Konsep dan Teknik Data Mining

Klasifikasi dan Prediksi

- Apa Itu Klasifikasi dan apa itu prediksi?
- Isu yang berkenaan dengan klasifikasi dan prediksi
- Klasifikasi dengan induksi pohon keputusan
- Klasifikasi bayesian

Klasifikasi vs. Prediksi

- Klasifikasi:
  - Meramalkan kategori label kelas (diskrit atau nominal)
  - Mengklasifikasikan (membuat suatu model) berdasarkan himpunan pelatihan dan nilai-nilai (label kelas) dalam suatu atribut klasifikasi dan menggunakannya didalam mengklasifikasikan data baru
- Prediksi:
  - Memodelkan fungsi bernilai kontinu, artinya menaksir nilai-nilai yang tidak diketahui atau nilai-nilai yang hilang

Klasifikasi vs. Prediksi

- Bentuk umum aplikasi
  - Persetujuan kredit
  - Target marketing
  - Diagnosa medis
  - Analisis keefektifan tindakan

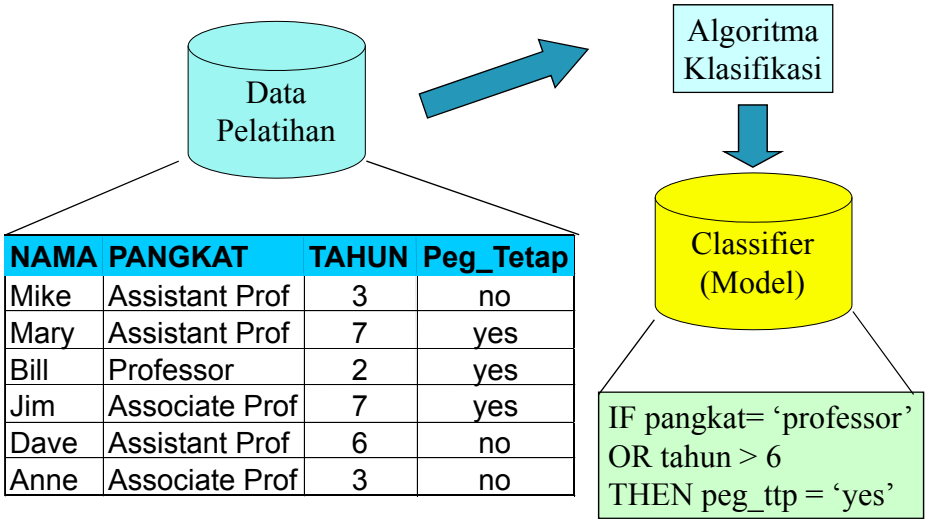
Klasifikasi—Suatu Proses 2 Step

1. Konstruksi model: menguraikan suatu himpunan kelas yang ditentukan sebelumnya
   - Setiap tuple/sample dimisalkan masuk kedalam suatu kelas yang didefinisikan sebelumnya, seperti yang ditetapkan melalui label atribut kelas
   - Himpunan dari tuple yang digunakan untuk konstruksi model dinamai himpunan pelatihan
   - Model disajikan sebagai kaidah klasifikasi, pohon keputusan, atau rumus matematika

Klasifikasi—Suatu Proses 2 Step

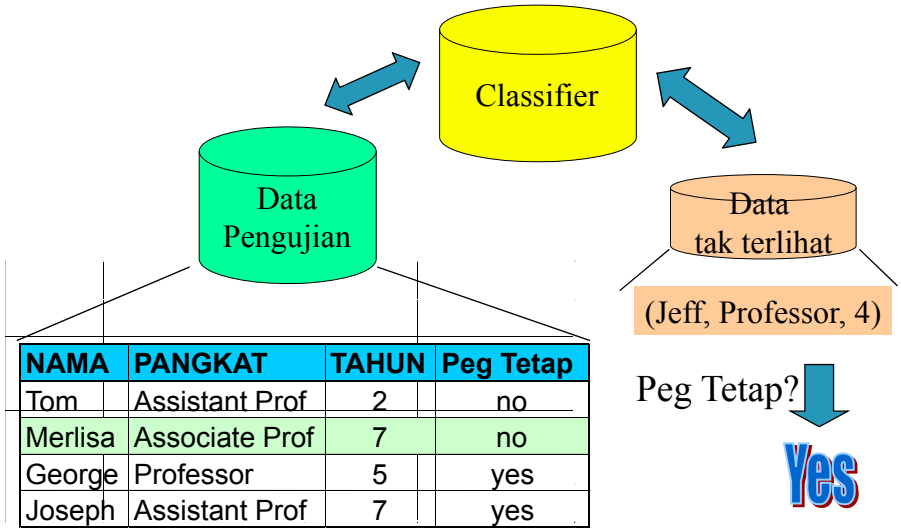
2. Kegunaan model: untuk klasifikasi kedepan atau untuk objek-objek yang tidak dikenal
   - Menaksir akurasi dari model
     - Label yang dikenali melalui uji sampel dibandingkan dengan hasil klasifikasi dari model
     - Tingkat akurasi adalah persentasi dari himpunan sampel uji yang secara benar diklasifikasikan oleh model tersebut
   - **Jika akurasi bisa diterima, gunakan model tersebut untuk mengklasifikasikan tuple data yang label kelasnya tidak diketahui**

Proses Klasifikasi (1): Konstruksi Model



NAMA	PANGKAT	TAHUN	Peg_Tetap
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Proses Klasifikasi (2): Menggunakan Model Dalam Prediksi



NAMA	PANGKAT	TAHUN	Peg Tetap
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Pembelajaran Dengan Supervisi vs. Tanpa Supervisi

- Pembelajaran dengan supervisi (klasifikasi)
  - Supervisi: data pelatihan (observasi, pengukuran, dsb.) disertakan dengan label yang menunjukkan kelas dari pengamatan
  - Data baru diklasifikasikan berdasarkan himpunan pelatihan
- Pembelajaran tanpa supervisi (clustering)
  - Label-label kelas dari data pelatihan tidak dikenal
  - Diberikan sekumpulan pengukuran, observasi, dsb. Dengan tujuan menetapkan keberadaan kelas atau cluster dalam data

Isu Yang Berkenaan dgn Klasifikasi & Prediksi (1):Penyiapan Data

- Pembersihan data
  - Memproses awal data dalam upaya untuk mengurangi noise dan menangani nilai-nilai yang hilang
- Analisis relevansi (seleksi fitur)
  - Membuang atribut yang tak relevan atau redundan
- Transformasi data
  - Menggeneralisasi dan/atau menormalisasi data (data punya skala yang sama).

Isu Yang Berkenaan dgn Klasifikasi & Prediksi (2):Evaluasi Metoda Klasifikasi

- Akurasi prediksi
- Kecepatan dan skalabilitas
  - Waktu untuk membuat model
  - Waktu untuk menggunakan model
- Kekokohan
  - Penanganan noise dan nilai-nilai yang hilang
- Skalabilitas
  - Efisiensi dalam database yang berada dalam disk

Isu Yang Berkenaan dgn Klasifikasi & Prediksi (2):Evaluasi Metoda Klasifikasi

- Penafsiran:
  - Pemahaman dan pengertian yang disediakan oleh model
- Kebaikan dari kaidah
  - Ukuran pohon keputusan
  - Kepadatan dari kaidah klasifikasi

Pohon Keputusan

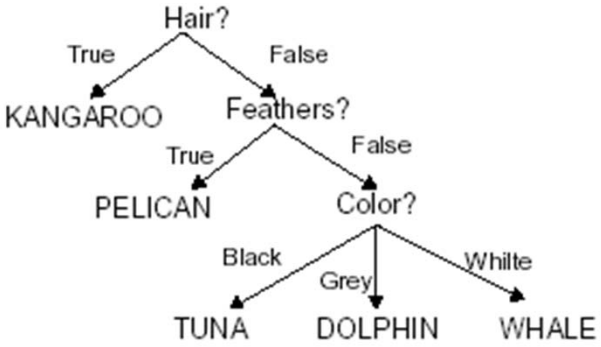
- Pohon keputusan merupakan salah satu tool paling populer untuk klasifikasi karena hasilnya yang bisa dipahami dalam bentuk kaidah keputusan
- Klasifikasi adalah proses penetapan suatu objek baru kedalam kategori atau kelas yang telah didefinisikan sebelumnya.
  - Diberikan sekumpulan record berlabel
  - Buat suatu model (pohon keputusan)
  - Taksir label-label untuk melabeli record-record tak berlabel

Pohon Keputusan

- Suatu **pohon keputusan** biasanya berupa suatu graph berarah dari simpul-simpul dan panah berarah.
- Simpul-simpul seringkali berhubungan dengan suatu pertanyaan atau suatu pengujian

Pohon Keputusan

- Sebagai contoh, suatu pohon keputusan digunakan untuk mengklasifikasikan suatu binatang baru apakah memiliki Hair, Feather (bulu), dan Color-nya



Struktur Dari Pohon Keputusan

- Suatu pohon keputusan dibangun dari 3 tipe dari simpul: simpul root, simpul perantara, dan simpul leaf.
- Simpul leaf memuat suatu keputusan akhir atau kelas target untuk suatu pohon keputusan
- Simpul root adalah titik awal dari suatu pohon keputusan
- Setiap simpul perantara berhubungan dengan suatu pertanyaan atau pengujian

Data Set Pelatihan (org beli komputer)

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Attribute Selection by Information Gain Computation

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"
- I(p, n) = I(9, 5) =0.940
- Compute the entropy for *age*:

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
30…40	4	0	0
>40	3	2	0.971

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$$E(age) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

$\frac{5}{14} I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's. Hence

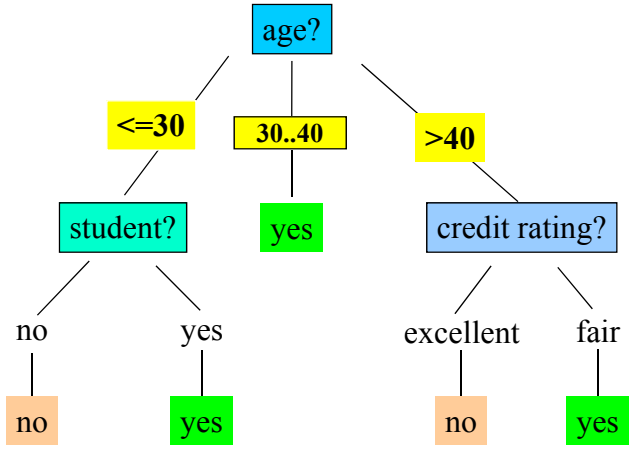
$$Gain(age) = I(p,n) - E(age) = 0.246$$

Similarly,

$$Gain(income) = 0.029$$
$$Gain(student) = 0.151$$
$$Gain(credit\_rating) = 0.048$$

16

Output: Pohon Keputusan Untuk "Membeli Komputer"



Algoritma Untuk Induksi Pohon Keputusan

- Algoritma dasar (suatu greedy algorithm)
  - Pohon dibangun dalam suatu metoda rekursif top-down divide-and-conquer (conquer = menaklukkan)
  1. Seluruh contoh pelatihan dimulai dari simpul root, lalu dilakukan pengujian
  2. Mencabang ke jalur yang benar berdasarkan hasil pengujian
  3. Apakah simpul leaf ditemukan? Jika yes, masukkan contoh ini ke kelas target, jika tidak kembali ke langkah 1.
  - Atribut-atribut berada dalam suatu kategori (jika bernilai kontinu, nilai-nilai tersebut didiskritkan terlebih dahulu)

Algoritma Untuk Induksi Pohon Keputusan

- Contoh-contoh dipartisi secara rekursif berdasarkan atribut terpilih
- Atribut-atribut uji dipilih berdasarkan heuristik atau pengukuran statistik (misal, information gain)

- Kondisi-kondisi penghentian partisi
  - Seluruh sampel untuk suatu simpul yang diberikan masuk ke kelas yang sama
  - Tidak ada atribut sisa untuk partisi selanjutnya -- majority voting diperlakukan untuk klasifikasi leaf
  - Tidak ada sampel tertinggal

Ukuran Seleksi Atribut: Information Gain

- Memilih atribut dengan information gain terbesar
- S memuat tupel-tupel s_i dari kelas C_i untuk i = {1, …, m}
- Ukuran-ukuran informasi info diperlukan untuk mengklasifikasikan tuple sebarang apapun

$$I(s_1,s_2,\ldots,s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} log_2 \frac{s_i}{s}$$

Ukuran Seleksi Atribut: Information Gain

- entropy dari atribut A dengan nilai-nilai $\{a_1,a_2,\ldots,a_v\}$

$$E(A)=\sum_{j=1}^{v}\frac{s_{1j}+\ldots+s_{mj}}{s}I(s_{1j},\ldots,s_{mj})$$

- Informasi yang diperoleh melalui pencabangan pada atribut A

$$Gain(A) = I(s_1,s_2,\ldots,s_m) - E(A)$$

Kaidah Ekstraksi Klasifikasi Dari Pohon

- Sajikan pengetahuan dalam bentuk kaidah IF-THEN
- Satu kaidah dibuat untuk setiap path dari root ke leaf
- Setiap pasangan nilai atribut sepanjang path membentuk suatu fabungan
- Simpul leaf menyimpan prediksi kelas
- Kaidah-kaidah lebih mudah dipahami manusia

Kaidah Ekstraksi Klasifikasi Dari Pohon

- Contoh-contoh

IF *age* = "<=30" AND *student* = "*no*" THEN *buys_computer* = "*no*"

IF *age* = "<=30" AND *student* = "*yes*" THEN *buys_computer* = "*yes*"

IF *age* = "31…40" THEN *buys_computer* = "*yes*"

IF *age* = ">40" AND *credit_rating* = "*excellent*" THEN *buys_computer* = "*yes*"

IF *age* = "<=30" AND *credit_rating* = "*fair*" THEN *buys_computer* = "*no*"

Contoh klasifikasi Decision Tree: Algo ID3

Database T :

Attribute1	Attribute2	Attribute3	Class
A	70	True	Class1
A	90	True	Class2
A	85	False	Class2
A	95	False	Class2
A	70	False	Class1
B	90	True	Class1
B	78	False	Class1
B	65	True	Class1
B	75	False	Class1
C	80	True	Class2
C	70	True	Class2
C	80	False	Class1
C	80	False	Class1
C	96	False	Class1

Klasifikasi Decision Tree: Algo ID3

Basis Data T yang diatas memiliki 14 sampel dengan tiga buah atribut penghitung yang akan dibagi menjadi dua kelas yaitu "Class1" dan "Class2". Class1 memiliki 9 sampel dan Class2 memiliki 5 sampel. Jadi entropy penghitungnya adalah

$$I_T(9,5) = -\sum_{i=1}^{2} \frac{Si}{S} Log_2 \frac{Si}{S}$$

$$= -\left[\left(\frac{9}{14} Log_2 \frac{9}{14}\right) + \left(\frac{5}{14} Log_2 \frac{5}{14}\right)\right]$$

$$= 0{,}94027$$

Hitungan Nilai Informasi

$$\frac{9}{14}\log_2\frac{9}{14} = \frac{9}{14}\log_2(0{,}64285$$

$$= \frac{9}{14}\,\frac{\log 0{,}64285}{\log 2}$$

$$= \frac{9}{14}\,\frac{-0{,}19189}{0{,}30103} = -0{,}40977$$

$$\frac{5}{14}\log_2\frac{5}{14} = \frac{5}{14}\,\frac{\log 0{,}35714}{\log 2}$$

$$= \frac{5}{14}\,\frac{-0{,}44716}{0{,}30103} = -0{,}53050 \quad +$$

$$= -0{,}94027$$

$$I_T(9,5) = -\left[\left(\frac{9}{14} Log_2 \frac{9}{14}\right) + \left(\frac{5}{14} Log_2 \frac{5}{14}\right)\right] = 0{,}94027$$

Entropy untuk atribute ke 1

Sesudah mendapatkan nilai diatas tersebut, maka penghitungan akan menghitung pada atribut masing-masing untuk mendapatkan nilai gain yang terbesar. Untuk nilai informasi Attribute1 adalah :

Attribute1	Banyaknya sampel pada "Class1"	Banyaknya sampel pada "Class2"	I(Class1,Class2)
A	2	3	0,9709
B	4	0	0
C	3	2	0,9709

Entropy dan Information Gain untuk atribute ke 1

$$E(Attribute1) = \frac{Sij}{|Sj|} I_A(2,3) + \frac{Sij}{|Sj|} I_B(4,0) + \frac{Sij}{|Sj|} I_C(3,2)$$

$$E(Attribute1) = \frac{5}{14}(0,9709) + \frac{4}{14}(0) + \frac{5}{14}(0,9709)$$

$$E(Attribute1) = 0,6935$$

$$Gain(Attribute1) = I(9,5) - E(Attribute1)$$

$$Gain(Attribute1) = 0,9402 - 0,6935$$

$$Gain(Attribute1) = 0,2467$$

Nilai Informasi untuk atribute ke 2 (numerik)

Untuk nilai informasi pada Attribute2, data harus diurutkan terlebih dahulu, seperti rumus yang telah kita ketahui diatas. Nilai Attribute2 sesudah diurut adalah: 65,70,75,78,80,85,90,95,96.

Attribute2	Banyaknya sampel pada "Class1"	Banyaknya sampel pada "Class2"	I(Class1,Class2)	E(Attribute2)	Gain (Attribute2)
<=65	1	0	0	-	-
>65	0	0	-		
<=70	3	1	0,8112	0,9251	0,0149
>70	6	4	0,9708		
<=75	4	1	0,7218	0,8947	0,0450
>75	5	4	0,8947		
<=78	5	1	0,6499	0,8499	0,0901
>78	4	4	1		
<=80	7	2	0,7641	0,8379	0,1020
>80	2	3	0,9708		
<=85	7	3	0,8812	0,9151	0,0240
>85	2	2	0,9151		
<=90	8	4	0,9182	0,9298	0,0102
>90	1	1	1		
<=95	0	1	0	-	-
>95	0	0	-		

Nilai Informasi untuk atribute ke 3 (logical)

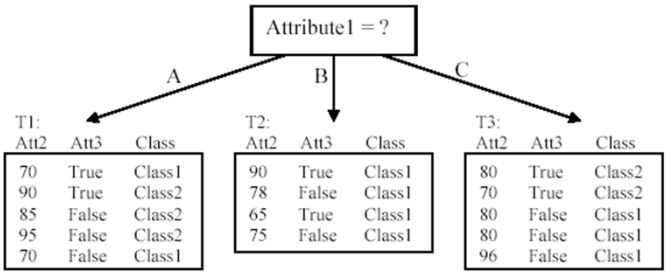
Setelah dihitung dan mendapatkan nilai *gain* masing-masing, maka nilai 80 yang akan dijadikan batas pembaginya, karena memiliki informasi *gain* yang tertinggi yaitu 0,1020.

Untuk nilai *gain* pada attribute3 adalah

Attribute3	Banyaknya sampel pada "Class1"	Banyaknya sampel pada "Class2"	I(Class1,Class2)	E(Attribute3)	Gain (Attribute3)
True	3	3	1	0,892	0,048
False	6	2	0,8112		

Klasifikasi Decision Tree: Algo ID3

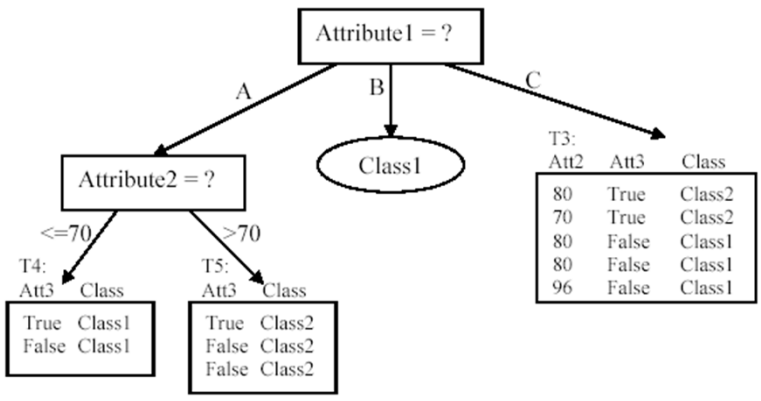
Sesudah menghitung informasi *gain* dari semua atribut, maka attribute1 yang dipilih untuk menjadi atribut pembagi karena memiliki nilai informasi *gain* yang tertinggi yakni 0,2467. Cabang pohon yang dibentuk adalah 3 cabang, karena attribute1 memiliki 3 nilai unik(kelas). Setelah dibagi, setiap anak cabang akan memiliki sampel yang nilai attribute1 nya sama dengan cabangnya, seperti yang dapat dilihat pada gambar 2.5.



Gambar 2.5 Pohon keputusan awal yang dibentuk

Klasifikasi Decision Tree: Algo ID3

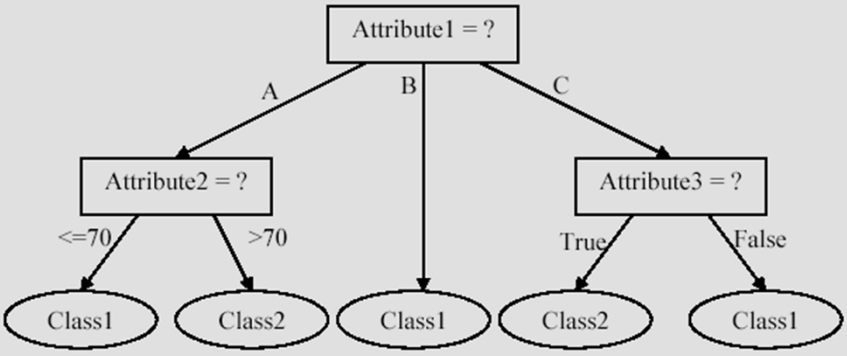
Nilai informasi *gain* yang tertinggi adalah pada attribute2 dengan batas nilai 70 yaitu 0,9709. Pohon keputusan yang dibentuk adalah seperti pada gambar 2.6.



Gambar 2.6 Pohon keputusan yang dibentuk setelah T1 dihitung

Klasifikasi Decision Tree: Algo ID3

Nilai informasi *gain* yang tertinggi adalah pada attribute3 yaitu 0,9709, sehingga didapat pohon keputusan yang baru yaitu seperti pada gambar 2.7.



Gambar 2.7 Pohon keputusan yang dibentuk dari sampel contoh

Decision Rule dan Kegunaannya

- Bila ada pertanyaan :
  - Atribute 1 = A
  - Atribute 2 = 75
  - Atribute 3 = F
- Maka menurut rule yang telah di training hasilnya adalah: masuk Class 2

```
IF   atribut1 = A
     THEN
          IF    atribut2 <= 70
                THEN
                     Klasifikasi = Class 1;
                ELSE
                     Klasifikasi = Class 2;
ELSEIF    atribut1 = B
     THEN
                     Klasifikasi = Class 1;
ELSEIF   atribut1 = C
     THEN
          IF    atribut3 <= True
                THEN
                     Klasifikasi = Class 2;
                ELSE
                     Klasifikasi = Class 1
```

Klasifikasi Dalam Database Besar

- Mengapa induksi pohon keputusan dalam data mining?
  - Kecepatan pembelajaran yang relatif lebih cepat(dari metoda klasifikasi lainnya)
  - Bisa dikonversi ke kaidah klasifikasi yang sederhana dan mudah dipahami
  - Bisa menggunakan query SQL untuk pengaksesan database
  - Akurasi klasifikasi yang bisa dibandingkan dengan metoda lainnya

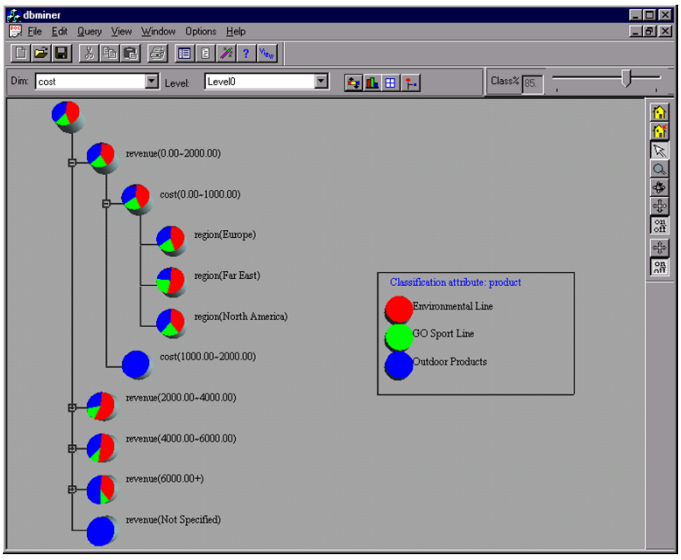
Studi Metoda Dalam Data Mining: Induksi Pohon Keputusan yg Skalabel

- SLIQ (EDBT'96 — Mehta dkk.)
  - Membangun suatu indeks untuk setiap atribut dan hanya daftar kelas dan daftar atribut sekarang yang ada dalam memori
- SPRINT (VLDB'96 — J. Shafer dkk.)
  - Membangun suatu struktur data daftar atribut
- PUBLIC (VLDB'98 — Rastogi & Shim)
  - Mengintegrasikan pohon terpisah dan pohon terpangkas: menghentikan pertumbuhan pohohn lebih awal

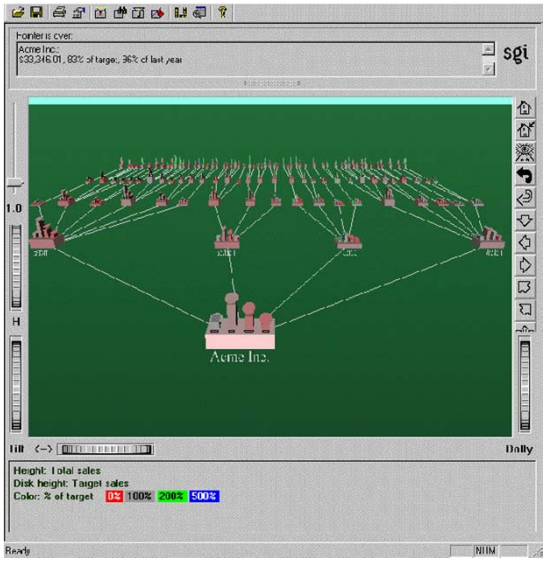
Studi Metoda Dalam Data Mining: Induksi Pohon Keputusan yg Skalabel

- RainForest (VLDB'98 — Gehrke, Ramakrishnan & Ganti)
  - Memisahkan aspek-aspek skalabilitas dari kriteria yang menentukan kualitas dari pohon
  - Membangun suatu daftar AVC (atribut, value, lebel class)

Presentasi Hasil Klasifikasi



Visualisasi Dari Suatu Pohon Keputusan Dalam SGI/MineSet 3.0



Teorema Bayesian: Dasar

- Misal X merupakan suatu sampel data yang label kelasnya tidak dikenal
- Misal H suatu hipotesa dimana X masuk ke kelas C
- Untuk problem klasifikasi, tentukan P(H/X): peluang dimana hipotesa berlaku untuk sampel data pengamatan X diberikan

Teorema Bayesian: Dasar

- P(H): peluang hipotesa sebelumnya H (artinya peluang awal sebelum kita mengamati data apapun, menggambarkan latar belakang pengetahuan)
- P(X): peluang bahwa sampel data diamati
- P(X|H) : peluang pengamatan sampel X, apabila diberikan bahwa hipotesa berlaku

Teorema Bayesian

- Diberikan data pelatihan *X, peluang yang mengikuti suatu hipotesa H, P(H|X) memenuhi teorema Bayes*

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

- Secara informal, ini bisa ditulis sebagai

  posterior =likelihood x prior / evidence
- Hipotesis MAP (maximum posteriori)

$$h_{MAP} \equiv \arg\max_{h \in H} P(h|D) = \arg\max_{h \in H} P(D|h)P(h).$$

- Prakteknya sukar: membutuhkan pengetahuan awal dari banyak peluang, biaya komputasi cukup berarti

Classifier Bayes Naif

- Suatu penyederhanaan asumsi: atribut-atribut bebas bersyarat:

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$

- Perkalian dari kejadian, katakan 2 elemen x_1 dan x_2, dimana kelas sekarang diberikan C, adalah perkalian dari peluang masing-masing elemen yang diambil terpisah, diberikan kelas yang sama $P([y_1,y_2],C) = P(y_1,C) * P(y_2,C)$

Classifier Bayes Naif

- Tak ada relasi kebergantungan diantara atribut
- Sangat besar mengurangi biaya komputasi, hanya menghitung distribusi kelas.
- Apabila peluang $P(X|C_i)$ sudah diketahui, berikan X ke kelas dengan $P(X|C_i)*P(C_i)$ masksimum

Dataset Pelatihan

Class:
C1:buys_computer= ‘yes’
C2:buys_computer= ‘no’

Sampel Data
X =(age<=30,
Income=medium,
Student=yes
Credit_rating=
Fair)

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
30…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Classifier Bayes Naif: Contoh

- Hitung P(X/Ci) untuk masing-masing kelas

P(age=“<30” | buys_computer=“yes”) = 2/9=0.222
P(age=“<30” | buys_computer=“no”) = 3/5 =0.6
P(income=“medium” | buys_computer=“yes”)= 4/9 =0.444
P(income=“medium” | buys_computer=“no”) = 2/5 = 0.4
P(student=“yes” | buys_computer=“yes)= 6/9 =0.667
P(student=“yes” | buys_computer=“no”)= 1/5=0.2
P(credit_rating=“fair” | buys_computer=“yes”)=6/9=0.667
P(credit_rating=“fair” | buys_computer=“no”)=2/5=0.4

X=(age<=30 ,income =medium, student=yes,credit_rating=fair)

P(X|Ci) : P(X|buys_computer=“yes”)= 0.222 x 0.444 x 0.667 x 0.667 =0.044
P(X|buys_computer=“no”)= 0.6 x 0.4 x 0.2 x 0.4 =0.019

P(X|Ci)*P(Ci) :
P(X|buys_computer=“yes”) * P(buys_computer=“yes”)=0.044*(9/14)=0.028
P(X|buys_computer=“no”) * P(buys_computer=“no”)=0.019*(5/14)=0.007

Jadi : X masuk ke kelas “buys_computer=yes”

Classifier Bayes Naif: Komentar

- Keuntungan:
  - Mudah diimplementasikan
  - Dalam banyak kasus, hasilnya baik
- Kerugian
  - Asumsi: kelas bebas bersyarat, jadi kehilangan akurasi
  - Dalam prakteknya, kebergantungan ada diantara variabel
  - Misal rumah sakit: pasien: Profil: usia, family history, misalkan DeBrito-Cowok, Stece-Cewek

    Symptom: demam,batuk dsb., Penyakit: lung cancer, diabetes dsb
  - Kebergantungan diantara variabel ini tidak dapat dimodelkan dengan Classifier Bayes Naif

Metoda Klasifikasi Lainnya

- k-nearest neighbor classifier
- Penalaran berbasis kasus
- Algoritma genetika
- Pendekatan himpunan rough
- Pendekatan himpunan Fuzzy