PENERAPAN SISTEM DATA MINING UNTUK DIAGNOSIS PENYAKIT KANKER PAYUDARA MENGGUNAKAN CLASSIFICATION BASED ON ASSOCIATION ALGORITHM

Herwanto[1)], Aniati Murni Arymurthy[1)]
Fakultas Ilmu Komputer, Universitas Indonesia
herwanto_p@yahoo.com, aniati@cs.ui.ac.id

ABSTRAK

Aplikasi system data mining untuk mengidentifikasi atribut-atribut penting yang berguna membantu pengambilan keputusan dari basis data rumah sakit akan dibahas dalam paper ini. Data-data medis pasien yang beresiko menderita penyakit kanker payudara dimasukkan kedalam data warehouse.

Metodologi model klasifikasi didasarkan pada tiga tahapan, yaitu a) menangani data yang tidak lengkap melalui ekstraksi, b) merubah data yang bernilai kontinyu menjadi data yang bernilai diskrit serta c) rule mining dan klasifikasi. Algoritma yang digunakan untuk proses data mining adalah Clasification based on Predictive Assiciation Rule(CPAR). Pada tahapan diskritisasi, terdapat masalah yang dikenal dengan istilah "sharp boundary". Paper ini mengusulkan proses optimalisasi menggunakan soft discretization, dimana fuzzy logic digunakan untuk mempartisi data.

Ada 2.767 pasien yang terpilih, masing-masing diambil 8 atribut: sex, umur dan hasil pemeriksaan laboratorium yaitu Hemoglobin (HB), Lekosit (Leko), Trombosit (Tromb), Hemotokrit (HCT), Red blood cell distribution width (RDW) dan RDW-SD. Tingkat akurasi maksimum untuk positif kanker payudara adalah 67% dan negatif kanker payudara 97%.

Kata kunci: *data mining, discretization, classification*, algoritma *Classification Predictive Association Rule, fuzzy logic*

1. PENDAHULUAN

Kanker payudara merupakan penyakit kronis dan untuk penyembuhan secara total masih sangat diragukan, dan memerlukan jangka waktu pengobatan yang lama serta biaya yang tinggi. Kanker payudara dapat didiagnosa dengan berbagai cara pemeriksaan antara lain dengan pemeriksaan mammografi yaitu suatu teknik pemeriksaan foto rontgen untuk jaringan lunak, yang terbukti efektif untuk memberikan suatu petunjuk adanya kelainan pada payudara. Akan tetapi gambar yang dihasilkan oleh mammografy tidak selalu cukup untuk menentukan keberadaan tumor ganas atau tidak. Penyakit ini oleh World Health Organization (WHO) dimasukkan dalam International Classification of Diseases (ICD) dengan nomor kode 174 (Tjahyadi, dkk, 1996).

Sampai saat ini faktor risiko utama penyakit kanker payudara belum dapat diketahui secara pasti. Penyebab kanker payudara adalah multifaktorial, artinya bahwa berbagai macam faktor yang saling berkaitan dan berakumulasi dapat menyebabkan risiko terkena penyakit tersebut. Faktor-faktor tersebut dapat berupa (a) faktor internal, yaitu faktor genetik (adanya riwayat keluarga kanker) dan ketidakseimbangan hormon endogenous (estrogen, progesteron, androgen, prolaktin); (b) faktor eksternal, yaitu faktor onkologi (virus, diet, obesitas, glucosa intolerance) dan kondisi lingkungan seperti pemakaian estrogen, rokok, karsinogen kimia yang berasal dari makanan (zat tambahan makanan), air minum, dan udara yang mungkin secara keseluruhan ikut berperan sebagai faktor risiko (Vorherr H., 1980).

Faktor-faktor lain yang berhubungan dengan kanker payudara bisa ditambahkan untuk screening guna meningkatkan kemungkinan deteksi dini kanker payudara salah satunya adalah hasil pemeriksaan patologi klinik. Dalam paper ini penulis akan menggali apakah hasil pemeriksaan laboratorium patologi klinik bisa berkontribusi dalam membantu diagnose penyakit kanker payudara.

Tujuan dari penelitian yang akan dilakukan adalah untuk mendapatkan classifier yang bisa membedakan apakah pasien terkena tumor jinak atau ganas berdasarkan dua sumber data yaitu,

data historis serta data klinis pasien. Dengan classifier yang terbentuk diharapkan bisa membantu dan memudahkan dokter dalam melakukan klasifikasi sehingga bisa menghindari biopsy yang tidak perlu.

2. TINJAUAN PUSTAKA

2.1. *Data Mining*

Data mining berkaitan dengan mencari pola dan relasi-relasi yang tersembunyi dalam sejumlah data yang besar dengan tujuan untuk prediksi. Ada dua jenis data mining, yaitu *directed data mining* dan *undirected data mining*. *Directed data mining* digunakan jika sudah diketahui secara pasti apa yang akan diprediksi, sehingga bisa secara langsung menambang data untuk diarahkan pada tujuan tertentu. Misalnya model prediksi yang digunakan untuk membuat prediksi tentang sesuatu yang tidak diketahui. Model prediksi menggunakan pengalaman untuk menentukan bobot dan tingkat kepercayaan. Salah satu kunci keberhasilan adalah adanya data yang cukup dengan hasil yang sudah diketahui untuk mengarahkan/melatih model.

Undirected data mining berkaitan dengan menelusuri pola-pola baru dalam data. Tidak seperti *directed data mining*, yang sudah mengetahui apa yang akan diprediksi. Pada *undirected data mining*, ingin diketahui bagaimana model memajukan/mengusulkan jawaban. Dalam prakteknya *data mining* sering berisi kombinasi dari keduanya. Misalnya saat membangun *predictive model*, sering berguna untuk mencari pola dalam data menggunakan teknik *undirected*.

2.1 Klasifikasi

Klasifikasi merupakan teknik data mining yang sangat umum digunakan. Contoh aplikasi klasifikasi adalah pengenalan pola, diagnosis penyakit dan lain-lain. Klasifikasi bisa didefinisikan sebagai memetakan data kedalam kelas-kelas. Kelas-kelas sebelumnya sudah ditentukan. Setiap record dalam database dimasukkan tepat ke hanya satu kelas.

Klasifikasi data dilakukan dengan dua tahapan. Pada tahap pertama, model dibentuk dengan menentukan kelas-kelas data. Model dibentuk dengan menganalisa *database tuples* yang dinyatakan dengan atribut. Setiap *tuple* dianggap memiliki kelas tertentu, seperti yang dinyatakan oleh salah satu atributnya yang disebut *class label atribute*.

Pada tahap kedua, model digunakan untuk klasifikasi. Pertama, akurasi model prediksi (atau *classifier*) di estimasi, menggunakan *data test*. *Sample* ini secara acak dipilih dan *independent* dengan *training sample*. Akurasi dari model pada *test set* adalah prosentase dari *sample test set* yang diklasifikasikan oleh model dengan benar. Untuk setiap *sample test*, label kelas yang telah diketehui dibandingkan dengan model kelas prediksi yang telah dilatih untuk *sample* tersebut. Jika akurasi dari model bisa diterima, maka model bisa digunakan untuk mengklasifikasikan *data tuples* dimana label kelasnya tidak diketahui. (Han & Kamber, 2001).

2.3. *Association Rule*

Association rule merupakan salah satu teknik *data mining* yang paling banyak digunakan dalam penelusuran pola pada sistem pembelajaran *unsupervised*. Metodologi ini akan mengambil seluruh kemungkinan pola-pola yang diamati dalam basis data. *Association rule* menjelaskan kejadian-kejadian yang sering muncul dalam suatu kelompok.

Bentuk umum *association rule* adalah $A1,A2,\ldots,An \rightarrow B1,B2,\ldots,Bm$

Yang berarti jika item Ai muncul, item Bj juga muncul dengan peluang tertentu.

Misalkan X adalah *itemset*. transaksi T dikatakan mengandung X jika dan hanya jika $X \subseteq T$. Rule $X \Rightarrow Y$ menyatakan himpunan transaksi basis data dengan tingkat kepercayaan (*confidence*) C, jika C % dari transaksi dalam D yang mengandung X juga mengandung Y. *Rule* $X \Rightarrow Y$ mempunyai *support* dalam transaksi *set* D jika s% dari transaksi dalam basis data berisi XUY. Tingkat kepercayaan menunjukkan kekuatan implikasi, dan *support* menunjukkan seringnya pola terjadi dalam *rule*.

Contoh : $A,B \Rightarrow C$ (S=0.01, C=0.8)

80% dari semua pelanggan yang membeli A dan B juga membeli C

1% dari semua pelanggan membeli ketiga item tersebut.

Problem dari *mining association rule* bisa dipecah dalam dua tahap.

1. Mencari semua *association rule* yang mempunyai *minimum support* S_{min} dan *minimum confidence* C_{min}.
   Itemset dikatakan sering muncul (*frequent*) jika *Support*(A) $\geq S_{min}$
2. Menggunakan *itemset* yang besar untuk menentukan *association rule* untuk basis data yang mempunyai tingkat kepercayaan C di atas nilai minimum yang telah ditentukan.

2.4. *Classification-Based Association*

Saat ini, teknik *data mining* telah dikembangkan antara lain dengan menerapkan konsep yang digunakan dalam *association rule mining* untuk masalah klasifikasi. Ada beberapa metode yang bisa digunakan, antara lain *association rule clustering system* (ARCS) dan

associative classification. Metode ARCS me-*mining association rule* didasarkan pada *clustering* kemudian menggunakan *rules* untuk klasifikasi. ARCS, me-*mining association rule* dalam bentuk $A_{quant1} \wedge A_{quant2} \Rightarrow A_{cat}$, dimana bentuk A_{quant1} dan A_{quant2} adalah *data test* yang atributnya punya rentang nilai, A_{cat} menunjukkan label kelas untuk *catagorical atribute* yang diberikan dari *training data*.

Metode *associative classification* me-*mining rules* dalam bentuk *condset*⇒*y*, dimana *condset* adalah sekumpulan item dan *y* adalah label kelas. *Rules* yang sesuai dengan *minimum support* tertentu disebut *frequent*. *Rule* mempunyai support s jika s% dari *sample* dalam *data set* yang mengandung *condset* dan memiliki kelas *y*. *Rule* yang sesuai dengan *minimum confidence* disebut *accurate*. *Rule* mempunyai *confidence* c jika c% dari *sample* dalam *data set* yang mengandung *condset* memiliki kelas *y*. Jika beberapa rule mempunyai *condset* yang sama, maka *rule* dengan *confidence* tertinggi dipilih sebagai *possible rule* (PR). Metode ini menggunakan algoritma *association rule*, seperti algoritma *Appriori* untuk menghasilkan *association rule*, kemudian memilih sekelompok *rule* yang mempunyai kualitas tinggi dan menggunakan *rules* tersebut untuk memprediksi data. Tetapi *associative classification* masih kurang efesien karena sering kali menghasilkan rules dalam jumlah yang besar (Yin & Han, 2003).

Metode *classification-based association* lainnya adalah CPAR (*Classification based on Predictive Association Rule*). Algoritma ini mengambil ide dari FOIL (*First Order Inductive Leaner*) dalam menghasilkan *rule* dan mengintegrasikannya dengan *associative classification*.

2.5. *Classification based on Predictive Association Rules* (CPAR)

Ide dasar CPAR berasal dari FOIL yang menggunakan algoritma greedy untuk mempelajari *rules* yang membedakan contoh positip dengan contoh negatif. FOIL secara berulang mencari *rule* terbaik dan memindahkan seluruh contoh positip yang dicakup oleh *rule* sampai seluruh contoh positip dalam *data set* tercakup. Algoritma FOIL diperlihatkan pada Gambar 1.

Masukan: *Training set* D = P ∪ N. (P dan N adalah himpunan contoh positip dan contoh negatip)
Keluaran: Himpunan *rule* untuk memprediksi label kelas dari contoh.

```
Procedure FOIL
  rule set R ← Φ
  while |P| > 0
    N' ← N, P' ← P
    rule r ← empty_rule
    while |N| > 0 and r.length < max_rule_length
      find the literal p that brings most gain according to P' and N'
      append p to r
      remove from P' all examples not satisfying r
      remove from N' all examples not satisfying r
    end
    R ← R ∪ {r}
    Remove from p all examples satisfying r's body
  end
  return R
```

Gambar 1. Algoritma FOIL

Definisi

- *Literal* p adalah pasangan nilai atribut, dalam bentuk (A_i,v), dimana A_i adalah atribut dan v adalah nilai. *Tuple* t memenuhi literal $p=(A_i,v)$ jika dan hanya jika $t_i = v$, dimana t_i adalah nilai atribut ke i.
- *Rule* r, dalam bentuk "$p_1 \wedge p_2 \wedge \ldots \wedge p_t \Rightarrow c$," *Tuple* t memenuhi rule *r's body* jika dan hanya jika *tuple* tersebut memenuhi setiap *literal* dalam *rule*. Jika t memenuhi *r's body*, r memprediksi bahwa t adalah dari kelas c.

Pada saat memilih *literals*, FOIL Gain digunakan untuk mengukur informasi yang diperoleh dari penambahan *literal* tersebut ke *current rule*. Misalkan terdapat contoh positip |P| dan contoh negatip |N| memenuhi *current rule r's body*. Setelah *literal* p ditambahkan ke r, terdapat |P*| contoh positip dan |N*| contoh negatiip yang memenuhi *rule's body* yang baru. FOIL gain p didefinisikan sebagai,

$$Gain(p) = |P^*| \left[\log \frac{|P^*|}{|P^*|+|N^*|} - \log \frac{|P|}{|P|+|N|} \right]$$

dimana |P| dan |N| adalah jumlah contoh positip dan jumlah contoh negatip yang memenuhi *current rule r's body*. |P*| dan |N*| adalah jumlah contoh positip dan jumlah contoh negatip yang memenuhi *rule's body* yang baru, yang dihasilkan dengan menambahkan p ke r.

Predictive Rule Mining (PRM)

PRM adalah suatu algoritma yang memodifikasi FOIL untuk mendapatkan akurasi dan efesiensi yang lebih baik. Pada PRM, setelah contoh yang benar tercakup dalam rule, selain mengeluarkannya, bobotnya dikurangi dengan faktor perkalian. Algoritma PRM diperlihatkan pada Gambar 2.

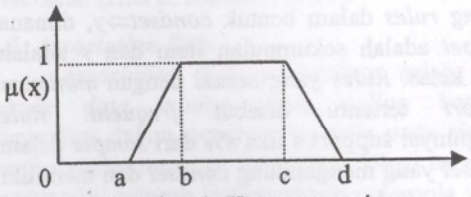
Masukan: *Training set* D = P ∪ N. (P dan N adalah himpunan contoh positip dan contoh negatip)
Keluaran: Himpunan *rule* untuk memprediksi label kelas dari contoh.
Procedure Predictive Rule Mining
Set the weight of every example to 1
Rule set R ← ϕ
Totalweight ← TotalWeight(P)
A ← Compute PNArray from D
While TotalWeight(P) > δ. totalWeight
N' ← N, P'← P, A' ← A
Rule r ← emptyrule
While true
Find best literal p according to A'
If gain(p) < min_gain then break
Append p to r
For each example t in P' ∪ N' not satifying r's body
Remove t from P' or N'
Change A' according to the removal of t
End
End
R ← R ∪ {r}
For each example t in P satisfying r's body
t.weight ← α.t.weight
change A according to the weight decreased
end
end
return R
Gambar 2. Algoritma PRM

Membuat *Rule* Dalam CPAR

Didalam PRM, setiap *rule* di-*generate* dari *dataset* yang tersisa, memilih hanya *literal* yang terbaik dan mengabaikan seluruh *literal* lainnya. CPAR membuat *rule* s dengan menambahkan *literals* satu per satu, yang mirip dengan PRM. Pada CPAR setelah menemukan *literal* terbaik p, *literal* lainnya q yang *gain*-nya mirip dengan p (misalnya hanya berbeda 1%) akan dicari. Selain terus membangun *rule* dengan menambahkan p ke r, q juga ditambahkan ke *current rule* r untuk membuat *rule* r' baru

2.6. Himpunan Fuzy

Pada himpunan fuzzy, nilai kebenaran suatu item tidak hanya bernilai benar atau salah, tetapi masih ada nilai-nilai yang terletak diantara benar dan salah. Misalnya pada atribut usia:
Muda: Usia < 35 Tahun
Setengah Baya:35 Tahun ≤ Usia ≤ 55 Tahun
Tua : Usia > 55 Tahun
Fungsi keanggotaan adalah suatu kurva yang menunjukkan pemetaan titik-titik input data kedalam nilai keanggotaannya (derajat keanggotaan) yang memiliki interval antara 0 sampai 1. Dengan menggunakan himpunan fuzzy, bisa dibuat suatu fungsi keanggotaan yang bersifat kontinyu. Misalnya orang dengan umur 50 tahun sudah mendekati tua, artinya mendekati tua dengan derajat keanggotaan μ=0.75.
Ada beberapa cara untuk menentukan fungsi keanggotaan, antara lain dengan menggunakan representasi segitiga, trapesium, Bel dan Gaussian. Representasi kurva trapezium ditunjukkan pada gambar 2.



Gambar 3. Kurva trapesium

$$\mu(x)\begin{cases}0, & x \le a \text{ atau } x \ge d\\ (x-a)/(b-a), & a \le x \le b\\ 1, & b \le x \le c\\ (d-x)/(d-c), & c \le x \le d\end{cases}$$

2.7. Fuzzy Association Rules

Pembentukan association rule untuk data-data kuantitatif bisa dilakukan dengan cara mempartisi data. Tetapi untuk data yang berada dekat batas mungkin saja diabaikan dalam proses data mining.

Dalam himpunan fuzzy, satu elemen bisa milik satu atau lebih himpunan, masing-masingnya dengan fungsi keanggotaan yang berbeda. Misalnya data pada table 1, fuzzy association rule bisa berbentuk, "O-Age & H-Rdw ⇒. Breast Cancer".

Tabel 1: Basisdata (D") dengan elemen-elemen fuzzy.

D"	Y-Age	M-Age	O-Age	N-Rdw	H-Rdw	BC
ID1	0.8	0.4	0.1	0.1	0.9	1
ID2	0.9	0.3	0	1	0	0
ID3	1	0	0	1	0	0
ID4	0.1	0.2	0.8	0	1	1

3. METODOLOGI

Sampel data hasil pemeriksaan laboratorium patologi klinik bersumber dari basis data dalam Sistem Informasi Manajemen Rumah Sakit X. Periode pengambilan data dari tanggal 1 Januari 2007 sampai dengan 31 Desember 2009. Data tersebut meliputi:

1. Data pasien yang beresiko menderita penyakit kanker payudara berjumlah 2.767 pasien. Pasien yang dipilih adalah pasien dengan minimal satu kali kunjungannya didiagnosa kanker payudara.
2. Data hasil pemeriksaan laboratorium yang dilakukan pada pasien selama periode 1 Januari 2007 sampai dengan 31 Desember

2009 baik yang berasal dari rawat jalan maupun rawat inap.

Persyaratan catatan medis yang dijadikan sampel mengacu pada *international classification of deseases tenth revision* (*ICD 10*) dimana penyakit kanker payudara diberi kode C50 *Malignant Neoplasm of Breast*, D05 *Carcinoma In Situ of Breast*, D05.9 *Carcinoma In Situ of Breast Unspecified*, D24 *Benign Neoplasm of Breast* dan N60.0 *Solitary Cyst of Breast*.

Dari 2.767 pasien yang terpilih didapat 211.694 hasil laboratorium. Untuk membentuk data *training* dan *testing*, diambil 6 jenis data pemeriksaan hasil laboratorium yang paling banyak dilakukan yaitu Hemoglobin (HB), Lekosit (Leko), Trombisit (Tromb), Hemotokrit (HCT), RDW dan RDW-SD. Terdapat 41.958 *record* hasil pemeriksaan laboratorium yang memenuhi kriteria tersebut.

Metodologi model klasifikasi didasarkan pada tiga tahapan, yaitu a) menangani data yang tidak lengkap melalui ekstraksi, b) merubah data yang bernilai kontinyu menjadi data yang bernilai diskrit serta c*) rule mining* dan klasifikasi. Pada tahap pertama, pemrosesan awal data dilakukan untuk menghapus data yang tidak lengkap dan mengekstrak data yang akan digunakan untuk mengelompokkan penyakit kanker payudara atau bukan. Pada tahap kedua dilakukan transformasi data kontinyu menjadi data bernilai diskrit. Pada tahap ketiga, algoritme *association rule mining* digunakan untuk menghasilkan aturan-aturan, yang berguna untuk mendeteksi penyakit kanker payudara atau bukan. Menurut Yin & Han (2003) algoritme yang efektif untuk digunakan dalam masalah klasifikasi adalah CPAR.

Pada algoritme CPAR proses dimulai dengan membaca data berupa sekumpulan bilangan *array* dua dimensi yang setiap kolomnya diberi atribut A dan atribut terakhir menunjukkan kelas. Data masukan selanjutnya dikelompokkan menjadi contoh positif P dan contoh negative N sesuai dengan kelasnya. Bobot contoh positif |P| dan bobot contoh negative |N| setiap atribut dijumlahkan untuk membentuk PN *array*, berupa *array* dua dimensi berisi daftar semua atribut, bobot contoh positif, dan bobot contoh negative. *Total weight threshold* (TWT) dihitung dengan mengalikan jumlah bobot positif dengan konstanta yang selama percobaan ditetapkan sama dengan 0.05.

Proses pembentukan aturan dilakukan berulang-ulang sampai jumlah bobot contoh positif lebih kecil dari TWT. Pada setiap proses dilakukan penyalinan P, N, A dan PN ke P', N', A' dan PN'. Menghitung *Gain* dan menyisipkan aturan ke *rule list*. Pada percobaan yang dilakukan konstanta minimum *gain* adalah 0.7, dan *decay factor* 1/3.

4. HASIL DAN PEMBAHASAN

4.1. Pembangunan Model Data Mining

Pembangunan model data mining dengan algoritma CPAR. Pada tahap awal data yang berasal dari relasi antara tabel master pasien, hasil laboratorium dan catatan medis ditransformasi ke dalam *working database* sampel data, selanjutnya diubah ke dalam bentuk kategori. Penentuan pembentukan nilai kategori terhadap data-data hasil laboratorium didasarkan atas referensi yang ada di SIM RSPP seperti diperlihatkan pada Tabel 1. Penentuan kelas positif kanker payudara atau negatif kanker payudara ditentukan oleh diagnosa yang terdapat pada tabel catatan medis. Kelas ditetapkan sebagai positif kanker payudara jika kode ICD pada diagnosa C50, D05, D05.9 atau D24. Kategori yang dibentuk dari atribut yang ada adalah sebanyak 28 kategori, seperti diperlihatkan pada Tabel 2

Tabel 2 Nilai referensi hasil laboratorium

Pemerik-saan	Keterangan	Satuan	Nilai Normal
HB	Hemoglobin	g/dl	12.0 – 14.0
Leko	Lekosit	Ribu/uL	5.0 – 10.0
Tromb	Trombosit	Ribu/uL	150 - 450
HCT	Hemotokrit	%	37 - 43
RDW	RDW	fL	11.5 – 14.5
RDW-SD	RDW-SD	fL	36.4 – 46.3

Tabel 3. Nilai kontinyu dan nilai diskrit

Atribut	Keterangan	Nilai Kontinyu	Nilai Diskrit
Umur	Umur Pasien	Umur < 20	1
Umur	Umur Pasien	20 <= Umur <= 40	2
Umur	Umur Pasien	Umur > 40	3
Sex	Jenis Kelamin	Sex = Laki-Laki	4
Sex	Jenis Kelamin	Sex = Perempuan	5
GDarah	Golongan Darah	A	6
GDarah	Golongan Darah	B	7
GDarah	Golongan Darah	O	8
GDarah	Golongan Darah	AB	9
GDarah	Golongan Darah	Lain-Lain	10
HB	Hemoglobin	HB<12	11
HB	Hemoglobin	12 <=HB <= 14	12
HB	Hemoglobin	HB>14	13
Leko	Lekosit	Leko < 5	14

Leko	Lekosit	5 <= Leko <= 10	15
Leko	Lekosit	Leko > 10	16
Tromb	Trombosit	Tromb < 150	17
Tromb	Trombosit	150 <= Tromb <=450	18
Tromb	Trombosit	Tromb > 450	19
HCT	Hemotokrit	HCT < 37	20
HCT	Hemotokrit	37 <= HCT <= 43	21
HCT	Hemotokrit	HCT > 43	22
RDW	RDW	RDW < 11.5	23
RDW	RDW	11.5<= RDW <= 14.5	24
RDW	RDW	RDW > 14.5	25
RDW-SD	RDW-SD	RDW-SD < 36.4	26
RDW-SD	RDW-SD	36.4<=RDW-SD<=46.3	27
RDW-SD	RDW-SD	RDW-SD> 46.3	28

Data yang sudah dalam bentuk diskrit selanjutnya ditransformasi kedalam bentuk array dua dimensi dan dilakukan proses *cleaning* dengan menghapus baris-baris yang tidak lengkap. Pada proses pembuatan model data mining sebagai contoh dipilih 20 baris sebagai data *training*.

Sampel data

03 04 06 11 15 17 20 25 28 30
03 05 07 11 15 17 20 25 28 30
03 04 06 11 14 18 20 25 27 30
03 04 09 11 16 18 20 25 28 30
01 04 07 11 16 17 20 23 26 31
02 05 07 11 14 17 20 23 27 31
01 05 07 12 16 18 20 24 27 31
03 05 08 11 16 18 20 24 27 31
03 04 06 12 14 18 21 25 27 30
03 04 06 12 14 18 21 25 27 30
03 04 08 12 15 18 21 25 28 30
03 04 06 12 15 18 21 25 28 30
03 04 06 11 15 18 21 25 28 30
02 05 07 12 14 18 21 23 27 31
02 05 07 12 14 17 21 23 27 31
02 05 07 12 14 17 21 23 27 31
02 05 07 12 15 18 21 24 27 31
03 04 07 12 15 18 21 24 27 31
02 04 08 12 14 18 21 24 27 31
01 05 06 13 15 18 22 24 27 31

Pembentukan sampel positip dan sampel negatip

Kolom terakhir pada sampel data menunjukkan kelas. Angka 30 menunjukkan kelas positip kanker payudara dan 31 menunjukkan kelas negatip kanker payudara.

Data *training* kemudian dipisahkan menjadi sampel positip (data pasien dengan diagnosa positip kanker payudara) dan sampel negatip (data pasien dengan diagnosa negatip kanker payudara).

Sampel positip

[03 04 06 12 14 18 21 25 27 30] [1.0]
[03 04 06 11 15 17 20 25 28 30] [1.0]
[03 04 06 12 14 18 21 25 27 30] [1.0]
[03 05 07 11 15 17 20 25 28 30] [1.0]
[03 04 08 12 15 18 21 25 28 30] [1.0]
[03 04 06 12 15 18 21 25 28 30] [1.0]
[03 04 06 11 15 18 21 25 28 30] [1.0]
[03 04 06 11 14 18 20 25 27 30] [1.0]
[03 04 09 11 16 18 20 25 28 30] [1.0]

Sampel negatip

[01 04 07 11 16 17 20 23 26 31] [1.0]
[02 05 07 11 14 17 20 23 27 31] [1.0]
[01 05 07 12 16 18 20 24 27 31] [1.0]
[0 3 05 08 11 16 18 20 24 27 31] [1.0]
[02 05 07 12 14 18 21 23 27 31] [1.0]
[02 05 07 12 14 17 21 23 27 31] [1.0]
[02 05 07 12 14 17 21 23 27 31] [1.0]
[02 05 07 12 15 18 21 24 27 31] [1.0]
[03 04 07 12 15 18 21 24 27 31] [1.0]
[02 04 08 12 14 18 21 24 27 31] [1.0]
[01 05 06 13 15 18 22 24 27 31] [1.0]

Pembentukan Gain

Pada awal proses setiap baris diberi bobot 1.0, dengan demikian angka 4 muncul 8 kali pada sampel positip (Wp) dan 3 kali pada sampel negatip (Wn), sehingga nilai gain pada atribut 4 adalah 3.14. Dalam penelitian ini *decay factor* ditentukan 0.3, *global minimum threshold* 0.7 .

Tabel 4. PN Array berisi informasi tentang atribut, bobot sampel positip, bobot sampel negatip dan Gain untuk setiap atribut.

Bobot total dari sampel positip pada awal proses adalah 9, dan *Total Weight Threshold* (TWT) dihitung berdasarkan rumus :

TWT = bobot total * total weight threshold

= 9 * 0.05 = 0.45

Sampel positip selanjutnya diproses sampai bobot totalnya lebih kecil dari 0.45. Tahap pertama adalah menghitung gain untuk setiap atribut. Gain terbesar adalah pada atribut 25 (gain = 2.37). *Local Gain Threshold* (LGT) dihitung berdasarkan rumus :

LGT = 2.37 * Gain_similarity_ratio

= 2.37 * 0.6

= 1.42

Tabel 4. PN Array ke 1

Atribut	Bobot Sampel Positip	Bobot Sampel Negatip	Gain
1	0	3	0.00
2	0	6	0.00
3	9	2	0.84
4	8	3	0.1
5	1	8	-0.8
6	6	1	0.77
7	1	8	-0.8
8	1	2	-0.38
9	1	0	0.26
10	0	0	0.00
11	5	3	-0.39
12	4	7	-1.38
13	0	1	0.00
14	3	5	-0.99
15	5	3	-0.39
16	1	3	-0.5
17	0	0	0.00
18	0	0	0.00
19	0	0	0.00
20	4	4	-0.79
21	5	6	-1.21
22	0	1	0.00
23	0	5	0.00
24	0	6	0.00
25	9	0	2.37
26	0	1	0.00
27	3	10	-1.59
28	6	0	1.58

Ada lima atribut yang nilai gain-nya diatas LGT yaitu 25 dan 28. Selanjutnya kedua atribut tersebut diproses satu persatu, yaitu dengan menyisipkan atribut 25 → 30 ke rule temporer, menyalin sampel positip dan sampel negatip dengan menghapus baris yang tidak berisi atribut 25.

Sampel positip
[03 04 06 12 14 18 21 25 27 30] [1.0]
[03 04 06 11 15 17 20 25 28 30] [1.0]
[03 04 06 12 14 18 21 25 27 30] [1.0]
[03 05 07 11 15 17 20 25 28 30] [1.0]
[03 04 08 12 15 18 21 25 28 30] [1.0]
[03 04 06 12 15 18 21 25 28 30] [1.0]
[03 04 06 11 15 18 21 25 28 30] [1.0]
[03 04 06 11 14 18 20 25 27 30] [1.0]
[03 04 09 11 16 18 20 25 28 30] [1.0]

Sampel negatip (kosong)

Karena sampel negatip kosong, maka atribut 25 → 30 disisipkan kedalam *rule list*. Bobot sampel positip selanjutnya direvisi dengan menggunakan decay factor.

Sampel positip
[03 04 06 12 14 18 21 25 27 30] [0.33]
[03 04 06 11 15 17 20 25 28 30] [0.33]
[03 04 06 12 14 18 21 25 27 30] [0.33]
[03 05 07 11 15 17 20 25 28 30] [0.33]
[03 04 08 12 15 18 21 25 28 30] [0.33]
[03 04 06 12 15 18 21 25 28 30] [0.33]
[03 04 06 11 15 18 21 25 28 30] [0.33]
[03 04 06 11 14 18 20 25 27 30] [0.33]
[03 04 09 11 16 18 20 25 28 30] [0.33]

Nilai bobot sampel positip sekarang adalah 2.97, masih lebih besar dari 0.45, sehingga atribut berikutnya yaitu 28 diproses dengan menyisipkan atribut 28 → 30 ke rule temporer dan menyalin sampel positip dan sampel negatip sebelumnya dengan menghapus baris yang tidak berisi atribut 28.

Sampel positip
[03 04 06 11 15 17 20 25 28 30] [1.0]
[03 05 07 11 15 17 20 25 28 30] [1.0]
[03 04 08 12 15 18 21 25 28 30] [1.0]
[03 04 06 12 15 18 21 25 28 30] [1.0]
[03 04 06 11 15 18 21 25 28 30] [1.0]
[03 04 09 11 16 18 20 25 28 30] [1.0]

Karena sampel negatip kosong, maka atribut 28 → 30 disisipkan kedalam rule list. Bobot sampel positip selanjutnya direvisi dengan menggunakan decay factor.

Sampel positip
[03 04 06 12 14 18 21 25 27 30] [0.33]
[03 04 06 11 15 17 20 25 28 30] [0.11]
[03 04 06 12 14 18 21 25 27 30] [0.33]
[03 05 07 11 15 17 20 25 28 30] [0.11]
[03 04 08 12 15 18 21 25 28 30] [0.11]
[03 04 06 12 15 18 21 25 28 30] [0.11]
[03 04 06 11 15 18 21 25 28 30] [0.11]
[03 04 06 11 14 18 20 25 27 30] [0.33]
[03 04 09 11 16 18 20 25 28 30] [0.11]

Setelah seluruh atribut diproses, langkah selanjutnya menghitung gain dengan bobot sampel yang sudah disesuaikan. Tabel 4 berisi informasi tentang atribut, bobot sampel positip, bobot sampel negatip dan Gain untuk setiap atribut yang telah disesuaikan.

Gain maksimum sekarang adalah 0.31 yang berarti lebih kecil dari global minimum 0.7, sehingga proses dihentikan. Cara yang sama digunakan untuk kelas 31. Pada akhir proses semua rule disisipkan kedalam rule list, dan dihitung akurasinya menggunakan Laplace accuracy sehingga didapat :

No.	Rule	L.A
1	25 → 30	0.91
2	28 → 30	0.88
3	7 → 31	0.75
4	5 → 31	0.80
5	23 → 31	0.86
6	24 → 31	0.83

Tabel 5. PN Array ke 2

Atribut	Bobot Sampel Positip	Bobot Sampel Negatip	Gain
1	0	3	0.00
2	0	6	0.00
3	1.65	2	0.01
4	1.54	3	-0.43
5	0.11	8	-0.38
6	1.32	1	0.31
7	0.11	8	-0.38
8	0.11	2	-0.24
9	0.11	0	0.09
10	0	0	0.00
11	0.77	3	-0.61
12	0.88	7	-1.23
13	0	1	0.00
14	0.99	5	-0.99
15	0.55	3	-0.59
16	0.11	3	-0.28
17	0	0	0.00
18	0	0	0.00
19	0	0	0.00
20	0.66	4	-0.76
21	0.99	6	-1.14
22	0	1	0.00
23	0	5	0.00
24	0	6	0.00
25	1.65	0	1.32
26	0	1	0.00
27	0.99	10	-1.59
28	0.66	0	0.53

4.2. Pelatihan dengan *Data Training*

Proses pelatihan model *data mining* menggunakan algoritme CPAR. Pelatihan dilakukan dengan mengambil sebanyak 3500 sampel, dengan perbandingan sampel positif kanker payudara dan negatif negatif kanker payudara 4 : 6. Sampel data memuat informasi tentang data input berupa umur, sex, hasil tes laboratorium, data *output* berupa diagnosa penyakit.

3500 sampel data yang mempunyai catatan medis lengkap dikumpulkan. Hasil proses *data mining* dari 3500 sampel data *training* dengan berbagai variasi *Gain similarity ratio* diperlihatkan pada Tabel 6 sampai dengan Tabel 8.

Tabel 6. Aturan yang dihasilkan dengan *Gain similarity ratio* 90%

No	Aturan	L.A
1	IF Sex= Perempuan Then Positif BC	0.57
2	IF RDW-SD >46.3 Then Positif BC	0.64
3	IF Umur > 40 Then Positif BC	0.52
4	IF RDW > 14.5 Then Positif BC	0.55
5	IF HB < 12 Then Positif BC	0.56
6	IF Sex= Laki-Laki Then Negatif BC	0.94
7	IF 11.5<=RDW<=14.5 Then Negatif BC	0.64
8	IF 36.4<=RDW-SD<=46.3 Then Negatif BC	0.61
9	If Umur < 20 Then Negatif BC	0.97

Tabel 7. Aturan yang dihasilkan dengan *Gain similarity ratio* 80%

No	Aturan	L.A
1	IF Sex= Perempuan Then Positif BC	0.57
2	IF RDW-SD >46.3 Then Positif BC	0.64
3	IF Umur > 40 Then Positif BC	0.52
4	IF RDW > 14.5 Then Positif BC	0.55
5	IF HB < 12 Then Positif BC	0.56
6	IF HCT < 37 Then Positif BC	0.55
7	IF Sex= Laki-Laki Then Negatif BC	0.94
8	IF 11.5<=RDW<=14.5 Then Negatif BC	0.64
9	IF 36.4<=RDW-SD<=46.3 Then Negatif BC	0.61
10	If Umur < 20 Then Negatif BC	0.97

Tabel 8. Aturan yang dihasilkan dengan *Gain similarity ratio* 50%

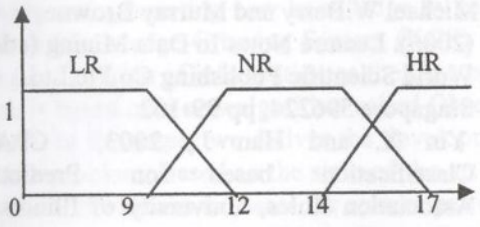
No	Aturan	L.A
1	IF Sex= Perempuan Then Positif BC	0.57
2	IF RDW-SD >46.3 Then Positif BC	0.64
3	IF Umur > 40 Then Positif BC	0.52
4	IF RDW > 14.5 Then Positif BC	0.55
5	IF HB < 12 Then Positif BC	0.56
6	IF HCT < 37 Then Positif BC	0.55
7	IF Sex= Laki-Laki Then Negatif BC	0.94
8	IF 11.5<=RDW<=14.5 Then Negatif BC	0.64
9	IF 36.4<=RDW-SD<=46.3 Then Negatif BC	0.61
10	If Umur < 20 Then Negatif BC	0.97

Data pada Tabel 6. memperlihatkan bahwa aturan IF RDW-SD >46.3 Then Positif BC mempunyai *Laplace Accuracy* (L.A) paling tinggi pada kelas positif kanker payudara yaitu 64%. Ini berarti bahwa pasien dengan hasil pemeriksaan RDW-SD lebih besar dari 46.3 mempunyai peluang terkena penyakit kanker payudara sebesar 64%. Aturan IF Umur < 20 Then Negatif BC mempunyai L.A paling tinggi pada kelas negatif kanker payudara yaitu 97%. Dengan menggunakan *Gain similarity ratio* 80%

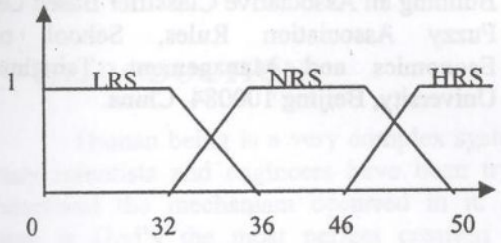
dihasilkan 10 aturan. Aturan tambahan adalah IF HCT < 37 Then Positif BC dengan LA 55%. Dengan menurunkan nilai *Gain similarity ratio* menjadi 50% didapat hasil yang sama dengan *Gain similarity ratio* 80%.

4.3. Proses Optimalisasi

Proses dilakukan dengan merubah kategori RDW dan RDW-SD yang menjadi penentu utama positif kanker payudara atau negatif kanker payudara. Proses optimalisasi terdiri dari empat tahap, yaitu: 1) merubah data input menjadi data fuzzy; 2) merubah variabel fuzzy menjadi nilai fuzzy menggunakan fungsi keanggotaan fuzzy logic seperti diperlihatkan pada gambar 30 dan 31; 3) merubah data menjadi array dua dimensi; 4) menerapkan algoritma CPAR. Tabel 9 adalah sampel data setelah proses optimalisasi. Aturan yang dihasilkan setelah proses optimalisasi dengan berbagai vaiasi *Gain similarity ratio* diperlihatkan pada Tabel 10, 11dan 12.



Gambar 4. Fungsi keanggotaan RDW



Gambar 5. Fungsi keanggotaan RDW-SD

Tabel 9. Contoh data fuzzy

RDW			RDW-SD			
L	N	H	L	N	H	Kelas
0	0.7	0.3	0	1	0	BC
0	0	1	0	0	1	BC
0	0.8	0.2	0	1	0	BC
0	0	1	0	0	1	BC
0	0.4	0.6	0	0.1	0.9	BC
0	0.7	0.3	0	0.8	0.3	BC
0	0.2	0.8	0	0.8	0.2	BC
0	0.7	0.3	0	1	0	BC
0	0.1	0.9	0	0	1	BC
0	0	1	0	0	1	BC
0	0	1	0	0	1	BC
0	0.8	0.2	0	1	0	BC
0	0.9	0.1	0	1	0	BC
0	0.4	0.6	0	0.1	0.9	BC
0	0.9	0.1	0	1	0	BC
0	0.5	0.5	0	0.6	0.4	Non BC
0	0.8	0.2	0	1	0	Non BC
0	1	0	0	1	0	Non BC
0	0.5	0.5	0	1	0	Non BC
0	0.7	0.3	0	1	0	Non BC
0	0.9	0.1	0	1	0	Non BC
0	0.1	0.9	0	1	0	Non BC
0	0.3	0.7	0	0.6	0.4	Non BC
0	0.4	0.6	0	0	1	Non BC
0	0.7	0.3	0	1	0	Non BC
0	0.9	0.1	0	1	0	Non BC
0	0.3	0.7	0	1	0	Non BC

Tabel 10. Aturan yang dihasilkan dengan *Gain similarity ratio* 90%

No	Aturan	L.A
1	IF Sex= Perempuan Then Positif BC	0.57
2	IF RDW-SD = High Then Positif BC	0.67
3	IF Umur > 40 Then Positif BC	0.52
4	IF RDW = High Then Positif BC	0.59
5	IF HB < 12 Then Positif BC	0.56
6	IF Sex= Laki-Laki Then Negatif BC	0.94
7	IF RDW = Normal Then Negatif BC	0.63
8	IF RDW-SD= Normal Then Negatif BC	0.59
9	If Umur < 20 Then Negatif BC	0.97

Tabel 11. Aturan yang dihasilkan dengan *Gain similarity ratio* 80%

No	Aturan	L.A
1	IF Sex= Perempuan Then Positif BC	0.57
2	IF RDW-SD = High Then Positif BC	0.67
3	IF Umur > 40 Then Positif BC	0.52
4	IF RDW = High Then Positif BC	0.59
5	IF HB < 12 Then Positif BC	0.56
6	IF HCT < 37 Then Positif BC	0.55
7	IF Sex= Laki-Laki Then Negatif BC	0.94
8	IF RDW = Normal Then Negatif BC	0.63
9	IF RDW-SD= Normal Then Negatif BC	0.59
10	If Umur < 20 Then Negatif BC	0.97

Tabel 12. Aturan yang dihasilkan dengan *Gain similarity ratio* 50%

No	Aturan	L.A
1	IF Sex= Perempuan Then Positif BC	0.57
2	IF RDW-SD = High Then Positif BC	0.67
3	IF Umur > 40 Then Positif BC	0.52
4	IF RDW = High Then Positif BC	0.59
5	IF HB < 12 Then Positif BC	0.56
6	IF HCT < 37 Then Positif BC	0.55
7	IF Sex= Laki-Laki Then Negatif BC	0.94
8	IF RDW = Normal Then Negatif BC	0.63
9	IF RDW-SD= Normal Then Negatif BC	0.59
10	If Umur < 20 Then Negatif BC	0.97

5. KESIMPULAN

Dari hasil yang diperoleh catatan sebagai kesimpulan dari penelitian ini yakni:

1. Pemeriksaan RDW-SD, RDW, HB dan HCT menjadi penentu utama untuk menentukan apakah pasien positif kanker payudara atau negatif kanker payudara.
2. Menurunkan nilai *Gain similarity ratio* dapat memunculkan aturan-aturan yang sebelumnya tersembunyi.
3. Algoritme CPAR hanya memilih nilai atribut yang memiliki nilai *Gain* terbaik, sehingga ada kemungkinan atribut yang mempunyai kekuatan prediksi yang tinggi tidak muncul dalam aturan.
4. Algoritme CPAR menerima input dalam bentuk kategori, sehingga proses penentuan data kontinyu menjadi data kategori sangat berpengaruh terhadap hasil prediksi.
5. Himpunan fuzzy bisa digunakan untuk meningkatkan akurasi hasil

6. DAFTAR PUSTAKA

[1] Berry M.J. and Linoff G.S (2000). Mastering Data mining : The Art and science of Customer Relationship Management, New York: John Wiley & Sons, Inc.

[2] Coenen F, 2004, The LUCS-KDD Implementations of CPAR (Classification Based on Predictive Association Rules), Department of Computer Science The University of Liverpool.

[3] Cox, Earl (1994). The Fuzzy Systems Handbook. Massachusetts. Academic Press, Inc.

[4] Jiawei H. and M. Kamber (2001). Data Mining Concepts and Techniques, The Morgan Kaufmann Publishers

[5] J.S.R. Jang, C.T. Sun and E. Mizutani (1997). Neuro-Fuzzy and Soft Computing, London: Prentice-Hall.

[6] Michael W.Berry and Murray Browne (2006). Lecture Notes In Data Mining (eds). World Scientific Publishing Co.Pte.Ltd. Singapore 596224, pp 99-102

[7] Yin X. and Han J., 2003, CPAR: Classification based on Predictive Association Rules, University of Illinois at Urbana-Champaign.

[8] Zuoliang Chen and Guoqing Chen, 2008, Building an Associative Classifier Based On Fuzzy Association Rules, School of Economics and Management, Tsinghua University, Beijing 100084, China.