SUPPORT VECTOR MACHINES YANG DIDUKUNG K-MEANS CLUSTERING DALAM KLASIFIKASI DOKUMEN

Ahmad Yusuf, Tirta Priambadha
Jurusan Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember
Kampus ITS Sukolilo, Surabaya 60111
Email: ahmad.yusuf11@mhs.if.its.ac.id

ABSTRAK

Dokumen dengan jumlah data yang besar dan bervariasi seringkali mempersulit proses klasifikasi. Hal ini dapat diperbaiki dengan mengatasi variasi data untuk menghasilkan akurasi yang lebih baik. Penelitian ini mengusulkan sebuah metode baru untuk kategorisasi dokumen teks berbahasa Inggris dengan terlebih dahulu melakukan pengelompokan menggunakan K-Means Clustering kemudian dokumen diklasifikasikan menggunakan multi-class Support Vector Machines (SVM). Dengan adanya pengelompokan tersebut, variasi data dalam membentuk model klasifikasi akan lebih seragam. Hasil uji coba terhadap judul artikel jurnal ilmiah menunjukkan bahwa metode yang diusulkan mampu meningkatkan akurasi dengan menghasilkan akurasi sebesar 88,1%, presisi sebesar 96,7% dan recall sebesar 94,4% dengan parameter jumlah kelompok sebesar 5.

***Kata Kunci**: Klasifikasi Dokumen, K-Means Clustering, Multi-class SVM.*

1. PENDAHULUAN

Banyaknya dokumen yang ditemui di berbagai media memungkinkan orang untuk mendapatkan segala jenis informasi. Akan tetapi kebanyakan dokumen tidak diklasifikasi atau digolongkan sesuai dengan kelompoknya sehingga dokumen-dokumen yang berhubungan sulit ditemukan. Untuk itu perlu dilakukan kategorisasi dokumen agar dokumen yang bertopik sama bisa ditemukan dengan mudah.

Variasi data terkadang dapat menyulitkan proses klasifikasi sehingga dapat dikelompokkan terlebih dahulu [1]. Hal yang sama dapat ditemukan dalam klasifikasi dokumen. Dokumen yang umumnya memiliki data dengan jumlah besar dan bervariasi dapat menyulitkan dalam membuat model klasifikasi. Oleh karena itu dokumen-dokumen tersebut dapat dikelompokkan menurut kemiripan satu sama lain agar dapat dengan mudah diklasifikasi, dicari dan ditemukan sesuai dengan permintaan yang ada.

Dalam proses klasifikasi dokumen, seringkali ditemukan hasil yang kurang baik dikarenakan jumlah data dokumen yang besar dan bervariasi. Maka dari itu perlu ada proses penggugusan sebelum dilakukan klasifikasi pada data dokumen. Penambahan proses penggugusan sebelum proses klasifikasi dalam pengelompokan dokumen diharapkan mampu memperbaiki hasil pengelompokan sesuai dengan kelas atau kategori spesifiknya.

Proses pengelompokan dokumen untuk memperbaiki hasil klasifikasi dokumen dapat dilakukan dengan proses penggugusan menggunakan metode *K-Means Clustering* [2]. Dari hasil penggugusan maka dilanjutkan dengan proses klasifikasi menggunakan *Multi-class Support Vector Machine* (SVM) [3].

2. METODE

Secara umum penelitian ini terdiri atas tiga tahapan. Tahap pertama ialah persiapan atau pra-proses. Tahap pra-proses mempersiapkan teks yang ada didalam dokumen agar siap untuk digunakan untuk proses selanjutya.

Tahap yang kedua melakukan proses pengelompokan atau kategorisasi dokumen. Proses pengelompokan dilakukan terhadap hasil pra-proses yang merupakan representasi data dalam bentuk model ruang vektor. Metode pertama ialah pengelompokan dokumen yang ada dengan *K-Means Clustering*. Kemudian setiap kelompok dokumen tersebut akan diklasifikasi dengan *Multi-Class SVM*.

2.1 Praproses

Tahap awal sebelum melakukan proses pengelompokan dokumen adalah mempersiapkan teks yang ada didalam dokumen. Pada tahap pra-proses ini dilakukan beberapa subproses agar dokumen dapat dipakai untuk melakukan proses pengelompokan.

Subproses yang pertama ialah *tokenizer*, yakni proses yang bertujuan untuk memisah teks menjadi beberapa *token* berdasarkan pembatas berupa spasi atau tanda baca. Proses selanjutnya adalah menghilangkan teks yang bersesuaian dengan teks

yang terdapat pada daftar *stopword*, karena teks tersebut dianggap tidak dapat mewakili konten dokumen.

Kemudian pada teks yang masih tersisa dilakukan proses *stemming*, yaitu proses pengubahan teks menjadi bentuk dasarnya. Selanjutnya, setiap kata tersebut disebut sebagai *term*.

Nantinya setiap *term* akan didaftar dan diberi bobot. Pembobotan masing-masing term dilakukan dengan metode TF-IDF (*Term Frequency – Inverse Document Frequency*). TF-IDF merupakan metode pembobotan *term* dengan menggunakan *term-frequency* (jumlah *term* yang terdapat pada tiap dokumen) serta *inverse document frequency* (invers jumlah dokumen yang memuat suatu *term*). Pembobotan TF*IDF dirumuskan dengan :

$$w_{i,j} = tf_{i,j}\ x \log(\frac{N}{df_i}) \qquad (1)$$

dengan $w_{i,j}$ merupakan bobot term ke-*i* pada dokumen ke-*j*, $tf_{i,j}$ merupakan jumlah term ke-*i* pada dokumen ke-*j*, df_i merupakan jumlah dokumen yang mengandung term ke-*i* dan *N* adalah jumlah dokumen keseluruhan.

2.2 Pengelompokan dengan *K-Means Clustering*

Salah satu metode dalam pengelompokan dokumen adalah *K-Means Clustering*. *K-Means Clustering* merupakan metode pengelompokan paling sederhana yang mengelompokkan data kedalam *k* kelompok berdasar pada *centroid* masing-masing kelompok [4]. Hanya saja hasil dari *K-Means* sangat dipengaruhi parameter *k* dan inisialisasi *centroid*. Pada umumnya *K-Means* menginisialisasi *centroid* secara acak. Namun metode yang diusulkan akan memodifikasi *K-Means* dalam inisialisasi *centroid* khususnya dalam memperbaiki performa dalam pengelompokan dokumen.

Berbeda dengan algoritma *K-Means* yang ada umumnya, *K-Means Clustering* dalam penelitian ini melakukan inisialisasi *centroid* dengan menggunakan pengukuran *Jaccard Distance* dengan rumus [2]

$$J_\delta(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \qquad (2)$$

Pengukuran ini dilakukan untuk menemukan perbedaan antara dua vektordokumen pada ruang *n* dimensi. Pertama-tama kami mengukur perbedaan matriks antara semua pasangan dokumen, dimana dokumen ini direpresentasikan sebagai vector di ruang *n* dimensi dengan menggunakan rumus

$$r'(d_1,d_2) = (\frac{|T(d_1) \cap T(d_2)|}{|T(d_1) \cup T(d_2)|}) \qquad (3)$$

dengan *T(d)* merupakan fitur yang muncul dari dokumen d dan $r'(d_1,d_2) = 1$ serta $r'(d_1,d_2) = r'(d_2,d_1)$.

Dengan menggunakan komplementer dari matriks *Jaccard similarity* $J' = 1 - r'$, maka akan didapatkan matriks yang mengindikasikan perbedaan diantara dokumen. Dari matriks perbedaan inilah nantinya akan dipilih *k* yang merupakan dokumen yang paling berbeda dan menginisialisasinya sebagai *k* pusat gugus. Langkah selanjutnya ialah untuk menempatkan dokumen agar masuk ke dalam kelompok yang benar. Kami menggunakan *Cosine similarity* untuk mengukur kesamaan diantara dokumen dengan *centroid* kelompok dengan rumus

$$similaritas = \cos(\theta) = (\frac{|A.B|}{\|A\|\|B\|}) \qquad (4)$$

lalu menempatkan setiap dokumen tersebut ke dalam kelompok yang paling mirip. Apabila semua dokumen telah didistribusikan ke semua kelompok dan sudah tidak ada lagi penambahan pada *centroid cluster*, maka iterasi dari penggugusan akan berakhir. Secara lengkap skema umum sistem ditunjukkan pada Gambar 1.

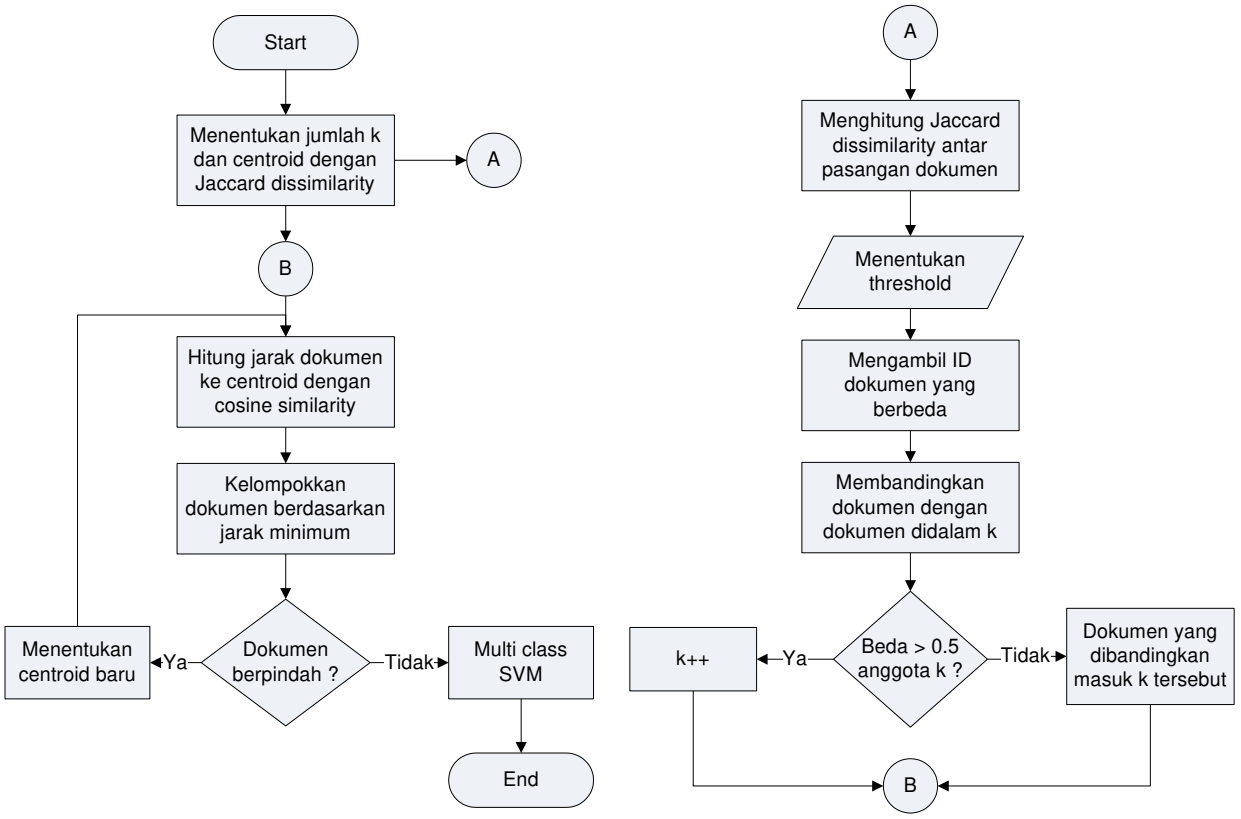
2.3 *Multi-Class SVM*

Support Vector Machine (SVM) adalah metode klasifikasi yang bekerja dengan cara mencari *hyperplane* dengan margin terbesar [5]. *Hyperplane* adalah garis batas pemisah data antar-kelas. Margin adalah jarak antara *hyperplane* dengan data terdekat pada masing-masing kelas. Adapun data terdekat dengan *hyperplane* pada masing-masing kelas inilah yang disebut *support vector*.

Pada dasarnya, SVM merupakan metode yang digunakan untuk klasifikasi dua kelas (*binary classification*). Pada perkembangannya, beberapa metode diusulkan agar SVM bisa digunakan untuk klasifikasi *multi-class* dengan cara mengombinasikan beberapa *binary classifier* [3]. Metode yang pernah diusulkan adalah metode *One-against-one*.

Adapun untuk metode *One-against-one*, akan dikonstuksi sejumlah *k(k-1)/2* model klasifikasi SVM dengan masing-masing model dilatih menggunakan data dari dua kelas yang berbeda. Dengan demikian, untuk data pada kelas *i* dan *j*, SVM menyelesaikan permasalahan klasifikasi biner untuk

$$\min_{w_{ij},b_{ij},\varepsilon_{ij}} \frac{1}{2} {w_{ij}}^T w_{ij} + C \sum_1 \varepsilon_{ij}^1$$
$$w_{ij}^T \emptyset(x_1) + b_{ij} \geq 1 - \varepsilon_{ij}^1 \ , if\ y_1 = i$$
$$w_{ij}^T \emptyset(x_1) + b_{ij} \leq -1 + \varepsilon_{ij}^1 \ , if\ y_1 = j$$
$$\varepsilon_{ij}^1 \geq 0 \qquad (5)$$



Gambar 1. Diagram Alir Metode

Decision function untuk fungsi di atas diambil melalui *voting*, jika hasil dari $sign\ (w^{ij})^T \Phi(x) + b^{ij}$ menyatakan bahwa data *x* berada di kelas *i*, maka nilai *vote* untuk kelas *i* ditambah satu. Selanjutnya, prediksi kelas dari data *x* adalah kelas dengan nilai *vote* tertinggi. Jika sebaliknya, nilai vote untuk kelas *j* ditambah satu [3].

Secara umum, tahapan klasifikasi dokumen dengan *multi-class SVM* pada penelitian ini dapat dijabarkan seperti tahap-tahap berikut ini. Pertama-tama dari hasil pra-proses dan penggugusan, maka dihasilkan *k*-kelompok data term dalam representasi model ruang vektor. Kemudian masing-masing kelompok dilatih menggunakan *Multi-class SVM* dengan metode *One-against-one*. Dengan demikian akan didapatkan model klasifikasi dokumen pada masing-masing kelompok.

3. SKENARIO UJI COBA

Dari *k* model klasifikasi yang telah ada, maka dapat dilakukan klasifikasi dokumen baru. Pengujian dilakukan dengan mengelompokkan dokumen baru kedalam kelompok yang ada menggunakan tetangga terdekat dari *centroid* pada masing-masing kelompok. Setelah didapatkan kelompok yang sesuai maka dilakukan proses klasifikasi dokumen baru dengan model *Multi-class SVM* pada kelompok yang bersangkutan.

Dataset dokumen yang digunakan dalam penelitian ini merupakan judul-judul artikel yang diambil dari beberapa jurnal ilmiah. Semua judul artikel yang dipakai dalam penelitian ini menggunakan bahasa Inggris. Artikel-artikel ini sebelumnya telah kami tentukan kategorinya terlebih dahulu, yaitu terbagi kedalam 6 kategori. Artikel-artikel tersebut kemudian dipakai sebagai data latih dan uji dengan menggunakan *cross validation*.

Uji coba dengan *cross validation fold* 5 seperti yang terdapat pada Tabel 1, menggunakan data latih sebanyak 240 artikel sedangkan untuk data uji memakai 300 data artikel.

4. HASIL PENGUJIAN

Tabel 1 menunjukkan hasil pengujian dengan *cross validation* dengan *fold* sama dengan 5, yang mempunyai arti bahwa setiap satu parameter kelompok diacak sebanyak 5 kali kemudian hasilnya dirata-rata.

Hasil uji coba menunjukkan semakin besar parameter jumlah kelompok maka akurasi, presisi, dan *recall* akan semakin baik sampai pada parameter tertentu yaitu titik optimal. Jika nilai parameter terus dinaikkan dari titik optimal maka nilai akurasi, presisi dan *recall* akan tetap atau menurun. Tabel 1 menunjukkan hasil terbaik pada jumlah kelompok

Tabel 1: Hasil nilai rata-rata uji coba dengan *cross validation* dengan *fold* = 5

Jumlah *cluster*	1	2	3	4	5	6	7	8	9	10
Akurasi (%)	87,6	87,9	88	88	88,1	87,6	87,6	87,6	87,6	87,6
Presisi (%)	96,7	96,7	96,7	96,7	96,7	96,4	96,4	96,3	96,3	96,3
Recall (%)	94	94	94	94,4	94,4	94,4	94,4	94	94	94

sebanyak 5, dengan akurasi sebesar 88,1%, presisi sebesar 96,7% dan *recall* sebesar 94,4%.

Nilai parameter jumlah kelompok terbaik didapatkan berdasar pada distribusi data pada dataset yang digunakan. Pada penelitian ini nilai parameter terbaik diperoleh berdasarkan pada hasil uji coba yang ditunjukkan pada tabel 1. Hal ini menunjukkan dataset yang digunakan memiliki distribusi data yang lebih cocok dibagi menjadi 5 kelompok dengan *K-Means Clustering* dalam pemodelan *Multi-class SVM* pada masing-masing kelompoknya.

5. KESIMPULAN

Dokumen dalam jumlah besar dan bervariasi terkadang menyulitkan dalam membuat model klasifikasi. Dalam penelitian ini dilakukan klasifikasi dokumen dengan menggunakan metode *Multi-class SVM* yang didukung dengan *K-Means Clustering* untuk pengelompokan data dokumen. Dengan adanya pengelompokan dokumen maka data yang akan dilatih dengan *Multi-class SVM* akan lebih seragam sehingga menghasilkan hasil yang lebih baik.

Berdasarkan hasil uji coba dan analisis yang telah dilakukan dalam penelitian ini, maka dapat diambil kesimpulan sebagai berikut:

1. Pengelompokan dokumen dengan *K-Means Clustering* sebelum melakukan klasifikasi mampu meningkatkan akurasi sebesar 0,5 % dan *recall* sebesar 0,4% pada data artikel yang digunakan.
2. Klasifikasi akan mencapai akurasi terbaik pada parameter jumlah kelompok tertentu. Hal ini dipengaruhi dengan variasi data artikel yang digunakan. Jika menggunakan uji coba dengan data berbeda besar kemungkinan parameter jumlah kelompok dengan akurasi terbaik berbeda pula.

Hasil ini kedepannya dapat digunakan sebagai landasan untuk pengembangan atau penelitian selanjutnya.

6. DAFTAR PUSTAKA

[1] S. Trivedi, A. Pardos and N. Sar. “Spectral Clustering in Educational Data Mining”. Department of Computer Science, Worcester Polytechnic Institute (2008).

[2] M. Shameem and R. Ferdous. “An efficient K-Means Algorithm integrated with Jaccard Distance Measure for Document Clustering”. IEEE (2009).

[3] J.Z. Liang. “SVM Multi-Classifier And Web Document Classification”, Proceedings of the IEEE Third International Conference on Machine Learning and Cybernetics (2004).

[4] M.E. Celebi, H.A. Kingravi and P.A. Vela. ‘”A comparative study of efficient initialization methods for the k-means clustering algorithm”. Elsevier Expert Systems with Applications (2012).

[5] J. Yunliang, F. Jing, S. Qing and Z. Xiongtao. “The Classification for E-Government Document Based on SVM”. 2010 International Conference on Web Information Systems and Mining (2010).