PAPER NAME

4 Jurnal Nasional Hartono.pdf

AUTHOR

Hartono

WORD COUNT

1707 Words

CHARACTER COUNT

10550 Characters

PAGE COUNT

5 Pages

FILE SIZE

1.3MB

SUBMISSION DATE

Feb 24, 2024 1:11 PM GMT+7

REPORT DATE

Feb 24, 2024 1:11 PM GMT+7

15% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 10% Publications database
- Crossref database
- Crossref Posted Content database
- 10% Submitted Works database

Excluded from Similarity Report

- Internet database
- Bibliographic material

JURNAL ILMIAH CORE IT Vol. 11 No. 4 e-ISSN: 2548-3528 p-ISSN: 2339-1766

Pengaruh Penentuan Natural Neighbors pada Hybrid Approach Di Dalam Penanganan Class Imbalance

Hartono

Program Studi Magister Ilmu Komputer, Universitas Potensi Utama

e-Mail: hartonoibbi@gmail.com

Abstrak

Permasalahan *class imbalance* merupakan salah satu permasalahan yang menarik di dalam *machine learning*. Untuk penanganan *class imbalance* sejumlah metode telah dikemukakan. Salah satunya adalah *Hybrid Approach* yang secara umum menggunakan penerapan *oversampling* pada *minority class*. Salah satu metode *oversampling* yang banyak digunakan adalah *Synthetic Minority Oversampling Technique* (SMOTE). Secara umum metode *oversampling* bekerja berdasarkan pada penentuan parameter *k* dan jumlah tetangga dari tiap *sample*. Salah satu metode yang cukup efektif di dalam penentuan nilai *k* secara adaptif adalah Natural Neighbors SMOTE (NaNSMOTE). Melalui penerapan NanSMOTE selain nilai *k* yang fleksibel juga *class center* dari tiap *sample* mempunyai jumlah *neighbors* yang lebih banyak dibandingkan dengan *border samples* sehingga dapat mengurangi kesalahan pada pembangkitan *sample*. Hasil penelitian menunjukkan bahwa penentuan Natural Neighbors dengan menggunakan NaNSMOTE yang diterapkan pada *Hybrid Approach* dapat memperoleh nilai akurasi, *precision*, dan *recall* yang lebih baik dibandingkan dengan SMOTE.

Kata Kunci: *Class Imbalance*, SMOTE, NaNSMOTE

1. PENDAHULUAN

Class imbalance merupakan salah satu topik yang paling sering dibahas oleh sejumlah peneliti yang mengkaji tentang *machine learning*. Permasalahan ini ditandai dengan adanya perbedaan jumlah *samples* yang cukup besar antara suatu *class* dibandingkan dengan *class* lainnya(Vuttipittayamongkol et al., 2021). Permasalahan ini didasari juga dengan adanya kenyataan bahwa pada *realtime dataset* cenderung mengalami permasalahan jumlah *sample* yang tidak sama untuk tiap *class*. Beberapa permasalahan nyata seperti *anomaly detection*(Chandola et al., 2009), *medical prediction*(Krawczyk et al., 2016), dan *object recognition*(X. Zhang et al., 2018) merupakan permasalahan yang pasti memiliki jumlah *sample* pada *class positive* (*minority class*) yang lebih sedikit dibandingkan *class negative* (*majority class*).

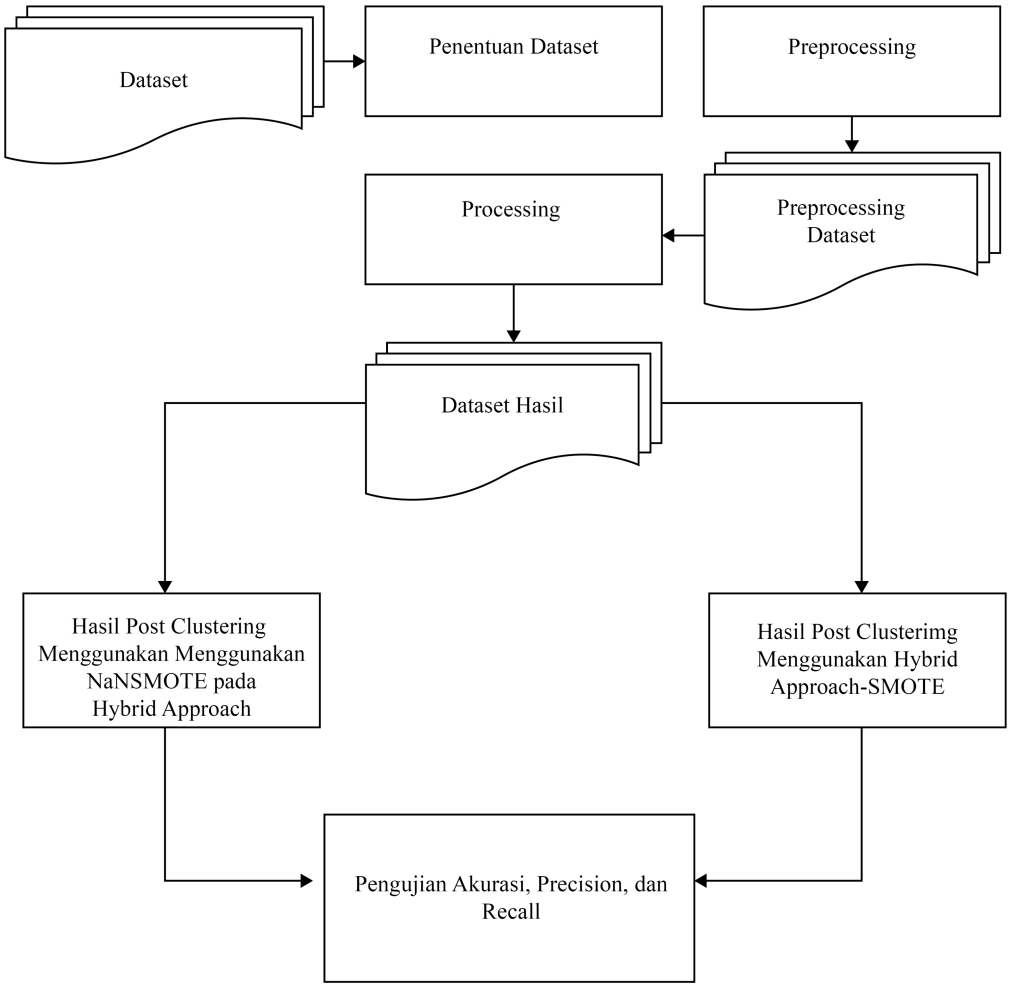
Permasalahan *class imbalance* seringkali menyebabkan hasil klasifikasi cenderung mengklasifikasn *samples* pada *majority class* dan mengabaikan *minority class* yang seringkali merupakan *class* yang mengandung informasi yang penting(Luque et al., 2019). Terdapat beberapa pendekatan yang umum digunakan di dalam menangani permasalahan *class imbalance*. Pendekatan tersebut dapat dikelompokkan menjadi *data-level, algorithm-level,* dan *Hybrid Approach*. *Hybrid Approach* secara umum menggunakan *oversampling* pada *minority class* dan juga pembangkitan sejumlah *classifier*(Zhao et al., 2020).

Diantar semua metode *oversampling* maka Synthetic Minority Oversampling Technique (SMOTE) merupakan metode yang paling umum digunakan(Chawla et al., 2002). Secara umum metode *oversampling* bekerja berdasarkan pada penentuan parameter *k* dan jumlah tetangga dari tiap *sample*. Salah satu metode yang cukup efektif di dalam penentuan nilai *k* secara adaptif adalah Natural Neighbors SMOTE (NaNSMOTE). Melalui penerapan NanSMOTE selain nilai *k* yang fleksibel juga *class center* dari tiap *sample* mempunyai jumlah *neighbors* yang lebih banyak dibandingkan dengan *border samples* sehingga dapat mengurangi kesalahan pada pembangkitan *sample*(Li et al., 2021).

Penentuan *Natural Neighbors* pada *Hybrid Approach* akan meningkatkan akurasi dari SMOTE di dalam penentuan parameter *k* dan *neighbor* pada tiap *class* sehingga dapat mengurangi kesalahan dan meningkatkan akurasi.

2. METODE

Adapun tahapan penelitian dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

Pada Gambar 1 dapat dilihat bahwa penelitian dimulai dengan penentuan *dataset*. Kemudian akan dilakukan proses *preprocessing* dataset. *Preprocessing dataset* kemudian akan menjalani tahapan *processing* dengan metode *NaNSMOTE* yang akan memanfaatkan keunggulan *NaNSMOTE* di dalam penentuan nilai *k* yang bersifat fleksibel dan juga penentuan *neighbors* pada tiap *class* yang lebih akurat. Hasil yang diperoleh kemudian akan dibandingkan dengan *Hybrid Approach* yang menggunakan SMOTE Adapun parameter untuk perbandingan di dalam penelitian ini menggunakan Akurasi, *Precision*, dan *Recall*.

2.1. *Hybrid Approach*

Adapun *pseudocode* dari *Hybrid Approach* adalah sebagai berikut(Galar et al., 2012).

$Input: D_T = \{x_1, x_2, \dots, x_n\}//Training\ Dataset$
$N = Number\ of\ Classifier$
$Output: Classification\ Prediction\ P$
$Method:$
$Step\ 1\ Preprocessing\ using\ Preprocessing\ Method$
$Step\ 2\ For\ i = 1\ to\ N\ do$
i. $Apply\ Machine\ Learning\ Classification\ Algorithm\ on\ The\ Attributes\ of\ D_T$
ii. $Obtain\ Classification\ Prediction\ P_i\ from\ machine$

$learning\ classification\ algorithm$
$End\ For$
$Step\ 3\ For\ i = 1\ to\ n$
$Apply\ processing\ using\ bagging, boosting\ or\ sampling$
$End\ For$

2.2. *SMOTE*

Adapun *pseudocode* dari *SMOTE* adalah sebagai berikut(Arafa et al., 2022).

$Input: X_{minor}, N_{percent}, K$
$Function\ SMOTE\ (X_{minor}, N_{percent}, K)$
1: $X_{SMOTE} \leftarrow \{\ \}$
2: $for\ i \leftarrow 1\ to\ len(X_{minor})do$
3: $nn \leftarrow K\ Nearest\ Neighbors\ (X_i, N_{Percent}, K)$
4: $p \leftarrow [Npercent/100]$
5: $while\ p! = do$
6: $X_{neighbour} \leftarrow select\ random\ (nn)$
7: $X_{SMOTE} \leftarrow X_i + rand\,(0,1)*|X_{neighbour}\text{-}X_i|$
8: $p \leftarrow p\text{-}1$
9: $end\ while$
10: $end\ for$
11: $return \leftarrow X_{SMOTE}$

2.3. *NaN SMOTE*

Adapun *pseudocode* dari *NaNSMOTE* adalah sebagai berikut(Li et al., 2021).

$Input: Smin\ (the\ set\ of minority\ class\ data), N\ (the\ number\ of\ samples\ generated\ for\ each\ selected\ base\ sample$
$output: setofsyntheticsamples(the\ set\ of\ synthetic\ samples)$
1: $NaNs = NaN\_search(Smin);$
2: $Smin = Smin\text{-}(x_i|x_i \in *MergeformatSmin\&\&|NaN(x_i| == 0);$
3: $Randomize\ Smin;$
4: $for\ i = 1\ to\ nmin$
5: $Base\_sample = Smin[i];$
6: $TempN = N;$
7: $while\ TempN \neq |*Mergeformat()$
8: $Nearest\ is\ one\ of\ NaNs\{i];$
9: $for\ j = 1\ to\ d$
10: $dif = nearest[j]\text{-}base\_sample[j];$
11: $gap = randomnumber\ between\ 0\ and\ 1;$
12: $SyntheticSamples[j] = base_{sample[j]} + gap\ x\ dif;$
13: $end\ for$
14: $SetOfSyntheticSamples = SetOfSyntheticSamples \cup |*MergeFormatSyntheticSamples;$
15: $TempN = TempN\text{-}1;$
16: $End\ While$
17: $End\ For$

2.4. *Dataset*

Dataset yang digunakan di dalam penelitian ini bersumber dari *KEEL Repository*(Alcalá-Fdez et al., 2009). Adapun *dataset* yang digunakan dalam penelitian ini dapat dilihat pada Tabel 1.

Tabel 1. *Dataset*

Dataset	#Ex	#Atts	Distribution of Class	IR
New-Thyroid	215	5	150/35/30	5
Balance	625	4	288/49/288	5.88
Car	1728	6	65/69/384/1210	18.61
Red Wine Quality	1599	11	10/53/681/638/199/18	68.1

2.5. Akurasi, *Precision*, dan *Recall*

Pengukuran akurasi, *precision*, dan *recall* didasarkan pada *confusion matrix*(L. Zhang et al., 2018). Adapun *confusion matrix* dapat dilihat pada Tabel 2.

Tabel 2. *Confusion Matrix*

	Predictive Positive Class	**Predictive Negative Class**
Actual Positive Class	True Positive (TP)	False Negative (FN)
Actual Negative Class	False Positive (FP)	True Negative (TN)

Adapun persamaan untuk menghitung akurasi, *precision*, dan *recall* dapat dilihat pada Persamaan 2-4(Watanabe et al., 2020).

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

3. HASIL DAN PEMBAHASAN

3.1. Hasil

Adapun nilai akurasi, *precision,* dan *recall* yang diperoleh oleh penentuan *Natural Neighbors* menggunakan *NaNSMOTE* pada *Hybrid Approach* dan *Hybrid Approach-SMOTE* dapat dilihat pada Tabel 3.

Tabel 3. Perbandingan Nilai Akurasi, *Precision*, dan *Recall*

Dataset	***Natural Neighbors* menggunakan *NaNSMOTE* pada *Hybrid Approach***			***Hybrid Approach-SMOTE***		
	Akurasi	***Precision***	***Recall***	**Akurasi**	***Precision***	***Recall***
New-Thyroid	0.91	0.87	0.88	0.88	0.88	0.86
Balance	0.89	0.89	0.87	0.88	0.87	0.84
Car	0.86	0.86	0.86	0.79	0.75	0.79
Red Wine Quality	0.79	0.87	0.85	0.77	0.74	0.73

Berdasarkan pada Tabel 3 dapat dilihat bahwa hasil yang diberikan oleh *Hybrid Approach* yang menggunakan penentuan *natural neighbors* pada proses *oversampling* menggunakan NaNSMOTE lebih baik jika dibandingkan dengan *Hybrid Approach-SMOTE* baik untuk nilai akurasi, *precision,* dan *recall*.

3.2. Pembahasan

Berdasarkan pada hasil penelitian dapat dilihat bahwa proses penentuan parameter *k* dan penentuan *natural neighbors* pada SMOTE sangat mempengaruhi penanganan *class imbalance*. Secara umum penentuan parameter *k* yang adaptif dan juga penentuan *sample* pada *class* dan *boundary sample* dapat mempengaruhi akurasi, *precision*, dan *recall* yang menyebabkan hasil yang diperoleh melalui penggunaan NaNSMOTE lebih baik bila dibandingkan dengan SMOTE.

4. KESIMPULAN

Berdasarkan hasil penelitian diperoleh bahwa nilai akurasi, *precision*, dan *recall* yang diberikan oleh penggunaan NaNSMOTE lebih baik bila dibandingkan dengan SMOTE. Hal ini menunjukkan bahwa proses *oversampling* perlu memperhatikan penentuan parameter *k* dan juga *neighbors* pada masing-masing *class* dan *boundary samples*. Penelitian ini perlu dikembangkan di dalam penanganan *multi-class imbalance*.

DAFTAR PUSTAKA

[1] Alcalá-Fdez, J., Sánchez, L., García, S., Jesus, M. J. del, Ventura, S., Garrell, J. M., Otero, J., Romero, C., Bacardit, J., Rivas, V. M., Fernández, J. C., & Herrera, F. (2009). KEEL: A software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, *13*(3), 307–318. https://doi.org/10.1007/s00500-008-0323-y

[2] Arafa, A., El-Fishawy, N., Badawy, M., & Radad, M. (2022). RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2022.06.005

[3] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 15:1-15:58. https://doi.org/10.1145/1541880.1541882

[4] 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.
[5] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). 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. https://doi.org/10.1109/TSMCC.2011.2161285
[6] Krawczyk, B., Galar, M., Jeleń, Ł., & Herrera, F. (2016). Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. *Applied Soft Computing*, *38*, 714–726. https://doi.org/10.1016/j.asoc.2015.08.060
[7] Li, J., Zhu, Q., Wu, Q., & Fan, Z. (2021). A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors. *Information Sciences*, *565*, 438–455. https://doi.org/10.1016/j.ins.2021.03.041
[8] Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, *91*, 216–231. https://doi.org/10.1016/j.patcog.2019.02.023
[9] Vuttipittayamongkol, P., Elyan, E., & Petrovski, A. (2021). On the class overlap problem in imbalanced data classification. *Knowledge-Based Systems*, *212*, 106631. https://doi.org/10.1016/j.knosys.2020.106631
[10] Watanabe, W. M., Felizardo, K. R., Candido, A., de Souza, É. F., Neto, J. E. de C., & Vijaykumar, N. L. (2020). Reducing efforts of software engineering systematic literature reviews updates using text classification. *Information and Software Technology*, *128*, 106395. https://doi.org/10.1016/j.infsof.2020.106395
[11] Zhang, L., Yang, H., & Jiang, Z. (2018). Imbalanced biomedical data classification using self-adaptive multilayer ELM combined with dynamic GAN. *BioMedical Engineering OnLine*, *17*. https://doi.org/10.1186/s12938-018-0604-3
[12] Zhang, X., Zhuang, Y., Wang, W., & Pedrycz, W. (2018). Transfer Boosting With Synthetic Instances for Class Imbalanced Object Recognition. *IEEE Transactions on Cybernetics*, *48*(1), 357–370. https://doi.org/10.1109/TCYB.2016.2636370
[13] Zhao, J., Jin, J., Chen, S., Zhang, R., Yu, B., & Liu, Q. (2020). A weighted hybrid ensemble method for classifying imbalanced data. *Knowledge-Based Systems*, *203*, 106087. https://doi.org/10.1016/j.knosys.2020.106087

15% Overall Similarity

Top sources found in the following databases:

- 10% Publications database
- Crossref database
- Crossref Posted Content database
- 10% Submitted Works database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Hartono Hartono, Erianto Ongko, Yeni Risyani. "Combining feature sele... Crossref	2%
2	"Soft Computing Applications in Industry", Springer Science and Busin... Crossref	2%
3	Erlin Erlin, Yenny Desnelita, Nurliana Nasution, Laili Suryati, Fransiskus ... Crossref	1%
4	Pradityo Utomo, Arief Budiman. "Application of Weighted Product (WP)... Crossref	1%
5	Universitas Pamulang on 2023-03-28 Submitted works	1%
6	Samsi Samsi. "UPAYA MENINGKATKAN HASIL BELAJAR MATEMATIK... Crossref	<1%
7	Associatie K.U.Leuven on 2017-01-10 Submitted works	<1%
8	Lukman Hakim, Nasrul Ilminnafik, Gaguk Jatisukamto, Moh. Nurkoyim ... Crossref	<1%
9	Sriwijaya University on 2020-07-23 Submitted works	<1%

10	Sriwijaya University on 2022-03-11 Submitted works	<1%
11	Institut Pertanian Bogor on 2023-04-06 Submitted works	<1%
12	J. Thomas Lindblad. "FOREIGN DIRECT INVESTMENT IN INDONESIA: F... Crossref	<1%
13	Universitas Brawijaya on 2019-01-03 Submitted works	<1%
14	Sriwijaya University on 2019-07-18 Submitted works	<1%
15	Zahid Ahmed, Sufal Das. "A Comparative Analysis on Recent Methods ... Crossref	<1%