Jurnal Sainstek Vol. II No. 1: 87-93, Juni 2010 ISSN: 2085-8019

PROSES DECISION TREE PADA DATAMINING DENGAN ALQORITMA ID3

Lita Sari Muchlis

Program Studi Manajemen Informatika STAIN Batusangkar
Jl. Sudirman No. 137 Kuburajo Lima Kaum Batusangkar, Sumatera Barat 27213
Email: Litasarimuchlis@yahoo.com

ABSTRACT

This article was study about decision process data meaning with algoritma ID3. Data meaning is an atomatic extraction process of large data. It was found with a contour data. Data meaning have a function to produce a different contour wich other. The function of classification of data meaning is helping write decision tree, the function with algoritma ID3.

Key words: data meaning, classification, decision tree

PENDAHULUAN

Suatu Aplikasi yang memudahkkan dalam penyimpanan dan pengaksesan data yang menyebabkan membengkaknya jumlah data yang tersedia. Sudah banyak orang yang menyadari bahwa data yang berukuran besar tersebut sebenarnya mengandung berbagai jenis pengetahuan tersembunyi yang berguna untuk proses pengambilan keputusan. Akan tetapi, pengetahuan akan sangat sulit ditemukan dengan cara menganaiisis data secara manual. Oleh karena itu, dilakukan *data mining* untuk mengekstraksi pengetahuan secara otomatis dari data berukuran besar dengan cara mencari pola-pola menarik yang terkandung di dalam data tersebut. *Data mining* memiliki banyak fungsionalitas, antara lain pembuatan ringkasan data, analisis asosiasi antar data, klasifikasi data, prediksi, dan pengelompokan data. Setiap fungsionalitas akan menghasilkan pengetahuan atau pola yang berbeda satu sama lain. Pada klasifikasi, akan dihasilkan sebuah model yang dapat memprediksi kelas atau kategori dari objekobjek di dalam basisdata. Sebagai contoh, klasifikasi dapat digunakan oleh petugas peminjaman uang di sebuah bank untuk memprediksi proses pengajuan kredit oleh nasabah (custumer) menjadi semakin mudah, baik untuk kredit barang maupun uang pemohon mana yang aman dan mana yang beresiko untuk diberi pinjaman, oleh periset di bidang medis untuk memprediksi jenis pengobatan apa yang cocok diberikan kepada seorang pasien dengan penyakit tertentu. Pada kasus-kasus tersebut, model klasifikasi dibuat untuk memprediksi kelas "aman" atau "beresiko" untuk data permohonan pinjaman; "beli" atau "tidak" untuk data pemasaran; dan "pengobatan-1", "pengobatan-2", atau "pengobatan-3" untuk data medis. Model klasifikasi dibuat dengan cara menganalisis *training data* (terdiri dari objek-objek yang kelasnya sudah diketahui). Model yang dihasilkan kemudian akan digunakan untuk memprediksi kelas dari *unknown data* (terdiri dari objek-objek yang kelasnya belum diketahui). Model klasifikasi dapat digambarkan dalam beberapa bentuk, seperti aturan klasifikasi (IF-THEN), pohon keputusan, rumus matematika, atau jaringan saraf tiruan. Pohon keputusan banyak digunakan karena mudah dipahami oleh manusia serta mampu menangani data beratribut banyak.

Penggunaan Model Pohon Keputusan (*Decision Tree*) adalah salah satu teknik klasifikasi sebagai bagian dari ilmu Data Mining. Data Mining melakukan penggalian pengetahuan (*knowledge*) terhadap data. Untuk

teknik klasifikasi ini, digunakan algoritma ID3 yang pertama kali dikembangkan oleh Ross Quinlan di Universitas Sydney pada tahun 1975 dalam bukunya: *Machine Learning*, vol. 1, no. 1. ID3 is based off the Concept Learning System (CLS) algorithm.

Kelebihan dari penggunaan model Pohon Keputusan ini selain mudah dipahami juga dapat digunakan untuk menemukan aturan atau syarat-syarat yang dapat dijadikan sebagai kriteria yang berguna untuk keperluan analisa dalam suatu proses pengambilan keputusan.

Untuk membuat alat bantu klasifikasi dan identifikasi data dengan model Pohon Keputusan menggunakan algoritma ID3 dan mendapatkan hasil yang optimal dengan pemahaman akan teknik yang digunakan terhadap model permasalahan yang dihadapi.

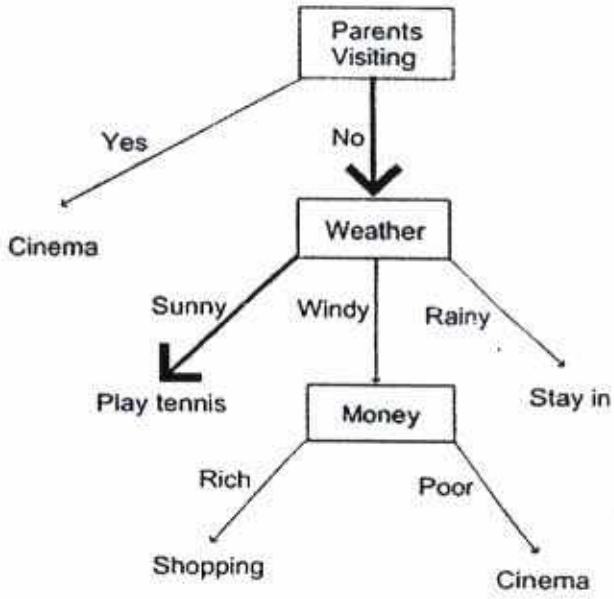
DECISION TREE MENGGUNAKAN ID3

Pohon Keputusan (*Decision Tree*)

Decision Tree merupakan suatu pendekatan yang sangat populer dan praktis dan *machine learning* untuk menyelesaikan permasalahan klasifikasi. Metode ini digunakan untuk memperkirakan nilai diskrit dari fungsi target yang mana fungsi pembelajaran direpresentasikan oleh sebuah *decision tree*. *Decision Tree* merupakan himpunan aturan IF..THEN. Setiap *path* dalam *tree* dihubungkan dengan sebuah aturan, dimana premis terdiri dari sekumpulan *node-node* yang ditemui, dan kesimpulan dari aturan terdiri atas kelas yang terhubung dengan *leaf* dan *parth*

Dalam Pohon keputusan *leafnodei* diberikan sebuah label kelas. *Non-terminal node*, yang terdiri atas *root* atas *internalnode* lainnya. Mengandung kondisi-kondisi uji atribut untuk memisahkan *record* yang memiliki karakteristik yang berbeda. *Edge-edge* dapat dilabelkan dengan nilai-nilai *numeric symbolic*. Sebuah atribut *numeric-symboilic* adalah atribut yang dapat bernilai *numeric* ataupun *symbolic* yang dihubungkan dengan dengan sebuah variabel kuantitatif. Sebagai contoh, ukuran seseorang dapat dituliskan " 1,75 meter", ataupun sebagai nilai *numeric-symbolic*: seperti "tinggi" yang berkaitan dengan suatu ukuran (size). Nilai seperti inilah yang menyebabkna perluasan dari *decesion tree* .

Pada sebuah pohon dapat digunakan untuk memetakan pilihan keputusan. Pohon yang memetakan pilihan-pilihan keputusan tersebut dinamakan *decision tree* atau pohon keputusan. Contoh bentuk dari pohon keputusan dapat dilihat pada gambar 1



Gambar 1. Pohon Keputusan (*Decision Tree*)

Pada Gambar 1 Pohon Keputusan dapat dibaca sebagai berikut :

1. Jika orang tua berkunjung, kemudian pergi ke bioskop
2. Jika orangtua tidak mengunjungi *dan* cerah, kemudian bermain tenis
3. Jika orangtua tidak mengunjungi *dan* ini berangin *dan* kau kaya, kemudian pergi berbelanja
4. Jika orangtua tidak mengunjungi *dan* ini berangin *dan* kau miskin, kemudian pergi ke bioskop
5. Jika orangtua tidak mengunjungi *dan* hari hujan, maka tinggal masuk

Representasi pohon berakar pada pohon keputusan adalah sebagai berikut:

- Simpul dalam merepresentasikan tes atribut
- Daun merepresentasikan klasifikasi

LEARNING POHON KEPUTUSAN

Menentukan Masalah

Bagaimana Pohon Keputusan mengidentifikasi masalah? Kita mempunyai sekumpulan contoh yang dikategorikan menjadi beberapa kategori (keputusan-keputusan). Kita juga mempunyai sekumpulan atribut yang menjelaskan contoh-contoh yang ada, dan setiap atrubut memiliki nilai yang terhingga. Kita ingin menggunakan contoh-contoh yang ada untuk mempelajari struktur dari sebuah pohon keputusan yang dapat digunakan untuk menentukan kategori dari contoh yang belum ada (*unseen example*).

Ide dasar

Pada gambar 1, simpul "parent visiting" diletakan paling atas (sebagai akar). Kita belum tahu mengapa, sebagaimana kita tidak tahu bagaimana proses pembuatan pohon keputusan tersebut. Namun, jika orang tua berkunjung, maka sudah pasti keputusannya adalah pergi ke cinema. Oleh karena itu, pilihan parent visiting bisa ditaruh paling atas tanpa memperdulikan kemungkinan yang lain. Pemikiran diatas yang mendasari algoritma ID3.

Entropy

Membuat pohon keputusan adalah perkara memilih atribut mana yang harus diuji pada setiap simpul pada Pohon untuk menentukan ukuran dimana proses ini disebut information gain, yang berguna untuk menentukan atribut mana yang akan digunakan pada setiap simpul. Information Gain itu sendiri didapatkan dari perhitungan yang menggunakan satuan yang disebut entropy. Mendefenisikan kasus pada keputusan biner dan kemudian menentukan kasus umum. Rumus Entropy adalah

$$Entropy\ (S) \equiv \sum_{i}^{c} - p_i \log_2 p_i$$

c : jumlah nilai yang ada pada atribut target (jumlah kelas klasifikasi).
p_i : porsi sampel untuk kelas i.

Information Gain

Informatin Gain adalah suatu nilai statistik yang digunakan untuk memilih atribut yang akan mengekspansi *tree* dan menghasilkan *node* baru pada alqoritme ID3. suatu *entropy* dipergunakan untuk mendefenisikan nilai *information gain.*

Entropy digunakan sebagai parameter untuk mengukur heterogenitas (keberagaman) dari kumpulan sampel data dan Jika kumpulan sampel data semakin heterogen, maka nilai *entropy*nya semakin besar. *Information Gain (IG)* dimanfaatkan sebagai Efektivitas atribut dalam mengklasifikasikan data Dihitung berdasarkan entropy dengan ketentuan rumus sebagai berikut:

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

- A : atribut
- V : menyatakan suatu nilai yang mungkin untuk atribut A
- $Values\ (A)$: himpunan nilai-nilai yang mungkin untuk atribut A
- $|S_v|$: jumlah sampel untuk nilai v

- $|S|$: jumlah seluruh sampel data
- Entropy(S_v) : *entropy* untuk sampel-sampel yang memiliki nilai *v*

Sebagai contoh, misalnya kita memiliki sekumpulan contoh S = {s1,s2,s3,s4} yang dikategorikan menjadi positif dan negatif, dimana s1 positif dan sisanya negative. Kita ingin menghitung information gain dari sebuah atribut A, dan A dapat memiliki nilai {v1,v2,v3}. Ditentukan bahwa :

s1 memiliki nilai v2 untuk A
s2 memiliki nilai v2 untuk A
s3 memiliki nilai v3 untuk A
s4 memiliki nilai v1 untuk A

Untuk mendapatkan information gain, pertama kita harus menghitung entropy dari S. Untuk menggunakan persamaan entropi pada persoalan ini, kita harus mengetahui jumlah positif dan negatifv pada S. Dari soal dapat diketahui bahwa positif = 1/4 dan negative = 3/4, sehingga kita bisa menghitung :

Entropy(S) = - (1/4)log2(1/4) – (3/4)log2(3/4) = - (1/4)(-2) – (3/4)(-0,415) = 0,5 + 0,311 = 0,811

Selanjutnya kita perlu menghitung Entropi(Sv) untuk setiap nilai v = v1, v2, v3, v4. Sv merupakan kumpulan dari contoh pada S yang memiliki nilai v pada atribut A, atau dapat ditulisakan sebagai berikut:

Sv1 = {s4}, sv2={s1, s2}, sv3 = {s3}

Sekarang, kita dapat menyelesaikan persamaan-persamaan berikut:

(|Sv1|/|S|) * Entropy(Sv1) = (1/4) * (-(0/1)log2(0/1) - (1/1)log2(1/1)) = (1/4)(-0 -(1)log2(1)) = (1/4)(-0 -0) = 0
(|Sv2|/|S|) * Entropy(Sv2) = (2/4) * (-(1/2)log2(1/2) - (1/2)log2(1/2)) = (1/2) * (-(1/2)*(-1) - (1/2)*(-1)) = (1/2) * (1) = 1/2
(|Sv3|/|S|) * Entropy(Sv3) = (1/4) * (-(0/1)log2(0/1) - (1/1)log2(1/1)) = (1/4)(-0 -(1)log2(1)) = (1/4)(-0 -0) = 0

Sekarang kita bisa menambahkan ketiga nilai tersebut dan mendapatkan hasil dari perhitungan Entropi(S) yang merupakan Informasi diperoleh melalui percabangan pada atribut A:

Gain(S,A) = 0.811 - (0 + 1/2 + 0) = 0.311

Informasi selanjuntnya digunakan untuk menggunakan Alqoritma !D3 dalam membangun Pohon keputusan

ALGORITMA ID3

Konsep dasar Algoritma ID3

Perhitungan untuk mendapat informasi adalah bagian sulit dari alqoritma. Algoritma ID3 melakukan pencarian dengan menggunakan Pohon Keputusan dan melibatkan penambah simpul ke pohon yang ada. Fungsi dari Algoritma ID3 adalah sebagai berikut :

- Sebuah algoritma matematika untuk membangun pohon keputusan.
- Diciptakan oleh Ross Quinlan J. pada tahun 1979.
- Teori Informasi Menggunakan ditemukan oleh Shannon pada tahun 1948.
- Membangun pohon dari atas ke bawah, dengan tidak mundur.
- Informasi Keuntungan digunakan untuk memilih atribut yang paling berguna untuk klasifikasi.

Algoritma ID3 (Contoh, Target_Attribute, Attributes) dapat dituliskan sebagai berikut ID3 :

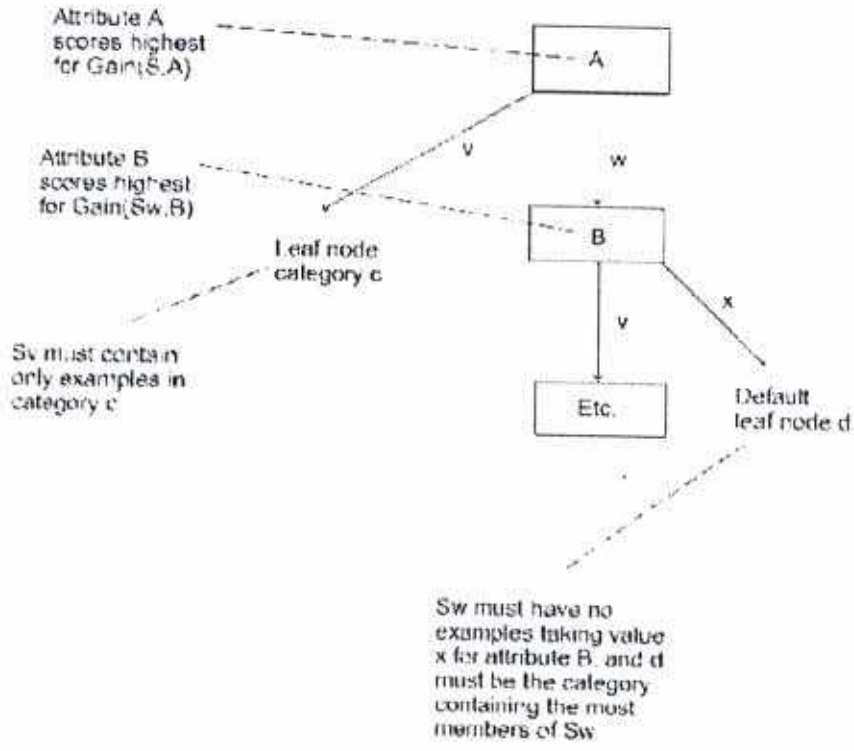
- Buat node root untuk pohon
- JIKA semua contoh positif, Kembali tunggal Root node pohon itu-, dengan label = +
- Jika semua contoh negatif, Kembali tunggal Root node pohon itu-, dengan label = -
- Jika jumlah memprediksi atribut kosong, maka Kembali node tunggal Root pohon, dengan label = nilai umum sebagian besar atribut target pada contoh
- Jika tidak Mulailah
  - J □ The Atribut yang paling mengklasifikasikan contoh
  - Keputusan atribut Pohon untuk Root □ A
  - Untuk setiap nilai positif, vi, A,
    - Tambah pohon cabang baru di bawah Root, sesuai dengan Tes = vi

- Contoh Biarkan (vi), menjadi bagian dari contoh yang memiliki nilai untuk A vi
- Jika contoh (vi) kosong
- Kemudian di bawah ini cabang baru menambahkan simpul daun dengan label = paling nilai umum target pada contoh
- Lain di bawah ini cabang baru menambahkan ID3 subpohon (Contoh (vi), Target_Attribute, Atribut - (A))
- Akhir
- Kembali Root

Dari Alqoritma diatas maka dapat kita misallkan, diberikan sekumpulan contoh S, dikategorikan dalam sekumpulan kategori ci, maka:

1. Tentukan simpul akar yang merupakan atribut A yang memiliki nilai information gain terbesar pada S
2. Untuk setiap nilai v yang mungkin dimiliki A, gambarkan cabang untuk simpul tersebut
3. Untuk setiap cabang dari A yang berkorespondensi dengan nilai v, hitung Sv.
   - Jika Sv kosong, pilih kategori cdefault yang mempunyai paling banyak contoh pada S, dan letakan pada simpul daun yang mengakhiri cabang tersebut
   - Jika Sv hanya mengandung contoh dari kategori c, maka letakan c sebagai simpul daun yang mengakhiri cabang tersebut
   - Selain itu, hilangkan A dari kumpulan atribut yang bisa diletakan pada simpul. Lalu, letakan simpul baru pada *decision tree*, dimana atribut baru yang diletakan adalah atribut yang memiliki information gain terbesar terhadap Sv. Ulangi langkah diatas dengan mengganti S dengan Sv

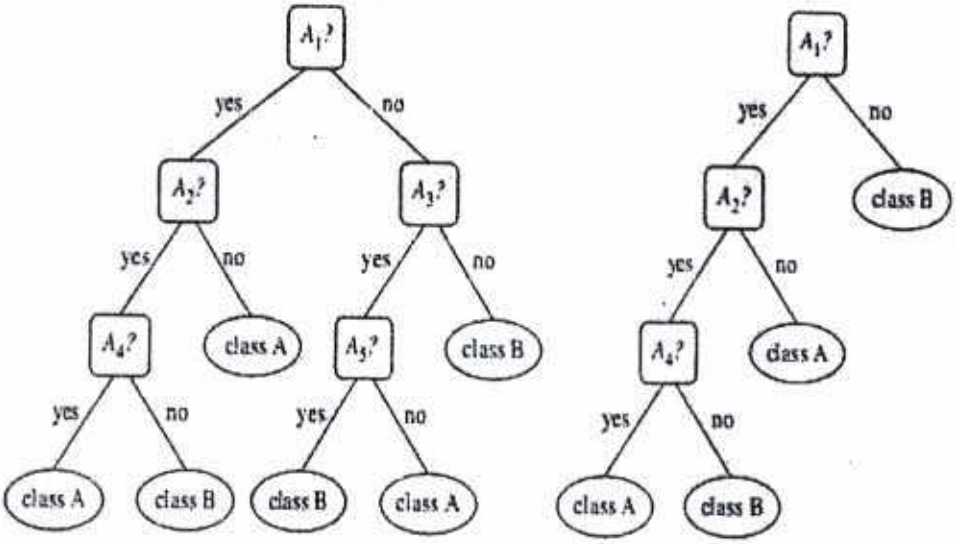
Algoritma di atas berhenti jika atribut sudah habis, atau *decision tree* sudah mengklasifikasi contoh dengan sempurna. Ilustrasi dari algoritma ID3 dapat dilihat Gambar 2.



Gambar 2. Ilustrasi dari algoritma ID3

Pemangkasan Pohon

Pada saat pembangunan pohon keputusan, banyaknya cabang mungkin mencerminkan adanya *noise* pada *training data*. Pemangkasan pohon dapat dilakukan untuk mengenali dan menghapus cabang-cabang tersebut. Pohon yang dipangkas akan menjadi lebih kecil dan lebih mudah dipahami. Pohon semacam itu biasanya juga menjadi lebih cepat dan lebih baik dalam melakukan klasifikasi terhadap *unknown data*. Melakukan proses klasisfikasi data disimpan pada database dengan langkah pengerjaan yaitu membangun Pohon keputusan dan menyederhanaan keputusan dalam bentuk pemangkasan dengan pendekatan *Prunning Desicion Tree* (Gambar 3).



Gambar 3. Pohon keputusan sebelum dan setelah dipangkas

Proses Kerja *Prunning Desicion Tree* dalam sebuah pohon keputusan jika sebuah *rule* telah dibuat berdasarkan pohon keputusan, maka prosesnya sebagai berikut:

1. Eliminasi anteseden yang tidak perlu, cara
   a. Buat tabel kontigensi untuk setiap rule yang mempunyai anteseden
   b. Sederhanakan rule dengan cara mengeliminasi anteseden
2. Eliminasi rule yang tidak perlu

KESIMPULAN

Teknologi data mining dengan menggunakan pemodelan Pohon Keputusan dapat digunakan untuk melakukan klasifikasi data. Dan hsil dari klasifikasi yang didapatkan dengan menggunakan alat bantu ini adalah dengan ditemukannnya aturan-aturan yang dapat digunakan sebagai kriteria-kriteria untuk mengidentifikasi dan mengetahui apakah suatu tindakan atau kegiatan harus dilakukan atau tidak.

Pohon Keputusan dengan Alqoritma ID3 dapat digunakan untuk memperoleh pengetahuan tentang pengolahan data dalam jumlah yang besar khususnya untuk mengklasifikasikan pemberian data pada pengambil keputusan.

Pemangkasan pohon dapat dilakukan untuk menghilangkan cabang-cabang tidak perlu yang terbentuk akibat adanya *noise* atau *outlier* pada *training data*.

DAFTAR KEPUSTAKAAN

Blake CL, Merz CJ. 1998. *UCI Repository of machine learning databases.Irvine*, CA: University of California, Department of Information and Computer Science.

Munir R. 2006. *Diktat Kuliah IF2153 Matematika Diskrit*. Program Studi Teknik Informatika, Institut Teknologi Bandung

Prayudi Y. 2002. *Knowledge Discovery In Medical Data*. Singapore. McGraw-Hill

Tom MM. 1997. *Machine Learning*. Singapore. McGraw-Hill

Russel N. 2003. *Artificial Inteligence – a Modern Approach 2nd Edition*. Pearson Education Inc, New Jersey