Blood Glucose Levels of Non Invasive Tool Using Support Vector Machine on Multi-Class Imbalanced Data

Selvi Annisa[1], Asep Saefuddin[2], Erfiani[3],

Abstract— Most of the classification method tends to be effective in the case of a balanced data. But, sometimes in many real-world applications, multi-class imbalance classification problems occured, such as in this case is blood glucose levels of non-invasive tools. One-vs-One strategies is a well-known tehcniques to address the classification problems involving multi-class. To handle the imbalance case in the data, we used SMOTE. The results of SVM with radial basis function kernel and SMOTE gave the avgacc 51.8% for predicting the classes of blood glucose levels from non-invasive measurements, with the optimal parameter C 2 and γ 0.002.

Index Terms— Blood Glucose Levels, Imbalanced Data, Multi-Class Classification, One-vs-One, Radial Basis Function,Support Vector Machine, SMOTE

1 Introduction

Support Vector Machine (SVM) is a classification method introduced in 1992 by Boser, Guyon and Vapnik [1]. SVM is a method that has the advantage of being able to process low and high dimensional data without experiencing significant performance degradation. Furthermore, SVM strategy in determining the maximum hyperplane margin can also reduce the classification errors [2]. Since SVM consider a balanced data set, it makes SVM less effevtive for imbalanced case. In recent years, many effort have been focused on the binary class imbalance problem which only contain two classes [6, 9]. However, multi-class imbalance classififcation is widely applied in many areas, including medical diagnosis [10]. One of the solution is to transform multi-class classification problems into binary class sub-problems, which are much easier to discriminate. One-vs-one (OVO) is a well-known approaches of decomposition strategies to deal with multi-class classification problem [5, 8]. From the binary subproblems, the imbalanced case will be handled by synthetic minority oversampling technique (SMOTE).

In this paper, we focus on multi-class imbalance classification problems using OVO and SMOTE with SVM algorithm to classify blood glucose levels of non invasive devices. The classes of blood glucose levels in this case are low, moderate and high.

2 Background

2.1 Non-Invasive Blood Glucose Level Monitoring Tool

Early detection for people with diabetes and pre-diabetes needs to be done so that patients can start managing the disease earlier, ie by checking blood sugar levels. Invasive method is a method that is widely used to determine blood sugar levels. But this method is considered to be still less effective because of the pain felt by the patient at the time of blood sampling and sterilization of the equipment used to take the blood sample. In addition, there have been many reports stating the occurrence of infection during blood sampling. This infection occurs because the body of diabetics is unable to produce insulin [11]. Therefore we need a tool to measure blood glucose levels without injuring the patient's body, which is non-invasive device. This tool using the principle of near-infrared spectroscopy.

2.2 One-vs-One

OVO decomposition scheme divides an m class problem into $m(m-1)/2$ binary problems. Each problem is faced by a binary classifier, which is responsible for distinguishing between a different pair of classes. The learning phase of the classifier is done using as training data only a subset of instances from the original training data set, that contains any of the two corresponding class labels, whereas the instances with different class labels are simply ignored [7].

In order to predict a new pattern, the voting strategy is the simplest but powerful aggregation, therefore, it is considered as the aggregation approach in this study. Vote which is also called binary voting and Max-Wins rule, considers a vote for the predicted class by the binary classifier. Votes received by each class are counted and the final class obtaining the largest number of votes is the predicted class.

2.3 Support Vector Machine

The concept of SVM can be explained simply as an effort in finding the best hyperplane that functions as a separator of two classes. Both classes are separated by lines called hyperplane with line equations is

- *Selvi Annisa is currently pursuing master degree program in applied statistics in Bogor Agriculture University, Indonesia, PH +6282250216538. E-mail: selviannisa92@gmail.com*
- *Asep Saefuddin is a Lecturer, Department of Statistics, Bogor Agriculture University, Bogor, Indonesia. E-mail: asaefuddin@gmail.com*
- *Erfiani is a Lecturer, Department of Statistics, Bogor Agriculture University, Bogor, Indonesia. E-mail: erfiani_ipb@yahoo.com*

$$w \cdot x + b = 0$$

w is a weight vector, **x** is an input vector and b is bias.

The best separator hyperplane between the two classes can be found by measuring the hyperplane margin and looking for the maximum point. Margin is the distance between the hyperplane and the closest pattern of each class. The pattern that has the closest distance to the hyperplane is called the support vector. SVM separates data using a hyperplane with the largest inter-class margin.

Generally problems in the real world are rarely linearly separable, mostly non-linear. So to solve this, SVM is modified by entering the Kernel function. Under normal circumtaces, the first consideration of choosing kernel function is the Gaussian radial basis function (RBF) because it has fewer parameters to select. The function of Gaussian radial basis function is

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = exp\left(-(\|\boldsymbol{x_i} - \boldsymbol{x_j}\|^p / 2\sigma^2)\right)$$

2.4 Synthetic Minority Oversampling Technique

The approach using the SMOTE method according to [3] 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 [12].

2.5 Performance Measures

Standard metrics such as accuracy rate should not be considered in the case of imbalanced datasets, since they do not distinguish between the number of correct classifications of the different classes, which may lead to erroneous conclusions. For this study we have decided to use the average accuracy metric. The average accuracy gives the same weight to each class. It achieves the accuracy rate of each class independently, and then the final result is obtained by theh average value. The average accuracy is computed as follows,

$$AvgAcc = \frac{1}{m}\sum_{i=1}^{m} TPR_i$$

where m is the number of classes and TPR_i for the True Positive Rate of the i-th class [4].

3 Method

3.1 Data

The data used in this study are part of the development research and prototype clinical trials of non-invasive monitoring tools for blood glucose levels. This research was conducted in April 2016 - January 2017 involving 118 respondents. Spectral data obtained from sensor of non-invasive monitoring tools for blood glucose levels with infrared 1600 nm.

The predictor variable in this study is the residual intensity value which is the output of a non-invasive blood glucose monitoring tool, while the response variable is the blood glucose level class from an invasive measurement carried out by Prodia's clinical laboratory team. Each respondent is measured 5 times for invasive and non-invasive blood glucose levels.

The residual intensity describes as intensity of light is continued and captured by the sensor. The resulting residual intensity forms two peaks for each measurement, so that a total of 10 peaks are formed from measurements using non-invasive tools. For each peak, 1 values will be taken as predictor values, so that there are a total of 10 predictors obtained. The value taken is the standard deviation of each peak.

3.2 Data Analysis

The analysis procedure in this study are:

1. Exploration data analysis on invasive blood glucose levels.
2. Exploration data analysis on non-invasive residual intensity of blood glucose levels.
3. The multi-class classification problem is split into 3 binary class subproblems. Low and Moderate, Moderate and High, Low and High.
4. Data set is split using 25 repeated 4-fold stratified cross-validation.
5. For each training dataset in subproblems, the imbalanced classes will be solved by SMOTE method with 500% oversampling percentages.
6. Build a classification model using the SVM method with RBF kernel in subproblems.
7. Predict the class of blood glucose levels in the test data by using majority vote.
8. Calculate the average accuracy obtained by means of the 4th step.
9. Find the optimal parameters of RBF SVM.
10. Calculate the average accuracy obtained by means of the 4th step using the optimal parameters of RBF SVM.

4 Discussion

In this study, moderate blood glucose levels has the most number of respondents, while low and high blood glucose levels much smaller than moderate as described in Figure 1.

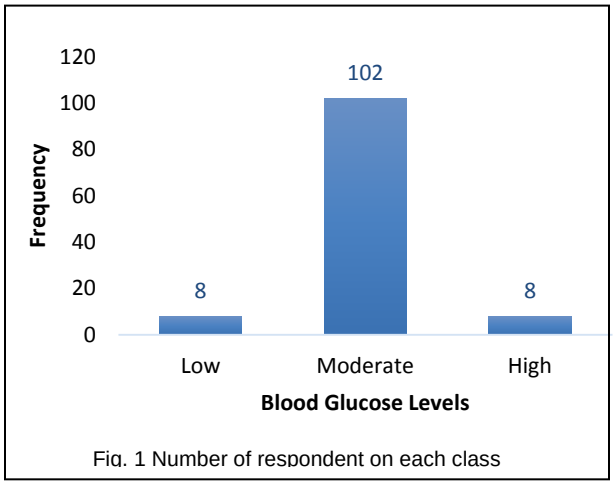

Fig. 1 Number of respondent on each class

Based on Figure 1, we conclude that the data is imbalance and this will be solved by using SMOTE on each training dataset. Both of low and high class in the training data will be generated with a percentage of 500.

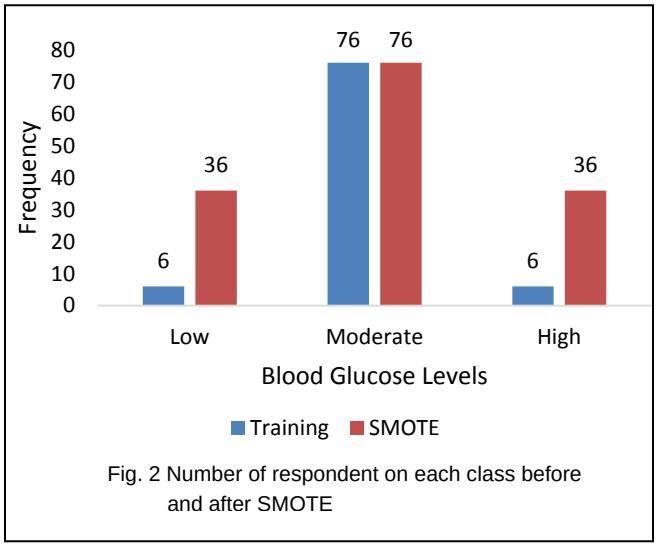

Fig. 2 Number of respondent on each class before and after SMOTE

As can be seen in Figure 2, the number of minority class observations becomes 6 times the number of observations of the initial minority class. The next step is to make the SVM classification model using SMOTE training data to predict the classes of non-invasive blood glucose levels and find the optimal parameter for RBF SVM model.

TABLE 1
AVGACC FOR EACH PARAMETER VALUE

Parameter		AvgAcc (%)
C	γ	
Default		38.1
0.02	0.002	33.3
	0.02	33.3
	0.2	33.3
	2	33.3
	20	33.3
	200	33.3
0.2	0.002	33.3
	0.02	47.9
	0.2	36.2
	2	33.3
	20	33.3
	200	33.3
2	0.002	**51.8**
	0.02	45.4
	0.2	35.1
	2	32.3
	20	33.3
	200	33.3
20	0.002	49.1
	0.02	43.4
	0.2	34.5
	2	30.3
	20	33.3
	200	33.3
200	0.002	45.4
	0.02	37.2
	0.2	32.9
	2	30.3
	20	33.3
	200	33.3

There are two parameters in RBF SVM that will be tune to find the optimal ones, which are the tolerance parameter (C) and γ. It is not known beforehand which C and γ are best for predicting blood glucose levels; consequently some kind of model selection (parameter search) must be done. The goal is to identify the optimal parameters so that RBF SVM can accurately predict blood glucose levels.

Various pairs of C and γ are tried and the one with the best cross-validation avgacc is picked. It was suggested that trying exponentially growing sequences of C and γ is a practical method to identify optimal parameters.

Based on Tabel 1, the highest avgacc is 51.8% that obtained from parameter C 2 and γ 0.002. We can see that there is a significant improvement from using the default paramaters, the avgacc increased from 38.1% to 51.8%, it increase by more than 13%.

5 CONCLUSION

5.1 Appendices

The classification of non-invasive blood glucose levels using RBF SVM and SMOTE gave the avgacc is 51.8%. the avgacc is obtained from the optimal parameter C 2 and γ 0.002. This avgacc is the mean or average avgacc from 25 repeated 4-fold stratified cross-validation.

5.2 Suggestion

The suggestion for the further research is to use other multiclass classification methods with imbalance case to classify the spectrum of non-invasive blood glucose levels to determine the better results than RBF SVM and SMOTE method.

REFERENCES

[1] Boser B, Guyon I, Vapnik V. 1992. A training algorithm for optimal margin classifier. *5th Annual ACM Workshop on COLT*, pp 144-152.

[2] Burges CJC. 1998. A Tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*. 2: 121-167.

[3] Chawla VN, Bowyer KW, Hall LO, Kegelmeyer WP. 2002.

SMOTE: Synthetic Minority Over-Sampling Technique. *J of Artif Intell Research*. 16: 321-357.

[4] Ferri C, Orallo HJ, Modroiu R. 2009. An experimental comparison of performance measures for classification. *Pattern Recogn. Lett*. 30: 27-38.

[5] Galar M, Fernandez A, Barrenechea E, Herrera F. 2014. Empowering difficult classes with a similarity-based aggregation in multi-class classification problems. *Information Sciences*. 264: 135-157.

[6] Haibo H, Garcia EA. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Enginerring*. 21 (9): 2163-1284.

[7] Hastie T, Tibshirani R. 1998. Classification by pairwise coupling. *Annals of Statistics*. 26(2): 451-471.

[8] Kang S, Cho S, Kang P. 2015. Constructing a multi-class classifier using one-against-one approach with different binary classifiers. *Neurocomputing*. 149: 677-682.

[9] Lopez V, Fernandez A, Garcia S, Palade V, Herrera F. 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intristic characteristics. *Information Sciences*. 250: 113-141.

[10] Krawczyk B, Galar M, Jelen L, Herrera F. 2015. Evolutionary undersampling boosting for imbalanced classififcation of breast cancer malignancy. *Applied Soft Computing*.

[11] Smith J. 2017. *The Pursuit of Noninvasive Glucose 5th ed.*.

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