International Journal of Research and Review
Vol.7; Issue: 4; April 2020
Website: www.ijrrjournal.com
E-ISSN: 2349-9788; P-ISSN: 2454-2237

Analysis of Performance Cross Validation Method and K-Nearest Neighbor in Classification Data

Sitefanus Hulu[1], Poltak Sihombing[2], Sutarman[2]

[1]Postgraduate Students at Universitas Sumatera Utara, Medan, Indonesia
[2]Postgraduate Lecturer at Universitas Sumatera Utara, Medan, Indonesia

Corresponding Author: Sitefanus Hulu

ABSTRACT

To produce data classifications that have 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 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. Based on the evaluation results of the Cross Validation algorithm on the effect of the number of K in the K-nearest Neighbor classification data. The data sharing with Cross Validation has better data recognition with a percentage of 100%. The results of the K-NN test results in the classification of data using iris data sets using variation test values 3, 4, 5, 6, 7, 8, 9, have 100% percentage accuracy with 75 true amount of data and 0 incorrect amount of data. Percentage of variation in 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

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 technique is K-Nearest Neighbor (KNN)..

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 neighbours. [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) dalam penelitiannya menggunakan metode k-NN untuk mengklasifikasi basis data gambar biometrik berbasis tangan (hand-based biometric) yang merupakan basis data sidik jari (fingerprint) dan vena jari (finger vein), serta melakukan optimasi pada metode K-NN untuk mendapatkan persentase yang lebih baik.

To produce data classifications that have 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 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]

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

RESEARCH METHODS

It is necessary to test in determining the distribution of training data and test data on the effect of the number of K in the K-Nearest Neighbor on a large dataset. Existing dataset uses the UCI Machine Learning Repository. The UCI Machine Learning Repository is a collection of databases, domain theory, and data generators used by communities who study machine learning, for the purposes of empirical analysis of machine learning algorithms. The dataset available at the UCI Machine Learning Repository is used by students, educators, and researchers worldwide as the main source of data sets in machine learning. The number of data sets available on the UCI Machine Learning Repository currently amounts to 320 data sets that can be used according to the needs of machine learning.

This research process has several activities, namely the activities contained in the study namely field observation, data collection and data analysis.

1. Observations made in this study are the most important things. Because the author can find out the level of visibility used. The data that has been collected becomes a monitoring point in this observation so that it gets the results used.
2. Collecting data on literature, journals, papers and other readings related to the classification algorithm for data mining. The researcher collected data by taking the UCI Machine Learning Repository dataset. Researchers use the iris dataset.
3. Preparation and selection of data obtained is like looking at the structure of tables in the database. The data selection is done because not all tables and data in the database are related to the research conducted, so only the data relating to the research is used.
4. Data cleaning is done to ensure that no data is duplicated, check for inconsistent data, and correct errors in the data. Data that has been cleared of errors can

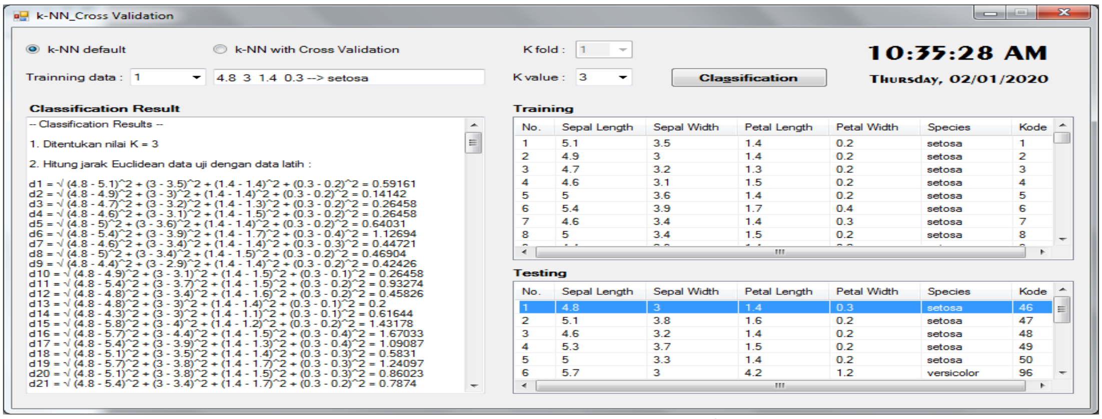
simplify research and prevent errors in research.

5. At this stage the analysis of the distribution of training data and test data with the Cross Validation algorithm on the effect of the K number of K-Nearest Neighbor.

RESULT

Cross Validation and K-NN Test Results.

In this test the next method is cross validation and K-NN. This test aims to see a discussion of the performance of 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. The following shows the test results.



● k-NN default ○ k-NN with Cross Validation

K fold : 1

Trainning data : 1 4.8 3 1.4 0.3 --> setosa

K value : 3

Classification Result

```
-- Classification Results --

1. Ditentukan nilai K = 3

2. Hitung jarak Euclidean data uji dengan data latih :

d1 = √ (4.8 - 5.1)^2 + (3 - 3.5)^2 + (1.4 - 1.4)^2 + (0.3 - 0.2)^2 = 0.59161
d2 = √ (4.8 - 4.9)^2 + (3 - 3)^2 + (1.4 - 1.4)^2 + (0.3 - 0.2)^2 = 0.14142
d3 = √ (4.8 - 4.7)^2 + (3 - 3.2)^2 + (1.4 - 1.3)^2 + (0.3 - 0.2)^2 = 0.26458
d4 = √ (4.8 - 4.6)^2 + (3 - 3.1)^2 + (1.4 - 1.5)^2 + (0.3 - 0.2)^2 = 0.26458
d5 = √ (4.8 - 5)^2 + (3 - 3.6)^2 + (1.4 - 1.4)^2 + (0.3 - 0.2)^2 = 0.64031
d6 = √ (4.8 - 5.4)^2 + (3 - 3.9)^2 + (1.4 - 1.7)^2 + (0.3 - 0.4)^2 = 1.12694
d7 = √ (4.8 - 4.6)^2 + (3 - 3.4)^2 + (1.4 - 1.4)^2 + (0.3 - 0.3)^2 = 0.44721
d8 = √ (4.8 - 5)^2 + (3 - 3.4)^2 + (1.4 - 1.5)^2 + (0.3 - 0.2)^2 = 0.46904
d9 = √ (4.8 - 4.4)^2 + (3 - 2.9)^2 + (1.4 - 1.4)^2 + (0.3 - 0.2)^2 = 0.42426
d10 = √ (4.8 - 4.9)^2 + (3 - 3.1)^2 + (1.4 - 1.5)^2 + (0.3 - 0.1)^2 = 0.26458
d11 = √ (4.8 - 5.4)^2 + (3 - 3.7)^2 + (1.4 - 1.5)^2 + (0.3 - 0.2)^2 = 0.93274
d12 = √ (4.8 - 4.8)^2 + (3 - 3.4)^2 + (1.4 - 1.6)^2 + (0.3 - 0.2)^2 = 0.45826
d13 = √ (4.8 - 4.8)^2 + (3 - 3)^2 + (1.4 - 1.4)^2 + (0.3 - 0.1)^2 = 0.2
d14 = √ (4.8 - 4.3)^2 + (3 - 3)^2 + (1.4 - 1.1)^2 + (0.3 - 0.1)^2 = 0.61644
d15 = √ (4.8 - 5.8)^2 + (3 - 4)^2 + (1.4 - 1.2)^2 + (0.3 - 0.2)^2 = 1.43178
d16 = √ (4.8 - 5.7)^2 + (3 - 4.4)^2 + (1.4 - 1.5)^2 + (0.3 - 0.4)^2 = 1.67033
d17 = √ (4.8 - 5.4)^2 + (3 - 3.9)^2 + (1.4 - 1.3)^2 + (0.3 - 0.4)^2 = 1.09087
d18 = √ (4.8 - 5.1)^2 + (3 - 3.5)^2 + (1.4 - 1.4)^2 + (0.3 - 0.3)^2 = 0.5831
d19 = √ (4.8 - 5.7)^2 + (3 - 3.8)^2 + (1.4 - 1.7)^2 + (0.3 - 0.3)^2 = 1.24097
d20 = √ (4.8 - 5.1)^2 + (3 - 3.8)^2 + (1.4 - 1.5)^2 + (0.3 - 0.3)^2 = 0.86023
d21 = √ (4.8 - 5.4)^2 + (3 - 3.4)^2 + (1.4 - 1.7)^2 + (0.3 - 0.2)^2 = 0.7874
```

Training

No.	Sepal Length	Sepal Width	Petal Length	Petal Width	Species	Kode
1	5.1	3.5	1.4	0.2	setosa	1
2	4.9	3	1.4	0.2	setosa	2
3	4.7	3.2	1.3	0.2	setosa	3
4	4.6	3.1	1.5	0.2	setosa	4
5	5	3.6	1.4	0.2	setosa	5
6	5.4	3.9	1.7	0.4	setosa	6
7	4.6	3.4	1.4	0.3	setosa	7
8	5	3.4	1.5	0.2	setosa	8

Testing

No.	Sepal Length	Sepal Width	Petal Length	Petal Width	Species	Kode
1	4.8	3	1.4	0.3	setosa	46
2	5.1	3.8	1.6	0.2	setosa	47
3	4.6	3.2	1.4	0.2	setosa	48
4	5.3	3.7	1.5	0.2	setosa	49
5	5	3.3	1.4	0.2	setosa	50
6	5.7	3	4.2	1.2	versicolor	96

Gambar 4.1 Test Result (k-NN & CV 1 fold)

Test Results (k-NN & CV 10 fold) K-NN method in the classification of data using iris data sets with variation test values of 3, 5, 7, 8, 9, has a percentage accuracy of 94.7% with 71 true amount of data and 4 amount of data is wrong. this is because there are different species in data 1 with k values 5 and 8, and data 4 at k values 8 and 9.

As for the results of testing using Cross Validation all datasets are used as test data and training data to produce a good classification. The examiner divides the data into 10 K-Fold Cross Validation from 15 data, where the results show that the random data has a better percentage of data classification than the dataset that has been determined for each piece.

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 the analysis show In this test the authors also analyzed variations in the value of K from the iris dataset. As shown below. This test using of 30 data test with 4 atribute dan 3 species in classification data.

Tabel 4.1 The variation result of K from K-NN Method dan Cross Validation

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

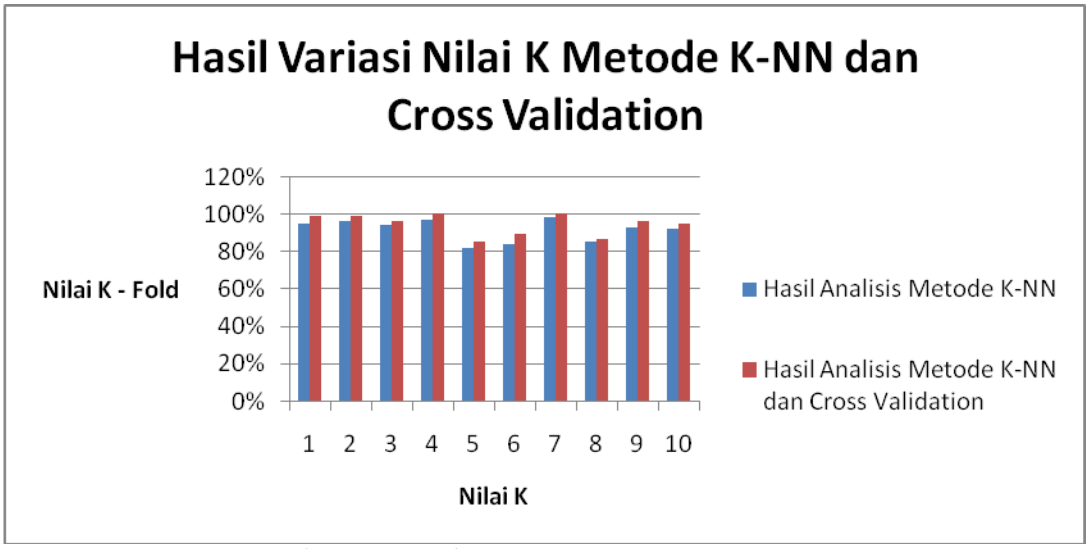
In this test use 135 test data with 4 attributes and 3 species in the data classification..

Tabel 4.2 Analysis result of K-NN Methode and Cross Validation

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

The analysis from Table 4.2 presents information on the accuracy of the specificity of the K-Nearest Neighbor and Cross Validation algorithms. Analysis is done by calculating Correct amount / amount of data * 100%.

Accuracy is the percentage of the total number of correct predictions in the classification process. This is done based on the table of Confusion for each class in the Confusion Matrix obtained on the results of training and testing.



Gambar 4.3 The Result of Variation Test from K-NN Methode and Cross Validation with 135 Data Test

In Figure 4.3 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 K value 4 and the 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. Has a very good percentage of accuracy compared to only K-NN. 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.

CONCLUSION

Based on the evaluation results of the Cross Validation algorithm on the effect of the number of K in the K-nearest

Neighbor classification data. The data sharing with Cross Validation has better data recognition with a percentage of 100%. K-NN test results in data classification using iris data sets using variation test values of 3, 4, 5, 6, 7, 8, 9, have a percentage accuracy of 100% with 75 correct amount of data and 0 incorrect amount of 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.

REFERENCES

1. Mulak, Punam. & Talhar, Nitin. 2015. Analysis of Distance Measures Using K-Nearest Neighbor Algorithm on KDD Dataset. International Journal of Science and 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

How to cite this article: Hulu S, Sihombing P, Sutarman. Analysis of performance cross validation method and K-nearest neighbor in classification data. International Journal of Research and Review. 2020; 7(4): 69-73.

******