論 文 の 内 容 の 要 旨

学位論文題目「大量・複雑なデータの解析と可視化に関する研究」

学位申請者 山田 実俊

キーワード：探索的データ解析　データの可視化　アソシエーションルール分析
対応分析　インタラクティブな可視化

データ解析の結果は通常数値で与えられるが、解析結果を理解するためには多くの値を同時に評価・解決する必要が生じることが多い。また、単に結果の値のみで判断するのは、データに対して誤った認識をする可能性もある(Anscombe, 1973)。ビッグデータと呼ばれる大量・複雑なデータの解析が行われ始めた近年において、データの可視化はデータの分析や解析の結果を表す優れた方法である。よって、大量・複雑なデータの解析結果を解釈しやすく表現する可視化を考える必要がある。また可視化は、結果を解釈しやすくするため、何らかの条件付けや写影が必要で、条件の変更が可能なインタラクティブ(対話的)な可視化が必要とされる場合も多い。

本論文では、多くのデータを価値のあるデータとして解釈しやすくしたいと考え、大量・複雑なデータの解析と可視化に関する研究を行い、多変量解析とデータマイニングのハイブリッド型の分析・可視化を提案する。データ解析においては、データの整形により従来のデータ解析の適用分野を拡げることも提案する。可視化においては、条件の変更によって変わる解析結果に対応させるためのインタラクティブな機能の実装の有用性についてシステムを提案することで示す。

論文の構成は以下の通りである。

第 1 章では統計的データ解析における古典的な分析手法である多変量データ解析を紹介し、その種の分析における探索的データ解析や、データの可視化の重要性や現状について紹介する。統計的データ解析の新しい手法としてのデータマイニングについても紹介する。2 章で提案する手法に用いた多変量データ解析の対応分析とデータマイニング手法のアソシエーションルール分析と可視化についても紹介し、本論文の構成も記述している。

第 2 章では用いるアンケートデータを紹介し、メディア層や回答項目の関係について探索的データ解析によりデータの理解を深め、年齢・性別の回答傾向を俯瞰する可視化として提案した、「属性特化型特徴抽出アソシエーションプロット」の紹介を行う。インタラクティブな可視化ができるアプリケーションにより、インタラクティブな機能が提案する可視化には重要であることを示す。年齢・性別に対する回答傾向の類似性や特異性について、全体の傾向を把握する方法として一般的に適用される対応分析を用いて、アンケートの結

果を可視化した。各回答項目がメディア層による年齢・性別の軸に対して関係のある位置に付置された。また、データの整形による通常とは異なる分析にアソシエーションルール分析を適用し、「メディア層⇒回答項目」という条件部がメディア層であるアソシエーションルールの抽出をし、ある年齢や性別に特徴的な回答項目を見つけた。そして、アソシエーションルールの可視化も行った。

「属性特化型特徴抽出アソシエーションプロット」は対応分析とアソシエーションルール分析のハイブリット型の可視化である。この可視化は対応分析とアソシエーションルール分析を相互補完し、全体を俯瞰し、特定の属性の関係も把握しやすくする可視化であるとして提案した。しかし、アソシエーションルール分析のパラメータは特徴付けしやすいルール数を表示するために試行錯誤する必要があるため、インタラクティブな機能を必要とした。そこで、RStudio の Shiny パッケージを用いて、パラメータを随時変更できるインタラクティブな可視化アプリケーションを作成した。インタラクティブな機能を含めて、アンケート結果とメディア層の関係を俯瞰する可視化の実装ができた。

第 3 章では属性特化型特徴抽出アソシエーションプロットで扱った対応分析の代わりに、可視化のための数量化手法である多次元尺度構成法と数量化Ⅲ類を用いた可視化を行い、それぞれの特徴を考察し、対応分析が今回の提案に適していることを示す。多次元尺度構成法では特定のメディア層と関係の強い回答項目がそのメディア層の付近に、回答率の多い回答項目が外側に付置された。数量化Ⅲ類では軸ごとに特定の項目とそれ以外を分けるように付置された。アソシエーションルール分析と合わせてデータの俯瞰がしやすい可視化としては対応分析が適していることを示した。

第 4 章では薬剤耐性菌と、薬剤耐性誘導の影響の分析の現状を紹介する。山本 他(2015)よりアソシエーションルール分析が有効ではないかと考え、医療データのデータ成形を行い、第 2 章で提案した属性特化型特徴抽出アソシエーションプロットに適用した。また、アソシエーションルール分析のリフトと医学界でよく使われる統計手法のオッズ比解析の関係を利用して、医学界で解釈しやすいルールの強さを表現した。さらに、ある期間での結果が見たい、アソシエーションルールを表でも見たい、図を拡大して見たい、ある抗菌薬から引かれるルールだけが見たい、重なりすぎて見づらいなど現場からの要望を考慮して、属性特化型特徴抽出アソシエーションプロットをより見やすくする、インタラクティブな可視化機能を持つアプリケーションを提案した。

第 5 章ではウツタインデータの救命率に注目し、搬送時間による地域差について可視化提案し、考察を紹介する。日本地図を用いた可視化によって、救命率は東低西高であり、搬送時間は地形が東西・南北に伸びている都道府県が長いなど、地域差の解釈しやすいシステムを作成した。しかし、搬送時間が救命率に影響を与えている因子ではなかった。

第 6 章ではスポーツデータの解析の 1 つとして、サッカーのアクションデータからボールの保持率、ホットゾーンの支配率について可視化を行い、ゲーム分析に役立つアプリケーションを提案し、考察を紹介する。ボールを保持していたチームのボールの保持時間と

し、ホットゾーン呼ばれるフィールドを 4×6 に分けた 24 区分について試合全体のボールの保持率(支配率)を求めた。その区分が使用された割合と、どちらのチームが多く使用していたかを色の濃度で表現した可視化を行った。また、試合を 6 つの時間帯に分けて可視化することで大まかな試合の流れを表現した。さらに、ある攻撃の起点から 10 分間の支配率を求め、どのタイミングで試合の流れが変わったかを推測できるような可視化も行った。

「属性特化型特徴抽出アソシエーションプロット」はクロス集計結果の可視化として、幅広い適用分野があり、各分野での活用が期待できる.現象の理解が困難であった薬剤耐性の出現と使用抗生剤の関係については東海大学病院における 10 余年間のデータに対してデータ成形の工夫により扱うことができた。ウツタインデータやサッカーのデータの探索的データ解析のためのインタラクティブなシステムを提案も、ビッグデータ時代の新たなデータ分析の可能性を示すものとして期待できる。

属性を持つデータであれば、属性特化型特徴抽出アソシエーションプロットを用いることで、データの全体の俯瞰と属性による特徴を把握することができると考える。様々なデータに適応させて, 基本の可視化手法の 1 つとして扱えるようにし、そのデータに合った解析・インタラクティブ的な可視化を研究していきたい。