Comparison of Two Main Approaches for Handling Imbalanced Data in Churn Prediction Problem

Nam N. Nguyen and Anh T. Duong
Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam
Email: nhatnamcse@gmail.com, dtanh@hcmut.edu

Abstract—**Customer churn is a major problem in several service industries such as banks and telecommunication companies for its profound impact on the company's revenue. However, the existing algorithms for churn prediction still have some limitations because the data is usually imbalanced. The commonly-used techniques for handling imbalanced data in churn prediction belong to two categories: resampling methods that balance the data before model training, and cost-sensitive learning methods that adjust the relative costs of the errors during model training. In this paper, we compare the performance of two data resampling methods: SMOTE and Deep Belief Network (DBN) against the two cost-sensitive learning methods: focal loss and weighted loss in churn prediction problem. The empirical results show that as for churn prediction problem, the overall predictive performance of focal loss and weighted loss methods is better than that of SMOTE and DBN.**

Index Terms—**churn prediction, deep belief network, SMOTE, focal loss, weighted loss**

I. Introduction

Timely prediction of customers' possibility to churn in several service industries, such as banks and telecommunication companies, has recently become a subject of focus for its impact on the business profit. The cost to retain existing customers is often lower than that of reaching out to new customers. This would also enhance the competitiveness of one service company among various service providers in the contemporary context of saturated customer base today. Recently, numerous data mining techniques have been applied for customer churn prediction, including traditional statistical methods, such as logistic regression [1], non-parametric models such as k-nearest-neighbors ([2], [3]), decision trees ([4], [5]), neural networks ([6], [7]), Support Vector Machines [8] and ensemble methods ([9], [10]).

In the churn prediction problem, the training data set used to train a prediction model often involves an inevitable quantitative imbalance between churn and non-churn groups; in particular, the leaving group only accounts for 2% of the total data. Classic algorithms often fail to handle this problem effectively due to its emphasis on the majority of non-churn customers, which leaves the prediction of churn customers vulnerable [11]. Thus, effective handling of the data imbalance problem is crucial in improving the model's predictive accuracy in practical applications.

During the last decade, many solving methods have been proposed to deal with imbalance data in classification, both for standard learning algorithms and for ensemble techniques ([12]-[14]). They can be grouped into categories:

1) Data resampling: In which the training instances are modified in such a way to produce a more or less balanced class distribution that enables classifiers to perform in a similar manner to standard distribution [15].
2) Cost-sensitive learning: This type of methods assigns higher costs for the misclassification of examples of the positive class with respect to the negative class, and therefore, trying to minimize higher cost errors [16].

In this work, we compare empirically the classification performance of data resampling approach against cost-sensitive learning approach in churn prediction problem. In the data resampling category, we select two methods: SMOTE [17], and the Deep Belief Network (DBN) generative model [18]. As for the cost-sensitive learning category, we shall examine two methods: Weighted Loss [19] and Focal Loss [20]. The empirical results on two datasets show that as for churn prediction problem, the overall predictive performance of Focal Loss and Weighted Loss methods is better than that of SMOTE and DBN.

The remainder of the paper is organized as follows. Section II presents the problems related to handling class imbalance. Section III explains the selected methods used in the performance evaluation process. Section IV reports the experimental results. The final section gives some conclusions and future works.

II. Handling Class Imbalance

There are two main problems that we have to handle when mining imbalanced classes: (1) how to select a proper evaluation metrics and (2) how to deal with the lack of data in minority class in comparison to the large amount of data in majority class. This subsection represents some ideas to address these two problems.

Manuscript received June 22, 2020; revised November 19, 2020.

doi: 10.12720/jait.12.1.29-35

A. *Appropriate Evaluation Metric*

The Area Under the Curve (AUC), which is derived from the Receiver Operating Characteristic (ROC) curve, is the commonly- used metric to evaluate the performance of the classifier with class imbalance. The AUC is used because it reflects the performance of the imbalanced data in churn prediction and the AUC does not depend on the predicted probability threshold between the two classes: churn and non-churn.

B. *Data Resampling and Cost-Sensitive Learning*

1) *Data resampling*

The data resampling methods include under-sampling and over sampling. Under-sampling eliminates a number of majority class examples while over-sampling duplicates minority class examples. Both of these sampling techniques reduce the overall level of class imbalance, thereby making the rare class less rare.

Some well-known methods in data resampling approach can be listed as follows. SMOTE (Synthetic Minority Oversampling Technique) proposed by Chawla *et al.* in 2002 [4] is an over-sampling method. To overcome the limitation of SMOTE, some other improved variants of SMOTE, such as Borderline-SMOTE (Han *et al.*, 2005 [21]), Safe-Level-SMOTE (Bunkhumporpat *et al.*, 2009 [22]), and ADYSIN (He *et al.*, 2008 [23]) were suggested. CUBE, proposed by Deville and Tille, in 2004 [24] is a popular method which is based on under-sampling.

2) *Cost-Sensitive learning*

In many data mining tasks, including churn prediction, it is the rare cases that are of primary interest. Evaluation function that does not take this into account often cannot perform well in these situations. One solving method is to use cost-sensitive learning methods. These methods can make use of the fact that the value of correctly identifying the rare class outweighs the value of correctly identifying the common class. For two-class problems this is done by assigning a greater cost to false negatives than with false positives.

Some well-known methods in cost-sensitive approach can be listed as follows. Chen *et al.* in 2004 [25] proposed a method which uses weighted random forests to classify imbalance data. Lin *et al.* in 2017 [20] proposed a cost-sensitive-learning method which employs Focal Loss, a kind of cross-entropy-loss to handle imbalance data. Harliman *et al.* in 2018 [26] proposed a Ripple-SMOTE method which employs both weighted loss function in deep neural network and oversampling synthetic data. Wang *et al.* in 2019 [19] proposed a method which employs both weighted (cross-entropy) loss and focal loss on the boosting machine to deal with imbalance data.

III. The Selected Comparative Methods

In this study, we will evaluate the performance of the two main approaches for handling imbalance data in churn prediction: data resampling and cost-sensitive learning. As for resampling approach, we investigate SMOTE and Deep Belief Network-based method. As for cost-sensitive learning, we examine two methods, one is based on focal loss [20] and the other is based on weighted cross-entropy loss [19].

A. *SMOTE*

SMOTE (Synthetic Minority Oversampling Technique) builds upon up-sampling the minority class [17].

SMOTE over-samples the minority class by generating synthetic minority examples in the neighborhood of observed ones. The idea is to form new minority examples by interpolating between examples of the same class. This has the effect of creating synthetic data around each minority observation. A simple example of SMOTE is shown in Fig. 1. An x_i minority class instance is selected as basis to create new synthetic data points. Based on a distance measure, several nearest neighbors of the same class (points x_{i1} to x_{i4}) are chosen from the training set. Then, a randomized interpolation is carried out to obtain new instances r_1 to r_4.

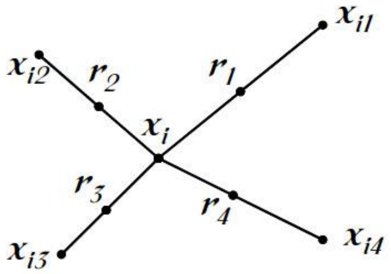

Figure 1. Illustration of how to create the synthetic data points by SMOTE [10].

The SMOTE method was implemented in scikit-learn framework with Python language.

B. *DBN-based Method*

Deep Belief Network (DBN) is an Unsupervised Probabilistic Deep Learning model ([18], [27]). A DBN is composed of multiple Restricted Boltzmann Machines (RBMs). These RBMs are stacked on top of each other, taking their inputs from the hidden layer of the previous RBM. RBMs follow the encoder-decoder paradigm. In this paradigm an encoder transform the input into a feature vector representation from which a decoder can reconstruct the original input. We use DBN as a generative model due to it ability to reconstruct the input data through the decoding-steps in RBMs ([18], [28]). In this work we apply DBN to generate synthetic data which belong to the churn class. To the best of our knowledge, this work is the first attempt to apply DBNs in data resampling for handling imbalanced data.

C. *Focal Loss*

Focal Loss, a cost-sensitive learning algorithm proposed by Lin *et al.* in 2017 [20], handles the imbalanced data by using the loss function to penalize more significant errors with minority class. In Focal Loss algorithm, the loss function is reshaped to down-weigh easy examples and thus focus training on hard negatives.

For convenience, let m denote the number of data samples, y_i denote the true label of the i-th sample, and $\hat{y}_i$

represent the probabilistic prediction for the i-th sample. And the focal loss, which is based on cross-entropy loss, is defined as follows:

$$L_f = -\sum_{i=1}^{m} y_i(1-\hat{y}_i)^{\gamma}\log(\hat{y}_i) + (1-y_i)\hat{y}_i^{\gamma}\log(1-\hat{y}_i)$$

In the above formula, a modulating factor $(1-\hat{y}_i)^{\gamma}$ is added to the cross entropy loss where γ is tested from the range [0, 5] in the experiment.

D. *Weighted Cross Entropy Loss*

Imbalance-XGBoost, proposed by Wang *et al.* in 2019 [19], is a cost-sensitive learning method which is adapted from XGBoost, a gradient tree boosting algorithm in order to handle the imbalanced data in binary classification. Imbalance-XGBoost emloys the loss function, called Weighted Cross Entropy Loss, which is defined as follows:

$$L_w = -\sum_{i=1}^{m}(\alpha y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i))$$

where α indicates the 'imbalance parameter'. Intuitively, if α is greater than 1, extra loss will be counted on 'classifying1 as 0'; on the other hand, if α is less than 1, the loss function will weight relatively more on whether data points with label 0 are correctly identified.

Focal Loss and Weighted Loss methods were implemented in Imbalance-XGBoost, a Python package that combines the XGBoost algorithm with weighted loss and focal loss to handling binary label-imbalanced classification tasks [19].

IV. Experimental Evaluation

A. *Experiment Scenario*

The evaluation experiment is based on the evaluation results on the test set, which is taken 30% from the original data set completely separated from the training set. AUC results are based on the best running result among several runs of the experiment. The optimal parameter through the grid search process was selected from an experiment with AUC metric to achieve the best results. All experiments are conducted on the most commonly-used datasets, and applying the same preprocessing process.

TABLE I. Characteristics of Telecom Datasets

Dataset	UCI	Cell2Cell
Source	UCI University	Duke Univerity
Feature	21	77
Samples	3333	51047
Missing Feature Values	No	Yes
Churn Class Samples	483	14711
Non-churn Class Samples	2850	36336

For customer churn prediction, the experiment is conducted on two datasets: UCI churn dataset and Cell2Cell dataset. The UCI churn dataset is from UCI Repository of Machine Learning Databases at the University of California, Irvine [29]. This churn dataset deals with cellular service provider's customers and the data pertinent to the calls they make. The Cell2Cell dataset is from Teradata Center for Customer Relationship Management of Duke University [30]. Cell2Cell is one of the largest wireless companies in the USA and its average monthly churn rate is 4%. Characteristics of these two telecommunication datasets are described in Table I.

The first purpose of the experiment is to compare the effectiveness of an imbalanced data processing approach (data sampling or cost-sensitive learning) to the classical approach that relies heavily on machine learning algorithms.

Second, we compare the effectiveness of a data resampling method to a cost sensitive learning method. The experiment also investigated the impact of parameters on the adjustment of the loss function to affect model performance.

B. *Experiment Results*

We applied Principle Component Analysis (PCA) projection on the two datasets UCI and Cell2Cell and the results are shown in Fig. 2 and Fig. 3. The two figures show that the data of the churn and non-churn groups on the Cell2Cell dataset are more ambiguous and interwoven than the UCI set. This observation implies that the churn prediction on Cell2Cell dataset is more difficult than on UCI dataset.

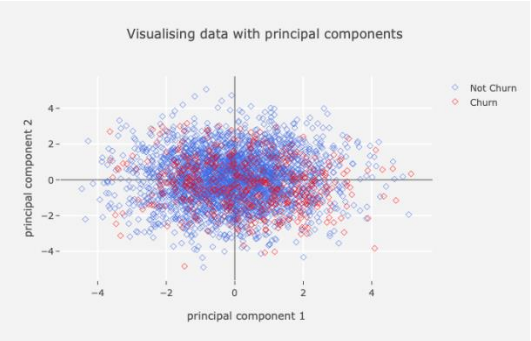

Figure 2. Visualising UCI dataset after applying PCA.

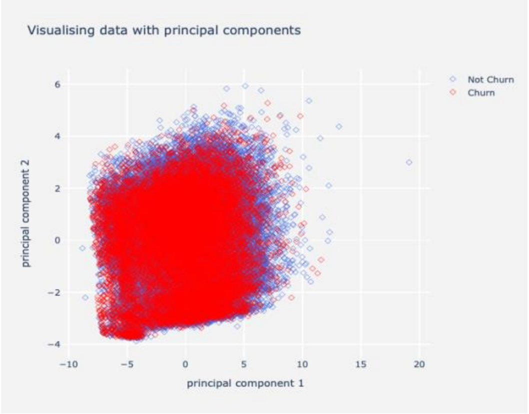

Figure 3. Visualising Cell2Cell dataset after applying PCA.

In this experiment, we will employ two classifiers based on different paradigms, namely logistic regression and XGBoost [31]. Logistic regression is a classical classifier and XGBoost is a gradient tree boosting algorithm (Chen and Guestrin, 2016 [31]). Then we develop a separate study for comparing data sampling approach with cost-sensitive learning approach in handling imbalanced data.

With respect to the evaluation metric, we will use the Area Under the ROC curve (AUC) as evaluation criteria.

We implemented logistic regression and XGBoost with *scikit-learn* framework. We implemented the DBN model with Tensorflow framework (using Python language).

In Table II and Table III, we show the average results for all comparative methods for handling imbalance data on the two datasets UCI and Cell2Cell respectively. In bold, we highlight the method that obtains the best performing average.

TABLE II. PERFORMANCE EVALUATION ON UCI DATASET

Method	Algorithms	AUC
No	Logistic Regression	0.5759
SMOTE	Logistic Regression	0.7568
DBN	Logistic Regression	0.6431
No	XGBoost	0.8455
SMOTE	XGBoost	0.8666
DBN	XGBoost	0.8714
Focal Loss	**XGBoost**	**0.8925**
Weighted Loss	**XGBoost**	**0.9115**
SMOTE + Focal Loss	XGBoost	0.8851
SMOTE + Weighted Loss	XGBoost	0.8703

TABLE III. PERFORMANCE EVALUATION ON CELL2CELL DATASET

Method	Algorithms	AUC
No	Logistic Regression	0.5248
SMOTE	Logistic Regression	0.5973
DBN	Logistic Regression	0.5247
No	XGBoost	0.5637
SMOTE	XGBoost	0.5676
DBN	XGBoost	0.5634
Focal Loss	**XGBoost**	**0.6618**
Weighted Loss	**XGBoost**	**0.6592**
SMOTE + Focal Loss	XGBoost	0.6542
SMOTE + Weighted Loss	XGBoost	0.6403

The results in Table II and Table III show that the classification performance when applying cost-sensitive learning methods with loss function for handling imbalanced data is better than the one when applying data resampling approach. In data resampling approach, SMOTE and DBN have equal performance. The results also reveal that the XGBoost outperforms the logistic regression in classification with imbalanced data. The combination of both data resampling and cost sensitive learning approaches does not bring out better results.

The experimental results on Cell2Cell dataset are quite similar to those on UCI dataset. The improvement in terms of AUC between the second approach (Focal Loss and Weighted Loss) and the first approach (SMOTE and DBN) on Cell2Cell dataset is higher the one on UCI dataset. The reason of this fact is that the Cell2Cell dataset has an interweaving structure between two classes, which makes it more difficult to separate them in comparison to the structure of UCI dataset, and therefore in this special case the loss function can bring out a better effectiveness in handling imbalanced data.

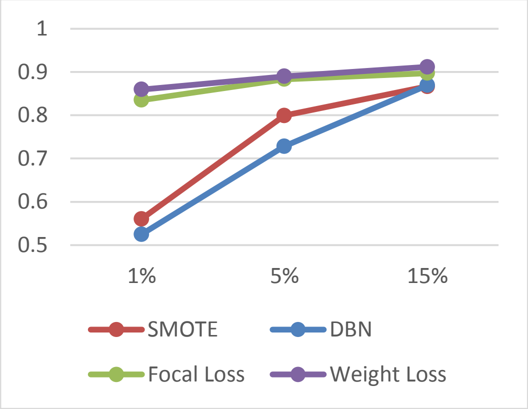

Figure 4. AUC to different churn rates on UCI dataset.

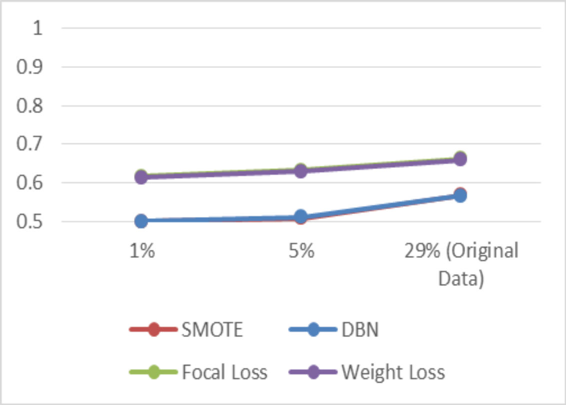

Figure 5. AUC to different churn rates on Cell2Cell dataset.

The variation of AUC to different rates of churn class on UCI dataset and Cell2Cell dataset are shown in Fig. 4 and Fig. 5, respectively. The curves in Fig. 4 and Fig. 5 show that the methods using the loss function have better performance than the methods using data resampling. Specially, both Focal Loss and Weighted Loss obtain the same results and have good performance even with churn rates in [1%, 5%]. This makes the methods Focal Loss and Weighted Loss capable of handling imbalanced data in practice with very low rate of churn class.

The variation of AUC to different values of γ in Focal Loss and three different churn rates on the UCI dataset is shown in Fig. 6. The variation of AUC to different values of α in Weighted Loss and three different churn rates on UCI dataset is shown in Fig. 7.

The variation of AUC to different values of γ in Focal Loss and three different churn rates on Cell2Cell dataset is shown in Fig. 8. The variation of AUC to different values of α in Weighted Loss and three different churn rates on Cell2Cell dataset are shown in Fig. 9.

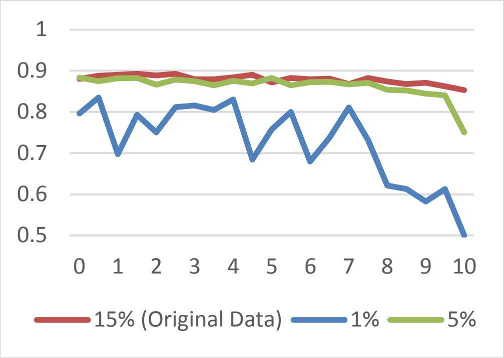

Figure 6. AUC to different γ values of focal loss and three different churn rates on UCI dataset.

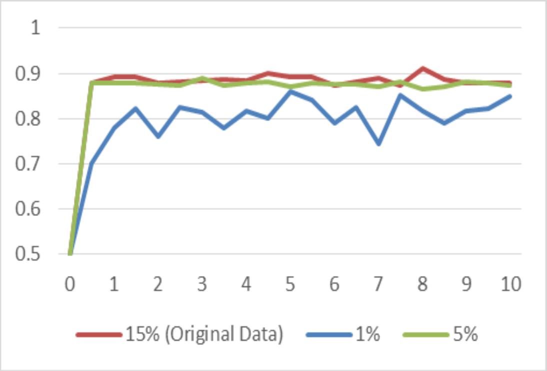

Figure 7. AUC to different α values of weighted loss and three different churn rates on UCI dataset.

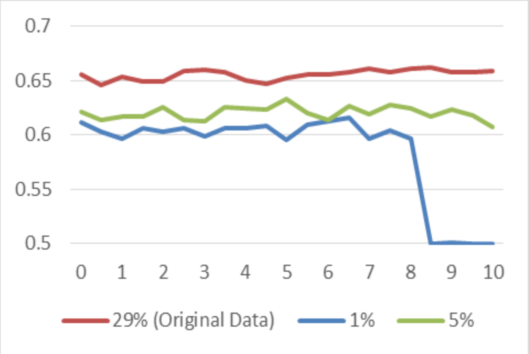

Figure 8. AUC to different γ values of focal loss and three different churn rates on Cell2Cell dataset.

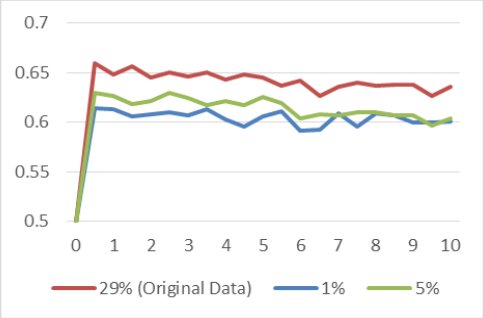

Figure 9. AUC to different α values of weighted loss and three different churn rates on Cell2Cell dataset.

From Fig. 6, Fig. 7, Fig. 8 and Fig. 9, we can see that the effect of the parameters γ and α in Focal Loss and Weighted Loss on the classification performance is very strong in the case churn rate is 1%. This observation implies that when applying Focal Loss and Weighted Loss techniques to a dataset with severe imbalance, we must pay much attention to tuning the penalty parameters for each group.

TABLE IV. TRAINING TIMES OF THE FOUR METHODS ON TWO DATASETS

Method	UCI	Cell2Cell
SMOTE	0.3960	46.3105
DBN	1.6687	98.4344
Focal Loss	0.9623	5.4266
Weighted Loss	0.1266	0.6092

The training times (in seconds) of the four methods on two datasets are reported in Table IV and Fig. 10. From Table IV and Fig. 10, we can see that the training time for Focal Loss or Weighted Loss method is remarkably lower than the other two methods (SMOTE and DBN) on Cell2Cell dataset with a large number of samples and attributes. This fact implies that the two methods Focal Loss or Weighted Loss have high practical applicability to the scenario of complex training dataset and large amount of data.

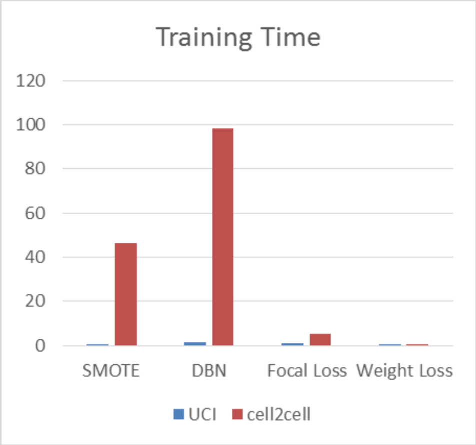

Figure 10. Training times (in seconds) of the four methods on two datasets.

V. Conclusion and Future Work

In this work, we compare the performance of two data resampling methods: SMOTE and DBN against two cost-sensitive learning methods: Focal Loss and Weighted Loss in churn prediction problem. The experimental results indicate that in this practical problem the method which concerns with handling the data imbalance proves to be more effective than the traditional one without handling the data imbalance. In both of the data resampling and cost-sensitive learning approaches, the latter approach, such as Focal Loss and Weighted Loss, is more effective than the data resampling approach, especially on the datasets with very small churn rates of the range [1%, 5%]. Fast training time of the two methods Focal Loss and Weighted Loss also imply their profound utility for practical application. As for comparing DBN with other methods, experimental results show that DBN's performance is comparable with SMOTE despite of longer training time and difficulties in tuning parameters.

In future, we intend to include the integration of focal loss and weighted loss method with some other efficient classification algorithms such as SVM, Random Forests [32], etc. To improve the generation of synthetic data, some data under-sampling techniques can be applied to select the appropriate samples for the training dataset.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

D. T. Anh proposed the main ideas of the research. N. N. Nam implemented the comparative methods and empirically evaluated the benefits of the proposed method.

References

[1] Y. Zhang, J. Qui, H. Shu, and J. Cao, "A hybrid KNN-LR classifier and its application in customer churn prediction," in *Proc. IEEE International Conference on Systems, Man and Cybernetics*, 7-12 Oct, 2017, pp. 3265-3269.

[2] P. Datta, B. Massand, D. Mani, and B. Li, "Automated cellular modeling and prediction on a large scale," *Artificial Intelligence Review*, vol. 14, pp. 485-502, 2000.

[3] Y. Wang and Z. Chen, "The application of classification algorithm combined with k-means in customer churning of telecom," *Journal of Jiamusi University (Natural Science Edition)*, vol. 28, no. 2, pp. 175-179, 2010.

[4] G. Li and X. Deng, "Customer churn prediction of China Telecom based on cluster analysis and decision tree algorithm," in *Communications in Computer and Information Science*, Springer, 2012, vol. 315, pp. 319-327.

[5] E. Lima, C. Mues, and B. Baesens, "Domain knowledge integration in data mining using decision tables: Case studies in churn prediction," *Journal of the Operational Research Society*, vol. 60, no. 8, pp. 1096-1106, 2009.

[6] S. Hung, D. Yen, and H. Wang, "Applying data mining to telecom churn management," *Expert Systems with Applications*, vol. 31, pp. 515-524, 2006.

[7] C. F. Tsai and Y. H. Lu, "Customer churn prediction by hybrid neural networks," *Expert Systems with Applications*, vol. 36, pp. 12547-12553, 2009.

[8] Y. Zhao, B. Li, X. Li, W. Liu, and S. Ren, "Customer churn prediction with improved one-class support vector machine," in *Lecture Notes in Computer Science*, Springer, 2005, vol. 3584, pp. 300-306.

[9] B. Lariviere and D. V. D. Poel, "Predicting customer retention and profitability by using random forests and regression forests techniques," *Expert Systems with Applications*, vol. 29, pp. 277-285, 2005.

[10] Y. Xie, X. Li, E. W. T. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," *Expert Systems with Applications*, vol. 36, pp. 5445-5449, 2009.

[11] J. Burez and D. V. D. Poel, "Handling class imbalance in customer churn prediction," *Expert Systems with Applications*, vol. 36, pp. 4626-4636, 2009.

[12] M. Galar, A. Fernandez, and F. Herreda, "A review on ensembles for class imbalance problem: Bagging, boosting and hybrid based approaches," *IEEE Transaction on Systems, Man and Cybernetics – Part C; Applications and Reviews*, vol. 42, no. 4, pp. 463-484, 2012.

[13] S. Wang and X. Yao, "Diversity analysis on imbalanced data sets by using ensemble models," in *Proc. IEEE Symposium on Computational Intelligence and Data Mining*, Nashville, TN, 2009, pp. 324-331.

[14] B. X. Wang and N. Japkowicz, "Boosting support vector machines for imbalanced data sets," *Knowl. Inf. Syst.*, vol. 25, no. 1, pp. 1-20, 2010.

[15] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress, challenges, marking the 15-year anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863-905, 2018.

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

[17] N. V. Chawla, K. W. Bowye, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, 321-357, 2002.

[18] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, 2006.

[19] C. Wang, C. Deng, and S. Wang, "Imbalance-XGBoost: Leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost," ArXiv:1908.01672v1 [cs.LG], 5 Aug., 2019.

[20] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object selection," in *Proc. IEEE Int. Conf. on Computer Vision and Applications*, 2018, pp. 1243-1248.

[21] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Proc. Internaional Conference on Advances in Intelligent Computing*, 2005, pp. 878-887.

[22] C. Bunkhumporpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level SMOTE: Safe level synthetic minority over-sampling technique for handling the class imbalanced problem," in *Proc. 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 2009, pp. 475-482.

[23] H. He, Y. Bai, E. A. Garcia, and S. Li. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1322-1328.

[24] J. C. Deville and Y. Tillle, "Efficient balanced sampling: The cube method," *Biometrika*, vol. 91, pp. 893-912, 2004.

[25] C. Chen, A. Liaw, and L. Breiman, "Using random forests to learn imbalance data," Technical Report 666, Statistics Department, University of California at Berkeley, 2004.

[26] R. Harliman and K. Uchida, "Data- and Algorithm-Hybrid approach for imbalanced data problems in deep neural network," *International Journal of Machine Learning and Computing*, vol. 8, no. 3, pp. 208-213, 2018.

[27] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504-507, 2006.

[28] A. R. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *Proc. NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, Vancouver, Canada, 2009, vol. 1, p. 39.

[29] C. L. Blake and C. J. Merz. (2019). Churn Data Set, UCI Repository of Machine Learning Databases, University of California, Department of Information and Computer Science,

Irvine, CA. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[30] CRM data in Teradata Center of Duke University. [Online]. Available: http://www.fuqua.duke.edu/centers/ccrm/index.html

[31] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proc. the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794, 2016.

[32] V. Effendy and Z. K. A. Baizal, “Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest,” in *Proc. 2nd International Conference on Information and Communication Technology*, 2014, pp. 325-330.

Nam N. Nguyen was born in Tien Giang, Vietnam in 1992. Now he is a graduate student in computer science Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology, where he also received B.Eng. in Computer Engineering in 2015. Mr. Nam’s research fields include machine learning and time series data mining.

Anh T. Duong was born in Quang Ngai, Vietnam in 1953. He received B.Eng. in Electronic Engineering from Ho Chi Minh City University of Technology University in 1976. He received his Master of Engineering and Doctorate of Engineering in Computer Science from the School of Advanced Technologies at Asian Institute of Technology, Bangkok, Thailand in 1989 and 1998, respectively. He is currently associate professor of computer science at Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology. His research is in fields of metaheuristics, machine learning and time series data mining. Dr. Anh is currently the Head of Time Series Data Mining Research.

Group in his faculty, he authored more than 100 scientific papers.