Journal of Computer Science

Original Research Paper

Comparison Between Selective Sampling and Random Undersampling for Classification of Customer Defection Using Support Vector Machine

[1]Heri Kuswanto, [1]Yogi Sarumaha and [2]Hayato Ohwada

[1]Department of Statistics, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia
[2]Department of Industrial Administration,
Graduate School of Science and Technology, Tokyo University of Science, Noda Chiba Japan

Article history
Received: 25-01-2017
Revised: 18-04-2017
Accepted: 16-08-2017

Corresponding Author:
Heri Kuswanto
Department of Statistics,
Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia
Email: heri_k@statistika.its.ac.id

Abstract: Quality of a product determines the customer loyalty and it can be measured by conducting a survey. Company 'X' that sells three kinds of product (low, medium and high price) collected very large dataset through an online survey and recorded customer defection and their characteristic. The measured variables are Update Accumulation, Product Price, Customer Type, Delivery Status and Customer Defection. The data has an imbalanced response that could mislead the accuracy of classification if it is analyzed using standard approaches. Selective Sampling (SS) and Random Undersampling (RU) have been applied to draw a sample from imbalance response in order to obtain more balance data. Furthermore, Support Vector Machine (SVM) has been applied to classify the sampled data. The performance of the SS-SVM and SS-RU to classify sampled data has been evaluated and compared with the result of classifying the raw dataset. The RU yields on exact balance (50%:50%) response class, while SS reduce the imbalance proportion significantly (around 52%:48%). Nevertheless, the SS-SVM outperforms RU-SVM in the sense that it is capable to run the process effectively, where the SS-SVM reduces the duration of classification process 3 to 20 h shorter than using RU-SVM, with slightly different accuracy rate. Moreover, the SS-SVM maintains the basic characteristics of raw data better than RU-SVM.

Keywords: Sampling, Imbalance, SVM, Defection

Introduction

Customer loyalty is highly influenced by the quality of product and service provided by the company. Studying customer loyalty can be conducted through an online survey. Hague and Hague (2015) argues that survey can be an efficient way to identify whether the consumer will tend to be loyal or not. Company "X" is a cloud based software company that sells antivirus products with three price categories i.e., Low Price (LP), Medium Price (MP) and High Price (HP). Statistic data shows that currently there has been a significant increase (around 36%) on the number of the cloud-based company in 2016 (Columbus, 2013). This growth shows that internet based company should think smartly to provide comfort and convenience service to the customers in order to maintain the customer loyalty. Company "X" is a big cloud company based in Japan which also has to maintain their customer loyalty and hence, the company conducts a survey to study their customer behavior towards defection case. One of the interesting information that needs to be gathered from the survey is about customer defection. The online survey that has been conducted are able to generate a big dataset with a very large number of sample. Furthermore, the collected data has characteristic of imbalanced response between defective and non-defective customers.

Big data phenomena create a challenge to researchers since statistical parametric approaches become less reliable when applied to very large data, that will lead to inferential statistics using a null hypothesis that tend to be insignificance. Lin *et al*. (2013) found that with very large data, p-value tends to fall to zero and this leads to bias conclusion. A computational approach is required to analyze big data in order to have faster, representative

and reliable result. Another challenge is the imbalance response. Balanced data is achieved when the response classes have balanced proportion. Sain (2013) observed that machine learning approaches applied to imbalanced data may lead to bias classification accuracy since higher accuracy will belong to the majority class. Meanwhile, the minor class will be underestimated. Choi (2010) suggests three approaches deal with imbalance data i.e., reduce the sample, adjust the classification method and combine few classification methods with ensemble learning.

Studies about predicting customer defection of company "X" have been carried out by several researchers. Prasasti *et al*. (2013) predicted customer defection of company 'X' using C4.5 Decision Tree and SVM. Using the same dataset, Martono (2014) used J48 Decision Tree classification method (J48), Random Forest (RF), Neural Network with Multi Layer Perception (MLP), as well as SVM with SMO algorithm. They found that J48, RF and SMO provide highest classification accuracy for HP product, meanwhile, MLP performs well for MP. However, these researches neglected imbalance issue on the data. Kuswanto *et al*. (2015) used logistic regression-based approach for the classification. The previous researches were conducted by classifying raw dataset and it is considered as inefficient with refers to the duration of running the process.

This paper applies SVM to the sampled data as one of the strategies to deal with the issue of very large data with imbalance response, where the data will be preprocessed using SS and RU. There have been many studies showed that the classifier such as SVM will generate biased classification output, due to neglecting the imbalance issue. The classifier is more sensitive to detecting the majority class and less sensitive to the minority class and hence, preprocessing stage is required for this case i.e., by undersampling the majority class or oversampling the minority class.

The Random Undersampling (RU) is one of the most popular approaches to reduce the number of samples. Dittman *et al*. (2014) proved that RU outperforms some other sampling approaches such as Random Oversampling and Synthetic Minority Oversampling Technique (SMOTE). They recommended to using Random Undersampling over Random Oversampling and SMOTE for the purposes of data sampling due to its computational costs and the end result of reducing the size of the dataset. D'Addabbo and Maglietta (2015) introduced sampling method by considering imbalanced data class namely selective sampling. The method performs very well in their case. Both papers found that RU and SS have good performance compared with other approaches to classify imbalance data. However, the comparison between them has never been investigated. Therefore, this paper will study the performance of both SS and RU to detect customer defection in company "X", combined with SVM as the classification method.

Literature Review

Selective Sampling

The selective Sampling method is applied to a very large dataset. Moreover, imbalanced data class can also be overcome using this method. Sampling is an under sampling method that reduces the majority class based on Tomek Links. Tomek Links removes negative class and positive class that shares the same characteristic. Given $\{E_1,\ldots,E_n\}\in R^k$, a pair of $\{E_i,E_j\}$ defined as Tomek Links if E_i and E_j have different labels without E_l where $d\{E_i, E_l\}<\{E_j, E_l\}$ or $d\{E_j, E_l\}<\{E_i, E_l\}$, where d(.,.) measures Euclidian distance. Tomek Links is used to reducing the majority class that has the same input space from the same class.

Given S = $\{(x_1,y_1),\ldots,(x_l,y_l)\}$ is a training where $x_1\in R^k$ and $y_i\in(0,1),\forall_i = 1\ldots l$. S_0 is defined as training dataset l_0 which also belong to the class $y = 0$ and S_1 is defined as training dataset l_1 which also belong to the class $y = 1$, with $l_0>>l_1$. The larger class can be found using Selective Sampling without response variable. Selective Sampling reduces the training data with the percentage of M% from the minority class to the total amount of sample. Population data that holds the major class and minor class is separated to decide which data have the larger number than another. The following procedure will be applied to reduce the major class to achieve a more balance data (D'Addabbo and Maglietta, 2015):

- *Tomek links.* Given that a set of T^n from each major class S_0^n is the first neighbour from the first sample in S_1, where $T^n=\left\{x\in S_0^n\,|\,(x,z)\right\}$ is Tomek Links for $\left\{S_1\cup S_0^n, z\in S_1\right\}$
- *Data reduction.* Given that $\bar{x}\in D^n = S_0^n\setminus T^n$, the data reduction steps are described as follows:
  - Tomek Links $(\bar{x},\bar{z})$ is calculated from dataset $\bar{x}\cup S_1$ with $z\in S_1$
  - Eucledian distance $d(\bar{x},x)$ is calculated for each $x\cup S_0^n$
  - Subset is defined as $L=\left\{x\cup S_0^n\middle| d(\bar{x},x)<d(\bar{x},\bar{y})\right\}$. Tomek Link $(x^*,\bar{z})$ from $\bar{z}\cup L$ measured, for example with x^* is defined as the first neighbour for L from $\bar{z}$
  - Defines $R=\left\{x\in L\middle| d(\bar{x},x)<\left[d(\bar{x},\bar{z})-d(x^*,\bar{z})\right]\right\}$, then for each point of R in Tomek Links as an

example $x \in R'$ with $R' = \{x \in R | x \notin T^n\}$. The remaining points of major class belongs to $s_0^{n'} = s_0^n / R'$

- If the class balance is not achieved, the algorithm follows steps c and randomly selects $\bar{x} \notin D^{n'} = S_0^{n'} \setminus T^n$ and repeats step b

- *Joining residual data*. The reduced majority class are joined, then continue to step d
- *Last Elimination*. Step c above is repeated until we obtain M% in the minor class

Random Undersampling

The Random Undersampling method can effectively handle classification case for imbalanced data. Different from the complex data sampling algorithm, RU simply removes training data set until balanced data achieved (Catal, 2012). Sampling method of Random Undersampling is begin with dataset selection and then continued to find the gap between majority class and minority class that have imbalanced class, if there are gap between the class, the major class will be removed until balanced are achieved for each of the class with the same amount for majority class and minority class.

Support Vector Machine

Support Vector Machine (SVM) method is a machine learning method developed by Boser, Guyon and Vapnik utilyzing computational theory such as kernel developed by Aronszajn in 1950, Lagrange Multiplier of Joseph Louis Lagrange in 1766 and other supporting theories (Vapnik, 1995). SVM is a prediction technique in regression and classification. SVM is used to obtain optimal hyperplane to distinguish observation that has target variable. Moreover, SVM is able to find the optimum solution in each running (Seiffert *et al*., 2010) This research applies SVM method because of the efficiency in solving classification for binary class (Miner *et al*., 2012). Even if this method effective in binary class, SVM have disadvantages for large data since it highly depends on the amount of data to be analyzed.

Denote a data in $\vec{x}_l \in R^d$ where each label noted as $y_i \in \{-1, +1\}$ for $i = 1,2,\ldots,l$ where l is the amount of data. For each class -1 and +1 which is separated completely by hyperplane d dimension, defined as:

$$\vec{w}.\vec{x} + b = 0 \tag{1}$$

The value of $\vec{x}_l$ which belongs to negative sample is formulated under the following equation:

$$\vec{w}.\vec{x} + b \leq -1 \tag{2}$$

while $\vec{x}_l$ which belongs to positive sample $+1$ is formulated with:

$$\vec{w}.\vec{x} + b \geq +1 \tag{3}$$

The largest margin is defined by maximizing the distance between hyperplane with the nearest point noted by $\frac{1}{\|\vec{w}\|}$. The *Quadratic Programming (Q)* is used to find an optimal point from Equation 4 with the constraint given in Equation 8:

$$\min_w \frac{1}{2} w^T Q w - e^T w \tag{4}$$

$$0 \leq w_i < C, \mathrm{i} = 1,\ldots,l, y^T w = 0 \tag{5}$$

where, e is a unit vector, C is the upper boundary and Q is a semidefinite matrix with a size of 1×1. The equation above can be solved using Lagrange Multiplier:

$$(w,b,a) = \frac{1}{2}\overline{\|w\|}^2 - \sum_{i=1}^{l}\left(\alpha_i\left(y_i\left((x_i)w_i + b\right) - 1\right)\right); \quad i = 1,2,\ldots,l \tag{6}$$

where, α_i is Lagrange multiplier with zero or positive value ($\alpha_i \geq 0$). The optimum value in Equation 6 is calculated by minimizing L to $\vec{w}$ and b and by maximizing L to α_i. By considering the optimal gradient point $L = 0$, Equation 6 can be modified so that only α_i remain as stated in Equation 7:

$$\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j \overrightarrow{x_i x_j} \tag{7}$$

with constraint as follow:

$$\alpha_i \geq 0 \left(i = 1,2,3,\ldots,l\right)\sum_{i=1}^{l}\alpha_i y_i = 0 \tag{8}$$

In general, α_i will be positive and the correlation between these positive α_i is called as support vector with the assumption that both classes will be separated perfectly by the hyperplane (Han *et al*., 2012). The linear equation of SVM is as follows:

$$\min\left(w,\gamma,\mathrm{y}\right) \in R^{R+1+mve^T y + \frac{1}{2}w^T w} \tag{9}$$

Where:

w: Normal vector $n \times 1$

γ: Value to decide the location of relative hyperplane to actual class

v: Slack vector variable $m \times 1$ measuring classification error and non negative
e: A vector with size nx1

The optimization process in SVM differs with some other optimization procedures applied in popular computational approaches such as neural network. Some interesting application of the neural network optimization can be found in Valipour (2016; Valipour and Gholami Sefidkouhi, 2017) among others.

Accuracy

To evaluate the performance of the classifier, it can be done by measuring accuracy and specificity (Baratloo *et al*., 2015; Astuti *et al*., 2014). The accuracy describes the total classification of the data that are classified correctly by the classifier, where higher accuracy means better classification. If the class number is two, the following Table 1 shows the classification between predictions and actual class.

The classification accuracy is measured by dividing the correct prediction with the total amount of prediction. The classification accuracy can be measured by the following criteria:

$$sensitivity = \frac{TP}{(TP + FN)} \tag{10}$$

$$specivicity = \frac{TN}{(FP + TN)} \tag{11}$$

$$Accuration = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{12}$$

The choice of the best method is evaluated by considering those three criteria as well as the runtime. In a case of predicting continuous response, the criteria are equivalent to the minimizing the root mean square errors among the compared models (Valipour *et al*., 2017 as an example).

Research Methodology

Data Source

The data used in this study are secondary data that has been used and pre-processed by Prasasti *et al*. (2013; Martono, 2014; Kuswanto *et al*., 2015). The negative class is the larger class and the positive class is the least class. The original data in the study owned by company "X" providing internet-based antivirus software collected within the period of 2007 to 2013. The following Table 2 provides information about the data structure.

The analyzed data is the record of consumer activities dealing with purchasing the products from the company 'X'. The variables consist of one response variable (Y) and four predictor variables (X). The detail of variables are provided as follows

Accumulation Update (X_1)

Accumulation Update variable is the update since purchase or renewal of a product. When the customer purchased or updated the product, the Accumulation Update increases by one.

Product Price (X2)

Product Price variable is the cost of purchased product (measured in Japan Yen (JPY))

Consumer Type (X3)

Consumer type is a type of consumer where 0 indicates personal use and 1 indicates organizational use

Delivery Status (X4)

Delivery Status is indicated with 0 if the email is failed to be sent and 1 if it is successfully sent.

Consumer Defection (Y)

Consumer Defection variable is a response variable with 1 represents consumer decided not to use the product anymore and 0 when consumer decided to continue using the product.

Table 1. Cross tabulation of prediction classification and actual classification

		Actual class	
		p(+)	*n*(-)
Prediction class	*p*(+)	*True Positive* (TP)	*False Positive* (FP)
	n(-)	*False Negative* (FN)	*True Negative* (TN)

Table 2. The amount of data all products

Product category	Number of raw data	Percentage of data class	Total of negative class	Total of positive class
LP	500000	58,6%: 41,4%	293102	206898
MP	408810	66,8%: 33,2%	273083	135727
HP	709989	54,4%: 45,6%	386431	323558

Analysis Steps

This research uses Selective Sampling and Random Undersampling to overcome imbalanced data to be further classified using SVM. SS-SVM and RU-SVM have been applied to Low Price, Medium Price and High Price data. Moreover, SVM classification for raw data will also be performed to be compared with reduced sample. The steps of the analysis are as follows:

- Selective Sampling for all products with SVM classification: Tomek Links, Data reduction, Combining residual data, Last elimination, Applying SVM from sampled data, Measuring duration and classification accuracy
- Random Undersampling for all products with SVM classification: Measuring difference between major class and minor class, Data reduction until balanced data obtained, applying classification with SVM from sampled data, measuring duration and classification accuracy
- SVM classification using raw data for all products and measure the running duration
- Select the best method based on the best classification accuracy as well as duration to run the process

Computer Specification

One of the indicators to assess the effectiveness of the method is by measuring the duration to run classification process using each approach. To deal with this, the device specification is an important factor. This research uses a computer with 3GB RAM of Windows 7 32-bit with 2.27 Ghz processor.

Results and Discussion

Sampling with Selective Sampling and Classification with SVM (SS-SVM)

Table 3 describes the percentage of the class response before and after sampling by SS. For the LP product, the raw data of 500000 is reduced into 293102 after sampling and it is more balance than the raw percentage i.e., 53.7%:46.3%. The MP product has the largest imbalanced raw data of 66.8%:33.2% with the amount of population data 408810 and after applying SS, the data is reduced to 135727 with the class percentage of 52.4%:47.6%. The HP product has the largest number of raw data i.e., 709989 with the percentage class of 54.5%:45.6%. After applying SS, the data is dropped to 323445 with class percentage of 52.4%:47.6%. For all categories, the raw data has been successfully reduced into more balanced class. The fact that the class proportion of the sampled data is not perfectly balanced (50%:50%) provides an interesting feature of the Selective Sampling procedure. It shows that the Selective Sampling algorithm involves of optimizing an objective function as indicated in the previous section.

Table 4 shows the runtime of carrying out sampling process and further continued with classification using SVM. The accuracy of each method is tabulated as well. Low Price product has the longest runtime compared to all product categories with the accuration of 65.29%. For Low Price product, the duration of the sampling and classification process is 9.40 h and the accuracy is 67.37% which is the highest accuracy. The Medium Price product has the runtime of 5.38 h with the accuracy 62.85%.

Sampling with Random Undersampling and Classification with SVM (RU-SVM)

Table 5 shows the amount of raw data for all categories with the comparison of percentage class from the data. Table 5 also provides Random Undersampling sampled data with its class percentage. Similar to SS, the sampling is effectively reduced the number of raw data into more balance class. Random Undersampling yields into more balanced class with the exactly 50%:50% to all product category. Under the Random Undersampling, the end results of the class proportion are simply determined by sampling in random the majority class. Consequently, the change in number of the sampled data is driven by the number of the minority class. The highest changing of class percentage holds by Medium Price.

Table 3. Comparison of raw data and selective sampling sample

Product category	Amount of raw data	Class percentage	Amount of sampled data	Sampled data class percentage
LP	500000	58.6%: 41.4%	293102	53.7%: 46.3%
MP	408810	66.8%: 33.2%	135727	52.4%: 47.6%
HP	709989	54.4%: 45.6%	323445	52.4%: 47.6%

Table 4. Duration and accuracy of process by SS-SVM

Product category	Duration (h)	Accuracy (%)
LP	11.25	65.29
MP	5.38	62.86
HP	9.40	67.37

Table 5. Comparison of data and random undersampling sample

Product category	Amount of raw data	Class percentage	Amount of sampled data	Sampled data class percentage
LP	500000	58.6%: 41.4%	413796	50%: 50%
MP	408810	66,8%: 33.2%	271454	50%: 50%
HP	709989	54.4%: 45.6%	646936	50%: 50%

Table 6. Duration and accuracy of process by RU-SVM

Product category	Duration (h)	Accuracy (%)
LP	21.34	64.94
MP	8.11	73.33
HP	29.27	68.38

Table 6 shows the duration and the classification accuracy of the RU-SVM method applied to classify LP, MP and HP products. Low Price has runtime of 21.34 h with the accuracy of 64.94%. This result is different with the Medium Price which has the least duration of 8.11 h with accuracy of 73.33%. Meanwhile, the High Price product has the longest running duration i.e., 29.27 h. This happens because the product has larger dataset than the others with the classification accuracy of 68.38%.

The long duration of the running process for all three cases in line with the number of sampled data obtained by Random Undersampling process. There is a slight improvement in term of the classification accuracy compared with SS-SVM, it might due to the fact that RU process yields on exactly balance sample, which fit with the basic idea of classification using SVM.

Classification of raw Data with SVM

All of the SVM results are performed with linear kernel specification. The results of the classification as well as the runtime is performed in Table 7.

Table 7 shows the duration and classification accuracy of applying SVM to raw data. No wonder that the duration of the process is much longer than the sample data. For LP, the duration of performing classification is 24.39 h with 67.02% total accuracy. Meanwhile, the runtime for MP reaches 12.09 h with accuracy of 68.92%. The lowest accuracy is for LP and the longest runtime belongs to the HP.

Table 7. Runtime and accuracy of process by SVM of raw data

Product category	Runtime (h)	Accuracy (%)
LP	24.39	67.02
MP	12.09	75.33
HP	30.06	68.92

Table 8. Comparison among methods

Product criteria		SS-VM	RU-SVM	SVM
LP	Duration	11.25 h	21.34 h	24.39 h
	Accuracy	65.29%	64.94%	67.02%
	Sensitivity	62.14%	56.43%	66.27%
	Specificity	76.72%	73.45%	67.28%
MP	Duration	5.38 h	8.11 h	12.09 h
	Accuracy	62.86%	73.33%	75.33%
	Sensitivity	67.11%	75.54%	62.13%
	Specificity	60.09%	71.46%	82.49%
HP	Duration	9.40 h	29.27 h	30.06 h
	Accuracy	67.37%	68.38%	69.92%
	Sensitivity	65.36%	62.25%	67.15%
	Specificity	71.06%	74.52%	70.22%

Best Method Selection

The best method in this research is characterized by the runtime (indicates the effectiveness) and the accuracy of doing classification. Table 8 summarizes the performance of all methods applied to those three different products. Sensitivity and Specificity will be compared as well. For LP product, the least duration belongs to SS-SVM with 11.25 h. For MP product, the least duration belongs to SS-SVM with 5.38 h. For HP product, the least duration belongs to SS-SVM with 9.40 h. From all results above, the runtime has positive correlation with the number of sampled data. The SS reduces the data significantly compared to the RU, which thus.

The accuracy of SVM without sampling yileds on slightly higher accuracy than the classification of sampled data. However, this could happen since SVM tends to neglect the least class from the data as for imbalanced data condition. In this case, we should consider the sensitivity and specificity. Therefore, classification with sampled data is better. The SS-SVM and RU-SVM compete each other, indicated by the inconsistency of the accuracy within those three products. Among all running processes, SS-SVM shows the shortest duration of performing the classification and data reduction process. The different is highly significant, where using SS-SVM will need only about half time than RU-SVM to run the process. For HP, SS-SVM reduces the duration much shorter with about 20 h.

Conclusion and Recommendation

Based on the analysis, we conclude as follows:

- Sampling methods are able to provide less amount of data compared to raw data with more balance class. Characteristics of raw data can be well

maintained with both Selective Sampling and Random Under sampling. In this case, the Selective Sampling is more efficient as the number of sampled data is much lower than obtained with Random Undersampling
- SS-SVM can reduce the runtime significantly compared to RU-SVM. The process using SS-SVM can be 20 h more efficient than SU-SVM. Using raw data might yield on higher accuracy but is inefficient
- The accuracy between SS-SVM and RS-SVM is case dependent and they compete each other with refers to the sensitivity and specificity

There are several sources of uncertainty that might influence the results of this study such as different time of the running process, the choice of SVM parameters which depend on the input of initial values, etc. In order to reduce the uncertainties, the authors recommend using cross validation approach during the classification process with SVM. This procedure will reduce the uncertainty by evaluating the classification accuracy among different sample classes. Another important thing is the computer specifications have to be exactly the same and must be run and evaluated at the same time.

Acknowledgement

Authors acknowledge the Ministry of Research, Technology and Higher Education Indonesia for the funding of this research through International Collaboration and Scientific Publication Research Grant.

Author's Contributions

Heri Kuswanto: Responsible for conducting the whole process of data analysis as well as manuscript preparation.

Yogi Sarumaha: Helps the first author for management as well as pre-processing the data.

Hayato Ohwada: Contributes on providing the dataset, advising the steps of analysis as well as manuscript preparation.

References

Astuti, A., N. Iriawan, Irhamah and H. Kuswanto, 2014. Kolmogorov-Smirnov and continuous ranked probability score validation on the Bayesian model averaging for microarray data. Applied Math. Sci., 8: 7277-7287. DOI: 10.12988/ams.2014.49760

Baratloo, A., M. Hosseini, A. Negida and G. El Ashal, 2015. Part 1: Simple definition and calculation of accuracy, sensitivity and specificity. Emergency (Tehran), 3: 48-49. DOI: 10.22037/emergency.v3i2.8154

Catal, C., 2012. Performance evaluation metrics for software fault prediction studies. Acta Polytechn. Hungarica, 9: 193-206.

Choi, J.M., 2010. A selective sampling method for imbalanced data learning on support vector machines. Graduate Theses, Iowa State University.

Columbus, L., 2013 Predicting enterprise cloud computing growth.

D'Addabbo, A. and R. Maglietta, 2015. Parallel selective sampling method for imbalanced and large data classification. Patt. Recogn. Lett., 62: 61-67. DOI: 10.1016/j.patrec.2015.05.008

Dittman, D.J., T.M. Khoshgoftaar, Wald, R., Napolitano, A., 2014. Comparison of data sampling approaches for imbalanced bioinformatics data. Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference.

Hague, P. and N. Hague, 2015. B2binternational.

Han, J., M. Kamber and J. Pei, 2011. Data Mining: Concepts and Techniques. 3rd Edn., Morgan Kaufmann, USA., ISBN-10: 9380931913, pp: 744.

Sain, H., 2013. Combine sampling support vector machine for imbalanced data classification. Proc. Comput. Sci., 72: 59-66. DOI: 10.1016/j.procs.2015.12.105

Kuswanto, H., A. Asfihani, Y. Sarumaha and H. Ohwada, 2015. Logistic regression ensemble for predicting customer defection with very large sample size. Proc. Comput. Sci., 72: 86-93. DOI: 10.1016/j.procs.2015.12.108

Lin, M., H.C.J. Lucas and G. Shmueli, 2013. Too big to fail: Large samples and the p-value problem. INFORMS.

Martono, N.P., 2014. Customer lifetime value and defection possibility prediction model using machine learning. Tokyo University of Science, Japan.

Miner, G., R. Nisbet, J. Elder, D. Delen and A. Fast *et al*., 2012. Practical Text Mining and Statistical Analysis for Unstructured Text Data Applications. 1st Edn., Academic Press, USA, ISBN-10: 0123870119, pp: 1000.

Prasasti, N., M. Okada, K. Kanamori and H. Ohwada, 2013. Customer lifetime value and defection possibility pre-diction model using machine learning: An application to a cloud-based software company. Proceedings of the 6th Asian Conference on Intelligent Information and Database Systems, (IDS' 13), Springer, pp: 62-71. DOI: 10.1007/978-3-319-05458-2_7

Seiffert, C., T.M. Khoshgoftaar, J. Van Hulse and A. Napolitano, 2010. RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Trans. Syst. Man Cybernet., 40: 185-197. DOI: 10.1109/TSMCA.2009.2029559

Valipour, M. and M.A. Gholami Sefidkouhi, 2017a. Temporal Analysis of reference evapotranspiration to detect variation factors. Int. J. Global Warm. Forthcom.

Valipour, M., M.A. Gholami Sefidkouhi and M. Raeini-Sarjaz, 2017b. Selecting the best model to estimate potential evapotranspiration with respect to climate change and magnitudes of extreme events. Agric. Water Manage., 180: 50-60. DOI: 10.1016/j.agwat.2016.08.025

Valipour, M., 2016. Optimization of neural networks for precipitation analysis in a humid region to detect drought and wet year alarms. Meteorol. Applic.. 23: 91-100. DOI: 10.1002/met.1533

Vapnik, V.N., 1995. Support-vector networks. Mach. Learn., 20: 273-297. DOI: 10.1023/A:1022627411411