Ensemble ROCK Methods and Ensemble SWFM Methods for Clustering of Cross Citrus Accessions Based on Mixed Numerical and Categorical Dataset

Alvionita¹*, Sutikno¹, and A Suharsono¹

¹Department of Statistics, Sepuluh Nopember Institute of Technology, Indonesia.

Email: alvionita110293@gmail.com

Abstract. Cluster analysis is a technique in multivariate analysis methods that reduces classifying data. This analysis has the main purpose to classify the objects of observation into groups based on characteristics. In the process, a cluster analysis is not only used for numerical data or categorical data but also developed for mixed data. There are several methods in analyzing the mixed data as ensemble methods and methods Similarity Weight and Filter Methods (SWFM). There is a lot of research on these methods, but the study did not compare the performance given by both of these methods. Therefore, this paper will be compared the performance between the clustering ensemble ROCK methods and ensemble SWFM methods. These methods will be used in clustering cross citrus accessions based on the characteristics of fruit and leaves that involve variables that are a mixture of numerical and categorical. Clustering methods with the best performance determined by looking at the ratio of standard deviation values within groups (S_w) with a standard deviation between groups (S_B). Methods with the best performance has the smallest ratio. From the result, we get that the performance of ensemble ROCK methods is better than ensemble SWFM methods.

1. Introduction
Cluster analysis is a technique in multivariate analysis that reduces classifying data. This analysis has main purpose to classify objects into groups based on the characteristics possessed by the obserativos. The analysis will classify objects so that each object with high degree of similarity to other objects are in the same group [1]. Problems encountered in many case is when the dataset contains mixed scale variables, i.e. it contains some variable with numerical scale while the other are categorical. In its development, analytical methods which are often performed for scale data clustering mixture, i.e. to transform categorical data into numeric data and vice versa. However, this analysis has weaknesses that led to the difficulty in determining the exact transformation so as not to lose a lot of information from the original data.

Apart from the clustering with the transformation method, it developed a ensemble method for clustering mixed data [2]. The ensemble clustering (clustering ensemble) is a clustering technique to combine the results of clustering several clustering algorithms to get a better group [2]. One approach that can be used is a ROCK (Robust Clustering using links) method. ROCK method is developed by [3]. ROCK method uses the concept of distance to measure the similarity / proximity between a pair of data points [3].
To overcome the clustering on mixed data can also be used clustering method based on the similarity ensemble of weight and filter method (SWFM). SWFM method was first developed by [4]. This method is an outgrowth of clustering ensemble has a difference in the final stages of cluster formation. There are two main concepts in this clustering is similarity weight and filter method [4].

In this study, the analysis was done comparing performance of clustering ensembles ROCK and ensemble SWFM method. The analysis in this study begins by separating the two types of data are then performed clustering independently. Clustering numerical data was calculated using agglomerative hierarchical cluster, while categorical data are grouped using methods ROCK (Robust Clustering using links). The criterion of better performance for the clustering is measure by the ratio of standard deviation values within groups ($S_W$) with a standard deviation between groups ($S_B$). The better the method, the smaller the ratio [5]. The purpose of this study is to compare performance of the ensemble ROCK methods with ensemble SWFM methods using a case study on citrus accessions clustering protoplasm fusion product.

2. Literature of Cluster Analysis for Mixed Numerical and Categorical Dataset
This section will explain the theories used in clustering to complete a case with a mixed numerical and categorical dataset. Cluster analysis for mixed data can’t be done directly, but in its analysis requires a stage ensemble. Ensemble method can be done in several approaches, including ensemble Robust Clustering using Links (ROCK) method and Similarity Weight and Filter Methods (SWFM) method. Some of the benefits of cluster analysis is a double variable data exploration, data reduction, and predicted state of the object.

2.1. Clustering Methods
The technique used for clustering includes methods of hierarchical and non-hierarchical methods. Clustering hierarchy starts with two or more objects that have the closest similarity, then the process is passed to other objects that have the proximity of the two. The analysis was performed by the group will be forming a "tree", where there is a hierarchy (levels) which clearly between objects, from the most similar to least similar [6]. There are two techniques in the analysis group clustering hierarchy, namely division techniques (divisive) and merging techniques (agglomerative). The advantages of using hierarchical method in the analysis group is speeding up processing and save time as input data will form a hierarchy or precipitated separately, making it easier to interpretation, but the weakness of this method is often contained errors in the data outliers, the difference measure of distance used, and the presence of variables that are not relevant.

2.2. Cluster Analysis for Mixed Numerical and Categorical Dataset
Cluster analysis for mixed data is done by ensemble clustering. In the ensemble method, the final formation of the cluster is done using clustering algorithm that similar with clustering method for categorical data. In addition, other method is developed for clustering mixed data, namely ensemble SWFM.

2.2.1 Ensemble Method
Ensemble Clustering is a method that combines several different algorithms to obtain a general partition of the data, which is aimed at the consolidation of the portfolios of individual clustering. Interest ensemble clustering is to combine the results of several clustering algorithms to get the better clustering and robust. Ensemble clustering consists of two stages of the algorithm. The first stage is the clustering with some algorithm and storing the results of these clustering. Second, use the function to determine the final cluster of groups of the first stage results [3].

Steps in the analysis of mixed data using a method called ensemble clustering algorithm CEBMDC has the following stages, [3]

a. Gathering data consisting of numerical and categorical variables, divided into two sub-data, namely pure numerical and pure categorical. Suppose that there are $m$ variables, with $m_{\text{numerical}}$
represent the number of numerical variables, and \( m_{\text{categorical}} \) represent the number of categorical variables, in order to obtain \( m = m_{\text{numerical}} + m_{\text{categorical}} \). Furthermore, clustering is done by clustering algorithm according to the type of data separately.

b. Clustering objects with numeric variables with numeric data clustering algorithm, and clustering objects with categorical variables with categorical data clustering algorithm.

c. Combining (combining) the results of the clustering of numerical and categorical variables, a process called ensemble.

d. Clustering ensemble using categorical data clustering algorithm to obtain the final group (final cluster).

2.2.2 Similarity Weight and Filter Method (SWFM)

SWFM clustering has a concept similar analysis by clustering ensemble in general. This method is an outgrowth of the clustering ensemble that has a difference algorithm to get the final stages of cluster formation. In the ensemble method, the final formation of the cluster is done using clustering algorithm similar to clustering on categorical data, while the ensemble SWFM using an algorithm based on the similarity weight and filter method. At the stage of weight similarity method used similarity measure that would include a weighting factor in the similarity formula.

The steps in clustering by SWFM are,

a. Dividing (splitting) that the original data,

b. Clustering accession by numerical variables using numeric data clustering algorithm,

c. Clustering accession by categorical variables using categorical data clustering algorithm,

d. Merging clustering results into the final cluster (stage ensembles). After getting the optimum group of rock and agglomerative method results, the next step is merging the group. At this stage, clustering is done by using the distance in equation (1), \[4\]

\[
sim(X_i, X_j) = \sum_{i \neq j} \frac{S_{ij}}{\max(n_i, n_j)}, \quad i \neq j
\]

where \( S_{ij} = \left| \frac{X_i \cap X_j}{X_i \cup X_j} \right| \)

The input for this stage are group of agglomerative method results (output 1) and group of ROCK methods results (output 2). Output 1 and output 2 is considered as categorical variables are used to compile the final cluster using filter algorithm in equation (2). \[4\]

\[
F(x_i, x_j) = \sum_{i \neq j} \sum_{w_{ij}} w_{ij} d(x_i, x_j)
\]

with \( w_{ij} \) is the weight between the groups to i and j.

2.3. Cluster Analysis for Numerical Dataset

Clustering for numerical data is done with the assistance of a distance, a distance that is often used is the euclidean distance. The concept of the euclidean distance measuring distances between observation \( x_i \) and \( x_j \), expressed in equation (3) as follows,[1]

\[
d_{ij} = \sqrt{(x_j - x_i)^T (x_j - x_i)}
\]

If the value of dissimilarity between two objects is greater, then the difference between the two objects will be even greater, so the two objects are increasingly not exist in the same group. Some techniques clustering between the groups stated as follows,[1]

a. Single linkage, this procedure is based on the smallest distance.

\[
d_w(u,v) = \min (d_{uw}, d_{wv})
\]
b. Complete linkage, this procedure is based on the greatest distance.
\[ d_{w(u,v)} = \max (d_{wu}, d_{wv}) \] (5)
c. Average linkage, this procedure is based on the average distance
\[ d_{w(u,v)} = \frac{n_u}{n_u + n_v} d_{wu} + \frac{n_v}{n_u + n_v} d_{wv} \] (6)

2.4. Cluster Analysis for Categorical Dataset

One algorithm for clustering categorical data is ROCK Method. ROCK methods developed from agglomerative hierarchical clustering method used for categorical data with a new concept that is links to measure similarity/proximity between a pair of data points. Observations that have great links will be combined into one group, while the other has a small links will be separated from the group. The steps undertaken in ROCK method is as follows, [3]

a. Initializing the object as a group with a single member,
b. Forming similarity between objects to the criteria using equation (7),
\[ \text{sim}(X_i, X_j) = \frac{|X_i \cap X_j|}{|X_i \cup X_j|}, \quad i \neq j \] (7)
where \( |X_i \cap X_j| \) is the number of intersection between \( X_i \) and \( X_j \), and \( |X_i \cup X_j| \) is the number of union between \( X_i \) and \( X_j \).
c. Determining the threshold (\( \theta \)) in the adjacency matrix, the threshold value (\( \theta \)) is a parameter defined by the user that can be used to control how close the relationship between objects. The value of the input \( \theta \) is \( 0 < \theta < 1 \),
d. Calculating the value of a link between the observations,
e. Calculating the local heap goodness measure value using equation (8),
\[ g(C_i, C_j) = \frac{\text{link} \left[ C_i, C_j \right]}{\left( n_i + n_j \right)^{1+2f(\theta)}} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)} \] (8)
where \( \text{link} \left[ C_i, C_j \right] = \sum_{X_i \neq X_j \neq X_k} \text{link}(X_i, X_j) \) stating the number of links from all possible pairs of objects that exist in \( C_i \) and \( C_j \), \( n_i \) and \( n_j \) each state the number of members in the group to \( i \) and \( j \), while \( f(\theta) = \frac{1-\theta}{1+\theta} \).
f. Determining the heap global value is the maximum goodness measure between columns in the ith row,
g. Repeat rare (e) and (f) in order to obtain the maximum value in the global heap and local heap,
h. During the data size> k, where k is the number of classes specified then did the merger group has a the largest value of local heap with the largest global heap into one group, and then add the link between the two groups are combined, then remove the group that emerged from the local heap and update values global heap with the result of merging,
i. Perform step (h) in order to obtain the expected number of groups or there are not any link between groups,
j. Repeating steps (a) through (i) the value of \( \theta \) is different,
k. Calculate the ratio between \( S_W \) and \( S_B \) for each value of \( \theta \),
l. Comparing the results of step (k) for each value of \( \theta \) and determine the optimum number of groups with a criteria the smallest ratio between \( S_W \) and \( S_B \).
2.5. Validation Clustering
Determining the optimum number of groups is an important step after the clustering process. This stage is referred to as validation clustering [7]. In addition determining the optimum number of clusters, cluster validation phase can also measure the performance results of the grouping. Good group is to have a high homogeneity between members in the group and the high heterogeneity between groups [6]. Performance results for the variable clustering with scale numerical data can be determined from the ratio of the value of the standard deviation in groups or within ($S_W$) and the standard deviation between groups or between ($S_B$). The performance of a method of clustering would be better if the smaller the ratio between $S_W$ and $S_B$, which means that there is a maximum homogeneity in the group and maximum heterogeneity between groups [5].

By using the average value of the variable, $S_W$ and $S_B$ can be formulated as in equation (9) and (10) [5].

$$S_W = \frac{1}{C} \sum_{c=1}^{C} S_c$$

(9)

with $S_c$ is a standard deviation of all groups C and C is the number of groups formed.

$$S_B = \left[ \frac{1}{C-1} \sum_{c=1}^{C} (\bar{X}_c - \bar{X})^2 \right]^{1/2}$$

(10)

with $\bar{X}_c$ is the average of all groups C and was average overall group. That performance measurement ratios $S_W$ and $S_B$ are only used for numeric data, while for categorical data is to use contingency tables which are equivalent to doing ANOVA (Analysis of Variance). Size diversity for categorical data one developed by [8]. So that the standard deviation within the group ($S_W$) and the standard deviation between groups ($S_B$) for categorical data can be written as in equation (11) and (12). [8]

$$S_W = \left[ \frac{WSS}{(n-c)} \right]^{1/2}$$

(11)

$$S_B = \left[ \frac{BSS}{(c-1)} \right]^{1/2}$$

(12)

with $WSS = \frac{n}{2} - \frac{1}{2} \sum_{c=1}^{C} \frac{1}{n_c} \sum_{k=1}^{k} n_{kc}^2$ and $BSS = \frac{1}{2} \left( \sum_{i=1}^{C} \frac{1}{n_i} \sum_{k=1}^{k} n_{ik}^2 \right) - \frac{1}{2n} \sum_{i=1}^{C} n_{ik}^2$.

3. Case Study
Citrus plant breeding with protoplasm fusion between two kinds of oranges has been started since 2006 in Indonesian Citrus and Subtropical Fruits Research Institute (ICSFRI) in cooperation with the Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development (ICABIOTEC). Expected crop yield is superior seedless orange with a sweet taste, skin easy to peel and has an attractive color [9]. Protoplast fusion is one of the techniques that use-right in the activities of crossbred plants by combining two different plant cells, so as to obtain hybrid plants. Plant cell that have incorporated parts of the cell wall is removed so that the genetic material contained in it is only protected by a plasma membrane [9]. Thus, it is possible the merger of the two parent protoplast, so it cans produce new individuals are more diverse than the result of a combination of characteristics of both parents.
Results of protoplasm fusion is a new of plants. The new plants thus obtained are then observed by citrus fruit produced. However, due to the incorporation of protoplasts can occur involving two or more protoplasts, then the newly generated individual diversity is very high. The results of breeding is 120 plants hereinafter referred accession, but there are only 25 of 120 accessions which can be observed due to the accession of 95 more unfruitful mature or crop failure. The number of accessions affect the selection process more difficult due to an accession have different characteristics, although there is also an accession that have the same characteristics.

Citrus fruit morphological characteristics have two types of variables, the scale variables categorical and numerical scale variables. Citrus fruit morphology numeric variables include variables that can be measured numerically like fruit diameter, fruit height, weight of the fruit, and others. While categorical variables include fruit morphological variables that can’t be measured numerically and can only be given a score such as skin color, fruit shape, color pulp and others.

The observed characteristics can’t be done partially analysis between numerical and categorical variables. It is because of these characteristics affect each other and come from the same research object, so the analysis should be performed multivariate and involving all good numerical and categorical variables simultaneously. If these characteristics are analyzed partially it will give maximum results. The results are obtained if the analysis is partially inconsistencies existing group members in the group that is formed, so that the occurrence of multiple groups of accession the same, so it can’t be done definite clustering of the accessions were observed. One method of multivariate analysis can be used for clustering is a cluster analysis. The data used in this research is secondary data obtained from the observed data at Indonesian Citrus and Subtropical Fruits Research Institute (ICSFRI). Here are the variable observations for analysis data are presented in Table 1.

| Numerical Variable     | Categorical Variable         |
|------------------------|------------------------------|
| Fruit Diameter         | (Diameter)                   |
| Insensitivity          | (Insensitivity)              |
| Total Segment          | (Segment)                    |
| Thick Skin Meat        | (Thick)                      |
| Diameter Axis          | (D. Axis)                    |
| Number Of Seeds        | (Seed)                       |
| Weight Fruit           | (Weight)                     |
| Brix                   | (Brix)                       |
| Fruit Shape            | (Shape)                      |
| Skin Color             | (Color)                      |
| The Surface Of The Skin| (Surface)                    |
| Fruit Axis             | (Axis)                       |
| Flesh Color Of Skin    | (Flesh C.)                   |
| Looked Axis            | (Looked A.)                  |
| Texture Of The Pulp    | (Pulp)                       |

4. Results and Discussion
The first step in this analysis is to separate the data into two groups. The first group is a group of numeric data and the second group is a group of categorical data. For numeric data, we get a breakdown into five groups. From some consideration for the number of clusters, we chose the clustering results that have outcomes that are grouped into five groups that have R-square value around 90%. This value indicates that the clustering is good enough and able to represent the whole observation. By looking at the ratio between the standard deviation value within a group with the standard deviation between groups, for clustering numeric data used single linkage methods that has a smaller ratio than average linkage and complete linkage. Ratio value that is formed can be seen in Table 2.

Based on the explanation above, it can be concluded that the analysis of numerical data is done by agglomerative hierarchical clustering methods with single linkage and the number of clusters formed as 5. Members of clusters formed are presented in Table 3. For categorical variable data is grouped using ROCK methods, with theta trial had values between 0.05 to 0.30. From the theta trial, we get
The largest number of the clusters is 3 clusters. The value of $\theta$ that produces the number of clusters as much as 3 are 0.15, 0.17, and 0.20, so that these groupings of 3 results calculated ratio value between $S_w$ and $S_B$. Ratio value that is formed can be seen in Table 4. Based on the explanation above, it can be concluded that the analysis for categorical data was conducted by ROCK with a value of $\theta$ at 0.15 and the number of clusters formed as 3. Members of the clusters that are formed can be seen in Table 5.

**Table 2.** Summary ratio value of single linkage method, complete linkage and average linkage with the number of clusters as much as 5 clusters.

| Method           | $S_W$   | $S_B$   | Ratio   |
|------------------|---------|---------|---------|
| Single linkage   | 0.681356| 1.764561| 0.386134|
| Complete linkage | 0.955187| 1.427995| 0.668901|
| Average linkage  | 1.029224| 1.361048| 0.756200|

**Table 3.** Cluster membership of the clustering on numerical data using hierarchical method agglomerative using single linkage with 5 clusters.

| Cluster | 1    | 2    | 3    | 4    | 5    |
|---------|------|------|------|------|------|
| FS01    | FS70 | FS07 | FS14 | FS16 | FS64 |
| FS10    | FS73 | FS22 | FS66 | FS18 |
| FS15    | FS89 | FS29 | FS32 |
| FS20    | FS103| FS56 |
| FS57    | P III| FS84 |
| FS68    |      |      | TL 3 |
|         |      |      | TL 5 |

Note: FS01, FS10, and the others are name code for the new citrus fruit.

**Table 4.** The ratio value of ROCK method for clustering categorical data with the number of clusters as much as 3 clusters.

| $\theta$ | WSS    | BSS    | $S_W$   | $S_B$   | Ratio   |
|----------|--------|--------|---------|---------|---------|
| 0.15     | 12.5263| 16.5137| 0.754571| 2.873474| 0.262599|
| 0.17     | 15.4673| 2.2927 | 0.838486| 1.070677| 0.783136|
| 0.20     | 20.8891| 6.8709 | 0.974425| 1.853497| 0.525723|

After obtaining optimum group of ROCK and agglomerative methods results, the next step is to merge the group. In the ensemble ROCK methods, this step just like clustering of numeric data using ROCK methods, wherein the input at this stage is the result of agglomerative hierarchical clustering methods as Output 1 and results clustering ROCK methods as Output 2. Output 1 and output 2 were expressed as categorical variables that will be used to draw up the final cluster using ROCK method. The optimum number of clusters is the number of groups that have smallest ratio between $S_w$ and $S_B$. Clustering is done with theta trial had values between 0.05 to 0.30. Ratio value that is formed can be seen in Table 6. Based on the ratio values in Table 6, it is known that optimum clustering done with a value of $\theta$ at 0.12.
Table 5. Cluster membership of the clustering on categorical data using ROCK method using $\theta = 0.15$ with 3 clusters.

| Cluster | 1         | 2          | 3         |
|---------|-----------|------------|-----------|
| FS07    | FS29      | FS01       | FS66      |
| FS14    | FS64      | FS10       | FS84      | FS73      |
| FS15    | FS68      | FS16       | FS89      |
| FS18    | FS107     | FS22       | FS103     |
| FS20    | TL5       | FS32       | P III i   |
|         | FS56      | TL3        |

Table 6. The ratio value of ensemble ROCK method for clustering categorical data with the number of clusters as much as 4 clusters.

| WSS   | BSS    | $S_W$    | $S_B$    | Ratio  |
|-------|--------|----------|----------|--------|
| 0.05  | 7.5090 | 13.8510  | 0.5842   | 2.6316 | 0.2220 |
| 0.07  | 4.3144 | 13.6856  | 0.4428   | 2.6159 | 0.1693 |
| 0.10  | 9.1380 | 26.6220  | 0.6445   | 3.6484 | 0.1766 |
| 0.12  | 4.4870 | 21.5130  | 0.4516   | 3.2797 | 0.1377 |
| 0.15  | 10.3820 | 25.6180  | 0.6870   | 3.5790 | 0.1919 |
| 0.17  | 10.3820 | 25.6180  | 0.6870   | 3.5790 | 0.1919 |
| 0.20  | 10.3820 | 25.6180  | 0.6870   | 3.5790 | 0.1919 |
| 0.22  | 11.2690 | 26.1710  | 0.7157   | 3.6174 | 0.1979 |
| 0.25  | 11.2690 | 26.1710  | 0.7157   | 3.6174 | 0.1979 |
| 0.27  | 11.2690 | 26.1710  | 0.7157   | 3.6174 | 0.1979 |
| 0.30  | 11.2690 | 26.1710  | 0.7157   | 3.6174 | 0.1979 |

For the ensemble SWFM methods, input to this step as the ensemble ROCK methods but these steps for clustering is used is the distance in equation (1) and filter algorithm in equation (2). Clustering is done by forming groups of 3 to 6, and then to determine the optimum number of clusters used ratio value between $S_W$ and $S_B$. Ratio value that is formed can be seen in Table 7.

Table 7. Summary ratio value of ensemble SWFM method.

| WSS   | BSS    | $S_W$    | $S_B$    | Ratio  |
|-------|--------|----------|----------|--------|
| k=3   | 14.7046 | 1.1354   | 0.817552 | 0.753459 | 1.085065 |
| k=4   | 16.9003 | 6.1397   | 0.876468 | 1.752099 | 0.500239 |
| k=5   | 17.2272 | 24.5328  | 0.884904 | 3.502342 | 0.252661 |
| k=6   | 34.9443 | 37.6157  | 1.260308 | 4.336802 | 0.290608 |

Results obtained from the final clustering of the ensemble ROCK methods is clustering the 25 observations into four groups with the ratio is 0.1377, while the results obtained from the clustering
ensemble SWFM methods is clustering the 25 observations into five groups with the ratio is 0.2526. Members of the clusters that are formed from these methods can be seen in Table 8 and Table 9. However, the results revealed from ensemble ROCK method that there is a grouping formed one’s own observations into clusters so must do more observations on those observations.

| Table 8. Cluster membership of the clustering on mixed data using ensemble ROCK method with 4 clusters. |
|--------------------------------------------------|
| Cluster 1 | 2 | 3 | 4 |
| FS14      | FS84 | FS07 | FS16 | FS70 | FS01 |
| FS22      | FS107 | FS18 | FS103 | FS10 |
| FS29      | TL 5 | FS20 | P III i | FS15 |
| FS32      | FS68 | FS57 | TL 3 | FS66 |
| FS56      | FS64 | FS89 |
| FS73      |

| Table 9. Cluster membership of the clustering on mixed data using ensemble SWFM method with 5 clusters. |
|--------------------------------------------------|
| Cluster 1 | 2 | 3 | 4 | 5 |
| FS01      | FS07 | FS16 | FS56 | FS57 |
| FS10      | FS14 | FS18 | FS84 | FS70 |
| FS20      | FS15 | FS22 | FS107 | FS89 |
| FS68      | FS32 | FS29 |
| FS103     | FS66 | FS64 |
| P III i   | FS73 | TL 3 |
| TL 5      |

5. Conclusion
The value of the ratio from ensemble ROCK method is smaller than ensemble SWFM method, which shows that the performance ensemble ROCK method is better than ensemble SWFM method. From the research conducted, it can be stated that ensemble ROCK method is more suitable for resolving existing case studies.

Acknowledgments
On this occasion, the author would like to thank to Indonesian Citrus and Subtropical Fruits Research Institute (ICSFRI) which has been granted permission to use their data as a case study. In addition, the authors also thank to other parties who contributed to the completion of this analysis.

References
[1] Johnson R A and Winchern D W 2007 Applied Multivariate Statistical Analysis (sixth ed.) (New Jersey: Pearson Education Inc)
[2] He Z, Xu X, and Deng S 2005 CoRR abs/cs/0509011
[3] Guha S, Rastogi R, and Shim K 2000 ROCK: A Robust Clustering Algorithm for Categorical Attributes Proceedings of the 15th International Conference on Data Engineering
[4] Reddy M V J and Kavitha B 2012 Clustering the Mixed Numerical and Categorical Dataset using Similarity Weight and Filter Method International Journal of Database Theory and Application 5 1
[5] Bunkers M J and James R M 1996 J.Climate 130-146
[6] Hair J F, Black W C, Babin J B, and Anderson E R 2009 Multivariate Data Analysis (seventh ed.) (New Jersey: Prentice Hall Inc)
[7] Halkidi M, Batistakis Y, and Vazirgiannis M 2001 Journal of Intelligent Information Systems 17 2/3 107-145
[8] Kader D G and Perry M 2007 Journal of Statistics Education 15 2
[9] Martasari C 2014 Kajian Genetik dan Percepatan Pembungaan Tanaman Hasil Fusi Protopena Jeruk Siam Madu dan Satsuma Mandarin [Dissertation] (Malang: Universitas Brawijaya Malang)