Applied Expectation Maximization (EM) Clustering for Local Variety Corn

Dwivayani Sentosa, Budi Susetyo, Utami Dyah Syafitri, Sutoro

Abstract— Corn plays an important role in food diversification since corn contain large amount of carbohydrate. One of the efforts to support the achievement of food diversification is to create a superior corn varieties, namely high yield and resistance to disease, through the cross-linking process. To get good hybrid seed, it also takes good parent inbred. There are 98 local varieties of maize Java based on morphologic characteristics. Cluster analysis is used in order to know superior morphology for the cluster of local varieties corn. EM algorithm is used as a comparison for hierarchical clustering. Based on the Pseudo-F test and Cluster Thightness Measure (CTM), the reult for clustering using EM algorithm is better than the hierarchical clutering.

Index Terms— clustering, EM algorithm, hierarchical clustering.

1 INTRODUCTION

Corn is one crop that plays an important role as a source of carbohydrate. There was increasing demand, but the supply almost decrease. So that goverment through their researcher develop superior varieties of corn that can be cultivated by people.

Source genes that can improve the productivity of maize needed to obtain superior varieties. Source genes can be obtained from the germplasm (genetic resources) from various local varieties. Local varieties can be used if the characteristic already known. The evaluation to determine the character of the germplasm requires an efficient way, given the amount of germplasm quite a lot. Efficiency is necessary, to minimize the use of time and labor charges. Clustering germplasm into homogeneous cluster based on the charateristic is used to help the effiency. Furthermore, from that cluster can be obtained the major collections (core collection) of corn, which can be used in the process of further research.

Cluster analysis is a statistical method that aims to group objects into a cluster that tend to have homogeneous characteristics within the cluster compared to between clusters. Similarities between objects in the cluster of analysis is determined using the information distance between two objects. In general there are two methods for distance-based clustering, hierarchical method and non hierarchical method. Hierarchical methods assume each object is a cluster, then merge the two adjacent objects into a larger cluster. The non hierarchical method assume all data observations is one large cluster and then separated into a number of small cluster.

Hierarchical clustering is the most often used by researchers because it is the simplest method, easily implemented and adapted to control every aspect of science. However, hierarchical clustering not consider the distribution of data, just pay attention to the proximity between observations. In addition, the results are determined based expertise's approach, so that the amount of the clustering result may differ among researchers.

One alternative method to overcome the limitation of hierarchical clustering offerd by Mclachan and Basford (1998) is model-based clustering, which use distribution of data. . It is better known as mixture model clustering. Cluster based mixed model can be applied to categorical data, continuous, or a combination of both. The purpose of this method is to optimize the similarity between objects by using probability distribution. Each distribution represents one cluster with certain parameters. Parameter estimation method used in this paper is Expectation Maximization algorithm (EM) that maximizes the log-likelihood function. Maximum likelihood requires the normal distribution data. Distribution of the data used in the application of the EM algorithm is Gaussian Mixture Models (GMM).

The purpose of this study is to apply the hierarchical clustering and EM algorithm in cluster local maize varieties from Java. Than compare the results obtained from each method.

2 RESEARCH METHOD

2.1 Data

The data used in this research is the data of local maize varieties from Java obtained from BB Biogen, Bogor. There are 18 morphological features variables observed in the form of 98 observations/varieties with different genotypes of each other, which describe in Table 1. Data collected from plants start blooming until harvest time.

2.2 Methods of Data Analysis

The stages of data analysis in this research are as follow:

1. Descriptive Analysis
   Descriptive analysis was performed to explore the general description of data patten that aimed to get the appropriate next analysis.

- *Dwivayani Sentosa is currently pursuing masters degree program in applied statistics in Bogor Agricultural University, Indonesia, PH +621382377454. E-mail: dwivayani@gmail.com*
- *Budi Susetyo is Lecturer, Departement of Statistics, Bogor Agricultural University, Bogor, Indonesia*
- *Utami Dyah Syafitri is Lecturer, Departement of Statistics, Bogor Agricultural University, Bogor, Indonesia*
- *Sutoro is Principal Researcher, BB Biogen, Bogor, Indonesia*

TABLE 1
LIST FOR OF VARIABLES

Variables	Explanation
X_1	Plant height (cm)
X_2	Leaft height (cm)
X_3	Leaft height (cm)
X_4	Number of leaves above the uppermost ear
X_5	Days to tasseling (day)
X_6	Days to silking (day)
X_7	Tassel length (cm)
X_8	Tassel peduncle length (cm)
X_9	Number of primary branches on tassel
X_{10}	Ear height (cm)
X_{11}	Days to mature (d)
X_{12}	Ear length (cm)
X_{13}	Ear diameter (mm)
X_{14}	Number of kernel rows
X_{15}	Number of kernel per ear
X_{16}	Number of ear per plot
X_{17}	1000 kernel weight (g)
X_{18}	Estimated yield (t/ha)

2. Checking the normality assumption
   Examine each variable by using the Kolmogorov-Smirnov test statistic.
3. Determine the principal component analysis score to overcome multicollinearity in data.
4. Perform hierarchical clustering.
   Do the hierarchical stages as follows:
   a. Calculating distance matrix using Euclidean distance as $d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})'(\boldsymbol{x}-\boldsymbol{y})}$.
   b. Do all hierarchical clustering methods: single linkage, average linkage, median linkage, complete linkage, and ward.
5. Perform EM algorithm for clustering.
   Stages in EM algorithm, i.e:
   a. Do hierarchical clustering to mapping the observation into clusters.
   $$z_{ik} = \begin{cases} 1; & \text{if } \boldsymbol{x}_i \text{ join cluster k} \\ 0; & \text{otherwise} \end{cases}$$
   b. M-step, calculate maximum likelihood parameter estimation: n_k the number of observation in cluster-k, $\widehat{\tau_k}$ propability to join cluster-k, $\widehat{\mu_k}$ mean vector of cluster-k, and $\widehat{\Sigma_k}$ covariance matrix for cluster-k.
   c. E-step, estimates the member of each cluster using parameter from M-step: $\widehat{z_{ik}} = \frac{\widehat{\tau_k} f_k(x_i|\widehat{\mu_k},\widehat{\Sigma_k})}{\sum_{j=1}^{c} \widehat{\tau_j} f_j(x_i|\widehat{\mu_j},\widehat{\Sigma_j})}$.
   d. Calculating BIC score $2l_{\mathcal{M}}(x,\hat{\theta}) - m_{\mathcal{M}}\log(n) \equiv BIC$
   e. Do iteration from a-d for the different number of cluster. The largest BIC score indicates the good result for the number of cluster.
6. Comparing the results of the cluster of the two methods. This are three methods to compare:
   a. Davies-Bouldien Index: $DB = \frac{1}{n}\sum_{i=1}^{n} \max_{i=j}[\frac{d'(c_i)+d'(c_j)}{d(c_i,c_j)}]$
   Based on Su [7], the smallest index indicates the number of optimum cluster.
   b. Pseudo-F Calinski-Harabasz: comparison between sum square between and sum square within. Largest score of Pseudo-F indicates optimum cluster with equation: $F = \frac{T-P_C/(G-1)}{P_C/(n-G)}$.
   c. Cluster Tightness Measure (CTM) shos optimum cluster if the score is minimum. CTM use standard deviation from each clusters, as $CTM = \frac{1}{k}\sum_{t=1}^{k}\left(\frac{1}{p}\sum_{m=1}^{p}\frac{s_{tm}}{s_m}\right)$

3 RESULT AND DISCUSSION

3.1 Descriptive Analysis

Data consists of 18 variables which represent morphological characteristics of observations. Information on preliminary data research can be seen at Table 2. Coefficient of variance on all variables are below 50%, indicating the diversity of data observations on each variable is relatively small.

TABLE 2
INFORMATION OF VARIABLES

Variables	Mean	Standard Deviation	Coeff. of Variance	Normality Test
X1	149.12	25.21	16.91%	lnX tranformation
X2	77.74	7.04	9.06%	Normal dist
X3	8.08	1.03	12.75%	Normal dist
X4	4.97	0.79	15.90%	Non-normal dist
X5	48.94	6.41	13.10%	1/X tranformation
X6	51.33	6.81	13.27%	Non-normal dist
X7	36.91	3.39	9.18%	Normal dist
X8	8.39	1.94	23.12%	Normal dist
X9	15.95	4.80	30.09%	Normal dist
X10	63.83	18.11	28.37%	lnX tranformation
X11	86.72	9.41	10.85%	Non-normal dist
X12	13.00	2.42	18.62%	Normal dist
X13	40.08	4.02	10.03%	1/X tranformation
X14	11.65	1.29	11.07%	Non-normal dist
X15	283.66	65.61	23.13%	Normal dist
X16	35.23	6.95	19.73%	Non-normal dist
X17	250.13	121.08	48.41%	Non-normal dist
X18	3.06	1.07	34.97%	lnX tranformation

Normality test is checked by Komogorov-smirnov test. There are 7 over 18 variables have normal dist. The examination of non-normal variable is using Box-Cox test. The Box-Cox result shows that 3 variables normalized by transforming using ln transformation, and variables using X^{-1} transformation. The other 6 variables not included in this research, because the data is discrete.

3.2 Hierarchical Clustering

Davies-Bouldien Index is used to checked optimum cluster for hierarchical clustering. The smaller the index, show the most optimum cluster. There are 2 cluster separeted for all hierarchical clustering methods. Then, from the result of 2 cluster, the evaluation for the best model is using Pseudo-F test. The larger the value of Pseudo-F, show the best model fo clustering. It can be seen from Table 3, Ward and Complete Linkage come as the best methodology for this hirarchical clustering. Then, we continu to use Ward Method and Complete Linkage for thi3 research.

TABLE 3
PSEUDO-F TEST FOR HIERARCHICAL METHODS

Methodology	Pseudo-F Test
Single linkage	2.1316
Average lingkage	2.1316
Median linkage	2.2327
Complete linkage	**13.2631**
Centroid linkage	2.5589
Ward	**14.8822**

The 2nd cluster has higher value for morphological characteristics than the 1st one, for both Complete Linkage and Ward Method. Based crosstabulation on Table 4, all member of 1st cluster from Complete Linkage also being a member of 1nd cluster in Ward Method. However, 39 observations of 77 observations that went into 2nd cluster in Complete Linkage, being the member of 1st cluster in Ward Method. It can be concluded that there is 39.80% missclassification observation between Complete Linkage and Ward Method.

TABLE 4
CROSSTABULATION RESULT BETWEEN COMPLETE LINKAGE AND WARD METHOD

		Ward Methods		Total
	Cluster	1	2	
Complete linkage	1	21	0	21
	2	39	38	77
	Total	38	60	98

3.3 Clustering Using EM Algorithm

The application process of EM algorithm clustering is using package Mclust in R. BIC value show that EM algorithm can devide data into 2 cluster, with 46 observations are in 1nd cluster and 52 observations on the 2nd cluster. The mean mixture and variance mixture for 1st cluster are 0.4621 and 0.3426. And mean mixture and variance mixture for 2nd cluster are 0.5378 and 0.4747. Based on Manova test, all variables contribute to differences between clusters. Than by using T^2 Hotteling test, F-test is greater than F-table (3.1584> 2.9498), which means the mean of 1st cluster is different from 2nd cluster. Based on t-test, the mean of each variable between 1st cluster and 2nd cluster is significantly different at significancy level 5%.

Characteristics of cluster presented in Table 5. The 1st cluster has smaller average value for morphological characteristics than the 2nd one. Only the average of tassel length from 1st cluster is greater value than the 2nd cluster.

TABLE 5
PSEUDO-F TEST FOR HIERARCHICAL METHODS

Variables	Algoritma EM	
	1st cluster	2nd cluster
X1	131.79	164.46
X2	73.69	81.32
X3	7.32	8.75
X5	43.96	53.35
X7	35.94	37.76
X8	9.60	7.32
X9	12.65	18.87
X10	51.84	74.44
X12	11.56	14. 27
X13	30.79	35.10
X15	236. 22	325.63
X18	2.48	3.58

3.4 Cluster Validation

Statistics Pseudo-F and CTM is used to get the best methods on clustering. An optimum cluster can be defined by the largest Pseudo-F value and the smallest CTM value (Table 6).

TABLE 6
CLUSTER VALIDATION

Methods	Pseudo-F	CTM
Complete linkage	55.8578	0.7390
Ward	43.0129	0.8037
EM Algorithm	**90.5562**	**0.3586**

Cross tabulation accuracy between Complete Linkage and EM algorithm is shown in Table 7. There are 75.51% observations from complete linkage which have same cluster in EM algorithm result. The accuracy between Ward and EM algorithm are 85.71%. It means only 14 observations from Ward which not classified in the same cluster with EM algorithm (Table 8). Since Ward has lager value of accuracy from EM algorithm, it means that Ward has the closest result

to EM algorithm result.

TABLE 7
CROSSTABULATION RESULT BETWEEN COMPLETE LINKAGE AND EM ALGORITHM

		EM Algorithm		Total
	Cluster	1	2	
Complete linkage	1	21	0	21
	2	25	52	77
	Total	46	52	98

TABLE 8
CROSSTABULATION RESULT BETWEEN WARD METHOD AND EM ALGORITHM

		EM Algorithm		Total
	Cluster	1	2	
Ward	1	46	14	21
	2	0	38	77
	Total	46	52	98

4 CONCLUSION

Local varieties corn form Java could be separated into two clusters. The result for comparison between hierarchical clustering and EM algorithm found that EM algorithm could saparated cluster better than hierarchical clustering, based on the highest Pseudo-F score and the lowest CTM score. Ward method from hierarchical clustering give the closest result to clustering using EM algorithm.

REFERENCES

[1] Calinski T, Harabasz J. 1974. A Dendrit Method for Cluster Analysis. Communications in Statistics Theory and Method. 3(1): 1-27 (Journal or magazine citation)

[2] Epps J, Ambikairajah E. Visualisation of Reduced Dimension Microarry Data Using Gaussian Mixture Model. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.59.1619. 2008. (URL link *include year)

[3] Fraley C, Raftery AE. 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. The Computer J 41(8). (Journal or magazine citation)

[4] Fraley C, Raftery AE. 2002. Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association 97:611 (Journal or magazine citation)

[5] Johnson RA, Wichern, DW. 2007. Applied Multivariate Statistical Analysis, 6th edition. New Jersey: Prentice-Hall (Book style)

[6] Sambandam, Rajan. 2003. Cluster Analysis Gets Complicated. Marketing Research, Vol 15 (1) (Journal or magazine citation)

[7] Su MC. A New Index of Cluster Validity. http://www.cs.missouri.edu/~skupicm/8820/ClusterValid.pdf. 2003. (URL link *include year)