多項ロジットモデル及び主成分分析を用いた統計的マッチング手法の提案

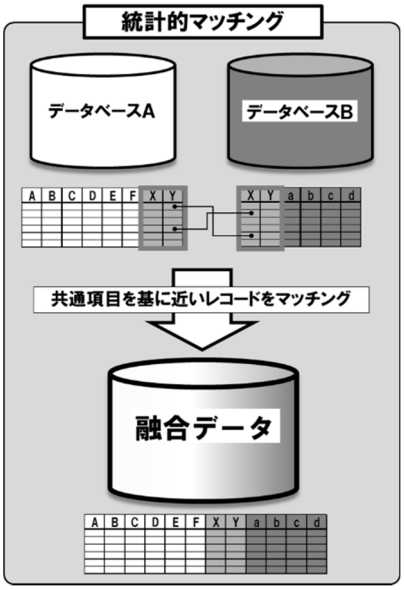
総合研究大学院大学　高部　　勲
統計数理研究所教授　山下　智志

１．統計的マッチングの概要と課題

データリンケージは、異なるデータベースを結合し、豊富な情報を持つ単一のデータベースを構築する技術である。データリンケージにより、新たな調査やデータ収集を行うことなく、有用なデータの作成が可能となることから、近年、様々な分野で利用されるようになってきている。

データリンケージを行う際に、各レコードを識別できる照合キー（共通一連番号、名称、所在地など）が存在する場合は完全照合（Exact Matching）が可能だが、そのようなキーが存在しない、あるいは利用できない場合には、レコード間の類似度を表す距離関数を定義し、近いレコード同士を結合する統計的マッチング（Statistical Matching）の手法が用いられる。

マッチングにはウエイト付き距離が用いられることが多いが、ウエイトの決定方法が恣意的であるとの指摘がある。また企業データは一般にレコード間の差が大きいため距離の定義が難しく、さらにデータ量も多く計算量も多くなるため、統計的マッチングの適用例はあまり多くない。



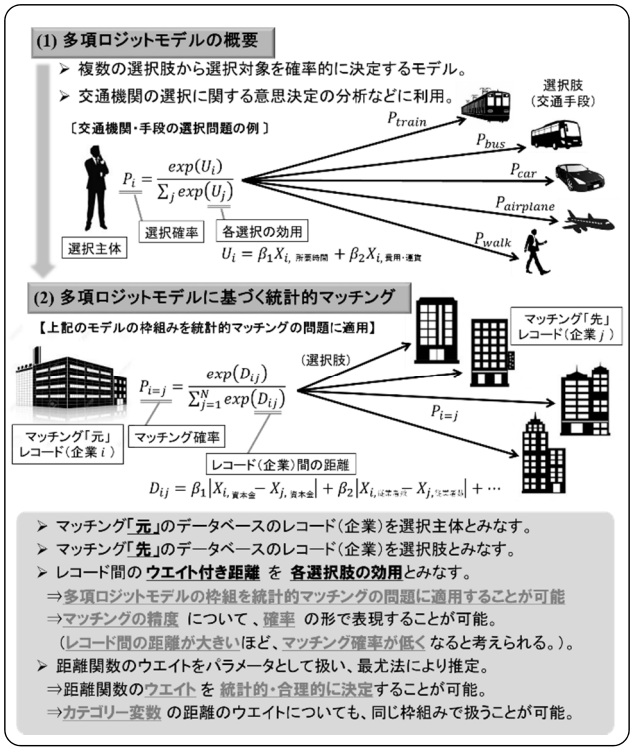
【ウエイト付き距離関数の例】　$d_{ij} = \sum_{p=1}^{P} \beta_p \left| X_{ip} - X_{jp} \right|$

２．提案手法と結果

本報告では、多項ロジットモデル及び主成分分析を用いた統計的マッチング手法を提案する。２種類の企業データの共通フィールドから算出したレコード間の距離を説明変数として多項ロジットモデルを構築することにより、ウエイトの合理的な決定及びマッチング精度の確率的な評価が可能となる。

ここで各データのレコード数が増加した場合、距離計算の対象となるレコードの組合せが飛躍的に増大し、計算が困難となることから、それらをうまく削減する必要がある。本報告では主成分分析によりデータを層化し、近隣のレコードのみを距離・尤度計算の対象とすることで計算の効率化を図っている。

提案手法を商用データ及び経済センサスのミクロデータに適用した結果、誤分類率などの点で、従前のデータマッチングの手法（Nearest Neighbor Methodなど）よりも優れていることが示された。



参考文献：

[1] D'Orazio, M., M. Di Zio & M. Scanu (2006), *Statistical Matching: Theory and Practice*, Wiley

[2] Rässler, S. (2002), *Statistical Matching*, Springer

本研究は科研費（16H02013及び15H03390）の助成を受けている。