MT法を用いたレコードリンケージのための特徴量選択

園田　桂子　総合研究大学院大学　統計科学専攻　博士課程（5年一貫性）5年

➢ 研究の目的

同一対象（人物や企業等）に関する記録（レコード）が異なるデータベースに含まれていた場合に、正しいレコードのペア（マッチング）を特定することをレコードリンケージと呼ぶ。これを行うことによって、同じ対象に行われた公的統計調査や民間企業調査によるミクロデータを結合することが可能となり、新たな統計調査やデータ収集を行うことなく情報量を増やすことができる。結合するデータベースそれぞれにレコードを識別できる照合キー（名称など）がある場合は、これを利用して照合することが可能であるが、調査主体が異なる場合は秘匿性の観点から使用できない場合が多い。そこで、データベースに共通して含まれる特徴量（変数）を基に、何らかの意味で類似したレコードを結合する手法（統計的マッチング）が用いられる。本研究は、共通変数として比較的多くの財務データを利用することができる企業のレコードリンケージを取り扱い、正しいレコードペアを特定するための効率的な特徴量選択の方法として、マハラノビス＝タグチ法（MT法）による特徴量選択の可能性を研究する。

➢ 先行研究と研究の意義

レコードリンケージに用いる特徴量選択を取り扱った先行研究には、全ての組み合わせについてマッチング実験を行って精度を比較した栗原（2013）があるが、研究例は多くない。これは、異なるデータベースでは共通する特徴量が少ないと想定されていたり、観測時点が正確には同じではないことから、時間的に一定、もしくは変動が少ないと想定される特徴量（資本金等）を使用するのがよいと考えられていることが理由にある。しかし、企業に関しては共通する特徴量として豊富な財務データが存在する場合が多く、また、レコードリンケージにおいて、時間的な変動が小さいと思われる特徴量を使用することの優位性は統計的に明らかではない。更に、最近は機械学習の手法を取り入れるなど、事前に統計モデルを特定することなく大量のデータからモデルを学習させる計算コストが高い統計的マッチング手法を用いることが増えており、計算にかかる時間や計算機の処理能力といった観点から事前にレコードマッチングに有効な特徴量を特定しておくことの実務的な意義は大きい。

➢ MT法を用いた特徴量選択

統計的マッチングを行う際、「何らかの意味での類似性」を測る尺度として、絶対値距離（Manhattan距離）、Eucrid距離、Mahalanobis（マハラノビス）距離など、レコード間の距離を用いることが多い。このうちマハラノビス距離は、多変数間でばらつきや相関がある場合を想定して確率分布局面上の勾配を考慮に入れたものと解釈することができる。品質管理、医療等の分野等では、マハラノビス＝タグチ（MT法）と呼ばれる一連の多変量解析手法の中で、マハラノビスの距離を用いて異常検知を行った事例が多くある。この際、直交表を用いて網羅性と効率性を両立させた特徴量の取捨選択パターンを定めて、これらについて、S/N比と呼ばれる異常度を測る上での有用性を示す評価尺度を計算して多くの特徴量の中から有用な特徴量の組み合わせを選択する。この手法をレコードマッチングに用いることで、誤ったマッチング間では距離が大きくなるような特徴量の組み合わせを、事前に簡便な方法で選択する事ができると期待される。

絶対値距離（ Manhattan距離）

$$D_{ij} = \sum_{k=1}^{p} \beta_k \left| X_{ik} - X_{jk} \right|$$

Eucrid距離

$$D_{ij} = \sqrt{\sum_{k=1}^{p} \beta_k \left(X_{ik} - X_{jk} \right)^2}$$

Mahalanobis 距離

$$D_{ij} = \left(X_i - X_j \right)^T \sum_{XX}^{-1} \left(X_i - X_j \right)$$

i, j：各データベースのレコード（iとj は異なるデータベースに属する）, k:両データベースの各共通変数, p:共通変数の数, D_{ij}:レコードiとレコードjの距離, β_k: k番目の共通変数のウェイト,

X_{ik}：レコードiに含まれるk番目の共通変数, $X_i = \left(X_{i1}, X_{i2}, \ldots, X_{ip} \right)^T$, $\sum_{XX}$　共通変数の分散共分散行列

➢ S/N比による特徴量の選択

正常データ（正しいマッチング間の距離）が圧倒的多数だと信じられるデータセット$D = \{x^{(1)}, \ldots, x^{(N)}\}$と、異常（誤ったマッチング間の距離）と判明しているデータセット$D' = \{x'^{(1)}, \ldots, x'^{(N')}\}$を用意する。

正常データセットDから計算した標本平均と標本共分散行列を利用して、異常データセットD'に含まれるデータの異常度をマハラノビスの距離で測る。S/N比は下記の式による（井手, 2015）。

$$SN_q \equiv -10log_{10} \left\{ \frac{1}{N'} \sum_{n=1}^{N'} \frac{1}{a_q\left(x'^{(n)}\right) \Big/ M_q} \right\}$$

q: 特徴量の組み合わせ

N':異常データの数

$a_q\left(x'^{(n)}\right)$: 特徴量の組み合わせqの際、異常データに含まれるある観測値x'の異常度（マハラノビスの距離）

M_q:特徴量の組み合わせqの際の変数の数

➢ 直交表とS/N比による有効な特徴量の組み合わせの選択

16の特徴量、2水準（1：使用する、0：使用しない）の直交表（イメージ）

特徴量 / 実験No.	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	SN比
1	0	0	0	1	0	0	1	1	1	1	0	0	1	0	1	1	η_1
2	0	0	1	1	0	1	0	0	1	1	1	0	0	0	0	1	η_2
3	0	1	1	1	0	1	0	1	1	1	1	1	0	0	0	0	η_3
4	1	1	1	0	1	0	1	1	1	0	0	0	1	1	1	0	η_4
5	0	0	1	0	0	1	1	1	0	0	1	1	0	1	0	0	η_5
6	1	1	0	0	1	1	1	1	1	1	1	0	0	0	1	1	η_6
7	1	1	0	0	1	0	1	1	0	0	1	0	0	1	0	1	η_7
8	0	0	1	0	0	0	0	0	1	0	1	0	1	0	0	1	η_8
9	1	0	1	1	0	0	1	1	0	1	0	1	0	1	1	0	η_9
10	0	1	0	1	1	0	1	1	1	0	0	1	0	1	0	0	η_{10}
11	0	0	1	1	0	1	1	0	0	1	0	1	1	1	1	1	η_{11}
12	1	0	0	1	0	1	1	1	0	1	1	0	0	0	1	1	η_{12}
⋮																	

SN比の利得＝$\eta_{B,1} - \eta_{B,0}$

$$\eta_{B,0} = \frac{\eta_1 + \eta_2 + \eta_5 + \eta_8 + \eta_9 + \eta_{11} + \eta_{12}}{7}$$

$$\eta_{B,1} = \frac{\eta_3 + \eta_4 + \eta_6 + \eta_7 + \eta_{10}}{5}$$

➢ 実証方法

2つのデータベースを使用。片方は日経NEEDs社FinancialQUESTの財務（短信・有報）データベース（有料のミクロデータベース、以下FQデータと呼ぶ）。収録されているのは上場企業が主だが一部非上場企業を含む。業種は限定されない。もう片方は、経済産業省企業活動基本調査のミクロデータ（以下企活データと呼ぶ）。収録されているのは、従業者数50人以上かつ資本金又は出資金額3，000万円以上の企業（大企業）で、日本標準産業分類における農業、漁業、建設、運輸・郵便、医療福祉、複合サービスを除く業種。独立行政法人統計センターに公的統計のミクロデータ利用申請を行って許可を受け、オンサイト施設で調査票情報を利用。データ時点は平成28年で揃えた。欠損値が存在せず、かつできるだけ多くの変数を検討対象とするために、両データベースに共通する変数のうち、欠損値の割合が相対的に低い16変数について検証する。

FQデータと企活データに共通して含まれる名称と所在地情報を照合キーとして、目視によって事前に正しいマッチングペアを特定する。FQデータについては、企活データと完全照合できたデータのみを使用することとし、FQデータは必ず企活データに含まれることとなっている。完全照合済みのデータから、名称及び所在地の情報を削除して分析用データとする。

MT法による特徴量選定のためのS/N比算出のために、正しいマッチングペアから算出した特徴量ごとの距離を正常データ、誤ったマッチングペアから算出した特徴量ごとの距離を異常データとする。異常データの算出するにあたっては、企活データのうち、FQデータと正しくマッチングしないレコードから、FQデータのレコード数と同数だけを復元無作為抽出して算出する。復元無作為抽出は複数回行い、S/N比の算出も複数回行う。

統計的マッチング手法として、高部・山下（2018）による、特徴量のウェイトを最尤推定で行い、FQデータのレコードと企活データのレコードの全組み合わせについて多項ロジットモデルでマッチング確率を計算するという、多項ロジットモデルを用いた統計的マッチング手法を用いる。同手法を用いるためには事前に特徴量を選択する必要はないが、計算量が膨大となるため、実務上、有効な特徴量を特定しておくことの意義が大きい手法である。また、マッチング確率上位20位の中に、正しいマッチング先が含まれる確率が他の手法よりも高いことが実証されている。

（結果の記述は公的統計のミクロデータ利用申請に伴う審査が必要なため許可が出るまで割愛する）