Support Vector Machine

–Teori dan Aplikasinya dalam Bioinformatika[1]–

Anto Satriyo Nugroho, Arief Budi Witarto, Dwi Handoko

asnugroho@ieee.org
http://asnugroho.net

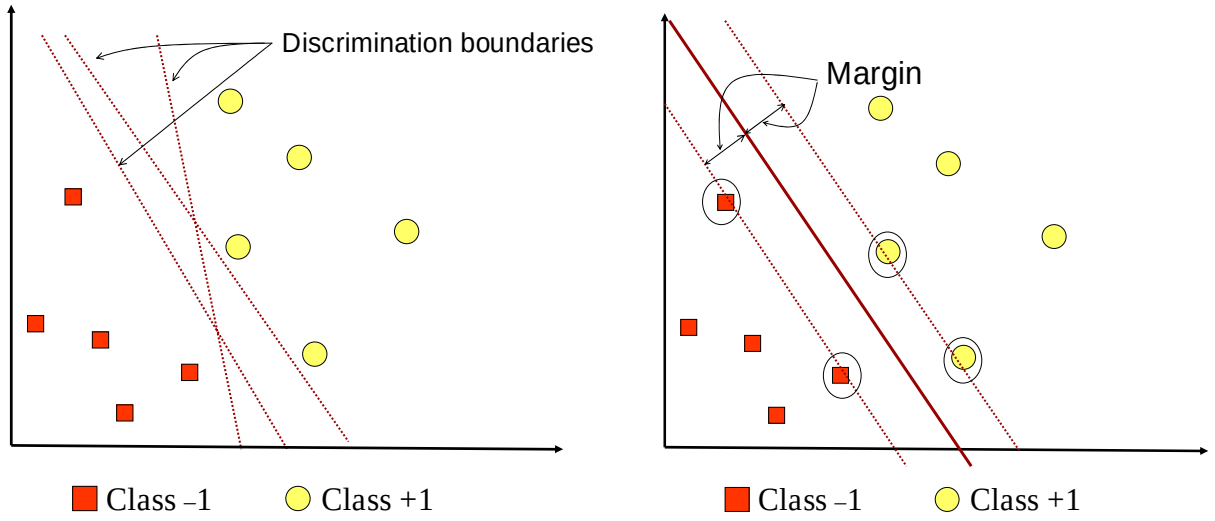
Abstrak:
Support Vector Machine (SVM) pertama kali diperkenalkan oleh Vapnik pada tahun 1992 sebagai rangkaian harmonis konsep-konsep unggulan dalam bidang pattern recognition. Sebagai salah satu metode pattern recognition, usia SVM terbilang masih relatif muda. Walaupun demikian, evaluasi kemampuannya dalam berbagai aplikasinya menempatkannya sebagai *state of the art* dalam pattern recognition, dan dewasa ini merupakan salah satu tema yang berkembang dengan pesat. SVM adalah metode learning machine yang bekerja atas prinsip Structural Risk Minimization (SRM) dengan tujuan menemukan hyperplane terbaik yang memisahkan dua buah class pada input space. Tulisan ini membahas teori dasar SVM dan aplikasinya dalam bioinformatika, khususnya pada analisa ekspresi gen yang diperoleh dari analisa microarray.

Keywords: pattern recognition, support vector machine, bioinformatika

1. PENDAHULUAN

Pattern Recognition merupakan salah satu bidang dalam komputer sains, yang memetakan suatu data ke dalam konsep tertentu yang telah didefinisikan sebelumnya. Konsep tertentu ini disebut *class* atau *category*. Aplikasi pattern recognition sangat luas, di antaranya mengenali suara dalam sistem sekuriti, membaca huruf dalam OCR, mengklasifikasikan penyakit secara otomatis berdasarkan hasil diagnosa kondisi medis pasien dan sebagainya. Berbagai metode dikenal dalam pattern recognition, seperti linear

[1] Bahan dalam makalah ini sebagian besar berasal dari makalah : Nugroho, A.S., Witarto, A.B., Handoko, D., "Application of Support Vector Machine in Bioinformatics", Proceeding of Indonesian Scientific Meeting in Central Japan, December 20, 2003, Gifu-Japan



(a) **(b)**

Gambar 1– SVM berusaha menemukan hyperplane terbaik yang memisahkan kedua class –1 dan +1

discrimination analysis, hidden markov model hingga metode kecerdasan buatan seperti artificial neural network. Salah satu metode yang akhir-akhir ini banyak mendapat perhatian sebagai *state of the art* dalam pattern recognition adalah Support Vector Machine (SVM) [1] [2]. Support Vector Machine (SVM) dikembangkan oleh Boser, Guyon, Vapnik, dan pertama kali dipresentasikan pada tahun 1992 di Annual Workshop on Computational Learning Theory. Konsep dasar SVM sebenarnya merupakan kombinasi harmonis dari teori-teori komputasi yang telah ada puluhan tahun sebelumnya, seperti *margin hyperplane* (Duda & Hart tahun 1973, Cover tahun 1965, Vapnik 1964, dsb.), kernel diperkenalkan oleh Aronszajn tahun 1950, dan demikian juga dengan konsep-konsep pendukung yang lain. Akan tetapi hingga tahun 1992, belum pernah ada upaya merangkaikan komponen-komponen tersebut [3][4].

Berbeda dengan strategi neural network yang berusaha mencari hyperplane pemisah antar class, SVM berusaha menemukan hyperplane yang terbaik pada input space. Prinsip dasar SVM adalah linear classifier, dan selanjutnya dikembangkan agar dapat bekerja pada problem non-linear. dengan memasukkan konsep *kernel trick* pada ruang kerja berdimensi tinggi. Perkembangan ini memberikan rangsangan minat penelitian di bidang pattern recognition untuk investigasi potensi kemampuan SVM secara teoritis maupun dari segi aplikasi. Dewasa ini SVM telah berhasil diaplikasikan dalam problema dunia nyata (real-world problems), dan secara umum memberikan solusi yang lebih baik dibandingkan metode konvensional seperti misalnya artificial neural network. Tulisan ini memperkenalkan konsep dasar SVM, dan membahas aplikasinya di bioinformatika, yang akhir-akhir ini merupakan salah satu bidang yang berkembang cukup pesat.

2. PATTERN RECOGNITION MEMAKAI SUPPORT VECTOR MACHINE

Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari hyperplane[2] terbaik yang berfungsi sebagai pemisah dua buah class pada input space. Gambar 1a memperlihatkan beberapa pattern yang merupakan anggota dari dua buah class : +1 dan –1. Pattern yang tergabung pada class –1 disimbolkan dengan warna merah (kotak), sedangkan pattern pada class +1, disimbolkan dengan warna kuning(lingkaran). Problem klasifikasi dapat diterjemahkan dengan usaha menemukan garis (hyperplane) yang memisahkan antara kedua

[2] hyperplane dalam ruang vector berdimensi *d* adalah affine subspace berdimensi *d-1* yang membagi ruang vector tersebut ke dalam dua bagian, yang masing-masing berkorespondensi pada class yang berbeda [4]

kelompok tersebut. Berbagai alternatif garis pemisah (*discrimination boundaries*) ditunjukkan pada gambar 1-a.

Hyperplane pemisah terbaik antara kedua class dapat ditemukan dengan mengukur *margin* hyperplane tsb. dan mencari titik maksimalnya. Margin adalah jarak antara hyperplane tersebut dengan pattern terdekat dari masing-masing class. Pattern yang paling dekat ini disebut sebagai *support vector*. Garis solid pada gambar 1-b menunjukkan hyperplane yang terbaik, yaitu yang terletak tepat pada tengah-tengah kedua class, sedangkan titik merah dan kuning yang berada dalam lingkaran hitam adalah support vector. Usaha untuk mencari lokasi hyperplane ini merupakan inti dari proses pembelajaran pada SVM.

Data yang tersedia dinotasikan sebagai $\vec{x}_i \in \Re^d$ sedangkan label masing-masing dinotasikan $y_i \in \{-1,+1\}$ untuk $i = 1,2,\cdots,l$, yang mana l adalah banyaknya data. Diasumsikan kedua class –1 dan +1 dapat terpisah secara sempurna oleh hyperplane berdimensi d, yang didefinisikan

$$\vec{w}.\vec{x} + b = 0 \qquad (1)$$

Pattern $\vec{x}_i$ yang termasuk class –1 (sampel negatif) dapat dirumuskan sebagai pattern yang memenuhi pertidaksamaan

$$\vec{w}.\vec{x}_i + b \le -1 \qquad (2)$$

sedangkan pattern $\vec{x}_i$ yang termasuk class +1 (sampel positif)

$$\vec{w}.\vec{x}_i + b \ge +1 \qquad (3)$$

Margin terbesar dapat ditemukan dengan memaksimalkan nilai jarak antara hyperplane dan titik terdekatnya, yaitu $1/\|\vec{w}\|$. Hal ini dapat dirumuskan sebagai *Quadratic Programming (QP) problem*, yaitu mencari titik minimal persamaan (4), dengan memperhatikan constraint persamaan (5).

$$\min_{\vec{w}} \quad \tau(w) = \frac{1}{2}\|\vec{w}\|^2 \qquad (4)$$

$$y_i(\vec{x}_i.\vec{w} + b) - 1 \ge 0, \quad \forall i \qquad (5)$$

Problem ini dapat dipecahkan dengan berbagai teknik komputasi, di antaranya Lagrange Multiplier.

$$L(\vec{w}, b, \alpha) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{l} \alpha_i (y_i((\vec{x}_i.\vec{w} + b) - 1))$$

$$(i = 1,2,\cdots,l) \qquad (6)$$

α_i adalah Lagrange multipliers, yang bernilai nol atau positif $(\alpha_i \ge 0)$. Nilai optimal dari persamaan (6) dapat dihitung dengan meminimalkan L terhadap $\vec{w}$ dan b, dan memaksimalkan L terhadap α_i. Dengan memperhatikan sifat bahwa pada titik optimal gradient L =0, persamaan (6) dapat dimodifikasi sebagai maksimalisasi problem yang hanya mengandung saja α_i, sebagaimana persamaan (7) di bawah.

Maximize:

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \vec{x}_i.\vec{x}_j \qquad (7)$$

Subject to:

$$\alpha_i \ge 0 \ (i = 1,2,\cdots,l) \quad \sum_{i=1}^{l} \alpha_i y_i = 0 \qquad (8)$$

Dari hasil dari perhitungan ini diperoleh α_i yang kebanyakan bernilai positif. Data yang berkorelasi dengan α_i yang positif inilah yang disebut sebagai support vector.

3. SOFT MARGIN

Penjelasan di atas berdasarkan asumsi bahwa kedua belah class dapat terpisah secara sempurna oleh hyperplane. Akan tetapi,

umumnya dua buah class pada input space tidak dapat terpisah secara sempurna. Hal ini menyebabkan constraint pada persamaan (5) tidak dapat terpenuhi, sehingga optimisasi tidak dapat dilakukan. Untuk mengatasi masalah ini, SVM dirumuskan ulang dengan memperkenalkan teknik *softmargin*. Dalam softmargin, persamaan (5) dimodifikasi dengan memasukkan *slack variabel* ξ_i $(\xi_i > 0)$ sbb.

$$y_i(\vec{x}_i.\vec{w}+b) \geq 1-\xi_i, \quad \forall i \qquad (9)$$

Dengan demikian persamaan (4) diubah menjadi :

$$\min_{\vec{w}} \quad \tau(\vec{w},\xi) = \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{l}\xi_i \qquad (10)$$

Paramater C dipilih untuk mengontrol tradeoff antara margin dan error klasifikasi ξ. Nilai C yang besar berarti akan memberikan penalti yang lebih besar terhadap error klasifikasi tsb.

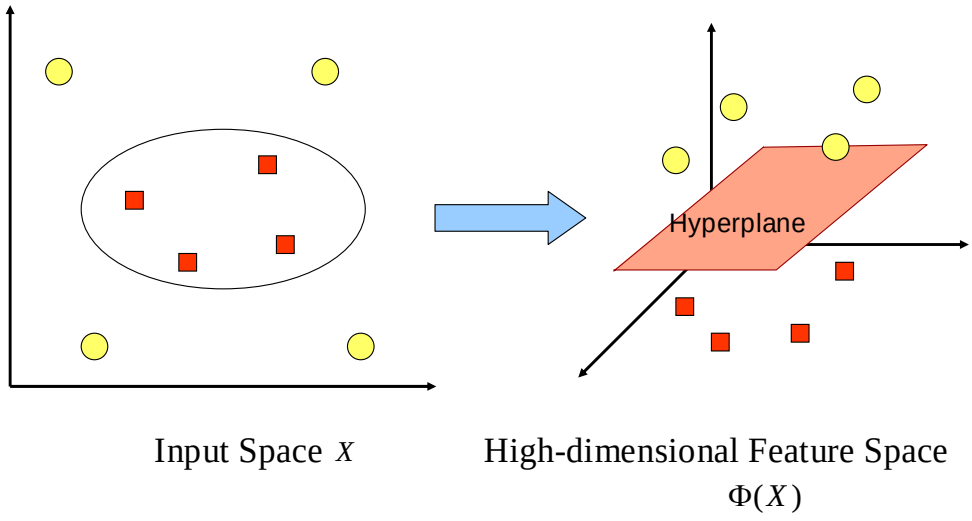
4. KERNEL TRICK DAN NON LINEAR CLASSIFICATION PADA SVM

Pada umumnya masalah dalam domain dunia nyata (real world problem) jarang yang bersifat linear separable. Kebanyakan bersifat non linear. Untuk menyelesaikan problem non linear, SVM dimodifikasi dengan memasukkan fungsi *Kernel*.

Dalam non linear SVM, pertama-tama data $\vec{x}$ dipetakan oleh fungsi $\Phi(\vec{x})$ ke ruang vektor yang berdimensi lebih tinggi. Pada ruang vektor yang baru ini, hyperplane yang memisahkan kedua class tersebut dapat dikonstruksikan. Hal ini sejalan dengan teori Cover yang menyatakan*"Jika suatu transformasi bersifat non linear dan dimensi dari feature space cukup tinggi, maka data pada input space dapat dipetakan ke feature space yang baru, dimana pattern-pattern tersebut pada probabilitas tinggi dapat dipisahkan secara linear"*.

Ilustrasi dari konsep ini dapat dilihat pada gambar 2. Pada gambar 2a diperlihatkan data pada class kuning dan data pada class merah yang berada pada input space berdimensi dua tidak dapat dipisahkan secara linear. Selanjutnya gambar 2b menunjukkan bahwa fungsi Φ memetakan tiap data pada input space tersebut ke ruang vektor baru yang berdimensi lebih tinggi (dimensi 3), dimana kedua class dapat dipisahkan secara linear oleh sebuah hyperplane. Notasi matematika dari mapping ini adalah sbb.

$$\Phi : \Re^d \rightarrow \Re^q \quad d < q \qquad (11)$$



Gambar 2– Fungsi Φ memetakan data ke ruang vektor yang berdimensi lebih tinggi, sehingga kedua class dapat dipisahkan secara linear oleh sebuah hyperplane

Tabel 1- Kernel yang umum dipakai dalam SVM

Jenis Kernel	Definisi
Polynomial	$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i.\vec{x}_j + 1)^p$
Gaussian	$K(\vec{x}_i, \vec{x}_j) = \exp(-\frac{\left\|\vec{x}_i - \vec{x}_j\right\|^2}{2\sigma^2})$
Sigmoid	$K(\vec{x}_i, \vec{x}_j) = \tanh(\alpha\vec{x}_i.\vec{x}_j + \beta)$

Pemetaan ini dilakukan dengan menjaga topologi data, dalam artian dua data yang berjarak dekat pada input space akan berjarak dekat juga pada feature space, sebaliknya dua data yang berjarak jauh pada input space akan juga berjarak jauh pada feature space.

Selanjutnya proses pembelajaran pada SVM dalam menemukan titik-titik support vector, hanya bergantung pada *dot product* dari data yang sudah ditransformasikan pada ruang baru yang berdimensi lebih tinggi, yaitu $\Phi(\vec{x}_i).\Phi(\vec{x}_j)$.

Karena umumnya transformasi Φ ini tidak diketahui, dan sangat sulit untuk difahami secara mudah, maka perhitungan dot product tersebut sesuai teori Mercer dapat digantikan dengan fungsi kernel $K(\vec{x}_i, \vec{x}_j)$ yang mendefinisikan secara implisit transformasi Φ.

Hal ini disebut sebagai *Kernel Trick*, yang dirumuskan

$$K(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i).\Phi(\vec{x}_j) \qquad (12)$$

Kernel trick memberikan berbagai kemudahan, karena dalam proses pembelajaran SVM, untuk menentukan support vector, kita hanya cukup mengetahui fungsi kernel yang dipakai, dan tidak perlu mengetahui wujud dari fungsi non linear Φ. Berbagai jenis fungsi kernel dikenal, sebagaimana dirangkumkan pada tabel 1.

Selanjutnya hasil klasifikasi dari data $\vec{x}$ diperoleh dari persamaan berikut :

$$f(\Phi(\vec{x})) = \vec{w}.\Phi(\vec{x}) + b \qquad (13)$$

$$= \sum_{i=1,\vec{x}_i \in SV}^{n} \alpha_i y_i \Phi(\vec{x}).\Phi(\vec{x}_i) + b \qquad (14)$$

$$= \sum_{i=1,\vec{x}_i \in SV}^{n} \alpha_i y_i K(\vec{x}, \vec{x}_i) + b \qquad (15)$$

SV pada persamaan di atas dimaksudkan dengan subset dari training set yang terpilih sebagai support vector, dengan kata lain data $\vec{x}_i$ yang berkorespondensi pada $\alpha_i \geq 0$.

5. METODE SEKUENSIAL

Hyperplane yang optimal dalam SVM dapat ditemukan dengan merumuskannya ke dalam QP problem dan diselesaikan dengan library yang banyak tersedia dalam analisa numerik. Alternatif lain yang cukup sederhana adalah metode sekuensial yang dikembangkan oleh Vijayakumar [5], sbb.

1. Initialization $\alpha_i = 0$

   Hitung matriks $D_{ij} = y_i y_j (K(x_i, x_j) + \lambda^2)$

2. Lakukan step (a), (b) dan (c) dibawah untuk $i = 1,2,\cdots,l$

(a) $E_i = \sum_{j=1}^{l} \alpha_j D_{ij}$

(b) $\delta\alpha_i = \min\{\max[\gamma(1 - E_i), -\alpha_i], C - \alpha_i\}$

(c) $\alpha_i = \alpha_i + \delta\alpha_i$

3. Kembali ke step 2 sampai nilai α mencapai konvergen

Pada algoritma di atas, γ adalah parameter untuk mengkontrol kecepatan proses learning. Konvergensi dapat didefinisikan dari tingkat perubahan nilai α.

6. KARAKTERISTIK SVM

Karakteristik SVM sebagaimana telah dijelaskan pada bagian sebelumnya, dirangkumkan sebagai berikut:

1. Secara prinsip SVM adalah linear classifier
2. Pattern recognition dilakukan dengan mentransformasikan data pada input space ke ruang yang berdimensi lebih tinggi, dan optimisasi dilakukan pada ruang vector yang baru tersebut. Hal ini membedakan SVM dari solusi pattern recognition pada umumnya, yang melakukan optimisasi parameter pada ruang hasil transformasi yang berdimensi lebih rendah daripada dimensi input space.
3. Menerapkan strategi *Structural Risk Minimization* (SRM)
4. Prinsip kerja SVM pada dasarnya hanya mampu menangani klasifikasi dua class.

7. KELEBIHAN DAN KEKURANGAN SVM

Dalam memilih solusi untuk menyelesaikan suatu masalah, kelebihan dan kelemahan masing-masing metode harus diperhatikan. Selanjutnya metode yang tepat dipilih dengan memperhatikan karakteristik data yang diolah. Dalam hal SVM, walaupun berbagai studi telah menunjukkan kelebihan metode SVM dibandingkan metode konvensional lain, SVM juga memiliki berbagai kelemahan. Kelebihan SVM antara lain sbb.

1. *Generalisasi*
   Generalisasi didefinisikan sebagai kemampuan suatu metode (SVM, neural network, dsb.) untuk mengklasifikasikan suatu pattern, yang tidak termasuk data yang dipakai dalam fase pembelajaran metode itu. Vapnik menjelaskan bahwa generalization error dipengaruhi oleh dua faktor: error terhadap training set, dan satu faktor lagi yang dipengaruhi oleh dimensi VC (Vapnik-Chervokinensis). Strategi pembelajaran pada neural network dan umumnya metode learning machine difokuskan pada usaha untuk meminimimalkan error pada training-set. Strategi ini disebut *Empirical Risk Minimization* (ERM). Adapun SVM selain meminimalkan error pada training-set, juga meminimalkan faktor kedua. Strategi ini disebut *Structural Risk Minimization* (SRM), dan dalam SVM diwujudkan dengan memilih hyperplane dengan margin terbesar. Berbagai studi empiris menunjukkan bahwa pendekatan SRM pada SVM memberikan error generalisasi yang lebih kecil daripada yang diperoleh dari strategi ERM pada neural network maupun metode yang lain.

2. *Curse of dimensionality*
   Curse of dimensionality didefinisikan sebagai masalah yang dihadapi suatu metode pattern recognition dalam mengestimasikan parameter (misalnya jumlah hidden neuron pada neural network, stopping criteria dalam proses pembelajaran dsb.) dikarenakan jumlah sampel data yang relatif sedikit dibandingkan dimensional ruang vektor data tersebut. Semakin tinggi dimensi dari ruang vektor informasi yang diolah, membawa konsekuensi dibutuhkannya jumlah data dalam proses pembelajaran. Pada kenyataannya seringkali terjadi, data yang diolah berjumlah terbatas, dan untuk mengumpulkan data yang lebih banyak tidak mungkin dilakukan karena kendala biaya dan kesulitan teknis. Dalam kondisi tersebut, jika metode itu “terpaksa” harus bekerja pada data yang berjumlah relatif sedikit dibandingkan dimensinya, akan membuat proses estimasi parameter metode menjadi sangat sulit.

   Curse of dimensionality sering dialami dalam aplikasi di bidang *biomedical engineering*, karena biasanya data biologi yang tersedia sangat terbatas, dan penyediaannya memerlukan biaya tinggi. Vapnik membuktikan bahwa tingkat generalisasi yang diperoleh oleh SVM tidak dipengaruhi oleh dimensi dari input vector [3]. Hal ini merupakan alasan mengapa SVM merupakan salah satu metode yang tepat dipakai untuk memecahkan masalah berdimensi tinggi, dalam keterbatasan sampel data yang ada.
3. Landasan teori
   Sebagai metode yang berbasis statistik, SVM memiliki landasan teori yang dapat dianalisa dengan jelas, dan tidak bersifat

black box.

4. *Feasibility*
   SVM dapat diimplementasikan relatif mudah, karena proses penentuan support vector dapat dirumuskan dalam QP problem. Dengan demikian jika kita memiliki *library* untuk menyelesaikan QP problem, dengan sendirinya SVM dapat diimplementasikan dengan mudah. Selain itu dapat diselesaikan dengan metode sekuensial sebagaimana penjelasan sebelumnya.

Disamping kelebihannya, SVM memiliki kelemahan atau keterbatasan, antara lain:

1. Sulit dipakai dalam problem berskala besar. Skala besar dalam hal ini dimaksudkan dengan jumlah sample yang diolah.
2. SVM secara teoritik dikembangkan untuk problem klasifikasi dengan dua class. Dewasa ini SVM telah dimodifikasi agar dapat menyelesaikan masalah dengan class lebih dari dua, antara lain strategi One versus rest dan strategi Tree Structure. Namun demikian, masing-masing strategi ini memiliki kelemahan, sehingga dapat dikatakan penelitian dan pengembangan SVM pada *multiclass-problem* masih merupakan tema penelitian yang masih terbuka.

8. APLIKASI SVM DALAM BIOINFORMATIKA

Pada paruh pertama tulisan ini, diskusi difokuskan pada dasar-dasar teori metode Support Vector Machine sebagai salah satu topik menarik yang tengah hangat dibicarakan dalam dunia komputer sains. Sebagaimana lazimnya perkembangan suatu teori, pertanyaan berikutnya adalah bagaimana teori tersebut diaplikasikan pada dunia nyata ? Apakah metode yang bagus secara teoritis itu mampu diaplikasikan untuk menyelesaikan suatu masalah nyata, ataukah teori tersebut hanya berhenti pada ujicoba dengan *toy problems* ? Dalam hal ini Vapnik memberikan ungkapan menarik yang perlu digarisbawahi : *"Nothing is more practical than a good theory"* [3]. Fakta yang membuktikan pernyataan Vapnik tersebut adalah semakin luasnya penelitian yang membuktikan kehandalan SVM dari sudut teori maupun aplikasi, dimana salah satu aplikasinya adalah dalam bidang bioinformatika.

Bioinformatika adalah suatu disiplin yang mengawinkan teknologi informasi dan teknologi biologi, untuk menjawab permasalahan kompleks dalam bidang biologi. Bioinformatika berkembang dari kebutuhan manusia untuk menganalisa data yang dewasa ini kuantitasnya makin meningkat. Akselerasi dari ketersediaan data biologi ini tidak terlepas dari peranan kerjasama harmonis teknologi informasi dan kemajuan di bidang bioteknologi. Sebagai contoh, pembacaan sekuen genom manusia oleh Celera Genomics dapat diselesaikan dalam waktu singkat, dibandingkan usaha konsorsium lembaga riset publik AS, Europa, dsb. [5]. Dengan melimpahnya data biologi tersebut, akan timbul pertanyaan : bagaimana kita memperoleh manfaat dari data ini ?

Rutherford D. Roger memberikan ungkapan menarik: *"We are drowning in information, but starving for knowledge"*. Ungkapan ini sejalan dengan situasi terkini di dunia bioteknologi. Melimpahnya ketersediaan data harus diikuti dengan tahapan mengekstrak informasi dari data tersebut. Selanjutnya informasi ini diolah agar dapat ditarik pengetahuan (knowledge) yang bermanfaat bagi masyarakat dan kemanusiaan. Misalnya dalam bidang klinis, pengetahuan yang diperoleh tersebut dipakai untuk mendesain obat atau terapi medis yang sesuai dengan kebutuhan sang pasien (*tailor made medicine*), untuk identifikasi agen penyakit baru, untuk diagnosa penyakit baru [6].

Untuk mewujudkan proses transformasi data-informasi-knowledge ini, teknologi informasi memiliki peranan penting. Hal ini terlihat dari banyaknya paper yang membahas aplikasi metode komputasi untuk menganalisa data biologi seperti statistical pattern recognition, artificial neural network, SVM, dsb.[8] Tiap metode memiliki sisi kelebihan dan kekurangan, dan metode yang tepat harus dipilih dengan memperhatikan karakteristik problem biologi tersebut.

Berbagai penelitian dilakukan untuk mengevaluasi potensi SVM dalam analisa data biologi, antara lain:

1. Analisa ekspresi gen [9]
2. Deteksi homologi protein [4]
3. Prediksi struktur protein [10]

Makalah ini mengambil contoh bahasan aplikasi SVM pada masalah pertama, yaitu analisa data ekspresi gen, karena kesesuaiannya dengan kemampuan SVM dalam mengolah informasi berdimensi tinggi. Data biologi yang diolah dan dianalisa oleh SVM diperoleh dari eksperimen microarray yang memungkinkan pengamatan ekspresi ribuan gen sekaligus, misalnya pada sel yang diambil dari penderita penyakit kanker. Pemanfaatan microarray membuka kemungkinan untuk mengetahui kuantitas maupun kualitas transkripsi satu gen, sehingga dapat diidentifikasikan : gen-gen apa saja yang aktif terhadap perlakuan tertentu, misalnya timbulnya kanker. Informasi ini merupakan pertimbangan penting bagi ahli medis untuk mengetahui mekanisme timbulnya penyakit, dan menentukan terapi mana yang paling tepat bagi si pasien.

Proses dalam analisa micorarray secara sederhana dapat diuraikan sebagai berikut. Pertama-tama mRNA yang disolasi dari sampel dikembalikan dulu dalam bentuk DNA menggunakan reaksi reverse transcription. Selanjutnya melalui proses hibridisasi, hanya DNA yang komplementer saja yang akan berikatan dengan DNA di atas chip. DNA yang telah diberi label warna berbeda ini akan menunjukkan pattern yang unik. Dengan memanfaatkan teknologi pengolahan citra (image processing), pattern ini selanjutnya ditransfer ke dalam ekspresi numerik untuk diolah dengan berbagai metode pattern recognition (dalam hal ini SVM).

Dalam studi analisa ekspresi gen, ada tiga hal yang merupakan bahasan menarik dari sudut pattern recognition [11]:

1. Mungkinkah dengan data ekspresi gen dari microarray, kita memprediksi suatu class, misalnya apakah seorang pasien tersebut terkena kanker atau tidak, atau menentukan status mutasi p53 pasien, dsb.
2. Kalau hal tersebut memungkinkan untuk dilaksanakan, berapakah tingkat akurasi yang mungkin dicapai ?
3. Bagaimana menentukan kandidat gen yang memiliki potensi kedokteran/farmasi ?

Bahasan dalam makalah ini dibatasi pada tema pertama, dengan mengevaluasi performa SVM dalam klasifikasikan ekspresi gen.Tema ini tergolong tema pattern recognition yang sangat sulit, karena memiliki karakteristik

1. Data observasi berdimensi tinggi : manusia memiliki sekitar 31 ribu jenis gen, sehingga setiap pengukuran memberikan satu titik pada ruang vektor berdimensi sekitar 31 ribu
2. Noisy
3. Unbalanced, dalam artian sampel class positif seringkali tersedia dalam jumlah yang jauh lebih sedikit daripada sampel class negatif.

Karakteristik ini menjadi latar belakang mengapa SVM mendapat perhatian besar dari kalangan bioinformatika. Potensi SVM sebagaimana diuraikan pada halaman yang terdahulu memberikan harapan untuk dapat menyelesaikan problem dengan karakteristik tersebut. Salah satu paper yang membahas aplikasi SVM dalam analisa data ekspresi gen adalah sebagaimana yang dilakukan oleh group Terrence S. Furey.

9. RISET GROUP TERRENCE S. FUREY : ANALISA EKSPRESI GEN MEMAKAI SVM

Salah penelitian bioinformatika mengenai aplikasi SVM dalam analisa gene-expression adalah sebagaimana yang dilakukan oleh group Terrence S. Furey, dimuat di journal Bioinformatics [9]. Group Furey memakai SVM dengan dot product kernel (linear SVM) untuk menganalisa vektor berdimensi ribuan yang dibentuk oleh ekspresi gen diperoleh dari eksperimen microarray. Evaluasi dilakukan terhadap tiga database : Ovarian tissue dataset, human acute leukemia (Golub dataset), dan yang ketiga adalah human tumour dan normal colon tissue dataset. Masing-masing eksperimen dapat dirangkumkan sebagai berikut:

1. Ovarian dataset

Sampel yang berasal dari ovarian cancer tissue, normal ovarian tissue dan normal tissue non-ovarian yang lain, total sebanyak 31 sampel.

Tiap data terdiri dari 97,802 cDNA untuk masing-masing tissue, dengan demikian membentuk ruang vektor berdimensi 97,802. Untuk mereduksi dimensi dari feature vector ini, dilakukan feature subset selection (FSS) dengan memilih sekumpulan feature yang paling signikan. Furey memilih strategy single best criterion, yaitu tiap feature dievaluasi secara terpisah dengan menentukan mana yang paling berpengaruh pada class separability. Walaupun metode ini memiliki banyak sisi lemah, dan mengabaikan kontribusi yang dimiliki secara berkelompok, tapi metode FSS ini mungkin paling mudah dilakukan dalam kondisi dimensi vektor yang hampir mencapai 100,000. Hasil dari FSS memperlihatkan bahwa dari 97,802 cDNA, cukup diperlukan 50 buah feature (cDNA) yang memiliki score signifikansi tertinggi. Selanjutnya estimasi parameter dilakukan dengan *leave-one-out cross validation.*

Hasil eksperimen menunjukkan bahwa satu sampel dari kelompok normal ovarian tissue selalu gagal diklasifikasikan. Hasil analisa dari kegagalan ini menunjukkan bahwa margin dari misclassification cukup besar. Hal ini berarti SVM sangat yakin, bahwa sampel ini tergolong cancerous tissue. Dengan mengeliminasi satu sampel dari non-ovarian normal tissue yang kualitasnya diragukan, total akurasi SVM 90% (misklasifikasi : 3 dari total 30 sampel).

2. Human acute leukemia

Data pada eksperimen ini berasal dari studi yang dilakukan oleh Golub [12], dan tersedia online di internet. Data diambil dari 72 pasien penderita myeloid leukimia (AML) dan acute lymphoblastic leukimia (ALL). Data ini dibagi dalam dua kelompok: training set (27 ALL dan 11 AML), dan test set (20 ALL dan 14 AML). Tiap sampel terdiri dari vektor berdimensi 7129 yang berasal dari ekspresi gen si pasien sebagai hasil analisa Affymetrix high-density oligonucleotide microarray. FSS dilakukan pada training set untuk menseleksi feature yang signifikan, dengan metode sebagaimana penjelasan sebelumnya.

SVM dilatih dengan data dari training set, dan performa-nya dievaluasi pada test set. Hasil dari eksperimen menunjukkan bahwa SVM mengklasifikan secara benar antara 30 sampai 32 dari total 34 sampel pada test set.

3. Human tumour dan normal colon tissue dataset

Data pada eksperimen ini berasal dari studi yang dilakukan oleh Alon [13], yang terdiri dari 40 tissue tumor dan 22 tissue normal colon. Tiap sampel berasal dari hasil analisa Affymetrix oligonucleotide arrays terhadap 6500 gen manusia. Dari 6500 gen ini, sebanyak 2000 diantaranya yang diseleksi terlebih dahulu berdasarkan kriteria tertentu, dipergunakan untuk keperluan klasifikasi.

Performa SVM dievaluasi dengan metode leave-one-out crossvalidation, dan sebagai hasil 56 sampel berhasil diklasifikasikan secara benar (misklasifikasi : 6 sampel). Selanjutnya percobaan diulangi dengan memakai subset yang terdiri dari 1000 dari total 2000 feature pada tiap vektor. Hasil pada eksperimen kedua ini sama dengan sebelumnya, yaitu 6 sampel saja yang tidak dapat diklasifikasikan secara benar. Dari ke-6 sampel ini tiga diantaranya normal tissue dan tiga yang lain tumor tissue.

Tabel 2- Sebagian dari hasil eksperimen Furey : komparasi SVM dan perceptron pada studi analisa ekspresi gen

Dataset	Dimensi	Error Num. of tumor tissue		Error Num. of normal tissue	
		SVM	Perceptron	SVM	Perceptron
Ovarian	97,802	3	4.8	5	4.6
Golub dataset	7,129	0	2.8	0	0.6
Colon dataset	2,000	3	3.7	3	3.8

Selanjutnya Furey menguji performa perceptron (artificial neural network), pada ketiga dataset yang sama. Sebagian hasil dari eksperimen tersebut dirangkumkan pada Tabel 2. Perhatikan bahwa tingkat error pada tabel adalah hasil rata-rata dari lima kali eksperimen, yang dilakukan dengan mengubah urutan sampel. Secara keseluruhan SVM memberikan hasil yang lebih baik, kecuali pada normal tissue data Ovarian.

Namun demikian, sebagaimana dijelaskan oleh Furey, karena evaluasi ini dilakukan pada data yang jumlahnya relatif sedikit, hasil pada Tabel 2 belum dapat dikatakan valid untuk memberikan kesimpulan bahwa SVM lebih superior dibandingkan metode yang lain. Dengan makin banyaknya ketersediaan data ekspresi gen hasil analisa microarray, diharapkan SVM dapat diujicoba pada eksperimen yang data skala besar, sehingga hasil komparasi pada eksperimen tersebut tidak diragukan validitasnya.

10. KESIMPULAN

Makalah ini memperkenalkan teori dasar Support Vector Machine (SVM), sebagai salah satu topik yang dewasa ini banyak mendapat perhatian sebagai state of the art dalam bidang pattern recognition. Kelebihan SVM dibandingkan metode yang lain terletak pada kemampuannya untuk menemukan hyperplane terbaik yang memisahkan dua buah class pada feature space yang ditunjang oleh strategi Structural Risk Minimization (SRM).

Pada paruh kedua dari makalah ini, dibahas aplikasi SVM pada bioinformatika, khususnya analisa ekspresi gen yang diperoleh dari eksperimen microarray terhadap pasien penderita penyakit kanker. Eksperimen ini dilakukan oleh group Terrence S. Furey, dengan tujuan memakai SVM untuk mengklasifikan apakah suatu pasien terkena penyakit kanker atau tidak, berdasarkan hasil analisa microarray terhadap sel pasien tersebut. Secara umum, SVM menunjukkan hasil yang lebih baik daripada perceptron. Walaupun demikian, dikarenakan jumlah sampel yang relatif sedikit, hasil eksperimen itu belum dapat memberikan kesimpulan final bahwa SVM superior terhadap perceptron.

Walaupun eksperimen dengan data microarray secara statistik masih terdapat kelemahan, terutama dari sudut keterbatasan data, dan mahalnya cost yang diperlukan untuk analisa, evaluasi SVM merupakan suatu usaha yang sangat berharga untuk mengklarifikasikan masalah yang timbul. Analisa pada data skala kecil ini akan memudahkan bagi kita untuk menemukan sisi-sisi lemah dari metode yang dipakai. Seiring dengan kemajuan IT dan bioteknologi modern yang mencengangkan akhir-akhir ini, diperkirakan pada tahun-tahun mendatang, dunia ilmu pengetahuan akan semakin terbanjiri dengan data biologi, sedangkan teknologi informasi pun akan melaju dengan kencang. Dalam situasi ini, dengan memanfaatkan teknologi informasi secara tepat, diharapkan data biologi tersebut dapat diolah menjadi suatu informasi, dan seterusnya ditransformasikan sebagai suatu pengetahuan yang dapat ditarik manfaatnya bagi kesehatan dan kesejahteraan umat manusia.

REFERENSI

1- Byun H., Lee S.W., “A Survey on Pattern Recognition Applications of Support Vector Machines”, International Journal of Pattern Recognition and Artificial Intelligence, Vol.17, No.3, 2003, pp.459-486

2- Tsuda K., “Overview of Support Vector Machine”, Journal of IEICE, Vol.83, No.6, 2000, pp.460-466 (in Japanese)

3- Vapnik V.N., “The Nature of Statistical Learning Theory”, 2nd edition, Springer-Verlag, New York Berlin Heidelberg, 1999

4- Cristianini N., Taylor J.S., “An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods”, Cambridge Press University, 2000

5- Vijayakumar S, Wu S, “Sequential Support Vector Classifiers and Regression”, Proc. International Conference on Soft Computing (SOCO'99),Genoa, Italy, pp.610-619, 1999

6- Witarto A.B, “Bioinformatika: Mengawinkan teknologi informasi dengan bioteknologi”, http://ilmukomputer.com , June 2003 (in Indonesian)

7- Utama A., “Peranan bioinformatika dalam dunia kedokteran”,

http://ilmukomputer.com , August 2003 (in Indonesian)

8- Nugroho A.S., “Bioinformatika dan pattern recognition”, http://ilmukomputer.com , July 2003 (in Indonesian)

9- Furey T.S, et al., “Support vector machine classification and validation of cancer tissue samples using microarray expression data”, Bioinformatics, Vol.16, No.10, 2000, pp.906-914

10- Ward J.J., et al., “Secondary structure prediction with support vector machine”, Bioinformatics, Vol.19, No.13, 2003, pp.1650-1655

11- Maeda E., “Gene expression analysis and feature selection”, IEICE Technical Report, PRMU-2003-37, Vol.103, No.150, 2003, pp.57-62 (in Japanese)

12- Golub T. et al., “Molecular classification of cancer : class discovery and class prediction by gene expression monitoring”, Science, Vol. 286, 1999, pp.531-537

13- Alon U. et al., “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays”, Proc. Natl. Acad. Sci. USA, No.96, 1999, pp.6745-6750