MT 法におけるしきい値設定法の提案と比較

安部 将成* 松田 眞一†
E-Mail: matsu@nanzan-u.ac.jp

品質工学の中に検査事象の異常判定を行う MT 法という方法がある。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』という数値に決めただけのものであった。現在，χ^2 分布を用いた方法やガンマ分布を用いた方法が提案されている。しかし，この 2 つの分布を用いた方法はどちらが優れているか研究されていない。そこで，F 分布を用いた MT 法のしきい値設定法を提案し，それも含めてどの方法がしきい値設定に適しているか適合度検定やシミュレーションによって比較を行った。結果として，確率分布を用いた方法にはそれぞれ良さがあるが，安定性の観点から χ^2 分布を用いた方法がよいと分かった。

1 はじめに

現在，製造業では品質第一・品質向上・品質基準などの観点から品質について重要視されるようになり，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活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利用されている。

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』という数値に決めただけのものであった（立林ら [8] 参照)。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 [4] が提案したガンマ分布を用いた方法がある。しかし，この 2 つの分布はどちらが優れているか研究されていない。また，マハラノビス距離の 2 乗は χ^2 分布に従うことが分かっており (田口 [6] 参照)，χ^2 分布とその分布に関連性のあるガンマ分布を用いた方法はあるが，同じく χ^2 分布に関連性のある F 分布を用いた方法は検討されていない。

そこで，本論文では F 分布を用いたしきい値設定法を提案し，しきい値を『4』とする方法や今までに提案されている確率分布を用いた方法と比較し，それぞれのしきい値設定法の性質について研究する。

2 MT 法の概要

MT 法は，検査事象の異常判定を行う方法であるが，そのために正例事象群と負例事象群が必要とされる。製造業でいうと，正例事象群とは正常な製造品（の測定データ）を指し，負例事象群とはいわゆる不良品（の測定データ）を指す。正例事象群からデータの平均と相関係数行列を算出し，それらで計算されたマハラノビス距離から異常判定を行う。以下にその概要を記す。(田口 [6]，立林ら [8]，中津川・大内 [4] 参照)

*南山大学大学院数理情報研究科数理情報専攻
†南山大学情報理工学部情報システム数理学科

2.1 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する。マハラノビス距離とは，正確にはマハラノビスの汎距離と呼ばれるもので，データの基準点および単位量に基づく多変量データの評価尺度である。正例事象群の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，MT 法の距離は正例事象群を構成する事象のマハラノビス距離の平均を 1 とするように基準化されて定義される (田口 [6] 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 は次の過程で求められる。正例事象群データ x_{ij} $(i = 1, 2, \cdots, n; j = 1, 2, \cdots, k)$ に基づき，各項目の平均 $\bar{x}_j$ および標準偏差 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 $\bar{x}_j$ および標準偏差 s_j を用いて，x_{ij} と y_{hj} $(h = 1, 2, \cdots, m)$ の基準化を行う。負例事象群も正例事象群の基準を用いて基準化されることに注意する。

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

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

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の 2 乗 d_x^2 と d_y^2 を以下の算出方法で求める。

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

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

マハラノビス距離 d_x と d_y の分布はそれぞれ項目数 k に依存するため，MT 法の距離 D_x と D_y は以下のように求められる。

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

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

2.2 項目選択

D_{yh} における MT 法の距離は k 項目のすべて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。（田口 [6] 参照）

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 η db を評価尺度として行う。それは，望大特性の SN 比であり，MT 法の距離 D_{yh} を用い D_{yh} が増加するほど SN 比 η db が高くなる評価値である。

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

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，それぞれの負例事象群の 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 比の水準平均によって 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。本論文では L_{12} 直交表を使用する。

2.3　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が，これは $10\log_{10} D^2$ に対するしきい値である (立林ら [8] 参照)。

3　確率分布を用いたしきい値設定法

3.1　χ^2 分布を用いたしきい値設定法

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に χ^2 分布を適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 k を用い，次式によりマハラノビス距離の 2 乗に対するしきい値 s を定める。

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

3.2　ガンマ分布を用いたしきい値設定法

中津川・大内 [4] は，MT 法の距離の 2 乗にガンマ分布を仮定することによって正例群の累積確率に基づくしきい値設定法を提案した。

MT 法の距離の 2 乗の実測値の分布はガンマ分布 $Ga(a,b)$ を用いて近似的にとらえることが可能となる。累積確率の設定値 α に対するガンマ分布のパーセント点より正例・負例の判別をするしきい値 s' が定まる。(中津川・大内 [4] 参照)

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

ただし，a，b は以下のように求める。

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

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

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

技術者により定められる α の値は，近似的にしきい値 s' 以下となる MT 法の距離の 2 乗を示す正例の割合に相当する。

3.3 F 分布を用いたしきい値設定法

現在，χ^2 分布とガンマ分布を用いたしきい値設定法はあるが，χ^2 分布に関連した F 分布を用いた方法はない。そこで，Penny[5] がマハラノビス距離の臨界値の設定について F 分布を用いて算出しているので，この方法をしきい値の設定に利用できないかと考えた。

Penny[5] による臨界値の設定は 3 通りあり，マハラノビス距離の 2 乗でのしきい値設定法を以下のようにして 3 種類算出する。下式の方法を以降順に F1，F2，F3 と呼ぶこととする。マハラノビス距離の 2 乗に対するしきい値 s を技術者の判断により累積確率 α から計算する。

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

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

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

4 しきい値設定法の比較

4.1 分析に用いるデータ

しきい値設定法を比較するために用いるデータは，事故分類別交通事故データ，気象データ，うつ病データである。

事故分別交通事故データは交通安全マップ [2] から作成された事故分類別交通事故数のものであり，都道府県別の 47 のデータを用いる。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。正例事象群は 37 サンプル，負例事象群は 10 サンプル，項目は以下の 10 項目である。

対人事故：	対面通行中	背面通行中	横断中		
対車事故：	正面衝突	出会い頭	右折時	左折時	追突
対物事故：	電柱標識	駐車車両			

気象データは気象庁ホームページ [3] から気象統計情報の毎日の全国データ一覧表より検索した 2012 年 7 月 1 日の全国 151 箇所のデータである。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。正例事象群は 121 サンプル，負例事象群は 30 サンプル，項目は以下の 10 項目である。

平均現地気圧	平均海面気圧	平均気温	最高気温	最低気温
平均湿度	最小湿度	風速平均	風速最大	風速最大瞬間

うつ病データは 10 項目のテスト結果データであり，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。正例事象群は 755 サンプル，負例事象群は 25 サンプルを使用する。

これらのデータは棚橋・松田 [7] の研究で使用されているデータを参考としている。

4.2 既存のしきい値設定法の単純比較

本節ではそれぞれ同じデータを用いて，まず既存の方法であるしきい値『4』を用いた方法，χ^2 分布を用いた方法，ガンマ分布を用いた方法の比較を行う。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 %を用いて検証することとする。

ここではうつ病データの結果のみを表 3 に示す。表中,「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,「負例」は負例誤判別率を示し,「平均」は正例誤判別率と負例誤判別率の算術平均を示す。

表 3: うつ病データの判別結果

	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

4.3 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 割を超えるということになる。

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。

4.4 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 乗に従っているかを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. 正例事象群に対し，マハラノビス距離の 2 乗を作成する。
2. 帰無仮説を「マハラノビス距離の 2 乗が対象とする分布に従う」，対立仮説を「マハラノビス距離の 2 乗が対象とする分布に従わない」とする。
3. χ^2 分布，ガンマ分布，F 分布のそれぞれで 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。
4. 求めた度数と 1/(刻み数) の割合との適合度検定を行い，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。

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さらに母数を 2 つ推定するため $(n-3)$ として検定を行う。そして，4.1 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ことが分かる。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サイズでなくデータ自身から推測しているためマハラノビス距離の 2 乗を捉える事ができたと考えられる。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マハラノビス距離の 2 乗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

分布	項目選択	事故	気象	うつ病
χ^2 分布	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	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	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	あり	0.0002	0.0068	6.5×10^{-11}
	なし	8.5×10^{-6}	2.2×10^{-9}	5.5×10^{-10}
F3	あり	0.0001	0.0018	1.8×10^{-9}
	なし	1.1×10^{-6}	5.5×10^{-10}	1.4×10^{-10}

4.5　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に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する。そして，残り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 万回とする。検証を行う%点は 90，92.5，95，97.5 であり，その結果を表 5〜表 8 に示す。

表 5: 各分布でのシミュレーション結果 90 %点

	項目選択なし			項目選択あり		
	事故	気象	うつ	事故	気象	うつ
χ^2 分布誤判別率の平均	0.3575	0.1796	0.1592	0.3372	0.1411	0.2155
χ^2 分布誤判別率標準偏差	0.0381	0.0294	0.0169	0.0444	0.0317	0.0470
ガンマ分布誤判別率の平均	0.3629	0.1942	0.1869	0.3390	0.1437	0.2241
ガンマ分布誤判別率標準偏差	0.0392	0.0290	0.0191	0.0457	0.0335	0.0462
F1 誤判別率の平均	0.2614	0.2152	0.1655	0.2754	0.1483	0.2175
F1 誤判別率標準偏差	0.0431	0.0300	0.0174	0.0469	0.0360	0.0467
F2 誤判別率の平均	0.3771	0.1761	0.1581	0.3518	0.1406	0.2151
F2 誤判別率標準偏差	0.0393	0.0285	0.0167	0.0470	0.0308	0.0471
F3 誤判別率の平均	0.2552	0.2160	0.1655	0.2714	0.1485	0.2175
F3 誤判別率標準偏差	0.0435	0.0300	0.0173	0.0475	0.0361	0.0467

表 6: 各分布でのシミュレーション結果 92.5 %点

	項目選択なし			項目選択あり		
	事故	気象	うつ	事故	気象	うつ
χ^2 分布誤判別率の平均	0.3525	0.1857	0.1704	0.3325	0.1461	0.2205
χ^2 分布誤判別率標準偏差	0.0376	0.0296	0.0172	0.0441	0.0362	0.0471
ガンマ分布誤判別率の平均	0.3579	0.2063	0.2153	0.3344	0.1547	0.2379
ガンマ分布誤判別率標準偏差	0.0387	0.0295	0.0185	0.0454	0.0376	0.0448
F1 誤判別率の平均	0.2556	0.2280	0.1817	0.2709	0.1613	0.2241
F1 誤判別率標準偏差	0.0434	0.0314	0.0181	0.0476	0.0398	0.0467
F2 誤判別率の平均	0.3746	0.1794	0.1683	0.3487	0.1444	0.2198
F2 誤判別率標準偏差	0.0393	0.0294	0.0173	0.0466	0.0350	0.0472
F3 誤判別率の平均	0.2516	0.2290	0.1817	0.2678	0.1617	0.2241
F3 誤判別率標準偏差	0.0440	0.0314	0.0181	0.0484	0.0398	0.0467

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 7: 各分布でのシミュレーション結果 95 %点

	項目選択なし			項目選択あり		
	事故	気象	うつ	事故	気象	うつ
χ^2 分布誤判別率の平均	0.3464	0.1953	0.1954	0.3272	0.1582	0.2323
χ^2 分布誤判別率標準偏差	0.0366	0.0289	0.0193	0.0437	0.0420	0.0467
ガンマ分布誤判別率の平均	0.3518	0.2232	0.2443	0.3282	0.1797	0.2634
ガンマ分布誤判別率標準偏差	0.0379	0.0316	0.0158	0.0452	0.0414	0.0429
F1 誤判別率の平均	0.2513	0.2457	0.2113	0.2672	0.1858	0.2391
F1 誤判別率標準偏差	0.0439	0.0318	0.0186	0.0491	0.0438	0.0457
F2 誤判別率の平均	0.3717	0.1859	0.1911	0.3452	0.1528	0.2306
F2 誤判別率標準偏差	0.0394	0.0296	0.0192	0.0467	0.0406	0.0469
F3 誤判別率の平均	0.2510	0.2469	0.2113	0.2661	0.1865	0.2391
F3 誤判別率標準偏差	0.0441	0.0318	0.0186	0.0493	0.0437	0.0457

表 8: 各分布でのシミュレーション結果 97.5 %点

	項目選択なし			項目選択あり		
	事故	気象	うつ	事故	気象	うつ
χ^2 分布誤判別率の平均	0.3377	0.2136	0.2347	0.3200	0.1887	0.2640
χ^2 分布誤判別率標準偏差	0.0358	0.0297	0.0161	0.0439	0.0488	0.0461
ガンマ分布誤判別率の平均	0.3428	0.2507	0.2876	0.3215	0.2325	0.3086
ガンマ分布誤判別率標準偏差	0.0368	0.0325	0.0194	0.0452	0.0424	0.0388
F1 誤判別率の平均	0.2531	0.2716	0.2462	0.2673	0.2336	0.2749
F1 誤判別率標準偏差	0.0444	0.0305	0.0149	0.0491	0.0428	0.0446
F2 誤判別率の平均	0.3679	0.1984	0.2305	0.3414	0.1751	0.2604
F2 誤判別率標準偏差	0.0393	0.0289	0.0163	0.0466	0.0487	0.0464
F3 誤判別率の平均	0.2586	0.2726	0.2462	0.2696	0.2345	0.2749
F3 誤判別率標準偏差	0.0449	0.0305	0.0149	0.0487	0.0426	0.0446

4.6 得点を付けた比較

次に，複数の%点に対し，どの分布を使用するのが良いかを総合的に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 9 点，8 点，…とし，最大値を含む一番大きい値を 1 点とする。90 %〜97.5 %の得点を合計したものを表 9 に示す。

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 が良いことが分かり，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サイズではなくデータ自身に適合させて算出されるため，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 [4] の提案では項目選択を前提としておりシミュレーション

結果から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。

また，前章の適合度検定の結果では「ガンマ分布の項目選択なし」や「F2 の項目選択あり」が良いが，シミュレーションの結果から分布に従っているかどうかで誤判別率が良くなるわけではないことがわかった。この理由として，正例事象群の MT 法の距離が分布に従うかどうかで判断しており負例事象群については SN 比の計算でしか使用しないため負例事象群の情報をあまり使わないことが挙げられる。そのため，ガンマ分布のようにパラメータを推定する方法だと今までとは異なる方向に飛び出た場合にうまく対応できていないと思われる。

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。

表 9: 得点のまとめ結果

	項目選択なし				項目選択あり			
	事故	気象	うつ	合計	事故	気象	うつ	合計
χ^2 分布誤判別率の平均	11	36	33	80	12	34	34	80
χ^2 分布誤判別率標準偏差	36	34	20	90	39	22	5	66
ガンマ分布誤判別率の平均	10	25	19	54	12	26	24	62
ガンマ分布誤判別率標準偏差	31	21	13	65	30	25	21	76
F1 誤判別率の平均	40	15	29	84	39	25	32	96
F1 誤判別率標準偏差	8	18	20	46	11	20	8	39
F2 誤判別率の平均	4	37	33	74	5	36	35	76
F2 誤判別率標準偏差	28	36	20	84	20	24	4	48
F3 誤判別率の平均	40	15	29	84	40	25	32	97
F3 誤判別率標準偏差	6	18	20	44	9	20	8	37

5　まとめ

本論文では，まず既存のしきい値設定方法間で比較を行った結果，しきい値『4』を用いた方法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 割を超えることが分かった。よって，しきい値『4』を使用するのは目安に留めるのが良いと述べた。

次に，F 分布を用いたしきい値設定法を含め 3 種類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確認した。適合度検定により他の分布に比べガンマ分布が実際のマハラノビス距離の 2 乗に近いものと分かり，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傾向が見られた。

最後にクロス・バリデーションによって誤判別率を比較する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

分かった。このシミュレーションにより提案した F 分布を用いた方法では F1，F3 が実用的であると言える。

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総合的には最も良い結果となった。

6　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，本論文では項目が 10 項目まで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 (1987): 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』, **40**(10), 57-64.

[2] 警察庁, 国土交通省 (2007): 交通安全マップ, http://www.kotsu-anzen.jp/.

[3] 気象庁 (2012): 毎日の全国データ一覧表,
http://www.data.jma.go.jp/obd/stats/data/mdrr/synopday/index.html/.

[4] 中津川雅史・大内東 (2001): 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』, **J84-A**(4), 519-527.

[5] Penny, Kay I (1996): Appropriate Critical Values when Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance, *Appl. Statist.*, **45**(1), 73-81.

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

[7] 棚橋誠・松田眞一 (2007)： MTS 法と各距離における分析法の比較, 南山大学紀要『アカデミア』情報理工編, **7**, 21-32.

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