Vol.2010-AL-128 No.6
2010/1/26

論文

集合基底問題の正規基底を求めるヒューリスティックアルゴリズム

A Heuristic Algorithm for the Normal Set Basis Problem

概要

データマイニングの手法の一つにアソシエーション分析がある．アソシエーション分析とは，同時に現れる事象を測定データから抽出し，アソシエーション・ルールを見いだす手法である．アソシエーション分析は，販売データから顧客の購買に関する傾向を見いだしたり，文書をトピックに分類する等の応用に用いられる．これら販売データや文書などの測定データを，できるだけ少ない個数のアソシエーション・ルールの集まりとして表す問題は，集合基底問題 (Set Basis Problem: SBP) として定式化される．本論文では，アソシエーション・ルールに対して正規条件を加えた問題を正規集合基底問題 (Normal Set Basis Problem: NSBP) として定式化し，正規基底を求めるヒューリスティックアルゴリズムを提案する．

Abstract

In data mining, association analysis is a popular method that finds association rules describing events occurring simultaneously in observation data. Association analysis is used for finding tendencies in purchase from transaction data, categorizing text data by topics, or etc. To decide the minimum size of set of association rules in observation data (transactions, texts, or etc) is formalized as the Set Basis Problem (SBP). We formalize the problem of deciding the minimum size of set of association rules with the normalized condition as the Normal Set Basis Problem (NSBP), and we present a heuristic algorithm for finding normal basis of a family of sets.

1. はじめに

アソシエーション分析とは，同時に現れる事象をデータから抽出し，アソシエーション・ルールを見いだす手法である[1)5)]．アソシエーション分析は，販売データから顧客の購買に関する傾向を見いだしたり，テキストデータをトピックに分類する等の応用に用いられる．

例えば，POS システムの取引データを調べ,「パン」を購入している取引のうち，80% が「ミルク」も購入している場合，アソシエーション・ルールとして,「パン」⇒「ミルク」という関係があると考える．このアソシエーション・ルールは確信度 0.8 を持ち,「パン」を条件部,「ミルク」を帰結部と呼ぶ．条件部と帰結部は，一般に複数の要素を持つこともある．アソシエーション・ルールは，店舗内レイアウトや棚割り，陳列の計画，商品仕入れなどを検討する際の参考にすることができる．

すべての測定データを，できるだけ少ない個数のアソシエーション・ルールの集まりとして表す問題は，集合基底問題 (Set Basis Problem: SBP)[8)] として定式化される．SBP とは以下のような問題である．

定義 1.1 $U = \{e_i | i = 1, \ldots, m\}$ を台集合とし，$C_j (j = 1, \ldots, n)$ と $B_l (l = 1, \ldots, k)$ を U の空でない部分集合とする．C_j の集合族を $\mathcal{C} = \{C_j | j = 1, \ldots, n\}$ とし B_l の集合族を $\mathcal{B} = \{B_l | l = 1, \ldots, k\}$ とする．それぞれの C_j が，B_l の幾つかの集合の和集合で表せるとき，集合族 $\mathcal{B}$ を $\mathcal{C}$ の集合基底と呼ぶ．

集合基底問題 (Set Basis Problem: SBP):

入力: 集合族 $\mathcal{C}$.

出力: 最小サイズの集合基底 $\mathcal{B}$.

次に集合基底問題を行列を用いて定式化する[5)]．集合族 $\mathcal{C}$ を次のように行列で表す．行列 $C = [c_{i,j}]$ をサイズ $m \times n$ のブール行列とする．集合 C_j が要素 e_i を含むとき $c_{i,j} = 1$ で，そうでないとき $c_{i,j} = 0$ とする．すると，集合基底問題は次のように表現できる．S を $m \times k$ ブール行列， B を $k \times n$ ブール行列とし，$C = S \circ B$ を満たすとする．ただし $\circ$ はブール演算に基づく行列積である．集合基底問題は、与えられた行列 C に対し k が最小のブール行列 B を見つける問題となる。SBP は NP 完全問題である[6)7)]．

2. 正規集合基底問題

集合族 $\mathcal{C}$ に対して、互いに共通要素を持たない $\mathcal{B}$ の要素で C_i を表す，という条件を加えた問題を正規集合基底問題(Normal Set Basis Problem: NSBP) と呼ぶ[3]．すなわち NSBP とは以下のような問題である．

定義 2.1 $U = \{e_i | i = 1, \ldots, m\}$ を台集合とし，$C_j (j = 1, \ldots, n)$ と $B_l (l = 1, \ldots, k)$ を U の空でない部分集合とする．C_j の集合族を $\mathcal{C} = \{C_j | j = 1, \ldots, n\}$ とし B_l の集合族を $\mathcal{B} = \{B_l | l = 1, \ldots, k\}$ とする．それぞれの C_j が，B_l の互いに共通要素を持たない幾つかの集合の和集合で表せるとき，集合族 $\mathcal{B}$ を $\mathcal{C}$ の正規集合基底と呼ぶ．

正規集合基底問題 (Normal Set Basis Problem: SBP):

入力: 集合族 $\mathcal{C}$.

出力: 最小サイズの正規集合基底 $\mathcal{B}$.

NSBP は 2 部グラフの完全 2 部グラフによる分割問題と等価であり，NP 完全問題である[3]．

SBP もしくは 2 部グラフの完全 2 部グラフによる被覆問題については多くの研究[2] があるが，NSBP についての研究は多くはない．本論文は，ブール行列を用い，多項式時間でその近似解を求めるアルゴリズムの提案を行う．

2.1 正規集合基底問題の行列表現の例

例として以下のような NSBP の問題例を考える．

$\mathcal{C} = \{C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8\}$,

$C_1 = \{e_3\}$, $C_2 = \{e_3, e_6\}$, $C_3 = \{e_6\}$, $C_4 = \{e_1, e_2, e_5\}$,

$C_5 = \{e_3\}$, $C_2 = \{e_3, e_6\}$, $C_7 = \{e_6\}$, $C_8 = \{e_2, e_4, e_5\}$.

上記の $\mathcal{C}$ を 8×6 ブール行列 C で表す．

$$C = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}.$$

ここで $(m, n, k) = (8, 6, 4)$ とし，8×4 ブール行列 S と 4×6 ブール行列 B を以下の様な行列とすると $C = SB$ となる．

$$S = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}.$$

行列 B の 4 つの行ベクトルは次の 4 つの集合を表している．

$$B_1 = \{e_3\}, B_2 = \{e_6\}, B_3 = \{e_1, e_2, e_5\}, B_4 = \{e_2, e_4, e_5\}.$$

つまり $\mathcal{C}$ の要素はそれぞれ

$$C_1 = B_1,\ C_2 = B_1 \cup B_2,\quad C_3 = B_2,\ C_4 = B_3,$$
$$C_5 = B_1,\ C_6 = B_1 \cup B_2,\quad C_7 = B_2,\ C_8 = B_4$$

と表すことができる．すなわち $\mathcal{B} = \{B_1, B_2, B_3, B_4\}$ が NSBP の解となる．

3. NSBP に対するアルゴリズム

τ を $(1, 0]$ の値を取るパラメータとする．$m \times n$ ブール行列 C に対して，$n \times n$ の相関行列 $A(\tau)$ を以下の様に構成する．ベクトル $C_{\cdot i}$ と $C_{\cdot i}$ に対して，確信度を $< C_{\cdot i}, C_{\cdot j} > / < C_{\cdot i}, C_{\cdot i} >$ と定義し，$\mathbf{c}(i \Rightarrow j)$ と表記する．ただし $< \cdot, \cdot >$ はベクトルの内積である．相関行列 $A(\tau)$ の要素 $A(\tau)_{ij}$ を，$\mathbf{c}(i \Rightarrow j) \geq \tau$ の場合 $A(\tau)_{ij} = 1$，$\mathbf{c}(i \Rightarrow j) < \tau$ の場合 $A(\tau)_{ij} = 0$ と定める．前の例の C の場合，

$\tau = 1.0$ とすると

$$A(\tau) = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

となる．

以下にアルゴリズムを説明する．まず $l = 1$，すなわち S が $m \times 1$ ， B が $1 \times n$ のブール行列の場合を考える．$A(\tau)$ の行ベクトル $A(\tau)_{1.}$ を $1 \times n$ 行列 B の行ベクトル $B_{1.}$ とする．$j = 1, \ldots, m$ に対し，$C_{1j} = 0$ かつ $B_{1j} = 1$ となる j が存在している場合は $S_{j1} = 0$ とし，それ以外の場合は，SB と C の一致する要素が最大になる様に S_{j1} の値を 0 または 1 に決定する． この手順で S，B は

$$S = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, B = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

となる．SB を計算し，C と一致している要素の数を $\mathrm{hit}(1)$ とする．すなわち $P \equiv C - SB$ とすると，$\mathrm{hit}(1)$ は P のゼロ要素の個数である．変数 max を用意し，$\max = \mathrm{hit}(1)$ と設定する．

次に $A(\tau)_{2.}$ を B の行ベクトル $B_{1.}$ とし，同様の手順で S，B を求める．SB を計算し，C と一致している要素の数 $\mathrm{hit}(2)$ と max を比較し，もし max が小さければ max の値を $\mathrm{hit}(2)$ で更新する．この手順を $A(\tau)_{m.}$ を $B_{1.}$ に代入するまで繰り返し，$j = 1, \ldots, m$ としたときの最大の $\mathrm{hit(j)}$ を与える $B_{1.}$ を決定する．

次に P の非ゼロの行ベクトル数 N_r と 非ゼロの列ベクトル数 N_c を求める．上の例では $N_r = 6$ ， $N_c = 5$ となる． $N_{\min} \equiv \min(N_r, N_c)$ とすると，P は

$m \times (1 + N_{\min})$ 行列 S' と $(1 + N_{\min}) \times n$ 行列 B' の積で表すことができる．ただし $S'_{\cdot 1}$ は S であり， $B'_{1\cdot}$ は B である．上の例では

$$S' = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}, B' = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

と求められる．この手順で $C = S'B'$ となる $n \times (1+N_{\min})$ 行列 S' と $(1+N_{\min}) \times m$ 行列 B' が求められた．この手順を $l = k$ となるまで繰り返す． $l = 1, \ldots, k$ に対して $\min(l + N(\tau, l)) \equiv k_{\min}$ を求めたとき，$l + N(\tau, l)$ が最小となるような $k_{\min} \times m$ 行列 B が，このアルゴリズムによる $\mathcal{C}$ の正規集合基底問題の解である．表1 にアルゴリズムを示す．

表1 アルゴリズム

入力: $m \times n$ ブール行列 C，$\tau \in (0, 1]$
出力: $l + N(\tau, l)$ を最小にするブール行列 S，B

1: **for** $i = 1, \ldots, m$ **do**// τ に関して相関行列 $A(\tau)$ を求める
2: **for** $j = 1, \ldots, n$ **do**
3: $c(i \Rightarrow j) \geq \tau$ のとき $A(\tau)_{ij} \leftarrow 1$， $c(i \Rightarrow j) < \tau$ のとき $A(\tau)_{ij} \leftarrow 0$
4: **for** $l = 1, \ldots, k$ **do** // $m \times l$ ブール行列 S，$l \times n$ ブール行列 B を求める．
B の行ベクトルは $A(\tau)$ の行ベクトルから選び
$P = C - SB$ のゼロ要素の個数を最大にするように B と S を決定する
5: P の非ゼロの行ベクトルの数と，非ゼロの列ベクトルのうち，小さい方を $N(\tau, l)$ とする
6: S，B に適切な行，列ベクトルを付け加え
$P = C - SB$ がゼロ行列になる様な S，B を出力する
7: $l + N(\tau, l)$ の最小値と，そのときの S，B を出力する

4. 実 験 結 果

行列 C において，0 を黒で，1 を白で表した図を示す (図 1，2，3)．行，列の相関が強い場合 (図 1)，行間の相関がある場合 (図 2)，さらに $(0,1)$ をランダムに与えた場合 (図 3) についての実験結果を以下に示す．ただし以下の実験では $\tau = 0.99$ としている．

行，列の相関が強い場合は，一般に少数の基底ベクトルで行列 C を分解することができる．図 1 の場合は，本アルゴリズムにより $k = 6$，すなわち 6 個の基底ベクトルによる分解を発見できる．行間の相関がある場合，図 2 の例では，本アルゴリズムにより $k = 24$ の分解を発見できる．$(0,1)$ をランダムに与えた場合，図 3 の例では $n = k = 48$ となり，本アルゴリズムでは自明な分解しか発見できない．

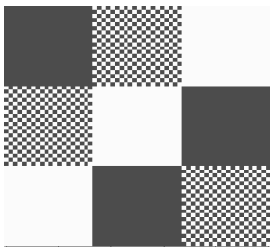

図 1　行，列の相関が強い場合． $((m,n) = (60,60),\ k = 6)$

Fig. 1　**The matrix with strong correlations between rows and columns.** $((m,n) = (60,60),\ k = 6)$

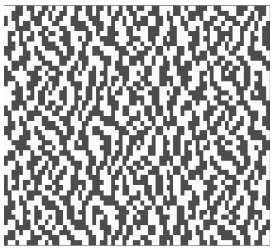

図 2　行間の相関がある場合． $((m,n) = (72,48),\ k = 24)$

Fig. 2　**The matrix with correlations between rows.** $((m,n) = (72,48),\ k = 24)$

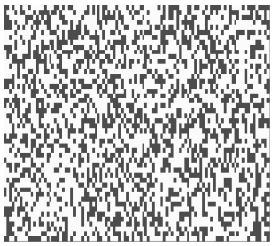

図 3 $(0,1)$ **をランダムに与えた場合．(**$(m,n)=(100,48)$**,** $k=48$**)**
Fig. 3 **A random binary matrix.(**$(m,n)=(100,48)$**,** $k=48$**)**

4.1 応 用 例

織機における綜絖枠枚数の最小化 織物の模様に対し，その織物を作成する織機の縦糸，横糸を制御する綜絖枠を考える．織物は 2 色で成り立っているとすると，模様はブール行列 C，縦糸はブール行列 S，横糸はブール行列 B に対応する．長目綜絖を導入した場合に，綜絖枠数 k を最小化する問題に対し，グラフ彩色アルゴリズムを利用した方法が提案されている[4]．本論文のアルゴリズムは，縦糸の通る綜絖が同時には上昇しないという条件のもとでの綜絖枠数の最小化問題に適用できる．図 1，図 2 の様な場合は，本アルゴリズムは綜絖枠数 k の減少に有効であることがわかる．

静止画像の可逆圧縮 静止画像におけるそれぞれの画素の色情報を，ブール行列 C の各行で表す．ブール行列 S と ブール行列 B によって $C=SB$ と分割すると，C が表している画像は， B の k 個の行の組み合わせによって表現されることになる．この場合，S の各行が画素に対応し，色情報は B の各行が表していると考えることができる．画素数 m の増加にくらべて k の増加が緩やかであれば，変換後のビットサイズは，元の画像のビットサイズに比べて小さくなる． 圧縮率を r とすると

$$r=\frac{mn}{k(m+n)}=\frac{n}{k}(\frac{m}{m+n})$$

である．この画像圧縮は，元の画像を可逆的に復元可能である．このアルゴリズムは，画像の局所的な性質ではなく，画像全体で相関を考えている．したがって，場所によらず似た色が多く使われている画像に有効である．

5. 結　　論

本アルゴリズムの計算量は $O(n^2m^2)$ であり，行列のサイズの多項式時間で実

行できる．アソシエーション分析では，ある条件部と帰結部が出現する割合を支持度と呼び，支持度もパラメータとして考慮する場合が一般的である．n が定数の場合は，支持度を考慮しても m の多項式時間で本アルゴリズムの実行が可能である．今後の課題として，このような状況が考えられる応用例等，実行時間を改善することを考えてみたい．

References

1) Agrawal, R., Imielinski, T. and Swami, A.: Database Mining: A Performance Perspective, *IEEE Transactions on Knowledge and Data Engineering*, Vol.5, No.6, pp.914–925 (1993).
2) Amilhastre, J., Janssen, P. and Vilarem, M.-C.: Computing a minimum biclique cover is polynomial for bipartite domino-free graphs, *SODA '97: Proceedings of the eighth annual ACM-SIAM symposium on Discrete algorithms*, Philadelphia, PA, USA, Society for Industrial and Applied Mathematics, pp.36–42 (1997).
3) Jiang, T. and Ravikumar, B.: Minimal NFA problems are hard, *SIAM J. Comput.*, Vol.22, No.6, pp.1117–1141 (1993).
4) Matsuura, I., Yagiura, M. and Hirata, T.: A Textile Design and the Boolean rank, *Proc. IADIS International Conference Applied Computing 2009*, pp.345–352 (2009).
5) Miettinen, P., Mielikäinen, T., Gionis, A., Das, G. and Mannila, H.: The Discrete Basis Problem, *IEEE Transactions on Knowledge and Data Engineering*, Vol.20, No.10, pp.1348–1362 (2008).
6) Müller, H.: On edge perfectness and classes of bipartite graphs, *Discrete Math.*, Vol.149, No.1-3, pp.159–187 (1996).
7) Orlin, J.: Contentment in graph theory: covering graphs with cliques., *Nederl. Akad. Wetenschappen, Proc. Ser. A 80, Indag, Math*, Vol.39, No.5, pp.406–424 (1977).
8) Stockmeyer, L.J.: The minimal set basis problem is NP-complete., *IBM Research Rep.RC 5431* (1975).