単位空間の汚染にロバストなマハラノビス・タグチ法

早稲田大学　大久保豪人
早稲田大学　永　田　靖

1. はじめに

タグチメソッドの代表的な手法であるマハラノビス・タグチ（MT）法（Taguchi and Jugulum (2002)）は我が国の製造業を中心に広く普及している異常検知のための多変量解析法である．本研究では，学習データに異常のラベルをもつ個体が混入したデータ（汚染データ）に MT 法を適用する場合の解析方法について考察する．現行 MT 法の解析プロセスは学習データのラベルが全て正常である仮定のもとで成立している．そのため，汚染データを対象とする場合，適切な分析ができるとは限らない．そこで本研究では，汚染データを対象とした MT 法の新たな解析プロセスを提案する．

2. MT 法の概要

MT 法では，均質な母集団を形成する群のことを「単位空間」と呼ぶ．MT 法では，この単位空間からの離れ具合をマハラノビス距離によって定量化する．そして，マハラノビス距離の大きさが事前に定めた閾値を超えるか否かによって，判定対象となる個体が単位空間に属するか判定する．

いま p 次元変数 $\boldsymbol{x}$ が母集団で観測され，母集団からの大きさ n の無作為標本を $\boldsymbol{x}_i$ (i = 1,2, ..., n)とする．また，$\boldsymbol{x}$ の母平均ベクトルおよび母共分散行列を $\boldsymbol{\mu}$ および $\boldsymbol{\Sigma}$ とする（以降，母数 $\boldsymbol{\theta}$ の推定量を $\hat{\boldsymbol{\theta}}$ と記す）．このとき，標本マハラノビス距離 $D(\boldsymbol{x}_i)$ (i = 1,2, ..., n)を次のように定義する．

$$D^2(\boldsymbol{x}_i)=(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})^T \tag{1}$$

現行 MT 法では母集団のラベルがすべて正常であること，すなわち，母集団と単位空間が一致することを前提として解析を行う．また，$\boldsymbol{\mu}$ および $\boldsymbol{\Sigma}$ の推定量には単位空間に属する n 個の個体から求めた標本平均ベクトルおよび標本共分散行列を用いる．

3. 汚染データへの適用方法

いま母集団の真の分布を $g(\boldsymbol{x})$とするとき，$g(\boldsymbol{x})$を近似する統計的モデル $f(\boldsymbol{x})$が与えられたとしよう．特に $f(\boldsymbol{x})$がパラメトリックなモデルである場合，$f_{\boldsymbol{\theta}}(\boldsymbol{x})$と記す．このとき，現行 MT 法は $g(\boldsymbol{x})$と $f_{\boldsymbol{\theta}}(\boldsymbol{x})$の Kullback-Leibler（KL）ダイバージェンスが最小となるように $\boldsymbol{\theta}$ を推定すると解釈できる．

ここで，単位空間に異常のラベルをもつ個体（ミスラベル・データ）が混入した状況を考える．この場合，KL ダイバージェンスはミスラベル・データが混入した単位空間に対する統計的モデル $f_{\boldsymbol{\theta}}(\boldsymbol{x})$のダイバージェンスとなる．すなわち，KL ダイバージェンスを最小化するようにパラメータを推定した場合，その混入状況に依存したバイアスを推定量はもってしまう．そこで本研究では，ミスラベル・データが混入した場合でも，真の単位空間に対する統計的モデル $f_{\boldsymbol{\theta}}(\boldsymbol{x})$のダイバージェンスが最小化されるようにパラメータを推定する方法論を MT 法に導入することを提案する．

具体的には，Fujisawa and Eguchi (2008)が提案した γ ダイバージェンスに基づくロバスト推定法を用いる．このロバスト推定法では，次式の γ ダイバージェンスを最小化するように $\boldsymbol{\theta}$ を推定する．

$$D_\gamma(g\,||\,f)=-d_\gamma(g\,||\,g)+d_\gamma(g\,||\,f) \tag{2}$$

ここで，右辺の第 1 項および第 2 項は各々γ 相互エントロピーと呼ばれる．すなわち，分布 $g(\boldsymbol{x})$に対する分布 $f(\boldsymbol{x})$の γ 相互エントロピーは正定数 γ を用いて次式で定義される．

$$d_\gamma(g\,||\,f)=-\frac{1}{\gamma}ln\int g(\boldsymbol{x})f(\boldsymbol{x})^\gamma d\boldsymbol{x}+\frac{1}{1+\gamma}ln\int f(\boldsymbol{x})^{1+\gamma}d\boldsymbol{x} \tag{3}$$

なお，当日の発表では数値実験を通して，汚染データを対象とする場合，提案プロセスが異常検知性能の向上に有用であることを確認する．特に大量のミスラベル・データが学習データに混入する場合や，ミスラベル・データの分布が未知の場合，従来のロバスト推定法を MT 法に導入した解析プロセスに比べて提案プロセスの導入効果が高いことも示す．

参考文献

Fujisawa, H., and Eguchi, S. (2008): Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, **99**(9), 2053-2081.

Taguchi, G. and Jugulum, R. (2002): *The Mahalanobis-Taguchi Strategy: A Pattern Technology System*. John Wiley and Sons.