Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 1, No. 1, March 2018, pp. 1-5
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Evaluation of F-Measure and Feature Analysis of C5.0 Implementation on Single Nucleotide Polymorphism Calling

[1]Lailan Sahrina Hasibuan, [2]Sita Nabila, [3]Nurul Hudachair, [4]Muhammad Abrar Istiadi
[1,2,3,4]Department of Computer Sciences, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University
[1,4]Bioinformatics Working Group, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University
Email: [1]lailan.sahrina@apps.ipb.ac.id, [2]nabila@apps.ipb.ac.id, [3]nurul_hudachair@apps.ipb.ac.id, [4]abrari@apps.ipb.ac.id

Article Info
Article history:
Received Feb 4th, 2018
Revised Feb 17th, 2018
Accepted Mar 5th, 2018

Keywords:
Analisis fitur
C5.0
Decision tree
NGS
SNP *calling*

ABSTRACT
Data growing in molecular biology has increased rapidly since Next-Generation Sequencing (NGS) technology introduced in 2000, the latest technology used to sequence DNA with high throughput. Single Nucleotide Polymorphism (SNP) is a marker based on DNA which can be used to identify organism specifically. SNPs are usually exploited for optimizing parents selection in producing high-quality seed for plant breeding. This paper discusses SNP calling underlying NGS data of cultivated soybean (*Glycine max [L]. Merr*) using C5.0, an improved rule-based algorithm of C4.5. The evaluation illustrated that C5.0 is better than the other rule-based algorithm CART based on f-measure. The value of f-measure using C5.0 and CART are 0.63 and 0.58. Besides of that, C5.0 is robust for imbalanced training dataset up to 1:17 but it is suffer in large training dataset. C5.0's performance may be increased by applying bagging or the other ensemble technique as improvement of CART by applying bagging in final decision. The other important thing is using appropriate features in representing SNP candidates. Based on information gain of C5.0, this paper recommends error probability, homopolymer left, mismatch alt and mean nearby qual as features for SNP calling.

Corresponding Author:
Lailan Sahrina Hasibuan
Department of Computer Sciences
Faculty of Mathematics and Natural Sciences
Bogor Agricultural University
Meranti Wing 20 Level 5, Bogor Indonesia
Email: lailan.sahrina@apps.ipb.ac.id

1. PENDAHULUAN

Pertumbuhan data di bidang biologi molekuler mengalami peningkatan yang signifikan dalam dua dekade terakhir, terutama sejak teknologi *Next-Generation Sequencer* (NGS) diperkenalkan pada tahun 2000 [1]. NGS merupakan teknologi *sequencing Deoxyribo Nucleic Acid* (DNA) yang dilakukan secara paralel dan masif sehingga mampu menghasilkan data sekuen dengan ukuran *gigabyte* dalam waktu beberapa hari, suatu pekerjaan yang sebelumnya membutuhkan waktu belasan tahun [2]. DNA manusia yang memiliki panjang 3 *megabase pairs* (mbs) dapat dibaca dalam waktu 2 minggu menggunakan teknologi NGS, sementara sebelumnya membutuhkan waktu 15 tahun untuk menyelesaikannya.

Data sekuen DNA sangat bermanfaat untuk mempelajari sistem kehidupan seperti: pewarisan sifat dari satu generasi ke generasi berikutnya, kerentanan suatu makhluk hidup terhadap penyakit yang belum diketahui ciri-cirinya dan perancangan *personalized medicine* [3]. Dalam bidang pertanian, data ini umumnya dimanfaatkan untuk menentukan penanda fenotipe pada level genomik. Penggunaan penanda ini dapat menghemat waktu pada proses seleksi tanaman untuk menghasilkan bibit unggul [4] [5].

Pada saat ini, penanda pada level genomik yang umum digunakan adalah *Single Nucleotide Polymorphism* (SNP). SNP dapat diartikan sebagai variasi basa nukleotida pada posisi yang sama di antara fragmen-fragmen DNA yang dijajarkan. Fragmen-fragmen DNA tersebut berasal dari individu yang berbeda pada populasi yang sama. Penanda SNP menjadi umum digunakan karena jumlahnya yang cenderung banyak dibandingkan penanda lainnya seperti *Simple Sequence Repeat* (SSR), *Random-Amplified Polymorphic* DNA

(RAPD), *Amplified Fragment Length Polymorphism* (AFLP) [6]. Oleh karena itu, identifikasi SNP merupakan penelitian yang memiliki peran strategis dalam bidang bioinformatika. Identifikasi SNP umumnya disebut SNP *calling* atau *variant calling*. Pada penelitian ini, istilah yang digunakan untuk menyatakan identifikasi SNP adalah SNP *calling*.

Berbagai penelitian mengenai SNP *calling* telah dilakukan sebelumnya. Pada tahun 2013, O'Fallon *et. al*. melakukan SNP *calling* menggunakan *Support Vector Machine* (SVM) dengan *Radial Basis Function* (RBF) sebagai *kernel* pada data *exome* wanita Kaukasian di Eropa Timur. Evaluasi berdasarkan *sensitivity* adalah 87.5% [7]. Pada tahun 2014, Hasibuan *et. al.* melakukan SNP *calling* menggunakan SVM dan RBF sebagai *kernel* dengan fokus masalah pada ketidakseimbangan data. Data yang digunakan adalah *whole-genome* dari *Glycine max* [*L.*] *Merr* sebanyak 1% dari kromosom 16. Hasil penelitian ini menunjukkan bahwa penerapan *undersampling* mampu meningkatkan *f-measure* hingga 50%, yaitu 57% menjadi 89% [8].

Kedua penelitian sebelumnya menggunakan pendekatan *black box* sehingga, proses SNP *calling* bersifat kabur. Pada tahun 2017, Hasibuan *et. al.* menggunakan pendekatan *rule-based classification and regression trees* (CART) untuk proses SNP *calling* pada data *whole-genome* dari *Glycine max* [*L.*] *Merr.* kromosom 16 [9]. Evaluasi model berdasarkan *f-measure* adalah 57%. Penelitian ini berhasil mengidentifikasi 4 buah fitur yang dominan dari 24 fitur yang digunakan, yaitu: kualitas maksimum alel minor, frekuensi alel minor, total alel varian dan total alel referen.

Berdasarkan data empris dari penelitian Loh pada tahun 2011, C4.5 memiliki akurasi yang lebih tinggi dibandingkan CART dalam mengidentifikasi jenis kendaraan [10]. Algoritme C5.0 adalah pengembangan dari C4.5, salah satunya dari sisi *pruning* [11]. Oleh karena itu, penelitian ini melakukan SNP *calling* pada data *whole-genome* dari *Glycine max* [*L.*] *Merr.* menggunakan pendekatan *rule-based* C5.0 untuk menganalisis pengaruhnya terhadap *f-measure* dan menganalisis fitur-fitur dominan yang berperan dalam proses SNP *calling*.

2. DATA DAN METODE

2.1. Data

Data yang digunakan pada penelitian ini adalah data fragmen DNA kedelai budidaya (*Glycine max* [*L.*] *Merr.*) hasil penelitian Lam *et al* pada tahun 2010 [11]. Praproses data telah dilakukan oleh [8] dan [12] . Tahap praproses data yang dilakukan meliputi: uji kualitas terhadap fragmen-fragmen DNA, penjajaran terhadap genom rujukan, pengelompokan berdasarkan kromosom, penentuan kandidat SNP dan ekstraksi 24 fitur kandidat SNP. Kelas kandidat SNP terdiri atas 2 yaitu: positif dan negatif. SNP positif berarti variasi basa karena adanya *polymophism*. SNP negatif berarti variasi basa palsu, yaitu kemunculan variasi yang disebabkan adanya kesalahan pada proses-proses sebelumnya.

2.2. Pelatihan dan Pengujian

Glycine max [*L.*] *Merr.* memiliki 20 kromosom yang dikode dengan Gm01-Gm20. Kandidat SNP yang paling banyak terdapat pada Gm18, sementara kandidat SNP yang paling sedikit terdapat pada Gm16. Tabel 1 menunjukkan jumlah kandidat SNP pada kromosom yang digunakan pada penelitian ini. Pada tahap awal, penelitian ini menggunakan 2 kromosom terpendek yaitu Gm11 dan Gm16 untuk pelatihan dan pengujian model C5.0. Setelah proses pelatihan dan pengujian, analisis fitur dilakukan untuk mencari fitur yang memiliki peran paling dominan dalam proses SNP *calling*. Analisis dilakukan berdasarkan *information gain* setiap fitur. Semakin besar nilai *information gain*, semakin besar nilai dominansi fitur tersebut. Menurut [13] nilai *information gain* suatu fitur F yang memiliki 2 kemungkinan nilai v_1 dan v_2 dapat dihitung menggunakan Persamaan 1.

$$Gain(S,F) = H(S) - \frac{S_{v_1}}{S}H(S_{v_1}) - \frac{S_{v_2}}{S}H(S_{v_2}) \quad (1)$$

Dengan S_{v_1} adalah jumlah kandidat SNP yang memiliki nilai v_1 untuk fitur F, dan S_{v_2} adalah jumlah kandidat SNP yang memiliki nilai v_2 untuk fitur F. Selanjutnya, $H(S)$ adalah nilai *entrophy*. Jika S adalah jumlah kandidat SNP yang terdiri atas SNP positif S^+ dan SNP negatif S^-, maka $H(S)$ dapat dihitung menggunakan Persamaan 2.

$$H(S) = -\frac{S^+}{S}log_2\frac{S^+}{S} - \frac{S^-}{S}log_2\frac{S^-}{S} \quad (2)$$

Pada tahap selanjutnya, metode yang sama diterapkan pada 2 kromosom yang paling panjang yaitu Gm18 dan Gm01. Hal ini dilakukan untuk menjawab pertanyaan "Apakah fitur-fitur yang dominan pada kromosom Gm11 memiliki dominansi yang sama pada kromosom lainnya?"

Tabel 1. Jumlah kandidat SNP pada kromosom yang diteliti

Kromosom	SNP positif	SNP negatif	Perbandingan	Total SNP
Gm16	154817	1369758	1:8	1524576
Gm11	87508	1566161	1:17	1653670
Gm01	138368	2216578	1:16	2354946
Gm18	233097	2377348	1:10	2610445

2.3. Evaluasi

Performa algoritme C5.0 dalam melakukan SNP *calling* dievaluasi menggunakan matriks konfusi yang terdiri atas: *true positive* (TP), *false positive* (FP), *false negative* (FN) dan *true negative* (TN). Berdasarkan matriks konfusi, dapat diturunkan beberapa metrik evaluasi lainnya yaitu: akurasi, *sensitifity/recall/true positive rate* (TPR), *specificity*, *precision* dan *f-measure* [14] [15]. Pada penelitian ini, *f-measure* digunakan sebagai metrik evaluasi akhir karena adanya ketidakseimbangan data yang signifikan antara kelas SNP positif dan SNP negatif seperti yang tertera pada Tabel 1. Persamaan *f-measure* menurut [16] tersaji pada Persamaan 3.

$$f - measure = \frac{2 * precision * recall}{precision + recall} \quad (3)$$

3. HASIL DAN PEMBAHASAN

Performa model C5.0 dalam melakukan SNP *calling* disajikan pada Tabel 2. Data pada tabel tersebut menunjukkan bahwa performa C5.0 lebih baik dibandingkan CART berdsarkan *f-measure*, yaitu 0.63 dan 0.58. Nilai *f-measure* 0.63 dibentuk dari nilai *recall* 0.60 yang memiliki arti bahwa setiap kandidat SNP kelas positif memiliki peluang 60% teridentifikasi sebagai SNP positif dan nilai *precision* 0.67 yang memiliki arti bahwa SNP *calling* yang dilakukan memiliki peluang 67% benar. Jika dibandingkan dengan metode CART yang dilengkapi teknik *bagging* pada keputusan akhirnya, ternyata performa C5.0 mampu menyamainya berdasarkan metrik *f-measure*, yaitu 0.63.

SNP *calling* merupakan kasus *binary classification* dengan data tidak seimbang. Ketidakseimbangan data SNP positif dan SNP negatif tersaji pada Tabel 1. Ketidakseimbangan yang paling tinggi terjadi pada Gm11. Pada dataset ini, SNP positif hanya diwakili 5% dari total data. Menurut [17], pelatihan model menggunakan data tidak seimbang akan menghasilkan model yang bias karena cenderung terhadap kelas mayor, yaitu SNP negatif.

Berdasarkan data yang disajikan pada Tabel 2, *f-measure* model yang dilatih menggunakan data dengan ketidakseimbangan yang tinggi, yaitu Gm11, dapat mengungguli *f-measure* model Gm18 dengan ketidakseimbangan yang lebih rendah. Hal ini menunjukkan bahwa C5.0 *robust* terhadap data tidak seimbang. Namun, C5.0 tidak *robust* terhadap ukuran data uji. Ukuran data uji yang besar dapat menurunkan nilai *f-measure*. Hal ini terlihat dari penurunan nilai *f-measure* sebesar 4% (0.03) antara model Gm11 yang diuji oleh Gm16 dan model Gm18 yang diuji oleh Gm01.

Pada penelitian sebelumnya [9], penerapan *bagging* untuk algoritme CART dapat meningkatkan *f-measure* model sebanyak 8% (0.05). Hal yang sama dapat terjadi pada algoritme C5.0, penggunaan *bagging* ataupun teknik *ensemble* lainnya memiliki peluang meningkatkan metrik *f-measure*.

Tabel 2. Evaluasi performa model dalam SNP *calling*

Dataset		Model	*Recall*	*Precision*	*f-measure*
Data Latih	Data Uji				
Gm11	Gm16	C5.0	0.60	0.67	**0.63**
		CART*	0.51	0.67	**0.58**
		Bagging CART*	0.6	0.66	**0.63**
Gm18	Gm01	C5.0	0.56	0.64	**0.60**

*Hasil berdasarkan penelitian [9]

Performa model dalam melakukan SNP *calling*, selain dipengaruhi oleh model atau algoritme yang digunakan, juga dipengaruhi oleh fitur yang digunakan untuk merepresentasikan setiap variasi basa. Dengan kata lain, penggunaan fitur yang tepat dapat meningkatkan performa model [19]. Berdasarkan *information gain*, dua fitur yang paling berpengaruh pada pelatihan model menggunakan Gm11 adalah *error probability* dan *homopolymer left*. Sementara itu, dua fitur yang paling berpengaruh pada pelatihan model menggunakan Gm18 adalah *mismatch alt* dan *mean nearby qual*.

Dengan asumsi bahwa fragmen-fragmen yang mengandung kandidat SNP memiliki distribusi binomial, berikut ini dijelaskan pengertian dari fitur-fitur yang dianggap dominan. *Error probability* menyatakan peluang kandidat SNP adalah *homozigot*, *heterozigot* dan *error* [7]. *Homopolymer left*

menyatakan banyaknya *homopolimer* genom rujukan pada bagian kiri kandidat SNP [7]. *Mismatch alt* menyatakan jumlah kandidat SNP pada fragmen yang mengandung kandidat SNP [7]. *Mean nearby qual* menyatakan rata-rata kualitas basa yang mengapit posisi kandidat SNP [7].

4. KESIMPULAN

Berdarkan hasil yang diperoleh pada penelitian ini, terdapat beberapa kesimpulan yang disajikan sebagai berikut:

(1) Secara empiris, C5.0 dan *bagging* CART memiliki performa yang sama dalam melakukan SNP *calling*. Oleh karena itu, penggunan *bagging* atau teknik *ensemble* lainnya memiliki peluang meningkatkan performa C5.0.
(2) Secara empiris, C5.0 *robust* terhadap data tidak seimbang namun tidak *robus* terhadap ukuran data uji yang besar.
(3) Berdasarkan *information gain*, penelitian ini merekomendasikan penggunaan fitur *error probability, homopolymer left, mismatch alt* dan *mean nearby qual* untuk melakukan SNP *calling*.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada ticntech yang telah menyediakan *server* untuk mengolah data pada penelitian ini

REFERENSI

[1] M. Barba, H. Czosnek, and A. Hadidi, "Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology," *viruses,* vol. 6, nº 2014, pp. 106-136, 2014.

[2] M. Gu`vic, "The History of DNA Sequencing," *J Med Biochem,* vol. 32, nº 2013, pp. 301-312, 2013.

[3] W. Kong, and K. W. Choo, "Predicting Single Nucleotide Polymorphisms (SNP) from DNA sequence by Support Vector Machine," *Frontiers in Bioscience,* vol. 12, nº 2007, pp. 1610-1614, 2007.

[4] J. Mammadov, R. Aggarwal, R. Buyyarapu, and S. Kumpatla, "SNP Markers and Their Impact on Plant Breeding," *International Journal of Plant Genomics,* vol. 2012, 2012.

[5] N. M. Boopathi, "Marker-Assisted Selection," em *Genetic Mapping and Marker Assisted Selection*, Springer India, 2013, pp. 173-186.

[6] L. K. Matukumalli, J. J. Grefenstette, D. L. Hyten, I. Y. Choi, P. B. Cregan, and C. P. V. Tassell, "Application of machine learning in SNP discovery," *BMC Bioinformatics,* vol. 7, nº 4, 2006.

[7] B. D. O'Fallon, W. W. Donahue, and D. K. Crockett, "A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data," *Bioinformatics,* vol. 29, nº 11, p. 1361–1366, 2013.

[8] L. S. Hasibuan, W. A. Kusuma, and W. B. Suwarno, "Identification of single nucleotide polymorphism using support vector machine on imbalanced data," em *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Jakarta, 2014.

[9] L. S. Hasbuan, N. Hudachair, and M. A. Istiadi, "Bootstrap Aggregating of Classification and Regression Trees in Identification of Single Nucleotide Polymorphisms," em *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Jakarta, 2017.

[10] W. Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 1, pp. 14-23, 2011.

[11] R. Pandya, and J. Pandya, "C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning," *International Journal of Computer Applications,* vol. 117, nº 16, pp. 18-21, 2015.

[12] H.-M. Lam, X. Xu, X. Liu, W. Chen, G. Yang, F.-L. Wong, M.-W. Li, W. He, N. Qin, B. Wang, and J. Li, "Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection," *Nature Genetics,* vol. 42, nº 12, p. 1053–1059, 2010.

[13] M. A. Istiadi, W. A. Kusuma, and I. M. Tasma, "Application of Decision Tree Classifier for Single Nucleotide Polymorphism Discovery from Next-Generation Sequencing Data," em *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Jakarta, 2014.

[14] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques, Elsevier, 2011.

[15] H. He, and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering,* vol. 21, nº 9, pp. 1263-1284, 2009.

[16] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE*

Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 39, nº 2, pp. 539-550, 2009.

[17] S.-J. Yen, and Y.-S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Systems with Applications,* vol. 36, nº 3, pp. 5718-5727, 2009.

[18] Cieslak, A. David, and V. Nitesh, "Learning decision trees for unbalanced data," em *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2008.

[19] U. Ojha, M. Jain, G. Jain, and R. K. Tiwari, "Significance of Important Attributes for Decision Making Using C5.0," em *International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, New Delhi, 2017.

BIBLIOGRAFI PENULIS

Penulis adalah pengajar di Departemen Ilmu Komputer Institut Pertanian Bogor (IPB). Selain itu, penulis juga bergabung di Bioinformatics Working Group, FMIPA IPB

Penulis adalah mahasiswa S1 Ilmu Komputer Institut Pertanian Bogor (IPB)

Penulis adalah mahasiswa S1 Ilmu Komputer Institut Pertanian Bogor (IPB)

Penulis adalah pengajar di Departemen Ilmu Komputer Institut Pertanian Bogor (IPB). Selain itu, penulis juga bergabung di Bioinformatics Working Group, FMIPA IPB