Performance Analysis of Distance Measures in K-Nearest Neighbor

Annisa Fadhillah Pulungan[1], Muhammad Zarlis[2], Saib Suwilo[3]
annisa.pulungan93@gmail.com[1], m.zarlis@yahoo.com[2], saibwilo@gmail.com[3]

Student in Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia[1]
Department of Computer Science, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia[2]
Department of Mathematics, Universitas Sumatera Utara, Medan, Indonesia[3]

Abstract. K-Nearest Neighbor (KNN) has important parameters that affect the performance of the KNN. The parameter is the k value and distance matrix. In KNN, the distance between two points is determined by the calculation of the distance matrix. In this paper we will analyze and compare the performance KNN using the distance function. The distance are Braycurtis, Canberra and Euclidean Distance. This study uses Confusion Matrix for evaluation of accuracy, sensitivity and specificity. The results showed that the Braycurtis distance had better performance than Canberra Distance and Euclidean Distance with accuracy values of 96%, sensitivity of 96.8% and specificity of 98.2%.

Keywords: Classification, K-Nearest Neighbor, Braycurtis Distance, Canberra Distance, Euclidean Distance, Confusion Matrix.

1 Introduction

Classification is a technique used to build a classification model from a sample of training data. Classification will analyze input data and build a model that will describe the class of data. Class labels from unknown data samples can be predicted using classification techniques [1]. One of the most popular classification techniques is K Nearest Neighbor (KNN).

KNN is known as an algorithm that is very simple and easy. Many researchers make the KNN algorithm as their research algorithm. This is because KNN is good at handling noise, simple, easy, and not complicated in implementation. The KNN algorithm aims to classify new objects based on attribute values and training data [2]. The KNN algorithm has important parameters that affect the performance of the KNN algorithm. The parameters are the K value and the distance measures. The parameter K value is used to determine the number of neighbors to be used compared to the predicted value.

In addition to the K value, the distance measures is an important factor that depends on collecting data in the KNN algorithm. The value of the resulting distance measures will affect the performance of the KNN. The distance between two data points is determined by the calculation of the distance matrix. Euclidean Distance is the most widely used distance matrix function in calculating distance matrices. There are several types of distance measures other

ICMASES 2019, August 03-04, Samarinda, Indonesia

DOI 10.4108/eai.3-8-2019.2290748

than Euclidean Distance namely Manhattan Distance, Minkowski Distance, Canberra Distance, Braycurtis Distance, Chi-Square and others.

2 Related Work

In conducting research, we use several relevant studies related to K-Nearest Neighbor and Calculation of distance as a reference in conducting research. Vashista & Nagar have done an experimental study by comparing the Euclidean distance, Manhattan distance, Canberra distance, and Hybrid distance on the LVQ algorithm. The conclusion of this study is that Hybrid Distance has the best ability in LVQ data recognition followed by Canberra Distance, Manhattan Distance and Euclidean Distance [3]. Alamri et al have also done an experimental study about satellite classification using distance matrices by comparing Braycurtis distance, manhattan distance, euclidean distance. This study shows that Braycurtis distance have the best accuracy of 85% and are followed by Manhattan Distance (City Block Distance) and Euclidean Distance of 71% [4].

Kaur have an experimental study by conducting a comparative study of several types of distance calculation methods to predict software errors using the K-means clustering method with three distance measures, namely Euclidean distance, Sorrensen distance and Canberra distance. The data used is a dataset collected from NASA MDP. This study produced K-Mean Clustering with a Sorrensen distance better than Euclidean Distance and Canberra Distance.[5].

The difference of this research with previous research is the use of the Braycurtis distance method and Canberra distance to measure distance in the K-Nearest Neighbor algorithm to get the best accuracy value by using the k-Fold Cross Validation method as a method of evaluating K-Nearest Neighbor performance in classification.

3 K-Nearest Neighbor

Cover and Hart introduced K-Nearest Neighbor in 1968. K-Nearest Neighbor is a classification method that is lazy learner because this algorithm stores all training data values and delays the process of forming classification models until the test data is given for prediction[1]. During the classification process of the test data, the KNN algorithm will immediately search through all training examples by calculating the distance between the test data and all training data to identify the nearest neighbor and produce a classification value. Specifically, the distance between the two data points is determined by the calculation of the distance matrix, where the most widely used distance matrix in the K-Nearest Neighbor algorithm is Euclidean Distance. Then, KNN will give a point to class between the k values of the nearest neighbor (where the value k is an integer)[6]. The steps for classifying the KNN algorithm are as follows:

1. Determine the parameter K value
2. Calculate the distance between the new data and all training data
3. Sort the distance and set the nearest neighbor based on the minimum distance to K
4. Check the class from the nearest neighbor
5. Set a majority of the closest neighbor class as the new data predictive value

4 Distance Measures

Distance Measures is widely used in determining the degree of similarity or dissimilarity between two vectors. So this method is widely used to carry out pattern recognition [7]. Some distance methods include: Euclidean Distance, Chebyshev, Angular Separation, Canberra Distance, Haming Distance, Sorrensen Distance and so on. In the K-Nearest Neighbor algorithm, the classification process uses the Euclidean Distance method.

The difference in measure similarity distance is very suitable for analyzing difference classes. Calculation of similarity distance using several matrix values is usually used to extract the similarity of data objects and assisted with the classification process using efficient algorithms.

4.1 Braycurtis Distance

Bray Curtis Distance is also called *Sorrensen Distance*. When a difference is added between normalized variables with addition variable of object, *sorrensen distance* will be *modified city block distance*[8].

$$d_{ij} = \sum_{k=1}^{n} \frac{|x_{ik} - x_{jk}|}{(x_{ik} + x_{jk})} \quad (1)$$

4.2 Canberra Distance

The Canberra Distance was introduced and developed first by G.N Lance and W.T William in 1966 and 1967. Canberra Distance is used to get the distance from the pair of points where the data is in the form of original data and is in a vector space. The Canberra Distance gives two output values, namely TRUE and FALSE[3].

$$d_{ij} = \sum_{k=1}^{n} \frac{|x_{ik} - x_{jk}|}{|x_{ik}| - |x_{jk}|} \quad (2)$$

4.3 Euclidean Distance

Euclidean Distance is the distance between points in a straight line. This distance method uses the Pythagorean theorem. And is the distance calculation that is most often used in machine learning processes [9]. The Euclidean distance formula is the result of the square root difference of two vectors.

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \quad (3)$$

5 Methodology

The steps of research conducted in this study are shown in **Figure 1**

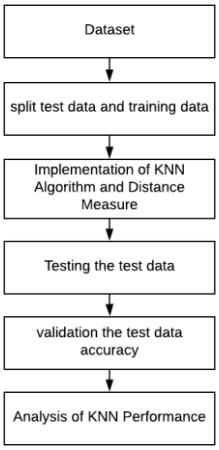

Fig. 1. Architecture KNN algorithm and Distance Measures

5.1 Dataset Used

In this study we used the Iris dataset from the UC Irvine Machine Learning Repository (UCI Machine Learning Repository). This data set has 5 attributes that will be used in the classification process using KNN. Four features of 5 features are measured from each sample length and width of sepals and petals in centimeters.

5.2 Split data with K-Fold Cross Validation

K-Fold Cross Validation is used to evaluate the performance of the KNN algorithm. The purpose of the K-fold cross validation is to validate the KNN algorithm to be more tested and the resulting performance is valid. K Value in K-Fold Validation is an integer that will be used to divide data [10].

In this study, researchers used K-Fold Cross Validation where the k value is 10. Then from 150 data in the Iris dataset, it will be divided into 10 subset of data sections. Each subset will have 15 data. And training process and testing process will be done 10 times.

5.3 Implementation KNN algorithm and Distance Measures

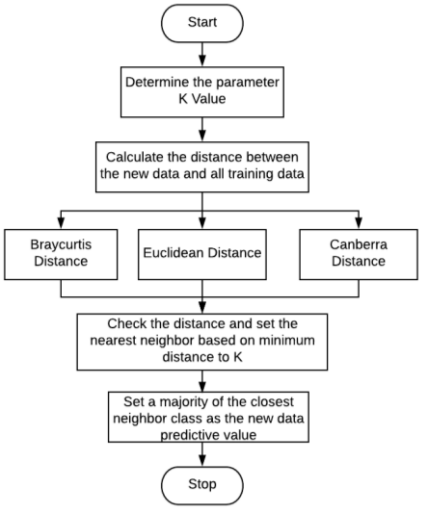

Fig. 2. Flowchart of KNN and Distance Measures

5.4 Evaluation of KNN Performance with Confusion Matrix

Confusion Matrix is a concept of machine learning techniques. Confusion Matrix has information about the actual data and the prediction results of a classification that has been done by the classification method. Confusion Matrix has two dimensions, namely dimensions that contain the actual data of an object and dimensions that contain the results of prediction of classification techniques.

		True Values	
		True	False
Prediction	True	TP Correct result	FP Unexpected result
	False	FN Missing result	TN Correct absence of result

Fig. 3 *Confusion Matrix*

In the confusion matrix, there are several terms used in the case of classification, namely :

1. *True Positive* (TP) is positive data detected correctly
2. *False Positive* (FP) adalah positive data detected incorrectly
3. *False Negative* (FN) adalah negative data detected incorrectly
4. *True Negative* (TN) adalah negative data detected correctly.

Some classification performance calculations can be explained from the confusion matrix. Some classification performance calculations in this study are

Accuracy
Accuracy is the percentage of the total number of correct predictions in the classification process [11]

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{n}} \quad (4)$$

Sensitivity
Sensitivity is often referred to as recall. Sensitivity is a percentage of positive data that is predicted as a positive value [12].

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5)$$

Specificity
Specivicity is the percentage of negative data predicted as negative data in the classification process [12].

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (6)$$

6 Experiment Result

In this study, the calculation of the distance of test data with training data was calculated using the distance method by Braycurtis Distance, Canberra Distance and Euclidean Distance. The scale of the k value given in the K-Nearest Neighbor algorithm is k = 2 to k = 10. **Table 1** shows the results of calculating the accuracy, sensitivity, and specificity of the KNN algorithm with the distance Braycurtis method.

Table 1. Performance KNN with Braycurtis Distance

K Value	Accuracy	Sensitivity	Specificity
2	94,47%	95,82%	97,69%
3	95,33%	95,98%	97,86%
4	95,33%	95,98%	97,86%
5	95,33%	96,27%	98,1%
6	96%	96,8%	98,2%
7	96%	96,8%	98,2%
8	96%	96,8%	98,2%
9	94,67%	95,87%	97,67%
10	96%	96,8%	98,23%

Table 1 shows that the best performance values in K-Neareast Neighbor and Braycurtis distance are K = 6, 7, 8 and 10 with accuracy values of 96%, sensitivity values of 96.8% and specificity values of 98.2%. **Table 2** shows the performance of the KNN with Canberra Distance.

Table 2. Performance KNN with Canberra Distance

K Value	Accuracy	Sensitivity	Specificity
2	93,28%	94,47%	97,01 %
3	94,70%	95,7%	97,65%
4	94%	95,14%	97,34%
5	94%	94,88%	97,25%
6	94,67%	95,43%	97,56%
7	94,67%	95,49%	97,56%
8	94,67%	95,49%	97,56%
9	94%	95,09%	97,31%
10	94%	95,09%	97,31%

Table 2 shows that the best performance values in the Neareast Neighbor and Canberra distance are K = 3 with accuracy value of 94.7%, sensitivity values of 95.7% and specificity values of 97.65%. **Table 3** shows the performance of the KNN with Euclidean Distance..

Table 3. Performance KNN with Euclidean Distance

Nilai K	Accuracy	Sensitivity	Specificity
2	95,09%	95,82 %	97,69 %
3	95,33%	95,98 %	97,86 %
4	95,33%	95,98 %	97,86 %
5	95,33%	96,04 %	97,87 %
6	95,33%	96,04 %	97,87 %
7	95,33%	95,88 %	97,82 %
8	95,33%	95,88 %	97,82 %
9	95,33%	95,88 %	97,82 %
10	95,33%	95,88 %	97,82 %

Table 3 shows that the best performance value in Neareast Neighbor and Euclidean distance is K = 5 and K = 6 with accuracy value of 95.33%, sensitivity value of 96.04% and value of specificity of 97.87%. **Figure 4** shows a graph of Accuracy, Sensitivity and Specificity Performance.

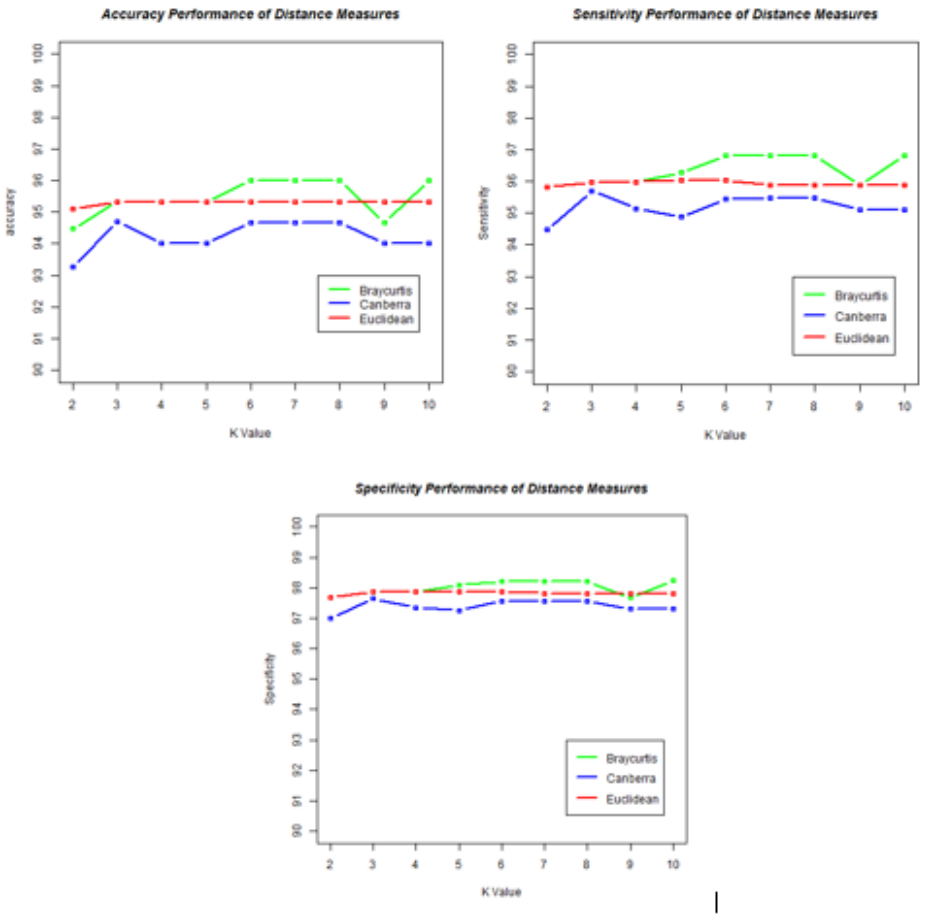

Fig. 4. Analysis Performance KNN and Distance Measures

7 Conclusion

Braycurtis Distance has a better performance than the Canberra and Euclidean distance methods at K = 6, K = 7, K = 8 and K = 10 with an accuracy value of 96%, sensitivity of 96.8% and specificity of 98.2%. Followed by the next best performance by Euclidean in the value of K = 5 and K = 6 with accuracy of 95.33%, Sensitivity of 96.04% and specificity of 97.82%. And the best performance Canberra distance method on the value of K = 3 is accuracy of 94.70%, sensitivity of 95.7% and specificity of 97.65% on the Iris dataset.

References

[1] Mulak, P. & Talhar, N. 2015. Analysis of Distance Measures Using K-Nearest Neighbor Algorithm on KDD Dataset. *International Journal of Science adn Research (IJSR)* **4**(7) : 2101-2104.

[2] Okfalisa., Mustakim., Gazalba, I., & Reza, N.G.I. 2017. Comparative Analysis of K-Nearest Neighbor and Modified K-Nearest Neighbor Algorithm for Data Classification. *International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE),* pp : 294-298.

[3] Vashistha, R., & Nagar, S.2017.An intelligent system for clustering using hybridization of distance function in learning vector quantization algorithm. *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pp. 1-7.

[4] Alamri, S. S. A., Bin-Sama, A. S. A., & Bin-Habtoor, A. S. Y. (2016). Satellite Image Classification by Using Distance Metric. *International Journal of Computer Science and Information Security* **14**(3) : 65.

[5] Kaur, D.2014. A comparative study of various distance measure for software fault prediction. *International Journal of Computer Trends and Technology (IJCTT)* **17**(3) :117-120.

[6] Hu, L.-Y., Huang, M.-W., Ke, S.-W., & Tsai C.-F., 2016. The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus* 5:1304.

[7] Wurdianarto, S.R., Novianto,S. & Rosyidah, U. 2014. Perbandingan euclidean distance dengan canberra distance pada face recognition. *Techni.COM* **13**(1): 31-37.

[8] Moghtadaiee, V., Dempster, A. 2015. Vector distance measure comparison in indoor location fingerprinting. *International Global Navigation Satellite Systems Society (IGNSS Symposium).*

[9] Viriyavisuthisakul, S., Sanguansat, P., Charnkeitkong, P., & Haruechaiyasak, C. 2015. A comparison of similarity measures for online social media Thai text classification. *2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON),* pp. 1-6.

[10] Jung, Y. 2017. Multiple Predicting K-Fold Cross-Validation for Model Selection. *Journal of Nonparametric Statistics* **30**(1) : 197-215.

[11] Deng, X., Liu, Q., Deng, Y.,& Mahadevan, S.2016. An Improved Method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences* 340-341.

[12] Saraswathi, D., & Sheela, L.M.I. 2016. Lung Image Segmentation Using K-Means Clustering Algorithm with Novel Distance Metric. *International Journal of Recent Trends in Engineering & Research* 2(12) : 236-245.