様式 C-19

科学研究費補助金研究成果報告書

平成２２年　６月　１日現在

研究種目：基盤研究（C）
研究期間：2007～2009
課題番号：１９５００１３２
研究課題名（和文）客観的信頼性と主観的興味を制御可能な知識発見支援システムの開発

研究課題名（英文）Development of a System Controls Objective Reliability and Subjective Interest for Supporting Knowledge Discovery in Databases

研究代表者
大崎　美穂（OHSAKI MIHO）
同志社大学・理工学部・准教授
研究者番号：３０３１３９２７

研究成果の概要（和文）：
本研究では，客観的信頼性を確保しつつ，主観的興味に合致するルールを得られる知識発見支援システムの開発を目的とした．ルールの客観的信頼性を自動評価する機能，ドメイン専門家の主観的興味を学習してモデル化する機能，客観的信頼性と主観的興味のバランスを調整する機能を開発し，システム本体に組み込んだ．また，現実の医療データを用いたシミュレーションと実験により，各機能の有効性を確認した．

研究成果の概要（英文）：
This research aimed at the development of a system which supports a domain expert to discover objectively reliable and subjectively interesting rules in databases. We developed the modules to automatically estimate the reliability of rules, model the subjective interest of the domain expert, and balance the objective reliability and subjective interest, and then embedded the modules into the main body of the system. We also confirmed the effectiveness of each module through some simulations and experiments using real clinical datasets.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	700,000	210,000	910,000
２００８年度	700,000	210,000	910,000
２００９年度	700,000	210,000	910,000
年度			
年度			
総　計	2,100,000	630,000	2,730,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：データマイニング，知識発見，興味深さ指標，信頼性，興味

１．研究開始当初の背景

データベースからの知識発見(Knowledge Discovery in Databases; KDD)のプロセスは，データに対してノイズ除去・欠損値補間等を行う『前処理』，データマイニング(Data Mining; DM)アルゴリズムを適用し，データからルールを導出する『(狭義の)データマイニング』，ルールを可視化しユーザに提示する『後処理』から成る．

DM や KDD の分野では,『データマイニング』の段階に主眼を置き，ルールの客観的信頼性を基準として DM アルゴリズムを提案，改善する研究が主流であった．しかし，現実の対象問題では信頼性が高いルールは既知である場合が多く，信頼性が高くても人間が理解困難なルールは有益でないという問題が生じた．そこで,『前処理』『後処理』にドメイン専門家が積極的に介入する試み,『データマイニング』の学習基準にドメイン専門家の知識を組み込む試みが活発になった．これらは，ルールの根拠付けを重視する自然科学分野の KDD で一定の成果を上げている．

しかしながら，「ドメイン専門家の主観的興味をどこまで許容するか」という新たな問題が浮上してきた．ドメイン専門家の考えや意図を反映し過ぎると，その人が求めるルールを恣意的に作り出す恐れがあるためである．以上の経緯より，本研究の開始当初において，客観的信頼性と主観的興味の両方を反映できる知識発見の仕組みが望まれていた．

２．研究の目的

上述の背景を踏まえて，本研究では，客観的信頼性を確保しつつ，主観的興味に合致するルールの獲得を可能とする知識発見支援システムの開発を目的とした(図１参照)．具体的には,『後処理』，および,『後処理』で得た情報を『データマイニング』にフィードバックすることに着目した．そして，ドメイン専門家が玉石混淆状態にあるルール群を評価し新知識を導出する作業を助け，客観的信頼性と主観的興味のバランスを考慮した学習基準を DM アルゴリズムに与える仕組みを目指した．

３．研究の方法

我々は目的達成に向けて，以下の３つの開発ステップを計画した．ステップ １：客観的信頼性を計算機上で扱えるように定式化し，本システムに組み込む．ステップ ２：主観的興味を計算機上で扱えるように形式化し，本システムに組み込む．ステップ ３：ドメイン専門家のルール評価過程を明示的で再現性がある形式で記録しながら，客観的信頼性と主観的興味のバランスを調整する機能を実現する．

ステップ１では，ルールに寄与する事例の生起確率や情報量を用いた客観的信頼性の定式化が可能である．過去に我々は，客観的信頼性を定式化したもの(興味深さ指標)を調査し，定義式の統一とソフトウェアライブラリ化を行った．なお，信頼性に限らず一般性や特異性を定式化した興味深さ指標も存在しており，本研究ではこれらの多種多様な興味深さ指標のライブラリを活用する．

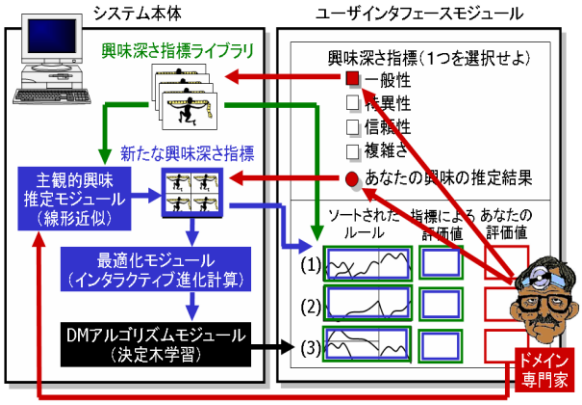

図１　提案する知識発見支援システム

ステップ２には，次の２つのアプローチがあり得る．１つは，一般的な主観的興味が存在するという前提のもと，多数のドメイン専門家を使った心理実験で一般的な主観的興味をモデル化するアプローチである．もう１つは，主観的興味には個人性・曖昧性・時間変動があるという前提のもと，システムの挙動を個々のドメイン専門家の主観的興味に適応させるアプローチである．

本研究では，これまでの研究の知見に基づき後者を採用する．個人性等を許容するため，明示的な言語表現によるドメイン知識の体系化ではなく，人間を評価系として取り込む最適化アルゴリズムであるインタラクティブ進化計算を用いる．過去に我々は，インタラクティブ進化計算を補聴器フィッティングに応用しており，この研究で得たノウハウを本研究において活用する．

ステップ３では，KDD プロセスにおけるドメイン専門家の思考の変化に着目する．過去の我々の研究では，思考を発散させる『仮説生成』と収束させる『仮説検証』の二段階を経て知識発見に至ると分かった．そこで，段階に応じて DM アルゴリズムの学習基準を変化させ,『仮説生成』では，客観的信頼性がやや低くても，ドメイン専門家が「仮説のヒントとなる」と感じ主観的興味を持つルールを優先的に生成する．一方,『仮説検証』では，ルールを絞り込み知識へと洗練化すべく，客観的信頼性が高いルールを優先的に生成する仕組みを考える．

我々が設計した本システムの詳細な機能と使用方法は次の通りである．図１に示すように，ドメイン専門家は主観的興味に応じて

様々な興味深さ指標を選択する．本システムはその興味深さ指標でルールを評価し，評価値順にソート表示する．また，ドメイン専門家は必要に応じて，主観的興味に基づきルールに評価値を与える．この機能１によりルールを多角的に見ることができ，円滑な『仮説生成』が可能になると考えられる．

本システムの内部では，興味深さ指標ライブラリによるルール評価値を説明変数，ドメイン専門家によるルール評価値を目的変数として両者の関係を線形回帰近似し，主観的興味の推定モデル(新たな興味深さ指標)を作る．この機能２により，新たな興味深さ指標を既存の興味深さ指標と同様に用いることができる．これは，ドメイン専門家が暗黙的な主観的興味を明示的に知る手助けになると考えられる．

さらに，本システムでは新たな興味深さ指標をDMアルゴリズムの学習基準に用い，ルールの再学習を行う．『仮説生成』における再学習では，主観的興味を強く反映するために新たな興味深さ指標をそのまま学習基準とする．一方，『仮説検証』では，新たな興味深さ指標を構成する説明変数のうち，信頼性を意味する興味深さ指標の重みを高く設定する．この機能３は，主観的興味と客観的信頼性の重み調整を可能とし，『仮説生成』から『仮説検証』を経てルールの質を段階的に高めると期待される．

４．研究成果

2007年度は主にステップ１を実施した．まず，興味深さ指標に関する文献の再調査を行い，過去に開発済みの興味深さ指標ライブラリに新規の指標を加えた．また，ライブラリに含まれる指標の一部については，その実装が汎用的ではないため，この点の改善も行った．

そして，知識発見支援システムの本体部分のフレームワークを設計し，これに，DMアルゴリズムモジュール(決定木学習)，主観的興味推定モジュール(線形回帰)，最適化モジュール(インタラクティブ進化計算)の各々を乗せられるように実装を進めた．開発に加え，興味深さ指標と主観的興味の関係を調べた実験の成果も得られたため，この成果は学術論文・学会発表で公表した．

2007 年度に公表した研究成果のうち主要なものとして，雑誌論文③の内容を以下に述べる．この研究では，病院で長期間に渡り収集された 2 種類の医療データセットを用い，医学知識発見における興味深さ指標の有効性を実験により検証した．加えて，実験結果に基づき，知識発見のための興味深さ指標の活用方法を検討した．

実験Ｉでは，髄膜炎に関する医療データから得られたルールに対し，40種類の代表的な興味深さ指標による評価結果，および，ドメイン専門家である医師による評価結果を求めた．そして，両者の評価結果の類似度に基づき，興味深さ指標が医師の興味をどの程度推定し得るかを調べた．実験 II では肝炎に関する医療データを用い，実験Ｉと同様の条件設定・手続きにて，興味深さ指標の推定性能を調べた．

実験 I，II の結果には同様の傾向が見られ，Accuracy，Uncovered Negative，Peculiarity等の興味深さ指標が医学ドメインにおける主観的興味の推定に役立つ可能性が示された．また，興味深さ指標の推定性能は，仮説生成，仮説検証の段階に応じて異なること，興味深さ指標の組合せにより，仮説生成から仮説検証に至る医師の思考プロセスを支援できることが示唆された．この研究成果は，本科研費研究におけるステップ１の基盤となった．

2008年度は主にステップ 1，2 における各モジュールの実装を進めた．ステップ１における各モジュールの実装が終了し，特に最適化モジュールに関しては，その成果を学術論文・学会発表で公表した．一方，システム本体のフレームワークへのモジュール組込みが想定以上に複雑であったため，2008年度末ではこの作業は途中段階に留まった．

2008 年度に公表した研究成果のうち主要なものとして，雑誌論文②の内容を以下に述べる．この研究では，最適化モジュールの要素技術であるインタラクティブ進化計算の改善を試みた．インタラクティブ進化計算は，評価関数の定式化が困難な最適化問題に有効であるが，繰り返し評価によるユーザの負担という問題がある．そこで，過去にユーザが評価した解候補とその評価値を用い，半自動的に新たな解候補の評価値を推定する手法を提案・評価した．

提案手法では，過去にユーザが評価した解候補を，一定個数だけデータベースに蓄えておく．未評価である新たな解候補が与えられると，その解候補とデータベース上の評価済み解候補とのユークリッド距離を求める．ユークリッド距離を重みとして評価済み解候補の評価値を加重平均し，これを新たな解候補の推定評価値とする．

人間による解候補の評価では，時間の経過や以前に評価した解候補の印象によって評価基準に変化が生じるため，この動特性を考慮した仕組みが必要である．そこで我々は，繰り返し評価の過程で求まる推定性能に基づき，ユーザが評価する解候補の数を変化させる機能を提案手法に組み込んだ．

インタラクティブ進化計算の応用は様々であるが，特に，高齢者や聴覚障害者のコミュニケーションを支援する補聴器への応用は社会的意義が大きい．そこで，我々が過去に開発したインタラクティブ進化計算に基づく補聴器フィッティングシステムに，今回提案した推定手法を適用し，その有効性を実験検証した．実験では，聴力損失の典型的パターン 2 種類を想定し，11 名の被験者が，各自 10〜20 セットのフィッティング作業を行った．

その結果，提案手法を組み込んでいない条件では,フィッティング 1 セットあたり約 230 回のユーザ評価が必要であった．一方，提案手法を組み込んだ条件では，約 210 回のユーザ評価となり，評価回数を 10%程度減少することができた．よって，一部の解候補を自動的に推定する提案手法が，ユーザの負担軽減に有効であることが確認された．この研究成果は，本科研費研究におけるステップ 2 に大きく貢献した．

2009 年度はステップ 2 の後半の開発を進め，システム本体のフレームワークに DM アルゴリズムモジュール，主観的興味推定モジュールを組み込んだ．加えて，主観的興味推定モジュールに用いる機械学習器の検証を行った．ステップ 3 については，過去の研究で得られたドメイン専門家のルール評価結果を集約し，客観的信頼性と主観的興味のバランスを調整する機能の評価に活用できるように形式化した．

2009 年度に公表した研究成果のうち主要なものとして，雑誌論文①の内容を以下に述べる．この研究では，知識発見支援システムの主要な 2 つの機能，すなわち，機能 1：興味深さ指標のルール評価値に基づくルールのソート提示，機能 2：興味深さ指標のルール評価値に基づくドメイン専門家の主観的興味の推定について，その改善と評価を行った．

2008 年度末までに開発した知識発見支援システムでは，機能 1 が次のように実装されていた．ドメイン専門家であるユーザが指定した 1 つの興味深さ指標についてルール評価値を求め，評価値の降順でルールをユーザに提示する．この動作のみでは，多くの興味深さ指標の評価結果を比較することが困難であった．そこで我々は，主成分分析により評価結果を縮約して提示する動作を機能 1 に追加した．

また，2008 年度末までは，興味深さ指標のルール評価値とユーザのルール評価値の対応関係をモデル化する機械学習器として，線形回帰を採用していた．しかし，ユーザの興味を高い精度で推定することは難しかった．そこで今回は，スタッキング，C4.5，C4.5 を弱学習器としたブースティングとバギング，ニューラルネットワーク，サポートベクターマシン，我々が過去に提案したメタ学習器である CAMLET など，9 種類の機械学習器を組み込んだ．

改善した機能 1，2 の有効性は，以下の分析とシミュレーションにより検証した．機能 1 に関しては，主成分分析を用いて 7 個の主成分を求めた．そして，各主成分に対する各興味深さ指標の寄与を基準に，興味深さ指標のグループ化と意味付けを行った．主成分全体の寄与率は 92%であり，40 種類の興味深さ指標の特徴を十分温存しつつ縮約できたと言える．また，人間が一度に認知可能な事物の最大個数である 7 にまで縮約したことで，機能 1 のユーザビリティが向上したと考えられる．

機能 2 に関しては，雑誌論文③でも使用した髄膜炎に関する医療データ，および，このデータから得られたルールと医師による評価結果を用いた．ただし，医師による評価値，言い換えると，主観的興味は『興味深い』『興味深くない』『理解不能』の 3 つのラベルで表現した．そして，各機械学習器に対してシミュレーションを行い，医師の主観的興味の推定性能を調べた．その結果，最も重要な『興味深い』というラベルについて，サポートベクターマシン 81.6%，スタッキング 81.1%，CAMLET80.3%の精度が得られた．よって，これらは機能 2 に活用できると考えられる．この研究成果は，本科研費研究におけるステップ 2 の改善とステップ 3 の基盤となった．

以上，3 年間に渡る本科研費研究において，知識発見支援システムの開発とその改善はほぼ計画通りに実施された．しかしながら，主観的興味と客観的信頼性の重みを調整する機能 3 については，当初の見通し以上に実装が困難であったため，モジュールの設計と開発に留まった．今後，機能 3 のモジュールを知識発見支援システムに組み込み，システム全体の完成と詳細な実評価に取り組んで行きたい．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計７件）

① Hidenao Abe, Shusaku Tsumoto, Miho Ohsaki, Takahira Yamaguchi, Improving a Rule Evaluation Support Method Based on Objective Indices, International Journal of Advanced Intelligence Paradigms, 査読有, vol.2, 2010, pp.180-197

② 渡辺芳信, 吉川大弘, 古橋武, 大崎美穂, 対話型進化計算における実評価数可変型評価値推論法の提案, 知能と情報(日本知能情報ファジィ学会論文誌), 査読有, vol.20, 2008, pp.810-816

③ Miho Ohsaki, Hidenao Abe, Shusaku Tsumoto, Hideto Yokoi, Takahira Yamaguchi, Evaluation of Rule Interestingness Measures in Medical Knowledge Discovery in Databases, International Journal of Artificial Intelligence in Medicine, 査読有, vol.41, 2007, pp.177-196.

〔学会発表〕（計９件）

① Masakazu Nakase, Miho Ohsaki, Shigeru Katagiri, Yukari Hatada, Clinical Time-series Data Interpolation Based on Regression Diagnosis, Proceedings of Joint International Conference on Soft Computing and Intelligent Systems and International Symposium on Advanced Intelligent Systems SCIS&ISIS-2008, 査読有, 2008, pp.1-5

② Hidenao Abe, Shusaku Tsumoto, Miho Ohsaki, Hideto Yokoi, Takahira Yamaguchi, Evaluation of Learning Costs of Rule Evaluation Models based on Objective Indices to Predict Human Hypothesis Construction Phases, Proceedings of IEEE International Conference on Granular Computing GrC-2007, 査読有, 2007, pp.458-464

③ Hidenao Abe, Shusaku Tsumoto, Miho Ohsaki, Takahira Yamaguchi, Evaluating Learning Algorithms to Construct Rule Evaluation Models Based on Objective Rule Evaluation Indices, Proceedings of IEEE International Conference on Cognitive Informatics ICCI-2007, 査読有, 2007, pp.212-221

６．研究組織

(1)研究代表者

大崎　美穂（OHSAKI　MIHO）
同志社大学・理工学部・准教授
研究者番号：30313927

(2)研究分担者

無し

(3)連携研究者

無し