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　元 年　 ６ 月 ２１ 日現在

機関番号：６２６０３
研究種目：基盤研究(B)（一般）
研究期間：2014～2018
課題番号：２６２８０００９

研究課題名（和文）内在的構造を持つ大規模高次元データ解析の理論と方法

研究課題名（英文）Thery and methods for high dimensional data analysis with internal structure

研究代表者
福水　健次（Fukumizu, Kenji）

統計数理研究所・数理・推論研究系・教授

研究者番号：６０３１１３６２

交付決定額（研究期間全体）：（直接経費）　　9,500,000円

研究成果の概要（和文）：高次元特有の問題としてハブ現象を研究し，その理論的な理解とハブ解消法の提案を行った．また，ツリーデータの解析や曲がった距離を用いたデータ解析など，ユークリッドベクトルとは異なるデータを扱う方法に関して研究を行った．さらに，高次元の密度関数を行う方法として，カーネル法に基づく指数分布族を提案し，その推定法を提案した．

研究成果の学術的意義や社会的意義
ビッグデータ時代になり高次元で複雑なデータを扱う必要性が高まったが，そのようなデータの性質や解析法の関して，理論的な知見や有効な方法が成果として得られた．今後，さまざまな分野に現れる高次元データを扱う際にこれらの成果が貢献できると考える．

研究成果の概要（英文）：The hub phenomenon has been analysed as an example of special properties of high-dimensional data, and a method of resolving hubs has been proposed. Methods of data analysis have been proposed for non-Euclidean data such as trees and points with skewed distance measures. Additionally, for a new method of density estimation of high-dimensional data, a kernel methods for constructing infinite dimensional exponential families has been proposed, and its estmation has been discussed.

研究分野：機械学習

キーワード：統計数学　データ解析　高次元　アルゴリズム

１．研究開始当初の背景

情報化社会の高度化や計測・観測技術の発達により多くの分野で大規模高次元データが大量蓄積されている現在、データから有益な知識を抽出するための基盤技術を確立することは、社会の活性化とさらなる発展のために統計科学に与えられた必須かつ緊喫の課題である。大規模高次元データには様々な数理的構造が内在すると考えられている。例えば、現実の高次元データは低次元多様体上に分布することが多い。また最近、高次元データでは、特異的に多数のデータ点の K 近傍に属するようなデータが存在しやすいという「ハブ現象」なども知られている。このようなデータからより有効に知識を獲得するためには、大規模高次元データのもつ非線形構造、多様体構造、ハブ構造などの数理的性質を解明し、それに根差したデータ解析の方法を体系化するアプローチが重要となる。

２．研究の目的

本研究では、大規模高次元データ解析のために、データが非線形構造、多様体構造、ハブ構造などの内在的構造を持つことを仮定して、その構造を反映したデータ解析の数理基盤を構築することを目的とする。具体的には以下の 3 課題を研究する。

【課題 1】内在的構造を持つ大規模高次元データの数理

多様体構造・ハブ構造などを仮定した場合の、大規模高次元データの統計的性質を明らかにし、データ解析手法の提案と理論解析のための数理的基礎付けを行う。

【課題 2】高次元データに対するカーネル法の理論と方法

次元が無限大になる場合のカーネル法の挙動に関する理論解析と、それに基づくカーネル選択などの方法を確立する。

【課題 3】超効率的アルゴリズムの理論と応用

データの持つ内在的構造を利用し、サブリニアなど超効率的なアルゴリズムの開発を行う。

３．研究の方法

【課題 1】内在的構造を持つ大規模高次元データの数理

内在的構造として低次元多様体構造とハブ構造に注目し、以下の課題を研究する。

- 潜在的多様体構造を有する場合のランダム行列理論：　データが低次元多様体構造を持つと仮定した場合に、次元が無限大の極限において、データの分散共分散行列やグラム行列が示す漸近的挙動を解明する。
- ハブ構造の出現原理の数理的解明とハブ解消法：　大規模データに対しては、効率的計算が可能な K 近傍に基づくデータ解析手法が適する場合が多いが、ハブの存在はこのような解析手法の適用を困難にする。そこで、過去の研究で得られた基礎的結果をさらに拡張することにより、ハブ現象の生じる数理的な原理を解明し、それを用いてハブ解消の方法を研究する。

【課題 2】高次元データに対するカーネル法の理論と方法

次元に依存したカーネルの理論解析

- ガウスカーネルのバンド幅パラメータ選択など、次元に依存するカーネルを用いた際の、高次元データに対するカーネル法の挙動を解明する。文献１の解析を発展させることによって研究を進める。
- 高次元データに対するカーネルアルゴリズム
  前項の結果に基づいて、カーネル法によるノンパラメトリック検定の検出力の解析を行う。

【課題 3】超効率的アルゴリズムの理論と応用

現在の大規模高次元データ解析においては、データ数や次元に対して線形の演算量よりも少ないサブリニアなアルゴリズムが必要となってきている。データの潜在的な低次元構造を利用した超効率的アルゴリズムを以下のように研究する。

- 高速なオンライン最適化アルゴリズムによるデータ解析技術の系統的研究
  双対座標降下法と交互方向乗数法の利点を組み合わせ、並列化可能で反復回数がサンプル数に依存しないオンライン学習アルゴリズムを開発する。
- グラム行列の効率的計算法
  カーネル法の効率的計算では、グラム行列を低ランク行列で近似する近似計算がよく用いられるが、これは行列サイズに対して線形の演算量のアルゴリズムである。これをデータの持つ低次元構造を用いてサブリニアに改良する方法を検討する。

４．研究成果

（１）ハブ現象の解析

遺伝研の原，鈴木氏らとともに，高次元データのハブ現象に関して研究を行い，ハブ現象が低次元でも起こる現象であることを発見し，クラスタリングと中心化によってそれを解消する方法を提案した．この結果を，人工知能分野のトップ国際会議である AAAI 2015 において発表した（論文１）.

（２）カーネル法による確率密度推定
高次元においては困難である確率密度関数の推定に関して，カーネル法によるノンパラメトリックな指数分布族を定義し，スコアマッチングによって正規化定数以外の関数部分を推定する方法を提案し，その数理的解析を行った．実験の結果，カーネル密度関数に比べて，高次元データに対して高精度な推定が可能であることがわかった．研究の結果をまとめた論文を機械学習分野のトップ論文誌である Journal of Machine Learning Research に発表した（文献 2）

（３）異なるドメイン間の共通構造の抽出法
２つの異なるドメインのデータにクラスタ構造があることを仮定して、カーネル法を用いて、これらのクラスタをマッチングさせる方法を研究した。特に、データ間の類似度などを全く与えられていない教師なしの状況をターゲットとした。研究の結果、カーネル平均とカーネル化ソーティングを組み合わせたグループ・カーネル化ソーティング法を提案した。これを Wikipedia の対応する項目グループの多言語間でのマッチングに適用したところ、既存の方法よりも良好な結果が得られることが分かった。この成果は、国際雑誌 Data Mining and Knowledge Discovery に掲載された（文献 3）。

（４）ツリーデータのクラスタリング解析
系統樹を表現するツリーデータ全体の空間の性質を利用したデータ解析手法に関して研究を行った．その結果，Billera-Holmes-Vogtmann によるツリーデータのなす距離空間の性質を利用したクラスタリング手法を確立し，系統樹解析に応用した．具体的には，遺伝子ごとに作成された多くの系統樹データを Billera-Holmes-Vogtmann 距離行列に基づいてクラスタリングし，遺伝子クラスタごとに異なる進化系統樹が構築される場合があることを確認し，実データに応用した．この結果を論文としてまとめ英文論文誌 Annals of Operations Research に発表した（文献４）.

（５）曲がった空間によるデータ解析
データが存在する空間として、平たんではない曲がった空間を表す CAT(k)空間を想定し、その距離を用いた距離行列に基づくデータ解析法を開発した。これにより、データ分布の異なるクラスタ構造などを抽出することが可能となった。この結果をまとめた論文が、一流国際誌 Statistics and Computing で出版された（文献 5）。

５．主な発表論文等

〔雑誌論文〕（計２４件）

[1] Kobayashi, K. and Wynn, H P. Empirical geodesic graphs and CAT(k) metrics for data analysis. Statistics and Computing. (2019) pp.1-18. 10.1007/s11222-019-09855-3. （査読有り）

[2] Yoshida, R., Fukumizu, K. & Vogiatzis, C. Ann Oper Res. Multilocus phylogenetic analysis with gene tree clustering (2019) pp.276-293. https://doi.org/10.1007/s10479-017-2456-9 （査読有り）

[3] Sriperumbudur, B., Fukumizu, K., Gretton, A., Hyvarinen, A., Kumar;, R. Density Estimation in Infinite Dimensional Exponential Families. Journal of Machine Learning Research 18(57):1−59, 2017. （査読有り）

[4] Iwata, T., M. Kanagawa, T. Hirao, K. Fukumizu. Unsupervised group matching with application to cross-lingual topic matching without alignment information. Data Mining and Knowledge Discovery (2016) pp 350–370. DOI: 10.1007/s10618-016-0470-1 （査読有り）

[5] Hara, K., I. Suzuki, M. Shimbo, and K. Kobayashi, K. Fukumizu, and M. Radomanovic. (2015) Localized Centering: Reducing Hubness in Large-Sample Data. Proc. AAAI-2015, pp.1659-1665 （査読有り）

〔学会発表〕（計１８件）

1) 福水健次 (2018) 教師ありクラスタマッチングとその隕石-小惑星分類体系への応用. 日本分類学会第 37 回大会（招待講演）
2) 鈴木大慈 (2018) 機械学習における構造を利用した確率的最適化技法. 2018 年電子情報通信学会基礎・境界ソサイエティ大会大会（招待講演）
3) Fukumizu, K. (2017) Local minima and saddle points in hierarchical structure of neural networks. Deep Learning: Theory, Algorithms, and Applications（招待講演）
4) 福水健次 (2017) データ構造を解明する機械学習アプローチ --グラフ構造、幾何構造--第 3 回 WIRP ワークショップ（早稲田大学，招待講演）
5) 鈴木大慈 (2014) スパース推定概観：モデル・理論・応用. 統計関連学会連合大会（招待

講演）

〔図書〕（計　0件）

〔産業財産権〕
○出願状況（計　0件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等
http://www.ism.ac.jp/~fukumizu/

６．研究組織

(1)研究分担者
研究分担者氏名： 鈴木大慈
ローマ字氏名： Taiji Suzuki
所属研究機関名： 東京大学
部局名：大学院情報理工学研究科
職名：准教授
研究者番号（8桁）：60551372

研究分担者氏名： 小林景
ローマ字氏名： Kei Kobayashi
所属研究機関名： 慶応義塾大学
部局名：理工学部
職名：准教授
研究者番号（8桁）：90465922

(2)研究協力者
研究協力者氏名：
ローマ字氏名：