Journal of the Korean Institute of Industrial Engineers
Vol. 39, No. 5, pp. 393-402, October 2013.
ISSN 1225-0988 | EISSN 2234-6457
http://dx.doi.org/10.7232/JKIIE.2013.39.5.393

<Original Research Paper>

마할라노비스-다구치 시스템과 로지스틱 회귀의 성능비교 : 사례연구

이승훈[1†] · 임 근[2]

[1]동의대학교 산업경영공학과 / [2]육군종합정비창 정비기술연구소

Performance Comparison of Mahalanobis-Taguchi System and Logistic Regression : A Case Study

Seung-Hoon Lee[1] · Geun Lim[2]

[1]Department of Industrial and Management Engineering, Dong-Eui University
[2]Maintenance Technology Laboratory, Korea Military Consolidated Maintenance Depot

The Mahalanobis-Taguchi System (MTS) is a diagnostic and predictive method for multivariate data. In the MTS, the Mahalanobis space (MS) of reference group is obtained using the standardized variables of normal data. The Mahalanobis space can be used for multi-class classification. Once this MS is established, the useful set of variables is identified to assist in the model analysis or diagnosis using orthogonal arrays and signal-to-noise ratios. And other several techniques have already been used for classification, such as linear discriminant analysis and logistic regression, decision trees, neural networks, etc. The goal of this case study is to compare the ability of the Mahalanobis-Taguchi System and logistic regression using a data set.

Keywords: Mahalanobis-Taguchi System(MTS), Logistic Regression, Classification, Diagnosis

1. 서 론

인도의 수학자 Mahalanobis는 1930년대에 한 집단과 이질의 집단을 구분하는 방법으로 마할라노비스 거리(Mahalanobis Distance : MD)를 이용하는 방법을 제안하였다. 그리고 Taguchi 는 1990년대 초반에 마할라노비스 거리(MD)와 Taguchi 방법을 접목시켜 Mahalanobis-Taguchi System(MTS) 기법을 제안하였다. MTS 기법에서는 먼저 정상 집단에 대하여 다차원의 단위공간으로 마할라노비스 공간(Mahalanobis Space)을 정의하고, 임의의 대상이 그 공간으로부터 얼마나 떨어져 있는가를 거리(MD)로 산정한다. 즉, 거리가 멀어질수록 공간으로서 선정된 정상 집단에서 멀리 떨어진 것을 의미한다. 그리고 선정된 많은 측정변수 중에서 MD에 영향이 큰 유용한 변수를 추출하기 위하여 Taguchi 방법을 이용한다. MTS 기법은 진단, 패턴 인식, 음성인식, 최적화 등의 분야에서 많은 적용 사례 연구가 이루어지고 있다(Taguchi and Jugulum, 2000; Taguchi *et al*., 2001; Taguchi and Jugulum, 2002; Woodall *et al*., 2003).

MTS 기법과 비슷한 용도로 사용되는 다변량분석(multivariate analysis) 및 데이터마이닝(data mining)에서의 분류기법은 판별분석(discriminant analysis), 로지스틱 회귀(logistic regression), 주성분 분석(principal component analysis), 나이브 베이지안 분류기(naive bayesian classifier), ROC(Receiver Operating Characteristic) curve 분석, 의사결정나무(decision tree), 신경망(neural network), 서포트 벡터 머신(support vector machine), 근접이웃 분류기(nearest neighbor classifier) 등등이 있다(Izenman, 2008; Tang *et al*., 2006; Hastie *et al*., 2009).

MTS와 분류기법간의 성능을 비교한 선행연구를 살펴보면 <Table 1>과 같다. Jugulum and Monplaisir(2002)는 독립변수가

† 연락저자 : 이승훈 교수, 614-714 부산광역시 부산진구 엄광로 176 동의대학교 산업경영공학과, Tel : 051-890-1656, Fax : 051-890-2627
E-mail : shlee@deu.ac.kr
2013년 6월 28일 접수; 2013년 8월 24일 수정본 접수; 2013년 9월 2일 게재 확정.

Table 1. Performance comparison studies

Authors	Comparison methods	Results
Jugulum and Monplaisir(2002)	MTS vs. NN	• MTS has better performance in small sample size
Wang *et al.*(2004)	MTS vs. LDA	• MTS has better performance
Cudney *et al.*(2007)	MTS vs. NN	• MTS has better performance in small sample size
Cudney *et al.*(2009)	MTS vs. PCA	• MTS has better performance for type II error
Vardhan *et al.*(2012)	MTS vs. ROC	• ROC Curve has better performance

15개인 이항분류인 의료 데이터 셋을 사용하여 분류정확도 관점에서 MTS와 신경망기법(NN : Neural Networks)의 성능비교 연구를 수행하였다. 이 연구에서는 대규모 샘플(large sample)에서는 두 가지 방법이 서로 비슷한 성능을 보였으며, 소규모 샘플(small sample)에서는 MTS가 우수하다는 결과를 보였다. Wang *et al.*(2004)은 MTS와 선형판별분석(LDA : Linear Discriminant Analysis)간의 성능비교 연구를 수행하였다. 2개의 데이터 셋(Iris data : 4개 변수, 다항분류 50개 데이터; Credit card data : 26개 변수, 이항분류 6000개 데이터)을 사용하였으며, 분류정확도 관점에서 MTS가 우수하다는 결과를 나타내었다. Cudney *et al.*(2007)은 1개의 데이터 셋(Wisconsin Breast Cancer Data : 9개 변수, 이항분류 683개 데이터)으로 MTS와 신경망기법간의 성능비교 연구를 수행하였는데, 분류정확도 관점에서 소규모 샘플(20～30개)에서는 MTS가 우수하였고, 대규모 샘플(100개)에서는 두 방법 모두 비슷한 성능을 나타내었다. Cudney *et al.*(2009)은 MTS와 주성분 분석(PCA : Principal Component Analysis)간의 성능비교를 위하여 시뮬레이션 데이터 셋(17개 변수, 이항분류 500개 데이터)를 이용하였으며, Type I error의 관점에서는 서로 비슷한 성능을 나타내었으며, Type II error의 관점에서는 MTS가 우수한 성능을 나타내었다. Vardhan *et al.*(2012)은 1개의 데이터 셋(3개 변수, 이항분류 100개 데이터)을 사용하여 MTS와 ROC(Receiver Operating Characteristic) Curve 분석기법간의 성능비교 연구를 수행하였다. 이 연구에서는 ROC Curve 분석기법이 MTS보다 우수하다는 결과를 나타내었다.

MTS와 분류기법 간의 성능비교에 관한 선행연구에서 로지스틱 회귀(logistic regression)와 비교연구가 없었으므로, 본 연구에서는 MTS 기법 및 로지스틱 회귀와 성능비교 연구를 수행하고자 한다. 이를 위하여 먼저 MTS 기법 및 로지스틱 회귀(logistic regression)의 절차에 대하여 살펴보고, 사례 데이터에 대하여 MTS 기법을 적용하여 분류 정확도를 파악하고, 로지스틱 회귀를 적용하여 분류 정확도를 파악하여, 분류 정확도의 관점에서 두 기법의 성능을 비교하고자 한다.

2. Mahalanobis-Taguchi System(MTS) 기법

MTS 기법은 다음 네 단계의 절차로 구성된다(Taguchi and Jugulum, 2002).

[1단계] 정상 집단의 MS(Mahalanobis Space) 구성

1단계에서는 정상 집단으로부터 판단에 적용될 측정변수들을 선정하고, 선정된 변수들에 의해 구성된 마할라노비스 공간(MS)을 구성한다. 이를 위하여 먼저 이상치가 없는 균일한 정상집단의 데이터를 구성하여야 한다. <Table 2>와 같이 측정변수가 k 개인 정상 집단의 데이터가 n 개로 구성되어 있다면, 다음 세 가지 과정으로 마할라노비스 거리(MD)를 계산한다.

Table 2. Healthy group data

No.	X_1	X_2	⋯	X_k
1	x_{11}	x_{21}	⋯	x_{k1}
2	x_{12}	x_{22}	⋯	x_{k2}
3	x_{13}	x_{23}	⋯	x_{k3}
4	x_{14}	x_{24}	⋯	x_{k4}
5	x_{15}	x_{25}	⋯	x_{k5}
⋮	⋮	⋮	⋮	⋮
n	x_{1n}	x_{2n}	⋯	x_{kn}
mean	$\overline{x_1}$	$\overline{x_2}$	⋯	$\overline{x_k}$
std. dev.	s_1	s_2	⋯	s_2

(1) 정상 집단 데이터의 표준화

정상 집단의 데이터 x_{ij}에서 평균 $\overline{x_i}$를 빼고 표준편차 s_i로 나누어 표준화시킨다.

$$z_{ij} = \frac{x_{ij} - \overline{x_i}}{s_i}, \quad i = 1, 2, \cdots, k;\ j = 1, 2, \cdots, n \qquad (1)$$

(2) 상관행렬 계산

정상 집단의 표준화 데이터 z_{ij}에 대한 상관행렬(correlation matrix) C를 구한다.

(3) 정상 집단 데이터의 MD 계산

정상 집단의 마할라노비스 거리(MD)를 다음 식을 이용하여 계산한다.

$$MD_j = D_j^2 = \frac{1}{k} Z_{ij}^T C^{-1} Z_{ij}, \quad j = 1, 2, \cdots, n \qquad (2)$$

참고로 MTS에서는 데이터의 분포는 따로 가정하고 있지 않으나, 만일 정규분포를 따른다고 가정하면 Johnson and Wichern (1992)에 의하여 $Z_{ij}^T C^{-1} Z_{ij}$는 $\chi^2(k)$ 분포를 따름을 증명할 수 있다. 따라서 $E(Z_{ij}^T C^{-1} Z_{ij}) = k$, $V(Z_{ij}^T C^{-1} Z_{ij}) = 2k$ 이므로 $E(MD_j) = 1$, $V(MD_j) = 2/k$가 된다. 그러므로 만일 정상집단의 데이터가 유효하게 잘 구성되어 있다면, 구해진 MD 값의 평균값은 대략 1에 근접한다. 이런 이유로 정상집단의 MD값으로 구성된 MS 공간을 단위공간(unit space)이라고 부른다.

그리고 MTS 절차의 MD 계산에서 상관행렬 C에 다중공선성(multicollinearity) 문제가 존재하면, 상관행렬의 역행렬 계산 시 정확도가 떨어져 부정확한 MD값이 나올 수 있다. Taguchi는 이러한 경우에 대하여 정확한 MD 계산을 위하여 Mahalanobis-Taguchi-Gram-Schmidt(MTGS) 절차를 개발하였다. MTGS 절차에서는 상관행렬의 역행렬을 이용하지 않고 다음과 같이 Gram-Schmidt 직교화 절차를 적용하여 MD값을 계산하며, 식 (2)의 MD값과 동일함을 증명하였다(Taguchi and Jugulum, 2002; 25-27).

$$U_1 = (u_{11}, u_{12,}\cdots, u_{1n}) = Z_1$$
$$U_2 = (u_{21}, u_{22,}\cdots, u_{2n}) = Z_2 - \frac{Z_2^T U_1}{U_1^T U_1} U_1$$
$$\vdots$$
$$U_k = (u_{k1}, u_{k2,}\cdots, u_{kn}) = Z_k - \frac{Z_k^T U_1}{U_1^T U_1} U_1 - \cdots - \frac{Z_k^T U_{k-1}}{U_{k-1}^T U_{k-1}} U_{k-1}$$
$$MD_j = D_j^2 = \frac{1}{k}\left(\frac{u_{1j}^2}{s_1^2} + \frac{u_{2j}^2}{s_2^2} + \cdots + \frac{u_{kj}^2}{s_k^2}\right)$$

여기서, $s_i = U_i$의 표준편차이다.

[2단계] 비정상 집단의 선정 및 MS 공간의 유효성 확인

2단계에서는 1단계에서 구해진 MS 공간의 유효성을 판단한다. 이를 위해서 MS 공간 밖의 측정값인 비정상 집단의 데이터를 이용한다. <Table 3>과 같이 측정변수가 k개인 비정상 집단의 데이터가 m 개로 구성되어 있다면, 다음 세 가지 과정으로 마할라노비스 거리(MD)를 계산하여 MS 공간의 유효성을 확인한다.

Table 3. Unhealthy group data

No.	X_1	X_2	⋯	X_k
1	y_{11}	y_{21}	⋯	y_{k1}
2	y_{12}	y_{22}	⋯	y_{k2}
3	y_{13}	y_{23}	⋯	y_{k3}
4	y_{14}	y_{24}	⋯	y_{k4}
5	y_{15}	y_{25}	⋯	y_{k5}
⋮	⋮	⋮	⋮	⋮
m	y_{1m}	y_{2m}	⋯	y_{km}

(1) 비정상 집단 데이터의 표준화

비정상 집단의 데이터 y_{ij}를 정상 집단의 평균 $\overline{x}_i$와 표준편차 s_i를 이용하여 표준화시킨다.

$$w_{ij} = \frac{y_{ij} - \overline{x_i}}{s_i}, \quad i = 1, 2, \cdots, k\,;\, j = 1, 2, \cdots, m \quad (3)$$

(2) 비정상 집단 데이터의 MD 계산

비정상 집단의 마할라노비스 거리(MD)를 다음 식과 같이 정상 집단의 상관행렬 C를 이용하여 계산한다.

$$MD_j = D_j^2 = \frac{1}{k} W_{ij}^T C^{-1} W_{ij}, \; j = 1, 2, \cdots, m \quad (4)$$

(3) MS 공간의 유효성 확인

MS 공간이 유효하다면 비정상 집단의 MD 값은 정상 집단의 MD 값보다 훨씬 크게 나타난다.

[3단계] 유용한 변수 선정

3단계에서는 측정변수 중에서 MD 값에 영향을 미치지 않거나 적게 미치는 변수를 찾아내고 제거하여 시스템을 쉽게 해석하는 일이다. 이러한 목적을 위하여 Taguchi 방법이 적용된다. 먼저, 모든 측정변수를 배치할 수 있는 적절한 2수준계 직교표를 선정하여 측정변수를 배치한다. 여기서 직교표의 수준 1은 해당 측정변수를 사용하는 경우를 의미하고, 수준 2는 해당 측정변수를 사용하지 않는 경우를 의미한다. 직교표 각 행의 실험조건으로 비정상 집단의 데이터 y_{ij}를 대상으로 하여 비정상 집단의 MD를 계산한다. 이렇게 하여 얻어진 비정상 집단의 MD값에 대한 SN비를 계산한다. 이항분류인 경우에는 SN비는 비정상 집단의 MD값이 크면 클수록 해당 단위 공간이 유용한 것을 나타내므로 망대특성의 SN비 공식을 사용하고, 범주의 수가 3개 이상인 다항분류인 경우에는 동특성의 SN비 공식을 사용한다. SN비에 대한 수준별 평균값을 산출하고, 측정변수의 사용 이득(gain)을 파악하여 유용한 측정변수를 선정한다.

[4단계] 유용한 변수로 구성된 진단 시스템 구축

4단계에서는 3단계에서 선정된 유용한 측정변수들로 구성된 미래의 진단 시스템을 구축한다. 이때 정상과 비정상을 구분하는 MD값의 임계값(threshold value)을 앞 단계에서의 결과를 토대로 설정하고, 적절한 행동 조치 사항을 정하여 모니터링 시스템을 구축한다.

3. 로지스틱 회귀

로지스틱 회귀(logistic regression)는 종속변수가 범주형인 경우에 적용되는 회귀분석 기법이다. 종속변수의 범주의 수가 2개

인 경우의 로지스틱 회귀가 이항 로지스틱 회귀이며, 범주의 수가 3개 이상인 경우의 로지스틱 회귀가 다항 로지스틱 회귀이다. 본 절에서는 이항 로지스틱 회귀의 절차를 간단히 살펴보도록 한다. 이항 로지스틱 회귀에서 사용되는 모형은 로짓모형(logit model), 프로빗모형(probit model), 곰핏모형(gompit model) 등 등이 있다(Hosmer and Lemeshow, 2000). 본 연구에서는 로짓모형을 적용하여 분석하므로 로짓모형에 대하여 기술한다. 이항 로지스틱 회귀의 로짓모형은 식 (5)와 같이 주어진다.

$$\log\left(\frac{\pi}{1-\pi}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k \quad (5)$$

여기서 $\pi=P[Y=1\mid X_1=x_1, X_2=x_2, \cdots, X_k=x_k]$

주어진 데이터을 이용하여 로짓모형을 최우추정법에 의하여 추정하여 식 (6)의 추정된 회귀모형을 얻는다.

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)=b_0+b_1x_1+b_2x_2+\cdots+b_kx_k \quad (6)$$

추정된 회귀모형으로부터 식 (7)에 의하여 주어진 독립변수에서 종속변수가 1일 확률 π를 추정하고, 확률이 높은 범주에 예측값을 부여한다.

$$\hat{\pi}=\frac{\exp(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}{1+\exp(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)} \quad (7)$$

4. MTS와 로지스틱 회귀의 성능 비교

본 절에서는 사례 데이터를 이용하여 MTS와 로지스틱 회귀의 성능을 비교하고자 한다. 본 연구에서 이용한 사례 데이터는 Johns Hopkins University Ionosphere data이다. 이 데이터 셋의 출처는 UC Irvine Machine Learning Repository(http://archive.ics.uci.edu/ml/ datasets.html)이며, 이곳의 수많은 데이터 셋 중에서 비교적 측정변수(독립변수)의 수가 많은 것을 선정하였다. 이 데이터 셋은 측정(독립)변수가 32개의 연속형 변수로만 구성되어 있으며, 종속(반응)변수는 이항분류로 good(225개)과 bad(126개)의 총 351개로 구성되어 있다.

본 연구에서는 총 351개 데이터 중에서 약 20%(good 48개, bad 32개 총 80개)를 훈련용 데이터(training data)로 랜덤하게 선정하였다(<Table 4>, <Table 5>). 그리고 나머지 약 80%(good 177개, bad 94개 총 271개)를 검증용 데이터(testing data)로 하였다.

Table 4. Healthy group data

No.	X1	X2	X3	X4	X5	⋯	X30	X31	X32	Y
1	0.99539	-0.05889	0.85243	0.02306	0.83398	⋯	-0.54487	0.18641	-0.45300	g
2	0.97588	-0.10602	0.94601	-0.20800	0.92806	⋯	-0.81318	-0.13832	-0.80975	g
3	1.00000	0.07380	1.00000	0.03420	1.00000	⋯	0.32492	1.00000	0.46712	g
4	1.00000	-0.08714	1.00000	-0.17263	0.86635	⋯	-0.92128	-0.13341	-1.00000	g
5	1.00000	-0.15899	0.72314	0.27686	0.83443	⋯	0.53372	1.00000	-0.57758	g
⋮	⋮	⋮	⋮	⋮	⋮	⋯	⋮	⋮	⋮	⋮
45	1.00000	1.00000	0.36700	0.06158	0.12993	⋯	0.00246	0.17758	0.79790	g
46	0.32789	0.11042	0.15970	0.29308	0.14020	⋯	0.20540	0.13376	0.26422	g
47	0.19466	0.05725	0.04198	0.25191	-0.10557	⋯	0.16794	-0.30153	-0.33588	g
48	0.66667	-0.01366	0.97404	0.06831	0.49590	⋯	0.24590	0.13934	0.48087	g

Table 5. Unhealthy group data

No.	X1	X2	X3	X4	X5	⋯	X30	X31	X32	Y
1	0.00000	0.00000	0.00000	0.00000	1.00000	⋯	1.00000	0.00000	0.00000	b
2	1.00000	-0.86701	1.00000	0.22280	0.85492	⋯	0.88428	1.00000	-0.18826	b
3	0.00000	0.00000	-1.00000	-1.00000	1.00000	⋯	-1.00000	1.00000	-1.00000	b
4	1.00000	1.00000	0.00000	0.00000	0.00000	⋯	1.00000	-1.00000	1.00000	b
5	-1.00000	1.00000	0.00000	0.00000	0.00000	⋯	1.00000	0.00000	0.00000	b
⋮	⋮	⋮	⋮	⋮	⋮	⋯	⋮	⋮	⋮	⋮
30	0.68729	1.00000	0.91973	-0.76087	0.81773	⋯	-0.06522	0.56522	0.23913	b
31	0.00000	0.00000	0.00000	0.00000	0.00000	⋯	0.00000	0.00000	0.00000	b
32	0.00000	0.00000	0.00000	0.00000	0.00000	⋯	0.00000	0.00000	0.00000	b

4.1 MTS 기법 적용 분석

본 연구에서의 사례 데이터인 Johns Hopkins University Ionosphere data에 대하여 먼저 MTS 기법을 적용하여 분석하였다. 분석 소프트웨어는 MINITAB 16을 사용하였다. MINITAB 16에는 MTS 분석 모듈이 없으므로 MD 계산을 위해서는 일련의 절차를 순차적으로 적용하여야 한다.

먼저 정상집단의 MD값을 구하기 위하여 <Table 4>의 정상집단의 데이터를 평균값을 빼고 표준편차로 나누어 표준화시켜 <Table 6>의 결과를 얻었다. 그리고 <Table 6>의 정상 집단의 표준화 데이터에 대한 상관행렬을 구한 결과가 <Table 7>에 주어져 있다.

그리고 식 (2)의 계산식에 의하여 정상 집단의 마할라노비스 거리(MD)를 계산한 결과가 <Table 8>에 나타나 있다. 계산된 정상 집단의 MD값들이 대체적으로 1부근의 값으로 구성되어 있음을 알 수 있다.

다음으로 비정상집단의 MD값을 구하기 위하여 <Table 5>의 비정상 집단의 데이터를 정상집단의 평균값을 빼고 표준편차로 나누어 표준화시키고, 비정상 집단의 마할라노비스 거리(MD)를 식 (4)를 이용하여 구한 결과가 <Table 9>와 같다. 계산된 비정상 집단의 MD값들이 정상 집단의 MD 값보다 훨씬 크게 나타났으므로, 정상집단의 MS 공간은 유효하다고 판단된다.

Table 6. Standardization of healthy group data

No.	Z1	Z2	Z3	Z4	Z5	…	Z30	Z31	Z32
1	0.74141	-0.28662	0.24926	-0.55373	0.52096	…	-1.06824	-0.42046	-0.94972
2	0.66305	-0.40417	0.60432	-1.15006	0.79631	…	-1.60922	-1.07432	-1.69552
3	0.75992	0.04433	0.80917	-0.52498	1.00687	…	0.68549	1.21773	0.97383
4	0.75992	-0.35708	0.80917	-1.05877	0.61570	…	-1.82718	-1.06443	-2.09324
5	0.75992	-0.53629	-0.24130	0.10129	0.52227	…	1.10649	1.21773	-1.21016
⋮	⋮	⋮	⋮	⋮	⋮	…	⋮	⋮	⋮
45	0.75992	2.35444	-1.59257	-0.45431	-1.53968	…	0.03533	-0.43824	1.66534
46	-1.93950	0.13567	-2.37911	0.14315	-1.50962	…	0.44451	-0.52647	0.54966
47	-2.47459	0.00305	-2.82577	0.03690	-2.22895	…	0.36898	-1.40294	-0.70487
48	-0.57884	-0.17381	0.71067	-0.43695	-0.46855	…	0.52617	-0.51524	1.00258

Table 7. Correlation matrix for healthy group data

	Z1	Z2	Z3	Z4	Z5	…	Z30	Z31	Z32
Z1	1	0.04124	0.68778	-0.00384	0.59987	…	-0.09944	0.38254	-0.00184
Z2	0.04124	1	-0.08943	0.33298	-0.21719	…	0.01581	-0.16276	0.25253
Z3	0.68778	-0.08943	1	-0.01604	0.74076	…	-0.08672	0.28679	-0.11306
Z4	-0.00384	0.33298	-0.01604	1	-0.19485	…	0.24807	-0.05416	0.25780
Z5	0.59987	-0.21719	0.74076	-0.19485	1	…	-0.10974	0.45373	-0.14973
⋮	⋮	⋮	⋮	⋮	⋮	…	⋮	⋮	⋮
Z30	-0.09944	0.01581	-0.08672	0.24807	-0.10974	…	1	0.01446	0.83352
Z31	0.38254	-0.16276	0.28679	-0.05416	0.45373	…	0.01446	1	-0.14637
Z32	-0.00184	0.25253	-0.11306	0.25780	-0.14973	…	0.83352	-0.14637	1

Table 8. MD values for healthy group data

0.55284	0.97011	0.35933	1.20755	1.31230	0.12435	1.17666	1.12109	1.27727	1.13340
1.10929	1.18669	1.32805	1.21893	0.75031	0.75230	0.78076	1.09667	1.24799	0.98338
0.95254	1.11828	0.98705	1.13244	1.34822	0.58055	1.33424	1.37487	1.02094	0.52136
0.93382	0.97566	1.22081	1.15027	1.15668	0.90579	0.54501	0.47266	0.76253	0.73279
0.96708	0.98129	0.75818	1.23499	1.18483	1.10546	0.94758	0.90479		

Table 9. MD values for unhealthy group data

35.280	6.822	166.854	120.288	119.151	10.435	11.005	106.484	118.581	9.887
2.665	13.394	40.279	16.145	10.654	22.527	47.840	30.447	4.077	76.161
63.294	128.812	123.512	126.497	16.756	55.159	26.388	17.377	25.518	2.856
44.250	28.164								

다음으로 유용한 변수를 선정하는 단계인데, 이를 위하여 적절한 2수준계 직교표를 선정하여 측정변수를 배치하여야 한다. 본 연구의 사례에서는 측정변수가 32개이므로 직교표 $L_{64}(2^{63})$에 배치하여야 한다. 이 경우 실험횟수가 너무 많아지는 관계로 직교표 $L_{32}(2^{31})$을 이용하기로 하였다. $L_{32}(2^{31})$에 배치하기 위해서는 영향이 없는 변수 1개를 미리 제거를 하여야 한다. 이를 위하여 각 변수를 1개씩 제거한 후에 비정상 집단의 MD값을 계산한 다음에 제거 전(모든 변수 사용)의 비정상 집단의 MD값과 비교하여 차이가 없는 변수를 삭제하기로 하였다. 각 변수를 1개씩 제거한 후에 비정상 집단의 MD값을 산출하여 모든 변수를 사용한 경우의 비정상 집단의 MD값과 쌍체 t-검정을 수행한 결과, X5에 대한 P-값이 0.998로 영향이 가장 작은 변수로 나와서 이를 제거하기로 하였다.

유용한 변수를 선별하기 위하여 X5를 제외한 31개의 변수를 $L_{32}(2^{31})$에 <Table 10>과 같이 배치하였다. 여기서 직교표의 수준 1은 해당 측정변수를 사용하는 경우를 의미하고, 수준 2는 해당 측정변수를 사용하지 않는 경우를 의미한다. 직교표 $L_{32}(2^{31})$ 각 행의 실험조건으로 비정상 집단의 데이터를 대상으로 하여 비정상 집단의 MD를 계산한 다음, 식 (8)의 망대특성의 SN비 공식을 이용하여 SN비값을 계산한다. 사례 데이터에 대하여 SN비값을 계산한 결과가 <Table 10>과 같다.

$$SN_i = -10\log\left\{\frac{1}{k}\sum_{j=1}^{k}\frac{1}{D_j^2}\right\} = -10\log\left\{\frac{1}{k}\sum_{j=1}^{k}\frac{1}{MD_j}\right\} \quad (8)$$

Table 10. $L_{32}(2^{31})$ array and SN ratios

No	X1	X2	X3	X4	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30	X31	X32	SN ratios
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**11.771**
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	**7.301**
3	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	**6.664**
4	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1	1	1	1	**7.641**
5	1	1	1	2	2	2	2	1	1	1	1	2	2	2	2	1	1	1	1	2	2	2	2	1	1	1	1	2	2	2	2	**8.263**
6	1	1	1	2	2	2	2	1	1	1	1	2	2	2	2	2	2	2	2	1	1	1	1	2	2	2	2	1	1	1	1	**6.914**
7	1	1	1	2	2	2	2	2	2	2	2	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	1	1	1	1	**5.781**
8	1	1	1	2	2	2	2	2	2	2	2	1	1	1	1	2	2	2	2	1	1	1	1	1	1	1	1	2	2	2	2	**4.328**
9	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2	**7.077**
10	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	1	**5.14**
11	1	2	2	1	1	2	2	2	2	1	1	2	2	1	1	1	1	2	2	1	1	2	2	2	2	1	1	2	2	1	1	**5.627**
12	1	2	2	1	1	2	2	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	1	1	1	2	2	1	1	2	2	**5.974**
13	1	2	2	2	2	1	1	1	1	2	2	2	2	1	1	1	1	2	2	2	2	1	1	1	1	2	2	2	2	1	1	**7.268**
14	1	2	2	2	2	1	1	1	1	2	2	2	2	1	1	2	2	1	1	1	1	2	2	2	2	1	1	1	1	2	2	**5.828**
15	1	2	2	2	2	1	1	2	2	1	1	1	1	2	2	1	1	2	2	2	2	1	1	2	2	1	1	1	1	2	2	**5.315**
16	1	2	2	2	2	1	1	2	2	1	1	1	1	2	2	2	2	1	1	1	1	2	2	1	1	2	2	2	2	1	1	**6.451**
17	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	**7.407**
18	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	**8.236**
19	2	1	2	1	2	1	2	2	1	2	1	2	1	2	1	1	2	1	2	1	2	1	2	2	1	2	1	2	1	2	1	**6.757**
20	2	1	2	1	2	1	2	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	1	2	1	2	1	2	1	2	**7.351**
21	2	1	2	2	1	2	1	1	2	1	2	2	1	2	1	1	2	1	2	2	1	2	1	1	2	1	2	2	1	2	1	**7.726**
22	2	1	2	2	1	2	1	1	2	1	2	2	1	2	1	2	1	2	1	1	2	1	2	2	1	2	1	1	2	1	2	**5.707**
23	2	1	2	2	1	2	1	2	1	2	1	1	2	1	2	1	2	1	2	2	1	2	1	2	1	2	1	1	2	1	2	**7.637**
24	2	1	2	2	1	2	1	2	1	2	1	1	2	1	2	2	1	2	1	1	2	1	2	1	2	1	2	2	1	2	1	**7.999**
25	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	**7.732**
26	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	2	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	**7.691**
27	2	2	1	1	2	2	1	2	1	1	2	2	1	1	2	1	2	2	1	1	2	2	1	2	1	1	2	2	1	1	2	**7.416**
28	2	2	1	1	2	2	1	2	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	1	2	2	1	1	2	2	1	**7.172**
29	2	2	1	2	1	1	2	1	2	2	1	2	1	1	2	1	2	2	1	2	1	1	2	1	2	2	1	2	1	1	2	**7.936**
30	2	2	1	2	1	1	2	1	2	2	1	2	1	1	2	2	1	1	2	1	2	2	1	2	1	1	2	1	2	2	1	**9.594**
31	2	2	1	2	1	1	2	2	1	1	2	1	2	2	1	1	2	2	1	2	1	1	2	2	1	1	2	1	2	2	1	**7.918**
32	2	2	1	2	1	1	2	2	1	1	2	1	2	2	1	2	1	1	2	1	2	2	1	1	2	2	1	2	1	1	2	**7.061**

Table 11. Average responses for SN ratios

	X1	X2	X3	X4	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17
Level 1	6.709	7.343	7.574	7.310	7.424	7.531	7.333	7.600	7.430	7.266	7.520	7.178	7.052	7.330	6.989	7.394
Level 2	7.584	6.950	6.719	6.983	6.870	6.762	6.960	6.693	6.863	7.027	6.773	7.115	7.241	6.963	7.304	6.899
Gain	-0.875	0.393	0.855	0.327	0.554	0.770	0.372	0.906	0.567	0.240	0.746	0.063	-0.190	0.366	-0.315	0.494
	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30	X31	X32	
Level 1	7.411	7.308	7.261	7.146	7.271	6.998	7.258	7.447	7.483	7.314	7.010	7.301	7.320	7.175	7.483	
Level 2	6.882	6.986	7.032	7.147	7.023	7.295	7.035	6.846	6.810	6.979	7.283	6.992	6.973	7.118	6.810	
Gain	0.529	0.322	0.228	-0.001	0.248	-0.298	0.223	0.602	0.673	0.335	-0.274	0.310	0.348	0.057	0.673	

Table 12. MD values for healthy group data using useful variables

0.39352	1.03290	0.36169	1.37403	1.56707	0.15102	0.77930	1.13045	1.26167	1.15847
1.13073	1.22036	1.33248	1.31832	0.82421	0.65658	0.91220	1.24394	1.36322	1.11302
0.49403	1.08543	1.04365	1.26667	1.65185	0.33186	1.72162	1.54934	0.50928	0.47697
1.00957	1.04022	1.10960	0.88811	1.38789	0.87608	0.51017	0.45384	0.78032	0.71971
0.95856	0.75557	0.52671	1.46348	1.24544	0.89203	1.03388	0.89295		

Table 13. MD values for unhealthy group data using useful variables

36.395	7.726	131.156	98.666	89.749	10.717	6.576	59.366	106.660	5.620
2.859	14.382	43.466	13.222	6.858	22.320	27.507	32.778	2.665	71.877
30.764	73.630	60.236	72.575	17.437	43.027	25.777	17.769	28.007	3.178
30.758	25.512								

SN비에 대한 수준별 평균값을 산출하고, 측정변수의 사용 이득(gain)을 계산한 결과가 <Table 11>에 주어져 있다.

<Table 11>로부터 측정변수의 사용 이득(gain)이 음수이거나 거의 차이가 없는 변수는 X1, X13, X14, X16, X21, X23, X28, X31로 파악되었고, 따라서 유용한 변수로 X2, X3, X4, X6, X7, X8, X9, X10, X11, X12, X15, X17, X18, X19, X20, X22, X24, X25, X26, X27, X29, X30, X32가 선정되었다. 선정된 유용한 변수만을 이용하여 계산한 정상집단의 MD값과 비정상집단의 MD값이 각각 <Table 12>와 <Table 13>에 주어져 있다.

다음으로 정상과 비정상을 구분하는 MD값의 임계값(threshold value)을 결정하기로 한다. 본 사례 연구에서는 판정오류를 고려하여 MD값의 임계값을 결정하는 방법을 사용하기로 한다. 이 방법은 임계값을 변화시켜 가면서 분류정확도를 검토하고, 가장 분류정확도가 높았을 때의 값을 임계값으로 정하는 것이다. 본 사례 연구에서는 MD값이 2를 기준으로 명확하게 정상과 비정상을 구분되므로 임계값을 2로 정하기로 하였다.

유용한 변수로 선정된 X2, X3, X4, X6, X7, X8, X9, X10, X11, X12, X15, X17, X18, X19, X20, X22, X24, X25, X26, X27, X29, X30, X32으로 구성된 Mahalanobis-Taguchi System(MTS)의 성능을 검증용 데이터(testing data)에 이용하여 분류정확도를 검증하기로 한다. 이 검증용 데이터는 정상그룹 데이터 177개, 비정상그룹 데이터 94개 총 271개로 구성되어 있다. 유용한 변수를 사용하여 검증용 데이터(testing data)에 대한 MD값을 계산한 결과의 그림이 <Figure 1>에 나타나 있다. 정상과 비정상을 구분하는 MD값을 2로 하여 분류정확도를 검토한 결과, 제 1종 오류(type I error)는 177개 중에서 1개로 나타나 오류율이 0.57%로 파악되었으며, 제 2종 오류(type II error)는 94개 중에서 11개로 나타나 오류율이 11.7%로 파악되었다. 따라서 전체 오분류 갯수는 총 271개 중에서 12개로 전체 오류율은 4.43%이며, 전체 분류정확도는 95.57%로 파악되었다.

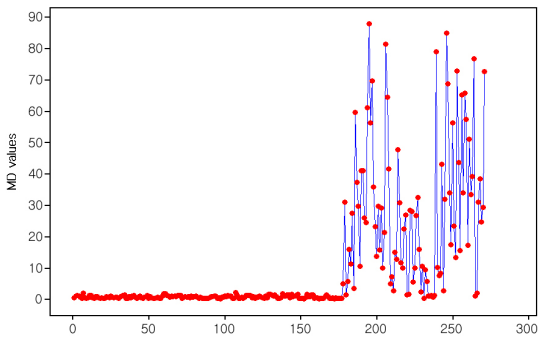

Figure 1. MD graph for testing data

참고로 X1-X32의 모든 변수를 사용한 MTS에서의 분류정확도를 검토한 결과, 제 1종 오류(type I error)는 177개 중에서 5개로 나타나 오류율이 2.82%로 파악되었으며, 제 2종 오류(type II error)는 94개 중에서 10개로 나타나 오류율이 10.64%로 파악되었다. 따라서 전체 오분류 갯수는 총 271개 중에서 15개로 전체 오류율은 5.53%이며, 전체 분류정확도는 94.47%로 파악되어, 유용한 변수로만 구성된 MTS의 성능보다 조금 낮게 나타났다.

4.2 로지스틱 회귀 적용 분석

본 연구에서의 사례 데이터인 Johns Hopkins University Ionosphere data에 대하여 이항 로지스틱 회귀의 로짓모형을 적용하여 분석하였다. 분석 소프트웨어는 SPSS 20을 사용하였다. 먼저 80개의 훈련용 데이터를 이용하여 이항 로지스틱 회귀모형을 추정하였다. 변수 선택 방법은 전진선택법과 후진소거법을 각각 사용하였으며, 선택기준은 LR(우도비) 기준, Wald 기준을 사용하였다. 기본적으로 변수진입 및 소거 기준은 각각 5%와 10%를 사용하였으나, 전진선택법에서는 선택된 변수가 너무 작은 관계로 추가로 진입기준을 10%로 늘려서 분석하여 보았다. 훈련용 데이터를 사용하여 분석한 이항 로지스틱 회귀분석의 결과가 <Table 14>에 나타나 있다.

훈련용 데이터에 의하여 구축된 이항 로지스틱 회귀모형을 이용하여 총 271개의 검증용 데이터를 대상으로 하여 분류 정확도를 파악한 결과가 <Table 15>에 주어져 있다. 결과를 살펴보면, 전진선택법에 의한 로지스틱 회귀모형의 분류정확도는 각각 83.8%와 84.5%로 나타났으며, 후진소거법에 의한 로지스틱 회귀모형의 분류정확도는 각각 78.6%와 80.8%로 나타났다. 그리고 X1-X32의 모든 변수를 다 포함하는 로지스틱 회귀모형의 분류정확도는 82.3%로 파악되었다.

본 사례 연구에서 MTS와 로지스틱 회귀를 적용한 성능 분석 결과를 요약하면 <Table 16>과 같다. 유용한 변수로 구성된 MTS의 분류 정확도가 95.57%, 모든 변수를 다 사용한 MTS가 94.47%로 나타났으며, 로지스틱 회귀에서는 가장 좋은 분류정확도를 준 경우가 84.50%로 나타났다. 따라서 본 사례의 경우 분류정확도 측면에서 MTS가 로지스틱 회귀보다 우수한 것으로 나타났다.

Table 14. Binary logistic regression analysis for training data

Methods	Critera	Selected variables
Forward selection	LR and Wald(enter = 0.05, leave = 0.1)	X1 X3 X16 X20
	LR and Wald(enter = 0.1, leave = 0.1)	X1 X3 X6 X14 X16 X20 X21 X32
Backward elimination	LR(enter = 0.05, leave = 0.1)	X2 X3 X5 X6 X7 X9 X10 X11 X12 X13 X14 X15 X16 X17 X18 X19 X21 X22 X24 X25 X26 X27 X28 X29 X32
	Wald(enter = 0.05, leave = 0.1)	X1 X3 X6 X14 X22 X24 X28 X29 X32

Table 15. Classification accuracy of binary logistic regression models for testing data

Models	Accuracy
$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -2.304+2.994x_1+2.575x_3+1.728x_{16}-3.250x_{20}$	83.8% (227/271)
$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -3.906+4.161x_1+3.808x_3+5.026x_6-3.105x_{14}+4.207x_{16}-2.949x_{20}+2.547x_{21}-5.648x_{32}$	84.5% (229/271)
$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -6.123-35.082x_2+84.711x_3-75.358x_5+60.066x_6+61.674x_7-70.897x_9+32.077x_{10}-82.804x_{11}+26.923x_{12}+104.300x_{13}-33.305x_{14}-20.032x_{15}+48.046x_{16}-27.284x_{17}-47.898x_{18}+52.415x_{19}+10.319x_{21}-54.787x_{22}+22.088x_{24}-41.242x_{25}-27.947x_{26}+39.760x_{27}+107.583x_{28}+28.350x_{29}-98.013x_{32}$	78.6% (213/271)
$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -8.284+5.858x_1+7.631x_3+11.351x_6-9.428x_{14}-5.319x_{22}+6.558x_{24}+6.985x_{28}+5.662x_{29}-10.310x_{32}$	80.8% (219/271)
$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -14.234+9.758x_1-33.764x_2+54.236x_3-16.717x_4-55.436x_5+\cdots+103.048x_{28}+39.178x_{29}+37.106x_{30}+5.182x_{31}-94.970x_{32}$	82.3% (223/271)

Table 16. Classification accuracy of MTS and binary logistic regression models

Models	Type I error	Type II error	Accuracy
MTS using useful variables	0.57% (1/177)	11.70% (11/94)	95.57% (259/271)
MTS using all variables	2.82% (5/177)	10.64% (10/94)	94.47% (256/271)
Logistic Regression(best model)	3.34% (6/177)	38.30% (36/94)	84.50% (229/271)

참고로 본 연구의 데이터를 최초로 사용한 연구인 Sigillito *et al*.(1989)에서는 신경망기법을 적용하여 분석하였는데, 훈련용 데이터로 200개의 데이터를 사용하였으며, 신경망기법의 선형 퍼셉트론 절차를 사용하였을 때 분류정확도가 90.7%, 비선형 퍼셉트론 절차를 사용하였을 때 분류정확도가 92%로 나타났었다.

5. 결론 및 토의

MTS 기법은 적용상의 한계점을 갖고 있지만 설비의 이상진단, 음성 패턴 인식, 문자 인식, 위조지폐 판별, 병의 진단, 계측 시스템 개발, 제품검사 등등의 분야에서 많은 적용 사례가 연구되고 있다. 본 연구에서는 MTS 기법 및 로지스틱 회귀의 절차에 대하여 살펴보고, 사례 데이터(Johns Hopkins University Ionosphere data)에 대하여 MTS 기법을 적용하여 분류 정확도를 파악하고, 로지스틱 회귀를 적용하여 분류 정확도를 파악하여, 분류 정확도의 관점에서 두 기법의 성능을 비교하였다.

본 연구의 총 351개 사례 데이터 중에서 약 20%(good 48개, bad 32개 총 80개)를 훈련용 데이터(training data)로 랜덤하게 선정하였고 나머지 약 80%(good 177개, bad 94개 총 271개)를 검증용 데이터(testing data)로 하여 분류 성능을 조사하였다. MTS와 로지스틱 회귀를 각각 적용하여 분석한 결과, MTS에서는 분류정확도가 95.57%로 파악되었고, 로지스틱 회귀에서는 가장 좋은 경우의 분류정확도가 84.5%로 파악되었다. 따라서 분류정확도 측면에서 MTS가 로지스틱 회귀보다 우수한 것으로 나타났다. 이는 훈련용 데이터의 수와 관련이 있는 것으로 판단된다. 훈련용 데이터의 수를 늘리면 로지스틱 회귀의 성능도 좋아질 것으로 판단된다. 그리고 로지스틱 회귀에서는 비정상집단 데이터의 영향이 더 큰 것으로 생각된다. 그러나 MTS에서는 정상집단의 데이터를 더 중요시 한다. Taguchi는 정상집단의 데이터를 대상으로 마할라노비스 공간(Mahalanobis Space)를 구성하여 비정상집단을 구분하라고 권고하였다. 그 이유는 정상집단은 그 패턴이 단순하고 균일하지만, 비정상집단은 그 원인이 다양하고 패턴 양상도 매우 다양하여 그 범위를 어디까지 인지 정하기 어려움이 있기 때문이다. 따라서 비정상집단의 원인을 밝히고 조치를 취하기 위해서는 비정상집단을 연구하여야 하지만, 올바른 판정을 위해서는 비정상집단을 연구하기 보다는 정상집단을 연구하는 것이 효율적이며, 이것이 MTS의 기본 철학이다.

MTS에서는 정상과 비정상을 구분하는 MD값의 임계값(threshold value)을 결정하는 것이 매우 중요하다. MTS에서 정상과 비정상을 구분하는 MD값의 임계값을 결정하는 방법은 본 연구에서와 같이 판정오류(Type I error, Type II error)를 고려하여 결정하기도 하지만, Taguchi는 다음과 같이 손실함수를 이용하는 절차를 제안하였다(Taguchi *et al*., 2001; 43-44).

$$\tau = \sqrt{A/A_0} \cdot D$$

여기서 τ는 정상과 비정상을 구분하는 MD값의 임계값이며, A는 이상유무를 감지하기 위한 완전검사(complete examination) 비용, A_0는 완전검사를 행하지 않았을 때 야기되는 손실, D는 자각 증상(subjective symptoms)을 갖는 비정상그룹의 MD의 중앙값(mid-value)이다. 그러나 자각 증상을 갖는 비정상그룹은 확실히 정의되지 않고 모호하여, 전문가의 지식에 크게 의존하게 되는 문제점이 있다(Woodall *et al*., 2003). 이러한 문제점을 해결하기 위한 한가지 방법으로 Kim *et al*.(2009)는 Hotelling의 T^2 관리도를 이용하는 절차, 즉 정상집단의 데이터를 이용하여 T^2 관리도의 관리한계선을 설정하고, 이를 임계값으로 대신하여 사용하는 절차에 대하여 논의하였다.

그리고 MTS에서는 유용한 변수를 선정하는 것이 차원축소와 분류정확도의 측면에서 중요하다. 이를 위하여 MTS에서는 2수준계 직교표를 선정하여 측정변수를 배치한 후에, SN비를 계산하여 SN비에 대한 수준별 평균값을 산출하고, 측정변수의 사용 이득(gain)을 파악하여 유용한 측정변수를 선정하는 절차를 사용한다. 이 방법이외에 Kim *et al*.(2009)은 직교표의 결과 중 SN비가 가장 좋은 조합으로 변수를 선정하는 방법, 측정변수의 모든 조합에 대한 SN를 출력하여 SN비가 가장 좋은 조합으로 변수를 선정하는 방법, 분류나무를 이용하는 방법, 주성분 분석을 이용하는 방법 등을 제안하여 비교 연구를 수행하였다. 이 이외에 MTS 기법의 한계점 및 비평에 대해서는 Woodall *et al*.(2003)과 Kim *et al*.(2009)의 연구를 참조하기 바란다.

본 연구에서는 사례 데이터 한 개에 대하여 분석한 결과이어서 한계점이 있을 것으로 판단된다. 따라서 많은 사례 데이터를 분석하여 종합적인 결론을 도출하여야 할 것으로 생각된다. 추후 연구로는 반응변수의 범주가 다항인 경우에 대하여 MTS와 로지스틱 회귀의 성능을 비교하고자 한다. 이 경우 MTS에서는 기준 집단을 정상집단으로 간주하고, 나머지 집단을 비정상집단으로 간주하여 MD값을 구하고, 유용한 변수를 선정하기 위한 SN비는 동특성의 SN비를 사용하여 한다. 로지스틱 회귀는 범주에 순서가 없는 경우는 명목형 로지스틱 회귀, 범주에 순서가 있는 경우는 순서형 로지스틱 회귀를 사용하여야 한다.

참고문헌

Cudney, E., Hong, J., Jugulum, R., Paryani, K., Ragsdell, K., and Taguchi, G. (2007), An Evaluation of Mahalanobis-Taguchi System and Neural Network for Multivariate Pattern Recognition, *Journal of Industrial and Systems Engineering*, **1**(2), 139-150.

Cudney, E., Hong, J., Drain, D., Paryani, K., Ragsdell, K., and Sharma, N. (2009), A Comparison of the Mahalanobis-Taguchi System to A Standard Statistical Method for Defect Detection, *Journal of Industrial and Systems Engineering*, **2**(4), 250-258.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*, 2nd

ed., Springer.
Hosmer, D.W. and Lemeshow, S. (2000), *Applied Logistic Regression*, 2nd ed, John Wiley and Sons, Inc.
Izenman, A. J. (2008), *Modern Multivariate Statistical Techniques : Regression, Classification and Manifold Learning*, Springer.
Johnson, R. A. and Wichern, D. W. (1992), *Applied Multivariate Statistical Analysis*, Englewood Ciffs, Prentice Hall.
Jugulum, R. and Monplaisir, L. (2002), Comparison between Mahalanobis-Taguchi-System and Artificial Neural Networks, *Journal of Quality Engineering Society*, **10**(1), 60-73.
Kim, S. B., Tsui, K.-L., Sukchotrat, T., and Chen, V. (2009), A Comparison Study and Discussion of the Mahalanobis-Taguchi system. *International Journal of Industrial and Systems Engineering*, **4**(6), 631-644.
Sigillito, V. G., Wing, S. P., Hutton, L. V., and Baker, K. B. (1989), Classification of Radar Returns from the Ionosphere using Neural Networks, *Johns Hopkins APL Technical Digest*, **10**, 262-266.
Taguchi, G., Chowdury, S., and Wu, Y. (2001), The Mahalanobis Taguchi System, McGraw Hill, New York.
Taguchi, G. and Jugulum, R. (2000), New Trends in Multivariate Diagnosis, *Indian Journal of Statistics*, **62**, Series B, 233-248.
Taguchi, G. and Jugulum, R. (2002), *The Mahalanobis-Taguchi Strategy : A pattern technology system*, John Wiley and Sons.
Tan, P.-N., Steinbach, M., and Kumar. V. (2006), *Introduction to Data Mining*, Addison-Wesley.
Vardhan, R. V., Sukanya, D. J. V., and Arthanari, T. S. (2012), Criteria of Classification and Measures of Performance, *International Journal of Advance Mathemetics and Mathematical Stiences*, **1**(1), 41-48.
Wang, H.-C., Chiu, C.-C., and Su, C.-T. (2004), Data Classification using the Mahalanobis-Taguchi System, *Journal of the Chinese Institute of Industrial Engineers*, **21**(6), 606-618.
Woodall, W. H., Koulelik, R., Tsui, K. L., Kim, S. B., Stoumbos, Z. G., and Carvounis, C. P. (2003), A Review and Analysis of the Mahalanobis Taguchi System, *Technometrics*, **45**(1), 1-30.