Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

On The Bernoulli Mixture Model for Bidikmisi Scholarship Classification with Bayesian MCMC

To cite this article: W Suryaningtyas *et al* 2018 *J. Phys.: Conf. Ser.* **1090** 012072

View the article online for updates and enhancements.

You may also like

- Bayesian Hierachical Clustering for Bidikmisi Environment: Results of successful and unsuccessful scholarship cluster
  W Suryaningtyas, N Iriawan, H Kuswanto et al.
- Investigating the influence of a cyclist on crowd behaviors on a shared road
  Chuanli Huang, Lu Wang, Hang Yu et al.
- JEANS ANALYSIS OF THE GALACTIC THICK DISK AND THE LOCAL DARK MATTER DENSITY
  F. J. Sánchez-Salcedo, Chris Flynn and J. A. de Diego

On The Bernoulli Mixture Model for Bidikmisi Scholarship Classification with Bayesian MCMC

W Suryaningtyas[1,2], **N Iriawan**[3], **K Fithriasari**[3], **BSS Ulama**[3], **I Susanto**[1], **AA Pravitasari**[1]

[1] Doctoral Student at 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; nur_i@statistika.its.ac.id

Abstract. This research has a purpose to develop Bernoulli Mixture model for Bidikmisi data modelling using Bayesian approach. Model development is done by considering the specificity in the data acceptance of Bidikmisi scholarship prototype in East Java Province. Bidikmisi acceptance status having a binary type (0 and 1) coupled with the main criteria factor of parent income and the number of dependents family produces a structure of Bernoulli mixture distribution with two components. The characteristics of each component can be identified through the Bernoulli Mixture modelling by involving the covariates of Bidikmisi scholarship recipients. The estimating parameter was performed using Bayesian Markov Chain Monte Carlo (MCMC) couple with the Gibbs Sampling algorithm. This model is applied to data registrants Bidikmisi districts/cities in the province of East Java as many as 44,489 students. This model shows the smallest value of Deviance Information Criteria (DIC) compared with Bayesian binary logistic regression.

1. Introduction

Bernoulli Mixture Model (BMM) is a model used to analyze the Bernoulli Mixture distributed data. Most of the references on binary data, the BMM research are mostly applied in the area of text mining [1]. BMM was first performed by Duda and Hart [2]. In its development, some research related to BMM were performed by Grim et al. [3], González, et al. [4], and Vidal [5,6], Patrikainen and Manilla [7], Zhu et al. [8], Sun et al. [9], Tikka et al. [10], Hollmen and Tikka [11], Myllykangas et al. [12], Bouguila [13] and Saeed et al. [14].

In this study, the development of BMM using Bernoulli Mixture regression analysis was applied in the social field by using local data of Bidikmisi scholarship. Bidikmisi tuition assistance program is an education assistance program by the Indonesian government through the Directorate General of Higher Education, which was launched in 2010. The government through Bidikmisi program aims to achieve equitable access and learning opportunities at university level and produce the independent, productive and having social care graduates who can play a role to solve the poverty chains to fill the needs of the Indonesian human resources and are ready for competing in the ASEAN Economic Community (MEA) [15].

This research employed the Bidikmisi prototype data of East Java Province. In 2015, the registrants of Bidikmisi in East Java were 44,489 students. The results of the data exploration showed that only 24.07% were awarded the scholarships, while those who were not successful were 75.93%, ie 33,780 student applicants. Bidikmisi grantee status with Binary type (0 and 1) was coupled with the main criteria factor, namely parent income and total of dependent. This combination produced the Bernoulli mixture data distribution with two components. Characteristics of each component of Bernoulli mixture can be identified through the Bernoulli Mixture modelling by involving the founding covariates of Bidikmisi scholarship grantee.

This study aims to perform the classification analysis of acceptance of Bidikmisi based on acceptance conditions with indicators of family ability. The classification is done by competing the Bayesian Bernoulli Mixture regression and Bayesian binary logistic regression. Both methods are analyzed using openBUGS software. Furthermore, the results of the analysis are used to compile the accuracy of Bidikmisi acceptance classification.

2. Literature Review

2.1. Bernoulli Mixture Model

If random sample Y is independently distributed Bernoulli Mixture derived from i -units, then there will be a vector $\mathbf{y}=\begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}^T$ with $i=1,2,\ldots,n$, which can contain L groups with the proportion $\boldsymbol{\pi}= \pi_1,\pi_2,...,\pi_L$. The finite mixture model of Y with L number components could have the density functions as follows [16]:

$$p(y_i)=\sum_{l=1}^{L}\pi_l p_l(y_i), \tag{1}$$

in which L is the number of mixture components and for each l, $p_l(y_i)$ is the mixture density component and π_l is the non-negative quantity which amounts to one, that is:

$$\sum_{l=1}^{L}\pi_l=1. \tag{2}$$

BMM based on model (1), if $\mathbf{Y}=(Y_1,Y_2,\ldots,Y_n)$ is a random sample. The goal is to partition $\mathbf{Y}$ into L (might be unknown, but limited) groups. The finite mixture density with L components can be written as [17]:

$$p(\mathbf{Y}\mid L,\boldsymbol{\pi},\boldsymbol{\Theta})=\sum_{l=1}^{L}\pi_l p_l(\mathbf{Y}\mid\theta_l), \tag{3}$$

where $p_l(.)$ is called with the mixture density l^{th} component, $\boldsymbol{\Theta}=(\theta_1,\theta_2,\ldots,\theta_L)$ is a mixture density component parameter, and $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_L)$ is the mixture proportion.

2.2. Bayesian MCMC

Bayesian Markov Chain Monte Carlo (Bayesian MCMC) is an approach, which applies Gibbs sampling process conducted through sampling by means of a series of Gibbs random variables based on the basic properties of Markov Chain [18]. Inference using the Bayesian approach to the parameter estimation process is done by integrating the posterior distribution. Numerically can be done integration through a simulation procedure commonly known as Markov Chain Monte Carlo (MCMC) method. The following are given the work steps of Markov Chain Monte Carlo method in general [19], namely:

1. Choosing a starting value $\boldsymbol{\theta}^{(0)}$.
2. Generating value $\boldsymbol{\theta}^{(m)}$, $m=1,...,M$ until they reached convergence in distribution.
3. Diagnosing the convergence of the chain of $\boldsymbol{\theta}$ used for monitoring. If the convergence diagnosis fails for length of *M*, then do some additional observation by increasing the larger *M*.
4. Cutting the first *B* observations as burn-in observations and dispose them.
5. Sampling some data $\left\{\boldsymbol{\theta}^{(B+1)},\boldsymbol{\theta}^{(B+2)},\ldots,\boldsymbol{\theta}^{(M)}\right\}$ as the estimated posterior parameters distribution.
6. Plotting each posterior parameter distribution (as the univariate marginal distribution each).
7. Finally, obtain summaries of the characteristics of the posterior distributions of $\boldsymbol{\theta}$.

3. Methodology

3.1 Source of Data

The data used in this research was sourced from Database of Ministry of Research, Technology and Higher Education through Bidikmisi channel, that was Bidikmisi data of all districts in East Java Province in 2015.

3.2 Variable Research

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

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

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

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

X_3 = Father's Education with three dummies – b_{31} as non educated, b_{32} as elementary to senior high school education, and b_{33} as higher education;

X_4 = Mother's Education with three dummies – b_{41}, b_{42}, and b_{43} defined as in Father's Education.

Each dummy variable has the value of 0 or 1.

3.3 Research Design

Classification analysis procedures using Bayesian Bernoulli mixture regression model and Bayesian binary logistic regression model are given the following research flows in Figure 1:

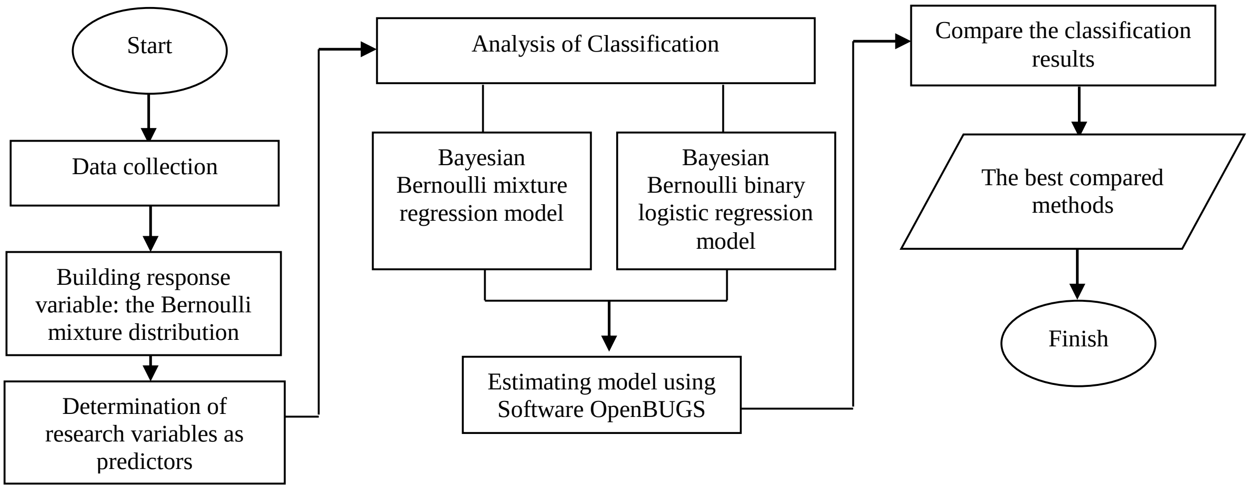

Start
Data collection
Building response variable: the Bernoulli mixture distribution
Determination of research variables as predictors
Analysis of Classification
Bayesian Bernoulli mixture regression model
Bayesian Bernoulli binary logistic regression model
Estimating model using Software OpenBUGS
Compare the classification results
The best compared methods
Finish

Figure 1. Flowchart Classification Bidikmisi using Bayesian Bernoulli Mixture Regression Model and Bayesian Binary Logistic Regression Model

4. Research Result

4.1. Pre-processing

The explanations of the techniques used in the pre-processing stages of identification for building the Bernoulli mixture distribution are 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 Mixture of Bidikmisi Scholarship 2015

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

The pre-processing result by involving founder covariate of the Bidikmisi scholarship shows response data of the Bernoulli mixture distribution with two components, namely component of wrong acceptance condition and component of right acceptance condition. Based on Table 1, it is obtained that *Bernoulli*-1 with θ_1=0.62 (AC is wrong) and *Bernoulli*-2 with θ_2=0.38 (AC is right).

4.2. *Bayesian Binary Logistic Regression*

The doodle of Bayesian binary logistic regression to model the acceptance of the Bidikmisi scholarship is presented in Figure 2.

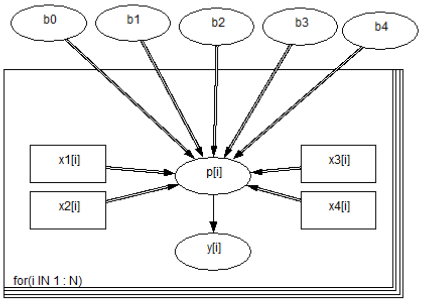


Figure 2. *Doodle* of Bayesian Binary logistic regression

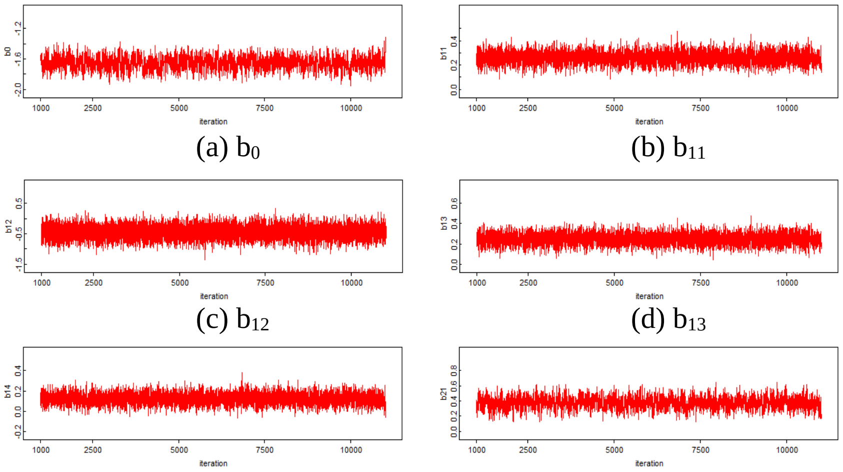
After compiling the doodle and its syntax, the next step is running program to get the estimated regression model. The serial historical sample values of each posterior parameter show a stable random pattern in a fixed domain, indicating the fulfillment of irreducible, aperiodic and recurrent properties. The historical sample series of the first six posterior estimation values (b_0, b_{11}, b_{12}, b_{13}, b_{14}, and b_{21}) are shown through the serial plot is presented in Figure 3. While their significant parameter estimates of the binary logistic regression using a link function logit is provided in Table 2.

Based on the significant parameter estimation by using Bayesian binary logistic regression in Table 2, the model can be expressed as follows:

$$\hat{\pi}(x)=\frac{\exp\left(-1,6540+0,2612\,b_{11}-0,4487\,b_{12}+0,2486\,b_{13}+0,1235\,b_{14}+0,3779\,b_{21}+0,3335\,b_{23}-0,1147\,b_{32}+0,0623\,b_{33}-0,1661\,b_{41}-0,1823\,b_{42}+0,0469\,b_{43}\right)}{1+\exp\left(-1,6540+0,2612\,b_{11}-0,4487\,b_{12}+0,2486\,b_{13}+0,1235\,b_{14}+0,3779\,b_{21}+0,3335\,b_{23}-0,1147\,b_{32}+0,0623\,b_{33}-0,1661\,b_{41}-0,1823\,b_{42}+0,0469\,b_{43}\right)}$$

4.3. *Bernoulli Mixture Bayesian Regression*

The doodle of regression of Bernoulli Mixture Bayesian to model the Bidikmisi scholarship acceptance is presented in Figure 4.



(a) b_0 (b) b_{11}

(c) b_{12} (d) b_{13}

(e) b_{14} (f) b_{21}

Figure 3. Serial Plot of the first six posterior parameters b_0, b_{11}, b_{12}, b_{13}, b_{14}, b_{21} with 100.000 Iterations (10.000 thin 10) for Bayesian Binary Logistic Regression

Table 2. Parameters Estimation of Bayesian Binary Logistic Regression

Parameter	Dummy Parameter	Mean	Standar Deviasi	2,5%	97,5%	Significant
Constant		-1,6540	0,0851	-1,8210	-1,4880	√
X_1	b_{11}	0,2612	0,0492	0,1641	0,3562	√
	b_{12}	-0,4487	0,2133	-0,8687	-0,0315	√
	b_{13}	0,2486	0,0535	0,1448	0,3528	√
	b_{14}	0,1235	0,0526	0,0219	0,2259	√
X_2	b_{21}	0,3779	0,0776	0,2277	0,5306	√
	b_{22}	-	-	-	-	-
	b_{23}	0,3335	0,0827	0,1725	0,4967	√
	b_{24}	0,1656	0,0862	-0,0027	0,3375	No
X_3	b_{31}	-0,0587	0,0483	-0,1533	0,0356	No
	b_{32}	-0,1147	0,02042	-0,1549	-0,0751	√
	b_{33}	0,0623	0,0201	0,0232	0,1012	√
X_4	b_{41}	-0,1661	0,0455	-0,2556	-0,0758	√
	b_{42}	-0,1823	0,0198	-0,2208	-0,1437	√
	b_{43}	0,0468	0,0209	0,0060	0,0877	√

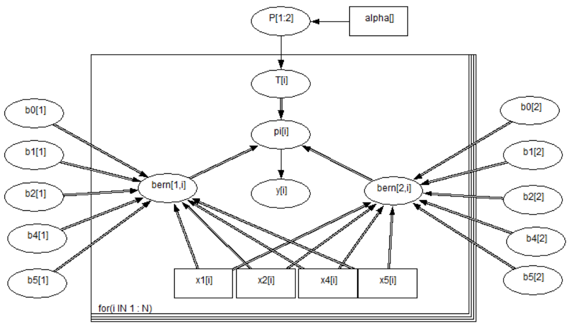


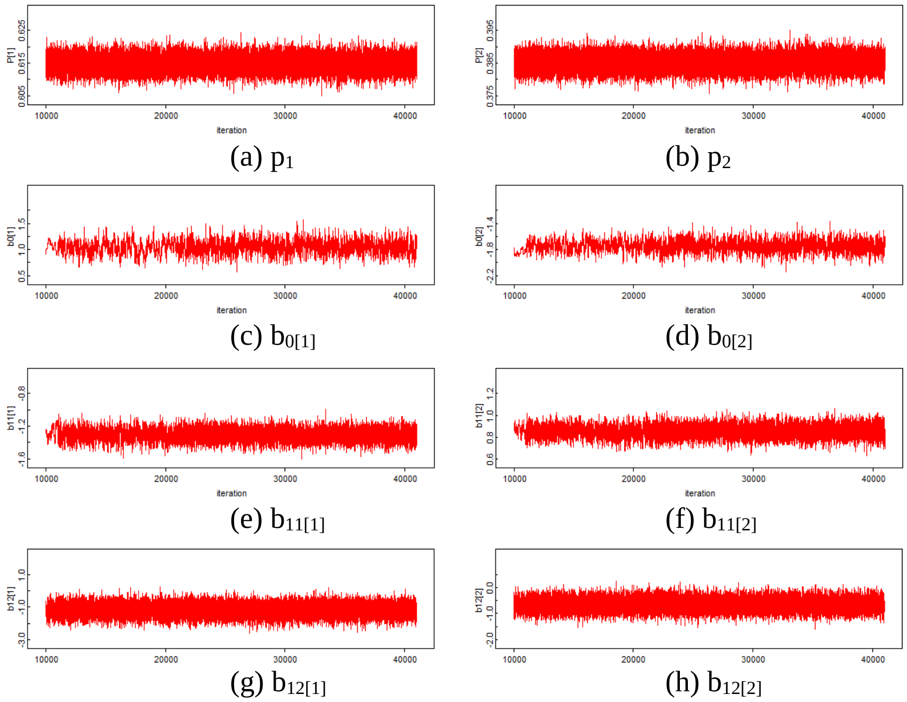
Figure 4. *Doodle* of Bernoulli Mixture Bayesian Regression

Figure 4 tells that the predicted parameters bern[1,i] and bern[2,i] have the same link function but are connected to different parameters. Node T[i] functions to separate data based on the acceptance condition of Bidikmisi that contains value 1 if the acceptance condition is false, and 2 if the acceptance condition is correct. Therefore, data in group 1 was analyzed on the bern[1,i] and data in group 2 was analyzed on the bern[2, i]. Prior parameters were normal distributions because the modeling method used a link function. Node P[1,2] was used to determine the proportions of group 1 and group 2 in the model. The total value of P[1] and P[2] is one. The parameter estimation is shown in Table 3.

Table 3. Parameters Estimation of Bernoulli Mixture Bayesian Regression

Parameter	Dummy Parameter	Mean	Standar Deviasi	2,5%	97,5%	Significant
P[1]		0,6148	0,0023	0,6103	0,6194	√
P[2]		0,3852	0,0023	0,3806	0,3897	√
b0[1]		1,0490	0,1256	0,8015	1,2930	√
b0[2]		-1,7490	0,0875	-1,9130	-1,5750	√
X_1	b_{11}[1]	-1,3060	0,0711	-1,4450	-1,1640	√
	b_{11}[2]	0,8516	0,0525	0,7482	0,9546	√
	b_{12}[1]	-1,1700	0,3535	-1,8760	-0,4949	√
	b_{12}[2]	-0,6488	0,2256	-1,0970	-0,2113	√
	b_{13}[1]	-1,1340	0,0792	-1,2890	-0,9791	√
	b_{13}[2]	0,5607	0,0591	0,4451	0,6775	√
	b_{14}[1]	-0,7730	0,0757	-0,9218	-0,6228	√
	b_{14}[2]	-0,0154	0,0578	-0,1298	0,0968	No
X_2	b_{21}[1]	-2,0240	0,1143	-2,2470	-1,7960	√
	b_{21}[2]	1,1030	0,0801	0,9451	1,2580	√
	b_{23}[1]	-1,5540	0,1213	-1,7920	-1,3140	√
	b_{23}[2]	0,5299	0,0879	0,3573	0,7015	√
	b_{24}[1]	-1,3330	0,1272	-1,5820	-1,0820	√
	b_{24}[2]	0,0041	0,0951	-0,1828	0,1900	No
X_3	b_{31}[1]	-0,6911	0,1137	-0,9169	-0,4709	√
	b_{31}[2]	0,0688	0,0560	-0,04112	0,1781	No
	b_{32}[1]	-0,2939	0,0378	-0,3692	-0,2201	√
	b_{32}[2]	-0,0917	0,0239	-0,1387	-0,0445	√
	b_{33}[1]	0,2806	0,0343	0,2139	0,3484	√
	b_{33}[2]	0,0415	0,0233	-0,0038	0,0870	No
X_4	b_{41}[1]	-0,2338	0,0963	-0,4225	-0,0467	√
	b_{41}[2]	-0,0854	0,0536	-0,1901	0,0197	No
	b_{42}[1]	-0,1130	0,0355	-0,1833	-0,0432	√
	b_{42}[2]	-0,1479	0,0228	-0,1921	-0,1028	√
	b_{43}[1]	0,0900	0,0354	0,0208	0,1598	√
	b_{43}[2]	0,0046	0,0241	-0,0429	0,0514	No

Historical serial sample values for parameters estimation of Bernoulli Mixture Bayesian Regression show a stable random pattern in a fixed domain. The historical sample series of the first 8 posterior estimation values (p_1, p_2, $b_{0[1]}$, $b_{0[2]}$, $b_{11[1]}$, $b_{11[2]}$, $b_{12[1]}$, $b_{12[2]}$) indicating the fulfillment of irreducible, aperiodic and recurrent properties as their convergence has been reached.



(a) p_1 (b) p_2

(c) $b_{0[1]}$ (d) $b_{0[2]}$

(e) $b_{11[1]}$ (f) $b_{11[2]}$

(g) $b_{12[1]}$ (h) $b_{12[2]}$

Figure 5. Serial Plot of the first 14 posterior estimation values (p_1, p_2, $b_{0[1]}$, $b_{0[2]}$, $b_{11[1]}$, $b_{11[2]}$, $b_{12[1]}$, $b_{12[2]}$) with 10.000 Iterations thin 30 for Bernoulli Mixture Bayesian Regression

4.4. Classification Comparison

The best combination of significant parameters for Bayesian Bernoulli mixture regression model and for Bayesian binary logistic regression model are compared to see which model can be more representative to explain the acceptance of the Bidikmisi data. The comparison is done using their classification values. The bigger classification percentage, the better the model explain the acceptance. Their classification results are tabled in Table 4. It can be seen that Bayesian Bernoulli mixture regression model with 71.70 percents shows more accurate than Bayesian binary logistic regression, which is just only able to accurately classify as big as 56.50 percents.

Tabel 4. Accepted Qualification Percentage

Model	% Classification
Bayesian binary logistic regression	56.50
Bayesian Bernoulli mixture regression	71.70

5. Conclusion

The new development approach, a Bayesian Bernoulli mixture regression model coupled with MCMC approach, has successfully worked to classify the Bidikmisi acceptance. This new development approach and the Bayesian binary logistic regression model can reach their posterior parameters estimation convergence perfectly. This proposed Bayesian Bernoulli mixture regression model coupled with MCMC approach can give a higher percentage of acceptance classification accuracy than the Bayesian binary logistic regression model. This model is more representative for explaining the classification of Bidikmisi acceptance.

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 PUPT research grant no. 608/PKS/ITS/2017.

References

[1] Wang X, Kabán A. Finding Uninformative Features in Binary Data. Intelligent Data Engineering and Automated Learning - IDEAL 2005. 2005; 3578: p. 40–47.

[2] Duda RO, Hart PE. Pattern Classification and Scene Analysis Wiley; 1973.

[3] Grim J, Pudil P, Somol P. Multivariate Structural Bernoulli Mixtures for Recognition. In ; 2000; Proceedings. 15th International Conference on, vol. 2, pp. 585–589.

[4] González J, Juan A, Dupont P, Vidal E, Casacuberta F. A Bernoulli Mixture Model for Word Categorization. In ; 2001; Benicassim, Spain: Proceedings of the IX Spanish Symposium on Pattern Recognition and Image Analysis.

[5] Juan A, Vidal E. On The Use of Bernoulli Mixture Models for Text Classification. Pattern Recognition. 2002; 35(12): p. 2705–2710.

[6] Juan A, Vidal E. Bernoulli Mixture Models for Binary Images. In ; 2004: Proceedings of the 17th International Conference on Pattern Recognition (ICPR’04).

[7] Patrikainen A, Mannila H. Sub Space Clustering of High-Dimensional Binary Data-A Probabilistic Approach. In ; 2004; Workshop on Clustering High-Dimensional Data and Its Applications: SIAM International Conference on Data Mining.

[8] Zhu S, Takigawa I, Zhang S, Mamitsuka H. A Probabilistic Model for Clustering Text Documents with Multiple Fields. In ; 2007; Berlin, Heidelberg: Advances in Information Retrieval, 29th European Conference on IR Research (ECIR2007).

[9] Sun Z, Rosen O, Sampson A. Multivariate Bernoulli Mixture Models with application to Postmortem Tissue Studies in Schizophrenia. Biometrics. 2007; 63: p. 901-909.

[10] Tikka J, Hollmen J, Myllykangas S. Mixture Modelling of DNA Copy Number Amplification Patterns in Cancer. In ; 2007; Berlin, Heidelberg: Proceedings of the 9th International Work-Conference on Artificial Neural Networks (IWANN' 2007), Springer-Verlag.

[11] Hollmen J, Tikka J. Compact and Understandable Descriptions of Mixture of Bernoulli Distributions. In ; 2007; Springer-Verlag, Berlin, Heidelberg: Proceedings of the 7th International Symposium on Intelligent Data Analysis (IDA2007).

[12] Myllykangas S, Tikka J, Böhling T, Knuutila S, Hollmén J. Classification of Human Cancers Based on DNA Copy Number Amplification Modelling. BMC Med. Genomics. 2008; 1: p. 1-13.

[13] Bouguila N. On Multivariate Binary Data Clustering and Feature Weighting. Comput. Stat.Data Anal. 2010; 54: p. 120-134.

[14] Saeed M, Javed K, Babri HA. Machine Learning Using Bernoulli Mixture Models: Clustering,Rule Extraction and Dimensionality Reduction. Neurocomputing. 2013; 119: p. 366–374.

[15] Direktorat Jenderal Pembelajaran dan Kemahasiswaan KRTdPT. Pedoman Penyelenggaraan Bantuan Biaya Pendidikan Bidikmisi Tahun 2016 Jakarta: Belmawa, Kemeristek Dikti; 2016.

[16] McLachlan G, Peel D. Finite Mixture Models New York: John Wiley and Sons, Inc.; 2000.

[17] Nadif M, Govaert G. Clustering for Binary Data and Mixture Models-Choice of The Model. Appl. Stochastic Models Data Anal. 1998; 13: p. 269-278.

[18] Casella G, George EI. Explaining Gibss Sampler. The America Statistical Association. 1992; 46(3): p. 167-174.

[19] Ntzoufras I. Bayesian Modeling Using WinBUGS New Jersey, USA: Wiley; 2009.