

非負行列分解を用いた銀行ディスクロージャーのテキスト解析

安川武彦 takehiko.yasukawa@jp.pwc.com

あらた基礎研究所

1. 研究の動機と目的

非負行列分解 (Nonnegative Matrix Factorization, 以下 NMF) は非負の行列を次元縮約しながら 2 つの非負の行列に分解する手法である。画像処理, 音声認識, 遺伝子発現解析, 文書クラスタリングなどへの適用が報告され, その有効性が示されている。

本研究では 2008 年度の銀行ディスクロージャーデータにおけるトップメッセージ項目を対象に NMF による文書クラスタリングを実施した。さらに, NMF の初期値依存性を評価するため, 初期値を変更し繰り返し計算した結果と初期値を固定して計算した結果を比較した。その結果, クラスタリングとして利用する場合には, 両者はほぼ同様の結果を示した。

2. 非負行列分解

テキストデータを形態素解析し, 語の重み付き頻度を集計した行列 A ($m \times n$) を考える。NMF では, 非負の行列 A を k 次元に次元縮約した非負の 2 つの行列 W ($m \times k$) と H ($k \times n$) に分解する (図 1)。

$$A \approx WH$$

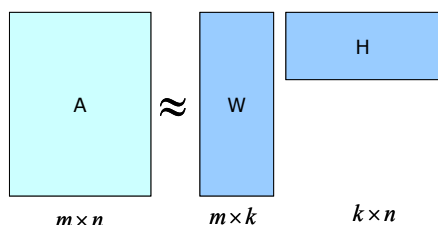


図 1 ; NMF による分解のイメージ

この分解を行うにあたり, 本研究では評価関数として次のノルムを最小化すアルゴリ

ズムを採用した (Lee and Seung (1999, 2001))。

$$\|A - WH\|^2$$

これ, 分解された行列は次のように解釈できる。

- W : 語とその重みからなる行列であり, k 個の列ベクトルは各クラスタにおける語の重みと解釈できる。なお, W の各列は直交しない。
- H : 列ベクトルは各文書がクラスタに所属する重みと解釈できる。最も大きな重みを持つ行番号を, 文書の所属するクラスタとすることで, クラスタリングに利用することができる。

3. 分析手順と結果の概要

NMF の計算結果は, 常に一意となるわけではない。また, 初期値の設定により計算結果は異なる。本研究では初期値をランダムに変え, 繰り返し計算した結果を統合することで, 最終的なクラスタリングとした。この結果, 初期値を固定して NMF を実施した結果と繰り返し計算した結果を統合したものはほぼ同様の分類結果となった。これにより, 安定的な分類結果を得ることができた。

具体的な分析手順および結果は発表当に詳細な解説を行う。

参考文献

- D. Lee. and H. Seung. (1999), *Nature*, **401**, 788-791.
- D. Lee and H. Seung. (2001), *Advanced in Neural Information Processing Systems* **13**, 556-562.