Analisis Perbandingan Metode Over-Sampling *Adaptive Synthetic-Nominal* (ADASYN-N) dan *Adaptive Synthetic-kNN* (ADSYN-kNN) untuk Data dengan Fitur *Nominal-Multi Categories*

Sri Rahayu[1], Teguh Bharata Adji[2], Noor Akhmad Setiawan[3]
Departemen Teknik Elektro dan Teknologi Informasi
Fakultas Teknik, Universitas Gadjah Mada
Jl. Grafika no.2 Yogyakarta-55281, Indonesia
ayu.ti14@mail.ugm.ac.id[1], adji@ugm.ac.id[2], noorwewe@ugm.ac.id[3]

Abstract—**This paper presented the comparasion of oversampling technique to overcome the imbalanced data problem on the datasets with nominal-multi categories featured between *Adaptive Synthetic-Nominal* (ADASYN-N) and *Adaptive Synthetic*-kNN (ADASYN-kNN) methods. There are 7 datasets with nominal-multi categories featured that have an unbalanced class distribution. The oversampled datasets with both methods are then classified using the Random Forests method. The accuracy between the original datasets and the datasets with ADASYN-N oversampling and ADASYN-kNN techniques are compared.**

Keywords- ADASYN; imbalanced data; nominal; k-NN; multi categories

Abstrak—**Pada penelitian ini disajikan tentang perbandingan teknik *oversampling* untuk mengatasi masalah ketidakseimbangan (*imbalanced*) kelas pada dataset dengan fitur *nominal-multi categories* antara metode *Adaptive Synthetic-Nominal* (ADASYN-N) dengan *Adaptive Synthetic*-kNN (ADASYN-kNN). Terdapat 7 dataset dengan fitur *nominal-multi categories* yang memiliki distribusi kelas yang tidak seimbang. Kemudian dataset yang telah di-*oversampling* dengan kedua metode tersebut dilakukan klasifikasi menggunakan metode *Random Forests*. Selanjutnya dilakukan perbandingan akurasi antara dataset asli dan dataset dengan teknik *oversampling* ADASYN-N serta ADASYN-kNN.**

Kata kunci-ADASYN; imbalanced data; nominal; k-NN; multi categories

I. PENDAHULUAN

Banyak permasalahan data mining, baik pada bisnis, ilmu pengetahuan, kesehatan atau teknik, melibatkan *imbalanced data* (ketidakseimbangan data). Ketidakseimbangan ini sering merupakan bagian integral dari masalah dan hampir pada setiap kasus entitas yang sedikit merupakan yang hal yang paling dibutuhkan.

Dataset dengan ketidakseimbangan kelas ini terjadi karena rasio yang tidak seimbang antara kasus yang satu dengan kasus yang lainnya. Ketidakseimbangan kelas ini akan merugikan pada penelitian bidang datamining karena machine learning pada datamining memiliki kesulitan dalam mengklasifikasikan kelas minoritas (jumlah *instance* yang kecil) dengan benar. Beberapa algoritme mengasumsikan bahwa distribusi kelas yang diuji adalah seimbang sehingga dalam beberapa kasus menjadikan kesalahan dalam mengklasifikasikan hasil pada tiap kelas. Pada algoritme seperti decision tree, nearest neighbor, dan Support Vector Machine (SVM) memiliki prinsip generalisasi data yang diuji sama kedudukannya dan menghasilkan hipotesis yang paling sederhana. Hal ini mengakibatkan error pada klasifikasi kelas minoritas dikarenakan ketidakseimbangan kelas yang cenderung fokus pada kelas mayoritas dan mengabaikan kelas minoritas pada saat klasifikasi.

Terdapat beberapa pendekatan untuk penanganan ketidakseimbangan, salah satunya dengan menggunakan metode sampling data asli. Pendekatan metode sampling yang pertama untuk mengatasi ketidakseimbangan kelas adalah under-sampling yang merupakan metode untuk menyeimbangkan kelas dengan cara mengurangi instance pada kelas mayoritas secara acak. Namun, pada metode *under-sampling* memiliki resiko hilangnya informasi dan data yang dianggap penting untuk proses pengambilan keputusan oleh *machine learning*.

Over-sampling merupakan metode penyeimbangan distribusi kelas dengan mereplikasi instance pada kelas minoritas secara acak. Namun, *over-sampling* meningkatkan kemungkinan munculnya *overfitting* karena menduplikasi *instance* secara sama persis. Chawla dkk [1] mengajukan solusi untuk menangani *overfiiting* pada metode over-sampling yaitu SMOTE (*Synthetic Minority Over-sampling Technique*). SMOTE memanfaatkan *nearest neighbors* dan jumlah *over-sampling* yang diinginkan. SMOTE ini digunakan untuk pendekatan data bertipe numerik.

Selain SMOTE, He, dkk mengajukan metode untuk pendekatan sampling pada pembelajaran dengan dataset tidak seimbang dengan fitur numerik yaitu ADASYN [2]. Ide utama dari ADASYN adalah menggunakan bobot distribusi untuk data pada kelas minoritas berdasarkan pada tingkat kesulitan belajar, dimana data sintesis dihasilkan dari kelas minoritas yang susah untuk belajar dibandingkan dengan data minoritas yang lebih mudah untuk belajar.

Untuk penanganan data dengan fitur nominal, Chawla mengajukan SMOTE-N yang merupakan pengembangan dari SMOTE[1]. Pada SMOTE-N, *nearest neighbor* dihitung menggunakan versi modifikasi dari Value Difference Metric (VDM) yang diajukan oleh Cost dan

Salzberg. Pada penelitian terbaru, Kurniawati [3] mengembangkan ADASYN-N dan ADASYN-KNN yang merupakan pengembangan dari metode ADASYN. ADASYN-N dan ADASYN-KNN ini disebut dapat menangani ketidakseimbangan data dengan fitur nominal. Kekurangan dari penelitian tersebut adalah ADASYN-N maupun ADASYN-KNN baru diuji pada satu dataset dengan kategori biner (misalnya atribut bernilai ya atau tidak) dan diuji klasifikasi menggunakan metode Naïve Bayes Classifier. Kedua metode tersebut kemudian dibandingkan dengan SMOTE-N dan menunjukkan bahwa ADASYN-N dapat meningkatkan akurasi lebih baik dari SMOTE-N sedangkan ADASYN-KNN menunjukkan performa akurasi dari kedua metode tersebut.

Berangkat dari masalah tersebut, maka penelitian ini bertujuan untuk menerapkan metode ADASYN-N dan ADASYN-KNN pada dataset dengan fitur nominal-*categorical* (memiliki kategori lebih dari 2) dan diuji dengan metode klasifikasi yang berbeda yaitu Random Forest, selanjutnya dibandingkan akurasi kedua metode tersebut.

II. METODOLOGI

A. Perhitungan KNN

Untuk menghitung KNN setiap data, perlu dilakukan perhitungan menggunakan persamaan euclidean distance yang tertera pada Persamaan (1):

$$D(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \quad (1)$$

Dengan kasus fitur *nominal multi categories*, maka rumusan *euclidean distance* menjadi Persamaan (2).

$$D(D_1, D_2) = \sqrt{\sum_{k=1}^{n}\left(D_{1,k} - D_{2,k}\right)^2} \quad (2)$$

Pada Persamaan (2), D_1 dan D_2 adalah data yang diukur jarak *euclidean*-nya, k adalah fitur yang terdapat pada data. Pada kasus *nominal multi categories* perhitungan $D_{1,k} - D_{2,k}$ menggunakan persamaan $\delta\left(F_{i,Ca}, F_{i,Cb}\right)$ di mana $F_{i,Ca}$ adalah fitur ke-i dengan kategori a. Selanjutnya, menghitung *distance* tiap fitur menggunakan persamaan (3).

$$\delta\left(F_{i,Ca}, F_{i,Cb}\right) = \sum_{i=1}^{n}\left|\frac{c_{1i}}{c_1} - \frac{c_{2i}}{c_2}\right|^{\mathrm{k}} \quad (3)$$

B. *Adaptive Synthetic* (ADASYN)

ADASYN merupakan metode untuk pendekatan *sampling* pada pembelajaran dengan dataset yang tidak seimbang yang diajukan oleh He, dkk. Ide utama dari ADASYN adalah menggunakan bobot distribusi untuk data pada kelas minoritas berdasarkan pada tingkat kesulitan belajar, sehingga data sintesis dihasilkan dari kelas minoritas yang susah untuk belajar dibandingkan dengan data minoritas yang lebih mudah untuk belajar. ADASYN meningkatkan pembelajaran dengan dua cara. Pertama, mengurangi bias yang diakibatkan oleh ketidakseimbangan kelas dan yang kedua secara adaptif menggeser batas keputusan klasifikasi terhadap kesulitan data.

1. ADASYN-Nominal (ADASYN-N)

ADASYN-N merupakan pengembangan dari ADASYN yang diajukan oleh Kurniawati, Y. E [3] dengan pendekatan data dengan tipe nominal. Nearest neighbor pada ADASYN-N dihitung menggunakan versi modifikasi dari *Value Difference Metric* (VDM) seperti pada SMOTE-N yang diajukan oleh Chawla, dkk [4]. VDM melihat pada nilai fitur yang *overlap* terhadap semua vektor fitur. Matriks mendefinisikan jarak antara nilai fitur yang sesuai untuk vektor fitur yang dibuat.. Berikut prosedur dari *multiclass* ADASYN-N:

Input

(1) Training dataset D_{tr} dengan m sampel $\{x_i, y_i\}, i = 1, \ldots, m$ dimana x_i adalah *instance* dalam n dimensional *feature space* $\boldsymbol{X}$ dan $y_s \in Y = \{1, \ldots, C\}$ adalah label identitas kelas dengan jumlah *instance* terbanyak. Tentukan m_s dan m_l sebagai jumlah *instance* kelas minoritas dan jumlah *instance* kelas mayoritas. Oleh karena itu, $m_{sc} \leq m_l$ dan $\sum m_{sc} + m_l = m$.

Prosedur

(1) Kalkulasi *degree of class imbalance* menggunakan Persamaan (4).

$$d_c = m_{sc}/m_l \quad (4)$$

Di mana $d \in [0,1]$

(2) Jika $d_c < d_{th}$ then (d_{th} adalah penetapan *threshold* untuk derajat tolerasi maksimum dari rasio *imbalance class*):

(a) Hitung jumlah *instance* data sintesis yang perlu di-*generate* untuk kelas minoritas ke-c dengan Persamaan (5).

$$G_c = (m_l - m_{sc}) \times \beta \quad (5)$$

Di mana $\beta \in [0,1]$ adalah parameter yang digunakan untuk menetapkan *level balance* yang diinginkan setelah generalisasi data sintesis. $\beta = 1$berarti data yang sepenuhnya seimbang dibuat setelah proses generalisasi.

(b) Untuk setiap *instance* $x_i \in minority\ class$, temukan k-*nearest neighbors* berdasarkan pada *Euclidean distance* pada n *dimensional space*, dan kalkulasi rasio r_i yangdidefinisikan oleh Persamaan (6).

$$r_i = \frac{\Delta_i}{K}, i = 1, \ldots, m_{sc} \quad (6)$$

Di mana $\Delta_i$adalah jumlah *instance* pada *nearest neighbor* yang termasuk kelas y_s (mayoritas) atau termasuk semua kelas kecuali y_{kc} (minoritas), oleh karena itu $x_i \in [0,1]$ Dimana y_{kc} adalah kelas yang dievaluasi.

(c) Normalisasi r_i dengan Persamaan (7), sehingga $\hat{r}_i$ adalah distribusi kerapatan (*density distribution*) ($\sum_i \hat{r}_i = 1$).

$$\hat{r}_i = r_i / \sum_{i=1}^{m_s} r_i \quad (7)$$

(d) Hitung jumlah dari *instance* data sintesis yang perlu dihasilkan pada setiap *instance* minoritas x_i menggunakan Persamaan (8)

$$g_i = \hat{r}_i \times G_c \quad (8)$$

Dimana G_c adalah total jumlah dari *instance* data sintesis yang perlu untuk dihasilkan untuk kelas minoritas ke- c yang dijelaskan pada Persamaan (5).

(e) Untuk setiap *instance* data kelas minoritas x_i, *generateinstance* data sintesis sebanyak g_i.

2. *Adaptive Synthetic* – KNN (ADASYN-KNN)

ADASYN-KNN merupakan pengembangan dari ADASYN-N dengan pengembangan pada Prosedur (2e) atau prosedur untuk menghasilkan instance data sintesis sebanyak g_i . Pada ADASYN-KNN, data sintesis dihasilkan dari *nearest neighbor instance* yang dievaluasi. Atribut sintesis dihasilkan dengan melakukan voting berdasarkan pada kemunculan atribut dari *nearest neighbor*. Kemudian, instance sintesis yang dihasilkan diduplikasi sebanyak g_i.

Prosedur

(1) Kalkulasi *degree of class imbalance*: persamaan (4)
(2) Jika $d_c < d_{th}$ then (d_{th} adalah penetapan *threshold* untuk derajat tolerasi maksimum dari rasio *imbalance class*): (Prosedur 2a sampai 2d sama dengan ADASYN-N)
(e) Untuk setiap *instance* data kelas minoritas x_i, *generate instance* data sintesis berdasarkan pada langkah berikut:
i. Cari *nearest neighbor* dari *instance* data kelas minoritas x_i.
ii. Lakukan *majority voting* untuk setiap atribut pada *instance nearest neighbor*.
iii. Hasilkan *instance* baru dengan atribut berdasarkan pada *majority voting*.
iv. Duplikasi *instance* baru sebanyak g_i.

C. *Random Forest*

Random forest pertama kali dikenalkan oleh Breiman pada Tahun 2001 [5]. Dalam penelitiannya menunjukkan kelebihan random forest antara lain dapat menghasilkan error yang lebih rendah, memberikan hasil yang bagus dalam klasifikasi, dapat mengatasi data training dalam jumlah sangat besar secara efisien, dan metode yang efektif untuk mengestimasi missing data.

Menentukan ketepatan klasifikasi dengan metode *Random Forest* [6]:

1. Menentukan *m* jumlah variabel prediktor yang diambil secara acak dan *k* pohon yang akan dibentuk untuk digunakan dalam klasifikasi random forest. Nilai k yang digunakan adalah 100. Umumnya k = 50 sudah memberikan hasil yang memuaskan untuk masalah klasifikasi [7]. Sementara itu k ≥ 100 cenderung menghasilkan tingkat misklasifikasi yang rendah.
2. Mengambil *n* sampel dengan teknik *resampling* dengan pengembalian sehingga diperoleh dataset baru D*
3. Membentuk *tree* model dari dataset D* dengan kombinasi *m* variabel prediktor yang diambil secara acak dan *k* buah ukuran pohon.
4. Melakukan voting mayoritas untuk setiap kali pohon.
5. Menentukan akurasi ketepatan klasifikasi.

III. JALANNYA PENELITIAN

Data yang diolah pada penelitian ini berupa 7 dataset dari sumber UCI-datasets dengan rincian seperti pada Tabel I.

TABEL I. DETAIL DATASET

Dataset	*Instances*	*Kelas*	*Distribusi Kelas*
Audiology	26	6	mixed_cochlear_age_fixation 1 cochlear_age 11 normal_ear 2 cochlear_poss_noise 4 cochlear_age_and_noise 4 mixed_cochlear_unk_fixation 4
Balance-Scale	626	3	L 288 B 49 R 288
Breast-Cancer	286	2	no-recurrence-events 201 recurrence-events 85
Car	1728	4	unacc 1210 acc 384 good 69 vgood 65
Lenses	24	3	hard-contact-lenses 4 soft-contact-lenses 5 no-contact-lenses 15
Lymphography	148	4	normal 2 metastases 81 malign_lymph 61 fibrosis 4
Nursery	12960	5	not_recom 4320 recommend 2 very_recom 328 priority 4266 spec_prior 4044

Untuk penelitian ini dilakukan proses *oversampling* pada setiap kelas minoritas dari masing-masing dataset yang bertujuan agar jumlah *instance* pada kelas minoritas dapat mendekati atau sama dengan jumlah *instance* kelas mayoritas untuk menyeimbangkan jumlah *instance* dalam semua kelas. Dari proses *oversampling* dihasilkan *instance* sintesis untuk setiap kelas pada kelas minoritas. Data hasil *oversampling* baik dengan algoritme

ADASYN-N maupun ADASYN-KNN, kemudian digabung dengan dataset asli sehingga membentuk dataset baru.

Dataset baru yang dihasilkan teknik *oversampling* ADASYN-N dan ADASYN-KNN kemudian diuji dengan menggunakan metode klasifikasi *Random Forests*. Implementasi dengan *classifier* tersebut dilakukan menggunakan 10-*Cross Fold Validation*. Yang dimaksud dengan 10-*Cross Fold Validation,* yaitu membagi dataset menjadi 10 bagian, dimana satu bagian akan menjadi *testing set* dan sembilan bagian sisanya digunakan sebagai *training set*, hal ini dilakukan bergantian sebanyak sepuluh kali.

Selanjutnya, akurasi hasil klasifikasi Random Forest dibandingkan antara dataset asli dengan dataset hasil oversampling ADASYN-N dan ADASYN-KNN. Hasil komparasi tersebut ditampilkan dalam Tabel II.

TABEL II. HASIL AKURASI KLASIFIKASI

Dataset	Dataset Asli	ADASYN-N	ADASYN-KNN
Audiology	80.8%	98.5%	87.7%
Balance-Scale	81.8%	90.5%	89.4%
Breast-Cancer	69.6%	83.7%	70.1%
Car	94.7%	99.1%	98.6%
Lenses	70.8%	93.5%	89.1%
Lymphography	81.1%	93.1%	91.8%
Nursery	99.1%	99.4%	99.3%

Performa teknik ADASYN-N maupun ADASYN-kNN dapat diketahui melalui uji hipotesis dengan menggunakan uji paired T-test dengan level signifikan 95%. Adapun hipotesis yang akan diuji adalah sebagai berikut:

H0 = teknik ADASYN-N maupun ADASYN-kNN tidak meningkatkan akurasi klasifikasi pada dataset dengan fitur *nominal-multi categories*

H1 = teknik ADASYN-N maupun ADASYN-kNN meningkatkan akurasi klasifikasi pada dataset dengan fitur *nominal-multi categories*.

Sebelumnya dilakukan analisis statistik deskriptif dari data pada Tabel dan hasilnya ditunjukkan oleh Tabel III.

TABEL III. HASIL STATISTIK DESKRIPTIF

Descriptive Statistics

	N	Mean		Std. Deviation	Variance
	Statistic	Statistic	Std. Error	Statistic	Statistic
Dataset Asli	7	82.557	4.1777	11.0530	122.170
ADASYN-N	7	93.971	2.1537	5.6982	32.469
ADASYN-kNN	7	89.429	3.6641	9.6943	93.979
Valid N (listwise)	7				

Selanjutnya, dilakukan pengujian Paired Sample t-Test dengan perbandingan antara hasil akurasi klasifikasi pada dataset asli dengan hasil akurasi pada dataset dengan teknik ADASYN-N, begitu pula antara dataset asli dengan dataset dengan teknik ADASYN-kNN. Hasil pengujian ditunjukkan pada Tabel IV.

TABEL IV. HASIL UJI PAIRED T-TEST

Paired Samples Test

			Pair 1	Pair 2
			Dataset Asli - ADASYN-N	Dataset Asli - ADASYN-kNN
Paired Differences	Mean		-11.4143	-6.8714
	Std. Deviation		7.6869	6.3211
	Std. Error Mean		2.9054	2.3891
	95% Confidence Interval of the Difference	Lower	-18.5235	-12.7174
		Upper	-4.3051	-1.0254
t			-3.929	-2.876
df			6	6
Sig. (2-tailed)			.008	.028

Hasil pengujian Paired Sample t-Test menunjukkan signifikansi antara dataset asli dengan dataset hasil teknik *oversampling* ADASYN-N adalah 0,008 atau < 0,05. Begitu pula signifikansi antara dataset asli dengan dataset hasil teknik *oversampling* ADASYN-kNN juga < 0,05 yaitu 0,028. Kedua hasil pengujian tersebut menolak H0 dan menerima H1, yaitu bahwa teknik oversampling ADASYN-N maupun ADASYN-kNN dapat meningkatkan akurasi klasifikasi pada dataset dengan fitur *nominal-multi categories*. Selain itu, hasil pengujian di atas juga menunjukkan bahwa teknik oversampling ADASYN-N menunjukkan performa yang lebih baik daripada teknik ADASYN-kNN dengan nilai signifikan 0,008 < 0,028.

IV. KESIMPULAN

Dari hasil perbandingan dan pembahasan di atas, dapat disimpulkan bahwa teknik *oversampling* dengan ADASYN-kNN menunjukkan peningkatan akurasi yang cukup signifikan dari dataset asli yang belum dilakukan proses *resampling*. Sedangkan teknik ADASYN-N menunjukkan akurasi yang lebih baik dari ADASYN-kNN dalam mengatasi ketidakseimbangan distribusi kelas pada data dengan fitur *nominal-multi categories*. Berbeda dengan hasil penelitian sebelumnya yang menunjukkan bahwa akurasi pada dataset dengan teknik ADASYN-kNN lebih baik daripada teknik ADASYN-N dalam penanganan data dengan fitur *nominal-binary* (hanya terdapat dua kategori pada masing-masing fitur).

REFERENCES

[1] N. Chawla and K. Bowyer, "SMOTE: Synthetic Minority Over-sampling Technique Nitesh," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[2] H. He and Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*, 1st ed. Wiley-IEEE Press, 2013.

[3] Y. E. Kurniawati, "Multilass Imbalanced Learning dengan Synthetic Minority Over Sampling Technique (SMOTE)

untuk Klasifikasi Hasil Tes Pap Smear," Tesis pada Departemen Teknik Elektro dan Teknologi Informasi, Fakultas Teknik, Universitas Gadjah Mada, 2017.

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

[5] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[6] Nidhomuddin and B. W. Otok, "Random Forest Dan Multivariate Adaptive Regression Spline (Mars) Binary Response Untuk Klasifikasi Penderita Hiv / Aids Di Surabaya," Statistika, Vol. 1, No. 3, Mei 2015.

[7] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 5, pp. 1–35, 1999.