprocessing and AIRS classifier to classify the heart disease dataset [13].Resul Dasa, Ibrahim Turkoglub, Abdulkadir Sengurb(2009) diagnosis of valvular heart disease through

978-0-7695-4495-3/11 $26.00 © 2011 IEEE
DOI 10.1109/ICSEng.2011.80

neural networks ensembles [11].M.Anbarasi et. al.(2010) predict heart disease with feature subset selection using genetic algorithm [14]. In this research,we would like to mention about classification using Artificial neural networks with feature selection.

A. Artificial neural networks(ANNs)

Artificial neural networks is inspired by attempts to simulate biological neural systems. The human brain consists primarily of nerve cells called neurons,linked together with other neurons via stand of fiber called axons. Axons are used to transmit nerve impulses from one neuron to another whenever the neurons are stimulated. A neuron is connected to the axons of other neurons via dendrites,which are extensions from the cell body of the neurons.The contact point between a dendrite and an axon is called a synapse.

Multilayer is feed-forward neural networks trained with the standard back-propagation algorithm.It is supervised networks so they require a desired response to be trained.It learns how to transform input data in to a desired response,so they are widely used for pattern classification.With one or two hidden layers,they can approximate virtually any input-output map. It has been shown to approximate the performance of optimal statistical classifiers in difficult problems.The most popular static network in the multilayer.The multilayer is trained with error correction learning,which is appropriate here because the desired multilayer response is the arteriographic result and as such known.Error correction learning works in the following way from the system response at neuron j at iteration t, $y_j(t)$,and the desired response $d_j(t)$for given input pattern an instantaneous error $e_j(t)$is defined by

$$e_j(t) = d_j(t) - y_j(t) \tag{1}$$

Using the theory of gradient descent learning, each weight in the network can be adapted by correcting the present value of the weight with a term that is proportional to the present input and error at the weight, i.e.

$$w_{jk}(t+1) = w_{jk}(t) + \eta\delta_j(t)x_k(t) \tag{2}$$

The $\eta(t)$ is the learning-rate patameter.The $w_{jk}(t)$ is the weight connecting the output of neuron k to the input neuron j at iteration t.The local error$\delta_j(t)$ can be computed as a weighted sum of errors at the internal neurons.

III. Classification accuracy among data mining technique.

A. Description of the data.

This database is taken from the Cleveland Clinic Foundation and was supplied by Robert Detrano, M.D., Ph.D. of the V.A. Medical Center, Long Beach, CA. It is part of the collection of databases at the University of California, Irvine(UCI) collected by David Aha [12]. The aim of the dataset is to classify the presence or absence of heart disease given the results of various medical tests carried out on a patient. The original dataset contains 13 numeric attributes and a fourteenth attribute in dicating whether the patient has a heart condition.This dataset is interesting because it represents real patient data and has been used extensively for testing various data mining techniques.We can use this data together with one of more data mining techniques to help us develop profiles for differentiating individuals with heart disease from those without known heart conditions.This study reviewed the literature and used the following 14 variables as explanatory variables in Table.1.Before building models,the data set were randomly split into two subsets,60 %(n=182) of the the data for training set and 40 %(n=121) of the data for validation set.

Table I
Cardiology Patient Data.

Attribute	Values	Numberic
Age	Numeric	Numeric
Sex	Male,Female	1,0
Chest Pain Type	Angina,Abnormal,NoTang,Asymp	1-4
Blood Pressure	Numeric	Numeric
Cholesterol	Numeric	Numeric
Fasting Blood Sugar	<120 True,False	1,0
Resting ECG	Normal,Abnormal,Hyp	0,1,2
Maximum Heart Rate	Numeric	Numeric
Induced Angina	True,False	1,0
Old Peak	Numeric	Numeric
Slope	Up,Flat,Down	1,2,3
Number Colored Vessels	0,1,2,3	0,1,2,3
Thal	Number,Fix,rev	3,6,7
Concept Class	Healthy,Sick	1,0

B. Classifier evaluation measures.

In our work, context selection is determined by the weight of each low-level context.Instead of learning weights through a generic algorithm or other machine learning method, we use the information gain of each attribute, as its weight. The basic motivation for this study comes from the power of ANN classification algorithm and it uses the concept of information gain as the criterion to select an attribute.IG is usually used for feature set selection so the method applied consists of computing the IG for each field.IG give attribute X with respect to the class attribute Y is the reduction in uncertainty about the value of Y when we know the value of X,I(Y;X).The uncertainty about the value of Y when we know the value of X is given by the conditional entropy of Y given X,H(Y|X).

$$IG(Y;X) = H(Y) - H(Y|X) \quad (3)$$

When Y and X are discrete variables that take values in $y_1...y_k$ and $x_1...x_l$ then the entroy of Y is given by:

$$H(Y) = -\sum_{i=1}^{i=k} P(Y = y_i) log_2(P(Y = y_i)) \quad (4)$$

The condition entropy of Y given X is:

$$H(Y|X) = -\sum_{j=1}^{j=l} P(X = x_j) H(Y|X = x_i)) \quad (5)$$

alternatively the information gain is given by:

$$IG(Y;X) = H(X) + H(Y) - H(X,Y) \quad (6)$$

The ranking of the attributes is then done with respect to the values of IG in a descending order, reflecting the intuition that the higher an IG value, the more information the corresponding attribute has to offer regarding the class. Note that to compute the information gain, data sets with numeric features are required to be discretized. Many alternative methods can be applied for this. In the present work, we simply handle the discretizaiton of continuous valued attributes by partitioning the range of values into a finite number of subsets.

We have trained the classifiers to classify the medical data set as either healthy or sick. The accuracy of a classifier can be computed using sensitivity and specificity.For the given two classes, we consider in terms of positive tuples (diagnosis =healthy) versus negative tuples (eg., diagnosis = sick).True positives refer to the positive tuples that were correctly labeled by the classifier, while true negatives are the negative tuples that were correctly labeled by the classifier. False positives are the negative tuples that were incorrectly labeled by the classifier, while false negatives are the positive tuples that were incorrectly labeled by the classifier. The sensitivity and specificity measures can be used for above purpose and precision is used for the percentage of samples labeled as healthy. The sensitivity(SEN) is given by:

$$SEN = \frac{t - pos}{pos} \quad (7)$$

t-pos is the number of true positives (i.e healthy samples that were correctly classified) and pos is the number of positive (healthy) samples.The specificity(SPE) is given by:

Table II
CONFUSION MATRIX

	Predicted class	
Actual class	C_0	C_1
C_0	$n_{0,0}$	$n_{0,1}$
C_1	$n_{1,0}$	$n_{1,1}$

$$SPE = \frac{t - neg}{neg} \quad (8)$$

t-neg is the number of true negatives (i.e sick samples that were correctly classified) an neg is the number of positive (sick) samples and f-pos is the number of false positives (sick samples that were incorrectly labeled as healty).The Accuracy(ACC) is given by:

$$ACC = SEN\frac{pos}{pos + neg} + SPE\frac{neg}{pos + neg} \quad (9)$$

The true positives, true negatives, false positives and false negatives are also useful in assessing the costs and benefits (or risks and gains) associated with a classification model.In Table 2,the alternatively the accuracy is given by:

$$Accuracy = \frac{n_{0,0} + n_{1,1}}{n} \quad (10)$$

where $n_{0,0}$ is number of C_0 cases classified correctly
$n_{1,1}$ is number of C_1 cases classified correctly

IV. EXPERIMENT AND RESULT.

Experiments were calculated information gain with Weka 3.6.4 tool. We adopt information gain to sort the features according to their importance for classification in the following experiment. To calculate information gain, the input must be discrete numbers. Since the inputs in our experiment are continuous real numbers, we handled the discretization of continuous valued attributes by partitioning the range of values into a finite number of subsets.A features information gain is proportional with its weight to deduction.The ranking information gains of 13 attributes are shown in table 3.We first directly used ANN with no information gain based feature selection function.The result is shown in table 4 is the accuracy and table 5 is the confusion matrix.The accuracy without IG,in training data set is 88.46%, in the validation data set is 80.17%.Then,we deduct feature which has the lowest IG and use ANN to classify.If the classify accuracy is higher or equal than the accuracy without IG.We deduct next feature which has the second lowest IG and use ANN to classify. We made the loop until the classify accuracy is less than the accuracy without IG.Table.4 shown that the left

Table III
INFORMATION GAIN IN EACH FEATURE

Item	Features	Information Gain
1	Thal	0.217395
2	Chest Pain Type	0.204599
3	Number Colored Vessels	0.190442
4	Old Peak	0.167595
5	Maximum Heart Rate	0.151654
6	Induced Angina	0.14221
7	Slope	0.116834
8	Age	0.072551
9	Sex	0.059138
10	Resting ECG	0.024075
11	Blood Pressure	0.2187
12	Cholesterol	0.20316
13	Fasting Blood Suga	0.000566

Table IV
ACCURACY OF ANN CLASSIFIER USED FEATURE SELECTION USING HEART DISEASE DATA SET

Features	Training	Validation
13	88.46%	80.17%
8	89.56%	80.99%

column is the feature number balance used in experiment.We deduct the features until it has 8 features.The accuracy is in training data set is 89.56%, in the validation data set is 80.99%.The accuracy differs between 13 features and 8 features in training data set is 1.1%, in the validation data set is 0.82%.

V. CONCLUSION.

In this paper,Information gain is used to filter features which do not contribute a lot for a given high-level.ANN is used to classify.This research shows that feature selection helps increase computational efficiency while improving classification accuracy.Besides,they decrease the complexity of the system by reducing the dataset. We argue that our proposed improve machine learning models. Simultaneously, it can decrease computational requirement,save repository size,save health checklist costing and reduces the number of attributes which is need to be taken from patients.

Table V
CONFUSION MATRIX OBTAINED FROM ARTIFICIAL NEURAL NETWORKS CLASSIFIER USING 8 FEATURES

	Predicted class	
Actual class	C_0	C_1
Training data set		
C_0	95	9
C_1	10	68
Validation data set		
C_0	54	7
C_1	16	44

REFERENCES

[1] Anamika Gupta, Naveen Kumar, and Vasudha Bhatnagar, "Analysis of Medical Data using Data Mining and Formal Concept Analysis", Proceedings Of World Academy Of Science, Engineering And Technology,Vol. 6, June 2005,.

[2] S Stilou, P D Bamidis, N Maglaveras, C Pappas, Mining association rules from clinical databases: an intelligent diagnostic process in healthcare, Stud Health Technol Inform 84: Pt 2. 1399-1403, 2001.

[3] Hian Chye Koh and Gerald Tan, "Data Mining Applications in Healthcare", Journal of healthcare information management, Vol. 19, No. 2, pp. 64-72, 2005.

[4] Frank Lemke and Johann-Adolf Mueller, "Medical data analysis using self-organizing data mining technologies," Systems Analysis Modelling Simulation, Vol. 43, No. 10, pp: 1399 - 1408, 2003.

[5] Andreeva P., M. Dimitrova and A. Gegov, Information Representation in Cardiological Knowledge Based System, SAER06, pp: 23-25 Sept, 2006.

[6] Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV, LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, May 2007.

[7] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008.

[8] Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. 1 (January - June 2007).

[9] Avci, E. and Turkoglu, I. 2009. An intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases. Expert Syst. Appl. 36, 2 (Mar. 2009), 2873-2878.

[10] van Gerven, M. A., Jurgelenaite, R., Taal, B. G., Heskes, T., and Lucas, P. J. 2007. Predicting carcinoid heart disease with the noisy-threshold classifier. Artif. Intell. Med. 40, 1 (May. 2007), 45-55.

[11] Das, R., Turkoglu, I., and Sengur, A. 2009. Diagnosis of valvular heart disease through neural networks ensembles. Comput. Methods Prog. Biomed. 93, 2 (Feb. 2009), 185-191.

[12] Kurgan, Lukasz A., Cios, Krzysztof J., Tadeusiewicz, Ryszard, Ogiela, Marek, & Doodenday, Lucy S. (2001). Knowledge discovery approach to automated cardiac SPECT diagnosis. Artificial Intelligence in Medicine, 149169.

[13] Polat, K., & GnesB S. (2007). A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS. Computer Methods and Programs in Biomedicine, 88(2), 164174.

[14] M.Anbarasi et. al. (2010). Enhanced Preditction of Heart Disease with Feature Subset Selection using Genetic Algorithm. International Journal of Engineering Science and Technology,Vol.2(10), 5370-5376.