不等間隔な標本化と量子化を用いた時系列パターン抽出手法の開発と評価

大崎美穂　同志社大学工学部
※阿部秀尚　島根大学医学部
北口真也　静岡大学大学院情報学研究科
横井英人　香川大学医学部
山口高平　慶應義塾大学理工学部

発表内容

1. 背景と目的

■時系列データマイニングのニーズは高い.

【応用例】 病気の発見や予後診断, 株価の変動予測,
交通流量の分析など

■時系列データマイニングには, 大別すると
記号値を扱う手法, 数値を扱う手法がある.

【数値を扱う手法】 データの数理構造を解析しモデル化して,
モデルとの合致度合いでパターンを抽出

これ以降, 簡略のため,
「数値時系列データ」を「データ」
「数値を扱う時系列データマイニング」を
「時系列データマイニング」と呼ぶ.

1. 背景と目的

■有益なパターンを得るには，数理的な正確さと
ユーザの主観の適切なバランスが必要である．

【問題点】 正確だが複雑すぎる手法は使われるか？
ユーザのバイアスが大きすぎる手法は信頼できるか？

【研究目的】
「データの数理構造」と「ユーザの主観」の両方を
考慮したパターン抽出手法を提案する．

【研究アプローチ】
- 標本化･量子化の間隔を調整し，
特徴的なデータ波形の取り出しを試みる．
- シンプルな処理，ユーザによる処理パラメータ設定
により，ユーザの理解しやすさと適度な介入を目指す．

発表内容

2. 関連研究

従来の時系列パターン抽出手法
時系列パターン抽出に適用可能な手法

A. **データが線形性・定常性・周期性などの性質を持つ場合**

厳密な数理モデルに基づく．自己回帰，ARIMAなどの統計手法，フーリエ・ウェーブレット変換，カオス・フラクタル解析などの信号処理手法が挙げられる．

B. **データがこのような性質を持たない，もしくは持っていても事前には分からない場合**

統計，信号処理，機械学習の考えを融合し，柔軟な数理モデルと領域知識の導入に基づく．以下の処理からなる時系列データマイニング手法が挙げられる．

1. **パターンの開始点・終了点を見つけ出す．**
2. **パターンの候補を挙げる．**
3. **パターン候補からパターンを絞り込む．**

2. 関連研究

時系列データマイニングにおける関連手法

【これらの立場】

-
-
-
- ユーザの主観（対象問題の領域知識や着眼点）に合う興味深いパターンを得られないことがある．

【我々の立場】

-
-
-
- バイアスを目指す．

我々の立場から問題解決するには，具体的なケーススタディに基づき，以下を議論する必要がある．

「どのようなユーザを対象とするか」
「ユーザにとって分かりやすい処理過程とは何か」
「客観的な正確さと主観的な興味のバランスを保つには，どの処理パラメータをユーザに開放すべきか」

GAによる処理パラメータ最適化

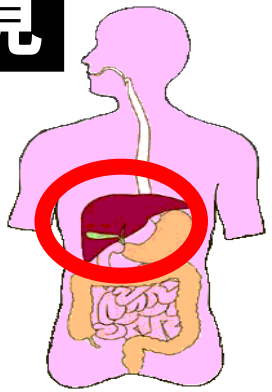
2. 関連研究

ケーススタディ：肝炎データからの知識発見

データセット

慢性肝炎
検査履歴データ

属性数: 957
患者数: 771
レコード数: 約1,600,000

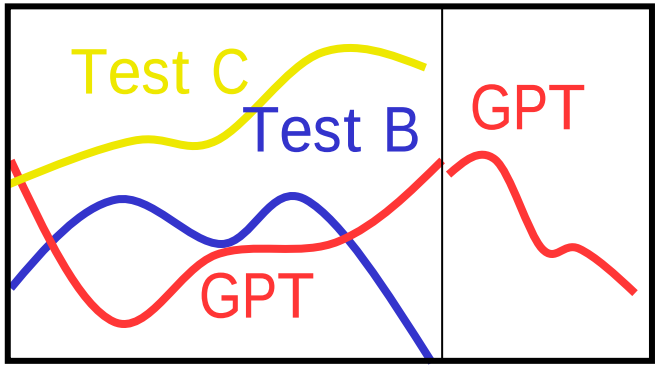


ルールの設計

属性:
過去の様々な検査値パターン群

クラス:
将来のGPT値パターン

IF (現在の病状)　THEN (将来の病状)



マイニングの過程

マイニングシステム改善⇔医師によるルール評価
を2セット繰り返し，ルールの洗練化を試みた．

2. 関連研究

先行研究から得られた知見

マイニングシステム ver.1

- 医学的に無意味なパターンが多い.
- パターン数やパターンを支持する事例数を医師が事前に決定する方が, 領域知識を反映できる.
- EMアルゴリズムは医師にとって直感的に理解しにくい.

マイニングシステム ver.2

- 医学的に無意味なパターンだけでなく, 医学的に興味深いパターンも得られた.
- パターンの波形が, 生データの波形とかなり異なる (ピークの繰り返しという特徴的な波形が損なわれる).
- K-Meansアルゴリズムは医師にとって直感的に理解しやすい.

発表内容

1. 背景と目的
2. 関連研究
3. 時系列パターン抽出手法の提案
   - 3.1 フレームワーク
   - 3.2 標本化
   - 3.3 量子化
   - 3.4 代表抽出
4. 評価実験
5. まとめと今後の課題

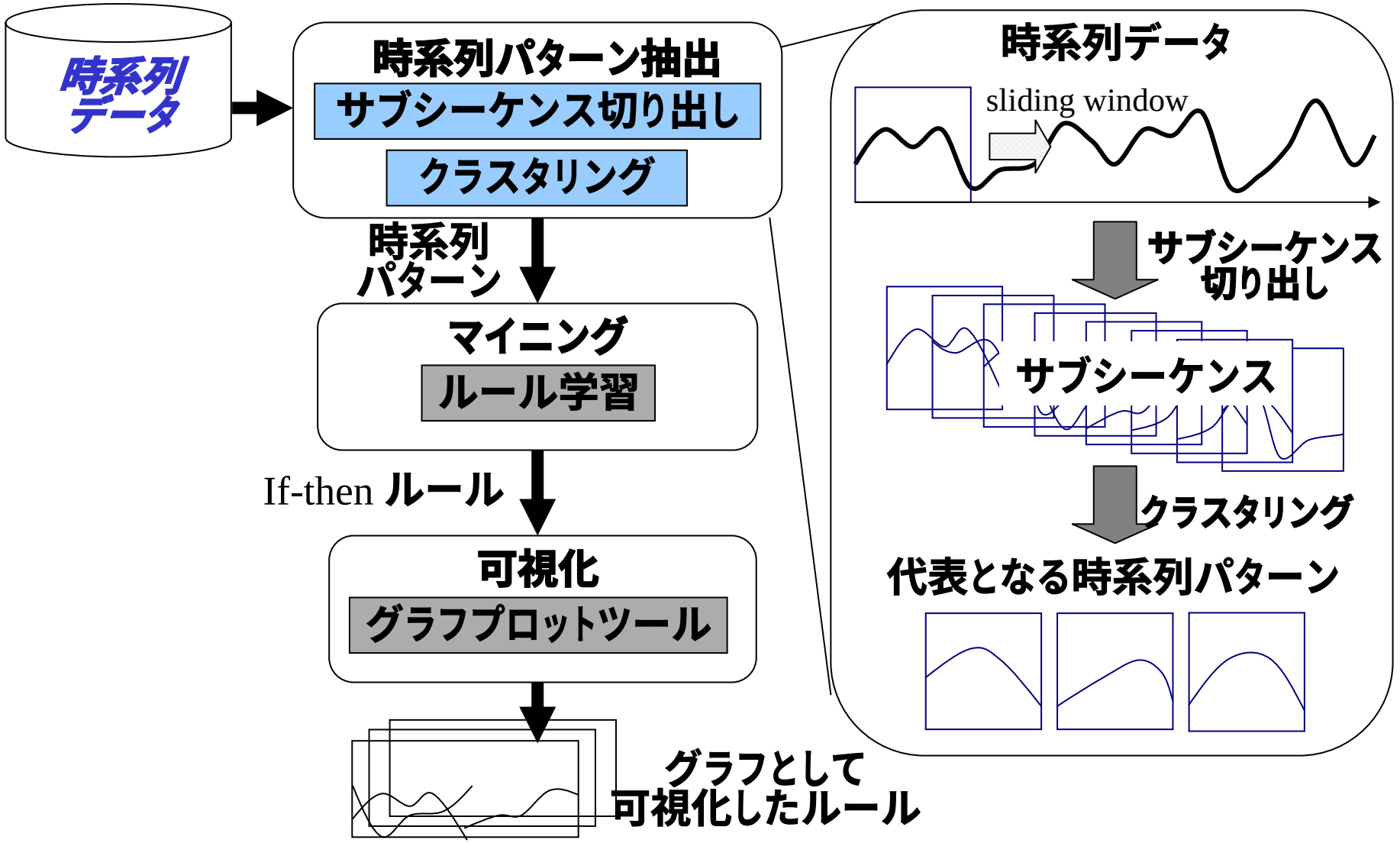
3. 時系列パターン抽出手法の提案

3.1 フレームワーク

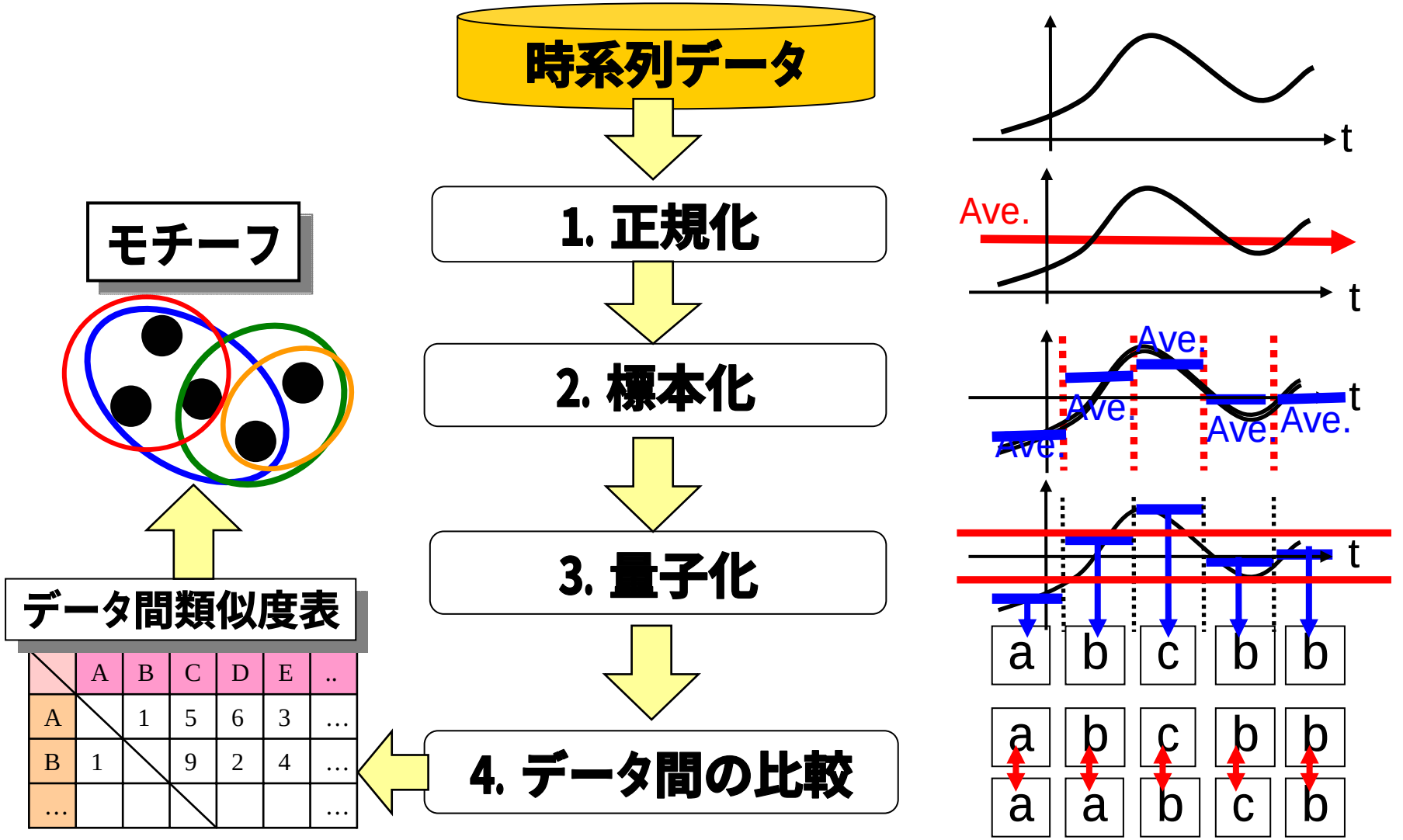
- **時系列データからのパターン抽出**[Das98]
  - **利点：波形の直感的把握**
  - **欠点：クラスタリングに起因するパラメータの不安定性** [Keogh03]
- PAA (Piecewise Aggregate Approximation)**の特徴**
  - **利点：微小変動の除去しつつ，波形を適度に近似**
  - **欠点：正規化によって絶対値情報が損なわれる**

波形の特徴（形状・絶対値）を保存した
時系列パターン抽出手法

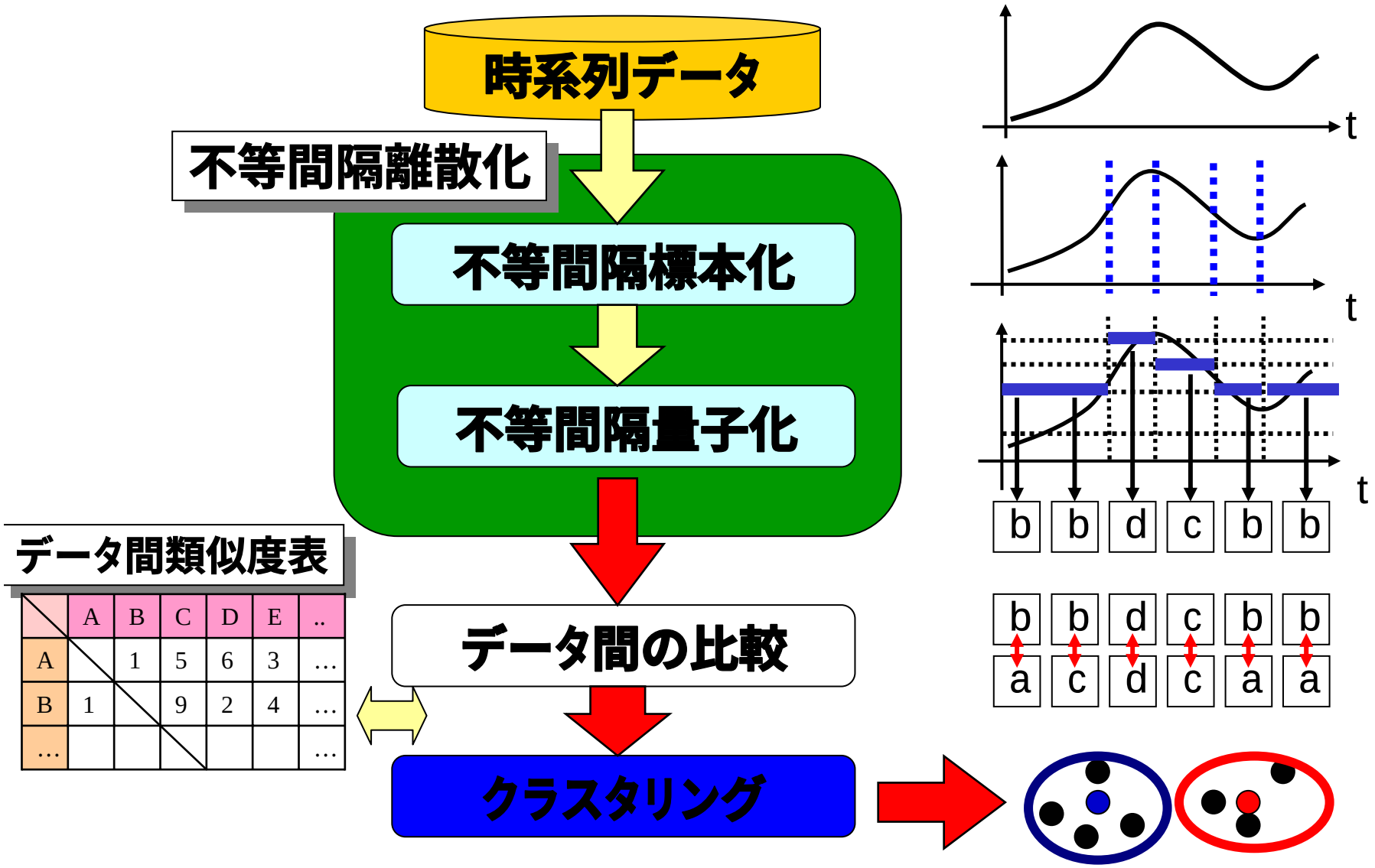
サブシーケンス切り出し＋クラスタリングによる時系列パターン抽出手法（マイニングシステムVer.1/2）



モチーフ抽出法 (Keogh, UC Riverside, 2002)

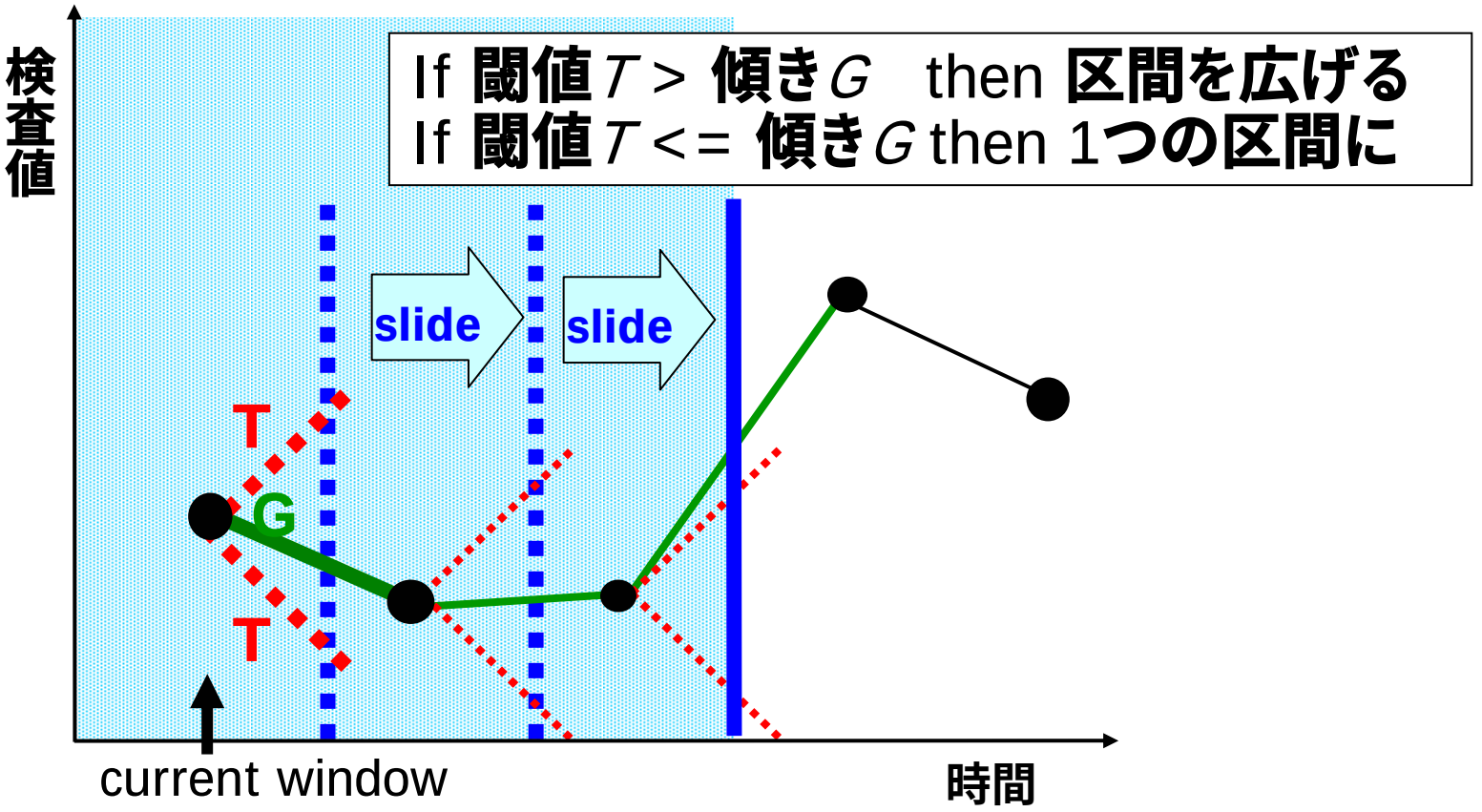


不等間隔離散化を用いたパターン抽出

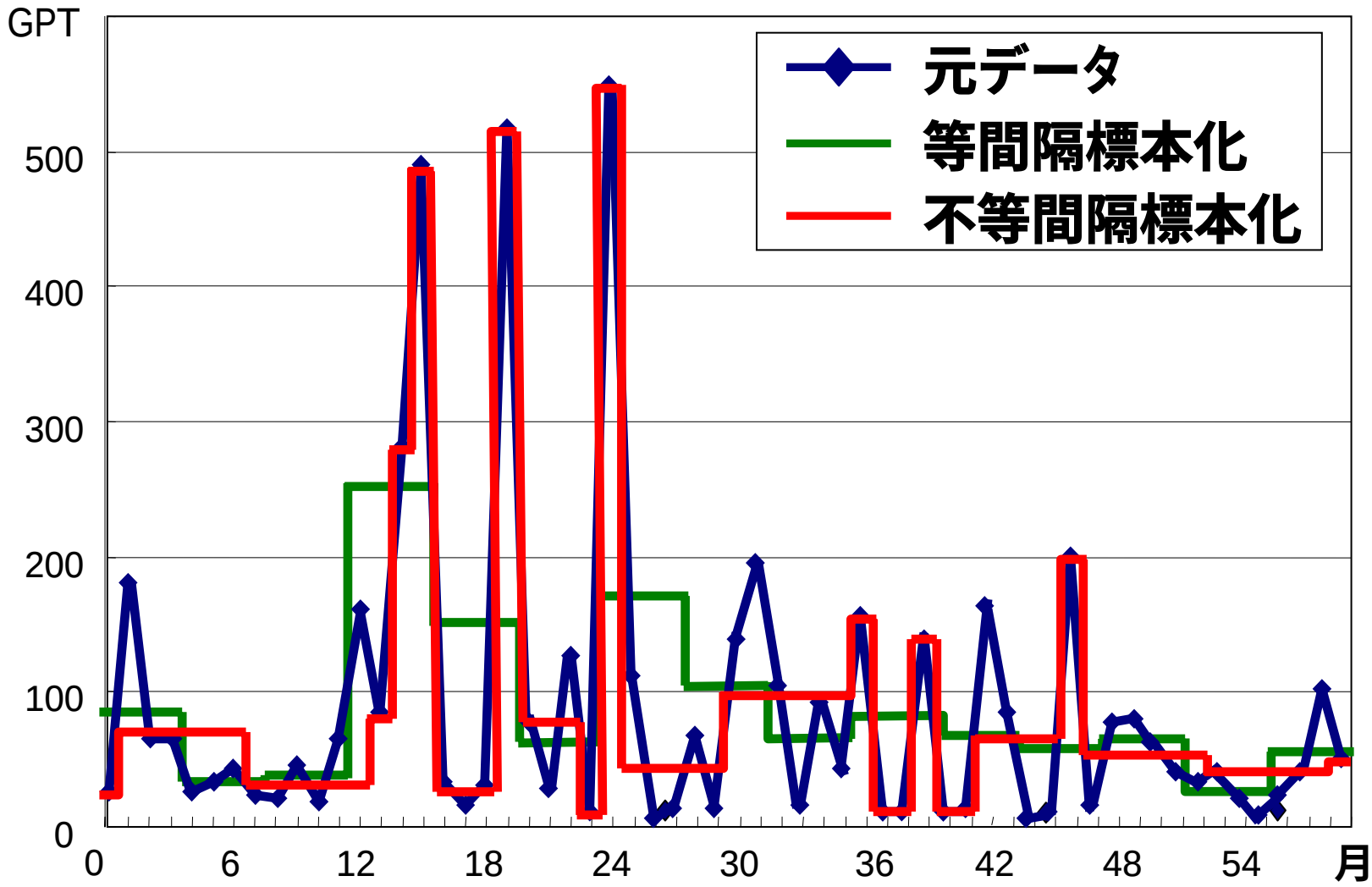


不等間隔標本化

- 区間決定の基準値：2点間の傾き
  - データの局所的形状を残すことができる
  - 閾値T は領域知識に合わせて調整可能



不等間隔標本化の結果


GPT
500
400
300
200
100
0
0
6
12
18
24
30
36
42
48
54
月
元データ
等間隔標本化
不等間隔標本化

不等間隔量子化

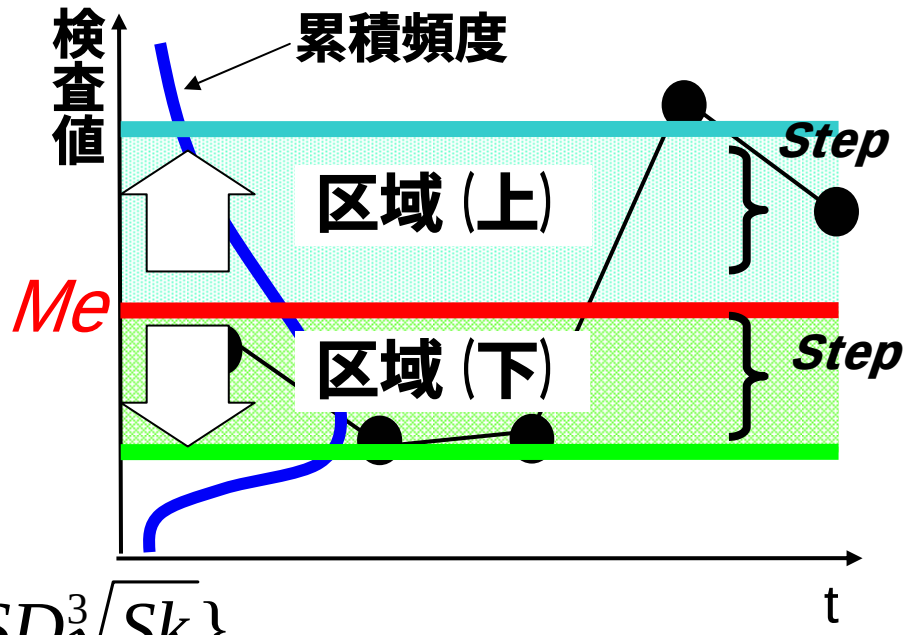
- データ全体の分布に注目する
  - 正規分布とは限らない
    - → 基準値：中央値 Me
    - → 分布の歪みを考慮した基準

区域（上），（下）の境界線：BUW, BLW
中央値：Me，標準偏差：SD，歪度：Sk
区間番号：k (k=1,…)
ステップ幅の調整係数：w

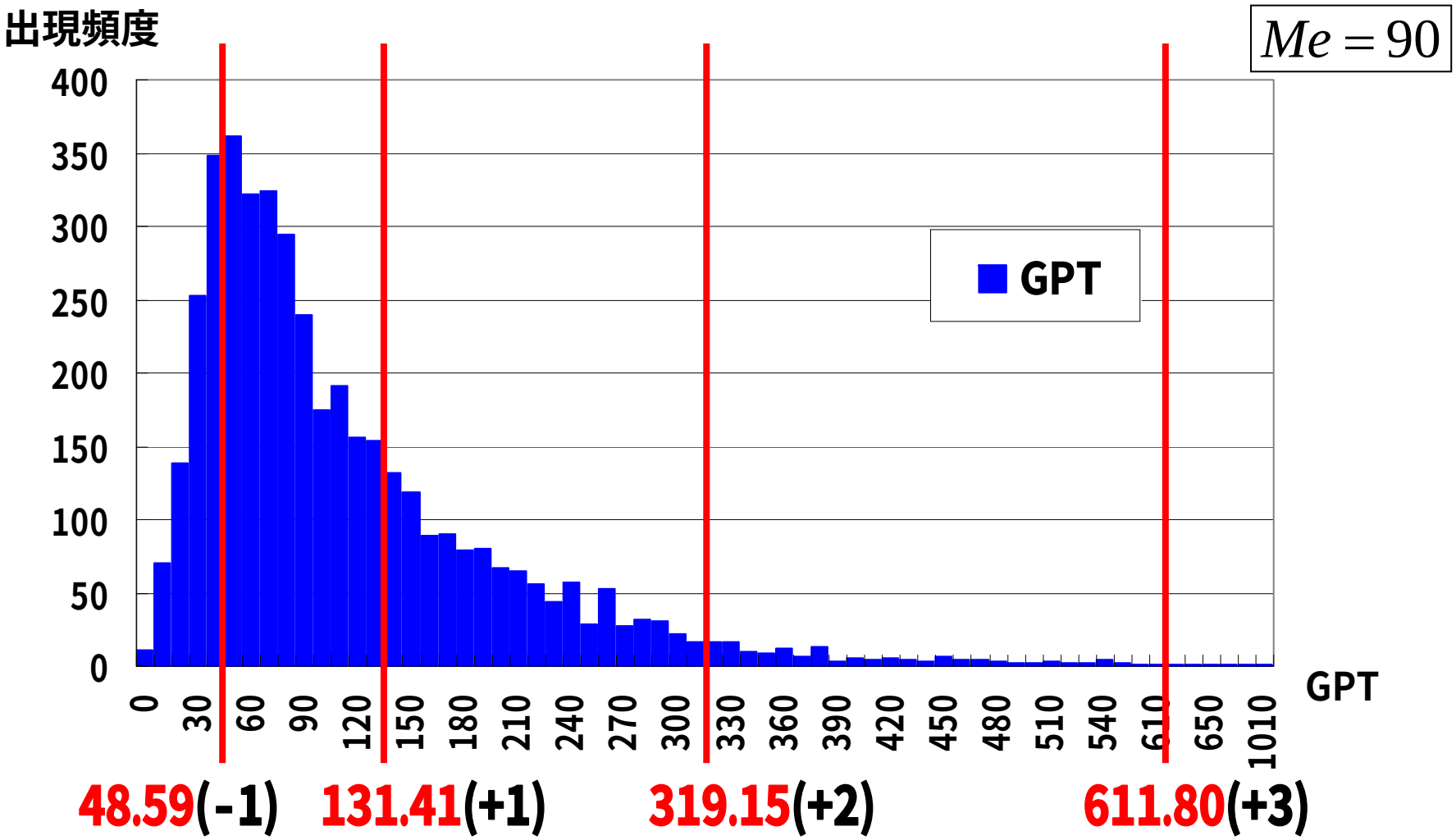
$$BUW = Me + Step$$

$$BLW = Me - Step$$

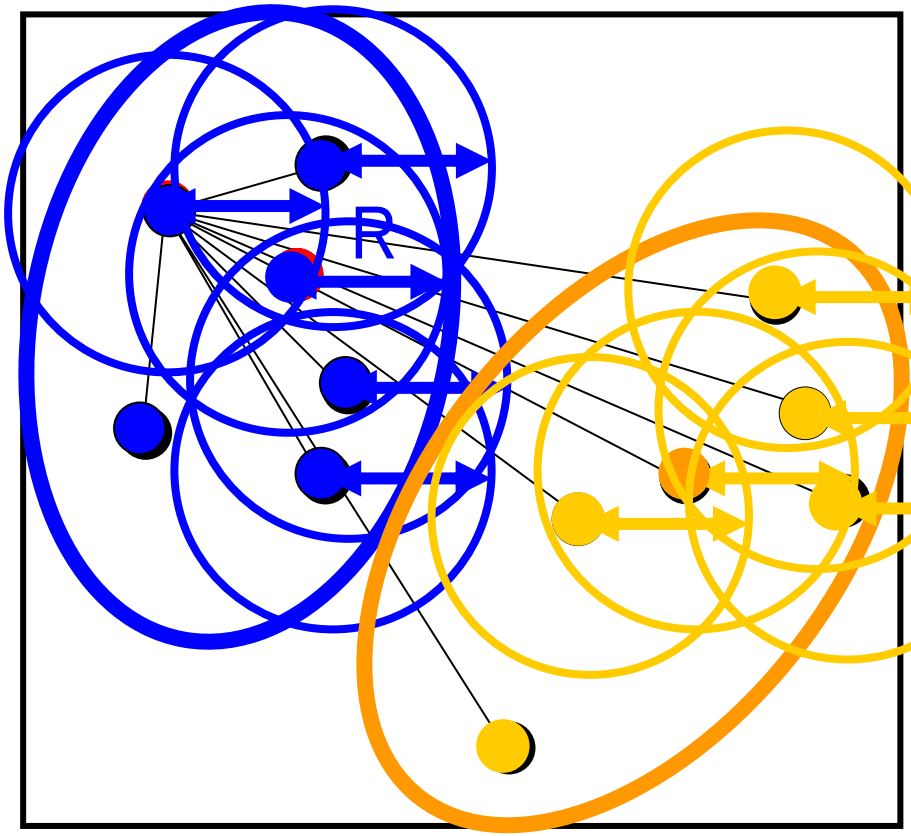
$$Step = w * \{(2k-1)\frac{SD}{2} + (k-1)SD\sqrt[3]{Sk}\}$$



不等間隔量子化の結果
(例：IFN投与前5年間のGPT検査データ)

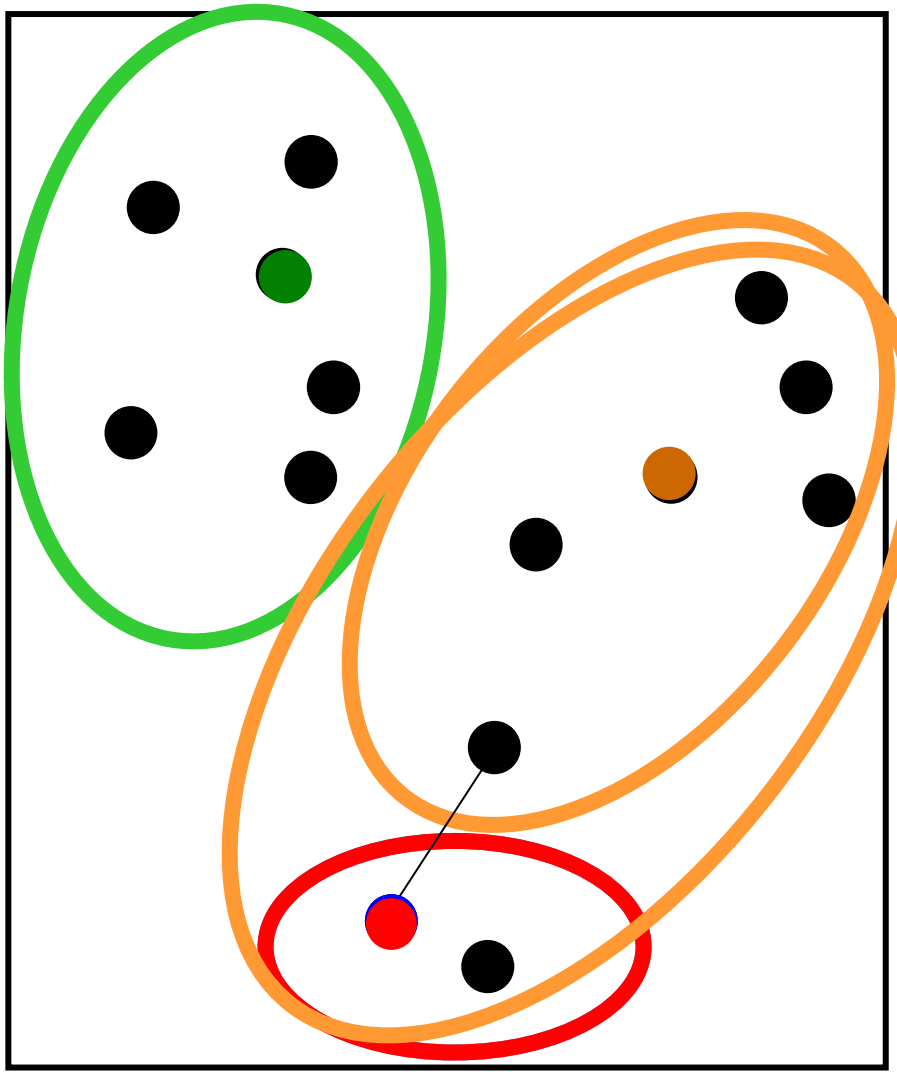


階層的･群平均に基づくクラスタリング (19)
(初期クラスタ生成)



1. 各データ間の距離を計算する
2. 半径Rの円内に他データを多く含むデータをクラスタの代表パターンとする
3. 半径Rに含まれたデータをクラスタの要素とし，追加されたデータに対し半径Rの円を描く
4. 3.をN回繰り返す
5. 2~4.を全データに対し繰り返す
6. どのクラスタにも含んでいないデータは，最も近い代表パターンのクラスタの要素とする
7. 各代表パターンの再設定を行う

階層的･群平均に基づくクラスタリング (初期クラスタ結合)



1. クラスタの要素数が最も少ないクラスタを選択
2. 選択したクラスタの代表パターンと類似データを最も多く含むクラスタを選択
3. 1.2.で選択したクラスタ同士を結合
4. 1.~3.を指定クラスタ数まで繰り返す
5. 各代表パターンの再設定を行う

発表内容

4. 評価実験

- **実験の目的**
  - **クラスタリング手法としての性能の評価**
    - **不等間隔離散化手法の有効性**
    - **類似度算出法の比較**
- **評価基準**
  - **正解率：本来の分類との一致度**

$$\text{正解率}[\%] = \frac{\text{正解データ数}}{\text{全データ数}} \times 100$$

ここでの「正解」は，クラスタの多数を占めるクラスと一致しているか

  - **一元配置分散分析F値：代表元（パターン）の分離度**

$$F = \frac{\text{パターン間の不偏分散}}{\text{各パターン内の不偏分散}}$$

4. 評価実験(cont.)

- **データセット**[UCR　TSDMA]：
  GunX,　EGC_znorm205, Tracedata, Leaf_all
- **比較対象**：K-means（**実装は**Weka**を利用**）(1)
- **実験結果の比較**
  - **比較1**：**不等間隔離散化の効果**（(1)vs.(2)vs.(3)）
  - **比較2**：**不等間隔離散化に加えるクラスタリングの比較**（(3)vs(5)）
  - **比較3**：**類似度算出法の比較**（(3)vs(4)，(5)〜(9)）

類似度（距離）の設定	差異	差の絶対値	差の2乗
量子値間の差異だけを規定	(2)(3)(5)		
量子値の間隔を1と規定		(6)	(7)
各区域内の平均値		(8)	(4)(9)

4. 評価実験

実験結果

正解率	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
GunX	50.00	70.00	57.00						
ECG_znorm 205	100.00	100.00	100.00						
Tracedata	53.00	74.00	52.50						
Leaf_all	32.13	35.29	28.51	28.73	36.65	32.81	33.25	32.81	32.81
平均値	58.78	69.82	59.50	59.43	63.04	61.70	62.56	62.45	62.70

(1)K-meansと比べて，時系列離散化を行った(2)(3)の性能が向上

(2)PAAによる離散化に比べて，(3)不等間隔離散化の性能が低下

一元配置分散分析F値	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
GunX	152.05	-	166.28	111.96	153.02	141.74	135.98	144.18	123.84
ECG_znorm 205	57.72	-	65.82	51.87	61.84	61.84	61.84	61.06	67.43
Tracedata	265.89	-	151.11	179.49	187.35	119.18	197.12	113.64	113.64
Leaf_all	18.38	-	9.29	15.80	44.89	48.10	46.36	44.63	45.91
平均値	123.51	-	98.13	89.78	111.78	92.72	110.32	90.88	87.71

4. 評価実験

実験結果

正解率	(1)	(2)	(3)	(4)	(5)				
GunX	50.00	70.00	57.00	56.00	64.00				
ECG_znorm 205	100.00	100.00	100.00	100.00	100.00				
Tracedata	53.00	74.00	52.50	53.00	51.50	52.50	55.50	52.50	52.50
Leaf_all	32.13	35.29	28.51	28.73	36.65	32.81	33.25	32.81	32.81
平均値	58.78	69.82	59.50	59.43	63.04	61.70	62.56	62.45	62.70

(3)PAA類似度＋K-means と(5)PAA類似度＋階層的・群平均に基づくクラスタリングでは，(5)の手法で性能が向上

一元配置分散分析F値	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
GunX	152.05	-	166.28	111.96	153.02	141.74	135.98	144.18	123.84
ECG_znorm 205	57.72	-	65.82	51.87	61.84	61.84	61.84	61.06	67.43
Tracedata	265.89	-	151.11	179.49	187.35	119.18	197.12	113.64	113.64
Leaf_all	18.38	-	9.29	15.80	44.89	48.10	46.36	44.63	45.91
平均値	123.51	-	98.13	89.78	111.78	92.72	110.32	90.88	87.71

4. 評価実験

実験結果

正解率	(1)	(2)	(3) ⟷ (4)		(5)	(6) ⟷ (7)		(8) ⟷ (9)	
GunX	50.00	70.00	57.00	56.00	64.00	61.50	61.50	64.50	65.50
			100.00	100.00	100.00	100.00	100.00	100.00	100.00
Tracedata	53.00	74.00	52.50	53.00	51.50	52.50	55.50	52.50	52.50
			28.51	28.73	36.65	32.81	33.25	32.81	32.81
			59.50	59.43	63.04	61.70	62.56	62.45	62.70

差の2乗を使う
(7)(9)の性能が向上

量子値に反映する
絶対値情報を抽象化
した順である(4)→(3),
(9)→(7)→(5)に性能が
向上

			(3)	(4)	(5)	(6)	(7)	(8)	(9)
GunX	152.05	-	166.28	111.96	153.02	141.74	135.98	144.18	123.84
ECG_znorm 205	57.72	-	65.82	51.87	61.84	61.84	61.84	61.06	67.43
Tracedata	265.89	-	151.11	179.49	187.35	119.18	197.12	113.64	113.64
Leaf_all	18.38	-	9.29	15.80	44.89	48.10	46.36	44.63	45.91
平均値	123.51	-	98.13	89.78	111.78	92.72	110.32	90.88	87.71

評価実験結果の考察

- **比較1** ((1)vs.(2)vs.(3))
  - **波形を離散化することによる微小変動の除去は有効**
  - **不等間隔離散化は各サブシーケンス毎の正規化による記号化ほど波形の特徴を保存していない**
- **比較2** ((3)vs.(5))
  - **階層的クラスタリングの有効性を確認**
- **比較3** ((3)vs.(4), (5)〜(9))
  - **差の2乗は大きな差を強調するフィルタの役割**
  - **データ間の類似度計算法は記号間の差異の総和が有効**

発表内容

1. 背景と目的
2. 関連研究
3. 時系列パターン抽出手法の提案
   - 3.1 フレームワーク
   - 3.2 標本化
   - 3.3 量子化
   - 3.4 代表抽出
4. 評価実験
5. まとめと今後の課題

5. まとめと今後の課題

- 不等間隔離散化に基づく時系列パターン抽出手法の提案
  - 波形の特徴に注目できるようパラメータを設定
  - クラスタリング手法として従来手法と比較して同等以上の性能
- 不等間隔離散化の改善
  - 2つの処理の影響度合いについて検証
  - 不等間隔量子化の検討
- 他の階層的クラスタリング手法の検討
- 領域専門家の主観に合致する時系列パターン抽出実験