ヒントによる局所相関推論

# Inference of Local Feature Correlations by Hints

ジェイ ホンジェ	原口 誠	今井 英幸
Hongjie ZHAI	Makoto HARAGUCHI	Hideyuki IMAI

### 北海道大学 大学院情報科学研究科

Graduate School of Information Science and Technology, Hokkaido University

We propose in this report a method for searching local correlations among features, where the locality means that the correlations are verified only for some parts of non-negative data matrix. As the number of such local correlations is large, we try to control the diversity by presenting a hint set, a small set of feature pairs demonstrating local correlations. Then a subspace with composite features as its axes is constructed so that the paired features become close on the subspace. The process of searching such a subspace is realized by applying NMF for shared subspace with a graph regularization term.

## 1. はじめに

データベースにおいて観察可能な属性やアイテムの相関 (correlation, association) 関係の検出は、データマイニングにおける基本問題の一つであり続けている.ここで、多数の個体により支持される大域的相関のみが興味の対象ではなく、相関を支持する個体数が比較的に少数であるがゆえに、「目立たない」 局所的相関も重要である.特定のユーザにとっては発想や発見のトリガーとなりえるからである [Taniguchi 2006].

似たような問題は、類似性を論じるときにも生じる、類似 性と同時に細かな差異はいくらでも存在し、よって可能な類似 性の中から妥当な類似性に絞り込む操作を必要とする、通常、 そうした絞り込み操作はタスクや文脈・観点に依存する、類推 学習における先駆的研究 [Greiner 1985] では、類推における 類似物の対をヒント  $f_i^{(1)} \sim f_i^{(2)}$ として与え、ヒントと整合的 な類似性のみに絞り込む、つまり、ヒントを文脈や観点の例示 として扱うわけである、類推図式としては

こうした類推的な操作は異なるデータセット間での属性の対応付け問題にも適用でき、実際、[Zhai 2017] で既に報告済である.本稿では、単一データセット内で成立する局所相関検出問題に対し、[Zhai 2017] のアイデアを適用する.すなわち、データセットにおいて対応付けられるべき属性 (feature)の対  $< f^{(1)}, f^{(2)} > 0$ いくつかをヒントとして与え、属性の無相関部分(行)を捨象できる部分空間の生成系で、対になった属性を近接したベクトルに射影できるものを求める.別の言い方をすれば、射影ベクトルは元の属性ベクトルの(行に関する)部分を表現させ、類推図式における条件 (SIM)を部分空間における近接性で捉える.一旦、こうした部分空間が得られれば、ヒントに出現しない属性対も、射影した結果が近接すればヒントが与える観点に従って対応付け可能だと推論・推測するわけである.

部分空間(厳密には凸多面体錘)の生成系は元の属性の線形 結合で得られる合成項である.線形結合による合成属性はデー タが非負の場合,非負行列分解によるデータ圧縮で得ることが

ニーカ レボゴールはの相人		$f_1$	$f_2$	$f_3$	$f_4$	$g_1$	$g_2$
アータ行列 X かノール値の場合. ヒント $\mathcal{H} = \{ < f_1, f_2 >, \}$	$x_1$	1 '	1			1	0
$< f_3, f_4 > \}$	$x_2$	1	1			1	0
右図は X の列をヒント属性	$x_3$	1	1	1	1	1	1
${f_1, f_2, f_3, f_4}$ に制限した部分行 列 属性対の無相関部分(行)け	$x_4$			1	1	0	1
生成系 $G = (\mathbf{q}_1 \ \mathbf{q}_2)$ が張る部分	$x_5$	1	•			0	0
空間への射影により無視できるが、	$x_6$	1	1	14	N	0	0
$H = (f_1 f_2 f_3 f_4) \mathcal{O}$ 圧縮と	$x_7$		1	1		0	0
しての G は、 悪相関部分の 括家 に 起因 する 近似 精度 劣化 を 伴う	$x_8$		1	i	1	0	0
	$x_9$	無相	開始	盼	1	0	0

#### 図 1: 局所相関と生成系の例

でき、ヒント属性対の近接性条件を正則化項として組み込んだ 非負行列分解を施せばよいことも直ちにわかる.具体的には、 共通空間法 [Gupta 2010] とグラフ正則化項 [Cai 2011] を伴う 非負行列分解を用いる.ただし、図1で示すように、相関部 分を表す生成系は無相関部分を考慮しないために、実際問題と しては、ヒントとの整合性と圧縮精度のトレードオフ問題とし て扱い、テクニカルには正則化項の重みであるラグランジェ定 数で調整を行うとする.

なお、用語としての「局所相関」は部分で成り立つ自己相関 を本稿では意味し、ブール値の場合はアイテム集合に対する 集合論的相関と概念的には同じである.また、本稿の枠組みは ヒント数を増加させれば「関係学習」ともみなせるが、個々の ユーザが興味を示すヒントを多数与えるのは骨のおれるタスク であり、「学習」ではなく、あくまでも比較的少数のヒントから の推測・推論問題として捉えている.

## 2. ヒントと局所相関

 $X = (f_1 \cdots f_N)$ を非負の  $M \times N$  データ行列とする.  $f_i$ は列  $f_i$  の M 次元列ベクトルである. 局所相関推論のためのヒ ント  $\mathcal{H}$  とは、列の対の集合  $\mathcal{H} \subseteq \{f_1, ..., f_N\}^2$ で、一般に列 の多対多関係であってよい.本稿では、前節で述べたように、

ヒントと一定の整合性を持ち,かつ,ヒントを構成 する列の圧縮表現をNMFにより求め,これを残り の列の圧縮の一部として組み込むことにより局所相

連絡先: 原口誠, 北海道大学大学院情報科学研究科, 札幌市北 区北14西9, 011-706-7106, mh@ist.hokudai.ac.jp

関推論を実現する.

ヒントに現れる属性を  $H = P_1(\mathcal{H}) \cup P_2(\mathcal{H})$  とし、以下、H中の属性を先行して配置した行列を同じ記号 H で記す. ここ で $P_i$  は属性の関係 R に対する射影である.また Y は X から H中の属性の列ベクトルを除去してできる部分行列である.

### ヒント属性からなる部分行列 H の非負行列分解

## 要請1:部分空間におけるヒント属性の近接性

<br/>
<  $f_i, f_j > \in \mathcal{H}$ とする.生成系 G に関する座標  $w_i \ge w_j$ <br/>は近接しなければならない.これは制約だが、ラグランジェ定数<br/>  $\lambda$ を用い、グラフ正則化項を伴う非負行列分解 [Cai 2011] のや<br/>
り方を踏襲する.すなわち、行列分解の近似誤差  $||H - GW||^2$ <br/>に下記の項の  $\lambda/2$  倍を加えた量の最小化を試みる.

$$\sum_{\langle f_i, f_j \rangle \in \mathcal{H}} \left\| \boldsymbol{w}_i - \boldsymbol{w}_j \right\|^2 = \sum_{\substack{f_i, f_j \in H}} s_{ij} \sum_{k=1}^{K_H} \left( w_{ki} - w_{kj} \right)^2$$
$$= 2 \operatorname{Trace} \left( WLW^t \right)$$

ただし、 $s_{ij}$ は  $f_i$  と  $f_j$ の結合重みで、ヒント中の対のときに 1、それ以外で0とする.また、L は ヒント属性  $f_i, f_j$ を頂 点とし、辺重み  $s_{ij}$ を持つグラフのラプラシアンである.

#### 要請2:共通部分空間制約

ヒントに出現しない Y の列を 生成系 G が張る部分空間 に射影するために、G とその「補空間」として残りの部分行 列 Y の圧縮 ( $G G^{(Y)}$ )を求める.

$$Y \approx \left( G \ G^{(Y)} \right) \begin{pmatrix} W^{(G)} \\ W^{(Y)} \end{pmatrix}$$

Y の圧縮として H と同じ圧縮 G を含ませた形が部分空間共 有制約を表していることに注意する.また,係数行列のうち, 部分行列  $W^{(G)}$  の第 j 列  $w_j^{(G)}$  が非ヒント属性  $f_{|H|+j}$  の共 通空間における座標ベクトルになる.

以上をまとめると、最適化(最小化)の式を下記で与えるこ とができる.反復式と収束性の詳細は省略するが、共通空間法 およびグラフ正則化項を持つ非負行列分解に対する論証を組み 合わせることにより示せる.

$$\begin{aligned} |H - GW||^2 + \lambda \operatorname{Trace} \left( WLW^t \right) \\ + \left\| Y - \left( G \ G^{(Y)} \right) \begin{pmatrix} W^{(G)} \\ W^{(Y)} \end{pmatrix} \right\|^2 \end{aligned}$$

### 局所相関推論

ヒント属性以外の属性  $f_{|H|+i}$ ,  $f_{|H|+j}$  に対しても共通空間の座標ベクトル  $\boldsymbol{w}_i^{(G)}$ ,  $\boldsymbol{w}_j^{(G)}$  が近接しているときに、ヒントと同じ観点で相関していると推論する.

#### 3. 実験

実験では、3千個の Reuters 記事(行)で約7千個の名詞 (列)を持つデータセットに対し、Gが張る部分空間において 近接してくる非ヒント属性について調べた. 局所相関実験とし ては、類似・相似関係にある名詞対をいくつか与え、それを説 明できる部分空間の軸の構成実験が最も適当であると考えるが, それらは記事における話題に必ずしも関連してくるとは限ら ず、今回の小規模実験においてはある観点から名詞間の結合強 度を結果的に推論する課題を選定した.具体的には、事故とい う観点から <bike,accident>, <motorcycle -- crash> 等 の12個のヒントを与え、ヒント以外の属性対を射影空間で の余弦類似度の降順に並べ、その上位のものについて調べた. その結果, <automobile,danger> <aircraft,dysfunction> <aircraft,scratch> など,ヒントが与える視点に照らして 関連づけられる語彙対が部分空間で近接することを確認でき た.これは、Hの圧縮次元を10に設定した場合だが、12個 の属性対を近接性制約を守る形で圧縮するためにどの程度の次 元が必要なのかが明らかになったわけではない.感覚的には, 12個の対に対して10次元だと、λの設定にもよるが、非相 関部分をかなり表現する余裕を残していると思われ、逆に言え ば,部分空間の生成系には研究意図とは外れたものが含まれて いる可能性も残っている.ただし、現在の最適化とパラメータ 設定でも、部分空間の非直交基底の中には例えば death に関 係する軸の形成に成功している.

## 4. 今後の課題と展望

最も重要な課題は、ヒントの属性対が共起・相関する行の規 模と複雑さの問題を明らかにすることである。今回の最適化は トレードオフ問題として安直に攻めたので、この課題が陽には 表現できていない。そもそも、局所相関部分があまりにも微小 な場合は、ヒントの証拠となる行が存在しないとみなせるから である。この目的のためには、(1) ヒント行列の圧縮生成系に 非相関部分の表現を許した形にする、あるいは(2) 所与のヒ ントと整合的な行を離散的な方法で直接求める、などが考え られ、これらを順次テストしていく予定である。さらには、今 回の方法はベクトル化されたデータであれば何でも良いので、 規模を大きくした問題、例えば、分散意味表現なども当然試行 すべき対象データとなるだろう。

## 参考文献

- [Taniguchi 2006] T. Taniguchi, M. Haraguchi: Discovery of Hidden Correlations in a Local Transaction Database Based on Differences of Correlations, Engineering Application of Artificial Intelligence, 19(4), pp. 419 – 428, Elsevier (2006).
- [Greiner 1985] R. Greiner. "Learning by understanding analogies". Ph.D.Thesis, Stanford University (1985).
- [Gupta 2010] S.K. Gupta, D.Phung, B.Adams, T. Tran and S. Venkatesh: Nonnegative Shared Subspace Learning and Its Application to Social Media Retrieval, KDD'10, 1169 – 1178 (2010).
- [Cai 2011] D. Cai, X.He, J.Han, T.S.Huang: Graph regularized nonnegative matrix factorization for data representation, IEEE Trans. Pattern Anal. Mach. Intell., 33(8), 1548-1560 (2011).
- [Zhai 2017] H. Zhai, M. Haraguchi: A Linear Algebraic Inference for Feature Association, 12th KICSS, IEEE CPS, 102 – 107 (2017).