Analisis Ekstraksi Fitur *Principle Component Analysis* pada Klasifikasi *Microarray* Data Menggunakan *Classification And Regression Trees*

Rizky Pujianto[1], Adiwijaya[2], Aniq A. Rahmawati[3]
[1,2,3]Prodi Ilmu Komputasi, Fakultas Informatika, Universitas Telkom, Bandung
[1]rizkyrp@students.telkomuniversity.ac.id, [2]adiwijaya@telkomuniversity.ac.id,
[3]aniqatiqi@telkomuniversity.ac.id

Abstrak
Di era yang sudah maju seperti saat ini pendeteksian kanker bisa dilakukan dengan beberapa cara, salah satunya dengan bioinformatika, yaitu dengan menggunakan teknologi *microarray*. Teknologi tersebut berupa DNA yang berbentuk *microchip* dengan ukuran dimensi yang sangat besar. Ukuran dimensi yang besar menyebabkan lamanya perhitungan komputasinya. Untuk mengurangi masalah komputasi maka dilakukan reduksi dimensi terlebih dahulu sebelum diklasifikasi menggunakan (*Classification and Regression Trees*) CART. Reduksi dimensi adalah pendekatan dengan memilih komponen, komponen ini dipilih karena tidak semua atribut pada data *microarray* dipilih, mengingat data pada *microarray* sangat banyak. Komponen yang paling memiliki ciri yang dipilih agar perhitungan bisa lebih menghasilkan hasil yang optimum. Reduksi dimensi yang digunakan pada penelitian ini adalah ekstraksi fitur dengan menggunakan algoritma *principle component analysis* (PCA). Ekstraksi fitur biasanya digunakan untuk data kontinu dengan cara mengekstrak atributnya sehingga tersisa atribut yang dapat mengoptimalkan hasilnya. Data kanker yang digunakan ada tiga yaitu, kanker usus besar, leukimia, dan kanker paru-paru. Akurasi yang dihasilkan dari penelitian ini rata-rata diatas 70% dengan algoritma PCA untuk reduksi dimensi dan CART sebagai klasifikasinya.

Kata kunci: kanker, *microarray*, reduksi dimensi, CART

Abstract
In an advanced era such as the current detection of cancer can be done in several ways, one of which is bioinformatics by using microarray technology. The technology consists of DNA that forms microchips with very large dimensions. Large size dimensions cause computational calculations. To reduce computational problems, the reduction is done before being classified using (Tree Classification and Regression) CART. Dimension reduction by selecting components, this component is selected because not all attributes in the microarray data are selected, considering that the data on the microarray is very large. Components that have the most characteristics are chosen so that calculations can produce optimal results. Dimension reduction used in this study is feature extraction using the principal component analysis (PCA) principle. Feature extraction is usually used for continuous data by extracting attributes so that they can produce attributes. There are three cancer data used, namely, colon cancer, leukemia, and lung cancer. The accuracy generated from this study averages over 70% with the PCA algorithm for reducing dimensions and CART as its classification.

Key word: cancer, microarray, dimention reduction, CART

1. Pendahuluan

Latar Belakang

Sel-sel didalam tubuh pada umumnya membelah diri sesuai dengan waktunya, namun pada orang yang terkena kanker, sel-sel didalam tubuhnya mengalami mutasi yang sangat cepat dimana sel-sel baru yang tidak dibutuhkan akan terus tumbuh sampai tidak terkendali. Untuk memperkecil risiko terkena kanker, dapat dilakukan pendeteksian kanker dengan beberapa metode yaitu, CT (*Computed Tomography*), MRI (*Magnetic Resonance Imaging*), PET (*Positron Emission Tomography*), *Ultra sound examination*, *Endoscopic examinations*, *In mammography*, *In isotopic diagnostics*, dan *gene expression* [1]. Selain itu pendeteksian dengan *gene expression* bisa dilakukan dengan teknologi *microarray*.

Microarray adalah teknologi yang digunakan untuk mempelajari ekspresi dari banyak gen, teknologi *microarray* ini lebih khusunya untuk menangani kanker walaupun bisa juga untuk berbagai macam penyakit. Teknologi *microarray* ini sudah digunakan dalam bidang kedokteran, teknologi ini digunakan karena pendeteksian kanker melalui tekologi *microarray* yang cepat, tetapi ada kendala dalam penggunaan *microarray*, yaitu besarnya dimensi yang dimiliki sehingga perlu dilakukan reduksi dimensi [2]. Dimensi yang dimaksud adalah ukuran data berupa atribut dan *sample*. Atribut berisi tentang informasi-informasi *gene expression* dari DNA dan *sample* berisi banyaknya DNA. Reduksi dimensi ada dua jenis yaitu *feature selection* dan *feature extraction*. *Feature* adalah atribut pada *microarray* data. Pada penelitian ini digunakan reduksi dimensi *feature extraction*. *Feature extraction* adalah salah satu cara untuk mereduksi dengan mengkompres atribut yang jumlahnya banyak menjadi lebih sedikit

agar mudah mendapatkan hasil yang optimum [3]. Salah satu metode yang ada di *feature extraction* adalah *Principal Component Analysis* (PCA). Pada tahun 2006 Jian J. Dai, Linh Lieu, dan David Rocke menganilisa *microarray* data menggunakan algoritma *Principal Component Algorithm* dan hasil error yang didapat adalah 0.042 untuk leukemia dan 0.162 untuk *colon* [9]. Algoritma PCA digunakan karena optimum untuk data yang jumlahnya banyak [2].

Dari penelitian ini ingin diketahui performansi *Principal Component Algorithm* sebagai reduksi dimensi dari *microarray* data dengan klasifikasi menggunakan *Classification and Regression Trees* (CART). Pada tahun 2014 Kun-Huang Chen et al menganalisa sepuluh *microarray* data dengan metode CART sebagai metode klasifikasinya dan algoritma *Binary Particle Swarm Optimization* (BPSO) sebagai seleksi fitur, didapatkan hasil akurasi rata-rata sebesar 70%. Algoritma CART digunakan karena memiliki akurasi yang tinggi dan waktu komputasinya yang rendah [2].

Topik dan Batasannya

Pengukuran kanker dengan teknologi *microarray* menggunakan DNA, dialam DNA tersebut menyimpan banyak atribut dan juga *sample*. Dari atribut dan juga *sample* tersebut akan dilakukan pendeteksian kanker, apakah positif atau negatif terkena kanker. Dengan banyaknya atribut-atribut yang tersimpan pada setiap DNA, maka atribut tersebut perlu direduksi untuk efisiensi dan menghindari multikolinearitas pada data yang akan mempengaruhi akurasi prediksi. Reduksi dimensi dilakukan untuk mengetahui pengaruh akurasi dan performansi yang dihasilkan dari klasifikasinya. Pada penelitian ini reduksi dimensi yang digunakan adalah PCA sebagai fitur ekstraksi dengan *proportion of variance* (PPV) sebagai parameternya. Setelah dimensi direduksi, data kanker tersebut diklasifikasi menggunakan algoritma CART dengan memperhatikan *maximum splits* sebagai parameternya. Dan data kanker yang digunakan adalah data kanker usus besar, kanker paru-paru, dan kanker leukemia yang berasal dari Kent-Ridge Bio-medical Data Set Repository.

Tujuan

Dengan menggunakan beberapa parameter pada reduksi dimensi dan klasifikasi, akan dilakukan analisis terhadap performansi klasifikasinya mengguanakan algoritma CART yang telah direduksi terlebih dahulu menggunakan algoritma PCA.

2. Studi Terkait

DNA *Microarray*

DNA microarray adalah teknologi penyimpanan layaknya hardisk pada komputer. DNA *microarray* berbentuk *microchip* yang berisi cDNA (*complimentary* DNA), yang berfungsi untuk melihat ekspresi gen dari makhluk hidup, terutama manusia. Teknologi ini bekerja dengan cara mengukur tingkat hibridasi pada mRNA pada cDNA didalam *microchip* tersebut, kemudian hasilnya sel akan berkorelasi dengan perubahan dalam tingkat mRNA. Biasanya yang digunakan adalah dua sampel. Dengan mengukur tingkat mRNA tersebut maka akan mendapatkan reaksi pada sel tersebut, jika diamati maka akan diketahui informasi biologikal yang terjadi dan dapat menentukan kesimpulan apa yang terjadi pada sel tersebut. Teknologi ini biasa digunakan untuk mendeteksi kanker dengan melihat dari tingkat ekspresi gen tersebut. Teknologi ini dikembangkan dan banyak digunakan karena bisa melihat ekspresi gen dengan jumlah banyak dalam satu waktu [2].

Reduksi Dimensi

Reduksi dimensi adalah pendekatan dengan memilih komponen yang paling penting agar perhitungan bisa lebih menghasilkan hasil yang optimum. *Microarray* memiliki dimensi yang sangat besar, untuk menghitung *microarray* tersebut maka dilakukan reduksi dimensi terlebih dahulu. Ada dua cara dalam mereduksi dimensi yaitu dengan seleksi fitur dan ekstrasi fitur dimana keduanya memiliki keunggulan dan kelemahannya masing-masing. Seleksi fitur biasanya digunakan untuk data diskrit dengan memilih fitur-fitur mana saja yang dianggap lebih mengoptimalkan hasil Ekstraksi fitur biasanya digunakan untuk data kontinu dengan cara mengekstrak sehingga tersisa data yang dapat mengoptimalkan hasilnya [5]. Pada penelitian ini digunakan ekstaksi fitur yaitu *Principal Component Analysis* (PCA).

Preprocessing

Sebelum *microarray data* diolah, akan dilakukan terlebih dahulu *preprocessing*. *Preprocessing* adalah proses yang dilakukan untuk menghindari perbedaan nilai yang sangat jauh. *Preprocessing* pada *microarray* data adalah normalisasi data, yaitu mengubah skala *range* nilai kedalam *range* 0 sampai 1. Normalisasi dibutuhkan karena *microarray* data memiliki perbedaan *range* yang signifikan. Pada normalisasi data ini digunakan persamaan sebagai berikut

$$Y_i = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2.1)$$

Dimana Y_i adalah nilai baru fitur dalam domain normalisasi, X adalah nilai sebelum di normalisasi, X_{min} adalah nilai terkecil dari seluruh data yang dinormalisasi dalam atribut, dan X_{max} adalah nilai terbesar dari suatu data dalam atribut yang dinormalisasi. Data yang telah dinormalisasi ini tidak akan terlihat perbedaan nilai yang sangat jauh, karena nilai terbesarnya menjadi satu dan nilai terkecilnya nol.

***Principal Component Analysis* (PCA)**

Algoritma *Principal Component Analysis* (PCA) adalah algoritma untuk mereduksi dimensi dengan mengubah kumpulan dimensi yang saling berkorelasi menjadi dimensi yang tidak saling berkorelasi. Algoritma ini nantinya akan menghasilkan nilai yang disebut sebagai *PrincipalComponent* (PC). Data PC tersebut adalah kombinasi linier dari nilai-nilai asli sebelum dilakukan reduksi.

Langkah pertama dalam menggunakan algoritma PCA ini adalah dengan mencari data $X^*_{i,j}$ yang berdimensi *m* x *n*, dimana *m* adalah jumlah *sample* dan *n* adalah jumlah atributnya. Dengan menggunakan teknik zero-mean yaitu dengan mengurangi semua nilai $X_{i,j}$ pada matriks X, dengan nilai rata-rata nilai matriks $\overline{X}$ [6]. Teknik zero-mean adalah proses supaya data tersebut menjadi berdistribusi normal standar. Ini dilakukan karena menurut teorema limit pusat, jika data yang diambil mendekati jumlah populasi maka datanya semakin mendekati distribusi normal. Jadi hasil perhitungan ini bisa mewakili data sejumlah populasi.

$$X^*_{i,j} = X_{i,j} - \overline{X} \quad (2.2)$$

Setelah itu adalah mencari nilai kovarian dari matriks $X_{i,j}$, C_x merupakan nilai dari kovarian yang dicari.

$$C_x = \frac{1}{m-1} \cdot X^{*T}_{i,j} \cdot X^*_{i,j} \quad (2.3)$$

C_x adalah matriks kovariansi dari *j x j*, dan *m* adalah jumlah dari *sample*. Kemudian setelah ini mencari nilai eigen

$$|C_x - \lambda \mathrm{I}| = 0 \text{ dan } \left(C_x - \lambda \mathrm{I}\right).v = 0 \quad (2.4)$$

dimana I adalah matriks identitas, $\boldsymbol{\lambda}$ adalah nilai eigen dan v adalah vektor eigen.
Vektor eigen inilah yang menjadi komponen utama untuk menentukan variabel baru.
Untuk penentuan jumlah variabel baru yang digunakan tergantung dari persentasi kontribusi komulatif variansi V_r,

$$V_r = \frac{\sum_{j=1}^{r} \lambda_j}{\sum_{j=1}^{D} \lambda_j} . 100\% \quad (2.5)$$

dimana *D* adalah jumlah atribut awal dan *r* adalah jumlah komponen yang dipilih [7].

***Classification and RegressionTrees* (CART)**

Algoritma CART adalah algoritma yang melibatkan proses regresi dan klasifikasi. Pada proses regresi melibatkan atribut yang ada, dan klasifikasinya menggunakan *decision* tree.

Cara kerja algoritma ini adalah dengan melakukan pemisahan secara rekursif dari nilai *sample*. Algoritma ini kemudian membentuk sebuah pohon kebenaran dan akan mencari kesemua variabel untuk mendapatkan nilai yang paling optimal dengan melihat nilai *goodness* [8]. Ada tiga tahapan dalam pembangunan pohon klasifikasi yaitu, pemilihan pemilah, penentuan simpul terminal, dan penandaan label kelas.

Pada bagian pemilah pemilih akan dicari nilai impuritasnya, nilai impuritas atau indeks Gini berfungsi untuk menentukan keheterogenan suatu simpul. Simpul yang dipilih adalah simpul dengan tingkat kehomogenan respon paling tinggi. Kesesuaian nilai *goodness* merupakan pemilah *s* pada simpul *t* [8]. Untuk mencari nilai *goodness* digunakan persamaan sebagai berikut:

$$\Phi(s,t) = \Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R) \quad (2.6)$$

Dimana $\Phi(s,t)$ adalah kriteria *goodness of split*, $P_L i(t_L)$ adalah proporsi pengamatan simpul *t* menuju simpul kiri, $P_R i(t_R)$ adalah proporsi pengamatan simpul *t* menuju simpul kanan, dan $i(t)$ adalah fungsi keheterogenan indeks Gini. persamaan dari $i(t)$ adalah sebagai berikut:

$$i(t) = \sum_{i \neq j} p(i|t)p(j|t) \tag{2.7}$$

Dengan $p(i|t)$ adalah proporsi kelas *i* pada simpul *t*, dan $p(j|t)$ adalah proporsi kelas *j* pada simpul *t* [10]. Kemudian setalah mendapatkan pemilah akan dicari simpul terminal. Simpul terminal dianggap tidak berarti apabila tidak ada lagi kelas prediksi yang signifikan terhadap kelas aktual atau sudah mencapai batas maksimum dari pohon. Jika kondisi tersebut terpenuhi maka simpul *t* tidak dipilah tetapi menjadi simpul terminal [8]. Kondisi berhenti juga bisa dioptimalkan dengan membatasi jumlah cabangnya. Pembatasan cabang disini dilakukan dengan *MaxSplits*. *MaxSplits* bernilai dua maka pembentukan simpul terminal dibatasi sebanyak dua kali *splits* saja Kemudian setelah penentuan simpul terminal, dilakukan penandaan label kelas. Penandaan label kelas menggunakan aturan jumlah terbanyak dengan rumus sebagai berikut,

$$P(J_0|t) = max_j P(j|t) = max_j \frac{N_j(t)}{N(t)} \tag{2.8}$$

Dimana :

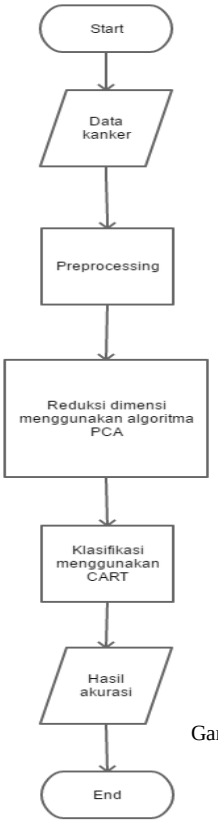
$P(j|t)$ = proporsi kelas *j* pada simpul *t*
$N_j(t)$ = jumlah pengamatan kelas *j* pada simpul *t*
$N(t)$ = jumlah pengamatan pada simpul *t*

Label kelas simpul terminal *t* adalah J_0 yang memberikan nilai dengan pengklasifikasian simpul *t* terbesar [8].

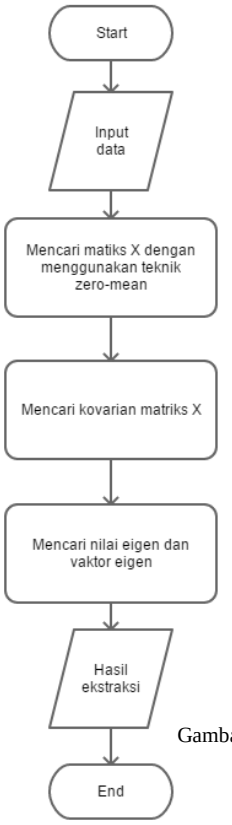
3. Sistem yang Dibangun

Gambaran Perancangan Sistem

Sistem perancangan pada gambar 3-1 adalah sistem yang nantinya dapat mengklasifikasikan *microarray* data. Sebelum data ini diproses *microarray* data memiliki dimensi yang sangat besar oleh karena itu, akan dilakukan reduksi dimensi. Metode yang digunakan untuk mereduksi dimensi pada penelitian ini adalah ekstraksi fitur *principal component analysis*. Kemudian, proses klasifikasi dilakukan untuk memprediksi kelas kanker dan yang bukan kanker. Selanjutnya hasil prediksi akan dihitung akurasinya menggunakan metode manual dengan membandingan kelas prediksi dengan kelas asli. Berikut merupakan gambaran umum pada tugas akhir ini:



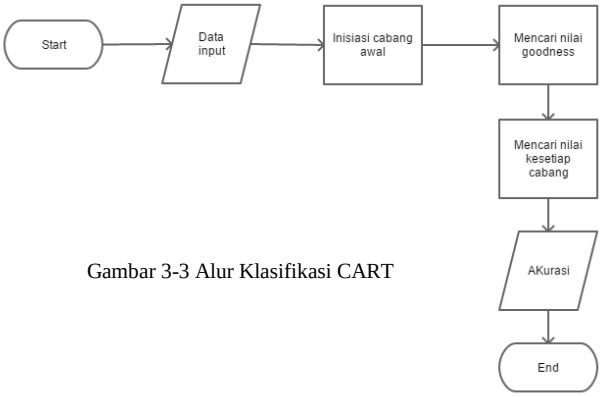
Gambar 3-1 Skema secara umum



Gambar 3-2 Skema PCA

Pada gambar 3-2 data kanker yang sudah di normalisasi akan direduksi menggunakan algoritma PCA, pada algoritma ini sebelum direduksi, data harus berupa matriks $i \times j$, dimana i adalah banyaknya baris atau disebut sebagai *sample*, dan j adalah banyaknya kolom atau disebut juga atribut. Setelah matriks terbentuk dilakukan pencarian nilai kovarian dari matriks data tersebut. Setelah itu dicari nilai eigen dan vector eigen.

Data yang sudah direduksi kemudian masuk kedalam proses klasifikasi. Pada penelitian ini proses klasifikasi digunakan metode *Classification and Regression Tress* (CART). Hal pertama yang dilakukan adalah mencari nilai *goodness* $\Phi(s|t)$, setelah nilai maka dilakukan pencarian terhadap cabang-cabangnya. Untuk alur dari sistemnya dapat dilihat pada gambar berikut:



Gambar 3-3 Alur Klasifikasi CART

Rencana Kebutuhan Data

Pada penelitian ini data yang digunakan adalah data kanker yang berasal dari Kent-Ridge yang dapat dari http://leo.ugr.es/elviraDBCRepository/. Dataset terdiri dari *kanker usus besar*, *Leukimia*, dan *kanker paru-paru*. Dari dataset inilah yang akan dinormalisasi dan direduksi dimensinya untuk mendapatkan klasifikasi mengenai kanker. Berikut merupakan distribusi data kanker.

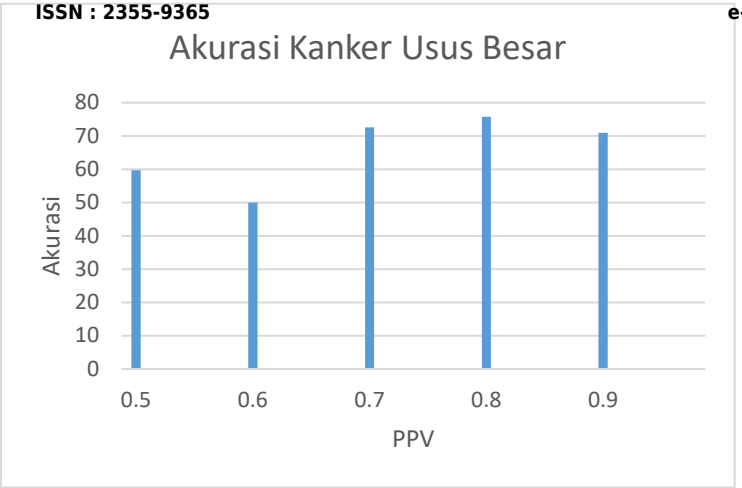
Tabel 3-1 Data kanker *Kent - Ridge*

Dataset	Jumlah gen	Jumlah kelas	Jumlah sampel	
Leukimia	7129	2	27 Positif	11 Negatif
Kanker paru-paru	12533	2	31 Positif	150 Negatif
Kanker usus besar	2000	2	22 Positif	40 Negatif

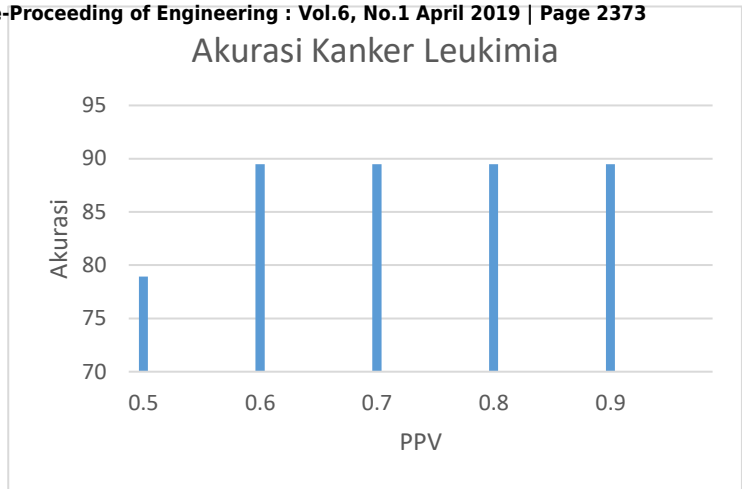
4. Evaluasi

4.1. Hasil Pengujian dan Analisis

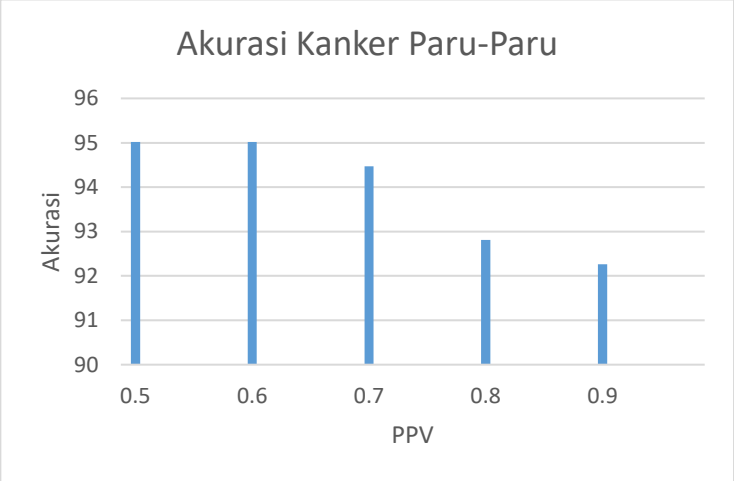
Pengujian dilakukan terhadap tiga data kanker, yaitu kanker usus besar, Leukimia, dan kanker paru-paru. Sebelum dilakukan klasifikasi, atribut tiap-tiap data kanker direduksi terlebih dahulu menggunakan algoritma PCA. Parameter yang digunakan dalam reduksi dimensi menggunakan PCA adalah proportion of variance (PPV). PPV adalah nilai perbandingan antara nilai eigen dengan jumlah keseluruhan nilai eigen. Semakin tinggi nilai PPV nya maka nilai eigen yang akan semakin besar, dan atribut yang terpilih semakin banyak, karena atribut yang terpilih semakin banyak maka akurasi cenderung naik. Parameter yang digunakan pada klasifikasinya adalah nilai dari maximum splits, yaitu nilai untuk membatasi level dari pohon klasifikasinya. Level ini dibatasi untuk mempercepat perhitungan komputasinya dan untuk pengoptimalan hasil akhirnya, karena jika tidak dibatasi maka, semakin banyak atribut yang terpilih semakin banyak juga split yang terbentuk, sehingga bisa saja terjadi kesalahan pelebelan jika split yang terbentuk semakin banyak. Besarnya akurasi yang dihasilkan tergantung berapa split yang dibentuk. Terkadang semakin banyak split yang dibentuk justru memperkecil nilai akurasi. Percobaan ini menggunakan dua parameter yaitu nilai *maximum splits* dan PPV.



Gambar 4-1 Grafik pengaruh PPV terhadap akurasi kanaker usus besar



Gambar 4-2 Grafik pengaruh PPV terhadap akurasi kanaker leukimia



Gambar 4-3 Grafik pengaruh PPV terhadap akurasi kanaker paru-paru

Diketahui bahwa nilai dari PPV mempengaruhi hasil dari akurasinya. Semakin besarnya PPV memiliki akurasi cenderung naik, itu disebabkan Karena jika PPV semakin besar nilai eigen yang terpilih juga semakin besar. Nilai eigen merepresentasikan keunikan nilai setiap atributnya, semakin besar nilai eigen maka atribut yang terpilih adalah atribut yang memiliki ciri paling berbeda dari atribut lainnya. Maka dari itu pemilihan PPV sangat nmempengaruhi hasil dari akurasi, pemilihan PPV yang besar bisa membuat akurasi menjadi besar juga. Untuk meningkatkan hasil akurasi klasifiasi, koreksi terhadap jumlah *split* masih dimungkinkan untuk dilakukan. Sehingga pemilihan *split* dari *system* masih memungkinkan dioptimasi oleh *user*. Besarnya PPV juga mempengaruhi akurasi pada klasifikasi menggunakan metode *optimal by system* maupun *optimal by user*. Besarnya PPV tidak selalu berbanding lurus dengan hasil akurasi. Hal tersebut terjadi karena bisa saja ada atribut yang terpilih masih berkorelasi dengan atribut lainnya walaupun memiliki keunikan atau ciri yang berbeda. Akurasi dari CART *optimal by user* juga tidak selalu lebih besar daripada CART dengan *optimal by system* dikarenakan CART dengan metode *optimal by system* sudah disesuaikan untuk pengoptimalan daripada *optimal by users*. *Optimal by users* bisa saja digunakan dan mendapatkan hasil optimal jika pemilihan *splits* yang bervariasi . Rata-rata akurasi yang dipengaruhi oleh besarnya PPV dan *maximum splits* tidak selalu naik. Pada nilai PPV 0.9 untuk setiap data kanker akurasinya cenderung turun. Itu disebabkan terlalu banyak atribut yang terpilih, dan bisa saja atribut yang terpilih bukan yang memiliki nilai eigen yang besar. Sehingga tidak memiliki ciri khusus untuk dikenali sebagai kelas kanker maupun tidak kanker.

5. Kesimpulan

1. Besarnya nilai PPV mempengaruhi akurasi dengan metode *optimal by user* maupun dengan metode *optimal by system*, semakin besarnya nilai PPV maka *sample* yang terpilih saat direduksi juga semakin banyak.
2. Rata-rata akurasi dari CART dengan metode *optimal by system* hampir selalu lebih besar daripada dengan metode *optimal by user*.
3. Analisis simulasi pada dataset *microarray* menunjukan bahwa prediksi klasifikasi yang melibatkan PCA menghasilkan rata-rata akurasi diatas 77%. Pada akhirnya sistem ini menunjukan pendeteksian kanker dengan menggunakan teknik *microarray* data dapat dilakukan dengan menggunakan algoritma CART sebagai proses klasifikasinya yang melibatkan algoritma PCA untuk mereduksi dimensi.

Daftar Pustaka

[1] https://www.allaboutcancer.fi/facts-about-cancer/detection/. [Accessed 20 February 2018].

[2] Yip, Wai-Ki., Amin, Samir B., Li, Cheng. 2011. Chapter 10: A Survey of Classification Techniques for Microarray Data Analysis. Springer Handbooks of Computational Statistics.

[3] M. Kusban, "Verifikasi dan Identifikasi Telapak Tangan dengan Karnel Gabor," Jurnal Nasional Teknik Elektro Teknik Informatika (JNTETI), vol. 4, no. 2, 2015.

[4] Dash, M. and H. Liu. 1997. Feature Selection for Classification, Intelligent Data Analysis,1(1-4).

[5] Liu H. Feature Extraction, Construction and Selection: A Data Mining Perspective, ISBN 0-7923-8196-3, Kluwer Academic Publishers, 1998.

[6] Nurfalah, A. Adiwijaya. and Suryani, A., 2016. CANCER DETECTION BASED ON MICROARRAY DATA CLASSIFICATION USING PCA AND MODIFIED BACK PROPAGATION. Far East Journal of Electronics and Communications, 16(2), p.269.

[7] Susetyoko, R. and Purwantini, E., Teknik Reduksi Dimensi Menggunakan Komponen Utama Data Partisi Pada Pengklasifikasian Data Berdimensi Tinggi dengan Ukuran Sampel Kecil. Dimensi, 1, p.500.

[8] Timofeev R. 2004. Classification and Regression Trees (CART) Theory and Application [tesis]. Berlin: Center of Applied Statistics and Economics, Humboldt University.

[9] Jian J. Dai, Linh Lieu, dan David Rocke. 2006. Dimension Reduction for Classification with Gene Expression Microarray Data. Statistical Application in Genetics and Molecular Biology. Vol 5 Issue 1.

[10] A. Syaukat, "Faktor – Faktor Yang Menentukan Pilihan Daerah Tujuan Migrasi Penduduk Jawa Barat," Tesis Pasca Sarjana, Program Studi Kajian Kependudukan dan Ketenagakerjaan, Universitas Indonesia, Jakarta, Indonesia (1997).

[11] Aydadenta, H., Adiwijaya, 2018. A Clustering Approach for Feature Selection in Microarray Data Classification Using Random Forest. JIPS (Journal of Information Processing Systems), 14(5), pp.1167-1175.

[12] Adiwijaya, U. N. Wisesty, E. Lisnawati, A. Aditsania, D. S. Kusumo, (2018). Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification, Journal of Computer Science 14.

[13] Adiwijaya, A., 2018. Deteksi Kanker Berdasarkan Klasifikasi Microarray Data. MEDIA INFORMATIKA BUDIDARMA, 2(4), pp.181-186.

[14] Aydadenta, Husna. Adiwijaya. "On the classification techniques in data mining for microarray data classification." Journal of Physics: Conference Series. Vol. 971. No. 1. IOP Publishing, 2018.

LAMPIRAN

Tabel Hasil dari data kanker usus besar

no	PPV	atribut	nilai eigen	Akurasi System
1	0.5	2	4.824	59.67
2			4.824	59.67
3			4.824	59.67
4			4.824	59.67
5	0.6	3	6.0306	50
6			6.0306	50
7			6.0306	50
8			6.0306	50
9	0.7	6	7.2916	72.58
10			7.2916	72.58
11			7.2916	72.58
12			7.2916	72.58
13	0.8	10	9.0166	75.8
14			9.0166	75.8
15			9.0166	75.8
16			9.0166	75.8
17	0.9	22	11.4205	70.96
18			11.4205	70.96
19			11.4205	70.96
20			11.4205	70.96

no	PPV	atribut	MaxSplit	nilai eigen	Akurasi User
1	0.5	2	2	4.824	58.06
2			5	4.824	54.83
3			7	4.824	58.06
4			10	4.824	59.67
5	0.6	3	2	6.0306	62.9
6			5	6.0306	58.06
7			7	6.0306	62.9
8			10	6.0306	64.51
9	0.7	6	2	7.2916	77.41
10			5	7.2916	66.12
11			7	7.2916	66.12
12			10	7.2916	66.12
13	0.8	10	2	9.0166	75.8
14			5	9.0166	74.19
15			7	9.0166	74.19
16			10	9.0166	74.19
17	0.9	22	2	11.4205	82.25
18			5	11.4205	72.58
19			7	11.4205	72.58
20			10	11.4205	72.58

Tabel Hasil dari data kanker paru-paru

no	PPV	atribut	nilai eigen	Akurasi System
1	0.5	15	18.8915	95.02
2			18.8915	95.02
3			18.8915	95.02
4			18.8915	95.02
5	0.6	25	23.333	95.02
6			23.333	95.02
7			23.333	95.02
8			23.333	95.02
9	0.7	41	28.7929	94.47
10			28.7929	94.47
11			28.7929	94.47
12			28.7929	94.47
13	0.8	68	34.9846	92.81
14			34.9846	92.81
15			34.9846	92.81
16			34.9846	92.81
17	0.9	93	1.8964	92.26
18			1.8964	92.26
19			1.8964	92.26
20			1.8964	92.26

no	PPV	atribut	MaxSplit	nilai eigen	Akurasi User
1	0.5	15	2	18.8915	96.13
2			5	18.8915	94.47
3			7	18.8915	94.47
4			10	18.8915	94.47
5	0.6	25	2	23.333	96.13
6			5	23.333	94.47
7			7	23.333	94.47
8			10	23.333	94.47
9	0.7	41	2	28.7929	96.13
10			5	28.7929	93.92
11			7	28.7929	93.92
12			10	28.7929	93.92
13	0.8	68	2	34.9846	96.13
14			5	34.9846	93.37
15			7	34.9846	93.37
16			10	34.9846	93.37
17	0.9	93	2	1.8964	90.6
18			5	1.8964	91.16
19			7	1.8964	91.16
20			10	1.8964	91.16

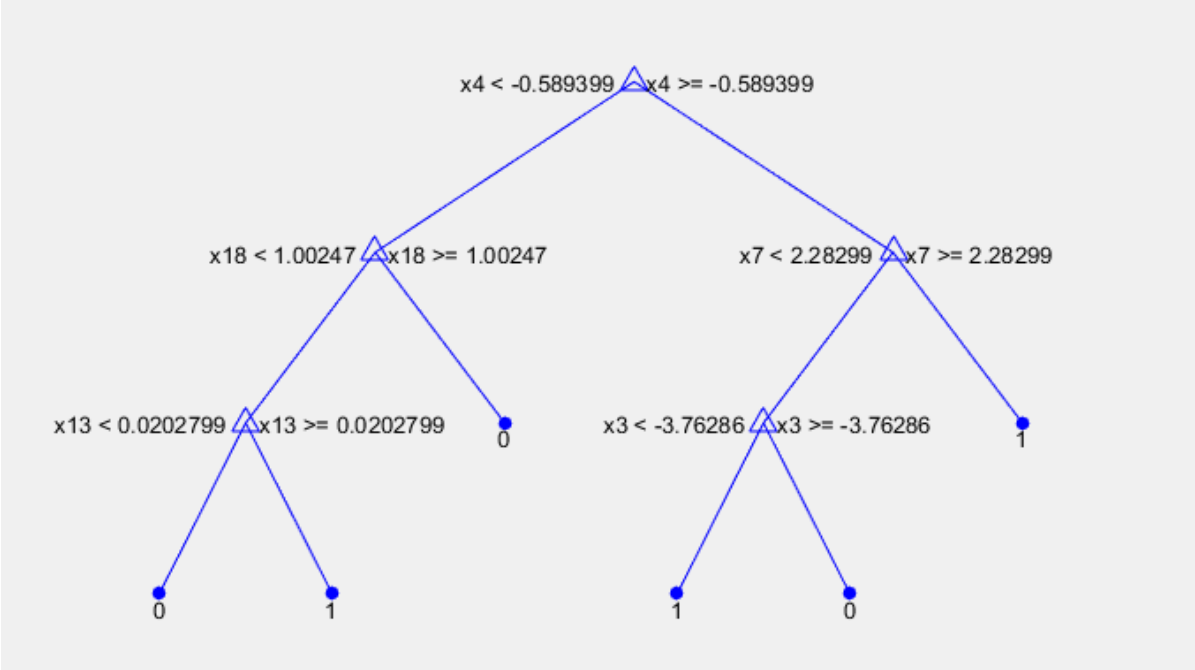
Tabel Hasil dari data laeukimia

no	PPV	atribut	nilai eigen	Akurasi System
1	0.5	8	7.8973	78.94
2			7.8973	78.94
3			7.8973	78.94
4			7.8973	78.94
5	0.6	11	9.8573	89.47
6			9.8573	89.47
7			9.8573	89.47
8			9.8573	89.47
9	0.7	16	11.843	89.47
10			11.843	89.47
11			11.843	89.47
12			11.843	89.47
13	0.8	22	14.3304	89.47
14			14.3304	89.47
15			14.3304	89.47
16			14.3304	89.47
17	0.9	29	17.5158	89.47
18			17.5158	89.47
19			17.5158	89.47
20			17.5158	89.47

no	PPV	atribut	MaxSplit	nilai eigen	Akurasi User
1	0.5	8	2	7.8973	78.94
2			5	7.8973	81.57
3			7	7.8973	81.57
4			10	7.8973	81.57
5	0.6	11	2	9.8573	86.84
6			5	9.8573	92.1
7			7	9.8573	92.1
8			10	9.8573	92.1
9	0.7	16	2	11.843	86.84
10			5	11.843	92.1
11			7	11.843	92.1
12			10	11.843	92.1
13	0.8	22	2	14.3304	86.84
14			5	14.3304	92.1
15			7	14.3304	92.1
16			10	14.3304	92.1
17	0.9	29	2	17.5158	86.84
18			5	17.5158	92.1
19			7	17.5158	92.1
20			10	17.5158	92.1

Tabel Hasil rata-rata akurasi dari data kanker

No	PPV	MaxSplits	Rata-Rata Akurasi Data Kanker			Rata-Rata
			Usus Besar	Paru-Paru	Leukimia	
1	0.5	2	58.865	95.575	78.94	77.793
2		5	57.25	94.745	80.255	77.417
3		7	58.865	94.745	80.255	77.955
4		10	59.67	94.745	80.255	78.223
5	0.6	2	56.45	95.575	88.155	80.060
6		5	54.03	94.745	90.785	79.853
7		7	56.45	94.745	90.785	80.660
8		10	57.255	94.745	90.785	80.928
9	0.7	2	74.995	95.3	88.155	86.150
10		5	69.35	94.195	90.785	84.777
11		7	69.35	94.195	90.785	84.777
12		10	69.35	94.195	90.785	84.777
13	0.8	2	75.8	94.47	88.155	86.142
14		5	74.995	93.09	90.785	86.290
15		7	74.995	93.09	90.785	86.290
16		10	74.995	93.09	90.785	86.290
17	0.9	2	76.605	91.43	88.155	85.397
18		5	71.77	91.71	90.785	84.755
19		7	71.77	91.71	90.785	84.755
20		10	71.77	91.71	90.785	84.755



Gambar Tree dengan nilai *maxsplits bernilai dua*