Journal of the Society of Korea Industrial and Systems Engineering
Vol. 32, No. 1, pp.20－25, March 2009.

Mahalanobis Taguchi System을 이용한 다변량 시스템의 해석에 관한 연구

홍정의* · 권홍규**†

*충주대학교 공과대학 산업경영공학과
**남서울대학교 산업경영공학과

Analysis of Multivariate System Using Mahalanobis Taguchi System

Jung-eui Hong* · Hong-kyu Kwon**†

*Department of Industrial and Management Engineering, Chungju National University
**Department of Industrial and Management Engineering, Namseoul University

Mahalanobis Taguchi System (MTS) is a pattern information technology, which has been used in different diagnostic applications to make quantitative decisions by constructing a multivariate measurement scale using data analytic methods without any assumption regarding statistical distribution. The MTS performs Taguchi's fractional factorial design based on the Mahahlanobis Distance (MS) as a performance metric. In this work, MTS is used for analyzing Wisconsin Breast Cancer data which has ten attributes. Ten different tests are conducted for the data to determine if the patient has cancer or not. Also, MTS is used for reducing the number of test to define the relationship between each attribute and diagnosis result. The accuracy of diagnosis is compare with two different previous research.

Keywords : MTS(Mahalanobis Taguchi System), Multivariate System, Medical Diagnosis

1. 서 론

대부분의 다 변량 시스템은 변수들의 변화에 따른 측정값의 변화를 얼마나 정확히 예측하는가를 그 분석의 기본으로 삼고 있다. 그러나 일반적으로 이러한 해석을 위해서는 많은 양의 데이터가 필요로 한다. 마할라노비스 거리(Mahalanobis Distance, MD)는 인도의 수학자 Mahalanobis에 의해 한 집단에서 이질의 집단을 구분하는 방법으로 1930년대에 소개되어 졌다. 강건 설계 방법을 고안해낸 다구찌는 어떤 집단의 평균값을 기초로 한 마할라노비스 공간(Mahalanobis Space, MS)를 설정하고 이를 기초로, 새로운 관측값이 이러한 공간으로부터 얼마나 벗어나 있는가를 측정하는 마할라노비스 다구찌 시스템(Mahalanonis Taguchi System, MTS) 방법을 고안해냈다[1]. MTS에서는 다차원의 단위 공간으로서 MS을 정의 하고 임의의 대상이 그 공간으로부터 얼마만큼 떨어져 있는가를 나타내는 방법으로 MD를 이용한다. 또한 설정된 측정공간으로부터의 진단의 정확도를 평가하기 위하여 다구찌 방법이 사용되어 진다. 이러한 MTS 방법의 장점은 다변량 함수해석에 매우 중요한 변수들간의 상관관계를 고려한다는 것이다[2].

이 연구의 목표는 MTS 방법을 의학적 검사 데이터의 해석에 적용하여 검사 항목과 진단과의 관계를 규명하여 불필요한 검사에 따른 시간적, 비용적 비효율요소

논문접수일 : 2008년 05월 21일　논문수정일 : 2008년 10월 06일　게재확정일 : 2008년 11월 03일
† 교신저자 hongkyuk@nsu.ac.kr
※ 이 논문은 충주대학교 대학구조개혁지원사업비(교육과학기술부 지원)의 지원을 받아 수행한 연구임.

를 배제하고 진단의 정확성을 높이는데 활용하고자 한다. 이를 통하여 보다 정확한 진단 방법을 제시 할 수 있을 뿐 아니라 환자의 고통 및 경제적 비용절감을 가능하게 할 수 있을 것이다[3].

2. Mahalanobis Taguchi System

MTS기법을 이용한 연구는 생물학분야[4], 최적의 생산조건 설정을 통한 생산성 향상[5], 반도체 공정의 최적화[6] 그리고 의학연구[7]등의 다양한 분야에서 이미 많은 연구가 진행되어 왔고 성공적으로 활용되어지고 있다.

MTS 기법은 다차원의 공간으로부터 MS를 정의 하고 임의의 측정대상이 그 공간으로부터 얼마만큼 떨어져 있는가를 데이터 해석학적인 방법으로 분석해 내는 해석 방법이다. MTS 방법을 이용한 정확한 예측을 위해서는 다차원의 공간을 대표하는 단위공간을 설정하는 것이다. 다차원 공간에서의 관측된 패턴은 변수들 간의 상호 상관관계에 영향을 받으며 종종 이러한 상관관계를 무시하고 독립된 변수로만 생각하여 잘못된 해석을 내리는 경우가 많이 발생하고 있다.

MTS에서 MS는 정상 또는 건강한 그룹의 표준화된 변수들을 이용하여 구할 수 있으며 이를 이용하여 건강한 그룹과 건강하지 않은 그룹을 구분하는 지표로 삼을 수 있다[8]. MS가 구해지면 변수들 중 측정치에 영향을 미치는 정도를 판단하기 위하여 SN비(Signal to Noise Ratio)와 OA(Orthogonal Array)를 이용한다. MTS의 일반적인 적용절차는

첫째, 표준이 되는 집단으로부터 판단에 적용될 변수들을 선정한다. 여기서 선정된 변수들에 의해 구성된 MS를 구성한다. 이를 위하여 정상 또는 건강한 그룹의 데이터가 사용되며 계산된 마할라노비스 거리의 평균값은 1에 근접한다.

$$z_i = \frac{X_i - m}{\sigma} \tag{1}$$

여기서 m은 변수의 평균 값이고, σ는 표준편차 그리고 X_i는 임의 측정 값이다. 다차원 공간에서의 마할라노비스 거리는 변수들 간의 상관관계를 계산함으로써 구할 수 있다. 이러한 MD의 통계학적인 의미는 임의의 측정값이 선택 집단의 중간 값으로부터 얼마나 근접해 있는가를 의미한다. 아래의 공식은 마할라노비스 거리를 계산하는 공식이다.

$$MD_i = D_i^2 = \frac{1}{k} Z_{ij}^T C^{-1} Z_{ij} \tag{2}$$

여기서 C^{-1}은 변수들 간의 상관계수를 포함한 상관행렬의 역행렬이고 는 표준 벡터의 transpose 벡터이다. 선택된 MS 공간에서 구해진 MD 값의 평균값은 대략 1에 근접한다. 따라서 이러한 MS 공간을 단위공간이라고 부른다.

두 번째 과정은 이렇게 구해진 MS 공간의 유효성을 판단한다. 이를 위해서 MS 공간 밖의 측정값 즉 비정상 또는 건강하지 않은 데이터를 이용한다. 이러한 집단의 MD를 구하기 위해서 정상 또는 건강한 집단으로 단위 MS를 구성하는 평균, 표준편차 그리고 상관행렬을 이용한다. 단위 공간이 유효하다면 비정상 또는 건강하지 않은 집단의 MD 값은 정상 또는 건강한 집단의 MD 값보다 훨씬 커서 구별이 뚜렷할 것이다.

세 번째 과정은 변수 중에서 측정값에 영향을 미치지 않거나 적게 미치는 변수를 찾아내서 제거 하여 시스템의 해석을 쉽게 하는 일이다. 이러한 목적을 위해서는 직교 배열표와 SN(Signal to Noise)비가 유용하게 활용되어 질수 있다. 직교배열표의 열은 실험변수들을 배열하였고 행은 실험의 조합을 나타낸다. 즉 직교 배열표의 Level 1은 변수를 사용하는 경우를 의미 하고 Level 2는 변수를 사용하지 않는 경우를 의미 한다. 따라서 직교배열표의 조건에 따라 변수들은 사용되어지거나 무시될 수 있으며 이를 바탕으로 SN비를 계산 할 수 있다. 망대 특성의 SN비를 구하는 식은

$$SN\text{비} = \eta = -10\log\left(\frac{1}{t}\sum_{i=1}^{t}\frac{1}{D_i^2}\right) \tag{3}$$

여기서 η는 SN ratio 이고
t는 비정상 그룹의 개수이며,
D_i^2는 i번째의 마할라노비스 거리(MD)이다.

3. MTS를 이용한 다변량시스템의 해석-위스콘신 유방암 진단 데이터이용

본 연구를 위해 University of California at Irvine의 Machining Learning Repository의 Wisconsin Breast Cancer 데이터를 이용하였다. Wisconsin Breast Cancer 데이터는 위스콘신 대학병원의 W. H. Wolberg에 의해 환자의 조직검사를 통해 얻은 의학적 데이터를 통해 9개의 의학적 검사 항목을 기본으로 유방암을 진단한 결과를 정리하여 놓은 데이터이다. 데이터는 699명의 환자의 자료를 가지고 있으나 16명의 자료는 일부데이터가 손실되어 있다. 따라서 본 연구의 목적상 손실된 데이터는 제거하여 활용 하였다. 본 데이터를 이용하여 많은 연구가 이루어져 왔고 이를 통한 진단의 정확성은 대략 90%초반

의 정확성을 보이고 있다.

<Table 1> Wisconsin Breast Cancer의 검사 항목과 Data 형태

Attribute	Domain
1. Clump Thickness	1-10
2. Uniformity of Cell Size	1-10
3. Uniformity of Cell Shape	1-10
4. Marginal Adhesion	1-10
5. Single Epithelial Cell Size	1-10
6. Bare Nuclei	1-10
7. Bland Chromatin	1-10
8. Normal Nucleoli	1-10
9. Mitoses	1-10
10. Class	2 for Benign, 4 for Malignant

3.1 변수 데이터

Wisconsin Breast Cancer Data는 <표 1>과 같은 항목의 데이터로부터 측정된 결과를 분석하여 얻을 수 있다. 본 연구는 Matlab의 Random Number 생성기를 이용하여 모든 데이터 값을 가지고 있는 683개의 데이터 중에서 임의의 30데이터를 선택 하였다. 선택된 데이터 중에서 암으로 진단된 데이터(11데이터)와 암이 아닌 것으로 진단된 데이터(19데이터)분류하여 MTS의 정상 그룹과 비정상 그룹으로 활용하기로 하였다.

3.2 MTS기법의 전개

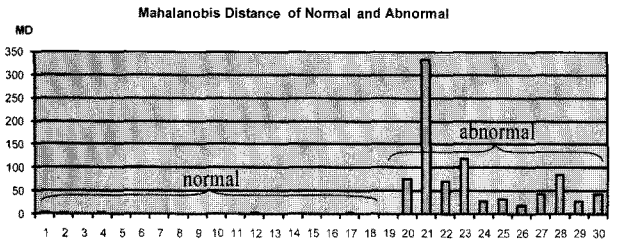
먼저 정상그룹의 데이터를 식 (1)을 이용하여 generalization과정을 수행하였다. 다음 relation Matrix와 Correla-

<Table 2> 정상그룹과 비정상그룹의 마할라노비스 거리

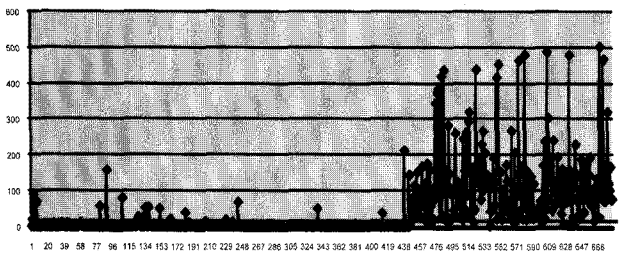
No	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class	MD
1	5	1	1	3	4	1	3	2	1	2	1.754
2	5	1	2	10	4	5	2	1	1	2	1.608
3	1	1	1	1	2	1	2	1	1	2	0.572
4	1	1	1	1	2	5	1	1	1	2	1.635
5	5	1	1	6	3	1	1	1	1	2	0.847
6	5	1	1	1	2	1	2	2	1	2	0.844
7	3	1	1	1	2	1	3	1	1	2	0.395
8	4	1	2	1	2	1	3	1	1	2	0.444
9	5	1	1	1	2	1	1	1	1	2	0.383
10	5	1	1	1	2	2	2	1	1	2	0.941
11	4	1	3	3	2	1	1	1	1	2	1.895
12	5	2	2	2	2	1	1	1	2	2	0.436
13	3	1	1	3	2	1	1	1	1	2	0.436
14	5	1	3	1	2	1	2	1	1	2	0.888
15	5	1	1	1	2	1	2	2	1	2	0.844
16	1	1	1	2	2	1	3	1	1	2	0.855
17	1	3	1	1	2	1	2	2	1	2	1.681
18	4	2	1	1	2	2	3	1	1	2	1.039
19	5	1	1	1	2	1	1	1	1	2	0.470
20	10	10	8	10	6	5	10	3	1	4	75.8
21	10	10	10	7	9	10	7	10	10	4	333
22	7	9	4	10	10	3	5	3	3	4	71.82
23	5	10	10	8	5	5	7	10	1	4	119.5
24	5	5	5	2	5	10	4	3	1	4	28.76
25	8	6	5	4	3	10	6	1	1	4	33.5
26	8	4	4	1	2	9	3	3	1	4	19.76
27	4	2	3	5	3	8	7	6	1	4	43.02
28	6	1	3	1	4	5	5	10	1	4	85.88
29	10	4	7	2	2	8	6	1	1	4	28.15
30	9	5	8	1	2	3	2	1	5	4	43.34

tion Matrix의 역행렬을 구하고 식 (2)를 이용하여 마할라노비스의 거리(MD)를 구한다. <표 2>는 정상그룹과 비정상그룹의 데이터 값과 이들 데이터의 마할라노비스 거리의 값을 나타내고 있다. <그림 1>은 정상그룹과 비정상 그룹의 MD의 크기를 보여주고 있다. 정상그룹(1번~19번)의 MD 값은 아주 작은 크기를 나타내고 있으며 비정상 그룹(20번~30번)의 MD 값은 정상그룹에 비해 상대적으로 아주 큰 값을 나타내고 있다. 따라서 임의로 선택된 정상그룹과 비정상 그룹은 효과적으로 암의 유무를 구별 해 낼 수 있다고 판단된다.

다음은 이미 선택된 정상그룹의 상관행렬과 표준편차 및 평균값을 이용하여 683개의 전체 데이터의 암 진단 결과를 실제 데이터와 비교하여 그 정확도를 비교 분석하고자 한다. 이러한 계산은 Matlab을 이용하여 계산되어 졌으며 그 진단 결과는 <그림 2>와 같다.



<그림 1> 정상그룹과 비정상그룹의 MD 값의 비교



<그림 2> 전체데이터의 MD 값

<그림 2>는 전체 683개 데이터의 MD 값을 계산하여 나타낸 그림이다. 진단의 정확도는 95.9%이고 19데이터의 type Ⅰ에러와 9데이터의 type Ⅱ에러를 나태내고 있다. Type Ⅰ에러(False Positive)는 암이 아닌 환자를 암인 환자로 진단한 경우이고 type Ⅱ에러(False negative)는 암인 환자를 정상으로 진단한 경우이다. Threshold 값의 결정은 type Ⅰ에러와 type Ⅱ에러의 어느 한 에러를 줄이는 것보다 이와 같은 에러에 의해 발생하는 사회적 손실을 최소화 하는 방향으로 결정되어 질 수 있다. 본 연구에서는 진단의 정확도를 최대화 하는 목적으로 threshold의 값을 결정 하였다.

3.3 샘플 데이터의 크기와 진단의 정확도와의 관계

본 연구는 샘플데이터의 크기가 진단의 정확도에 미치는 영향과의 관계를 정립하기 위하여 <표 3>과 같이 서로 다른 데이터의 크기를 이용하여 그 결과를 비교하였다. 즉 각각 30개, 40개 그리고 50개의 데이터를 이용하여 동일한 진단을 진행 하였고 그 진단의 정확도를 비교하였다. 일반적으로 데이터의 크기가 커질수록 진단의 정확도는 증가하는 결과를 얻을 수 있었다. 진단의 정확도는 결정된 threshold값을 이용하여 전체의 데이터를 진단하여 진단결과가 잘못된 경우의 비율을 나타낸다.

<표 3> 데이터의 크기와 진단의 정확도

Size of Data Set	Accuracy
30 data set	95.9%
40 data set	96.3%
50 data set	96.4%

4. Optimization

MTS의 마지막 단계로서 직교배열표와 SN비를 이용하여 각각의 검사항목과 이들이 진단에 미치는 영향과의 관계를 분석 한다. L12 직교 배열표를 이용하여 배열표상의 값이 1인 경우 해당 검사항목을 진단에 사용하고 값이 2인 경우 진단에 해당 검사항목을 사용하지 않는다. 즉 첫 번째 시뮬레이션의 경우 직교 배열표의 값이 모두 1이므로 모든 검사 항목 즉 9개의 검사 항목 모두를 진단에 활용하고 두 번째 시뮬레이션의 경우 처음 다섯 개의 검사 항목 값이 1이므로 이들 항목만 진단에 활용한다. 이와 같은 방법으로 12번의 시뮬레이션을 실행하고 식 (3)을 이용하여 각각의 SN 비를 구한다. SN 비를 계산하기 위해서는 비정상그룹의 데이터를 이용하는데 이는 비정상 데이터의 MD 값의 편차가 정상데이터의 그것보다 커서 SN비의 편차가 커지기 때문이다. <표 4>는 L12 직교 배열표와 SN비의 값을 보여주고 있다. <표 4>에서 각각의 검사항목별로 진단에 사용한 경우와 사용하지 않은 경우의 SN비의 합과 그 차이를 <표 5>와 <그림 3>에 나타내었다.

<그림 3>은 MTS를 이용한 최적치의 값을 보여주고 있다. 즉 검사 항목중에서 이득치의 값이 +인 6개의 검사항목 즉 Uniformity of Cell Size, Uniformity of Cell

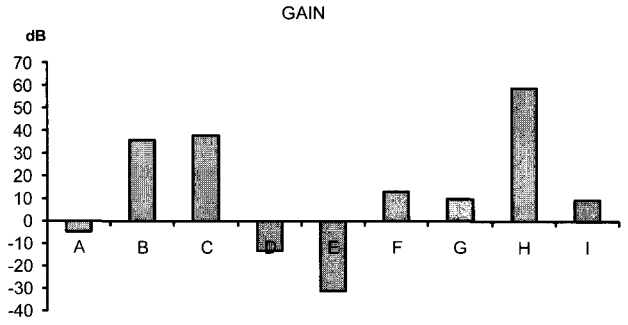
<표 4> L12 직교 배열표와 SN 비

	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses			SN Ratio
1	1	1	1	1	1	1	1	1	1	1	1	31.6
2	1	1	1	1	1	2	2	2	2	2	2	16.45
3	1	1	2	2	2	1	1	1	2	2	2	31.12
4	1	2	1	2	2	1	2	2	1	1	2	20.11
5	1	2	2	1	2	2	1	2	1	2	1	13.9
6	1	2	2	2	1	2	2	1	2	1	1	16.07
7	2	1	2	2	1	1	2	2	1	2	1	18.27
8	2	1	2	1	2	2	2	1	1	1	2	26.06
9	2	1	1	2	2	2	1	2	2	1	1	26.18
10	2	2	2	1	1	1	1	2	2	1	2	7.24
11	2	2	1	2	1	2	1	1	1	2	2	26.52
12	2	2	1	1	2	1	2	1	2	2	1	29.79

<표 5> SN 비의 수준평균 값과 and Gain

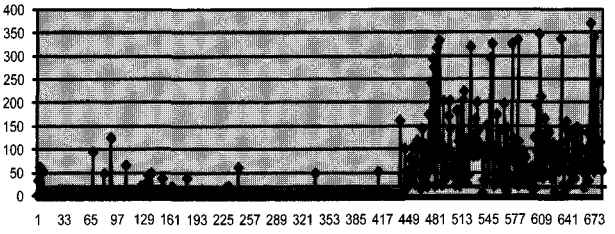
	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
Used	129.3	149.7	150.7	125.0	116.2	138.1	136.6	161.2	136.5
Unused	134.1	113.6	112.7	138.3	147.2	125.2	126.7	102.2	126.9
Gain	-4.8	36.04	38	-13.2	-31	12.96	9.82	59	9.6

Shape, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses은 진단에 영향을 미치는 것으로 판단 할 수 있고 이득치의 값이 마이너스인 Clump Thickness, Marginal Adhesion and Single Epithelial Cell Size는 상대적으로 진단에 영향 적게 미치는 것으로 판단 할 수 있다.



<그림 3> 검사 항목별 이득치

마지막으로 +이득치를 갖는 6개의 검사 항목만으로 암의 진단을 시도하여 그 결과치를 비교 하여 보았다. <그림 4>는 이 진단 결과를 보여준다. 시뮬레이션 결과 그 진단의 정확도는 95.61%이고 19개의 type Ⅰ에러와 11개의 type Ⅱ에러를 보여주고 있다.



<그림 4> 6개의 검사항목에 의한 진단

5. 결 론

다구찌 품질 공학의 기본개념은 복잡하고 시간이 많이 걸리는 다변량 시스템의 효과적인 해석에 있다. 다변량 시스템에서 multicolinearity와 partial correlation의 존재는 시스템해석을 복잡하게 하기도 하고 심지어는 해석 불가능하게 하기도 한다. 그러나 대부분의 경우, 이

러한 것들의 존재를 무시하고 그 대가로 많은 시간과 노력을 통해 시스템을 해석 하고자 한다. MTS 시스템에서 파라미터와 목표성능과의 관계는 이득치로서 측정되어 진다. 큰 값의 이득치는 그 파라미터가 시스템의 성능에 크게 영향을 미치는 것으로 판단 할 수 있다.

효과 적인 암의 진단방법을 제시하기 위한 이번 연구에서, 30명의 환자로부터 얻은 검사 데이터를 이용하여 95.9%의 진단 정확도를 얻을 수 있었다. 이 연구 결과는 Wolberg[9]와 Zhang[10]의 결과 보다 정확한 진단효과를 나타내고 있다<표 6>. 특히 본 연구는 아주 적은 환자의 데이터를 이용하여 높은 정확도를 얻을 수 있었다. 직교배열표와 SN비를 이용하여 파라미터와 진단과의 관계를 이득치로서 나타내는 최적화 과정을 통해 진단에 영향을 덜 미치는 것으로 판단되는 3개의 검사 항목을 제거하고 나머지 6개의 검사항목만으로 진단을 실시한 결과 진단정확도는 95.61%로 나타났다.

<Table 6> Compare to different approach

Researcher	No. of Test data	Accuracy
Wolberg[9]	369data set	93.5%
Zhang[10]	369data set	93.7%
This research	30data set	95.9%

따라서 암진 단을 위한 검사항목의 수를 9개 영역에서 6개영역으로 줄여도 그 진단의 정확도는 95.9%에서 95.61%로 0.29% 감소하는데 그쳤다. 본 연구는 MTS를 이용하여 다변량의 시스템을 해석하기 위한 새로운 대안을 제시 하고 있다. 첫째, 기존의 방법보다 적은수의 데이터를 이용하였으며 그 진단결과는 기존의 방법보다 우수 하였다. 둘째, 진단을 위한 파라미터의 수 즉 비용과 시간을 진단의 정확도 의 큰 손실 없이 현저히 줄일 수 있었다.

참고문헌

[1] Taguchi, G. and Jugulum, R.; "New Trends in Multivariate Diagnosis," *Indian Journal of Statistics*, 62, Series B, 2, 233-248, 2000.

[2] Jain, Anil K., Duin, Robert P. W., Mao, Jianchang; "Statistical Pattern Recognition : A Review," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 22(1) : Jan. 2000.

[3] Taguchi, S.; "Mahalanobis Taguchi System," *ASI Taguchi Symposium*, 2000.

[4] Lande, U.; "Mahalanobis Distance : A Theoretical and Practical Approach," http://biologi.uio.no/fellesavdelinger/finse/spatialstats/Mahalanobis%20distance.ppt, 2003.

[5] Hayashi, S., Tanaka, Y., and Kodama, E.; "A New Manufacturing Control System using Mahalanobis Distance for Maximizing Productivity," *IEEE Transactions*, 59-62, 2001.

[6] Asada, M.; "Wafer Yield Prediction by the Mahalanobis-Taguchi System," *IIE Transactions*, 25-28, 2001.

[7] Wu, Y.; "Pattern Recognition using Mahalanobis Distance," *TPD Symposium*, 1-14, 1996.

[8] Taguchi G., Chowdury, S., and Wu, Y.; The Mahalanobis Taguchi System, McGraw Hill Press New York, 2001.

[9] Wolberg, W. H. and Mangasarian O. L.; Multisurface method od pattern seperation for medical diagnosis applied to breast cytology, Porceeding of the national academy of Science, 9193-9196, 1990.

[10] Zhang, J.; Selecting typical instance in instance-based learning, Proceeding of ninth international machine learning conference, 470-479, 1992.