Forum Statistika dan Komputasi, April 2006, p:20-28 Vol. 11 No. 1
ISSN : 0853-8115

METODE KLASIFIKASI BERSTRUKTUR POHON DENGAN ALGORITMA CRUISE, QUEST, DAN CHAID

Yasmin Erika F.
Jurusan Teknik Mesin Politeknik Negeri Banjarmasin

Budi Susetyo dan Aam Alamudi
Departemen Statistika FMIPA IPB

RINGKASAN

Metode klasifikasi berstruktur pohon mulai banyak digunakan di berbagai bidang terutama karena hasilnya yang mudah diinterpretasikan. Tulisan ini mengangkat CRUISE sebagai metode pohon klasifikasi yang relatif baru, serta membandingkannya dengan dua metode serupa yang telah dikenal sebelumnya, yaitu CHAID dan QUEST. Data jamur tingkat tinggi (mushroom) genus Agaricus dan Lepiota digunakan untuk penerapan empat metode (CRUISE 1D, CRUISE 2D, QUEST, dan CHAID).

Analisis keempat metode tersebut menunjukkan bahwa peubah-peubah yang paling berkaitan dengan klasifikasi jamur yang 'dapat dimakan' atau 'beracun' adalah aroma dan warna sporanya. Untuk kasus ini, tampaknya CRUISE merupakan metode yang paling baik, dengan salah klasifikasi 0.0000 dan menghasilkan pohon terkecil. Berdasarkan metode ini, jamur yang dapat dimakan adalah jamur beraroma almond, adas, atau tidak beraroma, memiliki warna spora selain hijau, serta tidak ditemui secara bergerombol. Identifikasi jamur demikian cukup mendapat dukungan dari segi mikologi.

QUEST dan CHAID masing-masing menghasilkan salah klasifikasi 0.0014 dan 0.0028. Dari segi kecepatan proses, CRUISE 1D adalah yang tercepat. QUEST sedikit lebih lambat daripada CRUISE 1D untuk data berukuran besar dan peubah kategorik yang banyak, sedangkan CRUISE 2D memerlukan waktu pemrosesan (CPU time) paling lama. Dengan salah klasifikasi yang relatif kecil dan pohon yang pendek, CRUISE dapat menjadi alternatif yang baik bagi metode klasifikasi lainnya.

Kata kunci : Pohon klasifikasi, jamur

PENDAHULUAN

Penggunaan metode berstruktur pohon (*tree-structured methods*) sebagai metode klasifikasi telah menyebar luas di berbagai bidang dalam beberapa tahun terakhir, antara lain riset pemasaran (dalam hal segmentasi pasar), kedokteran (untuk diagnosis), ilmu komputer (untuk menyelidiki struktur data), botani (dalam hal klasifikasi), psikologi (teori pengambilan keputusan), dan linguistik.

Di antara kelebihan dari metode berstruktur pohon adalah dapat menghasilkan grafik pohon yang mudah diinterpretasikan, sifatnya yang fleksibel, non-parametrik dan nonlinear. Jika peubah responnya kategorik maka akan dihasilkan pohon klasifikasi. Algoritma pohon klasifikasi dapat dibagi ke dalam dua kelompok, yaitu yang menghasilkan pohon biner (misalnya CART dan QUEST) dan yang menghasilkan pohon non-biner (seperti CHAID, FACT, C4.5).

Agar dapat memberikan informasi yang bermanfaat, struktur pohon harus mudah dipahami dan pemilihan peubah tidak berbias. Jika pemilihan peubah berbias, peubah-peubah bebas akan memiliki peluang berbeda untuk dipilih sebagai peubah pemilah (*split*). Akibatnya, sulit diketahui apakah suatu peubah terpilih karena memang merupakan peubah penting (yang dapat menjelaskan peubah respon), ataukah terpengaruh oleh bias. Kim dan Loh (2001) mengembangkan algoritma yang dapat mengakomodasi kedua hal tersebut, yang dinamakan CRUISE (*Classification Rule with Unbiased Interaction Selection and Estimation*).

Tulisan ini akan menguraikan algoritma CRUISE dan membandingkannya dengan algoritma yang telah dikenal sebelumnya yaitu QUEST dan CHAID. Ilustrasi penggunaan ketiga juga diberikan dengan kasus identifikasi jamur besar.

TINJAUAN PUSTAKA

CHAID

Metode CHAID (*Chi-squared Automatic Interaction Detector*) yang diperkenalkan oleh Kass (1980) akan menghasilkan pohon non-biner. Split

yang dihasilkan berkisar antara dua sampai banyaknya jumlah kategori peubah split. Algoritma CHAID terutama sesuai untuk mengeksplorasi data berukuran besar. Toit *et al*. (1986) mengungkapkan bahwa CHAID tidak dapat diandalkan jika diterapkan pada data berukuran kecil. CHAID dapat mengakomodasi peubah respon ordinal maupun kontinu.

Pohon non-biner dapat dihasilkan oleh algoritma CHAID sebagai berikut:

1. Untuk setiap penduga X, cari pasangan kategori dari X yang memiliki nilai-p terbesar berdasarkan kelas peubah respon Y.
   - Jika Y kontinu, gunakan uji F.
   - Jika Y nominal, buat tabulasi silang dua-arah dengan kategori dari X sebagai baris dan kategori dari Y sebagai kolom. Gunakan uji khi-kuadrat Pearson.
   - Jika Y ordinal, buat model asosiasi Y. Gunakan uji nisbah kemungkinan.
2. Untuk pasangan kategori dari X dengan nilai-p terbesar, bandingkan nilai-p-nya dengan taraf α_{merge} yang telah ditentukan sebelumnya.
   - Jika nilai-p > α_{merge}, gabung pasangan ini ke dalam satu kategori baru.
   - Jika nilai-p ≤ α_{merge}, teruskan ke (3).
3. Hitung p^* (nilai-p-terkoreksi) untuk gugus kategori X dan kategori Y dengan menggunakan koreksi Bonferroni.
4. Pilih penduga X yang memiliki p^* terkecil. Bandingkan p^* tersebut dengan dengan taraf α_{split} yang telah ditentukan sebelumnya.
   - Jika $p^* \le \alpha_{split}$, node di-split berdasarkan gugus kategori X.
   - Jika $p^* > \alpha_{split}$, node tidak di-split. Node tersebut merupakan node akhir.

QUEST

QUEST (*Quick, Unbiased, Efficient Statistical Trees*) merupakan pengembangan dari FACT yang memiliki kecepatan tinggi (Loh dan Shih, 1997) dan menghasilkan pohon biner. QUEST menerapkan modifikasi analisis diskriminan kuadratik rekursif sebagai alternatif bagi metode-metode berstruktur pohon lain yang menggunakan pendekatan *exhaustive search*. QUEST tak berbias, namun tidak praktis untuk data berukuran sangat besar.

QUEST menangani pemilihan peubah dan pemilihan titik split secara terpisah. Algoritma pemilihan peubah split pada QUEST (SPSS Inc., 1998) dituliskan sebagai berikut.

1. Untuk setiap penduga X : jika X nominal, hitung nilai-p dari uji kebebasan khi-kuadrat Pearson antara X dan peubah respon Y. Jika X ordinal atau kontinu, gunakan uji F untuk menghitung nilai-p.
2. Bandingkan nilai-p terkecil dengan taraf α terkoreksi-Bonferroni.
   - Jika nilai-p ≤ α, maka pilih penduga yang bersesuaian untuk men-split node. Teruskan ke langkah (3).
   - Jika nilai-p > α, untuk setiap penduga X yang kontinu atau ordinal, gunakan uji Levene untuk ragam tak-homogen untuk menghitung nilai-p.
   - Bandingkan nilai-p terkecil dari uji Levene dengan taraf α terkoreksi-Bonferroni yang baru.
   - Jika nilai-p ≤ α, maka pilih penduga yang bersesuaian dengan nilai-p terkecil dari uji Levene untuk men-split node. Teruskan ke langkah (3).
   - Jika nilai-p > α, maka pilih penduga dari langkah (1) yang memiliki nilai-p terkecil (dari uji χ^2 maupun uji F) untuk men-split node. Teruskan ke langkah (3).
3. Misalkan X adalah penduga dari langkah (2). Jika X kontinu atau ordinal, teruskan ke langkah (4). Jika X nominal, X ditransformasi ke dalam peubah dummy, lalu diproyeksikan ke dalam koordinat diskriminan (*crimcoord*) terbesarnya.
4. Jika Y hanya memiliki dua kategori, teruskan ke langkah (5). Jika tidak, hitung rataan X untuk setiap kategori Y dan terapkan algoritma penggerombolan 2-mean terhadap masing-masing rataan untuk memperoleh dua kategori gabungan dari Y.
5. Lakukan analisis diskriminan kuadratik (QDA) untuk menentukan titik split.[1]

CRUISE

CRUISE merupakan pengembangan dari gabungan berbagai metode berstruktur pohon, terutama FACT, QUEST, dan GUIDE untuk pemilihan split, dan CART untuk pemangkasan (*pruning*). Algoritma ini menghasilkan dua sampai J split, di mana J adalah banyaknya kategori peubah respon. Metode ini memiliki sifat-sifat berikut: (1) pohon yang dihasilkan memiliki keakuratan pendugaan yang tinggi, (2) kecepatan komputasi tinggi, (3) bebas dari bias dalam pemilihan peubah, (4) sensitif terhadap interaksi lokal antar peubah, dan (5) keempat sifat di atas juga berlaku untuk data yang memiliki amatan hilang.

CRUISE dapat melakukan pemilahan tunggal maupun pemilahan kombinasi linear. Pemilahan kombinasi linear lebih fleksibel dan menghasilkan

[1] Prosedur lengkapnya dapat dilihat dalam Loh dan Shih (1997).

keakuratan pendugaan yang lebih baik, sehingga node akhir yang diperoleh juga lebih sedikit. Namun interpretasi pohon tidak mudah karena pemilahan kombinasi linear lebih sulit untuk dipahami. Dalam tulisan ini hanya diuraikan algoritma pemilahan tunggal.

Terdapat dua metode pemilihan peubah dalam pemilahan tunggal yang disebut metode 1D dan 2D. Ide pengembangan metode 1D diperoleh dari QUEST. Prosedur ini tak bias dalam pengertian bahwa jika penduga dan peubah respon saling bebas, setiap peubah memiliki peluang yang sama untuk terpilih.

Seperti pada QUEST, CRUISE menangani pemilihan peubah dan pemilihan titik split secara terpisah. Berikut algoritma Metode 1D.

Misalkan α adalah taraf kepercayaan yang dipilih (harga *default* adalah 0.05).

1. Untuk setiap penduga numerik, lakukan sidik ragam satu-arah dengan kategori Y sebagai perlakuan; kemudian hitung nilai-p dari uji-F. Misalkan X_{k1} memiliki nilai-p terkecil α_1.
2. Untuk setiap penduga kategorik, buat tabel kontingensi dengan nilai-nilai kategori sebagai baris dan nilai-nilai kelas (kategori peubah respon) sebagai kolom, dan hitung nilai-p dari uji χ^2. Misalkan nilai-p terkecil adalah α_2 dan peubah yang bersesuaian adalah X_{k2}.
3. Jika $\alpha_1 \le \alpha_2$, pilih peubah numerik X_{k1}; jika tidak pilih peubah kategorik X_{k2}. Anggap peubah terpilih dinamakan $X_{k'}$.
4. Jika $\min(\alpha_1,\alpha_2) < \alpha/K$ (koreksi Bonferroni pertama), maka pilih $X_{k'}$ sebagai peubah split.
5. Jika (4) tidak dipenuhi, hitung nilai-p untuk uji-F Levene untuk setiap peubah numerik. Misalkan $X_{k''}$ memiliki nilai-p terkecil $\overline{\alpha}$.
   - Jika $\overline{\alpha} < \alpha/(K+K_1)$, pilih $X_{k''}$ (koreksi Bonferroni kedua), di mana K = banyaknya penduga numerik dan K_1 = banyaknya penduga kategorik.
   - Jika tidak, pilih $X_{k'}$.
6. Untuk memilih titik split, teruskan ke langkah (7) jika peubah split tidak terpilih melalui uji Levene. Jika peubah split terpilih melalui uji Levene, lakukan transformasi Box-Cox (lihat Qu & Loh dalam Kim & Loh, 2001) pada peubah split yang terpilih. Jika X peubah kategorik, kategorinya terlebih dahulu diubah menjadi nilai *crimcoord* [2]. Jika ada nilai x yang negatif, tambahkan $2x_{(i+1)} - x_{(i)}$ pada nilai-nilai X. $x_{(i)}$ adalah statistik tataan ke-i pada X.
7. Lakukan analisis diskriminan linear (LDA). Jika X kategorik, setelah titik split diperoleh, nilai *crimcoord* diubah kembali menjadi kategori asal.

Metode 2D merupakan pengembangan masalah klasifikasi dari pendekatan Loh (2001) dalam mendeteksi interaksi berpasangan antar peubah pada pohon regresi. Algoritmanya adalah sebagai berikut.

Misalkan J_t = banyaknya kelas (kategori peubah respon) pada node t ; K = banyaknya penduga numerik; dan K_1 = banyaknya penduga kategorik.

1. Uji marjinal untuk setiap peubah numerik X. Misalkan X_{k1} memiliki nilai-p terkecil α_1.
   - Bagi data menjadi empat kelompok menurut kuartil contoh dari X.
   - Buat tabel kontingensi J_t x 4 dengan kelas sebagai baris dan kelompok sebagai kolom.
   - Hitung statistik χ^2 Pearson dengan derajat bebas $\nu = 3(J_t - 1)$.
   - Konversikan χ^2 menjadi nilai normal baku dengan transformasi Peizer-Pratt

$$z = \begin{cases} |W|^{-1}(W-1/3)\sqrt{(\nu-1)\log\left(\dfrac{\nu-1}{\chi^2}\right)+W}, & \nu > 1 \\ \sqrt{\chi^2}, & \nu = 1 \end{cases}$$

   di mana $W = \chi^2 - \nu + 1$.
   Misalkan z_n = maks $\{z_1, \ldots, z_{k1}\}$.
2. Uji marjinal untuk setiap peubah kategorik X. Misalkan C adalah banyaknya kategori pada X.
   - Buat tabel kontingensi J_t x C dengan kelas sebagai baris dan kategori C sebagai kolom.
   - Hitung statistik χ^2 Pearson dengan derajat bebas $\nu = (J_t - 1)(C - 1)$.
   - Gunakan transformasi Peizer-Pratt pada 1(d).

   Misalkan z_c = maks $\{z_{k+1}, \ldots, z_k\}$.
3. Uji interaksi untuk setiap pasang peubah numerik $(X_k, X_{k'})$:
   - Bagi ruang $(X_k, X_{k'})$ ke dalam empat kuadran menurut median contoh.
   - Buat tabel kontingensi J_t x 4 dengan kelas sebagai baris dan kuadran sebagai kolom.
   - Hitung statistik χ^2 Pearson dengan derajat bebas $\nu = 3(J_t - 1)$.
   - Gunakan transformasi Peizer-Pratt.

   Misalkan z_{nn} adalah nilai-z terbesar di antara $K_1 (K_1 - 1)/2$ nilai-z yang ada.
4. Uji interaksi untuk setiap pasang peubah kategorik : Jika pasangan peubah memiliki kategori sebanyak C_1 dan C_2, akan diperoleh tabel kontingensi J_t x C_1 C_2. Misalkan z_{cc}

[2] Prosedur ini merupakan modifikasi dari prosedur asli yang dipaparkan Gnanadesikan (dalam Loh & Shih, 1997)

adalah nilai-z terbesar di antara $(K - K_1)(K - K_1 - 1)/2$ nilai-z yang ada.

5. Uji interaksi untuk pasangan $(X_k, X_{k'})$ di mana X_k adalah peubah numerik dan $X_{k'}$ adalah peubah kategorik : Jika $X_{k'}$ memiliki C kategori, bentuk tabel kontingensi J_t x 2C. Misalkan z_{nc} adalah nilai-z terbesar di antara K_1 $(K - K_1)/2$ nilai-z yang ada.

   Misalkan f* adalah nilai bootstrap (lihat Kim & Loh, 2001) dan definisikan

   $$Z^* = \text{maks}\{f^* z_n, z_c, f^* z_{nn}, z_{cc}, z_{nc}\}.$$

   1. Jika $f^* z_n = z^*$, pilih peubah numerik dengan nilai-z terbesar.
   2. Jika $z_c = z^*$, pilih peubah kategorik dengan nilai-z terbesar.
   3. Jika $f^* z_{nn} = z^*$, pilih peubah numerik dalam pasangan yang nilai-z-nya lebih besar.
   4. Jika $z_{cc} = z^*$, pilih peubah kategorik dalam pasangan yang nilai-z-nya lebih besar.
   5. Jika $z_{nc} = z^*$, pilih peubah kategorik pada pasangan yang berinteraksi.

6. Untuk memilih titik split, lakukan transformasi Box-Cox pada peubah split yang terpilih. Jika X peubah kategorik, kategorinya terlebih dahulu diubah menjadi nilai *crimcoord*. Jika ada nilai x yang negatif, tambahkan $2x_{(i+1)} - x_{(i)}$ pada nilai-nilai X, di mana $x_{(i)}$ adalah statistik tataan ke-i pada X.
7. Lakukan analisis diskriminan linear (LDA). Jika X kategorik, setelah titik split diperoleh, nilai *crimcoord* diubah kembali menjadi kategori asal.

Tingkat salah klasifikasi metode 2D umumnya sama atau sedikit lebih baik dari metode 1D. Untuk data berukuran besar atau data dengan banyak peubah bebas, penerapan metode 2D memakan waktu CPU lebih lama daripada metode 1D karena algoritmanya yang lebih kompleks.

DATA

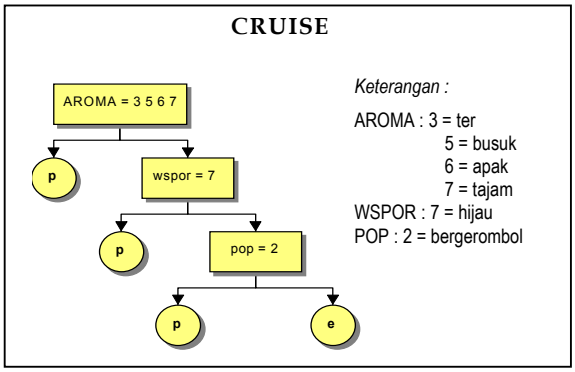
Data yang dianalisis adalah deskripsi contoh hipotetis dari jamur berinsang (*gilled mushroom*) dari genus Agaricus dan Lepiota yang diambil dari The Audubon Society Field Guide to North American Mushrooms (1981)[3]. Terdapat 8124 amatan di mana setiap spesies diidentifikasi sebagai "dapat dimakan" (*definitely edible*) atau "beracun atau tidak direkomendasikan untuk dimakan" (*definitely poisonous or of unknown edibility and not recommended*). Sebanyak 22 peubah bebas kategorik digunakan dalam identifikasi jamur beserta gambar bagian-bagian tubuh buah jamur dapat dilihat pada Lampiran 1.

Software yang digunakan untuk analisis data adalah program komputer CRUISE ver.1.09 (Kim & Loh, 2000) [4] untuk algoritma CRUISE dengan pemilahan tunggal (metode 1D dan 2D), SPSS Answer Tree v.2.0.1 untuk CHAID, dan QUEST (Loh & Shih, 1997)[5].

Options yang digunakan pada ketiga software adalah default, antara lain *estimated prior*, salah klasifikasi sama (*equal misclassification costs*). Pemilihan peubah split menggunakan $\alpha = 0.05$. Pemangkasan (*pruning*) dilakukan dengan validasi silang 10-lipat (*10-folds*), dan pohon yang dipilih adalah pohon 1-SE. Keterangan lebih jauh mengenai options ini dapat dilihat dalam Breiman *et al*. (1993) dan Statsoft, Inc. (2002).

HASIL DAN PEMBAHASAN

Cara penanganan amatan hilang oleh CHAID berbeda dengan CRUISE dan QUEST, karenanya hasil pengolahan terhadap data asli tidak dapat dibandingkan. Setelah amatan hilang disisihkan, sebanyak 5644 amatan dianalisis menggunakan keempat metode. Dari jumlah ini, 61.8% jamur (3488 amatan) diidentifikasi sebagai 'dapat dimakan', sedangkan sisanya (2156 amatan) beracun atau tidak dapat dimakan.



Gambar 1. Pohon CRUISE (1-SE).

Metode CRUISE 1D dan 2D memberikan hasil yang sama, dan peubah AROMA muncul sebagai peubah split pertama. Kedua metode CRUISE menghasilkan salah klasifikasi 0.000, yang berarti model atau pohon yang dihasilkan mampu

[3] Data diperoleh dari UC Irvine Machine Learning Repository (Blake & Merz, 1998)

[4] Program bisa diperoleh dari http://www.stat.wisc.edu/~loh/cruise.html [Juli 2002]

[5] Program CRUISE dan QUEST tersedia di http://www.stat.wisc.edu/~loh/ [Juli 2002]

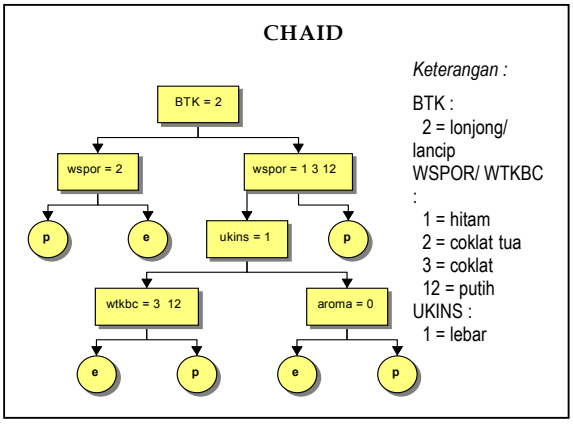
menempatkan seluruh amatan dalam klasifikasi yang benar.

Menurut CRUISE, peubah yang paling berpengaruh dalam membedakan jamur Agaricus/Lepiota yang dapat dimakan atau tidak adalah aromanya. Jamur yang beraroma ter, busuk, apak, dan tajam diklasifikasikan sebagai 'beracun atau tidak dapat dimakan'. Sedangkan jamur beraroma almond, adas, atau tidak beraroma belum tentu dapat dimakan; jika menemui jamur dengan aroma seperti ini, perlu diketahui informasi tambahan. Dari eksplorasi data diketahui bahwa jamur yang belum tentu dapat dimakan ini tidak beraroma; sehingga jamur yang juga diklasifikasikan sebagai beracun atau tidak dapat dimakan adalah jamur yang tidak beraroma dan memiliki warna spora hijau (WSPOR = 7).

Level ketiga pada pohon CRUISE memperlihatkan bahwa jamur tidak beraroma dan memiliki warna spora selain hijau serta ditemui secara bergerombol (POP = 2), diklasifikasikan sebagai beracun atau tidak dapat dimakan. Menurut kedua metode CRUISE, jamur yang diklasifikasikan sebagai 'dapat dimakan' adalah jamur yang beraroma almond, adas, atau tidak beraroma, memiliki warna spora selain hijau, serta tidak ditemui secara bergerombol (POP = 1, 3, 4, 5, 6).

Perbandingan Hasil

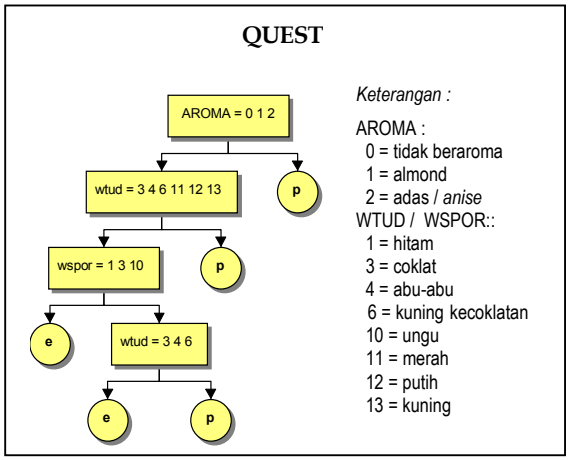
Empat metode yang digunakan menghasilkan dua peubah yang selalu muncul sebagai peubah split, yaitu AROMA dan warna spora (WSPOR). Dua metode CRUISE memberikan hasil yang sama dan peubah AROMA muncul sebagai peubah split pertama pada CRUISE dan QUEST. Ringkasan hasil pengolahan data dengan keempat metode dapat dilihat pada Lampiran 2.



Gambar 2. Pohon CHAID.

CHAID menghasilkan peubah split pertama yang berbeda, yakni bentuk tangkai (BTK). Peubah AROMA muncul pada level ke-4 bersamaan dengan warna tangkai bawah cincin (WTKBC), sedangkan peubah WSPOR muncul pada level ke-2. CHAID menghasilkan salah klasifikasi 0.0028, atau yang terbesar di antara keempat metode yang digunakan. Pohon CHAID lebih besar daripada pohon CRUISE dan QUEST, yaitu memiliki 13 node pada pohon akhir.

Menurut metode CHAID, jamur yang diklasifikasikan sebagai 'dapat dimakan' adalah jamur dengan bentuk tangkai lonjong/lancip (BTK = 2), dan memiliki warna spora hitam (WSPOR = 1), coklat (3), atau ungu (10). Atau jamur dengan bentuk tangkai membesar (BTK = 1), memiliki warna spora hitam (WSPOR = 1), coklat (3), atau putih (12), ukuran insang sempit (UKINS = 2), dan tidak beraroma. Jamur dengan bentuk tangkai membesar, memiliki warna spora hitam, coklat, atau putih, ukuran insang lebar (UKINS = 1), serta warna tangkai di bawah cincinnya coklat (WTKBC = 3) dan putih (12) juga diklasifikasikan sebagai 'dapat dimakan'.



Gambar 3. Pohon QUEST (1-SE).

Pengolahan dengan QUEST memunculkan peubah split WSPOR pada level ke-3, sedangkan peubah split ke-2 adalah warna tudung (WTUD). QUEST menghasilkan salah klasifikasi dua kali lebih kecil daripada CHAID, yakni 0.0014. Pohon QUEST sedikit lebih besar daripada pohon CRUISE, yaitu memiliki sembilan node pada pohon akhir.

Menurut metode QUEST, jamur yang diklasifikasikan sebagai 'dapat dimakan' adalah jamur beraroma almond, adas, atau tidak beraroma, memiliki tudung dengan warna coklat (WTUD = 3), abu-abu (4), kuning kecoklatan (6), merah (11), putih (12) atau kuning (13), serta memiliki warna spora hitam (WSPOR = 1), coklat (3), atau ungu (10). Atau jamur beraroma almond, adas, atau tidak beraroma, memiliki warna spora

hijau (WSPOR = 7) atau putih (12), serta memiliki tudung dengan warna coklat (WTD = 3), abu-abu (4), atau kuning kecoklatan (6).

Perbandingan algoritma

Langkah-langkah pemisahan. Pohon yang dihasilkan CRUISE (Gambar 1) ternyata adalah biner, serupa dengan pohon QUEST, karena hanya terdapat dua kelas peubah respon. CHAID juga menghasilkan pohon biner, namun dengan alasan yang berbeda. Ini dikarenakan nilai-p dari pasangan-pasangan kategori selalu lebih besar dari α_{merge}, sehingga pada akhirnya hanya tertinggal dua kategori baru dalam tiap peubah. Node akhir yang diperoleh pada umumnya 'murni', yaitu menunjukkan salah satu kelas peubah respon di mana tidak terdapat amatan yang termasuk kelas lainnya.

Pohon yang dihasilkan CRUISE 1D dan 2D dalam tulisan ini sama, karena walaupun algoritma pemilihan peubah split pada kedua metode agak berbeda, untuk data yang seluruh peubah penjelasnya kategorik metode 2D tidak melakukan bootstrap. Dalam hal pemilihan titik split, metode 1D tidak menerapkan transformasi Box-Cox (lihat Kim & Loh 2001), melainkan langsung melakukan analisis diskriminan linear (LDA).

QUEST mengolah data jamur ini dengan cara yang mirip dilakukan CRUISE 1D. Seluruh peubah penjelas yang kategorik dipetakan ke dalam nilai koordinat diskriminan terbesar, untuk kemudian diolah dengan analisis diskriminan kuadratik (QDA). Perbedaan inilah yang menyebabkan pohon hasil olahan QUEST berbeda dengan CRUISE, dan menyebabkan tingkat salah klasifikasinya lebih besar.

CHAID menghasilkan pohon yang relatif berbeda bila dibandingkan dengan ketiga metode lainnya. Algoritma CHAID tidak memisahkan pemilihan peubah dan titik split. Kategori-kategori peubah penjelas terlebih dahulu akan digabung, baru kemudian dihitung nilai-p terkoreksi untuk setiap peubah penjelas. Nilai-p terkecil menentukan peubah yang bersesuaian menjadi peubah split. Hal ini menyebabkan pemilihan peubah split yang berbeda dengan CRUISE maupun QUEST – dua metode takbias – dengan tingkat salah klasifikasi lebih besar daripada ketiga metode lainnya.

Waktu Pemrosesan. Waktu yang diperlukan CRUISE 2D untuk memproses data dalam tulisan ini relatif lambat bila dibandingkan dengan CRUISE 1D dan QUEST. Sebagai ilustrasi, pada komputer dengan prosesor Pentium III dan RAM 64 MB, data diolah CRUISE 1D dalam waktu 19 detik, QUEST dalam 126.17 detik, dan CRUISE 2D dalam 629 detik (10 menit 29 detik). Waktu pemrosesan dengan CHAID tidak diukur karena dilakukan dengan software lain dan dalam beberapa tahap terpisah.

Penelusuran algoritma memperlihatkan bahwa metode CRUISE 1D bisa selalu menjadi yang tercepat karena algoritmanya yang sederhana. Dengan peubah yang seluruhnya kategorik (dalam hal ini nominal), lima tahap pemilihan peubah split tereduksi menjadi hanya dua langkah sederhana saja. Penerapan QUEST memakan waktu lebih lama, karena dalam pemilihan titik split QUEST menerapkan analisis diskriminan kuadratik.

Waktu yang diperlukan CHAID tampaknya bisa sama atau lebih lama daripada CRUISE 1D karena harus menentukan kategori mana yang akan digabungkan dan menghitung nilai-p terkoreksi. Lamanya waktu pemrosesan yang diperlukan CRUISE 2D adalah hal yang wajar bila melihat algoritmanya yang cukup kompleks. Walaupun dalam mengolah data dalam tulisan ini ada langkah-langkah yang diabaikan (karena tidak ada peubah numerik), metode CRUISE 2D memerlukan 33.1 kali waktu lebih lama daripada metode CRUISE 1D karena harus melakukan uji marjinal dan transformasi Peizer-Pratt untuk setiap peubah, kemudian uji interaksi untuk setiap pasang peubah (dalam hal ini, (21 – 1) ! = 20! pasang peubah kategorik).

Tinjauan dari segi Mikologi

Dari ketiga pohon yang telah dibahas sebelumnya, hasil pohon CRUISE tampaknya paling mendekati kriteria beracun atau tidaknya jamur dari segi mikologi. Walau aroma saja tidak dapat menjadi petunjuk pasti apakah jamur dapat dimakan atau tidak, namun pada genus Agaricus/Lepiota petunjuk tersebut memegang peranan penting dalam mengenali jamur. Lain halnya dengan populasi jamur; beracun atau tidaknya jamur sulit bisa diketahui dari petunjuk ini, kecuali kita mengetahui spesies jamur yang bersangkutan.

KESIMPULAN

Algoritma CRUISE mengakomodasi dua hal penting dalam usaha memperoleh informasi yang bermanfaat dari data, yaitu struktur pohon yang mudah dipahami dan tidak adanya bias dalam pemilihan peubah. Interpretasi pohon yang lebih mudah diperoleh dengan membuat CRUISE menghasilkan pohon non-biner serta memilih peubah berdasarkan pengaruh satu-faktor dan dua-faktor.

Penerapan CRUISE pada data dalam tulisan ini menghasilkan salah klasifikasi terkecil yaitu 0.0000, diikuti oleh QUEST (0.0014) dan CHAID (0.0028). Pohon yang dihasilkan CRUISE adalah yang terkecil. Peubah split pertama yang dihasilkan CRUISE sama dengan yang dihasilkan QUEST, namun pada CHAID hasilnya berbeda.

Dari keempat metode yang diteliti, algoritma CHAID relatif berbeda dengan ketiga metode lainnya. Pengolahan secara statistik yang dilakukan CHAID hanyalah uji khi-kuadrat Pearson dengan teknik penggabungan kategori. Peubah penjelas numerik akan dikategorikan secara otomatis oleh CHAID. CRUISE dan QUEST menangani pemilihan peubah split dan titik split secara terpisah. Peubah kategorik akan dipetakan ke dalam koordinat diskriminan agar dapat diolah dengan analisis diskriminan dalam hal pemilihan titik split. Hal ini serta perbedaan peubah split yang dihasilkan menunjukkan bahwa CHAID dapat menghasilkan bias dalam pemilihan peubah, dan ini dapat mengubah interpretasi terhadap data.

Dari segi kecepatan pengolahan data, CRUISE 1D adalah yang tercepat di antara keempat metode yang dibicarakan. CHAID dapat menyamai kecepatan CRUISE 1D, QUEST sedikit lebih lambat untuk data berukuran besar dan peubah kategorik yang banyak, sedangkan CRUISE 2D memerlukan waktu pemrosesan (*CPU time*) paling lama.

Menurut keempat metode ini, peubah-peubah yang penting dalam mengklasifikasikan jamur sebagai 'dapat dimakan' atau tidak adalah aroma dan warna sporanya. Hal ini tidak jauh berbeda dengan dari segi mikologi, namun dalam pengonsumsian jamur tentu saja diperlukan kehati-hatian.

DAFTAR PUSTAKA

Alexopoulos CJ, Mims CW. 1979. *Introductory Mycology*. 3rd Ed. John Wiley, New York.

Blake CL, Merz CJ. 1998. UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Dept. of Information and Computer Science, University of California, Irvine, CA.

Breiman L, Friedman J, Olshen R, Stone C. 1993. *Classification and Regression Trees*. Chapman & Hall, New York.

Kass GV. 1980. An exploratory technique for investigating large quantities of categorical data. *Appl. Statist.* **29**: 119-127.

Kim H, Loh W.-Y. 2000. CRUISE User Manual. Revised ed. *Technical Report 989*. Dept. of Statistics, Univ. of Wisconsin, Madison.

Kim H, Loh W.-Y. 2001. Classification Trees with Unbiased Multiway Splits. *J. Am. Statist. Assoc*. **96**: 590-604.

Loh W.-Y. 2001. Regression Trees with Unbiased Variable Selection and Interaction Detection. *Statistica Sinica* **11**. http://www.stat.wisc.edu/~loh/

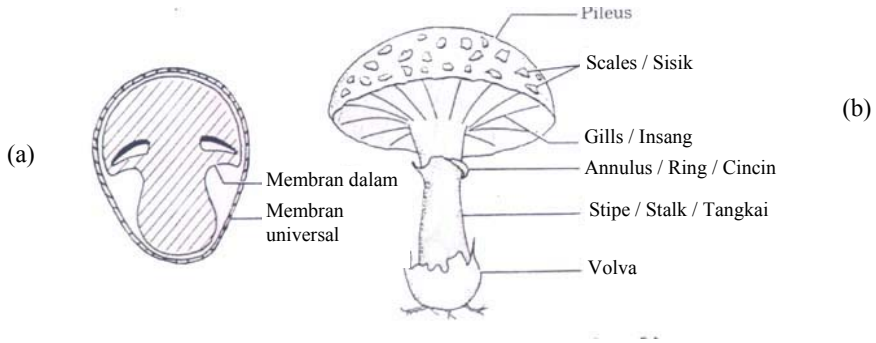
Loh W.-Y, Shih Y.-S. 1997. Split Selection Methods for Classification Trees. *Statistica Sinica* **7**: 815–840.

SPSS Inc. 1998. *AnswerTree™ 2.0 User's Guide*. SPSS Inc., Chicago, IL.

Statsoft, Inc. 2002. *Electronic Statistics Textbooks*. Statsoft, Tulsa, OK. http://www.statsoftinc.com/textbook/stathome.html. [Mei 2002]

Toit SHC du, Steyn AGW, Stumpf RH. 1986. *Graphical Exploratory Data Analysis*. Springer-Verlag, New York.

LAMPIRAN 1a. Bagian-bagian tubuh buah jamur



(a) Penampang tubuh buah (*basidiocarp*) jamur muda yang memperlihatkan membran pembungkus (*veil*);
(b) Tubuh buah jamur dewasa.

(Alexopoulos & Mims, 1979)

LAMPIRAN 1b. Peubah untuk data Jamur

Peubah	Deskripsi	Kategori
BTUD	Bentuk tudung (*cap*)	1=Lonceng, 2=kerucut, 3=cembung, 4=datar, 5=tombol (*knobbed*), 6=cekung
PMTUD	Permukaan tudung	1=Berserat, 2=berlekuk, 3=bersisik, 4=halus
WTUD	Warna tudung	3=Coklat, 4=abu-abu, 5=kekuningan, 6=kuning-kecoklatan, 7=hijau, 9=merah muda, 10=ungu, 11=merah, 12=putih, 13=kuning
GORES	Apakah jamur rusak / tergores?	1=Ya, 0=tidak
AROMA	Aroma jamur	1=Almond, 2=adas (*anise*), 3=ter kayu (*creosote*), 4=ikan/amis, 5=busuk, 6=apak, 7=tajam pedas (*pungent*), 8=pedas (*spicy*), 0=tidak beraroma
PLKINS	Pelekatan insang (*gill*)	1=Melekat, 2=*descending*, 3=*free*, 4=bertakik
KRINS	Kerapatan garis-garis insang	1=Dekat, 2=rapat, 3=jauh
UKINS	Ukuran insang	1=Lebar, 2=sempit
WINS	Warna insang	1=Hitam, 2=coklat tua, 3=coklat, 4=abu-abu, 5=kekuningan, 7=hijau, 8=jingga, 9=merah muda, 10=ungu, 11=merah, 12=putih, 13=kuning
BTK	Bentuk tangkai (*stalk*)	1=Membesar, 2=lonjong/lancip
BWTK*	Bentuk bagian bawah tangkai	1=Umbi, 2=*club*, 3=cangkir, 4=sama, 5=*rhizomorph*, 6=berakar
PTKAC	Permukaan tangkai di atas cincin (*ring*)	1=Berserat, 3=bersisik, 4=halus, 5=sutera
PTKBC	Permukaan tangkai di bawah cincin	1=Berserat, 3=bersisik, 4=halus, 5=sutera
WTKAC	Warna tangkai di atas cincin	3=Coklat, 4=abu-abu, 5=kekuningan, 6=kuning-kecoklatan, 8=jingga, 9=merah muda, 11=merah, 12=putih, 13=kuning
WTKBC	Warna tangkai di bawah cincin	3=Coklat, 4=abu-abu, 5=kekuningan, 6=kuning-kecoklatan, 8=jingga, 9=merah muda, 11=merah, 12=putih, 13=kuning
TPMEM	Tipe membran pembungkus	1=Parsial, 2=universal
WMEM	Warna membran pembungkus	3=Coklat, 8=jingga, 12=putih, 13=kuning
JCIN	Banyaknya cincin	0=Tidak ada, 1=satu, 2=dua
TPCIN	Tipe cincin	0=tidak ada, 1=Jaring laba-laba, 2=*evanescent*, 3=mengembang (*flaring*), 4=besar, 5=anting (*pendant*), 6=pelepah (*sheathing*), 7=*zone*
WSPOR	Warna cetakan spora	1=Hitam, 2=coklat tua, 3=coklat, 5=kekuningan, 7=hijau, 8=jingga, 10=ungu, 12=putih, 13=kuning
POP	Populasi jamur	1=Melimpah, 2=bergerombol, 3=banyak, 4=tersebar, 5=beberapa, 6=sendiri
HABITAT	Habitat jamur	1=Rumput, 2=dedaunan, 3=padang rumput, 4=jalan setapak, 5=perkotaan (*urban*), 6=sampah, 7=kayu

* terdapat amatan hilang sebanyak 2480 untuk peubah ini

LAMPIRAN 2. Ringkasan hasil pengolahan data

	Banyak amatan dengan klasifikasi benar	Salah klasi-fikasi	CPU Time (detik)	Banyaknya Node pada pohon akhir	Banyak Node akhir	Peubah split ke-			
						1	2	3	4
CRUISE 1D	5644	0.0000	19.00	7	4	AROMA	WSPOR	POP	–
CRUISE 2D	5644	0.0000	629.00	7	4	AROMA	WSPOR	POP	–
QUEST	5636	0.0014	126.17	9	5	AROMA	WTUD	WSPOR	WTUD
CHAID	5628	0.0028	n/a	13	7	BTK	WSPOR	UKINS	AROMA WTKBC