ISBN : 978.602.361.002.0

Algoritma AdaBoost
Dalam
Pengklasifikasian

Zulhanif

Staf Pengajar Jurusan Statistika FMIPA, Unpad Bandung

Email : dzulhanif@yahoo.com

ABSTRAK

Metode AdaBoost merupakan salah satu algoritma *supervised* pada data mining yang diterapakan secara luas untuk membuat model klasifikasi. AdaBoost sendiri pertama kali diperkenalkan oleh Yoav Freund dan Robert Schapire(1995). Walaupun pada awalnya algoritma ini diterapkan pada model regresi, seiring dengan perkembangan teknologi komputer yang cepat, metode ini juga dapat diterapkan pada model statistik lainnnya. Metode adaBoost merupakan salah satu teknik ensamble dengan menggunakan *loss function* fungsi *exponential* untuk memperbaiki tingkat akurasi dari prediksi yang dibuat. Pada makalah ini akan akan dijelaskan penerapan metode AdaBoostdalam masalah pengklasifikasian dengan tujuan untuk memperbaiki tingkat akurasi model yang dibentuk.

Kata Kunci : Boosting,Klasifikasi,AdaBoost.

ISBN : 978.602.361.002.0

1. PENDAHULUAN

Ada dua budaya dalam penggunaan pemodelan statistik untuk mencapai kesimpulan dari data. Budaya pertama merupakan budaya pertama adalah *The Data Modeling Culture* yang merupakan budaya sebagian besar ahli statistik saat ini, pada kelompok *Data Modelling Culture* statistikawan mengasumsikan model stokastik tertentu untuk menjelaskan mekanisme suatu data, sedangkan budaya kedua yang merupakan kelompok *Algorithmic Modeling Culture*merupakan kelompok statistikawan yang menggunakan model algoritma untuk menjelaskan mekanisme suatu data. Sebagai ilustrasi bahwa seorangEkonom dalam komunitas data mining akan berbeda dalam pendekatan mereka untukanalisis regresi. Para ekonom diluar komunitas data mining akan membangun sebuah model regresi dari teori dan kemudian menggunakan model tersebut untuk menjelaskan bentuk hubungan yang tejadi, hal ini berbeda untuk ekonom pada komunitas data mining, pada kominitas ini ekonom akan menggunakan prinsip "*kitchen sink*" yang mana suatu pendekatan yang menggunakansebagian besar regressor atau x-variabel yang tersedia akan dipergunakan dalam model regresi. Karena pemilihan x-variabel tidak didukung olehteori, validasi model regresi menjadi hal yang sangat penting. Pendekatan standar untukmemvalidasi model regresi pada data mining adalah dengan cara membagi data ke dalam data pelatihan dan datapengujian.Konsep dataset pelatihan versus dataset pengujian merupakan inti dari algoritma *supervised* pada data mining. Pada prinsipnya model yang dibentuk pada data pelatihan akan dipergunakan untuk membuat membuat prediksi pada data pengujin. Tujuan dari proses ini adalah agar model tidak *overfitted* serta dapat digeneralisasi. Pada makalah ini akan dijelaskan model boosting dalam pengklasifikasian berdasarakan algoritma AdaBoost yang dikembangkan oleh Yoav Freund dan Robert Schapire [3].

2. METODE PENELITIAN

Algoritma AdaBoost sendiri merupakan akronim dari Adaptive Boosting, algoritma ini diterapkan secara luas pada model prediksi dalam data mining. Inti dari algoritma AdaBoost adalah memberikan suatu bobot lebih pada observasi yang tidak tepat (*weak classification*). Boosting sendiri pada dasarnya adalah *m* buah kombinasi liniear dari $\mathrm{m}\, k_m(x_i)\, classifier$dengan fungsi *classifier* dimisalkan $k_m(x_i) \in \{-1,1\}$, sehingga kombinasi linear dari $mclassifier$ dapat dinyatakan sbb:

ISBN : 978.602.361.002.0

$$C_{(m-1)}(x_i) = \alpha_1 k_1(x_i) + \alpha_2 k_2(x_i) + \mathrm{K} + \alpha_{m-1} k_{m-1}(x_i) \quad (2.1)$$

Dalam bentuk yang lebih umum dapat dinyatakan dalam persamaan sbb:

$$C_m(x_i) = C_{(m-1)}(x_i) + \alpha_m k_m(x_i) \quad (2.2)$$

Pada algoritma AdaBoost didefinisikantotal*cost*, atau total*error*, dari*classifier*sebagai fungsieksponensial sbb:

$$E = \sum_{i=1}^{N} \exp(-y_i (C_{(m-1)}(x_i) + \alpha_m k_m(x_i)) \quad (2.3)$$

Persamaan 3 dapat ditulis juga dalam sebuah persamaan sbb:

$$E = \sum_{i=1}^{N} w_i^{(m)} \exp(-y_i (C_{(m-1)}(x_i) + \alpha_m k_m(x_i)) \quad (2.4)$$

Dengan

$$w_i^{(m)} = \exp(-y_i (C_{(m-1)}(x_i))$$

Persamaan 4 sendiri dapat dinyatakan dalam dua persamaan sbb:

$$E = \sum_{y_i = k_m(x_i)} w_i^{(m)} \exp(-\alpha_m) + \sum_{y_i \neq k_m(x_i)} w_i^{(m)} \exp(\alpha_m) \quad (2.5)$$

Jika persamaan pertama untuk kasus individu diklasifikasikan dengan benar dinyatakan $W_c \exp(-\alpha_m)$ sebagai dan yang tidak benar dinyatakan dnegan $W_e \exp(\alpha_m)$ maka persamaan 4 dapat ditulis sbb

$$E = W_c \exp(-\alpha_m) + W_e \exp(\alpha_m) \quad (2.6)$$

ISBN : 978.602.361.002.0

Dari persamaan 6 terihat bahwa untuk meminimumkan fungsi *E* akan dicari nilai bobot α_m yang optimum yang dihitung dengan cara sbb:

$$\frac{dE}{d\alpha_m} = -W_c \exp(-\alpha_m) + W_e \exp(\alpha_m) \tag{2.7}$$

$$-W_c + W_e \exp(2\alpha_m) = 0$$

$$\alpha_m = \frac{1}{2}\ln(\frac{W_c}{W_e})$$

$$\alpha_m = \frac{1}{2}\ln(\frac{W - W_e}{W_e}) = \frac{1}{2}\ln(\frac{1-e_m}{e_m}) \text{ dengan } e_m = \frac{W_e}{W}$$

ISBN : 978.602.361.002.0

Perumusan algoritma AdaBoost secara lengkap dapat dijabarkan sbb:

Untuk m=1 to M

1. Minimumkan fungsi error $W_e = \sum_{y_i \neq k_m(x_i)} w_i^{(m)} \exp(\alpha_m)$
2. Set $\alpha_m = \frac{1}{2}\ln(\frac{1-e_m}{e_m})$ yang mana $e_m = \frac{W_e}{W}$
3. Update nilai $w_i^{(m+1)} = w_i^{(m)} \exp(\alpha_m) = w_i^{(m)} \sqrt{\frac{1-e_m}{e_m}}$ jika pengamatan itu *missclasisfication* dan $w_i^{(m+1)} = w_i^{(m)} \exp(-\alpha_m) = w_i^{(m)} \sqrt{\frac{e_m}{1-e_m}}$ untuk lainnnya

3. HASIL DAN PEMBAHASAN

Sebagai ilustrasi dalam penggunaan algoritma Adaboost dengan fungsi *classifier*regression treeini akan dipergunakan data penelitian Database kanker payudara ini diperoleh dariRumah Sakit*University of Wisconsin* , Madison dari Dr. William H. Wolberg . Ia menilai biopsi tumor payudara untuk 699 pasien hingga 15 Juli 1992 ; masing-masing sembilan atribut telah dinilai pada skala 1 sampai 10 , sehingga terdapat 699 pengatamat celss dengan 9atribut, selanjutnayatahapan analisis dalam membuat model klasifikasinya adalah sbb :

1. Membagi data yang dipergunakan menjadi dua bagian yang terdiri atas data training dan data testing dengan perbandingan 80% dan 20%
2. Mengevaluasi besarnya kesalahan klasifikasi dari data training dan data testing
3. Membuat model prediksi

Dengan bantuan software didapat model klasifikasi dengan tingkat kepentingan variabel prediktor dan akurasi klasifikasi sbb:

ISBN : 978.602.361.002.0

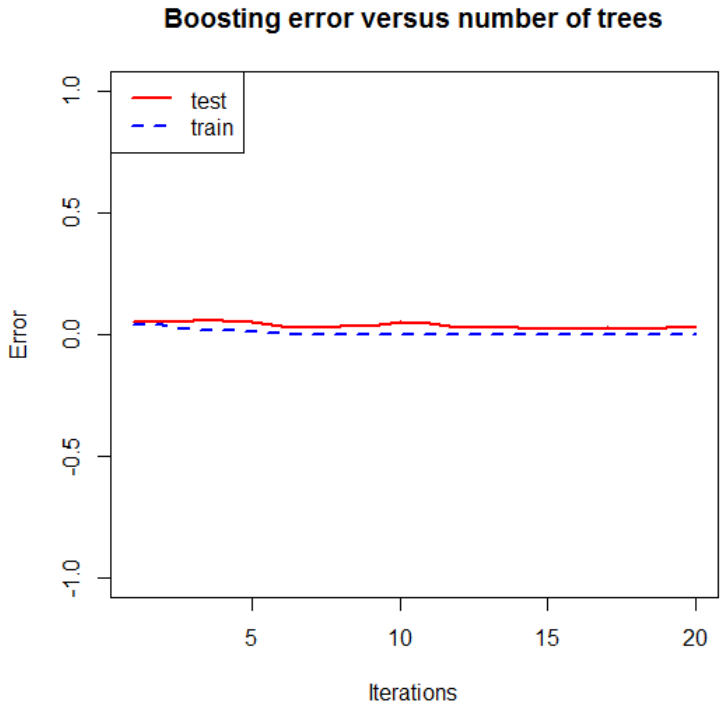
3.1.1)Jumlah Boosting Optimal

Tabel 1 Tingkat Kepentingan Variabel Prediktor

Variabel Prediktor	Impotance
Cell.size	20.923363
Bare.nuclei	20.239509
Normal.nucleoli	18.733969
Cl.thickness	10.158718
Bl.cromatin	9.387055
Marg.adhesion	8.721022
Epith.c.size	5.535636
Cell.shape	3.174272
Mitoses	3.126455

3.1.2) Jumlah Boosting Optimal

ISBN : 978.602.361.002.0



Gambar 1 Jumlah Optimal Boosting

3.2.1) Tingkat Akurasi Klasifikasi tanpa Boosting

Tingkat akurasi klasifikasi ditinjau dari data training dan data testing sbb:

3.2.1.1 Data Training

Tabel 2 Tabel Klasifikasi Data Training

	Benign	**Malignant**	**Total**
Benign	360	4	364
Malignant	17	179	196

ISBN : 978.602.361.002.0

Total	377	183	560

Dari data training pada tabel 2 dapat dilihat tingkat misklasifikasi sebesar 3.75%

3.2.1.2 Data Testing

Tabel 2 Tabel Klasifikasi Data Testing

	Benign	**Malignant**	**Total**
Benign	75	1	76
Malignant	6	57	63
Total	81	58	139

Dari data training pada tabel1 dapat dilihat tingkat misklasifikasi sebesar 5.04%

3.3.1 Tingkat Akurasi Klasifikasi dengan Boosting

Tingkat akurasi klasifikasi ditinjau dari data training dan data testing sbb:

3.3.1.1 Data Training

Tabel 3 Tabel Klasifikasi Data Training

ISBN : 978.602.361.002.0

	Benign	Malignant	Total
Benign	377	0	377
Malignant	0	183	183
Total	377	183	560

Dari data training pada tabel 3 dapat dilihat tingkat misklasifikasi sebesar 0%

3.2.1.2 Data Testing

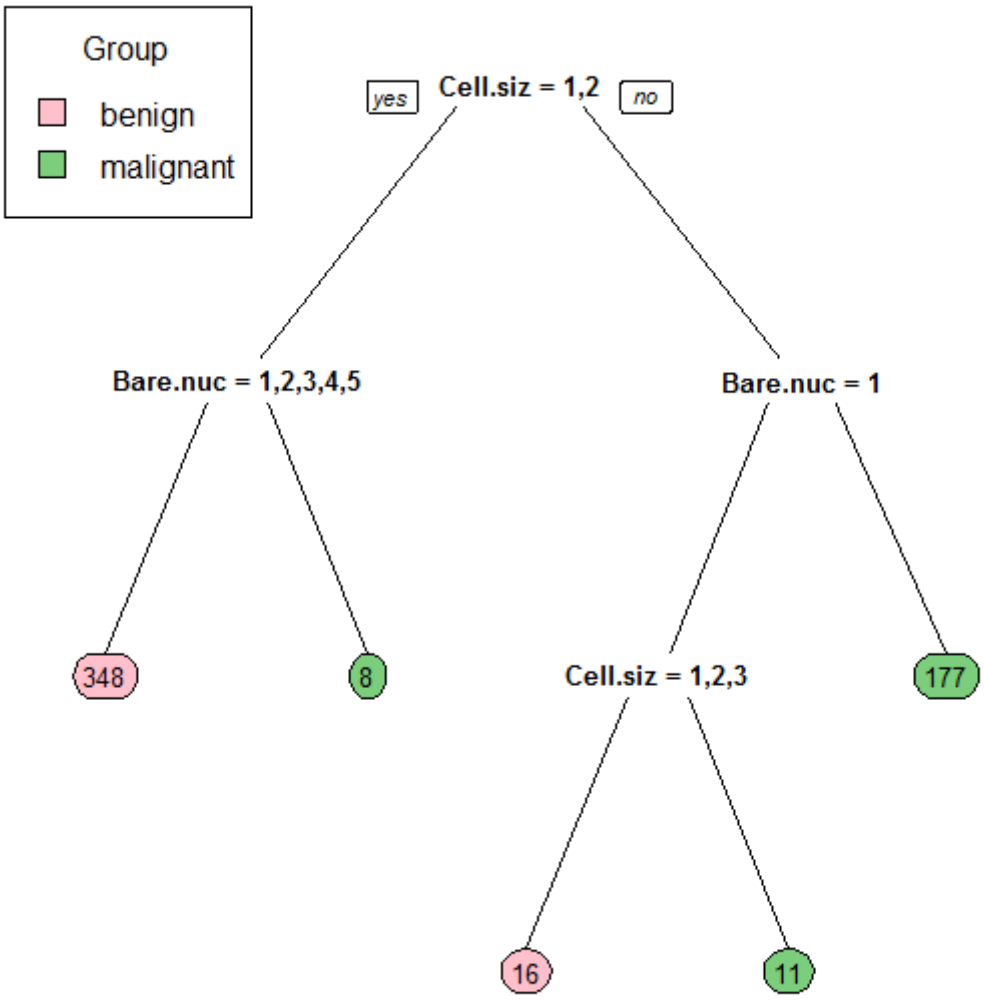
Tabel 4 Tabel Klasifikasi Data Testing

	Benign	Malignant	Total
Benign	77	0	77
Malignant	4	58	62
Total	81	58	139

Dari data training pada tabel 4 dapat dilihat tingkat misklasifikasi sebesar 2.88%

ISBN : 978.602.361.002.0

3.4 Model Boosting



4. SIMPULAN

ISBN : 978.602.361.002.0

Hasil analisis menunjukkan adanya kekurang akuratan hasil klasifikasi pada data testing yang berpotensi meyebabkan *over fiitng* dari model klasifikasi yang dibentuk.Pemodelan klasifikasi dengan metode ini perlu diuji lagi berkenaan dengan mempertimbangkan menggunakan *loss function* selain fungsi *exponential*.Pereduksian jumlah varabel prediktor menjadi hal yang dapat dipertimbangkan untuk megurangi kesalahan dari model kalaifikasi yang dibuat.

5.DAFTAR PUSTAKA

[1] Bauer, E. and R. Kohavi. 1999. An empirical comparison of voting classification algo- rithms: Bagging, boosting, and variants. Machine Learning 36: 105–139.

[2] Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. Classification and Regression Trees. Belmont, CA: Wadsworth.

[3] Freund, Y. and R. E. Schapire. 1997. A decision-theoretic generalization of online learning and an application to boosting. Journal of Computer and System Sciences 55(1): 119–139.

[4] Friedman, J. 2001. Greedy function approximation: a gradient boosting machine. An-nals of Statistics 29: 1189–1232.

[5] Friedman, J., T. Hastie, and R. Tibshirani. 2000. Additive logistic regression: a statis-tical view of boosting. Annals of Statistics 28: 337–407.

[6] Hastie, T., R. Tibshirani, and J. Friedman. 2001. The Elements of Statistical Learning New York: Springer.

[7] Long, J. S. and J. Freese. 2003. Regression Models for Categorical Dependent Variables Using Stata. rev. ed. College Station, TX: Stata Press.

[8] McCullagh, P. and J. A. Nelder. 1989. Generalized Linear Models. 2nd ed. London:Chapman & Hall.