Classification Spectrum Data of Non-Invasive Blood Glucose Level Measure Using Ordinal Logistic Regression Method on Unbalanced Data

Dwi Retno Puspita Sari[1], Erfiani[2], Bagus Sartono[3],

Abstract— The use of this classification method tends to be effective in the case of data with a balanced grouping. In the classification method, sometimes for a researcher to find an unbalanced data condition in his research. Whereas in the case of unbalanced data is necessary to approach other statistical methods before reaching the stage of logistic regression. One statistical method for handling unbalanced data class cases is SMOTE. In this research used a spectrum data of non-invasive glucose levels. Ordinal logistic regression method on unbalanced data using SMOTE results accuracy is 53% for training data and 56% for testing data.

Index Terms— Classification,Glucose,Logistic Regression,Non-Invasive, Ordinal Logistic Regression,SMOTE, Unbalanced Data

1 Introduction

Logistic regression is one of the classification methods using a parametric procedure. The use of logistic regression analysis is related for the categorical data response, both nominal and ordinal with independent variables as continuous or categorical [2]. In logistic regression analysis for ordinal responses, the final model is formed as the cumulative logit function opportunities with differentiation between classes in the form of opportunity values for each class [6]. In the case of ordinal response data more than two classes are called multi-class. Ordinal logistic regression is only effective for the balanced data case. While for the unbalanced data case, other statistical method is needed before reaching the stage of ordinal logistic regression. One statistical method to handle unbalanced data cases, i.e. synthetic minority oversampling technique (SMOTE).

In this research aims to apply ordinal logistic regression method to classify spectrum data of non-invasive blood glucose level measure devices according to low, normal, and high blood glucose levels in unbalanced data case.

2 Background

2.1 Diabetes

Diabetes mellitus (DM) is a disease that affects many people and cannot be cured. The number of diabetics in the world was 425 million with more than 10 million people living in Indonesia [1]. Therefore, it is important for diabetics to monitor the condition of their glucose levels regularly, including in the low, normal or high categories. Nowadays the popular method for checking blood glucose levels is invasive method, which takes the patient's blood sample and checks its glucose contents with a spectrophotometer. A patient who is positive for diabetes mellitus requires at least 4 times a day to check his blood glucose level [9]. In this case, it is very important to develop non-invasive measurement tools without needed to injure parts of body from DM patients.

2.2 Ordinal Logistic Regression Analysis

A statistical method analysis that is widely applied to nominal response data case is logistic regression. In ordinal data responses more than two groups were used ordinal logistic regression. Ordinal regression analysis method connects between one or more independent variables in the form of numerical and / or categorical to the response variable with two or more ordinal categories. The model used in the analysis is a cumulative logit model. In this logit model, the ordinal category of response variable is a cumulative opportunity, so if there is an ordinal category in the response variable Y where Xi is a predictor variable as much as p, then the cumulative opportunity to j is stated as follows [2]:

$$P(Y \leq j|X_i) = \pi_1(X_i) + \pi_2(X_i) + \cdots + \pi_j(X_i)$$

where $j = 1,2,\ldots,J, i = 1,2,\ldots,p$

Cumulative logit model obtained by comparing $P(Y \leq j|X_i)$ to $P(Y > j|X_i)$ for $j = 1,2,\ldots,J-1$ is [3]:

$$logit\,[P(Y \leq j|X_i)] = \log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right)$$

$$= log\left(\frac{\pi_1(X_i) + \pi_2(X_i) + \cdots + \pi_j(X_i)}{\pi_{j+1}(X_i) + \pi_{j+2}(X_i) + \cdots + \pi_J(X_i)}\right)$$

$$= \alpha_j + \sum_{i=1}^{p} \beta_i X_i$$

- *Dwi Retno Puspita Sari is currently pursuing master degree program in applied statistics in Bogor Agriculture University, Indonesia, PH +6285797320196. E-mail: dwiretno.ps@gmail.com*
- *Erfiani is Lecturer, Department of Statistics, Bogor Agriculture University, Bogor, Indonesia. E-mail: erfiani_ipb@yahoo.com*
- *Bagus Sartono is Lecturer, Department of Statistics, Bogor Agriculture University, Bogor, Indonesia. E-mail: bagusco@gmail.com*

In the ordinal logistic regression model, the model obtained only reaches the category J-1 because the cumulative opportunity of the last category is 1 with the assumption possibility of all categories. By using the cumulative probability equation above, the probability of each category of response variables can be obtained through the equation:

$$\pi_j = \frac{\exp(\alpha_j+\sum_{i=1}^{p}\beta_i x_i)}{1+exp\,(\alpha_j+\sum_{i=1}^{p}\beta_i x_i)} - \frac{\exp(\alpha_{j-1}+\sum_{i=1}^{p}\beta_i x_i)}{1+exp\,(\alpha_{j-1}+\sum_{i=1}^{p}\beta_i x_i)}$$

The opportunity value obtained from the equation is used as a classification guideline with the criteria for an observation into a category j based on the value of π_j to a certain extent [7].

2.3 Unbalanced Data

Ordinal logistic regression is only effective for the balanced data case. While for the unbalanced data case, other statistical method is needed before reaching the stage of ordinal logistic regression. Imbalance classes in classification method happened when there is a smaller data class than the other class. Classes with smaller data amount are called minority classes, while classes with higher data are called majority class. According to [8], the impact caused by imbalances class is the predicted results obtained to be unstable because the prediction leads to the majority class. Based on the case, it is necessary to handling of imbalances class in the classification method.

The problems with unbalanced data are often found in more than two class or multi-class data case. In binary data class, the problem of unbalanced data can be overcome by using a procedure at the data level, such as by increasing the amount of minor data classes (oversampling) or reducing the amount of major data classes (undersampling), while in multi-class data, it can be handled by approaching at the algorithm level [10].

2.4 Synthetic Minority Oversampling Technique

In the unbalanced data case, there are two approaches that can be used to handling these unbalanced data problems as procedure at the algorithm level and data [12]. One method at the algorithm level that is often used to handling unbalanced data class is synthetic minority oversampling technique (SMOTE).

The approach using the SMOTE method according to [5] is by generating data based on differences in data of minority classes with the nearest neighbors of the minority class into new synthetic data. The procedure of generating data performed on the SMOTE method is by calculating the difference data from the minority class that will be generated with the k-nearest neighbors, then multiplying the value obtained by random numbers 0 to 1 and added to the initial data [11]. For the mathematically, the procedure of artificial data using the SMOTE method is written as follows [4]:

$$s = x + u\,(x^R - x)$$

where:

u : weight of random number [0,1]
x^R : nearest neighbour data
x : minority data classes
s : new data

3 Method

3.1 Data

This research used primary data which are part of research development and clinical trial of non-invasive blood glucose level measure tool. The data were collected on 5-10 December 2016 of 118 respondents at the Laboratory of Biochemistry of IPB. The first step of data was collected by measuring respondent's height and weight, and then respondents were measured by non-invasive tool. The predictor variable (X) is the spectrum data of non-invasive tool in the form of residual spectrum data intensity of light. The residual intensity describes as intensity of light is continued and captured by the sensor. The spectrum produced by non-invasive devices forms as two peaks of residual intensity.

Standard deviation of residual intensity from every peak is used as the data approach of predictor variable. The standard deviation can describe how much variation intensity of light is captured by the sensor. Measurements were taken with 5 repetitions in one respondent, so there is 10 variables would be obtained to use as predictor variables (X). For the final stage, respondents were taken 4 ml of blood samples from the veins then measured it by invasive blood glucose device in Prodia Laboratory. The result of invasive measure used as a response variable (Y) are the grouping of glucose levels in low, normal, and high categories.

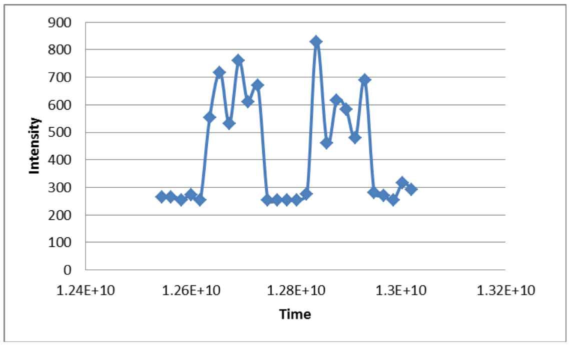

Fig. 1 Spectrum data of non-invasive device

3.2 Data Analysis

The stages in this research are:

1. Descriptive tables and histograms of invasive measurement.
2. Diagram the number of respondents in each category.
3. Divide the data randomly into 75% of training data and 25% testing data.

4. Resolve the unbalanced classes with SMOTE method on training data for low and high blood glucose levels.
5. Determine the estimator of ordinal logistic regression model on training data for 100 iterations with different result of SMOTE.
6. Calculate the classification accuracy of the training data obtained from 100 iterations iterations with different result of SMOTE.
7. Calculate the classification accuracy of testing data that has been applied to ordinal logistic regression model for 100 iterations with different result of SMOTE.
8. Calculate the classification accuracy of testing data that has been applied to ordinal logistic regression.

4 Discussion

4.1 Exploration

Table 1 shows that the results of invasive measurements from 118 respondents in this research, the lowest blood glucose level of respondent was 67 mg/dL with the highest blood glucose level was 276 mg/dL. Overall, the average blood glucose level of the respondents in this research was 82.64 mg/dL with the majority of glucose levels in the range of 80 mg/dL and most of respondent's glucose levels were 78 mg/dL. The slope of degree value 7.52 is more than zero, so tends of the data spread to the right.

TABLE 1
DESCRIPTIVE TABLE OF INVASIVE MEASUREMENT

Parameter	Min	Q1	Med	Mean	Q3	Max	Mod	Sknews
Kadar Glukosa	67	76	80	82.64	83	276	78	7.52

Based on Figure 2, the graphic of data distribution is not symmetrical. It can be seen from the distribution of invasive measurement data that spreads to the right. There is a data as high glucose levels far from other data distribution at the end of the charts. The average value of the data is also on the right side when a graphic of data distribution spreads to the right and it is higher than the median and mode data. The right tilt curve also shows that the graph of the frequency distribution has a positive slope.

The bar chart in Figure 3 shows the number of respondents in each category of low, normal and high blood glucose levels based on invasive measurement results. At low and high blood glucose levels, the number of respondents used as research samples respectively was 8 people. While the number of respondents in normal blood glucose levels who participated in this research was 102 people. So it can be concluded that the data is categorized as unbalanced data case because the number of respondents in low and high glucose levels is too much different than the number of respondents in normal blood glucose levels

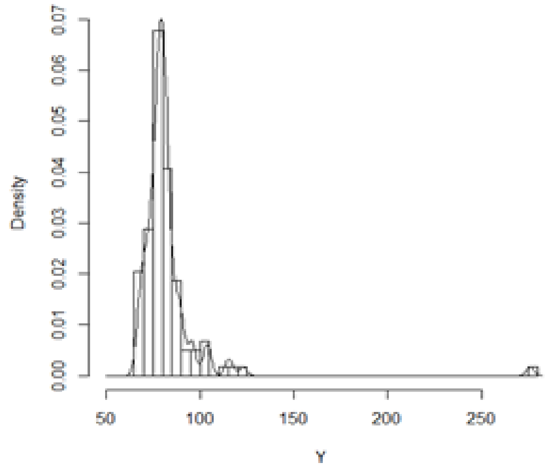

Fig. 2 Graphic of distribution data of invasive measurement result

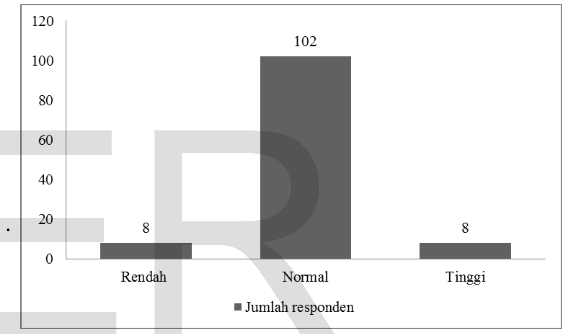

Fig. 3 Histogram of the number of respondents each category

4.2 Ordinal Logistic Regression Analysis with SMOTE

The SMOTE method is used because the data in this research is unbalanced. The initial data used consist of 3 data classes with 2 classes of minor data. In the minor class, the low and high blood glucose levels consisted of 8 respondents. As the first step is divide the data randomly into 75% of training data and 25% of testing data. The next stage is applied the SMOTE method to the training data. Both of minority data class in the training data group is duplicated to 36 data with a percentage of 500. In the majority data or normal blood glucose level there is no reduction number of respondents with the aim to avoid the reduction of information contained in normal data class. Final data is used in training groups for low, normal and high classes i.e. 36, 76 and 36.

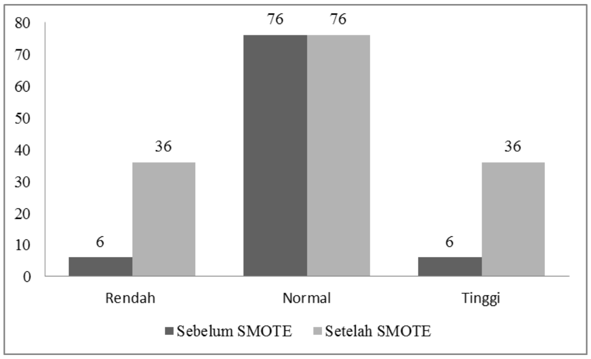

Fig. 4 Number of respondents before and after SMOTE

In this applied of model, all variables are used in the equation without an elimination system with the aim of being able to compare the classification accuracy values of ordinal logistic regression method with SMOTE and ordinal logistic regression method without SMOTE. The analysis phase is applied using 100 iterations with different result of SMOTE then calculates the mean of the average accuracy from the iterations. From the results of the analysis by applying the SMOTE method to ordinal logistic regression method, the classification accuracy of the training data was 53% and the testing data was 56%. One of the model equation obtained was

$$Logit[P(Y \le Rendah|X_i)] = -1{,}246 + 0{,}026\,X_1 - 0{,}025\,X_2 + 0{,}012\,X_3 - 0{,}025\,X_4 + 0{,}078\,X_5 + 0{,}017\,X_6 - 0{,}002\,X_7 - 0{,}013\,X_8 - 0{,}006\,X_9 - 0{,}037\,X_{10}$$

$$Logit[P(Y \le Normal|X_i)] = 2{,}456 + 0{,}026\,X_1 - 0{,}025\,X_2 + 0{,}012\,X_3 - 0{,}025\,X_4 + 0{,}078\,X_5 + 0{,}017\,X_6 - 0{,}002\,X_7 - 0{,}013\,X_8 - 0{,}006\,X_9 - 0{,}037\,X_{10}$$

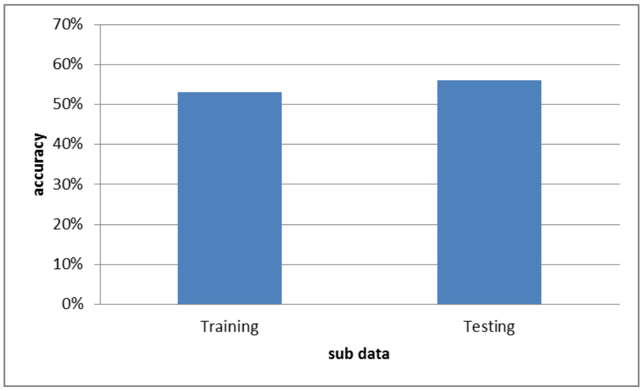

Fig.5 Histogram of classification accuracy

5 Conclusion

5.1 Appendices

The classification of Ordinal logistic regression with SMOTE using 100 iterations with different result of SMOTE then calculates the mean of the average accuracy from the iterations. Accuracy was obtained 53% for training data and 56% for testing data. Applied ordinal logistic regression method with SMOTE to classify spectrum data of non-invasive blood glucose level measuring device in low, normal, and high blood glucose level One of the model equation obtained was obtained by the model equation

$$Logit[P(Y \le Rendah|X_i)] = -1{,}246 + 0{,}026\,X_1 - 0{,}025\,X_2 + 0{,}012\,X_3 - 0{,}025\,X_4 + 0{,}078\,X_5 + 0{,}017\,X_6 - 0{,}002\,X_7 - 0{,}013\,X_8 - 0{,}006\,X_9 - 0{,}037\,X_{10}$$

$$Logit[P(Y \le Normal|X_i)] = 2{,}456 + 0{,}026\,X_1 - 0{,}025\,X_2 + 0{,}012\,X_3 - 0{,}025\,X_4 + 0{,}078\,X_5 + 0{,}017\,X_6 - 0{,}002\,X_7 - 0{,}013\,X_8 - 0{,}006\,X_9 - 0{,}037\,X_{10}$$

5.2 Suggestion

The suggestion for the further research is to use other Classification methods for unbalanced data case to classify the output spectrum data of non-invasive blood glucose levels to determine the possibility for better classification methods than ordinal logistic regression method and SMOTE method.

References

[1] International Diabetes Federation," IDF Diabetes Atlas Eighth Edition 2017," (http://diabetesatlas.org/ on 2018 Mei 18)

[2] A. Agresti, *Categorical Data Analysis*. New York: John Wiley and Sons, Inc. 2002.

[3] R. Azen , C. M. Walker, *Categorical Data Analysis for the Behavioral and Social Science*. New York: Routledge Taylor and Francis Group. 2011.

[4] R. Blagus, L. Lusa," SMOTE for High-dimensional Class-imbalanced Data," *Research Article BMC Bioinformatics 2013, 14:106*, 2013.

[5] V. N. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer,"SMOTE: Synthetic Minority Over-Sampling Technique," *J of Artif Intell Research*, 16: 321-357, 2002.

[6] B. W. Otok, M. S. Akbar, S. Guritno, Subanar,"Pendekatan Bootstrap pada Klasifikasi Pemodelan Respon Ordina," *Jurnal Ilmu Dasar*, Vol. 8 No.1, 54-67, 2007.

[7] R. T. Putranto, M. Mashuri,"Analisis Statistik Tentang Faktor-faktor yang Mempengaruhi Waktu Tunggu Kerja Fresh Graduate di Jurusan Statistika Institut Teknologi Sepuluh November (ITS) dengan Metode Regresi Logistik Ordinal," *Jurnal Sains dan Seni ITS* Vol.1, No.1, September 2012.

[8] Y. Sanguanmak, A. Hanskunatai,”Auto-Tuning of Parameters in Hybrid Sampling Method for Class Imbalance Problem,” *International Computer Science and Engineering Conference (ICSEC) ,* pp 1-5, 2016.

[9] E. Satria, Wildian,”Rancang Bangun Alat Ukur Kadar Glukosa Darah Non-invasif Berbasis Mikrokontroler AT89S51 dengan Mengukur Tingkat Kekeruhan Spesimen Urine Menggunakan Sensor Fotodioda,” *Jurnal Fisika Unand,* Vol.2, No.1, 2013.

[10] Y. Sun, A. K. C. Wong, M. S. Kamel,”Classification Of Imbalanced Data: A Review,” *Int J of Patt Recogn and Artif Intell*. Vol.23, No.4, 687-719, 2009.

[11] S. Wang, X. Yao,” Diversity Analysis on Imbalanced Data Sets by Using Ensemble Models,” *Symp. Computational Intelligence and Data Mining IEEE,* ISBN: 978-1-4244-2765-9, 2009.

[12] Z. Z. Zhang, Q. Chen, S.F. Ke, Y.J. Wu, F. Qi, Y. P. Zhang,”Ranking Potential Customers Based on Group Ensemble,” *Int J of Data Warehousing and Mining,* pp 79-89, 2008.