SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10207
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Analysis of Braycurtis, Canberra and Euclidean Distance in KNN Algorithm

Annisa Fadhillah Pulungan
North Sumatera University
Medan, Indonesia
annisa.pulungan93@gmail.com

Muhammad Zarlis
North Sumatera University
Medan, Indonesia
m.zarlis@yahoo.com

Saib Suwilo
North Sumatera University
Medan, Indonesia
saibwilo@gmail.com

***Abstract*—**Classification is a technique used to build a classification model from a sample of training data. One of the most popular classification techniques is The K-Nearest Neighbor (KNN). The KNN algorithm has important parameter that affect the performance of the KNN Algorithm. The parameter is the value of the K and distance matrix. The distance between two points is determined by the calculation of the distance matrix before classification process by the KNN. The purpose of this study was to analyze and compare performance of the KNN using the distance function. The distance functions are Braycurtis Distance, Canberra Distance and Euclidean Distance based on an accuracy perspective. This study uses the Iris Dataset from the UCI Machine Learning Repository. The evaluation method used id 10-Fold Cross-Validation. The result showed that the Braycurtis distance method had better performance that Canberra Distance and Euclidean Distance methods at K=6, K=7, K=8 ad K=10 with accuracy values of 96 %.

Keywords—Classification; K-Nearest Neighbor; Braycurtis Distance; Canberra Distance; Euclidean Distance

I. 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 (Mulak&Talhar, 2015). 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 (Okfalisa et al, 2017). 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, thedistance measures is an important factor that depends on collecting data in the KNN algorithm. The value of the resulting distance measureswill 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 than Euclidean Distance namely Manhattan Distance, Minkowski Distance, Canberra Distance, Braycurtis Distance, Chi-Square and others.

II. LITERATURE REVIEW

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 (Vashista& Nagar, 2017). Alamri et al have

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10207
e-ISSN : 2541-2019
p-ISSN : 2541-044X

also done an experimental study aboutsatellite 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% (Alamri et al, 2016).

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 (Kaur, 2014).

The difference of this research with previous research is the use of the Braycurtis distance method and Canberradistance 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.

III. PROPOSED METHOD

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

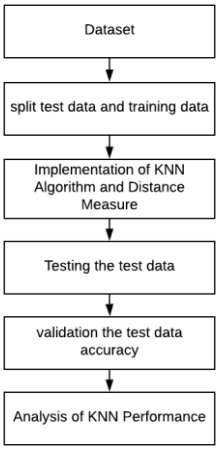

Figure 1. Architecture KNN algorithm and Distance Measures

1. Dataset

In this study used Iris dataset from 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 were measured from each sample length and width of sepals and flower petals in centimeters.

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 K-fold cross validation is to validate the KNN algorithm to be more tested and the resulting performance is valid. The K value in K-Fold Cross Validation is an integer that will be used to divide the data.

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

3. Classification With 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 kvalues of the nearest neighbor (where the value k is an integer)(Hu et al, 2016).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

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10207
e-ISSN : 2541-2019
p-ISSN : 2541-044X

4. Implementasion Measure Distance in KNN

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(Wudianarto et al, 2014). 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 similaritydistance 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.

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*modifiedcity block distance*(Moghtadaiee&Dempster, 2015).

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

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 (Vashista& Nagar, 2017).

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

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 (Viriyavisuthisakul et al, 2015). The Euclidean distance formula is the result of the square root difference of two vectors.

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

5. Evaluation of KNN Performance with Confusion Matrix

Confusion matrix is generally used to evaluate the performance of an algorithm. Confusion matrix has information about the actual data and the results of the prediction of a classification into matrix form.In obtaining the accuracy value of the KNN algorithm, the steps are as follows:

a. Determination of test data, training data and parameter values k on KNN
b. Modeling the classification with the KNN algorithm to get the KNN model
c. Conduct training and model testing processes using the 10-fold cross validation method.
d. Obtaining a confusion matrix
e. Change the confusion matrix into the table of confusion for each class
f. Calculate the accuracy of each class based on the table of confusion

IV. RESULT AND DISCUSSION

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 and Measure Distance

K Value	Braycurtis Distance	Canberra Distance	Euclidean Distance
2	94,47%	93,28%	95,09%
3	95,33%	94,70%	95,33%

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10207
e-ISSN : 2541-2019
p-ISSN : 2541-044X

4	95,33%	94%	95,33%
5	95,33%	94%	95,33%
6	96%	94,67%	95,33%
7	96%	94,67%	95,33%
8	96%	94,67%	95,33%
9	94,67%	94%	95,33%
10	96%	94%	95,33%

Figure 2 shows a graph of Accuracy performance in K-Nearest Neighbor.

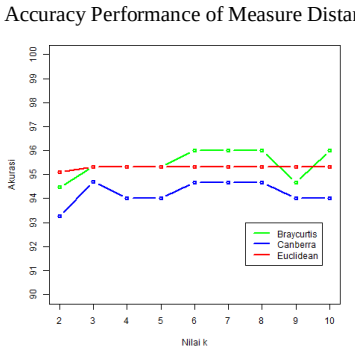

Figure 2. Accuracy Performance of Measure Distance

V. CONCLUSION AND SUGGESTION

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%. Followed by the next best performance by Euclidean in the value of K = 5 and K = 6 with accuracy of 95.33%. 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

VI. REFERENCES

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.*

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.*

Kaur, D. 2014. A comparative study of various distance measure for software fault prediction. *International Journal of Computer Trends and Technology (IJCTT).*

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

Mulak&Talhar, N. 2015. Analysis of Distance Measurs Using K-Nearest Neighbor Algorithm on KDD Dataset. *International Journal of Science and Research (IJSR).*

Okfalisa et al. 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).*

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.

Viriyavisuthisakul, S.,et al. 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.

Wurdianarto, S.R., Novianto,S. &Rosyidah, U. 2014. Perbandinganeuclidean distance dengancanberra distance pada face recognition. *Techni.COM***13**(1): 31-37.