SYSTEMIC
Vol. 1, No. 2, Desember 2015, 22-26

Seleksi Fitur Dua Tahap Menggunakan Information Gain dan Artificial Bee Colony untuk Kategorisasi Teks Berbasis Support Vector Machine

Khalid[1)], Bagus Setya Rintyarna[2)], Agus Zainal Arifin[3)]

[1)]Prodi Sistem Informasi, Universitas Islam Negeri Sunan Ampel Surabaya
[2)]Jurusan Teknik Elektro, Universitas Muhammadiyah Jember
[3)]Jurusan Teknik Informatika, Institut Teknologi Sepuluh Nopember Surabaya
e-mail: khalid@uinsby.ac.id [1)], bagus.setya@unmuhjember.ac.id [2)],
agusza@cs.its.ac.id [3)]

Abstrak

Salah satu problem yang dihadapi dalam kategorisasi teks adalah dimensi data yang besar yang menyebabkan terjadinya inefisiensi dalam aspek waktu komputasi. Untuk mengatasi hal tersebut, salah satu hal yang bisa dilakukan adalah seleksi fitur pada tahap pre- processing. Pada penelitian ini diusulkan seleksi fitur dua tahap dengan Information Gain dan Artificial Bee Colony. Kategorisasi teks dilakukan dengan Support Vector Machine. Hasil uji coba pada Dataset Reuter21578 menunjukkan adanya peningkatan Precision sebesar rata-rata 15% dan Recall sebesar rata-rata 13% dibandingkan metode pembanding yaitu PSO-SVM.

Kata Kunci : Seleksi Fitur, Information Gain, Artificial Bee Colony, Support Vector Machine, Kategorisasi Teks.

Abstract

One of the problems faced in text categorization is the dimension of data that cause inefficiencies in aspects of computing time. To face this, one of the things that can be done is a feature selection at the stage of pre-processing. this study proposed a two-stage feature selection with Information Gain and Artificial Bee Colony. Text categorization is done by Support Vector Machine. The results on the dataset Precision Reuter21578 showed an increase by an average of 15% and a recall by an average 13% over the comparison method that PSO-SVM.

Keyword : Feature Selection, Information Gain, Artificial Bee Colony, Support Vector Machine, text categorization.

1. PENDAHULUAN

Penelitian dengan topik kategorisasi teks memberikan kontribusi yang besar dalam memberikan solusi atas permasalahan-permasalahan yang muncul dalam sistem temu kembali informasi. Salah satu problem yang sering dihadapi dalam kategorisasi teks seiring dengan perkembangan dunia web adalah besarnya dimensi data yang harus diolah untuk mendapatkan output kategorisasi dengan akurasi yang tinggi [1]. Dimensi data yang besar menyebabkan terjadinya inefisiensi dalam aspek waktu komputasi, menyebabkan aplikasi temu kembali informasi tidak memberikan kepuasan yang optimal bagi user. User menginginkan sistem yang bekerja secara cepat dengan hasil yang memiliki nilai relevansi tinggi. Salah satu jenis penyelesaian yang bisa dilakukan untuk menghadapi permasalahan dimensi data yang besar adalah dengan melakukan seleksi fitur untuk mengurangi dimensi data dengan tujuan untuk mempercepat waktu komputasi.

Dalam kategorisasi teks, seleksi fitur adalah proses memilih subset yang terbaik dari fitur original-nya berdasarkan pada beberapa kriteria [2]. Beberapa penelitian menguji beberapa algoritma dan melakukan pengembangan pada algoritma tersebut untuk melakukan seleksi fitur dengan hasil yang lebih baik. Aghdam menguji dampak penggunaan Ant Colony Optimization untuk melakukan seleksi fitur pada kategorisasi teks. Dokumen teks yang sudah diseleksi fiturnya diklasifikasi dengan *classifier* yang sama kemudian dihitung precision dan recall-nya dan dibandingkan *performance*-nya dengan Information Gain (IG), x^2 test (CHI) dan Genetic Agorithm (GA). Secara rata-rata ACO menunjukkan performance yang lebih baik dibanding ketiga algoritma lain dalam aspek precision dan recall [3].

Dasgupta menguji performance algoritma Sampling for Regularized Least Squares (SRLS) dan membandingkannya dengan Super Structuring (SS), Information Gain (IG), Weighted Sampling (WS), dan Uniform Sampling (US) untuk seleksi fitur dalam kategorisasi teks [4]. Hasilnya menunjukkan keunggulan SRLS atas algoritma yang lain.

Basiri mengusulkan integrasi ACO dan GA pada tahap pre processing kategorisasi dokumen teks. Fitur diseleksi terlebih dahulu dengan menggunakan ACO. Fitur yang sudah diseleksi dengan menggunakan ACO diseleksi kembali dengan menggunakan GA. Berdasar fitur yang sudah diseleksi, dilakukan klasifikasi dan hasilnya dibandingkan dengan algoritma lain yaitu : IG, CHI dan ACO. Penelitian tersebut melaporkan bahwa ACO-GA menghasilkan performance klasifikasi yang lebih baik secara rata-rata dalam aspek precision dan recall dibandingkan ketiga algoritma lainnya [5].

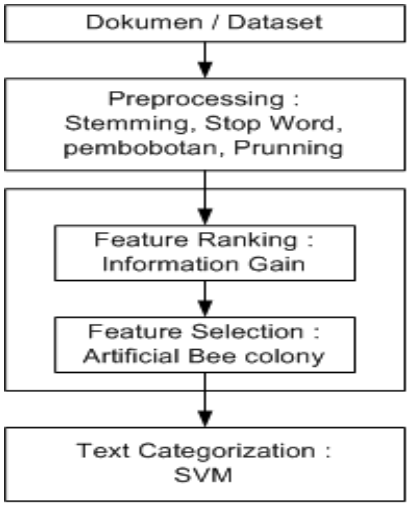
Sedangkan Zahran menggunakan Particle Swarm Optimization (PSO) pada tahap pre processing dalam kategorisasi dokumen teks. Hasil uji coba yang dilakukan menunjukkan PSO lebih baik secara rata-rata dalam aspek precision dan recall dibandingkan dengan CHI, DF dan TFxIDF [6].

Dengan semakin banyaknya dimensi data yang harus diolah dalam aplikasi temu kembali informasi maka diperlukan sebuah metode yang memiliki kinerja komputasi yang cepat sekaligus menghasilkan output temu kembali informasi dengan precision dan recall yang tinggi. Oleh karena itu diperlukan sebuah proses seleksi fitur dengan sebuah metode yang handal. Pada penelitian ini diusulkan seleksi fitur dua tahap dengan Information Gain dan Artificial Bee Colony. Kategorisasi teks dilakukan dengan Support Vector Machine.

2. ARTIFICIAL BEE COLONY DAN SUPPORT VECTOR MACHINE UNTUK KATEGORISASI TEKS

Gambaran tentang metode yang diusulkan untuk kategorisasi dokumen teks dalam penelitian ini bisa dijelaskan dalam Gambar 1.

Tiga tahap penting dalam penelitian ini adalah Pre-Processing, Seleksi Fitur dan Kategorisasi Teks. Tahap Pre-Processing dilakukan untuk mendapatkan term dengan nilai bobot untuk masing-masing term. Term ini pada tahap berikutnya adalah fitur yang akan diseleksi dengan Information Gain dan Artificial Bee Colony. Untuk mengetahui pengaruh seleksi fitur dengan metode yang diusulkan dalam kategorisasi teks maka selanjutnya dilakukan kategorisasi dengan metode Support Vector Machine. Setelah tahap kategorisasi, parameter uji bisa dihitung untuk mengevaluasi kinerja metode yang diusulkan. Pada penelitian ini, parameter uji yang digunakan adalah precision, recall dan fmeasure.



Gambar 1 Blok Diagram Penelitian

2.1 Seleksi Fitur dengan Artificial Bee Colony

Algoritma Artificial Bee Colony (ABC) yang diusulkan oleh Karaboga dan Basturk [7] termasuk cabang Artificial Intelligence yang disebut sebagai Swarm Intelligence [8] untuk menyelesaikan persoalan optimisasi NP-hard. Inspirasi munculnya algoritma Artificial Bee Colony adalah perilaku alami lebah madu dalam proses mendapatkan sumber makanan dengan mengetahui kualitas dan lokasi lebah madu.

Ada tiga jenis artificial bee dalam koloni bee yang digunakan dalam algoritma ABC, yaitu : employed bee, onlooker dan scout. Setengah jumlah koloni terdiri atas employed bee dan setengahnya adalah onlooker. Jumlah employed bee sama dengan jumlah sumber makanan di sekitar sarang karena dalam algoritma ini diasumsikan adanya satu employed bee untuk satu sumber makanan. Employed bee

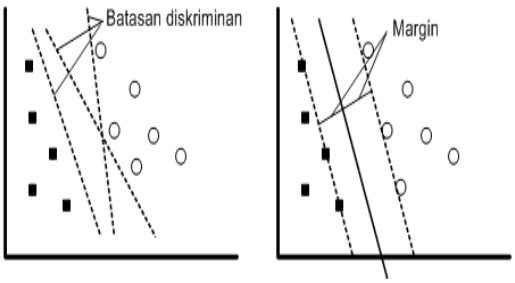
yang sudah meninggalkan sumber makanannya berubah menjadi scout [7].

Solusi yang mungkin dalam persoalan optimisasi direpresentasikan sebagai posisi sumber makanan alam algoritma ABC. Kualitas (fitness) dari solusi berkorespondensi dengan jumlah nectar dari suatu sumber makanan. Jumlah employed bee dan juga onlooker sama dengan jumlah solusi dari populasi. Awalnya, ABC meng-generate secara random populasi awal P(G=0) dari sejumlah solusi SN (posisi sumber makanan), dimana SN menyatakan ukuran populasi. Tiap solusi xi (di mana i = 1,2,3…, SN) adalah sebuah vector dengan D dimensi, di mana D adalah nilai parameter optimisasi. Setelah inisialisasi, populasi dari sumber makanan (solusi) dijadikan sebagai obyek untuk iterasi siklus C = 1,2,3,..., MCN dari proses pencarian yang dilakukan oleh employed bee, onlooker dan scout. Seekor employed bee menghasilkan modifikasi posisi (solusi) dalam memorinya tergantung pada informasi local (informasi visual) dan menguji jumlah nectar (nilai fitness) dari sebuah sumber makanan baru (solusi baru). Bee menyimpan data posisi sumber makanan dalam memorinya, akan tetapi jika jumlah nectar yang baru lebih tinggi dari pada jumlah *nectar* yang lama, maka bee menghapus data nectar yang lama dari dalam memorinya. Setelah semua employed bee menyelesaikan proses searching, mereka berbagi informasi tentang nectar dari sumber makanan. Seekor onlooker mengevaluasi informasi tentang nectar yang didapatkan dari employed bee dan memilih sumber makanan berdasarkan nilai probabilitasnya terhadap nectar.

Pada penelitian ini, Artificial Bee Colony digunakan untuk seleksi fitur. Seleksi fitur adalah proses mengurangi dimensi fitur dengan cara memilih fitur yang penting dan menghilangkan fitur yang *irrelevant*, *redundant* dan *noisy* untuk mendapatkan representasi data yang lebih akurat [6]. Tiga model pendekatan dalam seleksi fitur adalah : 1) Filter Method, 2) Wrapper Method dan 3) Embedded Method [9]. Filter Method mengevaluasi fitur dengan cara membuat rangking fitur secara independen terhadap kelas dalam training set. Sedangkan Wrapper Method menggunakan metode-metode Artificial Intelligence untuk mencari fitur yang terbaik dengan proses *searching* secara *iterative*. Metode Embedded Method menggunakan pendekatan prediksi linear secara simultan untuk meningkatkan goodness-of-fit dan menurunkan jumlah fitur inputnya. Berdasarkan ketiga pendekatan di atas, pendekatan yang digunakan dalam penelitian ini termasuk ke dalam kategori yang ke-2.

2.2 Kategorisasi Teks dengan Support Vector Machine

Konsep dasar SVM sebenarnya merupakan kombinasi harmonis dari teori-teori komputasi yang telah ada sebelumnya seperti margin hyperplane (Duda dan Hart, 1973) dan kernel (Aronsajn, 1950). Konsep SVM secara sederhana dapat dikatakan sebagai usaha untuk menemukan sebuah hyperplane terbaik yang berfungsi sebagai pemisah dua buah kelas pada input space.



Gambar 2. Prinsip kerja SVM

Berbeda dengan dengan strategi neural network yang berusaha mencari hyperplane pemisah antar class, SVM berusaha menemukan hyperplane yang terbaik pada input space. Pada dasarnya, SVM adalah linear classifier. Implementasi kernel trick memungkinkan SVM dapat bekerja pada problem non liear pada ruang kerja berdimensi tinggi. Ilustrasi kerja SVM seperti terlihat pada Gambar 2 bisa sebenarnya merupakan usaha menemukan hyperplane terbaik yang memisahkan dua kelas (pada kasus dalam gambar) pada input space.

Hyperplane terbaik antara kedua kelas dapat ditemukan dengan mengukur margin hyperplane dan mencari titik maksimalnya. Margin adalah jarak antara hyperplane tersebut. Margin adalah jarak antara hyperplane tersebut dengan pattern terdekat dari masing-masing kelas. Pattern yang paling dekat ini disebut support vector.

3. HASIL UJI COBA

Evaluasi kinerja metode yang diusulkan dilakukan dengan menghitung parameter uji berupa precision, recall dan f measure. Secara

Kelas	100 fitur			200 fitur		
	Prec	Rec	Fmeas	Prec	Rec	FMeas
Acq	0.953	0.953	0.953	0.937	0.938	0.938
Crude	0.975	0.976	0.976	0.972	0.973	0.973
Earn	0.980	0.980	0.980	0.975	0.975	0.975
Grain	0.959	0.960	0.959	0.978	0.979	0.979
Money	0.983	0.983	0.983	0.961	0.963	0.961

Table 1 Hasil Uji Coba dengan IG-ABC-SVM

umum, precision dihitung dengan menggunakan formula :

$$precision = \frac{tp}{tp+fi} \quad (1)$$

sedangkan recall dihitung dengan menggunakan rumus :

$$recall = \frac{tp}{tp+fi} \quad (2)$$

dan f measure dihitung dengan rumus :

$$f\ measure = 2x\ \frac{precision\ .recal}{precision\ +recal} \quad (3)$$

Dataset yang digunakan untuk evaluasi adalah Reuters-21578 Dataset yang disediakan oleh Reuters dan CGI untuk penelitian di bidang Temu Kembali Informasi pada Information Retrieval Laboratory of The Computer and Information Science Department di University of Masachusset. Hasil uji coba pada dataset menunjukkan adanya peningkatan Precision sebesar rata-rata 15% dan Recall sebesar rata-rata 13% dibandingkan metode pembanding yaitu PSO-SVM. Hasil uji coba pada beberapa kategori pada Dataset Reuter 21578 ditampilkan dalam Tabel 1 dan Tabel 2.

Kelas	Prec	Rec	FMeas
Acq	0.901	0.921	0.911
Crude	0.829	0.888	0.859
Earn	0.982	0.934	0.958
Grain	0.666	0.750	0.706
Money	0.685	0.689	0.692

Tabel 2 Hasil Uji Coba dengan Metode Pembanding (PSO-SVM)

4. PEMBAHASAN

Hasil uji coba menunjukkan IG-ABC-SVM secara rata-rata memiliki kinerja yang lebih baik dibandingkan metode pembanding (PSO-SVM) dalam aspek Precision, Recall dan F Measure. Dua tahap seleksi fitur dengan IG dan ABC menghasilkan fitur-fitur dengan bobot yang lebih baik yang memiliki pengaruh besar terhadap hasil kategorisasi teks sehingga didapatkan hasil yang lebih baik pula. Selain itu, berdasarkan penelitian yang dilakukan Karaboga didapatkan bahwa ABC menunjukkan hasil yang lebih baik dibandingkan PSO dalam penyelesaian optimasi numerik.

5. KESIMPULAN

Seleksi fitur dua tahap dengan IG dan ABC serta kategorisasi teks dengan SVM menunjukkan hasil yang lebih baik dibandingkan PSO-SVM dalam aspek Precision, Recall dan F Measure.

DAFTAR PUSTAKA

[1] Yan, J.,"OCFS : Optimal Orthogonal Centroid Feature Selection for Text Categorization", 28th Annual International ACM Sigir Conference on Research and Development in Information Retrieval, Page 122-129, 2007

[2] Mesleh, A.M., Kanaan, G.,"Support Vector Machine Text Classification System : Using Ant Colony Optimization Based Feature Subset Selection", Computer Engineering

and System, Issue 25-27, 2008
[3] Aghdam, M.H., Aghae, N.G., Basiri, M.E.,"Text Feature Selection Using Ant Colony Optimization", Expert System with Application, Vol 36, Page 6843-6853, 2009
[4] Dasgupta, A., Drineas, P., Harb, B.,"Feature Selection Method for Text Classification", Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007
[5] Basiri, M.E., Nemati, S.,"A Novel Hybrid ACO-GA Algorithm for Text Feature Selection", Proceedings of the Eleventh conference on Congress on Evolutionary Computation, 2009
[6] Zahran, B.M., Kanaan, G.,"Text Feature Selection Using Particle Swarm Optimization Algorithm", World Applied Sciences Journal 7 (Special Issue of Computer and IT), Page 69-74, 2009
[7] Karaboga, D., Basturk, B.,"Artificial Bee Colony (ABC) Optimization Algorithm for Solving Constrained Optimization Problem, Springer-Verlag, Page 789-798, 2007
[8] Shoukouhifar, M., Sabet, S.,"A Hybrid Approach for Effective Feature Selection Using Neural Network and Artificial Bee Colony Optimization", Proceeding of The 3rd International Conference on Machine Vision, 2010
[9] Forman, G.,"Feature Selection for Text Classification", Information Services and Process Innovation Laboratory, 2007