Journal of Soft Computing Exploration

Improved Accuracy of Naive Bayes Classifier for Determination of Customer Churn Uses SMOTE and Genetic Algorithms

Afifah Ratna Safitri[1], Much Aziz Muslim[2]
[1,2]Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received Aug 1, 2020
Revised Aug 14, 2020
Accepted Sept 5, 2020

Keywords:

Naive Bayes;
Customer Churn;
SMOTE;
Genetic Algorithms;

ABSTRACT

With increasing competition in the business world, many companies use data mining techniques to determine the level of customer loyalty. The customer data used in this study is the german credit dataset obtained from UCI. Such data have an imbalance problem of class because the amount of data in the loyal class is more than in the churn class. In addition, there are some irrelevant attributes for customer classification, so attributes selection is needed to get more accurate classification results. One classification algorithm is naive bayes. Naive Bayes has been used as an effective classification for years because it is easy to build and give an independent attribute into its structure. The purpose of this study is to improve the accuracy of the Naive Bayes for customer classification. SMOTE and genetic algorithm do for improving the accuracy. The SMOTE is used to handle class imbalance problems, while the genetic algorithm is used for attributes selection. Accuracy using the Naive Bayes is 47.10%, while the mean accuracy results obtained from the Naive Bayes with the application of the SMOTE is 78.15% and the accuracy obtained from the Naive Bayes with the application of the SMOTE and genetic algorithm is 78.46%.

Corresponding Author:

Afifah Ratna Safitri
Computer Science Departement
Faculty of Mathematich and Natural Sciences, Universitas Negeri Semarang,
Email: afifahratna25@students.unnes.ac.id

1. INTRODUCTION

The rapid development of technology, information systems, and science has resulted in increasingly tight competition in the business world. In the business world, customers are the main asset. Therefore, various ways have been taken by the company so that customers do not stop subscribing. The term that is often used for customers who stop subscriptions with one service provider and become a customer of another service provider is called customer churn [1]. Customer churn occurs because of customer dissatisfaction [2]. This happened in various industries including insurance, banking, and the telecommunications industry [3]. To prevent this from happening, one of the models used by the company is Customer Relationship Management (CRM) [4].

Journal of Soft Computing Exploration, Vol. 6, No. 1, May 2019 2 The concept of Customer Relationship Management (CRM) leads to the importance of maintaining customers and building long-term relationships with customers to keep customers from moving to the company's competitors [5]. The transfer of customers from one provider to another is due to better rates or services, or because of the different benefits offered by the company's competitors when registering [6].

With the increasing competition and diversity of offerings in the industrial market, many companies utilize data mining techniques to determine customer churn rates [6]. Data mining is an activity to find interesting patterns from a large number of data [7]. Data mining has been applied to many fields because of its ability to analyze large amounts of data and fast time [8]. Data mining has several techniques such as estimation, classification, association, and clustering [9]. Companies need customer classifications to determine the level of customer loyalty. Classification is the most important part in data mining [10]. Classification is a data mining technique that serves to predict classes in a data [11]. One classification algorithm is Naive Bayes. Naive Bayes has been used as an effective classification for years. Because Naive Bayes is easy to build and can handle a number of independent variables randomly, either continuously or categorically [12].

In the field of machine learning and data mining the classification of unbalanced data is a problem that often occurs. Data imbalances have a negative impact on classification results where minority classes are often incorrectly classified as the majority class [11]. The problem of class imbalance is a problem where data experiences significant differences between classes, where loyal classes are greater than the churn class. The problem of class imbalance can be overcome by using the Synthetic Minority Over Sampling Technique (SMOTE) method. The SMOTE method is often used to overcome class imbalance problems because the SMOTE method does not reduce the amount of data, so that no information is lost [13].

Classification on high dimensional data will reduce accuracy. High dimensional data is data that has a large number of attributes. To improve classification accuracy on high-dimensional data can be used attribute selection methods that function to understand the relevant attributes [14]. One algorithm that can be used for attribute selection is a genetic algorithm. Genetic algorithms are chosen because they can reduce attributes in high dimensional data. So that data that initially has many attributes is reduced to a few fewer attributes, without reducing information from the data [15]. The concept of genetic algorithms is to search for solutions based on the evolutionary process [16].

This study uses the German credit dataset. The dataset used in this study was taken from the UCI Machine Learning Repository. The purpose of this study is to improve the accuracy of the Naive Bayes algorithm by applying the SMOTE algorithm and attribute selection of Genetic Algorithms in classifying customers by seeing an increase in accuracy before and after the application of SMOTE and Genetic Algorithms.

2. METHOD

2.1 Dataset

The data used in this study are German Credit Data taken from the UCI Machine Learning Repository. The German Credit Data Collection has 20 attributes and 1000 instances. This dataset has 13 nominal type attributes and 7 numeric type attributes. This dataset has 1 class attribute of nominal type consisting of good and bad or loyal and churn. The description of the attributes of the German credit dataset can be seen in Table 1.

Table 1. German Credit Datasets Attributes

No	Attributes	Description	Attribute Type
1	*Status of existing checking account*	Status of current accounts / deposits held by debtors	Nominal
2	*Duration in month*	Credit duration in months	Numeric
3	*Credit history*	Credit history ever owned	Nominal
4	*Purpose*	The purpose of applying for credit	Nominal
5	*Credit Amount*	Amount of money credited	Numeric
6	*Saving account/bonds*	Savings account owned	Nominal
7	*Present employment since*	The length of time the debtor works	Nominal
8	*Installment rate in percentage of disposable income*	The installment rate in the percentage of disposable usage	Numeric
9	*Personal status and sex*	Personal status and gender	Nominal
10	*Other debtors / guarantors*	Other debtors / guarantor	Nominal
11	*Present residence since*	The length of stay in residence	Numeric

12	*Property*	Ownership property	Nominal
13	*Age in years*	Age in years	Numeric
14	*Other installment plans*	Other installment plans	Nominal
15	*Housing*	Status of residence inhabited	Nominal
16	*Number of existing credits at this bank*	The amount of credit in this bank	Numeric
17	*Job*	Job	Nominal
18	*Number of people being liable to provide maintenance for*	The number of people responsible for providing maintenance	Numeric
19	*Telephone*	Telephone ownership	Nominal
20	*Foreign worker*	Status of foreign workers	Nominal
21	*Class*	*Class*	Nominal

2.2 Experiment

In this study several algorithms were used to obtain a model to improve the accuracy of the Naive Bayes algorithm by using SMOTE and Genetic Algorithms. The experimental stages carried out in this study can be seen in Figure 1.

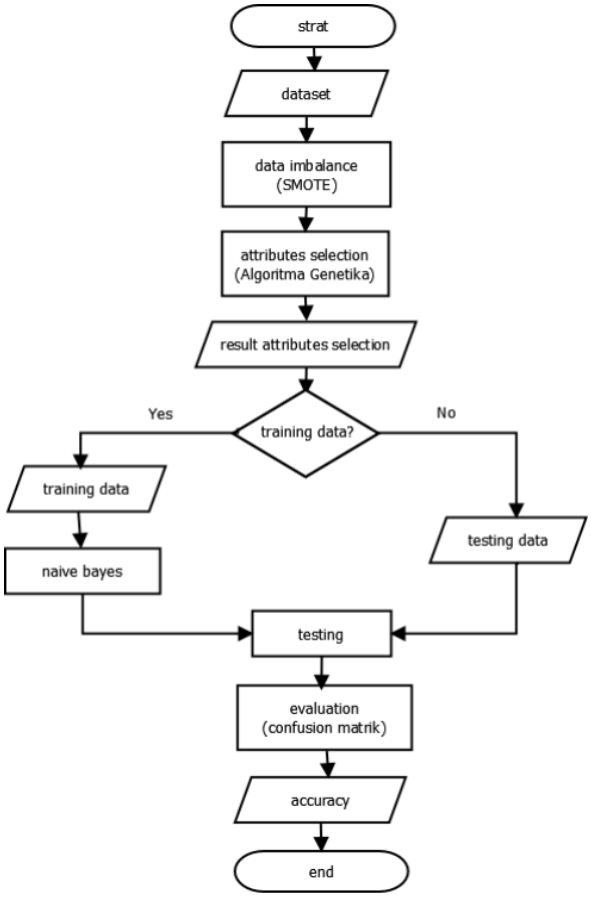

Figure 1. Experimental Stages of the Naive Bayes Method by Applying SMOTE and Genetic Algorithms

As seen in the picture above, the method used in this study is the application of the SMOTE algorithm and Genetic Algorithm to the Naive Bayes Classifier. The dataset that has been done by class balancing and attributes selection is then divided into training data and testing data for classification using the Naive Bayes Classifier. Data evaluation is done using cofussion matrix to calculate classification accuracy. The stages of each method can be seen as follows:

2.2.1 SMOTE

SMOTE is a technique used to expand minority sample data areas. This technique is made by making synthetic data for minority classes. Making synthetic data for minority classes in more detail can be seen as follows:

1. Enter the dataset and the amount of additional data that will be created. In this system, new minority class datasets generated as many as 300 new data.
2. Selecting minority class data, where in this dataset the minority class data is churn class data.
3. Separating minority data (churn) and majority class data (loyal). After the minority class data and the majority are separate and then remove the majority (loyal) class data.
4. Randomly select a minority dataset (churn) and calculate the selected k-nearest neighbor data. The k value used to calculate the k-nearest neighbor here is 3.
5. After that make new data based on randomly selected data and k- nearest neighbor by multiplying the distance that has been obtained in the fourth step with numbers chosen randomly between 0 and 1, then add the value of the original vector feature.
6. Repeat step 2 until the amount of new data corresponds to the number of additions to the desired data, where in this dataset the desired amount of new data is 300 data churn. 7. After all stages have been completed, 300 new minority data will be known so that there are 1300 sample data.

2.2.2 Genetic Algorithms

The stages of Genetic Algorithms can be seen as follows:

```
Awaken teh initial population of chromosome N.
Loop until the stop condition is fulfilled
    Loop for N chromosome N
        Individual = Decode (chromosome)
        Fitness    = Evaluation (individual)
    End
    Make one or two of the best chromosome copies
    Loop until you get a new N chromosome
        Select two chromosome as parents P1 and P2
        [parent1, parent2] = Recombination (P1, P2)
        [child1, child2]   = Mutation (child1, child2)
    End
    Change the old N chromosome with the new N chromosome.
End
```

2.2.3 Naïve Bayes Classifier

The stages of the Naive Bayes algorithm in classifying datasets are as follows:

1. Read training data.
2. Calculating probability in the following way:
   a. Calculates the average of each parameter with the following formula:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

Information:
- μ : mean
- *xi* : sample value i
- *n* : number of samples

   b. Calculates the standard deviation of each parameter with the following formula:

$$\sigma^2 = \frac{1}{n-1}\sum (x_i - \mu)^2 \tag{2}$$

Information:
- σ : Standard deviation, expresses the variance of all attributes
- n : Amount of data in the same class
- x_i : Value of attribute to i
- μ : mean

c. Look for probability values using the formula:

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{\frac{(x_i-\mu)^2}{2\sigma^2}} \quad (3)$$

Information:

σ : Standard deviation, expresses the variance of all attributes
x_i : Value of attribute to i
μ : mean
X_i : attribute to i
Y : Class sought
y_j : Y syb-class searched

3. Repeat step 2 until the probability of all parameters is calculated.
4. The calculation process will stop when the probability value of all parameters of each attribute has been calculated.

3. RESULT AND DISCUSSION

3.1 Results

This study uses a system created with the PHP programming language that is applied to German Credit datasets. Accuracy results obtained on the application of Naive Bayes without the pre-processing process which is equal to 73%. Whereas, the results of the average accuracy of ten executions obtained using SMOTE and the attibutes selection of Genetic Algorithms on Naive Bayes is 80.948%.

3.2 Discussion

Based on the results of the application of the SMOTE algorithm and the attributes selection of Genetic Algorithm in the Naive Bayes algorithm that has been carried out, it can be seen that the accuracy for determining customer churn using the German Credit dataset is taken from the UCI Machine Learning Repository. Data previously obtained has passed the pre-processing stage, namely the class balancing stage and attibutes selection stage.

At the stage of class balancing is done by applying the SMOTE algorithm. The SMOTE algorithm is applied to make new data more balanced. German Credit's initial dataset has 1000 samples with 700 loyal (good) classes and 300 churn (bad) classes. Therefore it is necessary to balance the class by creating new data in the churn class. The new dataset of the SMOTE algorithm results in 300 churn class data, so there are 1300 new sample data. This is done so that data can be classified optimally. The attribute selection stage is done by selecting attributes in the data used. In this attribute selection stage there is a dimension reduction in the data in order to optimize attributes that will affect the accuracy of the Naive Bayes algorithm. Dimension reduction in attributes is done by using Genetic Algorithms. Removal of attributes is done one by one from attributes that have the smallest fitness value and will be mining. The process of selecting attributes and mining will stop when the results of the accuracy have exceeded the specified minimum limit.

After going through the pre-processing stage, new data will go through the classification process using the Naive Bayes algorithm. From the results obtained, there is an increase in the accuracy of the Naive Bayes algorithm and the Naive Bayes algorithm by applying the SMOTE algorithm and attibutes selection of Genetic Algorithms.

4. CONCLUSION

In this study, testing the Naive Bayes algorithm by applying the SMOTE algorithm and attribute selection of Genetic Algorithms is done using the German Credit dataset taken from the UCI Machine Learning Repository to classify churn and loyal customers. Accuracy results obtained on the application of the Naive Bayes algorithm without the pre-processing process that is equal to 73%. Meanwhile, the average accuracy of ten executions obtained using the SMOTE algorithm in the Naive Bayes algorithm is 74.918% and the results of the average accuracy of ten executions obtained using the SMOTE algorithm and the attributes selection of the Genetic Algorithm of the Naive Bayes algorithm is 80.948%.

REFERENCES

[1] V. Mahajan, R. Misra, R. Mahajan. "Review on factors affecting customer churn in telecom sector", International Journal of Data Analysis Techniques and Strategies, 9(2), pp. 122-144, 2017.

[2] A. A. Q. Ahmed, D. Maheswari. "Churn prediction on huge telecom data using hybrid firefly based classification", Egyptian Informatics Journal, 18(3), pp. 215-220, 2017.

[3] R. Hejazinia, M. Kazemi. "Prioritizing Factors influencing customer churn", Interdisciplinary Journal of Contemporary Research in Business, 5(12), pp. 227-236, 2014.

[4] P. K. Banda, S. Tembo. "Application of System Dynamics to Mobile Telecommunication Customer Churn Management", Journal of Telecommunication, Electronic and Computer Engineering, 9(3), pp. 67-76, 2017.

[5] H. S. Soliman. "Customer Relationship Management and Its Relationship to the Marketing Performance", International Journal of Business and Social Science, 2(10), pp. 166-182, 2011.

[6] I. Brandusoiu, G. Toderean. "Churn prediction in the telecommunications sector using support vector machines", Annals of the Oradea University, 22(1), pp. 19-22, 2013.

[7] M. A. Muslim, A. J. Herowati, E. Sugiharti, B. Prasetiyo. "Application of the pessimistic pruning to increase the accuracy of C4.5 algorithm in diagnosing chronic kidney disease", Journal of Physics: Conf. Series, 983, pp. 1-9, 2017.

[8] M. A. Muslim, S. H. Rukmana, E. Sugiharti, B. Prasetiyo, S. Alimah "Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis", Journal of Physics: Conf. Series, 983, pp.1-7, 2017.

[9] P. Sinha, P. Sinha. "Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM", International Journal of Engineering Research & Technology (IJERT), 4(12), pp. 608-612, 2015.

[10] Makhtar, S. Nafis, M. A. Mohamed, M. K. Awang, M. N. A. Rahman, M. M. Deris. "Churn Classification Model for Local Telecommunication Company Based on Rough Set Theory", Journal of Fundamental and Applied Sciences, 9(6S), pp. 854-868, 2017.

[11] H. Lee, J. Kim, S. Kim. "Gaussian-Based SMOTE Algorithm for Solving Skewed Class Distributions", International Journal of Fuzzy Logic and Intelligent Systems, 17(4), pp. 229-234, 2017.

[12] M. H. A. Elhebir, A. Abraham. "A Novel Ensemble Approach to Enhance the Performance of Web Server Logs Classification", International Journal of Computer Information Systems and Industrial Management Applications, 7, pp. 189-195, 2015.

[13] M. Anis, M. Ali. "Investigating the Performance of Smote for Class Imbalanced Learning: A Case Study of Credit Scoring Datasets", European Scientific Journal, 13(33), pp. 341-353, 2017.

[14] L.Marlina, M. A. Muslim, A. P. U. Siahaan, "Data Mining Classification Comparison (Naive Bayes and C4.5 Algorithms)", International Journal of Engineering Trends and Technology (IJETT), 38(7), pp. 382-383, 2016.

[15] C. Kirui, L. Hong, W. Cheruiyot, H. Kirui. "Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining", International Journal of Computer Science Issues, 10(1), pp. 165-172, 2013.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, 16, pp. 321-357, 2002.