Performance of distance-based k-nearest neighbor classification method using local mean vector and harmonic distance

Dedi Candro Parulian Sinaga[1,4], Tulus[2,5], Poltak Sihombing[3,6]

[1]Student of Computer Science, University of North Sumatra, Medan, Indonesia
[2]Department of Mathematic, University of North Sumatra, Medan, Indonesia
[3]Department of Computer Science, University of Nort Sumatera, Medan, Indonesia

[4]dedisinaga27@gmail.com, [5]tulus_jp@yahoo.com, [6]poltakhombing@yahoo.com,

Abstract. K-Nearest Neighbor was one of the top ten algorithms data mining in the classification process. The low accuracy results in the K-Nearest Neighbor classification method was caused of this method used the system of majority vote which allowed the selection of outliers as the closest neighbors and in the distance model used as a method of determining similarity between data. In this process it is evident that local mean vector and harmonic distance can improve accuracy, where the highest increase in average accuracy obtained in the set data wine is equal to 6.29% and the highest accuracy increase for LMKNN is obtained in set data glass identification which is 16.18%. Based on the tests that had been conducted on all data sets used, it could be seen that the proposed method was able to provide a better value of accuracy than the value of accuracy produced by traditional K-Nearest Neighbor and LMKNN.

1. Introduction

The K-Nearest Neighbor method was first method which was introduced in the early 1950s. K-Nearest Neighbor was one of the lazy learning classification methods which was the most widely used in classification, pattern recognition, text categorization. Providing a solution to these weaknesses was done by replacing the traditional distance models that used a distance model based on similarity and feature value similarity features. In this study, the writer suggested to use a distance model harmonic as a substitute for the distance model Euclidean. Determination of the test data class Local Mean Based K-Nearest Neighbor used the measurement of the closest distance to each one using the distance eucllidean from each data class.

In addition, K-Nearest Neighbor worked by looking at the nearest K neighbor of each data where in the traditional K-Nearest Neighbor classification process uses the system voting most as the prediction class of the new data. The selection of a small K-Nearest Neighbor value caused the classification of noise or outliers to be sensitive, if the value of K is too large the number of closest neighbors may be too large, which could ultimately reduce the classification results. This study aimed to improve the accuracy of traditional K-Nearest Neighbor by using local mean vector as a class for new data using the distance model Harmonic in the process of calculating similarities between data

2. Problems

Based on the introduction above, it was necessary to increase the accuracy of the classification K-Nearest Neighbor at the variable average point. The results of the accuracy of

the traditional K-Nearest Neighbor classification method were caused because this method used the system majority vote which allowed the selection of outliers as the closest neighbors, and in the distance model used as a method of determining similarity between data, where traditional distance models were very fragile to similarity calculations . These things could increase errors in the classification process. This study used Local Mean Based K-Nearest Neighbor and Harmonic Distance to improve accuracy on the method K-Nearest Neighbor.

3. Distance Euclidean K-Nearest Neighbor

Traditional distance models were very fragile in determining the similarity. Moreover in traditional distance models, the value of attributes which were too large, it could cover the influence of other attributes, and most traditional distance models lack the difference between data, especially in large data samples.

In this research, the writer suggested to use a distance model Harmonic, where the distance model was considered better in describing the similarities between data.

$$D(x,y) = \frac{1}{\sum_{j=1}^{N} \frac{1}{|x-y|})} \quad (1)$$

The main idea of the distance model *Harmonic* was to take the average number of harmonics from the distance *Euclidean* between one particular data point to the point of another group of data. Compared to other distance models, distance of *Harmonic* was more focus on the influence of the closer data.

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

This method was classified as a simple, effective and resilient method. Stating the use of Local Mean was proven to improve performance and also to reduce the influence of outliers on traditional K-Nearest Neighbor methods, especially for small amounts of data.
The workflow of the LMKNN was as follows:

Determining the K Value, then calculated the distance of the test data throughout the data from each data class by using the distance model Euclidean. Classifying the distance data between the data from the smallest to the largest K from each class. Calculating the local mean vector of each class with the equation:

$$m_{w_j}^{k} = \frac{i}{k}\sum_{i=1}^{k} \quad y_{i,j}^{NN} \quad (2)$$

Determining the test data class by calculating the closest distance to the local mean vector of each data class with the equation:

$$w_c = argmin_{w_j} d\left(x, m_{w_j}^{k}\right), j = 1,2 \ldots M \quad (3)$$

Explaining the K value on LMKNN was very different from K- Traditional NN. LMKNN as the value of K was the number of closest neighbors of each data class, whereas in traditional K-Nearest Neighbor, the value of K was the number of closest neighbors of all data. LMKNN was equal to 1-NN if K value was 1

3.2 Classification

Classification was a process of assessing objects to include them in a particular class based on the characteristics possessed by that object. Knowing the amount of data that has

been successfully classified correctly could be seen from the level of accuracy and rate error of the prediction results in the classification system. Calculation of the level of accuracy could be seen from the equation below:

$$Accuracy = \frac{Amount\ of\ data\ is\ predictable\ right}{Amount\ of\ Prediction\ do} \quad (4)$$

As for measuring the rate of error used the equation:

$$The\ rate\ of\ error = \frac{Amount\ of\ data\ Predictable\ Wrong}{Amount\ of\ Prediction\ do} \quad (5)$$

All classification algorithms tried to create models with high accuracy (rate error low). The model which was built generally could predict the training data correctly, but when the model was evaluated with the test data then the performance of the classification model, surely it could be seen clearly.

4. Methodology

This study used a combination of several stages in Local Mean Based K-Nearest Neighbor and Harmonic Distance as a label for the test data. It was expected that by using a combination of the two methods can improve the accuracy of K-NN.
The general description of the stages of the method proposed in this study was shown in Figure 1

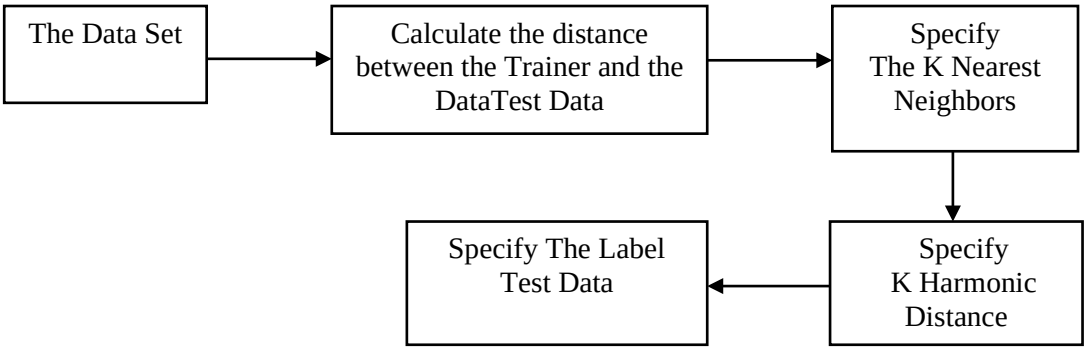

Figure 1. General architecture of the proposed method

Based on Figure 1, it could be seen that the proposed method had several stages, including:

i. Data set. In this process, the used data would be divided into 85% of the data which would be used as training data and 15% would be used as test data.
ii. Calculate the distance between training data and test data with *Euclidean*.
iii. Determine the nearest K neighbor, on the LMKNN the nearest neighbor was taken from each class of data. Whereas in traditional K-NN, the determination of the nearest K neighbor was taken from all data. In this process, the proposed method would follow the rules of the LMKNN.
iv. Specify the *HarmonicDistance* from each data class with Harmonic as determination of Labels for data test. Labels for test data were determined based on the value of the *Harmonic Distance*; the smaller value could indicate the similarity of closer data.

5. Results and Discussion

A dataset with 8 data records which showed that the data had 3 attributes and 2 classes. 85% of the data was used as training data and 15% was used as test data. The details of the dataset could be seen in table 1.

Table 1. Details data

No	X1	X2	X3	Class	Information
1	85	85	85	1	Training Data 1
2	87	73	70	1	Training Data 2
...	...	...	...	...	...
8	75	78	70	2	Test Data

After the Data were trained and data test was determined, then the classification process would be carried out by using the proposed method, LMKNN, and traditional K-NN. The first step in the classification process on the proposed method was to determine the K value, assuming the K value used was 2. Then, calculated the distance between the training data and the test data using *Euclidean*.

$$D(Test\ Data, Data\ Latih\ 1) = \sqrt{(75-85)^2 + (78-85)^2 + (70-85)^2}$$
$$D(Test\ Data, Data\ Latih\ 1) = \sqrt{374} = 19.34$$

Did this similar way for all other training data. The next step was to determine the nearest K neighbor from each data class. Next calculate the value of the *harmonic distance.* There were Harmonics for each class of data. The values *harmonic distance* of each data class could be seen in table 2.

Table 2. Harmonic distance to each class

Data	Class	*Harmonic Distance*
Test	1	12:37
	2	9:34

Stages in determining the class with the data test in combination of LMKNN and *Harmonic Distance* were to make the grade with values *Distance Harmonic* which showed that the highest as a class for the tested data. The highest value in the test data was found by class 2, so the tested data was in class 2

The first step in the LMKNN method was to determine the K value, in the previous sub-section K values were assumed to be 2, then calculated the distance of the test data to all training data by using the distance model *Euclidean.* The next stage was to sort the ascending distance as much as K for each class, at this stage 2 closest training data to the test data for each class will be sorted.

The next step was to calculate *local mean vector* for each data class, then calculate the distance of the test data to each local mean vector with *Euclidean.* The last step in LMKNN was to make *Local Mean Vector* from the closest class as a class for the test data. *The local mean vector* closest was found by class 2, so class 2 is used as a new class for the test data.

There was way to see clearly the average of the accuracy values found in each method for all data used in this study can be seen in Figure 2.

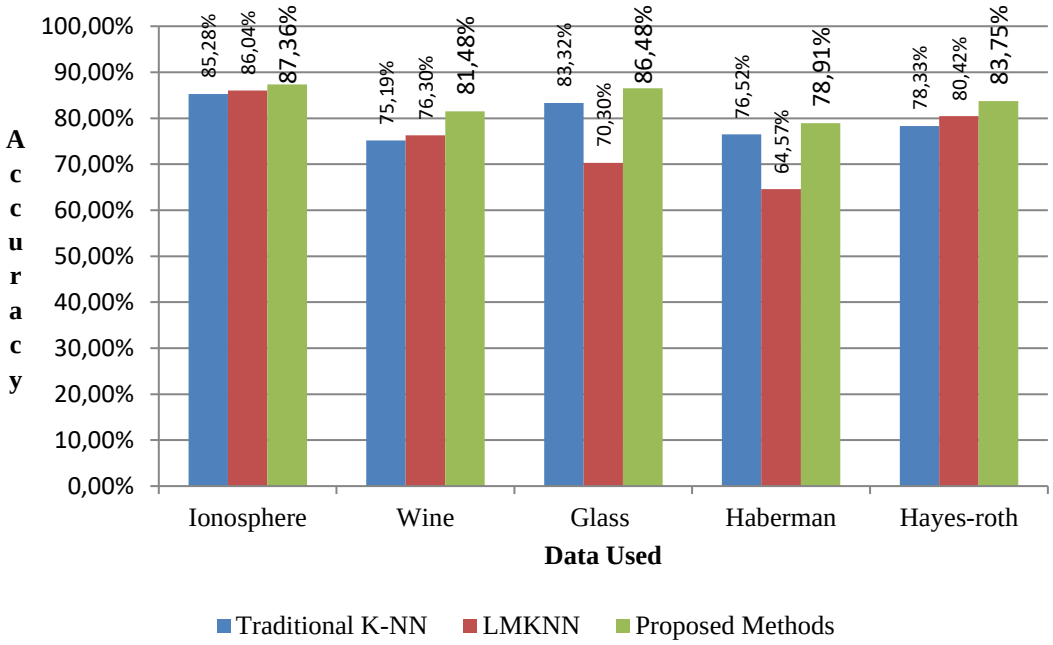

Figure 2. Graph of average accuracy values from all data

It could be seen that the proposed method was able to provide a value of better accuracy than traditional K-Nearest Neighbor and LMKNN. improving where the highest accuracy value to the traditional K-Nearest Neighbor found in the data set ionosphere that is equal to 6:29% and an increase in the highest accuracy of the method was found on dataset LMKNN toward glass identification that was equal to 16:18%. The lowest accuracy value increased between the methods proposed before traditional K-NNs of 2.08% and 1.32% for LMKNN, both of which were found in the set data ionosphere. The increase in the average accuracy value of all datasets used was 3.87% for traditional K-Nearest Neighbor and 8.07% for LMKNN.

6. Conclusion

While the lowest increase in average accuracy of conventional K-Nearest Neighbor was obtained at the data set, ionosphere which amounted to 2.08% for conventional K-Nearest Neighbor and 1.32% for LMKNN. The average increase in accuracy obtained from the entire dataset was 3.87% for conventional K-Nearest Neighbor and 8.07% for LMKNN. Based on the tests that had been carried out in the previous chapter, it could be concluded that local mean vectors and harmonic distances can improve accuracy in all data sets used.

7. Acknowledgment

The writer gave thank you greatly to the Research Institute of the University of Sumatra North (LP USU), the Graduate School of Computer Science at the USU Fasilkom-IT and rector of the University of North Sumatra, has s upported this research.

8. References

[1] García-Pedrajas, N. & Ortiz-Boyer, D. 2009. Boosting K-Nearest Neighbor Classifier By Means Of Input Space Projection. Expert System With Application 37(7): pp.10570-10582.

[2] Gou, J., Yi. Z., Du. L. &Xiong, T. 2012. A Local Mean-Based k-Nearest Centroid Neighbor Classifier. The Computer Journal 55(6): pp. 1058-1071.

[3] Han, J., Kamber, M. & Pei, J. 2011. Data Mining: Concepts and Techniques. 3rd Edition.Morga Kaufmann: Amsterdam.

[4] Iswarya, P. &Radha, V. 2015. Ensemble learning approach in Improved K Nearest Neighbor algorithm for Text Categorization. 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1-5.

[5] Jabbar, MA, Deekshatulu, BL & Chandra. P. 2013. Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm. International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA), pp. 85-94.

[6] Jo, T. 2017. Using K Nearest Neighbors for Text Segmentation with Feature Similarity. International \ Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE), pp. 1-5.

[7] Kalaivani, P. & Shunmuganathan, KL 2014. An Improved K-Nearest-Neighbor Algorithm Using genetic Algorithm For Sentiment Classification. 2014 International Conference on Circuit, Power and Computing Technologies (ICCPCT), pp. 1647-1651.

[8] Kataria, A. & Singh, MD 2013. A Review of Data Classification Using K-Nearest Neighbor Algorithm. International Journal of Emerging Technology and Advanced Engineering 3(6): 354-360.

[9] Kuhkan, M. 2016. A Method to Improve the Accuracy of K - Nearest Neighbor Algorithm. Internatonal Journal of Computer Engineering and Information Technology 8(6): 90-95.

[10] Lei, Z., Wang, S. & Xu, D. 2016. Sub-cellular Localization Protein Based on Noise-Intensity-Weighted Linear Discriminant Analysis and Improved K-Nearest-Neighbor Classifier. 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI 2016), pp. 1871-1876.

[11] Lidya, SK, Sitompul, OS & Efendi, S. 2015. Sentiment Analysis in Indonesian Language Texts Using Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN). National Seminar on Information and Communication Technology 2015, pp. 1-8.

[12] Loohach, R. & Garg, K. 2012. Effect of Distance Functions on K-Means Clustering Algorithm. International Journal of Computer Applications 49(6): 7-9.Prasetyo, E. (2014). Data Mining-Processing Data Into Information Using MATLAB. Yogyakarta: ANDI.

[13] Mitani, Y. & Hamamoto, Y. 2006. A Local Mean-Based Nonparametric Classifier. Patern Recognition Letter 27(10): 1151-1159.

[14] Nababan, A.A., Sitompul, O.S., &Tulus. April 2018. Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio. 2018. IOP Conf. Series: Journal of Physics: Conf. Series 1007 012007.

[15] Ougiaroglou, S. &Evangelidis, G. 2012. Fast and Accuratek-Nearest Neighbor Classification using Prototype Selection by Clustering. Panhellenic Conference on Informatics, pp. 168-173.

[16] Pan, Z., Wang, Y. & Ku, W. 2016. A New K-Harmonic Nearest Neighbor Classifer Based On The Multi-Local Means. Expert Systems With Applications 67: 115-125.
[17] Pandit, S. & Gupta, SA 2011. Comparative Study on Distance Measuring Approaches for Clustering. International Journal of Research in Computer Science 2(1): pp. 29-31.
[18] Prasetyo, E. (2014). Data Mining-Processing Data Into Information Using MATLAB. Yogyakarta: ANDI.
[19] Rui-Jia, W. & Xing, W., 2014. RWR / ESM Recognition in Airborne Emitter Radar Based on Improved K Nearest Neighbor Algorithm. 2014 IEEE International Conference on Computer and Information Technology (CIT), pp. 148-151.
[20] Sánchez, USA, Iglesias-Rodríguez, FJ, Fernándes, PR &Juez, FJde.C. 2015. Applying The K-Nearest Neighbor Technique To The Classification Of Workers According To Their Risk Of Suffering Musculoskeletal Disorders. International Journal of Industrial Ergonomics 52: 92-99.
[21] Saputra. M E., Mawengkang H., Nababan E B., Determination Value of K in Nearest Nieghbor Wit Local Mean Euclidean And Weight Gini Index. 3nd International Conference on Computing and Applied Informatics 2018 012098.
[22] Shirkhorshidi, AS, Aghabozorgi, S. &Wah, TY 2015. A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. PLUS ONE, pp. 1-20.
[23] Shaliman. KU, Sitompul. OS &Nababan, EB 2017. Improving The Accuracy Of K-Nearest Neighbor Using Local Mean Based And Distance Weight. 2nd International Conference on Computing and Applied Informatics 2017. Pp. 1-5.
[24] Song, Y., Liang, J., Lu, J. & Zhao, X. 2017. An Efficient Instance Selection Algorithm For K Nearest Neighbor Regression. Neurocomputing 251: 26-34.
[25] Wang. J., Neskovic. P. & Cooper LN, 2007. Improving Nearest Neighbor Rule With A Simple Adaptive Distance Measure. Pattern Recognition Letter 28: 207-213.
[26] Zheng, K. Si, G. Diao, L. Zhou, Z. Chen, J. & Yue W., 2017. Applications Of Support Vector Machine And Improved-Nearest Neighbor Algorithm In Fault Diagnosis and fault Degree Evaluation Of Gas Insulated Switchgear. 1st International Conference on Electrical Materials and Power Equipment - Xi'an - China, pp. 364-368.