



HOKKAIDO UNIVERSITY

Title	機械学習への計算論的アプローチ
Author(s)	杉山, 磨人
Description	ERATO湊離散構造処理系プロジェクト : 2010年度初冬のワークショップ (ERATO合宿) . 2010年11月29日 (月) ~12月1日 (水) . 札幌北広島クラッセホテル.
Relation	2010年度科学技術振興機構ERATO湊離散構造処理系プロジェクト講究録. p.459-466.
Issue Date	2011-06
Doc URL	https://hdl.handle.net/2115/48336
Type	conference presentation
File Information	21.sugiyama.pdf



ERATO 合宿 2010
December 1, 2010

機械学習への計算論的アプローチ

杉山 磨人

京都大学情報学研究科
知能情報基礎論分野 (山本章博研究室)
日本学術振興会特別研究員 DC2

1/17

自己紹介

- ・ 杉山磨人 (SUGIYAMA Mahito)
- ・ 京都大学 情報学研究科 知能情報学専攻
知能情報基礎論分野 (山本章博研究室)
- ・ 博士後期課程 3 年

2/17

自己紹介

- ・ 杉山磨人 (SUGIYAMA Mahito)
- ・ 京都大学 情報学研究科 知能情報学専攻
知能情報基礎論分野 (山本章博研究室)
- ・ 博士後期課程 3 年
- ・ 学部では山本研究室
- ・ 修士の 2 年間は、同専攻の生体情報処理分野で、生命科学 (生理学・電気生理学・分子生物学など) に従事
 - 書籍「生命科学研究に成功するための統計法ノート」(小林茂夫, 杉山磨人, 講談社, 2009) を出版
- ・ 博士後期課程から、再び山本研究室

2/17

キーワード

- ・ 離散, 連続, 位相, 距離, 計算
- ・ Gold 型学習 (極限同定), フラクタル, コンパクト集合, ハウスドルフ距離, ハウスドルフ次元, VC 次元
- ・ 2 進表現, グレイコード, 離散化, カントール空間, タイプ 2 マシン, 計算可能性解析, クラス分類
- ・ 形式概念分析 (FCA), 離散・連続混在データ, 束構造, ガロア対応, 閉集合

3/17

興味のあること

- ・ 機械学習や知識発見, データマイニングへの計算論的・離散的なアプローチ
 - 離散と連続を行ったり来たりしながら, 連続的な対象 (実数値データなど) のための計算可能な手法を提案したい
 - 特に, 位相的・代数的アプローチに興味がある

4/17

興味のあること

- ・ 機械学習や知識発見, データマイニングへの計算論的・離散的なアプローチ
 - 離散と連続を行ったり来たりしながら, 連続的な対象 (実数値データなど) のための計算可能な手法を提案したい
 - 特に, 位相的・代数的アプローチに興味がある
- ・ 連続的な対象が計算可能とは?

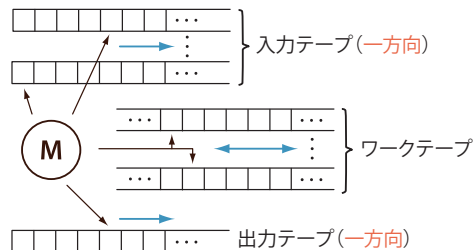
4/17

計算のモデル

- 離散の世界での計算モデル：
 - チューリングマシン，帰納関数，ラムダ計算
 - 有限時間で入力を読み込み，有限時間で出力して停止
- 連続の世界での計算モデル：
 - タイプ 2 マシン，実効的関数，...
 - 無限に入力を読み込み続け，無限に出力し続ける

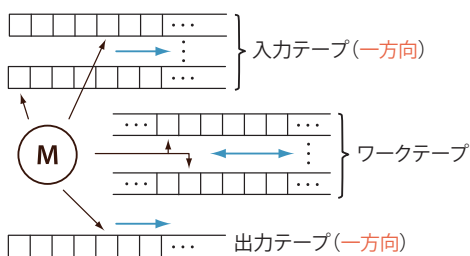
5/17

タイプ 2 マシン



6/17

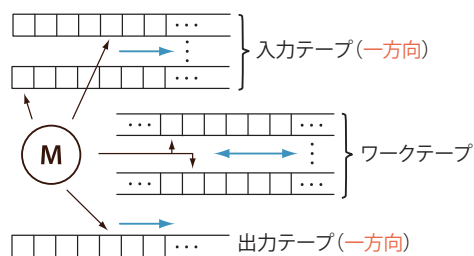
タイプ 2 マシン



- 計算可能性はコーディングの方法に依存する
 - 例：10 進表現や 2 進表現では $y = 3x$ が計算できない

6/17

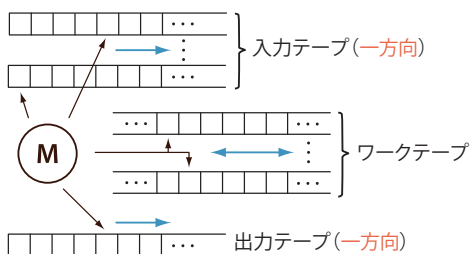
タイプ 2 マシン



- 0.33333333...

6/17

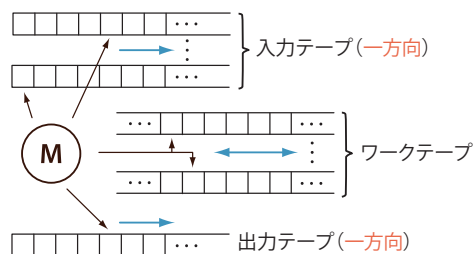
タイプ 2 マシン



- 0.33333333... → 0.99999999...

6/17

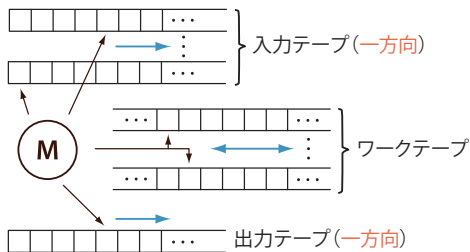
タイプ 2 マシン



- 0.333333334... → 0.99999999...

6/17

タイプ 2 マシン



• $0.33333334\dots \rightarrow 1.00000000\dots$

6/17

これまで（これから）の研究

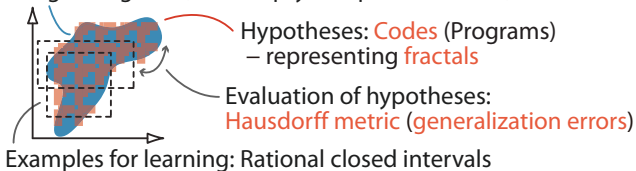
1. フラクタルによる図形（ユークリッド空間のコンパクト集合）の計算論的学習 [LLLL2009, ALT2010]
 - 廣渡栄寿先生（北九州大），立木秀樹先生（京都大）との共同研究
2. (グレイ) 符号化ダイバージェンスによる 2 つの集合の異なり具合の定量化，それを用いた機械学習 [ACML2010]
3. 離散量と連続量が混在するデータに対する形式概念分析を用いた半教師あり学習 [8othFPAI]

7/17

図形の学習の概要

- Constructing a computational learning model for analog data with discretization
- 1. Gold-style learning model as a base model
- 2. Computable Analysis to give theoretical support for discretizing process of analog data
- 3. Fractals to represent (and compute) continuous objects

Targets: Figures (non-empty compact sets in $\mathbb{R}^n$)

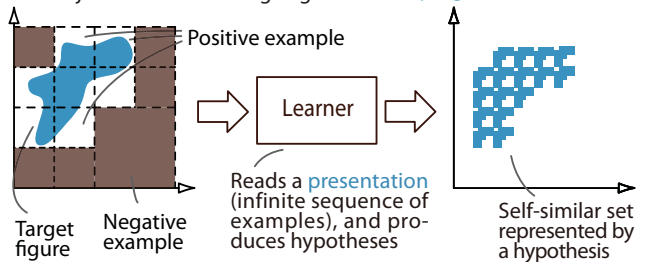


8/17

図形の学習の概要

Positive examples: Rational closed intervals intersecting the learning target
Negative examples: Rational closed intervals disjoint with the learning target

Hypotheses: Codes that represent self-similar sets (self-similar programs)



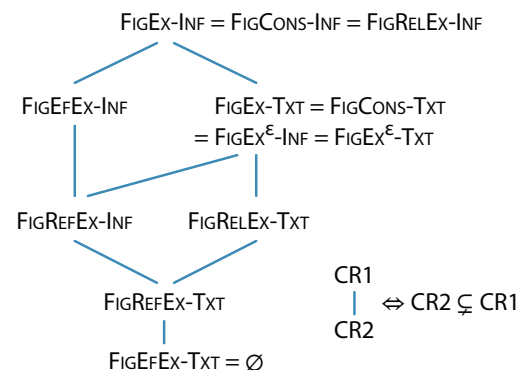
9/17

主な結果

1. We formulated learning of figures with self-similar sets (fractals) using the Gold-style learning model
 - Collage Theorem gives justification for self-similar sets
2. We analyzed the hierarchy of learnabilities (next slide)
3. We revealed the mathematical connection between Fractal Geometry and Computational Learning
 - The complexity of learning (sample size) is measured by using the Hausdorff dimension and the VC dimension
 - The Hausdorff dimension and the VC dimension are key concepts of Fractal Geometry and the Valiant-style learning model, respectively

10/17

学習基準間の階層



11/17

符号化ダイバージェンスの概要

• Main results:

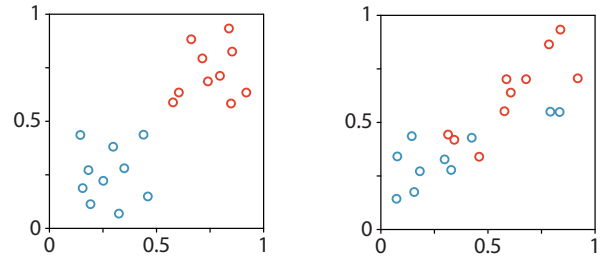
1. We propose the **coding divergence**, a novel measure of the similarity between two sets of continuous data
 - Measure the **complexity of separating** the two sets
2. We constructed the lazy learner, and showed the competitive performance in classification by experiments

• Key processes:

1. Embed continuous data in the **Euclidean space** $\mathbb{R}^d$ into the **Cantor space** Σ^ω topologically (**discretization**)
2. **Learn** the simplest model (**open set**) in Σ^ω
3. Count the **length of the code** encoding the model

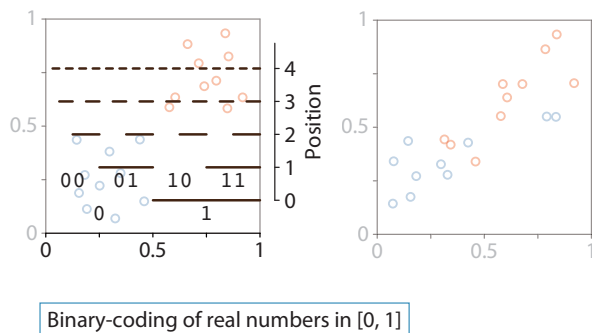
12/17

符号化ダイバージェンスの例



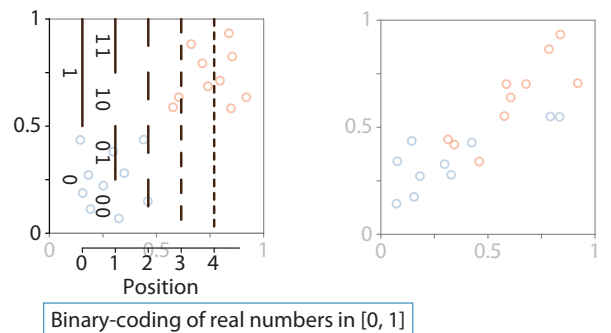
13/17

符号化ダイバージェンスの例



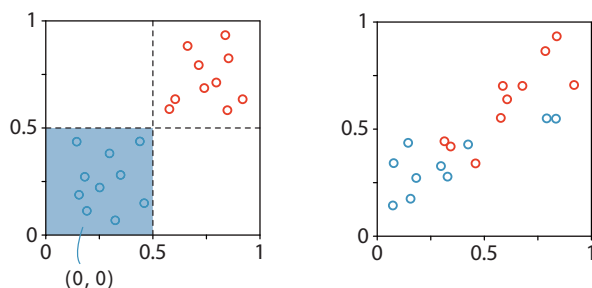
13/17

符号化ダイバージェンスの例



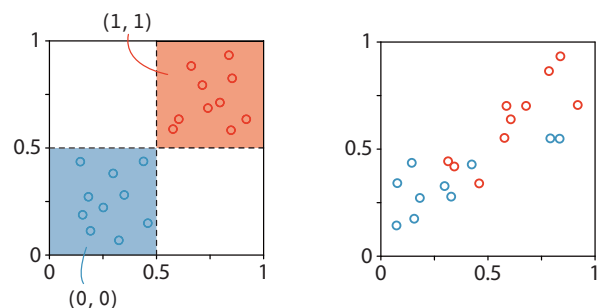
13/17

符号化ダイバージェンスの例



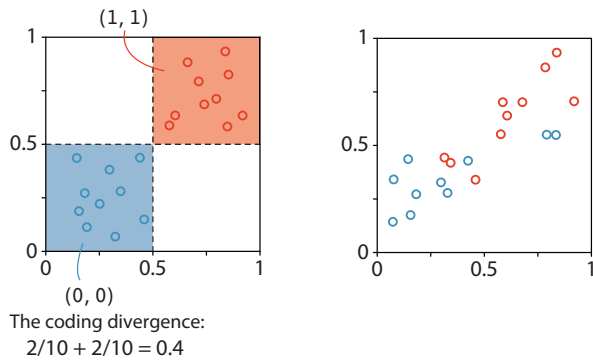
13/17

符号化ダイバージェンスの例



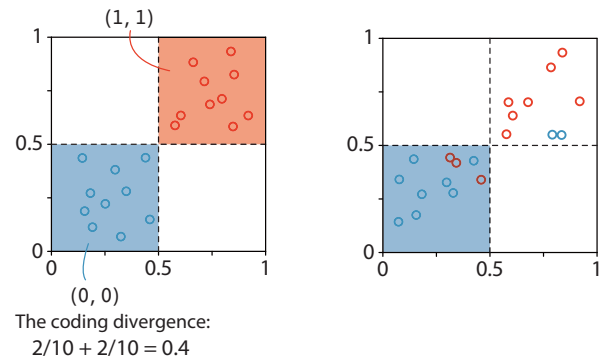
13/17

符号化ダイバージェンスの例



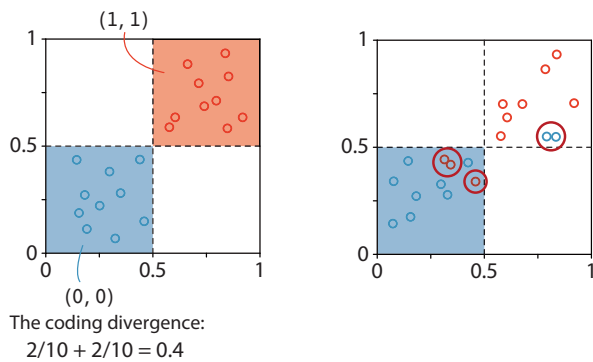
13/17

符号化ダイバージェンスの例



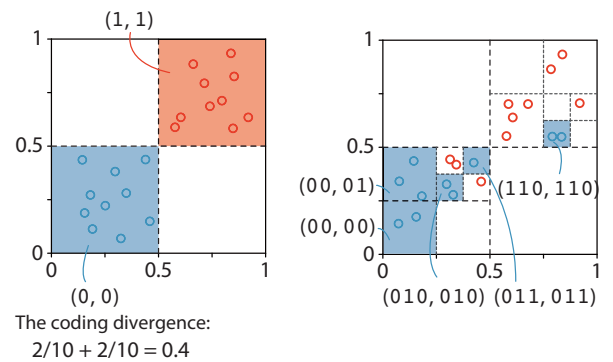
13/17

符号化ダイバージェンスの例



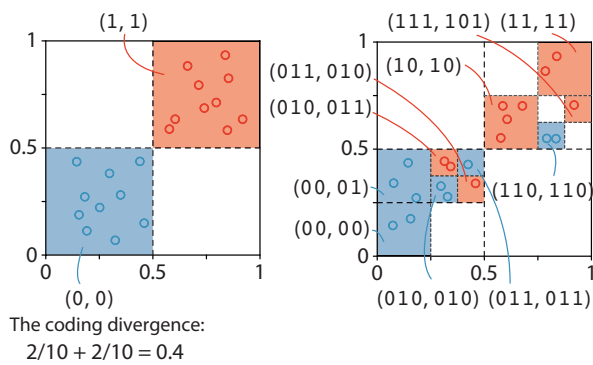
13/17

符号化ダイバージェンスの例



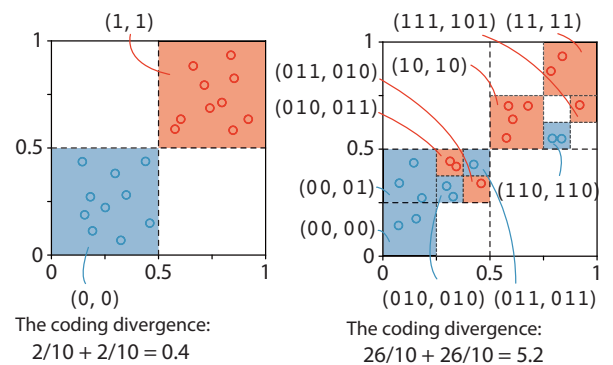
13/17

符号化ダイバージェンスの例



13/17

符号化ダイバージェンスの例



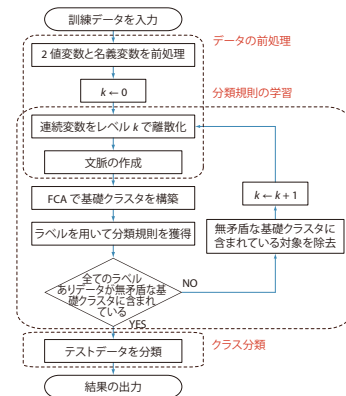
13/17

概要

- 離散量と連続量をともに含む**混在データ**に対する新規の**半教師あり学習手法 SELF** (SEmi-supervised Learning via FCA) を提案する
 - 形式概念分析 (FCA) を用いて、概念 (基礎クラスタ) からなる空間を構築
 - ラベル情報を用いて分類規則を学習する
- 主な成果
 - 混在データに対して直接適用可能な、初めての半教師あり学習手法
 - 混在データを扱える機械学習手法はあまりない
 - 不完全なデータセットをそのまま扱うことができる
 - 欠損値・ラベルの欠損・連続量が持つ誤差

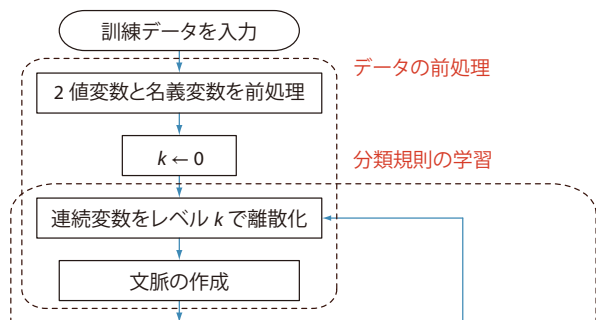
14/17

SELF の概要



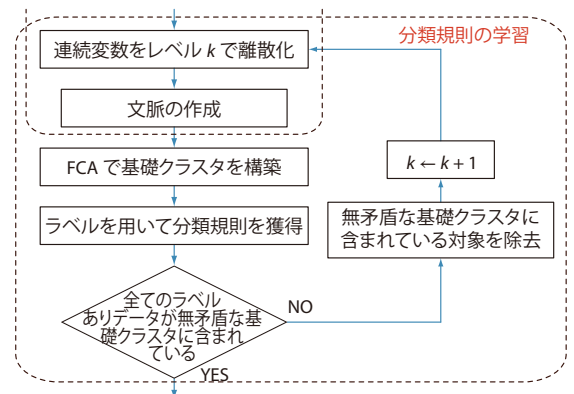
15/17

SELF の概要



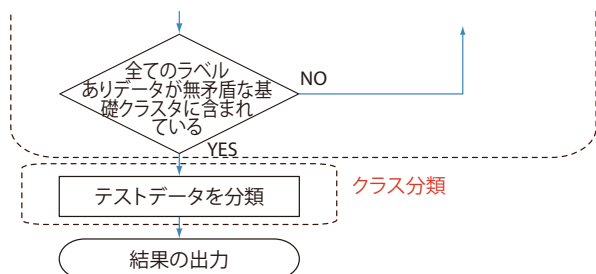
15/17

SELF の概要



15/17

SELF の概要



15/17

SELF による学習の例

- 以下のデータセットからの学習を考える (⊥ は欠損値)

	属性 1	属性 2	属性 3	ラベル
1	T	C	0.28	1
2	F	A	0.54	1
3	T	B	⊥	⊥
4	F	A	0.79	2
5	T	C	0.81	⊥

16/17

SELF による学習の例

- 以下のデータセットからの学習を考える (⊥ は欠損値)

	属性 1	属性 2	属性 3	ラベル
1	T	C	0.28	1
2	F	A	0.54	1
3	T	B	⊥	⊥
4	F	A	0.79	2
5	T	C	0.81	⊥

- まずデータ前処理で文脈を作成する

	1.T	2.A	2.B	2.C
1	×			×
2		×		
3	×		×	
4		×		
5	×			×

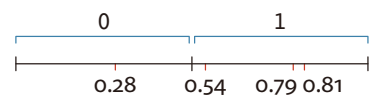
16/17

SELF による学習の例

- 以下のデータセットからの学習を考える (⊥ は欠損値)

	属性 1	属性 2	属性 3	ラベル
1	T	C	0.28	1
2	F	A	0.54	1
3	T	B	⊥	⊥
4	F	A	0.79	2
5	T	C	0.81	⊥

- 連続変数 (属性 3) は実数の 2 進表現に基づき離散化する



16/17

SELF による学習の例

- 以下のデータセットからの学習を考える (⊥ は欠損値)

	属性 1	属性 2	属性 3	ラベル
1	T	C	0.28	1
2	F	A	0.54	1
3	T	B	⊥	⊥
4	F	A	0.79	2
5	T	C	0.81	⊥

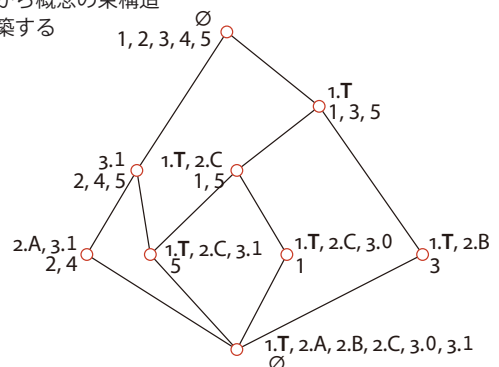
- まずデータ前処理で文脈を作成する

	1.T	2.A	2.B	2.C	3.0	3.1
1	×			×	×	
2		×				×
3	×		×			
4		×				×
5	×			×		×

16/17

SELF による学習の例

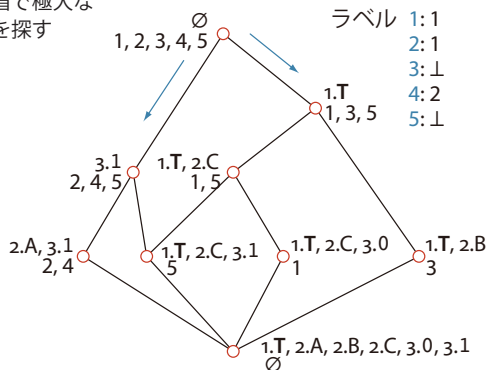
- 文脈から概念の束構造を構築する



16/17

SELF による学習の例

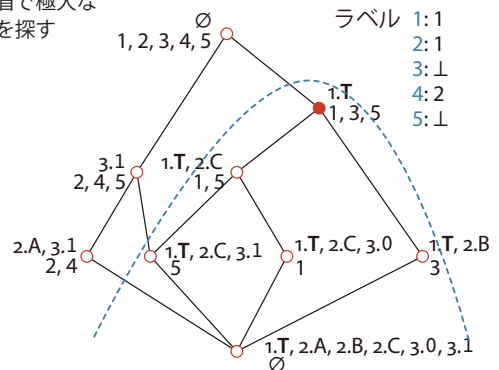
- 無矛盾で極大な概念を探す



16/17

SELF による学習の例

- 無矛盾で極大な概念を探す



16/17

SELF による学習の例

- 残ったデータから学習を進める

	属性 1	属性 2	属性 3	ラベル
2	F	A	0.54	1
4	F	A	0.79	2

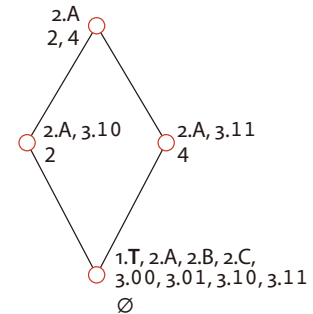
- 離散化を精密にすることで、以下の文脈を作成する

	1.T	2.A	2.B	2.C	3.00	3.01	3.10	3.11
2		×					×	
4		×						×

00	01	10	11
0.54 0.79			

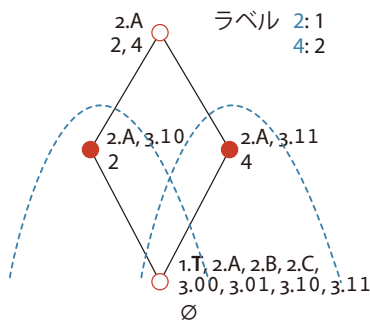
16/17

SELF による学習の例



16/17

SELF による学習の例



16/17

SELF による学習の例

- 獲得される分類規則: $\mathbb{R} = \{\mathbb{R}_1, \mathbb{R}_2\}$
 - $\mathbb{R}_1 = \{\{1.T\}, 1\}$
 - $\mathbb{R}_2 = \{\{2.A, 3.10\}, 1\}, \{2.A, 3.11\}, 2\}$
- 各 $\mathbb{R}_i$ は、組 (極大な概念の属性, 分類されるクラス) となる集合
- 未知データの分類の例
 - データ (T, B, 0.45) は、最初の変数が T なのでクラス 1
 - データ (F, A, 0.64) は、2 目目の変数が 1 で 3 目目が 0.5 ~ 0.75 に入っているのでクラス 2

16/17

今後の計画

- フラクタルを用いた図形の学習
 - 形式概念分析の枠組みで再構築
- 符号化ダイバージェンス
 - クラスタリング
 - オンラインでの異常値検出
 - ノイズデータの扱いや、VC 次元との理論的關係
- 混在データからの半教師あり学習
 - 理論的解析
 - 特徴選択との関連

17/17