COMBINE SAMPLING LEAST SQUARE SUPPORT VECTOR MACHINE UNTUK KLASIFIKASI MULTI CLASS IMBALANCED DATA

Oleh:

HANI KHAULASARI

IKIP Widya Darma

Abstrak : Kondisi *imbalanced* data menjadi masalah dalam klasifikasi *multi class*karena mesin *classifierlearning* 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 Combine Sampling (SMOTE+Tomek Links) LS-SVM untuk klasifikasi multi classimbalanced dengan menggunakan data medis. Data yang digunakan adalah data thyroid, kanker payudara dan kanker serviks. Percobaan tersebut menggunakan q-fold cross validation (q=5). LS-SVM *One Againt One* (OAO) digunakan untuk klasifikasi *multi class*.Optimasi parameter fungsi kernel RBF σ dan C menggunakan PSO-GSA. Hasil menunjukan bahwa klasifikasi dengan menggunakan penanganan*imbalanced* data (SMOTE, Tomek Links dan Combine Sampling) dapat meningkatkan nilai akurasi. Metode yang terbaik untuk digunakan dalam memprediksi status pasien penderita Thyroid, kanker payudara dan kanker serviks adalah metode *combine Sampling Least Square Support Vector Machine PSO-GS*A.

Kata Kunci : *Imbalanced data, LS-SVMMulticlass, SMOTE, Tomek Links, Combine Sampling, PSO-GSA.*

PENDAHULUAN

Klasifikasi merupakan salah satu bidang kajian dalam *machine learning*. Analisis Klasifikasi adalah proses menemukan model terbaik dari *classifier* untuk memprediksi kelas dari suatu objek atau data yang label kelasnya tidak diketahui (Han dan Kamber, 2001).

Selama dekade terakhir ini telah banyak metode *machine learning* yang dikembangkan untuk membantu klasifikasi tanpa terikat oleh asumsi dan memberikan fleksibilitas analisis data yang lebih besar tetapi tetap menghasilkan tingkat akurasi yang tinggi dan mudah dalam penggunaannya. Salah satu diantaranya adalah *Support Vector Machine* (SVM). Metode *Support Vector Machine* (SVM) merupakan metode *machine learning* yang baru, sangat berguna dan sangat berhasil dalam melakukan prediksi, baik dalam kasus klasifikai maupun regresi. Prinsip dasar SVM adalah *linier classifier* dan selanjutnya dikembangkan untuk masalah *nonlinier* dengan memasukkan konsep *kernel trick* pada ruang kerja berdimensi tinggi (Cortez dan Vapnik, 1995).

Penelitian sebelumnya tentang klasifikasi menggunakan SVM yaitu Rahman dan Purnami (2012) membandingkan klasifikasi data dengan regresi logistik ordinal dan SVM, hasilnya menunjukan bahwa SVM memiliki ketepatan klasifikasi yang lebih baik dibandingkan regresi logistik ordinal. Novianti dan Purnami (2012)membandingkan klasifikasi SVM dengan regresi logistik, hasil menunjukan bahwa akurasi klasifikasi dengan SVM lebih baik daripada regresi logistik. Haerdle, Prastyo dan Hafner (2014)melakukan prediksi kegagalan peminjaman kredit bank dengan membandingkan antara SVM, analisis diskriminan, probit dan logit, hasil menunjukan bahwa klasifikasi SVM lebih baik dibandingan dengan metode klasifikasi yang lain.

Pada SVMterdapat *quadratic programming* yang merupakan suatu kompleksitas komputasi dari algoritma SVM yang biasanya intensif untuk digunakan, karena dengan *quadratic programming* dapat diperoleh solusi optimal dalam menentukan variabel lagrange yang nantinya digunakan dalam perhitungan nilai bobot dan bias. *Quadratic programming*tidak efisien apabila diterapkan pada dimensi ruang yang lebih tinggi, oleh karena itu usulan metode yang diberikan untuk penelitian ini adalah menggunakan *Least Square Support Vector Machine* (LS-SVM).

LS-SVM merupakan pengembangan dari metode SVM. Jika SVM dikarakteristikkan dengan permasalahan *quadratic programming* dengan fungsi *constrain* berupa pertidaksamaan, LS-SVM sebaliknya, diformulasikan dengan menggunakan fungsi *constrain* yang hanya berupa persamaan. Sehingga solusi LS-SVM dihasilkan dengan menyelesaikan persamaan linier. Hal ini tentu berbeda dengan SVM dimana solusinya dihasilkan melalui penyelesaian *quadratic programming*.

Pada kehidupan nyata, khususnya di bidang medis seringkali ditemui klasifikasi dalam kasus *multi class*. Klasifikasi SVM dan LS-SVM yang semula untuk *binary class* akan dimodifikasi dengan menggunakan pendekatan *multi class.* Ada beberapa pendekatan yang sering digunakan untuk kasus *multi class*. Klasifikasi *multi class* Trapsilasiwi (2013), dengan pendekatan *One Againt One* (OAO) lebih baik dibandingkan dengan menggunakan pendekatan *One Againt All* (OAA).

Klasifikasi *multi class* seringkali ditemui kondisi himpunan data *imbalanced*. *Imbalanced data* merupakan kondisi data yang tidak berimbang antara kelas data satu dengan kelas data yang lain. Kelas data yang banyak merupakan kelas mayoritas atau kelas negatif sedangkan kelas data yang sedikit merupakan kelas minoritas atau kelas positif. Kondisi *imbalanced* data menjadi masalah dalam klasifikasi 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 (Chawla dkk, 2002),

Salah satu pendekatan metode *learning* untuk mengatasi masalah *imbalanced data*adalah *Sampling based approaches.* Metode *Sampling based approaches* dibedakan menjadi 2 yaitu *oversampling* dan *undersampling*. Salah satu metode *oversampling* adalah SMOTE (*Synthetic Minority Oversampling Technique*) yang diperkenalkan pertama kali oleh (Chawla dkk, 2002). Pendekatan ini bekerja dengan *synthetic* data yaitu data replikasi dari data minor. Pendekatan *oversampling* dilakukan dengan cara mereplikasi data minor sehingga tidak mengurangi banyak informasi. Metode *undersampling* dilakukan dengan cara mengurangi jumlah data kelas mayor. Akan tetapi, masalah yang ditimbulkan adalah banyak data yang dihilangkan, yang mengandung informasisehingga efektifitas klasifikasi menurun.Salah satu metode *undersampling*adalah Tomek Links (Chawla dkk, 2002),

Penelitian ini mengadopsi dari Trapsilasiwi (2013), menerapkan metode SMOTE *Least Square SVM* (LS-SVM*)* PSO-GSA pada data medis untuk menangani masalah *imbalanced data multi class*dengan menggunakan 5-*fold cross validation*.Hasil menunjukan bahwa metode SMOTE LS-SVM PSO-GSA lebih baik dibandingkan dengan metode LS-SVM tanpa adanya penambahan SMOTE dan PSO GSA. Akan tetapi, metode SMOTE LS-SVM ini belum memberikan hasil yang memuaskan pada kedua data percobaan. Pada hasil terlihat kalau masih terjadi *missclassification* yang cukup besar dan

overfitting. Akurasi tertinggi SMOTE LS-SVM PSO-GSA pada data kanker serviks sekitar 59,4%. jauh lebih rendah daripada akurasi data kanker payudara sekitar 96,9%. Akurasi pada *training*jauh lebih tinggi daripada akurasi pada *testing*atau terjadi *overfitting*. Oleh karena itu, dilakukan perbaikan klasifikasi pada tahap *preprocessing* menangani *imbalanced* data.

Batista dkk (2003)menggunakan metode gabungan *undersampling* dan *oversampling* (SMOTE+Tomek Links) pada klasifikasi masalah pengkajian protein dalam bioinformatika dengan *decision tree*. Penggunaan metode SMOTE+Tomek Links merepsentasikan hasil yang sangat baik untuk masalah *imbalanced* data.Sain (2013) menerapkan metode *Combine Sampling* (SMOTE+Timek Link) dengan metode SVM 5-*fold cross validation*, yang diterapkan pada data medis. Hasil dari Sain (2013) menunjukan bahwa dengan metode *combine sampling* (SMOTE+Tomek Links) secara umum lebih baik dari metode SMOTE dan Tomek Links.

Tujuan dari penelitian ini adalah melakukan penanganan mengenai persoalan *imbalanced data* pada kasus *multiclass* menggunakan algoritma *Combine Sampling* dan mendapatkan metode klasifikasi LS-SVM yang terbaik. Pada penelitian ini, dalam penentuan parameter menggunakan cara *trial error* dan optimasi PSO-GSA. Metode tersebut akan dicobakan pada data medis (thyroid, kanker payudara dan kanker serviks) dengan menggunakan CV 5-fold.

LANDASAN TEORI

Synthetic Minority Oversampling Technique

Synthetic Minority Oversampling Technique (SMOTE) merupakan salah satu metode *oversampling* yaitu teknik pengambilan sampel untuk meningkatkan jumlah data pada kelas positif dengan cara mereplikasi jumlah data pada kelas positif secara acak sehingga jumlahnya sama dengan data pada kelas negatif. Algoritma SMOTE pertama kali ditemukan oleh (Chawla dkk, 2002). Pendekatan ini bekerja dengan mereplikasi data minor. Metode yang terdapat pada algoritma SMOTE ini adalah *k-nearest neighbor* (ketetanggaan data). *Synthetic* data dilakukan dengan menggunakan persamaan sebagai berikut.

$$x_{syn} = x_i + \left(x_{knn} - x_i\right) \times \gamma \qquad (1)$$

dengan

x_{syn} adalah data hasil replikasi.

x_i adalahdata ke-i dari kelas minor.

x_{knn} adalah data dari kelas minor yang memiliki jarak terdekat dari x_i

γ adalah bilangan random antara 0 dan 1

Tomek Links

Tomek Links merupakan salah satu metode *undersampling*, yang diperkenalkan oleh (Tomek, 1998). Metode ini bekerja dengan menghapus data kelas negatif (mayoritas) yang merupakan kasus *borderline* atau yang memiliki kesamaan karakteristik. Tomek Links dapat digunakan sebagai metode pembersihan data dari *noise*.Untuk setiap data, jika satu tetangga yang paling dekat memiliki kelas label yang berbeda dengan data tersebut maka data mayor akan dihapus karena dianggap sebagai *noise* atau *misclassification*.Diberikan dua sampel **x** dan **z** milik kelas yang berbeda, dan d(***x***,***z***) adalah jarak antara **x**dan**z**. Sepasang (x,z) disebut Tomek Links jika tidak ada sampel **z***,sehingga $d(\mathbf{x},\mathbf{z}^*) < d(\mathbf{x},\mathbf{z})$ atau $d(\mathbf{z},\mathbf{z}^*) < d(\mathbf{z},\mathbf{x})$ [9]. Jika dua sampel membentuk *Tomek Links*, maka salah satu dari kedua sampel adalah data *noise* atau kedua contoh adalah *borderline.*

Least Square Support Vector Machine

Suykens dkk (1999a) mengusulkan sebuah versi *least squares* untuk algoritma pembelajaran *Support Vector Machine* (SVM) yang disebut *Least SquaresSupportVector Machine*. LS-SVM adalah formulasi ulangmetode SVMstandar yangmengarah padapemecahanliniersistem*Karush-Kuhn*-Tucker (KKT).DalamformulasiLS-SVM,perhitungankomputasi dariSVMyangdisederhanakan denganpelaksanaanversi*least squares* daripada *inequality constraints* danfungsibiaya penjumlahan kesalahankuadrat (*squared error*) yangdigunakan dalampelatihanjaringansaraf tiruan.Reformulasiini sangat menyederhanakanmasalahdalam memecahkansatu set persamaanlinier daripada pemrogramankuadratikdigunakan dalamSVMkonvensional.*Primal problem* pada LS-SVM atau fungsi tujuan dirumuskan sebagai berikut

$$\min_{w,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2} C \sum_{i=1}^{n} \xi_i^2 \qquad (6)$$

dengan, $y_i[\varphi(\mathbf{x}_i)^T \mathbf{w} + b] = 1 - \xi_i \;;\; i, j = 1, ..., n$

Fungsi kernel dengan $\Omega_{ij} = K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \varphi\left(\mathbf{x}_i\right)^T \varphi\left(\mathbf{x}_j\right)$

Meminimukan Fungsi Lagrange dari persamaan (6) adalah

$$\min_{\mathbf{w},b} \mathrm{L}_{\mathrm{pri}}(\mathbf{w},b,\alpha) = \min \frac{1}{2}\|\mathbf{w}\|^2 + C\frac{1}{2}\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i\left(y_i[\varphi(\mathbf{x}_i)^T\mathbf{w}+b]-1+\xi_i\right) \quad (7)$$

dengan α_i adalah pengali *lagrange*.

Meminimumkan fungsi *langrange* yaitu menurunkan persamaan(7) dengan **w**, *b*, ξ_i , dan α_i

adalah sebagai berikut. ,

$$\begin{cases} \dfrac{\partial \mathrm{L}_{pri}(\mathbf{w},b,\alpha)}{\partial \mathbf{w}} = \mathbf{0} \rightarrow \mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i \varphi(\mathbf{x}_i) \\ \dfrac{\partial \mathrm{L}_{pri}(\mathbf{w},b,\alpha)}{\partial b} = 0 \rightarrow \sum_{i=1}^{n}\alpha_i y_i = 0 \\ \dfrac{\partial \mathrm{L}_{pri}(\mathbf{w},b,\alpha)}{\partial \xi_i} = 0 \rightarrow \alpha_i = C\xi_i, \quad i = 1,....,n \\ \dfrac{\partial \mathrm{L}_{pri}(\mathbf{w},b,\alpha)}{\partial \alpha_i} = 0 \rightarrow y_i[\varphi(\mathbf{x}_i)^T\mathbf{w}+b] = 1-\xi_i, \quad i = 1,...,n \end{cases} \quad (8)$$

Dapat ditulis sebagai sistem linier sebagai berikut

$$\left[\begin{array}{ccc|c} \mathbf{I} & 0 & 0 & -\mathbf{Z}^{\mathbf{T}} \\ 0 & 0 & 0 & -\mathbf{y}^{\mathbf{T}} \\ 0 & 0 & \mathbf{CI} & -\mathbf{I} \\ \hline \mathbf{Z} & \mathbf{y} & \mathbf{I} & 0 \end{array}\right]\left[\begin{array}{c} \mathbf{w} \\ b \\ \xi \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ \hline \mathbf{1} \end{array}\right] \quad (9)$$

dengan

$\mathbf{Z} = [\varphi(\mathbf{x}_i)^T y_1 ,...., \varphi(\mathbf{x}_n)^T y_n]^T$, $\mathbf{y} = \left[y_1, \mathrm{y}_2 \ \ldots, y_n\right]^T$, $\mathbf{1} = [1,...,1]^T$, $\xi = [\xi_1,...,\xi_n]^T$, $\alpha = [\alpha_i,...,\alpha_n]^T$

C adalah matrik simetri ukuran *nxp* dan **I** adalah matrik identitas.Eliminasi **w** dan ξ menghasilkan Persamaan (10)

$$\left[\begin{array}{c|c} 0 & \mathbf{y}^{\mathrm{T}} \\ \hline \mathbf{y} & \mathbf{ZZ}^{\mathrm{T}} + \mathbf{C}^{-1}\mathbf{I} \end{array}\right]\left[\begin{array}{c} \mathrm{b} \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \mathbf{1} \end{array}\right] \quad (10)$$

Kondisi Mercer diaplikasikan ke matriks $\Omega = \mathbf{ZZ}^{\mathbf{T}}$ dengan

$$\Omega_{ij} = y_i y_j \varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}_j)$$
$$= y_i y_j \, K(\mathbf{x}_i, \mathbf{x}_j) \ ; \quad i, j = 1,2,...,\mathrm{n}$$

Particle Swarm Optimization-Gravitational Search Algorithm (PSO-GSA)

Algoritma PSO-GSA dikembangkan oleh (Mirjajili, 2010)untuk optimasi parameter sehingga menghasilkan solusi terbaik. Ide dasar dari PSO-GSA adalah untuk menggabungkan kemampuan global terbaik (*Gbest*) algoritma PSO dengan kemampuan pencarian lokal (*Pbest*) pada algoritma GSA. Dalam rangka untuk menggabungkan algoritma ini, maka algoritma tersebut dapat diformulasikan menggunakan persamaan (13).

$$\boldsymbol{V}_i^{\,p}(t+1) = \omega \times \boldsymbol{V}_i^{\,p}(t) + c_1 r \times ac_i^{\,p}(t) + c_2 r \times \left(\boldsymbol{Gbest}^{\,p}(t) - \theta_i^{p}(t)\right) \quad (13)$$

dengan $v_i^{\,p}(t)$ adalah kecepatan dari agen *i* pada iterasi *t*, c_1, c_2 adalah konstanta positif yang diboboti, ω adalah bobot inersia, *r* adalah bilangan random diantara 0 dan 1, $ac_i^{\,p}(t)$ adalah percepatan agen *i* pada iterasi *t*, dan *Gbest* adalah solusi global terbaik. Pada masing-masing iterasi, posisi $\theta_i^{p}(t)$ partikel di perbarui (*update*) sebagai berikut.

$$\theta_i^{p}(t+1) = \theta_i^{p}(t) + \boldsymbol{V}_i^{\,p}(t+1) \quad (14)$$

Evaluasi Performansi Model

Untuk mengevaluasi performansi suatu model klasifikasi (Han dkk, 2001)dapat dilakukan dengan menghitung jumlah dari data testing yang diprediksi benar (akurasi) oleh model tersebut.

$$\text{Akurasi Total} = \frac{\text{jumlah prediksi benar}}{\text{jumlah total prediksi}} \times 100\%$$

$$Sensitivity = \frac{TP}{(TP + FN)} \times 100\% \quad (15)$$

$$G - Mean = \sqrt{Sensitivity \times Specificity}$$

METODOLOGI PENELITIAN

Sumber Data

Studi kasus yang digunakan pada penelitian ini adalah klasifikasi *multiclass* dalam bidang medis. Peneliti menggunakan tiga jenis data yaitu data penderita thyroid, data penderita kanker payudara (*breast cancer*) dan kanker serviks (*cervical cancer*). Data kanker payudara dan kanker serviks diperoleh dari salah satu rumah sakit swasta di Surabaya. Data thyroid diambil dari dari UCI *Repository Of Machine Learning.* Variabelpenelitian yang digunakan dalam penelitian ini dapat dilihat pada Tabel 1.

Tabel 1 Variabel Penelitian

Thyroid		Kanker Payudara		Kanker Serviks	
Variabel	**Deskripsi**	**Variabel**	**Deskripsi**	**Variabel**	**Deskripsi**
X_1	Persentase hasil uji T3 Resin	X_1	Ukuran tumor	X_1	Usia
X_2	Total Serum Thyroxin	X_2	Nodus	X_2	Kontrasepsi

X_3	Total Serum Triidothyronine	X_3	Kemoterapi	X_3	Usia menstruasi pertama kali
X_4	Hormon Basal (TSH)	X_4	Tingkat keganasan	X_4	Usia pertama kali melahirkan
X_5	Perbedaan maximal absolute pada nilai TSH setelah disuntik	X_5	Letak kanker	X_5	Paritas
X_6	Kondisi Thyroid	X_6	Usia Pasien	X_6	Siklus menstruasi
				X_7	Riwayat keguguran
Y	Normal Hyperthyroid Hypothyroidi	Y	Stadium 1 Stadium 2 Stadium 3	Y	Kelas 1 Kelas 2 Kelas 3 Kelas 4 Kelas 5

A. Metode Analisis

Metode analisis yang dilakukan pada penelitian ini terdiri dari :

1. Melakukan *Preprocessing* data
2. Melakukan Deskripsi Data
3. Melakukan *Preprocessingimbalanced* data (SMOTE, Tomek Links dan Combine Sampling) . Jumlah data tyroid sebanyak 215, data kanker payudara sebanyak 178 dan data kanker serviks sebanyak 719.
4. Melakukan klasifikasi LS-SVM OAO untuk kasus klasifikasi *multi class*
   i. Membagi data menjadi data *training* dan *testing* dengan menggunakan 5 *fold crossvalidation*
   ii. Menentukan nilai parameter σ dan C (C=1,50,100 dan σ =1,10, 20)
5. Mengoptimisasi parameter kernel RBF (*Radial Basis Function*) dan nilai *C* (nilai pinalti) pada LS-SVM dengan PSO-GSA sehingga tidak melakukan *trial and error* untuk penentuan parameternya.
6. Mengevaluasi performansi metode klasifikasi*Combine Sampling* LS-SVM PSO-GSA berdasarkan nilai akurasi total, *Sensitivity*dan *Gmean.*
7. Melakukan perbandingan perfomansi metodepada setiap data berdasarkan nilai akurasi, *Sensitivity* dan *G-mean.* Metode yang diukur performansinya antara lain :

M1 : LS-SVM Original
M2 : LS-SVM SMOTE
M3 : LS-SVM Tomek Links
M4 : LS-SVM Combine Sampling
M5 : LS-SVM PSO-GSA Original
M6 : LS-SVM PSO-GSA SMOTE
M7 : LS-SVM PSO-GSA Tomek Links
M8 : LS-SVM PSO-GSA Combine Sampling

HASIL DAN PEMBAHASAN

Deskripsi Data

Sebelum dilakukan pengolahan maka langkah pertama adalah melakukan *preprocessing* data. Pada data kanker payudara dan kanker serviks tidak ditemukan data yang *missing* dan outlier. Pada data thyroid dideteksi ada data yang outlier. Hasil pengujian outlier secara *multivariate*, diperoleh nilai *P-value* kurang dari tingkat signifikan (α=0.00001) maka disimpulkan terdapat *outlier* pada observasi ke-156, 167, 193, 195, 196,199, 208 dan 204. Observasi ini dihilangkan. Data thyroid menjadi 207.

Setelah data terbebas dari *missing value* dan *outlier* maka selanjutnya akan dilakukan deskripsi data pada ketiga kasus yaitu kasus Thyroid, kanker payudara dan kanker serviks ditunjukkan pada Tabel 2.

Tabel2 Persentase Kelas
Pada Thyroid ,Kanker Payudara dan Kanker Serviks

Thyroid		Kanker Payudara		Kanker Serviks	
Kelas	**Persentase**	**Kelas**	**Persentase**	**Kelas**	**Persentse**
1	**72%**	1	6%	1	38%
2	16%	2	38%	2	**43%**
3	12%	3	**56%**	3	12%
				4	6%
				5	1%

Tabel 2 diketahui bahwa mayoritas pasien thyroid didiagnosa mengalami kondisi normal yaitu sebesar 72%. Mayoritas pasien kanker payudara berada pada Stadium III yaitu sebesar 56%. Sebagian besar pasien kanker serviks didiagnosa mengalami radang ringan non spesifik dan terdapat sel-sel abnormal yaitu 43%.

SMOTE Preprocessing Imbalanced Data

Pada ketiga data yang digunakan baik thyroid, kanker payudara maupun serviks diklasifikasikan menjadi beberapa kelas. Thyroid diklasifikasikan menjadi 3 kelas. Kanker payudara diklasifikasikan menjadi 3 kelas sedangkan kanker serviks diklasifikasikan menjadi 5 kelas. Dari masing-masing data, teknik *oversampling* yang dilakukan memiliki

tahapan yang sama. Deskripsi distribusi data pada kanker payudara dan serviks ditunjukkan pada Tabel 3.

Tabel 3 Deskripsi Distribusi Data Sebelum dan Setelah SMOTE

Data	Kelas Mayor	Kelas Minor	Replikasi	Kelas Mayor Baru	Kelas Minor Baru
Thyroid	(150*)(72 %**) (1***)	(33*)(16%**) (2***)	4 kali	(150*)(31%**) (1***)	(165*)(34%**) (2***)
		(24*)(12%**) (3***)	6 kali		**(168*)(35%**) (3***)**
Kanker Payudara	(100*) (56%**) (3***)	(11*)(6%**) (1***)	9 kali	(100*)(29%**) (3***)	(110*) (32%**) (1***)
		(67*)(38%**) (2***)	1 kali		**134 (39%) (2)**
Kanker Serviks	(340*) (43%**) (2***)	(7*) (1%**) (5***)	6 kali	(340*)(20%**) (2***)	#**49**
		#**49**	6 kali		(343*) (20%**) (5***)
		(50*) (6%**) (4***)	6 kali		(350*) (20%**) (4***)
		(98*)(12%**) (3***)	3 kali		**(392*)(23%**) (3***)**
		(299*)(38%**) (1***)	-		(299*) (17%**) (1***)

Keterangan: #) angka yang digunakan adalah sama *) jumlah data , **)persentase data, ***) kategori kelas

A. Tomek Links Preprocessing Imbalanced Data

Deskripsi distribusi data menggunakan tomek links dapat dilihat pada ditunjukkan pada Tabel 4.

Tabel 4 Deskripsi Distribusi Data Sebelum dan Setelah Tomek Links

Data	Mayor	Minor	Data Mayor Hapus	Mayor Baru	Minor Baru
Thyroid	(150*)(72 %**) (1***)	(33*)(16%**) (2***) (24*)(12%) (3)	3	(147*)(72%**) (1***)	(33*)(16%**) (2***) (24*)(12%**) (3***)
Kanker Payudara	(100*) (56%**) (3***)	(11*) (6%**) (1***) (67*)(38%**) (2***)	24	(76*)(49%**) (3***)	(11*)(7%**) (1***) (67*)(44%**) (2***)
Kanker Serviks	(340*) (43%**) (2***)	(7*) (1%**) (5***) (50*) (6%**) (4***) (98*)(12%**) (3***) (299*)(38%**) (1***)	176 -	(164*)(27%**) (2***)	(7*)(1%**) (5***) (50*)(8%**) (4***) (98*)(16%**) (3***) (299*)(48%*) (1***)

Metode Tomek Links yang digunakan dalam penelitian ini yaitu metode tomek links yang dapat digunakan sebagai metode *undersampling* yaitu hanya kelas mayoritas yang akan dieliminasi. Penerapan metode tomek links menggunakan data asli.

Combine Sampling Preprocessing Imbalanced Data

Metode *combine sampling* merupakan perpaduan metode *oversampling* dan *undersampling*yaitu antara metode SMOTE dan Tomek Links. Penggunaan kedua metode ini dilakukan secara berurutan yaitu penanganan menggunakan SMOTE terlebih dahulu selanjutnya hasil SMOTE dilanjutkan menggunakan penanganan Tomek Links. Deskripsi data sebelum dan setelah *Combine Sampling* dapat dilihat pada Tabel 5.

Tabel 5 Deskripsi Distribusi Data Sebelum dan Setelah Combine Sampling

Data	Kelas Mayor hasil SMOTE	Kelas Minor hasil SMOTE	Data Mayor Hapus	Kelas Mayor Baru	Kelas Minor Baru
Thyroid	(168*)(35%**) (3***)	(165*)(34%**) (2***) (150*)(31%**) (1***)	0	(168*)(35%**) (3***)	(165*) (34%**) (2***) (150**)(31%***) (1***)
Kanker Payudara	(134*)(39%**) (2***)	(110*)(32%**) (1***) 100 (29%) (3)	52	(82*)(28%**) (2***)	(110*) (38%**) (1***) 100(34%) (3)
Kanker Serviks	(392*)(23%**) (3***)	(343*)(20%**) (5***) (350*)(20%**) (4***) (340*)(20%**) (2***) (299*)(17%**) (1***)	44 - -	(348*)(21%**) (3***)	(343*)(20%**) (5***) (350*)(21%**) (4***) (340*)(20%**) (2***) (299*)(18%**) (1***)

Keterangan : *) jumlah data , **)persentase data, ***) kategori kelas

Klasifikasi Menggunakan LS-SVM

Rangkuman nilai rata-rata Akurasi klasifikasi tertinggi pada data testing di semua metode yang dicobakan dapat dilihat pada Tabel 6.

Tabel 6 Rangkuman nilai rata-rata Akurasi Total, Sensitivity, G-Mean, Pada Testing (q=5 Fold)

Metode	**Data (%)**			**Metode**	**Data (%)**		
	1 (Thyroid)	**2 (Kanker**	**3 (Kanker**		**1 (Thyroid)**	**2 (Kanker**	**3 (Kanker**

		Payudara)	Serviks)			Payudara)	Serviks)
	Akurasi Sensitivity G-Mean	**Akurasi Sensitivity G-Mean**	**Akurasi Sensitivity G-Mean**		**Akurasi Sensitivity G-Mean**	**Akurasi Sensitivity G-Mean**	**Akurasi Sensitivity G-Mean**
M1	86,466 86,466 88,963 (C=50) (σ=20)	89,249 89,248 92,095 (C=50) (σ=20)	46,727 46,161 59,772 (C=50) (σ=20)	M5	81,702 86,525 89,000 (C=48,58) (σ=1,27)	88,137 82,353 91,235 (C=11) (σ=1,61)	38,656 37,456 50,612 (C=100) (σ=1)
M2	**98,557*** 97,083 97,525 (C=50,100) (σ=1)	91,577 91,577 93,158 (C=1) (σ=1)	59,412 59,974 70,938 (C=100) (σ=1)	M6	**98,558*** **97,230**** **98,423*** (C=76,40) (σ=1,01)	90,638 92,636 93,164 (C=71,29) (σ=2,35)	**59,471*** 59,898 71,102 (C=100) (σ=1)
M3	86,024 86,024 88,430 (C=100) (σ=20)	92,817 82,817 94,791 (C=100) (σ=20)	56,140 55,513 67,968 (C=1) (σ=20)	M7	86,024 86,224 88,965 (C=57,56) (σ=1,93)	93,102 90,234 95,123 (C=33,68) (σ=2,82)	56,378 56,256 56,256 (C=98,89) (σ=1)
M4	98,557 97,083 97,525 (C=50,100) (σ=1)	92,782 **92,782**** 93,857 (C=1) (σ=1)	57,738 60,576 71,535 (C=100) (σ=1)	**M8**	**98,558*** **97,230**** **98,423*** (C=69,50) (σ=1)	**95,623*** **93,450** **96,423***** (C=48,98) (σ=1)	**59,621*** **61,422**** **72,133***** (C=100 (σ=1)

Ket :

***) Rata-rata Akurasi Tertinggi **) Rata-rata Sensitivity Tertinggi ***) Rata-rata G-Mean**

M1 : LS-SVM Original	M5 : LS-SVM PSO-GSA Original
M2 : LS-SVM SMOTE	M6 : LS-SVM PSO-GSA SMOTE
M3 : LS-SVM Tomek Links	M7 : LS-SVM PSO-GSA Tomek Links
M4 : LS-SVM Combine Sampling	M8 : LS-SVM PSO-GSA Combine Sampling

Pada klasifikasi LS-SVM, parameter kernel RBF (σ) dan (C) dilakukan *trial error*. *Trial error* dilakukan sebanyak 9 kali percobaan. Parameter kernel RBF (σ) yang dicobakan yaitu $(\sigma = 1, 10, 20)$ dan parameter (C) yang dicobakan yaitu $(C = 1, 50, 100)$.

Pada klasifikasi LS-SVM PSO-GSA, parameter kernel RBF (σ) dan (C) dioptimasi tidak menggunakan *trial error* dan berada pada range yang ditentukan. Penentuan range akan menentukan besar kecilnya akurasi. Parameter kernel RBF (σ) yang dicobakan yaitu *range* $\sigma = (1 - 20)$ dan parameter (C) yang dicobakan yaitu *range* $C = (1 - 100)$. Solusi optimum sebanyak 20 partikel, Iterasi maksimum (T*) sebesar 50, Bobot maksimum sebesar 0,9, bobot minimum sebesar 0,4, inisial G_0 pada GSA sebesar 10. Untuk validasi model klasifikasi digunakan data testing dengan (σ) dan (C) yang telah diperoleh selama proses training.

Tabel 6menunjukkan bahwa dengan metode *imbalanced* data dapat meningkatkan nilai akurasi kemudian metode terbaik atau unggul di semua data percobaan baik diukur performansi dari akurasi total, *Sensitivity* dan *Gmean* adalah *Combine Sampling-Least Square Support Vector Machine PSO-GSA*.

KESIMPULAN DAN SARAN

Kesimpulan

Berdasarkan hasil dan pembahasan dapat disimpulkan bahwa metode yang terbaik untuk kasus klasifikasi *imbalanced* dalam memprediksi status pasien penderita Thyroid, kanker payudara dan kanker serviks adalah metode *combine Sampling Least Square Support Vector Machine PSO-GSA*.

Saran

Berdasarkan kesimpulan yang diperoleh, saran yang dapat dipertimbangkan untuk penelitian selanjutnya adalah melakukan simulasi untuk setiap kategori*imbalanced data* yaitu kategori *imbalanced*tingkat tinggi, sedang, dan rendah serta menggunakan *Stratified Cross Validation* dalam membagi data training dan data testing agar proporsi kelas dapat seimbang.

DAFTAR PUSTAKA

Batista, G.E.A.P.A., Bazzan, A.L.C. dan Monard, M.C. 2003. "Balancing Training Data for Automated Annotation of Keyword: a Cese study". *Proceedings of the second Brazilian Workshop Bioinformatics.* Diakses darihttp://www.icmc.usp.br/gbatista/files/wob2003.pdf, pada Tanggal 16 Oktober 2015).

Chawla NV, Bowyer KW, Hall LO, dan Kegelmeyer WP. 2002. "SMOTE:Synthetic Minority Oversampling Technique", *Journal of Artificial Intelligence Research*, Vol. 16, Hal.321-357.

Cortez, C dan V. Vapnik. 1995. "Support Vector Networks", *Machine Learning*, Vol. 20, No. 3, Hal. 273–297.

Haerdle, WK, Prastyo, DD, and Hafner, CM. 2014. Support Vector Machines with Evolutionary Model Selection for Default Prediction," in *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, eds. Racine, JS, Su, L, and Ullah, A, Oxford University Press, 346-373.

Han, J dan Kamber, M. 2001.*Data Mining Concept and Tehniques,* USA, Academic Press.

Mirjajili S dan Hashim SZM,. 2010. "A New Hybrid PSOGSA Algorithm for Function Optimization", *International Conference on Computer and Information Application (ICCIA)*.

Novianti, A. Furina., Purnami, W. Santi. 2012. "Analisis Diagnosis Pasien Kanker Payudara Menggunakan Regresi Logistik dan Support Vector Machine (SVM) Berdasarkan Hasil Mammografi", *Jurnal SAINS dan SENI ITS,* Vol.1, No.1. (Sept. 2012) ISSN : 2301-928X.

Rahman, Farizi., Purnami, W. Santi. 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.

Sain, Hartayuni. 2013. *Combine sampling Support Vector Machine Untuk Klasifikasi Data Imbalanced*, Tesis, Statistika-FMIPA ITS, Surabaya.

Suykens, J.A.K., dan Vandewalle J. 1999a. "Least Squares Support Vector Machine Classifiers", *Neural Processing Letter*, Vol. 9, Hal. 293-300.

Tomek, I. 1998. "Two Modification of CNN". *IEEE Transactions on System Man and Communications*, SMC-6: 769-772, 1976.

Trapsilasiwi, R.K. 2013. *Klasifikasi Multiclass Untuk Imbalanced Data Menggunakan SMOTE Least Sqaure Support Vector Machine*, Tesis, Statistika FMIPA-ITS, Surabaya.