様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成２５年　５月２５日現在

機関番号：３２６８９
研究種目：挑戦的萌芽研究
研究期間：2011 ～ 2012
課題番号：23656072
研究課題名（和文）　情報量に基づく重み付きデータ縮約

研究課題名（英文）　Weighted data contraction based on information content

研究代表者
村田　昇（MURATA　NOBORU）
早稲田大学・理工学術院・教授
研究者番号：60242038

研究成果の概要（和文）：古典的な k-近傍法によるエントロピーのノンパラメトリック推定量を一般化し，重み付きデータの情報量の推定を効率良く行う方法を提案した．またこのデータ間の距離にもとづく推定法を広いクラスのデータに対して適用可能とするために，適切な距離を学習する方法を多重カーネル学習と JIT モデリングの枠組に基づいてを提案した．これらの手法をクラスタリング問題や集団学習に応用し，その有効性を確認した．

研究成果の概要（英文）：Generalizing the classical k-nearest neighbor method for entropy estimation, an computationally efficient method for estimating information contents of weighted data is proposed. For utilizing our distance-based method to various kinds of data sets, distance metric learning methods are considered in the framework of multiple kernel learning and just-in-time modeling. Validity of those proposed methods are confirmed by clustering problems and ensemble learning of real-world data.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	1,800,000	540,000	2,340,000

研究分野：数理工学
科研費の分科・細目：工学基礎
キーワード：データ縮約，情報量，多重カーネル学習，距離学習

１．研究開始当初の背景

計算機環境の急速な発展にともない，大規模なデータベースを用いた情報処理が多くの場面で用いられるようになっている．統計ではノンパラメトリック推定，制御では JIT(Just In Time)モデリングと呼ばれる手法がこれに相当し，強い非線形性のために厳密なモデル化が難しい問題において一定の成果を上げている．例えば，鉄鋼業における精錬プロセスでは，極めて非線形性の強いプロセスの特性を詳細にモデル化することが難しいため，過去の操業において得られた大量のデータの中から現在のプロセス状態に近いデータを探し出し，複数の近傍データに基づいた局所線形モデルによる制御が行われている．また，太陽光発電システムにおいては，雲などの影響により光量が変化し，それに伴ない出力側の電圧が変動する．これを一定に保つために蓄電池への充放電制御が行われるが，過去の電圧変動を計測したデータの中から現在の変動に近いデータを探し出し，これらの近傍データを用いて太陽電池パネルの電圧の変化を予測あるいは分類して制御に利用する方法が提案されている．

こうしたモデル化においては，日々蓄積されるデータによって予測精度は向上するものの，データの増加に伴なう近傍検索のための計算量の急速な増大はオンラインで高速に制御・予測を行う際の問題となっている．また，対象とするシステムや環境の経年変化

により，蓄積されたデータの全てが同じように信頼できるわけではないというデータの非一様性の問題もあり，蓄積されたデータを目的に応じて整理・縮約することが重要な課題となっている．

２．研究の目的

データベースに蓄えられた膨大な原データを縮約し，予測や判別を目的としたモデル構築に適切な少数のデータを収集する手法はこれまでにも提案されてきたが，収集された代表データは一般に原データの従う確率分布と異なる確率分布に従う．逆に原データの確率分布と同じ分布に従うように代表データを収集した場合，予測や判別のために重要なデータが取り零される確率が高くなるといった問題がある．

本研究課題では少数のデータによりデータベースに蓄えられた膨大な原データを少数のデータで適切に代表し縮約する問題を取り扱う．特に，データの各点の情報量を直接データ集合から推定する方法を提案し，これに基づいて代表データの各点に適切な重み付けを与えることで原データの確率構造を保持しつつ，モデル化のために重要な少数データを収集する方法論の確立に取り組む．

３．研究の方法

(1) 情報量の推定のためには，密度関数の推定が不可欠であり，データに基づく推定方法としては，これまでパラメトリックな確率モデルに基づく方法と，Parzen 窓によるカーネル密度推定や k-近傍法に基づくノンパラメトリックな方法が提案され用いられてきた．多次元で複雑なプロファイルを持つ密度関数の推定においては，パラメトリックな方法はそのモデルの選択に難点があると考えられる．特に最近では多自由度の混合正規モデルなどが好んで用いられるが，多次元の場合には多数の局所解の存在により解が不安定となることや，混合数の決定のために大規模な計算が必要となるなど，工学的な利用を考えた場合には多くの問題を含んでいる．一方ノンパラメトリック推定の代表であるカーネル密度推定は低次元の問題においては非常に強力であるが，多次元の問題においては所謂「次元の呪い」のため窓関数の台に含まれるデータ数が少なくなり，安定な推定が難しいことが知られている．安全な台の大きさをデータから推定する方法も提案されているが，その選択のためにはやはり多大な計算が必要である．

本研究では k-近傍法の考え方を発展させた方法を検討する．すなわち，密度を推定したい検査点を中心とする近傍を考え，この近傍内の密度関数の積分値(累積分布)と実際に含まれるデータ点の個数の比率(経験累積分布)の関係から密度関数の推定量を構成する．重み付きデータの経験累積分布関数は非一様なステップ幅の階段関数となるところが従来の方法と大きく異なり，この階段関数の統計的な性質を詳細に調べることにより，検査点周辺の密度関数の推定量の性質を論じる．この情報量の推定量から重み付きデータのエントロピー，2 つの重み付きデータの間のクロスエントロピーおよび KL-情報量の推定量の構成も検討する．

(2) 学習ベクトル量子化をはじめとして，大量データから歪み関数最小化を規準としたデータの収集・縮約の方法は既に様々なものが提案されている．歪み関数としてはユークリッド距離が多く用いられるが，本研究課題では既存のものに加え，以下のような歪み関数の導入を考慮したデータ縮約方法を検討する．

① サポートベクターマシンに代表されるカーネル多変量解析の考え方に基づくものを検討する．カーネル多変量解析は，データを適当な非線形変換によって高次元空間に埋め込んだ後で従来の多変量解析を行うものであるが，非線形変換から導かれる再生核の性質を利用して高次元空間内での演算を効率的に行う方法論である．適切な高次元空間を選べば原空間で複雑であったデータの分布は単純化され，線形演算に基づく推定方法が頑健な結果を導くことが知られている．前述のように歪み関数としてはユークリッド距離が多く用いられるが，本研究では多重カーネル学習の枠組を用いて適切な非線形変換を選択し，それにより埋め込まれた高次元空間でのユークリッド距離を用いることを検討する．

② JIT モデリングなどのノンパラメトリックな回帰問題において算出される目的変数間の距離と説明変数間の距離の関係を利用して，距離関数そのものを学習する枠組を考察する．このとき，一般化ユークリッド距離(マハラノビス距離)や，ダイバージェンス関数の凸結合などパラメトリックな表現を持つ距離関数の族を利用して，歪み関数の構成を試みる．

(3) 実問題における大規模データに提案手法の適用を試み，その有効性の検証や問題点の洗い出しを行う．

４．研究成果

(1)古典的な k-近傍法によるエントロピーのノンパラメトリック推定量の一般化することにより，重み付きデータの情報量の推定を効率良く行う方法を提案した．まず重み付きデータにおける近傍の定義を見直し，データから計算される経験累積分布関数の分位点にもとづくエントロピー，情報量および KL-情報量の推定量を構成した．これらの推定量の統計的性質を理論的に調べるとともに，計算量や安定性の観点から推定量を緩和し，大規模なデータに対しても高速に計算可能な方法を提案した．

(2) 本研究課題で提案した情報量の推定法は，データ間の距離のみに依存するため，比較的広いクラスのデータに対して適用可能である．一方その精度は採用する距離に強く依存するため，データの解析に適切な距離を何らかの基準のもとで選択する必要がある．このため，データ空間の距離構造をデータそのものから学習する 2 つの問題を考え，データ縮約に適した歪み関数を検討した．

① 非線形の回帰問題に着目し，sliced inverse regression の枠組を用いて多次元の説明変数の最適な次元縮約を行うカーネル関数の線形結合を求める多重カーネル学習の問題として定式化した．この枠組により学習されたカーネル関数は，データ空間の本質的な性質を低次元に効率よく圧縮して取り込んでいると考えられ，獲得されたカーネル関数により導かれる特徴空間が自然な距離構造を持っていることが期待される．次元縮約に際しての評価関数としては，特徴空間上でのデータの分布の正規性が重要な役割を果たすが，特徴空間上に定義される経験特性関数をカーネル関数により直接表現し，正規性を評価する方法を新たに提案した．

② JIT モデリングの枠組においても定式化を行なった．JIT モデリングにおいては予測対象の近傍データを説明変数間の距離を用いて抽出し，近傍データの目的変数の平均や中央値といった統計量を用いて予測を行う．目的変数間に適切な距離が定義されていれば，説明変数での近傍と目的変数での近傍が合致するように説明変数間の距離を学習する問題として定式化できる．距離の値そのものではなく，近傍関係を評価しているため，より柔軟な距離のクラスを考えてモデリングを行うことができる．具体的にはマハラノビス距離や距離関数の凸結合を利用したパラメトリックなモデルを用いて最適化する方法を提案した．

(3) これらの理論的な考察と平行して，実データ解析への応用展開についてもいくつか試みた．1 つは重み付きデータを対象とした情報量の推定法の応用として，複数の回帰関数を用いた条件付分布の粒子近似法を提案した．実際の市況データの予測問題に適用した結果，この分野で標準的に用いられている進化的プログラミングの手法と比較しても遜色なく，いくつかのデータにおいてはより良い結果を得ることができた．もう 1 つは提案する距離学習の枠組をエネルギー消費の分析に適用したものである．家庭の電力需要データのクラスタリングや短期予測，コジェネレーションシステムの最適運用問題に対して計算量の問題は残るものの一定の成果を挙げることができた．

５．主な発表論文等
（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 2 件）
(1) Hideitsu Hino, Keigo Wakayama, Noboru Murata, ″Entropy-Based Sliced Inverse Regression″,*Computational Statistics and Data Analysis*, 査読有,(掲載決定)

(2) Hideitsu Hino, Nima Reyhani, Noboru Murata, ″Multiple Kernel Learning with Gaussianity Measure″,*Neural Computation*, 査読有, Vol.24, 2012, pp.1853-1881, DOI:10.1162/NECO_a_00299

〔学会発表〕（計 4 件　以下抜粋）
① Haoyang Shen, Hideitsu Hino, Noboru Murata, Shinji Wakao, Yasuhiro Hayashi, ″Automatic Extraction of Basic Electricity Consumption Pattern in Households″, International Conference on Renewable Energy Research and Applications,
2012 年 11 月 13 日，長崎

② Hideitsu Hino, Kazuki Miura, Noboru Murata, ″Weight Optimization for Ensemble of Learners by Information Minimization″, The 2nd Institute of Mathematical Statistics Asia Pacific Rim Meeting,
2012 年 7 月 4 日，筑波

〔その他〕
ホームページ等
http://www.eb.waseda.ac.jp/murata/

６．研究組織

(1)研究代表者

村田 昇（MURATA NOBORU）
早稲田大学・理工学術院・教授
研究者番号：60242038

研究協力者

日野 英逸（HINO HIDEITSU）
早稲田大学・理工学術院・助教
研究者番号：10580079