Systematic literature review (SLR) for comparison of two classification models

By Aleksander Vanberg Eriksen

1. Introduction

This report will go through and review scientific articles that contains information about the topic I have chosen for my idea paper, comparison between multi-layer perceptron classifier and random forest classifier. This report will mainly contain information about MLP Classifier and Random Forest Classifier regarding data analysis, and how well they performed in terms of accuracy compared to data size.

The report will contain a background for why I wanted to research this topic, how I did my research, what type of information I gathered, a discussion part for how this is relevant to my topic, and my own code to compare to the information I have found.

2. Background

When doing my bachelor's degree, I took a course that was called Practical Machine learning which taught me about and how to use MLP classifier. From the course, I never really learned a lot about the other models that was usable. But when starting my master's degree, I took the course Machine Learning which has introduced me to the other models. The teacher has used the Random Forest model quite a lot, and then I wondered why we didn't learn about that in the bachelor's course.

The reason I chose to research about this topic is that I find it interesting that there are different models to use for classification, which means that there is a possibility that some models are better than others.

3. Systematic Review Protocol

To be able to fully answer my research questions, a SLR is done to gather relevant research articles that contain satisfying information. These articles will contain information about how a neural network functions in practice, and if there are any difference between the two distinct one.

3.1. Research questions

The research question that is raised is based on gaining information about how it works, and the difference between the two types.

RQ1: How does a neural network work when solving classification problems?

- RQ1a: How does MLP classifier work?
- RQ1b: How does Random Forest Classifier work?

RQ2: What's the difference in performance when solving a classification problem using MLP classifier and Random Forest Classifier?

3.2. Search strategy

This section contains the strategies I used to obtain the relevant research articles that contains information about the research questions.

3.3. Databases

Utilizing some known databases to find articles and a dataset.

- Kaggle: https://www.kaggle.com/
- Science direct: https://www.sciencedirect.com/
- ACM Digital: https://dl.acm.org/
- IEEE Xplore Digital Library: https://ieeexplore.ieee.org/
- Springer Link: https://link.springer.com/
- Google scholar: https://scholar.google.com/

3.4. Search term

A good search terms is important to have figured out so that it can get articles based on the research questions. My search terms tried to get most out of the research questions without excluding each other.
(Neural network AND Random Forest classifier AND MLP classifier) OR Random Forest Classifier OR MLP Classifier

3.5. Search procedure

All databases used the search terms. This section will give information about what kind of filters I applied when searching using the filter function.

Science direct: I went for only Research articles for the article type, Computer science for subject area.
IEE: I applied filters to only look for journals and keywords learning (artificial intelligence) and neural nets. Sorted by most cited.
Springer Link: I used the filter to only look for English articles containing artificial intelligence and computer science.
ACM Digital: Filters to see only research articles in pdf form.
Google scholar: Sorted by relevance

4. Study selection

When choosing the studies, I looked for
- Studies that used Random Forest classifier
- Studies that used MLP classifier
- Open access

Excluded articles
- No information about classifiers
- Articles that were in other languages than English

4.1. Data extraction

The technique I mostly used to find the most optimal studies were as followed:

- The title of the paper
- Main topic of the paper
- The abstract of the paper
- Passing the study selection requirements

4.2. Execution of the systematic literature review

Database	Articles where I applied the search terms	Articles that passed the data extraction	Articles used
Science Direct	3431	9	4
IEEE Xplore	451	9	7
Springer link	4593	5	5
ACM Digital	200k +	3	0
Google scholar	100k +	1	1
Total		27	17

Table 1. Articles that were found, checked and used

5. Results and findings

5.1. Results

Research questions	Publications
RQ1a	[2] [3] [4] [9] [12] [15] [18]
RQ1b	[5] [6] [10] [13] [14] [15] [16] [17]
RQ2	[7] [8] [11] [15]

Table 2. Overview of Articles that contained research questions.

This section contains an overview of the references which answers the research questions.

5.2. Findings

5.2.1.What are MLP classifiers, and how do they work?

Multilayer perceptron's are powerful classifiers that contains several free parameters [2, 18]. Even though they may provide superior performance, they need Cross validation which can be time-consuming and biased [2, 18]. Slow convergence and lack of guarantee of global minima are further drawbacks [2]. Still, they are commonly used feed forward network as classifiers for pattern classification approaches [3]. The parameters are number of hidden layers, number of hidden nodes, activation function, solver function and learning rate [4,15]. The accuracy of the MLP will then depend on the selection of these parameters which can require much effort in gaining the best architecture [4, 15]. The way that the classifier is set up is that is has three types of layers [10]. The first layer is called the input layer and is where the model receives the data [10]. Next layer is called the hidden layer and consists of neurons that map the data with mathematical functions [10]. The MLP can have more than one hidden layer [10]. The last layer is called the output layer which receives the data computed by the hidden layer and returns a result [10]. Each neuron has a weight attached which its purpose is for strengthening the linkage [10]. The neurons that are in the hidden layer and output layer also has a bias that works like a threshold for the activation of the neuron [10].

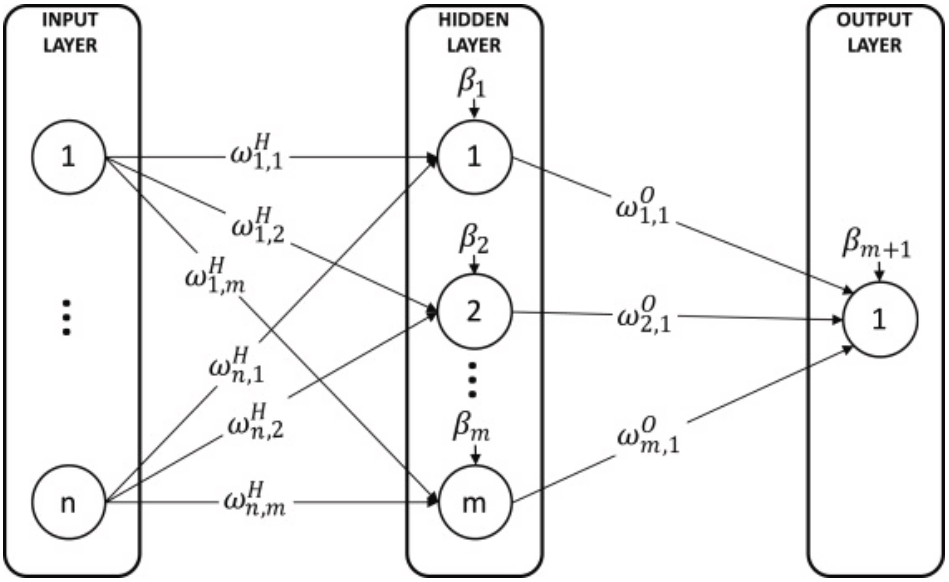

Fig 1: Example of a structure of the MLP classifier

To achieve an acceptable output value, the MLP classifier is trained with a method called Back Propagation [10]. Back Propagation adjusts the weights and biases based on the output value the model gets, until an acceptable error is reached [10]. Another technique you can use to improve the model, is by preventing overtraining [12]. Early stopping is one of these methods, and it works so that the training will stop when the generalization error begins to increase [12].

5.2.2.What are Random Forest classifiers, and how do they work?

The Random Forest classifiers is of the most popular machine learning techniques [6]. It is created by Breiman and presents many advantages for its application in remote sensing [17]. It performs several tasks which rises the accuracy of training and testing speed [5]. The classifier generates forests [9]. These forests consist of a huge number of unique trees [5,16]. These trees are known as decision trees, and each tree is divided into a class prediction [5]. The class with the most votes become the model's prediction [5, 16, 17]. The design of a decision tree requires the choice of an attribute selection measure and a pruning method [16, 17]. The most frequently used attribute selection is the information Gain Ratio criterion, Gini index and Chi-square [16, 17]. The classifier can be used by splitting a dataset for training, and one for testing [6]. The training set is used to train the classifier, while the testing set is used to simulate the model in terms of prediction [6]. The RF classifier only needs the definition of two parameters for generating a prediction model: the number of trees desired, and the number of prediction variables [17]. When increasing the number of trees, the generalization error always converges,

and over-training is not a problem [17]. If you would decrease the number of predicted variables, it would cause each individual tree of the model to be less strong but also have less correlation with each other which makes the model more accurate [17]. Further hyperparameter tuning is used to improve the model [6]. To reduce the possibility of over-fitting the training data, you can restrict the depth of any single trees and have a maximum number of trees. [14]. This is done if the data you are working with is small [14]. The model can either utilize bootstrap aggregating or bagging techniques to learn decision trees [13, 15]. Bootstrap aggregating is a technique used for creation of training data by randomly resampling the original dataset with replacement [17]. The technique offers a strong potential in medical domains [14].

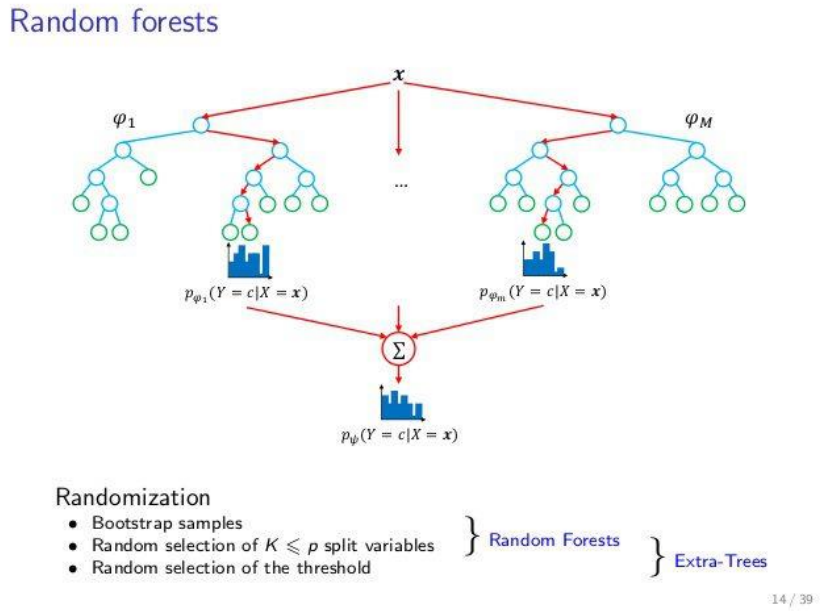

Fig 2: Example of a random forest structure

5.2.3. What's the difference in performance when solving a classification problem using MLP classifier and Random Forest Classifier?

A study done by A. Nasim, D. C. Nchekwube, F. Munir and Y. S. Kim about Arrhythmia Classification with Optimum Features Using Single-Lead Electrocardiogram used an optimization called MOPOS on MLP, k-nearest neighbor, support vector machine, Random Forest, and extra decision tree. This resulted in that the MLP consistently preformed the best with significantly reduced number of features for the targeted 15-class classification problem. The MLP classifier used optimizer Adam with a learning rate of 0.0001, activation function ReLU, Input size of 253, Output size of 15, and hidden layer

sizes of 220, 180, 120 and 60. The random forest model consisted of 25 trees with an Optimum split criterion of entropy [7].

Rachael Hagan, Charles J. Gillan and Fiona Mallett did a study about comparing machine learning methods for classification of cardiovascular disease. This study used two public datasets, UCI arrythmia dataset and Kaggle cardiovascular dataset, with significantly different characteristics to assess the potential differences in the uncertainty of the methods. For this study, the random forest classifier uses random subset of data and adds random selection of the available features. It had Gini criterion, 100 estimators and a max depth of 20 for the UCI dataset and Gini criterion, 100 estimators and max depth 10 for the Kaggle dataset. The MLP classifier model consisted of a tenfold cross validation for hyperparameter search. The model had an input layer of 15 nodes with 15 output nodes, a hidden layer with 15 input nodes and 50 output nodes, a drop out of 50 input nodes and 50 output nodes, another hidden layer with 50 input nodes and 25 output nodes, and an output layer of 25 input nodes and 2 output nodes. They got the best accuracy by using batch size 32 with 400 epochs, and Nadam optimizer function. The MLP classifier resulted in an accuracy of 0.74 for the UCI arrythmia dataset and 0.705 for the Kaggle cardiovascular dataset. The random forest classifier resulted in an accuracy of 0.95 for the UCI arrythmia dataset and an accuracy of 0.74 for the Kaggle cardiovascular dataset. This resulted in that the random forest classifier preformed the best [8].

Ashfaq Ahmad Najar & S. Manohar Naik did a study that was about detecting DDoS attack using MLP and Random Forest Algorithms. The random forest model was used for binary classification. The model checked if either a packet was normal or an attack packet, and got an accuracy of 0.9913 on train data, 0.9913 on validation data and an accuracy of 0.97 on test data with an F1 score of 0.9661. The MLP model was used to detect what type of attack it was, and got an accuracy of .9796 on train, 0.9853 on validation data and 0.74 on the test data [11].

A study done by A. J. Lado et al were comparing Neural Network and Random Forest Classifier on Dragon Fruit Disease. This study was an image classification problem and contained 41 images of healthy and sick fruit and leaf which was divided into four classes. The models used a 10-fold cross validation, and both obtained an accuracy in range of 70 to 82 % [15].

References

[1] fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [Date Retrieved] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

[2] T. Windeatt, "Accuracy/Diversity and Ensemble MLP Classifier Design," in *IEEE Transactions on Neural Networks*, vol. 17, no. 5, pp. 1194-1211, Sept. 2006,https://doi.org/10.1109/TNN.2006.875979.

[3] X. Zhai, A. A. S. Ali, A. Amira and F. Bensaali, "MLP Neural Network Based Gas Classification System on Zynq SoC," in *IEEE Access*, vol. 4, pp. 8138-8146, 2016,https://doi.org/10.1109/ACCESS.2016.2619181.

[4] M. Roy, D. Routaray, S. Ghosh and A. Ghosh, "Ensemble of Multilayer Perceptrons for Change Detection in Remotely Sensed Images," in *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 1, pp. 49-53, Jan. 2014, https://doi.org/10.1109/LGRS.2013.2245855.

[5] V. Geetha, A. Punitha, M. Abarna, M. Akshaya, S. Illakiya and A. P. Janani, "An Effective Crop Prediction Using Random Forest Algorithm," *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, 2020, pp. 1-5,https://doi.org/10.1109/ICSCAN49426.2020.9262311.

[6] N. A. Maung Maung, B. Y. Lwi and S. Thida, "An Enhanced RSS Fingerprinting-based Wireless Indoor Positioning using Random Forest Classifier," *2020 International Conference on Advanced Information Technologies (ICAIT)*, 2020, pp. 59-63,https://doi.org/10.1109/ICAIT51105.2020.9261776.

[7] A. Nasim, D. C. Nchekwube, F. Munir and Y. S. Kim, "An Evolutionary-Neural Mechanism for Arrhythmia Classification With Optimum Features Using Single-Lead Electrocardiogram," in *IEEE Access*, vol. 10, pp. 99050-99065, 2022, https://doi.org/10.1109/ACCESS.2022.3203586.

[8] Rachael Hagan, Charles J. Gillan, Fiona Mallett,

Comparison of machine learning methods for the classification of cardiovascular disease,

Informatics in Medicine Unlocked,Volume 24,2021,100606,ISSN 2352-9148, https://doi.org/10.1016/j.imu.2021.100606.

[9] Damodar Reddy Edla, Kunal Mangalorekar, Gauri Dhavalikar, Shubham Dodia,

Classification of EEG data for human mental state analysis using Random Forest Classifier,

Procedia Computer Science,Volume 132,2018,Pages 1523-1532,ISSN 1877-0509,

https://doi.org/10.1016/j.procs.2018.05.116.

[10] Matías Gabriel Rojas, Ana Carolina Olivera, Pablo Javier Vidal,

Optimising Multilayer Perceptron weights and biases through a Cellular Genetic Algorithm for medical data classification,Array,Volume 14,2022,100173,ISSN 2590-0056,

https://doi.org/10.1016/j.array.2022.100173.

[11] Najar, A.A., Manohar Naik, S. DDoS attack detection using MLP and Random Forest Algorithms. *Int. j. inf. tecnol.* **14**, 2317–2327 (2022).

https://doi.org/10.1007/s41870-022-01003-x

[12] Medeiros, C.M.S., Barreto, G.A. A novel weight pruning method for MLP classifiers based on the MAXCORE principle. *Neural Comput & Applic* **22**, 71–84 (2013). https://doi.org/10.1007/s00521-011-0748-6

[13] Khorshidpour, Z., Hashemi, S. & Hamzeh, A. Evaluation of random forest classifier in security domain. *Appl Intell* **47**, 558–569 (2017). https://doi.org/10.1007/s10489-017-0907-2

[14] Klement, W., Gilbert, S., Resende, V.F. *et al.* The validation of chest tube management after lung resection surgery using a random forest classifier. *Int J Data Sci Anal* **13**, 251–263 (2022). https://doi.org/10.1007/s41060-021-00296-8

[15] A. J. Lado et al., "Comparison of Neural Network and Random Forest Classifier Performance on Dragon Fruit Disease," 2021 International Electronics Symposium (IES), 2021, pp. 287-291, https://doi.org/10.1109/IES53407.2021.9593992

[16] M. Pal (2005) Random forest classifier for remote sensing classification, International Journal of Remote Sensing, 26:1, 217-222,https://doi.org/10.1080/01431160412331269698

[17] V.F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, J.P. Rigol-Sanchez, An assessment of the effectiveness of a random forest classifier for land-cover classification, ISPRS Journal of Photogrammetry and Remote Sensing,Volume 67,2012,Pages 93-104,ISSN 0924-2716, https://doi.org/10.1016/j.isprsjprs.2011.11.002. (https://www.sciencedirect.com/science/article/pii/S0924271611001304)

[18] Windeatt, T. (2008). Ensemble MLP Classifier Design. In: Jain, L.C., Sato-Ilic, M., Virvou, M., Tsihrintzis, G.A., Balas, V.E., Abeynayake, C. (eds) Computational Intelligence Paradigms. Studies in Computational Intelligence, vol 137. Springer, Berlin, Heidelberg.

https://doi.org/10.1007/978-3-540-79474-5_6

Figures

[1] Example of structure of a Multilayer Perceptron (MLP)

https://doi.org/10.1016/j.array.2022.100173

[2] Rishabh Mall, Jan 7, 2019, rand-forest-2.jpg

https://medium.com/@mallrishabh52/random-forest-67afc2ff884f

https://www.kdnuggets.com/wp-content/uploads/rand-forest-2.jpg