KLASIFIKASI DATA TIDAK SEIMBANG MENGGUNAKAN ALGORITMA *SMOTE* DAN *k-NEAREST NEIGHBOR*

Rimbun Siringoringo
Universitas Methodist Indonesia
E-mail : rimbun.ringo@gmail.com

ABSTRACT

Unbalanced data classification is a crucial problem in the field of machine learning and data mining. Data imbalances have a poor impact on classification results where minority classes are often misclassified as a majority class. k-Nearest Neighbor is one of the most popular and simple classification methods but it is not equipped with the ability to work on unbalanced datasets. In this study, the Synthetic Minority Over-Sampling Technique (SMOTE) was applied to solve the class imbalance problem on the Credit Card Fraud dataset. By applying the 10-cross-validation evaluation scheme, it was found that SMOTE increases the mean of G-Mean by 53.4% to 81.0% and the mean of F-Measure by 38.7 to 81.8%

Keywords*: Class imbalance, Synthetic Minority Over-sampling Technique, k-Nearest Neighbor*

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. k-Nearest Neighbor merupakan salah satu metode klasifikasi yang sangat populer dan sederhana tetapi, tidak dilengkapi dengan kemampuan untuk bekerja pada dataset tidak seimbang. Pada penelitian ini, Synthetic Minority Over-sampling Technique (SMOTE) diterapkan untuk menyelesaikan masalah ketidak seimbangan kelas pada dataset Credit Card Fraud. Dengan menerapkan skema evaluasi 10-cross fold validation diperoleh hasil bahwa SMOTE meningkatkan rata-rata G-Mean dari 53,4% ke 81,0% dan rata-rata F-Measure dari 38,7 ke 81,8%*Kata kuci : ketidakseimbangan kelas, *Synthetic Minority Over-sampling Technique, k-Nearest Neighbor*

***Kata kunci** : ketidakseimbangan kelas, Synthetic Minority Over-sampling Technique, k-Nearest Neighbor*

PENDAHULUAN

Data 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 lainnya [1]. Kelompok kelas data yang lebih sedikit dikenal dengan kelompok minoritas (*minority*), kelompok kelas data yang lainnya disebut dengan kelompok mayoritas (*mayority*).

Pada hakekatnya data *real*, data yang ditambang langsung dari *databas* adalah tidak seimbang. Kondisi tersebut menyulitkan metode klasifikasi dalam melakukan fungsi generalisasi pada proses *machine learning*. Hampir semua algoritma klasifikasi seperti *Naive Bayes*, *Decision Tree*, *K-Nearest Neighbor* dan yang lainnya menunjukkan performa yang sangat buruk ketika bekerja pada data dengan kelas yang sangat tidak seimbang. Metode-metode klasifikasi yang telah disebutkan di atas tidak delengkapi dengan kemampuan untuk menangani masalah ketidak seimbangan kelas.

Klasifikasi pada data dengan kelas tidak seimbang merupakan masalah utama pada bidang *machine learning* dan *data mining*, misalnya pada masalah medis [2], masalah klasifikasi teks [3], sosial media [4]. Jika bekerja pada data tidak seimbang, hampir semua algoritma klasifikasi akan menghasilkan akurasi yang jauh lebih tinggi untuk kelas mayoritas daripada kelas minoritas [5]. Perbedaan ini merupakan suatu indikator performa klasifikasi yang buruk. Pada beberapa kasus, kelas minoritas justru lebih penting untuk diidentifikasi daripada kelas mayoritas. Misalnya pada kasus transaksi dengan kartu kredit, kebanyakan status transaksi adalah transaksi yang normal, hanya sedikit kasus yang dapat ditemukan dimana terjadi transaksi yang *fraud*. Meskipun demikian, keberadaan transaksi yang *fraud* jauh lebih penting untuk di identifikasi daripada transaksi yang normal meskipun jumlah kasusnya jauh lebih sedikit. [5]

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.

Penelitian Terkait

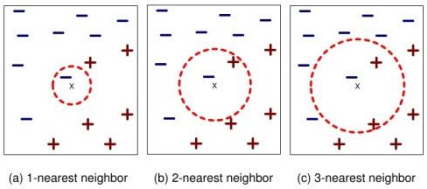
Penerapan SMOTE pada bidang klasifikasi menunjukkan perbaikan performa dari metode klasifikasi yang ada. Penerapan SMOTE pada masalah klasifikasi infertilitas menggunakan metode klasifikasi *Multi Level Perceptron* (MLP), *k-Nearest Neighbor* (k-NN), Naïve Bayes, *Random Forest*, dan *Support Vector Machine.* [6] SMOTE berhasil meningkatkan performa klasifikasi dengan hasil terbaik didapatkan dengan metode *Naive Bayes* sebesar 90,7%. Penerapan SMOTE pada masalah klasifikasi pada *dataset* menggunakan *Fuzzy C-Means* [7] dapat mengatasi ketidak seimbangan kelas pada data set tersebut. Pada masalah *clustering* dengan metode *k-Means Clustering*, Penerapan SMOTE berhasil menghilangkan *noise* serta menyelesaikan masalah ketidak seimbangan pada 71 *dataset*. Penerapan SMOTE dapat memperbaiki kualitas *clustering* [8].

Synthetic Minority Over-sampling Technique

Synthetic Minority Oversampling Technique (SMOTE) adalah salah satu turunan dari *oversampling*. SMOTE pertama kali diperkenalkan oleh Nithes V. Chawla [9]. 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* (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 *k-nearest 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 *Instance-Based 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. Jumlah tetangga yang dipakai adalah sebanyak *k* tetangga. Prinsip ketetanggaan dapat diilustrasikan pada gambar: [10]



Gambar 1. Ilustrasi *k-Nearest Neighbor*

Setelah mengambil *k* tetangga terdekat pertama kemudian dihitung jumlah data yang mengikuti kelas yang ada dari *k* tetangga tersebut. Kelas dengan data terbanyak yang mengikutinya menjadi kelas pemenang yang diberikan sebagai label kelas pada data X.

Pada kNN, nilai k dapat memberikan pengaruh terhadap performa klasifikasi yang dihasilkan. Jika nilai *k* terlalu kecil

METODE PENELITIAN

Pengukuran Performa

Metode pengukuran performa memiliki eranan yang sangat penting untuk mengevaluasi kinerja suatu metode klasifikasi. *Confusion matrix* merupakan alat yang paling populer dalam mengevaluasi performa klasifikasi. Pada tabel berikut ditampilkan *confusion matrix* untuk kelas biner, yaitu dataset dengan dua jenis kelas saja.

Tabel 1. *Confusion matrix* kelas biner

Class	*Predictive Positive*	*Predictive 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*, *F-Measure*, *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 TP_{rate}, PP_{value}, *F-Measure* dan *G-Mean* [11]. *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*.

$$Recall = TP_{rate} = \frac{TP}{TP + FN} \quad (1)$$

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

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

$$G - Mean = \sqrt{TP_{rate} - TN_{rate}} \quad (4)$$

HASIL DAN PEMBAHASAN

Dataset

Penelitian ini menggunakan dataset *Credit Card Fraud* yang bersumber dari *UCI repository*. Dataset ini terdiri dari 29.976 data, dimana 23.347 merupakan data positif (label 0) dan 6.629 merupakan data negatif (label 1). *Imbalance Ratio* (IR) dataset adalah 3, 521.

Tabel 2. Daftar atribut dataset

No	Atribut
1	ID
2	LIMIT_BAL
3	SEX
4	EDUCATION
5	MARRIAGE
6	AGE
7	PAY_0
8	PAY_2
9	PAY_3
10	PAY_4
11	PAY_5
12	PAY_6
13	BILL_AMT1
14	BILL_AMT2
15	BILL_AMT3
16	BILL_AMT4
17	BILL_AMT5
18	BILL_AMT6
19	PAY_AMT1
20	PAY_AMT2
21	PAY_AMT3
22	PAY_AMT4
23	PAY_AMT5
24	PAY_AMT6
25	payment_next_month

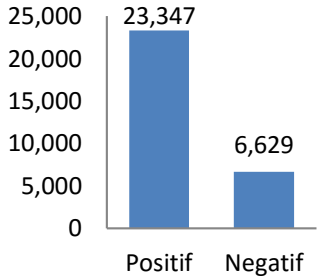
Prosedur SMOTE+kNN

1. Partisi dataset secara acak menjadi 10 bagian dengan skema 10-*fold cross validation*
2. Menerapkan penanganan kelas data tidak seimbang dengan SMOTE sebanyak dua kali pada data latih :
   i. Menentukan nilai tetangga dengan k=5,
   ii. Menghitung jarak antar data kelas minoritas dengan metode *eucledian*
   iii. Melakukan perhitungan untuk membangkitkan data buatan (*syntetic*)

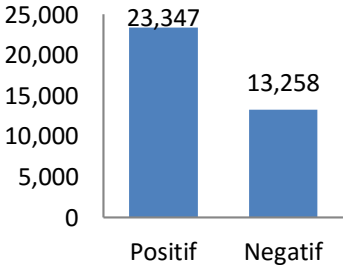
3. Menerapkan k-nearest neighbor untuk mengklasifikasi data uji :
   i. Menentukan nilai tetangga dengan k=1,2,3,5,7, dan 9
   ii. Menghitung jarak antar data kelas minoritas dengan metode *eucledian*
4. Membandingkan kinerja klasifikasi tanpa dan dengan diterapkannya SMOTE. Kinerja klasifikasi yang diterapkan adalah *G-Mean*

Penerapan SMOTE

Penerapan SMOTE meminimalisasi ketidak seimbangan kelas pada dataset *Credit Card Fraud* dengan membangkitkan data sintesis, sehingga total keseluruhan data terdiri dari 36.6056 data, yaitu 23.347 merupakan data positif (label 0) dan 13.258 merupakan data negatif (label 1).



Gambar 2. Grafik ketidak seimbangan awal



Gambar 3 . Grafik ketidak seimbangan akhir dengan SMOTE

Brdasarkan gambar di atas dapat dipahamai bahwa penerapan SMOTE membangkitkan data sintesis untuk data negatif sebesar 6.629 data.

Pada tabel berikut ditampilkan data perbandingan performa klasifikasi dengan kNN dan dengan kNN+SMOTE. Pada setiap percobaan menerapkan nilai *k* yang bervariasi yaitu 1,2,3,5,7, dan 9

Tabel 3. Hasil pengujian untuk nilai k=1

Performa	kNN+SMOTE	kNN
TP	21930	2551
TN	18436	19370
FP	4911	3977
FN	4586	3977
Accuracy	0,810	0,734
Recall	0,827	0,391
Precision	0,817	0,391
G-Mean	0,808	0,569
F-Measure	0,822	0,391

Untuk nilai k=1, pengaruh SMOTE berhasil meningkatkan *Accuracy, G-Mean dan F-Measure* masing-masing sebesar 7,6%, 23,9% dan 43,1%

Tabel 4. Hasil pengujian untuk nilai k=2

Performa	kNN+SMOTE	kNN
TP	21930	1305
TN	18436	22208
FP	4911	1139
FN	4911	5324
Accuracy	0,804	0,784
Recall	0,817	0,197
Precision	0,817	0,534
G-Mean	0,803	0,433
F-Measure	**0,817**	0,288

Pada tabel di atas, pengaruh SMOTE berhasil meningkatkan *Accuracy, G-Mean dan F-Measure* masing-masing sebesar 2,0%, 37,0% dan 52,1%

Tabel 5. Hasil pengujian untuk nilai k=3

Performa	kNN+SMOTE	kNN
TP	21463	2291
TN	19116	20918
FP	4231	2429
FN	5053	4338
Accuracy	0,814	0,774
Recall	0,809	0,346
Precision	0,835	0,485
G-Mean	**0,814**	0,556
F-Measure	**0,822**	0,404

Untuk nilai k=3, pengaruh SMOTE berhasil meningkatkan *Accuracy, G-Mean dan F-Measure* masing-masing sebesar 4,0%, 25,8% dan 41,8%

Tabel 6. Hasil pengujian untuk nilai k=5

Performa	kNN+SMOTE	kNN
TP	21158	2176
TN	19452	21573
FP	3895	1774
FN	5358	4453
Accuracy	0,814	0,792
Recall	0,798	0,328
Precision	0,845	0,551
G-Mean	0,815	0,551
F-Measure	0,821	0,411

Untuk nilai k=5, pengaruh SMOTE berhasil meningkatkan *Accuracy, G-Mean dan F-Measure* masing-masing sebesar 2,2%, 26,5% dan 40,9%

Tabel 7. Hasil pengujian untuk nilai k=7

Performa	kNN+SMOTE	kNN
TP	20822	2134
TN	19643	21837
FP	3704	1510
FN	5694	4495
Accuracy	0,812	0,800
Recall	0,785	0,322
Precision	0,849	0,586
G-Mean	0,813	0,549
F-Measure	0,816	0,415

Untuk nilai k=7, pengaruh SMOTE berhasil meningkatkan *Accuracy, G-Mean dan F-Measure* masing-masing sebesar 1,0%, 26,4% dan 40,0%

Tabel 8. Hasil pengujian untuk nilai k=9

Performa	kNN+SMOTE	kNN
TP	20509	2085
TN	19760	22009
FP	3587	1338
FN	6007	4544
Accuracy	0,808	0,804
Recall	0,773	0,315
Precision	0,851	0,609
G-Mean	0,809	0,545
F-Measure	0,810	0,415

Untuk nilai k=9, pengaruh SMOTE berhasil meningkatkan *Accuracy, G-Mean dan F-Measure* masing-masing sebesar 2,0%, 26,5% dan 39,6%

Tabel 9. Perbandingan *G-Mean* k-NN dengan SMOTE+k-NN

k	kNN+SMOTE	kNN
1	**0,808**	0,569
2	**0,803**	0,433
3	**0,814**	0,556
5	**0,815**	0,551
7	**0,813**	0,549
9	**0,809**	0,545

Tabel 10. Perbandingan *F-Measure* k-NN dengan SMOTE+k-NN

k	kNN+SMOTE	kNN
1	**0,822**	0,391
2	**0,817**	0,288
3	**0,822**	0,404
5	**0,821**	0,411
7	**0,816**	0,415
9	**0,810**	0,415

Pada tabel 9 dan 10 terlihat bahwa nilai *G-Mean* dan *F-Measure* untuk kNN lebih kecil dibandingkan nilai *G-Mean* serta *F-Measure* untuk kNN+SMOTE, hal tersebut membuktikan bahwa penerapan SMOTE memberian dampak yang signifikan terhadap perbaikan nilai G-Mean dan F-Measure

Untuk mengetahui apakah nilai *G-Mean* dan *F-Measure* pada k-NN berbeda secara signifikan dengan performa k-NN+SMOTE, maka pengujian dilakukan dengan metode *Wilcoxon Sing Rank Test* dengan taraf α=95%. Hasil pengujian ditampilkan pada tabel berikut.

Tabel 11. *Wilcoxon Sing Rank Test* untuk *G-Mean*

Exact confidence	Confidence interval	p-value	Asymptotic P-value
0.969	[0.116 , 0.212]	**0.031**	0.022

Tabel 12. *Wilcoxon Sing Rank Test* untuk *F-Measure*

Exact confidence	Confidence interval	p-value	Asymptotic P-value
0.906	[0.403 , 0.474]	**0.031**	0.021

Berdasarkan tabel di atas diperoleh nilai *p-value* sebesar (0,031). Nilai *p-value* tersebut lebih kecil dari 0,05. Berdasarkan nilai *p-value* tersebut dapat disimpulkan bahwa terdapat perbedaan *G-mean* dan *F-Measure* yang signifikan antara k-NN dan k-NN +SMOTE.

KESIMPULAN

Dataset *Credit Card Fraud* merupakan dataset yang memiliki ketidak seimbangan kelas. Penelitian ini menerapkan metode *Syntetic Minority Over Sampling Technique* (SMOTE) untuk menangani ketidak seimbangan kelas pada dataset *Credit Card Fraud*, serta metode *k-Nearest Neighbor* (kNN) untuk melaksanakan fungsi klasifikasi. Hasil eksperimen dengan menerakan nilai *k* yang bervariasi yaitu 1,2,3,5,7, dan 9 diperoleh rata-rata nilai performa *G-Mean* sebesar 81,0% untuk skema SMOTE + kNN dan 53,4% untuk skema kNN. Rata-rata performa *F-Measure* untuk skema SMOTE+kNN adalah sebesar 81,8% dan dengan skema kNN adalah 38,7%.

Berdasarkan hasil eksperimen di atas diperoleh kesimpulan bahwa penerapan SMOTE+kNN mampu menangani ketidak seimbangan kelas dataset *Credit Card Fraud* dengan menghasilkan nilai *G-Mean* dan *F-Measure* yang lebih tinggi dibandingkan dengan kNN saja. Hal tersebut membuktikan bahwa metode SMOTE efektif meningkatkan performa klasifikasi data tidak seimbang

DAFTAR PUSTAKA

[1]A. Ali, S. M. Shamsuddin, & A. L. Ralescu, "Classification with class imbalance problem: a review," *Int J Adv. Soft Compu Appl*, vol. 7, no. 3, 2015.

[2]R. Kothan&, "Handling class imbalance problem in miRNA dataset associated with cancer," *Bioinformation*, vol. 11, no. 1, pp. 6–10, Jan 2015.

[3]Q. Wu, Y. Ye, H. Zhang, M. K. Ng, & S.-S. Ho, "ForesTexter: An efficient random forest algorithm for imbalanced text categorization," *Knowl.-Based Syst.*, vol. 67, pp. 105–116, Sep 2014.

[4]C. Li & S. Liu, "A comparative study of the class imbalance problem in Twitter spam detection," *Concurr. Comput. Pract. Exp.*, pp. n/a-n/a.

[5]Q. Gu, X.-M. Wang, Z. Wu, B. Ning, & C.-S. Xin, "An improved SMOTE algorithm based on genetic algorithm for imbalanced data classification," *J Dig Inf Manag*, vol. 14, no. 2, pp. 92–103, 2016.

[6]B. Karlik, A. Yibre, & K. Barış, *Comprising Feature Selection and Classifier Methods with SMOTE for Prediction of Male Infertility*, vol. 3. 2016.

[7]R. Pruengkarn, K. W. Wong, & C. C. Fung, "Multiclass Imbalanced Classification Using Fuzzy C-Mean and SMOTE with Fuzzy Support Vector Machine," dalam *Neural Information Processing*, 2017, pp. 67–75.

[8]E. M. El Houby, N. I. Yassin, & S. Omran, "A Hybrid Approach from Ant Colony Optimization and K-nearest Neighbor for Classifying Datasets Using Selected Features," *Informatica*, vol. 41, no. 4, 2017.

[9]N. V. Chawla, K. W. Bowyer, L. O. Hall, & W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[10]N. C. Barde & M. Patole, "Classification and Forecasting of Weather using ANN, k-NN and Naïve Bayes Algorithms."

[11]W. Prachuabsupakij & P. Doungpaisan, "Matching preprocessing methods for improving the prediction of student's graduation," dalam *Computer and Communications (ICCC), 2016 2nd IEEE International Conference on*, 2016, pp. 33–37.