【研究ノート】

データ分割と最適輸送

田　中　研太郎

概要

この研究ノートでは，データを2つの群に分割する場合に，それぞれの属性に基づいて，2群がなるべく等質になるような分割をするにはどのようにすればいいのか，という問題について考える。等質の程度を測る指標として，最適輸送の考え方を使うことで，適切なものが構築できることを紹介する。

1. データ分割

例えば，10人の学生を学年と学部がなるべく等質に（似たようなものに）なるように2群に分割したいとする。よく行われる方法としては，右の表のように乱数を使って無作為に分ける，というものが考えられる。無作為に分割すれば，期待値的には確かに等質な2群に分割できるが，個々の分割においては，それほど適切でないものが生成されてしまうこともよくある。例えば，右の表1も乱数を使って無作為にグループAとグループBに分割しているが，実際には，学年や学部にかなり偏りが生じてしまっていることがわかる。

表１：乱数によるグループ分けの例

グループ	学年	学部	乱数
グループA	2	経営	0.02518
グループA	2	経済	0.05290
グループA	1	経営	0.22254
グループA	1	経営	0.30164
グループA	2	経営	0.35387
グループB	1	経済	0.35910
グループB	2	経済	0.36596
グループB	3	経済	0.59443
グループB	3	経営	0.64805
グループB	1	経済	0.91153

分割を1回だけ行ってみてうまくいかなければ，無作為に分割する操作を「うまくいく」まで何回も繰り返す，という手段も考えられる。ここで問題になるのは，はたして，どういう状態であれば「うまくいった」といえるのか，という基準が曖昧である，ということである。個々の問題に応じて何らかの基準を定めることもできるが，ここでは，なるべく多くの問題で使える一般的な基準を構築することを考えたい。そのための準備として，次の節では最適輸送について紹介する。

2. 最適輸送

一方からもう一方へ何かを輸送するとき，コストに関して最適な輸送方法を考えるのが，最適輸送の問題である。例えば，図1のように，20個の点からなるグループAの各点に1単位のものがあるとして，それらを20個の点からなるグループBに分配・輸送することを考える。このとき，輸送方法にはいろいろなパターンが考えられるが，輸送コストを最小にする輸送方法は，図2で与えられる。このような問題を一般的に考えると，次のように定式化できる。

まず，グループAにはm個の点があり，各点の設定量をベクトル$\boldsymbol{a} = (a_1, \cdots, a_m)$で表すことにする。また，グループBには$n$個の点があり，各点の設定量をベクトル$\boldsymbol{b} = (b_1, \cdots, b_n)$で表すことにする。$\boldsymbol{a}$と$\boldsymbol{b}$の要素はすべて正で，それぞれの合計は1になるように正規化されているとする。そして，グループAのi番目からグループBのj番目の場所に1単位のものを輸送するときのコストをM_{ij}で表すことにする。このとき，最小の輸送コストを考える問題は，以下の線形計画問題として定式化される。

$$\min_{\gamma \in R_+^{m \times n}} \sum_{i,j} \gamma_{ij} M_{ij}$$

$$s.t. \quad \boldsymbol{\gamma 1} = \boldsymbol{a}, \quad \boldsymbol{\gamma}^T \mathbf{1} = \boldsymbol{b}, \quad \boldsymbol{\gamma} \geq \mathbf{0}$$

ここで，$\mathbf{1}$は要素がすべて1のベクトルを表す。この線形計画問題を解いたときの目的関数の値が最小の輸送コストになるが，この値によってグループAとグループBの違いの大きさを表すことができる。このように計算される量は，確率分布間の距離を表すWasserstein距離というものの1つの形態（1-Wasserstein距離）になっている。

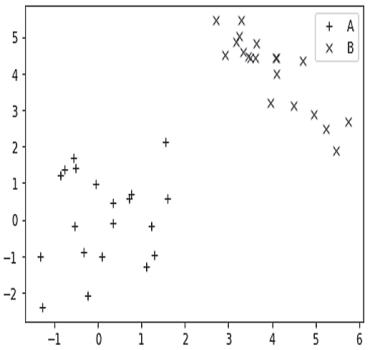

図1：グループA とグループB は各20 個の点からなるとする。

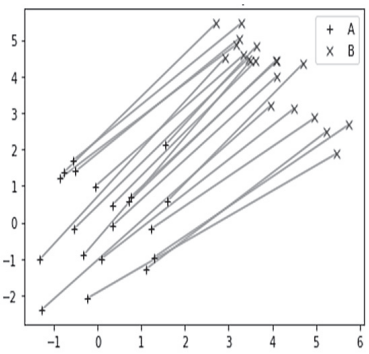

図2：グループA とグループB の間の最適輸送時の対応関係。

3. データ分割と最適輸送

この節では，いくつかの例で，データを2分割したときの最適輸送時のコストの値（1-Wasserstein距離）の計算結果を紹介していく。コンピューターでの計算においては，Pythonのライブラリ（参考文献[1] POT: Python Optimal Transport ）を利用した。

まず，2次元の量的データをAとBの2つのグループに分割したときに，その混ざり具合

の違いでコストがどのように変わるかを見てみると，図3のa)からc)のようになった。なお，コストとしては，ユークリッド距離を用いた。図3のa)は，AとBがあまり混ざっておらず，コストは0.3137になっている。図3のb)は，a)よりはAとBが混ざっており，コストは0.1324に減少している。図3のc)は，かなりAとBが混ざっており，コストは0.0188で，a)とb)のコストと比べて相対的にかなり小さな値になっている。これらの結果から，量的データを2分割したときに，それらがよく混ざっていればいるほどコストは小さな値になっており，グループ間の等質の程度を測る指標として最適輸送時のコストが役に立つことがわかる。

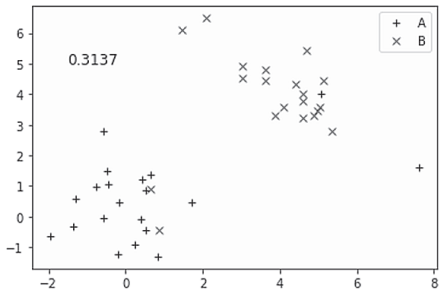

図3　a)：コスト 0.3137

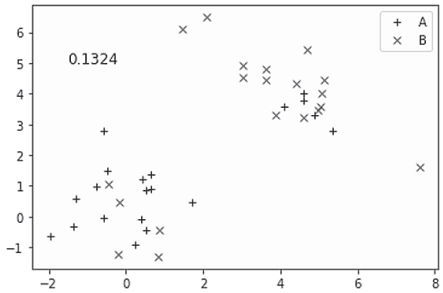

図3　b)：コスト 0.1324

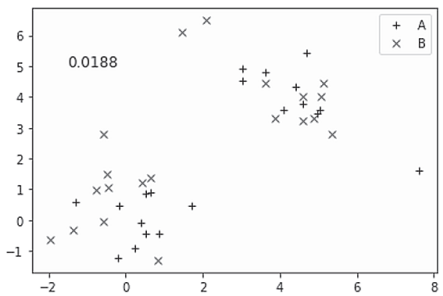

図3　c)：コスト 0.0188

次に，1節の表1と同様のデータでAとBの2つのグループに分割したときに，偏りの違いでコストがどのように変わるかを見てみると，表2のa)からc)のようになった。なお，コストとしては，質的変数をダミー変数で表したうえでユークリッド距離を用いた。表2のa)は，AとBで学年の偏りがかなりあるため，あまり混ざっておらず，コストは0.4667になっている。表2のb)は，a)よりは学年の偏りは少なく，また，学部の偏りも少ないので，コストは0.2000に減少している。表2のc)は，学年と学部の両方でかなりバランスの取れた分割になっており，結果としてコストは0.0667になった。a)とb)のコストと比べてc)の場合のコストは相対的にかなり小さな値になっている。これらの結果から，質的なデータを2分割したときにも，グループ間の等質の程度を測る指標として，最適輸送時のコストが役に立つことがわかる。

表2 a)：コスト 0.4667

グループ	学年	学部
A	1	経営
A	1	経営
A	1	経済
A	1	経済
A	2	経営
B	2	経営
B	2	経済
B	2	経済
B	3	経営
B	3	経済

表2 b)：コスト 0.2000

グループ	学年	学部
A	1	経営
A	1	経済
A	2	経営
A	2	経済
A	2	経済
B	1	経営
B	1	経済
B	2	経営
B	3	経営
B	3	経済

表2 c)：コスト 0.0667

グループ	学年	学部
A	1	経営
A	1	経済
A	2	経営
A	2	経済
A	3	経営
B	1	経営
B	1	経済
B	2	経営
B	2	経済
B	3	経済

4. まとめ

この研究ノートでは，データを2つのグループに分割する場合に，それぞれの属性に基づいて，グループ間がなるべく等質になるような分割をするにはどのようにすればいいのか，という問題について考えた。等質の程度を測る指標として，最適輸送時のコストが使用可能であることを数値例を用いて示した。

今後の課題として，以下の2つの課題が挙げられる。まず1つ目は，3つ以上のグループに分ける場合の適切な指標の構築である。最適輸送は，一方からもう一方への輸送を考えるものであり，基本的には，ペアとなる2つのデータに対する指標である。3つ以上のグループに分ける場合の指標の構築にあたっては，新たな概念の導入や工夫が必要になると考えられる。また，もう1つは，グループ分けのための効率的なアルゴリズムの構築である。この研究ノートにおいては，グループ分けをするアルゴリズムについては一切触れておらず，単に，指標の構築方法を説明しただけである。今回の数値例で用いたような小さなデータにおいては，無作為抽出を繰り返して最適輸送時のコストがなるべく小さくなるものを探す，という手法で十分であるが，実際には，もっと大きなデータの分割を行うことが多く，その場合には無作為抽出での探索は相当に効率が悪いと考えられる。今後は，グループ分けのための効率的なアルゴリズムについても研究を進めていきたい。

（成蹊大学経営学部教授）

参考文献

［1］POT: Python Optimal Transport (https://pythonot.github.io/)