BAGGING BASED ENSEMBLE ANALYSIS IN HANDLING UNBALANCED DATA ON CLASSIFICATION MODELING

Hartiny Pop Koapaha
Faculty of Economics and Business, Universitas Klabat
hartinikoapaha@unklab.ac.id

Niel Ananto
Faculty of Economics and Business, Universitas Klabat
niel@unklab.ac.id

Abstract

The purpose of this study is to Identify the algorithm of each method of handling the unbalanced class based on bagging based on the literature review. This study uses a bagging based ensemble method such as UnderBagging, OverBagging, UnderOverBagging, SMOTEBagging, Roughly Balanced Bagging and the last one is the Bagging Ensemble Variation. The data used is coded from the UCI Repository with 16 data, eight of which have class categories with low imbalance problems, and the rest are categorized as high imbalance problems. The number of classes used in this study amounted to two classes. The class with a small number is made into the minority class and the rest is made up as the majority class. The result of this research is the bagging based method gives better results when compared to classical methods such as the classification tree.

Keywords:* *Bagging, boosting, classification, class imbalance, ensemble

INTRODUCTION

In recent years, the class imbalance problem has emerged as one of the challenges in the data mining community. Unbalanced classes, known as imbalanced datasets, are a hot topic in the data mining environment. The class imbalance problem is a new problem that arises during machine learning. This imbalance occurs because there is a larger number of samples than the other examples, a large number of samples is called the majority class, while the small number of samples is called the minority class. Ramyachitra (2014). For example, in a data set consisting of two classes, the ratio of the number of samples in that class is 1: 100, 1: 1000, and/or 1: 10,000 Elrahman A. A. SM (2013). The impact resulting from this imbalance results in the classification being not optimal because classes with a higher number of samples have a very large effect on classification. Chawla K. N.V et.al. (2004). Several cases regarding the problem of imbalance are sometimes very important. For example, detecting fraud in banking operations, detecting network disruptions Galar H.F. M. et.al. (2012). Managing risks, and predicting failure of technical equipment (Longadge, 2013).

This study focuses on handling unbalanced classrooms using data levels and ensemble learning. The ensemble is a method that combines several single classifications to obtain a more accurate classification model Permatasari (2016). The working principle of the ensemble method is to produce many classification trees from a data set and then make an assumption based on the merging results of the allegations from each tree Schouts R (2015). The Ensemble method is designed to improve the accuracy of the classifier single by training several different classifiers. Then the prediction results of each classification are combined into the final prediction through the voting process Rodrigo (2010). The well-known ensemble methods are the bagging and boosting methods. Bagging and boosting methods have been successful in increasing the accuracy of the classification proce. The following Figure 1. Shows the development of research using the bagging method, accessed through.

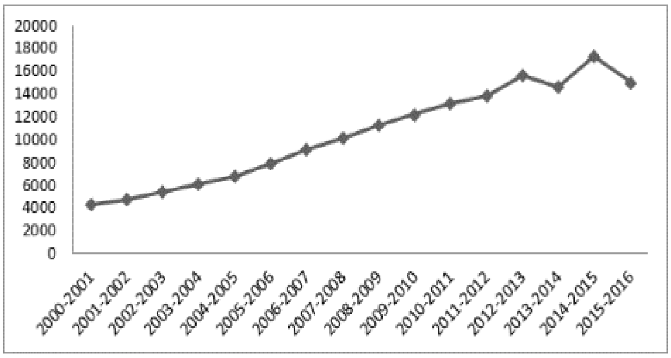

Figure 1
The development of articles using the bagging method (2000-2016)
Source: Bauer, 1999

In Figure 1, the development of bagging illustrates that on average each year scientific papers using the bagging method always increase. This shows the high interest of someone who applies the bagging method in various cases. Bagging is an ensemble method Breiman (1996), which is a combination of bootstrapping and aggregating. Meanwhile, the Boosting method (AdaBoost) is an ensemble learning method that can reduce variance, this happens because of the bias effect of the average ensemble to reduce variance from a set of classifications. Bisri (2015). In the literature in this study, there are several bagging-based methods used such as UnderBagging, OverBagging, SMOTE Bagging, Under OverBagging, Bagging Ensemble Variation, and Roughly Balanced Bagging. It is important to know these bagging-based approaches by comparing several scientific papers that discuss the ensemble method, especially those based on bagging to find the advantages and disadvantages of each method. And can help researchers in choosing a method that fits the data existing cases.

METHODS

This study discusses classification methods for handling unbalanced classes based on bagging, including OverBagging, SMOTEBagging, UnderBagging (UB), OverBagging (OB) UnderOverBagging (UOB), Bagging Ensemble Variation (BEV), and Roughly Balanced Bagging (RBB). The stages are carried out, namely:

Data Collection

The data used in this study were obtained from the UCI Repository consisting of 16 data. Eight of them have a low-grade ratio and the rest have a high-class ratio. A high-class ratio has an imbalance level of the majority class with a percentage of more than or equal to 90% and the rest as a minority class. Meanwhile, the class category with low imbalance has a majority class above 65% and below 90%, then the other classes are made into a minority class. The data used in this study consisted of 2 classes, namely the small class which was used as a minority class or positive class, while a large number of classes was used as a minority class or negative class. Some of the data used in this study have more than 2 classes, but the class is divided into 2 classes. The class that is made up as a minority class or positive class and the others are lumped together into a majority or negative class. The following Table 2.1 information about the data used.

Table 1
Information on data names, amount of data, number of variables in data, comparison of data classes and data class categories.

Name of Data	Amount of Data	Number of Variables	Class Comparison	Imbalance Category
Bank	41188	21	89:11	Low
Shuttle	43500	10	78:22	Low
Adult	30162	15	75:25	Low
Liver	583	11	71:29	Low
Credit	30000	24	77:23	Low
Transfussiun	748	5	76:24	Low
Pima	768	9	65:35	Low
White Wine	178	14	27:73	Low
Santimage	4435	37	91:9	High
Red Wine	1599	12	95:5	High
thiroyd	2030	29	92:8	High
Ann thiroyd	3772	22	92:8	High
Letter-A	20000	17	94:6	High
Car	1792	5	93:7	High
Glass	214	11	96:4	High
Htru	17898	9	91:9	High

Data Analysis

After the data is collected, analysis is carried out using R software with the help of several packages such as DMwR, rpart, ROSE, and caret. Then after that, examine the advantages and disadvantages of each of the bagging-based unbalanced class handling methods based on the accuracy, sensitivity, and specificity values.

Sampling Technique

The sampling approach is a technique used without having to change the algorithm. This technique is generally used to deal with unbalanced class problems. This technique changes the distribution of data or changes the size of the data from unbalanced to balanced. Yusof et.al. (2017). The resampling process is carried out at the pre-processing stage, before the modelling process. The sampling technique is divided into two, namely under sampling and oversampling (Sun, et.al., 2015).

Under sampling is removing samples of the majority class randomly to balance the data, the drawback of this technique is the loss of information in the data . Figure 2 shows an example of an under sampling technique. Meanwhile, oversampling is adding/replicating minority class samples randomly so that a balanced amount of data is obtained. The drawback of this technique is that there will be overfitting due to the large amount of data that will be generated. Figure 3 shows an example of the oversampling technique.

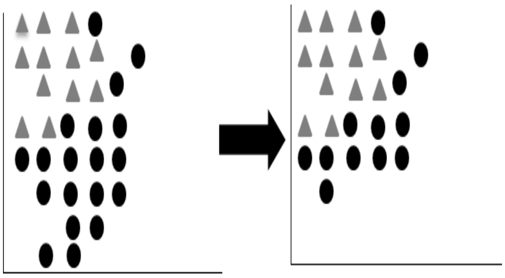

Figure 2
Random erasing majority class

Source: Chawla, 2004

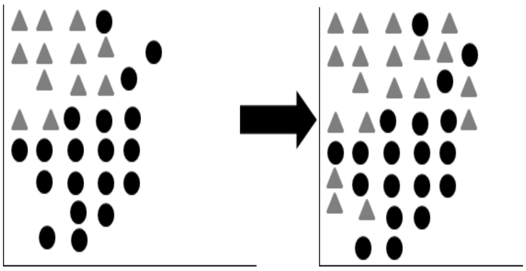

Figure 3
Increasing the number of classes minority

Source: Chawla, 2004

Unbalanced Class

Class imbalance is an imbalance that occurs when one of the two classes has a smaller number of samples than the other. The class that has the most examples is called the majority class, while the class that has few examples is called the minority class. For example, in a data set consisting of two classes, the ratio of the number of samples in that class is 1: 100, 1: 1000, and/or 1: 10,000. Elrahman (2013). The impact resulting from this imbalance results in the classification being not optimal because classes with a higher number of samples have a very large effect on classification. (Chawla, 2004).

In an unbalanced class, if classification is carried out, it tends to result in errors in the classification processor will cause bias in the classification process. This is because the minority samples tend to be

underrepresented. After all, their numbers are too small compared to the number of the majority class. Thus, several approaches are needed in dealing with this imbalance. Below are 3 approaches to dealing with class imbalances.

1. Algorithm level approach, this approach is carried out by utilizing existing algorithms to make classifications to consider minority classes. (Zhang, 2011).
2. The data level approach, this approach is done by modifying the number of samples so that the data is balanced. (Lopez et.al., 2013)
3. The cost-sensitive approach, this approach is a combination of the two previous approaches, namely the algorithm level approach and the data level approach to minimize the total cost of misclassification. (Zhu, 2017).

Some scientific papers that discuss class imbalance include Research conducted by (Barro et.al. 2013), discussing the unbalanced class in making herbal composition models by applying the Synthetic Minority Oversampling Technique (SMOTE) method. The data used are data on the status of plant use in the composition of herbs for certain properties. This data consists of 1002 types of herbal medicine in Indonesia that are registered with the Food and Drug Supervisory Agency. The comparison of the data used is 22: 980. The study compared two methods, namely the method without using the SMOTE technique in the data balancing process and the method using SMOTE in the data balancing process. The two models that have been obtained are compared with the level of accuracy with the AUC value or area under the ROC curve and the goodness of the model with Nagelkerke's R2. Nagelkerke's R2 value in the SMOTE model was 3.2% lower than that of the model without SMOTE. However, the AUC value for the model with SMOTE was 0.68% higher than the AUC value for the model without SMOTE. This shows that the model with SMOTE is more accurate than the model without SMOTE.

Bagging

Bagging is an acronym for bootstrap aggregating which was first. Based on its name, the bagging method consists of two main stages in analysis, namely bootstrapping which is nothing but sampling from the sample data that is owned (resampling) and aggregating which is combining many the estimated value becomes one estimated value. Sartono (2010). The algorithm of this method is taking samples of the original data n times with a return to creating training data Efron (2003). Then for each training data, a classification tree is made and the process of aggregating or selecting the most votes is carried out for classification cases and the average for regression cases. The use of bagging is very helpful in dealing with class imbalances. Bagging can reduce the misclassification rate in classification cases by repeating 50 times for classification cases and 25 times for regression cases. (Breiman, 1996).

UnderBagging (UB)

The UnderBagging method is a combined method of sampling under sampling and bagging which was first introduced by. The algorithm of UnderBagging is to build several training data b times containing all minority classes and randomly drawn the majority class with the same number as the minority class without or with returns. Barandela (2003). Then perform a combined prediction of the B classification that has been obtained by using the most votes rule. In general, the basic classification used in the UnderBagging method is the Classification Tree Method. However, several researchers developed this method using various classification methods of SVM, Logistic Regression, KNN, Random Forest, and others. Some related studies that explain UnderBagging include:

Research conducted by Permatasari (2016) discusses the handling of unbalanced classes using the RUSboost and UnderBagging techniques which are proven to provide better results compared to the classical classification tree. The RUSBoost and UnderBagging methods are more sensitive to unbalanced classes than using the classical classification tree.

OverBagging (OB)

The OverBagging method is a combined method of oversampling and bagging sampling techniques. This method was first introduced by Wang (2009). The algorithm of OverBagging is to build several training data b times containing all majority classes and randomly drawn the same number of minority classes in the majority class without or with returns.

Research conducted by Ralescu (2016), describes several classical classification methods and classification methods used for handling unbalanced classes. Classical classification methods are Decision Trees, Multilayers Perceptron (MLP), Bagging, and AdaBoostM1. While the classification methods used for handling unbalanced data are: C4.5 for imbalanced data, Multilayer Perceptron with Back Propagation Training Cost-Sensitive (NNCN), SMOTEBagging, SMOTEBoost, MSMOTEBagging, MSMOTEboosting, UnderBagging, and OverBagging. From this determination, the results show that the data mining technique developed for unbalanced data has much better.

SMOTEBagging (SBAG)

SMOTEBagging is a combination of SMOTE and bagging algorithms that involves the process of generating artificial data while constructing data clusters Wang (2009). Synthetic Minority Oversampling Technique (SMOTE) combined with bagging is one of the oversampling methods first introduced by Chawla (2004). SMOTE is an oversampling method that works by generating artificial data. The generated artificial data is made based on the characteristics of the object and k-nearest neighbor. The number of k-nearest neighbors is determined by considering the ease of implementation. In SMOTEBagging, each training data obtained comes from the bootstrapping process, balancing the data classes using SMOTE before modeling. The SMOTEBagging algorithm is different from the UnderBagging and OverBagging algorithms. The SMOTEBagging algorithm builds some training data that contains examples of all classes by resampling the original examples. with a return on the rate $\left(\frac{\boldsymbol{N_c}}{\boldsymbol{N_i}}\right)\boldsymbol{b}\%$ by using SMOTE (k, N) Where Ni is the number of examples of training data for class i, Nc is the number of examples of major class training data, b% is the value to control the number of new data generation (range from 10 to 100).

UnderOverBagging (UOB)

UnderOverBagging is an algorithm combination of undersampling, oversampling, and bagging, but the data generation process is not like the UnderBagging or OverBagging algorithms. The process of generating data in UnderOverBagging is similar to SMOTEBagging (Wang, 2009). The algorithm of the UnderOverBagging method generates several training data b times. Each sample in the training data was determined based on resampling (a%) which was arranged in each iteration ranging from (10% to 100%) Galar (2012). Then perform a combined prediction of the B classification that has been obtained by using the most votes rule.

Research conducted by Japkowicz (2014), discusses how to solve unbalanced class problems in binary classification. The amount of data used is 10 with unbalanced classes. Also, the data used will be checked with the first three conditions of pre-processing. Second, using 5 different oversampling techniques: ADAptive SYNthetic Sampling (ADASYN), Random over-sampling (ROS), Synthetic Minority Over-sampling Technique (SMOTE), Synthetic Minority Over-sampling Technique + Tomek's modification of Condensed Nearest Neighbor (SMOTE_TL) and Selective Preprocessing of Imbalanced Data 2 (SPIDER2). The third one uses 5 undersampling techniques: Neighborhood Cleaning Rule (CNNTL), Class Purity Maximization (CPM), Neighborhood Cleaning Rule (NCL), Undersampling Based on Clustering (SBC), and Tomek's modification of Condensed Nearest Neighbor (TL). From this research, it is found that the performance of the pre-processing model is influenced by class imbalance. Also, the Neighborhood Cleaning Rule (NCL) method has better performance on data with low imbalance levels. Then the SMOTE and SMOTE TL methods provide better performance on data with a high level of imbalance.

Bagging Ensembles Variation (BEV)

Bagging Ensemble Variation is part of the UnderBagging member, which is a combination of undersampling and bagging sampling. Bagging Ensemble Variation was first introduced by Wang (2009). The basic idea of this method is to maximize minority class data without making artificial data or making changes to the classification system. The algorithm of the Bagging Ensemble Variation method is to build several training data b times. Each training data contains all the number of minority classes, while the majority of classes differ from one training data to another. Then perform a combined prediction of the B classification that has been obtained by using the most votes rule. The following is related research using the Bagging Ensemble Variation method, including: Research conducted by Freund (2007), in his research which discusses the Bagging Ensemble Variation method applied to the railroad industry. The data used consisted of 30,686 wheels, of which 833 were labeled failed, and the rest were labeled safe. The method used in this research is 12 methods, 11 of which are single classification methods and 1 ensemble classification, namely Bagging Ensemble Variation. From this research, it was found that the overall level of accuracy, specificity, and sensitivity of the ensemble method was better than the 11 other single classification methods. This means that the Bagging Ensemble Variation method is quite effective in using unbalanced classes with various variations because overall it can improve the accuracy of the data.

Roughly Balanced Bagging (RBB)

Roughly Balanced Bagging is a new technique for handling unbalanced classes. Also, Roughly Balanced Bagging is part of the UnderBagging technique first introduced by Hido et.a.l., (2009). This method is very effective in balancing the average of each class. In contrast to other bagging techniques, Roughly Balanced Bagging involves a negative binomial distribution to balance classes on unbalanced data. Although basically, the number of each class is different, it is balanced on average. The algorithm of the Roughly Balanced Bagging method is to build several training data b times. Each training data contains all minority classes but the majority class is taken randomly using the negative binomial distribution approach with $q = 0.5$. The sample taken from the majority class may be more numerous when compared to the minority class which is not balanced in data but is balanced on average. Then perform a combined prediction of the B classification that has been obtained by using the most votes rule. The following is related research using the Roughly Balanced Bagging method, including:

Research conducted by Hido et.al. (2009), in this study explained the application of the Roughly Balanced Bagging method with several other methods. The data used are 9 data obtained from the UCI Repository, namely: Diabetes, Breast, German, E-Coli-4, Santimage, Flag, Glass, RealF, Letter-A. It is known that this study applies two methods as proposed, namely the Roughly Balanced Bagging method with either return or no return with K = 100. This method is compared with several other data mining methods, namely Roughly Balanced Bagging (K = 100), Roughly Balanced Bagging with returns (K = 100), Exactly Balanced Bagging (K = 100), Original Bagging (K = 100), C4.5 (pruned), AdaBoost (K = 100), AdaBoost (K = 200), RIPPER (Optimize = 2), and RIPPER (Optimize = 10). Overall, Roughly Balanced Bagging is superior to other methods. Some of the RIPPER's accuracy rate data outperforms Roughly Balanced Bagging and it seems to be overfitting. Also, some of the Roughly Balanced Bagging results matrices outperformed others such as AUC, ISE, or G-Mean. This means that the Roughly Balanced Bagging method can handle unbalanced classes properly, as evidenced by some of the data used.

RESULTS AND DISCUSSION

This section will discuss in detail the merits of the bagging-based ensemble method. The goodness of the method is seen from the accuracy, sensitivity, and specificity value generated for each data. It will also be seen the variance of the resulting accuracy, sensitivity, and specificity values.

The following is Table 1 regarding the level of accuracy of each data analyzed using various bagging-based methods.

Table 1
Comparison of the accuracy of predictions for each data with various methods

Data	classification tree	Under Bagging	Over Bagging	UnderOver Bagging	SMOTE Bagging	Roughly Balanced Bagging	Bagging Ensemble Variation
Bank	91.5%	83.6%	86.9%	85.9%	88.6%	84.8%	83.6%
Shuttle	99.8%	99.7%	99.8%	99.7%	99.7%	99.6%	99.7%
Adult	84.7%	79.9%	81.9%	81.3%	83.6%	80.8%	80.6%
Liver	73.3%	55.5%	63.7%	63.7%	58.9%	63.7%	67.8%
Credit	82.3%	74.6%	76.6%	75.7%	76.7%	73.6%	72.8%
Transfussiun	80.3%	64.4%	63.8%	63.8%	64.9%	67.0%	59.0%
Pima	68.2%	68.2%	71.3%	72.4%	73.9%	75.5%	70.3%
White wine	93.3%	71.1%	91.1%	93.3%	88.9%	91.1%	*
Santimage	93.4%	82.9%	88.1%	84.0%	84.7%	81.1%	79.7%
Red Wine	94.5%	62.1%	85.8%	75.1%	65.6%	62.8%	55.1%
Thiroyd	95.8%	95.5%	95.1%	92.9%	95.1%	96.1%	93.1%
Ann thiroyd	97.8%	97.9%	99.8%	97.6%	97.9%	98.2%	97.8%
Letter-A	98.9%	92.1%	98.7%	97.8%	95.0%	98.2%	50.5%
Car	97.1%	71.5%	96.2%	87.9%	88.8%	69.7%	88.4%
Glass	96.4%	52.7%	98.2%	94.5%	89.1%	83.6%	76.4%
Htru	97.9%	94.6%	96.5%	96.3%	96.0%	94.6%	92.8%
Overall Accuracy	90.3%	77.9%	87.1%	85.1%	84.2%	82.5%	77.8%
High Class Comparison	96.5%	81.2%	94.8%	90.8%	89.0%	85.5%	79.2%
Low Grade Comparison	84.2%	74.6%	79.4%	79.5%	79.4%	79.5%	76.3%

Table 1 shows the highest average overall accuracy value found in the Classification Tree method and the OverBagging method at 90.3% and 87.1%. Meanwhile, the lowest average was found in the Bagging Ensemble Variation and UnderBagging methods at 77.8% and 77.9%. Also, the UnderOverBagging method is still better when compared to the Roughly Balanced Bagging and SMOTEBagging methods. In the White Wine data, the Bagging Ensemble Variation method is unable to provide a final prediction because the number of trees formed is 2 trees and has the same opportunity. This is a drawback of the Bagging Ensemble Variation method. The highest and lowest average accuracy on data with high-class comparison categories is in the same method, namely the Classification Tree method, OverBagging, and the Bagging Ensemble Variation method. The highest average data with a low imbalance category is found in the Classification Tree method while the bagging method have almost the same average.

Table 2
Comparison of the sensitivity of each data with various methods

Data	classification tree	Under Bagging	Over Bagging	UnderOver Bagging	SMOTE Bagging	Roughly Balanced Bagging	Bagging Ensemble Variation
Bank	51.5%	93.1%	91.3%	93.0%	82.7%	93.9%	92.8%
Shuttle	99.6%	99.2%	99.9%	99.9%	100.0%	98.9%	99.3%
Adult	60.3%	84.4%	83.1%	83.4%	63.0%	84.2%	83.2%
Liver	26.2%	61.9%	73.8%	71.4%	47.6%	69.1%	54.7%
Credit	36.8%	63.8%	62.9%	62.5%	58.5%	63.8%	66.7%
Transfussiun	37.8%	77.8%	57.9%	82.2%	77.8%	86.7%	68.9%
Pima	67.2%	68.7%	67.2%	77.6%	73.1%	80.6%	62.7%
White wine	91.6%	66.7%	91.7%	100%	91.7%	91.7%	*
Santimage	46.1%	76.9%	82.7%	81.7%	80.7%	81.7%	84.6%
Red Wine	14.3%	71.4%	33.3%	38.1%	61.9%	61.9%	71.4%
Thiroyd	92.7%	90.2%	78.0%	82.9%	95.1%	90.2%	97.6%
Ann thiroyd	88.7%	100%	100%	100%	100%	100%	100%
Letter-A	76.3%	90.9%	97.5%	97.9%	93.9%	81.3%	96.9%
Car	87.8%	100%	100%	100%	100%	100%	100%
Glass	33.3%	100%	100%	100%	100%	100%	100%
Htru	86.6%	91.7%	89.5%	88.3%	89.7%	92.9%	93.2%
Overall Accuracy	62.3%	83.5%	81.8%	84.9%	82.2%	86.1%	84.8%
High Class Comparison	65.7%	90.1%	85.1%	86.1%	90.2%	88.5%	93.0%
Low Grade Comparison	58.9%	77.0%	78.5%	83.8%	74.3%	83.6%	75.5%

A good method is not only seen in the accuracy value but also considers the sensitivity value. Table 2 shows the highest overall sensitivity value found in the Roughly Balanced Bagging method of 86.1% and other bagging methods have almost the same value. Meanwhile, the overall classification tree method is not able to guess the minority class correctly on the data with low and high imbalance categories. The highest sensitivity values for data with a high imbalance category were found in the Bagging Ensemble Variation, SMOTEBagging, and UnderBagging methods at 92.0%, 90.2%, and 90.1%. Then the highest average value of data with a low imbalance category is found in the UnderOverBagging and Roughly Balanced Bagging methods.

Table 1 and Table 2 show the accuracy and sensitivity values for each method, then Table 3 shows the specificity value that describes the method's ability to predict the majority class. The highest average specificity values as a whole were found in the Classification Tree method and the OverBagging method, which were 96.9% and 87.6%. Likewise, the data with a high imbalance category, the method with the highest average specificity value is found in the same method, namely the Classification Tree and OverBagging. In the data with the low imbalance category, the highest specificity values were found in the Classification Tree method and the SMOTEBagging method. Meanwhile, the value with the lowest level of specificity as a whole is found in the UnderBagging method and the Bagging Ensemble method Variation of 77.3% and 77.9%. The lowest average specificity on extreme data is found in the Bagging Ensemble Variation and UnderBagging methods, which are 78.5% and 80.6%. Finally, the lowest average specificity for data with a low imbalance category is found in the same method of UnderBagging and Bagging Ensemble Variation of 73.9% and 77.3%.

Table 3
Comparison of specificity of each data by various methods

Data	classification tree	Under Bagging	Over Bagging	UnderOver Bagging	SMOTE Bagging	Roughly Balanced Bagging	Bagging Ensemble Variation
Bank	96.6%	82.4%	86.4%	85.1%	89.3%	83.8%	82.4%
Shuttle	99.8%	99.8%	99.8%	99.7%	99.7%	99.8%	99.8%
Adult	92.8%	78.4%	81.5%	80.5%	90.5%	79.7%	79.8%
Liver	92.3%	52.9%	59.6%	60.6%	63.5%	61.0%	73.1%
Credit	95.2%	77.6%	80.4%	79.4%	81.9%	76.3%	74.6%
Transfussiun	93.7%	60.1%	65.7%	58.0%	60.8%	60.8%	55.9%
Pima	68,8%	68.0%	73.6%	69.6%	74.4%	72.8%	75.2%
White wine	93.9%	71.7%	90.9%	90.9%	87.9%	90.9%	*
Santimage	98.3%	83.5%	88.6%	84.3%	85.2%	80.9%	79.2%
Red Wine	98.9%	61.6%	88.7%	77.1%	65.8%	62.9%	54.2%
Thiroyd	96.1%	95.9%	96.6%	93.8%	95.1%	96.6%	92.7%
Ann thiroyd	98.5%	97.9%	99.8%	97.4%	97.9%	98.1%	97.6%
Letter-A	99.8%	92.1%	98.7%	97.8%	95.1%	98.8%	48.6%
Car	97.8%	69.2%	95.9%	87.0%	87.9%	67.3%	87.5%
Glass	100.0%	50.0%	98.1%	94.2%	88.5%	82.7%	75.0%
Htru	99.1%	94.9%	97.2%	97.0%	96.6%	94.7%	92.8%
Overall Accuracy	96.9%	77.3%	87.6%	84.5%	85.0%	81.7%	77.9%
High Class Comparison	98.6%	80.6%	95.5%	91.1%	89.0%	85.3%	78.5%
Low Grade Comparison	94.9%	73.9%	79.7%	78.0%	81.0%	78.1%	77.3%

The following is Figure 4 regarding the level of stability of each method seen from the value of accuracy, sensitivity and specificity.

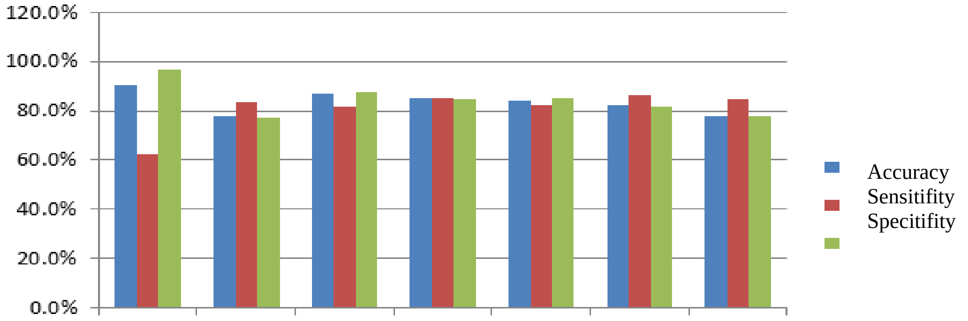

Figure 4
Comparison of the average overall accuracy, sensitivity and specificity values for each method
Source: Freund, 2007

Figure 4 shows the comparison of the bagging-based classification method with 16 data. Overall the bagging-based method is more stable than the tree classification method. The Roughly Balanced Bagging and Bagging Ensemble Variation methods can predict the minority class well and without having to ignore the majority class and its accuracy results. In the Transfusion data, the sensitivity value of the Roughly Balanced Bagging method is higher when compared to other bagging-based methods. This proves that the Roughly Balanced Bagging method is strong against data with various conditions.

The Bagging Ensemble Variation method is a method that can handle unbalanced class problems in the high imbalance comparison category but lacks the Bagging Ensemble Variation method when the number of trees obtained is even and each prediction has the same chance, it cannot be predicted. Meanwhile, overall, the method.

with the worst performance is the UnderBagging method. However, the computation process in the UnderBagging method is much faster when compared to other bagging-based methods. While the OverBagging and SMOTEBagging methods even though have good results, the computation process in both methods takes a longer time. Another disadvantage of the SMOTEBagging method is that it requires accuracy because when it is wrong to raise the minority class, this method will only focus on the minority class and ignore the accuracy and predictions of the majority class. Then the last one is a method that is no less good than other methods, namely UnderOverBagging. Overall, the UnderOverBagging method gives good results in terms of sensitivity, specificity, and accuracy. The goodness of the model is not only seen from the accuracy, sensitivity, and specificity values but also the variance of the good values of the model itself. The variances of the goodness of the model are shown in Figure 5, Figure 6, and Figure 7.

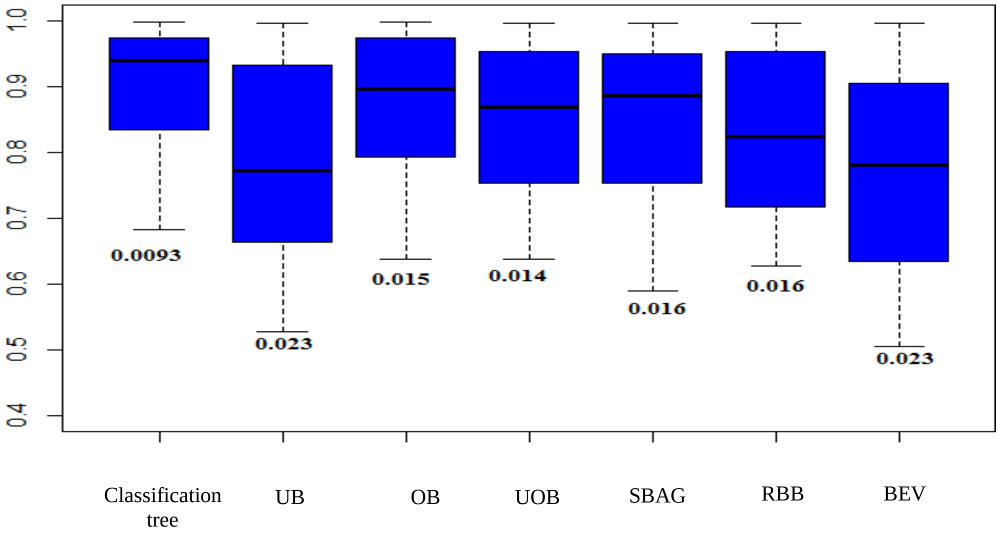

Figure 5
Value of the diversity of predictive accuracy of each ensemble method bagging based
Source: Analysis by RStudio

Figure 6 is intended to corroborate the information contained in Table 5. Here is Figure 6 information about the accuracy variance generated in each method. The level of accuracy of the UnderBagging and Bagging Ensemble Variation methods has a greater variance when compared to other bagging-based methods. Meanwhile, the OverBagging, UnderOverBagging, SMOTEBagging and Roughly Balanced Bagging methods have almost the same variance values. The accuracy variance in the Classification Tree method has a lower value compared to other methods.

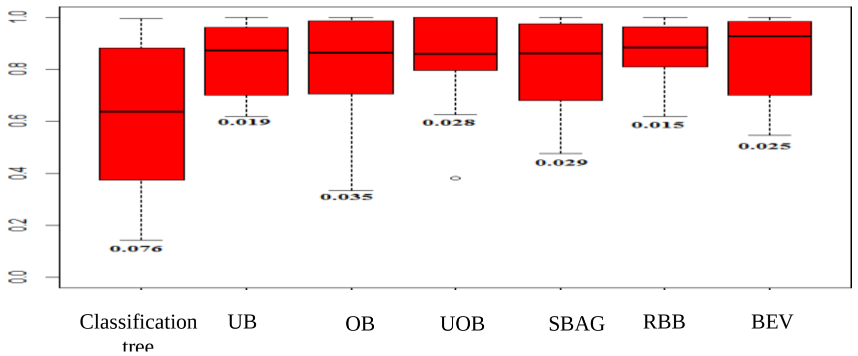

Figure 6
The value of the sensitivity diversity of each ensemble method bagging based
Source: Analysis by RStudio

The ability of each method is not only seen from the variance of accuracy but is also supported by the variance value of its sensitivity. Figure 7 provides information on the sensitivity variance of each method. The highest sensitivity variant is found in the Classification Tree Method. Whereas in the bagging-based ensemble method, the highest sensitivity variance was found in the OverBagging, SMOTEBagging, and Bagging Ensemble Variation methods. The lowest sensitivity variant is found in the Roughly Balanced Bagging, UnderBagging method and the last one is the UnderOverBagging method.

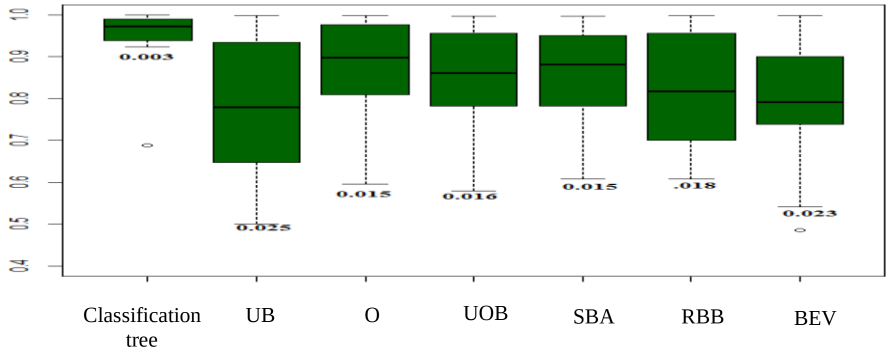

Figure 7
The value of the specificity diversity of each ensemble method bagging based
Source: Analysis by RStudio

Furthermore, the last is in Figure 7 information about the specificity variance value of each method. The lowest specificity variance value for each method is found in the Classification Tree Method. Then in the bagging-based ensemble method, the specificity variance values in the OverBagging, UnderOverBagging, SMOTEBagging, and Roughly Balanced Bagging methods have almost the same specificity variance values. However, the variance value in this method is better than the specificity variance value in the UnderBagging and Bagging Ensemble Variation methods.

CONCLUSION

Based on the results and discussion, it was concluded that: The Roughly Balanced Bagging method can predict minority classes well in various kinds of data without having to ignore the majority class and the accuracy of its prediction. Meanwhile, the Bagging Ensemble Variation method gives good results when the data imbalance category is high. However, this method will remove some of the majority class that does not belong to the minority class. Apart from that, other bagging based methods such as Over Bagging, Under OverBagging, and SMOTEBagging are still fairly stable in handling unbalanced classes on various kinds of data. It's just that the data Over Bagging and SMOTE Bagging methods take a long time in the computation process. Also, the SMOTEBagging method requires more accuracy in dealing with unbalanced class problems because if it generates too much synthesis data, the method only focuses on the minority class and ignores its accuracy and specificity. Then the last one, the UnderBagging method, is a method that has poor performance on all 16 data when compared to other bagging based methods. It's just that the computation process in this method is faster when compared to other bagging-based methods. Although it is considered less good when compared to some bagging based methods, the Under Bagging method can provide better results when compared to classical methods such as classification trees. Because the UnderBagging method can improve the classification results for minority classes while the Classification Tree method only focuses on accuracy and specificity values.

REFERENCES

Barandela V.R., Sanchez R. JS. "*New appllications of ensembles of classifiers*," *Pattern Anal*, vol. 6, pp. 245–256, 2003.

Barro A. F., R. Sulviant IDi, "*The application of synthetic minority oversampling technique (Smote) to unbalanced data in making herbal composition model*s," *J. Stat.*, vol. 1, no. 1–6, 2013.

Bauer E., "An empirical comparison of voting classification algorithms: bagging, boosting and variants.," vol. 36, pp. 15–139, 1999.

Bisri A, "Adaboost application to resolve class imbalances in determining student graduation using the decision tree method," *J. Intell. Syst.*, vol. 1, pp. 27–32, 2015.

Breiman L., "*Bagging predictors machine learning,*" vol. 24, pp. 123–140, 1996.

Chawla K. N.V, Japkowicz N. "*Special issue on learning from imbalanced data sets*.," *SIGKDD Explor. Newsl.*, vol. 6, pp. 1–6, 2004.

Efron T. R. B. *An introduction to the bootstrap*. New York: Chapman & Hall, 1993.

Elrahman A. A. SM, "*A review of class imbalance problem*," *J. Netw. Innov. Comput.*, vol. 1, pp. 332-340., 2013.

Freund Y., "Classifying imbalanced data using a bagging ensemble variation (BEV).," *ACM Southeast Conf.*, pp. 203–208, 2007.

Galar H.F. M., Fernandez A, Barrenechea E, Bustince H, "*A review on ensembles for the class imbalance problem: bagging.boosting and hybrid- based approaches.*," *IEEE Trans. Syst.*, vol. 42, pp. 463 – 484, 2012.

Gónzalez H. F. S, García S, Lázaro M, Vidal ARF, "*Class switching according to nearest enemy distance for learning from highly imbalanced data- sets*. Patern Recognition.," vol. 70, pp. 12–24, 2017.

Hido T.Y., Kashima H, "Roughly balanced bagging for imbalanced data. Stat.," *Stat. Anal. Data Min.,* vol. 2, pp. 412–426, 2009.

Japkowicz N., "Handling the class imbalance problem in binary classification," *Masdar Inst. Sci. Technol.,* 2014.

Longadge M.L. R, Dongre SS, "*Class imbalance problem in data mining:* Review.," *Int. J. Comput. Sci. Netw. (IJCSN).,* vol. 2, pp. 83–88, 2013.

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

Park Y, "*Ensembles of α-trees for imbalanced classification problems,*" *J. Latex Cl. Files.,* vol. 6, pp. 1–14, 2007.

Permatasari Y., *"Handling unbalanced class problems with rusboost and underbagging (Case Study: Drop Out Students of SPs IPB* Masters Program," 2016.

Ralescu A. A. "Predicting software aging related bugs from imbalanced datasets by using data mining techniques," *IOSR J. Comput. Eng.,* vol. 18, pp. 27–35, 2016.

Ramyachitra M.P. *"Imbalanced datasets classification and solutions*," *Int. J. Comput. Bus. Res.,* vol. 5, pp. 1–29, 2014.

Rodrigo. L., "*Ensemble-based classifiers.* Artif. Intell," vol. 33, pp. 1–39, 2010.

Sartono S.U. B. "*Combined tree method: the preferred solution to overcome the weaknesses of single regression and classification trees*," vol. 15, pp. 1–7, 2010.

Schouts R., "*An overview of the advantages of ensemble classification trees to improve the predictive ability of single classification trees*.," vol. 9, no. 33–38, 2015.

Sun Z, Song Q, Zhu X, Sun H, Xu B, "A novel ensemble method for classifying imbalanced data. Pattern Recognition," vol. 48, pp. 1623–1637, 2015.

Wang Y. X. S, "*Diversity analysis on imbalanced data sets by using ensemble models*," *IEEE Symp. Comput. Intell. Data Min.,* vol. 324–331, 2009.

Yusof R, Kasmiran KA, Mustapha A, Mustapha N, "Techniques for handling imbalanced datasets when producing classifier models," *J. Theor. Appl. Inf. Technol.,* vol. 95, pp. 1425–1440, 2017.

Zhang J. H. D, Liu W, Gong X, "*A novel improved smote resampling algorithm based on fractal. Article Computational Information Systems*," *Artic. Comput. Inf. Syst.,* pp. 2204–2211, 2011.

Zhu B., B. Baesens B, Seppe K.L.M, "*An empirical comparison of techniques for the class imbalance problem in churn prediction*," *Inf. Sci.,* vol. 408, pp. 84–99, 2017.