KLASIFIKASI VOTING ANN PSO BICLASS DENGAN SELEKSI FITUR GAIN RATIO

Fetty Tri Anggraeny [1)], Monica Widyasri [2)]
[1)] Jurusan Teknik Informatika, Fakultas Teknologi Industri, UPN Veteran Jawa Timur
[2)] Jurusan Teknik Informatika, Fakultas Teknik, Universitas Surabaya
email : fetty.ta@gmail.com [1)], monica@ubaya.ac.id [2)]

Abstrak: Seleksi fitur merupakan tahapan penting dalam proses klasifikasi. Proses ini menganalisa data (fitur) sehingga menghasilkan fitur yang berperan atau kurang berperan dalam proses klasifikasi. Peranan sebuah fitur dalam klasifikasi dapat dikalkulasi dengan suatu rumusan, dalam penelitian ini digunakan metode gain ratio untuk mendapatkan bobot atribut dalam proses klasifikasi. Metode seleksi fitur gain ratio menggunakan pendekatan seleksi fitur filter, karena dilakukan terlepas dari mesin klasifikasi. Mesin klasifikasi yang digunakan adalah ANNPSO, dimana mesin ini menggabungkan konsep kecerdasan buatan saraf manusia (neural network) dengan kecerdasan hewan (particle swarm intelligence). Metode yang diusulkan akan di uji coba terhadap 3 dataset UCI, antara lain iris, breast Wisconsin dan dermatology. Uji coba dengan variasi nilai batas gain ratio fitur menunjukkan nilai akurasi yang cukup tinggi terhadap 3 dataset yaitu 97,6%, 96,41%, dan 99,29%.

Keywords: gain ratio, voting klasifikasi, ANNPSO Biclass.

1. PENDAHULUAN

Dalam proses klasifikasi diperlukan fitur-fitur dari objek yang akan dianalisa. Jika suatu objek data memiliki banyak fitur, maka mesin klasifikasi akan membutuhkankan lebih banyak waktu untuk menciptakan garis pembatas data antar kelas. Seleksi fitur digunakan sebagai tahapan praproses klasifikasi yang bertujuan mengurangi dimensi dari fitur. Sehingga dari serangkaian fitur dapat diketahui fitur kuat dan fitur lemah peranannya dalam klasifikasi

Seleksi fitur yang digunakan pada penelitian sebelumnya [4] adalah adaptive feature selection discriminant ratio yang dilakukan secara Biclass (2 class berpasangan), sehingga jika ada n class, maka dilakukan seleksi fitur sebanyak $n*(n-1)/2$. Sehinggan setiap pasang class memiliki kumpulan fitur yang berbeda bobotnya dalam mengidentifikasi suatu data tergolong dalam class tertentu.

Jaringan Saraf Tiruan (*Artificial Neural Network*) merupakan salah satu mesin klasifikasi yang mengadopsi cara kerja saraf otak manusia. JST memerlukan pembelajaran agar neuron-neuron bisa menghasilkan keluaran sesuai dengan yang diinginkan.

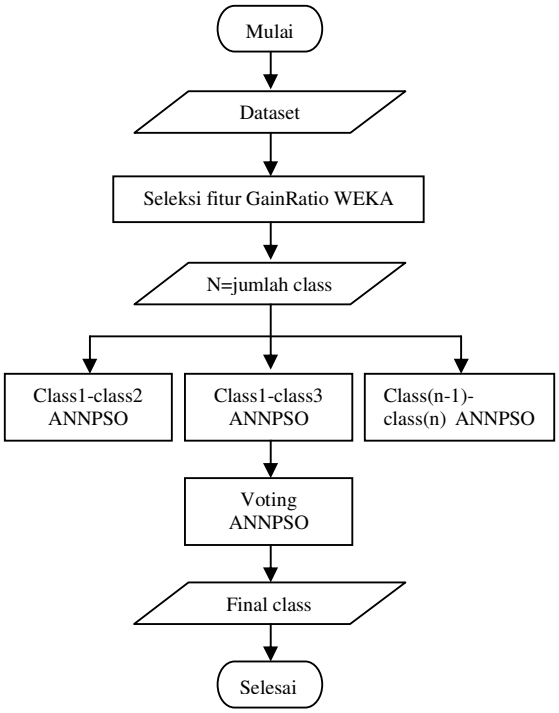
Ada 2 tujuan dalam pembelajaran JST, yaitu menghasilkan arsitektur JST dan menghasilkan nilai bobot neuron dalam arsitektur yang sudah ditentukan [2, 3]. Salah satu metode pembelajaran untuk memperoleh nilai bobot neuron yang umum digunakan adalah propagasi balik. Pembelajaran dengan metode kecerdasan buatan hewan saat ini semakin berkembang, salah satunya adalah metode optimasi *particle swarm* (PSO). PSO digunakan dalam teknik pencarian global untuk menghindari terjadinya local minima sehingga dapat mengoptimasi bobot neuron pada JST [3].

Dalam penelitian ini akan diterapkan 1 (satu) kali proses seleksi fitur untuk semua class, agar proses klasifikasi keseluruhan dapat meningkatkan akurasi mesin klasifikasi Voting ANNPSO Biclass.

2. MODEL, ANALISA, DESAIN, DAN IMPLEMENTASI

Gambar 1 menampilkan metodologi penelitian keseluruhan. Dimulai dari memasukkan dataset, kemudian seleksi fitur menggunakan software WEKA 3-6-9 dengan menggunakan seluruh data dataset. Keluaran dari WEKA dijadikan dasar perankingan fitur dan nantinya digunakan dalam mesin klasifikasi ANNPSO Biclass. Mesin klasifikasi yang terbentuk dalam 1 kali running adalah $n*(n-1)/2$, masing-masing bertugas untuk sepasang class. Karena terdapat beberapa mesin klasifikasi, maka diperlukan voting untuk mengetahui class final dari masing-masing data berdasarkan suara terbanyak. Suara terbanyak dihitung secara akumulasi, jumlah suatu dataset teridentifikasi dalam class tertentu oleh mesin klasifikasi.

Hasil seleksi fitur menggunakan WEKA 3.6.9 untuk masing-masing dataset dapat dilihat pada Gambar 2, Gambar 3, dan Gambar 4.



Gambar 1. Metode yang diusulkan.

2.1. Dataset

Dataset yang digunakan dalam penelitian ini antara lain iris, breast Wisconsin, dan dermatology. Adapun karakteristik masing-masing dataset dapat dilihat pada Tabel 1.

Tabel 1. Karakteristik dataset.

Dataset	Jumlah fitur	Jumlah class	Jumlah *instance*
Iris	4	3	150
Breast Wisconsin	9	2	699
Dermatology	34	15	366

Sumber: UCI Database

2.2. Seleksi Fitur

Seleksi fitur sebagai tahapan dalam klasifikasi terbagi menjadi 2 pendekatan, yaitu pendekatan *wrapper* dan pendekatan *filter* [1]. Seleksi fitur dengan pendekatan filter dilakukan secara terpisah dengan mesin klasifikasi, atau dengan kata lain seleksi fitur dijadikan tahapan pra-proses sebelum data dimasukkan ke dalam mesin klasifikasi. Sebuah data yang telah dilakukan seleksi fitur dengan pendekatan ini, dapat digunakan atau dikombinasikan dengan beberapa mesin klasifikasi. Berbeda dengan pendekatan filter, pendekatan wrapper menjadikan mesin klasifikasi sebagai bagian proses dalam menentukan penting atau tidaknya suatu Fitur. Mesin klasifikasi akan dilatih dan diuji dengan kombinasi fitur yang berubah, jika penambahan suatu fitur menurunkan akurasi mesin klasifikasi maka fitur tersebut tidak digunakan. Mesin akan memproses kombinasi fitur yang lain dan tidak menggunakan fitur yang akan memperlemah akurasi system.

Berdasarkan gambaran tersebut, tampak nyata bahwa pendekatan filter lebih sederhana dibandingkan pendekatan wrapper. Selain itu, waktu yang dibutuhkan untuk pendekatan filter lebih cepat, dibandingkan dengan pendekatan wrapper. Tetapi dengan menggunakan pendekatan wrapper, mesin klasifikasi akan dioptimasi berdasarkan fitur yang digunakan.

Dalam penelitian ini, proses seleksi fitur menggunakan software WEKA 3.6.9 dengan menggunakan metode GainRatioEvaluator (WEKA).

2.3. *GainRatio* Seleksi Fitur

Pohon keputusan adalah struktur tree dimana setiap internal node merepresentasikan fitur dan eksternal node (*leaf*) merepresentasikan konklusi class dari suatu data. Informai gain digunakan untuk menentukan fitur yang harus diletakkan di setiap internal node. Semakin tinggi posisi node fitur, maka fitur tersebut memiliki nilai informasi gain lebih tinggi. Semakin besar informasi gain, menunjukkan semakin besar suatu fitur memiliki peranan dalam menentukan keluaran, dalam hal ini konklusi class data.

Gain ratio merupakan pengembangan dari informasi gain. Informasi gain digunakan untuk membentuk induksi pohon keputusan (ID3), sedangkan gain ratio digunakan pada C4.5, yang merupakan pengembangan dari ID3 [1}. Informasi gain menghasilkan bias, informasi gain lebih memilih fitur dengan banyak variasi nilai daripada fitur yang memilik sedikit variasi nilai meskipun lebih informatif [1]. Contoh, fitur unik pada suatu data seperti id siswa dalam tabel siswa di database. Pemisahan menggunakan id siswa menghasilkan sangat banyak partisi, karena setiap record data memiliki nilai unik yaitu id siswa [5].

Misal *S* adalah himpunan data sampel dan *m* adalah class. Maka entropi atau perkiraan informasi untuk mengklasifikasi sample:

$$I(S) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (1)$$

dimana p_i adalah probabilitas sample dengan konklusi $class_i$. Misal fitur/atribut A memiliki variasi nilai sebanyak v. Misal s_{ij} adalah jumlah

sampel class C_i dalam subset S_j. S_j teridiri dari sampel dalam S yang memiliki nilai a_j dari A. Maka entropi berdasarkan pembagian menjadi subset atribut A:

$$E(A) = -\sum_{i=1}^{m} I(S) \frac{s_{1i}+s_{2i}+\cdots s_{mi}}{s}$$

(2)

Informasi gain untuk mencabangkan atribut A adalah:

Gain(A)=I(S)-E(A)

(3)

C4.5 menggunakan gain ratio dengan mengaplikasikan normalisasi terhadap informasi gain dengan nilai yang diperoleh dari:

$$SplitInfo(S) = -\sum_{i=1}^{v}(|S_i|/|S|)log_2(|S_i|/|S|)$$

(4)

Gain ratio dihitung menggunakan rumusan berikut:

GainRatio(A)=Gain(A)/SplitInfo(S)

(5)

Atribut dengan nilai gain ratio tertinggi terpilih sebagai atribut pemisah (*splitting attribute*).

```
=== Attribute Selection on all input data ===
Search Method: Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 5
class): Gain Ratio feature evaluator

Ranked attributes:
 0.871  4 petalwidth
 0.734  3 petallength
 0.381  1 sepallength
 0.242  2 sepalwidth
Selected attributes: 4,3,1,2 : 4
```

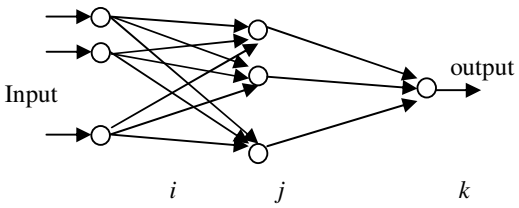
Gambar 2. Seleksi fitur dataset iris

```
=== Attribute Selection on all input data ===
Search Method: Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 10
Class): Information Gain Ranking Filter

Ranked attributes:
 0.675   2 Cell_Size_Uniformity
 0.66    3 Cell_Shape_Uniformity
 0.564   6 Bare_Nuclei
 0.543   7 Bland_Chromatin
 0.505   5 Single_Epi_Cell_Size
 0.466   8 Normal_Nucleoli
 0.459   1 Clump_Thickness
 0.443   4 Marginal_Adhesion
 0.198   9 Mitoses
Selected attributes: 2,3,6,7,5,8,1,4,9 : 9
```

Gambar 3. Seleksi fitur dataset breast wisconsin.

```
=== Attribute Selection on all input data ===

Search Method: Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 35 class): Gain Ratio feature evaluator

Ranked attributes:
 0.7715   31 perifollicular_parakeratosis
 0.7254   27 vacuolisation_and_damage_of_basal_layer
 0.7237   33 band-like_infiltrate
 0.7221    6 polygonal_papules
 0.7111   29 saw-tooth_appearance_of_retes
 0.7094   12 melanin_incontinence
 0.7019   15 fibrosis_of_the_papillary_dermis
 0.6829   25 focal_hypergranulosis
 0.6741    8 oral_mucosal_involvement
 0.628    30 follicular_horn_plug
 0.6012   22 thinning_of_the_suprapapillary_epidermis
 0.5919   20 clubbing_of_the_rete_ridges
 0.5303   21 elongation_of_the_rete_ridges
 0.5297   34 Age
 0.527     7 follicular_papules
 0.438     9 knee_and_elbow_involvement
 0.4291   24 munro_microabcess
 0.3993   10 scalp_involvement
 0.3707   28 spongiosis
 0.3251   14 PNL_infiltrate
 0.325    16 exocytosis
 0.3171   26 disappearance_of_the_granular_layer
 0.2941   23 spongiform_pustule
 0.2911   11 family_history
 0.2674    5 koebner_phenomenon
 0.1978    3 definite_borders
 0.1769    2 scaling
 0.1687   19 parakeratosis
 0.1599   13 eosinophils_in_the_infiltrate
 0.1491    4 itching
 0.098     1 erythema
 0.0959   18 hyperkeratosis
 0.0833   17 acanthosis
 0.0598   32 inflammatory_monoluclear_inflitrate

Selected attributes: 31,27,33,6,29,12,15,25,8,30,22,20,21,34,7,9,24,10,28,14,16,26,23,11,5,3,2,19,13,4,1,18,17,32 : 34
```

Gambar 4. Seleksi fitur dataset dermatology.



Gambar 5. ANNPSO Biclass topology

2.4. ANNPSO Biclass

Dalam penelitian ini digunakan mesin klasifikasi Artificial Neural Network yang dioptimasi dengan Particle Swarm. ANNPSO Biclass adalah ANNPSO yang mengklasifikasi data ke dalam 2 class. Jika dataset memiliki n class dan n>2, maka akan terbentuk n(n-1)/2 ANNPSO Biclass (Fetty, 2011).

Topology ANNPSO Biclass dapat dilihat pada Gambar 5. Node input disesuaikan jumlah fitur yang digunakan pada setiap klasifikasi, node keluaran hanya 1 karena digunakan untuk membedakan antara 2 class (0 dan 1). PSO digunakan untuk memperbaiki bobot jaringan sampai error klasifikasi minimum.

3. HASIL

Mula-mula dilakukan seleksi fitur menggunakan WEKA terhadap 3 dataset dengan fungsi GainRatio. Hasil seleksi fitur dapat dilihat pada Gambar 1 sampai Gambar 3. Table 2 menampilkan nilai rata, minimum, dan maksimum gain ratio masing-masing dataset. Kami melakukan 3 skenario uji coba dalam penelitian ini, antara lain:

a. Skenario 1, dilakukan ujicoba mesin klasifikasi dengan variasi fitur didasarkan pada nilai minimum gain ratio yang digunakan (limitRank). Uji coba ini akan menghasilkan perbandingan akurasi dan waktu komputasi dengan variasi limitRank.
b. Skenario 2, membandingkan metode yang diusulkan dengan penelitian sebelumnya (Fetty, 2011).

Tabel 2. Karakteristik nilai GainRatio dataset.

Dataset	Rata-rata	Minimum	Maksimum
Iris	0.5570	0.242	0.871
Breast Wisconsin	0.4450	0.198	0.675
Dermatology	0.4255	0.0598	0.7715

Skenario ujicoba yang pertama adalah menerapkan sebuah nilai limitRank yang digunakan untuk memfilter atribut berdasarkan gain ratio, yang kemudian diproses oleh mesin klasifikasi voting ANNPSO. Range limitRank mulai 0.05 sampai nilai 0.85 atau sampai nilai maksimum gain ratio. Masing-masing dataset dilakukan 3 kali uji coba. Nilai akurasi dan waktu komputasi yang ditampilkan pada Tabel 3 adalah rata-rata dari 3 uji coba yang telah dilakukan.

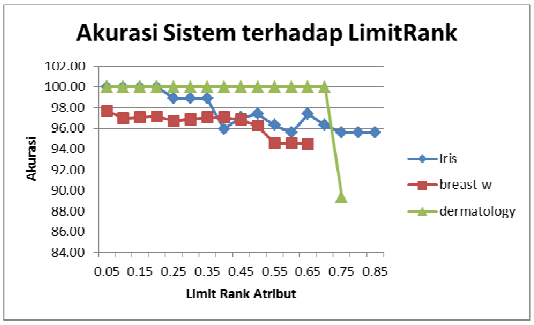
Berdasarkan Tabel 3, fitur petal width pada dataset iris sangat dominan. Hal ini ditampilkan dengan nilai akurasi mesin klasifikasi yang menggunakan fitur petal width saja (lihat baris 15-17 Tabel 3 kolom akurasi dataset iris) cukup tinggi, yaitu 95,56%. Sedangkan pada dataset dermatology, penggunaan 1 fitur dengan nilai gain ratio tertinggi tidak cukup memuaskan, karena menghasilkan akurasi sebesar 89,38%. Hal ini dikarenakan nilai gain ratio antara fitur ke-4 dan ke-3 pada dataset iris memiliki selisih yang cukup besar, yaitu 0,137% (diperoleh dari 0.871-0.734). Sedangkan pada dataset dermatology, 7 fitur dengan nilai gain ratio tertinggi memiliki silisih kecil, rata-rata sekitar 0,01%. Sedangkan pada dataset breast-w secara umum mengalami penurunan akurasi dengan semakin besar nilai limitRank, tetapi selisih akurasi antara penggunaan seluruh fitur dan paling sedikit fitur hanya berselisih sedikit, yaitu 3,19%. Sedangkan pada dataset iris dan dermatology adalah 4,44% dan 10,62%. Gambar 5 menunjukkan bahwa akurasi dataset dermatology merosot tajam, sedangkan dataset iris dan breast-w mengalami penurunan akurasi bertahap.

Waktu komputasi yang dibutuhkan relative sebanding dengan akurasi yang dihasilkan (lihat Gambar 6). Pada dataset dermatology dengan 1 fitur membutuhkan waktu yang jauh lebih besar dari yang lain, hal ini dikarenakan mesin klasifikasi ANNPSO lebih sulit untuk mencapai konvergen atau memenuhi limit error hanya dengan 1 fitur saja.

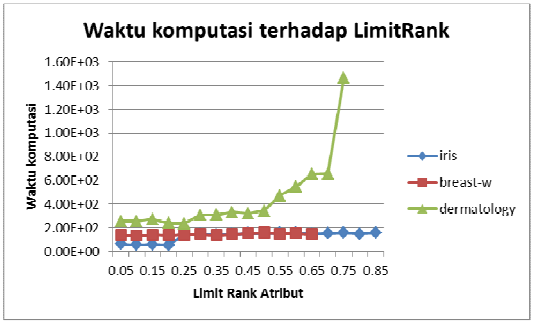
Skenario kedua, dilakukan dengan tujuan mengetahui perbandingan akurasi antara metode yang diusulkan dengan penelitian sebelumnya (Fetty, 2011).

Tabel 3. Rata-rata akurasi dan waktu komputasi

Nomor	limitRank	IRIS			BREAST WISCONSIN			DERMATOLOGY		
		Jumlah fitur	Akurasi	Waktu Komputasi	Jumlah fitur	Akurasi	Waktu Komputasi	Jumlah fitur	Akurasi	Waktu Komputasi
1.	0.05	4	100.00	63.03	9	97.71	141.61	34	100.00	255.75
2.	0.10	4	100.00	57.08	8	97.00	136.07	30	100.00	255.14
3.	0.15	4	100.00	60.87	8	97.04	143.06	29	100.00	273.02
4.	0.20	4	100.00	54.25	7	97.14	140.22	25	100.00	240.71
5.	0.25	3	98.89	146.81	7	96.71	143.70	25	100.00	237.56
6.	0.30	3	98.89	146.59	7	96.90	146.14	22	100.00	309.52
7.	0.35	3	98.89	145.23	7	97.04	139.51	19	100.00	311.11
8.	0.40	2	95.93	147.47	7	97.04	145.99	17	100.00	331.66
9.	0.45	2	97.04	151.95	6	96.85	151.60	15	100.00	321.97
10.	0.50	2	97.41	152.96	4	96.23	158.19	15	100.00	344.37
11.	0.55	2	96.30	159.87	2	94.56	148.62	12	100.00	474.19
12.	0.60	2	95.56	159.45	2	94.56	152.91	11	100.00	548.91
13.	0.65	2	97.41	150.98	2	94.52	146.83	9	100.00	655.48
14.	0.70	2	96.30	151.59	-	-	-	7	100.00	658.15
15.	0.75	1	95.56	157.59	-	-	-	1	89.38	1464.73
16.	0.80	1	95.56	149.94	-	-	-	-	-	-
17.	0.85	1	95.56	158.64	-	-	-	-	-	-
	RATA-RATA		97.60	130.25		96.41	145.73		99.29	445.49



Gambar 6. Grafik pengaruh limit gain ratio terhadap akurasi.



Gambar 7. Grafik pengaruh limit gain ratio terhadap waktu komputasi

Tabel 4. Perbandingan rata-rata akurasi.

Dataset	(Fetty, 2011)	Metode yang diusulkan
Iris	98,05%	97,6%
Breast-w	99,9%	96,41%
Dermatology	86,01%	99,29%

Tabel 4, menunjukkan metode yang diusulkan memberikan akurasi jauh lebih baik daripada metode sebelumnya pada dataset dermatology dengan nilai kenaikan sebesar 13,28%, sedangkan pada dataset iris dan breast-w mengalami penurunan akurasi sebesar 0,45% dan 3,49%.

Berdasarkan uji coba yang telah dilakukan, metode yang diusulkan secara umum dapat meningkatkan akurasi.

4. SIMPULAN

Berdasarkan uji coba yang telah dilakukan dapat disimpulkan bahwa metode yang diusulkan memiliki tingkat akurasi yang cukup tinggi dengan menggunakan beberapa fitur dengan nilai gain ratio tertinggi. Waktu komputasi yang dibutuhkan dengan sedikit fitur lebih lamadaripada dengan

banyak fitur, hal ini dikarenakan proses mendapatkan konfigurasi bobot dengan minimum error membutuhkan lebih banyak iterasi.

Nilai akurasi klasifikasi dengan menggunakan fitur yang memiliki gain ratio lebih besar 0,5 bernilai diatas 95% untuk 3 dataset uji. Hal ini sudah menunjukkan bahwa 0,5 dapat dijadikan batas minimum gain ratio fitur yang digunakan dalam proses klasifikasi. Selain itu, waktu komputasi yang dibutuhkan tidak terlalu lama untuk 3 dataset sekitar 3,64 menit.

5. DAFTAR PUSTAKA

[1] Asha Gowda Karegowda, A. S. Manjunath, M.A.Jayaram. Comparative Study of Attribute Selection Using Gain Ratio Correlation based feature selection. International Journal of Information Technology and Knowledge Management, July-December 2010, Volume 2, No. 2, pp. 271-277.

[2] Carvalho Marcio, Ludermir Teresa B., "Particle Swarm Optimization of Neural Network Architectures and Weights", Seventh International Conference on Hybrid Intelligent Systems, 2007.

[3] Carvalho Marcio, Ludermir Teresa B., "Particle Swarm Optimization of Feed-Forward Neural Networks with Weight Decay", Proceedings of the Sixth International Conference on Hybrid Intelligent Systems (HIS'06), 2006.

[4] Fetty Tri Anggraeny, Indriati, Heliza Rahmania Hatta. Voting of Biclass and Multiclass Artificial Neural Network. BISSTECH 2012 Proceesdings.

[5] Sri Harsha Vege. Ensemble of Feature Selection T echniques for High Dimensional Data. Masters Theses & Specialist Projects Western Kentucky University, 2012.