JOIV INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION
ISSN 2549-9610 (PRINT)
ISSN 2549-9904 (ONLINE)

#1069 Summary

SUMMARY REVIEW EDITING

Submission

Authors	Hairani Hairani, Anthony Anggrawan, Dadang Priyanto
Title	Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link
Original file	1069-2305-1-SM.DOCX 2022-08-01
Supp. files	None ADD A SUPPLEMENTARY FILE
Submitter	Hairani Hairani
Date submitted	August 1, 2022 - 12:34 PM
Section	Articles
Editor	Alde Alanda

Status

Status	In Editing
Initiated	2022-12-13
Last modified	2022-12-20

Submission Metadata

EDIT METADATA

Authors

Name	Hairani Hairani
ORCID iD	http://orcid.org/0000-0002-6756-5896
Affiliation	Universitas Bumigora
Country	Indonesia
Bio Statement	—

Principal contact for editorial correspondence.

Name	Anthony Anggrawan
Affiliation	Universitas Bumigora
Country	Indonesia
Bio Statement	—
Name	Dadang Priyanto
Affiliation	Universitas Bumigora
Country	Indonesia
Bio Statement	—

Title and Abstract

Title: Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link

Abstract

Most of the health data contained unbalanced data that affected the performance of the classification method. Unbalanced data causes the classification method to more easily classify the majority data and ignore the minority class. One of the health data that has unbalanced data is Pima Indian Diabetes. Diabetes is a deadly disease caused by the body's inability to produce enough insulin. Complications of diabetes can cause heart attacks and strokes. Early diagnosis of diabetes is needed to minimize the occurrence of more severe complications. In the diabetes dataset used, there is an imbalance of data between positive and negative diabetes classes. Diabetes negative class data (500 data) is more than diabetes positive class (268) so that it can affect the performance of the classification method. Therefore, this study aims to apply the Smote-Tomeklink and Random Forest methods in the classification of diabetes. The research methodology used is the collection of diabetes data obtained from Kaggle as many as 768 data with 8 input attributes and 1 output attribute as a class, pre-processing data is used to balance the dataset with Smote-Tomeklink, classification using the random

QUICK MENU

» Editorial Team
» Focus & Scope
» Indexing
» Author Guidelines
» Peer Review Process
» Author Fees
» Publication Ethics
» Online Submission
» Open Access Statement
» Plagiarism Policy
» Special Issues
» Licensing terms
» Contact

forest method, and performance evaluation based on accuracy, sensitivity, precision, and F1-score. Based on the tests carried out by dividing data using 10-fold cross-validation, the Random forest algorithm with Smote-TomekLink gets the highest accuracy, sensitivity, precision, and F1-score compared to Random Forest with Smote. The Random Forest algorithm with Smote-Tomeklink has 86.4% accuracy, 88.2% sensitivity, 82.3% precision, and 85.1% F1-score. Thus, using Smote-Tomeklink can improve the performance of the random forest method based on accuracy, sensitivity, precision, and F1-score.

Indexing

Keywords: Class Imbalance; Smote-Tomeklink;Random Fores Method;Diabetest Disease

Language: en

Supporting Agencies

Agencies: —

References

References

[1] O. Heranova, "Synthetic Minority Oversampling Technique pada Averaged One Dependence Estimators untuk Klasifikasi Credit Scoring," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 3, no. 3, pp. 443–450, 2019, doi: 10.29207/resti.v3i3.1275.

[2] T. Zhu, Y. Lin, and Y. Liu, "Synthetic minority oversampling technique for multiclass imbalance problems," Pattern Recognition, vol. 72, pp. 327–340, Dec. 2017, doi: 10.1016/j.patcog.2017.07.024.

[3] F. Last, G. Douzas, and F. Bacao, "Oversampling for Imbalanced Learning Based on K-Means and SMOTE," no. November, 2017, [Online]. Available: http://arxiv.org/abs/1711.00837.

[4] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, "Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning from Imbalanced Data," IEEE Access, vol. 9, pp. 74763–74777, 2021, doi: 10.1109/ACCESS.2021.3080316.

[5] M. Kamaladevi, V. Venkataraman, and K. R. Sekar, "Tomek link Undersampling with Stacked Ensemble classifier for Imbalanced data classification," vol. 25, no. 4, pp. 2182–2190, 2021.

[6] W. C. Lin, C. F. Tsai, Y. H. Hu, and J. S. Jhang, "Clustering-based undersampling in class-imbalanced data," Information Sciences, vol. 409–410, pp. 17–26, 2017, doi: 10.1016/j.ins.2017.05.008.

[7] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random Forest for medical imbalanced data," Journal of Biomedical Informatics, p. 103465, 2020, doi: 10.1016/j.jbi.2020.103465.

[8] E. AT, A. M, A.-M. F, and S. M, "Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method," Global Journal of Technology and Optimization, vol. 01, no. S1, 2016, doi: 10.4172/2229-8711.s1111.

[9] Z. Wang, C. Wu, K. Zheng, X. Niu, and X. Wang, "SMOTETomek-Based Resampling for Personality Recognition," IEEE Access, vol. 7, pp. 129678–129689, 2019, doi: 10.1109/ACCESS.2019.2940061.

[10] N. V Chawla, K. W. Bowyer, and L. O. Hall, "SMOTE : Synthetic Minority Over-sampling TEchnique," Journal of Artificial Intelligence Research, vol. 16, pp. 341–378, 2002.

[11] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," Jurnal Teknologi dan Sistem Komputer, vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.

[12] I. Tomek, "Tomek Link: Two Modifications of CNN," IEEE Trans. Systems, Man and Cybernetics, pp. 769–772, 1976, [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4309452.

[13] R. Kaur, "Predicting diabetes by adopting classification approach in data mining," International Journal on Informatics Visualization, vol. 3, no. 2–2, pp. 218–221, 2019, doi: 10.30630/joiv.3.2-2.229.

[14] A. Azrar, M. Awais, Y. Ali, and K. Zaheer, "Data mining models comparison for diabetes prediction," International Journal of Advanced Computer Science and Applications, vol. 9, no. 8, pp. 320–323, 2018, doi: 10.14569/ijacsa.2018.090841.

[15] S. Barik, S. Mohanty, S. Mohanty, and D. Singh, "Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques," Smart Innovation, Systems and Technologies, vol. 153, no. January, pp. 399–409, 2021, doi: 10.1007/978-981-15-6202-0_41.

[16] H. Hairani, M. Innuddin, and M. Rahardi, "Accuracy Enhancement of Correlated Naive Bayes Method by Using Correlation Feature Selection (CFS) for Health Data Classification," in 2020 3rd International Conference on Information and Communications Technology (ICOIACT), 2020, pp. 51–55, doi: 10.1109/ICOIACT50329.2020.9332021.

[17] C. Fiarni, E. M. Sipayung, and S. Maemunah, "Analysis and prediction of diabetes complication disease using data mining algorithm," Procedia Computer Science, vol. 161, pp. 449–457, 2019, doi: 10.1016/j.procs.2019.11.144.

[18] Erlin, Y. N. Marlim, Junadhi, L. Suryati, and N. Agustina, "Early Detection of Diabetes Using Machine Learning with Logistic Regression Algorithm," Jurnal Nasional Teknik Elektro dan Teknologi Informasi, vol. 11, no. 2, pp. 88–96, 2022.

[19] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus," Multimedia Systems, vol. 28, no. 4, pp. 1289–1307, 2022, doi: 10.1007/s00530-021-00817-2.

[20] X. Shi, T. Qu, G. Van Pottelbergh, M. van den Akker, and B. De Moor, "A Resampling Method to Improve the Prognostic Model of End-Stage Kidney Disease: A Better Strategy for Imbalanced Data," Frontiers in Medicine, vol. 9, no. March, pp. 1–9, 2022, doi: 10.3389/fmed.2022.730748.

[21] K. Wang et al., "Improving risk identification of adverse outcomes in chronic heart failure using smote +enn and machine learning," Risk Management and Healthcare Policy, vol. 14, no. May, pp. 2453–2463, 2021, doi: 10.2147/RMHP.S310295.

[22] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," ACM Computing Surveys, vol. 52, no. 4. Association for Computing Machinery, pp. 1–34, Aug. 01, 2019, doi: 10.1145/3343440.

[23] K. Guo, X. Wan, L. Liu, Z. Gao, and M. Yang, "Fault diagnosis of intelligent production line based on digital twin and improved random forest," Applied Sciences (Switzerland), vol. 11, no. 16, pp. 1–18, 2021, doi: 10.3390/app11167733.

[24] J. Chen, H. Huang, A. G. Cohn, D. Zhang, and M. Zhou, "Machine learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning," International Journal of Mining Science and Technology, vol. 32, no. 2, pp. 309–322, 2021, doi: 10.1016/j.ijmst.2021.08.004.

[25] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset," Sensors, vol. 22, no. 9, 2022, doi: 10.3390/s22093246.

[26] Y. Sun, H. Zhang, T. Zhao, Z. Zou, B. Shen, and L. Yang, "A New Convolutional Neural Network with

» Added to review: **May 29, 2020**

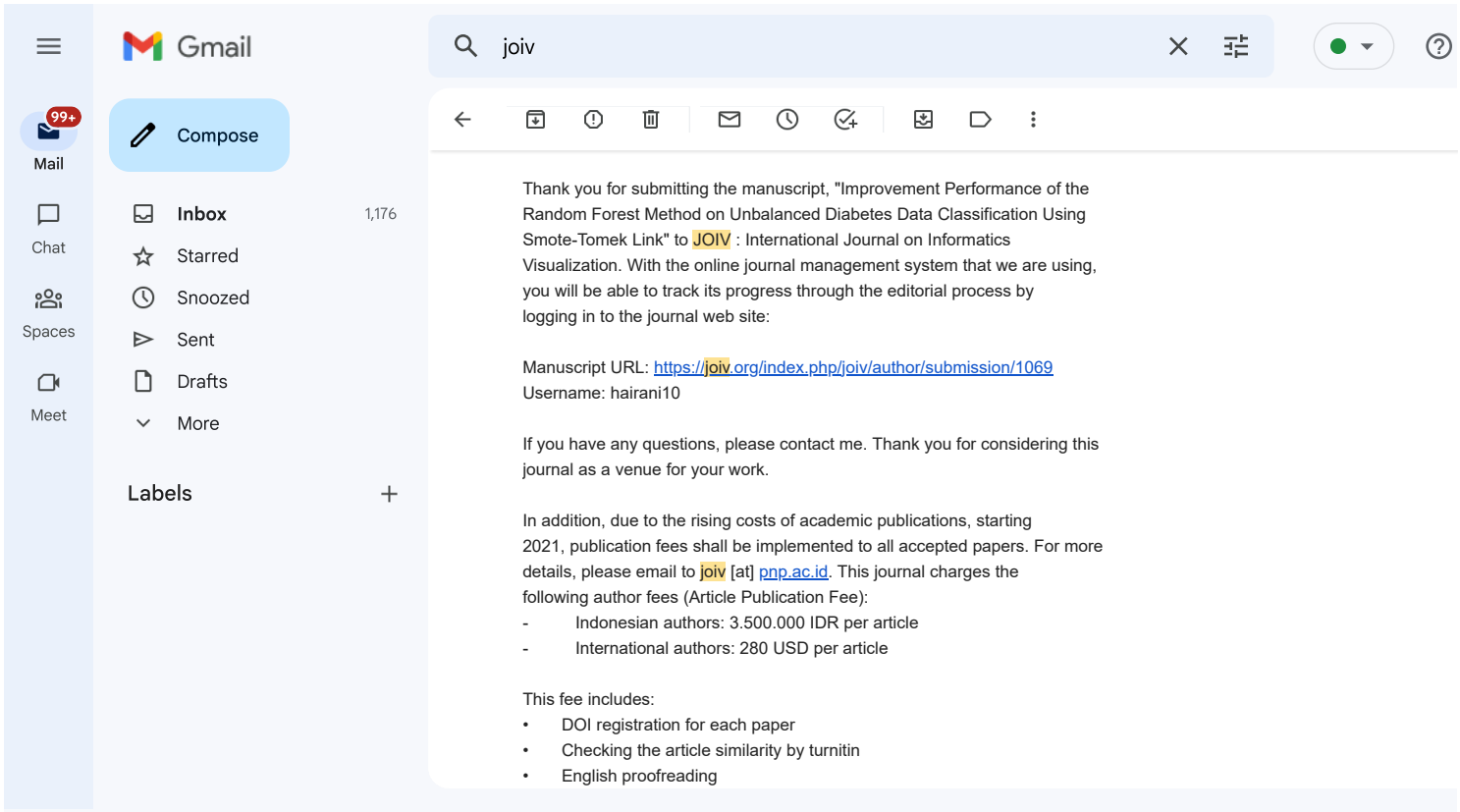
» **Ei COMPENDEX**

» Submission: **February 10, 2021**

[ ] , g, , , , g,
Random Forest Method for Hydrogen Sensor Fault Diagnosis," IEEE Access, vol. 8, pp. 85421–85430, 2020, doi: 10.1109/ACCESS.2020.2992231.
[27] H. Hartono and E. Ongko, "Avoiding Overfitting dan Overlapping in Handling Class Imbalanced Using Hybrid Approach with Smoothed Bootstrap Resampling and Feature Selection," International Journal on Informatics Visualization, vol. 6, no. June, pp. 343–348, 2022.
[28] H. Hairani, A. Anggrawan, A. I. Wathan, K. A. Latif, K. Marzuki, and M. Zulfikri, "The Abstract of Thesis Classifier by Using Naive Bayes Method," in Proceedings - 2021 International Conference on Software Engineering and Computer Systems and 4th International Conference on Computational Science and Information Management, ICSECS-ICOCSIM 2021, 2021, no. August, pp. 312–315, doi: 10.1109/ICSECS52883.2021.00063.
[29] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," Pattern Recognition, vol. 91, pp. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.
[30] H. Qteat and M. Awad, "Using Hybrid Model of Particle Swarm Optimization and Multi-Layer Perceptron Neural Networks for Classification of Diabetes," International Journal of Intelligent Engineering and Systems, vol. 14, no. 3, pp. 11–22, 2021, doi: 10.22266/ijies2021.0630.02.
[31] H. Hanafi, A. H. Muhammad, I. Verawati, and R. Hardi, "An Intrusion Detection System Using SDAE to Enhance Dimensional Reduction in Machine Learning," International Journal on Informatics Visualization, vol. 6, no. June, pp. 306–316, 2022.
[32] H. Hairani, A. S. Suweleh, and D. Susilowaty, "Penanganan Ketidak Seimbangan Kelas Menggunakan Pendekatan Level Data," MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 20, no. 1, pp. 109–116, 2020, doi: 10.30812/matrik.v20i1.846.
[33] M. Y. Thanoun, M. T. Yaseen, and A. M. Aleesa, "Development of Intelligent Parkinson Disease Detection System Based on Machine Learning Techniques Using Speech Signal," International Journal on Advanced Science, Engineering and Information Technology, vol. 11, no. 1, pp. 388–392, 2021.

JOIV : International Journal on Informatics Visualization
ISSN 2549-9610 **(print) |** 2549-9904 **(online)**
Organized by Department of Information Technology - Politeknik Negeri Padang, and Institute of Visual Informatics - UKM and Soft Computing and Data Mining Centre - UTHM
W : http://joiv.org
E : joiv@pnp.ac.id, hidra@pnp.ac.id, rahmat@pnp.ac.id

View JOIV Stats



Thank you for submitting the manuscript, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link" to JOIV : International Journal on Informatics Visualization. With the online journal management system that we are using, you will be able to track its progress through the editorial process by logging in to the journal web site:

Manuscript URL: https://joiv.org/index.php/joiv/author/submission/1069
Username: hairani10

If you have any questions, please contact me. Thank you for considering this journal as a venue for your work.

In addition, due to the rising costs of academic publications, starting 2021, publication fees shall be implemented to all accepted papers. For more details, please email to joiv [at] pnp.ac.id. This journal charges the following author fees (Article Publication Fee):

- Indonesian authors: 3.500.000 IDR per article
- International authors: 280 USD per article

This fee includes:

- DOI registration for each paper
- Checking the article similarity by turnitin
- English proofreading