ISBN- 978-978-57388-7-2

Trinity University, Nigeria
Society for Multidisciplinary & Advanced Research Techniques (SMART) Africa
IEEE Computer Society, Nigeria
The ICT University Foundations, USA.

LASUSTECH Multidisciplinary Innovations Conference (LASUSTECH-MIC)

16th – 18th April, 2022

Combination of Bagging and Boosting Ensemble Learning Methods: A Solution to Model Underfitting and Overfitting

Ezeora, Nnamdi Johnson, Abhadiomhen, Stanley Ebhohimhen, Nzeh, Royransom, Uzo, Blessing Chimezie, Anichebe, Gregory E., Asogwa, Caroline, Uzo, Izuchukwu & Ogbene, Nnaemeka E
Department of Computer Science
University of Nigeria
Nsukka, Nigeria

E-mails: nnamdi.ezeora@unn.edu.ng, stanley.abhadiomhen@unn.edu.ng, royransom.nzeh@unn.edu.ng, blessing.uzo@unn.edu.ng, gregory.anichebe@unn.edu.ng, caroline.asogwa@unn.edu.ng, izuchukwu.uzo@unn.edu.ng, nnaemeka.ogbene@unn.edu.ng

Phone No: +2349081000035

Proceedings Citation Format

Ezeora, N.J., Abhadiomhen, S.E, Nzeh, R., Uzo, B.C., Anichebe, G.E., Asogwa, C., Uzo, I., & Ogbene, N.E. (2022) Combination of Bagging and Boosting Ensemble Learning Methods: A Solution to Model Underfitting and Overfitting. Proceedings of the LASUSTECH 30th iSTEAMS Multidisciplinary Innovations Conference. Lagos State University of Science & Technology, Ikorodu, Lagos State, Nigeria May 2022. Pp 149-156. www.isteams.net/lasustech2022
DOI: https://doi.org/10.22624/AIMS/iSTEAMS/LASUSTECH2022V30P13

ISBN- 978-978-57388-7-2

Combination of Bagging and Boosting Ensemble Learning Methods: A Solution to Model Underfitting and Overfitting

Ezeora, Nnamdi Johnson, Abhadiomhen, Stanley Ebhohimhen, Nzeh, Royransom, Uzo, Blessing Chimezie, Anichebe, Gregory E., Asogwa, Caroline, Uzo, Izuchukwu & Ogbene, Nnaemeka E
Department of Computer Science
University of Nigeria
Nsukka, Nigeria

ABSTRACT

Training a model to avoid overfitting and underfitting is a top priority in machine learning. Hence, several efforts have been made in the last years to improve generalization performance in single learning settings. However, ensemble methods such as Decorate, Rotation Forest, Bagging, and Boosting have outperformed single learning methods. Besides, Bagging and Boosting methods proved stronger for overfitting and underfitting problem. In this paper, we propose a novel method by combining Bagging and Boosting ensemble methods to jointly overcome limitations of model overfitting and underfitting more effectively.

Keywords: Underfitting, overfitting, ensemble learning, bagging, boosting.

1. INTRODUCTION

Over the years, machine learning has grown as a subfield of artificial intelligence with several learning paradigms such as supervised, unsupervised, and reinforcement learning. Regardless of the learning method employed to solve machine learning problems, the level of accuracy of a learned model depends largely on how well it fits the datasets [1]. Besides, the dataset's quality can influence the accuracy of a model. Hence, there are continuous efforts to develop robust models that avoid overfitting and underfitting the datasets. However, the challenge of implementing relatively good models in a single learning setting is limited since a thoroughly developed model may still not be immune to overfitting and underfitting. The model may fit the dataset employed to train and validate but fails to generalize to an unseen dataset. Perhaps during training, it is important to learn in real-time when a model is performing poorly [2].

In that way, the time complexity of training and retraining machine learning models can be reduced. To this end, the study in [3] suggested a real-time method that utilized a diagnostic tool to provide insight into what is working or not with a learned model. But then again, a good knowledge of model overfitting or underfitting does not automatically eradicate the issue it presents. Although the diagnostic tool suggested above may provide insight and a practical guide on improving the model's performance, eradicating overfitting and underfitting problems requires robust techniques. This has been the focus of several pieces of research in machine learning. Recent studies in [4],[5],[6] gave reasonable proved that ensemble learning methods such as decorate [7], bagging [8], boosting [6], and random subspace [9] can surpass the performance of single learning models to improve generalization performance.

ISBN- 978-978-57388-7-2

Therefore, Ensemble learning aims to improve the generalization performance of single-task learning models such that multiple models are trained on a given dataset to obtain a unified model parameter. For example, in Heterogeneous ensemble learning, each model is created using a different algorithm on the same dataset. Illustratively, for a classification problem, one classifier can be trained using an Artificial neural network (ANN) and another using a Bayes network while equally using the Support vector machine (SVM) to train another base classifier. Then, the final model becomes an aggregate of several base classifiers, which may be done through stacking, cascading, simple majority voting averaging, and weighted averaging. Similarly, in Homogeneous ensemble learning, such as Bagging and Boosting methods, each base model is trained using the same algorithm on a different subset of the data.

More precisely, Bagging applies a combination of bootstrap and aggregation to train each base classifier in parallel. Whereas, for Boosting methods such as AdaBoost and Gradient Boosting, each base model is trained sequentially and the subsequent models are focused on the error made by the previous one. Even so, Bagging and Boosting methods are robust in solving overfitting and underfitting problems. For this reason, in this paper, a method for jointly overcoming overfitting and underfitting problems is proposed by combining Bagging and Boosting ensemble methods. First, we provide a brief review of related works in section 2. Then, we formulate the proposed model in section 3. Next, we discuss the benefit of combining Bagging and Boosting methods in section 4. Finally, section 5 presents the conclusion of this paper.

2. RELATED WORKS

Although several efforts are being made to improve information discovery using traditional machine learning methods, certain performance problems may arise in complex machine learning systems due to large dimensional or noisy datasets utilized to train them. These large dimensional datasets contain several sets of features that may be noisy and may cause the model to perform woefully in extreme extrapolation. The above could be attributed to the lapses in the traditional machine learning method occasioned by its inability to capture complementary knowledge in the original data structure. As a result, it is necessary to provide a method that enhances knowledge discovery through complementary efforts. One research area that has continued to receive immense interest is ensemble learning. It provides an opportunity to explore complementary efforts required to improve the performance accuracy of machine learning models.

The authors in [10] asserted that ensemble learning had become a trending research interest in artificial intelligence. Hence, several works on ensemble learning, especially those related to Bagging and Boosting methods, exist. In [11], Boosting, Bagging, and traditional dissolved gas analysis (DGA) were combined to interpret DGA. The performance of the proposed system was evaluated on models implemented either in their base form or through ensemble methods of bagging and boosting using four different machine learning algorithms: multilayer perceptron, Bayes network, J48 decision tree, and k-nearest neighbor. The experiment result showed that the combination methods of boosting and bagging performed better than the model in its based form. Also, an extensive algorithm evaluation demonstrated that the J48 algorithm had the best performance in most cases while the Bayes network algorithm and multilayer neural network algorithm trailed closely. Nevertheless, the robustness of their system may well not be tested considering that they had only trained their model using a dataset of only 347 samples, which may cause overfitting.

ISBN- 978-978-57388-7-2

Along the same line, [12] presented a system that enhances the prediction accuracy of a single model using bagging and boosting ensemble methods. However, their approach used an extended version of the online feature selection (OFS) technique to provide input to multiple classifiers. Unlike [11], they were able to evaluate the robustness of their proposed method using a combination of datasets collected from [13] and [14]. Their result disclosed that, in general, OFS based on boosting and OFS based bagging outperformed the single classifier. Similarly, [15] proposed a Chinese Dialogue intension system using a variant of boosting ensemble method named AdaBoost.

Additionally, studies by [4],[5],[6] observed that a single model with a relatively good overall accuracy for classification lesser than 80% might have an overall accuracy above 90% if Bagging and Boosting ensemble methods are used. Hence, they observed that an ensemble of classifiers could improve classification accuracy but may be computationally expensive. On the other hand, to reduce the computation complexity involved in training ensemble classifiers, Lipitakis and Kotsiantis [7] employed the feature selection method to remove redundant or noisy features before their ensemble was built. The proposed method combined several ensemble techniques in their base form, namely decorate, bagging, boosting, and random subspace, to prove the effectiveness of ensemble classifiers. Other efforts exist such as [16-19] based on low-rank representation. Nonetheless, to solve the issue posed by overfitting and underfitting, it is important to explore Bagging and Boosting methods in their combined form since they are considered stronger in solving overfitting and underfitting, respectively. Therefore, the CBB method proposed in this paper is built on the idea proposed in [7]. The following section describes the proposed method like [20].

Figure 1: Combining Bagging and Boosting ensemble methods

3. PROPOSED METHOD

This paper proposes a combined Bagging and Boosting (CBB) ensemble method (See figure 1 for illustration) as a joint solution to model the overfitting and underfitting problem. Boosting and Bagging methods are known to reduce underfitting and overfitting, respectively, and several researchers have explored their capabilities to improve the performance of learned models. For this reason, the proposed methodology combines both methods. First, to reduce the computational complexity of training base classifiers, the dataset is analyzed using dimensionality reduction techniques to remove redundant or noisy features.

ISBN- 978-978-57388-7-2

Then, bootstrap is applied to generate N training sets for the N training classifier. Besides, the parallel training method associated with bagging is maintained. However, to add the specialization effect of the boosting method, a single loop is introduced to allow the CBB procedure to be iterated to ensure that each classifier is focused on those training examples previously misclassified in subsequent iterations. The above is achieved by assigning each example in each N subset an equal initial weight that may change in the next iteration if an example is misclassified. For instance, if an example is misclassified in CBB round 1, the weight is increased to make it more likely to be classified correctly in the next iteration. Therefore, in the proposed method described in figure 2, the based classifiers are combined into a generalized model after CBB round 2 by averaging the probability of error of the individual classifier.

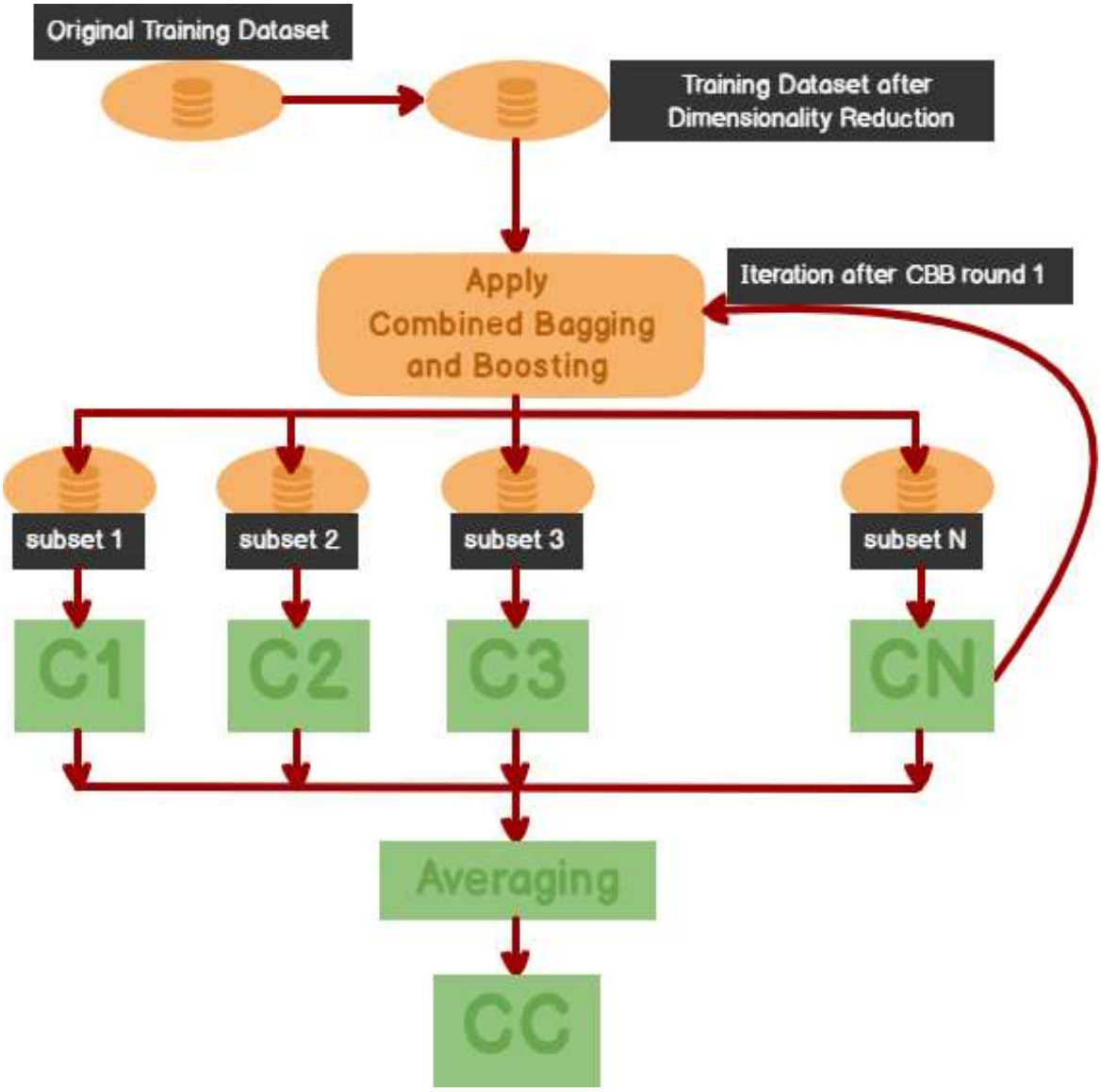

Figure 2: Proposed Method

ISBN- 978-978-57388-7-2

4. RESULTS

To demonstrate the effectiveness of the proposed method, it is important to illustrate the benefit of ensuring that the probability of error of each based classifier is lesser than 0.5 before their combination. First, we will assume that three binary classifiers were trained on three subsets of a dataset, and each has an error of 0.5. Secondly, we will assume that each error is independent, i.e., The probability that one event occurs does not affect the probability of the other event occurring. Thirdly, we will assume that averaging or simple voting combined these classifiers. Hence, the probability of error for an ensemble of the three classifiers is equal to the probability that two of the classifiers make an error plus the probability that all three classifiers make an error. We explain this below.

$$pe(ensemble) = classifier(3,2)\ pe\text{^}2(1-pe)\ +\ classifier(3,3)\ pe\text{^}3\text{s}$$
$$= 3 * (0.5)\text{^}2\ (0.5)\ +\ (0.5)\text{^}3$$
$$= 3/8\ +\ 1/8\ =\ 1/2\ =\ 0.5$$

Notice that each base classifier has an error that is not lesser than 0.5, so, an ensemble of the 3 classifiers did not provide better result than that of a single classifier.

On the other hand, what if each base classifier error is higher than 0.5? Let say each classifier has an error of 0.66.

$$pe(ensemble) = classfier(3,2)\ pe\text{^}2(1-pe)\ +\ classifier(3,3)\ pe\text{^}3$$
$$= 3 * (0.66)\text{^}2\ (0.34)\ +\ (0.66)\text{^}3$$
$$= 0.444\ +0.287\ =\ 0.731$$

The above illustrates the effectiveness of the proposed method and equally buttress the point made earlier that an ensemble of classifiers with each base classifier error lesser than 0.5 can improve performance of a model.

5. CONCLUSION

Ensemble models have been proved in several studies to outperform the single model. This paper proposes a method for solving overfitting and underfitting problems. Several factors were considered to make sure the proposed method is robust. However, to get better generalization performance, it is expected that the error of each base model should not be intensely related. Also, the error of the individual base model should be lesser than 0.5 (better than random guessing). In future research, we will further explore ways to ensure the error of the individual base model is much lesser than 0.5.

ISBN- 978-978-57388-7-2

REFERENCES

[1] N. Li, A. Martin and R. Estival, "Combination of Supervised Learning and Unsupervised Learning Based on Object Association for Land Cover Classification," in *2018 Digital Image Computing: Techniques and Applications (DICTA)*, 2018.

[2] A. Gavrilov, et al. "Convolutional Neural Networks: Estimating Relations in the Ising Model on Overfitting", in *2018 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, 2018.

[3] A. Ng, Machine Learning [Online]. Available at https://www.coursera.org/learn/machine-learning/home/welcome. Accessed on: Jan 1, 2020

[4] B. Wang and J. Pineau, "Online Bagging and Boosting for Imbalanced Data Streams*", IEEE Transactions on Knowledge and Data Engineering*, 2016, **28**(12): p. 3353-3366.

[5] H. R. Sanabila and W. Jatmiko, "Ensemble Learning on Large Scale Financial Imbalanced Data", in *2018 International Workshop on Big Data and Information Security (IWBIS)*,2018.

[6] S. Kulkarni and V. Kelkar, "Classification of multispectral satellite images using ensemble techniques of bagging, boosting and adaboost", in *2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)*, 2014.

[7] A. Lipitakis and S. Kotsiantis, "Combining ensembles algorithms of symbolic learners", in 2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA). 2015

[8] Y. Yu and H. Su, "Collaborative Representation Ensemble Using Bagging for Hyperspectral Image Classification", in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*. 2019.

[9] P.Shrivastava and M. Shukla,"Comparative analysis of bagging, stacking and random subspace algorithms". in *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*. 2015.

[10] X. Dong, et al., "A survey on ensemble learning. Frontiers of Computer Science", 2020. **14**(2): p. 241-258.

[11] M.E.A. Senoussaoui, M. Brahami and I. Fofana, "Combining and comparing various machine-learning algorithms to improve dissolved gas analysis interpretation", *IET Generation, Transmission & Distribution*, 2018. **12**(15): p. 3673-3679.

[12] G. Ditzler, et al., "Extensions to Online Feature Selection Using Bagging and Boosting*", IEEE Transactions on Neural Networks and Learning Systems*, 2018. **29**(9): p. 4504-4509.

[13] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?*", J. Mach. Learn*. Res., vol. 15, pp. 3133-3181, Jan. 2014.

[14] A. Frank, A. Asuncion, "UCI machine learning repository", 2010.

[15] M. Tu, B. Wang, and X. Zhao, "Chinese Dialogue Intention Classification Based on Multi-Model Ensemble", *IEEE Access*, 2019. **7**: p. 11630-11639.

[16] W. Gao, X. Li, S. Dai, X, Yin, and S.E. Abhadiomhen, "Recursive Sample Scaling Low-Rank Representation", Journal of Mathematics, 2021.

[17] K.F. Hui, X.J.Shen, S.E Abhadiomhen and Y.Z Zhan, "Robust low-rank representation via residual projection for image classification",Knowledge-Based Systems, 2022, 108230.

ISBN- 978-978-57388-7-2

[18] S.E. Abhadiomhen, R.C. Nzeh, E.D. Ganaa, H.C Nwagwu, G.E. Okereke, & S. Routray, "Supervised Shallow Multi-task Learning: Analysis of Methods", Neural Processing Letters, 2022, 1-18.

[19] H. Liang, H. Guan, S.E. Abhadiomhen, L. Yan, "Robust Spectral Clustering via Low-Rank Sample Representation", Applied Computational Intelligence and Soft Computing, vol. 2022, Article ID 7540956, 11 pages, 2022.

[20] S,E Abhadiomhen et al, "Design Of An Automated Home Security System With Remote Monitoring Capability", In Proceedings of the 28th SMART-iSTEAMS Interteriary Multidisciplinary Conference, The Gambia, 2021.