PENDEKATAN HIBRID PENAPIS DAN PEMBALUT MENGGUNAKAN ALGORITMA BAT SEARCH UNTUK MASALAH PEMILIHAN GEN

Chee Qi Jun
Prof. Dr. Salwani Abdullah

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Dalam kajian ekspresi gen microarray, tujuan utama pemilihan gen adalah memilih subset gen yang berbeza untuk membangunkan alat diagnostik untuk mengelas tisu kanser. Pemilihan gen bertujuan untuk memilih sebilangan kecil gen yang bersesuaian untuk mengelakkan risiko pengorbanan gen atau keturunan ketepatan pengelasan. Para penyelidik mendapati bahawa kajian mendapatkan set gen terkecil daripada set data microarray untuk diagnosis klinikal terbukti merupakan salah satu tugas yang paling sukar dalam pembelajaran mesin. Banyak penyelidik telah mengaji masalah ini dengan menggunakan pendekatan penapis atau pembalut. Banyak pendekatan meta-heuristik telah dicadangkan untuk masalah pemilihan gen kerana kecekapan mereka memperoleh hasil yang lebih baik dalam masa yang lebih pendek berbanding pendekatan lain. Terdapat dua kategori pendekatan meta-heuristik iaitu meta-heuristik berasaskan *Single-based* (Carian tempatan) dan meta-heuristik berasaskan populasi. Dalam kajian ini, algoritma carian kelawar (ACK) akan digunakan untuk mengkaji masalah pemilihan gen menggunakan pendekatan pembalut, pendekatan penapis menggunakan kolerasi atribut (KA) dan seterusnya menyelidik kesan penggabungan kolerasi atribut (KA) sebagai penapis dengan (ACK) sebagai pembalut (dikodkan sebagai KA-ACK) bagi mencapai ketepatan pengelasan yang lebih baik. Kajian ini terdiri daripada empat fasa ;pertamanya ialah pemahaman masalah dan kajian kesusateraan, prapemprosesan, pembinaan penyelesaian dan fasa penilaian. Objektif kajian ini adalah untuk menjalankan analisis ketepatan antara pendekatan penapis dan

pembalut dan juga penggabungan antaranya menggunakan algoritma pengelasan dalam pengelas iaitu Naïve Bayes.

1 PENGENALAN

Pemilihan gen untuk dataset microarray adalah tugas yang mencabar kerana data microarray dicirikan oleh kewujudan sejumlah besar gen (dalam ribuan) dan bilangan sampel yang terhad (ratusan atau kurang daripada seratus) yang tersedia untuk analisis; namun, kebanyakan gen ini adalah tidak relevan dan berlebihan. Fitur-fitur ini menimbulkan cabaran kepada semua algoritma pengelasan dan ini boleh memberi kesan buruk kepada ketepatan pengelasannya kerana proses pembelajaran (dalam pembelajaran mesin) adalah tugas berat yang dipengaruhi oleh 'kutukan dimensi' (Duval al.2009b). Oleh itu, pemilihan gen adalah proses teras yang perlu dilakukan sebelum sebarang tugas pengelasan sebagai langkah pra-pemprosesan untuk mengurangkan dimensi dataset ini dengan memilih set gen yang berguna dan menyingkir gen yang berlebihan dan tidak relevan (Talbi et al.2008). Dengan langkah ini, ia akan membantu meningkatkan ketepatan ramalan teknik pengelasan, memberikan pemahaman yang lebih baik apabila menganalisis data, dan mengurangkan kos pengiraan.

Berdasarkan ketergantungan pada algoritma pengelas, pendekatan pemilihan gen dapat dibahagikan kepada dua kategori: penapis (*filter*) dan pembalut (*wrapper*) (Kohavi dan John 1997b). Pendekatan penapis melaksanakan pemilihan gen secara bebas daripada algoritma pengelas, di mana mereka praproses dataset microarray sebelum dataset digunakan untuk analisis pengelasan. Dalam pendekatan pembalut, algoritma pengelas digunakan untuk menilai subset gen yang dihasilkan dalam setiap lelaran.

Dalam kajian ini, algoritma Bat Search asal yang dikaji oleh (Xin-She Yang 2010) digunakan untuk pendekatan pembalut untuk mengkaji pengurangan atribut. BSA adalah pendekatan meta-heuristik berasaskan pengoptimuman global yang digunakan dalam kajian ini. Pendekatan meta heuristik telah dicadangkan untuk masalah pemilihan gen kerana kecekapannya dalam mendapatkan penyelesaian yang lebih baik dalam masa yang munasabah. Selain itu, CA akan digunakan sebagai pendekatan penapis dalam cadangan kajian ini iaitu hibrid antara penapis dan pembalut untuk mengatasi kekurangan yang ditunjukkan oleh pendekatan pembalut.

2 PENYATAAN MASALAH

Walaupun pemilihan gen mempunyai banyak kelebihan, ia juga menghadapi risiko seperti ketepatan pengelasan dikurang atau kenaikan kebarangkalian *over-fitting* , yang disebabkan oleh gen dimensionaliti yang tinggi dalam dataset microarray dan bilangan sampel yang kecil. Oleh itu, pemilihan gen bertujuan untuk memilih sebilangan kecil gen yang sesuai untuk mengelakkan risiko ini dan meningkatkan ketepatan pengelasan.

Tujuan utama pemilihan gen adalah memilih subset gen yang berlainan untuk membangunkan alat diagnostik untuk mengelaskan tisu kanser. Pendekatan penapis memilih gen berdasarkan sifat-sifat asas data, dan bukannya menjadi berat sebelah terhadap pengelas tertentu. Penapis mencari gen yang berkaitan dan menghapuskan yang tidak relevan. Kelebihan pendekatan penapis adalah ia memerlukan masa pengiraan yang lebih kurang. Walau bagaimanapun, pendekatan penapis tidak berinteraksi dengan algoritma pengelasan. Manakala bagi pendekatan pembalut, algoritma pengelasan digunakan untuk menilai subset yang dihasilkan setiap kali.Tidak dinafikan bahawa keputusan yang diperoleh oleh pendekatan pembalut adalah lebih baik berbanding dengan keputusan yang diperolehi oleh pendekatan penapis,tetapi pendekatan pembalut memerlukan masa pengiraan yang lebih panjang kerana algoritma pengelas diperlukan untuk menilai setiap subset gen yang dihasilkan bagi setiap lelaran.

Oleh itu, apabila kedua-dua pendekatan ini digabungkan, pendekatan penapis dan pembalut dapat saling melengkapi untuk mendapatkan subset gen yang paling relevan dan mencapai ketepatan pengelasan yang lebih tinggi. Kelebihan hibridisasi ini mendorong kita untuk mengkaji pendekatan hibridisasi lain dalam usaha untuk menawarkan penyelesaian yang lebih baik untuk masalah pemilihan gen.

3 OBJEKTIF KA.JIAN

Sehubungan dengan itu,kajian ini bertujuan untuk mencapai objektif berikut:

i. Memahami konsep algoritma Bat Search yang berdasarkan pendekatan pembalut untuk masalah pemilihan gen.

ii. Mencadangkan pendekatan penggabungan antara BSA sebagai pendekatan pembalut dengan CA sebagai penapis untuk menambahbaik ketepatan pengelasan dan mengurangkan masa pengiraan. .

iii. Menjalankan analisis perbandingan penyampaian kaedah antara pendekatan penapis dan pembalut menggunakan BSA .

4 METOD KAJIAN

4.1 Set Data

Sembilan set data *microarray* telah dipilih untuk mengaji pendekatan yang dicadangkan. Set data ini boleh dimuat turun dari http://csse.szu.edu.cn/staff/zhuzx/Datasets.html. Ciri-ciri dataset ini telah ditunjukkan dalam bawah. Perbezaan besar antara bilangan gen dan bilangan sampel boleh dilihat antara set data .

Jadual 1.1 Penerangan Set data

Set data	Gen	Sampel	Kelas	Penerangan
ALL-AML	7129	72	2	Dua bentuk leukimia *acute*, i.e., *Acute Myelogenous Leukemia* (AML) dan *Acute Lymphoblastic Leukemia* (ALL)
ALL-AML-3C	7129	72	3	AML, *ALL B-cell*, dan ALL T-Cell
ALL-AML-4C	7129	72	4	*AML-Bone Marrow*, AML-*Peripheral Blood*, ALL B-cell, dan T-Cell
CNS	7129	60	2	Hasil rawatan untuk 60 pesakit kanser sistem saraf pusat (21 terselamat dan 39 gagal)
Lymphoma	4026	62	3	Tiga tumor limfoid dewasa yang paling lazim
MLL	12582	72	3	AML, ALL, dan *mixed-lineage leukemia* (MLL)
Breast	24481	97	2	97 sampel daripada pesakit kanser payudara (46 pesakit mengembangkan metastasis jarak jauh; 51 kekal sihat selepas diagnosis awal untuk sekurang-kurangnya lima tahun)

Ovarian	15154	253	2	Spektrum proteomik daripada 91 orang normal dan 162 pesakit kanser ovari
SRBCT	2308	83	4	Tumor sel kecil,biru bulat (SRBCT) dari zaman kecil

4.2 Fasa Perancangan

Fasa ini melibatkan proses pengenalpastian masalah, objektif, persoalan kajian dan menentukan skop. Kaedah yang diguna untuk masalah pemilihan gen dalam kajian ini adalah pembalut menggunakan BSA, dan pendekatan hibrid dengan penapis CA.

4.3 Fasa Analisis

Fasa ini menganalisis dan mengtafsir maklumat yang dikumpul dalam fasa perancangan . Analisis membentangkan kajian lepas berkaitan dengan masalah pemilihan gen yang telah dinilai; ditulis semula untuk tujuan memperoleh pandangan yang jelas latar belakang masalah, struktur penyelesaian masalah dan akhirnya menjalankan analisis perbandingan kaedah kajian yang ada.

4.3 Fasa Reka Bentuk

Fasa ini akan menggunakan pendekatan algoritma Bat Search untuk mengkaji masalah pemilihan gen. Fasa ini menggunakan WEKA untuk digunakan dengan algoritma BSA sebagai pembalut dan CA sebagai penapis dan juga penggabungan antara kedua-dua pendekatan ini dan dikaji atas tahap ketepatan pengelasan.

4.4 Fasa Penilaian

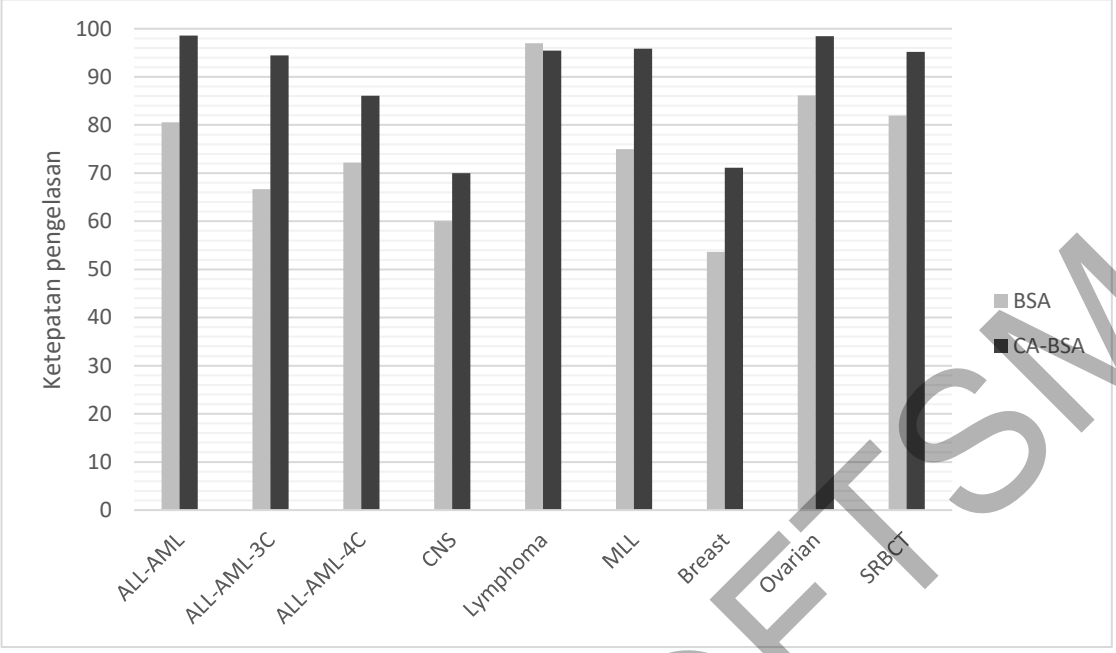
Fungsian utama fasa penilaian adalah untuk melakukan perbandingan antara penyampaian kaedah yang dikajikan dalam kajian ini untuk masalah pemilihan gen iaitu penapis dan pembalut. Prestasi setiap pendekatan diukur berdasarkan ketepatan pengelasan, bilangan gen minimum, dan masa pengiraan.

5 HASIL KA.JIAN

Bahagian ini membincang hasil daripada kajian menggunakan pendekatan pemilihan fitur iaitu pembalut dan hibrid penapis dan pembalut dalam set data gen. Hasil yang dilaporkan dalam bahagian ini diperoleh dengan menjalankan *cross-validation* 10-lipatan pada setiap set data, di mana set data dibahagikan kepada set latihan (90%) dan set ujian bebas (10%). Dengan kata lain, subset gen dipilih dari 90% daripada set latihan, dan ketepatannya diukur keatas 10% set ujian. Proses ini dilakukan 10 kali. Ketepatan dan bilangan gen terpilih yang dilaporkan dalam kajian ini berdasarkan *cross-validation* 10-lipatan berdasarkan data ujian.

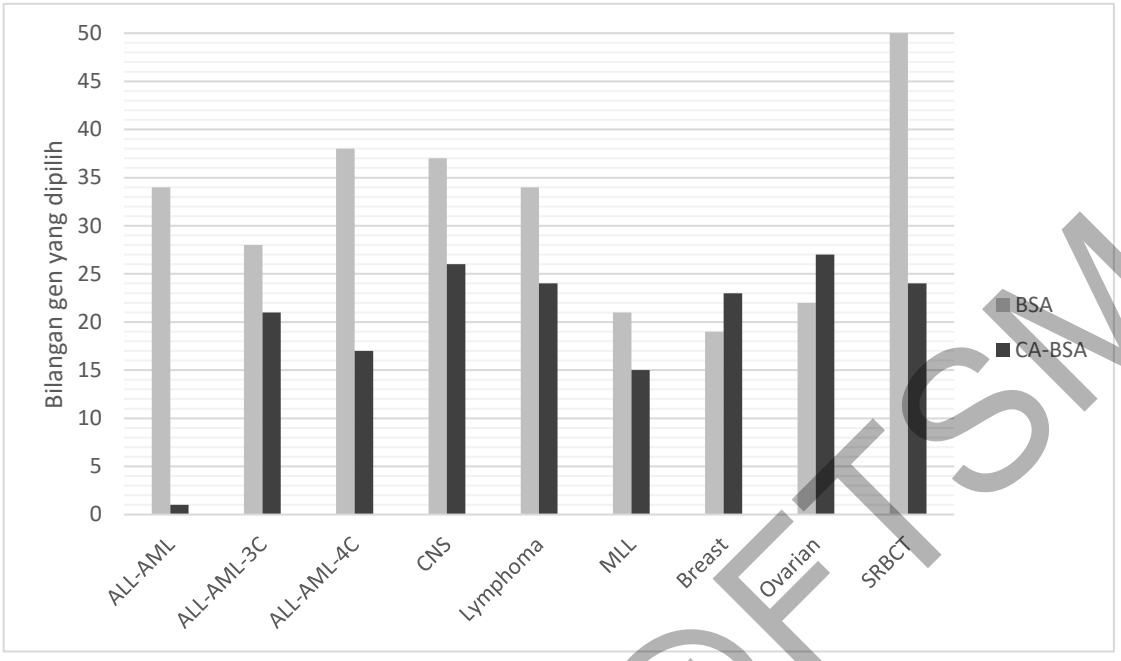
	Pembalut	Hibrid penapis dan pembalut
Set data	BSA	CA + BSA
ALL-AML	80.56	**98.61**
ALL-AML-3C	66.67	**94.44**
ALL-AML-4C	72.22	**86.11**
CNS	60.00	**70.00**
Lymphoma	**96.96**	95.45
MLL	75.00	**95.83**
Breast	53.61	**71.13**
Ovarian	86.17	**98.42**
SRBCT	81.93	**95.18**
Average	74.79	**89.46**

Jadual 1.2 Perbandingan nilai ketepatan pengelasan yang diperoleh oleh Algoritma Bat Search dan Hibrid Correlation Attribute dan Algoritma Bat Search



Rajah 1.1 Perbandingan nilai ketepatan pengelasan antara CA dengan CA-BSA

Keputusan yang ditunjukkan oleh pendekatan hibrid CA-BSA menpunyai nilai ketepatan pengelasan yang melebihi 70 peratus berbanding dengan pendekatan pembalut BSA. Dari segi bilangan minimal gen yang terpilih dalam subset, kita boleh lihat bahawa pendekatan CA-BSA dapat memilih kurang daripada 30 gen untuk kesemua sembilan set data microarray. Ini mencadangkan bahawa terdapat peningkatan dalam prestasi yang dicapai oleh CA-BSA bila berbanding dengan BSA dalam mencari subset gen yang kecil dengan ketepatan pengelasan yang tinggi. Hal ini disebabkan oleh penggunaan penapis CA untuk memilih gen yang paling berkesan (menurut penilaian CA) ketika menetap BPS bagi BSA. Keputusan kajian ini juga selaras dengan kajian kesusasteraan bagi menggabungkan penapis dengan pembalut dalam satu pendekatan untuk menyelesai masalah pemilihan gen, di mana ia dapat mendapat keputusan yang baik berbanding dengan pendekatan hanya mengguna pembalut sahaja. (Amaldi dan Kann 2013.).



Rajah 1.2 Perbandingan bilangan gen yang terpilih antara CA dengan CA-BSA

6 KESIMPULAN

Secara keseluruhannya, kajian ini dijalankan bertujuan untuk mengkaji masalah pemilihan gen menggunakan algoritma Bat Search dengan pendekatan pembalut dan juga hibridisasi dengan pendekatan penapis Correlation Attribute. Kajian ini juga dilaksanakan dengan menggunakan teknik perlombongan data yang mengandungi teknik pengelasan. Dalam teknik pengelasan, teknik Naïve Bayes digunakan dalam proses pengelasan data.

Pra-pemprosesan data adalah penting dalam menjalankan kajian ini kerana fasa ini merupakan fasa yang penting sebelum proses pengelasan dijalankan kerana fasa ini sangat berguna bagi memperolehi pengetahuan. Tujuan fasa ini dilakukan untuk memastikan set data difahami dan diguna untuk memilih subset set data microarray. Kajian ini juga merangkumi teknik pengelasan dengan menggunakan perisian Weka 3.8. Keputusan kajian ini memilih algoritma Naïve Bayes iaitu model yang paling sesuai untuk membina penyelesaian perbandingan antara pendekatan pembalut dan hibrid penapis dengan pembalut untuk membandingkan tahap ketepatan antara dua pendekatan .

7 RUJUKAN

Yang, 2011, X.-S. Yang, Bat algorithm for multi-objective optimisation, International Journal of Bio-Inspired Computation, 3 (2011), pp. 267-274

R. Agrawal, R. Bala, A hybrid approach for selection of relevant features for microarray datasets, International Journal of Computer and Information Science and Engineering 1 (2007) 196–202.

A. Duval, J.-K. Hao, J.C.H. Hernandez, A memetic algorithm for gene selection and molecular classification of cancer, in: Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation, 2009, pp. 201–208.

E.G. Talbi, L. Jourdan, J. Garcia-Nieto, E. Alba, Comparison of population based metaheuristics for feature selection: application to microarray data classification, in: International Conference on Computer Systems and Applications, AICCSA 2008, IEEE/ACS, 2008, pp. 45–52.

R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1997) 273–324.

Dueck, G. 1993. New optimization heuristics: The great deluge algorithm and the record-to-record travel. Journal of Computational Physics 104(1), pp. 86–92.

L.Y. Chuang, C.H. Yang, J.C. Li, A hybrid BPSO-CGA approach for gene selection and classification of microarray data, Journal of Computational Biology 19 (2011) 1–14.

Pawlak, Z.: Rough sets. Internat. J. Comput. Inform. Sci. 11(5), 341–356 (1982)

Lin, T. Y. 1997. Granular computing: From rough sets and neighborhood systems to information granulation and computing in words. European Congress on Intelligent Techniques and Soft Computing, pp. 1602–1606

L. Yu, H. Liu, Redundancy based feature selection for microarray data, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 737–742.

P. Bermejo, J.A. Gámez, J.M. Puerta, A GRASP algorithm for fast hybrid (filter–wrapper) feature subset selection in high-dimensional datasets, Pattern Recognition Letters 32 (2011) 701–711.

Z. Zhu, Y.S. Ong, M. Dash, Wrapper–filter feature selection algorithm using a memetic framework, IEEE Transactions on Systems, Man, and Cybernetics, Part B 37 (2007) 70–76.