科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１３　日現在

機関番号：　１０１０１
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５００３４０

研究課題名（和文）大規模な質的データの分類とデータ構造の可視化に関する研究

研究課題名（英文）Classification of large scale qualitative data and visualization of their structure.

研究代表者
今井　英幸（IMAI, Hideyuki）

北海道大学・情報科学研究科・教授

研究者番号：１０２１３２１６

交付決定額（研究期間全体）：（直接経費）　　4,000,000円、（間接経費）　　1,200,000円

研究成果の概要（和文）：ウェブ上に日々蓄積されるデータや各種のセンサーから取得されるデータなど、大規模で不定形なデータから知識発見を行う場合には、詳細な分析の前処理としてデータをある程度均質な属性をもつグループに分類することが必要である。大規模なデータは多数の項目を含むが、その中の小数の項目だけに着目することで、妥当な分類ができる場合が多い。本研究では質的データの分類のための変数選択法および正則化法に関して理論的な解析と数値的な検討を行った。この研究結果を赤外線センサーからのデータを用いて人間の行動を推定するシステムに応用した。

研究成果の概要（英文）：To perform knowledge discovery from large data sets as signals from various sensors or texts accumulated in the web, it is required to classify into groups which have some common attribute as a pretreatment for detailed analysis even for such large and irregular data sets. Though such data sets contain a large number of items, it is often sufficient to use a small number of items for reasonable classification. We have studied mainly variable selection and regularization method for classification of qualitative data both from theoretical analysis and numerical experiments. These results have applied to the system for estimating human behavior using the data from the infrared sensors placed in the ceiling of the room.

研究分野：　情報学

科研費の分科・細目：　情報学・統計科学

キーワード：　クラスタリング　ネットワーク　正則化　変数選択　標本化定理

１．研究開始当初の背景

連続的確率変数の解析においては藤越（中央大学）、若木（広島大学）、Ulyanov（モスクワ大学）などにより変数の次元とサンプルサイズの両方に着目した漸近展開法が提案されており、変数の次元が高い場合の漸近的な性質が明らかにされてきた。

データマイニングや機械学習などで解析が求められている質的データは確率変数の次元が高く個体数も多い。こうしたデータではそれぞれの特性をもつ観測度数が少数である場合が多く、全く観測されない場合も少なくない。統計的推測による未知母数の推定や漸近理論による仮説検定などは適切な正則条件の下で展開されており、ある特性をもつ観測度数がゼロである場合にはこの条件を満たさない。正則条件を満たす場合でも、推定すべき母数の個数も多いなどの理由で漸近的な解析では十分な近似が得られないことが多い。観測値（トレーニングデータ）を正確に分類するだけでなく、将来得られるデータ（テストデータ）を精度よく分類するためには分類に寄与する特徴量を選択することが必要である。 Imai et al. (2004) では、情報量規準をもちいて質的データの判別に必要な特徴量を検出する手法が提案されている。しかし、この手法は特徴量を総当りで調べる必要があり、計算量の点で大規模データに適用することは難しい。そのため実時間で大規模データの特徴量を効率的に選択する手法が求められる。

また特徴量対して十分な個体数が得られない場合には分散共分散行列が特異に近い構造をもつことが多く、従来用いられている数量化などの手法では解が安定せず、少数の個体による影響が非常に強く結果に影響を及ぼすことが指摘されている。判別分析においてはFriedman（スタンフォード大学）らによる正則化判別分析など多くの正則化法が提案されており、質的データの特性に基づく正則化法を取り入れた数量化法や対応分析が求められている。宮内他(2005) では、正則化判別分析において情報量規準により正則化パラメータを決定する方法を提案している。これは連続的変量に対して適応可能な手法であるが、質的データの分類においても同様な正則化とそのパラメータの選択が必要である。

データの構造を把握するためには人間のパターン認識特性を活用することも重要である。しかし、この能力は平面などの低次元空間に限られるため、高次元のデータを低次元に写像する必要がある。このような写像により低次元に射影されたデータは馬蹄形問題など見かけ上の構造が現われやすいことが指摘されている。Aoki and Sato(2007)では見かけ上の構造を取り除く方法も提案されているが、高次元データで有効に作用するかどうかは十分には検討されていない。

２．研究の目的

データマイニングや機械学習においてはサンプルサイズが大きいことに加え、確率変数の次元に該当する特徴量が従来の推測的統計学において想定されているものを超えて大規模なデータの解析が必要とされる場面が増えている。これはウェブ上のテキストデータが日々膨大に増加していることや、センサーデータが容易に入手できるようになったことが原因である。

このようなデータは綿密な計画を立てた上で収集される、質がそろった定型的なデータとは異なり、外れ値や欠損値を多く含む不均質なデータであり、また、量的な特徴量、質的な特徴量、言語による特徴量などが混在しているため、正規母集団からのランダムサンプリングのような厳密な家庭に基づく解析手法をそのまま用いることが適当であるとは限らない。

ウェブ上のデータなどからユーザが引き出したい情報は、パラメトリックモデルを仮定した未知母数の推定といった明確なものではなく、経時的に測定される特徴量の因果関係や、特徴量が不連続に変化する点（変化点）の検出といったものである。そうした情報は厳密な解析を用いる前に、探索的な手法と人間の持つパターン認識能力を活用することも必要である。

大規模なデータは従来統計学が仮定しているような単峰は分布をもつ一つの母集団が仮定できる場合は限られており、多くの場合にはいくつかのグループ（クラス）から構成される。データから質の良い情報を引き出すためにはデータを同じような属性をもつグループに分類し、グループに毎に詳細な分析をすることが求められる。

こうしたことから、本研究では、大規模データからの知識発見の手法として、部分クラス法、カーネル法、正則化法などの手法をもちいて質的データの分類手法の改良および提案を行う。

データをグループに分類するためには、特徴量の中からグループの特性を表す比較的少数の指標を構成し、その指標によって似た者同士を同じグループ集め、異質なものを違うグループに入れる。多変量のデータから少数の指標を構成する方法として射影が用いられることが多いが、もとの特徴量の数が非常に多い場合には、低次元への射影において元のデータにはない、見かけ上のデータ構造が表れやすいことが知られている。こうした状況を踏まえ、多次元尺度構成法や対応分析などによる分類指標の構成のための数理的な特性の解明を目指す。

３．研究の方法

大規模な質的データの分類手法および可視化手法を理論的解析と数値実験の両面から検討し、適用範囲や有効性を明らかにした上で従来手法の改良および新たな手法の提案を行う。また、変数選択法、部分クラス法、

正則化法、カーネル法、標本化などを適切に組み合わせることで精度の高い分類手法を開発するとともに、計算量の検討による手法の評価を行う。

（１）Imai et al.(2004)による質的データの特徴量選択手法を大規模データに適用可能な形式に定式化する。この手法を直接大規模データに適用することは計算量の観点から実用的ではないため、実用的な計算量による変数選択手法について検討する。最適な特徴量集合を選択できない場合でも、適当な近似アルゴリズムにより最適な特徴量に近いものを選択する手法の検討を行う。計算量と特徴量の選択の精度にはトレードオフの関係があることが多いので、計算量による手法の評価を行う。特徴量の定式化に基づき、数値的な検討を行う。データの構造を仮定した人工データやベンチマークに用いられるような典型的なデータ適用することにより手法の評価を行う。前進法や後退法、前進・後退法など従来提案されている様々な手法についても評価を行うことで、提案手法の有効な範囲を示すことが可能である。また、近似アルゴリズムによる特徴量の選択とも比較することにより手法の多方面からの検討が可能である

（２）観測度数が非常に少ない特徴量を含む場合に部分クラス法を用いて特徴量をあらかじめ適当な部分クラスに分割する手法を検討する。この手法においては適切な部分クラスが得られるかどうかが解析結果に大きな影響を与えると予想される。情報量規準は母集団の仮定が必要であるためそのままでは質的データに適用することはできないので、最短記述長原理などの適用を検討する。数量化分析において正則化法の適用に関する検討を行う。また正則化パラメータの選択手法として宮内他(2005)により提案された手法の質的データに対する適用に関する考察を行う。カーネル法はカーネル関数を用いることで変数の性質に関わらず適用可能な汎用的な方法であるが、カーネル関数の選択により解析結果が大きく異なることが指摘されている。そのため、できるだけ表現能力の高いカーネルを用いることが提案されている（Tanaka et al.(2007)）。 このカーネルの質的データ解析への適用を検討する。また、カーネルの積分範囲を適切に選ぶ手法を開発する。部分クラス法、正則化法を用いた特徴量の選択に関する数値的な検討を行い、その結果に基づき手法の改良を行う。適切な部分クラスの選択手法には情報量規準によるものの他、交差確認法やブートストラップ法などコンピュータの計算能力を活用した方法も適用することが可能であるため、こうした方法との比較により部分クラス法の有効性を示すことが可能であるまた、正則化法との比較も行う。カーネルの選択による分類性能への影響を数値的に検討する。カーネル法の選択は分類手法に大きな影響を与える可能性があるため様々なデータに対して広範囲の数値実験が必要であり、その結果からカーネルの選択方法への示唆が得られる。また、積分カーネルの積分範囲に関しても同様の数値実験が必要であり、その結果をもとに適切な積分範囲を選択することが可能になる。部分クラス法、正則化法、カーネル法による分類に関してこれらの特徴を統一的な始点から考察することで、手法の特徴を明確にすることができる。また、情報量規準を用いた方法と比較することでこれらの手法が持つデータに仮定に関する検討を行う。

（３）高次元の連続データを低次元に射影してからクラス分類する手法が Toyama et al.(2010)により提案されている。同様の手法を質的データについても可能であるか考察するとともに、手法の改良を行う。質的データ解析の一手法である数量化３類では馬蹄形問題が起こり得ることがしてきされているが、見かけ上の構造が大規模質的データに関しても同様に現われるかどうかの検討を行う。その結果に基づき、見かけ上の構造が現われる場合にはそれを回避する方法を、現われない場合にはより効果的にデータ構造が把握できるような写像を検討する。特に、データの分類を低次元で行う場合には見かけ上の構造は誤った分類結果を導く原因となる。そのため部分クラス法、正則化法がデータの可視化に及ぼす影響に関しても検討する。

４．研究成果

（１）質的データの分類に有効な特徴量を選択することを目的として提案された Imai et al. (2004)による変数選択法は、特徴量のすべての部分集合の中から最適なものを取り出すため、特徴量の次元が高い場合には現実的な時間での実行は不可能であった。Imai(2012)では L1 距離を用いた正則化項を付加した目的関数を導入することにより、分類に有効な特徴量を分類への寄与が大きい順に順次取り入れる手法を提案した。数値実験によりこの手法で選択された特徴量の集合は、総当たりで得られる最適な特徴量の集合と比較して、個数は多くなるものの必要な特徴量はすべて取り入れることが可能であり、また、計算時間は従来の総当たりにより選択する方法と比較して大幅に短縮されることが示された。

さらに、室内の天井に赤外線センサーを設置し、そのセンサーから取得された実データを用いて人間の行動様式、例えば椅子に着席している、数人で固まっている、ソファなどでじっとしている、移動しているなどを分類するためのモデル化を行い、解析および考察を行った（Tao et al.(2012)）。ビデオカメラなどによる監視では行動様式の分類には必要のない詳細な個人情報を取得できるため、ユーザの心理的な抵抗が大きい。それに対し、赤外線を用いる方法では、センサーの探知範

囲内に人がいる、いない、の二値データとして取得されるだけであるため、ユーザのプライバシーを侵害することなく行動様式を分類できる。こうした理由から、一人暮らしの高齢者の見守りなどへの応用が期待される。一方で、赤外線センサーだけでは複数の人間が同時に同じ部屋で生活するオフィスのような環境においては個人を特定することができない。コンピュータへの自動ログインなど、個人を特定することで可能となるサービスを提供するために、椅子の座面に圧力センサーを配置し、座り方の特徴から個人を識別する手法を提案した（Kudo et al(2012)）。また、圧力センサーから個人を識別に必要なセンサーの個数を配置場所を選択するために、判別分析の変数選択法を用いることを提案した。さらに、個人の識別にとどまらず、重心の移動パターンから疲労の度合いを推定する実験を行った。これは自動車の運転時などに疲労を自動的に感知し、休憩を促すための装置に応用できる。

監視カメラなど、動画を用いた行動分類においては、動画のすべてのフレームでのピクセル情報を用いると計算量が膨大になるため、リアルタイムでの対応が求められる侵入者の検知などに利用することが難しい。Luo et al. (2012) では、行動を分類するために必要なフレームだけを抜き出すことで行動様式を実時間で十分な精度を持って分類するアルゴリズムを提案した。また、連続したいくつかの動作を単一の動作に分割する手法を提案した。この手法を用いることで、一連の動画像から人間の行動の推移を推定することが可能となった。この手法は監視カメラによる侵入者の自動検知などに応用ができる。

（2）正則化判別分析は線形判別関数と二次判別関数を含む広いクラスの判別関数による判別を可能にする手法である。この分析法では二個のハイパーパラメータの値が判別性能を決定する。Imai(2012)では、ベイズ型情報量規準を用いることで、ハイパーパラメータをデータに応じて自動的に決定する手法を提案し、数値実験によりその有効性を示した。

正則化判別分析のハイパーパラメータは二次元の閉領域（単位閉区間の直積）の内部から選択する。ハイパーパラメータが領域の内部にある場合には、判別関数はハイパーパラメータに対して連続的に変化する。Imai(2012)では、領域の境界での判別関数の連続性について考察し、境界においては判別関数がハイパーパラメータに対して不連続に変化する可能性があることを示した。さらに、この不連続性は各クラスの分散共分散行列の性質によるもので、連続性が要請される場合には、もう一段階の正則化が必要であることを示した。

カーネル法を用いた判別分析においては、どのようなカーネル関数を選択するかが判別性能を決定する大きな要因である。Tanaka et al. (2012) では、階層構造をもつカーネル関数のクラスを構成し、その中から最小のカーネル関数の集合を選択する方法を提案した。この手法により、機械学習やパターン認識において、適切なカーネル関数を選択する基準を与えるものである。

（3）各変量での周辺分布と同時分布を関連付ける copula を用いた判別分析の判別領域の性質と誤判別確率に関する研究を行った。各変量での周辺分布が正規分布であるような同時分布を copula を用いて構成し、その母集団に基づく最適な判別領域する場合、判別領域が必ずしも連結な領域にはならないことが示された。判別領域が連結でない場合には誤判別確率を解析に求めることはもちろん、近似的に求めることも難しい。このため、マルコフ連鎖モンテカルロ法を用いて誤判別率を数値的に求めるアルゴリズムを提案し、誤判別率を求めた。また、copula のわずかな違いにより判別領域が大きく変わる可能性があることを示した。これは、各変量の周辺分布がほとんど同じ場合でも判別領域は大きく異なる場合があることを示す結果であり、低次元空間へ射影してから判別を行う手法を用いる場合には十分な検討が必要であることを示唆している。

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 11 件）

① A. Tanaka and H. Imai (2014). Parametric Wiener filter with linear constraints for unknown target signals, IEICE Transactions on Fundamentals, 査読有, E97-A, 322-330.

② G. Lu and M. Kudo (2013). Self-similarities in difference images: a new cue for single-person oriented action recognition, IEICE Transactions on Information and Systems, Vol. E96-D, 1238-1242.

③ G. Lu, M. Kudo and J. Toyama (2012). Selection of characteristic frames in video for efficient action recognition, IEICE Transactions on Information and Systems, 査読有, Vol. E96-A, 2066-2070.

④ S. Tao, M. Kudo and H. Nonaka (2012). Privacy-preserved behavior analysis and fall detection by an infrared ceiling sensor network, Sensors, 査読有, Vol. 12, 16920-16936.

⑤ S. Ohnishi, T. Yamanoi and H. Imai (2011). A fuzzy weights representation for inner dependence AHP, Journal of Advanced Computational Intelligence

and Intelligent Informatics, 査読有, vol. 15, 329-335.

〔学会発表〕（計 14 件）

① A. Tanaka and H. Imai (2013). Block-based image interpolation by linearly constrained least mean squares estimation. The 2nd Hokkaido University Korea University Joint Workshop in Statistics, 2013 年 5 月 26 日～31 日, Korea University, Seoul.

② H. Imai (2012). Regularization of discriminant analysis for a small number of samples with a large number of variables, The 1t Korea University Hokkaido University Joint Workshop in Statistics, 2012 年 7 月 6 日～10 日, 北海道大学, 札幌市.

③ S. Ohnishi, T. Yamanoi, and H. Imai (2011). A fuzzy representation for non-additive weights of AHP, 2011 IEEE International Conference on Fuzzy Systems, 2011 年 11 月 27 日～30 日, Taipei, Taiwan.

〔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：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕

ホームページ等

６．研究組織

(1)研究代表者

今井　英幸（IMAI, Hideyuki）
北海道大学・大学院情報科学研究科・教授
研究者番号：１０２１３２１６

(2)研究分担者

工藤　峰一（KUDO, Mineichi）
北海道大学・大学院情報科学研究科・教授
研究者番号：６０２０５１０１

田中　章（TANAKA, Akira）
北海道大学・大学院情報科学研究科・准教授
研究者番号：２０３３２４７１

(3)連携研究者

なし