Gini Index With Local Mean Based For Determining K Value In K-Nearest Neighbor Classification

M E Saputra¹, H Mawengkang², E B Nababan³

¹,²Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia
²Faculty of Computer Science and Mathematics, Universitas Sumatera Utara, Medan, Indonesia
Corresponding Author : ernabrn@usu.ac.id

Abstract. A process that explains and differentiates the data class is called Classification. The nearest neighbor is calculated based on the distance of each data, especially to determine the k value in the data. To fix K-Nearest Neighbor, it is necessary to test data class and train with Local Mean Based K-Nearest Neighbor using the closest distance measurement of Manhattan to each local mean of each data class. Gini Index is used in the process of calculating each weight in the data attribute. In this research, Gini Index, K-Fold Cross Validation and Local Mean are needed in the K-Nearest Neighbor classification. In Iris data the lowest k value is k=1, k=48, k=49, and k=50 accuracy is 94.67%, while the highest k value is k=12 and k=13 accuracy is 97.33%. So the result of the highest k value becomes the best k value in this study. Likewise with the Ionosphere data the lowest k value of k=50 accuracy is 86.92%, while the highest k value is k=2 with accuracy 92.89%, the Ionosphere data is the best k=2 and two Voice Gender and Lower Back data.

1. Introduction

K-Nearest Neighbor is one classification tool using all training samples in classifications that cause high level of computational complexity. where the nearest neighbor is calculated based on the k value in determining how many closest neighbors should be considered to decide the class of the sample data point, correcting the K-Nearest Neighbor based on the weight of the k value. Training is given weight according to distance from the sample data points, but computational complexity and memory remain a major concern [1-3].

Classification is part of data extraction, where data mining is used to explain knowledge discovery in data. K-Nearest Neighbor (KNN) is a way and regulation but very effective non parametric technique on classification patterns, but classification performance depends on the value of k [8-10]. k is used in each class, which can cause local high sensitivity with k values. If k is too small useful classification information may not be enough, while large k values can easily cause outliers including in the closest neighbors of the true class [5-7].

K-Nearest Neighbor goal is where the nearest neighbor is calculated based on the k value, which is the nearest neighbor to decide the data point class based on the determination of the weight of the k value [2-4]. Training points are given weight according to distance from data points. To further increase data points that do not affect the results of the training data set. In addition to the limitations of time and memory, another thing to consider is the value of k based on unknown sample categories. To choose the value of k by using the nearest neighbor. The model proposed to choose k values as well as many improvements are proposed to decide the value of k using the concept of rank [9,10]. The closest neighbor is located in the first class where the entire data is classified into training data and sample data points. The distance evaluated from all training points to the sample point with the distance of the nearest neighbor.

Class determination Local Mean Based K-Nearest Neighbor test data uses the closest distance measurement to each local mean of each data class, because this way can access effectively to overcome the negative effects of outlier effects [16]. showed that the combination of LMKNN and DWKNN was...
able to improve the classification accuracy of the KNN, where the average accuracy of the test data was 2.45% with the highest accuracy improvement of 3.71% accuracy that occurred in the lower back pain symptoms data set. For this data, accuracy increases were obtained as high as 5.16% [12].

Gini Index can be considered as the probability of two randomly selected data that have different classes. Gini Index is a method that has a good performance because it can cut the influence of outliers from handling the ability to measure divergence data and measure data impurity. Selecting the k value using a Gini Index model-based approach is needed to decide the value of k. In the first class where the entire data is classified into training data and test data. The distance is evaluated from all training facts to the test data at the lowest distance [1-17].

2. Problem

Based on the introduction above, in determining the k value for the K-Nearest Neighbor classification at the average point of the Outlier variable. The writer wants to overcome this problem, in determining the k value, the K-Nearest Neighbor method is used which is an effective method to be used in classification. The problem is that if the k value is too small it can cause insufficient information and if a large k value easily results in outliers. To overcome this problem, it is necessary to Gini Index process that is able to measure divergence and measure impurities in the attribute data also use Local Mean Based in determining good k value.

3. Manhattan k Value K-Nearest Neighbor

To find the closest distance between the data evaluated with k in the training data. Using calculation equations to find distances with Manhattan equations 1.

\[ d(x, y) = ||x - y||_1 = \sum_{i=1}^{n} |x_i - y_i| \]  \hspace{1cm} (1)

Sort remote results obtained where:
- d: Distance
- \( x_i \): data samples tested
- \( y_i \): test data
- \( n \): the number of features in the vectors of data

The steps in the classification of K-Nearest Neighbor are:
- a. Determine the k value.
- b. Calculate proximity based on Manhattan's distance model on the training data provided.
- c. Sort the results of the distance obtained descending
- d. Calculate the number of each class based on the nearest k neighbor.
- e. The highest k value is used as the best k value.

3.1. Local Mean Based K-Nearest Neighbor (LMKNN)

Local Mean K-Nearest Neighbor is a simple, effective and tough non parametric classification. It is proven to improve classification performance and cut the influence of outliers and in the size of a small amount of data [5-12]. Pay attention to the figure 1.

![Figure.1 Nearest Neighbors From Each Class](image)

This stage is a contribution from the Local Mean Based K-Nearest Neighbor (LMKNN) method. The value of k on LMKNN is very different from the value of k in conventional K-NN [17], where in conventional K-NN the k value is the number of closest neighbors of all sample data, while in LMKNN the k value is the number of closest neighbors of each sample data class [10-12].
As for the work steps in LMKNN in determining the k value, it is necessary to calculate the distance of the test data throughout the data from each data class by using the Manhattan distance model followed by sorting the distance between data from the smallest to the largest by k of each class. Calculate the local mean of each class with equation 2.

\[ m_{wj}^k = \frac{1}{k} \sum_{i=1}^{k} y_{ij} \]  
\[ \text{Local Mean} \]

Determine the test data class by calculating the closest distance to the local mean of each data class with equation 3.

\[ w_c = \arg \min_{w_j} d(x, m_{wj}^k), j = 1, 2, ..., M \]  
\[ \text{Minimum Distance} \]

Local Mean Based K-Nearest Neighbor k value is the number of closest neighbors selected from each class in training data.

### 3.2 Gini Index

Gini Index is generally used in Classification And Regression Trees (CART) and SPRINT Algorithms which say the size of how randomly selected objects are from training data. The size of the impurity reaches 0 when only 1 class is present at a point. But the opposite will reach the greatest when the class size at that point is balanced.

Class differences we can divide S into n subset (S_i, i = 1...n). Suppose S_i is a sample set included in class P_i, then the Gini Index of set S in equation 4.

\[ Gini(S) = 1 - \sum_{i=1}^{n} p_i^2 \]  
\[ \text{Gini Index} \]

Where P_i is the probability of each sample. The least Gini (S) is 0, that is, all members in that place are in the same class.

It can also be used to compare the advantages and disadvantages of the weight selection function in the equation 5.

\[ Gini(t) = 1 - \sum_{i,j=1}^{n} p(j|t)p(i|t), i \neq j \]  
\[ \text{Gini Index} \]

Where:
\[ p(j|t) = \text{proportion of class j on the symbol of t} \]
\[ p(i|t) = \text{proportion of class i on the symbol of t} \]

### 3.3 K-Fold Cross Validation

K-Fold Cross Validation is an alternative method of cross validation or cross validation which in each data is used in the same amount for training and is right for one test. Suppose the dataset is broken into two parts the same size. Part one for training data and one for test data, this approach is called two-fold cross validation. A special form of this method when k is set to k=N in the amount of data in the data set. Can also be called leave-one-out, that is the test dataset is only 1 data, while the training process is done as many as N times. The advantage is that almost everything in the dataset can be processed so that it gets an exact k value.

### 4. Methodology

In this study several steps will be taken to achieve the research objectives. The first step is to use the Gini Index Algorithm to adjust the k value with four main steps:

- a) Data in Peroses is data Iris, ionosphere, Voice Gender, and Lower Back data.
- b) Measurement of data weight based on the Gini Index on the equation 5.
- c) Divide Training data and Test data with 10-fold cross validation.
- d) Calculate the distance of the training facts to the test data and the distance of the test facts to the local mean using the distance Manhattan equation.1 with the Local Mean Based K-Nearest Neighbor.
- e) The highest k value becomes the best k value.
This step can be continued with the overall drawing of the study through the flow diagram of Figure 2.

![Flow Diagram](image)

**Figure 2. step to get the best k value**

The description of the research design can be explained by measuring the weight of each attribute based on the Gini Index. Determination of k value based on the distance of data with the process of k-fold cross validation and Local Mean Based K-Nearest Neighbor in the classification with K-Nearest Neighbor to get good results in accordance with the design in the study.

5. Results and Discussion

The iris dataset is one of the popular datasets and is good data. With data totaling 150 records with 4 attributes and consisting of 3 data classes. In this test, the training data amounted to 135 and the test data amounted to 15 data. By sharing data using 10-Fold Cross Validation on Iris data. Furthermore, the calculation of the distance between the training data and the test data using the Manhattan distance model using equation.1. As for the distance generated can be seen directly the data that has been sorted ascending, the order of the closest distance between data can be seen in table 1.

| Data Test | Sort the shortest distance |
|-----------|---------------------------|
| Test 1    | 2.293 0.137 0.088 0.076 0.069 ... 0.057 |
| Test 15   | 0.366 0.300 0.284 0.245 0.208 ... 0.161 |

For the Local Mean and Gini Index processes in the classification K-Nearest neighbor will calculate the average weight value of the nearest distance of k nearest k for each data class, and make the highest average weight as the k value of the test data class. Results obtained from Local Mean and Gini Index process can be seen in table 2 which is already in descending order.

| value of K | 12 | 13 | 11 | 14 | 15 | 16 | 17 | 18 | ... | 50 |
|-----------|----|----|----|----|----|----|----|----|-----|----|
|           | 0.973 | 0.973 | 0.966 | 0.966 | 0.966 | 0.966 | 0.966 | 0.966 | ... | 0.946 |

This graph is produced from the Gini Index process, 10 fold cross validation and Local Mean in the classification of the nearest neighbor. The weighting of each class with the Gini Index and dividing the 10-fold Cross Validation data are continued using the local mean to find the closest distance between the test data and the training data so that the gain can be seen in Figure. 3.

![Graph](image)

**Figure 3. Graph the results of the k value is good the Iris Dataset**
Not only iris data in this study, but there are some data about balanced data analysis data and unbalanced data, such as the ionosphere, Voice Gender, and Lower Back data. Each of them obtained a good k value in this processing, but each process must have its drawbacks and the results in this study still need further research, especially on unbalanced data. Data acquisition ionosphere The lowest k value k=50 accuracy 86.92%, while the highest k value k=2 accuracy 92.89%, Voice Gender data The lowest k value k=49 accuracy 91.41%, while the highest k value k=4 accuracy 94.22% and lower back data Lowest k value k=1 accuracy 67.41%, while highest k value K=6 accuracy 74.83% The results of the data can be seen in Figure 4.

![Graph](image)

Figure 4. Graph the results of the value of k is good on the data balance and unbalance

6. Conclusion
After the process of determining k values and getting good results. Judging from the joint graph, the results of the k value of the balance data such as Iris and Voice Genre are good data in this study compared to Unbalance Lower Back and ionosphere data. That each highest k value appears with a sudden high-rise, then the value of k will decrease and rarely will the value of k rise again due to the distance between the different data and the distance between the data. The farther distance of the data and each class, the Mean Based local process will be difficult to choose the closest distance in each class in the K-Nearest neighbor classification.

Acknowledgement
The great thank you given to Lembaga Penelitian Universitas Sumatera Utara (LP USU), the Graduate School of Computer Science Fasilkom-TI USU and rector of the Universitas Sumatera Utara has support this research can work with good and the facilities provided

References
[1] A A Nababan, O S Sitompul, and Tulus Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio IOP Conf. Series: Journal of Physics: Conf. Series 1007 (2018) 012007
[2] Bhatia, N. & Vandana., 2010. Survey of Nearest Neighbor Techniques. International Jurnal of Computer Science and Information Security (IJCSIS) 8(2): 302-305.
[3] Buana, P.W., Jannet. S.D.R.M., & Putra, I.K.G.D. 2012. Combination of K-Nearest Neighbor and K-Means based on Term Re-weighting for Classify Indonesian News. International Journal of Computer Applications 11(11):37-42.
[4] 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) 80: 340-355.

[5] 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.

[6] Gou, J. & Xiong, T. 2011. A Novel Weighted Voting for K-Nearest Neighbor Rule. Journal of Computer 6(5): 833-840.

[7] Gou, J., Zhang, Y., Rao, Y., Shen, X., Wang, X. & He, W. 2014. Improved Pseudo Nearest Neighbor Classification. Knowledge-Based Systems 70: 361-375.

[8] Hosein Alizadeh, Behrouz Minaei-Bidgoli and Saeed K. Amirgholipour A New Method for Improving the Performance of K Nearest Neighbor using Clustering Technique Journal of Convergence Information Technology Volume 4, Number 2, June 2009.

[9] Pan, Z., Wang, Y. & Ku, W. 2016. A New K-Harmonic Nearest Neighbor Classifier Based On The Multi-Local Means. Expert Systems With Applications 67: 115-125.

[10] Pan, Z., Wang, Y. & Ku, W. 2017. A New General Nearest Neighbor Classification Based On The Mutual Neighborhood Information. Knowledge-Based Systems 121: 142-152.

[11] Priyadarsini, R.P., Valarmathi, M.L., Sivakumari, S. 2011. Gain Ratio Based Feature Selection Method For Privacy Preservation. IJCTAC Journal on Soft Computing 1(4): 201-205.

[12] K U Syaliman, E B Nababan, and O S Sitompul Improving the accuracy of k-nearest neighbor using local mean based and distance weightInternational Conference on Computing and Applied Informatics (2017) Journal of Physics: Conf. Series 978 (2018) 012047

[13] Tyas Setiyorini, Rizky Tri Asmono Penerapan Gini Index Dan K-Nearest Neighbor Untuk Klasifikasi Tingkat Kognitif Soal Pada Taksonomi Bloom Jurnal Pilar Nusa Mandiri Vol. 13, No. 2 September 2017

[14] Wenqian Shang, Youli Qu, Haibin Zhu, Houkuan Huang, Yongmin Lin and Hongbin Dong An Adaptive Fuzzy kNN Text Classifier Based on Gini Index Weight Proceedings of the 11th IEEE Symposium on Computers and Communications (ISCC’06) 0-7695-2588-1/06 $20.00 © 2006 IEEE

[15] Y Mitani and Y Hamamoto 2006 A local mean-based nonparametric classifier Patern Recognition Letter pp 1151-115

[16] Zhibin Pan, Yidi Wang, Weiping Ku A new k-harmonic nearest neighbor classifier based on the multi-local means. Cina Expert Systems With Applications (2016).

[17] Zuguang Hu Bing XuLingling Pan, Shangfeng Zhang Juying Zeng The Dynamic KNN Clustering of Undergraduate Consumption With Gini Coefficient: A Case of Zhejiang 0-7695-2882-1/07 $25.00 ©2007 IEEE