ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 3, Issue 5 , May 2016

Performance Analysis of Neural Networks and Support Vector Machines using Confusion Matrix

Maria Navin J R, Balaji K

Assistant Professor, Department of CSE, Sri Venkateshwara College of Engineering, Bangalore, Affiliated to VTU, Karnataka, India

Assistant Professor, Department of CSE, Sri Venkateshwara College of Engineering, Bangalore, Affiliated to VTU, Karnataka, India

ABSTRACT: Confusion matrix plays an important role in describing the performance of classification models. This paper describes a comparison between neural network and support vector machine classification models by creating confusion matrix on Iris data set. The Iris data set contains three different classes namely Iris Setosa, Iris Vercicolor and Iris Virginica. Using the confusion matrix accuracy for each classifier and statistical parameters like kappa, sensitivity, specificity, positive prediction value, negative prediction value, prevalence, detection rate, detection prevalence and balanced accuracy can be obtained. The proposed paper gives a comparative study of the statistical parameters to show which classifier performs better classification on the Iris data set.

KEYWORDS: Confusion matrix, performance, support vector machines, neural network, accuracy, kappa, iris data set.

I. INTRODUCTION

In recent research the aim of machine learning is to find hidden patterns, unknown correlations, and useful information from data. Support vector machines (SVM) and neural networks (NN) are proved to be good classification tools. Support vector machine performs classification by constructing a flat boundary in the form of a hyper plane for maximizing the width of the margin between two classes in multidimensional space. The neural network consists of a network of artificial neurons. The types of neurons within the network are input neurons, hidden neurons and output neurons. Neurons are connected inside the network and weights are used to indicate the connection strength.

Input information are received by input neurons, the activation value will be greater for higher input values. Then, the activation value is passed through the network based on the weights and transfer functions in the graph. Now the activation values are summed up and modified with the transfer function by the hidden or output neurons. The activation value then flows through hidden neurons and reaches the output nodes. Now we can use the output value from the output neurons to classify the data. SVM are based on local approximation strategy and uses large number of hidden units whereas neural networks implement the global approximation strategy with very small number of hidden neurons. Recent works have shown that SVM can aid in the classification of different data sets and more will follow on further exploration of data [1].

Classification of brain MRI images by comparing SVM classifier and PNN classifier gives rise to increase in training and testing speed of feature extraction. The system provided better accuracy for PNN classifier than SVM classifier [2]. However, the performances of the two classifiers are different; SVM technique is more effective than ANN method which it gives the average classification rate more than 83% [3]. So a comparative study of SVM and NN using Confusion Matrix with many Statistical Parameters [4] for classification like accuracy, kappa, sensitivity, specificity, positive prediction value, negative prediction value, prevalence, detection rate, detection prevalence and balanced accuracy will be useful in analysing the performance of the classifiers.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 3, Issue 5 , May 2016

II. METHODOLOGY

2.1 Iris Data Set

It is a multivariate data set with four attributes. The data set contains three classes of Iris plant with 50 instances each. The classes are Iris Setosa, Iris Versicolour and Iris Virginica and the 4 attributes are sepal length, sepal width, petal length and petal width all measured in cm as shown in Fig. 1.

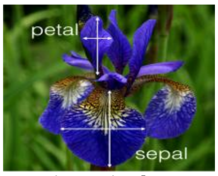

Fig. 1 Iris Plant

2.2 Confusion Matrix

A confusion matrix is a table that can be generated for a classifier on a Data Set and can be used to describe the performance of the classifier. This matrix is based on the terms
True Positives (TP) - These are cases were prediction and actual both are yes.
True Negatives (TN): These are cases were prediction is no and actual is yes.
False Positives (FP): These are cases were prediction is yes and actual is no.
False Negatives (FN): These are cases were prediction is no and actual is no.

2.3 Performance Measures

Positive predictive value (PPV) is the proportion of number of true positives by number of positives calls. **Negative predictive value** is the proportion of number of true negatives by number of negatives calls.
Kappa Score κ will be high if there is a big difference between the accuracy and the null error rate and can be calculated using

$$\kappa = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}m_{ij}}$$

where k=number of codes and w_{ij}, x_{ij}, and m_{ij} are elements in the weight, observed, and expected matrices respectively.
Accuracy indicates how often the classifier is correct.
True Positive Rate or **Sensitivity** is the percentage of True Positives to actual yes.
False Positive Rate is the percentage of False Positives to actual no.
Specificity is the percentage of True Positives to actual no.
Precision is the percentage of True Positives to predicted yes.
Prevalence is the percentage of actual yes to total number of instances.
Detection Rate is the rate of true events also predicted to be events.
Detection Prevalence is the prevalence of predicted events.
Balanced Accuracy is the sum of sensitivity and specificity divided by two.

III. EXPERIMENTAL RESULTS AND ANALYSIS

The comparison study using the Confusion Matrix on Iris Data Set was carried out using R which is a free software. R provides a wide range of statistical functions, allowing users to obtain the summary statistics of data, produce correlations and conduct statistical inferences.

In this paper we have used the following R libraries namely neuralnet, e1071 and caret on Iris Data Set. For modelling the Neural Network the neuralnet() function in the neuralnet library [5] is used with parameters namely data frame to be

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 3, Issue 5 , May 2016

modelled, number of neurons in the hidden layer for building the model. This function returns a neural network object that can be used for predictions.

Now predictions can be made using the compute() function with the neural network object as a parameter and the data frame containing the test data. This function returns a list of two components namely neurons for each layer in the network and the models predicted values.

In the case of modelling the support vector machine the svm() function in the e1071 library is used with the parameters for modelling namely data frame to be modelled and the cost of the support vector machine. This function returns a SVM object that can be used for predictions. Now predictions can be made using the predict() function with the SVM object as a parameter and the data frame containing the test data. This function returns a vector of predicted classes.

Now classAgreement() function can be used to calculate coefficients and the confusionMatrix() function from caret library in R [6] can be used to measure the prediction performance based on the classification table. The Confusion matrices & Overall Statistics for the classification of Iris Data Set using neural networks are shown in the Table 1 and Table 2. Similarly for support vector machine are shown in Table 3 and Table .4,

Table 1 Confusion Matrix for Iris Data Set using Neural Network

Classes	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	47	3
Virginica	0	3	47

Overall Statistics using neural network
Accuracy : 0.96
Kappa : 0.94

Table 2 Statistics by class using Neural Network

Classes	Setosa	Versicolor	Virginica
Sensitivity	1.0000	0.9400	0.9400
Specificity	1.0000	0.9700	0.9700
Pos Pred Value	1.0000	0.9400	0.9400
Neg Pred Value	1.0000	0.9700	0.9700
Prevalence	0.3333	0.3333	0.3333
Detection Rate	0.3333	0.3133	0.3133
Detection Prevalence	0.3333	0.3333	0.3333
Balanced Accuracy	1.0000	0.9550	0.9550

Table 3 Confusion Matrix for Iris Data Set using Support Vector Machine

Classes	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	48	2
Virginica	0	2	48

Overall Statistics using support vector machine
Accuracy : 0.9733
Kappa : 0.96

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 3, Issue 5 , May 2016

Table 4 Statistics by class using Support Vector Machine

Classes	Setosa	Versicolor	Virginica
Sensitivity	1.0000	0.9600	0.9600
Specificity	1.0000	0.9800	0.9800
Pos Pred Value	1.0000	0.9600	0.9600
Neg Pred Value	1.0000	0.9800	0.9800
Prevalence	0.3333	0.3333	0.3333
Detection Rate	0.3333	0.3200	0.3200
Detection Prevalence	0.3333	0.3333	0.3333
Balanced Accuracy	1.0000	0.9700	0.9700

From the confusion matrices and overall statistics we observe that the support vector machine performs better classification with an accuracy of 97.33% and kappa score of 96%. From the Table 2 and Table 4 we can also observe that the parameters namely sensitivity, specificity, positive predictive value, negative predictive value, detection rate, detection prevalence and balanced accuracy are better for support vector machine classifier.

IV. CONCLUSION

This paper presents a comparative study on the performance of neural network and support vector machine. The results were obtained using R statistical package with neuralnet, e1071 and caret libraries. Confusion matrices were created for neural network and support vector machine on Iris data set with three classes namely Iris Setosa, Iris Versicolor and Iris Virginica of 50 instances each. We conclude that the statistical parameters calculated from the respective confusion matrices indicated that support vector machine can perform better classification than neural network classifier in terms of accuracy, kappa, sensitivity, specificity, positive predictive value, negative predictive value, detection rate, detection prevalence and balanced accuracy.

REFERENCES

[1] B M Shawkat Ali,M Ameer Ali,G M Shafiullah, Colin Cole (2010). Smart Driving: A New Approach to Meeting Driver Needs A International Conference on Industrial Engineering and Operations Management.

[2] Ashwini, J. Sankhe, Prof. S.K.Bhatia (2016) Classification of brain MRI images by comparing SVM classifier and PNN classifier using Confusion Matrix, International Journal of Engineering Development and Research.

[3] Anith Adibah Hasseim, Rubita Sudirman, Puspa Inayat Khalid, Narges Tabatabaey, Comparison of ANN and SVM to Identify Children Handwriting Difficulties (2013) , Scietific Research, Engineering, 5, 1-5, (http://www.scirp.org/journal/eng).

[4] Vapnik, V. (1998) Statistical Learning Theory. John Wiley & Sons, Chichester.

[5] Günther, F., and Fritsch, S. (2010). neuralnet: Training of neural networks. The R journal, 2(1), 30-38.

[6] Kuhn, M. (2008), "Building predictive models in R using the caret package" Journal of Statistical Software.