IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Heart Disease Prediction System Using SVM and Naive Bayes

Ms. R.R.Ade[*1], Dhanashree S. Medhekar[2], Mayur P. Bote[3]
G. H. Raisoni Institute of Engineering and Technology, University of Pune, India
rosh513@gmail.com

Abstract

As large amount of data is generated in medical organisations (hospitals,medical centers)but as this data is not properly used. There is a wealth of hidden information present in the datasets. This unused data can be converted into useful data. For this purpose we can use different data mining techniques. This paper presents a classifier approach for detection of heart disease and shows how support vector machine(SVM) and Naive Bayes can be used for classification purpose. In our system, we will categories medical data into five categories namely no, low, average,high and very high. Also, if unknown sample comes then the system will predict the class label of that sample. Hence two basic functions namely classification (training) and prediction (testing) will be performed. Accuracy of the system is depends on algorithm and database used.

Keywords: data mining, Heart disease, Naive Bayes, SVM

Introduction

Data Mining

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Basic terms related to data mining

Classification

Classification is a data mining (machine learning) technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be “sunny”, “rainy” or “cloudy”.

Supervised learning

Supervised learning is the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier. The inferred function should predict the correct output value for any valid input object. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

Unsupervised learning

In machine learning, unsupervised learning refers to the problem of trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution. This distinguishes unsupervised learning from supervised learning .

Prediction

Models continuous-valued functions, that is predicts unknown or missing values.

Data Source

Description of Cleveland Dataset

This dataset contains information concerning heart disease diagnosis . The data was collected from the cleveland clinic foundation, and it is available at UCI Repository . Six instances containing missing values have been deleted from the original dataset .

Format :

A data frame with 303 observations on the following 14 parameters –

P1 - Age
P2 - Gender
P3 – CP (chest pain)
P4 - trestbps : resting blood pressure
P5 – cholesterol

P6 – fbs: fasting blood sugar>120 ? yes=1,no = 0
P7 – restecg: resting electrocardiographic results 0,1,2
P8 – thalach : maximum heart rate achieved
P9 – exang : exercise induced angina (1= yes ; 0= no)
P10 – oldpeak = ST depression induced by exercise relative to rest
P11 – slope : the slope of the peak exercise ST segment
P12 – ca: no. of major vessels (0 to 3) colored flurosopy
P13 – thal :3 =normal ,6=fixed defect ,7= reversable defect
P14 – diagnosis of heart disease

Related Work

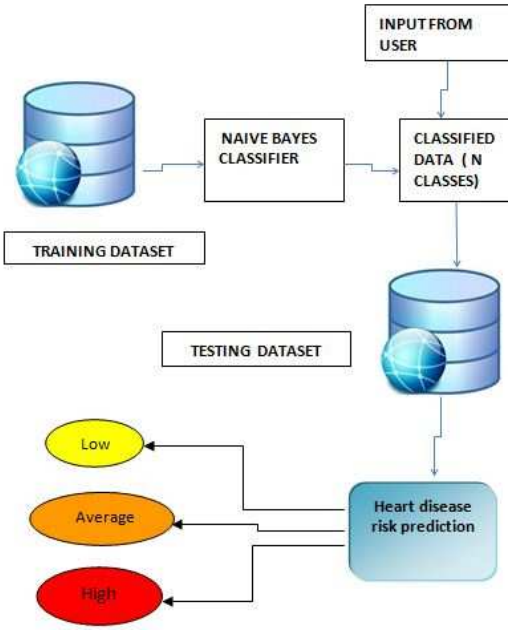


Fig.1 System Archtecture

As shown in Fig 2.1,the training dataset is given as input to the classifier.This classified data is further used for testing purpose.
We have used algorithm Naive Bayes.
Mainly system will work in two phases:
1)Training phase
2)Testing phase
Training Phase:
Classification assumes labeled data: we know how many classes there are and we have examples for each class (labeled data).

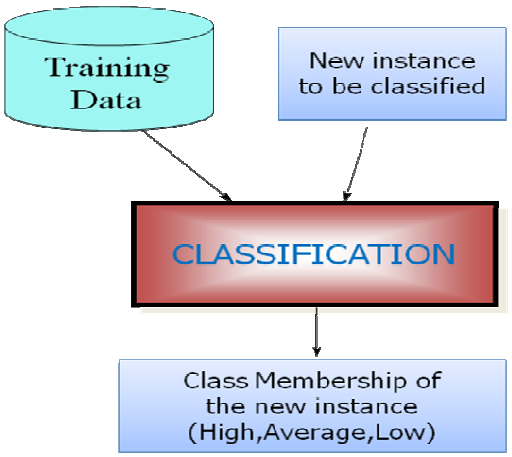
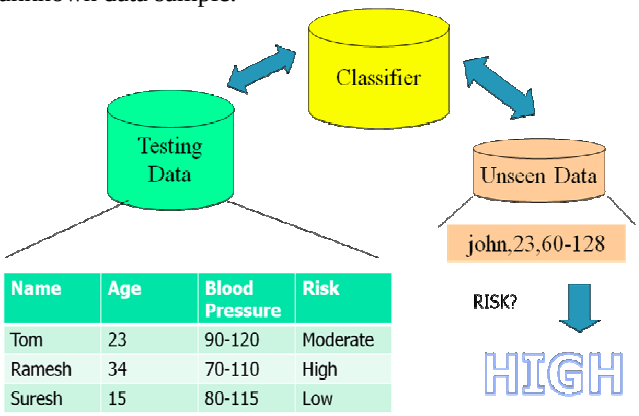


Fig.2 Classification

Classification is supervised
Classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
Testing Phase:
Testing phase involves the prediction of unknown data sample.



Name	Age	Blood Pressure	Risk
Tom	23	90-120	Moderate
Ramesh	34	70-110	High
Suresh	15	80-115	Low

Fig.3 Prediction

Models continuous-valued functions, i.e., predicts
unknown or missing values.
In testing we check those data that does not come under the dataset we have considered.
After the prediction, we will get the class labels

Techniques Used

Naive Bayes:
The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems.

Naive Bayes algorithm is based on Bayesian Theorem

Bayesian Theorem:

Given training data X, *posterior probability of a hypothesis H, P(H|X),* follows the Bayes theorem

P(H|X)=P(X|H)P(H)/P(X) (1.1)

Algorithm:

The Naive Bayes algorithm is based on Bayesian theorem as given by equation(1.1)

Steps in algorithm are as follows:

1. Each data sample is represented by an n dimensional feature vector, *X = (x1, x2….. xn),* depicting n measurements made on the sample from n attributes, respectively *A1, A2, An.*

2. Suppose that there are m classes, *C1, C2……Cm.* Given an unknown data sample, *X* (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability, conditioned if and only if:

P(Ci/X)>P(Cj/X) for all *1< = j< = m* and *j!= i*

Thus we maximize *P(Ci|X).* The class *Ci* for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes theorem,

3. As *P(X)* is constant for all classes, only P(X|Ci)P(Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, i.e. *P(C1) = P(C2) = …..= P(Cm),* and we would therefore maximize *P(X|Ci).* Otherwise, we maximize *P(X|Ci)P(Ci).* Note that the class prior probabilities may be estimated by *P(Ci) = si/s* , where Si is the number of training samples of class *Ci,* and s is the total number of training samples. on *X.* That is, the naive probability assigns an unknown sample *X* to the class *Ci* [2]

SVM

Support Vector machines(SVM) have gained popularity in the machine learning and pattern classification.

The aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data. For a linearly separable dataset, a linear classification function corresponds to a separating hyperplane *f (x).* SVM guarantees that the best such function is found by maximizing the margin between the two classes.

Thus,It is a linear classifier which constructs a separating hyperplane to maximize distance between data.

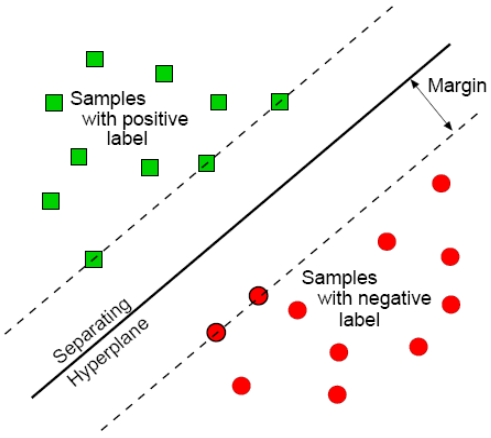


Fig:4Seperating the dataset with maximal margin (Optimal seperating hyperplane)

Classification is achieved by a cation,linear or non-linear separation surface in input space.

In this classification the separating function is expressed by linear combination of kernels associated with Support Vectors by

$f(x) = \sum \alpha_j y_j K(x_j,x) + b$

x_i where trainig patterns, {+1,-1} is class labels and S is the set of support vectors.

The advantage of the SVM algorithm it can be geometrically represented and simple to understand.

If S be the current candidate Support Vector set,and n the size of the dataset,the algorithms spend(n-s)|S| kernel evaluations to locate the maximum violator.

The dual formulation yields

min

$0 \leq \alpha_i \leq C$ $W = ½ \sum \alpha_i Q_{ij} \alpha_j - \sum \alpha_i + b \sum y_i \alpha_i$

Where α_i are corresponding coefficients b is the offset,

$Q_{ij} = y_i y_j K(x_i,x_j)$ is the symmetric positive definite kernel matrix and C is the parameter used to penalize error points in Support Vectors

.

SVM find out a linear separating hyperplane with maximal margin in higher dimensional space.Therefore kernel function is,

$K(x_i x_j) \equiv \Phi(x_i)^T \Phi(x_j)$

Researchers proposed new kernel functions but basic four kernels are:

i)linear: $K(x_i x_j)=x_i^T x_j$.

ii)polynomial: $K(x_i x_j)=(\gamma x_i^T x j_+ r)^d$ $\gamma>0$.

iii)radial basis function(RBF):$K(x_i x_j)=\exp(-\gamma||x_i - x_j||^2)$,$\gamma>0$.

iv)sigmoid: $K(x_i x_j)=\tanh(\gamma x_i^T - x_{j+} r)$.

Where γ,r and d are kernel parameters.

To implement SVM we have used LIBSVM package.

Karush-Kuhn-Tucker (KKT) Conditions in SVM

KKT conditions (a.k.a. Kuhn-Tucker conditions) are necessary conditions for the local minimum solutions
of problem (1).Let x! be a local minimum point for Problem (1) and suppose $x^*$is a regular point for the constraints. Then there is a vector $\mu \in R^m$ and a vector $\lambda \in R^p$ with $\lambda \geq 0$ such that

$$\nabla f(x^*) + \lambda \nabla^T h(x^*) + \mu^T \nabla g(x^*) = 0$$
$$g(x^*) = 0$$
$$\lambda_j h_j(x^*) = 0 \quad (j = 1,\ldots,p)$$

Convince yourself why the above conditions hold geometrically. It is convenient to introduce the Lagrangian
associated with the problem as
$L(x,\lambda,\mu) = f(x) + \lambda^T h(x) + \mu^T g(x)$
Where $\mu \in R^m$,$\lambda \in R^p$ and $\lambda \geq 0$ are Lagrange multipliers. Note that equation (2), (3) and (4) together
give a total of n + m+ p equations in the n + m + p variables x^*,λ and μ
From now on we assume that we only have inequality constrains for simplicity. The case with equality
constraints can be done in a similar way, except that μ does not have the nonnegative constraint as λ. So in our case we have the following optimization problem:
$\min f(x)$ s.t.$h(x) \leq 0$.

LIBSVM

LIBSVM is a library for support vector machines.
How to use Libsvm:
We must follow the procedure:
1.Data Preparation for SVM
2.Connvert data into SVM format.
3.Conduct simple scaling on the data.
4.Consider the RBF kernel K(x; y)=e
5.Use the best parameter C and gamma to train the whole training set.
6.Test

Advantages of SVM

1)The quality of generalization and ease of training of SVM is far beyond the capacities of these more traditional methods.
2)SVM can model complex, real-world problems
3)SVM performs well on data sets that have many attributes, even if there are very few cases on which to train the model Traditional neural networks do not perform well under these circumstances.

Results and Analysis

Results and analysis is done on Cleveland dataset.Results are shown in the form of pie charts,bar charts. Table 1 shows the accuracy obtained by changing the number of instances in the training dataset.

Results by Naive Bayes Classifier

Table.1 Accuracy(%)

Number Of records in Traning dataset	Number of records in Testing dataset	Number of Correctly classified instances	Number of Incorrectly classified instances	Accuracy (%)
303	276	245	31	88.76
303	240	215	25	89.58
303	290	258	32	88.96

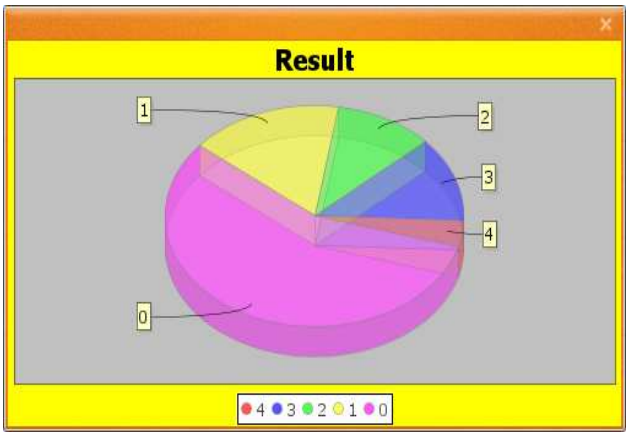


Fig.5 Classified data

Fig.5 shows the classified data in the form of Pie chart.0,1,2,3,4 represents the posssibility of heart disease.
0:No
1:Low
2:Average
3:High
4:Very high

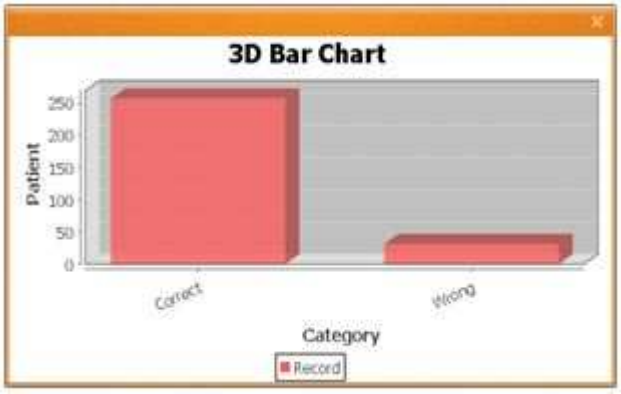


Fig.6 Correctly and wrongly classified data

Fig.6 shows the correctly and wrongly classified records in the form of bar chart.

Results by SVM

Analysis done by using SVM on Cleveland dataset is shown below.

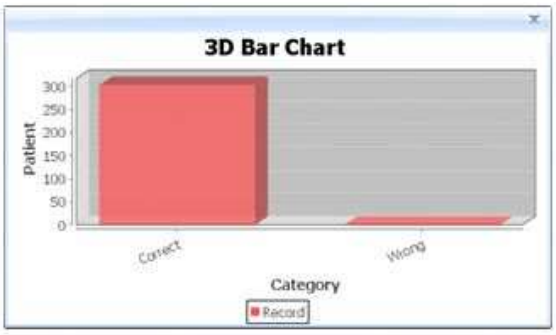


Fig.7 Correctly and wrongly classified data(SVM)

Fig.7 shows the correctly and wrongly classified records in the form of bar chart when we train the dataset using SVM.Here,we have achieved 100% accuracy.Zero records are wrongly classified.

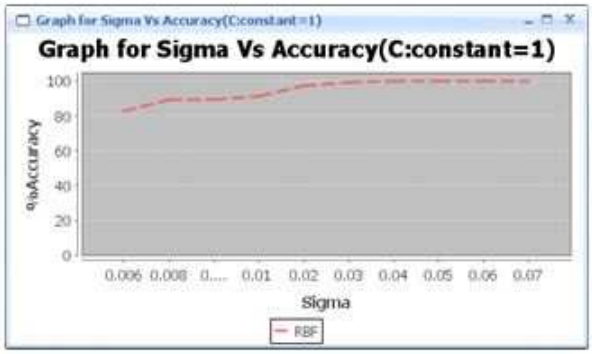


Fig.8 Graph for Sigma Vs Accuracy

Fig.8 shows the graph plotted between sigma and accuracy obtained. X-axis shows the values of parameter sigma and Y-axis shows the %accuracy.From fig.7 it is clear that as we are increasing the value of parameter sigma % accuracy also increases.After a certain value of sigma that is after 0.04 the %accuracy remains constant.We got 100% accuracy for the sigma value 0.04,0.05,0.06 and so on.

Conclusions and Future Work

This system classifies the given data into differerent categories and also predicts the risk of the heart disease if unknown sample is given as an input.The system can be served as training tool for medical students.Also,it will be helping hand for doctors.As we have developed generalised system,in future we can use this system for analysis of different datasets by only changing the name of dataset file which is given for training module.

References

[1] Mai Shouman, Tim Turner, Rob Stocker "USING DATA MINING TECHNIQUES IN HEART DISEASE DIAGNOSIS AND TREATMENT"Japan-Egypt Conference on Electronics, Communications and Computers 978-1-4673-0483-2 c_2012 IEEE

[2] N. Aaditya Sunder, P. PushpaLatha, "Performance analysis of classification data mining techniques over heart disease database" Inernational Journal Of Engineering Science and Advance Technology"-vol-2 issue-3,470-478,May-June 2012

[3] Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006

[4] IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August, 2008

[5] SellappanPalaniappan, RafiahAwang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, ©2008 IEEE.

[6] ShantakumarB.Patil, Y.S.Kumaraswamy, Intelligent andEffective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656 © EuroJournals Publishing, Inc. 2009.

[7] R. Bhuvaneswari and K. Kalaiselvi, Naive Bayesian Classification Approach in Healthcare Applications International Journal of Computer Science and Telecommunications [Volume 3, Issue 1, January 2012]

[8] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni," Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011

[9] Data mining concepts and techniques, second edition,Han Kamber.
[10] Masayuki Karasuyama, Student Member, IEEE, and Ichiro Takeuchi, Member, IEEE," Multiple Incremental Decremental Learning of Support Vector Machines"