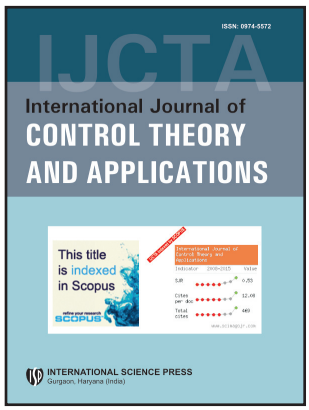

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 25 • 2017

Performance Comparison of SVM and C4.5 Algorithms for Heart Disease in Diabetics

Viswanathan K[a], Mayilvahanan K[b] and R. Christy Pushpaleela[c]

[a]*Technical Architect, Cognizant Technology, Chennai. Email: viswanathan_km@yahoo.co.in*
[b]*Head- Dean, Vels Univercity, Chennai. Email: hodmca@velsuniv.org*
[c]*Assistant Professor, Women's Christian College, Chennai. Email: leela_viswa@yahoo.com*

Abstract: The purpose of this research paper is to study and discuss the various classification algorithms applied on different kinds of medical data sets and also compares its performance. Among various classification algorithms, the performance analysis was done by considering an algorithm with maximum ac curacies on various kinds of medical data sets. Also this paper discusses the comparison of SVM and C4.5 algorithms on high dimensional patient data sets. In this paper, we will predict whether the diabetic patients will be suffered from heart disease or not.

Keywords: SVM, Weka, C4.5, Classification, Prediction, Medical Data set, and UCI, clustering, KDD, Diabetics, Heart Disease, KNN, Machine Learning (ML).

1. INTRODUCTION

Data mining is the process of predicting information from huge collections of data. Data mining uses mathematical analysis to derive patterns and trends that exist in data. Ideally, these pattern designs cannot be predicted by simple data study because the relationships are much complicated. These patterns and trends can be collected and defined as a model for data mining process. Data mining is also called as the procedure of mining knowledge from data. The following steps are more required for data mining process.

- Problem Definition
- Data prepare
- Data Exploring
- Model Building
- Exploring & validating models
- Deploying & updating models

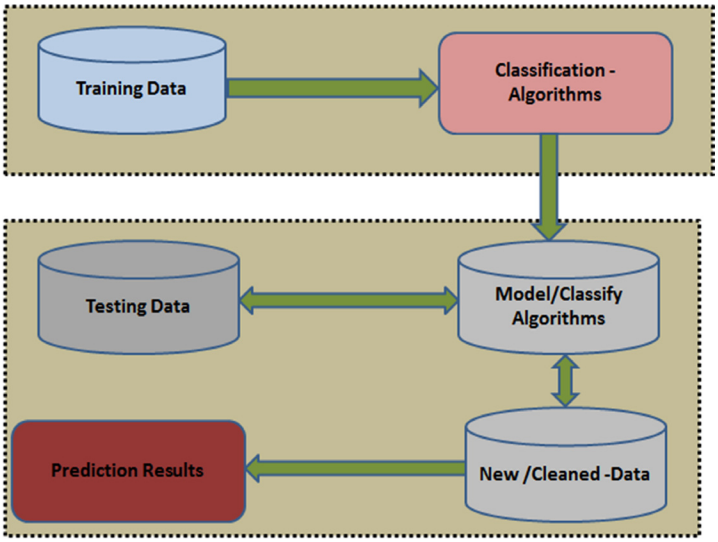

Figure 1: Data Mining –Process Flow

There are two forms of data analysis that can be used for extracting models.

- Classification
- Prediction

Classification: Method is a type of the data mining approaches and it is used to predict, classify the predetermined data for the specific class. The purpose of classification is to accurately predict the target dataset. A classification model could be used to identify target data by determining its various categories like low, medium, or high credit risk for loan applicants. There are two ways to classify the data

- Supervised - Set of possible classes is known in an advance.
- Unsupervised - Set of possible classes is not known called **Unsupervised**

Prediction: Model predicts continuous valued functions. Means that predicts **unknown or missing values.**

2. PRIMARY OBJECTIVE

Primary Intents of current work are proposed as below

- To Pre-Process the Patient data for Diabetics in Heart disease.
- To apply the classification algorithm for Diabetics in Heart disease.
- To classify the best algorithm for Diabetics in Heart disease.

3. WEKA-INTRODUCTION

Waikato Environment for Knowledge Analysis (**Weka**) is a popular machine learning software and it's written in Java language. It was developed at the University Of Waikato, New Zealand. It is open source software. Weka supports several standard data mining tasks and listed as below.

- Data pre-processing and Clustering
- Classification and Regression
- Visualization
- Feature selection

Figure 2: Weka Tool

3.1. Key Advantages of WEKA

- It is platform independent tool
- WEKA contains a GUI and an Open Source
- Large collections of different data mining algorithms

3.2. Data Pre-processing -Steps using Weka

The initial step to run the Weka tool is to launch the explorer window and select the "**Pre-process tab".** It is required toopen the data-sets in **.ARFF file format** and choose the attribute field in the data set (e.g. number of instances, attributes and classes etc). After datasets upload, need to run the **Visualization** button to see the pre-processed data.

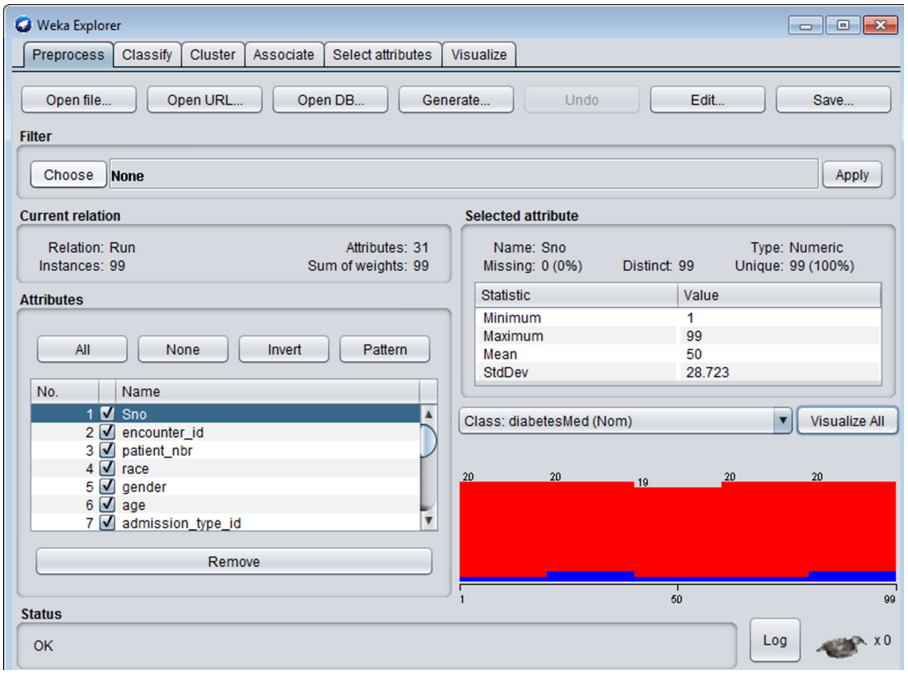

Figure 3: Weka - Pre-Process flow

Figure 4: Weka – Process -Visualization

4. CLASSIFICATION OF ALGORITHMS

There are different classification methods proposed by many researchers. The fundamental algorithms are considered and given as below,

- Support Vector Machine (for linear and nonlinear data- **SVM**)
- C4.5
- Decision Tree
- Bayesian classification
- Decision tree
- K-nearest neighbor classifier(KNN)

4.1. Support Vector Machine (SVM)

A support vector machine is a classification method used in datasets classification and regression.This is a non-linear classification algorithm and will help to perform data mining, text mining andpattern recognition. Also it often reported as better classification results over to other classification algorithms like C4.5, Decision Tree etc, It delivers good and precise solution to optimal dataset problems.

4.2. Decision Tree

Decision tree is a classic supervised learning algorithm. It's an easy classification method to understand and to perform data mining classifications. Decision Tree will build a predictive model which is mapped to a tree representation structure with a form of left and right child, root node compensations such that, every internal non-leaf node is labeled with values of the attributes.

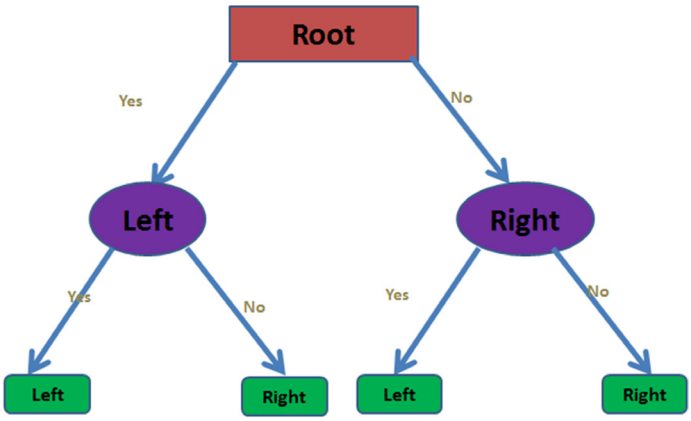

Figure 5: Decision Tree – Process flow

4.3. C4.5

It is a successor of ID3 algorithm which was used earlier and used to generate a decision tree. C4.5 algorithm is a greedy algorithm and it was developed by Ross Quinlan. It can be used for classification of the data and so referred to as statistical classifier.

4.4. Naïve Bayes

It is a Standard group of probabilistic classifier based on applying Bayes' theorem. It's highly scalable and an easy model to build very large data sets. Bayes is that it only requires a small amount of training data to estimate the parameters. Naïve Bayes and also It also perform well in multi class prediction. It is easy and fast to predict class of test data set An advantage of naive Bayes is a small number of training data to estimate the parameters for classification.

4.5. K-NN (Lazy Learning)

K-NN is a type of instance-based learning. KNN is a group of simple algorithm like as Classification and Regression. The main advantages of KNN are below:

- Easy implementation
- Robust
- Very low cost

4.6. LDA Linear Discriminant Analysis -LDA

Supervised learning algorithm called Linear Discriminant Analysis. This method is used in statistics, pattern recognition and ML to find a linear combination of features. LDA is simple, for each class to be identified, calculate linear function of the attributes. LDA is categorized in to two different methods.

1. Transformation with class dependency
2. Transformation with class independency

5. DISEASE OVERVIEW

5.1. Diabetes

Diabetes commonly referred as a group of metabolic diseases in which there are high blood sugar levels over a long period. There are Three type of diabetes like Type 1 (insulin diabetes), Type 2 (non-insulin diabetes) and Gestational Diabetes (Due to high sugar in blood at the time of pregnancy).

5.1.1. Diabetes-Symptoms

- Weight gain/strange loss
- Polyphagia
- Polyuria
- Blurred vision
- Fatigue
- Itchy skin

5.2. Heart Disease

Heart disease is the major cause of passing in the world. "**Heart disease**" involves narrowed or blocked blood vessels which might lead to a heart attack, chest pain (angina) or stroke. Heart disease is mostly produced by the following influences.

- Blood sugar (Diabetes)
- Smoking
- High/low Depression
- Low/High cholesterol
- High blood pressure
- Age

5.2.1. Heart Disease-Symptoms

- Rapid or irregular heartbeats
- Dizziness
- Sweating
- Indigestion
- Pain in the chest, arm, or below the breastbone

6. DATA SET FOR DIABETES WITH HEART PROBLEM

Table 1
Data Set attribute

S.No.	*Attribute*
1	Age
2	Heart Rate
3	Chest Pain
4	Obesity
5	Blood Sugar
6	Blood Pressure
7	Cholesterol
8	BMI-Body Mass Index
9	Triceps skin fold thickness
10	Number of times pregnant
11	Plasma glucose concentration
12	Class Variable – 0, 1

6.1. Classification Matrix

The basic parameters used to classify the Heart disease using a classifier is its performance and accuracy. Classification Matrix displays the frequency of correct and incorrect predictions. It compares the actual values in the test dataset with the predicted values in the trained model. In the below given example for 454 test patients, the dataset contained 208 patients with heart disease and 246 patients without heart disease.

Table 2
Confusion matrix

Predicted	*Classified Healthy(0)*	*Classified - not Healthy(1)*
Actual Healthy (0)	TP	FN
Actual not Healthy (1)	FP	TN

where,

TP – True Positive

TN – True Negative

FP – False Positive

FN – false Negative

For measuring accuracy rate, the following mathematical model is used.

$$\text{Accuracy} = TP + TN/TP + FP + TN + FN$$

7. PERFORMANCE COMPARISON RESULTS

The below is shown that various data mining Algorithms were hired to analyze the obtained Diabetes data.

Table 2
Data Set attribute

Algorithm used	*TP*	*FN*	*FP*	*TN*	*Acc.%*
SVM	25	34	56	296	91.22
C4.5	14	45	18	334	84.68
k-NN	22	42	24	323	83.95
Decision Tree	16	37	44	314	80.29

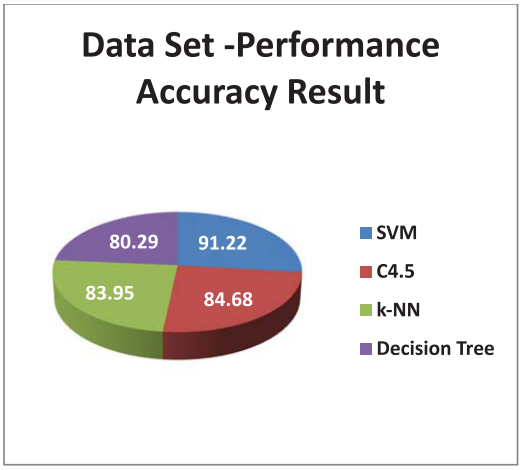

Figure 6: Performance Report for Classification Algorithm

8. CONCLUSION

This research work to classify the prediction of diabetes in heart disease considering the performance accuracy rate from the large datasets. Weka tool was made use to analyze the results for data validations. The result has been arrived with real data accuracy of performance comparison as the evaluating measurement. Based on the real data comparative results in the Table 2, it has been conclude that the top two classifiers namely SVM and C4.5 algorithms considered for data sets classification. Hence, the conclusion is that SVM tool based classification is much better than C4.5 algorithm.

REFERENCES

[1] International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, May 2013. A Study on WEKA Tool for Data Preprocessing, Classification and Clustering Swasti Singhal, Monika Jena.

[2] Payal Dhakate, Suvarna Patil, K. Rajeswari, Dr. V. Vaithiyananthan, Deepa Abin, - Preprocessing and Classification in WEKA using different classifiers||, Journal of Engineering Research and Applications www.ijera.com ISSN : 2248-9622, Vol. 4, Issue 8 (Version 1), August 2014.

[3] Remco R. Bouckaert, Eibe Frank, Mark A. Hall Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, —WEKA—Experiences with a Java Open-Source Project||, Journal of Machine Learning Research, November 2010.

[4] Thair Nu Phyu, "Survey of Classification Techniques in Data Mining" IMECS 2009.

[5] Fayyad, Piatetsky-Shapiro, Smyth and Uthurusamy, "Advances in Knowledge Discovery and Data Mining", (Chapter 1), AAAI/MIT Press 1996.

[6] Witten, I. and Eibe, F. Data mining practical machine learning tools and techniques.2nded, Sanfrancisco:Morgan Kaufmann series in data management systems., 2005.

[7] Pardha Repalli, "Prediction on Diabetes Using Datamining Approach".

[8] P. Padmaja, "Characteristic evaluation of diabetes data using clustering techniques", IJCSNS International Journal of Computer Science and Network Security, Vol. 8, No. 11, November 2008.

[9] http://www.openml.org/a/estimation-procedures/1

[10] http://www.medicalnewstoday.com/info/diabetes

[11] International Journal of Computer Science & Information Technology (IJCSIT) Vol. 3, No. 4, August 2011 "PERFORMANCE ANALYSIS OF VARIOUS DATAMINING CLASSIFICATION TECHNIQUES ON HEALTHCARE DATA" Shelly Gupta1, Dharminder Kumar2 and Anand Sharma3, 1AIM & ACT, Banasthali University, Banasthali, India shelly.gupta24@gmail.com 2Department of CSE, GJUS&T, Hisar, Indiadr_dk_kumar_02@yahoo.com 3Department of CSE, GJUS&T, Hisar, Indiaandz24@gmail.com.

[12] International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, May 2013 A Study on WEKA Tool for Data Preprocessing, Classification and Clustering Swasti Singhal, Monika Jena.

[13] Payal Dhakate, Suvarna Patil, K. Rajeswari, Dr. V. Vaithiyananthan, Deepa Abin, ―Preprocessing and Classification in WEKA using different classifiers‖, Journal of Engineering Research and Applications www.ijera.com, ISSN : 2248-9622, Vol. 4, Issue 8 (Version 1), August 2014.

[14] Remco R. Bouckaert, Eibe Frank, Mark A. Hall Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, ―WEKA―Experiences with a Java Open-Source Project‖, Journal of Machine Learning Research, November 2010.

[15] Performance Analysis of various Data mining classification (C4.5 Vs SVM) Techniques on Diabetics in Heart Problem. Viswanathan K, Dr. Mayilvahanan K, R. Christy Pushpaleela.

[16] Thair Nu Phyu, "Survey of Classification Techniques in Data Mining" IMECS 2009.

[17] Fayyad, Piatetsky-Shapiro, Smyth and Uthurusamy", Advances in Knowledge Discovery and Data Mining", (Chapter 1), AAAI/MIT Press 1996.

[18] Witten, I. and Eibe, F. Data mining practical machine learning tools and techniques.2nded, Sanfrancisco: Morgan Kaufmann series in data management systems. 2005.

[19] Pardha Repalli, "Prediction on Diabetes Using Datamining Approach".

[20] Joseph L. Breault., "Data Mining Diabetic Databases: Are Rough Sets a Useful Addition".

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

[22] P. Padmaja, "Characteristic evaluation of diabetes data using clustering techniques", IJCSNS International Journal of Computer Science and Network Security, Vol. 8, No. 11, November 2008.

[23] P. Yasodha, M. Kannan, ―Analysis of a Population of Diabetic Patients Databases in Weka Tool‖, Research Vol. 2, Issue 5, May-2011.

[24] Vikas Chaurasia, Saurabh Pal, ―Data Mining Approach to Detect Heart Dieses‖, International Journal of Advanced Computer Science and Information Technology Vol. 2.

[25] D. Lavanya and Dr. K. Usha Rani, ―Ensemble decision tree classifier for breast cancer data International Journal of Information Technology Convergence and Services, Vol. 2, No. 1. February 2011.

[26] Prof. K. Rajeswari, Dr. V. Vaithiyanathan and Shailaja V. Pede, ―Feature Selection for Classification in Medical Data Mining, International journal of emerging trends and technology in computer science. Vol 2, Issue 2, March – April 2013.

[27] J.S. Raikwal, Kanak Saxena, "Performance Evaluation of SVM and K-Nearest Neighbor Algorithm over Medical Data set", 2012, International Journal of Computer Applications.

[28] Mai Shouman, Tim Turner, and Rob Stocker, "Using Decision Tree For Diagnosing Heart Disease Patients", 2011, 9-Th Australasian Data Mining Conference (Ausdm'11), Ballarat, Australi.

[29] Ms. Rupali, R. Patil, "Heart Disease Prediction System Using Naive Bayes and Jelinek-Mercer Smoothing", 2014, International Journal of Advanced Research in Computer and Communication Engineering.

[30] Krati Saxena, Dr. Zubair Khan, Shefali Singh, "Diagnosis of Diabetes Mellitus Using K Nearest Neighbor Algorithm", 2014, International Journal of Computer Science Trends And Technology (Ijcst).

[31] Nongyao Nai-aruna, Rungruttikarn Moungmaia, "Comparison of Classifiers for the Risk of Diabetes Prediction", 2015, Procedia Computer Science 69.

[32] D. Sheela Jeyarani, G. Anushya, R. Rajarajeswari, A. Pethalakshmi, "A Comparative Study of Decision Tree and Naive Bayesian Classifiers on Medical Datasets", 2013, International Journal of Computer Applications.

[33] Nipjyoti Sarma, Sunil Kumar, Anupam Kr. Saini, "A Comparative Study on Decision Tree and Bayes Net Classifier for Predicting Diabetes Type 2", 2014, International Journal of Scientific Research Engineering & Technology (IJSRET).

[34] T. Revathi, S. Jeevitha, "Comparative Study on Heart Disease Prediction System Using Data Mining Techniques", 2013, International Journal of Science and Research (IJSR).

[35] Nathalie Villa and Fabrice Rossigive, "Support Vector Machine for Functional Data Classification", April 2005.

[36] Jaiwei Han, Micheline Kkamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006, pp 360-361.