データアクセスパターンに基づくデータマイニング手法の分類

秋岡　明香†　　村岡　洋一‡　　山名　早人‡　　中島　達夫‡
†電気通信大学　　‡早稲田大学

1　はじめに

情報爆発時代が到来し，膨大なデータからデータマイニング等の手法により，新たな知識・知見を高速に抽出することが重要視されている．大量の時系列データ（データストリーム）を高速に分析するストリームマイニングは，特に注目を浴びているが，その更なる高速化手段のひとつとして，クラウドなどの分散環境で並列分散実行する方法がある．

しかし，ストリームマイニングにおいては，一部の処理済みデータへのアクセスや，特徴量によるデータストリームの比較等が頻繁に生じるため，データアクセスパターンを把握せずにはストリームマイニングを効率良く並列分散実行できない．本稿では，各ストリームマイニングアルゴリズムのデータアクセスパターンに基づき，ストリームマイニング手法を分類する．

2　関連研究との比較

Turaga ら [1] は，複数の分類器を直列に連結しデータストリームを処理するシステムを提案し，各分類器のスループットと出力結果の精度を評価したが，分類器は直列に並ぶことが前提であり，処理の一部または全部を並列実行する場合については考察していない．

Thuraisingham ら [2] は，組み込みシステムにおける実時間データマイニングの実現を目的として，ストリームマイニングを含む複数のデータマイニング手法を比較し，データマイニング手法を並列化する上でデータアクセスパターンに注意する必要性を指摘してはいるが，具体的な方針や解決策は提案していない．

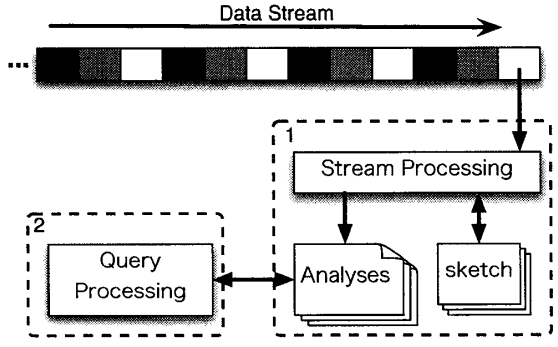


図 1: 一般的なストリームマイニング手法

Data Mining Algorithms Classified Based on Data Access Patterns
†Sayaka Akioka ‡Yoichi Muraoka ‡Hayato Yamana ‡Tatsuo Nakajima
†The University of Electro-Communications
‡Waseda University

個々のアプリケーションや環境に応じた高速化が重要であることは確かである．しかし，ストリームマイニングの重要性が高まる昨今，アルゴリズムのデータアクセスパターンという一般的な視点でアルゴリズムを分類し，高速化の指針を得ることが重要である．

3　ストリームマイニング

図 1 に，一般的なストリームマイニング手法のモデルを示す．本稿では以下のように，ストリームマイニング手法を 2 部（図 1 中の破線枠 1 と 2）に分ける．

- ストリーム処理部（図 1 中の破線枠 1）
  ストリーム処理部は，時系列で次々と到着するデータからスケッチと呼ばれる要約情報を生成する．処理を省メモリかつ高速に実行するため，スケッチ作成時に読み込んだオリジナルデータは破棄し，再利用することはない．充分な量のスケッチを得た後，該当データストリームの特徴量を得る為の計算を行なう．

- クエリ処理部（図 1 中の破線枠 2）
  クエリ処理部は，ストリーム毎の特徴量等の分析結果に基づき，類似シーケンスの検索や特定のデータパターンの検索を行なう．

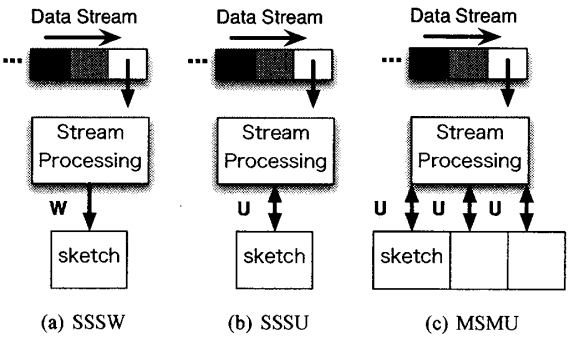
本稿では，代表的なストリームマイニング手法について，ストリーム処理部およびクエリ処理部のデータアクセスパターンを解析し，分類を行なう．

4　アルゴリズムの分類

4.1　ストリーム処理部

ストリーム処理部の主な処理内容は，該当ストリームの特徴量の計算である．したがって，処理はローカルで行なわれ，データアクセスパターンは厳密には計算する特徴量によって異なるが，図 2 のように大きく 3 種に分類できる．

(a) Single Sketch Single Write (SSSW)
Gilbert ら [3] の研究のように単一のスケッチに 1 回書き込む．



(a) SSSW (b) SSSU (c) MSMU

図 2: ストリーム処理部の分類

(b) Single Sketch Single Update (SSSU)
平均値をインクリメンタルに計算するなど，単一スケッチの内容を参照した後で 1 回更新する．

(c) Multiple Sketches Multiple Updates (MSMU)
Zhu ら [4] や Sakurai ら [5] の研究のように，ストリームデータの長期的傾向をとらえるために用意した，粒度の異なる複数のスケッチを更新する．

ここでは，各ストリームについて単一種類の特徴量を計算する前提で分類を行なったが，複数種類の特徴量を計算する場合は，ストリームをミラーリングして図 2(a)〜図 2(c) を組み合わせた形態となる．

4.2 クエリ処理部

クエリ処理部は，ストリーム処理部が解析結果として出力したストリームごとの特徴量，あるいはストリームの部分列ごとの特徴量を比較して，近似ストリームや近似部分ストリームを取り出す．基本的には，考え得るすべての組み合わせについての相関を計算し，相関が強い順にソートする処理であるため，MapReduce モデル [6] を多段に構えることで解決が可能ではある．

一方で，こうした問題空間を全探索する手法は，ストリームマイニングのような爆発的な量のデータを処理する上で，効率が悪く現実的でない．そこで探索空間を絞り込む手段として，Zhu ら [4] や Zhou ら [7] らの研究があるが，いずれも規模を縮小した多段の MapReduce モデルで実現可能である．ストリームマイニングにおいて並列分散化の対象となるのはクエリ処理部であるが，極めて並列度が高い部分であるため，こうした問題空間を小さくする工夫が性能向上の鍵となる．

5 まとめと今後の課題

本稿では，代表的なストリームマイニングアルゴリズムのデータアクセスパターンによる分類を試みた．今後は，さらに多くのストリームマイニングアルゴリズムを対象とした分類・検討を行なうと同時に，分類結果に基づいたストリームマイニングの効率的な並列分散手法についても検討を行なう予定である．

謝辞

本研究の一部は，文部科学省研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」（計画研究 A02-00-04，情報爆発に対応する高度にスケーラブルなモニタリングアーキテクチャ）による助成，ならびに文部科学省次世代 IT 基盤構築のための研究開発「Web 社会分析基盤ソフトウェアの研究開発」（多メディア Web 解析基盤の構築及び社会分析ソフトウェアの開発）の委託に基づいて行なわれた．

参考文献

[1] D. Turaga et al., “Resource Management for Networked Classifiers in Distributed Stream Mining Systems”, Proc. Sixth Int’l Conf. on Data Mining (ICDM’06), 2006.

[2] B. Thuraisingham et al., “Dependable Real-time Data Mining”, Proc. Eighth Int’l Sympo. on Object-Orient Real-Time Distributed Computing (ISORC2005), 2005.

[3] A. C. Gilbert et al., “Surfing Wavelets on Streams: One-Pass Summaries for Approximate Aggregate Queries”, Proc. 27th Int’l Conf. on Very Large Data Bases (VLDB2001), 2001.

[4] Y. Zhu et al., “StatStream: Statical Monitoring of Thousands of Data Streams in Real Time”, Proc. 28th Int’l Conf. on Very Large Data Bases (VLDB2002), 2002.

[5] Y. Sakurai et al., “BRAID: Stream Mining through Group Lag Correlation”, Proc. SIGMOD2005, 2005.

[6] J. Dean et al., “MapReduce: Simplified Data Processing on Large Clusters”, Proc. 6th Symp. on Operating Systems, Design, and Implementation (OSDI’04), 2004.

[7] M. Zhou et al., “Efficient Online Subsequence Searching in Data Streams under Dynamic Time Warping Distance”, Proc. 24th Int’l Conf. on Data Engineering (ICDE2008), 2008.