

符号化ダイバージェンスによる 2つの集合の異なり具合の定量化

杉山 磨人^{†,‡}, 山本 章博[†]

[†] 京都大学情報学研究科

[‡] 日本学術振興会特別研究員 DC2

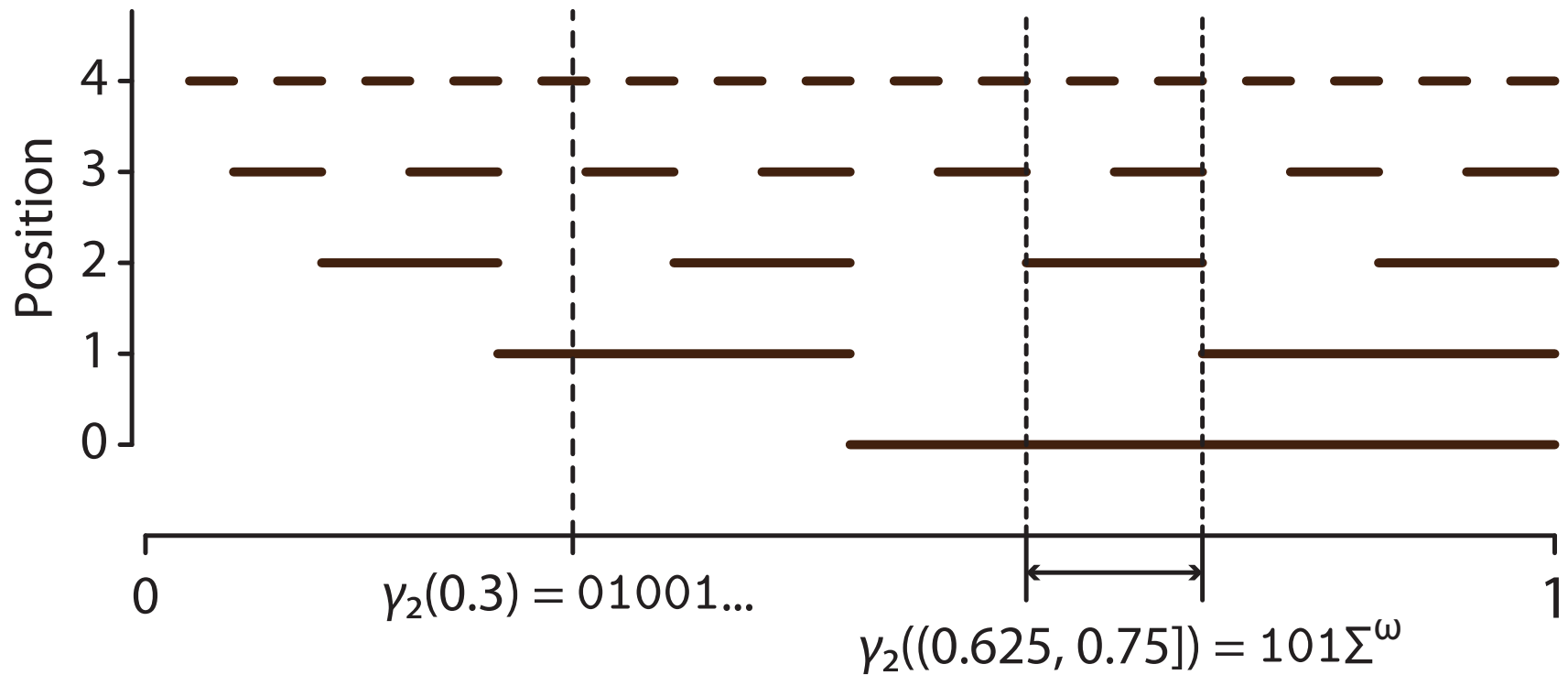
発表の概要

- 符号化ダイバージェンスという 2 つのデータ集合間の異なり具合を測る新規の尺度を提案する
 - ユークリッド空間 \mathbb{R}^d 上の実数値データをコントロール空間 Σ^ω へ位相的に埋め込む (離散化)
 - Σ^ω の中でデータに無矛盾なモデル (開集合) を学習する
 - 学習された開集合を表現するコードの長さで定量化
- 実験科学における仮説の検証に適用可能
 - コントロール実験で, データ集合間を比較する必要がある
 - 例: 新薬の効果と偽薬の効果の比較
 - 典型的な手法は統計的仮説検定 (t 検定など) だが, 現場で使いづらいことが多い [Johnson, 99]

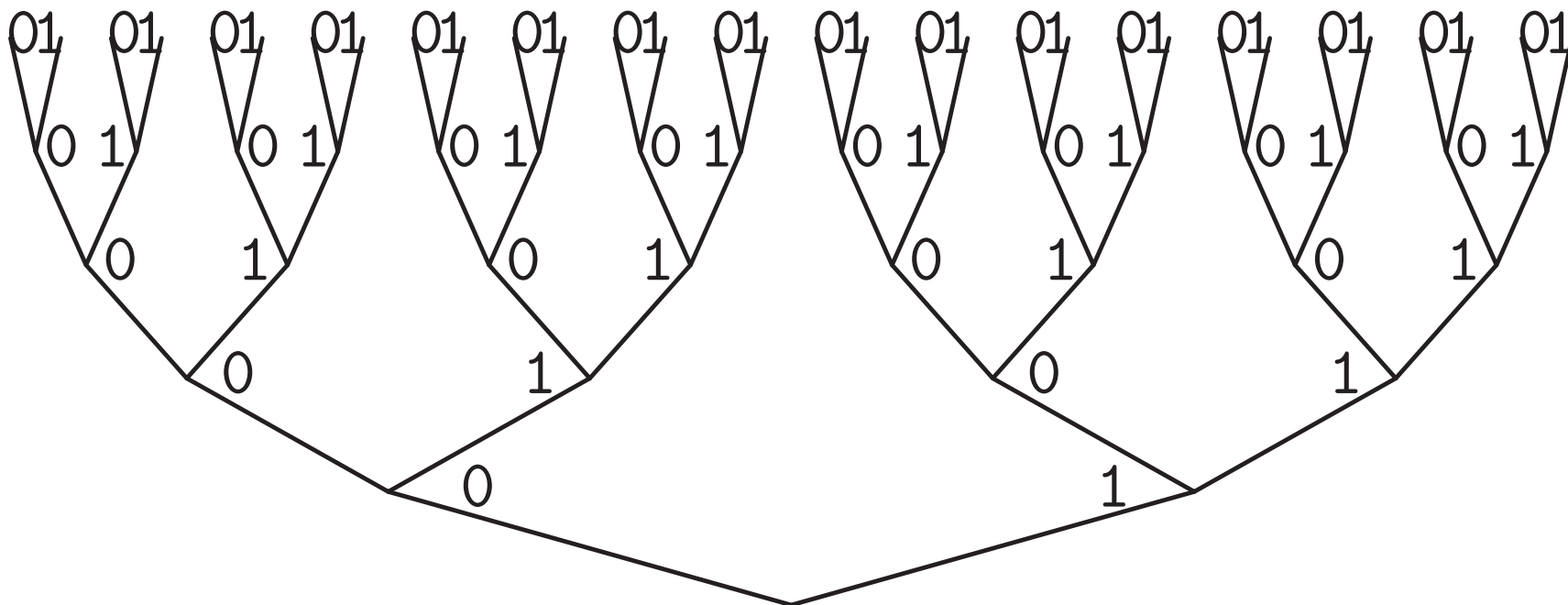
カントール空間への埋め込み

- **カントール位相** $\tau_{\Sigma^\omega} := \{W\Sigma^\omega \mid W \subseteq \Sigma^*\}$, 位相空間 $(\Sigma^\omega, \tau_{\Sigma^\omega})$ を**カントール空間**と呼ぶ
 - カントール空間とは, アルファベット Σ 上の無限列の集合 Σ^ω に導かれる標準的な**位相空間**
 - $w\Sigma^\omega = \{p \in \Sigma^\omega \mid w \sqsubseteq p\}$, $W\Sigma^\omega = \{p \in \Sigma^\omega \mid \exists w \in W. w \sqsubseteq p\}$
 - $\{w\Sigma^\omega \mid w \in \Sigma^*\}$ は基底となる
 - 集合 $P \subseteq \Sigma^\omega$ が**開集合**のとき, P は有限で観察可能
- d 次元ユークリッド空間 \mathbb{R}^d から**カントール空間**への**埋め込み** $\gamma: \subseteq \mathbb{R}^d \rightarrow \Sigma^\omega$ は, 実数値データの離散化に対応
 - 離散化されたデータは**開集合の基底**

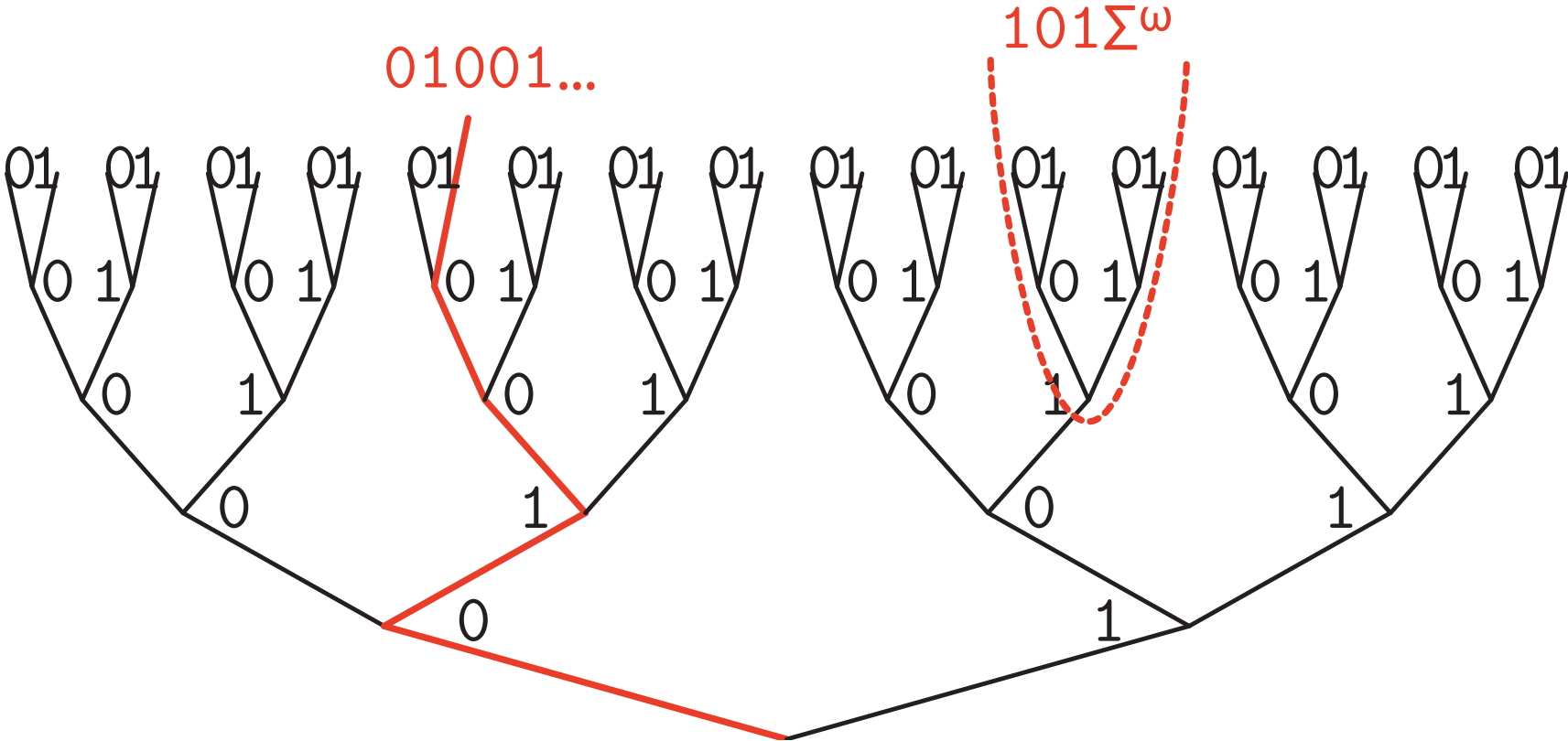
例：2進埋め込み γ_2



例：2進埋め込み γ_2



例：2進埋め込み γ_2



符号化ダイバージェンス

- 空でない有限集合 $X, Y \subset I$ (I は単位区間) に対して, γ に関する符号化ダイバージェンスを以下のように定義

$$C(X, Y) := \begin{cases} \infty & \text{if } X \cap Y \neq \emptyset, \\ D(X; Y) + D(Y; X) & \text{otherwise,} \end{cases}$$

- ここで, D は有向符号化ダイバージェンス

$$D(X; Y) := \frac{1}{\|X\|} \min \{ \sum_{w \in W} |w| \mid W \Sigma^\omega \text{ は } (\gamma(X), \gamma(Y)) \text{ に無矛盾} \}$$

- $\|X\|$ は X の要素数
- R が (P, Q) に無矛盾 $\iff R \supseteq R$ かつ $R \cap Q = \emptyset$
- 符号化ダイバージェンスはコントロール空間の位相的構造にのみ依存
 - 確率分布や統計的パラメータを全く使わない機械学習

学習アルゴリズム

function MAIN(X, Y, k_{\max})

$(H_1, H_2) \leftarrow$ LEARNING($X, Y, \emptyset, \emptyset, \mathbf{0}, k_{\max}$)

return $\frac{1}{\|X\|} \sum_{v \in H_1} |v| + \frac{1}{\|Y\|} \sum_{w \in H_2} |w|$

function LEARNING($X, Y, H_1, H_2, k, k_{\max}$)

$V \leftarrow$ OBSERVE(X, k), $W \leftarrow$ OBSERVE(Y, k)

$H_1 \leftarrow H_1 \cup \{v \in V \mid v \notin W\}$, $H_2 \leftarrow H_2 \cup \{w \in W \mid w \notin V\}$

$X \leftarrow \{x \in X \mid x \notin \rho(H_1 \Sigma^{\omega})\}$, $Y \leftarrow \{y \in Y \mid y \notin \rho(H_2 \Sigma^{\omega})\}$

if $X = \emptyset$ and $Y = \emptyset$ **then return** (H_1, H_2)

else if $k = k_{\max}$ **then return** $(H_1 \cup V, H_2 \cup W)$

else return LEARNING($X, Y, H_1, H_2, k + 1, k_{\max}$)

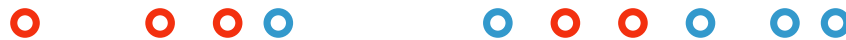
function OBSERVE(X, k)

return $\{\gamma(x)[n] \mid x \in X\}$ ($n = (k + 1)d - 1$)

符号化ダイバージェンスの学習

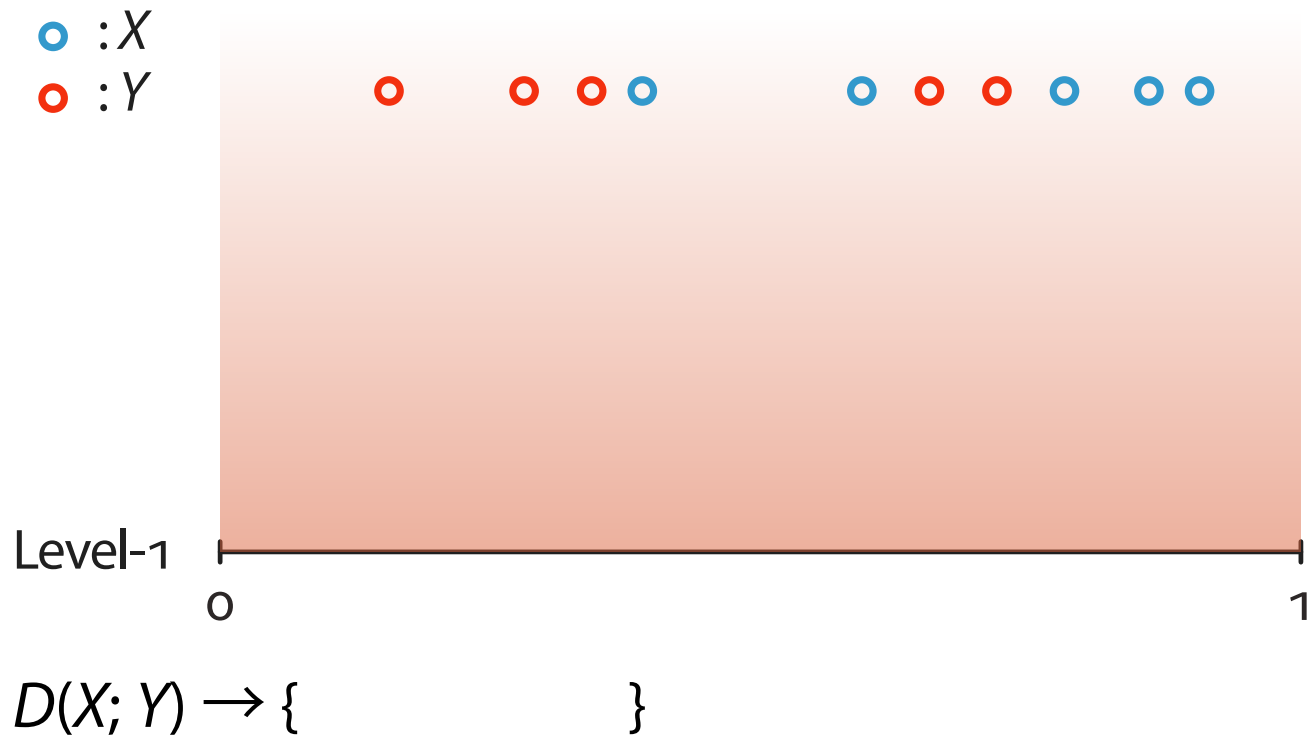
○ : X

○ : Y

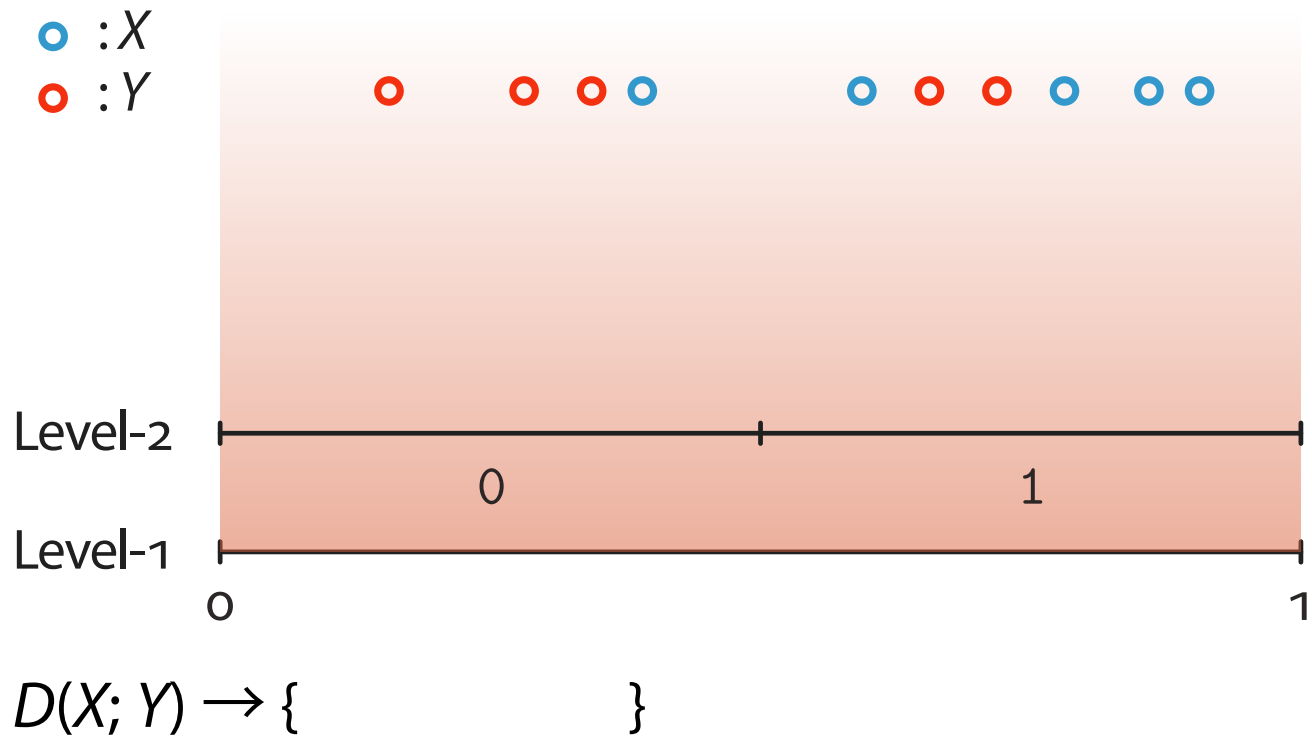


$D(X; Y) \rightarrow \{ \quad \}$

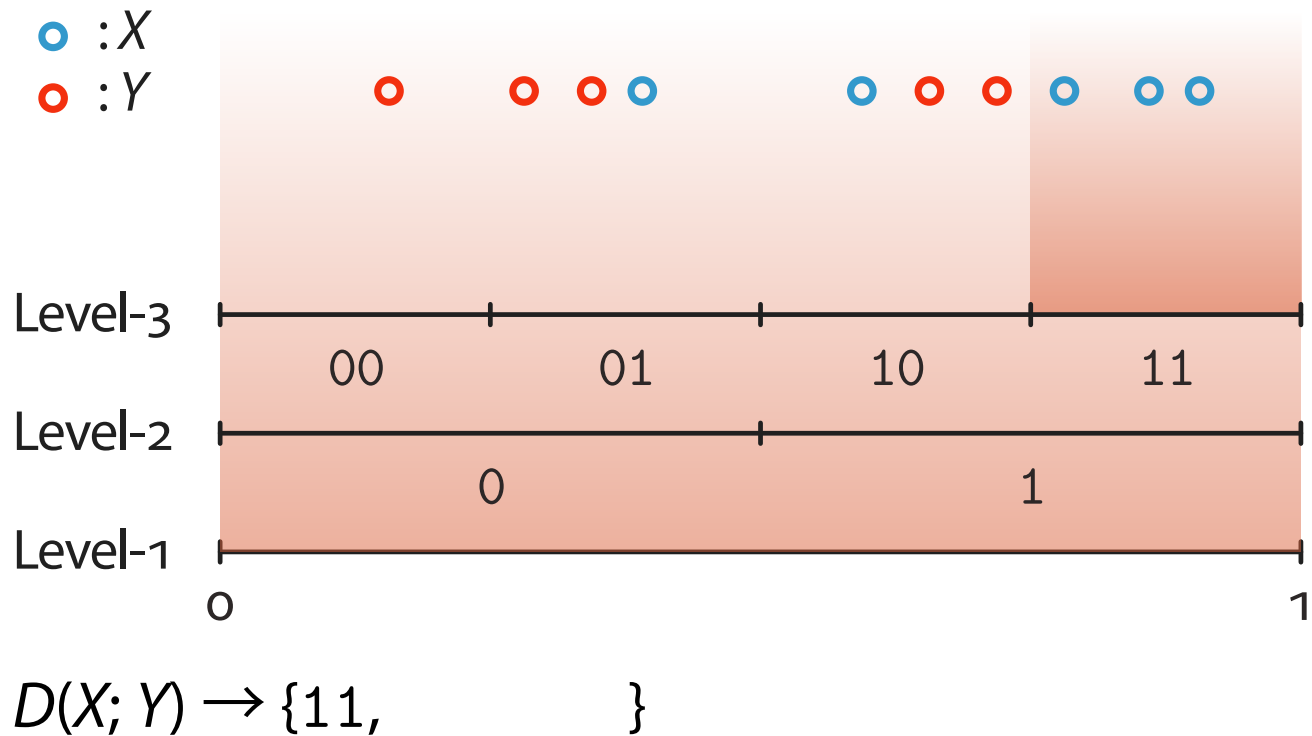
符号化ダイバージェンスの学習



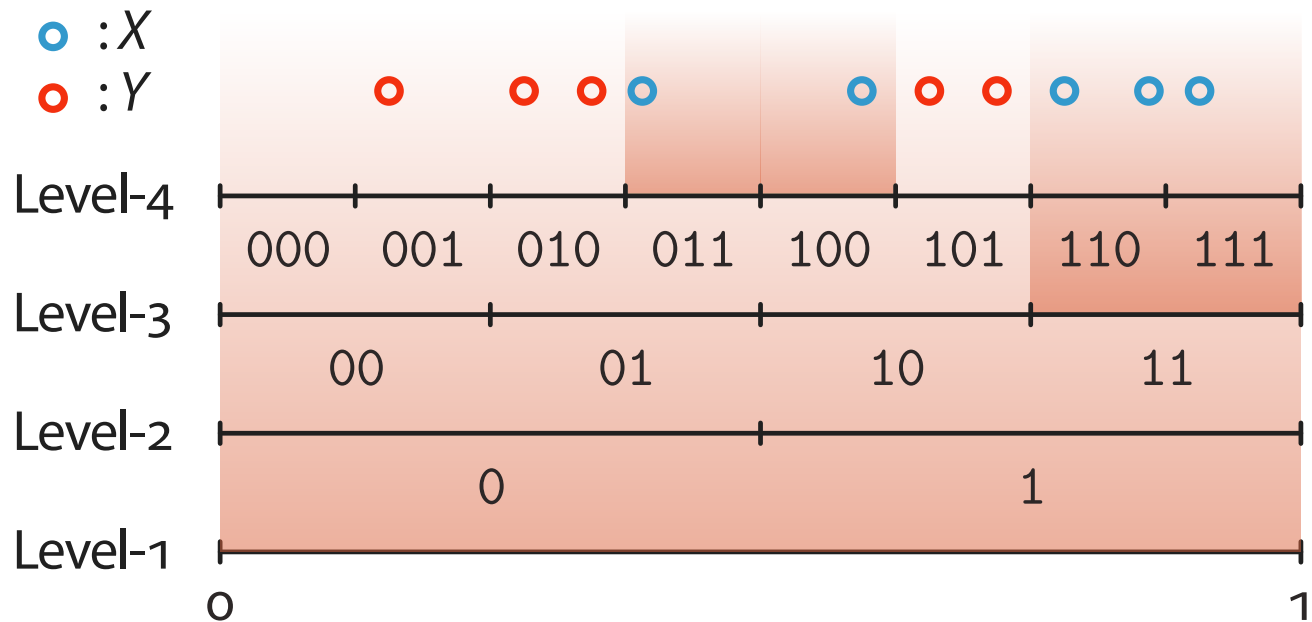
符号化ダイバージェンスの学習



符号化ダイバージェンスの学習

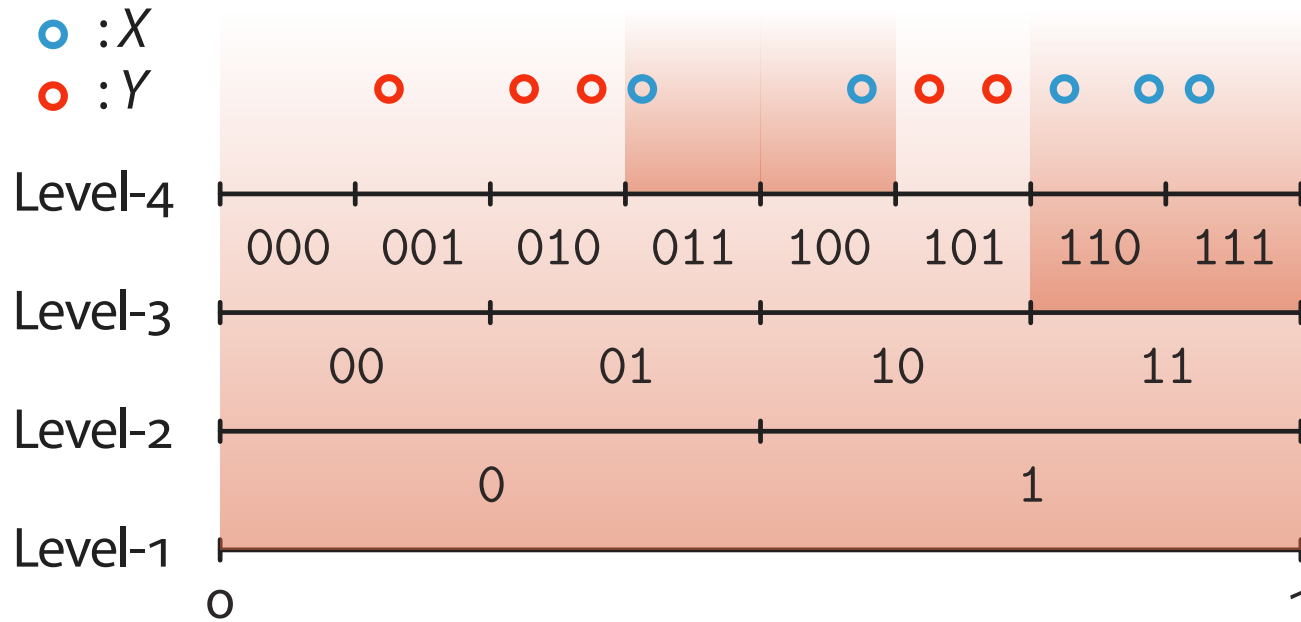


符号化ダイバージェンスの学習



$$D(X; Y) \rightarrow \{11, 011, 100\}$$

符号化ダイバージェンスの学習



$$D(X; Y) \rightarrow \{11, 011, 100\}$$

$$D(Y; X) \rightarrow \{00, 010, 101\}$$

$$C(X, Y) = 8/5 + 8/5 = 3.2$$

符号化ダイバージェンスを用いた分類

- 符号化ダイバージェンスを用いた怠惰学習によってクラス分類をおこなう
- この分類器は, 訓練データ X と Y (それぞれラベルは A と B とする) を受け取り, テストデータ Z を A か B へ分類する
 - 学習アルゴリズム M を用いる (k_{\max} は暗に与えられているとする)

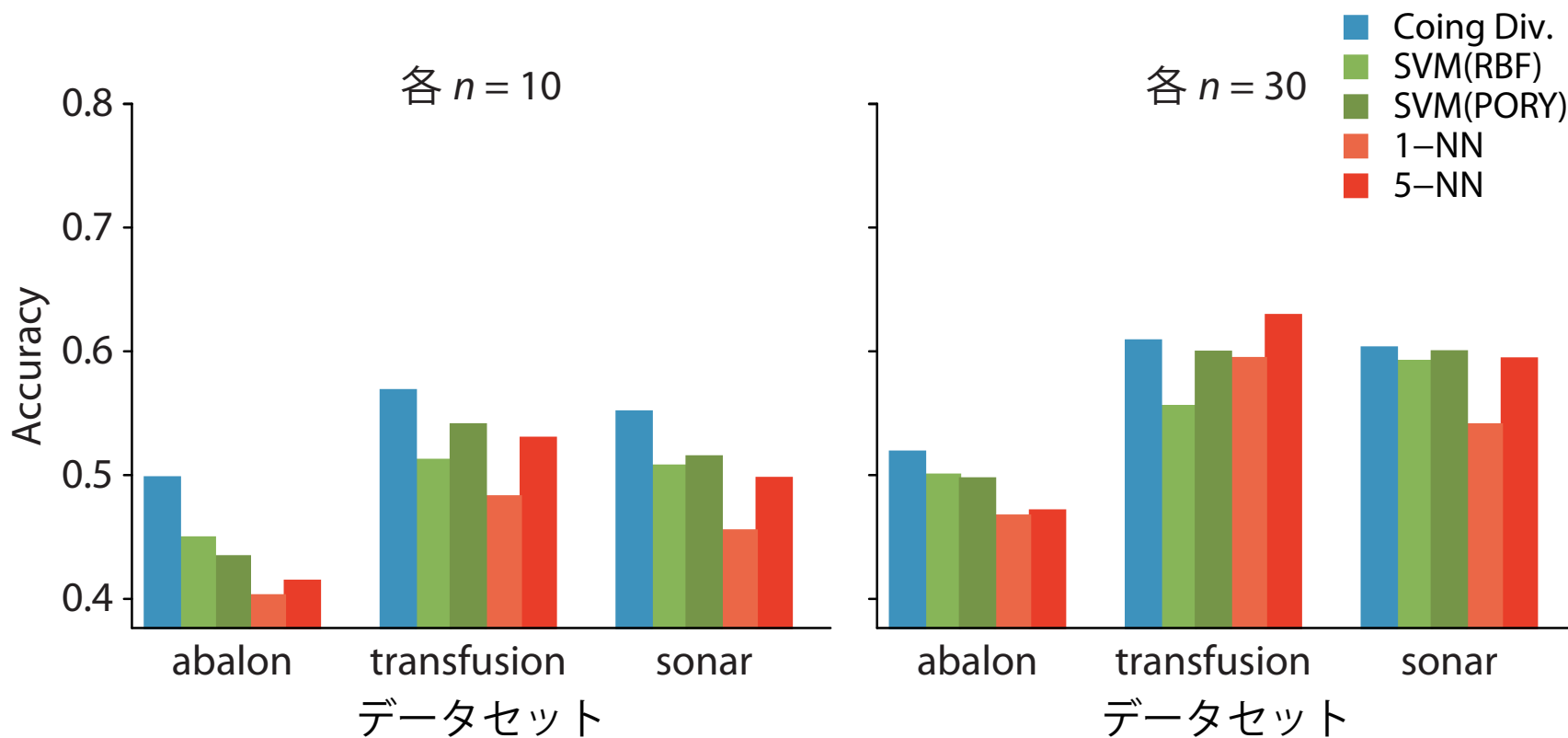
$$Z \text{ は } \begin{cases} A \text{ に属する} & \text{if } M(X, Z, k_{\max}) > M(Y, Z, k_{\max}), \\ B \text{ に属する} & \text{otherwise.} \end{cases}$$

- R 言語 (2.10.1) で実装した

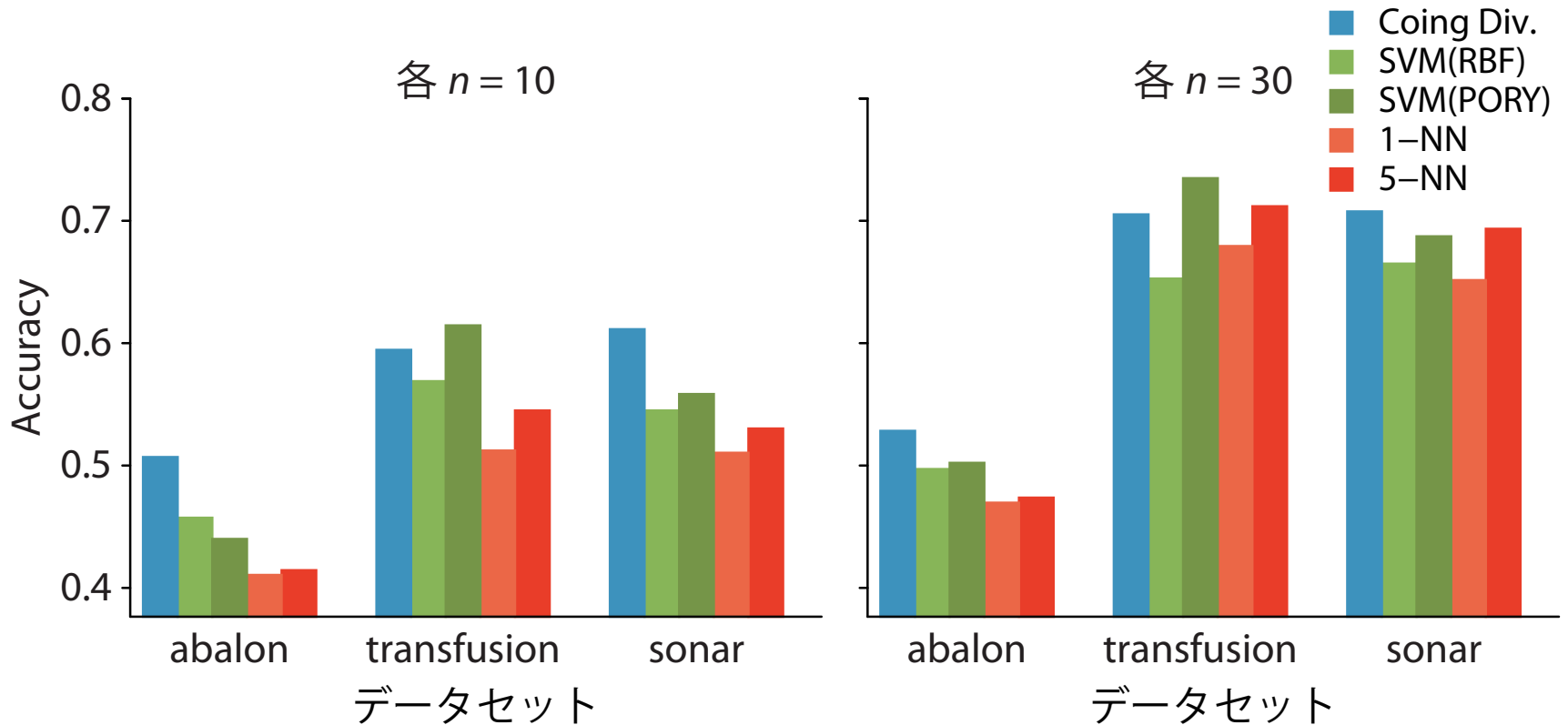
実験

- UCI のデータセット (abalone, sonar, ...) を用いる
- 以下を 10,000 回繰り返して, sensitivity と specificity から accuracy を求める
 - 識別に用いる属性をランダムに決める
 - 双方のラベルからそれぞれランダムに (非復元で) n 個サンプリングを 2 回 (X, T_+ と Y, T_-)
 - X, Y は訓練データ, T_+, T_- はテストデータ
 - データを正規化 (min-max normalization)
 - 符号化ダイバージェンス (と他の手法) を用いて T_+ が X と Y のどちらに近いかを判定して分類, T_- も同様に分類
- 得られた真陽性の数を t_{pos} , 真偽性の数を t_{neg} として, $(t_{\text{pos}} + t_{\text{neg}})/20000$ で accuracy を求める

実験結果 (属性数 1)



実験結果 (属性数 2)



まとめ

- 符号化ダイバージェンスという集合間の類似度を測る新規の尺度を提案
 - ユークリッド空間 \mathbb{R}^d 上の実数値データをコントロール空間 Σ^ω へ埋め込む (離散化)
 - 学習されるモデルはコントロール空間の開集合
 - 類似度はモデルを表現するコードの長さ
- 怠惰学習をおこなう分類器を構築
 - クラス分類の性能を実データを用いた実験で検証
 - SVM や k 近傍法に比べて遜色ない性能を持つことを確認