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 24 年 5 月 24 日現在

機関番号：３２６８９
研究種目：研究活動スタート支援
研究期間：2010～2011
課題番号：22800067
研究課題名（和文）　情報論的規準に基づくデータ分布空間の距離構造学習フレームワークの構築及び応用
研究課題名（英文）　A study on distance metric learning of data distribution space based on information theoretic criterion
研究代表者
日野　英逸（ HINO　HIDEITSU ）
早稲田大学・理工学術院・助教
研究者番号：10580079

研究成果の概要（和文）：
判別に適した形で特徴量が分布するように，データから特徴量を抽出する手法を開発した．本手法はデータが属するクラスを判別する問題一般に適用可能であるが，特に音声データによる話者判別問題に応用し，既存手法を大きく上回る判別精度を達成した．さらに，データが有する情報量の推定法を開発し，種々の応用を提案した．特に，太陽光発電量予測の信頼性評価へ応用し，再生可能エネルギー導入時に問題となる発電量の安定性評価に寄与した．

研究成果の概要（英文）：
We developed a method to optimize the distribution of features extracted from a given set of data. The features obtained by the method will follow a distribution, which is suitable for classification task. Though the method is applicable to any classification problem, we applied it to speaker recognition task, and achieved state-of-the-art performance. We also developed information estimators, which quantify the information contents in a datum. We applied the estimators to evaluate the credibility of prediction of solar power generated from solar panel.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	1,130,000	339,000	1,469,000
２０１１年度	960,000	288,000	1,248,000
年度			
年度			
年度			
総　計	2,090,000	627,000	2,717,000

研究分野：総合領域
科研費の分科・細目：情報学、統計科学
キーワード：統計的学習理論

１．研究開始当初の背景

統計的機械学習やデータマイニング手法はその適用範囲を広げ続けており，多種多様なデータから有用な情報を抽出する課題が日々生まれてきている．多くの学習アルゴリズムの性能は，入力データから抽出する特徴量と，その特徴量を用いて計算されるデータ間の距離に大きく依存している．例えば,類似したデータをグループに分類するクラスタリング問題では，通常はデータから分類に有用な特徴量を抽出し,その特徴量が属する特徴空間において定義した距離に従って分

類を行う.
望ましい結果を得るためには，有用な特徴量を生のデータから取り出し適切な距離を定義しなければならない．解析目的に適した特徴量をデータから抽出する手法は古くから盛んに研究されており，現在でも活発な研究分野である．情報論的な基準に基づく特徴抽出手法は独立成分分析などの分野では広く用いられてきた．しかし，多くの手法はデータの線形変換による特徴抽出であり，また，各データとそのデータに与えられたラベルとの関連性を保持することが主な目的であった.

２．研究の目的
本研究では，データから抽出した特徴量が，特徴空間においてどのように分布しているとデータ解析に適しているか，また，どのように特徴量の分布を最適化するか，という点に注目し，
① データから抽出した特徴空間の最適化
② データが有する情報量の定量的評価
のための手法の開発を目的とした.

３．研究の方法
①のデータから抽出した特徴空間の最適化については，カーネル法と呼ばれる枠組みで研究を行った．データを高次元の特徴空間に写像した上で従来の学習アルゴリズムを実行するカーネル法は，1990 年代のサポートベクターマシンの発明以来非常に盛んに研究されてきた．データが写像される特徴空間の構造はカーネル関数と呼ばれる関数によって決定される．カーネル法は現実的な多くの問題に適用されて成果をあげている一方，利用するカーネル関数を適切に選ばなければいけないという大きな問題があり，カーネル関数の最適な構成は機械学習分野における重要な課題として認識されている．本研究では，最も基本的な２クラス判別問題において，データが特徴空間の中で「クラス条件付きエントロピー」が最小になるように分布していることが，古典的な判別分析手法における判別性基準である Fisher の判別基準最小化の一般化に対応していることに着目し，カーネル関数の組み合わせの最適化により高精度な判別曲面を特徴空間上で学習する手法を検討した．また，視点を変えて，判別手法が特徴空間における Fisher の判別分析であるという状況では，データが特徴空間でどのように分布していればよいか，という問題に着目し，特徴空間においてデータが理想的な分布をするようにカーネル関数を学習する手法も検討した.

②の，データが有する情報量の定量的評価については，特に一つ一つのデータが異なる重みを有する状況において，情報理論や機械学習において最も基本的な量の一つである Shannon 情報量を推定する手法を検討した．Shannon 情報量の推定量が得られると，エントロピー，Kullback-Leibler ダイバージェンスといった重要な諸量を推定することができるため，その応用は多岐にわたる．情報量推定量の開発と，その基本的な性質の解析と並行して，推定量の種々の実問題への応用を検討した．具体的には，時系列の変化の検出，時系列予測の信頼性評価，異常・外れ値検出といった応用を検討した.

４．研究成果
①条件付きエントロピー最 小 化基準による学習の解析及び応 用 と ， 情報量およびエントロピー推定手法の展開を進めた.
　条件付きエントロピー最小化基準による学習方法に関しては，判別問題に適用した際に ， クラス内でのデータの変動に対してロバストであることを，音声認識システムに対する応用を通して実証した．その成果を，国内外の学会にて発表した.
具体的な成果としては，音声による話者の識別問題に取り組んだ．音声発話データは大量のデータを処理する必要があることと，一般にノイズが多いことから，判別手法の適用先としてチャレンジングな課題である．近年提案され良好な精度を示すことで知られている i-vector 方式と，本研究成果として提案した MCEM(Multiple kernel learning by Conditional Entropy Minimization)方式を比較したところ，図１に示すように，発話時間によらず常に提案手法が i-vector 方式を上回る精度を達成した．また，i-vector 方式と提案手法の組み合わせにより構成した判別器は，より高い性能を示した.

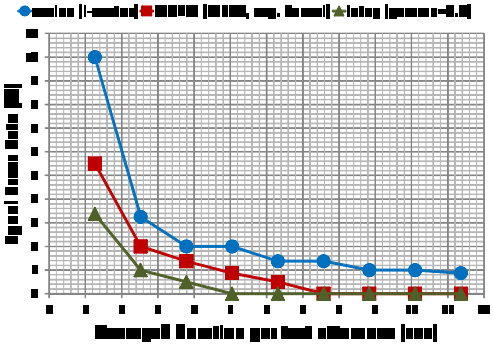

図 1: i-vector(●)，MCEM(■)，及びそれらの結合(▲)による話者識別誤り率.

これは既存の i-vector 方式と MCEM 方式で，データから抽出できる情報の質が異なり，これらを組み合わせることで音声による話者識別に有効な情報をより効果的に利用す

ることが出来たと言える.

CEM 基準による学習の枠組みをさらに推し進め，データから抽出した特徴量がその後のデータ解析に望ましい形で分布するように特徴空間の構造を学習する手法を提案した．本成果を国内学会にて発表し,論文誌にも論文掲載が決定された.特徴量の分布に特定の形を仮定するデータ解析手法は数多く有り,本研究成果を応用することで,そうした分布形に依存する手法の性能を底上げできると考えられる.

② 情報量は，統計科学，機械学習を始めとする多くの分野において基礎的な量であり，観測したデータの価値を測る尺度として多くの応用が考えられる．本研究では基礎原理である推定量の統計的性質の解明に努めた．具体的には，統計量のバイアスの算出，種々の既存統計量との比較実験を行ない，提案する情報量推定量が高次元,重み付きデータに対して有効であることを明らかにした.

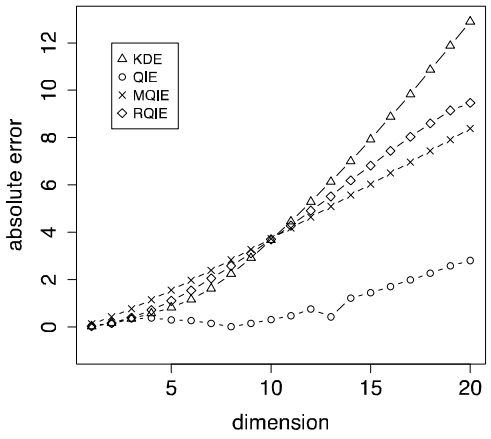

図 2: 各種の情報量推定量による推定誤差と,データの次元の関係. KDE(△)と RQIE(◇)が既存の代表的手法.提案手法(×と○でマーク)は，高次元になっても精度劣化が少ないことが分かる.

次に情報量推定手法の応用として，特に太陽光による発電の予測に付随する予測の信頼性の評価への応用を提案し,国内外の学会にて発表を行った．家庭への安定した電力の供給や，近年注目されている売電のためには，近い将来の時点で期待される発電量を予測する技術が不可欠である.しかし太陽光発電量は，上空に雲が差し掛かると急激に変化するため，精度の高い短期間予測は非常に困難であるとされている．本研究では，予測した値がどれだけ信頼出来るかの評価を，情報量の推定量を利用して行ったものである. Just-in-time モデリングの手法と組み合わせることで，実際に将来の発電量を観測する前に，その時点での発電量の予測値がどの程度信頼出来るかを評価することが可能となった.

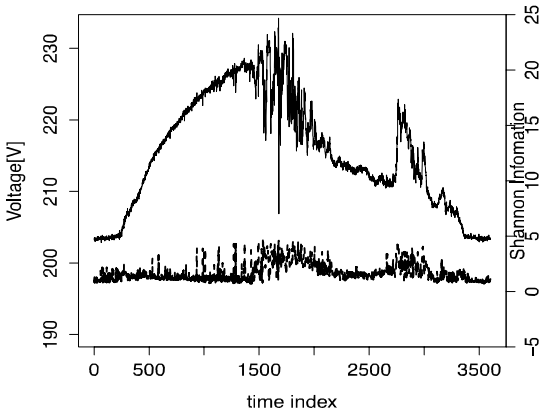

図 3: 太陽光発電実測値と，各時点における予測の不安定さの尺度

図 3 は，実際の発電量と，各時点における発電量予測値の不安定さを評価した値をプロットしたものである．本手法により，発電量が不安定なときには予測値も不安定であることがわかり，その不安定さが定量的に評価出来ていることが分る．本応用は,再生可能エネルギーの導入における問題点である発電量の不安定性の評価と対策に寄与するものである.

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 8 件）

① Hideitsu Hino, Nima Reyhani, Noboru Murata, Multiple Kernel Learning with Gaussianity Measures, Neural Computation, 査読有, vol.24, issue 7,2012

② Haoyang Shen, Hideitsu Hino, Noboru Murata, Shinji Wakao, Extraction of Basic Patterns of Household Energy Consumption, International Conference on Machine Learning and Applications, 査読有,2011,pp.275—280

③ Tetsuji Ogawa, Hideitsu Hino, Noboru Murata, Tetsunori Kobayashi, Speaker verification robust to talking style variation using multiple kernel learning based on conditional entropy minimization, INTERSPEECH,査読有,2011

④ Hideitsu Hino, Noboru Murata, A Computationally Efficient Information Estimator for Weighted Data, Artificial Neural Networks and Machine Learning (LNCS vol.6792),査読有2011, pp.301—308

⑤ Tetsuji Ogawa, Hideitsu Hino, Noboru Murata, Tetsunori Kobayashi, Speaker recognition using multiple kernel learning based on conditional entropy minimization, Acoustics, Speech and Signal Processing,査読有,2011,pp.2204—2207

⑥ Hideitsu Hino, Nima Reyhani, Noboru Murata, Multiple Kernel Learning by Conditional Entropy Minimization, International Conference on Machine Learning and Applications,査読有,2010,pp.223—228

〔学会発表〕(計 12 件)

① 沈浩洋, 日野英逸, 村田昇, クラスタリングによる家庭消費電力パターンの抽出, 数理モデル化と問題解決研究会, 2011, 12 月
② 日野英逸, 村田昇, ガウス性に基づく多重カーネル学習, 第 14 回情報論的学習理論ワークショップ, 2011 年 11 月
③ 小川哲司, 日野英逸, 村田昇, 小林哲則, クラス内変動に頑健なカーネルマシンと話者照合への適用, 日本音響学会 2011 年秋季研究発表会, 2011 年 9 月
④ 村田昇, 沈浩洋, 日野英逸, 太陽光発電量予測とその信頼性評価, 日本鉄鋼協会第 162 回秋季講演大会, 2011 年 9 月
⑤ 小川哲司, 日野英逸, 村田昇, 小林哲則, 条件付きエントロピー最小化基準に基づくマルチカーネル学習を用いた発話スタイル変動に頑健な話者照合, 第 87 回音声言語情報処理研究会, 2011 年 7 月
⑥ 日野英逸, 村田昇, 分位点に基づく重み付きデータの情報量推定手法とその応用, 第 13 回情報論的学習理論ワークショップ, 2010 年 11 月

６．研究組織

(1)研究代表者

日野　英逸 (HINO HIDEITSU)

早稲田大学・理工学術院・助教

研究者番号：10580079