文字列解析に基づくネットワークトラフィックデータからの異常発見

岡部 正幸[†]　　三輪 多恵子[‡]　　梅村 恭司[§]

[†] 豊橋技術科学大学 情報メディア基盤センター

[‡] 豊橋創造大学

[§] 豊橋技術科学大学 情報工学系

Anomaly Detection in Network Traffic based on String Analysis

Masayuki Okabe[†]　　Taeko Miwa[‡]　　Kyoji Umemura[§]

[†] Information and Multimedia Center, Toyohashi University of Technology

[‡] Toyohashi Sozo University

[§] Information and Computer Science, Toyohashi University of Technology

概要

本研究では，ネットワーク上の異常トラフィックを, 時系列データの文字列表現手法を用いて検出する方法を提案する．文字列表現による異常発見方法としてマルコフモデルによる方法が既に提案されているが，ネットワークデータに適用した場合，誤検知が多く，これをそのまま使用すると発見効率が悪い．このため，我々はマルコフモデルによって列挙された文字列集合にクラスタリングによるはずれ値検出手法を適用することで，誤検知を取り除く方法を提案する．本稿では，実際のトラフィックデータを使用した実験を行い，提案手法の有効性を示す.

1　はじめに

近年，P2P ソフトを介した個人情報や違法にコピーされた映画や音楽などのコンテンツの流出が問題となっている．また，ウイルスに感染した PC が不特定多数の相手に大量のメールを送りつけるなど，ネットワーク上で発生するこのような異常トラフィックを発見，対処することが安定的な運用を行う上で欠かせない.

異常発見方法には大別して，シグニチャ型とアノマリ型がある．シグニチャ型は，予め異常なパケットの特徴，例えばペイロード中に必ずある文字列が出現するといった特徴をルールとして記述しておき，パターンマッチを行って検出する．この方法は既知の振る舞いは確実に検出できるという利点があるが，ウイルスの亜種など微妙な特徴の変化には対応できないという欠点がある．最近，特定の P2P 通信を遮断するソフトが開発されているが，これらもシグニチャ型といえる．一方，アノマリ型では，予め正常な通信パターンをモ

デル化しておき，これに外れる振る舞いを示す場合に異常と判断する．判断基準のルール化が容易ではないという欠点はあるものの，未知の異常へ迅速に対応できるという利点がある．本研究で扱うのは，アノマリ型の異常検出方法である．

アノマリ型の発見手法としては，信号分析や時系列解析などによる方法 [1, 2, 3, 4] が研究されているが，近年，データマイニング分野において，時系列データを文字列によって表現する SAX (Symbolic Aggregation approXimation) [5] が提案され，いくつかの実験においてその有効性が示された．時系列データを文字列で表現することにより，自然言語処理分野で培われた文字列処理，言語解析などの有用な技術が適用可能となり，信号解析に新しい方法論を導くものと考えられる．本研究では，この文字列表現による解析手法の一つとして，マルコフモデルとクラスタリング技術を用いた方法を提案する．マルコフモデルを用いた異常発見は，SAX を提案した Keogh ら [6] によって提案された．マルコフモデルとは，ある時点で出現するアルファベットの確率が，過去のアルファベットの出現に依存して決まるような確率モデルである．正常な通信データ (訓練データ) を用いてモデルを生成することで，任意の部分文字列の出現回数の期待値が計算できる．検証したい通信データ (テストデータ) における出現回数と比較することで，異常かどうかを判断する．以上が，マルコフモデルを用いた異常発見方法の基本的なアイデアである．Keogh らはこの方法をオランダの電力需要データに適用した結果，Wavelet 変換や免疫機構を用いた従来手法では発見できなかった異常を見つけ出すことができることを示した．ただし，彼らが発見したのは，祝日に電力需要が減るといった程度のものである．ネットワーク上の異常通信を発見するにはより細かく，複雑なパターンを見つけ出す必要があると考えられ，それに伴い誤検知も生じることが予想される．実際に我々がこの方法を用いて行った予備実験では，誤検知，つまり正常であるのに異常であると判定されるデータが多く含まれてしまうことが分かった．

このため，我々はマルコフモデルによって列挙されたデータを異常候補集合として扱い，この集合にクラスタリングを適用することで，より異常データである可能性の高いものを絞り込むこととした．正常データには類似性があるためクラスタを形成しやすい，よってクラスタに取り込まれないはずれ値を見つけ出すことが目的となる．本研究では，はずれ値検出に適した階層型クラスタリングを用いる．マルコフモデルを用いなくてもクラスタリングを適用することは可能であるが，一般に階層型のクラスタリングの計算コストは高く，データ長が大きくなると部分文字列数も多くなり，莫大な計算時間を要してしまうことになる．マルコフモデルによる方法でクラスタリング対象を選別することで現実的な計算時間での処理を可能とすることができる．このように，2 手法を組み合わせることで，相互の欠点を補うことが期待される．

我々の提案手法は主に次の 3 つの処理からなる．

1. SAX によるネットワークトラフィックデータの文字列表現への変換.
2. マルコフモデルを用いた異常系列候補の列挙.
3. 階層型クラスタリングによる異常系列候補からのはずれ値検出.

以下の各章において，上記の処理それぞれについて詳しく説明する．

2　SAX による時系列データの変換

SAX は，離散フーリエ変換，離散ウェーブレット変換，特異値分解などと同じく時系列データの表現方法の一つである．SAX の特徴として，変換後に計算した 2 データ間の距離が元データの距離の下限となることが保障されるような距離尺度を定義できることが挙げられる．これにより，変換後の距離の相対的大小が変換前と比べて保存されるため，タスクによってはデータ処理に必要な計算量の削減に役立つ．また，変換後のデータは文字列によって表現されるため，自然言語処理分野における多くの文字列処理アルゴリズムが適用可能となる．

以下，Lin ら [5] に従って SAX の概要を説明する．まず最初に，Piecewise Aggregate Approximation (PAA) と呼ばれる操作を行う．PAA は，元の時系列データの時間軸を等間隔に区分し（図 1 左），各区間に含まれる値を平均化した値に置き換える（図 1 真ん中）．

今，時系列データ $X = x_1, x_2, ..., x_n$ があり，これを w 次元のデータ $\bar{X} = \bar{x}_1, \bar{x}_2, ..., \bar{x}_w$ に変換することを

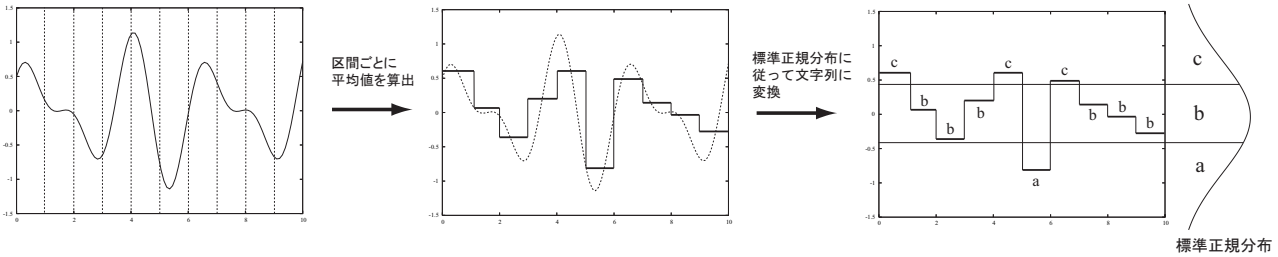

図 1　SAX による文字列への変換

考える．変換後の値は以下の式によって計算される．

$$\bar{x}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} x_j \tag{1}$$

PAA は単純な次元圧縮方法ではあるが，様々な実験において効果の高い方法であることが分かっている [5]．また，サンプリング数が膨大で扱えない場合などは，この変換による次元圧縮が必要となる．

次に PAA によって得られた値を実数値からアルファベット記号に変換する．SAX では，変換後の各記号が等確率で出現するように数値軸の分割区間を決定する．これを行うため，まず数値軸上に分布する標準正規分布を仮定し，分割区間の累積確率が等しくなるような数値軸上の分割点を離散化数（＝アルファベットサイズ）に応じて決めておく．次に PAA 変換後のデータを正規化し，正規化後の各値がどの分割区間に含まれるかを調べ，各区間に一意に割り当てられたアルファベットに変換する．図 1 右は，アルファベットサイズが 3 の場合の変換の様子を示している．最終的に図 1 の時系列データは，文字列 “cbbcacbbb” に変換される．この分割点は，アルファベットサイズ数ごとに予め計算しておくことができる．例えば，アルファベットサイズを 5 に設定した場合，4 つの分割点 $\beta_1 \sim \beta_4$ は，$(\beta_1, \beta_2, \beta_3, \beta_4) = (-0.84, -0.25, 0.25, 0.84)$ のように求めておくことができる．

Lin らは，SAX 表現に適した距離関数を以下のように定義している．

$$MINDIST(Q, C) = \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^{w} (dist(q_i, c_i))^2} \tag{2}$$

Q と C は，SAX 変換後の時系列データである．ここで関数 $dist(q, c)$ は，以下のルールによって計算される値である．

$$v_{i,j} = \begin{cases} 0 & if\ |i-j| \leq 1 \\ \beta_{\max(i,j)-1} - \beta_{\min(i,j)} & otherwise \end{cases} \tag{3}$$

$v_{i,j}$ はアルファベットを辞書順に並べた場合の i 番目と j 番目のアルファベット間の最小距離を表している．例えば，アルファベットサイズ 5 の場合の a と d の距離は，a が 1 番目のアルファベット，d が 4 番目のアルファベットであるので，分割点 $(\beta_1, \beta_2, \beta_3, \beta_4) = (-0.84, -0.25, 0.25, 0.84)$ を用いて，$dist(a, d) = v_{1,4} = \beta_3 - \beta_1 = 1.09$ となる．

Keogh らは，文字列による表現では，データの表現の誤差は離散化の粒度でコントロールできることを強調している．すなわち，先に述べたように式 (2) で計算した距離 $MINDIST(Q, C)$ は，Q と C の元のデータのユークリッド距離の下限と等しい．離散化数を増やし，離散化粒度を細かくしていくにつれ，$MINDISTD(Q, C)$ はユークリッド距離に近づいていく．注目すべき事実として，多くの時系列データを用いて類似度計算を行った実験結果によると，粒度がある程度荒くても，ユークリッド距離と同等，時にはそれ以上の性能を示すことが分かっている [5]．

3　マルコフモデルによる異常系列候補の列挙法

SAX によって時系列データを文字列に変換することにより，自然言語処理分野における様々な解析手法の適用が可能となる．本研究の目的は，時系列データ中の異常トラフィックが含まれる区間を特定することであるので，変換後の文字列データから異常トラフィックと対応する部分文字列集合を抽出することが必要と

なる．ここでは，Keogh らが提案したマルコフモデルに基づく抽出方法を適用する．この抽出方法の基本的なアイデアは，正常なトラフィックデータを用いて生成されたマルコフモデルにより，検証したい部分文字列の出現回数の期待値を計算し，実際の出現回数がこれをどれくらい上回っているかにより異常の度合いを判断するというものである．大幅に上回っている場合に異常と判断する．

マルコフモデルでは，ある時点における記号の生起確率が直前に生起した記号列に依存すると考える [7]．特に長さ M の記号列に依存するようなモデルは M 重マルコフモデルと呼ばれる．今，変換後の文字列データに含まれる長さ k の任意の部分文字列を $c_{[1,k]} = c_1, c_2, ..., c_k$ とすると，$c_{[1,k]}$ が文字列中の任意の位置に出現する確率 $P(c_{[1,k]})$ は以下のように表せる．部分文字列 $c_{[1,M]}$ の生起確率 $P(c_{[1,M]})$ と長さ M の部分文字列 $c_{[i,i+M-1]}$ が出現した後に文字 c_{i+M} が出現する条件付き確率 $P(c_{i+M}|c_{[i,i+M-1]})$ を用いて，以下のように表せる．

$$P(c_{[1,k]}) = \prod_{i=1}^{k-M} P(c_{[1,M]})P(c_{i+M}|c_{[i,i+M-1]}) \quad (4)$$

この確率は文字列中の任意の位置における確率であるから，長さ n の記号列中における出現回数の期待値 $E(c_{[1,k]})$ は以下のように計算できる．

$$E(c_{[1,k]}) = (n-k+1) \prod_{i=1}^{k-M} P(c_{[1,M]})P(c_{i+M}|c_{[i,i+M-1]}) \quad (5)$$

よって，長さ M の任意の文字列 $c_{[1,M]}$ について，その生成確率 $P(c_{[1,M]})$ と $c_{[1,M]}$ の後に任意の文字 c が出現する条件付確率 $P(c|c_{[1,M]})$ が分かっていれば，上記の期待値が計算できる．一般にこれらの確率は予め分からないので，訓練データを用いて推定する．本研究では正常なトラフィックデータを変換した文字列 r を用いて推定する．

$P(c_{[1,M]})$ の推定値 $\hat{P}(c_{[1,M]})$ と $P(c|c_{[1,M]})$ の推定値 $\hat{P}(c|c_{[1,M]})$ はそれぞれ以下のように推定できる．

$$\hat{P}(c_{[1,M]}) = \frac{f_r(c_{[1,M]})}{n-M+1} \quad (6)$$

$$\hat{P}(c|c_{[1,M]}) = \frac{f_r(c_{[1,M]}c)}{f_r(c_{[1,M]})} \quad (7)$$

ここで，$f_r(s)$ は文字列 r 中に部分文字列 s が出現した回数である．

モデルが生成できたので，異常系列とみなす部分文字列の列挙作業を行う．異常があるかどうかを調べたいトラフィックデータを変換した文字列 x をテストデータとし，x 中の任意の部分文字列 s について，以下の式で表されるスコアを計算する．

$$score(s) = f_x(s) - \alpha E(s) \quad (8)$$

このスコアの高いもの程，異常系列である可能性が高いとみなす．$E(s)$ は式 (5), (6), (7) を用いて計算した値である．また，$\alpha = \frac{|x|-|s|+1}{|r|-|s|+1}$ である．

一般に，M の値を大きくすればモデルの精度を上げることができると考えられるが，M の値を大きくしすぎると推定すべき状態数が増えるため，式 (6)，(7) の推定誤差が増え，逆にモデルの精度を下げてしまうこともある．例えば，文字列変換の際に用いるアルファベット数を 8，M を $5^8 = 262144$ 通りもの条件付確率の推定を行う必要がある．しかし実際には，訓練データの不足など，推定を十分に行えない場合もあるため，スムージング処理などの後処理を行う．ちなみに Keogh らが提案した方法では，スムージング処理を行わず，部分文字列ごとに式 (5) の値が 0 とならないような最大の M を決定することで，できるだけ M の値を大きくする方法をとっている．M の値を変化させるため様々な長さの部分文字列の出現頻度を計算する必要があるが，任意長の文字列の出現回数を高速に計算できる suffix tree と呼ばれるデータ構造を用いることで実用的な時間での計算が可能としている．

マルコフモデルによってトラフィックデータを完全にモデル化できるわけではないので，この方法を適用する場合にはモデルによる推定値には誤差があることを念頭に置かなければならない．後述の実験の部分でも述べるが，我々がこの方法を実際のトラフィックデータに適用したところ，頻度が高く，かつ同じ文字が連続するタイプの文字列について実際の出現回数と期待値に大きなずれが生じることが分かった．例えば “bbbbb” や “cbbbbbbbbbbbbbb” などの文字列が該当するが，これらの文字列は定常状態を表しており，異常とは呼べない．マルコフモデルによる方法では，これらの定常状態を示す文字列が上位に散らばったようにランクされてしまうため，それらを取り除く手段が必要となる．

4 階層型クラスタリングによるはずれ値検出

我々は，マルコフモデルを用いて選んだ部分系列集合から更に異常系列を絞り込むため，クラスタリングを適用することを新たに提案する．一般にクラスタリングでは，多数集合を形成するクラスタを発見することが目的であるが，異常発見においては，逆にクラスタを形成しないはずれ値を見つけ出すことが目的となる．このはずれ値を最終的な異常系列とみなす．

クラスタリングには主に分割型と階層型のものがある．分割型クラスタリングを用いてはずれ値を検出するには，クラスタ中心とクラスタ中心からどれくらい離れていたらはずれ値とみなすかを決定する閾値を決定する必要があるが，この決定は一般に難しい．一方，階層型クラスタリングでは，類似度の高い 2 つのクラスタを順に結合していくだけで残っていくはずれ値が決定される．階層型クラスタリングには，クラスタ間距離の計算方法の違いによって，最短距離法，最長距離法，群平均法，重心法，加重平均法，加重重心法，ウォード法など様々な方法があるが，我々が対象とするデータは文字列であるため，重心の計算ができないことなどを考慮して，群平均法による方法を用いる．この方法では，2 つのクラスタに属するメンバ間のすべての組み合わせの距離を求め、その平均値をクラスター間の距離とする．メンバ間の距離計算には，式 (2) を用いる．今 2 つのクラスタ $S = \{s_1, s_2, ..., s_l\}$, $T = \{t_1, t_2, ..., t_m\}$ があるとすると群平均法では，クラスタ間距離 $D(S,T)$ を以下のように計算する．

$$D(S,T) = \frac{\sum_{i=1}^{l}\sum_{j=1}^{m} MINDIST(s_i, t_j)}{l+m} \qquad (9)$$

長さの異なる文字列間では距離が計算できないので，実際にマルコフモデルによって列挙される部分文字列集合 S にクラスタリングを適用する場合には，まず S を同じ長さ k の文字列からなる部分集合 S^k に分類し，各 $S^k = \{s_1^k, s_2^k, ..., s_p^k\}$ についてクラスタリングを適用する．以下にクラスタリングによる異常発見の手順についてまとめる．

1. 初期クラスタ集合 $C^0 = \{c_1^0, c_2^0, ..., c_p^0\}$ の各要素クラスタに S^k の各要素 s_i^k をメンバーとして割り当て，$c_i^0 = \{s_i^k\}$ とする．
2. C^j の各要素クラスタ全ての組み合わせについて，式 (9) を用いてクラスタ間距離を計算し，距離が最小となる 2 つのクラスタを統合する．統合後のクラスタ集合を C^{j+1} とする．このとき，2 つのクラスタの内，要素数が 1 であるものをはずれ値リスト O に追加する．
3. 2 を要素数 1 のクラスタがなくなるまで繰り返す．
4. O に追加された順に異常度の低い文字列とみなす．つまり最後に追加された文字列ほど，異常データである可能性が高いものとみなす．

クラスタ統合時に 2 つのクラスタが共に要素数 1 である場合は，その 2 つのクラスタのはずれ値リストにおける順位は同じであることとする．

5 実験

5.1 設定

実験用の時系列データとして，豊橋技術科学大学における学外向け通信パケット約 3 週間分から送信先ホスト数を 1 分毎にまとめたものを利用した．藤井ら [8] は，P2P ソフト及びウイルスによる攻撃活動を異常トラフィックと位置づけ，これを検知する際の特徴量として送信先ホスト数が有効であることを示している．本研究の目的も同様なので同じ特徴量を用いる．図 2 と図 3 は，送信先ホスト数の推移をグラフ化したものである．図 2 の方は，マルコフモデル生成時の訓練データとして用いるもので，トラフィックの比較的安定した 9 日分を選んだ．図 3 の方はテスト用の 10 日分のデータである．訓練用データと比較して，値が急激に増加している箇所がいくつかあり，これらを検出することが目的となる．それぞれに予め SAX を適用し，文字列化しておく．時間軸方向の平均化は行わず，アルファベットサイズは 8 とした．

5.2 実験結果 1：マルコフモデルによる異常発見

表 1 は，マルコフモデルを用いた方法で，異常系列を文字列長別に上位 5 個を列挙したものである．モデルの次数 M は 2 に固定し，スムージングは行ってい

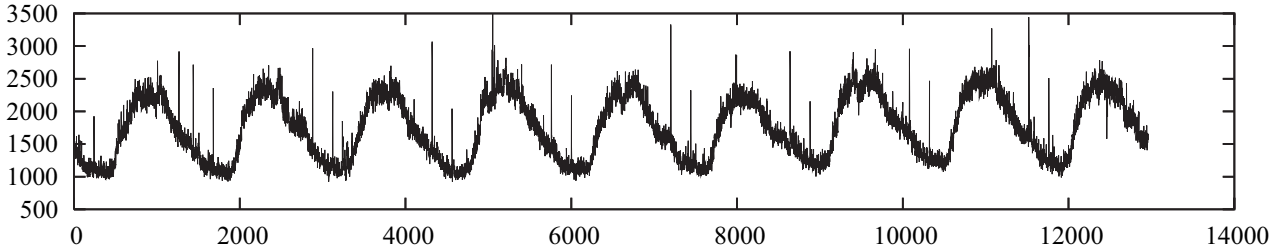

図 2　送信先ホスト数の推移（訓練データ）

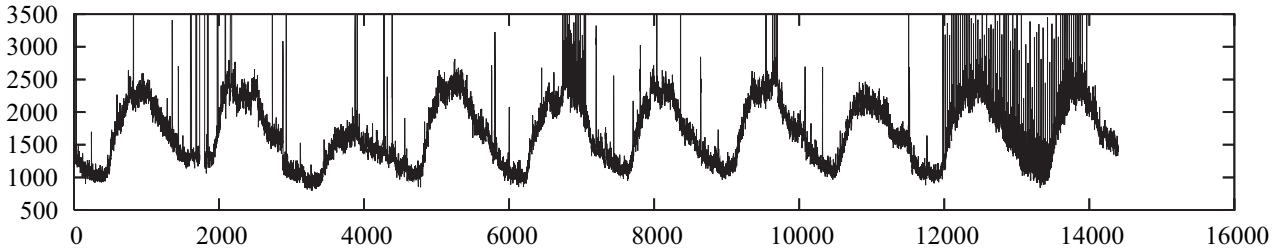

図 3　送信先ホスト数の推移（テストデータ）

ない．文字列横の数値はそれぞれ，テストデータにおける出現回数と訓練データから生成されたモデルによる期待値である．表から分かるように，どの文字列長においても同じ文字が連続したものが上位にランクしている．しかしながら，出現回数を見ても分かるようにこれらの文字列は異常系列ではなく正常な系列である．この傾向は，マルコフモデルの次数 M を大きくしても同様に観測され，マルコフモデルによる方法だけでは，異常系列を効率的に発見できないことを示している．

5.3　実験結果 2：階層型クラスタリングによるはずれ値検出

次に，マルコフモデルにより列挙された候補の上位 500 個の文字列にクラスタリングを適用した結果を表 2 に示す．クラスタリングによって検出された上位 5 個のはずれ値とその出現位置（文字列によってはその一部）を示している．出現位置とは，図 3 における時間軸における位置のことをいう．各文字列の出現位置付近でデータが実際にどのように推移しているかを調べるため，図 3 のグラフを使って，出現位置ごとに矢印を付けたものが図 4 である．図から，検出された文字列は主に 4 箇所を指し示していることが分かる．4 箇所ともホスト数が急激に増加しており，実際にダンプデータを解析すると，P2P ソフトまたはウイルスによる通信異常である可能性が高いものばかりであることが分かった．

表 2 に示された文字列は，マルコフモデルによる方法ではいずれも低い順位に位置している．例えば文字列 “hhehh” は，マルコフモデルでは 319 位に，文字列 “bbbbbbbbhb” は 125 位に位置している．少なくとも我々が用いたデータでは，マルコフモデルによる方法のみでは発見効率が悪く実用性に欠ける．本節で示したように，クラスタリングを行うことでマルコフモデルによる誤検知を取り除くことが期待でき，発見効率の向上に役立つと考えられる．

6　まとめ

本研究では，時系列データの文字列表現手法を用いてネットワーク上の異常トラフィックを自動的に発見する方法として，マルコフモデルによる異常候補の列挙とクラスタリングによるはずれ値検出を組み合わせた方法を提案した．時系列データの文字列変換方法には SAX を用い，SAX 表現からマルコフモデルに基づく文字列生成モデルを構築する方法について説明した．また，郡平均法による階層型クラスタリングを用いて

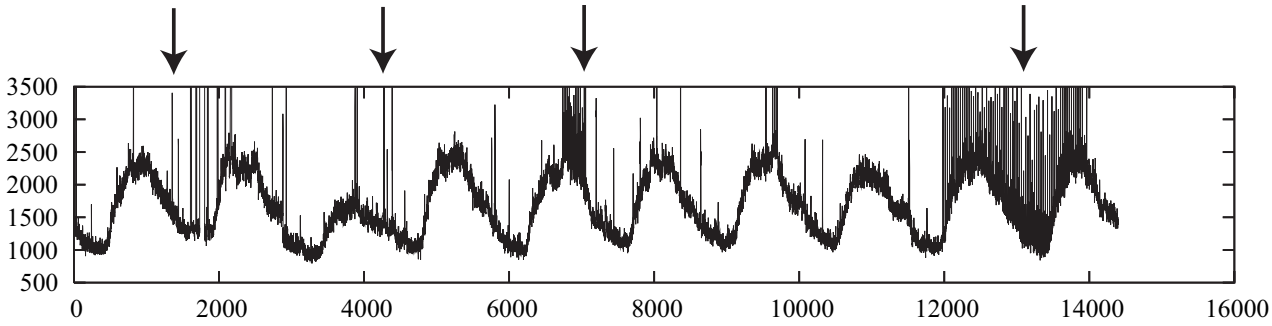

図 4　テストデータ中の異常が検出された位置

表 1　マルコフモデル ($M = 2$) による異常系列候補

順位	文字列長 5	出現回数	期待値
1	bbbbb	2345	668.90
2	ccccc	702	172.22
3	ddddd	571	132.22
4	fffff	525	101.11
5	eeeee	291	51.11
順位	文字列長 10	出現回数	期待値
1	bbbbbbbbbb	1723	129.26
2	cccccccccc	251	9.40
3	gggggggggg	263	55.76
4	dddddddddd	178	5.85
5	ffffffff	107	3.75
順位	文字列長 15	出現回数	期待値
1	bbbbbbbbbbbbbbb	1387	27.68
2	gggggggggggggg	126	7.29
3	ccccccccccccccc	101	0.63
4	ddddddddddddddd	59	0.30
5	cbbbbbbbbbbbbbb	41	3.55
順位	文字列長 20	出現回数	期待値
1	bbbbbbbbbbbbbbbbbbbb	1175	5.93
2	gggggggggggggggggggg	62	0.95
3	ccccccccccccccccccc	48	0.04
4	cbbbbbbbbbbbbbbbbbbb	25	0.76
5	bbbbbbbbbbbbbbbbbbbc	24	0.73

表 2　クラスタリング適用後の順位

順位	文字列長 5	出現位置
1	hhehh	7038
2	fcfcb	13202
3	cfcfc	13201
4	hfhgc	4389
5	fhhhe	7010
順位	文字列長 10	出現位置
1	bbbbbbbbhb	13116 13146 13176
2	bbbhbbbbbb	13121 13151 13181
3	bbhbbbbbbb	13122 13152 13182
4	bbbbbbbhbb	13117 13147 13177
5	bbbbhbbbbb	13120 13150 13180
順位	文字列長 15	出現位置
1	bbbbbbbbhbbbbbb	13116 13146 13176
2	bbbbbbbbbbhbbbb	13114 13174 13234
3	bbbbbbbbbbhbbbb	13115 13175 13235
4	bbbbbbbbbbbbhbb	13113 13173 13233
5	bbbbbhbbbbbbbbb	13149 13179 13239
順位	文字列長 20	出現位置
1	ccccchhhchhhcccdccc	4269
2	cccchhhchhhcccdcccc	4270
3	cccchhhchhhcccdccccd	4271
4	ccdhgfegddcccccdchcc	12957
5	cccchhhhhhhhhhhhhhh	1731

はずれ値を順位付けする方法を提案した．実際のトラフィックデータを用いて行った実験により，マルコフモデルを用いた方法のみでは誤検知が多いことを示し，クラスタリングを適用することで誤検知を取り除く効果が期待できることを示した．

ただし，本稿で発見した異常は飛びぬけて高い値を持ち，グラフ化した場合に人目により発見が容易なものばかりである．このため本手法が人手では見分けのつきにくい異常を発見できるかどうかより詳しく調べていく必要がある．また，定量的な性能評価も今後の課題である．

参考文献

[1] Lazarevic, A., Ertoz, L., Ozgur, A., Srivastava, J. and Kumar, V. “A comparative study of anomaly detection schemes in network intrusion detection.” In *Proceedings of the 3rd SIAM Conference on Data Mining* (2003)

[2] Yamanishi, K. and Takeuchi, J. “A Unifying Approach to Detecting Outliers and Change-Points from Non-stationary Data.” In *Proceedings of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pp.676–681 (2002)

[3] Dasgupta, D. and Forrest, S. “Novelty detection in time series data using ideas from immunology.” In *Proceedings of the Int'l Conf. on Intelligent Systems* (1999)

[4] Shahabi, C., Tian, X. and Zhao, W. “Tsa-tree: A wavelet-based approach to improve the efficiency of multi-level surprise and trend queries.” In *Proceeding of the 12th Int'l Conf. on Scientific and Statistical Database Management* (2000)

[5] Lin, J., Keogh, E., Lonardi, S. and Chiu, B. “A Symbolic Representation of Time Series, with Implications for Streaming Algorithms.” In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp.2–11 (2003)

[6] Keogh, E., Lonardi, S and Chiu, W. “Finding Surprising Patterns in a Time Series Database In Linear Time and Space.” In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*, pp.550–556 (2002)

[7] 北 研二, “確率的言語モデル”, 東京大学出版会 (1999)

[8] 藤井聖, 中村豊, 藤川和利, 砂原秀樹, “通信先ホスト数の変化に注目した異常トラフィック自動検出手法の提案と評価”, 電子情報通信学会論文誌 Vol.J88-B, No.10, pp.1922–1933 (2005)