科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１１　日現在

機関番号：　１０１０３
研究種目：　若手研究(B)
研究期間：　2012～2013
課題番号：　２４７００２０４

研究課題名（和文）多クラスの大規模感性データを対象とした概念マイニングシステムの開発

研究課題名（英文）Concept mining system in multi-class KANSEI data

研究代表者
岡田　吉史（OKADA, Yoshifumi）

室蘭工業大学・工学（系）研究科（研究院）・准教授

研究者番号：００４４３１７７

交付決定額（研究期間全体）：（直接経費）　　3,200,000円、（間接経費）　　960,000円

研究成果の概要（和文）：本研究の目的は、多クラスデータのための新規概念抽出法を開発することであった。本法の特徴は、１）多クラス用バイクラスタリング法に基づいてクラス特異的な概念の探索を行う点、２）オントロジーを用いて類似概念の融合と無意味な概念の排除を行う点、にある。これにより、クラス間の差異を表す意味のある概念の発見が可能となり、大量の概念から有用な概念のみを選別する解析者の負担を大幅に軽減できるようになった。

研究成果の概要（英文）：The aim of this study was to develop a new concept mining method for multi-class dataset. The features of this method are that 1) concepts appearing specifically in each class are mined using the new biclustering method for multi-class data and that 2) only meaningful concepts are extracted on the basis of an ontology database. This study hereby made it possible not only to discover differentially-expressed useful concepts among classes, but also to reduce user's workload in selecting only necessary concepts from large amount of output concepts.

研究分野：　総合領域

科研費の分科・細目：　情報学・感性情報学・ソフトコンピューティング

キーワード：　多クラス　概念　バイクラスタリング

１．研究開始当初の背景

感性工学研究において、観測されたデータから「概念」を抽出する作業は、データに内在する新しい規則や知識を発見する上で重要なプロセスである。概念抽出はこれまで、形式概念分析の分野で盛んに研究されてきた。形式概念分析では、表形式のデータから、共通の属性値を持つようなサンプル（行）の部分集合と属性（列）の部分集合の極大なペアを探索する。このようなサンプル群と属性のペアを「概念」と呼ぶ。

形式概念分析は有望なデータ解析手法の 1 つと考えられるが、感性工学研究で使用する際の問題点として、申請者は以下の２つの点に着目した：（１）感性工学研究ではしばしば、性質の異なる複数のグループ（多クラスのデータ）を比較分析することが必要となるが、従来の形式概念分析は単一クラスのデータのみを対象としている。（２）形式概念分析は組合せ探索問題であるが故に、大量の概念が出力される。その中には、サンプルと属性の大部分が重複した概念や、属性値が偶然一致した無意味な概念が含まれる。

２．研究の目的

本研究では、上記２つの問題を解決するため、以下の２つの課題を解決することを目標とする：（１）申請者らが以前独自に開発したバイクラスタリング法（以下, BiModule）を、多クラスデータを対象とした概念抽出手法へと拡張する。（２）意味的に共通または類似な概念を自動的に融合し、無意味な概念を排除する仕組みを導入する。以上により、個々のクラスで特徴的な概念あるいはクラス間の差異を表す概念の抽出が可能となり、さらに、有用な概念のみを選別する際の解析者の負担を大幅に軽減できるようになると期待される。

３．研究の方法

（１）多クラスデータを対象とした概念抽出法の開発：

提案法は、BiModule を多クラスを対象とした概念抽出手法へと拡張することにより実現される。これは、BiModule が抽出するバイクラスタの属性群（=飽和集合）とサンプル群は、概念における属性群とサンプル群に対応する、という性質に基づいている。以下、本稿では概念をバイクラスタに置き換えて説明する。提案法は、多クラスの表形式データを入力として、バイクラスタ（=概念）を抽出する。バイクラスタを構成する属性群にクラスラベルが含まれるならば、それは特定のクラスで出現するバイクラスタである。逆に、クラスラベルがないものはクラスを跨ぐバイクラスタである。提案法では、クラスラベルを持たないバイクラスタに対する枝刈り法を導入し、クラスを跨がないバイクラスタのみを探索することで高速化を図っている。また、多クラスデータの分析において、異なるクラス間の差異を発見することは重要な観点である。提案法では、増加率（Growth Rate）と呼ばれる指標を用いて、クラス間で差別的な（特異的）な属性値パタンを持つバイクラスタのみを抽出する機能を導入している。

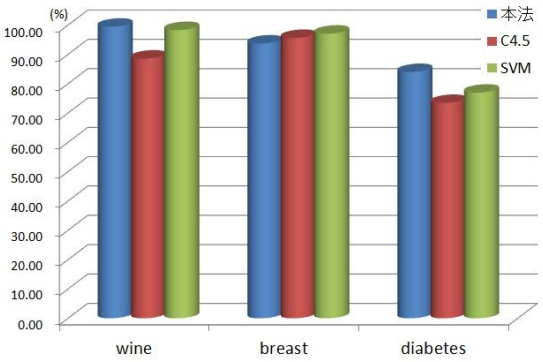

図１．分類精度の結果

（２）概念融合法の開発

概念融合、すなわちバイクラスタの融合はクラス毎に行われる。融合基準は、属性値パタンの重複度と概念的意味の 2 つである。前者はバイクラスタ間で共有される行列要素の割合である。後者は、属性集合（またはサンプル集合）に関するオントロジーに基づくバイクラスタの質的な類似度を意味する。融合処理は、バイクラスタを行列サイズで降順ソートし、上位のバイクラスタとそれより下位のバイクラスタとを比較することで行われる。この時、融合基準の条件を満たすならば上位のバイクラスタに下位のバイクラスタを順次融合していく。

４．研究成果

本研究では、（１）クラス特異的なバイクラスタが適切に抽出されているか（以下、実験１）、（２）融合処理が適切に行われ意味のあるバイクラスタが出力されているか（以下、実験２）、の 2 点について評価を行った。

（１）実験１の結果：

本研究では、機械学習用ベンチマークデータセット（UCI machine learning repository）を用いて、バイクラスタのクラス特異性を評価する実験を行った。具体的には、得られたバイクラスタの属性群を用いて、個々のサンプルを正しいクラスに分類できるか否かをテストした。各クラスから特異性の高いバイクラスタが抽出されているならば、サンプルの分類性能も高くなると期待される。分類手法として、k-近傍法を採用した。

分類精度の評価は交差検定法の１つである LOOCV を用いて行われた。図１は、３つのデ

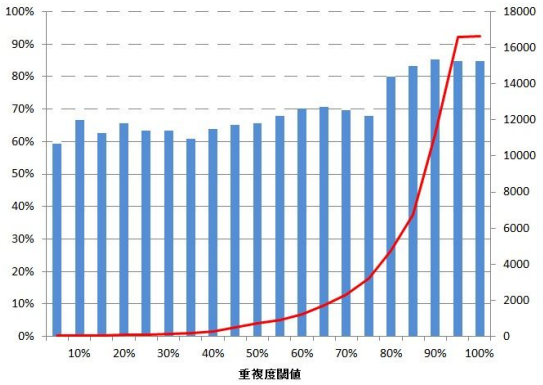

図２．バイクラスタの出力数（赤線）と有意なバイクラスタの割合（青棒）

ータセット（wine, breast, diabetes）を用いて行った分類精度の結果である。図には比較手法として C4.5（決定木に基づく分類法）と SVM（マージン最大化に基づく分類法）の結果も示している。図からわかるとおり、提案法は著名な分類法と比較して、若干ではあるが高い分類精度を示すことがわかった。これは、提案法により得られたバイクラスタ（の属性群）が、個々のクラスを特徴づけるだけでなく、クラス間を差別化しうる特異性を有することを意味している。

（２）実験２の結果：

この実験では、遺伝子発現データセットに提案法を適用し、出力されるバイクラスタ数と生物学的に意味のあるバイクラスタの割合を評価した。本稿では、7129 遺伝子（属性）×49 細胞検体（乳癌細胞 25 サンプル、正常細胞 24 サンプル）の 2 クラスで構成されるデータセットを用いた。バイクラスタの生物学的意味は、L2L と呼ばれる遺伝子機能解析法に基づいて決定される。L2L により、バイクラスタ内の遺伝子名リストをもとに、それと同数の遺伝子群を無作為に集めたときに、特定機能を持つ遺伝子が偶然出現する確率（P 値）を計算することができる。個々の遺伝子の機能は Gene Ontology と呼ばれるオントロジーを参照することで特定される。バイクラスタ内に、有意水準より小さい P 値の機能をもつ遺伝子が含まれるならば、そのバイクラスタはその機能で統計的有意に特徴づけられているとみなす。本実験では $P<0.01$ の機能遺伝子を含むバイクラスタを生物学的に意味のあるバイクラスタ、すなわち有意なバイクラスタとした。

図２は、バイクラスタの出力数（赤線）と有意なバイクラスタの割合（青棒）である。図では融合条件の一つである重複度の閾値を変化させたときの結果が示されている。ここで重複度 100%とは、バイクラスタの要素全てが一致したときに融合すること、すなわち融合処理を行わない場合を意味している。この図から重複度閾値を下げるに従い、バイクラスタの出力数が急激に減少することがわかる。特に、重複度閾値が高い場合においても多くのバイクラスタが融合されている。例えば、重複度閾値 80%のときには、融合しない場合と比べてバイクラスタ数は 3 割弱までに激減している。つまり、融合処理を行わない場合は、互いによく似たバイクラスタがいかに大量に出力されているかがわかる。一方、有意なバイクラスタの割合に関しては、特に高い重複度閾値では急激な減少は認められなかった。これは、単体では意味をなさなかったバイクラスタ同士が融合され、機能が同一の遺伝子群が集約されたバイクラスタが生成されたためと考えられる。逆に、重複度閾値を下げ過ぎると、バイクラスタの出力数は減るが、有意なバイクラスタの割合も減少してしまうことがわかった。

以上より、高い重複度閾値で融合処理を行えば、意味のあるバイクラスタの割合を下げずに出力数を大幅に抑えることが可能となることがわかった。

（３）今後の展望

本研究の特徴は、①多クラスへ拡張した新規バイクラスタリング法によりクラス特異的な概念を抽出する点、②重複度と概念的意味により有用な概念へと絞り込みを行う点、にある。①に関しては、ほぼ目標を達成できたと考えている。一方、②については、遺伝子発現データへの適用に留まっている。これは、体系的かつ詳細な Gene Ontology が整備されており、これを用いた解析手法が豊富に存在するため、提案法の評価には良い題材と考えたからである。本研究で開発した方法と解析プロトコルは、Gene Ontology と同様の形式でオントロジーを構成できれば、他分野へも容易に適用可能と考える。今後は、感性工学をはじめ様々な分野の多クラスデータに適用し、提案法の有用性を検討していく。

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　1 件）

① Takahiro Miura and Yoshifumi Okada, Detection of Linkage Patterns Repeating across Multiple Sequential Data, International journal of computer applications, 査読有, Vol.63, No.3, pp.14-17, 2013.

〔学会発表〕（計　14 件）

① Saerom Lee, Takahiro Miura and Yoshifumi Okada, "A New Method for Improving the Performance of Linkage Pattern Mining", IMECS 2014, 2014/3/12, The Royal Garden Hotel, Tsim Sha Tsui, Hong Kong.
② 芳賀凌太郎, 長久保大輔, 岡田吉史, ク

ラス特異的な飽和集合による多クラス行列データにおける欠損値補完，第 46 回 SICE 北海道支部学術講演会，2014/3/10，北海道大学．

③ 檜山奨，斉藤竜也，岡田吉史，アイテムベースのバイクラスタリングに基づく情報推奨手法，日本感性工学会 生命ソフトウェアシンポジウム 2013，2013/10/26，千葉工業大学．

④ 松本翔太，岡田吉史，階層的クラスタリングを用いた類似バイクラスタの融合法に関する研究，日本感性工学会 生命ソフトウェアシンポジウム 2013，2013/10/26，千葉工業大学．

⑤ Naoya Yokoyama and Yoshifumi Okada, Item Recommendation by Query-based Biclustering Method, 15th Int. Conf. KSE2013, 2013/10/17, Hanoi National University of Education, Hanoi, Vietnam.

⑥ Tatsuya Saito and Yoshifumi Okada, Bicluster-Network Method and Its Application to Movie Recommendation, 15th Int. Conf. KSE2013, 2013/10/17, Hanoi National University of Education, Hanoi, Vietnam.

⑦ Tatsuya Saito and Yoshifumi Okada, Recommendation Method Using Bicluster Network Method, IMECS 2013, 2013/3/13, The Royal Garden Hotel, Tsim Sha Tsui, Hong Kong.

⑧ Daisuke Nagakubo, Yasuo Kudo and Yoshifumi Okada, Recommendation Method based on Rough set for Sequential Data, IMECS 2013, 2013/3/13, The Royal Garden Hotel, Tsim Sha Tsui, Hong Kong.

⑨ 横山直也，斉藤竜也，川原光平，岡田吉史，クエリに基づくバイクラスタリング法を用いた情報推奨法，日本感性工学会 生命ソフトウェアシンポジウム 2012，2012/11/23，室蘭工業大学．

⑩ 斉藤竜也，川原光平，岡田吉史，バイクラスタネットワーク法を用いた情報推薦法の提案，日本感性工学会 生命ソフトウェアシンポジウム 2012，2012/11/23，室蘭工業大学．

⑪ 松田芳矩，大橋克哉，岡田吉史，Gene Ontology を用いた遺伝子モジュールの機能評価，日本感性工学会 生命ソフトウェアシンポジウム 2012，2012/11/23，室蘭工業大学．

⑫ 矢代耕平，大橋克哉，岡田吉史，クエリに基づくバイクラスタリング法による遺伝子発現データの欠損値補完，日本感性工学会 生命ソフトウェアシンポジウム 2012，2012/11/23，室蘭工業大学．

⑬ 大橋克哉,岡田吉史，遺伝子発現データからの機能遺伝子モジュールの抽出，第 14 回日本感性工学会大会，2012/8/31，東京電機大学 東京千住キャンパス．

⑭ 川原光平，岡田吉史，ユーザの着眼点に応じた結果の絞り込みを可能にする情報推奨システム，第 14 回日本感性工学会大会，2012/8/31，東京電機大学 東京千住キャンパス．

〔図書〕（計 1 件）

① Tomomasa Nagashima, Yoshifumi Okada and Yuzuko Nagashima, Springer, Biometrics and Kansei Engineering (Chapter11: Concepts of KANSEI and Aesthetic Thoughts), 2012, pp.191-210.

６．研究組織

(1)研究代表者

岡田　吉史（OKADA, Yoshifumi）
室蘭工業大学・工学研究科・准教授
研究者番号：００４４３１７７