ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

PENERAPAN ALGORITMA *SAFE-LEVEL-SMOTE* UNTUK PENINGKATAN NILAI *G-MEAN* DALAM KLASIFIKASI DATA TIDAK SEIMBANG

Resianta Perangin-angin, Eva J. G. Harianja, Indra Kelana Jaya, Benget Rumahorbo
Universitas Methodist Indonesia
Email: resianta88@gmail.com

DOI: https://doi.org/10.46880/jmika.Vol4No1.pp67-72

ABSTRAK

Klasifikasi data yang tidak seimbang merupakan masalah yang krusial pada bidang machine learning dan data mining. Ketidakseimbangan data memberikan dampak yang buruk pada hasil klasifikasi dimana kelas minoritas sering disalah klasifikasikan sebagai kelas mayoritas. Dimana kelompok kelas minoritas (*minority*) adalah kelompok kelas yang memiliki data lebih sedikit, dan kelompok kelas mayoritas (*mayority*) adalah kelompok kelas yang memiliki jumlah data lebih banyak. Data tidak seimbang adalah suatu kondisi dimana jumlah contoh dari salah satu kelas jauh lebih banyak dari kelas yang lain. Alasan buruknya kinerja metode klasifikasi biasa yang digunakan pada data tidak seimbang adalah bahwa tujuan metode klasifikasi dalam meminimumkan galat secara keseluruhan tidak dapat tercapai karena kelas minoritas hanya sedikit memberikan kontribusi, selain itu keputusan akhir yang dihasilkan tidak tepat karena terjadinya bias. Hal ini disebabkan oleh salah satu kelas mendominasi dalam hal jumlah. Dalam penelitian ini akan berfokus pada peningkatan nilai *G-Mean* dari dataset yang digunakan, dengan menerapkan algoritma *Safe-Level-Smote*. Dari hasil ujicoba yang dilakukan terhadap dua dataset yakni Abalon dan Vowel, untuk skema Smote + k-NN nilai G-Mean yang didapat yakni 0,47 untuk dataset Abalon dan 0.94 untuk dataset Vowel. Seletah dilakukan ujicoba terhadap dataset yang sama menggunakan skema *Safe-Level-Smote* menggunakan algoritma klasifikasi k-NN didapat hasil G-Mean 0,59 untuk dataset Abalon dan 1.00 Untuk dataset Vowel, rerata dari kenaikan nilai G-Mean terhadap algoritma Smote sebesar 12,68%. Hal ini membuktikan bahwasanya algoritma *Safe-Level-Smote* dapat meningkatkan nilai *G-Mean* pada klasifikasi data tidak seimbang menggunakan algoritma klasifikasi k-Nearst Neighbors.

Kata Kunci: Safe-Level-Smote, Smote, K-NN, Klasifikasi, Data Tidak Seimbang.

PENDAHULUAN

Data mining adalah ilmu yang mempelajari kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola dan hubungan dalam set data berukuran besar (Sugiato, 2015). Klasifikasi data yang tidak seimbang merupakan masalah yang krusial pada bidang machine learning dan data mining. Ketidakseimbangan data memberikan dampak yang buruk pada hasil klasifikasi dimana kelas minoritas sering disalah klasifikasikan sebagai kelas mayoritas (Siringoringo, 2018). Dimana kelompok kelas minoritas (minority) adalah kelompok kelas yang memiliki data lebih sedikit, dan kelompok kelas mayoritas (mayority) adalah kelompok kelas yang memilik jumlah data lebih banyak.

Data tidak seimbang adalah suatu kondisi dimana jumlah contoh dari salah satu kelas jauh lebih banyak dari kelas yang lain. Alasan buruknya kinerja metode klasifikasi biasa yang digunakan pada data tidak seimbang adalah bahwa tujuan metode klasifikasi dalam meminimumkan galat secara keseluruhan tidak dapat tercapai karena kelas minoritas hanya sedikit memberikan kontribusi, selain itu keputusan akhir yang dihasilkan tidak tepat karena terjadinya bias. Hal ini disebabkan oleh salah satu kelas mendominasi dalam hal jumlah amatan (Meidianingsih, Erfiani, & Sartono, 2017).

Pada hakekatnya data real atau data yang ditambang langsung dari database diastikan data tersebut dalam keadaan tidak seimbang. Dengan kondisi data seperti itu maka akan menyulitkan metode klasifikasi dalam melakukan fungsi generalisasi pada proses machine learning (Siringoringo, 2018). Banyak sekali algoritma-algoritma klasifikasi yang telah mengemukakan tekniknya masing-masing, namun hampir semua algoritma tersebut semisal Naive Bayes, KNearest Neighbor, Decision Tree dan algoritma yang lainnya pun selalu menunjukkan performa yang sangat buruk ketika bekerja pada d ata set yang tidak seimbang.

Dikarenakan algoritma yang disebutkan diatas tidak dilengkapi dengan kemampuan untuk menangani ketidakseimbangan kelas pada dataset. Klasifikasi pada data dengan kelas tidak seimbang

merupakan masalah yang sangat prioritas pada machine learning dan data mining, misalnya pada masalah medis (Kothandan, 2015) dan juga sangat pentig pada masalah klasifikasi teks (Wu, Ye, Zhang, Ng, & Ho, 2014), pada kegiatan sosial media (Li & Liu, 2017). Semuanya jika bekerja pada data kelas data yang tidak seimbang, hampir semua algoritma klasifikasi akan menghasilkan akurasi yang jauh lebih tinggi untuk kelas mayoritas dibandingkan dengan kelas minoritas (Gu, Wang, Wu, Ning, & Xin, 2016).

Metode Synthetic Minority Over-sampling Technique (SMOTE) merupakan metode yang populer diterapkan dalam rangka menangani ketidak seimbangan kelas. Teknik ini mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan dataset dengan cara sampling ulang sampel kelas minoritas (Siringoringo, 2018).

Algoritma yang umum digunakan untuk penanganan kelas data yang tidak seimbang adalah Metode Synthetic Minority Over-sampling Technique (SMOTE) dimana dalam beberapa penelitian metode SMOTE terbukti dapat menginkatkan G-Mean dan F-Measure dalam data kelas tidak seimbang, seperit pada penelitian kartu kredit metode smote dapat meningkatkan nilai 80,0% untuk nilai G-Mean dan 81,8% untuk nilai F-Measure.

Dalam penelitian ini akan berfokus pada nilai G-Mean dari dataset yang dipilih, dengan menerapkan algoritma Safe-Level-SMOTE, dimana algoritma ini merupakan dikembangkan dari algoritma SMOTE yang bertujuan untuk meningkatkan nilai G-Mean pada saat bekerja di kelas data tidak seimbang.

Synthetic Minority Over-sampling Technique Synthetic Minority Oversampling Technique (SMOTE) adalah salah satu turunan dari oversampling. SMOTE pertama kali diperkenalkan oleh Nithes V. Chawla (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Dimana Pendekatan ini bekerja dengan membuat replikasi dari data minoritas. Replikasi tersebut dikenal dengan data sintetis (syntetic data). Metode SMOTE bekerja dengan mencari k nearest neighbors untuk setiap data di kelas minoritas, setelah itu dibuat data sintetis sebanyak persentase duplikasi yang diinginkan antara data minor dan knearest neighbors yang dipilih secara acak (Siringoringo, 2018).

KAJIAN PUSTAKA

Safe-Level-SMOTE

Berdasarkan pada SMOTE, Safe-Level-SMOTE, Safe-Level-Synthetic Minority Oversampling TEchnique, yakni menetapkan setiap instance positif level terlebih dahulu sebelum membuat instance sintetis. Setiap instance sintetis diposisikan lebih dekat ke tingkat aman terbesar sehingga semua instance sintetis hanya dihasilkan di wilayah aman. Level aman (sl) didefinisikan sebagai rumus safe level (sl) = the number of a positive stances in k nearest neighbours…….(1)

Jika level aman instance dekat dengan 0, instance hampir noise. Jika dekat dengan k, instance dianggap aman. Itu rasio tingkat aman didefinisikan sebagai rumus. safe level ratio = sl of a positive instance / sl of a nearest neighbours……(2) Ini digunakan untuk memilih posisi aman untuk menghasilkan instance sintetis (Bunkhumpornpat, Sinapiromsaran, & Lursinsap, 2009).

K-Nearst Neighbor

Dalam klasifikasi Algoritma K-Nears Neigbors tidak memerlukan preprocessing dari data set sample sebelum penggunaannya. Gagasan ini dapat diperluas ke tetangga AT-terdekat dengan vektor yang ditugaskan itugaskan ke kelas yang diwakili oleh mayoritas di antara tetangga AT-terdekat. Tentu saja, ketika lebih dari satu tetangga dipertimbangkan, kemungkinan bahwa akan ada ikatan antara kelas dengan jumlah maksimum tetangga dalam kelompok tetangga terdekat yang ada (Keller , Gray, & Givens, 1985).

Yang menjadi fokus dan tujuan dari algoritma ini adalah mengklasifikasikan obyek berdasarkan atribut dan training sample. Clasifier tidak menggunakan apapun untuk dicocokkan dan hanya berdasarkan pada memori. Diberikan titik query, akan ditemukan sejumlah k obyek atau (titik training) yang paling dekat dengan titik query. Klasifikasi menggunakan voting terbanyak diantara klasifikasi dari k obyek.Algoritma K-Nearest Neighbor (K-NN) menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari query instance yang baru (Lidya, Sitompul, & Efendi, 2015)

METODE PENELITIAN

Untuk metode pengukuran performa dari algoritma yang akan di ujicoba terhadap dataset dalam kasus ini menggunakan Confusion matrix dikarenakan metode ini sangat populer dalam mengevaluasi performa algoritma klasifikasi. Untuk lebih jelasnya dapat dilihat pada Tabel 1 tampilan confusion matrix untuk kelas biner, yaitu dataset dengan dua jenis kelas saja.

Tabel 1. Confusion Matrix Kelas Biner

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

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. Namun untuk penelitian ini akan berfokus pada peningkatan nilai G-Mean dengan menggunakan Algoritma Safe-Level-SMOTE dengan persamaan dibawah ini

$$Recal = TP_{rate} = \frac{TP}{TP+FN} \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$Precision = PP_{value} = \frac{TP}{TP+FP} \ldots\ldots\ldots\ldots (2)$$

$$Specificity = TN_{rate} = \frac{TN}{TN+FP} \ldots\ldots\ldots\ldots (3)$$

$$G\text{-}Mean = \sqrt{TPrate - TNrate} \ldots\ldots (4)$$

HASIL DAN PEMBAHASAN

Berangkat dari metode penelitian, maka akan di uji semua dataset yang telah di pilih menggunakan skema Smote+k-NN dan Safe-Level-Smote + k-NN, untuk melihat hasil peningkatan nilai G-Mean yang dihasilkan oleh algoritma tersebut

Dataset

Penelitian ini menggunakan 2 dataset yang berbeda dan dengan Imbalancing Ratio (IR) yang berbeda pula, dimana dataset ini bersumber dari sci2s.ugr.es/keel/datasets.php. untuk dataset Abalon dengan IR. 9,18 dan Untuk Dataset Vowel dengan IR. 9,98, adapun atribut dari masing-masing dataset dapat dilihat pada tabel 2 dibawah ini.

Tabel 2. Atribut Dataset

Dataset	Atribut	IR
Abalon	Sex, Length, Diameter, Height, Whole_weight, Shucked_weight, Viscera_weight, Shell_weight	9,18
Vowel	TT, SpeakerNumber, Sex, F0, F1, F2, F3, F4, F5, F6, F7, F8, F9	9,98

Prosedur SMOTE+kNN

1. Membuat partisi dataset secara acak menjadi 5 bagian dengan skema 5-fold cross validation
2. Menerapkan penanganan kelas data tidak seimbang dengan SMOTE sebanyak dua kalipada data latih:
   a. Menentukan nilai tetangga dengan k=1
   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=1,2,3,5,7, dan 9
   b. Menghitung jarak antar data kelas minoritas dengan metode eucledian
4. Membandingkan kinerja klasifikasi dan dengan diterapkannya SMOTE dan Safe-Level-SMOTE. Kinerja klasifikasi yang diterapkan adalah G-Mean

Prosedur Safe-Level-SMOTE+kNN

Untuk prosedur Safe-Level-SMOTE + kNN 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 Safe-Level-SMOTE sebanyak 2 (dua) kali pada data latih :
   a. Menentukan nilai tetangga dengan k = 1
   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. Menghitung jarak antar data kelas minoritas dengan metode eucledian
   b. Melihat nilai kenaikan G-Mean

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

Hasil Pengujian Menggunakan Algoritma SMOTE + kNN

Selanjutnya akan dilakukan ujicoba untuk algoritma kNN+SMOTE, smote disini mempunyai fungsi sebagai preprosesing data dalam keadaan data tidak seimbang, oleh karenanya dalam pengujian ini akan dilakukan dengan dataset yang sama dan jumlah pengujian yang sama pulauntuk hasil pengujian dapat dilihat pada Tabel 3 dibawah ini.

Tabel 3. Hasil pengujian dataset Abalon

Performa	P-1	P-2	P-3	P-4	P-5	Rerata
TP	42	37	37	40	39	39
TN	554	555	556	557	542	553
FP	0	5	5	2	3	3
FN	135	134	133	132	147	136
Acc	0.82	0.81	0.81	0.82	0.79	0.81
Recal	0.24	0.22	0.22	0.23	0.21	0.22
Speci	1.00	0.99	0.99	1.00	0.99	0.99
Prec	1.00	0.88	0.88	0.95	0.93	0.93
G-M	**0.49**	**0.46**	**0.46**	**0.48**	**0.46**	**0.47**
F-M	0.38	0.35	0.35	0.37	0.34	0.36

Tabel 4. Hasil pengujian dataset Vowel

Performa	P-1	P-2	P-3	P-4	P-5	Rerata
TP	90	90	90	90	90	90
TN	886	891	882	885	888	886
FP	0	0	0	0	0	0
FN	12	7	16	13	10	12
Acc	0.99	0.99	0.98	0.99	0.99	0.99
Recal	0.88	0.93	0.85	0.87	0.90	0.89
Speci	1.00	1.00	1.00	1.00	1.00	1.00
Prec	1.00	1.00	1.00	1.00	1.00	1.00
G-M	0.94	0.96	0.92	0.93	0.95	0.94
F-M	0.94	0.96	0.92	0.93	0.95	0.94

Hasil Pengujian Menggunakan Algoritma Safe-Level-SMOTE + kNN

Tabel 5. Hasil pengujian dataset Abalon

Performa	P-1	P-2	P-3	P-4	P-5	Rerata
TP	8	9	8	9	10	9
TN	674	673	674	675	674	674
FP	34	33	34	33	32	33
FN	15	16	15	14	15	15
Acc	0.93	0.93	0.93	0.94	0.94	0.93
Recal	0.35	0.36	0.35	0.39	0.40	0.37
Speci	0.95	0.95	0.95	0.95	0.95	0.95
Prec	0.19	0.21	0.19	0.21	0.24	0.21
G-M	0.58	0.59	0.58	0.61	0.62	0.59
F-M	0.25	0.27	0.25	0.28	0.30	0.27

Tabel 6. Hasil pengujian dataset Vowel

Performa	P-1	P-2	P-3	P-4	P-5	Rerata
TP	90	90	90	90	90	90
TN	898	898	898	898	898	898
FP	0	0	0	0	0	0
FN	0	0	0	0	0	0
Acc	1.00	1.00	1.00	1.00	1.00	1.00
Recal	1.00	1.00	1.00	1.00	1.00	1.00
Speci	1.00	1.00	1.00	1.00	1.00	1.00
Prec	1.00	1.00	1.00	1.00	1.00	1.00
G-M	1.00	1.00	1.00	1.00	1.00	1.00
F-M	1.00	1.00	1.00	1.00	1.00	1.00

PEMBAHASAN

Dari hasil pengujian dataset Abalon dan Vowel, dilihat bahwasanya hasil rerata dari G-Mean mengalami kenaikan dengan menggunakan algoritma Safe-Level-Smote bila dibandingkan dengan penggunaan SMOTE saja, hal ini tentu akan menjadi sebuah kesimpulan dimana algoritma Safe-Level-SMOTE, untuk lebih jelasnya hasil dari Safe-Level-SMOTE dapat dilihat pada tabel berikut:

Tebel 7. Hasil G-Mean Abalon

Performa	Safe-Level	SMOTE
TP	9	39
TN	674	553
FP	33	3
FN	15	136
Acc	0.93	0.81
Recal	0.37	0.22
Speci	0.95	0.99
Prec	0.21	0.93
G-M	0.59	0.47
F-M	0.27	0.36

Tebel 8. Hasil G-Mean Vowel

Performa	Safe-Level	SMOTE
TP	90	90
TN	898	886
FP	0	0
FN	0	12
Acc	1.00	0.99
Recal	1.00	0.89
Speci	1.00	1.00
Prec	1.00	1.00
G-M	1.00	0.94
F-M	1.00	0.94

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

Dari tabel diatas bahwa hasil dari ujicoba yang dilakukan terhadap 2 dataset, menghasilkan peningkatan G-Mean yang cukup baik, hal ini membuktikan bahwa Safe-Level-SMOTE dapat meningkatkan nilai G-Mean dari Algoritma SMOTE, dengan mengimplementasikannya terhadap algoritma klasifikasi k-NN. Untuk dataset Abalon nilai G-Mean meningkat sebesar 25,53% sedangkan pada dataset Vowel nilai G-Mean Sebesar 6,38%. Untuk hasil rerata peningkatan nilai G-Mean dapat dilihat pada tabel dibawah ini.

Tabel 9. Rerata Performa

Performa	Safe-Level	SMOTE
TP	49.40	64.50
TN	786.00	719.60
FP	16.60	1.50
FN	7.50	73.90
Acc	0.97	0.90
Recal	0.68	0.55
Speci	0.98	1.00
Prec	0.60	0.96
GM	0.80	0.71
F-M	0.63	0.65

Dari hasil rerata nilai G-Mean terhadap 2 dataset yang di ujikan dengan menggunakan algoritma klasifkasi k-nn dan nilai k=1, didapat kenaikan nilai G-Mean sebesar 12,68%

KESIMPULAN

Dari hasil ujicoba yang dilakukan terhadap dua dataset yakni Abalon dan Vowel, untuk skema Smote + k-NN nilai G-Mean yang didapat yakni 0,47 untuk dataset Abalon dan 0.94 untuk dataset Vowel. Seletah dilakukan ujicoba terhadap dataset yang sama menggunakan skema Safe-Level-Smote menggunakan algoritma klasifikasi k-NN didapat hasil G-Mean 0,59 untuk dataset Abalon dan 1.00 Untuk dataset Vowel, rerata dari kenaikan nilai G-Mean terhadap algoritma SMOTE sebesar 12,68%. Hal ini membuktikan bahwasanya algoritma Safe-Level-Smote dapat meningkatkan nilai G-Mean pada klasifikasi data tidak seimbang menggunakan algoritma klasifikasi k-Nearst Neighbors.

DAFTAR PUSTAKA

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem. *Advances in Knowledge Discovery and Data Mining , 5476*, 475-482.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research , 16*, 321-357.

Gu, Q., Wang, X.-M., Wu, Z., Ning, B., & Xin, C.-S. (2016). An Improved SMOTE Algorithm Based on Genetic Algorithm for Imbalanced Data Classification. *Journal of Digital Information Management , 14* (2), 92-103.

Keller , J., Gray, M., & Givens, J. (1985). A fuzzy K-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics. SMC-15*, pp. 580-585. IEEE.

Kothandan, R. (2015). Handling class imbalance problem in miRNA dataset associated with cancer. *Bioinformation , 11* (1), 6-10.

Li, C., & Liu, S. (2017). A comparative study of the class imbalance problem in Twitter spam detection. *Concurrency and Computation* (Special).

Lidya, S. K., Sitompul, O. S., & Efendi, S. (2015). Sentiment analysis pada teks bahasa indonesia menggunakan support vector machine (SVM) dan K-nearest neighbor (K-NN). *Seminar Nasional Teknologi Informasi dan Komunikasi 2015 (SENTIKA 2015)*, *1*, pp. 1-8. Yogyakarta.

Meidianingsih, Q., Erfiani, & Sartono, B. (2017). *Kajian Metode Safe-Level SMOTE pada Kasus Klasifikasi Data Tidak Seimbang.* Bogor: Institut Pertanian Bogor.

Siringoringo, R. (2018). Klasifikasi data tidak seimbang menggunakan algoritma SMOTE dan k-nearest neighbor. *Jurnal ISD, 3* (1), 44-49.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

Sugiato, C. A. (2015). Analisis komparasi algoritma klasifikasi untuk menangani data tidak seimbang pada data kebakaran hutan. *Techno.COM , 14* (4), 336-342.

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.