Klasifikasi-Decision Tree

Data Mining

Decision Tree

- Untuk mengklasifikasikan suatu obyek, seringkali diajukan urutan pertanyaan sebelum bisa ditentukan kelompoknya.
- Jawaban pertanyaan pertama akan mempengaruhi pertanyaan berikutnya.

Pemilihan Atribut

- Tujuan pemilihan atribut adalah untuk mendapatkan decision tree yang paling kecil ukurannya.
- *Pure (bersih)* adalah apabila dalam satu cabang anggotanya berasal dari satu kelas . Semakin pure semakin cabang maka akan semakin baik.
- *Impurity* adalah ukuran purity suatu cabang.
- Salah satu kriteria impurity adalah Information Gain.
- Jadi dalam memilih atribut untuk pemecahan object ke dalam class-class harus dipilih atribut yang menghasilkan Information Gain yang paling besar.

Entropy

Entropi adalah nilai informasi yang menyatakan ukuran ketidakpastian(*impurity*) dari attribut dari suatu kumpulan obyek data dalam satuan bit.

$$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi$$

S : himpunan kasus
n: jml partisi S
Pi= proporsi dari Si terhadap S

Information Gain

- *Information Gain* adalah ukuran efektifitas suatu atribut dlm mengklasifikasikan data
- Digunakan untuk menentukan urutan atribut dimana attribut yang memiliki nilai *Information Gain* terbesar yang dipilih

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si)$$

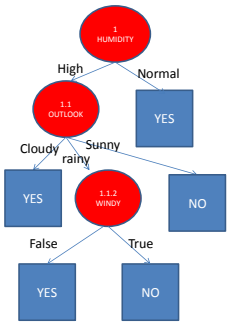
S: Himpunan kasus
A: Atribut
n : jml partisi atribut a
|Si| : jml kasus pada partisi ke-I
|S| : jml kasus dlm S

Decision Tree

- Sebuah ***Decision Tree*** adalah **struktur** yang dapat digunakan untuk membagi data yang besar menjadi himpunan-himpunan record yang lebih kecil dengan menerapkan serangkaian aturan keputusan. (Berry & Linoff)
- Proses pada ***Decision Tree*** adalah mengubah bentuk **data** (tabel) menjadi bentuk ***Tree***, Mengubah ***Tree*** menjadi ***Rule***, dan menyederhanakan ***Rule***(basuki&syarif,2003)

Decision Tree

- Tree merupakan struktur data yang terdiri dari simpul & rusuk. Simpul (root,branch,leaf)
- Algoritma yang digunakan untuk pembentukan ***Decision Tree*** diantaranya: **ID3,CART,C4.5**
- Algoritma **C4.5** merupakan pengembangan dari **ID3**



Contoh kasus rekomendasi Bermain Golf

NO	OUTLOOK	TEMPERATUR	HUMIDITY	WINDY	PLAY
1	Sunny	Hot	High	FALSE	NO
2	Sunny	Hot	High	TRUE	NO
3	Cloudy	Hot	High	FALSE	YES
4	Rainy	Mild	High	FALSE	YES
5	Rainy	Cool	Normal	FALSE	YES
6	Rainy	Cool	Normal	TRUE	YES
7	Cloudy	Cool	Normal	TRUE	YES
8	Sunny	Mild	High	FALSE	NO
9	Sunny	Cool	Normal	FALSE	YES
10	Rainy	Mild	Normal	FALSE	YES
11	Sunny	Mild	Normal	TRUE	YES
12	Cloudy	Mild	High	TRUE	YES
13	Cloudy	Hot	Normal	FALSE	YES
14	Rainy	Mild	High	TRUE	NO

Algoritma C4.5

1. Pilih atribut sebagai akar
2. Buat cabang untuk tiap-tiap nilai
3. Bagi kasus dalam cabang
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama

Rekomendasi bermain golf

NO	OUTLOOK	TEMPERATUR	HUMIDITY	WINDY	PLAY
1	Sunny	Hot	High	FALSE	NO
2	Sunny	Hot	High	TRUE	NO
3	Cloudy	Hot	High	FALSE	YES
4	Rainy	Mild	High	FALSE	YES
5	Rainy	Cool	Normal	FALSE	YES
6	Rainy	Cool	Normal	TRUE	YES
7	Cloudy	Cool	Normal	TRUE	YES
8	Sunny	Mild	High	FALSE	NO
9	Sunny	Cool	Normal	FALSE	YES
10	Rainy	Mild	Normal	FALSE	YES
11	Sunny	Mild	Normal	TRUE	YES
12	Cloudy	Mild	High	TRUE	YES
13	Cloudy	Hot	Normal	FALSE	YES
14	Rainy	Mild	High	TRUE	NO

Meringkas JML Kasus

NO	OUTLOOK	TEMPERATUR	HUMIDITY	WINDY	PLAY
1	Sunny	Hot	High	FALSE	NO
2	Sunny	Hot	High	TRUE	NO
3	Cloudy	Hot	High	FALSE	YES
4	Rainy	Mild	High	FALSE	YES
5	Rainy	Cool	Normal	FALSE	YES
6	Rainy	Cool	Normal	TRUE	YES
7	Cloudy	Cool	Normal	TRUE	YES
8	Sunny	Mild	High	FALSE	NO
9	Sunny	Cool	Normal	FALSE	YES
10	Rainy	Mild	Normal	FALSE	YES
11	Sunny	Mild	Normal	TRUE	YES
12	Cloudy	Mild	High	TRUE	YES
13	Cloudy	Hot	Normal	FALSE	YES
14	Rainy	Mild	High	TRUE	NO

		JML Kasus	NO	YES
TOTAL				
OUTLOOK				
	Cloudy			
	Rainy			
	Sunny			
TEMPERATUR				
	Cool			
	Hot			
	Mild			
HUMIDITY				
	High			
	Normal			
WINDY				
	FALSE			
	TRUE			

Menghitung Entropy Total

Node1 Node			JML Kasus	NO	YES	Entropy	Gain
1	TOTAL		14	4	10	0.863121	
	OUTLOOK						
		Cloudy	4	0	4		
		Rainy	5	1	4		
		Sunny	5	3	2		
	TEMPERATUR						
		Cool	4	0	4		
		Hot	4	2	2		
		Mild	6	2	4		
	HUMIDITY						
		High	7	4	3		
		Normal	7	0	7		
	WINDY						
		FALSE	8	2	6		
		TRUE	6	2	4		

$$\text{Entropy}(S) = \sum_{i=1}^{n} - pi * \log_2 pi$$

Entropy(Total)=(-4/14*$\log_2$(4/14)) -(10/14 * $\log_2$(10/14))

Entropy(Total)=0.863121

Menghitung Gain

Node			JML Kasus	NO	YES	Entropy	Gain
1	TOTAL		14	4	10	0.8631206	
	OUTLOOK						0.258521
		Cloudy	4	0	4	0	
		Rainy	5	1	4	0.7219281	
		Sunny	5	3	2	0.9709506	
	TEMPERATUR						
		Cool	4	0	4		
		Hot	4	2	2	1	
		Mild	6	2	4	0.9182958	
	HUMIDITY						
		High	7	4	3	0.9852281	
		Normal	7	0	7	0	
	WINDY						
		FALSE	8	2	6	0.8112781	
		TRUE	6	2	4	0.918296	

$$Gain(S,A)=Entropy(S)-\sum_{i=1}^{n}\frac{|Si|}{|S|}*Entropy(Si)$$

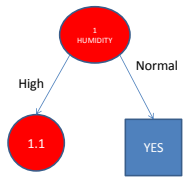
$$Gain(OutLook)=Entropy(Total)-\sum_{i=1}^{n}\frac{|OutLook|}{|Total|}*Entropy(OutLook)$$

Gain(OutLook)= 0.8631206 – ((4/14*0)+(5/14*0.722)+(5/14*0.97))

Gain(Total,OutLook)=0.258521

Memilih Atribut sebagai Akar

Node			JML Kasus	NO	YES	Entropy	Gain
1	TOTAL		14	4	10	0.863121	
	OUTLOOK						0.258521
		Cloudy	4	0	4	0	
		Rainy	5	1	4	0.721928	
		Sunny	5	3	2	0.970951	
	TEMPERATUR						0.183851
		Cool	4	0	4		
		Hot	4	2	2	1	
		Mild	6	2	4	0.918296	
	HUMIDITY						0.370507
		High	7	4	3	0.985228	
		Normal	7	0	7	0	
	WINDY						0.005978
		FALSE	8	2	6	0.811278	
		TRUE	6	2	4	0.918296	

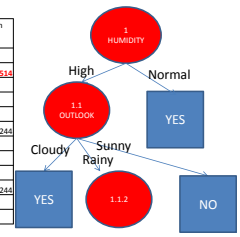


NO	OUTLOOK	TEMPERATUR	HUMIDITY	WINDY	PLAY
3	Cloudy	Hot	High	FALSE	YES
12	Cloudy	Mild	High	TRUE	YES
4	Rainy	Mild	High	FALSE	YES
14	Rainy	Mild	High	TRUE	NO
1	Sunny	Hot	High	FALSE	NO
2	Sunny	Hot	High	TRUE	NO
8	Sunny	Mild	High	FALSE	NO
7	Cloudy	Cool	Normal	TRUE	YES
13	Cloudy	Hot	Normal	FALSE	YES
5	Rainy	Cool	Normal	FALSE	YES
6	Rainy	Cool	Normal	TRUE	YES
10	Rainy	Mild	Normal	FALSE	YES
9	Sunny	Cool	Normal	FALSE	YES
11	Sunny	Mild	Normal	TRUE	YES

Memilih Atribut sebagai Node 1.1

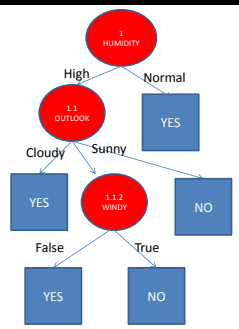
Node1.1

Node			JML Kasus	NO	YES	Entropy	Gain
1.1	HUMIDITY-HIGH		7	4	3	0.985228	
	OUTLOOK						0.699514
		Cloudy	2	0	2	0	
		Rainy	2	1	1	1	
		Sunny	3	3	0	0	
	TEMPERATUR						0.020244
		Cool	0	0	0		
		Hot	3	2	1	0.918296	
		Mild	4	2	2	1	
	WINDY						0.020244
		FALSE	4	2	2	1	
		TRUE	3	2	1	0.918296	



NO	OUTLOOK	TEMPERATUR	HUMIDITY	WINDY	PLAY
3	Cloudy	Hot	High	FALSE	YES
12	Cloudy	Mild	High	TRUE	YES
4	Rainy	Mild	High	FALSE	YES
14	Rainy	Mild	High	TRUE	NO
1	Sunny	Hot	High	FALSE	NO
2	Sunny	Hot	High	TRUE	NO
8	Sunny	Mild	High	FALSE	NO

Memilih Atribut sebagai Node 1.1.2



Node			JML Kasus	NO	YES	Entropy	Gain
1.1.2	HUMIDITY-HIGH &OUTLOOK RAINY		2	1	1	1	
	TEMPERATUR						0
		Cool	0	0	0	0	
		Hot	0	0	0	0	
		Mild	2	1	1	1	
	WINDY						1
		FALSE	1	0	1	0	
		TRUE	1	1	0	0	

Tugas

Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	hypermetrope	no	normal	soft

Klasifikasi- Pengukuran Kinerja

Data Mining

Pengukuran Kinerja *Klasifikasi*

- Pengukuran kinerja klasifikasi dilakukan dengan Confusion Matrix

Confusion Matrix

f_{ij}		class hasil klasifikasi (j)	
		Class=1	Class=0
Class asli (i)	Class=1	f_{11}	f_{10}
	Class=0	f_{01}	f_{00}

- $f_{11} + f_{00}$
  Jumlah data dari masing-masing class yang diprediksi secara benar.
- $f_{10} + f_{01}$
  Jumlah data dari masing-masing class yang diprediksi secara salah.

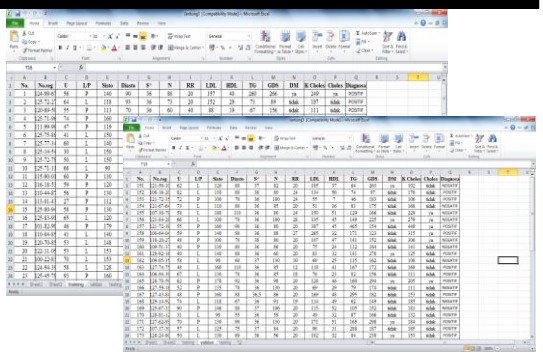
Confusion Matrix

f_{ij}		class hasil prediksi (j)	
		Class=1	Class=0
Class asli (i)	Class=1	f_{11}	f_{10}
	Class=0	f_{01}	f_{00}

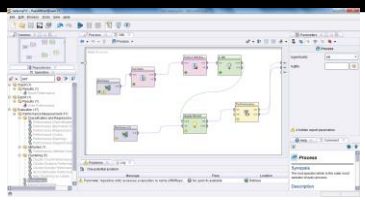
$$\text{Akurasi} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

$$\text{Laju Error} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

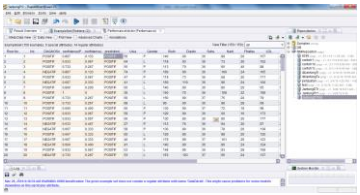
Contoh:



Proses KNN



Hasil



Akurasi