대한산업공학회/한국경영과학회 2000 춘계공동학술대회 논문집. 2000년 4월 21-22일. 경남대학교
Session PA10.2

다구찌 디자인을 이용한 데이터 퓨전 및 군집분석 분류 성능 비교

Comparison Study for Data Fusion and Clustering Classification Performances

연세대학교 컴퓨터 과학.산업시스템공학과
신형원, 손소영

Abstract

In this paper, we compare the classification performance of both data fusion 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. Since the relationship between input & output is not typically known, we use Taguchi design to improve the practicality of our study results by letting it as a noise factor. Experimental study results indicate the following: Clustering based logistic regression turns out to provide the highest classification accuracy when input variables are weakly correlated and the variance of data is high. When there is high correlation among input variables, variable bagging performs better than logistic regression. When there is strong correlation among input variables and high variance between observations, bagging appears to be marginally better than logistic regression but was not significant

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 데이터간의 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 작업중 가장 많이 사용되는 분류 모델링은 학습용(Training) 데이터로부터 입력과 출력간의 관계를 학습하고 이를 바탕으로 새로운(Test) 데이터를 분류하는데 적용된다. 이러한 분류 모델링의 가장 큰 이슈는 분류정확성의 향상이며 이를 위하여 몇 개의 부트스트랩 샘플에 단일모형을 여러 번 적용하여 분류한 결과를 융합해 주는 앙상블 방법에 대한 많은 연구가 있어왔다[1][3][4][5][9][11]. 또한 데이터의 분산이 큰 경우, 기존의 앙상블 방법과는 반대로 데이터를 특성에 따라 군집으로 나누고 각 군집별로 분류모형을 학습하는 Clustering 방법이 연구되었다[2][14]. 이처럼 주어진 데이터의 특성에 따라 적절한 앙상블 또는 군집분석 방법을 선택하는 일종의 메타모형은 그 중요성에도 불구하고 연구가 많이 되어 있지 않은 상황이다. 따라서 본 연
구에서는 Monte Carlo Simulation 을 이용하여 데이터의 특성을 나타내는 인자들과 앙상블, Clustering 방법간의 교호작용을 분류정확성의 관점에서 분석하고자 한다. 이를 위하여 데이터의 특성을 (1)입력변수간의 상관관계 (2)데이터의 분산 (3)데이터의 크기 (4) 입출력변수간의 함수로 나누고 (5)분류방법(로지스틱 회귀분석, Bagging, Variable Selection Bagging, Parameter Combining, Clustering)에 따라 이진 출력값에 대한 분류정확성을 비교하였다. 이들 요인중 입력변수와 출력변수간의 함수는 주어진 데이터에서 실제 알 수 없는 성격이므로 다구찌 실험계획법을 이용하여 비제어 인자로 간주하였다.

본 논문의 구성은 다음과 같다. 2 장에서는 앙상블 및 군집분석기법과 이에 관련된 기존문헌을 고찰하였으며 3 장에서는 본 연구에 사용된 실험계획법 및 실험 가설에 대하여 설명 하였다. 4 장에는 다구찌 실험계획을 이용 하여 실험한 결과를 정리하였으며 5 장에는 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 Tree 는 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 여러가지 앙상블 기법을 비교하였다. 앙상블 기법이란 다중 분류기들로부터 얻

은 예측값들을 결합하는 방법으로써 많은 연구자들이 하나의 분류기를 사용하는 경우보다 높은 분류성능을 얻기 위한 노력을 해왔다. 지금까지 보편적으로 알려져 있는 앙상블로는 Bagging, Arcing 을 등을 들 수 있다. 앙상블 방법을 이용한 분류에 대한 기존의 연구결과는 분류정확성을 높인 경우도 있었으며 오히려 낮춘 경우도 있었다. Breiman[1]은 Bagging 방법을 제안하고 시뮬레이션 데이터와 실제 데이터에 적용하여 분류정확성의 향상을 보였으며 Optiz &Maclin[9]은 14 개의 실제 데이터를 대상으로 Bagging 과 Boosting 을 이용하여 신경망 앙상블과 Decision Tree 앙상블을 만들어 분류하였다. 이들의 연구에서, Bagging 은 전반적으로 단일모형을 사용한 경우보다 분류정확성이 향상되었고 Boosting 은 경우에 따라 다른 분석결과를 보였다. 이밖에 분류정확성 향상을 위하여 Bagging, Boosting 을 이용한 방법 외의 다양한 방법이 시도 되어왔다. Christodoulou & Pattichis[3]은 근전도(EMG:Electromyographic) 신호로 3 가지 유형의 질병을 분류하기 위하여 72 개의 설명변수를 그 특징에 따라 6 개의 변수집합으로 만들어 6 개의 자기조직화 신경망(Self Organizing Map)으로 분류한 뒤 신뢰구간을 바탕으로 결합 하였다. Nezafat et al.[8]은 최근 이웃 학습법(K-Nearest Neighborhood), MLP(Multi Layer Perceptron) 신경망, RBF(Radial Basis Function) 신경망 등 6 가지 분류기의 특성에 따라 적절한 변수를 선택하여 학습한 후, 융합하는 방법을 연구하였다. Guvenir & Sirin[6]은 연속형 설명변수를 대상으로 설명변수의 값에 따라 구간별로 나눈 뒤 구간별 예측값을 구하고 각 설명변수의 예측값을 Voting 하는 휴리스틱을 개발하였다. Shannon & Banks[11]는 전체 데이터중 N 번의 샘플을 취하여 N 개의 Decision Tree 를 추정하고 N 개의 Decision Tree 와 가장 가까운 하나의 Decision Tree 를 최우추정법(Maximum Likelihood Estimation)을 이용하여 추정하는 Parameter Combining 방법을 제시하였다. Cao et al.[2]은 문자인식을 위해 비지도 신경망으로 데이터를 군집화 한 후, 역전파 신경망으로 군집별 학습을 하는 방법을 사용하였다. 손소영,이성호[14]는 교통사고 분류분석에 Bagging, Arcing, Demster-Shafer 이론 등, 다양한 앙상블 방법을 사용하여 분류정확성을 향상시키고자 하였으며 Clustering 분석을 이용한 군집별 학습 방법이 가장 분류정확성을 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[10][13].

3. 실험 디자인

본 장에서는 데이터의 특성에 비추어 예측능력이 높은 분류기법을 찾기 위한 실험의 인자 (Factor)와 수준(Level)을 정하였다. 실험에 사용된 모든 데이터는 5 개의 입력변수 이진값(Binary)을 가지는 출력변수를 가지고 있으며 그 특성을 (1)입력 변수간의 상관관계 (2) 데이터의 분산 (3)데이터의 크기 (4)입력과 출력변수사이의 함수로 나누고 각 시나리오 별로 (5) 앙상블모형에 따른 분류정확성을 분석하였다.

디자인에 사용된 각 요인별 수준을 자세히 살펴보면 다음과 같다.

← 입력변수간의 상관관계
5 개 입력변수간 상관관계가 약할 때의 피어슨 상관계수는 각각 0.05~0.09 사이이며 강할때의 피어슨 상관계수는 0.95~0.99 사이로 가정하였다.
↑ 데이터의 분산
다섯개 입력 변수의 평균을 0, 분산 공분산 행렬은 ←의 상관행렬(Correlation Matrix)에 각 각 1 과 100 을 곱하여 다중 정규(Multivariate Normal) 분포를 따르도록 하였다.
→ 입출력 변수간의 연결함수
시뮬레이션을 위하여 실제 모델로 사용한 입출력 변수간의 함수는 모수의 관점에서 로지스틱 선형인 경우와 로지스틱 비선형인 경우로 나누었다.
↓ 데이터의 크기
데이터 크기의 첫번째 수준은 "상대적으로 작은" 2000 개의 관측치를 가진 경우와 "상대적으로 많은" 10000 개의 관측치를 가진 데이터로 나누었다. 전체 데이터의 60%는 학습용 자료로, 40%는 검증용 자료로 사용하였다.

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

° 분류방법
실험에 사용된 분류방법은 전통적 통계분석 방법으로 오랜 기간 사용된 로지스틱 회귀분석과 앙상블 방법으로 가장 널리 알려진 Bagging, 일부 변수만을 번갈아 사용하므로 경제적인 분류 방법인 Variable Selection Bagging, Shanon&Banks[11]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[11].

- Bagging

본 실험에서는 16 개의 부트스트랩 샘플로 16 개의 로지스틱 분류기를 Bagging 하였다[1].

- Variable Selection Bagging

분류분석을 하는데 있어서 가능하면 작은 수의 변수를 사용하는 것이 경제적이다. 따라서 Variable Selection Bagging 분류기는 실험에서 사용된 5 개 설명변수 중 3 개씩 랜덤하게 사용하여 16 번의 부트스트랩 샘플을 바탕으로 16 개의 로지스틱 분류기를 Bagging 하는 방법이다[8].

- Parameter Combining

Parameter combining 분류기는 16 번의 부트스트랩 샘플로 16 개의 로지스틱 분류기 을 $\hat{p}_i$ 만들어 이 때 추정된 모수를 바탕으로 오차를 줄이는 대표적인 로지스틱 회귀모형(P)을 재 추정하여 이를 바탕으로 자료를 다시 분류하는 방법이다[11].

- Clustering

이 방법은 학습용 데이터를 K-평균법을 이용하여 4 개의 군집으로 나누고 각 군집별로 로지스틱 회귀분석을 이용한 학습을 하는 방법이다[2][14]. 분류정확

성의 측정은 검증용 데이터를 학습용 데이에 근거하여 4 개의 군집으로 나누고 군집별 로지스틱 회귀분석으로 측정했다.

다음은 앞서 언급된 ←~° 요인과 각각의 수준을 고려하여 실험계획법을 이용한 가설검정을 하였다. 이 중 ↓ 입출력 함수는 주어진 데이터에서 알 수 없는 성격이므로 비제어인자로 간주하였다. 실험과정은, $2^{3-1}\times5^{1}$ 일부 요인 실험계획법을 사용하여 20개의 treatment 마다 각 인자와 수준을 고려하여 난수 발생시켜 얻은 데이터를 학습용 데이터에 60%, 검증용 데이터에 40% 할당하고 분류 정확성의 신호대 잡음비(Signal to Noise Ratio)를 측정하였다.
이와 같은 실험 결과를 이용하여 검정하려는 아홉 개의 실험가설은 다음과 같다.

· 주효과에 의한 가설
Ha1: 입력변수간의 상관관계는 분류정확성에 영향을 미친다.
Ha2: 데이터의 분산은 분류정확성에 영향을 미친다.
Ha3: 데이터의 크기는 분류정확성에 영향을 미친다.
Ha4: 분류모형의 사용방법은 분류정확성에 영향을 미친다.
· 교호작용에 의한 가설
Ha5: 입력변수간의 상관관계가 크면 Variable Selection Bagging 방법은 로지스틱 회귀분석과 분류성능에 차이가 없다.
Ha6: 데이터의 분산이 크면 Clustering 방법은 다른 네 가지 방법보다 분류정확성이 높다.
Ha6: 데이터의 분산이 크면 Parameter Combining 방법은 다른 네 가지 방법보다 분류정확성이 높다.
Ha7: 입력변수간의 상관관계가 높고 데이터의 분산이 크면 Variable Bagging 이나 Bagging 은 로지스틱 회귀분석보다 분류정확성이 높다.

4. 다구찌 실험결과

본 장에서는 위와 같은 Ha1~:Ha7 의 가설검정을 위한 들을 실험결과를 바탕으로 분산분석을 하여 유의수준 10%에서 가설검정 결과 모형사용방법, 입력변수간의 상관관계, 데이터의 분산이 주효과가 있으며 데이터의 크기는 분류 정확성에 유의한 영향을 주지 않는 것으로 나타났다. 한편 모형사용방법× 입력변수간의 상관관계, 모형사용방법× 데이터의 분산, 모형사용방법× 입력변수간의 상관관계× 데이터 간에는 교호작용이 있는 것으로 나타났다. 유의한 주효과와 교호작용을 바탕으로 데이터의 특성에 따른 적합한 분류방법을 선택하기 위하여 고차 교호작용을 중심으로 던칸 검정을 하였다.

<표 1> 분류방법× 입력변수간의 던칸검정 결과

던칸 그룹핑	분류방법	입력변수간의 상관관계	신호대 잡음비
A A	Clustering	Weak	36.58
B A B A	Variable Bagging	Strong	36.23
B A	Logistic Regression	Weak	36.19
B A	Regression		
B A C B A C	Bagging	Weak	36.11
B A C B A C	Clustering	Strong	35.96
B A C B A C	Bagging	Strong	35.94
B D A C B D C	Logistic Regression	Strong	35.73
B D C D C	Variable Bagging	Weak	35.61
D C D	Parameter Combining	Strong	35.25
D	Parameter Combining	Weak	34.78

<표 1>에 나타난 던칸 검정결과에 의하면 입력변수간의 상관정도가 강할 경우 Variable Bagging 방법이 로지스틱 회귀분석 보다 분류정확성이 높은 것으로 나타났다. 반면에 입력변수간의 상관관계가 약할 경우는 Variable Bagging 이 로지스틱 회귀분석에 비하여 유의하게 낮은 분류정확성을 보였다. 이는 데이터의 특성이 입력변수간에 강한 상관관계를 가지는 경우 모든 변수를 이용하지 않아도 분류정확성을 저해하지 않는 것을 의미한다. 따라서 교통량 추정, 품질 예측문제에 있어서 센서의 설치비용을 절감할 수 있는 가능성을 제시한다. 또한 Parameter Combining 방법은 입력변수간 상관관계가 약할 때 나머지 네 가지 방법에 비하여 분류정확성이 떨어지는 것으로 나타났다.

<표 2> 분류방법× 입력변수간의 상관관계× 데이터의 교호작용에 대한 던칸검정 결과

던칸 그룹핑	분류방법	입력변수간의 상관관계	데이터의 크기	신호대 잡음비
A A	Clustering	Weak	Large	36.9965
B A B A	Variable Bagging	Strong	Small	36.5481
B A B A	Bagging	Weak	Large	36.4338
B A B A	Logistic Regression	Weak	Large	36.4109
B A B A	Variable Bagging	Weak	Large	36.3182
B A B A	Parameter Combine	Strong	Large	36.1867
B A B A	Clustering	Weak	Small	36.1649
B A C B A C	Clustering	Strong	Small	36.0940
B A C B A C	Bagging	Strong	Small	36.0104
B A C B A C	Logistic Regression	Weak	Small	35.9660
B A C B A C	Variable Bagging	Strong	Large	35.9139
B A C B A C	Bagging	Strong	Large	35.8784

B B	A A	C C	Clustering	Strong	Large	35.8173
B B	A	C C	Bagging	Weak	Small	35.7867
B B		C C	Logistic Regression	Strong	Small	35.7396
B B		C C	Logistic Regression	Strong	Large	35.7173
B	D D	C C	Parameter Combine	Weak	Large	35.5526
E E	D D	C	Variable Bagging	Weak	Small	34.8963
E E	D		Parameter Combine	Strong	Small	34.3229
E			Parameter Combine	Weak	Small	34.0141

<표 2>에 나타난 던칸 검정 결과에 의하면 입력변수간의 상관관계가 약하고 데이터의 분산이 크면 Clustering 방법은 나머지 네 가지 분류방법보다 분류정확성이 높은 것으로 나타났다. 이는 데이터가 분산이 클 경우, 같은 특성을 가진 군집별로 학습을 하는 것이 효과가 있음을 의미한다. 또한 입력변수간 상관관계가 강하고 데이터의 분산이 큰 경우는 Variable Bagging 과 Bagging 이 로지스틱 회귀분석 보다 다소 높은 분류정확성을 나타내기는 했으나 통계적으로 유의한 차이는 나지 않았다. 이는 기존의 많은 연구에서 Bagging 을 비롯한 앙상블 방법이 분류정확성을 향상시킨다는 결과가 통계적으로 유의한 성능 차이를 보이는 것이지 검증해볼 필요가 있음을 제시한다. 이밖에 Parameter Combining 방법은 기대했던 것처럼 데이터의 분산이 클 경우 높은 분류정확성 보였다.

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. 분류정확성에 영향을 미치는 인자로 네가지를 선택하고, 이중 입출력 변수간의 연결함수는 주어진 자료에서 파악할 수 없는 성격이므로 다구찌 디자인을 이용하여 비제어 인자로 간주하고 실험하였다. 일부요인 실험계획 결과, 유의수준 10% 에서 가설검정 결과 입력변수간의 상관관계, 데이터의 분산은 분류정확성에 영향을 주며 데이터의 크기는 분류 정확성에 유의한 영향을 주지 않는 것으로 나타났다. Clustering 방법은 입력변수간의 상관관계가 약하고 데이터의 분산이 크면 나머지 네 가지 분류방법보다 분류정확성이 높은 것으로 나타났으며 입력변수간의 상관정도가 강할 경우는 Variable Bagging 방법이 로지스틱 회귀분석 보다 분류정확성이 높은 것으로 나타났다. 이 결과는 여러 입력값을 동시에 감지하는데 많은 비용을 소요되는 분야에서 유용히 활용할 수 있을 것으로 보인다. 예를 들어 교통량 예측 분야에서는 여러 도로 상황변수를 동시에 센싱하기 위하여 한 지점의 다량에 센서를 설치함으로써 발생하는 비용 문제를 해결할 수 있는 대안이 될 수 있을 것이다. 또한 입력변수간 상관관계가 강하고 데이터의 분산이 큰 경우는 Bagging 방법이 단일 모형을 사용한 로지스틱 회귀분석과 비교하여 분류 정확성이 다소 높게는 나타났으나 유의수준 10%에서 유의한 성능 차이는 나지 않았다. 향후 연구방향으로, 본 연구에서 상대적으로 높은 분류정확성을 보인 Clustering 방법의 군집 개수를 다양하게 변화시켜 실험함으로써 더욱 분류성능을 높일 수 있는가 검증하고, 로지스틱 회귀분석이외에 신경망, Decision Tree 사용한 결과와 비교할 것을 과제로 하고 있다.

참고문헌

[1] Breiman, L.(1994), "Bagging Predictor", Technical Report No.421, University of California at Berkeley.

[2] Cao , J. Ahmadi, M., Shridhar, M.(1995), "Recognotion of Handwritten Numericals with Multiple Feature and Multistage Classifier", Pattern Recognition, 28(2), pp153~160.

[3] Christodoulou, C.,I. & Pattichis, C.,S.(1998), "Combining Neural Classifications in EMG Diagonosis", EUFIT '98, pp1837-1841.

[4] Domingos, P.(1999), "MetaCost: A General Method for Making Classifiers Cost-Sensitive", KDD-99 San Diego, CA USA, pp155-164.

[5] Freund, Y., Schapire, R. E.(1996), "Experiment with a New Boosting Algorithm" , In Machine Learning:Proceedings of the Thirteenth International Conference, pp 148~156.

[6] Guvenir, H.A. &Sirin, I.(1996), "Classification by Feature Partitioning", Machine Learning, 23, pp47-67.

[7] 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), pp226~239.

[8]Nezafat, R., Tabesh, A., Akhavan, S.,Lucas, C., 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.

[9]Opitz, D. W., 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, pp1401~1405.

[10] 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), pp.949-960.

[11] Shannon, W.D.& Banks, D.(1999), "Combining Classification Trees Using MLE", JSM, Baltimore, USA,

[12] Sohn , S. Y. & Shin, H. W.(1998), "Data Mining for Road Traffic Accident Type Classification ", Journal of Korean Society of Transportation, 16(4), pp.187-194,.

[13] Sohn , S. Y. & Shin, H. W.(1999), "Comparison of Data Mining Classification Algorithms for Categorical Feature Variables" Korea IE Interface, 12(4), pp551~556. ,

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