Pencarian Fitur Optimal Halaman Web Menggunakan Kombinasi Algoritme Genetika Dan *Naivebayes* Untuk Klasifikasi

Hendri Noviyanto

Sistem Komputer Universitas Surakarta

hendrinoviyantoo@gmail.com

ABSTRACT

The process of classification of web pages has problems in the selection of relevant features that affect the acquisition of value accuracy. These problems can be handled using feature selection techniques. Feature selection method works by evaluating and selecting relevant and informative features of each web page document. Features are a token or informative words that often appear on web pages. In this study, the method used is the genetic algorithm and decision tree incorporated in the wrapper technique. Genetic algorithms are used as subset selection and decision tree as attribute evaluators. The proposed method is able to reduce features by 45.39% for the WebKB dataset and 56.71% for the r8 dataset. Overall, the results of the classification process increased although not too significant.

Keywords *: Algoritme genetika, decision tree, klasifikasi halaman web, naivebayes, seleksi fitur, wrapper.*

I. PENDAHULUAN

Halaman web terus berkembang menjadi semakin besar, dengan kata lain jumlah dokumen akan semakin banyak. Hal semacam ini menyebabkan proses pencarian halaman web yang relevan menjadi lebih sulit. Jumlah halaman web yang semakin banyak membutuhkan proses klasifikasi untuk mengelompokkan halaman tersebut sesuai dengan kategorinya. Pengelompokan akan memudahkan proses pencarian, namun banyaknya fitur yang terdapat pada dokumen halaman web tersebut menyebabkan masalah baru yaitu bagaimana memilih fitur yang relevan dan informative pada setiap halaman web. Seperti yang telah disebutkan bahwa pengelompokkan dapat memudahkan pencarian, akan tetapi jika jumlah dokumennya sangat besar. Maka fitur yang relevan dan informatiflah yang dapat membantu mempercepat proses pencarian. Oleh Karena itu, perlu adanya penerapan teknik seleksi fitur untuk mendapatkan fitur yang optimal dari setiap halaman web. Seleksi fitur akan mengevaluasi dan memilih fitur yang penting kemudian menghapus fitur yang tidak berguna, berlebihan atau jarang muncul (Il-Seok Oh, Jin-Seon Lee, & Byung-Ro Moon, 2004). Hal tersebut dapat mengurangi data yang akan diproses, sehingga mampu mendapatkan fitur yang optimal. Fitur dalam dokumen halaman web adalah sebuah token atau kata-kata yang sering muncul dalam dokumen halaman web.

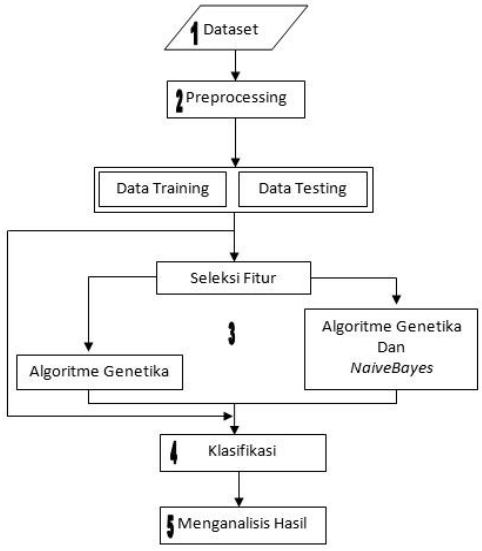
Proses klasifikasi halaman web, penanganan untuk proses seleksi fitur dirasa membutuhkan penganganan yang khusus. Karena banyaknya fitur yang terdapat di dokumen halaman web menjadi salah satu penyebab rendahnya nilai akurasi yang didapatkan. Oleh Karena itu pada penelitian ini diterapkan sebuah teknik seleksi fitur untuk memilih fitur yang relevan dan informative. Teknik tersebut menggunakan metode kombinasi algoritme yaitu algoritme genetika dan naivebayes. Teknik ini dikenal dengan sebut teknik wrapper.

II. 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-earest 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 mengusulkan metode seleksi fitur dengan kombinasi algoritme genetika dengan algoritme naivebayes. Kedua metode tersebut dibungkus dalam satu teknik yang dikenal dengan sebut teknik wrapper. Proses yang dilakukan algoritme genetika dalam penelitian ini sebagai subset selection, sedangkan peran algoritme naivebayes sebagai attribute evaluatornya. Penelitian ini akan membandingkan nilai akurasi dari hasil klasifikasi dengan memanfaat fitur yang telah diseleksi yang dibandingkan dengan fitur yang belum diseleksi.

III. METODOLOGI



Gambar 1. Diagram Alur Penelitian

Penelitian ini memiliki beberapa tahapan yang dilakukan sebelum proses klasifikasi dikerjakan. Proses pertama yang dilakukan adalah preprocessing dataset yang akan digunakan

sebagai data input, kemudian menentukan data yang akan digunakan untuk proses testing dan training, proses seleksi fitur dan proses analisis hasil. Beberapa proses tersebut dituangkan dalam bentuk flowchart diagram alur yang dapat dilihat pada gambar 1.

Diagram alur penelitian diatas, dijabarkan sebagai berikut.

A. Dataset

Dataset yang digunakan dalam penelitian ini adalah WebKB ("WebKB," 2016) dan r8 ("R52 dan R8," 2016). Dataset tersebut diperoleh dari website penyedia dataset. 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

B. 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.

C. Seleksi Fitur

Pada proses seleksi fitur dilakukan sebanyak 2 kali, yaitu proses dengan menggunakan GA dan proses menggunakan GA yang dikombinasikan dengan NaiveBayes. Pada teknik wrapper, GA bekerja sebagai subset selection, sedangkan NaiveBayes bekerja sebagai attribut evaluator. Hal tersebut dilakukan untuk mendapatkan fitur yang optimal dari dokumen halaman web.

D. 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 teknik wrapper yang terdiri dari GA dan NaiveBayes. Pada proses klasifikasi algoritme pengklasifikasi yang digunakan adalah Decision Tress (J48).

E. 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 teknik 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 teknik wrapper yang mengkombinasikan algoritme GA dan NaiveBayes lebih baik daripada algoritme GA tanpa kombinasi.

1) Seleksi Fitur

a. Teknik Wrapper

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

b. 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
8. Kembali ke langkah 2

c. Algoritme Naïve Bayes

Algoritme NiaveBayes merupakan metode yang menerapkan teorema Bayes yang berdasarkan nilai probabilitas. Algoritme NaiveBayes memiliki kemampuan yang cukup baik dalam mengklasifikasikan data, dengan catatan data latih yang cukup besar. Algoritme NaiveBayes merupakan sebuah algoritme analisis yang berkerja dengan cara mengolah data numerik.

Algoritme NaïveBayes dapat dirumuskan ke dalam persamaan 1.1 Dimana p(A|B) merupakan sebuah conditional probability A yang mempengaruhi B dan p(B|A) merupakan conditional probability B yang mempengaruhi kejadian A. Sedangkan p(A) dan p(B) merupakan probabilitas A yang merupakan kategori dan B yang merupakan dokumen berdasarkan persamaan 1.1

$$p(A \mid B) = \frac{p(B \mid A) \times p(A)}{p(B)} \qquad 1.1$$

d. Decision Tree (J48)

Decision trees (J48) merupakan algoritme yang memiliki struktur seperti pohon, dimana setiap internal node-nya menotasikan attribute pengujian dan pada setiap cabangnya merepresentasikan kelas yang dimiliki. Algoritme J48 menurut (Anik, n.d.) memiliki 3 node, yaitu root node, internal node dan leaf node.

Cara kerja pemilihan attribute algorime J48 adalah dengan cara menggunakan nilai gain tertinggi. Sehingga setiap attribute akan dihitung nilain gainnya untuk dipilih menjadi attribute proses pengujian. Nilai ini akan dipilih menjadi nilai parent dari node selanjutnya.. proses perhitungan gain membutuhkan nilai entropy yang dapat dihitung menggunakan rumus 1.2. Entropy merupakan parameter yang digunakan untuk mengukur tingkat keberagaman suatu dataset. Jika dalam suatu dataset memiliki

keberagaman yang tinggi maka nilai entropy yang dihasilkan juga akan semakin besar.

$$\text{Entropy}(S) = \sum_{i=1}^{n} - \text{pi}\log_2\text{pi} \quad 1.2$$

Dimana:

S = Himpunan kasus

n = Jumlah partisi S

Pi = Proporsi Si terhadap S atau jumlah sampel untuk kelas i

Setelah nilai dari entropy didapatkan maka proses selanjutnya adalah menghitung nilai gain untuk mengukur efektifitas dari suatu attribut. Proses menghitung gain dapat menggunakan rumus 1.3.

$$Gain(S, A) = entropy\,(S) - \sum_{i=1}^{n} \frac{|S_i|}{S} * Entropy\,(S_i) \quad 1.3$$

Dimana:

S = Himpunan kasus

A = Fitur

n = Jumlah partisi attribut A

Field Name	Data Type
user	Short Text
password	Short Text

|Si| = Proporsi Si terhadap S

|S| = Jumlah kasus dalam S

IV. HASIL DAN PEMBAHASAN

Pada penelitian ini menggunakan aplikasi machine learning WEKA (Garner, 1995) untuk mengklasifikasikan dokemen halaman web. Proses klasifikasi yang dilakukan menggunakan bantuan algoritme Decision Tree. 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.

Penelitian ini memeliki beberapa skenarion pengujian, diantaranya adalah komparasi metode seleksi fitur antara GA tanpa kombinasi (GA murni) dengan teknik wrapper yang mengkombinasikan GA dengan NaiveBayes. 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 teknik wrapper pada saat proses klasifikasi.

A. Pengujian metode seleksi fitur

Hasil pengujian metode seleksi fitur menggunakan GA dan kombinasi GA dengan NaiveBayes dapat dilihat pada Tabel 2.

Tabel 2.Hasil proses seleksi fitur

Dataset	Jml fitur awal	Seleksi fitur %	
		GA	Wrapper
WebKB	1002	73.85%	45.39%
R8	1011	75.15%	56.71%

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 teknik wrapper yang memanfaatkan algoritme GA dan NaiveBayes mampu menyeleksi fitur sebesar 45.39% dan 56.71%. Menurut presentase yang telah didapatkan, GA mampu menyeleksi fitur lebih banyak dibandingkan dengan teknik wrapper.

B. 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

Metode Seleksi	Dataset WebKB	Dataset r8
Tanpa Seleksi Fitur	78.02%	88.52%
GA	76.64%	87.09%
Wrapper	78.35%	88.67%

Pada Tabel 3, dapat dilihat hasil komparasi proses klasifikasi dokumen halaman web. Klasifikasi dengan data WebKB tanpa seleksi fitur memeliki nilai akurasi 78.02%, GA 76.64%, dan wrapper 78.35%. Nilai akurasi yang didapatkan meningkat, namum terpaut sangat tipis sekali. Sedangkan pada dataset r8, nilai akurasi yang didapatkan tanpa seleksi fitur 88.52%, GA 87.09%, dan wrapper 88.67%. Dalam proses pengujian menggunakan dataset WebKB dan r8, seleksi fitur menggunakan GA mengalami penurunan nilai akurasi. Hal tersebut disebabkan karena jumlah fitur yang didapatkan sangat sedikit sehingga menyebabkan penurunan nilai akurasi. fitur merupakan point penting dalam sebuah dokumen halaman web, jika informasi yang terdapat didalamnya tidak mencakup seluruh isi dokumen, maka hasilnya akan menurun. Sedangkan pada saat penggunaan teknik wrapper nilai akurasi tetap meningkat, meskipun hanya sedikit. Hal ini membuktikan bahwa metode yang diusulkan mampu meningkatkan nilai akurasi.

V. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, dihasilkan beberapa kesimpulan, yaitu:

1) Hasil perolehan nilai akurasi menggunakan teknik kombinasi mengalami peningkatan untuk proses klasifikasi halaman web. Hal ini dipengaruhi Karena fitur yang didapatkan sangat informative. Dengan kata lain penggunaan metode seleksi yang benar dapat membantu meningkatkan hasil akurasi.
2) Hasil proses seleksi fitur menggunakan GA dan NaiveBayes memliki keunggulan daripada seleksi fitur hanyak menggunakan GA saja. Hal ini dibuktikan pada perolehan fitur yang didapatkan dan perolehan nilai akurasi. GA kombinasi dengan NaïveBayes terbukti lebih baik daripada hanya GA saja.
3) Proses seleksi fitur menggunakan teknik kombinasi menyebabkan kebutuhan waktu komputasi menjadi lebih lama, hal ini disebabkan fitur harus diproses oleh dua algoritme sebelum keluar sebagai output. Dari temuan ini, maka penelitian yang akan datang dapat membahas atau meneliti bagaimana teknik mendapatkan fitur yang informative namun tidak membutuhkan waktu komputasi yang lama.

REFERENSI

Anik, A. (n.d.). *Penerapan Algoritma C4.5 pada Program Klasifikasi Mahasiswa Dropout*. Presented at the AMIK BSI Jakarta. AMIK BSI Jakarta.

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. *Proceedings of the New Zealand Computer Science Research Students Conference*, 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. *Computational Intelligence and Design (ISCID), 2014 Seventh International Symposium on*, *1*, 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

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.

Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2014 11th International Computer Conference on, 403–406. https://doi.org/10.1109/ICCWAMTIP.2014.7073436

Platt, J. C. (1998). *Fast Training of Support Vector Machines using Sequential Minimal Optimization*. Retrieved from http://www.research.microsoft.com/~jplatt

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

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

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.

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