Journal of the Korean Institute of Industrial Engineers
Vol. 27, No. 1, pp. 47-53, March 2001.

다구찌 디자인을 이용한 앙상블 및 군집분석 분류 성능 비교

신형원 · 손소영†

연세대학교 컴퓨터 과학 · 산업시스템공학과

Comparing Classification Accuracy of Ensemble and Clustering Algorithms Based on Taguchi Design

Hyung Won Shin · So Young Sohn

Dept. of Computer Science and Industrial Systems Engineering, Yonsei University, Seoul

In this paper, we compare the classification performances of both ensemble and clustering algorithms (Data Bagging, Variable Selection Bagging, Parameter Combining, Clustering) to logistic regression in consideration of various characteristics of input data. Four factors used to simulate the logistic model are (1) correlation among input variables (2) variance of observation (3) training data size and (4) input-output function. In view of the unknown relationship between input and output function, we use a Taguchi design to improve the practicality of our study results by letting it as a noise factor. Experimental study results indicate the following: When the level of the variance is medium, Bagging & Parameter Combining performs worse than Logistic Regression, Variable Selection Bagging and Clustering. However, classification performances of Logistic Regression, Variable Selection Bagging, Bagging and Clustering are not significantly different when the variance of input data is either small or large. When there is strong correlation in input variables, Variable Selection Bagging outperforms both Logistic Regression and Parameter combining. In general, Parameter Combining algorithm appears to be the worst at our disappointment.

Keywords : Classification, Taguchi Design, Ensemble, Clustering

1. 연구배경

현대의 고도 산업사회는 컴퓨터 하드웨어의 기술 발달로 데이터 저장 비용이 저렴해졌으며 시장 경쟁의 심화로 인하여 빠르고 정확한 데이터 분석능력을 요구하고 있다. 따라서 대용량의 데이터 간의 관계, 패턴, 규칙 등을 찾아내어 모형화 함으로써 유용한 정보를 고객에게 제공할 능력이 요구되고 있다. 이를 위하여 최근 대용량의 자료를 빠르고, 정확하고, 다양하게 분석할 수 있는 데이터 마이닝 기법들이 대두되고 있다. 데이터 마이닝 모델링 작업은 연관규칙(Association Rule), 세분화(Clustering, Segmentation), 분류(Classification), 값 예측(Value Prediction) 등이 있으며 이들은 마케팅, 통신, 제조, 교통 등 다양한 분야에서 활용되고 있다. 데이터 마이닝 작업 중 가장 많이 사용되는 분류 모델링은 학습용(Training) 데이터로부터 입력과 출력 간의 관계를 학습하고 이를 바탕으로 새로운(Test) 데이터를 분류하는 데 적용된다. 이러한 분류 모델링의 가장 큰 이슈는 분류정확성의 향상이며, 이를 위하여 몇 개의 부트스트랩 샘플에 단일모형을 여러번 적용하여 분류한 결과를 융합해주는 퓨전 기법의 일종인 앙상블 방법에 대한 많은 연구가 있어왔다(Breiman, 1994; Christodoulou and Pattichis, 1998; Freund and Shapire, 1996; Ho *et al.*, 1994; Kittler *et al.*, 1998; Optiz and Maclin, 1997; Shanon and Banks, 1999). 또한 데이터의 분산이 큰 경우, 기존의 앙상블 방법과는 반대로 데이터를 특성에 따라 군집으로 나누고 각 군집별로 분류모형을 학습하는 Clustering 방법이 연구되었다(Cao, 1995; Sohn and

본 논문은 한국과학재단 특정기초연구(1999-1-303-0053)의 지원을 받음.
† Corresponding author : Professor So Young Sohn, Dept. of Computer Science and Industrial Systems Engineering, Yonsei University, Shinchondong 134, Seoul, Korea, Fax : 82-2-364-7807, e-mail : sohns@yonsei.ac.kr
2000년 4월 접수, 2회 수정 후, 2000년 12월 게재 확정.

Lee, 1999). 이처럼 주어진 데이터의 특성에 따라 적절한 앙상블 또는 군집분석 방법을 선택하는 일종의 메타모형은 그 중요성에도 불구하고 연구가 많이 되어 있지 않은 상황이다. 따라서 본 연구에서는 Monte Carlo Simulation을 이용하여 데이터의 특성을 나타내는 인자들과 앙상블, Clustering 방법 간의 교호작용을 분류정확성의 관점에서 분석하고자 한다. 이를 위하여 데이터의 특성을 (1) 입력변수 간의 상관관계, (2) 데이터의 분산, (3) 데이터의 크기, (4) 입출력변수 간의 함수로 나누고, (5) 분류방법(로지스틱 회귀분석, Bagging, Variable Selection Bagging, Parameter Combining, Clustering)에 따라 이진 출력값에 대한 분류정확성을 비교하였다. 이들 요인 중 입력변수와 출력변수 간의 함수는 주어진 데이터에서 실제 알 수 없는 성격이므로 다구찌 실험계획법을 이용하여 비제어 인자로 간주하였다.

본 논문의 구성은 다음과 같다. 2절에서는 앙상블 및 군집분석기법과 이에 관련된 기존문헌을 고찰하였으며 3절에서는 본 연구에 사용된 실험계획법 및 실험 가설에 대하여 설명 하였다. 4절에는 다구찌 실험계획을 이용하여 실험한 결과를 정리하였으며 5절에는 논의된 내용을 종합하고 향후 연구방향을 제시하였다.

2. 분류성능 향상을 위한 앙상블 및 군집분석 기법

데이터 마이닝 작업 중 가장 일반적으로 사용되는 분류 모형에는 신경망, Decision Tree, 로지스틱 회귀분석 등이 있다. 인공신경망은 여러 패턴 추출방법 중 일반적으로 예측 능력에 높은 정확성을 가지고 있고 비선형 모형에 적합하다고 평가되고 있다. Decision Tree는 범주형 자료에 높은 분류 정확성을 가지고 있고 대상이 되는 결과에 대하여 그 원인을 나뭇가지 형태로 찾아가 사용자가 이해하기 쉬운 장점이 있다. 또한 로지스틱 회귀분석은 범주형 자료분석에 오랜 기간 이용해 온 전통적 통계분석 기법이다. 본 연구에서는 로지스틱 회귀분석을 바탕으로 분류정확성 향상을 위한 여러 가지 앙상블 기법을 비교하였다. 앙상블 기법이란 다중 분류기들로부터 얻은 예측값들을 결합하는 방법으로써 많은 연구자들이 하나의 분류기를 사용하는 경우보다 높은 분류성능을 얻기 위한 노력을 해왔다. 지금까지 보편적으로 알려져 있는 앙상블로는 Bagging, Arcing을 등을 들 수 있다.

Bagging(Bootstrap AGGregatING)의 기본 아이디어는 각 분류기들의 편차를 줄이기 위해 부트스트랩 샘플링을 바탕으로 구한 예측치들의 값을 Voting 하는 데 있다. Bagging 기법의 분류진행순서를 보면 다음과 같다.

① 본래의 데이터와 동일한 크기를 갖는 부트스트랩 샘플 b를 만들어 원래의 데이터를 대체하여 분류기 Cb를 학습시킨다 ($b=1,\cdots, B$).

② 입력 X와 출력 Y로 구성된 부트스트랩 샘플 i에 대한 분류기의 가능한 예측 클래스의 종류를 K라 한다($Cb(x)=1,\cdots, K$).

③ 임의의 관측치 X의 분류결과는 B개의 분류기 Cb의 K개 예측 클래스 중 가장 많은 Voting 결과를 가진 클래스로 정한다.

Arcing(Adaptive Resampling and Combining)은 분류관련 문제에서만 독보적으로 다루어진 기법으로 Schapire(1990)에 의해 Boosting이라는 기법으로 처음으로 개발되었으나, Breiman (1996)이 이것을 Arcing으로 발전시켰다. Arcing의 기본 아이디어는 Bagging과 같으나 샘플링할 때, 분류 정확성을 높이는 것으로 나타난 관측치가 샘플링 될 확률이 높도록 부트스트랩 리샘플링 한다.

이상의 앙상블 방법을 이용한 분류에 대한 기존의 연구결과는 분류정확성을 높인 경우도 있었으며 오히려 낮춘 경우도 있었다. Breiman(1994)은 Bagging 방법을 제안하고 시뮬레이션 데이터와 실제 데이터에 적용하여 분류 정확성의 향상을 보였으며, Optiz and Maclin(1997)은 14개의 실제 데이터를 대상으로 Bagging과 Boosting을 이용하여 신경망 앙상블과 Decision Tree 앙상블을 만들어 분류하였다. 이들의 연구에서, Boosting은 전반적으로 단일모형을 사용한 경우보다 분류 정확성이 향상되었고 Bagging은 경우에 따라 다른 분석결과를 보였다. 이밖에 분류 정확성 향상을 위하여 Bagging, Boosting을 이용한 방법 외의 다양한 방법이 시도되어 왔다. Christodoulou and Pattichis (1998)은 근전도(EMG : Electromyographic) 신호로 3가지 유형의 질병을 분류하기 위하여 72개의 설명변수를 그 특징에 따라 6개의 변수집합으로 만들어 6개의 자기조직화 신경망(Self Organizing Map)으로 분류한 뒤 신뢰구간을 바탕으로 결합 하였다. Nezafat *et al.*(1998)은 최근 이웃 학습법(K-Nearest Neighborhood), MLP(Multi Layer Perceptron) 신경망, RBF(Radial Basis Function) 신경망 등 6가지 분류기의 특성에 따라 적절한 변수를 선택하여 학습한 후, 융합하는 방법을 연구하였다. Ho *et al.* (1994)은 4개의 분류기를 사용하고 각 분류기마다 임의의 변수를 선택하여 분류한 뒤(Variable Selection Bagging), Borda Count, 로지스틱 회귀분석으로 융합하였다. Guvenir and Sirin(1996)은 연속형 설명변수를 대상으로 설명변수의 값에 따라 구간별로 나눈 뒤 구간별 예측값을 구하고 각 설명변수의 예측값을 Voting하는 휴리스틱을 개발하였다. Shannon and Banks(1999)는 전체 데이터중 B번의 샘플을 취하여 B개의 Decision Tree를 추정하고 B개의 Decision Tree와 가장 가까운 하나의 Decision Tree를 최우추정법(Maximum Likelihood Estimation)을 이용하여 추정하는 Parameter Combining 방법을 제시하였다. Cao *et al.*(1995)은 문자인식을 위해 비지도 신경망으로 데이터를 군집화 한 후, 역전파 신경망으로 군집별 학습을 하는 방법을 사용하였다. 손소영, 이성호(2000)는 교통사고 분류분석에 Bagging, Arcing, Demster-Shafer 이론 등, 다양한 앙상블 방법을 사용하여 분류 정확성을 향상시키고자 하였으며 Clustering 분석을 이용한 군집별 학습 방법이 가장

분류 정확성을 향상시키는 것으로 결과를 보였다. 그러나 이상의 다양한 앙상블 방법에 대한 연구들은 데이터의 특성을 중심으로 된 것이라기 보다는 경험적(empirical) 연구의 측면이 강하다. 따라서 본 논문은 기존의 연구에서 수행된 Bagging, Variable Selection Bagging, Parameter Combining 방법과 더불어 여러 분류기 예측 결과를 융합하는 기존의 앙상블 방법과는 반대로 데이터를 특성에 따라 군집으로 나누고 각 군집별 분류를 하는 Clustering 방법의 성능을 평가하고자 한다. 분류 방법에 따른 성능평가의 현실성을 높이기 위하여 다구찌 디자인을 바탕으로 데이터로부터 성격을 파악할 수 있는 제어인자와 파악할 수 없는 비제어 인자를 동시에 고려한 시뮬레이션 성능을 연구하였다.

3. 실험 디자인

본 장에서는 데이터의 특성에 비추어 예측능력이 높은 분류기법을 찾기 위한 시뮬레이션을 시행하였다. 시뮬레이션 데이터는 다중 정규(Multivariate Normal) 분포를 따르는 5개의 입력변수와 이진값(Binary)을 가지는 출력변수로 이루어져 있으며, 이들의 특성을 나타내는 실험의 인자(Factor)와 수준(Level)을 정하였다. 실험에 사용한 인자는 기존 연구에서 사용된 바 있는 '데이터의 크기'와 '입출력 변수 간의 함수'와 더불어, 본 연구에서는 입력 변수 간의 상관관계와 데이터의 분산을 새로이 추가하여 분류정확성에 미치는 영향을 파악하였다(Peterson *et al.*, 1995 ; Sohn and Shin, 1999). 디자인에 사용된 각 요인별 수준을 자세히 살펴보면 다음과 같다.

3.1 입력변수 간의 상관관계

5개 입력변수 간 상관관계가 약할 때의 ρ^2 값은 각각 0.05~0.3 사이이며, 중간일 때는 0.4~0.7, 강할 때는 0.7~0.96 사이로 가정하였다. 실험에 사용된 각 입력변수 간 상관관계는 <표 1>, <표 2>, <표 3>과 같다.

3.2 데이터의 크기

데이터의 크기는 변수의 수에 비하여 400배의 관측치 수를 가지는 '상대적으로 작은' 2000개 데이터 셋과 변수의 수에 비하여 2000배의 관측치 수를 가지는 '상대적으로 많은' 10000개의 데이터 셋으로 나누었다. 전체 데이터의 60 %는 학습용 자료로, 40 %는 검증용 자료로 사용하였다.

- small 2000(학습용 1200, 검증용 800)
- large 10000(학습용 6000, 검증용 4000)

표 1. 입력변수 간의 상관관계가 약함

ρ^2	X1	X2	X3	X4	X5
X1	1.00	0.12	0.08	0.22	0.05
X2	0.12	1.00	0.07	0.20	0.04
X3	0.08	0.07	1	0.05	0.26
X4	0.22	0.20	0.05	1	0.17
X5	0.05	0.04	0.26	0.17	1

표 2. 입력변수 간의 상관관계가 중간 정도임

ρ^2	X1	X2	X3	X4	X5
X1	1.00	0.56	0.46	0.45	0.35
X2	0.56	1.00	0.66	0.42	0.48
X3	0.46	0.66	1.00	0.60	0.50
X4	0.45	0.42	0.60	1.00	0.62
X5	0.35	0.48	0.50	0.62	1.00

표 3. 입력변수 간의 상관관계가 강함

ρ^2	X1	X2	X3	X4	X5
X1	1.00	0.90	0.79	0.94	0.92
X2	0.90	1.00	0.94	0.90	0.81
X3	0.79	0.94	1.00	0.75	0.96
X4	0.94	0.90	0.75	1.00	0.73
X5	0.92	0.81	0.96	0.73	1.00

3.3 데이터의 분산

다섯 개 입력 변수의 평균을 0, 분산 공분산 행렬은 <표 1, 2, 3>의 상관행렬(Correlation Matrix)에 1, 10,100을 곱한 세 수준으로 다중 정규(Multivariate Normal) 분포를 따르도록 하였다. 데이터의 분산 수준에 따른 이진 출력에 대한 확률값 P의 산포정도를, 입력변수 간의 상관관계가 약하고 데이터의 크기가 작을 때의 예를 들어 <그림 1>에 나타내었다.

3.4 입출력 변수 간의 연결함수

시뮬레이션을 위하여 실제 모델로 사용한 입출력 변수 간의 함수는 모수의 관점에서 로지스틱 선형인 경우와 로지스틱 비선형인 경우로 나누었다.

- 선형

$$P_{i(x)} = \frac{\exp[0.5+0.01X_1+0.02X_2+0.03X_3+0.04X_4+0.05X_5]}{1+\exp[0.5+0.01X_1+0.02X_2+0.03X_3+0.04X_4+0.05X_5]} \quad (3)$$

- 비선형

$$P_{i(x)} = \frac{\exp(0.2\sin(x1/x2)^2+0.5\sqrt{(|x2-3x3|)}\cos|x4/x5|)}{1+\exp(0.2\sin(x1/x2)^2+0.5\sqrt{(|x2-3x3|)}\cos|x4/x5|)} \quad (4)$$

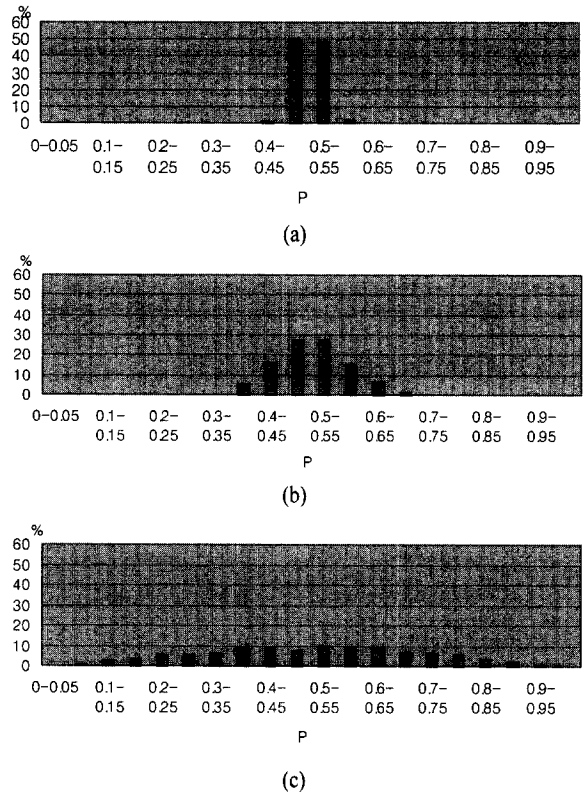

그림 1. 데이터의 분산 정도에 따른 이진출력의 확률 P의 산포도(입력변수 간 상관관계가 약하고 데이터의 크기가 작을 때의 예) (a) 분산이 작을 때 (b) 분산이 중간일 때 (c) 분산이 클 때.

3.5 분류방법

실험에 사용된 분류방법은 전통적 통계분석 방법으로 오랜 기간 사용된 로지스틱 회귀분석과 앙상블 방법으로 가장 널리 알려진 Bagging(Breiman, 1994), 일부 변수만을 번갈아 사용하므로 경제적인 분류 방법인 Variable Selection Bagging(Ho *et al.*, 1994)), Shanon and Banks(1999)에 의하여 제안된 Parameter Combining, 데이터의 분산이 클 때 효과적일 수 있는 Clustering 방법을 사용하였다.

• 로지스틱 회귀분석

로지스틱 회귀분석은 출력변수가 범주형일 때 그 변화를 입력변수(X)의 함수로 예측 할 때 사용되는 모수적인 방법이다 (Sohn and Shin, 1999).

• Bagging

본 실험에서는 16개의 부트스트랩 샘플에 16개의 로지스틱 분류기를 적용한 후 Bagging 하였다(Breiman, 1994).

• Variable Selection Bagging

분류분석을 하는 데 있어서 가능하면 작은 수의 변수를 사용하는 것이 경제적이다. 따라서 Variable Selection Bagging 분류기는 실험에서 사용된 5개 설명변수 중 3개씩 불규칙하게 사용하여 16번의 부트스트랩 샘플을 바탕으로 16개의 로지스틱 분류기를 Bagging하는 방법을 취하였다(Ho *et al.*, 1994).

• Parameter Combining

Parameter combining 분류기는 16번의 부트스트랩 샘플로 16개의 로지스틱 분류기 $\hat{p}_i$를 만들어 이 때 추정된 모수를 바탕으로 식 (3)과 같이 오차를 줄이는 대표적인 로지스틱 회귀모형(P)을 재 추정하여 이를 바탕으로 자료를 다시 분류하는 방법이다 (Shanon and Banks, 1999).

$$\text{Min} \sum_{i=1}^{16} (\hat{P}_i - P)^2 \tag{5}$$

• Clustering

이 방법은 학습용 데이터를 K-평균법을 이용하여 몇 개의 군집으로 나누고 각 군집별로 분류기법을 이용한 학습을 하는 방법이다(Cao, 1995; Sohn and Lee, 2000). 본 연구에서 사용된 데이터의 적절한 군집 개수를 결정하기 위하여 학습용 데이터를 이용한 사전 실험을 바탕으로 2~5개로 변화시켜 시험해본 결과, 4개의 군집수가 가장 적절한 것으로 나타났다. 분류정확성의 측정은 검증용 데이터를 학습용 데이터에 근거하여 4개의 군집으로 나누고 군집별 로지스틱 회귀분석으로 측정했다.

다음은 앞서 언급된 (1)~(5) 요인과 각각의 수준을 고려하여 실험계획법을 이용한 가설검정을 하였다. 이 중 (3) 입출력 함수는 주어진 데이터에서 알 수 없는 성격이므로 비제어인자로 간주하였다. 실험과정은, $5^1 \times 3^{2-1} \times 2^1$ 일부 요인 실험계획법을 사용하여 30개의 수준조합(treatment)마다 각 인자와 수준을 고려하여 난수 발생시켜 얻은 데이터를 학습용 데이터에 60 %, 검증용 데이터에 40 % 할당한 후, 비제어 인자를 고려한 실험디자인을 위하여 분류 정확성의 신호대 잡음비(S/N ratio : Signal to Noise Ratio)를 측정하였다. 여기서 신호대 잡음비는 실험디자인의 출력이 분류 정확성이므로 큰 값을 가질수록 바람직한 경우에 적용하는 망대특성((Lager-is-better characteristic)을 사용하였다. 다구찌 디자인을 이용한 실험결과는 <표 4>와 같다.

표 4. 일부 요인 실험계획법

	내측 배열				외측 배열		
	모형사용 방법	입력변수 간의 상관관계	데이터의 분산	데이터의 크기	분류정확성 (%) 함수의 종류 선형	분류정확성 (%) 함수의 종류 비선형	신호대 잡음비
1	Clustering	Strong	Large	Large	69.4	61.2	36.24
2	Clustering	Strong	Large	Small	68.3	62.9	36.31
3	Clustering	Medium	Medium	Large	635.	66.1	36.22
4	Clustering	Medium	Small	Large	60.4	59.0	35.51
5	Clustering	Weak	Medium	Small	62.7	61.4	35.85
6	Clustering	Weak	Small	Small	56.9	54.1	34.87
7	Bagging	Strong	Large	Small	69.1	61.6	36.26
8	Bagging	Strong	Medium	Large	58.9	64.6	35.78

9	Bagging	Medium	Medium	Small	64.3	62.3	36.02
10	Bagging	Medium	Small	Small	53.7	57.1	34.85
11	Bagging	Weak	Large	Large	70.9	68.6	36.86
12	Bagging	Weak	Small	Large	62.3	59.5	35.68
13	Parameter Combining	Strong	Large	Small	59.7	56.7	35.28
14	Parameter Combining	Strong	Medium	Large	52.3	50.1	34.17
15	Parameter Combining	Medium	Large	Small	70.8	51.0	35.34
16	Parameter Combining	Medium	Small	Small	52.7	51.8	34.36
17	Parameter Combining	Weak	Medium	Large	52.3	50.1	34.17
18	Parameter Combining	Weak	Small	Large	52.2	50.6	34.21
19	Variable Selection Bagging	Strong	Medium	Small	70.8	67.8	36.80
20	Variable Selection Bagging	Strong	Small	Small	60.8	58.8	35.53
21	Variable Selection Bagging	Medium	Large	Large	71.7	65.4	36.69
22	Variable Selection Bagging	Medium	Medium	Large	58.3	61.2	35.51
23	Variable Selection Bagging	Weak	Large	Large	68.3	64.2	36.41
24	Variable Selection Bagging	Weak	Small	Small	52.8	58.7	34.88
25	Logistic Regression	Strong	Small	Large	54.4	60.2	35.12
26	Logistic Regression	Strong	Small	Large	54.4	60.2	35.12
27	Logistic Regression	Medium	Large	Large	71.4	65.4	36.67
28	Logistic Regression	Medium	Medium	Small	57.9	62.4	35.56
29	Logistic Regression	Weak	Large	Small	68.6	64.6	36.45
30	Logistic Regression	Weak	Medium	Small	68.9	64.6	36.47

이와 같은 실험 결과를 이용하여 검정하려는 실험가설은 네 가지 주요인 효과와 더불어 다음과 같은 교호작용 효과이다.

Ha1 : 데이터의 크기는 분류 정확성에 유의한 영향을 준다.
Ha2 : 데이터의 분산이 작으면 로지스틱 회귀분석은 다른 네 가지 방법보다 분류 정확성이 높다.
Ha3 : 데이터의 분산이 중간이면 Bagging과 Parameter Combining 은 다른 네 가지 방법보다 분류 정확성이 높다.
Ha4 : 데이터의 분산이 크면 Clustering 방법과 다른 네 가지 방법보다 분류 정확성이 높다.
Ha5 : 입력변수 간의 상관관계가 높으면 Variable Selection Bagging은 다른 네 가지 방법보다 분류 정확성이 높다.

Ha1~Ha5와 같은 가설설정의 이유는 다음과 같다. 주 효과로 사용된 데이터의 크기는 기존의 시뮬레이션 연구에서 분류 정확성에 미치는 영향이 서로 다른 결과를 보였다. 따라서 이의 검증이 필요하다(Peterson *et al*, 1995; Sohn and Shin, 1999). 데이터의 분산이 클 때는 부트스트랩 샘플을 바탕으로 하기 때문에 일반적으로 강건한(robust) 모형을 만들 수 있다고 알려진 앙상블 방법(Bagging, Parameter Combining)이 우수한 성능을 발휘 할 것으로 기대되며 특히 Clustering 방법은 관측치의 산포가 클 때 군집을 이룰 가능성이 크므로 더욱 우수한 성능을 보일 것으로 예상된다. 또한 Variable Selection Bagging은 부트스트랩 샘플마다 입력변수 중 일부를 선택함으로 입력변수 간의 상관관계가 높을 경우, 분류 정확성이 높을 것으로 기대된다.

4. 다구찌 실험결과

위와 같은 Ha1~Ha4의 가설들에 대해 <표 4> 실험결과를 바탕으로 교호작용인 분류기법 × 입력변수 간의 상관관계 × 데이터의 분산을 오차항으로 두고 분산분석을 하여 유의수준 10 %에서 가설검정한 결과 분류기법, 입력변수 간의 상관관계, 데이터의 분산이 주 효과가 있으며 데이터의 크기는 분류 정확성에 유의한 영향을 주지 않았다(Ha1). 또한 분류기법 × 입력변수 간의 상관관계, 분류기법 × 데이터의 분산 등의 교호작용이 유의하였다(<표 5> 참고). <표 5>에 나타내지 않은 나머지 교호작용은 교락(confounding)된 것이나, 검정하고자 하는 가설이 디자인에 포함되도록 할당하였으므로 Ha1~Ha5의 가설검정에 문제가 없었다.

가설 Ha1~Ha5의 관점에서 유의한 주 효과와 교호작용을 바탕으로 데이터의 특성에 따른 적합한 분류방법을 선택하기 위하여 고차 교호작용을 중심으로 <표 6>, <표 7>과 같이 던칸 검정을 하였다.

<표 6>에 나타난 던칸 검정결과에 의하면 데이터의 분산 크기에 관계없이 모든 경우에 Parameter Combining 방법이 다른 방법에 비해 낮은 분류 정확성을 보이고 있으며, 가설 Ha2~Ha4

표 5. 실험 인자와 교호작용 대한 분산분석표

요 인	DF	Sum of Square	Mean Square	F-value	P-value
분류기법	4	8.335	2.0833	1488.07	*0.019
입력변수간의 상관관계	2	0.045	0.0227	16.21	0.172
데이터의 분산	2	5.3414	2.6707	1907.64	*0.016
데이터의 크기	1	0.0023	0.0023	1.64	0.422
분류기법× 입력변수간의 상관관계	8	4.2662	0.5332	380.85	*0.039
분류기법 × 데이터의 분산	6	0.7461	0.1243	88.78	*0.08
입력변수간의 상관관계 ×데이터의 분산	4	0.5147	0.1286	91.85	*0.07
E(분류기법×입력변수 간의 상관관계 × 데이터의 분산	1	0.0014	0.0014		

* *P* 값 < 0.1

표 6. 분류방법×데이터의 분산 간의 교호작용에 대한 던칸 검정 결과($\alpha = 0.1$)

던칸 그룹핑		분류방법	데이터의 분산	신호대 잡음비
	A A	Logistic Regression	Large	36.57
	A A	Bagging	Large	36.56
	A A	Variable Bagging	Large	36.55
B B	A A	Clustering	Large	36.28
B B	A A	Variable Bagging	Medium	36.16
B B	A A	Clustering	Medium	36.04
B B	A	Logistic Regression	Medium	36.02
B	C C	Bagging	Medium	35.90
D D	C	Parameter Combining	Large	35.32
D D		Bagging	Small	35.27
D D		Variable Bagging	Small	35.21
D D		Clustering	Small	35.20
D		Logistic Regression	Small	35.13
	E E	Parameter Combining	Small	34.29
	E	Parameter Combining	Medium	34.18

표 7. 분류방법×입력변수 간의 상관관계간의 교호작용에 대한 던칸검정 결과($\alpha = 0.1$)

던칸 그룹핑			분류방법	데이터의 분산	신호대 잡음비
	A A		Logistic Regression	Weak	36.46
B B	A A		Clustering	Strong	36.28
B B	A A		Bagging	Weak	36.27
B B	A	C C	Variable Selection Bagging	Weak	36.16
B B		C C	Logistic Regression	Medium	36.12
B B		C C	Variable Selection Bagging	Medium	36.10
B B		C C	Bagging	Strong	36.02
	D D	C	Clustering	Medium	35.87
E E	D		Variable Selection Bagging	Weak	35.64
E E	F F		Bagging	Medium	35.44
E	F F		Clustering	Weak	35.36
G G	F		Logistic Regression	Strong	35.12
G	H H		Parameter Combining	Medium	34.85
	H		Parameter Combining	Strong	34.73
	I		Parameter Combining	Weak	34.19

의 관점에서, 분산이 크거나 작을 때는 로지스틱 회귀분석, Variable Selection Bagging, Bagging, Clustering 방법 간에 유의한 분류성능 차이를 보이지 않았다. 그러나 분산이 중간 정도일 때는 Bagging과 Parameter Combining이 로지스틱 회귀분석, Variable Selection Bagging, Clustering에 비하여 상대적으로 떨어지는 것으로 나타났다. 이는 기존의 많은 연구에서 Bagging을 비롯한 앙상블 방법이 분류 정확성을 향상시킨다는 결과가 통계적으로 유의한 성능 차이를 보이는 것인지 검증해볼 필요가 있음을 제시한다. 또한 부트스트랩 샘플링에 기초한 앙상블 방법들이 분석에 소요되는 시간을 감안할 때 로지스틱 회귀분석과 같은 개별모형(Individual Model)에 비하여 효과적이지 못하다는 결론을 내릴 수 있다.

다섯번째 가설(Ha5)의 관점에서, Variable Selection Bagging 방법은 입력변수 간의 강한 상관관계를 가질 때 로지스틱 회귀분석이나 Parameter Combining 방법보다 상대적으로 우수한 분류 성능을 보였으며 Clustering이나 Bagging과는 유의한 성능 차이가 나지 않았다. 이는 입력 변수 간에 강한 상관관계를 가지는 경우 모든 변수를 이용하지 않아도 분류 정확성을

저해하지 않는 것을 의미한다. 따라서 교통량 추정, 품질 예측 문제에 있어서 센서의 설치비용을 절감할 수 있는 가능성을 제시한다. 이상의 결과는 기존의 경험적(empirical) 연구 중에서 앙상블 방법이 분류성능을 향상시키는 경우도 있으며 그렇지 못한 경우도 있었던 이유가 연구에 사용된 데이터의 변수간 상관관계와 분산의 정도에 따른 특징에 기인하는 것으로 볼 수 있다.

5. 결 론

본 연구에서는 로지스틱 회귀분석, Bagging, Variable Selection Bagging, Parameter Combining, Clustering 방법을 이용하여 분류분석을 할 때, 분류성능에 잠재적으로 영향을 미치는 데이터의 특성에 따라 적합한 분류방법을 알아보았다. 분류 정확성에 영향을 미치는 인자로 네 가지를 선택하고, 이 중 입출력변수 간의 연결함수는 주어진 자료에서 파악할 수 없는 성격이므로 다구찌 디자인을 이용하여 비제어 인자로 간주하고 실험하였다. 일부 요인 실험계획 결과, '분류방법×데이터의 분산', '분류방법×입력변수 간의 상관관계' 등의 교호작용이 분류성능에 유의한 영향을 미치는 것으로 나타났다($\alpha=0.1$). 이들 교호작용을 중심으로 분석결과를 정리하면, Parameter Combining방법이 낮은 분류 정확성을 보인 것을 제외하고, 나머지 앙상블 방법들과 로지스틱 회귀분석 사이에 유의한 성능차이가 나지 않았다. 이는 앙상블 방법이 부트스트랩 샘플링을 취하고 분류기를 만드는 데 소요되는 시간과 작업량을 고려하면 비효율적 일수도 있다는 결론을 내릴 수 있다. Variable Selection Bagging은 입력변수 간의 상관관계가 높을 경우, 로지스틱 회귀분석이나 Parameter Combining보다 우수하며 모든 변수를 사용하는 Bagging, Clustering 방법과 유의한 차이가 나지 않았다(Ha5). 이 결과는 여러 입력값을 동시에 감지하는 데 많은 비용이 소요되는 분야에서 유용히 활용할 수 있을 것으로 보인다. 예를 들어, 교통량 예측 분야에서는 여러 도로 상황변수를 동시에 센싱하기 위하여 한 지점에 다량의 센서를 설치함으로써 발생하는 비용 문제를 해결할 수 있는 대안이 될 수 있을 것이다. 이상의 시뮬레이션 연구에서 사용된 인자 및 수준은 데이터가 가질 수 있는 다양한 특성을 모두 포함하고 있다고 말할 수 없다. 또한 부트스트랩 샘플의 수에 따라 앙상블 모형의 성능이 달라질 수 있음을 고려하지 않았다. 따라서 향후 연구방향으로, 더욱 다양한 인자와 수준을 이용한 성능비교가 요구되며, 부트스트랩 샘플의 수 하나의 인자로 사용하여 앙상블의 효과를 파악할 필요가 있다.

참고문헌

Breiman, L. (1994), Bagging Predictor, *Technical Report,* 421, University of California at Berkeley.

Breiman, L. (1996), Arcing Classifiers, *Technical Report*, 486, University of California at Berkeley.

Cao, J., Ahmadi, M. and Shridhar, M. (1995), Recognition of Handwritten Numericals with Multiple Feature and Multistage Classifier, *Pattern Recognition*, **28**(2), 153~160.

Christodoulou, C. I. and Pattichis, C. S. (1998), Combining Neural Classifications in EMG Diagonosis, *EUFIT '98*, 1837-1841.

Domingos, P. (1999), MetaCost : A General Method for Making Classifiers Cost-Sensitive, *KDD-99* San Diego, CA USA, 155-164.

Freund, Y. and Schapire, R. E. (1996), Experiment with a New Boosting Algorithm, *Proceedings of the Thirteenth International Conference Machine Learning*, 148~156.

Guvenir, H. A. and Sirin, I. (1996), Classification by Feature Partitioning, Machine Learning, **23**, 47-67.

Ho, T. K., Hull, J. J. and Srihari, S. N. (1994), Decision Combination in Multiple Classifier Systems, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **16**(1), 66-75.

Kittler, J., Hatef, M., Duin Robert, P.W. and Matas, J. (1998), On Combining Classifiers, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **20**(3), 226~239.

Montgomery, D. C. (1997), *Design and Analysis of Experiments, Wiley, 4th*, 318~327.

Nezafat, R., Tabesh, A., Akhavan, S., Lucas, C. and Zia, M. A. (1998), Feature Selection and Classification for Diagnosing Breast Cancer *Proceedings of the IASTED International Conference Artificial Intelligence and Soft Computing*, 310~313.

Opitz, D. W. and Maclin, R. F. (1997), An Empirical Evaluation of Bagging and Boosting for Artificial Neural Netowrks, *Proceedings of the 1997 International Conference on Neural Networks*(ICNN'97), **3**, 1401~1405.

Peterson, G. E., Clair, D. 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.

Schapire, R. E. (1990), The Strength of Weak Learnability, *Machine Learning*, **5**(2), 197~227.

Shannon, W. D. and Banks, D. (1999), Combining Classification Trees Using MLE, *Proceeding of JSM*, Baltimore, U.S.A,

Sohn, S. Y. and Shin, H. W. (1998), Data Mining for Road Traffic Accident Type Classification, *Journal of Korean Society of Transportation*, **16**(4), 187-194,.

Sohn, S. Y. and Shin, H. W. (1999), Comparison of Data Mining Classification Algorithms for Categorical Feature Variables *Korea IE Interface*, **12**(4), 551~556.

Sohn, S. Y. and Lee, S. H. (2000), Data Fusion and Clustering for the Severity Classification of Road Traffic Accident in Korea, *Proceedings of NIEMS 2000.*