Implementasi Integrasi Seleksi Data dan Extreme Learning Machine (IDELM) Untuk Klasifikasi DNA Bakteri Patogen

by Umi Mahdiyah

Submission date: 09-Apr-2021 02:29AM (UTC+0700)
Submission ID: 1553916940
File name: earning_Machine_IDELM_Untuk_Klasifikasi_DNA_Bakteri_Patogen.docx (296.51K)
Word count: 1839
Character count: 11803

Prosiding Seminar Nasional Integrasi Matematika dan Nilai Islami
Vol.3, No.1, September 2019, pp. xx~xx
p-ISSN: 2580-4596; e-ISSN:2580-460X

Implementasi Integrasi Seleksi Data dan Extreme Learning Machine (IDELM) Untuk Klasifikasi DNA Bakteri Patogen

Umi Mahdiyah[1], Lilia Sinta Wahyuniar[2]
[1,2]Jurusan Teknik Informatika, Universitas Nusantara PGRI Kediri
umimahdiyah@gmail.com, li2asint@gmail.com

Info Artikel

Riwayat Artikel:

Diterma:
Direvisi:
Diterbitkan:

Kata Kunci:

Pertama
Kedua
Ketiga
Keempat
Kelima

ABSTRAK

PCR adalah salah satu metode untuk mendeteksi keberadaan mikroba dalam tubuh. Dibanding metode lain, metode ini tergolong akurat, cepat, dan bisa diandalkan. Keuntungan PCR lainnya adalah sekuens DNA dari mikroba atau strain infeksi yang baru ditemukan. Selanjutnya data DNA yang ditemukan tersebut dilakukan pencarian masuk pada jenis DNA apa, sehingga dibutuhkan metode yang optimal. Sekuensing DNA dapat dimanfaatkan untuk menentukan identitas maupun fungsi gen atau fragmen DNA lainnya dengan cara membandingkan sekuens-nya dengan sekuens DNA lain yang sudah diketahui. Integrasi Seleksi data dan Extreme Learning Machine ini dipilih sebagai metode untuk klasifikasi DNA karena data DNA merupakan data yang besar serta karakteristik datanya yang kebanyakan adalah data yang *imbalance*. Pada proses penelitian data yang akan diolah terlebih dahulu diuraikan fragmennya, selanjutnya dilakukan ekstraksi fitur, kemudian dilakukan proses klasifikasi. Hasil dari pengklasifikasian tersebut diperoleh precision, recall, gmean, dan akurasi masing-masing sebesar 0,917, 0,824, 0,793, dan 0,943.

Korespondensi:

Penulis,
Jurusan Matematika,
UIN Maulana Malik Ibrahim Malang,
Jl. Gajayana No. 50 Malang, Jawa Timur, Indonesia 65144
penulis@gmail.com

1. PENDAHULUAN (10 PT)

DNA merupakan unsur yang sangat penting dan mendasar pada setiap organisme, DNA sendiri tuliskan dalam bentuk sebuah urutan symbol-simbol yang mengkodekan ciri dari suatu organisme tersebut. Urutan tersebut dikenal sebagai sequence DNA, yang merupakan informasi paling mendasar suatu gen atau genom karena mengandung instruksi yang dibutuhkan untuk pembentukan tubuh makhluk hidup[1].
PCR(Polymerase Chain Reaction) [2] adalah salah satu metode untuk mendeteksi keberadaan mikroba dalam tubuh. Dibanding metode lain, metode ini tergolong akurat, cepat, dan bisa diandalkan. Keuntungan PCR lainnya adalah sekuens DNA dari mikroba atau strain infeksi yang baru ditemukan. Selanjutnya data DNA yang ditemukan tersebut dilakukan pencarian masuk pada jenis DNA apa, sehingga dibutuhkan metode yang optimal. Sekuensing DNA dapat dimanfaatkan untuk menentukan identitas maupun fungsi gen atau fragmen DNA lainnya dengan cara membandingkan sekuens-nya dengan sekuens DNA lain yang sudah diketahui. Pembandingan yang banyak dilakukan adalah dengen proses pensejajaran 2 atau lebih sequence DNA. Proses penseajran ini membutuhkan waktu yang cukup lama. Sehingga dalam penelitian ini dibahas

pengkalsifikasian DNA dengan menggunakan Machine Learning[3]. Dalam proses kalsifikasi dengan machine learning tidak mensejajarkan, tetapi lebih ke pembuatan model dari masing- masing jeis DNA sehingga jika ada DNA baru yang diujikan dapat dikenali.

Dalam penelitian ini dilakukan pengelompokan DNA dengan metode klasifikasi, yakni metode IDELM(Integrasi Selekesi Data dan *Extreme Learning Machine*) [4]. Metode IDELM dipilih dikarenakan data DNA yang sangat besar dan merupakan data dengan tipe *imbalance data*, maka perlu dilakukan seleksi data terlebih dahulu. Seleksi data ini dilakukan untuk mengatasi masalah *imbalanced data*.

Konsep dari IDELM sendiri adalah melakukan proses seleksi data dan klasifikasi secara bersamaan. Seleksi data dilakukan untuk memilih mana fitur yang perlu digunakan dan mana fitur yang sudah tergantikili dengan fitur yang terpilih sebelumnya. Proses seleksi data dan klasifikasi dilakukan secara bersamaan untuk menghindari masalah inkonsistensi antara proses seleksi data dan klasifikasi. Sehingga dengan menerapkan metode IDELM pada proses klasifikasi untuk pengelompokan DNA ini diharapkan hasil yang dicapai lebih optimal.

2. METODE PENELITIAN (10 PT)

Langkah-langkah dari penelitian ini adalah sebagai berikut:

1) Pengumpulan Data

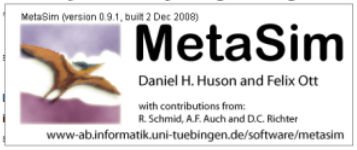
Dalam tahap ini juga dilakukan proses pengumpulan data dengan cara melakukan pengabilan data berupa FASTA dari DNA pada web NCBI. Data yang telah digunakan dalam penelitian ini dapat dilihat pada tabel 1.

Tabel 1. Data Penelitian

No	Bakteri Patogen	Bakteri Nonpatogen
1	*Bartonella bacilliformis*	*Bifidobacterium animals*
2	*Bordetella pertussis*	*Bifidobacterium bifidum*
3	*Borrelia recurrentis*	*Bifidobacterium breve*
4	*Haemophilus influenzae*	*Bifidobacterium adolescentis*
5	*Haemophilus ducreyi*	*Bifidobacterium longum*
6	*Streptococcus salivarius*	*Lactobacillus delbrueckii*
7	*Streptococcus Pyogenes*	*Lactobacillus acidophilus*
8	*Streptococcus mutans*	*Lactobacillus brevis*
9	*Streptococcus agalactiae*	*Lactobacillus fermentum*
10	*Brucella abortus*	*Lactobacillus reuteri*

Selanjutnya data disiapkan sedemikian rupa (dibuat dalam bentuk fragmen DNA) untuk diklasifikasikan. Pengumpulan data ini dilakukan oleh ketua peneliti. Adapun proses penyiapan data adalah:

a. Penguraian Data Menjadi Beberapa Fragmen
   Pada proses penguraian data DNA menjadi beberapa fragmen digunakan aplikasi MetaSim[5].



Gambar 1. Simulator MetaSim

   Simulator ini menguraikan DNA menjadi masing-masing fragmennya.

b. Ekstraksi Fitur dengan *n-mers*
   Selanjutnya dari data tersebut dilakukan ekstraksi fitur dengan menggunakan *n-mers*. Metode ini digunakan untuk mengetahui banyaknya kemunculan *substring* tertentu pada sebuah *string*[6]. Artinya untuk data DNA yang tersusun dari 4 jenis basa (A,C,G,T), sehingga jika n=3 makan akan ada 4^3 yaitu membentuk 64 *substring*, sedangkan jika n=4 maka akan ada 4^4 yaitu 256 *substring*, dalam hal ini n yang digunakan adalah 3, 4, dan 5. Pada setiap data dicari pula nilai rata-rata dan standar deviasinya untuk menjadi bagian dari fitur yang digunakan. Selanjutnya setiap data diberikan label sesuai dengan jenisnya, *patogen* dilabeli dengan 1 dan *nonpatogen* dilabeli dengan 0.

c. Normalisasi Data
   Proses normalisasi data pada penelitian ini adalah menggunakan normalisasi minmax dengan rumus dapat dilihat pada persamaan 1.

$$x_{norm} = \left(\frac{x - x_{min}}{x_{max} - x_{min}} \times (max_{new} - min_{new})\right) + min_{new} \quad (1)$$

Keterangan:
x_{norm}= data hasil normalisasi

x_{min} = nilai minimum dari data per kolom

x_{max}= nilai maximum dari data per kolom

min_{new} = adalah batas minimum yang kita berikan

max_{new} = adalah batas maximum yang kita berikan

d. Pembagian data training dan testing
Dalam penelitian ini, data training dan data testing dibagi menggunakan *k-fold cross validation* yang diilustrasikan pada tabel 2. Data training dikelompokkan menjadi 5 kelompok, jika 1 kelompok untuk data testing maka sisanya untuk training. Sehingga berdasar ilustrasi pada tabel 2 proses training dan testing sebanyak 5 kali.

Tabel 2. Ilustrasi *5-fold cross validation*

Kelompok 1	Kelompok 2	Kelompok 3	Kelompok 4	Kelompok 5
Kelompok 1	Kelompok 2	Kelompok 3	Kelompok 4	Kelompok 5
Kelompok 1	Kelompok 2	Kelompok 3	Kelompok 4	Kelompok 5
Kelompok 1	Kelompok 2	Kelompok 3	Kelompok 4	Kelompok 5
Kelompok 1	Kelompok 2	Kelompok 3	Kelompok 4	Kelompok 5

Keterangan:

: Data Training

: Data Testing

2) Proses Training

Tahap desain dan perancangan sistem akan menerjemahkan syarat kebutuhan ke sebuah perancangan perangkat lunak sebelum dibuat *coding*. Pada proses ini dibuat perancangan arsitektur perangkat lunak, struktur data, dan algoritma prosedural. pada tahap desain dan perancangan sistem ini dilakukan dengan diskusi bersama antara ketua peneliti dan anggota.

3) Pengujian Sistem (*Testing & Integration*)

Pada tahap ini dilakukan pengujian terhadap program yang telah dibuat dengan cara melakukan uji coba terhadap semua fungsi pada sistem. Tahap pengujian sistem juga dilakukan sendiri oleh ketua peneliti, pengujian dilakukan dengan mengunakan berbagai macam data DNA, sehingga dapat diketahui seberapa baik kemampuan sistem yang dibuat.

Untuk pengujian kemampuan dilakukan analisis confusion matriks, krena data yang digunakan termasuk dalam *imbalanced data*. *Imbalanced data* merupakan kasus khusus dalam *Machine Learning*. Ukuran evaluasi memainkan peran penting dalam *machine learning*. Ukuran tersebut digunakan untuk mengevaluasi dan mengarahkan algoritma pembelajaran. Jika pilihan ukuran mengabaikan kelas minoritas, maka algoritma pembelajaran tidak akan mampu menangani masalah *imbalanced data* dengan baik. Ukuran yang umum digunakan untuk dalam penelitian biasanya adalah tingkat klasifikasi keseluruhan yaitu akurasi. Namun pada *imbalanced* dataset, tingkat klasifikasi keseluruhan tidak lagi menjadi ukuran yang cocok, karena kelas minoritas tidak berpengaruh pada akurasi dibandingkan dengan kelas mayoritas.

Oleh karena itu, ukuran lainnya telah dikembangkan untuk menilai kinerja *classifier* untuk data yang *imbalanced*. Berbagai ukuran yang umum didefinisikan berdasarkan *confusion matrix*. *Confusion matrix* untuk klasifikasi biner ditunjukkan dalam Tabel 3.

Tabel 3 *Confusion matrix* untuk klasifikasi biner

		Nilai Sebenarnya	
		True	*False*
Prediksi	*True*	TP *(True Positive)*	FP *(False Positive)*
	False	FN *(False Negative)*	TN *(True Negative)*

Diantara berbagai kriteria evaluasi, ukuran yang paling relevan dengan data yang *imbalanced* yaitu *precision*, *recall*, *sensitifity*, *specificity*, dan *geometric mean* (*G-mean*)[7].

Precision dalam artikel ini adalah berapa persen bakteri yang benar patogen dari keseluruhan bakteri yang diprediksi patogen. Sedangkan *recall* dalam penelitian ini adalah ketepatan berapa persen bakteri yang diprediksi patogen dibandingkan keseluruhan bakteri yang sebenarnya patogen.

$$precision = \frac{TP}{TP + FP}.$$

$$recall = \frac{TP}{TP + FN}.$$

Ukuran *Geometric mean* (*G-mean*) ini digunakan ketika performa dari kedua kelas yang bersangkutan dan diharapkan tinggi secara bersamaan. *Geometric mean* telah digunakan beberapa peneliti untuk mengevaluasi *classifier* pada dataset yang *imbalanced* . *G-mean* mengindikasikan keseimbangan antara kinerja klasifikasi pada kelas mayoritas dan minoritas. Ukuran *G-mean* diambil berdasarkan s*ensitifity* (akurasi dari data positif) dan s*pecificity* (akurasi data negatif).

$$sensitifity = recall$$

$$specificity = 1 - \frac{TP}{FP + TN}$$

$$G - mean = \sqrt{sensitivity \times specificity}$$

3. HASIL DAN PEMBAHASAN(10 PT)

Hasil penelitian dapat dilihat pada tabel 4. Dari tabel 4 ditunjukkan seberapa besar nilai *precision, recall, g-mean,* dan *akurasi* dari pengklasifikasian data patogen dan non patogen.

Tabel 4. Hasil Penelitian

	Precision	***Recall***	***G-mean***	***Akurasi***
1000bp				
3-mers	0,950	0,905	0,829	0,970
4-mers	0,900	0,783	0,774	0,930
5-mers	0,850	0,680	0,725	0,890
Rata-rata	0,900	0,789	0,776	0,930
2000bp				
3-mers	0,950	0,905	0,829	0,970
4-mers	0,900	0,818	0,793	0,940
5-mers	0,900	0,750	0,757	0,920
Rata-rata	0,917	0,824	0,793	0,943

Dari tabel sangat terlihat akurasi yang dimiliki cukup baik, dapat dilihat dari rata-rata *precision, recall, G-mean, dan akurasi*, masing masing pada sequence DNA yang panjang fragmennya 1000bp adalag 0,900, 0,789, 0,776,dan 0,930. Sedangkan untuk yang ukuran fragmennya 2000bp masing-masing nilai *precision, recall, G-mean, dan akurasi*nya adalah 0,917, 0,824, 0,793, dan 0,943. Perhitungan akurasi yang digunakan tidak hanya perhitungan akurasi biasa karena data DNA termasuk dalam kategori data yang berkarakter *imbalance*. Hal tersebut terjadi karena meskipun sama-sama menggunakan 10 data masing - masing untuk data *patogen dan non patogen* saat dilakukan penguraian fragmen dengan panjang fragmen 1000bp dan 2000bp banyak fragmen yang terbentuk memiliki perbandingan yang jauh kberbeda nata data patogen dan non patogen, perbandingannya sekitar 1:10. Dari tabel di atas juga dapat dilihat pada masing masing panjang fragmen yang berbeda, jika dilihat dari proses ektraksi fitur sequence DNA tersebut nika nili n pada n-mers semakin bersar maka akurasi yang di dapat semakin sedikit. Sedangkan jika dilihat dari panjanganya fragmen semakin panjang fragmen yang dibuat, dalam penelitian ini nilai akurasinya semakin bagus.

4. KESIMPULAN (10 PT)

Dari uraian yang telah dituliskan dapat disimpulkan bahwa klasifikasi sequence DNA bakteri patogen non patogen memiliki performansi yang cukup baik. Dari hasil yang ada dapat dilihat bahwa rata-rata seluruh hasil melebihi 0,75, artinya akurasinya sekirat 75%.

5. UCAPAN TERIMAKASIH (10 PT)

Termakasih saya sampaikan kepada instansi UNP Kediri yang telah memberikan support yang sangat optimal. Serta diucapkan pula terimakasih pada pihak KEMENRISTEK Diki telah memberikan bantuan dana untuk terlaksannanya penelitian ini. Terimakasih pula kepada seluruh pihak yang telah mendukung dan membantu terlaksananya penelitian ini.

DAFTAR PUSTAKA (10 PT)

[1] Dolanski A, Kimmel M. 2007. *Bioinformatics*. Berlin (DE): Springer.
[2] Alberts, Bruce; Johnson, Alexander; Lewis, Julian; Morgan, David; Raff, Martin; Roberts, Keith; Walter, Peter (2014). *Molecular Biology of the Cell, Sixth Edition*. Garland Science. United State of America
[3] Wang J, Zaki M, Toivonen H, Shasha D. 2005. *Data Mining in Bioinformatics*. London (UK): Springer.
[4] Mahdiyah, Umi, Imah, E. M., Irawan, M. I.. 2017. *Integrating Data Selection And Extreme Learning Machine To Predict Protein-Ligand Binding Site*, Contemporary Enggineering Science, vol. 9.
[5] Richter DC, *et al*. 2009. *User manual for MetaSim* V0.9.5 [Internet]. [diunduh 2014 Juni 5]. Tersedia pada: ab.informatik.uni.tuebigen.de/software/metasim/download/V 0 9 5/manual.pdf
[6] Erbert M, Rechner S, Müller-Hannemann M. Gerbil. 2017. *a fast and memory-efficient k-mer counter with GPU-support*. Algorithms for Molecular Biology. Springer Nature. doi:10.1186/s13015-017-0097-9
[7] Narwat, A.,2018. Classification Assessment Methods, *Applied Computing and Informatics*
[8] Huang, G., Zhu, Q. dan Siew, C., (2006a), "*Extreme Learning Machine*: Theory and Applications", *Neurocomputing*, Vol. 70, 489–501

Implementasi Integrasi Seleksi Data dan Extreme Learning Machine (IDELM) Untuk Klasifikasi DNA Bakteri Patogen

ORIGINALITY REPORT

20%	19%	8%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	simanis.uin-malang.ac.id Internet Source	1%
2	Submitted to Universitas Gunadarma Student Paper	1%
3	english.rejbrand.se Internet Source	1%
4	distan.jogjaprov.go.id Internet Source	1%
5	eprints.undip.ac.id Internet Source	1%
6	artikataku.blogspot.com Internet Source	1%
7	es.scribd.com Internet Source	1%
8	Elvismary Molina de Armas, Maristela Holanda, Daniel de Oliveira, Nalvo F. Almeida, Sérgio Lifschitz. "Chapter 1 A Classification of de Bruijn Graph Approaches for De Novo	1%

Fragment Assembly", Springer Science and Business Media LLC, 2020
Publication

9	Submitted to Universitas Brawijaya Student Paper	1 %
10	conferences.uin-malang.ac.id Internet Source	1 %
11	idoc.pub Internet Source	1 %
12	iptek.its.ac.id Internet Source	1 %
13	firlyamalia.wordpress.com Internet Source	1 %
14	Submitted to iGroup Student Paper	1 %
15	sinta.ristekbrin.go.id Internet Source	1 %
16	tampub.uta.fi Internet Source	1 %
17	digilib.uinsby.ac.id Internet Source	1 %
18	download.garuda.ristekdikti.go.id Internet Source	1 %
19	ejournal3.undip.ac.id Internet Source	1 %

20	Submitted to UIN Sultan Syarif Kasim Riau Student Paper	1 %
21	Submitted to Universitas Sumatera Utara Student Paper	1 %
22	ipi.portalgaruda.org Internet Source	1 %
23	repository.its.ac.id Internet Source	1 %
24	unsri.portalgaruda.org Internet Source	1 %
25	worldwidescience.org Internet Source	1 %

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		