Classification on Different Databases Using Artificial Neural Network (ANN) and 10-Fold Cross-Validation

Ngong Ivoline-Clarisse Kieleh

Computer Engineering, Faculty of Engineering and Technology
Selcuk University, Alaeddin Keykubat Yerleskesi Selcuklu, Konya, 42075 Turkey
ivolinengong@gmail.com

Abstract

In this paper, the author introduces a classification approach using multilayer perceptron with the backpropagation learning algorithm. This paper presents a comparative study on the performance of artificial neural networks on 3 different datasets. An artificial neural network with 10-fold cross validation was applied to UCI datasets – Iris, Wisconsin breast cancer and BUPA liver disorder datasets. The performance measures (accuracy, sensitivity and specificity) were measured and the values obtained were compared to other studies on the same databases. 97.33% , 97.38% and 71.53% accuracies were obtained for Iris, Wisconsin breast cancer and BUPA liver disorder datasets respectively.

Introduction

Nowadays, the purpose of machine learning has been to find masked structures, unknown relationships and important details from data.[1] This is exactly what Artificial Neural Networks do. They are computational models that are inspired by the brain. Simply put, a neural network is an algorithm that tries to mimic the biological structure or functioning of the brain. It is usually built by using electronic components or simulating a software on a computer. [2]. Neural networks have an exceptional capacity to acquire meaning from complex data, can observe trends and extract hidden patterns that are too complex to be done by other computational models or by humans.[5] This is reason why today is being applied in many diverse fields such as medicine, science, commerce, industries, economics and so on. In artificial neural networks we try to build a mathematical model that can simulate the working and design of a biological neural network.[4] A neural network is made up of artificial neurons which are inter connected with other neurons to form a network. These neurons have weights which determine the impact one neuron has on another. [5] It is composed of 3 layers; an input layer, hidden layer and output layer. Through weighted links the neurons in the input layer receive the data and transfer them to neurons in the hidden layer. After the data is processed in the hidden layer (which may be made up of many layers), the result is then transferred to the output layer. In the hidden layer the formula below is used for processing where the weighted sum is calculated and the bias(θj) is added to it.

$$\sum_{i=1}^{n} xi*wij+\theta j\ j=(1.2.3\ldots n)[5]$$

In summary, an artificial neural network is defined by the inputs, weight, bias, the activation function (transfer function) and the output

.

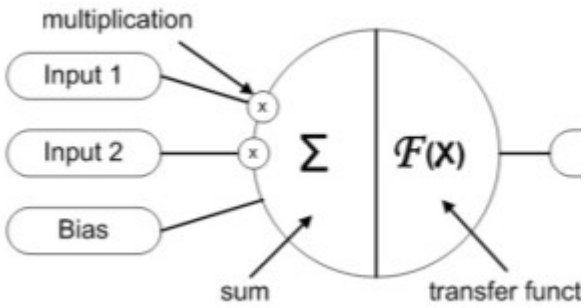

Figure 1 Artificial Neural Network [6]

Three features of ANN makes them stand out. Firstly, ANN acquires knowledge about a given problem domain in a very uncomplicated and effortless way through a training phase which is the opposite of what happens in most knowledge engineering AI systems. Secondly, the noncomplex manner in which the information is stored and can be easily accessed is an important characteristic of ANN. And very importantly in the presence of noise in the input data, ANN is very robust. Also, ANN are known to have a high degree of accuracy when used.[3]

It is also important we understand the meaning of cross-validation as it has been applied in the study. Cross validation is a statistical technique applied on algorithms where data is divided into 2 portions; one for training a model and the other to test and validate the model. All of the data points will be used for both training and testing but not at the same time. K-fold cross validation is a typical cross validation method. In this method the data is first divided into k sub segments (folds). Then k iterations of training and tests are performed such that within each iteration a different fold of the data is held-out for testing while the remaining k-1 folds are used for training.

Literature Review

Zheng, Huo, Guo et Fang[7] proposed a new supervised adaptive incremental clustering(SAIC) algorithm that can automatically cluster dynamic datasets of arbitrary shapes and sizes. It is made up of 2 phases; the learning and pre-processing phase. They made use of fourteen UCI datasets and four synthetic datasets to evaluate to evaluate the performance of SAIC. The results show that SAIC outperforms some supervised and unsupervised clustering algorithms.

P. H. Kassani, A. B. J. Teoh, E. Kim[8] presented an improved version of the k nearest neighbour and fuzzy algorithm called the multi-objective-genetic-algorithm-modified FNN(MOGA-MFNN). They introduced 2 new objective functions which were used to improve the generalization capability of MFNN for unobserved data. 20 datasets obtained from UCI were evaluated and the proposed method performed competitively with existing methods.

B. V. Ramana, M.S. P. Babu and N. B. Venkateswarlu [9] evaluated a selected classification algorithms for the classification of some liver patient datasets. Four algorithms were considered and they were evaluated based on accuracy, sensitivity, precision and specificity.

A. S. Aneeshkumar and C. J. Venkateswaran[10] in their paper used classification to predict previously unknown class of objects. They also carried out predictions for liver disorder disease as it is usually a difficult task for medical practitioners;

M. Swain, S. K. Dash, S. Dash and A. Mohapatra[11] focused on implementing a

neural network for classification on IRIS plants. The IRIS plant species were identified on the basis of plant attribute measurements.

S. T. Halakatti and S. T. Halakatti [12] as well identified IRIS flower species using a semi-automated extraction of knowledge of data. They made use of scikit learn tools for this classification and identigied flower species on the basis of glower attribute measurements.

R. Setiono[13] presented a new algorithm for pruning in neural networks. The algorithm ensures that networks which have a small number of connections and high accuracy rates for breast cancer diagnosis were obtained. A 95% accuracy was obtained for both training and test data.

A. M. Abdel-Zaher and A. M. Eldeib[14] propose a Computer-Aided Diagnosis(CAD) scheme for detection of breast cancer using belief networks and back propagation. The classifier complex gave an overwhelming accuracy of 99.68% indicating promising results compared to previously published studies.

B. V. Ramana and M.S. P. Babu[15] proposed a classification technique which is tree-based. They presented a modified Random Forest algorithm for UCI liver dataset which was done with multilayer perceptron classification algorithm and random subset feature selection technique.

R. Lin [16] employed a classification and regression tree with case-based reasoning techniques to build an intelligent liver disease diagnosis model. The proposed model raises the accuracy of existing models. The final model could be used as a supporting system in making decisions regarding liver disease diagnosis and treatment.

K. Patel, J. Vala and J. Pandya [17] in their paper implemented 4 algorithms with the IRIS dataset and compated the precision, recall, FP-rate, TP-rate and ROC curve parameters using Weka.

M. Navin JR and Balaji K[1] compared support vector machine and neural network classification models on the Iris dataset. They created a confusion matrix where the got statistical parameters for their comparative study.

W. Yue,Z. Wang, H. Chen, A. Payne and X. Liu [18] in their paper reviewed machine learning techniques and their applications in diagnosis and prognosis. They investigated the Wisconsin Breast Cancer dataset with applications built from 4 algorithms and provided a final healthcare system of their model.

A. Roshanpoor, M. Ghazisaeidi, S. Niakan, K. Maghooli and R. Safdari [19] also conducted a study on the Wisconsin Breast Cancer dataset. They made use of Statified 10-fold cross validation and their validation was repeated 100times to increase correctness of the outcome. The algorithm used here was KNN.

Methodology

The classification was implemented in the Python programming language using Jupyter Notebooks and Anaconda. Python packages like numpy(package for scientific computing), pandas(package that offers data structures and operations that make it easy to perform data manipulation and analysis) , matplotlib(visually represent the data for both exploration and reporting) and scikitlearn(contains various classification, regression and clustering algorithms) were used. Also made use of Keras (designed to enable fast experimentation with deep neural

networks) library which runs on TensorFlow. The classification was divided into 3 parts:

1. Data Preprocessing
2. Building ANN
3. Evaluating the model

In this section we talk about the first two phases.

Data preprocessing has a significant effect on ANN networks. ANN perform better and learn more quickly if the input variables are pre-processed before using them to train the network. All preprocessing techniques carried out on the train set should be carried out on the test set as well. A couple of preprocessing techniques were employed. First we checked for null values and removed any columns that had null values. For example in the Wisconsin Breast Cancer dataset the last column in the attributes were all null, so that column was excluded from the dataset. We then checked for any categorical variables and converted them into numeric variables that would be understood by our model. This was done using a feature called one-hot encode in Scikit Learn(this was applied to the Iris dataset). Next, we checked for missing values. No datasets had any missing values. We also checked for imbalance in the dataset. A dataset was considered to be balanced if the proportion of the class values was 60:40 or 50:50. Our datasets were also balanced. And finally , the last preprocessing technique carried out was feature scaling. This was done to equalize the importance of variables in the dataset. With all these preprocessing done, our data was now ready to be trained.

Now we go to building the model. Here, we look at each dataset separately:

- IRIS Dataset

  The IRIS dataset contains150samples, 4 features and 3 classes: setosa, versicolor and Virginian. Hence, this was a multiclass classification. As earlier mentioned in the paper, neural networks are defined as a sequence of layers. Here, we had the input layer, one hidden layer and output layer. The activation function used from the input to the hidden layer was the relu function(rectified linear unit) which sums the function from each neuron. The softmax activation function was used for the output. This was used because we have more than 2 classes in this dataset. The network was then compiled with a categorical_crossentropy loss function (again because it has multiple classes) and the adam was the optimizer.

- Wisconsin Breast Cancer and Liver Disorders Datasets

  The Wisconsin Breast Cancer dataset contains 569 samples with 30 features and 2 classes – Benign and Malignant while the Liver Disorders dataset has 345 samples, 6 features and 2 classes as well. Positive and negative disease diagnosis.

  For these 2 datasets, two hidden layers were used. The activation function used was relu for the hidden layers but sigmoid was used for the output layer (since they have just 2 classes). The binary_crossentropy loss function was used and adam was the optimizer.

10-fold Cross validation was then performed dividing the data into training

and test sets. The model was fitted on the training set and then evaluation was performed.

Results and Discussions

The models were then evaluated and we got the confusion matrix. The confusion matrix is a technique for summarizing the performance classification of algorithm. When dealing with confusion matrix it is common to hear terms like true positive (TP), true negative (TN), false positive (FN) and false negative(FN). From the confusion matrix we got the accuracy, sensitivity and specificity. These can be expressed in terms of TP, TN, FN and FP.

Sensitivity = TP / (TP + FN) = (Number of true positive assessment) / (Number of all positive assessment)

[20]

Specificity = TN / (TN + FP) = (Number of true negative assessment)/(Number of all negative assessment)

[20]

Accuracy = (TN + TP) /(TN+TP+FN+FP) = (Number of correct assessments)/Number of all assessments)

[20]

The confusion matrixes we got for classification on out datasets are shown below.

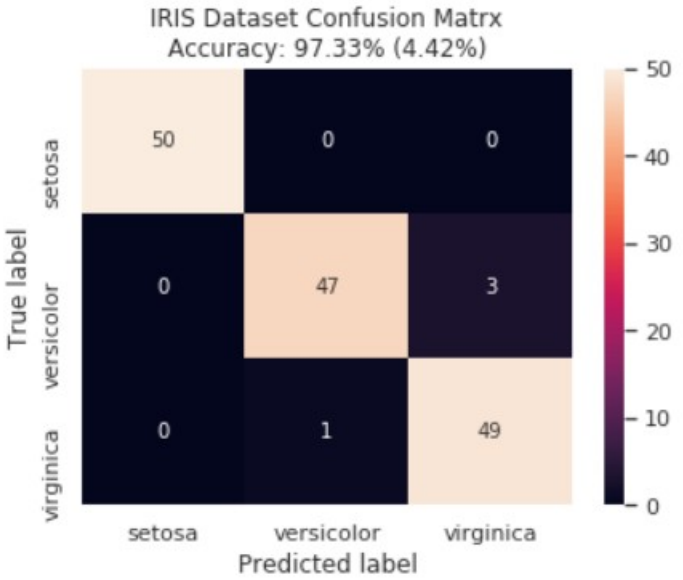

Figure 2 Confusion Matrix For Iris Dataset

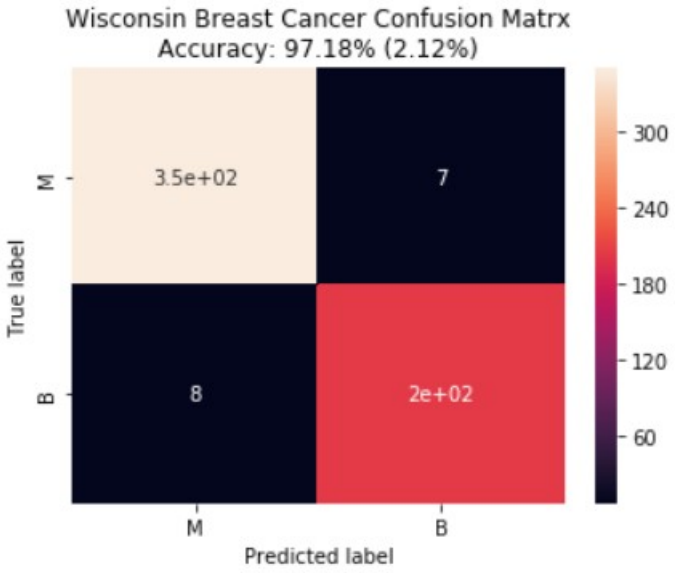

Figure 3 Confusion Matrix For Wisconsin Breast Cancer Dataset

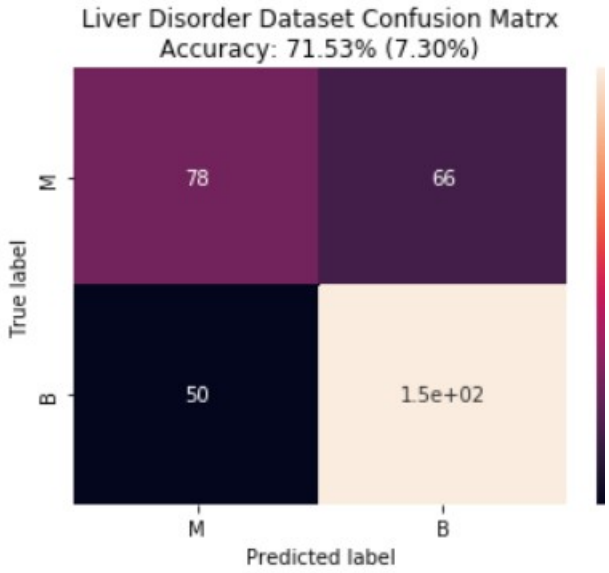

Figure 4 Confusion Matrix For Liver Disorder Dataset

The results are summarized in Table 1 below:

Table 1 Performance Measures From Our Study

UCI DataSets	Iris	Wisconsin Breast Cancer	Liver Disorder
Accuracy	97.33	97.38	71.53
Sensitivity	1.0	0.98	0.54
	0.94		
	0.98		
Specificity	1.0	0.96	0.75
	0.97		
	0.99		

We compare the result of our studies with other studies.

Table 2 Comparison of Accuracies of This Study with Other Studies

DataSets		n Breast Cancer	Disorder
UCI	**Iris**	**Wisconsi**	**Liver**
[1]	93.8		
[7]	97.8	94.5	68.3
[8]			95.73
[9]			88.38
[10]	96.6		
[11]	96.6		
[12]		95	
[13]		99.68	
[14]			73.3
[15]			90
[16]	96		
[17]	97.3		
[18]		94.74	
This Study	**97.33**	**97.38**	**71.53**

Conclusions

This paper presents a comparative study on the performance of ANN on different datasets. The results were obtained using Python Jupyter Notebook, with Keras, Tensorflow, Numpy and Pandas.

In this paper, we have provided explanations for the ANN approach, cross validation and its application to 3 different datasets. Several studies including this one has shown that ANN has remarkable ability to improve classification and prediction accuracy. The results from various studies have been show in Table 2 with references and classification accuracies. We can see that lots of algorithms have achieved very high accuracy these datasets. The Iris and Wisconsin Breast Cancer performed very well in this study, however there is still room for improvement. The accuracy for Liver disorder obtained was 71.53% which is not very high, so one can say the algorithm needs to be improved. In future studies, we will try to not only improve the accuracy rate obtained for Liver Disorder

but for all the other datasets. We also intend to add more datasets to the study.

References

[1] M. Navin JR and Balaji K, "Performance Analysis of Neural Networks and Support Vector Machines using Confusion Matrix", "International Journal of Advanced Research in Science,Engineering and Technology", vol. 3, Issue 5, 2016. (Introduction too)

[2] S. Haykin,"Neural Networks and Learning Machines","Pearson Prentice Hall", 3rd Edition,2008. (haykin)

[3] R. Andrews, J. Diederich and A. B. Tickle, "Survey and Critique of Techniques For Extractng rules from Trianed Artificial Neural Networks", "Knowledge-Based Systems", vol. 8, 1995 (info)

[4] S.Sathish Kumar, Dr N Duraipandian, "Artificial Neural Network Based Method for Classification of Gene Expression Data of Human Diseases along with Privacy Preserving", "International Journal of Computers & Technology",vol. 4,2013. (ANN)

[5] M. Zakaria, M. AL-Shebany, S. Sarhan, "Artificial Neural Network : A Brief Overview", "Int. Journal of Engineering Research and Applications",vol. 4,2014.

[6] A. Krenker, J. Bester and A. Kos, Introduction to the Artificial Neural Networks, Artificial Neural Networks - Methodological Advances and Biomedical Applications, Prof. Kenji Suzuki (Ed.), ISBN: 978- 953-307-243-2, InTech, Available from: http://www.intechopen.com/books/artificial-neural-networksmethodological-advances-and-biomedical-applications/introduction-to-the-artificial-neural-networks , 2011

[7] L. Zheng, H. Huo, Y. Guo and T. Fang," Supervised Adaptive Incremental Clustering for data stream of chunks", "Neurocomputing", http://dx.doi.org/10.1016/j.neucom.2016.09.054, 2016.

[8] P. H. Kassani, A. B. J. Teoh and E. Kim, "Evolutionary-modified fuzzy nearest-neighbor rule for pattern classification", "Elsevier Ltd.", 2017.

[9] B. V. Ramana, M.S. P. Babu and N. B. Venkateswarlu, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis", "International Journal of Database Management Systems" (IJDMS), Vol.3, No.2, 2011.

[10] A. S. Aneeshkumar and C. J. Venkateswaran, " Estimating the Surveillance of Liver Disorder using Classification Algorithms", "International Journal of Computer Applications (0975 – 8887) " Volume 57– No.6,2012.

[11] M. Swain, S. K. Dash, S. Dash and A. Mohapatra, "An Approach For Iris Plant Classification Using Neural Network", "International Journal on Soft Computing (IJSC)", Vol.3, No.1,2012.

[12] S. T. Halakatti and S. T. Halakatti , "Identification Of Iris Flower Species Using Machine Learning" , "IPASJ International Journal of Computer Science (IIJCS)", Volume 5, Issue 8,2017.

[13] R. Setiono, "Extracting Rules from Pruned Neural Networks for Breast Cancer Diagnosis", "Appears in Artificial

Intelligence in Medicine", Vol. 8, No. 1, 1996.

[14] A. M. Abdel-Zaher and A. M. Eldeib, "Breast Cancer Classification Using Deep Belief Networks", "Expert Systems With Applications",2016.

[15] B. V. Ramana and M.S. P. Babu, " Liver Classification Using Modified Rotation Forest"," International Journal of Engineering Research and Development",vol.1,2012.

[16] R. Lin, "An intelligent model for liver disease diagnosis", "Artificial Intelligence in Medicine",2009.

[17] K. Patel, J. Vala and J. Pandya, "Comparison of various classification algorithms on iris datasets using WEKA", "International journal of Advance Engineering and Research Development (IJAERD)", Vol. 1 Issue 1,2014.

[18] W. Yue,Z. Wang, H. Chen, A. Payne and X. Liu, "Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis", "MDPI",2018

[19] A. Roshanpoor, M. Ghazisaeidi, S. Niakan, K. Maghooli and R. Safdari, "The Performance of K-Nearest Neighbors on Malignant and Benign Classes: Sensitivity, Specificity, and Accuracy Analysis for Breast Cancer Diagnosis"," International Journal of Computer Applications (0975 – 8887)", Volume 180 – No.8, 2017

[20] W. Zhu, N. Zeng, N. Wang, "Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations", "NESUG", 2010.