Improving the Accuracy of Features Weighted k-Nearest Neighbor using Distance Weight

K. U. Syaliman[1], Ause Labellapansa[2], Ana Yulianti[2]
[1]*Department of Informatics, Politeknik Caltex Riau, Pekanbaru, Indonesia*
[2]*Department of Informatics, Universitas Islam Riau, Pekanbaru, Indonesia*

Keywords: Accuracy, Distance Weight, FWk-NN, K-NN, Vote Majority

Abstract: FWk-NN is an improvement of k-NN, where FWk-NN gives weight to each data feature thereby reducing the influence of features that are less relevant to the target. Feature weighting is proven to be able to improve the accuracy of k-NN. However, the FWK-NN still uses the majority vote system for class determination to new data. Whereby the majority vote system is considered to have several weaknesses, it ignores the similarity between data and the possibility of a double majority class. To overcome the issue of vote majority at FWk-NN, the research will change the voting majority by using distance weight. This study uses a dataset obtained from the UCI repository and a water quality data set. The data used from the UCI repository are iris, ionosphere, hayes-Roth, and glass. Based on the tests carried out using UCI repository dataset it is proven that FWk-NN using distance weight has averaged an increase about2%, with the highest increase of accuracy of 4.23% in the glass dataset. In water quality data, FWk-NN using distance weight can achieve an accuracy of 92.58% or has increased 2% from FWk-NN. From all the data tested, it is proven that the distance weight is able to increase the accuracy of the FWk-NN with an average increase about 1.9%.

1 INTRODUCTION

k-Nearest Neighbor or commonly known as kNN is one of the popular classification methods for dealing with problems in the field of mining data, including text categorization, pattern recognition, classification, etc (Bhatia and Vandana, 2010; Jabbar et al., 2013; Rui-Jia and Xing, 2014; Sánchez et al., 2016; Zheng et al., 2017). This is because kNN has advantages including simple methods, quite interesting, easy to implement, intuitive, can be exploited in various domains, and is quite efficient (Wang et al., 2007; Garca-Pedrajas and Ortiz-Boyer, 2009; Ougiaroglou and Evangelidis, 2012; Feng et al., 2016; Pan et al., 2017; Sánchez et al., 2016; Song et al., 2017).

kNN still has weaknesses that make the results of accuracy remain relatively low, even more so when compared with other classification algorithms. (Danades et al., 2016; Tamatjita and Mahasta, 2016). The low accuracy value of kNN is caused by several factors. One of them is because each feature has the same effect on determining the similarity between data. The solution is to give weight to each data feature or commonly called Feature Weight k-NN (Kuhkan, 2016; Duneja and Puyalnithi, 2017; Nababan et al., 2018).

FWk-NN is proven to improve the accuracy of the kNN method. It can be seen in the research conducted by Duneja (2017) and Nababan, et al (2018) which gives weights for each data feature using the Gain Ratio. In determining the class for new data, FWk-NN still adopts the votes system, where the majority vote system ignores the similarity between data, and another problem is the possible emergence of a double majority class(Gou and Xiong, 2011; Yan et al., 2015; Syaliman et al., 2017).

The solution to the majority vote system problem has been done by Mitani et al. (2006) . In this research, it was proposed to make a method change in class determination for new data, initially used the voting majority to be exchanged using local mean, so the class for new data is no longer based on the majority class, but is determined based on the similarity of the local mean vector.The results of this research proved that the local mean was able to reduce misclassification caused by the vote majority system.

Another solution to overcome the weaknesses in the vote majority system is to use the method proposed by Batista & Silva (2009). In this research it is recommended to use a distance weight while to

Syaliman, K., Labellapansa, A. and Yulianti, A.
Improving the Accuracy of Features Weighted k-Nearest Neighbor using Distance Weight.
DOI: 10.5220/0009390903260330
In *Proceedings of the Second International Conference on Science, Engineering and Technology (ICoSET 2019)*, pages 326-330
ISBN: 978-989-758-463-3

determine the new data class is based on the weight of the distance between the data, it has proved that it is able to overcome the problem in the majority vote system which ignored the similarity between data (Gou and Xiong, 2011; Syaliman et al., 2017).

Based on previous studies, the authors see that the accuracy of the FWk-NN method can be improved, where to improve the accuracy of FWk-NN, in this research the author will replace the vote majority system with a distance weight system. It is expected that using distance weight is able to increase the results of the classification.

2 FEATURE WEIGHTED K-NN (FWK-NN)

FWk-NN is a method developed to overcome problems in kNN that are sensitive to distance functions because of the sensitivity inherent in irrelevant features. FWk-NN is based on feature weighting (Chen and Hao, 2017).The details of the FWk-NN algorithm are as follows :

Step 1: Compute the weight of each feature using the Gain Ratio. (Duneja and Puyalnithi, 2017; Nababan et al., 2018).

Step 2: Determine the value of k, k is the number of nearest neighbor (Syaliman and A., 2015).

Step 3: Compute distance using equations :

$$D(x-y) = \sqrt{\sum_{i=1}^{f} fw_i \times (x_i - y_i)^2} \tag{1}$$

where $D(x-y)$ is the euclidean distance from x and y, f is the number of features, fw is the weight of the features.

Step 4: Sort the distance between data from the smallest to the largest (ascending) depend on the number of k.

Step 5: Compute the number of each class based on the nearest neighbor k.

Step 6: Make the majority class a new data class.

FWk-NN gives each feature a different weight, where features that have a greater influence on the class will be given a feature weight greater than the weight of other features. Thus the less relevant weight can be reduced by its influence (Kuhkan, 2016).

3 DISTANCE WEIGHTED K-NN (DWK-NN)

DWk-NN is also one of the improvements of k-NN.Improve to the DWk-NN were carried out to overcome the problem of the vote majority system from k-NN (Lidya et al., 2015).In k-NN, each nearest neighbor has the same influence in class determination for new data, this is considered irrational when viewed based on the similarity between data (Pan et al., 2016). The details of the FWk-NN algorithm are as follows:

Step 1: Determine the value of k

Step 2: Compute distance using equations :

$$D(x-y) = (\sum_{i=1}^{f} (x_i - y_i)^r)^{\frac{1}{r}} \tag{2}$$

$D(x-y)$ is the distance between x and y, f is the number of features, r is lambda value (a positive integer). r = 1 is known as Manhattan / City Block distance, r = 2 is known as Euclidean distance and if r = infinity is known as Chebyshev distance (Merigó and Casanovas, 2008; Labellapansa et al., 2016; Koteswara Rao, 2012).

Step 3: Sort the distance between data from the smallest to the largest (ascending) depend on the number of k

Step 4: Compute the weight of the distance between data using equation (Batista and Silva, 2009):

$$dw = \frac{1}{d(x,y)} \tag{3}$$

Step 5: Compute the average weights each data class based on closest k neighbors using the equation (4).

$$sum\_w_c = \sum_{i=1}^{k^{NN}} w_i, (c = c_i^{NN}) \tag{4}$$

Step 6: Select the class with the highest average weight value, then make it as a class for new data.

The workflow of DWk-NN is quite similar to k-NN. In K-NN class determination is based on majority vote while in DWk-NN uses the highest number of average distance weight values between data.

4 PROPOSE METHOD

To further described the changes made to FWk-NN using the distance weight will explain step by step in this sub-chapter. The stages are in figure 1.

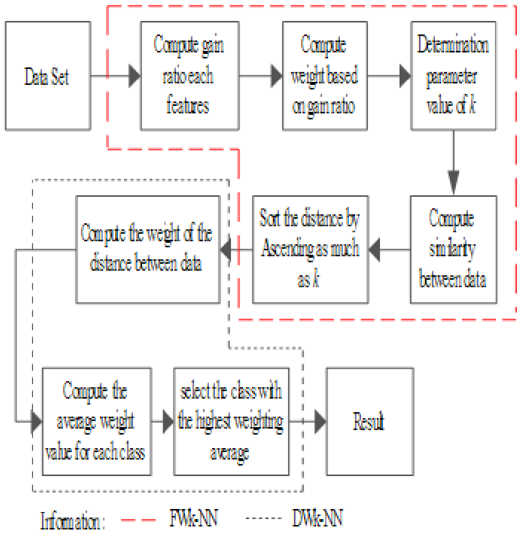

Figure 1: Proposed Method

From figure 1, the modified Feature Weight K-Nearest Neighbor (FWk-NN) and Distance Weight have several stages,which are :

Step 1: Compute influence of features by using Gain Ratio

Step 2: Compute the weight based on the gain ratio using equation (5)

$$fw_i = \frac{(G_i - Min(G))}{Max(G) - Min(G)} \times 1 \quad (5)$$

where fw_i is features weight-i, G_i is Gain Ratio-i, $Min(G)$ is the minimum gain ratio, and $Max(G)$ is the maximum gain ratio.

Step 3: Determine the value of k

Step 4: Compute distance using equations (1).

Step 5: Sort the distance between data from the smallest to the largest (ascending) based on the number of k

Step 6: Compute the weight of the distance between data sorted by equation(3).

Step 7: Calculate the average weight for each class based on the nearest neighbor kusing equation(4).

Step 8: Select the class with the highest average weight value, then make it as a class for new data.

Step 1 to 5 is the contribution from FWk-NN, while step 6 to step 8 are the steps of the distance weight to determine the class for new data.

5 RESULT AND DISCUSSION

This research uses several datasets from the UCI Machine Learning repository, such as ionosphere, Haberman, hayes, glass, and iris. In addition, the proposed method is also tested using real data from a water quality status of Indonesia (Danades et al., 2016). The detail of the data can be seen in table 1.

Table 1: Detail of Data

Data	Features	Class	Total Data
Ionosphere	34	2	351
Iris	4	3	150
Hayes	4	3	160
Glass	10	6	214
Water Quality Status	8	4	120

In this study used 10-fold cross-validation, and the value of k is only worth 1 to 10. The average accuracy of each data can be seen in figure 2.

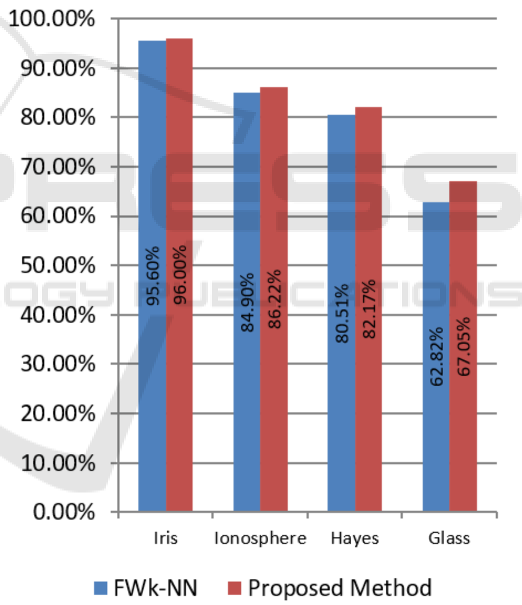

Figure 2: Accuracy from Dataset

Based on figure 2,the proposed method or FWk-NN using distance weight has higher accuracy than FWk-NN, where the highest improved of accuracy obtained in Glass dataset is worth 4.23%, and the lowest improved of accuracy obtained in the iris dataset of 0.4%. From all dataset, the accuracy increase is 2%.

Based on the testing by using the dataset from UCI was knowing, the proposed method is better than the original FWk-NN. To know with certainty whether the proposed method is better to make predictions in the real data from water quality status in Indonesia, it will be compared with original FWk-NN. Details of the test results can be seen in table 2.

Table 2: Comparison of Accuracy

k	Accuracy		Best Method
	FWk-NN$^{(1)}$	Proposed Method$^{(2)}$	
1	94.20%	96.67%	(2)
2	92.50%	96.67%	(2)
3	95.00%	95.83%	(2)
4	90.00%	93.33%	(2)
5	92.50%	93.33%	(2)
6	87.50%	90.83%	(2)
7	90.80%	89.17%	(1)
8	88.30%	90.00%	(2)
9	89.20%	90.00%	(2)
10	85.80%	90.00%	(2)
Avg	90.58%	92.58%	

Based on table 2, the proposed method gives the best prediction results in determining of Data. Although when the value of *k* is 7, the accuracy of the proposed method is decreased by 1.63%, overall the proposed method was able to improve the accuracy worth 2%, whereby the highest difference of accuracy is 4.20% when *k* is 10.

6 CONCLUSIONS

Referring result and discussion in the previous chapter can be concluded that distance weights can improve the accuracy of FWk-NN. Based on the test, the highest accuracy is obtained at about 4.23% in the glass data.Distance weights also have proven to be successful in improving accuracy on water quality status data. The highest accuracy occurs when k is ten by 4.2% with the average increase is 2%.In all tests that have been carried out, it has proven that the distance weights applied to FWk-NN provide better accuracy results than the majority vote system with the average accuracy of all data used is 1.9%.

ACKNOWLEDGMENTS

High appreciation should be given to Universitas Islam Riau (UIR) especially Department of Informatics for their support on the dissemination of this research work as well as the facilities provided.

REFERENCES

Batista, G. and Silva, D. F. (2009). How k-nearest neighbor parameters affect its performance. *Argentine Symposium on Artificial Intelligence*, pages 95–106.

Bhatia, N. and Vandana (2010). Survey of nearest neighbor techniques. *International Jurnal of Computer Science and Information Security (IJCSIS)*.

Chen, Y. and Hao, Y. (2017). *A Feature Weighted Support Vector Machine and K-Nearest Neighbor Algorithm for Stock Market Indices Prediction*. Expert Systems With Applications : 1-29.

Danades, A., Pratama, D., Anggraini, D., and Anggraini, 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.

Duneja, A. and Puyalnithi, T. (2017). *Enhancing Classification Accuracy of K-Nearest Neighbors Algorithm Using Ga in Ratio*. International Research Journal of Engineering and Technology (IRJET).

Feng, Y., Jian-Chang, L., and ming L., D. (2016). *An Approach for Fault Diagnosis Based on an Improved k-Nearest Neighbor Algorithm*. Control Conference (CCC).

Garca-Pedrajas, N. and Ortiz-Boyer, D. (2009). *Boosting K-Nearest Neighbor Classifier By Means Of Input Space Projection*. Expert System With Application37(7):.

Gou, J. and Xiong, T. (2011). *A Novel Weighted Voting for K-Nearest Neighbor Rule*. Journal of Computer6(5): 833-840.

Jabbar, M. A., Deekshatulu, B., and Chandra, P. (2013). Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm. *Procedia Technology*, 10:85–94.

Koteswara Rao, M. (2012). Face Recognition Using Different Local Features with Different Distance Techniques. *International Journal of Computer Science, Engineering and Information Technology*, 2(1):67–74.

Kuhkan, M. (2016). A method to improve the accuracy of k - nearest neighbor algorithm. *Internatonal Journal of Computer Engineering and Information Technology*.

Labellapansa, A., Efendi, A., Yulianti, A., and Kadir, E. A. (2016). *Lambda Value Analysis on Weighted Minkowski Distance Model in CBR of Schizophrenia Type Diagnosis*. Fourth International Conference on Information and Communication Technologies (ICoICT).

Lidya, S. K., Sitompul, O. S., and Efendi, S. (2015). Sentiment analysis pada teks bahasa indonesia menggunakan support vector machine (svm) dan k-nearest neighbor (k-nn). In *Seminar Nasional Teknologi Informasi dan Komunikasi*. (In Bahasa).

Merigó, J. M. and Casanovas, M. (2008). The induced minkowski ordered weighted averaging distance operator. In *Proceedings of the ESTYLF Conference*, pages 35–41.

Mitani, Y. and Hamamoto, Y. (2006). *A Local Mean-Based Nonparametric Classifier*. Patern Recognition Letter27(10): 1151-1159.

Nababan, A. A., Sitompul, O. S., and Tulus (2018). Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio. *Journal of Physics: Conference Series*, 1007(1).

Ougiaroglou, S. and Evangelidis, G. (2012). *Fast and Accuratek-Nearest Neighbor Classification using Prototype Selection by Clustering*. Panhellenic Conference on Informatics.

Pan, Z., Wang, Y., and Ku, W. (2016). *A New K-Harmonic Nearest Neighbor Classifer Based On The Multi-Local Means*. Expert Systems With Applications67:115-125.

Pan, Z., Wang, Y., and Ku, W. (2017). A new general nearest neighbor classification based on the mutual neighborhood information. *Knowledge-Based Systems*, 121:142–152.

Rui-Jia, W. and Xing, W. (2014). *Radar Emitter Recognition in Airborne RWR/ESM Based on Improved K Nearest Neighbor Algorithm*. Conference on Computer and Information Technology (CIT).

Sánchez, A. S., Iglesias-Rodríguez, F. J., Fernández, P. R., and de Cos Juez, F. (2016). 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.

Song, Y., Liang, J., Lu, J., and Zhao, X. (2017). *An Efficient Instance Selection Algorithm For K Nearest Neighbor Regression*. Neurocomputing251:26-34.

Syaliman, K. U. and A., L. (2015). *Analisa Nilai Lamda Model Jarak Minkowsky Untuk Penentuan Jurusan SMA (Studi Kasus di SMA Negeri 2 Tualang)*. Jurnal Teknik Informatika dan Sistem Informasi (JuTISI) 1(2) : 163-171 (In Bahasa).

Syaliman, K. U., Sitompul, O. S., and Nababan, E. B. (2017). *Improving the accuracy of k-nearest neighbor using local mean based and distance weight*. 2nd International Conference on Computing and Applied Informatics 2017 (ICCAI).

Tamatjita, E. N. and Mahasta, A. W. (2016). *Comparison Of Music Genre Classification Using Nearest Centroid Classifier And K-Nearest Neighbors*. International Conference on Information Management and Technology (ICIMTech).

Wang, J., Neskovic, P., and Cooper, L. N. (2007). Improving nearest neighbor rule with a simple adaptive distance measure. *Pattern Recognition Letters*, 28(2):207–213.

Yan, Z. H. C., L., Y., and Zhao, J. T. (2015). *Using Improved K-nearest Neighbor Method to Identify Anti- and Pro-apoptosis Proteins*. 8th International Conference on BioMedical Engineering and Informatics (BMEI 2015).

Zheng, K. S., Diao, G., Zhou, L., Chen, Z., J., and W., Y. (2017). *Applications Of Support Vector Machine And Improvedk-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.