

論理値行列因子分解を用いた試験結果からのスキル構造の抽出

清野真理子†

木更津工業高等専門学校情報工学科†

大枝真一‡

木更津工業高等専門学校情報工学科‡

1. まえがき

試験は学生のスキルを測るために用いられる。設問に対して正解ならば1、不正解ならば0とすれば、試験結果を行列で表すことができる。

設問には必要なスキルが設定されている。設問と、設問を解くために必要となるスキルの関係を行列で表したものをQ-matrixという[1]。本研究では、試験結果の行列を因子分解することによって、Q-matrixを得るを試みる。

行列因子分解に用いられる手法として、NMF(Non-negative Matrix Factorization)[2]やBMF(Boolean Matrix Factorization)[3]がある。本研究では、BMFを用いて行列因子分解を行う。また、BMFによる行列因子分解では、パラメータによって分解精度が異なる。このパラメータを自動決定する手法として、MDL(Minimum Description Length)[4]を用いる。

2. BMF (Boolean Matrix Factorization)

2.1 BMF とは

NMFは非負値の行列に分解するが、行列が0,1の論理値の時は、論理値に対応した手法を用いることが良いと考えられる。BMFは、論理演算を用いて行列因子分解を行う手法である。NMFが非負値の実数値に行列因子分解するのに対し、BMFでは要素が論理値である行列に分解する。

図1に示すように、 $n \times m$ 行列 $R \in \{0,1\}$ を、 $n \times k$ 行列 $Q \in \{0,1\}$ と $k \times m$ 行列 $S \in \{0,1\}$ に分解する。なお、 $\circ$ は行列の論理積和演算子を表す。

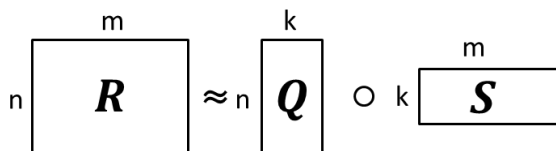


図1 BMF の概念図

2.2 BMF のアルゴリズム

R の各列同士の相関値から、相関行列 X を作成する。 i 番目の列を r_i とおくと、 X の要素である x_{ij} は、Association rule によって式(1)のように表すことができる。 τ は閾値であり、相関値が τ 以上ならば1を、 τ 未満なら0を x_{ij} に格納する。なお、 $\langle, \rangle$ は内積を表す。

$$x_{ij} = 1(\langle r_i, r_j \rangle / \langle r_i, r_i \rangle \geq \tau) \quad (1)$$

Q の列は貪欲法によって、式(2)に示す cover 関数の値が最大になるように X を元に作られる。また、 Q を作り出すと同時に S も作り出す。

$$\begin{aligned} \text{cover}(Q, S, R) = & |(i, j) : r_{ij} = 1 \wedge (Q \circ S)_{ij} = 1| \\ & - |(i, j) : r_{ij} = 0 \wedge (Q \circ S)_{ij} = 1| \end{aligned} \quad (2)$$

式(2)では、 R の要素と $Q \circ S$ の要素の違いを調べ、等しい要素の数から異なっている要素の数を引いている。2つの数の差が大きいほど、cover 関数の値は大きくなる。

図2に、BMF のアルゴリズムを示す。

```

Input : matrix  $R \in \{0,1\}^{n \times m}$  for data, a positive integer  $k$ , and a threshold value  $\tau \in (0,1]$ .
Output : Matrices  $Q \in \{0,1\}^{n \times k}$  and  $S \in \{0,1\}^{k \times m}$  such that  $Q \circ S$  approximates  $R$ .
1: function Asso( $R, k, \tau$ )
2: for  $i = 1, \dots, n$  do
3:    $x_i = j : r(i \Rightarrow j) \geq \tau$ 
4:  $Q \leftarrow \square, S \leftarrow \square$ 
5: for  $l = 1, \dots, k$  do
6:    $q_l \leftarrow x_i$  and  $s_{,l} \leftarrow \{0,1\}^m$  maximizing  $\text{cover}(Q, S, R)$ 
7: return  $Q$  and  $S$ 

```

図2 BMF のアルゴリズム

Extraction of the skill structure from an examination result using boolean matrix factorization

†Mariko Seino, National Institute of Technology, Kisarazu College

‡Shinichi Oeda, National Institute of Technology, Kisarazu College

3. MDL(Minimum Description Length)

MDL とは、モデルセット $\mathcal{H}$ が与えられたとき、

$$\text{MDL} = L(H) + L(D|H) \quad (3)$$

を最小にするモデル $H \in \mathcal{H}$ を選択する手法である。 $L(H)$ は H の記述長、 $L(D|H)$ は H を符号化した時のデータ記述長を示す。

BMF において、データセット R が与えられたとき、 R の分解 (Q, S) と R の異なる要素を

$$E = R \oplus (Q \circ S) \quad (4)$$

と表す。式 (4) を用いれば、 R は次のように定義することができる。

$$R = (Q \circ S) \oplus E \quad (5)$$

また、この様子の例を図 3 に示す。

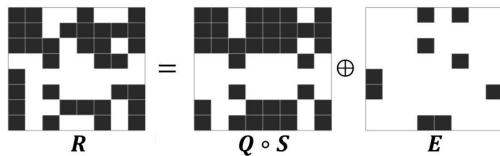


図 3 式 (5) の図

BMF において、 $H = (B, C)$ とし、MDL を用いて最小化する式を

$$L(R, H) = L(H) + L(E)$$

と定義する。この時、 B が $n \times k$ 行列、 C が $k \times m$ 行列であるとし、 $L(H)$ を式 (6) と定義する。

$$L(H) = L_{\mathbb{N}}(n) + L_{\mathbb{N}}(m) + L(k) + L(Q) + L(S) \quad (6)$$

4. 計算機実験

論理値行列 Q および S をそれぞれ 15×8 行列、 8×100 行列とし、要素をランダムに決定する。また、 k を MDL によって決定した。 $Q \circ S$ によって R を求め、 R に対して BMF を用いて行列因子分解を行った。

実験の結果、MDL を用いて $k = 8$ と求めることができた。行列因子分解を行ったところ、求められた $\hat{Q}$ と元の Q のフロベニウスノルムは 14 となった。

5. 実データによる実験

5.1 実験方法

実データとして、統計言語 R の CDM パッケージにある『fraction.subtraction.data』および『fraction.subtraction.qmatrix』を用いた。このデータは、

分数の問題 20 問を 536 人が解いた結果である。この試験における Q-matrix はエキスパートによってスキルは 8 つであると定義されているため、定義されている Q-matrix と推定による Q-matrix の比較を行う。

5.2 MDL を用いた実験

MDL を用いて k を決定したところ、 $k = 2$ となり、定義されているものよりもきわめて小さくなった。

k が小さくなってしまった原因として、8 つあるとされているスキルの中に、少ない問題にしか使われていないものがあるためということが考えられる。

5.3 考察

分数の問題に設定されているスキルは非常に細かくなっている。エキスパートから見ればそのように設定されているとしても、実際にはこれらの細分化スキルが正しくないかもしれない。本研究において、MDL によって $k = 2$ となったが、それは「分数の問題を解くことができるスキル」と「凡ミスをしないスキル」に分けることもできると考察する。

6. まとめ

MDL を用いて BMF による行列因子分解を行い、設問とスキルの関係を表す Q-matrix を求めた。実験の結果、計算機実験ではうまく MDL でスキル数を求めることができたが、実データではエキスパートの決定したスキル数よりも小さくなった。

しかし、エキスパートによって定義された Q-matrix が正しいとは限らないため、得られた Q-matrix に意味づけができるかどうか重要であると考えられる。

謝辞 本研究は JSPS 科研費 25750095 の助成を受けたものです。

参考文献

- [1] Shinichi Oeda, Yu Ito, Kenji Yamanishi, Extracting Latent Skills from Time Series of Asynchronous and Incomplete Examination, EDM2014.
- [2] Daniel Lee, H. Sebastian Seung, Algorithms for non-negative matrix factorization, in Adv.NIPS, 2000.
- [3] Pauli Miettinen, Taneli Mielikainen, Aristides Gionis and, Gauntan Das, Heikki Mannila, The Discrete Basis Problem, PKDD, pp.335–346, 2006.
- [4] Pauli Miettinen, Jiles Vreeken, Minimum Description Length for Boolean Matrix Factorization, MPI-I-2012-5-001.