様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　５　月　２６　日現在

機関番号：　３７１１４
研究種目：　基盤研究(C)
研究期間：　2012～2014
課題番号：　２４５００１９１

研究課題名（和文）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

研究課題名（英文）Research on knowledge discovery from incomplete database based on evolutionary computation with accumulation mechanism

研究代表者

嶋田　香（SHIMADA, Kaoru）

福岡歯科大学・口腔歯学部・准教授

研究者番号：２０４５４１００

交付決定額（研究期間全体）：（直接経費）　　3,900,000円

研究成果の概要（和文）：本研究では、相関ルール抽出を組み込んだ世代継続型進化論的計算手法による不完全データベースの欠損値推定アルゴリズムを提案した。医療系データなどを用いた評価実験により、その有効性を確認した。欠損値を含むデータからの種々のルール発見手法を提案するとともに、欠損値を含むデータにおけるルール指標の特性を評価した。また、ルールベースの連続値予測方法を提案し、人工的欠損値の利用法とあわせて推定性能の向上への有効性を大規模データを用いて確かめた。

研究成果の概要（英文）：In this research, a method of missing value prediction for incomplete database was proposed based on the evolutionary computation with accumulation mechanism of association rule mining. Its validity was confirmed by experiments using medical data sets and so on. Methods for rule discovery from incomplete databases were proposed and characteristics of rule measurements of the extracted rules were evaluated. In addition, a rule-based continuous value prediction method was proposed adopting an application of artificial missing values, and its effectiveness was confirmed by large real data sets.

研究分野：　知能情報学

キーワード：　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

データマイニング手法の一つである相関ルール（Association Rule :IF～THEN～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、これらには欠損値を考慮したルールの指標算出ができないという課題がある。一方、詳細な条件を付けて相関ルール抽出を行う手法として遺伝的アルゴリズム(GA)を用いることも考えられる。GA の適合度関数に最適な少数のルールが得られるが、適合度関数の決定法が課題となる他、推定・分類問題に応用しようとする場合等に、必要とされる数のルールを得ることが困難である。

研究代表者らは、上記の課題を解決する相関ルール発見手法を、有向グラフ構造を特徴とした進化論的計算手法を用いて提案している。とくに、欠損値を含むデータから直接に相関ルールを抽出する方法の提案では、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と考えられるルールが得られる。また、従来の進化論的計算手法が、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が、進化の過程を通して世代継続的な課題解決をしていく成果蓄積方式である。

研究代表者らは、欠損値を含むデータベースにおける特定のクラス属性に注目し、この属性値に基づいてレコードの分類を相関ルール群を発見して行なう方法を提案している。この手法では、ルール抽出に用いる学習データと、分類を行うテストデータの両方に欠損値を含む場合に対応可能である。さらに、クラス属性が不明のデータを分類しながら学習データに取り込んで自己拡大的にクラス分類を行っていく手法を提案している。

２．研究の目的

(1) データベースの欠損値推定アルゴリズムの提案

研究代表者らがこれまでに提案した欠損値を含むデータベースから IF～THEN～タイプのルールを発見する手法を応用して、データベースの欠損値をルールベースで推定する手法の研究を行う。ルール抽出と欠損値推定を反復的に行う独自の進化型計算により、世代継続的にデータベース全体に存在する欠損値の推定を行っていく手法を開発する。欠損値を含むことの多い医療系データの解析に資する基礎技術としてルール発見法とともに確立する。実用化に向けて連続値を含むデータなどへの対応や、離散化方法による推定への影響評価、データの特徴に応じたパラメータ設定法を検討する。

(2) 人工的な欠損値を用いたルール抽出時の情報保護方式の検討

人工的な欠損値を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方式に関する研究を行う。この方式は、不完全データベースからのルール発見法が確立されることで、方式の検討・評価がはじめて可能になる。ランダムに欠損値を発生する方式やデータの特性を反映した情報の隠し方など、人工的欠損値の発生法を検討することで、欠損値を含むデータの取り扱いに関する知見を獲得する。

３．研究の方法

(1) データベースの欠損値推定アルゴリズムの提案

相関ルール抽出を組込んだ世代継続型進化論的計算手法による欠損値推定アルゴリズムを開発する。推定は、各欠損値を個々に定めていくのではなく、推定済みの欠損値を用いた状態でルール抽出を行い、既知のデータへのあてはまりを確認しながらデータベース全体の欠損値を順次埋めていく進化計算型の戦略を応用する。

欠損値を含むデータベースから相関ルールを抽出し、これらのルールを用いた欠損値推定のための進化型計算手法を応用した基本アルゴリズムを確立し、その有効性を検証する。２値データまたはカテゴリデータの離散値からなるデータベースに人工的にランダムな欠損値を発生させ、これらを提案方式により推定することで、推定精度などの評価を行う。推定精度の評価の観点から、すべての欠損値を推定するのではなく、ルールの当てはまり状況、進化時の適合度指標の状況の評価により、一部の欠損値の推定を行わない方式についても検討する。また、大学・研究機関等で公開されているデータについて提案方式を適用し、実用性を検証する。

また、ルール発見アルゴリズムおよび欠損値推定アルゴリズムの高速化・簡易化等を行う。データベースの属性値として、連続値についても扱う。連続値属性の閾値の決定、ファジィの利用の検討を行う。離散化方法による欠損値の推定への影響を評価し、実用化に向けた応用研究を進める。

(2) 人工的な欠損値を用いたルール抽出時の情報保護方式の検討

データベース保有者が人工的な欠損値を用いてデータを保護する場合を想定する。ルール抽出を行う者は、データの全容を知ることはないが、どの程度の正確さでルールを抽出可能か検討・評価する。上述の欠損値推定アルゴリズムと組合せて、情報保護への信頼度を評価する。研究代表者らは、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ルールを差異ルールと定義し、差異ルールの発見手法を提案している。本研究では、この差異ルール抽出法を応用し、欠損値の推定前・推定後のデータベースの差異の評価、人工的欠損値利用時と完全時の差異の評価に用いる。

提案方式を用いて欠損値発生率と推定精度の関連等を評価し、欠損値を含むデータにおけるルール指標の特性を分析する。人工的な欠損値を発生させるときのパラメータ設定法・運用法を検討し、評価を行う。

また、人工的欠損値を発生させる場合に、属性に関連する統計量や相関ルールの知見を反映させた場合の結果を分析することで、データの特性に応じた調整機能をもつ欠損値推定法を検討する。

４．研究成果

(1) データベースの欠損値推定アルゴリズムの提案

欠損値を含むデータベースから相関ルールを抽出してルールベースの欠損値推定を行うための進化型計算手法を応用した基本アルゴリズムを確立し、評価実験用のプログラムを作成してその有効性を検討した。提案手法は、データベースのある属性の欠損値を推定するための相関ルール集合を抽出し、これを用いてその属性における欠損値を推定し、さらにその推定値を含めて別の属性の欠損値推定のためのルール抽出を行い、欠損値を推定、推定値を更新することを繰り返していくことでデータベース全体の欠損値推定を最適化しようとする方式である。

評価実験では、公開されている２値データまたはカテゴリデータの離散値からなるデータベースを用いて、人工的にランダムな欠損値を発生させ、これらを提案方式により推定することで、欠損値発生率と推定精度の関連等の評価を行った。推定対象となる属性の中には、推定開始当初の抽出ルール数が少なく属性値の推定が行えないが、データベースにおける他の属性群の欠損値推定が進行することで、推定が実行可能となる場合がみられるなど、提案手法の有効性が明らかとなった。また、属性ごとの推定精度がデータベース全体の推定の進行に伴って向上することからデータベース全体の欠損値推定を最適化しようとする方式の有効性が示された。

また、欠損値を含むデータにおけるルール指標の特性についての検討、属性による推定の難易の状況の把握方法の検討を併せて行った。欠損値を含むレコードを削除する従来手法では、欠損値発生率の増大により、利用可能なレコード数が減少するため、ルール抽出が困難となる課題があったが、提案方式では、ルール抽出を利用した欠損値の推定が可能であることが明らかとなった。

提案方式のうち欠損値の推定を特定の属性に対して行おうとする場合について、医療系データを用いたアルゴリズムの評価結果と応用例を国際会議で発表した。また、提案方式におけるルール抽出技術を応用した欠損値の推定前・推定後のデータベースの差異の評価、人工的欠損値利用時と完全時の差異の評価をルール抽出状況の観点から行うことのできるルール抽出法について国際会議で発表した。

欠損値推定アルゴリズムの改善と連続値属性へ対応するための拡張を進め、これらの評価実験を行った。また、進化型計算手法で個体の構造変更時のルール抽出状況への影響の評価により、ルール発見効率のよい個体構造に関する知見を得た。さらに、ビッグデータへの対応の観点から、属性数・レコード数の規模を拡大して、パラーメータの最適化や高速化方法を検討し評価実験を行った。

さらに、属性ごとの推定の困難さの違いに対応した改善をするために、推定に用いるルール集合がカバーできる欠損値の割合の増大方法の検討と推定精度への影響の評価した。推定に用いるルール集合がカバーできる欠損値の割合の増大方法の一つとして、人工的欠損値を連続値の離散化時の境界値付近に用いる方法が考えられ、この方法と大規模データを用いた評価実験結果を国内学会および国際会議で発表した。

(2) 人工的な欠損値を用いたルール抽出時の情報保護方式の検討

公開されている２値データまたはカテゴリデータの離散値からなるデータベースを用いて、人工的にランダムな欠損値を発生させ、欠損値発生率とルール指標の関連等の評価を行った。医療系データにおける相関ルールの指標として用いられるカイ自乗値では、5%程度のランダムな欠損値の発生によりルールの興味深さの判別に影響を受けることがわかった。

また、人工的に発生させた欠損値を利用した情報保護等の応用として、人工的欠損値を連続値離散化時の境界値付近に用いることでのルールベースでの欠損値推定法の推定効率改善や精度向上に関する手法を検討し、その有効性を評価した。実験結果から人工的欠損値を利用することでの連続値の推定に用いるルール集合のカバーできる欠損値の割合の向上、汎化能力の獲得に関する知見が得られた。カバーできる欠損値の割合と推定精度はトレードオフの関係にある結果を得ているが、推定条件の制御への応用も考えられた。研究課題の申請時においては、情報保護方式を想定していたが、新しいデータ解析方法開発への基礎研究といえる内容となった。

(3) 本研究課題の成果について、国内外における位置付けとインパクト、今後の展望

本研究では、相関ルール抽出を組み込んだ世代継続型進化論的計算手法による欠損値推定アルゴリズムを提案した。推定は、各欠損値を個々に定めていくのではなく、推定済みの欠損値を用いた状態でルール抽出を行い、既知のデータへの当てはまりを確認しながらデータベース全体の欠損値を順次埋めていくことを特徴としている。提案手法はルールベースであるため推定のメカニズムの可

読化が期待できる。
　また、本研究の成果として、欠損値推定方法の提案に加えて従来困難であった欠損値を含むデータからの種々のルール発見方法や人工的欠損値の利用法の提案等によるルールベースの解析方法の提供、欠損値を含むデータへの理解の深化があげられる。さらに、人工的欠損値を連続値離散化時の境界値付近に用いる手法のように、推定性能の改善や管理といったデータ解析の新技術開発への知見の獲得があげられる。
　今後の展望としては、種々のデータベースにおける利用の観点から、欠損値推定時や人工的欠損値発生時の種々のパラメータの自動最適化方法の研究があげられる。

５．主な発表論文等

〔雑誌論文〕（計４件）

① Kaoru Shimada, Takaaki Arahira, Takashi 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 2015, 392-399, 査読有, 2015.

② Kaoru Shimada, Takashi Hanioka, An Evolutionary Method for Exceptional Association Rule Set Discovery from Incomplete Database, Lecture Notes in Computer Science, Vol.8649, 133-147, 査読有, 2014.

③ Kaoru Shimada, Takashi Hanioka, An Evolutionary Method for Associative Local Distribution Rule Mining, Lecture Notes in Computer Science, Vol.7987, 239-253, 査読有, 2013.

④ Kaoru Shimada, Takashi Hanioka, An Evolutionary Associative Contrast Rule Mining Method for Incomplete Database, Proc. of the 2013 International Conference on Data Mining, 160-166, 査読有, 2013.

〔学会発表〕（計７件）

① Kaoru Shimada, Takaaki Arahira, Takashi Hanioka, An Evolutionary Rule Mining Method for Continuous Value Prediction from Incomplete Database and Its Application Utilizing Artificial Missing Values, 1st IEEE International Conference on Big Data Computing Service and Applications, 2015.4.2, San Francisco (USA)

② 嶋田香，荒平高章，埴岡隆．不完全データに対応したルールベースの連続値予測法と人工的欠損値の利用，第 42 回知能システムシンポジウム，2015.3.18，北野プラザ六甲荘（神戸市）

③ Kaoru Shimada, Takashi Hanioka, An Evolutionary Method for Exceptional Association Rule Set Discovery from Incomplete Database, 5th International Conference on Information Technology in Bio- and Medical Informatics (ITBAM 14), 2014.9.2, Munich (Germany)

④ 嶋田香，埴岡隆．不完全データベースからの相関ルール抽出とルールベースの欠損値推定を世代継続的に行う進化計算手法，2013 年度統計関連学会連合大会，2013.9.10，大阪大学（大阪市）

⑤ Kaoru Shimada, Takashi Hanioka, An Evolutionary Method for Associative Local Distribution Rule Mining, 13 th Industrial Conference on Data Mining (ICDM 2013), 2013.7.18, New York (USA)

⑥ Kaoru Shimada, A Contrast Rule Mining Method for Incomplete Database Based on Evolutionary Rule Accumulation Mechanism, 26International Biometric Conference (IBC 2012), 2012.8.30, 神戸国際会議場（神戸市）

⑦ Kaoru Shimada, An Evolving Associative Classifier for Incomplete Database, 12th Industrial Conference on Data Mining (ICDM2012), 2012.7.19, Berlin (Germany)

〔産業財産権〕
○出願状況（計０件）
○取得状況（計０件）

〔その他〕
なし

６．研究組織
(1)研究代表者
嶋田　香（SHIMADA, Kaoru）
福岡歯科大学・口腔歯学部・准教授
研究者番号：２０４５４１００

(2)研究分担者
なし

(3)連携研究者
なし