

マイノリティクラスを重視した ルールアンサンブル法の研究

情報認識学研究室

横山 祐也

(yoko-yokodayo@ist.hokudai.ac.jp)

背景

- ・パターン認識ではたびたび、クラスインバランスが問題となる。

[Chawla et al., 2002] [Drummond et al., 2003] [Qi et al., 2018]

例: 癌が陽性か否かの分類問題

識別モデル1

		予測	
		陽性	陰性
真値	陽性	1	9
	陰性	1	89
識別率		90%	

識別モデル2

		予測	
		陽性	陰性
真値	陽性	7	3
	陰性	27	63
識別率		70%	

陽性: マイノリティクラス

陰性: マジョリティクラス

医療の分野などでは、マイノリティクラスの発見が重要である。

従来手法

データセットの修正	Minority Oversampling [Chawla et al., 2002] Majority Under Sampling [Drummond et al., 2003] など
学習アルゴリズムの修正	決定木の改良 [Dietterich et al., 1996] ニューラルネットワークの改良 [Qi et al., 2018] など

それぞれの問題として、

過学習の恐れ, 情報の損失.

マイノリティクラスの識別を重視した設計・評価をできていない

結果の解釈が困難

目的と提案

- ・マイノリティクラスの識別を重視した学習アルゴリズムの修正と評価を行う。
 - 識別モデルに、識別率ではなくバランスされた精度を最大化する工夫をする。
- ・解釈可能性の向上
 - 解釈可能性のため識別モデルはルールアンサンブル法 [Friedman et al., 2008]を用いる。
 - インバランスデータに適したルール評価を行う。(詳細は後述)

バランスされた精度 [Qi et al., 2018]

$$\text{B.A.} = \frac{1}{|C|} \sum_{c \in C} P(c \rightarrow c) \quad (C = \{c_1, c_2, \dots\}, c_i \text{ は各クラス})$$

例: 癌が陽性か否か.

B.A.はバランスされた精度

識別モデル1

		予測	
		陽性	陰性
真値	陽性	1	9
	陰性	1	89

識別モデル2

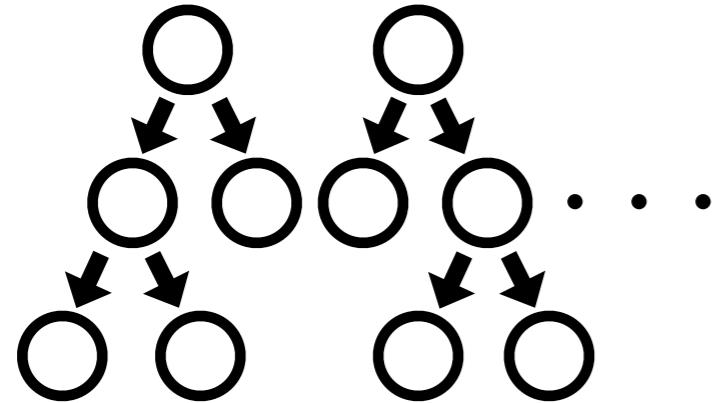
		予測	
		陽性	陰性
真値	陽性	7	3
	陰性	27	63

	陽性の識別率	陰性の識別率	B.A.	識別率
識別モデル1	10%	99%	55%	90%
識別モデル2	70%	70%	70%	70%

ルールアンサンブル [Friedman et al.,2008]

トレーニングデータを $\{x_i, y_i\}_1^N$ とする.

($x_i \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$)



木アンサンブル構築

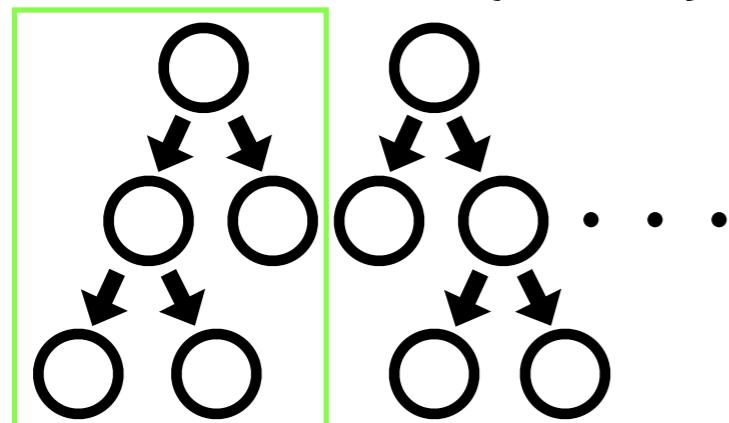
*各決定木は、データ N 個のうち
ある割合でサブサンプリングする。

分割基準は情報エントロピーを
用いて計算する。

ルールアンサンブル [Friedman et al.,2008]

トレーニングデータを $\{x_i, y_i\}_1^N$ とする.

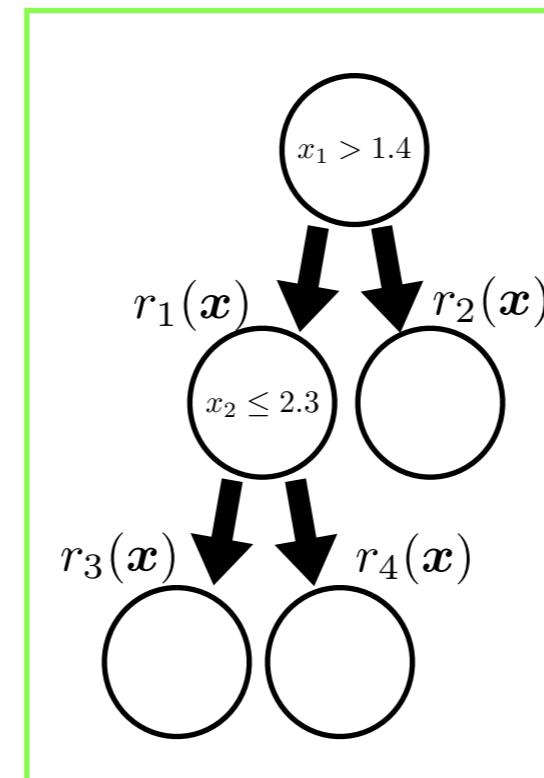
($x_i \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$)



木アンサンブル構築

*各決定木は、データ N 個のうちある割合でサブサンプリングする。

分割基準は情報エントロピーを用いて計算する。



ルール抽出

$$r_1(\mathbf{x}) = I(x_1 > 1.4)$$

$$r_2(\mathbf{x}) = I(x_1 \leq 1.4)$$

$$r_3(\mathbf{x}) = I(x_1 > 1.4) \cdot I(x_2 \leq 2.3)$$

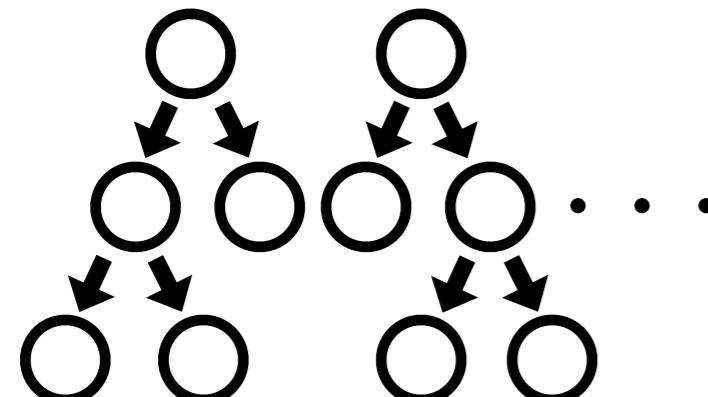
$$r_4(\mathbf{x}) = I(x_1 > 1.4) \cdot I(x_2 > 2.3)$$

$$* r_k(\mathbf{x}) \in \{0, 1\}$$

ルールアンサンブル [Friedman et al.,2008]

トレーニングデータを $\{\underline{x}_i, y_i\}_1^N$ とする.

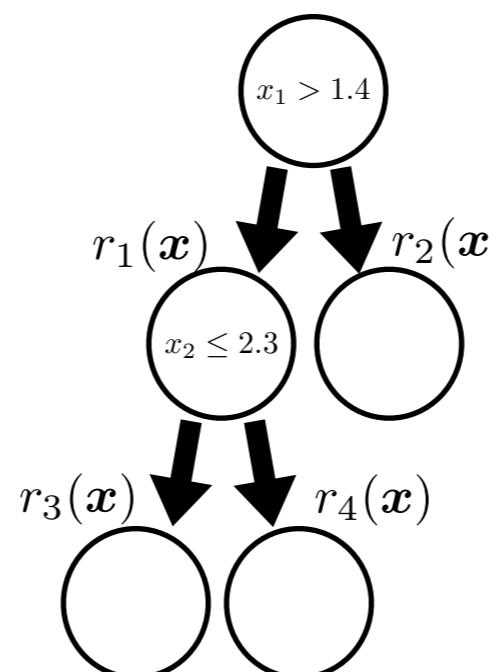
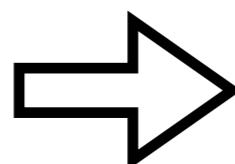
($x_i \in \mathbb{R}^n, y_i \in \{-1, 1\}$)



木アンサンブル構築

*各決定木は、データ N 個のうちある割合でサブサンプリングする。

分割基準は情報エントロピーを用いて計算する。



ルール抽出

$$r_1(\mathbf{x}) = I(x_1 > 1.4)$$

$$r_2(\mathbf{x}) = I(x_1 \leq 1.4)$$

$$r_3(\mathbf{x}) = I(x_1 > 1.4) \cdot I(x_2 \leq 2.3)$$

$$r_4(\mathbf{x}) = I(x_1 > 1.4) \cdot I(x_2 > 2.3)$$

* $r_k(\mathbf{x}) \in \{0, 1\}$

予測式: F

$$F(\mathbf{x}; \{a_k\}, \{b_j\}) = a_0 + \sum_{k=1}^K a_k \underline{r_k(\mathbf{x})} + \sum_{j=1}^n b_j \underline{l_j(x_j)}$$

回帰係数 a_k, b_j は $\sum_{i=1}^N ||y_i - F(x_i; \{a_k\}, \{b_j\})||^2$ を最小化するように決定する。

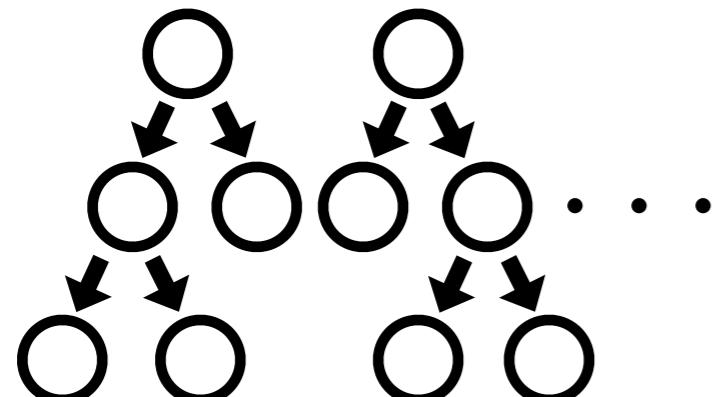
$$l_j(x_j) = \min(\delta_j^+, \max(\delta_j^-, x_j))$$

δ_j^+, δ_j^- は x_j の $1 - q, q$ 分位数

ルールアンサンブルの改良

トレーニングデータを $\{x_i, y_i\}_1^N$ とする.

$$(x_i \in \mathbb{R}^n, y_i \in \{-1, 1\})$$



木アンサンブル構築

*各決定木は、データ N 個のうちある割合でサブサンプリングする。分割基準は情報エントロピーを用いて計算する。

予測式: F

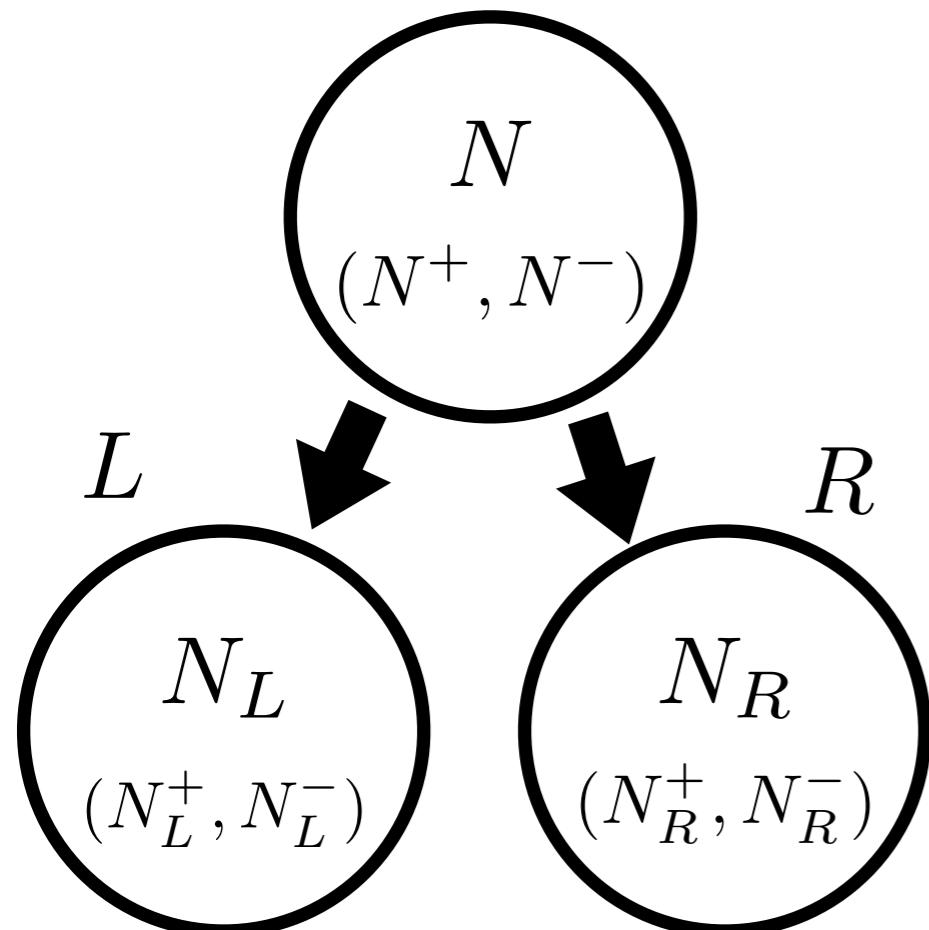
$$F(x; \{a_k\}, \{b_j\}) = a_0 + \sum_{k=1}^K a_k r_k(x) + \sum_{j=1}^n b_j l_j(x_j)$$

回帰係数 a_k, b_j は $\sum_{i=1}^N ||y_i - F(x_i; \{a_k\}, \{b_j\})||^2$ を最小化するように決定する.

$$l_j(x_j) = \min(\delta_j^+, \max(\delta_j^-, x_j)) \quad \delta_j^+, \delta_j^- \text{ は } x_j \text{ の } 1-q, q \text{ 分位数}$$

分割基準のための損失関数

$C \in \{+, -\}$, $D \in \{L, R\}$, L : 左の子ノード , R : 右の子ノードとする.



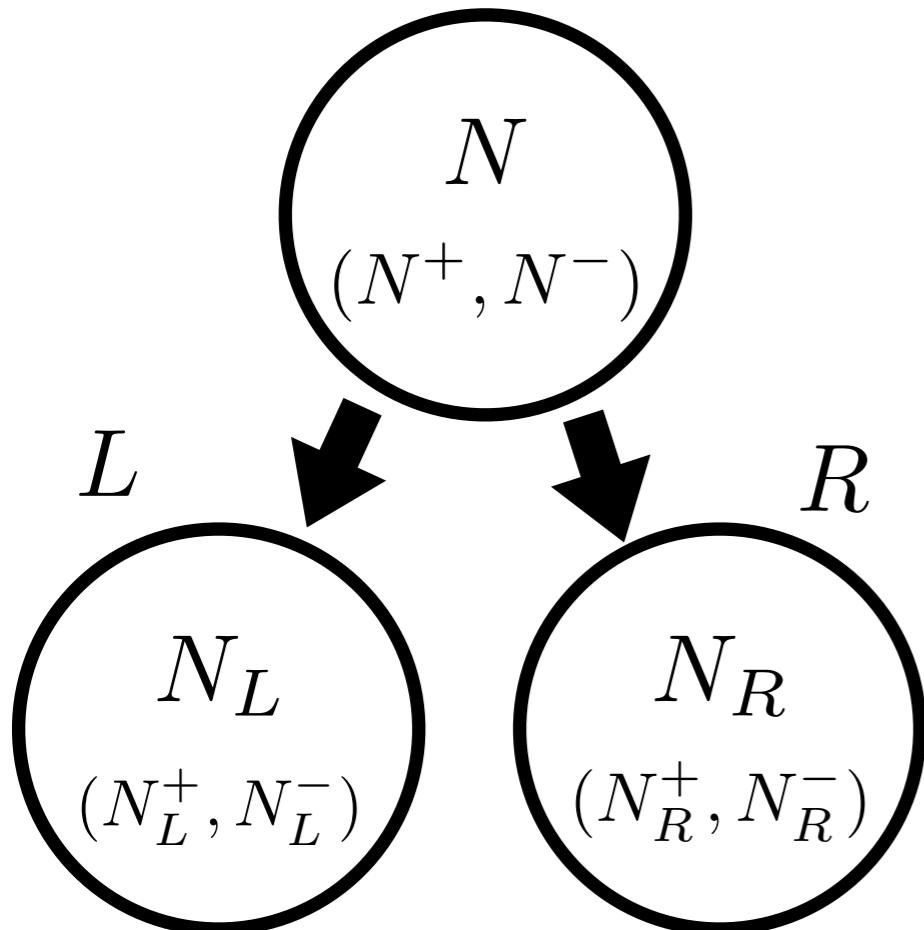
従来 $I(C; D) = H(C) - H(C|D)$

提案 $I(C; D) = H(D) - H(D|C)$ を利用

*ただし, $H(p, q) = -p \log_2 p - q \log_2 q$ とする.

分割基準のための損失関数

$C \in \{+, -\}$, $D \in \{L, R\}$, L : 左の子ノード , R : 右の子ノードとする.



従来 $I(C; D) = H(C) - H(C|D)$

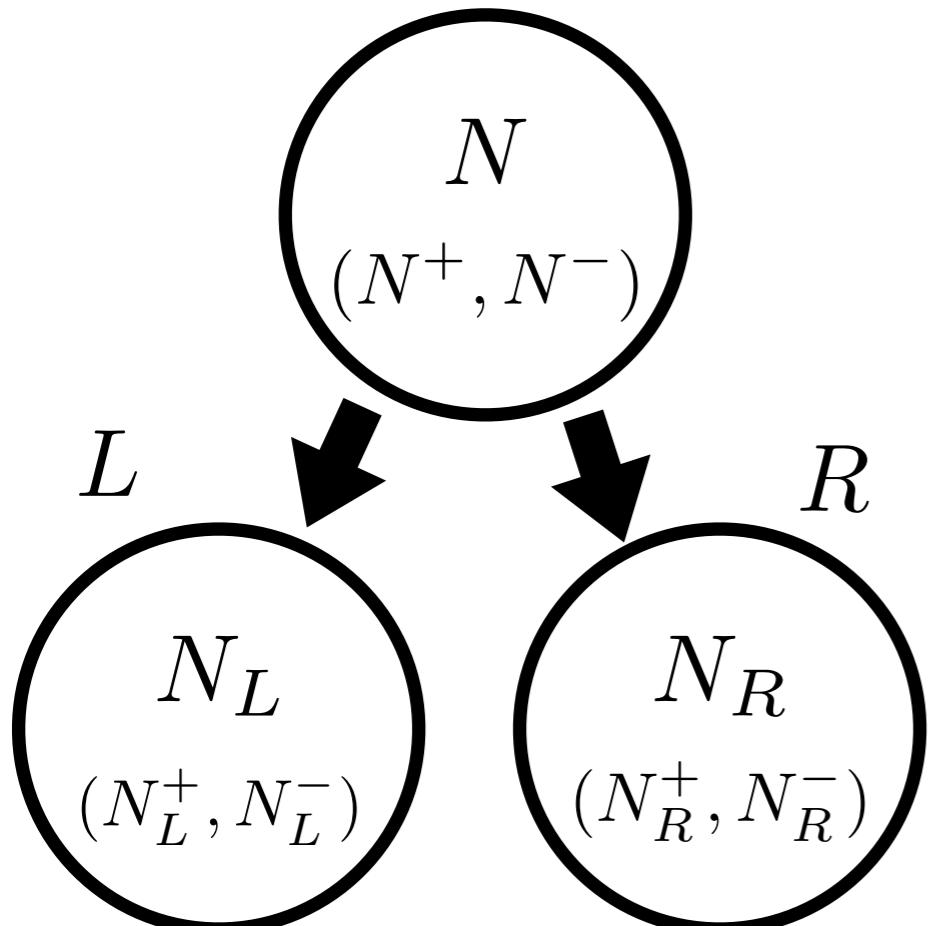
$$= H\left(\frac{N^+}{N}, \frac{N^-}{N}\right) - \left(\frac{N_L}{N} H\left(\frac{N_L^+}{N_L}, \frac{N_L^-}{N_L}\right) + \frac{N_R}{N} H\left(\frac{N_R^+}{N_R}, \frac{N_R^-}{N_R}\right)\right)$$

提案 $I(C; D) = H(D) - H(D|C)$ を利用

*ただし, $H(p, q) = -p \log_2 p - q \log_2 q$ とする.

分割基準のための損失関数

$C \in \{+, -\}$, $D \in \{L, R\}$, L : 左の子ノード, R : 右の子ノードとする.



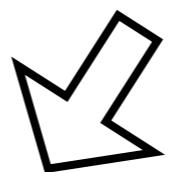
$$H \left(\frac{1}{2} \left(\frac{N_L^+}{N^+} + \frac{N_L^-}{N^-} \right), \frac{1}{2} \left(\frac{N_R^+}{N^+} + \frac{N_R^-}{N^-} \right) \right) - \left(\frac{1}{2} H \left(\frac{N_L^+}{N^+}, \frac{N_R^+}{N^+} \right) + \frac{1}{2} H \left(\frac{N_L^-}{N^-}, \frac{N_R^-}{N^-} \right) \right)$$

従来 $I(C; D) = H(C) - H(C|D)$

$$= H\left(\frac{N^+}{N}, \frac{N^-}{N}\right) - \left(\frac{N_L}{N} H\left(\frac{N_L^+}{N_L}, \frac{N_L^-}{N_L}\right) + \frac{N_R}{N} H\left(\frac{N_R^+}{N_R}, \frac{N_R^-}{N_R}\right) \right)$$

提案 $I(C; D) = H(D) - H(D|C)$ を利用

$$= H\left(\frac{N_L}{N}, \frac{N_R}{N}\right) - \left(\frac{N^+}{N} H\left(\frac{N_L^+}{N^+}, \frac{N_R^+}{N^+}\right) + \frac{N^-}{N} H\left(\frac{N_L^-}{N^-}, \frac{N_R^-}{N^-}\right) \right)$$

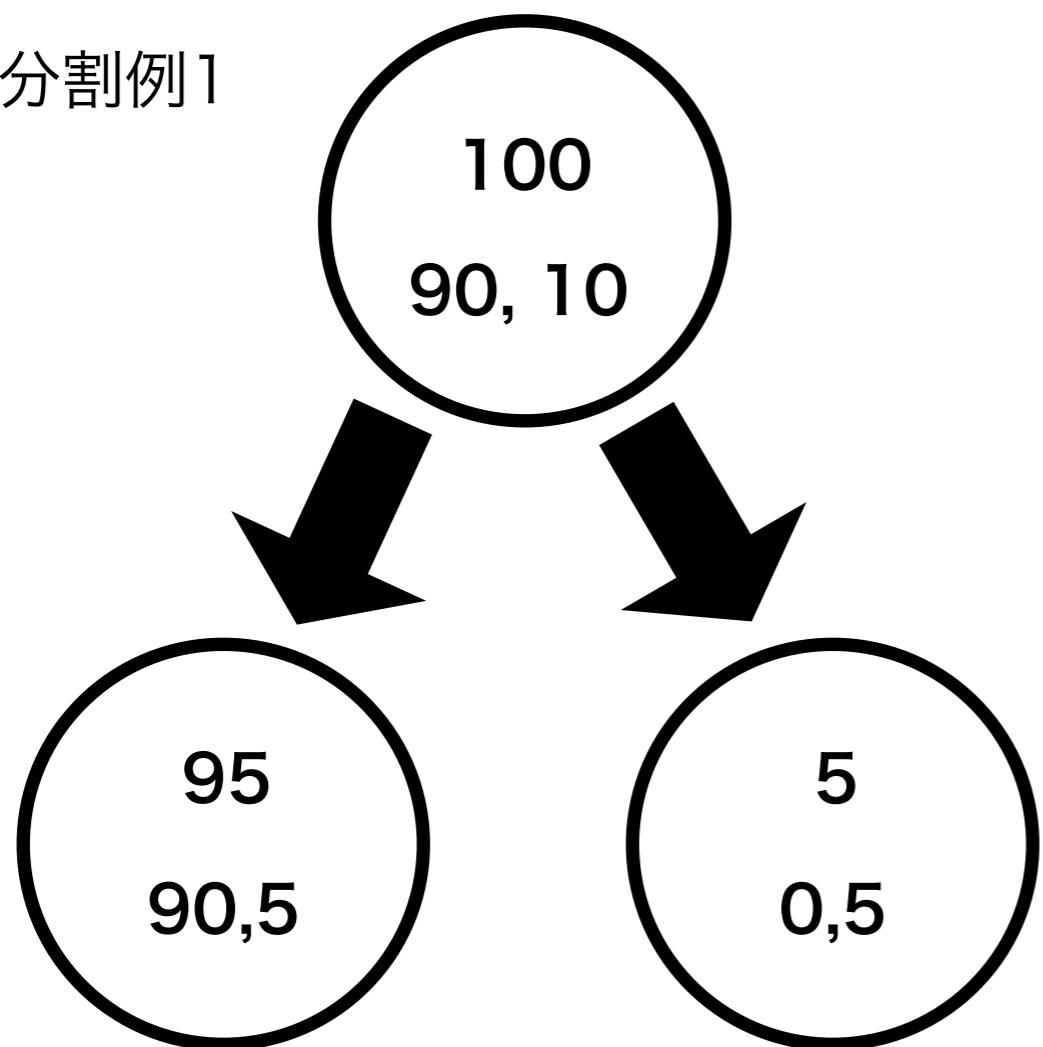


$$\frac{N^+}{N} = \frac{N^-}{N} = \frac{1}{2}$$
 (事前確率)を均等とするため,

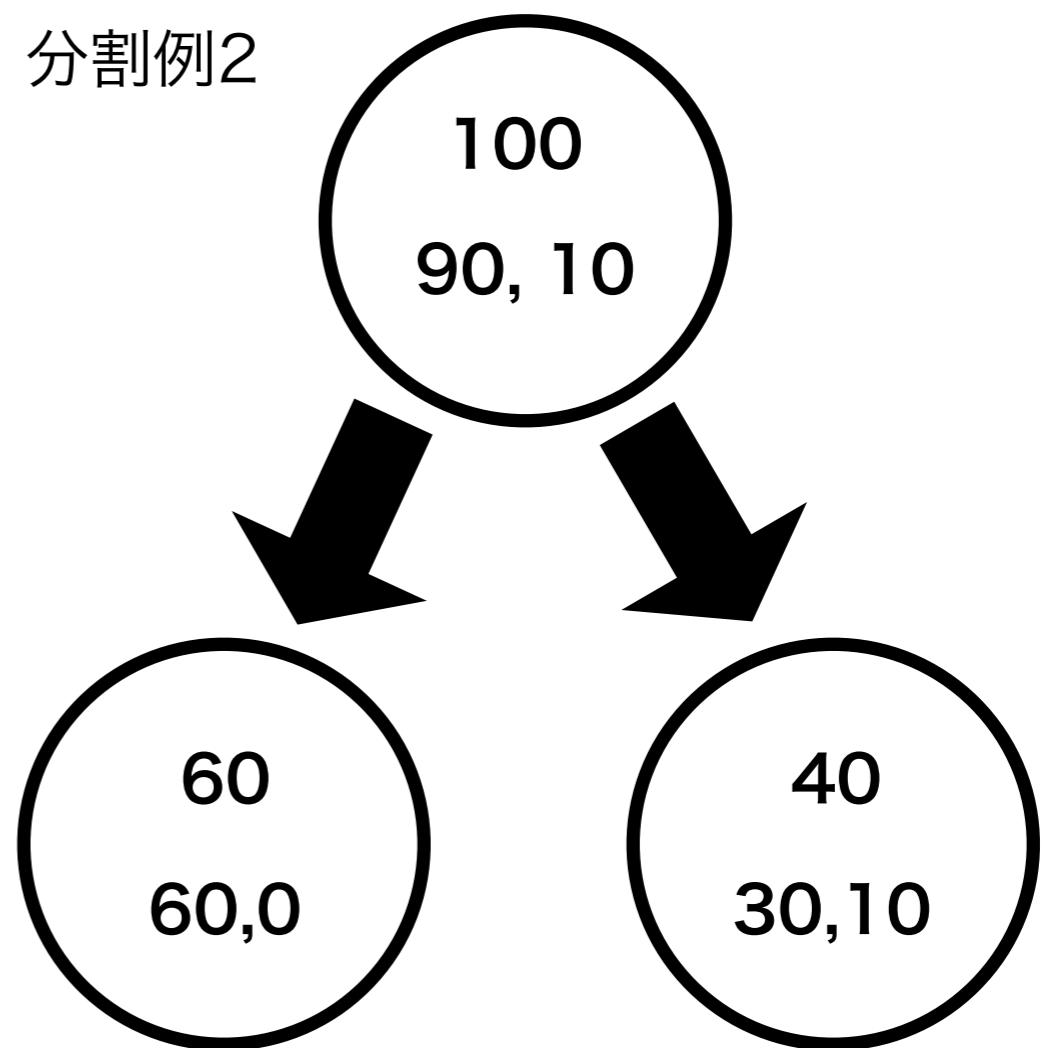
*ただし, $H(p, q) = -p \log_2 p - q \log_2 q$ とする.

損失関数の計算例

分割例1



分割例2



分割基準値

従来法

0.186

>

分割基準値

0.144

提案法

0.311

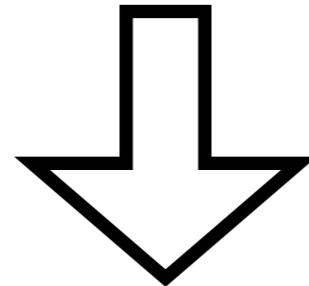
<

0.459

回帰係数の決定

従来

$$\sum_{i=1}^N ||y_i - F(x_i; \{a_k\}, \{b_j\})||^2$$



提案

$$\sum_{i=1}^N \frac{1}{P(y_i)} ||y_i - F(x_i; \{a_k\}, \{b_j\})||^2$$

F : 予測式

a_k, b_j : 回帰係数

y_i : 真値 ($y_i \in \{-1, 1\}$)

$P(y_i)$: y_i のデータ数

重要度の尺度

従来のルールの重要度の尺度 [Friedman et al., 2008]

$$I_k = |a_k| \sigma$$

$$\bar{r}_k = \frac{1}{N} \sum_{i=1}^N r_k(\mathbf{x}_i)$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (r_k(\mathbf{x}_i) - \bar{r}_k)^2$$

提案するルールの重要度の尺度

$$I_k = |a_k| \sigma$$

$$\bar{r}_k = \frac{\sum_{i=1}^N \frac{1}{P(y_i)} r_k(\mathbf{x}_i)}{\sum_{i=1}^N \frac{1}{P(y_i)}}$$

$$\sigma^2 = \frac{\sum_{i=1}^N \frac{1}{P(y_i)} (r_k(\mathbf{x}_i) - \bar{r}_k)^2}{\sum_{i=1}^N \frac{1}{P(y_i)}}$$

$$r_k(\mathbf{x}) : ルール \quad (r_k(\mathbf{x}) \in \{0, 1\})$$

実験

使用したUCIのデータセット [Newman et al., 2007]

データ名	クラス数	データ数	特徴の数	クラス比(M:m)
Glass	6	214	9	0.32:0.32: 0.13
				0.07:0.06:0.04
Block	5	5473	10	0.91:0.06:0.01
				0.01:0.01

パラメータ設定

*太字はマイノリティクラス

[Friedman.et al.,2003], [Friedman.et al,2008]を参考に

木の数: **500**

葉ノードの数: **8**

サブサンプリング数: $\min\left\{\frac{N}{2}, 100 + 6\sqrt{N}\right\}$

結果

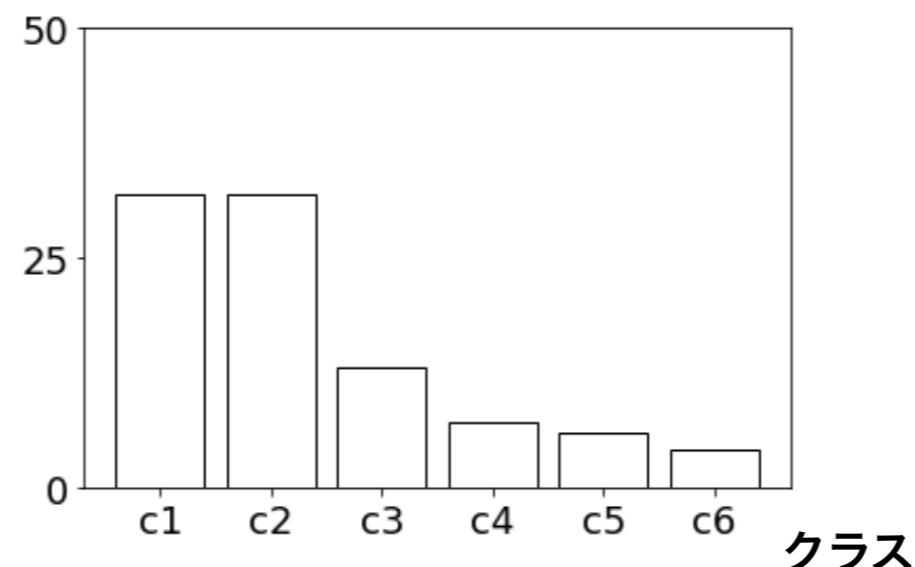
glassデータセット

(B.A.はバランスされた精度)

*不要なインバランスを避けるため, 1 vs. 1で学習を行なった.

B.A.(%)		クラス毎識別率					
		Majority			Minority		
		c_1	c_2	c_3	c_4	c_5	c_6
通常	66	0.79	0.58	0.21	0.62	0.93	0.86
提案	67	0.74	0.60	0.21	0.62	1.00	0.86

データ数の割合



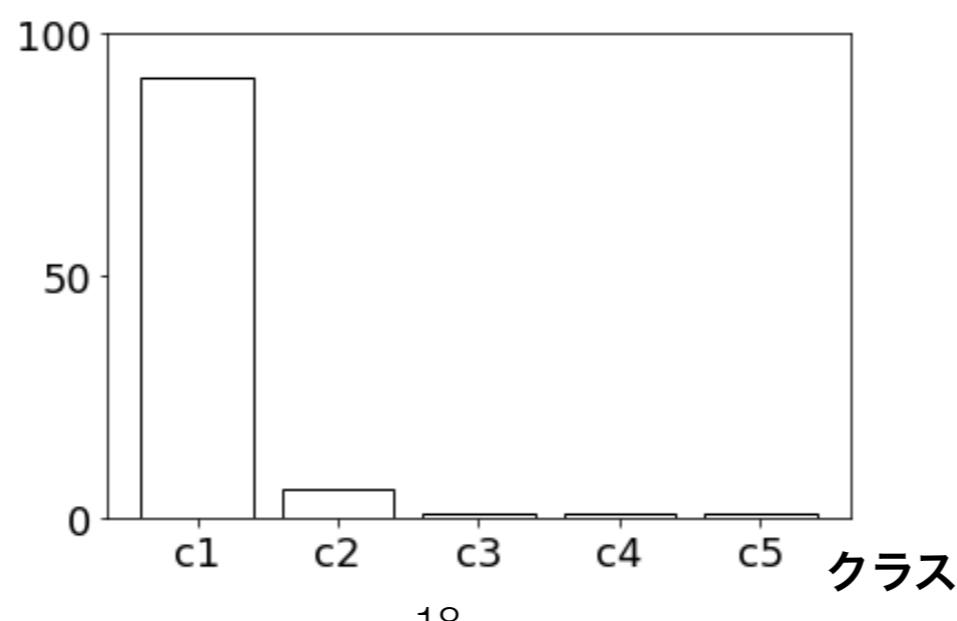
結果

blockデータセット (B.A.はバランスされた精度)

*不要なインバランスを避けるため, 1 vs. 1で学習を行なった.

B.A.(%)	Majority	クラス毎識別率				
		c_1	c_2	c_3	c_4	c_5
通常	78	0.98	0.86	0.63	0.85	0.61
提案	82	0.95	0.90	0.76	0.85	0.64

データ数の割合



ルールの重要度

blockデータセットを用いたc2 vs Restから得られたルールの重要度上位5つの表
*+1をマジョリティクラス, -1をマイノリティクラスとして学習している。

提案法				従来法			
重要度順位	重要度	回帰係数の符号	ルール	重要度	回帰係数の符号	ルール	
1	8.16	+	$x_6 \geq 2.9, x_1 < 4$	1.68	+	$x_1 \geq 3.5, x_7 < 37$	
2	7.54	-	$x_0 < 4.5, x_7 \geq 23$	0.56	+	$x_1 < 4.5, x_7 < 12.5$	
3	5.94	-	$x_0 < 3.5, x_7 \geq 34$	0.54	-	$x_1 < 3.5, x_7 \geq 1.32, x_1 \geq 1.5$	
4	3.96	+	$x_0 < 4.5, x_6 \geq 1.5$	0.49	+	$x_1 < 3.5, x_7 \geq 1.32, x_1 < 1.5$	
5	3.25	-	$x_0 < 3.5, x_6 < 34$	0.27	+	$x_1 \geq 3.5, x_7 \geq 29.5, x_1 < 41$	

*重要度は見やすさのため、100倍している。

ルールの重要度

blockデータセットを用いたc2 vs Restから得られたルールの重要度上位5つの表
*+1をマジョリティクラス, -1をマイノリティクラスとして学習している.

提案法				従来法			
重要度順位	重要度	回帰係数の符号	ルール	重要度	回帰係数の符号	ルール	
1	8.16	+	$x_6 \geq 2.9, x_1 < 4$	1.68	+	$x_1 \geq 3.5, x_7 < 37$	
2	7.54	-	<u>$x_0 < 4.5, x_7 \geq 23$</u>	0.56	+	$x_1 < 4.5, x_7 < 12.5$	
3	5.94	-	<u>$x_0 < 3.5, x_7 \geq 34$</u>	0.54	-	<u>$x_1 < 3.5, x_7 \geq 1.32, x_1 \geq 1.5$</u>	
4	3.96	+	$x_0 < 4.5, x_6 \geq 1.5$	0.49	+	$x_1 < 3.5, x_7 \geq 1.32, x_1 < 1.5$	
5	3.25	-	<u>$x_0 < 3.5, x_6 < 34$</u>	0.27	+	$x_1 \geq 3.5, x_7 \geq 29.5, x_1 < 41$	

*重要度は見やすさのため, 100倍している.

まとめ

ルールアンサンブル法に、バランスされた精度を最大化する工夫として分割のための損失関数と回帰係数の決定方法に改良を加えた。

実験の結果、

1. バランスされた精度とマイノリティクラスの識別率ともに改善された。
2. 重要なルールの比較を行った結果、従来とは異なるルールの検出とマイノリティクラスの識別が十分に加味されていることが確認できた。

分割基準第1項の計算(付録)

提案 $I(C; D) = H(D) - H(D|C)$ を利用

$$= H\left(\frac{N_L}{N}, \frac{N_R}{N}\right) - \left(\frac{N^+}{N} H\left(\frac{N_L^+}{N^+}, \frac{N_R^+}{N^+}\right) + \frac{N^-}{N} H\left(\frac{N_L^-}{N^-}, \frac{N_R^-}{N^-}\right)\right)$$

$P(+)$ = $\frac{1}{2}$ (事前確率を均等)にするために,

$\hat{N}^+ = \alpha^+ N^+ \left(\alpha^+ = \frac{N}{2N^+} \right)$ とする.

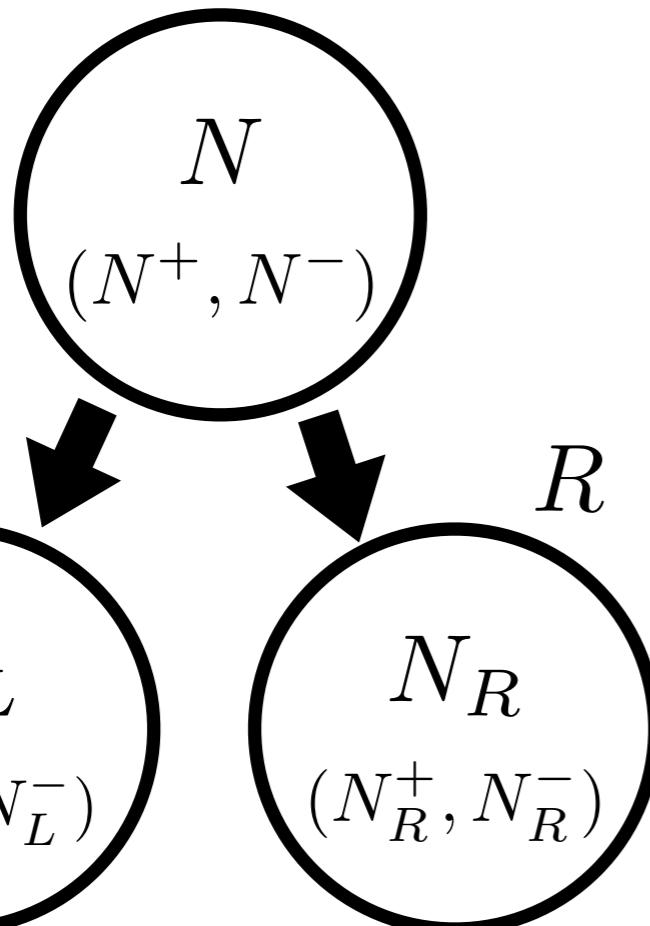
\hat{N}^- についても同様に計算して、これらを利用すると、

$$\hat{N}_L = \hat{N}_L^+ + \hat{N}_L^- = \alpha^+ N_L^+ + \alpha^- N_L^-$$

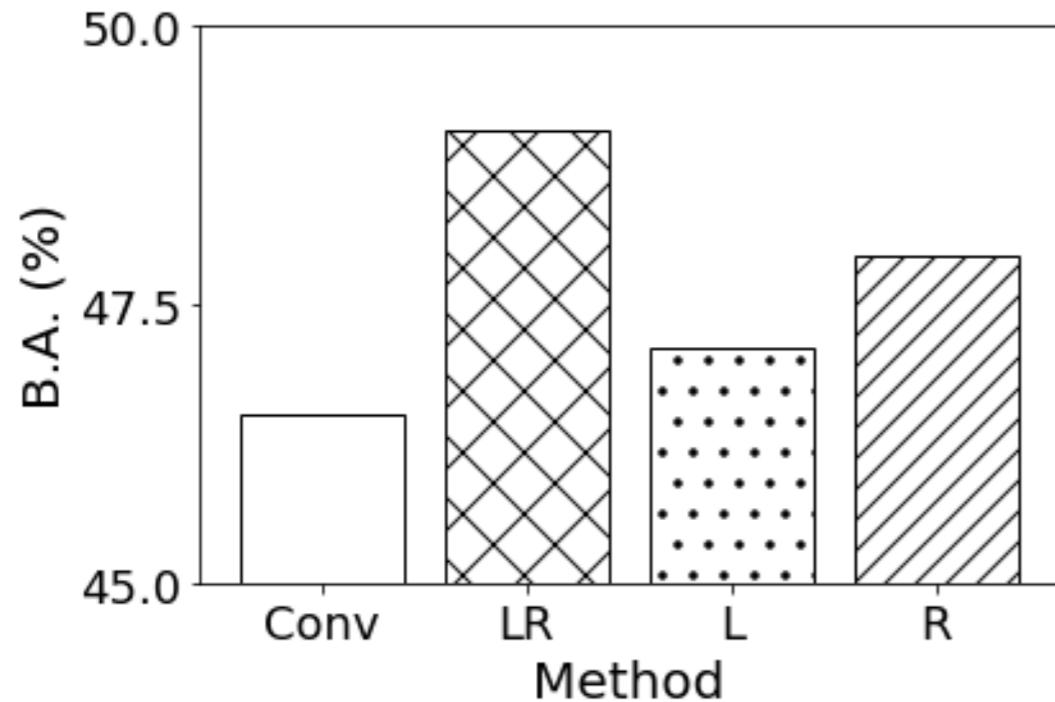
$$\hat{N}_R = \hat{N}_R^+ + \hat{N}_R^- = \alpha^+ N_R^+ + \alpha^- N_R^-$$

上の2式を第一項に適用すると、

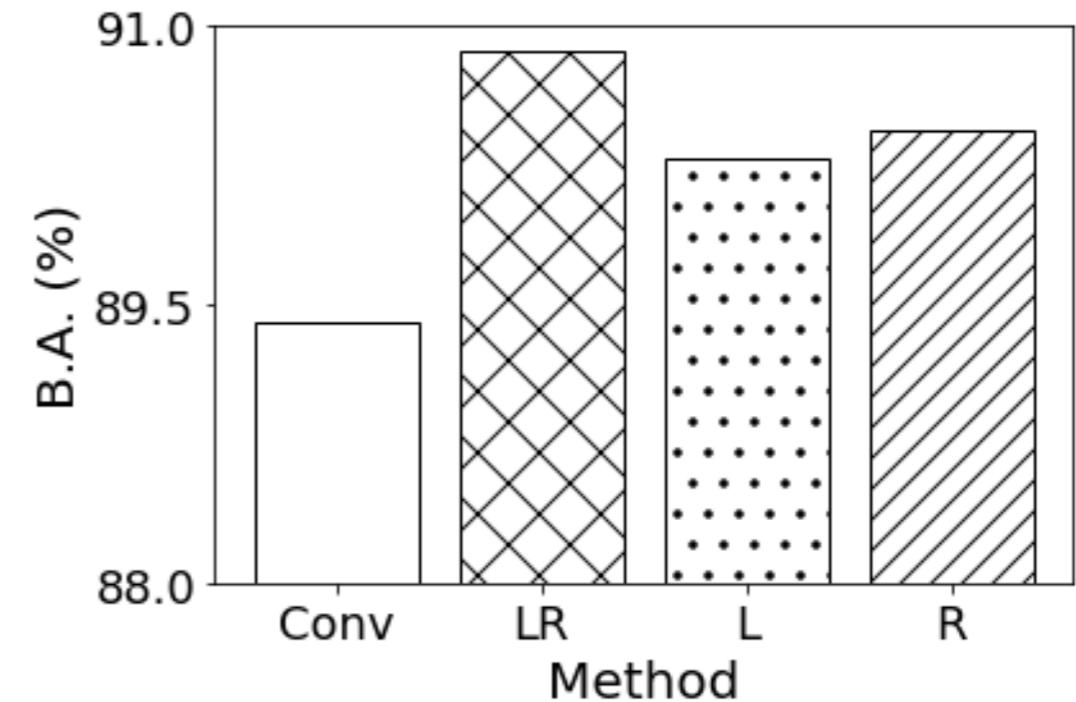
$$H\left(\frac{1}{2}\left(\frac{N_L^+}{N^+} + \frac{N_L^-}{N^-}\right), \frac{1}{2}\left(\frac{N_R^+}{N^+} + \frac{N_R^-}{N^-}\right)\right) - \left(\frac{1}{2} H\left(\frac{N_L^+}{N^+}, \frac{N_R^+}{N^+}\right) + \frac{1}{2} H\left(\frac{N_L^-}{N^-}, \frac{N_R^-}{N^-}\right)\right)$$



2つの改良の個別の効果(付録)



Balanceデータセット



Ionosphereデータセット

Conv: 通常のルールアンサンブル

LR: 損失関数と回帰係数の決定法に改良を加えたもの

L: 損失関数にのみ改良を加えたもの

R: 回帰係数の決定法にのみ改良を加えたもの

従来法との比較

glassデータセット

(B.A.はバランスされた精度)

B.A. (%)	クラス毎識別率					
	Majority			Minority		
	c_1	c_2	c_3	c_4	c_5	c_6
通常	66	0.79	0.58	0.21	0.62	0.93
提案	67	0.74	0.60	0.21	0.62	1.00
RUS	61	0.49	0.32	0.30	0.76	1.00
ROS	66	0.82	0.58	0.21	0.62	0.93

RUS: Random Under Sampling

ROS: Random Oversampling

データ数の割合

