Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 2, April 2022
DOI : https://doi.org/10.33395/sinkron.v7i2.11427
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Comparative analysis of resampling techniques on Machine Learning algorithm

Tri Suci Amelia[1)*], Mila Nirmala Sari Hasibuan[2)], Rahmadani Pane[3)]
[1)2)3)]Universitas Labuhanbatu, Indonesia
[1)]sucirambe7@gmail.com, [2)] milanirmalasari7@gmail.com, [3)] rahmadanipane@gmail.com

Submitted : Apr 30 | **Accepted** : Apr 30, 2022 | **Published** : May 2, 2022

Abstract:. Generally, classification algorithms in the field of data science assume that the classes of training data are equally distributed. However, datasets on real problems often have an unbalanced class distribution. Unbalanced dataset classes make up the majority class and the minority class. In general, minority classes are more attractive and more important to identify. In this case, the correct classification for the minority class sample is more valuable than the majority class. The unbalanced class distribution causes the classification algorithm to have difficulty in classifying minority class samples correctly. If the performance of the algorithm model is good for the majority class sample but bad for the minority class then this imbalance problem is a crucial thing to be addressed. Many solutions are offered for this problem, namely by oversampling techniques in the minority class and/or undersampling techniques in the majority class. In this study, the authors tried various sampling techniques and tested them on various machine learning classification algorithms to find out the combination of resampling techniques and algorithms that have high recall in classifying minority class samples and still considering the majority class classification.

Keywords: ADASYN; Imbalanced data; KNN; Machine Learning; Resampling.

INTRODUCTION

In general, machine learning algorithms, in this case on classification problems, work with the main goal of maximizing accuracy (Provost, 2000). This makes a lot of sense, because the high accuracy means that the model algorithm does its job well, classifying data classes correctly with few errors. However, accuracy only provides general information, what if the algorithm model works on an unbalanced dataset, and is only able to correctly classify the majority class but cannot classify the minority class. If the comparison between the minority and majority classes is only one in one hundred, then the accuracy that will be obtained is greater than 99%, with an error of less than 1% which is almost entirely the minority class. This problem biases the performance of classification algorithms, especially if the priority class to classify correctly is a minority class, such as spam email, medical diagnosis, fraudulent credit card detection and others (Visa & Ralescu, 2005) (Rahman & Davis, 2013). This shows that in the case of unbalanced datasets, more attention is needed to preprocess the data before it is entered into the model.

Many ways have been found to overcome this unbalanced dataset, such as resampling the existing data. Resampling is a technique of taking samples repeatedly from the original data sample (Statistic Solutions, 2016). The resampling technique consists of oversampling, which is taking samples repeatedly from the minority class; and undersampling, which is taking a random sample from the majority class (Burnaev, Erofeev, & Papanov, 2015). These two techniques can be used separately or in combination (Anand, Pugalenthi, Fogel, & Suganthan, 2010) (More, 2016) (Yen & Lee, 2006). SMOTE is the most popular oversampling technique, with Borderline-SMOTE an extension of SMOTE. One resampling technique that is quite popular is ADASYN which is able to adjust the amount of synthetic data.

In several related studies (More, 2016) (Batista, Prati, & Monard, 2004) (Amin et al., 2016) (Burnaev et al., 2015), various experiments have been carried out to overcome the problem of unbalanced datasets, but these methods have The resampling methods and machine learning algorithms used do not vary to find out the best method to solve this problem. As research conducted by Amin (Amin et al., 2016) only examined oversampling techniques. Burnaev, More, and Batista et al (Batista et al., 2004) (Burnaev et al., 2015) researched oversampling and undersampling techniques but only used one machine learning algorithm, while Diri (Diri & Albayrak, 2008) only examined several machine algorithms. learning without considering unbalanced datasets.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 2, April 2022 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v7i2.11427 p-ISSN : 2541-044X

Meanwhile, to find out the best resampling method and machine learning algorithm for this problem, combinations of resampling techniques are needed, as well as between machine learning algorithms. Each of these combinations (pairs), such as SMOTE with Support Vector Machine, or Tomek Links with Logistics Regression, will be tested for its performance against a given dataset, then from these combinations conclusions can be drawn regarding the best combination of algorithms and resampling techniques, and machine learning algorithms. with the best performance, and the best performing resampling technique. Each combination or pair is evaluated for results not only on one unbalanced dataset, but with several additional datasets to get more general results.

Based on the description above, this study aims to perform a comparative analysis of resampling techniques on machine algorithms on unbalanced datasets. The formulation of the problem in this research is a combination of machine learning algorithms and resampling techniques which have good performance to overcome unbalanced datasets..

LITERATURE REVIEW

Machine learning is programming a computer to optimize a performance measure using sample data or based on experience (Alpaydin, 2014). Machine learning uses an algorithm to analyze data. In supervised learning or supervised learning, the classifier will be given a certain input and relate it to an output. The case where the goal is to classify the input data into a certain discrete category is called classification, and the case where the output is a continuous variable is called regression. In unsupervised learning, the classifier is given input and left alone to find patterns in the data. The case of unsupervised learning where the goal is to group similar observations is called clustering, if determining the distribution of data on the input is called density estimation. In reinforcement learning, the computer system receives input continuously and tries to choose the most optimal decisions based on environmental conditions. Each type of learning has many algorithms that have been developed with different approaches (I. & M., 2015).

A dataset is a collection of data in the form of a table, where each column represents a feature, attribute or feature. Each line represents the observation of an individual, record or sample (Snijders, Matzat, & Reips, 2012). A dataset usually has one additional column that represents the class of the observations, this column is called the class column. This class column is also referred to as the dependent variable on the independent variables which are the characteristics (attributes) of a particular observation.

This imbalance problem will bias the performance of the classifier because the number of samples in a particular class cannot provide sufficient information to the classifier based on the given characteristics (Bishop, 2021) (Pedro, 2012).

Random oversampling, or random oversampling is an oversampling technique in which members of a minority class are randomly selected and duplicated into a new dataset until equilibrium is reached (Liu, 2004). The minority data can be duplicated several times. This technique usually causes overfitting of the model (Alpaydin, 2014).

Random undersampling, atau undersampling secara acak adalah teknik undersampling di mana anggota dari kelas mayoritas dipilih secara acak dan dihapus dari dataset training hingga tercapai keseimbangan. Kekurangan dari teknik ini adalah tidak ada cara untuk mengatur informasi apa saja yang dihilangkan dari dataset tersebut, informasi yang berguna bisa saja hilang (Amin et al., 2016).

METHOD

To complete this research, several stages of research were carried out, namely: Pre-research, exploration and preprocessing of data, model tuning and fitting, and analysis of results.

At the pre-research stage, the research theme is determined, the problem to be researched, collects reference sources or literature such as journals and books that support the research, and determines the method used and the limitations of the problem. Then the researchers looked for data that matched the research theme as the object of research. At the data exploration and preprocessing stage, the characteristics of each dataset are described as information for making decisions at the preprocessing stage. Identify the problems contained in the dataset then take an approach to solve the related problems. Normalization and attribute reduction are included at this stage. At the tuning and fitting model stage, the best parameters for the model to be used are searched based on the results of data exploration and trial and error to get the best results. Tuning is also carried out on several resampling techniques that require parameters. Then the model will provide predictive results which will be analyzed at a later stage.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 2, April 2022
DOI : https://doi.org/10.33395/sinkron.v7i2.11427
e-ISSN : 2541-2019
p-ISSN : 2541-044X

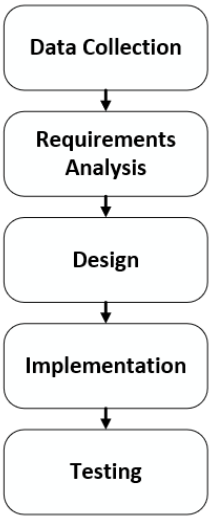

Fig 1. Research Framework

Data were taken from the official Kaggle website (kaggle.com), UCI Machine Learning Repository (archive.ics.uci.edu/ml/) and KEEL (sci2s.ugr.es/keel/imbalanced.php). The data are in the form of three datasets, namely:

1. Credit Card Fraud Dataset (Kaggle), which consists of 30 attribute columns with 1 class column, 284,807 rows. 284,315 the number of samples of the majority class and 492 the number of samples of the minority class with an imbalanced ratio of 577:1. This dataset is the most popular dataset at Kaggle because of the large amount of data with a very high imbalanced ratio.
2. Spambase Dataset (UCI), which consists of 57 attribute columns with 1 class column, 4,601 rows. 2788 the number of samples of the majority class and 1813 the number of samples of the minority class with an imbalanced ratio of 1.5:1.
3. Image Segmentation Dataset (KEEL), which consists of 19 attribute columns with 1 class column, 2308 rows. 1962 the number of samples of the majority class and 346 the number of samples of the minority class with an imbalanced ratio of 6:1.

All datasets only have continuous attributes with binary class labels that match the research theme and the methods used.

RESULT

Image Segmentation dataset is a dataset about images where each pixel has a class based on the results of manual segmentation of the outdoor image. Each observation of this dataset is a composite of pixels measuring 3x3 (called a region). In the original dataset, there were six different classes, but this dataset was modified by Keel where there were only 2 classes, namely positive and negative. Positive classes are class 0 in the original dataset, and negative classes are classes 1, 2, 3, 4 and 5 in the original dataset.

Image Segmentation This dataset belongs to the unbalanced dataset category, where the majority class is a negative category region with a total of 1962 and the minority class is a positive category region with a total of 346.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 2, April 2022
DOI : https://doi.org/10.33395/sinkron.v7i2.11427
e-ISSN : 2541-2019
p-ISSN : 2541-044X

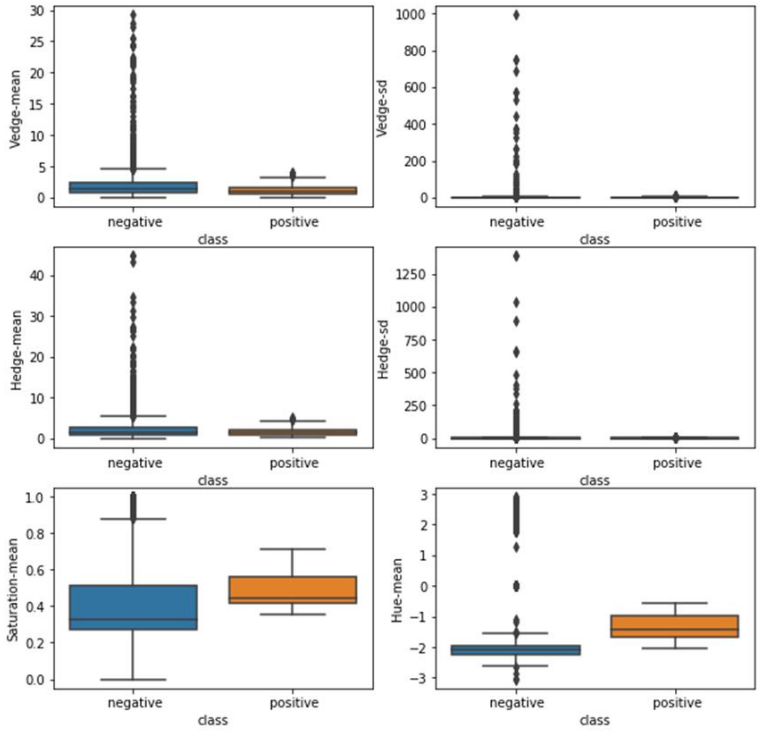

Fig 2. Boxplot Image Segmentation Dataset

Figure 2 shows that there are many extreme outliers that are far from 90% of the data distribution. However, these extreme outliers are only spread over the negative class (the majority), which means that removing some of these outliers has no impact on the positive class (the minority)..

Table 1. Capital Distribution

	capital_run_length_average	capital_run_length_longest	capital_run_length_total
count	4601.000000	4601.000000	4601.000000
mean	5.191515	52.172789	283.289285
std	31.729449	194.891310	606.347851
min	1.000000	1.000000	1.000000
25%	1.588000	6.000000	35.000000
50%	2.276000	15.000000	95.000000
75%	3.706000	43.000000	266.000000
max	1102.500000	9989.000000	15841.000000

Table 1 shows that the maximum value of each attribute differs greatly from the range of values [1, 100]. Then these attributes need to be normalized to scale [1, 100] using MinMax. However, there is a very large jump from the third quartile (75%) to the maximum value of each of the above attributes. This indicates that there are extreme outliers that bias the attribute standard deviation, and have a negative impact if the data is scaled with outliers. Then normalization will be performed after the extreme outliers have been removed from the dataset.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 2, April 2022
DOI : https://doi.org/10.33395/sinkron.v7i2.11427
e-ISSN : 2541-2019
p-ISSN : 2541-044X

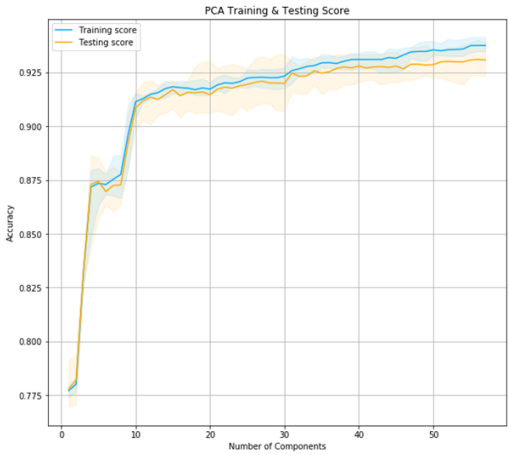

Fig 3. PCA Training and Testing Score

Figure 3 shows that using 58 components ensures the highest training and testing accuracy. Figure 11 also shows that the first 10 components contribute greatly to the accuracy of the model, with an average accuracy of 91% with a difference of less than 2% compared to using all components. The model using 58 components obtained a high testing score with only a slight difference from its training score. but the difference between the training score and the testing score is clearly visible after the first 30 components, with the biggest difference or bias occurring when all components are used.

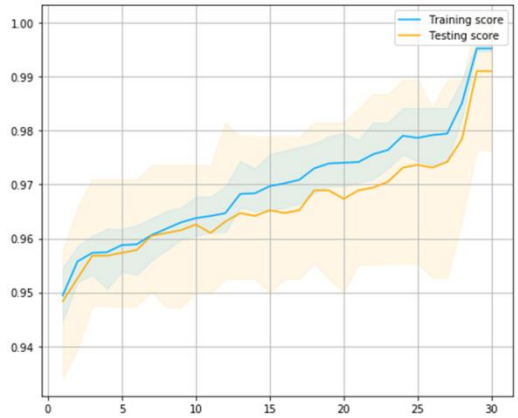

Fig 4. Training & Testing Score Credit Card Fraud

Figure 4 shows that using all components ensures the highest training and testing accuracy. Figure 14 also shows that the first 2 components contribute greatly to the accuracy of the model, with an average accuracy of 95.5% with a difference of less than 4% compared to using all components. Running time is the most important aspect of this dataset. Figure 15 shows that running time grows linearly based on the number of components. Using 2 main components of PCA has a difference in running time of 3.73 times than using all components.

Table 2. Time difference from the total fitting on the original dataset and the 200 datasets approach

	original dataset	*200 datasets approach*
time(s)	>15000	~6000

Table 2 shows that the 200 dataset approach (200 datasets approach) can complete 30 combinations on the Credit Card Fraud dataset in 6,000 seconds, while without the rebalancing dataset it takes more than 15,000 seconds.

Table 3. Machine learning algorithm performance

clf	*accuracy*	*precision-0*	*precision-1*	*recall-0*	*recall-1*	*f1-0*	*f1-1*
lr	0.948	0.972	0.902	0.948	0.942	0.960	0.919

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 2, April 2022
DOI : https://doi.org/10.33395/sinkron.v7i2.11427
e-ISSN : 2541-2019
p-ISSN : 2541-044X

svm	0.955	0.975	0.916	0.957	0.946	0.966	0.930
mlp	0.956	0.976	0.921	0.958	0.946	0.966	0.931
dt	0.921	0.947	0.893	0.935	0.870	0.938	0.870
knn	0.949	0.977	0.895	0.944	0.951	0.960	0.921

Table 3 shows that based on recall-1, all machine learning algorithms have similar performance (0.942 – 0.951) with the exception of the Decision Tree algorithm which has a recall-1 of 0.870..

Table 4. Resampling technique performance

clf	accuracy	precision-0	precision-1	recall-0	recall-1	f1-0	f1-1
lr	0.948	0.972	0.902	0.948	0.942	0.960	0.919
svm	0.955	0.975	0.916	0.957	0.946	0.966	0.930
mlp	0.956	0.976	0.921	0.958	0.946	0.966	0.931
dt	0.921	0.947	0.893	0.935	0.870	0.938	0.870
knn	0.949	0.977	0.895	0.944	0.951	0.960	0.921

Table 4 shows that based on recall-1, ADASYN is a resampling technique that has the highest detection rate for the minority class with a value of 0.953.

By averaging the results from the three datasets, Table 5 below is the 10-highest combination of machine learning algorithms sorted by recall-1 (minority class detection rate).

Table 5. 10-highest combination of machine learning algorithms and resampling techniques sorted by recall-1

Res	clf	accuracy	precision-0	precision-1	recall-0	recall-1	f1-0	f1-1
adasyn	mlp	0.936	0.984	0.852	0.921	0.969	0.951	0.904
adasyn	lr	0.927	0.981	0.819	0.909	0.967	0.943	0.885
adasyn	svm	0.939	0.983	0.855	0.927	0.966	0.953	0.905
adasyn	knn	0.931	0.984	0.840	0.913	0.966	0.946	0.897
bsmote	lr	0.929	0.981	0.824	0.912	0.965	0.945	0.888
bsmote	mlp	0.943	0.982	0.870	0.932	0.964	0.956	0.912
bsmote	svm	0.944	0.982	0.869	0.934	0.963	0.957	0.913
bsmote	knn	0.941	0.982	0.869	0.928	0.963	0.953	0.912
Rus	knn	0.945	0.979	0.875	0.939	0.954	0.958	0.912
smote	knn	0.951	0.977	0.900	0.948	0.951	0.962	0.924

DISCUSSIONS

Borderline-SMOTE and ADASYN have the highest recall in predicting minority classes, but at the expense of accuracy. This is because Borderline-SMOTE and ADASYN create a lot of synthetic data around the majority class, giving clear boundaries between minority and majority classes. In the three datasets, Borderline-SMOTE and ADASYN consistently have the highest recall.

Tomek Links has the lowest recall in predicting minority classes when compared to other resampling techniques. This is because Tomek Links cannot balance the number of majority and minority classes. Practically, this technique is only an outlier removal technique when other resampling techniques have been performed.

KNN has a good recall because the number of minority neighbors will be very dominant when the class has been balanced with resampling techniques. But it has a very bad impact on the precision of the minority class, resulting in a lot of false positives.

MLP and SVM performance is the best on machine learning classification algorithms, but with the highest running time. These two methods consistently have the best recall across the three given datasets, making them ideal classifiers for data imbalance problems.

Decision Tree is a machine learning algorithm that has the worst performance in these three datasets, this is because the decision tree is a classification algorithm that is better used on datasets with categorical attributes, while the three datasets used all have continuous attributes..

*name of corresponding author

CONCLUSION

ADASYN and Borderline-SMOTE are the best resampling techniques for recognizing minority classes, but with a slight decrease in accuracy. SMOTE is below ADASYN and Borderline-SMOTE on recall-1, but SMOTE is far superior to precision-1 and accuracy in general, which results in higher f1-1 values than ADASYN and Borderline-SMOTE. SMOTE is also better at recall-0, which is ideal if false positives are also highly undesirable in a dataset. KNN has the best recall compared to all other classifiers when the class has been balanced by resampling technique, but MLP and SVM have higher accuracy with very thin recall differences than KNN (2%). MLP and SVM also have a much higher precision-1 than KNN, resulting in a higher f1-1 value. In general, MLP and SVM performed better on the three given datasets. The best combination of machine learning algorithms and resampling techniques is adasyn_mlp for the highest recall, smote_knn for the highest f1 value (balance between recall and precision), and tl_mlp for accuracy in general. Datasets with classes with a very high level of imbalance, such as the Credit Card Fraud Dataset, are easier to process if the datasets are grouped as in section 4.1.3.3 Dataset Rebalancing. This approach shows the large difference in recall-1 when compared to data processed directly before being grouped.

REFERENCES

Alpaydin, E. (2014). *Introduction to Machine Learning (third edition)*.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., … Hussain, A. (2016). Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study. *IEEE Access*, *4*, 7940–7957. https://doi.org/10.1109/ACCESS.2016.2619719

Anand, A., Pugalenthi, G., Fogel, G., & Suganthan, P. (2010). An approach for classification of highly imbalanced data using weighting and undersampling. *Amino Acids*, *39*, 1385–1391. https://doi.org/10.1007/s00726-010-0595-2

Batista, G., Prati, R., & Monard, M.-C. (2004). A Study of the Behavior of Several Methods for Balancing machine Learning Training Data. *SIGKDD Explorations*, *6*, 20–29. https://doi.org/10.1145/1007730.1007735

Bishop, C. M. (2021). Pattern Recognition and Machine Learning. In *EAI/Springer Innovations in Communication and Computing*. https://doi.org/10.1007/978-3-030-57077-4_11

Burnaev, E., Erofeev, P., & Papanov, A. (2015). Influence of resampling on accuracy of imbalanced classification. *Eighth International Conference on Machine Vision (ICMV 2015)*, *9875*, 987521. https://doi.org/10.1117/12.2228523

Diri, B., & Albayrak, S. (2008). Visualization and analysis of classifiers performance in multi-class medical data. *Expert Systems with Applications*, *34*(1), 628–634. https://doi.org/https://doi.org/10.1016/j.eswa.2006.10.016

I., J. M., & M., M. T. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Liu, A. Y. (2004). The Effect of Oversampling and Understanding on CLassifying Imbalanced Text Datasets.

More, A. (2016). *Survey of resampling techniques for improving classification performance in unbalanced datasets*. *10000*, 1–7. Retrieved from http://arxiv.org/abs/1608.06048

Pedro, D. (2012). A Few Useful Things to Know About Machine Learning. *Communications of the ACM*, *55*(10), 9–48. Retrieved from https://dl.acm.org/citation.cfm?id=2347755

Provost, F. (2000). *Machine Learning from Imbalanced Data Sets 101 Extended Abstract*.

Rahman, M., & Davis, D. N. (2013). Addressing the Class Imbalance Problem in Medical Datasets. *International Journal of Machine Learning and Computing*, *3*, 224. https://doi.org/10.7763/IJMLC.2013.V3.307

Snijders, C., Matzat, U., & Reips, U.-D. (2012). "Big Data" : Big Gaps of Knowledge in the Field of Internet Science. *International Journal of Internet Science*, *7*, 1–5.

Statistic Solutions. (2016). Resampling. Retrieved April 10, 2022, from statisticsolutions.com website: https://www.statisticssolutions.com/dissertation-resources/sample-size-calculation-and-sample-size-justification/resampling/

Visa, S., & Ralescu, A. (2005). Issues in Mining Imbalanced Data Sets - A Review Paper. *Proc. 16th Midwest Artificial Intelligence and Cognitive Science Conference*.

Yen, S.-J., & Lee, Y.-S. (2006). *Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset BT - Intelligent Control and Automation: International Conference on Intelligent Computing, ICIC 2006 Kunming, China, August 16–19, 2006* (D.-S. Huang, K. Li, & G. W. Irwin, Eds.). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-37256-1_89

*name of corresponding author