Research Journal of Applied Sciences 14 (4): 110-116, 2019
ISSN: 1815-932X

Implementation of k-Means Clustering Algorithm based on Rice Productivity Level in Subdistrict Area (Case Study: In Special Region of Yogyakarta)

Indrawan Risangaji, Muhammad Nasrun and Anggunmeka Luhur Prasasti
Major of Computer System, Faculty of Electrical Engineering, Telkom University,
Jl. Telekomunikasi No. 1, 40257 Bandung, Indonesia
risangaji7@gmail.co.id

Abstract: The need for people to consume rice continues to increase which causes the price of rice in the market to rise and fall in the Yogyakarta Special Region while the government still lacks attention to rice production in each year. This research focuses on making regional groupings in Yogyakarta Special Region based on the level of rice productivity. Data mining process is done by clustering using k-means to classify sub-districts based on production data and land area. Evaluation of the results of clusters uses elbow method and comparison of cluster results with other programs. Regional grouping based on the level of rice productivity produces 3 optimal clusters which are divided into low productivity, medium productivity and high productivity. With the results of this study, it is expected to help the Yogyakarta agricultural service in an effort to increase rice productivity more evenly in each region.

Key words: Clustering, regional grouping, k-means, rice productivity, data mining process, elbow method

INTRODUCTI ON

Food is the basic human need. According to UU RI No. 7 year 1996 on food, it states that food is the human rights for each individual in Indonesia. Nature achieves high quality human resource to build national development.

In Indonesia, food is identical with rice because almost all or most of Indonesian people consume rice as staple foodstuffs and main carbohydrate source. Rice also becomes staple food stuffs of people of most countries in Asia and even most of people in the world.

Food sustainability for Indonesia is closely related to the sufficiency of rice availability. With the increased national development, especially, in fulfilling the food needs, so, the demand food stuffs of people of most countries in Asia and even most of people in the world. Food sustainability for Indonesia is closely related to the sufficiency of rice availability. With the increased national development, especially, in fulfilling the food needs, so, the demand of foodstuffs is also increased, considering a great natural resource in agricultural sector, so, in the future this sector will still become an important sector in giving contribution in national economic growth.

Central Statistics Agency (BPS) reported that DI Yogyakarta is one of best producing regions in national scale after Nusa Tenggara Barat. As one of national rice barn, Special Region of Yogyakarta is expected to be able to keep and increase the production of rice every year, so, it can counterbalance a large number of people that keep increasing.

To meet rice need, Agricultural Agency of Special Region of Yogyakarta tries to keep optimizing result of rice production. It needs a method to make grouping on harvest result based on land area of rice field and harvest production of every sub district. The aim is to know the region with the result of highest productivity of rice until the lowest. So, the government can control the region with high productivity of rice and increase the region with low productivity of rice by getting the attention and effective handling because it is related to policy making of help distribution which isconducted by Agricultural Agency of Special Region of Yogyakarta.

Based on that explanation, the researcher aimed to use the data from BPS DI Yogyakarta to make grouping of rice productivity and categorize it in order to ease the analysis of agriculture product and know the performance of that method. Parameter of comparator that is used is test of variable of k value to evaluate how good the algorithm is in predicting grade of instance after the training is conducted. The best result of k will be used to validate program that is made and used for clustering all data that will be used in grouping region of rice productivity in every year.

Corresponding Author: Indrawan Risangaji, Major of Computer System, Faculty of Electrical Engineering, Telkom University, Jl. Telekomunikasi No. 1, 40257 Bandung, Indonesia, risangaji7@gmail.co.id

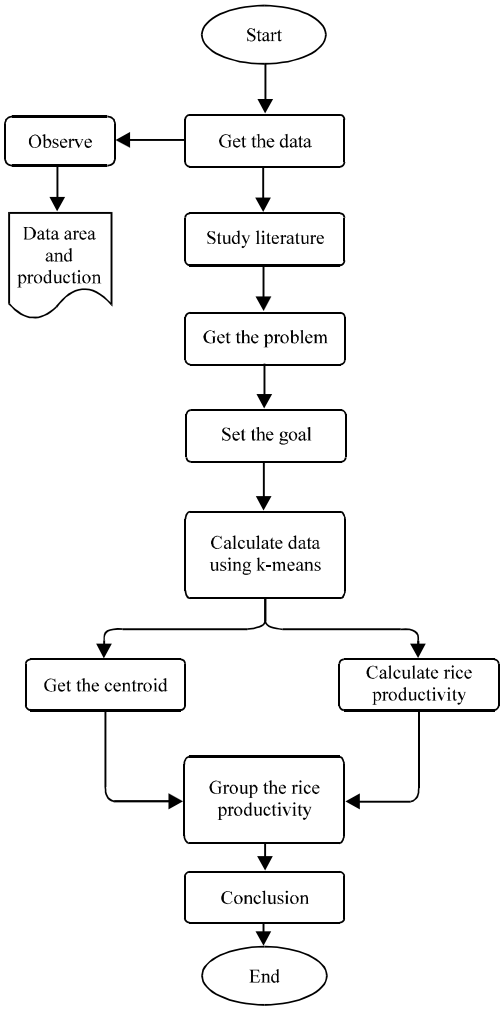

Fig. 1: Research structure

MATERIALS AND METHODS

The flow of this study follows the framework of study in Fig. 1. The study was conducted in Agriculture Agency of Yogyakarta. The study was begun by doing analysis and interview to the staff in agriculture agency, then suggested by agriculture agency to find the data in local BPS and it obtained the data of production and area of rice field from 2006 until 2016. This analysis obtained data of rice productivity in every sub district of Yogyakarta.

The study was continued by formulating problems in order this study is guided and has problem that obtain goal. The goal is that in order the process of subdistrict grouping uses algorithm of k-means based on level of rice productivity in Special Region of Yogyakarta from 2006 until 2016. Then the result of grouping of algorithm k-means was analyzed by using elbow method, program comparison and determining which regions included in the level of rice productivity.

Machine learning: As we know, machine learning is the main point of big data. By applying machine learning the patterns of saved data will be shown. Those patterns are actually the expected result by applying big data.

Algorithm k-means is included in algorithm type used for clustering. Clustering is categorizing data into some groups or clusters. Algorithm which is related to machine learning, there is one more type of algorithm whose function is almost the same with clustering namely classification (Bholowalia and Kumar, 2014).

Clustering and classification are different. Classification is supervised learning while clustering is unsupervised learning (Feldman and Sanger, 2006; Han *et al*., 2012). The example of classification is grouping the email. As we know, email provider such as gmail categorizes the emails by some groups such as spam, inbox and priority. We define for ourselves how the criteria for spam email, priority criteria and so on. We only need to define these criteria at the beginning or occasionally add these criteria, for example, when an important email is included as spam. Under normal conditions when there is a new email, the classification of algorithm will automatically put this email into spam or other priorities or groups.

Clustering or unsupervised learning means grouping data without user intervention (Manning *et al*., 2009). Each incoming data will be grouped according to similarity with the data that has been entered previously. Without interference, here does not mean that there is absolutely no definition before the algorithm is run. In clustering algorithm, we definefeatures that will be used as guidelines for grouping data. However, we do not manually group data into certain groups such as classification.

Clustering method: Clustering is the process of partitioning a set of data objects into subsets called clusters (Wu, 2012). Objects in the cluster have similar characteristics between one and another and they are different from other clusters. Partitions are not done manually but with a clustering algorithm. Therefore, clustering is very useful and can find groups or unknown groups in the data. Clustering is widely used in various applications such as business intelligence, image pattern recognition, web search, biological sciences and security. In business intelligence, clustering can manage many customers into many groups, for example, grouping

customers into several clusters with similarity of strong characteristics. Clustering is also known as segmentation data because clustering partition many data sets into many groups based on their similarity.

Calculation of data distance: Distance is a commonly used approach to determine the similarity or inequality of two features of vectors that are stated by ranking. If the resulted ranking value is smaller, the similarity between the two vectors is closer/higher. Distance measurement technique with Euclidean method is one of the most commonly used methods to calculate the distance between data and centroid (Srivastava and Sahami, 2009). Euclidean distance is often used in calculating distance, this is because the result obtained is the shortest distance between the two calculated points. Distance measurement by using Euclidean method can be written with the following Eq. 1:

$$Dij = \sqrt{\sum_{k=1}^{p}(Xik - Xjk)^2} \quad (1)$$

Where:
Dij = Distance of object between object i and j
p = Data dimension
Xik = Coordinate of object i in dimension of k
Xjk = Coodinate of object j in dimension of k

k-means clustering: k-means is one method of grouping non-hierarchical data, trying to partition data in the form of two or more groups. This method will partition data in groups, so that, data with the same characteristics are included in the same group and vice versa for different groups. The purpose of grouping is to minimize the objective function set in the processing, simply trying to minimize variation in a group and maximize variation between groups, the following is the k-means algorithm on the grouping of rice production data. The flow of the k-means algorithm in Fig. 2 (Handoyo *et al*., 2014). Then updating a centroid point can be done with the following Eq. 2:

$$\mu_k = \frac{1}{N_k}\sum_{q=1}^{N_k} x_q \quad (2)$$

Where:
μ_k = The centroid point of the k-cluster
N_k = The amount of data in the k-cluster
x_q = q-data in the k-cluster

Elbow method: Cluster evaluation is used is the Elbow method to produce information in determining the best number of clusters by looking at the percentage of the comparison result between the number of clusters that will form an Elbow at a point (Kodinariya and Makwana, 2013).

The percentage results which are different from each cluster values can be shown using graphs as the source of information. If the value of the first cluster with the value of the second cluster gives the angle in the graph or the value has the biggest decrease then that cluster value is the best (Bholowalia and Kumar, 2014).

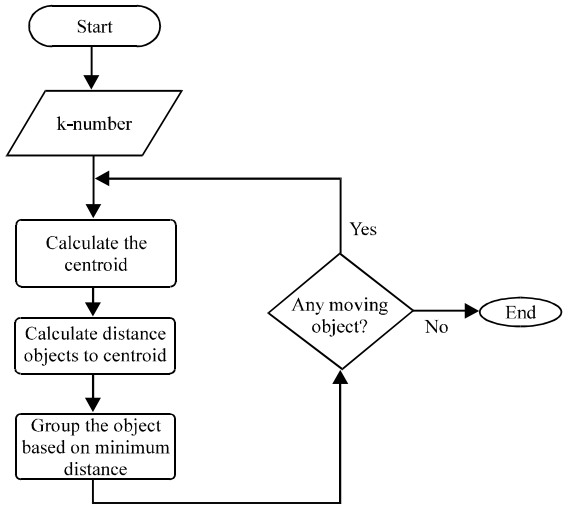

Fig. 2: Flow of k-means

Table 1: Implementation of k-means classes

Id object	Types	Details
getAssignmets	Int	To decide class or group
getCounts	Int	To find index of lowest value
getDistances	Double	To calculate the distance between the data and centroid
getPoints	Int	To take data of production and land area
getCentroids	Int	To decide first centroid

To get the comparison is to calculate SSE (Sum of Square Error) from each cluster value. Because the greater the number of cluster k, the smaller SSE value will be. SSE Eq. 3 on k-means:

$$SSE = \sum_{i=1}^{k}\sum_{p\in Ci} d(p, mi)^2 \quad (3)$$

After being seen, there will be some k-values that experience the biggest decrease and then the results of the k-value will decrease slowly until the results of the k-value are stable. For example, the value of cluster k = 2 to k = 3, then from k = 3 to k = 4, it can be seen a drastic decrease in forming an elbow at the point k = 3 then the ideal value of cluster k is k = 3.

Implementation of k-means: In implementing k-means, there are several classes that are formed namely getAssignmets, getCounts, getDistances, getPoints, getCentroids to be called. Detailed specifications of the k-means class can be seen in the following Table 1.

Centroid used as the median of each group for calculation was explained as follows.

```
...
public Kmeans calcKmeans(int
k, boolean equal, double[][]
points,
            double[][]
minmaxlens[]
        if(k==0) return null;
        double[][] centroids =
new double[k][2];
        for (int i = 0; i < k;
i++) {
            centroids[i][X] =
minmaxlens[MIN][X] +
(minmaxlens[LEN][X] / 2d);
            centroids[i][Y] =
minmaxlens[MIN][Y] +
(minmaxlens[LEN][Y] / 2d); }

AbstractEKmeans.Listener
listener = null;
        listener = new
AbstractEKmeans.Listener() {
                public void
iteration(int iteration, int
move) {
...
```

Calculation of the distance between data and centroid obtained by using Eucledian distance was explained as follows.

```
...
public static final
DoubleDistanceFunction
EUCLIDEAN_DISTANCE_FUNCTION =
new DoubleDistanceFunction() {
        public double
distance(double[] p1, double[]
p2) {double s = 0;
    for (int d = 0; d <
p1.length && d < p2.length;
d++) {s +=
Math.pow(Math.abs(p1[d] -
p2[d]), 2); }
            return
Math.sqrt(s); }
    };
...
```

Interface design of this document grouping system was designed based on the diagram design that had been described previously. The grouping system requires a system interface that can perform every function that is needed. The need for this system is divided into 3 parts. First, the need to display the data to be grouped. Second, the need to cluster data that already exists in the system. And the last, the need to see the results of clustering data in the form of tables and plots that have been done. The appearance of the interface system of data grouping of rice productivity can be seen as follows (Fig. 3):

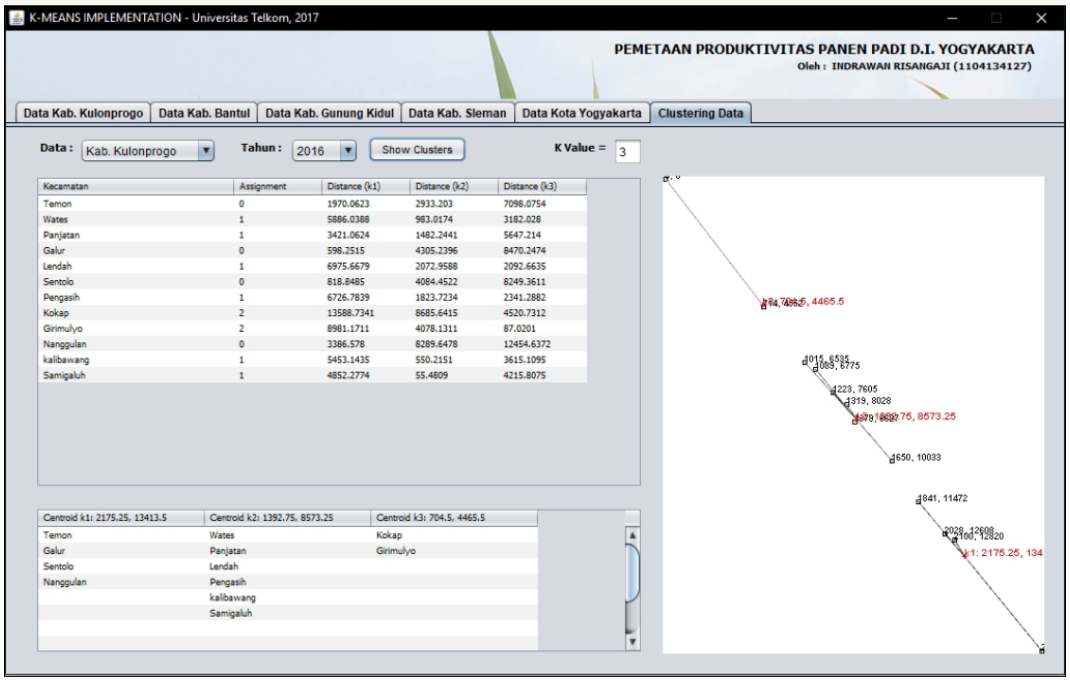

Fig. 3: Interface clustering program

RESULTS AND DISCUSSION

Results of elbow method: Based on data in 5 regencies in 2016 with the number of input k values between 2-6 clusters, the results of the sum of square error which experienced the biggest decrease occurred in Kulonprogo, Sleman, Bantul, Gunung Kidul and Yogyakarta city on k = 3 because there is a large decrease of SSE results between the value of k = 2 and k = 3, so that, elbow is seen, then for the value of k = 4 to k = 6, the SSE value tends to increase and decrease which tends to be unstable.

So, for this case the ideal number of clusters is k = 3 because the SSE results between k = 4 to k = 6 tend to be unstable which should be the larger number of clusters k, the smaller the value of SSE will be (Kodinariya and Makwana, 2013). k = 3 is used as the default cluster to determine the grouping of rice productivity areas based on data on land area and production. The best SSE results are seen in Kulonprogo Regency in Fig. 4.

Algorithm validation results: The results of the algorithm validation testing between the programs built by comparing with Weka tools resulted in the same cluster labeling, namely "C0 ","C1 " and "C3 " but the results of the centroid value were different but the members of each cluster still had relationships with the clusters generated by Weka tools (Fig. 5).

Looking at the results of the cluster in Bantul Regency, it obtained the same membership in each cluster but it is only different in the centroid location on the label. For the cluster results of Kulonprogo Regency, the members of each cluster are not all the same when compared to the cluster results of Weka tools, there are one to three members who move between clusters as well as Sleman Regency.

The cluster results of the calculation of Gunung Kidul Regency from the three clusters, there is only 1 cluster formed that has the similarity of members, 2 other clusters are different only from one to three members as well as the results of cluster calculations in the city of Yogyakarta.

Test results of program: One of the data successfully processed in Kulonprogo Regency in 2016 with centroid 0 is at 2017.25, 13413.5 number of members is 4, centroid 1 is at 1392.75, 8573.25 number of members is 6 and centroid 2 is at 704.5, 4465.5 with the number of members is 2, the distribution of data from cluster calculations can

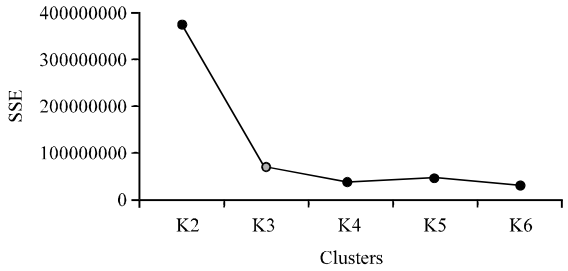

Fig. 4: Elbow method on Kulonprogo

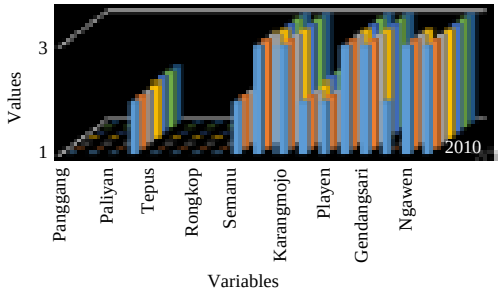

Fig. 5: Cluster result on Gunung Kidul

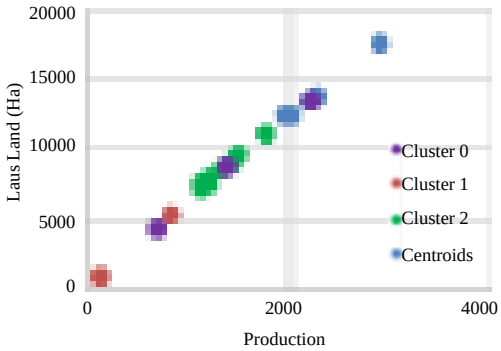

Fig. 6: Data spread Kulonprogo

be seen in Fig. 6. It mean the program can calculate and display the table and mapping of cluster results as shown in Fig. 7.

By looking at the results of the grouping of rice productivity areas from the five regencies in the Special Region of Yogyakarta that Gunung Kidul and Yogyakarta city have stable rice productivity while Kulonprogo, Sleman and Bantul are areas of unstable rice productivity in the previous few years (Table 2 and 3).

The results of data testing from 2006-2016 were obtained as follows with the example in Gunung Kidul Regency with cluster results that are clearly enough in illustrating the low, medium and high level of rice productivity in Fig. 5 and Table 4.

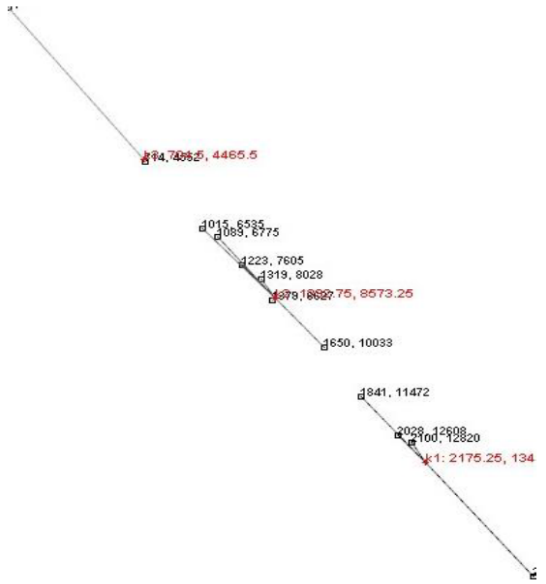

Fig. 7: Mapping cluster result

Table 2: Results of Kulonprogo cluster

High productivity	Medium productivity	Low productivity
Temon, Galur, Sentolo, Nanggulan	Wates, Panjatan, Lendah Kalibawang, Samigaluh	Pengasih, Kokap, Girimulyo

Table 3: Results of Bantul cluster

High productivity	Medium productivity	Low productivity
Bambanglipor, Jetis, Piyungan, Banguntapan	Sanden, Kretek, Pundong, Pandak, Bantul, Imogiri, Pleret, Sewon, Sedayu	Srandakan, Dlingo,Kasihan, Pajangan

Table 4: Results of Gunung Kidul cluster

High productivity	Medium productivity	Low productivity
Ponjong, Karangmojo, Patuk, Gendangsari, Ngawen, Semin	Saptopsari, Semanu, Wonosari, Playen	Panggang, Purwosari, Paliyan, Tepus, Tanjungsari, Rongkop Girisubo

Regions in Kulonprogo which are included in the high rice productivity regions are Temon, Galur, Sentolo and Nanggulan while regions with lowrice productivity are Kokap and Girimulyo.

The regions in Bantul which are included in the high rice productivity areas are Bambanglipuro, Jetis and Banguntapan while areas with low rice productivity are Srandakan, Kasihan and Pajangan.

The regions in Gunung Kidul which are included in high rice productivity areas are Ponjong, Karangmojo, Patuk, Gendangsari and Semin while areas with low rice productivity are Panggang, Purwosari, Paliyan, Tepus, Tanjungsari, Rongkop and Girisubo.

Regions in Sleman which are included in the regions with high productivity of rice are Ngemplak, Ngaglik, Sleman and Tempel while regions with low rice productivity are Depok and Turi (Table 5).

Table 5: Result of Sleman cluster

High productivity	Medium productivity	Low productivity
Ngemplak, Sleman	Moyudan, Minggir, Seyegan, Godean, Gamping, Mlati, Berbah, Prambanan, Kalasan, Ngaglik, Tempel	Depok, Turi

Table 6: Result of Yogyakarta cluster

High productivity	Medium productivity	Low productivity
Umbulharjo, Kotagede Tegalrejo	Mantrijeron, Mergangsan	Kraton, Gondokusuma, Danurejan, Pakualaman, Gondomanan, Ngampilan, Wirobrajan, Gedongtengen, Jetis

Region in Yogyakarta that is included in the regions with high productivity of rice namely Umbulharjo, Kotagede and Tegalrejo while the regions with low productivity of rice namely Kraton, Gondokusuman, Danurejan, Pakualaman, Gondomanan, Ngampilan, Wirobrajan, Gedongtengenand Jetis which does not produce rice because there is no rice field (Table 6).

Every district in 5 regencies and cities in Special Region of Yogyakarta had have their own label of productivity level, the analysis that had been done could be the evaluation material for agriculture agency as the reference to optimize the effort of increase of rice productivity for the regions that are still lacked for.

Accuracy test: Validation was conducted to ensure accuracy of clustering result. Validation process was done by comparing the mean of center of last cluster from the program developed with validation data obtained from calculation result by using Weka tools namely validation data in the form of area of field and rice production.

Validation result can be seen in the accuracy test that obtains value at 62.9%, if comparing centroid value of developed program with calculation result in Weka tools. That percentage was only in the form of reference because the used comparator was the calculation result in Weka and not the calculation of result in the field by the government of Yogyakarta Regency related to the cluster center of rice production and area of field.

CONCLUSION

After doing test to the application program of regional groupings based on rice productivity, so, it can be concluded as follows. Validation result of program obtained membership in each cluster that is almost similar and centroid value obtained was different but it had same labeling. In contrast, accuracy level of centroid value of program, if compared with Weka tools had value at 62.9%.

System could process data of production and area of rice field in Special Region of Yogyakarta for each in 2010 until 2016 to be classified into 3 regional status of rice productivity namely high, medium and low in every year.

Clustering with k-means can help agriculture party in Special Region of Yogyakarta in recommending the region that still need an effort of increase of rice productivity or control region that had highly produced the rice.

REFERENCES

Bholowalia, P. and A. Kumar, 2014. EBK-means: A clustering technique based on elbow method and K-means in WSN. Intl. J. Comput. Appl., 105: 17-24.

Feldman, R. and J. Sanger, 2006. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, New York, USA., ISBN: 9781139457835, Pages: 410.

Han, J., M. Kamber and J. Pei, 2012. Data Mining: Concepts and Techniques. 3rd Edn., Elsevier, Amsterdam, Netherlands, USA., ISBN:9789380931913, Pages: 703.

Handoyo, R., R. Mangkudjaja and S.M. Nasution, 2014. [Comparison of clustering methods using the single linkage and K-means method on document grouping (In Indonesian)]. J. Sifo Mikroskil, 15: 73-82.

Kodinariya, T.M. and P.R. Makwana, 2013. Review on determining number of cluster in K-Means clustering. Intl. J. Adv. Res. Comput. Sci. Manage. Stud., 1: 90-95.

Manning, C.D., P. Raghavan and H. Schutze, 2009. Introduction to Information Retrieval. Cambridge University Press, Cambridge.

Srivastava, A.N. and M. Sahami, 2009. Text Minning Classification, Clustering and Applications. CRC Press, Boca Raton, Florida, USA., ISBN:9781420059458, Pages: 328.

Wu, J., 2012. Advances in K-means Clustering: A Data Mining Thinking. Springer, Berlin, Germany, ISBN:9783642298073, Pages: 180.