Dataset Weighting Features Using Gain Ratio To Improve Method Accuracy Naïve Bayesian Classification

1Novriadi Antonius Siagian, 2Sutarman Wage, 3Sawaluddin

Graduate Program of Computer Science, Faculty of Computer Science and Information Technology

E-mail: novriadi.antonius95@gmail.com, sutarman@usu.ac.id, sawal@usu.ac.id

Abstract. The Naïve Bayes method is proven to have a high speed when applied to large datasets, but the Naïve Bayes method has weaknesses when selecting attributes because Naïve Bayes is a statistical classification method that is only based on the Bayes theorem so that it can only be used to predict the probability of the class membership of a class independently. Independent without being able to do the selection of attributes that have a high correlation and correlation between one attribute with other attributes so that it can affect the value of accuracy. Naïve Bayesian Weight has been able to provide better accuracy than conventional Naïve Bayesian. Where an increase in the highest accuracy value obtained from the Water Quality dataset is equal to 88.57% in the Weight Naïve Bayesian classification model, while the lowest accuracy value is obtained from the Haberman dataset which is 78.95% in the conventional Naïve Bayesian classification model. The increase in accuracy of the Weight Naïve Bayesian classification model in the Water Quality dataset is 2.9%. While the increase in accuracy value in the Haberman dataset is 1.8%. If done the average accuracy of each dataset using the Weight Naïve Bayesian classification model is 2.35%. Based on the testing that has been done on all test data, it can be said that the Weight Naïve Bayesian classification model can provide better accuracy values than those produced by the conventional Naïve Bayesian classification model.

Keywords: Naïve Bayes, Weight Naïve Bayesian, Water Quality, Haberman, Python, Gain Ratio

1. Introduction

Naïve Bayesian Classifier (NBC) is a classification technique for calculating a set of probabilities by adding up the frequency and combination of values from a given dataset and giving strong (naïve) independence assumptions. Data classification is done by analyzing training samples and giving a categorization (giving class) to the data based on the predicted value of an attribute. Attributes are characteristics or characteristics of data. The expected prediction of an attribute is a value that produces an accurate category description [1-3]. To improve the classification accuracy of a predictor, the algorithm that can be developed is by giving weight to the attribute before the data classification stage [4].

The weighting of attributes from data samples before the classification stage. This aims to increase the accuracy of the conventional Naïve Bayesian method by applying attribute reduction through weighting the attributes of the data sample using the Information Gain method. The application of the Information Gain Weighted Naïve Bayesian Classifier (IGWNBC) method produces a significant correctness rate better than conventional Naïve Bayesian methods. Correctness rates generated from the car, zoo, and mushroom datasets have an average accuracy of over 97% [5-6].
Propose the weighting to the attributes of the data sample using the Multivariable Linear Regression model to produce a coefficient (weight coefficient) then classified using the Naïve Bayesian Classifier (NBC). The application of the MLRM and NBC models produces a value of accuracy (correctness rate) which is significantly better than conventional Naïve Bayesian methods. The correctness rate generated from the UCI Machine Learning dataset is 80% [7].

Propose local attribute weighting using the K-Nearest Neighbors Classifier (KNN) method. This method is carried out to produce a number of K nearest neighbors from the sample data and calculate the probability values of each attribute and then weight each attribute then classified using the Naïve Bayesian Classifier (NBC). The sample data used comes from historical bus route data and weather conditions data from August to December 2014. The accuracy obtained from the Naïve Bayesian Local Attribute Weighted KNN method was 89% [8].

2. Research Methods

2.1. Gain Ratio
C4.5 algorithm is a decision tree method where the attribute selection is based on Gain Ratio. Gain Ratio (GR) is a modification of Information Gain that reduces the bias. Determination of Gain Ratio is as follows:

1. Calculate the Entropy value for each attribute
   \[
   \text{Entropy}(S) = \sum_{i=1}^{n} -p_i \log_2 p_i
   \]

2. Calculate the information gain value for each attribute
   \[
   \text{InformationGain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^{|S|} \frac{|S_i|}{|S|} \times \text{Entropy}(S_i)
   \]

3. Calculate the Split Information value for each attribute
   \[
   \text{SplitInfo}(D) = -\sum_{j=1}^{r} \frac{|D_j|}{|D|} \times \log_2 \left( \frac{|D_j|}{|D|} \right)
   \]

4. Calculate the Gain Ratio value for each attribute
   \[
   \text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}(A)}
   \]

2.2. Naïve Bayesian Method (Naïve Bayesian Classifier)
Naïve Bayes is a classification that calculates a set of probabilities by adding up the frequency and combination of values from a given dataset. Naïve Bayes is based on the simplification assumption that attribute values are conditionally mutually independent if output values are given. The advantage of using Naïve Bayes is that this method only requires a small amount of training data to determine the estimated parameters needed in the classification process. Naïve Bayes often work far better in most complex real-world situations than expected. Following are the basic equations of the Bayes theorem [9]:

1. \[
P(C|X) = \frac{P(X|C)P(C)}{P(X)}
\]

2.3. Problems Identification
Naïve Bayes method still has a level of weakness when going to do attribute selection, because Naïve Bayes itself is a statistical classification method based only on the Bayes theorem so that it can only be used with the aim of predicting the probability of membership in a group or class. Therefore we need attribute weighting to increase accuracy more effectively.
3. Results And Discussion

To implement the performance of the proposed method, two sets of data are used, the Water Quality dataset and the Haberman dataset. The Dataset of Water Quality originates from the result of research [3], where the data is the result of data collection by the Ministry of Environment on Water Quality Provisions which are classified into four categories. The Haberman dataset comes from the KEEL-Dataset Repository with the urladdress: https://sci2s.ugr.es/keel/dataset.php?cod=62, where the Haberman data description is a number of breast cancer patient data that is predicted whether the patient these can survive for 5 years or more (Positive) or vice versa the patient dies within a period of 5 years (negative) after surgery [3].

The cleaning process is done by removing data duplication so that the number of Water Quality Status datasets which were originally 120 instances to 117 instances and the same thing also applies to the number of Haberman datasets which originally amounted to 306 instances to 289 instances.

3.1. Acquisition Result Weight Gain Ratio

The highest Gain Ratio for the Total Coliform and Pij attribute is 1. The lowest Gain Ratio is for the DO attribute, which is 0.257.

| Attribute            | Weight |
|----------------------|--------|
| TSS (mg/L)           | 0.441  |
| DO (mg/L)            | 0.257  |
| COD (mg/L)           | 0.325  |
| BOD (mg/L)           | 0.488  |
| Total phospat (mg/L) | 0.274  |
| Fecal Coliform (mg/L)| 0.624  |
| Total Coliform (mg/L)| 1      |
| Pij                  | 1      |

The outputs of weight gain result from the Gain Ratio process to the Haberman dataset can be seen in the following table:

| Attribute | Weight |
|-----------|--------|
| Age       | 0.219  |
| Year      | 0.005  |
| Positive  | 0.195  |

3.2. Accuracy Result of Conventional NaivesBayesian Models (Dataset Water)

The Water Quality Status dataset has 8 attributes, 4 classes and 120 instances, class distribution in the form of good condition (30 instances), lightly polluted (30 instances), medium polluted (30 instances) and heavily polluted (30 instances). The data is divided by 70% from the data will be used as training data and as much as 30% of the data will be used as test data conducted randomly.
3. Accuracy Result of Conventional Naïve Bayesian Models (Haberman Dataset)

Dataset of Haberman has 3 attributes, 2 classes and 289 instances. Data shared by 80% of the data will be used as training data and as much as 20% of the data will be used as test data. Result of Conventional Naïve Bayesian Accuracy (Haberman Dataset) are as follows:

\[
\text{Accuracy} = \frac{2 + 43}{2 + 43 + 1 + 11} = \frac{45}{57} = 0.7895 \times 100\% = 78.95\%
\]

\[
\text{Classification error} = \frac{1 + 11}{2 + 43 + 1 + 11} = \frac{12}{57} = 0.2105 \times 100\% = 21.05\%
\]

The output of the accuracy value produced by the Weight Naïve Bayesian model uses the Water Quality dataset. While the result of the performance and the classification of Weight Naïve Bayesian using the Confusion Matrix calculation table method:

| Actual Class | Predicted |
|--------------|-----------|
| Lightly Polluted | Predicted |
| Heavily Polluted | Predicted |
| Medium Polluted | Predicted |
| Good Condition | Predicted |

Table 3. Confusion Matrix Model Weight Naïve Bayesian (Water quality)
Actual. Lightly Polluted (0) & 12 & 0 & 0 & 0 \\
Actual. Heavily Polluted (1) & 2 & 6 & 1 & 0 \\
Actual. Medium Polluted (2) & 0 & 0 & 10 & 1 \\
Actual. Good Condition (3) & 0 & 0 & 0 & 3 \\

Continue to calculate the Accuracy value and the classification error level (Classification error) of the Weight Naïve Bayesian model using the Water Quality dataset, the following are the result of the calculation:

\[
\text{Accuracy} = \frac{12 + 6 + 10 + 3}{12 + 6 + 10 + 3 + 2 + 1 + 1} = \frac{31}{35} = 0.8571 \times 100\% = 88.57\% \\
\text{Classification error} = \frac{2 + 1 + 1}{12 + 6 + 10 + 3 + 2 + 1 + 1} = \frac{4}{35} = 0.1143 \times 100\% = 11.43\% 
\]

Table 4. Weight Naïve Bayesian Confusion Matrix model (Haberman)

| Performa Classification | Predicted Class | 
|-------------------------|----------------|
| Actual Class            |                |
| Actual. Negative (0)    | Predicted Negative (0) | 44 |
| Actual. Positive (1)    | Predicted Positive (1) | 11 | 2 |

Then proceed to calculate the value of Accuracy and the classification error rate (Classification Error) of the Weight Naïve Bayesian model using the Haberman dataset, the following result of the calculation:

\[
\text{Accuracy} = \frac{44 + 2}{44 + 2 + 11} = \frac{46}{57} = 0.8070 \times 100\% = 80.70\% \\
\text{Classification error} = \frac{11}{44 + 2 + 11} = \frac{11}{57} = 0.1143 \times 100\% = 19.30\% 
\]
Naïve Bayesian Weight has been able to provide better accuracy than conventional Naïve Bayesian. Where an increase in the highest accuracy value obtained from the Water Quality dataset is equal to 88.57% in the Weight Naïve Bayesian classification model, while the lowest accuracy value is obtained from the Haberman dataset which is 78.95% in the conventional Naïve Bayesian classification model. The increase in accuracy of the Weight Naïve Bayesian classification model in the Water Quality dataset is 2.9%. While the increase in accuracy value in the Haberman dataset is 1.8%. If done the average accuracy of each dataset using the Weight Naïve Bayesian classification model is 2.35%. Based on testing that has been done on all test data, it can be said that the Weight Naïve Bayesian classification model is able to provide better accuracy values than those produced by the conventional Naïve Bayesian classification model.

4. Conclusion
In the research that has been done, the authors produce several conclusions as follows:

1. In the research conducted on the Water Quality dataset and the Haberman dataset, it is done by producing predictions from the conventional Naïve Bayesian classification model and the Weight Naïve Bayesian classification model by doing the highest level of accuracy that is good.
2. Success in predicting using Naïve Bayesian Weight which is able to provide better accuracy than conventional Naïve Bayesian. Where an increase in the highest accuracy value obtained from the Water Quality dataset is equal to 88.57% in the Weight Naïve Bayesian classification model, while the lowest accuracy value is obtained from the Haberman dataset which is 78.95% in the conventional Naïve Bayesian classification model.
3. Increased accuracy of the Weight Naïve Bayesian classification model in the Water Quality dataset is 2.9%. While the increase in accuracy value in the Haberman dataset is 1.8%. If done the average accuracy of each dataset using the Weight Naïve Bayesian classification model is 2.35%.
4. The Weight Naïve Bayesian classification model is able to provide better accuracy values than those produced by the conventional Naïve Bayesian classification model.

REFERENCES
[1]. Amra, Ihsan & Maghari, Ashraf. 2017. Students Performance Prediction using KNN and Naïve Bayesian. 8th International Conference on Information Technology (ICIT). pp. 909-913.
[2]. Han, Xiaoyan., Xu, Liancheng., Ren, Min & Gu, Weiping. 2015. A Naïve Bayesian Network Intrusion Detection Algorithm Based on Principal Component Analysis. 7th International Conference on Information Technology in Medicine and Education. pp. 325-328.
[3]. Danades, A., Pratama, D., Anggraini, D., Anggriani, D. 2016. Comparison of Accuracy Level K-Nearest Neighbor Algorithm and Support Vector Machine Algorithm in Classification
Water Quality Status. *International Conference on System Engineering and Technology*, pp. 137-141.

[4]. Wang, Xingang & Sun, Xiu. 2016. An Improved Weighted Naïve Bayesian Classification Algorithm based on Multivariable Linear Regression Model. *9th International Symposium on Computational Intelligence and Design*. pp. 219-222.

[5]. Mao, Xin, Zhao, Gang & Sun, Ruoying. 2017. Naïve Bayesian Algorithm Classification Model with Local Attribute Weighted based on KNN. pp.

[6]. Han, J., Kamber, M. & Pei, J. 2012. *Data Mining: Concepts and Techniques*. 3rd Edition. Morgan Kaufmann Publishers: San Francisco.

[7]. Johnson, W.A. & Wichern, D.W. 2007. *Applied Multivariate Statistical Analysis*. 6th Edition. Pearson Prentice Hall: New Jersey.

[8]. Duan, Weil & Lu, Xiang. 2010. Weighted Naïve Bayesian Classifier Model Based on Information Gain. *International Conference on Intelligent System Design and Engineering Application*. pp. 819-822.

[9]. Repaka, A. N., Ravikanti, S. D. & Franklin, R. 2019. Design and Implementing Heart Disease Prediction using Naïve Bayesian. *International Conference on Trends in Electronics and Informatics (ICOEI)*. pp. 292-297.