Implementasi Algoritma C4.5 Pada Sistem Evaluasi Perijinan Pembukaan Program Studi

Ibnu Akil
Akademi Sekretari dan Manajemen BSI Jakarta
Jalan Raya Jatiwaringin no 18, Jakarta Timur
Ibnu.ial@bsi.ac.id

Abstract: C4.5 algorithm is one of machine learning algorithms which have been used widely for expert systems. C4.5 algorithm has been admitted and has a high accuracy in decision making cases based on classes and multiple attributes. The algorithm which is proposed by J. Ross Quinlan in 1993 will be implemented in one of expert system that is permission evaluation of new study program in Kemenristek Dikti.

Keyword: C4.5, artificial intelligent, expert system.

Abstrak: Algoritma C4.5 yang merupakan salah satu algoritma machine learning yang banyak digunakan untuk 80tatis-sistem pakar. Algoritma C4.5 sudah diakui dan memiliki ketepatan yang tinggi dalam memecahkan kasus-kasus pemilihan yang berbasiskan kelas-kelas dan multiple atribut-atribut. Algoritma yang diprakarsai oleh J. Ross Quinlan pada tahun 1993 ini akan dicoba untuk diimplementasikan dalam salah satu kasus system pakar yaitu; evaluasi perijinan pembukaan program studi baru di Kemenristek Dikti.

Kata kunci: C4.5, kecerdasan buatan, sistem pakar.

I. Pendahuluan

Dengan melonjaknya kebutuhan akan informasi yang cepat dan akurat, semakin meningkat juga kebutuhan akan artificial intelligent (AI) yang dapat menggantikan peran manusai sebagai decision maker tunggal. Dengan semakin banyaknya data yang akan diproses maka semakin menurun kemampuan manusia untuk melakukannya (memproses data tersebut).

AI yang dapat berperan sebagai suatu machine learning menjadi hal yang mutlak diperlukan untuk memerankan kepakaran suatu bidang ilmu dari manusia. Adalah algoritma C4.5 yang merupakan salah satu algoritma machine learning yang banyak digunakan untuk 80tatis-sistem pakar. Algoritma C4.5 sudah diakui dan memiliki ketepatan yang tinggi dalam kasus-kasus pemilihan yang berbasiskan kelas-kelas dan multiple atribut-atribut.

Dalam konteks ini penulis mencoba untuk mengimplementasikan algoritma C4.5 dalam studi kasus system evaluasi untuk menentukan kelulusan suatu proposal pembukaan program studi baru di Kementerian Riset, Teknologi dan Pendidikan Tinggi.

II. Landasan Teori

A. Artificial Inteligent

Menurut Russel & Norvig: “Selama ribuan tahun, kita telah mencoba untuk memahami bagaimana cara kita berpikir, yaitu bagaimana segelintir materi dapat melihat, memahami, memprediksi, dan memanipulasi dunia yang lebih besar dan lebih rumit dari dirinya sendiri. Bidang AI masih lebih jauh lagi, AI bukan hanya mencoba untuk memahami akan tetapi untuk membuat entitas yang pintar” (Akil).

Dalam bukunya Russel dan Norvig membagi definisi AI menjadi empat kategori, yaitu;

- Thinking Humanly; suatu usaha yang luar biasa untuk membuat bagaimana sebuah mesin dapat berpikir seperti layaknya manusia.
- Acting Humanly; sebuah seni dari membuat sebuah mesin yang menjalankan fungsi yang membutuhkan kecerdasan apabila dilaksanakan oleh manusia.
- Thinking Rationally; sebuah kajian tentang komputasi yang membuatnya menjadi mampu mempersepsikan, berpikir dan bertindak.
- Acting Rationally; adalah suatu kajian dari merancang agen (mesin) yang cerdas.

B. Expert System

 ISSN: 1410-5063

Menurut Ibnu Akil: "Selama beberapa 81tatis terakhir, Sistem Pakar telah menjadi aplikasi praktek yang utama dari riset AI. Dewasa ini, ada banyak system yang berguna dalam 81tatis setiap bidang operasional diseluruh dunia. Mulai dari gadget sederhana seperti handphone sampai robot-robot dalam 81tatisti manufaktur dan medis" (Akil).

C. ID3

ID3 adalah algoritma yang ditemukan oleh Ross Quinlan, yang digunakan untuk menghasilkan pohon keputusan dengan menggunakan konsep *information Entropy*. "Entropy adalah parameter statistic yang mengukur dengan cara tertentu, berapa banyak informasi yang dihasilkan dalam rata-rata untuk setiap huruf dari satu text dalam bahasa" (Shannon). Pada setiap node dari pohon keputusan tersebut, ID3 memilih atribut dari data yang paling efektif memisahkan kumpulan contoh-contohnya menjadi sub kumpulan yang diperkaya dalam satu kelas. Kriteria yang dipisahkan adalah normalisasi dari *information gain* (perbedaan didalam entropy). Atribut dengan hasil normalisasi information gain tertinggi dipilih untuk mengambil keputusan.

Entropy

Rumus entropy yang di usulkan oleh Shannon:

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

Dimana Entropie***(P)*** adalah entropy dari kumpulan data p, ***i*** adalah kumpulan kelas-kelas, ***n*** adalah jumlah data.

Information Gain

Rumus dari information gain adalah:

$$Gain(p,T) = Entropy(P) - \sum_{j=1}^{n} (Pj \times Entropy(Pj))$$

Dimana nilai Pj adalah sekumpulan nilai-nilai yang mungkin untuk atribut T. kita dapat menggunakan pengukuran ini untuk membuat rangking dari atribut-atribut dan menghasilkan pohon keputusan.

Algoritma ID3

Berikut adalah pseudo code dari algoritma ID3 (Squire).

```
Function ID3 (I, O, T) {
/* I is the set of input attributes
* O is the output attribute
* T is a set of training data
*
* function ID3 returns a decision tree
*/
if (T is empty) {
      return a single node with the
value "Failure";
}
if (all records in T have the same
value for O) {
      return a single node with that
value;
}
if (I is empty) {
      return a single node with the
value of the most frequent value of
      O in T;
      /* Note: some elements in this
node will be incorrectly classified */
}
      /* now handle the case where we
can't return a single node */
      compute the information gain for
each attribute in I relative to T;
      let X be the attribute with
largest Gain(X, T) of the attributes
in I;
      let {x_j| j=1,2, .., m} be the
values of X;
      let {T_j| j=1,2, .., m} be the
subsets of T when T is partitioned
      according the value of X;
      return a tree with the root node
81tatist X and
      arcs 81tatist x_1, x_2, .., x_m,
where the arcs go to the
      trees ID3(I-{X}, O, T_1), ID3(I-
{X}, O, T_2), .., ID3(I-{X}, O, T_m);
}
```

D. C4.5

Algoritma C4.5 masih diajukan oleh Ross Quinlan pada tahun 1993 untuk mengatasi kekurangan yang ada pada algoritma ID3.

Salah satu kekurangan dari ID3 adalah terlalu sensitive untuk bekerja dengan atribut yang memiliki banyak nilai. Hal ini perlu diatasi jika anda ingin menggunakan ID3 sebagai algoritma search engine di Internet. (Hssina, Merbouha and Ezzikouri). Contoh lain adalah nomor

pasien di database rumah sakit. Atribut seperti itu akan memberikan nilai maksimum yang memungkinkan dari *information gain*, sejak semua data training dapat diklasifikasikan secara tepat dengan menguji nilainya. Hal ini akan menghasilkan di pohon keputusan dimana semua node dibawah node root adalah node-node daun. Pohon ini, bagaimanapun, akan menjadi tidak berguna untuk mengklasifikasikan data baru. Tidak 82tatisti curva yang terkait dengan nilai dari atribut nomor pasien (Squire).

Dalam bukunya Quinlan menegaskan *key requirements* untuk algoritma C4.5 dapat bekerja yaitu (Quinlan):

1) **Attribute-value description**: data yang akan dianalisa haruslah dalam bentuk *flat-file* dimana semua informasi mengenai satu objek atau kasus harus dapat diekspresikan dalam bentuk kumpulan dari 82tatisti-properti atau atribut-atribut.
2) **Predefine-class**: adalah kategori-kategori kemana kasus akan ditugaskan harus sudah disediakan sebelumnya. Dalam terminology machine learning ini adalah termasuk dalam kategori supervised learning.
3) **Discrete classes**: adalah kebutuhan-kebutuhan yang berhubungan dengan kemana kasus akan ditugaskan. Kelas-kelas tersebut haruslah secara jelas digambarkan sebagai suatu kasus yang dimiliki atau tidak dimiliki oleh satu kelas tertentu.
4) **Sufficient data**: adalah generalisasi induktif yang diproses dengan mengidentifikasikan pola didalam data. Pendekatan yang digunakan jika valid, pola-pola yang kuat tidak bisa dibedakan dari kemungkinan hanya kebetulan. Dimana perbedaan ini biasanya bergantung pada pengujian 82tatistic efektif.
5) **Logical classification models**: adalah rancangan program hanya mengklasifikasikan apa yang bisa diekspresikan sebagai pohon keputusan atau sekumpulan rule.

Menurut (Quinlan) ID3 yang asli menggunakan kriteria yang disebut Gain, dijelaskan sebagai berikut. Informasi yang disampaikan melalui pesan bergantung pada probabilitasnya dan bisa dihitung dalam satuan bits sebagai minus dari logaritma dasar 2 dari probabilitas tersebut. Sebagai contoh jika ada 8 pesan yang memiliki kemungkinan yang sama, informasi yang disampaikan dari salah satunya adalah $-\log_2(1/8)$ atau 3 bits. Bayangkan memilih salah satu kasus secara acak dari satu set kasus ***S*** dan dinyatakan bahwa ia adalah milik dari beberapa class C_j. pesan ini memiliki probabilitas:

$$\frac{freq(C_j, S)}{S}$$

Dan informasi yang disampaikan adalah:

$$-log_2\left(\frac{freq(C_j, S)}{S}\right)$$

Untuk mencari informasi yang diharapkan dari pesan yang terkait satu kelas seperti itu, kita menjumlahkan semua kelas-kelas dalam proporsi mereka terhadap frekuensi mereka di dalam ***S***.

$$info(S) = -\sum_{j=1}^{k} \frac{freq(C_j, S)}{S} \times log_2\left(\frac{freq(C_j, S)}{S}\right)$$

Ketika diterapkan pada satu set kasus uji coba, $info(T)$ mengukur jumlah rata-rata yang diperlukan untuk mengidentifikasi kelas dari satu kasus ***T***. Atau disebut juga entropy dari set ***S***.

Sekarang coba pertimbangkan pengukuran yang sama setelah ***T*** telah dipartisi sesuai dengan sejumlah keluaran ***n*** dari test ***x***. kebutuhan informasi yang diharapkan dapat ditemukan sebagai bobot jumlah keseluruhan subset, sebagai

$$info_x(T) = \sum_{i=1}^{n} \frac{T_i}{T} \times info(T_i)$$

Sedangkan kuantitasnya adalah:

$$Gain(x) = info(T) - info_x(T)$$

Untuk menghindari informasi yang bias pada atribut-atribut yang masing-masing value-nya bersifat unik, seperti contoh atribut NIP pada data karyawan, mempartisi atribut tersebut akan menghasilkan banyak subset atribut, jadi informasi gain yang didapatkan dari atribut tersebut menjadi maksimal. Bias ini dapat diperbaiki dengan menyesuaikan gain yang didapatkan, menggunakan:

$$Split\ Info(x) = -\sum_{i=1}^{n} \left(\frac{T_i}{T}\right) \times log_2\left(\frac{T_i}{T}\right)$$

 ISSN: 1410-5063

Hal ini merepresentasikan informasi yang dihasilkan dari membagi ***T*** menjadi ***n*** subset. Dimana informasi gain diukur dengan menggunakan rasio secara proporsional menggunakan:

$$Gain\ Ratio(x) = \frac{Gain(x)}{SplitInfo(x)}$$

III. Pembahasan

A. Sistem Evaluasi Pembukaan Program Studi

System ini digunakan oleh Kemenristek Dikti untuk mengevaluasi proposal pengajuan pembukaan program studi baru yang diajukan oleh perguruan tinggi yang ingin menambah program studi di perguruan tingginya. System ini diawali ketika perguruan tinggi melakukan registrasi untuk pengajuan ijin secara online. Kemudian mengunggah berkas-berkas yang disyaratkan secara online. Setelah itu kemenristek dikti menugaskan kepada evaluator untuk menilai proposal tersebut apakah direkomendasi atau ditolak.

Table 3.1 Data Training

Skor	Range	Legalitas	Dosen	Class
136,821	<250	Lengkap	Memenuhi	No
28,289	<250	Belum Lengkap	Belum Memenuhi	No
251,094	>=250	Lengkap	Memenuhi	Yes
222,577	<250	Lengkap	Memenuhi	No
191,274	<250	Lengkap	Memenuhi	No
242,105	<250	Belum Lengkap	Belum Memenuhi	No
180,379	<250	Belum Lengkap	Belum Memenuhi	No
244,786	<250	Lengkap	Memenuhi	No
111,056	<250	Lengkap	Memenuhi	No
136,309	<250	Lengkap	Belum Memenuhi	No
249,821	<250	Lengkap	Memenuhi	No
205,091	<250	Lengkap	Memenuhi	No
167,909	<250	Belum Lengkap	Memenuhi	No
254,914	>=250	Lengkap	Memenuhi	Yes
56,3014	<250	Belum Lengkap	Memenuhi	No
158,146	<250	Belum Lengkap	Belum Memenuhi	No
296,827	>=250	Lengkap	Memenuhi	Yes
145,935	<250	Belum Lengkap	Memenuhi	No
195,366	<250	Lengkap	Belum Memenuhi	No
274,218	>=250	Belum Lengkap	Memenuhi	No
185,038	<250	Lengkap	Memenuhi	No
203,38	<250	Lengkap	Belum Memenuhi	No
260,109	>=250	Lengkap	Memenuhi	Yes
161,84	<250	Lengkap	Belum Memenuhi	No
274,012	>=250	Lengkap	Memenuhi	Yes
224,269	<250	Lengkap	Memenuhi	No
271,066	>=250	Lengkap	Memenuhi	Yes
233,375	<250	Lengkap	Memenuhi	No
227,493	<250	Lengkap	Memenuhi	No
253,066	>=250	Lengkap	Memenuhi	Yes

Table diatas adalah data training yang terdiri dari 30 item. Atribut skor adalah continues numeric value, agar tidak terjadi bias kita kelompokan dalam range >=250 dan <250. Sedangkan classes-nya adalah "Yes" dan "No". Dari 30 item class Yes = 7, dan No = 23.

1) Menghitung Entropy / Info

$$info(S) = -\sum_{j=1}^{k} \frac{freq(C_j, S)}{S} \times log_2\left(\frac{freq(C_j, S)}{S}\right)$$

$$info(S) = -\left(\frac{7}{30}\right) \times log_2\left(\frac{7}{30}\right) - \left(\frac{23}{30}\right) \times log_2\left(\frac{23}{30}\right)$$

$$= 0{,}7833$$

2) Menghitung Partisi / Subset Atribut Skor

$$info_x(T) = \sum_{i=1}^{n} \frac{T_i}{T} \times info(T_i)$$

Disini kita akan menggunakan atribut skor untuk membagi T menjadi dua subset yaitu <250 dan >=250.

Classes	Yes	No	Total
<250	0	22	22
>=250	7	1	8

$$info_x(T) = \frac{22}{30} \times \left(\frac{0}{22} \times log_2 \left(\frac{0}{22} \right) - \frac{22}{22} \times log_2 \left(\frac{22}{22} \right) \right)$$

$$+ \frac{8}{30} \times \left(-\frac{7}{8} \times log_2 \left(\frac{7}{8} \right) - \frac{1}{8} \times log_2 \left(\frac{1}{8} \right) \right)$$

$$= 0{,}1450$$

3) Menghitung Information Gain Atribut Skor

$$Gain(x) = info(T) - info_x(T)$$

$$Gain(x) = 0{,}7838 - 0{,}1450$$

$$= 0{,}6388$$

4) Menghitung Split Info Atribut Skor

$$Split\ Info(x) = -\sum_{i=1}^{n} \left(\frac{T_i}{T} \right) \times log_2 \left(\frac{T_i}{T} \right)$$

$$Split\ Info(x) = -\frac{22}{30} \times log_2 \left(\frac{22}{30} \right) - \frac{8}{30} \times log_2 \left(\frac{8}{22} \right)$$

$$= 0{,}8366$$

5) Menghitung Gain Ratio Atribut Skor

$$Gain\ Ratio(x) = \frac{Gain(x)}{SplitInfo(x)}$$

$$Gain\ Ratio(x) = \frac{0{,}6388}{0{,}8366} = 0{,}7636$$

Untuk atribut-atribut yang lain perhitungannya caranya sama seperti diatas.

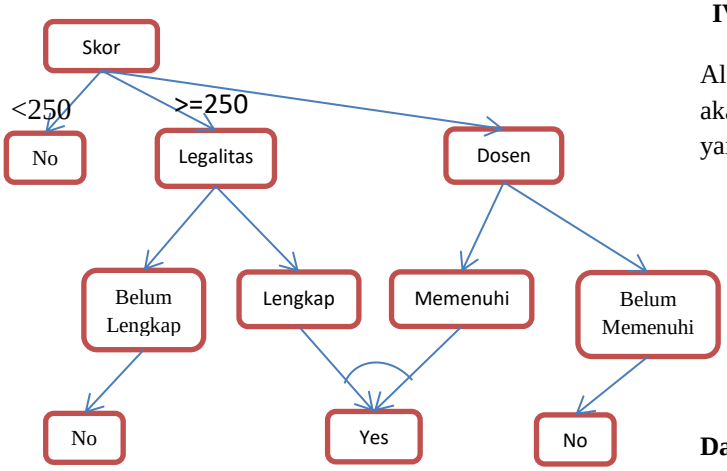
B. Hasil Runing Algoritma C4.5 di Program

```
attribute: skor
 value: <250,
       classes: No,
       counts: 22,
 value: >=250,
       classes: Yes, No,
       counts: 7, 1,

Information Gain:
0.6388264292981419
Split Info: 0.8366407419411673
Gain Ratio: 0.7635612243983503

---------------------------------
attribute: legalitas
 value: Lengkap,
       classes: No, Yes,
       counts: 15, 7,
 value: Belum Lengkap,
       classes: No,
       counts: 8,

Information Gain:
0.12202187343504978
Split Info: 0.8366407419411673
Gain Ratio: 0.14584739580328776

---------------------------------
attribute: dosen
 value: Memenuhi,
       classes: No, Yes,
       counts: 15, 7,
 value: Belum Memenuhi,
       classes: No,
       counts: 8,

Information Gain:
0.12202187343504978
Split Info: 0.8366407419411673
Gain Ratio: 0.14584739580328776
```

C. Decision Tree

Decision tree atau pohon keputusan dibentuk dari hasil running diatas dimana yang menjadi root node adalah atribut dengan gain ratio tertinggi, yaitu skor.

ISSN: 1410-5063



Gambar 3.1 Decision Tree

D. Decision Rule

Berikut adalah decision rule yang dihasilkan dari decision tree gambar 3.1.

```
If(skor>=250){
 If(legalitas==”Lengkap” &&
dosen==”Memenuhi”){
       Return “Yes”;
 }else{
       Return “No”;
   }
}else{
 Return “No”;
}
```

IV. Kesimpulan

Algoritma C4.5 bertindak sama seperti algoritma ID3 akan tetapi memperbaiki beberapa kekurangan dari ID3 yaitu:

1. Memungkinkan untuk menggunakan data yang continues.
2. Dapat memprediksi nilai yang hilang.
3. Mampu menggunakan atribut dengan bobot yang berbeda.
4. Memperbaiki pohon keputusan yang telah ada.

Daftar Pustaka

Akil, Ibnu. "Analisa Efektifitas Metode Forward Chaining Dan Backward Chaining." Pilar (2017): 35-42.

Hssina, Badr, et al. "A comparative study of decision tree ID3 and C4.5." International Journal of Advanced Computer Science and Applications (2014).

Quinlan, J. Ross. C4.5: Programs For Machine Learning. San Mateo: Morgan Kaufmann Publishers, 1993.

Shannon, Claude. E. "Prediction and Entropy of Printed English." Bell Sistem Technical Journal (1951).

Squire, David McG. "The ID3 Decision Tree Algorithm." CSE5230 Tutorial. 2004.