

コンセプトチェンジを考慮した回帰分析とフィルタリングへの応用

A Regression Method Adapted to Concept Change and its Application to Filtering

島崎 智史† 安村 禎明‡ 関 和広† 上原 邦昭†
Satoshi Shimazaki Yoshiaki Yasumura Kazuhiro Seki Kuniaki Uehara

あらまし

連続的にデータが取得されるストリーム環境では、データの特性が時間的に変化するコンセプトチェンジという特徴があり、この変化への適応が重要な課題となっている。本稿では、データストリームを対象とする回帰分析において、コンセプトチェンジに適応するための手法を提案する。具体的には、データストリームをチャンクというまとまりに分割し、学習における訓練事例の範囲を更新し続けることで回帰分析をコンセプトチェンジに適応させる。そして、人工データ及び大規模なテキストデータを用いたフィルタリング問題に本手法を適用し、その有効性を評価する。

1. はじめに

近年、高速ネットワークと大規模センシング技術の発達に伴い、流れてくる大量のデータをリアルタイムで分析し、そこから有用な規則やパターンを発見するデータストリームマイニングの研究が盛んになっている。連続的にデータが取得されるストリーム環境では、データの特性が時間的に変化するコンセプトチェンジという特徴がある。コンセプトチェンジには、急激な変化と緩やかな変化の2種類が存在する。データストリームマイニングを行う上で、これらのコンセプトチェンジへの適切な処理は重要な課題となっている。また、従来のデータストリームマイニングでは数値推定に有効な回帰分析についての研究はあまり進んでいない。

回帰分析をコンセプトチェンジに適応させるための既存手法である FIMT [1]や、その改良手法である FIRT-DD [2]では、推定値と実値との誤差を記憶しながら訓練データの範囲を拡大していく時間窓の概念を導入している。しかし、この方法では、誤差の増加が激しい急激な変化は検出できるものの、誤差の増加が少ない緩やかな変化の検出は難しい。また、これらの既存手法は、変化を検出した際、学習した回帰木の枝狩りを行うことで不要な事例の除去し、変化への適応を試みている。しかし、枝狩りだけでは不要な事例を全て除去することができず、変化への適応も迅速かつ正確ではない。

そこで本稿では、複数の回帰木を用いるアンサンブル手法によってこれらの問題の解決を目指す。まず、時間窓の概念ではなく、チャンクと呼ぶ概念を利用する。チャンクとは、データストリームを一定の大きさに分割したまとまりである。本手法では、時間窓のように訓練データを増やすのではなく、学習に用いるチャンクの移動によって訓練データを推移させることで緩やかな変化への適応を可能とする。また、学習に用いるチャンクは複数であり、それぞれのチャンクから異なる回帰木を生成する。これらの複数の回帰木でアンサンブル学習を行う

ことで、精度の向上を試みる。そして、急激な変化を検出した際には、変化前の不要なデータを含むチャンクを学習から切り離すことで、コンセプトチェンジへの即応を目指す。

2. 提案手法

2.1 概要

前述した既存手法の問題点の解決を目指し、回帰分析にアンサンブル手法を組み合わせる。図 1 にアンサンブル手法の概要を示した。まず、既に取得したストリームデータを新しいものから一定数の事例集合（チャンク）に分割する。分割したチャンクの内、利用するチャンク 1 つにつき回帰木を 1 つずつ構築し、最大 K 個の回帰木によってアンサンブル学習を行う。回帰木が 1 つの時はアンサンブル学習を行わない。

この手法では緩やかなコンセプトチェンジへの対応を考え、時間窓ではなくチャンクの推移によって、常に新しいデータのみを訓練に使用する。急激な変化を検出した際には、最新のチャンク以外の回帰木を削除することでコンセプトチェンジへ即応させる。また、複数の回帰木によってアンサンブルを行うので、各回帰木の訓練事例数が少なくなった為に精度が落ちることもなく、精度は向上するものと考えられる。そして、増加し続けるデータを生データでなく学習器として保存するため容量の節約になるほか、各回帰木が生成された時点における事例の分布情報を保持しているため、分布を考慮した情報の取捨選択を行うことができるという利点もある。

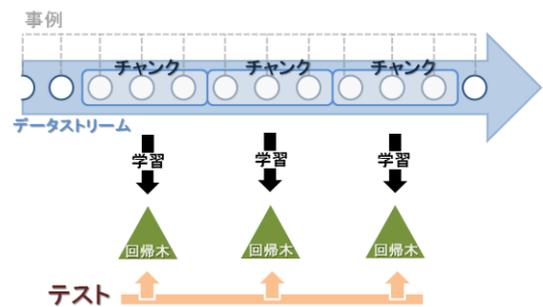


図 1: アンサンブル手法の概要

2.2 アンサンブル手法

以下では、データストリームにおいて既に取得したデータは全て訓練に利用できるものとする。訓練に使用可能な事例数を T 、チャンク 1 つの事例数を M 、実際に訓練に使用するチャンク数を k とする。また、初めの事例 M

†神戸大学, Kobe University

‡芝浦工業大学, Shibaura Institute of Technology

個は訓練データとしてのみ利用し、テストには利用しない。アルゴリズムを以下に示す。

1. 新たな事例が到着した際、現在保持している回帰木によって、その事例の目的変数の推定値を算出する。もし木を複数個もっている場合は、後述の重みを利用して推定結果のアンサンブルを行う。
2. 事例毎に推定値と真の値との誤差を記憶しておき、その和と閾値 ϵ を比較する。
(ア) 閾値 ϵ を超えている場合：コンセプトチェンジが発生したと判断し、今までに持っていた木を全て破棄し、最新のチャンクのデータのみを訓練データとする回帰木を生成する。また、今回破棄した事例は今後訓練データとして利用することはない。($T=M, k=1$)
(イ) 閾値 ϵ を超えていない場合：コンセプトチェンジが発生していないと判断し、チャンクを 1 事例分移行させ、保持している各回帰木を更新する。
3. $T > (k+1) \cdot M$ かつ $k < K$ の場合は、さらなる回帰木を作れると判断し、新たな回帰木の生成を行う。

以上の 1 から 3 を繰り返すことで、コンセプトチェンジを含んだデータストリームに対して回帰分析を行っていく。

2.3 複数の木に対する重み付け

アンサンブルを行う際の重み付けについて述べる。それぞれの回帰木は、テスト事例を処理する毎に、真の値 o と推定値 y との差を記憶する。複数の回帰木で推定を行う際には、その差が小さい回帰木ほど重みを大きくすることで、予測精度の向上を試みる。 t 番目の回帰木の平均誤差を g_t とする (式(1))。アンサンブルに使用する回帰木 S 個の内、 t 番目の誤差率 d_t は式(2)で表される。誤差率の大小関係を反転させた値を、それぞれの木の重みとして与える。

$$g_t = \frac{1}{M} \sum_{m=1}^M (o - y)^2 \quad (1)$$

$$d_t = \frac{g_t}{\sum_{i=1}^S g_i} \quad (2)$$

3. 評価実験

本節では、提案手法の有効性を示す実験として FIRT-DD [2] との比較実験を行った。実験には、人工的にコンセプトチェンジを発生させた線形データセットを用いた。新聞記事データ RCV1 [3][4] については、時間の都合上、実験が終了しなかったため、実験の設定についてのみ述べる。

3.1 実験設定

まず人工データセットについて述べる。本研究では、参考文献 [2] で用いられた線形データセットを用いた。事例数は 10,000 個、説明変数 ($x_1 \sim x_5$ [0 ~ 1 の間]) は 5 個である。事例の目的変数 y の値は、式(3)で与えられる。

$$y = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 \quad (3)$$

コンセプトチェンジは説明変数の領域を指定することと、目的変数の式を変化させることで起こした。チャンク 1 つの事例数 M は 200、アンサンブルに用いる回帰木の最大数 K は 3 とした。

次にテキストデータ RCV1 について述べる。これは TREC 2002 Filtering Track で用いられたロイター社のニュース記事である。データは ID:2286 から ID:86967 までを訓練データ、ID:86968 から ID:810597 までをテストデータとする約 800,000 の事例数を持つ。ニュース記事とトピックはそれぞれが組になっている。記事 47,0236 個の説明変数 (単語) を持つ数値ベクトル (TF-IDF 値) に置き換え、目的変数はニュース記事とトピックとの関連性を示す。関連があれば 1、なければ 0 である。トピックは 100 個あり、評価には MAP (Mean Average Precision) を用いる。ここでは、データストリームのある時点において、考慮するトピック数を増減することで、擬似的にコンセプトチェンジを起こす。つまり、新しく興味のあるトピックが加わることによって、ランキング結果にも変化が生じるものと仮定する。また、トピック毎に重みを設けることによって、緩やかな変化と急激な変化を実現する。この場合は、目的関数が 2 値でなくなるので、評価には NDCG (Normalized Discounted Cumulative Gain) [5] を用いる。

3.2 結果と考察

表 1, 2 は人工データにおける実験結果を示している。表 1 は、全事例の 4 分の 1, 2 分の 1 の時点で、目的関数の式を変更し、急激な変化を起こしている。また、全事例を 2 つに分割し、それぞれの説明変数の取る値を指定することで、全体として緩やかな変化を実現している。また、表 2 は全事例の 2 分の 1, 4 分の 3 の時点で目的変数の式を変更し、緩やかな変化を起こしている。

表 1 においては、既存手法が変化への適応が遅いのに対し、提案手法は変化の検出から適応までが迅速であることがわかる。また表 1, 2 と共に、既存手法は緩やかな変化を検出するまでにとっても時間がかかるのに対し、提案手法は緩やかな変化への適応できていることがわかる。

4. おわりに

本稿では、データストリームにおけるコンセプトチェンジへの適応を指向したアンサンブル手法を提案した。チャンクを用いた回帰木の更新によって、緩やかな変化への適応や、急激な変化への迅速な適応を可能にした。

また、複数の回帰木を用いたアンサンブルを行うことで、精度の向上を試みた。人工データにおける実験では、緩やかな変化と急激な変化ともに既存手法よりもよい精度を示し、本手法の有効性が確認できた。

今後の課題は、大規模な実データにおける本手法の有効性を検証することである。本稿で述べた RCV1 データは、説明変数の数が膨大であり、また正例の数が非常に少ない。このようなデータに対しても頑健な学習アルゴリズムを開発することが望まれる。

Rank Using Multiple Classification and Gradient Boosting”, In Proceedings of *Advances in Neural Information Processing Systems*, pp.897-904, 2008.

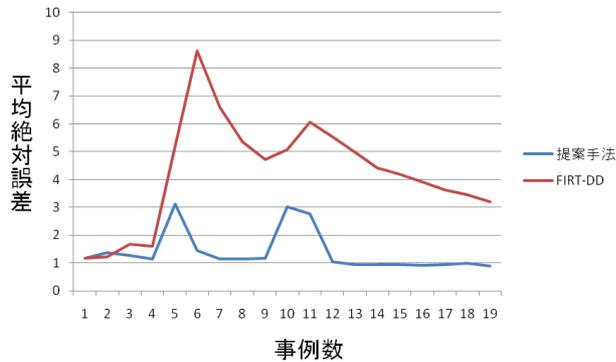


図 2：急激なコンセプトチェンジへの適応

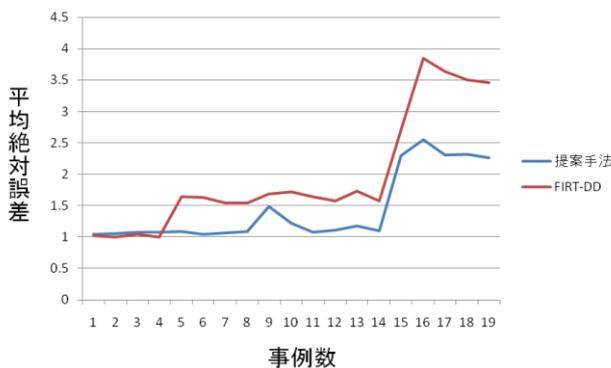


図 3：緩やかなコンセプトチェンジへの適応

参考文献

- [1] Ikonomovska, E., Gama, J, “Learning Model Trees from Data Streams”, In Proceedings of the 12th International Conference on Discovery Science, pp.52-63, 2008.
- [2] Ikonomovska, E., Gama, J, Raquel Sebastiao, Dejan Gjorgjevik, “Regression Trees from Data Streams with Drift Detection”, In Proceedings of the 13th International Conference on Discovery Science, pp.121-135, 2009.
- [3] Stephen Robertson, Ian Soboroff, “The TREC 2002 Filtering Track Report”, In Proceedings of TREC 2002, 2002.
- [4] David D. Lewis, Yiming Yang, “RCV1: A New Benchmark Collection for Text Categorization”, Proceedings of the 14th annual international ACM SIGIR conference on research and development in information retrieval, pp.361-397, 2004.
- [5] Ping Li, Christopher J.C. Burges, “McRank: Learning to