4F1-3in

異粒度データ解析のための非負値行列分解手法

Non-negative Matrix Factorization for Inconsistent Resolution Data Analysis

幸島 匡宏 Kohjima Masahiro　松林 達史 Matsubayashi Tatsushi　澤田 宏 Sawada Hiroshi

日本電信電話株式会社 NTT サービスエボリューション研究所
NTT Service Evolution Laboratories, NTT Corporation

Difficulty of a comprehensive data collection and consideration of privacy protection leads to an increase of the needs to analyze a set of inconsistent resolution data such as a pair of user's individual information and attribute-based statistical information simultaneously. In this paper, we focus on the problem of analyzing retail purchase log data which consist of the log of membership users and that of non-membership users. By carefully considering the relation between these logs, we propose a new method based on non-negative matrix factorization. We confirm its effectiveness using real purchase log dataset.

1. はじめに

近年のデータ分析では, 網羅的なデータ収集の困難さやプライバシー保護等の観点から, ユーザ個人単位の情報と属性単位の統計情報といった, 異なる粒度の情報が混在するデータを扱う機会が増えている. 代表例のひとつには, スーパーマーケットや衣料品店などの小売店のデータ解析が挙げられる. 多くの小売店では, マーケティング等における活用のために, (1) 利用顧客に対して会員カードを発行し, 会員顧客の個人単位の購買情報・属性情報を収集すること, (2) 会計時にレジ係が利用顧客の外見から性別, 年代を判定・入力することで, 顧客の属性単位の購買情報を収集すること, の 2 点が実施される. したがって, 小売店において収集されるデータから, 各個人・各属性単位の購買傾向を抽出するためには, 個人単位の購買情報である会員顧客購買履歴と属性単位の購買情報である非会員顧客購買履歴を組み合わせて解析する必要がある.

そこで本研究では, 会員顧客購買履歴と非会員顧客購買履歴を統合的に解析する手法を提案する. 提案手法は, 各購買履歴における同一属性での総購入数に関係性を明示的に導入し, 購買履歴を表現する 2 つの行列の低ランク表現を求める手法として定式化される. 提案手法の出力を用いて, 顧客の個人単位・属性単位の購買パターンの把握と, 顧客の個人単位のデータがスパースな場合であっても, 商品推薦等を行う際に有益な行列中の欠損値補完が可能となる.

提案手法のベースとなる非負値行列分解 (Non-negative Matrix Factorization, NMF) [2] 手法は, 教師なし学習手法の 1 つであり, 近年広くその有効性が確認されている. ユークリッド距離や板倉斉藤距離, 一般化カルバックライブラーダイバージェンス (一般化 KL 距離) 等を用いることで, 映画等のレーティング履歴, 時系列, 文書集合, 購買履歴など多様なデータに適用可能である. 様々な拡張手法も近年提案されており [3, 6], NMF の拡張は教師なし学習手法の研究・応用を考えるうえで非常に重要でもある. 本稿では, 実購買履歴データに対して提案手法を適用し, 上記既存手法を上回る性能が確認されたことを報告する.

連絡先: 幸島匡宏. 日本電信電話株式会社 NTT サービスエボリューション研究所, 〒 239-0847 神奈川県横須賀市光の丘 1-1. E-mail: kohjima.masahiro@lab.ntt.co.jp

2. 提案手法

はじめに入力データについて述べる. ユーザ (会員) 数, 商品数, ユーザ属性数をそれぞれ I, J, K, 会員購買履歴を行列 $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{I} = \{x_{ij}\}_{i,j=1}^{I,J}$ と書く. 行列の要素 x_{ij} がユーザ i の商品 j の購買数を表す. 同様にユーザ会員属性を行列 $\boldsymbol{V} = \{v_{ik}\}_{i,k=1}^{I,K}$ で表現し, ユーザ i が属性 k である時 v_{ik} の値は 1, そうでなければ 0 を取るとする. 非会員購買履歴は行列 $\boldsymbol{Y} = \{\boldsymbol{y}_k\}_{k=1}^{K} = \{y_{kj}\}_{k,j=1}^{K,J}$ と書く. 行列の要素 y_{kj} が属性 k の商品 j の購入数を表す. 提案手法は行列 $\boldsymbol{V}$ を用いて $\boldsymbol{X}$ と $\boldsymbol{Y}$ の低ランク表現を求める手法として定式化される. 行列 $\boldsymbol{X}$ の低ランク表現を, 因子行列 $\boldsymbol{A} = \{a_{ir}\}_{i,r=1}^{I,R}$, $\boldsymbol{B} = \{b_{jr}\}_{j,r=1}^{J,R}$ を用いて, $\hat{\boldsymbol{X}} = \boldsymbol{A}\boldsymbol{B}^T$ で定義する. R でランク数を表す. 本稿では低ランク表現をもとめる際の損失関数 $\mathcal{D}$ として, 購買履歴等の離散値をとるデータを解析する際に用いられることの多い一般化 KL 距離を用いる.

$$\mathcal{D}(\boldsymbol{X}|\hat{\boldsymbol{X}};\boldsymbol{A},\boldsymbol{B}) = \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}\log\frac{x_{ij}}{\hat{x}_{ij}} - x_{ij} + \hat{x}_{ij}. \quad (1)$$

なお, 式 (1) を因子行列の非負値制約 ($\boldsymbol{A}, \boldsymbol{B} \geq 0$) のもと最小化する手法が NMF である.

我々の提案手法は行列 $\boldsymbol{X}$ と $\boldsymbol{Y}$ の間に導入する関係から自然に導かれる. その関係とは, 属性 k の会員顧客購買数 と 属性 k の非会員購買数 が 比例する というものである. 会員購買履歴 $\boldsymbol{X}$ と属性情報 $\boldsymbol{V}$ が利用可能であるから, 属性 k の会員顧客による商品 j の総購買数は積 $\boldsymbol{V}^T\boldsymbol{X}$ の要素として表現できる. したがって導入する比例関係は $\boldsymbol{y}_k \propto \sum_{i=1}^{I} v_{ik}\boldsymbol{x}_i$ と書くことができる. もし会員数が十分に多ければ, この関係は "ほぼ" 満たされると想定することは自然である. それゆえ我々はこの比例関係を, 低ランク行列 $\hat{\boldsymbol{X}}$ と $\hat{\boldsymbol{Y}}$ が満たすように因子分解することを考える. 対角行列 $\boldsymbol{C} := diag(\{c_k\}_{k=1}^{K})$ をその要素 c_k が属性 k の比例定数を表すものとして定義する. すると満たすべき比例関係は次の等式で表現される.

$$\hat{\boldsymbol{Y}} = \boldsymbol{C}\boldsymbol{V}^T\hat{\boldsymbol{X}}. \quad (2)$$

したがって NMF と同じく $\hat{\boldsymbol{X}} = \boldsymbol{A}\boldsymbol{B}^T$ という因子分解を考え, 式 (2) に代入すると,

$$\hat{\boldsymbol{Y}} = \boldsymbol{C}\boldsymbol{V}^T\boldsymbol{A}\boldsymbol{B}^T \quad (3)$$

が導かれる. したがって提案手法は次のように定式化できる.

$$\arg\min_{\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}} \left\{ \mathcal{D}(\boldsymbol{X}|\boldsymbol{A}\boldsymbol{B}^T) + \alpha\mathcal{D}(\boldsymbol{Y}|\boldsymbol{C}\boldsymbol{V}^T\boldsymbol{A}\boldsymbol{B}^T) \right\} \quad (4)$$
$$s.t. \quad \boldsymbol{A},\boldsymbol{B},\boldsymbol{C} \geq 0, \quad \boldsymbol{C} = diag(\{c_k\}).$$

ただし $\alpha(\geq 0)$ は行列 $\boldsymbol{Y}$ へのフィッティングの重要度を表す重みパラメタである. 上記の最適化問題を解くアルゴリズムには勾配法などが利用可能である.

3. 関連研究

本研究の問題設定は collective matrix factorization (CMF) と呼ばれる因子行列を共有させながら複数の行列を同時に分解する手法 (e.g. [3, 5, 6]) と類似している. しかし, CMF で解析対象となる行列の組は, ユーザのアーティストに対する楽曲視聴履歴とアーティストのタグ情報の組 [6] や絵本中の単語情報とその対象年齢情報 [7] など異なる種類のデータを表す行列の組であることが多い. それに対し, 本研究の解析対象となる行列 $\boldsymbol{X},\boldsymbol{Y}$ はともに購買情報であり, これらの行列同士の関係性は CMF よりも考慮される必要がある. 我々の提案手法は行列間に比例関係を導入することで, 行列同士の関係性を因子分解形に表現することを可能としている.

4. 実験

4.1 実験設定

実験には, 株式会社インテージの「SCI」データを利用する. このデータには 2013 年 1 月 1 日から 2013 年 12 月 31 日までの 1 年間における, ユーザのスーパーマーケット, コンビニエンスストア等での購買履歴情報とデモグラフィック属性情報が含まれる. データ中に含まれている全ての購買履歴は, ユーザを識別可能なユーザ ID が含まれる. したがって本実験では, データ中のユーザからランダムに抽出した 10%, 20%, 30%, 40%, 50%を会員, 残りを非会員として取り扱う. すなわち会員の購買履歴より行列 $\boldsymbol{X}$ を作成し, 非会員の購買履歴より, ユーザ ID 情報を利用せず, 年齢, 性別に関するユーザの属性情報 (男性 20 代, 女性 30 代等) を用いることで, 行列 $\boldsymbol{Y}$ を作成する. 解析対象とするユーザ・商品はそれぞれ 50 回以上, 30 回以上購買履歴のあるユーザ・商品に限定した. これにより, (会員と非会員分離前の) 全体ユーザ数は約 5000 人, 商品数 J, 属性数 K はそれぞれ約 7000 商品と 12 属性となった.

評価尺度にはテストデータに対する平均絶対誤差 (Mean Absolute Error, MAE) を用いる. 具体的には, 行列 $\boldsymbol{X}$ からランダムに抽出された非ゼロ要素の 5%をテストデータとして学習データから取り除き, MAE を次のように計算する.

$$MAE = \frac{1}{|\mathcal{T}|} \sum_{(i,j)\in\mathcal{T}} |x_{ij} - \hat{x}_{ij}|.$$

ただし, $\mathcal{T}$ でテストデータに含まれる行列要素のインデックス全体, $|\cdot|$ で集合中の要素数を表す. 比較手法には NMF [2] と NMF に基づく CMF 手法である NMMF [6] を用いる. 因子行列のサイズ R と NMMF と提案手法の重みパラメタ α は $R = 5, 10, 15, 20, 25, 30$, $\alpha = 0.1, 0.5, 1.0$ と設定し, 作成した学習データとテストデータの組 10 通りに対して MAE を算出する. 次節で各手法における MAE の平均値の最も優れた値を報告する.

表 1: 会員比率毎の性能比較.

会員比率		比較手法群		
会員	非会員	NMF	NMMF	PROPOSED
5%	95%	0.685±0.153	**0.230**±0.062	0.299±0.063
10%	90%	0.313±0.215	0.338±0.151	**0.249**±0.093
20%	80%	0.452±0.374	0.385±0.132	**0.318**±0.139
40%	60%	0.424±0.50	0.326±0.071	**0.239**±0.242

4.2 実験結果

表 1 に実験結果を示す. 提案手法は NMF を会員比率によらず上回っていることが確認できる. また, 会員比率が 5%の時には, 提案手法と NMMF は同等程度の性能となることが分かる. これは会員比率 5%(会員数約 250 人) では提案手法で導入した比例関係が十分成り立っていないことが要因になっていると考えられる. しかし, 会員比率 10%以上の設定では提案手法は安定して NMMF を上回る性能を示していることが分かる. したがって, 低ランク表現行列間に明示的に比例関係を導入することで精度向上が可能であることが分かり, 提案手法の妥当性が確認された.

5. まとめ

本研究では, 会員顧客購買履歴と非会員顧客購買履歴を統合的に解析する手法を提案し, 実購買データを用いてその有効性を検証した. 本手法によるアプローチは小売店データの解析のみならず, 粒度の異なる複数のデータを解析する問題に対して有効であると考えられる. さらなる発展・検証が今後の課題である.

参考文献

[1] A. Cichocki, R. Zdunek, A. H. Phan, and S. I. Amari. Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. John Wiley & Sons, 2009.

[2] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. In *Nature*, 1999.

[3] H. Lee and S. Choi. Group nonnegative matrix factorization for EEG classification. In *AISTATS*, 2009.

[4] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *NIPS*, 2007.

[5] A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In *KDD*, 2008.

[6] K. Takeuchi, K. Ishiguro, A. Kimura, and H. Sawada. Non-negative Multiple Matrix Factorization. In *IJCAI*, 2013.

[7] 竹内考, 石黒勝彦, 小林哲生, 藤田早苗, 平博順. 複合非負値行列因子分解 (NM2F) による絵本データセットからの多角的パターン抽出. In *JSAI*, 2014.