Research Article

AStudy on GBW-KNN Using Statistical Testing

SeowonSong[1], Young Sang Kwak[2], Myung-ho Kim[3], Min Soo Kang*[4]

[1,2]Medical IT & Marketing, EuljiUniv.,Seongnam, Republic of Korea,
[3]Equipment & Fire Protection Engineering, GachonUniv, Seongnam, Republic of Korea,
[*4]Medical IT, EuljiUniv.,Seongnam, Republic of Korea,
songst32@gmail.com[1], ysk1188@naver.com[2], ibs@gachon.ac.kr[3], mskang@eulji.ac.kr*[4]
Corresponding Author*

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 4 June 2021

Abstract: In the 4th industrial revolution, big data and artificial intelligence are becoming more and more important. This is because the value can be for by applying artificial intelligence techniques to data generated and accumulated in real-time. Various industries utilize them to provide a variety of services and products to customers and enhance their competitiveness. The KNN algorithm is one of such analysis methods, which predicts the class of an unlabeled instance by using the classes of nearby neighbors. It is used a lot because it is simpler and easier to understand than other methods. In this study, we proposed a GBW-KNN algorithm that finds KNN after assigning weights to each individual based on the KNN graph. In addition, a statistical test was conducted to see if there was a significant difference in the performance difference between the KNN and GBW-KNN methods. As a result of the experiment, it was confirmed that the performance of GBW-KNN was excellent overall, and the difference in performance between the two methods was significant.

Keywords:Classification, K Nearest Neighbour, WKNN, Machine Learning.

1. Introduction

In the 4th industrial revolution, the boundaries of existing industries are blurring, centering on the development of ICT technology [1]. This is called the big blur phenomenon, and it is accelerating as technologies such as IoT, artificial intelligence, and big data emerge. In particular, new services are emerging through the combination of artificial intelligence and big data, and are changing the form of existing business. A large amount of data is accumulated in real-time due to the advancement of technology capable of storing and processing data. In addition, advances in data mining, machine learning, and deep learning technologies allow us to discover patterns and new values from this data. If these values are applied to products or services, competitiveness can be enhanced, so countries and companies are paying attention. Big data refers to a large amount of data and data analysis technology beyond the capabilities of existing databases. In the past, information was simply individual-centered and at the level of structured data, but now it includes unstructured data such as text [2,28]. The characteristics of big data can be expressed as 4V, which means volume, variation, velocity, and value. Techniques for finding or analyzing such patterns of big data include statistical techniques, data mining, and artificial intelligence. Machine learning is a field of artificial intelligence that develops algorithms or techniques so that computers can learn, and it means making predictions based on given attributes through training data. The process of creating a model based on data is called learning. Machine learning is classified into supervised learning and unsupervised learning depending on the presence or absence of a label. Supervised learning is to train a computer in a given state of a label and predict the result value of new instance, and regression and classification are representative. If the classification result value is fixed, it is classification, whereas in regression, the result value is not fixed. Regression is a model that is mainly used in the form of guessing data through functional expressions. Unsupervised learning is a method of learning with unlabelled data, finding hidden patterns or structures of data, and clustering is representative [3, 4,29]. The goal of the cluster model is to group similar entities by analyzing the characteristics of unlabelled data.

1.1 K Nearest Neighbor

KNN is a representative classification algorithm that classifies new individuals with labeled data. The KNN algorithm is a technique that classifies entities with unknown categories into the class of the most similar entities among labeled entities. It is a popular technique in many fields because of its intuitive and simple advantages [5]. The pseudo-code of KNN is as follows [6].

Algorithm K Nearest Neighbor

```
    #Input
#training dataset: X, class labels of X: Y, unknown
sample: x
    #Output: class labels of x: y
    #Classify (X, Y, x)
```

for i = 1 **to** length of X **do**
Compute distance d(X_i, x)
end
Compute the k smallest distances
get class labels of k-nearest-neighbors Y
Compute majority label of k-nearest-neighbors and assign the label of x

In order to classify a new object that has not been classified, the most similarly labelled data is used. Methods of measuring similarity include Euclidean distance, Manhattan distance, and cosine similarity, and the Euclidean distance is widely used. The formula to find the distance d between point A (x1, y1) and point B(x2, y2) in the Euclidean method is as follows [7].

$$d(A,B) = \sqrt{(x1 - x2)^2 + (y1 - y2)^2} \quad (1)$$

The distance between data is calculated according to the similarity measurement method, and k neighbors from the nearest data to the k-th nearest neighbor are obtained. And the most class among the k neighbor's classes becomes the new object class. Various metrics such as accuracy, precision, and recall are used to measure classification performance [8]. The advantages and disadvantages of KNN are summarized in the table below.

Table 1. Advantages and Disadvantages of KNN

Pros	Cons
-Easy to implement -Few parameters (distance metric, k)	-Sensitive to noise -Requires large storage -Difficult to find optimal k value

2. Related Research

The KNN algorithm is one of the simplest classification techniques and is a popular machine learning technique. KNN has been developed in various ways to improve performance, and a method of weighting has been mainly proposed [9, 10].Weighted KNN refers to a KNN that is weighted according to the importance of a feature or object. In addition to the method of taking the inverse of similarity, various methods have been proposed [11, 12, 13, 14].

Dudani proposed a WKKN methodology, which is a KNN that added a method of giving larger weights to nearby objects [15]. Dudani assigned the weight w as shown below.

$$wj = \begin{cases} \frac{d_k - d_j}{d_k - d_1}, & d_k \neq d_1 \\ 1, & d_k = d_1 \end{cases} \quad (2)$$

Jianping Gou and Lan Du et al. proposed DWKNN, which is a complementary form of WKNN, which uses a new weighting formula based on equation (1) [16]. This is a method of calculating the dual weight by squaring the weight obtained through Equation 1 suggested by Dudani, and using it as a new weight. Dual weight can be calculated as the following equation.

$$\overline{w}_i = \begin{cases} \frac{d(\bar{x},{x_k}^{NN}) - d(\bar{x},{x_i}^{NN})}{d(\bar{x},{x_k}^{NN}) - d(\bar{x},{x_1}^{NN})} \times \frac{d(\bar{x},{x_k}^{NN}) - d(x,{x_1}^{NN})}{d(\bar{x},{x_k}^{NN}) - d(\bar{x},{x_i}^{NN})}, \\ \quad if d(x', {x_k}^{NN}) \neq d(x', {x_1}^{NN}), \\ \quad 1 \quad , \\ \quad if d(x', {x_k}^{NN}) = d(x', {x_1}^{NN}) \end{cases} \quad (3)$$

In addition, WKNN, which applied various methods or weighted by taking the reciprocal of similarity, were introduced, and KNN derivative studies in the form of fusion of other techniques and KNN were conducted [17, 18, 19, 20]. There was also a study that improved the performance of KNN by fusion of Genetic Algorithm and KNN. GA is a technique commonly used as stochastic search methods [21]. Yan Xuesong et al conducted a study on applying GA to KNN, and was able to derive higher performance than KNN. Daniel Mateos-García, Jorge Garcia

Gutierrez, and Jose C. proposed the Simultaneous Weighting of Attributes and Neighbors (SWAN) method [22]. Unlike other WKNN algorithms, this method considers the contribution of neighbors and the importance of data attributes. Regardless of the analysis method, pre-processing is required and KNN also has been used for this purpose. KNN-based studies for pre-processing were conducted. Hautamaki, V proposed a method for detecting and processing outliers using KNN graph [22].Two-dimensional KNNG was created and the outliers were determined by counting the Indegree number of each object. The minimum number of indegrees is determined, and outliers are determined according to this criterion.

In this study, we propose a KNN that uses a KNN graph to give weights to each object and predicts a class using it. Using this method, neighbors can be selected by considering the relationship between the object and the surrounding data. Performance was compared using public data, and statistical tests were performed to confirm whether the difference was significant.

3. GBW-KNN Algorithm

In this paper, we propose GBWKNN to improve accuracy. In general, the KNN algorithm calculates similarity with other objects to classify unclassified objects, and finds the K neighbors with the highest similarity among them. Allocate objects that are not classified into the most classes by referring to the labels of K neighbors. It is used a lot because it is an easy to understand and intuitive method, but it is affected a lot by outliers, and there is a burden to determine k in advance. Outliers mean values that are far from other observations. If neighbors are selected using the conventional KNN method, data that are farther apart than many data can be selected if it is close. Therefore, in this paper, we propose GBWKNN that can supplement this problem.

First, perform KNN and count “how many times as KNN” for each point. For example, in the case of K=3, there are 3 extending lines for each point. The weight is calculated by counting the number of times connected to other points. If weights are assigned to each point by calculating the weight in this way, the point that receives more selections from other neighbors has a smaller weight. Also, when classifying unknown points, the weight given to the distance obtained is multiplied by the similarity update the similarity. When selecting a neighbor in this way, a point that has received many selections from other data groups is selected as a neighbor, even if the absolute distance is close but a little farther than a point apart from other data. Therefore, it can be regarded as a method of selecting a neighbor by considering the surrounding characteristics of the data, rather than simply judging by the similarity measure.

4. Experiment

4.1 Experiment design

In order to compare the performance of GBW-KNN and KNN proposed in this paper, an experiment was designed as follows. Wisconsin Breast Cancer data and Pima Indian Diabetes data were used for the experiment [23, 24]. Each data can be downloaded from the UCI repository and Kaggle. Breast Cancer data consists of 32 attributes and 569 patient data, and Pima Indian data consists of 9 attributes and data of 768 patients.

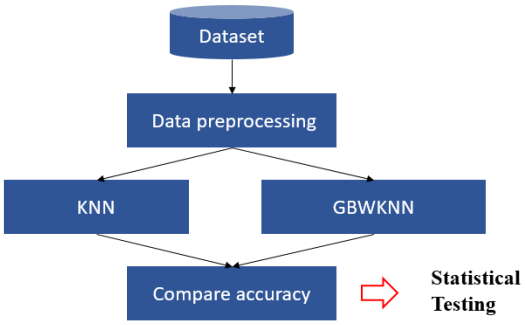

Figure 1. Experiment Design

Normalization and feature selection were performed to apply the dataset to the experiment, and GBWKNN and KNN were implemented in Python. The experiment was repeated 30 times for each K value while changing the k value, and accuracy was used as a performance evaluation scale in this paper.

4.2 Experiment Result

The experimental results of each data are as follows. This is a table and graph that summarizes the average accuracy of 30 times for each K value.

Table 2. Result of Breast Cancer Dataset

	KNN	GBW-KNN	difference
3	92.96	93.70	0.74
5	92.94	93.76	0.82
7	93.27	94.08	0.81
9	93.84	94.56	0.72
11	94.07	94.58	0.51
avg	93.416	94.136	0.72

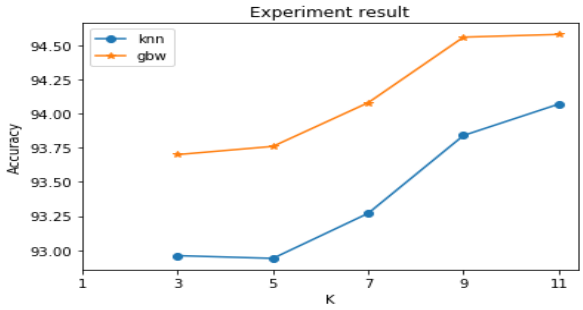

Figure2.Result of Breast Cancer Dataset

Table 3. Result of Diabetes dataset

	KNN	GBW-KNN	difference
3	70.72	71.83	1.11
5	71.47	72.53	1.05
7	72.02	72.97	0.95
9	73	73.87	0.64

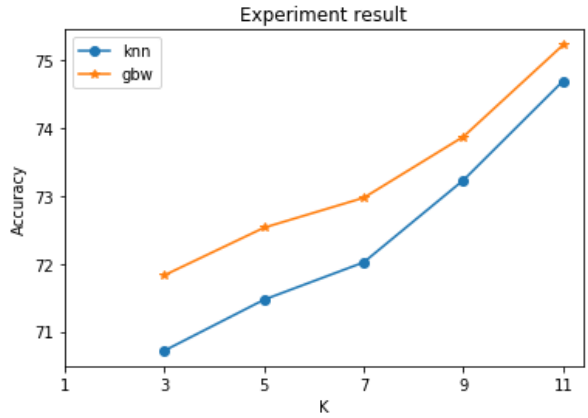

Figure3. Result of DiabetesDataset

4.3 Statistical Testing

Paired t-test is a statistical test technique that is widely used when comparing the performance of classification algorithms [25]. Unlike the independent T-test, the paired T-test is a method of analyzing differences between the same groups [26]. In the experiment of this paper, the GBWKNN algorithm is applied immediately after the

accuracy of the existing KNN algorithm is derived under the same random seed value, so the difference between before and after the application of the proposed algorithm can be verified with a paired t-test.

In order to perform the T-Test, it is necessary to check whether the collected data follows a normal distribution. Depending on whether or not, an analysis method is selected from the parameter method and the nonparametric method [27]. There is also a method to check normality using a graph, but in this paper, shapiro. test() functions were used. In the Shapiro test result, if the p-value is greater than or equal to the significance level, it can be considered to follow the normal distribution. The hypothesis here is as follows. The null hypothesis is 'Sample follows the normal distribution' and 'Sample does not follow the normal distribution' is the alternative hypothesis. Here, if the p-value is less than 0.05, the null hypothesis is rejected and the alternative hypothesis is adopted. Conversely, if the p value is greater than 0.05, the null hypothesis that the normal distribution is followed cannot be rejected.

```
> with(pima_data, shapiro.test(gbw_3nn-ori_3nn))

        Shapiro-Wilk normality test

data:  gbw_3nn - ori_3nn
W = 0.89572, p-value = 0.00661
```

Figure 4. Result of Shapiro normality test 1

The above figure is the result of Shapiro test when k=4 in the Pima dataset. In this case, the null hypothesis can be rejected because the p-value is less than the significance level of 0.05. Therefore, since normality is not satisfied, the Wilcoxon signed rank test should be performed instead of the paired t-test.

```
> with(bc_data, shapiro.test(ori_5nn-gbw_5nn))

        Shapiro-Wilk normality test

data:  ori_5nn - gbw_5nn
W = 0.93944, p-value = 0.08787
```

Figure 5. Result of Shapiro Normality Test 2

The above figure is the result of Shapiro test when k=5 in breast cancer. Because the P-value is greater than the significance level, the null hypothesis cannot be rejected. Therefore, since normality is satisfied, the paired-t test can be performed. The table below shows the Shapiro-Wilk normality test result of the breast cancer dataset. When K was 5 and 9, the paired t-test was applied, and when K was 3, 7, and 11, the Wilcoxon signed rank technique was applied.

Table 4. Result of Shapiro Normality Test of Breast Cancer Dataset

K	p-value	Normality
3	0.494	X
5	0.087	O
7	0.0031	X
9	0.075	O
11	2.348e-05	X

The table below shows the Shapiro-Wilk normality test result of the Diabetes dataset. Paired t-test was applied when K is 5, 9, 11, and Wilcoxon signed rank technique was applied when K is 3 and 7.

Table 5. Result of Shapiro Normality test of Diabetes dataset

K	p-value	Normality
3	0.00661	X
5	0.09996	O
7	0.004815	X
9	0.2879	O
11	0.3148	O

If the number of data is too small or the normality assumption is not satisfied, the test can be performed using a nonparametric method instead of a parametric method. The analysis method was applied depending on whether each data satisfies the normality. The null hypothesis (H0) and the alternative hypothesis (H1) are as follows. The null hypothesis is 'the difference between the medians between the two groups is 0', and the alternative hypothesis is 'the difference between the medians between the two groups is not 0'.

```
> t.test(pima_data$ori_5nn, pima_data$gbw_5nn, paired = TRUE)

        Paired t-test

data:  pima_data$ori_5nn and pima_data$gbw_5nn
t = -3.3062, df = 29, p-value = 0.002526
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.7050329 -0.4017671
sample estimates:
mean of the differences
                -1.0534
```

Figure 6. Result of Paired T test

The above figure is the result of applying Paired t-test when k=5 of Pima data. As a result of the Shapiro test, the p-value was 0.09996, which was greater than the significance level of 0.05, so the paired t-test was performed as shown in the figure above. As a result of the paired t-test, the p-value was 0.002526, which is less than 0.05, so the null hypothesis was rejected at the significance level of 0.05, and it can be said there was a significant difference in accuracy before and after weighting each point. Like the paired t-test, the Wilcoxon signed rank method is also tested using the difference between the paired data.

```
> wilcox.test(bc_data$ori_3nn, bc_data$gbw_3nn, paired = TRUE)

        Wilcoxon signed rank test with continuity
        correction

data:  bc_data$ori_3nn and bc_data$gbw_3nn
V = 29, p-value = 0.0001936
alternative hypothesis: true location shift is not equal to 0
```

Figure 7. Result of Wilcoxon signed rank test

The picture above is the result of applying Wilcoxon signed rank when k=3 of breast cancer data. As a result of the paired t-test, the p-value was 0. 0001936, which is less than 0.05, so the null hypothesis was rejected at the significance level of 0.05, and it can be said there was a significant difference in accuracy before and after weighting each point. For breast cancer, the paired t-test was performed when K is 5 or 9, and each p-value was less than 0. 05. When K is 3, 7, 11, Wilcoxon signed-rank technique was applied and each p-value was less than 0.05. For the Pima data, the paired t-test was applied when K is 5, 9, and 11, and each p-value was all less than 0.05. When K is 3 and 7, Wilcoxon signed-rank technique was applied and each p-value was less than 0.05. Therefore, it can be said that there is a significant difference in the accuracy comparison experiment of KNN and GBW-KNN using the two data sets.

5. Conclusion

In the 4th industrial revolution era, the boundaries between industries are blurring. Artificial intelligence is being applied to big data in each industry, and it is expected to increase gradually. Each country and companies try to actively use it because the value that can enhance competitiveness can be found through analysis. The KNN algorithm predicts class of a new object that is not classified as a representative machine learning algorithm. Although it has the advantage of being simple and easy to understand, it has the disadvantage that k must be determined in advance and is sensitive to outliers. In this paper, GBWKNN was developed to improve classification accuracy, and statistical tests were performed to find out whether there was a significant difference when performing performance comparison experiments with the existing KNN. For verification, Wisconsin Breast Cancer dataset and Pima Indian diabetes dataset, which can be downloaded from the Internet, were used, and experiments were conducted under five k-value conditions, accuracy was measured, and statistical tests were performed. The Paired t-test, which is widely used when comparing the performance of classification algorithms, was conducted, and when normality was not satisfied, Wilcoxon signed-rank technique was applied. Through the experiment, it was possible to verify the difference before and after weighting through GBWKNN. As a result of the experiment, the overall accuracy of the proposed GBWKNN was superior to that of the KNN. The normality satisfaction test was performed for each data set and condition of K value, and the statistical test technique was applied differently according to the result. As a result, the p-value in all cases was less than the significance level of 0.05, and the null hypothesis was rejected. Therefore, it could be said that there is a significant difference in the accuracy of KNN and GBWKNN.

Acknowledgement

This research was supported by Eulji University in 2020.

References

1. https://ko.wikipedia.org/wiki/%EB%B9%85%EB%B8%94%EB%9F%AC
2. Byun JH, Lee SC,(2017) Big data processing and statistical analysis, Seoul, Comone media
3. Kelleher, J. D., Mac Namee, B., D'arcy (2017) AFundamentals of Machine Learning for Predictive Data Analytics, Seoul, Acorn.
4. O'Neil, C., & Schutt, R., (2014)Doing Data Science. Seoul, Hanbit.
5. Belur V. Dasarathy.(1991) Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. Mc GrawHill, Computer Science Series. IEEE Computer Society Press. Las Alamitos, California, 217- 224.
6. Tay, B., Hyun, J. K., & Oh, S. (2014). A machine learning approach for specification of spinal cord injuries using fractional anisotropy values obtained from diffusion tensor images. Computational and mathematical methods in medicine, 2014.
7. https://en.wikipedia.org/wiki/Euclidean_distance
8. Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology, 5(5), 241-266
9. F. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, and Z. Protopapas. (1996) Fast Nearest-Neighbor Search in Medical Image Databases. In proceedings of the 22nd International Conference on Very Large Data Base. Bombay, India. Sept 3-6, 215-226.
10. K. L. Cheung & A. W. Fu.(1998) Enhanced Nearest Neighbour Search on the R-Tree. SIGMOD Record. vol. 27. pp. 16-21.
11. E. Corchado , M. Wozniak , A. Abraham, (2014) A. de Carvalho , V. Snásel , Recent trends in intelligent data analysis., Neurocomputing 126, 1-2.
12. Abraham, A. (2012). Hybrid approaches for approximate reasoning, *Journal of Intelligent & Fuzzy Systems*, 23(2, 3), 41-42.
13. Samworth, R. J. (2012) Optimal weighted nearest neighbour classifiers. The Annals of Statistics, 40(5), 2733-2763.
14. Bicego, M., &Loog, M. (2016). Weighted K-nearest neighbor revisited. In 2016 23rd International Conference on Pattern Recognition (ICPR) (pp. 1642-1647)..
15. Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, (4), 325-327.
16. Gou, J., Du, L., Zhang, Y., &Xiong, T. (2012). A new distance-weighted k-nearest neighbor classifier. *J. Inf. Comput. Sci*, *9*(6), 1429-1436.
17. Lin, Jessica, David Etter, David DeBarr (2008) "Exact and approximate reverse nearest neighbor search for multimedia data." Proceedings of the 2008 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics.
18. Radovanović, Miloš, Alexandros Nanopoulos, and Mirjana Ivanović., (2014)Reverse nearest neighbors in unsupervised distance-based outlier detection., IEEE transactions on knowledge and data engineering 27.5, 1369-1382.
19. Yan, X., Li, W., Chen, W., Luo, W., Zhang, C., Wu, Q., & Liu, H. (2013). Weighted K-nearest neighbor classification algorithm based on Genetic Algorithm. *Telkomnika*, *11*(10), 6173-6178.
20. Mateos-García, D., García-Gutiérrez, J., &Riquelme-Santos, J. C. (2019). On the evolutionary weighting of neighbours and features in the k-nearest neighbour rule. *Neurocomputing*, *326*, 54-60.
21. J Holland. (1975) Adaptation in natural and artificial systems. University of Michigan press.
22. Hautamaki, V., Karkkainen, I., &Franti, P. (2004). Outlier detection using k-nearest neighbour graph. In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR. Vol. 3, 430-433.
23. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
24. https://www.kaggle.com/uciml/pima-indians-diabetes-database
25. Bramer, M., (2007) Principles of data mining, Vol. 180, London: Springer
26. .https://nittaku.tistory.com/459
27. https://mansoostat.tistory.com/51.
28. Bae, Y., & Han, S. (2019). Academic Engagement and Learning Outcomes of the Student Experience in the Research University: Construct Validation of the Instrument. Educational Sciences: Theory & Practice, 19(3).

29. Hadi, N. U., & Muhammad, B. (2019). Factors Influencing Postgraduate Students' Performance: A high order top down structural equation modelling approach. Educational Sciences: Theory & Practice, 19(2).