ANALISA INFORMASI DIMENSI TINGGI PADA BIOINFORMATIKA MEMAKAI SUPPORT VECTOR MACHINE

Anto Satriyo Nugroho[1) 3)], Dwi Handoko[1)] dan Arief B. Witarto[2)]
Email : asnugroho@ieee.org, dwih@inn.bppt.go.id, witarto@yahoo.com

1. Pusat Pengkajian & Penerapan Teknologi Informasi & Elektronika, BPP Teknologi, Gedung II Lt.21, Jalan M.H. Thamrin No. 8, Jakarta
2. Laboratorium Rekayasa Protein, Pusat Penelitian Bioteknologi, LIPI, Jalan Raya Bogor Km.46, Cibinong, Kab.Bogor,
3. School of Life System Science & Technology, Chukyo University, Japan

ABSTRACT

Bioinformatika timbul dari kebutuhan akan metode untuk mengolah data-data biologi yang dewasa ini sangat pesat bertambah. Berbagai penemuan di bidang bioteknologi modern memacu pertumbuhan ketersediaan data, yang pada gilirannya memerlukan support dari teknologi informasi untuk mengubah data tersebut menjadi suatu informasi, dan selanjutnya menjadi suatu pengetahuan yang bermanfaat bagi kemanusiaan. Salah satu metode pattern recognition yang dipakai adalah Support Vector Machine (SVM). Metode ini 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 aplikasi menempatkannya sebagai state of the art dalam pattern recognition. Tulisan ini membahas teori dasar SVM dan aplikasinya dalam bioinformatika, khususnya pada analisa ekspresi gen dimensi tinggi yang diperoleh dari analisa microarray.

Keywords: pattern recognition, support vector machine, bioinformatika

1. PENDAHULUAN

Bioinformatika merupakan salah satu topik yang sedang hangat dibicarakan dewasa ini. Sebagai suatu disiplin ilmu, bioinformatika melibatkan dua aspek, yaitu aspek teknologi informasi (TI) dan aspek biologi. Aplikasi dari bioinformatika ini meliputi berbagai bidang, antara lain bidang farmasi, kedokteran dan pertanian.

Paper ini dipublikasikan di Proc. of National Conference on Information & Communication Technology (ICT) for Indonesia/e-Indonesia Initiatives-II 2005, pp.427-435

Bioinformatika timbul dari kebutuhan akan metode untuk mengolah data-data biologi yang dewasa ini sangat pesat bertambah. Berbagai penemuan di bidang bioteknologi modern memacu pertumbuhan ketersediaan data, yang pada gilirannya memerlukan support dari teknologi informasi untuk mengubah data tersebut menjadi suatu informasi, dan selanjutnya menjadi suatu pengetahuan yang bermanfaat bagi kemanusiaan.

Berkaitan dengan arus data yang deras mengalir ini, sangat menarik untuk mencermati perbandingan perkembangan pesat semikonduktor dan genetika. Di bidang komputer, dikenal Moore's law yang memprediksi bahwa setiap 18 bulan, jumlah transistor per satuan area pada IC, selalu berlipat dua. Dengan kata lain, kemampuan komputer akan berlipat jadi dua kali setiap 18 bulan. Pengamatan Gordon Moore ini dikeluarkan tahun 1965, dan secara ajaib selalu terbukti berlaku, setidaknya selama dua dekade ini. Hal ini menjadi motivasi dan misi Intel Corporation untuk selalu memenuhi tuntutan dari ramalan Moore [1].

Analog dengan Moore's law, di dunia biologi, dikenal juga ramalan menarik dari Prof. R. Dawkins (Oxford University), yang mencoba menarik korelasi antara jumlah nucleotide-base yang bisa dibaca dengan dana £ 1000, terhadap waktu. Pada tahun 1965, diperlukan £ 1 untuk membaca 1 huruf pada RNA bakteri. Tahun 1975, £ 10 untuk satu huruf pada virus code. Pada 1985, membaca 1 huruf pada nematode memerlukan £ 1, dan pada tahun 2000 diperlukan 0.10 £ untuk membaca 1 huruf pada human genome project. Kalau ramalan ini benar, maka pada tahun 2012 diprediksi dengan £ 1000 dapat dianalisa E.coli yang terdiri dari 200 ribu bases. Sedangkan pada tahun 2050, diperlukan £ 1000 untuk membaca seluruh nucleotide base pairs manusia. Prediksi ini dikenal sebagai "Son of Moore's law for genetics", menggambarkan perkembangan yang pesat di dunia bioteknologi [2].

Berangkat dari ketersediaan data genome dalam jumlah besar ini, terminologi *biological-datamining* menjadi sangat populer. Datamining didefinisikan sebagai proses otomatis mengekstrak suatu informasi dari sekumpulan data yang berjumlah besar. Salah satu aplikasi dari penerapan datamining di bioinformatika ini adalah pengembangan industri farmasi dan kedokteran. Informasi yang diekstrak ini dapat dimanfaatkan dalam industri medis, misalnya menekan resiko timbulnya efek samping dari terapi kanker.

Tulisan ini membahas potensi metode support vector machine yang sering dipakai dalam analisa data dan datamining di bioinformatika. Teori SVM dibahas pada bab-bab awal, dan contoh aplikasinya dibahas pada bab 6. Data yang diolah diperoleh dari analisa microarray yang menghasilkan data pada ruang dimensi tinggi. Bagian akhir tulisan ini berupa kesimpulan dari studi yang dilakukan.

2. PATTERN RECOGNITION MEMAKAI SVM

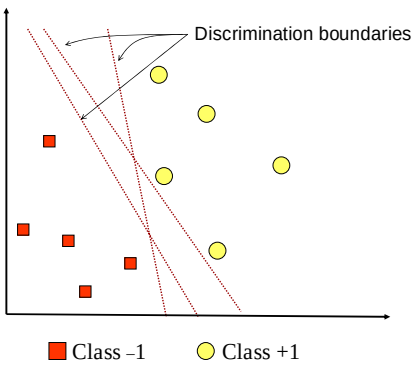
Pattern Recognition merupakan salah satu bidang dalam komputer sains, yang memetakan suatu data ke dalam konsep tertentu yang telah didefinisikan sebelumnya. Konsep tertentu ini disebut *class* atau *category*. Aplikasi pattern recognition sangat luas, di antaranya mengenali suara dalam sistem sekuriti, membaca huruf dalam OCR, mengklasifikasikan penyakit secara otomatis berdasarkan hasil diagnosa kondisi medis pasien dan sebagainya. Berbagai metode dikenal dalam pattern recognition, seperti linear 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) [3][4]. 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 [5][6].

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.

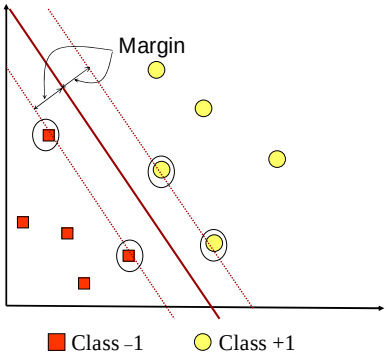
3. KONSEP SUPPORT VECTOR MACHINE

Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari hyperplane 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 kelompok tersebut. Berbagai alternatif garis pemisah (*discrimination boundaries*) ditunjukkan pada gambar 1-a.

(a)



(b)



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

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.

$$L(\vec{w}, b, \alpha) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{l} \alpha_i (y_i((\vec{x}_i.\vec{w} + b) - 1)) \quad (i = 1,2,\cdots,l) \qquad (6)$$

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.

4. SOFTMARGIN

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.

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

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.

5. 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.

$$\Phi : \Re^d \rightarrow \Re^q \quad d < q \tag{11}$$

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

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

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.

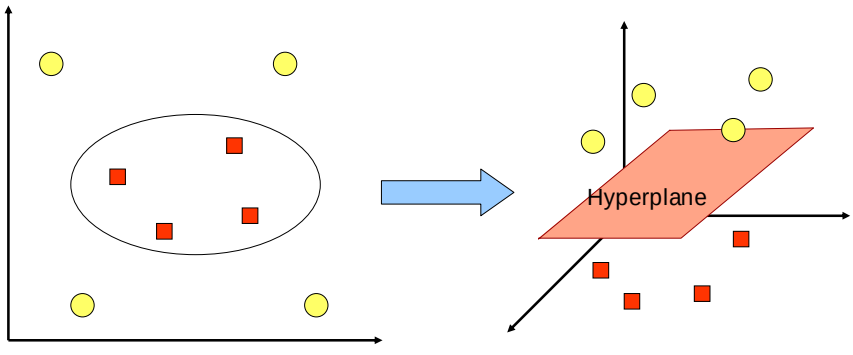
Selanjutnya hasil klasifikasi dari data $\vec{x}$ diperoleh dari persamaan berikut :

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

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

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

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 $\alpha_i \geq 0$.



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

6. APLIKASI SVM DALAM BIOINFORMATIKA

6.1 Ruang vektor berdimensi tinggi pada data microarray

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

1. Analisa ekspresi gen [7]
2. Deteksi homologi protein [6]
3. Prediksi struktur protein [8]

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 [9]:

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

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. Imbalanced, 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.

6.2 Aplikasi SVM pada analisa ekspresi gen memakai database Human Acute Leukemia

Data pada eksperimen ini berasal dari studi yang dilakukan oleh Golub [10], 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.

Dua eksperimen dilakukan. Pertama-tama, prediksi dilakukan berdasarkan seluruh informasi ekspresi gen pasien, dengan kata lain input vektor terdiri dari 7129 dimensi. Selanjutnya, pada eksperimen kedua, Feature Subset Selection (FSS) dilakukan pada training set untuk menseleksi feature yang signifikan, dengan metode Sequential Forward Selection (SFS).

Experimen 1

Pada eksperimen ini, tiga metode dipakai untuk melakukan prediksi, masing-masing 1 Nearest Neighbor Classifier, Multilayer Perceptron dengan hidden neuron 100, dan Support Vector Machine. SVM dalam eksperimen ini memakai Kernel Polynomial pada derajat 1.

Tiap feature menunjukkan rentang nilai yang sangat bervariasi. Hal ini dapat menyebabkan terlalu dominannya pengaruh feature tertentu yang

memiliki rentang nilai besar terhadap feature lain yang memiliki rentang nilai jauh lebih kecil. Untuk mengatasi hal tersebut, sebelumnya dilakukan normalisasi dengan mengurangkan data pada feature tertentu dengan rata-rata nilai pada feature tersebut dan dibagi dengan standard deviasi-nya. Dengan cara demikian, dapat dihindarkan pengaruh yang terlalu besar dari feature tertentu.

$$x_i' = \frac{x_i - x_{i\,ave}}{\sigma_i} \qquad i = 1,2,...,7129 \qquad (16)$$

Tabel 1- Recognition Rate class Acute Lymphoblastic leukimia (ALL) dan Acute Myeloid Leukimia (AML) pada tiga jenis metode : 1 Nearest Neigbour Classifier, Support Vector Machine dan Multilayer Perceptron

Method	**RR ALL**	**RR AML**	**RR TOTAL**
1-NNClassf.	95%	64%	78%
SVM	85%	100%	92.2%
MLP	55%	57.1!	56.1%

Hasil eksperimen ini ditunjukkan pada Tabel 1. RR Total merupakan geometrical mean dari RR ALL dan RR AML. Geometrical Mean ini dipilih karena terdapat ketidaksetimbangan jumlah pattern masing-masing class dalam test set (ALL : 20 sample, AML: : 14 sample). Dari tabel di atas diketahui bahwa hasil terbaik dicapai oleh Support Vector Machine. Jumlah support vector yang terpilih sebanyak 30 dari total 38. Tiga pattern yang gagal diklasifikasikan oleh SVM berasal dari kategori Acute Lymphoblastic Leukimia.

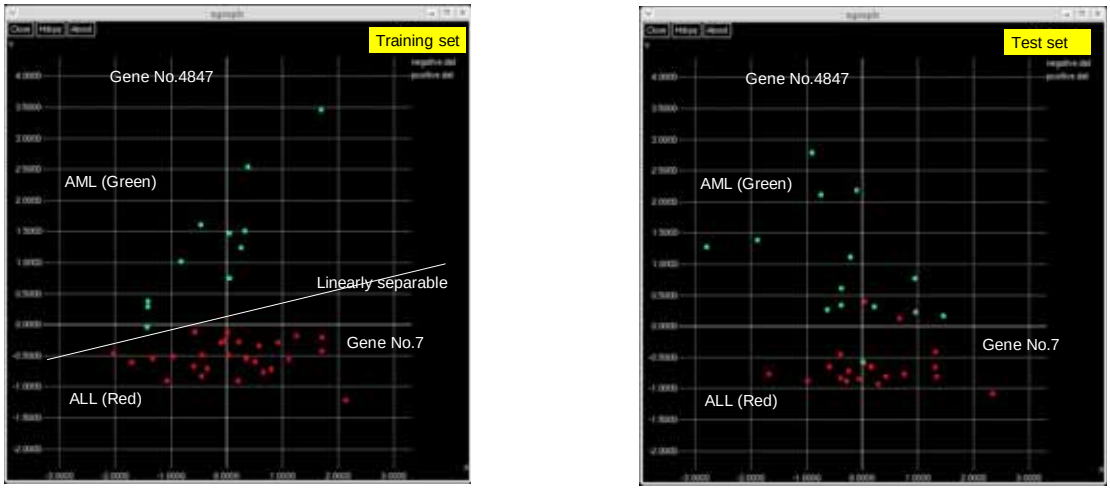
Experimen 2

Pada eksperimen kedua, dilakukan Feature Subset Selection untuk memilih gene-gene yang signifikan. Algoritma yang dipakai adalah Sequential Forward Selection, yaitu memilih satu persatu gene yang paling signifikan diukur berdasarkan kriteria subset tertentu terhadap kombinasi gene tersebut dengan yang telah terpilih sebelumnya. Algorithma Sequential Forward Selection dapat dijabarkan sebagai berikut. Misalnya sekumpulan feature telah diseleksi dari semesta pengukuran $Y = \{y_j \mid j = 1,2,...,D\}$ untuk membuat feature set X_k. Selanjutnya feature ke $k+1$ dipilih dari $Y - X_k$, sedemikian hingga

$$J(X_{k+1}) = \max \quad J(X_k \cup y_j)$$
$$y \in Y - X_k \qquad (17)$$

Dalam persamaan di atas $J(X_k)$ adalah nilai fitness dari subset X_k yang diukur berdasarkan kriteria tertentu. Inisialisasi algoritma ini adalah $X_o = \phi$. Pada studi ini, kriteria evaluasi suatu subset diukur berdasarkan leave one out 1 nearest neighbour classification rate

Sebagai hasilnya, saat algoritma ini dijalankan, dua buah gen terseleksi sebagai yang paling signifikan, dan menghasilkan score classification rate 100% (leaveone out 1 nearest neighbour) terhadap training set. Kedua gen tersebut adalah No.7 (AFFX-BioDn-3_at) dan No. 4847 (X95735_at). Distribusi data pada bidang yang dibentuk oleh kedua gen tersebut adalah sebagaimana ditunjukkan pada gambar 3. Gambar 3a menunjukkan bahwa data pada kedua class pada training set dapat terpisah sempurna secara linier. Gambar 3b menunjukkan bahwa pada data pada test set tidak dapat terpisah secara linier.
Support Vector Machine dilatih dengan data pada training set dan diperoleh 3 support vector, dengan classification rate 100% . Selanjutya saat diujicobakan terhadap data pada test set, diperoleh recognition rate 86.4%. Score masing-masing class adalah ALL: 95%, AML: 78.6%.



Gambar 3– Distribusi data dalam training set (a) dan test set (b) pada bidang yang dibentuk oleh gen No.7 dan gen No.4847

6.3 Diskusi

Dari eksperimen 1 dan eksperimen 2 di atas, dirangkumkan beberapa hal sbb.

1. Dalam percobaan diatas kami tetap memakai polynomial kernal derajat 1. Kami mencoba juga kernel yang lain, akan tetapi tidak memberikan hasil yang lebih baik dibandingkan kernel derajat 1. Diketahui bahwa kemampuan Support Vector Machine (SVM) sangat sensitif terhadap kernel yang dipilih.
2. Secara umum, SVM menunjukkan hasil yang lebih baik daripada perceptron.maupun 1 nearest neighbour classifier. Walaupun demikian, dikarenakan jumlah sampel yang relatif sedikit, hasil eksperimen itu belum dapat memberikan kesimpulan final bahwa SVM superior terhadap dua metode yang lain.
3. Walaupun Feature Subset Selection telah dilakukan, tetapi tidak terlihat adanya peningkatan classification rate, sebaliknya terlihat sedikit menurun. SFS dalam hal ini tidak berhasil memberikan kontribusi positif terhadap classification rate, karena nilai maksimal $J(X_k)$ telah dicapai di periode awal ($k = 2$), sehingga proses seleksi berhenti. Walaupun demikian, kedua gen yang terseleksi memiliki daya tarik untuk dianalisa lebih lanjut korelasinya dengan kedua jenis penyakit leukimia.
4. MLP bekerja dengan memetakan data kepada ruang vektor berdimensi yang lebih rendah yang dibentuk oleh neuron pada hidden layer. Selanjutnya classification boundary dicari pada ruang vektor dimensi rendah tersebut. Terhadap pendekatan ini, SVM memperlihatkan cara kerja yang bertolak belakang. Data dipetakan terhadap ruang vektor baru yang dibentuk oleh kernel, yang berdimensi jauh lebih tinggi. Data pada ruang terbaru ini memiliki kemungkinan lebih mudah untuk dapat dipisahkan secara linear (Teori Cover). Classification boundary pada ruang vektor baru tersebut dicari dengan metode optimisasi tertentu.
5. Analisa teoretik dari tingkat generalisasi SVM menunjukkan bahwa batas maksimal expected error classification rate tidak dipengaruhi oleh dimensi dari data. Dengan demikian, pemakaian SVM pada dimensi tinggi, tidak akan menyebabkan pengaruh negatif yang terjadi karena curse of dimensionality.

7. 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.

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 akhir-akhir ini demikian pesat, 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.

8. REFERENSI

1- Moore's Law http://www.intel.com/research/silicon/mooreslaw.htm

2- Son of Moore's Law http://myweb.tiscali.co.uk/royalphil/rps/summaries/evolution.htm

3- 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

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

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

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

7- 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

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

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

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