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　５ 年　 ６ 月 ２７ 日現在

機関番号：１２３０１
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ１１９６４

研究課題名（和文）自己展開型の知識発見による大規模データからの説明可能な知識創出

研究課題名（英文）Evolutionary knowledge discovery to create explainable knowledge from large data sets

研究代表者
嶋田　香（Shimada, Kaoru）

群馬大学・情報学部・教授

研究者番号：２０４５４１００

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：大規模データから個別の事例の特徴を最大限に説明すると考えられる知識を従来のモデル構築の過程を経ずに発見する局所的知識発見方法を提案した。説明可能な局所的知識表現として統計的な特徴を背景として持つアイテム集合を定義するとともに、その特性や発見実施における信頼性・再現性を評価した。また、発見された局所的知識集合体からの大域的知識創出方法の開発に取組み、従来手法とは異なる発想のクラスタリング的な技術、統計的な特徴を有する小集団を構成してその組合せで大域的な知識を表現しようとする技術の開発を行い、公開されている医療系データ等を用いて説明性や個別性からの知識創出の可能性を見出した。

研究成果の学術的意義や社会的意義
説明可能な局所的知識表現として提案した統計的な特徴を背景として持つアイテム集合（ItemSB：Itemsets with Statistically Distinctive Backgrounds）はデータ分析におけるアイテム集合ベース手法と統計学的方法を橋渡しして扱うことができるため、大規模データの分析に統計学的手法を効果的に導入可能とする技術と位置付けられる。大規模データの分析を統計的な背景をもつ小集団の組合せで扱おうとする新たなクラスタリング手法や、小集団にみられる統計的な特徴の連結により大域的な知識を扱おうとする方法の開発などの実用的で発展性のある独自方式による技術開発である。

研究成果の概要（英文）：In this study, we proposed a local knowledge discovery method to discover knowledge from large-scale data that is thought to best explain the characteristics of individual cases without going through the process of conventional model building. We defined an itemset with statistical characteristics as an explanatory local knowledge representation, and evaluated their properties, reliability, and reproducibility in implementing the discovery. We also developed a method for global knowledge creation from the local knowledge set that was discovered, and developed clustering techniques that are different from conventional methods, and techniques that attempt to express global knowledge by combining small groups of statistical characteristics. It was found to have the potential for knowledge creation from explanatory and individual characteristics using publicly available medical data and other data.

研究分野：知能情報学

キーワード：知識発見　データマイニング　知識創出　進化計算　アイテム集合　説明可能性

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

人工知能（AI）技術が急激に進化し次の技術ブレークスルーへの期待も高まっていた。技術的には医療診断や自動運転を含む様々なAI応用システムで人間を上回る精度や高度なプロセスの自動化が実現しつつあるものの、最大の課題はそのブラックボックスにあり、その解決が求められていた。研究代表者らは成果蓄積型の問題解決を特徴とする独自の進化計算による If～Then～型ルール表現知識の発見手法群を提案しており、これらを基礎としてAI技術の最大の課題とされる説明可能なAIの一つの在り方をルール表現の拡張の観点から検討し、とくに大規模データの分析の基礎となる技術開発を展開することを構想した。研究代表者らは、If～Then～型ルール表現知識の後件部に１つの連続変数の分布に関する条件が入る場合についての知識発見法を提案していた。この方法の拡張や一般化にあたり、大規模データを扱うにあたって全体的な分析を一気に実施しようとするのではなく、はじめにルール表現の特性から局所的な知識を扱うこととし、獲得された局所的知識群から大域的な知識を構成していくことが考えられた。また、研究代表者らの方法は、ルール発見の成果蓄積を継続的に実施して問題解決に必要なルール集合体を発見しようとする特徴をもち、ルール発見の実行条件設定を柔軟に行うことができることから、大規模データから個別の事例の特徴を最大限に説明すると考えられる知識を従来のモデル構築の過程を経ずに発見することが可能であることを見出していた。

２．研究の目的

本研究課題の申請時における当初の研究目的として、大規模データから個別の事例の特徴を最大限に説明すると考えられる知識を従来のモデル構築の過程を経ずに発見する局所的な知識発見方法を提案すること、獲得された局所的知識群から説明可能な大域的な知識の創出のための進化計算を応用した手法を開発すること、物理的・機械的データ（強い判別性・再現性あり）ではなく人間的・ライフ支援的データ（強い多様性・個別性あり）を対象とした説明可能なAIの基礎技術を開発することとした。

３．研究の方法

研究代表者らは、本研究課題に応用可能と考えられるAI基礎技術として、進化計算を用いたIf～Then～型ルールの知識発見手法群を提案している。ルール表現に連続変数の特徴的な分布の条件を含む知識の発見、３個のルールの組を用いて例外事象を表現する知識の発見、２つの異なる設定条件下での差異を表すルールの発見など、実用上柔軟な知識表現を実現する特徴をもっている。本研究課題においては、これらの手法の拡張や一般化が基礎となる。

局所的な知識発見手法として、個別事例に関する説明可能な知識発見手法を確立する。この方法は、ある目的のために構築されたデータベースがあるとき、同形態の新たな事例に対して、発見されたルール（属性の組合せ）を根拠にして、この事例と類似のものがどの程度あるのかを知ることができる。発見されるルールの個数が少なければ、個性の強い事例と考えることができ、ルールが発見できない場合（発見されても極少数の場合）は外れ事例と判断できるものであり、個別対応の知識発見を実現する方法である。この知識表現の定義と発見方法･評価方法を検討し、説明可能な状態で出力するプログラムを作成して公開データによる評価を実施する。また、自己展開型の大域的な知識創出アルゴリズムを検討し、プログラムを作成する。局所的な知識を入力データとして得られる結果について分析・評価する。とくに、公開されている大規模データ等を用いて幅広く評価する。また、開発したアルゴリズムの拡張・応用法を検討する。とくに個別性の観点からの知識発見が期待される人間的・ライフ支援的データからの知識発見を実施・評価するために、これらのデータから獲得されると期待される知識の表現および説明性を検討する。その実施結果について、説明可能な知識となっているか、経験知の見える化となっているか、等を各分野の専門家の協力を得て検証する。これらの結果から実際の複雑なデータを用いることで知識創出手法としての提案手法が説明可能なAIとなっているかを評価する。

４．研究成果

（1）統計的に特徴的な背景を持つアイテム集合（ItemSB: Itemsets with Statistically Distinctive Backgrounds）の提案

頻出アイテム集合の抽出とアソシエーションルールの利用はデータマイニングの基礎技術となっている。アソシエーションルールは、P → Q（if P then Q）の表現をとり、データベースの事例が前件部Pを満たせば後件部Qも満たすであろうと解釈される。従来手法では、Qは属性の出現頻度や注目クラスの占める割合を扱うにとどまっていたが、Q の部分を拡張一般化して「興味深いルール P → Q」を「Qという統計的な背景をもつアイテム集合P」を考えることができることを提案し、このP、Qの組を発見する方法を提案した。アイテム集合とこれを含む事例に注目することで、大規模データにおける局所的知識を表現することができる。また、この知識表現は、従来の頻出アイテム集合に基づいたデータ分析と統計学的分析を橋渡しするものと位置づけられた。ルールベースであることからの説明性と統計的なデータに基づく根拠を有する表現となる。

具体例として、後件部 Q について２次元に限定して扱うものとし、データベースの属性を A_i $(1 \leqq i \leqq n)$、X、Y とする。A_i は１または０の値をとり、X、Y は連続値をとり、X、Y には欠損値はないものとする。属性に関する組合せ$(A_j=1) \wedge \cdots \wedge (A_k=1)$ をアイテム集合 I として、I を満たすデータの小集団に注目する。この小集団における X、Y の平均値、標準偏差をそれぞれ $m_X(I)$、$s_X(I)$、$m_Y(I)$、$s_Y(I)$とし、X と Y の相関係数を $R_{X,Y}(I)$とする。また、I の出現頻度を support(I) とする。$m_X(I)$、$s_X(I)$、$m_Y(I)$、$s_Y(I)$、$R_{X,Y}(I)$、support(I) について予め設定しておいた条件が満たされるとき、I を統計的に特徴的な背景をもつアイテム集合として ItemSB（Itemsets with Statistically Distinctive Backgrounds）とよぶことにし、$(A_j=1) \wedge \cdots \wedge (A_k=1) \rightarrow (m_X, s_X, m_Y, s_Y, R_{X,Y})$ と表す。図１に基本的なアイデアとして$(A_1=1) \wedge (A_2=1)$が ItemSB となり X と Y の間に正の相関がある例を示す。

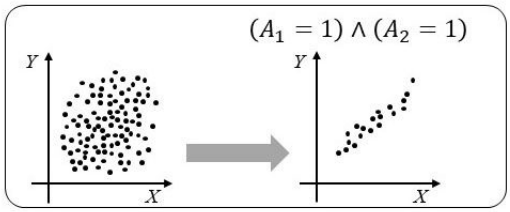

図1 ItemSB の基本的なアイデア

ItemSB の発見法として、有向グラフによるネットワーク構造を特徴とする GNP（Genetic Network Programming）の構成と各世代で獲得した成果を蓄積していく進化戦略をとる進化計算手法を GNMiner として拡張・一般化して提案した。GNMiner は一般的な進化計算手法と異なり最終世代の最優良個体を解とするのではなく、興味深いルール(ItemSB)の候補を進化計算における個体内に多数表現して探索した上で、興味深いルールが見つかった時点で、ルールライブラリに蓄積をしていくものである。このため、進化計算手法としては、進化操作時における適合度の設定方法が性能に与える影響が小さいこと、成果蓄積型の課題解決であるためルール発見の打ち切りが随時可能であること、個体のネットワーク構造の柔軟性により多様なルール表現が可能であること、などの特徴がある。ItemSB を抽出するためには、ItemSB の候補集団の作成、ItemSB の候補の評価を行う必要があり、GNMiner における GNP の個体内のノード接続による ItemSB の表現、および ItemSB の平均値や標準偏差、相関係数を求めるために、統計値の算出のための各ノードにおける諸値の記憶・管理の仕方を提案した。公開されている事例数 10 万件程度のデータと医療系のデータを用いて ItemSB の発見実施を行い、提案手法の性能や特性を評価した。

研究成果は、2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering で発表し、２つのデータベース間での同一 ItemSB の差異の評価に関する研究成果を含めて International Journal of Semantic Computing, 16 (2022) に掲載された。

（2）局所的知識としての ItemSB の２次元予測問題への応用

２つの連続属性が局所的な分布となる場合について、ItemSB の発見と、発見された ItemSB を用いた２変数 X、Y の値の組を予測する実験を行った。ItemSB の統計的特徴として、２次元の極めて狭い範囲となるように事前に条件設定するものであり、具体的問題として地理的に狭い領域にみられる音楽的特徴を与える属性の組合せを発見しようとする問題を扱った。１個の ItemSB は位置情報としての m_X、m_Y の値とその位置と関係する音楽的特徴の組合せの情報をもつため、その位置の示す地域に特徴的な音楽的背景の分析に利用可能と考えられた。図２に発見された音楽的情報の組とその地理的位置を表す ItemSB の例を示す。

また、２次元の回帰を扱う問題として、１個抜き交差検証を用いて、1059 曲について 1058 曲から得られた ItemSB とのマッチングにより１曲の X、Y の値の組を予測する（曲の由来となる地図上の位置を予測する）実験を行った。音楽的属性情報の全体を用いて発見された 2000 個の ItemSB の利用では、予測時のカバー率は 14.5%であったが、予測対象曲の持つ音楽的属性情報のみを対象として個別の事例に対応した ItemSB 発見に基づいた予測では、他の条件を同一にしたところカバー率が 93.4%と向上し、個別の事例の特徴を最大限に説明する方法の利用の優位性が示された。なお、予測の精度に関しては曲が複数の音楽的背景を有する場合には複数地点の予測地候補が生じることから予測の困難さが考えられた。

研究成果は、The 33rd IEEE International Conference on Tools with Artificial Intelligence で発表した。

$support(I) \geq 0.007, s_X(I) \leq 5, s_Y(I) \leq 5$
(2,000 rules)
図2 発見された ItemSB の例

（3）大域的知識発見の実施方法

ItemSB の発見は、大規模データにおいて統計的に特徴的な小集団（グループ）の発見を目的とする場合においても有用と考えられる。大規模データにおいては、データをいくつかの小集団にわけて分析や予測に用いることが考えられるが、こうした小集団の選別をどのように実施するかは扱う問題によって異なってくる。本研究課題では、小集団ごとに成立する一次の回帰式などに注目して大域的知識発見法を検討した。ItemSB では、小集団としてのグループが構成された根拠がアイテム集合のもつ統計的背景（関連性など）として貼り付けられている。この小集団

は、クラスタリングによって得られる小集団とは異なる特性をもつものであり、大規模データ全体をいくつかのItemSBで分類してカバーすることは目的が異なっている。従来のクラスタリングにおいてはクラスタリング後に各小集団のラベル付けを行うことになるが、ItemSB 利用の場合には説明可能なラベル付きでの小集団が得られる。各事例が単一の小集団にのみ属するのではなく重複が生じ得ることも相違点である。

ItemSB の発見は大規模データにおける小集団の発見をしようとするものとも考えられるが、その小集団のもつ信頼性・再現性の評価が課題と考えられた。同様の目的や方法で収集されたデータベースが複数あるとき、１つのデータベースにおいてItemSBとされたアイテム集合が別のデータベースにおいて与えられた条件を満たしているかどうかはItemSBの応用において重要になると考えられたことから、２つのデータベース間での同一ItemSBの統計値の差異を評価するためのコントラストItemSB発見法を提案し、アイテム集合の出現頻度に基づく、信頼性・再現性の傾向を見出すことができた（図３）。この評価実験結果を含む研究成果について大規模データの分析を専門分野とする国際会議に論文投稿を行った。

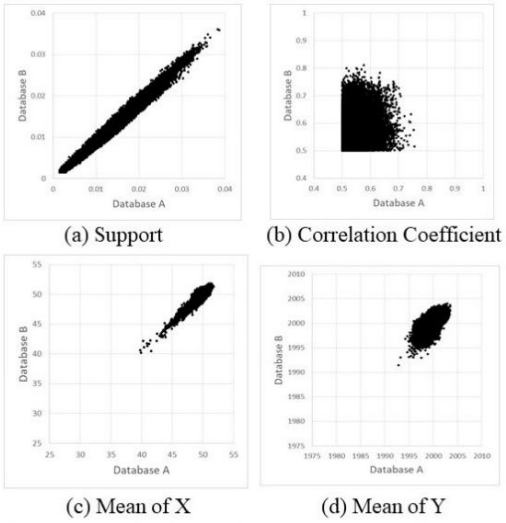

図3　2つのデータベース間での ItemSB の差異の例

（4）GNMinerの進化操作と発見性能特性に関する指標の提案

知識発見手法である GNMiner は、ルール発見の成果蓄積を継続的に実施して問題解決に必要なルール集合体を発見しようとする進化計算であり、ルール発見の実行条件設定を柔軟に行うことができる。進化操作時の確率設定などのパラメータについては経験上の効果的な代表的設定値が報告されていたが、各値の設定がルール発見等に及ぼす効果や影響についての研究報告はほとんどなかった。本研究課題では、大規模データを想定していることからデータベースの属性数が多くなることがあり、ルール発見を乱数系列を変えて繰り返すとき進化の過程にバラツキがみられるケースが生じた。具体的には、成果蓄積が効果的に開始される世代数が遅れ気味となるケースがあること、早い世代で成果蓄積が効果的に開始されてもその後の新しいルールの発見効率が低下するケースがあることである。これらの性能特性を特徴づける指標として、候補となったルールが新規かどうかの判定を行う際の問合せ時の処理に関する値を定義したところ、ルール発見の効率と進化の過程の関係を特徴づけることができることがわかった。また、GNMinerにおける発見の効率に関する進化の過程が３つのステージで把握できることを見出した。この研究成果は、The Genetic and Evolutionary Computation Conference (ACM GECCO'22) で発表した。多数の進化パラメータの設定を用いた進化の過程についての評価実験と特性の分析を加えて国際論文誌に投稿した。

（5）得られた成果の国内外における位置づけ

本研究課題で提案した ItemSB（統計的に特徴的な背景を持つアイテム集合）は、従来の頻出アイテム集合に基づいたデータ分析と統計学的分析を橋渡しするものであり、その発見法とあわせて大規模データの分析における新たな基礎技術となり得ると評価されている。ItemSB の発見法である GNMiner は、独自の進化計算に基づいた手法であり欠損値を含んだままでのルール発見の実施が可能であることなどから大規模データの分析への適応性がある。得られた成果は、大規模データからの説明性と統計的表現を有する新たな知識創出技術の提案と位置づけられる。

（6）今後の展望

実際に複雑なデータを収集して知識発見・知識創出を実施し、その実施結果について説明可能な知識となっているか、経験知の見える化となっているか等の検証をすることは、コロナ禍による影響のためできていない。こうした各分野の専門家の協力を得て提案手法を検証するというような発展的な内容の実施が今後の課題として考えられる。また、ItemSB を用いた大規模データからの知識手法の応用展開として、これまで研究されてきた頻出アイテム集合ベースの応用手法を参考とした、より高度な応用分析法の開発などが考えられる。例えば、アイテム集合を基礎としたデータ分析法はテキスト分析などとの融合がみられることから、テキストデータを含む大規模データの分析への応用が考えられる。多様なデータで構成される高次元の統計的な特徴的空間の発見法の開発、小集団での特徴的な交絡を発見的に扱うことや時系列のアイテム集合などの導入により因果推論問題を扱おうとする技術の開発など、本研究課題の成果を基礎として新たな研究課題を設定して取組んでいくことが期待される。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Shimada Kaoru、Arahira Takaaki、Matsuno Shogo	４．巻 16
２．論文標題 ItemSB: Itemsets with Statistically Distinctive Backgrounds Discovered by Evolutionary Method	５．発行年 2022年
３．雑誌名 International Journal of Semantic Computing	６．最初と最後の頁 357～378
掲載論文のDOI（デジタルオブジェクト識別子） 10.1142/S1793351X22420028	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Matsuno Shogo、Shimada Kaoru	４．巻 Companion
２．論文標題 Evolutionary operation setting for outcome accumulation type evolutionary rule discovery method	５．発行年 2022年
３．雑誌名 Proc. of the Genetic and Evolutionary Computation Conference (ACM GECCO'22) Companion	６．最初と最後の頁 451～454
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3520304.3528974	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Shimada Kaoru、Arahira Takaaki、Matsuno Shogo	４．巻 －
２．論文標題 Evolutionary Method for Two-dimensional Associative Local Distribution Rule Mining	５．発行年 2021年
３．雑誌名 Proc. of The 33rd IEEE International Conference on Tools with Artificial Intelligence	６．最初と最後の頁 1018～1025
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ICTAI52525.2021.00163	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Kaoru Shimada, Takaaki Arahira, Shogo Matsuno	４．巻 －
２．論文標題 Evolutionary Method to Discover Itemsets with Statistically Distinctive Backgrounds	５．発行年 2021年
３．雑誌名 Proc. of 2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering	６．最初と最後の頁 113～120
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/AIKE52691.2021.00024	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計5件（うち招待講演　0件／うち国際学会　3件）

１．発表者名
Shogo Matsuno, Kaoru Shimada

２．発表標題
Evolutionary Operation Setting for Outcome Accumulation Type Evolutionary Rule Discovery Method

３．学会等名
Genetic and Evolutionary Computation Conference (ACM GECCO'22)（国際学会）

４．発表年
2022年

１．発表者名
嶋田香，松野省吾，荒平高章

２．発表標題
統計的に特徴的な背景を持つアイテムセットを発見するための進化計算方法

３．学会等名
2022年度 人工知能学会全国大会（第36回）

４．発表年
2022年

１．発表者名
嶋田香

２．発表標題
２つの連続変数間の統計的特性を背景とするアイテム集合

３．学会等名
2022年度統計関連学会連合大会

４．発表年
2022年

１．発表者名
Kaoru Shimada, Takaaki Arahira, Shogo Matsuno

２．発表標題
Evolutionary Method for Two-dimensional Associative Local Distribution Rule Mining

３．学会等名
The 33rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)（国際学会）

４．発表年
2021年

１．発表者名 Kaoru Shimada, Takaaki Arahira, Shogo Matsuno
２．発表標題 Evolutionary Method to Discover Itemsets with Statistically Distinctive Backgrounds
３．学会等名 2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)（国際学会）
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

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	荒平　高章 (Arahira Takaaki) (30706958)	九州情報大学・経営情報学部・准教授 (37120)	

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