Clustering Based Undersampling for Handling Class Imbalance in C4.5 Classification Algorithm

Wahyu Nugraha¹, Muhammad Sony Maulana, Agung Sasongko

¹Department of Engineering and Informatics, University Bina Sarana Informatika, Jakarta, Indonesia
E-mail: wahyu.whn@bsi.ac.id, muhammad.sony.mom@bsi.ac.id, agung.ako@bsi.ac.id

Abstract. Machine Learning is very difficult to make an effective learning model if the distribution of classes in the training data set that is used is not balanced. The problem of class imbalance is mostly found during classifications in the real world where one class is very small in number (minority class) while the other classes are very numerous (majority in class). Building a learning algorithm model without considering the problem of class imbalance causes the learning model to be flooded by majority class instances so that it ignores minority class predictions. Random undersampling and oversampling techniques have been widely used in various studies to overcome class imbalances. In this study using the undersampling strategy with clustering techniques while the classification model uses C4.5. Clustering is used to group data and the undersampling process is performed on each data group. The goal is that sample samples that are useful are not eliminated. Statistical test results from experiments using 10 imbalance datasets from KEEL-repository dan Kaggle dataset with various sample sizes indicate that clustering-based undersampling produces satisfactory performance. Improved performance can be seen from the sensitivity and AUC values that increased significantly.

1. Introduction

In data mining and machine learning, it is very difficult to create an effective learning model if the class distribution in the training data set used is not balanced. This imbalance problem is known as a class imbalance problem [1]. Reports from academics and industry show unequal distribution data from the data collection and assessment of the increase in data allocation in accordance with the agreed class distribution [2]. The problem of class imbalance is often found in real-world classification problems such as fault diagnosis, anomaly detection, medical diagnosis, and email folding [3]. Based on these problems, the class that is considered more important (primary class of interest) in data mining is the minority class (positive class) or the rare class [4][1].

The standard learning algorithm is designed to generalize training data so that it does not pay attention to rare problems on dates that do not meet. Therefore, to predict larger classes is more difficult than predicted classes [2]. Build a learning model without considering the problem of class imbalance that causes the learning model to be flooded by instances Class imbalance has a serious impact on improving performance. For example, data sets with a class number of 99% and a minority class of only 1% will increase. However, in this case all minority class samples will be predicted wrongly when classified [4][5].
Class imbalances are very different in high-risk application domains such as software defect predictions or quality estimates, medical data sets, fraud detection, and risk management. In this domain the misclassification of positive class instances can be approved for higher costs compared to errors in negative classes [6]. For example in bankruptcy predictions, the number of bankruptcy cases (ie minority classes) is much smaller than non-bankruptcy cases (ie acquisition classes) [1]. The prediction model error in classifying bankruptcy cases to non-bankruptcy classes is more important than the average classification level. Because of this error, increasing the number of bad loans for financial institutions.

The method for overcoming class imbalance is divided into four discussed, namely: algorithmic level, cost-sensitive, data level, and ensemble of classifiers [7]. The data level method is recommended for preprocessing data before the classification model is built and the pre-existing classification model is not approved. The level of the data level is much taken into consideration for research in various literatures [1]. Because, the preprocessing technique is done separately or independently with the algorithm level (algorithmic level) and looks like an advantage. However, it will usually be very difficult to determine the optimal resampling ratio automatically [7].

The data preprocessing method is categorized as an external method, known as the data balancing method [8]. This method is based on the process of resampling the data training conducted in class imbalance and conducted before the classification model training. To make a balanced training set can be done by using oversampling techniques in the minority class and undersampling in the majority class [9]. Several other studies using a combination of oversampling and undersampling methods in preprocessing data and combining them also use the classifier ensemble method such as boosting and bagging techniques [1].

Data preprocessing with a sampling strategy is used to overcome class imbalances by eliminating some data from the agreed class (undersampling) or adding some data using the results of the resulting process or duplicate data to the right class (oversampling) [10]. Of these two resampling strategies, undersampling has proven to be a better choice than oversampling. This is because oversampling can increase the likelihood of overfitting the data compilation process of the training model [1]. However, the undersampling strategy can cause some useful data (useful data) for the learning process of the model to be eliminated [4]. To overcome the disadvantages of undersampling, in this study a sampling method is used which is based on undersampling for random undersampling methods. Grouping is used to help classify the approved class instances into data groups that have supporting characters [8]. Then in each cluster random data samples are taken and the rest will be eliminated to reduce the majority class.

2. Related Work

Research on class imbalance (class imbalance problem) has long been conducted and many research results have been published. This aims to find out the state of the art about research that addresses the problem of class imbalance, especially in research that uses the undersampling random data level approach.

Research by Lin [1] class imbalance is a problem in various data in the world. Imbalance can cause machine learning to be dominated by majority classes and neglect minority classes. Strategies to overcome class imbalances such as undersampling can cause some useful data for the model learning process to be eliminated. So this study applies the K-means algorithm as a clustering-based undersampling technique to overcome class imbalance with two approach strategies, namely cluster center and NN (nearest neighbors) from the cluster center. The experimental results show that clustering-based undersampling using the NN approach strategy results in better performance. The experiment used 44 small-scale datasets with an imbalance ratio of 1.8-129 and two large-scale datasets from kdd-cup.
Research by Devi [11] the classification model tends to be biased towards the majority class, which results in errors in the prediction of minority classes. Existing resampling techniques such as undersampling can eliminate the example of majority classes to a level that can provide equal majority with minority classes. However, the random elimination of majority class cases increases the loss of useful information while endangering the classification process. So the data level approach is used to deal with the problem of eliminating majority cases effectively without losing valuable information. The Nearest Neighbor technique and the Tomers-undersampling link are used to improves the undersampling algorithm. The dataset uses 10 real life datasets taken from the University of California, Irvine (UCI) machine learning repository. The results obtained are satisfactory because the proposed model provides an efficient undersampling mechanism to eliminate the majority of cases with less information loss and higher performance.

The problem of class imbalance when observed from several previous studies shows that the data level approach is the most important approach in dealing with class imbalance. There are two approaches, namely undersampling and oversampling. This study tries an undersampling technique with a clustering-based undersampling approach using the CLARA clustering algorithm (clustering large application) while the classification algorithm in this study uses the C4.5 model.

3. The Proposed Method
The undersampling based clustering model proposed in this study uses a cluster technique that aims to group objects that have similar characters into the same cluster. In the cluster method the first stage is to determine the best K cluster value. The Silhouette value method is used to determine the optimal value of the K cluster. The cluster or medoid center of these cluster groups will be used as a representation of all data that will be used as a sample of the majority of classes. Samples in each cluster group will be taken randomly and eliminate the remaining data (undersampling). Furthermore, the majority class sample will be combined with the minority class sample to form a new balanced training dataset. Following are the steps of the proposed undersampling based clustering algorithm is given below [12].

| Table 1. The steps of the proposed undersampling based clustering algorithm. |
|---------------------------------------------------------------|
| **INPUT Training Set**                                        |
| divide positive and negative class from training set into variable; |
| determine number of K cluster using silhouette value;          |
| clustering negative class using CLARA algorithm;              |
| **Implementation CLARA Algorithm**                            |
| *Input:* A set $S$ of $d$-dimensional patterns, the number of clusters $K$, number of samples $T$ |
| *Output:* $K$ mediods                                        |
| **begin**                                                     |
| for $i = 1$ to $T$ repeat the following                       |
| **begin**                                                     |
| draw a sample of size $40 + 2K$ randomly from the entire data set; |
| execute the PAM algorithm over this sampled subset to find mediods; |
| assign each object in the original data set to its nearest mediod; |
| calculate the average distance between data items and their respective mediods; |
| save the best result obtained so far;                        |
| **end**                                                       |
| from each cluster, pick random sample;                        |
| merge/combine sample from all cluster into single training set; |
| add negative and positive class;                             |
| **end**                                                       |
| **OUTPUT** New Training Set with balanced distribution of class |


The classification method in this study uses the C4.5 classification algorithm, while for the performance evaluation using confusion matrix measurements and AUC values. The design of the proposed algorithm show in Figure 1. This study uses a public dataset which is data that was previously made by someone whether published or not [13]. The dataset used in this study are data related to class imbalance problems. The selected dataset varies and has a imbalance ratio of 2.5 to 129. 10 dataset used comes from KEEL-dataset repositories (abalone9-18, abalone19, segment0, suttle0vs4, yeast1, yeast3, yeast4, yeast5, yeast6) available from https://sci2s.ugr.es/keel/imbalanced.php and Kaggle dataset repository (creditcardfraudrisk) from https://www.kaggle.com/mlg-ulb/creditcardfraud. The dataset used in this study needs to be carried out several data pre-processing techniques in order to get good quality data. Some data preparation techniques are performed such as data validation to overcome the problem of Incompleteness or data that experiences missing value and data transformation [14]. The software used in processing data using R language with IDE using Rstudio.

4. Results and Discussion
This section will discuss the results obtained from experiments or testing models. Measurements will not focus on Accuracy values, because if measuring the performance of unbalanced data using only Accuracy, then this can be a misleading indicator. Because, Accuracy only places general class weights instead of rare classes or minorities [15]. This research focuses on detecting minority classes or positive classes so that sensitivity and specificity can be used to show the performance of two classes. The cutoff of sensitivity and specificity can be used to make ROC curves [16].

Figure 1. Clustering-based undersampling procedure.
The results of tests C4.5 with undersampling when compared to the C4.5 without undersampling indicated a considerable significant increase in the sensitivity and AUC, but the accuracy and specificity values decreased significantly, shown in figure 2 and 3. This shows that the model built from the unbalanced dataset causes the model to focus on the prediction of the majority class and ignore the minority class so that the value of accuracy and specificity increases but sensitivity and AUC decreased significantly. Learning algorithm model that is built without considering the problem of class imbalance causes the learning model to be flooded by majority class instances, thus ignoring minority class [5]. From Figure 4 dan 5 has shown that clustering based undersampling+C4.5 model shows better performance based on the sensitivity and AUC.

The results of tests the proposed model when compared with C4.5+undersampling indicated an overall improvement. The values of sensitivity and AUC increased significantly, the accuracy and specificity values also increased but not significantly. Undersampling+C4.5 model is already good in predicting the minority class when compared to C4.5 without undersampling. However, this model sacrifices useful majority class that cause specificity and accuracy decreases significantly. Undersampling strategy can cause some useful data for the learning process of the model to be eliminated [4].

From Figure 4 dan 5 has shown that clustering based undersampling+C4.5 model shows better performance based on the sensitivity and AUC values that increase significantly when compared to another model. The accuracy and specificity values that increase are quite good compared to the undersampling model, indicating that the process of eliminating the majority class sample in the proposed model is quite successful in keeping some useful data from being eliminated. This means that clustering based undersampling is quite successful in overcoming the weaknesses of random undersampling to eliminate majority classes in unbalanced datasets.
5. Conclusions and Future Work
Most of the research related to class imbalance uses data level approach techniques, namely random undersampling and oversampling. In this study the undersampling technique with clustering-based undersampling approach is introduced. The purpose of this study is to develop an undersampling-based clustering technique to overcome class imbalances so as to improve the performance of the C4.5 classification model. Seeing the results of the comparison of the three models can be concluded that the classification model that was built using the clustering-based undersampling technique using the CLARA algorithm is quite effective to overcome the class imbalance, especially against the weaknesses of the random undersampling (RUS) technique. This can be seen from the increased value of sensitivity, specificity and AUC. From the results of this study it can be concluded that the clustering based undersampling technique can improve the performance of the C4.5 classification for prediction of minority classes in unbalanced datasets.

Most likely the results of this study can encourage further research in the future, related to the results that show both positive and negative values. One way to improve the performance of this model might be to increase the effectiveness of CLARA. Because the effectiveness of CLARA depends on the sampling technique and sample size, it is difficult to determine the best sample size [17]. So as to improve clustering performance, other K-medoid-based techniques can be used such as CLARANS Algorithm or CLATIN: An Efficient k-Medoids Clustering Method.

References
[1] Lin W C Tsai C F Hu Y H and Jhang J S, 2017 Clustering-based undersampling in class-imbalanced data Inf. Sci. (Ny). 409–410 p. 17–26.
[2] Sun Y Kamel M S Wong A K C and Wang Y, 2007 Cost-sensitive boosting for classification of imbalanced data Pattern Recognit. 40, 12 p. 3358–3378.
[3] Galar M Fernandez A Barrenechea E Bustince H and Herrera F, 2012 A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 42, 4 p. 463–484.
[4] Sun Z Song Q Zhu X Sun H Xu B and Zhou Y, 2015 A novel ensemble method for classifying imbalanced data Pattern Recognit. 48, 5 p. 1623–1637.
[5] Liu X Y Wu J and Zhou Z H, 2009 Exploratory under-sampling for class-imbalance learning Proc. - IEEE Int. Conf. Data Mining, ICDM p. 965–969.
[6] Sowah R A Agebure M A Mills G A Koumadi K M and Fiawoo S Y, 2016 New Cluster Undersampling Technique for Class Imbalance Learning Int. J. Mach. Learn. Comput. 6, 3 p. 205–214.
[7] Beyan C and Fisher R, 2015 Classifying imbalanced data sets using similarity based hierarchical decomposition Pattern Recognit. 48, 5 p. 1653–1672.
[8] Rayhan F Ahmed S Mahbub A Jani R Shatabda S and Farid D M, 2017 CUSBoost: Cluster-Based Under-Sampling with Boosting for Imbalanced Classification in 2nd International Conference on Computational Systems and Information Technology for Sustainable Solutions, CSITSS 2017 September p. 1–6.
[9] Chawla N Bowyer K Hall L and Kegelmeyer W, 2002 SMOTE: synthetic minority over-sampling technique J. Artif. Intell. Res 16 p. 321–357.
[10] Rahman M M and Davis D N, 2013 Cluster Based Under-Sampling for Unbalanced Cardiovascular Data Proc. World Congr. Eng. 2013 3 p. 1–6.
[11] Devi D Biswas S kr and Purkayasta B, 2016 Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance Pattern Recognit. Lett. 93 p. 1339–1351.
[12] Pakhira M K, 2008 Fast Image Segmentation Using Modified CLARA Algorithm in 2008 International Conference on Information Technology p. 14–18.
[13] Kothari C, 2004 Research methodology: methods and techniques 2nd ed. New Delhi: New Age International (P) Limited.
[14] Vercellis C, 2009 Business Intelligence: Data Mining and Optimization for Decision Making John Wiley & Sons, Ltd.
[15] Bekkar M Djennaa H K and Alitouche T A, 2013 Evaluation measures for models assessment over imbalanced data sets J. Inf. Eng. Appl. 3, 10 p. 27–38.
[16] Kuhn M, 2008 Building Predictive Models in R Using the caret Package J. Stat. Softw. 28, 5 p. 1–26.
[17] Zhang Q and Couloigner J, 2005 A New and Efficient K-Medoid Algorithm for Spatial Clustering in ICCSA 2005: Computational Science and Its Applications p. 181–189.