Pembandingan Stabilitas Algoritma Seleksi Fitur menggunakan Transformasi Ranking Normal

Taufik Djatna
Grad.School of Engineering, Dept. Information Engineering –Hiroshima University
Kagamiyama 1-7-1 Higashi Hiroshima 739-8521

dan
Dept. Teknologi Industri-FATETA IPB
Po Box 220 Kampus Darmaga Bogor 16002
taufikdjatna@ipb.ac.id

Yasuhiko Morimoto
Grad.School of Engineering, Dept. Information Engineering –Hiroshima University
Kagamiyama 1-7-1 Higashi Hiroshima 739-8521 Japan
morimoto@mis.hiroshima-u.ac.jp

Abstract: We address the need for evaluating the ranking robustness on different classifiers in feature selection algorithm. We propose a normalized rank transformation metric to compare the stability on classifying target class on training and testing dataset. Using stability comparison is a promising effort to improve the choice decision on deploying any feature selection algorithms. We also show relationship between stability and scalability.

Keyword: feature selection, normalized rank transformation

I. Pendahuluan

Seleksi fitur adalah salah teknik terpenting dan sering digunakan dalam pre-processing data mining [1], khususnya untuk *knowledge discovery* maupun *discovery science*. Teknik ini mengurangi jumlah fitur yang terlibat dalam menentukan suatu nilai kelas target, mengurangi fitur irelevan, berlebihan dan data yang menyebabkan salah pengertian terhadap kelas target yang membuat efek segera bagi aplikasi. Hasilnya, aplikasi data mining bias dipercepat, mempertinggi kinerja mining seperti akurasi peramalan. Seleksi atau seleksi fitur merupakan proses yang memilih subset dari ukuran fitur asalnya. Tugas utama dalam seleksi fitur adalah menentukan fitur mana yang dipilih dan dipakai dalam rangka peramalan atribut atribut fitur. Fitur dianggap relevan bila nilainya bervariasi secara sistematis dengan keanggotaan kategori [2]. Dengan demikian algoritma seleksi fitur mempunyai peran kritis dalam banyak aplikasi machine learning, CRM, data mining secara umum dan tentu saja dalam analisis genomic.

Dengan sedemikian luasnya bidang aplikasi yang memerlukan keunggulan algoritma seleksi fitur, sangat penting untuk membuat kerangka evaluasi peringkat (rank) berbasiskan pada kesamaan (similarity) antara klasifier yang dipakai sebagai inti algoritma seleksi fitur. Pada dasarnya subset yang akan didapat dari seleksi fitur merupakan urutan berperingkat dari semua fitur yang dipilih oleh masing masing algoritma. Dalam paper ini kami mengusulkan suatu cara untuk mengevaluasi kestabilan pemilihan fitur tersebut berbasis pada asumsi bahwa seleksi fitur melakukan kuantifikasi perankingan stabilitas yang dinamakan sebagai S_R . Untuk itu, kami usulkan sebagai satu ukuran yang memanfaatkan transformasi peringkat kestabilan normal (*normalized rank transformation*) serta mengevaluasi keakuratan klasifikasinya menggunakan algoritma klasifikasi C4.5. Diharapkan cara ini juga berguna untuk lebih jauh memahami tentang pentingnya kestabilan dalam algoritma seleksi seleksi sehingga selanjutnya dapat digunakan sebagai dasar pengambilan keputusan algoritma mana yang cocok untuk satu domain masalah.

II. Motivasi

Kami mendefinisikan suatu metrik-pengukuran untuk mengevaluasi kestabilan yang tangguh (*robust*) bagi algoritma seleksi fitur, dengan motivasi dalam dua hal:

1) Mempercepat pengambilan pilihan penggunaan algoritma seleksi fitur pada domain permasalahan tertentu. Bila mempertimbangkan memilih algoritma seleksi fitur tertentu, sementara tidak ada dasar pengetahuan yang bisa dipakai secara eksak tentang perilaku data-algoritma berdasarkan kesesuaian (*fitness*). Mterik stabilitas perankingan ini bisa dipakai pada permasalahan yang melibatkan banyak atribut atau fitur (*multi attributes problem*).
2) Menjamin kestabilan pola seleksi subset yang dipilih pada skala yang berbeda dari seluruh record data. Secara teoritis dapat dinyatakan bahwa algoritma seleksi fitur mestinya melingkupi kestabilan dalam

skala ukuran manapun dalam database. Untuk menjamin prasyarat ini, algoritma tersebut mestinya dilengkapi juga dengan kemampuan mengukur kestabilan hasil subset terpilih antara data training dan testingnya.

III. Latar Belakang Masalah Seleksi Fitur

Pada bagian ini algoritma yang dipakai dalam seleksi fitur dibahas secara singkat. Seleksi fitur, kita bisa deskripsikan dengan cara formal sebagai berikut: suatu masalah dengan banyak fitur $f_i \in$ n dengan F={f_1,f_2,..,f_k}, bila fitur bernilai riil (R) bisa dinyatakan sebagai satu himpunan contoh subset V={v_1,v_2,..v_n} dengan n < k merupakan subset kelas C dengan klasifier K: $R^k \rightarrow$ C didefinisikan sebagai:

$$\forall_{vi} \in V_{i,j} \in (1,..,k),\ v_{i,j} \in f_j \qquad (1)$$

Tugas algoritma seleksi fitur adalah untuk menginduksi hipotesis klasifier yang pada akurasi tertentu mampu memprediksi label nilai dari contoh baru [3]. Pembelajaran klasifier dapat diturunkan dengan nilai fitur yang tertentu. Misalkan G adalah sembarang subset F dan f_G adalah vector nilai G. secara umum, tujuan seleksi fitur dapt diformalkan sebagai (P(C|F=f)), dimana (P(C|G=f_G)) merupakan distribusi peluang dari berbagai kelas nilai fitur dalam G and ((P(C|F=f)) merupakan distribusi asal nilai fitur dalam F. Asumsi umum yang dipakai adalah bahwa fitur yang berguna bagi kelas target adalah jika dan hanya jika berkolerasi dengan kelas target; sebaliknya fitur diangap irelevan. Dengan kata lain dalam seleksi fitur secara umum, subset fitur yang baik sangat berkolerasi dengan target kelas. Seleksi fitur untuk tugas klasifikasi dalam machine learning bias dijalankan berbasis pada korelasi antar fitur. Rasionalnya adalah bahwa fitur-fitur relevan bila nilainya bervariasi secara sistematis dengan keanggotaan kategori kelas target.

Dalam paper ini teknik seleksi fitur yang dipakai dalam evaluasi kestabilan hasil adalah metode Information gain (IG) [4], Gain ratio (GR) [5], Chi-Square (CHI) [6], Symmetrical Uncertainty (SU) [5], ReliefF [3] dan Correlation-based Feature Selection (CFS) [2]. IG, GR, CHI dan SU merupakan teknik seleksi fitur yang memakai metode scoring untuk nominal ataupun pembobotan atribut kontinus yang didiskretkan menggunakan maksimasi entropy [6]. ReliefF mengirimkan pembobotan fitur saat melakukan pengambilan frekuensi interaksi antar fitur dalam contoh dalam data training.

3.1. Kriteria Klasifikasi Teknik Seleksi Fitur

Kriteria klassifikasi untuk kegunaan penentuan fitur dapat didefinisikan sebagai fungs*i* $x(S)=x_1, x_2,..x_k$ bagi suatu kaidah klasifikasi dimana atribut target memiliki nilai beda sebanyak k (a_1,a_2,..a_k). Berikut adalah criteria klasisikasi yang dipakai dalam paper ini didefinisikan secara ringkas. Anggap R adalah himpunan semua record dalam database and R = S $\cup$ Š, dimana Š adalah komplementer S. Anggap p(S) merupakan peluang target nilai ke-i dalam himpunan S, i.e xi(S)/(|S|). Diketahui $x_i(S)$ merupaka jumlah record dalam S yang memiliki nilai target sebagai nilai ke-i dalam domain target.

Definisi 1 : Fitur terpilih berdasarkan kriteria seleksi fitur tertentu adalah fitur yang berasal dari subset keseluruhan atribut dalam dataset yang memiliki skor lebih besar dari batas minimal. Berdasarkan definisi ini maka perankingan yang dilakukan berdasarkan hasil skor masing masing teknik seleksi fitur dan bersifat unik terhadap metode yang dipakai.

3.2 Fungsi Entropy Gain

Fungsi Entropy gain berikut membandingkan infomasi yang didapat oleh segmentasi S. Ukuran ini menunjukkan berapa banyak informasi yang diberikan oleh segmentasi S.

$$Ent(x(S)) = Ent(S;\hat{S}) = -\sum_i^2 p_i(R)\log_2 p_i(R) + \frac{|S|}{|R|}\sum_i^2 p_i(S)\log_2 p_i(S) + \frac{|\hat{S}|}{|R|}\sum_i^2 p_i(\hat{S})\log_2 p_i(\hat{S}) \qquad (2)$$

Dimana p_i merupakan nilai harapan per porsi S atau Š terhadap R

3.3 Korelasi χ^2 – Chi Squared

Fungsi berikut menunjukkan bagaimana kuatnya hipotesa statistik bahwa nilai S dan Š yang tidak berbeda dengan R dapat diabaikan. Mutu kaidah klasifikasi dengan Chi square menggunakan segmentasi informasi S dan Š (pada kasus klasifikasi), semakin tinggi nilai fungsi tujuannya, semakin baik mutu kaidah yang dibangun berdasarkan criteria tertentu. Formula yang dipakai pada klasifikasi dua segmentasi adalah sebagai berikut:

$$\chi^2(x(S)) = \chi^2(x(S;\hat{S})) = \sum_{i=1}^2 \frac{|S|(p_i(S) - p_i(|\hat{S}|(p_i(|\hat{S}|) - p_i(|R|)^2)}{p_i(|R|} \qquad (3)$$

3.4 Symmetrical Uncertainty

Suatu model probabilistic yang memiliki nilai fitur Y bisa dibentuk melalui meramalan probabilitas individual nilai y $\in$ Y dari data training. Apabila model ini digunakan untuk memprakirakan nilai Y dari suatu sample baru yang diambil dari distribusi sumber data training yang sama, maka entropy model (juga atributnya) merupakan jumlah bit yang akan diambil, rata-rata untuk memperbaiki output model.

Entropy adalah ukuran ketidakpastian dalam sistem. Dimana bentuk dasar entropy adalah:

$$H(Y) = -\sum_{y \in Y} p(y)\log_2(p(y)) \quad (4)$$

Bila nilai Y yang diamati dalam data training dipartisi berdasarkan nilai fitur kedua X, dan entropy Y berdasarkan pada partisi yang diinduksi oleh X akan lebih kecil daripada entropy Y sebelum dipartisi. Hubungan yang terdapat antara fitur Y dan X diformulasi sebagai:

$$H(Y \mid X) = \sum_{x \in X} p(x) \sum_{y \in Y} p(y \mid x)\log_2(p(y \mid x)) \quad (5)$$

Jumlah entropy pada Y menurun mencermin penambahan informasi tentang Y yang disediakan oleh X dan dinamakan sebagai Information Gain [7]. Sebagai alternatif juga dinamai sebagai mutual information yang memiliki formula modifikasi sebagai berikut :

$$\text{Gain}=H(Y)-H(Y|X) = H(X) - H(X|Y) = H(Y) + H(X) - H(X,Y) \quad (6)$$

Information gain merupakan ukuran simetris yaitu sebagai jumlah informasi yang didapat tentang Y setelah mengamati X, setara dengan jumlah informasi yang didapat tentang X setelah mengamati Y. Sifat simetri diinginkan bagi pengukuran inter-korelasi fitur ke fitur. Sayangnya, information gain memiliki bias terhadap atribut. Dari hal inilah ketidakpastian simetris (Symmetrical Uncertainty) dikenalkan sebagai ukuran dengan nilai normal dalam rentang [0,1] dan ditetapkan sebagai

$$SU = 2.0\left(\frac{Gain}{H(Y)+H(X)}\right) \quad (7)$$

3.5 ReliefF

ReliefF [3] merupakan algoritm pembobotan fitur yang sensitive terhadap interaksi fitur. Pembobotan aproksimasi menghitung beda probabilititas berikut bagi bobot fitur X:

$$\mathbf{W}_x = P(\text{Beda Nilai } X \mid \text{Contoh terdekat kelas yang berbeda}) - P(\text{Beda Nilai } X \mid \text{Contoh terdekat kelas yang sama}) \quad (8)$$

Dengan menghilangkan sensitivitas konteks yang disediakankondisi contoh terdekat , atribut dilatih secara independent terhadap satu dan lainnya. Formulasi ReliefF adalah sebagai berikut:

$$\text{Re}lief_x = \left(\frac{Gini' \sum_{x \in X} p(x)^2}{1 - \sum_{c \in C} p(c)^2 \sum_{c \in C} p(c)^2}\right) \quad (9)$$

Dimana C merupakan variabel kelas dan

$$Gini' = \left(\sum_{c \in C} p(c)(1-p(c))\right) - \sum_{x \in X}\left(\frac{p(x)^2}{\sum_{x \in X} p(x)^2} \sum_{c \in C} p(c \mid x)(1-p(x))\right) \quad (10)$$

Gini' merupakan modifikasi fitur mutu atribut lain yang dinamai Gini index [8], yang digunakan dalam pembuatan classification and regression tree (CART). Baik Gini Index and Gini' sama dengan information gain dalam hal bias terhadap kepentingan atribut yang memiliki nilai banyak. Agar bisa menggunakan Relief secara simetris pada dua fitur, pengukuran bisa dihitung dua kali—masing masing fitur diperlakukan dalam dua kelas--- untuk kemudian hasilnya dirata-ratakan. Dengan demikian setiap kali ReliefF dipakai maka sudah dijamin sifat simetrisnya.

*3.6 Correlation based Feature Selection **(CFS)***

Sebagai inti dalam CFS adalah teknik heuristic untuk mengevaluasi nilai atau harga subset fitur [2]. Teknik ini mempertimbangkan kegunaan fitur individual bagi prakiraan label kelas dengan level interkorelasi di antara fitur-fitur. Fitur secara individual menguji mana ukuran yang berkaitan dengan variable yang diamati (sebagai kelas target). Persamaan berikut adalah formalisasi nilai harga heuristic yang dimaksud:

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)r_{ff}}} \quad (11)$$

Dimana $Merit_s$ merupakan harga heuristic subset fitur S yang berisi k fitur r_{cf} yang merupakan rata-rata korelasi fitur-kelas, r_{ff} adalah rata-rata interkorelasi fitur ke fitur. Pada kenyataannya semua variable distandardisasi sesuai rumus korelasi Pearson. Numerator dianggap telah dipahami sebagai indikasi bagaimana sifat prediksi suatu fitur kelompok, sedangkan denominator menunjukkan bagaimana redundansi data antara fitur.

Definisi 2: Perbandingan kestabilan data training dan testing—Dataset D_T sebagai data training dan D_S sebagai data testing dipartisi menjadi densitas yang sama untuk diuji menggunakan algoritma seleksi fitur. Pada densitas tersebut dipilih fitur yang paling penting menggunakan skor masing masing algoritma untuk melihat stabilitas fitur terpilih pada densitas data yang diset.

IV. Evaluasi Perankingan yang *Tangguh (Robust)*

Pada bagian ini, evaluasi terhadap kinerja perankingan dilakukan dengan focus pada kesamaan (similarity) hasil seleksi fitur khususnya dipandang dalam kaitannya dengan urutan kestabilan ranking terhadap sejumlah besar data yang dikelola. Kestabilan yang dimaksud di sini mencerminkan secara langsung kepada trend terhadap kemungkinan overfitting dan tentu saja tergantung pada bagaimana dekatnya nilai ranking yang dihasilkan bagi masing masing atribut yang dipilih dalam subset.

Berdasarkan pada pengukuran yang dilakukan dalam algorithma seleksi fitur, kami usulkan suatu formulasi evaluasi perankingan yang *robust* membanding dua kelompok ranking r dan r'. Dalam pendekatan ini, koefisien transformasi ranking normal dimodifikasi menggunakan jarak Euclidean untuk menghasilkan ukuran S_R yaitu *Stability ranking measure*:

$$S_R = \sqrt{\sum_{p=1}^{\Pi}\left(\frac{r_p}{(\Pi-1)} - \frac{r_p'}{(\Pi-1)}\right)^2} \qquad (12)$$

Dimana r_p dan r'_p merupakan ranking fitur (pada masing masing atribut) dalam dataset masing-masing untuk data training dan data testing, *II* adalah jumlah atribut dalam dataset. Apabila didapat output S_R bernilai null berarti kedua algoritma seleksi fitur identik. Semakin besar nilai S_R dihasilkan semakin rendah derajat stabilitas algoritma yang bersangkutan.

V. Percobaan dan Diskusi

Tujuan percobaan ini adalah untuk memahami algoritma seleksi fitur yang mana yang menunjukkan kinerja yang baik dengan evaluasi terhadap kerobustan stabilitasnya dalam menangani semua data set. Asumsi dasar evaluasi ini adalah bahwa kapasitas dan kemampuan olah data mencerminkan kerobustan stabilitas algoritma. Melalui penentuan stabilitas algoritma seleksi fitur ini juga diharapkan didapat cara praktis memilih algoritma yang tepat bagi pengambilan keputusan.

5.1. Setup dataset

Permasalahan penentuan fitur yang tepat berkaitan erat dengan akses berfrekuensi tinggi dalam pemantauan basisdata seperti dalam kasus warehouse CRM atau yang sejenis. Di samping masalah akses berfrekuensi tinggi, dalam memilih dataset juga mempertimbangkan ada kondisi acak dalam basis data yang menjamin keragaman stabilitas data antara fase training dan testing. Berikut adalah pertimbangan lainnya:

1. Dataset mestinya berisikan dua-kelas untuk memudahkan proses evaluasi dan memungkin berbagai klasifier seleksi fitur terlibat. Salah satu yang memenuhi adalah sumber data dari UCI repository dan PAKDD 2007
2. Dataset semestinya tidak berisi terlalu banyak nilai yang tidak diketahui atau hilang agar bias menjamin perbandingan akurasi yang lebih objektif di antara algoritma yang dibandingkan
3. Pemeriksaan lebih jauh bagi atribut individu dalam kaitannya dengan kemungkinan korelasi langsung yang bias merusak hasil pembandingan antar algoritma

Dari pertimbangan di atas CoIL2000 dataset yang berisi 5900 record data training serta 4000 record data testing dipilih untu melengkapi data contest PAKDD 2007 yang berisikan 40K record data training dan 8K record data testing. Implementasi algoritma seleksi fitur ini dijalankan menggunakan data mining library pada Java dalam lingkungan Weka [6] dan library Orange [9]. Mesin CPU menggunakan Pentium-4 3.06 GHz berjalan dalam lingkungan Win32 dan RAM 1.5 GB serta Java SE 1.6x.

5.2. Hasil dan pembahasan

Kami menggunakan CoIL2000 dan PAKDD data contest untuk menguji kinerja semua algoritma seleksi fitur. Dalam kaitannya menentukan rangkaian urutan ranking yang paling relevan, kami menggunakan fitur yang paling penting dalam subset hasil seleksi. Pada Tabel 1. dijelaskan hasil dari empat algoritma yang dipakai (IG, SU, GR dan CHI) yang memberikan rentang S_R antara 2.0 – 3.0 yang berarti terdapat beda signifikan antara hasil perankingan pada data training dan data testing. Terdapat kesamaan pada nilai akurasi, ukuran classification tree dan durasi proses klasifikasi yang berarti juga terdapat kesamaan jumlah fitur yang dipilih sebanyak 38. Dari hasil ini bisa kita lihat bahwa algoritma CFS merupapak algoritma paling stabil dengan perolehan S_R paling rendah, durasi klasifikasi tercepat dan tingkat akurasi tertinggi.

Tabel 1. Ranking stabilitas CoIL2000 dan akurasi menggunakan C4.5

No	Algoritma	S_R	#Atrib	Akuras	Ukuran Tree
1	IG	2.097	38	93.89	Leaf=6; root=11
2	GR	2.800	38	93.89	Leaf=6; root=11
3	SU	2.620	38	93.89	Leaf=6; root=11
4	CHI	2.210	38	93.89	Leaf=6; root=11
5	REL	1.021	61	93.90	Leaf=6; root=11
6	CSF	0.430	10	94.02	Leaf=6; root=11

Hasil pada Tabel 1., juga menunjukkan bahwa algoritma ReliefF cukup stabil dalam perankingan pada S_R sebesar 1.021, namun tidak cukup handal dalam seleksi subset terkecil dengan 61 atribut dibandingkan algoritma lain. ReliefF juga memerlukan waktu paling lama sehingga menjadikan algoritma yang paling mahal secara komputasi. Hasil detil penghitungan durasi hitung dapat dilihat pada Tabel 2. Untuk menguji lebih jauh pendekatan stabilitas ini, kami terapkan pada dataset yang lebih besar dari kontes PAKDD 2007. Dataset ini terdiri atas 40K record training data dan 8K record testing data. Kami menghitung ranking stabilitas berdasarkan porsi training dan testing ini.

Sebagaimana telah didapatkan gambaran awal kinerja stabilitas tiap algoritma menggunakan dataset CoIL2000, dari Tabel 2 kita bisa tambahkan pengertian perilaku stabilitas masing-masing algoritma tersebut. Dari sisi internal pembangun, IG,GR, dan SU terdiri atas unit entropy menghasilkan jumlah fitur terpilih yang sama sehingga juga memakai durasi klasifikasi yang sama menggunakan algoritma C4.5.

Dari sisi durasi seleksi dan perankingan atribut, algoritma ReliefF menjadi yang terlama sedangkan CFS tercepat menghasilkan pilihan fitur. Hasil ini menunjukkan beban pembandingan antar fitur yang dilakukan dalam ReliefF jauh lebih berat dengan proses rekursif *missed* dan *hit*. Sementara pada CFS dengan pembandingan antar fitur yang berbasis pada korelasi Pearson terlihat lebih ringan dan sangat cepat.

Tabel 2. Akurasi klasifikasi hasil seleksi fitur menggunakan C4.5 pada dataset PAKDD 2007

No	Algoritma	Durasi olah(det)	#Atribut	Ukuran Tree
1	IG	81.69	37	Leaf=78;size=113
2	GR	81.69	37	Leaf=78;size=113
3	SU	81.69	37	Leaf=78;size=113
4	REL	>9000	37	Leaf=78;size=113
5	CFS	7.09	11	Leaf=6;size=11
6	CHI	81.69	37	Leaf=78;size=113

Pada Tabel 3. kami membandingkan durasi pengklasifikasian C4.5 dengan algoritma penklasifikasian populer lain. Tujuannya adalah untuk mengetahui rentang akurasi yang mungkin didapat dari semua fitur terpilih. Dari hasil pada Tabel 3, kita lihat algoritma 1-R (one R) merupakan algoritma tercepat dengan 2.53 detik masa olah, namun menghasilkan tingkat akurasi lebih rendah daripada pengklasifikasi lain. Penghitungan akurasi klasifikasi dataset PAKDD tertinggi didapat pada penggunaan algoritma Adaboost dengan inti decision stump sebesar 99.86%, disusul oleh algoritma Bayes naïve sebesar 99.79%. Trade off durasi pengolahan yang diberikan oleh Bayes naïve cukup signifikan, yaitu sekitar 1/5 waktu yang dibutuhkan oleh Adaboost.

Tabel 3. Perbandingan akurasi hasil seleksi fitur terbaik dataset PAKDD 2007 menggunakan beberapa klasifier

No	Klasifier	Durasi (det)	Akurasi (%)
1	Bayes Net	6.41	99.79
2	Ensemble Naïve Bayes	548.32	98.61
3	AdaBoost—Decision Stump	31.58	99.86
4	1-R	2.53	98.70

Selanjutnya untuk menjamin konsistensi stabilitas algoritma, homogenitas ranking yang dihasilkan setiap algoritma seleksi fitur dievaluasi berdasarkan densitas dataset secara acak. Pendekatan ini merupakan ekstensi dari validasi silang. Asumsi dari evaluasi ini adalah bahwa rasio fitur terpilih pada masing-masing tingkat kerapatan data yang berbeda dapat memberikan gambaran sebaran kestabilan yang dihasilkan oleh algoritma pilihan fitur. Algoritma yang paling stabil akan menunjukkan rasio yang relatif seragam pada densitas dataset yang berbeda.

Kami mengacak record dataset menjadi tiga tingkat densitas pada CoIL2000 (1K, 2K dan 4K records) dan empat tingkat pada PAKDD 2007 (1K, 2K, 4K dam 8K records). Selanjutnya proses seleksi fitur dilakukan pada setiap kelompok densitas dan membandingkan besarnya fitur terpilih pada tahapan training dengan hasil yang diperoleh pada data testing. Hasil dari proses ini dapat dilihat pada Tabel 4.

Tabel 4. Rasio jumlah fitur terpilih dalam densitas berbeda pada data training terhadap data testing

	Algoritma					
Dataset-densitas	IG	GR	REL	SU	CFS	CHI
CoIL20001K	15/25	15/25	84/84	15/25	9/6	15/25
CoIL2000-2K	21/25	21/25	84/84	21/25	5/12	21/25
CoIL2000-4K	39/38	39/38	84/84	39/38	14/10	39/38
PAKDD1K	14/21	14/21	40/40	14/21	2/7	14/21
PAKDD2K	19/24	19/24	40/40	19/21	7/9	19/24
PAKDD4K	25/23	25/23	40/40	25/23	7/9	25/23
PAKDD8K	34/33	34/33	40/40	34/33	9/12	34/33

Hasil yang didapat pada Tabel 4 menunjukkan CFS rasio terkecil dari rata rata semua porsi densitas yang dipakai.

Algoritma ReliefF secara konsisten memberikan nilai 1 untuk semua tipe densitas pada PAKDD 2007 maupun CoIL2000. Temuan pola yang sama pada empat algoritma (IG, GR, SU dan CHI) pilihan fitur, kembali bisa diperkuat dengan sebaran fitur terpilih pada training terhadap testing. Pada masing masing densitas keempat algoritma sepakat memilih sejumlah hasil di training dan pada dataset testing.
Pada Tabel 5., kami tampilkan detil hasil ranking stabilitas pada skala densitas sesuai definisi 2.

Tabel 5. Ranking stabilitas pada beberapa skala densitas

	Algoritma					
Dataset-densitas	IG	GR	REL	SU	CFS	CHI
CoIL20001K	3.40	2.32	3.48	3.03	1.65	2.01
CoIL2000-2K	3.71	4.41	2.78	3.43	0.92	4.00
CoIL2000-4K	2.78	3.18	2.78	4.44.	1.74	3.55
Rata-rata	3.30	3.30	3.01	3.63	1.43	3.18
PAKDD1K	1.77	1.58	2.34	1.64	0.57	1.64
PAKDD2K	1.64	1.89	1.73	1.44	0.52	1.04
PAKDD4K	1.72	1.31	2.22	1.86	0.38	2.08
PAKDD8K	2.55	2.23	2.22	2.78	0.78	2.62
Rata-rata	1.92	1.76	2.13	1.93	0.56	1.85

Hasil pada Tabel 5 di atas memperkuat hasil awal yang sebelumnya telah didapat. Dapat dilihat bahwa algoritma CFS adalah algoritma yang paling stabil dalam menyelesaikan proses seleksi fitur pada dua dataset dengan nilai rata-rata S_R terkecil pada semua skala densitas. Dengan demikian bisa dinyatakan bahwa definisi 2 tentang evaluasi homogenitas bisa diterapkan untuk evaluasi detil stabilitas algoritma seleksi fitur. Melalui pengukuran homogenitas kestabilitas algoritma juga menjamin kompleksitas acak isi data dalam dataset pada permasalahan berfitur banyak.

VI. Kesimpulan dan Peluang Pengembangan

Pada paper ini kami mengenalkan transformasi ranking normal bagi evaluasi kestabilan algoritma selesi fitur. Pengukuran yang dibuat untuk kepentingan ini menunjukkan bagaimana evaluasi ketangguhan (robustness) masing-masing algoritma bisa mendukung proses pemilihan dan penggunaan algoritma yang paling tepat pada kasus-kasus spesifik.
Hasil evaluasi ini juga menunjukkan bahwa semua algoritma berbasis entropy mempunyai kecenderungan memilih atribut yang sama dengan jumlah yang sama juga. CFS merupakan algoritma paling stabil pada semua tingkat densitas yang diuji.
Pada riset mendatang kami akan menguji saling ketergantungan antar fitur yang mempengaruhi kestabilan algoritma seleksi fitur pada database berdimensi sangat besar.

Daftar Pustaka

1. Kira, K., L.A.Rendell: The Feature Selection Problem: Traditional methods and a New Algorithm. AAAI Press/The MIT Press (1992) 129-134
2. Hall, M.A.: Correlation-based feature selection for discrete and numeric class machine learning. In Proceedings of the 17th Intl. Conf. Machine Learning (2000) 359-366
3. Kononenko, I., Robnik-Sikonja, M., Pompe, U.: ReliefF for estimation and Discretization of Attributes in Clasification, Regression and ILP. In Proceedings of the the European Conf. on Machine Learning (1994) 171-182
4. Han, J., Kamber, M.: Data Mining: Concept and Techniques. Morgan Kaufmann San Fransisco (2005)
5. Maimon, O., Rokach, L.: The Data Mining and Knowledge Discovery Handbook. Morgan Kaufmann San Fransisco (2005)
6. Witten, I.H., Frank, E.: Data Mining-Practical Machine Learning Tools and Techniques in Java Implementation. Morgan Kaufmann San Fransisco (2000)
7. Quinlan, J.R.: C4.5 The Program for Machine Learning. (1993)
8. Breiman, L., Friedman, M., Stoner, M., Olshen, M.: Classification and Regression Tree. John Wiley and Co (1984)
9. Demsar, J., Zupan, B.: Orange: From Experimental Machine Learning to Interactive Data Mining. White paper (ww.ailab.si/orange) (2004)