JU-KOMI : Jurnal Komputer Indonesia
e-ISSN : 2963-0460 (media online)
Vol. 1, No. 2, 2023, pp. 30-33
https://jurnal.seaninstitute.or.id/index.php/jukomi/login

Comparison of Naive Bayes Classifier with Feature Selection Gain Ratio on Data Classification

Ahmad Rozi[1]
[1]Sistim Informasi, Universitas Mahkota Tricom Unggul, Indonesia

Article Info

Article history:

Received, March 8, 2023
Revised, March 9, 2023
Accepted, March 10, 2023

Keywords:

Naïve Bayes,
Relief-F,
Gain Ratio,
Feature Selection,
Accuracy

ABSTRACT

In this study, the authors propose a process of increasing accuracy in Naïve Bayes with a combination of feature selection using the gain ratio and Relief-F methods. The cause of the less than optimal accuracy in Naïve Bayes compared to other classification methods is due to the less significant influence of features and the relatively low percentage of influence of data in determining the class of new data. The Gain Ratio and Relief-F methods are used to select features that have a poor correlation with the data being tested. The test of the proposed method is to compare the accuracy obtained from the Naïve Bayes method without using feature selection with Naïve Bayes using Gain Ratio and Relief-F feature selection. The test results obtained were the proposed Gain Ratio and Relief-F methods, the gain ratio method did not increase while the Relief-F method was able to increase the classification accuracy of naïve Bayes with an increase obtained of 0.2928% when compared to the Naïve Bayes test without feature selection.

This is an open access article under the CC BY-NC license.

Corresponding Author:

Ahmad Rozi,
Sistem Informasi,
Universitas Mahkota Tricom Unggul,
Gedung Jati Junction Lt.24 Jl. Perintis Kemerdekaan No.3A, Medan - Sumatera Utara.
Email: zeerozy737@gmail.com

1. INTRODUCTION

Naive Bayes Classifier (NBC) is a method in data mining that is used for data classification. This method aims to build a classification model using a training dataset to determine a class [1]. In addition, this method is also a simple method for building a classification model by assigning class labels as examples of problems, then representing them as feature value vectors by taking class labels from several finite sets [2].

Several studies on data classification using NBC have been published by researchers. In the [3] study, predicting chronic kidney disease used the Random Forest, Naïve Bayes, and K-Nearest Neighbor (KNN) classification methods. The results of this study show that the accuracy of the Random Forest method is better than the Naïve Bayes and K-Nearest Neighbor (KNN) methods.

Furthermore, research [4], which compared the K-Nearest Neighbor (KNN), Naive Bayes, and Decission Tree (J-48) methods for predicting creditworthiness, obtained results stating that the Decission Tree (J-48) method has a higher level of accuracy. higher (92.21%) than the Naive Bayes method (81.83%) and KNN (81.82%).Literature review that has been done author used in the chapter "Introduction" to explain the difference of the manuscript with other papers, that it is innovative, it are used in the chapter "Research Method" to describe the step of research and used in the chapter "Results and Discussion" to support the analysis of the results [2]. If the manuscript was written really have high originality, which proposed a new method or algorithm, the additional chapter after

the "Introduction" chapter and before the "Research Method" chapter can be added to explain briefly the theory and/or the proposed method/algorithm [4].

In data mining, the classification process is carried out with one of the stages, namely the preprocessing stage to anticipate data from logs that may be incomplete or have noise and inconsistent data used [5]. At the data preprocessing stage, a method can be used that has a direct effect on the classification results, namely by selecting features [6].

To obtain features that are significant to the accuracy value of a classification method in selecting several features which are a subset of the original features, it is known as feature selection. The best solution that can be used to reduce the dimensions of the data we use is to do the feature selection. According to research [7], states that feature selection used in classification can improve performance by removing or removing features that are irrelevant to the classification results. Then in the research conducted [8] also states that feature selection can reduce high data dimensions and can improve the performance of a classification method

2. RESEARCH METHOD

A. Research Analysis Steps

The steps applied in this study are as follows:

1. Determine the dataset to be used.
2. Perform a selection of dataset features with gain ratio and relief-F.
3. Performs Naive Bayes accuracy calculations without feature selection.
4. Perform calculations of Naive Bayes accuracy by selecting the gain ratio and F-relief features.
5. Naïve Bayes classification accuracy with gain ratio and F-relief feature selection.

B. Data Used

The data used in this study is the Wisconsin breast cancer dataset obtained at UCI Machine Learning. The data consists of 699 data with 10 attributes and 1 class attribute.

Table 1. Gain Ratio Weight Value

No	Feature	Weight Value
1	*Sample_Code_Number*	0.0996
2	*Clump_Thickness*	0.1522
3	*Uniformity_of_Cell_Size*	0.2996
4	*Uniformity_of_Cell_Shape*	0.2719
5	*Marginal_Adhesion*	0.2099
6	*Single_Epithelial_Cell_Size*	0.2333
7	*Bare_Nuclei*	0.3027
8	*Bland_Chromatin*	0.2005
9	*Normal_Nucleoli*	0.2375
10	*Mitoses*	0.1876

Then the process of calculating the weight values on the dataset using *Relief-F* using *Weka Waikato tools*, so that the weights of 10 criteria with different weight values are obtained. The results of the weight calculation using the *Relief-F method* can be seen in Table 2 below:

Table 2. Relief-F Weight Value

No	Feature	Weight Value
1	*Sample_Code_Number*	0.00527
2	*Clump_Thickness*	0.47218
3	*Uniformity_of_Cell_Size*	0.53719
4	*Uniformity_of_Cell_Shape*	0.54261
5	*Marginal_Adhesion*	0.24436
6	*Single_Epithelial_Cell_Size*	0.29605
7	*Bare_Nuclei*	0.60117
8	*Bland_Chromatin*	0.42635
9	*Normal_Nucleoli*	0.28243
10	*Mitoses*	0.06896

Next is to calculate the accuracy of the dataset used using the *Naive Bayes method* without feature selection. The accuracy obtained by the *Naive Bayes classification method* on the *Wisconsin Breast Cancer dataset* is 97.3646. The accuracy results obtained using the *Naïve Bayes calculation method* can be seen in Table 3 below:

Table 3. Naive Bayes Accuracy Value

No	Method	accuracy
1	*Naive Bayes*	97.3646

Then calculate the accuracy of the Naïve Bayes method with the selection of the gain ratio feature and the selection of the F-relief feature. The feature weight value limit used in this calculation is 0.15. So that in the gain ratio feature selection and the relief-F feature selection, only features that have a weight value of > 0.15 are used. From the existing provisions, the features obtained in the new dataset can be seen in Table 4 below:

Table 4. Gain Ratio Feature Selection

No	Feature	Weight Value
1	Clump_Thickness	0.1522
2	Mitoses	0.1876
3	Bland_Chromatin	0.2005
4	Marginal_Adhesion	0.2099
5	Normal_Nucleoli	0.2375
6	Single_Epithelial_Cell_Size	0.2333
7	Uniformity_of_Cell_Shape	0.2719
8	Uniformity_of_Cell_Size	0.2996
9	Bare_Nuclei	0.3027

While the new features obtained from feature selection using the Relief-F method can be seen in table 5 below:

Table 5. Relief-F Weight Value

No	Feature	Weight Value
1	Marginal_Adhesion	0.24436
2	Normal_Nucleoli	0.28243
3	Single_Epithelial_Cell_Size	0.29605
4	Bland_Chromatin	0.42635
5	Clump_Thickness	0.47218
6	Uniformity_of_Cell_Size	0.53719
7	Uniformity_of_Cell_Shape	0.54261
8	Bare_Nuclei	0.60117

The level of accuracy obtained by the Naïve Bayes method by using the Gain Ratio feature selection and the Relief-F feature selection can be seen in table 6 below:

Table 6. Accuracy Value of Naive Bayes with Feature Selection

No	Method	Feature Selection	accuracy
1	*Naive Bayes*	*Gain Ratio*	97.3646
2	*Naive Bayes*	*Gain Ratio*	97.6574

3. RESULTS AND DISCUSSION

Based on the results of tests conducted on the Wisconsin Breast Cancer dataset with a total of 699 data with 10 criterion features and 1 class feature, an increase in accuracy was obtained from the Naïve Bayes method by using feature selection using the relief-f method. While the naïve Bayes method did not increase after feature selection was carried out using a Gain ratio with a threshold of 0.15. The weight of each feature which is calculated using the gain ratio and relief-F can make it easier for decision-makers because the determined weights are not calculated manually but are calculated systematically and objectively. The test for calculating the accuracy of the Naïve Bayes classification without using feature selection obtained a result of 97.3646%, while the test carried

out for accuracy of Naïve Bayes with the feature selection gain ratio was 97.3646%, the accuracy of the classification of Naïve Bayes using the relief-f feature selection was 97.6574%. The increase in the level of accuracy is obtained from the Naïve Bayes method with the selection of relief-F features with an increased accuracy rate of 0.2928%.

4. CONCLUSION

Based on the results of testing the accuracy of the Naïve Bayes method without feature selection and with the gain ratio and F-relief feature selection on the Wisconsin Breast Cancer dataset, it can be concluded that the feature selection method can increase accuracy in Naïve Bayes. The increased accuracy of the Naïve Bayes classification method using Relief-F feature selection increased by 0.2928% from the accuracy of the Naïve Bayes method without using feature selection with an accuracy rate of 97.3646%.

REFERENCES

[1]. Seth, H. R., & Banka, H. (2016). Hardware implementation of Naive Bayes classifier: A cost effective technique. 2016 3rd International Conference on Recent Advances in Information Technology, RAIT 2016, 264–267.

[2]. Granik, M., & Mesyura, V. (2017). Fake news detection using naive Bayes classifier. 2017 IEEE 1st Ukraine Conference on Electrical and Computer Engineering, UKRCON 2017 - Proceedings, 900–903.

[3]. Devika, R., Avilala, S. V., & Subramaniyaswamy, V. (2019). Comparative study of classifier for chronic kidney disease prediction using naive bayes, KNN and random forest. Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019, Iccmc, 679–684.

[4]. Wahyuningsih, S., & Utari, D. R. (2018). Perbandingan Metode K-Nearest Neighbor, Naive Bayes dan Decision Tree untuk Prediksi Kelayakan Pemberian Kredit. Konferensi Nasional Sistem Informasi 2018, 619–623.

[5]. Samsani, S. (2017). An RST based efficient preprocessing technique for handling inconsistent data. 2016 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2016.

[6]. Zhang, X., Shi, Z., Liu, X., & Li, X. (2018). A Hybrid Feature Selection Algorithm For Classification Unbalanced Data Processsing. 2018 IEEE International Conference on Smart Internet of Things (SmartIoT), 269–275.

[7]. Rafei, N. S. I. M., Hassan, R., Saedudin, R. D. R., Raffei, A. F. M., Zakaria, Z., & Kasim, S. (2019). Comparison of feature selection techniques in classifying stroke documents. Indonesian Journal of Electrical Engineering and Computer Science, 14(3), 1244–1250.

[8]. Mohana C. P., & K., P. (2017). On Feature Selection Algorithms and Feature Selection Stability Measures: A Comparative Analysis. International Journal of Computer Science and Information Technology, 9(3), 159–168.

[9]. Alharan, A. F. H., Fatlawi, H. K., & Ali, N. S. (2019). A cluster-based feature selection method for image texture classification. Indonesian Journal of Electrical Engineering and Computer Science, 14(3), 1433–1442.

[10]. Yusra, R. N., Sitompul, O. S., & Sawaluddin. (2021). Kombinasi K-Nearest Neighbor (KNN) dan Relief-F Untuk Meningkatkan Akurasi Pada Klasifikasi Data. InfoTekJar: Jurnal Nasional Informatika Dan Teknologi Jaringan, 1, 0–5.