2007 年 9 月 12 日(水)
1031060322 松本 亮介

統計的手法による
時系列からの外れ値と変化点の検出

1 はじめに

近年，コンピューターネットワークシステムが社会に普及していくにつれ，ネットワークセキュリティの確保が必要不可欠となっている．そこで，インターネット上でおこるセキュリティインシデントを分析する「インシデント分析システム」が注目されている．

インシデント分析に関する重要な要件に，各種ログデータからのインシデント候補のリアルタイム検知があり，時系列データからの外れ値特定による変化点検出が提案されている．この種の研究は，これまで，統計量の分野で広く扱われてきている．

情報源の外れ値や変化点の特定は，ログデータから異常行動や不正行為につながるデータや，新しいトレンドを示す重要なデータを発見することができるとされている．そのため，データマイニング[1][2][3]や統計量に関する研究において，最も注目されている問題である．

本稿では，データマイニングの観点から，外れ値と変化点に明確な関係を与え，外れ値検出と変化点検出，両方を同一の学習アルゴリズムに基づいて取り扱うための統計的手法[4]を紹介する．この手法は，時系列データを当てはめる確率モデルが過去の統計量を次第に忘れていくことによって，リアルタイムで時系列データの特徴をうまく抽出し追跡できる学習アルゴリズムである．

以降 2 章において紹介手法の概要と従来の研究との関係について述べ, 3 章で紹介手法の 2 段階学習について説明し, 4 章でその他の関係する手法を紹介し, 5 章で実験と考察，6 章でむすびとする．

2 手法

2.1 紹介手法の概要

これまで，外れ値検出に関する手法[1][5][6]が，研究されてきたが, 時系列データにも対応できるような明確な統計的手法は，提案されていない．

そこで，時系列データを用いて，変化点と外れ値の検出に明確な関係を持たせ，両方を取り扱うための統計的手法を示す．

2.1.1 目的

ネットワークトラフィックにおける，変化点検出の問題は，アクセスパターンの統計的な周期規則に対して，大きな変化が生じた時間や，外れ値が現れた時間を識別することにある．

本稿では，以下に示すような方向性に沿って，文献[5][7]を拡張し，時系列データを用いて変化点を検出する．

第一に，ガウス混合モデルのような独立モデル[5][7]の代わりに，AR(自己回帰)モデルのような時系列モデル[8]を用いる．オンラインで用いるために，時系列モデルで学習する際の計算量を減らしたアルゴリズムを開発する．このアルゴリズムでは，学習モデルからの外れ具合によってある特定のデータに対するスコアを計算し，そのスコアが高い程，外れ値になる高い可能性を示している．

第二に，変化点検出と外れ値検出を関連づけて考えた時，一定サイズのウィンドウ内のスコアの平均を計算し，ウィンドウをスライドすることによって，移動平均スコアによる新しい時系列データを得る．変化点検出は，これらの時系列データからさらに外れ値を求める問題に置き換えることができる．本稿では，この機構を ChangeFinder と命名する．

本稿では，実験によって，提案手法がネットワークセキュリティにおけるインシデント検出に対して有効であることを示す．

2.1.2 関連する理論と研究

時系列の統計的な振る舞いを表すために，AR モデルを用いる．AR モデルは，統計学の分野で広く利用されてきた時系列の最も典型的な統計モデルである[8][9].

統計量の変化点検出の一般的な方法は，先験的に変化点の数を測定して，連続的な変化点の間で，その範囲に合う定常モデルを決めることである[8][10][11][12]．しかし，実際の適用において，局部的に定常であるという仮定は，統計的な周期が変化する可能性があるため，除去しなければならない．

文献[2]では，変化点検出の問題は，データが局部的に定常であるという仮定を使わずに扱われている．そのように定義するかわりに，変化点を連続的な区間同士の境の点と定義し，時系列データにあてはまるように，断片的に区分していく機能を用いている．その方法であれば，変化点の前後のデータ区間にあてはまる局部的なモデルの誤差を全て最小化するような点を見つけることで，変化点を検出できる可能性がある．しかし，この方法では，区間毎に局部的なモデルに対応した処理が何度も必要になる．そのため，変化点を見つけるには，膨大な計算量が必要となる．さらに，断片的に区分することができないようなデータに関しては，処理できない可能性がある．

それに対し，本稿ではより一般的な方法を取る．AR モデルが，新たなデータを 1 つ読み込むごとに，過去の情報に対する重みが次第に割り引かれるように，パラメータの推定値を更新する．そして，各データにスコアを与え，より高いスコアほど，外れ値が変化点になる可能性が高くなるようにする．これによって，[2]の手法のような，断片的に区分する機能では表すことができないような時系列データに対しても，有効なデータとして取り扱うことができる．

2.1.3 従来手法との比較

今までの研究では，外れ値検出および変化点検出が明確に関連付けられていなかったが，本稿ではそれらに明確な関係を与え，両方を取り扱うための統計的手法を示す．この手法では，同一の学習アルゴリズムに基づいて，外れ値および変化点をリアルタイムで検出することができる．

提案する変化点検出のアルゴリズムは，計算量の点で有効であり，高い精度で検出することができる．

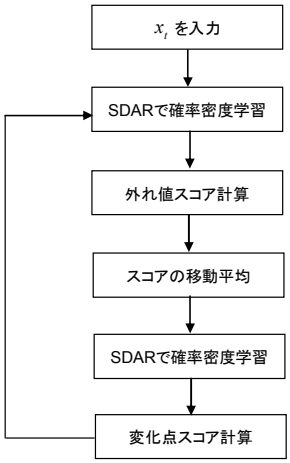
3 2 段階学習アルゴリズム

ここでは，ChangeFinder と命名した 2 段階の時系列を用いた変化点検出の理論を述べる．ChangeFinder の特筆すべき点は，2 段階の学習過程を繰り返すところにある．最初に，第 1 段階で学習したモデルを利用して外れ値を検出する．次に，第 2 段階で学習したモデルを用いて変化点検出を行う．

3.1 ChangeFinder

第一段階学習では，まず各時刻 t において，AR モデルを SDAR アルゴリズムによって学習する．そして，時系列データに対する外れ値らしさを示す，外れ値スコアを計算する．

第二段階学習では，外れ値スコアに対して，再度 AR モデルをあてはめ，これを学習して，変化点スコアを計算する．変化点スコアが大きいほど，t が変化点である度合いが高い．



フローチャート 1：ChangeFinder の流れ

ChangeFinder の特徴は，第一段階学習では時系列中の外れ値しか検出できないところを，外れ値スコアの平滑化を通じて，本質的なモデルの変動を検出しているところにある．計算量に関しても，データ数 n に対して，統計的検定に基づく方式が $O(n^2)$ であるのに対して，ChangeFinder の計算量は $O(n)$ で済むため，明らかに効率がよいことがわかる．

さらに，ChangeFinder は平均値の変化だけでなく，AR モデルのパラメータ（AR 係数や分散）の変化も原理的には検出できる．実際に，分散が突然変化する場合でも，十分な効果が得られるという報告がある．

3.2 外れ値の検出

最初に，時間 t に対して，$\{x_t : t = 1,2,\cdots\}$ で表すことができる時系列データを考える．t が変化するときの x_t が，本稿における重要な意味を示す d 次元の実数値ベクトルである．ChangeFinder は，第 1 段階で，$\{p_t : t = 1,2,\cdots\}$ として表される時系列データの確率密度関数を計算する．データ x_t が入力されると，この時系列データは $\{x_t\}$ から徐々に学習していく．一般的に，それぞれの p_t が確率過程の密度を表すと考える．確率過程 p に対して，$x^t = x_1 x_2 \cdots x_t$ を与える x_{t-1} の

条件付き確率密度関数を考えるために，$p(x_{t-1} \mid x^t)$のような表記法を用いる．学習方法には，確率過程 p を推定するために，各入力 x_t に対して，次式(1)を用いて，x_t の外れ値スコアを計算する:

$$\mathrm{Score}(x_t) = -\log p_{t-1}(x_t \mid x^{t-1}) \qquad (1)$$

式(1)の左辺は，確率密度関数 $p_{t-1}(\cdot \mid x^{t-1})$ に対する x_t の対数予測損失を表し，対数損失スコアと呼ぶことにする．

また，対数損失ではなく，2 次的損失に基づく別のスコアを次式(2)のように定義する．

$$\mathrm{Score}(x_t) = (x_t - \hat{x}_t)^2 \qquad (2)$$

また，以下のように，$\hat{x}_t$ は学習モデル p_{t-1} に基づいて，ある x に対する x^{t-1} からの予測を表している．

$$\hat{x}_t := E_{p_{t-1}}\left[x_t \mid x^{t-1}\right] \stackrel{\mathrm{def}}{=} \int x p_{t-1}(x \mid x^{t-1}) dx$$

これを，二次損失スコアと呼ぶことにする．$\mathrm{Score}(x_t)$ が高いほど，x_t が外れ値である可能性が高いことを示す．

3.2.1 AR モデル

例えば，一連の確率過程 $\{p_t\}$ の系列に，AR（自己回帰）モデルを用いるとする．初期値が 0 であるような，定常時系列 $\{z_t : t = 1,2,\ldots\}$ を考える．それぞれの z_t は d 次元の縦ベクトルを表す．このとき，以下の式より，k 次の AR モデルを与えることができる．

$$z_t = \sum_{i=1}^{k} A_i z_{t-i} + \varepsilon$$

ただし，データ z_t は n 次元のベクトル，$A_i (i = 1,\ldots,k)$ は n 元正方行列，ε は期待値 0，共分散行列 Σ のガウス分布 $\mathrm{N}(0,\Sigma)$ に従うノイズ項であるとする．

実際に観測されるデータを

$$x_t = z_t + \mu$$

で表す．

これより，期待値が μ，$x_{t-k}^{t-1} = (x_{t-1} \cdots x_{t-k})$，であるとすると，$x_t$ の確率密度関数は，

$$p(x_t \mid x_{t-k}^{t-1} : \theta) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{\xi^T \Sigma^{-1} \xi}{2}\right) \quad (1')$$

で与えられる．ただし，

$$\xi = x_t - \left(\sum_{i=1}^{k} A_i z_{t-i} + \mu\right)$$

であり，$\theta = (A_1,\ldots,A_k,\mu,\Sigma)$ とする．

AR モデルに関する通常の推定アルゴリズムについて，次式(2')，(3')を定義する．

$$\hat{\mu} = \frac{1}{t-k} \sum_{i=k+1}^{t} x_i \qquad (2')$$

$$C_j = \frac{1}{t-k} \sum_{i=k+1}^{t} (x_i - \hat{\mu})(x_{i-j} - \hat{\mu})^T \qquad (3')$$

式(2')は μ の推定値，(3')は $x_1,\ldots,x_t$ の相関関数の推定値を表す．さらに A_i の推定値は，以下の $\overline{A_i}$ を未知数とする連立方程式を解くことで得られる．

$$C_j = \sum_{i=1}^{k} \overline{A_i} C_{j-i} \quad (j = 1,\ldots,k) \qquad (4')$$

式(4')の解 $\overline{A_i}$ より，Σ の推定値は，

$$\hat{\Sigma} = C_0 - \sum_{i=1}^{k} \overline{A_i} C_i \qquad (5')$$

によって求めることができる．しかし，この手続きでは，情報源が定常であると仮定されており，いわゆるバッチ学習方式になっている．

3.2.2 AR モデルの改良

ここで，AR モデルを改良して，以下のように考える．θ_t は，x_t が与えられたときの θ の推定値を表し，$p_t = (\cdot \mid \cdot \theta_t)$ とする．θ の評価のために，以下の量を最大にする θ の値を計算するアルゴリズムを提案する．

$$\sum_{i=1}^{t} (1-r)^{t-i} \log p(x_i \mid x^{i-1}, \theta)$$

これは，オンラインで使用するための，最尤推定法の変形である．このとき，重さが時間 t で指数的に減少するところで，尤度が最大になる．これを，SDAR (sequentially discount-ing AR model estimating)アルゴリズムと呼ぶことにする．

SDAR アルゴリズムは，バッチ学習方式の AR モデルを改良した，逐次型学習方式である．SDAR アルゴリズムでは，逐次学習と忘却機能という 2 つのポイントがある．

逐次学習とは，新たなデータを 1 つ読み込むごとにパラメータの推定値を更新する．

忘却機能とは，i 時点前のデータの影響が $(1-r)^i$ 倍に減少するようにパラメータの推定値を更新する．これによって，非定常な情報源に対応できる．アルゴリズムのパラメータ r を忘却パラメータと呼び，$1/r$ 個程度の過去データの情報を蓄積するようにする．以下に SDAR アルゴリズムを示す．

SDAR アルゴリズム $(0<r<1$:所与$)$
STEP 1. 初期化
Set $\hat{\mu}, C_j, \hat{A}_j (j=1,\ldots,k), \hat{\Sigma}$.

STEP 2. パラメータ更新
For $t=1,2,\ldots,$
x_t を読み込む:

$$\hat{\mu} := (1-r)\hat{\mu} + r x_t$$

$$C_j := (1-r)C_j + r(x_t - \hat{\mu})(x_{t-j} - \hat{\mu})^T$$

以下の連立方程式を A_i について解く:

$$C_j = \sum_{i=1}^{k} A_i C_{j-i} \quad (j=1,\ldots,k). \qquad (6')$$

方程式(6')の解を $\overline{A}_1,\ldots,\overline{A}_k$ とし，以下を計算

$$\hat{x}_t := \sum_{i=1}^{k} \overline{A}_i (x_{t-k} - \hat{\mu}) + \hat{\mu}$$

$$\hat{\Sigma} := (1-r)\hat{\Sigma} + r(x_t - \hat{x}_t)(x_t - \hat{x}_t)^T$$

このアルゴリズムにおいて t 番目のデータまで用いて得られる確率密度関数(1')を，p_t と書く．

3.3 変化点検出

次に変化点スコアを求める．T を正の整数とする．データ列 $\{x_t\}$ に対して，T 移動平均スコア y_t を次式(3)で定義する．

$$y_t = \frac{1}{T}\left(\sum_{i=t-T+1}^{t} \mathrm{Score}(x_i) \right) \qquad (3)$$

ただし，$\mathrm{Score}(x_t)$ は式(1)より求める．この計算によって，新たな時系列 $\{y_t : t=1,2,\cdots\}$ を得る．

次に，$\{y_t\}$ を入力データとして，再度 SDAR アルゴリズムを用いて AR モデルの学習を行う．q_t を y_t まで用いて得られる確率密度関数で表すと，AR モデルによる確率密度関数の列 $\{q_t : t=1,2,\cdots\}$ が得られる．

さらに，対数損失式(1)と 2 次損失式(2)と同様に，T 移動平均スコアを次式(4)で定義する．

$$\mathrm{Score}(t) = \frac{1}{T} \sum_{i=t-T+1}^{t} \left(-\log q_{i-1}\left(y_i \mid y_{i-k'}^{i-1}\right) \right) \qquad (4)$$

式(4)は，時間 t の変化点らしさを示す指標となる．すなわち，$\mathrm{Score}(t)$ が大きければ，変化度合いが大きいと解釈することができる．これを，変化点スコアと呼ぶ．

4 その他の手法

3 章で述べた手法のほかに，比較のために Guralnik と Srivastava[2]による手法を用いる．この手法は，任意の時刻の前後に，モデルをあてはめ，次に，誤差の合計が，変化点がまったくない場合と比べて，大きく減少していれば，変化点が存在すると考える．データに対しては，多項式関数をあてはめ，2 乗誤差を用いて誤差を測定する．この手法[2]を GS 手法とする．

GS の計算量は，2 乗誤差を繰り返し計算しなければならないため，AR モデルによる ChangeFInder よりもはるかに多くなる．GS 計算量はそれぞれのパラメータの増加により，総計算量は n のデータを扱った場合，最悪 $O(n^2)$ のような指数的な量になってしまう．しかし，CF（ChangeFinder）では，n に対して計算量は $O(n)$ のように直線的になる．次の章では，2 つの手法（CF，GS）の性能を比較する．

5 実験と考察

5.1 シミュレーション

提案手法を数値シミュレーションによって評価した．ここでは，2 種類のデータセット，データセット 1 およびデータセット 2 を用意した．両者とも 10,000 個のデータからなり，1,000 番目毎に平均値が不連続に変化するように生成した．

データセット 1 では，次式(6)，(7)のモデルに従ってデータを発生させた．

$$z_t = 0.6 z_{t-1} - 0.5 z_{t-2} + \varepsilon_t \qquad (6)$$

$$x_t = z_t + \mu_t \qquad (7)$$

図 1 は，データセット 1 を示しており，横軸は時間 t，縦軸は x_t である．

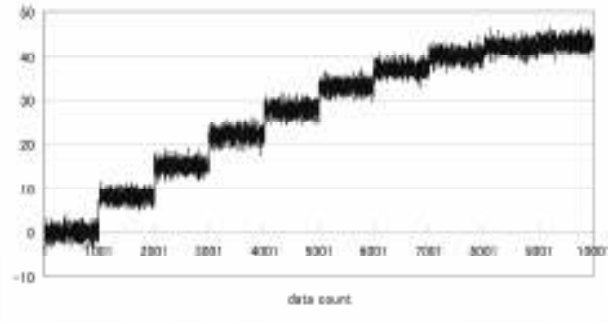


図 1：データセット 1

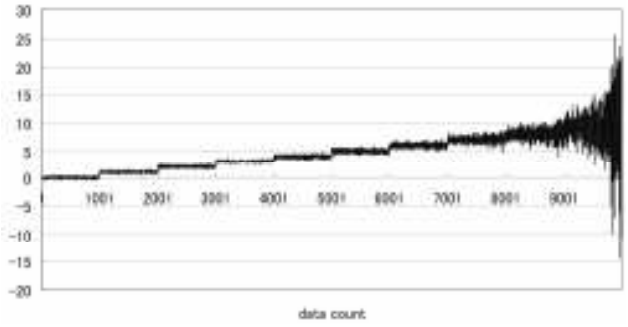


図 3：データセット 2

ただし，ノイズ項 ε_t は，期待値 0，分散 1 の独立ガウスノイズである．また，μ_t は，$\mu_1 = 0$ とし，1,000 データ毎に配置された変化点以外では一定の値をとり，$N = 1$ 番目の変化点において，$\mu_t = 10$ とし，N が増加するごとに，1 減少するように設定した．

このデータに対し，2 段階の SDAR アルゴリズムに AR モデルを適用して，変化点スコアの計算を行った．SDAR アルゴリズムでは，1 段目 2 段目ともに 2 次の AR モデルを用いた．また，忘却パラメータは全て $r = 0.005$ とし，$T = 5$ と設定した（(3)参照）．

図 2 は SDAR による変化点スコアをグラフに示したものである．横軸は時刻を表し，縦軸は変化点スコアである．グラフから，変化の大きさがスコアの大きさによく反映されていることがわかる．

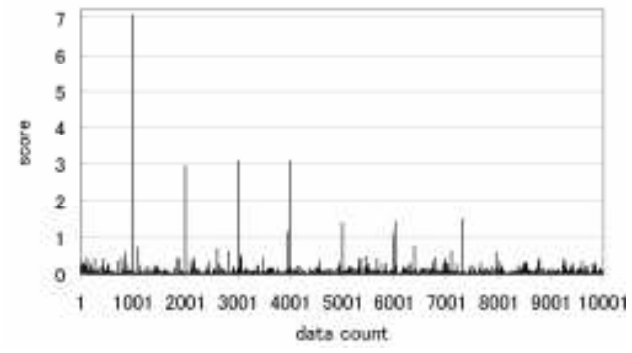


図 2：SDAR による変化点スコア（データセット 1）

次に，データセット 2 でも同様のシミュレーションを行った．

データセット 2 も，変化点の間のデータは基本的に式(6)の AR モデルで生成したが，ノイズ項の標準偏差を時刻に応じて $1000 \times t$ のように変化させた．また，1,000 データごとに設定した変化点における期待値 μ_t は，$\mu_1 = 0$ とし，変化点 N 番目ごとに 1 増加するように設定した．図 3 にデータセット 2 を示す．

図 4 に SDAR によるデータセット 2 に関する変化点スコアのグラフを示す．データセット 2 では，変化点と変化点の間のデータ列は分散が変化するという意味で，非定常なモデルで生成されているが，定常な場合とほぼ同等の検出結果が得られていることがわかる．

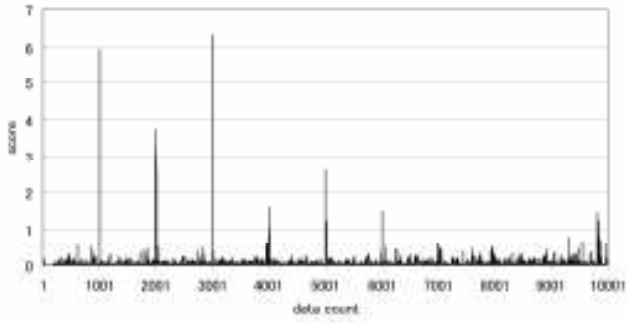


図 4：SDAR による変化点スコア（データセット 2）

5.2 実データによる実験

ここでは実データを使用して，CF をあるネットワークセキュリティ問題に適用して実験を行った．使用したデータは，インターネットに接続されているサーバの，あるポートに対する接続頻度の時系列データである．例えば，サーバとクライアントが接続をする時に，サーバ側が処理することができるよりも早く，クライアント側が TCP 接続要求を発信する「SYN-FLOOD 攻撃」と呼ばれる DoS 攻撃がある．このような攻撃は，接続頻度が多くなるため，変化点検出を適用することで，攻撃を早期に発見することができると考えられる．

図 6 はあるサーバの Port135 に対する接続頻度の時系列を表している．縦軸は，CF によって求めた，毎時の接続頻度に対する変化点スコアを表しており，横軸は，時間（2003 年のグリニッジ標準時-7 月 31

日午後3時から8月26日の午後2時)を表している.

ここで，CF において，1 段目に 2 次の AR モデルを適用し, 2 段目には 3 次の AR モデルを適用して実験を行った. 変化点スコアには, 対数損失を使用し, 忘却パラメータは $r = 0.0.2$ とし，$T = 5$ とした（式(3)参照）.

図 6 の，変化点スコアには，2 つの大きなピークがあり，そのピークは MS.Blast ウィルスの 1 回目と 2 回目の大発生を表している．変化点スコアのグラフから，3.0 を閾値として設定した．MS.Blast は 2003 年 8 月 11 日に発見されており, Trend Micro 社によるレポートでは, グリニッジ標準時 2003 年 8 月 11 日午後 9 時 44 分に発表されている.

それに対して，CF では，そのレポートよりも 1 時間 44 分前に発見することができた．しかし，2 回目の大発生に関しては，Trend Micro 社がグリニッジ標準時 2003 年 8 月 17 日午前 5 時 31 分にレポートを発表したのに対して，CF では約 1 日遅れて発見するに至った.

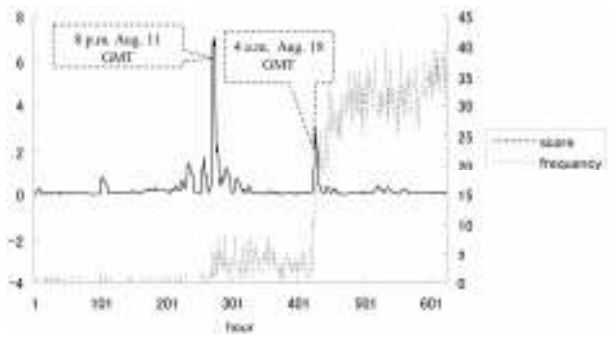


図 6 : CF による MS.Blast に対するスコア

また GS に関して, 同じデータセットを用いて評価した．その結果を，表 1 に示す．この結果からも，CF が GS と同じように，変化点を正確に検出していることが分かる．また，アラート検知時間に関しては，GS は CF よりも，最長約 5 日以上遅れていることが分かった．従って，CF は GS よりも非常に早く発見できていることが分かる.

表 1 : GS による変化点検出時刻

検出した変化点	**変化点1**	変化点2	変化点3
変化点を通知した時刻	8月11日午後9時	8月13日午後10時	8月23日午後10時

6 むすび

本稿では，非定常な時系列データから外れ値と変化点を統一的に扱って検出する手法（ChangeFinder）を紹介した．これは，統計的モデリングとスコア付けという 2 つの手続きからなる．モデリングにおいては，データ列から逐次的に確率密度関数を学習する SDAR アルゴリズムを用いた．特に，本稿では AR モデルを採用した．スコア付けに関しては，学習したモデルに基づいて各データに対してスコアを与える．特に，変化点検出の問題を，外れ値スコアの系列の移動平均における外れ値検出に還元した．この手法の特徴は, 2 段階で学習過程を繰り返すところにある．まず，第 1 段階で外れ値を検出し，第 2 段階で第 1 段階において求めた外れ値スコアの時系列移動平均を用いて，変化点を検出する．これにより，外れ値と変化点の両方を同時に取り扱う統一的枠組みが得られた．また，従来の方法より，計算量の面で非常に効率がよく，精度に関しても従来法と同程度であることがわかった.

今後の課題としては，以下が挙げられる.

1, 外れ値スコアの移動平均 T の適応的決定

2, ARMA モデルの採用

3, マルコフモデルとの融合

1 に関して，本稿で説明した実験において，移動平均 T は経験的に決定した．その値が，うまくいく理由，または，データセットによってどのように T の最適な値が決まるのか．これらは，紹介手法において重要な要件であるといえる.

2 に関して, ARMA モデルとは AR モデルを一般化し，時系列データに対して，より柔軟に対応できるモデルである．こういったモデルに対処できることで，検出精度を上げることが可能であると考えられる.

3 に関して, 本稿の枠組みではデータは全て連続値であると仮定している．より一般的なデータを扱うためには，離散値でのデータを扱えるようにする必要がある．例えばマルコフモデルや隠れマルコフモデルといった離散値データを扱う時系列モデルを取り入れることで，連続値と離散値を同時に扱えることができると考えられる.

参考文献

[1] P. Burge and J. Shaw-Taylor, “Detecting Cellular Fraud Using Adaptive Prototypes,” Proc. AI Approaches to Fraud Detection and Risk Management, pp. 9-13, 1997.

[2] V. Guranlink and J. Srivastava, “Event Detection from Time Series Data,” Proc. ACM-SIGKDD Int’l Conf. Knowledge Discovery and Data Minig, pp. 33-42, 1999.

[3] E.M. Knorr and R.T. Ng, “Algorithms for Mining Distance Based Outliers in Large Data Sets,” Proc. 24th Very Large Data Bases Conf.,pp. 392-403, 1998.

[4] J. Takeuchi and K. Yamanishi, “A Unifying Framework for Detecting Outliers and Change Points from Time Series,” IEEE transactions on Knowledge and Data Engineering, , pp.482-492, 2006.

[5] K. Yamanishi, J. Takeuchi, G. Williams, and P. Milne, “Online Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms,” Data Mining and Knowledge. Discovery J., vol. 8, no. 3, pp. 275-300, May 2004.

[6] U. Murad and G. Pinkas, “Unsupervised Profiling for Identifying Superimposed Fraud,” Proc. Third European Conf. Principles and Practice of Knowledge Discovery in Databases, pp. 251-261, 1999.

[7] K. Yamanishi and J. Takeuchi, “Discovering Outlier Filtering Rules from Unlabeled Data,” Proc. Fourth Workshop Knowledge. Discovery and Data Mining, pp. 389-394, 2001.

[8] 赤池弘次，北川源四郎 編，時系列解析の実際Ⅰ,Ⅱ，朝倉書店，1994,1995.

[9] G. Kitagawa and W. Gersch, “Smoothness Priors Analysis of Time Series,” Lecture Notes in Statistics, vol. 116, Springer-Verlag, 1996.

[10] S.B. Guthery, “Partition Regression,” J. Am. Statistical Assoc., vol. 69, no. 348, pp. 945-947, Dec. 1974.

[11] D.M. Hawkins, “Point Estimation of Parameters of Piecewise Regression Models,” J Royal Statistical Soc. Series C, vol. 25, no. 1,pp. 51-57, 1976.

[12] M. Huskova, “Nonparametric Procedures for Detecting a Change in Simple Linear Regression Models,” Applied Change Point Problems in Statistics, 1993.