Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio

A A Nababan¹, O S Sitompul²*, and Tulus³

¹Graduate School of Computer Science
²Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia
³Department of Mathematics, Universitas Sumatera Utara, Medan, Indonesia

E-mail: *opim@usu.ac.id

Abstract. K-Nearest Neighbor (KNN) is a good classifier, but from several studies, the result performance accuracy of KNN still lower than other methods. One of the causes of the low accuracy produced, because each attribute has the same effect on the classification process, while some less relevant characteristics lead to mis-classification of the class assignment for new data. In this research, we proposed Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio as a parameter to see the correlation between each attribute in the data and the Gain Ratio also will be used as the basis for weighting each attribute of the dataset. The accuracy of results is compared to the accuracy acquired from the original KNN method using 10-fold Cross-Validation with several datasets from the UCI Machine Learning repository and KEEL-Dataset Repository, such as abalone, glass identification, haberman, hayes-roth and water quality status. Based on the result of the test, the proposed method was able to increase the classification accuracy of KNN, where the highest difference of accuracy obtained hayes-roth dataset is worth 12.73%, and the lowest difference of accuracy obtained in the abalone dataset of 0.07%. The average result of the accuracy of all dataset increases the accuracy by 5.33%.

1. Introduction
K-Nearest Neighbor (KNN) is one of the effective, simple and performs well method for classification [1-3], but from several studies, the result performance accuracy of KNN is lower than other methods. One of them is in the study by [4] which compared performance between support vector machine (SVM) and KNN. The result of their research seen that performance of SVM better than KNN, where the value accuracy obtained by SVM of 82.54% while the value obtained from KNN of 79.22%. Another study by [5] compared K-Nearest Neighbor (KNN) and Artificial Neural Network (ANN). The results of their study seen that the performance of ANN (with 5 hidden layers) better than KNN, where the result of accuracy is 90.5%. Research by [6] compared the performance of Naïve Bayes, Decision Tree and K-Nearest Neighbor (KNN). The results show that Naïve Bayes has the best accuracy in classification compared to Decision Tree Tree and K-Nearest Neighbor (KNN) with an average accuracy of 73.7%, while the average accuracy of Decision Tree and KNN respectively 58.9% and 56.7%. In the study by [7] compared KNN and Naïve Bayes in diagnosing heart disease. Results obtained by the value of accuracy Naïve Bayes of 79.62% while the average value of accuracy on KNN is only 64.85% when k = 10.
From some research above seen that the accuracy of KNN method is still lower than other methods. One of the causes of the low value of accuracy produced, because each attribute has the same effect on the classification process. The solution to this problem is to give weight to each attribute [8]. In this research, we propose attribute weighting based K-Nearest Neighbor using Gain Ratio to increase the accuracy value of K-Nearest Neighbor (KNN) method by giving weight to each attribute [8].

In this research, we propose attribute weighting based K-Nearest Neighbor using Gain Ratio to increase the accuracy value of K-Nearest Neighbor (KNN) method by giving weight to each attribute, where the weights obtained are calculated using the normalization equation min-max, then calculated the data similarity using the distance model proposed by [1]. The rest of this paper is structured as follows. Section 2 will summarize previous studies on the theoretical foundation regarding the topic. In Section 3 we will provide the result and discussions and in Section 4 will provide with conclusions.

2. Feature Weight K-Nearest Neighbor (FWKNN)
K-Nearest Neighbor (KNN) is a nonparametric learning method and sensitive to distance function due to the inherent sensitivity of irrelevant attributes [1]. So to get FWKNN modeling based on attribute weighting, FWKNN determines the weight of the attribute by identifying the nearest k-neighbor that has the highest similarity with the test data. Therefore the FWKNN requires a model to measure the distance between test data and training sample data.

To measure the distance between data, one of the most popular options is the Euclidean distance model [1]. FWKNN uses a modified distance using the weighted feature for standard Euclidean distance expansion. The equations used in FWKNN are as follows:

$$d_w(X' - X) = \sqrt{\sum_{i=1}^{y} W_y \times (x'_i - x_i)^2}$$  \hspace{1cm} (1)

where $d_w(X' - X)$ is the Euclidean distance from the test data and training data, $y$ is the number of attributes, the attribute weights, $x'_i$ and $x_i$ the attribute values of test data and training data. The detail of FWKNN algorithm is as follows:

Step 1: Determine the weight of each feature
Step 2: Determine the value parameter $k$
Step 3: Calculate the distance using equation (1)
Step 4: Sorting results obtained by ascending (sequentially from high to low).
Step 5: Calculate the number of each class based on the $k$-nearest neighbor.
Step 6: The majority class is used as a class for test data.

FWKNN makes a difference to the characteristics by giving weight to its attributes, meaning that the more important attributes have a greater effect on distance determination [8]. Thus the error in the classification process can be reduced [1].

3. Gain Ratio
Algorithm C4.5 is a decision tree method wherein the selection of attributes based on Gain Ratio. Gain Ratio (GR) is a modification of Information Gain which reduces its bias. Gain Ratio improves the information gain by taking intrinsic information from each attribute [9]. The steps in determining the Gain Ratio are as follows:

Step 1: Calculate the value of Entropy on each attribute, with the equation:

$$\text{Entropy} (S) = \sum_{i=1}^{n} p_i \times \log_2 p_i$$  \hspace{1cm} (2)

where :

$S$ = Set of Cases
$n$ = Number of Partitions $S$
$p_i$ = Proportions of $S_i/S$

Step 2: Calculate the value of the information gain in each attribute by the equation:

$$\text{Information Gain} (S, A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \text{Entropy}(S_i)$$  \hspace{1cm} (3)

where :

$S$ = Overall Dataset
A = Subset Attribute  
|Si| = Subset Size of Dataset  
|S| = Subset Size of Dataset owned attribute on A i-partition

Step 3: Calculate the value of Split Information for each attribute with the equation below:

\[ SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2 \left( \frac{|D_j|}{|D|} \right) \]  

where:

- \( D \) = Overall Dataset  
- \( A \) = Subset Attribute  
- \( v \) = Number of Partition Attributes \( A \)  
- \( |D_j| \) = Subset Size of Dataset Owned attribute on \( A \) j-partition  
- \( |D| \) = Size of Number of Cases in Dataset

Step 4: Calculate the Gain Ratio of each attribute with the equation:

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

The Gain Ratio is found in the C4.5 algorithm, where the gain ratio is used to calculate the effect of attributes on the target of a data [11]. Gain Ratio is the development of information gain, where the gain ratio eliminates the bias value of each attribute.

4. Previous Research

Research work by [12] comparing K-Nearest Neighbor (KNN) with Support Vector Machine (SVM) for water quality status classification. The test results show the highest accuracy value found SVM, which is 92.40% while the average value of accuracy on KNN only 71.28% with \( k=7 \). The accuracy of KNN method is still lower than other methods. One of the causes of the low value of accuracy produced, because each attribute has the same effect on the classification process.

Weighting attribute has been proposed by [8] However, the weighting of attributes is less effective, because it gives the same weight for each of characteristics, while in research by [13], his research suggests consideration of KNN method by using different weighting functions.

In addition, attribute weighting is also performed on research work by [1]. in his research suggested the method of Information Gain as a basis for weighting attributes on the KNN algorithm known as Feature Weight K-Nearest Neighbor (FWKNN). This method is considered more effective than original KNN.

Finally, [14] in his research using Gain Ratio as the basic weighting attribute on KNN. From their research, it is seen that KNN by using Gain Ratio is considered more intuitive and easy to understand. The results obtained in the study, weighting attributes by using Gain Ratio able to increase the highest accuracy of 5%.

5. The Proposed Method

To describe the proposed method, it will be explained step by step in this chapter. The stages can be seen in figure 1:

![Figure 1. Detail stages of the proposed method.](image)

From figure 1 it can be explained that the proposed method has several stages, among others:

Step 1: Weighting Attributes by Gain Ratio

The weight is calculated using the normalization equation min-max [10], where the lowest weight after normalized is 0.1 and the highest weight after normalized is 1. The equation used is as follows:
\[ W_i = \frac{(G_i - \text{Min}(G))}{\text{Max}(G) - \text{Min}(G)} \times (0.9) + 0.1 \]  

(6)

where:
- \( W_i \) = Weight of the \( i \)-Attribute
- \( G_i \) = The Gain ratio \(-i\)
- \( \text{Min}(G) \) = The Lowest Value of The Gain Ratio
- \( \text{Max}(G) \) = The Highest Value of The Gain Ratio

Step 2: Classification with KNN

Furthermore, the process of classification with KNN, after the weight of each attribute is obtained it will be classified by KNN. In determining the proximity between data, each attribute will have different effects based on the weight value of the attribute using the equation (1).

This research will use the gain ratio as a parameter to see the correlation between each attribute in the data and the gain ratio also will be used as the basis for weighting each attribute of the data set. The higher the gain ratio of an attribute the correlation to the data class will be greater, so the weight of the attribute is also higher. This research will give analysis result to the performance of KNN method with weighting attribute using gain ratio on the performance of KNN method conducted by [11].

6. Result and Discussion

This research uses data sets obtained from UCI Machine Learning Repository and KEEL-Dataset Repository such as abalone, glass identification, haberman, hayes-roth and one dataset of water quality status derived from research conducted by [11]. The details of the data used can be seen in table 1:

| Data                     | Attributes | Type                      | Class | Total Data |
|--------------------------|------------|---------------------------|-------|------------|
| Abalone                  | 8          | Categorical, Integer, Real| 3     | 4177       |
| Glass Identification     | 10         | Real                      | 6     | 214        |
| Haberman                 | 3          | Real, Integer, Nominal    | 2     | 306        |
| Hayes-roth               | 4          | Real, Integer, Nominal    | 3     | 160        |
| Water Quality Status     | 8          | Integer, Real             | 4     | 120        |

6.1 Testing Using Data Set

This study used a 10-fold cross validation on datasets, so the total dataset was randomly divided into 10 datasets with the same comparison values. Then do 10 trials, where each trial using a data partition to-kas the testing of data and utilize the remaining more partitions as training data. The average accuracy of each data can be seen in figure 2:
Figure 2. The average result of the accuracy of the dataset.

The figure 2 shows the KNN using Gain Ratio has a higher accuracy value than the original KNN, where the highest difference of accuracy obtained hayes-roth data set is worth 12.73%, and the lowest difference of accuracy obtained in the abalone dataset of 0.07%. The average result of the accuracy of all dataset increases the accuracy of 3.79%.

6.2 Testing Using Water Quality Status
To know the weighting of attributes using gain ratio can improve the accuracy of KNN, then the next test will be conducted by using the water quality status datasets derived from research by [11]. The result of comparison between KNN and KNN with Gain Ratio can be seen in table 2.

| k   | Comparison Of Accuracy (%) | Difference of Accuracy |
|-----|-----------------------------|------------------------|
|     | KNN     | KNN+Gain Ratio |                      |
| 1   | 87.5%   | 94.2%         | 6.7%                  |
| 2   | 87.5%   | 92.5%         | 5%                    |
| 3   | 85.8%   | 95%           | 9.2%                  |
| 4   | 85.8%   | 90%           | 4.2%                  |
| 5   | 85.8%   | 92.5%         | 6.7%                  |
| 6   | 87.5%   | 87.5%         | 0%                    |
| 7   | 82.5%   | 90.8%         | 8.3%                  |
| 8   | 83.3%   | 88.3%         | 5%                    |
| 9   | 83.3%   | 89.2%         | 5.9%                  |
| 10  | 83.3%   | 85.8%         | 2.5%                  |
|     | Average | 85.23%        | 90.58%                |
|     |          |               | 5.35%                 |

Results of testing using KNN and KNN with Gain Ratio classification contained in table 2. The table shows original KNN with the highest difference of accuracy grade are in grades $k = 1, 2, 6$ that is 87.5% and the lowest difference of accuracy grade are in grade $k = 7$ that is 82.5%, while the results KNN with Gain Ratio is obtained the highest difference of accuracy grade are in grades $k = 3$ that is 95% and the lowest difference of accuracy grade are in grade $k = 10$ that is 85.8%. The comparison of the accuracy of KNN and KNN with Gain Ratio shows that the average result of the accuracy of all $k$ increases the accuracy by 5.35% against accuracy value of original KNN.

7. Conclusion
Based on the findings and discussion that has been described previously, it can be concluded as follows:
1. The test using 10-fold cross validation proved that Attribute weighting based K-Nearest-Neighbor using Gain Ratio has a higher accuracy value than the original KNN, where the highest difference of accuracy obtained hayes-roth data set is worth 12.73%, and the lowest difference of accuracy obtained in the abalone dataset of 0.07%. The average result of the accuracy of all dataset increase the accuracy by 4.09%.
2. The test results weighting of attributes using gain ratio on the water quality dataset can improve the accuracy of KNN with the highest difference of accuracy grade are in grades $k = 3$ that is 9.2% and the average result of the accuracy of all $k$ increase the accuracy by 5.35% against accuracy value of original KNN.
3. KNN is a good classifier but when we implement this method over textual data (nominal data) it’s all performance parameters are varied according to the size of the dataset. KNN performs poor results as the size of data set increases it is the best fit for a small dataset.
References

[1] Chen, Y., Hao, Y. 2017. A Feature Weighted Support Vector Machine and K-Nearest Neighbor Algorithm for Stock Market Indices Prediction. Expert Systems With Applications (2017), pp. 340-355.

[2] Song, Y., Liang, J., Lu, J. & Zhao, X. 2017. An Efficient Instance Selection Algorithm For K Nearest Neighbor Regression. Neurocomputing, pp. 26-34.

[3] Suguna, N., Thanushkodi, K. 2010. An Improved k-Nearest Neighbor Classification Using Genetic Algorithm. ICJSI International Journal of Computer Science Issues, pp. 18-21.

[4] Raikwal, J.S., Saxena, K. 2012. Performance Evaluation of SVM and K-Nearest Neighbor Algorithm over Medical Data set. International Journal of Computer Applications, pp. 35-39.

[5] Moosavian, A., Ahmadi, H., Tabatabaeefar, A., Khazaei, M. 2012. Comparison of two classifiers; K-nearest neighbor and artificial neural network, for fault diagnosis on a main engine journal-bearing. Shock and Vibration, pp. 263-272.

[6] Ashari, A., Paryudi, I., Tjou, A.M. 2013. Performance Comparison between Naïve Bayes - Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an En: Simulation Tool. (IJACSA) International Journal of Advanced Computer Science & Applications, pp. 33-39.

[7] Meiyar, V.K., Shanmugasundaram, D. 2015. The Comparative Study for Diagnosing Heart Disease Using KNN and Naïve Bayes. (IJARCSMS) International Journal Advanced Research in Computer Science and Management Studies, pp. 9-19.

[8] Kuhkan, M. 2016. A Method to Improve the Accuracy of K - Nearest Neighbor Algorithm. International Journal of Computer Engineering and Information Technology, pp. 90-95.

[9] Priyadarsini, R.P., Valarmathi, M.L., Sivakumari, S. 2011. Gain Ratio Based Feature Selection Method For Privacy Preservation. ICTACT Journal on Soft Computing, pp. 201-205.

[10] Saranya, C & Manikandan, G. 2013. A Study on Normalization Techniques for Privacy Preserving Data Mining. International Journal of Engineering and Technology (IJET), pp. 2701-2704.

[11] Mitchell, T.M. 1997. Machine Learning. McGraw-Hill Science/Engineering/Math; (March 1, 1997). ISBN: 0070428077.

[12] 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. DOI 10.1109/FIT.2016.7857553.

[13] Demidova, L., Sokoloba, Y. 2017. A Novel SVM-kNN Technique for Data Classification. 2017 6th Mediterranean an Conference on Embedded Computing (MECO), pp. 11-15.

[14] Duneja, A., Puyalnithi, T. 2017. Enhancing Classification Accuracy of K-Nearest Neighbours Algorithm Using Gain Ratio. International Research Journal of Engineering and Technology (IRJET), pp. 1385-1388.