Change Sign Detection with Two-Stage MDL Change Statistics
(2段階MDL変化統計量による変化予兆検知の研究)

数理情報学専攻 48-196233　結城 凌
指導教員　山西 健司 教授

1　はじめに

本研究では，変化予兆検知という問題に 2 段階 MDL 変化統計量でアプローチする．変化予兆検知の前身となる問題として，変化検知という問題がある [1]．これは，時系列データを考えた際に，データ発生の確率分布が変化する時刻を検知する問題を指す．ここで，変化は，1) 確率分布のパラメータ，2) 確率分布自体が持つ自由パラメータの数や潜在構造 (構造的変化と呼ぶ) の二種類に分けることができる．いずれの変化も，その背後に人間にとって重要なイベントに結びつくことが多く，実用的に重要であることから機械学習・データマイニングの文脈で古くから研究されてきた．例えば，web アクセスログが SQL インジェクションに対応すること [2] や，顧客の購買トランザクション [3] における類似購買パターンが市場トレンドに対応することが報告されている．従来変化検知に関する研究では，統計的に扱いやすいことから，突発的な変化の発生が頻繁に仮定されていた．しかし，現実の状況下においては，変化の漸進的発生を仮定する方が自然であり，この場合従来手法では漸進的変化の検知遅延・見逃しが起こりうる．そのため，漸進的変化開始点検知や，より一般に変化予兆の検知を試みる研究が増えてきている [3, 4, 5]．[5] で指摘されているとおり，変化予兆検知の目的は二つある．一つ目は，変化予兆を捉えることで，その直後に起こる変化を予測することであり，二つ目はデータが一括で与えられた下で既に発生した変化の原因を解析するためである．

本研究では，パラメータ変化および構造的変化において漸進的変化を仮定し，その開始点を変化予兆と考える (図 1)．そして，2 段階の変化検知で漸進的変化の開始点を検知する，2 段階 MDL 変化統計量を提案する．

2　従来研究

2 段階 MDL 変化統計量は MDL 変化統計量 [2, 6] を基にしており，まずこちらを説明する．$\mathcal{F} = \{p(X;\theta) : \theta \in \Theta\}$ を確率変数 X の従う確率分布のクラスとする．$\theta \in \Theta$ は確率分布のパラメータである．時系列データ

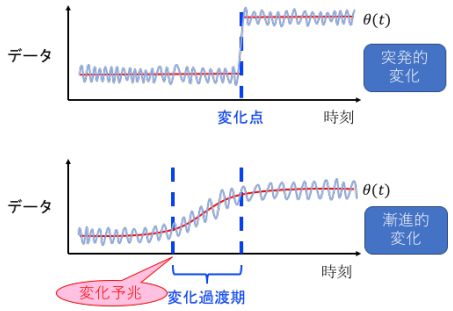

図 1. パラメータ変化における，突発的変化と漸進的変化の例

$x_1^n = x_1, \ldots, x_n$ が与えられており，時刻 t が変化点のただ一つの候補点であるとする．MDL 原理 [7] に基づき，与えられたデータを 1) 変化がないと仮定した時のデータの記述長，2) 時刻 t が変化点であると仮定したときのデータの記述長，を比べ変化点かどうかを検定する．具体的には，与えられた正のパラメータ ϵ を用いて

$$\Phi(x_1^n)_t \stackrel{\text{def}}{=} L(x_1^n) - \{L(x_1^t) + L(x_{t+1}^n)\} - n\epsilon,$$

を時刻 t における MDL 変化統計量とし，これが 0 を超過した場合にのみ変化点であると判断する．ここで，$L(\cdot)$ はデータの記述長を出力する関数であり，NML 符号長 [8] が良好な性質をもつのでよく用いられる．また，パラメータ・構造的変化の両方における MDL 変化統計量の計算方法はすでに確立されている [2, 6].

3　提案手法

提案手法である 2 段階 MDL 変化統計量について説明する．1 段階目では，change probability p_{change}^t という各時刻が変化点である確率をモデリングした統計量を計算する．これが 1 段階目の変化検知に対応する．2 段階目では，change probability の負の対数の増加率 $r_t = (-\log p_{change}^t)/(-\log p_{change}^{t-1}) - 1$ を計算し，この統計量に対し MDL 変化統計量を用いて再度変化検知を行う．パラメータ・構造的変化両方において，漸進的変化の過渡期に入ると r_t の分散が時間に沿って増加することを確認し，その分散の変化を検知することで漸進的変化の開始点を検知を実現する．まず時刻 t での change probability を，与えられた正のパラメータ

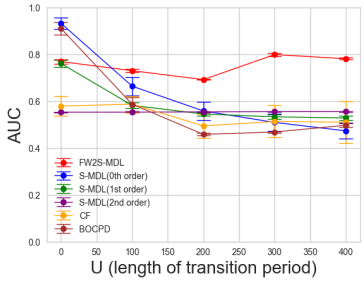

図 2. 誤検知率と遅延の曲線下面積．赤線が提案手法であり，他色は対抗手法．変化過渡期の長さが長い時，提案手法が最も良い性能を出すことが確認できた．縦軸: 曲線下面積．横軸: 変化過渡期の長さ．

β を用いて次の形で定義する．

$$p^t_{change} = \frac{\exp(-\beta(L(x^t_1) + L(x^n_{t+1})))}{\exp(-\beta L(x^n_1)) + \exp(-\beta(L(x^t_1) + L(x^n_{t+1})))},$$

ここで，change probability はパラメータ・構造的変化のどちらにおいても定義できることに注意したい．次に，change probability の負の対数の増加率 $r_t = (-\log p^t_{change})/(-\log p^{t-1}_{change}) - 1$ という新たな統計量を考える．変化過渡期になると，この統計量の分散が増加することが，実験的，さらに簡単な数理的考察から確認した．r_t が 1 変数ガウス分布に従うことを仮定した上で，MDL 変化統計量を用いて

$$\Phi^t_{2-stage} \stackrel{\text{def}}{=} L(r^n_1) - \{L(r^t_1) + L(r^n_{t+1})\} - n\epsilon,$$

をシステムの最終的な出力とする．

4 実験結果

提案手法および対抗手法を人工データおよび実データを用いて評価した．人工データでは，パラメータ変化および構造的変化の両方において，漸進的変化を複数持つデータセットを生成した．パラメータ変化においては，独立ガウス分布のパラメータが変化するデータセットを生成し，さらに漸進的変化の変化過渡期の長さを変化させた．ここで，変化過渡期前後のパラメータは固定したため，変化過渡期が長くなるほど変化予兆の検知は難しくなることが予想される．図 2 は提案手法および対抗手法の結果の一部である．変化が突発的変化に近い場合，対抗手法が高い性能を出すことが出来ているものの，変化過渡期が長くなるにつれて，提案手法がより高い性能を出せていることが確認できる．構造的変化においては，混合ガウス分布に従い，混合数が漸進的に変化するデータを，漸進的変化の特徴を変えつつ複数種類生成した．こちらにおいても同様に，提案手法が最も良い性能を出すことを確認できた．

実データにおいては，Dow Jones 平均株価，工場における生産ラインの多次元データ，感染流行データという多様な性質を持つデータセットにおいて提案手法を適用した．それらのデータセットにおいて，現実でのイベント (平均株価では政治的イベント，感染流行データでは感染流行対策など) と提案手法があげたアラームを比較した結果，イベントの予兆と考えられるアラームを複数上げていることが確認できた．

5 まとめ

本研究ではパラメータ・構造的変化を対象にした変化予兆検知手法，2 段階 MDL 変化統計量を提案した．また，実用的なアルゴリズムを 2 種類提案した．人工データによる実験では，提案手法が定量的に良い性能を示せることを確認し，実データでは，様々な性質を持つデータにおいて提案手法が良い変化予兆検知を行えることを確認した．今後の展望としては，提案手法の予兆検知の意味での性能保証解析などが考えられる．

参考文献

[1] 山西健司編著. **データサイエンスの数理 数理で読み解くデータの価値 数理科学** *2019* 年 *06* 月号. サイエンス社, 2019 年 6 月.

[2] Kenji Yamanishi and Kohei Miyaguchi. Detecting gradual changes from data stream using mdl-change statistics. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 156–163. IEEE, 2016.

[3] So Hirai and Kenji Yamanishi. Detecting model changes and their early warning signals using mdl change statistics. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 84–93. IEEE, 2019.

[4] Kohei Miyaguchi and Kenji Yamanishi. Online detection of continuous changes in stochastic processes. *International Journal of Data Science and Analytics*, 3(3):213–229, 2017.

[5] Kenji Yamanishi and So Hirai. Detecting signs of model changes with continuous model selection. *submitted to IEEE Transactions on Knowledge and Data Engineering*, 2020.

[6] Kenji Yamanishi and Shintaro Fukushima. Model change detection with the mdl principle. *IEEE Transactions on Information Theory*, 64(9):6115–6126, 2018.

[7] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[8] Yurii Mikhailovich Shtar'kov. Universal sequential coding of single messages. *Problemy Peredachi Informatsii*, 23(3):3–17, 1987.