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

令和　５ 年　 ６ 月 １５ 日現在

機関番号：８２６２６
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ１２０３４

研究課題名（和文）汎化性能向上に資する大規模データセット構築のためのサンプル選択手法に関する研究

研究課題名（英文）Sample selection method for large scale datasets to improve robustness in recognition tasks

研究代表者

渡辺　顕司（Kenji, Watanabe）

国立研究開発法人産業技術総合研究所・情報・人間工学領域・研究員

研究者番号：５０５７１０６４

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：近年、AIという呼称の元、機械学習手法が様々な分野で利用されるようになり、汎化性の意味における分類・識別性能向上に大きな関心が集まるようになった。この問題を解決するには、爆発的に増加し続ける収集データ量の適切な削減にも資する、学習データの取捨選択、すなわちサンプル選択を行うことが有効な対応の一つである。
そこで本研究では、入力データの統計的性質を踏まえた再生成データの構築と、これら入力データと再生性データの誤差基準などから、大多数のデータが示す基準値から大きく外れる、すなわち例外となるデータを検出・削除するために、特に因子分解手法に着目した検討を実施した。

研究成果の学術的意義や社会的意義
本研究で着目した因子分解手法は古典的な多変量解析手法の一つであり、昨今の隆盛を極める深層学習手法を検討対象とすることをあえて避けたのは、一定の理論的基準と確信を持って、汎化性能の向上に臨めるからである。これは、現在の学術・商用を問わず一定の性能が望めるという一点のみで、「なぜ、所望の性能を達成できたのか？」という理論的解析が困難な深層学習手法を軽々と利用する風潮に一石を投じる意味で学術的・社会的意義のある研究であるものと考える。

研究成果の概要（英文）：Machine learning methods have been applied to solve recognition tasks in many academic and commercial fields, and the methods are demanded for the improvement of robustness to solve the tasks. Overcoming this problem, training datasets should be re-constructed only using favorable samples which are subtracted to outliers.
In this research, we studied a matrix factorization which is applied in a sample selection framework for large scale and unknown dataset. Because we may be able to subtract the outliers from the datasets by measuring distances and/or simple criteria in the feature space for the input (original) samples and obtained samples from the factorization.

研究分野：パターン認識　多変量解析

キーワード：多変量解析　因子分解

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年、AI という呼称の元、機械学習手法が商用利用も含めて様々な分野で利用されるようになった。このような機械学習手法の大衆利用の進展により、汎化性の意味における識別性能向上に大きな関心が集まるようになった。この問題を解決するには、爆発的に増加し続ける収集・収録データ量の適切な削減にも資する、学習データの取捨選択、すなわちサンプル選択を行うことが有効な対応の一つである。サンプル選択の問題を解決するための手法は、古くから様々な提案がなされており、高度・複雑化してきた経緯がある[1, 2]。しかし、数理的な知識に明るくない技術者等が容易に利用できるほど簡素・明解な手法であり、かつ汎化性能向上効果を実験的にも示した手法は多いとは言えず、サンプル選択手法の普及には至っていない。

２．研究の目的

本研究における当初の目的は、1. に記載した背景を踏まえ、簡素かつ有意なサンプル選択を実現する数理的手法の提案と、提案手法を適用して再構成したデータセットを用いることで汎化性の意味における分類・識別性能向上効果を実験的に示し、学習データの必要十分な収集・収録に資することであった。

３．研究の方法

機械学習手法を用いて識別問題を解く場合、その識別性能を向上させるためには、多量、かつ問題設定に沿った学習データ（すなわち「学習に適したデータ」）が必要不可欠となる。この「学習に適したデータ」の準備に関する問題は、学術的には古くから認識されている課題の一つであるサンプル選択問題として、様々な手法が提案され、高度化・複雑化してきた。

これら高度化・複雑化したサンプル選択手法は、数理的な知見を持った研究者・技術者が利用すれば、十全の性能を発揮する。しかし、数理的知見に明るくない手法利用者も多数存在する昨今、サンプル選択手法の有効活用には至っていない現状がある。

本研究では、サンプル選択問題を解くために、特徴空間上のサンプル分布に関する幾何的考察を行い、この考察から得られた知見を定式化することで、例外サンプルの選択基準を意味的にも解釈可能な因子分解手法を提案する。

４．研究成果

本研究の実施当初、まずは代表的な特異値問題を解くことで因子負荷量および因子を推定する因子分析、および既存の行列因子分解手法を実装・検証したところ、因子分析のような強力な正規直行制約を持つ手法以外、特に推定行列を疎にする制約条件を持つ手法では、推定行列の要素の値域が大きく偏り、その意味的解釈が困難となることが確認された。さらにこれらの行列因子分解手法は初期値依存性が非常に高く、好意的に解釈すれば問題設定ごとに初期値決定方式を任意に選択することができるものの、簡素かつ容易に利用できる手法を提案するという目的にはそぐわないことが確認できた。

また、実際のサンプル選択では、研究申請時に計画していたような、推定した因子を例外サンプルのインジケータ的に利用する方法では、簡素かつ確実なサンプル選択実現が極めて困難であることが判明した。

以上の知見を踏まえ、外部発表には至らなかったが、本研究により得られた知見を用いて検討した、初期値依存性が低く、かつクラスタリング的な運用も可能な行列因子分解手法を（１）項に記載する。また、特に多量なサンプルの蓄積がなされている時系列データセットへの適用を実現するための検討結果を（２）項に記載する。

（１）例外データを含む一般的なデータセットの解析

学習用データセットの構築を考えた場合、特に実計測データセットは一般的な特性とは異なるサンプルを多少なりとも含む事が多く、この特性の異なるサンプル（以降例外サンプルと呼称する）の存在が、機械学習手法の学習効率や汎化性能の向上に負の影響を与えることが多い。したがって、収集したデータセットにおける例外サンプルの検出・削除は、機械学習手法を用いた分類・識別問題を解く際の汎化性能向上に大きく貢献することとなる。さらに、多量のデータに対し、正しい教示ラベルを事前に付与することは多大なコストがかかる事から、教師なし機械学習手法の枠組みで、例外サンプルの検出を実現することが望ましい。

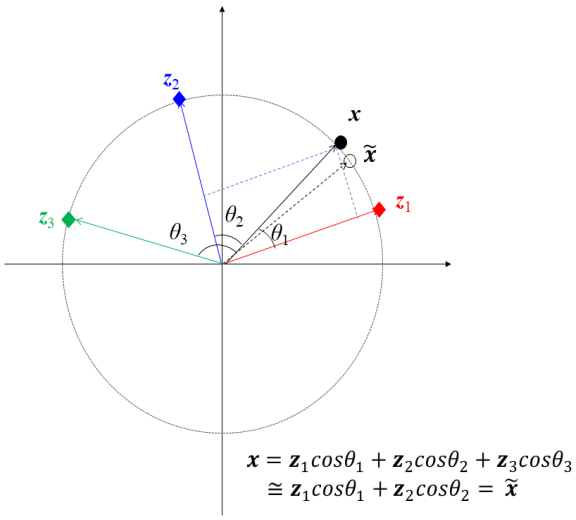

図 1 :検討手法における因子負荷量と因子の関係

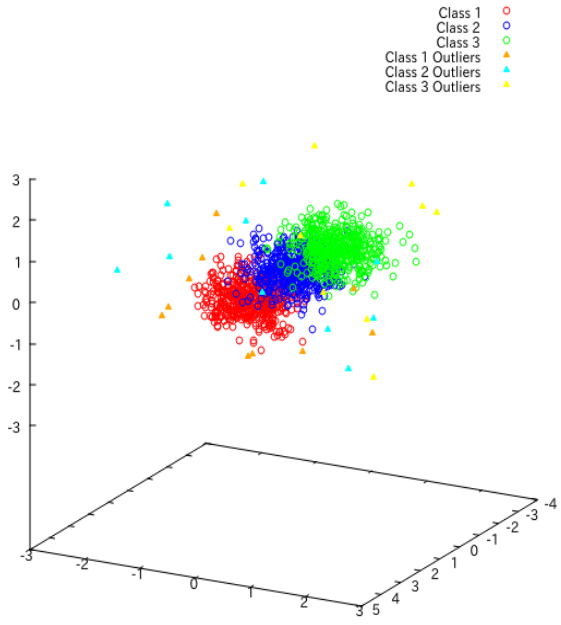

図 2: 生成データセットのサンプル分布

表 1: 生成データセットを用いた評価結果

Method	NMI
K-means (given rank: 3)	0.9205
Spherical K-means (given rank: 3)	0.5635
Semi-NMF (given rank: 3)	0.6865
Ours (given rank: 3, estimated rank: 3)	0.9081
Ours (given rank: 10, estimated rank: 3)	0.9081
Ours (given rank: 100, estimated rank: 3)	0.9081

そこで本研究では因子分解手法、およびクラスタリング手法に着目し、初期値依存性も低く、推定した行列の意味的解釈が容易な手法の検討を行った。この検討手法では、例外サンプルの影響を抑制した各クラスタの代表ベクトルを推定できる Spherical K-means [3] を参考に、この分類手法において各クラスタの平均ベクトルを因子負荷量とみなし、かつ各サンプルを推定クラスタに割り振るインジケータを 0 または 1 の二値から、0 以上 1 以下の連続値に緩和した表現に拡張することで因子とみなし、Semi-nonnegative matrix factorization (Semi-NMF) [4]の枠組みで定式化した手法を検討した（図 1）。さらに検討手法では、下式に示すとおり評価関数に因子に関する L0 正則化項を導入することで、クラスター数（因子分解手法としてはランク数）を自動推定できるものとした。

$$f = \|X - ZH\|_F^2 + \alpha\|H\|_{L0}$$

ここで $X = [\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_n]$は入力データセットであり、$Z = [\boldsymbol{z}_1 \ \cdots \ \boldsymbol{z}_k]$および $H = [\boldsymbol{h}_1 \ \cdots \ \boldsymbol{h}_n]$は、それぞれ因子負荷量および因子である。評価関数の右辺第 2 項の L0 正則化は、あるサンプルの因子ベクトル $\boldsymbol{h}$ の L0 ノルムの和とすることで、疎なランク数 r の自動推定を実現する[5]。

まず、検討手法のランク推定性能の検証を図 2 に示す例外サンプルを含む 3 クラス生成データセットに対して行った。このデータセットは混合ガウス分布で表現される。本評価では、全評価手法の初期値はランダムな値とした。結果は表 1 に示すとおりである。検討手法は、十分に大きなランクを初期値として与えれば、正しいランクを推定できる可能性が高く、かつ正規化相互情報量（Normalized Mutual Information）の意味で、K-means と同様な分類性能を示した。

さらに実データを用いた手法評価として、2011 年の第一および第二四半期における米国株式市場における 30 銘柄の株価推移データセット[5]の分類問題に適用し、分類結果の妥当性を検証した。当該データセットは、週毎の株価変動等に関する数値情報を銘柄毎に保有する。検討手法を含む因子分解手法は、各ランクの因子負荷量が各クラスの代表ベクトルと等価であると仮定し、各サンプル因子において最大値となるランクが帰属するクラスタであるとしたクラスタリングを行った。Semi-NMF の初期値は原著と同様に K-means のクラスタリング結果であり、他の手法は全てランダムな値とした。

表 2: 実データセットを用いた評価結果

Method	Cumulative percentage change (Cluster 1)	Cumulative percentage change (Cluster 2)
K-means (given rank: 2)	175.89 %	-36.71 %
Spherical K-means (given rank: 2)	110.88 %	28.30 %
Semi-NMF (given rank: 2)	175.89 %	-36.71 %
Ours (estimated rank: 2)	169.13 %	-29.94 %

表２は、対象データセットにおける、各サンプルの次週の株価変動率に対して、分類結果として得られた各クラスタのサンプルごとの変動率の和を累積値として示したものである。この累積値がクラスタ毎に正負それぞれに大きな値を示していれば、株価が上昇、および下落する銘柄を明確に分類できているであろうという仮定のもと評価を行った。表２の結果より、K-means clustering、Semi-NMF および検討手法では、株価が上昇、または下落するサンプルを分類することができたと考えられる。さらに検討手法では、初期クラスタ数を 100 と大きな間を与えても、最終的なクラスタ数は 2 となり、他の比較手法では任意にクラス数（本実験では２クラス）と設定しなければならない場合でも、自動で有意なランク数を推定することができた。

以上の結果から、検討手法は分類手法の意味で解釈しやすい因子負荷量、および因子推定することが可能であり、かつ一般的な因子分解手法では任意に決定ければならないランク数も自動推定することができる手法となる。しかしながら、推定した行列を用いたサンプル選択は、実現が困難であった。当該手法は過去に査読付き国際会議に投稿するも再録拒否となったままであるが、さらなる改善を行い、世に公表したいと考えている。

（２）例外データを含む多次元時系列情報の解析

本節では、多量・高詳細に計測・収集されるデータの中でも、特に多次元時系列信号の例外値検出の可否を判断するべく、因子分解手法を用いた因子負荷量、および因子の推定を脳波の計測データである EEG Eye State [6]に対して行った。このデータセットでは、瞼の開閉動作に関して二値のクラスラベルが付与されており、さらに人間を対象とした計測データであることから、例外値も含まれている事を期待して選出した。

適用する因子分解手法は、時系列変化を加味しない手法として一般的な因子分析を用いたほか、Recurrent neural networks (RNN) を用いた。RNN は厳密には因子分析手法ではないが、2 層の RNN において、推定する因子は入力サンプルの（非）線形変換で表現するという制約条件を与えた場合、中間層の出力は因子とみなすことが可能であり、出力層の変換係数行列は因子負荷量とみなすことができる。さらに、既存研究[7]において例外値検出にも用いられていることから本実験で適用した。

図３は全データに因子分析を適用した場合の因子負荷量、およびランク１からランク３までの因子を時系列でプロットしたものである。実験結果より、因子分析を用いると、計測開始から 9000 ms 程度の間までは、因子の絶対値が周辺時刻でピーク値となる時刻（サンプル）を検出することで、瞼の開閉が行われる時刻を推定することができそうではある。しかし、特に計測時刻後半では、このピークと開閉動作にズレが見られる。したがって、計測時刻後半に何らかの脳波自体、または計測環境など起因した例外値が存在すると予想できるが、因子分析では、例外値を明確に分離することは困難であった。

図４は、RNN を適用した場合の因子負荷量、および因子のプロット図である。RNN を適用した場合も、因子分析と同様の特徴次元（計測 ch）に対して正負大きな値となるが、ランクに対して、より密な因子負荷行列となる。時系列プロットした因子では、基本的に因子分解と同様の位置にピークが見られ、動作とピーク位置に同様の傾向が見られる。特にランク３およびランク４において 10000 ms 付近に大きなピークを持ち、このピーク以降は因子のピークパターンと瞼の開閉動作パターンに大きな誤差が生じることから、この 10000 ms 付近の入力サンプルが例外値となっている可能性がある。

以上の結果から、因子分析、および RNN を用いることで、多量・多次元な時系列データセットを識別に有意な特徴量に変換できる可能性のほか、RNN のような非線形変換も含む手法を用いることで、識別器の学習に不適なサンプの検出可能性が示唆された。

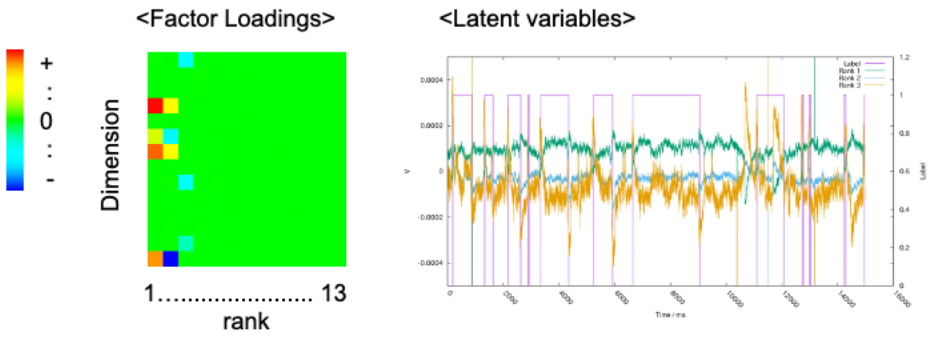

図 3: 因子分析の適用結果

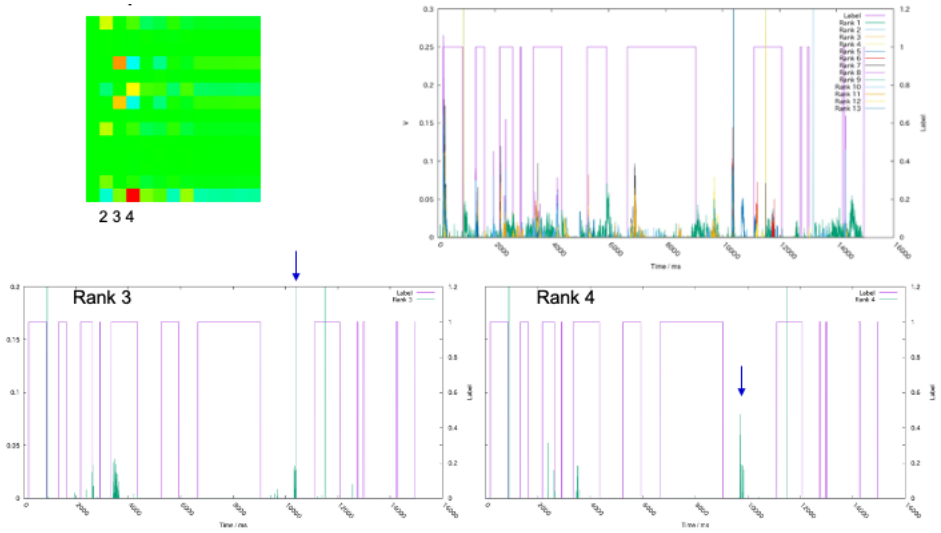

図 4: RNN の適用結果

<引用文献>

[1] K. Millard, and M. Richardson, *Remote Sens.,* 7, pp. 8489-8515, (2015).
[2] M. Gupta, et al., *IEEE T-KDE*, 26(9), pp. 2250-2267, (2014).
[3] I.S. Dhillon, and D.S. Modha, *Machine Learning*, 42(1), pp.143–175, (2001).
[4] C.H.Q. Ding, T. Li, and M.I. Jordan, *IEEE Trans. on PAMI*, 32(1), pp.45–55, (2010).
[5] D. Dua and C. Graff, UCI Machine Learning Repository. http://archive.ics.uci.edu/ml, (2019).
[6] O. Roesler, UCI Machine Learning Repository. https://doi.org/10.24432/C57G7J, (2013).
[7] G. Williams, et al., *IEEE Int. Conf. on Data Mining*, Japan, 2002, pp. 709-712

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 渡辺 顕司
２．発表標題 時系列信号解析のための因子分解法の検討
３．学会等名 福岡大学数理情報学セミナー
４．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関