ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Probabilistic Classification for Prediction of Heart Disease

[1]John Peter and [2]Somasundaram

[1]*VTU, SAIT, Dept. of ISE, India.*
[2]*Anna University JEC, Dept. of CSE, India*

ARTICLE INFO

Article history:
Received 17 January 2015
Received in revised form 26 28 February 2015
Accepted 17 March 2015
Available online 14 April 2015

Keywords:
Data Mining, Classification, Heart Disease, Probabilistic Classification

ABSTRACT

In this IT mechanized world large amount of Data flow from various resources. Maintaining these Big Data becomes a challenge. Segregating the data into various concerned repositories and make it available for future use is a hectic task. Data stored in repositories when retrieved for research purpose for prediction without preprocessing or data cleansing may give erroneous result. Proposed is the Probabilistic classification algorithm for prediction of heart disease with the preprocessed Big data available. Map Reduce algorithm using Hadoop is used to handle big data. This application is made globally accessible by deploying it on cloud and can be accessed through any browser. It is user friendly, parallel processing is performed on the big data for faster computation.

To Cite This Article: John Peter and Somasundaram., Probabilistic Classification for Prediction of Heart Disease. ***Aust. J. Basic & Appl. Sci.,*** *9(7): 639-643, 2015*

INTRODUCTION

In present world situation there are many scientific technologies which help doctors in taking clinical decisions which might not be accurate. Heart disease prediction system can assist medical professionals in predicting status of heart disease, based on the clinical data of patients. Doctors may sometimes fail to take a correct decision in predicting heart disease risk level, therefore heart disease prediction systems are useful in such cases to get accurate results. There are many tools available which perform this task but all of them have some flaws. Most of the tools cannot handle big data and hence predicting heart disease would be a tedious task. There are many hospitals and healthcare industries which collects huge amounts of patient data which becomes difficult to handle with currently existing systems. In this paper we are predicting the risk levels of patients from a huge data set.

Data mining techniques are used for preprocessing, machine learning algorithms are used for implementation. Popular machine learning algorithms have been implemented to determine the heart disease risk level and to help the doctors correctly predict the same. Finally a comparison between the algorithms is done which helps the user to determine which algorithm shows the highest accuracy.

This paper is divided into five phases first phase gives the theoretical background for reducing attributes from a data set, second phase gives the implementation of machine learning algorithms for predicting heart disease risk level, third phase deals with processing of Big Data using Hadoop Map Reduce programming, and fourth phase gives conclusion and future scope of the project.

Reduction of Attributes Using Data Mining Techniques:

Data Mining is a process of extracting useful and important knowledge from huge data set. Data Preprocessing is a important process in Data Mining and Machine learning. Dimensionality reduction is an effective method for downsizing data. The important techniques for Dimensionality reduction are Feature Selection and Feature Extraction.

Feature selection is the process of selecting a subset of relevant features. Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature Subset selection is one of the method for Feature selection. In subset selection we find the best subset of the set of features. The best subset contains the least number of dimensions that most contribute to accuracy. We discard the remaining, unimportant dimensions.

Corresponding Author: John Peter, VTU, SAIT ISE, Bangalore-97, India.
Ph: +91 7829759155, E-mail: tjpeter.cse@gmail.com

There are two approaches in subset selection
a. Forward Selection
b. Backward Selection

Forward Selection starts with no variables and we add them one by one, at each step adding the one that decreases the error the most, until any further addition does not decrease the error.

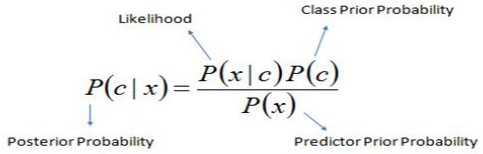
In backward selection, we start with all variables and remove them one by one, at each step removing the one that decreases the error the most, until any further removal increases the error significantly. Let us denote by F, a feature set of input dimensions, xi, i = 1,...,d. E(F) denotes the error incurred on the validation sample when only the inputs in F are used. In sequential backward selection, we start with F containing all features and we remove one attribute at a time from F, and we remove the one that causes the least error.
$j = \text{argmin}_i E(F - x_i)$ and we remove x_j from F if $E(F - x_j) < E(F)$.We stop if further removal does not decrease the error. On theoretical basis Backward Selection is more favorable in our case, since we are considering only 13 out of 76 heart parameters.

The dataset under consideration has been taken from University of California Irvin (UCI). Thirteen attributes are involved in prediction of heart Disease. These thirteen attributes have been shown in Table 1. This data set is fed into the classification model i.e. Naïve Bayes Classification and Probabilistic Analysis and Classification. The Big data file containing patients record is given as the input and the result is processed.

Machine Learning Algorithms:

A. Naive Bayes Classifier:

Naïve Bayes classifier is used as the first method for prediction of heart disease. Here we use the preprocessed 13 attributes as input. With Naïve Bayes' assumption all attributes are independent of each other, this significantly reduces the calculations shown later. The Naïve Bayes formula is given by



p(c|x) is the posterior probability of class (target) given predictor (attribute).
P(c) is the prior probability of class, also called prior. It is the probability of observing a class in general.
P (x|c) is the likelihood which is the probability of predictor given class.
P(x) is the prior probability of predictor also called evidence.

The term evidence is constant for all the class values hence the posterior probability of a class is proportional to product of likelihood and prior value for that class only. Here x is a tuple of 13 reduced attributes and can be expressed as (x1, x2, x3…, x13).Using Naïve Bayes' assumption likelihood can be split into continuous product of class conditional probabilities of 13 attributes as shown below.

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Hence given as input a patient record of 13 attributes we can calculate posterior probability for all risk levels. Patient has the risk level for which the posterior probability is maximum. Training data set is used for calculation of class conditional probabilites. Given an attribute xi we can calculate P(xi|Cj) for class Cj. For this we can use basic definition of probability that is.

$$P(x_i|Cj) = \frac{\text{Number of times } x_i \text{ occurs in rows of training data set } X^t \text{ for class Cj}}{\text{Number of times Cj occurs in training data set } X^t}$$

xi € X and j=0,1,2,3,4. Hence for the calculation of likelihood entire training dataset is used. However this method of calculation holds good if and only if variables are discrete in nature like sex, chest pain type etc. for patient record. In this dataset exactly 5 attributes i.e., age, cholesterol, resting blood pressure, thalach and oldpeak are continuous. Hence the initial approach is calculation of class conditional densities using probability density function assuming Normal distribution for all the continuous variables as shown

$$p(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

Here σ_c^2 is the variance for variable x given class C. µc is the mean for variable x given class C.

Using normal distribution for age, cholesterol and thalach is an approximately good assumption. However resting blood pressure and old peak do not fit into this distribution and these results in over fitting of model. As a result, we get partially accurate results leading to low accuracy. To avoid dealing with distribution of variables, we can use another approach, i.e. assuming these variables to be discrete. In this case, calculation of class conditional probabilities for these variables is done in the same way as done for other discrete variables. This assumption holds good in this case since the dataset is large and also leads to improved results and high accuracy.

B. Probabilistic Classification and Analysis:

Probabilistic Analysis and Classification is a supervised machine learning algorithm. It uses the concept of weighted average probability calculation over the entire training data set {Xt} It is built over Naïve Bayes model to overcome the shortcomings of Naïve Bayes algorithm. One advantage is complete reduction of continuous variables to discrete variables using discretization technique, hence the hard work to find suitable distribution for continuous variables is not required which else was resulting in "over fitting the model". Another

advantage is due to complete conversion of continuous variable to discrete variable, Laplacian smoothing used in Naïve Bayes Classification is not required, which in turn reduces unnecessary comparisons and instructions.

The main idea of algorithm is to use weighted average calculation for all heart disease attributes until unless we find an exact same tuple in the training data set, in this case the risk level of tuple is assigned the risk level of the input patient record. However this doesn't happen often and so we have to use weighted average calculation for the entire training data set and calculate the contribution of each and every attribute for that particular risk level and what the different contributions for entire training data set are. For considering the entire data set we have used number of supporting tuples for various risk levels in the training data set. This is similar to concept of "Prior" in Naïve Bayes algorithm but in Naïve Bayes algorithm where prior probabilities give more weight to risk levels on the basis of their own values. In PAC it substantially reduces this weight, resulting in error, due to difference in percentage increase in numerator and denominator in the term αi. So to overcome this drawback, we multiply by normalizing factor to reduce this error and give appropriate results. Finally the maximum term μi among all risk levels is returned as the risk level for the patient.

Other variances from Naïve Bayes implementation are feeding the training data and Big Data files, which have to be parsed in pre-initial step to convert continuous variables to discrete variables. Another step is conversion of continuous variables supplied by user to discrete form so that algorithm can read and process them.

Hadoop Map Reduce Programming for Processing Big Data:

In this paper we have successfully designed an algorithm for accurate prediction of heart disease risk level. PAC algorithm is built using existing machine learning algorithms. It covers up the disadvantages of the existing algorithms and in turn increases accuracy of prediction of disease risk level. Many hospitals and health care industries have huge amounts of patient data. With the tremendously growing population, the doctors and experts available are not in proportion with the population. Doctors may sometime fail to correctly diagnose the severity of the disease. Hadoop single node cluster is used to process Big Data. Map Reduce code is implemented for the designed algorithms. Figure 1.1 shows the flow diagram for implementation of Map Reduce Programming for the uploaded data set.

Mapper:

Inside Mapper function each line from input file is taken as input to map phase and is fed to different map-tasks in parallel, considering multi-node cluster each node follows the same procedure in parallel. If there are N lines in input file and we have default M map tasks then number of lines processed by each map task is N/M.

The mapper function executes our algorithm on each and every map task of node and hence on each and every node in a multi-node cluster. Every time it takes single line from Big Data as input and processes machine learning algorithm to calculate risk level. Here the line number is taken as key and entire line is taken as value. The risk level is supplied as key to reducer and value is assigned to whatever attribute we wish to evaluate with. The context file is the intermediate output given by mapper function as input to reducer function.

Reducer:

The reducer shuffles the risk level provided by context file and sorts them according to key values provided to reducer function in ascending order and stores the sorted output in a file. The map-reduce jobs are used to process Big Data in both the algorithms. Different Map-Reduce functions are implemented to calculate the graphs for different attributes versus number of people with and without disease. This can be employed for various population surveys.

RESULT AND DISCUSSION

The output of the project can be either a report of a single patient for form based input or graphical output if Big Data file is provided as input. A comparative study of machine learning algorithms explained above is made and an accuracy graph is plotted to determine the best algorithm for disease prediction. This includes multiple aspects of the study such as the total number of patients who have and do not have heart disease, number of patients of a particular age who have and do not have disease etc. all these aspects are shown in graphical format so that it is easier for the user to understand. Figure 1.2 shows the comparative study of the Machine Learning algorithms as explained in the paper Naïve Bayes for continuous variables (red), Naïve Bayes for discrete variables (blue) and PAC Algorithm (green). And Figure 1.3 shows a sample patient report.

Fig 1.4 shows Risk versus Number of People.This graph gives the total number of people without and with disease.Red color represents people without disease and blue color represents total number of people with disease.

Conclusion and Future Scope:

Health care related data are huge in nature and they arrive from various birthplaces which are not suitable in structure or quality. These days, the utilization of knowledge and experience of copious

specialists and medical screening data of patients collected in a database during the diagnosis process, has been widely accepted. Implementing accurate machine learning algorithms to determine the heart disease risk and comparison of algorithms is done to determine the accuracy using graphs. It is easier to understand the graphs and the user can also determine his own risk level and get the report for the same. The project can be used in hospitals and research centers to analyze the heart disease attributes and how they contribute to the disease and its prediction. Multiple efficient machine learning algorithms can be included in this to more accurately predict the heart disease and compare it with other algorithms. Statistical study can be done precisely by focusing on important attribute for medical study of heart disease over big population set. The project for prediction of heart disease can be extended for prediction of similar disease by collecting the medical data from hospitals and medical centers and create a similar diagnosis web forum as done in heart disease. Depending on the increasing requirement multi nodes can be added to the cluster to decrease the execution time and process more data.

Table 1: Input attributes.

Predictable attributes:
Value=0 (safe level)
Value=1 (low risk level)
Value=2(medium risk level)
Value=3 (high risk level)
Value=4(Very high risk level)
Input Attributes
1.Age in Year
2. Sex (value 1: Male; value 0: Female)
3. Chest Pain Type (value 1:typical type 1 angina, value 2: typical type angina, value 3:non-angina pain; value 4: asymptomatic)
4. Fasting Blood Sugar (value 1: >120 mg/dl; value 0: <120 mg/dl)
5. Restecg – resting electrographic results (value 0:normal; value 1: having ST-T wave abnormality; value 2: showing probable or definite left ventricular
6. Exang - exercise induced angina (value 1: yes; value 0: no)
7. Slope – the slope of the peak exercise ST segment (value 1:unsloping; value 2: flat; value 3: down sloping)
8. CA – number of major vessels colored by fluoroscopy (value 0-3)
9. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
10. Trest Blood Pressure (mm Hg on admission to the hospital)
11. Serum Cholesterol (mg/dl)
12. Thalach – maximum heart rate achieved
13. Oldpeak – ST depression induced by exercise

PAC algorithm:

PAC (input.csv)
{ fp=input.csv
fw=training.csv
fq=discretize (fp)
while fq!=EOF
for each line in fq
for each line in fw
$\alpha i = \sum 1$ (for each matching attribute)
13 Where i= different risk levels
End For
$\beta i = \alpha i/SPi$
Where SPi is the number of supporting cases.
For each risk level
End For
μi= Normalizing Factor j x βi
r=maximum (μi)
Output r as the risk level
End While
}maximize (μi)
{ Return index for which μi is maximum for all i=0, 1,2,3,4
}discretize (file pointer)
{Assign continuous variables discrete values Vi by splitting into equal intervals with varying ranges
Return the dataset which has maximum accuracy }

Fig. 1: Proposed PAC algorithm.

Australian Journal of Basic and Applied Sciences, 9(7) April 2015, Pages: 639-643

Table 2: Performance Comparison.

Machine Learning Algorithms	Accuracy
NAÏVE BAYES CONTINUOUS VARIABLE	67.4%
NAÏVE BAYES DISCRETE VARIABLE	78.2%
PROBABILISTIC ANALYSIS & CLASSIFICATION	88.74%

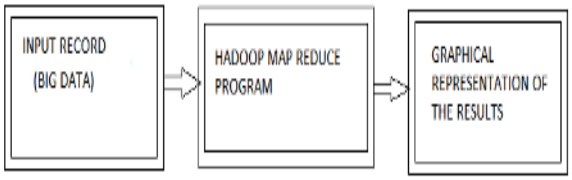


Fig. 2: Flow diagram of Big Data processing.

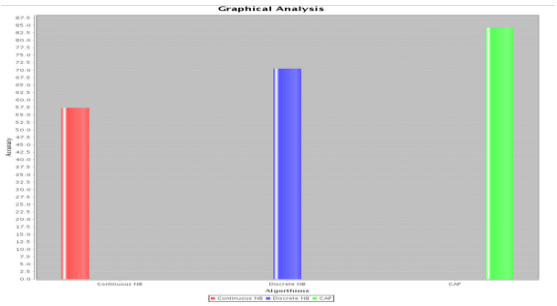


Fig. 3: Comparison of Machine learning Algorithms.

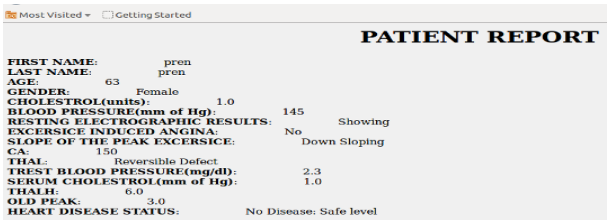


Fig. 4: Sample Patient Report.

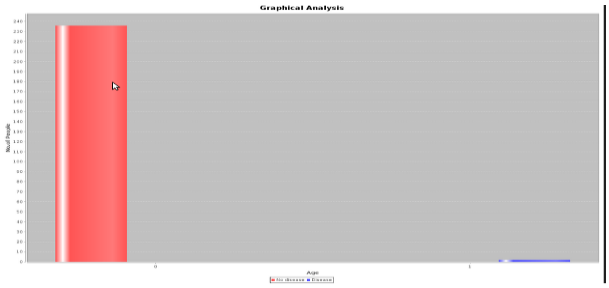


Fig. 5: Performance measure.

REFERENCES

http://www.rohitmenon.com/index.php/introducing-mapreduce-part-i/

http://www.javacodegeeks.com/2013/08/writing-a-hadoop-mapreduce-task-in-java.html

http://developer.yahoo.com/hadoop/tutorial/module4.html

http://nxhoaf.wordpress.com/2013/01/04/hadoop-mapreduce-word-count-using-eclipse/

http://bigdatacircus.com/2012/09/09/hadoop-map-reduce-introduction-and-internal-data-flow/

https://www.harding.edu/fmccown/r/#barcharts

http://stackoverflow.com/questions/19510656/how-to-upload-files-on-server-folder-using-js

http://www.uniweimar.de/medien/webis/teaching/lecturenotes/machine-learning/unit-en-decision-trees-algorithms.pdf

Prediction System for heart disease using Naïve Bayes *Shadab Adam Pattekari and Asma Parveen Department of Computer Science and Engineering Khaja Banda Nawaz College of Engineering, Rouza Buzurg, Gulbarga-585 104, Karnataka, India.

.