Jurnal Teknik Elektro Vol. 8 No. 2 ISSN 1411 - 0059
Juli - Desember 2016

Klasifikasi Kanker Payudara Menggunakan Algoritma *Gain Ratio*

Balqis Aisyah Farahdiba[1] dan Yusuf Sulistyo Nugroho[2]

Program Studi Teknik Informatika Fakultas Komunikasi dan Informatika Universitas Muhammadiyah Surakarta
Jl. Ahmad Yani Tromol Pos I Pabelan, Surakarta, Indonesia
balqisaisyahf@gmail.com[1], yusuf.nugroho@ums.ac.id[2]

Abstrak— Kanker payudara merupakan jenis kanker yang menempati urutan kedua sebagai penyakit yang paling umum ditemui. Seperlima dari wanita penderita kanker adalah mereka yang didiagnosa mengidap kanker payudara. Kanker secara umum dibagi dua yaitu jinak dan ganas, begitupun kanker payudara. Pada status ganas, kanker dapat berakibat buruk bagi penderitanya bila terlambat diketahui. Oleh sebab itu, deteksi dini pada penyakit kanker sangatlah penting agar penderitanya dapat menerima penanganan yang tepat. Penelitian ini dilakukan dalam rangka untuk melakukan klasifikasi jenis kanker berdasarkan variabel-variabel yang mempengaruhi menggunakan teknik *data mining*. Klasifikasi kanker payudara dilakukan menggunakan metode *decision tree* dengan algoritma *gain ratio*. Atribut-atribut yang digunakan dalam klasifikasi yaitu ketebalan rumpun, keseragaman ukuran sel, keseragaman bentuk sel, adhesi marjinal, ukuran sel epitel tunggal, ukuran asli nuclei, kromatin, keadaan nucleoli normal dan mitosis. Hasil evaluasi kinerja algoritma *gain ratio* diperoleh nilai *recall, accuracy* dan *precision* masing-masing sebesar 92,55%, 95,17% dan 93,76%. Nilai tersebut menunjukkan bahwa algoritma *gain ratio* sangat baik digunakan dalam klasifikasi ini. Berdasarkan skema *decision tree*, variabel keseragaman ukuran sel merupakan variabel yang paling signifikan mempengaruhi jenis kanker.

Kata kunci— *data mining, decision tree, gain ratio,* kanker payudara, klasifikasi

I. PENDAHULUAN

Menurut data dari World Cancer Research Found International pada tahun 2012 telah ditemukan lebih dari 1,7 juta kasus kanker payudara. Kanker ini merupakan jenis penyakit kanker dengan penderita terbanyak kedua setelah kanker kulit. Selama dua dekade terakhir kasus kanker payudara lebih banyak ditemukan di negara-negara berkembang jika dibandingkan dengan negara-negara yang maju. Lebih dari 21% wanita penderita kanker adalah mereka yang terkena kanker payudara dan dan lebih dari 60% kematian akibat kanker payudara terjadi di negara berkembang [1].

Kanker atau sering disebut dengan tumor secara umum dibedakan menjadi dua macam yaitu jinak (benign) dan ganas (malignant). Pada level jinak, tumor akan memiliki kondisi dan perkembangan yang tidak bersifat kanker, dimana penyakit ini dapat terdeteksi namun tidak menyebar dan merusak jaringan lain di sekitarnya. Sedangkan pada level ganas, tumor akan menyebar dan merusak jaringan dan organ di sekitarnya [2]. Kanker ganas jika tidak cepat ditangani akan berdampak buruk bagi penderita sehingga deteksi dini pada penyakit kanker sangatlah penting bagi keselamatan penderita. Dengan mengetahui jinak atau ganasnya kanker yang diderita pasien sedini mungkin maka dapat segera ditangani dengan sesuai dan dapat diantisipasi karena bukan tidak mungkin dapat berujung pada kematian bagi penderita.

Jinak atau ganasnya kanker payudara dapat diketahui melalui beberapa variabel seperti ketebalan rumpun, keseragaman ukuran dan bentuk sel, adhesi marjinal, ukuran sel epitel tunggal, ukuran asli nuclei, kromatin, keadaan nucleoli normal serta mitosisnya. Berdasarkan variabel tersebut, dalam penelitian ini akan digunakan algoritma *gain ratio* untuk mengklasifikasikan kanker payudara menjadi dua kelas yaitu kanker payudara jinak dan ganas. *Data mining* merupakan proses penggalian informasi penting dan pola-pola yang tersembunyi dalam data dengan jumlah yang besar [3]. Salah satu algoritma dalam *data mining* yang dapat digunakan untuk pengklasifikasian adalah *gain ratio*. Penerapanan algoritma *gain ratio* pada *decision tree* (pohon keputusan) dapat meningkatkan akurasi perhitungan hingga 76% dengan waktu komputasi yang lebih singkat [4].

Klasifikasi kanker payudara sebelumnya juga telah dilakukan. Rashmi dkkmenganalisis kinerja algoritma naïve bayes dalam mengklasifikasi dan memprediksi kanker payudara [2]. Penelitian tersebut membandingkan luaran dari algoritma untuk melakukan klasifikasi dan prediksi yang menghasilkan nilai sama. Hal ini dilakukan untuk menemukan tingkat keberhasilan dan kesalahan dari algoritma yang digunakan. Hasil dari penelitian mengungkapkan bahwa kedua algoritma tersebut memilki tingkat kesuksesan dan kesalahan yang sama, yaitu tingkat kesuksesan berkisar 85-95% dan tingkat kesalahan 10-15%. Namun algoritma ini masih memiliki nilai akurasi yang kecil ketika data *training* yang digunaan sangat sedikit jumlahnya.

Shah dan Jivani membandingkan tingkat akurasi dan waktu eksekusi data dari tiga algoritma klasifikasi dalam *data mining* yaitu random forest, klasifikasi naïve bayes dan k-nearest neighbor dalam mengklasifikasi kanker payudara [5]. Penelitian tersebut menyimpulkan bahwa naïve bayes sebagai

algoritma paling unggul diantara ketiganya karena tingkat akurasi yang dimilikinya menjadi yang paling besar yaitu 95,9943% dengan waktu eksekusi data hanya 0,02 detik.

Perbedaan penelitian ini dengan kedua penelitian sebelumnya terletak pada algoritma *data mining* yang digunakan. Klasifikasi kanker payudara dalam penelitian ini menggunakan algoritma *gain ratio* untuk membangun *decision tree*. Sedangkan dalam mengevaluasi kinerjanya dilakukan dengan mencari nilai *recall, accuracy* dan *precision*.

II. METODOLOGI PENELITIAN

A. Penentuan Variabel

Seperti dalam penelitian yang dilakukan oleh [6], variabel yang digunakan terdiri dari dua jenis yaitu variabel bebas (disimbolkan dengan X) dan variabel terikat (disimbolkan dengan Y). Sementara itu, total data yang untuk proses klasifikasi kanker payudara adalah sebanyak 683 data dengan 9 variabel bebas dan 1 variabel terikat. Tabel I menunjukkan variabel-variabel yang digunakan dalam penelitian untuk mengklasifikasi jenis kanker.

TABEL I. VARIABEL DALAM KLASIFIKASI JENIS KANKER

Variabel	Nama Atribut
X1	Ketebalan Rumpun
X2	Keseragaman Ukuran Sel
X3	Keseragaman Bentuk Sel
X4	Adhesi Marjinal
X5	Ukuran Sel Epitel Tunggal
X6	Ukuran Asli *Nuclei*
X7	Kromatin
X8	Keadaan *Nucleoli* Normal
X9	Mitosis
Y	Kelas : (2 untuk Jinak, 4 untuk Ganas)

B. Data *Preprocessing*

Data *preprocessing* merupakan tahap yang dilakukan sebelum mengolah data. Tahap ini untuk mengubah data menjadi format yang sesuai dengan kebutuhan analisis. Perubahan data diperlukan karena pada saat menentukan *decision tree* nilai data yang real akan membuat pohon memiliki cabang yang rumit. Beberapa variabel yang digunakan mengandung data yang bernilai real maka data tersebut dikelompokkan berdasarkan skala tertentu. Hasilnya nilai tiap variabel dibagi menjadi tiga kelas yaitu 1-3, 4-6 dan 7-10.

C. Implementasi Algoritma *Gain Ratio*

Decision tree yang dibentuk didasarkan pada nilai atribut yang memiliki nilai *gain ratio* tertinggi untuk ditempatkan menjadi node pertama [7]. Untuk mencari nilai *gain ratio* digunakan rumus yang terdapat pada persamaan 2 sebagaimana yang dinyatakan pada [8].

$$gain\ ratio\ (S,A) = \frac{information\ gain\ (S,A)}{split\ information\ (S,A)} \quad (1)$$

Sementara itu, nilai *information gain* dan *split information* dapat dihitung dengan persamaan 2 dan 3.

$$information\ gain\ (Y,A) = entropy(Y) - \sum(|Y_v|/|Y|)\ entropy\ Y_v \quad (2)$$

$$split\ information\ (Y,A) = \sum_{i=1}^{c}(|Y_i|/|Y|)\ log_2(|Y_i|/|Y|) \quad (3)$$

$$entropy(Y) = \sum_{i} p\ (i|Y)\ log_2\ p(i|Y) \quad (4)$$

III. HASIL DAN PEMBAHASAN

A. Perhitungan Algoritma *Gain ratio*

1) Penentuan *Entropy* Y

Tahap pertama dalam perhitungan adalah mencari nilai *entropy* variabel terikat (Y) menggunakan persamaan 4.

$$\text{Entropy y } [444,239] = -\left(\frac{444}{683}.\log_2\frac{444}{683}\right) - \left(\frac{239}{683}.\log_2\frac{239}{683}\right) = 0{,}934003$$

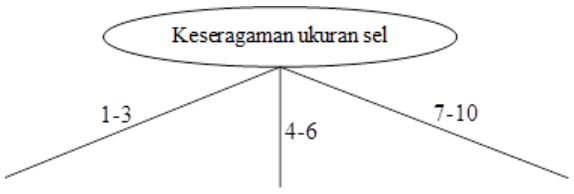
2) Menentukan *Root node*

Root node (simpul akar) ditentukan dengan menghitung *gain ratio* dari setiap variabel bebas (X) yang digunakan. Hasil perhitungan dengan menerapkan *gain ratio* ditunjukkan dalam Tabel II.

TABEL II. NILAI GAIN RATIO PADA *ROOT NODE*

Variabel	*Entropy* Total	*Information Gain*	*Split Information*	*Gain ratio*
X1	0,509762	0,424241	1,534277	0,276509
X2	0,360885	0,573118	1,203542	0,476193
X3	0,368947	0,565056	1,245212	0,453782
X4	0,542623	0,391380	1,071057	0,365414
X5	0,531231	0,402772	1,116546	0,360730
X6	0,400607	0,533596	1,169618	0,456213
X7	0,437029	0,496974	1,192445	0,416769
X8	0,549569	0,384434	1,050959	0,365793
X9	0,875598	0,058405	0,462500	0,126281

Tabel II menunjukkan bahwa X2 (keseragaman ukuran sel) memiliki nilai *gain ratio* tertinggi, sehingga X2 ditentukan sebagai simpul akar (*root node*). Gambar 1 menunjukkan *root node* yang terbentuk dalam *decision tree*.



Gambar 1. *Root node* pada *decision tree* untuk klasifikasi kanker payudara

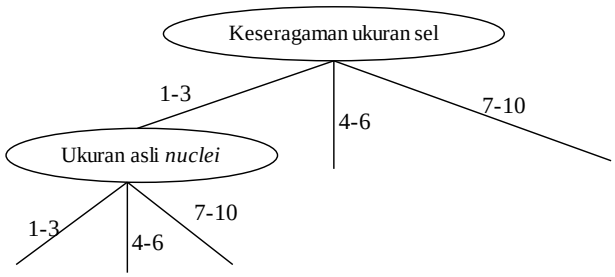
3) Penentuan *Internal Node* Cabang Kelas 1-3 pada Simpul Akar

Untuk mencari node pada cabang pertama dari simpul akar (kelas 1-3) adalah dengan mencari nilai *gain ratio* dari setiap variabel selain X2 yang telah menempati simpul akar. Tabel III menunjukkan hasil perhitungan *gain ratio* untuk menentukan simpul pada cabang pertama setelah simpul akar.

TABEL III. NILAI *GAIN RATIO* PENENTUAN *INTERNAL NODE CABANG 1 ROOT NODE*

Variabel	*Entropy* Total	*Informaion Gain*	*Split Information*	*Gain Ratio*
X1	0,246950	0,150714	1,180638	0,127654
X3	0,274316	0,123348	0,320109	0,385331
X4	0,316428	0,081236	0,328599	0,247219
X5	0,303167	0,094497	0,416032	0,227139
X6	0,176082	0,221582	0,551069	0,412982
X7	0,261728	0,135936	0,462345	0,294014
X8	0,259459	0,138205	0,383338	0,360530
X9	0,379511	0,018153	0,126390	0,143627

Berdasarkan Tabel III diperoleh X6 (ukuran asli nuclei) sebagai simpul dari cabang pertama (kelas 1-3) *root node* seperti yang ditunjukkan pada Gambar 2.

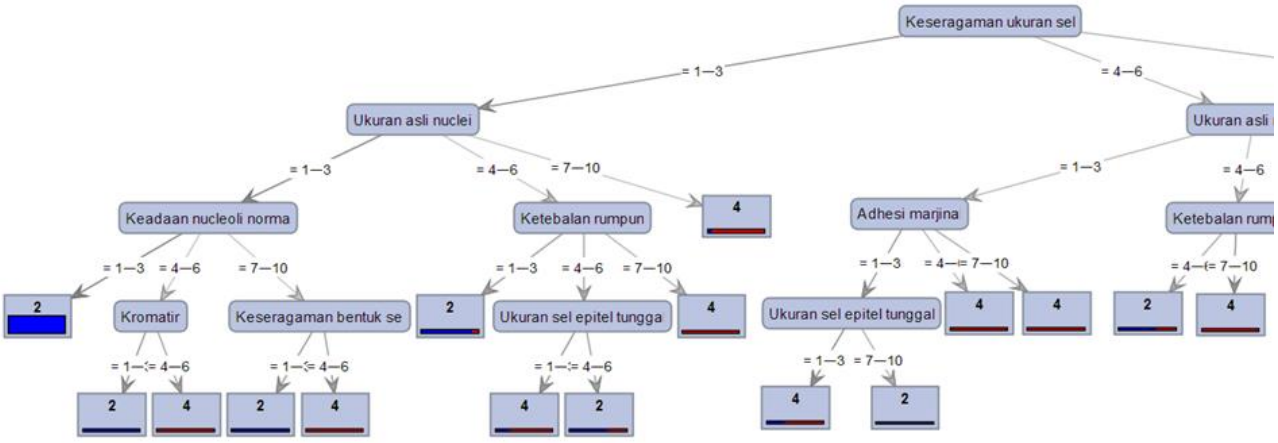


Gambar 2. *Internal node* cabang kelas 1-3 dari variabel X2

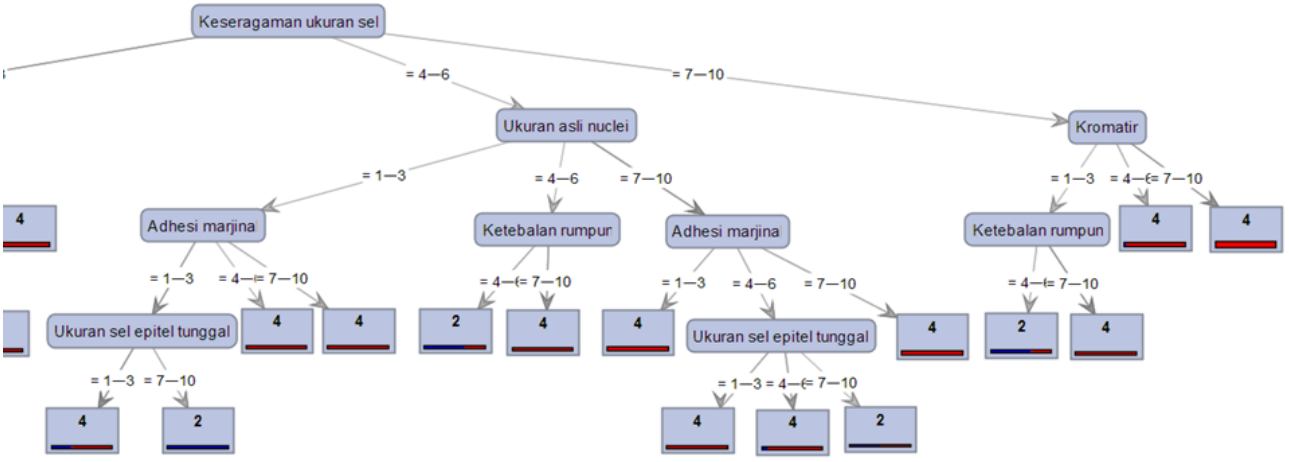
Perhitungan *gain ratio* dilakukan secara kontinyu hingga ditemukan simpul daun (*leaf node*) pada *node* level terendah untuk semua cabang *root node* dan *internal node*. Gambar 3 dan Gambar 4 merupakan potongan dari *decision tree* utuh yang terbentuk.

B. Penentuan *Decision tree*

Hasil perhitungan algoritma *gain ratio* secara lengkap dapat digambarkan dalam bentuk *decision tree* yang ditunjukkan pada Gambar 3 dan Gambar 4.



Gambar 3. *Decision tree* bagian 1



Gambar 4. *Decision tree* bagian 2

Berdasarkan Gambar 3 dan Gambar 4 dapat dilihat bahwa variabel keseragaman ukuran sel (X2) berada di posisi *root node* dalam pohon keputusan, sehingga dapat diinterpretasikan bahwa keseragaman ukuran sel merupakan variabel yang paling berpengaruh dalam menentukan klasifikasi kanker payudara.

C. Performa Algoritma

Kinerja dari algoritma *gain ratio* untuk klasifikasi kanker payudara dapat dilihat dari nilai *recall, accuracy* dan *precision* yang dimilikinya. Berdasarkan perhitungan nilai *recall, accuracy* dan *precision* diperoleh nilai sebesar 92,55%, 95,17% dan 93,76%.

IV. PENUTUP

Hasil klasifikasi kanker payudara menggunakan algoritma *gain ratio* ini dapat diambil kesimpulan bahwa performa algoritma yang diukur berdasarkan tingkat recall, accuracy dan precision yang masing-masing memiliki nilai 92,55%, 95,17% dan 93,76% menunjukkan bahwa algoritma *gain ratio* sangat baik digunakan dalam penelitian ini.

Sementara itu, berdasarkan *decision tree* yang dihasilkan dapat dilihat bahwa variabel keseragaman ukuran sel merupakan variabel yang paling berpengaruh dalam klasifikasi kanker payudara. Hal ini ditunjukkan dari variabel tersebut menempati sebagai simpul akar (*root node*).

REFERENSI

[1] Ma, J., & Jemal, A. (2013). Breast Cancer Statistics. Breast Cancer Metastasis and Drug Resistance, 143–159. http://doi.org/10.1007/978-1-4614-5647-6.

[2] Rashmi, G. D., Lekha, A., & Bawane, N. (2015). Analysis of Efficiency of Classification and Prediction Algorithms (Naïve Bayes) for Breast Cancer Dataset. In 2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT) (pp. 108–113). http://doi.org/10.1109@ERECT.2015.7498997.

[3] Demigha, S. (2015). *Data mining* for Breast Cancer Screening. 2015 10th International Conference on Computer Science & Education (ICCSE), (Iccse), 65–69. http://doi.org/10.1109/ICCSE.2015.7250219.

[4] Prasad, N., & Naidu, M. M. (2015). *Gain ratio* as Attribute Selection Measure in Elegant *Decision tree* to Predict Precipitation Narasimha. In Proceedings - 8th EUROSIM Congress on Modelling and Simulation, EUROSIM 2013 (pp. 141–150). http://doi.org/10.1109/EUROSIM.2013.35.

[5] Shah, C., & Jivani, A. G. (2013). Comparison of *Data mining* Classification Algorithms for Breast Cancer Prediction. 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), 1–4. http://doi.org/10.1109/ICCCNT.2013.6726477

[6] Nugroho, Y. S., & Gunawan, D. (2016). *Decision tree* Induction for Classifying the Cholesterol Levels. The 2nd International Conference on Science, Technology, and Humanity, 231–240.

[7] Nagpal, A., & Gaur, D. (2015). A New Proposed Feature Subset Selection Algorithm Based on Maximization of *Gain ratio*. 4th International Conference, DBA 2015, 181–197. http://doi.org/10.1007/978-3-319-03689-2_13.

[8] *Kotsiantis, S. B. (2013). Decision trees: A Recent Overview. Artificial Intelligence Review, 39(4), 261–283. http://doi.org/10.1007/s10462-011-9272-4.*