Improving the accuracy of k-nearest neighbor using local mean based and distance weight

Improving the accuracy of k-nearest neighbor using local mean based and distance weight

K U Syaliman[1], E B Nababan[2*], and O S Sitompul[2]

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

ernabrn@usu.ac.id

Abstract. In k-nearest neighbor (kNN), the determination of classes for new data is normally performed by a simple majority vote system, which may ignore the similarities among data, as well as allowing the occurrence of a double majority class that can lead to misclassification. In this research, we propose an approach to resolve the majority vote issues by calculating the distance weight using a combination of local mean based k-nearest neighbor (LMKNN) and distance weight k-nearest neighbor (DWKNN). The accuracy of results is compared to the accuracy acquired from the original k-NN method using several datasets from the UCI Machine Learning repository, Kaggle and Keel, such as ionosphare, iris, voice genre, lower back pain, and thyroid. In addition, the proposed method is also tested using real data from a public senior high school in city of Tualang, Indonesia. Results shows that the combination of LMKNN and DWKNN was able to increase the classification accuracy of kNN, whereby the average accuracy on test data is 2.45% with the highest increase in accuracy of 3.71% occurring on the lower back pain symptoms dataset. For the real data, the increase in accuracy is obtained as high as 5.16%.

1. Introduction

The k-NN method is among the most widely used method in data mining and machine learning research, such as text categorization, pattern recognition, and classification (see for example the works of [1-8]). The kNN is recognized as an attractive, easy to apply, intuitive, simple and could be exploited in various application domains. However, it is found that the accuracy of kNN is still relatively low, especially when compared to other classification algorithms. In [9] the authors measured the accuracy of support vector machine (SVM) and kNN to obtain an average accuracy of kNN and SVM equal to 71.28% and 92.40%, respectively. Another study by [10] compared nearest centroid classifier (NCC) and kNN method. The result of their research revealed that NCC reach a highest accuracy of 96.67% and a lowest accuracy of 33.33%, whereas the kNN method was only capable to produce a highest accuracy of 26.7% and a lowest accuracy of 22.5%. A recent study by [11] found that the kNN method gave a best result of 48.78% with k = 8 when applied on a dataset which has 395 records, 30 attributes, and 4 classes.

The relatively low accuracy of kNN is caused by several factors. One of them is that every characteristic of the method has the same result on calculating distance. The solution of this problem is to give weight to each data characteristic [12]. Another factor that cause to the low accuracy of the kNN is the determination of new data classes which is based on a simple vote majority system [13-14], where the majority vote system ignores the closeness between data [15], this is unacceptable when the distance of each nearest neighbor differs greatly against the distance of the test data [16]. Furthermore, there will be a possibility of double majority class caused by the class determination system for new data based on the vote majority and determination of the number of nearest neighbors, where the number of nearest neighbors are chosen according to the desired level of success [17]. However, as suggested in [18], problems in determining new data classes with a vote majority system that ignores the closeness between data, which resulting in a misclassification, can be overcome by using distance weight. By using the method of distance weight the class determination for new data is based on the weights obtained through the distance between data.

In this research, we propose a replacement of the vote majority system in kNN using distance weight method, whereby in order to find the weight between data we combine the local mean based k-nearest neighbor (LMKNN) and distance weight k-nearest neighbor (DWKNN) methods, it is expected that the merging of these two methods is able to improve the accuracy result in the classification process. 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. Local Mean Based K-Nearest Neighbor

In [18] a method called local mean based k-nearest neighbor (LMKNN) had been proposed. This method is a simple, effective and resilient nonparametric classification. This LMKNN has been proven to improve classification performance and also reduce the effect of existing outliers, especially in small data size [16]. The of LMKNN process could be describe as follows :

Step 1: Determination of Value k

Step 2: Compute of the distance between test data to each all training data using the Euclidean distance using the equation:

$$D(x,y) = ||x-y||_2 = \sqrt{\sum_{j=1}^{N} |x-y|^2} \qquad (1)$$

Step 3: Sort distance of data from the smallest to the largest as much as k for each data class

Step 4: Calculate local mean vector of each class with the equation [16] :

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

Step 5: Define the test data class by calculating the closest distance to the local mean vector of each data class using the equation :

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

The LMKNN classification is equal to 1-NN if the value of k = 1 [18]. The value of k in LMKNN is different from the original kNN, in original kNN value of k is the number of nearest neighbors from all training data, whereas in LMKNN the value of k is the number of nearest neighbors from each class in the training data [16].

In the class determination of the test data, LMKNN uses the closest distance measurements to each local mean vector of each data class, which is considered effective in overcoming the negative effects of outliers [16].

3. Distance Weight K-Nearest Neighbor

In [19] a method called distance weight k-nearest neighbor (DWKNN) had been proposed. This method specifies a new data class based on the weight value obtained from the distance between data, so that misclassification occurs due to ignoring the proximity between data can be overcome. This weighting method had a good performance because it can reduce the influence of outliers and distribution of unbalanced data sets [15]. The of DWKNN process could be describe as follows:

Step 1: Determination the value of k.

Step 2: Calculate the test data distance for each data in each class using the Euclidean distance

Step 3: Sort the distance between data from the smallest to the largest according to number of k.

Step 4: Calculate the weights from the distances between the ordered data.

In [19] solutions to count weight based on the distance between data, one of which may be use equation :

$$w_i = \frac{1}{d(x_q,x_i)} \tag{4}$$

Or

$$w_i = 1 - d(x_q, x_i) \tag{5}$$

In [15] give another option to calculate weight based on the distance between data, the weights given by using the equation :

$$w_i = \begin{cases} \frac{d_k^{NN} - d_i^{NN}}{d_k^{NN} - d_1^{NN}} \times \frac{1}{i}, & d_k^{NN} \neq d_i^{NN}, \\ 1 \quad , & d_k^{NN} = d_i^{NN} \end{cases} \tag{6}$$

Where :
w_i is the weight of *i* from the nearest neighbor.
$d(x_q, x_i)$ is the distance between test data and training data.
d_k^{NN} is distance of *k* nearest neighbor.
d_i^{NN} is distance of *i* from *k* nearest neighbor.

4. Proposed Method

To further describe the combination of local mean based k-nearest neighbor (LMKNN) and distance weight k-nearest neighbor (DWKNN) methods, it will be explained step by step in this sub-chapter. The stages in outline can be seen in figure 1:

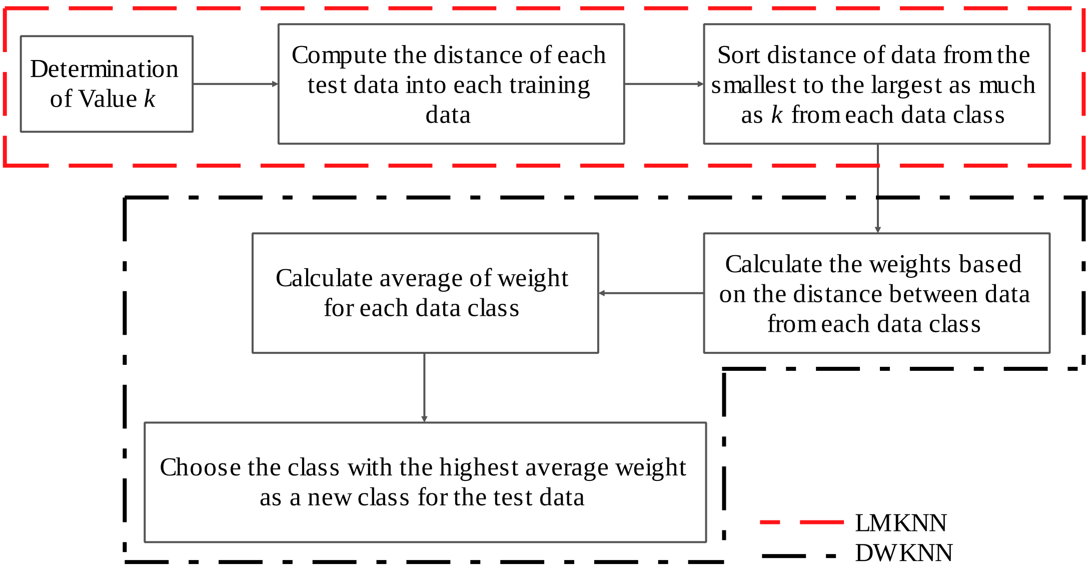
Determination of Value *k*
Compute the distance of each test data into each training data
Sort distance of data from the smallest to the largest as much as *k* from each data class
Calculate the weights based on the distance between data from each data class
Calculate average of weight for each data class
Choose the class with the highest average weight as a new class for the test data
LMKNN
DWKNN

Figure 1 Details Combined Stages of Local Mean Based K-Nearest Neighbor (LMKNN) and Distance Weight K-Nearest Neighbor (DWKNN)

From figure 1 it can be explained that the combined method of Local Mean Based K-Nearest Neighbor (LMKNN) and Distance Weight K-Nearest Neighbor (DWKNN) has several stages, among others :

Step 1 : Determination of value *k*, which is the number of nearest neighbors.
Step 2 : Compute the distance from each test data to each training data using the Euclidean using equation (1)

Step 3 : Sort the distance data from smallest to largest as many as k of each class.
Step 4 : Calculate the weights using equation (4)
Step 5 : Calculate average weight for each data class.
Step 6 : Choose class with the highest average weight as a new class for the test data.

Step 1 to step 3 is a contribution from LMKNN, and step 4 to step 6 is a contribution from DWKNN. The value of *k* in this proposed method follows the rules of LMKNN, where the value of *k* is the number of nearest neighbors from each class.

5. Result and Discussion

5.1. Result of the classification from data set

To knowing the proposed method is better than original kNN, it will be tested using several datasets from the UCI Machine Learning repository, Kaggle and Keel, such as ionosphere, iris, voice genre, lower back pain, and thyroid (new thyroid). In addition, the proposed method is also tested using real data from a public senior high school in city of Tualang, Indonesia. Details of data can be seen in the table 1 :

Table 1 Detail of Data

Data	Attributes	Class	Total Data	Data	Attributes	Class	Total Data
Ionosphere	34	2	351	Lower Back Pain Symptomps	13	2	310
Iris	4	3	150	Thyroid	5	3	215
Voice Genre	21	2	3168	Majors of Student	9	2	167

Data will be divided randomly into two category, 80% of the data will be used as training data and 20% of the data will be used as test data. Then it will be calculated using original kNN and proposed method. In this study, K is only worth 1 to 10. The average accuracy of each data can be seen in the figure 2 :

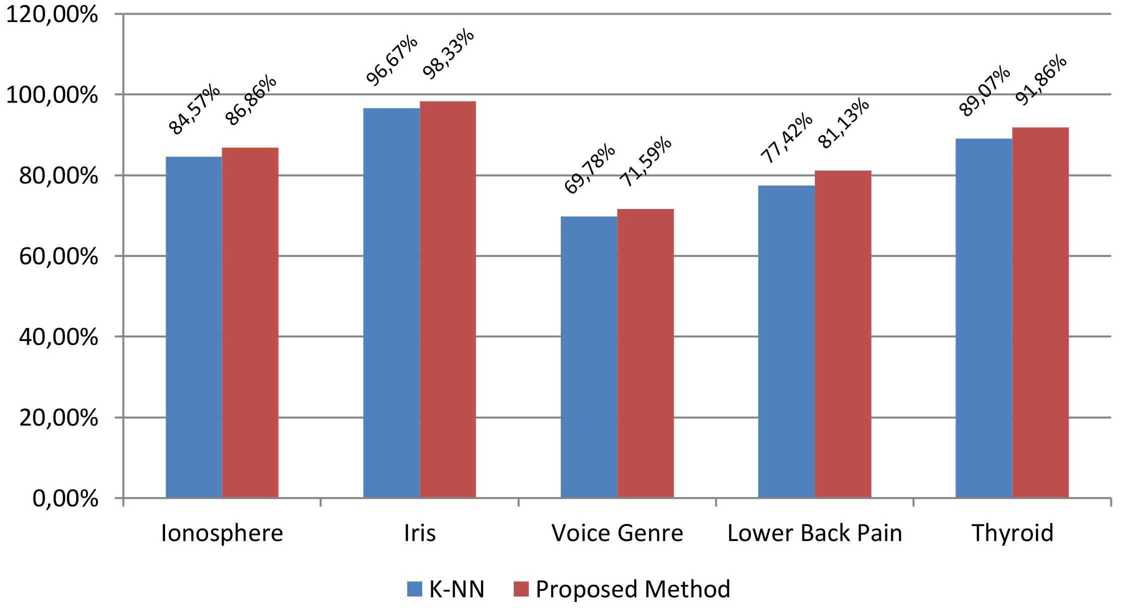

Figure 2 Average Accuracy from All Data

Referring to figure 2 it can be seen that the proposed method has a higher accuracy value than the original kNN, where the highest difference of accuracy obtained in Lower Back Pain Symptoms data is worth 3.71%, and the lowest difference of accuracy obtained in the iris data set of 1.66%. Of all tested data the average increase the accuracy is 2.452%.

5.2. Result of the classification from real data

From the results of the tests performed in the previous sub-section was knowing, the proposed method is better than original kNN. To know with certainty whether the proposed method is better to make predictions in the real data from public high school in city of Tualang, Indonesia, it will be recomparison with original kNN. Details from devide of real data can be seen in table 2 and details of the test results can be seen in table 3 :

Table 2 Details of Real Data

No	Category	Total Data for Each Class		Total of Data
		Science	Social	
1	Training	64	70	134
2	Test	15	18	33

Table 3 Comparison of Accuracy from Real Data

No	K	K-NN Conventional			Combine LMKNN+DWKNN			Method With Highest Accuracy
		True Prediction		Accuracy	True Prediction		Accuracy	
		Science	Sosial		Science	Sosial		
1	1	14	14	84.85%	14	14	84.85%	Both
2	2	14	14	84.85%	15	14	87.88%	Proposed Method
3	3	14	13	81.82%	15	13	84.85%	Proposed Method
4	4	14	13	81.82%	15	14	87.88%	Proposed Method
5	5	13	13	78.79%	15	14	87.88%	Proposed Method
6	6	14	13	81.82%	15	14	87.88%	Proposed Method
7	7	14	13	81.82%	15	14	87.88%	Proposed Method
8	8	14	13	81.82%	15	14	87.88%	Proposed Method
9	9	14	13	81.82%	15	14	87.88%	Proposed Method
10	10	14	14	84.85%	15	15	90.91%	Proposed Method
Average				82.42%			87.58%	Proposed Method

Table 3 shows that the proposed method gives the best prediction results in determining of majors public senior high school in city Tualang, Indonesia. The highest score reaching 90.91% when k = 10 and the lowest being 84.85% when k = 1. The proposed method are able to increase the average accuracy by 5.16% against accuracy value of original kNN.

6. Conclusion

Based on the findings and discussion in the previous section it can be concluded that:

- The proposed method is able to improve accuracy, where the average increase in accuracy of all test data is 2.492%. Highest average value of accuracy obtained in the data Lower Back Pain Symptoms of 3.91%, and the average value of the lowest accuracy value found in iris data that is equal to 1.66%.
- In the case of predicting majors of student in public senior high school in city of Tualang, Indonesia, the researcher proposes using the proposed method, because the proposed method proved to be better than the original kNN with average difference of accuracy is 5.16%, whereby highest accuracy value is 90.91% when k = 10.

Acknowledgement

High appreciation should be given to Lembaga Penelitian Universitas Sumatera Utara (LP USU), Graduate School of Computer Science Fasilkom-TI USU and Rector of Universitas Sumatera Utara for their support on the dissemination of this research work as well as the facilities provided.

References

[1] Nitin B and Vandana 2010 Survey of nearest neighbor techniques *Int. J. of Computer Science and Information Security* **8**(2) 302-5

[2] M Akhil j, Bulusu L D and Priti C 2013 Classification of heart disease using k-nearest neighbor and genetic algorithm Procedia Technology **10** 85-94

[3] A S Sánchez, F J Iglesias-Rodríguez, P R Fernándes & Francisco J de C J 2015 Applying the k-nearest neighbor technique to the classification of workers according to their risk of suffering musculoskeletal disorders Int. J. of Industrial Ergonomics **52** 92-9

[4] Zhibin P, Yidi W and Weiping K 2017 A new general nearest neighbor classification based on the mutual neighborhood information *Knowledge-Based System* **121** 142-52

[5] Nicolás G-P and Domingi O-B 2009 Boosting k-nearest neighbor classifier by means of input space projection *Expert Systems with Application* **36** 10570-82

[6] Jigang W, Predrag N and Leon N C 2007 Improving nearest neighbor rule with a simple adaptive distance measure *Pattern Recognition Letters* **28** 207–213

[7] Stefanos O and Georgios E 2012 Fast and accuratek-nearest neighbor classification using prototype selection by clustering The 16th Panhellenic Conference on Informatics (PCI)

[8] Yunsheng S, Jiye L, Jing L and Xingwang Z 2017 An efficient instance selection algorithm for k nearest neighbor regression *Neurocomputing* **251** 26-34

[9] Amri D, Devie P S, Dian A and Dini A 2016 Comparison of accuracy level k-nearest neighbor algorithm and support vector machine algorithm in classification water quality status *The 6th Int. Conf. on System Engineering and Technology (ICSET)* 137-41

[10] Elizabeth N T and Aditya W M Comparison of music genre classification using nearest centroid classifier and k-nearest neighbours *Int. Conf. on Information Management and Technology (ICIMTech)* 118-23

[11] Jessica M B 2017 Math test scores using k-nearest neighbor *IEEE Integrated STEM Conference (ISEC)*

[12] Maryam K 2016 A method to improve the accuracy of k-nearest neighbor algorithm *Int. J. of Computer Engineering and Information Technology* **8**(6) 90-5

[13] Syahfitri K L, Opim S S and Syahril E 2015 Sentiment analysis on Indonesian language text using support vector machine (svm) and k-nearest neighbor (kNN) (in Bahasa) *Seminar Nasional Teknologi Informasi dan Komunikasi* (SENTIKA) pp. 1-8

[14] Aditya D and Thendral P 2017 Enhancing classification accuracy of k-nearest neighbours algorithm using gain ratio *International Research Journal of Engineering and Technology (IRJET)* **4**(9) pp 1385-1388

[15] Jianping G and Taisong X and Yin K 2011 A novel weighted voting for k-nearest neighbor rule *J. of Computers* **6**(5) 833-40

[16] Zhibin P, Yidi W and Weiping K 2016 A new k-harmonic nearest neighbor classifier based on the multi-local means *Expert Systems with Applications* **67** 115-25

[17] Ö F Ertuğrul & M E Tağluk 2017 A novel version of k nearest neighbor: dependent nearest neighbor *Applied Soft Computing Journal* **55** 480-490

[18] Y Mitani and Y Hamamoto 2006 A local mean-based nonparametric classifier *Patern Recognition Letter* pp 1151-1159

[19] Gustavo E A P A B and Diego F S 2009 How k-nearest neighbor parameters affect it's performance *38º JAIIO - Simposio Argentino de Inteligencia Artificial* (ASAI) pp 95-106