

学習数理情報学研究室の紹介

東京大学 大学院情報理工学系研究科
数理情報学専攻
第6研究室

<http://www.ibis.t.u-tokyo.ac.jp/>

入試説明会 2021/5/22

学習数理情報学研究室

- 機械知能の本質を数理で捉える -



教授 山西健司

専門: 情報論的学習理論、データマイニング、異常検知
講義: 機械学習の数理(夏:学部)
数理情報学特別講義I データマイニングによる
異常検知(冬:大学院)



准教授 鈴木大慈

専門: 機械学習, 数理統計学, 確率的最適化
講義: 確率数理工学(夏:学部),



講師 久野遼平 (MIセンター本務、数理兼務)

専門: 経済学、ビッグデータ解析



特任助教 木脇太一

専門: ニューラルネット講義:
データマニング実践演習



助教 池祐一

専門: 位相的データ
解析(TDA)

研究室の特徴

- しっかりとした基礎理論に立脚しつつ、現場に生きる学習数理情報学をめざします
- 幅広い産業界との連携のもとで、研究を進めます
以下の企業とはすでに協力を開始しています。

Microsoft



hd 東大病院
The University of Tokyo Hospital

三菱商事

NTT

HAKUHODO

LAC
Little eArth Corporation

sas | THE POWER TO KNOW.

YAHOO!
JAPAN

ネットリサーチなら、マクロミル
macromill

VALUES
Consulting & Creation Group

TORAY
Innovation by Chemistry

MIZUHO

みずほ銀行

HONDA
The Power of Dreams

MODEC

SOMPO 損保ジャパン

NIPPON STEEL

山西研究室の概要

コア技術: **情報論的学習理論**

目標: **ディープナレッジの発見**

基礎理論

潜在的ダイナミクス

潜在構造最適化

予兆情報学

Sign Informatics

主要応用

医学応用

経済応用

マーケット
解析

量子
機械学習

潜在的ダイナミクス1

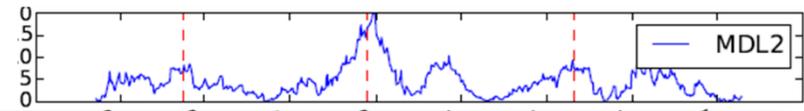
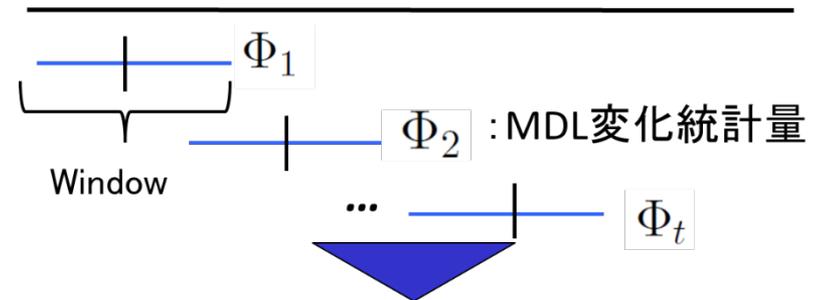
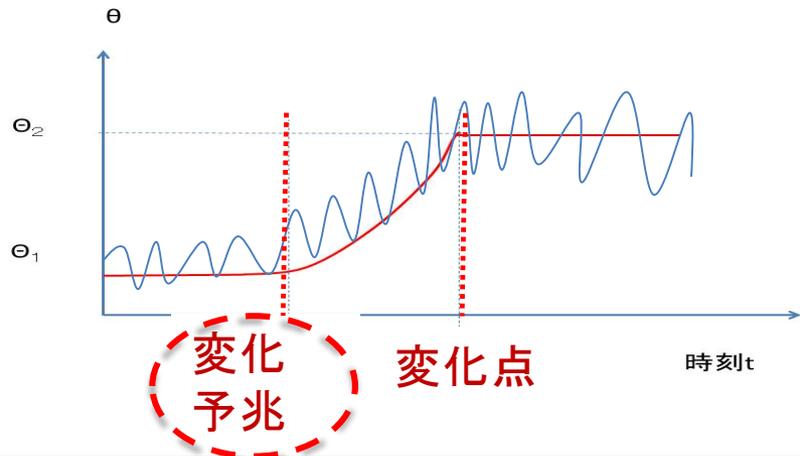
MDL変化統計量による変化予兆検知

データ圧縮度合いで漸進的変化の開始点(予兆)を検知

[Yamanishi, Miyaguchi BigData2016] [Kaneko, Miyaguchi, Yamanishi BigData2017]

[Yamanishi, Fukushima IEEE Trans Inform Theory 2018]

$$x^n = x_1 \dots x_n,$$



スコア
曲線

技術コア: MDL変化統計量によるオンライン変化検知を提案

MDL: Minimum Description Length

$$\Phi_t = \min_M L_{\text{NML}}(x_1^n; M) - \min_{M_1, M_2} \{L_{\text{NML}}(x_1^t; M_1) + L_{\text{NML}}(x_{t+1}^n; M_2)\}$$

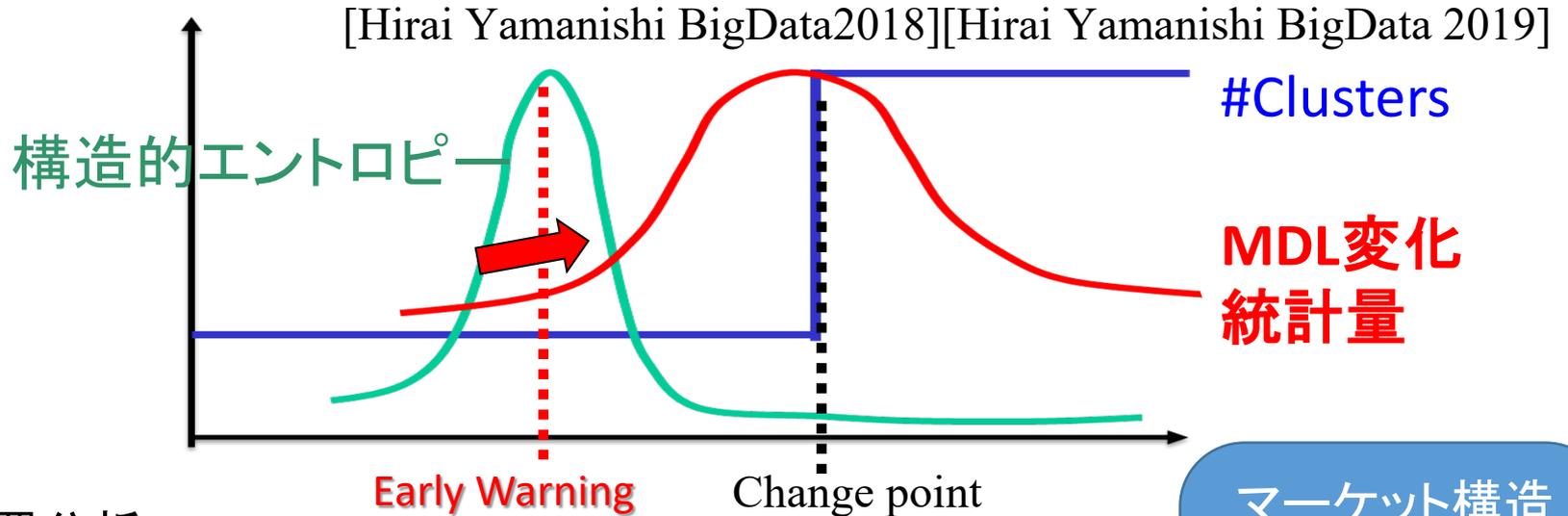
変化した場合と
しない場合の
記述長の差

- ⇒
- ・固定窓で移動計算、漸進的変化検知 [BigData 2016]
 - ・可変窓で移動計算、検知精度を大幅改善 [BigData 2017]
 - ・工場のボイラー事故の予兆を検知(東レとの実証実験)

潜在的ダイナミクス2

潜在構造変化予兆検知

連続量(構造エントロピー、MDL変化統計量)で変化過渡期を検知

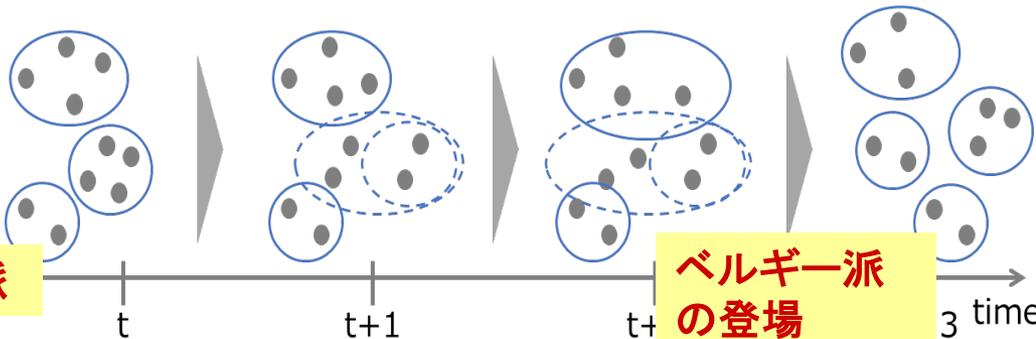


ビール購買分析

発泡酒派

地ビール派

プレミアム派



マーケット構造
変化予兆検知
電力消費パタン
変化予兆検知
で有効性検証

$$H_t = - \sum_k p(k|\mathbf{x}) \log p(k|\mathbf{x}_t) \quad p(k|\mathbf{x}_t) = \frac{\exp(-\beta L_{\text{NML}}(\mathbf{x}_t; k))}{\sum_{k'} \exp(-\beta L_{\text{NML}}(\mathbf{x}_t; k'))}$$

潜在的ダイナミクス3

階層的変化検知

変化はどこから来るのか？ 要因を分析し、変化の感度分析を実現

潜在変数
モデル

$$(x_t, z_t) \sim f(X, Z; \theta, M),$$

[Fukushima Yamanishi ICDM2020]

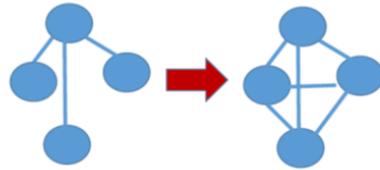
X : observed variable

Z : latent variable

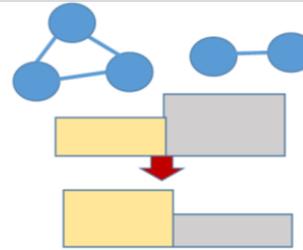
$$p_j(X, Z; \theta, M) = f(X|Z; \theta_1, M) f(Z; \theta_2, M) \quad \theta = (\theta_1, \theta_2)$$

Level1
パラメータ
の変化

θ_1



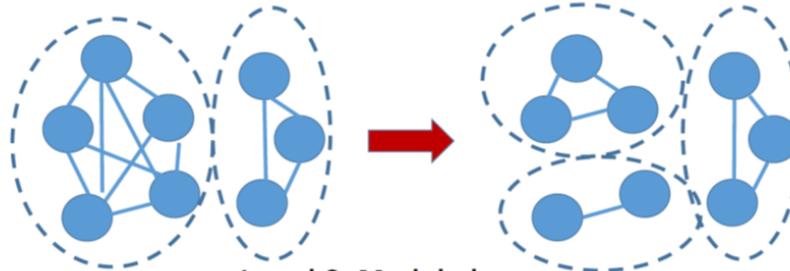
Level 1. Data distribution change



Level 2. Latent variable change

Level2
潜在変数
の変化

θ_2



Level 3. Model change

Level3
モデルの変化

M

人口移動
データ分析、
センサー
データ解析で
有効性実証

分解型
変化スコア
(MDL変化統計量
+ DNML基準)

$$\Phi_t = \Phi_t^{X|Z} + \Phi_t^Z + \Delta L_t(\hat{M}, \hat{M}_1, \hat{M}_2)$$

↑
パラメータのMDL
変化統計量

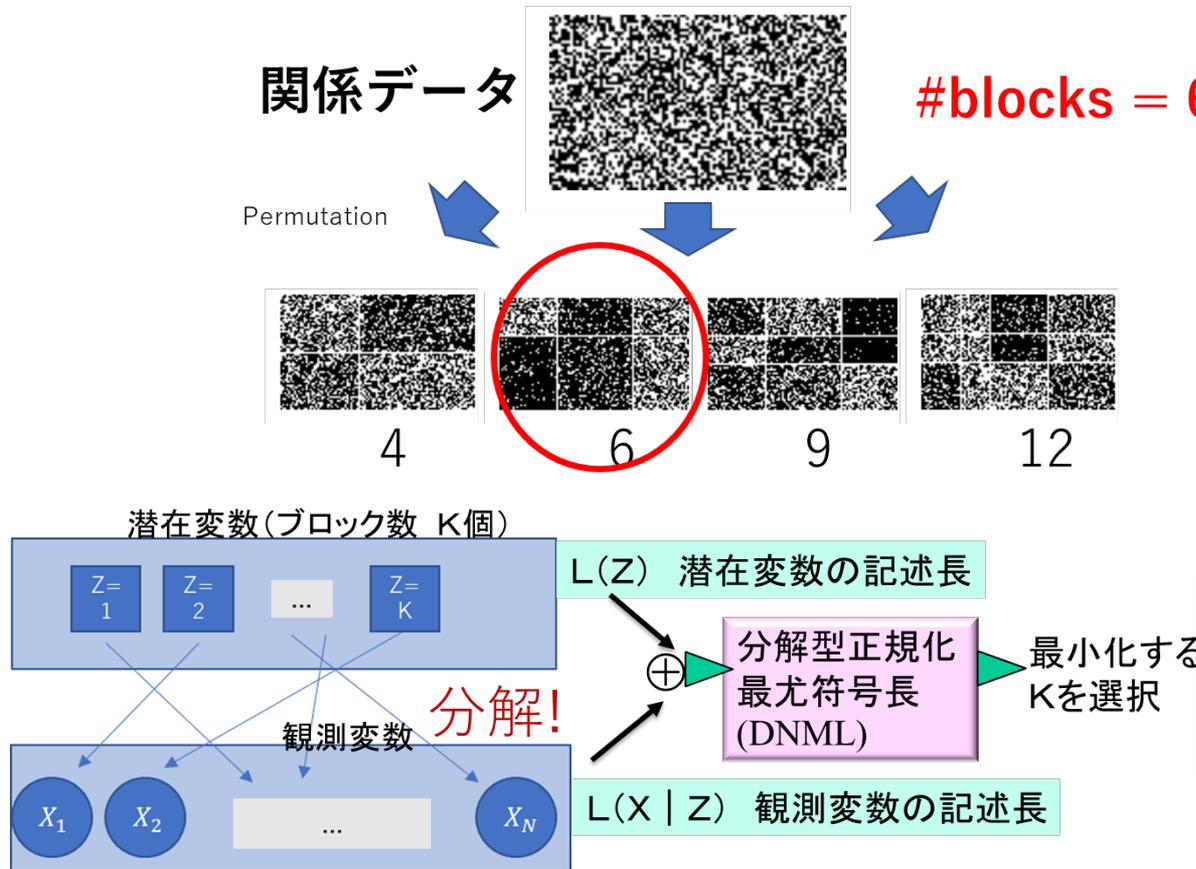
↑
潜在変数のMDL
変化統計量

モデルのMDL変
化統計量(全体)

潜在構造最適化 1

DNMLによる潜在変数モデル選択

分解型正規化最尤符号長規準 (DNML) で潜在変数の数を最適化



[Yamanishi, Wu, Sugawara, Okada DAMI 2019]

[Wu, Sugawara, Yamanishi KDD2017]

[Okada, Yamanishi, Masuda RSOS2019]

- 1) 効率計算可能
- 2) 広い適用可能性
- 3) 高精度かつロバストなモデル選択

$$P(X; K) = \sum_Z P(X, Z; K)$$

観測変数

潜在変数

$$\min_K \{L_{\text{NML}}(X|Z; K) + L_{\text{NML}}(Z; K)\}$$

潜在構造最適化 2

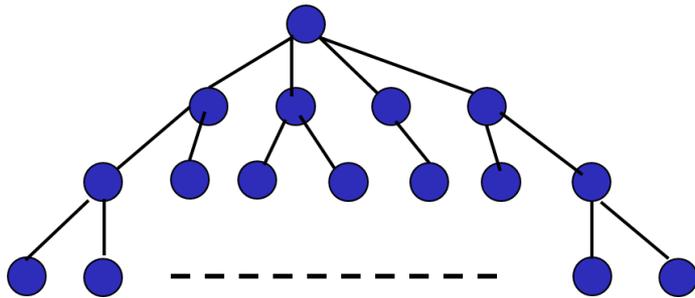
双曲順序付き埋め込み

順序関係を保ちつつ双曲空間に埋め込むには？

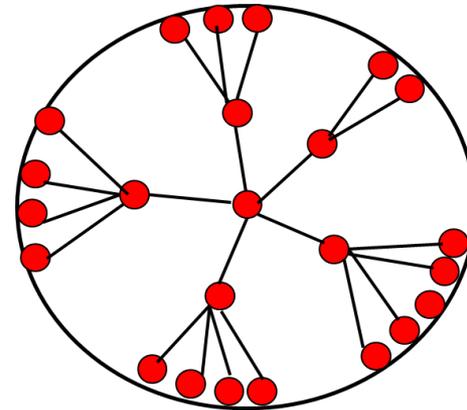
[Suzuki, Wang, Tian, Nitanda, Yamanishi ACML19][Wang, Suzuki, Xu, Fei, Yamanishi IJCAI2019]

[Suzuki, Nitanda, Xu, Wang, Yamanishi, Cavazza ICML2021]

- **Embedding**: map of objects to a metric space



高次元ユークリッド空間



低次元非ユークリッド空間
(双曲空間)

- **Hyperbolic Ordinary Embedding**

$i \leftrightarrow j$ is more similar than $i \leftrightarrow k \Rightarrow d_{\mathbb{H}^D}(x_i, x_j) \leq d_{\mathbb{H}^D}(x_i, x_k)$

- Distance function $d_{\mathbb{H}^D}(\mathbf{x}, \mathbf{y}) = \operatorname{arcosh} \left(1 + \frac{2(\mathbf{x}-\mathbf{y})^\top(\mathbf{x}-\mathbf{y})}{(1-\mathbf{x}^\top\mathbf{x})(1-\mathbf{y}^\top\mathbf{y})} \right)$

効果的な可視化と
ユークリッド空間
よりも高精度な
リンク予測を実現

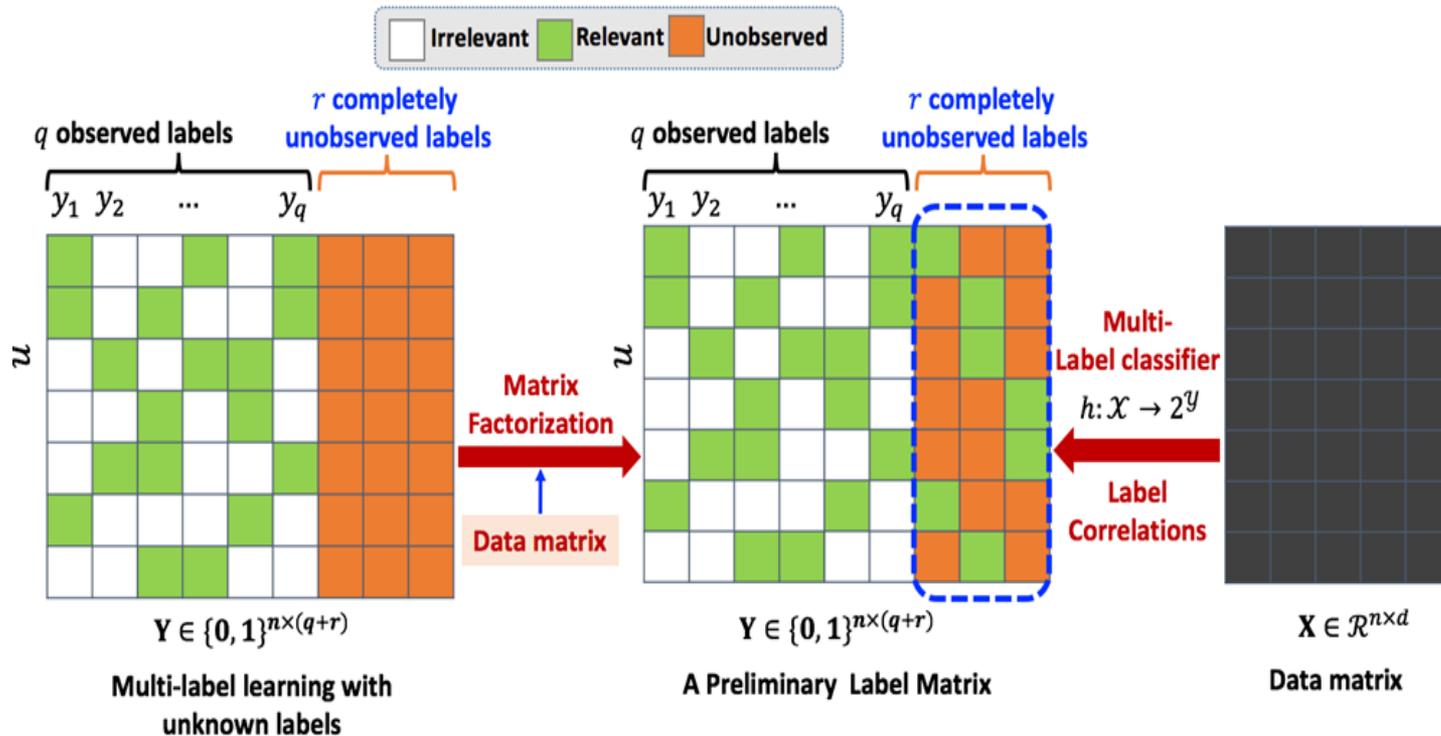
潜在構造最適化 3

潜在ラベルを推定するマルチラベル学習方式

観測ラベルと潜在ラベルの混合データから潜在ラベル推定を初めて実現

[Hunag, Xu, Wang, Fe, Yamanishi IJCAI2019]

[Hunag, Xu, Qian, Wang, Yamanishi DAMI2021]



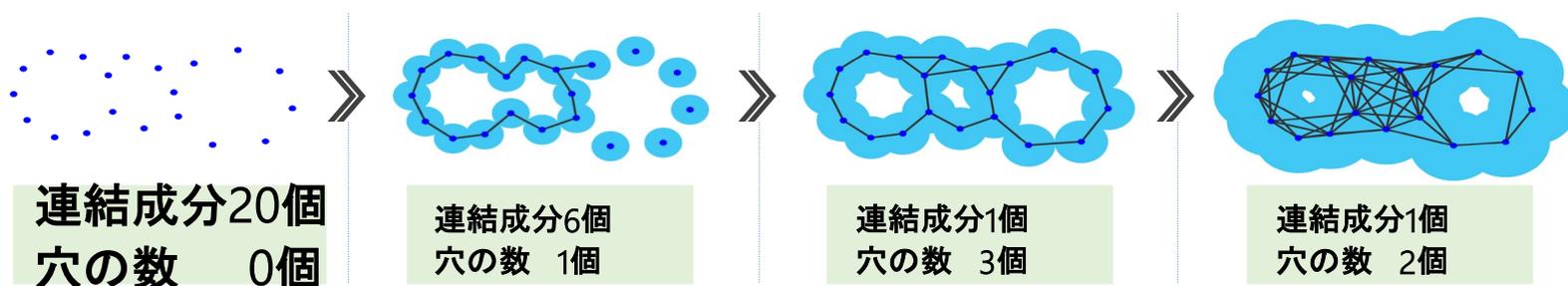
マルチラベル学習に、顕在ラベルと潜在ラベルの相関学習を取り込み潜在ラベル推定を実現

$$\min_{\mathbf{W}, \mathbf{U}, \mathbf{V}} \frac{1}{2} \|\mathbf{X}\mathbf{W} - \hat{\mathbf{Y}}\|_F^2 + \frac{\lambda_1}{2} \|\mathbf{X} - \mathbf{U}\mathbf{V}\|_F^2 + \frac{\lambda_2}{2} \|\mathbf{U}\mathbf{P} - \mathbf{Y}\|_F^2 + \frac{\lambda_3}{2} \sum_{i,j}^l r_{ij} c_{ij} d_{ij}^2 + \lambda_4 \|\mathbf{W}\|_1$$

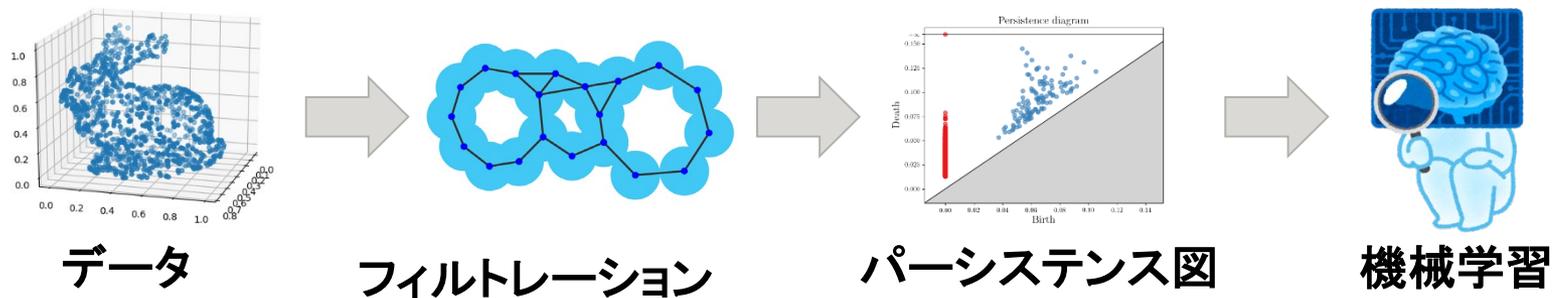
位相的データ解析 (Topological Data Analysis)

■データの「形(トポロジー)」を抽出して解析する手法

■データ点中心の球の半径を大きくしていき和集合のトポロジー変化を記録



■トポロジー情報をパーシステンス図として出力し機械学習等と組み合わせる

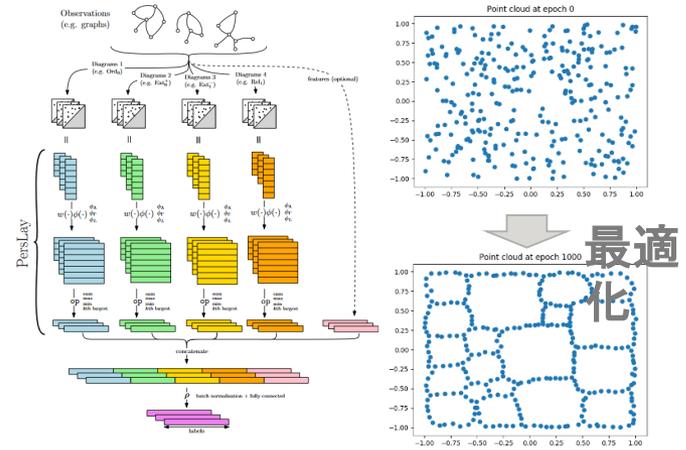


■位相的データ解析はおおまかな「形」に着目するのでノイズ耐性がある

位相的データ解析 研究課題

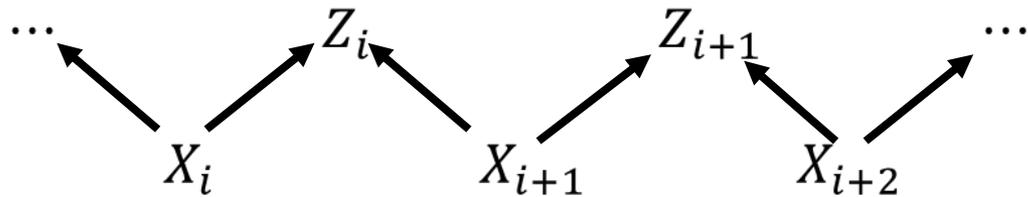
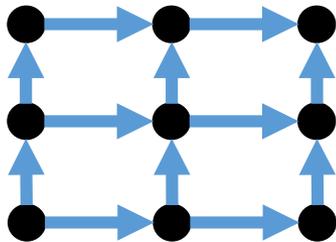
■機械学習との組合せ

- 画像データ・グラフデータの解析への応用
- パーシステンス図のベクトル化手法
- トポロジ的損失関数
- 学習器からのトポロジー情報抽出



■理論的側面

- 複数パーシステンスパラメータを持つパーシステンス加群
- 箆 (quiver) 表現・層理論からのアプローチ



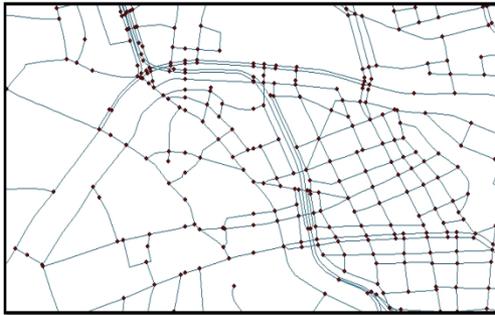
交通リスクマイニング

地形データをネットワーク中心性で数値化、危険パターンを特徴づけ

[Kobayashi, Lee, Matsushima, Yamanishi BigData 2017]

[Lee, Matsushima, Yamanishi DAMI 2019]

地形
データ



高速パターンマイニングに基づく分類

$$\min_{\mathbf{w}} C \sum_{i=1}^m L \left(\sum_{\phi \in \Phi(k)} w_{\phi} \phi(\mathbf{x}_i), y_i \right) + \|\mathbf{w}\|_1$$

ただし, $L(f(\mathbf{x}), y) = \log(1 + \exp(-yf(\mathbf{x})))$



Network中心性
で数値表現

(次数、近傍ノード数、
ページランク、近傍中心
性、平均リンク数 etc)



低コストデータ



- ・危険パターンの抽出
- ・潜在的危険個所の特定

ビッグデータから
コンパクトなパターン
を高速に抽出

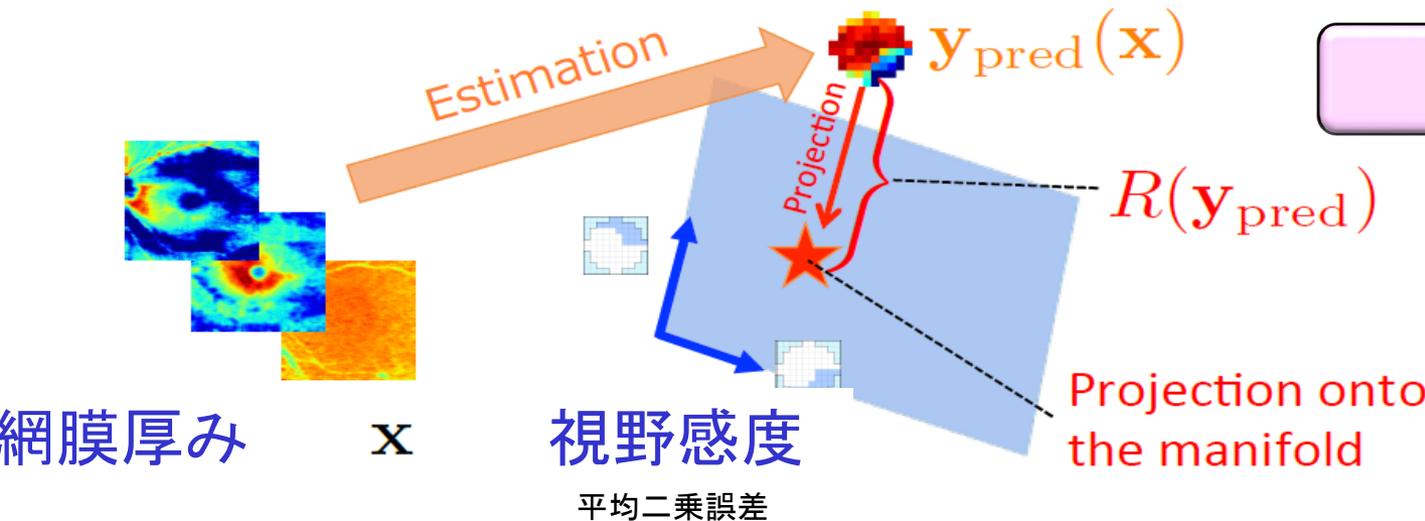
緑内障進行予測 1

パターン正則化

パターン正則化により網膜厚みから視野感度を高精度に推定

[Sugiura, Kiwaki, Siamak, Murata, Asaoka, Yamanishi KDD2018]

[Xu, et al. American Jr. Opht 2020][Hashimoto et al. British Jr. Opht. 2020]



パターン正則化

教師なしで学習した視野パターンからずれないように正則化

	Median	Mean	Worst
CNN [30]	6.68	6.76	21.7
CNN + DA	6.70	6.52	21.7
CNN + DA + PBR1(NMF)	6.41	6.30	23.5
CNN + DA + PBR1(PCA)	6.31	6.26	24.1
CNN + DA + PBR1(AE)	6.39	6.26	24.2
CNN + DA + PBR1(PCA) + PBR2	6.00	6.16	21.4
CNN + DA + PBR1(AE) + PBR2	6.02	6.16	21.5

提案手法の最新結果

$$\frac{1}{N} \sum \|\mathbf{y}_{\text{pred}}(\mathbf{x}_n) - \mathbf{y}_n\|^2 + \frac{\beta}{2} R(\mathbf{y}_{\text{pred}}(\mathbf{x}_n)),$$

$$\mathbf{y}_{\text{proj}} \stackrel{\text{def}}{=} \Theta^T (\Theta \Theta^T)^{-1} \Theta (\mathbf{y}_{\text{pred}} - \mathbf{b}) + \mathbf{b}.$$

東大附属病院眼科教室との共同研究

緑内障進行予測 2

DLLR(deeply regularized latent space linear regression)

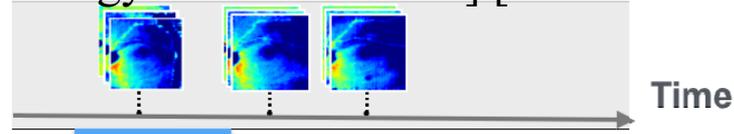
網膜厚と視野感度の時系列の双方を潜在空間で統合、予測

[Tei, Xu, Kiwaki, Wang, Murata, Asaoka, Yamanishi KDD2019]

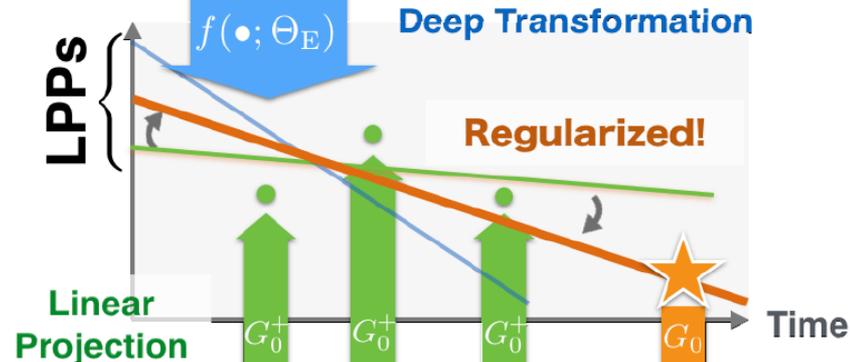
[Xu, et al. Ophthalmology Glaucoma 2020] [Xu et al. KDD2021]



網膜厚時系列



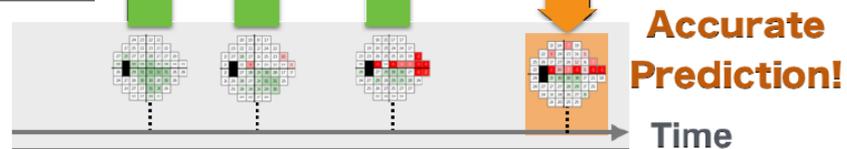
潜在空間



潜在進行
パターン



視野感度時系列



Accurate
Prediction!

予測誤差
(RMSE)比較

Methods	S						
	2	3	4	5	6	7	8
LR	50.66	21.11	13.21	9.76	7.98	6.52	6.06
TSLR	6.27	6.01	5.65	5.54	5.52	5.09	5.02
MTMC	5.65	5.47	4.99	4.82	4.83	4.67	4.62
DLLR	4.96	4.76	4.61	4.49	4.48	4.46	4.43

視野感度と網膜厚
の双方の活用で
最高精度達成

2015－2016の修論テーマ

- ・ネットワーク中心性を用いたテンポラルネットワークの変化検知
DSAA16採択
- ・ヘテロデータに基づく緑内障進行予測のための個々の患者に特化した病状のモデル化
Healthinf16採択 2015研究科長賞(創造)
- ・記述長最小化原理に基づく関係データ学習の研究
SDM16採択
- ・Support Vector Machine 及びトピックモデルを用いた位置情報による個人識別の研究
- ・完全変数化NML 符号長に基づく正準相関分析のランク推定の研究
DSAA17採択
- ・Study on Learning from Nonstationary Time Series
DSAA15採択
- ・サポートベクトルマシンの最適化アルゴリズムの研究
- ・野球データ解析に対する機械学習アプローチ
DSAA16採択
- ・交通データからのリスク因子発見に関する研究
2016研究科長賞(創造)
- ・緑内障眼における網膜層厚視野感度間マルチビュー学習
KDD17採択
- ・A Study on Model Selection for Hierarchical Mixture Models
2016研究科長賞(数理)
- ・畳み込み行列分解のパラメータ推定と構造選択
KDD17採択
ICDM16採択

2017-2020の修論テーマ

- Model Selection for Non-negative Tensor Factorization with MDL **Entropy採択**
- パターン正則化学習を用いた網膜層厚からの視野感度推定と知識発見 **KDD18採択**
- 分解型正規化最尤符号長を用いた連続値潜在変数モデルの選択 **DAMI採択** **RSOS採択**
- 統計的モデル選択を用いた勾配ブースティング木の早期停止ルールの研究 **KDD19採択**
- A machine learning approach to glaucoma progression prediction **KDD19採択**
- 双曲空間における測地線更新アルゴリズムの研究 **2016研究科長賞(創造) AIPラボ長賞**
- グラフ埋め込みにおける正則化の研究 **2016研究科長賞(創造) AIPラボ長賞**
- 敵対的学習における目的関数の設計及び解析に関する研究
- Dynamic Nondiagonal Mixture of Dirichlet Distributions for Economic Network
- Network Change Detection Based on Random Walk in Latent Space
- Cluster Structure Analysis in Finite Mixture Models With Component Overlap
- Change Sign Detection with Two-Stage MDL Change Statistics

2017-2020の博論テーマ

- Learning High-dimensional Models with the Minimum Description Length Principle

SDM18採択

MLJ採択

AISTATS19採択

- Machine Learning over Space Forms

ISIT18採択

IEEE IT採択

ACML18採択

- Detecting Model Changes and their Early Warning Signals with the Minimum Description Length Principle

BigData18採択

BigData19採択

- An Information-Theoretic Approach to Hierarchical Change Detection

Entropy採択

ICDM20採択

2016-2020 の卒論テーマ

- ・可変窓上の記述長最小化に基づく時系列データの変化検知
- ・ネットワーク情報を用いた道路の潜在的危険予測
- ・オンライン非定常ロバスト学習とその異常検知への応用
- ・複数の関係パターンを持つマルコフ確率場のMDL原理に基づく構造学習
- ・緑内障診断のための網膜厚に対する視野感度の微分値の異常検知
- ・混合分布を用いるパターン正則化学習とその眼科学への応用
- ・潜在空間埋め込みを用いるベイズ実験計画の研究
- ・有限混合モデルの変化予兆検知とCOVID-19の感染爆発検知への応用

BigData17採択

BigData17採択

2017工学部長賞(計数)