JURNAL TIMES
TECHNOLOGY INFORMATICS & COMPUTER SYSTEM

Volume IX No 1, Juni 2020
*p*ISSN : 2337 – 3601
*e*ISSN : 2549 – 015X

Tersedia *online* di http://ejournal.stmik-time.ac.id

PENDEKATAN LEVEL DATA UNTUK MENANGANI KETIDAKSEIMBANGAN DATA MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR

Resianta Perangin-angin[1], Eva Julia Gunawati Harianja[2] Indra Kelana Jaya[3]
Progam Studi Komputerisasi Akuntansi, Teknik Informatika
Universitas Methodist Indonesia
Jl. Hang Tuah No.8 Medan
Email: [1]resianta88@gmail.com
[2]graziedamanik@gmail.com
[3]indrakj_sagala@yahoo.com

Abstrak

Dalam penelitian ini digunakan dataset yang memiliki tingkat ketidakseimbangan yang berbeda beda mulai dari 16.40, 8.60, 2.06, 2.78, 1.87, tentu hal ini dapat menurunkan kinerja algoritma klasifikasi. Secara umum ketidakseimbangan kelas dapat ditangani dengan dua pendekatan, yaitu level data dan level algoritma. Pendekatan level data ditujukan untuk memperbaiki keseimbangan kelas, sedangkan pendekatan level algoritma ditujukan untuk memperbaiki algoritma atau menggabungkan (ensemble) pengklasifikasi agar lebih konduktif terhadap kelas minoritas. Pada penelitian ini diusulkan pendekatan level data dengan resampling, yaitu random oversampling (ROS), dan random undersampling (RUS), Pengklasifikasi yang digunakan adalah k-near neighbors. Hasil penelitian menunjukkan bahwa model ROS+KNN dan RUS+KNN didapat dengan selisih G-Means sebesar 13% dan F-Measure 2,08%, dari, hal ini menunjutkan bahwa RUS+KNN dan ROS+KNN bisa meningkatkan akurasi dari G-Mean dan F-Measure namun tidak memiliki perbedaan yang signifikan.

Kata kunci : ***Ketidakseimbangan Kelas, Random Over Sampling, Random Under Sampling, k-Nearest Neighbor.***

1. Pendahuluan

Ketidakseimbangan kelas telah menjadi masalah berkelanjutan di bidang Machine Learning dan Klasifikasi, dimana tidak seimbang merupakan suatu keadaan dimana distribusi kelas data tidak seimbang, jumlah kelas data (instance) yang satu lebih sedikit atau lebih banyak dibanding dengan jumlah kelas data.(Ali et al., n.d.; Cordón et al., 2018; Juan Carbajal-Hernández et al., 2016; J. Lee et al., 2015; J. Sun et al., 2018, 2020; Y. Sun et al., 2009). Kelompok kelas data yang lebih sedikit dikenal dengan kelompok minoritas (minority), kelompok kelas data yang lainnya disebut dengan kelompok mayoritas (mayority).(Ali et al., n.d.; Bolón-Canedo et al., 2014; Cordón et al., 2018; Douzas et al., 2018; Juan Carbajal-Hernández et al., 2016; Rout et al., 2018; J. Sun et al., 2018). Ketika kita menambang data dari datase langsung data tersebut sudah bisa dipastikan adalah data tidak seimbang, dimana kondisi tersebut tentu akan menyulitkan metode klasifikasi apapun, oleh sebab itu perlu dilakukan penyeimbangan data terlebih dahulu.(Bolón-Canedo et al., 2014; Huang & Fitzmaurice, 2005; Juan Carbajal-Hernández et al. - 2016—Classification of unbalance and misalignment in in.pdf, n.d.; Khalilpour Darzi et al., 2019; Searle, 1994; Y. Sun et al., 2009). Banyak metode klasifikasi yang umumnya di gunakan dalam kasus-kasus klasifikasi seperti, Naïve Bayes, K Neares Neighbor, Adaboost support vector machine, namun umumnya semua algoritma tersebut memiliki performa yang buruk pada saat bekerja di data tidak seimbang Dikarenakan algoritma-algoritma yang disebutkan tersebut belum memiliki kemampuan untuk menangani masalah ketidakseimbangan data. (Chung et al., 2011; Duan et al., 2016; Farquad & Bose, 2012; Verbiest et al., 2014)

Klasifikasi pada data dengan kelas tidak seimbang merupakan fokus masalah dalam bidang machie learning dan data mining, misalnya dalam masalah medis [19], [20], masalah finansial(J. Sun et al., 2020), masalah geometris(Douzas & Bacao, 2019), masalah klasfikasi teks(C.-Y. Lee & Lee, 2012; Sundarkumar & Ravi, 2015; Wu et al., 2014), semua algoritma yang disebutkan tadi jika berkerja pada situasi data tidak seimbang maka akan memberikan performa yang buruk seperti misaslnya, jika sealgoritma tersebut bekerja pada data yang tidak seimbang, hampir semua algoritma klasifikasi akan menghasilkan akurasi yang jauh lebih tinggi untuk kelas mayoritas daripada kelas minoritas(Maldonado et al., 2019; Qiong, 2016; Raghuwanshi & Shukla, 2019). Perbedaan ini merupakan suatu indikator yang menjadikan performa klasifikasi yang buruk pada data tidak seimbang. Pada beberapa kasus, kelas minoritas justru lebih penting untuk diidentifikasi daripada kelas mayoritas itu sendiri.

Misalnya pada kasus transaksi dengan kartu kredit, kebanyakan status transaksi adalah transaksi yang normal, hanya sedikit kasus yang dapat ditemukan dimana terjadi transaksi yang tidak normal atau fraud. Meskipun demikian, keberadaan transaksi yang yang tidak normal ini jauh lebih penting untuk di identifikasi daripada transaksi yang normal meskipun jumlah kasusnya jauh lebih sedikit dikarenakan akan lebih mudah dalam menganalisis penyebab dari transaksi yang tidak normal tersebut untuk keamanan selanjutnya(Qiong, 2016)

Banyak algoritma metode yang bisa dilakukan dalam hal pengangan ketidakseimbanga kelas dalam data tidak seimbang, dalam penelitian ini akan digunakan metode Synthetic Minority Over-sampling Technique (SMOTE), metode ini

sangat sering atau populer digunakan dalam menangani masalah ketidakseimbangan data.(Douzas & Bacao, 2019; Prusty et al., 2017; Raghuwanshi & Shukla, 2019; Wang et al., 2014; Zhang et al., 2016). Metode ini mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan dataset dengan cara sampling ulang sampel kelas minoritas. Telah banyak dilakukan penelitian terkait dengan metode SMOTE salah satunya Penerapan SMOTE pada masalah klasifikasi pada masalah prediksi Tipe Ion Channel-targeted.(Zhang et al., 2016). Dan pada masalah clustering dengan metode k-Means Clustering, Penerapan SMOTE berhasil menghilangkan noise serta menyelesaikan masalah ketidak seimbangan pada 71 dataset dengan penerapan SMOTE dapat memperbaiki kualitas clustering (A Hybrid Approach from Ant Colony Optimization ..., n.d.) dan pada penelitian akan menggunakan algoritma Random Over Sampling (ROS) dan Random Under Sampling (RUS) dengan algoritma klasifikasi k-nn

Synthetic Minority Over-sampling Technique

Synthetic Minority Oversampling Technique (SMOTE) merupakan metode turunan dari oversampling. SMOTE pertama kali diperkenalkan oleh Nithes V. Chawla(Chawla et al., 2002). Metode ini berkerja dengan membuat replikasi dari data minoritas. Replikasi tersebut dikenal dengan data sintetis (syntetic data). Metode SMOTE bekerja dengan mencari k nearest neighbors (yaitu ketetanggaan terdekat data sebanyak k) untuk setiap data di kelas minoritas, setelah itu dibuat data sintetis sebanyak prosentase duplikasi yang diinginkan antara data minor dan knearest neighbors yang dipilih secara acak.

k-Nearest Neighbor

Metode k-Nearest Neighbor (k-NN) merupakan metode klasifikasi klasik yang paling sederhana. Metode k-NN sering juga disebut dengan InstanceBased Learning, k-NN melakukan klasifikasi terhadap objek berdasarkan jarak antara objek tersebut dengan objek lain.[8] Metode k-NN menggunakan prinsip ketetanggaan (neighbor) untuk memprediksi kelas yang baru. Jumlahtetangga yang dipakai adalah sebanyak k tetangga.

2. Methodology

2.1 Pengukuran Performa

Dalam Mengukur Performa sebuah algoritma dalam hal ini algoritma klasifikasi maka umumnya sebagai acuan adalah tingkat akurasi nya, namun untuk ketidakseimbangan data Metode pengukuran performa menggunakan Confusion matrix dimana confusion matrix merupakan alat yang paling populer dalam mengevaluasi performa klasifikasi. ada tabel berikut ditampilkan confusion matrix untuk kelas biner, yaitu dataset dengan dua jenis kelas saja.

Tabel 1. Confusion Matrix

Class	Predictive Positive	Predictive Negative
Actual Positive	TP	FP
Actual Negative	TN	FN

True Positive (TP) dan True Negative (TN) merupakan jumlah kelas positif dan negatif yang diklasifikasikan dengan tepat, False Positive (FP) dan False Negative (FN) merupakan jumlah kelas positif dan negatif yang tidak diklasifikasikan dengan tepat. Berdasarkan confusion matrix tersebut dapat ditentukan kriteria performa seperti Accuracy, Precision, Recall, specificity, FMeasure, G-Mean dan yang lainnya. Akurasi (accuracy) merupakan kriteria yang paling umum untuk mengukur kinerja klasifikasi, tetapi jika bekerja pada kelas tidak seimbang, kriteria ini kurang tepat karena kelas minoritas akan memiliki sumbangsih yang kecil pada kriteria accuracy. Kriteria Penilaian yang disarankan adalah TPrate, FPvalue, F-Measure dan G-Mean F-Measure digunakan untuk mengukur klasifikasi kelas minoritas pada kelas tidak seimbang, dan indeks G-mean digunakan untuk mengukur performa keseluruhan (overall classification performance). Pada penelitian ini, performa klasifikasi menggunakan F-Measure dan G-Mean.

Recal = TPrate = ……..…….(1)
Precision = PPvalue = …….(2)
Specificity = TNrate = …….(3)
G - Mean = ……(4)

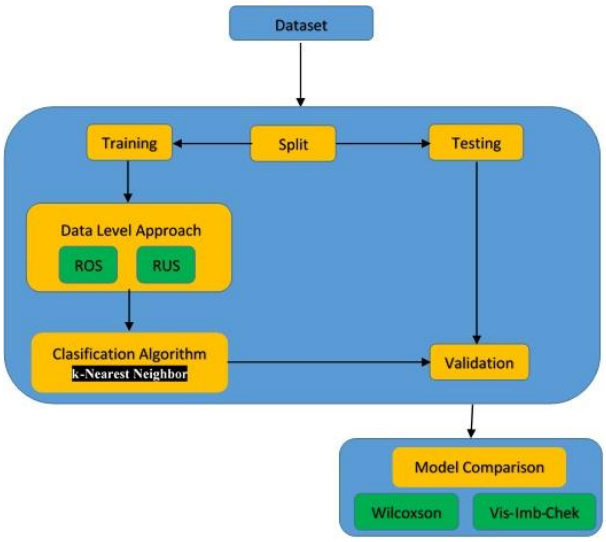
2.2 Kerangka Yang Diusulkan

Penelitian ini dilakukan dengan mengusulkan model, mengembangkan aplikasi untuk mengimplementasikan model yang diusulkan, dengan menggunakan keel dataset “https://sci2s.ugr.es/keel/datasets.php” dan mengukur kinerja dari algoritma klasifikasi tersebut. Untuk menangani masalah ketidakseimbangan kelas pada dataset diusulkan model menggunakan pendekatan level data yang dilakukan dengan resampling, dan mensintesis data latih. Algoritma resampling yang digunakan adalah random oversampling (ROS), dan random undersampling (RUS). Algoritma pengklasifikasi yang digunakan adalah K-NN. Validasi pada pengukuran kinerja digunakan 5-fold cross validation. Hasil pengukuran dianalisa menggunakan wilcoxon. Kerangka kerja model yang diusulkan ditunjukkan pada Gambar 2.

JURNAL TIMES
TECHNOLOGY INFORMATICS & COMPUTER SYSTEM

Volume IX No 1, Juni 2020
*p*ISSN : 2337 – 3601
*e*ISSN : 2549 – 015X

Tersedia *online* di http://ejournal.stmik-time.ac.id



Gambar 2. Kerangka Yang Diusulkan

3. Hasil Dan Pembahasan

3.1 Dataset

Penelitian ini menggunakan 5 jenis dataset yang bersumber dari UCI repository. Dataset ini terdiri Abalone, Ecoli, Glass, Haberman, dan Pima, dengan tingkat Imbalance Ratio (IR) masing masing yakni 16.40, 8.60, 2.06, 2.78, 1.87, attribut dari dataset dapat dilihat pada tabel 3.1

Tabel 3.1 Attribut Dataset dan IR

No.	Dataset	Attribut	IR
1.	Abalone	Sex, Length, Diameter, High, Whole_wight Shucked_weight Viscera_weight Shell_weight	16,40
2.	Ecoli	Mcg, Gvh, Lip, Chg, Aac, Alm1, Alm2	8,60
3.	Glass	Na, Mg, Al, Si, K, Ca, Ba, Fe	2,06
4.	Haberman	Age, Year, Posotive	2,78
5	Pima	Preg,Plas, Pres,Skin, Insu,Mass, Pedi,Age	1,87

3.2 Prosedur ROS + KNN dan RUS + KNN

Untuk prosedur dari algoritma ada beberapa tahapan yang untuk melihat performansi dari algoritma, beberapa tahapan tersebut yakni:

1. partisi dataset secara acak menjadi 5 bagian dengan skema 5-fold cross validation

2. Menerapkan penanganan kelas data tidak seimbang pada ROS dan RUS sebanyak 2 (dua) kali pada data latih :
   a. Menentukan nilai tetangga dengan k = 5
   b. Menghitung jarak antar data kelas minoritas dengan metode eucledian
   c. Melakukan perhitungan untuk membangkitkan data buatan (syntetic)

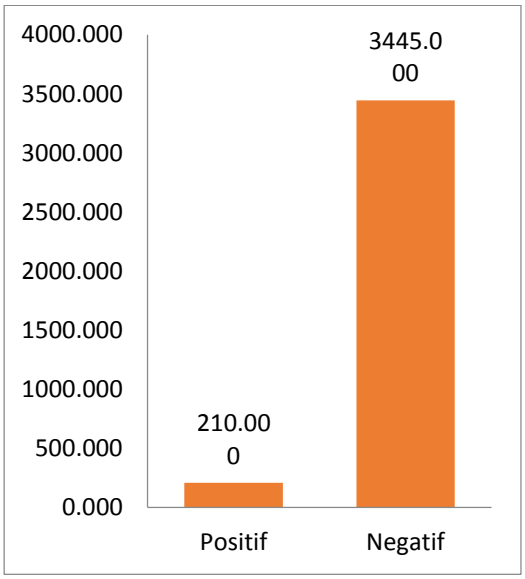
3. Menerapkan k-nearest neighbor untuk mengklasifikasi data uji :
   a. Menentukan nilai tetangga dengan k = 1
   b. Menghitung jarak antar data kelas minoritas dengan metode eucledian

c. Membandingkan kinerja klasifikasi dengan diterapkannya ROS dan RUS kinerja klasifikasi yang diterapkan adalah G-Mean dan F-Mean.

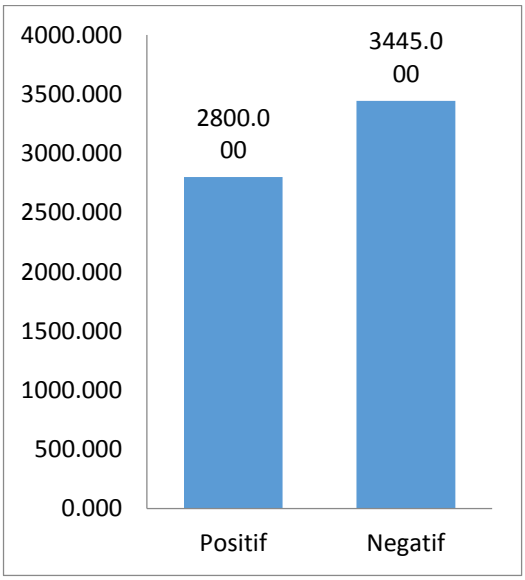
3.3 Hasil Pengujian

Penerapan ROS + KNN meminimalisasi ketidak seimbangan kelas pada semua dataset yang diuji pengujian pertama dilakukan menggunakan dataset abalone, terlihat hasil dari pengujian dilihat pada tabel 3.2

Tabel 3.2 Data Awal



Tabel 3.3 Hasil Pengujian ROS+KNN



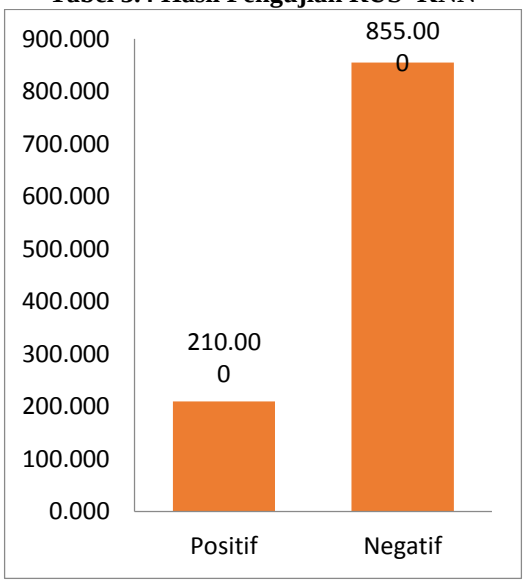
Terlihat kenaikan IR yang sangat signifikan setelah dilakukan pengujian terhadap dataset Abalone dengan menerapkan algoritma ROS+KNN

JURNAL TIMES
TECHNOLOGY INFORMATICS & COMPUTER SYSTEM

Volume IX No 1, Juni 2020
*p*ISSN : 2337 – 3601
*e*ISSN : 2549 – 015X

Tersedia *online* di http://ejournal.stmik-time.ac.id

Tabel 3.4 Hasil Pengujian RUS+KNN



Dari hasil pengujian RUS+KNN juga didapat hasil pengingkatan IR dari dataset abalone yang di ujikan, dari sini didapat sebuah hipotesa bahwasanya dengan menambahkan algoritma RUS ataupun ROS dapat meningkatkan perbadingan *Imbalacing Ratio* dari dataset yang diuji

3.4 Performa Algoritma

Selanjutnya akan cari performa dari algoritma yang akan diujikan kedalam 5 dataset yang telah disiapkan, untuk pengujian performa ini akan dilihat nilai dari G-Mean F-Measure nya pada masing-masing algoritma, untuk lebih jelasnya dapat dilihat pada tabel 3.5

Tabel 3.5 Pengujian Dataset Abalon dengan IR 16.40

Performa	**ROS+KNN**	**RUS+KNN**
TP	9	27
TN	674	464
FP	33	15
FN	15	225
Acc	0.93	0.67
Recal	0.37	0.11
Speci	0.95	0.97
Prec	0.21	0.64
GM	**0.59**	**0.32**
F-M	**0.27**	**0.18**

Dari tabel diatas dilihat performa dari Ros sedikit lebih tinggi daripada performa Rus, dan pengujian akan dilanjutkan terus terhadap 5 dataset yang telah disiapkan

Tabel. 3.6 Pengujian Dataset Ecoli dengan IR 8.60

Performa	**ROS+KNN**	**RUS+KNN**
TP	18	30
TN	284	253
FP	17	5
FN	17	48
Acc	0.90	0.84
Recal	0.51	0.39
Speci	0.94	0.98
Prec	0.51	0.86
GM	**0.69**	**0.62**
F-M	**0.51**	**0.54**

Tabel 3.7 Pengujian Dataset Glass dengan IR 2.06

Performa	**ROS+KNN**	**RUS+KNN**
TP	53	58
TN	123	108
FP	17	12
FN	21	36
Acc	0.82	0.78
Recal	0.72	0.62
Speci	0.88	0.90
Prec	0.76	0.83
GM	**0.80**	**0.75**
F-M	**0.74**	**0.71**

Tabel 3.8 Pengujian Dataset Haberman dengan IR 2.78

JURNAL TIMES
TECHNOLOGY INFORMATICS & COMPUTER SYSTEM

Volume IX No 1, Juni 2020
*p*ISSN : 2337 – 3601
*e*ISSN : 2549 – 015X

Tersedia *online* di http://ejournal.stmik-time.ac.id

Performa	**ROS+KNN**	**RUS+KNN**
TP	28	43
TN	178	127
FP	53	38
FN	47	98
Acc	0.67	0.56
Recal	0.37	0.31
Speci	0.77	0.77
Prec	0.35	0.53
GM	0.54	0.49
F-M	0.36	0.39

Tabel 3.9 Pengujian Dataset Pima dengan IR 1.87

Performa	**ROS+KNN**	**RUS+KNN**
TP	144	175
TN	396	347
FP	124	93
FN	104	153
Acc	0.70	0.68
Recal	0.58	0.53
Speci	0.76	0.79
Prec	0.54	0.65
GM	0.67	0.65
F-M	0.56	0.59

Dari semua hasil pegujian dilakukan, maka akan di akumulasi untuk melihat rerata performa yang dimiliki setiap algoritma yang terhadap 5 dataset yang telah di ujikan, untuk lebih jelasya dapat dilihat pada tabel 3.8

Tabel. 3.8 Hasil Performa Algoritma

Performa	**ROS+KNN**	**RUS+KNN**
TP	50	67
TN	331	260

FP	49	33
FN	41	112
Acc	0.81	0.71
Recal	0.51	0.39
Speci	0.86	0.88
Prec	0.47	0.70
GM	**0.66**	**0.57**
F-M	**0.49**	**0.48**

Dari hasil semua pengujian terlihat Ros+KNN dapat meningkatkan *Accuracy* Sebesar 0.09 atau 15.79% untuk performa dari G-Mean dan 0,01 untuk F-Measure 2.08%. Untuk mengetahui apakah nilai G-Mean dan FMeasure pada k-NN berbeda secara signifikan dengan performa k-NN+SMOTE, maka pengujian dilakukan dengan metode Wilcoxon Sing Rank Test dengan taraf Hasil pengujian ditampilkan pada tabel berikut.

4. Kesimpulan

Penelitian ini berfokus pada pendekatan level data untuk mengurangi pengaruh ketidakseimbangan kelas menggunakan dua algoritma resampling, yaitu random oversampling (ROS) dan random undersampling (RUS), dan satu algoritma klasifikasi, yaitu KNN yang diujicobakan terhadap 5 dataset yang memiliki tingkat ketidakseimbangan yang berbeda-beda. Dataset baru yang dihasilkan dari masing-masing algoritma tersebut digunakan untuk melatih pengklasifikasi KNN Kinerja yang dihasilkan diukur dan dilakukan uji statistik. Uji statistik dilakukan dengan uji Wilcoxon untuk mengetahui signifikansi perbedaan antarmodel. Pada pengukuran yang memiliki perbedaan tidak begitu signifikan, meliputi perbandingan berpasangan, menghitung p-value, dan membuat tabel signifikansi, serta dibuatkan diagram, hasil dari ujicoba tersebut mendapatkan kesimbupan Algoritma ROS dan RUS dapat mengatasi tingkat ketidakseimbangan data menjadi lebih baik, untuk kedua algoritma tersebut dari ujicoba yang dilakukan didapat bahwa algoritma ROS memiliki kemampuan yang lebih baik untuk menyeimbangkan data dibandingkan dengan algoritma RUS dalam 5 dataset yang di ujicobakan.

Datar pustaka

A Hybrid Approach from Ant Colony Optimization ... (n.d.). 13.

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (n.d.). Classification with class imbalance problem: A review. 31.

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2014). Data classification using an ensemble of filters. Neurocomputing, 135, 13–20. https://doi.org/10.1016/j.neucom.2013.03.067

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Chung, H.-Y., Ho, C.-H., & Hsu, C.-C. (2011). Support vector machines using Bayesian-based approach in the issue of unbalanced classifications. Expert Systems with Applications, 38(9), 11447–11452. https://doi.org/10.1016/j.eswa.2011.03.018

Cordón, I., García, S., Fernández, A., & Herrera, F. (2018). Imbalance: Oversampling algorithms for imbalanced classification in R. Knowledge-Based Systems, 161, 329–341. https://doi.org/10.1016/j.knosys.2018.07.035

Department of Biological Sciences, BITS PILANI K K Birla Goa Campus, Zuarinagar, Vasco Da Gama, India, & Kothandan, R. (2015). Handling class imbalance problem in miRNA dataset associated with cancer. Bioinformation, 11(1), 6–10. https://doi.org/10.6026/97320630011006

Douzas, G., & Bacao, F. (2019). Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE. Information Sciences, 501, 118–135. https://doi.org/10.1016/j.ins.2019.06.007

JURNAL TIMES
TECHNOLOGY INFORMATICS & COMPUTER SYSTEM

Volume IX No 1, Juni 2020
*p*ISSN : 2337 – 3601
*e*ISSN : 2549 – 015X

Tersedia *online* di http://ejournal.stmik-time.ac.id

Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Information Sciences, 465, 1–20. https://doi.org/10.1016/j.ins.2018.06.056

Duan, L., Xie, M., Bai, T., & Wang, J. (2016). A new support vector data description method for machinery fault diagnosis with unbalanced datasets. Expert Systems with Applications, 64, 239–246. https://doi.org/10.1016/j.eswa.2016.07.039

Farquad, M. A. H., & Bose, I. (2012). Preprocessing unbalanced data using support vector machine. Decision Support Systems, 53(1), 226–233. https://doi.org/10.1016/j.dss.2012.01.016

Han, W., Huang, Z., Li, S., & Jia, Y. (2019). Distribution-Sensitive Unbalanced Data Oversampling Method for Medical Diagnosis. Journal of Medical Systems, 43(2). https://doi.org/10.1007/s10916-018-1154-8

Huang, W., & Fitzmaurice, G. M. (2005). Analysis of longitudinal data unbalanced over time. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(1), 135–155. https://doi.org/10.1111/j.1467-9868.2005.00492.x

Juan Carbajal-Hernández et al. - 2016—Classification of unbalance and misalignment in in.pdf. (n.d.).

Juan Carbajal-Hernández, J., Sánchez-Fernández, L. P., Hernández-Bautista, I., Medel-Juárez, J. de J., & Sánchez-Pérez, L. A. (2016). Classification of unbalance and misalignment in induction motors using orbital analysis and associative memories. Neurocomputing, 175, 838–850. https://doi.org/10.1016/j.neucom.2015.06.094

Khalilpour Darzi, M. R., Niaki, S. T. A., & Khedmati, M. (2019). Binary classification of imbalanced datasets: The case of CoIL challenge 2000. Expert Systems with Applications, 128, 169–186. https://doi.org/10.1016/j.eswa.2019.03.024

Lee, C.-Y., & Lee, Z.-J. (2012). A novel algorithm applied to classify unbalanced data. Applied Soft Computing, 12(8), 2481–2485. https://doi.org/10.1016/j.asoc.2012.03.051

Lee, J., Wu, Y., & Kim, H. (2015). Unbalanced data classification using support vector machines with active learning on scleroderma lung disease patterns. Journal of Applied Statistics, 42(3), 676–689. https://doi.org/10.1080/02664763.2014.978270

Maldonado, S., López, J., & Vairetti, C. (2019). An alternative SMOTE oversampling strategy for high-dimensional datasets. Applied Soft Computing, 76, 380–389. https://doi.org/10.1016/j.asoc.2018.12.024

Prusty, M. R., Jayanthi, T., & Velusamy, K. (2017). Weighted-SMOTE: A modification to SMOTE for event classification in sodium cooled fast reactors. Progress in Nuclear Energy, 100, 355–364. https://doi.org/10.1016/j.pnucene.2017.07.015

Qiong, G. (2016). An Improved SMOTE Algorithm Based on Genetic Algorithm for Imbalanced. 14(2), 12.

Raghuwanshi, B. S., & Shukla, S. (2019). SMOTE based class-specific extreme learning machine for imbalanced learning. Knowledge-Based Systems. https://doi.org/10.1016/j.knosys.2019.06.022

Rout, N., Mishra, D., & Mallick, M. K. (2018). Handling Imbalanced Data: A Survey. In M. S. Reddy, K. Viswanath, & S. P. K.M. (Eds.), International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications (Vol. 628, pp. 431–443). Springer Singapore. https://doi.org/10.1007/978-981-10-5272-9_39

Searle, S. R. (1994). Analysis of Variance Computing Package Output for Unbalanced Data from Fixed-Effects Models with Nested Factors. The American Statistician, 48(2), 148. https://doi.org/10.2307/2684275

Sun, J., Lang, J., Fujita, H., & Li, H. (2018). Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. Information Sciences, 425, 76–91. https://doi.org/10.1016/j.ins.2017.10.017

Sun, J., Li, H., Fujita, H., Fu, B., & Ai, W. (2020). Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. Information Fusion, 54, 128–144. https://doi.org/10.1016/j.inffus.2019.07.006

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). CLASSIFICATION OF IMBALANCED DATA: A REVIEW. International Journal of Pattern Recognition and Artificial Intelligence, 23(04), 687–719. https://doi.org/10.1142/S0218001409007326

Sundarkumar, G. G., & Ravi, V. (2015). A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. Engineering Applications of Artificial Intelligence, 37, 368–377. https://doi.org/10.1016/j.engappai.2014.09.019

Verbiest, N., Ramentol, E., Cornelis, C., & Herrera, F. (2014). Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection. Applied Soft Computing, 22, 511–517. https://doi.org/10.1016/j.asoc.2014.05.023

Wang, K.-J., Makond, B., Chen, K.-H., & Wang, K.-M. (2014). A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients. Applied Soft Computing, 20, 15–24. https://doi.org/10.1016/j.asoc.2013.09.014

Wu, Q., Ye, Y., Zhang, H., Ng, M. K., & Ho, S.-S. (2014). ForesTexter: An efficient random forest algorithm for imbalanced text categorization. Knowledge-Based Systems, 67, 105–116. https://doi.org/10.1016/j.knosys.2014.06.004

Zhang, L., Zhang, C., Gao, R., Yang, R., & Song, Q. (2016). Using the SMOTE technique and hybrid features to predict the types of ion channel-targeted conotoxins. Journal of Theoretical Biology, 403, 75–84. https://doi.org/10.1016/j.jtbi.2016.04.034