Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 5, No. 1, February 2020, Pp. 55-60

Comparative study of decision tree, k-nearest neighbor, and modified k-nearest neighbor on jatropha curcas plant disease identification

Triando Hamonangan Saragih*[1], Diny Melsye Nurul Fajri[2], Alfita Rakhmandasari[3]
Computer Science, Faculty of Mathematics and Nature Science, Lambung Mangkurat University, Indonesia[1]
Magister of Computer Science, Faculty of Computer Science, Brawijaya University, Indonesia[2]
Magister of Computer Science, Faculty of Computer Science, Brawijaya University, Indonesia[3]

Article Info
Keywords:
Decision Tree, Identification, Jatropha Curcas, K-Nearest Neighbor, Modified K-Nearest Neighbor

Article history:
Received 17 December 2019
Revised 18 December 2019
Accepted 17 January 2020
Published 06 February 2020

Cite:
Saragih, T., Fajri, D., & Rakhmandasari, A. (2020). Comparative Study of Decision Tree, K-Nearest Neighbor, and Modified K-Nearest Neighbor on Jatropha Curcas Plant Disease Identification. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 5(1). doi:https://doi.org/10.22219/kinetik.v5i1.1012

*Corresponding author.
Triando Hamonangan Saragih
E-mail address:
triando.saragih@ulm.ac.id

Abstract
Jatropha Curcas is a very useful plant that can be used as a bio fuel for diesel engines replacing the coal. In Indonesia, there are few plantation that plant Jatropha Curcas. But there is so limited farmers that understand in detail about the disease of Jatropha Curcas and it may cause a big loss during harvesting when the disease occured with no further action. An expert system can help the farmers to identify the lant diseases of Jatropha Curcas. The objective of this research is to compare several identification and classification methods, such as Decision Tree, K-Nearest Neighbor and its modification. The comparison is based on the accuracy. Modified K-Nearest Neighbor method given the best accuracy result that is 67.74%.

1. Introduction
Jatropha Curcas is a plant categorized as shrub plant, that can live in dry area, which has low rainfall. Southeast Asia, Central and South India, and also Southern Africa are the areas that suitable for Jantropha Curcas, so we can find a lot of them there [1]. For industrial usage, Jantropha Curcas can be used as diesel engine bio fuel [2]. Like the other plant, there are a few plat disease that can attack this plant such as Bacterial Wilt, Anthracnose, Fusarium Wilt, Leaf Blight, Charcoal Rot, and Powdery Mildew. Unfortunately, the farmers have just a few knowledge about this plant's disease so it can be harmful when the diseases occured. The plant diseases need to be identified fast so the farmer know futher action in order to avoid big loss in harvesting season. An expert system that implements expert knowledge can be a solution for robust plant diseases identification.

There are few methods that can be used for identification, such as Dempster-Shafer [3], Fuzzy Neural Network [4], Optimized Fuzzy Neural Network [5], Extreme Learning Machine [6] and Naive Bayes [7] which work based on certainty and probabilistic method. Dhempster-Shafer method use belief value for making decision. The preliminary search of the author show 86.77% of accuracy. For obtaining the higher accuracy, authors used several classification methods on data mining, such as Decision Tree, K-Nearest Neighbor (KNN) and Modified K-Nearest Neighbor (M-KNN). Different with KNN which classify the class only use the Euclidian distances, in M-KNN the classification is based on the validity of each training data and the Euclidian distance to calculate the weight voting value. We want to observe and compare the accuracy of these methods.

There are some previous research that use classification methods on data mining for identification and detection case, such as C4.5 algorithm, KNN, fuzzy and neural network. Hashi et al [8] use c4.5 algorithm and KNN method for predict diabetes disease. The dataset claimed by The Pima Indians Diabetes Database and Kidney Diseases, with 768 sample with 8 numerical value attribute and the result show that accuracy of Decision Tree: C4.5 get 90.43% and 76.96% for KNN method. Meanwhile Vadovsky and Paralic [9] in their case to classification Parkinson patient used speech signal applied C4.5 algorithm get accuracy about 60,71%. Baihaqi et al [10] combining C4.5 algorithm with fuzzy expert system get the accuracy result 81.82% in diagnose coronary arteri disease case. The combination between fuzzy and genetic algorithm can also be used to optimized membership function adjustment of few parameters for determining the quality of river water [11] and get better accuracy up to 5%. Whereas Subba et al [12] compare various intrusion

detection models with C4.5 algorithm and another methods. It raised accuracy about 99.25% in selection of feature. Neural network can also be used for classification, for example malware detection case [13], which determine the malware based on the training process of few parameters such as the list of manifest files, permission battery rating, and the size of application file.

K-Nearest Neighbor (KNN) also can be used to solved classification problems. Research conducted by Tan [14] for text classification used KNN method found that the use of DragPushing on KNN gives a better performance than the Centroid Based algorithms, Winnow, Naïve Bayes and C4.5 Algorithm. In testing this algorithm using case studies such as Reuter-21 578, Sector-48, and TDT-5. The use DragPushing on KNN itself can also be used for other object classification. The other study was conducted by Liu, Lee and Lin on fall detection by the silhouette image using KNN method [15]. In this study, three testing data compared with training data that is between 1: 9, 1: 3 and 1: 1 with a view to see the validity of the data used. Based on this study the accuracy of the values obtained 84.44% and can reduce the effect of the activity of the human upper body.

M-KNN is a method of determining the weighting of parameter k in the KNN and weighting parameter is determined by using a different procedure. The first thing to do is calculating the validity of all training data. Furthermore, the calculation of weight voting on all the test data using the data validity [16]. MKNN has been implemented by Simanjuntak et al [17] to identify diseases on soybean plants. By using soybean disease data set consist of 266 training data and the K value was determined using the Brute Force. The highest accuracy results in this test at 100% with a value of k = 1 and the average accuracy of 5 trials of 98.83%.

Based on previous researches, we want to know what the best method that implied to Jatropha Curcas disease identification between Decision Tree, KNN and M-KNN.

2. Research Method

In this paper, authors implemented three different approaches (Decision Tree, K-NN, and Modified K-NN) to identify Jantropha Curcas diseases from same datasets. The classification result from each approach than will be compared based on accuracy and computing performance. This section will show the individual algorithm of each approach.

2.1 Decision Tree

Decision Tree classification used to change the big of fact into a decision tree which represent the rules into database language such as Structured Query Language (SQL) to find record in certain category. Tree classification method is able to classify and show relation between atributtes. This method shows the factors that influence alternative results of the decision with final result estimated if the decision be accepted. The advantage of decision tree is able to break down the complex decision making process becomes simple decision making. So, decision maker will be interpreting solution of the problem. The concept of decision tree is change the data to be a decision which represented by tree and rule.

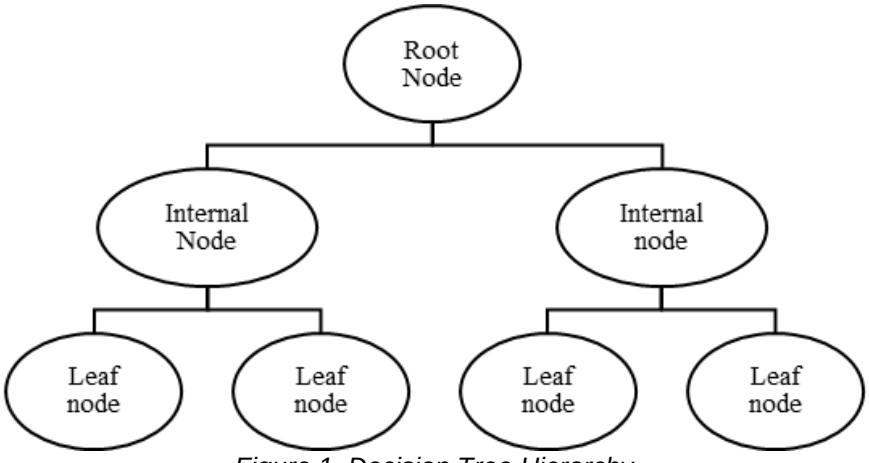

Figure 1. Decision Tree Hierarchy

Decision tree using hierarchy structure consist of root node, internal node, and leaf node which illustrated in Figure 1. Root node is node which located in the top. Internal node is node branch which has only one input and has at least two output. Leaf node is the last node which has one input and has no output. There are some algortihm which can use to form decision tree, such as ID3, CART, Sprint, SLIQ Public, Cis, Random Forest, Random Tree, ID3+, Oci, Clouds and C4.5 algorithm. C4.5 algorithm is develompment from ID3 algorithm [18].

2.2 K-Nearest Neighbor

Nearest Neighbor is the approach to looking for a case to calculate the closeness between the new cases with old cases based on matching the weight of a number of existing features. The classification does not use any model to be matched and only based on similarity with previous cases. The following is step-by-step on solving problem using KNN algorithm [19]:

1. Determine parameter k, which show the number of nearest neighbor
2. Calculate the square of eucledian distance between objects and data samples
3. Sort the object by eucledian distance in ascending mode
4. Classify predicted class based on the previous step based on the class that most appear on the k-nearest neighbour

2.3 Modified K-Nearest Neighbor (M-KNN)

To implement MKNN, we make sure the validity of training data first before calculating the weight voting [13].

2.3.1 Training Data Validity

Validity is used to calculate the number of points with the same label for all the data in the training data. The validity of any data depended on any nearest neighbors. After validating the data, then the data is used as information about the data. The Equation 1 used to calculate the validity of each training data.

$$validity\ (x) = \frac{1}{H}\sum_{i=1}^{H} S\ (lbl(x), lbl(N_i(x))) \quad (1)$$

Where H is the number of nearest points, lbl(x) is class of X, Ni(x) is the class label of nearest point x. S function is used to calculate the similarity between point a and b-nearest neighbor datas. S function is defined in the Equation 2.

$$S(a,b) = \begin{cases} 1, a = b \\ 0, a \neq b \end{cases} \quad (2)$$

where a is data in the training set, while b is other class except a in the training set.

2.3.2 Weight Voting

In MKNN method, the weight of each neighbor first calculated using 1 / (de + 1). Then, the validity of any data on the training set multiplied by weight based on Euclidean distance. So, the weight voting of each neighbor is calculated as follows Equation 3.

$$W(x) = validity\ (x) \times \frac{1}{d_e + 0{,}5} \quad (3)$$

where W(x) is the weight voting calculation, and de is the eucledian distance.

3. Results and Discussion

In this research are grouped into two that is training data and test data. The training data was obtained from experts of Jatropha plant disease at Indonesian Crops and Fiber Crops Research Institute. Data obtained from the results of direct interviews conducted in 2015 in the form of 9 diseases and 30 symptoms with the value of each symptom is worth between 0-1. The test data were obtained from direct observation data at Jatropha plantation in Situbondo. The test data obtained amounted to 166 datas that divided into 30 testing datas and 136 training datas. The obtained data will be the limitations in this study. In this section, testing for each method and obtained results of the accuracy described as follows:

3.1 Decision Tree

Testing has done 10 times using 31 testing data. In constructing a decision tree using a genetic algorithm that has 549 genes on each chromosome.

Based on Figure 2 can be seen that chromosome shaped matrix with size 9x61. Each line on the matrix shows a rule for 1 type of disease. There are 9 lines of rule that show the values of 9 types of Jantropha Curcas plant disease. Each rule consist of 61 genes, in which each pair of genes (2 adjacent genes) show the observed variables or symptoms.

0	93	1	67	4	90	…	4	90	0
1	45	2	56	2	49	…	3	100	1
3	89	2	99	3	94	…	4	21	2
2	4	3	23	4	92	…	3	43	3
4	91	1	65	2	25	…	2	66	4
1	39	1	36	3	50	…	3	74	5
2	98	4	77	1	74	…	2	98	6
0	77	2	39	3	38	…	1	99	7
2	25	3	59	2	90	…	0	77	8

Figure 2. Chromosome

Figure 3 shows the detail of the shape of chromosome on each line. Each gene pair consist of marks and values. The genes that show the marks are between 0-4, where each number represent the meaning shown in Table 1.

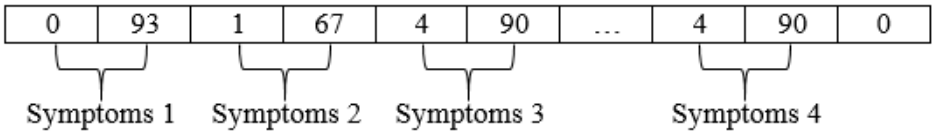

Figure 3. Chromosome Details

Table 1. Marks of Genes

Number	Meaning
0	The symptoms are ignored/not used
1	Operator <
2	Operator ≤
3	Operator >
4	Operator ≥

While genes that contain values are in the range 0-100. And the most recent genes in each line indicate the type of disease.

Table 2. Test Result

Testing Number	Accuracy (%)
1	0.00
2	22.58
3	3.00
4	3.00
5	3.00
6	9.70
7	9.70
8	10.00
9	5.60
10	3.00

Based on Table 2, the highest accuracy results when the second test is 22.58%. Accuracy obtained is low because the use of this method is often obtained a trait that can be categorized more than 1 type of disease with the suitability of all variables. It is also very difficult to describe the criteria of other types of diseases so that some diseases tend to appear more often as a classification

3.2. K-Nearest Neighbor

Testing has done by testing the value of k and testing of training data. W used 135 training data and 31 testing data. The k value test aims to see how many neighbors are drawn to the nearest which gives a good accuracy value. This test uses the k value of multiple 5 and uses 20% training data by the available total training data.

Table 3. Value of k Testing

K Value	Accuracy (%)
5	35.48
10	38.71
15	45.16
20	35.48
25	41.94

Based on the Table 3 can be seen that the value of k by 15 has the highest accuracy at 45.16%. K value of 15 proves the culmination and after than 15 are impairment values. The next test is using test data variant. Test data used multiplication 20% using value k = 15.

Table 4. Testing of Training Data

Number of Data	Accuracy (%)
20% data training and 31 testing data	45.16
40% data training and 31 testing data	45.16
60% data training and 31 testing data	45.16
80% data training and 31 testing data	41.94
100% data training and 31 testing data	51.61

Testing of training data in the Table 4 proves that the training data as much as 100% gives a maximum accuracy at 51.61%. This proves that in the training data 100% has the spread of data train evenly and provide a better solution.

3.2 Modified K-Nearest Neighbor (MKNN)

In MKNN method, a computational experiment has been done by testing the value of k and testing of training data. The training data used is 135 and the test data is 31. The k test aims to see how many neighbors are drawn to the nearest which gives a good accuracy value. This test uses the k value of multiple 5 and uses training data of 20% of the total training data available.

Table 5. Value of k Testing

K Value	Accuracy (%)
5	51.61
10	54.84
15	61.29
20	54.84
25	51.61

Table 5 shows that the value of k by 15 has the highest accuracy at 61.29%. K value of 15 proves the culmination and after than 15 are impairment values. The next test is using test data variant. Test data used multiplication 20% using value k = 15.

Table 6. Testing of Training Data

Number of Data	Accuracy (%)
20% data training and 31 testing data	61.29
40% data training and 31 testing data	41.94
60% data training and 31 testing data	67.74
80% data training and 31 testing data	51.61
100% data training and 31 testing data	61.29

Testing of training data in the Table 6 proves that the training data as much as 60% gives a maximum accuracy at 67.74%. This proves that in the training data 60% has the spread of training data evenly and provide a better solution. The uneven spread data training causes decreased accuracy when using all training data.

Table 7 shows that the best accuracy of all methods is 67.74% resulted by modified k-nearest neighbor method. M-KNN method can provide better accuracy because this method provides the calculation of validity value of each training data based on some nearest neighbors which spread between search space and have the same class label in

the first step. This can be a guarantee that if the data has a big value of validity then it will be a good reference to classify its class label based on the Euclidian distance.

Table 7. The Best Accuracy in Each Method

Methods	Accuracy (%)
Decision Tree	22.58
K-Nearest Neighbor	51.61
Modified K-Nearest Neighbor	67.74

4. Conclusion

The methods that have been tested for Jantropha Curcas identification cases are obtained with the highest accuracy is the M-KNN method with an accuracy of 67.74% using 60% training data and 31 testing data. These result can still be improved for the better by trying to use other methods for further research such as hybridization with genetic algorithm, neural network algorithm, or Support Vector Machine.

References

[1] J. Rodrigues et al., "Storage Stability of Jatropha Curcas L. Oil Naturally Rich in Gamma-Tocopherol," Ind. Crops Prod., vol. 64, pp. 188–193, 2015. https://doi.org/10.1016/j.indcrop.2014.10.048

[2] C. M. Fernandez, L. Fiori, M. J. Ramos, A. Perez, and J. F. Rodriguez, "Supercritical Extraction and Fractionation of Jatropha Curcas L. Oil for Biodiesel Production," J. Supercrit. Fluids, vol. 97, pp. 100–106, 2015. https://doi.org/10.1016/j.supflu.2014.11.010

[3] T. H. Saragih, A. A. Soebroto, and T. Yulianti, "Sistem Pakar Diagnosa Penyakit Tanaman Jarak Pagar (Jantropha Curcas L.) menggunakan Metode Dempster-Shafer (Expert System Diagnosis of Jatropha Curcas L. using the Dempster-Shafer Method)," DORO Repos. J. Mhs. PTIIK Univ. Brawijaya, vol. 7, no. 4, 2016.

[4] T. H. Saragih, D. M. N. Fajri, A. Hamdianah, and W.F Mahmudy, "Jatropha curcas disease identification using Fuzzy Neural Network," 2017 International Conference on Sustainable Information Engineering and Technology (SIET), pp 305-309, 2017. https://doi.org/10.1109/SIET.2017.8304153

[5] D. M. N. Fajri, T. H. Saragih, A. Hamdianah, and W.F Mahmudy, "Optimized fuzzy neural network for Jatropha Curcas plant disease identification," 2017 International Conference on Sustainable Information Engineering and Technology (SIET), pp 297-304, 2017. https://doi.org/10.1109/SIET.2017.8304152

[6] T. H. Saragih, D. M. N. Fajri, Y. P. Anggodo, A. Latief and W.F Mahmudy, "Jatropha Curcas Disease Identification With Extreme Learning Machine," Indonesian Journal of Electrical Engineering and Computer Science, vol 12, no.2, 2018. http://doi.org/10.11591/ijeecs.v12.i2.pp883-888

[7] A. V Sitanggang, A. A. Soebroto, and T. Yulianti, "Sistem Pakar Identifikasi Penyakit Tanaman Jarak Pagar (Jatropha Curcas L.) dengan menggunakan Metode Naive Bayes (Expert System of Jatropha Curcas L. Identification by using Naive Bayes Method)," DORO Repos. J. Mhs. PTIIK Univ. Brawijaya, vol. 7, no. 4, 2016.

[8] E. K. Hashi, M. Shahiduzzaman, and M. R. Hasan, "An Expert Clinical Decision Support System to Predict Disease using Classification Techniques," in International Conference on Electrical, Computer and Communication Engineering, 2017. https://doi.org/10.1109/ECACE.2017.7912937

[9] M. Vadovsky and J. Paralic, "Parkinson's Disease Patients Classification Based on The Speech Signals," in IEEE 15th International Symposium on Applied Machine Intelligence and Informatics, 2017. https://doi.org/10.1109/SAMI.2017.7880326

[10] W. M. Baihaqi, N. A. Setiawan, and I. Ardiyanto, "Rule Extraction for Fuzzy Expert System to Diagnose Coronary Artery Disease," in 1st International Conference on Information Technology, Information Systems and Electrical Engineering, 2016. https://doi.org/10.1109/ICITISEE.2016.7803062

[11] Q. Kotimah, W. F. Mahmudy, and V. N. Wijayaningrum, "Optimization of Fuzzy Tsukamoto Membership Function using Genetic Algorithm to Determine the River Water Quality," Int. J. Electr. Comput. Eng., vol. 7, no. 5, pp. 2838–2846, 2017. http://doi.org/10.11591/ijece.v7i5.pp2838-2846

[12] B. Subba, S. Biswas, and S. Karmakar, "A Neural Network based System for Intrusion Detection and Attack Classification," in 22nd National Conference on Communication, 2016. https://doi.org/10.1109/NCC.2016.7561088

[13] F. Al Huda, W. F. Mahmudy, and H. Tolle, "Android Malware Detection Using Backpropagation Neural Network," Indones. J. Electr. Eng. Comput. Sci., vol. 4, no. 1, pp. 240–244, 2016. http://doi.org/10.11591/ijeecs.v4.i1.pp240-244

[14] S. Tan, "An Effective Refinement Strategy for KNN Text Classifier," Expert Syst. Appl., vol. 30, no. 2, pp. 290–298, 2006. https://doi.org/10.1016/j.eswa.2005.07.019

[15] C.-L. Liu, C.-H. Lee, and P.-M. Lin, "A Fall Detection System using k-Nearest Neighbor," Expert Syst. Appl., vol. 37, no. 10, pp. 7174–7181, 2010. https://doi.org/10.1016/j.eswa.2010.04.014

[16] H. Parvin, H. Alizadeh, and B. Minaei-Bidgoli, "MKNN: Modified K-Nearest Neighbor," in Proceedings of the World Congress on Engineering and Computer Science, 2008.

[17] T. H. Simanjuntak, W. F. Mahmudy, and Sutrisno, "Implementasi Modified K-Nearest Neighbor dengan Otomatisasi Nilai K pada Pengklasifikasian Penyakit Tanaman Kedelai (Implementation of Modified K-Nearest Neighbor with Automation of K Value on Classification of Soya Plant Disease)," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 1, no. 2, pp. 75–79, 2017.

[18] D. T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining. Hoboken, New Jersey: A John Wiley & Sons, Inc., 2005.

[19] J. Han, M. Kamber, and J. Pei, Data Mining: Concept and Techniques. Waltham, USA: Morgan Kaufmann, 2012.