

USULAN TUGAS AKHIR

1. IDENTITAS PENGUSUL

Nama : **RADEN RISKY DWI LISTYO FIRMANSYAH**
NRP : **510810031**
Dosen Wali : **DR. AGUS ZAINAL ARIFIN. S.KOM, M.KOM**

2. JUDUL TUGAS AKHIR

***"Implementasi Metode Elastic SCAD dengan Support Vector Machine (SVM) untuk
Prediksi Kanker Payudara berdasarkan Data DNA Microarray"***

3. LATAR BELAKANG

Kanker payudara adalah suatu penyakit dimana terjadi pertumbuhan berlebihan atau perkembangan tidak terkontrol dari sel-sel (jaringan) payudara. Dari seluruh penjurur dunia, penyakit kanker payudara (*Carcinoma Mammar*) merupakan salah satu penyakit kanker terbesar kelima yang menyebabkan kematian setelah kanker paru-paru, kanker rahim, kanker hati, dan kanker usus. Pertumbuhan jumlah penderita penyakit ini pun terus bertambah setiap tahunnya. Umumnya seseorang baru diketahui menderita penyakit kanker payudara setelah menginjak stadium lanjut dan cukup parah. Tindakan yang biasa dilakukan bagi penderita kanker payudara yang cukup parah adalah dengan operasi pengangkatan kanker tersebut dan hal ini dihindari oleh sebagian besar wanita. Oleh sebab itu, prediksi atau klasifikasi seseorang yang menderita penyakit kanker payudara atau tidak sebaiknya dilakukan sedini mungkin, sehingga dapat mengurangi akibat yang timbul setelahnya.

DNA Microarray merupakan sekumpulan titik DNA mikroskopis. *DNA Microarray* ini juga dikenal sebagai *chip-gen*, *DNA chip*, atau juga *biochip*. Para ilmuwan menggunakan *DNA Microarray* untuk menghitung nilai ekspresi dari beberapa gen yang

berbeda tetapi masih dalam satu jenis. Misalnya saja dengan membandingkan sel normal dengan sel kanker dan melihat perbedaan yang terjadi pada ekspresi genetik antara dua jenis sel tersebut. Kemudian, penelitian yang dilakukan oleh Van't Veer tahun 2002 menunjukkan bahwa *DNA Microarray* dapat membantu klasifikasi penderita kanker payudara^[3]. Sebelumnya, data *DNA Microarray* ini direpresentasikan menjadi vektor. Vektor hasil representasi inilah yang kemudian diolah dan diklasifikasi dengan metode-metode yang sudah ada.

Banyak metode klasifikasi data yang sudah berkembang, diantaranya regresi, *decision tree*, *genetic algorithm*, *ant colony optimization* (ACO), *partial swarm optimization* (PSO), dan lainnya. Salah satu metode yang unggul dalam penanganan data dengan dimensi tinggi adalah *Support Vector Machine* (SVM), yang dikembangkan oleh Vapnik^[4]. SVM memisahkan sampel data sesuai dengan kelasnya masing-masing melalui *hyperplane* dengan nilai margin terbesar.

Perkembangan *Support Vector Machine* (SVM) selanjutnya ada pada seleksi fitur dan nilai penalti. Seleksi fitur ini digunakan untuk menemukan fitur yang paling relevan untuk sebuah prediksi tertentu. Tujuan seleksi fitur itu untuk meningkatkan nilai prediksi dari prediktor, *cost* prediktor lebih cepat dan lebih efektif, dan mendapatkan informasi tentang data yang diproses jauh lebih banyak. *LASSO*, *SCAD*, dan *Elastic Net* merupakan seleksi fitur yang lebih fleksibel. *Elastic SCAD SVM* merupakan perkembangan seleksi fitur dan penalti yang lebih kompleks lagi pada *Support Vector Machine* (SVM) yang dikembangkan oleh Becker et al^[2].

Tugas akhir ini bertujuan mengimplementasikan *Elastic SCAD SVM*. Implementasi ini kemudian diharapkan dapat bermanfaat di bidang kedokteran dalam diagnosa awal penyakit kanker payudara.

4. RUMUSAN MASALAH

Berdasarkan latar belakang yang telah diuraikan di atas, terdapat beberapa permasalahan yang dapat dirumuskan sebagai berikut:

1. Bagaimana melakukan representasi data *DNA Microarray* menjadi sebuah vektor sehingga dapat diolah dengan lebih mudah.
2. Bagaimana membuat modul *Elastic SCAD*.

3. Bagaimana menggabungkan modul *Elastic SCAD* pada metode klasifikasi *Support Vector Machine* (SVM).
4. Bagaimana memodelkan *Elastic SCAD SVM* pada data *DNA Microarray* dalam kasus klasifikasi penyakit kanker payudara.

5. BATASAN MASALAH

Permasalahan yang dibahas dalam Tugas Akhir ini memiliki beberapa batasan, antara lain sebagai berikut :

1. Sistem perangkat lunak segmentasi retina ini bangun dengan menggunakan perangkat lunak MATLAB 7.6.
2. Dataset yang digunakan pada Tugas Akhir ini yaitu *NKI breast cancer*.

6. TUJUAN TUGAS AKHIR

Tugas akhir ini bertujuan mengimplementasikan *Elastic SCAD SVM* yang kemudian dapat digunakan dalam kasus klasifikasi penyakit kanker payudara.

7. MANFAAT TUGAS AKHIR

Tugas akhir ini dikerjakan dengan harapan dapat memberikan manfaat yang besar pada bidang kedokteran dalam mendiagnosa penyakit kanker payudara lebih dini.

8. RINGKASAN TUGAS AKHIR

Pembuatan Tugas Akhir ini dikerjakan dalam beberapa modul. Modul-modul itu antara lain:

▪ *Support Vector Machine* (SVM).

SVM merupakan metode klasifikasi data. Data yang digunakan adalah data vektor. SVM mencari *hyperplane* dengan nilai margin terbesar yang dapat membagi kelas dengan baik.

Secara umum, SVM dapat didefinisikan sebagai berikut:

$$\{ x : f(x) = w \cdot x + b = 0 \}$$

Dengan $x \in R^p$ merupakan data vektor. $w = (w_1, w_2, \dots, w_p)$ adalah nilai koefisien vektor dari *hyperplane*, dan b adalah nilai *intercept* dari *hyperplane*.

- **Smoothly Clipped Absolute Deviation Penalty (SCAD).**

Pengembangan selanjutnya pada SVM adalah dengan menambahkan fitur seleksi dan penalti. Salah satunya yaitu, penalti SCAD. Penalti SCAD ini dianggap lebih baik daripada metode lainnya, misalnya LASSO. Karena SCAD dapat mengatasi permasalahan nilai koefisien yg besar, yang tidak bisa ditangani oleh LASSO^[1].

$$p_{scad(\lambda)}(w_j) = \begin{cases} \lambda |w_j|, & \text{if } |w_j| \leq \lambda, \\ -\frac{|w_j|^2 - 2a\lambda|w_j| + \lambda^2}{2(a-1)}, & \text{if } \lambda < |w_j| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & \text{if } |w_j| > a\lambda, \end{cases}$$

Dimana $w_j, j = 1, \dots, p$ adalah nilai koefisien vektor *hyperplane*. Bentuk penalti SCAD SVM seperti di bawah ini:

$$pen_{\lambda}(w) = \sum_{j=1}^p p_{scad(\lambda)}(w_j)$$

- **Elastic SCAD.**

Penggunaan metode penalti SCAD masih memungkinkan terlalu ketat dalam menyeleksi fitur-fitur dengan kondisi data tidak tersebar (*non-sparse data*). Modifikasi SCAD ini bisa dilakukan dengan menambahkan kelebihan dari metode penalti lainnya, yaitu penalti *Ridge*. Penalti *Ridge* dapat menghindari keterbatasan *sparsity* yang terlalu membatasi data tidak tersebar (*non-sparse data*).

$$pen_{\lambda}(w) = \lambda \|w\|_2^2 = \lambda \sum_{j=1}^p w_j^2$$

Kombinasi penalti SCAD dan penalti *Ridge* yang kemudian disebut *Elastic SCAD* adalah

$$pen_{\lambda}(w) := \sum_{j=1}^p p_{SCAD(\lambda_1)}(w_i) + \lambda_2 \|w\|_2^2, \quad \text{dimana } \lambda_1, \lambda_2 \geq 0.$$

Elastic SCAD dapat meningkatkan metode SCAD pada data *sparse* yang sedikit. Berdasarkan sifat SCAD dan penalti *Ridge*, seharusnya *Elastic SCAD* menunjukkan ketepatan prediksi yang baik pada kedua data, data tersebar dan tidak tersebar (*sparse* dan *non-sparse*).

Bentuk persamaan dari permasalahan optimisasi *Elastic SCAD SVM* adalah

$$\min_{b,w} \frac{1}{n} \sum_{i=1}^n [1 - \gamma f(x_i)]_+ + \sum_{j=1}^p p_{SCAD(\lambda_1)}(w_j) + \lambda_2 \|w\|_2^2$$

▪ ***Elastic SCAD SVM : Algoritma.***

Elastic SCAD SVM dapat diimplementasikan melalui algoritma iteratif berikut ini:

Langkah 1

Set $k = 1$ dan tetapkan nilai awal $b^{(1)}, w^{(1)}$ dengan standar SVM penalti *Ridge* menurut Zhang et al^[5].

Langkah 2

Simpan hasil iterasi ke-k: $(b_0, w_0) = b^{(k)}, w^{(k)}$

Langkah 3

Minimisasi $\tilde{A}(b, w) = \frac{1}{2} \begin{pmatrix} b \\ w \end{pmatrix}^T Q \begin{pmatrix} b \\ w \end{pmatrix} - P \begin{pmatrix} b \\ w \end{pmatrix}$ dengan menyelesaikan persamaan

$Q \begin{pmatrix} \hat{b} \\ \hat{w} \end{pmatrix} = P$, dan notasikan hasil sebagai $b^{(k+1)}, w^{(k+1)}$.

Dengan $P = \frac{1}{2n} (y + r)^T X$ dan $Q = X^T D_0 X + Q_1 + Q_2$.

Dimana $r = \left[\frac{\gamma_1}{|\varepsilon_1|}, \dots, \frac{\gamma_n}{|\varepsilon_n|} \right]^T$ dan $D_0 = \frac{1}{2n} \text{diag} \left[\frac{1}{|\varepsilon_1|}, \dots, \frac{1}{|\varepsilon_n|} \right]$

$$Q_1 = \text{diag} \left[0, \frac{p'_{\lambda}(|w_{10}|)}{|w_{10}|}, \dots, \frac{p'_{\lambda}(|w_{d0}|)}{|w_{d0}|} \right] \text{ dan } Q_2 = \text{diag}[0, 2\lambda_2, \dots, 2\lambda_2]$$

Sebelumnya, definisikan matriks-matriks berikut ini,

$$y = [\gamma_1, \dots, \gamma_n]^T$$

$$w = [w_1, \dots, w_p]^T$$

$$\varepsilon = [\varepsilon_1, \dots, \varepsilon_n]^T \text{ dimana } \varepsilon_i = \gamma_i - (b_0 + w_0 \cdot x_i)$$

$X = [1, x_1, \dots, x_p]$ dimana 1 adalah vektor pertama dengan panjang n dan x_j adalah vektor masukan ke-j.

Langkah 4

Set $k = k + 1$. Kembali ke langkah 2 hingga konvergen.

Jika elemen $w_j^{(k)}$ mendekati nol, misalnya, lebih kecil dari 10^{-4} , maka variable ke-j dianggap redundan dan langkah selanjutnya akan dihapus dari model. Algoritma berhenti setelah $b^{(k)}, w^{(k)}$ konvergen.

9. METODOLOGI

Metodologi yang akan dilakukan dalam Tugas Akhir ini memiliki beberapa tahapan, diantaranya sebagai berikut:

1. Penyusunan Proposal Tugas Akhir

Tahap awal untuk memulai pengerjaan Tugas Akhir adalah penyusunan Proposal Tugas Akhir. Pada proposal ini, penulis mengajukan gagasan metode klasifikasi yang dapat diimplementasikan pada bidang bioinformatika (dalam hal ini deteksi penyakit kanker payudara melalui *DNA Microarray*).

2. Studi Literatur

Pada tahap ini dilakukan pencarian, pengumpulan, penyaringan, pembelajaran dan pemahaman literatur yang berhubungan dengan metode klasifikasi Support Vector Machine (SVM) dan pengembangannya, dataset *DNA Microarray* dan pengolahannya, serta pengaplikasian metode klasifikasi pada dataset. Literatur yang digunakan dalam pengerjaan Tugas Akhir ini sebagian besar berasal dari internet berupa makalah ilmiah, tesis, artikel, materi kuliah, serta beberapa buku referensi.

3. Implementasi

Implementasi merupakan tahap untuk membangun sistem tersebut.

4. Pengujian dan Evaluasi

Pada tahap ini dilakukan uji coba terhadap sistem yang telah dibuat, mengamati kinerja sistem yang baru dibuat, serta mengidentifikasi kendala yang mungkin timbul.

5. Penyusunan Buku Tugas Akhir

Tahap terakhir merupakan penyusunan laporan yang memuat dokumentasi mengenai pembuatan serta hasil dari implementasi perancangan dan *prototype* yang telah dibuat.

10. JADWAL PEMBUATAN TUGAS AKHIR

No.	Kegiatan	2011												2012
		September			Oktober			Nopember			Desember			Januari
1	Penyusunan Proposal Tugas Akhir													
2	Studi Literatur													
3	Perancangan Perangkat Lunak													
4	Implementasi Perangkat Lunak													
5	Pengujian dan Evaluasi													
6	Penyusunan Buku Tugas Akhir													

11. DAFTAR PUSTAKA

^[1]Fan J, Li R: **Variable selection via nonconcave penalized likelihood and its oracle properties.** *Journal of American Statistical Association* 2001, 96:1348-1360.

^[2]Becker et al: **Elastic SCAD as a novel penalization method for SVM classification tasks in high-dimensional data.** *BMC Bioinformatics* 2011, 12:138

^[3]Van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao HL, Mand Peterse, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts PS, Cand Linsley, Bernards R, H FS: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, 415:530-536.

^[4]Vapnik V: *The Nature of Statistical Learning Theory* New York: Springer; 1995.

^[5]Zhang HH, Ahn J, Lin X, Park C: **Gene selection using support vector machines with non-convex penalty.** *Bioinformatics* 2006, 22(1):88-95.

LEMBAR PENGESAHAN

Surabaya, 14 Oktober 2011

Menyetujui,

Dosen Pembimbing I,

Dosen Pembimbing II,

(Prof. Ir. Handayani Tjandrasa, M.Sc, Ph.D)
(NIP. 19490823 197603 2001)

(Isye Ariesianti, S.Kom. M. Phil.)
(NIP. 19780412 200604 2001)