ISSN 2085-4552

Review of Multiple Input Multiple Output Causal Strategies for Gene Selection

Ranny
Universitas Indonesia, Depok, Indonesia

Diterima 02 Desember 2011
Disetujui 12 Desember 2011

***Abstract*—Feature extraction is one of a problem in bioinformatics. Bioinformatics research using many feature in their dataset. In this research we develop a method to find the interaction of the feature in the dataset. The method using multiple input and multiple output (MIMO) to select the feature of gene. The purpose of the experiment is to measure the effective and correctness of the MIMO method. The breast cancer dataset was used in the experiment. The result of the experiment show that the MIMO can improve the gene selection as the feature.**

***Index Terms*—Bioinformatics, breast cancer dataset, gene selection.**

I. PENDAHULUAN

Analisa pada bidang bioinformatik memiliki beberapa jenis analisa dataset, misalnya dengan menganalisa ekspresi gen pada microarray[1]. Selain itu, juga dapat dilakukan dengan melakukan perbandingan array hybridisasi genomic. Jenis-jenis analisa ini menggunakan dataset dengan jumlah fitur yang banyak tapi memiliki jumlah sampel yang terbatas. Untuk itu dilakukan beberapa pendekatan analisa, yaitu dengan melakukan regulasi atau dengan melakukan beberapa strategi untuk menentukan fitur pada dataset. Salah satu strategi pemilihan fitur pada dataset bioinformatik adalah menggunakan strategi pengurutan yaitu menentukan urutan variabel berdasarkan skor relevansi. Ukuran dari skor relevansi yang biasa digunakan berdasarkan keterkaitan antar informasi, korelasi atau nilai-p antara input dengan output target. Salah satu kekurangan dari strategi pengurutan adalah tidak mampu menentukan hubungan sebab dan akibat antara data input dengan target. Strategi pengurutan hanya terbatas pada penentuan skor relevansinya saja.

Kasus yang dikembangkan pada paper ini adalah mengelompokkan data bioinformatik (microarray) berdasarkan perbedaan kelas tumor atau prediksi efek sebuah terapi yang dilihat dari profil gen ekspresinya. Pada kasus ini varibel input yang digunakan merepresentasikan jumlah dari probe gen. Variabel input tersebut memiliki jumlah yang sangat besar sedangkan jumlah sempel (merepresentasikan jumlah pasien tumor) sangat terbatas. Hal ini menyebabkan penentuan gen relevan yang tepat sangatlah sulit. Dengan mampu melihat hubungan sebab akibat antara varibel pada data analis yang dibangun bukan hanya akan meningkatkan akurasi prediksi pada dunia kedoketeran tapi juga akan mampu menentukan terapi yang tepat digunakan pada sebuah penyakit.

Pada umumnya penelitian yang dilakukan adalah dengan menentukan skor pasangan gen. Setelah mendapatkan skor dari pasangan gen tersebut akan dihitung tingkat kebenarannya untuk menentukan keterkaitan antar gen. Namun, metode tersebut memiliki kelemahan pada masalah-masalah yang memiliki dimensi tinggi seperti analisis microarray. Untuk menangani masalah ini maka digunakan sebuah perhitungan estimasi interaksi dengan menggunakan pendekatan multiple input-multiple output. Umumnya pada dunia kedokteran pendekatan multiple input-multiple output digunakan untuk menentukan tingkat keparahan dari sebuah penyakit dan menentukan terapi pengobatan yang tepat.

Pada paper ini akan digunakan contoh kasus dari pasien kanker payudara. Tujuan dari paper ini adalah mengukur interaksi antara input beberapa gen set dengan data biological (ukuran tumor dan tingkat histological). Interaksi diukur menggunakan nilai probabilitas dari semua data, baik data gen set dengan data bilogical. Dengan adanya pengukuran interkasi tersebut akan dipilih gen-gen yang mempengaruhi perkembangan tumor pada tubuh seorang pasien. Data yang digunakan pada eksperimen diambil dari data microarray dan data biological penderitan kanker payudara.

II. MULTIPLE INPUT MULTIPLE OUTPUT FOR GENE SELECTION

Pada paper pengukuran sebuah interksi antar variabel input dan target menjadi tujuan dari penelitian. Banyaknya variabel input yang digunakan akan menentukan seberapa keterkaitan antara variabel input

dan target tersebut. Untuk mendapatkan jumlah yang variabel yang optimal maka akan didapatkan dengan mendapatkan nilai interaksi yang maksimal dari setiap variabel input dengan target. Untuk itu digunakan rumus (1) untuk menentukan jumlah dari variabel yang optimal (x^*_{d+1}).

$$x^*_{d+1} = arg \max_{x_k \in X - X_s} s(x_k) = arg \max_{x_k \in X - X_s} I(x_k; y_1) \quad (1)$$

Penentuan jumlah variabel yang tepat pada multiple input dan multiple output dengan dimensi yang besar, maka berdasarkan rumus (1) terlebih dahulu dilakukan pencarian variabel optimal untuk multiple input dan multiple output. Dengan menemukan jumlah variabel yang tepat akan didapat nilai interaksi yang juga optimal dari semua input dan target.

Variabel input yang digunakan adalah data pada microarray, sedangkat target adalah data biological berupa ukuran tumor serta tingkat histological dari beberpa pasien penderita kanker payudara.

Bagan umum dari algoritma yang digunakan pada paper terdapat pada Gambar 1. Algoritma dimulai dengan memberikan input berupa variabel input dan variabel target. Kemudian dilanjutkan dengan melakukan proses perhitungan nilai interaksi. Perhitungan dilanjutkan dengan menambahkan nilai λ. Pada percobaan nilai λ yang digunakan dimulai dari 0, 0.2, 0.4, 0.6, 0.8, 0.9, 1, dan 2. Tahap selanjutnya adalah dengan melihat apakah seiring dengan penambahan nilai λ, maka nilai skor interaksi juga bertambah. Jika nilai skor interaksi juga bertambah, maka variabel input dan target terdapat interaksi dan kandidat gen terpilih.

III. EKSPERIMEN

Algoritma yang digunakan pada paper ini adalah algoritma MIMO (Gambar 1) dengan menggunakan nilai λ: 0.1;0.2;0.4;0.6;.0.8;1;2. Data yang digunakan pada eksperimen ini diambil dari data microarray sebanyak enam dataset pendirita kanker payudara. Berikut table dataset yang digunakan [2]:

Tabel 1 Dataset untuk eksperimen

Dataset	Pasien
UPP	251(110)
STK	159
VDX	344
UNT	137(92)
MAINZ	200
TRANSBIG	198

Eksperimen yang dilakukan pada paper dibagi menjadi beberapa kelompok. Berdasarkan validasi meta-analisis, eksperimen dibagi menjadi dua kelompok, yaitu

1. *Holdout*: pada eksperimen *holdout* ini, digunakan sebanyak 100 data untuk pelatihan dan pengujian. Setangah data (50) digunakan untuk pelatihan dan sisannya digunakan untuk pengujian.

2. *Leave one dataset out*: eksperimen *leave one data set out* mengguanakan 100 data untuk palatihan. Pada saat pengujian, data yang telah terpilih akan dikeluarkan sebagai data pelatihan yang kemudian sisa datanya akan dilakukan pengujian kembali, sampai data pelatihan habis.

Dari kedua kelompok tersebut dengan niliai λ yang telah ditentukan di atas, masing-masing hasil kualitas pemilihan tersebut akan diuji. Pengujian kualitas pemilihan menggunakan Naive Bayes dengan empat kriteria pengujian, yaitu dengan *Area Under the Roc curve* (AUC*), Root Mean Squared Error* (RMSE), skor F sebagai nilai ketelitian dari pemilihan dan terakhir menggunakan teknik W-L (*Win-Loss*).

Eksperimen dilakukan dengan meningkatkan nilai λ. Hasil dari perhitungan ranking pada dataset dengan λ = 0.1 dan λ =2 dapat dilihat pada Gambar 2 dan Gambar 3. Sedangkan hasil dari semua eksperimen dapat dilihat pada Lampiran Tabel 2 [2] dan Tabel 3 [2].

IV. ANALISA DAN KESIMPULAN

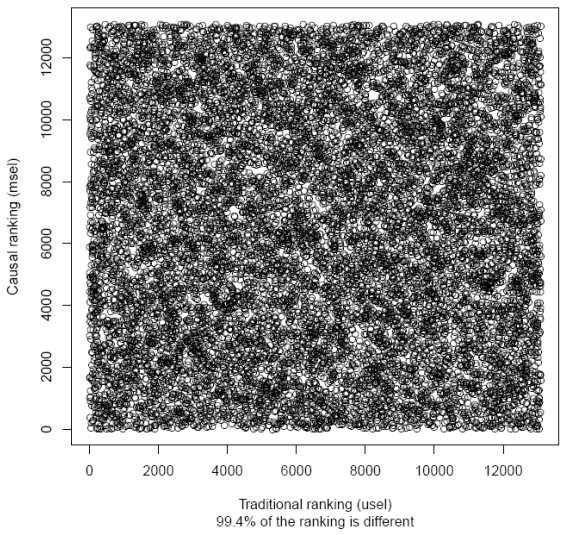
Dari hasil eksperimen dapat dilihat bahwa skoring mengalami peningkatan seiring dengan penambahan nilai λ. Hal ini membuktikan bahwa terdapat interaksi antara variabel-variabel input dengan variabel-variabel output yang diguanakan. Eksperimen membuktikan dengan menggunakan multiple input dan multiple output untuk perhitungan skor, akan mampu melihat hubungan sebab akibat antar variabel. Dengan adanya kemampuan untuk melihat sebab akibat dengan penggunaan mulitple input multiple output, maka pemilihan gen tidak hanya berdasarkan skoring antar variable saja. Hubungan antar variabel lain dengan jumlah yang sama akan terlihat dengan penggunaan algortimya MIMO ini.

Kemampuan melihat hubungan sebab akibat antar variabel akan mempermudah proses pradiagnosis pada bidang kedokteran. Diharapkan dengan mampu melihat hubungan antar variabel pada microarray dengan target-target yang telah ditentukan sesuai dengan ilmu kedokteran, akan mempercepat proses diagnosis. Dengan mempercepat proses diagnosis tentu akan mempercepat proses penyusunan terapi pengobatan

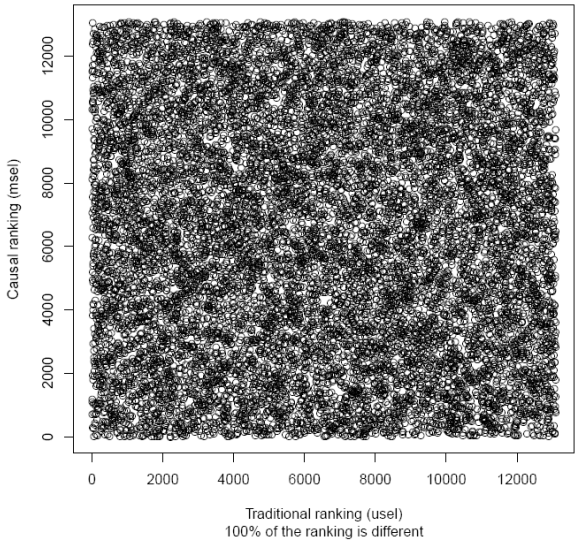
pada pasien kanker. Selain mampu melihat hubungan atau interaksi, algoritma ini terbukti mampu menangani keterbatasan dataset yang jumlahnya sedikit. Dengan memanfaatkan secara maksimal jumlah dataset yang sedikit ini, namun dapat tetap memlihat hubungan antara dataset melalui algoritma MIMO ini.

LAMPIRAN

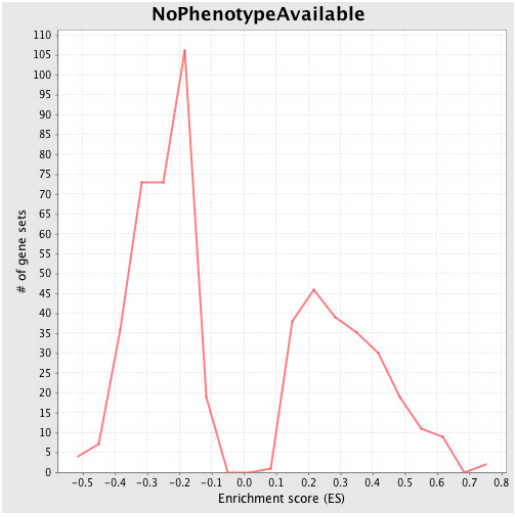
Lamda (λ) = 0.1



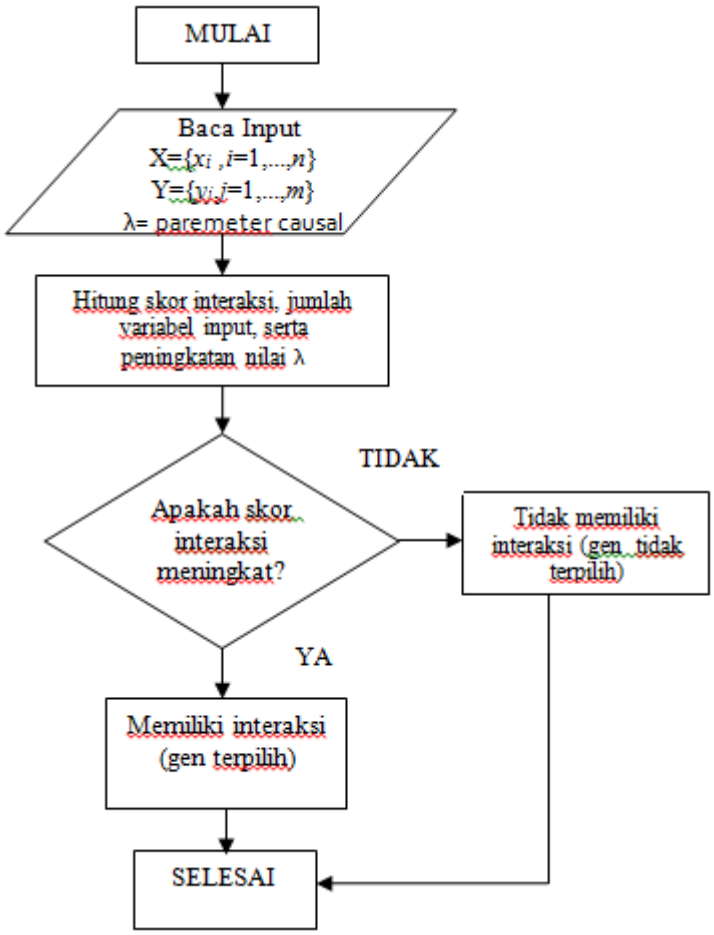
Lamda (λ) = 2



Gambar 2. kiri: Pengukuran ranking dari data set dengan lambda 0.1.
kanan: Pengukuran dari data set dengan lambda 2
Sumber code: http://compbio.dfci.harvard.edu/pubs/mimocausal



Gambar 3: Skor enrichment dari gen set tanpa phonotype.
Sumber gambar: http://compbio.dfci.harvard.edu/pubs/mimocausal



Gambar 1. Diagram flowchart algoritma MIMO gene selection

Tabel 2 Hasil eksperimen dengan *Holdout*

v = 20	**λ = 0**	**λ = 0.2**	**λ = 0.4**	**λ = 0.6**	**λ = 0.8**	**λ = 0.9**	**λ = 1**	**λ = 2**
AUC	0.688	0.688	0.694	0.699	0.703	0.704	0.705	0.707
1-RMSE	0.460	0.466	0.481	0.493	0.504	0.510	0.515	0.542
SAR	0.559	0.561	0.569	0.575	0.580	0.583	0.583	0.595
F	0.255	0.254	0.254	0.262	0.265	0.265	0.265	0.274
W-L		1-0	3-0	5-0	6-0	5-0	5-0	5-0
v =50	**λ = 0**	**λ = 0.2**	**λ = 0.4**	**λ = 0.6**	**λ = 0.8**	**λ = 0.9**	**λ = 1**	**λ = 2**
AUC	0.693	0.698	0.703	0.706	0.709	0.710	0.711	0.715
1-RMSE	0.451	0.458	0.465	0.471	0.477	0.479	0.482	0.503
SAR	0.552	0.556	0.556	0.567	0.571	0.572	0.574	0.583
F	0.263	0.265	0.265	0.270	0.272	0.271	0.273	0.277
W-L		2-0	2-0	3-0	2-0	2-0	3-0	4-0
v = 100	**λ = 0**	**λ = 0.2**	**λ = 0.4**	**λ = 0.6**	**λ = 0.8**	**λ = 0.9**	**λ = 1**	**λ = 2**
AUC	0.699	0.704	0.708	0.711	0.714	0.715	0.715	0.716
1-RMSE	0.454	0.457	0.459	0.463	0.467	0.472	0.472	0.487
SAR	0.545	0.549	0.553	0.557	0.561	0.563	0.564	0.573
F	0.272	0.271	0.272	0.274	0.274	0.274	0.275	0.284
W-L		1-0	1-0	1-0	2-0	3-0	4-1	4-1

Tabel 3 Hasil Eksperimen dengan *Leave one data set out*

v = 20	**λ = 0**	**λ = 0.2**	**λ = 0.4**	**λ = 0.6**	**λ = 0.8**	**λ = 0.9**	**λ = 1**	**λ = 2**
AUC	0.678	0.674	0.678	0.680	0.682	0.682	0.680	0.669
1-RMSE	0.447	0.448	0.467	0.469	0.482	0.528	0.544	0.556
SAR	0.553	0.552	0.560	0.561	0.566	0.582	0.586	0.586
F	0.280	0.275	0.275	0.281	0.279	0.283	0.287	0.276
W-L		1-1	5-1	2-0	4-0	5-0	4-0	4-0
v = 50	**λ = 0**	**λ = 0.2**	**λ = 0.4**	**λ = 0.6**	**λ = 0.8**	**λ = 0.9**	**λ = 1**	**λ = 2**
AUC	0.681	0.687	0.692	0.693	0.698	0.700	0.700	0.693
1-RMSE	0.428	0.438	0.453	0.457	0.464	0.473	0.490	0.516
SAR	0.542	0.551	0.559	0.561	0.565	0.569	0.576	0.582
F	0.284	0.284	0.281	0.281	0.285	0.291	0.298	0.303
W-L		3-0	4-0	5-1	3-0	5-0	4-0	6-0
v = 100	**λ = 0**	**λ = 0.2**	**λ = 0.4**	**λ = 0.6**	**λ = 0.8**	**λ = 0.9**	**λ = 1**	**λ = 2**
AUC	0.687	0.694	0.704	0.708	0.711	0.706	0.708	0.676
1-RMSE	0.430	0.436	0.449	0.457	0.463	0.463	0.476	0.477
SAR	0.537	0.545	0.556	0.562	0.586	0.565	0.571	0.561
F	0.290	0.292	0.294	0.296	0.299	0.294	0.304	0.288
W-L		1-0	4-0	6-0	4-0	4-0	5-0	5-1

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Bapak Ito Wasito,PhD selaku pengarah dalam menyusun makalah ini, juga kepada seluruh anggota kelas *Artificial Intelligence* Magister Ilmu Komputer Universitas Indonesia semester Ganjil 2011/2012 yang telah memberikan masukan serta saran selama pembuatan makalah ini.

Daftar Pustaka

[1] Y. Saeys, I. Inza, P. Larranaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, 23:2597-2517, 2007

[2] B. Gianluca, H.K. Benjamin, D. Christine, S. Christos, Q. John, "Multiple Input Output Causal Strategies for Gene Selection," BMC Bioinformatics, 12:458, 2011.

[3] B. Gianluca, H.K. Benjamin, D. Christine, S. Christos, Q. John, "Multiple Input Output Causal Strategies for Gene Selection," http://compbio.dfci.harvard.edu/pubs/mimocausal. (2011).