統計的有意性を担保するパターンマイニング技術

杉山　麿人

パターンマイニングの原理について紹介し，特に統計的に有意なパターンを発見する技術について詳しく紹介する．パターンマイニングとは，データに基づいて，パターンと呼ばれる組合せ的規則の中から重要なもののみを網羅的に見つけ出す解析技法である．この技術を統計的仮説検定と組み合わせることで，発見されたパターンに含まれてしまう偽陽性の割合を適切に制御することができる．また，パターンマイニングと情報幾何との関連についても紹介する．

キーワード：パターンマイニング，半順序集合，統計的有意性，多重検定補正，情報幾何

1. はじめに

パターンマイニング (pattern mining) とは，パターン (pattern) と呼ばれる膨大な数の組合せ的規則の中から，重要なもののみを効率的に取り出すことを目的とする技術の総称であり [1]，データマイニングやビッグデータ解析において欠かせない．たとえば，店舗の商品の全組合せの中からよく買われた組合せのみを発見するアイテム集合マイニング (itemset mining) [2, 3] や，化合物の集合において共通して出現している部分構造を見つける部分グラフマイニング (subgraph mining) [4, 5]，頻出するアイテムの系列を列挙する系列パターンマイニング (sequential pattern mining) [6, 7] などがある．Uno et al. [8–11] が開発したアルゴリズム LCM は世界最速のアイテム集合マイニングアルゴリズムとして知られており（コンペティション FIMI04 で優勝），Inokuchi et al. [4] が世界で初めて部分グラフマイニングのアルゴリズム構築に成功するなど，日本のコミュニティも存在感を発揮してきた．

本稿では，まずパターンマイニングについて，なるべく基本的な原理を取り出し定式化する（2 節）．そして，パターンマイニングにおける最先端のトピックである統計的有意パターンマイニング (significant pattern mining) について紹介する（3 節）．これは，発見されたパターンの p 値を計算することで，それらに含まれている偽陽性の割合を適切に制御する技術である．Terada et al. [12] が提案した LAMP という手法に端を発し，部分グラフマイニングへの適用 [13] や省メモリかつ高速な手法 Westfall-Young light [14] などが開発されている．生命科学などの科学的発見への応用が期待されており，ゲノム解析への適用 [15] やソフトウェアの整備 [16] なども進んでいる．現在は，LAMP を開発した東京大学の津田らを中心としたグループと，LAMP の発展をはじめとしたさまざまな手法を開発している ETH Zürich の Borgwardt らのグループ，そして，このトピックに最も長く取り組んでいる Monash 大学の Webb と Aalto 大学の Hämäläinen らのグループが，精力的に研究を進めている．

最後に，筆者の最近の取り組みとして，情報幾何的な解析とパターンマイニングの関連 [17] について紹介する（4 節）．パターンマイニングが扱う空間には，情報幾何で知られている双対平坦な構造 [18–20] が潜んでおり，情報幾何的な解析が可能となる．

2. パターンマイニングの定式化

パターン全体からなる集合を S とし，その要素 $x \in S$ をパターンとする．集合 S は帰納的可算と仮定しておく．与えられるデータは S の部分多重集合 $D \subseteq S$ として扱われ，$\mathbf{1}_D : S \to \mathbb{N}$ を D の重複度関数とすると $\mathbf{1}_D(x)$ は D 中の x の個数を表す．また，$|D|$ は D の要素数（サイズ）を表し，$|D| = \sum_{x \in S} \mathbf{1}_D(x)$ である．パターンマイニングの目的は，何らかの尺度に基づいてパターン $x \in S$ に重要度を与える実数値関数 $\xi : S \to \mathbb{R}$ を定めて，各パターン x に対して重要度 $\xi(x)$ がデータ D を用いて計算できるとき，$\xi(x)$ が与えられたしきい値 σ を超えるパターン x をすべて見つけることである．

パターンマイニング問題：

集合 $F = \{x \in S \mid \xi(x) \geq \sigma\}$ の列挙．

パターンマイニングは一見簡単そうに見える．特に，S は多くの場合有限集合なので，S の要素 x を順に

すぎやま　まひと
大阪大学産業科学研究所
〒 567–0047 大阪府茨木市美穂ヶ丘 8–1
独立行政法人科学技術振興機構，さきがけ
〒 332–0012 埼玉県川口市本町 4–1–8
mahito@ar.sanken.osaka-u.ac.jp

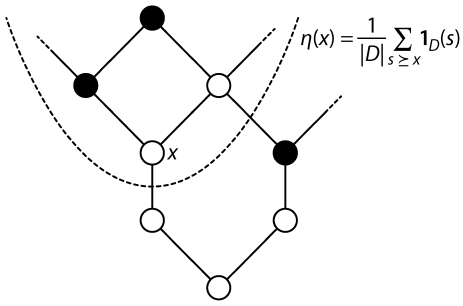

図 1　パターン空間 S

チェックして，重要度 $\xi(x)$ を計算し，しきい値を超えていれば x を出力する，という最も素朴な生成テストアルゴリズムで原理的には解ける．しかし，パターンマイニングの困難さは，S の巨大さにある．たとえば，アイテム集合マイニングの場合，アイテムの集合 $V = \{1, 2, \ldots, n\}$ とすると，それらの全組合せの中から重要なものを出力することが目的となる．したがって，S は V のべき集合 2^V となるため，たとえば $n = 10000$ だと S のサイズは $|S| = 2^{|V|} = 2^{10000}$ となり，生成テストアルゴリズムの適用は現実的には不可能である．

そこで，このパターン空間の組合せ爆発を回避し，現実的な時間でパターンマイニングを達成するための鍵が，パターン空間の構造を利用した列挙の効率化である．一般には，パターンの構造として半順序 (partial order) を仮定し，パターン全体の空間 S を半順序集合 (partially ordered set; poset)$(S, \preceq)$ として扱う．集合 S は小さい要素から順に枚挙可能であるとする．ここで，半順序 $\preceq$ とは，以下の三つを満たす関係であり，数学や計算機科学における基本的な構造である [21, 22]．任意の $x, y, z \in S$ に対して，

1. $x \preceq x$（反射律）．
2. $(x \preceq y$ かつ $y \preceq x) \Rightarrow x = y$（反対称律）．
3. $(x \preceq y$ かつ $y \preceq z) \Rightarrow x \preceq z$（推移律）．

パターン空間 $(S, \preceq)$ の概念図を図 1 に示す．黒丸は D に含まれているパターンを表し，点線で囲まれた部分が要素 x の上方集合を表す．二つのパターン（丸）が線で結ばれているとき，下のパターンは上のパターンより半順序 $\preceq$ に関して小さい．

パターン x, y が $x \preceq y$ の関係をもつとき，y は x よりも精密であり，逆に x は y よりも一般的である．ここで，x から y の導出を精密化 (refinement) といい，y から x の導出を汎化 (generalization) と呼ぶ．また，パターン x よりも精密なパターン全体の集合を x の上方集合 (upper set) といい，$\uparrow x = \{s \in S \mid s \succeq x\}$ と書く．前述のアイテム集合マイニングの場合は，集合の包含関係で順序が定まり S に構造が入る．すなわち，パターン（アイテム集合）$x, y \in S$ に対して，$x \subseteq y$ なら $x \preceq y$ である．

さらに，パターン空間の構造を利用するために，重要度 ξ に関する大小関係 $\xi(x) \leq \xi(y)$ が，パターン間の順序 $x \preceq y$ と一致するように定める．すなわち，ξ が順序準同型写像 (order homomorphism) [23] であることを要求し，$x \preceq y \Rightarrow \xi(x) \leq \xi(y)$ の関係が成り立つと仮定する．これは，構文的な世界での構造 $x \preceq y$ と，意味的な世界での構造 $\xi(x) \leq \xi(y)$ が一致することを意味する．ただし実際には，ξ が大きいものを取り出したいので，たとえば $\xi'(x) = 1/\xi(x)$ として ξ' が順序準同型写像になるように定める．重要度 ξ がパターン間の順序 $\preceq$ に関して $x \preceq y \Rightarrow \xi(x) \geq \xi(y)$ を満たすとき，ξ は $\preceq$ に関して逆単調 (anti-monotonic) という [24]．最もよく用いられる重要度 ξ は頻度 (frequency)η であり，以下のように定義される．

$$\eta(x) = \frac{1}{|D|} \sum_{s \succeq x} \mathbf{1}_D(s)$$

また，$\eta'(x) = |D|\eta(x) = \sum_{s \succeq x} \mathbf{1}_D(s)$ をサポート (support) と呼ぶ．定義から明らかなように，$\xi = \eta$ のときは必ず逆単調である．パターン x の上方集合 $\uparrow x$ とデータ D の交わり $D \cap \uparrow x$，すなわちデータ D の中で x より精密なパターン $s \succeq x$ は，支持集合 (supporting set) と呼ばれ，パターン x のサポート $\eta'(x) = |D \cap \uparrow x|$ である．

パターンの順序関係 $\preceq$ と逆単調な ξ を利用することで，初めてパターンマイニングが可能となる．具体的なアルゴリズムは，S 中の最も小さい要素からスタートして，S の順序 $\preceq$ に従ってパターンを列挙していき，あるパターン x に対してもし $\xi(x) < \sigma$ ならば，それ以降の要素 $s \succeq x$ はチェックしない（アルゴリズム 1）．ここで，$x \lessdot s$ は s による x の被覆 (cover) を表し，$(x \prec s$ かつ $x \preceq y \prec s) \Rightarrow x = y$ を満たす．この戦略は，初めて逆単調性を利用してパターンマイニングを達成したアルゴリズム Apriori [25] にちなんで，Apriori 原理としても知られている．

パターンマイニングの代表例を以下に挙げる．

- **アイテム集合マイニング** [2, 3]：アイテムの有限集合 V に対して，$S = 2^V$ または $S = \{0, 1\}^{|V|}$．各パターン $x \in S$ はアイテム集合 (itemset) と呼ばれ，順序は $x \subseteq y$ ならば $x \preceq y$ として定まる．
- **部分グラフマイニング** [4, 5]：S はグラフ全体からなる集合．通常は連結グラフのみを対象とし，実際

アルゴリズム **1**　Apriori アルゴリズム

```
PatternMining(σ)
    PatternEnumeration(⊥, σ)
PatternEnumeration(x, σ)
    for each s ≻ x
        if ξ(s) ≥ σ
            s を出力する
            PatternEnumeration(s, σ)
```

の問題に応じて，ノードやエッジにラベルがあるかないかなどの種類がある．データとして与えられたグラフの集合 $D \subseteq S$ の部分グラフ発見が目的のため，各パターン $x \in S$ は部分グラフ (subgraph) と呼ばる．順序は x が y の部分グラフなら $x \preceq y$ として定まる．

- **系列パターンマイニング** [6, 7]：S はアルファベット Σ 上の文字列全体の集合．すなわち $S = \{a_1 a_2 \dots a_k \mid a_i \in \Sigma\}$ で多くの場合 $\Sigma = 2^V$．各 $x \in S$ は系列パターン (sequence) と呼ばれ，x が y の部分列，すなわち $y = vxw \in S$ となる $v, w \in S$ があるとき $x \preceq y$ と定まる．

部分グラフマイニングや系列パターンマイニングでは S の要素数が無限大となりうるが，たとえば $\xi = \eta$ のときは $\eta(x) > 0$ を満たすパターン x は高々有限個しかないため，アルゴリズム 1 で問題なくパターンマイニングが達成できる点に注意されたい．

3. 統計的有意パターンマイニング

データ $D \subseteq S$ があらかじめ二つのクラスに分割されている，すなわち，クラス $C \subseteq D$ と $\bar{C} = D \setminus C$ があると仮定する．以下では，常に $|C| \leq |\bar{C}|$ を仮定する．これは，機械学習の教師あり学習に対応する問題設定であり，コントラストパターンマイニング (contrast pattern mining) と呼ばれている．また，出現パターンマイニング (emerging pattern mining) や識別パターンマイニング (discriminative pattern mining) とも呼ばれている [26]．コントラストパターンマイニングの目的は，クラス C または $\bar{C}$ を特徴づけるパターンを見つける，すなわち，特定のクラスのみで重要度が高いパターンを見つけることである．各クラスにおける頻度 $\eta(x)$ の差や比など，さまざまな重要度には，各クラスにおける尺度が用いられてきた．

本稿では，頻度 $\eta(x)$ に関する分割表 (contingency table) を利用してパターンの重要度を測ることで，偽陽性 (false positive) を制御する手法を紹介する．偽陽性とは，クラス分類と関連があると判断したが実際には関連がないパターンのことであり，偽陽性の制御は科学的発見の正当性を統計的に担保するための基本的戦略として用いられている．パターンが偽陽性となりうる確率は，統計的仮説検定 (statistical hypothesis test) によって p 値として定量化でき，この p 値が十分小さく偽陽性となる確率があらかじめ定めた水準（有意水準）α よりも小さいと保証できるとき，そのパターンは統計的に有意 (statistically significant) という．p 値を用いて偽陽性を制御し，統計的に有意なパターンをすべて列挙するコントラストパターンマイニング問題を，特に統計的有意パターンマイニング (significant pattern mining) と呼ぶ [13, 14].

3.1 仮説検定

あるパターン x に着目したとき，x とクラス分類との関連は，x によって定まる上方集合 $\uparrow x$ とその補集合 $S \setminus \uparrow x$ への S の分割と，クラスによる C または $\bar{C}$ への D の分割，という二つの分割から定まる以下の分割表で表現される．

	$\succeq x$	$\not\succeq x$	S
C	$\lvert C \cap \uparrow x\rvert$	$\lvert C \setminus \uparrow x\rvert$	$\lvert C\rvert$
$\bar{C}$	$\lvert \bar{C} \cap \uparrow x\rvert$	$\lvert \bar{C} \setminus \uparrow x\rvert$	$\lvert \bar{C}\rvert$
D	$\lvert D \cap \uparrow x\rvert$	$\lvert D \setminus \uparrow x\rvert$	$\lvert D\rvert$

ここで，$|D \cap \uparrow x| = \eta'(x)$ である．この分割表はフィッシャーの正確確率検定 (Fisher's exact test) によって検定でき，計算された p 値が α 以下のとき，偽陽性の割合が α 以下であると保証できる．

分割表の周辺合計値，すなわち $\eta'(x), |C|, |D|$ を固定し，$k = |C \cap \uparrow x|$ とする．このとき，この分割表を観測する確率 $q(k)$ は，二項係数を用いて

$$q(k) = \binom{|C|}{k}\binom{|\bar{C}|}{\eta'(x)-k} \Big/ \binom{|D|}{\eta'(x)}$$

と計算でき，$q(k)$ は超幾何分布と呼ばれる離散確率分布となる．したがって，k が与えられたときの左側確率 P_{L} と右側確率 P_{R} は，

$$P_{\mathrm{L}} = \sum_{X=k_{\min}}^{k} q(X), \quad P_{\mathrm{R}} = \sum_{X=k}^{k_{\max}} q(X),$$
$$k_{\min} = \max\{0, \eta'(x) - |\bar{C}|\},$$
$$k_{\max} = \min\{\eta'(x), |C|\}$$

である．片側検定の場合は，その方向に応じてこれらの値のどちらかが p 値となり，両側検定の場合は，分布が左右対称でないため，パターン x の p 値は

$$\mathrm{pval}(x) = 2\min\{P_{\mathrm{L}}, P_{\mathrm{R}}\}$$

として定まる．

3.2 多重検定

単一のパターン x の検定では，p 値 $\leq \alpha$ であれば偽陽性となる確率が α 以下であることが保証できる．ところが，もしこの検定をすべてのパターンに対して繰り返し適用し，p 値 $\leq \alpha$ を満たすパターンをすべて出力すると，その中の $|S|\alpha$ 個のパターンが偽陽性になりうることを意味する．ただし，ここでは簡単のため各パターンは独立と仮定している．パターンマイニングでは S が巨大なため，$|S|\alpha$ も巨大であり，大量のパターンが偽陽性になってしまう．さらに，部分グラフマイニングなどでは，全パターン数 $|S|$ が無限大となる場合があり，その場合は偽陽性の個数が無限大となりうることを許容してしまっている．

この問題を解決し，全パターン S にわたって偽陽性が生じる確率を適切に制御するために，多重検定補正 (multiple testing correction) が必要となる．具体的には，α をより小さい値 δ に設定することで，少なくとも一つのパターンが偽陽性である確率 FWER (family-wise error rate) を制御する．p 値のしきい値を $\delta \leq \alpha$ に設定したときの FWER を $\mathrm{FWER}(\delta)$ と書くとき，$\mathrm{FWER}(\delta_{\mathrm{opt}}) = \alpha$ を満たす δ_{opt} を求めればよいが，この問題は解析的に解くことができない．$\mathrm{FWER}(\delta)$ は δ に関して単調増加なので，$\mathrm{FWER}(\delta) < \alpha$ が保証できるなるべく大きい δ をできるだけ簡単に探すことが，多重検定補正の目的となる．

最もよく用いられている多重検定補正は Bonferroni 補正 [27] であり，$\delta_{\mathrm{Bon}} = \alpha/|S|$ と設定すると，必ず $\mathrm{FWER}(\delta_{\mathrm{Bon}}) < \alpha$ となることが知られている．しかし，パターンマイニングでは $|S|$ が巨大なため，$\delta_{\mathrm{Bon}} \ll \delta_{\mathrm{opt}}$ となってしまい，$\mathrm{FWER}(\delta_{\mathrm{Bon}})$ が小さくなりすぎる．さらに，S が有限集合でない場合は，Bonferroni 補正の適用が本質的に不可能である．

3.3 検定可能性

パターンマイニングで多重検定を達成するための鍵となるのが，Tarone [28] によって導入された仮説の検定可能性 (testability) で，検定可能でない仮説は取り除いてしまっても FWER が変化しない，という性質である．この手法は，統計的有意パターンマイニングにおいて欠かせない手法である．それ以前にも統計的有意性に着目したパターンマイニング手法は存在していたが [29–31]，Terada et al. [12] が Tarone 法を再発見し，パターンマイニングと融合したことによって，厳格な FWER の制御が可能となった．

パターン x に対して，分割表の周辺合計値が固定されたとき，すなわち，頻度 $\eta(x)$ と各クラスおよびデータのサイズ $|C|, |\bar{C}|, |D|$ が固定されたときの，パターン x が取りうる p 値の下限値を $\psi(x)$ と表記する．そして，全パターン S を，この p 値の下限値 $\psi(x)$ に従って昇順に並べたときのパターン列を $x_1, x_2, x_3, \ldots$ とする．すなわち，

$$\psi(x_1) \leq \psi(x_2) \leq \psi(x_3) \leq \ldots \tag{1}$$

かつ $S = \{x_1, x_2, \ldots\}$ となる．このとき，ちょうど

$$l \cdot \psi(x_l) < \alpha \text{ かつ } (l+1) \cdot \psi(x_{l+1}) \geq \alpha \tag{2}$$

を満たす l 番目のパターン x_l に対して，パターン $x_1, x_2, \ldots, x_l$ を検定可能 (testable) と呼ぶ．Tarone は，これら検定可能なパターン（仮説）をしきい値

$$\delta_{\mathrm{Tarone}} = \frac{\alpha}{l}$$

で検定すればよく，残りの検定可能でないパターンは無視してしまっても $\mathrm{FWER}(\delta_{\mathrm{Tarone}}) < \alpha$ が保証できることを示した．実際に，これら検定可能なパターンに対する FWER は，

$$\mathrm{FWER}(\delta_{\mathrm{Tarone}}) \leq \sum_{i=1}^{l} \psi(x_i) \leq l \cdot \psi(x_l) < \alpha$$

なので，$\mathrm{FWER}(\delta_{\mathrm{Tarone}}) < \alpha$ となっていることが確認できる．また，定義から必ず以下が成り立つ．

$$\delta_{\mathrm{Bon}} \leq \delta_{\mathrm{Tarone}} \leq \delta_{\mathrm{opt}}.$$

分割表を用いたフィッシャーの正確確率検定においては，最も極端な状況，すなわち $k = |C \cap \uparrow x|$ が $k_{\min}$ または $k_{\max}$ のときの p 値が $\psi(x)$ と一致するので，$\eta'(x)$ の値に応じて以下の 4 通りに分けて計算することができる．

$$\psi(x) = \begin{cases} \dbinom{|C|}{\eta'(x)} \Big/ \dbinom{|D|}{\eta'(x)} & \text{if } 0 \leq \eta'(x) \leq |C|, \\ \dbinom{|\bar{C}|}{\eta'(x) - |C|} \Big/ \dbinom{|D|}{\eta'(x)} & \text{if } |C| < \eta'(x) \leq \dfrac{|D|}{2}, \\ \dbinom{|\bar{C}|}{\eta'(x)} \Big/ \dbinom{|D|}{\eta'(x)} & \text{if } \dfrac{|D|}{2} < \eta'(x) \leq |\bar{C}|, \\ \dbinom{|C|}{\eta'(x) - |\bar{C}|} \Big/ \dbinom{|D|}{\eta'(x)} & \text{if } |\bar{C}| < \eta'(x) \leq |D|. \end{cases}$$

図 2 に，$|D| = 100, |C| = 20$ のときの $\psi(x)$ を示す．図からもわかるように，$\psi(x)$ は $k = |D|/2$ に関して

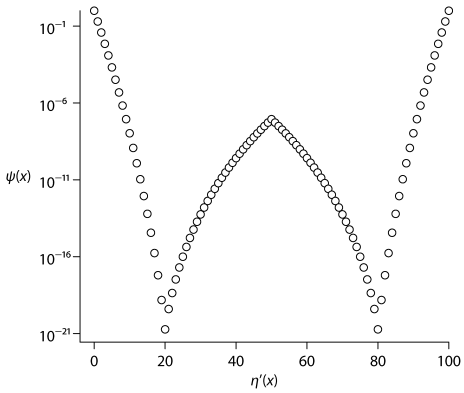

図 2　p 値の下限 $\psi(x)$

対称であり，$k > |D|/2$ に対して $\psi(k) = \psi(|D| - k)$ である．

ここでの目的は，$\psi(x)$ が小さい順にパターンを取り出すことなので，パターンマイニングのアルゴリズムを活用するために $\psi(x)$ の下限値を用いて順序準同型写像を設計する．具体的には，以下のように $\psi(x)$ の代理関数 $\psi'(x)$ を定める．

$$\psi'(x) = \begin{cases} \binom{|C|}{\eta'(x)} \Big/ \binom{|D|}{\eta'(x)} & \text{if } 0 \le \eta'(x) \le |C|, \\ 1 \Big/ \binom{|D|}{\eta'(x)} & \text{otherwise.} \end{cases}$$

すると，ψ' はサポート $\eta'(x)$ に関して単調減少で，$\eta'(x)$ はパターン間の順序 $\preceq$ に関して逆単調なので，ψ' は順序準同型写像となる．あとは，アルゴリズム 1 を用いて $\psi'(x)$ の小さい順にパターンを列挙し，各パターンに対して本来の p 値の下限 $\psi(x)$ を計算することで，不等式 (1) を満たすパターン列を見つけることができる．これは，$\xi(x) = 1/\psi'(x)$ として，アルゴリズム 1 中のしきい値 σ を大きい順，もしくは小さい順に変化させていけば可能である．詳しくは文献 [12, 13, 32] を参照されたい．

最後に，不等式の条件 (2) を満たすような l 番目のパターンを見つけたら，パターンマイニングを中断し，各検定可能パターン $x_1, x_2, \ldots, x_l$ に対してフィッシャーの正確確率検定を用いて p 値を計算し，その p 値が $\delta_{\text{Tarone}} = \alpha/l$ より小さいパターンを出力すれば，完了となる．

3.4　ランダム置換法

Tarone 法では，各パターンが独立と仮定している．しかし，パターン x と y の間に順序関係があるときは，一般に独立でない．このため，Tarone 法で求まる δ_{Tarone} は，最適値 δ_{opt} よりも大幅に小さくなってしまい，FWER の制御が保守的になりすぎてしまう傾向がある．そこで以下では，Westfall and Young のランダム置換による多重検定補正 [33] を紹介する．この多重検定補正は，最初に Terada et al. の手法 FastWY [34] で統計的有意パターンマイニングに導入され，その後 Llinares-López et al. の手法 Westfall-Young light [14] で高速化かつ省メモリ化された．

ランダム置換法では，クラスのサイズ $|C|$ と $|\bar{C}|$ を保ったまま，データ中の各パターン $x \in D$ をどちらかのクラスにランダムに割り当てる．この操作を h 回繰り返すことで，h 個の新たなクラス $C_1, \bar{C}_1, C_2, \bar{C}_2, \ldots, C_h, \bar{C}_h$ を作成する．実際には，パターンそのものを複製する必要はなく，クラスのラベルのみを保持しておけばよい．そして，各置換 $i \in \{1, 2, \ldots, h\}$ に対して，全パターンの中で最小の p 値を計算する．これを $p^i_{\min}$ と書くと，h 回の置換で求まる FWER(δ) の推定値 $\widehat{\text{FWER}}(\delta)$ は，

$$\widehat{\text{FWER}}(\delta) = \frac{|\{i \mid p^i_{\min} \le \delta\}|}{h}$$

となる．したがって，$p^1_{\min}, p^2_{\min}, \ldots, p^h_{\min}$ を小さい順に並べたときの α 分位点を δ_{perm} とすれば，これが求めるしきい値である．実用的には，反復数 h を 1,000 から 10,000 くらいに設定すれば，FWER(δ_{perm}) はほぼ α と一致し [14]，以下の関係が成り立つ．

$$\delta_{\text{Bon}} \le \delta_{\text{Tarone}} \le \delta_{\text{perm}} \approx \delta_{\text{opt}}.$$

あとは，各置換 i における最小 p 値 $p^i_{\min}$ が計算できればよい．ここで，p 値の下限 $\psi(x)$ の代理関数 $\psi'(x)$ を用いると，パターン x, y に対して $x \preceq y$ のとき，

$$\begin{aligned} &\text{pval}(x) < \psi'(y) \Rightarrow \\ &\text{任意の } s \succeq y \text{ に対して } \text{pval}(x) < \text{pval}(s) \end{aligned}$$

の関係が成り立つ．したがって，パターンを順序 $\preceq$ にしたがって小さいものから順に調べていけば，最小値 $\min_{x \in S} \text{pval}(x)$ を求めることができる．

4.　情報幾何とのつながり

実は，パターン全体からなる空間である半順序集合 $(S, \preceq)$ において，パターン x の頻度 $\eta(x)$ は，この空間がもつ情報の一つの側面しか評価しておらず，対となるもう一つの尺度 θ が見過ごされている．ここでは，情報幾何的な観点からの解析 [17] を紹介し，このことを示す．

まず，データ $D \subseteq S$ に対して

$$p(x) = \frac{1}{|D|}\mathbf{1}_D(x)$$

と定めると，p は D 上の確率分布となる．すなわち，$0 < p(x) \leq 1$ かつ $\sum_{x\in D} p(x) = 1$ を満たし，頻度 $\eta(x)$ に対して

$$\eta(x) = \sum_{s\succeq x} p(s)$$

の関係がある．このとき，

$$\log p(x) = \sum_{s\preceq x} \theta(x) \tag{3}$$

を満たす関数 $\theta : D \to \mathbb{R}$ を導入する．ただし，常に最小元 $\perp \in D$ の存在を仮定する．これは，対数線形モデル (log-linear model) [35] の一般化となっており，$D = \{0,1\}^n$, $\boldsymbol{x} = (x^1, x^2, \ldots, x^n) \in D$ の場合は

$$\log p(\boldsymbol{x}) = \sum_i \theta^i x^i + \sum_{i<j} \theta^{ij} x^i x^j + \cdots + \theta^{1\ldots n} x^1 \ldots x^n - Z$$

と一致する（Z は分配関数）．

すると，簡単な式変形によって，p で定まる確率分布は必ず指数型分布族となり，θ は指数型分布族の自然パラメータ (natural parameter)，η は十分統計量 (sufficient statistics) と一致することが確認できる．すなわち，θ と η が双対の関係にあり，

$$\mathbb{E}\left[\frac{\partial}{\partial\theta(x)}\log p(x)\frac{\partial}{\partial\eta(y)}\log p(y)\right] = \delta_{xy}$$

となる．ただし，δ_{xy} はクロネッカーのデルタであり，$x = y$ ならば $\delta_{xy} = 1$，そうでなければ $\delta_{xy} = 0$ である．この事実は，任意のパターン空間において，Amari et al. [18–20] が情報幾何の立場から研究してきた双対平坦構造 (dually flat structure) が入っていることを意味しており，情報幾何で発展してきたさまざまな手法がパターンマイニングに適用できることを示唆している．たとえば，KL ダイバージェンスに関する拡張ピタゴラスの定理を用いることで，各パターンに対してエントロピーなどの情報量を分離し，取り出すことができる．

統計や機械学習において研究されてきたほかの確率モデルとも関連が深い．代表的なモデルは，深層学習の基本的なモデルとしても知られているボルツマンマシン (Boltzmann machine) である．$D = 2^V$ とし，各パターン $x \subseteq V$ に対して，$|x| > 2$ ならば $\theta(x) = 0$ とすると，数式 (3) の対数線形モデルはボルツマンマシンと一致する．

5. おわりに

本稿では，データマイニングの分野で主に研究が進んできたパターンマイニング技術について紹介し，その最先端のトピックとして統計的に有意なパターンのマイニング技術を紹介した．その中で，パターン間の順序関係 $\preceq$ と順序同型となるような重要度 ξ をいかに設計するかが鍵となることを示した．統計的有意パターンのマイニングは，現在も盛んに研究が行われており，最新の成果としては，データが複数のカテゴリに分類されている場合のマイニング手法 [36, 37] や，より大規模なデータでのパターンマイニングを可能とするための並列化手法 [38] などがある．

パターンマイニングは，重要度が大きいパターンを列挙するという一見単純な問題である．しかし，その背後には豊富な数理的構造が潜んでおり，その一端として，情報幾何とのつながりがある．本稿を通して，パターンマイニングの魅力が伝われば幸いである．

謝辞　本研究は JSPS 科研費 JP16754296 の助成を受けたものです．

参考文献

[1] C. C. Aggarwal and J. Han (eds.), *Frequent Pattern Mining*, Springer, 2014.
[2] R. Agrawal, T. Imieliński and A. Swami, “Mining association rules between sets of items in large databases,” In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pp. 207–216, 1993.
[3] J. Han, J. Pei and Y. Yin, “Mining frequent patterns without candidate generation,” In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 1–12, 2000.
[4] A. Inokuchi, T. Washio and H. Motoda, “An apriori-based algorithm for mining frequent substructures from graph data,” *Principles of Data Mining and Knowledge Discovery*, **1910**, pp. 13–23, 2000.
[5] X. Yan and J. Han, “gSpan: Graph-based substructure pattern mining,” In *Proceedings of 2002 IEEE International Conference on Data Mining*, pp. 721–724, 2002.
[6] R. Agrawal and R. Srikant, “Mining sequential patterns,” In *Proceedings of the 11th International Conference on Data Engineering*, pp. 3–14, 1995.
[7] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M.-C. Hsu, “PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth,” In *Proceedings of the 17th International Conference on Data Engineering*, pp. 215–224, 2001.
[8] T. Uno, Y. Uchida, T. Asai and H. Arimura, “LCM: An efficient algorithm for enumerating frequent closed item sets,” In *Proceedings of Workshop on Frequent Itemset Mining Implementations (FIMI03)*, 2003.

[9] T. Uno, M. Kiyomi and H. Arimura, "LCM ver.2: Efficient mining algorithms for frequent/closed/maximal itemsets," In *Proceedings of Workshop on Frequent Itemset Mining Implementations (FIMI04)*, 2004.
[10] T. Uno, T. Asai, Y. Uchida and H. Arimura, "An efficient algorithm for enumerating closed patterns in transaction databases," *Discovery Science*, **3245**, pp. 16–31, 2004.
[11] T. Uno, M. Kiyomi and H. Arimura, "LCM ver.3: Collaboration of array, bitmap and prefix tree for frequent itemset mining," In *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*, pp. 77–86, 2005.
[12] A. Terada, M. Okada-Hatakeyama, K. Tsuda and J. Sese, "Statistical significance of combinatorial regulations," *PNAS*, **110**, pp. 12996–13001, 2013.
[13] M. Sugiyama, F. Llinares-López, N. Kasenburg and K. M. Borgwardt, "Significant subgraph mining with multiple testing correction," In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pp. 37–45, 2015.
[14] F. Llinares-López, M. Sugiyama, L. Papaxanthos and K. M. Borgwardt, "Fast and memory-efficient significant pattern mining via permutation testing," In *Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 725–734, 2015.
[15] F. Llinares-López, D. G. Grimm, D. A. Bodenham, U. Gieraths, M. Sugiyama, B. Rowan and K. M. Borgwardt, "Genome-wide detection of intervals of genetic heterogeneity associated with complex traits," *Bioinformatics*, **31**, pp. i240–i249, 2015.
[16] A. Terada, R. Yamada, K. Tsuda and J. Sese, "LAMPLINK: Detection of statistically significant SNP combinations from GWAS data," *Bioinformatics*, **32**, pp. 3513–3515, 2016.
[17] M. Sugiyama, H. Nakahara and K. Tsuda, "Information decomposition on structured space," In *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 575–579, 2016.
[18] S. Amari, "Information geometry on hierarchy of probability distributions," *IEEE Transactions on Information Theory*, **47**, pp. 1701–1711, 2001.
[19] H. Nakahara and S. Amari, "Information-geometric measure for neural spikes," *Neural Computation*, **14**, pp. 2269–2316, 2002.
[20] S. Amari, *Information Geometry and its Applications*, Springer, 2016.
[21] B. A. Davey and H. A. Priestley, *Introduction to Lattices and Order*, 2nd edition, Cambridge University Press, 2002.
[22] G. Gierz, K. H. Hofmann, K. Keimel, J. D. Lawson, M. Mislove and D. S. Scott, *Continuous Lattices and Domains*, Cambridge University Press, 2003.
[23] P. D. Laird, *Learning from Good and Bad Data*, Springer, 1988.
[24] M. J. Zaki and W. Meira Jr., *Data Mining and Analysis*, Cambridge, 2016.
[25] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," In *Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 487–499, 1994.
[26] G. Dong and J. Bailey, *Contrast Data Mining: Concepts, Algorithms, and Applications*, Chapman & Hall/CRC, 2012.
[27] C. E. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilità," *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, pp. 3–62, 1936.
[28] R. E. Tarone, "A modified Bonferroni method for discrete data," *Biometrics*, **46**, pp. 515–522, 1990.
[29] G. I. Webb, "Discovering significant patterns," *Machine Learning*, **68**, pp. 1–33, 2007.
[30] X. Yan, H. Cheng, J. Han and P. S. Yu, "Mining significant graph patterns by leap search," In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pp. 433–444, 2008.
[31] W. Hämäläinen, "Kingfisher: An efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures," *Knowledge and Information Systems*, **32**, pp. 383–414, 2012.
[32] S. Minato, T. Uno, K. Tsuda, A. Terada and J. Sese, "A fast method of statistical assessment for combinatorial hypotheses based on frequent itemset enumeration," *Machine Learning and Knowledge Discovery in Databases*, **8725**, pp. 422–436, 2014.
[33] P. H. Westfall and S. S. Young, *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*, John Wiley & Sons, 1993.
[34] A. Terada, K. Tsuda and J. Sese, "Fast Westfall-Young permutation procedure for combinatorial regulation discovery," In *2013 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 153–158, 2013.
[35] A. Agresti, *Categorical Data Analysis*, 3rd edition, Wiley, 2012.
[36] L. Papaxanthos, F. Llinares-Lopez, D. Bodenham and K. M. Borgwardt, "Finding significant combinations of features in the presence of categorical covariates," *Advances in Neural Information Processing Systems*, **29**, pp. 2271–2279, 2016.
[37] A. Terada, D. duVerle and K. Tsuda, "Significant pattern mining with confounding variables," *Advances in Knowledge Discovery and Data Mining (PAKDD 2016)*, **9651**, pp. 277–289, 2016.
[38] K. Yoshizoe, A. Terada and K. Tsuda, "Redesigning pattern mining algorithms for supercomputers," arXiv:1510.07787, 2015.

