JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i3 (211-216)

ISSN : 1979-9292
E-ISSN : 2460-5611

ANALISIS VARIATION K-FOLD CROSS VALIDATION ON CLASSIFICATION DATA METHOD K-NEAREST NEIGHBOR

Ridha Maya Faza Lubis[1)], Zakarias Situmorang[2)] , Rika Rosnelly[3)]
[1,3]Universitas Potensi Utama, Jl. KL YosSudarso Km 6.5 No 3A, Medan, 59391, Indonesia
email: ridhamayafazalubis@gmail.com, rikarosnelly@gmail.com
[2]Universitas Katolik Santo Thomas,Jl. Setia Budi Kp. Tengah, Medan, 20135, Indonesia
email: zakarias65@yahoo.com

Abstract

To produce a data classification that has data accuracy or similarity in proximity of a measurement result to the actual numbers or data, testing can be done based on accuracy with test data parameters and training data determined by Cross Validation. Therefore data accuracy is very influential on the final result of data classification because when data accuracy is inaccurate it will affect the percentage of test data grouping and training data. Whereas in the K-Nearest Neighbor method there is no division of training data and test data. For this reason, researchers analyzed the determination of training data and test data using the Cross validation algorithm and K-Nearest Neighbor in data classification. The results of the study are based on the results of the evaluation of the Cross Validation algorithm on the effect of the number of K in the K-nearest Neighbor classification of data. The author tests using variations in the value of K K-Nearest Neighbor 3,4,5,6,7,8,9. While the training and test data distribution using Cross validation uses variations in the number of K-Fold 1,2,3,4,5,6,7,8,9,10.

Keywords: Classification Data, Cross Validation, K-Nearest Neighbor

INTRODUCTION

This Proccessing of classification data refers to artificial intellegience methode on focus for machine learning. Many other method in machine learning that are used for classification proccess include K-Nearest Neghbor and Naive Bayes Classifier. Classification is a grouping of object classes based on the characteristic of similarities or differences.

Classification is a tehnic that used for making classification models from training data sample. The classification will analyse data input and build the model that describe of class from data. The class label of unknown sample data can be predicted by classification techniques [1]. One of the most popular classification

K-Nearest Neighbor (k-NN) is a classic classification method that does not require prior knowledge, the new sample label is only determined by its closest neighbors [2]. K-NN can also be interpreted as a non-parametric classification method and has been widely used in the pattern classification process. The classification results are based on the process of making the most votes [3].

Jaafar et al. (2016) in his research using the k-NN method to classify a hand-based biometric image database which is a fingerprint and finger vein database, and to optimize the K-NN method for get a better percentage.

To produce data classifications that have data accuracy or similarity in proximity of a measurement result to the actual

Submitted : 30 Agt 2020 – Accepted : 26 Sep 2020 – Published : 29 Sep 2020

ISSN : 1979-9292
E-ISSN : 2460-5611

numbers or data, testing can be done based on accuracy with test data parameters and training data specified by Cross Validation. Therefore data accuracy is very influential on the final result of data classification because when data accuracy is inaccurate it will affect the percentage of test data grouping and training data. Whereas in the K-Nearest Neighbor method there is no division of training data and test data. [4] in his research to determine the ability of the accuracy of data classification and also to find out the optimal data patterns obtained in the distribution of training and testing data can be done with Cross Validation.

Cross Validation is the most commonly used method for evaluating the predictive performance of models. Data is usually divided into two parts and based on this separation in one section, training is carried out while predictive is tested in another [5].

METHOD

KNN classification is a simple non-parametric method for classification. Apart from the simplicity of the algorithm, the performance is very good, and is an important benchmark method. KNN classification requires positive metrics and integers K [42]. KNN rules hold the position of their training samples and classes. When deciding about new incoming data. The purpose of this algorithm is to classify new objects based on attribute values and training data [4].

Cover and Hart gave rise to KNN in 1968. KNN is a classification method which is also called lazy because this KNN algorithm stores all training data and makes the process delayed for the formation of classification models until the test data is given for prediction. (Mulak & Talhar, 2015). The idea in the k-Nearest Neighbor method is to identify the sample k in the training set whose independent variable x is similar to u, and use this sample k to classify this new sample into classes, v. F is a smooth function, a sensible idea is to look for samples in our training data that are nearby (in terms of independent variables) and then calculate v of the value of y for this sample. Distances or measures of inequality can be calculated between samples by measuring distances using Euclidean distances. The Euclidean distance between the points is [5].

RESULTS BACKGROUND

Modification of K-Nearest Neighbor is done with the aim of knowing the ability of data classification accuracy. This modeling is used as training data and test data to be tested with K-Nearest Neighbor.

K-Nearest Neighbor is how to determine the appropriate value of K. The general value of K is usually not optimal for all instances.Cross Validation in using a dataset, with Cross Validation can determine a large K value (but smaller than the number of instances) in data sharing. The results of the study give us about 0.1-3% more accurate results.Unbalanced data is a serious problem in machine learning. The results of his research show that Cross Validation can balance data with structured division.

As for some previous studies, [4] K-Nearest Neighbor method has problems in the provision of test data distribution and training data for the data classification process. Sanjay Yadav (2016) Cross Validation can determine a large K value

Submitted : 30 Agt 2020 – Accepted : 26 Sep 2020 – Published : 29 Sep 2020

JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i3 (211-216)

ISSN : 1979-9292
E-ISSN : 2460-5611

(but smaller than the number of instances) in the distribution of test and training data.

RESULTS AND DISCUSSION

In this study the authors analyzed the test with variations in the value (K) K-NN and the number of CV K-Fold. From the results of this test the authors also analyzed variations in the K value of the iris dataset. As shown below. From the results of the analysis using the method of Cross Validation and K-Nearest Neighbor in the classification of data that has data accuracy or similarity of closeness it is good for truly random data compared to using a dataset.

In this test using 30 test data with 4 attributes and 3 species in the data classification.

	Jumlah CV K-Fold	Result Analysis Method K-NN	Result Analysis Method K-NN and Cross Validation
Dataset Iris	1	85%	88.7%
	2	86%	98.7%
	3	77.3%	86%
	4	77%	80%
	5	81%	85%
	6	73.6%	80%
	7	68%	80%
	8	73%	81%
	9	83%	96%
	10	92%	94.7%

Table 1. Result of Variation of K-KNN and Cross Validation Method K Value

In this test using 50 test data with 4 attributes and 3 species in the data classification.

Dataset Iris	Jumlah CV K-Fold	Result Analysis Method K-NN	Result Analysis Method K-NN and Cross Validation
	1	82%	88.7%
	2	83%	88.7%
	3	88.3%	96%
	4	71%	80%
	5	81%	85%
	6	73.6%	80%
	7	68%	80%
	8	73%	95.7%
	9	84%	81%
	10	81.5%	88.7%

Table 2. Result of Variation of K-KNN and Cross Validation Method K Value

In this test use 135 test data with 4 attributes and 3 pecies in the data classification.

	Jumlah CV K-Fold	Result Analysis Method K-NN	Result Analysis Method K-NN and Cross Validation
Dataset Iris	1	95%	98.7%
	2	96%	98.7%
	3	94.3%	96%
	4	97%	100%
	5	82%	85%
	6	83.6%	89%
	7	98%	100%
	8	85%	86.7%
	9	93%	96%
	10	92%	94.7%

Table 3. Result of Variation of K-KNN and Cross Validation Method K Value

The analysis from Table 3 presents information on the accuracy of the specificity of the K-Nearest Neighbor and Cross Validation algorithms. The analysis is done by calculating the correct amount / amount of data * 100%. Accuracy is the percentage of the total number of correct predictions in the classification process [4]. This is done based on the table of Confusion for each class in the Confusion Matrix obtained on the results of training and testing.

Submitted : 30 Agt 2020 – Accepted : 26 Sep 2020 – Published : 29 Sep 2020

JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i3 (211-216)

ISSN : 1979-9292
E-ISSN : 2460-5611

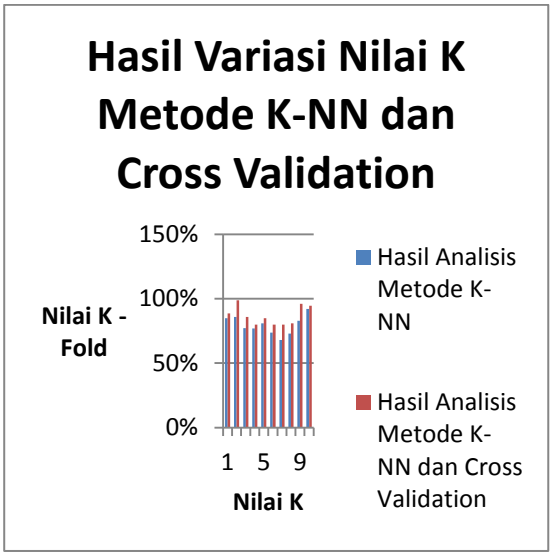

Figure 1. Test Results Variation of K Value for K-NN Method and Cross Validation with 30 Test Data

In Figure 1 above it can be seen that from the K 1 to 10 values tested the percentage of the results of the K-NN analysis method and cross validation is higher than the results of the K-NN method analysis. And from the K value that has been tested the value of K 2 and K value 9 has the largest percentage so that the accuracy is also more precise.

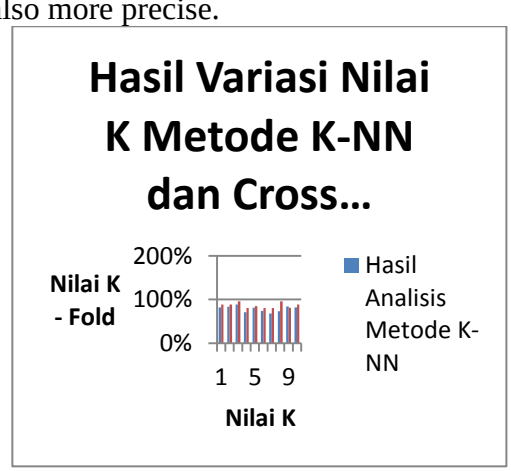

Figure 2. Test Results Variation of K Value for K-NN Method and Cross Validation with 50 Test Data

In Figure 2 above it can be seen that from the K 1 to 10 values tested the percentage of the results of the K-NN analysis method and cross validation is higher than the results of the K-NN method analysis. And from the K value that has been tested the value of K 3 and K value 8 has the largest percentage so that the accuracy is also more precise.

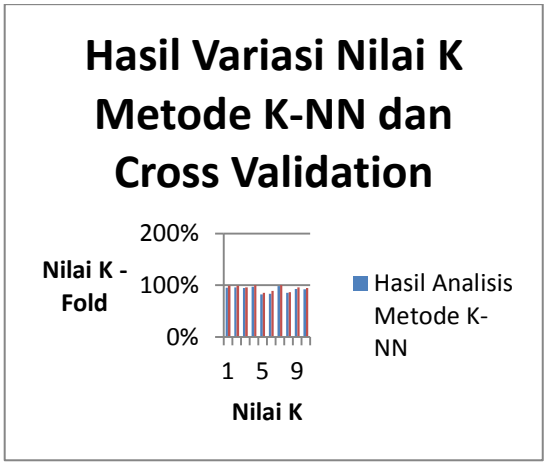

Figure 3. Test Results Variation of K Value for K-NN Method and Cross Validation with 135 Test Data

In Figure 2 above, it can be seen that from the K 1 to 10 values tested the percentage of the results of the K-NN analysis method and cross validation is higher than the results of the K-NN method analysis. And from the K value that has been tested the value of K 4 and K value 7 has the largest percentage so that the accuracy is also more precise.

As for the results of testing the K-Nearest Neighbor and Cross Validation methods in data classification. The author tests using variations in the value of K K-Nearest Neighbor 3,4,5,6,7,8,9. While the training and test data distribution using Cross validation uses variations in the number of K-Fold 1,2,3,4,5,6,7,8,9,10.

Submitted : 30 Agt 2020 – Accepted : 26 Sep 2020 – Published : 29 Sep 2020

ISSN : 1979-9292
E-ISSN : 2460-5611

CONCLUSIONS

Based on testing and evaluation of the method of determining data classification with the Cross Validation algorithm on the effect of the number of K in the K-nearest Neighbor, the conclusions that can be drawn include: Sharing data with Cross Validation has better data recognition with a percentage of 100%. From the results of the analysis using the Cross Validation method and K-Nearest Neighbor in data classifiers have good data accuracy for truly random data compared to using a dataset. The test results show the K-Nearest Neighbor and Cross Validation methods in data classification have a good percentage accuracy when using random data. Percentage of variation in the value of K K-Nearest Neighbor 3,4,5,6,7,8,9. and variations in the number of K-Fold 1,2,3,4,5,6,7,8,9,10. has a percentage of 100% on K-Fold 4 and 7.

ACKNOWLEDGMENT

Thank you to those who have helped both substantially and financially. To UniversitasPotensiUtama lecturers through the Faculty of Engineering and Computer Science who have provided theories to compile this research and we also like to thank our friends and family who supported us and offered deep insight into the research.

REFERENCES

[1] Mulak, Punam. & Talhar, Nitin. 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] Dongyin, Pan., Zhongyi Zhao., Liao Zhang., & Changzhong Tang. 2017. Recursive Clustering K-Nearest Neighbors Algorithm and the Application in the Classification of Power Quality Disturbances. ISSN (Online): 2319-7064. IEEE. pp : 1 - 5

[3] Haryati Binti Jaafar., Nordiana binti mukahar., & Dzati Athiar binti Ramli. A Methodology of Nearest Neighbor: Design and Comparison of Biometric Image Database. IEEE Student Conference on Research and Development (SCOReD). pp : 1 – 6.

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

[5]Sanjay Yadav., & Sanyam Shukla. 2016. Analysis of K-Fold Cross Validation Over Hold-Out Validation On Colossal Datasets For Quality Classification. IEEE 6th International Advanced Computing. pp : 78-83

Submitted : 30 Agt 2020 – Accepted : 26 Sep 2020 – Published : 29 Sep 2020

ISSN : 1979-9292
E-ISSN : 2460-5611

AUTHOR(S)BIOGRAPHY

Ridha Maya Faza Lubis, born on Februari 01, 1996. LabuhanBatu regency, North Sumatera.She's comes from simple family that's a culture mixture of the mandailing's andbataktoba. Graduated from high school in 2013, studied at the Department Of Industrial engineering, Faculty Of Engineering Syiah Kuala University Banda Aceh. And finished in 2018. After can graduated bachelor's degree in enginerring, in 2019 She's continued master's education at the PotensiUtama University, Department of Computer Science.

Submitted : 30 Agt 2020 – Accepted : 26 Sep 2020 – Published : 29 Sep 2020