KLASIFIKASI DATA MICROARRAY MENGGUNAKAN DISCRETE WAVELET TRANSFORM DAN NAIVE BAYES CLASSIFICATION

MICROARRAY DATA CLASSIFICATION USING DICRETE WAVELET TRANSFORM AND NAIVE BAYES CLASSIFICATION

Rizma Nurviarelda[1], Adiwijaya[2], Aniq A Rohmawati[3]

[1]Prodi S1 Informatika, Fakultas Informatika, Universitas Telkom
[2,3]Prodi S1 Ilmu Komputasi, Fakultas Informatika, Universitas Telkom
[1]nurviareldarizma@student.telkomuniversity.ac.id, [2]adiwijaya@telkomuniversity.ac.id, [3]aniqatiqi @telkomuniversity.ac.id

Abstrak

Saat ini, kanker adalah salah satu penyakit paling mematikan. Sehingga, dibutuhkan sebuah program untuk deteksi kanker secara akurat. Pada data kanker biasanya data berupa data *microarray*. Dimana, atribut terdiri dari informasi gen seorang individu dan data objek adalah individu-individu yang terdeteksi kanker. Informasi gen terdiri dari jumlah yang sangat banyak hingga mencapai puluhan ribu. Sedangkan, jumlah individu berdasarkan jenis kanker namun hanya berkisar puluhan hingga ratusan individu. Tugas akhir ini bertujuan untuk melakukan proses klasifikasi deteksi kanker dengan mereduksi atribut menggunakan *Discrete Wavelet Transform family daubechies4* (db4) kemudian dilakukan proses klasifikasi menggunakan *Naive Bayes*. Lalu hasil akan dibandingkan dengan menggunakan seleksi atribut *Minimum-Redundancy Maximum-Relevance* jenis *F-Test Correlation Difference* dengan metode klasifikasi *Naive Bayes*. Pengujian yang dilakukan mengambil jumlah atribut terbaik pada metode db4. Sistem yang dibuat menggunakan db4 dengan metode klasifikasi *Naive Bayes* mendapatkan hasil yang baik. Dimana, nilai akurasi mencapai 98,4126%.

Kata kunci : Kanker, data microarray, daubechies4, Minimum-Redundancy Maximum-Relevance, Naive Bayes.

Abstract

Nowadays, cancer is one of the deadliest diseases. Thus, a program for cancer detection is required. In cancer's data usually data onto the form of data microarray. Where the attributes consist of an individual's gene information and data object are individuals detected by cancer. Gene information consists of a very large number up to tens of thousands information. Meanwhile, the number of individuals based on the type of cancer but only ranged from tens to hundreds of individuals. This type of research aims to process the classification of cancer detection by reducing attributes using Discrete Wavelet Transform family daubechies4 (db4) and then classification processes using Naive Bayes. Then the results will be compared using the Minimum-Redundancy Maximum-Relevance attribute type F-Test Correlation Difference with the Naive Bayes classification method. The tests performed to take the best number of attributes on the db4 method. The system created using db4 with the Naive Bayes classification method gets good results. Where the accuracy values reached 98.4126%.

Keywords: Cancer, data microarray, daubechies4, Minimum-Redundancy Maximum-Relevance, Naive Bayes.

1. Pendahuluan

Data adalah sekumpulan keterangan atau bahan nyata yang dapat dijadikan dasar kajian (analisis atau kesimpulan) [1]. Menurut Kamus Besar Bahasa Indonesia, pengertian data dalam dunia komputer adalah informasi dalam bentuk yang dapat diproses oleh komputer, seperti representasi digital dari teks, angka, gambar grafis, atau suara. Selain bentuknya yang bermacam-macam, data memiliki berbagai jenis dan struktur, salah satunya adalah data *microarray*. Data *microarray* adalah jenis data yang memiliki struktur data set yang berbeda daripada data set pada umumnya. Pada umumnya, data set terdiri dari banyak data objek dan beberapa atribut. Namun pada data *microarray*, data objek biasanya hanya terdiri kurang dari seratus sedangkan atributnya dapat berjumlah ribuan hingga puluhan ribu [2]. Data *microarray* adalah teknik yang efisien digunakan pada deteksi kanker [3]. Data objek pada data set menunjukkan individu yang akan dideteksi apakah terklasifikasi kanker atau tidak. Atribut menunjukkan gen-gen dari setiap individu tersebut [2].

Kanker adalah salah satu penyakit mematikan yang hingga saat ini masih belum terdapat obat untuk menghilangkan sel kanker yang ada pada tubuh. Dalam proses deteksi kanker, diperlukan reduksi atribut supaya dari ribuan atau puluhan ribu atribut sebelumnya dapat berkurang menjadi ratusan atribut. Sehingga dapat memudahkan untuk proses selanjutnya, yaitu proses klasifikasi. Sudah banyak metode yang digunakan untuk proses seleksi ataupun ekstraksi atribut pada data *microarray*. Beberapa metode sudah terbukti cocok sebagai seleksi atribut, ekstraksi atribut, dan klasifikasi pada data *microarray* untuk mendeteksi kanker. Menurut penelitian sebelumnya, dengan menggunakan

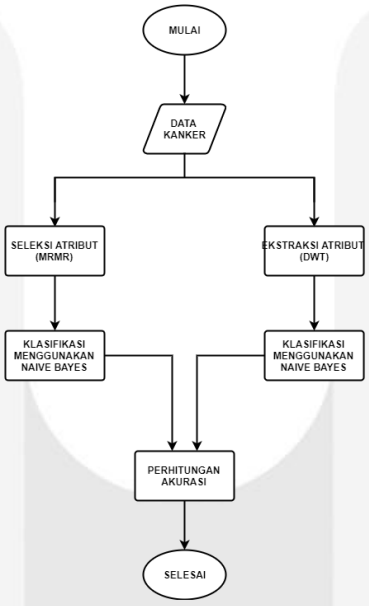
seleksi atribut *Information Gain* dan klasifier *Support Vector Machine* nilai akurasi deteksi kanker mencapai 98% [4]. Pada tahun 2008, dilakukan penelitian deteksi kanker menggunakan metode MRMR untuk seleksi atribut dan menghasilkan nilai akurasi yang cukup baik [5]. Saat ini mulai banyak penelitian menggunakan metode *Discrete Wavelet Transform (DWT)* sebagai ekstraksi atribut pada data *microarray*. DWT biasanya digunakan dalam pengolahan citra karena bentuk pengklasifikasian DWT berupa sebuah sinyal. Pada penelitian yang dilakukan oleh Ikka Damayana dengan judul "Deteksi Kanker Kulit Melanoma berbasis Pengolahan Citra menggunakan Wavelet Transform" menghasilkan nilai akurasi sebesar 76% [6].

Pada penelitian ini dilakukan proses *pre-processing* dengan ekstraksi dan seleksi atribut menggunakan DWT dan MRMR. Kemudian, proses klasifikasi dilakukan menggunakan metode *Naive Bayes*. Ekstraksi atribut dilakukan menggunakan DWT karena kemampuannya dalam mengekstraksi atribut pada data *microarray* hingga level tertentu. Seleksi atribut dilakukan menggunakan metode MRMR karena metode ini biasa digunakan dalam mengidentifikasi karakteristik gen. Seleksi atribut dilakukan berdasarkan nilai maksimum dari relevansi data antar atribut dengan minimum data yang redundan. Metode klasifikasi *Naive Bayes* dipilih karena metode ini merupakan metode yang stabil dalam melakukan klasifikasi karena sifat antar data yang saling independen. Pada penelitian ini akan membuktikan ekstraksi atribut menggunakan DWT dengan metode klasifier *Naive Bayes* cocok untuk data *microarray* dalam deteksi kanker dan perbandingannya dengan menggunakan metode MRMR.

2. Metodologi Penelitian

2.1 Gambaran Umum Sistem

Berikut adalah gambaran umum dari sistem yang dibuat pada penelitian ini



Gambar 1. Gambaran Umum Sistem

2.2 Dataset

Dataset yang digunakan pada penelitian ini adalah data kanker dengan jenis data *microarray*. Data kanker yang digunakan adalah *colon cancer* yang terdiri dari 62 data *object* dan 2000 atribut, *lung cancer* terdiri dari 32 data *object* dan 12.533 atribut, dan *ovarian cancer* yang terdiri dari 253 data *object* dan 15154 atribut. Data terlebih dahulu dilakukan normalisasi menggunakan *min-max normalization* sebelum dilakukan ekstraksi atau seleksi atribut [7].

2.3 Discrete Wavelet Transform

Wavelet adalah sebuah fungsi yang digunakan untuk menganalisis sinyal bergerak [8]. Salah satu pengembangan transformasi *wavelet* adalah *Discrete Wavelet Transform (DWT)* [9]. *Discrete Wavelet Transform (DWT)* adalah metode pemrosesan sinyal untuk memilih gen mana saja yang akan diproses [10]. DWT biasa digunakan pada data *microarray* karena kemampuannya untuk mengolah data multiresolusi dalam sinyal pengolahan [11]. Pada penelitian ini akan digunakan keluarga *wavelet daubechies* yaitu *daubechies4* (db4). Setiap fungsi *wavelet* memiliki *Scaling Function (father wavelet* (ϕ)) dan *wavelet function (mother wavelet* (ψ)) [12]. *Wavelet* db4 memiliki empat *scaling (h)* dan *wavelet function coefficient (g)*. Nilai-nilainya adalah sebagai berikut:

$$h_0 = \frac{1+\sqrt{3}}{4\sqrt{2}} \qquad g_0 = h_3 = \frac{1-\sqrt{3}}{4\sqrt{2}}$$

$$h_1 = \frac{3+\sqrt{3}}{4\sqrt{2}} \qquad g_1 = -h_2 = -\left(\frac{3-\sqrt{3}}{4\sqrt{2}}\right)$$

$$h_2 = \frac{3-\sqrt{3}}{4\sqrt{2}} \qquad g_2 = h_1 = \frac{3+\sqrt{3}}{4\sqrt{2}}$$

$h_3 = \frac{1-\sqrt{3}}{4\sqrt{2}}$ $g_3 = -h_0 = -\left(\frac{1-\sqrt{3}}{4\sqrt{2}}\right)$

2.4 Minimum-Redundancy Maximum-Relevance

Minimum-Redundancy Maximum-Relevance (MRMR) adalah metode yang sering digunakan dalam mengidentifikasi karakteristik gen [13]. Konsep dari metode MRMR adalah mencari nilai *relevancy* yang maksimum dengan nilai *redundancy* yang minimum. Terdapat empat skema yang dapat digunakan pada MRMR yaitu terdapat pada Tabel 1 berikut [14].

Tabel 1. Skema MRMR

Tipe Data	Akronim	Nama Metode	Rumus
Diskrit	MID	*Mutual Information Difference*	$\max\left[I(i,h) - \frac{1}{\lvert S\rvert}\sum_{j\in S} I(i,j)\right]$
	MIQ	*Mutual Information Quotient*	$\max\left\{I(i,h)/\left[\frac{1}{\lvert S\rvert}\sum_{j\in S} I(i,j)\right]\right\}$
Kontinu	FCD	*F-Test Correlation Difference*	$\max\left[F(i,h) - \frac{1}{\lvert S\rvert}\sum_{j\in S} \lvert c(i,j)\rvert\right]$
	FCQ	*F-Test Correlation Quotient*	$\max\left\{I(i,h)/\left[\frac{1}{\lvert S\rvert}\sum_{j\in S} \lvert c(i,j)\rvert\right]\right\}$

Pada penelitian ini akan digunakan metode *F-Test Correlation Difference* (FCD) karena jenis data adalah data kontinu. Nilai *Relevancy* diperoleh dari perhitungan nilai F-Test antara atribut dengan kelasnya [15].

2.5 Naive Bayes Classification

Machine Learning adalah sebuah teknologi yang sedang marak saat ini [16]. *Naive Bayes* adalah metode *classification* yang berdasarkan nilai probabilitas dari data dimana setiap data memiliki sifat saling independen [17]. Perhitungan prediksi pada metode *Naive Bayes* menggunakan aturan *Bayes's Rule* yaitu sebagai berikut [18]:

$$P(C|X) = \frac{p(X|C)p(c)}{P(X)} \quad (1)$$

C adalah nilai probabilitas dari setiap kelas yang ada dan nilai X adalah nilai probabilitas setiap atribut pada suatu kelas tertentu. P(C = 1) adalah nilai *prior probability* dari setiap atribut pada kelas C [19]. P(X|C) adalah nilai *class likelihood* dari atribut X pada kelas C. P(X) adalah nilai *evidence* dari data tetapi, nilai *evidence* dapat diabaikan karena sama dengan satu [20]. Dalam kasus numerik, nilai *likelihood* dapat diukur dengan menggunakan distribusi normal dengan rumus sebagai berikut

$$P(X_j \mid C = c_i) = \frac{1}{\sigma_{ji}\sqrt{2\pi}} \exp\left(-\frac{(X_j-\mu_{ji})^2}{2\sigma_{ji}^2}\right) \quad (2)$$

Nilai μ_{ji} adalah *mean* dari X_j dengan kelas = c_i dan σ_{ji} adalah standar deviasi dari X_j dengan kelas = c_i.

2.6 Evaluation Measure

Untuk mengukur apakah sistem yang dikembangkan dengan metode yang diusulkan sudah efisien maka digunakan tabel *confusion matrix* yang tertera pada tabel berikut.

Tabel 2. *Confusion Matrix*

		Kelas Prediksi	
		Positif	Negatif
Kelas Asli (*Actual Class*)	Positif	TP	FN
	Negatif	FP	TN

Dari Tabel 2, dapat dihitung besar akurasi dengan perhitungan seperti pada rumus (3):

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

3. Pembahasan

Pengujian dilakukan sebanyak tiga kali, yaitu hasil ekstraksi digunakan sebanyak 500 atribut, 1000 atribut, dan 1500 atribut dimana presentase data train dan data test adalah 70%-30%, 75%-25%, dan 80%-20%. Tabel berikut adalah nilai akurasi dari hasil pengujian dimana (x) adalah pengujian daubechies4 dan Naive Bayes, (y) adalah pengujian MRMR dan Naive Bayes, (a) adalah presentase data 70% data train dan 30% data test, (b) adalah presentase data 75% data train dan 25% data test, dan (c) adalah presentase data 80% data train dan 20% data test.

Tabel 3. Hasil Pengujian

Data 500 Atribut	**Akurasi (%)**					
	(a)		**(b)**		**(c)**	
	(x)	**(y)**	**(x)**	**(y)**	**(x)**	**(y)**
Colon	68,42	78,95	66,67	73,33	58,33	75
Lung	30	100	62,5	100	66,67	100
Ovarian	68,42	96,05	65,08	100	54,9	100
Data 1000 Atribut	**Akurasi (%)**					
	(a)		**(b)**		**(c)**	
	(x)	**(y)**	**(x)**	**(y)**	**(x)**	**(y)**
Colon	78,95	73,68	60	66,67	58,33	66,67
Lung	40	100	50	100	83,33	100
Ovarian	90,79	98,68	90,48	96,83	90,2	96,08
Data 1500 Atribut	**Akurasi (%)**					
	(a)		**(b)**		**(c)**	
	(x)	**(y)**	**(x)**	**(y)**	**(x)**	**(y)**
Colon	63,16	73,68	53,33	60	58,33	58,33
Lung	50	100	75	100	83,33	100
Ovarian	97,37	97,37	98,41	95,24	98,03	96,08

Dari Tabel 3 didapat bahwa penggunaan 500 atribut pada db4 dan NB nilai akurasi tertinggi diperoleh pada data colon dan data ovarian dengan nilai sebesar 68,4211%. Untuk data colon dan ovarian nilai akurasi dengan menggunakan metode db4 dan NB cenderung menurun saat bertambahnya data train. Sedangkan, dengan menggunakan MRMR dan NB cenderung stabil. Pada data lung nilai akurasi meningkat seiring dengan peningkatan jumlah data train. Hal ini terjadi karena sedikitnya jumlah data objek pada dataset Lung.

Jika menggunakan 1000 Atribut, metode db4 dan NB menghasilkan nilai akurasi tertinggi sebesar 90,7894% yaitu pada data ovarian dengan presentase data train 70% dan 30% data test. Nilai akurasi pada tiga data lebih baik saat menggunakan 1000 atribut daripada dengan 500 atribut karena, jumlah atribut awal data cukup banyak yaitu 2000 untuk data colon, 12.533 untuk data lung, dan 15.154 untuk data ovarian jika digunakan 500 atribut maka terlalu banyak atribut yang diekstraksi.

Pada penggunaan 1500 atribut, hasil akurasi tertinggi diperoleh dengan menggunakan MRMR dan NB dengan nilai sempurna untuk data Lung. Sedangkan, untuk metode db4 dan NB nilai akurasi tertinggi diperoleh dengan nilai 98,4126% pada data ovarian dengan presentase data test 25%. Pada data colon nilai akurasi dengan mengambil 1500 atribut mengalami penurunan daripada menggunakan 1000 atribut. Hal ini terjadi karena jumlah atribut yang dapat digunakan untuk ekstraksi atribut adalah tidak lebih dari nilai q dimana $q = 2^M$. Semakin mendekati q maka ekstraksi akan semakin baik.

4. Kesimpulan

Penggunaan metode db4 dan NB menghasilkan nilai akurasi yang baik yaitu mencapai 98,4126% dengan pengambilan 1500 atribut pada data ovarian. Pada data colon nilai akurasi tertinggi adalah 78,95% dan pada data lung adalah 83,33%. Semakin banyak atribut maka nilai akurasi yang didapatkan akan semakin baik. Namun, harus diketahui berapa banyak atribut yang dapat menghasilkan nilai terbaik. Sedangkan, seleksi atribut menggunakan metode MRMR dapat menghasilkan nilai akurasi yang lebih baik daripada menggunakan metode db4 karena, metode ini hanya mengambil atribut dengan nilai relevansi yang tinggi dan nilai redundansi yang rendah sehingga setelah proses klasifikasi dapat menunjukkan hasil yang baik.

Daftar Pustaka:

[1] Badan Pengembangan dan Pembinaan Bahasa, Kementerian Pendidikan dan Kebudayaan Republik Indonesia. 2016. *Data*. https://kbbi.kemdikbud.go.id/entri/data, diakses 25 April 2017.

[2] M.Dashtban and Mohammadali Balafar. *"Gene Selection for Microarray Cancer classification using a new evolutionary method employing artificial intelligence concept"*. University of Tabriz, Department of Computer Engineering. 2017.

[3] Mukesh Kumar, Nitish Kumar Rath, Amitav Swain, and Santanu Kumar Rath. *"Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor"*. National Institute of Technology Rourkela, Department of Computer Science and Engineering. 2015.

[4] World Health Organization. 2017. *Cancer*. http://www.who.int/mediacentre/factsheets/fs297/en/, diakses 25 April 2017.

[5] The American Cancer Society medical and editorial content team. 2015. *What is Cancer?*. https://www.cancer.org/cancer/cancer-basics/what-is-cancer.html, diakses 25 April 2017.

[6] Ikka Damayana, Ratri Dwi Atmaja, S.T., M.T., Hilman Fauzi, S.T., M.T. "*Deteksi Kanker Kulit Melanoma berbasis Pengolahan Citra menggunakan Wavelet Transform"*. Telkom University, Indonesia. 2016.

[7] Adiwijaya. 2014. Aplikasi Matriks dan Ruang Vektor. Yogyakarta: Graha Ilmu.

[8] Aniq Atiqi Rohmawati, Ir. Elly Anna, M.Si., Nur Chamidah, S.Si., M.Si., "*Transformasi Wavelet Diskret dan Partial Least Squares dalam Pemodelan Kalibrasi serta Implementasinya dengan OSS-R"*. Departemen Matematika Fakultas Sains dan Teknologi, Universitas Airlangga, 2010.

[9] Dean Fathony Alfatwa. *"Watermarking pada Citra Digital menggunakan Discrete Wavelet Transform"*. Institut Teknologi Bandung. Program Studi Teknik Informatika.

[10] Jaison Bennet, Chilambuchelvan Arul Ganaprakasam, and Kannan Arputharaj. *"A Discrete Wavelet based Feature Extraction and Hybrid Classification Technique for Microarray Data Analysis"*. Anna University, Department of Computer Science and Engineering, 2014.

[11] Bagaskara, K., Adiwijaya, & Rohmawati, A., A. "*Implementasi Discrete Wavelet Transform untuk Prediksi Kandungan Kurkumin pada Temulawak dengan menggunakan Pendekatan Kalibrasi"*. School of Computing, Telkom University, 2016.

[12] Adiwijaya, Maharani, M., Dewi, B.K., Yulianto, F.A. and Purnama, B., 2013. Digital Image Compression using Graph Coloring Quantization based on Wavelet-SVD. In Journal of Physics: Conference Series (Vol. 423, No. 1, p. 012019). IOP Publishing.

[13] Waletta Dinda, Adiwijaya, dan Angelina Prima K. "*Analisis dan Implementasi Minimum-Redundancy-Maximum-Relevance (mRMR) Feature selection pada Klasifikasi Data"*. Departmen Teknik Informatika Institut Teknologi Telkom, Bandung, 2008.

[14] Chris Ding and Hanchuan Peng. "*Minimum Redundancy Feature Selection from Microarray Gene Expression Data"*. Computational Research Division and Life Science/Genomics Division, Laurence Berkeley National Laboratory, University of California, 2004.

[15] Adiwijaya, 2016, Matematika Diskrit dan Aplikasinya, Bandung: Alfabeta.

[16] Wisesty, U.N., Nasri, J., and Adiwijaya. 2016. Modified Backpropagation Algorithm for Polycystic Ovary Syndrome Detection Based on Ultrasound Images. In International Conference on Soft Computing and Data Mining (pp. 141-151). Springer, Cham.

[17] Mubarok, M. S., Widiastuti, K. C., & Adiwijaya, A. (2017). Implementasi Mutual Information Dan Naive Bayes Untuk Klasifikasi Data Microarray. *eProceedings of Engineering*, *4*(2).

[18] Mubarok, M.S., Adiwijaya and Aldhi, M.D., 2017. Aspect-based sentiment analysis to review products using Naïve Bayes. In AIP Conference Proceedings (Vol. 1867, No. 1, p. 020060). AIP Publishing.

[19] Putri, L., Mubarok, M., & Adiwijaya, A. (2017). Klasifikasi Sentimen Pada Ulasan Buku Berbahasa Inggris Menggunakan Information Gain Dan Naïve Bayes. *eProceedings of Engineering*, *4*(3)

[20] Ethem Alpaydin. 2014. *Intoduction to Machine Learning Third Edition*. The MIT Press Cambridge, Massachusetts London, England.