2K2-5

クラスラベルとユーザタグ情報を反映した次元圧縮に基づく分類学習の実験的考察

A Method of Clasiffication Based on Class-Dependent Masked Nonnegative Matrix Factorization

大久保 好章　　原口 誠
Yoshiaki OKUBO　Makoto HARAGUCHI

北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

In this paper, we discuss a classification method for high-dimensional data based on a low-rank basis matrix identified by Masked Non-negative Matrix Factorization (Masked-NMF), where a mask matrix works as a structural constraint on the basis to be obtained. Given a training dataset (samples) with class labels, we first try to construct a mask matrix reflecting the class information. Intutively speaking, the mask consists of boolean vectors each of which corresponds to a set of original features frequently cooccuring in the samples with a certain class label. By Masked-NMF with the mask, we can obtain a low-rank basis in which each basis vector has the same strucure as its corresponding mask vector's. Thus, each basis vector is associated to some class for which its mask is constructed. As a data with unknown class label can be approximated as a linear combination of the basis vectors, the weighting would provide us a useful hint on which class it could belong to. We present a method of classification based on this idea.

1. はじめに

本稿では，非負実数の高次元ベクトルとして与えられるデータを対象としたクラス分類問題のための次元圧縮手法について議論する．

クラス分類は，**機械学習**における主要なタスクのひとつであり，クラスラベルが既知のデータ (オブジェクト) 集合をもとに，クエリとして与えられたクラスが未知のデータが属する適切なクラスを予測する**教師あり学習問題**のひとつとされている．その身近な具体例として，**文書分類** [1]，**手書き文字の認識** [2]，**音楽のジャンル分類** [3] 等が挙げられる．

機械学習の様々なタスクにおいて，対象となるデータが高次元ベクトルで表現される場合，次元の呪いを回避すべく，一般には**次元圧縮**が不可欠である．こうした処理を経ずに満足のいく学習結果を得ることは困難であり，通常，高次元データは，何らかの情報を (できるだけ) 保存したまま，より低次元に近似された後に処理される．PCA [4] や SVD [5] をはじめとする様々な次元圧縮手法が知られているが，近年，非負値データを対象とした**非負値行列因子分解** (*NMF: Nonnegative Matrix Factorization*) が注目されている [6]．その一番の特徴は，元の高次元空間のデータが，(より少数の) **非負値基底ベクトル**の**非負重みによる線形結合**として近似される点にある．すなわち，ここでの基底ベクトルをひとつの部品と捉えると，元のデータは幾つかの部品の加算のみで近似され，減算がないという意味で自然な解釈を与え得るものと期待できる．

データを解釈する立場から NMF は極めて望ましい特徴を有する次元圧縮手法であるが，個々の部品 (基底ベクトル) に対するより明確な意味付けができれば，その有用性はさらに増すと考えられる．例えば，クラス分類問題を考えた場合，各部品がそれぞれ特定のクラスに特徴的なものであれば，部品の線形和によるデータ $\boldsymbol{x}$ の近似表現は，$\boldsymbol{x}$ がどの様なクラスの部品をどの様な重要度で用いて構成されているかを表すものとなり，そのデータが属するクラスを予測する際の有用な情報となることが期待できるであろう．つまり，$\boldsymbol{x}$ の近似表現において，あるクラス c に関する部品の重みが他と比べて大きな場合，$\boldsymbol{x}$ のクラスは c であると考えるひとつの根拠となろう．

本稿では，こうしたクラス分類に有用な次元圧縮を，**マスク制約付き非負値行列因子分解** (*Masked-NMF: Masked Nonnegative Matrix Factorization*) [7] の枠組みを基礎に実現することを試みる．マスク制約は，次元圧縮により得られる各基底ベクトル (部品) の構造を明示的に定めるものであり，元の高次元空間で観測される各クラスにおける属性間の共起関係をマスクに反映させることで，それぞれのクラスに関連した部品群を得ることが可能となる．

連絡先: 大久保 好章
北海道大学大学院情報科学研究科
〒 060-0814 札幌市北区北 14 条西 9 丁目
yoshiaki@ist.hokudai.ac.jp

2. 非負値行列因子分解

非負実数の m 次元列ベクトルで表現されるデータ (オブジェクト) $\boldsymbol{x}_i \in \mathbb{R}_+^m$ $(i \in \{1,\ldots,n\})$ から成る，$m \times n$ の非負値行列 $X = (\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_N)$ を考える．X の**非負値行列因子分解** (*NMF: Nonnegative Matrix Factorization*) とは，$W \in \mathbb{R}_+^{m\times k}$ $H \in \mathbb{R}_+^{k\times n}$ なるふたつの非負値行列 W と H の積により X を近似する操作である．すなわち，

$$X \approx WH \tag{1}$$

であり，W を**基底行列**，H を**係数行列** と呼ぶ．ここで 基底ベクトルの数を表すパラメータ $k \in \mathbb{N}_+$ は通常ユーザが設定する．

$W = (\boldsymbol{w}_1 \ \cdots \boldsymbol{w}_k)$ とすると，式 (1) より，各データ $\boldsymbol{x}_j$ は，

$$\boldsymbol{x}_j \approx \sum_{i=1}^{k} \boldsymbol{w}_i h_{ij} \tag{2}$$

となり，$\boldsymbol{x}_j$ は W を構成する各基底 (列) ベクトル $\boldsymbol{w}_i$ を h_{ij} で重み付けした線形結合として近似されることがわかる．つまり，$H = (\boldsymbol{h}_1 \ \cdots \boldsymbol{h}_n)$ の第 j 列ベクトル $\boldsymbol{h}_j$ は，j 番目のデータ $\boldsymbol{x}_j$ を近似する際の各基底に対する重みベクトルを与える．

一般に $k < \min\{m, n\}$ と設定することで，m 次元空間のデータがより低次の k 次元空間のデータとして近似されることから，次元圧縮が実現される．

近似の度合いは多くの場合，行列のフロベニウスノルム $\|\cdot\|_F$ を用いて $\| X - WH \|_F^2$ により評価される．よって，所与の X に関する非負値行列因子分解は，非負性制約のもとで $\| X - WH \|_F^2$ を最小化する W および H を求める最適化問題として定式化される．

この最適化問題に対する真の最適解を得ることは困難であることから，次に示す局所最適解を得るための**乗法的更新式** [6] が広く知られている．

$$H_{ij} \leftarrow H_{ij} \frac{(W^T X)_{ij}}{(W^T W H)_{ij}} \tag{3}$$

$$W_{ij} \leftarrow W_{ij} \frac{(X H^T)_{ij}}{(W H H^T)_{ij}} \tag{4}$$

ここで，行列 A について，その (i, j) 成分を A_{ij} と表す．

実際の NMF の計算処理では，最初に W と H の各成分を適当な値によって初期化し，式 (3) と (4) を用いてそれらを交互に繰返し更新する．これらの値は更新処理に伴い非増加であることが示されていることから，値が収束した後，もしくは，十分な回数の更新が行われた後の W と H が局所最適解として出力される．しかし，これら更新式は行列積演算を含むことから，高次元かつ大規模なデータ行列に対する繰返し計算の負荷は大きい．また，W と H の初期値設定が，得られる局所最適解の品質や収束速度に大きく影響するという問題もある．

2.1 マスク制約付き NMF

先に述べた通り，NMF により，高次元データ行列 X の各データは，より少数の基底ベクトル (部品) の非負重みによる線形結合で近似される．基底ベクトルの加算のみでデータを表現できることは，データの自然な解釈を得る上で極めて有用であり，これが次元圧縮に NMF を用いる一番の利点と言っても過言ではない．

一方，上述した標準的な NMF の枠組みにおいて，解釈のもととなる部品群を定める W に関する制約は非負性のみであり，次元圧縮後の基底ベクトル (部品) 数 k 以外に構造的な制約は一切ない．しかし，W の各基底ベクトルは，その部品における元の高次元空間を定める属性の重要度を示すものであり，W の構造は部品そのものの解釈に大きく影響する．よって，データのより良い解釈を得るためには，従来の非負性に加え，何らかの構造的な制約を W に課すことが望ましい．

そのひとつのアプローチとして，**マスク制約付き非負値行列因子分解** (*Masked-NMF: Masked Nonnegative Matrix Factorization*) が提案されている [7].

所与のデータ行列 $X \in \mathbb{R}_+^{m \times n}$ と 圧縮次元数 k について，Masked-NMF では，W と同じ型のブール行列 $M \in \{0, 1\}^{m \times k}$ を用いて $X \approx (M \odot W)H$ なる X の近似を考える．ここで M を**マスク行列**と呼び，$\odot$ は行列のアダマール積 (成分毎の積) を表す．すなわち，元のデータを近似する際に用いる部品は，$(M \odot W)$ の列ベクトル $\boldsymbol{p}_j$ として与えられ，その構造はマスク行列 M の対応する列ベクトル $\boldsymbol{m}_j$ に従う．つまり，$\boldsymbol{p}_j$ で非零の値をとれる成分は，対応する $\boldsymbol{m}_j$ の成分が必ず 1 でなければならない．この様に，マスク行列は，次元圧縮後の基底行列の構造を明示的に制約するものであり，そこにユーザの興味を反映させることで，各部品に何らかの意味を持たせることが可能となる．

標準的な NMF と同様，Masked-NMF も，非負性制約のもとで $\| X - (M \odot W)H \|_F^2$ を最小化する W および H を求める最適化問題として定式化され，次に示す局所最適解を得るための乗法的更新式が与えられている [7] [*1].

$$H_{ij} \leftarrow H_{ij} \frac{((M \odot W)^T X)_{ij}}{((M \odot W)^T (M \odot W) H)_{ij}} \tag{5}$$

$$W_{ij} \leftarrow W_{ij} \frac{(M \odot (X H^T))_{ij}}{((M \odot W) H H^T)_{ij}} \tag{6}$$

式 (5) および (6) による繰返し更新処理で得られる W と H がその初期値設定に影響される点は，標準的な NMF と同様である．

3. クラスラベルを反映したマスク制約付き NMF によるクラス分類

本節では，Masked-NMF の枠組みを利用したクラス分類手法について議論する．

3.1 基本アイデア

前節で述べた通り，Masked-NMF では，適切なマスク行列を用意することで，次元圧縮後の基底ベクトル (部品) にユーザの意図を反映させることができる．所与のデータ $\boldsymbol{x}$ は，これら部品群の非負重みによる線形結合として近似されるが，このことは，$\boldsymbol{x}$ を構成する部品群とそれらの重要度がわかることを意味する．よって，各部品が特定のクラスに深く関わるものであれば，こうした近似表現によって，データ $\boldsymbol{x}$ が属するクラスに関する有用な情報が得られると期待できる．例えば，$\boldsymbol{x}$ の近似表現において，あるクラス c に関連する部品の重みが他と比べて大きな場合，$\boldsymbol{x}$ はクラス c に属すると考えることは極めて自然であろう．

本稿では，上記の考え方に基づくクラス分類手法を提案する．具体的には，クラスラベルが既知の (高次元) データ行列 X を訓練データとし，Masked-NMF により $X \approx (M \odot W)H$ なる，基底行列 W と 係数行列 H を求める．ここで M は，訓練データの各クラス毎に共起頻度の大きな属性集合族を抽出することで，その次元数 (基底ベクトル数) を含めて自動的に構築するものとする．

こうして得られた W をもとに，クラスが未知の (高次元) データ $\boldsymbol{x}$ の近似表現，すなわち，$\boldsymbol{x} \approx W\boldsymbol{h}$ における係数 (列) ベクトル $\boldsymbol{h}$ を求め，そこでの最大重みを有する部品 (基底ベクトル) が関係するクラスを，$\boldsymbol{x}$ が属するクラスであると予測する．

以下では，これら処理の詳細について議論する．

3.2 クラスを反映したマスク行列

ここでは，分類タスクに有用な次元圧縮を実現すべく，データに付与されたクラスラベル情報をもとに，圧縮後の基底行列の構造を制約するマスク行列を自動構築する手法を与える．

NMF を用いた次元圧縮により得られる各基底ベクトルは，一言で述べると，関連するいくつかの属性群をひとつの部品としてまとめたものである．あるデータにおいてその関連性が観測される場合，その部品はデータを近似する際に大きな重みのもとで使われる．よって，あるクラスのデータを近似する際に主に使われる部品は，そのクラスにおいて関連性が観測される属性群をもとに構成すればよい．具体的な関連性としては，

*1 実際の文献 [7] では，罰則項を含めた目的関数の最小化問題を議論しているが，以降の議論では不要なためここでは除外している．

各属性を確率変数とみなした場合の統計的な相関や，データ中の属性の有無に着目した共起度などが挙げられる．以下では，文書データ等の扱いに馴染む後者について議論を進める．

m 次元 (非負) 列ベクトル $\boldsymbol{x}$ について，その第 ℓ 成分を，$\boldsymbol{x}$ の第 ℓ 属性 A_ℓ の値と考え，これを $A_\ell(\boldsymbol{x})$ で参照する．すなわち，$\boldsymbol{x} = (x_1 \cdots x_m)^T$ とすると，$A_\ell(\boldsymbol{x}) = x_\ell$ である．特に，$A_\ell(\boldsymbol{x})$ が非零の時，$\boldsymbol{x}$ は属性 A_ℓ を有する，あるいは，属性 A_ℓ は $\boldsymbol{x}$ に関連すると言う．

いま，m 次元 (非負) 列ベクトル $\boldsymbol{x}_i$ から成るデータ行列 $X = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_n)$ について，属性 A_ℓ が関連する列ベクトルの集合を $V_{A_\ell}(X)$ で参照する．つまり，

$$V_{A_\ell}(X) = \{\boldsymbol{x}_i \mid \boldsymbol{x}_i \text{ in } X, A_\ell(\boldsymbol{x}_i) > 0\}$$

である．

X における属性 A_i と A_j について，それらの**共起度**を $CoOcc_X(A_i, A_j)$ と表記し，

$$CoOcc_X(A_i, A_j) = \frac{|V_{A_i}(X) \cap V_{A_j}(X)|}{n}$$

と定義する．共起度は区間 $[0, 1.0]$ の値をとり，両属性が共に関連する X 中のベクトル集合が多いほど大きな値をとる．

いま，共起度の下限閾値を δ とし，各属性を頂点，共起度が δ 以上の (異なる) 属性間を辺で接続した m 頂点の無向グラフ

$$G_\delta = (\{A_i\}_{i=1}^m, \{(A_i, A_j) | CoOcc_X(A_i, A_j) \geq \delta, i \neq j\}).$$

を考える．すると，G_δ の**クリーク** (完全部分グラフ) は，共起度が互いに閾値以上である属性集合を与え，特に，クリークがある程度の大きさを有する場合は，その属性集合に関連するデータが X 中にそれなりの数存在することを期待できる．よって，G_δ の極大クリークを構成する属性集合をもとに基底ベクトルを作成すれば，X を近似する際の有用な部品となろう．

Masked-NMF による次元圧縮でこうした基底ベクトルを得るためには，比較的サイズが大の極大クリークを構成する属性集合 $\mathcal{Q}$ に対応するブーリアンベクトル $\boldsymbol{b}_\mathcal{Q} = (b_1, \ldots, b_m)^T$ をマスク行列の列ベクトルとすればよい．ここで，

$$b_i = \begin{cases} 1, & \text{if } A_i \in \mathcal{Q} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

である．

所与のグラフにおける極大クリークは一般に多数存在するが，サイズが大きなものは比較的少数である．ここでは，特に，X 中の一定の割合以上のデータと実際に関連する極大クリーク族 [*2] をもとに，マスク行列を作成するものとする．よって，圧縮後の次元数 k をユーザが明示的に事前に与える必要はない．なお，極大クリークの抽出処理については，大規模グラフにおいても実用上十分高速に動作するアルゴリズムが既にいくつか提案されている (例えば [9])．

本稿では特に，クラス分類タスクに有用な次元圧縮を考えるため，クラスラベル情報を付与されたデータ行列 X をクラス毎に分割し，各部分データ行列に対して上述した処理を行なうものとする．

*2 これは，X をトランザクションデータと見做した場合の，極大頻出パターン [8] の一部に相当する．

具体的には，m 次元列ベクトル $\boldsymbol{x}_i$ から成る非負値データ行列 $X = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_n) \in \mathbb{R}_+^{m \times n}$ について，各 $\boldsymbol{x}_i$ には，クラスラベル集合 $\mathcal{CL} = \{C_1, \ldots, C_\ell\}$ 中のひとつのラベルが付与されていると仮定し，それを $class(\boldsymbol{x}_i)$ で参照する．こうしたクラスラベルの情報をもとに，X を各クラス毎の部分データ行列 $X_{C_1}, \ldots, X_{C_\ell}$ に分割する．ここで，X_{C_i} は列ベクトルの集合 $\{\boldsymbol{x} \in X | class(\boldsymbol{x}) = C_i\}$ を行列表現したものである．

先の手順に従い，各データ行列 X_{C_i} に対して，属性を頂点とする共起度に基づく無向グラフ $G_{C_i,\delta}$ を構築し，そこでの (サイズ大の) 極大クリーク族をもとに (部分) マスク行列 $M_{C_i} \in \{0, 1\}^{m \times k_{C_i}}$ を作成する．ここで，k_{C_i} はマスク行列作成に用いた極大クリークの数である．

こうして得られた ℓ の部分マスク行列を用いて，元のデータ行列 X に Masked-NMF を適用する際のマスク行列 M を

$$M = (M_{C_1} \cdots M_{C_\ell})$$

と定義する．すなわち，圧縮後の次元数 (基底ベクトル数) は $\sum_{i=1}^{\ell} k_{C_i}$ となる．

3.3 未知データのクラス予測

クラスラベルが付与されたデータ行列 $X \in \mathbb{R}_+^{m \times n}$ に，上述したクラスラベル情報を反映したマスク行列 $M = (M_{C_1} \cdots M_{C_\ell})$ を用いて Masked-NMF を適用した結果，基底行列 $W \in \mathbb{R}^{m \times K}$ が得られたとする．ここで，$K = \sum_{i=1}^{\ell} k_{C_i}$ である．ここでは，W を用いてクラスが未知の所与のデータが属するクラスを予測する手続きを与える．

W の各基底ベクトルは，特定のクラスのデータに関連する属性群を部品としてまとめたものに相当する．よって，W を用いてあるデータを近似する際，各部品に対する重みは，そのデータが属するクラスに関する有用な情報を与える．すなわち，あるクラスに対して作られた部品の重みが大きいことは，データがそのクラスに属すると考えるひとつの合理的な根拠となるであろう．本稿では，この考え方に従って所与のデータが属するクラスを予測する．

m 次元非負ベクトルで表現されたクラスが未知のデータ $\boldsymbol{x}$ を考える．いま，$\boldsymbol{x}$ が $W = (\boldsymbol{w}_1 \cdots \boldsymbol{w}_K)$ を用いて $\boldsymbol{x} \approx W\boldsymbol{h}$ と近似できたとする．具体的には，$\| W\boldsymbol{h} - \boldsymbol{x} \|^2$ を最小化する $\boldsymbol{h}$ は，W の一般逆行列 (ムーア・ペンローズ逆行列) W^+ を用いて，$\boldsymbol{h} = W^+\boldsymbol{x}$ として求めることができる．この時，列ベクトル $\boldsymbol{h} = (h_1 \cdots h_K)^T$ の各成分 h_i は，データ $\boldsymbol{x}$ を近似する際の部品 $\boldsymbol{w}_i$ の重みを与える．よって，ここでは，クラス毎にそこでの各部品に対する重みの総和をとり，それを $\boldsymbol{x}$ の近似表現における各クラス部品 (群) の重要度と考える．

$W = (\boldsymbol{w}_1 \cdots \boldsymbol{w}_K)$ の構造は，マスク行列 $M = (M_{C_1} \cdots M_{C_\ell})$ により制約されており，各基底ベクトルについて，その対応するクラスが一意に決まる．いま，各クラス C_i に対応する W の列インデックスを，行ベクトル $\boldsymbol{c}_i \in \{0, 1\}^K$ で与える．すなわち，$\boldsymbol{c}_i = (c_1 \cdots c_K)$ において，

$$c_j = \begin{cases} 1, & \text{if } (\sum_{p=1}^{i-1} k_{C_p}) < j \leq (\sum_{p=1}^{i} k_{C_p}) \\ 0, & \text{otherwise} \end{cases}$$

である．こうした $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_\ell$ を上から順に縦に並べた行列を $C = (\boldsymbol{c}_1^T \ldots \boldsymbol{c}_\ell^T)^T$ とすると，$\boldsymbol{x} \approx W\boldsymbol{h}$ における，クラス毎の部品重み総和の分布 $\boldsymbol{v}$ は，$\boldsymbol{v} = (v_1 \cdots v_\ell)^T = C\boldsymbol{h}$ で与えられる．その最大成分が v_α である時，$\boldsymbol{x}$ の予測クラスを C_α とする．

4. おわりに

訓練データに付与された各クラス情報を反映した次元圧縮を実現すべく，本稿では，各クラス毎の基底ベクトル構造を与えるマスク行列を自動構築し，それを用いた Masked-NMF の結果をもとに，未知のデータが属するクラスを予測する枠組みを与えた．

現在，既存の分類システムとの比較を含め，提案システムの有効性の実験的検証を進めており，その結果は口頭発表時に報告したい．

未知データに対するクラス毎の部品重み総和の分布情報を利用した**マルチラベル分類** [10] への拡張は，興味深い今後の課題のひとつである．

参考文献

[1] Sebastiani, F.: Machine Learning in Automated Text Categorization, ACM Computing Surveys, 34(1), pp. 1 – 47, 2002.

[2] Smith, S. J., Bourgoin, M. O., Sims, K. and Voorhees, H. L.: Handwritten Character Classification Using Nearest Neighbor in Large Databases, IEEE Transactions on Pattern Analysis and Machine Intelligence, 16(9), pp. 915 – 919, 1994.

[3] Pampalk, E., Flexer, A., Widmer, G.: Improvements of Audio-Based Music Similarity and Genre Classificaton, Proc. of the 6th International Conference on Music Information Retrieval - ISMIR'05, pp. 628 – 633, 2005.

[4] Jolliffe, I. T.: Principal Component Analysis (2nd ed.), Springer, 2002.

[5] Golub, G. H. and Loan, C. F. V.: Matrix Computations (4th ed.), Johns Hopkins University Press, 2012.

[6] Lee, D. D. and Seung, H. S.: Algorithms for Non-negative Matrix Factorization, Proc. of NIPS 2000, pp. 556 – 562, 2000.

[7] Casalino, G., Buono, N. D. and Mencar, C.: Part-Based Data Analysis with Masked Non-negative Matrix Factorization, Proc. of ICCSA 2014, Part VI, LNCS 8584, pp. 440 – 454, 2014.

[8] Aggarwal, C. C. and Han, J. (Eds.): Frequent Pattern Mining, Springer, 2014.

[9] Tomita, E., Tanaka, A. and Takahashi, H.: The Worst-Case Time Complexity for Generating All Maximal Cliques and Computational Experiments, Theoretical Computer Science, 363(1), pp. 28 – 42, Elsevier, 2006.

[10] Tsoumakas, G., Katakis, I. and Vlahavas, I.: Mining Multi-Label Data, Maimon, O. and Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook (2nd ed.), pp. 667 – 685, Springer, 2010.