博　士（情報科学）　ミルザル　アンドゥリ

学 位 論 文 題 名

A Study on Clustering and Latent Semantic Indexing Aspects of the Nonnegative Matrix Factorization Techniques

（非負行列分解によるクラスタリングと潜在意味インデックス化）

学位論文内容の要旨

In this thesis, clustering and latent semantic indexing (LSI) aspects of the nonnegative matrix factorizations (NMFs), techniques that factorize a nonnegative matrix into a pair of other nonnegative matrices, will be studied. The standard method for clustering is the spectral clustering. And the method associated with LSI is the singular value decomposition (SVD). However, as real datasets usually are nonnegative, the mixed signs entries produced by the spectral methods and the SVD are unintuitive. Hence, it is natural to utilize the NMFs which produce nonnegative factors that can offer more interpretable results.

Clustering aspect of the NMF even though numerically well studied, it is not theoretically well explained. So far the best approaches to explain this aspect are either by showing the equivalence of the NMF objective to the k-means clustering objective, or by applying zero gradient conditions to the NMF objective to show its equivalence to graph clustering objective, i.e., ratio association. The problem with the first approach is there is no obvious way to incorporate the nonnegativity constraints into the k-means objective. And the problem with the second approach is it discards the nonnegativity constraints, thus is equivalent to finding stationary points on the unbounded feasible region. In the first part this thesis, we will provide a theoretical support for clustering aspect of the NMF by analyzing the objective at the stationary points on the feasible region without any assumption. We will show that at these stationary points, the NMF objective is equivalent to the relaxed ratio association objective, therefore clustering aspect of the NMF has a solid justification. Moreover, the capability and limitation of the NMF as a clustering method, which implied by the theoretical result, will be also studied subsequently.

To the best to our knowledge, there is still no work that study LSI aspect of the NMF. Previous studies indirectly imply the possibility of using the NMF as a LSI method, but no standard metric has been used to measure its LSI capability. Instead the studies show exactly co-clustering capability of the NMF. In the second part of this thesis, we will provide such study by comparing performances of the NMF and the SVD in some standard datasets.

Convergence guarantee is an important issue in NMF algorithms as the most popular NMF algorithm

due to Lee and Seung which is based on the multiplicative update rules only guarantees the non-increasing of the update sequences. Some researchers propose other algorithms that both converge and are faster. But due to the complexity of the algorithms, it is not clear how to incorporate auxiliary constraints like orthogonality constraints, locality constraints, and smoothness constraints into the algorithms. In the third part of this thesis, we develop orthogonal NMF algorithms with rigorous convergence proofs. We choose orthogonal NMFs because they tend to have better clustering capability than the standard NMF objective. The convergence proofs presented in this thesis is not trivial since they are developed in matrix form, and thus providing a framework for developing converged algorithms for other NMF objectives that have matrix based auxiliary constraints.

Also, in the process of developing the proofs, the objectives need to be decomposed into the Taylor series. When the objectives have only up to second order derivatives, the nonincreasing properties can be proven by showing the positive-definiteness of the Hessians of the objectives. But in general cases, the objectives can have more than second order derivatives. And in particular, the orthogonality constraints make the objectives have more than second order derivatives. Thus, we introduce a strategy to deal with this kind of objectives. Note that the proofs presented here are sufficiently general to be a framework for developing converged algorithms for other NMF objectives with well-defined partial derivatives up to second order.

Overall speaking, we provide a theoretical support for clustering aspect of the NMF by analyzing the objective at the stationary points on the feasible region, study LSI aspect of the NMF, and propose converged algorithms for orthogonal NMFs which tend to have better clustering capability than the standard NMF.

The thesis is organized into following chapters:

Chapter 1 is the introduction that gives brief review on clustering and LSI terms, followed by motivations and background of the researches.

In Chapter 2, a concise explanation on the logic behind the SVD as a clustering method will be discussed. The numerical evaluation on clustering aspect of the SVD will be provided by using synthetic datasets to show its capability in clustering linearly inseparable datasets, and Reuters-21578 text corpus to evaluate its performances in some standard clustering metrics. LSI aspect of the SVD will demonstrated by using synthetic datasets to show its capability and limitation in solving the synonymy and polysemy problems first, then a more intensive evaluation will be done by using standard datasets in LSI researches.

Chapter 3 discusses clustering and LSI aspects of the NMF. First a theoretical support for clustering aspect of the NMF will be given. And its limitation in clustering linearly inseparable data points which implied by the theoretical work will be shown by using synthetic datasets. Then, a comprehensive evaluation will be given by using Reuters-21578 text corpus. The LSI aspect of the NMF will also be studied, and the results will be compared to the results of the SVD.

Chapter 4 presents converged algorithms for orthogonal NMFs with rigorous convergence proofs. The numerical results that show the stability of the algorithms in guaranteeing the convergences will also be given, and the clustering capability of the algorithms will be also be studied subsequently.

Finally the summary and conclusions are given in chapter 5.

学位論文審査の要旨

主　査　教　授　古　川　正　志
副　査　教　授　栗　原　正　仁
副　査　教　授　鈴　木　恵　二
副　査　教　授　小　野　哲　雄
副　査　准教授　山　本　雅　人

学位論文題名

A Study on Clustering and Latent Semantic Indexing Aspects of the Nonnegative Matrix Factorization Techniques

(非負行列分解によるクラスタリングと潜在意味インデックス化)

要素が非負値をとる行列 (非負行列) は二つの非負行列の積に分解することができる. 本研究では非負行列を二つの非負行列の積に分解し, その時に得られたに行列に基づくデータセットのクラスタリングとクラスタリングされたデータに存在する潜在意解析 (LSI) を研究している. 標準的なクラスタリングの手法にはスペクトル法が使用されるが, これは LSI の場合において特異値分解 (SVD) の方法となる. 実際の多くのデータセットは非負値で表されるため,SVD を適用した場合, それによって分解された行列には正負の値が混在し, クラスタリングの解析が難しくなる. 従って, 非負行列分解を行う NMF を使用できれば, クラスタリングと LSI の解析が容易になる.

NMF とクラスタリングの関係は数値計算的にはよく研究されているが, 理論的には余り説明されていない.NMF は, 与えられた非負行列と分解された非負行列の積との自乗和を評価関数とし, その評価関数を最小にする二つの分解行列を求める最適化問題として定式化される. これまでに.NMF 結果と k-平均法の結果が等価であること,NMF の評価関数の勾配が 0 になる条件がグラフ理論による ratio association クラスタリングの結果と等価であること, の二つが主に説明されている. しかしながら, 両者ともデータセットが非負値である制約条件を満たしていない.

また,NMF の数値計算上での収束性についての理論的な解明も余り明確にされていない. 収束性に関する研究としては,Lee と Seung が未定乗数法を更新することによって逐次減少する探索が可能であること, 更に,NMF で得られる非負行列に直交性を与えることで, 収束が可能になることを示している. しかし, この更新にどのような制約を具体的に付加すれば,NMF が収束するかについては明らかにしていない. これは彼らが示した KKT(Karush Kuhn Tucker の条件) に基づくアルゴリズムの複雑性によるものである. 更に, これらの研究が LSI に応用された事例は, ほとんどない.

本研究では第 1 に,NMF に制約を置かずにその実行可能解の停留点を調べることによって, 理

論的に NMF とクラスタリングの関係を説明している. すなわち,NMF によって得られる停留点は,relaxed ratio association の結果と等価となることを明らかにし,NMF をクラスタリングに使用する正当性を明らかにしている. 第 2 に,NMF をクラスタリングに使用する時の可能性とその限界を理論的に示している. 第 3 に,NMF の解が未定乗数法による KKT の更新アルゴリズムによって収束することを証明している. ここでは NMF によって分解される二つの非負行列に直交性の性質を条件として与え, 評価関数をテーラー展開で 2 次近似した場合に, 導入した直交性の条件が 2 次導関数に現れるヘッセ行列を正定値行列とし, 収束が保証されることを説明している. この方法は,2 次導関数が明確に保証されれば, 一般的な NMF にも拡張可能である.

ここで述べられてきた NMF は, クラスタリングの潜在意味解析に関して適用された応用例がほとんどない. そのために本研究では第 4 に実データセットに NMF を適用し,SVD と比較することにより, 本研究で提案した NMF の有効性を検証している.

本論文は, 以下のように構成されている.

第 1 章では, クラスタリングに関する研究動向を述べ, 本研究の背景と動機を及び目的を述べている. 同時に, 本論文で用いている用語の説明も行っている.

第 2 章では, これまでにクラスタリング手法として使用されてきた SVD の理論的な取り扱いをまず説明している. ついで, 実際に SVD の能力を実データ (ロイター/コーパス 21578) に適用し, 従来使用されているクラスタリング評価基準に基づいてその結果を評価している. その後, 同義語/多義語問題に SVD を適用し,LSI の能力と限界を実験に基づいて示した上で,LSI の標準データセットに SVD を適用し,SVD の評価を行っている.

第 3 章では,NMF をクラスタリングに適用した時の LSI との関連について述べている. まず最初に,NMF がクラスタリングに関してどのように定式化されるかを理論的に導入し, 理論的に示される線形分離のクラスタリングが不可能な合成データセットに NMF を適用した場合の限界について説明している. ついで,NMF を実データ (ロイター/コーパス 21578) に適用し, その結果得られた LSI を SVD の結果と比較し,NMF の能力評価を実施している.

第 4 章では,NMF が厳密に収束する条件を求めている. この条件は, 分解された二つの非負行列に直交性を付与することであり, この直交性に基づいて NMF の評価関数の 2 次導関数に表れるヘッセ行列が正定値行列となり, よって収束が保証されることを理論的に示している. このことは, クラスタリングが安定して行われることを保証している.

第 5 章では, 本研究のまとめを行い, 結論としている.

これを要するに, 本研究はクラスタリングと LSI に NMF を適用できる要件と限界を理論的に明らかにし, また NMF 適用時のクラスタリング収束条件を明確にした上で LSI に適用し, その有用性を示したものでもので, ここで得られた知見は複雑系工学に貢献するところが大である. よって, 著者は北海道大学博士 (情報科学) の学位を授与される資格があるものと認める.