様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１４　日現在

機関番号：　３０１１５
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０２５９

研究課題名（和文）競合学習による高速高精度な非負値行列因子分解法の確立

研究課題名（英文）Improvement of Nonnegative Matrix Factorization method using competitive learning

研究代表者

内山　俊郎（Uchiyama, Toshio）

北海道情報大学・経営情報学部・教授

研究者番号：８０７０８６４４

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：本研究では、従来研究と比較し高速・高精度な非負値行列因子分解（NMF）手法を確立することを目的とし、理論的な検討と実験による検証により、高速・高精度なNMF手法の導出と検証に至った。理論的には「情報理論的クラスタリングの解が一般化KLダイバージェンスを目的関数とするNMFの制約付き（ハードクラスタリング制約）の解であること」を示した。そして、情報理論的クラスタリングの解を利用した初期値設定により、従来行われていたランダムな初期値設定よりも、より優れた解に到達できることを実験により示した。また、ベクトルの部分集合を勝者とする新たな競合学習アルゴリズムを提示し、実験により有効性を示した。

研究成果の概要（英文）：This study investigated novel methods to improve the accuracy of nonnegative matrix factorization (NMF) from both theoretical and experimental side. Theoretically, it has shown the equivalence between information-theoretic clustering (ITC) and NMF based on generalized KL divergence. Then, it proposed a novel initialization method for NMF using ITC and experimentally showed the effectiveness of the method compared to conventional methods. It also proposed another algorithm for NMF using competitive learning which selects a subset of vectors as winner and showed the effectiveness.

研究分野：　データマイニング

キーワード：　競合学習　情報理論的クラスタリング　非負値行列因子分解　トピックモデル

１．研究開始当初の背景

世の中のデータの多くは、文書や画像のように、要素が非負値の行列として表現されることが多い。非負値行列因子分解のための代表的なアルゴリズムとして Multiplicative algorithm が知られている。これは求めるべき行列の要素に係数を掛けて更新し、これを繰り返す方法である。ここで、非負値行列を X(M ×N)、分解結果を W (N × K) と H (K × N) とすると、

$\boldsymbol{X} = \boldsymbol{WH}$;

が成り立てば、誤差０の理想的な分解となる。実際には誤差が発生するため、行列XとWHを引数とする目的関数D(X,WH)を最小化することが、アルゴリズムの目指すところである。しかし、前記アルゴリズムは初期値に依存する。適切な初期値設定問題は、非常に難しい問題である。

２．研究の目的

本研究の目的は、ＮＭＦの目的関数を小さくするという意味で優れた因子分解の方法を確立することである。

３．研究の方法

非負値行列因子分解（ＮＭＦ）のためのアルゴリズムである Multiplicative アルゴリズムは初期値に依存する。従来は、たとえばランダムな非負値を初期値とする方法が取られていたが、優れた解へ収束するとは限らない。そこで、ＮＭＦの目的関数と等価なクラスタリングを行い、それによって得られた解を使って、Multiplicative アルゴリズムの初期値設定を行うことを考えた。これを実現するには、ＮＭＦの目的関数と等価なクラスタリング基準とそのためのアルゴリズムを明確にする理論的な分析と、実際に実験による効果の検証が必要である。これら検討を通じて、目的の達成を目指した。

４．研究成果

ＮＭＦの目的関数の基準として、一般化ＫＬダイバージェンスの場合について検討した。この場合、画像、文書、行動履歴など様々な非負値データに対して適用されている、トピックモデル（ＰＬＳＡ、ＬＤＡ、など）におけるパラメータ推定の問題と等価な問題を解くことになる。さらに、発表論文の中で、情報理論的クラスタリングが、トピックモデルにおけるハードクラスタリング制約解であることを示し、この情報理論的クラスタリングを用いて、初期値設定を行う方法を提案した。厳密には、ＮＭＦやトピックモデルにおいては、特徴の出現を等しく扱い、クラスタリングでは、特徴を含むデータの出現を等しく扱うという違いがある。そのため、重み付き情報理論的クラスタリングがトピックモデル（ＮＭＦ）の制約解であると言える。以下、等価になることの説明を記す。

文書などデータを生成する K 個のトピック $k\{k = 1,\ldots,K\}$ があるとする．トピックモデルでは，１つのデータは複数のトピックから生成されるとするが，ハードクラスタリングでは，単独のトピック k から生成されるとし，そのとき「データはクラスタ $C^k\{k = 1,\ldots,K\}$ に属する」という．ここでトピック k とクラスタ C^k は呼び方が変わるだけで同一の概念である．

データが持つ特徴を $m\{m = 1,\ldots,M\}$ と表し，特徴の出現が確率変数の取り得る値であるような確率分布 P と Q を考え，特徴 m の出現確率を p_m, q_m で表すとき，P の Q に対する KL（Kullback-Leibler）ダイバージェンス $D_{\mathrm{KL}}(P\|Q)$ は，

$$D_{\mathrm{KL}}(P\|Q) = \sum_{m=1}^{M} p_m \log \frac{p_m}{q_m}, \tag{1}$$

と書ける．今，N 個のデータについて，確率分布 $P^i(i = 1,\ldots,N)$ があり，そのうち，クラスタ $C^k(k = 1,\ldots,K)$ に属する確率分布の平均確率分布 Q^k を考えると，P^i の Q^k に対する KL ダイバージェンスの平均すなわち JS ダイバージェンス $D_{\mathrm{JS}}\left(\{P^i|P^i \in C^k\}\right)$ は，クラスタ C^k に属するデータ数 N_k を使って

$$D_{\mathrm{JS}}\left(\{P^i|P^i \in C^k\}\right) = \frac{1}{N_k} \sum_{P^i \in C^k} D_{\mathrm{KL}}(P^i\|Q^k), \tag{2}$$

と表せ，ITC のクラスタリング基準であるクラスタ内 JS ダイバージェンス JS_W は，これを全クラスタについて平均化した

$$JS_W = \sum_{k=1}^{K} \frac{N_k}{N} D_{\mathrm{JS}}\left(\{P^i|P^i \in C^k\}\right), \tag{3}$$

$$= \frac{1}{N} \sum_{k=1}^{K} \sum_{P^i \in C^k} D_{\mathrm{KL}}(P^i\|Q^k), \tag{4}$$

となる．ここまでは，すべてのデータについての確率分布 P^i を同等に扱ったが，i 番目のデータ $(i = 1,\ldots,N)$ に含まれる特徴 m の個数を x_m^i，それを特徴 m について足しあわせた特徴の総数を $t_i = \sum_{m=1}^{M} x_m^i$ と表し，この「特徴数 t_i」で重み付けをする．すなわち，クラスタの平均確率分布 Q^k は，t_i で重み付けされた平均確率分布であるとし，新たに重み付けされた JS ダイバージェンス D'_{JS} として，t_i により重み付き平均された KL ダイバージェンスである

$$D'_{\mathrm{JS}}\left(\{P^i|P^i \in C^k\}\right) = \frac{1}{B_k}\sum_{P^i \in C^k} t_i D_{\mathrm{KL}}(P^i \| Q^k \tag{5}$$

を考え，これを全クラスタについて平均化した

$$JS'_W = \sum_{k=1}^{K} \frac{B_k}{B} D'_{\mathrm{JS}}\left(\{P^i|P^i \in C^k\}\right), \tag{6}$$

$$= \frac{1}{B}\sum_{k=1}^{K}\sum_{P^i \in C^k} t_i D_{\mathrm{KL}}(P^i \| Q^k), \tag{7}$$

$$= \frac{1}{B}\sum_{k=1}^{K}\sum_{P^i \in C^k}\sum_{m=1}^{M} t_i p_m^i \log \frac{p_m^i}{q_m^k}, \tag{8}$$

$$= \frac{1}{B}\sum_{k=1}^{K}\sum_{P^i \in C^k}\sum_{m=1}^{M} x_m^i \log \frac{p_m^i}{q_m^k}, \tag{9}$$

を，重み付き情報理論的クラスタリングの基準（目的関数）とする．ここで，$x_m^i = t_i p_m^i$ という関係を使い，B_k と B は，それぞれクラスタ内および全クラスタでの重み t_i の和

$$B_k = \sum_{P^i \in C^k} t_i,\ B = \sum_{k=1}^{K} B_k, \tag{10}$$

である．式 (9) の内部は

$$x_m^i \log \frac{p_m^i}{q_m^k} = \left(-x_m^i \log q_m^k\right) - \left(-x_m^i \log p_m^i\right), \tag{11}$$

と変形でき，第 2 項はクラスタリング結果に依存しないので，第 1 項を最小化すべき目的関数と考えればよい．

一方，トピックモデルにおいて，i 番目データの特徴 m が x_m^i 個出現する確率は，トピック分布（トピック k を選ぶ確率）を θ_k^i とすれば，トピック k における特徴の確率分布 ϕ_m^k を用いて

$$\prod_{i=1}^{N} A^i \prod_{m=1}^{M}\left(\sum_{k=1}^{K}\theta_k^i \phi_m^k\right)^{x_m^i}, \tag{12}$$

と書ける．ここで，A^i は出現の組み合わせ数である．対数をとって定数となる部分を除くと，

$$\sum_{i=1}^{N}\sum_{m=1}^{M} x_m^i \log\left(\sum_{k=1}^{K}\theta_k^i \phi_m^k\right), \tag{13}$$

となる．トピック k における特徴分布 ϕ_m^k は，クラスタリングにおいてはクラスタ C^k の平均確率分布 q_m^k（t_i で重み付きされた）に対応するから，この式は，重み付き ITC の基準の大小関係を決める式 (11) の第 1 項を一般化したものといえ，

$$\theta_k^i = \begin{cases} 1 & P^i \in C^k \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

のとき、一致する（符号は逆で、最小化ではなく最大化になる）．つまり，トピックモデルの目的関数にハードクラスタリングの制約を加えたものの最大化は、重み付き情報理論的クラスタリングの基準（目的関数）の最小化と等価になる。

上記のような理論的背景の下、競合学習を用いた情報理論的クラスタリングによって得られたベクトル集合をWの初期値とし（細かくは、微小な値を全体に加えた）、一様な値を係数行列Hに与え、ＰＬＳＡのアルゴリズムを適用することで、トピックモデルのパラメータを推定する方法を提案した。

実験では、いくつかの文書データを用い、単語数が４０以上の文書集合のBag-of-words を作成し、10%の文書のうちの50%の単語をテストデータ、残りを学習データとした。学習データでモデルパラメータを得て、テストデータの単語の予測性能すなわちパープレキシティで評価を行った。実験に用いた文書データセットの特徴を表２に示す。

表 2　文書データセットの特性
Table 2　Property of text data sets

Data	Size(N)	Feature(M)
20Newsgroups	14111	60149
RCV1	685132	284338
Enron	23959	28058
Nips	1491	12375
Kos	3088	6906
NYtimes	296829	101631
Pubmed	7478052	141043

トピック数（＝クラスタ数）K=10 と K=20 の場合について、初期値を変えてトピックモデルパラメータのＭＡＰ推定（Bayesian PLSA）を行い、前述のパープレキシティにより評価を行った。

実験結果を表３と表４に示す。表において、“ Rnd ”は従来法であるランダムな初期値を用いることを表し、“ sdCL ”と“ sdCLS ”は、競合学習による「重み付き情報理論的クラスタリング」の結果を利用して初期値設定を行う提案手法である。“ sdCLS ”は分岐型の競合学習を使うという意味である。また、太字の数字は、ランダムな初期値設定を用いる従来法と比較（t 検定）し、統計的な意味（有意水準 5%）で提案手法が優れていることを表す。ほとんどの場合に、提案手法が優れていることが読み取れる。競合学習を使う両手法の結果には差がないことが多いが、データセットによっては、分岐型の場合のみ従来法よりも優れることがある。

実験結果から、提案手法は、トピックモデル（すなわち本研究の目的であるＮＭＦ）のより良いパラメータ推定をする上で有効であるといえる。

表 3　初期値設定方法の違いによるパープレキシティの比較（$K = 10$）
「太字：有意水準 5%でランダムな初期値設定よりも値が小さい」
Table 3　Comparison of perplexities with different initialization ($K = 10$)

Data	Rnd	sdCL	sdCLS
20Newsgroups	5669.9	**5637.9**	**5631.1**
RCV1	1802.1	**1789.4**	**1791.7**
Enron	2937.4	**2924.1**	**2022.1**
Nips	1991.1	**1986.4**	**1988.3**
Kos	1794.4	**1788.6**	**1789.8**
NYtimes	5982.6	**5953.0**	**5956.7**
Pubmed	4919.1	**4877.7**	**4876.1**

表 4　初期値設定方法の違いによるパープレキシティの比較（$K = 20$）
「太字：有意水準 5%でランダムな初期値設定よりも値が小さい」
Table 4　Comparison of perplexities with different initialization ($K = 20$)

Data	Rnd	sdCL	sdCLS
20Newsgroups	4762.5	**4683.3**	**4708.6**
RCV1	1542.0	**1529.8**	**1534.5**
Enron	2549.6	2552.0	**2541.8**
Nips	1812.6	1813.0	1814.9
Kos	1680.8	**1676.2**	**1673.9**
NYtimes	5158.8	5165.9	**5132.1**
Pubmed	4198.6	**4143.0**	**4145.3**

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）
[1]内山俊郎、「情報理論的クラスタリングを用いた確率的潜在意味解析の性能向上」、電子情報通信学会論文誌 D、vol.J100-D, No.3, pp.419-426, 2017
[2] Toshio Uchiyama 、"Information-Theoretic Document Clustering using Skew Divergence”、北海道情報大学紀要、vol.27-2, pp. 19-25, 2016

〔学会発表〕（計６件）
[1] 内山俊郎、「正規化相互情報量を用いたクラスタリング解の分布解析」、情報処理学会北海道シンポジウム 2016, pp.195-200, 2016
[2] 内山俊郎、「重み付き情報理論的クラスタリングを用いた確率的潜在意味解析の性能向上」、電子情報通信学会技術研究報告、PRMU, vol.116, no.89, pp.47-52, 2016
[3] 内山俊郎、「非負値行列因子分解の高精度化とＰＬＳＡへの応用」、電子情報通信学会技術研究報告、PRMU, vol.115, no.224, pp.117-122, 2015
[4] 内山俊郎、「競合学習を用いた非負値行列因子分解」、電子情報通信学会技術研究報告、IBISML, vol.114, no.198, pp.7-12, 2014

〔図書〕（計１件）
T. Uchiyama, “Information-theoretic clustering and algorithms,” Advances in Statistical Methodologies and Their Application to Real Problems, pp.93-119, InTech, 2017.

〔産業財産権〕

○出願状況（計１件）

名称：計算処理装置、計算処理方法、計算処理システム、及びプログラム
発明者：内山俊郎
権利者：学校法人電子開発学園
種類：特許出願公開
番号：特開 2016-38680
出願年月日：平成 26 年 8 月 6 日
国内外の別： 国内

○取得状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
内山俊郎（UCHIYAMA TOSHIO）
北海道情報大学・経営情報学部・教授
研究者番号：80708644

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
（　　）