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　６ 年　 ６ 月 ２０ 日現在

機関番号：１３５０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2023
課題番号：１９Ｋ１２０９６

研究課題名（和文）潜在的規則の抽出を目的とした負の相関ルールの抽出の効率化と一般化

研究課題名（英文）Efficiency and generalization of the extraction of negative association rules for the extraction of latent rules

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）：本研究では巨大データに潜在するルール型知識の抽出を目的として，負の相関ルールの効率的な抽出法について研究を行った．多数のルールの圧縮は，本質的にルールの一般化・抽象化を行うために極めて重要である．そのため負ルール集合の幾つかの圧縮原理，圧縮した形で負ルール集合を高速抽出するための原理とそれらの高速実行アルゴリズムを開発した．例えば，極小生成子の理論的な性質の解明と，その高速抽出アルゴリズムを開発している．また強飽和集合の高速オンライン抽出アルゴリズムなども開発している．更に，極小生成子の下方閉包性を利用した列挙木の上で，負ルールを圧縮した形で直接列挙するアルゴリズムを開発している．

研究成果の学術的意義や社会的意義
現実の世界を考えるとき，潜在因子も考慮した法則の発見抽出は重要な課題である．しかし，これまでのデータマイニングの研究では殆ど考慮されてこなかった．潜在的ルールの発見抽出は，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問題であり，重要である．本研究の成果である負ルール集合の圧縮原理と直接抽出アルゴリズムにより，抽出計算が大幅に効率化・高速化でき，実用レベルの潜在関係規則のマイニングへ近づくことができた．ルール集合の圧縮は内在する共通な現象を発見して一つにまとめる作業と考えられる．これはルールを一般化することに相当し有用である．この圧縮に基づくマイニングはこれまで殆ど研究されていなかった．

研究成果の概要（英文）：In this study, we investigate an efficient method for extracting negative association rules for the purpose of extracting rule-type knowledge latent in huge data. Compression of a large number of rules is extremely important because it is considered to be essentially a generalization and abstraction of rules. Therefore, we developed several compression principles for negative rule sets, principles for fast extraction of negative rule sets in compressed form, and algorithms for fast execution of these principles. For example, we have theoretically clarified the properties of minimal generators and developed algorithms for their fast extraction. We have also developed a fast on-line extraction algorithm for strongly closed itemsets. Furthermore, we have developed an algorithm for directly enumerating negative rules in a compressed form on an enumeration tree, which is formed on the downward closure property of minimal generators.

研究分野：人工知能基礎

キーワード：データマイング　負の相関ルール　圧縮　極小生成子　アルゴリズム　一般化　飽和集合

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

現実世界を考えるとき，潜在因子を考慮した法則の発見と抽出は非常に重要な課題である．これまでのデータマイニングの研究では，潜在的因子を考慮したルールの抽出は殆ど研究されてこなかった．潜在的ルールの予測と抽出は，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り，その抽出計算はかなり難しい．潜在的ルールのマイニングに取り組んだ研究は負の相関ルールマイニングしかない．発生事象をアイテムと呼ぶとき，負の相関ルールとは，アイテムの集合 X と Y に対する

$\neg X \Rightarrow Y$（左否定形），$X \Rightarrow \neg Y$（右否定形）または　$\neg X \Rightarrow \neg Y$（両否定形）

の形のルールのことである．$\neg X$ や $\neg Y$ は負のアイテム集合と呼ばれる．負ルール $\neg X \Rightarrow Y$ 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ることができる．負ルールは正ルールでは表現が困難な共起関係を記述でき，潜在的な関係を表現するために極めて有用である．例えば，$e_1, e_2, e_3, \cdots, e_n, e_c$ を現在考慮しているアイテム（事象）の全てとするとき，$n-1$ 個の正ルール

$$\{e_2\} \rightarrow \{e_c\}, \quad \{e_3\} \rightarrow \{e_c\}, \quad \ldots, \quad \{e_n\} \rightarrow \{e_c\}$$

の全体で表す性質は，ただ一つの負ルール $\neg\{e_1\} \rightarrow \{e_c\}$ でコンパクトに表現することが可能になる．稠密なデータセット中にはこのような潜在的な関係が多数存在している．

このような負ルールの基底集合 $X \cup Y$ は非頻出なアイテム集合となる．このため負ルールの抽出には，陽にはあまり出現しないアイテム集合の検出が本質的に必要となる．そのため探索空間は正ルールに比べて格段に大きくなり，抽出されるルールも非常に多くなるため，効果的な抽出計算は極めて難しい．負の相関ルールは 1990 年代末から研究されているが，その殆どが膨大な負ルールを絞り込むための評価尺度についての研究 [1,2,4] である．負ルール集合の圧縮や，抽出計算の高速化ついてはあまり研究されていない．先行研究 [2] では Apriori 流のボトムアップ抽出法を提案しているが，妥当な負ルールの候補を求めるために，膨大な数の基底集合 $X \cup Y$ を明示的に生成しており非常に効率が悪い．これに対して [3] では，基底集合の明示的な生成を避けて，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っている．しかし，Apriori 流ボトムアップ計算を行うために探索空間の枝刈が難しく，高速化には限界がある．

負ルール集合の圧縮と，圧縮した形の負ルールを直接抽出できれば，計算全体が大幅に効率化･高速化でき，実用レベルの潜在関係のマイニングをかなりの程度実現できる．また本質的に，圧縮はデータに内在する共通な部分現象を発見して一つにまとめる作業であり，極めて重要である．抽出した負ルールの一般化が可能になり，より有用な潜在規則が発見できる可能性も高まる．しかし，これまで圧縮に基づく抽出計算は殆ど研究されていない．また，アイテム集合 X の否定集合 $\neg X$ は，正と負のアイテムの混在を認めない形式であり，潜在因子の表現能力に本質的な限界がある．正負双方のアイテムの混在を許す一般化アイテム集合を用いれば，表現能力は格段に向上する．ただし一般化アイテム集合の総数は，これまでの $O(2^n)$ から $O(3^n)$ に増加（但し，n はアイテムの種類数）し，一般化ルールの総数も $O(3^m)$ から $O(5^m)$ に跳ね上がるため，その効率的な抽出は負ルールと比較しても更に難しくなる [1,4]．飽和集合などの圧縮技術も全く研究されておらず，一般化ルールの妥当性の基準なども殆ど研究されていない．

２．研究の目的

本研究では，大規模データに隠れる潜在的法則の抽出を目的として，まず負ルール集合の多段階の圧縮法と直接抽出技術について研究を行う．我々はこれまでに，正の相関ルールの圧縮に有効な飽和アイテム集合は，負ルールの集合の圧縮には利用できないことを明らかにし，その解決法として極小生成子を用いた負ルール集合の無損失圧縮法を提案している [7]．本研究では，極小生成子のより高速な抽出法や，極小生成子で圧縮表現した負ルール集合を直接抽出する高速アルゴリズムを開発する．更に，抽出した負ルール集合を圧縮する多段圧縮法を可能とするために，強飽和集合 [5] およびストリームマイニングで開発された分位数近似サマリ技術 [6] に基づく技術について研究を行う．次に，より複雑な形の潜在因子や規則を発見するために，一般化アイテム集合とその上の一般化ルールの圧縮抽出法を開発する．具体的には，一般化アイテム集合の飽和集合と極小生成子の効率的な抽出法を開発する．開発システムは幾つかの実データを用いた実証的な評価実験を通して有効性を確認する．研究代表者の知る限りにおいて，これまで本課題と類似した研究は無く，独自性が高い試みでと考えられる．

３．研究の方法

研究目的を達成するめに具体的には以下の項目について研究を行った.

(1) 極小生成子の高速かつメモリ節約型の抽出アルゴリズムの開発： これまでの研究から, 極小生成子抽出の計算上のボトルネックは, アイテムの出現頻度の計算であることが判明している. 本研究では極小生成子の性質を考察し, 頻度計算を陽には行わずに生成子および生成子の極小性を判定できる計算原理をそれぞれ新たに解明し, その高速実行アルゴリズムを開発する.

(2) 極小生成子で圧縮表現した負の相関ルールの直接抽出： これまでの研究から, 極小生成子は部分集合に関して閉じている（下方閉包性）ことが判明している. これを用いれば極小生成子だけを頂点に持つ探索木が自然に構成できるので, この探索木から極小生成子上の負ルールが直接的に抽出できると考えている. 抽出を高速化するために, 効果的な上界関数を新たに導出して分枝限定法を用いたアルゴリズムの開発を行う. また抽出する相関ルールは互いに矛盾せず, また非冗長であって欲しいので, そのための抽出法を検討する

(3) 抽出した負の相関ルールの集合の圧縮： 抽出した負ルールは膨大な数に及ぶことから, 基底となるアイテム集合の更なる圧縮法として, 強飽和集合や分位数近似サマリ技術の利用を考察する. 一般に, 近似計算を導入すると, 頻度などに誤差を許容しなければならないが, その見返りとして非常に高い圧縮効率が期待できる. 誤差の精度を理論保証する枠組を鋭意検討していく.

(4) 一般化したアイテム集合上の負の相関ルールの抽出：一般化アイテム集合上のルールの効率的な計算法はまだ殆ど確立していない. そこでまず, 飽和集合と極小生成子の効率的な抽出法について研究を行う. 一般化集合は正負のアイテムが混在することから, 最初に高速抽出技術が確立している正アイテムの飽和集合を求めて, 次に負アイテムを順次追加し拡張する手法を考察する.

４．研究成果

研究を通じて, 負ルール集合を多段階で圧縮するための幾つかの圧縮原理とその効果的な実行アルゴリズムを開発した. また, 圧縮を行いながら負ルール集合を高速に抽出するための計算原理とアルゴリズムを開発した. 圧縮は, 相関ルールを一般化・抽象化して本質的に重要なルールに集約することを意味しており, 極めて重要な概念である. より具体的な成果は, 以下の通りである.

(1) 負ルールの基底となる頻出な極小生成子の高速列挙を目的として, 頻度計算を全く用いない生成子の判定原理を新たに導出した. また, その数学的な正当性を証明した. この判定原理を高速実行するために, 探索済みのアイテム集合の全てをハッシュ表に記録する下降型探索法を開発した. ただ, このハッシュ表が巨大になりがちでメモリ負荷が大きいことから, 生成子の極小性の判定を前述のハッシュ表も頻度計算も全く利用せずにできる計算原理を新しく導出した. これを用いて省メモリ型の高速計算法としてバケット型垂直配置表を新に考案した. これに基づく計算アルゴリズムを実装し, 性能評価実験を通して良好な性能を確認した.

(2) 極小生成子に基づく正負の相関ルールの集合を列挙するために, 極小生成子の下方閉包性を利用して, 極小生成子のみを頂点としてもつ探索木上で正負のルールを直接列挙する高速アルゴリズムを開発した. 抽出する正負のルールの集合はデータに潜在する知識を表現することから, それ自身は無矛盾かつ非冗長であることが望ましい. 抽出した多数の正負のルールを, 後から組み合わせて無矛盾かつ非冗長なルール集合を構成することは極めて非効率的である. そのため正負のルールに優先順位を仮定して, その優先順位に従って互いに無矛盾かつ非冗長なルールを逐次抽出するルール集合の漸近的構築アルゴリズムを開発した. 評価実験の結果, 良好な性能を確認した.

(3) 極小生成子集合を更に圧縮する手法として強飽和集合に着目し, 全ての強飽和集合を列挙するε近似オンライン型計算法を開発した. 先読みに基づく探索木の枝刈法などを開発し, 疎なデータと密なデータの双方において安定的に高速計算ができることを, 評価実験を通して確認している. 同時に強飽和集合とは別種の圧縮原理となりえる ε 近似分位数に着目し, その近似誤差を軽減するカウンタを用いたオンライン型高速列挙法を開発した. 最大誤差等に関する理論的性能解析と実装システムを用いた性能評価を行った. 評価実験を通して良好な性能を確認した.

(4) 負ルールの表現能力を本質的に上げる目的で, 正負のアイテムが混在する一般化アイテム

集合とその圧縮と高速抽出について研究を行った．まず一般化アイテム集合の飽和性について考察し，通常のアイテム集合よりも飽和性による圧縮・一般化が効果的に働くこと実証実験により明らかにした．次に拡張された接頭木探索の上に直接的に閉包計算を行うオフライン型アルゴリズムを開発し，実証実験により有用性を確認した

＜引用文献＞

[1] S.Brin, R.Motwani and C.Silverstein: Beyond Market Baskets: Generalizing Association rules to Correlations, *Proc. ACM SIG-MOD*, pp. 265-276 (1997)
[2]. X.Wu, C.Zhang and S.Zhang: Efficient Mining of Both Positive and Negative Association Rules, *ACM Transactions on Information Systems*, 22(3), pp.381–405 (2004)
[3] H.Wang, X.Zhang and G.Chen: Mining a Complete Set of Both Positive and Negative Association Rules from Large Databases, *Proc. of PAKDD ’08*, pp.777-784 (2008)
[4] M.-L. Antonie and O. R. Za ¨ıane: Mining Positive and Negative Association Rules: An Approach for Confined Rules. *Proc. Euro. Conf. on Principles and Prac. of Knowl. Discovery in Databases*, pp.27–38 (2004)
[5] M.Boley T, Horv_ath S, Wrobel: Efficient Discovery of Interesting Patterns Based on Strong Closedness, *Proc. SIAM 2009*, pp.1002-1012 (2009)
[6] V. Braverman et al.: BPTree: an L2 heavy hitters algorithm using constant memory, *Proc. PODS*, pp.1-15 (2016)
[7] 岩沼宏治，佐生隼一，黒岩健歩，山本泰生：負の相関ルール集合の極小生成子に基づく圧縮表現．情報処理学会論文誌, Vol.57, No.8, pp.1845-1849 (2016)

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1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 －

〔学会発表〕　計13件（うち招待講演　0件／うち国際学会　6件）

１．発表者名 Kento Yajima, Koji Iwanuma and Yoshitaka Yamamoto
２．発表標題 A Bottom-Up Enumeration Algorithm of Minimal Generators without Support Counting for Compressing Negative Association Rules
３．学会等名 2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI)（国際学会）
４．発表年 2022年

１．発表者名 Koji Iwanuma and Ryo Hinata
２．発表標題 A Fast On-Line ε-Approximation Algorithm for Mining Strongly Closed Itemsets
３．学会等名 2022 IEEE International Conference on Big Data (Big Data2022)（国際学会）
４．発表年 2022年

１．発表者名 前田浩丞，岩沼宏治
２．発表標題 カウンタを用いたε近似φ分位数のオンライン抽出の高性能化
３．学会等名 第21回情報科学技術フォーラム(FIT2022)
４．発表年 2022年

１．発表者名
Yuta Ando and Koji Iwanuma

２．発表標題
A Preliminary Study of Closed Generalized Itemsets and Their Enumeration Algorithms

３．学会等名
2021 10th International Congress on Advanced Applied Informatics (IIAI-AAI)（国際学会）

４．発表年
2021年

１．発表者名
Koji Iwanuma, Kento Yajima and Yoshitaka Yamamoto

２．発表標題
Enumerating Minimal Generators from Closed Itemsets: Toward Effective Compression of Negative Association Rules

３．学会等名
2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)（国際学会）

４．発表年
2021年

１．発表者名
日向涼，岩沼宏治，仁科拓巳

２．発表標題
ストリームデータからの強飽和集合をオンライン抽出するε-近似アルゴリズムの高速化

３．学会等名
2021年度人工知能学会全国大会（第35回）

４．発表年
2021年

１．発表者名
Koji Iwanuma, Kento Yajima, Yoshitaka Yamamoto

２．発表標題
Mining Consistent, Non-Redundant and Minimal Negative Rules Based on Minimal Generators

３．学会等名
2020 IEEE International Conference on Big Data (Big Data 2020)（国際学会）

４．発表年
2020年

１．発表者名
前田 浩丞、岩沼 宏冶

２．発表標題
カウンタを用いたε近似分位数サマリ構築の高速化に関する研究

３．学会等名
2020年度人工知能学会全国大会

４．発表年
2020年

１．発表者名
鍋島 崇宏、岩沼 宏冶

２．発表標題
極小生成子とその閉包アイテム集合のペアの高速列挙法

３．学会等名
2020年度人工知能学会全国大会

４．発表年
2020年

１．発表者名
日向涼，岩沼宏治，仁科拓巳

２．発表標題
ストリームデータからのΔ-強飽和集合のオンライン抽出

３．学会等名
第19回情報科学技術フォーラム（FIT2020）

４．発表年
2020年

１．発表者名
渡井慎一郎，岩沼宏治

２．発表標題
巨大イベント系列データの構築を目的としたオンライン型系列マイニングとその高速化

３．学会等名
第19回情報科学技術フォーラム（FIT2020）

４．発表年
2020年

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

〔図書〕　計0件

〔産業財産権〕

〔その他〕

潜在的規則の抽出を目的としたデータマイニングアルゴリズムの研究
http://www.kki.yamanashi.ac.jp/~iwanuma/data_mining_for_latent_rules

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

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関