Handling of unbalanced LC-MS medicinal plant data using Near-Miss Undersampling tested with Gaussian Naive Bayes and K-Nearest Neighbors

Iwan Binanto*
Informatics Department
Sanata Dharma University
Yogyakarta, Indonesia
*iwan@usd.ac.id

Rosalia Arum Kumalasanti
Informatics Department
Sanata Dharma University
Yogyakarta, Indonesia
rosalia.santi@usd.ac.id

Nesti Fronika Sianipar
Biotechnology Departement, Faculty of Engineering, Bina Nusantara University.
11480 Jakarta, Indonesia
Research Center Food Biotechnology, Bina Nusantara University, 11480 Jakarta, Indonesia
nsianipar@binus.edu

***Abstract*— Imbalanced data refers to data with classes that have extreme majority and minority data. Such data can lead to inaccurate results. The dataset used in this study came from LC-MS data of medicinal plants that had previously been labeled using the webscraping method and unbalance. There are several resampling algorithms to balance the data. This study used near-miss undersampling with consideration for being more robust against overfitting. The balanced data was split for training and testing with a ratio of 70:30, which will be tested using Gaussian Nave Bayes and K-Nearest Neighbors classification algorithms. The results showed that Near Miss version 1 sampling with the Gaussian Naive Bayes algorithm provided better accuracy and faster execution time.**

Keywords—imbalance data, LC-MS, classification, machine learning, undersampling

I. INTRODUCTION

Imbalanced data is data where one class has significantly more observations compared to the other, causing the majority class to be given priority over the minority class in machine learning algorithms. The majority-minority class ratio can reach 100:1, 1000:1, or even 10000:1. Further to binary-class data, this issue also affects multi-class data (more than two classes). The majority class is usually referred to as the positive label, and the minority class is known as the negative label [1]. This can lead to inaccurate results.

This problem is very interesting because it may be seen in many classification problems in the real world, such fraud, risk management, the identification of contaminants, and remote sensing [2]. In addition, it also arises in cancer diagnosis [3], computer network security [4–6], detecting hard drive failures [7], and other fields [8–11]. These studies used existing resampling algorithms [3–10, 12–13] with their own data. There are two algorithms used in that studies which are SMOTE and Near Miss Undersampling.

SMOTE works by increasing the observations of the minority class, which can lead to overfitting and other issues [5, 8, 9, 12, 13]. On the other hand, some advantages of the near-miss undersampling method include preventing overfitting, preserving information, addressing the problem of classifying minority classes, being effective for large majority classes, increasing accuracy, reducing bias, and improving the representation of minority classes. However, near-miss undersampling also has some disadvantages, such as not ensuring that the samples left from the majority class are the most representative and may cause problems in the model's performance on new data [8, 10]. NearMiss version 1 selects the three closest examples from the majority class and skips over the ones with the least average distance to them.

As a result, the NearMiss version 1 samples selected are similar to certain minority occurrences. NearMiss v2 selects from the majority class instances with the smallest average distance to the three furthest minority classes.

In simple terms, NearMiss version 2 looks for majority samples that are similar to all minority situations. In NearMiss version 3, there are k-instances of the majority class surrounding each instance of the minority class. For each minority occurrence, a considerable number of the nearest majority samples are picked [14, 15].

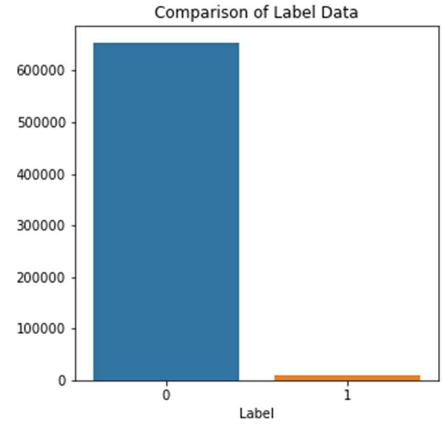

Fig. 1. Unbalanced LCMS data of medicinal plants

979-8-3503-2666-6/23/$31.00 ©2023 IEEE
DOI 10.1109/ICCMSO59960.2023.00033

This study used a dataset from LC-MS data of medicinal plants that had been labeled using the webscraping technique in previous studies [16, 17]. This dataset consists of 663,228 rows with 7 columns. This dataset is prepared for supervised learning. The label used is a binary label. This dataset is unbalanced because there is a label "1," which is a minority, as shown in Fig. 1. This makes machine learning results inaccurate. This study will use the Near Miss Undersampling algorithm to handle data imbalances. As a test, supervised machine learning will be used using the Gaussian Nave Bayes and K-Nearest Neighbors algorithms. The use of these two algorithms is important because they are very affected by unbalanced data.

II. Literature Review

The data used in this research originated from the work of Sianipar et al. [18-22], who obtained their LC-MS data using tuber sections. The mutant plant outperformed the regular plant in this experiment. When processed using the data labeling employed in prior studies [16, 17], this data yields opposite outcomes. There are several methods for integrating the data. Using various data balancing techniques, Bagui et al. conclude that undersampling shortens training time and oversampling lengthens it; oversampling and undersampling both significantly improve recall when the data is highly unbalanced; and resampling has little effect if the data is not very balanced [5]. Uneven data, according to Haixiang et al., will impact classification results since it causes bias. The minority class can cause bias, which is known as noise [8]. Johnson et al. demonstrated this when investigating deep learning issues using unbalanced data [9]. Tanimoto et al. claimed that the Near Miss technique can improve classification by decreasing the majority sample size while retaining data information [10]. They show how the Near Miss technique enhances classification algorithm performance on unbalanced data sets including medical and network security datasets. [10]. The problem of psychological data, according to Rekha et al., is a complex and tough topic in machine learning. They concluded that there are numerous unsolved questions and numerous avenues for future research on how to address the problem of data inequality [11]. Several studies [23-26] compared the results of Naive Bayes and K-Nearest Neighbors classification techniques. Safri et al. used Nave Bayes with K-Nearest Neighbors to improve accuracy [26]. Nave Bayes is frequently employed due of its competitive accuracy and processing economy, according to Anand et al. [27]. Based on these considerations, this study will compare the two approaches using data that has been balanced by undersampling Near Misses.

The Naive Bayes [28] model is simple to develop and may be applied to a wide range of data sets. Naive Bayes can handle extremely complex classification procedures. Our hypothesis (h) in the classification job could be the class to be allocated to the new data instance (d Based on our prior information, we can determine the probability of a hypothesis using Bayes' theorem. The following is an illustration of Bayes' theorem:

$$P(h|d) = (P(d|h) * P(h)) /p(d) \tag{1}$$

P(h|d) represents the likelihood that hypothesis h will arise from data d. If hypothesis h is true, P(d|h) reflects the chance that data d exists. P(h) denotes the likelihood that hypothesis h is correct regardless of the data. The goal of our job, as we can see, is to estimate the future likelihood P(h|d) given the prior probability P(h) and the inputs P(d) and P(d|h). It is possible to compose:

$$MAP(h) = max(P(h|d)) \tag{2}$$

or

$$MAP(h) = max((P(d|h) * P(h)) / P(d)) \tag{3}$$

or

$$MAP(h) = max(P(d|h) * P(h)) \tag{4}$$

Probabilities can be defined using P(d), the normalized term. We may disregard it because it is a constant and only use it to normalize if we are only interested in the most likely hypothesis. If our training data has the same number of samples from each class, the probability (P(h)) for each class will be the same. We may remove this component because it is a constant in the equation, resulting in the formula (5).

$$MAP(h) = max(P(d|h) \tag{5}$$

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{6}$$

K-Nearest Neighbors has the simplest approach in statistical classification and is also the smartest technique ever discovered. The data feeder (Euclidean equation) will be trained into an algorithm in this way. Data sets can be gathered from classes that have been set up to be used as training data. The Euclidean equation is used to determine the distance between the test and training data points, where i denotes the attribute value and specifies the number of attribute dimensions [29].

III. Method

The dataset that has been obtained from previous research [16, 17] is visualized so that it is well seen that the dataset is unbalanced data, as shown in Fig. 1. This dataset is balanced with the Near Miss Undersampling algorithm versions 1 and 3. The data distribution is illustrated using the results. The data balance is then separated into training and testing data in a 70:30 ratio. Training data is tested with Gaussian Naive Bayes and K-Nearest Neighbors machine learning algorithms to obtain a model. This model is then tested with test data. This test will involve the confusion matrix to see accuracy, precision, recall, and F1-score. This research method is shown in Fig. 2. Near Miss version 2 was not used because it was not successful in obtaining sampling. This is due to the limited computer's memory used.

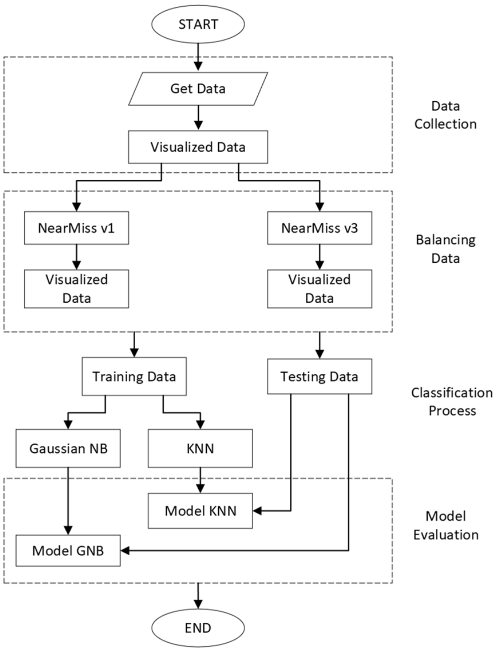

Fig. 2. Research Method

IV. Results And Discussions

After balancing the data, the number of datasets decreases from 663,228 rows to 19,660 rows using the Near Miss version 1 and version 3 algorithms. Indeed, there is a reduction in the total amount of data needed to get balanced data. The visualization of the unbalanced data distribution is shown in Fig. 3. Fig. 4 shows a visualization of balanced data distribution using Near Miss version 1, while Fig. 5 shows a visualization of balanced data distribution using Near Miss version 3.

Balanced data using Near Miss version 3 shows a more uniform distribution of majority and minority data than balanced data using Near Miss version 1.This shows that Near Miss version 3 seems to be better because the sample distribution becomes even.

Gausian Naive Bayes and K-Nearest Neighbors are then used to model the balanced data in order to determine the accuracy, precision, recall, and F1-score. The generated confusion matrix is used as the basis for calculating accuracy, precision, recall, and the F1-score.

TABLE I. Results

	Gaussian Naïve Bayes		K-Nearest Neighbors	
	Near Miss version 1	Near Miss version 3	Near Miss version 1	Near Miss version 3
Accuracy	0.961	0.578	0.938	0.710
Precision	0.996	0.559	1.00	0.723
Recall	0.925	0.705	0.875	0.673
F1 Score	0.959	0.623	0.933	0.697
Process Time	0.012	0.011	0.508	0.469

K-Nearest Neighbors uses optimal k, with value of 140, which is the square root of the amount of data that is already balanced. Table 1 shows the experimental outcomes.

From table 1 can be seen that the Near Miss version 1 tested with Gaussian Naïve Bayes has a better accuracy of 96% compared to that tested with K-Nearest Neighbors which is 93%. Likewise, the F1 Near Miss version 1 score with Gaussian Naïve Bayes is better than K-Nearest Neighbors. Process Time Gaussian Naïve Bayes is also faster, which is 0.012 seconds compared to K-Nearest Neighbors, which is 0.508 seconds. This is in accordance with the opinion of Anand et al that Naïve Bayes has computational efficiency [27].

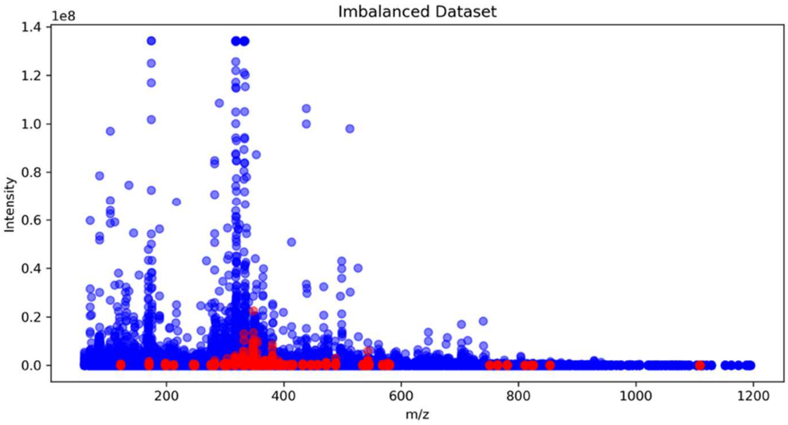

Fig. 3. Graph of unbalanced original data

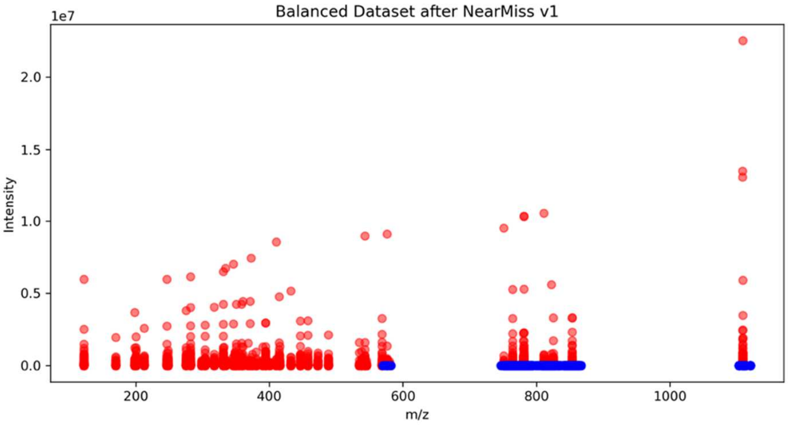

Fig. 4. Graph of balanced data with NearMiss version 1 prior to inclusion in both classification algorithms

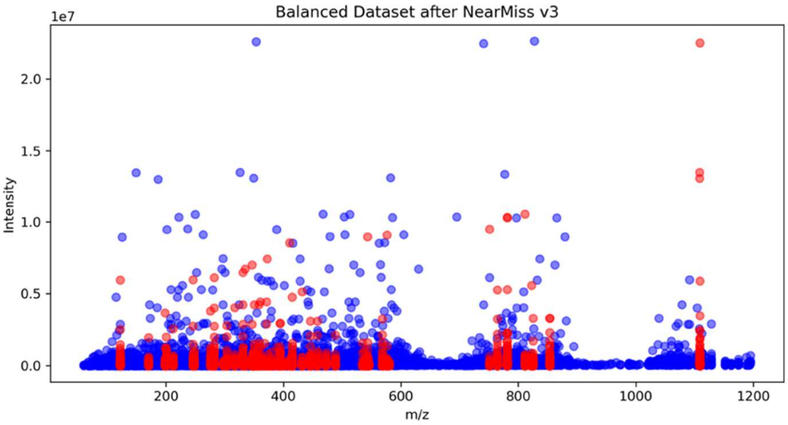

Fig. 5. Graph of balanced data with NearMiss version 3 prior to inclusion in both classification algorithms

V. Conclusions

The LC-MS data of medicinal plants is used in this investigation. It must be balanced because it is not currently balanced. The two methods that were selected are Near Miss versions 1 and 3, with the undersampling technique, which is more resistant to overfitting, being taken into consideration. Near Miss version 2 was not used because it failed to generate sampling. This is due to the limitations of the devices used.

For test the sampling, Gaussian Naive Bayes and K-Nearest Neighbors algorithms are used with consideration of their popularity and ease of use.

The results of this study indicate that Near Miss version 1 tested with Gaussian Naive Bayes is better in terms of accuracy and execution time than tested with K-Nearest Neighbors.

This research raises a new problem with Near Miss version 3, which has a more even distribution but poor accuracy, even though the execution time is faster. We'll be working on this in the future, along with conducting the experiment with different classification systems.

References

[1] N. Rout, D. Mishra, M.K. Mallick, "Handling imbalanced data: A survey," Advances in Intelligent Systems and Computing, **628**, 431–443, 2018, doi:10.1007/978-981-10-5272-9_39.

[2] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, **42**(4), 463–484, 2012, doi:10.1109/TSMCC.2011.2161285.

[3] S. Fotouhi, S. Asadi, M.W. Kattan, "A comprehensive data level analysis for cancer diagnosis on imbalanced data," Journal of Biomedical Informatics, **90**(January), 103089, 2019, doi:10.1016/j.jbi.2018.12.003.

[4] R. Zuech, J. Hancock, T.M. Khoshgoftaar, "Detecting web attacks using

random undersampling and ensemble learners," Journal of Big Data, **8**(1), 2021, doi:10.1186/s40537-021-00460-8.

[5] S. Bagui, K. Li, "Resampling imbalanced data for network intrusion detection datasets," Journal of Big Data, **8**(1), 2021, doi:10.1186/s40537-020-00390-x.

[6] S. Sapre, P. Ahmadi, K. Islam, "A Comprehensive Data Sampling Analysis Applied to the Classification of Rare IoT Network Intrusion Types," in 2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA: 1–2, 2021.

[7] J. Ahmed, R.C. Green II, "Predicting severely imbalanced data disk drive failures with machine learning models," Machine Learning with Applications, **9**(June), 100361, 2022, doi:10.1016/j.mlwa.2022.100361.

[8] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, G. Bing, "Learning from class-imbalanced data: Review of methods and applications," Expert Systems with Applications, **73**, 220–239, 2017, doi:10.1016/j.eswa.2016.12.035.

[9] J.M. Johnson, T.M. Khoshgoftaar, "Survey on deep learning with class imbalance," Journal of Big Data, **6**(1), 2019, doi:10.1186/s40537-019-0192-5.

[10] A. Tanimoto, S. Yamada, T. Takenouchi, M. Sugiyama, H. Kashima, "Improving imbalanced classification using near-miss instances," Expert Systems with Applications, **201**(March), 117130, 2022, doi:10.1016/j.eswa.2022.117130.

[11] G. Rekha, A.K. Tyagi, N. Sreenath, S. Mishra, "Class Imbalanced Data: Open Issues and Future Research Directions," 2021 International Conference on Computer Communication and Informatics, ICCCI 2021, 2021, doi:10.1109/ICCCI50826.2021.9402272.

[12] N. V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, **16**, 321–357, 2002, doi:10.1002/eap.2043.

[13] S. He, H., Bai, Y., Garcia, E., & Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In IEEE International Joint Conference on Neural Networks, 2008," IJCNN 2008.(IEEE World Congress on Computational Intelligence) (Pp. 1322– 1328), (3), 1322–1328, 2008.

[14] J. Zhang, I. Mani, "KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction," in International Conference on Machine Learning (ICML 2003): Workshop on Learning from Imbalanced Datasets II, Washington DC, 2003.

[15] A. Kasem, A.A. Ghaibeh, H. Moriguchi, "Empirical study of sampling methods for classification in imbalanced clinical datasets," Advances in Intelligent Systems and Computing, **532**, 152–162, 2017, doi:10.1007/978-3-319-48517-1_14.

[16] I. Binanto, H.L.H.S. Warnars, N.F. Sianipar, W. Budiharto, "Anticancer Compound Identification Model Of Rodent Tuber's Liquid Chromatography-Mass Spectrometry Data," ICIC Express Letters, **16**(1), 9–16, 2022, doi:10.24507/icicel.16.01.9.

[17] I. Binanto, H.L.H.S. Warnars, N.F. Sianipar, W. Budiharto, "Webscraping Data Labeling System On Liquid Chromatography-Mass Spectrometry Of Rodent Tuber For Efficiency Of Supervised Learning Preprocessing," ICIC Express Letters Part B: Applications, **13**(1), 107–114, 2022, doi:10.24507/icicelb.13.01.107.

[18] D. Laurent, N.F. Sianipar, Chelen, Listiarini, A. Wantho, "Analysis of Genetic Diversity of Indonesia Rodent Tuber (Typhonium flagelliforme Lodd.) Cultivars Based on RAPD Marker)," in The 3rd International Conference on Biological Science 2013 (The 3rd ICBS-2013), 139–145, 2015.

[19] N.F. Sianipar, R. Purnamaningsih, D.L. Gumanti, Rosaria, M. Vidianti, "Analysis Of Gamma Irradiated Fourth Generation Mutant Of Rodent Tuber (Typhonium Flagelliforme Lodd.) Based On Morphology And RAPD Markers," Jurnal Teknologi, **78**(5–6), 41–49, 2016.

[20] N.F. Sianipar, R. Purnamaningsih, Rosaria, "Bioactive compounds of fourth generation gamma-irradiated Typhoniumflagelliforme Lodd . mutants based on gas chromatography-mass spectrometry," in 2nd International Conference on Agricultural and Biological Sciences (ABS 2016), 2016, doi:10.1088/1755-1315/41/1/012025.

[21] N.F. Sianipar, K. Assidqi, R. Purnamaningsih, T. Herlina, "in Vitro Cytotoxic Activity of Rodent Tuber Mutant Plant (Typhonium Flagelliforme Lodd.) Against To Mcf-7 Breast Cancer Cell Line," Asian Journal of Pharmaceutical and Clinical Research, **12**(3), 185–189, 2019, doi:10.22159/ajpcr.2019.version 12i3.29651.

[22] N.F. Sianipar, A. Wantho, Rustikawati, W. Maarisit, "The Effects of Gamma Irradiation on Growth Response of Rodent Tuber (Typhonium flagelliforme Lodd .) Mutant in In Vitro Culture," HAYATI Journal of Biosciences, **20**(2), 51–56, 2013, doi:10.4308/hjb.20.2.51.

[23] M.J. Islam, Q.M.J. Wu, M. Ahmadi, M.A. Sid-Ahmed, "Investigating the Performance of Naive- Bayes Classifiers and K- Nearest Neighbor Classifiers," in International Conference on Convergence Information Technology, 1541–1546, 2007, doi:10.1109/iccit.2007.148.

[24] K. Chandel, V. Kunwar, S. Sabitha, T. Choudhury, S. Mukherjee, "A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques," CSI Transactions on ICT, **4**(2–4), 313–319, 2017, doi:10.1007/s40012-016-0100-5.

[25] M. Irfan, W. Uriawan, O.T. Kurahman, M.A. Ramdhani, I.A. Dahlia, "Comparison of Naive Bayes and K-Nearest Neighbor methods to predict divorce issues," IOP Conference Series: Materials Science and Engineering, **434**(1), 2018, doi:10.1088/1757-899X/434/1/012047.

[26] Y.F. Safri, R. Arifudin, M.A. Muslim, "K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor," Scientific Journal of Informatics, **5**(1), 18, 2018, doi:10.15294/sji.v5i1.12057.

[27] M.V. Anand, B. Kiranbala, S.R. Srividhya, K. C., M. Younus, M.H. Rahman, "Gaussian Naïve Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer," Mobile Information Systems, **2022**, 2022, doi:10.1155/2022/2436946.

[28] K. V Lakshmi, N.S. Kumari, "Survey on Naive Bayes Algorithm," International Jpurnal of Advance Research in Science and Engineering, **07**(03), 240–246, 2018.

[29] Purwanto, D.S.S. Sahid, "Using KNN Algorithms for Determining the Recipient of Smart Indonesia Scholarship Program," Jurnal Komputer Terapan, **7**(2), 163–173, 2021..