Journal of Information Technology and Computer Science
Volume 6, Number 1, April 2021, pp. 1-8
Journal Homepage: www.jitecs.ub.ac.id

Application of Density Based Spatial Clustering Application With Noise (DBSCAN) in Determining the Quality of Keprok Orange and Siam Orange Hybrid in the Research Center of Orange and Subtropic Plants Batu City

Faiz Alqorni*[1], Wayan Firdaus Mahmudy[2], Agus Wahyu Widodo[3],

[1,2,3]Brawijaya University, Computer Science Faculty, Malang
[1]faisalqorni@student.ub.ac.id ,[2,3]{Wayanfm, a_wahyu_w }@ub.ac.id
*Corresponding Author

Received 21 October 2020; accepted 15 February 2021

Abstract. One of the tasks of the Indonesian Citrus and Subtropical Research Institute is research on crossing between citrus varieties to produce saplings with the best quality products through observation of the fruit produced. Because the amount of fruit production studied is very large, it requires a fast and accurate observation process, one of which is the clustering method of data mining. Observations were made using a clustering process or grouping Density Based Spatial Clustering Application with Noise (DBSCAN) on fruit characteristics that indicate quality. DBSCAN works by grouping data based on density, so that it is expected to find several data groups that are close to each other which shows the tendency of the quality of the observed fruit data as well as labeling outlays for data that are too far from the crowd. The results of the grouping will be analyzed to find out the number and characteristics of the groups formed where the results of the grouping are assessed using the Silhouette Coefficient method to determine the best parameter values. The results obtained in this study are obtained three group results which will be divided into medium quality, good, and not so good. The quality of grouping using the Silhouette Coefficient value of 0.69.

Keyword : data, group, DBSCAN, research

1 Introduction

The Research Institute for Citrus and Subtropical Fruits or commonly known as Balitjestro has the task of carrying out research activities on citrus plants and subtropical fruits such as oranges, apples, grapes, longan.

One of the objects of research conducted by Balitjestro Kota Batu is a cross between various types of Siamese citrus and tangerine varieties. Crosses are carried out with the aim of producing seeds that produce fruit of the desired quality. In 2015, citrus fruits were divided into 6 accession codes which indicated the types of tangerines and Siamese oranges that were crossed. Accession code P1 which is a cross between *siam banja* and tangerine *satsuma*, accession code P2 of *siam honey* cross with tangerine *satsuma*, accession code P3 with *siam mamuju* cross with *afternoon tangerine*, accession code P4 *siam mamuju* cross with *satsuma tangerine*, accession

code P5 *siam pontianak* cross with *soe* tangerines, and accession code P6 crossing *pontianak siam* with *satsuma* tangerines.

The difference between tangerines and siam oranges is that tangerines have a sweet, slightly sour and fresh taste, attractive skin color and are easy to peel. The weight of tangerines is 125-274 grams, the shape of tangerines is generally round, some are flattened, has a characteristic condition, the surface texture is rather rough, the skin color on the highlands can be up to orange. Has a thick fruit wall with a stiff outer skin layer, the thickness of the skin is 3.13-4.63 mm. This chayote is popular in the community. This fruit is commonly found in traditional markets and is generally widely circulated everywhere. Its sweet taste, thin skin and easy to peel is its characteristic. The size of siam oranges is relatively smaller than tangerines, with a range of 99.8-112.2 grams. The fruit is round with a round fruit tip. The fruit skin is yellowish green, shiny. The skin is about 1.8 - 2.5 mm thick, thinner than tangerines. The surface texture of the siam fruit skin is smoother because the pores are tighter and 0.8 mm small, tangerine pores are rarer with large sizes ≥1.2 mm [1]. The purpose of crossing is to produce the best hybrid results.

The different characteristics of each citrus variety, it takes an easy method to determine the quality of the crosses. One easy way to determine the distribution of quality of each cross is to use the clustering method. The results of clustering will produce data groups based on an imaginary line that divides the data, using the proximity of data point centers, using relationships between data to form a hierarchy, using a radius in each data point as group members [2].

The results of grouping using clustering can show the tendency of the characteristics of a group which can later be used to see the average quality of the crosses. Because the number of research results is very large, the time and effort required to perform manual clustering is not efficient, so we need a data mining model that is capable of clustering the research data.

One of the clustering algorithms that can be used in this study is Density-based spatial clustering of applications with noise (DBSCAN), DBSCAN works on the basis of data density or density. DBSCAN was developed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 1996 [3]. DBSCAN's performance in classifying the quality of agricultural products compared to several other grouping algorithms such as Partitioning Around Medoids (PAM) and Clustering Large Applications (CLARA) generally achieves better performance based on evaluation metrics such as purity, homogeneity, completeness, V-measure, recall, F-measure, and the Rand Index (Majumdar, et al., 2017)[18]. In addition to the agricultural sector, DBSCAN has also been successfully used in grouping galaxy cluster members and determining points that are not part of the galaxy cluster (Zhang, 2019)[19].

K-Means clustering is a grouping algorithm based on a group center which is the average value of each feature and member of these groups) [4] [5]. In this study, the K-means grouping method will be used as a comparison of the performance of DBSCAN.

This study assumes that there is a clustering tendencies in data's populace that divided the data based on provided features. The result of this study is limited to the scale and feature provided by data used and orange's variety.

2 Basic Theory

2.1 Citrus and Hybridization

Siamese oranges in general tend to be smaller than oranges in general. According to Endarto and Martini in 2016 [6], Siamese oranges have thin skin and are easy to peel, the fruit is not hollow, has a sweet taste, and has a high water content. Another characteristic of the Siamese orange is the yield of vitamin C which is quite high and the yield ranges from 1000-2000 fruits per tree per year. Outwardly, chayote tends to be green and bright yellow if it is too ripe.

According to Endarto & Martini (2016) [6], tangerines have superior properties than Siamese oranges in the field of fruit yields. Tangerines are bigger, and fresher, apart from the skin that is easy to peel because there is a cavity between the skin and the flesh of the tangerine has a fresh sweet taste and contains a high water content. However, tangerines only produce about 200-300 fruits per tree in one year.

According to Hasibuan, Rohani, Ridwan Suprima, Rasul, & Saputra (2014) [7], hybridization is a technique of multiplying varieties or genetic diversity of plant populations. This technique increases traits and productivity by combining the genetics of crossed plants.

2.2 Citrus Quality Features

According to Lado, Rodrigo, and Zacarias (2014) [8], in citrus or citrus plants several features or characteristics that can be used as a basis for determining the quality of oranges include the color of the fruit skin and the size of the fruit. This was also agreed upon by Abouzari and Nehzad (2016) [9], where other features were also discussed, namely the absence of seeds (seedlessness) and the ease of peeling as properties that affect the quality of citrus fruit.

2.3 Clustering

Clustering / Cluster Analysis / grouping is a process of dividing a population of data into sub-sections of data called clusters. Clustering does grouping without using labels in the learning phase, this characteristic is what causes clustering to be called unsupervised learning or learning without guidance. Clustering works by making observations based on the similarity of each data, mostly using the distance between data [10]. The use of the clustering method can also shorten the pattern recognition time compared to the classification method because in some cases it is very easy to collect data but it is not easy to label large amounts of data [11].

2.4 Normalization

Data inequality can be biased or calculations using algorithms will give one column more importance than other columns, because the small value and similarity of data

distances can help create an optimal algorithm model and a more efficient training phase [12] .

Normalization often changes data values into a small range eg [-1,1] (minus one to one) or [0,1] (zero to one). One of the normalization methods is the min-maxm normalization. Min-max normalization can work well in many areas of computing and research [13]. The min max normalization notation is :

$$x_i' = \frac{x_i - min_A}{max_A - min_A} \quad (1)$$

x_i = Original value of an attribute A

x_i' = Standardized value

min_A = Minimal value of an attribute A

max_A = Maximal value of an attribute A

2.5 Heterogenous Euclidean Overlap Measures (HEOM)

One way to calculate the distance on data with mixed attributes is to use the HEOM algorithm, HEOM is a distance calculation algorithm based on the Euclidean algorithm but the distance calculation on different data features has different treatments to get the optimal distance [14]. HEOM is applied to several cases, for example in the case of distance-based ecological case classification [15] and to the optimization of the number of clustered squares [16]. HEOM uses standardized or scaled data using the min-max method. Suppose E is the HEOM value of two data points x and y, where x and y have features A as many as M and dist represents the distance between features, then E can be denoted as

$$E(x,y) = \sqrt{\sum_{A=1}^{M} dist(x_A - y_A)^2} \quad (2)$$

Where dist denoted as

$$dist(x,y) = \begin{Bmatrix} 1, if\ null \\ overlap(x,y), if\ category \\ diff(x,y)\ if\ numerical \end{Bmatrix} \quad (3)$$

And Overlap stated as

$$overlap(x,y) = \begin{Bmatrix} 0, if\ x = y \\ 1, if\ x \neq y \end{Bmatrix} \quad (4)$$

And diff stated as

$$diff(x,y) = \begin{Bmatrix} x - y, normalized \\ \frac{(x-y)}{(max(A) - min(A))}, not\ normalized \end{Bmatrix} \quad (5)$$

2.6 Density Based Spatial Clustering Application with Noise

DBSCAN is a data density-based grouping algorithm where DBSCAN will check the degree of proximity of each data to determine whether the data belongs to a group and other parameters, namely the minimum number of members to determine whether a group is a valid cluster or an outlier [10].

2.7 Silhouette Coefficient

The classification algorithm works with different levels of performance for each different case, so we need a method of evaluating the results of grouping to find out how well an algorithm handles clustering cases. Evaluation of the grouping algorithm model is divided into two methods, namely the external method and the internal method, the external method is used when the ground truth or actual label of data is used, while the internal method is used without using ground truth [10]. The internal method most often used in evaluating a clustering model is the Silhouette Coefficient or Silhouette Score [17].

How to calculate the Silhouette coefficient, namely, from a data population of D, a number of n objects, for example D has been divided into k groups with the notation C1, C2, ..., Ck. For each object o ϵ D, calculate a (o) as the average distance between object o against all other objects in the group o is in. Not much difference, b (o) is the minimum average between o and groups other than groups o are located with dist is the distance of the data discussed in section 2.5 (two points five) and | C_i | is the number of members of the cluster i. Suppose o ϵ C_i (1≤i <k), then:

$$a(o) = \frac{\sum_{o' \epsilon C_{i, o \neq o'}} dist(o,o')}{|C_i|-1} \quad (6)$$

And b as

$$b(o) = \min_{C_j : 1 \leq j \leq k, j \neq i} \left\{ \frac{\sum_{O' \epsilon C_j} dist(o,o')}{|C_j|} \right\} \quad (7)$$

So Silhouette Coefficient of an object o defined as:

$$s(o) = \frac{b(o)-a(o)}{\max\{a(o),b(o)\}} \quad (8)$$

3 Research Methods

3.1. Research data

The data in the study were obtained from the results of the measurement of tangerine and tangerine crossing data carried out at the Citrus and Subtropical Research Institute of Batu City in 2015. The amount of data obtained for this study was 600 individual citrus fruits.

3.2. Research Design

The design of the research phase starts from problem analysis used to understand and formulate problems, then literature studies, model formation and optimization using the brute force method and then a discussion of the results of the optimized model. To achieve this the research started with problem analysis which function to define the scope of research, then literature study on similar problem or topic as a base and provide deeper insight on the problem domain, optimizing and creating models to analyze the problem and offer insight, and lastly analyzing results of optimized model to achieve the goals of research and gain insights on domain problem's citrus hybridization.

4 Analysis

Analysis of the characteristics of each group formed is as follows

Group one has 96 members with P1 accession and 89 members for P5. Group one has the following properties:

- Fruit diameter and height ranging from 46 to 59 millimeters and 40 to 59 millimeters, with an average diameter of 53.6 millimeters and a height of 48.9 millimeters
- Ease of peeling in group one which is similar, namely moderate
- The number of seeds in the range of 10 seeds to 31 seeds per fruit with an average of 19 seeds per fruit
- All group members have a yellow orange color

Group two has members with P2 accessions as many as 81 members and P6 as many as 70 members. Group two has the following properties:

- Fruit diameter and height ranging from 51 to 69 millimeters and 50 to 64 millimeters, with a mean diameter of 60.4 millimeters and a height of 55.04 millimeters
- Ease of peeling in group one which is similar, namely easy to peel
- Number of seeds in the range of 11 seeds to 21 seeds per fruit with an average of 13 seeds per fruit
- All group members have an orange color

Group three has members with P3 accessions of 73 members and P4 with 42 members. Group three has the following properties:

- Fruit diameter and height ranging from 35 to 49 millimeters and 20 to 33 millimeters, with an average diameter of 42.4 millimeters and a height of 25.03 millimeters
- Ease of peeling in group one which is similar, namely moderate
- The number of seeds in the range of 14 seeds to 31 seeds per fruit with an average of 24 seeds per fruit
- All group members have a green yellow color

The results of grouping using DBSCAN which have been optimized generally show the following patterns:

1. There are three groups of quality detected by the algorithm. In general, usually the division of these three groups means that there will be three levels of quality, namely good, medium, and bad. If we look at the literature review regarding the quality features of oranges and the results of extracting each group of data, it can be seen that it is true that there are good, medium and bad quality results of crosses where the good results of crosses fall into group 2 (two), the group that is included in the group 1 (one) and the worst group into group 3 (three). This quality division can be used as a cross reference for the Citrus and Subtropical Crops Research Institute (Balitjestro) of Batu city where plants with accessions that fall into group two are prioritized as superior crosses or types of tangerines and Siamese oranges

that can be further researched. continue.

2. Based on the group division accessions are group 1 (one) which is dominated by accessions P1 and P6, group 2 (two) which is dominated by accessions P2 and P5, while group 3 (three) is dominated by accessions P3 and P4. In accordance with point 1 (one), this shows that citrus plants with crosses coded P2 and P5 are crosses that can be used to produce superior products or crosses that can be investigated further.

3. Based on the quality of the groups, the order of the clusters from the best to the worst are 2 (two), 1 (one), and 3 (three).

4. The conclusion of point 3 (three) is based on the results of analysis of each cluster which shows the range of data and the mean of each attribute that affects the quality of oranges.

5. Based on the analysis of the results of grouping the best crosses are citrus plants with P2 and P5 accessions, which shows a cross between Siamese Honey and Satsuma (P2) and a cross between Pontianak Siamese and Soe Tangerines.

5 Conclusion

Based on the range of assessments using the Silhouette Coefficient metric, the Density Based Spatial Clustering Application with Noise (DBSCAN) algorithm has a fairly good value, reaching a value of 0.6903 which has been discussed in chapter two, the value range for the Silhouette Coefficient value is {-1.1}. Compared to the Kmeans evaluation results which were only able to reach a value of 0.57. The difference in general from each group is the mean and distance of each input feature used as a quality determinant. Group two is the group with the best quality with accessions P2 and P6, followed by group one with accessions P1 and P5, and the last. is group three with accessions P3 and P4.

References

1. Litbang. (2019). Badan Litbang Pertanian. Retrieved February 28, 2020, from http://www.litbang.pertanian.go.id/info-teknologi/3455/
2. Suyanto. (2017). Data Mining untuk Klasifikasi dan Klasterisasi Data (1st ed.). Bandung: Penerbit Informatika.
3. Ester, M., Kriegel, H.-P., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Atlantic: AAAI Press.
4. Alfiyatin, A., Mahmudy, W., & Anggodo, Y. (2018). K-Means Clustering and Genetic Algorithm to Solve Vehicle Routing Problem with Time Windows Problem. Indonesian Journal of Electrical Engineering and Computer Science, 11(2), 462-468.
5. Auliya, Y., Mahmudy, W., & Sudarto. (2019). Improve Hybrid Particle Swarm Optimization and K-Means for Clustering. Journal of Information Technology and Computer Science, 4(1), 42-56.
6. Endarto, O., & Martini, E. (2016). Budidaya Jeruk Sehat. Bogor: Balai Penelitian Tanaman Jeruk dan Subtropika (Balitjestro).
7. Hasibuan, H., S, F. R., Suprima, R., Rasul, M., & Saputra, B. (2014). Persilangan (Hibridasi). Jurnal Praktikum Pemuliaan Hibrida, I(2), 1-6.
8. Lado, J., Rodrigo, M. J., & Zacarias, L. (2014). Maturity Indicators and citrus fruit quality. Stewart Postharvest Review, 2(2), 1-6.
9. Abouzari, A., & Nezad, N. M. (2016). The Investigation of Citrus Fruit Quality Popular Characteristic and Breeding. Acta Universitatis Agriculturae Et Silviculturae Mendelianae Brunensi, 64(3), 726 - 740.
10. Kaufmann, M. (2012). Data Mining Concepts and Techniques (3rd ed.). Waltham: Elsevier.

11. Duda, R. O., Hart, P. E., & Stork, D. G. (2000). Pattern Classification (2nd ed.). New York: Wiley-Interscience.
12. Mohammad, I., & Usman, D. (2013). Standardization and Its Effects on K-Means Clustering Algorithm. Research Journal of Applied Sciences, Engineering and Technology, 6(17), 3299-3303.
13. Patro, S. K., & Sahu, K. K. (2015, April 01). ResearchGate. Retrieved 03 18, 2020, from https://www.researchgate.net/publication/274012376_Normalization_A_Preprocessing_Stage
14. Wilson, D. R., & Martinez, T. R. (1997). Improved Heterogeneous Distance Functions. Journal of Artificial Intelligence Research, 6(1), 1-34.
15. Spencer, M. R., Prins, S. C., & Beckom, M. S. (2010). Heterogeneous Distance Measures and Nearest-Neighbor Classification in an Ecological Setting. Missouri Journals of Mathematic and Science, 22(2), 108-123.
16. Kettleborough, G., & Rayward-Smith, V. (2013). Optimising sum-of-squares measures for clustering multisets defined over a metric space. Discrete Applied Mathematics, 161(16-17), 2499-2513.
17. Cady, F. (2017). The Data Science Handbook. New Jersey: John Wiley & Sons, Inc.
18. Majumdar, J., Naraseeyappa, S. & Ankalaki, S., 2017. Analysis of agriculture data using data. Journal of Big Data, IV(20), pp. 1-15.
19. Zhang, M., 2019. Use Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Algorithm to Identify Galaxy Cluster Members. Guangzhou, IOP Publishing.