様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 25 年　4 月　2 日現在

機関番号：14401
研究種目：基盤研究(B)
研究期間：2010～2012
課題番号：22300054
研究課題名（和文）　超高次元データに関する統計的推定原理確立と大規模データマイニングへの適用
研究課題名（英文）Establishment of Statistical Estimation Principle for Super High Dimensional Data and Its Application to Large Scale Data Mining
研究代表者
鷲尾　隆（WASHIO TAKASHI）
大阪大学・産業科学研究所・教授
研究者番号：00192815

研究成果の概要（和文）：次元の呪い効果を分析し、超高次元データが中心から[r,r+ε]の距離に分布する「球面集中効果」、超高次元空間の局所に確率が集中する「確率密度集中効果」、超高次元データが広大な体積内に分布する「スパース化効果」等を特徴付けし、前者２効果を打ち消す人工的歪みをデータ・状態分布に与える高精度、ロバストな推定法を提案した。

研究成果の概要（英文）：Upon analysis of dimensionality curse, we characterized “hyper-sphere concentration effect”, “probability concentration effect” and “sparsity effect” of super high dimensional data, and proposed an accurate and robust estimation method against the former two effects.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	4,500,000	1,350,000	5,850,000
２０１１年度	5,000,000	1,500,000	6,500,000
２０１２年度	3,800,000	1,140,000	4,940,000
総　計	13,300,000	3,990,000	17,290,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：知能発見，データマイニング，統計的推定，超高次元データ，ビッグデータ

１．研究開始当初の背景

近年のネットワークセンシングや生物情報科学、環境科学の実験測定技術、リモートセンシングの進歩を通じ、超高次元ベクトルで表現されるデータが大量蓄積される時代となった。一方、多くのデータマイニングでは、解析目的に応じた諸条件でデータを場合分けした沢山のデータセグメントを作り、セグメント毎に傾向分析を行う。このため、各セグメント内の超高次元少数ベクトルデータから目的情報を高精度、ロバストに推定し、かつ全セグメントを高速処理することが重要となる。しかし、数十次元以上では後述する「次元の呪い」効果で推定精度やロバスト性が大きく損なわれ、かつこれを避けるには $O(r^{d/4+1})$ (d は次元数,r>>1 は定数)の大量データが必要となり、データ収集や計算のコストの面でも回避困難であることが分かっている。

この問題に対し、特異値分解などによる次元圧縮を用いて問題軽減を図る手法が研究されてきた。しかし、これは情報削減による近似推定であり、また次元の意味が不明確となるため結果解釈が難しい。一方、ある種のデータ分布については、解析パラメータを適応させる直接推定手法が提案されているが、一般性のある研究は世界的に皆無である。これに対し、本研究では超高次元データに関する一般的非圧縮直接推定を目指すこととした。

２．研究の目的

本研究では、「次元の呪い」の詳細機構を明らかにし、超高次元データについて高精度、ロバスト、低計算量な統計的推定原理の確立を目指した。更に、それを海洋温度分布推定など社会的に重要な理学の大規模データマイニング問題で検証した。

より具体的には、３年の研究期間内で以下の項目を遂行した。

(1) 次元の呪いの詳細機構の探求

各種次元の呪い現象の洗い出し、相互関係の整理、類型化を行い、各々の発生機構の解明と数理的性質の特徴づけを行った。

(2) 次元の呪い克服原理の探求

次元の呪いのうち、「球面集中効果」及び「確率密度集中効果」を克服する一般的統計的推定原理の確立を目指した。特に、精度、ロバスト性、効率性で超高次元データに適用可能な世界的ブレークスルーを目指した。

(3) 大規模データマイニング問題への適用

海洋温度分布推定など社会環境科学上重要な大規模問題に適用し性能検証を行った。

３．研究の方法

(1) 第１に次元の呪いの詳細機構の探求に取り組んだ。

これは、更に

① 各次元の呪い現象を洗い出し、それらの相互関係の整理、類型化を行う研究

② 次元の呪いの根本的な各発生機構の解明と数理的特徴づけの研究

に分けられる。①について、「球面集中効果」及び「確率密度集中効果」以外にも、d→∞での確率密度無限大の特異領域の出現や全てのデータ間距離がほぼ等しくなり確率密度推定が困難になるなど、他にも統計的推定に影響する次元の呪いが存在する。また幾何学分野でも、d→∞で任意の半径を持つ超球の体積や表面積がゼロに収斂する現象や超立方体の体積や表面積にも同様の収斂や発散が見られる現象が知られている。これらの中には相互に密接に関係しているないしは同根の現象も含まれる。一見、バラバラに考察されている現象の知見を収集し、それらが別個の現象なのか、同じ原因の異なる形での表れだけなのかを調べ、更にそれら現象を原因に基づいて類型化する作業を実施した。また、②を一般的な枠組みで行うには、任意の確率密度関数に関し厳密な解析を行う必要がある。このためには、超高次元ないしは次元 d→∞の極限における任意の確率密度関数とそれに従うデータ分布の数理的な挙動を調べる方法論を確立しなければならない。確率密度関数をパラメトリックなものに限定するならば、この解析は比較的容易であるが、任意の確率密度関数に関する解析を行うために、確率密度関数をノンパラメトリックに扱い、その漸近性を確認しつつ d→∞での関数挙動を解析した。

(2) 次に次元の呪い克服原理の探求に取り組んだ。

これは、更に

①次元の呪い１，２を克服する一般的な統計的推定原理の確立

②他の次元の呪いを克服する一般的な統計的推定原理の確立

に分けられる。①については、前記(1)の①及び②で解明された「球面集中効果」及び「確率密度集中効果」の発生機構や数理的特徴に基づいて、統計的推定の何れの過程においてこれらの効果が影響するか検討し、その各過程において次元の呪いの効果を回避ないしは軽減する原理を考案し、それを体現する計算手法を構築した。②では、「球面集中効果」及び「確率密度集中効果」以外の発生機構や数理的特徴が解明された呪いについて、高精度、ロバスト、高効率な統計的推定を可能にする原理を確立を目指した。これも上記と①同様に、各呪いの発生機構や数理的特徴が影響を及ぼす統計的推定過程について、次元の呪いの効果を回避ないしは軽減する原理を考案し、それを体現する計算手法を構築した。

４．研究成果

高次元データ分布からある種の条件下で平均値や確率密度などの統計量を推定する際には、前述の３種類の次元の呪い現象が相まって、深刻な推定精度低下が起こる。たとえば、高次元ベクトルデータの確率密度 $p(X)$ とその事例 X の下でのある他の情報 Y の確からしさ $p(Y|X)$ から、ベイズ推定によって逆に Y の下で事例 X が得られる確率 $p(X|Y) \propto p(Y|X)p(X)$ を計算し、Y の下での X の期待値を推定する場合を考える。一般に $p(X)$ は不明でデータのみが知られていることが多く、そのデータ分布で $p(X)$ を代用して各事例ベクトル Xi に関する $p(Y|Xi)$ から Xi の重み $p(Xi|Y) \propto p(Y|Xi) / \Sigma p(Y|Xi)$ を計算し、$E_Y(X) = \Sigma Xip(Xi|Y)$ を得る。各 Xi が高次元事例ベクトルである場合には、その分布はスパースでかつある球面上に集中している。一方，$p(Y|X)$ は確率密度の集中によってその極大点以外では極めてゼロに近い。したがって、ほとんどの $p(Y|Xi)$ はゼロに近く、重み $p(Xi|Y)$ に $p(Y|X)$ の極大点などの形状が反映されず、Y の下での X の期待値 $E_Y(X)$ の推定精度が著しく低下する。この効果は、観測ベクトル Y の系列から逐次的に観測対象の高次元状態ベクトル X をベイス推定する粒子フィルタなどの計算において、深刻な精度低下をもたらすことが分かった。

このような次元の呪い現象を解消ないし軽減するために，データマイニングや統計的

推定において用いられる代表的方法に次元圧縮が挙げられる。しかし、次元圧縮法では高次元データ空間内で事例が占める部分空間の次元である本質次元と、対象とする特徴の保存に求める精度によって、圧縮可能な次元が決まる。したがって、本質次元が高いデータや高い特徴保存精度が求められる応用では、十分に低い次元にまで圧縮できず次元の呪いを解消することが困難であることが分かった。

そこで、高次元事例ベクトル Xi の下でのある情報 Y の確からしさ p(Y|Xi)から、ベイズ推定によって Y の下での X の期待値 $E_Y(X)$ を推定する場合に、次元の呪いを軽減する方法としてプロポーザル分布を用いる IEP(intensive and extensive proposal)手法を提案した。この手法では図１に示されるように、実データから大まかに予想されるベイス推定分布 p(X|Y)の中心付近と裾野に人工データを付加して、実データの球面集中現象とスパース化現象を緩和した新たなプロポーザル分布 q(X)に従うデータを生成する。ただし、新たなデータに元データ分布を反映させるため、新たなデータの各事例を w(Xi)=p(Xi)/q(XI)によって重み付ける。これにより、Y の下での事例 Xi の確率を p(Xi|Y)=w(Xi)p(Y|Xi)/ Σ w(Xi)p(Y|Xi)によって重み付き推定し、それを基に $E_Y(X)$= Σ Xip(Xi|Y)を得る。IEP 手法を粒子フィルタに適用することで、本質次元が非常に高いカオスダイナミクスを有する系の観測ベクトル Y の系列から、系の高次元状態ベクトル を高精度にベイス推定可能であることを示した.

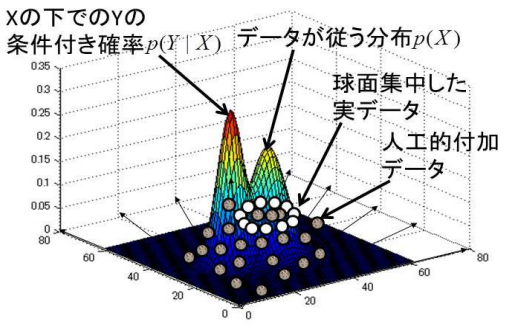

図１　高次元ベクトルデータに関するベイズ推定。データの球面集中現象、スパース化現象及び p(Y|X)の確率密度集中現象により，p(Y|X)の形状が推定結果に反映されないため、予想されるベイス推定分布 p(X|Y)の中心付近と裾野に人工的にデータを付加し、プロポーザル分布 q(X)を生成する。

実データに関する検証結果の詳細を図２及び表１に示す。米国の National Oceanographic Data Center において公開されている図示の領域の巨視的な海洋波動に関する人工衛星リモートセンシング時系列データを対象とし、波高の高精度予測推定を行った。これは海面の２次元的広がりを持つセンシングデータであり、観測ベクトル Y(t)は 200 次元、波動の状態ベクトルは 400 次元である。２次元の広がりを持つ海洋波動は、システム方程式として Kadomtsev-Petviashvili equation に従うことが知られている。また、観測方程式はリモートセンシング時系列データのサンプリング過程から導かれる。これに標準的 PF と上記 IEP を適用したマージナル推定値の精度比較結果が表１である。提案手法は高次元データに対して標準的 PF よりも遥かに高い精度を達成している。また、両者の計算時間にはそれほど大きな違いはない。

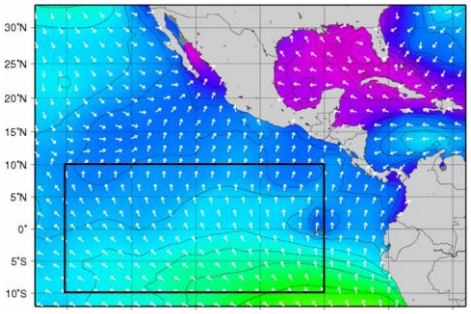

図２　海洋波動のセンシング範囲

表１　波高推定誤差の標準偏差(feet)

粒子数	PF	IEPF
256	4.68	1.53
1024	4.56	1.44

以上より、各種次元の呪い現象の洗い出し、相互関係の整理、類型化と、各々の発生機構の解明を行うことができた。また、次元の呪いのうち、「球面集中効果」及び「確率密度集中効果」を克服する一般的統計的推定原理を確立した。特に、精度、ロバスト性、効率性で超高次元データに適用可能な１つのブレークスルーを実現し、それを海洋温度分布推定などの大規模問題に適用し、その優れた性能を確認した。

５．主な発表論文等

〔雑誌論文〕（計５件）

① H.Kuwajima, T.Washio, L.Ee-Peng, Fast and accurate PSD matrix estimation by row reduction; IEICE TRANSACTIONS on Information and Systems, E95-D (11) [査読有], (2012)

② S.Hara, Y.Kawahara, T.Washio, P.von Bunau, T.Tokunaga and K.Yumoto, Separation of stationary and non-stationary sources with a generalized eigenvalue problem; Neural Networks, 33, 7-20 (2012) [査読有]

③ Y.Sogawa, S.Shimizu, T.Shimamura, A.Hyvarinen, T.Washio and S.Imoto, Estimating exogenous variables in data with more variables than observations; Neural Networks, 24(8): 875-880 (2011) [査読有]

④ V.P.Nguyen, T.Washio and T.Higuchi, A new particle filter for high-dimensional state-space models based on intensive and extensive proposal distribution; Int. J. Knowledge Eng. and Soft Data Paradigms, 2(4), 284-311 (2010) [査読有]

〔学会発表〕(計６件)

① T.Ueno, K.Hayashi, T.Washio and Y.Kawahara, Weighted likelihood policy search with model selection; Advances in Neural Information, Processing Systems 25, Proc. Neural Information Processing Systems Conference 2012 (NIPS2012), Dec., 5, 2012, (Lake Tahoe, USA) [査読有]

② T.Washio, A new approach to bayesian estimation over the curse of dimensionality; Thirty-first SGAI International Conference on Artificial Intelligence (AI2011), Workshop on Machine Learning and Intelligent Autonomous Systems, Dec., 12, 2011 [Invited,査読無]

③ K.M.Ting, T.Washio, J.Wells and T.Liu, Density estimation based on mass; Proc. 11th IEEE Int. Conf. on Data Mining (ICDM2011), 715-724, Dec., 11, 2011, (Vancuver, Canada) [査読有]

④ S.Hara and T.Washio, Common Substructure Learning of Multiple Graphical Gaussian Models; Proc. European Conf. on Machine Learning and Principle and Practice of Knowledge Discovery in Databases 2011 (ECML/PKDD2011), Lecture Notes in Computer Science, 2011, 6912, 1-16, Sep., 7, 2011, (Athens, Greece) [査読有]

⑤ T.Inazumi, T.Washio, S.Shimizu, J.Suzuki, A.Yamamoto and Y.Kawahara, Discovering causal structures in binary exclusive-or skew acyclic models; Proc. 27th Conf. on Uncertainty in Artificial Intelligence (UAI2011), 373-382, July, 16, 2011, (Barcelona, Spain) [査読有]

⑥ S.Hara, Y.Kawahara, T.Washio and P.Bunau, Stationary subspace analysis as a generalized eigenvalue problem ; Lecture Note in Computer Science, 6443, Proc. of 17th Int. Conf. on Neural Information Processing (ICONIP2010), 422-429, Nov., 23, 2010, (Sydney, Australia) [査読有]

６．研究組織

(1)研究代表者

鷲尾　隆（WASHIO TAKASHI）
大阪大学・産業科学研究所・教授
研究者番号：00192815

(2)研究分担者

該当なし

(3)連携研究者

樋口　知之（HIGUCHI TOMOYUKI）
統計数理研究所・モデリング研究系・教授
研究者番号：70202273

猪口　明博（INOKUCHI AKIHIRO）
大阪大学・産業科学研究所・助教
研究者番号：70452456

河原　吉伸（KAWAHARA YOSHINOBU）
大阪大学・産業科学研究所・助教
研究者番号：00514796

清水　昌平（SHIMIZU SHOHEI）
大阪大学・産業科学研究所・助教
研究者番号：10509871

中野　慎也（NAKANO SHINYA）
統計数理研究所・モデリング研究系・助教
研究者番号：40378576