WRAPPER FEATURE SELECTION PADA PENGKATEGORISASIAN ARTIKEL BERITA BERBAHASA INDONESIA WRAPPER FEATURE SELECTION ON CATEGORIZATION OF INDONESIAN LANGUAGE NEWS ARTICLE

Debby Damishu[1], Yanuar Firdaus A.w.[2], Zk. Abdurahman Baizal[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Pesatnya penggunaan Internet mengakibatkan pertumbuhan dan pertukaran informasi yang sangat cepat. Informasi yang terdapat pada Internet sangat berlimpah dan beragam, sehingga mengakibatkan jumlah informasi terus meningkat secara eksponensial. Perkembangan informasi ini memungkinkan informasi dapat diakses dengan mudah oleh pengguna. Namun, dengan jumlah informasi yang terus bertambah setiap harinya menimbulkan masalah dan tantangan yang cukup besar. Oleh karena itu, diperlukan suatu pengkategorisasian terhadap artikel berita yang memudahkan pengguna untuk mencari artikel yang diinginkan. Salah satu cara yang dapat mengkategorikan dokumen adalah dengan menggunakan teknik kategorisasi dalam data mining. Akan tetapi jumlah dimensi yang besar membuat performansi classifier kurang baik. Untuk mengatasinya digunakan teknik feature selection. Pada tugas akhir ini, digunakan pendekatan feature selection dengan wrapper feature selection. Sedangkan metoda pencarian subset untuk wrapper adalah hill-cilmbing search dan best first search dengan menggunakan teknik klasifikasi Naive Bayes dari tools WEKA 3.5. Pencarian feature subset dilakukan dengan menghitung nilai macro average F-measure dari setiap node dan akan dihasilkan feature terbaik.

Kata Kunci : feature selection, wrapper feature selection, best first search, hill

Abstract

Internet usage that grow rapidly makes information develop and exchange very fast. Various kind of information are available on Internet, so it makes the number of information rises exponentially. This development makes the information is able to access easily by user. However, the number of information that increase more and more every day make big problem and challenge. Because of that, news article categorization is needed to make article serching more easy for user. One way to categorize the document is categorization technique on data mining. However, high dimensionality makes classifier performance not good. To solve this problem, we use feature selection technique. This final project uses wrapper feature selection approach. While subset search method for wrapper are hill-climbing search and best first search with Naive Bayes classifier from WEKA 3.5. Feature subset searching is done by calculate macro average Fmeasure from each node and will be produced the best feature.

Keywords : feature selection, wrapper feature selection, best first search, hill

BAB I
PENDAHULUAN

1.1 Latar Belakang

Dewasa ini, informasi merupakan hal yang sangat dibutuhkan oleh banyak orang. Internet merupakan salah satu sumber informasi terbesar dan tercepat. Pesatnya penggunaan Internet mengakibatkan pertumbuhan dan pertukaran informasi yang sangat cepat dibandingkan sebelumnya. Begitu juga dengan jumlah informasi berupa artikel berita berbahasa Indonesia berbasiskan web yang terdapat pada Internet juga sangat berlimpah dan beragam, sehingga mengakibatkan jumlahnya meningkat secara eksponensial. Perkembangan informasi ini memungkinkan informasi dapat diakses dengan mudah oleh pengguna. Namun, dengan jumlah informasi yang terus bertambah setiap harinya menimbulkan masalah dan tantangan yang cukup besar, karena dapat menyulitkan pembaca dalam mencari topik berita yang diinginkan.

Oleh karena itu, diperlukan suatu pengkategorisasian terhadap informasi yang berupa artikel berita yang memudahkan pembaca untuk mencari topik berita yang mereka inginkan. Salah satu cara yang dapat mengkategorikannya adalah dengan menggunakan teknik kategorisasi. Sekumpulan dokumen artikel berita tidak hanya memiliki *noise* tapi juga memiliki *feature space* yang berdimensi tinggi. Hal ini dapat menjadi masalah karena dapat menyebabkan rendahnya tingkat keakuratan dalam pengkategorisasian teks. Untuk mengatasinya digunakan teknik *feature selection* yang dapat mengurangi dimensionalitas data yang besar, membuang data yang tidak relevan, dan meningkatkan akurasi hasil.

Feature selection terdapat beberapa pendekatan, salah satunya adalah pendekatan *wrapper feature selection*. Melalui pendekatan ini dapat menghasilkan akurasi yang tinggi. *Wrapper* bergantung kepada clasifier untuk mengevaluasi setiap feature subset. Akan dilakukan dengan menggunakan algoritma induksi *Naive Bayes* sebagai *classifier*. Proses pemilihan feature dilakukan dengan melakukan pencarian terhadap feature subset yang kemudian akan dievaluasi dengan menggunakan pengukuran performansi dari *classifier* itu sendiri.

Pada tugas akhir ini akan dibahas metoda pencarian feature subset untuk proses *feature selection* pada artikel berbahasa Indonesia dengan pendekatan *wrapper*, yaitu dengan metoda *Hill Climbing search* dan *Best First Search*. Pencarian ini memiliki komputasi yang tinggi dengan cost yang banyak. Tugas akhir ini merupakan bagian dari riset text mining.

1.2 Perumusan Masalah

Permasalahan yang akan diselesaikan dalam tugas akhir ini yaitu:

1. Bagaimana implementasi metoda pencarian feature subset dengan menggunakan algoritma pencarian *Hill Climbing search* dan *Best First Search*?
2. Bagaimana menganalisis pencarian feature subset dengan menggunakan algoritma pencarian heuristik yaitu *Hill Climbing search* dan *Best First search*?

Adapun batasan masalah pada tugas akhir ini yaitu:

1. Dataset yang akan digunakan merupakan artikel berita berbahasa indonesia yang diperoleh dari web dan data bersifat *offline* yang disimpan dalam file berekstensi .txt.
2. Klasifikasi yang digunakan dengan menggunakan algoritma induksi *Naive Bayes* yang ada di WEKA.
3. Feature hanya berupa kata bukan berupa frasa.

1.3 Tujuan

Tujuan dari Tugas Akhir ini adalah :

1. Mengimplementasikan metode *feature selection* untuk memilih atribut secara tepat dengan pendekatan *wrapper*.
2. Menganalisis pengaruh algoritma pencarian feature subset menggunakan *Hill Climbing search* dan *Best First Search* terhadap performansi hasil klasifikasi berdasarkan nilai *macro-average F-measure*.

1.4 Metodologi Penelitian

Metodologi yang digunakan untuk menyelesaikan masalah dalam tugas akhir ini adalah :

1. Studi literatur.
   Mencari dan mengumpulkan informasi serta memahami dan mempelajari konsep *wrapper* pada *feature selection* serta konsep *text mining* melalui literatur berupa makalah, buku, atau jurnal yang berhubungan dengan *feature selection, wrapper feature selection, text categorization*.
2. Pencarian dan pengumpulan data.
   Data yang akan digunakan berupa artikel berita berbahasa Indonesia yang diperoleh dari web.
3. Analisis kebutuhan dan implementasi aplikasi yang akan dibangun.
   Analisis kebutuhan dilakukan dengan merancang sistem kebutuhan perangkat lunak. Sedangkan implementasi akan dilakukan terhadap hasil analisis sistem kebutuhan perangkat lunak.

4. Pengujian
   Melakukan pengujian dan analisis terhadap dataset yang diklasifikasikan dengan perangkat lunak yang dibangun serta mengukur performansi dari *classifier*.
5. Pengambilan kesimpulan dan penyusunan laporan tugas akhir.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Dari hasil pengujian dan analisis pada bab sebelumnya dalam tugas akhir ini, maka diperoleh kesimpulan:

1. Feature selection dapat meningkatkan performansi dari klasifikasi dimana feature selection menghasilkan feature yang lebih sedikit dengan hasil macro average F-measure yang sama.
2. Pada wrapper feature selection, jika data yang digunakan mempunyai jumlah feature atau term-nya banyak, maka waktu pencarian akan semakin lama. Hal ini dikarenakan pada wrapper menggunakan metoda pencarian.
3. Berdasarkan hasil klasifikasi menggunakan clasifier Naive Bayes yang terdapat di WEKA, diperoleh hasil feature subset dengan nilai akurasi yang tinggi.

5.2 Saran

1. Untuk metoda pencarian dapat dilakukan dengan metoda pencarian AI yang lain. Misal metoda pencarian *simulated annealing*, dll.
2. Wrapper feature selection dapat dipakai untuk kategorisasi dokumen berita selain bahasa Indonesia dengan mengganti daftar stopword dan stemming sesuai dengan bahasa yang digunakan.

DAFTAR PUSTAKA

1. Adiwijaya Phd, Igg. *Text Mining dan Knowledge Discovery*. Kolokium Bersama Komunitas Datamining Indonesia dan Soft-Computing Indonesia. 2006. [6 Desember 2007]
2. Asian, Jelita, Hugh E.Williams, and S.M.M Tahaghohi. *Stemming Indonesian*. School of Computer Science Technology : Australia. 2005.[11 Mei 2008]
3. Guiterrez, Ricardo and Osuna. *Lecture 11: Introduction to Pattern Analysis*. Texas A&M University. www.research.cs.tamu.edu/prism/lectures/pr/pr_l11.pdf [3 Januari 2008].
4. Guyon, Isabelle. *Lecture 8: Wrappers*. www.clopinet.com/**isabelle**/Projects/ETH/**lecture8**.pdf [20 Juli 2008].
5. Ian H. Written and Eibe Frank. 2005. *Data Mining : Practical Machine Learning Tools and Techniques 2nd edition*. San Francisco : Morgan Kaufmann Publisher.
6. Kohavi, Ron and George H.Jhon. *Wrappers for Feature Subset Selection*. AIJ Special Issue on Relevance. 1997.[1 Januari 2008]
7. Kohavi, Ron and Dan Sommerfield. *Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology*. The First International Conference on Knowledge Discovery and Data Mining. 1995.[1 Januari 2008]
8. Kotsiantis, S.B, D.Kanellopoulos, and P.E Pintelas. *Data Preprosessing for Supervised Learning*. International Journal of Computer Science Volume 1 Number 2 2006 ISSN 1306-4428. pp.111-117. 2006. www.waset.org/pwaset/v12/v12-54.pdf [24 Juni 2008].
9. Kusumadewi, Sri. *Artificial Intelligence: Teknik dan Aplikasinya*. Yogyakarta: Penerbit Graha Ilmu. 2003.
10. Liu, Huan and Lei Yu. *Toward Integrating Feature Selection Algorithms for Classification and Clustering*. Department of Computer Science and Engineering: Arizona State University.[3 Januari 2008]
11. Mannning, Christopher D, Prabhakar Raghavan and Hinrich Schutze. *An Introduction to Text Mining* (online book). Cambridge University Press.[15 Maret 2008]
12. Sebastiani, Fabrizio. *A Tutorial on Automated Text Categorization*. Pisa, Italy: Istituto di Elaborazione dell'Informazione. 1999.[3 Februari 2008]
13. Tala, Fadillah Z. *A Study of Stemmings Effects On Information Retrieval in Bahasa Indonesia*. Institute for Logic, Language and Computation : Universiteit van Amsterdam The Netherlands. [15 Maret 2008].
14. Tan, Steinbach, Kumar. *Slide: Introduction to Data Mining, Chapter 1*.
15. www.en.wikipedia.org/wiki/Naïve_Bayes_classifier [26 Juni 2008]