7ZL-04

NMFのデータ前処理におけるIRMの有効性の検証

飯棲 俊介 †　大枝 真一 ‡

木更津工業高等専門学校 制御・情報システム工学専攻 †　木更津工業高等専門学校 情報工学科 ‡

1. まえがき

e-Learning システムを効果的に運用するためには学習者のレベルに応じた設問を提供するべく，設問に含まれる潜在スキルを特定し，その内どのスキルを学習者が修得しているかを把握することが重要となる．設問を解くために必要な潜在スキルは Q-Matrix により定義されるが，Q-Matrix の手動作成には有識者の知見が必要となるほか，学習者のスキル特定に口頭試問の実施が必要となり時間的コストがかかってしまう．そこで，先行研究では NMF による試験結果から Q-Matrix の自動抽出が考案されている [1]．その後，時系列から不変の Q-Matrix を得る Online NMF[2] といった NMF の改良手法が提案されている．

本研究では，NMF の精度向上をデータの前処理の観点から図る．そのための手法として，共クラスタリング手法である IRM を用いてデータの前処理を行い，NMF の精度向上に繋がるかどうかを人工データを用いた実験により検証する．

2.Q-Matrix

Q-Matrix は設問 (Items) と，それを解くために必要なスキル (Skills) の関係を表す行列である．Q-Matrix には，設問を解くために必要なスキル数が 1 つのみである Additive-Model と，複数必要とする Conjunctive-Model がある．図 1 に Q-Matrix の例を示す．図 1 の Conjunctive-Model に注目すると，設問 1 を解くためにはスキル 1 が必要，設問 2 を解くためにはスキル 1,2 が必要ということを意味している．Q-Matrix は学習者 (Examinees) と修得しているスキルの関係を表す S-Matrix とともに使用される．この Q-Matrix,S-Matrix から学習者が設問に正答できるかどうかを表す R-matrix が式 (1) によって算出される．

$$\neg R = Q \cdot (\neg S) \tag{1}$$

ただし，$\cdot$ は論理積，$\neg$ は論理否定を表す．Q,S から R を算出した例を図 2 に示す．図 2 において，R_{22}, R_{23} に着目すると，学習者 2 はスキル 1 を有していないため設問 2 を解くことができない ($R_{22} = 0$)，学習者 3 はスキル 1,2 を有しているため設問 2 を解くことができる ($R_{23} = 1$) という関係を表している．

3.NMF

非負値行列因子分解 (NMF:Non-Negative Matrix Factorization) とは，次元縮約手法の 1 つであり，非負

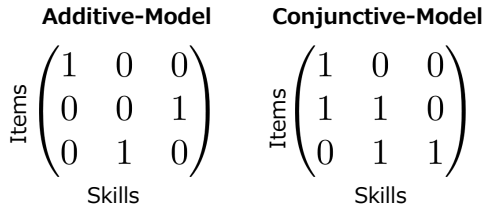

図 1: Q-Matrix

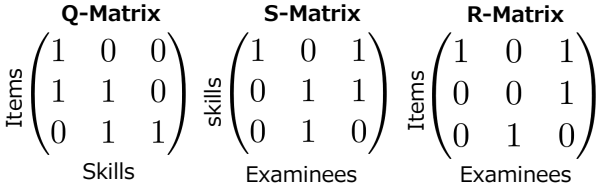

図 2: R-Matrix の算出

値行列を 2 つの非負値行列に因子分解する手法である．

非負値行列 $X(m \times n)$ が得られたとき，2 つの非負値行列 $U(m \times k), V(k \times n)$ に因子分解することを考える．分解誤差である目的関数を式 (2) と設定し，これを最小化することを考えると，U, V に初期値を与え，式 (3),(4) の更新式により U, V を近似計算することができる [3].

$$||X - UV||_F^2 \equiv \sum_i^m \sum_j^n \left(X_{ij} - \sum_l^k U_{il} V_{lj} \right)^2 \tag{2}$$

$$U'_{ij} \leftarrow U_{ij} \frac{(XV^T)_{ij}}{(UVV^T)_{ij}} \tag{3}$$

$$V'_{ij} \leftarrow V_{ij} \frac{(U^T X)_{ij}}{(U^T UV)_{ij}} \tag{4}$$

ここで $0 < k < \min(m, n)$ とし，分解の際に任意に決める次元数となる．

4.IRM

無限関係モデル (IRM:Infinite Relational Model) とは，異なるオブジェクト同士の関係データに対する確率モデルであり，関係の類似性から異種オブジェクトを同時にクラスタリング (共クラスタリング) する手法である [4]．関係データには例えば顧客同士が互いに知り合いであるか否か，顧客と購入した商品の関係などがある（ともに二値データ）．例えば図 3 左のデータを IRM により共クラスタリングを行い，各クラスタが隣接するように 2 つのオブジェクトをそれぞれソーティングすることで，図 3 右のように 1 もしくは 0 の要素が固まったデータ群を抽出することができる (太線はクラスタの境界線を表す)．

Verification of the Effectiveness of IRM in Data Preprocessing for NMF

†Shunsuke Iizumi, Advanced Course of Control and Information Engineering, National Institute of Technology, Kisarazu College

‡Shinichi Oeda, Department of Information and Computer Engineering, National Institute of Technology, Kisarazu College

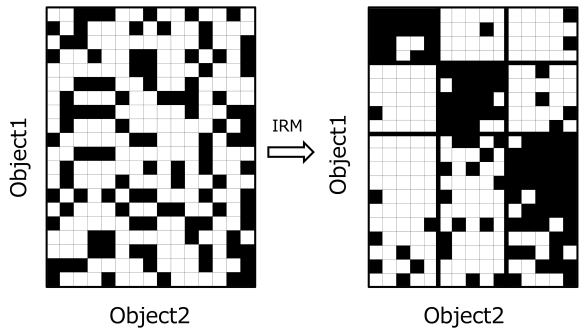

図 3: IRM によるクラスタリング

4.1.IRM の生成モデル

IRM では，行クラスタ i, 列クラスタ j に所属するオブジェクトは生起パラメータ θ_{ij} が与えられ，同じ生起確率分布から生起されるものとする．二値データの場合，関係データ R の k 行 l 列の要素 $R_{kl} \in \{0,1\}$ が行クラスタ i，列クラスタ j に所属するものとすると，生起確率 $P(R_{kl};\theta_{ij})$ はパラメータ $\theta_{ij}(0 \leq \theta_{ij} \leq 1)$ をもつ式 (5) のベルヌーイ分布に従うと仮定する．

$$P(R_{kl};\theta_{ij}) = \theta_{ij}^{R_{kl}}(1-\theta_{ij})^{1-R_{kl}} \tag{5}$$

これは確率 θ_{ij} で 1，確率 $1-\theta_{ij}$ で 0 が生起される確率分布である．また，このパラメータ θ_{ij} の生起確率分布（事前分布）$p(\theta_{ij})$ として，式 (6) のベータ分布により θ_{ij} が生成されるとする．

$$p(\theta_{ij}) = \frac{\theta_{ij}^{a-1}(1-\theta_{ij})^{b-1}}{\int_0^1 \theta_{ij}^{a-1}(1-\theta_{ij})^{b-1}d\theta_{ij}} \tag{6}$$

ここで，a, b はハイパーパラメータであり，$a, b = 1$ のとき，$p(\theta_{ij}) = 1$，すなわち一様分布となりパラメータ生成の偏りがないことを意味する．

4.2. クラスタリング方法

IRM によるクラスタリングでは，クラスタ数は未知のものとし，行クラスタ，列クラスタの所属クラスタの組 $\mathbf{s}^1, \mathbf{s}^2$ を予測するものとなる．しかし，行クラスタ数，列クラスタ数ともに未知であるため，全通りの事後確率を計算するには計算量が膨大となってしまう．そこで，ギブスサンプリング [5] の学習アルゴリズムを用いて逐次的に解を求める．

5. 提案手法

本研究では，NMF による Q-Matrix の抽出精度を上げるべく，データの前処理として IRM を用いたソーティングを行う．これにより学習者と解ける設問の関係を結びつけ，NMF の精度向上を図ることができると考える．

6. 計算機実験

人工データによる実験を行い提案手法の有効性の検証を行う．人工データとして，答えた学習者，設問，設問を解くために必要なスキル，正答できたか否かの項目からなるレコードデータを作成する．レコードデータの例を表 1 に示す．正答の可否は項目反応理論を用いて確率的に決定する．回答する学習者とその学習者が回答する設問はランダムに選択する．加えて実際の e-Learning システムを想定し，学習者による回答数の頻度，能力に応じた設問の選択を偏らせたレコードデータも作成する．作成されたレコードデータから最終的に学習者が設問に正解できたかどうかを表す関係行列 R，及び学習者が設問に答えた回数を表す行列 C を作成する．R の各要素は正答した回数が回答数の半分以上であれば正答したとみなし，値 1 とする．

表 1: レコードデータの例

No	User_id	Item_id	Skill	Result
1	15	13	A	1
2	8	11	B,C	0
3	11	1	A,D	0
4	20	11	B,C	1
5	2	5	A,D,E	1

得られた行列 C に対して IRM によりクラスタリング，ソーティングを行い，R を C のソーティング結果と同じ順序に並び替えた行列 R' を作成する．そして R 及び，R' を NMF で $R \approx QS$，$R' \approx Q'S'$ へと因子分解し，式 (2) により分解精度を比較する．R には 1 回も回答されず正答したのかが分からない欠損箇所が現れることが考えられるため，欠損値に対応させた NMF の手法である WeightedNMF[6] を用いる．分解誤差から IRM による共クラスタリングで誤差が小さくなるかどうかを検証する．

謝辞：本研究は JSPS 科研費 19H01728 の助成を受けたものです．

参考文献

[1] Michel C. Desmarais, “ Conditions for effectively deriving a Q-Matrix from data with Non-negative Matrix Factorization”, EDM2011, pp.41-50, 2011.

[2] 大枝真一, 天野恵理子, 山西健司, “ 行列因子分解を用いた時系列試験結果からの潜在スキル構造の抽出”, 信学技法:Technical report of IEICE, Vol.113, No.286, pp.123-130, 2013.

[3] 亀岡弘和, “ 非負値行列因子分解 ”, 計測と制御, 第 51 巻, 第 9 号, pp.835-844, 2012.

[4] 桑田修平, 山田武士, 上田修功, “ディレクレ過程混合モデルに基づく離散データの共クラスタリング”, 情報処理学会論文誌 数理モデル化と応用, Vol.1, No.1, pp.60-73, 2008.

[5] Radford M. Neal, “Markov Chain Sampling Methods for Dirichlet Process Mixture Models ”, Journal of Computational and Graphical Statistics, Vol.9, No.2, pp.249-265, 2000.

[6] 大野泰己, 大枝真一, “ 問題推薦のための行列因子分解を用いたスキル階層構造の可視化 ”, 情報処理学会第 80 回全国大会論文集, pp.883-884, 2018.