Improving The Performance of K-Nearest Neighbor Algorithm by Reducing The Attributes of Dataset Using Gain Ratio

N Hasdyna¹*, B Sianipar¹* and E M Zamzami²

¹Graduate Program of Computer Science, Universitas Sumatera Utara, Medan, 20155, Indonesia.
²Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan 20155, Indonesia.

*Email: noviahasdyna@gmail.com, aniparbingin87@gmail.com, elvi_zamzami@usu.ac.id

Abstract. Data that has many attributes or higher dimensions will affect the performance of the K-NN classification algorithm. In this study, the Gain Ratio implemented for selecting and reducing the dataset attributes to form a new dataset for the classification process is carried out with the K-NN. The dataset used in this study are the Breast Cancer Coimbra dataset and Hepatitis C Virus dataset obtained from the UCI Machine Learning Repository. The results showed that the Breast Cancer Coimbra dataset, Gain Ratio can improve the performance of K-NN with average value $\text{TPR} = 0.535596$, $\text{TNR} = 1$, $\text{NPV} = 0.608279$, $\text{FNR} = 1$, $\text{FOR} = 0.391721$, $\text{Accuracy} = 72.85\%$. In Hepatitis C Virus dataset also managed to improve the performance of K-NN with average value $\text{TPR} = 0.665596$, $\text{TNR} = 0.876667$, $\text{NPV}=0.738279$, $\text{FNR}=0.88$, $\text{FOR}=0.521721$, and $\text{Accuracy}=86.25\%$.

1. Introduction

One of the Machine Learning classification algorithms is the K-Nearest Neighbor [1]. High-dimensional data can reduce the performance of K-NN [2]. Gain Ratio can be applied to reduce the high-dimensional data [3]. Gain Ratio is a feature selection method in machine learning that used in the Pre-Processing state [4]. In this study we implement the Gain ratio + K-NN methods in Breast Cancer Coimbra Dataset and Hepatitis C Virus Dataset. The focus of this study is to analyze the performance of the K-NN algorithm before the reduced of the attributes and after the reduced attributes using Gain Ratio.

2. Literature Review

2.1 Gain Ratio

Gain ratio is the development of information gain [5]. The steps of the gain ratio calculation are as follows.
2.1.1 Calculating Entropy

\[ E_{\text{entropy}}(s) = \sum_{i} - p_i \log_2 p_i \]  

\( C \) = The number of values in the target features (number of classes)
\( P_i \) = The ratio between the number of samples in class \( i \) against all the samples in the data set

2.1.2 Calculate the Value of Information Gain

\[ \text{Gain } (S, A) = \text{Entropy} - \sum_{v \in \text{values } (A)} \frac{|S_v|}{|S|} \text{entropy } (S_v) \]  

2.1.3 Calculate the Value of Gain Ratio. To calculate the required value of Gain Ratio split information, as in Equation 3.

\[ \text{SplitInfo}(S, A) = - \sum_{j=1}^{c} \frac{|S_j|}{S} \log_2 \frac{|S_j|}{S} \]  

Furthermore, Gain Ratio is calculated by Equation 4.

\[ \text{GainRatio} = \frac{\text{Gain } (S, A)}{\text{SplitInfo}(S, A)} \]  

2.2 Classification with K-Nearest Neighbor

K-Nearest Neighbor (KNN) is seeking the shortest distance between the data that will be evaluated by the closest K-Nearest Neighbor in the training data [6]. The steps of the method of use of K-NN:

- Determine the parameter of K
- Calculate the distance between the data to be evaluated with all the training, the study used Euclidean Distance, as in equation 5.

\[ d_i = \sqrt{\sum_{i=1}^{p} (x_{2i} - x_{1i})^2} \]  

- Sort range formed (the order of the value of the smallest to the largest value).
- Determine the shortest distance to the order of K
- Pair the corresponding class

2.3 Performance Analysis

To calculate performance classification algorithms can use the confusion matrix [7]. Some performance measures classification algorithm is as follows:

- Sensitivity, recall, hit rate or True Positive Rate (TPR)

\[ \text{TPR} = \frac{TP}{TP + FN} \]  

- Specificity, selectivity or True Negative Rate (TNR)

\[ \text{TNR} = \frac{TN}{TN + FP} \]
• Negative Predictive Value (NPV)

\[ \text{NPV} = \frac{TN}{TN + FN} \]  

(8)

• Miss Rate or False Negative Rate (FNR)

\[ \text{FNR} = \frac{FN}{FN + TP} \]  

(9)

• False Omission Rate (FOR)

\[ \text{FOR} = \frac{FN}{FN + TN} \]  

(10)

• Accuracy (ACC)

\[ \text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \]  

(11)

3. Methodology

Figure 1 shows the sequence of the attribute reduction process by using the Gain Ratio. The original dataset will be calculated using the gain ratio, they will be ranked the highest value to the lowest attributes. The attribute with the lowest value will be reduced and will form the new data. The new dataset will be processed with the K-Nearest Neighbor algorithm. After the classification with \( k = 3 \), they will be measured and analyzed the accuracy of the K-NN algorithm performance.

Figure 1. Block Diagram of Reduction the Attributes Dataset with Gain Ratio + KNN.
4. Result and Discussion

4.1 Description of Dataset
The dataset used two types of datasets, the dataset Coimbra Breast Cancer and Hepatitis C virus dataset obtained from the UCI Machine Learning Repository. Details of datasets in this study can be seen in Table 4.1

| Dataset            | Amount | Attributes | Class | Data types |
|--------------------|--------|------------|-------|------------|
| Breast Cancer Coimbra | 116    | 10         | 2     | integer    |
| Hepatitis C Virus   | 1385   | 29         | 4     | integer    |

4.2 Attributes Reduction in Breast Cancer Coimbra Dataset
The attribute reduction process with Gain Ratio is by calculating the entropy, information gain, Split Info and Gain Ratio. Gain Ratio results are shown in Table 4.2.

| No. | Attributes | Information | Gain  | Split Info | Gain Ratio  |
|-----|------------|-------------|-------|------------|-------------|
| 1   | Age        | 0.962037723 | 0.0302289 | 1.57837906 | 0.019151871 |
| 2   | BMI        | 0.980267275 | 0.0119994 | 1.56140271 | 0.00768499  |
| 3   | glucose    | 0.863920247 | 0.1283464 | 1.57974362 | 0.081245077 |
| 4   | Insulin    | 0.928129614 | 0.064137  | 1.484448855| 0.043205951 |
| 5   | HOMA       | 0.916172531 | 0.0760941 | 1.573889068| 0.048347822 |
| 6   | leptin     | 0.988534366 | 0.0037323 | 1.583578821| 0.003457725 |
| 7   | resistin   | 0.944827115 | 0.00768499| 1.56140271 | 0.003457725 |
| 8   | MCP.1      | 0.986791059 | 0.0054756 | 1.573889068| 0.009963614 |

The first reduction process is conducted to reduce one attribute with the smallest gain ratio value is 0.002357847. Having reduced the attribute, it will be processed with K-NN classification and analyzed its performance. The second reduction process, the smallest gain ratio value is 0.003457725 and new dataset after reduced performed with K-NN classification. Likewise with the third attribute reduction value is 0.00768499. Based on Table 4.2 can be seen in Figure 2.

![Figure 2. Graph of Gain Ratio Value In Breast Cancer Coimbra Dataset.](image-url)
4.3. Performance Calculation Results of Gain Ratio + KNN In Breast Cancer Coimbra Dataset
In this research, the performance measurement K-Nearest Neighbor algorithm with \( k = 3 \). The result of the calculation of the K-NN performance calculated by confusion matrix such as table below.

| Methods                                         | TPR     | TNR     | NPV     | FNR     | FOR     | ACCURACY  |
|------------------------------------------------|---------|---------|---------|---------|---------|-----------|
| K-NN                                           | 0.46546 | 1       | 0.6511  | 0.3489  | 0.7285  | 0.728571  |
| Gain Ratio, Reduction One Attribute + K-NN     | 0.51984 | 1       | 0.6234  | 0.3766  | 0.7285  | 0.728571  |
| Gain Ratio, Reduction Two Attributes + K-NN    | 0.52139 | 1       | 0.62    | 0.38    | 0.7285  | 0.728571  |
| Gain Ratio, Reduction Three Attributes + K-NN  | 0.56556 | 1       | 0.5814  | 0.4186  | 0.7285  | 0.728571  |
| Performance Reduction One, two and three reduction attributes | 0.53559 | 1       | 0.6083  | 0.3917  | 0.7285  | 0.728571  |

The average value of K-NN Performance Reduction One, two and three reduction attributes 0.66556 0.87666 0.738279 0.521721 0.862571

![Comparison of K-NN, Gain Ratio + K-NN in Breast Cancer Coimbra Dataset](image)

Figure 3. Comparison chart K-NN Performance Value, Gain Ration + K-NN In Breast Cancer Coimbra Dataset.

4.4. Performance Calculation Results of Gain Ratio + KNN In Hepatitis C Virus Dataset
The results of the calculation of the performance of K-NN without reduction attributes and Gain Ratio + K-NN with attribute reduction can be seen in a comparison table below.

| Methods                                         | TPR     | TNR     | NPV     | FNR     | FOR     | ACCURACY  |
|------------------------------------------------|---------|---------|---------|---------|---------|-----------|
| K-NN                                           | 0.59546 | 0.87    | 0.7811  | 0.4789  | 0.8206  | 0.820571  |
| Gain Ratio, Reduction One Attribute + K-NN     | 0.64984 | 0.88    | 0.7534  | 0.5066  | 0.8585  | 0.858571  |
| Gain Ratio, Reduction Two Attributes + K-NN    | 0.65139 | 0.88    | 0.75    | 0.51    | 0.8595  | 0.859571  |
| Gain Ratio, Reduction Three Attributes + K-NN  | 0.69556 | 0.87    | 0.7114  | 0.5485  | 0.8695  | 0.869571  |
| Performance Reduction One, two and three reduction attributes | 0.66559 | 0.87666 | 0.738279| 0.521721| 0.862571| 0.862571  |
The performance comparison chart K-NN can be seen below.

![Graph of Gain Ratio Values In Hepatitis C Virus Dataset](image_url)

**Figure 4.** Graph of Gain Ratio Values In Hepatitis C Virus Dataset.

5. Conclusion and Future Work

Based on the research that has been done on two different datasets, the Gain Ratio method can improve the performance of the K-Nearest Neighbor algorithm. In Breast Cancer Coimbra Dataset, K-NN after reduced can increase the value of TPR, FOR, and its accuracy is stable in 72.85%. In the Hepatitis C Virus dataset after reduced attribute, the KNN Performance value of TPR, TNR, FNR, FOR, and accuracy were increased from 82% to 85%. In the future scope, Gain Ratio can be a recommendation for other researchers in reducing the attribute data to improve the performance of an algorithm.

References

[1] Suguna N and Thanushkodi K 2010 An Improved K-Nearest Neighbor Classification Using Genetic Algorithm *International Journal of Computer Science Issues* Vol.7 No.2 pp. 18-21.

[2] Priyadarsini R P, Valarmathi M L and Sivakumari S 2011 Gain ratio Based Feature Selection Method for Privacy Preservation *ICTACT Journal on Soft Computing* Vol.1 No.4 pp. 201-205.

[3] Doshi M 2014 Correlation Based Feature Selection (CFS) Technique to Predict Student Performance *International Journal of Computer Networks & Communications* Vol.6 No.3 pp. 197.

[4] Karegowda A G Manjunath S and Jayaram M A 2010 Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection *International Journal of Information Technology and Knowledge Management* Vol.2 No.2 pp.271-277.

[5] Syaliman K U Nababan E B and Sitompul O S 2018 Improving The Accuracy of K-Nearest Neighbor Using Local Mean Based and Distance Weight *Journal of Physics: Conference Series* Vol. 978 No. 1 pp. 012047.

[6] Dinata R K Fajriana F & Khairunnisa K 2018 Penerapan Algoritma Classification and Regression Trees (CART) pada Penerimaan Anggota Baru Unit Kegiatan Mahasiswa (UKM) di Universitas Malikussaleh Berbasis Web *Techsi-jurnal teknik informatika* Vol.10 No.2 pp. 74-81.

[7] Cohen S Rokach L and Maimon O 2007 Decision-tree Instance-space Decomposition with Grouped Gain-Ratio *Information Sciences* Vol. 177 No.17 pp. 3592-3612.