様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月 １９ 日現在

機関番号：１３５０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2019
課題番号：１６Ｋ００２９８

研究課題名（和文）潜在的相関ルール抽出を目的したオンライン型近似計算法の開発と仮説推論との統合

研究課題名（英文）An On-line Approximation Algorithm for Mining Latent Association Rules and its Integration with Hypothetical Reasoning

研究代表者
岩沼　宏治（IWANUMA, Koji）

山梨大学・大学院総合研究部・教授

研究者番号：３０１７６５５７

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：　本研究では，データストリームの中の潜在的法則の抽出を目的として，オンライン型の負の相関ルール抽出のための基盤技術を開発した．更により深いレベルの潜在規則の抽出を目的として，正負のルールを統一的に扱う一般化相関ルールを新しく提案し，仮説推論の実現を試みた．
　まず極小生成子に基づく負ルール集合の無損失圧縮を新しく提案し，良好な圧縮性能を確認した．また併せて極小生成子の高速抽出法を提案した（人工知能学会2017年度研究会優秀賞受賞）．次に極小生成子生成の母体となる飽和集合をオンライン抽出する新しい圧縮抽出法を提案した．更に正負のアイテムが混在する一般化アイテム集合を考察し，高速抽出法を開発した．

研究成果の学術的意義や社会的意義
本研究の学術的意義は，これまで殆ど着目されてこなかった負の相関ルールと潜在的規則関係に着目し，種々の技術的開発を行ったことにある．これにより巨大データに埋もれる多数の潜在的性質を，負ルールの形で抽出発見することがある程度可能になった．昨今のセンサーネットワークの飛躍的な発展に伴い増大する一方の巨大データの利用方法を，より一層高度化し深化をさせる試みであり，昨今の高度情報化社会における社会的貢献を行うものと考えられる．

研究成果の概要（英文）：In this research, we proposed several new methods for online mining of negative association rules in order to discover important latent rules embedding in a data stream. Moreover, we gave a new framework, called a generalized association rule, for treating positive and negative association rules in a uniform way, and studied a hypothetical reasoning based on this proposed framework.
　More concretely, we first proposed a novel effective lossless compression method of negative association rules by using minimal generators, and gave a fast algorithm for extracting minimal generators from closed itemsets (we got the 2017 research award from Japan Artificial Intelligence Association). Next we construct an efficient compression methods for online mining of closed itemsets in a data stream. Furthermore, we proposed a concept of generalized itemsets consisting of both positive and negative items, and studied a fast extraction method of the generalized itemsets for a transaction database.

研究分野：人工知能基礎

キーワード：データマイニング　潜在的規則　負の相関ルール　圧縮　極小生成子　オンライン計算　一般化アイテム集合　一般化相関ルール

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

現実の世界を考えるとき，潜在因子までも考えた相関ルールの自動抽出は非常に重要である．しかし，研究を開始する前年度の 2015 年度の時点のデータマイニングの研究では，潜在的相関ルールは殆ど研究されていなかった．潜在因子の予測と抽出は，統計的学習での潜在パラメータの推定問題等とは異なる種類の問題である．我々の知る限り，潜在的相関ルールの発見に関係する研究は，負の相関ルールマイニングしかない．負の相関ルールとは， X と Y をアイテム集合とするとき，¬X ⇒ Y または Y ⇒ ¬X の形のルールのことであり，¬X は負のアイテム集合と呼ばれる．ルール ¬X ⇒ Y は「X が出現しない場合に Y がよく出現する」ことを意味しており，X は Y の出現に影響を及ぼす潜在因子と考えられる．負ルールの抽出は，陽には殆ど出現しないアイテム集合の検出が本質的に必要であり，効果的な計算は容易なことではない．

負の相関ルールは 1990 年代の末から研究が開始されている[文献①]．負ルール ¬X ⇒ Y（または Y ⇒ ¬X）の基底となるアイテム集合 X∪Y は非頻出となるため，負ルールは正ルールよりも数が本質的に非常に多い．このため，これまでは負ルールの有効性の評価尺度について多くの研究 [文献①，②] されてきたが，それを利用した高速な抽出法については殆ど研究されていなかった．[文献 A2] では，Apriori 型の抽出法を提案しているが，負ルール候補の膨大な基底アイテム集合を明示的に生成しており，非常に効率が悪い．[文献③]では基底集合の生成を避けて，頻出アイテム集合 X と Y の組合せから負ルールの抽出を行うが，Apriori 型ボトムアップ計算を行うために高速化には限界がある．これに対して，我々は新しく頻出集合のみを用いるトップダウン型の負ルール抽出手法を開発した．また，抽出した正と負のルールの無矛盾性，即ち X ⇒ Y と X ⇒ ¬Y の同時抽出の禁止問題はこれまで見過ごされてきたが，これを防止する枠組みも新しく提案している．実装評価を行い，100~1000 倍程度の高速化を確認している[文献⑤]．

上記の「頻出集合の組合せだけを使って負ルールを抽出」する方式は，トランザクション・ストリーム上でのオンライン型処理に本質的に適した性質である．事前に頻出アイテム集合をストリームからオンラインで抽出する必要があるが，組合せ爆発が生じるために通常のオンライン処理では対応が難しい．この問題に対して，我々は計算資源指向型のオンライン型近似抽出法 Skip LC-SS を開発した．Skip LC-SS 法は，使用可能なメモリ量を定数値に固定し，メモリ使用状況を常に観察しながら，適応的に候補アイテム集合を積極的にスキップする．この研究成果は，データ科学・工学の分野で世界のトップ会議である 2014 ACM SIGMOD に regular paper として採択されている [文献④]．また更に圧縮技術の積極的な利用を図ることの重要性に鑑み，頻出アイテム集合の圧縮表現である飽和アイテム集合をストームデータから直接抽出するオンライン型近似計算法を提案し，人工知能学会 2014 年度研究会優秀賞を受賞している．

２．研究の目的

観測される現象の関係や法則を深く考察するには，データの中に隠れている潜在因子を考慮することは極めて重要である．我々は負の相関ルールマイニングに着目し，データストリームの中に隠れる潜在的法則の抽出を目的として，各種の圧縮原理を積極的に活用したオンライン型の負ルール抽出アルゴリズムを開発する．更により深いレベルに隠れる潜在因子や規則の抽出を目的として，仮説推論技術の導入を試みる．データストリーム中の論理ルールの効果的な抽出手法を開発し，それらを背景知識とした仮説推論や結論発見計算の実現を試みる．提案手法の有用性について理論的考察を行うと共に，幾つかの実データを用いて実証的に評価を行う．

３．研究の方法

本研究では，負の相関ルール抽出に関する幾つかの先進的技術を開発し，潜在規則オンライン型マイニングの基盤となる技術の確立を図る．具体的には以下の研究課題に順次取り組んだ．

（１） 膨大な負の相関ルール全体の圧縮を目的として，極小生成子に基づく圧縮復元技術の開発と，圧縮したデータから直接的に負ルールを抽出する手法の開発．
（２） トランザクション・ストリームから負ルール集合を抽出するオンライン型近似計算アルゴリズムの開発．
（３） トランザクション・データベースから論理型ルールを抽出する基盤技術の開発と仮説推論の導入

４．研究成果

以下に研究成果の概略を述べる．

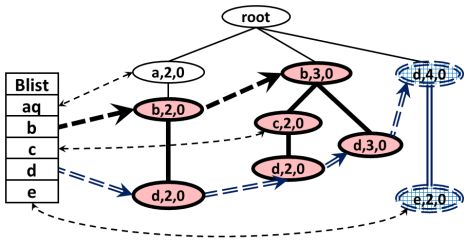
（１）膨大な数の負ルールの圧縮は非常に重要である．頻出アイテム集合の圧縮によく用いられる飽和性は，負の相関ルールの圧縮には全く無力である．本研究では極小生成子に基

づく負ルール集合の圧縮原理について考察し，提案法の無損失圧縮性を理論的に証明した．また実証実験により密なデータセットに関して十分な圧縮性能を持つことを確認した．

（2）上記の極小生成子は高速な抽出が難しいために，本研究では飽和アイテム集合からの高速抽出法を新しく開発した．新しく極小生成子の下方閉包性を発見し，これを枝刈に用いた抽出法を提案した．実験の結果，アイテム集合の出現頻度計算が高速化の大きな障害であることが判明したので，頻度計算を陽に行わない抽出法を新たに開発した．以上の研究成果により人工知能学会 2017 年度研究会優秀賞を受賞している．

（3）圧縮された極小生成子の集合だけを用いて（即ち，膨大な数におよぶ頻出アイテム集合は全く用いずに）妥当な負ルールを全て抽出するためのオフライン型アルゴリズムを開発した．極小生成子が持つ下方閉包性から，全ての極小生成子を頂点とする接尾辞木を構成し，その上で確信度の上界関数の逆単調性等を基いた枝刈を駆使するルール抽出法である．

（4）極小生成子の母体となる飽和アイテム集合を，データストリームからオンライン高速抽出するための新しい技術を開発した．高速にオンライン抽出するためには，これまでに抽出した多数の飽和アイテム集合を効果的に保持し，更にその上の集合積計算を高速に行う必要がある．本研究では多数の飽和アイテム集合を圧縮保持するために Han らが開発した FP 木に着目し，そこにスキップ走査を導入し，圧縮した形のままで集合積計算を非常に高速，即ち木のサイズの線形時間で行うことを可能にした．実証実験の結果，提案したスキップ走査 FP 木は「疎と密の両方のデータを効果的に圧縮し，高速に集合積を計算できる」という従来には無かった有用な特徴をもつことが確認できた． 以下は {a,b,d}, {b,c,d}, {b, d}, {d, e} の 4 つの飽和アイテム集合を保持する FP 木の例である．このときアイテム集合{b,c,d,e}と，この 4 つの飽和集合それぞれとの集合積演算の履歴が，FP 木の頂点の背景色で表現されている．全体の集合積演算が FP 木の頂点数に対する線形時間で完了することは，全ての頂点が高々1 色しか持たないことがその証拠になっている．



（5）より大きなサイズの飽和集合を高速に抽出することは，実用上，極めて重要である．そのため本研究では，抽出する飽和アイテム集合のサイズに最小サイズ制限を設け，スキップ走査 FP 木の上の集合積演算を最小サイズ制限に基づき高速化する技術を開発した．右グラフは，Mushroom データセットに対する実証実験の結果であり，高速化効果が確認できる．

（6）仮説推論を導入する上で，相関ルールを正と負のアイテムが混在する形へ一般化することは必要不可欠である．そのために，その基盤となる正負のアイテムが混在する一般化アイテム集合の基本性質を考察した．その中で発見した正負分離可能定理に基づく効率的なその抽出法を提案した．更に一般化アイテム集合の飽和性について考察を進め，閉包演算と一般化された接頭木探索に基づく高速抽出法を開発した．更に一般化アイテム集合上の相関ルールの枠組みと高速抽出法を考察した．

<引用文献>

① S.Brin, R.Motwani and C.Silverstein, “Beyond Market Baskets: Generalizing Association rules to Correlations,” Proc. ACM SIGMOD Conf., pp. 265-276, May 1997.

② X.Wu, C.Zhang and S.Zhang: “Efficient Mining of Both Positive and Negative Association Rules, ACM Trans. on Information Systems, 22(3), pp.381-405, July 2004.

③ H.Wang, X.Zhang and G.Chen: “Mining a Complete Set of Both Positive and Negative Association Rules from Large Databases”, Proc. PAKDD’08, pp.777-784, 2008.

④ Yoshitaka Yamamoto, Koji Iwanuma and Shoshi Fukuda: Resource-oriented Approximation for Frequent Itemset Mining from Bursty Data Streams. Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data (SIGMOD'14), pp: 205-216, (2014)

⑤ 井出典子, 岩沼 宏治, 山本 泰生：負の相関ルールを抽出する高速トップダウン型アルゴリズム, 人工知能学会論文,Vol.29, No.4, pp.406--415 (2014)

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 Yoshitaka Yamamoto, Yasuo Tabei, Koji Iwanuma	４．巻 -
２．論文標題 PARASOL: a hybrid approximation approach for scalable frequent itemset mining in streaming data	５．発行年 2019年
３．雑誌名 Journal of Intelligent Information Systems	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10844-019-00590-9	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 岩沼宏冶，佐生隼一，黒岩健歩，山本泰生	４．巻 57
２．論文標題 負の相関ルール集合の極小生成子に基づく圧縮表現	５．発行年 2016年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 1845-1849
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 岩沼宏冶，山本泰生，福田翔士	４．巻 31
２．論文標題 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	５．発行年 2016年
３．雑誌名 人工知能学会論文誌	６．最初と最後の頁 1-10
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計16件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 Koji Iwanuma, Takumi Nishina, Yoshitaka Yamamoto
２．発表標題 Accelerating an On-Line Approximation Mining for Large Closed Itemsets
３．学会等名 2019 IEEE International Conference on Big Data (Big Data)（国際学会）
４．発表年 2019年

１．発表者名
安藤 祐太，岩沼 宏冶

２．発表標題
閉包計算に基づく一般化飽和集合の高速な列挙法：相関ルールの一般化を目指して

３．学会等名
人工知能学会 第112回人工知能基本問題研究会

４．発表年
2020年

１．発表者名
Takumi Nishina ; Koji Iwanuma ; Yoshitaka Yamamoto

２．発表標題
A Skipping FP-Tree for Incrementally Intersecting Closed Itemsets in On-Line Stream Mining

３．学会等名
2019 IEEE International Conference on Big Data and Smart Computing (BigComp)（国際学会）

４．発表年
2019年

１．発表者名
雨宮 晶良，岩沼 宏治，谷島 健斗，山本 泰生

２．発表標題
正負の相関ルールの妥当性の再考察と正負ルールの高速抽出手法

３．学会等名
人工知能学会研究会第 115回知識ベース研究会

４．発表年
2018年

１．発表者名
Takumi Nishina, Koji Iwanuma and Yoshitaka Yamamoto

２．発表標題
Frequent Closed Itemsets More than Size K and Its On-Line Approximation Mining

３．学会等名
3rd International Conference on Big Data, Cloud Computing, and Data Science Engineering (BCD 2018)（国際学会）

４．発表年
2018年

１．発表者名
谷島健斗，岩沼宏冶，山本泰生

２．発表標題
負の相関ルールマイニングの効率化のための飽和アイテム集合からの極小生成子の高速抽出

３．学会等名
人工知能学会　第112回知識ベース研究会

４．発表年
2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
負の相関ルールマイニングとオンライン近似計算
http://www.kki.yamanashi.ac.jp/~iwanuma/Kaken2018/

本研究の一部の成果に対して人工知能学会2017年度研究会優秀賞を受賞している．

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	山本　泰生 (YAMAMOTO Yoshitaka) (30550793)	静岡大学・情報学部・准教授 (13801)	