Handling imbalance data in churn prediction using combined SMOTE and RUS with bagging method

Eka Pura Hartati¹, Adiwijaya², Moch Arif Bijaksana³

School Of Computing, Telkom University, Indonesia

ekapura@telkom.co.id¹, adiwijaya@telkomuniversity.ac.id², arifbijaksana@telkomuniversity.ac.id³

Abstract. Customer churn has become a significant problem and also a challenge for Telecommunication company such as PT. Telkom Indonesia. It is necessary to evaluate whether the big problems of churn customer and the company’s managements will make appropriate strategies to minimize the churn and retaining the customer. Churn Customer data which categorized churn Atas Permintaan Sendiri (APS) in this Company is an imbalance data, and this issue is one of the challenging tasks in machine learning. This study will investigate how is handling class imbalance in churn prediction using combined Synthetic Minority Over-Sampling (SMOTE) and Random Under-Sampling (RUS) with Bagging method for a better churn prediction performance’s result. The dataset that used is Broadband Internet data which is collected from Telkom Regional 6 Kalimantan. The research firstly using data preprocessing to balance the imbalanced dataset and also to select features by sampling technique SMOTE and RUS, and then building churn prediction model using Bagging methods and C4.5.

1. Introduction

Customer churn has become a significant problem and also a challenge for Telecommunication company. It is necessary to evaluate whether the big problems of churn customer and the company’s managements will make appropriate strategies to minimize the churn from its own data and retaining the customer to save the revenue. Churn prediction is considered as one of data mining application that reflecting imbalanced problems. Based on data from PT. Telkom Indonesia, the average churn rate Atas Permintaan Sendiri (APS) for Internet broadband customer of PT. Telkom Indonesia Regional 6 Kalimantan was less than 1% per month. The imbalanced data-set problem occurs when one class, usually the one that refers to the concept of interest (positive or minority class), is underrepresented in the data-set and the number of negatives (majority) instances have outnumbers of positive class instances [1]. The objective of this research is to handle imbalanced data problems using combine sampling SMOTE and RUS, and measure the performance of churn prediction using a classifier C4.5 with bagging approach. The dataset used in this research is dataset from PT. Telkom Indonesia Regional 6 Kalimantan which has 1,08% churn data (imbanced dataset). A higher performance has been obtained by applying combine sampling technique SMOTE and RUS. The implementation of bagging approach in classification is able to improve the F-score of a single classifier (C4.5).
2. Literature Review

2.1. Related Work
Many researchers have attempted to address the imbalanced data. Because this learning task is quite challenging over the past 15 years. A lot of techniques have been developed for dealing with imbalanced data according to their categories approach. There were researchers who focus on algorithm level, called Internal approach, some others focus on the data level, called external approach and the last cost-sensitive learning falls between data and algorithm approach. Based on three approaches above, algorithm level tries to adapt the existing classifier to bias the learning toward the minority class [2]. On data level approach aims to rebalance the class distribution by resampling the data [1], [3]. These techniques were independent of the classifier used and usually more versatile. The last cost-sensitive level incorporates both data level transformation by adding cost to an instance and modify the algorithm so accepts the additional cost [4]. Ensembles based methods that used in this research was bagging. This technique consists in a combination of bagging and data level approach to preprocess the data before training each classifier.

2.2. Customer Churn Management
Churning Customers can be divided into two groups, Voluntary and Non-Voluntary churner (Mattison, 2005). Non-voluntary churners are easier to identify, They have had their service withdrawn by Company. This group includes Customers that churned for fraud and non-payment Customer. Voluntary churners are more difficult to identify because these churners occur when the customer makes a decision to terminate their service to a provider. Voluntary churners can be sub-divided into two groups, an incidental churner, and a deliberate churner. an incidental churner usually happens when there is a change in circumstances prevents the Customer from further requiring the provided service. Deliberate churner happens for reason of technology (there is a new technology), economics (customer's packet price), service quality, psychological factor and other reasons.

2.3. Methods and Classifier
2.3.1. Random Under Sampling, SMOTE and Bagging. Sampling is a popular methodology to handle the problem of class imbalance. There are three types of sampling methods, namely Random Under-Sampling (RUS), Random Over Sampling and Hybrid methods that combine both sampling methods. RUS method, is a nonheuristic method that aims to balance class distribution through the random elimination of majority class examples. The majority class instances are discarded at random until a more balanced distribution is reached. Its major drawback is that it can discard potentially useful data, which could be important for the induction process. Random Over Sampling methods, which create a superset of the original data-set by replicating minority class instances but it can also increase the likelihood of occurring the overfitting, since it makes exact copies of existing instances. SMOTE is an oversampling method which the main idea is to create new minority class examples by interpolating several minority class instances that lie together. Its create instances by randomly selecting one or more of the nearest neighbors (kNN) of a minority class instance and the generation of the new instance value from a random interpolation of both instances. Thus, the overfitting problem is avoided and causes the decision boundaries for the minority class to be spread further into the majority class space. Bagging is one of ensemble methods which manipulate the training data set by sampling with replacement. Each new training data set has the same size with the original data training. Bagging methods use diverse training samples (bags) to train independent base learner capable of handling imbalanced data in a parallel way. Bagging method is very suitable for unstable classifier [1], through their ability technique to do sampling with a replacement which can make lower variance of dataset.
2.3.2. Classifier This research used C4.5 Decision Tree as the weak learner, because C4.5 gives a good performance for binary classification [1]. C4.5 is an improvement of ID3 which it is a possibility to use continuous data, using unknown (missing) values, having the ability to use attributes with different weights, pruning the tree after being created, pessimistic prediction error, and also sub-tree raising [2]. Moreover, its classification model is easy to understand and have high precision but C4.5 is also an unstable classifier like regression tree, artificial neural network, and rule-base classifier. This classifier is very sensitive to the change of data training, if the data is changed so does the classifier while ensemble learning can effectively improve its stability and generalization performance [3].

3. Material and Proposed Method

3.1. Proposed Scheme
The combined sampling SMOTE with RUS was proposed in this research to balance the dataset and using bagging approach to classify the churn class. The design process of this research was as described below.

![Block Diagram of Research Design](image)

3.1.1. Data Preparation. After collecting data from database resource then those data were combined and selected the fitting attributes. These were seven attributes that were used in this process, namely CUST_OLD, CUST_CAT, CUST_BILL, PACKET_NAME, KWAD, TROUBLE_TIC, STAT_CHURN. The Attribute with a missing value and had redundant information was not used in this research and then all dataset transformed into numerical type for the next step.

3.1.2. Data Preprocessing. Data preprocessing was a process before data were processed. Sampling became one of standard approach for improving classification. The training data was altered to create more balanced class distribution before they were used in the learning stage. This research used
combined RUS and SMOTE as sampling methods. Therefore the dataset could be balanced by neither loosing information too much (i.e., under sampling too many majority class instances), nor suffering from over-fitting (i.e., oversampling too heavily).

3.1.3. Learning Stage. There were two processes in this stage, they were training and testing. Bagging was one of ensemble methods which manipulate the training dataset by sampling with replacement. Each new training dataset has the same size with the original data training. Bagging methods use diverse training samples (bags) to train independent base learner capable of handling imbalanced data in a parallel way. Bagging method is very suitable for an unstable classifier, through their ability technique to do sampling with a replacement which can make lower variance of dataset. It also used voting to choose the best prediction that was resulted from basic classifiers. In this stage, C4.5 was used for a basic classifier to predict the churn. The decision trees were built, using a set of a training dataset. At each node of the tree, C4.5 taken one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. After the churn model has built then it was tested using prepared data test by 10-fold cross-validations technique.

4. Experiments

4.1. Parameter Measurement

Precision is the number of True Positives (TP) divided by the number of True Positives (TP) and False Positive (FP). It is the number of positive prediction divided by the total number of positives class value predicted (1). Recall is the number of True Positives (TP) divided by the number of True Positives (TP) and the number of False Negatives (FN). It is the number of positive prediction divided by the number of positive class values in the test data (2).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{F-Score} = \frac{(2 \times \text{precision} \times \text{recall})}{(\text{precision} + \text{recall})} \quad (3)$$

$$\text{F-Score} = \frac{(2 \times TP)}{2 \times (TP + FP + FN)}$$

Precision and recall can be summarized into another metric known as the F-measure. It is also called the F Score. The F Score conveys the balance between the precision and the recall (3).

### Table 1. Confusion Matrix for Two-Class

| ACTUAL CLASS | PREDICTIVE CLASS |
|--------------|------------------|
| POSITIVE (CHURN) | POSITIVE (CHURN) | False Negative (FN) |
| NEGATIVE (NOT CHURN) | False Positive (FP) | True Negative (TN) |

4.2. Experiment’s Scenario

The data set that use in this reseach is data broadband Internet from PT. Telekomunikasi Indonesia Area Regional 6 Kalimantan (Kalsel, Kalbar, Kalteng, Balikpapan, Samarinda, and Tarakan) in October 2016.
There were three steps of Experiment's Scenario:
1) Scenario 1: the first scenario aims to show the performance of classification without using data sampling and only use single classifier C4.5. F-Score was measured to see the performance of churn prediction model.
2) Scenario 2: the second scenario shows the performance of classification using data preprocessing by sampling with RUS and SMOTE in the learning process. It was also using single classifier C4.5. The F-Score was measured with combined SMOTE (find the best N% SMOTE) and RUS (find the best N% RUS). Then the F-score performance was evaluated.
3) Scenario 3: the last scenario aims to show the performance of classification with data preprocessing combined SMOTE and RUS, and using bagging methods in classification. This scenario used C4.5 for its classifier and investigate how many bags (Nmodel) that showed the best performance.

4.3. Experiment’s Result and Discussion

The first scenario aims to show the performance of classification without using data sampling and only use single classifier C4.5. F-Score was measured to see the performance of churn prediction model with F-score.

### Table 2. F-Score Value of Scenario 1

| Measurements        | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|---------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| TESTING (F-Score)   | 0.011| 0.016| 0.019| 0.014| 0.015| 0.020| 0.019| 0.014| 0.017| 0.017|
| TIME (sc)           | 14.710| 13.539| 14.257| 13.922| 13.809| 13.611| 13.437| 14.201| 13.454| 13.434|

There were not any input parameters at this first scenario because it just wants to show the real result of C4.5 as a single classifier to build a churn model (base line) and to measure the F-Score value. The average F-Score of Training in this scenario was 0.053 and Testing was 0.013. The result of this scenario still had a small value of F-score. The small value of F-Score because the dataset that used was still imbalanced and in the next scenario, the dataset will be rebalanced by using sampling process.

The second scenario showed the performance of the classification using data preprocessing by sampling with SMOTE and RUS in the learning process. It also used single classifier C4.5. The F-Score was measured with combined SMOTE (find the best N% SMOTE) and RUS (find the best N% RUS). The amount of minor data were 1.502 and the variable N SMOTE used in this experiments were N = 5x, N = 10x and N = 20x, thus, the minority class was generated as many as 5x (minor data will become 7.510),...
10x (minor data will become 15.020) and 20x (minor data will become 30.040) in the experiment. By SMOTE processing, the nearest neighbors (kNN) with k = 3 was used to avoid overfitting problem and caused the decision boundaries for the minority class to spread further into the majority class space. The N RUS variables that follow this experiments were N = ½ and N = ¾, the amount of major data were 136.404. It then reduced the major data by N= ½ means the major data became 68.202. The F-score performance was evaluated. The results were shown in the tables below:

**Table 3. F-Score Value of Scenario 2A, N RUS=1/2, N SMOTE=20x**

| Measurements    | Number of Experiments | AVERAGE |
|-----------------|------------------------|---------|
| F-SCORE TESTING | 0.0444, 0.0452, 0.0450, 0.0443, 0.0448 | 0.045   |
| CH RATE TESTING (%) | 0.5921%, 0.5792%, 0.5860%, 0.5921%, 0.5869% |         |
| TIME (sc)       | 343,499, 344,739, 342,649, 343,485, 341,581 | 343,629 |

Table 4-3 shows there is an increase of F-Score value, with average value is 0,045 with N SMOTE = 20x and N RUS = ½.

**Table 4. F-Score Value of Scenario 2A, N RUS=1/2, N SMOTE=10x**

| Measurements    | Number of Experiments | AVERAGE |
|-----------------|------------------------|---------|
| F-SCORE TESTING | 0.081, 0.081, 0.082, 0.080, 0.085 | 0.081   |
| CH RATE TESTING (%) | 0.8140%, 0.8173%, 0.8121%, 0.8143%, 0.8129% |         |
| TIME (sc)       | 340,342, 342,646, 327,057, 343,395, 341,700 | 336,682 |

In the table 4-4, the N SMOTE = 10x and N RUS = ½, the F-Score value gets a higher value with average 0,081.

**Table 5. F-Score Value of Scenario 2A, N RUS=1/2, N SMOTE=5x**

| Measurements    | Number of Experiments | AVERAGE |
|-----------------|------------------------|---------|
| F-SCORE TESTING | 0.087, 0.090, 0.091, 0.086, 0.093 | 0.089   |
| CH RATE TESTING (%) | 0.8339%, 0.8300%, 0.8268%, 0.8316%, 0.8267% |         |
| TIME (sc)       | 338,688, 349,440, 337,294, 338,936, 339,061 | 341,807 |

Table 4-5 shows the F-Score value still increase with average 0,089 and comparing to the tables above, there is a significant improvement of F-Score especially with N SMOTE = 5x and N RUS = ½. The average value of baseline F-Score was 0,013 and had increased to 0,089. Since the implementation of SMOTE and RUS, the churn rate value also increased the churn rate. It indicated that there was a decreasing number of majority data in order to balance the dataset. As the resume, the values of F-Score Testing of N RUS = ½ were presented by the graphic below.
From variable $N = \frac{1}{2}$, the research continued to check another value of $N_{\text{RUS}} = \frac{3}{4}$. The major data reduced to 102.303 and $N_{\text{SMOTE}} = 20x, 10x$ and $5x$.

**Table 6. F-Score Value of Scenario 2B, N_{\text{RUS}} = 3/4, N_{\text{SMOTE}} = 20x**

| Measurements         | Number of Experiments | AVERAGE |
|----------------------|-----------------------|---------|
|                      | 1         | 2        | 3       | 4       | 5       |         |
| F-SCORE TESTING      | 0.066     | 0.068    | 0.069   | 0.072   | 0.071   | 0.068   |
| CHURN RATE TESTING   | 0.8132%   | 0.8123%  | 0.8118% | 0.8098% | 0.8126% |
| TIME (sc)            | 375,355   | 374,932  | 375,345 | 375,965 | 375,888 | 375,211 |

Table 6 shows there is an increase of F-Score value, with average value is 0.068 with $N_{\text{SMOTE}} = 20x$ and $N_{\text{RUS}} = \frac{3}{4}$.

**Table 7. F-Score Value of Scenario 2B, N_{\text{RUS}} = 3/4, N_{\text{SMOTE}} = 10x**

| Measurements         | Number of Experiments | AVERAGE |
|----------------------|-----------------------|---------|
|                      | 1         | 2        | 3       | 4       | 5       |         |
| F-SCORE TESTING      | 0.089     | 0.086    | 0.087   | 0.092   | 0.088   | 0.087   |
| CHURN RATE TESTING   | 0.8264%   | 0.8279%  | 0.8284% | 0.8237% | 0.8272% |
| TIME (sc)            | 387,647   | 375,395  | 376,106 | 379,773 | 372,092 | 379,716 |

In the table 7, the $N_{\text{SMOTE}} = 10x$ and $N_{\text{RUS}} = \frac{3}{4}$, the F-Score value gets a higher value with average 0.087.
Table 8. F-Score Value of Scenario 2B, N RUS=3/4, N SMOTE=5x

| Measurements          | Number of Experiments | Average |
|-----------------------|-----------------------|---------|
| F-SCORE TESTING      | 0.093 0.089 0.089 0.094 0.090 0.090 |
| CHURN RATE TESTING   | 0.8302% 0.8332% 0.8324% 0.8302% 0.8382% |
| TIME (sc)             | 372,364 371,603 371,809 372,068 372,518 371,925 |

Table 8 shows the F-Score value still increase with average 0.090 and comparing to the three tables above, there is a significant improvement of F-Score specially with N SMOTE = 5x and N RUS = 3/4. SMOTE created instances by interpolating minor data in 5x ratio and RUS decreased the major data by N= ¾, these best parameters would be used in the next scenario 3 to obtain the better F-Score value with bagging approach. The F-Score’s tests of scenario 2B were figured in the graphic below as the resume from the tables above.

![F-Score Testing](image)

**Figure 4.** F-Score Testing of N RUS = 3/4

The combination of SMOTE and RUS had been able to reduce the probability of overfitting problem and rebalance the data distribution of majority class without loosing its substantial data. The best result from those parameters will be used in the next scenario 3 to aims the better F-Score value with bagging approach.

The testing’s result with bagging compared to without bagging and with a number of bags = 5 or 7. It was shown the better F-Score resulted from bagging process with numbers of bags 7 than 5. The best average F-Score value was 0.140 as explained in the table below.

Table 9. F-Score Value of Scenario 3, N RUS=3/4, N SMOTE=5x, Bags=5

| Measurements          | Number of Experiments | Average |
|-----------------------|-----------------------|---------|
| F-SCORE TESTING      | 0.159 0.138 0.118 0.126 0.126 0.138 |
| CH RATE TESTING (%)   | 0.8308% 0.8308% 0.8315% 0.8351% 0.8336% |
| TIME (sc)             | 1.875,62 1.878,40 1.872,86 1.727,26 1.882,81 1.875,63 |
Table 9 shows there is an increase of F-Score value, with average value is 0.138 with N SMOTE = 5x, N RUS = ¾ and bags = 5.

**Table 10.** F-Score Value of Scenario 3, N RUS=3/4, N SMOTE=5x, Bags=7

| Measurements          | Number of Experiments | 1    | 2    | 3    | 4    | 5    | AVERAGE |
|----------------------|-----------------------|------|------|------|------|------|---------|
| F-SCORE TESTING      |                       | 0.178| 0.146| 0.132| 0.141| 0.136| 0.152   |
| CH RATE TESTING (%)  |                       | 0.8328%| 0.8329%| 0.8351%| 0.8303%| 0.8322%|         |
| TIME (sc)             |                       | 2.640,42| 2.627,27| 2.645,73| 2.628,35| 2.629,72| 2.637,80 |

In the table 4-10, the N SMOTE = 5x, N RUS = ¾ and bags = 7, the F-Score value gets a higher value with average 0.152.

![Figure 5. F-Score Testing of Scenario 3](image)

The graphic as described in figure 4 conveys that the more bags used in the research the better F-Score we had since the new dataset was formed to train each single classifier of C4.5 by randomly drawing (with replacement) instances from the original dataset and the best prediction from each bag then to be chosen to infer the class.

### 5. Conclusions & Future Work

**5.1. Conclusions**

Based on the result this research, the performance of F-score is quite good, but it still needs more research to obtain the increase of F-score. The combined SMOTE and RUS were able to handle the imbalanced data and improved the churn prediction performance. SMOTE was used to generate the
synthetic data from the churn class in order to increase the probability of drawing the churn data (churn is the minority class). RUS was used to reduce the probability of overfitting problem caused by the oversampling in SMOTE. This combined method improved the F-Score to 571% from dataset without sampling. The Average F-Score Value was 0.090. The implementation of sampling and bagging methods in prediction process by C4.5 improved the performance of F-Score. The F-Score increased about 56% from without bagging. The average F-Score was 0.152, this combined method was very suitable for unstable classifier like C4.5, through their ability technique to do sampling with replacement to reduce the variance of dataset. The number of bags also influenced the improvement of F-Score since there were more option values for the voting process. The best F-Score resulted from a number of bags = 7.

5.2. Future Work

What most attributes that can influence customer is also needed to be investigated in attempt to identify why they churn and also an additional comparison for other ensemble methods such as boosting is needed to be evaluated

References

[1] N. V. Chawla, N. Japkowicz and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," ACM Sigkdd Explorations Newsletter, pp. 1-6, 2004.

[2] G. M. Weiss, "Mining with rarity: a unifying framework," ACM SIGKDD Explorations Newsletter, p. 7–19, 2004.

[3] H. Haibo, M. Yuqian and others, "Imbalanced Datasets: from Sampling to Classifiers," Imbalanced Learning, Foundation, Algorithms, and Application, pp. 52-69, 2013.

[4] E. Dwiyanti, Adiwijaya and A. Ardiyanti, "Handling Imbalanced Data in Churn Prediction Using RUSBoost and Feature Selection," International Conference Soft Computing and Data Mining, pp. 376-385, 2016.

[5] Tan, P. Ning, M. Steinbach and others, "Introduction on Datamining," Boston, Pearson Addison Wesley, 2006.

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

[7] J. Tang, S. Alelyani and H. Liu, "Feature selection for classification: A review," Data Classification: Algorithms and Applications, p. 37, 2014.

[8] Q. R.J, "Bagging, Boosting, and C4.5," in in AAAI/IAAI : Proceeding of the 13th National Conference on Artificial Intelligence and 8th Innovative Applications of Artificial Intelligence Conference, Portland, Oregon, 1996.

[9] H. Haibo, M. Yuqian and others, Imbalanced Learning, Foundations, Algorithms, and Applications, Wiley, 2013, pp. 22-51.

[10] L. S. Salzberg, "C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993," Machine Learning, vol. 16, pp. 235-240, 1994.