SELEKSI FITUR MENGGUNAKAN METODE KOMBINASI ALGORITME GENETIKA DAN *SEQUENTIAL MINIMAL OPTIMIZATION* UNTUK KLASIFIKASI HALAMAN WEB

Hendri Noviyanto[1] , Teguh Bharata Adji[2] , Noor Akhmad Setiawan[3]
[1]Teknik Elektro dan Teknologi Informasi, Universitas Gadjah Mada, Yogyakarta Indonesia
nhendri0@gmail.com
[2] Teknik Elektro dan Teknologi Informasi, Universitas Gadjah Mada, Yogyakarta Indonesia
adji@mti.ugm.ac.id
[3] Teknik Elektro dan Teknologi Informasi, Universitas Gadjah Mada, Yogyakarta Indonesia
noorwewe@ugm.ac.id

Abstract

Klasifikasi halaman web memiliki masalah dalam pemilihan fitur yang relevan sehingga mempengaruhi nilai akurasi yang didapatkan. Masalah tersebut dapat ditangani menggunakan metode seleksi fitur. Metode seleksi fitur bekerja dengan cara mengevaluasi dan menyeleksi fitur yang relevan dan informatif dari setiap dokumen halaman web. Fitur yang dimaksud merupakan sebuah token atau kata-kata yang muncul dalam halaman web. Pada penelitian ini, metode seleksi fitur wrapper yang digunakan untuk mengkombinasikan algoritme genetika sebagai subset selection dengan pengklasifikasi sequential minimal optimization sebagai attribute evaluator. Metode yang diusulkan mampu mereduksi fitur sebesar 45.20% untuk dataset WebKb dan 42.73% untuk dataset r8. Secara keseluruhan, nilai akurasi pada proses klasifikasi halaman web juga meningkat setelah penerapan metode seleksi fitur dari 76.35% menjadi 78.05% untuk dataset webkb dan 86.47% menjadi 86.61% untuk dataset r8.

Keywords **:** *Algoritme genetika, klasifikasi halaman web, seleksi fitur, SMO, wrapper.*

1. Pendahuluan

Jumlah dokumen halaman web berkembang semakin besar. Hal ini menyebabkan pengguna mengalami kesulitan dalam menemukan halaman web yang relevan dalam pencarian. Banyaknya jumlah halaman web membutuhkan proses klasifikasi untuk mengelompokkan halaman web sesuai dengan kategorinya. Namun, banyaknya fitur pada setiap halaman web menyebabkan munculnya masalah baru, yaitu bagaimana memilih fitur yang relevan dan informatif pada setiap halaman web. Oleh karena itu, perlu diterapkan teknik seleksi fitur untuk mendapatkan fitur yang optimal dari setiap halaman web. Seleksi fitur akan mengevaluasi dan memilih fitur yang penting serta menghapus informasi yang tidak berguna, berlebihan, atau jarang muncul (Il-Seok Oh, Jin-Seon Lee, & Byung-Ro Moon, 2004). Hal tersebut dapat mengurangi jumlah data yang akan diproses, sehingga mampu mendapatkan fitur yang optimal. Fitur merupakan sebuah token atau kata-kata yang muncul dalam halaman web.

Pada klasifikasi halaman web, proses seleksi fitur membutuhkan penanganan khusus. Banyaknya fitur yang terdapat di halaman web menjadi salah satu penyebab rendahnya nilai akurasi yang didapatkan. Oleh karena itu, pada penelitian ini diterapkan sebuah teknik seleksi fitur untuk memilih fitur yang relevan dan informatif. Teknik tersebut menggunakan metode *wrapper* dengan mengkombinasikan algoritme genetika (GA) dengan algoritme *sequential minimal optimization* (SMO)

2. Tinjauan Pustaka

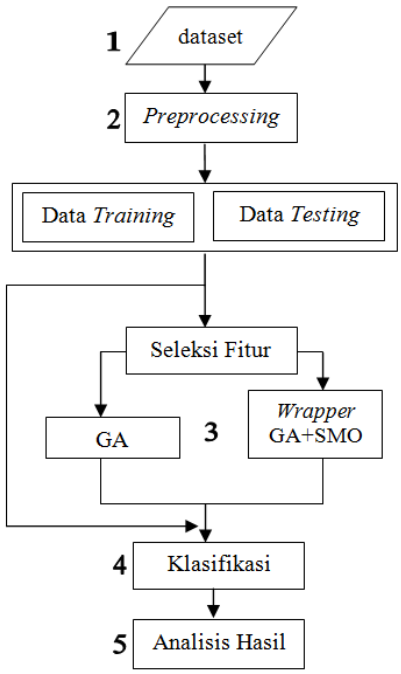
Penelitian mengenai pengklasifikasian secara lebih luas telah banyak dilakukan dalam beberapa tahun belakangan dengan menawarkan

berbagai macam teknik, metode dan algoritme yang berbeda. Penelitian (H. Ge & T. Hu, 2014), menggunakan algoritme genetika (GA) untuk melakukan proses seleksi fitur pada klasifikasi halaman web. Penelitian ini mengkombinasikan GA dengan *mutual information* (FSGM) sebagai usulan metode yang akan digunakan. Penelitian tersebut menggunakan dataset dari UCI *machine learning* yaitu *Iris*, *Brecan*, *Customer*, *Wine*. Setelah dilakukan seleksi fitur menggunakan metode FSGM, hasil akurasi klasifikasi yang didapatkan adalah *Iris* 100%, *Brecan* 97.5%, *Customer* 93.9%, *Wine* 98.1%. Penelitian (P. S. Tang, X. L. Tang, Z. Y. Tao, & J. P. Li, 2014) melakukan kombinasi seleksi fitur menggunakan GA dan *mutual information* (MI-GA) sama dengan penelitian (H. Ge & T. Hu, 2014). Penelitian ini membandingkan algoritme seleksi fitur seperti GA, *relief*, dan MI-GA. Hasilnya adalah MI-GA mampu mendapatkan akurasi terbaik sebesar 0.87. Dataset yang digunakan oleh (P. S. Tang et al., 2014) meliputi *Anneal*, *Audiology*, *German*, *Ionosphere*, *Sick*, *Sonar*, *Splice*, dan *Waveform*. Penelitian (S. A. Özel, 2011) menggunakan GA sebagai algoritme seleksi fitur. Penelitian ini diuji coba menggunakan beberapa algoritme pengklasifikasi seperti *Naïve Bayes Multinomial* (NBM), *k-Nearest Neighbour, dan Decision Trees*. Hasil tertinggi yang didapatkan mampu meningkatkan akurasi sebesar 96% dengan algoritme NBM. Dalam penelitian (S. A. Özel, 2011) dataset yang digunakan berasal dari WebKB dan DBLP dengan 5 kategori. Proses klasifikasi dilakukan dengan cara melakukan pembagian data *training* 75% dan *testing* 25%.

Pada penelitian ini, proses seleksi fitur yang diusulkan menggunakan metode *wrapper* yang mengkombinasikan GA dengan SMO. Pada proses yang dilakukan, GA digunakan sebagai *subset selection* dan SMO digunakan untuk *attribute evaluator*. Penelitian ini bertujuan untuk melakukan proses seleksi fitur guna mendapatkan fitur yang optimal serta membandingkan efek penggunaan seleksi fitur dengan tanpa menggunakan seleksi fitur terhadap perolehan tingkat akurasi.

3. Metode Penelitian

Dalam penelitian ini, terdapat beberapa tahapan yang dilakukan, mulai dari proses *preprocessing* dataset yang digunakan sebagai masukan, penentuan data *training* dan *testing*, proses seleksi fitur, klasifikasi dan analisis hasil. Diagram alur penelitian dapat dilihat pada Gambar 1 berikut.



Gambar 1. Diagram Alur Penelitian.

Berdasarkan pada Gambar 1 diatas, alur penelitian dapat dijelaskan sebagai berikut:

1) Dataset

Penelitian ini menggunakan sejumlah dataset yang diperoleh dari website. Dataset pertama adalah WebKB ("WebKB," 2016) dan r8 ("R52 dan R8," 2016). Jumlah Dataset yang digunakan dapat dilihat pada Tabel 1 berikut.

Tabel 1. Dataset Penelitian

Nama Dataset	Jumlah Dokumen	Jumlah Fitur	Jumlah Kategori
WebKb	2803	1002	4
R8	5485	1011	8

2) *Preprocessing*

Pada tahap ini dataset yang telah didapatkan diproses dengan tujuan untuk mendapatkan data yang bersih dari *noise* dan sesuai dengan format data masukan agar dapat digunakan sebagai data *training* dan *testing*. Pada tahap *preprocessing* metode yang digunakan meliputi *Minimal Term Frequency*,

Tokenizing, dan *Stoplist*. *Minimal Term Frequency* merupakan sebuah metode untuk menghilangkan *term* yang kurang dari nilai "*n*" ("*n*" adalah nilai masukan yang ditentukan). *Tokenizing* adalah sebuah metode untuk memisahkan sebuah kalimat menjadi sebuah *term-term* dengan tujuan memudahkan proses *learning*. *Stoplist* adalah metode yang digunakan untuk menghilangkan kata yang kurang penting dan tidak memiliki makna, contohnya : "*and*", "*or*", "*are*", dan sebagainya.

3) Seleksi Fitur

Pada tahap ini, proses seleksi fitur dilakukan sebanyak 2 kali, yaitu proses dengan menggunakan GA tanpa kombinasi dan *wrapper* yang terdiri dari kombinasi GA dengan SMO. Pada metode *wrapper*, GA bekerja sebagai *subset selection*, sedangkan SMO bekerja sebagai *attribut evaluator*. Hal tersebut dilakukan untuk mendapatkan fitur yang optimal dari dokumen halaman web.

4) Klasifikasi

Klasifikasi merupakan sebuah proses yang digunakan untuk mengelompokkan sebuah data atau dokumen ke dalam kategori yang memiliki kemiripan data.

Proses klasifikasi dilakukan sebanyak 3 kali. Hal ini digunakan untuk mendapatkan nilai akurasi dari masing-masing metode yang diterapkan. Proses pertama klasifikasi dilakukan tanpa seleksi fitur, proses klasifikasi kedua menggunakan GA, dan proses klasifikasi ketiga menggunakan metode *wrapper* yang terdiri dari GA dan SMO. Pada proses klasifikasi algoritme pengklasifikasi yang digunakan adalah *NaiveBayes*.

5) Analisis Hasil

Pada tahap ini dilakukan proses analisis terhadap skenario proses pengujian klasifikasi halaman web. Skenario pertama adalah membandingkan kinerja algoritme seleksi fitur dengan menghitung presentase keberhasilan dalam menyeleksi fitur. Skenario kedua adalah membandingkan nilai akurasi dari proses klasifikasi dengan tanpa seleksi fitur, dengan seleksi fitur GA tanpa kombinasi, dan dengan metode *wrapper*. Hasil akhir dari proses analisis dapat digunakan untuk mengambil kesimpulan apakah penggunaan seleksi fitur mampu meningkatkan nilai akurasi dari proses klasifikasi halaman web dan apakah hasil dari metode *wrapper* yang mengkombinasikan algoritme GA dan SMO lebih baik daripada algoritme GA tanpa kombinasi.

3.1. Seleksi Fitur

3.1.1 Metode Wrapper

Menurut (P. S. Tang et al., 2014) *wrapper* merupakan metode yang digunakan untuk mengevaluasi *subset* melalui proses pembelajaran. Untuk mengevaluasi *subset* metode *wrapper* perlu melatih sebuah algoritme pengklasifikasi untuk mendapatkan hasil yang bagus. Metode *wrapper* bekerja dengan menggunakan teknik *subset selection*, kemudian akan dievaluasi oleh algoritme pengklasifikasi.

3.1.2 Algoritme Genetika

Algoritme Genetika (GA) merupakan algoritme yang dikembangkan dari konsep teori evolusi makhluk hidup oleh (Goldberg & Holland, 1988). Dengan menggunakan elemen-elemen dasar dari evolusi makhluk hidup seperti reproduksi, kawin silang, dan mutasi GA mencoba mendapatkan solusi optimal dari masalah yang dihadapi (Santosa & Willy, 2011). Dalam GA prosedur pencarian nilai optimal hanya berdasarkan dari nilai fungsi tujuan, tidak ada pemakaian teknik *gradient* atau teknik kalkulus(Santosa & Willy, 2011). GA dalam kasusnya banyak digunakan untuk menyelesaikan masalah *TSP*, *VRP*, dan *crew scheduling* untuk *airline*. Namun, penelitian (S. A. Özel, 2011),(H. Ge & T. Hu, 2014),(P. S. Tang et al., 2014),(Chaikla & Qi, 1999) menggunakan GA sebagai metode seleksi fitur. Dari beberapa penelitian tersebut, hasil akurasi yang didapatkan meningkat beberapa persen sehingga mampu meningkatkan kinerja dari algoritme pengklasifikasi. Secara garis besar GA dapat di jelaskan sebagai berikut (Santosa & Willy, 2011):

1) Bangkitkan populasi awal
2) Set iterasi t =1
3) Pilih individu terbaik untuk mengantikan individu yang lain
4) Lakukan seleksi untuk memilihi induk yang akan dikawin silangkan
5) Lakukan proses kawin silang antar induk yang telah terpilih
6) Menentukan jumlah individu dalam populasi untuk proses mutasi
7) Jika belum konvergen set t = t + 1, kembali ke langkah 2.

3.1.3 Algoritme *Sequential Minimal Optimization* (SMO)

Algoritme *Sequential Minimal Optimization* (SMO) ditemukan oleh (Platt, 1998) dalam riset yang dilakukan di *microsoft*. Menurut (Platt, 1998) SMO merupakan algoritme klasifikasi yang bekerja secara sederhana dan mudah diaplikasikan. SMO dimanfaatkan untuk menutup kekurangan dari algoritme *Support Vector Machine* (SVM) pada masalah *Quadratic Programming* (QP). SMO termasuk dalam metode *Decomposition* yang bekerja berdasarkan prinsip "*Working Set*" (Fan, Chen, & Lin, 2005). Metode ini hanya mengubah beberapa *Multiplier* α_i dalam jumlah tertentu pada setiap iterasi, sementara nilai yang lainnya tetap (Santosa, 2010).Tidak seperti metode sebelumnya, SMO memilih menyelesaikan masalah kecil yang mungkin pada setiap langkahnya. Pada standart masalah QP, masalah optimasi pada algoritme SMO melibatkan dua *Lagrange Multipliers*. SMO melibatkan *Working Set* berelemen dua sehingga pencarian solusi optimal dapat dilakukan secara analitis. Hal ini tentu akan mengakibatkan jumlah iterasi semakin bertambah, akan tetapi karena waktu yang dibutuhkan dalam setiap iterasi sangat kecil maka waktu total pelatihan menjadi lebih singkat (Santosa, 2010). Secara singkat algoritme SMO bekerja sebagai berikut:

1) Temukan *lagrange multiplier* yang melanggar kondisi *karush-kuhn-tucker* (KTT) dalam optimasi.
2) Pilih *multiplier* kedua dan optimalkan pasanganya.
3) Ulangi langkah 1 dan 2 hingga konvergen. Ketika semua *lagrange multipliers* memenuhi kondisi KTT maka masalah terselesaikan.

4. Eksperimen dan Analisis

Penelitian ini menggunakan aplikasi *machine learning* WEKA (Garner, 1995) untuk proses klasifikasi. Proses klasifikasi dilakukan menggunakan algoritme pengklasifikasi *NaiveBayes*. Skema pengujian proses pembagian data *training* dan *testing* menggunakan 10-*fold cross validation*. Dataset yang digunakan bersumber dari WebKB ("WebKB," 2016) dengan jumlah 4 jumlah kategori atau *class* dan dataset r8 ("R52 dan R8," 2016) dengan jumlah 8 kategori.

Dalam penelitian ini, terdapat beberapa skenario pengujian yang meliputi proses komparasi antara metode yang digunakan.

Skenario pengujian pertama pada penelitian ini meliputi komparasi metode seleksi fitur antara GA tanpa kombinasi (GA murni) dengan metode *wrapper* yang mengkombinasikan GA dengan SMO. Skenario pengujian kedua adalah proses komparasi terhadap nilai akurasi. Proses komparasi tersebut meliputi nilai akurasi yang dihasilkan pada saat tidak menggunakan metode seleksi fitur dengan pada saat menggunakan metode seleksi fitur, baik menggunakan GA maupun metode *wrapper* pada saat proses klasifikasi.

4.1. Pengujian metode seleksi fitur

Hasil pengujian metode seleksi fitur menggunakan GA dan metode *wrapper* dapat dilihat pada Tabel 2:

Tabel 2**.** Hasil Proses Seleksi Fitur

Dataset	**Jumlah Fitur Awal**	**Seleksi Fitur %**	
		GA	***Wrapper***
WebKB	1002	73.85%	45.20%
R8	1011	75.15%	42.73%

Pada Tabel 2 diatas, jumlah fitur awal adalah 1002 dan 1011. Setelah proses seleksi fitur tersebut berkurang sebesar 73.85% dan 75.15% ketika diseleksi menggunakan GA. Proses seleksi menggunakan metode wrapper mampu menyeleksi fitur sebesar 45.20% dan 42.73%. Menurut presentase yang telah didapatkan, GA mampu menyeleksi fitur lebih banyak dibandingkan dengan metode *wrapper*.

4.2. Pengujian Proses Klasifikasi

Pengujian kedua adalah mengkomparasi nilai akurasi yang didapatkan pada saat proses klasifikasi dilakukan. Hasil dari proses klasifikasi ditabelkan dan dapat dilihat pada Tabel 3.

Tabel 3**.** Nilai Akurasi Hasil Proses Klasifikasi

Dataset	**Tanpa Seleksi Fitur**	***GA***	***Wrapper***
WebKB	76.34%	77.09%	78.05%
R8	86.47%	85.56%	86.61%

Pada Tabel 3 diatas, dapat dilihat komparasi nilai akurasi hasil proses klasifikasi. Klasifikasi pada dataset WebKB tanpa seleksi fitur memiliki nilai akurasi dengan presentase 76.34%, GA 77.09%, dan *wrapper* 78.05%. Nilai akurasi yang didapatkan meningkat mulai dari tanpa seleksi fitur, GA, dan *wrapper*. Sedangkan pada dataset r8, nilai akurasi yang didapatkan tanpa seleksi fitur 86.47%, GA 85.56%, dan *wrapper* 86.61%. Dalam pengujian dataset r8, seleksi fitur menggunakan GA mengalami penurunan nilai akurasi. Hal tersebut disebabkan karena jumlah data r8 lebih besar dan hanya memiliki fitur yang sangat sedikit sehingga menyebabkan penurunan nilai akurasi. Sedangkan pada saat penggunaan metode *wrapper* nilai akurasi tetap meningkat, meskipun hanya sedikit. Hal ini membuktikan bahwa metode yang diusulkan mampu meningkatkan nilai akurasi

5. Kesimpulan

Berdasarkan hasil penelitian ini dapat disimpulkan bahwa penggunaan metode seleksi fitur yang tepat dapat membantu meningkatkan nilai akurasi yang didapatkan dalam membantu proses klasifikasi halaman web.

Klasifikasi halaman web dengan setelah menggunakan seleksi fitur *wrapper* memiliki nilai akurasi lebih unggul daripada ketika menggunakan GA.

Daftar Pustaka

Chaikla, N., & Qi, Y. (1999). Genetic Algorithms in Feature Selection. *Computer Science and Information Management Program Asian Institute of Technology*.

Fan, R.-E., Chen, P.-H., & Lin, C.-J. (2005). Working set selection using second order information for training support vector machines. *Journal of Machine Learning Research*, *6*(Dec), 1889–1918.

Garner, S. R. (1995). Weka: The waikato environment for knowledge analysis. In *Proceedings of the New Zealand computer science research students conference* (pp. 57–64). Citeseer.

Goldberg, D. E., & Holland, J. H. (1988). Genetic Algorithms and Machine Learning. *Machine Learning*, *3*(2), 95–99.

H. Ge, & T. Hu. (2014). Genetic Algorithm for Feature Selection with Mutual Information. In *Computational Intelligence and Design (ISCID), 2014 Seventh International Symposium on* (Vol. 1, pp. 116–119). https://doi.org/10.1109/ISCID.2014.122

Il-Seok Oh, Jin-Seon Lee, & Byung-Ro Moon. (2004). Hybrid genetic algorithms for feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(11), 1424–1437. https://doi.org/10.1109/TPAMI.2004.105

Platt, J. C. (1998). Fast Training of Support Vector Machines using Sequential Minimal Optimization (pp. 42–64). USA: Microsoft Research. Retrieved from http://www.research.microsoft.com/~jplatt

P. S. Tang, X. L. Tang, Z. Y. Tao, & J. P. Li. (2014). Research on feature selection algorithm based on mutual information and genetic algorithm. In *Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2014 11th International Computer Conference on* (pp. 403–406). https://doi.org/10.1109/ICCWAMTIP.2014.7073436

R52 dan R8. (2016). Retrieved from http://csmining.org/index.php/r52-and-r8-of-reuters-21578.html

Santosa, B. (2010). Tutorial Support Vector Machine. *Teknik Idustri, ITS.[Online]. Tersedia: Http://www. Google. Co. Id/url*.

Santosa, B., & Willy, P. (2011). *Metoda Metaheuristik Konsep dan Implementasi*. Surabaya: Guna Widya.

S. A. Özel. (2011). A genetic algorithm based optimal feature selection for Web page classification. In *Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on* (pp. 282–286). https://doi.org/10.1109/INISTA.2011.5946076

WebKB. (2016). [Organisasi]. Retrieved June 10, 2016, from htpp://csmining.org/index.php/webkb.html

Biodata Penulis

Hendri Noviyanto, memperoleh gelar S1 di Universitas Muhammadiyah Surakarta. Proses menyelesaikan gelar S2 di Universitas Gadjah Mada.

Teguh Bharata Adji, memperoleh gelar S1 di Universitas Gadjah Mada. Memperoleh gelar S2 di Universitas Doshisha, Jepang. Memperoleh gelar S3 di Universitas Teknologi Petronas, Malaysia. Saat ini menjadi pengajar tetap di Teknik Elektro dan Teknologi Informasi Universitas Gadjah Mada.

Noor Akhmad Setiawan, memperoleh gelar S1 di Universitas Gadjah Mada. Memperoleh gelar S2 di Universitas Gadjah Mada. Memperoleh gelar S3 di Universitas Teknologi Petronas, Malaysia. Saat ini menjadi pengajar tetap di Teknik Elektro dan Teknologi Informasi Universitas Gadjah Mada.

BERITA ACARA PELAKSANAAN HASIL SEMINAR SESI PARALEL KNASTIK 2016

Judul : Seleksi Fitur menggunakan Metode Kombinasi Algoritme Genetika dan Sequential Minimal Optimization untuk Klasifikasi Halaman Web

Pemakalah : Hendri Noviyanto, Teguh Bharata Adji, Noor Akhmad Setiawan

Moderator : Gloria Virginia, S.Kom., MAI, Ph.D

Notulis : Yoas

Peserta : 11 orang di ruang : D.3.2

Tanya Jawab :

Masukan & Pertanyaan (oleh sdr. Alz):

1. Mungkin bisa lebih dijelaskan bahwa penelitian bisa lebih baik dari yang ada, terutama pada algoritma pencariannya (Signifikasi/Posisi Penelitian berdasarkan kalimat pembuka awal presentasi penyaji).
2. Sebenarnya yang perlu ditekankan di kombinasi algoritmanya, kalau hanya tingkat prosentase masih sedikit kurang jelas.
3. Pernyataan awal penyaji perlu dikoreksi, karena menyatakan *Google* ada kekurangan dan penyaji meneliti berdasar kekurangan tersebut.

Jawaban:

1. Para *programmer* dan peneliti pada *Google* adalah orang-orang yang sudah tergolong *experts*. Kalau ditekankan punya saya lebih baik, mustahil, pasti kurang baik karena hanya diteliti oleh penyaji seorang yang masih *newbie*.
   Namun yang saya temukan, *Google* terkadang masih belum bisa menemukan dokumen berisi informasi yang saya inginkan pada 10 halaman pertama yang saya telusuri.
2. Menerapkan GA + SMO karena belum ada *paper* yang menggunakan algoritma tersebut, dan merupakan pembanding dengan pencarian yang tidak menggunakan seleksi fitur. GA + SMO memberi hasil yang lebih baik. GA + SMO memberikan juga hasil komputasi klasifikasi lebih baik. Langkah seleksi fitur dilakukan sebelum klasifikasi. Setelah proses seleksi selesai, data akan di *save* menjadi model. Proses paling terakhir adalah klasifikasi. Algoritma ini menggunakan *time* dan *cost* lebih kecil. Saya memerlukan waktu setengah hari untuk seleksi fitur dengan algoritma ini.

Masukan Seminar :

Signifikasi penelitian terutama berhubungan dangan posisi penelitian ini terhadap aplikasi yang sudah ada, yaitu *Google*, kurang jelas.

Yogyakarta, 19 November 2016

Moderator Kelas	Penyaji Makalah
Gloria Virginia, S.Kom., MAI, Ph.D.	Hendri Noviyanto