Knowledge Engineering and Data Science (KEDS)
Vol 5, No 2, December 2022, pp. 137–142

pISSN 2597-4602
eISSN 2597-4637

Performance of Ensemble Classification for Agricultural and Biological Science Journals with Scopus Index

Nastiti Susetyo Fanany Putri [a,1], Aji Prasetya Wibawa [a,2,*], Harits Ar Rosyid [a,3], Agung Bella Putra Utama [a,3], Wako Uriu [b,5]

[a] *Department of Electrical Engineering, Faculty of Engineering, Universitas Negeri Malang, Jl Semarang 5, Malang, East Java 65145, Indonesia*
[b] *Department of English, Chikushi Jogakuen University, 2-chōme-12-1 Ishizaka, Dazaifu, Fukuoka 818-0118, Japan*
[1] *nastiti.susetyo.2005348@students.um.ac.id;* [2] *aji.prasetya.ft@um.ac.id*;* [3] *harits.ar.ft@um.ac.id; agungbpu02@gmail.com* [4] *; ue2017119@chikushi-u.ac.jp* [5]
** corresponding author*

ARTICLE INFO

Article history:
Received 3 October 2022
Revised 29 October 2022
Accepted 30 November 2022
Published online 30 December 2022

Keywords:
Quartile Journals
Ensemble Classification
Bagging
Boosting

ABSTRACT

The ensemble method is considered an advanced method in both prediction and classification. The application of this method is estimated to have a more optimal output than the previous classification method. This article aims to determine the ensemble's performance to classify journal quartiles. The subject of agriculture was chosen because Indonesia is an agricultural country, and the interest of researchers in this field shows a positive response. The data is downloaded through the Scimago Journal and Country Rank with the accumulation in 2020. Labels have four classes: Q1, Q2, Q3, and Q4. The ensemble applied is Boosting and Bagging with Decision Tree (DT) and Gaussian Naïve Bayes (GNB) algorithms compiled from 2144 instances. The Boosting meta-ensembles used are Adaboost and XGBoost. From this study, the Bagging Decision Tree has the highest accuracy score at 71.36, followed by XGBoost Decision Tree with 69.51. The third is XGBoost Gaussian Naïve Bayes with 68.82, Adaboost Decision Tree with 60.42, Adaboost Gaussian Naïve Bayes with 58.2, and Bagging Gaussian Naïve Bayes with 56.12 results. This paper shows that the Bagging Decision Tree is the ensemble method that works optimally in this subject classification. This result suggests that the ensemble method can still fail to produce an ideal outcome that approaches the SJR system.

I. Introduction

The agricultural sector is one of the research areas that is expanding yearly. The journal grew, demonstrated by a considerable increase in the total number of papers published in this discipline each year at SCImago. Figure 1 depicts this industry's increased journals over the previous 20 years. Figure 1 illustrates this industry's increased number of journals over the last 20 years. This increase dramatically impacts the increasing number of literature sources for further research. Scimago ranks the journal itself. There are four journal classes, namely Q1, Q2, Q3, and Q4. However, in the provision of quartiles, there are some differences in values in the same journal in different fields.

Therefore, it is necessary to have data processing methods, such as classification. The technique for finding models or functions that explain and differentiate ideas or classes of data is known as classification [1]. This technique can predict the class label of an object whose label is unknown [2]. Therefore, we attempt to utilize a classification technique using the idea of an ensemble in this paper. Where Bagging and Boosting comprise the ensemble, this research aims to evaluate the ensemble classification mechanism's performance using quartile data from agricultural and biological science periodicals.

https://doi.org/10.17977/um018v5i22022p137-142

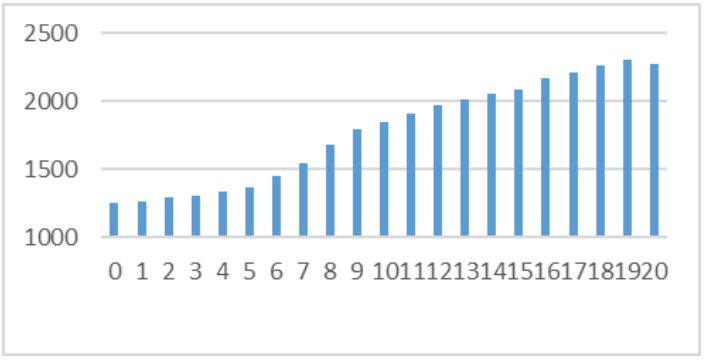

Fig. 1. The growth of agricultural and biological sciences journal

The ensemble model is a further development of the usual classification method. The working principle of this method is to combine the same two algorithms with a specific pattern [3] and decide the final result by the voting system [4]. The fundamental objective of using an ensemble is to achieve superior outcomes to a conventional single classifier. This is due to the method's ability to combat overfitting [5] and noise data [6].

The purpose of this study is to assess the effectiveness of the ensemble classification using Bagging and Boosting. Agricultural and biological science journal quartiles, particularly data accumulating for 2020, are the data sources. The research questions cover these points: Out of all the strategies used, which ensemble mechanism performs best? Are the publications in the domains of agriculture and biology ranked differently, and can the chosen ensemble solve this issue?

II. Method

This research is divided into four stages. The acquiring of datasets is the initial step. Data preprocessing, which aims to provide clean data suited for classification, comes next. The classification stage is the third. Ensemble Bagging and Boosting is the technique employed. The Confusion Matrix evaluation stage is the final step. In Figure 2, the research procedure is displayed.

Fig. 2. Method Process

A. Data Collecting

The first process carried out in this research is data collecting. Secondary data is collected from the SCImago page for journal and country rankings. The data subject used in agriculture and biological science in 2020. It was composed in February 2022. It consists of 2164 instances, with details listed in Table 1. Twenty qualities are present. However, just nine were applied. This is because these nine attributes are visible on the SCImago home page, leading one to believe that these are the ones that decide the journal quartiles [7][8]. SJR Best Quartile is the name of the label. This study falls under the multi-class classification because it includes the four classes Q1, Q2, Q3, and Q4.

H index, Total Docs (2020), Total Docs (3 Years), Total Refs, Total Cites (3 Years), City Docs (3 Years), Cite/ Doc (2 Years), and Ref.Doc are some of the attributes used. The label class, or the journal quartiles, are predicted using this feature as an independent variable.

B. Pre-processing

The data must be prepared in such a way as to produce accurate predictions. The data preparation stage to suit the needs of this process is called preprocessing [9]. Preprocessing can raise a classification method's predictive value [10]. Data cleansing, integration, transformation, reduction, feature selection, and resampling are a few examples of preprocessing [11][12]. However, not all types of preprocessing are used here. The technique used in this article is data cleaning.

Table 1. List of data set attribute

Attribute	Data Type	Range
Rank	Integer	1-2164
Sourceid	Real	12016-21101020133
Title	Nominal	Annual Review of Plant Biology, Ecology Letters, ISME Journal, etc
Type	Nominal	Journal
Issn	Nominal	995444, 00015342, etc
SJR	Real	0.1-11695
SJR Best Quartile	Nominal	Q1, Q2, Q3, Q4, NQ
H Index	Integer	0-342
Total Docs. (2020)	Integer	0-3921
Total Docs (3 Years)	Integer	0-6917
Total Refs.	Integer	0-251461
Total Cites. (3 years)	Integer	0-42304
Citable Docs. (3 years)	Integer	0-6322
Cite/ Doc (2 years)	Real	0-25.28
Ref.Doc.	Real	0-326.27
Country	Nominal	Indonesia, Hungary, Poland, etc
Region	Nominal	Northern America, Western Europe, the Asiatic region, etc
Publisher	Nominal	SEJANI Ltd, CSIC, EM International, etc
Coverage	Nominal	1988-2020, 1978-2020, 1977, 1996-2020 etc
Categories	Nominal	Agricultural and Biological Sciences, Ecology. Evolution Behavior and Systematic Cell Biology etc

Data cleaning eliminates extraneous data, such as missing values or noise [13]. Several instances in the agricultural and biological sciences data lack class labels. Therefore, the instances are removed to prevent incorrect classification. After this process, 2144 instances in the dataset are used. Table 2 contains information on the quantity of data in each class of labels following preprocessing.

Table 2. Label class summary

Class Label	Before Cleaning	After cleaning
q1	603	603
q2	551	551
q3	519	519
q4	471	471
-	20	-
Sum	2164	2144

C. Classification

The third stage that is passed is the classification process. There are two ensemble mechanisms in this stage. The first is Boosting with the Adaboost and XGBoost meta-ensembles. The second is the Bagging ensemble. Ensemble techniques use decision tree (DT) and Gaussian Naïve Bayes (GNB) algorithms as base learners. The scenario in this experiment is shown in Figure 3.

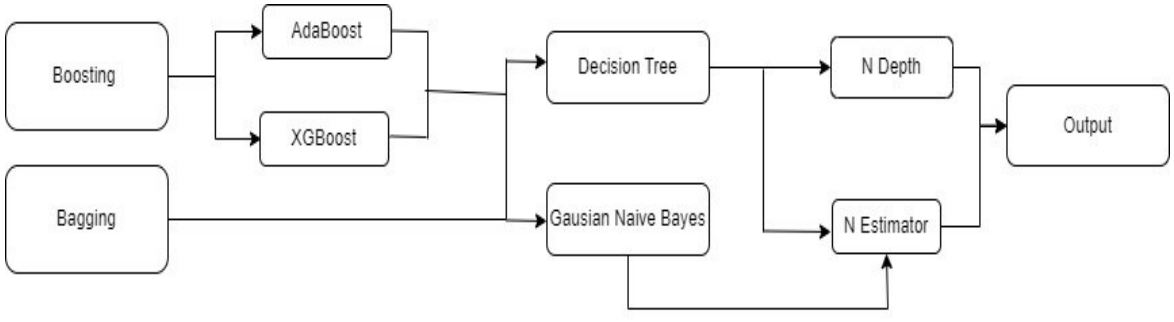

Fig. 3. Research scenario

Stage one is to break the dataset into training and testing data using the split test training command. The setting used is 20%: 80%. This comparison was chosen because this value produced sound output in several similar studies [14][15]. In addition, this value is often used [16]. The ensemble method's quartile classification of agricultural journals comes next. For both DT and GNB, this algorithm uses a base-learner repetition setting of 100. Regarding the 50 times set for the N depth DT, these numbers were selected randomly, understanding that they would be sufficient for this investigation.

D. Evaluation

The evaluation procedure used is the Confusion Matrix [17]. Information on predictable classifications and actual values using the classification system is contained in the Confusion Matrix [18]. Classification performance evaluation comprises six aspects: accuracy, precision, recall, specificity, f-score, and error rate [19][20]. However, not all of them will be applied in this study. The terms accuracy, precision, and recall will all be used in this essay.

III. Result and Discussion

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

The method has undergone various revisions during the classification phase. Adaboost DT, Adaboost GNB, XGBoost DT, XGBoost GNB, Boosting DT, and Boosting GNB are a few of them. Table 3 includes a list of the classification's outcomes. Figure 4 shows the table's results graphically.

Table 3. Classification results

Ensemble	Meta-Ensemble	Algorithm	Accuracy (%)	Precision (%)	Recall (%)
Boosting	Adaboost	DT	60.54	60.34	60.79
		GNB	59.58	46.76	47.96
	XGBoost	DT	69.97	76.96	63.31
		GNB	69.75	76.93	62.75
Bagging	-	DT	71.59	76.43	67.21
		GNB	56.12	47.78	46.29

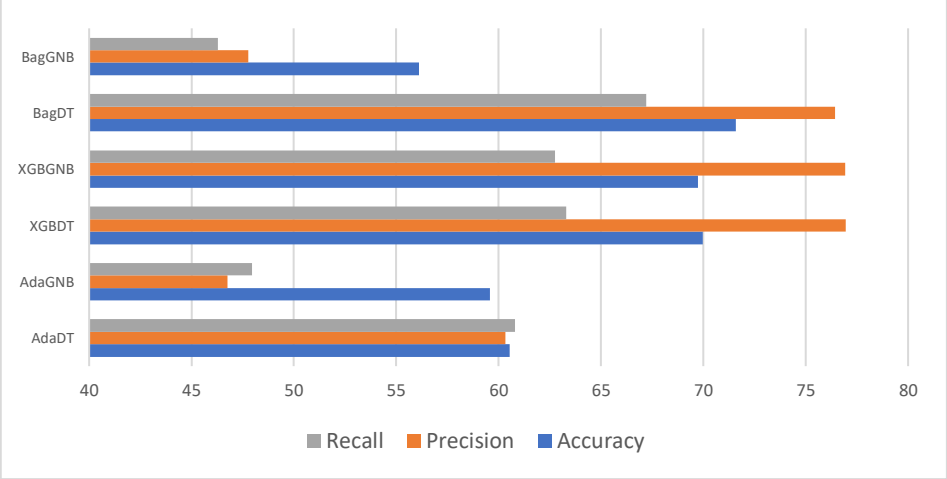

Fig. 4. Classification performance

Table 3 and Figure 4 show that the ensemble mechanism that works optimally, in this case, is Bagging DT, with an accuracy score of 71.59%. The second-best value is the XGBoost meta-ensemble with base learner DT with an accuracy value of 69.97%. If sorted from optimal to less than optimal performance, this classification process is Bagging DT, XGBoost DT, XGBoost GNB, AdaBoost DT, Adaboost GNB, and finally, Bagging GNB.

It is also seen that the XGBoost method has the slightest difference between the two algorithms, only 0.22%, as opposed to the Bagging approach, where there is a significant 15.47% difference. The difference in base-learner accuracy in the Adaboost meta-ensemble is 0.96%.

The ratio of correct optimistic predictions to the total number of positive predicted outcomes is known as precision [21]. In this realm, the values are XGBoost DT, XGBoost GNB, Bagging DT, Adaboost DT, Adaboost GNB, and Bagging GNB in that order from lowest to highest. Bagging DT had the highest recall score, coming up at 67.21%. Out of the six cases, Adaboost GNB has the lowest value. Recall quantifies the ratio of correctly predicted positive facts to actual positive facts [22].

This study produces a prediction accuracy value with an average of above 60%. These results indicate that all scenarios can be used to assess the journal quartiles because the results are still more than 50%.

Bagging can work better because Bagging extracts additional data for training from the dataset [23]. Each data component has the same chance of being selected. This data set is used to conduct model training simultaneously. The more training data obtained, the better knowledge of algorithms for classifying [24] and can reduce the variance of the classification process [25].

DT is a derivative of the independent variable, where each node has its conditions for features [26]. This node determines which node to go to in the following state. The proper sequence of nodes can produce the best output. DT does not make assumptions on the distribution of data [27], overcomes collinearity efficiently [28], and does not require data preprocessing [29]. However, this method can give overfitting if it uses too many branches. In this article, not too many branches are used so that the model can work optimally. In the case of Naïve Bayes often working by chance, this case cannot measure the accuracy of the prediction. On the other hand, Naïve Bayes is also weak in selecting attributes that can affect accuracy [30].

The data used is only quartile data for agricultural and biological science journals in the 2020 accumulation. This study also only uses simple settings in preprocessing. This action affects the performance of the classifier.

IV. Conclusion

In conclusion, the classification using ensemble models is applicable. According to the research findings, the Bagging Decision Tree is a method with reasonable accuracy, precision, and recall. Thus, it can be inferred that this approach may be used to resolve problems of a similar nature. The XGBoost meta-ensemble performs better in terms of the Boosting mechanism. XGBoost can indirectly minimize variance by lowering overfitting. The outcomes, nevertheless, can be improved. Therefore, it is essential to investigate other ensemble approaches, such as stacking, for future research. Using meta-ensemble and other base learners is strongly advised to create a better prediction score.

Declarations

Author contribution

All authors contributed equally as the main contributor of this paper. All authors read and approved the final paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare no known conflict of financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Reprints and permission information are available at http://journal2.um.ac.id/index.php/keds.

References

[1] D. Zhang and S. Lou, "The application research of neural network and BP algorithm in stock price pattern classification and prediction," Futur. Gener. Comput. Syst., vol. 115, pp. 872–879, Feb. 2021.
[2] Y. Zhang, Y. Wang, X.-Y. Liu, S. Mi, and M.-L. Zhang, "Large-scale multi-label classification using unknown streaming images," Pattern Recognit., vol. 99, p. 107100, Mar. 2020.
[3] J. Lin, H. Chen, S. Li, Y. Liu, X. Li, and B. Yu, "Accurate prediction of potential druggable proteins based on genetic algorithm and Bagging-SVM ensemble classifier," Artif. Intell. Med., vol. 98, pp. 35–47, Jul. 2019.
[4] G. Kaur, "A comparison of two hybrid ensemble techniques for network anomaly detection in spark distributed environment," J. Inf. Secur. Appl., vol. 55, p. 102601, Dec. 2020.
[5] F. Liu, M. Cai, L. Wang, and Y. Lu, "An Ensemble Model Based on Adaptive Noise Reducer and Over-Fitting Prevention LSTM for Multivariate Time Series Forecasting," IEEE Access, vol. 7, pp. 26102–26115, 2019.
[6] A. B. Shaik and S. Srinivasan, "A Brief Survey on Random Forest Ensembles in Classification Model," 2019, pp. 253–260.
[7] A. P. Wibawa, "International Journal Quartile Classification Using the K-Nearest Neighbor Method," 2019.
[8] A. P. Wibawa et al., "Naïve Bayes Classifier for Journal Quartile Classification," Int. J. Recent Contrib. from Eng. Sci. IT, vol. 7, no. 2, p. 91, 2019.
[9] M. J. Willemink et al., "Preparing Medical Imaging Data for Machine Learning," Radiology, vol. 295, no. 1, pp. 4–15, Apr. 2020.
[10] B. Sekeroglu, K. Dimililer, and K. Tuncal, "Student Performance Prediction and Classification Using Machine Learning Algorithms," in Proceedings of the 2019 8th International Conference on Educational and Information Technology, Mar. 2019, pp. 7–11.
[11] I. Cordón, J. Luengo, S. García, F. Herrera, and F. Charte, "Smartdata: Data preprocessing to achieve smart data in R," Neurocomputing, vol. 360, pp. 1–13, Sep. 2019.
[12] X. Shi, Y. D. Wong, M. Z.-F. Li, C. Palanisamy, and C. Chai, "A feature learning approach based on XGBoost for driving assessment and risk prediction," Accid. Anal. Prev., vol. 129, pp. 170–179, Aug. 2019.
[13] T. Wang, H. Ke, X. Zheng, K. Wang, A. K. Sangaiah, and A. Liu, "Big Data Cleaning Based on Mobile Edge Computing in Industrial Sensor-Cloud," IEEE Trans. Ind. Informatics, vol. 16, no. 2, pp. 1321–1329, Feb. 2020.
[14] Q. H. Nguyen et al., "Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil," Math. Probl. Eng., vol. 2021, pp. 1–15, Feb. 2021.
[15] R. Patgiri, H. Katari, R. Kumar, and D. Sharma, "Empirical Study on Malicious URL Detection Using Machine Learning," 2019, pp. 380–388.
[16] A. Mirbolouki, S. Heddam, K. Singh Parmar, S. Trajkovic, M. Mehraein, and O. Kisi, "Comparison of the advanced machine learning methods for better prediction accuracy of solar radiation using only temperature data: A case study," Int. J. Energy Res., vol. 46, no. 3, pp. 2709–2736, Mar. 2022.
[17] S. Ruuska, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen, and J. Mononen, "Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle," Behav. Processes, vol. 148, pp. 56–62, Mar. 2018.
[18] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," Inf. Sci. (Ny)., vol. 507, pp. 772–794, Jan. 2020.
[19] N. Khare et al., "SMO-DNN: Spider Monkey Optimization and Deep Neural Network Hybrid Classifier Model for Intrusion Detection," Electronics, vol. 9, no. 4, p. 692, Apr. 2020.
[20] O. S. Albahri et al., "Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: Taxonomy analysis, challenges, future solutions and methodological aspects," J. Infect. Public Health, vol. 13, no. 10, pp. 1381–1396, Oct. 2020.
[21] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," BioData Min., vol. 14, no. 1, p. 13, Feb. 2021.
[22] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview," Aug. 2020.
[23] P. Yariyan et al., "Improvement of Best First Decision Trees Using Bagging and Dagging Ensembles for Flood Probability Mapping," Water Resour. Manag., vol. 34, no. 9, pp. 3037–3053, Jul. 2020.
[24] A. M. Abdi, "Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data," GIScience Remote Sens., vol. 57, no. 1, pp. 1–20, Jan. 2020.
[25] S. Alelyani, "Stable bagging feature selection on medical data," J. Big Data, vol. 8, no. 1, p. 11, Dec. 2021.
[26] G. Pappalardo, S. Cafiso, A. Di Graziano, and A. Severino, "Decision Tree Method to Analyze the Performance of Lane Support Systems," Sustainability, vol. 13, no. 2, p. 846, Jan. 2021.
[27] S. Park, S.-Y. Hamm, and J. Kim, "Performance Evaluation of the GIS-Based Data-Mining Techniques Decision Tree, Random Forest, and Rotation Forest for Landslide Susceptibility Modeling," Sustainability, vol. 11, no. 20, p. 5659, Oct. 2019.
[28] B. Aronov, E. Ezra, and M. Sharir, "Testing Polynomials for Vanishing on Cartesian Products of Planar Point Sets: Collinearity Testing and Related Problems," Mar. 2020.
[29] H. Fujita and D. Cimr, "Decision support system for arrhythmia prediction using convolutional neural network structure without preprocessing," Appl. Intell., vol. 49, no. 9, pp. 3383–3391, Sep. 2019.
[30] M. Li and K. Liu, "Causality-Based Attribute Weighting via Information Flow and Genetic Algorithm for Naive Bayes Classifier," IEEE Access, vol. 7, pp. 150630–150641, 2019.