International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

http://ijpsat.ijsht-journals.org
Vol. 10 No. 1 August 2018, pp. 158-162

Classification Accuracy of K-Nearest Neighbours Algorithm to Predict Rice Quality

Andria Rezki, Herman Mawengkang, Syahril Efendi, Husnul Khair

Fakultas Ilmu Komputer danTeknologi Informasi Universitas Sumatera Utara

Abstract - **This study aims to produce a system that is efficient and effective in predicting the results of a useful decision to determine the accuracy of ricequality with the selection of features/attributes by using Information Gain applied to K-Nearest Neighbor algorithm. One function of the k-nearest neighbor algorithm is as a classification. In its function as a classification in algorithm pattern recognition, k-nearest neighbor is a method of grouping objects based on examples of the nearest training data. This k-nearest neighbor algorithm is the most basic classification technique and is also very simple. In the search for these solutions, several attributes are used for classification. But in its performance there is the possibility of an attribute that does not have a suitable value in data mining, when the attribute is entered into the classification process can cause confusion in the data mining process. Therefore, it is necessary to do attribute selection so that it can identify attributes with values that are irrelevant or relevant and then discard these attributes. In this case, the author uses the information gain method as an identification and also performs attribute selection.**

Keywords - **K-Nearest Neighbor, Classification, Information Gain, Selection, Entropy.**

I. INTRODUCTION

One function of the k-nearest neighbor algorithm is as a classification. In accordance with one of its uses as a classification in recognizing patterns, the k-nearest neighbor algorithm is a method of grouping objects based on examples of the closest training data. K-nearest neighbor algorithm is the most basic and simple classification method when there is little or no knowledge of the distribution of data that is known beforehand. KNN is a lazy algorithm, meaning that the training phase is minimal and the full dataset is used as the test phase and it is also fast, independent of the classification algorithm.

When we have to find a pattern by analyzing large data it will be difficult to compare it by visualizing it with raw data. So here the author uses the K-Nearest Neighbor algorithm method as a classification method.

To carry out the classification process, several studies have been conducted on the application of data mining in determining the quality of rice quality using the k-nearest neighbor algorithm.

K-nearest neighbor algorithm is usually very related to a sustainable attribute, and from this study also shows that k-nearest neighbor is one way to mine a data that is most widely used to classify because of itssimplicity, convergence and very high speed. (JICT, Application of k-nearest neighbor classification vol 4 2014).

The k-nearest neighbor algorithm is a method which uses a set of new instances, which is juxtaposed with the weight of a number of features that have been formed. For example, the desire to seek the best results for the quality of new rice by using a solution of the quality of the previous rice. To find out which rice quality cases to use, the closeness of the new rice case was calculated with all the old rice cases. In the search for these solutions, several attributes are used to be classified. But in its performance there is a possibility that happens is an attribute that does not have a match value in data mining; if the attribute is entered into the classification process can cause confusion in the data mining process. Therefore it is necessary to select attributes so that they can identify attributes with values that are not suitable or relevant and can eliminate these attributes.

To identify relevant factors that affect the quality of rice, the Information Gain method is implemented as a feature selection technique. Information Gain uses entropy in getting the best attributes. Entropy is a measure of inequality. More the information gain, stronger the attribute.

Corresponding Author: Andria Rezki

So, the partition takes place on the attribute which has the highest information gain, so the attribute is clear for the quality of rice.

According to Mehmed Kantardzic (Data Mining. Concepts,Models,Methods and Algorithms) In practice, the two primary goals of data mining tend to be prediction and description. Prediction involves using some variables or fields in the data set to predictunknown or future values of other variables of interest. Description, on the other hand,focuses on finding patterns describing the data that can be interpreted by humans.Therefore, it is possible to put data - mining activities into one of two categories:predictive data mining, which produces the model of the system described bythe given data set, or descriptive data mining, which produces new, nontrivial information based onthe available data set.

On the predictive end of the spectrum, the goal of data mining is to produce amodel, expressed as an executable code, which can be used to perform classification. It canimprove marketing campaigns and the outcomes can be used to provide customers withmore focused support and attention. Data - mining techniques can be applied to problemsof business process reengineering, in which the goal is to understand interactions andrelationships among business practices and organizations.The success of a data - mining engagement depends largely on the amount of energy,knowledge, and creativity that the designer puts into it. In essence, data mining is likesolving a puzzle.

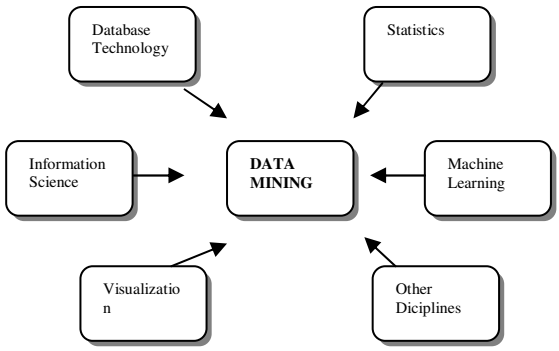
The principle of basic modeling in data mining also has roots in control theory applied to engineering systems and industrial processes. One of the main tasks of data mining is classification where a new data is matched or also called data testing with training data or pre-existing data, so as to provide effective and accuracy results.

A. Data Mining Classification

Data mining has its origins in various disciplines, of which the two most important are statistics and machine learning, visualization and scientific information. The system also relies on data mining approaches used such as neural networks, fuzzy, representations of knowledge, and high-level computing. Data mining classification systems can also integrate techniques from a data and information analysis, recognition to image analysis patterns and so on.

Data mining procedure is also expected to produce a variety of systems from mining a data. Therefore, it is very necessary to provide a clear system of data mining classification, which can later help to distinguish a system from the process and identify which are most appropriate based on needs.

The following is an image of data mining based on several combinations of categories:



(**DATA MINING,** Concepts and Techniques.*Jiwai Han and Micheline Kamber* 2nd Edition)

B. K - Nearest Neighbor algorithm

K-nearest neighbor algorithm or it can be called the K-NN algorithm is a method commonly used to classify objects that become the basis of the nearest learning data on the object. KNN includes supervised learning algorithms or predetermined, where from a new data are classified based on many parts of the categories in the KNN. The intuition behind this method is that when the distanceis calculated, the point with the least distance is selectedas the neighbor(k times). As the distance calculated is thesum of the squares of distances between individualattributes, minimizing the effective distance of an attributecontributes in that point being selected as a neighbor. So,we are taking the strength of the attribute intoconsideration while we calculate the distance. That is whyif the distance of an attribute is more but it is a strongerattribute than others(that is, if we choose that value of anattribute, we can tell with maximum certainty (relative toother attributes) that a particular class is going to beassigned to the test example). The KNN algorithm will work based on the shortest distance from the data sample data to be searched using training data.

The best neighbor k value in the k-nearest neighbor algorithm will depend on the data. Based on high k values, it reduces the result of blurring in a classification, but makes the boundary between each classification even more blurred. One particular problem where the classification results are based on nearest training data , k = 1 is called the nearest neighbor algorithm (T. Hastie and R. Tibshirani, 1996)

C. Research Problem

How is the accuracy of rice quality with the selection of features/attributes by using information gain applied to K-Nearest Neighbor algorithm?

II. Material And Method

A. Research Methodology

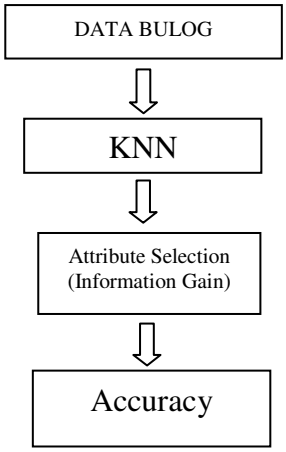


Figure 1 Research stage diagram

The research methodology is carried out in three stages, namely data collection, algorithm analysis and attribute / feature selection and testing.

Data collected from literature, journals, papers and references related to this research.Data were analyzed using PHP and MySQL based programming.

Testing using the holdout method, research data will be divided into two, 2/3 as training data and then 1/3 as data testing (Han and Kamber, 2006).

III. Result

The steps of testing, as follows:

1. The first stage of training data will be processed using the k-Nearest Neighbor algorithm for the entire attribute. From the training data it will be classified. Then the testing data is tested to get predictive results with accuracy and error.
2. In the second stage of training data is the information gain algorithm process. Each attribute will be calculated for its Gain information and then it will be sorted from the highest to the lowest value. The lowest attribute is discarded, and the rest is done by using the K-nearest Neighbor algorithm for predictive results with accuracy and error values.
3. The process of the results of the second stage, see the change in the value of accuracy and error from the first stage process.
4. Testing will be implemented on the application that will be created. The research data was also tested in a data mining application that is Waikato Environment for Knowledge Analysis (WEKA) to ascertain whether the algorithm is implemented correctly.
5.

A. Test result

Measurement of the value of accuracy and speed from testing training data and testing data is done using a classification table, confusion matrix, and measured process speed in the length of time needed in the training and testing of data. The number of training data consists of 20 rows (2/3 of the number of rows of research data and the number of testing data consists of 15 rows of data (1/3 of the row of research data).

B. First Stage Testing Results

The test results for the first stage are carried out without discard attributes from the dataset and the classification process is carried out by KNN algorithm. From the results of 15 training data tested in the testing data, the prediction results shown in Table 1:

Table1 Prediction Value

		Prediction Class	
		Good	Poor
Actual Class	Good	4	1
	Poor	2	3

From the table above, we get the correct predictive value, 4, good quality and 3, poor quality. While the wrong prediction consists of 2 good and 1 poor quality (which in actual not poor). Value of accuracy and error, as follows:

First Stage Accuracy: $\frac{TP+TN}{P+N} = \frac{4+3}{10} = 0.7 = 70\ \%$

First Stage Error:

$\frac{FP+FN}{P+N} = \frac{2+1}{10} = 0.3 = 30\ \%$

The time of the testing process from data testing involving all attributes of table 1 is: 0.043 seconds.

C. Second Stage Testing Results

The results of the second stage of testing are carried out without removing the attributes of the existing datasets and performing the classification process with the K-NN algorithm. Unlike the first stage of testing, this time there is 1 attribute to be discarded, the degree milling attribute has the smallest information gain value. From the results of 15 of the training data, testing of 10 rows of testing data and the results of predictions in the matrix confusion table were conducted. For calculating these criteria we used the confusion matrix in our calculation process. In confusion matrix the predicted class is the class that is predicted by the classifier and the actual class isthe class that is given in the data set.

Table 2 Second stage testing

		Prediction Class	
		Good	Poor
Actual Class	Good	4	1
	Poor	2	3

From the table above we have actual, good quality, 4, and poor quality predictions, 3. While 2 predictions that were wrongly predicted as poor quality rice and 1 predicted poor (in actual not poor). Value of accuracy and error, as follows:

Second stage of accuracy: $\frac{TP+TN}{P+N} = \frac{4+3}{10} = 0.7 = 70\ \%$

Second stage of error:

$\frac{FP+FN}{P+N} = \frac{2+1}{10} = 0.3 = 30\ \%$

The duration of testing for all attributes in table 2 is: 0.043 seconds.

D. Third Stage Testing Results

The results of the third stage testing were carried out without removing the attributes of dataset, and the classification process was carried out with the K-NN algorithm. Unlike the first stage of testing, this time there are 2 attributes that will be discarded, the attributes of the milling degree and brokengrain attribute, the value of information gain is 0.009 and 0.009. From the results of 15 of the training data, testing of 10 rows of testing data and predictions are shown in the confusion matrix table, as follows:

Table3 Confusion Matrix

		Prediction Class	
		Good	Poor
Actual Class	Good	4	1
	Poor	2	3

From the table above the accurate predictive values for good quality, 4, and poor quality rice, 3. While the wrong prediction, 2, predicted poor quality rice and 1 predicted poor (in actual not poor). Value of accuracy and error, as follows:

Third stage of accuracy:

$\frac{TP+TN}{P+N} = \frac{4+3}{10} = 0.7 = 70\ \%$

Third stage of error : $\frac{FP+FN}{P+N} = \frac{2+1}{10} = 0.3 = 30\ \%$

The duration of testing for all attributes in table 2 is: 0.037 6 seconds.

E. Fourth Stage Testing Results

The fourth stage of the test results is done without removing the attributes of the dataset, and then the classification process with the K-NN algorithm is carried out. Unlike the first stage of testing, 3 attributes were discarded, the degree of milling, broken grain, and water content which had information gain values of 0.009, 0.0506, and 0.1779. From the results of 15 of the training data, testing of 10 rows of testing data was conducted, so the predictions in the confusion matrix table are as follows:

Table 4 Confusion Matrix

		Prediction Class	
		Poor	Buruk
Actual Class	Good	4	1
	Poor	3	2

From the table above, we get an accurate prediction value for good quality 4 and for poor quality rice, 3. While the wrong prediction, 2, predicted poor quality rice and 1 predicted poor (actual not poor). Value of accuracy and error, as follows:

Fourth stage of accuracy: $\frac{TP+TN}{P+N} = \frac{4+2}{10} = 0.6 = 60\ \%$

Fourthstage of error: $\frac{FP+FN}{P+N} = \frac{3+1}{10} = 0.3 = 40\ \%$

The duration of testing for all attributes in table 2 is: 0.0362 seconds.

IV. DISCUSSION

From the results of the above tests can be seen a comparison of the value of accuracy, error and time.

The process in the first to fourth stages is shown in table 5:

Table 5 Accuracy values

Testing	Discarded Attribute	Attribute	Accuracy	Error	Speed
First	-	8	70 %	30 %	0.043
Second	Milling degree	7	70 %	30 %	0.039
Third	Milling degree and Broken grain degree	6	70 %	30 %	0.037
Fourth	Milling degree,broken grain degreeand moisture degree	5	60 %	40 %	0.0362

From figure2, graphs of the test results can be seen that the results of testing in the first, second, and third stages get the same accuracy and have no effect in the classification process with 70% accuracy and 30% error. Unlike the fourth stage when 3 attributes were discarded, the accuracy of the classification process decreased by 10% to an accuracy rate of 60% and the error rate to 40%.

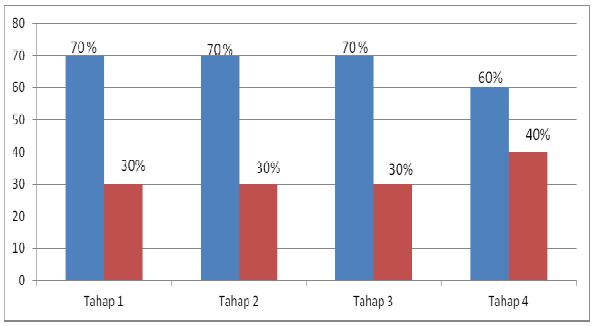


Figure 2 Testing Result Graph

V. CONCLUSIONS AND SUGGESTIONS

A. Conclusions

The conclusions that can be drawn from this study are as follows:

1. From the attribute selection process that has been carried out using the information gain method, the recommended attribute not to be included is the degree milling attribute which has the smallest information gain value of 0.008954 and produces the same accuracy value of 70% when not included and produces an accuracy value same if the attribute is not discarded during the classification process with the K-NN algorithm
2. When the three attributes are removed, the degree of error, grain break, and water content, accuracy decreases to 60%. Information Gain value from moisture degree attribute 0.17790. However, if the two attributes are removed, the grain is broken and the degree of milling is discarded, the accuracy does not decrease and has an accuracy of 70%. It can be concluded that if the discarded moisture degree attribute has an impact on accuracy due to differences in information gain values, the information gain value of the milling degree is 0.008954 and broken grain 0.050587.
3. By using the attribute selection process, the processing time for training and testing data is slightly faster than using the overall attribute. This can be seen in the first phase of the experiment which takes 0.043 seconds, the second stage requires the same time as the first stage 0.043 seconds, the third stage takes 0.037 6 seconds, and the fourth stage takes 0.0362 seconds.

B. Suggestion

The limitations of this study are expected to be developed in the future by using other methods and algorithms that can support the use of applications.

REFERENCE

[1] A Generalized k-Nearest Neighbor Rule, E. A. PATRICK AND F. P. FISCHER, III
School of Electrical Engineering, Purdue University, Lafayette, Indiana

[2] Jiawei Han and Micheline Kamber, Data Mining Concept and Techniques 2nd Edition

[3] Cover, T. and Hart, P. (1967) Nearest neighbor pattern classification, IEEE Trans Information Theory 13, page(s): 21–27.

[4] Domingos, P. (1999) MetaCost: A general method for making classifiers costsensitive. In Proceedings of the Fifth International Conference on Knowledge Discoveryand Data Mining, pp. 155–164.