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 ２５ 日現在

機関番号：６２６０３
研究種目：基盤研究(C)（一般）
研究期間：2014～2018
課題番号：２６３３００５４

研究課題名（和文）集約的シンボリックデータ解析の基礎構築

研究課題名（英文）Establishing the foundation of aggregated symbolic data analysis

研究代表者
中野　純司（Nakano, Junji）

統計数理研究所・モデリング研究系・教授

研究者番号：６０１３６２８１

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：個体データが大量にある場合、それらの意味のある集合に着目する場合が多い。例えば、数千羽の鳥の体長や体重、色、種などのデータが得られているとき、個体の集合としての種の特徴を知りたい。その場合、グループの特徴を少数の記述統計量で表したものをシンボリックデータと呼ぶ。本研究では個々のデータが連続変数とカテゴリー変数を持つ場合を考える。カテゴリー変数はダミー変数を用いて表し、連続変数とともに2次までのモーメント統計量を構成し、それを集約的シンボリックデータと呼ぶ。そして集約的シンボリックデータ間の非類似度や、可視化を提案した。

研究成果の学術的意義や社会的意義
現在、データ量が爆発的に増加しており、その解析のために新しい手法が必要となっている。個々のデータではなくデータのグループを対象とするシンボリックデータ解析は、超大量データを縮約し、人間が現実的に扱い理解することを可能にする手法である。最近は実用的で強力だが人間が解釈・理解することの難しい手法も機械学習や人工知能の分野で多く開発されている。ただ、社会的にも学術的にも、現象の本質を人間が理解することは重要である。コンピュータ、通信、センサーなどの技術の発展により、観測されるデータの量は莫大になっている。そのようなデータを人間に近づけるために集約的シンボリックデータ解析は有用である。

研究成果の概要（英文）：When we have a large amount of individual data, we are often interested in meaningful groups of individuals. For example, when we have data such as body length, weight, color and species of several thousand birds, we hope to know the characteristics of species as groups of individual birds. Characteristics of the group may be represented by small number of descriptive statistics. We call them symbolic data. We consider the case where each individual data has continuous variables and categorical variables. Categorical variables are expressed by dummy variables. Such dummy variables and continuous variables can be summarized by up to second order moment statistics, which are called aggregated symbolic data. We proposed dissimilarities and the visualization method among aggregated symbolic data.

研究分野：計算機統計学

キーワード：カテゴリー変数　シンボリックデータ　分類　連続変数

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

シンボリックデータは“コンセプト”を表現するために提案された。“コンセプト”は集合的な概念であり、例として鳥を取ると、鳥の個体ではなく、雀、ツバメ、鳩、などのようなグループを考える。これを統計解析のためのデータにしようとすると、雀の多数の個体のデータから雀というコンセプトを表現するための新しいデータ形式を考えなくてはならない。シンボリックデータ解析では、そのための変数の値として、区間、ヒストグラム、集合、離散確率分布のようなものを考える。この手法はヨーロッパの研究者を中心として 1980 年代から開発されており、その成果はいくつかの書籍にまとめられている。

ただ、これまでの研究ではグループを表現するために周辺分布の情報だけを考えることが多かったが、その妥当性はあまり考慮されなかった。さらにほとんどの場合、シンボリックデータがすでに所与であり個々の個体データには立ち返らない、という立場で研究が進められている。これは超大量の個体データにも少なくとも一度は現実的な時間でアクセス可能となっている現在の計算機技術を考えると、不十分である。

本研究ではグループを多変量分布の実現と考え、それを表現するためにグループに属する個体データから計算される適切な記述統計量を用いることを提案し、集約的シンボリックデータと呼ぶ。そして集約的シンボリックデータの 1) 情報損失の少ない表現、2) これまでの研究との関係、3) 種々の数理統計的手法、を研究・開発することにした。

２．研究の目的

ビッグデータを意味のあるグループに分割することにより、人間が理解・解釈・解析しやすくするためのひとつの手法を確立することが目的である。ビッグデータにおいては個体数が多くなるだけではなく、変数も多くなり、かつ、その構造も複雑になる。特に、カテゴリー変数と連続変数が同時に観測されることが普通である。これまでの統計手法では連続変数とカテゴリー変数を同時に同等に扱うことは容易ではない。そこで、それが可能なる手法を集約的シンボリックデータとして開発する。そしてその妥当性を理論的にも示すことが目標である。また、開発した手法を実際のビッグデータに対して適用しなければならない。そのためにはデータの集約、可視化などのための計算機ソフトウェアの開発も必要である。

３．研究の方法

個体データがグループを示すカテゴリー変数を含む場合に、それを用いてグループを構成する。すると各グループは同じ変数を有するので、それを同じように縮約することによりグループ間の比較ができる。カテゴリー変数はダミー変数で表現し、連続変数はそのまま利用すると、行列計算により 2 次までのモーメント統計量を構成することができ、それを集約的シンボリックデータとする。グループ内の個体の個数は 0 次モーメント統計量、連続変数の標本平均とカテゴリー変数の標本周辺分布は 1 次モーメント統計量、連続変数の標本分散共分散とカテゴリー変数のペアごとの分割表は 2 次モーメント統計量である。さらに 2 次モーメント統計量の中には連続変数とカテゴリー変数のペアの情報も含まれる。それらに対して可視化の方法を考え、また、適当な距離を定義することによって、これまでの統計手法が利用できるようにする。

４．研究成果

グループの特徴を 2 次までのモーメント統計量を用いて表すことで、その特徴がある程度記述できることがわかった。当然の事ながら、3 次、4 次とより高次のモーメント統計量を用いればさらに正確なグループの特徴を記述できる。しかしそうすると用いる記述統計量の数が増大する上、人間が理解しにくくなる。解析した実データでは、変数変換を行い、対称分布に近づけることにより、その 2 次までのモーメント統計量でかなりの情報を縮約できることがわかった。また、グループ内の変数の関係を考えると、周辺分布だけを考えていたこれまでのシンボリックデータ解析より詳細な情報が得られる事もわかった。

そのような集約的シンボリックデータの可視化を行う場合にはさらなる縮約が必要である。特にカテゴリー変数に対しても、連続変数の平均と分散に相当する量を定義する必要がある。それは全部の個体に関する周辺分布を基準として、そこからの位置と集中の程度を定義することによって得られる。また、2 つのカテゴリー変数間、連続変数とカテゴリー変数の間の相関関係を示す統計量も提案し、それが連続変数の標本相関係数に相当することを示した。これらを用いると、拡張した平行座標プロットで可視化することが可能になる。そしてそのためのソフトウェアを Java 言語によって開発し、さらにそれを統計解析ソフトウェア R から利用できるようにした。また、一台のコンピュータで扱うことが難しいような超大量データも扱えるように分散処理環境 Hadoop との連携も試みた。

グループ間の距離としてはカテゴリー変数に関しては分割表の集合である Burt 行列間の距離（正確には非類似度）を考える事になる。そのためにはカイ 2 乗統計量が利用できる。ただ連続変数も同じように扱うためには連続変数をカテゴリー化することを考える。それによって連続変数とカテゴリー変数の 2 次までのモーメント統計量の間の距離を定義できた。さらに二つの集約的シンボリックデータが同じ性質を持つという仮説を、そうでないという仮説に対して検定するための疑似尤度比検定統計量を非類似度と考えることも提案した。

５．主な発表論文等

〔雑誌論文〕（計　1　件）

① 清水信夫, 中野純司, 山本由和 (2018) 集約的シンボリックデータのカイ 2 乗統計量を用いた非類似度とその不動産情報データへの適用, 統計数理, 66,2, 279-294　（査読有り）

〔学会発表〕（計　13　件）

① 清水信夫，中野純司，山本由和（2018）集約的シンボリックデータの変数選択，2018 年度統計関連学会連合大会

② Yamamoto, Y., Nakano, J. and Shimizu, N. (2017) Interactive visualization of aggregated symbolic data, NZSA-IASC 2017

③ Shimizu, N., Nakano, J. and Yamamoto, Y. (2017) Dissimilarities between groups of data, NZSA-IASC2017

④ 清水信夫，中野純司，山本由和（2017）カテゴリー変数を含む集約的シンボリックデータのカイ 2 乗統計量，2017 年度統計関連学会連合大会

⑤ Shimizu, N., Nakano, J. and Yamamoto, Y. (2017) Dissimilarity by chi-squared statistic for aggregated symbolic data with continuous and categorical variables, The 2017 conference of the International Federation of Classification Societies

⑥ Yamamoto, Y., Nakano, J. and Shimizu, N. (2017) Interactive visualization of characteristics of groups, The 2017 conference of the International Federation of Classification Societies

⑦ 清水信夫，中野純司，山本由和（2016）カテゴリー変数を含む集約的シンボリックデータの非類似度の性質，2016 年度統計関連学会連合大会

⑧ Nakano, J., Shimizu, N. and Yamamoto, Y. (2015) Visualizing aggregated symbolic data with continuous and categorical variables, 60th World Statistics Congress ISI2015

⑨ Nakano, J., Shimizu, N. and Yamamoto, Y. (2015) Dissimilarity between aggregated symbolic data with categorical variables, 2015 International Workshop for JSCS 30th anniversary in Okinawa

⑩ 中野純司，山本由和，清水信夫（2015）連続変数とカテゴリー変数を含む集約的シンボリックデータの特徴抽出，2015 年度統計関連学会連合大会

⑪ Nakano, J., Yamamoto, Y., Shimizu, N. and Fujiwara, T. (2014) Aggregated symbolic data description including real and categorical variables, COMPSTAT 2014

⑫ 中野純司，清水信夫，山本由和，藤原丈史（2014）カテゴリー変数を含む集約的シンボリックデータの記述と可視化，2014 年度統計関連学会連合大会

⑬ Nakano, J., Yamamoto, Y. and Shimizu, N. (2014) Visualization of aggregated symbolic data with real and categorical variables, 2014 Workshop in Symbolic Data Analysis

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）
○取得状況（計　0　件）

〔その他〕
ホームページ等
　なし

６．研究組織

(1)研究分担者
　なし

(2)研究協力者

研究協力者氏名：　清水信夫
ローマ字氏名：　(SHIMIZU, Nobuo)

研究協力者氏名：　山本由和
ローマ字氏名：　(YAMAMOTO, Yoshikazu)