Bulletin of Electrical Engineering and Informatics
Vol. 9, No. 1, February 2020, pp. 326~338
ISSN: 2302-9285, DOI: 10.11591/eei.v9i1.1464

Optimization of distance formula in K-Nearest Neighbor method

Arif Ridho Lubis[1], Muharman Lubis[2], Al-Khowarizmi[3]
[1,3]Department of Computer Enginnering and Informatics, Politeknik Negeri Medan, Indonesia
[2]Telkom University, Bandung, Indonesia
[3]Department of Information System, Universitas Muhammadiyah Sumatera Utara, Indonesia

Article Info

Article history:

Received Dec 29, 2018
Revised Mar 2, 2019
Accepted Sep 26, 2019

Keywords:

Distance formula
K-Nearest Neighbor
Optimization

ABSTRACT

K-Nearest Neighbor (KNN) is a method applied in classifying objects based on learning data that is closest to the object based on comparison between previous and current data. In the learning process, KNN calculates the distance of the nearest neighbor by applying the euclidean distance formula, while in other methods, optimization has been done on the distance formula by comparing it with the other similar in order to get optimal results. This study will discuss the calculation of the euclidean distance formula in KNN compared with the normalized euclidean distance, manhattan and normalized manhattan to achieve optimization results or optimal value in finding the distance of the nearest neighbor.

Corresponding Author:

Arif Ridho Lubis,
Department of Computer Enginnering and Informatics,
Politeknik Negeri Medan, Indonesia.
Email: arifridho@polmed.ac.id

1. INTRODUCTION

Classification techniques in conducting the process to find a model or function explaining and characterizing the concepts or data classes, for specific purposes [1]. Many techniques or methods applied in the classification, which one of them is the K-Nearest Neigbor (KNN) method for classifying objects based on learning data of which the closest distance to the object. Actually, classification means the attempt to predict certain case fall under specific category or class, differs with regression that focus on number alue a variable will have [2, 3]. Learning data is projected into a large dimension space, where each dimension represents a feature of the data, which is divided into sections based on instance-based learning or lazy learning where the function is only approximated locally and all computation is deferred until classification [4, 5]. Furthermore, learning in the KNN method passes through a point in space of a space marked from a class if it is the most commonly found classification in the nearest neighbor data closest to that data. The distance of the neighbors in learning the KNN method is usually calculated based on Euclidean distance [6]. Therefore, the regulation and policy should be considered first before the implementation takes place due to decision point leading to quality of the result as well as effectiveness and efficiency [7-9].

Academician conducted a research [10] applying KNN where in the learning process also apllied Euclidean distance in classifying recruiting prospective teachers and employees at vocational high schools by combining Weighted Product (WP) methods with the results of several criteria weight values. It was obtained the value of accuracy is 94%, 80% precision, and 80% recall value. While [11] also conducted research with

KNN where it applied a prediction system classification for the students' achievement, in the system also applied the Euclidean distance formula in the learning process with the results through the Euclidean distance formula applied in predicting the students'achievement scores resulted in accurate value of 82%. On the other hand, one research [12] applying KNN, in their learning they also applied the Euclidean distance formula to make new weighting. In addition, conducted [13] a classification study applying KNN and support vector mechane (SVM) with results of accuracy of more than 99.83%, sensitivity more than 0.995 and specificity of more than 0.998.

From those studies the average research with KNN applied the Euclidean distance formula in the learning process. In a different method, several researchers conducted an optimization in the method by performing an optimization on the distance formula. Conducted [14] an optimization of the Simple Evolving Connectionist Systems (SeCOS) method by testing the Normalized Euclidean distance formula, Normalized Manhattan and Normalized Hamming. Furthermore, optimized [15] configuration of discrete wavelet frame (DWF) applying the texture feature extraction method in images involving manhattan distance, euclidean distance, normalized manhattan distance and normalized euclidean distance. Based on the previous research conducted by the KNN method, it was necessary to optimize the search for the closest distance by comparing several distance formulas. The optimization process replaces the euclidean distance formula with the normalized euclidean distance formula, manhattan and normalized Manhattan to obtain optimal calculation results. The sample data used were creditcard payment usage data with 30000 datasets and 23 attributes achieved from UCI Machine Learning. This study implement k-Nearest Neigbor for big data analytics as the action of making the best or most effective use of a situation or resource to help identify the optimal value that allow the process to be simplified in certain cases.

2. METHODOLOGY

KNN is a method of classifying objects based on learning data that is closest to the object. This method aims at classifying new objects based on attributes and training samples. Given a query point, then it will find a number of K objects or training points closest to the query point. The predicted value of the query will be determined based on the neighbor classification. Before performing calculations using the K-Nearest Neighbor method, the training and test data must firstly be determined. Then the calculation process will be carried out to find distances applying the Euclidean distance formula. It is a very simple technique and easy to implement. Similar to clustering techniques, grouping new data based on their distance to some of the closest data/neighbors. The similarity function will produce a value determining whether there are similarities between the new cases and those in the case base. To determine the similarity can be done with several functions, i.e. with the similarity euclidean distance function. The disadvantage of this Euclidean distance function is that if one attribute input has a relatively large range, it can defeat other attributes. Consequently, distance is often normalized by dividing the distance for each attribute with the range (i.e. the maximum value-minimum value) of the attribute so that the values for each attribute have a normalized new range of 0 to 1. The (1) is a formula for normalizing data, i.e. data is range 0 to 1.

$$y = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where; x=value of the data
y=value of normalisation
x_{min}=minimum value i.e. 0
x_{max}=maximum value i.e. 0

The types of this method, if seen from its N value are as follows:

- 1-NN, predictions are made on 1 closest labeled data.
- Calculate the distance between new data to each labeled data.
- Determine 1 labeled data that has the most minimum distance.
- Predicting the new data into labeled data.

After normalization, the data then calculates the proximity value. This calculation process is applied in finding predictions using the Euclidean distance formula. The equation of 2 formulas for calculating proximity between two cases is as follows.

$$similarity(T, S) = \frac{\sum_{i=1}^{n} f(T_i, S_i)x}{w_i} \quad (2)$$

where: t=new case
s=the value of the closeness of the case in storage
n=number of attributes in each case

i=individual attributes between 1 to n
f=similarity attribute function of i between case T and case S
w=the weight given to the attribute i

The (3) is applied on the new data calculated with all old ones. In the calculation process of them, it was obtained the length of time calculated by the system. However, in this study optimization of the KNN method was carried out by changing or replacing the euclidean distance formula with the hamming distance formula and the distance distance formula in order to find a more optimal value of closeness. The hamming distance formula is seen in 3.

$$D_n = \frac{\sum_i^K |I_i - W_i|}{\sum_i^K |I_i + W_i|} \quad (3)$$

where: K=number of attributes in each case
I=new case
W=the value of the proximity of the case in storage

After obtaining the results of the two distance values, then comparing with the Manhattan distance formula. The manhattan distance formula is seen in (4). After obtaining the value of the three distance formulas, the next step is to discuss them in the KNN method.

$$D_n = \sum_i^K \frac{|I_i - W_i|}{k} \quad (4)$$

2.1. Data used

The data used are Marketing Bank data obtained from UCI Machine Learning. They are those of prospective bank customers who would be predicted to make credit. The data consists of 41188 customers consisting of 6 attributes, namely work, marital status, education, owning a home loan, owning bank loans and agreeing to the credit initiated by the bank. The data are research data that have been done [16], namely research predicting bank telemarketing.

2.2. General architecture

The general implemented architecture of the method is illustrated in Figure 1. That can be explained in stages as follows:

- The having achieved dataset is stored in the database by entering all old data.
- Input new data to get proximity value before processing the data to be trained it is normalized to have a range from 0 to 1.
- Calculate the value of the proximity of the new case with the entire old case by applying the euclidean distance formula.
- Calculate the value of the proximity of the new case with the entire old case by applying the hamming distance formula.
- Calculate the value of the proximity of the `new case with the entire old case by using the Manhattan distance formula.
- Displays the results of the proximity value based on the three distance formulas used.
- Give conclusions which distance formula is more optimal to use with existing case data.

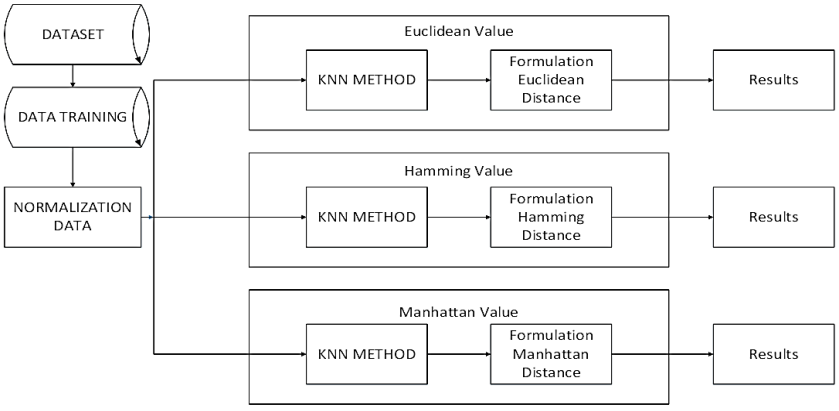

Figure 1. General architecture

3. RESULTS AND DISCUSSION

3.1. Analysis of the nearest neighbor method

Nearest neighbor is an approach to search cases by calculating the closeness between new cases and old cases, which is based on matching weights of a number of existing features [5]. For example like finding a solution for a new customer using a solution from a previous customer. To find out which customers will be used, then the closeness of the case of new customers is calculated with all cases of old customers. Interestingly, instance-based learning has several advantages over rule based classification methods such as it is very robust that can outperform conventional parametric classifiers when the actual distribution of data is different from the assumed distribution. It also establishes the decision boundary automatically based on a traning set that can be incrementally refined when new training samples are added to the existing samples [17]. However, there are many techniques in *k*-Nearest Neighbor such as weighted kNN, condensed kNN, reduced kNN, model based kNN, rank kNN, modified kNN, pseudo/generalized NN, clustered kNN, Ball Tree kNN, k-d tree, nearest feature line neighbor (NFL), local NN, tunable NN, center based NN, principal axis tree NN and othogonal search tree NN [18]. Mostly, those techniques focused on good performance, less computation time, fast search and effective for large data sets. Therefore, when there is littele or no prior knowledge about the distribution of the data, the KNN method should be one of the first choices for classification, because it is a powerful non-parametric classification system which bypasses the problem of probability desitiees completely [19, 20]. Accuracy of KNN is kept high in most of the cases but as size of dataset increases lead to the decreases, so with the time taken to calculate all required values for result that increases as the dataset become larger [21]. The case of old customers with the greatest closeness will be taken to be used in the case of new customers. From Figure 2, there are 4 old customers, namely A, B, C, and D. When there are new customers, the solution will be taken by finding the distance between new customers and all old customers. With the closest distance is the solution from the old customer, from the figure above the old customer solution B will be used because it has the shortest distance. In this study we will test the proximity value with 3 distance formulas, including the Euclidean, hamming and manhattan distance formulas. Of the three distance formulas, the optimal value is achieved so that the result of changing the distance formula can optimize the KNN method with the test data are the bank telemarketing data.

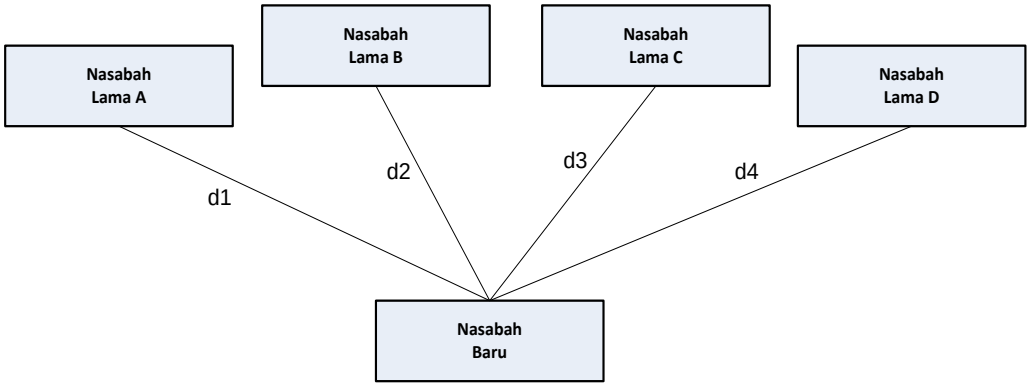

Figure 2. Ilustration

3.2. Problem solving analysis

Examples of cases, for example, to predict whether the new bank customers have problems or not based on the data.

a. Case table

The following Table 1 is an example of case of old customers

Table 1. Example of case table of old customers

Old Customers' case						
Name	Education	Status	Home Credit	Bank Credit	Occupation	Agree
A	>=Bachelor	Single	No	No	Enterpreneur	no
B	<=High School	Married	Yes	No	Enterpreneur	yes
C	>=Bachelor	Single	No	Yes	Private employees	yes
D	D1-D3	Married	No	Yes	Civil Servant	no

b. Determine the weight of each attribute
The following Table 2 is weight of each attribute

Table 2. Weight of each attribute

Attribute	Weight
Education	0.5
Status	0.5
Home Credit	1
Bank Credit	1
Occupation	0.75

c. Next step to determine the closeness of the value in the attribute
- The closeness of the value in the attribute. The following Table 3 is an example of of the closeness of the value in the attribute of education.

Table 3. Tabel of the closeness of the value in the attribute of education

Education	Education	Closeness
<=High School	<=High School	1
Diploma	Diploma	1
>=Bachelor	>=Bachelor	1
<=High School	Diploma	0.5
Diploma	<=High School	0.5
<=High School	>=Bachelor	0.4
>=Bachelor	<=High School	0.4
Diploma	>=Bachelor	0.75
>=Bachelor	Diploma	0.75

- The Closeness of the Status Value. The following Table 4 is an example of of the closeness of the status value.

Table 4. The tabel of the closeness of the status value

Status	Status	Closeness
Single	Single	1
Married	Married	1
Divorced	Divorced	1
Single	Married	0.5
sMarried	Single	0.5
Single	Divorced	0.4
Divorced	Single	0.4
Married	Divorced	0.75
Divorced	Married	0.75

- The closeness of the Home Credit Value. The following Table 5 is an example of the closeness of the home credit value.

Table 5. The tabel of the closeness of the home credit value

Home credit	Home credit	Closeness
Yes	Yes	1
No	No	1
Yes	No	0.7
No	Yes	0.7

- The Closeness of Bank Credit Value. The following Table 6 is an example of the closeness of bank credit value.

Table 6. The tabel of the closeness of bank credit value

Bank Credit	Bank Credit	Closeness
Yes	Yes	1
No	No	1
Yes	No	0.7
No	Yes	0.7

- The Closeness of Occupation Value. The following Table 7 is an example of the closeness of occupation value.

Table 7. The tabel of the closeness of occupation value

Occupation	Occupation	Closeness
Private Employees	Private Employees	1
Enterpreneur	Enterpreneur	1
Civil Servant	Civil Servant	1
Private Employees	Enterpreneur	0.5
Enterpreneur	Private Employees	0.5
Private Employees	Enterpreneur	0.4
Enterpreneur	Private Employees	0.4
Civil Servant	Enterpreneur	0.75
Enterpreneur	Civil Servant	0.75

d. Examples of problem solving are new customers with the following attribute values:
   - Education: Diploma
   - Status: Single
   - Home Loans: No
   - Credit Debt: No
   - Occupation: Entrepreneur

To predict whether the customer will agree or not, the following steps are taken.

e. Next to determine the weight of each attribute.

3.3. Analysis by using formula of a euclidean distance

- Calculate the closeness between the cases of new customer and case A. The following Table 8 is an example of the closeness of nes case and case A.

Table 8. Examples of the table of the closeness of nes case and case A

Attribute	New Case	Old Case	The Closeness Value	Weight of Atribute
Education	Diploma	>=Bachelor	0.75	0.5
Status	Single	Single	1	0.5
Home Credit	No	No	1	1
Bank Credit	No	No	1	1
Occupation	Enterpreneur	Enterpreneur	1	0.75

The closeness of new case and Case A is calculated by applying (4):

$$\frac{0.75\,x\,0.5 + 0.7\,x\,0.5 + 1\,x\,1 + 1\,x\,1 + 1\,x\,0.75}{0.5 + 0.5 + 1 + 1 + 0.75} = \frac{2,653125}{3,75} = 0,7075$$

- Calculate the closeness between the case of new customer and case B. The following Table 9 is an example of new case proximity table with B.

Table 9. Example of new case proximity table with B

Atribute	New Case	Old Case	The Closeness Value	Weight of Atribute
Education	Diploma	<=High School	0.5	0.5
Status	Single	Married	0.5	0.5
Home Credit	No	No	1	1
Bank Credit	No	No	1	1
Occupation	Enterpreneur	Enterpreneur	1	0.75

The closeness of the new case with case B was calculated by applying (4):

$$\frac{0.5\,x\,0.5 + 1\,x\,0.5 + 0.4\,x\,1 + 1\,x\,1 + 1\,x\,0.75}{0.5 + 0.5 + 1 + 1 + 0.75} = \frac{2,26875}{3,75} = 0,605$$

- Calculate the closeness between new customer cases and case C. The following Table 10 is an example of new case proximity table with C.

Table 10. Example of new case proximity table with C

Attribute	New Case	Old Case	Value Closeness	Weight Atribute
Education	Diploma	>=Bachelor	0.75	0.5
Status	Single	Single	0.7	0.5
Home Credit	No	No	1	1
Bank Credit	No	Yes	0.4	1
Occupation	Enterpreneur	Private Employees	0.4	0.75

The closeness of new case and case C was calculated by applying the formula from (4) as follow:

$$\frac{0.75\,x\,0.5 + 0.7\,x\,0.5 + 1\,x\,1 + 0.4\,x\,1 + 0.4\,x\,0.75}{0.5 + 0.5 + 1 + 1 + 0.75} = \frac{1,753125}{3,75} = 0,4675$$

- Calculate the closeness between new customer cases and case D. The following Table 11 is an example of the closeness table of new case with case D.

Table 11. Example of the closeness table of new case with case D

Attribute	New Case	Old Case	Value of Closeness	Weight of Attribute
Education	Diploma	Diploma	1	0.5
Status	Single	Married	1	0.5
Home Credit	No	No	0.5	1
Bank Credit	No	Yes	0.4	1
Occupation	Enterpreneur	Civil Servant	0.6	0.75

The closeness of new case and case D was calculated by applying the formula from (4) as follows:

$$\frac{1\,x\,0.5 + 1\,x\,0.5 + 0.5\,x\,1 + 0.4\,x\,1 + 0.6\,x\,0.75}{0.5 + 0.5 + 1 + 1 + 0.75} = \frac{1,6875}{3,75} = 0,45$$

From the calculation of the closeness between new cases with cases A, B, C and D, it can be found that the greatest closeness value is obtained in case A, then the prediction in case A will be used, i.e. new customers to agree or disagree in the bank offer.

3.4. Analysis by appling hamming distance formula

- Calculating the closeness of new customer cases and case A. The following Table 12 is an example of new case with case B.

Table 12. Example of new case with case B

Attribute	New Case	Old Case	Value of Closeness	Weight of Attribute
Education	Diploma	>=Bachelor	0.75	0.5
Status	Single	Single	1	0.5
Home Credit	No	No	1	1
Bank Credit	No	No	1	1
Occupation	Enterpreneur	Enterpreneur	1	0.75

The closeness of new case with case A was calculared by applying formula from (4) as follows:

$$\frac{0.75x0.5 - 0.7x0.5 - 1x1 - 1x1 - 1x0.75}{0.75x0.5 + 0.7x0.5 + 1x1 + 1x1 + 1x0.75} = \frac{2.75}{3.5} = 0.785$$

- Calculating the closeness of new customer cases and case B. The following Table 13 is an example of new case with case B.

Table 13. Example of new case with case B

Attribute	New Case	Old Case	Value of Closeness	Weight of Attribute
Education	Diploma	<=Bachelor	0.5	0.5
Status	Single	Married	0.5	0.5
Home Credit	No	No	1	1
Bank Credit	No	No	1	1
Occupation	Enterpreneur	Enterpreneur	1	0.75

The closeness of new case with case B was calculated by applying the formula from (4) as follows:

$$\frac{0.5x0.5 - 1x0.5 - 0.4x1 - 1x1 - 1x0.75}{0.5x0.5 + 1x0.5 + 0.4x1 + 1x1 + 1x0.75} = \frac{2.4}{2.9} = 0.827$$

- Calculating the closeness of new customer cases and case C. The following Table 14 is an example of new case with case C.

Table 14. Example of new case with case C

Attribute	New Case	Old Case	Value of Closeness	Weight of Attribute
Education	Diploma	>=Bachelor	0.75	0.5
Status	Single	Single	0.7	0.5
Home Credit	No	No	1	1
Bank Credit	No	Yes	0.4	1
Occupation	Enterpreneur	Private Employees	0.4	0.75

The closeness of new case with case C was calculated by applying the formula from (4) as follows:

$$\frac{0.75x0.5 - 0.7x0.5 - 1x1 - 0.4x1 - 0.4x0.75}{0.75x0.5 + 0.7x0.5 + 1x1 + 0.4x1 + 0.4x0.75} = \frac{1.675}{2.425} = 0.690$$

- Calculating the closeness of new customer cases and case D. The following Table 15 is an example of new case with case C.

Table 15. Example of new case with case C

Attribute	New Case	Old Case	Value of Closeness	Weight of Attribute
Education	Diploma	Diploma	1	0.5
Status	Single	Married	1	0.5
Home Credit	No	No	0.5	1
Bank Credit	No	Yes	0.4	1
Occupation	Enterpreneur	Civil Servant	0.6	0.75

The closeness of new case with case D was calculated by applying the formula from (4) as follows:

$$\frac{1x0.5 - 1x0.5 - 0.5x1 - 0.4x1 - 0.6x0.75}{1x0.5 + 1x0.5 + 0.5x1 + 0.4x1 + 0.6x0.75} = \frac{1.35}{2.35} = 0.574$$

From the calculation of the closeness between new cases with cases A, B, C and D, it can be found that the greatest closeness value is obtained in case A, then the prediction in case A will be applied, i.e. new customers to agree or disagree in the bank offer.

3.5. Analysis by using formula of a euclidean distance

- Calculating the closeness of new customer cases and case A. The following Table 16 is an example of new case with case A.

Table 16. Example of new case with case A

Attribute	New Case	Old Case	Value of Closeness	Weight of Attribute
Education	Diploma	>=Bachelor	0.75	0.5
Status	Single	Single	1	0.5
Home Credit	No	No	1	1
Bank Credit	No	No	1	1
Occupation	Enterpreneur	Enterpreneur	1	0.75

The closeness of new case with case A was calculated by applying the formula from (4) as follows:

$$\frac{0.75x0.5 - 0.7x0.5 - 1x1 - 1x1 - 1x0.75}{0.5 + 0.5 + 1 + 1 + 0.75} = \frac{2.75}{3.75} = 0.733$$

- Calculating the closeness of new customer cases and case B. The following Table 17 is an example of new case with case B.

Table 17. Example of new case with case B

Attribute	New Case	Old Case	Value of Closeness	Weight of Attribute
Education	Diploma	<=High School	0.5	0.5
Status	Single	Married	0.5	0.5
Home Credit	No	No	1	1
Bank Credit	No	No	1	1
Occupation	Enterpreneur	Enterpreneur	1	0.75

The closeness of new case with case B was calculated by applying the formula from (4) as follows:

$$\frac{0.5x0.5 - 1x0.5 - 0.4x1 - 1x1 - 1x0.75}{0.5 + 0.5 + 1 + 1 + 0.75} = \frac{2.4}{3.75} = 0.640$$

- Calculating the closeness of new customer cases and case C. The following Table 18 is an example of new case with case C.

Table 18. Example of new case with case C

Attribute	New Case	Old Case	Value of Closeness	Weight of Attribute
Education	Diploma	>=Bachelor	0.75	0.5
Status	Single	Single	0.7	0.5
Home Credit	No	No	1	1
Bank Credit	No	Yes	0.4	1
Occupation	Enterpreneur	Private Employees	0.4	0.75

The closeness of new case with case C was calculated by applying the formula from (4) as follows:

$$\frac{0.75x0.5 - 0.7x0.5 - 1x1 - 0.4x1 - 0.4x0.75}{0.5 + 0.5 + 1 + 1 + 0.75} = \frac{1.675}{3.75} = 0.446$$

- Calculating the closeness of new customer cases and case C. The following Table 19 is an example of new case with case D.

Table 19. Example of new case with case D

Attribute	New Case	Old Case	Value of Closeness	Weight of Attribute
Education	Diploma	Diploma	1	0.5
Status	Single	Married	1	0.5
Home Credit	No	No	0.5	1
Bank Credit	No	Yes	0.4	1
Occupation	Enterpreneur	Civil Servant	0.6	0.75

The closeness of new case with case D was calculated by applying the formula from (4) as follows:

$$\frac{1x0.5 - 1x0.5 - 0.5x1 - 0.4x1 - 0.6x0.75}{0.5 + 0.5 + 1 + 1 + 0.75} = \frac{1.35}{3.75} = 0.360$$

From the calculation of the closeness between new cases with cases A, B, C and D, it can be found that the greatest closeness value is obtained in case A, then the prediction in case A will be used, namely new customers to agree or disagree in the bank offer.

3.6. The results of optimizing the KNN method with euclidean distance formulas

From the 18,000 number of old case data, the closeness value with the new data will be calculated by applying the euclidean distance formula. The display of the system in Figure 3. From Figure 3 it is clear that all the data used are close to the value of using the euclidean distance formula with the amount of training data of 18,000 old cases which will be calculated by applying the KNN method. From the new cases that are inputted, the value of proximity is similar or the value of proximity is 1 with 8 old cases with attributes agree no. It means that the new case data applying the KNN method calculated the value of its proximity using the euclidean distance formula as a result with 8 old case data located in the listview of the proximity of the KNN within 17 minutes 45 seconds.

3.7. The results of optimizing the KNN method with hamming distance formulas

From the 18,000 number of old case data, the closeness value with the new data will be calculated by applying the hamming distance formula. The display of the system in Figure 4. From Figure 4 it is clear that all the data used are searched for the value of proximity by applying the hamming distance formula with the amount of training data of 18,000 old cases which will be calculated by using the KNN method. From the input new cases the value of closeness is similar or the value of closeness is 0.884 with 29 old cases with attributes agree no. It means the new case data applying the KNN method, the value of proximity is calculated by applying the hamming distance formula, then the results are not worth 1 (similar) but approaching the value of 1 with the value of 0.884 with 29 old case data located in the listview of the KNN proximity with 13 minutes and 13 seconds.

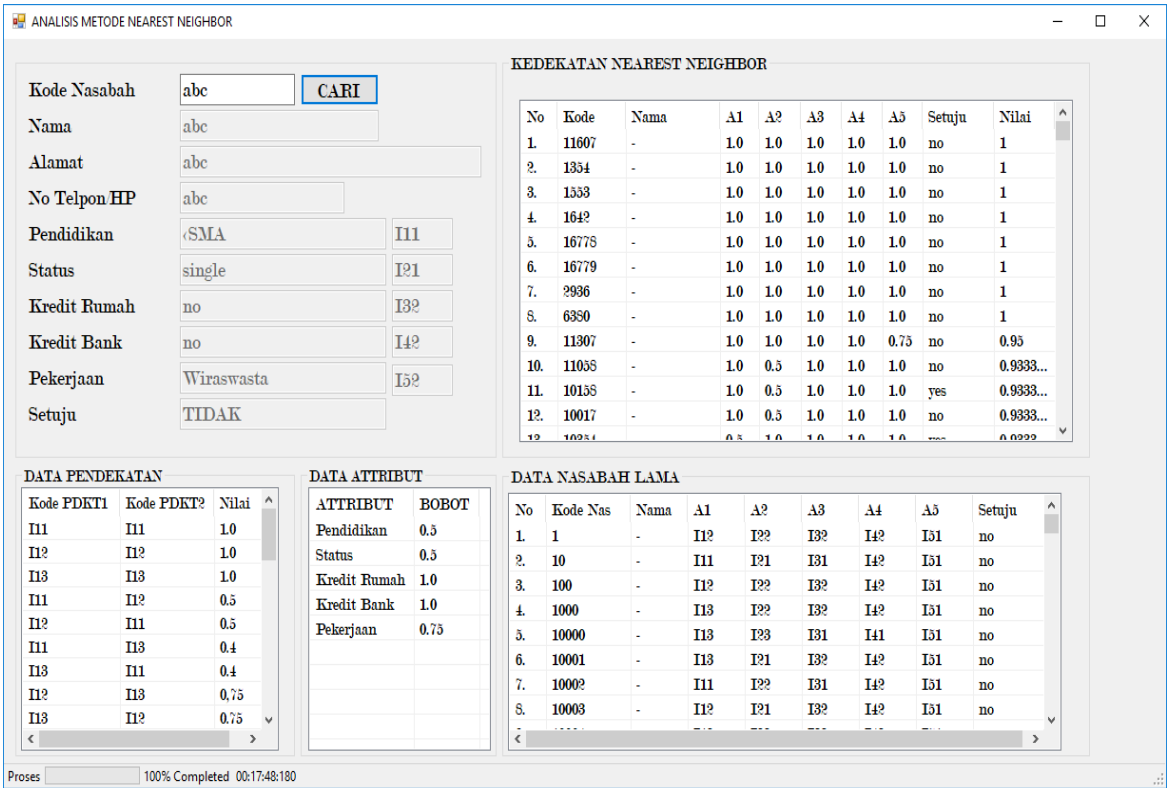

Figure 3. The results of optimizing the KNN method with euclidean distance formulas

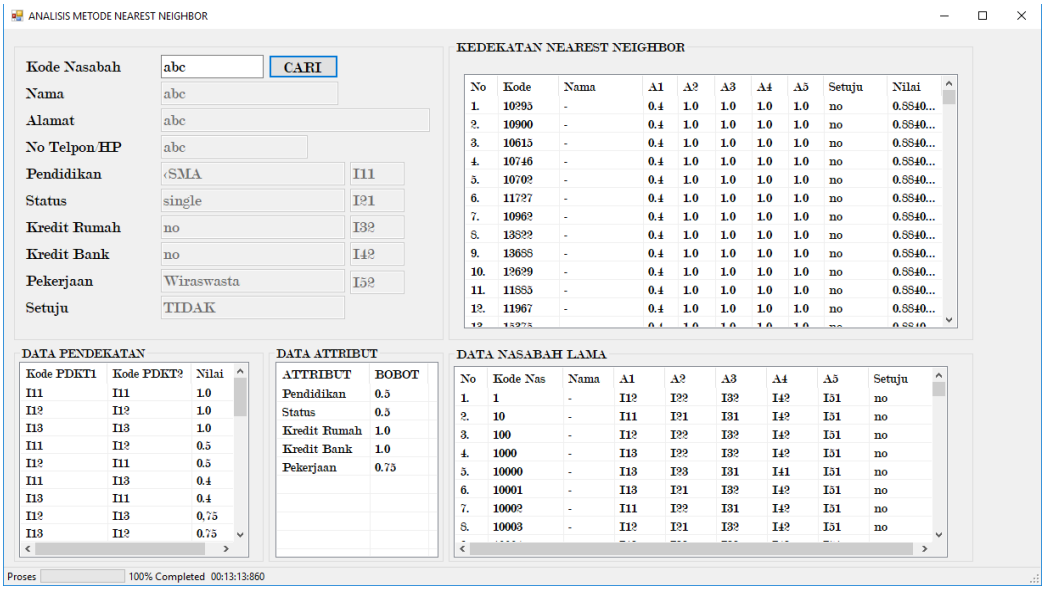

Figure 4. The results of optimizing the KNN method with the hamming distance formula

3.8. The results of optimizing the KNN Method with manhattan distance formulas

From the 18,000 number of old case data, the closeness value of the new data will be calculated by applying the Manhattan distance formula. The display of the system in Figure 5. From Figure 5 it is clear that all the data used, the value of proximity is calculated by applying the Manhattan distance formula with the amount of training data of 18,000 old cases which will be calculated by using the KNN method. From the input new cases, the value of closeness is similar or the value of closeness is 0.8133 with 29 old cases with attributes agree no. It means the new case data using the KNN method, the value of proximity is calculated by applying the Manhattan distance formula results with none having a proximity value of 1 but 0.8133 with 29 old case data located in the listview of the proximity of the KNN within 16 minutes 11 seconds.

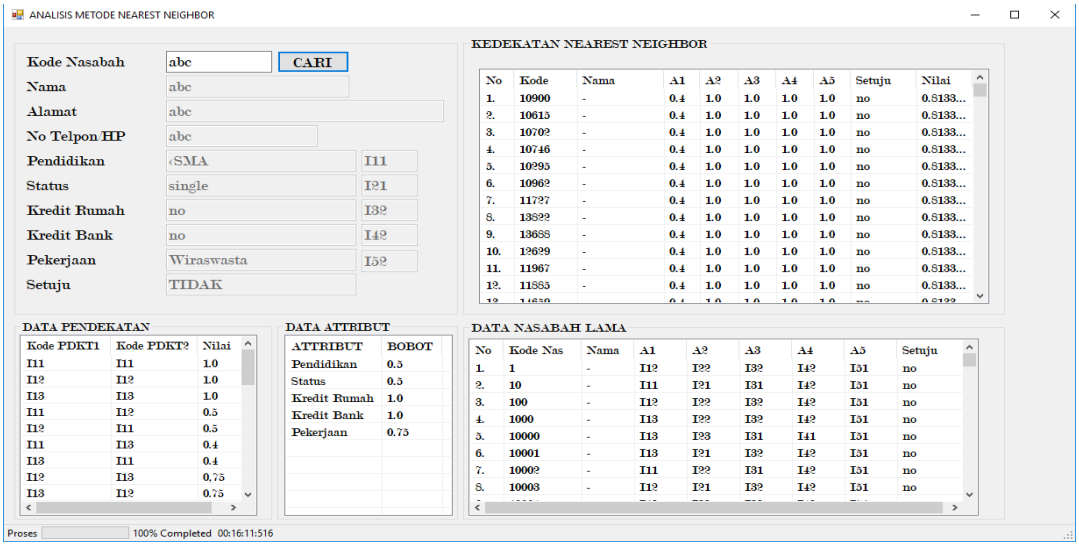

Figure 5. Results of optimizing the KNN method with manhattan distance formula

3.9. Discussion

After testing by 18,000 old case data then new case data is input and the value of the closeness between the new case and the old case is calculated by optimizing the distance formula by using

the euclidean distance formula with the hamming distance formula and the Manhattan distance formula where each of the formulas is obtained the results applying the euclidian distance formula require 17 minutes 45 seconds to calculate the value of proximity with the results of the proximity value of 1 (similar) to 8 old cases. Whereas by applying the hamming distance formula with the same amount of data and the same new case requires 13 minutes and 13 seconds in the system. From the results of the calculation the value of proximityis 0.884 which has a value close to 1 (similar) with the number of old cases as many as 29 cases. Meanwhile, by applying the Manhattan distance formula calculates the value of proximity to the same old case data and the new case takes 16 minutes 11 seconds to calculate it in the system. But from this calculation there is no closeness value 1 (similar) but approaching the value 1, which is 0.8133 which consists of 29 old case data. From the data above, it can be seen the optimal comparison results in the following table 20. Therefore, the optimization should deliver the suitable solution to the problem within the context by considering various factors such as total cost, aggregation value, maximum loading point, consistency of performance, system losses, category accuracy, feature extracted and so on [22-25]. The following Table 20 is result of optimization comparison.

Table 20. The result of optimization comparison

No	Distance Formula	Number of old case	Time	Closeness Value	Number of Data Closeness Value
1	Euclidean distance	18.000	17Min 45 Sec	1	8
2	Hamming distance	18.000	13 Min 13 Sec	0.884	29
3	Manhattan distance	18.000	16 Min 11 Sec	0.8133	29

4. CONCLUSION

By applying the euclidean distance formula the closeness value is 1 (similar) with 8 the number of old case data with the processing time on the system with the number of old case data is 18,000 requires 17 minutes 45 seconds. With the distance hamming formula there is no 1 (similar) value of closeness but 0.884 with 29 the number of old case data with the process time on the system with the 18,000 old case data requires 13 minutes 13 seconds. With the Manhattan distance formula there is no 1 (similar) value of closeness but 0.8113 with 29 the number of old case data with the process time on the system with the 18,000 old case data requiring 16 minutes 11 seconds.

REFERENCES

[1] Varun and R. Nisha, "Knowledge Discovery from Database Using an Integration of Clustering and Classification," *Int. Journal of Advanced Computer*, vol 2, no 3, pp. 29-33, 2011.

[2] N. Krisandi, et al., "The k-nearest neighbor algorithm in the classification of palm oil production data at PT. Minamas, Parindu sub-district (in Indonesia: Algoritma *k-nearest neighbor* dalam klasifikasi data hasil produksi kelapa sawit pada PT. Minamas kecamatan Parindu)," *Bimaster,* vol 02, no 1, 2013.

[3] S. B. Imandoust and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoritical Background," *Int. Journal of Engineering Research and Applications*, vol. 3, no. 5, pp. 605-610, 2013.

[4] W. Sun and B. Trevor, "Combining K-nearest neighbor models for annual peak breakup flow forecasting," *Cold Regions Science and Technology*, vol. 143, pp. 59-69, November 2017,

[5] Kusrini, et al., *"*Comparison of Nearest Neighbor Method and C4.5 Algorithm to Analyze Possible Resignation of Prospective Students at STMIK AMIKOM Yogyakarta (in Indonesia Perbandingan Metode Nearest Neighbor dan Algoritma C4.5 untuk Menganalisis Kemungkinan Pengunduran Diri Calon Mahasiswa di STMIK AMIKOM Yogyakarta)*," Jurnal DASI*. vol. 10, no 1. pp. 114-132. 2009.

[6] X. Pan, et al., "K-nearest neighbor based structural twin support vector machine," *Knowledge-Based Systems*, vol. 88, pp. 34-44, 2015.

[7] M. Lubis, *et al.*, "Current State of Personal Data Protection in Electronic Voting: Criteria and Indicator for Effective Implementation," *TELKOMNIKA (Telecommunication, Computing, Electronics and Control)*, vol. 16, no. 1, pp. 290-301. 2018.

[8] L. Hakim, *et al.*, "Text Mining of UU-ITE Implementation in Indonesia," *Journal of Physics: Conference Series*, vol. 1007, vol. 1. 2018.

[9] M. Lubis, *et al.*, "The Indonesia Public Information Disclosure Act (UU-KIP): Its Challenges and Responses," *Int. J. of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 1, pp. 94-103. 2018.

[10] N. Dzikrulloh, *et al.*, "Application of K-Nearest Neighbor (KNN) Method and Weighted Product (WP) Method in Accepting Prospective Teachers and Employees with Technology-Based New Administration (Case Study: Muhammadiyah Vocational High School 2 Kediri (in Indonesia Penerapan Metode K–Nearest Neighbor (KNN) dan Metode Weighted Product (WP) dalam Penerimaan Calon Guru Dan Karyawan Tata Usaha Baru Berwawasan Teknologi (Studi Kasus: Sekolah Menengah Kejuruan Muhammadiyah 2 Kediri))," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 1, no. 5, pp. 378-385. 2017.

[11] Mustakim & G. F. Oktaviani, "Algoritma *K-Nearest Neighbor Classification* Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa," *Jurnal Sains, Teknologi dan Industri*, vol 13, no 2, pp. 195-202, 2016.
[12] D. Meteos-Garcia, *et al.*, "On the evolutionary weighting of neighbours and features in the k-nearest neighbour rule," *Elsevier: Neurocomputing*, vol. 326–327, pp. 54-60, 31 January 2019.
[13] X. Li, *et al.*, "Discrimination of soft tissues using laser-induced breakdown spectroscopy in combination with k nearest neighbors (kNN) and support vector machine (SVM) classifiers," *Elsevier: Optic and Laser Technology*, vol. 102, pp. 233-239, 2018.
[14] Al-Khowarizmi, *et al.*, "Measuring the Accuracy of Symple Evolving Connectionist System with Varying Distance Formulas," *Journal of Physics: Conference Series*, vol 930, pp. 1-6. 2017.
[15] M. F. A. Fauzi, "Optimal Discrete Wavelet Frames Features for texture-based image retrieval applications," *Proceedings of the First International Visual Informatics Conference, IVIC 2009*, held in Kuala Lumpur, Malaysia, in November 2009, pp. 66-77.
[16] S. Moro, et al., "A Data-Driven Approach to Predict the Success of Bank Telemarketing," *Decision Support Systems*, vol. 62, pp. 22-31, 2014.
[17] T. Cho, R. W. Conners and P. A. Araman, "A comparison of rule-based, k-nearest neighbor, and neural net classifiers for automated industrial inspection," *[1991] Proceedings of the IEEE/ACM International Conference on Developing and Managing Expert System Programs*, Washington, DC, USA, 1991, pp. 202-209.
[18] N. Bhatia and Vandana, "Survey of Nearest Neighbor Techniques", *Int. Journal of Computer Science and Information Security*, vol. 8, no. 2. pp, 302-305, 2010.
[19] H. Parvin, *et al.*, "MKNN: Modified K-Nearest Neigbor," *Proceedings of the World Congress on Engineering and Computer Science (WCECS)*, 2008.
[20] B. V. Darasay, "Nearest Neighbor Pattern Classification Tecniques," Las Alamitos, LA: IEEE Computer Society Press. 1991.
[21] A. Tomar and A. Nagpal, "Comparing Accuracy of K-Nearest-Neigbor and Support-Vector-Machines for Age Estimation," *Int. J. of Eng. Trends and Tech. (IJETT)*, vol. 38, no. 6, pp. 326-329, 2016.
[22] E. E. Hassan, *et al.*, "Maximum Loadability Enhancement with a Hybrid Optimization Method," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 7, no. 3, pp. 323-330, 2018.
[23] N. M. N. Mathivanan, *et al.*, "Improving Classification Accuracy using Clustering Technique," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 7, no. 3, pp. 465-470, 2018.
[24] R. Fauzi, *et al.*, "Defense Behavior of Real Time Strategy Games: Comparison between HFSM and FSM," *Indonesia Journal of Electrical Engineering and Computer Science*, vol. 13, no. 2, pp. 634-642, 2019.
[25] Julham, *et al.*, "Development of Soil Moisture Measurement with Wireless Sensor Web-Based Concept," *Indonesia Journal of Electrical Engineering and Computer Science*, vol. 13, no. 2, pp. 514-520, 2019.