매장문화재 예측을 위한 통계적 분류 분석

(Classification Analysis for the Prediction of Underground Cultural Assets)

유 혜 경*, 이 진 영**, 나 종 화***
(Hye-Kyung Yu, Jin-Young Lee, Jonghwa Na)

요 약 본 논문에서는 통계적 분류방법을 이용하여 문화재 자료의 분석을 수행하였다. 분류방법으로는 선형판별분석, 로지스틱회귀분석, 의사결정나무분석, 신경망분석, SVM분석을 사용하였다. 각각의 분류방법에 대한 개념 및 이론에 대해 간략히 소개하고, 실제자료 분석에서는 국내 I시 자료를 사용하여 매장문화재에 대한 분류방법별 적합모형을 구축하였다. 구축된 모형에 대한 성능비교와 함께, 새로운 자료에 대한 적용성 평가를 위해 모의실험을 수행하였다. 분석에 사용된 도구로는 최근 가장 관심을 갖는 R 언어를 사용하였으며, 구체적 분석과정을 제시하였다.

핵심주제어 : 선형판별분석, 로지스틱회귀분석, 의사결정나무분석, 신경망분석, SVM

Abstract Various statistical classification methods have been used to establish prediction model of underground cultural assets in our country. Among them, linear discriminant analysis, logistic regression, decision tree, neural network, and support vector machines are used in this paper. We introduced the basic concepts of above-mentioned classification methods and applied these to the analyses of real data of I city. As a results, five different prediction models are suggested. And also model comparisons are executed by suggesting correct classification rates of the fitted models. To see the applicability of the suggested models for a new data set, simulations are carried out. R packages and programs are used in real data analyses and simulations. Especially, the detailed executing processes by R are provided for the other analyser of related area.

Key Words : Linear Discriminant Analysis, Logistic Regression Analysis, Decision Tree Analysis, Neural Network Analysis, Support Vector Machines

1. 서 론

현재 우리나라에서는 다양하고 복잡한 자료들에 대한 통계적인 분석이 이루어지고 있다. 그 중 우리나라의 문화유적 분야에서는 매장문화재 예측형으로 여러 통계적 분류방법(선형판별분석, 로지스틱 회귀분석, 의사결정나무분석, 신경망분석, SVM)이 사용되고 있다. 이들 모형의 특징은 문화재 자료(1:매장, 0:지상)와 같은 이진반응변수(binary response variable)의 모형화에 효과적으로 이다. 선형판별분석(Linear Discriminant Analysis)은 통계적으로 수학적인 판별식의 해석적인 모형구축이 가능하며 모형 갱신이 용이한 점을 가지나 범주형 입력변수의 사용에 제한이 따르는 단점을 가진다. 로지스틱회귀분석(Logistic Regression Analysis)은 명확한 수학적 모형을 제시하며 유용한 추가정보(회귀계수, 오즈비 등)를 제공하고, 모형의 해석적인 측면에는 유리하지만 최적의 변수선택을 위

* 충북대학교 자연과학대학 정보통계학과 박사과정
** 한국지질자원연구원 지구환경연구본부 선임연구원
*** 충북대학교 자연과학대학 정보통계학과 교수(교신저자 : cherin@cbu.ac.kr)

한 노력과 분석절차에 대한 전문성이 요구되며, 비선형적인 분석에 대한 한계를 가지고 있다. 의사결정나무분석(Decision Tree Analysis)은 빠르고 편리한 예측 기준을 제시하지만, 계층적 구조 때문에 결과가 불안정할 수 있다[6]. 신경망분석(Neural Network Analysis)은 예측력의 측면에서는 우수하나, 망의 구축에 전문성이 요구되며, 결과에 대한 설명력이 떨어진다. SVM(Support Vector Machines)은 명확한 이론적 근거를 가지며 예측력이 뛰어난 장점을 가지는 반면 모형식의 구체적 표현에 어려움이 있어, 모형의 이식성이 결여되며 복잡한 이론으로 인해 적용성에 한계가 있다. 이러한 5가지 분류방법에 대해 2장에서는 개념과 이론에 대하여 설명하였고, 3장에서는 문화재 자료를 근거로 R을 이용한 실제 자료 분석을 통해 각각의 분석 결과를 정리 및 해석하였다. 4장에서는 모형의 성능에 대한 측도에 대해 간략한 소개와 3장에서 구축된 모형과 모의실험의 결과를 통해 모형의 성능을 비교하였다. 5장에서는 본 연구의 결론을 제시하였다.

2. 분류방법

2.1 선형판별분석

간격척도 이상인 독립변수의 선형결합과 케이스가 속한 집단을 나타내는 명목척도인 종속변수와의 관계를 규명하여 독립변수의 값을 이용하여 케이스가 속한 집단을 예측하여 집단을 분류하는 방법이다. 여기서 독립변수의 선형결합을 표현하면

$$Z=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p \qquad (1)$$

와 같고, 판별점수(Z)의 집단간 분산과 집단내 분산을 분산 비율을 최대화 하는 절편과 계수를 도출하여, 각각의 케이스의 판별점수를 구한 후 이를 이용해 종속변수의 규정된 각 집단의 판별점수의 평균과 비교하여 거리가 가까운 집단으로 분류한다. 이때, 다수의 독립변수 값의 조합이 정규분포를 이루고, 각 집단들의 변수간 분산-공분산 행렬은 동일하다는 가정이 필요하다[3].

2.2 로지스틱회귀분석

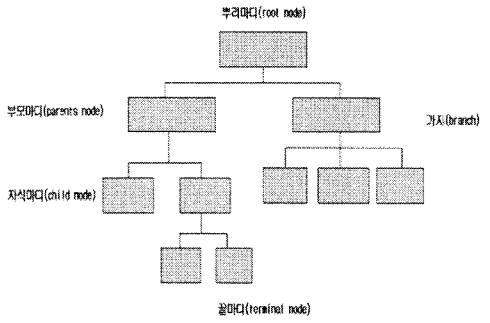
종속변수가 이진형(또는 범주형)이고, 독립변수가 범주형이나 연속형 또는 혼합형인 경우에 적용되는 회귀분석 기법으로 선형판별 분석과 유사한 결과를 준다[6]. 만약 주어진 x에서 반응변수(y)가 성공할 확률은 $p=P\{y=1|x\}$이라 할 때, 로지스틱회귀 분석은 $\ln(p/1-p)$를 설명변수들의 선형결합으로 모형화 하는 기법이다. 즉, 로지스틱회귀 모형은

$$\ln\left(\frac{p_i}{1-p_i}\right)=\beta_0+\beta_1x_{1i}+\cdots+\beta_kx_{ki},\ i=1,\cdots,n \quad (2)$$

으로 표현된다. 식 (2)의 좌변은 반응변수가 성공할 오즈(odds)의 로그값을 나타내며 이를 로짓(logit)함수 즉, $logit(p_i)$로 나타내기도 한다.

2.3 의사결정나무분석

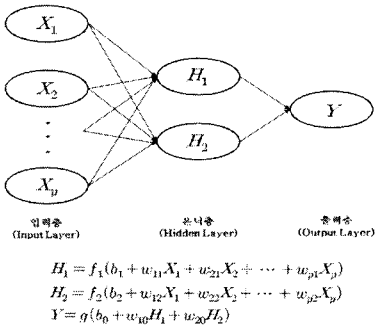
의사결정규칙(decision rule)을 나무구조로 도표화하여 분류와 예측을 수행하는 방법이다. 의사결정나무 모형의 적합과정은 분석의 목적과 자료구조에 따라 적절한 분리기준과 정지규칙을 지정하여 모형을 설정하고, 분류오류(classification error)를 크게 할 위험이 높거나 부적절한 추론규칙을 가진 가지(branch)를 제거한다[5]. 다음으로 이익도표(gains chart)나 위험도표, 검증용 자료에 대한 교차타당성을 이용하여 타당성 평가를 통해 의사결정나무를 해석하고 분류 및 예측모형을 설정하게 된다. 의사결정나무 모형의 구조는 다음의 (그림 1)과 같다.



(그림 1) 의사결정나무 모형

2.4 신경망분석

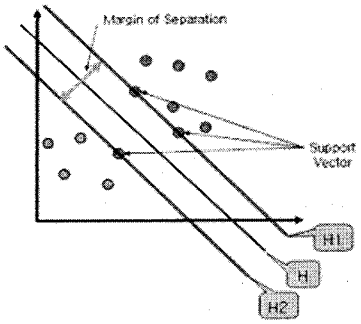
신경세포를 모방한 개념으로 과거에 수집된 자료로부터 반복적인 학습과정을 거쳐 자료에 내제되어있는 패턴을 찾아내는 분석기법이다. 이러한 신경망 모형 중 MLP(Multilayer perceptron)신경망이 가장 많이 사용되며 그 구조는 다음의 (그림 2)와 같다. 이러한 구조를 통해 은닉마디에서 입력층의 입력변수들의 결합을 수신하여 목표변수에 전달한다[4]. 이 때 결합에 사용되는 계수(연결강도)를 활성함수를 통해 입력값을 변환하고 이를 입력으로 사용하는 다른 마디로 출력하게 된다. 여기서 입력층 또는 은닉층의 마디들을 결합해주는 함수를 결합함수(H_1, H_2)라 하고, 입력변수 또는 은닉마디의 결합을 변환하는 함수를 활성함수라고 한다. 활성함수(f_1, f_2, g)에는 로지스틱함수와 쌍곡탄젠트 함수가 사용된다.



(그림 2) MLP 신경망 모형

2.5 Support Vector Machines

분류문제를 해결하기 위한 최적의 분리경계면을 제공하는 비모수적 방법으로 구조적 위험(structural risk)을 최소화 하려는 원칙을 구현한 방법이다. SVM 알고리즘은 두 개의 클래스가 있을 때, 다음의 (그림 3)과 같이 초평면에서 가장 가까운 데이터까지의 거리(margin)를 최대화 시키는 방법으로 여기서 초평면을 결정하는 입력패턴을 서포트 벡터(support vector)라고 한다.



(그림 3) SVM의 초평면

3. 문화재 자료 분석

3.1 분석자료 및 R 소개

본 연구에 사용된 자료는 "지역별 문화재 통계분석 및 모형개발 연구(2008)[1]"에 사용된 문화재 자료로 그 중 일부인 I시 자료를 사용하였다. I시 자료를 정리하면 아래의 <표 1>과 같다.

<표 1> 문화재 자료 구분

	구 분	범 주 설 명	빈 도(%)	전체%
문화재 자료 (n=173)	매장 문화재 (n_1=83)	무덤유적	13(15.66%)	7.51%
		유물산포지	63(75.90%)	36.42%
		산업생산유적	7(8.44%)	4.04%
		소 계	83(100%)	-
	지상 문화재 (n_2=90)	건축/건축물	27(30.00%)	15.61%
		종교/신앙유적	42(46.67%)	24.28%
		정치/국방유적	15(16.67%)	8.68%
		위인선현유적	6(6.66%)	3.46%
		소 계	90(100%)	-
전체			173	100%

문화재 자료에 포함된 변수는 총 53개(범주형: 26, 연속형: 27)로 분석을 위해 변수선택법(stepwise, forward, backward)과 주요변수 추가를 통해 분석에 사용될 변수는 다음의 <표 2>와 같이 선택되었다. 이를 통해 분류방법별 매장문화재 예측모형을 구축하고자 한다.

<표 2> 분석에 사용된 변수

구분		변수명	변수설명
목표변수		CULTURE	문화재 구분: 매장문화재(1), 지상문화재(0)
설명변수	범주형	SOIL_CAT	주토양분: 6개 범주
		GEO_CAT	주지질명: 2개 범주
	연속형	ALT_AVG	유적의 평균고도 (단위: m)
		SLP_MIN	유적의 최소경사 (단위: °, 범위: 0°~90°)
		SLP_MAX	유적의 최대경사 (단위: °, 범위: 0°~90°)
		W_ALTMIN	수계로부터의 최소고도 (단위: m)
		W_DISTMIN	수계로부터의 최소거리 (단위: m)
		W_DISTMAX	수계로부터의 최대거리 (단위: m)
		DENSITY	유적의 밀도(유적지의 1km 반경내 유적수)

3.2 R을 통한 실제자료 분석 결과

3.2.1 선형판별분석

R을 통한 선형판별 분석을 수행하기 위해 MASS 패키지에 있는 lda() 함수를 사용한다.

```
lda.res<-lda(CULTURE ~ ALT_AVG+ SLP_MIN+ SLP_MAX+
 W_ALTMIN+ W_DISTMIN+ W_DISTMAX+ SOIL_CAT+ GEO_
 CAT+ DENSITY, data=x)
```

위의 결과를 통해 얻어진 선형판별 모형의 판별함수는 다음과 같다.

$$Z = -0.0088 \cdot ALT\_AVG - 0.0632 \cdot SLP\_MIN - 0.0115 \cdot SLP\_MAX + 0.0191 \cdot W\_ALTMIN - 0.0074 \cdot W\_DISTMIN + 0.0058 \cdot W\_DISTMAX - 0.8399 \cdot SOIL\_CAT2 - 1.8263 \cdot SOIL\_CAT4 + 0.3865 \cdot SOIL\_CAT5 - 0.5277 \cdot SOIL\_CAT6 + 0.4531 \cdot GEO\_CAT2 + 0.0486 \cdot DENSITY$$

3.2.2 로지스틱회귀분석

로지스틱회귀분석을 수행하기 위해 nnet 패키지에 있는 glm() 또는 multinom() 함수를 사용한다 [8].

```
>logis.res<-glm(CULTURE ~ ALT_AVG+ SLP_MIN+ SLP_
 MAX+ W_ALTMIN+ W_DISTMIN+ W_DISTMAX+ SOIL_
 + CAT+ GEO_CAT+ DENSITY,data=x, family=binomial(link
 ="logit"))
```

glm() 함수를 수행한 결과는 다음과 같다.

```
                Estimate    Pr(>|z|)
(Intercept)   -1.103e+00   0.318870
ALT_AVG       -1.761e-02   0.019371*
SLP_MIN       -1.934e-01   0.074810
SLP_MAX       -1.487e-02   0.530751
W_ALTMIN       3.409e-02   0.113410
W_DISTMIN     -1.086e-02   0.000652***
W_DISTMAX      8.987e-03   0.000655***
SOIL_CAT2     -1.042e+00   0.365417
SOIL_CAT4     -1.623e+01   0.990012
SOIL_CAT5      5.061e-01   0.604901
SOIL_CAT6     -4.550e-01   0.700487
GEO_CAT2       5.407e-01   0.396362
DENSITY        8.766e-02   0.091572

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

위의 결과를 통해 ALT_AVG, W_DISTMIN, W_DISMAX 변수가 가장 유의한 변수로 나타났고, 로지스틱회귀 모형식은 다음과 같다.

$$logit(p)(\text{또는} = \log[p/(1-p)]) = -1.1030 - 0.0176 \cdot ALT\_AVG - 0.1934 \cdot SLP\_MIN - 0.0149 \cdot SLP\_MAX + 0.0341 \cdot W\_ALTMIN - 0.0108 \cdot W\_DISTMIN + 0.0090 \cdot W\_DISTMAX - 1.0420 \cdot SOIL\_CAT2 - 16.2300 \cdot SOIL\_CAT4 + 0.5061 \cdot SOIL\_CAT5 - 0.4550 \cdot SOIL\_CAT6 + 0.5407 \cdot GEO\_CAT2 + 0.0877 \cdot DENSITY$$

3.2.3 의사결정나무분석

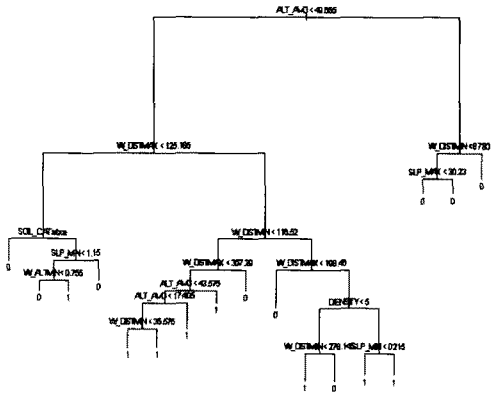
의사결정나무분석을 수행하기 위해 tree 패키지에 있는 tree() 함수를 사용한다.

```
>tree.res<-tree(CULTURE ~ ALT_AVG+ SLP_MIN+ SLP_
 MAX+ W_ALTMIN+ W_DISTMIN+ W_DISTMAX+ SOIL_
 CAT+ GEO_CAT+ DENSITY, data=x)
```

tree() 함수를 수행한 결과는 다음과 같다.

```
node), split, n, deviance, yval, (yprob)
      * denotes terminal node
 1) root 173 239.500 0 ( 0.52023 0.47977 )
   2) ALT_AVG < 49.665 128 170.300 1 ( 0.38281 0.61719 )
                          <중략>
     13) SLP_MAX > 30.23 7    0.000 0 ( 1.00000 0.00000 ) *
   7)W_DISTMIN >87.83 27 0.000 0 ( 1.00000 0.00000 ) *
  Number of terminal nodes:  17
  Misclassification error rate: 0.1098 = 19 / 173
```

위의 결과를 통해 적합된 의사결정나무 모형은 끝마디(*: terminal node)가 17개이고, 오분류율이 10.98%라는 것을 알 수 있었다. 적합된 의사결정나무 모형을 그림으로 표현하면 다음의 (그림 4)와 같다.



(그림 4) 구축된 의사결정나무 모형

3.2.4 신경망분석

신경망분석을 수행하기 위해 nnet 패키지에 있는 nnet() 함수를 사용한다.

```
>nn.res<-nnet(CULTURE~ALT_AVG+SLP_MIN+SLP_
 MAX+W_ALTMIN+W_DISTMIN+W_DISTMAX+SOIL_
 CAT+GEO_CAT+DENSITY,    data=x,    decay=0.01,
size=2,
 maxit=100)
```

nnet() 함수를 수행한 결과는 다음과 같다.

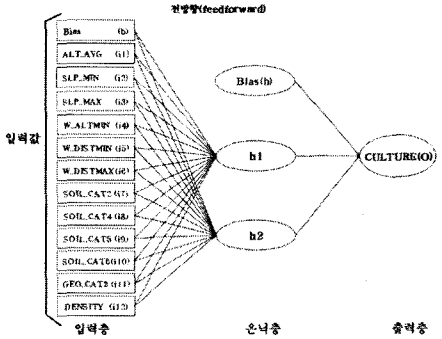
```
 a 12-2-1 network with 29 weights
inputs : ALT_AVG    SLP_MIN    SLP_MAX    W_ALTMIN
         W_DISTMIN  W_DISTMAX  SOIL_CAT2  SOIL_CAT4
         SOIL_CAT5  SOIL_CAT6  GEO_CAT2   DENSITY
output(s) : CULTURE
```

위의 결과를 통해 입력층의 입력변수가 총12개이고, 하나의 은닉층에 2개의 은닉마디가 있고, 출력층을 갖는 MLP모형임을 확인하였다. 또한 신경망 모형식은 다음과 같다.(H_2는 생략)

$$f(x)=\frac{1}{1+e^{-x}}\text{(로지스틱 함수)},\ g(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}\text{(쌍곡탄젠트 함수)}$$

$$\begin{cases} H_1 = f(-1.00-1.13\cdot ALT\_AVG+1.68\cdot SLP\_MIN+0.58\cdot SLP\_MAX \\ \quad +0.09\cdot W\_ALTMIN-0.70\cdot W\_DISTMIN-0.14\cdot W\_DISTMAX \\ \quad +1.06\cdot SOIL\_CAT2+0.30\cdot SOIL\_CAT4-2.42\cdot SOIL\_CAT5 \\ \quad +0.07\cdot SOIL\_CAT6-1.00\cdot GEO\_CAT2-0.27\cdot DENSITY) \\ \hat{P}(Y=1)=g(-2.37-1.95H_1+3.95H_2) \end{cases}$$

적합된 신경망 모형을 그림으로 표현하면 다음의 (그림 5)와 같다.



(그림 5) 구축된 신경망 모형

3.2.5 Support Vector Machines

SVM을 수행하기 위해 e1071패키지에 있는 svm 함수를 사용한다[7]. 우선 모형 적합에 사용할 gamma(커널모수)값과 cost(정규화상수)값을 tune() 함수를 이용해 구한다.

```
>tune<-tune.svm(CULTURE~ALT_AVG+SLP_MIN+SLP_
 MAX+W_ALTMIN+W_DISTMIN+W_DISTMAX+SOIL_
 CAT+GEO_CAT+DENSITY,  data=x,  gamma=seq(0.1,  1,
0.1)
 ,cost=seq(0.1,1,0.1))
```

tune() 함수를 수행한 결과 gamma(커널모수)가 0.4, cost(정규화상수)가 1.0으로 선택되었다. 이를 이용하여 svm() 함수를 수행한다.

```
>svm.res<-svm(CULTURE~ALT_AVG+SLP_MIN+SLP_
 MAX+W_ALTMIN+W_DISTMIN+W_DISTMAX+SOIL_
 CAT+GEO_CAT+DENSITY,                    data=x,
type="C-classification"
 , kernel="radial", cost=1, gamma=0.4)
```

svm() 함수를 수행한 결과는 다음과 같다.

```
Number of Support Vectors:  125
    ( 71 54 )

Number of Classes:  2

Levels:
 0 1
```

위의 결과를 통해 데이터의 총 케이스 173개 중 서포트 벡터의 수는 총 125개로 모형의 분리에 어려움이 있고, 자료에 적합지 않은 모형으로 평가

할 수 있다.

4. 모형의 성능비교

4.1 적합모형의 성능 비교

4.1.1 모형의 성능 비교 측도

다음의 <표 3>을 이용하여 모형의 적합도를 평가하는 측도인 정분류율, 예측정확도를 평가하는 민감도와 특이도를 구한다. 만약 적합된 모형이 유의한 경우라면 이 세 가지 측도가 동시에 높아야 한다.

<표 3> 관측결과와 예측결과의 분류표

관측결과 \ 예측결과	$\hat{Y}=1$	$\hat{Y}=0$	행 합계
$Y=1$	n_{11}	n_{10}	n_{1+}
$Y=0$	n_{01}	n_{00}	n_{0+}
열 합계	n_{+1}	n_{+0}	n

- 정분류율 $= \left(\frac{n_{11}+n_{00}}{n}\right)\times 100\%$
- 민감도 $= \left(\frac{n_{11}}{n_{1+}}\right)\times 100\%$ · 특이도 $= \left(\frac{n_{00}}{n_{0+}}\right)\times 100\%$

4.1.2 적합모형의 성능 비교결과

3장에서 적합된 모형의 정분류율, 민감도, 특이도를 정리하면 다음의 <표 4>와 같다.

<표 4> 적합모형의 성능 비교표

구 분	실제집단	예측된 집단		정분류율	민감도	특이도
		매장문화재 (1)	지상문화재 (0)			
선형판별 모형	매장문화재 (1)	69 (83.13%)	14 (16.87%)	80.92%	83.13%	78.89%
	지상문화재 (0)	19 (21.11%)	71 (78.89%)			
로지스틱 회귀 모형	매장문화재 (1)	70 (84.34%)	13 (15.66%)	83.82%	84.34%	83.33%
	지상문화재 (0)	15 (16.67%)	75 (83.33%)			
의사결정 나무 모형	매장문화재 (1)	73 (87.95%)	10 (12.05%)	89.01%	87.95%	90.00%
	지상문화재 (0)	9 (10.00%)	81 (90.00%)			
신경망 모형	매장문화재 (1)	74 (89.16%)	9 (10.84%)	84.39%	89.16%	80.00%
	지상문화재 (0)	18 (20.00%)	72 (80.00%)			
SVM	매장문화재 (1)	80 (96.39%)	3 (3.61%)	92.49%	96.39%	88.89%
	지상문화재 (0)	10 (11.11%)	80 (88.89%)			

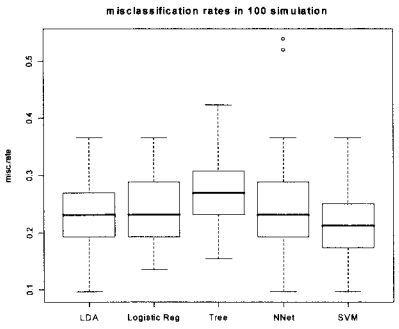
이를 통해 실제 데이터를 적합한 모형의 성능은 SVM이 다른 모형에 비해 정분류율, 민감도, 특이도가 우수한 것을 알 수 있었다.

4.2 모의실험을 통한 비교

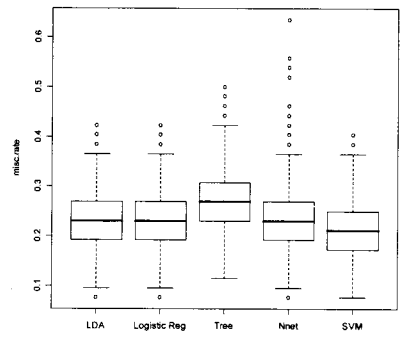
모의실험은 구축된 모형의 새로운 데이터에 대한 적용성을 확인하기 위해 실시하였다. 데이터를 모형 적합용 데이터와 모형 예측용 데이터를 각각 70%, 30%로 나누고 모의실험 횟수(rep)를 100, 1000, 10000번을 수행하여 얻어지는 각 모형의 오분류율의 평균을 통해 모형간의 성능을 비교하였다[2]. 다음의 <표 5>는 모형의 오분류율에 대해 정리하였으며, (그림 6), (그림 7), (그림 8)은 모의실험 횟수별 오분류에 대한 상자그림을 나타낸 것이다.

<표 5> 모의실험 및 적합모형의 오분류율

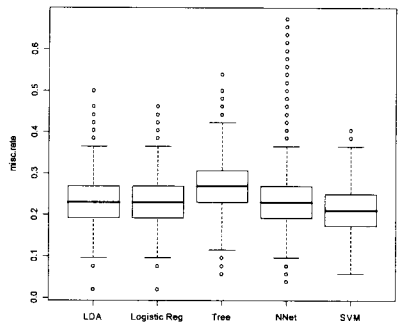
구 분	모의실험 (rep=100)	모의실험 (rep=1000)	모의실험 (rep=10000)	적합모형
선형판별 모형	23.27%	22.97%	22.79%	19.08%
로지스틱회귀 모형	23.98%	23.25%	23.16%	16.18%
의사결정나무 모형	27.19%	27.49%	27.51%	10.99%
신경망 모형	24.19%	23.08%	23.12%	15.61%
SVM	21.61%	21.18%	21.55%	7.51%



(그림 6) 오분류율 비교(rep=100)

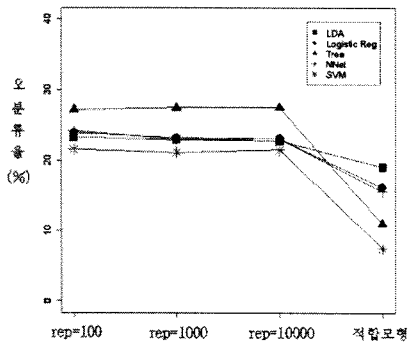


(그림 7) 오분류율 비교(rep=1000)



(그림 8) 오분류율 비교(rep=10000)

위의 결과를 통해 모의실험 횟수에 대한 모형의 오분류율 및 분산의 차이는 크지 않았으며 SVM 모형이 다른 모형에 비해 오분류율이 가장 낮게 나타났고, 의사결정나무모형이 다른 모형에 비해 높게 나타났다. 선형판별 모형, 로지스틱 회귀모형, 신경망 모형의 정분류율은 미묘한 차이를 보였으며, 또한 실제 데이터에 적합한 모형의 오분류율이 모의실험 결과를 통해 모두 과소적합 된 것을 다음의 (그림 9)를 통해 확인하였다.



(그림 9) 오분류율 비교

5. 결 론

본 연구에서는 다양한 통계적 분류방법을 이용한 매장문화재 예측모형을 개발하였다. 분석에 사용된 지역은 국내 I시의 자료를 사용하였다. 분석에 사용된 통계적 분류방법으로는 선형판별분석, 로지스틱회귀분석, 의사결정나무분석, 신경망분석, SVM 등이 사용되었으며, 모든 분석 절차는 최근 통계학분야에서 많은 관심을 받고 있는 R (http://www.r-project.org)을 이용하였다. 또한 분석의 전 과정에서 사용된 R 프로그램을 자세히 제시하여 유사 분석에 활용할 수 있도록 하였다. 마지막으로 모의실험을 통해 적합모형의 새로운 자료에 대한 적용성을 비교 분석하였다.

참 고 문 헌

[1] 나종화, *지역별 문화재 통계분석 및 모형개발 연구 1차*, 한국지질자원연구원, 2008.

[2] 송종우, "R의 분류방법을 이용한 신용카드 승인 분석 비교," *한국품질경영학회*, 2008.

[3] 양병화, *다변량 데이터 분석법의 이해*, 커뮤니케이션 북스, 2006.

[4] Anderson, J. A. *An Introduction to Neural Networks*, MIT press, Massachusetts, 2006.

[5] Breiman, L., Friedman, J. H., Olshen, R. A.,and Stone, C. J., *Classification and Regression Trees*, Wadsworth and Brooks, Pacific Grove, CA.

[6] Hastie, T., Tibshirani, R. and Friedman, J., *The Elements of Statistical Learning*, Springer, New York, 2001.

[7] Karatzoglou, A. and Meyer, D. "Support vector Machines in R," *Journal of Statistical Software*, 15, Article9, 2006.

[8] Yee, T. W. and Wild, C. J. "Vector Generalized Additive models," *Journal of the Royal Statisticla Society*, Series B, Vol. 58, No. 3, pp.481-493, 1996.

유 혜 경 (Hye-Kyung Yu)

- 2005년 2월 : 충북대학교 통계학과 (이학학사)
- 2007년 2월 : 충북대학교 통계학과 (이학석사)
- 2008년 2월 ~ 현재: 충북대학교 정보통계학과 (박사과정)
- 관심분야 : 데이터마이닝, 전산통계, 자료분석

이 진 영 (Jin-Young Lee)

- 2001년 2월 : 충남대학교 지질학과 (이학석사)
- 2006년 2월 : 충남대학교 지질학과 (이학박사)
- 현한국지질자원연구원 지구환경연구본부 선임연구원
- 관심분야 : 공간통계, 지리정보통계

나 종 화 (Jonghwa Na)

- 정회원
- 서울대학교 계산통계학과 이학박사
- 공군사관학교 전산통계학과 전임강사
- 서울대학교 통계연구소 특별연구원
- (미) PSU 방문교수
- 현 충북대학교 정보통계학과 교수
- 관심분야 : 전산통계, 데이터마이닝, 수리통계

논문접수일 : 2009년 7월 6일
논문수정일 : 2009년 8월 14일
게재확정일 : 2009년 8월 20일