MEBoost: Mencampur Estimator dengan Boosting untuk Klasifikasi Data Tidak Seimbang

Farshid Rayhan, Sajid Ahmed, Asif Mahbub, Md. Rafsan Jani,
Swakkhar Shatabda, Dewan Md. Farid dan Chowdhury Mofizur Rahman
Jurusan Ilmu Komputer dan Teknik, Universitas Internasional Bersatu, Bangladesh
Surel: frayhan133057@bscse.uiu.ac.bd

Abstrak—Masalah ketidakseimbangan kelas telah menjadi masalah penelitian yang menantang di bidang pembelajaran mesin dan penambangan data karena sebagian besar kumpulan data kehidupan nyata tidak seimbang. Beberapa algoritma pembelajaran mesin yang ada mencoba memaksimalkan klasifikasi akurasi dengan mengidentifikasi sampel kelas mayoritas dengan benar sambil mengabaikan kelas minoritas. Namun, konsep contoh kelas minoritas biasanya lebih menarik daripada kelas mayoritas. Baru-baru ini, beberapa metode yang sensitif terhadap biaya, model ansambel, dan teknik pengambilan sampel telah digunakan dalam literatur untuk mengklasifikasikan kumpulan data yang tidak seimbang. Dalam makalah ini, kami mengusulkan MEBoost, algoritma peningkatan baru untuk kumpulan data yang tidak seimbang. MEBoost memadukan dua pembelajar lemah yang berbeda dengan peningkatan untuk meningkatkan kinerja pada kumpulan data yang tidak seimbang. MEBoost adalah alternatif untuk teknik yang ada seperti SMOTEBoost, RUSBoost, Adaboost, dll. Kinerja MEBoost telah dievaluasi pada 12 kumpulan data tidak seimbang tolok ukur dengan metode ansambel canggih seperti SMOTEBoost, RUSBoost, Easy Ensemble, EUSBoost, DataBoost. Hasil eksperimen menunjukkan peningkatan yang signifikan dibandingkan metode lainnya dan dapat disimpulkan bahwa MEBoost merupakan algoritma yang efektif dan menjanjikan untuk menangani ketidakseimbangan dataset. Versi kode dalam bahasa python tersedia di sini: https://github.com/farshidrayhanuiu/

Kata Kunci—Peningkatan; Ketidakseimbangan kelas; Ensemble; Klasifikasi biner

Aku. AkuPENGANTAR

Dalam pembelajaran terbimbing, pembelajaran mesin adalah proses mengidentifikasi sampel baru atau tidak dikenal yang menggunakan algoritma klasifikasi berdasarkan sekelompok contoh [1], [2], [3], [4]. Dalam dunia nyata, kumpulan data sering kali berdimensi tinggi, multikelas, dan tidak seimbang. Algoritma pembelajaran mesin yang umum sering kali gagal mendapatkan akurasi klasifikasi yang baik pada kumpulan data ini. Ada dua jenis metode untuk menangani kumpulan data yang tidak seimbang: (a) metode internal dan (b) metode eksternal. Metode internal memodifikasi algoritma yang sudah ada sebelumnya untuk mengurangi kepekaannya terhadap rasio ketidakseimbangan kumpulan data. Metode eksternal menerapkan berbagai teknik penyeimbangan data untuk mengurangi rasio ketidakseimbangan kumpulan data.

Terutama ada dua jenis metode sampling untuk memodifikasi distribusi asli dari dataset, yaitu over sampling dan under sampling. Metode under sampling mengurangi instance dari kelas utama berdasarkan intuisi atau hanya secara acak. Beberapa metode under sampling arus utama adalah neighborhood cleaning rule [5], near miss [6], clustered under sampling [7], [8], One sided selection [9]. Metode over sampling bekerja dengan cara yang berlawanan dengan under sampling. Alih-alih menghilangkan instance dari kelas utama, metode ini menghasilkan sampel kelas minoritas menggunakan kelas minoritas itu sendiri menggunakan berbagai teknik seperti AdaSyn [10], SMOTE [11] atau secara acak. Kedua teknik ini memiliki beberapa kekurangan. Metode pengambilan sampel yang kurang berpotensi kehilangan data informatif karena mengurangi sampel dari data kelas utama. Di sisi lain, pengambilan sampel yang berlebihan menghasilkan sampel dari minoritas yang menciptakan potensi risiko overfitting [12].

Dalam kasus dataset yang tidak seimbang, contoh kelas minoritas sering kali kalah jumlah. Meskipun konsep yang mereka wakili biasanya lebih penting daripada kelas mayoritas. Algoritma pembelajaran mesin tradisional untuk penambangan data seperti k-nearest neighbor [1], pohon keputusan [3], Support Vector Machine [13], Random Forest [14] biasanya mencoba untuk memaksimalkan akurasi tingkat klasifikasi sambil mengabaikan biaya kesalahan klasifikasi kelas minoritas. Berbagai metode yang peka terhadap biaya telah diusulkan untuk menangani masalah ketidakseimbangan kelas. Pembelajaran yang peka terhadap biaya menerapkan biaya yang berbeda untuk kesalahan klasifikasi yang salah pada setiap kelas. Tujuannya adalah untuk menetapkan biaya sedemikian rupa sehingga biaya kesalahan klasifikasi untuk kelas minoritas akan tinggi dan rendah untuk kelas mayoritas. Klasifikasi yang stabil sulit ditemukan menggunakan metode yang peka terhadap biaya karena sangat sulit untuk menetapkan biaya kesalahan klasifikasi yang benar untuk setiap kelas. Model ensemble seperti bagging dan boosting biasanya digunakan untuk klasifikasi yang tidak seimbang [10], [15]. Pengklasifikasi ensemble adalah jenis algoritma di mana beberapa pembelajar digunakan untuk meningkatkan kinerja klasifikasi individual dengan menggabungkan hipotesis setiap pembelajar [3].

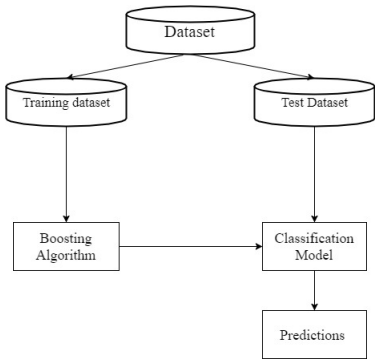
Dalam makalah ini, kami menyajikan algoritma boosting baru yang disebut MEBoost, yang menggabungkan dua penduga lemah secara bergantian pada set pelatihan. Sebagai penduga lemah, kami telah menggunakan pohon keputusan dan pengklasifikasi pohon tambahan. Dengan cara ini, kami memanfaatkan kedua pembelajar sekaligus menghindari keterbatasan dengan menggunakan pengklasifikasi basis tunggal dalam model boosting. Kami menguji kinerja MEBoost dengan pengklasifikasi boosting canggih lainnya seperti Adaboost, RUSBoost, SMOTEBoost, DataBoost, EUSBoost, Easy ensemble pada 12 set data tidak seimbang tolok ukur standar. Dari hasil eksperimen, dapat divalidasi bahwa penggunaan dua pembelajar seperti pohon keputusan dan pengklasifikasi pohon tambahan secara bergantian dengan Adaboost secara signifikan meningkatkan kinerja algoritma lain dan merupakan teknik yang menjanjikan untuk menangani masalah ketidakseimbangan kelas.

Sisa makalah ini disusun sebagai berikut, Bagian II merupakan karya terkait; Bagian III menyajikan rincian algoritma MEBoost yang diusulkan; Bagian IV menunjukkan hasil eksperimen, dan Bagian V menyimpulkan makalah.

II. RPEKERJAAN YANG MENYENANGKAN

Sepanjang dekade terakhir berbagai metode pengambilan sampel, metode ensemble, metode ensemble berdasarkan bagging dan boosting

telah menjadi fokus utama untuk menangani masalah klasifikasi dengan kumpulan data ketidakseimbangan kelas. Gambar 1 menggambarkan sketsa umum penerapan algoritma penguat pada masalah klasifikasi.



Gbr. 1: Peningkatan untuk mengklasifikasikan data yang tidak seimbang.

Beberapa metode ensemble diusulkan dalam literatur untuk menangani dataset yang tidak seimbang [10], [15]. Metode ensemble diusulkan oleh Sun et al. [12] yang mengubah masalah kelas biner yang tidak seimbang menjadi proses pembelajaran berganda. Metode yang diusulkan membagi instance kelas mayoritas menjadi beberapa sub dataset. Di sini setiap sub dataset memiliki jumlah instance kelas minoritas yang hampir sama. Dengan demikian, beberapa dataset seimbang dibuat dan digunakan untuk membuat pengklasifikasi biner. Kemudian kombinasi pengklasifikasi tersebut digunakan untuk mempelajari pengklasifikasi ensemble.

Boosting adalah meta classifier yang menggabungkan prediksi beberapa estimator dasar dan menggunakan teknik pemungutan suara untuk klasifikasi. Boosting menetapkan bobot pada instance berdasarkan seberapa sulitnya mereka untuk diklasifikasikan. Jadi, ia menetapkan bobot tinggi pada instance yang sulit. Di sini, bobot yang disumbangkan setiap estimator digunakan oleh estimator berikutnya. Kemudian, berdasarkan akurasi prediktif pembelajar dasar, bobot ditetapkan padanya. Bobot tersebut dipertimbangkan untuk prediksi instance baru. Meskipun boosting tidak ditujukan untuk masalah ketidakseimbangan kelas, karena karakteristik ini, boosting telah menjadi sangat ideal untuk masalah ketidakseimbangan kelas.

RUSBoost [16] adalah algoritma boosting hybrid yang menggunakan Adaboost dengan random under sampling sebagai metode sampling. Dari data yang tidak seimbang, random under sampling secara acak menghilangkan instance dari kelas utama di setiap iterasi. Adaboost [17] digunakan dengan random under sampling untuk membuat algoritma RUSBoost. Demikian pula SMOTEBoost dibuat menggunakan Adaboost dan teknik over sampling yang disebut SMOTE. Metode ini diusulkan dalam [18]. Ini melakukan over-sample instance kelas minoritas menggunakan teknik over sampling yang disebut SMOTE [11]. Dengan menggunakan*aku*tetangga terdekat dari kelas minoritas, instance sintetis dihasilkan dengan beroperasi di ruang fitur. Metode yang disebutkan di atas yang menggunakan pengambilan sampel di dalam Adaboost menunjukkan kinerja yang mengesankan dalam hal area di bawah kurva Receiver Operating Characteristic (ROC). Investigasi tentang perilaku SMOTEBoost dilakukan oleh Blagus dan Lusa [19] pada dataset yang tidak seimbang dengan dimensi tinggi. Di sini dataset dengan dimensi tinggi berarti di mana terdapat lebih banyak fitur daripada instance. Mereka datang sampai pada kesimpulan bahwa, karena SMOTE mengarahkan pengklasifikasi ke arah kelas minoritas, maka perlu dilakukan pemilihan fitur.

De Souza et al. mengusulkan algoritma Adaboost dinamis baru [20] di mana 10 estimator berbeda digunakan secara bergantian di setiap iterasi. Karena Adaboost menyimpan estimator yang lebih baik dan membuang estimator dengan kesalahan tinggi, dengan mengizinkan 10 estimator berbeda secara bergantian, hal itu mengurangi beban pengguna untuk memilih pembelajar. Algoritma ini tidak mengikuti konsep pembelajar lemah karena menggunakan estimator seperti Random forest, SVM, Neural Network sehingga juga membuatnya sangat mahal secara komputasi. Algoritma peningkatan ensemble evolusioner diusulkan oleh Galar et al. [21]. Ia menggunakan metode undersampling evolusioner sehingga disebut EUSBoost. Algoritma ini membuat beberapa sub dataset yang dihasilkan dengan metode under sampling acak untuk menemukan under-sampled terbaik dari dataset asli. EUSBoost juga dibangun berdasarkan algoritma AdaBoost [17].

Algoritma DataBoost atau metode DataBoost-IM diperkenalkan oleh Hongyu Guo [22]. Ia mengusulkan model ensemble yang menggunakan pembuatan data. Dalam algoritma ini, contoh kelas mayoritas dan minoritas yang sulit diidentifikasi selama pelaksanaan boosting. Kemudian contoh-contoh sulit tersebut dipilih secara terpisah dan digunakan untuk membuat contoh sintetis dari kelas masing-masing. Setelah itu contoh yang dibuat tersebut ditambahkan ke dataset utama. Easy Ensemble adalah metode ensemble yang diusulkan oleh Xu-Yung Liu [23]. Mereka membuat beberapa subset contoh kelas mayoritas. Kemudian menggunakan masing-masing dataset tersebut, ia melatih seorang pembelajar. Subset ini dibuat menggunakan random under sampling. Namun, ia membuat beberapa sub dataset untuk mengatasi keterbatasan utama random under sampling yaitu membuang contoh dari kelas mayoritas secara acak terlepas dari kepentingannya.

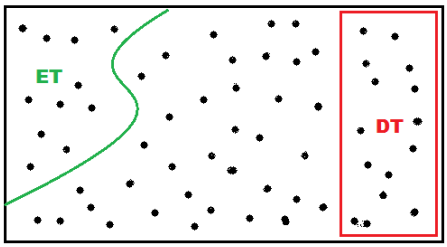
III. MEBBahasa Indonesia: OOSTÁLGORITMA

Sebagian besar algoritma boosting yang dibahas di Bagian II menggunakan satu estimator lemah untuk membuat model ensemble. Dalam metode MEBoost yang kami usulkan, alih-alih menggunakan satu estimator, kami menggunakan dua estimator berbeda secara bergantian. Untuk setiap iterasi, ia menggunakan pohon keputusan atau pengklasifikasi pohon tambahan sebagai pembelajarnya. Dengan melakukan ini, algoritma mengambil manfaat dari kedua pengklasifikasi tersebut. Algoritma ini juga membuang pembelajar dengan kinerja yang buruk berdasarkan desain prosedur boosting. Algoritma MEBoost tidak melakukan pengambilan sampel apa pun pada set kereta. Untuk setiap iterasi, kami menggunakan pohon keputusan dan pengklasifikasi pohon tambahan secara bergantian pada set data kereta. Pohon keputusan dibangun dari set kereta menggunakan entropi informasi [24]. Data kereta$S = S_1, S_2, S_3, S_4 \ldots$ *Bahasa Indonesia:*adalah sampel yang diklasifikasikan. SampelS_{Saya}terdiri dari vektor$X = X_1$ *Bahasa Indonesia:*$_{Saya}$*Bahasa Indonesia:*X_2*Bahasa Indonesia:*$_{Saya}$*Bahasa Indonesia:* X_3*Bahasa Indonesia:*$_{Saya}$*Bahasa Indonesia:*X_4*Bahasa Indonesia:*$_{Saya}$. Di Sini X_{Saya}*J* adalah representasi fitur atau nilai atribut dari sampel. Untuk setiap node, fitur/ atribut dipilih sedemikian rupa sehingga membagi dataset train menjadi subset dari setiap kelas dengan paling efektif. Extra tree adalah algoritma klasifikasi pohon acak. Ketika mencari pemisahan terbaik untuk memisahkan instance dari node ke dalam kelompok, Extra tree menggambar pemisahan acak untuk setiap jumlah fitur yang dipilih secara acak dan di antara mereka pemisahan terbaik dipilih [25]. Extra Tree Classifier bertindak seperti pohon keputusan jika jumlah fitur yang dipilih secara acak adalah 1.

Pseudo-code dari metode MEBoost yang kami usulkan diberikan dalam Algoritma 1. Ini adalah modifikasi dari Adaboost dasar

pengklasifikasi. Jumlah pengklasifikasi dasar tidak dibatasi di sini. Namun, pada setiap iterasi algoritma, MEBoost menguji setiap penduga lemah yang dipelajari dan membuangnya jika gagal menjadi pengklasifikasi lemah atau tingkat kesalahannya lebih besar atau sama dengan 0,5. Pengklasifikasi meta diuji pada data uji yang dipisahkan dan menyimpan kombinasi terbaik menurut skor auROC. Pengklasifikasi terus menambahkan pembelajar lemah ke model hingga tidak ada perubahan signifikan dalam auROC pada data uji.

Intuisi di balik ide ini adalah bahwa algoritma pohon seperti pohon tambahan dan pohon keputusan biasanya lebih cocok untuk skema peningkatan karena ketidakstabilannya. Dalam kumpulan data tertentu, sejumlah SVM lebih mungkin membuat batasan keputusan yang serupa. Tetapi ada kemungkinan besar bahwa pada kumpulan data itu setiap algoritma pohon akan menghasilkan pohon yang berbeda dengan cara yang berbeda dan mencakup subruang yang berbeda dari total kumpulan data. Jadi karena mereka mencakup subruang yang berbeda, menyisirnya di bawah skema peningkatan adalah resep yang sangat baik untuk algoritma klasifikasi yang baik. Untuk lebih memaksimalkan keragaman, kami menggunakan 2 algoritma pohon yang berbeda, yaitu Pohon tambahan dan pohon keputusan, di bawah skema peningkatan. Pada Gambar 2, persegi panjang mewakili seluruh kumpulan data dan setiap titik hitam mewakili sebuah contoh. Di sini titik-titik di dalam kotak merah mewakili bagian dari kumpulan data yang telah dieksplorasi oleh Pohon Keputusan (DT). Demikian pula batas hijau mewakili bagian dari kumpulan data yang telah dieksplorasi oleh Pohon Tambahan (ET). Dengan menggunakannya di dalam mekanisme penguat, kami mengambil karakteristik ketidakstabilan algoritma pohon sebagai keuntungan untuk hasil klasifikasi yang lebih baik.



Gbr. 2: Contoh kasus yang dicakup oleh algoritma pohon yang berbeda

Konvergensi algoritma bergantung pada parameter jendela stagnasi *Kami*Dalam makalah ini, kami membahas*Kami*=10. Oleh karena itu setelah kombinasi terbaik ditemukan maka akan dilanjutkan dengan penambahan *Kami*estimator dalam model dan jika tidak ada peningkatan signifikan dalam skor tes maka model ensemble akan mengembalikan meta classifier terbaik yang dipelajari. Kriteria penghentian awal ini[26] pertama kali digunakan oleh Bühlmann pada tahun 2003 [27] dan kemudian dipelajari oleh Jiang [28]. Algoritma MEBoost merupakan kombinasi penggunaan estimator alternatif De Souza dan kriteria penghentian awal Bühlmann dalam algoritma Adaboost.

IV. Bahasa IndonesiaEKSPERIMENTALRHASIL

Bagian ini menetapkan hasil eksperimen dan analisis kinerja algoritma MEBoost.

Algoritma 1Meningkatkan ME

Masukan:Data tidak seimbang,*D*Ukuran jendela *Kami* Keluaran:Model ansambel*H*. Metode:

1. Bahasa Indonesia:mengatur*T*ke0,*skor*$_{terbaik}$keangka 0Dan*tidak membaik*keangka 0
2. Bahasa Indonesia:inisialisasi berat,*aku*$_T$ *Saya*ke$\frac{1}{n}$untuk masing-masing$x_{Saya} \in D$Bahasa Indonesia:
3. Bahasa Indonesia:ketika*BENAR*Mengerjakan
4: meningkatkan*T*Bahasa Indonesia:
5: pilih jenis pohon penaksir*T* pelajari penaksir$H(\eta)$bertipe*T*
6: hitunglah tingkat kesalahan$H(\eta)$Bahasa Indonesia:
nomor 7 *kesalahan*($H(\eta)$) jika*kesalahan*($H(\eta)$)≥angka 0.5Kemudian
nomor 8:
9: kembali ke langkah 4 dan coba lagi;
10: berakhir jika
11: untuksetiap$x_{Saya} \in D$Mengerjakan
12: perbarui bobot*aku*$_T$ *Saya*
13: akhir untuk
14: *sebuah*$_T$=$\alpha_{catatan_{bahasa\ Inggris}}$$\frac{1\text{-}kesalahan(H(\eta))}{kesalahan(H(\eta))}$
15: pelajari meta classifierH_T=*tanda*($\sum_{H_t(x_{Saya})=1}$*skor*=
16: *skor auROC*(H_T,X_{tes}, *Kamu*$_{tes}$) jika*skor*$_{atas}$
17 tahun: ≤*skor*Kemudian
18: *skor*$_{terbaik}$=*skor*
19: $H_{terbaik}$=H_T
20 tahun: berakhir jika
21: jika*skor*≤*skor*$_{atas}$Kemudian
22: *tidak membaik*=*tidak membaik*+1 jika
23: *tidak membaik*=*Kami*Kemudian
24 tahun: *H*=$H_{terbaik}$
25: merusak
26 tahun: berakhir jika
27: berakhir jika
28:berakhir sementara

A. Dataset Tolok Ukur

Kumpulan data dengan rasio ketidakseimbangan yang berbeda dipilih dari repositori kumpulan data KEEL [29]. TABEL I berikut menyajikan ringkasan kumpulan data. Setiap kumpulan data yang tidak seimbang dengan rasio ketidakseimbangan berada dalam kisaran 1,87 hingga 41,03.

TABEL I: Deskripsi himpunan data.

Nama	Rasio ketidakseimbangan	contoh	fitur
ikan	1.87	768	8
kaca-0-1-2-3vs4-5-6	3.2	214	9
tiroid baru2	5.14	215	5
tiroid baru1	5.14	215	5
segmen0	6.02	tahun 2308	19
kaca6	6.38	214	9
ragi-2 vs 4	Tanggal 9.08	514	8
blok halaman-1-3-vs 4	15.86	472	10
kaca5	22.78	214	9
*ragi*4	28.1	tahun 1484	8
*ragi*5	32.72	tahun 1484	8
*ragi*6	41.03	tahun 1484	8

B. Metrik Evaluasi

Beberapa metrik evaluasi digunakan dalam literatur untuk mengukur kinerja algoritma klasifikasi. Dalam makalah ini kami menggunakan, area under Receiver Operating Characteristics curve

TABEL II: Kinerja rata-rata metode AdaBoost, EUSBoost, EasyEnsemble, SMOTEBoost, RUSBoost, DataBoost dan MEBoost pada 12 set data yang tidak seimbang.

Kumpulan data	Penguat Ada	EUSBoost	Ensemble Mudah	SMOTEBost meningkatkan	RUSBoost	Peningkatan Data	Meningkatkan ME
ikan	0.68	0.73	0.71	0.73	0.82	0.72	0.71
kaca-0-1-2-3	0,89	0,91	0.9	0,91	0,96	0.9	0,98
tiroid baru2	0,95	0,94	0,95	0.9	0,98	0,93	0,99
tiroid baru1	0,94	0,96	0,96	0,98	0,98	0,97	0,99
segmen0	0,96	0,97	0,98	0,99	0,99	0,99	0,99
kaca6	0.84	0.9	0,88	0,85	0,89	0,88	0,99
ragi-2 vs 4	0,88	0.9	0,89	0,89	0,94	0,89	0,98
blok halaman	0.92	0,98	0,93	0,95	0,97	0.9	0,99
kaca5	0,97	0,97	0,98	0,98	0,97	0,98	0,99
*ragi*4	0.62	0,85	0,78	0.71	0,89	0,75	0,91
*ragi*5	0.81	0,94	0,95	0,94	0,96	0,85	0,99
*ragi*6	0,74	0.82	0.84	0.81	0,93	0,87	0,95

(auROC) sebagai metrik perbandingan yang telah banyak digunakan sebagai standar untuk perbandingan kinerja dalam literatur kumpulan data yang tidak seimbang. Kurva ROC merupakan representasi batas keputusan terbaik untuk biaya antara rasio positif benar (TPR) dan rasio positif salah (FPR). Kurva ROC memetakan TPR terhadap FPR. TPR dan FPR didefinisikan sebagai berikut:

$$TPR = \frac{T.P.}{T.P. + \text{Bahasa Inggris}} \quad (1)$$

$$FPR = \frac{\text{Bahasa Inggris}}{\text{Bahasa Inggris} + \text{Bahasa Inggris}} \quad (2)$$

Di sini, TP menunjukkan jumlah sampel positif yang diklasifikasikan dengan benar, TN menunjukkan jumlah sampel negatif yang diklasifikasikan dengan benar, FP menunjukkan jumlah sampel negatif yang diklasifikasikan secara salah dan FN menunjukkan jumlah sampel positif yang diklasifikasikan dengan benar oleh penduga. Titik pada kurva au-ROC dibatasi antaraangka 0.angka 0hingga1 tahunDi manaangka 0.angka 0 berarti semua contoh salah diklasifikasikan dan1berarti semua contoh positif diklasifikasikan dengan benar. Garis*kamu*=*X*adalah ambang batas minimum karena garis tersebut mewakili skenario di mana kelas ditebak secara acak. Area di Bawah Kurva ROC sangat berguna sebagai metrik kinerja untuk masalah ketidakseimbangan kelas. Karena tidak bergantung pada kriteria keputusan yang dipilih dan probabilitas sebelumnya. Hubungan dominan dapat dibuat antara pengklasifikasi menggunakan perbandingan AUC.

C. Hasil

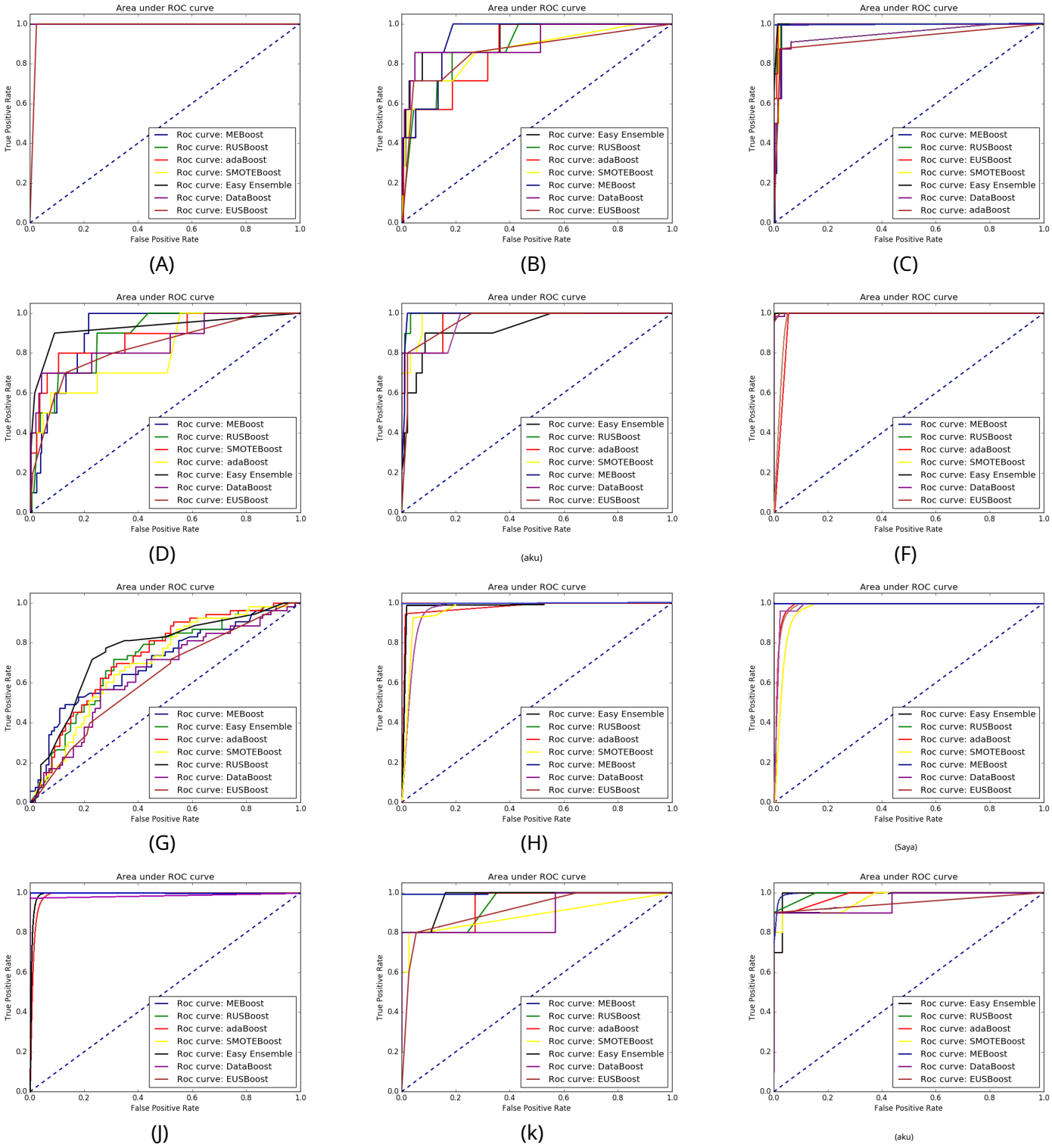
1) MEBoost vs algoritma peningkatan lainnya:Kami membandingkan kinerja MEBoost dengan metode RUSBoost, AdaBoost, Easy Ensemble, DataBoost, SMOTEBoost dan EUS-Boost. Kinerja klasifikasi pada dataset diukur dalam bentuk auROC. Untuk metode lain, pohon keputusan C4.5 digunakan sebagai pembelajar dasar. Dalam kasus MEBoost, metode yang kami usulkan, kami menggunakan pengklasifikasi C4.5 dan Extra Tree. Implementasi repositori Keel-dataset [29] digunakan untuk algoritma DataBoost, AdaBoost, RUSBoost, SMOTEBoost, EUSBoost dan Easy Ensemble. Semua dataset dibagi menjadi 3 bagian: Set pelatihan, Set pengujian dan Set validasi. Set validasi berisi5%dari kumpulan data. Pengklasifikasi dilatih dan diuji pada set pelatihan dan pengujian menggunakan validasi silang 5 kali lipat. Rata-rata Skor auROC pada set validasi dari 10 percobaan ditunjukkan pada TABEL II.

TABEL II menunjukkan kinerja pengklasifikasi MEBoost terhadap algoritma Boosting canggih lainnya seperti Adaboost, EUSBoost, EasyEnsemble, SMOTEBoost, RUSBoost dan DataBoost. MEBoost mampu mencapai skor auROC tertinggi di semua dataset kecuali pima di mana RUSBoost mencapai nilai tertinggi. Perhatikan bahwa rasio ketidakseimbangan dataset pima adalah yang terendah di antara semua kumpulan data.

2) Estimator jamak vs estimator tunggal :

Kami juga membandingkan kinerja MEBoost dengan boosting menggunakan estimator basis tunggal seperti Decision tree, Extra tree, Random forest, Support Vector Machine. Tabel III menyajikan hasil untuk percobaan ini. Di sini kami menggunakan algoritma AdaBoost. Sebagai pembelajar, kami menggunakan decision tree, extra tree, Support vector machine, Random Forest dan semuanya telah dibandingkan dengan metode yang kami usulkan di mana kami menggunakan beberapa estimator, decision tree, dan extra tree secara bergantian. Hasil pada Tabel III menunjukkan bahwa manfaat penggunaan beberapa estimator cukup menjanjikan. Meskipun dalam beberapa kasus metode yang diusulkan tidak dapat memperoleh hasil terbaik, melatih pembelajar seperti decision tree atau extra tree membutuhkan daya komputasi yang jauh lebih sedikit daripada melatih support vector machine (SVM) atau random forest. Dalam dataset glass-0-1-2-3 dan yeast6, estimator MEBoost dan Random Forest mencapai skor auROC tertinggi. Demikian pula dalam dataset newthyroid2 dan yeast5, estimator MEBoost dan Random Forest mencapai skor auROC teratas. Pada MEBoost newthyroid1 dan segment0, estimator Random Forest dan estimator SVM mencapai skor auROC tertinggi secara bersamaan. Dari segi komputasi, Support Vector Machine dan Random jauh lebih kompleks daripada Extra tree dan decision tree secara bersamaan. Jadi, meskipun keduanya memberikan skor yang sama, MEBoost lebih disukai mengingat biaya komputasinya.

Dari hasil tabel pada TABEL II, III kami telah menemukan bahwa teknik multiple estimator menunjukkan kinerja yang lebih baik dibandingkan dengan algoritma boosting canggih lainnya dan juga weak or strong learner lainnya. Kami juga menyajikan analisis ROC untuk semua kumpulan data yang berbeda dan plotnya diberikan pada Gambar 3.



Gbr. 3: Analisis ROC untuk kumpulan data berbeda yang digunakan dalam makalah ini: (a) glass5, (b) yeast6, (c) yeast5, (d) yeast4, (e) yeast-2 vs 4, (f) segment0, (g) pima, (h) page-blocks-1-3 vs 4, (i) newthyroid2, (j) new-thyroid1, (k) glass6 dan (l) glass-0-1-2-3 vs 4-5-6.

TABEL III: Perbandingan kinerja MEBoost dengan Adaboost menggunakan estimator basis tunggal yang berbeda pada 12 set data yang tidak seimbang.

Kumpulan data	Keputusan pohon	Tambahan pohon	Acak hutan	Bahasa Indonesia: SVM	Meningkatkan ME
ikan	0.68	0.64	0.66	0.69	0.71
kaca-0-1-2-3	0,89	0,97	0,98	0,96	0,98
tiroid baru2	0,95	0,97	0,98	0,99	0,99
tiroid baru1	0,94	0,98	0,99	0,99	0,99
segmen0	0,96	0,99	0,99	0,99	0,99
kaca6	0.84	0,97	0,93	0,93	0,99
ragi-2 vs 4	0,88	0,95	0,95	0,91	0,98
blok halaman	0.92	0,96	0,98	0,98	0,99
kaca5	0,97	0,97	0,95	0,93	0,99
*ragi*4	0.62	0.92	0.8	0,89	0,91
*ragi*5	0.81	0,97	0,98	0,99	0,99
*ragi*6	0,74	0,91	0,95	0,94	0,95

V.C. Bahasa IndonesiaKESIMPULAN

Sebagian besar algoritma klasifikasi terutama berfokus pada contoh kelas mayoritas daripada contoh kelas minoritas yang lebih penting. Jadi, tugasnya cukup menantang untuk membangun pengklasifikasi yang dapat mengklasifikasikan contoh kelas minoritas dengan benar. Dengan harapan untuk mengatasi masalah ketidakseimbangan kelas ini dalam makalah ini disajikan algoritma baru yang disebut MEBoost, atau Boosting dengan beberapa pelajar. MEBoost telah dibandingkan dengan teknik boosting yang efektif seperti algoritma SMOTEBoost, RUSBoost, Adaboost, DataBoost, EUSBoost dan Easy Ensemble. Dari hasil eksperimen, kami telah menyimpulkan bahwa MEBoost berkinerja lebih baik dibandingkan dengan teknik serupa.

Algoritma MEBoost berbeda dari semua metode boosting lainnya karena menggunakan C4.5 dan Extra tree classifier secara bergantian, bukan hanya menggunakan salah satunya. Hal ini memungkinkan untuk memanfaatkan karakteristik kedua pembelajar dan membuang kelemahan masing-masing. Baik C4.5 maupun Extra tree classifier memiliki kelebihan dan kekurangan satu sama lain. Hasilnya menunjukkan bahwa penggunaan 2 estimator yang berbeda, bukan 1, memiliki dampak yang besar pada skor auROC. Di masa mendatang, kami bermaksud untuk melakukan eksperimen ekstensif untuk terus menyelidiki kinerja MEBoost dengan pembelajar lain.

RReferensi

[1] DM Farid, MA Al-Mamun, B. Manderick, dan A. Nowe, "Pengklasifikasi berbasis aturan adaptif untuk menambang data biologis besar," *Sistem Pakar dengan Aplikasi*, vol. 64, hlm. 305–316, Desember 2016.

[2] DM Farid, L. Zhang, CM Rahman, M. Hossain, dan R. Strachan, "Pohon keputusan hibrida dan pengklasifikasi bayes naif untuk tugas klasifikasi multi-kelas," *Sistem Pakar dengan Aplikasi*, vol. 41, no. 4, hal. 1937–1946, Maret 2014.

[3] DM Farid, L. Zhang, A. Hossain, CM Rahman, R. Strachan, G. Sexton, dan K. Dahal, "Pengklasifikasi ansambel adaptif untuk menambang aliran data pergeseran konsep," *Sistem Pakar dengan Aplikasi*, vol. 40, no. 15, hal. 5895–5906, November 2013.

[4] DM Farid, A. Nowé, dan B. Manderick, "Metode penyeimbangan data baru untuk mengklasifikasikan data genomik multi-kelas yang tidak seimbang," *Konferensi Belgia-Belanda ke-25 tentang Pembelajaran Mesin (Benelearn)*, hlm. 1– 2, 12-13 September 2016.

[5] J. Laurikkala, "Meningkatkan identifikasi kelas kecil yang sulit dengan menyeimbangkan distribusi kelas," *Kecerdasan Buatan dalam Kedokteran*, hal. 63–66, 2001.

[6] I. Mani dan I. Zhang, "pendekatan knn terhadap distribusi data yang tidak seimbang: studi kasus yang melibatkan ekstraksi informasi," dalam *Prosiding lokakarya tentang pembelajaran dari kumpulan data yang tidak seimbang*, jilid 126, 2003.

[7] S.-J. Yen dan Y.-S. Lee, "Pendekatan under-sampling berbasis cluster untuk distribusi data yang tidak seimbang," *Sistem Pakar dengan Aplikasi*, vol. 36, no. 3, hal. 5718–5727, 2009.

[8] F. Rayhan, S. Ahmed, S. Shatabda, DM Farid, Z. Mousavian, A. Dehzangi, dan MS Rahman, "idti-esboost: Identifikasi interaksi target obat menggunakan fitur evolusi dan struktural dengan boosting," *arXiv pracetak arXiv:1707.00994*Bahasa Indonesia: Tahun 2017.

[9] M. Kubat, S. Matwin*dan lain-lain*., "Mengatasi kutukan set pelatihan yang tidak seimbang: pemilihan satu sisi," dalam *Bahasa Inggris ICML*, vol. 97. Nashville, Amerika Serikat, 1997, hlm. 179–186.

[10] H. He dan EA Garcia, "Belajar dari data yang tidak seimbang," *Transaksi IEEE pada Rekayasa Pengetahuan dan Data*, vol. 21, no. 9, hal. 1263–1284, Juni 2009.

[11] NV Chawla, KW Bowyer, LO Hall, dan WP Kegelmeyer, "SMOTE: Teknik over-sampling minoritas sintetis," *Jurnal Penelitian Kecerdasan Buatan*, vol. 16, hal. 321–357, Juni 2002.

[12] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, dan Y. Zhou, "Metode ensemble baru untuk mengklasifikasikan data yang tidak seimbang," *Pengenalan Pola*, vol. 48, no. 5, hal. 1623–1637, Mei 2015.

[13] C. Cortes dan V. Vapnik, "Mesin vektor pendukung," *Pembelajaran mesin*, vol. 20, no. 3, hal. 273–297, 1995.

[14] A. Liaw dan M. Wiener*dan lain-lain.*"Klasifikasi dan regresi dengan randomforest," *Berita R*, vol. 2, no. 3, hal. 18–22, 2002.

[15] Y. Sun, AKC Wong, dan MS Kamel, "Klasifikasi data ketidakseimbangan: Sebuah tinjauan," *Jurnal Internasional Pengenalan Pola dan Kecerdasan Buatan*, vol. 23, no. 4, hal. 687–719, Juni 2009.

[16] C. Seiffert, TM Khoshgoftaar, JV Hulse, dan A. Napolitano, "Rusboost: Pendekatan hibrida untuk mengurangi ketidakseimbangan kelas," *Transaksi IEEE pada Sistem, Manusia, dan Sibernetika - Bagian A: Sistem dan Manusia*, vol. 40, no. 1, hal. 185–197, Januari 2010.

[17] Y. Freund dan RE Schapire*dan lain-lain.*"Eksperimen dengan algoritma peningkatan baru," dalam *bahasa inggris*, vol. 96, 1996, hlm. 148–156.

[18] NV Chawla, A. Lazarevic, LO Hall, dan KW Bowyer, "Smoteboost: Meningkatkan prediksi kelas minoritas dalam meningkatkan," *Konferensi Eropa ke-7 tentang Prinsip dan Praktik Penemuan Pengetahuan dalam Basis Data*, hlm. 107–109, 22-26 September 2003.

[19] R. Blagus dan L. Lusa, "SMOTE untuk data kelas tidak seimbang berdimensi tinggi," *Bioinformatika BMC*, vol. 14, no. 106, hlm. 1–16, Maret 2013.

[20] É. de Souza dan S. Matwin, "Memperluas adaboost untuk memvariasikan pengklasifikasi dasarnya secara berulang," *Kemajuan dalam Kecerdasan Buatan*, hlm. 384–389, 2011.

[21] M. Galar, A. Fernández, E. Barrenechea, dan F. Herrera, "Eusboost: Meningkatkan ensemble untuk set data yang sangat tidak seimbang dengan undersampling evolusioner," *Pengenalan Pola*, vol. 46, no. 12, hal. 3460–3471, Desember 2013.

[22] H. Guo dan HL Viktor, "Belajar dari kumpulan data yang tidak seimbang dengan peningkatan dan pembuatan data: pendekatan databoost-im," *Buletin Eksplorasi ACM Sigkdd*, vol. 6, no. 1, hal. 30–39, 2004.

[23] X.-Y. Liu, J. Wu, dan Z.-H. Zhou, "Undersampling eksploratif untuk pembelajaran ketidakseimbangan kelas," *Transaksi IEEE pada Sistem, Manusia, dan Sibernetika, Bagian B (Sibernetika)*, vol. 39, no. 2, hal. 539–550, 2009.

[24] JR Quinlan, "Peningkatan penggunaan atribut berkelanjutan di c4. 5," *Jurnal penelitian kecerdasan buatan*, vol. 4, hal. 77–90, 1996.

[25] P. Geurts, D. Ernst, dan L. Wehenkel, "Pohon yang sangat acak," *Pembelajaran mesin*, vol. 63, no. 1, hal. 3–42, 2006.

[26] Y. Yao, L. Rosasco, dan A. Caponnetto, "Tentang penghentian dini dalam pembelajaran penurunan gradien," *Perkiraan Konstruktif*, vol. 26, no. 2, hal. 289–315, 2007.

[27] P. Bühlmann dan B. Yu, "Peningkatan dengan kehilangan l 2: regresi dan klasifikasi," *Jurnal Asosiasi Statistik Amerika*, vol. 98, no. 462, hal. 324–339, 2003.

[28] W. Jiang, "Konsistensi proses untuk adaboost," *Catatan Statistik*, hlm. 13–29, 2004.

[29] J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, dan F. Herrera, "Alat perangkat lunak penambangan data Keel: Repositori kumpulan data, integrasi algoritma dan analisis eksperimental

kerangka,"*Jurnal Logika Bernilai Ganda dan Komputasi Lunak*, vol. 17, no. 2-3, hal. 255–287, 2011.