

신경망 이론과 Mahalanobis Distance 이상치 탐색방법을 이용한 고강도 콘크리트 강도 예측 모델 개발에 관한 연구

홍 정 의[†]

충주대학교 공과대학 산업경영공학과

Modeling of Strength of High Performance Concrete with Artificial Neural Network and Mahalanobis Distance Outlier Detection Method

Jungeui Hong[†]

Department of Industrial and Management Engineering, Chungju National University

High-performance concrete (HPC) is a new terminology used in concrete construction industry. Several studies have shown that concrete strength development is determined not only by the water-to-cement ratio but also influenced by the content of other concrete ingredients. HPC is a highly complex material, which makes modeling its behavior a very difficult task. This paper aimed at demonstrating the possibilities of adapting artificial neural network (ANN) to predict the compressive strength of HPC. Mahalanobis Distance (MD) outlier detection method used for the purpose increase prediction ability of ANN. The detailed procedure of calculating Mahalanobis Distance (MD) is described. The effects of outlier compared with before and after artificial neural network training. MD outlier detection method successfully removed existence of outlier and improved the neural network training and prediction performance.

Keywords : Mahalanobis Distance, Neural Network, Multiple Regression

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)는 인도의 수학자 Mahalanobis 에 의해 한 집단에서 이질의 집단을 구분하는 방법으로 1930년대에 소개되어 졌다. 강건설계 방법을 고안해낸

다구찌는 어떤집단의 평균값을 기초로한 MS를 설정하고 이를 기초로한 새로운 관측값이 이러한 공간으로부터 얼마나 벗어나 있는가를 측정하는 마할라노비스 MD 방법을 고안해 냈다. MD를 계산하기 위해서는 다차원의 단위 공간인 마할라노비스 공간(Mahalanobis Space)을 정의 하고 임의의 대상이 그 공간으로부터 얼마만큼 떨어져 있는가를 나타내는 방법으로 MD를 이용한다[1, 2].

콘크리트는 오늘날 건설 분야에서 가장 광범위하게 활용되고 있는 재료로서 시멘트, 물, 모래, 혼화제, 골재

등의 구성요소들의 긴밀한 상관관계에 의해 강도가 결정된다. 특히 고강도 콘크리트(High Performance Concrete, HPC)는 물, 시멘트, 모래 이외에 보조 첨가제로 Fly ash, Blast furnace slag 그리고 화학첨가제인 Superplasticizer 등이 첨가된다. 따라서 고강도 콘크리트의 강도특성 분석은 일반 콘크리트보다 해석이 더 어려울 수 밖에 없다 [3, 4].

고강도 콘크리트의 첨가제와 강도와와의 관계 모델을 정의 하기위해 UCI Machining Learning Repository[1]의 Data를 이용하여 최적의 강도예측모델을 개발 하고자 하였다. 먼저 MD를 이용하여 이상치를 제거 한후 이상치 제거 전후의 강도 예측 결과를 비교 하였다. 연구의 목적은 최소의 실험을 통한 예측모델의 개발을 하고자 한다. 따라서 전체의 실험 데이터 중에서 무작위로 선정된 일부 데이터를 신경망의 학습데이터로 활용하여 신경망 학습 후 예측결과를 비교 하였다. 또한 콘크리트 강도에 영향을 미치는 주요 변수들의 상호 관계를 비교하여 보았다.

2. Mahalanobis Distance

MTS 기법은 다차원의 공간으로부터 MS를 정의 하고 임의의 측정대상이 그 공간으로부터 얼마만큼 떨어져 있는가를 데이터 해석학적인 방법으로 분석해 내는 해석 방법이다[3]. 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 MS(기준그룹)은 정상 또는 건강한 그룹의 표준화된 변수들을 이용하여 구할 수 있으며 이를 이용하여 건강한 그룹과 그렇지 못한 그룹을 구분하는 지표로 삼을 수 있다. MS가 구해지면 변수들 중 측정치에 영향을 미치는 정도를 판단하기 위하여 S/N비(Signal to Noise Ratio)와 OA(Orthogonal Array)를 이용한다. MTS의 일반적인 적용절차는[5~7]

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.

$$z_i = \frac{X_i - m}{\sigma} \quad (1)$$

여기서 m 은 변수의 평균값이고, σ 는 표준편차 그리고 X_i 는 임의 측정값이다.

다차원 공간에서의 마할라노비스 거리는 변수들간의 상관관계를 계산함으로써 구할 수 있다. 이러한 MD의 통계학적인 의미는 임의의 측정값이 선택집단의 중간값으로부터 얼마나 근접해 있는가를 의미한다. 아래의 공식은 마할라노비스 거리를 계산하는 공식이다.

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

여기서 C^{-1} 은 변수들 간의 상관계수를 포함한 상관행렬의 역행렬이고

T는 표준 벡터의 transpose 벡터이다.

선택된 MS 공간에서 구해진 MD 값의 평균값은 대략 1에 근접한다. 따라서 이러한 MS 공간을 단위공간이라고 부른다.

3. MD를 이용한 이상치의 제거

실험을 통한 데이터는 실험의 오류, 측정오류 등의 결과에 따라 다른 데이터와 구별되는 결과를 얻을 수도 있다. 이상치는 대부분의 데이터와 다른 특성을 보이거나 특정 속성의 값이 다른 개체들과 달리 유별난 값을 가지는 데이터를 의미한다. 이러한 이상치는 데이터의 분석에 부정적인 영향을 미칠 수 있다. 따라서 이상치를 포함한 경우와 포함하지 않은 경우 각각의 신경망학습을 통한 예측의 결과를 비교하고자 한다.

3.1 변수 데이터

고강도 콘크리트 실험 데이터는 <표 1>과 같은 항목의 변수들의 값을 변화시키면서 실험을 통해 얻은 콘크리트의 강도 데이터이다. 원본 데이터는 9개의 요소와 103가지의 서로 다른 실험 결과가 포함되어 있다.

3.2 이상치 제거

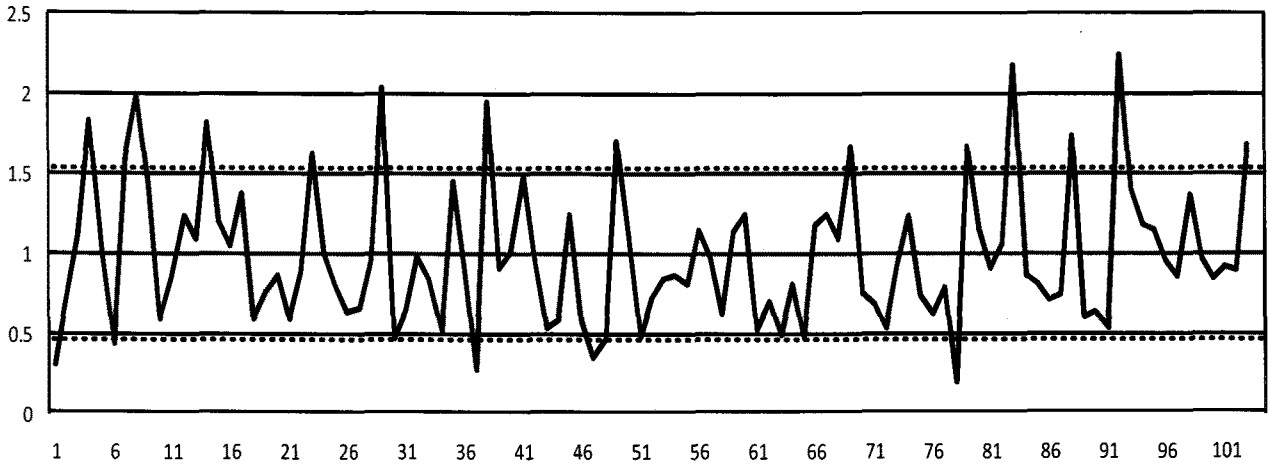
본 연구는 시멘트, Blast furnace slag, Fly ash, 물, Superplasticizer, Coarse Aggregate, Fine Aggregate, Slump, 그리고 Flow의 9개의 변수를 이용하여 MD 값을 계산하였고 MD 값이 1.5이상인 14개의 데이터와 0.5이하인 9개의 데이터를 이상치로 분류 하고 이상치를 제거한 80개의 데이터를 이상치를 포함하지 않은 집단으로 선정 하였다.

〈표 1〉 콘크리트 양생조건

변수	항목	Range of Component
A1	Cement	137~374(kg/m ³)
A2	Blast Furnace Slag	0~193(kg/m ³)
A3	Fly Ash	0~260(kg/m ³)
A4	Water	160~240(kg/m ³)
A5	Superplasticizer	4.4~19(kg/m ³)
A6	Coarse Aggregate	708~1050(kg/m ³)
A7	Fine Aggregate	640~902(kg/m ³)
A8	Slump	0~29(cm)
A9	Flow	20~78(cm)

〈표 2〉 고강도 콘크리트 강도(Sample)

	Cement	Blast Furnace Slag	Fly Ash	Water	Superplasticizer	Coarse Aggregate	Fine Aggregate	Slump	Flow	Strength
1	273	82	105	210	9	904	680	23	62	34.99
2	163	149	191	180	12	843	746	0	20	41.14
3	162	148	191	179	16	840	743	1	20	41.81
4	162	148	190	179	19	838	741	3	21.5	42.08
5	154	112	144	220	10	923	658	20	64	26.82
6	147	89	115	202	9	860	829	23	55	25.21
7	152	139	178	168	18	944	695	0	20	38.86
8	145	0	227	240	6	750	853	14.5	58.5	36.59
9	152	0	237	204	6	785	892	15.5	51	32.71
10	304	0	140	214	6	895	722	19	51	38.46
11	145	106	136	208	10	751	883	24.5	61	26.02
12	148	109	139	193	7	768	902	23.75	58	28.03
13	142	130	167	215	6	735	836	25.5	67	31.37
14	354	0	0	234	6	959	691	17	54	33.91
15	374	0	0	190	7	1013	730	14.5	42.5	32.44
16	159	116	149	175	15	953	720	23.5	54.5	34.05
17	153	0	239	200	6	1002	684	12	35	28.29
18	295	106	136	206	11	750	766	25	68.5	41.01
19	310	0	143	168	10	914	804	20.5	48.2	49.3
20	296	97	0	219	9	932	685	15	48.5	29.23
21	305	100	0	196	10	959	705	20	49	29.77
22	310	0	143	218	10	787	804	13	46	36.19
23	148	180	0	183	11	972	757	0	20	18.52
24	146	178	0	192	11	961	749	18	46	17.19
25	142	130	167	174	11	883	785	0	20	36.72
26	140	128	164	183	12	871	775	23.75	53	33.38
27	308	111	142	217	10	783	686	25	70	42.08
28	295	106	136	208	6	871	650	26.5	70	39.4
29	298	107	137	201	6	878	655	16	26	41.27
30	314	0	161	207	6	851	757	21.5	64	41.14
31	321	0	164	190	5	870	774	24	60	45.82
32	349	0	178	230	6	785	721	20	68.5	43.95
33	366	0	187	191	7	824	757	24.75	62.7	52.65
34	274	89	115	202	9	759	827	26.5	68	35.52
35	137	167	214	226	6	708	757	27.5	70	34.45



<그림 1> MD값과 이상치

4. 신경망을 이용한 고강도 콘크리트 강도 예측 모델

신경망 이론이 가지는 추론 기능을 이용하여 최소한의 데이터를 이용하여 예측모델의 개발 가능여부를 판단하고자 한다. 신경망 설계를 위하여 9개의 입력 Node, 한 개의 Hidden layer, 그리고 출력층에 한 개의 Node를 가진 역전파 신경망(Back-Propagation Neural Network)을 사용하였다. Hidden Layer node의 개수는 학습데이터의 수에 따라 전체 weight matrix의 인자의 수가 학습데이터의 개수를 넘지 않는 범위내에서 결정했다. 전이 함수로는 Sigmoid함수를 사용하였으며 이를 위해 입력 데이터의 Generalization을 다음과 같은 식을 이용하여 변환 하였다.

$$Tar = Tar_{min} + \frac{X - X_{min}}{X_{max} - X_{min}} \times (Tar_{max} - Tar_{min}) \quad (3)$$

먼저 이상치를 제거 하지 않은 103개의 데이터 중 랜덤함수를 이용하여 각각 30, 40, 50개의 데이터를 선택하여 각각 200,000회 이상의 Iteration중 최소의 SSE값을 가지는 Weight Matrix를 저장하고 이를 이용하여 103개의 데이터에 대한 강도를 예측하였다. <그림 2>은 학습 데이터의 수에 따른 예측값과 실제 강도와 의 그래프 이다. 예측된 강도 값을 실제의 데이터와 비교하기 위하여 상관분석을 이용하여 상관계수의 값으로 비교하였다. 상관계수는 예측값과 실제의 측정값과의 관계를 나타내는 척도이다. 상관계수는 -1과 1사이에 존재하고 상관계수의 값이 1에 가까울수록 예측값과 실제값이 일치함을 나타내고 -1인 경우 부정적인 상관관계를 나타낸다. 또한 상관계수가 0인 경우 두 값사이의 상관관계가 없음을 의미한다.

<표 3>은 이상치를 제거하지 않은 data중에서 임의로 학습을 위해 선택된 데이터를 가지고 학습과 예측을 통해 얻은 예측 값을 원 데이터와 비교하여 얻은 상관계수를 나타내고 있다.

<표 3> 이상치 제거전의 상관계수 비교

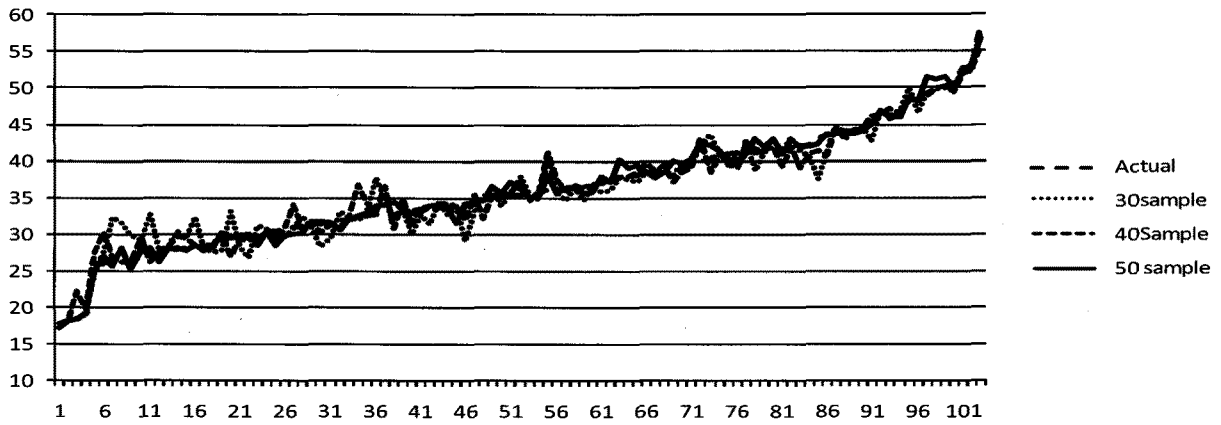
	전체 Data	30 Data	40 Data	50 Data
상관계수	1	0.967	0.982	0.993

<표 4>는 이상치를 제거한 후 data 중에서 이상치 제거 전과 마찬가지로의 과정을 통해 얻은 예측 값의 상관계수를 보여주고 있다.

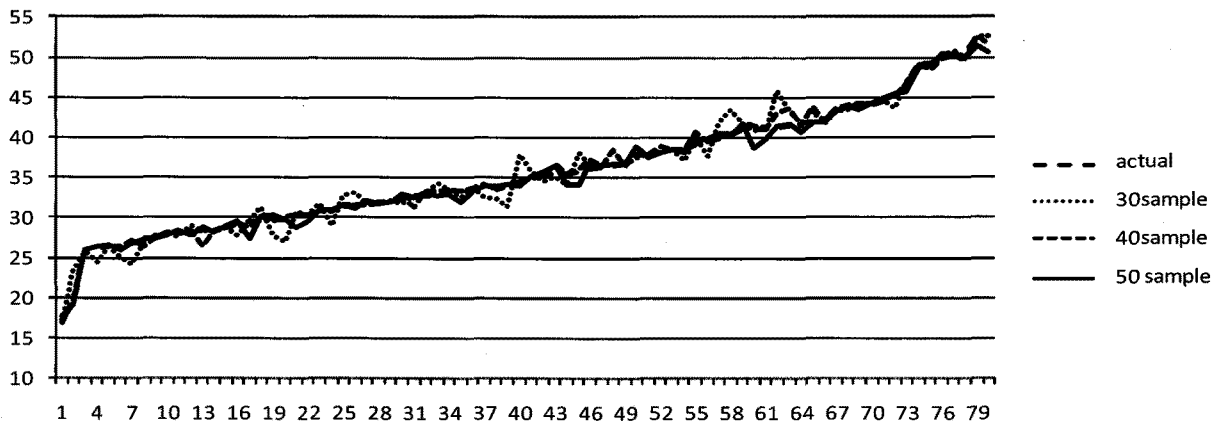
<표 4> 이상치 제거후의 상관계수 비교

	전체 Data	30 Data	40 Data	50Data
상관계수	1	0.984	0.996	0.996

<표 3>, <표 4>에서와 같이 이상치를 제거하지 않은 전체의 데이터 중 임의로 학습데이터를 선정하여 신경망을 학습한 후 강도값을 예측한 결과 학습데이터의 수가 20개 인 경우 상관계수의 값이 0.967로 나타났다. 이 결과는 이상치를 제거한 경우의 상관계수 0.984보다 상관도가 낮음을 알 수 있다. 즉 이상치를 제거한 경우 신경망 학습을 통한 강도예측의 결과가 이상치를 제거 하지 않은 경우 보다 높은 상관관계를 나타냄을 알 수 있다. <표 4>는 이상치를 제거한 후 임의로 데이터 개수를 변경하면서 실제의 강도값과 예측 값을 상관분석을 이용하여 상관계수로 나타내었다. 이상치를 제거한 데이터는 학습데이터의 수에 관계없이 거의 일정한 상관계수를 나타내고 있다.



<그림 2> 이상치 제거전의 예측 결과

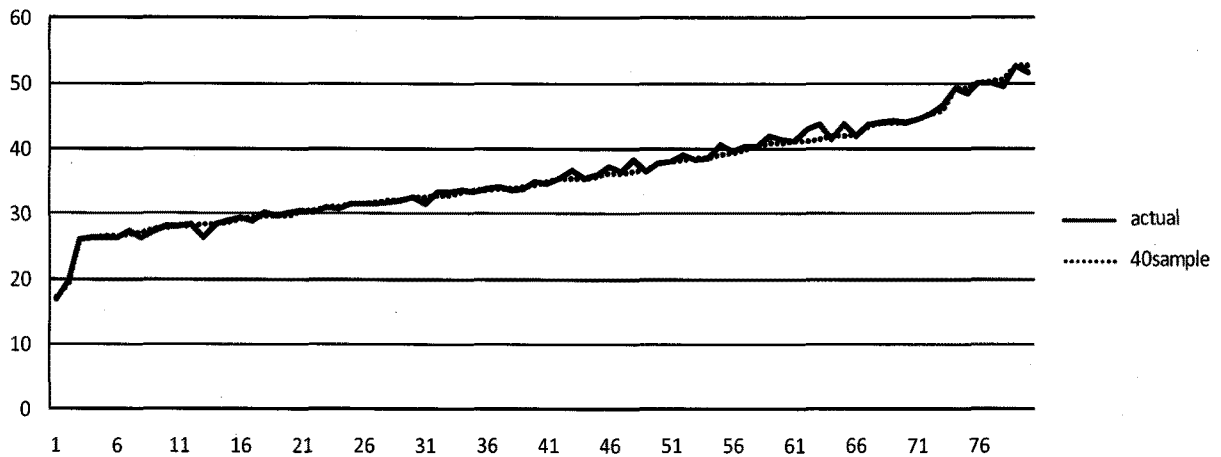


<그림 3> 이상치 제거후의 예측 결과

5. HPC 강도와 변수와의 관계

본 연구에서 가장 효율적인 신경망을 이용한 예측모델은 이상치를 제거한 후 40개의 학습데이터를 이용하여 학습한 경우 실제 측정 데이터와의 상관계수가 0.996으로

나타났다. 학습을 통해 얻은 신경망의 Weight Matrix를 이용하여 실제의 강도를 예측한 결과를 <그림 4>에 나타 내었다. 최고의 강도를 나타낸 실제 실험에 의한 변수들의 조건과 신경망을 이용한 실험 조건은 <표 5>, <표 6>과 같다.



<그림 4> 최적 예측 결과

<표 5> 측정된 최적강도 조건

case	Cement	Blast Furnace Slag	Fly Ash	Water	Superplasticizer	Coarse Aggregate	Fine Aggregate	Slump	Flow	Strength
50	326	0	167	174	6	884	792	21.5	42	52.65
33	366	0	187	191	7	824	757	24.75	62.7	52.65

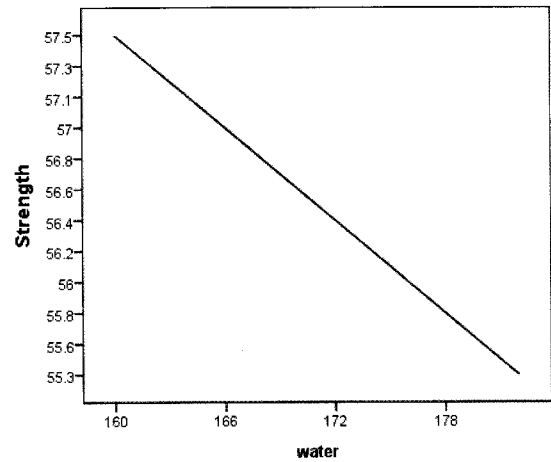
<표 6> 신경망을 이용한 예측된 최적 강도조건

case	Cement	Blast Furnace Slag	Fly Ash	Water	Superplasticizer	Coarse Aggregate	Fine Aggregate	Slump	Flow	Strength
1	366	0	211	190	5.86	932	738.25	21.75	49	54.92
2	366	0	211	180	5.86	932	735.25	21.75	49	54.88
3	366	0	211	170	5.86	932	738.25	21.75	49	54.7
4	366	0	211	165	5.86	932	735.28	21.75	49	54.55
5	366	0	211	160	5.86	932	735.25	21.75	49	54.36

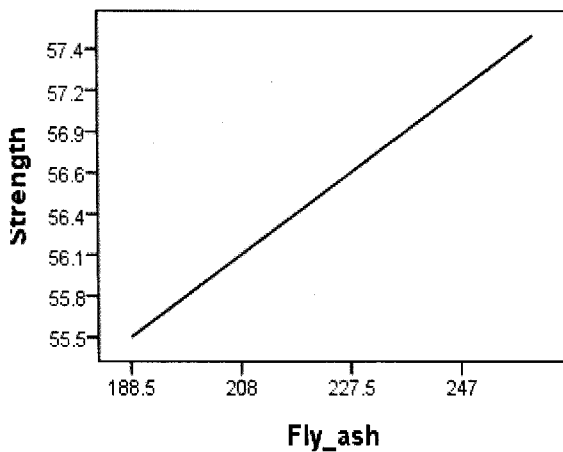
<표 6>은 최적의 콘크리트 강도 조건은 최대의 시멘트 양과 적절한 양의 첨가제를 배합했을 경우 얻을 수 있었으며 또한 이조건하에서 물의 양이 강도에 큰 영향을 미치는 것으로 나타났다.

<그림 5>~<그림 7>은 40개의 데이터를 이용하여 학습한 신경망의 Weight Matrix를 이용하여 콘크리트, Fly ash 그리고 물과 강도와의 관계를 예측하여 보았다. 예측을 위하여 다른 변수들의 값을 고정 값으로 하고 하나의 변수 값 만을 변화시켜 얻은 예측 값을 그래프로 나타 내었다.

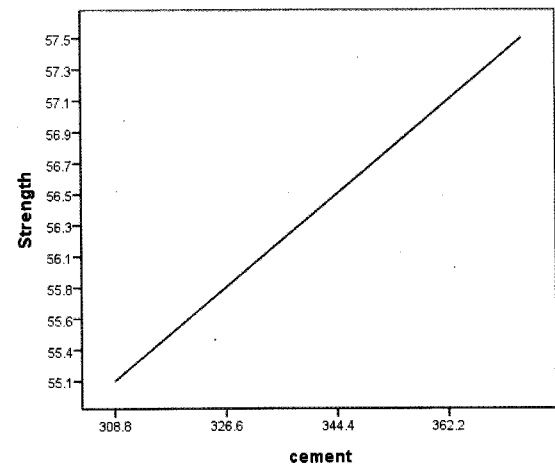
<그림 8>~<그림 10>콘크리트의 강도와 물, 시멘트 그리고 Fly ash와의 관계를 나타낸다. 시멘트와 Fly ash는 첨가량을 증가시킬수록 콘크리트의 강도를 증가시키는 것을 알 수 있고, 물의 양과 콘크리트 강도와는 반비례 관계임을 알 수 있다.



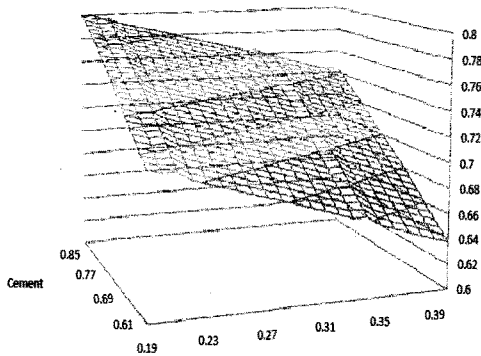
<그림 6> 물의 양과 강도



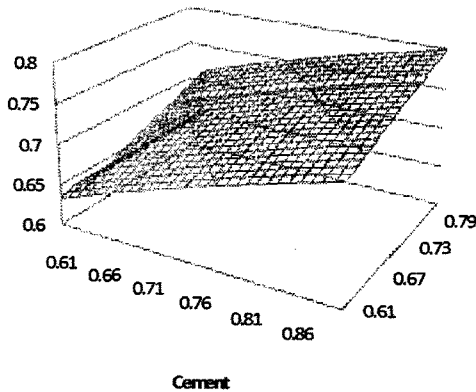
<그림 5> Fly ash 강도



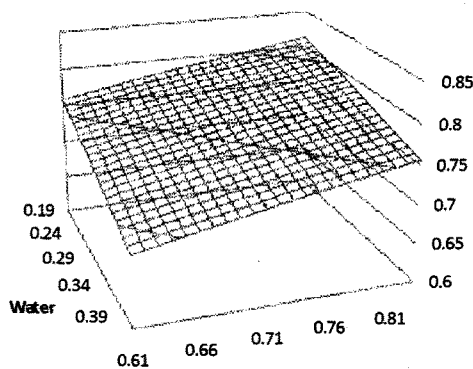
<그림 7> 시멘트양과 강도



<그림 8> Water vs Fly ash



<그림 9> Water vs Cement



<그림 10> Cement vs Fly ash

6. 결 론

다변량 시스템의 효과적인 해석은 실험데이터를 얻기 위한 시간과 비용을 효율적으로 절감 할 수 방법이 될 수 있다. 본 연구는 9개의 변수를 가지는 고강도 콘크리트 실험 데이터를 기반으로 가장 효과적인 강도 예측모델을 개발하고자 MD를 이용한 이상치 제거, 선형회귀분석 그

리고 신경망을 이용한 변수와 강도와의 관계를 예측하는 연구를 진행 하였다. 선형회귀분석을 이용한 예측 모델은 0.05의 유의 확률을 가진 변수만을 고려 할때 0.923의 R 값을 가지는 것으로 나타 났다.

신경망이론은 이용한 예측은 SSE를 줄여나가는 목적으로 Weight Matrix의 값을 변화 시켜나가는 트레이닝이라는 과정을 통해 목적 값에 가장 근접한 Weight Matrix를 저장하고 이를 이용하여 강도 값을 예측 할 수 있다. 본 연구에서 이상치를 제거한 데이터에서 40개의 학습 데이터를 이용하여 학습한 후 강도를 예측한 결과 0.996의 상관계수를 얻을 수 있었다.

본 연구는 현장에서 데이터의 수집에 많은 시간과 비용이 소요됨을 고려하여 최소한의 데이터를 이용하여 콘크리트의 강도 값을 예측하고자 하였다. 따라서 전체 데이터 중에서 임의의 정해진 학습데이터만 선정하고 이를 이용하여 신경망의 학습을 수행 한 후 전체 데이터에 대한 예측 실험을 반복하여 시행 하였다. 일반적으로 신경망을 이용한 예측은 신경망의 학습조건이외에도 선택된 학습데이터에 따라 그 결과가 달라 질 수 있다. MD를 이용한 이상치를 제거하기 전과후의 결과, 이상치를 제거 한 후의 예측은 적은양의 학습데이터임에도 불구하고 일정한 수준의 예측효과를 보여주고 있다. 이는 이상치를 제거한 데이터를 이용한 신경망을 이용한 강도 예측이 보다 높은 상관관계를 가지는 예측모델을 만들 수 있음을 알 수 있다.

본 연구는 최소의 데이터를 이용한 다변량 시스템의 효율적인 해석을 위해 응용되어 질 수 있을 것으로 판단된다.

참고문헌

- [1] Wu, Y.; "Pattern Recognition using Mahalanobis Distance," TPD Symposium, 1-14, 1996.
- [2] Lande, U.; "Mahalanobis Distance : A Theoretical and Practical Approach," <http://biologi.uio.no/fellesavdelinger/finse/spatialstats/Mahalanobis%20distance.ppt>, 2003.
- [3] I. C. Yeh; "Modeling of Strength of High-Perfprance Concrete Using Artificial Neural Networks," *Cement and Concrete Research*, 28(12) : 1797-1808, 1998.
- [4] 최희복, 김광희, 강경인; "재생골재 적용 콘크리트의 강도 예측", *대한건축학회논문집*, 21(6) : 49-55, 2005.
- [5] Taguchi, G and R. Jugulum; "New Trends in Multivariate Diagnosis," *Indian Journal of Statistics*, 62(B, 2) : 233-248, 2000.
- [6] Taguchi, G. and R. Jugulum; *The Mahalanobis-Taguchi Strategy : A Pattern Technology System*, John Wiley and Sons, Inc., 2002.

- [7] Hayashi, S., Y. Tanaka, and E. Kodama; "A New Manufacturing Control System using Mahalanobis Distance for Maximizing Productivity," IEEE Transactions, 59-62, 2001.
- [8] Hong, Jung-Eui, Cudney E. A, Taguchi G, Jugulum R., Paryani K., Ragsdell K.; "A Comparison study of Mahalanobis Taguchi System and neural network for multivariate pattern recognition," 2005, ASME IMECE Proceedings, Orlando FL.
- [9] E. A. Cudney, Jung-Eui Hong, Rajesh Jugulum, Kioumars Paryani, Kenneth M. Ragsdell, Genishi Taguchi; "An Evaluation of Mahalanobis Taguchi System and Neural Network for Multivariate Pattern Recognition," *Journal of Industrial and System Engineering*, 1(2) : 139-150, 2007.
- [10] R. Jugulum and L. Monplaisir; "Comparison between Mahalanobis-Taguchi-System and Artificial Neural Networks", *Journal of Quality Engineering Society*, 10(1) : 60-73, 2002.