A Comparison of Ordinal Logistic Regression Analysis and Kernel Discriminant Analysis in Case of Classification of Higher Education in Indonesia

Amelia Crystine, Muhammad Nur Aidi, I Made Sumertajaya

Abstract— Ordinal logistic regression analysis is a classification method that provides a cumulative opportunity value model of the relationship between independent variables and dependent variables. Some connect functions are used to define the cumulative opportunity function of each classification group, where the use of this connect function is based on the data distribution. Discriminant analysis is a multivariate technique focusing on the separation of new objects at defined groups. Kernel discriminant analysis is used in the classification process with non-normal distributed independent variable characteristics. It focuses on the selection of smoothing parameters (bandwidth). The Normal Kernel discriminant analysis, which has the classification efficiency of ≈ 0.9512, was used in this study. In the case of Classification of Higher Education in Indonesia in 2015, the Normal Kernel discriminant analysis resulted in the classification accuracy which was higher than ordinal logistic regression analysis. It was 90.82% for training data and 88.98% for testing data with sensitive classification accuracy for each group. Large data diversity affected the accuracy of performance of a method in a classification process.

Index Terms— Classification, Ordinal Logistic Regression, Normal Kernel Discriminant

1 Introduction

In several studies, it is common to find that the response variables in the regression model are not numerical, but categorical. Logistic regression model is one model that can be used to overcome this. In the model of Ordinal Logistic Regression Analysis, the response variable has an ordinal scale, and the independent variables can be in the form of data with categorical and / or numeric scales [2]. There are 5 connect functions that can be used in logistic regression analysis, namely logit, complementary log-log, negative log-log, probit, and cauchit [5]. The use of these connect functions is based on each data distribution used. The logit connect function was used in the analysis of this study; therefore, the model that was formed was a cumulative opportunity that can indicate a tendency of an observation coming into a group. The probability for an observation to come into in a group is based on a relationship model between independent variables and dependent variables formed. Thus, this logistic regression model can also be used to measure the accuracy of a classification process.

Besides the logistic regression model, another analysis that can be used to measure the accuracy of classification is discriminant analysis. Discriminant analysis is a multi variable technique that focuses on object separation or observation and allocates new objects to defined groups [9]. In this case, the dependent variables used are nonmetric data (nominal or ordinal) while the independent variables are the metric data (interval or ratio). The approach of linear discriminant analysis is generally used in populations whose independent variables are multivariate normal distributions. However, it is common when the classification process is required in the data whose independent variables are not normally distributed. In this case, the nonparametric discriminant analysis of kernel discriminant method can be used because this analysis does not take into account certain assumptions [7]. There are several functions in kernel discriminant analysis: Epanechnikov, Biweight, Triangular, Gaussian (Normal) and Rectangular. The selection of kernel functions is based on the level of efficiency of the kernel functions. This study used a normal function kernel that has an efficiency of ≈ 0.9512 [7].

The case study in this study used data from Classification and Ranking of Universities in Indonesia in 2015. Based on academic texts published by the Ministry of Research, Technology and Higher Education of Indonesia, the classification and ranking activities should be able to fully understand the characteristics of data of Higher Education Database. Accordingly, the factors that influence the classification of Higher Education were analyzed using Ordinal Logistic Regression Analysis and Wilk's Lambda Test, and the classification accuracy between the method of logistic regression and the normal kernel discriminant analysis was compared.

2 Research Method

2.1 Data

Data of Classification and Ranking of Higher Education in Indonesia consist of 4 groups that become dependent variables

- *Amelia Crystine is currently pursuing masters degree program in statistics in Bogor Agriculture University, Indonesia, PH +6285218330863. E-mail: ameliacrystine@gmail.com*
- *Muhammad Nur Aidi is Lecturer, Departement of Statistics, Bogor Agriculture University, Bogor, Indonesia. E-mail: nuraidi@yahoo.com*
- *I Made Sumertajaya is Lecturer, Departement of Statistics, Bogor Agriculture University, Bogor, Indonesia. E-mail: imsjaya@yahoo.com*

(response variables) with 10 independent variables in accordance with the Academic Text of Classification and Ranking of Universities in Indonesia in 2015, which are:

X_1 ratio of the number of lecturers who have doctoral degree to the total number of lecturers
X_2 ratio of the number of Senior Lecturers and Professors to the total number of lecturers
X_3 ratio of the number of permanent lecturers to the total number of lecturers
X_4 the number of students
X_5 the score of accreditation
X_6 the ratio of the number of A and B accredited study programs to the total number of study programs
X_7 the number of achievements in the National Student Scientific Week
X_8 the score of research
X_9 the ratio of the number of scopus indexed documents to the number of permanent lecturers
X_{10} the ratio of the number of scopus indexed scientific articles to the number of permanent lecturers

2.2 Methods of Data Analysis

The steps of analysis were as follows:

1. The Multicolinearity test was conducted by using Pearson correlation test
2. The data were divided into 2 parts: Data Training, as many as 2431 universities, and Data Testing, as many as 608 universities
3. The Ordinal Logistic Regression Analysis was conducted
   a. The model of ordinal logistic regression analysis for each k-group was written as follows [2]:

$$logit\,[P(Y \le k)] = \log \frac{P(Y \le k)}{1-P(Y \le k)}$$
$$= \beta_{0k} + \boldsymbol{\beta}' \boldsymbol{x}_i = \beta_{0k} + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots$$

   b. The simultaneous assessment of all independent variables using the Likelihood Ratio test. The test statistic used was: G = -2 ln $[L_0/L_M]$ with L_0: likelihood function without independent variables; L_M : likelihood function with independent variables. If the value of G > $\chi^2_{(\alpha;db)}$ then it could be concluded that the model with independent variables were better than the model without independent variables.
   c. The partial assessment of all partial variables using Wald test. The test statistic used was $W^2 = \left(\frac{\hat{\beta}_t}{\widehat{SE}(\hat{\beta}_t)}\right)^2$ if the value of $W^2 > \chi^2_{(\alpha;db)}$ then it could be concluded that the independent variables affected the grouping.
   d. The grouping was done on the basis of calculating the cumulative opportunity value of the model of ordinal logistic regression analysis of each k-group:

$$\pi_k(\boldsymbol{x}_i) = \frac{e^{g_k(x_i)}}{1+e^{g_k(x_i)}}$$

$$g_k(\boldsymbol{x}_i) = \beta_{0k} + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

   e. Interpretation of beta coefficient value was conducted using odds ratio value. If the odds ratio is > 1 then the odds when X_1 was greater than the odds at X_2.
4. Kernel Discriminant Analysis was conducted
   a. The assessment of the independent variable effects on grouping was conducted using Wilk's Lambda Test
   b. The model of kernel discriminant analysis in each k-group was generally written as follows [4]:

$$\hat{f}_k(\boldsymbol{x}_0) = \frac{1}{n_k}\sum_{i=1}^{n_k} K_k(\boldsymbol{x}_0 - \boldsymbol{x}_i)$$

   c. The normal kernel function equation with an average of 0, and the range, $h^2\boldsymbol{V}_k$, was written as follows [4]:

$$K_k(\boldsymbol{z}) = \frac{1}{c_0(k)} exp\left(-\frac{1}{2h^2}\boldsymbol{z}'\boldsymbol{V}_k{}^{-1}\boldsymbol{z}\right)$$

   Where $\boldsymbol{z} = \boldsymbol{x}_0 - \boldsymbol{x}_i$
   d. The classification of the kernel discriminant analysis was based on the calculation of the probability value of the kernel discriminant model influenced by the smoothing parameter (bandwidth). The selection of smoothing parameters was based on the calculation of the optimum bandwidth calculation, which is [7]:

$$h = \left(\frac{A(K_k)}{n_k}\right)^{1/(p+4)}$$

   with:

$$A(K_k) = \frac{4}{2p+1}$$

   The optimum bandwidth value was used as the starting point of the trial process, resulting in the best classification accuracy at one particular bandwidth point.
   e. The grouping of observations was based on the posterior probability of each group for each observation [9]:

$$P(\pi_k|\boldsymbol{x}_0) = \frac{p_k\hat{f}_k(\boldsymbol{x}_0)}{\sum_{k=1}^{g} p_k\hat{f}_k(\boldsymbol{x}_0)}$$

   The observed unit was to be classified to the group with the greatest $P(\pi_k|\boldsymbol{x}_0)$ probability value.
5. The accuracy of the classification of each method was calculated

3 Result and Discussion

3.1 Overview of Data

The Classification and Ranking of Universities in Indonesia was based on the final score of the assessment of each indicator according to the academic texts. Table 1 shows an imbalance of the amount of data in each group.

TABLE 1
DATA DESCRIPTION OF FINAL SCORE

Group	N	Minimum	Maximum	Average	Variety
1	11	2,9779	3,7430	3,2706	0,2784

IJSER

2	55	2,0059	2,7534	2,2764	0,1994
3	644	1,0002	1,9830	1,3342	0,8486
4	2329	0,0040	0,9986	0,4971	0,2360

The imbalance data might cause the variety of each group to be not equal. This could be supported with Box's M test result that showed p-value of 0.000 which caused rejection of H_0 with the final conclusion that the variety of each group was not equal.

3.2 Multicollinearity Test

The assessment of multicollinearity assumptions can be reviewed using Pearson correlation. Pearson correlation resulted in some variables which had high correlation to other variables, namely the variable X_5 and X_6 that was 0,951 and variable X_9 and X_{10} that was 0,950. This indicated that there was multicollinearity in those independent variables. One of ways to overcome the multicollinearity is to remove one of the correlated variables, in this case variables X_5 and X_{10} were going to be removed in the next analysis.

3.3 Ordinal Logistic Regression Analysis

In this study, the initial stage in conducting Ordinal Logistic Regression Analysis was to change the scale of independent variables from metric to non-metric. Here is a description of the category of each independent variable:

TABLE 2
CATEGORY EXPLANATION OF ORDINAL LOGISTIC REGRESSION ANALYSIS

Variable	Category	Explanation	Dummy
X_1	0	There is / are lecturer(s) with doctoral degree	1
	1	There is no lecturer with doctoral degree	-1
X_2	0	There is / are lecturer(s) serving as senior lecturer(s) and professor(s)	1
	1	There is no lecturer serving as senior lecturer and professor	-1
X_3	1	The ratio of the number permanent lecturers to the total number of lecturers > 0,9677	1 0 0
	2	0,8333 < the ratio of the number of permanent lecturers to the total number of lecturers ≤ 0,9677	0 1 0
	3	0,6363 < the ratio of the number of permanent lecturers to the total number of lecturers ≤ 0,8333	0 0 1
	4	The ratio of the number of permanent lecturers to thetotal number of lecturers ≤ 0,6363	-1 -1 -1
X_4	1	The number of students > 1067,75	1 0 0
	2	353,5 < the number of students ≤ 1067,75	0 1 0
	3	123,25 < the number of students ≤ 353,5	0 0 1
	4	the number of students ≤ 123,25	-1 -1 -1
X_6	0	A and B accredited study programs	1
	1	A and B unaccredited study programs	-1
X_7	0	There is achievement in the National Student Scientific Week	1
	1	There is no achievement in the National Student Scientific Week	-1
X_8	0	There is achievement in research performance	1
	1	There is no achievement in research performance	-1
X_9	0	There is / are scopus indexed publication document(s)	1
	1	There is no scopus indexed publication document	-1

The assessment using the method of Ordinal Logistic Regression Analysis was based on training data. The test results showed that there were some independent variables that had no significant effect on dependent variables, namely variables X_1 and X_2 which had p-value greater than 0.05. Consequently, reassessment by eliminating the two independent variables was required.

After variables X_1 and X_2 were removed, the calculation of likelihood ratio test resulted in a conclusion that there was rejection of H_0 with G statistic value of 1565,8880 which was bigger than table value $\chi^2_{(0,05;10)}$ that was 18,307 and p-value less than 0,05. This indicated that the model with independent variables was better than the model without independent variables. Wald's test showed that all independent variables generally had p-value < 0.05. There were several categories of independent variables that had p-value > 0.05 but they were still incorporated into the model and the result could be interpreted. At last, the model that was formed was:

Logit [P(Y≤1)] =-7,3613+0,9015X_{31}+0,6526X_{32}-

IJSER

$0,0782X_{33}+0,0630X_{41}+0,1560X_{42}+0,6190X_{43}+2,9816X_{60}+0,9688X_{70}+0,4359X_{80}+1,8802X_{90}$

Logit $[P(Y\leq 2)] = -4,6975+0,9015X_{31}+0,6526X_{32}-0,0782X_{33}+0,0630X_{41}+0,1560X_{42}+0,6190X_{43}+2,9816X_{60}+0,9688X_{70}+0,4359X_{80}+1,8802X_{90}$

Logit $[P(Y\leq 3)] = -0,0201+0,9015X_{31}+0,6526X_{32}-0,0782X_{33}+0,0630X_{41}+0,1560X_{42}+0,6190X_{43}+2,9816X_{60}+0,9688X_{70}+0,4359X_{80}+1,8802X_{90}$

The coefficient of independent variables showed the change of response value with the change of value in the independent variables. Coefficients that were marked positive resulted in an addition to probability value, in contrast to the coefficients that were marked negative resulted in a reduction in probability value. However, in the Ordinal Logistic Regression Analysis with non-metric independent and dependent variables, coefficients might be mathematically difficult to interpret. Therefore odds ratio value was used to know the interpretation of each coefficient.

The odds ratio of Permanent Lecturers variable (X_3) was greater than 1 for each category, with category 4 as comparison category. This showed that universities with categorized 1 permanent lecturers had the probability of 10,776 times higher compared to universities with categorized 4 permanent lecturers to be grouped in high score universities. Universities with categorized 2 permanent lecturers had the probability of 8,402 times higher compared to universities with categorized 4 permanent lecturers to be grouped in high score universities. Universities with categorized 3 permanent lecturers had the probability of 4,046 times higher compared to universities with categorized 4 permanent lecturers to be grouped in high scores universities.

The odds ratio of Student variable (X_4) was greater than 1 for each category. Universities with categorized 1 students had the probability of 2,462 times higher compared to universities with categorized 4 students to be grouped in high score universities. Universities with categorized 2 students had the probability of 2,702 times higher compared to universities with categorized 4 students to be grouped in high score universities. Universities with categorized 3 students had the probability of 4,293 times higher compared to universities with categorized 4 students to be grouped in high score universities.

The odds ratio of Study Program variable (X_6) was greater than 1 indicating that A and B accredited universities had the probability of 388,845 times higher compared to A and B unaccredited universities to be grouped in high score universities.

The odds ratio of variable of achievement in the National Student Scientific Week (X_7) was 6,941 which showed that these universities had the probability of 6,941 times to be grouped in high score universities compared to universities which had no achievement in Scientific Week National Students.

The odds ratio of research variable (X_8) was 2.391, which indicated that universities that conducted research activities had the probability of 2,391 times higher compared to universities that did not conduct research activity to be grouped in high score universities.

The odds ratio of Scopus Document variable (X_9) was 42,965 which indicated that the scopus indexed research documents made the universities have the probability of 42,965 times to be grouped in high score universities.

Table 3 shows the accuracy of classification using ordinal logistic regression method :

TABLE 3
ACCURACY OF CLASSIFICATION OF ORDINAL LOGISTIC REGRESSION

Accuracy	*Training*	*Testing*
Whole	0,8893	0,8586
Group 1	0,7500	0,6667
Group 2	0,2759	0,3846
Group 3	0,7821	0,8352
Group 4	0,9252	0,9007

3.4 Normal Kernel Discriminant Analysis

The test results of Wilk's Lambda showed that all independent variables, namely X_1, X_2, X_3, X_4, X_6, X_7, X_8 and X_9 had p-value of 0.000 and caused rejection of H_0 which meant there was a factual average difference in each group for every independent variable. The normal kernel discriminant analysis is a kernel discriminant analysis with a normal function that has no functional limit; therefore the approach of data distribution that uses the normal kernel function is sort of easier than other kernel functions. The bandwidth value was obtained by using the optimal bandwidth equation as follows:

$$h_1 = \left(\frac{0,2353}{8}\right)^{1/(8+4)} = 0,7438$$

$$h_2 = \left(\frac{0,2353}{29}\right)^{1/(8+4)} = 0,6695$$

$$h_3 = \left(\frac{0,2353}{468}\right)^{1/(8+4)} = 0,5310$$

$$h_4 = \left(\frac{0,2353}{1926}\right)^{1/(8+4)} = 0,472$$

$$h_{opt} = \frac{0,7438 + 0,6695 + 0,5310 + 0,472}{4} = 0,6045 \approx 0,6$$

From the above calculation, the optimal bandwidth value obtained was 0.6, which was the initial bandwidth of trial and error process. In the calculation process using kernel discriminant analysis, the observations that could not be classified for each bandwidth value were not found. However, it became known that bandwidth 0.1 resulted in the best classification accuracy for data training and data testing. Table 4 shows the accuracy of the classification as a whole and in each group was good, however, in the data testing, it appears that the accuracy of classification of group 1 was 0 because the samples taken in the data testing were too few, to be exact as many as 3 observations allocated to group 2.

TABLE 4
ACCURACY OF CLASSIFICATION OF NORMAL KERNEL BANDWIDTH

0.1

Accuracy	*Training*	*Testing*
Whole	0,9083	0,8898
Group 1	1	0
Group 2	0,7241	0,7692
Group 3	0,7479	0,8181
Group 4	0,9496	0,9354

In general, data of Classification and Ranking of Higher Education in 2015 can be classified by the method of Ordinal Logistic Regression Analysis and Normal Kernel Discriminant Analysis. Both methods resulted in different classification accuracy, either as a whole or as a group. Here is a summary of the classification accuracy of both methods:

TABLE 5
SUMMARY OF DATA ACCURACY OF CLASSIFICATION AND RANKING OF HIGHER EDUCATION IN 2015

Accuracy	*Training*	*Testing*
Ordinal Logistic Regression Analysis	88,94%	85,86%
Normal Kernel Discriminant Analysis	90,82%	88,98%

Table 5 shows that the method of Ordinal Logistic Regression Analysis and the method of Normal Kernel Discriminant Analysis had high overall classification accuracy. However, when viewed from the accuracy of each group, the normal kernel discriminant analysis provided more sensitive accuracy values than the Ordinal Logistic Regression Analysis for the case of data of Classification and Ranking of Higher Education with the characteristics of data that were unbalanced and the varieties that were significantly different.

4 CONCLUSION

Based on the discussion of case studies of Classification and Ranking of Higher Education in 2015 it can be concluded that:

a. The independent variables that affected the classification of the method of Ordinal Logistic Regression Analysis were: the ratio of the number of permanent lecturers to the total number of lecturers, the number of students, the ratio of the number of accredited A and B study programs to the total number of study programs, the number of achievements in the National Student Scientific Week, scores of research and the ratio of the number of scopus indexed documents to the number of permanent lecturers. Meanwhile, based on Wilk's Lambda test, the independent variables that affected the average group were: the ratio of the number of lecturers with doctoral degree to the total number of lecturers, the ratio of the number of senior lecturers and professors to the total number of lecturers, the ratio of the number of permanent lecturers to the total number of lecturers, the number of students, the ratio of the number of accredited A and B study programs to the total number of study programs, the number of achievements in the National Student Scientific Week, the score of research, and the ratio of the number of scopus indexed documents to the number of permanent lecturers. Both methods resulted in different assessment on independent variables, but the results of the Ordinal Logistic Regression Analysis can be interpreted more broadly and in detail based on the categories formed.
b. Normal kernel discriminant analysis resulted in the best classification accuracy compared to Ordinal Logistic Regression Analysis for Higher Classification and Ranking Data in 2015 that was 90.82% for training data and 88.98% for testing data. The normal kernel discriminant analysis also provided good classification accuracy for each group.

REFERENCES

[1] Agresti A. 2010. Analysis Of Ordinal Categorical Data. 2nd Edition. New Jersey (US):John Wiley and Sons.
[2] Hosmer D W, Lemeshow S, Sturdivant R X. 2013. Applied Logistic Regression. 3rd Edition. New Jersey (US): John Wiley and Sons.
[3] [Kemenristekdikti] Kementerian Riset, Teknologi, dan Pendidikan Tinggi. 2015. Naskah Akademik Klasifikasi dan Pemeringkatan Perguruan Tinggi Indonesia Tahun 2015. Jakarta (ID):Kemenristekdikti.[in Indonesian Language]
[4] Khattree R, Naik D N. 2000. Multivariate Data Reduction and Discrimination with SAS® Software. North Carolina (US) : SAS Institute Inc.
[5] Norusis MJ. 2010. SPSS Statistics Guides: Ordinal Regression. http://www.norusis.com/pdf/ASPC_v13.pdf
[6] Republik Indonesia. 2012. Undang-Undang Republik Indonesia Nomor 12 Tahun 2012 Tentang Pendidikan Tinggi. Jakarta (ID):Kemenristekdikti. [in Indonesian Language]
[7] Silverman B W. 1986. Density Estimaton for Statistics and Data Analysis. London (US):Chapman and Hall.
[8] Tarno. 2008. Estimasi Model Untuk Data Dependen Dengan Metode Cross Validation. Media Statistika 1(2):75-82. [in Indonesian Language]
[9] Johnson, Wichern. 2007. Applied Multivariate Statistical Analysis. 6th Edition. London (US):Pearson Education.