

**UNIVERSITAS INDONESIA**

**JUDUL**

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

**PROPOSAL TESIS**

**MUKHLIS AMIEN**

**1406522102**

**FAKULTAS ILMU KOMPUTER**

**PROGRAM MAGISTER ILMU KOMPUTER**

**UNIVERSITAS INDONESIA**

**DEPOK**

**DESEMBER 2015**

ABSTRAK

Nama : Mukhlis Amien

Program Studi : Magister Ilmu Komputer

Judul : METODE SELEKSI FITUR BERBASIS PERANKINGAN BOBOT SECARA MULTI STEP MENGUNAKAN 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

Study Program : Magister Ilmu Komputer

Judul : MULTI STEP RANK BASED WEIGHT DEEP LEARNING METHOD TO SEARCH BIOMARKER IN MICROARRAY DATA

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.

**Key Words :** *Microarray*, *ekspresi gen*, *Algoritma Clustering*, *feature selection*, *deep learning*, *unsupervised learning.*

DAFTAR ISI

[ABSTRAK ii](#_Toc440349268)

[ABSTRACT iii](#_Toc440349269)

[DAFTAR ISI iv](#_Toc440349270)

[BAB 1 1](#_Toc440349271)

[PENDAHULUAN 1](#_Toc440349272)

[1 1](#_Toc440349273)

[1.1 Latar Belakang 1](#_Toc440349274)

[1.2 Perumusan Masalah 2](#_Toc440349275)

[1.3 Tujuan Penelitian 3](#_Toc440349276)

[1.4 Manfaat Penelitian 3](#_Toc440349277)

[1.5 Batasan Penelitian 3](#_Toc440349278)

[BAB 2 4](#_Toc440349279)

[LANDASAN TEORI 4](#_Toc440349280)

[2 4](#_Toc440349281)

[2.1 Ekspresi Gen 4](#_Toc440349282)

[2.2 Pemrosesan Awal dan Normalisasi 7](#_Toc440349283)

[2.3 Algoritma Clustering 8](#_Toc440349284)

[2.4 Reduksi Dimensi dan Ekstraksi Fitur 9](#_Toc440349285)

[2.5 Reduksi Dimensi dengan Pendekatan Deep Learning 10](#_Toc440349286)

[2.6 Evaluasi Hasil Clustering dengan Adjusted Rand Index 12](#_Toc440349287)

[BAB 3 14](#_Toc440349288)

[METODOLOGI PENELITIAN 14](#_Toc440349289)

[3 14](#_Toc440349290)

[3.1 Pengumpulan Data dan Pengolahan Awal 14](#_Toc440349291)

[3.2 Penelitian Penempatan Metode Reduksi Dimensi dan Clustering **Error! Bookmark not defined.**](#_Toc440349292)

[3.3 Perancangan Algoritma Clustering **Error! Bookmark not defined.**](#_Toc440349293)

[3.4 Perancangan Algoritma Reduksi Dimensi **Error! Bookmark not defined.**](#_Toc440349294)

[3.5 Melakukan Testing dan Visualisasi Reduksi Dimensi 17](#_Toc440349295)

[3.6 Evaluasi Hasil Kombinasi Clustering dan Reduksi Dimensi 17](#_Toc440349296)

BAB 1

PENDAHULUAN

# 

## Latar Belakang

Data ekspresi gen pada percobaan *microarray* memiliki ciri khas yaitu dimensi fitur gen yang jauh lebih besar, dibandingkan dengan sampel. Oleh karena itu metode seleksi fitur dan *clustering* dibutuhkan untuk mempermudah melakukan analisa gen yang informatif (Yoon, 2006). Menurut penelitian (Yoon, 2006) dan (Bandyopadhyay, 2014) tidak semua gen yang didapat dari eksperimen *microarray* tersebut adalah gen yang informatif. Bahkan, jumlah ekspresi gen yang informatif untuk kasus yang diinginkan sangat sedikit dibandingkan dengan keseluruhan ekspresi gen yang didapat dalam eksperimen (Bandyopadhyay, 2014). Ekspresi gen yang tidak informatif dapat mengurangi performa dalam proses pengenalan pola seperti *supervised* dan *unsupervised learning*. Akan tetapi, beberapa gen yang informatif berpengaruh secara signifikan terhadap pengenalan pola tersebut. Sebagai contoh, untuk mendiagnosa kanker paru-paru, hanya dibutuhkan 50 gen saja. Gen yang paling informatif ini disebut dengan *Biomarker* (Bing, 2006).

Pendekatan *unsupervised learning* sering dilakukan untuk analisa pengenalan pola data *microarray*, dikarenakan pengetahuan tentang gen yang masih terbatas, yaitu ada sekitar 26% fungsi gen yang belum diketahui. (Häggström, 2014). Oleh karena itu pendekatan *unsupervised* yang akan dipilih pada penelitian ini.

Pada penelitian ini, akan dilakukan seleksi fitur terhadap data *microarray*. Dari hasil seleksi fitur gen tersebut akan diterapkan algoritma *clustering*. *Cluster* yang dihasilkan berupa kumpulan gen yang paling informatif berdasarkan rankingnya. Untuk seleksi fitur tersebut menggunakan pendekatan arsitektur deep learning dengan modifikasi algoritma seleksi fitur untuk *logistic regression* (Shevade, 2003). 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.

## Perumusan Masalah

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

## Tujuan Penelitian

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

## Manfaat Penelitian

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

## Batasan Penelitian

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

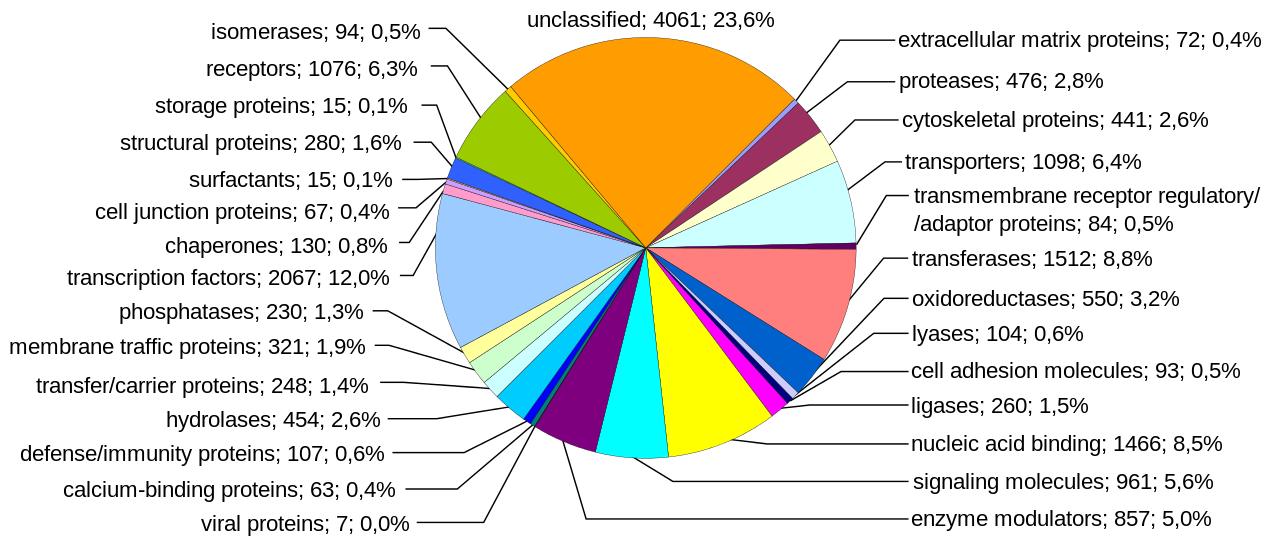
BAB 2

LANDASAN TEORI

# 

## 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, 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, data 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 mengidentifikasikan 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)



Gambar 2.2. Proses Keseluruhan Percobaan Microarray. [Babu, 2004]

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 2.3: Perbandingan Ekspresi gen yang relevan dan informatif dibandingkan dengan gen yang tidak relevan

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.

## Pemrosesan Awal dan Normalisasi

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]

1. 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.

1. 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.

1. *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.

1. 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.

1. Clustering

Proses clustering adalah proses pengelompokan data berdasarkan kesamaan sifat. Hal ini digunakan untuk membantu peneliti untuk mendapatkan gambaran struktur dari data yang diolah. Ada banyak metode clustering, pada penelitian ini akan digunakan tiga metode clustering yang paling sering dipakai untuk menganalisa data microarray, yaitu : *Cluster Affinity Search Technique (CAST), K-Means Clustering* dan *Hierarchical Clustering.*

## Algoritma Clustering

**A. *Cluster Affinity Search Technique (CAST)***

CAST adalah algoritma yang di ajukan oleh Ben-dor (Ben-dor,1999) untuk melakukan clustering data ekspresi gen. Input dari algoritma ini merupakan pasangan kesamaan dari gen dan parameter *cutoff* t (yang merupakan nilai real antar 0 sampai 1). Cluster dikonstruksi satu-satu. Cluster yang sedang direkonstruksi dinamakan dengan **.** Affinity dari gen g, yaitu a(g) didifinisikan sebagai jumlah dari nilai *similiarity* antara g dan semua gen didalam **.** Sebuah gen g dikatakan memiliki affinity yang tinggi jika **.** Jika sebaliknya, maka gen dinamakan memiliki affinity yang rendah.

**B. Hierarchical Clustering (Average-Link)**

Metode hierarchical clustering dibangun secara *bottom-up*. Awalnya, setiap obyek adalah memiliki cluster dirinya sendiri. Pada setiap langkah, jika ada dua cluster dengan kemiripan yang tinggi, akan disatukan. Proses ini terus dilakukan sampai didapatkan *k-cluster* yang diinginkan. Pada average-link, kriteria kesamaan cluster adalah merupakan rata-rata kesamaan pasangan antar obyek pada dua cluster. (jain, 1988), (Anderberg, 1973)

**C. K- Means Clustering**

Jumlah cluster k merupakan input dari algoritma k-means clustering. Cluster dideskripsikan oleh sebuah centroid, dimana merupakan titik tengah cluster. Implementasi algoritma k-means pada penelitian ini, digunakan centroid awal yang merupakan hasil dari average-link. Setiap obyek di berikan nilai centroid (yang merupakan anggota cluster dengan centroid tersebut) dengan jarak Eucledian yang terdekat. Kemudian, centroid baru dihitung setelah semua obyek mendapatkan anggota cluster. Langkah pemberian anggota obyek terhadap centroidnya berhenti setelah tidak ada obyek yang berpindah cluster.

## Ekstraksi Fitur

Teknik ekstraksi dan seleksi fitur pada data microarray sudah sering dipakai pada penelitian sebelumnya. Tabel 1 dibawah merupakan perbandingan metode yang sudah pernah dipakai untuk tujuan reduksi dimensi :

|  |  |  |  |
| --- | --- | --- | --- |
| **Pengarang** | **Judul Paper** | **Metode** | **Dataset** |
| C. Aliferis et al. 2003 | *Machine learning models for classication of lung cancer and selection of genomic markers using array gene expression data.* | Reduksi fitur secara rekursif dan melakukan filter secara asosiasi univariate | Lung Cancer Microarray |
| Ramaswamy, S. et al. 2001 | *Multiclass cancer diagnosis using tumor gene expression signatures.* | Pengurangan fitur secara rekursif dengan mengguanakan SVM | Various Microarray |
| Wang et al., 2005 | *Gene-expression proles to predict distant metastasis of lymph-node-negative primary breast cancer.* | Mengkombinasikan seleksi fitur yang berbasis korelasi dengan pendekatan assosiasi. | Various Microarray |
| Sharma et. Al, 2012 | *Combining multiple approaches for gene microarray classification.* | Mengkombinasikan banyak pendekatan ekstraksi fitur | Various Microarray |

## Deep Learning

**A. Deep Belief Network (hinton,2006)**

Data yang memiliki dimensi tinggi, dapat dikonversi menjadi suatu kode yang berdimensi rendah dengan cara melakukan pelatihan neural network yang memiliki banyak layer dengan layer pusat yang kecil untuk melakukan rekonstruksi dimensi tinggi pada vector input. Metode gradient descent yang digunakan pada back propagation digunakan untuk melakukan fine-tune. Dan metode ini bekerja dengan baik jika bobot awal dekat dengan solusi yang baik. Hinton[11] menunjukkan bahwa metode Deep RBM ini memiliki performa yang lebih baik untuk melakukan reduksi dimensi dibandingkan dengan PCA seperti yang telah dilakukan pada penelitian.

Menurut Hinton, auto encoder adalah metode generalisasi dari PCA yang menggunakan “encoder” yang adaptif untuk mentransformasikan data berdimensi tinggi ke dimensi rendah, dan kemudian digunakan “decoder” untuk mengubah kembali data dimensi rendah tersebut ke dimensi asli dengan sesedikit mungkin terjadi kehilangan data.

Dimulai dengan memberikan bobot random diantara dua network, yang dapat di latih dengan cara meminimalkan perbedaan antara data asli dengan data rekonstruksi nya. Gradien dapat didapatkan dengan *chain rule* untuk melakukan penurunan error back propagation melalui network decoder dan kemudian network encoder (gambar 1.) Sistem keseluruhan ini dinamakan dengan “auto encoder” seperti yang digambarkan pada gambar 2.5.



Gambar 2.5: Arsitektur auto encoder [Hinton, 2006]

Karena sulit untuk mengoptimasi bobot pada auto encoder yang nonlinear. Dikarenakan sistem neural network tersebut memiliki banyak hidden layer (2-4) dengan bobot awal yang juga banyak, sehingga biasanya dapat terjebak kedalam lokal minima; dengan bobot awal yang kecil dan gradien pada layer awal yang kecil, maka pelatihan auto encoder dengan hidden layer yang banyak sangat sulit. Oleh karena itu dikenalkanlah metode “*pretraining*”, yaitu prosedur untuk data biner yang menggeneralisir data asli. Yang telah dibuktikan oleh Hinton, sangat bagus untuk digunakan di berbagai macam dataset.

Rumus fungsi energy :

(1)

Cross Entropy – Error

 (2)

Change of Weight :

(3)

Dalam merancang vektor biner dapat dimodelkan dengan dua layer network yang disebut dengan “Restricted Boltzmann machine” atau disingkat dengan RBM.

## Evaluasi Hasil Clustering dengan Adjusted Rand Index

Untuk membandingkan hasil clustering dengan kriteria eksternalnya, dibutuhkan pengukuran tingkat kecocokan. Adjusted Rand Index (ARI) (Hubert, 1985) menghitung tikat kecocokan antara dua partisi pada himpunan obyek. Milligan dan Copper (1986) merekomendasikan ARI sebagai pengukuran tinkat kecocokan sebuah cluster dengan cluster luarnya, walaupun, cluster luarnya memiliki lebih dari satu cluster.

Diberikan himpunan n obyek , misalkan dan . U dan V merepresentasikan dua obyek yang berbeda pada himpunan S. Pada kasus ini, partisi V merupakan cluster eksternal. U adalah hasil cluster. Misalkan a adalah jumlah pasangan obyek yang ada pada elemen yang sama dari partisi U dan juga elemen yang sama pada partisi V. Dan misalkan d adalah jumlah pasangan pasangan obyek yang ada pada elemen berbeda dari partisi U dan V. Maka Rand Index (rand, 1971) adalah merupakan fraksi kesamaan antar dua partisi tersebut. Rand Index memiliki nilai diantara 0 dan 1. Masalahnya Rand index untuk duap artisi acak tidak memiliki nilai yang konstan. Oleh karena itu, (Hubert,1985) mengenalkan Adjusted Rand Index (ARI) yang memperbaiki hal ini dengan mengasumsikan bentuk generalisasi dari Rand Index yaitu (index – index harapan) / (index maksimum – index harapan). Nilai maksimumnya adalah 1 dan nilai harapan jika cluster tersebut random adalah 0. Artinya, semakin tinggi nilai ARI maka semakin tinggi kesesuaian antara dua cluster tersebut.

****

BAB 3

METODOLOGI PENELITIAN

# 

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

## Gambaran Umum Penelitian

Pengumpulan Data Dan Pengolahan Awal

Merancang Algoritma

Testing dan Evaluasi

Visualisasi

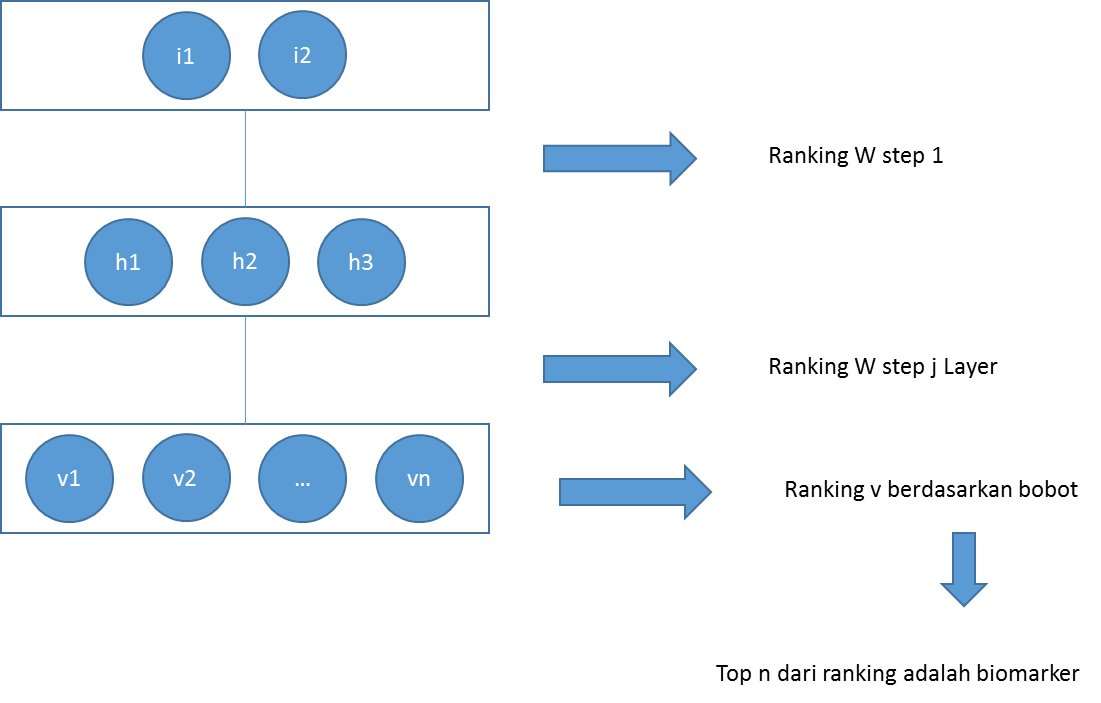
Gambar 3.1. Overview Penelitian

## 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.

Gambar 3.1. Pengumpulan data dan Pengolahan Awal

## Perancangan Algoritma



Gambar 3.2. Perangkingan multi step bobot untuk mendapatkan biomarker

Gambar 3.3. Proses clustering setelah didapatkan biomarker

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.

## 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.

Apakah hasil untuk data training bagus?

YA

Apakah hasil untuk data testing bagus?

SELESAI

YA

TIDAK

TIDAK

Ubah arsitektur / Perbesar Network

Tambahkan data yang lebih besar / Perbaiki parameter

Gambar 3.3. Proses Pengujian Arsitektur Deep Learning (Ng, 2012)

## Evaluasi Hasil Kombinasi Seleksi Fitur dan Clustering

Untuk mengevaluasi seberapa baik hasil dari percobaan ini, dilakukan dengan menghitung *Adjusted Rand Index (ARI)* (Hubert, 1985). ARI ini digunakan untuk mengukur seberapa baik 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 bagus dibandingkan dengan kriteria luar cluster, maka nilai ARI mendekati 1.

Untuk metode evaluasi hasil, ada beberapa hal yang perlu dibuktikan, tidak terjadi *curse of dimensionality* dan *over fitting*, dibuktikan dengan error saat training dan testing harus lebih kecil dibandingkan dengan error saat tidak dilakukan reduksi dimensi.

**Referensi**

Babu, M. Madan. "Introduction to microarray data analysis." *Computational Genomics: Theory and Application* (2004): 225-249.

Bandyopadhyay, Supriyo, 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 (2014): 95-115.

Elloumi, Mourad, and Albert Y. Zomaya. *Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications*. Vol. 21. John Wiley & Sons, 2011.

Fakoor, Rasool, et al. "Using deep learning to enhance cancer diagnosis and classification." *Proceedings of the International Conference on Machine Learning*. 2013.

Häggström, M. "Medical gallery of Mikael Häggström 2014." *Wikiversity Journal of Medicine1* 2 (2014).

Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." *Science* 313.5786 (2006): 504-507.

Liang, Ming, et al. "Integrative Data Analysis of Multi-platform Cancer Data with a Multimodal Deep Learning Approach." (2014).

Mallik, Saurav, Anirban Mukhopadhyay, and Ujjwal Maulik. "RANWAR: Rank-Based Weighted Association Rule Mining From Gene Expression and Methylation Data." *NanoBioscience, IEEE Transactions on* 14.1 (2015): 59-66.

Shevade, S. K., & Keerthi, S. S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, *19*(17), 2246-2253.

Smyth, Gordon K. "Limma: linear models for microarray data."*Bioinformatics and computational biology solutions using R and Bioconductor*. Springer New York, 2005. 397-420.

Stacklies, Wolfram, et al. "pcaMethods—a bioconductor package providing PCA methods for incomplete data." *Bioinformatics* 23.9 (2007): 1164-1167.

Wang, Wei, et al. "Generalized Autoencoder: A Neural Network Framework for Dimensionality Reduction." *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*. IEEE, 2014.

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

Xu, Rui, and Donald Wunsch. "Survey of clustering algorithms." *Neural Networks, IEEE Transactions on* 16.3 (2005): 645-678.