Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 2, pp. 447 ~ 456

Development of Discriminant Analysis System by Graphical User Interface of Visual Basic[1)]

Yongkyun Lee[2)] · Youngjae Shin[3)] · Kyungjoon Cha[4)]

Abstract

Recently, the multivariate statistical analysis has been used to analyze meaningful information for various data. In this paper, we develope the multivariate statistical analysis system combined with Fisher discriminant analysis, logistic regression, neural network, and decision tree using visual basic 6.0.

Keywords : Decision Tree, Fisher Discriminant Analysis, Logistic Regression, Neural Network

1. 머리말

기하급수적으로 증가하는 자료와 컴퓨터 시스템의 발전으로 현대사회는 정보화시대로의 발전을 가속화 하였다. 이와 같은 정보의 홍수 속에서, 어떠한 방법으로 자료를 분석하느냐에 따라 다양한 결과가 도출되고, 결과 또한 다르게 해석된다. 그러므로 복잡하고 다양한 자료를 통계적 절차에 따라 처리 및 분석하면 보다 정확한 결과의 도출이 가능할 것이다. 현재 통계 분석 및 처리를 위해 국내에서 많이 사용되는 SAS, SPSS, MINITAB 등은 많은 모듈로 구성되어 있고 다양한 통계분석이 가능하다. 그러나 프로그램 사용법과 결과를 바르게 해석하기 위해 통계적 이론의 학습과 부단한 노력이 필요하다. 더불어 유사한 방법들이 하나의 진행 과정에서 수행되지 않기 때문에 비교 분석이 용이하지 않다.

---

1) This work was supported by a grant No. R01-2005-000-10866-0 from KOSEF.

2) First Author : Instructor, Dept. of Mathematics, Republic of Korea Airforce Academy, P.O. Box, 335, Ssangsu-ri, Namil-myeon, Cheongwon-gun, Chungbuk, 363-849, Korea. E-mail : mathyouth@hanyang.ac.kr

3) M.Sc., Dept. of Mathematics, Hanyang Univ., 17 Haengdang-dong, Seongdong-Gu, Seoul, 133-791, Korea.

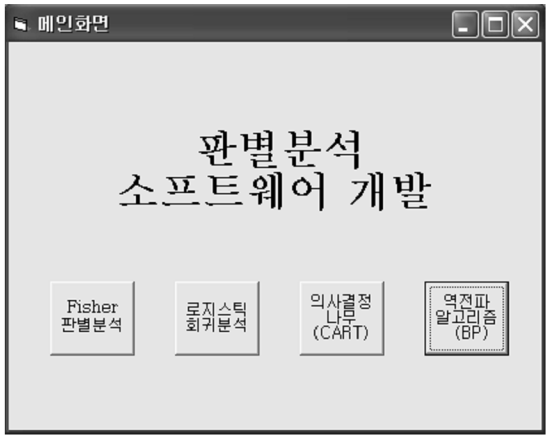
4) Corresponding Author : Professor, Dept. of Mathematics, Hanyang Univ., 17 Haengdang-dong, Seongdong-Gu, Seoul, 133-791, Korea.
E-mail : kjcha@hanyang.ac.kr

현재 다양한 자료에서 유의미한 정보를 탐색하는 방법으로 사회, 경제, 자연과학 등에서 활용되는 다변량 분석이 있으며, 다변량 분석과 관련하여 시스템을 구현하는 연구도 이루어져 상용화되기도 하였다. 서혜선 등(1999)은 허명회(1999)에 의해 연구된 다변량 수량화분석을 시스템화 하였으며, 현기홍, 최용석(2000)은 이원표 자료행렬의 행과 열을 그래프에 나타내어 관계와 패턴을 분석하는 다변량 그래프적 행렬도를 구현하였고, 한상태 등(2001)은 SAS/AF(application frame)와 SCL(screen control language)을 이용하여 SAS의 명령어 방식의 분석 과정을 메뉴방식으로 제공하여 활용이 어려웠던 다변량 기법을 일반 사용자들도 쉽게 사용할 수 있도록 하였다.

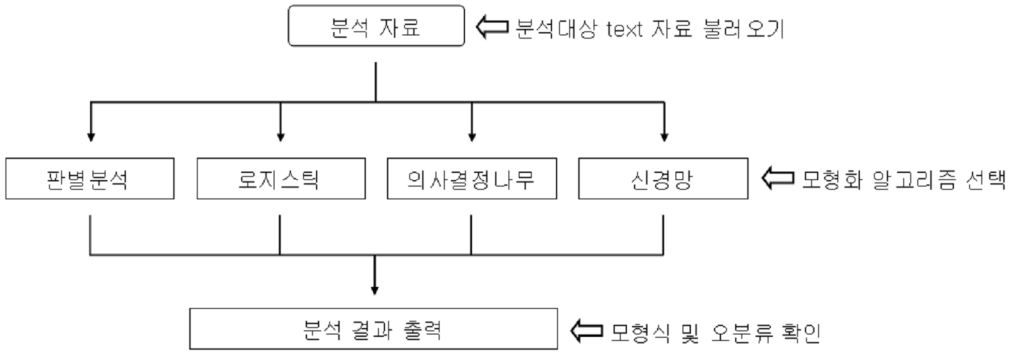
본 논문에서는 다변량 분석 중 이미 알려진 상호배반적인 몇 개의 집단에 속하는 다변량 관측치로부터 중복을 최소화하면서 구분할 수 있는 함수를 추정하는 판별분석을 비주얼 베이직 6.0(visual basic 6.0)을 사용하여 개발하였다. 비주얼 베이직은 프로그램 개발이 쉬우며, 구성화면이 사용자가 편리하게 활용 가능한 GUI(graphical user interface) 시스템으로 개발이 가능하다. 더불어 SAS와 같은 통계 도구 없이 단독으로 구동이 가능한 시스템이다. 개발에 사용되는 판별분석 방법으로 고전적인 Fisher의 판별분석(Bahong 등, 2004; Sugiyama, 2006), 의학 분야 중 질병 진단에 많이 활용되는 로지스틱 회귀분석(Zhu 등, 2004; Shen 등, 2005), 기계학습 방법인 신경망 중 다층 퍼셉트론(Yao, 1999) 그리고 결과 해석이 용이한 의사결정나무(Safavial 등, 1991; Du 등, 2002)를 판별분석 모형으로 선택하였다.

2. 판별분석 시스템의 구성

본 연구에서 개발한 판별분석 시스템은 비주얼 베이직을 사용하여 GUI에서 누구나 쉽게 사용할 수 있도록 하였으며, 전체 시스템은 메뉴 방식으로 구성하였다. 개발된 시스템은 분석에 필요한 자료의 구성이 선결되었다는 가정을 전제로 최종적인 모형이나 결론 도출이 가능한 시스템이다. 즉, 분석 과정에서 모형화와 결과 도출에 초점을 두고 개발하였다.



<그림 1> 판별분석 시스템 초기화면



<그림 2> 판별분석 시스템 자료처리 순서 및 결과 출력

다변량 분석 시스템 메인화면의 구성은 <그림 1>처럼 이루어졌으며, Fisher의 판별분석, 로지스틱 회귀분석, 신경망의 다층 퍼셉트론 중 오류 역전파 알고리즘, 그리고 의사결정나무 중 CART(classification and regression trees)를 실행하는 시스템이다. 또한 각 분석방법은 텍스트 형식으로 저장된 자료를 모형화 전에 파일로 구성하여 입력하므로 적용이 간단하고, 사용자의 선택에 따라 분석 방법별로 결과가 도출되어 오분류 및 모형식으로 최종적인 결과를 출력한다.

전체적인 분석 절차는 <그림 2>처럼 우선 분석 대상이 되는 자료를 텍스트 형식으로 저장하여 입력한다. 다음으로 입력된 자료를 4가지 판별분석 방법 중 사용자 선택에 따라 모형화하며 최종적으로 모형과 오분류 등으로 결과를 확인하게 된다. 각 방법에 따른 옵션의 선택과 결과의 도출은 조금씩 차이를 보인다.

3. 판별분석시스템 구성 알고리즘과 실행

본 연구의 판별분석시스템에 이용된 알고리즘은 Fisher의 판별분석, 로지스틱 회귀분석, 오류 역전파 알고리즘 그리고 CART이다. 이 방법들은 지금까지도 다양한 분야에서 널리 사용되는 방법(Fisher의 판별분석과 로지스틱 회귀분석)과 최근에 개발되어 판별분석의 효과 및 결과 해석이 용이한 방법(오류 역전파 알고리즘과 CART)이다. 판별분석시스템의 성능 확인을 위해 활용한 자료는 통계적 방법의 효율성 평가에 사용되는 붓꽃자료이다(Anderson, 1935). 붓꽃자료는 3가지 종류(versicolor, setosa, virginica)의 붓꽃에 대해 꽃의 종류(species), 꽃받침 너비(petal width), 꽃받침 길이(petal length), 꽃잎 너비(sepal width) 그리고 꽃잎 길이(sepal length)를 측정한 것인데, 꽃의 각 종류마다 50개씩 전체 150개의 자료로 구성되어 있다.

3.1 Fisher의 판별분석과 로지스틱 회귀분석

Fisher에 의해 제안된 방법으로 Fisher's between-within method라고 불리는 방법은 변수의 유용한 정보를 모두 포함한 정준(canonical)변수를 이용하여 판별분석한다. 변수의 수가 너무 많아 결과 해석이 곤란한 경우 고차원 공간 개체들의 집단 평균들

을 저차원 공간으로 변환하는 방법이다.

변수벡터 $X=(X_1, X_2, \cdots, X_p)$를 다음과 같이 선형 변환된 Y를 고려할 수 있다.

$$Y=d'X=d_1X_1+d_2X_2+\cdots+d_pX_p,$$

여기서, $d'=(d_1, d_2, \cdots, d_p)$이고 2표본 t-검정 통계량의 제곱(Hoteling T^2)을 최대화하는 선형변환으로, $d=S_p^{-1}(X_1-X_2)$을 구하게 되고, 이렇게 구해진 d를 정준계수벡터, $d'X$를 정준판별함수라고 한다. 부분집단이 2개인 경우 새로운 개체 X_0에 대하여

$$d'X_0 > \frac{d'\overline{X_1}+d'\overline{X_2}}{2} \rightarrow G_1,\ d'X_0 \leq \frac{d'\overline{X_1}+d'\overline{X_2}}{2} \rightarrow G_2.$$

이와 같이 각각 G_1, G_2로 분류한다.

일반적으로 3개 이상인 집단을 분류하기 위한 경우 Fisher의 판별함수는

$$F=\frac{d'Bd/(g-1)}{d'Wd/(N-g)}$$

를 최대화 하는 d를 구하면 된다. 여기서 T는 전체 평균 수정제곱합과 교적합의 행렬을 나타내고 W는 집단내 제곱합과 교적합의 행렬을 나타내고 B는 집단간 제곱합과 교적합의 행렬을 나타낸다.

로지스틱 회귀모형은 입력변수 $x_1, x_2, \cdots, x_p$에 대해서 식 (1)과 같다.

$$\ln\frac{p(y=1|x_1,\cdots,x_p)}{1-p(y=1|x_1,\cdots,x_p)}=\alpha+\beta_1x_1+\cdots+\beta_px_p. \tag{1}$$

식 (1)과 같은 모형을 로지스틱 판별(logistic discrimination) 모형이라 하고, 추정된 회귀계수 $a, b_1, \cdots, b_p$를 이용하여 식(2)와 같이 사후확률을 구한다.

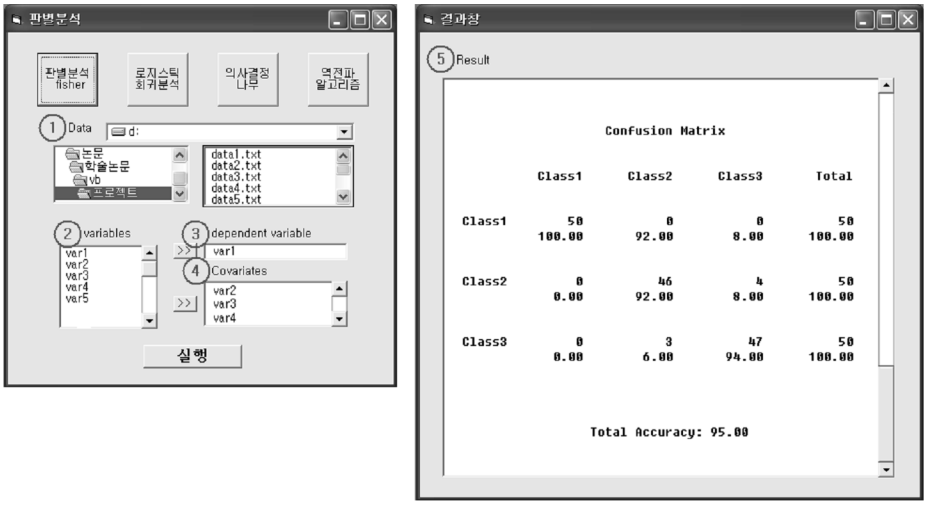
$$\hat{p}(y=1|x_1,\cdots,x_p)=\frac{\exp(a+b_1x_1+\cdots+b_px_p)}{1+\exp(a+b_1x_1+\cdots+b_px_p)}. \tag{2}$$

식 (2)에서 구한 각 개체에 대한 사후확률(posteria probability)은 그 개체를 분류하기 위해 사용된다$\left(\hat{p}(y=0|x_1,\cdots,x_p)=1-\hat{p}(y=1|x_1,\cdots,x_p)\right)$.

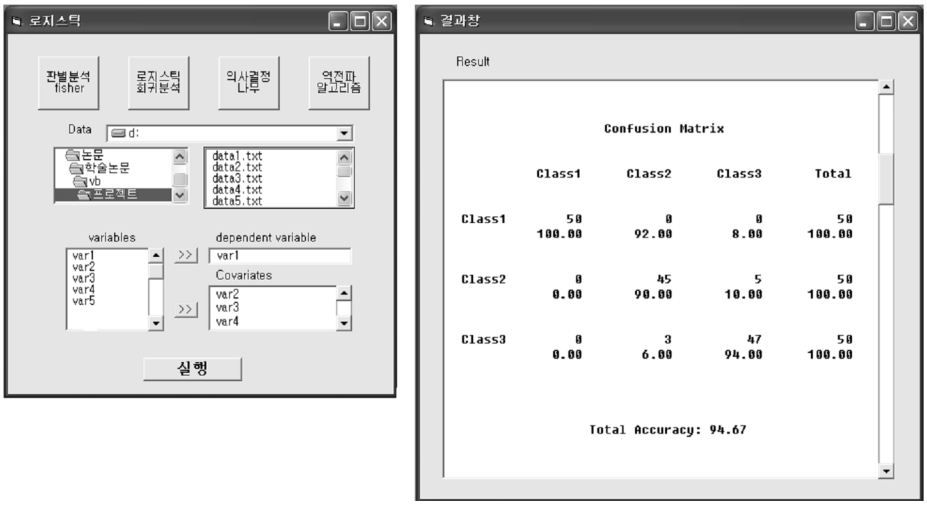
판별집단이 $k(k\geq 3)$인 경우 각 집단에 속할 사후확률 $q_i(\vec{x})$과 비교대상 집단에 속할 사후확률 $q_k(\vec{x})$의 자연대수 비율은 독립변수들의 선형결합으로 표시되고, 이러한 사후확률에 대한 자연대수비율을 집단 g_i와 집단 g_k를 판별하는 판별함수하고 한다.

Fisher 판별분석과 로지스틱 회귀분석의 실행은 다음과 같다.

우선 메인 메뉴에서 판별분석을 선택하면 <그림 3>의 왼쪽과 같이 판별분석을 실행하는 기본화면이 나타난다. <그림 3>의 ①은 분석대상이 되는 파일을 불러오는 메뉴이고 ②는 파일에 존재하는 모든 변수의 명을 나타내는 부분이다. 다음으로 ③은 분류를 위한 종속변수를 선택하는 부분이고 ④는 관측변수를 선택하여 각 집단을 판별하는데 사용한다. 이렇게 선택한 후 마지막으로 실행을 선택하여 판별분석을 실행하게 되며 결과는 ⑤와 같은 화면으로 나타나게 된다. 결과 출력물로는 오분류 행렬과 오분류율이 나타난다.



<그림 3> Fisher 판별분석 실행 및 결과



<그림 4> 로지스틱 실행 및 결과

로지스틱 회귀분석의 실행화면은 <그림 4>와 같으며 기본화면은 판별분석과 같은 형식으로 이루어져 있으며 사용방법도 동일하게 구성하였다. 실행 결과 역시 오분류 행렬과 오분류율로 출력된다.

3.2 신경망 - 오류 역전파 알고리즘

신경망 학습은 뉴런간의 연결강도를 조정하여 변화시키는 과정으로 신경망 학습 방법은 크게 지도학습(supervised learning)과 비지도 학습(unsupervised learning)으로 구분된다(Michie 등, 1994).

지도학습은 주어진 입력에 대해 올바른 출력이 어떤 것인지를 제공해주는 학습 방법으로 반드시 입력 x와 원하는 목표치 y의 쌍 (x,y)가 필요하며, 이를 학습형태 쌍(training pattern pair)이라 한다(Murray, 1996).

다층퍼셉트론은 입력층과 출력층 사이에 하나 이상의 중간층이 존재하며, 이 중간층을 은닉층(hidden layer)이라 한다. 이 다층퍼셉트론은 입력층, 은닉층, 출력층으로 연결되며, 각 층 내의 연결과 출력층에서 입력층으로 직접적인 연결이 존재하지 않는 전방향(feedforward) 신경망이다.

다층퍼셉트론은 단층, 2층, 3층 퍼셉트론의 구조를 가지며, 각 층에 따라 결정 구역도 다르게 표현된다. 다층퍼셉트론은 중간층과 각 노드의 입출력 특성을 비선형으로 연결하여 신경망의 능력을 향상시켰다. 다층퍼셉트론은 층의 개수가 증가할수록 퍼셉트론이 형성하는 결정 구역의 특성이 고급화된다.

신경망의 은닉층과 출력층은 다음과 같이 정의된다.

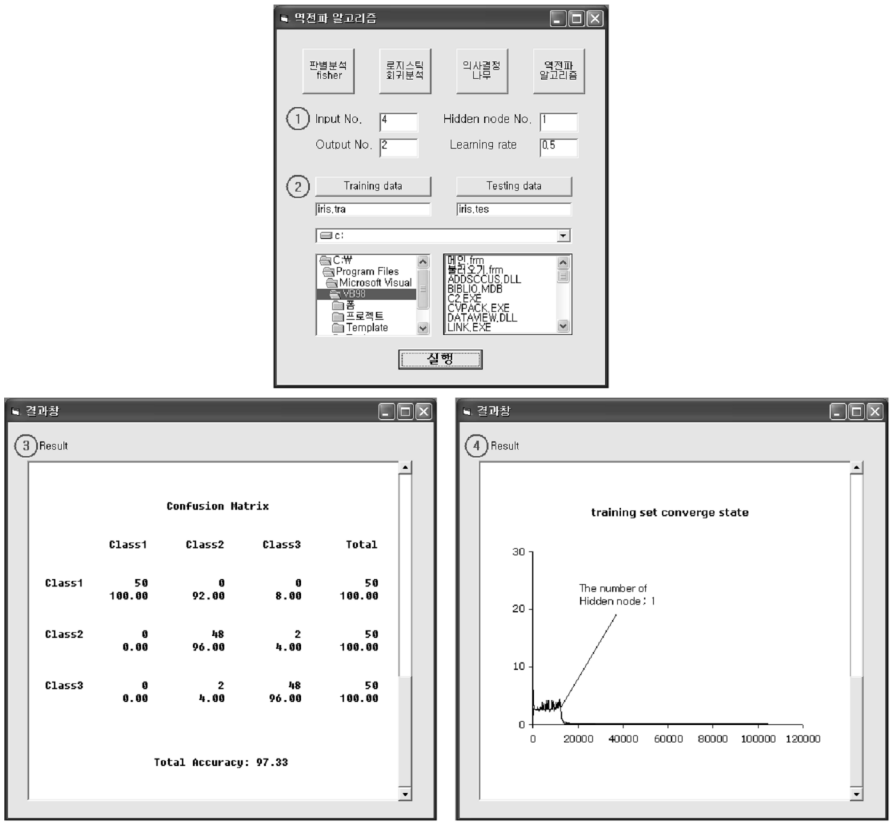
$$y_i^{(H)} = f(\sum_i w_{ij} x_i) \ , \ y_k^{(O)} = f(\sum_j w_{jk} y_j).$$

이는 입력 벡터 x에서 은닉 벡터 $y^{(H)}$를 경유하여 출력 벡터 $y^{(O)}$로의 사상을 의미한다. 여기서, $f(\cdot)$는 입력변수 또는 은닉마디의 결합을 변환하는 함수를 의미하며, 활성화 함수(activation function)라 한다.

활성화 함수는 뉴런의 입력에 대한 활성화 여부를 출력 신호로 사상시키며, 주로 로지스틱 함수(logistic function)로써 식 (3)와 같고 진폭은 $0 \le y_k^{(O)} \le 1$의 범위 위에 놓이고 출력은 확률로 표현된다.

$$f(x) = \frac{1}{1+e^{-x}}. \tag{3}$$

<그림 5>는 신경망 중 오류역전파 알고리즘의 실행메뉴와 이를 적용한 결과화면이다. 신경망 실행화면에서 ①은 신경망 분석에서 사용되는 입력층, 은닉층, 그리고 출력층의 수를 설정 가능하게 하였다. 그리고 ②에서는 훈련(training) 자료와 검정(testing) 자료를 선택하여 분석과정에서 모형화와 검증의 과정을 분리하였다. 이와 같이 선택된 자료를 실행하면 ③과 같이 오분류 행렬과 오분류율이 출력되고 마지막으로 ④의 결과 화면에서 보는 것과 같이 훈련(training) 과정에서 모형의 수렴 과정에서 반복수행하는 학습의 횟수가 출력된다.



<그림 5> 신경망 실행 및 결과

3.3 의사결정나무 - CART 알고리즘

의사결정나무는 데이터마이닝에서 결과를 예측하거나 자료를 분류하고자 할 때 매우 효과적인 기법으로, 의사결정규칙을 나무구조로 도표화하여 분류(classification)와 예측(prediction)을 수행하는 방법이다.

본 연구에서는 의사결정나무 알고리즘 중 CART(classification and regression trees) 알고리즘을 시스템화 하였다. CART는 목표변수가 이산형인 경우 지니 지수(gini index)를, 목표변수가 연속형인 경우 분산의 감소량을 이용하여 이지분리를 수행한다(Breiman 등, 1984).

지니 지수는 마디에서의 불순도를 측정하는 하나의 지수로 최고득점 규칙에 의해 각 마디의 범주가 결정된다. $P(j|t)$가 t번째 마디에서 임의의 한 개체가 목표변수의 j번째 범주에 속할 확률이라 할 때 오분류 확률의 추정치로 식 (4)과 같이 정의된다.

$$G = \sum_{j=1}^{c} p(j|t)(1-p(j|t)) = 1 - \sum_{j=1}^{c} p(j|t)^2 = 1 - \sum_{j=1}^{c} (n_j^{(t)}/n^{(t)})^2, \quad (4)$$

여기서, $n^{(t)}$은 마디 t에 포함되어 있는 관찰치의 수를 말하고, $n_j^{(t)}$는 마디 t에서 목표변수의 j번째 범주에 속하는 관찰치의 수이다. 식 (4)에서 지니 지수의 감소량($\triangle G$)을 식 (5)와 같이 정의할 수 있으며 이 값이 최대가 되는 설명변수와 그 변수의 최적분리를 분리점으로 선택한다.

$$G = G - \frac{n_L^{(t)}}{n^{(t)}} G_L - \frac{n_R^{(t)}}{n^{(t)}} G_R, \tag{5}$$

여기서, $n^{(t)}$은 부모마디의 관측치 수, $n_R^{(t)}$과 $n_L^{(t)}$은 자식마디의 관측치 수, G_L은 왼쪽 자식마디의 지니 지수, G_R은 오른쪽 자식마디의 지니 지수를 말한다.

목표변수가 연속형인 경우 다음과 같이 분산을 고려한다.

$$V = \frac{1}{n^{(t)}} \sum_{i=1}^{n^{(t)}} (y_i - \bar{y})^2.$$

<그림 6> 의사결정나무 실행 및 결과

이는 마디 t의 목표변수의 평균을 그 마디에 속하는 모든 개체의 예측값으로 사용하여 예측오차를 최소화하는 것이라고 볼 수 있으며 이런 분산의 감소량은 다음과 같이 정의된다.

$$V = V - \frac{n_L^{(t)}}{n^{(t)}} V_L - \frac{n_R^{(t)}}{n^{(t)}} V_R,$$

여기서, $n^{(t)}$은 부모마디의 관측치 수를 말하고, $n_R^{(t)}$과 $n_L^{(t)}$은 자식마디의 관측치 수를 말한다. 그리고 V_L과 V_R은 자식마디의 분산을 의미한다. 이러한 분산의 감소량($\triangle V$)를 최대로 하는 설명변수와 그 변수의 최적분리를 분리점으로 선택한다.

<그림 6>은 의사결정나무의 실행 메뉴와 결과 화면이다. 메뉴 화면 중 ①에서는 의사결정나무의 정지 규칙을 설정하는 부분 및 의사결정나무 알고리즘을 선택할 수 있는 부분으로 구성되어 있다. 현재 CART 알고리즘만 실행 가능하다. 그리고 ②와 ③에서는 분석이 되는 자료를 선택가능하게 하였다. 이처럼 입력된 자료를 CART 알고리즘을 적용하여 실행한 결과 ④에서 오분류 행렬과 오분류율이 출력되고 ⑤와 같이 나무구조의 결과도 출력된다.

4. 결론

본 논문에서는 유의미한 정보를 분석하기 위한 방법으로 다변량 분석 중에서 판별분석을 비주얼 베이직 6.0을 사용하여 개발하였다. 개발에 사용된 분석 방법은 Fisher의 판별분석, 로지스틱 회귀분석, 다층퍼셉트론, 그리고 의사결정나무가 이용되었고, 다음과 같은 특징을 갖는다.

첫째, 다양한 통계 분석 모듈에서 제공하는 판별분석 방법은 사용법과 결과를 바르게 해석하기 위해 통계적 이론의 학습과 부단한 노력이 필요하나, 본 연구에서 개발한 시스템은 판별분석만을 위한 손쉬운 절차를 갖는다.

둘째, 유사한 판별분석 간의 비교와 전문적인 판별분석을 수행하기 전에 좋은 방법의 선택에도 도움이 되는 시스템이라고 할 수 있다.

셋째, 비주얼 베이직을 사용하여 GUI에서 누구나 쉽게 클릭만으로 사용할 수 있게 하여 분석을 위한 절차 및 분석에 대한 내용보다 최종적인 모형이나 결론 도출에 활용할 수 있는 특징을 갖는다. 그러므로 판별분석을 쉽게 접근할 수 있어 교육적으로도 가치가 있다고 판단된다.

현재 다양한 자료에서 필요한 정보를 파악하는 데이터마이닝은 경제, 자연과학, 공학 그리고 의학 등 사회전반에 걸쳐 필요한 실정이다. 본 논문의 판별분석시스템은 데이터마이닝 도구로써 여러 방법을 적용할 수 있는 시스템이며 활용도가 높다고 할 수 있다. 차후 신경망과 의사결정나무의 다른 알고리즘의 추가 개발과 알고리즘간의 연계가 가능하도록 한다면 완성도 높은 시스템이 될 것이다.

참고 문헌

1. 서혜선, 김미경, 허명회 (1998). SAS AF/SCL로 구현한 다변량 수량화 시스템, *한국분류학회*, 제 3권, 1-11.
2. 허명회 (1999). *사회과학을 위한 다변량자료분석*, 서울 : 자유아카데미.
3. 현기홍, 최용석 (2000). 행렬도 시스템(Biplots System)의 개발, *응용통계연구*, 13권 2호, 297-306.
4. 한상태, 강현철, 이성건, 장명석, 이덕기, 유동균 (2001). Development of Multivariate Analysis System by Using SAS/AF and SCL, *The Korean Communications in Statistics*, 8, 507-514.
4. Anderson, E. (1935). The irises of the Gaspé peninsula, *Bulletin of the American Iris Society*, Vol 59, 2-5.
6. Bahong, J., Chang, C.I., Jensen, J.L. and Jensen, J.O. (2004). Unsupervised constrained linear Fisher's discriminant analysis for hyperspectral image classification, *SPIE,* Vol. 5546, pp. 344-353.
7. Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A. (1984). *Classification and Regression Trees*, CRC Pr I Llc.
8. Du, W. and Zhan, Z. (2002). Building Decision Tree Classifier on Private Data, *Australian Computer Society Inc. IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining. Conferences in Research and Practice in Information Technology,* Vol. 14. pp. 1-8.
9. Michie, D., Spiegelhalter, D.J. and Taylor, C.C. (1994). *Machine Learning, Neural and Statistical Classification,* Ellis Horwood, New York.
10. Murray, S. (1996). *Neural Networks for Statistical Modeling,* John Wiley & Sons, New York.
11. Safavial, S.R. and Landgrebe, D. (1991). A survey of decision tree classifier mothodology, IEEE Trans. *Systems, Man, and Cybernetics,* Vol. 21, No. 3, pp. 660-674.
12. Shen, L. and Tan, E.C. (2005). Dimension Reduction-Based Penalized Logistic Regression for Cancer Classification Using Microarray Data, IEEE/ACM Trans. *Computational Biology and Bioinformatics(TCBB) archive,* Vol. 2, No. 2, pp. 166-175.
13. Sugiyama, M. (2006). Local Fisher Discriminant Analysis for Supervised Dimensionality Reduction, *23rd International conference on Machine Learning.* pp. 905-912.
14. Yao, X. (1999). Evolving Artificial Neural Networks, *IEEE,* Vol. 87, No. 9. pp. 1423-1447.
15. Zhu, J. and Hastie, T. (2004). Classification of Gene microarrays by penalized Logistic Regression, *Biostatistics,* Vol. 5, No. 3, pp. 427-443.

[2007년 4월 접수, 2007년 5월 채택]