INSTITUT KEGURUAN DAN ILMU PENDIDIKAN
WIDYA DARMA SURABAYA

STATUS : " TERAKREDITASI "

Jl. Ketintang 147 – 151 Telp. / Fax : 031 – 827 3446 Surabaya 60243

Email : ikip.widya.darma@gmail.com. Website : www.ikipwidyadarma.ac.id.

FPBS : Jurusan / Program Studi
Pendidikan Bahasa dan Sastra Indonesia
Pendidikan Bahasa Inggris

FPMIPA : Jurusan / Program Studi
Pendidikan Matematika

FPIPS : Jurusan / Program Studi
Pendidikan Pancasila dan Kewarganegaraan
Pendidikan Ekonomi

SURAT PERNYATAAN

Yang bertanda tangan di bawah ini, Pimpinan Perguruan Tinggi :

Nama	: Dr. Hari Purwanto, SE., MM.
NIP	: 195402241986061001
Pangkat/Golongan Ruang	: IVb / Pembina Tingkat I
Jabatan Fungsional	: Lektor Kepala
Jabatan	: Rektor
Unit Kerja	: IKIP WIDYA DARMA SURABAYA

Dengan ini menyatakan bahwa dokumen pelaksanaan **Penelitian dan Karya Ilmiah Dosen** dalam pengajuan jabatan akademik ini telah dilakukan secara plagiasi secara daring (online).

Jika dikemudian hari ternyata ditemukan data, informasi dan berkas yang tidak benar maka saya bertanggungjawab sepenuhnya dan bersedia diberikan sanksi administrasi oleh Kementerian Riset, Teknologi dan Pendidikan Tinggi atau Kementerian/Lembaga lain yang berwenang. Selain itu, jika ternyata di kemudian hari ditemukan hal-hal yang berimplikasi terhadap masalah hukum, saya bertanggungjawab penuh dan tidak melibatkan pihak lain, baik secara personal maupun kelembagaan.

Demikian pernyataan ini dibuat tanpa paksaan atau tekanan dari pihak lain.

Surabaya, 30 Juli 2018

Rektor

METERAI TEMPEL 2A0D0AEF496477842 6000 ENAM RIBU RUPIAH

Dr. HARI PURWANTO. SE, MM

NIP. 19540224 198606 1 001

Klasifikasi Multi Class Imbalanced Data “Smote Spport Vector Machine” Untuk Diagnosis Penyakit Kanker

by Hani Khaulasari

Prosiding Seminar Nasional Hasil Penelitian dan Pengabdian dengan tema: ((Inovasi, Teknologi dan Pendidikan Guna Mewujudkan Indonesia Sejahtera di Era Industrialisasi 4.0), 21 Desember 2018, hal: 187-193 ISSN: 978-602-5793-40-0

PROSIDING SEMINAR NASIONAL HASIL PENELITIAN DAN PENGABDIAN KEPADA MASYARAKAT (*Tema: Sains dan Kesehatan Industrialisasi 4.0*), 21 Desember 2018, (hal:1-5)

Artikel Ilmiah (Hasil Penelitian)

KLASIFIKASI MULTI CLASS IMBALANCED DATA "SMOTE SUPPORT VECTOR MACHINE" UNTUK DIAGNOSIS PENYAKIT KANKER

Hani Khaulasari[1]
Jurusan Pendidikan Matematika, IKIP Widya Darma Surabaya[1]
Email : hanikhaulasari@gmail.com

ABSTRAK

Analisis Klasifikasi adalah proses menemukan model terbaik dari *classifier* untuk memprediksi kelas dari suatu objek atau data yang label kelasnya tidak diketahui. Pada kehidupan nyata, khususnya di bidang medis sering kali ditemui klasifikasi *multi class* dengan kondisi himpunan data *imbalanced*. Kondisi *imbalanced* data menjadi masalah dalam klasifikasi *multi class* karena mesin *classifier learning* akan condong memprediksi ke kelas data yang banyak (mayoritas) dibanding dengan kelas minoritas. Akibatnya, dihasilkan akurasi prediksi yang baik terhadap kelas data *training* yang banyak (kelas mayoritas) sedangkan untuk kelas data *training* yang sedikit (kelas minoritas) akan dihasilkan akurasi prediksi yang buruk. Oleh Karena itu, pada penelitian ini akan diterapkan metode *SMOTE Support Vector Machine* untuk klasifikasi *multi class imbalanced* dengan menggunakan data kanker. Data yang digunakan adalah data kanker tiroid, kanker payudara dan kanker serviks. Percobaan tersebut menggunakan q-fold cross validation (q=5) dan (q=10). SVM *One Againt One* (OAO) digunakan untuk klasifikasi *multi class*. Optimasi parameter fungsi kernel RBF σ dan C. Hasil menunjukan bahwa metode yang terbaik untuk digunakan dalam memprediksi status pasien penderita kanker tiroid, kanker payudara dan kanker serviks adalah metode *SMOTE Support Vector Machine* dengan (q=5).

Kata kunci: *Imbalanced Data,* SMOTE, SVM

ABSTRACT

Classification analysis is the process of finding the best model of a classifier for predicting the class of an object or data class label is unknown. In the real life, especially in the medical field often encountered multi-class classification with imbalanced data sets conditions. Imbalanced condition of the data at issue in multi-class classification as machine learning classifier will be inclined to predict that a lot of data classes (the majority) compared with a minority class. As a result, generated a good prediction accuracy of the data class training that many (the majority class), while for class training data bit (the minority) will produce a poor prediction accuracy. Hence, this research will apply the method Smote SVM. SVM for multi-class classification imbalanced using cancer data. The data used is data thyroid cancer, breast cancer and cervical cancer. The experiment using a q-fold cross validation (q = 5) and (q = 10). SVM One againt One (OAO) is used for multi-class classification. Parameter optimization RBF kernel function (σ) and C. Results showed that the best method to use in predicting the status of patients with thyroid cancer, breast cancer and cervical cancer is the SMOTE Support Vector Machine method with (q=5).

Keyword:ImbalancedData,SMOTE,SVM

PENDAHULUAN

Analisis Klasifikasi adalah proses menemukan model terbaik dari *classifier* untuk memprediksi kelas dari suatu objek atau data yang label kelasnya tidak diketahui. Pada kehidupan nyata, khususnya di bidang medis sering kali ditemui klasifikasi *multi class* dengan kondisi himpunan data *imbalanced*. Kondisi *imbalanced* data menjadi masalah dalam klasifikasi *multi class* karena mesin *classifier learning* akan condong memprediksi ke kelas data yang banyak (mayoritas) dibanding dengan kelas minoritas. Akibatnya, dihasilkan akurasi prediksi yang baik terhadap kelas data *training* yang banyak (kelas mayoritas) sedangkan untuk kelas data *training* yang sedikit (kelas minoritas) akan dihasilkan akurasi prediksi yang buruk. Beberapa strategi telah diusulkan untuk menangani masalah ini, salah satunya adalah teknik *Synthetic Minority Oversampling Technique* (SMOTE) yang diusulkan oleh [1] sebagai skema *preprocessing* untuk dataset yang tidak seimbang.

Setelah langkah *preprocessing* dilakukan, metode klasifikasi dapat diterapkan pada data yang sudah diolah sebelumnya. Pada peelitian ini, Support Vector Machines (SVM) digunakan sebagai teknik klasifikasi yang mendasari. SVM adalah *state-of-the-art* di bidang teori pembelajaran statistik yang mengikuti prinsip *Structural Risk Minimization* (SRM). Metode ini bekerja dengan menemukan classifier yang dapat memisahkan observasi kedalam kelas yang berbeda. Menurut teori *Structural Risk Minimization* (SRM), SVM telah memperlihatkan performa sebagai metode yang bisa mengatasi masalah *overfitting* dengan cara meminimalkan batas atas pada *generalization error*, yang menjadi alat yang kuat untuk *supervised learning*. SVM dapat menangani sampel besar atau kecil, *nonlinier*, *high dimensional*, *over learning* dan masalah lokal minimum [2].

Penelitian sebelumnya tentang klasifikasi menggunakan SVM yaitu dilakukan oleh [3] menggunakan klasifikasi SVM untuk diagnosis *breast cancer,*hasilnya menunjukan bahwa SVM menghasilkan akurasi yang tinggi. Penelitian oleh [4] membandingkan klasifikasi data dengan regresi logistik ordinal dan SVM, hasilnya menunjukan bahwa SVM memiliki ketepatan klasifikasi yang lebih baik dibandingkan regresi logistik ordinal. Penelitian oleh [5] melakukan klasifikasi dengan Combine Sampling (SMOTE+Tomek Link) SVM, hasil menunjukan bahwa metode Combine memberikan hasil performansi yang tinggi pada beberapa dataset tetapi tidak pada dataset kanker serviks, untuk dataset kanker serviks metode yang menghasilkan performansi terbaik adalah SMOTE SVM.

Pada penelitian ini, SVM diterapkan pada data yang tidak seimbang setelah metode SMOTE digunakan sebagai *preprocessing*. Strategi prediksi multiclass dengan menggunakan *One-Against-One (OAO)*. Metode yang diusulkan diterapkan pada tiga dataset yang tidak seimbang (imbalanced data): (i) kanker serviks, (ii) kanker payudara, (iii) tiroid.

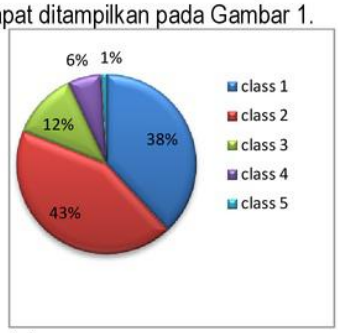
METODE

Data yang digunakan dalam penelitian ada 3 dataset diagnosis penyakit kanker yaitu kanker payudara , kanker serviks dan kanker tiroid. Dataset diagnosis kanker payudara dan kanker serviks diperoleh dari sebuah rumah sakit di Surabaya dan dataset diagnosis kanker tiroid diperoleh dari *UCI Machine Learning Repository*. Data Kanker payudara terdiri dari 6 variabel dan 178 observasi, datakanker serviks terdiri dari 7 variabel

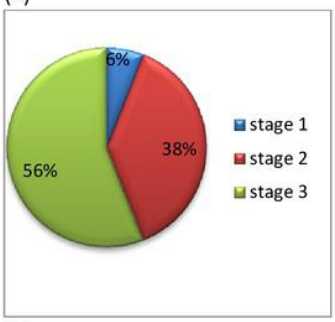
Prosiding Seminar Nasional Hasil Penelitian dan Pengabdian dengan tema: ((Inovasi, Teknologi dan Pendidikan Guna Mewujudkan Indonesia Sejahtera di Era Industrialisasi 4.0), 21 Desember 2018, hal: 187-193 ISSN: 978-602-5793-40-0

dan 794 observasi dan data kanker tiroid terdiri dari 5 variabel dan 215 observasi.

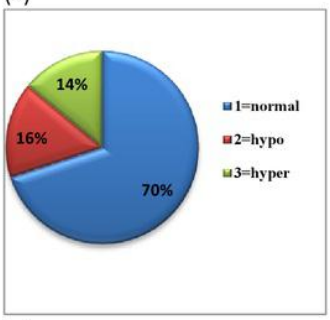
Distribusi kelas label pada dataset dapat ditampilkan pada Gambar 1.



(a)



(b)



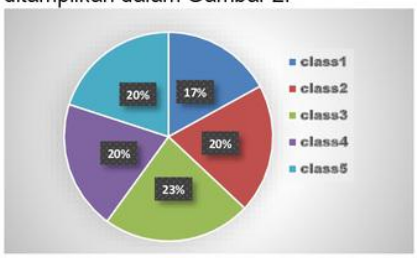
(c)

Gambar 1 Dataset distribusi Kelas (a) Kanker Serviks; (b) kanker payudara; (c) kanker tiroid

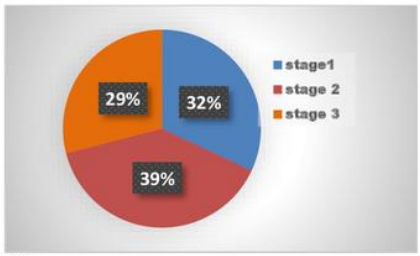
Pada penelitian ini membandingkan performa metode SVM data asli dan SVM data *preprocessing* dengan SMOTE. Klasifikasi menggunakan q-fold Cross Validation (q=5) dan (q=10). Performa klasifikasi berdasarkan Akurasi, *Sensitivity*, *Specificity*, *Precision*. Komputasi menggunakan Matlab R2009.

HASIL

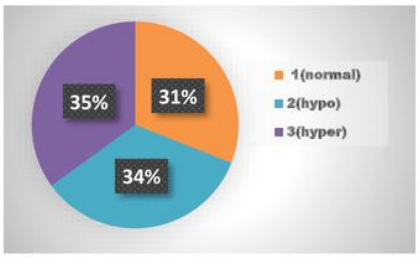
Metode SMOTE merupakan metode *oversampling* yang digunakan untuk meningkatkan jumlah kelas minoritas dengan mereplikasi data secara acak sesuai dengan persentase yang diinginkan sehingga jumlahnya mendekati jumlah data mayor [1] dan [5]. Penerapan metode *oversampling* pada data *imbalanced* menyebabkan tingkat *imbalanced* data semakin kecil dan klasifikasi dapat dilakukan dengan tepat. Hasil dari penanganan metode SMOTE terhadap masing-masing data *imbalanced* yang digunakan dalam dalam penelitian ini ditampilkan dalam Gambar 2.



(a)



(b)



(c)

Gambar 2 Dataset distribusi Kelas SMOTE (a) Kanker Serviks; (b) kanker payudara; (c) kanker tiroid

Pada klasifikasi SVM, parameter kernel RBF (σ) dan (C) ,dicobakan dengan $(\sigma = 1, 10, 20)$ dan parameter (C) yang dicobakan yaitu $(C = 1, 50, 100)$. Hasil akurasi, *Sensitivity, Specificity, Precision* klasifikasi SMOTE SVM dan SVM dari *training* dan *testing* dapat dilihat pada Tabel 1-2.

Tabel 1. Akurasi Klasifikasi dan Waktu Komputasi (Q=5 dan Q=10)

Data	**(i) (q=5)**	**(i) (q=10)**	**(ii) (q=5)**	**(ii) (q=10)**
	Hyperparameters (C ; σ)			
	Akurasi Training (%)			
	Akurasi Testing (%)			
	Waktu Komputasi (detik)			
A	(100;1)	(100;1)	(100;1)	(100;1)
	82	90,02	93,88	93,95
	40,30	39,16	**59,41**	59,41
	(3,6)	(3,79)	**(3,5)**	(3,6 s)
B	(100;1)	(100;1)	(100;1)	(100;1)
	90,25	95,37	95,276	95,07
	62,35	62,44	**90,43**	90,27
	(0,29)	(0,28)	**(0,24)**	(0,26)
C	(100;1)	(100;1)	(100;1)	(100;1)
	92,88	100	99,95	100
	65,25	65,70	**90,56**	90,56
	(1,2)	(1,13)	**(0,27)**	(0,29s)

Keterangan :

a) Kanker Serviks
b) Kanker Payudara
c) Kanker Tiroid

i) SVM
ii) SMOTE SVM

Tabel 2. *Sensitivity, Specificity, Precision* Klasifikasi (Q=5 dan Q=10)

Data	**(i) (q=5)**	**(i) (q=10)**	**(ii) (q=5)**	**(ii) (q=10)**
	Hyperparameters (C ; σ)			
	Sensitivity Testing (%)			
	Specificity Testing (%)			
	Precision Testing (%)			
A	(100;1)	(100;1)	(100;1)	(100;1)
	41,15	39,02	59,44	**59,56**
	40,30	39,16	**59,41**	59,41
	41,39	39,27	**59,62**	59,61
B	(100;1)	(100;1)	(100;1)	(100;1)
	62,25	62,37	**90,27**	90,25
	62,35	62,44	**90,13**	90,12
	62,37	62,45	**90,16**	90,12
C	(100;1)	(100;1)	(100;1)	(100;1)
	64,39	65,73	**90,39**	90,25
	64,25	65,70	**90,28**	90,28
	65,23	66,21	**90,33**	90,32

Keterangan :

a) Kanker Serviks
b) Kanker Payudara
c) Kanker Tiroid

i) SVM
ii) SMOTE SVM

Tabel 3 Akurasi Klasifikasi dengan Regresi Logistik (RL) dan Analisis Disriminan (AD) Sebelum dan Sesudah SMOTE pada Testing Data

Data	**Method RL**		**Method AD**	
	(i) (q=5) (q=10)	**(ii) (q=5) (q=10)**	**(i) (q=5) (q=10)**	**(ii) (q=5) (q=10)**
	44,97%	**46,5%**	35,5%	**45,1%**
A	45,96%	**47,5%**	35,9%	**45,4%**
	83,26%	**89,2%**	87,5%	**87,6%**
B	83,36%	**89,2%**	87,8%	**88,4%**
	85,5%	**87,7%**	85,2%	**86,1%**
C	85,7%	**88,7%**	85,9%	**86,2%**

Keterangan :

a) Kanker Serviks
b) Kanker Payudara
c) Kanker Tiroid

i. Sebelum SMOTE
ii. Sesudah SMOTE

Pada Tabel 1 dan Tabel 2 dapat disimpulkan bahwa metode SMOTE SVM dengan (q=5), parameter C=100 dan $(\sigma) = 1$ menghasilkan performa klasifikasi yang terbaik pada tiga dataset.

Pada Tabel 3 dapat disimpulkan bahwa metode SMOTE dapat meningkatkan akurasi (performansi) pada semua dataset dengan metode klasik yang telah dicobakan, yaitu metode regresi logistik dan analisis diskriminan. Akan tetapi metode klasik tersebuh masih memiliki performansi yang lebih rendah dibandingkan dengan menggunakan metode SVM.

SIMPULAN

Metode yang terbaik untuk kasus klasifikasi *imbalanced* dalam memprediksi status pasien penderita kanker tiroid, kanker payudara dan kanker serviks adalah metode *SMOTE Support Vector Machine*.

UCAPAN TERIMA KASIH

Ucapan terimakasih disampaikan kepada Jurusan Statistika ITS atas kerjasamanya dalam proses pengerjaan penelitian ini dan kepada Institusi kampus IKIP Widya Darma Surabaya atas supportnya untuk menjadi pemakalah.

DAFTAR RUJUKAN

[1] Chawla N.V, Bowyer K.W, Hall L.O, dan Kegelmeyer, W.P., 2002,"SMOTE: Synthetic Minority Oversampling Technique", *Journal of Artificial Intelligence Research*,Vol.16, Hal.321-357.

[2] Guo, J.,Yi, Ping.,Wang, R., Ye, Qiaolin.,Zhao, Chunxia. "Feature Selection for Least Sqaure Projection Twin Support Vector Machine". *Neurocomputing*, Vol. 14, Hal. 174-183.

[3] S. Zhang, C. Zhu, J. K. O. Sin, and P. Cheng, Hui-Ling., B. Yang, J.Liu dan D.Y.Liu, 2011, "A Support Vector Machine Classifier with rough set based feature selection for breast cancer diagnosis". *Expert System with Application*, Vol. 38, No 7, Hal 9014-9022.

[4] Rahman, Farizi dan Santi,W.Purnami., 2012, "Perbandingan Klasifikasi Tingkat Keganasan Breast Cancer Dengan Menggunakan Regresi Logistik Ordinal Dan Support Vector Machine (SVM)". *Jurnal SAINS dan Seni ITS*, Vol.1, No.1, (September 2012) ISSN : 2301-928X

[5] Sain, Hartayuni dan Santi, W. Purnami., 2015, "Combine Sampling Support Vector Machine For Imbalanced Data Clasification". *Jurnal Science Direct, Procedia Computer Science*, Vol.72, Hal. 59-66.

Klasifikasi Multi Class Imbalanced Data “Smote Spport Vector Machine” Untuk Diagnosis Penyakit Kanker

ORIGINALITY REPORT

30%	29%	5%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	repository.its.ac.id Internet Source	20%
2	addi.ehu.es Internet Source	2%
3	Hartayuni Sain, Santi Wulan Purnami. "Combine Sampling Support Vector Machine for Imbalanced Data Classification", Procedia Computer Science, 2015 Publication	1%
4	se.aist-nara.ac.jp Internet Source	1%
5	docobook.com Internet Source	1%
6	Submitted to Higher Education Commission Pakistan Student Paper	1%
7	albertusgarut4.blogspot.com Internet Source	1%

8	www.ijritcc.org Internet Source	1%
9	www.bdigital.unal.edu.co Internet Source	1%
10	blokimia.blogspot.com Internet Source	<1%
11	id.scribd.com Internet Source	<1%
12	repository.ipb.ac.id Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		