

PERBANDINGAN KINERJA METODE PRA-PEMROSESAN DALAM PENGKLASIFIKASIAN OTOMATIS DOKUMEN PATEN

Budi Nugroho¹⁾, Asep Denih²⁾

¹⁾Pusat Penelitian Informatika, LIPI,

Jl. Sangkuriang/Cisitu No.21/154D Komplek LIPI Gedung 20 Lantai 3, Bandung

²⁾Program Studi Ilmu Komputer, FMIPA, Universitas Pakuan, Jl. Pakuan, P.O. Box 452, Bogor

Corresponding Author: budi.nugroho@lipi.go.id

Abstrak

Makalah ini menyajikan analisis kinerja dan perbandingan beberapa metode pra-pemrosesan yang digunakan dalam klasifikasi paten otomatis dengan kernel graf untuk Support Vector Machine (SVM). Metode pra-pemrosesan didasarkan pada teknik transformasi data, yaitu penskalaan data, pemusatan data, standarisasi data, normalisasi data, transformasi Box-Cox dan transformasi Yeo-Johnson. Pengklasifikasi paten otomatis dirancang untuk mengklasifikasikan input graf sitasi paten ke dalam satu dari 10 kelas International Patent Classification (IPC) yang memungkinkan. Input diambil dengan berbagai kondisi latar belakang. Percobaan menunjukkan bahwa hasil terbaik dicapai ketika metode pra-pemrosesan adalah normalisasi data, mencapai akurasi klasifikasi hingga 85,33,15 % untuk kernel KEHL dan 93,80 untuk KVHL. Sebaliknya, untuk KEHG, aplikasi metode preprocessing menurunkan akurasi.

Kata kunci: pra-pemrosesan data, kernel graf, support vector machine, pengklasifikasin otomatis dokumen paten

Abstract

This paper presents a performance analysis and comparison of several pre-processing methods used in automatic patent classification with graph kernels for Support Vector Machine (SVM). The pre-processing methods are based on the data transform techniques, namely data scaling, data centering, data standardization, data normalization, the Box-Cox transform and the Yeo-Johnson transform. The automatic patent classification is designed to classify an input of patent citation graphs into one of 10 possible classes of the International Patent Classification (IPC). The input is taken with various background conditions. The experiments showed that the best result is achieved when the pre-processing method is data normalization, achieving a classification accuracy of up to 85.33.15% for the KEHL and 93.80% for the KVHL. In contrast, for the KEHG, the preprocessing method application decreased the accuracy.

Keywords: pre-processing, graph kernels, support vector machine, automatic patent classification

1. Pendahuluan

Ketika menerapkan algoritma *machine learning* pada suatu dataset, agar mendapatkan hasil yang terbaik, dapat dilakukan dengan berbagai cara, diantaranya adalah penyiapan data. Penyiapan data ini dikenal sebagai tahap pra-pemrosesan. Sebagian algoritma *machine learning* memerlukan pra-pemrosesan terhadap dataset, tetapi sebagian lainnya tidak. Dataset yang tidak memenuhi format atau struktur yang ideal seringkali memerlukan tahap pra-pemrosesan, khususnya bila menggunakan algoritma *Support Vector Machine* (SVM). Dalam proses pengklasifikasian dataset dokumen paten berdasarkan graf sitasi, tahap pra-pemrosesan merupakan tahapan penting yang sebaiknya dilakukan agar hasil pengklasifikasian memberikan akurasi yang terbaik.

Beberapa penelitian terdahulu mengenai metode pra-pemrosesan telah dilakukan dalam beberapa aspek. [1] membahas klasifikasi tersupervisi untuk dataset tak seimbang (*imbalance*). Masalah yang hadapi dalam dataset tak seimbang bila dalam klasifikasi biner, salah satu kelas jumlahnya jauh lebih banyak dari kelas lainnya. Penelitian [1] ini menggunakan pendekatan pra-



pemrosesan data untuk mengurangi dampak tumpang tindih kelas dan pergeseran dataset dalam klasifikasi tak seimbang [1]. Sedangkan [2] mengusulkan metode pra-pemrosesan data untuk meningkatkan kinerja analisis komponen utama (*Principal Component Analysis /PCA*) untuk menyelesaikan masalah pengklasifikasian. Metode yang dilakukan adalah dengan pendekatan dua tahap, yaitu menghitung bobot tiap fitur dari dataset lalu menentukan ambang batas fitur yang akan digunakan berdasarkan hasil perhitungan bobot tersebut [2]. Di pihak lain, [3] melakukan studi untuk meningkatkan efisiensi prosedur training data menggunakan *Support Vector Clustering* (SVC). Metode pra-pemrosesan yang digunakan yaitu *Heuristic for Redundant-point Elimination* (HRE) dan *Shared Nearest Neighbor* (SNN) [3].

Kemungkinan untuk meningkatkan kinerja pengklasifikasi berbasis SVM dilakukan pula oleh Adamiak dkk (2016) dengan memperkenalkan langkah pengurangan dimensionalitas untuk input data. Ekstraksi fitur dilakukan dengan dua metode kernel yang berbeda, yaitu kernel *Principal Component Analysis* (kPCA) dan kPCA tersupervisi. Pendekatan pra-pemrosesan ini mampu menyediakan variabilitas yang rendah dalam kinerja pengklasifikasi [4]. Adapun penelitian yang menggunakan graf sebagai dataset dilakukan pula oleh [5] dengan pendekatan proposisionalisasi graf yang digunakan untuk mentransformasi data relasi dan terstruktur menjadi vektor fitur. Metode ini diuji dengan tiga algoritma pengklasifikasi yaitu *random forest*, SVM dan *nearest neighbor*. Dalam aspek lainnya, Farquad (2012) menggunakan SVM sebagai metode untuk pra-pemrosesan data tak seimbang. Setelah mendapatkan data hasil training di SVM, data ini dimasukkan ke sejumlah teknik uji komputasi cerdas, diantaranya *Multilayer Perceptron* (MLP), *Logistic Regression* (LR), dan *Random Forest*. Berdasarkan hasil pengukuran sensitivitas, metode SVM mampu menghasilkan data yang lebih seimbang dan jumlah contoh yang efektif untuk kelas minoritas [5].

Untuk mengetahui seberapa efektif tahap pra-pemrosesan terhadap hasil pengklasifikasian, Almuhaideb dan Menaii (2016) menerapkan metode ini dalam penelitian data medis. Tujuan dari pra-pemrosesan ini adalah untuk menentukan karakteristik data medis yang mencakup *noise*, data tak lengkap, dan fitur tak relevan. Metode pra-pemrosesan dilakukan dengan operasi diskretisasi atribut numerik, pemilihan atribut subset dan penanganan nilai hilang [6]. Termasuk dalam hal pra-pemrosesan data, [7] mengusulkan perbaikan algoritma kNN dengan menggunakan matrik korelasi untuk merekonstruksi titik-titik data tes untuk mendapatkan nilai k yang paling tepat mengacu kepada matrik korelasi kNN [7].

Penerapan algoritma pengklasifikasi dapat dilakukan terhadap berbagai macam dataset diantaranya dataset dokumen paten. Paten merupakan salah satu hasil penelitian dan pengembangan dilakukan oleh banyak lembaga, termasuk lembaga litbang, perguruan tinggi dan industri, yang mengandung unsur kebaruan dalam teknologi yang perlu dilindungi. Untuk mendaftarkan sebuah paten, pendaftar harus memasukkan aplikasi dokumen patennya ke kantor paten dan menentukan paten tersebut termasuk ke dalam salah satu kategori yang ada. Proses pengklasifikasian ini biasanya dilakukan oleh pendaftar dan pemeriksa secara manual. Dengan banyaknya dokumen aplikasi paten yang didaftarkan, proses menentukan klasifikasi ini dapat dilakukan secara otomatis dengan menerapkan algoritma machine learning. Dalam penelitian ini digunakan graf sitasi paten sebagai basis untuk melakukan klasifikasi dan memilih SVM sebagai algoritma pengklasifikasi. Metode pra-pemrosesan diperlukan untuk mendapatkan hasil akurasi yang terbaik berdasarkan perhitungan matriks kernel dari graf sitasi paten.

Data graf terstruktur saat ini semakin melimpah keberadaannya. Untuk menganalisis dan memahami data ini, dibutuhkan metode analisis data dan *machine learning* yang mampu menangani dataset graf berskala besar. Sebagai contoh, persoalan *learning* atas graf meningkat dalam klasifikasi dokumen paten. Menurut [10], sitasi paten mengindikasikan keterhubungan pengetahuan dan sangat penting untuk penggunaan yang luas dari metode analisis sitasi paten [8]. Dalam persoalan ini, ketika didapati graf sitasi paten yang besar, yang direpresentasikan oleh simpul (*node*) dan sisi (*edge*) berlabel, yang mempunyai kode klasifikasi tertentu. Tugas selanjutnya adalah bagaimana secara akurat memprediksi apakah satu paten baru akan diklasifikasikan ke dalam kode klasifikasi ini atau tidak. Sebuah asumsi umum yang dibuat untuk mengatasi persoalan ini adalah bahwa graf sitasi dengan struktur yang mirip akan mempunyai kode klasifikasi yang mirip pula. Persoalan mengukur kemiripan graf karenanya menjadi inti dari *learning* dalam riset graf.

Metode kernel menawarkan penyelesaian terhadap kasus-kasus yang mana suatu sampel tidak dapat divektorkan. Trik untuk mendefinisikan fungsi kemiripan objek-objek yang sesuai dikenal dengan fungsi kernel. Lalu matriks pasangan kemiripan dapat dilewatkan melalui

algoritma berbasis kernel tersupervisi seperti SVM untuk melakukan klasifikasi. Dengan kernel yang sesuai, pendekatan dua langkah ini memberikan hasil yang banyak dipakai atas macam-macam dataset dan menjadi baku serta banyak dilakukan oleh para peneliti. Salah satu keterbatasan utama pendekatan kernel graf dan SVM adalah, representasi dan learning merupakan dua langkah yang independen. Dengan kata lain, fitur dihitung awal secara terpisah dari tahap *training* dan tidak dioptimasi untuk tahap selanjutnya. Yang menyebabkan bisa jadi hasil akhir klasifikasi bersifat bias. Dalam penelitian ini, dilakukan pra-pemrosesan terhadap nilai matriks kernel dari hasil perhitungan.

Menentukan metode pra-pemrosesan yang benar-benar tepat cukup sulit dilakukan. Bila atribut input dataset mempunyai skala yang sama, maka metode pra-pemrosesan yang efektif digunakan adalah metode berdasarkan sampel. Sedangkan metode regresi akan lebih tepat diterapkan untuk atribut input yang terstandar. Eksperimen dengan menggunakan macam-macam transformasi data dan algoritma *machine learning* yang berbeda untuk mendapatkan hasil terbaik pada dataset yang ada dapat dilakukan.

Dalam makalah ini, diuraikan penerapan transformasi data untuk menyiapkan data yang digunakan sebagai input bagi pengklasifikasi SVM dalam bentuk kernel matrik untuk pengklasifikasian otomatis dokumen paten. Kontribusi utama yang disampaikan dalam makalah ini mencakup penyajian beberapa metode pra-pemrosesan data numerik dan perbandingan hasil klasifikasi dokumen paten yang dapat dipakai sebagai salah satu acuan untuk menentukan metode pra-pemrosesan yang tepat dalam penerapan algoritma SVM untuk pengklasifikasian otomatis.

Lanjutan makalah ini disusun berdasarkan urutan sebagai berikut. Dalam Bagian 2 Metodologi, dipaparkan tahapan penelitian dan metode yang diterapkan. Pada Bagian 3 Hasil dan Pembahasan, disajikan hasil eksperimen dan pembahasannya yang disertai dengan elaborasi dari hasil-hasil riset yang telah dilakukan. Akhirnya dirangkum dan disimpulkan hasil penelitian ini dalam Bagian 4 Kesimpulan.

2. Metode Penelitian

Eksperimen untuk membandingkan beberapa metode pra-pemrosesan data dengan penerapan fungsi kernel dalam mengklasifikasi dokumen paten dilakukan melalui eksploitasi graf sitasi paten. Penggunaan kernel graf dilandasi oleh penelitian Sugiyama dan Borgwardt (2015) yang menjelaskan tentang macam-macam kernel graf [9] serta penelitian Zhu dan Gao (2016) tentang pengaruh pra-pemrosesan data terhadap kinerja algoritma pengklasifikasi [10]. Diterapkan enam metode pra-pemrosesan standar didasarkan pada teknik transformasi data, yaitu penskalaan data, pemusatan data, standardisasi data, normalisasi data, transformasi Box-Cox dan transformasi Yeo-Johnson. Transformasi data diterapkan pada matriks kernel yang dihitung dengan tiga macam kernel graf, yaitu: kernel Gaussian RBF antara histogram label sisi (KEHG), kernel linier antara histogram label sisi (KEHL), dan kernel linier antara histogram label simpul (KVHL). Pengklasifikasi SVM dilatih menggunakan matriks kernel data dalam dataset training. Dalam penelitian ini, pengklasifikasi yang digunakan adalah SVM karena telah terbukti kinerjanya dalam beberapa penelitian sebelumnya [9] [4] [11]. Sesuai dengan tujuan penelitian untuk mengklasifikasi dokumen paten, ukuran kemiripan merupakan sebuah fungsi yang berhubungan dengan nilai numerik dari pasangan graf sitasi dengan konsep bahwa nilai yang lebih tinggi menunjukkan kemiripan yang lebih dengan antara graf. Terdapat hubungan positif antara matriks kernel dengan matriks kemiripan berdasarkan jarak (*distance based*) [12]. Seluruh proses perhitungan dalam penelitian ini menggunakan Bahasa pemrograman R dengan *library caret* [13] dan *kernlab* [14].

Dataset yang digunakan dalam penelitian ini adalah koleksi dokumen paten dalam bentuk daftar sitasi yang diperoleh dari Stanford Large Network Dataset Collection (URL <https://snap.stanford.edu/data/>). Dataset paten dari Kantor Patent Amerika Serikat dikelola oleh National Bureau of Economic Research (<http://www.nber.org/>). Dataset ini menghimpun paten sepanjang 37 tahun (1 Januari 1963 s.d 30 Desember 1999) dan mencakup semua paten utilitas terdaftar sebanyak 3.923.922 buah. Graf sitasi memasukkan keseluruhan sitasi yang dibuat oleh patent terdaftar dalam rentang waktu tahun 1975 s.d 1999. Tabel 1 berikut ini menyajikan informasi statistik dari dataset graf sitasi paten yang digunakan.

Tabel 1. Statistik Dataset Graf Sitasi Paten

Statistik	Nilai
Nodes	3.774.768
Egdes	16.518.948
Nodes in largest WCC	3.764.117 (0,997)
Edges in largest WCC	16.511.741 (1,000)
Nodes in largest SCC	1 (0.000)
Edges in largest SCC	0 (0.000)
Average clustering coefficient	0,0757
Number of triangles	7.515.023
Fraction of closed triangles	0,02343
Diameter (longest shortest path)	22
90-percentile effective diameter	9,4

Dalam penelitian ini, dilaporkan akurasi sebagai ukuran kinerja pengklasifikasi. Proses klasifikasi dalam penelitian ini merupakan klasifikasi multi kelas. Setelah mendapatkan hasil klasifikasi dalam bentuk matriks hasil prediksi dan hasil sebenarnya, maka akurasi dihitung dengan rumus sebagai berikut:

$$Akurasi = \frac{\sum_{i \in C} D_i}{N}$$

dimana C = satu set kelas, N = jumlah keseluruhan sampel, dan D_i = jumlah sampel yang terklasifikasi dengan benar pada masing-masing kelas i .

Metode pra-pemrosesan data matrik kernel menggunakan fungsi `preProcess()` dari *library caret*. Transformasi pra-pemrosesan dapat dilakukan terhadap data training dari dataset manapun dengan variabel yang sama. Berikut ini adalah penjelasan dari masing-masing metode pra-pemrosesan data yang digunakan.

- Penskalaan data, metode ini menghitung simpangan baku dari setiap atribut dan membagi masing-masing nilai dengan simpangan baku tersebut.
- Pemusatan data, metode ini menghitung nilai rata-rata dari setiap atribut dan mengurangnya dari masing-masing nilai.
- Standardisasi data, metode ini menggabungkan penskalaan dan pemusatan data. Nilai masing-masing atribut akan mempunyai rata-rata antara 0 dan 1.
- Normalisasi data, metode ini melakukan penskalaan nilai atribut menjadi rentang antara 0 dan 1.
- Transformasi Box-Cox, metode ini melakukan operasi agar nilai atribut mendekati distribusi Gaussian. Metode ini berlaku jika semua atribut bernilai positif.
- Transformasi Yeo-Johnson, metode ini mirip dengan transformasi Box-Cox, yaitu mengubah data agar mendekati distribusi Gaussian, hanya nilai atribut bias sama dengan 0 atau nilai negative.

3. Hasil dan Pembahasan

Graf sitasi paten menggambarkan keterkaitan antar dokumen paten. Melalui tahapan membandingkan pola sitasi yang tercermin dalam graf sitasi, dikelompokkan dokumen paten tersebut sesuai dengan skor kemiripannya. Skor kemiripan dalam graf sitasi paten dapat dihitung dan dilambangkan dalam nilai kernel matriknya. Dalam penelitian ini, setelah melakukan tahap perhitungan nilai matriks kernel berdasarkan tiga buah kernel yang telah dipilih, lalu diterapkan beberapa metode pra-pemrosesan. Tabel 2 di bawah ini menyajikan kinerja rata-rata dari pengklasifikasi SVM dengan kernel graf yang berbeda dengan metode pra-pemrosesan yang diterapkan.

Tabel 2. Rata-rata Kinerja Pengklasifikasi SVM pada Metode Pra-pemrosesan yang Berbeda

Metode Pra-pemrosesan	Kernel Graf		
	KEHG	KEHL	KVHL
Original	-	-	-
Penskalaan data	B	B	A
Pemusatan data	B	B	B



Standardisasi data	B	B	A
Normalisasi data	B	A	A
Transformasi Box-Cox	B	B	B
Transformasi Yeo-Johnson	B	B	B

A = Kinerja di atas rata-rata original, B = Kinerja di bawah rata-rata original

Berdasarkan hasil di atas, ingin dipastikan bahwa tahap pra-pemrosesan data benar-benar berlaku untuk kernel graf yang berbeda. Hasil di atas menunjukkan bahwa untuk kernel graf KEHL dan KVHL, tahap pra-pemrosesan memberikan pengaruh terhadap kinerja pengklasifikasi SVM. Sedangkan pada kernel graf KEHG, proses pra-pemrosesan justru mengurangi kinerja pengklasifikasi SVM. Gambar 1 di bawah ini menampilkan nilai kernel graf pada masing-masing matriks kernel sebelum tahap pra-pemrosesan. Nilai yang ditampilkan hanya mencakup matriks 5 x 5 dari keseluruhan yang berukuran 6795 x 6795.

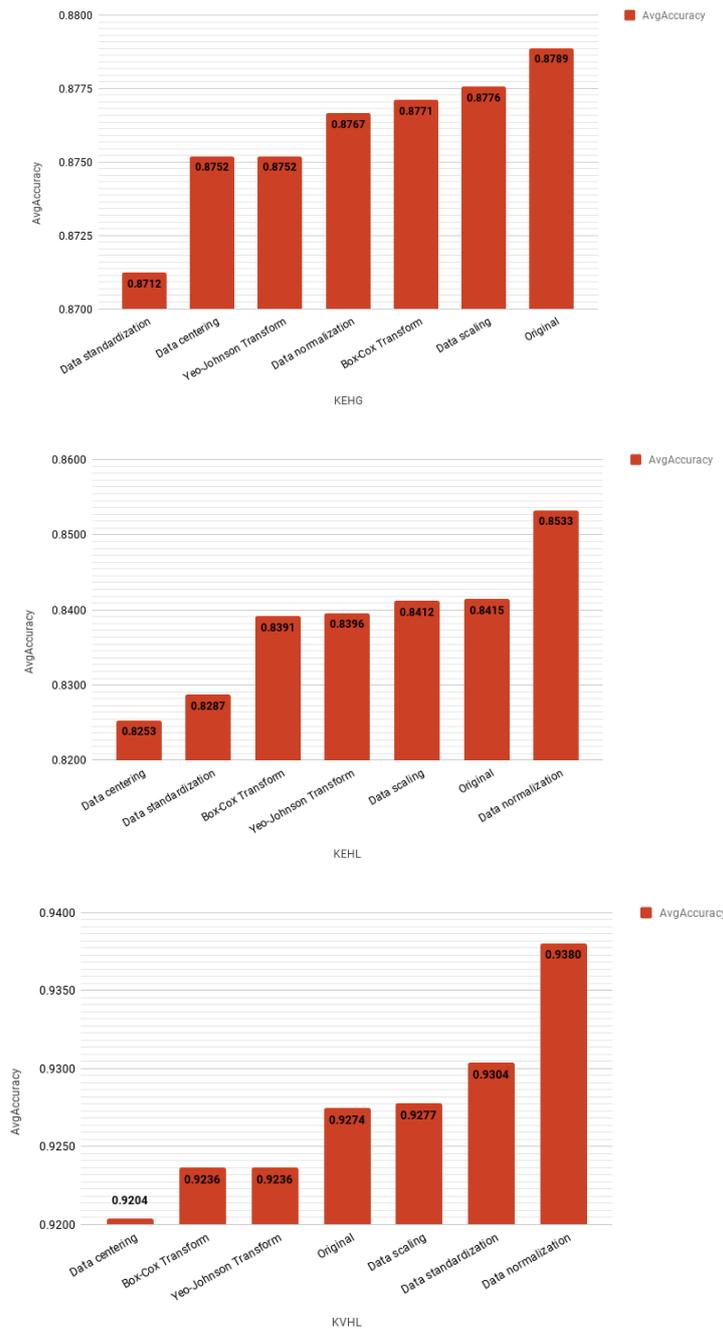
```
> ep_KEHG[1:5,1:5]
      [,1] [,2] [,3]      [,4]      [,5]
[1,] 1.00000e+00 0 0 0.00000e+00 1.91517e-174
[2,] 0.00000e+00 1 0 0.00000e+00 0.00000e+00
[3,] 0.00000e+00 0 1 0.00000e+00 0.00000e+00
[4,] 0.00000e+00 0 0 1.00000e+00 1.91517e-174
[5,] 1.91517e-174 0 0 1.91517e-174 1.00000e+00
> ep_KEHL[1:5,1:5]
      [,1] [,2] [,3] [,4] [,5]
[1,] 1 0 0 0 0
[2,] 0 16 6 6 2
[3,] 0 6 18 2 1
[4,] 0 6 2 5 2
[5,] 0 2 1 2 1
> ep_KVHL[1:5,1:5]
      [,1] [,2] [,3] [,4] [,5]
[1,] 2 0 0 0 0
[2,] 0 8 0 4 0
[3,] 0 0 9 0 0
[4,] 0 4 0 4 0
[5,] 0 0 0 0 2
```

Gambar 1. Tampilan 5 x 5 nilai matriks kernel KEHG, KEHL dan KVHL

Dari tampilan di atas, terlihat bahwa nilai matriks kernel KEHG relatif sangat kecil dibandingkan dengan kedua matriks kernel lainnya. Hal ini menyebabkan pra-pemrosesan data matriks kernel KEHG tidak memberikan pengaruh yang signifikan dalam peningkatan kinerja pengklasifikasi SVM. Sebaliknya, pengklasifikasi SVM memberikan hasil yang berbeda antara data original dengan data hasil pra-pemrosesan.

Selanjutnya untuk melihat kinerja pengklasifikasi SVM secara lebih detail, dapat dilihat pada Gambar 2 berikut. Akurasi klasifikasi dinyatakan dinyatakan sebagai persen setelah terhadap matriks kernel KEHG, KEHL dan KVHL dilakukan tahap pra-pemrosesan.





Gambar 2. Nilai Akurasi Klasifikasi setelah tahap pra-pemrosesan.

Dari hasil di atas, percobaan yang dilakukan menunjukkan bahwa hasil terbaik dicapai ketika metode pra-pemrosesan adalah normalisasi data, mencapai akurasi klasifikasi hingga 85,33,15 % untuk KEHL dan 93,80 untuk KVHL. Sebaliknya, untuk KEHG, aplikasi metode pra-pemrosesan menurunkan akurasi.

4. Kesimpulan

Dalam penelitian ini, didiskusikan pengaruh beberapa metode pra-pemrosesan data untuk kernel graf yang berbeda. Menentukan metode pra-pemrosesan yang benar-benar tepat cukup sulit dilakukan. Bila atribut input dataset mempunyai skala yang sama, maka metode pra-pemrosesan yang efektif digunakan adalah metode berdasarkan sampel. Sedangkan metode regresi akan lebih tepat diterapkan untuk atribut input yang terstandar. Eksperimen dengan

menggunakan macam-macam transformasi data dan algoritma *machine learning* yang berbeda untuk mendapatkan hasil terbaik pada dataset yang ada dapat dilakukan. Hasil percobaan ditunjukkan bahwa kernel graf yang berbeda dengan metode pra-pemrosesan yang ada memberikan hasil akurasi yang berbeda pula.

Referensi

- [1] V. López, A. Fernández, J. G. Moreno-Torres and F. Herrera. 2012. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. *Expert Systems with Applications*, vol. 39, pp. 6585-6608.
- [2] S. Yazdani, J. Shanbehzadeh and M. T. Manzuri Shalmani. 2012. RPCA: A Novel Preprocessing Method for PCA. *Advances in Artificial Intelligence*, vol. 2012, pp. 1-7.
- [3] P. Jha, K. K. Lavania, D. Dembla and H. Arora. 2013. SVC-ACO Architecture : An Efficient Data Preprocessing. *International Journal of Electronics Engineering*, vol. 5, no. 1, pp. 5-8.
- [4] K. Adamiak, P. Duch and K. Ślot. 2016. Object Classification Using Support Vector Machines with Kernel-based Data Preprocessing. *Image Processing & Communications*, vol. 21, pp. 45-53, 2016.
- [5] T. Karunaratne and H. Boström, "Pre-Processing Structured Data for Supervised Graph Propositionalization - a Case Study with Medicinal Chemistry Datasets," in *2010 Ninth International Conference on Machine Learning and Applications (ICMLA)*, Washington.
- [6] S. Almuhaideb and M. E. B. Menai. 2016. Impact of preprocessing on medical data classification. *Frontiers of Computer Science*, vol. 10, no. 6, pp. 1082-1102.
- [7] Z. Yang, A. J. Smola, L. Song and A. G. Wilson. 2014. A la Carte - Learning Fast Kernels. vol. 38, pp. 1098-1106.
- [8] L. X. Chen. 2017. Do patent citations indicate knowledge linkage? The evidence from text similarities between patents and their citations. *Journal of Informetrics*, vol. 11, no. 1, pp. 63-79.
- [9] M. Sugiyama and K. M. Borgwardt. 2015. Halting in Random Walk Kernels. *Advances in Neural Information Processing Systems*, no. Section 2, pp. 1639 - 1647.
- [10] C. Zhu and D. Gao. 2016. Influence of data preprocessing. *Journal of Computing Science and Engineering*, vol. 10, pp. 51-57.
- [11] O. Devos, G. Downey and L. Duponchel. 2014. Simultaneous data pre-processing and SVM classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils. *Food Chemistry*, vol. 148, pp. 124-130.
- [12] B. Nugroho and E. Marlina. 2017. Pengkategorian Otomatis Artikel Ilmiah dalam Pangkalan Data Perpustakaan Digital Menggunakan Metode Kernel Graph. *Jurnal IPTEK-KOM : Jurnal Ilmu Pengetahuan dan Teknologi Komunikasi*, vol. 19, no. 2, pp. 95 - 106.
- [13] M. Kuhn, "Building Predictive Models in R Using the caret Package. 2008. *Journal of Statistical Software, Articles*, vol. 28, no. 5, pp. 1-26.
- [14] A. Karatzoglou, A. Smola, K. Hornik and A. Zeileis. 2004. Kernlab - An S4 Package for Kernel Methods in R," *Journal of Statistical Software*, vol. 11, no. 9, pp. 1-20.

