The 13th Industrial Electronics Seminar 2011 (IES 2011)
Electronic Engineering Polytechnic Institute of Surabaya (EEPIS), Indonesia, October 26, 2011

Seleksi Fitur Menggunakan Random Forest Dan Neural Network

Wahyu S. J. Saputra[1, 2], Arif Rahman Sujatmika[1], Agus Zainal Arifin [3]
[1] Program Magister Jurusan Teknik Informatika ITS Surabaya
[2] Jurusan Informatika, Fakultas Teknologi Industri, UPN "Veteran" Jawa Timur
[3] Laboratorium Vision and Image Processing, Jurusan Teknik Informatika, ITS Surabaya
wahyu.s.10@mhs.if.its.ac.id

Abstrak

Seleksi fitur merupakan sebuah tahapan penting dalam proses klasifikasi, karena fitur yang terseleksi sangat mempengaruhi tingkat akurasi dari klasifikasi. Pada dataset yang memiliki banyak fitur membutuhkan proses untuk mereduksi fitur sebanyak mungkin. Pada makalah ini diusulkan sebuah metode seleksi fitur menggunakan Ensemble Random Forest dan Neural Network. Dataset yang telah terbagi menjadi dua kelompok data, secara parallel akan dilakukan seleksi fitur menggunakan Neural Network, dan Ensemble Random Forest yang dilanjutkan Neural Network. Fitur hasil keluaran dari proses paralel tersebut, dilakukan pemilihan fitur menggunakan metode Voting-by-Majority. Percobaan seleksi fitur dengan menggunakan datataset iris, lung cancer, dan semeion handwritten digit. Dari hasil uji coba pada dataset iris dengan seleksi fitur, didapatkan akurasi 94,4%. Hasil uji coba menunjukkan metode seleksi fitur yang diusulkan dapat memperoleh hasil baik pada dataset yang memiliki tingkat variasi data yang tinggi, dan berlaku kebalikannya.

Keywords:Seleksi fitur,*neural network, ensemble random forest*

1. Pendahuluan

Fitur merupakan hal terpenting dalam proses klasifikasi, karena pemilihan fitur menentukan hasil dari proses klasifikasi. Berbagai metode dilakukan untuk melakukan seleksi fitur, karena tidak semua fitur mampu memberikan hasil klasifikasi baik. Jumlah fitur yang digunakan juga mempengaruhi waktu dalam proses klasifikasi. Proses seleksi fitur merupakan suatu usaha untuk mereduksi jumlah fitur pada sebuah dataset, dengan tetap mempertahankan nilai *error* seminimum mungkin. Dengan seleksi fitur mungkin dapat mengurangi biaya dalam proses koleksi data.

Dataset dengan jumlah fitur yang banyak sering dijumpai pada aplikasi pengenalan pola. Namun tidak semua fitur digunakan karena terdapat beberapa fitur yang redundan. Biasanya perfoma yang baik didapatkan dengan menghilangkan beberapa variable [1, 2, 3, 4]. Sebagaimana jumlah fitur yang bertambah tentu dibutuhkan jumlah data *training* yang lebih banyak lagi secara eksponensial[5]. Sehingga pada banyak penerapan dibutuhkan sebuah proses untuk mereduksi fitur.

Principal Component Analysis (PCA) [1, 6] serta *Linear Discriminant Analysis* (LDA) [1], dengan cara menciptakan fitur baru yang didapatkan dari kombinasi secara linear beberapa fitur, sehingga dapat digunakan untuk mereduksi dimensi dari fitur.

Esemble Random Forest to Trees (ERFTrees) merupakan sebuah metode untuk seleksi fitur pada dataset yang memiliki dimensi fitur yang besar namun memiliki jumlah data yang kecil [7], dengan menggunakan *Decision Tree* untuk melakukan proses seleksi. Dalam ERFTrees untuk mengatasi dataset yang kecil namun memiliki fitur yang banyak dilakukan proses *enlarge data* dengan menggunakan *Ensamble Random Forest* yang juga merupakan perkembangan dari *Decision Tree*. Dengan melakukan voting terhadap hasil dari keluaran *C.45 Decision Tree* dengan dataset yang belum dan yang sudah di *enlarge*, maka didapatkan nilai masing-masing fitur dari dataset yang di uji. Dari nilai itulah dilakukan pengurutan dari fitur yang memiliki nilai yang tinggi (memiliki nilai dua) sampai yang terendah (memiliki nilai nol). Bisa dikatakan bahwa fitur dengan nilai tertinggi adalah fitur yang paling direkomendasikan untuk digunakan dalam proses klasifikasi.

Sebenarnya selain menggunakan *C.45 Decision Tree* sebagai metode untuk seleksi fitur, dapat pula menggunakan mesin klasifikasi yang lain yaitu *Neural Network* [8].Telah didapatkan hasil pula bahwa pada kondisi dataset tertentu,*Ensamble Decision Tree* sebanding dengan *Single Neural Network* [9]. Pada proses klasifikasi sebuah *Feedforward Neural Network* memeriksa seluruh data *training* pada setiap perulangan untuk memperbaiki bobotnya [10]. Fokus dalam paper ini adalah peningkatan akurasi dalam melakukan seleksi fitur pada dataset yang mem[unyai banyak fitur.

Makapada makalah ini diusulkan sebuah metode untuk melakukan seleksi fitur dengan menggabungkan *Ensamble Random Forest* dan *Neural Network*. Penggunaan *Ensamble Random Forest* digunakan untuk proses enlarge dataset, sedangkan *Neural Network* digunakan untuk seleksi fitur.

2. Ensemble Random Forest

*Randomforest*merupakan pengembangan dari *Decision Tree* dengan menggunakan beberapa *Decision*

Tree, dimana setiap *DecisionTree* telah dilakukan *training* menggunakan sampel individu dan setiap atribut dipecah pada *tree* yang dipilih antara atribut subset yang bersifat acak. Dan pada proses klasifikasi, individunya didasarkan pada *vote* dari suara terbanyak pada kumpulan populasi *tree*.

Random Forest yang dihasilkan memiliki banyak *tree*, dan setiap *tree* ditanam dengan cara yang sama. *Tree* dengan variabel *x* akan ditanam sejauh mungkin dengan *tree* dengan variabel *y*. Dan dalam perkembangannya, sejalan dengan bertambahnya data set, maka *tree* pun ikut berkembang. Penempatan tree yang saling berjauhan membuat apabila terdapat *tree* disekitar *treex* berarti pohon tersebut merupakan perkembangan dari tree *x*[7]. Beberapa fungsi *learning* yang dihasilkan *random forest* digunakan strategi *ensemble* "*bagging*" untuk mengatasi masalah *overfitting* apabila dihadapkan data set yang kecil. Pada makalah ini *Ensemble* digunakan untuk melakukan *resample*data dengan mengklasifikasi ulang data *outlayer*sehingga merubah struktur data set yang asli.

3. Seleksi Fitur Dengan Neural Network

Neural Network merupakan sebuah mesin klasifikasi yang dimodelkan meniru dari struktur biologi pada saraf manusia. Pada makalah ini *neural network* yang digunakan termasuk dalam jenis MLP (*Multi Layer Perceptron*) dimana terdapat *hidden layer* sebelum *output* dari *Perceptron* diproses. MLP juga dikenal sebagai FFNN (*Feed-Forward Neural Network*).

Secara umum model ini bekerja dengan menerima suatu vektor input *I* dan kemudian menghitung suatu respon atau output *O* dengan memproses (*propagating)I* melalui beberapa elemen-elemen proses yang saling terkait.

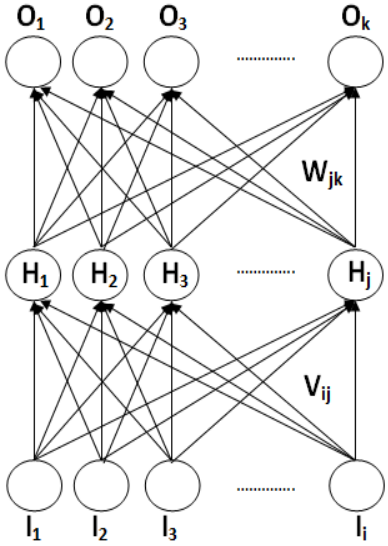
Pada FFNN elemen-elemen proses tersusun dalam beberapa lapis (*layer*) dan data input mengalir dari satu lapis ke lapis berikutnya secara berurutan. Pada tiap lapisan *input*data ditransformasikan kedalam lapis berikutnya secara *nonlinear* oleh elemen-elemen proses dan kemudian diproses ke lapis berikutnya. Akhirnya nilai *output O* yang didapat berupa nilai scalar atau vektor, yang dihitung pada lapisan output [11].

Untuk seleksi fitur *Neural Network* akan digunakan sebagai mesin klasifikasi yang kemudian akan di hitung selisih nilai *error* yang dihasilkan.

Dapat dikatakan bahwa jika sebuah fitur ketika dihapus membuat nilai *error* semakin bertambah besar maka fitur tersebut merupakan fitur yang baik dan direkomendasikan untuk digunakan, dan sebaliknya.

4. Algoritma Seleksi Fitur

Dari pembahasan sebelumnya, diketahui bahwa proses pemilihan fitur sangat mempengaruhi hasil dari proses klasifikasi. Makalah ini mengusulkan sebuah metode untuk melakukan seleksi fitur dengan menggabungkan metode *Ensemble Random Forest* dan *Neural Netwok*. Struktur dari metode yang diusulkan dapat dilihat pada Gambar 2.



Gambar 1. Neural network feedforward

4.1. Strategi Seleksi Fitur

Pada proses seleksi fitur, dataset dibagi menjadi tiga kelompok data dengan jumlah yang sama. Kelompok pertama adalah data yang digunakan untuk melakukan proses *training* dari *Neural Netwok*, kelompok data yang kedua digunakan untuk proses validasi pada *Neural Network*, kelompok data yang ketiga merupakan kelompok data yang digunakan untuk proses testing setelah fitur dari dataset terseleksi.

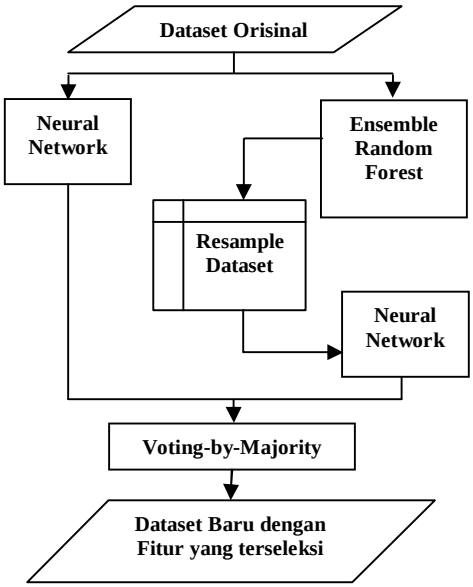
Pada *Neural Network* dua kelompok dataset yang kemudian dilakukan proses *training* dengan menggunakan data *training* untuk mendapatkan bobot yang optimal. Setelah didapatkan bobot yang paling optimal maka dilakukan proses validasi dengan menggunakan dataset validasi untuk mendapatkan nilai *error*. Nilai *error* didapatkan dengan membandingkan hasil dari *output* yang didapat dari *Neural Network* dengan target yang harus terpenuhi, dengan menggunakan formula

$$MSE = \frac{\sum(X-\bar{X})^2}{n} \quad (1)$$

Dimana $\bar{X}$merupakan *output* dari *Neural* Network dan X target, dan *n* adalah jumlah dari data pada dataset.

Sebagai contoh, terdapat sebuah dataset yang telah dibagi menjadi tiga buah kelompok dataset, dinotasikan

dengan D_1, D_2, dan D_3.Semua fituryang dimiliki D_1 akan digunakan oleh *Neural Network* pada proses *learning* untuk mencari nilai bobot yang optimal. Setelah nilai optimal didapatkan maka *Neural Network* akan menggunakan D_2 untuk proses validasi sehingga didapatkan nilai *MSE*. Nilai *MSE* tersebut akan disimpan untuk perbandingan di iterasi berikutnya. Kemudian akan di uji ulang dengan menghilangkan fitur dari D_1 dan D_2satu persatu. Dengan membandingkan nilai *MSE* yang dihasilkan dari proses validasi maka dapat dikatakan bahwa, jika sebuah fitur di hilangkan membuat nilai *MSE* semakin bertambah maka fitur tersebut merupakan fitur yang direkomendasikan untuk dipilih. Jika *MSE* semakin berkurangmaka dapat dikatakan bahwa fitur tersebut boleh dihilangkan, dan nilai MSE yang baru (lebih kecil) akan disimpan untuk perbandingan pada iterasi berikutnya.



Gambar 2. Model Metode Yang Diusulkan.

Terdapat nilai *treshold* dari selisih MSE yang masih di toleransi. Sehingga jika sebuah fitur dihilangkan mambuat nilai MSE semakin bertambah, tetapi masih selisih MSE yang didapat masih lebih kecil daripada *threshold* maka fitur akan dianggap masih layak untuk dihilangkan.

Ensemble Random Forest merupakan metode lain yang dilakukan untuk proses *resample data*. *Resample data* yang dimaksud dalam proses ini adalah membuat sebuah dataset baru berdasarkan dataset lama dan hasil klasifikasi *Ensemble Random Forest*. Dengan kata lain data *resample* merupakan data yang dibangun dengan *Reclustering* (pengelompokan ulang) berdasarkan *Ensemble Random Forest*. Kelompok dataset yang diproses dengan *Ensemble Random Forest* adalah kelompok data *training* dan validasi.

4.2. Strategi Voting

Strategi pemilihan fitur pada makalah ini dilakukan dengan membandingkan kedua hasil seleksi fitur, sesuai dengan strategi seleksi fitur yang sudah dibahas sebelumnya. Pada makalah ini menggunakan metode *Voting-by-majority* yaitu dengan memberikan nilai pada setiap fitur yang terdapat di sebuah dataset. Dengan memberikan nilai 0 (nol) untuk fitur yang tidak pernah muncul dari kedua metode seleksi fitur. Nilai 2 (dua) pada fitur yang muncul dari kedua proses seleksi fitur. Sebagai contoh, pada metode yang diusulkan terdapat sebuah *Neural Network* dan *Neural Network* yang mengambil data *resample* dari *EnsembleRandom Forest*. F_{NN} adalah fitur yang telah terseleksi yang dihasilkan oleh *Neural Network*, dan F_{RN} adalah fitur yang terseleksi yang dihasilkan oleh *Ensemble Random Forest* yang dilanjutkan dengan *Neural Network*. Jika sebuah fitur $f_i$muncul di F_{NN} dan F_{RN} maka dapat dikatakan bahwa $f_i$mendapatkan nilai 2 (dua), dan bisa dikatakan pula bahwa fitur tersebut sangat direkomendasikan untuk dipilih. Jika terdapat sebuah fitur yang muncul hanya di salah satu F_{NN} atau F_{RN} saja maka fitur tersebut memiliki nilai 1 (satu) yang masih memungkinkan untuk dipilih, daripada fitur yang memiliki nilai 0 (nol), yang tidak pernah muncul di keduanya.

5. Hasil dan Pembahasan

Pada bagian ini akan di evaluasi hasil dari metode yang diusulkan. Dengan menggunakan 3 (tiga) dataset yaitu *iris*, *lung cancer*, dan *semeion hand written digit* yang didapatkan dari *UCI (University of California) Machine Learning Repository*.

Dataset *iris* merupakan dataset yang paling terkenal dan dapat ditemukan di berbagai literatur pengenalan pola. Makalah *fisher* merupakan makalah klasik namun masih sering dirujuk sampai hari ini. Dataset *iris* memiliki tiga buah kelas yang masing-masing kelas memiliki 50 macam data (kasus), dimana setiap kelas mengacu pada jenis tanaman iris.

Tabel 1. Hasil Proses Seleksi Fitur Dari Dataset

Dataset	Tree Bagger	Threshold	
		0,05	0,10
Iris (4 fitur)	30	2	2
	50	2	2
Lung Cancer (56 fitur)	30	8	8
	50	6	6
Semeion (256 fitur)	30	20	19
	50	118	118

Tabel 2. Hasil Pengukuran

Dataset	Fitur Asal	Fitur Klasifikasi	MSE	Akurasi (%)
Iris	4	4	0,07	92,5
		2	0,04	94,4
Lung Cancer	56	56	0,22	91,6
		8	0,33	83,3
		6	0,27	75,0
Semeion	256	256	3,94	26,0
		20	6,06	20,0
		19	7,53	10,7
		118	7,35	10,4

Sedangkan dataset *lung cancer* merupakan data yang digunakan oleh Hong and Young untuk mengilustrasikan kemampuan dari *discriminant plane* yang optimal bahkan pada pengaturan yang buruk. Pada dataset *lung cancer* terdapat tiga jenis kanker paru-paru patologis. Pada dataset asli empat nilai untuk atribut kelima bernilai '-1' dan pada atribut ke-39 terdapat satu data yang bernilai '4' data tersebut nilainya telah dirubah menjadi '?' (*unknown*). Perubahan nilai pada dataset tersebut ditujukan agar mendapatkan nilai normal *integer* antara 0-3. Namun pada uji coba dalam makalah ini, nilai-nilai pada dataset menggunakan nilai-nilai awal agar tidak ada data yang bernilai '?' atau *missing value*.

Pada dataset *semeion handwritten digit* terdapat 256 atribut yang didapatkan dari kolom *pixel* sebuah gambar 16x16 *pixel* persegi. Dataset *semeion* diambil dari 1593 angka tulisan tangan dari 80 orang yang telah dipindai, membentang dalam 16x16 *pixel* persegi dengan 256 derajat keabuan. Kemudian setiap *pixel* pada gambar dirubah kedalam bentuk *binner* (0/1) menggunakan nilai *threshold* yang tetap. Setiap orang menulis diatas kertas semua angka dari 0 (nol) sampai 9 (sembilan) sebanyak dua kali.

Dengan menggunakan ketiga dataset tersebut, didapatkan hasil dari seleksi fitur seperti terlihat pada Tabel 1. Pada dataset *iris* terseleksi dua fitur pada setiap percobaan. Pada dataset *lung cancer* terseleksi delapan fitur dari percobaan dengan jumlah *tree bagger*=30, dan dengan jumlah*tree bagger*=50 terseleksi enam fitur. Dari 256 fitur yang terdapat pada dataset *semeion* terseleksi 20 fitur pada percobaan dengan *threshold*=0,05 dan jumlah *tree bagger*=30.

Hasil klasifikasi dengan menggunakan fitur yang terseleksi dari setiap dataset dapat dilihat pada Tabel 2. Hasil klasifikasi menggunakan dataset *iris* dengan fitur yang telah terseleksi memiliki tingkat akurasi 94,4%. Terlihat bahwa akurasi lebih tinggi dibandingkan hasil klasifikasi dengan menggunakan semua fitur yaitu 92,5%.

Anomali terjadi pada dataset *lung cancer*. Hasil klasifikasi menggunakan dataset *lung cancer* dengan semua fitur memiliki tingkat akurasi 91,6%. Tingkat akurasi tersebut lebih tinggi dibandingkan dengan tingkat akurasi yang didapatkan dari proses klasifikasi menggunakan fitur yang telah terseleksi. Diantaranya yaitu 83,3% untuk delapan fitur yang terseleksi, dan 75,0% untuk enam fitur yang terseleksi. Anomali yang terjadi pada proses seleksi fitur, disebabkan karena pada dataset *lung cancer* terdapat tiga variasi nilai untuk setiap data. Terbukti bahwa pada dataset *semeion* juga terjadi

anomali. Karena pada dataset *semeion* terdapat dua variasi nilai pada setiap data yaitu 0 (nol) dan 1 (satu).

Terjadinya anomali pada dataset *lung cancer*, dan *semeion* dikarenakan setiap fitur pada dataset memiliki variasi nilai yang memiliki kesamaan, sehingga antara fitur yang satu dengan fitur yang lain terdapat irisan yang cukup besar bahkan hampir berhimpitan. Hal inilah yang membuat hampir semua fitur pada kedua dataset tersebut dapat dianggap sebagai fitur yang penting, sehingga proses seleksi menjadi semakin rumit. Proses seleksi fitur yang semakin rumit, semakin sulit pula diproses oleh metode yang diusulkan sehingga hasil keluaran, yaitu fitur-fitur yang terseleksi merupakan fitur yang dianggap paling baik untuk proses seleksi pada dataset *training* dan validasi. Dan ketika diproses dengan menggunakan data yang lain maka nilai akurasi yang didapat akan semakin menurun karena ketika seleksi fitur terhadap kedua dataset (*semeion* dan *lung cancer*) tersebut,fitur yang dihasilkan akan menurunkan tingkat generalisasi proses klasifikasi terhadap dataset tersebut. Dengan kata lain *over fitting* akan terjadi terhadap data *training* dan data validasi dari kedua dataset tersebut.

6. Kesimpulan

Pada makalah ini telah diusulkan sebuah metode baru dalam seleksi fitur dengan menggunakan *ensemble random forest* dan *neural network*. Dan dari hasil percobaan, dapat disimpulkan bahwa metode seleksi fitur dengan menggunakan *Ensemble random forest* dan *neural network* dapat meningkatkan akurasi pada dataset yang memiliki tingkat variasi data yang tinggi, dan berlaku kebalikannya, pada dataset memilki tingkat variasi rendah, maka tingkat akurasi dari metode seleksi fitur yang diusulkan justru menurun. Hal ini disebabkan karena proses seleksi fitur akan semakin rumit pada dataset yang dengan fitur yang memiliki irisan cukup lebar satu sama lain bahkan hampir berhimpit. Dan fitur hasil seleksi akan menurunkan generalisasi dari proses klasifikasi, dengan kata lain *over fitting* terjadi pada dataset *training* dan validasi.

Penelitian selanjutnya adalah dengan memberikan variasi dataset yang lebih banyak, baik dari segi jumlah fitur, ukuran, maupun nilai data.

Daftar Pustaka

[1] Fukunaga, K., *Introduction To Statistical Pattern recognition. Academic Press*, New York, 1972.

[2] Mucciardi, A., Gose, E.E., *A Comparison Of Seven Techniques for Choosing Subsets of Pattern Recognition Properties*, IEEE Trans. Comput. 20 (9), 1023-1031, 1971.

[3] Steppe, J. M., Bauer, K. W., *Improved Feature Screening in FeedForward Neural Networks*, Neurocomputing 13, 47-58, 1996.

[4] Steppe, J. M., Bauer, K. W., *Integrated Feature and Architecture Selection*, IEEE Trans. Neural Networks 7 (4), 1007-1014, 1996.

[5] Duda, R.O., Hart, P.E., *Classification and Scene Analysis.Wiley*, New York, 1973.

[6] Bishop, C.M., *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, 1995.

[7] Rong Jia, Li Gang, Chen Yi-Ping P., *Accoustic Feature Selection For Automatic Emotion Recognition From Speech*, Information Processing and Management 45 (2009) 315–328, 2009.

[8] Verikas, A., Bacauskiene, M., *Feature Selection With Neural Network*, Pattern Recognition Letters 23 (2002) 1323–1335, 2001.

[9] Hall Lawrence, O., Bowyer Kevin, W., *Bandfield Robert, Why are Neural Network Sometimes Much More Accurate than Decision Trees: An Analysis on a Bio-Informatics Problem*, IEEE International Conference of Systems, Man & Cybernetics. Washington D. C., pp. 2851-2856, 2003.

[10] Martin Anthony and Peter Bartlett, *Neural Network Learning: Theoretical Foundations*, Cambridge University, 1999.

[11] Suhartono, *Feedforward Neural Network Untuk Pemodelan Runtun Waktu*, Universitas Gadjah Mada, Yogyakarta, 2007.