Naïve Bayes Classifier for Journal Quartile Classification

https://doi.org/10.3991/ijes.v7i2.10659

Aji Prasetya Wibawa (✉), Ahmad Chandra Kurniawan, Della Murbarani Prawidya Murti, Risky Perdana Adiperkasa, Sandika Maulana Putra, Sulton Aji Kurniawan, Youngga Rega Nugraha
Universitas Negeri Malang, Malang, Indonesia
aji.prasetya.ft@um.ac.id

Abstract—Classification is a process for distinguishing data classes, with the aim of being able to estimate the class of an object with unknown label. One popular method that used for classifying data is Naïve Bayes Classifier. Naïve Bayes Classifier is an approach that adopts the Bayes theorem, by combining previous knowledge with new knowledge. The advantages of this method are the simple algorithm and high accuracy. In this study, it will show the ability of Naïve Bayes Classifier to classify the quality of a journal commonly called Quartile. This study use a dataset of 1491 instances. The results show an accuracy of 71.60% and an error rate of 28.40%.

Keywords—Classification, Naive Bayes Classifier, Journal Quartile.

1 Introduction

One of the reference sources in writing scientific papers is articles. Articles are one type of scientific work produced from research studies or literature studies. Scientific articles are divided into research and non-research articles. Research article are sourced from research reports, while the non-research articles contain the thoughts, arguments, and authors opinions with the support of scientific sources. In general, the contents of scientific articles include titles, lines of author, abstracts, keywords, contents or body of the text, and literature reviews.

People can find various articles in the journal. A journal is a special scientific work which contains articles in a particular field of science. The role of the journal in the preparation of scientific work is as a provider of reference materials. Each journal has different qualities, including Q1, Q2, Q3, Q4 and NQ. The value of Q1 shows the highest value. In the Scimago Journal and Country Rank, rankings from various types of journals can be found, but there are still unbalanced values in the journal ranking system [1]. The problem is that the Quartile category of each field of study is not the same. for example, a journal in the X field could be lower than the Quartile value in another field

Based on these problems, a data processing method is required, namely classification. Classification is one method for finding a model or function that explains and distinguishes concepts or data classes. The model is derived based on the analysis of a series of training data and can be used to predict the class label of an object with unknown label [2].

Naïve Bayes is one of the data classification algorithms. This algorithm belongs to the top 10 algorithms in data mining [3]. The Naive Bayes algorithm is simple probabilistic classification. This algorithm calculates a set of probabilities by calculating the frequency and combination of values in a particular data set [4]. The probability of certain features in the data appears as members in a probability sequence and it obtained by calculating the frequency of each feature value in the class from the training data set. The training data set is a subset used to train classification algorithms. The training process uses known values to predict unknown values [5].

Several comparative studies of classification methods said that the Naïve Bayes Classifier algorithm shows the best area under the curve (AUC) value than LogR and LC algorithms [6]. The results in other studies show that the Naïve Bayes Classifier algorithm has the best accuracy compared with the Lazy-IBK algorithm, Zero-R and Decision Tree-J48 [7]. The Naïve Bayes Classifier algorithm has also proven effective and potentially good in many practical applications, including text classification, medical diagnosis and system performance management [8][9].

This study uses the Naïve Bayes Classifier Algorithm to process numerical data. The data used is a journal ranking data obtained from Scimago Journal and Country Rank. Through this research, it is expected to be able to produce quality classification of journals in Q1, Q2, Q3, Q4 and NQ label with optimal accuracy.

2 Method

2.1 Dataset

The dataset was taken from the Journal Rangkings in the Scimago Journal and Country Rank on November 5, 2018. The dataset specified in Computer Science as the subject area and only uses instances with journal types. The dataset consists of 1491 instances and 10 attributes show in Table I.

The classification output is determined based on a label class attribute. From the dataset, label class attribute is SJR Best Quartile show in Table II.

The proportion of each class member as output is shown in Table III.

Table 1. List of attributes in the dataset

Attribute Name	Attribute Explanation	Data Type	Range of Values
SJR Best Quartile	Journal's best quartile	Polynomial	Q1, Q2, Q3, Q4, NQ
SJR	The average number of weighted quotes received in 2017 with articles published in journals in 2016, 2015 and 2014.	Real	
H index	Journal's number of articles (h) that have received at least h citations over the whole period	Integer	
Total Docs.(2017)	Journal's published articles in 2017. All type of documents are considered.	Real	

Total Docs. (3years)	Journal's published articles in 2016, 2015 and 2014. All type of documents are considered.	Integer	
Total Refs.	Number of references included in the journal's published articles in 2017.	Real	
Total Cites (3years)	Citations in 2017 received by journal's documents published in 2016, 2015 and 2014.	Integer	
Citable Docs. (3years)	Journal's citable documents in 2016, 2015 and 2014. Citable documents include: articles, reviews and conference papers.	Integer	
Cites/Doc. (2years)	Average citation per document in a 2 year period. This metric is widely used as impact index.	Real	
Ref. / Doc.	Average amount of references per document in 2017.	Real	

Table 2. Detail of label class attribute

Label	**Attribute Explanation**	**Data Type**	**Range of Values**
SJR Best Quartile	Journal's best quartile	Polynomial	Q1, Q2, Q3, Q4, NQ

Table 3. Proportion of dataset class members

Class Name	**Number of Instances**	**Percentage**
SJR Best Quartile "Q1"	407	27.30 %
SJR Best Quartile "Q2"	376	25.22 %
SJR Best Quartile "Q3"	338	22.67 %
SJR Best Quartile "Q4"	320	21.46 %
SJR Best Quartile "NQ"	50	3.35 %

2.2 Data preprocessing

Before the main classification process, the dataset needs to be prepared to optimize it. From the data above, there is an unbalanced proportion between the number of label instances of each class. Class Q1 has more sample proportions than classes Q2, Q3, Q4, and NQ. This study uses the Undersampling technique to balance data between class labels based on the smallest proportion. The implementation of the Undersampling technique shown in Table IV.

Table 4. Implementation of the undersampling technique on the proportion of label classes

Label Class Name	**Number of Instance**			
	Without Undersampling		***With Undersampling***	
	Total	***Percentage***	***Total***	***Percentage***
Q1	407	27.30 %	50	20%
Q2	376	25.22 %	50	20%
Q3	338	22.67 %	50	20%
Q4	320	21.46 %	50	20%
NQ	50	3.35 %	50	20%

The Undersampling technique in RapidMiner software can be applied with Sample operators. By applying the Undersampling concept where classes will be balanced based on minority classes. In Table IV, before Undersampling, the number of label class data is not balanced. Q1 label is the majority class with 407 instance while NQ is a minority class with 50 instance. By taking the Undersampling technique, the proportion of label class data will be balanced based on the minimum number of proportions. So, the proportion of the five label class data is balanced to 50 instances each class with a percentage of 20%.

2.3 Naive bayes algorithm

Naive Bayes is a classifier using probability and statistical methods proposed by a British scientist, Revered Thomas Bayes [10]. Naive Bayes often works much better in many complex real-world situations than might be expected [11].

Naïve Bayes is a popular model in Machine Learning applications because of its simplicity in allowing all attributes to contribute to the final decision equally. This simplicity is equivalent to computational efficiency, which makes the Naïve Bayes technique attractive and suitable for various fields [12]. The main element of Naïve Bayes Classifier is about three aspects, they are prior, posterior dan class conditional probability [13].

The advantages of the Naive Bayes algorithm as show in [10] and [14]-[16] are as follows:

- Small training data
- Simple computing
- Easy to implement
- Time efficiency
- Can handle big data
- Can handle incomplete data (*missing value*)
- Not sensitive to irrelevant features
- Not sensitive to data noise

The Bayes Theorem formula is as follows [16]:

$$P(Q|X) = \frac{P(X|Q).P(Q)}{P(X)} \tag{1}$$

With

X Data with unknown class

Q The hypothesis X is a specific class

P(Q|X) The probability of the Q hypothesis refers to X

P(Q) Probability of the hypothesis Q (prior probability)

P(X|Q) Probability X in the hypothesis Q

P(X) Probability X

To explain the Naïve Bayes theorem, it must be known that the classification process requires various clues to determine the class according to the sample analyzed. Therefore, the Bayes theorem above is adjusted as follows:

$$P(Q|X_1 \dots X_n) = \frac{P(Q)P(X_1 \dots X_n|Q)}{P(X_1 \dots X_n)} \tag{2}$$

Where Q variable is a representation of class, while variable X_1 ... X_n represents the characteristics of the instructions needed for the classification process. Then the formula explains that the probability of the entry of certain characteristic samples in class Q (Posterior) is the opportunity for the emergence of the Q class (before the entry of the sample is called prior), multiplied by the probability of the appearance of sample characteristics in class Q (also called likelihood). Then divided by the chance of the appearance of sample characteristics globally (also called evidence). Therefore, the formula above can also be written simply as follows:

$$Posterior = \frac{Prior\ x\ likelihood}{evidence}$$

Evidence values are always fixed for each class in one sample. The value of the posterior will be compared with other class posterior values, to determine the class of a sample to be classified. Further explanation of the Bayes formula is done by describing $(Q|X_1 \dots X_n)$ using the multiplication rule as follows:

$$\begin{aligned} P(Q|X_1 \dots X_n) &= P(Q)P(X_1 \dots X_n|Q) \qquad (3) \\ &= P(Q)P(X_1|Q)P(Q|X_1 \dots X_n|Q,X_1) \\ &= P(Q)P(X_1|Q)P(X_2|Q,X_1)P(X_3 \dots X_n|Q,X_1,X_2) \\ &= P(Q)P(X_1|Q)P(X_2|Q,X_1)P\big(X_3\big|Q,X_{1,}X_2\big),P(X_4 \dots X_n|Q,X_1,X_2,X_3) \\ &= P(Q)P(X_1|Q)P(X_2|Q,X_1)P\big(X_3\big|Q,X_{1,}X_2\big) \dots P(X_n|Q,X_1,X_2,X_3,\dots,X_{3n-1}) \end{aligned}$$

It can be seen that the results of the translation cause more and more complex factors that affect the probability value, which is almost impossible to analyze one by one. As a result, the calculation becomes difficult to do. This is where the assumption of very high independence (naive) is used, that each of the instructions $(X_1, X_2 \dots X_n)$ are free from each other. With these assumptions, a formula applies as follows :

$$P\big(P_i\big|X_j\big) = \frac{P\big(X_i \cap X_j\big)}{P\big(X_j\big)} = \frac{P(X_i)P\big(X_j\big)}{P\big(X_j\big)} = P(X_i)$$

For $i \neq j$, so that

$$\text{Arq max} : P(X_i|Q,X_j) = P(X_i|Q) \qquad (4)$$

From the above equation it can be concluded that the assumption of naïve independence makes the conditions of opportunity simpler, making the calculation possible to do. Next, the translation of $P(Q|X_1, \dots, X_n)$ can be simplified to become:

$$\begin{aligned} arq\ max: P(Q|X_1, \dots, X_n) &= P(Q)P(X_1|Q)P(X_2|Q)P(X_3|Q) \\ &= P(Q) \textstyle\prod_{i=1}^{n} P(X_n|Q) \qquad (5) \end{aligned}$$

The equation above is a model from Naïve Bayes which will be used in the classification process. For classification with numerical data can be handled using the Standard Probability Density function, where the function represents the distribution of known data [17]. The formula from Gauss Density show in equation (6).

$$P(X_i = x_i|Q = q_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(X_i-\mu_{ij})^2}{2\sigma^2{}_{ij}}} \qquad (6)$$

Keterangan :
P : Chance
X_i : i-attribute
x_i : i-attribute value
Q : Related class
q_i : Related subclass Q
μ : Mean, the average of all attributes
σ : Standard Deviation, the variance of all attributes

2.4 Confusion Matrix

Confusion matrix contains information about predictable and actual classifications with a classification system. System performance is generally evaluated using data in the form of a matrix [18]. Through Confusion Matrix, the accuracy, error rate, precision and recall values can be known. *Confusion Matrix* show in Table V.

Table 5. Confusion matrix

Class	Positive Classified	Negative Classified
Positive	TP (*True Positive*)	FN (*False Negative*)
Negative	FP (*False Positive*)	TN (*True Negative*)

Calculation of Confusion Matrix is done with TP, FN, FP and TN.

TP (True Positive) is the amount of positive data that has a true truth value.

FN (False Negative) is the amount of negative data that is considered by the system to have the truth value false.

FP (False Positive) is the amount of positive data that is considered by the system to have the truth value false.

TN (True Negative) is the amount of negative data that is considered by the system to have true truth value.

3 Result and Discussion

The use of Cross Validation with k-fold=5 show the best accuracy produce by Naïve Bayes Classifier which reaches 71.60%. Table VI shows the results of classification consisting of the value of accuracy, error, precision and recall from the Naïve Bayes Classifier.

Table 6. Result of classification

Stratified Split-validation	
Summary	
Accuracy	71.60 %
Classification Error	28.40 %
Precision	75,12%
Recall	71,60%

Naïve Bayes Classifiers' performance isn't too bad because the accuracy is still above 50%. This shows that the Naïve Bayes Classifier algorithm can be used to classify the quality of journals by only requiring a small amount of training data to determine the estimation of parameters needed in the classification process. According to Syarifah and Muslim (2015), Naïve Bayes Classifier has several advantages, such as fast in the calculation process, simple algorithms and high accuracy [19]. However, according to Muhammad (2017), the probability on the Naïve Bayes Classifier algorithm cannot measure the accuracy of a prediction. In addition, Naïve Bayes Classifier also has weaknesses in attribute selection that can affect on the value of accuracy [10]. Therefore, the researcher recommends using the Naïve Bayes Classifier algorithm with an optimization method to increase the accuracy of the Naïve Bayes Classifier algorithm.

In the next study, researchers will combine the Naïve Bayes Classifier algorithm with several optimization methods, such as Particle Swarm Optimization or Genetic Algorithms. In addition, researchers will also use other classification algorithms such as Support Vector Machine and k-Nearest Neighbor. Preprocessing techniques such as the imputation method can also be done by randomly removing data to compare the value of accuracy generated after the imputation method than without imputation method.

4 Conclusion

Based on the results of the discussion in this study, the data are classified into several labels, namely Q1, Q2, Q3, Q4 and NQ. The variable used in this study is H index, SJR, Total Docs. (2017), Total Docs. (3years), Total Refs, Total Cites (3years), Citable Docs. (3years), Cites / Doc. (2years), and Ref. / Doc. The classification of quality journals can make it easier for people to choose quality journals. In this study, the researchers also concluded that Naïve Bayes Classifier algorithm was able to classify the quality of journals, even though the value of accuracy is not too optimal. For better accuracy, journals quartile classification using the Naive Bayes Classifier algorithm needs to be optimized with other algorithms.

5 References

[1] J. Mañana-Rodríguez, "A critical review of SCImago Journal & Country Rank," Res. Eval., pp. 1–12, 2014.

[2] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, Third. Waltham, USA: Elsevier Inc., 2012.

[3] X. Wu et al., Top 10 algorithms in data mining, vol. 14, no. 1. 2008.

[4] T. R. Patil, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification," Int. J. Comput. Sci. Appl. ISSN 0974-1011, vol. 6, no. 2, pp. 256–261, 2013.

[5] G. Dimitoglou, J. a Adams, and C. M. Jim, "Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability," J. Neural Comput., vol. 4, no. 8, pp. 1–9, 2012.

[6] R. Entezari-Maleki, R. Arash, and M. Behrouz, "Comparison of Classification Methods Based on the Type of Attributes and Sample Size," J. Converg. Inf. Technol., vol. 4, no. 3, pp. 94–102, 2009.

[7] S. Fitri, "Perbandingan Kinerja Algoritma Klasifikasi Naïve Bayesian , Lazy-Ibk , Zero-R , Dan Decision Tree- J48," Dasi, vol. 15, no. 1, pp. 33–37, 2014.

[8] I. Rish, "An empirical study of the naive Bayes classifier," Empir. methods Artif. Intell. Work. IJCAI, vol. 22230, no. January 2001, pp. 41–46, 2001.

[9] I. Journal, S. Engineering, S. C. Applications, S. C. View, and M. Learning-based, "Is Naïve Bayes a Good Classifier for Document Classification ?," Int. J. Softw. Eng. Its Appl., vol. 5, no. January, pp. 37–46, 2011.

[10] H. Muhamad et al., "Optimasi Naive Bayes Classifier dengan Menggunakan Particle Swarm Optimization pada Data Iris," Teknol. Inf. Dan Pendidik., vol. 4, no. 3, pp. 180–184, 2017. https://doi.org/10.25126/jtiik.201743251

[11] M. Dhanashree S, B. Mayur P, and D. Shruti D, "Prediction System For Heart Disease Using Naive Bayes," Int. J. Adv. Comput. Math. Sci., vol. 3, no. 3, pp. 290–294, 2012.

[12] D. Xhemali, C. J. Hinde, and R. G. Stone, "Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," Int. J. Comput. Sci., vol. 4, no. 1, pp. 16–23, 2009.

[13] A. Jamain and D. J. Hand, "The Naive Bayes Mystery: A classification detective story," Pattern Recognit. Lett., vol. 26, no. 11, pp. 1752–1760, 2005. https://doi.org/10.1016/j.patrec.2005.02.001

[14] C. Science and S. Engineering, "Comparative analysis of Naive Bayes and J48 Classification Algorithms," IJARCSSE, vol. 5, no. 12, pp. 813–817, 2015.

[15] R. P. Rajeswari, K. Juliet, and Aradhana, "Text Classification for Student Data Set using Naive Bayes Classifier and KNN Classifier," Int. J. Comput. Trends Technol., vol. 43, no. 1, pp. 8–12, 2017. https://doi.org/10.14445/22312803/ijctt-v43p103

[16] Bustami, "Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi," J. Inform., vol. 8, no. 1, pp. 884–898, 2014.

[17] R. J.Roiger, Data Mining A Tutorial-Based Primer, 2nd ed. London: CRC Press, Taylor & Francis Group, 2017.

[18] A. K. Santra and C. J. Christy, "Genetic Algorithm and Confusion Matrix for Document Clustering," IJCSI Int. J. Comput. Sci. Issues, vol. 9, no. 1, pp. 322–328, 2012.

[19] A. Syarifah and M. A. Muslim, "Pemanfaatan Naïve Bayes untuk Merespon Emosi dari Kalimat Berbahasa Indonesia," UNNES J. Math., vol. 4, no. 2, pp. 147–156, 2015.

6 Authors

Aji Prasetya Wibawa works at the Electrical Engineering Department of the Universitas Negeri Malang in Indonesia. aji.prasetya.ft@um.ac.id

Ahmad Chandra Kurniawan works for Universitas Negeri Malang at the department of Electrical Engineering, in Malang, Indonesia. ahmadchandra52@gmail.com

Della Murbarani Prawidya Murti works at the Electrical Engineering Department, Universitas Negeri Malang, Malang, Indonesia. dellamurbarani4@gmail.com

Risky Perdana Adiperkasa works for UNM at the Dept of EE in Malang, Insonesia. riskyperdana999@gmail.com

Sandika Maulana Putra works at the Electrical Engineering Department of the Universitas Negeri Malang in Indonesia. maulanasandika12@gmail.com

Sulton Aji Kurniawan works at the Electrical Engineering Department of the Universitas Negeri Malang in Indonesia. sulton.aji18@gmail.com

Youngga Rega Nugraha works at the Electrical Engineering Department of the Universitas Negeri Malang in Indonesia. younggarega@gmail.com

Article submitted 2019-04-14. Resubmitted 2019-05-21. Final acceptance 2019-05-21. Final version published as submitted by the authors.