Journal of Information Technology Management

Print ISSN: 2008-5893

Online ISSN: 2423-5059

Determining Journal Rank by Applying Particle Swarm Optimization-Naive Bayes Classifier

Aji Prasetya Wibawa*

*Corresponding author, Associate Professor, Department of Electrical Engineering, University of Negeri Malang, Malang, Indonesia. E-mail: aji.prasetya.ft@um.ac.id

Sulton Aji Kurniawan

BSc., Department of Electrical Engineering, University of Negeri Malang, Malang, Indonesia. E-mail: sulton.aji18@gmail.com

Ilham Ari Elbaith Zaeni

Assistant Professor, Department of Electrical Engineering, University of Negeri Malang, Malang, Indonesia E-mail: ilham.ari.ft@um.ac.id

Abstract

SCImago Journal Rank (SJR) is one indicator of a journal's reputation. The value is calculated based on several published journals, such as scholarly journals' scientific impact, representing the number of quotes sent to a journal and the relevance or reputation of journals from which the quotations originate. A high SJR value means that the corresponding journal has a high reputation. This study aims to approach the SJR classification by implementing a machine learning approach. A simple yet powerful method Naïve Bayes Classifier (NBC), is selected. NBC utilizes probability calculations based on Bayes' theorem. However, NBC has an assumption that the attribute values do not depend on each other. This method is optimized using Particle Swarm Optimization (PSO) to overcome this weakness. This study used SJR data of the computer science domain from 2014 to 2017. Publication without Q rank is filtered for better performance. As a result, the accuracy of the proposed method is higher than the baseline. The use of PSO significantly improves the NBC performance based on the performed T-test. The PSO-NBC selects four of eight features: H index, Cites/ Doc (2 Years), and Ref. / Doc. Overall results show that using PSO-NBC is closer to SJR rather than using mere NBC.

Keywords: Naive Bayes Classifier, SCImago Journal Rank, Journal Quartile, Classification

DOI: http:// 10.22059/JITM.2021.305435.2559 Document Type: Research Paper

Received: December 06, 2020 Accept: June 28, 2021

Introduction

Researchers and academicians need to publish their manuscripts in scholarly journals. They may find the list of journals on various sites such as SCImago Journal & Country Rank, Impact Factor, and Google Scholar (Delgado-López-Cózar & Cabezas-Clavijo, 2013; Falagas, Kouranos, Arencibia-Jorge, & Karageorgopoulos, 2008). The portal provides a scientific indicator, SCImago Journal Rank (SJR), that ranks Scopus indexed journals in specific orders (Colledge et al., 2010). SJR measures scholarly journals' scientific influence based on the number of citations and sources.

The SJR measures scholarly journals' scientific impact, representing the number of quotes sent to a journal and the relevance or reputation of journals from which the quotations originate. A journal's SJR is a numerical value representing the total amount of weighted citations earned in the last three years for a chosen year per article published in the journal (Falagas et al., 2008; Mañana-Rodríguez, 2015). More SJR values should show a higher reputation in the journal.

The SJR indicator is a variant of the eigenvector centrality measure used in network theory (Roldan-Valadez, Salazar-Ruiz, Ibarra-Contreras, & Rios, 2019). Such measures establish the node's importance in a network based on the principle that connections to high-scoring nodes contribute more to the node's score. The SJR indicator has been used in extensive and heterogeneous journal citation networks. It is a size-independent indicator, and its values order journals by their "average prestige per article" and can be used for journal comparisons in science evaluation processes. Academics believe that this indicator is valid to depict journal quality in quartile categorization. Since classification or categorization is a part of the machine learning approach (Maxwell, Warner, & Fang, 2018). The implementation of such computational techniques is beneficial for SJR classification.

This research explores machine learning techniques as a new way to classify journals based on SJR categories. The proposed method is based on a Naïve Bayes classifier (NBC), a widespread and well-known learning problem algorithm that predicts a classified type output classifier (Sendari, Zaeni, Lestari, & Hariyadi, 2020). The classification approach is then optimized using Particle Swarm Optimization (PSO) for more efficient performance.

Methdology

Research Design

Figure 1 shows the research design. The first step is dataset collection. The second, preprocessing stage consisted of two elements: data cleaning and feature selection. The data cleaning deletes unused data values. The second element selects the most influence classification feature based on PSO. NBC classifies preprocessed data according to quartile categories. Finally, the evaluation stages generated three indicators, namely accuracy, precision, and recall.

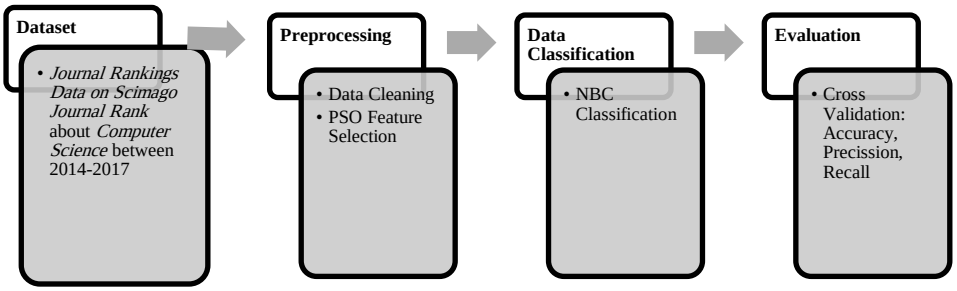

Figure 1. Research Design

We use the Scimago Journal Rank in the Computer Science domain for PSO-NBC journal classification, which was taken on January 3rd, 2019. The data has 7191 instances from 2014-2017. Table 1 shows 18 attributes of the dataset. We only use nine yellow highlighted attributes (8 features, one label) for classification purposes since the other features do not influence the journal quartiles (Q1-Q4) (SCImago, 2007).

Table 1. Dataset Features and Characteristics.

Attribute	Data Type	Value Range
Rank	Integer	(1-7191)
Sourceid	Real	(12125-21100855883)
Title	Nominal	(Journal of Statistical Software, Bioinformatics, IEEE Network, etc.)
Type	Nominal	(journal, book series, conference and proceedings, trade journal)
Issn	Nominal	(0149144X, 1573689X, 0219581X, etc.)
SJR	Real	(0.1-13.802)
SJR Best Quartile	Nominal	(Q1, Q2, Q3, Q4)
H index	Integer	(0-318)
Total Docs. (2017)	Integer	(0-20858)
Total Docs. (3 years)	Integer	(0-66063)
Total Refs.	Integer	(0-415920)
Total Cites (3 years)	Integer	(0-58176)
Citable Docs. (3 years)	Integer	(0-61823)
Cites / Doc. (2 years)	Real	(0-19.990)
Ref. / Doc.	Real	(0-269)
Country	Nominal	(United States, China, France, etc.)
Publisher	Nominal	(Elsevier BV, Springer Verlag, IOS Press, etc)
Categories	Nominal	(Software, Information Systems, Artificial Intelligence, etc.)

Preprocessing

The dataset within SJR has several problems in the form of missing values and attributes that do not influence the classification process. In this research, two preprocessing procedures are described as follows.

Data Cleaning

In the SJR dataset, several data is presented, and the data selected was journal-type data. Accordingly, the selected data was journal-type and comprised of a quartile value. Then, it was necessary to clean the data. Data Cleaning aims to eliminate errors and inconsistencies within the data to improve data quality (Blake, 2011; Lomet, 2001). In this research, data cleaning removes empty values contained in the dataset.

Particle Swarm Optimization Feature Selection

Not all features in SJR affect the classification process. Hence, it requires an approach to select beneficial features. This research uses Particle Swarm Optimization (PSO),) a population-based stochastic optimization technique inspired by the social behavior of flocks of birds and fish schools (Das, Jena, Nayak, Naik, & Behera, 2015). PSO also constitutes a swarm intelligence algorithm, namely the study of computational systems inspired by collective intelligence (Bratton & Kennedy, 2007; Cho & Hoang, 2017; Du & Swamy, 2016). Collective intelligence occurs because of population or homogeneous cooperation in a particle environment. The particle's environment is assumed to have a specific size where each particle has a random initial position (Moradi & Gholampour, 2016). The position of a random particle is in one place location. Each particle in one place is assumed to have two characters: position and speed. If each particle's position finds its best position, the information will be conveyed to other particles. Then, the particle speed process is carried out, which is stated in the following equation.

$$v_j(i+1)=v_j(i)+c_1r_1\left(P_{best}-x_j(i)\right)+c_2r_2\left(G_{best}-x_j(i)\right) \quad (1)$$

The position of the particle was updated to obtain the output of a new particle. The purpose of the update of particle position was to find the result used in the NBC parameter. The particle position update process was carried out using the following equation.

$$x_j(i+1) = x_j(i) + v_j(i+1) \quad (2)$$

Where

j = *particle index*

i = iteration

v_j = particle speed

x_j = *particle position*

P_{best} = the highest particle value

G_{best} = the highest iteration value

c_1 = *learning rates*

r_1 = random number

Naïve Bayes Classification (NBC)

The gigantic SJR dataset grouping will be challenging without using any classification approach. The data were separated according to four classes: Q1, Q2, Q3, and Q4. Here, the employed classification process is the Naive Bayes Classifier (NBC) algorithm. Naive Bayes often works far better in numerous complex real-world situations than might be expected. NBC is a popular model in Machine Learning applications because of its simplicity in allowing all attributes to contribute to the final decision equally (Muhamad, Prasojo, Sugianto, Surtiningsih, & Cholissodin, 2017; Wu et al., 2008). This simplicity is equivalent to computational efficiency, making the Naive Bayes technique attractive and suitable for various fields. The formula describes the Bayes' Theorem equation.

$$P(Q|X) = \frac{P(X|Q).P(Q)}{P(X)} \quad (3)$$

Where

X: Data with unknown class

Q: Hypothesis X as a specific class

P(Q|X): Probability of Q, depends on X

P(Q): Probability of Q (prior probability)

P(X|Q): Probability X in Q

P(X): Probability of X

Output & Evaluasi

The output of this study was a prediction model of class variables. Furthermore, accuracy, precision, and recall were calculated with the following equations.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \text{ x } 100\% \quad (4)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \text{ x } 100\% \quad (5)$$

$$\text{Recall (Sensitivity)} = \frac{\text{TP}}{\text{TP+FN}} \text{ x } 100\% \quad (6)$$

The evaluation was done by looking for scenarios that provide the best classification results. The best classification results were indicated by the highest value of accuracy, precision, and recall.

Results

This study took eight attributes in determining the class quality of journals, namely H index, Total Docs. (2017), Total Docs. (3 years), Total Refs., Total Cites (3 years), Citable Docs. (3 years), Cites/Doc. (2 years), and Ref./Doc. These eight attributes are the primary attributes used to classify the journal quartile. Several experiments were conducted with various NBC models and then evaluated using accuracy, precision, and recall evaluator to obtain the best model.

The first test was tested using the NBC algorithm without using PSO optimization with the k-fold value of 10 and NBC algorithm testing using PSO optimization with the k-fold value of 10 and population size of 8. Table 2 shows the result of this scenario.

Table 2. Accuracy of NBC Classification using k-fold (k=10)

Accuracy	Error Rate
52,67%	47,33%

Table 3. Accuracy of PSO-NBC using k-folds and population size of 8

Accuracy	Error Rate
59,96%	40,04%

The second test of the PSO-NBC classification included the size of the PSO population (Table 3). The population is the number of individuals or particles each generation. Then for the variable, the default value is used inertia weight using the value of 1.0, local best weight using the value of 1.0, global best weight using the value of 1.0, min weight using the value of 0.0, and max weight using the value of 1.0. Population testing was done with ten tests with the value of 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 with a maximum generation of 30. Table 4 shows the results of these tests. From Table 4, the weighting of each attribute occurs. The weight value amounted to – up to 1. If the weight value is close to or equal to one, the attribute strongly influences the classification.

Table 4. The Weight Values Based on Population Addition

Attribute	Population									
	10	20	30	40	50	60	70	80	90	100
H Index	1	1	1	1	0,8	0,8	0,9	1	1	1
Total Docs (2017)	1	0	0	1	1	0	0	0	0	0
Total Docs (3 years)	0	0	0	0	0	0	0	0	0	0
Total Refs	1	0	0	1	0	0	0	0	0	0
Total Cites (3 years)	0	0	0	0	0	0	0	0	0	0
Citable Docs (3 years)	0	1	0	0	0	0	0	0	0	0
Cites/Doc (2 Years)	0,7	0,5	1	0,7	1	0,7	1	0,6	1	1
Ref. / Doc	0	1	1	0	1	1	1	1	1	1

changes inaccuracy in each population added. The level of accuracy changes that occur, the value was not too far from the previous test. Table 5 shows the To find out the range referred to the researchers, it can be seen in the following table 5:

Table 5. Accuracy of PSO-NBC in Various Population

Population	Accuracy (%)	Error Rate (%)	Precision (%)	Recall (%)
10	57,67	42,33	59,31	58.33
20	59,33	40,67	69,08	59,91
30	60,79	39,21	61,47	61,37
40	59,95	40,05	60,89	60,58
50	60,86	39,14	61,59	61,45
60	60,93	39,07	61,73	61,50
70	60,93	39,07	61,75	61,51
80	60,86	39,14	61,62	61,45
90	60,93	39,07	61,59	61,45
100	60,86	39,14	61,59	61,42

From the results, we conclude that the accuracy of the PSO-NBC outperforms the original NBC approach. The results of the PSO-NBC accuracy turned out to acquire higher accuracy results than the NBC. Then, to increase the accuracy of PSO-NBC, adding a population size value can be performed further. Table 5 shows that the highest accuracy is 60.93%, with a population size of 60, 70, and 90. However, other population indicators are not the best of the variety: 61.75% of precision and 61.51% of recall.

Population addition also impacted the weight value of each class selected from the PSO feature selection. The increasing population of PSO also affects the execution time when the classification process is running. The execution time needed for the classification process takes longer. Although the execution time is longer than its baseline, the research shows that the optimal accuracy value is found in PSO-NBC with a population value of 70. Due to overall results, we assumed that PSO could optimize the classification performance using the NBC method.

Discussion

Particle Swarm Optimization Feature Selection can optimize NBC as proven by increasing the average value of accuracy, recall, and precision. The proposed method is efficient because of the reduced feature dimensions used in the classification process. The use of Particle Swarm Optimization on NBC can improve the accuracy of classifications that use NBC only.

The results obtained are that the NBC classification gets an accuracy of 52.60%, while the PSO-NBC has an accuracy of 60.93%. We use a T-test with a 0.05 significance level to prove the influence of PSO on NBC. As a result, the PSO implementation is significant to NBC accuracy since the T-value is -20.57226, significant at $p<0.05$.

The population size of the PSO algorithm affects the weighting of features. The number of PSO-selected features indicates this. The selected features show a population range of 10 to 100; of the eight chosen features, only three to 4 features. Features that are considered influential by PSO are the H index, Cites/ Doc (2 Years), and Ref. / Doc.

We compare the result with our previous research, which uses an inter-correlation aspect between features (Adiperkasa, Wibawa, Zaeni, & Widiyaningtyas, 2019). The best accuracy of intercorrelated NBC is only 59.14%. PSO-NBC again outperforms the baseline technique. Thus, it is more applicable than both NBC and intercorrelated NBC in terms of quartile classification.

Conclusion

Based on the statistical and computational test, it can be concluded that PSO-NBC can classify the journals in the quartile category as in SJR. Researchers may use the PSO-NBC as an alternative to Scopus journal classification. Furthermore, it can show the efficiency of general classification algorithm performance in classifying the journal database. In other words, everyone can research by comparing the classification result with the Scimago list of the scientific journal. Further research should consider a broader domain knowledge instead of the computer science category. The future action should consider the Scopus list of discontinued journals for smoother performance.

Acknowledgement

The researchers expressed our most generous gratitude to Universitas Negeri Malang, who support this research. We also thank the Knowledge Engineering and Data Science research center who shares the resources and ideas.

Conflict of interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Adiperkasa, R. P., Wibawa, A. P., Zaeni, I. A. E., & Widiyaningtyas, T. (2019). International Reputable Journal Classification Using Inter-correlated Naïve Bayes Classifier. *Proceedings - 2019 2nd International Conference of Computer and Informatics Engineering: Artificial Intelligence Roles in Industrial Revolution 4.0, IC2IE 2019*. https://doi.org/10.1109/IC2IE47452.2019.8940887

Blake, C. (2011). Text mining. In *Annual Review of Information Science and Technology* (Vol. 45). https://doi.org/10.1002/aris.2011.1440450110

Bratton, D., & Kennedy, J. (2007). Defining a Standard for Particle Swarm Optimization. *2007 IEEE Swarm Intelligence Symposium*, 120–127. https://doi.org/10.1109/SIS.2007.368035

Cho, M. Y., & Hoang, T. T. (2017). Feature Selection and Parameters Optimization of SVM Using Particle Swarm Optimization for Fault Classification in Power Distribution Systems. *Computational Intelligence and Neuroscience*, Vol. 2017. https://doi.org/10.1155/2017/4135465

Colledge, L., De Moya-Anegón, F., Guerrero-Bote, V., López-Illescas, C., El Aisati, M., & Moed, H. F. (2010). SJR and SNIP: two new journal metrics in Elsevier's Scopus. *Serials*, *23*(3), 215–221. https://doi.org/10.1629/23215

Das, H., Jena, A. K., Nayak, J., Naik, B., & Behera, H. S. (2015). A Novel PSO Based Back Propagation Learning-MLP (PSO-BP-MLP) for Classification. In *Computational Intelligence in Data Mining - Volume 2* (pp. 461–471). https://doi.org/10.1007/978-81-322-2208-8_42

Delgado-López-Cózar, E., & Cabezas-Clavijo, Á. (2013). Ranking journals: could Google Scholar Metrics be an alternative to Journal Citation Reports and Scimago Journal Rank? *Learned Publishing*, *26*(2), 101–113. https://doi.org/10.1087/20130206

Du, K.-L., & Swamy, M. N. S. (2016). Particle Swarm Optimization. In *Search and Optimization by Metaheuristics* (pp. 153–173). https://doi.org/10.1007/978-3-319-41192-7_9

Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R., & Karageorgopoulos, D. E. (2008). Comparison of SCImago journal rank indicator with journal impact factor. *The FASEB Journal*, *22*(8), 2623–2628. https://doi.org/10.1096/fj.08-107938

Lomet, D. B. (2001). Bulletin of the Technical Committee on Data Engineering. *Bulletin of the Technical Committee on Data Engineering*, *24*(4), 1–56.

Mañana-Rodríguez, J. (2015). A critical review of SCImago Journal & Country Rank. *Research Evaluation*, *24*(4), 343–354. https://doi.org/10.1093/reseval/rvu008

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing*, *39*(9), 2784–2817. https://doi.org/10.1080/01431161.2018.1433343

Moradi, P., & Gholampour, M. (2016). A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. *Applied Soft Computing Journal*, *43*, 117–130. https://doi.org/10.1016/j.asoc.2016.01.044

Muhamad, H., Prasojo, C. A., Sugianto, N. A., Surtiningsih, L., & Cholissodin, I. (2017). Naive Bayes Classifier Optimization by Using Particle Swarm Optimization on Iris Data. *Teknologi Informasi Dan Pendidikan*, *4*(3), 180–184.

Roldan-Valadez, E., Salazar-Ruiz, S. Y., Ibarra-Contreras, R., & Rios, C. (2019). Current concepts on bibliometrics: a brief review about impact factor, Eigenfactor score, CiteScore, SCImago Journal Rank, Source-Normalised Impact per Paper, H-index, and alternative metrics. *Irish Journal of Medical Science (1971 -)*, *188*(3), 939–951. https://doi.org/10.1007/s11845-018-1936-5

SCImago. (2007). *Description of Scimago Journal Rank Indicator*. 1–4. Retrieved from http://bit.ly/1tNwvj6

Sendari, S., Zaeni, I. A. E., Lestari, D. C., & Hariyadi, H. P. (2020). Opinion Analysis for Emotional Classification on Emoji Tweets using the Naïve Bayes Algorithm. *Knowledge Engineering and Data Science*, *3*(1), 50–59. https://doi.org/10.17977/um018v3i12020p50-59

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., … Steinberg, D. (2008). Top 10 algorithms in data mining. In *Knowledge and Information Systems* (Vol. 14). https://doi.org/10.1007/s10115-007-0114-2

Bibliographic information of this paper for citing:

Wibawa, A.; Kurniawan, A. & Zaeni, I. (2021). Determining Journal Rank by Applying Particle Swarm Optimization-Naive Bayes Classifier. *Journal of Information Technology Management*, 13(4), 116-125.