D_009

制約充足確率による非逆単調な制約に基づくパターンマイニング
Pattern mining with non anti-monotone constraints by constraint satisfaction probability

北原 洋一† 折原 良平† 櫻井 茂明†
Youichi Kitahara Ryohei Orihara Shigeaki Sakurai

1．まえがき

データマイニングの活躍が期待されている分野の一つとして医療分野がある．近年，医療分野においては EBM や EBH の推進にも見られるように，データに基づいた治療や健康指導に関心が高まっている．EBM に関するデータマイニングの適用事例としては，心不全の投薬内容と心機能検査成績に対する決定木解析[5]などがある．

さらに，医療費増大の問題を背景に，長年採取されてきた健康診断データから有用な知識を発見する試みもなされている．時系列に蓄積された医療データから，最終的に発症や治癒につながりやすいパターンを発見することができれば，効果的な疾病予防や改善指導に有効であると思われる．

パターンを効率的に発見するアルゴリズムとしては，Apriori や PrefixSpan がよく知られている[1][2]．これらのアルゴリズムは，制約が有する逆単調性を利用することで探索空間を削減させている．しかしながら，医療時系列データから特徴的なパターンを発見する際に用いる制約は非逆単調であることが多い．医療時系列データのマイニングを効率的に行うためには，非逆単調な制約を満たすパターンを効率的に発見する手法が望まれる．

本研究の目的は，逆単調な性質を満たさない制約に基づくパターンマイニングを効率的に行うことである．逆単調性の定義を拡張して制約充足度を導入し，これを利用することで効率的に制約を満たすパターンを発見する方法を提案する．２章では，医療時系列データにおける特徴的パターンについて述べる．３章では，制約に基づくパターンマイニングについて説明する．４章では，制約充足確率を利用したマイニング方法を示す．５章では，まとめをする．

2．医療時系列データにおける特徴的パターン

本研究で想定しているデータは，医療時系列データである．さらに限定すると，多数の被験者について長年にわたって蓄積された健康診断データを想定している．このデータは，問診アンケートの回答や検査結果を属性データとする時系列データと考えられる．

健康診断データから，最終要素に含まれる疾病リスクが大きくもしくは小さくなりやすいパターンを発見することができれば，疾病予防や治療に活用できる．このようなパターンは，オッズや平均値，χ二乗値などについての制約として表現することができる．本研究では，オッズを用いることにする．オッズは，発症したときとしなかったときの頻度の比であり，発症確率の研究などで用いられることが多い．任意の系列パターンを S とし，S の最終要素に付随する発症の有無に関する属性値が発症を意味していれば $S_{r=1}$，そうでなければ $S_{r=0}$ と表現する．また，S の支持度を $f(S)$ で表せば，オッズが閾値 t 以上もしくは t' 以下でなければならないという制約 $C_{O(t,t')}$ は

$$C_{O(t,t')}(S)=\left\{S\left|\frac{f(S_{r=1})}{f(S_{r=0})}\geq t \cup \frac{f(S_{r=1})}{f(S_{r=0})}\leq t'\right.\right\} \quad (1)$$

と表すことができる．本研究では，この制約(1)を満たすパターンを効率的に発見することを課題とした．

3．制約に基づくパターンマイニング

パターンマイニングでは，興味深い特徴をもつパターンを発見するのが目的である．特に，頻出パターンを効率的に発見するアルゴリズムの研究が進んだため，パターンの支持度と関連する問題を扱うことが多い．このようなアルゴリズムとしてよく知られたものとして，Apriori アルゴリズムがある．このアルゴリズムでは，発見されるパターンの支持度が特定閾値以上でなければならないという制約が満たす次のような性質を利用している．

系列パターン S について，$S' \supseteq S$ となる系列パターン S' を S の上位パターンと呼ぶ．なお，系列同士の含有関係では，順序が保持されているものとする．また，発見すべきパターン S が制約を満たしていないときを C_a と表現すると

$$\overline{C_a(S)} \Rightarrow \overline{C_a(S')} \quad (2)$$

と言う性質を満たす C_a は逆単調な制約と呼ばれる．逆単調な制約に基づくマイニングでは，S が制約 C_a を満たさないことが判明した時点で，S の全ての上位パターン S' の探索を中断することができる．そのため，探索空間を大幅に削減することができ，効率的なマイニングが可能になる．しかしながら，非逆単調な制約を満たすパターンを発見する場合には，このような極めて扱いやすい性質を有することは多くなく，処理時間が増大しやすい．

非逆単調な制約を満たすパターンを発見する問題では，解決方法は大きく三種類に分けられる．一つ目は，頻出パターン発見処理と制約充足パターン発見処理を二段階で別途行う方法である．二つ目は，制約を満たすパターンを完全に列挙する方法である．三つ目は，完全に列挙せずとも主要な制約充足パターンを効率的に発見する方法である．

二段階で行う方法がもっともシンプルでよく利用される．これは，支持度の高いパターンを発見した後に，制約を満たすパターンを再度抽出する方法である．この方法には，支持度が低い制約充足パターンを発見するのが難しいという問題がある．

完全に列挙する手法としては，AprioriSMP がある．これは，制約を構成する関数の凸性を利用することで算出される，上位パターンが取り得る制約関数の上限値を用いることで，逆単調な制約と同様の効率的な処理を可能

†（株）東芝研究開発センター

にさせている．しかしながら，扱う問題によっては上限値が非常に大きくなり，計算量が膨大になることが想定される．また，現実的な問題では，完全列挙せずとも効率的に主要なパターンのみを発見できればよいケースも多い．そこで，本研究では，三つ目の方法を採用した．

4. 制約充足確率を利用したパターンマイニング

まず，本研究にて扱う非逆単調な制約は，逆単調な性質をもつ変量によって構成されていることとする．支持度が閾値以上でなければならないというのは逆単調な制約であるから，支持度を構成要素に含む制約(1)は，この要請を満たしている．

S について確率 P_S で成立する条件を $\Rightarrow_{P_S}$ と表すとき

$$\overline{C_P(S)} \Rightarrow_{1-P_S} \overline{C_{O(t,t')}(S')} \quad (3)$$

となる前提条件 C_P を制約充足条件，P_S を制約充足確率と呼ぶことにする．制約充足条件 C_P と制約充足確率 P_S を導入することにより，逆単調性が確率 P_S で緩和されて成り立つ．このため，S が制約 C_P を満たさないことが判明した時点で，S の上位パターン S' の探索優先度を下げることができるため，効率的なマイニング処理が可能になると思われる．

非逆単調な制約を扱う場合，一般に制約充足確率はわからないが，制約に含まれる変量の逆単調性を利用することで推定可能である．パターン S が定まれば $f(S_{r=1})$，$f(S_{r=0})$ も定まるから，パターンと支持度の関係を，支持度の軸で構成される平面にて，図1のように表すことができる．また，支持度は制約(2)を満たすから，S が定まると上位パターン S' の取り得る範囲は，支持度軸と S から支持度軸に引いた垂線とで囲まれる範囲 R_f に限定される．さらに，制約(1)を満たす範囲は図1において R_C に限定される．したがって，S' は S が定まったときに取り得る範囲の状態を等確率に取ると仮定すると，S の制約充足確率 P_S は R_f と R_C の面積比となるから，制約充足確率に関する閾値 T_C を用いて P_S と C_P は

$$P_S = \begin{cases} 1 - \dfrac{(t-t')f(S_{r=0})}{2f(S_{r=1})} & \left(\dfrac{f(S_{r=1})}{f(S_{r=0})} \geq t \text{のとき}\right) \\ \dfrac{f(S_{r=1})}{2tf(S_{r=0})} + \dfrac{t'f(S_{r=0})}{2f(S_{r=1})} & \left(t > \dfrac{f(S_{r=1})}{f(S_{r=0})} > t' \text{のとき}\right) \\ 1 - \dfrac{(t-t')f(S_{r=1})}{2tt'f(S_{r=0})} & \left(t' \geq \dfrac{f(S_{r=1})}{f(S_{r=0})} \text{のとき}\right) \end{cases}$$

$$C_P = \{P_S \mid 1 - P_S <= T_C\}$$

となる．

次に，制約充足条件を利用したマイニングアルゴリズムについて説明する．図2に，PrefixSpan アルゴリズムのフレームワークにて利用される再帰関数の擬似コードを示す．これに加え，処理の初期においては，制約充足確率を大きく設定し，上位パターンが制約を満たしやすいパターンから処理を進めてパターンを発見する．特定の制約充足確率を満たす全ての上位パターンを処理し終えたら，徐々に制約充足確率を小さくしていくことで，制約を満たしやすいパターンの上位パターンから順次処理を行うことができる．

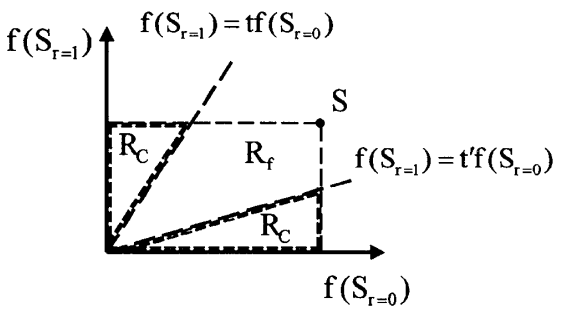

図1　パターンと支持度の関係

```
miner(prefix α, projected database D|α){
  find next items b by minimum scanning D|α;
  S ← α;
  append b to S;
  f0, f1 ← support of S0 and S1 by scanning D|α;
  if(C_O(t,t')(S)){output S;}
  calculate P_S;
  if(C_P(S)){call miner(S,D|S)} else{buffer B ← α'};
  get next prefix β;
  miner(β,ner(β, D|β);
}
```

図2　再帰関数の擬似コード

5. まとめおよび今後の課題

本稿では，上位パターンの制約の満たしやすさである制約充足確率を導入して緩和された探索打ち切り条件を利用し，制約が満たされやすいと想定されるパターンから探索を行う手法を提案した．制約充足確率は，制約に含まれる変量と制約関数から構成される空間において，上位パターンが取り得る範囲から算出される．なお，提案手法の有効性は実験にて確認する予定である．

今後の課題として，より簡便で適切な制約充足確率の算出法の開発などが挙げられる．本稿にて提案した手法を平均値やχ二乗値などに適用することもできるが，制約定義が複雑になると確率算出のオーバーヘッドが大きくなるという問題がある．より適用範囲の広い手法とするためには，制約充足確率算出方法の改善が必要である．

6. 参考文献

[1] R. Agrawal et.al., Mining Sequential Patterns, Proc. Of the 11th Int. Conf. Data Engineering, pp.3-14, 1995

[2] J. Pei et. al. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth, Proc. Of the 12th IEEE Int. Conf. On Data Engineering, 2001

[3] R. Ng et.al., Exploratory Mining and Pruning Optimizations of Constrained Association Rules, Proc. Of the 1998 SIGMOD Conf., 1998

[4] Morishita et.al., Traversing Lattice Itemset with Statistical Metric Pruning, Proc. Of PODS'00, 2000

[5] 金智隆ら，科学的根拠に基づく医療(EBM:Evidence-Based Medicine)におけるデータマイニングの適用事例と今後の展望，課題について，人工知能学会誌，19 巻，6 号，pp710-711, 2004