研究レポート　No.7　～データフュージョン～

2020 年 10 月 21 日　株式会社アイズファクトリー　https://bodais.com/company/

概要

データ形式や収集条件に違いのある複数のデータを分析のできる一つのデータに統合する技術をデータフュージョンという。ビッグデータを活用するためにも、データフュージョンによるデータの統合が利用されている。本稿では、データフュージョンの対象となるデータと手法、応用例について概説する。

1.　はじめに

アンケート調査の統計分析を行う際に、調査実施期間により調査対象者が異なる、質問の項目や形式に違いがある等、全てのデータを同条件で分析することが困難な場合がある。そのようなデータに対して、データを個体ベースで結び付けて融合させることをデータフュージョンという。データフュージョンにより、データの母数や情報量を増やし、より有用な分析を可能とする。近年では、インターネット技術やセンシング技術の発達により、ビッグデータと呼ばれる大規模データ[1]が分析に利用される。これらのデータは形式や収集条件が異なることが多く、統合した分析を行うためにデータフュージョンの技術が用いられる。例えば、商品の購買歴とインターネットの広告閲覧履歴を結び付けて分析する際には、両データ間での直接的な個人対応は通常困難で、データフュージョンにより個々のデータを対応させる必要がある。また、センサーデータは互いにデータの形式や計測時刻の間隔が異なることが多いため、複数のセンサーデータを統合させた分析を行うためにもデータフュージョンが利用される。

統計分析に使われるデータには、一つの情報源から得られるシングルソースデータ、異なる情報源から得られるマルチソースデータがある。商品購買に対する広告効果解析例におけるシングルソースデータとマルチソースデータの違いの概念図[2]を図 1 に示す。シングルソースデータでは、広告接触に関する項目 A と購買実績に関する項目 B の変数が同じ対象者から得られるため、そのままの形で項目 A と項目 B を統計分析やモデルのインプットに使うことができる。一方、マルチソースデータは項目 A の対象者と項目 B の対象者が別々に分かれている。このような場合では、項目 A と項目 B の関係性を分析することができない。このようなマルチソースデータの変数を使った分析をする場合、データフュージョンによりシングルソースデータへと統合する必要がある。

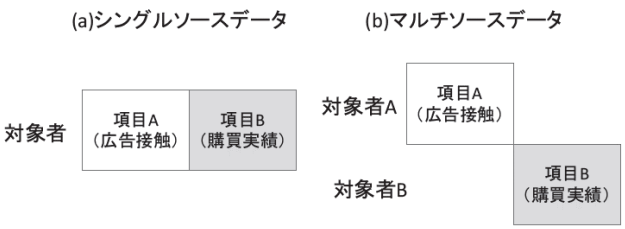

図 1. シングルソースデータとマルチソースデータの概念図
(出所: 文献[2])

2.　データフュージョン

データフュージョンが必要になる分析のケースには、上記の様にマルチソースデータの別々に分かれた変数を分析する場合の他に、データの計測時刻に差異がある場合等がある。これらのデータにおいて、データ間で対応しない項目は欠測と見なせる。データフュージョンの目的はこれらのデータの欠測を補完して統合したひとつのデータとして分析することで、分析の精度の向上や新たな知見を得ることにある。

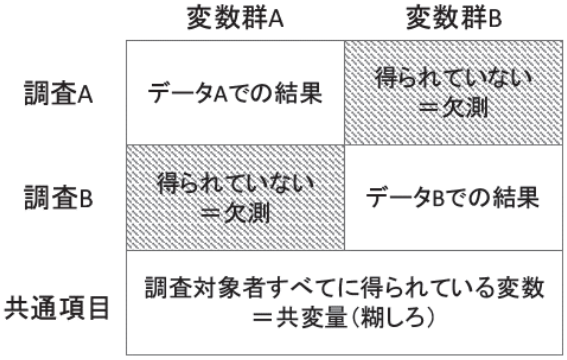

図 2. データフュージョン概念図 (出所: 文献[2])

図 2 にデータフュージョンの概念図を示す[2]。調査 A と調査 B から得られるデータをそれぞれデータ A、データ B とする。調査 A では変数群 A、調査 B では変数群 B が得られ、その他に調査対象者すべてに共通項目の変数が割り当てられている。この 2 データに共通する変数を共変量と呼び、両データの欠測を補完するための糊しろとして利用する。データフュージョンは共変量を基に変数値を推定する。推定方法によって、①回帰モデルのようにモデルを仮定する方法（パラメトリック推定）、②マッチングのようにモデルを仮定しない手法（ノンパラメトリック推定）、③これらの中間的な手法（セミパラメトリック推定）に分類される[3]。代表的なものに、マハラノビス距離を使ったマハラノビスマッチング法、潜在変数を推定式から得る重回帰モデル（共分散分析モデル）[4]がある。文献[2]では、これらの方法は予測精度が低いことやモデル推定の柔軟性がないことを指摘しており、より高い精度が期待されるセミパラメトリック法を提唱している。セミパラメトリック法には、傾向スコアやベイズモデルを使った回帰分析等がある。

以下の章では、ノンパラメトリック手法の代表例であるマハラノビス法と、セミパラメトリック法について文献上での使用例とともに説明する。

3.　マハラノビスマッチング法

マハラノビスマッチング法は融合させるデータ間で共変量のマハラノビス距離を算出し、距離が小さくなるデータ同士を結合させる。データ A の 1 要素の共変量ベクトルを x_A、データ B の 1 要素の共変量ベクトルを x_B、分散共分散逆行列を Σ^{-1} とすると、マハラノビス距離 d は以下のように表わせられる。

$$d = \sqrt{(x_A - x_B)^T \Sigma^{-1} (x_A - x_B)} \quad (1)$$

マハラノビスマッチング法では距離計算に使う共変量しかデータの結び付けに使われず、他の変数は結び付けに考慮されず、また、結合する他の変数に推定量等は使われず、元のデータに含まれる値のみ補完される[5]。文献[5]ではマハラノビス法とベイズモデル

を使ったベイジアン回帰補完法によるデータフュージョンを行い、融合した後の変数の相関を比較することにより、ベイジアン回帰補完法の方がよりよい相関が得られることを示している。また、文献[6]ではマハラノビス距離算出時の共変量が多いと次元の呪いの影響を受けること、元データの共変量に偏りがあると分析の予測精度が悪くなることを指摘し、カーネル正準相関分析とカーネルマッチング法を組み合わせたデータフュージョンを提案している。このようにマッチングによるデータフュージョンでは、元のデータの変数の分布や使える共変量によってはその後の分析の精度を低下させる恐れがある。

4. セミパラメトリック法

セミパラメトリック法として、傾向スコアを使った回帰モデルによるデータフュージョンが挙げられる。傾向スコアの具体的な手法は、文献[7]に記載されている。大まかな手順としては、データの共変量を使ったロジスティック回帰等の回帰分析により、共変量を 1 次元化した傾向スコアを得る。共通でない変数に対して、傾向スコアの逆数の重み付け平均値で補正を行うことデータを補完して融合する。文献[8]では、傾向スコアによるデータフュージョンにおいて回帰分析時の共変量の選択が重要で、探索的な共変量の選択により有用な補正値を得られることを報告している。しかし、傾向スコアによるデータフュージョンを適用する条件として、十分な共変量の情報が得られている必要があることが指摘されている[9]。共変量の情報が少ないデータに傾向スコアを使うことは難しい。文献[9]では国民性調査のデータに対して、潜在共変量をセミパラメトッリクのディリクレ過程混合モデルで設定し、パラメータを変化させた感度分析により調査不能の標本データの補完をした。セミパラメトリックによる補完データは、同一回答を想定した補完の場合よりも妥当な 95%信頼区間幅を得ることを示している。

5. データフュージョンの応用例

データフュージョンを用いた分析は、インターネットやクラウドデータ、センサー技術の発達により利用される機会が増えている。インターネットを介したログやセンサー計測により収集されるビックデータは、特定の分析に合わせて収集されていないことから、データの形式や収集時刻の間隔が異なることが多い。そのようなデータを分析するには、データフュージョンによりマルチソースからシングルソースのデータに統合させる必要がある。このようなデータフュージョンの応用例を以下に挙げる。

文献[10]では、自動運転制御のために複数のセンサーデータを統合して解析する技術が示されている。使用している車載のセンサーはカメラや GPS 等の 10 種類であり、収集されるデータはフォーマットや収集時刻、座標系等に違いがある。それらの違いを階層構造による統合する方法を開発し、統合に要する時間と労力の短縮をしている。文献[11]は、スマートフォンやカーナビ等の時刻と位置の情報を共変量として様々なデータと統合させて、分析に利用するリアルタイムデータフュージョンによるプラットフォームを開発している。統合したデータの活用例として、子供が一定距離親から離れた場合アラートを知らせるスマートフォンのみまもりアプリや、テーマパークの混雑状況把握システムを挙げている。文献[12]では、交通系 IC カードの乗降車の記録データと移動目的等も記録されている交通計画調査のデータを統合することで IC カードのデータに移動目的等の情報を付与し、交通行動のパターンを解析に利用できるデータフュージョン手法を開発している。これらの例のようにビックデータを利用した分析においてデータフュージョンが用いられることが多い。

6. まとめ

データフュージョンの技術を使うことで、調査の条件や対象が異なることでそのままでは分析に扱えないデータの補完をし、データの有効利用や分析方法の幅を広げることができる。データフュージョンの手法は、補完するデータの共変量や欠測値、母集団の分布によって選択する必要があり、マッチング法は直接的にデータの補完ができるが補完するデータによっては分析の精度が悪くなり、モデルベースによる推定ではデータに合うモデルや共変量を吟味する必要がある。

データフュージョンの応用活用範囲は、ビックデータの利用とともに広がっていくことが期待される。インターネットを介したログデータやセンサーの計測データは、一般に情報量やデータのカバーしている範囲は広いが、データの形式の違いや収集時刻の違いから、正確な分析結果を得るためにはデータフュージョンが必要となる。今後も、ビックデータの活用とともにデータフュージョンの技術は発展していくであろう。

7. 参考文献

[1]「研究レポート No. 6 ～ビックデータの解析手法～」, 株式会社アイズファクトリー (2014)

[2] 星野ら 「集計マクロレベル情報とミクロレベルデータを融合した広告効果推定法の開発と応用」, 吉田秀雄記念事業財団助成研究報告書 (2017)

[3] 星野崇宏 「調査観察データの統計科学: 因果推論・選択バイアス・データ融合」, 岩波書店 (2009)

[4] Deaton et al. Statistical models for zero expenditures in household budgets, Journal of Public Economics, 23, 59-80 (1984)

[5] 栗原由紀子 「統計的マッチングにおける推定精度とキー変数選択の効果 ― 法人企業統計調査ミクロデータを対象として ―」, 統計学 108, 1-15 (2015)

[6] 光廣ら 「正準変量のカーネルマッチングによるデータ融合法」, 人工知能学会全国大会論文集 2P304 (2018)

[7] 「研究レポート No. 1 ～傾向スコアによる調査データ補正～」, 株式会社アイズファクトリー (2011)

[8] 星野ら 「傾向スコアを用いた補正法の有意抽出による標本調査への応用と共変量の選択法の提案」, 統計数理 54(1), 191-206 (2006)

[9] 星野ら 「調査不能がある場合の標本調査における セミパラメトリック推定と感度分析: 日本人の国民性調査データへの適用」, 統計数理 58(1), 3-23 (2010)

[10] 福元ら「自動運転向けマルチセンサフュージョン処理プラットフォームの開発」, 組込みシステムシンポジウム 2018 論文集, 67-70 (2018)

[11] 北橋ら「データフュージョン技術を用いたデータ活用ソリューション」, 株式会社インテック, 技術情報誌, 15, (2015)

[12] 日下部ら「データフュージョンによる行動データマイニングのための基礎分析」, 土木計画学研究・講演集, 45, 286 (2012)