Classification accuracy on the family planning participation status using kernel discriminant analysis

Dian Kurniawan¹, Suparti¹, Sugito¹

¹Department of Statistics, Faculty of Science and Mathematics, Diponegoro University
Jl. Prof. Soedharto, SH, Tembalang, Semarang 50275, Indonesia
E-mail: diankurniawan@gmail.com

Abstract. Population growth in Indonesia has increased every year. According to the population census conducted by the Central Bureau of Statistics (BPS) in 2010, the population of Indonesia has reached 237.6 million people. Therefore, to control the population growth rate, the government hold Family Planning or Keluarga Berencana (KB) program for couples of childbearing age. The purpose of this program is to improve the health of mothers and children in order to manifest prosperous society by controlling births while ensuring control of population growth. The data used in this study is the updated family data of Semarang city in 2016 that conducted by National Family Planning Coordinating Board (BKKBN). From these data, classifiers with kernel discriminant analysis will be obtained, and also classification accuracy will be obtained from that method. The result of the analysis showed that normal kernel discriminant analysis gives 71.05 % classification accuracy with 28.95 % classification error. Whereas triweight kernel discriminant analysis gives 73.68 % classification accuracy with 26.32 % classification error. Using triweight kernel discriminant for data preprocessing of family planning participation of childbearing age couples in Semarang City of 2016 can be stated better than with normal kernel discriminant.

1. Introduction

Indonesia is one of the most populous countries in the world. Based on the data from BPS RI in 2010, the population of Indonesia has reached 237.6 million people. This significant increase in population is inseparable from the high presence of fertility rates in Indonesia and of course, this is a serious problem that needs to be handled immediately, even by Indonesia. Fertility is one of the factors that affecting population growth in addition to mortality and migration, because fertility is a contributing factor to the high rate of birth. High population growth leads to another problem in the other fields such as social, economic, health, education, and enviromental problems. The government have done various efforts and solutions to overcome problems arising from the rate of population growth, one of their solution are family planning program aimed for couples of childbearing age by the use of contraceptives [1]. Family planning program is an attempt to regulate the number of births in such a way for the mother and her baby and for the father or his family or the concerned society so it will not cause any harm as a direct result of the birth. The purpose of family planning program is to improve the welfare of mothers and children in order to manifest prosperous society by controlling births while ensuring control of population growth [2].

Several studies on family planning program that have done before, mentioning that the variables that affecting the participation of childbearing age couples in family planning program are the mother’s age, last child’s age, has ever get counseling about family planning program from the
authorities or not, welfare status, number of children, and visit frequency from authorities. Meanwhile, to analyze childbearing age couples that participating in family planning program or not, kernel discriminant analysis can be used, because the dependent variable (family planning program participation) is qualitative. The independent variables are used in this study consists of husband’s age, wife’s age, last child’s age, number of children and then the discriminant analysis aims to separate data and allocate observation objects into groups so that each object becomes a member of one group and no object becomes a member of more than one group. Discriminant analysis with a flexible nonparametric method is a kernel discriminant analysis because it does not have to meet certain assumptions as in parametric discriminant analysis [3]. Therefore, in this study, kernel discriminant analysis will be used for identifying the status of family planning program participation and measuring the classification accuracy.

2. Methodology

2.1. Parametric and Nonparametric Discriminant Analysis
Discriminant analysis is one method in multivariate analysis with dependency method (where the relationship between variables can be distinguished between the response variable and the predictor variable). This means, there are variables which their result depends on the predictor variable data. Discriminant analysis is used in cases where predictor variables are metric data (interval or ratio) and response variables are nonmetric data (nominal or ordinal). Grouping on discriminant analysis is mutually exclusive, meaning that if an observation has entered one of the groups, it is not a member of another group [4].

Discriminant analysis approach is very diverse, from the parametric and nonparametric approach. But in the practice, the parametric discriminant approach often involves variables that do not follow the normal distribution pattern, resulting in a non-optimal discriminant classification result [5]. Therefore, to overcome it, discriminant analysis of nonparametric approach with kernel method which serves to classify data can be used without considering certain assumptions [6]. For example, $X_1, X_2, ..., X_{n_t}$ is a random sample of $\pi_t$ population and $x$ is an additional observation of the $\pi_t$ population, which probability density function $f_t(x)$ is unknown. The probability density function $f_t(x)$ can be estimated by:

$$\hat{f}_t(x) = \frac{1}{n_t} \sum_{i=1}^{n_t} K_t(x - x_i)$$

where the quantity of $K_t(z)$ is called the kernel function of the t group.

Some of the multiple variable kernels that often used in the probability density function $f_t(x)$ are:

- Normal kernel (mean 0, varian $h^2 \cdot S_t$)
  $$K_t(z) = \frac{1}{C_0(t)} \exp \left(- \frac{1}{h^2} z^T S_t^{-1} z \right)$$

  with $C_0(t) = (2\pi)^{d/2} h^d |S_t|^{1/2}$

- Triweight Kernel
  $$K_{(1)}(z) = \begin{cases} 
  C_3(t) \left( \frac{1 - z^T S_t^{-1} z}{h^2} \right)^3 & : \text{if } z^T S_t^{-1} z \leq h^2 \\
  0 & : \text{for others}
  \end{cases}$$
with $C_3(t) = \left(1 + \frac{d}{4}\right)\left(1 + \frac{d}{4}\right)\left(1 + \frac{d}{2}\right)\frac{\pi^d h^d}{\pi(1+\frac{d}{2})}\left|S_t\right|^\frac{1}{2}$

where $z = x - x_i$, $t=1,2,...,g$ denotes the $t^{th}$ group, $h$ denotes the value of the smoother parameter (bandwidth), $d$ denotes the size of the independent variable, and $S_t$ is the covariant variant matrix in the $t$ group. The selection of bandwidth values can be done subjectively by trying various bandwidth options on the data (trial and error). The optimal bandwidth value is the bandwidth value that produces the highest classification accuracy [7].

Classification rules on kernel discriminant analysis are using the Bayes theorem based on the largest posterior probabilities. Based on the probability density function, then the posterior probability of the $x$ group can be calculated. For example, $\hat{f}_t(x)$ is the kernel function estimator of the $\pi_t$ group and $P_t$ is the initial chance of the $\pi_t$ group, for $t=1,2,...,g$, then the posterior probability of an $x$ observation comes from the $\pi_t$ group are [3]:

$$P(\pi_t|x) = \frac{P_t\cdot\hat{f}_t(x)}{\sum_{i=1}^{g} P_t\cdot\hat{f}_t(x)}$$

where $P_t$ is obtained from

$$P_t = \frac{n_t}{\sum_{i=1}^{g} n_t}$$

observation $x$ is allocated to $\pi_t$, which having the largest posterior probability.

2.2. Classification Accuracy

An error classification in an observation is most likely to occur. This is happened sometimes because there are some observations that do not belong to a particular group but put in that group. The calculation of APER (Apparent Error Rate) value can be done by using confusion matrix as follows [3]:

$$APER = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}}$$

From the calculation of APER value that has been described, to find the value of classification accuracy can be used $1 - APER$.

2.3. Type of Data, Data Source, and Research Variable

The data used in this study is secondary data obtained from family data collection conducted by BKKBN Central Java Province. The data is childbearing age couples data based on their status as the participant of family planning program or not, that taken from January to June 2016. And then the dependent variable (Y) that used in this study is the Family Planning Program participation status which categorized by notation 1 for childbearing age couples not participating in family planning program and notation 2 for childbearing age couples participating in family planning program. While the independent variables are husband’s age ($X_1$), wife’s age ($X_2$), last child’s age ($X_3$), and the number of children ($X_4$).

2.4. Step of Data Analysis

Data analysis is done with the following steps:

1) Input the data of childbearing age couples who participating in family planning program or not as response variable and husband’s age, wife’s age, last child’s age, and the number of children as the predictor variable.
2) Do the descriptive analysis of childbearing age couples data based on family planning program participation status.
3) Divide data into training and testing data.
4) Perform discriminant analysis of normal kernel and triweight kernel with steps as follows:
   a) Specifies the kernel density function.
   b) Enter the value of the smoothing parameter or bandwidth h.
   c) The classification of childbearing age couples with varying bandwidth values.
   d) Calculate the size of classification accuracy based on APER criteria.
   e) Repeat the step (b) – (d) several times for different values of h.
   f) Choosing the optimal bandwidth based on the size of classification accuracy.
5) Compare the classification accuracy that has been calculated based on the normal kernel discriminant with the triweight kernel discriminant on the family planning program participation status.

3. Experiment Result
3.1. Descriptive Analysis of Family Planning Program Participation Data in Semarang City
Descriptive analysis is used to obtain an overview of data in general. The data used in this study are 379 data of childbearing age couples with the percentage of 68.07% of them are not participating in family planning program, while the remaining 31.93% are participating in family planning program. For a summary, can be seen in the following table:

| Status                                      | Frequency | Percentage |
|---------------------------------------------|-----------|------------|
| Childbearing Age Couples who do not participate in Family Planning Program | 258       | 68.07      |
| Childbearing Age Couples who participate in Family Planning Program        | 121       | 31.93      |
| Total                                                  | 379       | 100.0      |

Furthermore, the author divides the data into 2 parts, namely training and testing data, before using the normal kernel and triweight kernel discriminant analysis. Training data is the data used for the formation of probability density function, while testing data is the data used in validating the function that has been generated by training data. This study used 341 pairs of childbearing age couples as training data and 38 pairs of childbearing age couples as testing data.

3.2. Kernel Discriminant Analysis
In this section, the normal and triweight kernel functions are used. There are two groups for response variable Y, that is \( t = 1,2 \). Group 1 represent childbearing age couples that not participating in family planning program and group 2 represent childbearing age couples that participating in family planning program. The sample sizes in the testing data for each group are :
\[ n_1 = 26 \text{ and } n_2 = 12 \]

The density function used in the kernel discriminant classification is as follows:
\[ \hat{f}_1(x) = \frac{1}{26} \sum_{i=1}^{26} K_1(x - x_i) \text{ and } \hat{f}_2(x) = \frac{1}{12} \sum_{i=1}^{12} K_2(x - x_i) \]

1) Kernel Discriminant Analysis with Normal Kernel Function
The normal kernel function that used as follows :
\[ K_1(z) = \frac{1}{c_0(z)} \exp \left(-\frac{1}{h^2} z^T S_1^{-1} z \right) \text{ and } K_2(z) = \frac{1}{c_0(z)} \exp \left(-\frac{1}{h^2} z^T S_2^{-1} z \right) \]
with \( C_0(1) = (2\pi)^{\frac{4}{3}} h^4 |S_1|^\frac{1}{2} \) and then \( C_0(2) = (2\pi)^{\frac{4}{3}} h^4 |S_2|^\frac{1}{2} \). Then, the classification result of kernel discriminant analysis using normal kernel function with bandwidth 1 to 2.5 can be seen in the following figure:

**Figure 1. The Classification Accuracy of Normal Kernel Discriminant Analysis with Bandwidth 1 to 2.5 in Testing Data**

Based on the figure, the highest classification accuracy value is kernel discriminant analysis with normal kernel function using the bandwidth value of 1 and 1.1, but in this study, the author only uses bandwidth value of 1, where the result of classification accuracy is 0.7105 or 71.05%. Furthermore, the details of the classification accuracy of testing data using bandwidth 1 can be obtained in Table 2.

**Table 2. The Result of Classification Accuracy from Normal Kernel Discriminant Analysis with Bandwidth 1 in Testing Data**

| Actual Group | Predicted Group | Total |
|--------------|-----------------|-------|
| 1            | 26              | 0     | 26 |
| 2            | 11              | 1     | 12 |
| Total        | 37              | 1     | 38 |

Table 2 shows that from 38 analyzed testing data, the childbearing age couples that classified exactly as non-family planning program participant are 26 couples and childbearing age couples that classified as having family planning program participant status are 12 couples. In addition, childbearing age couples who should have been in family planning program participant status but they are classified into childbearing age couples with no family planning program participant status are 11 couples. Therefore, APER value and classification accuracy are obtained as follows:

\[
\text{APER} = \frac{11}{38} = 0.2895 \quad \text{and} \quad \text{Accuracy} = 1 - 0.2895 = 0.7105
\]

2) Kernel Discriminant Analysis with Triweight Kernel Function

The triweight kernel function that used as follows:

\[
K_1(z) = C_3(1) \left( \frac{1-x^2S_1}{h^2} \right)^3 \quad \text{and} \quad K_2(z) = C_3(2) \left( \frac{1-x^2S_2}{h^2} \right)^3
\]

with \( C_3(1) = \left( 1 + \frac{4}{4} \right) \cdot \left( 1 + \frac{4}{4} \right) \cdot \frac{\pi h^4}{\pi (1+\frac{4}{2})} |S_1|^\frac{1}{2} \) and then \( C_3(2) = \left( 1 + \frac{4}{4} \right) \cdot \left( 1 + \frac{4}{4} \right) \cdot \left( 1 + \frac{4}{4} \right) \cdot \frac{\pi h^4}{\pi (1+\frac{4}{2})} |S_2|^\frac{1}{2} \). Then, the classification result of kernel discriminant analysis using triweight kernel function with bandwidth 1 to 2.5 can be seen in the following figure:
Figure 2. The Classification Accuracy of Triweight Kernel Discriminant Analysis with Bandwidth 1 to 2.5 in Testing Data

Based on the figure, the highest classification accuracy value is kernel discriminant analysis with triweight kernel function using the bandwidth value of 1.4 to 1.7, but in this study, the author only uses bandwidth value of 1.4, where the result of classification accuracy is 0.7368 or 73.68%. Furthermore, the details of the classification accuracy of testing data using bandwidth 1.4 can be obtained in Table 3.

Table 3. The Result of Classification Accuracy from Triweight Kernel Discriminant Analysis with Bandwidth 1.4 in Testing Data

| Actual Group | Predicted Group | Total |
|--------------|-----------------|-------|
| 1            | 26              | 0     | 26   |
| 2            | 10              | 2     | 12   |
| Total        | 36              | 2     | 38   |

Table 3 shows that of 38 testing data analyzed, the childbearing age couples that classified exactly as non-family planning program participant are 26 couples and childbearing age couples that classified as having a family planning status program participant are 2 couples. In addition, childbearing age couples who should have been in family planning program participant status but they are classified into childbearing age couples with no family planning program participant status are 10. Therefore, APER value and classification accuracy are obtained as follows:

\[
\text{APER} = \frac{10}{38} = 0.2632 \quad \text{and} \quad \text{Accuracy} = 1 - 0.2632 = 0.7368
\]

3.3. Comparison of Normal and Triweight Kernel Discriminant

Based on the previous explanations, the classification accuracy of normal kernel discriminant in testing data is 71.05% with APER value of 28.95%, while the classification accuracy of triweight kernel discriminant is obtained equal to 73.68% with APER value equal to 26.32%. Thus, this indicates that the classification of family planning program participation data of childbearing age couples in Semarang City of 2016 using triweight kernel discriminant can be expressed better than normal kernel discriminant.

4. Conclusion

Based on the results, then obtained the conclusions that the classification accuracy using normal kernel discriminant is obtained equal to 71.05 % with APER value equal to 28.95 %, while classification accuracy with triweight kernel discriminant is obtained equal to 73.68 % with APER value equal to 26.32 %. Furthermore, the classification of family planning program participation data of childbearing age couples in Semarang City of 2016 using triweight kernel discriminant so that can be expressed better than normal kernel discriminant.
References
[1] BKKBN. 2014. *Tatacara Pelaksanaan Pencatatan dan Pelaporan Program Kependudukan dan Keluarga Berencana Nasional*. Semarang: BKKBN.
[2] Irianto, K. 2014. *Pelayanan Keluarga Berencana*. Bandung: Alfabeta.
[3] Johnson, R. A. and Wichern. D. W., 2007. *Applied Multivariate Statistical Analysis*. Sixth Edition. New Jersey: Prentice Hall International, Inc.
[4] Hair, J. F., Black. W. C., Babin, B. J. and Anderson R. E., 2006. *Multivariate Data Analysis*. Seventh Edition. Pearson Education Prentice Hall, Inc.
[5] Dillon, W and Goldstein, M., 1984. *Multivariate Analysis Methods and Applications*. New York: John Wiley & Sons, Inc.
[6] Silverman, B. W., 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
[7] Khattree, R. and Naik, D. N., 2000. *Multivariate Data Reduction and Discrimination with SAS Software*. Cary, NC: SAS Institute, Inc.