2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC)

Comparison of XGBoost and the Neural Network model on the class-balanced datasets

Jincenzi Wu[1, *, †]
[1]Tandon School of Engineering
New York University
New York, USA
[*]Corresponding author's e-mail:
jw6217@nyu.edu

Yuanyuan Li[2, †]
[2]Department of Computer and Mathematical Sciences
University of Toronto Scarborough
Toronto, Canada
tuanzi.li@mail.utoronto.ca

Yizhou Ma[3, †]
[3]Jinan-Birmingham Joint Institute
Jinan University
Guangzhou, China 1806240115@qq.com
[†]These authors contributed equally.

***Abstract*—The Extreme Gradient Boosting method and Deep learning methods are classical machine learning methods widely used in many fields. However, the advantages and disadvantages of these two methods have been a long-debated topic. This paper evaluates the classification performance of the XGBoost and Multiple-Layer Perceptron Neural Network on the structured data. We compare the classification performance of both methods, using the large scale, public datasets, and show the overall trend with the different percentages of datasets used. The experiment validated the higher accuracy in all datasets obtained through the XGBoost method. We concluded that the XGBoost method overcomes the complex data distribution with the feature space to better classify performance on the structured data.**

Keywords - Extreme Gradient Boosting, Neural Network, binary classification

I. INTRODUCTION

Nowadays, deep learning empowered some stunning developments in many fields such as computer vision, speech recognition, medical image analysis, electronic sports, and other fields. However, the problems of interpretability and robustness of deep learning keep puzzling many researchers [1]. For example, the lack of metrics for interpretation methods makes it impossible to measure the interpretation results of models in a more rigorous way [2]. Secondly, existing interpretation methods are often targeted at a single model. The effectiveness of model-independent interpretation methods still needs to be further improved [3]. Thirdly, the effectiveness of existing deep learning methods depends on the high-quality requirements of the training datasets [4]. When the training datasets present problems such as significantly complex noise, abnormal point invasion, and class imbalance, its high effectiveness is often not guaranteed [5].

Moreover, due to the lack of theoretical basis, tuning hyper-parameters and network designing are considerable challenges for deep learning [6]. Moreover, Deep learning algorithms rely heavily on high-performance machines because the deep learning algorithm includes the GPU, an integral part of its work. Because deep learning performs genetic operations through massive matrix multiplication, these operations can be efficiently optimized using a GPU explicitly built for this purpose. To achieve high performance, deep learning algorithms often require extensive datasets. However, for many applications, such large datasets are not readily available. While in low-dimensional discrete datasets, deep learning algorithms often do not significantly advantage against traditional machine learning algorithms.

Compared to deep learning methods, traditional machine learning methods usually have better interpretability [7]. It is easier to tune the hyper-parameters and change the model design because we completely understand both the datasets and the underlying algorithms. Furthermore, traditional machine learning algorithms can run on low-performance machines. Because they are not computationally expensive, they can iterate faster and try many different technologies in a shorter period.

This paper explores the difference in the performance of deep learning methods and traditional machine learning algorithms on discrete binary classification datasets. This paper compares accuracy by applying five different discrete binary classification datasets between Extreme Gradient Boosting (XGBoost) and neural networks. Moreover, K-Nearest Neighbor (KNN) [8] and Principal Component Analysis (PCA) [9] are used as the baseline. The result shows that Neural Network does not give a promising result than traditional machine learning models XGBoost [10] in these five datasets. With the increase of the samples overlapped, the neural network model tends to have worse performance. This kind of sample distribution harms XGBoost performance as well. However, XGBoost performed significantly better than Neural Network in one of the datasets. The conclusion is that XGBoost creates the feature space with all its weak learners based on the features, which helps classify the samples well. At the same time, it is hard for the Neural Network to narrow out suitable feature space when the space in high dimension is created to classify the samples with many layers and neurons built.

II. METHODOLOGY

In this section, the methods in the experimentation adopted to do the analysis are described. The experiment was carried out on five public datasets, and four algorithms are enlisted in the task.

A. Dataset used in the experimentation

Cardiovascular Disease dataset --- A balanced binary classification dataset contains 70 thousand samples and 11 features. The objective is to predict whether a person has cardiovascular disease [11].

978-1-6654-0605-5/21/$31.00 ©2021 IEEE

2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC) | 978-1-6654-0605-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICFTIC54370.2021.9647373

Heart Disease Dataset --- A balanced binary classification dataset contains about one thousand samples and 11 features with the objective is to predict whether a person has heart disease [12].

Airline Passenger Satisfaction --- A balanced binary classification dataset contains about 100 thousand samples and 24 features with the objective is to predict whether a customer satisfies the airline service [13].

Health Insurance Cross-Sell Prediction --- A imbalanced binary classification dataset contains about 380 thousand samples and 10 features with the objective is to predict whether the customer would be interested in vehicle insurance [14].

L&T Vehicle Loan Default Prediction --- A balanced binary classification dataset contains about 100 thousand samples and 34 features with the objective is to predict whether a client can be offered a loan [15].

B. Models used for experimentation

The research figures out the classification performance of the machine learning algorithm and deep learning algorithm on the binary datasets with various statistical factors included. Extreme Gradient Boosting works as the training algorithm to classify the samples based on a series of features. A neural network has been set to recognize the relationship in the discrete data to classify all changing inputs adopted. The K-Nearest Neighbors algorithm works as a baseline classification algorithm and figures out the sample distribution to highlight the learning process of those algorithms. Moreover, Principal Component Analysis projects the samples to 2D dimensionality to show the statistical information of the datasets.

As one of the most popular machine learning classification methods, the K-Nearest Neighbors (KNN) is straightforward [16]. With the value of the k nearest neighbors assigned, the class of the predicted sample is determined by the k closest training samples. The distance between the predicted and training samples is used to obtain the k nearest neighbors and conduct the predicted sample class. There is no training stage of the K-Nearest Neighbors method. So that KNN is used as a baseline method in the research. Compared with the KNN, the performance of the other classifications methods shows how the model learns from the features and how feature spaces affect classification performances.

Principal Component Analysis (PCA) is one of the most widely used methods to interpret a large dataset. It reduces the dimensionality of a dataset while preserving statistical information as possible [17]. Several datasets used in the research have more than 100 thousand samples. PCA is used to consider the dataset in two dimensions, and all samples are plot in the plane to illustrate the similar or different statistical information within datasets.

Extreme Gradient Boosting is an efficient model based on the gradient boosting decision tree [18]. An ensemble of trees is built based on the features and summed sequentially to predict the sample class. Each tree tries to recover the difference between the target and prediction predicted by the prior ensemble of trees. As one of the popular classification models in machine learning, XGBoost does binary classification on five datasets in the research. Furthermore, it works as the training algorithm and shows the classification performance of the machine learning algorithm on the discrete binary datasets.

Multiple-Layer Perceptron Neural Network is one of the most significant models in the artificial neural network [19]. In our research, the Multiple-Layer Perceptron Neural Network model consists of one input layer, three hidden layers, and one output layer.

TABLE I. BASIC INFORMATION OF NEURAL NETWORK

Layer	Neurons #
Input	64
Dense 1	32
Dense 2	16
Dense 3	8
Dense 4	2

III. EXPERIMENT

A. Dataset Summary

The experiment enlists five distributive binary classification datasets from the Kaggle. After balancing datasets, each dataset's basic information (instance number, feature number, and target number) is summarized in TABLE II. 25% of each dataset was used for testing. Also, to visualize data distribution, the experiment used the PCA method to reduce the dimensionality of each dataset to a 2D layer and recorded the f1-score and accuracy of KNN on testing data for them, as shown in TABLE III.

TABLE II.BASIC INFORMATION OF THE DATASET

Dataset #	Instance #	Feature #	Target #
Dataset 1	50557	12	1
Dataset 2	1190	11	1
Dataset 3	103594	22	1
Dataset 4	67532	10	1
Dataset 5	98000	34	1

TABLE III. F1-SCORE AND ACCURACY ON TESTING DATA FOR KNN

	D1	D2	D3	D4	D5
0	0.76	0.75	0.59	0.62	0.57
1	0.65	0.79	0.75	0.57	0.56
Accuracy	0.71	0.77	0.69	0.60	0.57

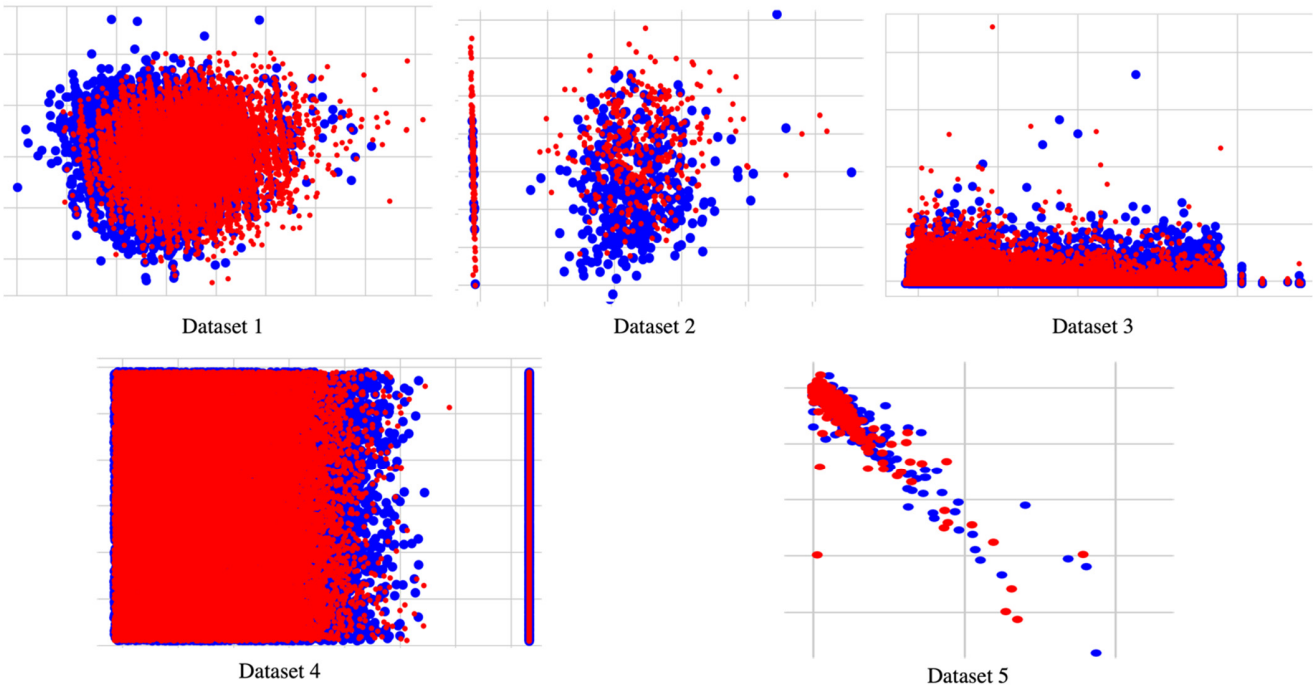

Figure 1. 2D-PCA

TABLE IV. F1-SCORE ON TESTING DATA FOR XBG AND NN

	D1		D2		D3		D4		D5	
	XGB	**NN**	**XGB**	**NN**	**XGB**	**NN**	**XGB**	**NN**	**XGB**	**NN**
0	0.78	0.77	0.86	0.73	0.93	0.55	0.77	0.60	0.61	0.27
1	0.68	0.66	0.88	0.77	0.95	0.74	0.82	0.70	0.61	0.64

In general, for all five datasets, the PCA pictures show that the data from both classes largely overlap with each other in a particular area. Since the PCA on the 2D layer shows a high overlap for every dataset, it is safe to conclude that both classes' data likely distribute evenly (at least in a particular space) in the hyper dimension. In other words, the data quality of these five datasets is not perfect. The KNN result in Table III also proved this conclusion.

Since KNN takes the K-th nearest data points and chooses the class of the majority group, it will reflect the data distribution status in hyperspace. For all five datasets, the KNN performance on Dataset 1, 2, and 3 is better than on Dataset 4 and 5. Therefore, Dataset 1,2, and 3 have a more characterized data distribution than Dataset 4 and 5.

B. XGBoost and NN Default Setting Analysis

The experiment used the default setting of the XGBoost model in the sklearn ensumble package and the fully connected neural network described in the methodology section. Then we recorded the performance reports on testing data for each dataset respectively in TABLE IV. For choosing epoch number of NN, all the datasets tend to converge after 30 – 45 epochs; selecting a number within that range is reasonable.

Comparing with the KNN, both XGBoost and NN learned character distribution and improved the prediction accuracy to a certain degree on Dataset 1, 2, and 4. The F1-score of the XGBoost is higher than the F1-score of the NN, around 15% - 20% on Dataset 2 and 4, which indicates that the XGBoost performs better than the NN on those datasets. Meanwhile, only the XGBoost increases the accuracy on Dataset 3 and 5, while the neural network's performance is even worse than the result of the baseline model.

This experiment result was explained by the difference in choosing feature subspace for the XGBoost and the NN.

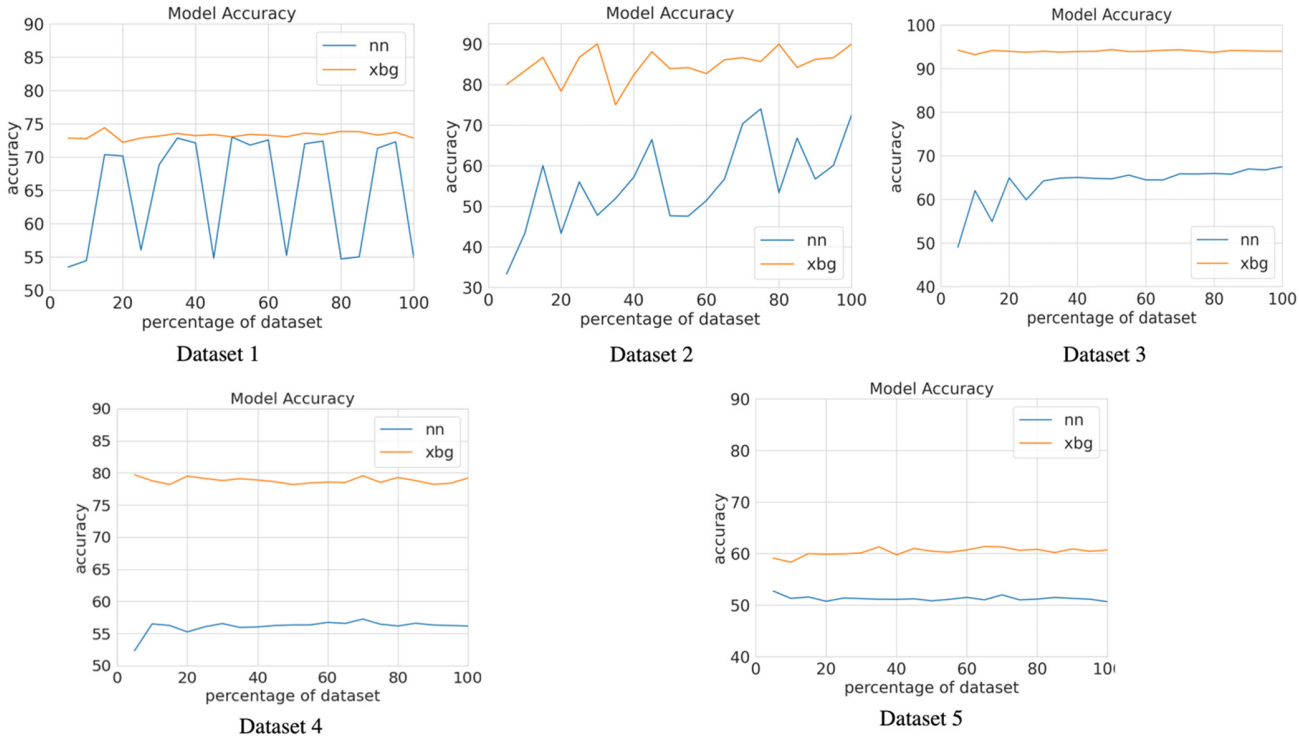

Figure 2. Accuracy in the different percentage

For the XGBoost, due to the algorithm design and the tree-structured weak learner, the choice of a feature subspace is restricted by the dataset's features, which helps produce a high-quality classification. However, for the NN, because the layer construction may expand potential feature subspaces, the choice of a feature subspace may end up in a cartesian product of the number of nodes and the number of layers. This assumption needs to be further discussed.

C. *Comparison of model performance with different percentages of the dataset used*

When taking a certain percentage amount of data to form a whole dataset in increasing order, the experiment recorded the performance of the XGBoost and the NN for all the datasets. The XGBoost has a better actuary prediction than the NN for all five datasets, as shown in Figure 2.

The comparison is made among Dataset 1, 3, 4 and 5, because of the relatively small instance number of Dataset 2. Regardless of dataset 2, the XGBoost accuracy results are stable (fluctuation range is within 3%) and do not show an increasing tendency to increase the amount of data. On Dataset 3, the accuracy of NN positively correlates with the amount of data; however, the NN result remains stable on Dataset 4 and Dataset 5. Moreover, on Dataset 1, the NN's accuracy fluctuates strongly when the percentage goes up.

Combining with the PCA picture from the Dataset Summary section implies that there is no direct relationship between the learning model accuracy and the increasing amount of data when the data quality is not ideal.

IV. Conclusion

This research has made various comparisons between the XGBoost model and the fully connected neural network model for five public binary classification datasets. As one of the ensemble classifiers, XGBoost increases its confidence in classification gradually. It has a greater classification performance than the neural network model commonly. A neural network cannot extract features from the data and give a proper performance based on the relationship of the discrete samples efficiently. A more significant percentage of the samples is used to train the model; the neural network model shows only about 50% accuracy on Dataset 5.

Based on the KNN classification performance and two-dimensional PCA figures, the samples from both classes in five datasets overlap each other by varying degrees. With an increasing number of the samples overlapped, the neural network model has worse performance. This kind of sample distribution has an adverse impact on the XGBoost performance as well. However, XGBoost shows better than 90% accuracy on Dataset 3 which almost all samples are overlapped. We conclude that the ensemble classification algorithm XGBoost creates the feature space with all its weak learners based on the features, which helps classify the samples well. For the neural network model, it is hard for the Neural Network to narrow out suitable feature space when the space in high dimensionality is created to classify the samples with many layers and neurons built. We suppose that the space created by the neural network could be narrowed based on the current feature information so that the neural network could create the feature space more effectively and have a better classification performance.

REFERENCES

[1] Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

[2] Sato, M. and Tsukimoto, H., 2001, July. Rule extraction from neural networks via decision tree induction. In IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222) (Vol. 3, pp. 1870-1875). IEEE.

[3] Zilke, J.R., Mencía, E.L. and Janssen, F., 2016, October. Deepred–rule extraction from deep neural networks. In International Conference on Discovery Science (pp. 457-473). Springer, Cham.

[4] Schmitz, G.P., Aldrich, C. and Gouws, F.S., 1999. ANN-DT: an algorithm for extraction of decision trees from artificial neural networks. IEEE Transactions on Neural Networks, 10(6), pp.1392-1401.

[5] Gallant, S.I., 1988. Connectionist expert systems. Communications of the ACM, 31(2), pp.152-169.

[6] Tsukimoto, H., 2000. Extracting rules from trained neural networks. IEEE Transactions on Neural networks, 11(2), pp.377-389.

[7] D.V. Carvalho, E.M. Pereira and J.S. Cardoso, 2019. Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), p.832.

[8] Cunningham, P. and Delany, S.J., 2020. k-Nearest neighbour classifiers: (with Python examples). arXiv preprint arXiv:2004.04523.

[9] Jolliffe, I.T. and Cadima, J., 2016. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), p.20150202.

[10] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y. and Cho, H., 2015. Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), pp.1-4.

[11] Svetlana U. 2018. *Cardiovascular Disease Dataset*. Kaggle. [Online]. [Accessed May 2021]. Available from: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

[12] Manu S. 2019. *Heart Disease Dataset.* Kaggle. [Online]. [Accessed May 2021]. Available from: https://www.kaggle.com/sid321axn/heart-statlog-cleveland-hungary-final

[13] TJ K. 2020. *Airline Passenger Satisfaction.* Kaggle. [Online]. [Accessed May 2021]. Available from: https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction

[14] Mamta D. 2019. *L&T Vehicle Loan Default Prediction.* Kaggle. [Online]. [Accessed May 2021]. Available from: https://www.kaggle.com/mamtadhaker/lt-vehicle-loan-default-prediction

[15] Anmol K. 2020. *Health Insurance Cross Sell Prediction.* Kaggle. [Online]. [Accessed May 2021]. Available from: https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction?select=test.csv

[16] Jiang, S., Pang, G., Wu, M. and Kuang, L., 2012. An improved K-nearest-neighbor algorithm for text categorization. Expert Systems with Applications, 39(1), pp.1503-1509.

[17] Abdi, H. and Williams, L.J., 2010. Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), pp.433-459.

[18] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30, pp.3146-3154.

[19] Li, Y. and Liang, Y., 2018. Learning overparameterized neural networks via stochastic gradient descent on structured data. arXiv preprint arXiv:1808.01204.