MT 法におけるしきい値設定法の提案と比較

M2011MM001 安部将成
指導教員：松田眞一

1　はじめに

現在，企業では品質第一・品質向上・品質基準など品質について重要視され品質管理がよく知られるようになった。さらに，1950 年代から田口玄一博士によって提案されてきた品質工学が使用されるようになってきている。その品質工学の中にマハラノビス・タグチシステム（以下 MT 法と呼ぶ）という方法があり，判別・予測・パターン認識といった場面で活用される。

2　研究の目的

品質工学の MT 法のしきい値設定は統計的な面からあまり研究されておらず，初めはしきい値を『4』という数値を決めただけのものであった。今では兼高 [1] が提案した χ^2 分布を用いた方法や中津川・大内 [2] が提案したガンマ分布を用いた方法が考えられている。しかし，この 2 つの分布はどちらが優れているかは研究されていない。また，マハラノビス距離の 2 乗は χ^2 分布に従う事がわかっており，χ^2 分布とその分布に関連性のあるガンマ分布を用いているが χ^2 分布に関連性のある F 分布を用いた方法は研究されていない。

そこで，本研究では F 分布を使用したしきい値設定法を考え，しきい値『4』と今までに提案されている方法を比較し現時点でベストなしきい値設定法はなにか研究をする。

3　MT 法の概要

MT 法は，検査事象の異常判定を行うために正例事象群と負例事象群が必要とされる。正例事象群からデータの平均を求め，正例事象群のデータから相関係数行列を算出しそれらで計算された距離から異常判定を行う。以下にその概要を記す。(田口 [4]，立林ら [5]，中津川・大内 [2] 参照)

3.1　MT 法の距離の算出方法

MT 法は異常判定を行うためにマハラノビス距離を用いて検査をする。マハラノビス距離は，正例事象群より定まる基準点及び単位量に基づく多変量データの評価尺度である。正例事象群は項目ごとの平均値により形成されるベクトルを基準点とし，正例事象群を構成する事象のマハラノビス距離の平均値を 1 とするように定義される (田口 [4] 参照)。

正例事象群は表 1 のように n 事象 k 項目の多変量データとし，負例事象群は表 2 のように m 事象 k 項目の多変量データとする。正例事象群のデータを用いマハラノビス距離の 2 乗 d_x^2，d_y^2 は次の過程で求める。

正例事象群データ $x_{ij}(i = 1, 2, \cdots, n; j = 1, 2, \cdots, k)$ に基づき，各変数の平均 $\bar{x}_j$ 及び標準偏差 s_j を求める。

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \tag{1}$$

$$s_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} \tag{2}$$

表 1　正例事象群データ

	項目			
事象番号	x_1	x_2	$\cdots$	x_k
1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

表 2　負例事象群データ

	項目			
事象番号	y_1	y_2	$\cdots$	y_k
1	y_{11}	y_{12}	$\cdots$	y_{1k}
2	y_{21}	y_{22}	$\cdots$	y_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
m	y_{m1}	y_{m2}	$\cdots$	y_{mk}

平均 $\bar{x}_j$ 及び標準偏差 s_j を用いて，x_{ij} と $y_{hj}(h = 1, 2, \cdots, m)$ の基準化を行う。

$$u_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{3}$$

$$v_{hj} = \frac{y_{hj} - \bar{x}_j}{s_j} \tag{4}$$

基準化されたデータ u_{ij} を用い，正例事象群の相関行列 R を求める。

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{pmatrix} \tag{5}$$

相関行列 R と基準化された u_{ij} と v_{hj} を用い，$X_i = [u_{i1}u_{i2}\cdots u_{ik}]$，$Y_h = [v_{h1}v_{h2}\cdots v_{hk}]$ としてマハラノビス距離 d_x と d_y を求める。算出方法は以下のようになる。

$$d_{xi}^2 = X_i R^{-1} X_i^{\mathrm{T}} \tag{6}$$

$$d_{yh}^2 = Y_h R^{-1} Y_h^{\mathrm{T}} \tag{7}$$

マハラノビス距離 d_x と d_y の分布はそれぞれの変数の数 k に依存する。これから，MT 法の距離 D_x と D_y は以下のように求められる。

$$D_{xi}^2 = \frac{1}{k} X_i R^{-1} X_i^{\mathrm{T}} \tag{8}$$

$$D_{yh}^2 = \frac{1}{k} Y_h R^{-1} Y_h^{\mathrm{T}} \tag{9}$$

3.2 項目選択

D_{yh} における MT 法の距離は k 項目のすべての項目を用いて算出されている。しかし，項目を選択することで余分なノイズをなくし本質的な要因のみを抽出することが考えられる。また，正例と負例の判別精度を向上させ，データの計測コストを削減することができる。

MT 法における項目選択は，2 水準系の直交表に基づき式 (10) の SN 比 ηdb を評価尺度として行う。それは，MT 法の距離 D_{yh} を用い D_{yh} が増加するほど SN 比 ηdb が高くなる評価値である。このことから望大特性の SN 比と呼ばれる。

$$\eta = -10 \log \frac{1}{m} \left(\frac{1}{D_{y1}^2} + \cdots + \frac{1}{D_{ym}^2} \right) \tag{10}$$

正例事象群の各項目を直交表の第 1 列から順に割り当て，それぞれの負例の MT 法の距離から SN 比を算出する。そして，直交表に割り当てた制御因子ごとに SN 比の水準平均をもとに SN 比が高くなる水準を選択し，そこで得られた結果から項目選択を行う。

3.3 しきい値の設定

これまでに計算を行った MT 法の距離を用い，正例か負例かを判別するためのしきい値を決め正例事象群と負例事象群それぞれ判別を行う。そのしきい値 s の設定方法は技術者の判断に基づいて決めるとされている。一般的に，しきい値の目安として『4』という数値が良いとされている (立林ら [5] 参照)。

4 確率分布を用いたしきい値設定法

4.1 χ^2 分布でのしきい値設定法

兼高 [1] は，マハラノビス距離の 2 乗が項目数を自由度とする χ^2 分布に従うことから，χ^2 値を使用したしきい値設定法を試みた。正例事象群のデータから χ^2 分布を使用し技術者の判断により累積確率 α を設定する。そして，正例事象群の項目数を k とし次式からしきい値 s を定める。

$$s = \chi_k^2(\alpha) \tag{11}$$

4.2 ガンマ分布でのしきい値設定法

中津川・大内 [2] は，マハラノビス距離の 2 乗にガンマ分布を仮定することによって正常群の累積確率にもとづくしきい値設定法を提案した。

ガンマ分布 $Ga(a, b)$ により，マハラノビス距離の 2 乗の実測値の分布を確率分布を用いてとらえることが可能となる。ガンマ分布により累積確率の設定値 α に対する正例・負例の判別をするしきい値 s が定まる (中津川・大内 [2] 参照)。

$$P(D_x^2 \leq s) = \int_0^s \frac{b^a}{\Gamma(a)} z^{a-1} \mathrm{e}^{-bz} dz = \alpha \tag{12}$$

ただし，a，b は次のように求める。

$$a = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2} \tag{13}$$

$$b = \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2} \tag{14}$$

$$\hat{\mu}_m = \frac{1}{n} \sum_{i=1}^{n} (D_{xi})^m \quad (m = 1, 2) \tag{15}$$

技術者により定められる α の値は，しきい値 s 以下の MT 法の距離を示す正例の割合に相当する。

4.3 F 分布でのしきい値設定法

現在，χ^2 分布とガンマ分布を用いたしきい値設定法があるが，χ^2 分布に関連した F 分布を用いた方法はまだ試されていない。そこで，Penny[3] がマハラノビス距離の臨界値の設定を F 分布を用いて算出しており，それをしきい値の設定に利用できないかと考えた。

Penny[3] による臨界値の設定が 3 通りあり MT 法の距離でのしきい値設定法を以下のようにして算出する。下式の方法を以降順に F1，F2，F3 と呼ぶこととする。しきい値 s を技術者の判断により累積確率 α を定め，n はサンプルサイズ，k は項目数とする。

$$s = \frac{k(n^2 - 1)}{n(n - k)} F_{k,n-k}(\alpha) \tag{16}$$

$$s = \frac{k(n-1)^2 F_{k,n-k-1}(\alpha)}{n(n - k - 1 + kF_{k,n-k-1}(\alpha))} \tag{17}$$

$$s = \frac{nk(n-2)}{(n-1)(n-k-1)} F_{k,n-k-1}(\alpha) \tag{18}$$

5 しきい値設定法の比較

5.1 分析に用いるデータ

しきい値設定法を比較する為に用いるデータは，事故分類別交通事故データ，2012 年 7 月 1 日の気象データ，うつ病データである。

事故分類別交通事故データは正例事象群を東名高速道路および名神自動車道，または国道 1 号線を通らない都道府県とし，負例事象群をそれら以外の都道府県とする。気象データは正例事象群を北海道と沖縄県を除く地域とし，負例事象群を北海道と沖縄県の地域とする。うつ病データは正例事象群を医師の診断によって正常と診断された人とし，負例事象群を医師の診断によってうつ病と診断された人とする。

5.2 既存のしきい値設定法の単純比較

ここではそれぞれ同じデータを用いてしきい値『4』，χ^2 分布，ガンマ分布の比較を行う。正例事象群の誤判別率と負例事象群の誤判別率，そして，これら 2 つの誤判別率の平均誤判別率をみて比較を行う。ここで，χ^2 分布，ガンマ分布の累積確率は 95 %点で検証する。

ここではうつ病データの結果のみを表 3 に示す。項目選択はそれぞれのしきい値設定法で項目選択をするかしないかを示す。正例は正例誤判別率を示し，負例は負例誤判別率を示す。最後に，平均は正例誤判別率と負例誤判別率を足して 2 で割った数値となっている。

表 3 うつ病データ

	項目選択	正例	負例	平均
しきい値 4	あり	0.0291	0.5200	0.2746
しきい値 4	なし	0.0291	0.5200	0.2746
χ^2 分布	あり	0.0768	0.2800	0.1784
χ^2 分布	なし	0.0887	0.3600	0.2244
ガンマ分布	あり	0.0477	0.4800	0.2638
ガンマ分布	なし	0.0464	0.4400	0.2432

5.3 しきい値『4』の欠点

うつ病データでは「χ^2 分布の項目選択あり」が良い結果となり，交通事故データでは「ガンマ分布の項目選択なし」，気象データでは「χ^2 分布の項目選択あり」が良い結果となった。このことからしきい値『4』は他の確率分布を用いた方法よりも劣っていることがわかる。さらに，表 3 のうつ病データのしきい値『4』では負例の誤判別率が 5 割を超えている。これは負例事象群がうつ病と診断された人のデータ群であるので，実際にうつ病と診断された人が MT 法ではうつ病ではないとして第 2 種の過誤のように誤って診断してしまう人が 5 割ということになる。よって，目安としてしきい値『4』を使用するのは良いが完全な判別として MT 法に使用することはとても危険と言える。

5.4 適合度検定による比較

95 %点だけの比較では各分布によるしきい値設定法の優劣がつけられないためマハラノビス距離の 2 乗がその分布に従っているかどうかをみて評価を行う。そこで，F 分布を用いた方法を含めどの分布がマハラノビス距離の 2 乗に従っているかみるため適合度検定を行う。

以下に，適合度検定の検定手順を説明する。

1. マハラノビス距離の 2 乗を作成する。
2. χ^2 分布，ガンマ分布，F 分布で 10 %刻みなど区間を設けその各区間にいくつマハラノビス距離の 2 乗のサンプルが入るか度数を求める。
3. 求めた度数と 1/(刻み数) の割合とを適合度検定を行い作成したマハラノビス距離の 2 乗がその分布にあてはまっているかをみる。
4. 帰無仮説を「マハラノビス距離の 2 乗が分布に従う」，対立仮説を「マハラノビス距離の 2 乗が分布に従わない」とし検定を行う。

ただし，自由度は，刻み数を n とすると χ^2 分布と F 分布の場合マハラノビス距離の 2 乗の尺度を推定しているため $(n-1)$ とし，ガンマ分布の場合は母数を 2 つ推定するため $(n-3)$ として検定を行う。そして，5.1 節に記載したデータを用い，各分布との適合度検定で求めた p 値の結果を表 4 に示す。F3 は F1 と同じ結果であったため省略する。

表 4 からガンマ分布では項目選択ありなし共に比較的高い数値をとっている。そして,「ガンマ分布の項目選択なし」は $\alpha = 0.05$ で事故データのみ棄却されない。次に良いと考えられる分布は気象データで一番高い数値をとっている F2 であり，項目数に依存してしきい値の設定をする χ^2 分布や F 分布では項目選択ありのほうが全体的に良いことがわかる。ガンマ分布では χ^2 分布や F 分布と違い項目数やサンプル数でなくデータ自身から推測しているため MT 法の距離を捉える事ができたのだと考えられる。また，項目数に依存する χ^2 分布や F 分布では項目選択ありの方が良く，項目選択により余分な項目を削除することで MT 法の距離に近づくことがわかった。

表 4 各分布での適合度検定の p 値

χ^2 分布			
項目選択	事故	気象	うつ病
あり	0.0008	0.0017	2.5×10^{-10}
なし	2.3×10^{-5}	3.5×10^{-7}	1.7×10^{-9}
ガンマ分布			
項目選択	事故	気象	うつ病
あり	0.0016	8.3×10^{-7}	9.5×10^{-5}
なし	0.1822	5.6×10^{-5}	6.7×10^{-5}
F1			
項目選択	事故	気象	うつ病
あり	0.0001	0.0018	1.8×10^{-9}
なし	1.1×10^{-6}	5.5×10^{-10}	1.4×10^{-10}
F2			
項目選択	事故	気象	うつ病
あり	0.0002	0.0068	6.5×10^{-11}
なし	8.5×10^{-6}	2.2×10^{-9}	5.5×10^{-10}

5.5 クロス・バリデーションによる比較

クロス・バリデーションを用いてシミュレーションを行う。クロス・バリデーションの具体的な方法としてはデータを無作為で半分抽出しその半分をしきい値作成のための解析データとして使用し，もう一方の半分のデータを検証用として判別する。また，解析用のデータと検証用のデータを逆にしたときについても判別する。試行回数は 1 万回実施する。検証を行う%点は 90，92.5，95，97.5 である。ここでは 95 %についてのみ結果を表 5 と表 6 に示す。項目選択ありとなしのとき全体の結果では標準偏差の結果から明らかに項目選択なしのときのほうがばらつきが少なかった。誤判別率の平均をみると気象データでは項目選択ありのほうが良いが，うつ病データでは項目選択なしのほうが断然良い。今回使用したデー

タとしては正例・負例の位置づけがしっかりしているうつ病データを優先的にみると項目選択ありのときのばらつきが項目選択なしのときに比べかなりばらついており誤判別率も劣っている。このことから項目選択については項目選択なしのほうが良いと考えられる。

表 5 シミュレーション結果 95 %項目選択なし

	事故	気象	うつ病
χ^2 分布平均	0.3464	0.1953	0.1954
χ^2 分布標準偏差	0.0366	0.0289	0.0193
ガンマ分布平均	0.3518	0.2232	0.2443
ガンマ分布標準偏差	0.0379	0.0316	0.0158
F1 平均	0.2513	0.2457	0.2113
F1 標準偏差	0.0439	0.0318	0.0186
F2 平均	0.3717	0.1859	0.1911
F2 標準偏差	0.0394	0.0296	0.0192
F3 平均	0.2510	0.2469	0.2113
F3 標準偏差	0.0441	0.0318	0.0186

表 6 シミュレーション結果 95 %項目選択あり

	事故	気象	うつ病
χ^2 分布平均	0.3272	0.1582	0.2323
χ^2 分布標準偏差	0.0437	0.0420	0.0467
ガンマ分布平均	0.3282	0.1797	0.2634
ガンマ分布標準偏差	0.0452	0.0414	0.0429
F1 平均	0.2672	0.1858	0.2306
F1 標準偏差	0.0491	0.0438	0.0457
F2 平均	0.3452	0.1528	0.2306
F2 標準偏差	0.0467	0.0406	0.0469
F3 平均	0.2661	0.1865	0.2391
F3 標準偏差	0.0493	0.0437	0.0457

5.6 得点を付けた比較

次に，どの分布を使用するのが良いかをみるためシミュレーションで得られた結果に 1〜10 点の得点を割り振り比較する。得点の割り振り方は，まず項目選択ありなしを区別してすべての%点を含めて各データの誤判別率と標準偏差の最大と最小を求める。求めた最大と最小から (最大 – 最小)/10 によって区間を算出する。この区間により最小値を含む一番小さい値を 10 点とし順に最大値を含む一番大きい値を 1 点とする。

その結果，誤判別率の平均のみで比較すると項目選択ありなしどちらでも F1，F3 が良い。誤判別率の平均と標準偏差を併せて比較すると，項目選択ありなしどちらでもバランスのとれている χ^2 分布が良い結果となった。ガンマ分布では項目選択ありでばらつきが少なく χ^2 分布や F 分布のように項目数やサンプル数ではなくデータ自身から算出されるためガンマ分布は項目選択ありのとき他に比べばらつきが少ないと考えられる。中津川・大内 [2] の提案では項目選択を前提としているがガンマ分布を用いた方法は項目選択ありのほうが良いと分かった。

総合的に，χ^2 分布は項目選択なしでも項目選択ありでもバランス良く使用できるため χ^2 分布を用いる方法が最も良い。また，もう 1 つの理由として F 分布の 90 %点が最も良い誤判別率の数値をとっていたが，χ^2 分布もあまり変わらず良い数値を取っていることも挙げられる。

6 まとめ

本研究では，既存のしきい値設定方法で比較を行いしきい値『4』は分布を用いた方法より誤判別率が悪く，特にうつ病データでは患者の診断ミスが 5 割を超える結果となった。よって，しきい値『4』を使用するのには目安として使用することが良いと述べた。そして，F 分布を用いたしきい値設定法を含め 3 つのしきい値設定法に対し，適合度検定によって分布のあてはまり具合をみた。適合度検定により他の分布に比べガンマ分布がマハラノビス距離の 2 乗に近いものとわかり，χ^2 分布や F 分布についても項目選択をすることであてはまりが良くなることがわかった。次にクロス・バリデーションによってシミュレーションを行いその結果から適合度検定で分布のあてはまりが良いと誤判別が良くなるわけではないことがわかった。シミュレーションにより F 分布を用いた方法では F1，F3 が実用的であることがわかった。

MT 法におけるしきい値設定では誤判別率の良い F 分布を用いたり，項目選択をする際ばらつきを抑えるためガンマ分布を用いることができるが，どちらにも対応できる χ^2 分布を用いる方法が最も良い結果となった。

7 おわりに

項目選択を行うことにより誤判別率がばらつくことがわかり項目選択の弱点を知ることができた。しかし，本研究では項目が 10 変数のデータしか取り扱っておらず，文字認識のような項目が多いものについては触れていない。項目選択をしないといけないものについての議論が必要であり項目が多いときどのようにして項目選択を行うかも考えなければならない。また，負例事象群のデータも最大で 30 サンプルでありもっと十分に多い場合の安定性も調べていない。その場合には項目選択が有利に働く可能性があることにも注意する。

参考文献

[1] 兼高達貳:マハラノビスの汎距離の応用例　特殊健康診断の事例, 『標準化と品質管理』, pp.57-64, 東京, 1987.

[2] 中津川雅史, 大内東:MTS アルゴリズムにおけるしきい値設定法に関する考察,『電子情報通信学会論文誌』, pp.519-527, 東京, 2001.

[3] Penny, Kay I: Appropriate Critical Values when Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance, *Appl. Statist.*, vol. 45, No. 1, pp. 73-81, 1996.

[4] 田口玄一:『MT システムにおける技術開発』, 日本規格協会，東京，2002.

[5] 立林和夫, 手島昌一, 長谷川良子:『入門 MT システム』, 日科技連出版社, 東京, 2008.