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３月３１日現在

機関番号：３２６１２
研究種目：基盤研究（Ｃ）
研究期間：2009～2011
課題番号：２１５００１４６
研究課題名（和文）　構造混合分布モデルの提案－応用と基礎解析－

研究課題名（英文）　A proposal of structural mixture distribution model - its application and basic analysis -

研究代表者
櫻井 彰人（SAKURAI　AKITO）
慶應義塾大学・理工学部・教授
研究者番号：00303339

研究成果の概要（和文）：学習モデルとしての混合モデルにおいて、その係数の決定方法に関する提案を行った。実データを用いた検証によれば、従来法と同等以上の精度が得られる。ほぼ効率的と考えられる金融市場の価格形成において、要素モデルの存在を示唆するシミュレーション結果を得た。

研究成果の概要（英文）：　For the structural mixture model as a learning model, new methods to estimate mixture coefficients were proposed. It outperformed or was equivalent in performance to existing methods on real datasets evaluation. On price formation in financial markets which are considered efficient, existence of an element model was indicated by simulation trading in the market.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,200,000	360,000	1,560,000
2010 年度	1,100,000	330,000	1,430,000
2011 年度	1,100,000	330,000	1,430,000
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

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：知的システムアーキテクチャ

１．研究開始当初の背景

分布関数の混合を用いて対象をモデル化する際、当該分布関数（以下、要素と呼ぶ）は同一構造でパラメータのみが異なるとすることが多い。しかし、異なる構造を用いた方が対象をよりよくモデル化できることが実験的にわかっていた。しかしながら、どのように混合するか、混合比率を決めるか、要素分布はどのように決めるか、実データを対象として実際に要素モデルが意味を持つか等多くの疑問があった。

２．研究の目的

(1) 構造混合分布（造語である。定義は略す。構造を異にする分布の混合分布）の適切な構成方法を提案し、その有効性と限界を、データをもとに、明らかにする。
(2) 連続混合分布（定義は略す。混合係数を連続確率分布とした混合分布）の適切な構成方法を提案し、その有効性と限界を、現実データをもとに、明らかにする。

３．研究の方法

(1) 要素分布として ODE を例にあげその荷重

付き和に関して、荷重値の定め方を様々試みる。結果を他の要素分布へと拡張する。
(2) 外国為替や株式指標等の市場データを用いて、連続混合分布により実データをより正確に表現することを目指す

４．研究成果
(1) 構造混合に関して得られた結果をまとめる。
① 構造混合の最も簡単な例として ODE (one dependence estimator)の荷重付き和に関して、等荷重、Bayesian model averaging、相互情報量、Kullback-Leibler divergence (KLd)に基づく荷重を比較した。さらに EM 類似の繰り返し計算を伴う準最適化方法を提案し、比較した。相互情報量については、厳密値を用いるより、クラス変数との相互情報量という近似値を用いる方がよい結果を与えた。また、KLd の場合は KLd そのものよりデータ一個あたりの平均値の方がよいことを確認した（次元解析により正当化できるが、厳密な証明は未完である)。相互情報量と KLd に基づく荷重は、その値は全く異なるが、ほぼ同等の汎化能力を有する。EM 類似の方法を用いることにより、近似精度を向上させることができる。通常の EM と異なり、状況が限定されるため、グローバルな最適値への収束が推定されるが、未確認である。なお、UCI machine learning repository の実データを 25 件用い、汎化誤差がほぼすべての場合に有意に減少することを確認した。提案手法中特に意味があるのは、Bayesian model averaging 法を改良したもの（以下 cBMA）と EM 類似の方法である（未公刊のため具体的記述は割愛する)。cBMA は Bayesian model averaging と同様、データから直ちに計算できる。これに対し EM 類似の方法では、任意の初期値から繰り返し計算で求める。両者の値はまったく類似性がないと言ってよいほど異なるが、結果として得られる混合分類器の精度は類似している。実際、上記 UCI の 25 データセットを用いると、分類精度に関する、EM 類似 vs. cBMA の勝ち負け数は、EM 類似の 1 勝 24 引分 0 敗である。なお、現在最も精度がよいと考えられている Weightily ODE に対しては、両者とも 2 勝 22 引分 1 敗である。
② 時系列データに対しても同様の構造混合の手法を適用するための方法を検討し、実用化可能か否かを検討した。直接的な適用は、計算時間の制約と当該時系列データ（外国為替取引データ）の時間方向の非一様性によるさらなる計算時間の増大により、実際上適用は困難である。そこで、ODE 以外の要素分布を自動的に構成する方法として、変数（時系列データを定義域とする関数を含む）と単純な分布関数との組み合わせを試みる GP (genetic programming) ・ GA (genetic algorithm) を用いて、関数形の推定とパラメータ推定を行った。多くの研究で行われているように、時系列データのもののみを用いる方法もあるが、本研究においては、現場の専門家（本研究においてはトレーダー）の知識・経験は有用であると考え、彼らの知識を援用することとした。実際には、いわゆるテクニカル指標を説明変数として用いることである。学術的な研究では、こうした、理論的根拠の乏しい経験知は用いられることが少ない。また実際、テクニカル指標は時系列データの関数であるので、非線形関数を用いた回帰を行えば、等価な表現が暗黙的に学習されても不思議はない。しかし、実験を行ったところ、テクニカル指標を用いた方が学習結果はよかった。テクニカル指標としては、oversold/overbought を表すという RSI (relative strength index)，BIAS, Williams %R 等のテクニカル指標を用いた。現在のところ、GA によるある特定の関数が最も良い推定値を出している。次にはこうして得られた性能上位の分布を組み合わせる試みを行う。
公刊された範囲内での具体的結果を示す。外国為替取引データ中、USD/JPY（米国ドル vs 日本円)の取引履歴を用いた。学習・パラメータ調整期間 6 週間、out-of-sample テスト期間を後続する 2 週間とし、この合計 8 週間の期間を 2 週間ずつずらしながら、2007 年～2009 年の一時間足で実験を行った。この間、交換比率の予測ではなく、模擬取引の収益を測定し、安全資産の金利を考慮した収益・リスク比で評価した。3 年間とも正の収益であり、統計的にも有意となった。GA を用いているため、また実データに現れる、市場の変化のため、得られた取引規則は非常にバラエティに富んでいる。逆に言えば、市場がもつ多数の要素分布中から、多くのものが顕現し、それが観測されたと考えることができる。得られた取引規則を分析し得られた（演繹ではなく帰納による推測で得た）要素分布を用いて、価格の確率過程を、構造混合を用いて記述したところ、予測精度が random walk 仮説が与える予測精度よりよくなった（この部分は未公刊)。予測精度の計測方法が異なるため、簡単な比較は控えなければいけないが、従来手法による予測では、予測誤差が random walk 仮説の予測誤差より小さくなる結果は報告されてない現状を鑑みれば、この結果は、インパクトのある結果だと考える。
③ 半教師あり学習を構造混合の問題ととらえなおし、2 つの基本的な方法（要素分布に対応）に適用すること検討した。すなわち元の形式のままでは半教師あり学習に拡張できない教師付学習アルゴリズムに対して構造混合の考えを用いて半教師あり学習に拡張することを試みた。Minimax probability

machine に対して、manifold regularization を近似する形を得た。
Minimax probability machine (Lanckriet et al. 2002)では、所与の平均ベクトル・分散共分散行列のもとでの最悪の場合の誤判別率が最小になるような線形判別関数を得る。Lanckriet 等はこれを二次錐計画問題に帰着し、効率的な計算法を提示し、さらに実際の問題に対して有効であることを示した。しかし、半教師あり学習への拡張は試みられていない。本研究では、導出した二次錐計画問題に対して（導出の結果として）manifold regularization 相当(同等ではない)の項の導入を可能とするとともに、効率的な計算を可能としている。さらにこの結果に基づき、半教師あり学習を可能とした。本手法を UCI machine learning repository のデータに適用し、よい結果を得た。半教師あり学習を行わせた場合、正則化項にデータの隠れた構造が現れるためと考えられる。引き続きその解析を行っている。
(2) 連続混合に関する結果をまとめる。
①ランダムに見える実際の時系列データに構造があるか否かを、予測が可能（random guessing よりよい）か不可能（random guessing と同等）かで調べた。これは外国為替取引市場では、市場の効率性が成立している可能性が高く、どのような予測手法でも予測はできない、すなわち、構造がないと考えられているが、一方、市場参加者の考えることには共通性があり、いわゆるトレンドが存在する可能性もあるからである。
（a) 移動平均を用いる方法について記す。移動平均を用いる様々な方法を試し、短期と長期に構造が見えやすいフィルター関数を発見した。KZ filter に基づくものである。実際、フィルター適用後の値を用いた予測の予測精度が random walk 仮説に基づく予測精度を超えることを確認した。また市場・時期・観測時間単位によっては構造が発見しやすくなり実際それに基づき取引手数料以上の利益が上がりうることが分かった。この結果は、上記 1. ②の研究結果と関連している。一方ではここに述べるように文献サーベイと数値実験から要素候補を抽出し、組み合わせることでその妥当性を検証した。他方、GA・GP を用いて、組み合わせのランダム探索を行い、やはり数値評価を用いて妥当な要素モデルを抽出した。結果的に両結果は重複しており、目立つものは、何等かの形で移動平均に関連するものである。
(b) 状態を volatility の大きさとして、レジームスイッチングを HMM でモデル化し SVM を用いて次状態予測を行わせることを試みた。データとしては外国為替市場のデータを用い、特に 2001 年から 2009 年までの USDJPY の値(日足)を用いた。情報量基準に基づき、状態数としては 2 が最もよくデータを説明することが分かった。この状態の時系列を、SVM を用いて学習させた結果、予測誤差が random walk 仮説より得られる誤差より小さくなることが確認できた。こうしたことからも実際の volatility の時系列はランダムな系列とは言い難いことが分かる。この結果から容易に連続混合モデルを適用して、volatility の分布モデル、さらに、交換レートの分布モデルを作成することができる。通常モデルとして使われる truncated Levy process より、よい近似となっていると考えている。
(c) 模擬取引では、RSI (relative strength index) , BIAS, Williams %R 等の oversold/overbought を表すというテクニカル指標を組み合わせた。一方、予測にも MACD-value, MACD-signal と呼ばれるテクニカル指標を組み合わせた。また新たな説明変数として他通貨ペアの交換レート及びそれらに基づくテクニカル指標を用いた。多くの研究では、相関を示す研究もあるが、外国為替市場は通貨ペアごとに別個のものとして扱っている（実務としては相関を用いて、予測やそれに基づく取引が行われている)。本研究においては、GBP/USD, EUR/USD, AUD/USD, USD/CHF も説明変数とした。USD/JPY に関する変数だけを用いた場合より、予測誤差は減少した（これは自明ではない。相関はあっても因果関係は必ずしも存在しないからである)。こうしたことから、外国為替取引データには、さらに多くの要素構造があると考えられ、構造混合と連続混合の両者が存在していると考えられる。こうした知見を確認する必要がある。
② Markov regime switching モデルを用いた volatility 予測のため、モデル内のパラメータを積分消去する連続混合を用いて外国為替市場の volatility を予測するモデルを構成した。この時、連続分布で積分消去するモデル（実際にはその近似モデル）ではよい結果が得られず、単純な離散分布との混合では良い結果が得られている。これは、情報量基準を用いて状態数が少ない場合（状態数が 2 の場合）が最良であるという結論が得られていることから、当然な結論かとも思われるが、より詳細な検討が必要と考えている。
③また、上記の minimax probability machine において分類クラスを連続値分布に置き換え、これを積分消去し、連続混合分布と考える regression も考察している。

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 0 件）

〔学会発表〕(計５件)
① Deng Shangkun and Sakurai Akito. Combining Technical Analysis with Sentiment Analysis for Stock Price Prediction. International Conference on Social Computing and its Applicaitons (SAC 2011), Sydney, Australia, December 13, 2011.
② Deng Shangkun and Sakurai Akito. Multiple Kernel Learning on Time Series Data and Social Networks for Stock Price Prediction. ICMLA 2011 Special Session on Learning on the Web, Honolulu, Hawaii, December 19, 2011.
③ Kei Shioda, Deng Shangkun, and Sakurai Akito. Prediction of Foreign Exchange Market States with Support Vector Machine The Tenth International Conference on Machine Learning and Applications 2011, Honolulu, Hawaii, December 20, 2011.
④ K Yoshiyama and A Sakurai. Manifold-Regularized Minimax Probability Machine. IAPR Workshop on Partially Supervised Learning (PSL 2011), Universität Ulm, Germany, September 16, 2011.
⑤ Deng Shangkun and Sakurai Akito Combining Multiple Kernel Learning and Genetic Algorithm for Forecasting Short Time Foreign Exchange Rate. The Eleventh IASTED International Conference on Artificial Intelligence and Applications AIA 2011, Innsbruck, Austria, February 15, 2011.

〔図書〕(計 １件)
① K Ishikawa, Y Shinozawa, and A Sakurai. Self-Organization and Aggregation of Undisclosed Knowledge. In Self Organizing Maps - Applications and Novel Algorithm Design, 143-172 (Chap.9), INTECH, 2011.

６．研究組織
(1)研究代表者
櫻井 彰人 (SAKURAI AKITO)
慶應義塾大学・理工学部・教授
研究者番号：00303339

(2)研究分担者
該当なし

(3)連携研究者
該当なし