

ALGORITMA C4.5 UNTUK SIMULASI PREDIKSI KEMENANGAN DALAM PERTANDINGAN SEPAKBOLA

Marwana*)

Abstract : Abstract-This study is a simulation for memeriksa victory in a football game using the C4.5 data mining algorithm, the data is assumed to be the result of the previous game, then the data is processed by the principles of data mining algorithm C4.5, while the stages are used that determines the attribute-attributes, and entropy calculate the gain of each attribute, determine the root node, and the latter makes a decision tree.

Keyword; Simulation; Prediction; Data mining; Football; Algorithm C.45

PENDAHULUAN

Prediksi bola, sering sekali menjadi bahan perbincangan hangat dari masyarakat pecinta bola. Setiap pecinta bola pasti ingin sekali menyaksikan partai-partai bola yang enak ditonton dan seru. Apabila ketika team kesayangan kita yang akan bermain, pastinya hanya ada satu hasil yang diinginkan yaitu Menang.

Karena banyaknya komentar-komentar yang dilontarkan, banyak yang mengatakan prediksi bola pasti. Tidak ada satu orangpun yang dapat memprediksi bola pasti menang, pasti kalah atau pun pasti seri atau mengetahui apa yang terjadi sebelum pluit berbunyi tanda berakhirnya pertandingan.

Memang banyak faktor yang mempengaruhi dalam menentukan prediksi sepakbola, penelitian ini bertujuan untuk menyelesaikan masalah diatas yaitu mencoba menggunakan algoritma C4.5 untuk menggali informasi dari data pertandingan sebelumnya untuk melakukan prediksi pertandingan sepakbola.

TINJAUAN PUSTAKA

A. Pohon Keputusan

Di dalam kehidupan manusia sehari-hari, manusia selalu dihadapkan oleh berbagai macam masalah dari berbagai macam bidang. Masalah-masalah yang

dihadapi oleh manusia memiliki tingkat kesulitan dan kompleksitas yang sangat bervariasi, mulai dari masalah yang teramat sederhana dengan sedikit faktor-faktor yang terkait, sampai dengan masalah yang sangat rumit dengan banyak sekali faktor-faktor yang terkait dan perlu untuk diperhitungkan. Untuk menghadapi masalah-masalah ini, manusia mulai mengembangkan sebuah sistem yang dapat membantu manusia agar dapat dengan mudah mampu untuk menyelesaikan masalah-masalah tersebut. Adapun pohon keputusan ini adalah sebuah jawaban akan sebuah sistem yang manusia kembangkan untuk membantu mencari dan membuat keputusan untuk masalah-masalah tersebut dan dengan memperhitungkan berbagai macam faktor yang ada di dalam lingkup masalah tersebut. Dengan pohon keputusan, manusia dapat dengan mudah mengidentifikasi dan melihat hubungan antara faktor-faktor yang mempengaruhi suatu masalah dan dapat mencari penyelesaian terbaik dengan memperhitungkan faktor-faktor tersebut. Pohon keputusan ini juga dapat menganalisa nilai resiko dan nilai suatu informasi yang terdapat dalam suatu alternatif pemecahan masalah. Peranan pohon keputusan sebagai alat bantu dalam mengambil keputusan (*decision support tool*) telah dikembangkan oleh manusia sejak perkembangan teori pohon yang

dilandaskan pada teori graf. Kegunaan pohon keputusan yang sangat banyak ini membuatnya telah dimanfaatkan oleh manusia dalam berbagai macam sistem pengambilan keputusan.

Pohon dalam analisis pemecahan masalah pengambilan keputusan adalah pemetaan mengenai alternatif-alternatif pemecahan masalah yang dapat diambil dari masalah tersebut. Pohon tersebut juga memperlihatkan faktor-faktor kemungkinan/probabilitas yang akan mempengaruhi alternatif-alternatif keputusan tersebut, disertai dengan estimasi hasil akhir yang akan didapat bila kita mengambil alternatif keputusan tersebut.

Pohon keputusan adalah salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi oleh manusia. Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhirarki. Konsep dari pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan. Manfaat utama dari penggunaan pohon keputusan adalah kemampuannya untuk mem-*break down* proses pengambilan keputusan yang kompleks menjadi lebih simpel sehingga pengambil keputusan akan lebih menginterpretasikan solusi dari permasalahan. Pohon Keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target. Pohon keputusan memadukan antara eksplorasi data dan pemodelan, sehingga sangat bagus sebagai langkah awal dalam proses pemodelan bahkan ketika dijadikan sebagai model akhir dari beberapa teknik lain.

Kelebihan dari metode pohon keputusan adalah:

1. Daerah pengambilan keputusan yang sebelumnya kompleks dan sangat global, dapat diubah menjadi lebih simpel dan spesifik.

2. Eliminasi perhitungan-perhitungan yang tidak diperlukan, karena ketika menggunakan metode pohon keputusan maka sample diuji hanya berdasarkan kriteria atau kelas tertentu.
3. Fleksibel untuk memilih fitur dari internal node yang berbeda, fitur yang terpilih akan membedakan suatu kriteria dibandingkan kriteria yang lain dalam node yang sama. Kefleksibelan metode pohon keputusan ini meningkatkan kualitas keputusan yang dihasilkan jika dibandingkan ketika menggunakan metode penghitungan satu tahap yang lebih konvensional.
4. Dalam analisis multivariat, dengan kriteria dan kelas yang jumlahnya sangat banyak, seorang pengujian biasanya perlu untuk mengestimasi baik itu distribusi dimensi tinggi ataupun parameter tertentu dari distribusi kelas tersebut. Metode pohon keputusan dapat menghindari munculnya permasalahan ini dengan menggunakan kriteria yang jumlahnya lebih sedikit pada setiap node internal tanpa banyak mengurangi kualitas keputusan yang dihasilkan.

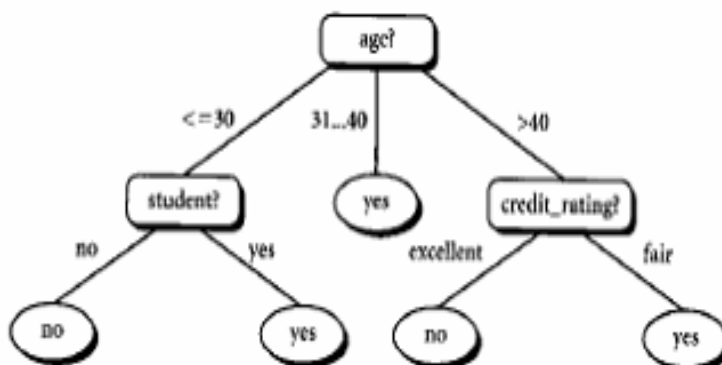
Kekurangan Pohon Keputusan

1. Terjadi overlap terutama ketika kelas-kelas dan kriteria yang digunakan jumlahnya sangat banyak. Hal tersebut juga dapat menyebabkan meningkatnya waktu pengambilan keputusan dan jumlah memori yang diperlukan.
2. Pengakumulasian jumlah eror dari setiap tingkat dalam sebuah pohon keputusan yang besar.
3. Kesulitan dalam mendesain pohon keputusan yang optimal.
4. Hasil kualitas keputusan yang didapatkan dari metode pohon keputusan sangat tergantung pada bagaimana pohon tersebut didesain.

Model Pohon Keputusan

Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur

berhierarchy. Contoh dari pohon keputusan dapat dilihat di Gambar 1 berikut ini.



Gambar 1. Model Pohon Keputusan (Pramudiono, 2008)

Disini setiap percabangan menyatakan kondisi yang harus dipenuhi dan tiap ujung pohon menyatakan kelas data. Contoh di Gambar 1. adalah identifikasi pembeli komputer, dari pohon keputusan tersebut diketahui bahwa salah satu kelompok yang potensial membeli komputer adalah orang yang berusia di bawah 30 tahun dan juga pelajar. Setelah sebuah pohon keputusan dibangun maka dapat digunakan untuk mengklasifikasikan *record* yang belum ada kelasnya. Dimulai dari *node root*, menggunakan tes terhadap atribut dari *record* yang belum ada kelasnya tersebut lalu mengikuti cabang yang sesuai dengan hasil dari tes tersebut, yang akan membawa kepada *internal node* (*node* yang memiliki satu cabang masuk dan dua atau lebih cabang yang keluar), dengan cara harus melakukan tes lagi terhadap atribut atau *node* daun. *Record* yang kelasnya tidak diketahui kemudian diberikan kelas yang sesuai dengan kelas yang ada pada *node* daun. Pada pohon keputusan setiap simpul daun menandai label kelas. Proses dalam pohon keputusan yaitu mengubah bentuk data (tabel) menjadi model pohon (*tree*) kemudian mengubah model pohon tersebut menjadi aturan (*rule*)

B. Algoritma C.45

Algoritma C4.5 yaitu sebuah algoritma yang digunakan untuk membangun *decision tree* (pengambilan keputusan). Algoritma C.45 adalah salah satu algoritma induksi pohon keputusan yaitu ID3 (*Iterative Dichotomiser 3*). ID3 dikembangkan oleh J. Ross Quinlan. Dalam prosedur algoritma ID3, input berupa sampel *training*, label *training* dan atribut. Algoritma C4.5 merupakan pengembangan dari ID3.

Pohon dibangun dengan cara membagi data secara rekursif hingga tiap bagian terdiri dari data yang berasal dari kelas yang sama. Bentuk pemecahan (*split*) yang digunakan untuk membagi data tergantung dari jenis atribut yang digunakan dalam *split*. Algoritma C4.5 dapat menangani data numerik (kontinyu) dan diskret. *Split* untuk atribut numerik yaitu mengurutkan contoh berdasarkan atribut kontinyu A, kemudian membentuk minimum permulaan (*threshold*) M dari contoh-contoh yang ada dari kelas mayoritas pada setiap partisi yang bersebelahan, lalu menggabungkan partisi-partisi yang bersebelahan tersebut dengan kelas mayoritas yang sama. *Split* untuk atribut diskret A mempunyai bentuk *value* (A) X dimana $X \subset domain(A)$. Jika suatu set data mempunyai beberapa pengamatan dengan *missing value* yaitu *record* dengan

beberapa nilai variabel tidak ada, Jika jumlah pengamatan terbatas maka atribut dengan *missing value* dapat diganti dengan nilai rata-rata dari variabel yang bersangkutan. Dalam melakukan pemisahan obyek (*split*) dilakukan tes terhadap atribut dengan mengukur tingkat ketidakmurnian pada sebuah simpul (*node*).

Secara singkat logika algoritma C4.5 yang digunakan adalah sebagai berikut:

- Pilih atribut sebagai akar
- Buat cabang untuk masing-masing nilai
- Bagi kasus dalam cabang
- Ulangi proses untuk masing-masing cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut sebagai akar, didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada. Untuk menghitung gain digunakan rumus sebagai berikut :

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

Dimana :

S : himpunan kasus

A : atribut

n : jumlah partisi atribut A

|S_i| : jumlah kasus pada partisi ke-i

|S| : jumlah kasus dalam S

$$Entropy(s) = \sum_{i=1}^n - p_i * \log_2 p_i \quad \text{Rumus} \quad \text{Sedangkan} \quad \text{untuk}$$

No.	Pelatih	Kandi
1	Pengertian	
2	Pengertian	
3	Pengertian	
4	Pengertian	
5	Pengertian	
6	Pengertian	
7	Pengertian	
8	Menekan	
9	Menekan	
10	Menekan	
11	Menekan	
12	Menekan	

$$\log_2(x) = \frac{\ln(x)}{\ln(2)}$$

menghitung entropi adalah sebagai berikut:

Dimana :

S : himpunan Kasus

n : jumlah partisi S

P_i : Jumlah kasus pada partisi ke-i

Kemudian menghitung perolehan informasi dari output data atau variabel dependent y yang dikelompokkan berdasarkan atribut A, dinotasikan dengan *gain* (y,A). Perolehan informasi, *gain* (y,A), dari atribut A relative terhadap output data y adalah:

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

Dimana :

S : himpunan kasus

A : atribut

n : jumlah partisi atribut A

|S_i| : jumlah kasus pada partisi ke-i

|S| : jumlah kasus dalam S

HASIL DAN PEMBAHASAN

Penelitian ini menggunakan data simulasi, diasumsikan bahwa data berasal dari hasil pertandingan sebelumnya untuk memprediksi menang atau kalah pada pertandingan sepakbola, dengan memakai data berikut ini :

Dari data diatas dihitung nilai *Entropy* dan *Gain* untuk masing-masing attribut, dengan menggunakan rumus yang telah disebutkan diatas, seperti pada tabel berikut ini

Node		Jumlah	Ya	Tidak	Entropy	Gain
1	Total	12	7	5	0,979868757	
	Pelatih					0,0005896
	Pengertian	7	4	3	0,985228136	
	menekan	5	3	2	0,970950594	
	Kandang Sendiri					0,01024609
	Ya	8	5	3	0,954434003	
	Tidak	4	2	2	1	
	Latihan					0,20944272
	Rutin	8	6	2	0,811278124	
	Jarang	3	1	2	0,918295834	
	Tidak Ada	1	0	1	0	
	Stamina					0,10434897
	Kuat	5	4	1	0,721928095	
	Lemah	7	3	4	0,985228136	
	Mental					0,61749246
	PD	8	7	1	0,543564443	
	Groggi	4	0	4	0	

Tabel 2. Perhitungan Node 1

Kunci pencarian *Entropy*

- Jika diantara kolom “Ya” atau “Tidak” ada yang bernilai 0 (nol) maka *entropy*-nya dipastikan juga bernilai 0 (nol)
- Jika kolom “Ya” dan “Tidak” mempunyai nilai yang sama maka *entropy*-nya dipastikan juga bernilai 1 (satu)

$$\text{Entropy (total)} = (-\frac{7}{12} \cdot \log_2 \frac{7}{12}) + (-\frac{5}{12} \cdot \log_2 \frac{5}{12}) = 0,979869$$

$$\text{Entropy (Pelatih (Pengertian))} = (-\frac{4}{7} \cdot \log_2 \frac{4}{7}) + (\frac{3}{7} \cdot \log_2 \frac{3}{7}) = 0,985228136$$

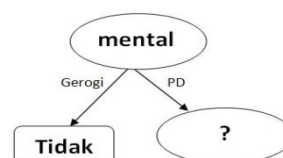
$$\text{Entropy (Pelatih (Menekan))} = (-\frac{3}{5} \cdot \log_2 \frac{3}{5}) + (\frac{2}{5} \cdot \log_2 \frac{2}{5}) = 0,970950594 \text{ ,,,,dst}$$

$$\text{Gain (Pelatih)} = 0,979869 - (\frac{7}{12} \cdot 0,985228136) + (\frac{5}{12} \cdot 0,970950594) = 0,0005896$$

Dst...

Dari hasil pada Tabel 2. dapat diketahui bahwa atribut dengan Gain tertinggi adalah **Mental**, sebesar 0,61749246. Dengan demikian, **Mental** dapat menjadi node akar. Ada dua nilai atribut dari **Mental** yaitu **PD** dan **Groggi**. Dari kedua nilai atribut tersebut, nilai atribut **Groggi** sudah mengklasifikasi kasus menjadi satu dengan keputusan **tidak**, sehingga tidak perlu dilakukan perhitungan lebih lanjut, tetapi untuk nilai atribut **PD** masih perlu perhitungan lagi.

Dari hasil tersebut dapat digambarkan pohon keputusan sementaraanya tampak seperti pada Gambar 2.



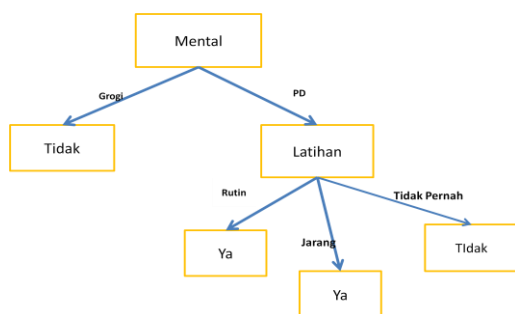
Gambar 2. Pohon Keputusan Hasil Perhitungan Node 1

Selanjutnya dihitung nilai *Entropy* dan *Gain* untuk masing-masing atribut untuk dapat menjadi node akar dari atribut **PD**, Seperti dibawah ini :

Node		Jumlah	Ya	Tidak	Entropy	Gain
1.1	Mental- PD	8	7	1	0,543564443	
	Pelatih					0,09235938
	Pengertian	5	4	1	0,721928095	
	menekan	3	3	0	0	
	Kandang Sendiri					0,05604763
	Ya	6	5	1	0,650022422	
	Tidak	2	2	0	0	
	Latihan					0,54356444
	Rutin	6	6	0	0	
	Jarang	1	1	0	0	
	Tidak Ada	1	0	1	0	
	Stamina					0,13792538
	Kuat	4	4	0	0	
	Lemah	4	3	1	0,811278124	

Tabel 3. Perhitungan Node 1.1

Dari hasil pada *Tabel 3*, dapat diketahui bahwa atribut dengan *Gain* tertinggi adalah **Latihan**, yaitu sebesar 0,54356444. Dengan demikian **Latihan** dapat menjadi node cabang dari nilai atribut **PD**. Ada tiga nilai atribut dari **Latihan**, yaitu **Rutin**, **Jarang** dan **Tidak Ada**. Dari tiga atribut tersebut, nilai atribut **Rutin** sudah mengklasifikasi kasus menjadi 1, yaitu keputusannya **Ya**, nilai atribut **Jarang** juga sudah mengklasifikasi kasus menjadi 1, yaitu keputusannya **Ya** serta nilai atribut **Tidak Ada** juga sudah mengklasifikasi kasus menjadi 1, yaitu keputusannya **Tidak**, sehingga tidak perlu dilakukan perhitungan lebih lanjut. Dengan demikian diperoleh pohon keputusan tampak seperti pada *Gambar 3*.



Gambar 3. Pohon Keputusan terakhir yang terbentuk

KESIMPULAN

Dari hasil penelitian ini dapat ditarik kesimpulan sebagai berikut :

1. Algoritma C4.5 dapat digunakan untuk memudahkan dalam pengambilan

keputusan dengan memproyeksikan data-data yang ada ke dalam bentuk pohon keputusan, berdasarkan nilai *entropy* dan *gain* yang dimiliki masing-masing atribut data.

2. Untuk hasil prediksi yang lebih akurat dibutuhkan data dalam jumlah besar, artinya semakin besar jumlah data yang digunakan maka semakin akurat hasil prediksi yang dihasilkan.

DAFTAR PUSTAKA

- Hanik, Umi. 2011. *Fuzzy Decision Tree* dengan Algoritma C4.5 pada Data Diabetes Indian Pima
- Kusrini dan Emha Taufiq Luthfi. 2009. *Algoritma Data Mining*. Penerbit Andi Offset, Yogyakarta.
- Sunjana. Klasifikasi Data Nasabah sebuah Asuransi Menggunakan Algoritma C4.5 Seminar Nasional Aplikasi Teknologi Informasi 2010 (SNATI 2010). Yogyakarta.
- Ariadni, Ratih dan Arieshanti Isye, Implementasi Metode Pohon Keputusan untuk Klasifikasi Data Dengan Nilai Fitur yang tidak pasti. Surabaya.

*) *Dosen Teknik Informatika STIMED Nusa Palapa Makassar, Indonesia*