Sistem Pakar Dengan Aplikasi 134 (2019) 93-101

Daftar isi tersedia di ScienceDirect

Sistem Pakar Dengan Aplikasi

beranda jurnal: www.elsevier.com/locate/eswa

Ulasan

Perbandingan metode pemilihan variabel random forest untuk pemodelan prediksi klasifikasi

Jaime Lynn Speiser* , Michael E. Miller, Janet Tooze, Edward IP

Departemen Ilmu Biostatistik, Wake Forest School of Medicine, Medical Center Boulevard, Winston-Salem, NC 27157, AS

ARTIKLE INFO

Riwayat artikel:
Diterima 11 Oktober 2018
Direvisi 21 Mei 2019
Diterima 22 Mei 2019
Tersedia secara online pada
23 Mei 2019

Kata kunci: Hutan
acak Pemilihan
variabel
Pengurangan fitur
Klasifikasi

ABSTRACT

Klasifikasi hutan acak adalah metode pembelajaran mesin yang populer untuk mengembangkan model prediksi di banyak lingkungan penelitian. Seringkali dalam pemodelan prediksi, tujuannya adalah untuk mengurangi jumlah variabel yang diperlukan untuk mendapatkan prediksi untuk mengurangi beban pengumpulan data dan meningkatkan efisiensi. Beberapa metode pemilihan variabel tersedia untuk pengaturan klasifikasi hutan acak; namun, ada kekurangan literatur untuk memandu pengguna tentang metode mana yang lebih disukai untuk berbagai jenis set data. Dengan menggunakan 311 set data klasifikasi yang tersedia secara online, kami mengevaluasi tingkat kesalahan prediksi, jumlah variabel, waktu komputasi, dan luas area di bawah kurva operasi penerima untuk berbagai metode pemilihan variabel random forest. Kami membandingkan metode pemilihan variabel random forest untuk berbagai jenis dataset (dataset dengan hasil biner, dataset dengan banyak prediktor, dan dataset dengan hasil yang tidak seimbang) dan untuk berbagai jenis metode (standard random forest versus metode conditional random forest dan metode berbasis tes versus metode berbasis kinerja). Berdasarkan penelitian kami, metode pemilihan variabel terbaik untuk sebagian besar dataset adalah metode Jiang dan metode yang diimplementasikan dalam paket *VSURF* R. Untuk dataset dengan banyak prediktor, metode yang diimplementasikan dalam paket R, yaitu *varSelRF* dan *Boruta,* lebih disukai karena efisiensi komputasi. Kontribusi yang signifikan dari penelitian ini adalah kemampuan untuk menilai teknik pemilihan variabel yang berbeda dalam pengaturan klasifikasi acak untuk mengidentifikasi metode yang lebih disukai berdasarkan aplikasi dalam sistem pakar dan cerdas.

Isi

* Penulis korespondensi.
Alamat email: jspeiser@wakehealth.edu (J.L. Speiser), mmiller@wakehealth.edu (M.E. Miller), jtooze@wakehealth.edu (J. Tooze), eip@wakehealth.edu (E. Ip).

https://doi.org/10.1016/j.eswa.2019.05.028

1. Int roduction

Random forest adalah prosedur pembelajaran mesin yang populer yang dapat digunakan untuk mengembangkan model prediksi. Pertama kali diperkenalkan oleh Breiman pada tahun 2001 (Breiman, 2001), random forest merupakan kumpulan pohon klasifikasi dan regresi (Breiman, Friedman, Olshen, & Stone, 1984), yang merupakan model sederhana yang menggunakan pemisahan biner pada variabel pra-diktor untuk menentukan prediksi hasil. Pohon keputusan mudah digunakan dalam praktiknya, menawarkan metode intuitif untuk memprediksi hasil yang membagi nilai "tinggi" versus "rendah" dari pra-diktor yang terkait dengan hasil. Meskipun menawarkan banyak manfaat, metodologi pohon keputusan sering kali memberikan akurasi yang buruk untuk set data yang kompleks (misalnya set data yang besar dan set data dengan interaksi variabel yang kompleks). Dalam pengaturan random forest, banyak pohon klasifikasi dan regresi dibangun menggunakan dataset pelatihan yang dipilih secara acak dan himpunan bagian acak dari variabel prediktor untuk hasil pemodelan. Hasil dari setiap pohon digabungkan untuk memberikan prediksi untuk setiap pengamatan. Oleh karena itu, random forest sering kali memberikan akurasi yang lebih tinggi dibandingkan dengan model pohon keputusan tunggal dengan tetap mempertahankan beberapa kualitas yang menguntungkan dari model pohon (misalnya kemampuan untuk menafsirkan hubungan antara prediktor dan hasil) (Speiser, Durkalski, & Lee, 2015). Random forest secara konsisten menawarkan akurasi prediksi tertinggi dibandingkan dengan model lain dalam hal klasifikasi (Fernández-Delgado, Cernadas, Barro, & Amorim, 2014).

Manfaat utama dari penggunaan random forest untuk pemodelan prediksi adalah kemampuannya untuk menangani dataset dengan jumlah variabel prediktor yang banyak; namun, sering kali dalam praktiknya, jumlah variabel prediktor yang diperlukan untuk mendapatkan prediksi hasil harus diminimalkan untuk meningkatkan efisiensi. Sebagai contoh, daripada menggunakan semua variabel yang tersedia dalam rekam medis elektronik, seseorang dapat memilih untuk menggunakan hanya sebagian dari variabel yang paling penting ketika mengembangkan model prediksi medis. Dalam pemodelan prediksi, sering kali yang menjadi perhatian adalah menentukan prediktor terpenting yang harus dimasukkan dalam model yang telah direduksi dan sederhana. Hal ini dapat dicapai dengan melakukan pemilihan variabel, di mana prediktor waktu diidentifikasi berdasarkan karakteristik statistik seperti kepentingan atau akurasi. Mengembangkan model prediksi dengan menggunakan seleksi variabel dapat mengurangi beban pengumpulan data dan dapat meningkatkan efisiensi prediksi dalam praktiknya. Karena banyak set data modern memiliki ratusan atau ribuan kemungkinan prediktor, pemilihan variabel sering kali merupakan bagian penting dari pengembangan model prediksi.

Pemilihan variabel dalam kerangka kerja hutan acak merupakan pertimbangan penting untuk banyak aplikasi dalam sistem pakar dan aplikasi. Secara umum, tujuan keseluruhan dari banyak sistem pakar adalah untuk membantu pengambilan keputusan untuk masalah yang kompleks. Hal ini sesuai dengan tujuan pemodelan prediksi, di mana kami menggunakan dataset untuk mengembangkan model (random forest dalam penelitian ini) yang akan memberikan prediksi hasil yang diinginkan. Untuk meningkatkan efisiensi dalam memperoleh prediksi model, pemilihan variabel dapat digunakan untuk mengidentifikasi subset variabel prediktor yang akan dimasukkan ke dalam model akhir yang lebih sederhana. Ada banyak aplikasi yang menggunakan hal ini dalam pengembangan sistem pakar, misalnya, mengembangkan alat pendukung keputusan medis, model proyeksi harga pasar saham, dan model analisis bisnis untuk memaksimalkan keuntungan. Ada beberapa metode yang tersedia untuk melakukan pemilihan variabel dalam pengaturan klasifikasi random forest. Banyak paket R yang menyediakan prosedur pemilihan variabel random forest, termasuk *boruta* (Kursa & Rudnicki, 2010), *varSelRF* (Díaz

Uriarte & De Andres, 2006), *VSURF* (Genuer, Poggi, & Tuleau-Malot, 2015), *caret* (Kuhn, 2008), *party* (Hothorn, Hornik, Strobl, & Zeileis, 2010), *randomForestSRC* (Ishwaran & Kogalur, 2014), *RRF* (Deng & Runger, 2013), *vita* (Janitza, Celik, & Boulesteix, 2015), *AUCRF* (Urrea & Calle, 2012), dan *fuzzyForest* (Conn, Ngun, Li, & Ramirez, 2015). Beberapa metode lain telah diusulkan dalam literatur (misalnya Hapfelmeier, 2013; Jiang dkk., 2004; Altmann, Tolos¸ i, Sander, dan Lengauer, 2010; Svetnik, Liaw, Tong, dan Wang, 2004). Meskipun ada banyak metode untuk pemilihan variabel random forest untuk masalah klasifikasi yang tersedia, ada kekurangan panduan dalam literatur tentang metode mana yang lebih disukai dalam hal tingkat kesalahan prediksi (out-of-bag), parsimoni (jumlah variabel), waktu komputasi dan luas area di bawah kurva operasi penerima (AUC) untuk berbagai jenis set data. Cadenas, Garrido, dan MartíNez (2013) dan Hapfelmeier (2013), Degenhardt, Seifert, dan Szymczak (2017), Sanchez-Pinto, Venable, Fahrenbach, dan Churpek (2018) menilai metode pemilihan variabel untuk klasifikasi hutan acak, tetapi sebagian besar makalah ini hanya membandingkan beberapa metode. Selain itu, makalah-makalah ini memiliki cakupan yang terbatas karena penggunaan data simulasi sintetis yang tidak selalu mewakili dataset dunia nyata atau sejumlah kecil dataset aplikasi. Keterbatasan terakhir dari literatur pemilihan variabel random forest saat ini adalah waktu komputasi untuk prosedur yang berbeda jarang dilaporkan. Dengan adanya keterbatasan ini, ada kebutuhan untuk membandingkan prosedur pemilihan variabel untuk sejumlah besar masalah klasifikasi hutan acak untuk memberikan rekomendasi tentang prosedur mana yang sesuai untuk berbagai jenis dataset.

Sisa dari makalah ini disusun sebagai berikut. Bagian 2 merangkum metode dan implementasi untuk pemilihan variabel hutan yang dijalankan untuk klasifikasi dalam literatur saat ini. Desain dari penelitian saat ini disajikan pada Bagian 3, termasuk dataset yang digunakan dan metrik evaluasi untuk prosedur pemilihan variabel. Bagian 4 menyajikan ringkasan hasil yang membandingkan tingkat kesalahan, parsimoni, dan waktu komputasi untuk prosedur pemilihan variabel. Diskusi dan kesimpulan disajikan di Bagian 5.

2. Metode untuk pemilihan variabel random forest untuk klasifikasi

Metode pemilihan variabel untuk klasifikasi hutan acak dijelaskan secara menyeluruh dalam literatur (misalnya Cadenas dkk., 2013; Cano dkk., 2017; Degenhardt dkk., 2017; Hapfelmeier & Ulm, 2013; Sanchez-Pinto dkk., 2018); oleh karena itu, demi kepentingan ringkas, kami hanya meringkas ide utama dari setiap metode dan memberikan pengaturan parameter yang digunakan dalam penelitian ini. Secara kronologis berdasarkan tahun publikasi, metode-metode yang kami bandingkan disajikan pada Tabel 1. Metode yang menggunakan pendekatan eliminasi mundur dengan conditional inference forest antara lain Jiang dkk. (2004), Svetnik dkk. (2004), dan metode Hapfelmeier (2013). Beberapa metode menggunakan prosedur eliminasi mundur dengan implementasi standar dari random forest, termasuk *varSelRF* (Díaz-Uriarte & De Andres, 2006), *caret* (Kuhn, 2008), dan random-ForestSRC (Ishwaran & Kogalur, 2014). Prosedur seleksi bertahap diimplementasikan dalam *VSURF* (Genuer et al., 2015), sedangkan *RRF* (Deng & Runger, 2013) menggunakan prosedur hutan acak yang teregulasi dan pendekatan seleksi ke depan. Altmann dkk. (2010), *Boruta* (Kursa & Rudnicki, 2010) dan Janitza dkk. (2015) menggunakan ukuran kepentingan acak untuk melakukan seleksi variabel. Serupa dengan kategorisasi Hapfelmeier

Tabel 1
Ringkasan metode pemilihan variabel untuk klasifikasi hutan acak.

Singkatan di kertas	Publikasi	Paket/implementasi R	Pendekatan	Jenis metode hutan	Ringkasan	Pengaturan parameter
RF	Breiman, 2001	*randomForest*	N/A	Hutan acak	Tidak ada pemilihan variabel	Default
RFtuned	Breiman, 2001	*randomForest*	N/A	Hutan acak	Tidak ada pemilihan variabel, disetel dengan fungsi *tuneRF()*	Default
Svetnik	Svetnik et al, 2004	Menggunakan *pesta,* kode dari Hapfelmeier	Kinerja Berbasis	Bersyarat Hutan Inferensi	Menggunakan eliminasi berbasis eliminasi mundur pada ukuran kepentingan dan k-lipat validasi	# jumlah pohon = 100, jumlah lipatan = 5, # pengulangan = 20
Jiang	Jiang et al, 2004	Menggunakan *pesta,* kode dari Hapfelmeier	Kinerja Berbasis	Bersyarat Hutan Inferensi	Mirip dengan Svetnik tetapi menyediakan mekanisme untuk mencegah pemasangan yang berlebihan	# pohon = 1000
varSelRF	Diaz Uriarte, 2007	*varSelRF*	Kinerja Berbasis	Hutan acak	Menggunakan eliminasi mundur, kriteria untuk menghapus variabel berdasarkan mempertahankan tingkat kesalahan yang sama dengan model lengkap	Default
Caret	Kuhn, 2008	*caret*	Kinerja Berbasis	Hutan acak	Menggunakan eliminasi fitur rekursif, kriteria untuk menghapus variabel berdasarkan untuk mempertahankan tingkat kesalahan yang sama dengan model lengkap	Default
Altmann	Altmann et al, 2010	*vita*	Berdasarkan Tes	Hutan acak	Berdasarkan uji parametrik terhadap permutasi berulang dari langkah-langkah penting	Default
Boruta	Kursa 2010	*Boruta*	Berdasarkan Tes	Hutan acak	Berdasarkan uji permutasi menggunakan pendekatan yang mengedepankan kepentingan tindakan	Default
Hapfelmeier	Hapfelmeier 2013	Menggunakan *pesta,* kode dari Hapfelmeier	Berdasarkan Tes	Bersyarat Hutan Inferensi	Mirip dengan Altmann, tetapi menggunakan ukuran kepentingan yang tidak bias	# permutasi = 100, # trees=100, alpha=0.05
RRF	Deng 2013	*RRF*	Kinerja Berbasis	Hutan Acak	Berdasarkan acak yang teratur hutan, yang menggunakan forward pilihan untuk menambahkan secara berurutan variabel sampai tidak ada lagi perolehan informasi	Default
SRC	Ishwaran 2014	*randomForestSRC*	Kinerja Berbasis	Hutan Acak	Menggunakan eliminasi berbasis eliminasi mundur pada kedalaman prediktor yang minimal	Default
VSURF	Genuer et al, 2015	*VSURF*	Kinerja Berbasis	Hutan Acak	Prosedur pemilihan bertahap yang mengimplementasikan eliminasi mundur kemudian meneruskan seleksi berdasarkan langkah-langkah penting dan kesalahan nilai	Default
Janitza	Janitza et al, 2015	*Vita*	Berdasarkan Tes	Hutan acak	Mirip dengan Altmann, tetapi juga menggunakan validasi silang	Default

metode sebagai berbasis pengujian atau berbasis kinerja. Pendekatan berbasis kinerja memilih variabel berdasarkan perubahan dalam akurasi pra-diksi ketika variabel ditambahkan atau dihapus dari model, dan termasuk metode oleh Svetnik dan Jiang, varSelRF, caret, RRF, SRC, dan VSURF. Pendekatan berbasis uji memilih variabel berdasarkan uji statistik atau permutasi, dan mencakup metode oleh Altmann, Hapfelmeier dan Janitza, serta Boruta.

Selain metode-metode tersebut, kami mempertimbangkan untuk menggunakan dua metode lainnya tetapi akhirnya memutuskan untuk tidak menggunakannya dalam penelitian ini. Metode tersebut adalah metode eliminasi mundur berdasarkan area di bawah kurva operasi penerima yang diimplementasikan dalam paket *AUCRF* R oleh Urrea (Urrea & Calle, 2012), yang tidak disertakan karena terbatas pada hasil biner, dan metode pemilihan variabel dengan adanya variabel berkorelasi yang diimplementasikan dalam paket *fuzzy-Forest* R oleh Conn (Conn dkk, 2015), yang tidak disertakan karena memerlukan spesifikasi struktur korelasi oleh pengguna. Tujuan kami adalah untuk menjadi seinklusif mungkin dalam hal menggunakan semua metode pemilihan variabel yang tersedia untuk klasifikasi hutan acak untuk mengevaluasi dan membandingkan metode secara menyeluruh.

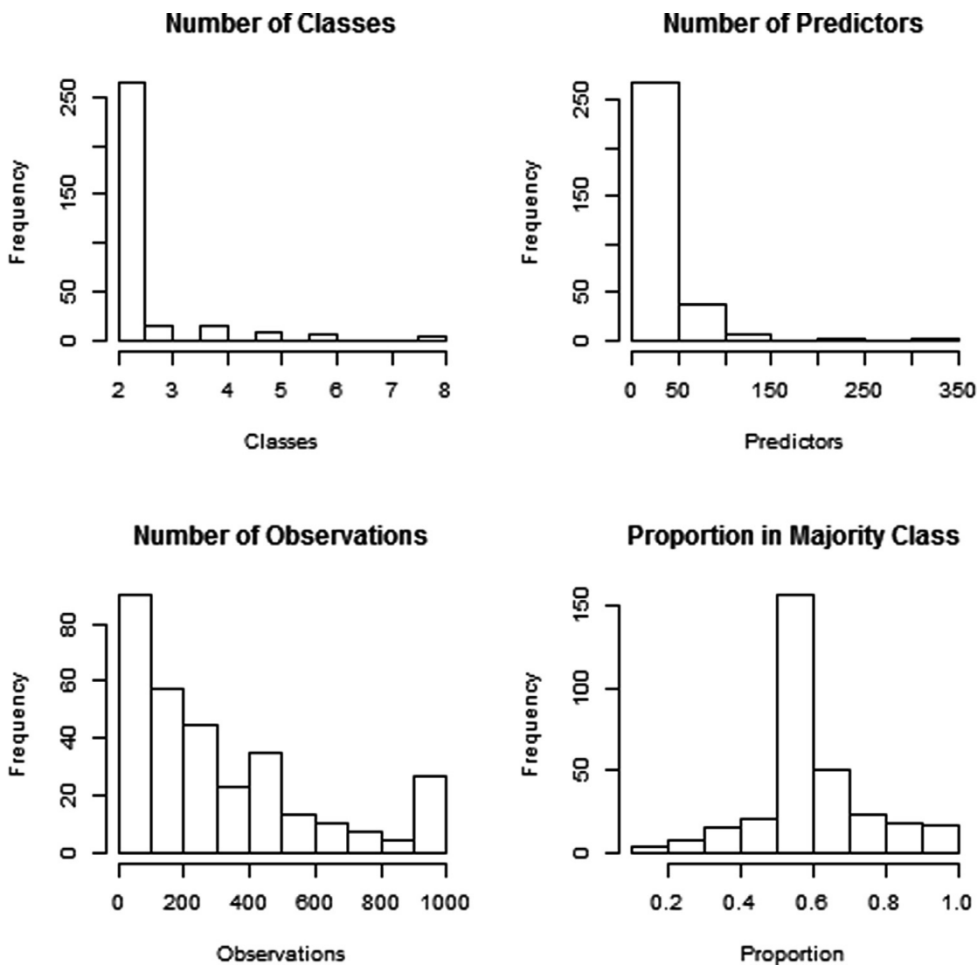
3. Desain studi

Kami menganalisis set data yang tersedia secara bebas di perangkat lunak R menggunakan paket *OpenML* R (situs web: https://www.openml.org/home) (Casalicchio et al., 2017). Kami menggunakan batasan berikut untuk memilih dataset untuk penelitian ini: pertanyaan penelitian untuk dataset ditetapkan sebagai klasifikasi yang diawasi (yaitu tugas klasifikasi yang diawasi) dan tidak ada prediktor yang hilang.

tor atau nilai hasil dalam set data. Karena keterbatasan komputasi yang terkait dengan beberapa metode, kami membatasi dataset pada dataset dengan hasil yang memiliki 8 level atau kurang (kategori dalam variabel hasil), 1000 prediktor atau kurang, dan 1000 pengamatan atau kurang. Kami membatasi ukuran dataset karena penelitian sebelumnya oleh Degenhardt dkk. (2017) menyelidiki pemilihan variabel random forest untuk dataset besar di bidang omics, sehingga kami memutuskan untuk mengecualikannya dari penelitian kami. Kami juga menghilangkan dataset di mana random forest standar menghasilkan pesan kesalahan menggunakan R pack- age *randomForest*, meskipun hal ini cukup jarang terjadi (2 dataset).

Secara keseluruhan dengan spesifikasi ini, kami menganalisis 311 set data total, yang memiliki rata-rata (standar deviasi (SD)) jumlah kelas hasil 2,4 (1,0). Terdapat rata-rata (SD) 22 (33) prediktor dalam dataset, dengan rata-rata (SD) 322 (285) observasi. Gbr. 1 menampilkan distribusi jumlah kelas hasil, prediktor, dan observasi untuk dataset, serta proporsi observasi di kelas hasil mayoritas. Rincian tentang karakteristik dataset yang digunakan dalam penelitian ini disertakan dalam Supplementary Dataset Listing File. Dataset berasal dari berbagai bidang aplikasi sistem pakar, termasuk kedokteran, bisnis, ilmu lingkungan, ilmu dasar, pertanian, psikologi, pendidikan, ilmu komputer, dan nutrisi.

Untuk setiap set data yang dijelaskan di atas, kami melakukan seleksi variasi menggunakan metode yang tercantum dalam Tabel 1 dan mengembangkan model hutan acak standar menggunakan paket R *randomFor- est* (Liaw & Weiner, 2002) dan hutan acak dengan parameter *mtry* yang disetel untuk perbandingan. Metode-metode pada Tabel 1 yang tersedia dalam paket R digunakan dengan nilai default. Sisanya



Gambar 1. Gambar ini menampilkan karakteristik dataset yang digunakan untuk penelitian, termasuk jumlah kelas hasil, jumlah variabel prediktor, jumlah observasi, dan proporsi kelas hasil mayoritas.

Metode yang menggunakan *paket* R diimplementasikan berdasarkan kode yang disediakan oleh Hapfelmeier (Hapfelmeier & Ulm, 2013). Untuk setiap metode pemilihan variabel dan dataset, kami mencatat tingkat kesalahan prediksi (didefinisikan sebagai proporsi prediksi yang salah untuk data yang tidak sesuai), jumlah variabel yang digunakan, waktu komputasi, dan AUC. Untuk mendapatkan estimasi AUC, kami menggunakan paket R *multiROC* (Wei, Wang, & Jia, 2019). Kami menggunakan versi R 3.4.1 pada komputer dengan Intel® Core™ i1-7700 CPU 3.60 GHz, 3600 Mhz, 4 Core, 8 Logical Processor dan 16.0GB RAM. Kode yang digunakan untuk mengimplementasikan dan mengevaluasi metode pemilihan variabel disediakan dalam File Kode Tambahan.

4. R esults

4.1. Hasil untuk semua set data

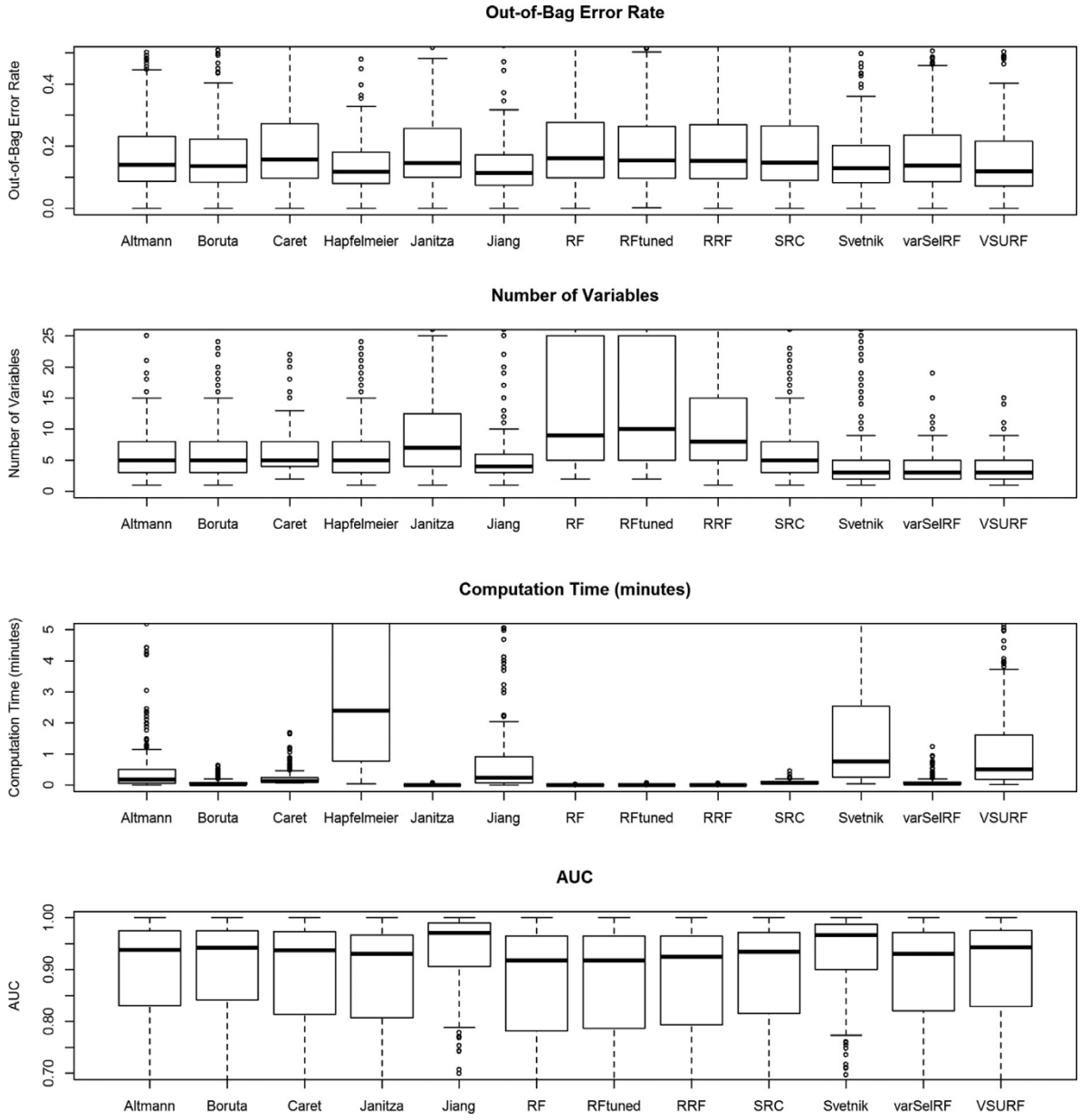
Kami membandingkan distribusi tingkat kesalahan di luar kantong, jumlah variabel yang termasuk dalam model yang direduksi, waktu komutasi dan AUC dari prosedur pemilihan variabel pada Gbr. 2 dan Tabel 2. Rata-rata tingkat kesalahan di luar kantong berkisar antara 16% hingga 23% di seluruh metode pemilihan variabel dan semuanya memiliki distribusi miring ke kanan. Tiga metode teratas dengan tingkat kesalahan di luar kantong rata-rata terendah adalah metode VSURF, metode Boruta, dan metode Altmann. Beberapa model memiliki tingkat kesalahan di luar kantong yang cukup besar (lebih besar dari 0,5) untuk beberapa dataset, yang mengindikasikan bahwa model-model ini tidak memprediksi hasil dengan baik.

Distribusi jumlah variabel yang termasuk dalam model juga miring ke kanan. Sebagian besar metode memiliki median jumlah variabel sekitar tiga hingga sebelas. Metode VSURF menghasilkan jumlah variabel rata-rata terendah (3,4), diikuti oleh metode varSelRF (3,9) dan metode Jiang (5,1). Standard ran- dom forest memiliki rata-rata 21,4 variabel dan tuned random forest memiliki rata-rata 21,8 variabel, yang merupakan yang terbesar karena tidak ada seleksi variabel yang dilakukan.

Hutan acak standar memiliki waktu komputasi terendah, yang serupa dengan RRF dan tuned random for- est. Metode-metode dengan waktu komputasi terbesar (rata-rata lebih dari satu menit) adalah metode Jiang (1,3 menit), VSURF (2,9 menit), metode Svetnik (3,2 menit), dan metode Hapfelmeier (16,2 menit). Metode-metode ini juga memiliki variabilitas terbesar dalam waktu komputasi, mulai dari beberapa detik hingga beberapa menit. Metode oleh Hapfelmeier memiliki waktu komputasi tertinggi dengan selisih yang besar.

Nilai AUC untuk model-model tersebut berkisar antara 0,865 hingga 0,929, yang mengindikasikan bahwa sebagian besar model menawarkan kecocokan yang baik. Metode Jiang dan metode Svetnik memiliki rata-rata AUC yang sedikit lebih tinggi dibandingkan dengan model lainnya. Kami menghilangkan metode Hapfelmeier dari hasil AUC karena waktu komputasinya yang besar.

Perlu dicatat bahwa beberapa metode untuk dataset tertentu menghasilkan pesan kesalahan atau tidak memberikan model fisual. Tabel 2 berisi jumlah dataset (*N*) yang digunakan untuk masing-masing metode. Metode SRC adalah satu-satunya metode yang memberikan prediksi untuk setiap dataset. Sebagian besar metode memiliki kurang dari dua puluh

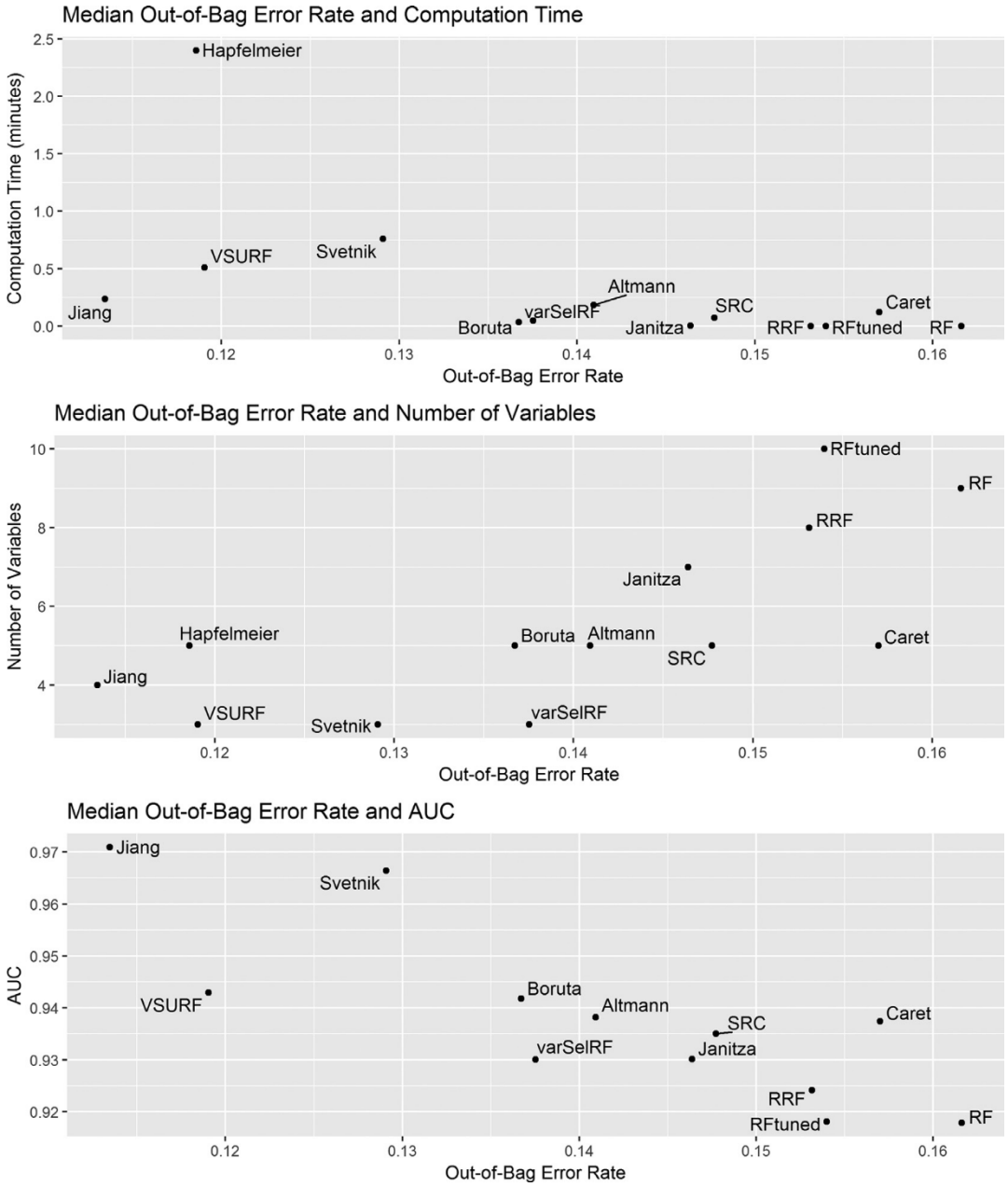


Gbr. 2. Gambar ini menampilkan boxplot dari tingkat kesalahan di luar kantong, jumlah variabel, waktu komputasi (menit) dan AUC untuk metode pemilihan variabel untuk klasifikasi hutan acak.

Tabel 2
Distribusi tingkat kesalahan, waktu komputasi dan jumlah variabel yang digunakan untuk prosedur pemilihan variabel.

Tingkat kesalahan OOB		Jumlah variabel		Waktu komputasi (menit)		AUC	
Model (*N*)	Rata-rata (SD)	Model (*N*)	Rata-rata (SD)	Model (*N*)	Rata-rata (SD)	Model (*N*)	Rata-rata (SD)
VSURF (251)	0.156 (0.136)	VSURF (251)	3.442 (2.312)	RF (308)	0.003 (0.005)	Jiang (291)	0.929 (0.106)
Boruta (292)	0.17 (0.142)	VarSelRF (308)	3.877 (3.703)	RRF (297)	0.004 (0.008)	Svetnik (292)	0.924 (0.103)
Altmann (297)	0.179 (0.148)	Jiang (291)	5.131 (5.405)	RF yang disetel (301)	0.004 (0.009)	Boruta (292)	0.89 (0.133)
VarSelRF (308)	0.18 (0.148)	Svetnik (292)	6.182 (8.479)	Janitza (183)	0.007 (0.01)	Caret (298)	0.888 (0.133)
SRC (311)	0.188 (0.154)	SRC (311)	7.19 (8.089)	Boruta (292)	0.061 (0.098)	Janitza (183)	0.888 (0.13)
Janitza (183)	0.196 (0.147)	Boruta (292)	7.86 (9.777)	SRC (311)	0.089 (0.054)	VarSelRF (308)	0.887 (0.13)
RF yang disetel (301)	0.196 (0.151)	Hapfelmeier (290)	8.021 (10.05)	VarSelRF (308)	0.106 (0.165)	SRC (311)	0.885 (0.137)
RRF (297)	0.196 (0.154)	Altmann (297)	9.421 (18.898)	Caret (298)	0.225 (0.264)	Altmann (297)	0.883 (0.147)
Caret (298)	0.2 (0.157)	RRF (297)	10.603 (7.95)	Altmann (297)	0.554 (0.971)	VSURF (251)	0.881 (0.138)
RF (308)	0.203 (0.158)	Caret (298)	11.718 (25.606)	Jiang (291)	1.314 (3.396)	RRF (297)	0.873 (0.13)
Jiang (291)	0.22 (0.488)	Janitza (183)	14.914 (27.864)	VSURF (251)	1.754 (2.937)	RF yang disetel (301)	0.87 (0.131)
Hapfelmeier (290)	0.23 (0.503)	RF (308)	21.418 (32.653)	Svetnik (292)	3.2 (7.794)	RF (308)	0.865 (0.133)
Svetnik (292)	0.232 (0.41)	RF yang disetel (301)	21.77 (32.942)	Hapfelmeier (290)	16.16 (42.77)	Hapfelmeier	NA

N: Jumlah set data yang dikompilasi untuk model. SD:
Standar deviasi.
OOB: Di luar kantong.



Gbr. 3. Gambar ini menampilkan plot median tingkat kesalahan di luar kantong berdasarkan median waktu komputasi, median jumlah variabel, dan median AUC.

model prediksi yang hilang. Metode Janitza memiliki pra-dikte yang paling banyak dihilangkan untuk penelitian kami karena tidak menyediakan pra-dikte yang dipilih untuk 128 set data, sehingga menghasilkan nilai yang hilang untuk kesalahan di luar kantong, jumlah variabel, waktu komputasi, dan AUC.

Tingkat kesalahan di luar kantong rata-rata diplot berdasarkan waktu komutasi rata-rata, jumlah variabel rata-rata, dan AUC rata-rata dalam Gbr. 3. Metode Jiang, metode Hapfelmeier, dan metode VSURF memiliki tingkat kesalahan di luar kantong dan jumlah variabel yang rendah; namun, metode Hapfelmeier memiliki waktu komputasi yang jauh lebih tinggi dibandingkan dengan metode lainnya. Metode VSURF memiliki waktu komputasi yang sedikit lebih tinggi dibandingkan dengan metode Jiang. Metode Jiang menghasilkan AUC tertinggi, sedangkan metode VSURF juga menawarkan AUC yang baik dibandingkan dengan sebagian besar metode lainnya. Metode Boruta, metode varSelRF, dan metode Altmann untuk pemilihan variabel memiliki tingkat kesalahan di luar kantong yang cukup rendah, waktu komputasi yang rendah, dan jumlah variabel yang moderat. Hutan run- dom standar (tidak ada pemilihan variabel yang dilakukan) memiliki waktu komputasi terendah.

Waktu komputasi dan jumlah variabel tertinggi, sedangkan hutan acak yang disetel tanpa pemilihan variabel juga memiliki jumlah variabel tertinggi dengan tingkat kesalahan yang sedikit lebih rendah. Median waktu komutasi, tingkat kesalahan di luar kantong, jumlah variabel dan AUC serupa untuk hutan acak standar, hutan acak yang disetel dan RRF. Tidak ada kelompok metode yang jelas dalam hal membandingkan tingkat kesalahan dan waktu komputasi, tingkat kesalahan dan jumlah variabel atau tingkat kesalahan dan AUC.

Karena beberapa metode tidak dapat dikompilasi untuk beberapa dataset, kami juga menyelidiki kinerja metode berdasarkan dataset yang dapat dikompilasi oleh semua metode. Tabel 3 menampilkan distribusi tingkat kesalahan di luar kantong, jumlah variabel yang dimasukkan ke dalam model yang direduksi, waktu komputasi dan AUC dari prosedur pemilihan variabel yang tidak termasuk semua dataset yang setidaknya salah satu metodenya tidak memberikan model akhir. Terdapat 141 set data yang digunakan dalam analisis ini. Meskipun ada 170 dataset yang memiliki setidaknya satu prosedur pemilihan variabel yang tidak

Tabel 3
Distribusi tingkat kesalahan, waktu komputasi dan jumlah variabel yang digunakan untuk prosedur pemilihan variabel tidak termasuk semua dataset yang setidaknya salah satu metodenya tidak menghasilkan model akhir (141 dataset yang digunakan dalam analisis ini).

Tingkat kesalahan OOB		Jumlah variabel		Waktu komputasi (menit)		AUC	
Model	Rata-rata (SD)	Model	Rata-rata (SD)	Model	Rata-rata (SD)	Model	Rata-rata (SD)
VSURF	0.155 (0.113)	VSURF	3.489 (1.999)	RF	0.003 (0.005)	Jiang	0.929 (0.106)
VarSelRF	0.163 (0.111)	VarSelRF	3.972 (2.775)	RRF	0.004 (0.005)	Svetnik	0.924 (0.103)
Boruta	0.166 (0.115)	Jiang	5.014 (4.559)	RF disetel	0.005 (0.007)	Boruta	0.89 (0.133)
Altmann	0.172 (0.12)	Svetnik	6.823 (9.546)	Janitza	0.008 (0.01)	Caret	0.888 (0.133)
Janitza	0.174 (0.115)	SRC	8.532 (9.917)	Boruta	0.082 (0.1)	Janitza	0.888 (0.13)
SRC	0.181 (0.114)	Boruta	8.823 (10.603)	SRC	0.089 (0.051)	varSelRF	0.887 (0.13)
RRF	0.185 (0.116)	Hapfelmeier	9.206 (10.608)	VarSelRF	0.139 (0.185)	SRC	0.885 (0.137)
RF disetel	0.187 (0.117)	Altmann	12.894 (24.502)	Caret	0.264 (0.276)	Altmann	0.883 (0.147)
Caret	0.191 (0.121)	RRF	14.248 (6.905)	Altmann	0.67 (1.063)	VSURF	0.881 (0.138)
Jiang	0.193 (0.316)	Caret	16.099 (35.16)	VSURF	1.76 (2.401)	RRF	0.873 (0.13)
RF	0.196 (0.115)	Janitza	17.319 (30.93)	Jiang	1.944 (4.325)	RFtuned	0.87 (0.131)
Hapfelmeier	0.204 (0.322)	RF disetel	34.22 (41.728)	Svetnik	5.008 (10.413)	RF	0.865 (0.133)
Svetnik	0.23 (0.366)	RF	34.22 (41.728)	Hapfelmeier	25.179 (55.017)	Hapfelmeier	NA

N: Jumlah set data yang dikompilasi untuk model. SD: Standar deviasi.
OOB: Di luar kantong.

Dalam menghasilkan model akhir, hasil untuk kesalahan di luar kantong, jumlah variabel , waktu komputasi dan AUC cukup mirip dengan Tabel 2 (yaitu analisis yang menyertakan semua hasil yang tidak hilang untuk prosedur pemilihan variabel). Perbedaan utamanya adalah bahwa metode dengan tingkat kesalahan terkecil adalah VSURF, varSelRF, Boruta, kemudian Altmann dalam analisis tanpa data yang hilang (Tabel 3), tetapi urutannya adalah VSURF, Boruta, Altmann, kemudian varSelRF dalam analisis termasuk data yang hilang (Tabel 2). Namun, tingkat kesalahan di seluruh metode ini cukup mirip.

4.2. *Membandingkan metode yang dikelompokkan berdasarkan karakteristik set data*

4.2.1. *Hasil untuk set data dengan hasil biner*

Kami juga menganalisis hasil untuk subset dataset yang berisi hasil biner, atau dua kelas, (Tabel Tambahan 1 dan Gambar Tambahan 1). Terdapat 264 set data dengan hasil biner. Hasilnya serupa dengan yang disajikan pada Bagian 4.1 untuk semua dataset. Metode Hepfelmeier, metode Jiang, metode VSURF, dan metode Svetnik memiliki waktu komputasi yang tinggi dan tingkat kesalahan di luar kantong yang rendah. Metode Boruta, metode Altman, dan metode varSelRF memiliki tingkat kesalahan out-of-bag dan waktu komputasi yang cukup rendah, dan varSelRF menawarkan jumlah variabel yang paling rendah di antara metode-metode ini.

4.2.2. *Hasil untuk set data dengan banyak variabel prediktor*

Analisis dilakukan untuk subset set data dengan banyak prediktor, yang kami definisikan sebagai lebih dari 50 variabel prediktor (Tabel Tambahan 2 dan Gambar Tambahan 2). Terdapat 44 set data dengan lebih dari 50 prediktor. Serupa dengan hasil sebelumnya, metode Hapfelmeier sejauh ini memiliki waktu komputasi yang paling tinggi dibandingkan dengan metode lainnya. Boruta dan varSelRF memiliki tingkat kesalahan di luar kantong yang cukup rendah, waktu komputasi yang rendah, dan jumlah variabel yang dimasukkan juga rendah. VSURF memiliki tingkat kesalahan dan jumlah variabel yang paling rendah, tetapi hal ini dibarengi dengan waktu komputasi rata-rata yang relatif tinggi, yaitu sekitar empat menit.

4.2.3. *Hasil untuk set data dengan hasil yang tidak seimbang*

Analisis akhir dilakukan untuk subset dataset dengan hasil yang tidak seimbang, yang kami definisikan sebagai kelas hasil mayoritas yang mengandung lebih dari 60% pengamatan (Tabel Tambahan 3 dan Gambar Tambahan 3). Terdapat 107 set data yang memiliki hasil yang tidak seimbang. Sekali lagi, metode Hapfelmeier memiliki rata-rata waktu komputasi tertinggi. Tingkat kesalahan di luar kantong untuk metode ini berada dalam kisaran yang lebih ketat dibandingkan dengan analisis lainnya, yang berkisar antara 12% hingga 16% secara rata-rata. Sementara metode Jiang

memiliki waktu komputasi yang sedikit lebih tinggi dibandingkan dengan metode lainnya, metode ini menawarkan jumlah variabel yang paling rendah.

4.3. Membandingkan metode yang dikelompokkan berdasarkan karakteristik metode

4.3.1. Hasil untuk membandingkan metode hutan acak standar dan hutan acak bersyarat

Jenis hutan acak yang diterapkan berbeda di antara metode-metode seleksi yang ada (Tabel 1). Mayoritas metode menggunakan hutan acak standar; ini termasuk varSelRF, Caret, metode Altmann, Boruta, RRF, SRC, VSURF, dan metode Janitza. Prosedur pemilihan variabel lainnya menggunakan hutan acak bersyarat, termasuk metode Svetnik, Jiang, dan Hapfelmeier. Secara umum, metode yang menggunakan conditional random forest memiliki waktu komputasi yang lebih tinggi dan tingkat kesalahan yang lebih rendah dibandingkan dengan metode yang menggunakan standard random forest (Gbr. 3). Pengecualian untuk hal ini adalah VSURF, yang memiliki waktu komputasi yang sedikit lebih tinggi dan tingkat kesalahan yang lebih rendah dibandingkan dengan metode lain yang juga menggunakan hutan acak standar.

4.3.2. Hasil untuk membandingkan metode berbasis tes dan berbasis kinerja

Prosedur pemilihan variabel random forest juga berbeda dalam hal pendekatannya, apakah berdasarkan uji permutasi atau kinerja (akurasi). Metode yang memilih variabel berdasarkan uji permutasi termasuk metode Altmann, Boruta, metode Hapfelmeier dan metode Janitza, sedangkan metode yang memilih variabel berdasarkan kinerja dalam hal akurasi termasuk metode Svetnik, metode Jiang, varSelRF, Caret, RRF, SRC, dan VSURF. Ketika mengelompokkan berdasarkan metode pengujian atau metode berdasarkan kinerja, tidak ada pola yang terlihat dalam hal tingkat kesalahan, waktu komputasi, jumlah variabel yang dipilih atau AUC.

5. Diskusi

Dalam makalah ini, kami memberikan perbandingan metode yang tersedia untuk pemilihan variabel random forest dalam pengaturan klasifikasi dengan menggunakan 311 set data yang tersedia secara online. Metode dengan tingkat kesalahan di luar kantong terendah, serta waktu komputasi dan jumlah variabel terendah lebih disukai. Yang paling penting adalah nilai AUC yang tinggi. Secara keseluruhan, metode oleh Jiang dan VSURF memiliki jumlah kesalahan terendah dan parsimoni terbaik (jumlah variabel paling sedikit), tetapi hal ini juga dibarengi dengan waktu komputasi yang lebih tinggi. Alasan mengapa metode Jiang memiliki waktu komputasi yang tinggi adalah karena metode ini menggunakan validasi k-fold untuk memilih variabel, yang

dapat menjadi mahal secara komputasi. Metode VSURF kemungkinan memiliki waktu komputasi yang tinggi karena menggunakan prosedur bertahap untuk memilih variabel, di mana variabel dieliminasi dan kemudian ditambahkan kembali, yang mungkin terkait dengan waktu komputasi yang lebih tinggi. Pada setiap analisis, metode Hapfelmeier memiliki waktu komputasi tertinggi, yang mungkin terjadi karena metode ini harus mengumpulkan permutasi untuk setiap variabel yang membutuhkan biaya komputasi yang tinggi. Metode Althann, varSelRF, dan Boruta secara umum memiliki kinerja yang serupa untuk berbagai jenis dataset, dengan waktu komputasi yang rendah, tingkat kesalahan yang cukup rendah, dan keseragaman yang sedang hingga baik. Hal ini secara umum serupa ketika menganalisis subset dataset yang memiliki hasil biner. Waktu komputasi cukup baik untuk model dengan hasil biner, dengan median satu menit atau kurang untuk sebagian besar model, kecuali untuk metode Hapfelmeier dan metode Svetnik. Untuk set data dengan jumlah variabel prediktor yang lebih banyak, VSURF memiliki waktu komputasi rata-rata sekitar lima menit, sedangkan varSelRF dan Boruta memiliki waktu komputasi yang jauh lebih rendah dengan tingkat kesalahan yang sedikit lebih tinggi. Waktu komputasi yang lebih rendah dari varSelRF dan Boruta kemungkinan besar disebabkan oleh prosedur eliminasi mundur, yang dilakukan lebih cepat daripada pemilihan bertahap atau pendekatan pemilihan validasi k-lipatan yang digunakan oleh VSURF. Meskipun VSURF menawarkan tingkat kesalahan median terendah untuk dataset, VSURF juga tidak memberikan model (yaitu tidak memilih variabel) untuk 60 dataset. Metode Janitza, bagaimanapun, tidak memberikan model untuk 128 dataset, yang sejauh ini merupakan yang terburuk dalam hal kemampuannya untuk diterapkan ke berbagai dataset. Ringkasan kelebihan dan kekurangan metode yang digunakan dalam penelitian kami disajikan dalam Tabel Tambahan 4.

Sangat menarik bahwa tingkat kesalahan dari banyak metode serupa (rata-rata berkisar antara 16 hingga 23% secara keseluruhan), sementara parsimoni, waktu komputasi, dan AUC sangat berbeda jika dibandingkan dengan metode yang berbeda. Perbedaan kecil dalam tingkat kesalahan mungkin disebabkan oleh banyak metode yang serupa, atau bahkan bersarang di dalam satu sama lain dengan menyertakan penyesuaian pada algoritme untuk memilih variasi. Meskipun kami mengharapkan metode-metode yang serupa untuk mengelompok dalam kelompok (misalnya, pada plot Gbr. 3), hal ini tidak terjadi pada analisis keseluruhan ketika membandingkan metode berbasis tes dan berbasis kinerja. Hal ini menunjukkan bahwa jenis metode (tes atau kinerja) tidak secara berbeda berdampak pada kinerja keseluruhan metode pemilihan variabel dalam hal tingkat kesalahan, waktu komputasi, parsimoni, dan AUC. Namun, metode yang menggunakan hutan acak konvensiona- l biasanya berperilaku serupa: metode ini cenderung memiliki tingkat kesalahan yang lebih rendah dibandingkan dengan metode hutan acak standar, meskipun disertai dengan waktu komputasi yang jauh lebih tinggi. Metode untuk pemilihan variabel yang menggunakan conditional random forest mungkin lebih baik untuk dataset dengan asosiasi yang mendasari antara prediktor dan hasil yang diketahui karena conditional random forest biasanya lebih baik dalam mengidentifikasi dengan benar hubungan yang signifikan antara prediktor dan hasil dibandingkan dengan standard random forest. Untuk dataset yang tidak memiliki banyak informasi sebelumnya tentang prediktor hasil yang diketahui, metode pemilihan variabel hutan acak standar mungkin lebih disukai karena berfokus pada pengoptimalan akurasi daripada mengidentifikasi hubungan prediktor-hasil yang benar.

Hasil penelitian kami harus dipertimbangkan dalam konteks yang lebih besar dari literatur sebelumnya yang membandingkan metode pemilihan variabel untuk klasifikasi hutan acak. Konsisten dengan temuan dari Sanchez-Pinto dkk. (2018), penelitian kami menemukan bahwa VSURF memiliki kesalahan prediksi yang sedikit lebih baik dibandingkan dengan Boruta dan RRF. Degenhardt dkk. (2017) berfokus pada dataset omics yang cukup besar, tetapi menyimpulkan bahwa Boruta dan Altmann cocok untuk data berdimensi rendah; namun, penelitian ini tidak memasukkan metodologi RRF, Hapfelmeier, Jiang, Svetnik, SRC, atau Caret untuk pemilihan variabel. Studi oleh Cadenas dkk. (2013) menggunakan dua puluh empat dataset untuk membandingkan metode pemilihan variabel random forest, dengan fokus pada data microarray. Mirip dengan penelitian kami,

Mereka menemukan bahwa VSURF memiliki tingkat kesalahan yang paling rendah. Meskipun berfokus pada pengusulan prosedur pemilihan variabel baru, Hapfelmeier (2013) juga membandingkan metode-metode dalam studi simulasi, yang menghasilkan kesalahan yang lebih rendah dibandingkan dengan Altmann, Jiang, varSelRF, Svetnik, dan VSURF. Dalam penelitian kami, metode Hapfelmeier memiliki waktu komputasi yang tinggi, parsimoni yang moderat, dan tingkat kesalahan rata-rata yang lebih tinggi dibandingkan dengan metode lainnya.

Ada beberapa keterbatasan dari penelitian kami yang harus disadari. Pertama, kami hanya menyertakan dataset dengan 1000 observasi atau kurang, sehingga hasil kami mungkin tidak dapat digeneralisasi ke dalam pengaturan data dengan jumlah data yang besar. Kami memasukkan batasan ini karena biaya komputasi-beberapa set data yang lebih besar yang tersedia membutuhkan waktu beberapa jam untuk menyelesaikannya, dan karena penelitian sebelumnya telah membahas masalah pemilihan variabel dalam pengaturan data berdimensi tinggi (Degenhardt et al., 2017). Penelitian di masa depan dapat mengulangi analisis kami dalam pengaturan data dimensi tinggi, mungkin dengan menggunakan komputasi paralel untuk mempercepat waktu pengerjaan. Kedua, kami hanya menyertakan data yang tidak memiliki nilai yang hilang. Akan menarik untuk mengulangi analisis ini dengan data yang diperhitungkan untuk menentukan pengaruh data yang hilang terhadap pemilihan variabel dalam pengaturan klasifikasi hutan acak. Meskipun kami membatasi set data kami dengan cara ini, kami dapat menganalisis 311 set data dengan jumlah prediktor dan observasi yang bervariasi. Oleh karena itu, hasil ini dapat digunakan sebagai panduan untuk memilih jenis prosedur pemilihan variabel mana yang lebih baik tergantung pada jenis hasil (biner atau multi-kelas; seimbang atau tidak seimbang) dan set data (jumlah prediktor yang besar atau kecil).

Ada beberapa jalan untuk penelitian selanjutnya yang berasal dari penelitian ini. Selain melakukan penelitian ini dalam dataset besar seperti yang disarankan di atas, seseorang juga dapat mempertimbangkan untuk memasukkan data yang hilang dalam konteks pemilihan variabel. Dalam penelitian ini, kami hanya menyertakan set data tanpa nilai yang hilang, sehingga akan sangat menarik untuk menilai jumlah data yang hilang dan bagaimana dampak imputasi terhadap pemilihan variabel dalam kerangka kerja random forest. Selain itu, seseorang dapat melakukan penelitian serupa dengan menggunakan hasil kontinu dalam kerangka random forest (penelitian kami hanya berfokus pada hasil kategorikal). Terakhir, teknik pemilihan variabel di luar kerangka random forest dapat ditambahkan sebagai pembanding untuk menentukan apakah metode pemilihan variabel random forest lebih baik daripada metode lainnya.

Kontribusi utama dari penelitian kami adalah kemampuan untuk menilai teknik pemilihan variabel yang berbeda dalam pengaturan klasifikasi hutan acak. Secara khusus, penelitian kami memberikan waktu komputasi untuk model, yang mengatasi kesenjangan penting dalam literatur pemilihan variabel sewa. Berdasarkan hasil kami untuk mengoptimalkan tingkat kesalahan, parsimoni, waktu komputasi dan AUC, kami merekomendasikan penggunaan VSURF atau Jiang untuk dataset yang berisi hasil biner, dataset dengan hasil yang tidak seimbang, dan dataset yang memiliki kurang dari lima puluh prediktor. Kelemahan dari VSURF adalah bahwa VSURF mungkin tidak memilih variabel apa pun untuk model akhir, sehingga metode ini mungkin tidak ideal untuk set data yang berisik (yaitu set data dengan data yang berantakan) atau set data dengan prediktor yang lemah untuk hasilnya. Untuk set data dengan banyak prediktor, kami merekomendasikan penggunaan varSelRF atau Boruta karena lebih efisien secara komputasi dibandingkan dengan metode lain.

Pendanaan

Penelitian ini didukung oleh National Institutes of Health National Center for Advancing Translational Sciences Grant (KL2 TR001421). Sponsor penelitian ini tidak terlibat dalam desain penelitian, analisis, atau penulisan laporan.

Pernyataan minat

Tidak ada.

Materi tambahan

Materi tambahan yang terkait dengan artikel ini dapat ditemukan, dalam versi online, di doi: 10.1016/j.eswa.2019.05.028.

Pernyataan kontribusi kepenulisan kredit

Jaime Lynn Speiser: Konseptualisasi, Analisis formal, Akuisisi dana- ing, Metodologi, Penulisan - draf asli. **Michael E. Miller: Perolehan** dana, Metodologi, Penulisan - draf asli. **Janet Tooze: Perolehan** dana, Metodologi, Penulisan - naskah asli. **Edward IP: Perolehan** dana, Metodologi, Penulisan - draf asli - inal.

Referensi

Altmann, A., Tolos, i, L., Sander, O., & Lengauer, T. (2010). Permutasi impor- tance: Ukuran kepentingan fitur yang dikoreksi. *Bioinformatics, 26*(10), 1340- 1347.

Breiman, L. (2001). Hutan acak. *Machine Learning, 45*(1), 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Klasifikasi dan pohon regresi*. Monterrey, CA: Wadsworth and Brooks.

Cadenas, J. M., Garrido, M. C., & MartíNez, R. (2013). Fitur subset seleksi fil- ter-wrapper berdasarkan data berkualitas rendah. *Sistem Pakar dengan Aplikasi, 40*(16), 6241-6252.

Cano, G., Garcia-Rodriguez, J., Garcia-Garcia, A., Perez-Sanchez, H., Benediktsson, J. A., & Thapa, A. (2017). Pemilihan otomatis deskriptor molekuler menggunakan random forest: Aplikasi untuk penemuan obat. *Sistem Pakar dengan Aplikasi, 72*, 151-159.

Casalicchio, G., Bossek, J., Lang, M., Kirchhoff, D., Kerschke, P., & Hofner, B. (2017). OpenML: Paket R untuk terhubung ke platform pembelajaran mesin OpenML. *Statistika Komputasi*, 1-15.

Conn, D., Ngun, T., Li, G., & Ramirez, C. (2015). Hutan kabur: Memperluas hutan acak untuk berkorelasi. *Data Dimensi Tinggi*.

Degenhardt, F., Seifert, S., & Szymczak, S. (2017). Evaluasi metode pemilihan variabel untuk hutan acak dan kumpulan data omics. *Briefings in Bioinformatics, 20*(2), 492-503.

Deng, H., & Runger, G. (2013). Seleksi gen dengan for- masi acak teratur terpandu. *Pattern Recognition, 46*(12), 3483-3489.

Díaz-Uriarte, R., & De Andres, SA (2006). Seleksi gen dan klasifikasi data microarray menggunakan hutan acak. *BMC Bioinformatics, 7*, 3.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Apakah kita membutuhkan ratusan pengklasifikasi untuk memecahkan masalah klasifikasi dunia nyata. *Jurnal Penelitian Pembelajaran Ma- chine, 15*, 3133-3181.

GenuerGenuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2015). VSURF: Paket R untuk pemilihan variabel menggunakan hutan acak. *The R Journal, 7*(2), 19-33.

Hapfelmeier, A., & Ulm, K. (2013). Pendekatan pemilihan variabel baru menggunakan hutan acak. *Statistika Komputasi & Analisis Data, 60*, 50-69.

Hothorn, T., Hornik, K., Strobl, C., & Zeileis, A. (2010). *Paket R: Pesta: Laboratorium untuk partisi rekursif*.

Ishwaran, H., & Kogalur, U. (2014). *Random forests untuk kelangsungan hidup, regresi dan* pengelompokan *(RF-SRC)*. Paket R versi 1.6 http://CRANR-project.org/package= randomForestSRC.

Janitza, S., Celik, E., & Boulesteix, A.-L. (2015). Uji signifikansi variabel yang cepat secara komputasi untuk hutan acak untuk data berdimensi tinggi. *Kemajuan dalam Analisis dan Klasifikasi Data*, 1-31.

Jiang, H., Deng, Y., Chen, H.-S., Tao, L., Sha, Q., & Chen, J. (2004). Analisis gabungan dari dua set data ekspresi gen microarray untuk memilih gen penanda adenokarsinoma paru. *BMC Bioinformatika, 5*, 81.

Kuhn, M. (2008). Paket Caret. *Jurnal Perangkat Lunak Statistik, 28*, 1-26.

Kursa, M. B., & Rudnicki, W. R. (2010). Pemilihan fitur dengan paket Boruta. *Jurnal Perangkat Lunak Statistik, 36*, 1-13.

LiawLiaw, A., & Weiner, M. (2002). Klasifikasi dan Regresi dengan randomForest. *R News, 2*, 18-22.

Sanchez-Pinto, LN, Venable, LR, Fahrenbach, J., & Churpek, MM (2018). Perbandingan metode pemilihan variabel untuk pemodelan prediktif klinis. *Jurnal Internasional Informatika Medis, 116*, 10-17.

Speiser, J. L., Durkalski, V. L., & Lee, W. M. (2015). Klasifikasi hutan acak etiologi untuk penyakit yatim piatu. *Statistics in Medicine, 34*(5), 887-899.

Svetnik, V., Liaw, A., Tong, C., & Wang, T. (2004). Penerapan hutan acak Breiman untuk memodelkan hubungan struktur-aktivitas molekul farmasi. In *Lokakarya internasional tentang sistem pengklasifikasi ganda* (pp. 334-343). Springer.

Urrea, V., & Calle, M. L. (2012). *AUCRF: Pemilihan variabel dengan random forest dan area di bawah kurva*. Paket R versi 1.1.

Wei, R., Wang, J., & Jia, W. (2019). *Paket R: MultiROC*.