研　究　報　告　書

「高次知識を獲得するリソース指向型オンラインマイニング法の開発」

研究タイプ：通常型
研究期間：平成26年10月～平成30年3月
研 究 者：山本　泰生

1．　研究のねらい

本研究の目的は、ストリーム型ビッグデータの利活用に資するデータマイニング法の基盤技術開発である。ストリームデータとは、観測系から生成され続ける無限長のデータ系列のことである。近年のクラウドサービスやIoT の発展に伴い、気象、小売、製造、インフラ、観光、医療、スポーツなどの多岐にわたる分野において，多種多様なストリームデータが生み出されている。

ストリームデータの特徴は、時間経過とともに蓄積されるデータ総量が急速に増加する点にある。よって、ストリームデータを扱う計算タスクでは、ボトルネックとなるデータ全体のスキャンを極力避けながら、新規データを逐次的にインメモリ処理することが求められる。このような“On-the-fly”なオンライン処理はビッグデータの管理・運用技術における重要な研究課題と位置づけられている。

ストリームデータの社会的インパクトは観測系に対するリアルタイム分析にある。特に、科学的発見のような観測系の「理解」やマーケットバスケット分析（MBA）にみられる観測系の「改善」を目的とする分析タスクでは、生成モデルだけでなく、再現性を担保する解釈可能な形式表現を獲得することが不可欠である。このような、人の認知が介在するデータ活用をリアルタイム分析のコンテキストで実現することをねらい、そのために不可欠なオンライン処理技術を開発する。

本研究では、形式知として、観測系を特徴づける説明変数間の関係表現（パターン）を扱う。観測系や分析タスクにより、説明変数とパターンはそれぞれ異なるが、他方、パターンの組み合わせ爆発現象を抱える点は共通する。計算科学としてのパターン抽出の本質的課題は、この組み合わせ爆発にある。特にアイテムの出現分布が常に変化していくストリームデータにおいては、これに立ち向かうコア技術が十分確立されているとは言い難い現状であった。

そこで本研究では、近似・圧縮・並列化の 3 つの要素技術を導入し精錬することで、ストリームデータからパターンを抽出するスケーラブルなオンライン計算法を確立する。具体的には従来法に比べ 100 倍の高速化とスケーラビリティの改善を計り、バスケット長（バスケットに含まれる説明変数の数）10,000 のデータの実時間処理を目指す。また開発したコア技術を利用し、より複雑な高次のパターン（集合系列等）抽出にチャレンジする。さらに、ストリーム型ビッグデータの省スペース管理のための新しいオンライン計算法の開発にも取り組む。

2.　研究成果

（1）概要

本研究では、(A) オンライン型パターンマイニング法について、Δ-圧縮法と呼ぶ非可逆圧縮表現に基づく新たなソリューションを提案し、目標としたスケーラビリティの改善を実現した[主な業績 1,3,4][論文 1,2][受賞 1,3]。また (B) Δ-圧縮法を拡張し、集合系列を扱うオンライン型パターンマイニング法を開発した [主な業績 2][論文 12]。また負の相関ルール抽出のためのフレームワークを整備し、アルゴリズムの効率化を進めた [論文 11]。さらに (C) 順序関係に基づくストリームデータのサマリ計算法を検討し、時系列データのオンライン符号化への応用可能性を示した [論文 8]。主な研究成果を以下の 3 項目にまとめる。

(A) オンライン型パターンマイニング法のスケーラビリティ向上

アイテム集合を抽出対象とするデータマイニング分野の代表的問題に対し、従来法の性能を大きく上回るオンライン計算法を開発した。従来法は、各パターンの頻度情報に誤差Δを許容するオンライン近似計算をベースとしている。これに対し、提案法では、Δを利用した非可逆圧縮を新たに導入し、管理すべきパターンを大幅に削減している。また近似計算の従来アプローチを統合した新たな近似法を考案した。これにより、任意の長さ Lのバスケットデータを計算時間 O(k×L) (ただし k は消費メモリの上限を示すサイズ定数) によりオンライン処理することを可能とした。実証実験の結果、従来法に比べ、平均して 200 倍の高速化、20分の1の省メモリ化、85%の誤差Δの軽減に成功している。

(B) 高次パターンへの拡張: 集合系列パターンのオンラインマイニング

(1) のコア技術をもとに集合系列のΔ-圧縮表現を抽出するオンラインマイニング法を開発した。集合系列は関係表現の中でも最も複雑なパターンであり、従来のオンライン計算の枠組みでは実問題へ適用することは原理的に困難とされてきた。この問題に対するソリューションを世界に先駆けて提案している。また本手法を系列予測問題に適用し可用性を検証した。Yahoo! Hadoop Grid のログデータを用いた実験において、Variable-order Markov Model (VMM) の既存手法を上回る予測性能を確認した。

(C) ストリームデータのサマリ構築

順序関係に基づくストリームデータの要約表現 (サマリ) に注目し、半順序及び全順序を満たすストリームデータのε劣サマリを求める決定論的オンラインアルゴリズムを開発した。また本手法を、事前分布仮定を必要としないデータ符号化に適用し可用性を検証した。ブラックスワンを含む高感度地震観測網 (Hi-net) の地震計データを用いた予備実験において、既存法に比べ符号化誤差を 30 分の 1 まで削減できることを確認している。

（2）詳細

（A）オンライン型パターンマイニング法のスケーラビリティ向上

オンライン計算の枠組みで組み合わせ爆発問題に対処するコア技術を開発した。

1. 軽量なオンライン処理により実現可能な非可逆圧縮法の開発

アイテム集合を対象とするマイニングでは、飽和性に基づくパターンの可逆圧縮がよく利用される。しかし一方、データによっては可逆圧縮の効果が限定的となる場合がある。例えば、機械学習で対象となる稠密なデータでは、飽和圧縮は有効に働かないことが多い。本研究では、Δカバーと呼ぶ飽和性の拡張概念に着目し、元のソリューションである全頻出アイテム集合族の非可逆圧縮表現を求めるオンライン計算法を提案している。この圧縮表現から任意の頻出アイテム集合と頻度を誤差Δの範囲内に復元することが可能である。Δカバーの概念自体は、パターンのクラスタリングや素性選択の指標としてすでに有用性は確認されていた。他方、オンライン計算の枠組みにて、Δカバーに基づく圧縮表現をいかに求められるかわかっていなかった。本研究では、この有用な圧縮表現が、集合積演算と近似削除と呼ばれる逐次処理により獲得できることを理論的に明らかにしている［主な業績1,4][受賞 1,3]。

2. オンライン近似計算の新たなパラダイム構築

近似削除は、許容誤差Δ以下の低頻度パターンを削除する操作である。パターン抽出を扱うオンライン近似計算の代表的な省スペース技術であるが、普通、ユーザーが事前に設定する誤差パラメータ ε（$0 \leqq \varepsilon < 1$）をもとに、$\varepsilon \times i$（i は任意の時刻）以下のパターンを削除していく。このアプローチをパラメータ指向（PO）型近似と呼ぶと、PO 削除法では、常に頻度誤差は $\varepsilon \times i$ 以下となる。この手法とは別に、研究代表者らが提案したリソース指向（RO）型の近似アプローチでは、ε の代わりにサイズ定数 k を利用する。RO 削除法では、頻度 top-k のパターンのみを管理し、それ以外を削除する。PO 削除と RO 削除はそれぞれ長所と短所を持つ。すなわち、PO 削除では許容誤差Δを ε により直接コントロールできる反面、突発的な候補パターンの増加に応じて性能が著しく低下する。他方、RO 削除では k によりメモリ消費の上限を直接コントロールできる反面、最終的な誤差計数は不明であり出力の品質を事前に担保することができない。

そこで本研究では、PO 型と RO 型を統合した新しい近似アプローチを提案している。提案法では、各時刻における誤差率と登録パターン数に応じて、動的に PO 削除と RO 削除をスイッチする。図1は、あるベンチマーク問題における提案法の誤差率と登録パターン数の時間推移を表している。サイズ定数 k = 10,000、誤差パラメータ ε = 0.015 と設定している。

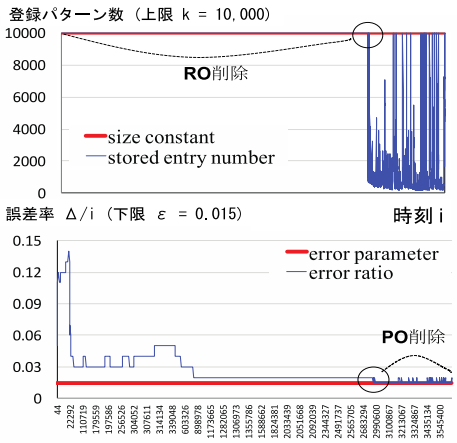

図1: PO 型と RO 型削除の動的スイッチ

図1をみると、最初の段階では、top-k 以外のパターンが削除されており（RO 削除）、結果として誤差率が ε を超える。ところが一定時間が過ぎた段階で、許容誤差の値（$\varepsilon \times i$）が大きくなり PO 削除が適用されている。その結果、最終的な誤差率は ε 以内に保たれるようになる。このような

誤差率の自律的な回復現象は、PO 型と RO 型を統合することで生じた新しい知見であり、既存法への波及効果も期待できる［論文投稿中］。

本研究では上記のコア技術を導入した新しいオンライン型パターンマイニング法を実装し、様々な実データと合成データを用いて性能評価を行なった。実験の結果、速度・使用メモリ・誤差計数（すなわち解の品質）の観点から、既存法の性能を大幅に改善していることを確認した。結果として当初目標とした 100 倍の高速化とスケーラビリティ改善に成功している。図2は、バスケット長 L の変化に対する速度と誤差率の推移を表している。図2より、L に対して速度と誤差率が線形に推移することがわかる。

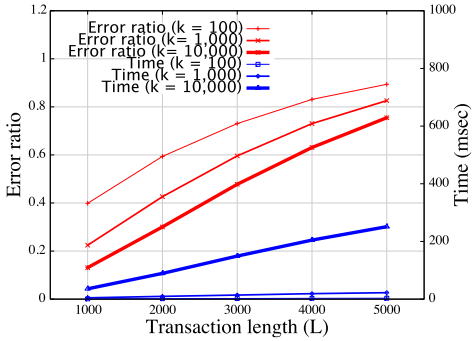

図2: 提案法のスケーラビリティ

提案法のオンライン処理は、各登録パターンと新規データとの共通部分を求める集合積演算と近似削除の2操作により実現されるが、これらの操作は各パターンに対して独立して実行することができる。また登録パターン数の上限はサイズ定数 k により固定されるため、並列分散化が適用できることができる。そこで本研究では、OpenMP によるマルチスレッド処理により提案法のオンライン計算を実現した［主な業績 4］。また Xeon E5-2698v3（32 コア）マシンを用いた実験の結果、16 倍の速度改善の効果を確認している。

(B) 高次パターンへの拡張: 集合系列パターンのオンラインマイニング

(A) で開発したコア技術を用いて、複雑な関係表現を扱うパターンマイニング法を開発した。

1. 集合系列パターンの抽出と系列予測への応用

対象パターンは、図3のように、ストリーム上に出現する説明変数（アルファベットに相当）の集合列である。図の例のように、パターンを構成する変数の出現は不連続でも構わないが、最大系列幅 W 以内に必ず出現するものとする。よって、バスケット長を L とすると $2^{W \times L}$ 個もの集合系列パターンが候補となる。集合パターンの場合に比べ、さらに多くのパターンを管理する必要がある。このため、オンライン計算の枠組みでは集合系列を抽出することは困難と考えられてきた。本研究では、Δ-圧縮法と PO-RO 近似法を用いることで、従来困難とされたこの問題に対する新しいソリューションを与えている。

系列特有の難しさは、系列間の（極大）共通部分がユニークに定まらない点にある（例えば (a)->(b)->(c) と (a)->(c)->(b) の極大共通部分は (a)->(b) と (a)->(c) の 2 つ存在する）。Δ-圧縮法におけるオンライン処理には、登録パターンと新規データの共通部分を求める積演算が必要である。このため、集合から系列へΔ-圧縮法を拡張するには、系列間の極大共通部分を列挙することが求められた。本研究では、この列挙問題を効率的に解くアルゴリズムを考案し、Δ-圧縮法に

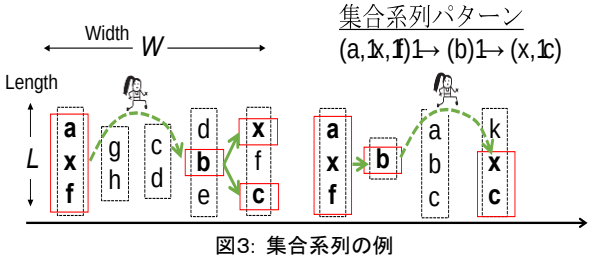

図3: 集合系列の例

組み込むことで、集合系列族の非可逆圧縮表現を高速・省スペースに求めるオンライン計算法を実現した[主な業績 2]。

提案法により求められる圧縮表現は、任意の集合系列の頻度情報を誤差Δの範囲内で復元することができる。すなわち、これをストリームデータの解釈可能な生成モデルの一つとみることができる。そこで本研究では提案法の可用性を検証するため、提案法を系列のオンライン予測に適用する実証実験を行なった。系列予測器としての特徴は 3 つある。第 1 に RNN とは異なり、予測結果に対する説明を与えることができる。第 2 に HMM とは異なり、事前に適切なモデルの構造を与える必要がない。これらの特徴を持つ代表的な生成モデルに VMM があるが、第 3 の特徴として VMM とも異なり「飛び飛び」の不連続なコンテキストを考慮した予測が可能である。

これらの特徴が活用される場面として、多様なコンテキストがごった煮となって入れ混じるデータのオンライン予測が考えられる。ケーススタディとして、Yahoo! Research が提供する Hadoop Grid Log データを用いた実証実験を行った。このデータは 5,000 を超えるクライアントユーザーによる HDFS サーバのアクセスログ記録（Audit）が保存されている。同時多発的に書き込まれるため、ユーザーの一連の操作はデータ上不連続でだが、代表的なパターンが存在する（例えば、listStatus -> mkdir -> rename -> remove)。このようなパターンを抽出し利用することで精度の高い予測と説明を与えることができる。

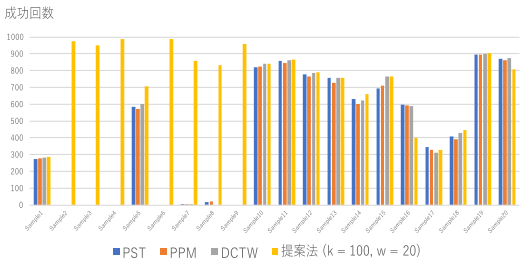

図4: 系列予測の精度（20 サンプル）

図4は 20 個のサンプル（訓練イベント 199,000 個とテストイベント 1,000 個のセット）での提案法と VMM の代表的な既存法との予測精度の比較結果である。全体的に提案法が良好な結果を得ていることがわかる。また各予測の説明するパターンを検証したところ、実際ユーザーの飛び飛びの操作列を意味しているものを確認している [論文投稿中]。また VMM より高い精度を得る本実験結果は、算術符号としての応用可能性を示唆しており、今後の波及効果も期待できる。

3.　今後の展開

本研究では、ストリームデータから解釈可能な形式知を獲得する効率的なオンライン処理技術を開発している。形式知として、データを特徴づける説明変数上の関係表現（パターン）に着目し、複雑なパターンにも拡張可能なオンライン計算のスキームを提案している。Δ−圧縮法とPO−RO 統合型近似法を用いて、従来困難とされてきた集合系列パターンのオンライン抽出を実現している。これにより、既存法より優れた予測精度と説明を与えうる解釈可能な生成モデルを新たに提供することが可能となっている。

オンライン型パターンマイニングの本質的な難しさは、管理すべきパターンの組み合わせ爆発にある。本研究で提案したオンライン計算のスキームでは、ソリューションとして、元の抽出対象パターンではなく、その非可逆圧縮表現を求める。誤差計数を許容した非可逆圧縮を導入することで、目標とした 100 倍の高速化とスケーラビリティの改善を達成している。

今後これらの新技術をもとに、人の認知が介在するリアルタイム分析の実応用に繋げていく。

4.　評価

（1）自己評価

本研究には 2 つの目標があった。すなわち第 1 の目標は、オンライン型パターンマイニングの高速化とスケーラビリティの大幅な改善であり、第 2 の目標は、オンライン型パターンマイニングの新しい応用可能性を示すことであった。第 1 の目標については、上述した通り、目標を上回る成果を得ることができている。また集合系列に拡張可能なオンラインパターンマイニングのソフトウェアは世界的にもユニークなものであり、公開後、ストリーム型ビッグデータを扱うリアルタイム分析への応用展開が見込まれる。第 2 の目標について、系列予測器としての応用可能性を実証的に示している。VMM は算術符号によく利用される代表的な生成モデルであり、ベンチマークによるさらなる比較実験を通して、解釈可能な新しい生成モデルとして発展することが期待できる。他方、データ駆動科学であるビッグデータ分野を新しく切り開くという点では、本研究期間内に十分な成果を示すことができなかった。今後、災害・セキュリティ等の人の認知が介在するリアルタイム分析の分野において本研究で培った知見を活かし、社会に資する研究成果を出していきたい。

（2）研究総括評価（本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

ビッグデータとともに IoT が注目を集め、大量のセンサーから生成されるストリームデータのリアルタイム解析が、さまざまな応用分野で重要な課題と認識される時代となった。高速、省メモリ、高精度のオンライン解析技術が求められている。

本研究では、ストリームデータに対する効率的なオンライン頻出パターンマイニングのために、非可逆圧縮を用いる方式を提案した。圧縮結果から頻出パターンを復元する際に、頻度の誤差を一定値まで許容することで、従来法の 100 倍の高速化とバスケット長 10,000 に対応できるスケーラビリティという目標を達成している。さらに、許容誤差とメモリ使用量のトレードオフを考えて 2 種類のポリシーを動的に切り替える方式、非可逆圧縮の手法を拡張して集合系列パターンの抽出や系列予測にも適用可能な方式を考案した。論文は国際会議などで多数発表しており、学術的には十分なレベルに達している。

さきがけ研究期間中には、コア技術の基本性能や表現力を大きく進歩させることができた。今後は、この成果を使いやすい形で提供し、インパクトのある応用分野を切り拓いていくことを期待したい。

5. 主な研究成果リスト

(1)論文(原著論文)発表

1. Yoshitaka Yamamoto and Koji Iwanuma: Online pattern mining for high-dimensional data streams, Proceedings of the 2015 IEEE International Conference on Big Data, pp. 2880-2882 (2015.11)
2. 山本泰生, 山内夏美, 岩沼宏治: Incremental intersection for frequent sequential pattern mining, 第 100 回人工知能基本問題研究会 Vol. B05-03. pp.80-85 (2016.3)
3. 山本泰生, 岩沼宏治: オンライン頻出パターンマイニングの並列分散化に向けて, 第 30 回人工知能学会全国大会 Vol. 3K4-1in1, pp.1-4 (2016.5)
4. 岩沼宏治, 山本泰生, 福田翔士: ストリーム中の頻出飽和集合を抽出するオンライン型 ε-近似アルゴリズムの完全性, 人工知能学会論文誌 Vol. 31, pp. 1-10 (2016.8)
5. 山本泰生: 高次知識を獲得するリソース指向型オンラインマイニング法の開発, 日本オペレーションズ・リサーチ学会 Vol. 62, pp. 246-252 (2017.3)

(2)特許出願

なし

(3)その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

学会発表

1. 山本泰生, 岩沼宏治: リソース指向近似計算に基づくオンライン頻出アイテム集合マイニング, 情報処理学会第 77 回全国大会, pp. 443-444 (2015.3)
2. 山本泰生, 福田翔士, 岩沼宏治: トランザクションストリーム上のオンライン型頻出飽和集合マイニング, 人工知能学会合同研究会優秀賞記念講演 (2015.11)
3. Yoshitaka Yamamoto: Approximation and compression in streaming data mining, 23rd International Symposium on String Processing and Information Retrieval (2016.9) [Invited talk]
4. 山本泰生: リソース指向型ストリームデータマイニング法の高度化に向けて, 第 15 回自動チューニング研究会 (2016.10) [招待講演]
5. 山本泰生: リソース指向型計算に基づくストリームデータマイニングの研究, 人工知能合同研究会 2017 知識ベースシステム研究会 招待講演 (2017.11)

受賞

1. 2014 年度人工知能学会研究会優秀賞 (2015.3)
2. 山梨大学優秀教員奨励制度特別表彰 (2016.6)
3. 山梨科学アカデミー奨励賞 (2017.5)

著作物

1. データ構造とアルゴリズム（電子情報通信レクチャーシリーズ）7 章及び 8 章担当(印刷中)