様式 C-7-2

自己評価報告書

平成２２年３月３１日現在

研究種目：基盤研究（Ｂ）
研究期間：2007～2010
課題番号：19300097
研究課題名（和文）　高次元大規模データのモデル化を助けるデータヴィジュアリゼーションの理論と実際

研究課題名（英文）Theory and Practice of Data Visualization for Modeling Complex Large Scale Data

研究代表者
柴田　里程（SHIBATA　RITEI）
慶應義塾大学・理工学部・教授
研究者番号：60089828

研究代表者の専門分野：総合領域
科研費の分科・細目：統計科学
キーワード：データヴィジュアリゼーション，Textile Plot，大規模複雑データ，モデリング

１．研究計画の概要

大規模で高次元なデータのモデル化を助けるための効果的なデータヴィジュアリゼーション環境の確立を目指して，Textile Plot を中心に研究を進める．研究は，環境の実現から始め，それを，ファイナンス分野，バイオインフォマティックス分野，社会調査分野，環境分野など，大規模で高次元なデータを日常的に扱わなければならない分野での実際データからのモデル化の実践を通して，本環境の有効性あるいは問題点を明らかにし，改善を図るだけでなく，それを裏付ける理論の構築を図る．本研究は扱う対象の点ではデータマイニングと類似しているが，データマイニングがあえて避けているデータの全体像をモデル化することを目標としている点が大きく異なる.また，実装に際しても，データの持つ属性の最大限の利用と，データ変容のシステマティックな記述を柱としており，これまでの DandD プロジェクトの成果も十分に活用した統合環境の構築を目的とする．

２．研究の進捗状況

(1) Textile Plot に関しては，新たな発想により実質的に次元数（変量数）の制約から自由なアルゴリズムの開発に成功した．

(2) ゲノム解析への適用．数十万次元に及ぶヒトゲノム系列のヴィジュアリゼーション手法として，Textile Plot が有効であることを，これまで伝統的に用いられてきた LD ディスプレイと比較することにより確かめた．

(3) 金融データベースシステムへの展開．ヘッジファンドの収益率を主な対象として，時間を変量に取る Textile Plot と時間ごとに変量を導入する Textile Plot の併用により，その姿を総合的に理解できるシステムを ICS FinAnalyzer として構築した．

(4) 海洋生物の豊富さ検証への応用．前年に引き続き，オーストラリアの CSIRO(Commonwealth Science and Industry Research Organisation)との共同プロジェクトとして実施した．これまで開発した様々なデータヴィジュアリゼーション手法を用いることにより，海底生物のドレッジデータについては，種ごとの，個体数については Thomas 分布が，重量に関しては確率微分方程式の定常解として現れるガンマ分布が適切であることが判明し，その確証実験を行った．

３．現在までの達成度

①当初の計画以上に進展している．

(理由)

Textile Plot について次元数の制約から自由なアルゴリズムの開発に成功したことにより，適用範囲が飛躍的に増大した．また，さまざまな分野でのデータヴィジュアリゼーションの実践により，これまで見過ごされてきた有効なモデル構築に欠かせない共通な課題も明らかになってきた．たとえば，モデルの適合度を確認するためのヴィジュアリゼーション手法と伝統的な検定統計量のギャップなどがあげられる．

４．今後の研究の推進方策

(1) これまでの成果を統合し，大規模複雑データからのモデル化支援環境としてまとめる．

(2) これまでの成果が，電力取引価格や天候ディリバティブなど，さらに複雑度の高いデータに関してもどれだけ適用可能かどうか検証する．

５． 代表的な研究成果

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計５件）

① R.Miura, Y. Aoki, D.Yokouchi, A Note on Statistical Models for Individual Hedge Fund Returns, Mathematical Methods of Operations Research, 69, 553-557, 2009, 査読有

② N.Kumasaka and R. Shibata, High Dimensional Data Visualisation: the Textile Plot, Computational Statistics and Data Analysis, 52, 3636-3644,2008,査読有

③ H.Shimadzu, R.Shibata and Y.Ohgi, Modelling Swimmer’Speeds Over the Course of a Race, J.Biomechanics,41, 549-555, 2008,査読有

〔学会発表〕（計９件）

④ M. Naka, R. Shibata, Goodness of fit of Gamma distribution to sea fauna weights, *The International Biometric Society Australasian Region (New Zealand),* 2009年11月30日, Lake Taupo, New Zealand

⑤ 柴田里程，菅谷勇樹，独立異分布標本のPPプロットによる分布適合性の検証，*統計関連学会連合大会,* 2009年９月８日，同志社大学