IE Interfaces
Vol. 12, No. 4, pp. 551-556, December 1999.

범주형 자료에 대한 데이터 마이닝 분류기법 성능비교†

손소영 · 신형원

연세대학교 산업시스템공학과

Comparison of Data Mining Classification Algorithms for Categorical Feature Variables

So-Young Sohn · Hyung-Won Shin

In this paper, we compare the performance of three data mining classification algorithms(neural network, decision tree, logistic regression) in consideration of various characteristics of categorical input and output data. 2^{4-1} 3 fractional factorial design is used to simulate the comparison situation where factors used are (1) the categorical ratio of input variables, (2) the complexity of functional relationship between the output and input variables, (3) the size of randomness in the relationship, (4) the categorical ratio of an output variable, and (5) the classification algorithm. Experimental study results indicate the following: decision tree performs better than the others when the relationship between output and input variables is simple while logistic regression is better when the other way is around; and neural network appears a better choice than the others when the randomness in the relationship is relatively large. We also use Taguchi design to improve the practicality of our study results by letting the relationship between the output and input variables as a noise factor. As a result, the classification accuracy of neural network and decision tree turns out to be higher than that of logistic regression, when the categorical proportion of the output variable is even.

1. 연구의 배경과 목적

현대의 고도 산업사회는 컴퓨터 하드웨어의 기술발달로 데이터 저장비용이 저렴해졌으며 시장경쟁의 심화로 인하여 빠르고 정확한 데이터 분석능력을 요구하고 있다. 따라서 대용량의 데이터간의 관계, 패턴, 규칙 등을 찾아내어 모형화함으로써 유용한 정보를 고객에게 제공할 능력이 요구되고 있다. 이를 위하여 최근 대용량의 자료를 빠르고, 정확하고, 다양하게 분석할 수 있는 데이터 마이닝 기법들이 대두되고 있다. 데이터 마이닝의 모델링작업은 기술적 모델링(Descriptive Modeling)과 예측적 모델링(Predictive Modeling)으로 분류된다. 기술적 모델링이란 주어진 데이터를 설명하는 패턴을 찾아내는 것이 주목적이며 찾아낸 패턴을 사용자 이해를 위해 표현, 설명하는 작업으로써 연관규칙 발견(Association), 세분화(Clustering, Segmentation) 등이 이에 해당한다. 예측적 모델링은 주어진 데이터에 근거하여 모델을 만들고 이 모델을 이용하여 새로운 경우에 대한 예측을 하는 작업이며 분류(Classification) 및 예측(Prediction) 등이 이에 해당된다(한국 SAS Software, 1998; Hand, 1998). 여러 모델링 작업 중 가장 많이 사용되는 분류 모델링은 Training set를 근거로 Test set를 분류하는 것으로써 신경망분석, Decision Tree, 로지스틱 회귀분석 등 여러 분류기법에 따라 분류정확성에 차이가 있을 것으로 예상된다. 이와 같은 분류 성능비교를 한 연구 중 Brokett *et al.*(1997)은 기업의 파산여부를 예측하기 위하여 로지스틱 회귀분석과 유사한 판별분석과 신경망분석을 이용하였다. Leshno와 Spector(1997)는 Training 자료의 크기, 입출력변수간의 함수가 신경망과 판별분석의 분류정확도에 미치는 영향을 분석하였다. Cherkassky *et al.*(1996)은 입출력변수간에 함수, 데이터 크기, 에러의 크기에 따라 신경망을 비롯한 여러 가지 비모수 분류기법들을 비교하였다. 그러나 이들의 연구에서 고려되지 않은 것은 범주형 입력변수 성격이 분류정확성에 미치는 영향이다. 최근들어 마케팅, 통신, 제조, 교

†본연구는 한국과학재단 특정기초연구(1999-1-303-005-3)지원으로 수행되었음.

통 등 여러 분야에서 데이터 마이닝 기법들을 이용하여 자료를 효과적으로 분석하고자 하는 시도가 증가하고 있다. 이들 자료 중 많은 경우는 범주형 입력변수를 포함하고 있으며 분류를 분석목적으로 하고 있는 경우 또한 적지 않다(Bigus, 1996; Brokett, *et al.*, 1997; Gupta, *et al.*, 1997). 그러나 분석하고자 하는 범주형 자료에 적합한 데이터 마이닝 기법을 선택함에 있어 분석자의 기호에 의지하는 경우가 많다. 범주형 자료의 특성에 따른 기법별 분류성능의 차이에 대한 통계적 검증은 향후 데이터 마이닝을 이용한 자료분석자의 분류기법 선택에 유용한 정보를 제공할 것으로 보인다(손소영, 신형원, 1998). 따라서 본 연구에서는 데이터를 범주형으로 난수발생시켜 데이터의 특성을 나타내는 인자들의 효과가 각 데이터 마이닝의 분류성능에 미치는 영향을 실험계획법을 이용하여 검정하였다. 이를 위하여 범주형 데이터의 특성을 (1)입력변수의 범주별 비율 (2) 입출력변수간의 함수, (3) 에러의 크기 (4) 출력변수의 범주별 비율로 나누고, (5) 세 가지 데이터 마이닝 기법(인공신경망, Decision Tree, 로지스틱 회귀분석)을 이용하여 분류정확성을 비교하였다. 인공신경망은 여러 패턴 추출방법 중 일반적으로 예측능력에 높은 정확성을 가지고 있고 비선형 모형에 적합하다고 평가되고 있으며 Decision Tree는 범주형 자료에 높은 분류 정확성을 가지고 있고 대상이 되는 결과에 대하여 그 원인을 나뭇가지 형태로 찾아가 사용자가 알아보기 쉬운 장점이 있다. 또한 로지스틱 회귀분석은 범주형 자료분석에 오랜 기간 이용해 온 전통적 통계분석기법으로 알려져 있어 이들간의 성능비교는 의의가 있다고 본다.

본 논문의 구성은 다음과 같다. 2장에서 범주형 자료에 대한 분류기법별 성능비교를 위해 사용된 실험계획법과 분석결과를 설명하였다. 3장에는 본 연구결과의 실용성을 높이기 위해 실제로 성능비교시 제어할 수 없는 요소를 비제어인자로 보고 다구찌 디자인을 이용하여 재실험한 결과를 정리하였다. 4장에는 논의된 내용을 종합하고 향후 연구방향을 제시하였다.

2. 분류기법 성능비교를 위한 실험

본 장에서는 범주형 자료의 특성을 (1) 입력변수의 범주별 비율, (2) 입출력변수간의 연결함수, (3) 출력값의 잡음(noise)에 해당하는 에러의 크기, (4) 출력변수의 범주별 비율로 나누고 각 시나리오별로 신경망, Decision Tree, 로지스틱 회귀분석의 분류정확성을 분석하고자 한다.

본 연구에서 범주형 자료의 특성에 비추어 예측능력이 높은 데이터 마이닝 기법을 찾기 위하여 실험에 사용된 각 인자들과 수준은 <표 1>과 같다.

① 입력변수의 범주별 비율

실험시행마다 입력변수는 3개($X1$, $X2$, $X3$)이며 각 변수는 4개의 범주를 가지고 있다. 세 개 입력변수의 범주별 비율이 1:1:1:1로 균등한 경우와 7:1:1:1로 어느 범주에 자료가 치우쳐 있는 경우로 나누었다.

표 1. 연구자료의 인자와 수준

인자	수준	Coding
입력변수의 범주별 비율	1:1:1:1	-1
	7:1:1:1	1
함수	$Y_1=2X_1+0.5X_2+3X_3$	-1
	$Y_2=\exp((2X_1)+\sin(\pi(X_2))+10\log(X_3))+(3\cdot X_1)(\exp(X_2X_3)$	1
에러의 크기	$E_1 = 2 \times N(0,1)$	-1
	$E_2 = 8 \times N(0,1)$	1
출력변수의 범주별 비율	1:1	-1
	9:1	1
분류기법	역전파 신경망	-1
	Decsion Tree	0
	로지스틱 회귀분석	1

② 함수

모의 자료를 만들기 위하여 입력변수와 출력변수를 연결하는 함수는 간단한 선형 함수와 복잡한 비선형 함수로 나누었다.

③ 에러의 크기

일반적으로 자료에는 에러가 포함되어 있으므로 표준 정규분포를 따르는 데이터를 난수 발생시켜 이것에 2배와 8배가 되는 경우의 두 수준으로 나누어 ②에서 구한 함수에 더해 주었다. 함수의 두 수준과 에러의 크기를 나타내는 두 수준과의 조합으로 나오는 값의 상관계수는 <표 2>와 같다. <표 2>로 부터, 입출력 변수를 연결하는 함수에 표준 정규분포에 8배가 되는 에러(E_2)를 더하면 함수에 의한 설명력이 약 50%(1-0.752)에 불과해 지는 반면 2배가 되는 에러(E_1)를 더하면 설명력이 약 91%(1-0.292)가 되는 것을 알 수 있다.

표 2. 함수의 종류와 에러의 크기간 조합에 따른 피어슨 상관계수

	Y_1+E_1	Y_1+E_2	Y_2+E_1	Y_2+E_2
$E\sim N(0,1)$	0.29712	0.75989	0.25849	0.73462

④ 출력변수의 범주별 비율

하나의 출력변수에 2개의 범주를 가지고 있도록 하며 범주별 비율은 임계값을 정하여 균등하게 1:1인경우와 9:1인 경우로 나누었다.

⑤ 데이터 마이닝 기법

역전파 신경망, CART(Classification and Regression Tree)를 이용한 Decision Tree, 로지스틱 회귀분석이 사용되었다. 인공신경망 중 가장 보편적으로 사용되는 역전파 신경망은 입력층, 하나

표 3. 일부 요인계획법(Mixed Level Fractional Factorial Design)

	입력변수의 범주별 비율	함수의 종류	에러의 크기	출력변수의 범주별 비율	분류기법	분류정확성(%)
1	-1	-1	-1	-1	-1	96.00
2	1	-1	-1	1	-1	97.50
3	-1	1	-1	1	-1	89.75
4	1	1	-1	-1	-1	79.25
5	-1	-1	1	1	-1	85.25
6	1	-1	1	-1	-1	74.50
7	-1	1	1	-1	-1	69.50
8	1	1	1	1	-1	91.67
9	-1	-1	-1	-1	0	95.25
10	1	-1	-1	1	0	99.50
11	-1	1	-1	1	0	89.75
12	1	1	-1	-1	0	78.75
13	-1	-1	1	1	0	85.25
14	1	-1	1	-1	0	74.50
15	-1	1	1	-1	0	69.50
16	1	1	1	1	0	87.50
17	-1	-1	-1	-1	1	88.50
18	1	-1	-1	1	1	94.75
19	-1	1	-1	1	1	98.25
20	1	1	-1	-1	1	91.00
21	-1	-1	1	1	1	87.75
22	1	-1	1	-1	1	72.50
23	-1	1	1	-1	1	73.25
24	1	1	1	1	1	88.25

이상의 은닉층, 출력층으로 이루어져 있으며 은닉층은 뉴런(neuron) 또는 노드라 불리는 요소로 구성되어 있다. 이들은 다음 층의 요소와 연결강도 또는 가중값을 갖는 링크(link)로 연결되어 있으며 전 단계의 출력값을 입력값으로 받아 특정 활성함수(Activation Function)에 의해 출력값을 생성하게 된다. 본 연구에서는 신경망이 최대의 성능을 발휘하도록 하기 위하여 반복실험 후, 2개의 은닉층에 각각 2, 3개의 뉴런을 갖는 구조를 선택하고 활성함수로는 Hyperbolic Tangent를 사용하였다.

Decision Tree란 대상이 되는 집단을 나뭇가지처럼 몇 개의 소집단으로 구분하여 분류하고 예측하는 기법이다. Decision Tree중 본 연구에 사용된 CART는 주어진 데이터에 대하여 각 노드마다 이진분류(binary tree)로 진행한다. 이를 위하여 모든 설명변수에 대하여 혼합도 값이 최소가 되는 설명변수를 반복적으로 검색하여 분류에 사용할 변수로 선택한다. 이 과정을 전 단계의 혼합도값과 유의한 차이가 나지 않을 때까지 계속한다.

로지스틱 회귀분석은 종속변수가 범주형일 때 그 변화를 설명변수(X)의 함수로 하여 예측하고자 할 때 사용되는 모수적인 방법이다. 일반적으로 로지스틱 회귀분석은 다음과 같이 정의된다.

$$P_{i(x)} = \frac{\exp[\alpha_i + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n]}{1 + \exp[(\alpha_i + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)]}$$

여기서 $P_i(x)$는 주어진 설명변수 $(X_1,\ldots,X_n)$ 하에서 종속변수의 분류수준이 i 보다 작거나 같을 확률을 말한다. 일반적으로 설명변수가 분류 결정에 미치는 영향은 β에 대한 추론으로 가능하다. 미지의 모수인 α_i, β_i는 보통 최우추정법(Maxmun Likelihood Estimation)이나 반복적인 가중 최소 자승법(Iterative Weighted Least Squares Method)을 이용하여 추정이 된다.

다음은 앞서 언급된 ①~⑤ 인자와 각각의 수준을 고려한 $2^{4-1} \cdot 3$ 일부요인 실험계획법을 이용하여 분류정확성을 측정하였다. 실험과정은 24개의 treatment마다 2000개의 난수를 발생시켜 얻은 데이터를 Training에 60%, Validation에 40%로 할당하고 해당 분류기법의 정확성을 측정하였다. 각 실험결과는 <표 3>과 같다.

이와 같은 실험 결과를 이용하여 검정하려는 9개의 실험가설은 다음과 같다.

· 주효과에 의한 가설

Ha1: 입력변수의 범주별 비율은 분류정확성에 영향을 미친다.

표 4. 인자와 교호작용에 대한 분산분석표

요인	DF	Sum of Square	Mean Square	F_0	$Pr>F$
Model	14	1734.8972	123.9212	3.23	0.0411
입력변수의 범주별 비율	1	0.1162	0.1162	2.45	0.2581
함수의 종류	1	83.7837	83.7837	1764.62	0.0006
에러의 크기	1	803.0737	803.0737	16923.14	0.0001
출력변수의 범주별 비율	1	733.3887	733.3887	15454.67	0.0001
분류 기법	2	13.8353	6.9176	145.78	0.0068
입력변수의 범주별 비율×분류 기법	2	0.8424	0.4212	8.88	0.1013
함수의 종류×분류 기법	2	94.5424	47.2712	996.14	0.0010
에러의 크기×분류기법	2	5.2649	2.6324	55.47	0.0177
e(출력변수의 범주별 비율×분류 기법)	2	0.0949	0.0474		

Ha2: 입출력변수간의 함수는 분류정확성에 영향을 미친다.
Ha3: 에러의 크기는 분류정확성에 영향을 미친다.
Ha4: 출력변수의 범주별 비율은 분류정확성에 영향을 미친다.
Ha5: 데이터 마이닝기법은 분류정확성에 영향을 미친다.

- 교호작용에 의한 가설

Ha6: 입력변수의 범주별 비율과 데이터 마이닝기법들 간의 교호작용은 분류정확성에 영향을 미친다.
Ha7: 입출력변수간의 함수와 데이터 마이닝기법들 간의 교호작용은 분류정확성에 영향을 미친다.
Ha8: 에러의 크기와 데이터 마이닝기법 간의 교호작용은 분류정확성에 영향을 미친다.
Ha9: 출력변수의 범주별 비율과 데이터 마이닝기법들 간에 교호작용은 분류정확성에 영향을 미친다.

이와 같은 가설들을 <표 3>의 실험결과를 바탕으로 분산분석을 하여 <표 4>와 같이 검정하였다. 검정하고자 하는 가설들의 효과는 다른 효과와 교락(Confounding)되지 않는 것으로 나타났다.

표 5. 함수의 종류와 데이터 마이닝 기법간 교호작용에 의한 분류정확성 (α=0.05)

Duncan Grouping	함수의 종류 수준	분류 기법 수준	N	Mean
A A	-1	Decision Tree	4	88.6250
B A B	-1	신경망	4	88.3125
B B	1	로지스틱회귀분석	4	87.6875
C	-1	로지스틱회귀분석	4	85.8750
D	1	신경망	4	82.5425
E	1	Decision Tree	4	81.3750

표 6. 에러의 크기와 데이터 마이닝 기법간 교호작용에 의한 분류정확성 (α=0.05)

Duncan Grouping	에러의 크기 수준	분류 기법 수준	N	Mean
A	-1	로지스틱회귀분석	4	93.1250
B	-1	Decision Tree	4	90.8125
B	-1	신경망	4	90.6250
C	1	로지스틱회귀분석	4	80.4375
C	1	신경망	4	80.2300
D	1	Decision Tree	4	79.1875

유의수준 5%에서 가설검정결과 함수의 종류, 에러의 크기, 출력변수의 범주별 비율, 사용한 기법은 주효과가 있는 것으로 나타났으며 함수의 종류×사용한 기법, 에러의 크기×사용한 기법은 교호작용이 있는 것으로 나타났다. 이상의 연구결과를 바탕으로 데이터의 특성에 따라 적합한 분류기법을 선택하기 위하여 함수의 종류×사용한 기법, 에러의 크기×사용한 기법에 대한 유의수준 5%에서 Duncan 검정결과는 <표 5>, <표 6>과 같다.

<표 5>에 나타난 Duncan검정결과에 의하면, α=0.05에서 함수의 종류가 간단할 때 Decision Tree는 로지스틱 회귀분석에 비하여 높은 분류정확성을 가지고 있는 것으로 나타났다. 또한 함수의 종류가 복잡할 때 로지스틱 회귀분석, 신경망, Decision Tree순으로 분류정확성이 높은 것으로 나타났다. <표 6>의 Duncan 검정결과에 의하면, α=0.05 에서 에러의 크기가 작을 때 로지스틱 회귀분석이 다른 기법에 비하여 분류정확성이 높은 것으로 나타났으며 신경망과 Decision Tree는 유의한 차

표 7. 비제어 인자를 고려한 실험설계

요인배치	외측배열				내측배열		
	입력변수의 범주별 비율	에러의 크기	출력변수의 범주별 비율	분류 기법	분류정확성(%) 함수의 종류(B)		$SN_i=-10\log[\frac{1}{2}(1/y_{i1}^2+1/y_{i2}^2)]$
					-1(Y_1)	1(Y_2)	SN 비
1	-1	-1	-1	-1	96.00	95.00	39.60
2	1	-1	1	0	99.50	98.00	39.89
3	-1	1	1	1	87.75	88.75	38.91
4	1	1	-1	1	72.50	69.75	37.04
5	1	1	-1	0	74.50	77.50	37.61
6	-1	1	1	-1	85.25	90.25	38.85
7	1	-1	1	-1	97.50	98.00	39.80
8	-1	-1	-1	0	95.25	95.00	39.57
9	-1	-1	-1	1	88.50	91.25	39.07
10	1	-1	1	1	94.75	98.00	39.68
11	-1	1	1	0	85.25	89.84	38.84
12	1	1	-1	-1	74.50	77.50	37.61

표 8. 비제어 인자를 고려한 분산분석

요인	DF	Sum of Square	Mean Square	F_0	$Pr>F$
Model	7	10.04419	1.43488	53.62	0.0009
입력변수의범주별 비율(A)	1	0.8586	0.8586	350.48	0.0340
에러의 크기 (C)	1	6.3802	6.3802	2604.17	0.0125
출력변수의 범주별 비율 (D)	1	2.4934	2.4934	1017.72	0.0199
분류 기법 (E)	1	0.1682	0.1682	68.65	0.0765
$A\times E$	1	0.0060	0.0060	2.47	0.3608
$D\times E$	1	0.1352	0.1352	55.18	0.0852
에러($C\times E$)	1	0.02450	0.02450		

이가 나지 않는 것으로 나타났다. 에러의 크기가 큰 경우는 Decision Tree가 다른 기법에 비하여 분류 정확성이 떨어지며 신경망과 로지스틱 회귀분석간에는 유의한 차이가 나지 않는 것으로 나타났다.

3. 다구찌 디자인

본 연구에서 범주형 자료의 특성으로 선택한 네 가지 인자 중 입출력변수간에 관계를 나타내는 '함수의 종류'는 실질적으로 데이터가 주어졌을 때 간단하고 복잡한 정도를 알 수 없으므로 이에 따라 적합한 데이터 마이닝 기법을 찾는다는 것에 큰 의미가 없다고 할 수 있을 것이다. 따라서 '함수의 종류'를 비제어인자로 보고 다구찌 디자인을 이용하여 망대특성을 위한 실험설계를 하면 <표 7>과 같다(박성현, 1995).

함수의 종류를 비제어인자로 하여 분산 분석을 수행한 결과 <표 8>과 같이 $\alpha=0.1$에서 입력 변수의 범주별 비율, 에러의 크기, 출력변수의 범주별 비율, 사용한 분류기법, 출력변수의 범주별 비율과 사용한 기법간의 교호작용이 분류정확성에 유의하게 영향을 미치는 것으로 나타났다. 이 실험에서 검정하려는 효과는 다른 효과와 교락되지 않는 것으로 나타났다.

<표 9>에 나타난 Duncan 테스트결과에 의하면 $\alpha=0.05$에서 출력변수의 범주별 비율이 어느 한쪽으로 치우치지 않았을 경우 신경망과 Decision Tree가 로지스틱 회귀분석에 비하여 분류정확성이 높은 것으로 나타났으며 출력변수의 범주별 비율이 한쪽으로 치우친 경우 기법간 분류정확성에 유의한 차이가 나지 않는 것으로 나타났다.

4. 결 론

표 9. 출력변수의 범주별 비율과 데이터 마이닝 기법의 교호작용에 의한 분류정확성

Duncan Grouping	출력변수의 범주별 비율	데이터 마이닝 기법 수준	N	SN비의 Mean
A A	1	Decision Tree	4	39.37
A A	1	신경망	4	39.33
A	1	로지스틱 회귀분석	4	39.30
B B	-1	신경망	4	38.61
B	-1	Decision Tree	4	38.59
C	-1	로지스틱 회귀분석	4	38.06

본 연구에서는 신경망, Decision Tree, 로지스틱 회귀분석을 이용하여 범주형 자료에 대한 분류분석을 할 때 분류 정확성에 영향을 미치는 인자를 알아보았다. 분류정확성에 영향을 미치는 인자로 다섯 가지를 선택하여 분산분석을 한 결과, 함수의 종류와 에러의 크기, 출력변수의 범주별 비율, 데이터 마이닝 기법의 선택이 α=0.05에서 유의하게 분류정확성에 영향을 미치는 것으로 나타났다. 범주형 자료의 특성에 따라 적합한 데이터 마이닝 기법을 찾기 위하여 교호작용을 통하여 분석한 결과, 함수의 종류가 간단할 때는 Decision Tree, 함수의 종류가 복잡할 때 로지스틱 회귀분석이 적합한 것으로 나타났다.

이는 일반적으로 신경망이 로지스틱 회귀분석에 비하여 비선형 함수관계를 갖는 자료에 더 높은 분류정확성을 갖는다고 알려진 것과는 다른 실험결과를 보이고 있다. 또한 로지스틱 회귀분석은 에러의 크기가 작을 때와 클 때 모두 적합하며 신경망은 에러의 크기가 클 때 적합한 것으로 나타났다. 한편, 4가지 인자중 함수의 종류는 주어진 자료에서 파악 할 수 없는 성격이므로 다구찌 디자인을 이용하여 비제어 인자로 간주하여 실험한 결과, 출력변수의 범주별 비율이 어느 한 쪽으로 치우치지 않았을 경우, 신경망과 Decision Tree의 분류정확성이 로지스틱 회귀분석에 비하여 높은 것으로 나타났다. 이와 같이 '함수의 종류'를 비제어 인자로 한 실험은 네 가지 인자를 모두 사용한 실험보다 현실적이라 할 수 있을 것이다. 본 연구에서는 데이터 마이닝기법 별로 분류성능의 차이를 파악하기 위해 네 가지 인자를 대상으로 실험하였으나 이밖의 다른 인자와 다른 수준을 대상으로 하는 폭넓은 연구가 진행되어야 할 것이다.

따라서 향후 연구방향으로, 본 연구와 같이 모의자료를 이용하여 데이터 마이닝 기법의 분류성능을 측정한 연구를 취합하여 메타분석을 함으로써 좀더 다양한 인자와 수준을 고려한 데이터 마이닝 기법의 성능비교를 제시할 수 있겠다.

참고문헌

박성현 (1995), *다구찌 방법과 통계적 공정관리를 중심으로 한 품질관리*, 민영사, 217-234.

손소영, 신형원 (1998), 데이터 마이닝을 이용한 교통사고 심각도 분류분석, *대한교통학회*, 16(4), 187-194.

한국 SAS Software (1998), *Enterprise Miner 2.0 Seminar 자료집*, 4-5.

Berry, M. J. A. and Linoff, G (1997)., *Data Mining Techniques*, John Wiley & Sons, 243-285.

Bigus, J. P (1996)., *Data Mining with Neural Networks*, Mcgraw-Hill, 61-97.

Brokett, P. L., Cooper, W. W., Golden, L. L. and Xia, X. (1997), A Case Study in Applying Neural Networks to Predicting Insolvency for Property and Casualty Insurers, *Journal of the Operational Research Society*, 48, 1153-1162.

Cherkassky, V., Gehring, D. and Mulier, F. (1996), Comparison of Adaptive Methods for Function Estimation from Samples, *IEEE Transaction on Neural Networks*, 7(4), 969-984.

Gupta, V. K., Chen, J. G. and Murtaza, M. B. (1997), A Learning Vector Quantization Neural Network Model for the Classification of Industrial Construction Project', Omega, Int. J. Mgmt Sci., 25, 715-727.

Hand, D. J.(1998), Data Mining: Statistics and More?, *Journal of American Statistical Association*, 52(2), 112-118.

Leshno, M. and Spector, Y. (1997), The Effect of Training Data Set Size and the Complexity of the Separation Function on Neural Network Classification Capability: The Two-Group Case, *Naval Research Logistics*, 44, 699-717.

Michie, D., Spiegelhalter, D. J. and Taylor, C. C. (1994), Machine Learning, *Neural and Statistical Classification*, Ellis Horwood, 84-96.

Peterson, Generald E., Clair, Daniel C., Aylward, S. R., and Bond, W. E. (1995), Using Taguchi's Method of Experimental Design to Control Error in Layered Perceptrons, *IEEE Transaction on Neural Network*, 6(4), 949-960.

손소영

1981년 연세대학교 수학과 학사
1983년 한국과학기술원 산업공학과 석사
1989년 Univ. of Pittsburgh 산업공학과 박사
현재: 산업시스템공학과 교수로 재직중
관심 분야: 품질 및 신뢰도 공학모형 추정, 마케팅 분석 등

신형원

1998년 연세대학교 산업시스템공학과 학사
현재: 연세대학교 산업시스템공학과 석사 과정 재학중
관심 분야: 데이터 마이닝, 마케팅분석