Journal of Engineering and Applied Sciences 14 (1): 247-252, 2019
ISSN: 1816-949X

Data Processing of Laboratory Recruitment Using K-Nearest Neighbor Algorithm

Mustabshiroh, Roswan Latuconsina and Tito Waluyo Purboyo
Department of Computer Engineering, Faculty of Electrical Engineering, Telkom University, Bandung, Indonesia

Abstract: The development of science and technology increasingly popular in the society. Computer technology is growing very rapidly according to the level of human needs to solve various of problems. One of the rapid development of computer science is the field of data processing that is capable of processing data and provide information to support decision, one of which is data mining. Data mining helps to extract patterns from large data to be interpreted as human-readable information. A simple example of the use of data mining is the recommendation system used in the recruitment of laboratory assistants by using the k-nearest neighbor method. The training data is in the form of the grade from previous recruitment which will be used as a reference in determining whether the candidate is declared accepted or not.

Key words: Data mining, k-nearest neighbor, training data, reference, candidate, method

INTRODUCTION

Along with the development of the era, science and technology are more commonly used by the community. The technology that is very likely to be used is computer technology. Computer technology is growing very rapidly according to the level of human needs to solve various problems. One of computer science's fastest development is the field of data processing capable of processing data and provide information to support the decision. One of them is data mining.

Data mining is the mining of data to search a particular pattern of a very large amounts of data to make it easier to understand. Data mining helps to interpret information from large data that is easily understood by humans. The volume of data increases every year. In 2011, the data volume reached 1.8 trillion gigabytes in 2012 up to 2.8 ZB (Gantz and Reinsel, 2012). Therefore, data mining is needed in data reading or digging, so that, the data is easily understood by humans. Many scientists already use data mining in their research, data mining is used from simple to complex problems. However, in Indonesia the use of data mining is still rare. Therefore, this study will discuss the use of data mining in a simple way which is data processing for laboratory assistant recruitment.

Data mining has many supporting methods, one of which is the k-nearest neighbor algorithm. k-nearest neighbor is one of the simple algorithms in the classification method in data mining. Therefore, the data processing of laboratory assistant recruitment in this study will use the k-nearest neighbor algorithm.

MATERIALS AND METHODS

Data mining: A process that combines statistical techniques, artificial intelligence, mathematics and machine learning to extract information stored in large databases for easy understanding to humans (Turban and Aronson, 2001). Data mining can analyze large data by sorting them into more concise form, so, it can be processed further as needed.

Data mining can process data with several variables by determining the pattern of the data to predict other variables that have unknown value. Data mining can find important characteristics in a data, so, the data is easily recognizable or identifiable. Data mining helps to describe important data that is hard to describe, especially in large data such as big data. Data mining can find a relationship between several attributes in the data.

K-Nearest Neighbor (KNN): KNN is one of the algorithms in classification techniques which is studying a set of data to generate rules that can recognize new data that has never been studied (Arboleda *et al*., 2018). k-nearest neighbor classifies the data closest to the test data by determining the number of k objects closest to the test

Corresponding Author: Mustabshiroh, Department of Computer Engineering, Faculty of Electrical Engineering, Telkom University, Bandung, Indonesia

Table 1: Training data to be used

No.	Name	ID	Paper and pencil test	Coding test	Apprentice test	Teaching test	Coding test (12 h)	Results
1	N.A. Elvia	4086	71	60	72.8	52	57.5	Yes
2	M. Rizqi M	1076	57.2	100	86.2	48	76.3	Yes
3	N.P. Gietha	0031	59	20	62.1	74	56.3	Yes
4	D.I. Tirtha	3075	50	30	76.3	50	43.8	Yes
5	B.J. Aqil	0165	64	30	41.3	38	71.9	Yes
6	M. Khalifah	4207	66	20	75.7	42	62.1	Yes
7	B.P.A. Ricky	0138	55	30	67.1	66	60.5	No
8	H. Julian	0093	56	40	72.2	66	50	Yes
9	R. Assyifa	4051	42	20	69.5	39	40	Yes
10	R. Rery	4192	73	40	80.1	56	82.1	Yes
11	M. Arief	4188	81	65	83.9	48	52.5	Yes
12	M. Rico A	0018	74	60	82.4	60	41.3	Yes
13	N. Herman	4005	29.2	30	4.40	62	60.5	No
14	P. Arif	4164	59	30	57.6	40	50	No
15	M. Wahyu PI	2097	64	30	66.1	74	60.5	Yes

Table 2: Testing data for manual calculation

Testing data	
Name	Nuraini SW
ID	4190
Paper and pencil test	60
Coding test	40
Apprentice test	70
Teaching test	80
Coding test (12 h)	45
Result	Yes

data. Calculation of distance between train data and test data using Euclidean formula (Aprilia *et al*., 2018):

$$d(a, b) = \sqrt{\sum_{i=0}^{n} (X_i - Y_i)^2} \quad (1)$$

d(a, b) = Euclidean distance
X = Data 1
i = Data to-i
Y = Data 2
n = Amount of data

Euclidean distance is a more modern version of the Naive Bayes method by making it simpler but having the same probability and matrix (Hammam *et al*., 2017). One of the largest international data mining locations, IEEE ICDM which ended on December 22, 2006 has announced 10 best algorithms used in data mining and the k-nearest neighbor algorithm is included (Wu and Kumar, 2009).

K-nearest neighbor has several advantages including its simplicity and very suitable for data that has many classes. In addition, the classification method is more suitable to use in analysis that uses a of data compared to clustering methods (Kumar and Rathee, 2011).

The general stage in k-nearest neighbor is to determine the parameter of k first which is the number of the nearest neighbors as a reference. Then, the calculation is done to determine the distance between training data and testing data using Euclidean formula. Once the distance on each data is found, ascending ordering is done to determine the minimum distance according to the number of parameters k. To determine the results, use the data that is the majority (Cai *et al*., 2017; Arboleda *et al*., 2018).

Microsoft Excel: An application that is part of the Microsoft Office installation package that works to process the numbers. Microsoft Excel consists of columns and rows used as a place to enter numbers and calculation formulas. Microsoft Excel features easy-to-learn graphical calculations and graphs. Therefore, in this study we will use Microsoft Excel for calculation. Microsoft Excel has several functional formulas that can be used to perform k-nearest neighbor calculations.

RESULTS AND DISCUSSION

In this study, the laboratory assistant data processing will be carried out using the k-nearest neighbor algorithm in two ways, namely manual calculation and using Microsoft Excel. Data processing is done using two different methods to ensure the accuracy of the two methods. Due to the manual calculation, data that is not too large is used. The training data will be used as a reference in the form of data of previous registrants who have passed and have not passed, the test data is data of registrants who are following the laboratory assistant's receipt. The data used is in the form of value at each stage of the receipt. In this study, using the training data of 15 applicants in the previous year where 12 applicants were passed and using the example of one new registrant that will be processed to find out the final results. In Table 1, it can be seen that there are several

data that will be used as a reference consisting of the results of the assessment in several stages of recruitment such as paper and pencil tests, coding tests, apprentice tests, teaching tests, coding test in 12 h and the results of each participants who were presented in the form of yes for the participants stated to pass the selection and the number for participants who were declared as not passing the selection.

Manual calculation: The calculation process using Euclidean formula is done manually by doing calculation of paper and using calculator tool by applying k-nearest neighbor algorithm.

In Table 2 one of the participant data that will be used as testing data. The data which consists of the results of the evaluation at each stage of the recruitment will be calculated the proximity distance with the training data. Calculations using the Euclidean formula as in the calculation as:

Looking for euclidean:

$$\begin{aligned} d &= \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2} \\ &= \sqrt{(71-60)^2+(60-40)^2+(72.8-70)^2+(52-80)^2+(57.5-45)^2} \\ &= \sqrt{(11)^2+(20)^2+(2.8)^2+(-30)^2+(12.5)^2} \\ &= \sqrt{121+400+7.84+900+156.25} \\ &= \sqrt{1585.09} \\ &= 39.8 \quad \text{(Yes)} \end{aligned} \tag{2}$$

$$\begin{aligned} d &= \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2} \\ &= \sqrt{(57.2-60)^2+(100-40)^2+(86.2-70)^2+(48-80)^2+(76.3-45)^2} \\ &= \sqrt{(-2.8)^2+(60)^2+(16.2)^2+(-32)^2+(31.3)^2} \\ &= \sqrt{7.84+3600+262.44+1024+979.69} \\ &= \sqrt{5873.97} \\ &= 76.64 \quad \text{(Yes)} \end{aligned} \tag{3}$$

$$\begin{aligned} d &= \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2} \\ &= \sqrt{(59-60)^2+(20-40)^2+(62.1-70)^2+(74-80)^2+(56.3-45)^2} \\ &= \sqrt{(-1)^2+(-20)^2+(-7.9)^2+(-6)^2+(11.3)^2} \\ &= \sqrt{1+400+62.41+36+127.69} \\ &= \sqrt{627.1} \\ &= 25.04 \quad \text{(Yes)} \end{aligned} \tag{4}$$

$$\begin{aligned} d &= \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2} \\ &= \sqrt{(50-60)^2+(30-40)^2+(76.3-70)^2+(50-80)^2+(43.8-45)^2} \\ &= \sqrt{(11)^2+(10)^2+(6.3)^2+(-30)^2+(1.2)^2} \\ &= \sqrt{100+400+39.69+900+1.44} \\ &= \sqrt{1141.13} \\ &= 39.8 \quad \text{(Yes)} \end{aligned} \tag{5}$$

$$\begin{aligned} d &= \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2} \\ &= \sqrt{(64-60)^2+(30-40)^2+(41.3-70)^2+(38-80)^2+(71.9-45)^2} \\ &= \sqrt{(-4)^2+(-10)^2+(-28.7)^2+(-42)^2+(26.9)^2} \\ &= \sqrt{16+100+823.69+1764+715.54} \\ &= \sqrt{3419.23} \\ &= 58.47 \quad \text{(Yes)} \end{aligned} \tag{6}$$

$$\begin{aligned} d &= \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2} \\ &= \sqrt{(66-60)^2+(20-40)^2+(75.7-70)^2+(42-80)^2+(62.1-45)^2} \\ &= \sqrt{(-6)^2+(-20)^2+(-5.7)^2+(-38)^2+(17.1)^2} \\ &= \sqrt{36+400+32.49+1444+292.41} \\ &= \sqrt{2204.9} \\ &= 46.956 \quad \text{(Yes)} \end{aligned} \tag{7}$$

$$\begin{aligned} d &= \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2} \\ &= \sqrt{(55-60)^2+(30-40)^2+(67.1-70)^2+(66-80)^2+(60.5-45)^2} \\ &= \sqrt{(-5)^2+(-10)^2+(-2.9)^2+(-14)^2+(15.5)^2} \\ &= \sqrt{25+100+8.41+196+240.25} \\ &= \sqrt{269.66} \\ &= 23.866 \quad \text{(No)} \end{aligned} \tag{8}$$

$$\begin{aligned} d &= \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2} \\ &= \sqrt{(56-60)^2+(40-40)^2+(72.7-70)^2+(66-80)^2+(50-45)^2} \\ &= \sqrt{(-4)^2+(0)^2+(2.7)^2+(14)^2+(5)^2} \\ &= \sqrt{16+0+7.29+196+25} \\ &= \sqrt{244.29} \\ &= 15.63 \quad \text{(Yes)} \end{aligned} \tag{9}$$

$$d = \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2}$$
$$= \sqrt{(42-60)^2+(20-40)^2+(69.5-70)^2+(39-80)^2+(40-45)^2}$$
$$= \sqrt{(-18)^2+(-20)^2+(-0.5)^2+(-41)^2+(-5)^2}$$
$$= \sqrt{324+400+0.25+1681+25}$$
$$= \sqrt{2430.25}$$
$$= 49.3 \quad \text{(Yes)}$$

(10)

$$d = \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2}$$
$$= \sqrt{(73-60)^2+(40-40)^2+(80.1-70)^2+(56-80)^2+(82.1-45)^2}$$
$$= \sqrt{(-13)^2+(0)^2+(10.1)^2+(-24)^2+(37.1)^2}$$
$$= \sqrt{169+0+102.01+576+1376.41}$$
$$= \sqrt{2223.42}$$
$$= 47.15 \quad \text{(Yes)}$$

(11)

$$d = \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2}$$
$$= \sqrt{(81-60)^2+(65-40)^2+(83.9-70)^2+(48-80)^2+(52.5-45)^2}$$
$$= \sqrt{(-21)^2+(-25)^2+(-13.9)^2+(-32)^2+(-7.5)^2}$$
$$= \sqrt{441+625+193.21+1024+56.25}$$
$$= \sqrt{2339.46}$$
$$= 48.37 \quad \text{(Yes)}$$

(12)

$$d = \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2}$$
$$= \sqrt{(74-60)^2+(60-40)^2+(82.4-70)^2+(60-80)^2+(41.3-45)^2}$$
$$= \sqrt{(-14)^2+(20)^2+(-12.4)^2+(-20)^2+(-3.7)^2}$$
$$= \sqrt{196+400+153.76+400+13.69}$$
$$= \sqrt{1163.45}$$
$$= 34.1 \quad \text{(Yes)}$$

(13)

$$d = \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2}$$
$$= \sqrt{(29.2-60)^2+(30-40)^2+(4.4-70)^2+(62-80)^2+(60.5-45)^2}$$
$$= \sqrt{(-30.8)^2+(-10)^2+(-65.6)^2+(-18)^2+(-15.5)^2}$$
$$= \sqrt{948.64+100+4303.36+324+240.25}$$
$$= \sqrt{5916.25}$$
$$= 76.9 \quad \text{(No)}$$

(14)

$$d = \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2}$$
$$= \sqrt{(59-60)^2+(30-40)^2+(57.6-70)^2+(40-80)^2+(50-45)^2}$$
$$= \sqrt{(-1)^2+(10)^2+(-12.4)^2+(-40)^2+(-5)^2}$$
$$= \sqrt{1+100+153.76+1600+25}$$
$$= \sqrt{1879.76}$$
$$= 43.36 \quad \text{(No)}$$

(15)

$$d = \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2}$$
$$= \sqrt{(64-60)^2+(30-40)^2+(66.1-70)^2+(74-80)^2+(60.5-45)^2}$$
$$= \sqrt{(-4)^2+(-10)^2+(-3.9)^2+(-6)^2+(-15.5)^2}$$
$$= \sqrt{16+100+15.21+36+240.25}$$
$$= \sqrt{4.7.46}$$
$$= 20.1856 \quad \text{(Yes)}$$

(16)

In Table 1, the result of distance calculation using Euclidean formula in each training data with test data. The calculation results are ranked noble ascending from the smallest to the biggest results. The use of ratings makes it easy to determine the dependency of the test data.

In Table 2, the label of the rank which has been ascending. The k value is 7 which means it will take 7 pieces with the smallest value. Here, is the role of ranking in facilitating the search for the smallest value. From the results obtained the results in the form of one to No and six for Yes. So that, it can be concluded that the results of the test data are Yes.

Use of Microsoft Excel: Microsoft Excel was chosen because of the use of simple and easy-to-understand formula functions. Microsoft Excel has several formulas that can be used in calculating Euclidean formulas and some functional formulas that can be applied to the k-nearest neighbor algorithm. As for some functions that will be used in this paper that is:

SQRT: To find the square root of a number.

$$= SQRT((E4-SMS5)^2+(F4-SMS6)^2+$$
$$(G4-SMS7)^2+(H4-SMS8)^2+(14-SMS9)^2)$$

The use of the SQRT function Eq. 1 to calculate the distance between the train data and the test data.

RANK: To rank or order lists ascending or descending. k-nearest neighbor performs sorting data ascending.

Table 3: Results based on Euclidean distance (ascending)

No.	-------Euclidean distance-------		Rank
1	39.800	YES	7
2	76.640	Yes	14
3	25.040	Yes	4
4	33.780	Yes	5
5	58.470	Yes	13
6	46.956	Yes	9
7	23.866	No	3
8	15.630	Yes	1
9	49.300	Yes	12
10	47.150	Yes	10
11	48.370	Yes	11
12	34.100	Yes	6
13	76.900	No	15
14	43.360	No	8
15	20.1856	Yes	2

Table 4: Prediction of results with K = 7

K (rank)	Label
1	Yes
2	Yes
3	No
4	Yes
5	Yes
6	Yes
7	Yes

Table 5: a) An example of testing data used and b) Prediction of results based on the top 5 ranking that has the shortest distance

Tasting data

Name	Nuraini SW	K	Label
ID	4190	1	Yes
Paper and pencil test	60	2	Yes
Coding test	40	3	No
Apprentice test	70	4	Yes
Teaching test	80	5	Yes
Coding test (12 h)	45	-	-
Result	Yes	-	-

= RANK(P5, SPS2 : SPS16.1)

The use of the RANK function Eq. 2 to sort the data ascending.

VLOOKUP: To search for labels from Table 3. In this study, VLOOKUP is used to determine the label of the data that has the least value of the number of parameters k where the parameter k will be determined. The most dominant data is the predicted result.

= VLOOKUP(S4, SOS2 : SOS16, 3, FALSE)

The use the VLOOKUP function Eq. 3. In Table 3-5, the formula writing is done. The layout of tables in formulas can change depending on where the data table is used. In Table 4 use the No. 1 at the end of the formula as a sign that the data will be ranked ascending.

In Table 6 is one of the participant data that will be used as data testing as in manual calculations. Unlike the manual calculation, the proximity will be calculated with

Table 6: The result of calculating the distance of training data with test data using SQRT

Rank	Euclideans distance	Label
7	38.32870987	Yes
14	76.64182931	Yes
4	25.04196478	Yes
5	33.78015740	Yes
13	58.54314648	Yes
9	46.95636272	Yes
3	23.86755119	No
1	15.62977927	Yes
12	49.29756586	Yes
10	47.15315472	Yes
11	48.36796460	Yes
6	34.10938287	Yes
15	76.91716323	No
8	43.35619910	No
2	20.18563846	Yes

the training data using calculations performed by Microsoft Excel. In Table 6 is a label of the rank that has been ascending in order with a value of k = 5, so that, the top five is ranked with the smallest neighbor. The results of calculating the value of the constancy using the Euclidean formula can be seen in Table 7.

From the results of the ranking with the value of K = 5 as shown in Table 6, it can be concluded that the testing data was passed with the details of one label for No and 4 labels for Yes.

In chapter 1 and Table 7, it can be seen that the results of historical calculations with Euclidean are not very different. The difference is only seen in some calculation results that are influenced by the difference in nominal rounding.

In this study, the use of a different k value that is k with a value of 7 in the manual calculation and k with a value of 5 in the calculation using. Microsoft Excel has no effect on the final result in the test data on behalf of Nuraini SW is estimated to be accepted in the recruitment of the acceptance of laboratory assistants. In the final results shown in Table 2 and 6 said Yes labels dominate.

CONCLUSION

Manually calculating and using Microsoft Excel each of which has a different k value does not significantly affect the predicted results. Both ways provide the same results, so, for calculations with training data that is not too much Microsoft Excel can facilitate data processing rather than manual calculations.

From these calculations can be concluded that the prospective laboratory assistant on behalf of Nuraini SW is expected to pass because the number of data passed more than not pass. For small data Microsoft Excel can still be used. However, for large data it is suggested to use the database and to apply k-nearest neighbor algorithm to the application or decision-making system.

REFERENCES

Aprilia, Y.D., R. Latuconsina and T.W. Purboyo, 2018. A review of several algorithms for data mining. J. Eng. Appl. Sci., 13: 6157-6161.

Arboleda, E.R., A.C. Fajardo and R.P. Medina, 2018. Classification of coffee bean species using image processing, artificial neural network and K nearest neighbors. Proceedings of the 2018 IEEE International Conference on Innovative Research and Development (ICIRD), May 11-12, 2018, IEEE, Bangkok, Thailand, ISBN:978-1-5386-6295-3, pp: 1-5.

Cai, Y., H. Huang, H. Cai and Y. Qi, 2017. A K-nearest neighbor locally search regression algorithm for short-term traffic flow forecasting. Proceedings of the 2017 9th International Conference on Modelling, Identification and Control (ICMIC), July 10-12, 2017, IEEE, Kunming, China, ISBN:978-1-5090-6576-9, pp: 624-629.

Gantz, J. and D. Reinsel, 2012. The digital universe in 2020: Big data, bigger digital shadows and biggest growth in the far East. International Data Corporation Framingham, Massachusetts, USA. https:// trends. ifla.org/node/88

Hammam, S.A., T.W. Purboyo and R.E. Saputr, 2017. Texture analysis using gray level run length and euclidean distance. J. Theor. Appl. Inf. Technol., 95: 6915-6923.

Kumar, V. and N. Rathee, 2011. Knowledge discovery from database using an integration of clustering and classification. Intl. J. Adv. Comput. Sci. Appl., 2: 29-33.

Turban, E. and J.E. Aronson, 2001. Decision Support Systems and Intelligent Systems. 6th Edn., Prentice Hall, Upper Saddle River, New Jersey.

Wu, X. and V. Kumar, 2009. The Top Ten Algorithms in Data Mining. CRC Press, Boca Raton, Florida, USA., ISBN-13:978-1-4200-8964-6, Pages: 214.