Journal of Information Technology and Computer Science
Volume 3, Number 2, 2018, pp. 175-183
Journal Homepage: www.jitecs.ub.ac.id

# Classification of Physical Soil Condition for Plants using Nearest Neighbor Algorithm with Dimensionality Reduction of Color and Moisture Information

Dahnial Syauqy[1], Hurriyatul Fitriyah[2], Khairul Anwar[3],

[1,2,3]Fakultas Ilmu Komputer, Universitas Brawijaya
[1]dahnial87@ub.ac.id, [2]hfitriyah@ub.ac.id, [3]khairulanwarr@hotmail.com

Received: 31 August 2018; accepted: 5 November 2018

**Abstract**. Determining the quality of soil is an important task to perform especially on newly opened agricultural land since it may provide a significant impact on the growth of plants. One alternative to determine physical soil quality is by visually observe the color of the soil and measure its moisture. This paper designed an embedded system to classify soil condition for plants according to the dimensionality reduction of color and moisture information from the soil using k-NN algorithm. The dimension of attribute information was reduced using correlation analysis to achieve lower computational time and lower memory usage on embedded system. In this study, 39 sample of soil from various location were collected and categorized by soil expert using visual observation. In the accuracy testing on the system that used four attributes, 100% accuracy was given by 60:40 ratio with 7 neighbors. In contrast, the system that used only two attributes, 100% accuracy was given by 60:40 ratio with 5 nearest neighbors. The resource usage testing shown that by using reduced attributes dimension, the resource usage can be lowered as many as 188 bytes on program storage and 192 bytes on global variable usage. Moreover, the average of computation time performed by the system using reduced attribute dimension was 5.4 ms compared to the system that used all attributes which was 6.2 ms.
**Keywords:** classification, nearest neighbor algorithmn, dimensionality reduction

## 1 Introduction

As an agricultural country, Indonesia has many types of agricultural lands, such as wetlands, dry field, and shifting cultivation lands. According to BPS-Statistics Indonesia on 2015, Indonesia has 8.092.906,80 Ha of wetlands, 11.861.675,90 Ha of dry fields, and about 5.190.378,40 Ha of shifting cultivation land. Other than that, there were also temporarily unused land of about 12.340.270,20 Ha [1]. Geographical condition may also inflict varying soil condition throughout the land [2].

The quality and the fertility of the soil affect the growth of plants planted on it [3]. Thus, determining the quality of soil is an important task to perform especially on newly opened agricultural land. Generally, the parameters of soil fertility can be categorized in three areas; physical, biological, and chemical [4].

The quality of soil can be examined visually based on its appearance. Soil that has darker color contains more organic matter compared to brighter soil [5]. A massive amount of nutrition and water made fertile soil appears darker. The usual method involves the comparison of the soil with soil color chart as standard. In traditional way, the task can be done by visually observing the color of the soil and compare it toward a standardized color chart [6]. That means, the physical appearance of soil can be easily observed without additional complex procedure such as using biological or chemical material. However, it relatively needs time and the decision is affected by the condition of light and individual's color perception [7].

Another parameter that defines the quality of soil is its moisture. The moisture of soil represents the quantity of water contained in the soil. In dry season, soil tends to have less water than in rainy season. The soil in location which is far form water source also tends to have less water contents in it. The ability of soil to contain water is one of important factor for the growth of plants above it [8].

In 2017, Prasetyo et.al. designed low budget system to detect soil fertility especially in Cihaur village. He only used moisture data using soil hygrometer sensor and only used simple value thresholding comparison which achieved 75% accuracy [9]. In our previous research, naïve Bayes had been used as the classification algorithm for color and moisture information. It used four dimensions of data as the input attributes. It had 100% accuracy but relatively longer time to compute. It also required to perform offline pre-processing computation before inputting the training dataset [10].

Based on the previous statements about the importance of determining the quality of soil for plants, this study proposed a system that is able to classify soil condition for plants according to the color and moisture information from the soil. The proposed system used color and moisture sensor to extract the visual information of soil. Then, the dimension of all information would be reduced to achieve lower computational time and lower memory usage on embedded system.

## 2 Methodology

The system uses TCS3200 to sense color and FC-28 to measure the moisture of soil, respectively. Both sensors were connected to Arduino Uno that would classify the data into fertile or non-fertile category. The category was then displayed in an LCD 16x2 as an output. The block diagram of the system is shown in Figure 1.

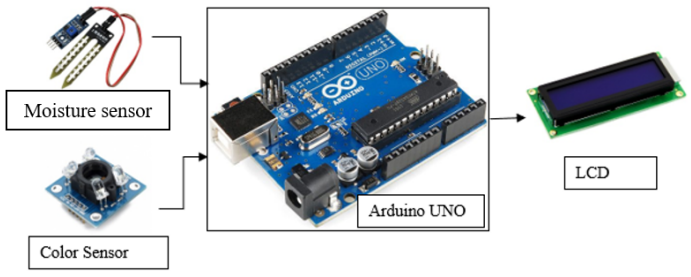



Fig. 1. Block diagram of the proposed system.

Moisture sensor FC-28 was chosen as it is designed specifically to measure moisture of soil. The sensor has two tips that are inserted into soil during data acquisition. It does not require specific condition for acquiring data. Whilst the color sensor TCS3200 is sensitive to illumination thus require special case during data acquisition. Different ambient lighting could cause different color value for a soil sample. Hence the system was developed in special case where the soil must be put inside a jar and its bottom was inserted into a fully covered black case. Lighting inside the case was only came from the sensor's lighting. Figure 2 shows the implementation of the hardware system using moisture sensor and color sensor. The hardware system had been implemented and used in the previous research using naïve Bayes system [10].

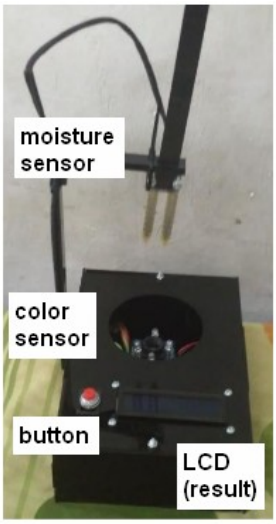



Fig. 2. Hardware implementation of the proposed system.

Color sensor extracts three components of color which were value of Red, Green and Blue (RGB). Hence the classification utilized four features, three from color sensor and one from moisture sensor. All four features are a numeric data type, hence k-NN (k-Nearest Neighbors) is suitable to be used for classification. The k-NN also popular in classification as it is does not require dimensionality enlargement as in SVM to increase the accuracy. The classification was embedded in a microcontroller Arduino that has small computation memory hence keeping dimension to the least is important. The k-NN simply measures distance between a data to all the training data and sort it ascendingly. It then chooses k number of nearest data training and assign the data to the majority class.

### 2.1 Data Acquisition

In this study, 39 sample of soil from various location were collected. Each sample was categorized using visual observation by expert in Soil Science laboratory. The RGB color and moisture values were used to classify the soil condition. The color sensor TCS3200 has digital output of 8-bits, resulting value between 0 – 255. The moisture sensor FC-28 has digital output of 10-bits, resulting value between 0 – 1023. The output class were separated into two: high organic level which is correlated to "good"

for plants, and low organic level which is correlated into "bad" for plants. The number of samples for each class is shown in Table 1.

Table. 1. The number of samples of training dataset and their classes

| Organic Level Class | Number of Samples on dataset |
|---|---|
| High | 24 |
| Low | 15 |
| TOTAL | 39 |

### 2.2 Data Normalization

k-NN utilizes distance between data to define nearest neighbors. Distance is a measure that is sensitive to data range. Different data range between features could yield a classification that depends only in the larger data range. This is because distance between objects in smaller range feature is insignificant when it is added to distance of larger range feature. This study perform normalization to the data set by scaling each feature in a range of 0 to 1.

### 2.3 Dimensionality reduction Using Correlation Between Feature-Pairs

As the classification was embedded in Arduino that has limited computational memory, reducing dimension of data set is beneficial. Simple dimensionality reduction method usually carried out by analyzing correlation between feature-pairs. Finding two or one features from all of four features would reduce computational time in the embedded system.

A correlation analysis is used in feature selection to determine relation between two different features [11]. Since the data are numeric, then Pearson's correlation is applied. The correlation uses linear approximation between two variables. The coefficient is calculated using Equation 1.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}} \quad (1)$$

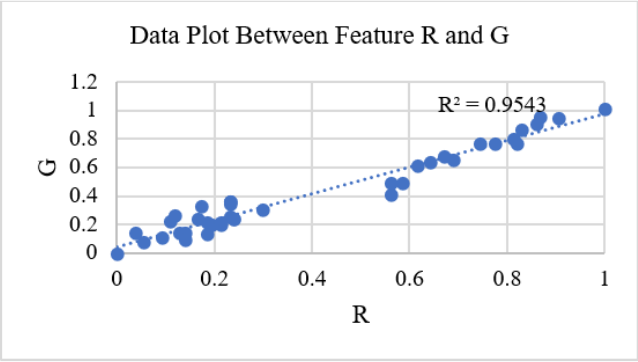
Table. 2. Correlation Coefficient, r, between features

| | R | G | B | Moist |
|---|---|---|---|---|
| **R** | 1.00 | 0.98 | 0.80 | -0.85 |
| **G** | 0.98 | 1.00 | 0.80 | -0.82 |
| **B** | 0.80 | 0.80 | 1.00 | -0.70 |
| **Moist** | -0.85 | -0.82 | -0.70 | 1.00 |

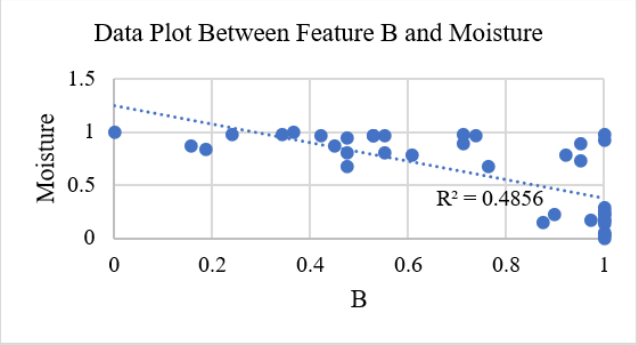
Equation 1 results coefficients that has value of -1 to +1 where positive sign means both features is directly proportional and negatives means both features are inversely proportional. The magnitude shows how strong the correlation between both features is, where 1 means strong and 0 means weak. This study used linear

approximation of correlation analysis. The coefficients r of each pair of features are shown as matrix in Table 2.

High magnitude of r value means the feature pair has high similarity. Both features has similar information of each data and fluctuate similarly thus eliminating one of them would not affect the classification. In opposite, low r value means the feature pair has diverse information hence both should be kept as features. Feature selection usually choose to keep one from feature pair with high r value. In this study, two selected features was simply chosen to be the feature pairs with lowest r value, which was -0.70 of B and Moisture feature pair. The data plot between R and G (r = 0.98), B and moisture (r = -0.70) is shown in Figure 3(a) and 2(b), respectively. As can be seen visually, plot of R and G appears to be similar compared to B and moisture.




(a)




(b)

Fig.3. Data Plot between features; (a) R and G, (b) B and Moisture

### 2.4 K-Nearest Neighbors (k-NN)

K-nearest neighbors is one of classification method that categorizes data based on the class of its *k* number of neighbors. Neighbors are identified as the closest data with the nearest distance. Euclidian distance is one of the most common distance approximations which is based on distance measurement in vector space. The data are

classified to majority of classes in the *k*-neighbors data. The pseudocode for k-NN is shown in Figure 4.

```
Input (Data training; A Data testing, k)
       For each data in Data Training
              Measure distance between a data testing to it
       End
Sort distance ascendingly
Vote majority class from k nearest distance's data
Output Class of a Data testing
```

Fig.4. Pseudocode of KNN

### 2.5 Testing and Analysis

Two tests were conducted in term of classification accuracy and computational time. Both test are needed to justify whether data reduction has advantage in computation time but still maintain high accuracy.

a. Classification accuracy

Testing on the k-NN classification was conducted by using all four features (R, G, B, and Moisture) and two features (B, moisture). The test was performed on ratio of training-testing data and number of neighbors, *k*. The result is shown in Table 3. As can be seen in the Table, using two features with the least correlation coefficient gave similar accuracy compared to when using all four features. In four features, 100% accuracy was given by 60:40 ratio with 7 neighbors. In two features, 100% accuracy was given by 60:40 ratio with 5 nearest neighbors. The two features even excel in term of data number. In two features, using 60% of data as training data means that only $23\times2$ vector data are used. It reduces the number of distance calculation from testing data to training data. Using 5 neighbors instead of 7 also reduces the number of voting input.

Table 3. Accuracy of Various Data ratio and number of neighbors *K*

| Ratio (Training : Testing) | k | Accuracy (%) | |
|---|---|---|---|
| | | 4 features | 2 features (B and Hum) |
| 20:80 | 3 | 85.47 | 86.32 |
| | 5 | 76.92 | 58.97 |
| | 7 | 58.97 | 67.52 |
| 40:60 | 3 | 99.15 | 98.29 |
| | 5 | 98.29 | 97.44 |
| | 7 | 99.14 | 88.89 |
| 60:40 | 3 | 99.15 | 98.29 |
| | 5 | 99.15 | 100.00 |
| | 7 | 100.00 | 100.00 |
| 80:20 | 3 | 100.00 | 100.00 |

| | | |
|---|---|---|
| 5 | 100.00 | 100.00 |
| 7 | 100.00 | 100.00 |

b. Resource usage and Computational time

According the previous accuracy result, the system was tested using two scenarios. First, the system that used four attributes as input parameter and k=7. Second, the system that used two attributes (only B and moisture) and k=5. Using 60:40 ratio, the dataset was divided into 23 training data and 16 testing data. In the first scenario, it can be seen in Figure 5, that the system using four attributes consumed exactly 7674 bytes and used 763 bytes of global variable.

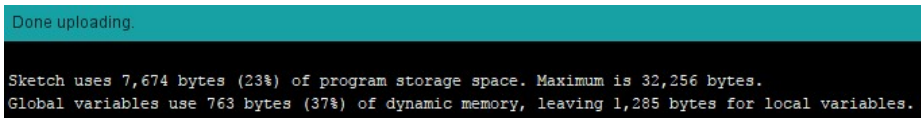



Fig.5. Resource usage on 4 attributes 7-NN classification

In other hand, the system that used dimensionality reduction (only used B and moisture information) took space as many as 7486 bytes and 571 bytes of global variable as it is shown in Figure 6. That means, by using reduced attributes dimension, the resource usage can be lowered especially when it comes to embedded system.

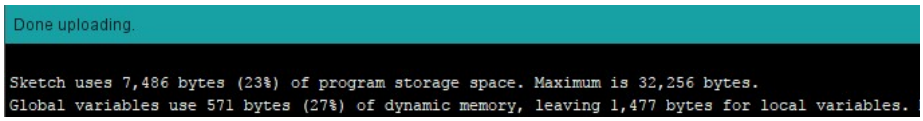



Fig.6. Resource usage on 2 attributes 5-NN classification

In k-NN algorithm, the number of attributes provides a significant impact on computation time. That happened because Euclidean distance computation become more complex on higher dimension. In this test, both scenarios were tested on Arduino UNO board. Computation time was started and finished exactly when k-NN computation took place. The comparison result of both scenarios is displayed in Table 4. It can be implied that the average of computation time performed by the system using attributes dimensionality reduction was relatively shorter than the system that used all attributes.

Table. 4. Comparison of computation using 4 attributes and 2 attributes

| R | G | B | Moist (%) | Organic Level | Computation time using 4 attributes (ms) | Computation time using 2 attributes (ms) |
|---|---|---|---|---|---|---|
| 27 | 81 | 58 | 603 | High | 5 | 4 |
| 43 | 113 | 198 | 504 | High | 5 | 4 |
| 23 | 93 | 102 | 698 | High | 5 | 4 |
| 40 | 107 | 128 | 503 | High | 5 | 4 |
| 47 | 93 | 128 | 667 | High | 6 | 5 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 85 | 120 | 185 | 635 | High | 5 | 4 |
| 47 | 83 | 128 | 576 | High | 5 | 4 |
| 69 | 111 | 141 | 680 | High | 5 | 6 |
| 71 | 110 | 243 | 543 | High | 5 | 4 |
| 58 | 105 | 160 | 572 | High | 5 | 4 |
| 229 | 237 | 255 | 205 | Low | 8 | 7 |
| 153 | 156 | 255 | 238 | Low | 8 | 7 |
| 173 | 189 | 255 | 103 | Low | 9 | 8 |
| 198 | 206 | 255 | 201 | Low | 8 | 7 |
| 218 | 230 | 255 | 283 | Low | 7 | 7 |
| 211 | 222 | 255 | 136 | Low | 8 | 7 |
| AVERAGE | | | | | 6.19 | 5.37 |

## 4 Conclusion

This paper proposed a system that is able to classify soil condition for plants according to the color and moisture information from the soil. The proposed system used color and moisture sensor to extract the visual information of soil. Then, the dimension of all information would be reduced using correlation analysis to achieve lower computational time and lower memory usage on embedded system. Finally, k-NN was used to classify the soil with particular attributes into two class; high organic and low organic. In this study, 39 sample of soil from various location were collected. Each sample was categorized by expert using visual observation. In the accuracy testing on the system that used four attributes, 100% accuracy was given by 60:40 ratio with 7 neighbors. However, in the system that used only two attributes, 100% accuracy was given by 60:40 ratio with 5 nearest neighbors. The resource usage testing also shown that by using attributes dimensionality reduction, the resource usage can be lowered. Moreover, the average of computation time performed by the system using attributes dimensionality reduction was relatively shorter than the system that used all attributes.

## References

1. Ministry of Agriculture: Buku Statistik Data Lahan Tahun 2012-2016. BPS-Statistics Indonesia (2017) [accessed online at http://epublikasi.setjen.pertanian.go.id/epublikasi/statistik%20data%20lahan/Buku%20Statistik%20Data%20Lahan%20Tahun%202012-2016/files/assets/basic-html/page1.html]
2. Alam, S, Sunarminto, B.H., Siradz, S.A: Karakteristik Kesuburan Tanah Pada Kondisi Iklim Berbeda Di Sulawesi Tenggara. Agriplus vol 23 (1) (2013)
3. Kadarwati, F.T. Evaluation of Soil Fertility to Sugarcane at Rembang District, Central Java. Jurnal Littri 22(2) (2016)
4. Warudkar, G, Dorle, S: Review on sensing the fertility characteristics of agriculture soils. International Conference on Information Communication and Embedded Systems (ICICES) (2016).
5. Brown, P.E., O'neall, A.M: The Color of Soils in Relation to Organic Matter Content. Soil Chemistry and Bacteriology (1923)
6. Pendleton, R.L.; Nickerson, Dorothy. Soil Colors And Special Munsell Soil Color Charts.

Soil Science: Volume 71 - Issue 1 - ppg 35-44 (1951)
7. Centeri, C, Vona, M, and Bíró, Z: Deviation in soil colour determination based upon students visual perception. World Congress of Soil Science, Soil Solutions for a Changing World (2010)
8. Wahyunie, E.D., Baskoro, D.P.T and Sofyan, M. Kemampuan retensi air dan ketahanan penetrasi tanah pada sistem olah tanah intensif dan olah tanah konservasi. Jurnal Tanah Lingkungan 14 (2), 73-78 (2012)
9. Prasetyo, T.F., Frastya, E.A., and Enceng Enda S. Sistem Pendeteksi Kesuburan Tanah Pada Desa Cihaur Kelompok Tani Bina Mandiri. Seminar Nasional Energi & Teknologi (SINERGI) (2017).
10. Anwar, K., Syauqy, D., & Fitriyah, H. Sistem Pendeteksi Kandungan Nutrisi dalam Tanah Berdasarkan Warna dan Kelembapan dengan Menggunakan Metode Naive Bayes. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 2, no. 9, p. 2491-2498, (2018).
11. Gogtay, N.J, Thatte, U.M. Principles of Correlation Analysis. Journal of The Association of Physicians of India Vol. 65 (2017)