Comparison of Under-Sampling and Over-Sampling Techniques in Diabetic Mellitus (DM) Patient Data Classification by Using Naive Bayes Classifier (NBC)

Ristu Saptono, Winarno, and Dewi Prasetyan Drajati
Department of Informatics, Universitas Sebelas Maret, Surakarta, Indonesia.
ristu.saptono@staff.uns.ac.id

***Abstract*—Imbalance dataset is a big problem inside a classification process. Most of the classification algorithms tend to classify the majority instances and ignore the minority ones. It can cause the misclassification of the minority instances and make the precision and recall of this minority data become low. In order to resolve this kind of problem there will be done both undersampling and oversampling process to make the dataset balance. In this proposed research there will be used undersampling and oversampling techniques to balance the number of majority and minority instances from diabetic patient data. The other techniques used in this research are backward greedy stepwise for features selection and Naive Bayes Classifier (NBC) for data classification. The conclusion, oversampling techniques give significantly higher precision and recall than oversampling, although the accuracy fairly equal.**

***Index Terms*—Backward Greedy Stepwise; Naive Bayes Classifier Oversampling; Undersampling.**

I. INTRODUCTION

Imbalance number of instances from some categories inside a dataset is a big problem. It can cause data misclassification that can cause invalid classification result. Category or class with big number of instances is called majority class, on the other hand category or class with small number of instances is called minority class. The classification algorithms tend to ignore to classify the minority class instances so the classification result of this class become low especially the values of precision and recall [1].

There are two methods can be used for balancing the number of majority and minority class instances, they are undersampling and oversampling methods. Undersampling method will eliminate some majority instances randomly to make the number of this majority class not too far away from the number of minority class instances. Oversampling method will replicate the number of minority class instances in order to make the number of this minority instances balance with the number of majority class instances [2].

The undersampling usage to resolve the imbalance data problem had ever done by [3] using diabetic patient data from University California Irvine (UCI) repository. The UCI's diabetic data contains diabetic patient data from some American hospitals between 1999 until 2008 with two main data categories like Otherwise and Readmitted. The number of Otherwise data is 64141 and the number of Readmitted data is 6293. The Otherwise data explains that diabetic patients didn't do outpatient since 30 days after they had been discharged from the hospital, readmitted data explains that diabetic patients still did outpatient after undergoing hospitalization counted 30 days after they had been discharged from the hospital [4].

The result of the previous research by [3] there was increasing of precision and recall values, but the increasing of precision and recall values wasn't too significant. In order to increase the values of precision and recall, in this proposed research will be used oversampling method by using Synthetic Minority Oversampling Technique (SMOTE). The classification result from diabetic data that processed by using undersampling method will be compared with the classification result that processed by using SMOTE. SMOTE had ever used inside research by [5] for comparing three classification algorithms such as Probabilistic Neural Network (PNN), Naive Bayes (NB), and Decision Tree (DT). The best performance of diabetic patient data from Tehran Lipid Glucose Study (TLGS) classification showed by Naive Bayes algorithm. Inside this proposed research, another technique will be used is Backward Greedy Stepwise for selecting the most influenced attribute.

II. SAMPLING METHOD

The data imbalance problem occurs between two or more classes in a set of data. The majority class is a class with high amount of instances whereas minority class is a class with low amount of instances. The imbalance between majority and minority class data amount can cause invalid of classification results. Invalid results of classification process caused by misclassification of class instances. The conventional classification algorithm tends to classify the instances belong to majority class and ignore the classification process on minority class [5]. It will cause the classification result of minority class to become low. To resolve this kind of problem, there must be a method that could balance both majority and minority class data amount. This proposed method is sampling method. Sampling method is a method that could balance data distribution inside majority and minority classes with some certain procedures. There are two variations of sampling method such as undersampling and oversampling methods [6].

A. Undersampling Method

Undersampling method is the way to resolve the imbalance data set problem by eliminating the amount of majority data.

Spreadsubsample is one of undersampling technique that generates random data distribution in majority class with maximum distribution ratio of majority class to minority class is 10:1. The amount of instances on majority class will be cut randomly. The negative effect of undersampling method is the loss of important information inside the majority class data [7]. Undersampling method is illustrated in Figure 1.

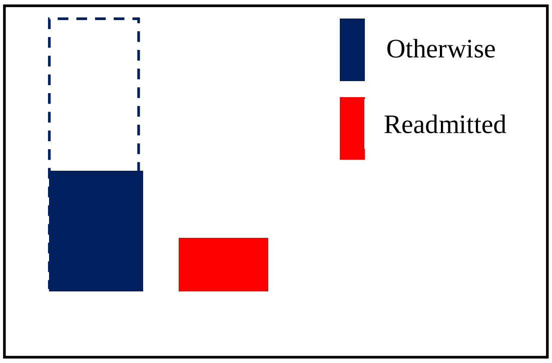

Figure 1: Undersampling Process

B. *Oversampling Method*

Oversampling method is the other way to resolve the imbalance data set problem by increasing the amount of minority data. SMOTE is one of oversampling method that generates new instances inside the minority class by calculating the value of the nearest linear neighborhood data. The distance of the nearest neighborhood data usually symbolized with k that set to 5 then the other new data will be taken from the previous data randomly [8]. Oversampling method is illustrated in Figure 2.

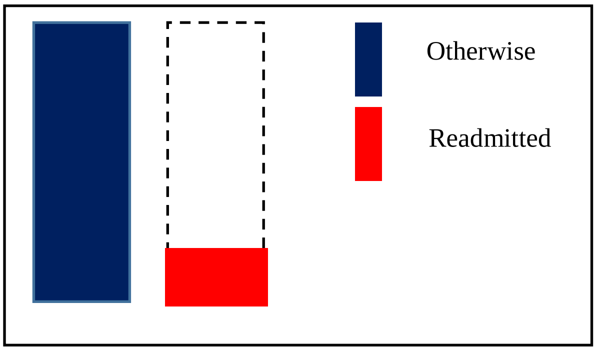

Figure 2: Oversampling Process

III. Backward Greedy Stepwise

Backward greedy stepwise is one of wrapper feature selection technique. Backward greedy stepwise will search the relevant attribute from a data set greedily [9]. The searching process will be started by eliminating the irrelevant attribute followed by evaluation of the data performance until only the relevant data attributes left. Relevant data attribute means that the presence of an attribute can give positive impact to the data classification result [10].

IV. Naive Bayes Classifier (NBC)

Naive Bayes Classifier (NBC) is a classification algorithm based on Bayesian Theorem. This algorithm simplifies the data training process by assuming the independence of the feature inside a class. NBC can work effectively inside supervised learning environment by searching the biggest value from some class probabilities inside a data set. There are some advantages of using NBC algorithm such as the simplicity and its ability to handle the high number of data [11]. Classification using NBC is given by:

$$P(C|X) = \frac{P(X|C) \; x \; P\,(C)}{P(X)} \qquad (1)$$

where P(C|X) is probability of data C inside class X, P(C) is probability of data C, P(X) is probability of data X, and P(X|C) is probability of data X inside class C.

V. Research Methodology

Data set used in this proposed research is diabetic patient data taken from University California Irvine (UCI) repository [12]. This data contains American diabetic patient data from 1999 until 2008. The amount of diabetic patient with readmission status Readmitted is 6293 and the amount rest of data with readmission status Otherwise is 64141. Diabetic data with readmission status Otherwise are grouped as the majority data and the other diabetic data with Readmitted status are grouped as the minority data. To solve the imbalance data problem between majority and minority data, there are some to do steps as figured in Figure 3.

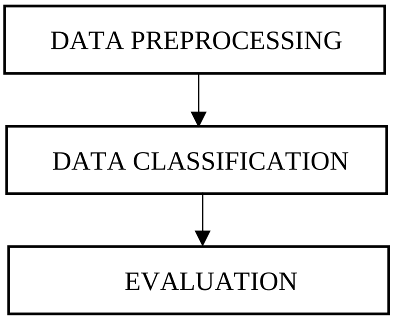

Figure 3: Research Methodology

A. *Data Preprocessing*

The initial process in this research starts with data preprocessing. In preprocessing step, the imbalanced DM data will be duplicated and grouped into undersampling and oversampling groups. Each group will handle their imbalance data by using Spreadsubsample for undersampling and SMOTE for oversampling.

1) Undersampling
   Diabetic patient data in majority class will be eliminated by using Spreadsubsample technique. Spreadsubsample technique includes cutting process of the majority data amount, after that the data will be categorized into several levels of data distribution. The proposed distribution levels of the majority data are six, seventh, eight, and nine times higher than the amount of minority class data.
2) Oversampling
   Diabetic patient data in minority class will be increased by using Synthetic Minority Oversampling Technique (SMOTE). Oversampling process with SMOTE will be started by calculating value of some nearest linear neighborhood data along k distances (if the replication number is n, so the k value is the same with n - 1) then followed by taking some random data from the previously available data . The proposed oversampling level of the minority class are six, seven,

eight, and nine times higher than the amount of the previous amount of minority data without oversampling.

The attributes of the balanced data set then selected by using Backward Greedy Stepwise feature selection to search the most relevant attributes in undersampling and oversampling groups. The most relevant attribute will give positive impact for data classification result. DM data with selected attributes on both undersampling and oversampling group, then separated into ten groups randomly. Each of tenth group will be labeled with "Test_n" (n = 1, 2, ... , 10) shown in Figure.4. The data inside Test_n groups then divided into some percentages of training and testing data. The percentages of training data are 66%, 75%, 80%, and 90%, and the rest of each previous percentages will be grouped as testing data.

B. Data Classification

Diabetic patient data then classified by using NBC algorithm. The classification result used to predict the classification result of testing data. There are some components of classification result used as measuring elements such as accuracy, precision, and recall. Those three measuring elements from undersampling group then compared with the oversampling group elements.

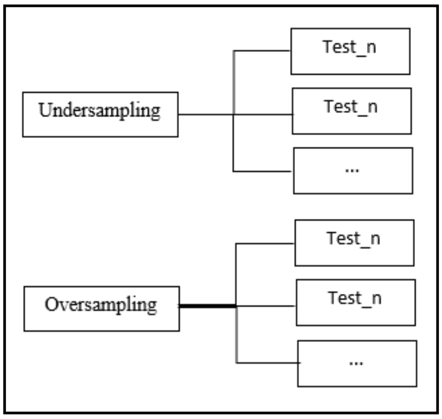

Figure 4: Data Grouping Process

C. Evaluation

The evaluation of data classification performance is based on the values of accuracy, precision, and recall. The accurate, precision, and recall values from undersampling group are compared with accurate, precision, and also recall values from oversampling group.

VI. RESULT

The average number of the classification result from training data belong to undersampling and oversampling group are shown in Table 1.

Table 1
Accuracy comparison in training set for undersampling (US) and oversampling (OS) in percentage (%)

Training set	Data Distribution							
	6 times		7 times		8 times		9 times	
	US	OS	US	OS	US	OS	US	OS
66%	84.9	88.0	87.6	88.7	88.9	88.5	90.1	89.0
75%	84.7	88.1	87.7	88.6	89.0	88.6	90.1	88.9
80%	84.7	88.1	88.0	88.7	89.0	88.6	89.9	88.9
90%	84.9	88.1	87.6	88.7	89.3	88.7	89.9	89.0

Despite around 3.5% different in sixth-time data distribution, Table 1 shows fairly equal in accuracy from seventh until nine times data distribution. The highest value of accurate average belongs to undersampling group shown in Table 1 is 90.1% when the amount of instances in majority class is nine times higher than the amount of instances in minority class with 66% and 75% training data percentages. The highest value of accurate average belongs to oversampling group is 89.0% when the amount of instances in minority class is increased by nine times higher than the amount of the previous minority data without oversampling with 66% and 90% training set.

Table 2
Precision comparison in training set for undersampling (US) and oversampling (OS) in percentage (%)

Training set	Data Distribution							
	6 times		7 times		8 times		9 times	
	US	OS	US	OS	US	OS	US	OS
66%	52.4	84.5	54.6	87.0	50.3	88.9	51.2	89.9
75%	50.1	84.8	53.6	87.3	51.8	88.6	49.8	89.5
80%	49.9	85.5	58.2	87.4	55.1	88.9	50.7	89.5
90%	50.2	86.9	46.4	87.8	55.4	88.4	53.4	88.9

Table 2 shows significant differences in precision average. Oversampling techniques have almost double compare to undersampling technique. The highest precision average for undersampling techniques is shown in Table 2 is 58.2% when the amount of the majority class instances is seven times higher than the amount of minority class instances with 80% training set. The highest precision average shown for oversampling technique in Table 2 is 88.9% when the amount of minority class instances is increased by nine times higher than the amount of the previous minority data without oversampling.

Table 3
Recall comparison in training set for undersampling (US) and oversampling (OS) in percentage (%)

Training set	Data Distribution							
	6 times		7 times		8 times		9 times	
	US	OS	US	OS	US	OS	US	OS
66%	10.5	83.9	6.7	85.8	3.6	88.4	3.4	89.6
75%	50.1	84.0	6.0	85.6	3.9	88.4	3.0	89.6
80%	9.1	83.9	6.1	86.0	3.7	88.4	3.4	89.7
90%	8.7	84.0	6.0	85.8	3.8	88.4	3.3	89.5

Table 3 shows the recall averages of oversampling almost 21 times higher than undersampling technique, except in sixth times and 75% training set. The highest recall average for undersampling shown in Table 3 is 50.1% when the amount of instances in majority class is six times higher than the amount of instances in minority class with 75% training set. The highest value of recall average shown in Table 3 is 89.7% when the amount of instances in minority class is increased by nine times higher than the amount of the previous minority class instances without oversampling with 80% training set.

In testing set, accuracy average of undersampling is slightly higher than oversampling as shown in Table 4. The highest accuracy average of undersampling shown in Table 4 is 89.5% when the amount of majority class instances is nine times higher than the amount of minority class instances with 10% testing data percentage. The highest accuracy average of oversampling shown in the table is 89.9% when the amount of minority class instances is increased by nine times higher than the amount of minority class instances without oversampling with 34% testing data percentage.

Table 4
Accuracy comparison in testing set for undersampling (US) and oversampling (OS) in percentage (%)

Testing set	Data Distribution							
	6 times		7 times		8 times		9 times	
	US	OS	US	OS	US	OS	US	OS
34%	88.2	86.8	88.3	87.9	88.2	87.4	89.0	88.4
25%	88.5	86.8	88.6	87.9	88.1	87.6	89.1	88.3
20%	88.7	86.8	88.5	88.0	87.8	87.6	89.1	88.3
10%	88.4	86.9	88.0	88.1	88.0	87.7	89.5	88.9

Table 5
Precision comparison in testing set for undersampling (US) and oversampling (OS) in percentage (%)

Testing set	Data Distribution							
	6 times		7 times		8 times		9 times	
	US	OS	US	OS	US	OS	US	OS
34%	32.2	86.3	32.6	87.5	32.8	87.1	29.0	88.7
25%	35.5	86.0	38.8	87.8	45.7	87.2	32.3	88.8
20%	35.6	86.2	41.1	87.7	38.6	87.0	31.5	88.7
10%	43.5	86.2	36.4	87.7	32.1	87.4	34.5	89.3

Table 5 shows that precision average of oversampling technique is more than twice higher compared to undersampling. The highest precision average of undersampling shown in Table 5 is 45.7% when the amount of majority class instances is eight times higher than the amount of minority class instances with 25% testing data percentage. The highest precision average of oversampling shown in Table 5 is 89.3% when the amount of minority class instances is increased by nine times higher than the amount of the previous minority class instances without oversampling with 10% testing data percentage.

Table 6 shows the significant different the recall averages of oversampling compared to undersampling group in every cell data distribution and testing set. The highest average of recall for undersampling is 10.8% when the amount of majority class instances is nine times higher than the amount of minority class instances. The highest value of recall average shown in the table is 89.5% when the number amount of minority class instances is increased by nine times higher than the amount of the previous minority class instances without oversampling with 10% testing data percentage.

VII. Conclusion

Based on the classification result of the training data from both undersampling and oversampling groups, it can be concluded that oversampling process by using SMOTE give significantly higher precision and also recall values than Spreadsubsample undersampling. The accuracy average of SMOTE oversampling is fairly equal with Spread subsample undersampling with around 1% to 4% different.

Table 6
Recall comparison in testing set for undersampling (US) and oversampling (OS) in percentage (%)

Testing set	Data Distribution							
	6 times		7 times		8 times		9 times	
	US	OS	US	OS	US	OS	US	OS
34%	6.6	84.3	4.2	85.6	2.6	87.8	2.0	89.0
25%	6.6	84.7	4.2	86.1	3.3	87.3	1.6	89.3
20%	7.3	84.7	4.4	86.0	2.7	87.1	1.6	89.2
10%	8.2	84.1	4.3	85.1	3.5	87.1	10.8	89.5

References

[1] 08 September 2013. [Online]. Available: http://www.depkes.go.id/article/view/2383/diabetes-melitus-penyebab-kematian-nomor-6-di-dunia-kemenkes-tawarkan-solusi-cerdik-melalui-posbindu.html. [Accessed 12 June 2017].

[2] "PENDERITA DIABETES AKIBAT KOMPLIKASI." 16 07 2017. [Online]. Available: http://penderitadiabetes.com/jumlah-penderita-diabetes-di-dunia-menurut-who/.

[3] M. R. Wijaya, R. Saptono and A. Douwes. "The Effect of Best First and Spreadsubsample on Selection of a Feature Wrapper With Naive Bayes Classifier for The Classification of The Ratio Inpatiens." *Scientific Journal of Informatics*, vol. 3 pp. 41-50, 2016.

[4] L.M. Taft, R.S. Evans, C.R. Shyu, M.J. Egger, N. Chawla, J.A.Mitchell, S. Thornton, B. Bray, and M. Varner, "Countering imbalanced datasets to improve adverse drug event predictive models in labor delivery." *Journal of Biomedic Informatics*, vol. 42, p. 356-364, 2009.

[5] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review." *GESTS International Transactions on Computer Science and Engineering*. vol. 30, pp. 25-36, 2006.

[6] P. Yildirim, "Pattern classification with imbalanced and multiclass data for the prediction of albendazole adverse event outcomes." *Procedia Computer Science*. vol. 83, pp. 1013-1018, 2016.

[7] A. Sonak, and R.A.Patankar. "A Survey on Methods to Handle Imbalance Dataset", *International Journal of Computer Science and Mobile Computing*, vol. 4, no. 11, 2015.

[8] D. D. Ramyachitra, and P.Manikandan. "Imbalanced Dataset Classification and Solutions: A Review." *International Journal of Computing and Bussiness Research (IJBR)*. vol. 5,. no. 4, 2014.

[9] E. Roglia, and R. Meo. "A Composite Wrapper for Feature Selection", In *Proceedings of Workshop on Data Mining and Bioinformatics in AI* IA-Intelligenza Artificiale e Scienza della Vit a (DMBIO08) Cagliari (Italy)*, 2008, p. 14.

[10] J. Landy. "Stepwise regresion for unsupervised learning." arXiv. p. 9. 2017.

[11] A.Ambica, S. Gandi and A. Kothalanka, "An Efficient Expert System For Diabetes By Naive Bayesian Classifier",. *International Journal of Engineering Trends and Technology (IJETT)*. vol. 4. no. 10. 2013.

[12] "UCI Machine Learning Repository." [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/diabetes/. [Accessed 28 July 2017].

[13] Muniroh and A. Suharsono. "Klasifikasi Dynamic Financial Distress Perusahaan Manufaktur yang Terdaftar di Bursa Efek Indonesia Tahun 2012-2014 Menggunakan Regresi Logistik Biner dan Classification Analysis & Regression Tree (CART)." *Jurnal Sains dan Seni ITS*. vol. 5, no. 2, pp. 311-316, 2016.