Improvisasi Teknik *Oversampling* MWMOTE Untuk Penanganan Data Tidak Seimbang

Pramana Yoga Saputra, Moch Zawaruddin Abdullah*, Annisa Puspa Kirana

Jurusan Teknik Informatika, Politeknik Negeri Malang, Malang, Indonesia

Email: ¹pramana.yoga@polinema.ac.id, ^{2,*}zawaruddin@polinema.ac.id, ³puspakirana@polinema.ac.id

Email Penulis Korespondensi: zawaruddin@polinema.ac.id

Abstrak—Data tidak seimbang (*imbalance data*) merupakan keadaan dimana terjadi perbedaan jumlah data yang mengakibatkan terjadinya kelas mayoritas (kelas dengan anggota sangat banyak) dan kelas minoritas (kelas dengan anggota sangat sedikit). Hal ini dapat menyulitkan proses klasifikasi karena metode algoritma *machine learning* dirancang untuk melakukan klasifikasi pada data yang sudah seimbang. Pendekatan metode *oversampling* digunakan untuk mengatasi ketidakseimbangan data dengan menambahkan data sintetis pada kelas minoritas sehingga memiliki kuantitas data yang sama dengan kelas mayoritas. MWMOTE merupakan metode *oversampling* yang melakukan *generate* data sintetis berdasarkan anggota cluster kelas minoritas yang berdekatan dengan kelas mayoritas. Metode ini mampu menghasilkan data sintetis secara baik. Akan tetapi data sintetis yang dihasilkan masih terpadu pada area berdekatan dengan kelas mayoritas dan terlalu rapat densitasnya pada area tepi cluster, sehingga memungkinkan data sintetis yang dihasilkan masuk pada klasifikasi kelas mayoritas. Penelitian ini bertujuan untuk melakukan improvisasi proses *generate* data sintetis pada metode MWMOTE agar data yang dihasilkan tersebar luas dalam kelas minoritas. Hasil dari pengujian menunjukkan metode yang diusulkan mampu meningkatkan performa klasifikasi untuk klasifikasi KNN dan C4.5 Decision Tree sebesar 0.46% dan 0.96% dibandingkan dengan metode MWMOTE.

Kata Kunci: Improvisasi; *Oversampling*; MWMOTE; Data Tidak Seimbang; Data Mining

Abstract—Imbalance data is a condition which there is a distinction in the quantity of data that results within the majority class (classes with very many members) and minority class (classes with very few members). It can complicate the classification process since the *machine learning* algorithm method is designed to classify already balanced data. The oversampling process technique is used to resolve data imbalance by applying synthetic data to the minority class in such a manner that it has the same volume of data as the majority class. MWMOTE is an oversampling technique that generates synthetic data based on members of the minority class clusters that are close to the majority class. This approach is capable of generating synthetic data well. The resulting synthesis data remains in the nearby majority region and too dense on the border of the cluster. It is hence permitting the resulting synthetic data to go into the majority class classification. This study is objectives to improve the process of generating synthetic data on MWMOTE so that the resulting data is extensively dispensed within the minority class. The outcomes of the test show that the proposed method is capable of enhancing the classification performance for KNN and C4.5 Decision Tree classification sequentially by 0.46% and 0.96% compared to MWMOTE.

Keywords: Improvising; Oversampling; MWMOTE; Imbalance Data; Data Mining

1. PENDAHULUAN

Data tidak seimbang (*imbalance data*) merupakan keadaan dimana terjadi perbedaan jumlah data pada kelas data. Kelas data yang tidak seimbang dikelompokkan menjadi 2 yaitu data kelas minoritas dan data kelas mayoritas [1]. Data tidak seimbang dapat menyulitkan proses klasifikasi karena metode algoritma *machine learning* dirancang untuk melakukan klasifikasi pada data yang seimbang. Ketidakseimbangan data dapat menyebabkan pembuatan data model yang tidak sesuai [2] serta menyebabkan *overfitting* dan penurunan akurasi sebuah klasifikasi [3], sehingga dataset dianggap memiliki kuantitas data yang sama atau seimbang (*balance*). Apabila terdapat data pada kelas minoritas, maka akan dijadikan sebagai *noise/outlier* [4], [5].

Misklasifikasi menjadi suatu permasalahan yang sering terjadi dalam proses klasifikasi pada data yang tidak seimbang. Hal tersebut dapat terjadi karena *classifier* akan cenderung mengarah ke data kelas mayoritas dan mengakibatkan tingkat akurasi terhadap data kelas minoritas menjadi rendah [6]. Untuk menangani data tidak seimbang, beberapa penelitian melakukan manipulasi sampel data (pembuatan data sintetis) dan mengaplikasikan suatu algoritma [7], [8].

Pendekatan metode *oversampling* digunakan untuk mengatasi ketidakseimbangan data dengan menambahkan data sintetis pada kelas minoritas sehingga memiliki kuantitas data yang sama dengan kelas mayoritas. Beberapa metode *oversampling* yang ada antara lain *Synthetic Minority Over-sampling Technique* (SMOTE) [7], *Borderline-SMOTE* [9], *Adaptive Synthetic* (ADASYN) [10], *Majority Weighted Minority Oversampling Technique* (MWMOTE) [1], dan lain sebagainya.

Oversampling merupakan suatu metode yang dilakukan untuk menangani permasalahan ketidakseimbangan data dengan melakukan pembagian data yang seimbang. Pembagian data dilakukan dengan cara melakukan replikasi data kelas minoritas secara acak dan berulang-ulang sampai memiliki kuantitas yang sama dengan data kelas mayoritas [6]. Mekanisme pembuatan data sintetis pada *oversampling* memiliki kekurangan dengan munculnya *overfitting*. *Overfitting* terjadi karena data sintetis yang dibuat terlalu fokus pada *training dataset* tertentu sehingga tidak bisa melakukan prediksi dengan tepat apabila diberikan dataset lain yang serupa.

MWMOTE merupakan metode yang men-*generate* data sintesis berdasarkan anggota *cluster* kelas minoritas yang berdekatan dengan kelas mayoritas. Metode ini dapat menangani *overfitting* dengan men-*generate* data sintesis secara baik dan mampu menghindari data sintesis yang dihasilkan menjadi data *noise*. Metode MWMOTE melakukan tiga tahapan dalam membuat data sintesis, yaitu 1) melakukan proses identifikasi sampel kelas minoritas dan kelas mayoritas pada dataset; 2) melakukan pembobotan pada kelas minoritas; 3) melakukan proses *clustering*. Tahapan proses tersebut mampu menghasilkan data sintesis dengan tingkat akurasi lebih baik dan menurunkan derajat bias (*noise*) pada klasifikasi [1].

Metode MWMOTE menghasilkan data sintesis dengan tingkat akurasi lebih baik, akan tetapi data sintesis yang dihasilkan masih terpaku pada area tepi *cluster*, sehingga data sintesis berada di area *borderline minorities* atau data anggota terluar kelas minoritas yang paling dekat dengan kelas mayoritas. Hal ini dapat menjadikan data sintesis yang dihasilkan terlalu rapat densitasnya pada area tepi *cluster* dibandingkan area lain pada satu *cluster* kelas minoritas. Akibatnya dapat terjadi pembiasan klasifikasi karena data sintesis untuk kelas minoritas dapat di-*cluster*-kan ke kelas mayoritas ataupun data sintesis dianggap sebagai *noise/outlier*.

Berdasarkan permasalahan di atas munculah gagasan untuk meningkatkan (improvisasi) proses *generate* data sintesis pada MWMOTE agar data yang dihasilkan tersebar luas dalam kelas minoritas dan tidak hanya tersebar pada area *borderline minorities*, sehingga meningkatkan hasil suatu klasifikasi.

2. METODOLOGI PENELITIAN

2.1 Imbalance Data

Data tidak seimbang (*imbalance data*) adalah data yang memiliki jumlah data yang tidak seimbang antara data kelas satu dengan data kelas lainnya, disebut juga *imbalanced*. Dalam *data mining*, *imbalanced* diartikan sebagai rasio jumlah data kelas mayoritas dengan kelas minoritas tidak sama. Kelas mayoritas memiliki jumlah data lebih banyak dibandingkan dengan data pada kelas minoritas. Sebagai contoh terdapat 1000 data profil mahasiswa perguruan tinggi tertentu yang diklasifikasikan dengan dua kelas yaitu kelas jenis kelamin perempuan dan kelas jenis kelamin laki-laki. Dimana jumlah data kelas jenis kelamin perempuan adalah 775 data, sedangkan untuk kelas jenis kelamin laki-laki hanya 225 data. Sehingga kelas mahasiswa jenis kelamin perempuan mendominasi dan menjadi kelas mayoritas dibandingkan dengan kelas mahasiswa jenis kelamin laki-laki yang menjadi kelas minoritas.

Peristiwa ketidakseimbangan data dapat terjadi pada *machine learning*, sehingga mengakibatkan *misclassification* dan berdampak pada penurunan nilai akurasi prediksi [11]. Penurunan akurasi pada data tidak seimbang dapat disebabkan karena banyaknya ditemukan *noise/outlier* pada *dataset* uji yang berasal dari kelas minoritas [12]. Distribusi data yang tidak seimbang menghasilkan klasifikasi yang lebih cenderung ke data mayoritas dibandingkan dengan data minoritas [13]. Salah satu cara dalam menangani data tidak seimbang adalah menggunakan metode klasifikasi dengan mengaplikasikan algoritma [7], [8]. Data tidak seimbang dapat diselesaikan dengan penambahan data sintesis pada kelas minoritas dengan metode *oversampling*, ataupun pengurangan data pada kelas mayoritas dengan metode *undersampling*.

2.2 MWMOTE

Majority Weighted Minority Oversampling Technique (MWMOTE) [1] merupakan metode yang men-*generate* data sintesis berdasarkan anggota *cluster* kelas minoritas yang berdekatan dengan kelas mayoritas. Metode ini mampu melakukan *generate* data sintesis secara baik dan mampu menghindari data sintesis yang telah dihasilkan menjadi data *noise*. Metode MWMOTE memiliki dua tujuan, yaitu: meningkatkan proses pemilihan sampel dan meningkatkan proses pembuatan sampel data sintesis. Metode MWMOTE memiliki tiga tahap dalam prosesnya seperti ilustrasi yang ditunjukkan Gambar 1 yaitu: (1) MWMOTE melakukan identifikasi terhadap data kelas minoritas yang berada di dalam area *cluster* data mayoritas, data kelas minoritas yang berdekatan (*borderline*) dengan data mayoritas, serta data kelas mayoritas yang berada berdekatan (*borderline*) pada kelas minoritas. (2) Setiap anggota sampel informatif minoritas diberi bobot seleksi *sample weight* (S_w), sesuai dengan kepentingan data. (3) Proses *clustering* bertujuan untuk menjaga sampel data sintesis yang dihasilkan berada dalam satu kelompok *cluster* [6]. Terdapat enam data parameter yang ada pada MWMOTE yang digunakan untuk melakukan *generate* data sintesis, yaitu:

1. S_{maj} : kumpulan data mayoritas
2. S_{min} : kumpulan data minoritas
3. N : jumlah data sintesis yang akan di-*generate*
4. $k1$: jumlah *node* tetangga yang digunakan untuk memprediksi *noise* pada kelas minoritas
5. $k2$: jumlah *node* tetangga dari kelas mayoritas yang digunakan untuk membangun informasi mengenai kelas minoritas
6. $k3$: jumlah *node* tetangga dari kelas minoritas yang digunakan untuk membangun informasi mengenai kelas minoritas.

Gambar 1. Tahapan MWMOTE

2.3 Tahapan Penelitian

Data sintetis yang dihasilkan MWMOTE masih terdapat di sekitar area *borderline minorities* yaitu area tepian kelas minoritas yang berdekatan dengan kelas mayoritas. Hal ini dapat menjadikan data sintetis yang dihasilkan terlalu rapat densitasnya pada area tepi *cluster* dibandingkan area lain pada satu *cluster* kelas minoritas. Sehingga dapat terjadi pembiasan klasifikasi karena data sintetis untuk kelas minoritas dapat dianggap sebagai *noise* bahkan dapat di-*cluster*-kan ke kelas mayoritas.

Penelitian ini bertujuan untuk melakukan improvisasi proses *generate* data sintetis pada MWMOTE berdasarkan titik tengah (*centroid*) dari *cluster* minoritas dan *borderline minorities* yang disebut *Centroid-Borderline MWMOTE* (CB-MWMOTE). CB-MWMOTE diusulkan agar data yang dihasilkan tersebar luas dalam kelas minoritas dan tidak hanya tersebar pada area *borderline minorities*. Tahapan yang dilakukan pada penelitian ini adalah

2.3.1 Identifikasi Kelas Minoritas Yang Paling Penting S_{imin} (paling dekat dengan kelas mayoritas)

Proses identifikasi kelas minoritas masih mengacu pada metode MWMOTE, dimana proses yang dilakukan adalah untuk mengidentifikasi S_{imin} . Dalam melakukan pengidentifikasian S_{imin} terdapat 3 variabel yang ditunjukkan untuk 3 sub proses, yaitu variabel $k1$, $k2$, dan $k3$.

- 1) Sub proses hapus data *noise* pada kelas minoritas berdasarkan $k1$. $k1$ merupakan jumlah node tetangga yang digunakan untuk memprediksi *noise* pada kelas minoritas. Gambar 2 menunjukkan mengenai proses yang dilakukan, dimana data kelas minoritas akan dicari terlebih dahulu *node* tetangganya (contoh *node* tetangga terdekat yang dicari sebanyak 5 *node*, $k1 = 5$), dan apabila *node* tetangga tersebut semuanya adalah kelas mayoritas, maka data tersebut diindikasikan *noise*, karena tidak ada informasi yang didapatkan dari satu data minoritas di dalam area mayoritas dan bisa dihapus data kelas minoritasnya.

Gambar 1. Data *noise* pada kelas minoritas

- 2) Identifikasi *borderline majorities* (S_{bmaj}) berdasarkan $k2$. $k2$ merupakan jumlah node tetangga dari kelas mayoritas yang digunakan untuk membangun informasi mengenai kelas minoritas. Gambar 3 mengilustrasikan data kelas mayoritas akan dicari berdasarkan *node* dari kelas minoritas yang berdekatan dengan kelas mayoritas. Data-data kelas mayoritas yang paling dekat dengan kelas minoritas akan disimpan dan disebut *borderline majorities* (S_{bmaj}).

Gambar 2. *Borderline majorities*

- 3) Identifikasi *borderline minorities* (S_{imin}) berdasarkan $k3$. $k3$ merupakan jumlah node tetangga dari kelas minoritas yang digunakan untuk membangun informasi mengenai kelas minoritas. Gambar 4 mengilustrasikan data kelas minoritas akan dicari berdasarkan *node* dari *borderline majorities* (S_{bmaj}) pada kelas mayoritas yang berdekatan dengan kelas minoritas. Data-data kelas minoritas yang paling dekat dengan S_{bmaj} akan disimpan dan disebut *borderline minorities* (S_{imin}).

Gambar 3. S_{imin} untuk kelas minoritas

2.3.2 Identifikasi Kelas Minoritas S_{bmin} (paling dekat dengan *centroid cluster* kelas minoritas)

MWMOTE menghasilkan data sintesis pada area S_{imin} yang berada di *borderline minorities*. Proses ini merupakan proposal improvisasi metode MWMOTE dengan menambahkan informasi data S_{bmin} dengan penambahan parameter $k4$. S_{bmin} merupakan *mini-cluster* dari *cluster* kelas minoritas, yaitu area antara *centroid cluster* minoritas dengan *borderline minorities* yang menjadi area *generate* data sintesis baru dan $k4$ merupakan jumlah node tetangga dari *centroid cluster* minoritas. Data S_{bmin} menjadi kunci pada metode ini agar data sintesis yang dihasilkan tidak terlalu rapat ke tepi (*borderline minorities*) akan tetapi menyebar dalam area *cluster* minoritas yang seperti ditunjukkan pada Gambar 5.

Gambar 4. Penentuan S_{bmin}

2.3.3 Generate Data Sintesis Berdasarkan S_{bmin}

Generate data sintesis pada penelitian ini dilakukan dengan menambahkan parameter $k4$ pada MWMOTE untuk menentukan area *generate* data sintesis (S_{bmin}). Tahapan *generate* dilakukan dengan cara:

1. Menentukan *cluster-cluster* minoritas (L_M) dari kelas minoritas S_{imin}
2. Inisialisasi *oversample* data $S_{omin} \rightarrow S_{omin} = S_{imin}$
3. Lakukan perhitungan sebanyak N data sintesis sampling, dengan mengacu pada Persamaan 1.

$$S = x + \alpha \times (y - x) \quad (1)$$

Dimana,

x : data sampel random dari S_{imin} dalam *cluster* L_k , $1 \leq k \leq M$

y : data sampel random dari S_{bmin} pada *cluster* yang sama dengan x

α : nilai random $[0, 1]$

3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan 8 dataset dari *UC Irvine Machine learning Repository* untuk pengujian, yaitu dataset *abalone*, *ecoli*, *glass*, *page blocks*, *pima*, *image segmentation*, *vehicle* dan *wine*. Dataset yang didapat memiliki banyak fitur sehingga perlu dilakukan *preprocessing* untuk mengubahnya menjadi *binary dataset* yaitu dataset yang hanya memiliki dua fitur, fitur kelas minoritas dan fitur kelas mayoritas.

Tabel 1. Deskripsi Dataset

Dataset	Kelas Minoritas	Kelas Mayoritas	Jumlah Fitur	Total Data	Data Minoritas	Data Mayoritas	Rasio (%)
Abalone	'18'	'9'	7	731	42	689	06 : 94
Ecoli	'im'	kelas lainnya	7	336	77	259	23 : 77

Dataset	Kelas Minoritas	Kelas Mayoritas	Jumlah Fitur	Total Data	Data Minoritas	Data Mayoritas	Rasio (%)
Glass	'5', '6', '7'	kelas lainnya	9	214	51	163	24 : 76
Page Blocks	'Graphic', 'vertical line', 'picture'	kelas lainnya	10	5476	231	5245	04 : 96
Pima	'1'	'0'	8	768	268	500	35 : 65
Image Segmentation	'Grass'	kelas lainnya	19	2310	330	1980	14 : 86
Vehicle	'1'	'0'	18	940	219	721	23 : 77
Wine	'3'	kelas lainnya	13	178	48	130	27 : 73

Tabel 1 menunjukkan karakteristik dari *dataset* yang akan digunakan dalam pengujian. Penentuan kelas minoritas dan mayoritas mengacu pada penelitian [1]. *Dataset* yang diperoleh akan dilakukan normalisasi sebelum dilakukan proses pengujian. Hal ini dilakukan untuk mendapatkan skala data yang sama yaitu rentang nilai antara 0-1. Pengujian dilakukan dengan menggunakan aplikasi *The Waikato Environment for Knowledge Analysis* atau WEKA dengan versi 3.8. WEKA merupakan kumpulan algoritma *machine learning* dan *tools* untuk *preprocessing* data [14]. Algoritma klasifikasi yang digunakan pada pengujian ini adalah algoritma klasifikasi *K-Nearest Neighbor (KNN)* dan *C4.5 decision tree*.

K-Nearest Neighbour (K-NN) adalah metode yang dikenal paling sederhana dan merupakan metode yang memiliki akurasi sangat tinggi [15]. K-NN melakukan klasifikasi data berdasarkan tingkat kemiripan yang dihitung berdasarkan jarak terdekat. Parameter *k* pada K-NN merupakan parameter ketetanggaan dan digunakan untuk menunjukkan jumlah tetangga terdekat yang berperan dalam memprediksi label kelas pada data yang diuji [16]. Pada pengujian ini nilai ketetanggaan yang digunakan adalah 3, yaitu $k = 3$.

Salah satu metode lain dapat digunakan dalam klasifikasi *data mining* adalah *decision tree* (pohon keputusan). *Decision Tree* digunakan untuk mempelajari klasifikasi dan prediksi pola dari data, dan menggambarkan relasi dari variabel atribut *x* dan variabel target *y* dalam bentuk pohon [17]. Salah satu algoritma *decision tree* yang sering digunakan adalah *C4.5 decision tree*. Algoritma C4.5 merupakan algoritma turunan dan hasil pengembangan dari algoritma ID3 dengan peningkatan seperti penanganan untuk atribut-atribut numerik, *missing value*, dan *noise* pada *dataset*; dan aturan-aturan yang dihasilkan dari model pohon (*tree*) yang terbentuk [18].

Tabel 1. Performa akurasi MWMOTE dan metode usulan (CB-MWMOTE)

Dataset	Metode	Total Data	<i>k-NN (k = 3)</i>			<i>C4.5 Decision Tree</i>		
			Klasifikasi Minoritas	Klasifikasi Mayoritas	Akurasi (%)	Klasifikasi Minoritas	Klasifikasi Mayoritas	Akurasi (%)
Abalone	MWMOTE	1378	655/689	601/689	91.15	627/689	589/689	88.24
	CB-MWMOTE	1378	658/689	632/689	93.61	645/689	644/689	93.54
Ecoli	MWMOTE	518	249/259	230/259	92.47	249/259	233/259	93.05
	CB-MWMOTE	518	254/259	229/259	93.24	243/259	233/259	91.89
Glass	MWMOTE	326	162/163	155/163	97.24	158/163	153/163	95.40
	CB-MWMOTE	326	162/163	154/163	96.93	161/163	149/163	95.09
Page Block	MWMOTE	10484	5206/5242	5015/5242	97.49	5187/5242	5101/5242	98.13
	CB-MWMOTE	10484	5220/5242	5024/5242	97.71	5200/5242	5127/5242	98.50
Pima	MWMOTE	1000	427/500	355/500	78.20	388/500	339/500	72.70
	CB-MWMOTE	1000	427/500	348/500	77.50	408/500	352/500	76.00
Image Segment	MWMOTE	3960	1978/1980	1980/1980	99.95	1980/1980	1978/1980	99.95
	CB-MWMOTE	3960	1977/1980	1980/1980	99.92	1976/1980	1980/1980	99.90
Vehicle	MWMOTE	1268	592/634	456/634	82.65	549/634	486/634	81.62
	CB-MWMOTE	1268	599/634	465/634	83.91	553/634	500/634	83.04
Wine	MWMOTE	260	130/130	125/130	98.08	127/130	124/130	96.54

Dataset	Metode	Total Data	<i>k</i> -NN (<i>k</i> = 3)			<i>C4.5</i> Decision Tree		
			Klasifikasi Minoritas	Klasifikasi Mayoritas	Akurasi (%)	Klasifikasi Minoritas	Klasifikasi Mayoritas	Akurasi (%)
	CB-MWMOTE	260	130/130	124/130	97.70	125/130	121/130	94.61

Tabel 2 menunjukkan tingkat akurasi yang diperoleh MWMOTE dan metode usulan CB-MWMOTE. Nilai akurasi merupakan persentase jumlah data yang diklasifikasikan secara benar oleh sebuah algoritma [19]. Total data pada Tabel 2 merupakan jumlah total data pada *dataset* yang sudah dilakukan *oversampling* untuk kelas minoritas baik menggunakan metode MWMOTE maupun metode yang diusulkan. Nilai yang dibalikkan merupakan hasil yang terbaik diantara kedua metode pada tiap *dataset*. Akurasi yang diperoleh menunjukkan bahwa metode CB-MWMOTE unggul pada *dataset abalone*, *page blocks*, *vehicle* dan bersaing ketat dengan metode MWMOTE pada *dataset* lainnya. Metode yang diusulkan memiliki rata-rata akurasi klasifikasi K-NN (*k* = 3) sebesar 92.57% dan *C4.5 decision tree* sebesar 91.57%, sedangkan metode MWMOTE memiliki rata-rata akurasi klasifikasi KNN (*k* = 3) sebesar 92.15% dan *C4.5 decision tree* sebesar 90.70%. Metode yang diusulkan memperoleh peningkatan klasifikasi dibandingkan dengan metode MWMOTE sebesar 0.46% untuk KNN dan 0.96% untuk *C4.5 decision tree*.

Tabel 3 menunjukkan detail akurasi pada metode MWMOTE dan metode CB-MWMOTE. Detail akurasi yang digunakan dalam penelitian ini adalah *Precision / confidence*, *Recall / true positive rate (TPR) / sensitivity*, *F-Measure*, dan *receiver operating characteristic (ROC) Area* atau yang sering disebut juga *AUC (Area Under the ROC Curve)*. ROC Area atau AUC digunakan untuk merangkum akurasi diagnostik keseluruhan dari suatu pengujian. ROC Area bernilai antara 0 hingga 1, di mana nilai 0 menunjukkan pengujian yang sangat tidak akurat dan nilai 1 mencerminkan pengujian yang sangat akurat [20], [21].

Tabel 2. Detail akurasi MWMOTE dan metode usulan (CB-MWMOTE)

Dataset	Metode	<i>k</i> -NN (<i>k</i> = 3)				<i>C4.5</i> Decision Tree			
		Precision	Recall	F-measure	ROC	Precision	Recall	F-measure	ROC
Abalone	MWMOTE	0.914	0.911	0.911	0.952	0.884	0.882	0.882	0.914
	Proposed	0.937	0.936	0.936	0.973	0.935	0.935	0.935	0.944
Ecoli	MWMOTE	0.927	0.925	0.925	0.963	0.932	0.931	0.930	0.934
	Proposed	0.936	0.932	0.932	0.960	0.920	0.919	0.919	0.931
Glass	MWMOTE	0.973	0.972	0.972	0.982	0.954	0.954	0.954	0.959
	Proposed	0.970	0.969	0.969	0.979	0.953	0.951	0.951	0.962
Page Block	MWMOTE	0.976	0.975	0.975	0.986	0.981	0.981	0.981	0.988
	Proposed	0.978	0.977	0.977	0.987	0.985	0.985	0.985	0.985
Pima	MWMOTE	0.788	0.782	0.781	0.843	0.729	0.727	0.726	0.761
	Proposed	0.782	0.775	0.774	0.839	0.763	0.760	0.759	0.794
Image Segment	MWMOTE	0.999	0.999	0.999	1.000	0.999	0.999	0.999	1.000
	Proposed	0.999	0.999	0.999	1.000	0.999	0.999	0.999	1.000
Vehicle	MWMOTE	0.842	0.826	0.824	0.876	0.819	0.816	0.816	0.839
	Proposed	0.855	0.839	0.837	0.887	0.833	0.830	0.830	0.845
Wine	MWMOTE	0.981	0.981	0.981	0.993	0.966	0.965	0.965	0.959
	Proposed	0.978	0.977	0.977	0.985	0.947	0.946	0.946	0.952

Tabel 3 menunjukkan metode yang diusulkan CB-MWMOTE unggul dalam klasifikasi *dataset abalone*, *page blocks*, dan *vehicle* baik pada algoritma klasifikasi *K-NN* maupun *C4.5 decision tree*. Metode yang diusulkan unggul dengan rata-rata *precision*, *recall*, *f-measure*, dan *ROC area* secara berurutan adalah 0.921, 0.917, 0.917, 0.937. Pada metode MWMOTE nilai rata-rata yang didapat adalah 0.903, 0.899, 0.898, 0.926. Pada *dataset* ini, metode CB-MWMOTE meningkatkan proses klasifikasi rata-rata sebesar 1.825%.

CB-MWMOTE juga bersaing ketat dengan metode MWMOTE dalam klasifikasi pada *dataset ecoli* dan *pima*. CB-MWMOTE unggul dalam klasifikasi pada *dataset ecoli* dengan algoritma *K-NN* dan pada *dataset pima* dengan algoritma *C4.5 decision tree*, sedangkan MWMOTE unggul dalam klasifikasi pada *dataset ecoli* dengan

algoritma *C4.5 decision tree* dan pada *dataset pima* dengan algoritma *K-NN*. Namun secara rata-rata *precision*, *recall*, *f-measure*, dan *ROC area* untuk *dataset ecoli* dan *pima*, metode CB-MWMOTE sedikit lebih baik dibandingkan MWMOTE dengan perolehan secara berurut 0.851, 0.847, 0.846, 0.882 dan untuk MWMOTE sebesar 0.845, 0.842, 0.841, 0.876. Pada *dataset ecoli* dan *pima* ini, metode CB-MWMOTE meningkatkan proses klasifikasi rata-rata sebesar 0,646%.

Metode CB-MWMOTE menghasilkan data sintesis yang hampir sama dengan metode MWMOTE sehingga pada proses klasifikasi K-NN dan *C4.5 decision tree* untuk *dataset image segmentation* tidak terlihat perbedaannya. Akan tetapi metode yang diusulkan yaitu CB-MWMOTE masih kurang baik pada klasifikasi untuk *dataset glass* dan *wine* dibandingkan dengan metode MWMOTE. Nilai rata-rata *precision*, *recall*, *f-measure*, dan *ROC area* yang dihasilkan oleh metode CB-MWMOTE adalah 0.962, 0.961, 0.961, 0.97 lebih rendah dibandingkan dengan rata-rata yang dihasilkan metode MWMOTE, yaitu 0.969, 0.969, 0.969, 0.974. Metode CB-MWMOTE mengalami penurunan hasil klasifikasi rata-rata sebesar 0,696% untuk *dataset glass* dan *wine*.

Dari kedelapan *dataset* yang telah dilakukan pengujian, metode CB-MWMOTE memiliki keunggulan dalam klasifikasi beberapa *dataset* dan begitu pula dengan metode MWMOTE. Dalam ringkasan akhir percobaan didapatkan bahwa secara global metode CB-MWMOTE sedikit lebih baik dalam men-generate data sintesis sehingga meningkatkan hasil klasifikasi yang didapat. Dimana metode CB-MWMOTE secara rata-rata memberikan peningkatan hasil klasifikasi untuk K-NN sebesar 0.431% dan *C4.5 decision tree* sebesar 0.989%.

4. KESIMPULAN

Penelitian ini mengusulkan improvisasi metode *oversampling* MWMOTE untuk menangani data yang tidak seimbang. Data set yang digunakan adalah 8 *dataset* yaitu *dataset abalone*, *ecoli*, *glass*, *page blocks*, *pima*, *image segmentation*, *vehicle* dan *wine*. *Preprocessing* dilakukan pada seluruh *dataset* sehingga fitur yang digunakan dalam penelitian ini terdiri dari 2 fitur yaitu fitur kelas minoritas dan fitur kelas mayoritas. Berdasarkan hasil penelitian menunjukkan bahwa metode yang diusulkan yaitu CB-MWMOTE unggul dalam klasifikasi *dataset abalone*, *page blocks*, dan *vehicle* baik pada algoritma klasifikasi K-NN maupun *C4.5 decision tree*. Pada ketiga *dataset* tersebut, metode CB-MWMOTE meningkatkan proses klasifikasi rata-rata sebesar 1.825%. CB-MWMOTE juga bersaing ketat dengan metode MWMOTE dalam klasifikasi pada *dataset ecoli* dan *pima*. Metode CB-MWMOTE menghasilkan data sintesis yang hampir sama dengan metode MWMOTE sehingga pada proses klasifikasi K-NN dan *C4.5 decision tree* untuk *dataset image segmentation* tidak terlihat perbedaannya. Akan tetapi metode yang diusulkan CB-MWMOTE masih kurang baik pada klasifikasi untuk *dataset glass* dan *wine* dibandingkan dengan metode MWMOTE. Metode CB-MWMOTE secara rata-rata keseluruhan *dataset*, memberikan peningkatan hasil klasifikasi untuk K-NN sebesar 0.431% dan *C4.5 decision tree* sebesar 0.989%. Dari kedelapan *dataset* yang telah dilakukan pengujian, metode CB-MWMOTE memiliki keunggulan dalam klasifikasi beberapa *dataset* dan begitu pula dengan metode MWMOTE. Metode CB-MWMOTE yang diusulkan terbukti mampu memberikan peningkatan untuk penanganan hasil klasifikasi data yang tidak seimbang. Akan tetapi, perlu penelitian lebih lanjut untuk mengetahui performa dan tingkat konsistensi metode yang diusulkan jika menggunakan data yang berukuran besar atau dengan jumlah fitur yang banyak. Selain itu perlu juga dilakukan perbandingan performa dengan beberapa teknik *oversampling* lainnya sehingga bisa mendapatkan metode yang paling efisien untuk menangani data tidak seimbang.

REFERENCES

- [1] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE - Majority weighted minority *oversampling* technique for imbalanced data set learning," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 2, pp. 405–425, 2014.
- [2] J. Gong and H. Kim, "RHSBoost: Improving classification performance in imbalance data," *Comput. Stat. Data Anal.*, vol. 111, pp. 1–13, 2017.
- [3] I. Fakhruzi, "An artificial neural network with bagging to address imbalance datasets on clinical prediction," *2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018*, vol. 2018-Janua, no. 1, pp. 895–898, 2018.
- [4] K. Napierała, "Improving Rule Classifiers For Imbalanced Data," Poznan University of Technology, 2012.
- [5] P. Phoungphol, "A Classification Framework for Imbalanced Data," Georgia State University, 2013.
- [6] M. C. Untoro and J. L. Buliali, "Penanganan imbalance class data laboratorium kesehatan dengan Majority Weighted Minority *Oversampling* Technique," *Regist. J. Ilm. Teknol. Sist. Inf.*, vol. 4, no. 1, p. 23, 2018.
- [7] Nitesh V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, no. Sept. 28, pp. 321–357, 2002.
- [8] C. Seiffert, T. M. Khoshgoftaar, and J. Van Hulse, "Hybrid sampling for imbalanced data," *Integr. Comput. Aided. Eng.*, vol. 16, no. 3, pp. 193–210, 2009.
- [9] H. Han, W. Wang, and B. Mao, "Borderline-SMOTE : A New Over-Sampling Method in," *Int. Conf. Intell. Comput. ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I*, pp. 878–887, 2005.
- [10] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *Proc. Int. Jt. Conf. Neural Networks*, no. 3, pp. 1322–1328, 2008.
- [11] S. Guo, D. Guo, L. Chen, and Q. Jiang, "A centroid-based gene selection method for microarray data classification," *J. Theor. Biol.*, vol. 400, pp. 32–41, 2016.

- [12] J. A. S. Almeida, L. M. S. Barbosa, A. A. C. C. Pais, and S. J. Formosinho, "Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering," *Chemom. Intell. Lab. Syst.*, vol. 87, no. 2, pp. 208–217, 2007.
- [13] J. A. Sáez, B. Krawczyk, and M. Woźniak, "Analyzing the *oversampling* of different classes and types of examples in multi-class imbalanced datasets," *Pattern Recognit.*, vol. 57, pp. 164–178, 2016.
- [14] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine learning Tools and Techniques"*, Fourth Edi. Morgan Kaufmann, 2016.
- [15] H. Brawijaya, S. Samudi, and S. Widodo, "Komparasi Algoritma K-Nearest Neighbor dan Naiive Bayes Pada Pengobatan Penyakit Kutil Menggunakan Cryotheraphy," *JUITA J. Inform.*, vol. 7, no. 2, p. 93, 2019.
- [16] S. Sukanto, Y. Adriyani, and R. Aulia, "Prediksi Kelompok UKT Mahasiswa Menggunakan Algoritma K-Nearest Neighbor," *JUITA J. Inform.*, vol. 8, no. 1, p. 121, 2020.
- [17] N. YE, *Data Mining Theories, Algorithms, and Examples*, vol. 16, no. 4. 2013.
- [18] I. Sutoyo, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik," *J. Pilar Nusa Mandiri*, vol. 14, no. 2, p. 217, 2018.
- [19] G. S. Mahendra and K. Y. E. Aryanto, "SPK Penentuan Lokasi ATM Menggunakan Metode AHP dan SAW," *J. Nas. Teknol. dan Sist. Inf.*, vol. 05, no. 01, pp. 49–56, 2019.
- [20] J. N. Mandrekar, "Receiver operating characteristic curve in diagnostic test assessment," *J. Thorac. Oncol.*, vol. 5, no. 9, pp. 1315–1316, 2010.
- [21] M. E. Rice and G. T. Harris, "Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r," *Law Hum. Behav.*, vol. 29, no. 5, pp. 615–620, 2005.