III.B.1.a5_TURNITIN_Bayesian Hierachical Clustering for Bidikmisi Environment Results of Successful and Unsuccessful Scholarship Cluster

by W Suryaningtyas

Submission date: 23-Nov-2022 01:35PM (UTC+0700)
Submission ID: 1961879702
File name: t_Results_of_Successful_and_Unsuccessful_Scholarship_Cluster.pdf (1.26M)
Word count: 2200
Character count: 12167

Bayesian Hierachical Clustering for Bidikmisi Environment: Results of successful and unsuccessful scholarship cluster

IOP Conf. Series: Earth and Environmental Science **469** (2020) 012092 doi:10.1088/1755-1315/469/1/012092

Bayesian Hierachical Clustering for Bidikmisi Environment: Results of successful and unsuccessful scholarship cluster

W Suryaningtyas[1,2], N Iriawan[3], H Kuswanto[3], I Zain[3]

[1]Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia.
[2] Mathematic Education Program Study, Faculty of Teacher Training and Education, Muhammadiyah University of Surabaya, Surabaya, Indonesia.
[3]Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia.
E-mail: wahyuni.pendmat@fkip.um-surabaya.ac.id

Abstract. This research focuses on conducting Bidikmisi data clusters. Group analysis or cluster analysis is one of the multivariate techniques used with the aim of classifying an observation that has certain characteristics. This analysis groups observations into a group so that in one group there are similarities or homogeneity with each other. In contrast, intergroup is expected to have differences or heterogeneity. The clustering process uses the Bayesian Hierachical Clustering (BHC) approach, aiming to present a complete taxonomy of grouping evaluation measurements that will be used for empirical studies.

1. Introduction

Bidikmisi data comes from the registration process data which is done online by accessing information on the Bidikmisi scholarship official website page: http://bidikmisi.belmawa.ristekdikti.go.id [1]. Registration (registration) of Bidikmisi scholarships annually shows massive growth of heterogeneous data, this is due to the large number of Bidikmisi registrants from all districts in Indonesia. The massive data phenomenon of Bidikmisi provides a variety of information. Therefore, information management is needed that contains important things in decision making. The information obtained is called data. Bidikmisi data displays many characters, so that it can cause problems in decision making. Therefore a data analysis technique is needed as an effort to process data so that data characteristics can be easily understood and useful in answering related problems.

This study focuses on conducting Bidikmisi data clusters. Group analysis or cluster analysis is one of the multivariate techniques used with the aim of classifying an observation that has certain characteristics. This analysis groups observations into a group so that in one group there are similarities or homogeneity with each other. In contrast, intergroup is expected to have differences or heterogeneity [2]. The clustering process uses the Bayesian Hierachical Clustering (BHC) approach [3], aiming to present a complete taxonomy of grouping evaluation measurements that will be used for empirical studies.

This study aims to analyze the Bidikmisi admission cluster based on acceptance conditions with indicators of family ability. The method was analyzed using R. software. Furthermore, the results of the analysis were used to compile the Bidikmisi acceptance cluster.

2. Literature Review

2.1. Bayesian Hierarchical Clustering

Bayesian Hierarchical Clustering [4] presents a new algorithm for agglomerative hierarchical clustering using Bayesian methods by performing hierarchical grouping using marginal likelihood in evaluating cluster membership to determine the merging of profitable clusters and avoiding overfitting. Bayesian Hierarchical Clustering (BHC) algorithm for each potential cluster clustering $D_k = D_i \dot{E} D_j$ using a hypothesis:

H_1 : all data in D_k comes from one cluster

H_2 : data in D_k comes from other groupings that are consistent with trees T_i, T_j

The first hypothesis H_1^k indicates that all data is in D_k produced independently and identically from the same probabilistic model, $p(x|\theta)$ with unknown parameters θ. In order to evaluate the probability of the data, more prior determination is needed than the model parameters, $p(q|b)$ with hyperparameter b. Data probability D_k on the hypothesis H_1^k given as follows:

$$p(D_k|H_1^k) = \int p(D_k|q)p(q|b)db = \int \left[\prod_{x^{(i)}\in D_k} p(x^{(i)}|q)\right] p(q|b)db \,. \quad (1)$$

Determine the probability that all data is in D_k generated from the same parameter values assuming the model of the form $p(x|\theta)$.

2.2. *Bayesian MCMC Marginal Probability*

The combination of the probability of the data in the hypothesis H_1^k and H_2^k, weighted by prior all the inner points D_k become one cluster, with $p_k \overset{def}{=} H_1^k$ given the marginal probability of data on the tree T_k, that is:

$$p(D_k|T_k) = p_k\, p(D_k|H_1^k) + (1- p_k)p(D_i|T_i)p(D_j|T_j). \quad (2)$$

Equations can be used to reduce the marginal likelihood estimation of the Dirichlet Process Mixture (DPM) model and provide a new lower limit of marginal likelihood. Posterior probability of the combined hypothesis $r_k \overset{def}{=} p(D_k|H_1^k)$ obtained using Bayes rules:

$$r_k = \frac{p_k p(D_k|H_1^k)}{p_k p(D_k|H_1^k) + (1- p_k)p(D_i|T)p(D_j|T_j)}. \quad (3)$$

This quantity is used to decide which two trees are combined and determine the merge in the final structure of the hierarchy.

3. Methodology

3.1 *Source of Data*

The data used in this study are Bidikmisi data from all provinces in Indonesia in 2015, with several provinces as follows: Gorontalo, Bengkulu, Bangka Belitung, East Kalimantan, Riau Island, Jambi, West Sumatra, South Sumatra, North Sumatra, Banten, Yogyakarta Special Region (DIY), NTT and NTB. Bidikmisi data is sourced from the Ministry of Research, Technology and Higher Education Database through the Bidikmisi channel.

3.2 *Variable Research*

Research variables used in this study consisted of the response variable (*Y*) and the predictor variable).

Y = The acceptance Status of Bidikmisi Scholarship (1 = accepted, 0 = not accepted)

X_1 = Father's job with four dummies – d_{11} as an agricultural sectors, b_{12} as the government employee, d_{13} as an entrepreneur, and d_{14} as a private employee;

X_2 = Mother's Job with four dummies d_{21}, d_{22}, d_{23}, and d_{24} defined as in Father's job;

X_3 = Father's Education with three dummies – d_{31} as non educated, d_{32} as elementary to senior high school education, and d_{33} as higher education;
X_4 = Mother's Education with three dummies – d_{41}, d_{42}, and d_{43} defined as in Father's Education.
Each dummy variable has the value of 0 or 1.

There are still many variables in the registration form of Bidikmisi, but these four variables selected above are more fundamental variables in considering the acceptance of these grantees, in accordance with one of the rules of acceptance of Bidikmisi is that the income per-capita in the family is no more than certain values **[5]**.

3.3 Research Design

Cluster analysis procedures using Bayesian Hierarchical Clustering (BHC) is given the following research flows in Figure 1:

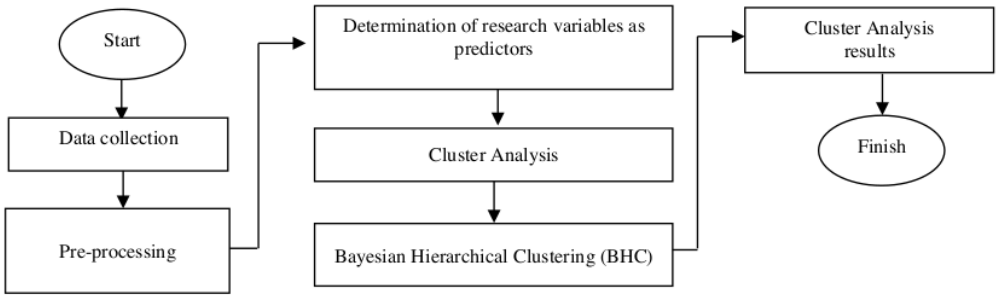


Figure 1. Flowchart Cluster Analysis using Bayesian Hierarchical Clustering

4. Research Result

4.1. Pre-processing

Explanation of the techniques used in the pre-processing stage identification component Polytomous of Bidikmisi receipts given as follows:

Step 1. take response variable (Y)

Step 2. select covariate "father's income", "mother's income" and "family dependent"

Step 3. Create a new covariate by counting the amount of "dad's income" and "maternal income" divided by "the number of family dependents", then name it with "Code Category (CC)".

Step 4. coding the covariate "CC" with the following criteria:

0 = if CC> Rp. 750,000 per head in the family included in the category of wealthy family

1 = if CC <Rp. 750,000 per head in the family fall into the category of poor families.

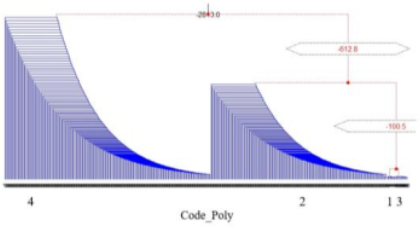
Step 5. match the response variable (Y) to the CC in Step 4 to the AC (Acceptance Condition) with the Bidikmisi acceptance classification table of "wrong" and "right" which are given as follows:

Table 1. Identification Components Polytomous Bidikmisi Scholarship 2015

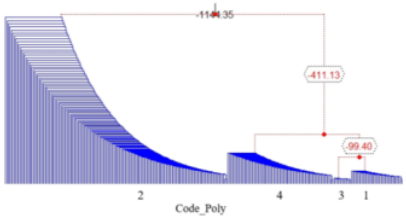
Y	CC	AC	Code_Poly	Interpretation
0	0	1	1	Acceptance Condition is right (AC = 1) if the grantee (Y = 0) is followed with the category of wealthy family (CC = 0)
0	1	0	2	Acceptance Condition is wrong (AC = 0) if the grantee (Y = 0) is followed with the category of poor family (KK = 1)
1	0	0	3	Acceptance Condition is wrong (AC = 0) if the grantee (Y = 1) is followed with the category of wealthy family (CC = 0)
1	1	1	4	Acceptance Condition is right (AC = 1) if the grantee (Y = 1) is followed with the category of poor family (CC = 1)

4.2. Bayesian Hierarchical Clustering

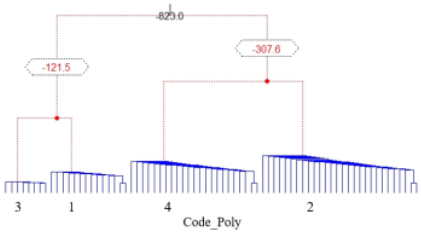
Cluster analysis uses Bayesian Hierarchical Cluster by utilizing R software through BHC packages. Resulting in grouping based on the BHC method with several provinces: Gorontalo, Bengkulu, Bangka Belitung, East Kalimantan, Riau Island, Jambi, West Sumatra, South Sumatra, North Sumatra, Banten, Yogyakarta Special Region (DIY), NTT and NTB as follows in Figure 1:



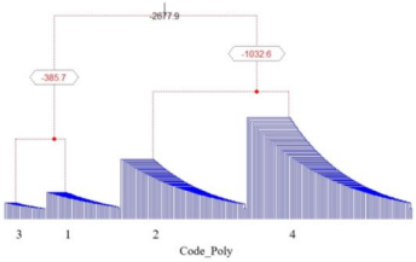
(a) Gorontalo Province



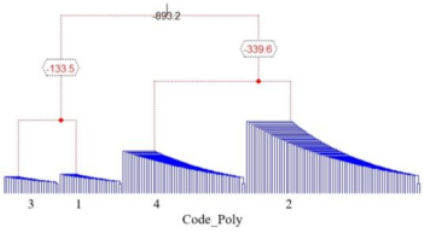
(b) Bengkulu Province



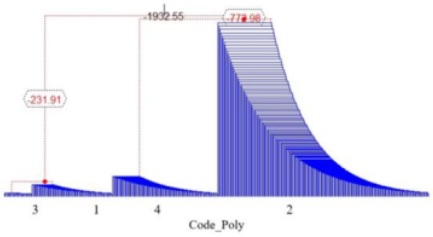
(c) Bangka Belitung Province



(d) East Kalimantan Province

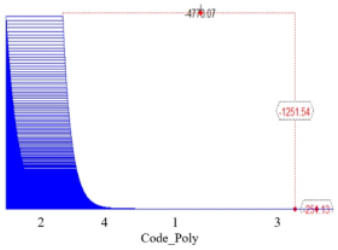


(e) Riau Island Province

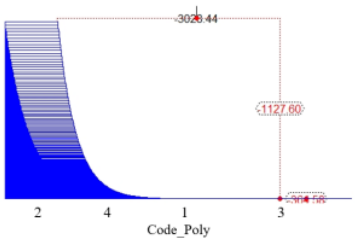


(f) Jambi Province

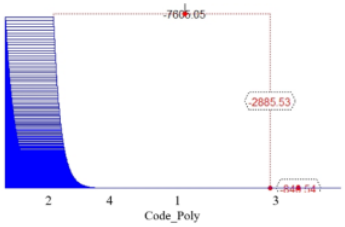
Continued Figure 1



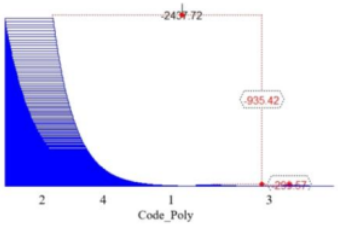
(g) West Sumatra Province



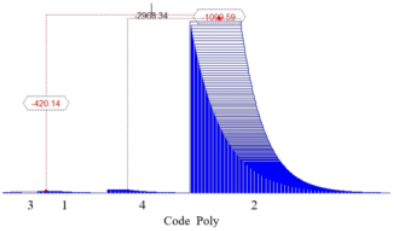
(h) South Sumatra Province



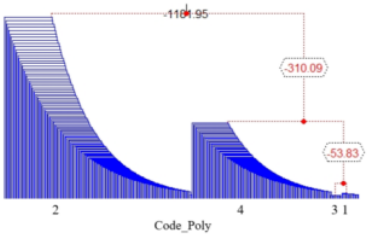
(i) North Sumatra Province



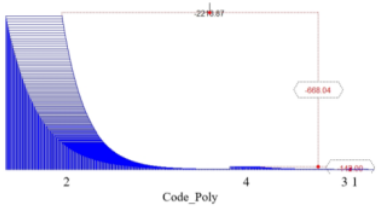
(j) Banten Province



(k) Yogyakarta Special Region (DIY) Province



(l) NTT Province



(m) NTB Province

Figure 1. BHC Result Cluster Analysis for several provinces

Based on Figure 1 above, it can be seen that the BHC method did not succeed in classifying groups that were able and unable (poor) in several provinces there are Gorontalo, Bengkulu, West Sumatra, South Sumatra, North Sumatra, Banten, NTT and NTB. Cluster analysis using the BHC method was successful in classifying groups that were able and poor for 5 (five) provinces namely: Bangka Belitung, East Kalimantan, Riau Island, Jambi, and Yogyakarta Special Region (DIY). BHC methods in classifying groups in several Indonesian provinces if presented in table form are as follows:

Table 2. BHC Methods In Classifying Groups Able and Unable (Poor) in Several Indonesian Provinces

Province	BHC Method
Gorontalo Province	Not Success
Bengkulu Province	Not Success
Bangka Belitung Province	Success
East Kalimantan Province	Success
Riau Island Province	Success
Jambi Province	Success
West Sumatra Province	Not Success
South Sumatra Province	Not Success
North Sumatra Province	Not Success
Banten Province	Not Success
Yogyakarta Special Region (DIY) Province	Success
NTT Province	Not Success
NTB Province	Not Success

5. Conclusion

Bidikmisi acceptance cluster analysis using the BHC method gives two results, namely successful and unsuccessful in classifying data based on acceptance conditions with indicators of ability of capable and unable families. The BHC method successfully classifies capable and poor groups in 5 (five) provinces, namely: Bangka Belitung, East Kalimantan, Riau Islands, Jambi, and Yogyakarta Special Region (DIY). Whereas for the other 8 (eight) provinces, they were not successfully grouped based on indicators of family ability.

Acknowledgements

The Authors are grateful to Directorate for Research and Community Service (DRPM) Ministry of Research, Technology and Higher Education Indonesia which support this research under PDD research grant no. 022/II.3.SP/L/IV/2018.

6. References

[1] K. R. T. d. P. T. Direktorat Jenderal Pembelajaran dan Kemahasiswaan, Pedoman Penyelenggaraan Bantuan Biaya Pendidikan Bidikmisi Tahun 2016, Jakarta: Belmawa, Kemeristek Dikti, 2016.

[2] . J. Hair , W. C. Black , B. J. Babin and Anderson, Multivariate Data Analysis, 7th Edition ed., New York: Prentice Hall International, Inc., 2010.

[3] K. Heller and Z. Ghahraman, Bayesian Hierarchical Clustering (Technical Report 2005-002), London: Gatsby Computational Neuroscience Unit, 2005.

[4] K. Heller and Z. Ghahraman, Bayesian Hierarchical Clustering (Technical Report 2005-002), London: Gatsby Computational Neuroscience Unit, 2005.

[5] N. Iriawan, K. Fithriasari, B. S. S. Ulama, W. Suryaningtyas, I. Susanto and A. A. Pravitasari, "Bayesian Bernoulli Mixture Regression Model for Bidikmisi Scholarship Classification," *Journal of Computer Sciences and Information,* vol. 11, no. 2, DOI: http://dx.doi.org/10.21609/jiki.v11i2 , 2018.

III.B.1.a5_TURNITIN_Bayesian Hierachical Clustering for Bidikmisi Environment Results of Successful and Unsuccessful Scholarship Cluster

ORIGINALITY REPORT

10%	%	9%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Cambridge Student Paper	4%
2	"Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)", Springer Science and Business Media LLC, 2019 Publication	2%
3	P Indrawati, N Subekti. "The effectiveness of the Macrotermes gilvus termite queen for sperm repair in infertile mice", Journal of Physics: Conference Series, 2020 Publication	2%
4	Wahyuni Suryaningtyas, Nur Iriawan, Heri Kuswanto, Ismaini Zain. "On the Hierarchical Bernoulli Mixture Model Using Bayesian Hamiltonian Monte Carlo", Symmetry, 2021 Publication	2%

Exclude quotes	On	Exclude matches	< 30 words
Exclude bibliography	On		