様式 C-19

科学研究費補助金研究成果報告書

平成 22 年　6 月　9 日現在

研究種目：若手研究(B)
研究期間：2007～2009
課題番号：19700261
研究課題名（和文）適応的分位点回帰分析によるネットワークトラフィックの確率的予測に関する研究
研究課題名（英文）A study on probabilistic prediction of network traffic using adaptive quantile regression
研究代表者
竹内　一郎　(TAKEUCHI ICHIRO)
名古屋工業大学・工学研究科・准教授
研究者番号：40335146

研究成果の概要（和文）：ネットワークトラフィック量の異常を検知するため，適応的に分位点回帰モデルを推定するアルゴリズムを開発した．開発したアルゴリズムを現実のさまざまな異常検出問題に適用し，その有用性を検証した．

研究成果の概要（英文）：We developed an adaptive algorithm quantile regression algorithm for detecting the anomaly in computer network traffic. We demonstrated the effectiveness of the algorithm by applying it to various practical problems.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	600,000	180,000	780,000
2008 年度	600,000	180,000	780,000
2009 年度	600,000	180,000	780,000
年度			
年度			
総　計	1,800,000	540,000	2,340,000

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：データマイニング，機械学習，異常値検出，分位点回帰分析

１．研究開始当初の背景

(1) コンピュータウイルス感染，不正アクセス，個人情報の流出など，コンピュータセキュリティ被害による経済的/社会的損失は深刻なものとなっている．これに対処すべくさまざまなネットワークセキュリティ技術が開発されているが，その大半はセキュリティ専門家の地道な作業や人海戦術によって成り立っている．コンピュータネットワーク上のデータは時々刻々と増加しており，これらのデータの異常を適切に自動検知できるような汎用システムの方法論の構築が望まれている．

(2) ネットワークセキュリティ対策の効率化や自動化を目的とした従来技術としては，例えば，ウイルスや不正アクセスの早期発見のため，ネットワークトラフィックを常時監視し，特定の閾値を越えた時点で管理者に警告を与えるシステムなどが実用化されている．しかしながら，これらのシステムでは，閾値などの設定がセキュリティ管理者の試行錯誤によって行われるものが多

く，未検出や誤検出を適切に制御し評価することができない．ネットワークセキュリティの異常は悪意を持った cracker に対処することが必要であるため，ロバスト性の高い異常値検出システムが必要とされている．また，人々の活動は時々刻々と変化するものであるため，システム自らが適応的に更新されるような仕組みが必要とされている．

２．研究の目的

(1) 本研究では，ネットワークトラフィックの確率モデルを構築し，異常値を確率的に評価できるような方法論の構築する．これが可能となれば，異常検出システムの未検出率と誤検出率を適切に制御でき，対処方法などの意志決定を適切に行える．システムの異常を「確率的に起こりにくい事象」と定義することにより汎用的な異常値検出の枠組みを構築することが可能となる．コンピュータネットワークトラフィック量を確率的にモデル化するために以下の課題に取り組んだ．

1. コンピュータネットワークでは膨大なデータが高速に行き交うため，大量のデータを高速に処理できるような方法論が必要である．また，ネットワークの状態は時々刻々と変化するため，適応的にモデルを更新する方法を考案する必要がある．大量データを扱う統計モデルの適応的更新アルゴリズムに関しては統計的機械学習の分野で精力的に研究されており，当分野の知見を本問題に適切に導入することが重要となる．

2. コンピュータネットワークの状態は時系列として計測されるため，時系列モデルを構築する．時系列モデルは AR モデル，ARMA モデルなど様々なものが考案され，その予測性能や前提条件などが詳しく研究されている．時系列モデルは主に計量経済学の分野で発展しており，その分野の知見を導入する必要がある．経済時系列データからの異常検出問題を考察することにより有益な知見が得られると期待できる．

3 未検出率と誤検出率を正確に見積もるには，確率分布の裾(テイル)を適切に推定する．従来の確率的異常検出の枠組みではデータの分布が正規分布にしたがうことを暗に仮定したものが多いが，実際のネットワークトラフィックデータはより裾の厚い分布を持つことが予備実験などによりわかっており，分布形状を仮定しないノンパラメトリックな確率モデルを用いる必要がある．

(2) 構築した適応的な異常値検出アルゴリズムを実際のコンピュータネットワークトラフィックデータに適用し，その有効性を検証する．本研究で構築するアルゴリズムの有効性を異常値検出性能と計算速度の観点から評価する必要がある．実際のコンピュータネットワークからデータを取得し，本研究で構築するアルゴリズムを適用することにより有効性の検証を行う必要がある．

３．研究の方法

(1) 本研究では，上記の目的を満たすようなツールとして分位点回帰分析と呼ばれる方法に着目した．通常の回帰分析を本問題に適用すると，様々な条件のもとでのネットワークトラフィック量の期待値と分散を推定することができる．ネットワークトラフィックが正規分布にしたがうことがわかっていれば従来の回帰分析を使うことで十分である．しかし，上述のように，ネットワークトラフィックデータの分布は未知であり，正規分布に比べると裾が厚いものであるため，ノンパラメトリックにトラフィック量の分布を推定できるツールが必要となる．分位点回帰分析を用いると，ネットワークトラフィック量の任意の条件付分位点を推定することができるため，例えば，条件付 99%点などを予測することにより異常値検出を行うことができる．

(2) 本研究では，分位点回帰分析のうち，多次元のデータを効率的に扱うことが可能なカーネル分位点回帰分析を利用する．研究代表者はこれまでもカーネル分位点回帰分析の研究を行っており，関連する最適化アルゴリズムや推定性能の理論解析などに多くの実績がある．カーネル分位点回帰モデルをデータから推定する問題は凸二次計画問題として定式化される．

(3) 本研究では時々刻々と新しいデータが取得される状況においてカーネル分位点回帰モデルを適応的に更新していく数値計算アルゴリズムを開発した．上述のように，カーネル分位点回帰分析の推定は凸二次計画問題として定式化されるため，複数の新しいデータがモデルに加わり，複数の古いデータがモデルから削除されるたびにモデルの再推定を行う必要が生じてしまう．この問題を回避するため，この凸二次計画問題ををパラメトリック計画法(Parametric Programming)の問題として定式化した．このパラメトリック計画法を解くことにより，カーネル分位点回帰モデルの更新を高速に行えるように従来の方法を拡張した．

(4) 構築したアルゴリズムを(研究代表者の以前の所属先である)三重大学情報工学科のネットワークデータに適用し，異常値検出制度とアルゴリズムの速度を検証する.

４．研究成果

(1) 本研究では，まず，データが逐次的に得られる状況でカーネル分位点回帰分析の更新を逐次的に更新する問題をパラメトリック計画問題として定式化した．具体的にはサポートベクトルマシンの逐次更新アルゴリズム(Incremental Decremental Support Vector Machine)の技術をカーネル分位点回帰分析の適応させることによりこれを実現した．その後，パラメトリック計画問題を解くため，数理最適化の分野で研究の進んでいる区分線形パス追跡法を本問題へ適用した．区分線形パス追跡アルゴリズムは問題設定によっては数値的に不安定であるため，これを安定化するための様々な工夫と改良を行った．その成果は，統計的機械学習分野の主要雑誌である Neural Computation に掲載された.

(2) 続いて，(1)において構築した適応的分位点回帰分析アルゴリズムをさまざまな異常値検出問題に適用し，その検出性能を検証した．具体的には，経済時系列データに対して本アルゴリズムを適用し，市場の異常現象を同定できるかどうかを確認した．データの正規性を仮定した従来の異常値検出方法と比較し，本方法の有効性が確認された．この成果は，ソフトコンピューティング分野の論文誌である Jounral of Advanced Intelligent Informatics に掲載された.

(3) さらに，適応的分位点回帰分析アルゴリズムを実際のコンピュータネットワークトラフィックデータに適用し，適応的更新の計算速度が十分であるかどうかを検証した．具体的には，コンピュータネットワーク上のパケットに関する特徴量を多数抽出した．その特徴量を入力変数とし，ネットワークトラフィック量を出力変数とするカーネル分位点回帰分析を構築し，その適応的更新アルゴリズムをオフラインでデータに適用した．その結果，処理すべきデータ量が多い場合に，実用上の十分な計算速度が得られない場合があることを確認した.

(4) (3)の明かになった問題を解決するため，複数のデータをまとめて更新できるよう問題の再定式化を行った．具体的はパラメトリック計画法を多次元に拡張したマルチパラメトリック計画問題となるように本問題を再定義した．続いて，マルチパラメトリック計画法のアルゴリズムを本問題に適合させ，従来のように個別のデータを更新するよりも高速に更新できるよう改良を行った．この改良により，同時に更新するデータ量の二乗根のオーダで高速化が可能となることを理論的にも証明した.この成果は統計的機械学習の主要国際会議である Neural Information Processing Systems (NIPS)に採択されるとともに，当分野の主要雑誌である IEEE Transactions on Neural Networks に掲載され，国内外から注目された.

(5) (4)で改良した複数データに対する分位点回帰分析の更新アルゴリズムを実際のコンピュータネットワークトラフィックデータに適用し，その有効性と計算速度の検証を行った．その結果 ,大量のデータを扱う場合にも十分な計算速度性能が保てることが実証された.

(6) 現在は，パケットからの特徴抽出に関して試行錯誤を繰り返しながら，異常検出性能の向上を図る取り組みを行っている.

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 13 件）

① M. Karasuyama, I. Takeuchi, Multiple Incremental Decremental Learning of Support Vector Machines, IEEE Transactions on Neural Networks, 2010 (to appear).

② M. Sugiyama, I. Takeuchi, T. Suzuki, T. Kanamori, H. Hachiya, and D. Okanohara, Least-Squares Conditional Density Estimation, IEICE Transactions on Information and Systems, 2010 (to appear).

③ Y. Ishikawa, I. Takeuchi, R. Nakano, Multi-directional search from the primitive initial point for Gaussian mixture estimation using variational Bayes method, Neural Networks, vol. 23(3), pp. 356-364, 2010.

④ M. Karasuyama, I. Takeuchi, R. Nakano, Efficient leave-m-out cross-validation of support vector regression by generalizing decremantal algorithm, New Generation Computing, vol.27, no.4, pp.307-318, 2009.

⑤ M. Sugiyama, T. Kanamori, T. Suzuki, S. Hido, J. Sese. I. Takeuchi, L. Wang, A density-ratio framework for statistical data processing, IPSJ Transactions on Computer Vision and Applications, vol.1, pp.183-208, 2009.

⑥ I. Takeuchi, K. Nomura and T. Kanamori, Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression, Neural Computation, vol.21, no.2, pp.533-559, 2009.

〔学会発表〕(計 10 件)
① M. Sugiyama, I. Takeuchi, T. Suzuki, T. Kanamori, H. Hachiya, and D. Okanohara, Conditional density estimation via least-squares density ratio estimation, the 13th International Conference on Artificial Intelligence and Statistics, 2010 年 5 月 15 日, Sardinia, Italy.

② Y. Ishikawa, I. Takeuchi and R. Nakano, Variational Bayes from the Primitive Initial Point for Gaussian Mixture Estimation, International Conference on Neural Information Processing 2009, 2009 年 12 月 3 日, Bangkok, Thailand.

③ N. Harada, Y. Ishikawa, I. Takeuchi and R. Nakano, A Bayesian Graph Clustering Approach Using Degree Distribution Prior, International Conference on Neural Information Processing 2009, 2009 年 12 月 3 日, Bangkok, Thailand.

④ H. Moriguchi and I. Takeuchi, Adaptive kernel quantile regression for anomaly detection of time series, SCIS and ISIS 2008, 2008 年 9 月 21 日, Nagoya, Japan.

〔図書〕(計 0 件)

〔産業財産権〕

○出願状況 (計 0 件)

〔その他〕
特になし.

６．研究組織
(1)研究代表者
竹内　一郎 (TAKEUCHI ICHIRO)
名古屋工業大学・工学研究科・准教授
研究者番号：40335146