UNIVERSITAS INDONESIA

METODE SELEKSI FITUR BERBASIS PERANKINGAN BOBOT SECARA MULTI STEP MENGGUNAKAN DEEP LEARNING UNTUK PENCARIAN BIOMARKER PADA DATA MICROARRAY

THESIS

MUKHLIS AMIEN
1406522102

FAKULTAS ILMU KOMPUTER
PROGRAM STUDI MAGISTER ILMU KOMPUTER
DEPOK
JUNI 2016

UNIVERSITAS INDONESIA

METODE SELEKSI FITUR BERBASIS PERANKINGAN BOBOT SECARA MULTI STEP MENGGUNAKAN DEEP LEARNING UNTUK PENCARIAN BIOMARKER PADA DATA MICROARRAY

THESIS
Diajukan sebagai salah satu syarat untuk memperoleh gelar
Master

MUKHLIS AMIEN
1406522102

FAKULTAS ILMU KOMPUTER
PROGRAM STUDI MAGISTER ILMU KOMPUTER
DEPOK
JUNI 2016

HALAMAN PERSETUJUAN

Judul	:	Metode Seleksi Fitur Berbasis Perankingan Bobot Secara Multi Step Menggunakan Deep Learning untuk Pencarian Biomarker pada Data Microarray
Nama	:	Mukhlis Amien
NPM	:	1406522102

Laporan Thesis ini telah diperiksa dan disetujui.

XX Januari 2016

Ito Wasito PhD.

Pembimbing Thesis

HALAMAN PERNYATAAN ORISINALITAS

Thesis ini adalah hasil karya saya sendiri,
dan semua sumber baik yang dikutip maupun dirujuk
telah saya nyatakan dengan benar.

Nama	:	**Mukhlis Amien**
NPM	:	**1406522102**
Tanda Tangan	:	
Tanggal	:	**XX Januari 2016**

HALAMAN PENGESAHAN

Thesis ini diajukan oleh	:	
Nama	:	Mukhlis Amien
NPM	:	1406522102
Program Studi	:	Magister Ilmu Komputer
Judul Thesis	:	Metode Seleksi Fitur Berbasis Perankingan Bobot Secara Multi Step Menggunakan Deep Learning untuk Pencarian Biomarker pada Data Microarray

Telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Master pada Program Studi Magister Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Indonesia.

DEWAN PENGUJI

Pembimbing	:	Ito Wasito PhD.	(	)
Penguji	:	Prof. XXX	(	)
Penguji	:	Prof. XXXX	(	)
Penguji	:	Prof. XXXXXX	(	)

@todo

Jangan lupa mengisi nama para penguji.

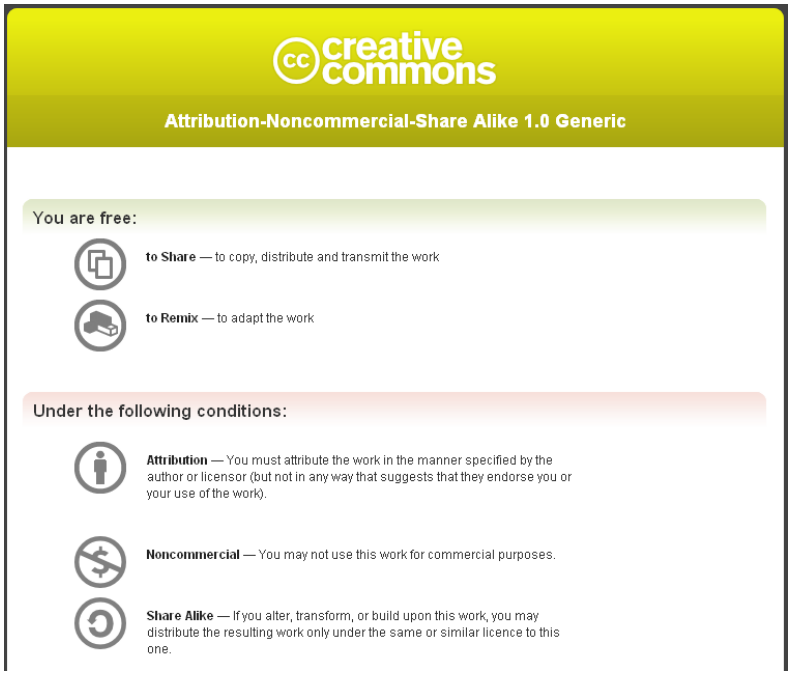
Ditetapkan di	:	Depok
Tanggal	:	XX Januari 2016

KATA PENGANTAR

Template ini disediakan untuk orang-orang yang berencana menggunakan LaTeX untuk membuat dokumen tugas akhirnya. Mengapa LaTeX? Ada banyak hal mengapa menggunakan LaTeX, diantaranya:

1. pertama
2. kedua
3. ketiga

1. LaTeX membuat kita jadi lebih fokus terhadap isi dokumen, bukan tampilan atau halaman.
2. LaTeX memudahkan dalam penulisan persamaan matematis.
3. Adanya automatis dalam penomoran caption, bab, subbab, subsubbab, referensi, dan rumus.
4. Adanya automatisasi dalam pembuatan daftar isi, daftar gambar, dan daftar tabel.
5. Adanya kemudahan dalam memberikan referensi dalam tulisan dengan menggunakan label. Cara ini dapat meminimalkan kesalahan pemberian referensi.



Gambar 1: *Creative Common License 1.0 Generic*

Gambar 1 diambil dari http://creativecommons.org/licenses/by-nc-sa/1.0/deed.en_CA. Jika ingin mengentahui lebih lengkap mengenai *Creative Common License 1.0 Generic*, silahkan buka http://creativecommons.org/licenses/by-nc-sa/1.0/legalcode. Seluruh dokumen yang dibuat dengan menggunakan template ini sepenuhnya menjadi hak milik pembuat dokumen dan bebas didistribusikan sesuai dengan keperluan masing-masing. Lisensi hanya berlaku jika ada orang yang membuat template baru dengan menggunakan template ini sebagai dasarnya.

Dokumen ini dibuat dengan LaTeX juga. Untuk meyakinkan Anda, coba lihat properti dari dokumen ini dan Anda akan menemukan bagian seperti Gambar 2. Dokumen ini dimaksudkan untuk memberikan gambaran kepada Anda seperti apa mudahnya menggunakan LaTeX dan juga memperlihatkan betapa bagus dokumen yang dihasilkan. Seluruh url yang Anda temukan dapat Anda klik. Seluruh referensi yang ada juga dapat diklik. Untuk mengerti template yang disediakan, Anda tetap harus membuka kode LaTeX dan bermain-main dengannya. Penjelasan dalam PDF ini masih bersifat gambaran dan tidak begitu mendetail, dapat dianggap sebagai pengantar singkat. Jika Anda merasa kesulitan dengan template ini, mungkin ada baiknya Anda belajar sedikit dasar-dasar LaTeX.

Gambar 2: Dokumen Dibuat dengan PDFLatex

Semoga template ini dapat membantu orang-orang yang ingin mencoba menggunakan LaTeX. Semoga template ini juga tidak berhenti disini dengan ada kontribusi dari para penggunanya. Kami juga ingin berterima kasih kepada Andreas Febrian, Lia Sadita, Fahrurrozi Rahman, Andre Tampubolon, dan Erik Dominikus atas kontribusinya dalam template ini.

Depok, 30 Desember 2009

Mukhlis Amien

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Indonesia, saya yang bertanda tangan di bawah ini:

Nama	:	Mukhlis Amien
NPM	:	1406522102
Program Studi	:	Magister Ilmu Komputer
Fakultas	:	Ilmu Komputer
Jenis Karya	:	Thesis

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Indonesia **Hak Bebas Royalti Noneksklusif (Non-exclusive Royalty Free Right)** atas karya ilmiah saya yang berjudul:

Metode Seleksi Fitur Berbasis Perankingan Bobot Secara Multi Step Menggunakan Deep Learning untuk Pencarian Biomarker pada Data Microarray

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Indonesia berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (database), merawat, dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyatan ini saya buat dengan sebenarnya.

Dibuat di : Depok
Pada tanggal : XX Januari 2016

Yang menyatakan

(Mukhlis Amien)

ABSTRAK

Nama	:	Mukhlis Amien
Program Studi	:	Magister Ilmu Komputer
Judul	:	Metode Seleksi Fitur Berbasis Perankingan Bobot Secara Multi Step Menggunakan Deep Learning untuk Pencarian Biomarker pada Data Microarray

Data ekspresi gen pada percobaan microarray memiliki ciri khas yaitu jumlah sampel yang sedikit dengan dimensi fitur yang sangat banyak. Algoritma *clustering* dan algoritma *deep learning* merupakan dua algoritma *unsupervised learning*, yang bisa membantu menganalisa data ekspresi gen. Algoritma pemilihan fitur digunakan untuk mendapatkan fitur gen yang paling penting. Dan kemudian akan digunakan algoritma *clustering* untuk mendapatkan struktur cluster dari data ekspresi gen. Pemilihan fitur yang paling informatif dari suatu kasus percobaan microarray, merupakan masalah yang ada pada pemrosesan data ekspresi gen. Sehingga diperlukan eksplorasi lebih lanjut untuk pemilihan fiturnya. Seleksi fitur gen, berdasarkan ranking bobot yang dihasilkan oleh *deep learning*, diharapkan dapat memecahkan masalah seleksi fitur tersebut. Sedangkan metode clustering yang dipakai adalah: Cluster Affinity Search Technique (CAST) , K-Means Clustering dan Hierarchical Clustering. Deep learning adalah metode pembelajaran mesin yang merupakan bagian dari algoritma neural network, metode deep learning yang sering dipakai adalah arsitektur deep believe network (DBN). Pendekatan unsupervised learning pada deep learning dan clustering, diharapkan dapat digunakan untuk membantu peneliti dalam menganalisa data ekspresi gen-nya.

Kata Kunci:
Microarray, ekspresi gen, Algoritma Clustering, feature selection, deep learning, unsupervised learning.

ABSTRACT

Name	:	Mukhlis Amien
Program	:	Magister Ilmu Komputer
Title	:	FEATURE SELECTION METHOD BASED ON MULTI STEP WEIGHT RANKING USING DEEP LEARNING TO SEARCH BIOMARKER IN LUNG CANCER MICROARRAY DATA SET

Microarray technology has made possible the profiling of gene expressions of the entire genome in a single hybridization experiment. Since microarray data acquire tens of thousands of gene expression values simultaneously. However, the number of sample usually small. Feature selection and clustering algorithm for microarray data analysis is useful to extract cluster structure and to reduce the high dimensional microarray data and reconstruct to lower dimensional with minimum error possible. Deep learning and clustering is a machine learning method. In this research we will investigate the effectiveness of clustering after or prior dimensionality reduction. The most common deep learning architecture used for dimensionality reduction is deep believe network (DBN) and stacked auto encoder (SAE). Pre training unsupervised learning and greedy layer wise training approach are expected for better dimensionality reduction in microarray datasets compared with other methods.

Keywords:
Microarray, ekspresi gen, Algoritma Clustering, feature selection, deep learning, unsupervised learning.

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Data ekspresi gen pada percobaan *microarray* memiliki ciri khas yaitu dimensi fitur gen yang jauh lebih besar dibandingkan dengan sampel pasien yang sedikit dikarenakan oleh mahalnya percobaan dan terbatasnya pasien. Hal ini menyebabkan masalah pada penerapan teknik machine learning untuk pengenalan pola penyakit yang diinginkan. Oleh karena itu, dalam menyederhanakan data ekspresi gen tersebut, dibutuhkan metode seleksi fitur untuk mempermudah melakukan analisa gen dengan menyeleksi gen-gen yang dibutuhkan saja (Yoon et al., 2006). Menurut penelitian Yoon et al. (2006) dan Bandyopadhyay et al. (2014) tidak semua gen yang didapatkan dalam percobaan microarray tersebut adalah gen yang informatif, bahkan jumlah ekspresi gen yang informatif untuk kasus yang diinginkan misalnya untuk pengenalan sel kanker, sangat sedikit dibandingkan dengan keseluruhan ekspresi gen yang didapatkan dalam sebuah eksperimen (Bandyopadhyay et al., 2014). Data ekspresi gen yang tidak informatif tersebut dapat mengurangi performa proses pengenalan pola secara signifikan pada teknik *machine learning* yang diterapkan. Akan tetapi, beberapa gen yang informatif berpengaruh secara signifikan terhadap pengenalan pola tersebut. Sebagai contoh, untuk mendiagnosa kanker paru-paru, hanya dibutuhkan sekitar 50 gen saja dari 22 ribu gen yang didapatkan dalam percobaan. Gen-gen yang paling informatif ini disebut dengan Biomarker (Belinsky, 2004). Sehingga, hanya dengan menggunakan data biomarker yang ditemukan, bisa dikenali penyakit yang diderita oleh pasien.

pengetahuan manusia tentang gen sampai saat ini masih terbatas, yaitu ada sekitar 26% dari keseluruhan gen yang belum diketahui kegunaannya (Häggström, 2014). Oleh karena itu pendekatan teknik machine learning secara unsupervised sering dilakukan untuk analisa pengenalan pola data microarray. Pada penelitian ini, akan dilakukan seleksi fitur terhadap data microarray secara unsupervised learning dengan menggunakan teknik deep learning. Dari hasil seleksi fitur gen tersebut akan diterapkan algoritma supervised learning yang digunakan untuk melakukan evaluasi seberapa baik keakurasian seleksi fitur tersebut dalam pengenalan pola pendeteksian penyakit kanker paru-paru pada sample pasien sakit dan normal.

Untuk teknik seleksi fitur tersebut akan digunakan metode de-

ngan cara melakukan modifikasi algoritma seleksi fitur untuk logistic regression dilakukan oleh Shevade and Keerthi (2003).

@todo

bagian bawah ini masih ruwet

Dikarenakan algoritma seleksi fitur menggunakan logistic regression merupakan bagian dari supervised learning dan linier, maka dianggap kurang cocok untuk data microarray yang fungsinya belum diketahui secara lengkap dan memiliki karakteristik yang kompleks. Dan logistic regression memiliki masalah dalam mengeliminasi fitur, dikarenakan koefisien bobot ditempatkan disetiap fitur. Oleh karena itu, disini akan diajukan arsitektur deep learning. Arsitektur deep learning, yang akan digunakan pada penelitian ini adalah arsitektur Deep Belief Network (DBN). DBN merupakan jaringan Restrictive Boltzmann Machine (RBM) yang dijajarkan. Dimulai dengan memberikan bobot random diantara dua network, yang dapat di latih dengan cara meminimalkan perbedaan antara data asli dengan data rekonstruksinya. Gradien didapatkan dengan chain rule untuk melakukan penurunan error dengan teknik Contrastive Divergence (CD). Untuk dicari bobot (W) dengan maximum likelihood learning secara greedy per layer-nya (greedy layer wise training) (Hinton, 2006). Pada penelitian ini, untuk mencari perangkingan bobotnya, menggunakan modifikasi dari cara yang digunakan oleh Shevade (Shevade, 2003) dalam teknik seleksi fitur berbasis weight menggunakan Sparse logistic regression (Shevade, 2003). Sehingga teori perankingan weight ini akan dimodifikasi dan digunakan untuk meranking fiturnya secara multi step yang akan diterapkan pada DBN. Tahap selanjutnya fitur yang telah didapatkan pada tahap seleksi fitur, akan digunakan sebagai data untuk penerapan clustering. Algoritma yang akan dipakai adalah : Cluster Affinity Search Technique (CAST) , K-Means Clustering dan Hierarchical Clustering (yeung, 2001). Untuk mengevaluasi dan menganalisa seberapa baik hasil dari percobaan ini, dilakukan dengan menghitung Adjusted Rand Index (ARI) (Hubert, 1985). ARI ini digunakan untuk mengukur mutu cluster dari hasil clustering. ARI menghitung derajat kesesuaian antara dua partisi, yaitu menghitung cluster yang dihasilkan, dibandingkan dengan kriteria eksternal. Nilai ARI berada di antara 0 dan 1. Jika mutu cluster yang dihasilkan memiliki keterpisahan yang baik dibandingkan dengan kriteria luar cluster, maka nilai ARI mendekati 1. Jika sebaliknya, nilai ARI mendekati 0. Untuk mengetahui gen yang dipilih tersebut informatif dan tidak, dilakukan literatur review.

1.2 Perumusan Masalah

Dikarenakan karakteristik sedikitnya sampel dan besarnya fitur. Serta tidak lengkapnya informasi kita terhadap gen. Apakah pendekatan unsupervised pada deep learning untuk mencari biomarker dengan perankingan bobot secara multi step cocok dipakai pada data microarray?

$\frac{1}{|\mathcal{D}|}\mathcal{L}(\theta = \{W, b\}, \mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{i=0}^{|\mathcal{D}|} \log(P(Y = y^{(i)}|x^{(i)}, W, b))$
$\ell(\theta = \{W, b\}, \mathcal{D})$

1.2.1 Definisi Permasalahan

@todo

Tuliskan permasalahan yang ingin diselesaikan. Bisa juga berbentuk pertanyaan

1.2.2 Batasan Permasalahan

- Dataset microarray
- Data yang digunakan adalah dataset yang sudah dilakukan preprocessing standar dan sudah dinormalisasi.

1.3 Tujuan

Penelitian ini bertujuan untuk membangun metodologi dalam pencarian Biomarker Gen yang paling penting untuk percobaan microarray. Dengan menghitung bobot ranking gen secara multi step.

1.4 Posisi Penelitian

@todo

Posisi penelitian Anda jika dilihat secara bersamaan dengan peneliti-peneliti lainnya. Akan lebih baik lagi jika ikut menyertakan diagram yang menjelaskan hubungan dan keterkaitan antar penelitian-penelitian sebelumnya

1.5 Manfaat Penelitian

Mendapatkan framework cara perankingan data ekspresi gen menggunakan arsitektur deep learning. Sehingga membantu dalam pencarian biomarker pada penelitian data microarray.

1.6 Sistematika Penulisan

Sistematika penulisan laporan adalah sebagai berikut:

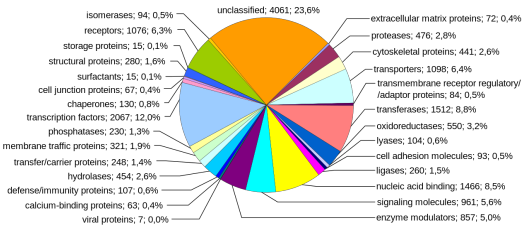
- Bab 1 PENDAHULUAN
  Berisi gambaran permasalahan dan metodologi apa yang akan diterapakan
- Bab 2 LANDASAN TEORI
  Landasan teori dipakainya metodologi yang akan diterapkan dalam eksperimen
- Bab 3 METODOLOGI PENELITIAN
  Penjelasan detail metodologi yang akan diterapkan dalam penelitian
- Bab 4 HASIL PENELITIAN DAN PEMBAHASAN
  Pembahasan hasil dari eksperimen
- Bab 5 KESIMPULAN DAN SARAN

@todo
Tambahkan penjelasan singkat mengenai isi masing-masing bab.

BAB 2
LANDASAN TEORI

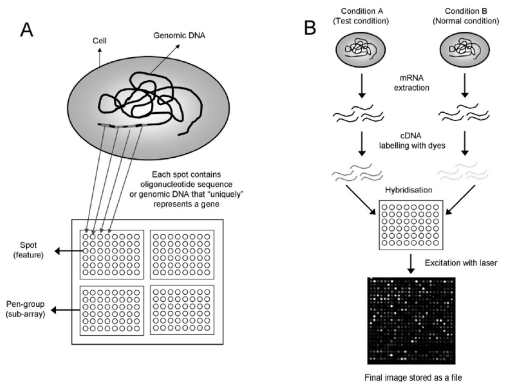
2.1 Ekspresi Gen

Percobaan microarray, mengukur tingkat aktivitas gen di dalam sebuah jaringan sel. Sehingga dapat memberikan informasi berdasarkan aktivitas di dalam jaringan yang bersangkutan. Data ini didapatkan dengan cara mengukur banyaknya mRNA yang diproduksi pada saat proses transkripsi DNA, dimana dapat diukur seberapa aktif atau seberapa berfungsinya gen tersebut dalam sebuah jaringan (Elloumi and Zomaya, 2011). Karena kanker berhubungan dengan berbagai macam aktivitas penyimpangan regulasi pada sel, maka data ekspresi gen pada kanker merefleksikan penyimpangan regulasi tersebut. Untuk menangkap keabnormalan ini, percbaan microarray, dimana dapat mengukur secara simultan dari level ekspresi ratusan bahkan ribuan ekspresi gen dapat digunakan untuk mengidentifikasi kanker. Percobaan microarray sering dipakai untuk membandingkan profil ekspresi gen pada sel yang terkena kanker, dibandingkan dengan sel yang normal pada berbagai macam percobaan. Percobaan microarray digunakan untuk mengidentifikasi ekspresi yang berbeda pada dua percobaan, yang biasanya berupa data tes dan data kontrol.



Gambar 2.1: Ada 23,6% dari keseluruhan fungsi gen yang belum diketahui, sehingga pengetahuan tentang fungsi gen masih belum lengkap. (Häggström, 2014)

Data ekspresi gen yang masih mentah didapatkan dari percobaan di laboratorium menggunakan alat yang dinamakan dengan alat Genchip microarray. Data tersebut kemudian dilakukan pemrosesan awal untuk mendapatkan sebuah matriks ekspresi gen. Matriks ini memiliki data kolom dan baris, dimana kolom berisi data eksperimen, dan baris berisi nilai ekspresi pada tiap-tiap gen (gambar 2.2) [Babu, 2004].



Gambar 2.2: Proses Keseluruhan Percobaan Microarray.(Babu, 2004)

@todo
tambahkan dua gambar yang digambar sendiri

Karena data microarray yang didapatkan dapat mencapai ribuan ekspresi dalam satu waktu secara simultan, maka data ini dapat sangat membantu dalam mengidentifikasi penyakit. Akan tetapi, hasil yang didapat dengan menganalisa beberapa data microarray yang dilakukan oleh dua percobaan yang berbeda tetapi dengan tujuan yang sama, dapat menghasilkan hasil yang sangat berbeda. Salah satu alasannya adalah terbatasnya sampel dan terlalu banyaknya profil ekspresi gen. Sehingga diperlukan metode testing statistik untuk memastikan bahwa data microarray tersebut memiliki tingkat signifikansi yang cukup, dan dipastikan bahwa perbedaan tersebut memang karena eksperimen, bukan karena kerusakan alat atau kesalahan prosedur eksperimen.

2.2 Pemrosesan Data Microarray

Data yang dihasilkan dari alat microarray ini berupa citra yang perlu diproses lebih lanjut. Sebelum data ekspresi gen dapat dianalisa lebih lanjut, perlu dilakukan pemrosesan awal yang berupa (i) perbaikan background, (ii) normalisasi data dan kemudian (iii) penyaringan data.

1. **Perbaikan Background**

   Perbiakan background ini ditujukan untuk menghilangkan titik-titik noise yang tidak berasal dari proses hibridisasi. Metode untuk perbaikan background ini banyak diajukan dalam penelitian. [15]

2. **Normalisasi**

   Tujuan dari normalisasi adalah untuk mengatur bias yang dihasilkan oleh

variasi proses percobaan microarray. Metode normalisasi data microarray ada banyak, dan pada penelitian ini akan digunakan normalisasi standar untuk data microarray.

3. **Penyaringan data** Tidak semua data yang didapat dari percobaan microarray bagus, kadangkala terjadi kesalahan alat dan noise yang diakibatkan oleh alat, oleh karena itu perlu disaring, mana data yang disebabkan oleh proses biologi, dan mana yang disebabkan oleh noise alat.

4. **Missing Value Imputation**
   Tidak semua data ekspresi gen dapat kita dapatkan, dikarenakan rumitnya percobaan microarray, kadangkala data tidak kita dapatkan, oleh sebab itu diperlukan metode untuk melakukan pendekatan statistic dalam memberikan perkiraan isi data dalam titik data yang hilang tersebut.

5. **Seleksi Fitur**
   Setelah proses diatas, diperlukan teknik untuk menseleksi fitur pada data microarray. Ada banyak metode yang sudah diusulkan oleh para peneliti. Seperti pada table 1 dibawah. Dan pada titik inilah penelitian ini dijalankan. Diharapkan penelitian ini menghasilkan metode reduksi dimensi untuk data microarray.

2.3 Ekstraksi Fitur dan Seleksi Fitur Pada Penelitian Sebelumnya

@todo
tulis tabel perbandingan seleksi fitur

2.4 Deep Learning

@todo
tuliskan secara singkat tentang deep learning

2.5 Restricted Boltzmann Machine

@todo

tuliskan secara detail tentang rbm

2.6 Deep Believe Network

@todo

tuliskan secara detail tentang dbn

2.7 Maximum Likelihood Estimation

@todo

tuliskan secara detail tentang MLE

$$\mathcal{L}(\theta\,|\,x_1,\ldots,x_n) = f(x_1,x_2,\ldots,x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Thus, the extrema of $\mathcal{L}$ are equivalent to the extrema of $\log \mathcal{L}$:

$$\log \mathcal{L}(\theta\,|\,x_1,\ldots,x_n) = \sum_{i=1}^{n} \log f(x_i|\theta) \tag{2.1}$$

From which the maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$ is defined as:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \sum_{i=1}^{n} \log f(x_i|\theta) \tag{2.2}$$

As an aside, Bayesians will remind us we can generalized into a MAP estimator, given uniform prior $g(\theta)$:

$$\arg\max_{\theta} \sum_{i=1}^{n} \log f(x_i|\theta) = \arg\max_{\theta} \log(f|\theta) = \arg\max_{\theta} \log(f|\theta)g(\theta) = \hat{\theta}_{\text{MAP}} \tag{2.3}$$

From which optimization and real analysis reminds us of the following equivalence, for all x:

$$\arg\max_{x}(x) = \arg\min_{x}(-x) \tag{2.4}$$

Thus, the following are equivalent:

$$\arg\max_{\theta} \sum_{i=1}^{n} \log f(x_i|\theta) = \arg\min_{\theta} -\sum_{i=1}^{n} \log f(x_i|\theta) = \hat{\theta}_{\text{MLE}} \tag{2.5}$$

2.8 Multi Layer Perceptron

@todo
tuliskan secara detail tentang dbn

2.9 Logistic Regression

@todo
tuliskan secara detail tentang dbn

BAB 3
METODOLOGI PENELITIAN

@todo

tambahkan kata-kata pengantar bab 3 disini

Penelitian ini dibagi menjadi tiga tahap: (1) Mendapatkan data microarray dan pengolahan awal; (2) Perancangan algoritma; (3) Testing dan kemudian dilanjutkan dengan evaluasi.

3.1 Gambaran Umum Penelitian

@todo

bikin dan tambahkan overview penelitian terbaru

Gambar overview penelitian

3.2 Pengumpulan Data dan Pengolahan Awal

Data microarray tersedia secara bebas di geo [http://www.ncbi.nlm.nih.gov/geo/], dan dapat diunduh, untuk digunakan sebagai data penelitian. Kemudian dilakukan normalisasi standar yang sering di pakai pada data microarray, proses normalisasi ada banyak metode, dan akan digunakan satu metode standar untuk pengolahan awal microarray agar mendapatkan data konsisten dan dapat dibandingkan. Proses pengolahan awal dan normalisasi digunakan tools standar dan tersedia bebas yaitu R-Bioconductor.

@todo

tambahkan bagan pengumpulan data dan pengolahan awal

3.3 Perancangan Algoritma

@todo

bawah masih amburadul tambahkan bagan2 algoritma pada presentasi

Algoritma deep learning yang dipakai adalah DBN dengan teknik ranking multi

step bobot adalah modifikasi dari algoritma seleksi fitur Sevade (Shevade, 2006) Ada tiga algoritma clustering yang diusulkan untuk diteliti, yaitu : (1) Cluster Affinity Search Technique (CAST) ; (2) K-Means Clustering ; (3) Hierarchical Clustering. Metode evaluasi yang dipakai adalah : Adjusted Rand Index yang digunakan untuk mengevaluasi baik tidaknya hasil clustering. 3.4 Melakukan Testing Arsitektur Deep Learning Hasil dari unsupervised learning yang dilakukan oleh deep learning, harus diuji dahulu dengan dengan data testing, apakah error rekonstruksinya masih baik atau rekonstruksi tersebut lebih jelek. Setelah dilakukan perankingan biomarker, diperlukan pengujian apakah apakah seleksi fitur tersebut menggambarkan hasil yang diinginkan, dengan membandingkan biomarker yang dihasilkan dengan literature.

3.4 Evaluasi Hasil Kombinasi Seleksi Fitur Secara Unsupervised dan Klasifikasi MLP Secara Supervised

@todo
pikirkan lagi judul section

BAB 4
HASIL PENELITIAN DAN PEMBAHASAN

@todo
tambahkan kata-kata pengantar bab 4

4.1 Pengolahan Awal

@todo
buat brainstorming untuk hasil eksperimen dan pembahasan

BAB 5
KESIMPULAN DAN SARAN

@todo
Tambahkan kesimpulan dan saran terkait dengan perkerjaan yang dilakukan.

5.1 Kesimpulan

5.2 Saran

DAFTAR REFERENSI

M Mwanadan Babu. Introduction to microarray data analysis. *Computational genomics: Theory and application*, pages 225–249, 2004.

Supriyo Bandyopadhyay, Saurav Mallik, and Amit Mukhopadhyay. A survey and comparative study of statistical tests for identifying differential expression from microarray data. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 11(1):95–115, 2014.

Steven A Belinsky. Gene-promoter hypermethylation as a biomarker in lung cancer. *Nature Reviews Cancer*, 4(9):707–717, 2004.

Mourad Elloumi and Albert Y Zomaya. *Algorithms in computational molecular biology: techniques, approaches and applications*, volume 21. John Wiley & Sons, 2011.

Mikael Häggström. Diagram of the pathways of human steroidogenesis. *Medicine*, 1:1, 2014.

Shirish Krishnaj Shevade and S Sathiya Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.

Youngmi Yoon, Jongchan Lee, and Sanghyun Park. Building a classifier for integrated microarray datasets through two-stage approach. In *BioInformatics and BioEngineering, 2006. BIBE 2006. Sixth IEEE Symposium on*, pages 94–102. IEEE, 2006.

LAMPIRAN

LAMPIRAN 1

@todo
Membuat todolist apa yang akan dikerjakan untuk thesis 1. belajar copy paste code 2. belajar membuat git 3. belajar buat bagan 4. belajar pseudo code 5. kumpulan dalam sebuah totorial dan link dengan cepat secara offline jika diperlukan 6. belajar machine learning nando 7. xxx