Comparison of algorithm Support Vector Machine and C4.5 for identification of pests and diseases in chili plants

M Irfan[1,*], N Lukman[2], A A Alfauzi[1] and J Jumadi[2]

[1] Department of Informatics Engineering, Sunan Gunung Djati State Islamic University Bandung, Jl. A.H. Nasution 105, Bandung, Indonesia
[2] Department of Information and Communication Technology, Asia E University Kuala Lumpur, Malaysia

[*]irfan.bahaf@uinsgd.ac.id

Abstract. Data from the Central Bureau of Statistics of the population working in the agricultural sector continued to decline from 39.22 million in 2013 to 38.97 million in 2014, the number dropped back to 37.75 million in 2015. According to the MIT G-Lab Team (global entrepreneurship program) concludes five factors that make it difficult to raise agricultural productivity to compete in the domestic market, namely the low education of farmers in dealing with pests, the difficulty of access to finance for rural areas, lack of skills, lack of access to information and lack of application of agricultural technology. Chili plants are plants that are very susceptible to pests so BPS noted a decrease in chili production reaching 25%. Information about chili pests is collected so that it becomes a database that can be used to identify disease pests using the data mining method. The use of data mining algorithms is expected to help in the identification of pests and diseases in chili plants. In this study comparing the performance classification techniques of Support Vector Machine (SVM) and C4.5 algorithms. The attributes used consist of Leaves, Stems, and Fruits. By using each training data and testing data as many as 30 data. The results of the study were conducted, based on the accuracy of SVM, which was 82.33% and C4.5 89.29 %%. The final result of this study was that the accuracy of the C4.5 method was better.

1. Introduction

Information from the Central Statistics Agency (BPS), the number of people working in the agricultural sector continued to decline from 39.22 million in 2013 to 38.97 million in 2014, the number dropped back to 37.75 million in 2015 [1]. According to the MIT Team G-Lab (global entrepreneurship program) concluded five factors that make it difficult to raise agricultural productivity to compete in the domestic market, namely the low education of farmers in overcoming pests, difficulty in access to finance for rural areas, lack of skills, lack of access to information and lack of application of agricultural technology [2].

The factors that influence the income of these farmers are none other than pests and diseases. Pests are a group of plant disrupting organisms that can damage crops both physically and physiologically [3,4]. The lack of information can cause problems for farmers, with alternative information such as systems that can identify pests and diseases, maybe farmers can go out and act faster to deal with pests and the disease. The entry of technology into the world of agriculture is expected to be an alternative for farmers so that farmers can have a lot of information on the world of agriculture which has become

their daily life. The SVM and C4.5 algorithms each have advantages and disadvantages [5]. Therefore, in this study we will make a comparison between the two algorithms to obtain the maximum algorithm in the classification of pests and diseases in chili plants. The comparative parameters of the two algorithms are the level of system accuracy, and algorithm processing time [6].

This article consists of several parts, the first part of the introduction that explains the background of the problem, the two methodologies, the three discussions and the results of the study and the final conclusions of the results of the study.

2. Research methodology

2.1. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a learning machine method that works on the principle of Structural Risk Minimization (SRM) with the aim of finding the best hyperplane that separates two classes in input space [5,7]. The best hyperplane is a hyperplane located in the middle between two sets of objects from two classes. the best separator hyperplane between the two classes can be found by measuring the margin of the hyperplane and looking for the maximum point [8,9]. Margin is the distance between the hyperplane and the closest pattern of each class. The closest pattern is called Support Vector.

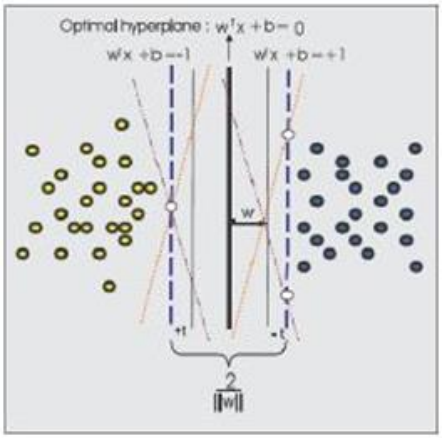

Figure 1. SVM concept to search the best hyperplane.

2.2. Algorithm C4.5

The C4.5 algorithm is a well-known algorithm that is used for data classification that has numerical and categorical attributes [6,10]. The results of the classification process in the form of rules can be used to predict the value of the discrete type attribute of the new record [11–14]. C4.5 algorithm itself is a development of the ID3 algorithm, where development is done in terms of, can overcome missing data, can handle continuous data and pruning. In general, the C4.5 algorithm for building decision trees is as follows:

- Select the attribute as root.
- Create a branch for each value.
- Share cases in branches.
- Repeat the process for each branch until all the cases in the branch have the same class [4].

To choose the root attribute, it is based on the highest gain value of the existing attributes. To calculate the gain, use the formula as shown in equation 1 below:

Gain (S.A) = Entropy (S) - Zi; =i Entropy (Si)

S : set of cases
Si : number of cases on the i-partition
A : attribute
| Sr | : number of cases in S

N : number of partition attributes A

3. Discussions

3.1. Data analysis

Following is the implementation of the Support Vector Machine (SVM) and C 4.5 methods in the system of identifying pests and diseases in chili with the case and sample data as follows [5,6,10,15–17]:

Table 1. Data training

Leaf	Stem	Fruit	Statement
Spotted Yellow	Not Grow	Not Change	Plant Disease
Spotted Black	Become Small	Spotted Black	Plant Disease
Become Yellow	Spotted Black	Coloured Yellow	Disease
Become Yellow	Become Small	Coloured Yellow	Disease
Become Yellow	Spotted Black	Not Change	Disease
Spotted Black	Not Grow	Coloured Yellow	Plant Disease
Become Yellow	Spotted Black	Not Change	Disease
Spotted Yellow	Become Small	Spotted Black	Plant Disease
Become Small	Spotted Black	Not Change	Disease
Spotted Yellow	Not Change	Coloured Yellow	Plant Disease
Spotted Black	Become Small	Not Change	Plant Disease
Become Yellow	Spotted Black	Not Change	Disease
Spotted Black	Become Small	Coloured Yellow	Plant Disease
Spotted Yellow	Not Grow	Spotted Black	Plant Disease
Become Yellow	Not Grow	Coloured Yellow	Disease
Spotted Black	Become Small	Not Change	Plant Disease
Spotted Yellow	Not Grow	Coloured Yellow	Plant Disease
Spotted Yellow	Not Grow	Not Change	Plant Disease
Become Yellow	Become Small	Not Change	Plant Disease
Spotted Black	Not Grow	Coloured Yellow	Plant Disease

3.2. Calculation of method C 4.5

Step 1: Change the data into a model tree:

To determine the initial node, the Gain value is calculated for each attribute using the formula

$$(S, A) = Entropy\ (S) - \sum_{Vevslustion\ (A)} \frac{|Sv|}{|S|} entropy_{(Sv)}$$

Previously it was calculated the entropy value of each attribute using the entropy formula

$$\sum_{i}^{c} -p_i\ log_2 p_i$$

Calculate the total entropy of the case first as follows:

$$Entropy \text{ total} = -\left(\frac{19}{30}\right) \text{x } \log 2\left(\frac{19}{30}\right) + -\left(\frac{11}{30}\right)\log 2\left(\frac{11}{30}\right) = 0{,}9480$$

After getting the entropy from the whole case, then analyse each attribute and its values and calculate the gain.

$$\text{Leaf Gain} = 0{,}9480-\left(\left(\frac{10}{30}\right)x0 + \left(\frac{9}{30}\right)x0{,}50325 + \left(\frac{11}{30}\right)x0{,}43949 = 1{,}2602$$

$$\text{Gain} = 0{,}9480-\left(\left(\frac{11}{30}\right)x0{,}6840 + \left(\frac{11}{30}\right)x0{,}6840 + \left(\frac{8}{30}\right)x0{,}54356\right) = 0{,}30149$$

$$\text{Fruit Gain} = 0{,}9480-\left(\left(\frac{11}{30}\right)x0{,}94566 + \left(\frac{5}{30}\right)x0 + \left(\frac{14}{30}\right)x0{,}98522 = 0{,}14156$$

From the search for the gain value above, we get a table of gain values in table 2.

Table 2. Gain value data

Attribute	Value	Sum (value)	Sum (Pest)	Sum (Diseases)	Entropy
Leaf	Yellow Spots	10	10	0	0
	Black spots	9	8	1	0,50325
	Yellow	11	1	10	0,43949
Stem	Not growing	11	9	2	0,6840
	Shrinking	11	9	2	0,6840
	Black spots	8	1	7	0,5435
Fruit	Fruitlessness	11	7	4	0,9456
	Black spots	5	5	0	0
	Orange	14	6	8	0,9822

From the table above, the greatest gain value is obtained, namely the leaf gain value. So that the Leaf is chosen as the initial node.

Step 2: Arrange the tree

Arrange the tree starting from the selected node in step 1 as in figure 2:

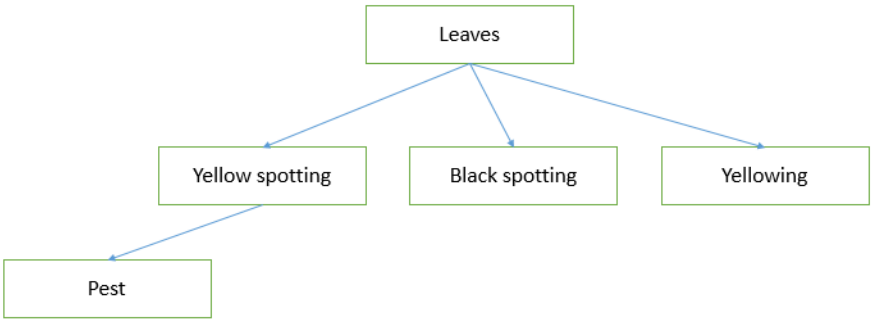

Figure 2. First decisions tree.

From the tree above the next leaf node is determined. The node to be branched is a node on the Black Spotting and Yellowing leaves. To determine the decision node attribute under the attributes of the Black Spots Leaves, and Yellowing, the gain values of the remaining attributes are calculated, namely Stems

and Fruits. The biggest gain value will be the decision knot under the Leaves of Black Spots and Yellowing. Both steps are carried out repeatedly until the results of the data are obtained.

3.3. Calculation of the SVM method

Support Vector Machine (SVM) is a learning machine method that works on the principle of Structural Risk Minimization (SRM) with the aim of finding the best hyperplane that separates two classes in input space.

Table 3. Criteria and weighting according to samples from plantation.

Leaf	Stem	Fruit
50%	20%	30%

Operational variables and equations used:

Table 4. Training data samples.

No	Leaf	Stem	Fruit	*Goal*
1	2	1	0	1
2	1	0	0	0
3	0	1	1	1
4	0	2	1	1
5	2	0	0	1
6	1	0	1	1
7	1	2	2	1
8	2	1	1	1
9	0	0	0	0
10	1	1	1	1

Table 5. Description criteria.

Name Criteria	Characters	Number
Leaf	Spotted Yellow	2
	Spotted Black	1
	Become Yellow	0
Stem	Not Grow	2
	Become Small	1
	Spotted Black	0
Fruit	Not Fruit	2
	Spotted Black	1
	Yellow	0
Goal	Plant Disease	1
	Disease	0

Recommendations for diseases and pests with three dividing lines Y1, Y2, Y3

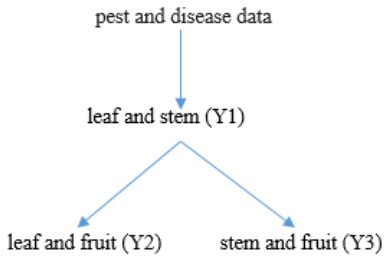

Figure 3. Three separating rows Y1, Y2, Y3.

Table 6. Y1 separator.

Leaf	Stem	Target
2	1	1
1	0	1
0	1	1
0	2	1
2	0	1
1	0	1
1	2	1
2	1	1
0	0	1
1	0	1

```
data1    = data (1:2);
target1  = [1;1;1;1;1;1;1;1;1;1]
y1       = SVM train (data1, target1)
```

4. Results

Confusion matrix is one method that can be used to measure the performance of a classification method. Basically confusion matrix contains information that compares the results of the classification carried out by the system with the results of the classification that should be.

Table 7. Confusion matrix.

		Prediction class	
		1	0
In Fact Class	1	TP	FN
	0	FP	TN

The information for the following table is stated as follows:

- True POSTIVE (TP), which is the number of documents from class 1 correct and classified as class 1.
- True Negative (TN), which is the number of documents from class 0 that are correctly classified as class 0.
- False Positive (FP), which is the number of documents from class 0 that are incorrectly classified as class 1.
- False Negative (FN), which is the number of documents from class 1 that are incorrectly classified as class 0.

Calculations to find accuracy can be formulated by: $accuration = \frac{TP + FN}{TP + FN + FP + TN} \times 100\%$

Table 8. Comparison of system results and actual results.

No	SVM	C4.5	EXPERT
1	Plant Disease	Plant Disease	Plant Disease
2	Plant Disease	Plant Disease	Plant Disease
3	Disease	Disease	Disease
4	Disease	Disease	Disease
5	Disease	Plant Disease	Disease
6	Plant Disease	Plant Disease	Plant Disease
7	Disease	Disease	Disease
8	Plant Disease	Plant Disease	Plant Disease
9	Disease	Disease	Disease
10	Plant Disease	Plant Disease	Plant Disease
11	Plant Disease	Plant Disease	Plant Disease
12	Disease	Disease	Disease
13	Plant Disease	Plant Disease	Plant Disease
14	Plant Disease	Plant Disease	Plant Disease
15	Plant Disease	Plant Disease	Disease
16	Plant Disease	Plant Disease	Plant Disease
17	Disease	Plant Disease	Plant Disease
18	Plant Disease	Plant Disease	Plant Disease
19	Plant Disease	----	Plant Disease
20	Plant Disease	Plant Disease	Plant Disease
21	Plant Disease	Plant Disease	Plant Disease
22	Plant Disease	Plant Disease	Plant Disease
23	Plant Disease	Plant Disease	Plant Disease
24	Plant Disease	Plant Disease	Disease
25	Plant Disease	Plant Disease	Plant Disease
26	Plant Disease	Plant Disease	Plant Disease
27	Disease	Disease	Disease
28	Plant Disease	----	Disease
29	Disease	Disease	Disease
30	Disease	Disease	Plant Disease

Based on the comparison table above, the calculation of SVM and C4.5 results is obtained as follows:

4.1. SVM algorithm

Table 9. Confusion matrix SVM.

		EXPERT RESULTS	
		Plant Disease	Disease
EXPERT SVM	Plant Disease	17	3
	Disease	2	8

The validity of the system is assessed by counting TP, TN, FP, and FN values from Table 9.

TP = 17 + 8 = 25
TN = 8 + 17 = 25
FP = 2 + 3 = 5
FN = 3 + 2 = 5

$$System\ performance = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

$$System\ performance = \frac{25 + 25}{25 + 25 + 5 + 5} \times 100\%$$

$$System\ performance = \frac{50}{60} \times 100\%$$
$$System\ performance = 83.33\%$$

4.2. Algorithm C4.5

Table 10. Confusion matrix C4.5.

		Expert Result	
		Plant Disease	Disease
C4.5 RESULT	Plant Disease	18	2
	Disease	1	7

The validity of the system is assessed by counting TP, TN, FP, and FN values from Table 10.

TP = 18 + 7 = 25
TN = 8 + 17= 25
FP = 1 + 2 = 3
FN = 2 + 1 = 3

$$System\ performance = \frac{TP + TN}{TP + TNFP + FN} \times 100\%$$
$$System\ performance = \frac{25 + 25}{25 + 25 + 3 + 3} \times 100\%$$
$$System\ performance = \frac{50}{56} \times 100\%$$
$$System\ performance = 89.29\%$$

5. Conclusion

The results of a comparative study of Support Vector machine and C4.5 algorithms in Identifying Pests and Diseases in Chili Plants (Case Study of Cintarasa Village Plantation) can be seen that C4.5 algorithm is more accurate than SVM algorithm with C4.5 89.29% accuracy while SVM Algorithm 82.33 %. In SVM algorithm speed is faster than c4.5 algorithm because SVM data processing is simpler than C4.5.

As the end of the research, the researcher presented several suggestions that were expected to be useful for the interests of the parties concerned. These suggestions are other researchers, namely the discussion in this study is limited to only two algorithms namely SVM and C4.5, it is expected that the next researcher can examine other algorithms or other datasets.

References

[1] BPS 2013 *Proyeksi Penduduk Indonesia Indonesia Population Projection* No 6
[2] Pujiawati L 2016 *Statistik Daerah Kota Bandung* (Bandung: BPS Kota Bandung)
[3] Astuti I and Sutarno H 2017 The Expert System of Children's Digestive Tract Diseases Diagnostic using Combination of Forward Chaining and Certainty Factor Methods *Int. Conf. Sci. Inf. Technol.* 608–612
[4] Irfan L P, Mohamad and Ayuningtias 2017 Analisa Perbandingan Logic Fuzzy Metode Tsukamoto , Sugeno , Dan Mamdani (Studi Kasus: Prediksi Jumlah Pendaftar Mahasiswa Baru Fakultas Sains Dan Teknologi Universitas Islam Negeri Sunan Gunung Djati Bandung) Laras P : Analisa Perbandingan Logic ... L *J. Tek. Inform.* **10** 1
[5] Athoillah M 2018 Pengenalan Wajah Menggunakan SVM Multi Kernel dengan Pembelajaran yang Bertambah *J. Online Inform.* **2** 2 84
[6] Irfan M, Uriawan W and Lukman N *Implementation of Algorithm C 45 for Classification of al-Quran Letters* 105 3–6

[7] Khosyi'Ah S, Irfan M, Maylawati D S and Mukhlas O S 2018 Analysis of Rules for Islamic Inheritance Law in Indonesia Using Hybrid Rule Based Learning *IOP Conference Series: Materials Science and Engineering* **288** 1
[8] Fauzan R, Indrasary Y and Muthia N 2018 Sistem Pendukung Keputusan Penerimaan Beasiswa Bidik Misi di POLIBAN dengan Metode SAW Berbasis Web *J. Online Inform.* **2** 2 79
[9] Charim A, Basuki S and Akbi D R 2019 Detect Malware in Portable Document Format Files (PDF) Using Support Vector Machine and Random Decision Forest *JOIN (Jurnal Online Inform.* **3** 2 99–102
[10] Maylawati D S A, Ramdhani M S, Rahman A and Darmalaksana W 2017 Incremental technique with set of frequent word item sets for mining large Indonesian text data *5th Int. Conf. Cyber IT Serv. Manag. CITSM 2017* 1–6
[11] Elisa E 2017 Analisa dan Penerapan Algoritma C4 . 5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT . Arupadhatu Adisesanti *JOIN (Jurnal Online Inform.* **2** 1 36–41
[12] Revathy R 2017 *C4.5 Algorithm*
[13] Darmawan E 2018 C4.5 Algorithm Application for Prediction of Self Candidate New Students in Higher Education *J. Online Inform.* **3** 1 22
[14] Muslim M A, Nurzahputra A and Prasetiyo B 2018 *Improving Accuracy of C4 . 5 Algorithm Using Split Feature Reduction Model and Bagging Ensemble for Credit Card Risk Prediction* **1996** 141–145
[15] Liu K, Xu L and Zhao J 2015 Co-extracting opinion targets and opinion words from online reviews based on the word alignment model *IEEE Trans. Knowl. Data Eng.* **27** 3 636–650
[16] Herliani M 2015 *Aplikasi Pencarian Buku Dengan Menggunakan Metode tf-idf dan Vector Space Berbasis Web Pada Perpustakaan Sekolah Menengah Atas Negeri 2 Pangkal Pinang* 8
[17] Setiawati D, Taufik I, Jumadi and Budiawan W Z 2016 Klasifikasi Terjemahan Ayat Al-Quran Tentang Ilmu Sains Menggunakan Algoritma Decision Tree Berbasis Mobile *J. Online Inform.* **1** 1 24–27