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　　元 年　　６ 月 １４ 日現在

機関番号：１２６０１
研究種目：若手研究(B)
研究期間：2017～2018
課題番号：１７Ｋ１２７３６

研究課題名（和文）発見に関する統計的保証のあるパターンマイニング

研究課題名（英文）Statistically Sound Pattern Mining

研究代表者
小宮山　純平（Komiyama, Junpei）

東京大学・生産技術研究所・助教

研究者番号：２０７８００４２

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：パターンマイニングのアルゴリズムは、出現頻度の比率が一定以上大きい組合せ特徴量（パターン）集合を効率的に列挙する。しかし、データの確率的な偏りについては考慮がされず、興味のあるパターンが偶然の偏りによって出てきたものなのかが検証されない。本研究では、誤った発見をする確率に関して統計的保証を満たしたパターンマイニングの手法を提案し、データマイニングのトップ国際会議であるKDD2017において発表を行った。また、統計的保証の頑健性などについて研究を行った。いくつかのパターンがあるときに、そのうちのたまたまうまくいったものを選ぶ出版バイアスがどの程度の大きさになるか定量化することができた。

研究成果の学術的意義や社会的意義
データマイニングは知識発見を求める分野であるが、発見が統計的にどの程度の確からしさがあるのかは多くの場合考慮されていない。とくに、パターンマイニングはパターン（特徴量）の組合せの中から興味があるものを探すが、パターン数が多い場合には出版バイアスが発生し、得られたパターンが偶然の偏りなのか再現可能なものかの判断がつかない。この現状を鑑みて、本研究は得られたパターンのうち統計的に有意なものを探すアルゴリズムや、出版バイアスがどの程度大きくなりうるのかを定量化することで、データマイニングの知識発見としての健全性を保証するための基礎的な結果が得られたと考える。

研究成果の概要（英文）：Pattern mining algorithms enumerate all the combinatorial patterns with their frequency larger than a given threshold. Existing algorithms output many patterns that characterizes a dataset, they do not address how significant the found patterns are in terms of statistical significance. To address this issue, we propose a method that guarantees the rate of false discovery in found patterns while keeping its computational efficiency. The proposed method is presented in a top-tier data mining / artificial intelligence conference (KDD2017).

研究分野：機械学習

キーワード：パターンマイニング　統計検定　多重検定　データマイニング　機械学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

２つのデータセット D+,D-間での出現頻度が異なる特徴量の組合せ(組合せ特徴量)を考える。エマージングパターンマイニングのアルゴリズムは、出現頻度の比率が一定以上大きい組合せ特徴量（パターン）集合を効率的に列挙する。しかし、データの確率的な偏りについては考慮がされていないため、興味のあるパターンが偶然の偏りによって出てきたものなのか、統計的な有意性があるのかが検証されていない。とくに、特徴量の数が多くなると組合せの数も増えていくため、偏りがたまたま起こる組合せは高確率で存在すると予想される。本研究では、誤った発見をする確率に関して統計的保証を満たしたパターンマイニングの手法を提案する。また、提案手法を具体的なデータセットに活用するための枠組みを提案し、実装・検証を行う。

	x	y
A:	{1, 2}	+
B:	{1, 2, 3}	-
C:	{1, 2, 3, 4}	-
D:	{1, 3, 4}	+
E:	{1, 2, 6}	-

図 1: パターンマイニングの例。特徴量を x、D-と D+のどちらのデータセット由来かを y∈{-,+}と表現する。たとえば、パターン{1,2}は D-データセットに 3 度（B,C,E)、D+データセットに１度（A）現れているため、出現比は 1/3 となる。このように、２つのデータセットで出現比が異なる特徴的なパターンは、データセットの分類問題などに広く用いられる。

研究分野：データマイニング、統計

キーワード：パターンマイニング、多重検定、誤り確率、再現性、出版バイアス、順序統計

２．研究の目的

全体として､知識発見が偶然なのか､再現性があるものなのかを検証するという課題に挑んだ。データマイニングは知識発見を求める分野であるが、発見が統計的にどの程度の確からしさがあるのかは多くの場合考慮されていない。とくに、パターンマイニングはパターン（特徴量）の組合せの中から興味があるものを探すが､パターン数が多い場合には出版バイアスが発生し、得られたパターンが偶然の偏りなのか再現可能なものかの判断がつかない｡この現状を鑑みて、本研究は得られたパターンのうち統計的に有意なものを探すアルゴリズムや、出版バイアスがどの程度大きくなりうるのかを定量化することで、データマイニングの知識発見としての健全性を保証するための基礎的な結果が得られたと考える。また、この研究の発展として、出版バイアスの定量化に挑んだ。

３．研究の方法

本研究が完成するために必要なものは、フレームワーク、統計的保証のためのアルゴリズム、パターンマイニング・実装技術の３つであると考える。フレームワークとは、目的を含めた全体の枠組みであり、統計的保証は FDR を所与の水準に保つために個々の仮説の有意水準をコントロールし、データを代表する特徴量をどのように出力するかである。また、パターンマイニングは指数的に増えうる組み合わせをどのように効率的に探索するかの探索・実装技術である。これら３要素を協力研究者とのディスカッションにより進めていった。

４．研究成果

主要な実績としては、(i)KDD2017 において発表した統計的な保証のあるパターンマイニング手法、および(ii)発見の出版バイアスに関する論文公開レポジトリ arXiv での発表の２つである。パターンマイニングのアルゴリズムは、出現頻度の比率が一定以上大きい組合せ特徴量（パターン）集合を効率的に列挙する。しかし、データの確率的な偏りについては考慮がされていないため、興味のあるパターンが偶然の偏りによって出てきたものなのか、統計的な有意性があるのかが検証されていない。とくに、特徴量の数が多くなると組合せの数も増えていくため、偏りがたまたま起こる組合せは高確率で存在すると予想される。本研究では、誤った発見をする確率（False Discovery Rate, FDR）に関して統計的保証を満たしたパターンマイニングの手法を提案し、データマイニングのトップ国際会議である KDD2017 において発表を行った。
提案手法は、パターンマイニングの state-of-the-art 手法である LCM 法 [Uno+ 2004]を利用することによって計算的な効率を確保しつつ、FDR のコントロールを行う初のアルゴリズムであり、統計的な発見をこれまで知られている手法（LAMP 法 [Terada+ 2013]）よりも多く得られることが特長である。提案手法の良さは分類問題の public dataset における組み合わせパターンの発見で検証した。提案手法のソースコードをオープンソースソフトウェア共有サイト github で公開し、再現性の確保などに努めた。KDD2017 での発表の後の発展性としては、統計的保証の頑健性などについて着目した。これは、いくつかのパターンがあるときに、そのうちのたまたまうまくいったものを選ぶ出版バイアスがどの程度の大きさになるか定量化することが重要という考えに基づく。この成果を論文公開レポジトリ arXiv に投稿した。

５．主な発表論文等

〔雑誌論文〕（計　２件）

Junpei Komiyama, Masakazu Ishihata, Hiroki Arimura, Takashi Nishibayashi, Shin-ichi Minato. “Statistical Emerging Pattern Mining with Multiple Testing Correction.” Proceedings of the 23rd ACM {SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017 897-906. (査読あり、採択率 21%)
Junpei Komiyama, Takanori Maehara. “A Simple Way to Deal with Cherry-picking.” CoRR abs/1810.04996 (2018).

〔学会発表〕（計　２件）

〔図書〕（計　０件）

〔産業財産権〕
○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等

６．研究組織

(1)研究分担者
研究分担者氏名：小宮山　純平
ローマ字氏名：Komiyama Junpei

所属研究機関名：東京大学
部局名：生産技術研究所
職名：助教
研究者番号（8桁）：20780042

(2)研究協力者
研究協力者氏名：石畠 正和
ローマ字氏名：Ishihata Masakazu
研究協力者氏名：有村 博紀
ローマ字氏名：Arimura Hiroki
研究協力者氏名：西林 孝
ローマ字氏名：Nishibayashi Takashi
研究協力者氏名：湊 真一
ローマ字氏名：Minato Shin-ichi
研究協力者氏名：前原 貴憲
ローマ字氏名：Takanori Maehara