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 ２４ 日現在

機関番号：３７１２９
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００３１６

研究課題名（和文）人工的欠損値を導入した大規模データにおける知識発見

研究課題名（英文）Knowledge discovery method in large scale data using artificial missing values

研究代表者
嶋田　香（Shimada, Kaoru）

福岡看護大学・看護学部・教授

研究者番号：２０４５４１００

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：データベースに人工的欠損値を導入することを特徴とする知識発見方法の研究開発を実施した。人工的欠損値は、ある方策によってデータベースの情報を間引いて隠すなどしたものである。IF～THEN型ルールベースの予測・分類問題における数値属性の離散化時の境界値付近の値を欠損値として扱うことを特徴とする方式を提案し評価した。また、人工的欠損値の利用の配置の方策によっては、取得されたデータの個々の値に関しての信頼性評価への応用の可能性があること、データベースにおける特定の属性あるいは属性の組合せが、目的とする属性と因果関係にあるとした場合の評価指標の算定への応用の可能性があることを見出した。

研究成果の学術的意義や社会的意義
判断根拠を解釈しやすいルールベース手法であること、予測・分類精度の向上や対応事例数の制御ができることなどから、予測・分類問題における信頼性・有効性の改善が望める他、情報の選択的利用への理解の深化が期待される。人間の発想に似た知識発見法の提案に係わる研究であり、人間に似た判断法という観点からの知能情報処理分野における波及効果が予想される。ビッグデータにおけるデータの取捨選択や構造理解を実現する基礎技術になりえること、実社会における判断支援システムへの応用が検討できることなど、実用的で発展性のある技術開発である。

研究成果の概要（英文）：In this study, we proposed a rule-based classification method that uses artificial missing values to improve the effectiveness and precision of data analysis. We apply artificial missing values to avoid the sharp boundary problem encountered when discretizing continuous variables. In discretization, we treat attribute values near the boundary as missing values. We evaluated the performance of the proposed artificial missing value-based classification method and our experimental results using medical data showed this method to be effective for classification. In addition, we obtained new knowledge that led to a method of applying artificial missing values to causal analysis and a method finding similar cases in a large scale 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　ソフトコンピューティング　人工知能　欠損値

１．研究開始当初の背景

実世界のデータには、欠損値が含まれることが多くある。例えば、アンケート調査における収入・年齢等の未記入、実験時の不具合や未測定による場合、あるいは、同種のデータの統合時に欠損が生じる場合がある。従来の情報処理では、こうした欠損値を含むレコードを削除すること、あるいは、その属性の平均値や頻出事例を欠損値に変えて用いることがよく用いられる。欠損値をどのように補填するかという課題はデータの構造解析とともに重要な課題であり、多くの研究成果が発表されている。その一方で近年話題となっているビッグデータ解析においては、膨大な情報の活用の観点から欠損値を含んだままデータを扱う解析手法の開発が求められている。ビッグデータ解析においては、知識発見や機械学習の研究が活発に進められているが、欠損値の扱いを課題の中心に据えた研究はほとんど見られていない。

データマイニング技法の一つである相関ルール（Association Rule :IF～THEN～型ルール）では、アプリオリ法など、頻出アイテム集合の抽出をベースとした手法が主流であるが、これらには、欠損値の扱いが困難であるという課題がある。研究代表者らは、欠損値を含むデータから直接に相関ルールを高速に発見する方法[*1]を有向グラフ構造を特徴とした進化型計算手法を用いて提案している。これは欠損値を含むレコードの欠損値以外の属性の情報を利用するため、ルール指標が本来のデータの性質通りのものが得られる。また、従来の進化型計算手法が、進化の最終世代における最優良個体を解として課題解決を行う方式であるのに対し、研究代表者らのルール発見手法は、進化の過程を通して世代継続的に成果を蓄積して課題解決をしていく独自方式である。

研究代表者は、科研費課題 24500191・基盤研究（C）（H24-H26）「世代継続的な進化型計算手法による欠損値を含むデータからの知識発見に関する研究」において、データベースの欠損値推定アルゴリズムおよび人工的な欠損値を用いたルール抽出時の情報保護に関する研究を実施した。とくに人工的な欠損値をランダムに発生させることでデータベースの一部情報を隠し、その上でのルール抽出を行う情報保護に関する研究において、欠損値の発生割合がルール指標に及ぼす影響についての知見[*2]を得ている。この研究課題では、人工的欠損値をデータベースにランダムに適用した場合を扱っているが、ある方策のもとに人工的欠損値を用いることで、革新的な情報処理の実現の可能性があることを着想するに至った。具体的には、連続変量の離散化時の境界値付近を人工的欠損値として扱うことで処理から外すという着想を得た。例えば、予測・分類問題で学習データにおいては境界値付近の属性値を持つレコードはルール指標の算出からその都度除外し、一方、テスト事例のルールマッチングにおいては、境界値付近の値を用いたルールによる判断は避けて別のルールを用いて判断しようとするものである。研究代表者らが 40 万件程度のデータ（270 属性）を用いて試験的に評価したところ、人工的欠損値の導入により連続値の推定精度向上とルールマッチングで予測可能なテスト事例数割合増大を制御できる可能性を示す結果[*3]を得ている。

人工的欠損値を利用しようとする場合、人工的欠損値の量的・範囲的な与え方がルールの興味深さ指標に与える影響や、予測・分類での精度の向上、最適な組合せとなる属性ごとの人工的欠損値生成の設定法などの工学的な課題がある。新概念としての理論的な体系の整備や種々の指標の定義についても求められる。また、予測・分類の精度向上などに有効な人工的欠損値発生条件を見出すことは、データベースの属性間の関連性を含めたデータ構造的特性を説明する知識発見の基礎となるものと期待される。なお、人工的欠損値の導入が、データが本来持っている性質を正しく反映しているかどうかの検討には、研究代表者らがすでに開発している同一形態の２つのデータベース間のコントラストを特徴づける差異ルールの発見手法[*4]が応用可能であると考えられた。

予測・分類を行うに当たり、情報の一部を選択的に用いずに、よりよい情報に基づこうとする方策は、人間の思考様式に近いものと考えられる。ビッグデータでは情報の取捨選択などによるデータの有効活用が課題であり、人間の思考様式に似たデータの扱われるべき理想形や処置法の観点から、人工的欠損値の活用方策の検討も有益と考えられる。情報のあいまいな部分の効果的な利用法としては、ファジィを用いる方法があり、本研究で提案する方式との特性の比較を行う。なお、人工的欠損値によって本来利用可能であった情報を使用しないことに倫理面での検討の必要性も指摘できることから、情報の選択利用に関する信頼性指標の導入も課題と考えられた。

*1 K. Shimada and K. Hirasawa, A Method of Association Rule Analysis for Incomplete Database Using Genetic Network Programming, Proc. of the ACM Genetic and Evolutionary Computation Conference (ACM GECCO'10), pp.2673-2680, 2010.

*2 K. Shimada and T. Hanioka, An Evolutionary Method for Exceptional Association Rule Set Discovery from Incomplete Database, Lecture Notes in Computer Science (Information Technology in Bio- and Medical Informatics), Vol.8649, pp.133-147, 2014.

*3 K. Shimada, T. Arahira and T. Hanioka, An Evolutionary Rule Mining Method for Continuous Value Prediction from Incomplete Database and Its Application Utilizing Artificial Missing Values, Proc. of the First IEEE International Conference on Big Data Computing Service and Applications, pp.392-399, 2015.

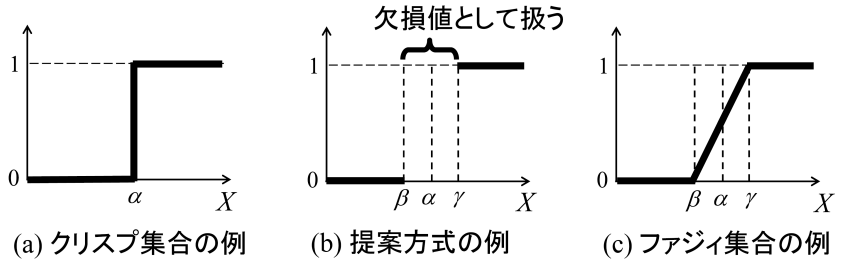
*4 K. Shimada and T. Hanioka, An Evolutionary Associative Contrast Rule Mining Method for Incomplete Database, Proc. of the International Conference on Data Mining (DMIN 2013), pp.160-166, 2013.

２．研究の目的

研究代表者らが提案した欠損値を含むデータベースから IF～THEN～型のルールを発見する手法群を基礎として、人工的欠損値を利用した連続変量の離散化時における最適化技術を開発する。あいまいな情報を人工的欠損値として扱うことで人間の思考・判断に似た柔軟な予測・分類を行おうとするシステムを提案し、ビッグデータの予測・分類問題における信頼性指標を向上させたルールベースの手法を提案する。与えられたデータベースに最適な人工的欠損値の配置を実現する手法を提案するとともにその有効性を実データを用いて検証する。また、実社会に応用可能な人工的欠損値を用いた人間の思考様式に似た知識発見法の開発に取り組む。ルールの生成時における離散化境界値付近の値の扱いや、ルール利用時の手元の数値が離散化境界値付近である場合等、判断に迷うと考えられる場合において、判断を一旦保留しておき、よりよい情報に基づいて処理しようとするといった、人間の発想に似た知識発見法を開発する。さらに、人工的欠損値の最適利用となる条件設定を獲得することから、逆にデータベースにおける属性相互の関係性や、ある属性のとる値の領域の予測・分類に与える影響の把握といったデータ構造に関する知識発見法を提案する。ビッグデータにおける選択的な情報利用に関する属性特性の指標を導入して評価することで、データベースにおける構造的特性を説明する知識発見方法を開発する。

３．研究の方法

人工的欠損値は､ある方策によってデータベースの情報を間引いて隠すなどしたものである。例えば、連続変量 X の離散化において、図１(a)では、境界値αを境にして属性値が２値化されているが、人工的欠損値利用の具体例として、図１(b)に示すような連続変量の離散化時の境界値付近の値（β＜X＜γ）を欠損値として扱うことができる[*3]。



(a) クリスプ集合の例　(b) 提案方式の例　(c) ファジィ集合の例

図1　連続変量の離散化

はじめに、与えられたデータベースに最適な人工的欠損値利用の方法を提案し評価する。ここで、最適であるか否かは、予測・分類精度の向上、テスト事例における予測・分類可能な事例数の増加などの目的に応じた評価指標に基づいて判断する。具体的には、人工的欠損値の量的・範囲的な与え方がルールの興味深さ指標に与える影響や推定での精度の向上を評価し、最適な人工的欠損値生成の設定法、属性間の欠損値発生条件の組合せ最適化などのアルゴリズムを提案する。評価用プログラムを作成し、これらの検証を公開データを用いて実施する。これは、従来より境界値の設定法が広く予測・分類問題における課題として認識されており、ルールベースのシステムを想定した場合には、その性能に大きな影響を及ぼすことになるためである。また、数学的表現や評価指標の確立、人間の判断に似ているかどうか等の評価を行う。従来手法として､境界値付近をあいまいさを含む情報として図１(c)のように確率的に扱う方式としてファジィの利用があり、開発手法の評価比較対象として検討する。

ルール発見手法については、研究代表者らがこれまでに提案している世代継続的な進化型計算を応用したルール発見アルゴリズム群を利用する。IF～THEN～型ルールについては、柔軟なルール表現法とその指標を採用し、大規模データにおける特徴的な小分布の発見[*3]や３個のルールの組合せにより例外的なパターンをルール表現とする手法[*2]などを発展的に用いる。

研究の方向性として、ビッグデータにおける情報の選択的利用法の開発、および、データベースの構造を説明するための解析手法の開発を目指すものとする。最適化アルゴリズムの簡易化・高速化や手法の拡張を行いつつ、大規模データにおけるデータの取捨選択に資する技術を開発する。IF～THEN～型ルールやその拡張されたルール表現による知識表現は可読性を有するため、判断根拠の説明が重要視される分野のデータからの知識発見や知識表現に適するものと考えられる。医療・福祉分野における大規模調査や判断支援システム構築等を想定した情報の選択利用を検討するとともに、こうした分野の実データを用いた手法の検証を行う。

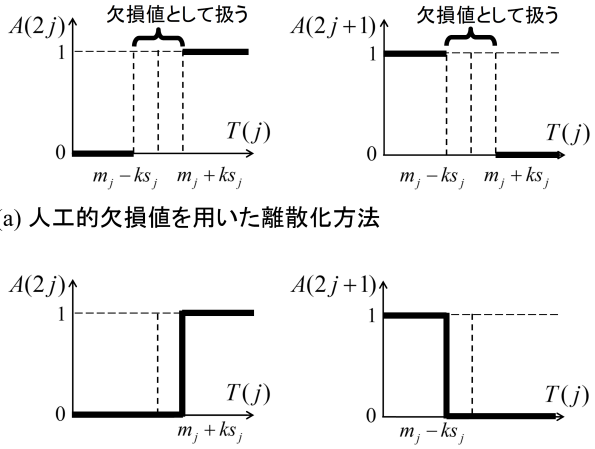
また、人工的欠損値を利用した予測・分類問題を扱う過程で、予測・分類の精度向上などに有効な人工的欠損値発生条件を見出すことにより、データベースの属性間の関連性を含めたデータ構造に関する知識の獲得が期待される。特定の属性のとる値のある領域、あるいはこの組合せが、別の属性のとる値の領域に与える影響等を把握することでデータベースの構造特性を説明する方法を開発する。ここでは、データが人間の思考様式と同様に扱われるべき理想形といったものの考察を含めていくとともに、情報の選択利用に関する信頼性の指標や説明方法を提案する。

４．研究成果

（1）人工的欠損値を用いたルールベースの予測・分類問題の性能改善アルゴリズムの提案

ルールベースの予測・分類問題における数値属性の離散化において離散化時の境界値付近の値を欠損値として扱うことを特徴とする方式を比較的小規模の医療データ等を用いた評価した。これまでの研究では、人工的欠損値をデータベースにランダムに導入した場合を評価実験として扱っており、情報の一部秘匿などの利用可能性が考えられていた。本研究課題では、数値属性の離散化時の境界値付近を人工的欠損値として扱うことで処理から外すことに利用する。具体的には、学習データにおいては境界値付近の属性値を持つ事例は該当するルールの指標算出から外し、一方、テスト事例のルールマッチングにおいては、境界値付近の値を用いたルールによる判断は避けて別のルールを用いた判断により分類を実行する。

予測・分類手法の評価に UCI ML Repository の公開データ(http://www.ics.uci.edu/~mlearn/MLRepository.tml.)を用いた。ここでは、Pima Indians Diabetes Database を用いた評価の例を示す。このデータベースは、768人についての8個の数値属性と２つのクラスをもつクラス属性に関するもので、分類精度は 76%程度と報告されており、人工的欠損値の導入が分類性能に与える影響を評価するのに適しているものと考えられた。データベースから、90%を学習データ、残りの10%をテストデータとする評価実験用のデータをランダムに30組作成し、評価結果はこれらの平均を用いることとした。なお、従来の手法として、境界値付近をあいまいさを含む情報として図１(c) のように確率的に扱うファジィの利用があるが、医療系データなどの場合には、確率的な扱いが好ましくないと考えられる場合もある。



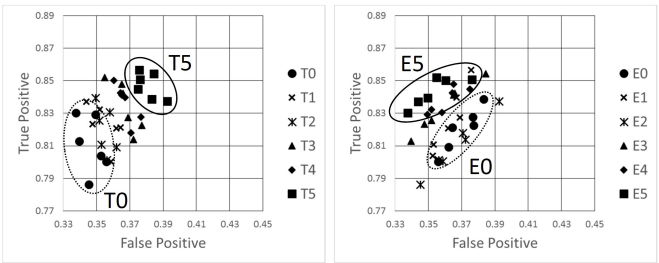
(a) 人工的欠損値を用いた離散化方法

(b) 比較のための離散化方法

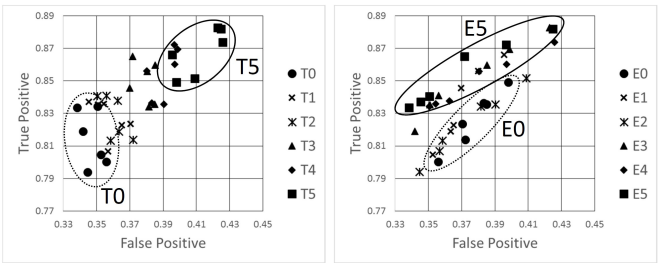
図2： 人工的欠損値の設定

数値属性T(j) の平均 m_j と標準偏差 s_j を用いて図２(a) のように人工的欠損値を導入して離散化を行った。ここで、k は人工的欠損値を導入する範囲に関する定数で 0、0.1、0.2 、0.3、0.4、 0.5 の 6 通りとした。学習データとテストデータについて各 6 通りを設定することで、36 個の組合せによる評価を行った。なお、この離散化は人工的欠損値の導入による分類性能などの評価を目的としており、臨床的な観点からの意味はない。

IF～THEN～型ルールの結論部をクラスの属性値としたルールを定義し、ルール抽出とクラス分類器構築は代表者らが従来用いている方法によった。進化計算手法における個体数や進化操作時の確率、最終世代数の設定はについても同一とした。抽出するルールの指標の満たす条件として、最小支持度 0.03、最小カイ二乗値 6.63 を用いた。また、人工的欠損値を利用しない従来の方式と比較する目的で、図２(b) で示される離散化を用いた実験を行った。結果は以下の通りであった。



(a) 人工的欠損値を用いた場合



(b) クリスプ集合として離散化した場合

図3： 分類問題における性能評価

・学習データから抽出されるルール数は、k の値が大きくなるに従って減少するが、人工的欠損値を導入した場合の方が、従来の方式よりも抽出されるルール数は多かった。

・分類の正確さ（クラスを正しく分類できた事例数の割合）は、従来の方式では k の値が大きくなるに従って低下するが、人工的欠損値を導入した場合は k の値によらずほぼ一定であった。

・人工的欠損値を導入した場合、従来の方式と比較して少ないルール数で未知の事例を分類可能な分類器を構築できる可能性が示された。

・人工的欠損値を導入した場合、従来の方式と比較して少ないルール数で同等の精度を有する分類器を構築できる可能性が示された。

図３は、人工的欠損値を導入した場合とそうでない場合の分類の精度について 36 個の組合せ

の結果を示したものである。左側が学習データに注目したもの、右側がテストデータに注目したものである。学習データをT、テストデータをEとして、k=0の場合をT0、E0、k=0.5の場合をT5、E5と表している。人工的欠損値の導入により真陽性の割合を大きくできる可能性があることがわかる。また、人工的欠損値を導入した場合の方が従来の方式より精度のばらつきが小さく、偽陽性の割合が小さい傾向にあることがわかる。

これらのことから、人工的欠損値の導入により予測・分類方法の性能向上を実現可能であることが示された。これは、判断に迷うと考えられる場合において、判断を一旦保留しておき、よりよい情報に基づいて処理しようとする、人間の発想に似た知識発見法の開発と位置づけられる。国内外の学会発表では、多くの参加者から新しい発想によるデータの扱いに対して興味が示され、技術的な内容について議論を深めることができた。

（2）実社会に応用可能な人工的欠損値を用いた人間の思考様式に似た知識発見法の提案

本研究課題においては、実社会のデータの分析への応用に向けて、ルール発見法の改善・拡張も進めた。提案方式においては、IF X THEN Y 型のルールの評価時にIF (not X) THEN Y 型のルールの評価を同時に実施することができる。このためルールの評価指標にオッズ比などの医療分野で利用される指標の算出を実現できる。また、IF X1 THEN Y と IF X2 THEN Y は興味深いとはいえないが、IF (X1 かつ X2) THEN Y は興味深いというような例外的な属性の組合せ(X1 かつ X2)の発見が可能となる。従来の手法においては、属性の組合せを扱おうとする場合、属性間の独立性を仮定した確率計算が用いられることがある。提案手法では、属性の組合せに関する評価を組合せごとに行うことができるため、ルール表現における属性の組合せの評価に対する信頼性とその理解のしやすさを検討することが可能となった。実社会における属性数の多いデータベースで利用しようとする場合について拡張した方法の検討･評価を行うとともに、与えられたデータベースに最適な人工的欠損値生成の設定法、属性間の欠損値発生条件の組合せ最適化などのアルゴリズムの検討を行った。種々の検討の結果、属性間の関連性の扱いなど新たな課題が具体化したことから、課題を整理して発展的に取り組むこととした。

また、実社会への応用の観点から、本研究課題で作成・整備したプログラム群を用い、所属機関の研究者らと協力して比較的大規模なデータを用いた医療専門家の視点からの発見されたルール集合の解釈や分類結果等の検討・評価を行った。口腔の健康評価に関するデータ、在宅高齢者の健康情報に関するデータを用いた知識発見の実施例について国際会議での発表を行った。

（3）ビッグデータにおける情報の選択的利用法とデータ構造に関する知識発見手法の提案

予測・分類の精度向上などに有効な人工的欠損値の導入条件を見出すことによるデータベースの属性間の関連性を含めたデータ構造に関する知識発見方法を含めた検討・評価を行ったほか、人工的欠損値の利用の配置の方策によっては、取得されたデータの個々の値に関しての信頼性評価への応用の可能性があること、データベースにおける特定の属性あるいは属性の組合せが、目的とする属性と因果関係にあるとした場合の評価指標の算定への応用の可能性があることについて検討した。因果関係理解への応用に繋がると期待される方法の基本的なアイデアは次のようなものである。データベース（DB）における属性Xiは値として0、1、m（欠損値）のいずれかの値をとるものとした場合、

・何も操作をしないオリジナルのDB
・Xi=0の事例を削除したDB　　・Xi=0の事例の属性値をXi=mで置換したDB
・Xi=1の事例を削除したDB　　・Xi=1の事例の属性値をXi=mで置換したDB

を考えて、属性Yとの関連性の状況から因果関係に関する指標を定義して扱おうとするものである。

また、データベースにおける類似事例の発見の方法についての知見も得た。これは、データベースの１事例に注目するとき、これと類似の事例群を探索し、類似性の根拠となる属性組合せを発見する方法であり、具体的には、注目する事例の属性値に応じて、データベースの属性値を人工的欠損値を活用しながら変換して､注目事例以外を探索するものである。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、同形態の新事例に対して、これの類似事例の発見、これが外れ事例であるかの判断に応用可能であり、個人対応の知識発見の基礎となりえる。また、予め分類器を構築せずに注目事例のクラス分類の参考となる根拠発見を期待できる。ビッグデータにおいて類似事例となる小集団を属性組合せというラベル付きで発見することができることになることから、ビッグデータにおける情報の選択的利用法とデータ構造に関する知識発見手法に繋がるものと考えらる。これらの新たな知見に関しては、検討・評価をすすめて結果を国内外において発表していく予定である。

５．主な発表論文等

〔雑誌論文〕（計３件）

① Kaoru Shimada, Hisae Aoki, Keiko Kubota, Satoru Haresaku, Shinsuke Mizutani, Toru Naito, Michio Ueno, Exceptional association rule set discovery from community-dwelling elderly people database, Proc. of 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp.3285-3290 (2018)
DOI:10.1109/SMC.2018.00558
② Kaoru Shimada, Satoshi Noguchi, Michiko Makino, Toru Naito, Exceptional Association

Rule Set Mining from Oral Health Assessment Database, Advances in Intelligent Systems and Computing, Vol.659, pp.429-438 (2018)
DOI:10.1007/978-3-319-67792-7_42
③ Kaoru Shimada, Takaaki Arahira, Takashi Hanioka, Association Rule-based Classifier Using Artificial Missing Values, Lecture Notes in Artificial Intelligence, Vol.10357, pp.57-67 (2017)
DOI:10.1007/978-3-319-62701-4_5

〔学会発表〕（計５件）
① Kaoru Shimada, Hisae Aoki, Keiko Kubota, Satoru Haresaku, Shinsuke Mizutani, Toru Naito, Michio Ueno, Exceptional association rule set discovery from community-dwelling elderly people database, 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2018.10.10
② 嶋田 香, 青木 久恵, 窪田 惠子, 晴佐久 悟, 水谷 慎介, 内藤 徹, 上野 道雄、在宅高齢者の主観的幸福感低値群における影響要因の分析 －AI を用いた影響因子組合せの発見－、第19回医療情報学会看護学術大会、2018.7.7
③ Kaoru Shimada, Satoshi Noguchi, Michiko Makino, Toru Naito, Exceptional Association Rule Set Mining from Oral Health Assessment Database, 5th International Conference on Man-Machine Interactions, 2017.10.5
④ Kaoru Shimada, Takaaki Arahira, Takashi Hanioka, Association Rule-based Classifier Using Artificial Missing Values, 17th Industrial Conference on Data Mining, 2017.7.13
⑤ 嶋田 香、荒平 高章、埴岡 隆、数値属性の離散化に人工的欠損値を導入したルールベースのクラス分類手法、第79回情報処理学会全国大会、2017.3.17

６．研究組織
(1)研究分担者
なし

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。