International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 3– No.7, August 2012 – www.ijais.org

Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients

G. Parthiban
Research Scholar,
Dr. MGR Educational Research and Institute,
Maduravoyal,Chennai, India.

S.K.Srivatsa
Phd, Sr. Professor, Dept of E & I,
St.Joseph's College of Engineering,
Chennai,

ABSTRACT
Classifying data is a common task in Machine learning. Data mining plays an essential role for extracting knowledge from large databases from enterprises operational databases. Data mining in health care is an emerging field of high importance for providing prognosis and a deeper understanding of medical data. Most data mining methods depend on a set of features that define the behaviour of the learning algorithm and directly or indirectly influence the complexity of resulting models. Heart disease is the leading cause of death in the world over the past 10 years. Researches have been using several data mining techniques in the diagnosis of heart disease. Diabetes is a chronic disease that occurs when the pancreas does not produce enough insulin, or when the body cannot effectively use the insulin it produces. Most of these systems have successfully employed Machine learning methods such as Naïve Bayes and Support Vector Machines for the classification purpose. Support vector machines are a modern technique in the field of machine learning and have been successfully used in different fields of application. Using diabetics' diagnosis, the system exhibited good accuracy and predicts attributes such as age, sex, blood pressure and blood sugar and the chances of a diabetic patient getting a heart disease.

Keywords
Data Mining, Diabetes, Heart Disease, Machine Learning Methods, Naïve Bayes Method and Support Vector Machines.

1. INTRODUCTION
To extract hidden patterns and relationships from large data bases, Data mining merges statistical analysis, machine learning and database technology.[1]
Diabetes is a chronic disease which causes serious health complications including heart disease, kidney failure and blindness. [4], [5] It is a major risk factor for cardiovascular disease (disease of the heart and circulatory system). Diabetes also increases the risk of micro-vascular damage and macro-vascular complications. Thus diabetes is found to be one of the leading causes of global death by disease. Around 366 million people have diabetes world wide according to statistics taken in the year 2011. Also it has been projected that the people with diabetes will increase to around 552 million by the year 2030.
Heart disease is a term for variety of disease that affecting the heart such as chest pain, shortness of breath, heart attack and other symptoms. It encompasses the diverse diseases that affect the heart. [6] Chest pains arise when the blood received by the heart muscles is inadequate. Heart disease refers to numerous problems that distress the heart and the blood vessels in the heart. The term 'cardiovascular disease' that represents a category of heart disease comprises a broad variety of conditions that upset the heart and the blood vessels and the way in which blood is pumped and circulated in the body [7]. Heart disease is the leading cause of death in the world over the past 10 years. The World Health Organization reported that heart disease is the first leading cause of death in high and low income countries [8]. There are several methods in the literature individually to diagnosis diabetes or heart disease. There is no automated diagnosis method to diagnose Heart disease for diabetic patient based on diabetes diagnosis attributes to our knowledge.
This research paper is related to our previous work, diagnosis of heart disease for diabetic patients using Naïve bayes method [23] and Diagnosing Vulnerability of Diabetic Patients to Heart Diseases using Support Vector Machines [24] to predict the heart disease for diabetic patients using diabetic diagnosis attributes.

2. DATA MINING TECHNOLOGY
Data mining technology is useful for extracting non trivial information from medical databases. It is the intelligent computational analysis of large sets of data by using a combination of machine learning, statistical analysis and database technology, with the objective to discover patterns and rules useful for guiding decisions about future activities [2], [3] The goal of data mining is predicting and generalizing a pattern to other data. Medical data mining is becoming increasingly important in health care. Data mining is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses [9]. Data mining tools predict future trends and behaviours, help organizations to make proactive knowledge-driven decisions [10]. There are various data mining techniques available with their suitability dependent on the domain application. Data mining application in health can have tremendous potential and usefulness. It automates the process of finding predictive information in large databases.
Data mining classification technology consists of two models such as classification model and evaluation model. The classification model makes use of training data set in order to build classification predictive model. Testing data set is used for testing the classification efficiency. Patient dataset is collected from diabetes healthcare institute who have symptoms of heart disease. The classification algorithm like Naive bayes and Support vector machine used for prediction

to find whether diabetic patient is suffering from heart disease with indicating levels.

2.1 NAÏVE BAYES METHOD

Naïve Bayes Classifier is a term dealing with simple probabilistic classifier based on applying Bayes Theorem with strong independence assumptions. It assumes that the presence or absence of particular feature of a class is unrelated to the presence or absence of any other feature **[11].** The Naive Bayes algorithm is based on conditional probabilities. It uses Bayes' theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data.

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. If `B` represents the dependent event and `A` represents the prior event, Bayes' theorem can be stated as follows.

Prob (B given A) = Prob(A and B)/Prob(A)

To calculate the probability of `B` given `A`, the algorithm counts the number of cases where `A` and `B` occur together and divides it by the number of cases where `A` occurs alone.

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. Since independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire. It can be used for both binary and multi class classification problems.

The diabetes attributes used in our proposed system and their descriptions are shown in Table 1.

Table 1 Diabetes attributes used in the experimentation

Attribute	Description
Sex	A classification of the sex of the person
Age	Age of the patient
Family Heredity	Previous history (Father / Mother)
Weight	Patient's weight
BP	Blood pressure
Fasting	Sugar level after fasting
PP	Post Prandial blood glucose level
A1C	HbA1c level Glycosylated Last 4 months sugar level
LP Tot Cholesterol	Total cholesterol level

Weka is a collection of machine learning algorithms for data mining tasks, written in Java and it contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. **[12]** The key features of Weka are it is open source and platform independent. It provides many different algorithms for data mining and machine learning **[13].**

We have used Naïve bayes method to perform the mining and classification process. We have used 10 folds cross validation to minimize any bias in the process and improve the efficiency of the process.

The results of our experimentation are shown in Figure1.

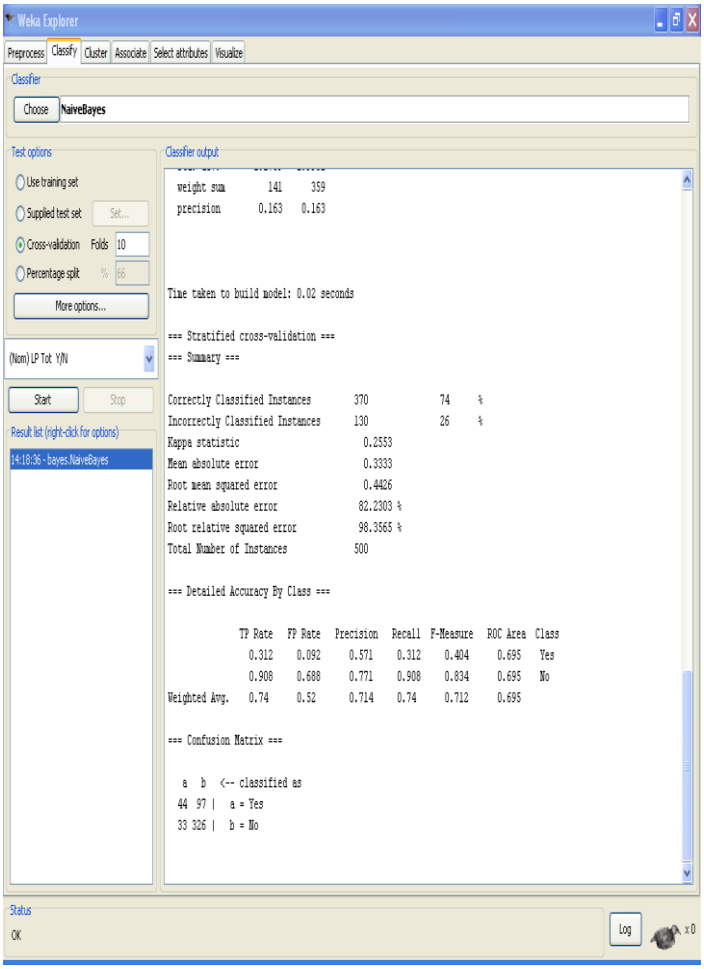

Figure 1 Result window of the data mining process

The proposed naïve bayes model was able to classify 74% of the input instances correctly. It exhibited a precision of 71% in average, recall of 74% in average, and F-measure of 71.2% in average. The results show clearly that the proposed method performs well compared to other similar methods in the literature, taking into the fact that the attributes taken for analysis are not direct indicators of heart disease.

2.2 SUPPORT VECTOR MACHINES (SVM)

A Support Vector Machine (SVM) is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize pattern introduced by Corinna Cortes and Vladimir Vapnik used for classification and regression analysis. SVM have shown good performance in a number of application areas. It constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. [14] SVM's are very much useful in data classification. SVM's classify data by finding an optimal hyper plane separating the d – dimensional data into its two classes with a maximum interclass margin. SVM's use so called kernel functions to cast data into a higher dimensional space where the data is separable. [15], [16]

Classifying data is a common task in machine learning. Suppose in some given data points each belong to one of two classes and the goal is to decide which class a new data point will be in. In the case of support vector machines, a data point is viewed as a p-dimensional vector (a list of p numbers), and we want to know whether we can separate such points with a (p − 1)-dimensional hyper plane. This is called a linear classifier [17].

SVM is a learning machine that plots the training vectors in high dimensional space and labels each vector by its class. [18] SVM based on the principle of risk minimization which aims to, minimize the error rate. [19], [20] SVM uses a supervised learning approach for classifying data. That is, SVM produces a model based on a given training data which is then used for predicting the target values of the test data. Given a labelled training set (x_i, y_i), SVM require the solution of the following optimization problem to perform classification [22].

$$\min_{w,b,\varepsilon} \frac{1}{2} W^T W + C \sum_{i=1}^{l} \varepsilon_i \qquad 1$$

Subject to,

$$y_i(W^T \phi(x_i) + b) \geq 1 - \varepsilon_i \qquad 2$$

where,

$\varepsilon_i \geq 0$, a slack variable to allow for errors in the classification

x_i – training vectors, $x_i \in R^n$

ϕ - function mapping x_i into a higher dimension space,

C – penalty parameter of the error term (usually C>0),

y_i – Class label, $y_i \in \{1,-1\}^l$

3. BASIC METHODOLOGY

The methodology described in this paper is diagnosing vulnerability of diabetic patients to heart diseases and we had collected 500 records of diabetic patients to perform the experimentation. The attributes making up each record is shown in Table 2.

Table 2. Attributes used for the diagnosis

Attribute Role	Attribute Name	Attribute Type	Description
Regular	Sex	binominal	Sex of the patient. Takes the following values: Male, Female
Regular	Age	integer	Age of the patient
Regular	Fam/Heri	polynomial	Indicates whether the patient's parents were affected by diabetes. Takes the following values: Father, Mother, Both
Regular	Weight	numeric	Weight of the patient
Regular	BP	polynomial	Blood Pressure of the patient
Regular	Fasting	integer	Fasting Blood Sugar
Regular	PP	integer	Post Prondial Blood Glucose
Regular	A1C	numeric	Glycosylated Hemoglobin Test
Regular	LDL	integer	Low Density Lipoprotein
Regular	VLDL	integer	Very Low Density Lipoprotein
Label	Vulnerability	nominal	Indicates the vulnerability of the patients to heart disease. Takes the following values: High, Low

Out of the 500 records, 142 records were pertaining to patients highly vulnerable to heart diseases. The remaining 358 records were pertaining to patients less vulnerable to heart disease. Since SVM processes only numeric attributes, the nominal were converted to numeric attributes by replacing each value by a unique integer. For example, the attribute Sex values are converted as follows: Male – 1 and Female – 0.

The values of the attributes were then normalized to the range 0 to 1. These records were then given as input to the SVM classifier.

SVM uses kernel functions to map the data set to a high dimensional data space for performing classification. The different types of kernel functions are as follows [22]:

Linear: $K(x_i, x_j) = x_i^T x_j$ (3)

Polynomial:

$$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0 \qquad (4)$$

Radial Basis Function:

$$K(x_i, x_j) = \exp\left(\gamma \| x_i - x_j \|^2\right), \gamma > 0 \quad (5)$$

Sigmoid:

$$K(x_i, x_j) = \tanh\left(\gamma x_i^T x_j + r\right) \quad (6)$$

where γ, r, d are kernel parameters. The choice of the kernel depends on whether the relationship between the class labels and attributes are linear or nonlinear. For nonlinear relationships, the radial basis function (RBF) kernel has been found to be a good choice as it has lesser number of hyper parameters than other nonlinear kernels. Also RBF kernel has fewer numerical difficulties [22]. Hence we have used RBF kernel in our SVM classifier.

4. RESULT ANALYSIS

The data set used for training the classifier comprises of 500 diabetic patient records out of which 142 records are of those having heart disease (positive cases) and the remaining 358 records are of those not having heart disease (negative cases). These records after sufficient pre-processing was given as input to train the SVM classifier.

The SVM classifier was trained for different values of the RBF kernel parameters, C and γ. The models thus obtained for each of the values of C and γ where then tested for accuracy. A good classifier should be able to exhibit high accuracy for datasets unseen rather than the training data. Hence we have used 10 fold cross validation for testing the accuracy of the classifier.

In 10-fold cross-validation, we first divide the training set into 10 subsets of equal size. Sequentially one subset is tested using the classifier trained on the remaining 9 subsets. Thus, each instance of the whole training set is predicted once so the cross-validation accuracy is the percentage of data which are correctly classified. The cross validation tests prevents over fitting problem. Based on the exhaustive trials conducted, we found that for C = 5.0 and γ = 1.0 the classifier exhibited the best accuracy of 94.60%. The accuracy obtained for a few values of C and γ in our trials is shown in the Table 3.

Table 3. Partial results of the trials conducted

C Value	γ Value	Accuracy of the
.	.	.
2	0.125	89.60%
2	0.75	92.40%
4	2.5	93.20%
4	2	93.60%
4	1.5	93.80%
4	1	94.20%
5	**1**	**94.60%**
6	1.25	94%
.	.	.

The ROC curve for the classifier characteristics is shown in Fig. 2

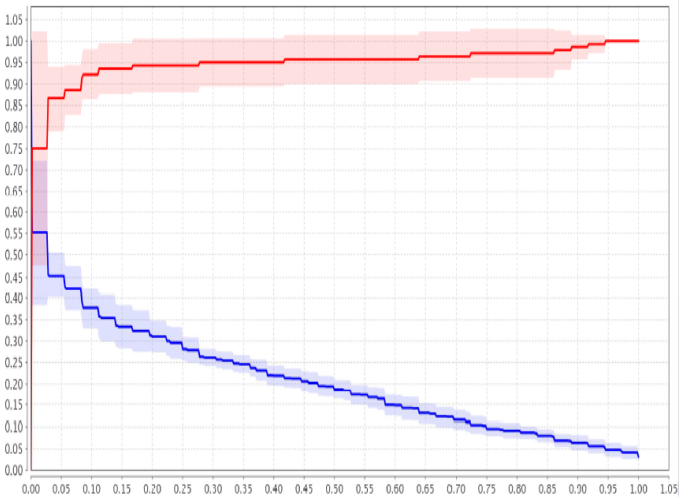

Fig 2: ROC curve for the classifier characteristics

R.O.C chart [25] is a useful visual tool for comparing classification methods. It shows the trade-off between the true positive rate and the false positive rate for a given model. ROC chart is based on the conditional probabilities sensitivity and specificity [26].

The confusion matrix indicating the accuracy of the SVM classifier for the given data set is shown in Table 4. The confusion matrix is a Visualization tool used in supervised learning which contains actual and predicated classification. Each column represents instance in a predicated class and each row represents instance in an actual class.

Table 4. The confusion matrix of the classifier

	True low	**True high**	**Class**
pred. low	355	24	93.67%
pred. high	3	118	97.52%
class recall	99.16%	83.10%	
Overall accuracy: 94.60% +/- 2.01% (mikro: 94.60%)			

From the results obtained, it can be seen that the classifier exhibits a very high classification accuracy i.e 94.60% overall. It also shows a very high precision for the positive class (97.52%) and also the recall of the positive class is quite good (83.10%). In the case of negative classes, the classifier exhibits high precision (93.67%) as well as high recall (99.10%).

5. CONCLUSIONS

Application of Data mining in analysing the medical data is a good method for considering the existing relationships between variables. From our proposed approach we have shown that mining helps to retrieve useful correlation even from attributes which are not direct indicators of the class we are trying to predict.

In our work we have tried to predict the chances of getting a heart disease using attributes from diabetic's diagnosis and we have shown that it is possible to diagnose heart disease vulnerability in diabetic patients with reasonable accuracy. Classifiers of this kind can help in early detection of the vulnerability of a diabetic patient to heart disease. There by the patients can be forewarned to change their lifestyle. This will result in preventing diabetic patients from being affected by heart disease, there by resulting in low mortality rates as well as reduced cost on health for the state. SVM's have proven to be a classification technique with excellent predictive performance and also been investigated with the help of ROC curve for both training and testing data. Hence this SVM model can be recommended for the classification of the diabetic dataset.

6. ACKNOWLEDMENTS

We are grateful to Dr.V.Shesiah, Chairman and Managing director of Dr.V.Shesiah Diabetic Research Institute, Chennai for providing an access to medical diabetic data and for his involvement in this domain.

7. REFERENCES

[1] Thuraisingham, B.: "A Primer for Understanding and Applying Data Mining", IT Professional, pp: 28-31, 2000

[2] J. Han Kamber, M. 2006. Data Mining: Concepts and Techniques, 2nd ed. San Francisco: Morgan Kaufman.

[3] U. Fayyad, G.Piatetsky-Shapiro, and P.Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, Vol.17, pp.37-54, 1996.

[4] World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: Available: http://www.who. int/ diabetes/en

[5] World Health Organization. Available: http: // www.who.int/topics/ diabetes mellitus/en/

[6] K.Srinivas et al. / "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", International Jounal on Computer Science and Engineering (IJCSE) Vol..02, No.02, 2010, 250-255. Available:http://www.enggjournals.com/ijcse/doc /IJCSE10-02-02-25.pdf

[7] "Hospitalization for Heart Attack, Stroke, or Congestive Heart Failure among Persons with Diabetes", Special report: 2001 – 2003, New Mexico.

[8] World Health Organization. (July 2007-Febuary 2011). [Online]. Available:http://www.who.int/mediacentre/factsheets/fs310.pdf.

[9] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation 10500 Falls Road, Potomac, MD 20854 (U.S.A.),1999.

[10] L.A.Rose, D.T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN O-471-66657-2, ohn Wiley & Sons, Inc, 2005.

[11] Naïve bayes classifier based on applying bayes theorem: http://en.wikipedia.org/wiki/Naive bayes classifier

[12] Weka Data mining software http://www.cs.waikato .ac.nz/ml/weka

[13] An Introduction to the WEKA Data mining system – http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf

[14] Introduction to Support Vector Machine Available: http://en. wikipedia. org/wiki/Support_vector_machine

[15] Colin Campbell and Yiming Ying, Learning with Support Vector Machines, 2011, Morgan and Claypool. Available: http://www. morganclaypool.com/doi/abs/10.2200/S00324ED1V01Y201102 AIM010?journalCode=aim

[16] H.Barakat, Andrew P.Bradley and Mohammed Nabil H.Barakat (2009) "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus", IEEE Transactions on Information Technology in Bio Medicine, Volume 14, Issue 4, pp 1-7, 2009. Available: http://ieeexplore. ieee.org /xpls/ abs_ all.jsp?arnumber=5378519 Digital Object Identifier: 10.1109 /TITB. 2009.2039485

[17] Mertik M., Kokol P., Zalar B.Gaining Features in Medicine Using Various Data-Mining Techniques //Computational Cybernetics ICCC 2005, IEEE rdInternational Conference. – 2005. – P. 21–24.

[18] G.Suganya, D.Dhivya "Extracting Diagnostic rules from SVM" , Journal of Computer Applications (JCA), 2011.

[19] N. Barakat and A.P.Bradley, "Rule Extraction from Support Vector Machines: A Sequential Covering Approach " IEEE Transactions on Knowledge and Data Engineering, Volume 19,no.6,pp 729-741, 2007. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04161896 Digital Object Identifier no. 10.1109/TKDE.2007.1023.

[20]S.Balakrishnan, R.Narayanaswamy, N.Savarimuthu, R.Samikannu "SVM Ranking with Backward Search for Feature Selection in Type II Diabetes Databases" 2008 IEEE International Conference on Systems, Man and Cybernetics. Available: http://ieeexplore. ieee.org /xpls /abs_all.jsp?arnumber=4811692 Digital Object Identifier: 10.1109 /IC SMC.2008.4811692

[21] K. E. Heikes, B. Arondekar, D. M. Eddy, and L. Schlessinger, "Diabetes Risk Calculator,A simple tool for detecting undiagnosed diabetes and pre-diabetes," Diabetes Care, vol. 31, no. 5, pp. 1040-1045, 2008

[22] W. Kong, L. Tham, K. Y. Wong, and P.Tan, "Support vector machine approach for cancer detection using amplified fragment length polymorphism (AFLP) method," Proc. the 2nd Asia-Pacific Bioinformatics Conference (APBC2004), Dunedin, New Zealand, 2004.

[23] G.Parthiban, A.Rajesh, S.K.Srivatsa, "Diagnosis of Heart Disease for Diabetic Patients using Naïve Bayes Method", International Journal of Computer Applications (IJCA) Volume 24-No.3, June 2011, 0975-8887.Available:

http://www.ijcaonline.org/archives/volume24/number3/2933-3887 doi 10.5120/2933-3887

[24]G.Parthiban, A.Rajesh, S.K.Srivatsa, "Diagnosing Vulnerability of Diabetic Patients to Heart Diseases using Support Vector Machines", International Journal of Computer Applications (IJCA) Volume 48-No.2, June2012,0975-888.Available:

http://www.ijcaonline.org/archives/volume48/number2/7324-0149 doi 10.5120/7324-0149

[25] SPSS Clementine help file. http//www.spss.com

[26] Kelly H. Zou, PhD; A. James O"Malley, PhD; Laura Mauri, MD, M.Sc "ROC Analysis for Evaluating Diagnostic Test and Predictive Models.