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-ch ôme-12-1 Ishizuka, 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; ue2017119@chikushi-u.ac.jp
* corresponding author

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.

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.

This is an open-access article under the CC BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/).
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.

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.
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>
<thead>
<tr>
<th>Class Label</th>
<th>Before Cleaning</th>
<th>After cleaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>q1</td>
<td>603</td>
<td>603</td>
</tr>
<tr>
<td>q2</td>
<td>551</td>
<td>551</td>
</tr>
<tr>
<td>q3</td>
<td>519</td>
<td>519</td>
</tr>
<tr>
<td>q4</td>
<td>471</td>
<td>471</td>
</tr>
<tr>
<td>-</td>
<td>20</td>
<td>-</td>
</tr>
<tr>
<td>Sum</td>
<td>2164</td>
<td>2144</td>
</tr>
</tbody>
</table>

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.
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

<table>
<thead>
<tr>
<th>Ensemble</th>
<th>Meta-Ensemble</th>
<th>Algorithm</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boosting</td>
<td>Adaboost</td>
<td>DT</td>
<td>60.54</td>
<td>60.34</td>
<td>60.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GNB</td>
<td>59.58</td>
<td>46.76</td>
<td>47.96</td>
</tr>
<tr>
<td></td>
<td>XGBoost</td>
<td>DT</td>
<td>69.97</td>
<td>76.96</td>
<td>63.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GNB</td>
<td>69.75</td>
<td>76.93</td>
<td>62.75</td>
</tr>
<tr>
<td>Bagging</td>
<td>-</td>
<td>DT</td>
<td>71.59</td>
<td>76.43</td>
<td>67.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GNB</td>
<td>56.12</td>
<td>47.78</td>
<td>46.29</td>
</tr>
</tbody>
</table>

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.

Publisher’s Note: Department of Electrical Engineering - Universitas Negeri Malang remains neutral with regard to jurisdictional claims and institutional affiliations.
References


