978-1-5386-5525- 2/18/$31.00 ©2018 IEEE

Ensemble Learning on Large Scale Financial Imbalanced Data

H.R Sanabila
Faculty of Computer Science, Universitas Indonesia
hadaiq@cs.ui.ac.id

Wisnu Jatmiko
Faculty of Computer Science, Universitas Indonesia
wisnuj@cs.ui.ac.id

***Abstract*—This study focused on evaluating the performance of ensemble learning on handling imbalanced data. Imbalanced data is a special problem in classification task where the class distribution is not uniformed. Resampling (SMOTE and ENN) is employed to improve the classifier performance. Four metrics is applied for performance evaluation i.e. precision, recall, specificity, and F-1 score. Based on the experiments, Bagging has a superior performance compared to baseline classifiers (Naïve Bayes and Log Regression) and other ensemble learnings (Boosting and Random Forest). In addition, the combination of SMOTE and ENN successfully increase the classification performance and avoiding biased to the majority class.**

Keywords—ensemble learning; imbalanced data; big data; spark (key words)

I. Introduction

Sobolevsky et al. predicted regional economic indices from individual spending behavior [1]. They used supervised learning approach (Generalized Learning Model) to identify characteristics of the category of individual transaction. Particle Component Analysis (PCA) was used as a mechanism for feature dimensionality reduction. For the dataset, they used bank card transactions from BBVA (Banco Bilbao Vizcaya Argentaria). This dataset is quite large, with 4,500,000 customer data, with approximately 178,000,000 number of transactions.

Rajwani et al. utilized several algorithms such as LSTM (Long Short-Term Memory Model), Linear Regression, Bayesian Ridge Regression, Random Forest Regression, and ARIMA to estimate the amount of cash in ATM [2]. They used 2.5 years of transaction historical data in a busy area. It consists of 120,246 rows and contains some important features, such as: withdrawal, change of pin, task transfer, bill payment, and balance inquiry. Some researchers utilized optimization in the form of evolutionary computation algorithms and spark to solve machine learning problems [3][4]. Data stream algorithm can also be used to solve big data problems [5][14].

G. Archana et al. evaluated the classification of customer's credit score (good or bad) based on the customer data [6]. They used several classification algorithms to classify the credit score (good or bad) based on the available dataset. The dataset consists of 1000 instances with 18 features. Some of the features were age, credit, duration, balance, and occupation. They experimented with some classification models such as, Random Forest, Linear Regression, Support Vector Machine, Adaptive Boosting, Decision Tree, and Neural Network. The random forest algorithm achieved the best classification accuracy with 100% accuracy.

The prediction of credit defaulters in credit risk evaluation was introduced by Suhamathy G [7]. This dataset contains 1000 rows with 21 features. Before they train the dataset, they utilize SMOTE as a solution for the imbalanced dataset. Random forest was used as a method to predict whether the features are categorized as a defaulter or not. The result showed that random forest algorithm achieved 94% accuracy. A mathematical model of stock price prediction by using social network has been introduced by Wang [8]. The author used Fuzzy Opinion Networks to predict the stock price. Fuzzy set become its output and the investor acts as the fuzzy input. Wang evaluated the mathematical model by using Monte Carlo simulations. The dataset that is used is the stock price of 15 top banking companies for 2 years (2013-2015). This model can identify combined uncertainty from social network to give a reliable signal in stock price prediction.

The prediction of Bank Failure was introduced by Vahid Behbood et.al [9]. They incorporate Multistep Fuzzy Bridged Refinement Domain Adaptation Algorithm to enhance the prediction accuracy of bank failure data. The utilization of similarity concept use to refine the labels and performing shift-unaware prediction model. They used 4 different datasets to test the proposed algorithm. Based on the evaluation by using some prediction models, SVM, FNN, and NN, the proposed model gives a significant improvement in accuracy. J. Xue et al. solved financial time series prediction by using RF-ELM (Random Fourier Mapping based Extreme Learning Machine) [10]. This method can prune the redundant and irrelevant hidden neurons to obtain a simpler hidden layer. In addition, it also gives differentiation in the hidden layer. The results showed that RF-ELM is comparable with other methods such as ANN (Artificial Neural Network) and SVM (Support Vector Machine).

G.Song et.al utilize Double Deep ELMs(DD-ELM) method to handle problems in time series forecasting [11]. They presented SA-ReTSP pruning technique in DD-ELM. Based on the error of the evaluation performance. The results show that DD-ELMS gives a better accuracy measurement compared to basic deep ELM models. Thomas Fischer introduced Deep learning with long short-term memory (LSTM) network to predict financial market. LSTM was tested and compared with other methods such as random forest, deep neural network (DNN), and logistic regression [12]. The dataset of financial market is gathered from December 1992 to October 2015. Based on the experiment, LSTM is able to get useful information from noisy data. It also gives a better accuracy prediction compared to random forest, DNN, and logistic regression. A cloud computing framework has been built as a framework for drug discovery simulation [14].

II. Dataset

This paper use costumer behavior of banking transactions that were obtained in 1.5 year from January 28^{th}, 2015. The initial data consists of 12 million rows and 48 features. The raw data is prepared in two stages i.e. data cleansing and feature selection.

1. Data Cleansing
   The data cleansing is conducted in Hive. The cleansing is focused on removing blank value and outlier. Subsequently, any decimal value in the data is rounded into integer. It is aimed to make it simpler to be processed.
2. Feature Selection
   The data have 48 features and it should be reduced into a more specific number that substantially represents the data. By employing Principal Component Analysis (PCA) in H2O, the features that has eigen value above 0.1 are preserved. There are 5 features and 1 label that is the combination of 3 products, i.e. current accounts, e-accounts, and direct debits. Then, the selected features are adjusted by using normalization to ensure that each feature has an equal weight in the learning process. The details of the selected feature can be seen in table 1.

The class distribution of this dataset is not uniformed. The majority is class 4 (60.4%) and class 0 (25.83%). Meanwhile, other classes have a low percentage and some of them are even lower than 1% such as class 7 and class 3. The distribution of the classes is illustrated in figure 1. For a classification task, it will certainly affect the model that is acquired from the training phase. The generated model will be biased to the majority class.

III. Methodology

In this study, we are focused on analyzing the classification task using ensemble method on an imbalance dataset. Ensemble method is a learning algorithm that use a set of base classifiers and classify the data by taking a vote for the predictions. Multiple base classifiers should tackle the learning problems in the imbalanced data. This study uses three Ensemble methods, namely Boosting, Random Forest, and Bagging. Meanwhile Naïve Bayes and Logistic Regression is employed as the baseline method. The experimental pipeline of this study is shown in figure 3.

Logistic regression is a learning algorithm that use statistical approach. Logistic regression use weighted linear combination of each feature to predict the class. By employing logistic function, the real value of the feature's linear combination is mapped into binary class (0 or 1). Naïve Bayes is one of the learning algorithm baseline that consider the probability of each class. Naïve Bayes learns from the historical data and reflects how the data is generated. Naïve Bayes assumes that a specific feature is independent of the other feature on the given class variable. Despite the model is generated on a naïve assumption, Naïve Bayes works well in a real-world complexity. The model is constructed based on the formula 1.

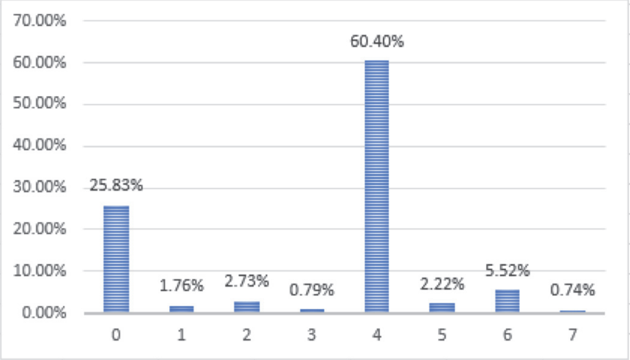

Figure 1 Initial (Imbalanced) Class Distribution

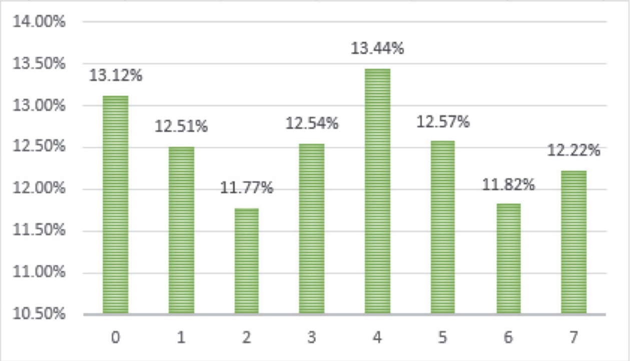

Figure 2 Class Distribution after performing SMOTE-ENN

$$\hat{y} = \underset{k\in\{1.....K\}}{\operatorname{argmax}} P(C_k) \prod_{i=1}^{n} p(x_i|C_k) \qquad (1)$$

Bagging (Bootstrapped Aggregating) use sampling by replacement to build the model. Bagging exploits many independent learner and combine it using averaging technique. It works well to reduce the variance and bias. In this work, we use decision tree as the weak learner.

Boosting using a weak learner to build a strong learner. During the training phase, a weight (w_t) is assigned in each weak leaner (k_i). The weight will be adjusted subsequently by using the error of the previous weak learner as illustrated in formula 2 and 3. In this study, decision tree is chosen as the weak learner because of its simplicity and it also performs well in large scale data.

$$C_{m-1}(x_i) = \alpha_1 k_1(x_i) + \cdots + \alpha_{m-1} k_{m-1}(x_i) \qquad (2)$$

Table 1 The selected feature

No	Feat	Desc	Value
1	sexo	*Customer's sex*	0: H (*female*) 1: V (*male*)
2	age	*Age*	*Integer*
3	antiguedad	*Customer seniority (in months)*	*Integer*
4	renta	*Gross income of the household*	*Integer*
5	segmento	*Customer segmentation*	*0: VIP* *1: Individuals* *2: College graduated*
6	Results	*Product recommendation*	0: didn't purchase 1: buy direct debits 2: buy e-accounts 3: buy direct debits & e-accounts 4: buy current accounts 5: buy current accounts & direct debits 6: buy current accounts & e-accounts 7: buy 3 products

$$C_m(x_i) = C_{m-1}(x_i) + \alpha_m k_m(x_i) \qquad (3)$$

Random Forest is a well-known ensemble method that use decision tree as the weak learner. This learning algorithm can handle overfitting, biased data, and run efficiently in large scale data. The construction of the model applies the general rule of bagging. Random forest adds a procedure called feature bagging. This procedure aims to choose the most representative feature for model building. We use 51 trees as the weak learner.

Imbalanced data is one of the main challenge in classification task. Classification performance relies highly on the training data. Imbalanced data will generate the model that is bias to the majority rather than minority class. On the previous work [15], We have examined the imbalanced data by generating artificial data based on the particular distribution. For this problem, we employ the combination of oversampling and under-sampling. For the oversampling we use SMOTE meanwhile for the under-sampling we use Edited Nearest Neighbors. The class distribution after performing SMOTE-ENN is illustrated in figure2.

IV. Result

All the experiments were conducted in Spark environment but the Bagging. The Bagging experiment used python's sk-learn library. In this experiment, we are evaluating the performance of 2 baseline learning and 3 ensemble learning. Furthermore, we also examine the performance of SMOTE-ENN on handling the imbalanced data. The experiments were conducted in three training-testing data ratios i.e. 70-30, 80-20, and 90-10.

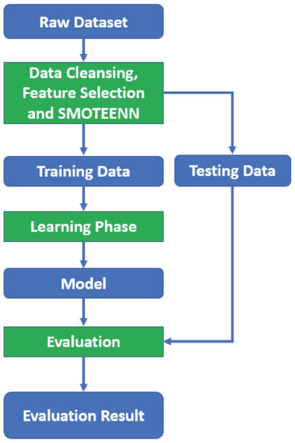

Figure 3 The Experimental pipeline for over- and under-sampling using SMOTE and Edited Nearest Neighbors

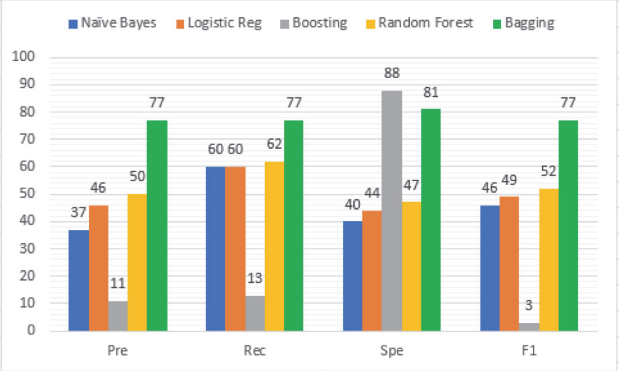

Figure 4 The experiment result of in imbalanced initial data on 70-30 training-testing ratio

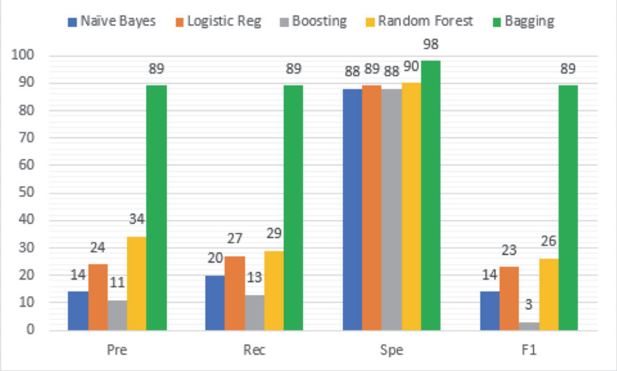

Figure 5 The experiment result of resampling data on 70-30 training-testing ratio

For the performance evaluation, we are using 4 metrics namely precision, recall, specificity, and F1-score. In addition, confusion matrix is used to see the details of system confusion on recognizing common class. Based on the experiment results in all training-testing ratio, Bagging achieved a superior performance compared to the other ensemble and baseline learnings. However, Bagging has a slightly lower specificity score than Boosting in the initial data. The experiment results depicted in figure 4-9.

Naïve Bayes, Logistic Regression, Boosting, and Random Forest tends to bias to the majority class (class 4 and class 0) in the imbalanced data. As can be seen in the confusion matrix, most of the testing data is recognized as class 4 or class 0. Meanwhile, Bagging has a more robust result and less misclassification. It verifies that Bagging has a better performance on handling extreme imbalanced data rather than other ensemble learning approach. The confusion matrix on imbalanced data are depicted in figure 10-12.

For the balanced data, Bagging achieved a superior performance compared to other ensemble and baseline learnings. Furthermore, the resampling (SMOTE and ENN) successfully improved the performance of the Bagging. Meanwhile, the resampling hinders the performance of other learning methods. However, resampling demonstrated that it helps to minimize the bias to the majority class. As can be seen on the confusion matrix in figure 13-15, the learning method is able to recognize the minority class despite having numerous misclassifications.

Bagging has a superior performance in imbalanced and resampling dataset. It is demonstrated that Bagging is robust enough to handle imbalanced data and avoid the bias. By using averaging approach, it will avoid the bias and variance that affects the learning algorithm performance. However, Naïve Bayes and Logistic Regression still has a better performance compared to Boosting. In fact, the classifier causes most predictions to be on the majority class and not the minority class. Meanwhile, the Boosting performance relies highly on the initial base learner performance. If the initial leaner is biased to the majority class, the next learner will keep on predicting the minor class into majority class.

Random Forest has a similar approach to Bagging but with additional procedure on feature selection. This procedure didn't achieve a good result because the dataset is already pre-processed using data cleansing and feature selection. Furthermore, the feature bagging aggravates the learning that leads to poor model building.

V. Conclusion

Ensemble learning uses set of weak classifiers and classify new data by taking (weighted) vote of the predictions. Imbalanced data is a special case in classification where the distribution of the class is not uniformed and will affect the generated model that biased to the majority class. In this study, we are evaluating the performance of ensemble learning on imbalanced data.

Based on the experiments, Bagging performed well in imbalanced data compared to other baseline learning and ensemble methods. Average weight approach is suited to handle bias and variance in data training. Resampling can be exploited to improve the learning algorithm performance on imbalanced data. Based on the experiment, it can gain the performance more than 10% on each metrics.

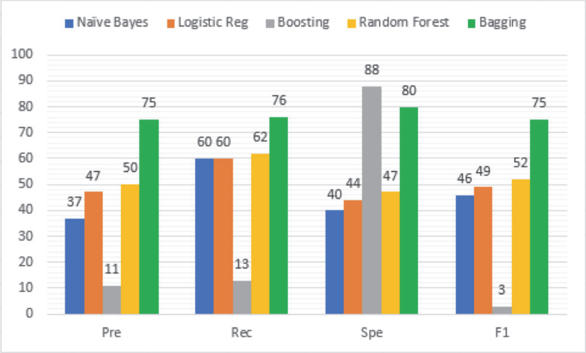

Figure 6 The experiment result of imbalanced data on 80-20 training-testing ratio

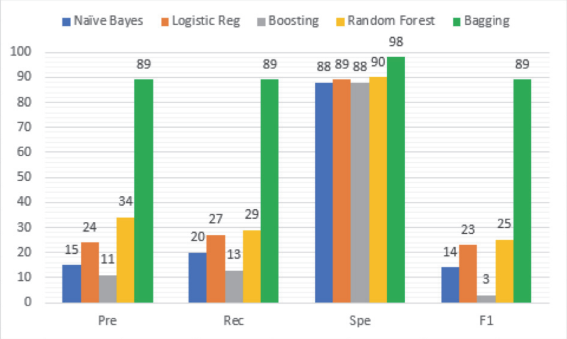

Figure 7 The experiment result of resampling data on 80-20 training-testing ratio

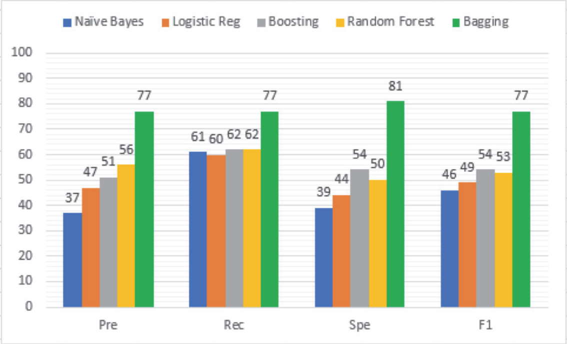

Figure 8 The experiment result of imbalanced data on 90-10 training-testing ratio

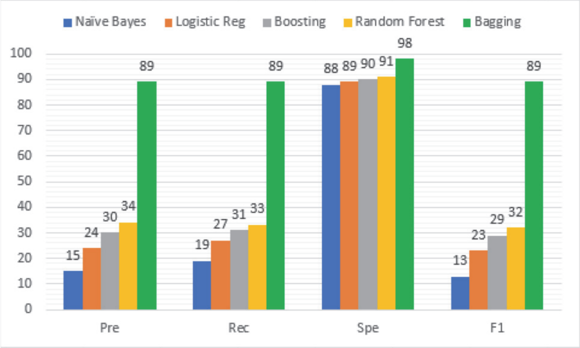

Figure 9 The experiment result of resampling data on 90-10 training-testing ratio

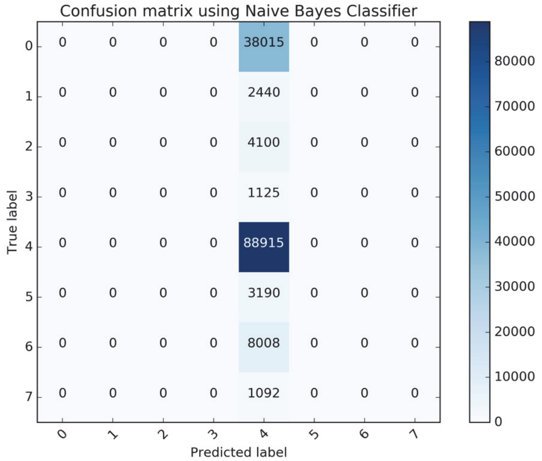

Figure 10 The confusion matrix of Naïve Bayes on imbalanced data using 90-10 training testing ratio

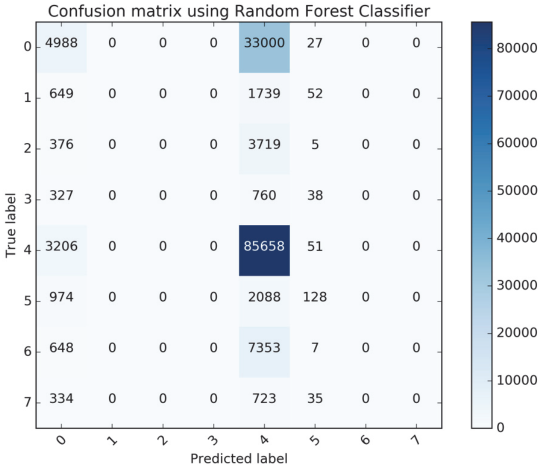

Figure 11 The confusion matrix of Random Forest on imbalanced data using 90-10 training testing ratio

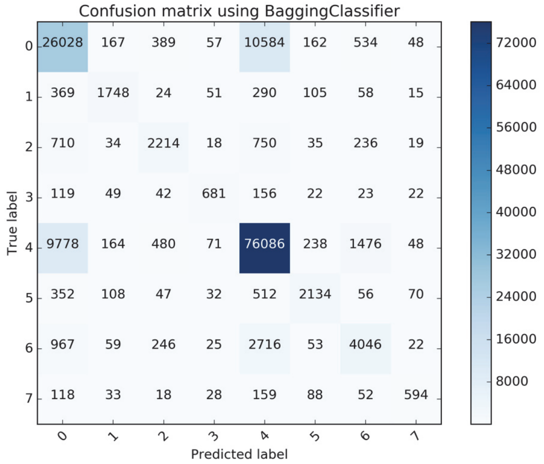

Figure 12 The confusion matrix of Bagging on imbalanced data using 90-10 training testing ratio

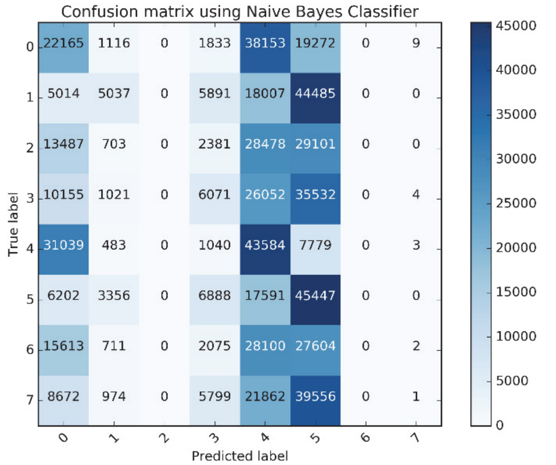

Figure 13 The confusion matrix of Naïve Bayes on resampled data using 90-10 training testing ratio

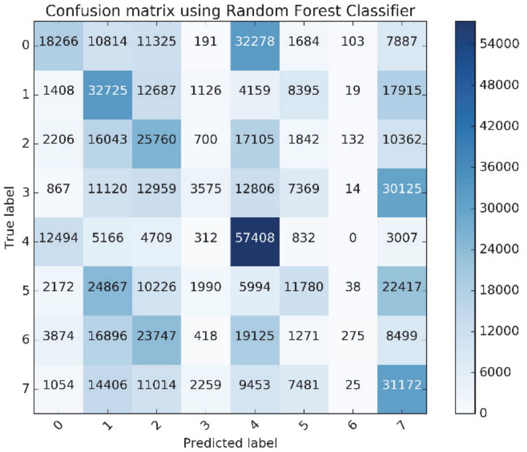

Figure 14 The confusion matrix of Random Forest on resampled data using 90-10 training testing ratio

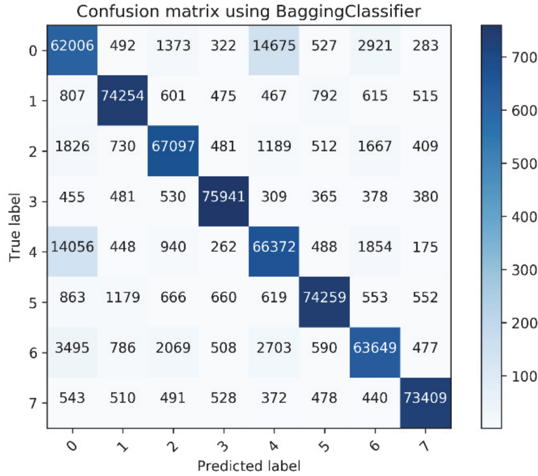

Figure 15 The confusion matrix of Bagging on resampled data using 90-10 training testing ratio

Acknowledgment

This article's publication is partially supported by the United States Agency for International Development (USAID) through the Sustainable Higher Education Research Alliance (SHERA) Program for Universitas Indonesia's Scientific Modeling, Application, Research and Training for City-centered Innovation and Technology (SMART CITY) Project, Grant #AID-497-A-1600004, Sub Grant #IIE-00000078-UI-1.] No:0141/UN2.R3.SC/HKP.05.01/2018. We would thank Ari Wibisono for his support in supplying the dataset, infrastructure (both GPU and Spark environment), source codes, and giving tutorials.

References

[1] S. Sobolevsky, E. Massaro, I. Bojic, J. M. Arias, and C. Ratti, "Predicting Regional Economic Indices using Big Data of Individual Bank Card Transactions," arXiv, vol. 1506.00036, pp. 1313–1318, 2015.

[2] A. Rajwani, T. Syed, B. Khan, and S. Behlim, "Regression Analysis for ATM Cash Flow Prediction," 2017 Int. Conf. Front. Inf. Technol., pp. 212–217, 2017.

[3] W. Jatmiko, Rochmatullah, B. Kusumoputro, H. R. Sanabila, K. Sekiyama and T. Fukuda, "Visualization and statistical analysis of fuzzy-neuro learning vector quantization based on particle swarm optimization for recognizing mixture odors," 2009 International Symposium on Micro-NanoMechatronics and Human Science, Nagoya, 2009, pp. 420-425.

[4] S. C. Purbarani, H. R. Sanabila, A. Bowolaksono and B. Wiweko, "A survey of whole genome alignment tools and frameworks based on Hadoop's MapReduce," 2016 International Workshop on Big Data and Information Security (IWBIS), Jakarta, 2016, pp. 65-70.

[5] A. Wibisono, H. A. Wisesa, W. Jatmiko, P. Mursanto and D. Sarwinda, "Perceptron rule improvement on FIMT-DD for large traffic data stream," 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, 2016, pp. 5161-5167.

[6] A. Gahlaut and P. K. Singh, "Prediction analysis of risky credit using Data mining classification models," 2017.

[7] G. Sudhamathy and C. Jothi Venkateswaran, "Analytics using R for predicting credit defaulters," 2016 IEEE Int. Conf. Adv. Comput. Appl. ICACA 2016, pp. 66–71, 2017.

[8] L. X. Wang, "Modeling Stock Price Dynamics with Fuzzy Opinion Networks," IEEE Trans. Fuzzy Syst., vol. 25, no. 2, pp. 277–301, 2017.

[9] V. Behbood, J. Lu, G. Zhang, and W. Pedrycz, "Multistep Fuzzy Bridged Refinement Domain Adaptation Algorithm and Its Application to Bank Failure Prediction," IEEE Trans. Fuzzy Syst., vol. 23, no. 6, pp. 1917–1935, 2015.

[10] J. Xue, S. H. Zhou, Q. Liu, X. Liu, and J. Yin, "Financial time series prediction using ℓ2,1RF-ELM," Neurocomputing, vol. 277, pp. 176–186, 2017.

[11] G. Song and Q. Dai, "A novel double deep ELMs ensemble system for time series forecasting," Knowledge-Based Syst., vol. 134, pp. 31–49, 2017.

[12] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," Eur. J. Oper. Res., vol. 0, pp. 1–16, 2018.

[13] Wibisono, A., Suhartanto, H. Cloud computing model and implementation of molecular dynamics simulation using Amber and Gromacs (2012), 2012 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2012 - Proceedings, art. no. 6468763, pp. 31-36.

[14] Wiska, R., Habibie, N., Wibisono, A., Nugroho, W.S., and Mursanto, P. Big sensor-generated data streaming using Kafka and Impala for data storage in Wireless Sensor Network for CO2 monitoring (2017) 2016 International Workshop on Big Data and Information Security, IWBIS 2016, art. no. 7872896, pp. 97-101.

[15] Sanabila, H. R., Kusuma, I., & Jatmiko, W. (2016, October). Generative oversampling method (GenOMe) for imbalanced data on apnea detection using ECG data. In Advanced Computer Science and Information Systems (ICACSIS), 2016 International Conference on (pp. 572-579). IEEE.