Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 7, Issue. 4, April 2018, pg.135 – 140

COMPARING CLASSIFICATION MODELS FOR PREDICTING LIVER DISEASES

Mudit Wadhwa, Shallu Juneja

Computer Science and Engineering Department, Maharaja Agrasen Institute of Technology, India

Assistant Professor, Computer Science and Engineering Department, Maharaja Agrasen Institute of Technology, India

wmudit@gmail.com; shallujuneja9@gmail.com

Abstract— There is a constant rise in the number of liver disease patients with the most common reason being alcohol overdose, intake of drugs and medicines. This research work focuses to classify the given data using Support Vector Machine and Artificial Neural Network and label the patients as having liver disease or not. SVMs are a non-probabilistic binary classifier which label the new data into one of the two belonging categories. Artificial Neural Networks are designed to solve problems like the human brain can. They consist of nodes (known as artificial neurons) which are arranged in layers and can transmit information in the form of real number. The objective of this research is to compare the performance of SVMs and ANNs to classify the patients and to improve the accuracy of ANN using k-fold cross validation and hyperparameter tuning. The experimental results show that ANNs outperform the SVMs. The results can help the doctors in diagnosing liver diseases.

Keywords— Liver Disease, Data Mining, Classification, Support Vector Machine, Artificial Neural Network

I. INTRODUCTION AND RELATED WORK

Medical data mining involves extracting and analyzing the medical data in building certain prediction model to increase the accuracy of diagnosis in any specific disease [1]. Liver disease is a tricky disease to diagnose given the subtlety of symptoms while in the early stage. Problems with liver diseases are not discovered until it is often too late as the liver continues to function even when partially damaged [2].

The aim of this paper is to compare the performance of 2 classification techniques. The first classification technique is Support Vector Machines which represent the given examples in as points in space and divide them into 2 categories by a clear gap and then the new examples are also mapped in the same space and predicted which category they belong to based on which side of the gap they fall into.

The second technique used is the Artificial Neural Networks which is a layered connection of artificial neuron which are activated by an activation function and their weights are adjusted as the learning proceeds to try to learn the correlations among the data and classify the examples accurately.

Bahramirad in [1] used various machine learning classification algorithms to diagnosis liver diseases such as Logistic, Linear Logistic Regression, Bayesian Logistic Regression, Logistic Model Trees, Multilayer Perceptron, K-STAR, Ripper, Neural Net, Rule Induction, Support Vector Machine (SVM) and Classification and Regression Trees (CART) and obtained the accuracy of 66.52% and 71.24% for Neural Nets and SVM respectively.

II. DATASET USED

The dataset has been taken from University of California, Irvine machine learning dataset repository [3]. The data contains the records from North East of Andhra Pradesh, India and contains 583 records out of which 416 are liver patients and 167 are non-liver patients.

TABLE I. Attributes in the dataset

Attribute	Type
Age of patient	Integer
Gender	Categorical
Total Bilirubin	Real Number
Direct Bilirubin	Real Number
Alkaline Phosphatase	Real Number
Alamine Aminotransferase	Real Number
Aspartate Aminotransferase	Real Number
Total Proteins	Real Number
Albumin	Real Number
Albumin and Globulin Ratio	Real Number
Dataset (Class)	Binary

III. LITERATURE REVIEW

Liver diseases are mainly caused by obesity, hepatitis infections and alcohol misuse. Bahramirad in [1] used the Liver Disorder dataset (BUPA) and Indian Liver Patients dataset (ILPD) and used used 11 classification models to predict liver diseases. Accuracy of 66.52% was achieved by Neural Nets and 71.24% by the Support Vector Machine for ILPD dataset and for BUPA and for IPLD 73.91% by Neural Nets and 69.23% by Support Vector Machine. Further brute force optimization and Bayesian boosting was applied to get better results. The results achieved by brute force optimization by SVM is 78.29% accuracy for ILPD and 72.12% accuracy for BUPA and by Bayesian boosting 90% accuracy was obtained for both ILPD and BUPA.

Sonetakke in [2] also worked on the Indian Liver Patient Dataset compiled by University of California, Irvine and used machine learning models like Support Vector Machine and Backpropagation Neural Networks. Accuracy obtained by SVM was 71% and by neural network was 73.2%.

IV. CLASSIFICATION EXPERIMENT

The experiment was carried out in Python using Scikit-learn library for Support Vector Machine and Keras library using TensorFlow backend for Artificial Neural Network.

A. *Data pre-processing*

There were 4 records with missing Albumin and Globulin Ratio which were filled by the average value of all the records. The Gender attribute was in the form of strings (“Male”, “Female”) which was converted into categorical variable containing 1 or 0 using LabelEncoder.

B. *Classification Models*

- Support Vector Machine (SVM) – It is a non-probabilistic linear binary classifier which builds a model by training on the given labelled examples belonging to two classes and assigns new examples one of the class.

- Artificial Neural Network (ANN) – It is formed by interconnect nodes called artificial neurons which transmit signals to one another (analogous to synapse). The output of each neuron is a real number calculated by its activation function and the connections of the neurons have an assigned weight which are adjusted by the process of back propagation as the learning proceeds.

C. *Performance Comparison*

The metrics used to compare the performance of the models are accuracy, precision, recall and F score.

- Accuracy – It is the number of correct predictions made by the classifier divided by the total number of predictions made by the classifier.

$$\text{accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

- Precision – It is the number of positive predictions made by the classifier divided by the total number of positive predictions made by the classifier.

$$\text{precision} = \frac{TP}{TP+FP}$$

- Recall – It is the number of correctly identified positive predictions made by the classifier divided by total number of positive predictions.

$$\text{recall} = \frac{TP}{TP+FN}$$

- F Score - It is the harmonic mean of precision and recall.

$$\text{f score} = \frac{2 \times precision \times recall}{precision+recall}$$

where TP are the True Positives, TN are the True Negatives, FP are the False Positives and FN are the false negatives.

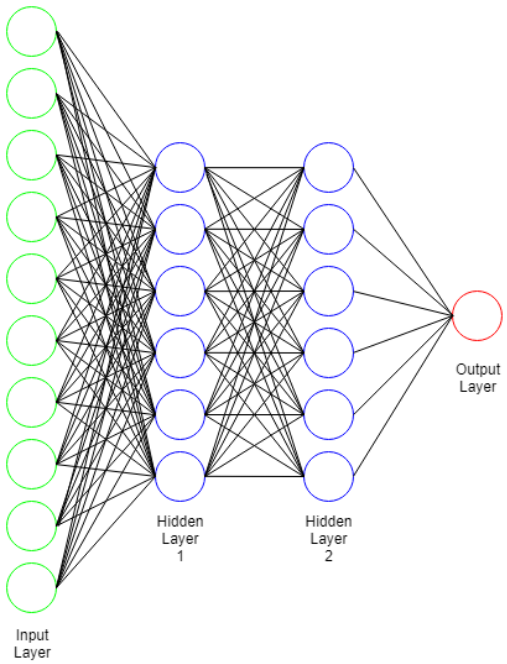


Fig. 1 Structure of ANN used

V. RESULTS AND DISCUSSION

The dataset was imported through the Pandas library and the Support Vector Machine and Artificial Neural Network algorithms were imported from Scikit-learn and Keras libraries respectively. The ANN has an input layer of 10 nodes and 2 hidden layers with 6 neurons each and rectifier activation function. There is a single output node with sigmoid activation function. The ANN was trained over a batch size of 10 with 100 epochs.

In a further attempt to reduce the overfitting during the training of the ANN 10-fold cross validation was applied. The mean accuracy after 10-fold cross validation was found out to be 72.73% and a variance of 5.7% which indicated the classifier has low bias and low variance. In the next step hyperparameter tuning was done to find out the best parameters for the training of the ANN. The results of hyperparameter tuning are shown in Table III.

The main result of 10-fold cross validation is that there was a considerable improvement in the precision of the ANN of 8.56% indicating that number of false positives were reduced. Hyperparameter tuning did not improve the original rather we saw a slight decrease in its overall performance.

Table II. Results of Classification

Model	Accuracy	Precision	Recall	F score
SVM	67.52%	78.16%	67.52%	55.27%
ANN	70.94%	69.99%	70.94%	70.27%

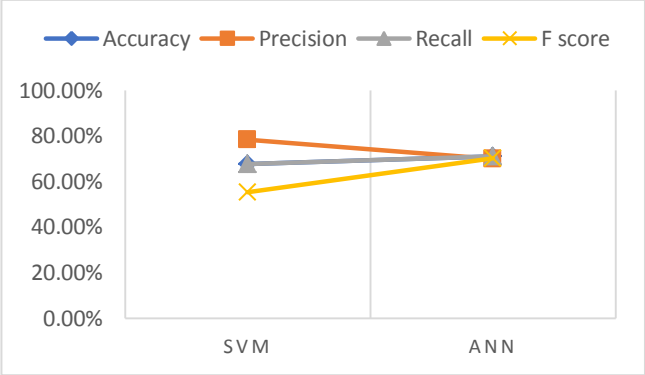
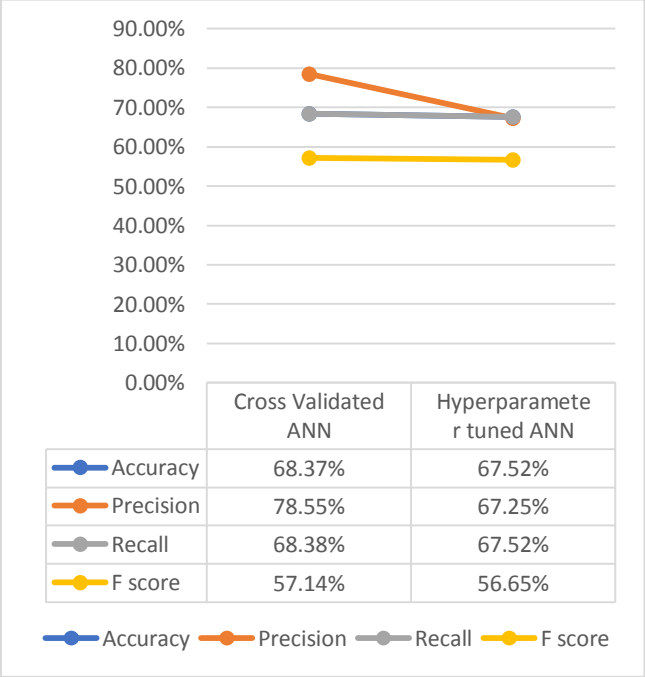


Fig. 1 Comparison of performance

TABLE III. Results of Hyperparameter tuning

Hyperparameter	Values tested	Result
Batch size	10, 25, 32	32
Epochs	100, 250	250
Optimizer	adam, rmsprop	rmsprop

TABLE IV. ANN Optimization Results



	Cross Validated ANN	Hyperparameter tuned ANN
Accuracy	68.37%	67.52%
Precision	78.55%	67.25%
Recall	68.38%	67.52%
F score	57.14%	56.65%

VI. Conclusion and Future Work

In this study classification models like Support Vector Machine and Artificial Neural Networks were used on the Indian Liver Patients Dataset [3] and was found out that ANN had a greater accuracy than the SVM by 3.42% but SVM was more precise than the ANN by 8.17% indicating that ANN had more false positives than the SVM.

The SVM used in [1] had a greater accuracy than the SVM in this paper but it had a lower precision and recall. However, after Bayesian Boosting Optimization the SVM in [1] outperformed the SVM in this paper in precision and recall too. The Neural Net in [1] was outperformed by the ANN in this paper in all performance metrics accuracy, precision and recall.

In the future, optimization of hyperparameters that have not been done in this paper can be done to improve the accuracy of the Artificial Neural Network and k-fold cross validation can be carried out with a value other than 10 to produce more accurate results. Other optimization techniques can be employed to make predictions on the dataset more accurately and thus helping in diagnosis of liver problems.

References

[1] S. Bahramirad, A. Mustapha, M. Eshraghi, "Classification of Liver Disease Diagnosis: A Comparative Study" in *Informatics and Applications(ICIA), 2013 Second International Conference*, pp. 42–46.

[2] S. Sontakke, J. Lohakare, R. Dani, "Diagnosis of liver diseases using machine learning", Emerging Trends & Innovations in ICT(ICEI), 2017, pp. 129-133.

[3] ILPD (Indian Liver Patients Dataset) Data Set, UCI Machine Learning Repository, 2012, available at https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)

[4] S. Alfisahrin, T. Mantoro, "Data Mining Techniques for Optimization of Liver Disease Classification", International Conference on Advanced Computer Science Application and Technologies,2013, pp. 379-384

[5] A. Banu, H. Ganesh, "A hybrid approach for an efficient classification using Decision Tree and SVM", International Journal of Computer Science and Mobile Computing (IJCSMC), 2018, pp. 42-48

[6] X. Jin, Q. Jin, X. Yang, "A Disease Detection Method of Liver Based on Improved Back Propagation Neural Network", International Symposium on Computational Intelligence and Design, 2015, pp. 111-113

[7] A. Alivar, H. Daniali, M. Helfroush, "Classification of liver disease using ultrasound image based on feature combination", International Conference on Computer and Knowledge Engineering (ICCKE), 2014, pp. 669-672
[8] M. Vasinek, J. Platos, V. Snasel, "Limitations on low variance k-fold cross validation in learning set of rule inducers", International Conference on Intelligent Networking and Collaborative Systems, 2016, pp. 207-214
[9] G. Diaz, A.Nkoutche, G. Nannicini, H. Samulowitz, "An effective algorithm for hyperparameter optimization of neural network", IBM Journal of Research and Development, 2017, pp 9:1-9:11

