Prediction of Heart Disease using SVM

Nagaraj M. Lutimath, Arathi B N,Shona M

Abstract*: Support Vector Machine (SVM) is an important classification method in data mining. It is a supervised classification technique. It finds a hyperplane for classification of the target classes. The heart disease consists set of disorders affecting the heart. It includes blood vessel problems such as irregular heart beat issues, weak heart muscles, congenital heart defects, cardio vascular disease and coronary artery disease. Coronary heart disorder is a familiar type of heart disease. It reduces the blood flow to the heart leading to a heart attack. In this paper the UCI machine learning repository data set consisting of patients suffering from heart disease is analyzed using support vector machines. The classification accuracy of the patients suffering from heart disease is predicted. Implementation is done using R language.

Keywords: Support Vector Machines, UCI machine learning repository data set, Data Mining, R Studio

I. INTRODUCTION

Data Mining plays a vital role in diagnosing a heart disease. Some of the data mining techniques are decision trees, neural networks, Naïve Bayes classification, genetic algorithms, regression and support vector machines. The decision tree algorithm is used for extracting rules in predicting heart disease. C5.0 decision tree procedure was accomplished using Cleveland data set. Its accuracy value of 85.33% was compared to the rest of the algorithms[1] [2]. It found to better than other data mining algorithms. A graphical user based interface was used to input the patient data and predict whether the patient is suffering from heart disease or not, using Weighted Association rule based Classifier. Results showed that Weighted Associative Classifier was providing improved accuracy as compare to other already existing Associative Classifiers. Naïve Bayes is a probability based classifier [3]. Medical attributes such as blood pressure, age, sex were used for prediction of heart disease. MatLab was used for implementation. A prediction model that uses combination of both pre pruning and post pruning of decision tree learning improved the classification accuracy by reducing the tree size [4].Other techniques in data mining such as regression, neural networks, support vector machines and genetic algorithms can also be utilized for prediction analysis.

Revised Manuscript Received on July 22, 2019.

Nagaraj M. Lutimath, Department of Computer Science and Engineering, Sri Venkateshwara College of Engineering, Bengaluru, India, Email: nagarajlutimath@gmail.com

Arathi B N,Shona M, Department of Computer Science and Engineering, Sri Venkateshwara College of Engineering, Bengaluru India, Email : aradvg@gmail.com, sonasuresh04@gmail.com

This paper provides a comparison of support vector machines with linear and sigmoid kernel function. The dataset used is UCI machine learning data set repository.
This paper is structured as follows, related works is explained in section II, methodology and data set analysis is described in section III section IV illustrates the feature engineering, section V presents prediction analysis and lastly section VI with conclusion.

II. RELATED WORKS

Heart disease is a vital disease explored by the researchers, in predicting the patients suffering from this disease. Generally a knowledge discovery process is used for extraction of hidden patterns in the data set. Data Mining is a knowledge discovery process of extraction information. It plays an important part in identification of the disease. The approaches utilized in data mining are decision trees, neural networks, Naïve Bayes classification, support vector machines and genetic procedures [5] for classification of the data set. Decision tree C4.5 and Fast Decision trees were studied [6] using a suitable medical data. Medical data sets was used from UCI repository. An accuracy of 69.5% was achieved for decision trees and an accuracy 78.54% was achieved for fast decision tree.

Analyses and predicting of coronary artery heart disease was done, utilizing a data set consisting of 335 records indicating the various 26 attributes [7]. The data set was pre-processed using correlation concept. Features selection and extraction was completed using particle swarm optimization (PSO) approach. The neural network, regression, fuzzy and decision tree models were modeled. The data set was applied to neural network model. An accuracy of 77% was found. It was further applied to regression model. This resulted in the accuracy of 83.5%. The other fuzzy and decision tree model did not show any major changes.

The data set was then optimized by utilizing the pre-processing approach. Correlation, feature selection and extraction with PSO, K-means clustering were used. The classification of the data set using one of the procedure or a combination of them was done. An accuracy of 88.4% was result for the regression model. The data set was further applied to hybrid model. The accuracy of classification procedures improved from 8.3 % to 11.4 %.

Another work on prediction of the heart disease was accomplished by pre-processing of the data by feature selection utilizing Ginni Index and support vector machine [8]. Classification of the data was further completed using suitable classification techniques. The algorithms used for classification were the Naive Bayes probability classification, Sequential Minimal Optimization (SMO) algorithm. SMO with bagging

and artificial neural network models were also added for analysis. An accuracy of 93.4% was obtained for SMO with bagging. 75.51% accuracy for Naive Bayes probability classification. 94.08 accuracy for SMO and 88.11 accuracy for the neural network models. Verification of the results was completed by 10-cross fold validation method.

An Apriori procedure using the Transaction Reduction Method (TRM) was applied to in diagnosing the heart disease using a suitable medical data set [9] [10]. The obtained results were compared it with some of the classical methods. An accuracy of 93.75% was achieved using the algorithm. When SMO was utilized 92.09% an accuracy was obtained. When SVM was used 89.11% accuracy was achieved. C4.5 decision tree resulted in 83.85% accuracy and Naïve Bayes probability classification an accuracy of 80.15% accuracy was the result.

All the techniques mentioned above deal with the predictive analysis using classical methods. The classification approaches like decision trees, Naive Bayes, Support Vector Machines or neural networks are the models for consideration using suitable medical data sets.

III. METHODOLOGY AND DATASET ANALYSIS

Data mining act of knowledge discovery from hidden data sets. Multi-dimensional data is collected from various sources and pre-processed and transformed into a suitable format. Then data mining procedures are applied on this data for further classification.

A. Experimental Procedures

SVM is a significant method for supervised classification. A hyperplane is utilized in classification of the target classes. Classification is performed by identifying the hyperplane that divides one class with the other classes. Training time for the SVM is very slow but it is very accurate in predicting the target classes.

IV. H FEATURE ENGINEERING

For studying classification process data set from UCI machine learning repository for heart disease at Cleveland is considered. The dataset is divided into two sets, the test data set and the training data set. The related feature engineering is done on the training data, and model thus obtained is utilized on the test data to predict the results.

The problem statement is defined as, *"To predict and analyze the value for the patients suffering from heart disease using support vector machine"*

To group the features with heart disease data set in order to analyze the number of patients with heart disease disorder.

Data Set used is the *"Heart disease diagnosis from the Cleveland dataset taken from UCI Machine Repository"*. The variables are defined as data fields as shown below.

Data attributes are,

v_age- age attribute given in years

v_sex- sex attribute categorized into male values 1 and female with value 0.

v_cp- chest pain attribute is categorized into values 1, 2, 3 and 4 in for angina, atypical angina, non-anginal pain, asymptomatic respectively.

v_trestbps-resting blood pressure (BP) attribute expressed in mm Hg, when the person is admitted to the hospital.

v_chol- serum cholesterol expressed in mg/dl

v_fbs- Fasting blood sugar > 120 mg/dl attribute with true and false indicated numerically by 1, 0.

v_restecg- attribute for resting electrocardiographic outcome expressed with values 0,1 for normal and S T-T wave abnormality(T wave inversions and/or ST elevation or depression of > 0.05 mV), 2= showing probable or definite left ventricular hypertrophy by Estes' criteria)

v_thalach- attribute for maximum heart rate of the patient.

v_exang- attribute for exercise induced angina indicated numerically by 1 and 0 for yes and no categorical values.

v_oldpeak- attribute for ST depression induced by exercise relative to rest

v_slope- attribute for the slope of the peak exercise ST segment expressed in terms of up sloping, flat and down sloping with values 1, 2 and 3 respectively.

v_ca- attribute for count of major vessels with a range from (0-3) with flourosopy coloring.

v_thal- attribute for type of heart defect with value 3 for normal, 6 for fixed defect and 7 for reversable defect

v_num- attribute for predicting the patients suffering from heart disease.

The input data set of 303 tuples is distributed into 227 tuples for training data set and 76 tuples into test data set. The dataset for training is executed in R and is taken using the equation 1 and 2.

split <-subset (dataset, SpiltRatio=0.75) (1)

training_set=subset (dataset, split=TRUE) (2)

The test data set is then calculated using equation 2.

The formula is computed using the equation 3 below,

formula = (v_num~v_age+v_sex+v_cp+ v_trestbps+ v_chol+ v_fbs+v_restecg+v_thalch+v_exang+v_oldpeak+ v_slope+ v_ca+ v_thal) (3)

In the equation (3) v_num is the predictor attribute, v_age, v_sex, v_cp, v_trestbps, v_chol, v_fbs, v_restecg, v_thalach, v_exang, v_oldpeak, v_slope, v_ca and v_thal are the response attributes.

The svm model is then constructed using the equations 4.

fit = svm (formula, data = training_set, type = 'C-classification', kernel = 'linear') (4)

The parameters used in the svm function of equation 4 are the formula which is used from equation 3, data is the training_set calculated from equation 2, type is C-classification used for classification analysis, and kernel is used for partitioning with values linear, polynomial, sigmoid and radial. In the above equation 4, kernel is used with linear value. The kernel with sigmoid is also used for further analysis.

Performance Measures

Some of the important parameters used in the performance analyses of the data set are the Mean Absolute Error (MAE), Sum of Squared Error (SSE) and Mean Squared Error (MSE). MAE is the mean of the absolute value of actual values minus the predicted values of the instances in the data set. SSE is summation of the squares of the actual instance values minus the predicted instance values of the data set MSE is

the mean of the squares of the actual instance values minus the predicted values in the data set.

V. PREDICTION ANALYSIS

Before prediction analyses the data is preprocessed and missing data are evaluated using mean of the attribute. The MAE, SSE and MSE are calculated for the test dataset heart disease data set and are listed in Table I.

In the Table I the values of MAE, SSE and MSE are calculated for linear and sigmoid kernel. The values of MAE, SSE and MSE are lower in case of sigmoid kernel than linear kernel. Now analyzing Table II, we find the lowest value of MAE is 0.77 for v_sex is male. MSE 1.26 for v_sex is female. We also observe that SSE is lower when v_sex is female, which also supports the evidence that the model predicts with higher accuracy when v_sex is female.

TABLE I. MAE, SSE AND MSE FOR OVERALL TEST DATA SET

Error Type	Value of Kernel=Linear	Value of Kernel = Sigmoid
MAE	0.84	0.61
SSE	100	98
MSE	1.32	1.29

TABLE II. MAE , SSE AND MSE FOR MALE AND FEMALE FOR v_sex for Kernel = Linear

v_sex	MAE	SSE	MSE
male	0.77	76	1.33
female	1..05	24	1.26

TABLE III. . MAE , SSE AND MSE FOR MALE AND FEMALE FOR v_cp for Kernel = Linear

Type of Error	Value of v_cp=1	Value of v_cp=2	Value of v_cp=3	Value of v_cp=4
MAE	0.63	0.83	1.11	0.76
SSE	5	10	29	56
MSE	0.63	0.83	1.53	1.51

TABLE IV. . MAE , SSE AND MSE FOR MALE AND FEMALE FOR v_slope for Kernel=Linear

Type of Error	v_slope=1	v_slope=2	v_slope=3
MAE	0.76	0.85	1.11
SSE	31	47	22
MSE	0.93	1.38	2.44

TABLE V. MAE , SSE AND MSE FOR MALE AND FEMALE FOR v_sex for Kernel = Sigmoid

v_sex	MAE	SSE	MSE
male	0.51	79	1.39
female	0.89	19	1

TABLE VI. . MAE , SSE AND MSE FOR MALE AND FEMALE FOR v_cp for Kernel = Sigmoid

Type of Error	Value of v_cp=1	Value of v_cp=2	Value of v_cp=3	Value of v_cp=4
MAE	0.63	0.83	0.74	0.46
SSE	5	10	26	57
MSE	0.63	0.83	1.37	1.54

TABLE VII. . MAE , SSE AND MSE FOR MALE AND FEMALE FOR v_slope for Kernel= Sigmoid

Type of Error	v_slope=1	v_slope=2	v_slope=3
MAE	0.76	0.5	0.44
SSE	31	51	16
MSE	0.94	1.51	1.78

Now observing Table III we see that minimum value of MAE and MSE is 0.63. This occurs when v_cp has 1 as its value. Thus the model predicts better in this case. We also observe that the highest value of MAE and MSE are 1.11 and 1.53. Thus the prediction model deviates from the actual values in this case.

Now observing Table IV we see that the lowest value of MAE and MSE are 0.76 and 0.93. This occurs when the v_slope has 1 as its value. Hence the prediction accuracy of the model is better in this case. The highest value of the MAE and MSE in Table IV are 1.11 and 2.44, this happens when the value of v_slope is 3, thus the prediction model deviates from the actual values in this case. The prediction model behaves moderately when the value of v_slope is 1.Using the tables Table II, Table III and Table IV, we find that the minimum MAE considering the attributes v_sex, v_cp and v_slope we get 0.63. This occurs for attribute v_cp for value 1.Thus the model predicts better for this value of the attribute v_cp. From tables Table II, Table III and Table IV, the minimum value of SSE is 5 for the same value of attribute v_cp. Hence the model predicts better for the attribute v_cp with value 1.

We now analyse the svm for sigmoid kernel. Consider Table V, we find the lowest value of MAE is 0.51 for v_sex is male. MSE 1 for v_sex is female. We also observe that SSE is lower when v_sex is female, which also supports the evidence that the model predicts with higher accuracy when v_sex is female.

Now observing Table VI we see that minimum value of MAE is 0.56 for v_cp is 4, and MSE is 0.63 for v_cp is 1. The minimum value for SSE is 5 when v_cp is 1. Thus the model predicts when v_cp is 1. We also observe that the highest value of MAE and MSE are 0.83.

Thus the prediction model deviates from the actual values when v_cp is 2.

Now observing Table VII we see that the lowest value of MAE and MSE are 0.44 and 0.94. This occurs when the v_slope has 2 and 1 respectively. The lowest value SSE is 16. This occurs when v_slope is 3. Hence the prediction accuracy of the model is better in this case. The highest value of SSE is 51, this happens when the value of v_slope is 3, thus the prediction model deviates from the actual values in this case. Using the tables Table V, Table VI and Table VII, we find that the minimum MAE considering the attributes v_sex, v_cp and v_slope we get 0.44. This occurs for attribute v_cp for value 3.Thus the model predicts better for this value of the attribute v_cp.

From tables Table II, Table III and Table IV, the minimum value of SSE is 5 for the same value of attribute v_cp. Hence the model predicts better for the attribute v_cp with value 3.

Now analyzing Tables II to VI. We find the lowest value for attributes v_sex, v_cp and

v_slope. We see that the lowest value for MAE, considering Kernel with linear and sigmoid values, we get 0.51. This occurs when SVM's kernel value is sigmoid for attribute v_sex is male. The lowest value for MSE, considering kernel with linear and sigmoid values, we get 0.63. The value is true when SVM is linear and sigmoid. The lowest value of SSE are 0.63 for both linear and sigmoid kernel SVM for v_cp is 1. The lowest value of SSE is 5 for both linear and sigmoid kernel for v_cp is 1. So better accuracy for SVM's is executed for v_cp attribute for value 1.

For the attribute v_sex the minimum value for MAE is 0.51. This occurs when v_sex is male and SVM is sigmoid. The minimum value for MSE for the v_sex is 1. This occurs when v_sex is female for sigmoid kernel SVM. The SVM with sigmoid performs better in this case. The minimum value for SSE for v_sex is 19. This occurs when SVM is sigmoid. Thus SVM with sigmoid kernel predicts better v_sex. Now for v_slope the minimum value for MAE is 0.44 for sigmoid kernel SVM. The sigmoid kernel SVM predicts better than linear SVM in this case. The lowest value of MSE for v_slope is 0.93. This is true for linear kernel SVM. Linear kernel performs better in this case. Considering the minimum value of SSE for v_slope, we get 16. This true for sigmoid kernel. Considering v_sex, v_cp and v_slope attributes, sigmoid kernel performs better for v_sex and v_slope. We find sigmoid kernel SVM better than linear SVM. Observing the values of MAE for v_sex for linear and sigmoid kernels, we find that MAE is lower for male than female. Hence males are affected more than the females in this case. Observing the value of MSE for a given linear and sigmoid kernel, we find MSE and for female is lower than male. Thus females are affected by heart disease than male in this case. Using the consistency measure MSE, we predict that females are affected by heart disease than males.

VI. CONCLUSION

In this paper the SVM is used for prediction for the heart disease taking the UCI Cleveland data set repository. The SVM in terms of linear and sigmoid kernels is analyzed. MAE, SSE and MSE are calculated utilizing suitable attributes of the data set. Sex, cp and slope features are taken for analyses and prediction of the heart disease. In comparison with SVM with linear kernel and sigmoid kernel. SVM with sigmoid kernel offers better accuracy than linear kernel. The data set containing male and female attributes is also analyzed. We find female are more affected by the heart disease than male using the consistency measures MSE. In future other data mining techniques such as deep learning, association rule analysis and genetic algorithms will be studied in predicting the accuracy with suitable performance parameters.

REFERENCES

[1] Moloud Abdar,"Using Decision Trees in Data Mining for Predicting Factors Influencing of Heart Disease", Carpathian Journal of Electronic and Computer Engineering 8/2 , 2015,pp. 31-36.

[2] Jyoti, S., U. Ansari and D. Sharma, Sunita Soni, "Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers", .International Journal on Computer Science and Engineering (IJCSE),3: 23852392, 2011, pp. 2385-2392.

[3] Rupali, M and R.Patil, "Heart Disease Prediction System using Naive Bayes and Jelinek-mercer smoothing",.International Journal of Advanced Research in Computer and Communication Engineering, May 2014. Vol. 3, Issue 5,pp. 6787-6789.

[4] Ali Mirza Mahmood1, 2* Mrithyumjaya Rao Kuppa, "Early detection of clinical parameters in heart disease by improved decision tree algorithm", Second Vaagdevi International Conference on Information Technology for Real World Problems, 2010, pp. 2429.

[5] František Babič, Jaroslav Olejár, Zuzana Vantová, Ján Paralič, "Predictive and Descriptive Analysis for Heart Disease Diagnosis", Proceedings of the Federated Conference on Computer Science and Information Systems, Prague, 2017, ISSN 2300-5963 ACSIS, Vol. 11,, DOI: 10.15439/2017F219, pp. 155–163.

[6] R. El-Bialy, M. A. Salamay, O. H. Karam, and M. E. Khalifa, "Feature Analysis of Coronary Artery Heart Disease Data Sets", Procedia Computer Science, ICCMIT 2015, vol. 65, pp. 459–468, doi: 10.1016/j.procs.2015.09.132.

[7] L. Verma, S. Srivastaa, and P.C. Negi, "A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data", Journal of Medical Systems, vol. 40, no. 178, 2016, doi: 10.1007/s10916-016-0536-z.

[8] R. Alizadehsani, J. Habibi, M. J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, and Z. A. Sani, "A data mining approach for diagnosis of coronary artery disease", Computer Methods and Programs in Biomedicine, vol. 111, no. 1, 2013, pp. 52-61, doi: 10.1016/j.cmpb.2013.03.0.

[9] Ch. Yadav, S. Lade, and M. Suman, "Predictive Analysis for the Diagnosis of Coronary Artery Disease using Association Rule Mining", International Journal of Computer Applications, vol. 87, no. 4, 2014, pp. 9-13.

[10] František Babič, Jaroslav Olejár, Zuzana Vantová, Ján Paralič, "Predictive and Descriptive Analysis for Heart Disease Diagnosis", Proceedings of the Federated Conference on Computer Science and Information Systems, Prague, 2017, ISSN 2300-5963 ACSIS, Vol. 11,, DOI: 10.15439/2017F219, pp. 155–163.t.