KLASIFIKASI PADA DATASET TAK SEIMBANG

SEMESTER 3

D4 SAINS DATA TERAPAN

RONNY SUSETYOKO, S.Si., M.Si.

Masalah Ketidakseimbangan Kelas

- Masalah
  - Ketidakseimbangan Kelas: contoh dalam data pelatihan milik satu kelas jauh lebih banyak daripada contoh di kelas lain.
  - Sebagian besar sistem pembelajaran menganggap set pelatihan seimbang.
- Hasil:
  - mempengaruhi kinerja yang dicapai oleh sistem pembelajaran yang ada.
  - Sistem pembelajaran mungkin mengalami kesulitan untuk mempelajari konsep yang berkaitan dengan kelas minoritas.

Mengapa Belajar Kumpulan Data yang Tidak Seimbang mungkin sulit?

- Database rekam medis tentang penyakit langka
- Tugas pemantauan kesalahan terus menerus
- Komunitas ML tampaknya setuju dengan hipotesis bahwa ketidakseimbangan kelas adalah hipotesis utama dalam mendorong pengklasifikasi dalam domain yang tidak seimbang.

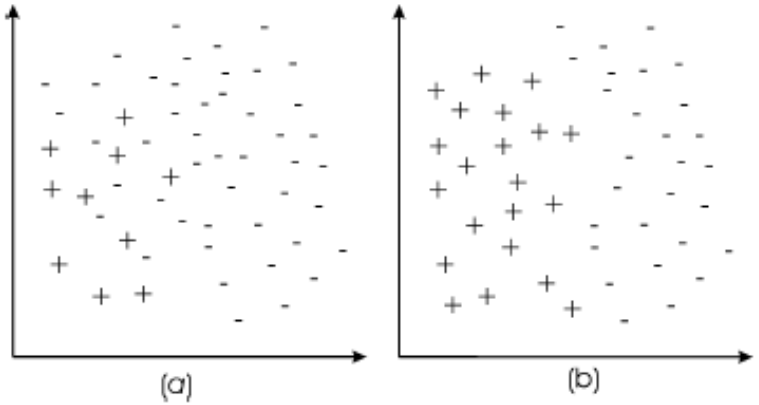
Mengapa Belajar dari Kumpulan Data yang Tidak Seimbang mungkin sulit?

- Namun, algoritme ML standar mampu mendorong pengklasifikasi yang baik bahkan dalam set pelatihan yang sangat tidak seimbang, misalnya dalam set data yang orang sakit.
  - Ketidakseimbangan kelas bukan satu-satunya masalah penurunan kinerja dalam algoritma pembelajaran
  - Jadi, apa faktor lainnya???

Mengapa Belajar dari Kumpulan Data yang Tidak Seimbang mungkin sulit?

- Kasing cadangan dari kelas minoritas dapat membingungkan pengklasifikasi seperti *k-nearest Neighbor (k-NN)*

Figure 1: Many negative cases against some spare positive cases (a) balanced data set with well-defined clusters (b).



Motivasi dan Metode

Jawaban: Derajat tumpang tindih data antar kelas

Motivasi:

- Seimbangkan data pelatihan
- Hapus contoh noise yang terletak di sisi yang salah dari batas keputusan

Metode:

- Metode pengambilan sampel berlebih: SMOTE
- Metode pembersihan data: Tomek links, and Edited Nearest Neighbor Rule

Beberapa Metode

Baseline Methods

- Random over-sampling
- Random under-sampling

Under-sampling Methods

- Tomek links
- Condensed Nearest Neighbor Rule
- One-sided selection
- CNN + Tomek links
- Neighborhood Cleaning Rule

Over-sampling Methods

- Smote

Combination of Over-sampling method with Under-sampling method

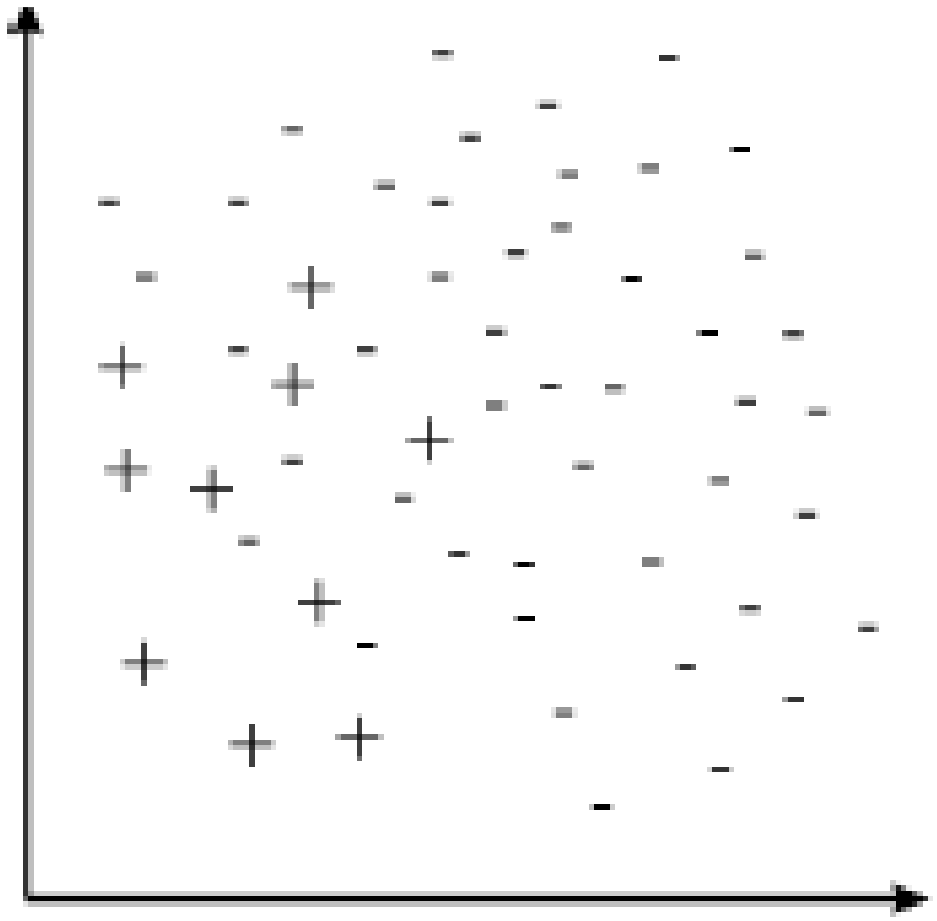
- Smote + Tomek links
- Smote + ENN

Baseline Methods

- Baseline methods
  - Random over-sampling
    - replikasi acak dari contoh kelas minoritas
    - Dapat meningkatkan kemungkinan terjadinya overfitting
  - Random under-sampling
    - eliminasi acak contoh kelas mayoritas
    - Dapat membuang data yang berpotensi berguna yang mungkin penting untuk proses induksi

Four Groups of Negative Examples

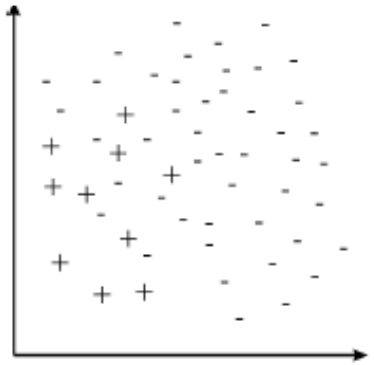
- Noise examples
- Borderline examples
  - Borderline examples : tidak aman karena sejumlah kecil noise dapat membuat mereka jatuh di sisi yang salah dari decision border
- Redundant examples
- Safe examples

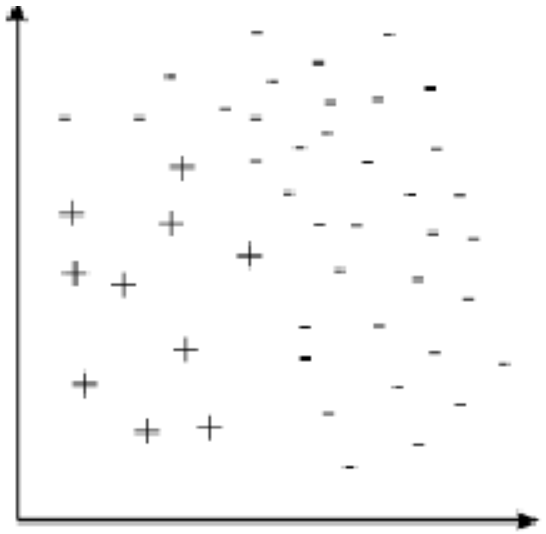


Tomek links [1]

- Untuk menghilangkan contoh noise dan borderline
- tautan tomek
  - Ei, Ej milik kelas yang berbeda, d (Ei, Ej) adalah jarak antara mereka.
  - Pasangan A (Ei, Ej) disebut hubungan Tomek jika tidak ada contoh El, sehingga d(Ei, El) < d(Ei, Ej) atau d(Ej , El) < d(Ei, Ej).

Tomek links





Condensed Nearest Neighbor Rule (CNN rule) [2]

Untuk memilih titik di dekat batas antara kelas

Untuk menemukan subset contoh yang konsisten.

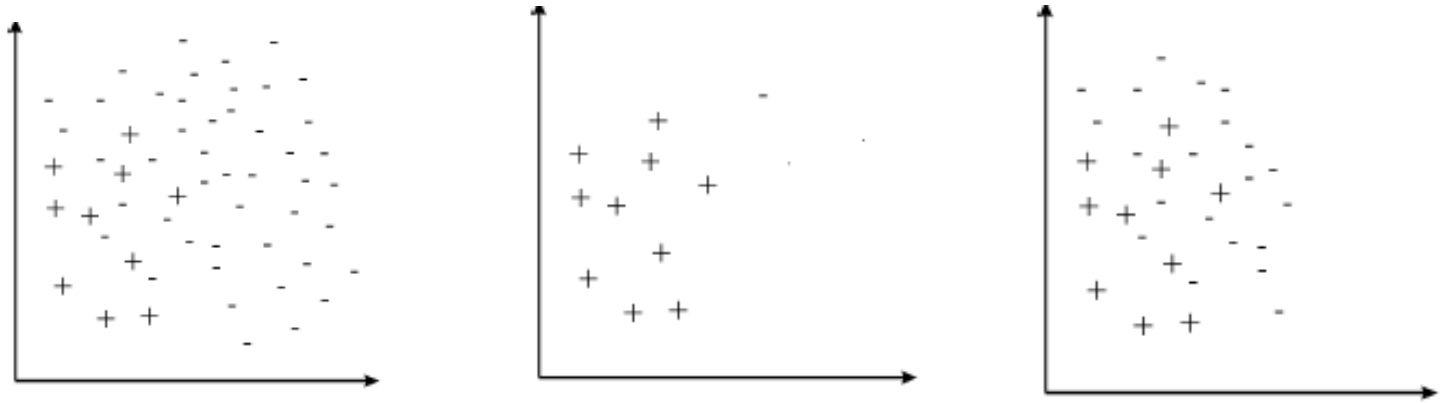
- Subset E’⊆E konsisten dengan E jika menggunakan 1-tetangga terdekat, E’ mengklasifikasikan contoh-contoh di E dengan benar

Algoritma:

- Biarkan E menjadi set pelatihan asli
- Misalkan E’ berisi semua contoh positif dari S dan satu contoh negatif yang dipilih secara acak
- Klasifikasikan E dengan aturan 1-NN menggunakan contoh pada E’
- Pindahkan semua contoh yang salah diklasifikasikan dari E ke E’

Sensitif terhadap kebisingan. Contoh-contoh yang berisik kemungkinan besar akan salah diklasifikasikan, banyak dari mereka akan ditambahkan ke set pelatihan.

Condensed Nearest Neighbor Rule (CNN rule)



One-sided selection [3] vs CNN+Tomek links

- One-sided selection
  - Tomek links + CNN
- CNN + Tomek links
  - Diusulkan oleh penulis
  - Menemukan tautan Tomek sangat menuntut komputasi, akan lebih murah secara komputasi jika dilakukan pada kumpulan data yang dikurangi.

Neighborhood Cleaning Rule [4]

- Untuk menghapus contoh kelas mayoritas
- Berbeda dari OSS, lebih menekankan pembersihan data daripada reduksi data
- Algoritma:
  - Temukan tiga tetangga terdekat untuk setiap contoh Ei di set pelatihan
  - Jika Ei termasuk dalam kelas mayoritas, & tiga tetangga terdekat mengklasifikasikannya sebagai kelas minoritas, maka hilangkan Ei
  - Jika Ei termasuk kelas minoritas, dan tiga tetangga terdekat mengklasifikasikannya sebagai kelas mayoritas, maka hilangkan tiga tetangga terdekat

Smote: Synthetic Minority Over-sampling Technique [6]

- Untuk membentuk contoh kelas minoritas baru dengan cara interpolasi antara beberapa contoh kelas minoritas yang terletak bersama-sama.
- di ``ruang fitur'' daripada ``ruang data''
- Algoritma: Untuk setiap contoh kelas minoritas, perkenalkan contoh sintetik di sepanjang segmen garis yang menghubungkan setiap/semua k kelas minoritas tetangga terdekat.
- Catatan: Tergantung pada jumlah over-sampling yang diperlukan, tetangga dari k tetangga terdekat dipilih secara acak.
- Sebagai contoh: jika kita menggunakan 5 tetangga terdekat, jika jumlah over-sampling yang dibutuhkan adalah 200%, hanya dua tetangga dari lima tetangga terdekat yang dipilih dan satu sampel dihasilkan ke arah masing-masing.

Consider a sample (6,4) and let (4,3) be its nearest neighbor.

(6,4) is the sample for which k-nearest neighbors are being identified

(4,3) is one of its k-nearest neighbors.

Let:

f1_1 = 6 f2_1 = 4 f2_1 - f1_1 = -2

f1_2 = 4 f2_2 = 3 f2_2 - f1_2 = -1

The new samples will be generated as

(f1',f2') = (6,4) + rand(0-1) * (-2,-1)

rand(0-1) generates a random number between 0 and 1.

Smote: Synthetic Minority Over-sampling Technique

- Sampel sintetis dihasilkan dengan cara berikut:
  - Ambil perbedaan antara vektor fitur (sampel) yang dipertimbangkan dan tetangga terdekatnya.
  - Kalikan perbedaan ini dengan angka acak antara 0 dan 1
  - Tambahkan ke vektor fitur yang sedang dipertimbangkan.

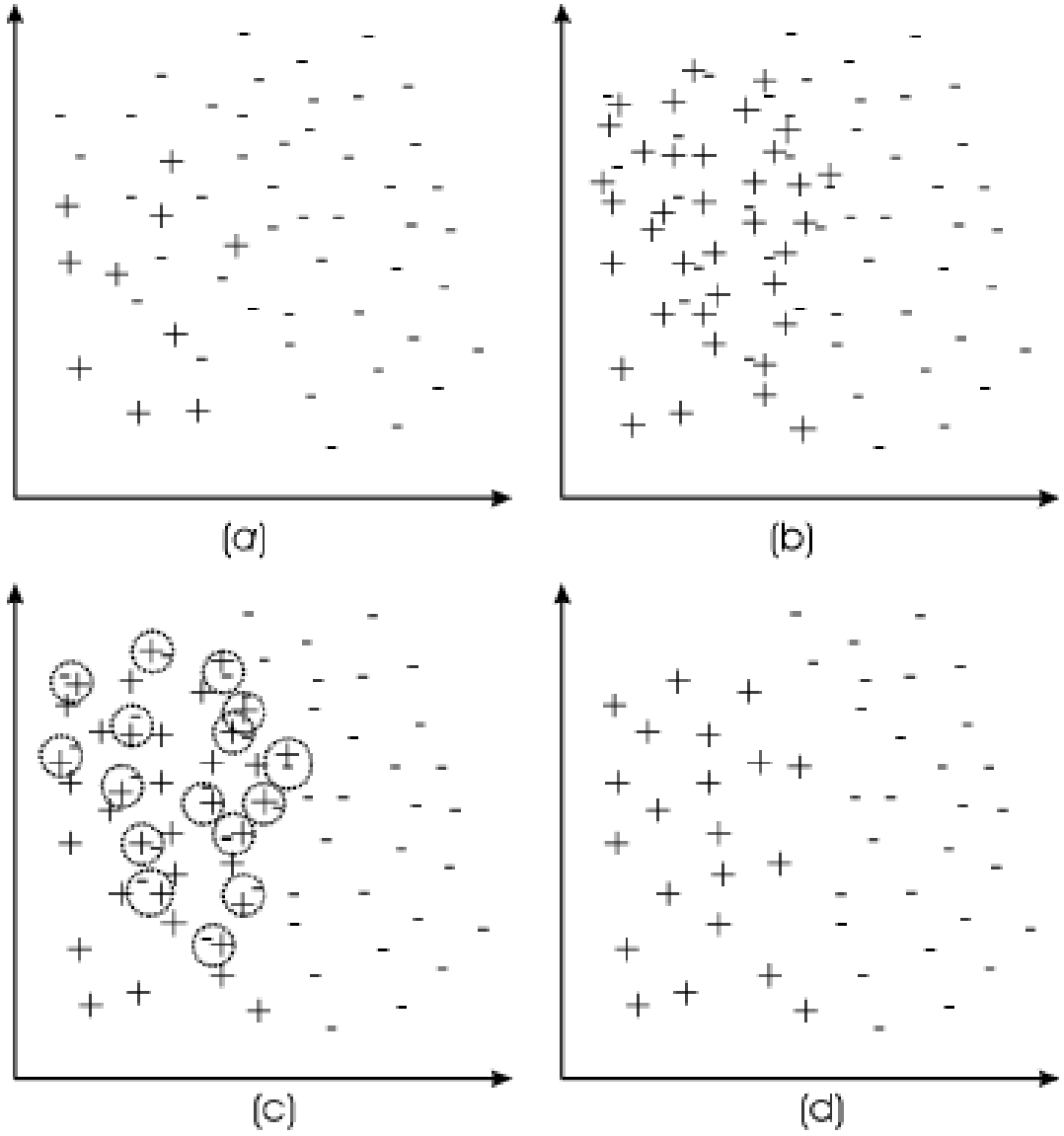
Smote + Tomek links

Masalah dengan Smote: mungkin memperkenalkan contoh kelas minoritas buatan terlalu dalam di ruang kelas mayoritas.

Tautan Tomek: pembersihan data

Alih-alih menghapus hanya contoh kelas mayoritas yang membentuk tautan Tomek, contoh dari kedua kelas dihapus

Smote + Tomek links



Smote + ENN

ENN menghapus setiap contoh yang label kelasnya berbeda dari kelas setidaknya dua dari tiga tetangga terdekatnya.

ENN menghapus lebih banyak contoh daripada tautan Tomek

ENN menghapus contoh dari kedua kelas

Experimental Evaluation

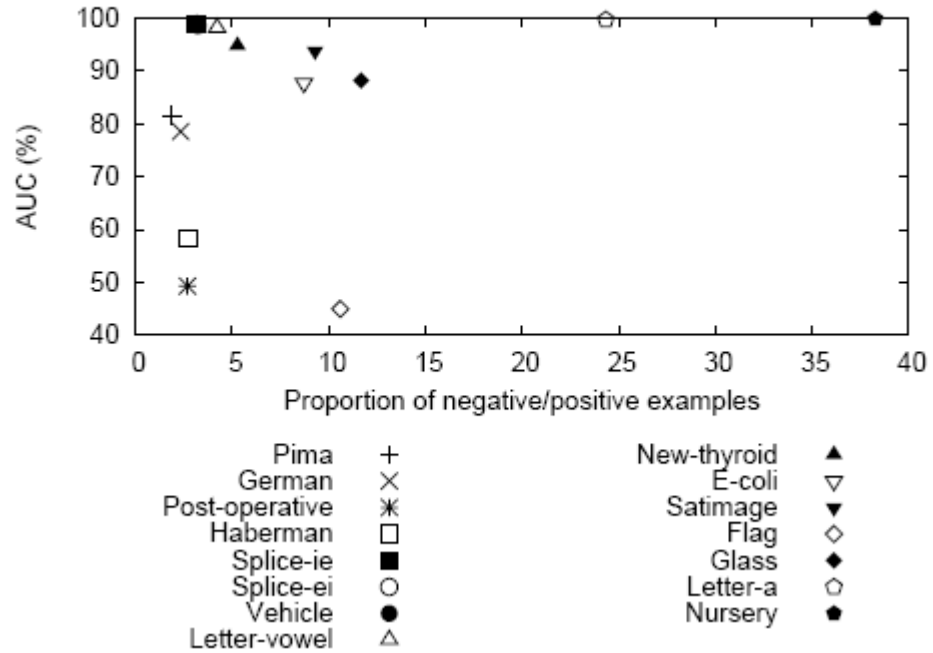
Table 3: Data sets summary descriptions.

Data set	#Examples	#Attributes (quanti., quali.)	Class (min., maj.)	Class % (min., maj.)	Majority Error
Pima	768	8 (8,0)	(1, 0)	(34.77%, 65.23%)	65.23%
German	1000	20 (7,13)	(Bad, Good)	(30.00%, 70.00%)	70.00%
Post-operative	90	8 (1,7)	(S, remainder)	(26.67%, 73.33%)	73.33%
Haberman	306	3 (3,0)	(Die, Survive)	(26.47%, 73.53%)	73.53%
Splice-ie	3176	60 (0,60)	(ie, remainder)	(24.09%, 75.91%)	75.91%
Splice-ei	3176	60 (0,60)	(ei, remainder)	(23.99%, 76.01%)	76.01%
Vehicle	846	18 (18,0)	(van, remainder)	(23.52%, 76.48%)	76.48%
Letter-vowel	20000	16 (16,0)	(all vowels, remainder)	(19.39%, 80.61%)	80.61%
New-thyroid	215	5 (5,0)	(hypo, remainder)	(16.28%, 83.72%)	83.72%
E.Coli	336	7 (7,0)	(iMU, remainder)	(10.42%, 89.58%)	89.58%
Satimage	6435	36 (36,0)	(4, remainder)	(9.73%, 90.27%)	90.27%
Flag	194	28 (10,18)	(white, remainder)	(8.76%, 91.24%)	91.24%
Glass	214	9 (9,0)	(Ve-win-float-proc, remainder)	(7.94%, 92.06%)	92.06%
Letter-a	20000	16 (16,0)	(a, remainder)	(3.95%, 96.05%)	96.05%
Nursery	12960	8 (8,0)	(not_recom, remainder)	(2.55%, 97.45%)	97.45%

- 10 methods
- 13 set data UCI yang memiliki tingkat ketidakseimbangan yang berbeda

First Stage

Figure 3: Proportion of negative/positive examples versus AUC.



- Ran 4.5 atas set data asli yang tidak seimbang menggunakan validasi silang 10 kali lipat

First Stage

Fakta:

- Terlepas dari tingkat ketidakseimbangan yang besar, set data Letter-a dan Nursery memperoleh hampir 100% AUC

Menyimpulkan:

- domain dengan kelas yang tidak tumpang tindih tampaknya tidak bermasalah untuk dipelajari terlepas dari tingkat ketidakseimbangannya
- Tetapi ketika bersekutu dengan kelas yang sangat tumpang tindih, itu dapat secara signifikan mengurangi jumlah contoh kelas minoritas yang diklasifikasikan dengan benar.
- Hubungan antara ukuran set pelatihan dan kinerja
  - Untuk kumpulan data kecil yang tidak seimbang, ketika ada tingkat tumpang tindih kelas yang besar dan kelas dibagi lagi menjadi subcluster, kelas minoritas kurang terwakili oleh jumlah contoh yang dikurangi secara berlebihan.
  - Untuk kumpulan data yang besar, efek dari faktor-faktor rumit ini tampaknya berkurang, kelas minoritas lebih baik diwakili oleh lebih banyak contoh.

Second Stage

- Menerapkan metode over dan under-sampling ke set data asli
- Fakta:
  - Pemangkasan jarang mengarah pada peningkatan AUC untuk set data asli dan seimbang.
  - Semua hasil terbaik (hasil dalam huruf tebal) diperoleh dengan metode over-sampling.
  - Metode over-sampling memiliki peringkat yang lebih baik daripada metode under-sampling
  - Pengambilan sampel berlebih secara acak khususnya memiliki peringkat yang baik di antara metode sisanya
  - Dua dari metode yang diusulkan: Smote+Tomek dan Smoke+ENN umumnya berada di antara yang terbaik untuk kumpulan data dengan sejumlah kecil contoh positif
- Penjelasan
  - Hilangnya kinerja secara langsung berkaitan dengan kurangnya contoh kelas minoritas dalam hubungannya dengan faktor rumit lainnya.
  - Over sampling adalah metode yang paling langsung menyerang masalah kurangnya contoh kelas minoritas.

Second Stage - hasil untuk kumpulan data asli dan lebih dari sampel

Table 4: AUC results for the original and over-sampled data sets.

Data set	Pruning	Original	Rand Over	Smote	Smote+Tomek	Smote+ENN
Pima	yes	81.53(5.11)	85.32(4.17)	**85.49(5.17)**	84.46(5.84)	83.66(4.77)
	no	82.33(5.70)	**86.03(4.14)**	85.97(5.82)	85.56(6.02)	83.64(5.35)
German	yes	79.19(5.84)	**84.65(3.80)**	80.74(5.43)	81.75(4.78)	80.91(4.36)
	no	**85.94(4.14)**	85.56(4.31)	84.51(4.55)	84.02(3.94)	83.90(3.70))
Post-operative	yes	49.29(2.26)	**68.79(23.93)**	55.66(24.66)	41.80(16.59)	59.83(33.91)
	no	**78.23(15.03)**	71.33(23.43)	68.19(26.62)	47.99(16.61)	59.48(34.91)
Haberman	yes	58.25(12.26)	71.81(13.42)	72.23(9.82)	75.73(6.55)	**76.38(5.51)**
	no	67.91(13.76)	73.58(14.22)	75.45(11.02)	**78.41(7.11)**	77.01(5.10)
Splice-ie	yes	98.76(0.56)	**98.89(0.47)**	98.46(0.87)	98.26(0.51)	97.97(0.74)
	no	**99.30(0.30)**	99.09(0.27)	99.19(0.28)	99.13(0.31)	98.88(0.34)
Splice-ei	yes	98.77(0.46)	98.80(0.44)	**98.92(0.44)**	98.87(0.44)	98.85(0.60)
	no	99.47(0.61)	**99.52(0.60)**	**99.52(0.26)**	99.51(0.32)	99.49(0.16)
Vehicle	yes	98.49(0.84)	**99.14(0.73)**	98.96(0.98)	98.96(0.98)	97.92(1.09)
	no	98.45(0.90)	**99.13(0.75)**	99.04(0.85)	99.04(0.85)	98.22(0.90)
Letter-vowel	yes	98.07(0.63)	98.80(0.32)	98.90(0.20)	98.90(0.20)	**98.94(0.22)**
	no	98.81(0.33)	98.84(0.27)	99.15(0.17)	99.14(0.17)	**99.19(0.15)**
New-thyroid	yes	94.73(9.24)	98.39(2.91)	98.91(1.84)	98.91(1.84)	**99.22(1.72)**
	no	94.98(9.38)	98.89(2.68)	98.91(1.84)	98.91(1.84)	**99.22(1.72)**
E.Coli	yes	87.64(15.75)	93.24(6.72)	95.49(4.30)	**95.98(4.21)**	95.29(3.79)
	no	92.50(7.71)	93.55(6.89)	95.49(4.30)	**95.98(4.21)**	95.29(3.79)
Satimage	yes	93.73(1.91)	95.34(1.25)	95.43(1.03)	95.43(1.03)	**95.67(1.18)**
	no	94.82(1.18)	95.52(1.12)	95.69(1.28)	95.69(1.28)	**96.06(1.20)**
Flag	yes	45.00(15.81)	**79.91(28.72)**	73.62(30.16)	79.30(28.68)	79.32(28.83)
	no	76.65(27.34)	79.78(28.98)	73.87(30.34)	**82.06(29.52)**	78.56(28.79)
Glass	yes	88.16(12.28)	92.20(12.11)	91.40(8.24)	91.40(8.24)	**92.90(7.30)**
	no	88.16(12.28)	92.07(12.09)	91.27(8.38)	91.27(8.38)	**93.40(7.61)**
Letter-a	yes	99.61(0.40)	99.77(0.30)	**99.91(0.12)**	**99.91(0.12)**	**99.91(0.12)**
	no	99.67(0.37)	99.78(0.29)	**99.92(0.12)**	**99.92(0.12)**	99.91(0.14)
Nursery	yes	99.79(0.11)	**99.99(0.01)**	99.21(0.55)	99.27(0.36)	97.80(1.07)
	no	99.96(0.05)	**99.99(0.01)**	99.75(0.34)	99.53(0.31)	99.20(0.51)

Second Stage - hasil untuk set data asli dan di bawah sampel

Table 5: AUC results for the under-sampled data sets.

Data set	Pruning	Rand Under	CNN	CNN+Tomek	Tomek	OSS	NCL
Pima	yes	81.17(3.87)	79.60(6.22)	80.30(3.86)	82.56(5.11)	77.89(5.37)	81.61(4.48)
	no	81.49(4.29)	80.08(5.82)	81.71(3.69)	83.11(4.65)	79.23(4.81)	82.55(3.53)
German	yes	79.85(3.05)	79.85(5.56)	79.48(5.01)	78.87(4.27)	79.20(3.15)	77.89(3.85)
	no	84.54(3.32)	82.25(5.59)	81.70(4.00)	85.90(3.99)	82.96(3.22)	85.07(3.54)
Post-operative	yes	49.11(14.07)	49.20(8.91)	49.02(11.34)	46.16(5.89)	46.31(18.77)	42.34(28.12)
	no	55.52(24.47)	65.69(21.64)	75.79(16.86)	66.45(23.29)	64.44(20.88)	45.62(32.71)
Haberman	yes	66.07(10.26)	58.36(10.26)	55.73(14.31)	64.46(10.95)	62.70(11.50)	68.01(13.99)
	no	68.40(10.17)	58.36(10.26)	55.73(14.31)	69.59(13.30)	62.03(11.82)	69.29(14.13)
Splice-ie	yes	97.46(1.10)	98.39(0.64)	97.55(0.46)	98.69(0.51)	97.37(0.84)	98.38(0.57)
	no	98.80(0.40)	99.17(0.36)	98.82(0.32)	99.18(0.43)	98.93(0.30)	99.15(0.36)
Splice-ei	yes	98.74(0.46)	98.78(0.46)	98.85(0.42)	98.78(0.46)	98.83(0.45)	98.77(0.47)
	no	99.25(0.48)	99.27(0.77)	99.47(0.27)	99.44(0.60)	99.33(0.66)	99.40(0.66)
Vehicle	yes	97.25(1.95)	98.62(0.67)	98.34(1.32)	98.26(0.90)	98.79(0.67)	97.94(1.05)
	no	97.80(0.94)	98.64(0.63)	98.42(1.02)	98.41(0.90)	98.71(0.97)	98.17(1.12)
Letter-vowel	yes	97.69(0.43)	98.03(0.37)	97.97(0.46)	98.18(0.53)	97.66(0.30)	98.17(0.30)
	no	98.26(0.28)	98.49(0.31)	98.39(0.22)	98.90(0.18)	98.27(0.19)	98.81(0.17)
New-thyroid	yes	94.87(5.00)	94.79(10.14)	94.54(10.10)	94.73(9.24)	92.72(10.55)	93.44(9.74)
	no	94.87(5.00)	94.79(10.14)	94.54(10.10)	94.98(9.38)	92.72(10.55)	93.69(9.90)
E.Coli	yes	88.75(12.45)	80.32(19.96)	80.34(19.85)	91.57(7.81)	83.97(21.27)	91.73(8.00)
	no	88.64(12.46)	81.13(20.00)	81.95(19.90)	94.03(5.56)	83.76(21.17)	92.04(8.15)
Satimage	yes	92.34(1.27)	92.25(1.45)	92.73(1.38)	94.21(1.76)	92.85(1.19)	94.42(1.53)
	no	92.86(1.29)	92.35(1.35)	92.90(1.38)	95.11(1.29)	92.84(1.22)	95.06(1.27)
Flag	yes	71.13(28.95)	49.12(21.57)	75.85(30.26)	45.00(15.81)	45.00(15.81)	44.47(15.71)
	no	78.35(29.98)	78.90(28.63)	75.64(29.37)	78.59(28.75)	81.73(29.51)	76.13(27.80)
Glass	yes	82.44(8.99)	58.44(13.15)	72.69(14.07)	87.15(16.47)	72.16(16.84)	91.67(12.76)
	no	80.47(13.25)	64.31(14.21)	75.44(11.61)	87.00(16.75)	78.76(12.52)	91.67(12.76)
Letter-a	yes	99.35(0.48)	99.60(0.37)	99.61(0.37)	99.61(0.40)	99.66(0.46)	99.60(0.40)
	no	99.46(0.42)	99.66(0.37)	99.65(0.38)	99.67(0.37)	99.67(0.45)	99.67(0.37)
Nursery	yes	97.52(0.82)	99.55(0.21)	98.77(0.35)	99.80(0.08)	99.47(0.19)	99.79(0.12)
	no	98.76(0.22)	99.84(0.13)	99.57(0.21)	99.89(0.08)	99.83(0.08)	99.89(0.09)

Second Stage
- peringkat kinerja untuk set data asli dan seimbang untuk yang dipangkas decision trees

- Warna abu-abu muda: hasil yang diperoleh dengan metode over sampling
- Warna abu-abu gelap: hasil yang diperoleh dengan kumpulan data asli
- Metode yang ditandai dengan tanda bintang memperoleh hasil yang lebih rendah secara statistik jika dibandingkan dengan metode peringkat teratas

Table 6: Performance ranking for original and balanced data sets for pruned decision trees.

Data set	1°	2°	3°	4°	5°	6°	7°	8°	9°	10°	11°
Pima	Smt	RdOvr	Smt+Tmk	Smt+ENN	Tmk	NCL	Original	RdUdr	CNN+Tmk	CNN*	OSS*
German	RdOvr	Smt+Tmk	Smt+ENN	Smt	RdUdr	CNN	CNN+Tmk*	OSS*	Original*	Tmk*	NCL*
Post-operative	RdOvr	Smt+ENN	Smt	Original	CNN	RdUdr	CNN+Tmk	OSS*	Tmk*	NCL*	Smt+Tmk*
Haberman	Smt+ENN	Smt+Tmk	Smt	RdOvr	NCL	RdUdr	Tmk	OSS*	CNN*	Original*	CNN+Tmk*
Splice-ie	RdOvr	Original	Tmk	Smt	CNN	NCL	Smt+Tmk	Smt+ENN*	CNN+Tmk*	RdUdr*	OSS*
Splice-ei	Smt	Smt+Tmk	Smt+ENN	CNN+Tmk	OSS	RdOvr	Tmk	CNN	NCL	Original	RdUdr
Vehicle	RdOvr	Smt	Smt+Tmk	OSS	CNN	Original	CNN+Tmk	Tmk	NCL*	Smt+ENN*	RdUdr*
Letter-vowel	Smt+ENN	Smt+Tmk	Smt	RdOvr	Tmk*	NCL*	Original*	CNN*	CNN+Tmk*	RdUdr*	OSS*
New-thyroid	Smt+ENN	Smt+Tmk	Smt	RdOvr	RdUdr	CNN	Original	Tmk	CNN+Tmk	NCL	OSS
E.Coli	Smt+Tmk	Smt	Smt+ENN	RdOvr	NCL	Tmk	RdUdr	Original	OSS	CNN+Tmk*	CNN*
Satimage	Smt+ENN	Smt	Smt+Tmk	RdOvr	NCL	Tmk	Original*	OSS*	CNN+Tmk*	RdUdr*	CNN*
Flag	RdOvr	Smt+ENN	Smt+Tmk	CNN+Tmk	Smt	RdUdr	CNN*	OSS*	Tmk*	Original*	NCL*
Glass	Smt+ENN	RdOvr	NCL	Smt	Smt+Tmk	Original	Tmk	RdUdr	CNN+Tmk*	OSS*	CNN*
Letter-a	Smt+Tmk	Smt+ENN	Smt	RdOvr	OSS	Original	Tmk	CNN+Tmk	NCL	CNN	RdUdr*
Nursery	RdOvr	Tmk	Original	NCL	CNN*	OSS*	Smt+Tmk*	Smt*	CNN+Tmk*	Smt+ENN*	RdUdr*

Second Stage
- peringkat kinerja untuk set data asli dan seimbang untuk yang tidak dipangkas decision trees

Table 7: Performance ranking for original and balanced data sets for unpruned decision trees.

Data set	1°	2°	3°	4°	5°	6°	7°	8°	9°	10°	11°
Pima	RdOvr	Smt	Smt+Tmk	Smt+ENN	Tmk	NCL	Original	CNN+Tmk	RdUdr	CNN*	OSS*
German	Original	Tmk	RdOvr	NCL	RdUdr	Smt	Smt+Tmk	Smt+ENN	OSS	CNN	CNN+Tmk
Post-operative	Original	CNN+Tmk	RdOvr	Smt	Tmk	CNN	OSS	Smt+ENN	RdUdr	Smt+Tmk*	NCL*
Haberman	Smt+Tmk	Smt+ENN	Smt	RdOvr	Tmk	NCL	RdUdr	Original	OSS*	CNN*	CNN+Tmk*
Splice-ie	Original	Smt	Tmk	CNN	NCL	Smt+Tmk	RdOvr	OSS*	Smt+ENN*	CNN+Tmk*	RdUdr*
Splice-ei	RdOvr	Smt	Smt+Tmk	Smt+ENN	Original	CNN+Tmk	Tmk	NCL	OSS	CNN	RdUdr
Vehicle	RdOvr	Smt	Smt+Tmk	OSS	CNN	Original	CNN+Tmk	Tmk	Smt+ENN	NCL	RdUdr*
Letter-vowel	Smt+ENN	Smt	Smt+Tmk	Tmk*	RdOvr*	NCL*	Original*	CNN*	CNN+Tmk*	OSS*	RdUdr*
New-thyroid	Smt+ENN	Smt	Smt+Tmk	RdOvr	Original	Tmk	RdUdr	CNN	CNN+Tmk	NCL	OSS
E.Coli	Smt+Tmk	Smt	Smt+ENN	Tmk	RdOvr	Original	NCL	RdUdr	OSS	CNN+Tmk*	CNN*
Satimage	Smt+ENN	Smt	Smt+Tmk	RdOvr	Tmk	NCL	Original	CNN+Tmk*	RdUdr*	OSS*	CNN*
Flag	Smt+Tmk	OSS	RdOvr	CNN	Tmk	Smt+ENN	RdUdr	Original	NCL	CNN+Tmk	Smt
Glass	Smt+ENN	RdOvr	NCL	Smt	Smt+Tmk	Original	Tmk	RdUdr	OSS*	CNN+Tmk*	CNN*
Letter-a	Smt	Smt+Tmk	Smt+ENN	RdOvr	Tmk	OSS	NCL	Original	CNN	CNN+Tmk	RdUdr*
Nursery	RdOvr	Original	NCL	Tmk	CNN	OSS*	Smt*	CNN+Tmk*	Smt+Tmk*	Smt+ENN*	RdUdr*

Third Stage

- Untuk mengukur kompleksitas sintaksis dari model yang diinduksi.
- Kompleksitas sintaksis diberikan oleh dua parameter utama:
  - jumlah rata-rata aturan yang diinduksi
  - jumlah rata-rata kondisi per aturan
- Fakta
  - Pengambilan sampel yang berlebihan menyebabkan peningkatan jumlah aturan yang diinduksi dibandingkan dengan yang diinduksi dengan set data asli
  - Over-sampling acak dan Smote+ENN memberikan peningkatan yang lebih kecil dalam jumlah rata-rata aturan
  - Smote+ENN memberikan peningkatan yang lebih kecil dalam jumlah rata-rata kondisi per aturan
- Penjelasan
  - Pengambilan sampel yang berlebihan meningkatkan jumlah total contoh pelatihan, yang biasanya menghasilkan pohon keputusan yang lebih besar

Table 8: Number of rules (branches) for the original and over-sampled data sets and unpruned decision trees.

Data set	Original	Rand Over	Smote	Smote+Tomek	Smote+ENN
Pima	**29.90(6.06)**	63.80(13.15)	57.70(11.52)	54.20(12.91)	47.50(8.76)
German	315.50(21.41)	410.60(28.64)	367.30(20.85)	355.10(24.20)	**261.00(28.08)**
Post-operative	**20.40(3.86)**	36.80(3.05)	38.60(4.35)	32.70(5.87)	25.90(4.09)
Haberman	**7.80(3.79)**	25.20(10.94)	23.20(9.61)	25.00(7.70)	30.30(4.92)
Splice-ie	**203.50(7.78)**	258.70(13.07)	443.20(16.69)	340.60(21.34)	307.90(17.21)
Splice-ei	**167.80(9.40)**	193.30(7.41)	374.50(20.41)	283.90(14.90)	248.80(12.90)
Vehicle	**26.20(3.29)**	28.90(2.60)	34.90(3.38)	34.90(3.38)	29.20(2.82)
Letter-vowel	**534.50(11.92)**	678.80(19.07)	1084.50(19.61)	1083.20(20.12)	1022.00(26.34)
New-thyroid	5.40(0.84)	**5.10(0.32)**	6.90(1.29)	6.90(1.29)	6.90(0.99)
E-coli	**11.60(3.03)**	17.70(2.91)	16.70(3.20)	16.50(3.84)	12.70(3.23)
Satimage	**198.80(11.04)**	252.70(9.33)	404.60(12.97)	404.60(12.97)	339.40(13.80)
Flag	**28.60(6.52)**	46.30(7.72)	52.50(12.47)	46.50(13.36)	40.30(9.09)
Glass	**9.40(2.22)**	13.00(1.33)	17.70(1.77)	17.70(1.77)	15.50(1.58)
Letter-a	**59.10(3.45)**	88.00(5.56)	257.60(15.42)	257.60(15.42)	252.60(18.23)
Nursery	**229.40(4.65)**	282.50(5.34)	1238.30(28.91)	1204.70(27.94)	766.30(77.24)

Third Stage
- jumlah rata-rata aturan yang diinduksi

- Hasil terbaik ditampilkan dalam huruf tebal
- Hasil terbaik yang diperoleh dengan metode over-sampling disorot dalam warna abu-abu muda

Table 9: Mean number of conditions per rule for the original and over-sampled data sets and unpruned decision trees.

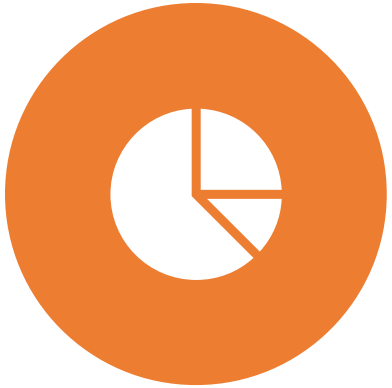
Data set	Original	Rand Over	Smote	Smote+Tomek	Smote+ENN
Pima	**6.21(0.61)**	7.92(0.64)	7.74(0.44)	7.59(0.54)	7.27(0.67)
German	**6.10(0.17)**	6.89(0.25)	10.27(0.51)	9.68(0.32)	7.35(0.58)
Post-operative	**3.61(0.41)**	4.86(0.26)	5.36(0.37)	4.75(0.52)	4.46(0.50)
Haberman	**3.45(1.36)**	5.71(1.43)	5.61(1.27)	5.81(1.02)	6.45(0.60)
Splice-ie	6.04(0.09)	6.15(0.04)	6.08(0.08)	6.00(0.09)	**5.58(0.11)**
Splice-ei	5.46(0.14)	5.70(0.08)	5.51(0.07)	5.41(0.09)	**4.91(0.09)**
Vehicle	7.21(0.70)	7.03(0.44)	7.09(0.50)	7.09(0.50)	**6.63(0.38)**
Letter-vowel	20.96(1.19)	19.32(0.82)	18.78(0.40)	18.78(0.40)	**18.32(0.43)**
New-thyroid	**2.76(0.39)**	2.85(0.17)	3.12(0.26)	3.12(0.26)	3.08(0.20)
E-coli	4.43(0.79)	5.48(0.41)	4.98(0.60)	4.92(0.65)	**4.15(0.49)**
Satimage	**12.13(0.46)**	15.93(0.42)	13.89(0.64)	13.89(0.64)	12.54(0.36)
Flag	**3.92(0.70)**	5.42(0.55)	9.43(1.04)	8.75(1.53)	6.71(1.23)
Glass	**4.20(0.61)**	5.80(0.51)	5.92(0.50)	5.92(0.50)	5.51(0.32)
Letter-a	**7.30(0.22)**	10.35(0.64)	10.97(0.38)	10.97(0.38)	10.86(0.36)
Nursery	6.51(0.01)	6.84(0.03)	6.87(0.03)	6.84(0.03)	**6.41(0.12)**

Third Stage
- jumlah rata-rata kondisi per aturan

Kesimpulan

- Ketidakseimbangan kelas tidak secara sistematis menghambat kinerja sistem pembelajaran.
- Selain ketidakseimbangan kelas, tingkat tumpang tindih data antar kelas merupakan faktor lain yang menyebabkan penurunan kinerja algoritma pembelajaran.
- Eksperimen menunjukkan bahwa secara umum, metode over-sampling memberikan hasil yang lebih akurat daripada metode under-sampling mengingat berada di bawah kurva ROC.
- Random over-sampling sangat bersaing dengan metode over-sampling yang lebih kompleks.
- Pengambilan sampel berlebih secara acak biasanya menghasilkan peningkatan terkecil dalam jumlah rata-rata aturan yang diinduksi, jika dibandingkan di antara metode pengambilan sampel berlebih.
- Smote+ENN menghasilkan peningkatan terkecil dalam jumlah rata-rata kondisi per aturan, jika dibandingkan di antara metode over-sampling.

Title Lorem Ipsum



LOREM IPSUM DOLOR SIT AMET, CONSECTETUER ADIPISCING ELIT.

NUNC VIVERRA IMPERDIET ENIM. FUSCE EST. VIVAMUS A TELLUS.

PELLENTESQUE HABITANT MORBI TRISTIQUE SENECTUS ET NETUS.