분류 알고리즘의 효율성에 대한 경험적 비교연구

The empirical comparison of efficiency in classification algorithms

전홍석[1)]
이주영[2)]

요약

We may be given a set of observations with the classes or clusters. The aim of this article is to provide an up-to-date review of different approaches to classification, compare their performance on a wide range of challenging data-sets. In this paper, machine learning algorithm classifiers based on CART, C4.5, CAL5, FACT, QUEST and statistical discriminant analysis are compared on various datasets in classification error rate and algorithms.

1. 서 론

일상생활에서 일어나는 많은 일들이 분류(Classification) 또는 예측(prediction)의 과제와 관련되어 있다. 두 길 중에 어느 길을 갈 것인가, 냄새로 와인의 품종과 생산년도를 구별해 내는 것으로부터 시작해서 병의 증상으로 병명을 판단하고자 할 때등 분류는 여러 분야에 걸쳐 일어나고 있다. 통계학 분야에서는 R. A. Fisher가 1930년대 붓꽃에 대한 자료를 분석할 때에 판별분석 기법을 처음으로 적용하였다. 또한 기계학습(Machine Learning)분야에서는 Thornquist와 Morgan은 자동적으로 상호작용을 찾아주는 방식의 AID(Automatic Interaction Detection) 기법을 제안하였고, Friedman은 AID기법을 비모수적 분류에 의한 반복적인 분할 결정 규칙으로 발전시켰으며, Breiman등은 이미 잘 알려진 분류와 회귀 트리(Classification And Regression Trees, CART)를 완성하였다. 기계학습이란 일반적으로 논리적이거나 이진 연산을 기초로 하여 자동으로 계산해 주는 과정을 다 포함한다. 분류 문제를 해결하기 위한 방법으로 판별분석을 비롯하여, 기계학습(Machine Learning)의 분야에서는 여러 가지 기계학습 알고리즘이 소개되었는데 이 논문에서는 C4.5, CAL5, FACT, QUEST의 기계학습 알고리즘을 다룰 것이다.

정보이론(Information Theory)을 기초로 하여 Friedman이 Breiman등과 함께 CART(1984) 시스템에 대한 방법을 개선한 C4.5는 Quinlan(1992)이 완성하였다. CAL5알고리즘은 CART시스템에서 노드가 두 개씩 나누어지는 단점을 보안하여 만들

*(402-751) 인천광역시 남구 용현동 253번지, 인하대학교 통계학과, 교수
**(122-701) 서울시 은평구 녹번동 5번지, 국립보건원, 박사후 연수생

어졌으며 Loh와 Vanichsettaki(1988), Loh와 Shih(1997)는 통계학적 판별분석을 기계학습 알고리즘에 적용한 FACT와 이를 개선한 QUEST 기계학습 알고리즘을 개발하였다. 본 논문의 목적은 현재까지 개발 된 여러 가지 분류 방법에 대해서 알아보고, 그들 방법의 장단점과 사용 방법 등에 토대로 하여 효율성을 비교하고자 한다. 이때, 효율성의 비교는 주어진 자료로 학습을 시킨후에 새로운 자료가 들어왔을 때 잘 분류를 하여야 좋은 알고리즘일 것이므로, 잘못 분류한 비율, 즉 오분류율이 각 분류알고리즘의 효율성을 나타내는 것으로 판단할 수 있다.

2. 판별분석(Discriminant Analysis)

2.1. 분류방법

분류에 영향을 미치는 것으로 조사된 k개의 확률변수 $X_1, X_2, \cdots, X_k$ 로 이루어진 관측벡터를 $X' = [X_1, ..., X_k]$ 라 하자. 고객들은 신용불량 집단 (ω_1)과 신용양호 집단 (ω_2)으로 이루어져 있는데, 특정 신청자의 관측값을 x라면 이 신청자가 ω_1에 속하는지 아니면 ω_2에 속하는지를 판단하여야 한다. 이때, ω_1의 확률밀도 함수를 $f_1(x)$, ω_2의 확률밀도함수를 $f_2(x)$라 표기하자. 관측 값 x가 이루는 표본공간을 Ω라 할 때 이것을 두 개의 서로 배반인 부분집합 R_1, R_2로 나누어 $x \in R_1$일 때 x를 ω_1으로, $x \in R_2$일 때에는 ω_2로 분류하게 된다. 주어진 관찰치를 가지고 관찰치가 사실은 ω_i으로부터 온 것인데 ω_j라고 잘못 분류할 확률은

$$P(j \mid i) = P(X \in R_j \mid \omega_i) = \int_{R_j} f_i(x)\,dx \qquad (2.1)$$

이다($i \neq j$, $i,j=1,2$). 다음 그림2.1은 $k=1$ 인 경우의 오분류 확률을 나타낸 것이다.

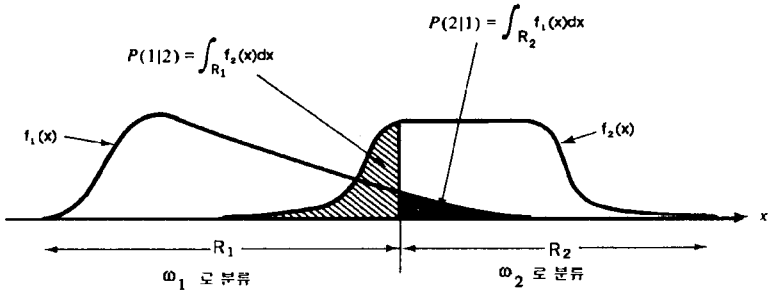

그림2.1. $k=1$ 일 때의 분류영역에 대한 오분류 확률

π_i을 관찰치 X가 ω_i에 속할 사전확률이라 하면, 주어진 관찰치 X가 ω_i로 틀리게 또는 맞게 분류할 확률은 다음과 같다.

$$P(\omega_i \text{로 틀리게 분류}) = P(X \in R_i \mid \omega_j)P(\omega_j) = P(i \mid j)\pi_j$$

$$P(\omega_i \text{로 맞게 분류}) = P(X \in R_i \mid \omega_i)P(\omega_i) = P(i \mid i)\pi_i$$

오 분류 비용은 다음과 같은 손실 비용행렬로 정의 될 수 있다.

		분류 결과	
		ω_1	ω_2
실제 모집단	ω_1	0	c(2|1)
	ω_2	c(1|2)	0

위의 표에서 ω_1로 분류 되어야하는 관찰치나 ω_2로 분류 되어야하는 관찰치가 ω_1, ω_2로 맞게 분류되었을 경우의 비용은 0, ω_1으로 분류되어야 하는 관찰치가 ω_2로 잘못 분류된 경우의 비용은 $c(2 \mid 1)$, ω_2로 분류되어야 하는 관찰치가 ω_1으로 잘못 분류된 경우의 비용은 $c(1 \mid 2)$으로 나타내었다. 따라서, 어떤 분류법칙에 대하여 오 분류 평균 비용(expected cost of misclassification, ECM)은 다음과 같이 정의한다,

$$ECM = c(2 \mid 1)P(2 \mid 1)\pi_1 + c(1 \mid 2)P(1 \mid 2)\pi_2 \quad . \tag{2.2}$$

합리적인 분류규칙은 ECM값이 가능한 한 매우 작은 값을 가져야 한다.

2.2. 가능도 비 법칙

$f_1(x), f_2(x)$ 은 각각 평균벡터는 μ_1, μ_2이고, 공분산 행렬은 Σ_1, Σ_2인 다변량 정규분포 밀도함수라 하자. 이때에 모집단 ω_1과 ω_2에 대하여 $X' = [X_1, \cdots, X_k]$ 의 분포는 다음과 같이 주어진다,

$$f_i(x) = \frac{1}{(2\pi)^{k/2} \mid \Sigma_i \mid^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right), \; i = 1,2. \tag{2.3}$$

이때, 모집단의 모수 μ_1, μ_2 그리고 $\Sigma_1 = \Sigma_2$ 이고 이 값이 Σ로 알려져 있다고 가정하면 영역 R_1과 R_2은 조건식의 양변에 log를 취하여 다음과 같이 표시할 수 있다;

$$R_1 = \left\{ (\mu_1 - \mu_2)'\Sigma^{-1}x - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq \ln\left(\frac{c(1 \mid 2)}{c(2 \mid 1)}\right)\left(\frac{\pi_2}{\pi_1}\right)\right\} \tag{2.4}$$

$$R_2 = \left\{ (\mu_1 - \mu_2)'\Sigma^{-1}x - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) < \ln\left(\frac{c(1 \mid 2)}{c(2 \mid 1)}\right)\left(\frac{\pi_2}{\pi_1}\right)\right\}.$$

따라서, 새로운 관찰치 x_0에 대하여 $\Sigma_1 = \Sigma_2$ 이라고 가정할 때,

$$(\mu_1 - \mu_2)'\Sigma^{-1}x - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq \ln\left(\frac{c(1 \mid 2)}{c(2 \mid 1)}\right)\left(\frac{\pi_2}{\pi_1}\right)$$

을 만족하면 ω_1 으로 분류하고 아니면 ω_2 로 분류한다.

그러나, 판별분석은 모집단의 분포 $f_i(x)$ 를 알아야 하고, 범주형 자료에 대해서는 적용을 할수 없다는 단점이 있다. 이러한 문제를 해결할수 있는 것으로 다음의 CART분류방법을 들수 있다.

3. CART (Classification And Regression Tree)

관측값 $x_n \in X$ 가 속한 집단을 $j_n \in \{c_1, ..., c_J\}$ 이라고 할 때, N개의 학습자료 (learning data, L)를 L이라 하면 학습자료는 $L = \{(x_1, j_1), ..., (x_N, j_N)\}$ 이 된다.

3.1. 트리 구조의 구축과정

이진 분류자는 측정벡터 X 자신으로부터 시작해서 X의 작은 부분집합으로 잘게 분리하는 일을 반복함으로서 구축할 수 있다. CART의 트리를 구축하는 데 있어서 첫 번째 문제는 주어진 학습자료를 두 개씩 가능한 한 작게 자르기 위하여 자료를 어떻게 사용하는가 이다. 이러한 기본적인 생각은 트리를 구축하여 자료를 부모 부분집합(parent subset)보다 더 '순수한' 각각의 자손 부분집합으로 분류되도록 선택하는 것이다. 노드의 분할을 찾는 방법은 그 뒤에 따라오는 노드를 만들기 위해 다음과 같은 방법으로 실행한다.

1. 각 노드(t)에 속한 자료가 집단 c_j에 속할 확률, $p(c_j \mid t), j = 1, 2, ..., J$ 을 계산한다. 이때 각각의 $p(c_j \mid t)$의 합은 $p(c_1 \mid t) + ... + p(c_J \mid t) = 1$ 이다.
2. 기니지수(Gini Index) $i(t) = \sum_{i \neq j} p(c_i \mid t)p(c_j \mid t)$ 는 노드 t에서의 불순(impurity)정도를 나타낸다. 각 변수에 대하여 기니지수 값을 계산하여 가장 작은 값을 갖는 변수를 선택하여 분할의 조건 s를 찾으면 노드 t는 두개의 노드 t_L, t_R로 나눈다. 노드 t에 속한 개체는 p_L과 p_R의 확률로 t_L과 t_R로 분할된다. 임의의 개체가 노드 t에 올 확률을 $p(t)$라고 하면, p_L, p_R는 각각 $p_L = p(t_L)/p(t)$, $p_R = p(t_R)/p(t)$ 으로 계산할 수 있어 불순 정도가 감소한 정도 $\Delta i(s,t)$ 는 $\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$ 으로 정의할 수 있다.
3. 각 노드(t_1)에서 혼합정도를 가장 많이 감소시키는 이진분할 s^*(

$\Delta i(s^*,t_1) = \max_{s \subset S} \Delta i(s,t_1)$)에 의하여 그 질문에 '예'인 개체는 모두 t_L로 '아니오'인 개체는 t_R로 보내는 방법으로 트리는 형성된다. 마지막 노드의 계급은 확률이 높은 쪽에서 결정되는 다수법칙(plurality rule)에 의해서 결정된다. 따라서, $p(c_{j_0} \mid t) = \max_j p(c_j \mid t)$ 이면, 마지막 노드 t의 계급은 c_{j_0}로 결정한다. 만약 어느 계급에 속하는지 잘 모르는 관찰치가 계급 c_j에서 마지막 노드가 되었다면 그 노드는 계급 c_j로 분류한다.

2장에서 언급한 판별분석과 비교해 보면, CART분류방법은 수치적 자료뿐만 아니라 순서형 변수와 범주형 자료를 모두 다룰 수 있는 장점이 있다. 또한 위의 판별분석의 결과보다는 해석하기 쉽고 자동으로 진행하므로 추가적인 노력 없이도 관찰치에 대한 분류뿐만 아니라 오분류확률을 추정 할 수 있다. 더욱이 가장 중요한 것은 판별분석에서는 반드시 필요한 모집단의 분포를 필요로 하지 않는다. 그러나 이진분할을 진행함으로서 트리의 크기가 커지는 단점이 있다. 이러한 단점을 보완하기 위하여 가지치기를 진행하는데 이때 최적 트리의 크기의 결정 또한 매우 중요한 연구의 한 분야이다. 그러나 이 논문에서는 최적 트리를 결정하기 위하여 가지치기를 시행하였으나 알고리즘 구축에 중점을 두고 언급하고자 한다.

4. C4.5

CART 분류방법이 Breiman et. al.이 발표한 결정트리 알고리즘(decision tree algorithm)인데 반하여 C4.5는 정보이론(Information Theory)에 기초하여 만들어진 알고리즘으로서 Quinlan(1979, 1983a, 1986b)의 ID3를 개선하여 만든 알고리즘이다.

4.1. 결정 트리의 구축

자료의 부분 집합을 Tr이라고 하자. Tr에 속하는 각 개체들이 계급 $\{c_1, c_2, ..., c_J\}$에 의하여 $Tr_1, Tr_2, ..., Tr_n$ 와 같은 부분 집합으로 나뉘어질 경우 Hunt(1966)는 정보량에 의하여 트리를 구축하는 방법을 제안하였다.

Hunt(1966)와 Quinlan(1992)에 의하여 제안된 C4.5 알고리즘은 결정 트리를 구축하는 방법은 어느 변수를 이용하여 트리를 구축하는 것이 최대의 정보를 주는 가를 계산하여 트리를 구축해 나간다. Tr이 주어졌을 때 정보량 $info(Tr)$을 다음과 같이 정의하였다.

$$info(Tr) = -\sum_{i=1}^{n} \frac{|Tr_i|}{|Tr|} \times \log_2\left(\frac{|Tr_i|}{|Tr|}\right) \qquad (4.1)$$

여기서 $|Tr|$은 Tr에 속하는 자료의 개수이고, i는 Tr의 각 경우의 수이므로

$|Tr_i|$는 각 경우에 속하는 개수이다. Tr이 어떤 변수 X에 의하여 $Tr_1, Tr_2, ..., Tr_n$ 으로 나누어지면 Tr의 변수 X에 따른 기대 정보는 각 부분집합 Tr에 대하여 다음 식 $info_X(Tr)$을 정의할 수 있다,

$$info_X(Tr) = \sum_{i=1}^{n} \frac{|Tr_i|}{|Tr|} \times info(Tr_i). \quad (4.2)$$

위 식을 이용하여 다음과 같은 정보의 측도, 즉 X를 선택하여 분류함으로서 얻을 수 있는 정보의 이득 $gain(X)$는 $gain(X) = info(Tr) - info_X(Tr)$ 이다. 변수 중 정보의 이득(4.3)을 최대로 하는 것에 의하여 분류를 계속하게 된다. 즉 주어진 자료에 대하여 $info_X(Tr)$을 구하고 각 $info(Tr)$와의 차를 구해 그값이 큰 쪽으로 분류를 해나간다.

변수 X의 속성 A에 기초하여 자료집합 Tr을 분류할 때 Tr에서 속성 A를 만족하는 비율을 F라 하면 정보의 이득량을 계산한 후에 정보의 이득을 $gain\ ratio(X) = gain(X)/split\ info(X)$ 로 다시 쓸 수 있으며 이 경우 $info_X(Tr)$을 $split\ info(X)$로 표기한다. 이득율 $split\ info(X)$ 값을 기준으로 가지치기를 실행한다.

4.2. 이득율 기준

이러한 분류방법은 자료의 상태에 따라 결과들이 변수에 의존하여 편의가 심해진다. 즉 편의가 심한 하나의 변수가 정보이득이 크면 그 변수로 먼저 분할을 하여 트리를 형성하기 때문이다. 또한 모든 식별은 하나의 경우만을 갖도록 진행되기 때문에, 이 속성들의 값은 자료를 임의로 분할하는 것은 각 하나의 경우만을 포함하는 부분집합의 개수가 많아지게 된다.

그러나, 이득 기준에 의하여 야기되는 편의는 실험에 타당한 이익을 주는 여러 가지 속성들을 정규화 함으로서 수정되어 질 수 있다. 각각 자료가 속하는 계급뿐만 아니라 학습자료, 테스트 자료에 따라 계급의 정보를 잘 지적해 주는 그러한 정보가 있다고 가정해보자.

그때 같은 분할로 생겨나는 분류의 적절한 정보에 대하여 $split\ info(X)$를 $info_X(Tr)$의 정의를 이용하여 계산하게 되고, 주어진 자료에서 Tr를 n개의 부분집합으로 나누어줌으로서 일반화되는 잠재적 정보이다. 또한 $gain\ ratio(X) = gain(X)/split\ info(X)$ 는 정보 비가 유용한 즉, 분류하는데 도움이 되는 분할에 의하여 일반화됨을 표현하고 있다. 만약 분할이 거의 같다면, 분할정보는 매우 작고, 이 값 또한 안정적이지 못할 것이다. Quninlan(1988b)은 경험적으로 정보량의 이득만 가지고 트리를 구하는 것보다는 이득률기준을 이용하여 트리를 형성한 것이 더 좋은 결과를 얻을 수 있다는 것을 보였고, Minger(1989)도 여러 가지 기준

의 테스트에서 이득율을 사용하여 트리를 형성하는 것이 더 간단한 트리를 형성한다는 것을 보였다.

본 논문에서 C4.5를 이용하여 트리를 구축하고 오분류율을 구할 때 이득율 기준에 따라 최적의 트리를 구하였다.

이러한 분류방법은 CART방법과 마찬가지로 수치적 변수뿐만 아니라 범주형 변수까지도 포함하여 분류를 할수 있는 장점이 있고, 모집단의 분포를 가정하지 않아도 된다. 또한 CART가 이진분할을 하는데 반하여 C4.5는 실험집합에 대하여 모두 분할을 하므로 트리의 크기가 적어지는 장점이 있다.

5. CAL5

CAL5는 C4.5와 같이 정보이론을 이용하여 만들어진 알고리즘이나 정보이론중에서 엔트로피 이론을 이용하여 구축된 알고리즘이다. O를 분류되는 자료들의 전체 집합이라고하고, k개의 변수가 존재한다고 하자. CAL5는 n차원으로부터 부분집합 $O_j \in O\,(j=1,2,...,k)$ 에 의해 표현되는 영역으로 나눈다. 이때, 계급 $c_j,(j=1,2,...,J)$ 가 $P(c_j)>\beta$로 존재하고, $\beta<1$은 분기점이다.

순차적 변수는 트리 구축의 성과에 영향을 많이 주게 되므로 첫 번째 단계에서 좋은 판별성질을 가지고 있는 변수를 선택하게 되고 분류를 잘하기 위하여 변수를 적정한 구간으로 나누어주는 알고리즘도 매우 중요하다. 모든 자료들은 값이 증가함에 따라 축의 속성 A에 따라 순서화 되어 있다고 하고, 또한 적정한 구간 I로 나누어져 있다고 하자. 구간 I에서 변수의 순서화 된 자료를 포함하는 그러한 구간은 왼쪽에서 오른쪽으로 자료를 계급을 결정하거나 계급을 결정 할 수 없다고 판단할 때까지 순차적으로 진행한다. 계급의 결정은 구간 I에서 현재노드를 t라 할 때 조건부 확률 $p(c_i \mid t)$의 추정량을 사용하여서 구간 I에서 다음 가설을 검정한다.

H_1: 구간 I에서 $p(c_i \mid t) \geq \beta$을 만족하는 계급 c_i가 존재한다.

H_2: 구간 I에서 생겨나는 모든 계급 c_i에 대하여 부등식 $p(c_i \mid t)<\beta$가 성립한다.

즉 각 계급 c_i에 대하여 신뢰구간이 미리 결정된 분기점보다 커지면 가설 H_1을 선택하고, 신뢰구간이 분기점보다 작으면 H_2를 채택하게 된다는 것이다. 따라서, 위 가설에 대하여 검정하게 된다.

1. 만약 임의의 계급 c_i가 존재하면 즉, H_1 가설이 구간 I에서 c_i가 존재하는 것이 맞다면, 구간 I에서 과정은 끝나게 된다. 따라서 트리의 결과로서 계급 c_i라고 결정하게 된다.

2. 만약 구간 I에서 모든 계급에 대하여 H_2의 가설이 맞다면 구간 I에서는 계급을 정할 수 없다. 이 경우에는 다음 단계에서 가지치기를 진행하여 분할을 계속하게 된다.
3. 만약 1,2 둘 다 자료의 부족으로 인하여 결론을 내리지 못했다면 구간 I는 현재의 변수에서 다음 관찰치를 선택함으로서 구간이 확장되어진다. 이것은 H_1과 H_2를 결정하기 위해 자료가 증가하게 된다. 만약 더 이상 구간 I를 확장하기 위한 자료가 없다면 계급 결정은 각 계급일 확률이 높은 곳에서 계급을 결정하게 된다.

그러나, CAL5분류방법은 구간을 어떻게 설정하는가에 따라 오분류율의 차이가 많이나고 트리를 해석하기 어렵다는 단점이 있다. 실제로 본 논문에서 실제 분류에 적용되는 자료중에서 숫자인식 자료의 경우 자료의 순서나 구간의 차이에 따라 오분류율의 차이가 커 인식률이 안정적이지 못한 결과를 보였다. 또한 분기점 β 값을 어떻게 주느냐에 따라 트리 결과도 많은 차이를 보였다.

6. FACT와 QUEST

Loh와 Vanichsettaki(1988)는 FACT(A Fast Algorithm for Classification Trees) 알고리즘을 제안하였다. 이 알고리즘에서는 각 노드에서 분할을 하기 위한 변수를 선택하기 위해 통계적 가설검정을 도입하여, 그를 통해 변수를 선택한 후 판별분석을 실행한다. 따라서, 각 노드에서는 두 개이상의 가지가 형성될 수 있다. 결정트리의 크기는 미리 정한 정지규칙에 의해 결정된다. FACT알고리즘은 CART(Breiman et. al.,1984)처럼 하나의 변수를 이용한 분할도 가능하고 선형분할도 가능하다는 장점이 있다.

6.1. 트리 구축과정

k차원의 자료가 존재하고 이들 자료의 계급을 c_j, $j=1,\cdots,J$, 주어진 자료에서 계급 c_j에 해당하는 자료의 수를 N_j라 하면 전체 표본의 수는 $N=\sum_j N_j$ 이다. 또한 자료로부터 임의의 노드 t에서 자료가 속하는 계급의 수는 $J_t(J_t \le J)$로 표기하고 계급 c_j의 수를 $N_j(t)$라고 표기하자. k번째 변수는 X_k, 관찰치는 x_k라 하자. 이때 계급 c_j에 대한 사전 확률 $\pi_j = N_j/N$ 이라 하자. 또한 임의의 개체가 노드 t에서 계급 c_j에 속할 확률 $p(c_j,t)$ 은 $p(c_j,t)=\pi_j N_j(t)/N_j$ 이 되고, 노드 t에 속한 개체가 계급 c_j일 조건부 확률 $p(c_j|t)=p(c_j,t)/\sum_i p(c_i,t)$ 이 된다. 노드 t에서 혼합정도인 기니지수 $i(t)$는 $i(t)=1-\sum_j p^2(c_j \mid t)$ 로 정의한다.

6.2. 분할점 선택

임의의 노드를 분할하기 위하여 선택한 순서형 변수를 X라고 하자. FACT는 분할을 구축하기 위하여 선형판별 분석을 시행하는데 이것은 두 가지 단점이 있다. 첫째로, 노드에서 계급을 결정할 때까지 수많은 부노드를 만들어 낸다. 두 번째는 계급간의 분산이 동일하지 않을 수도 있다는 사실이 무시되었다는 것이다.

이들 두 가지 문제를 QUEST가 해결하였는데 그 과정을 살펴보면 다음과 같다. $J>2$일 때, 이진 분할을 잘하기 위해 전체 계급을 두 개의 초 계급(superclass)으로 나누어준다. 분산이 같지 않을 수도 있음을 고려하여 두 개의 초 계급에 이차 판별함수를 사용한다.

첫 번째 두 개의 계급만이 존재한다고 가정하자. 전통적인 이차 판별분석은 정규분포에서 계급의 밀도함수를 추정한다. 이때, 정규분포의 평균과 분산은 자료로부터 추정한다. 표본에서 $\bar{x}^{(j)}$와 s_j^2를 $c_j(j=1,2)$ 계급의 평균과 분산이라고 하자. $\phi(x)$를 $\phi(x)=(2\pi)^{\frac{1}{2}}\exp(-\frac{x^2}{2})$ 인 표준 정규분포 밀도함수라고 하자. 이차 판별분석은 X축을 세 개의 구간으로 나누는데 그 구간을 살펴보면 $(-\infty,d_1),(d_1,d_2)$, (d_2,∞)이다. 이때, d_1,d_2는 $p(c_1 \mid t)s_1^{-1}\phi((x-\bar{x}^{(1)})/s_1)=p(c_2 \mid t)s_2^{-1}\phi((x-\bar{x}^{(2)})/s_2)$ 방정식의 근 이다. 이진분할을 하기 위하여, QUEST는 분할점으로 두 개의 근 중에서 하나만을 사용한다. 두 개의 근 중에서 각 계급의 표본평균에 가까운 것을 분할점으로 선택한다. 두 개의 초 계급으로 예비 집단화하는 과정에서는 2-평균 군집화 알고리즘을 적용하여서 실행한다. 만약 계급의 평균이 동일하다면, 대부분의 경우가 초 계급 A가 되고 다른 형태는 초 계급 B가 된다. 이 알고리즘을 요약하면 다음과 같다.

1. 2-평균 군집화 알고리즘을 사용하여 계급을 두개의 초 계급 A와 B로 나눈다.
2. $\bar{x}_A$와 s_A^2를 초 계급 A의 평균과 분산이라고 하자. 계급 B의 평균과 분산을 $\bar{x}_B$와 s_B^2이라고 하자. 또한 $p(A \mid t)=\sum_{c_j \in A} p(c_j \mid t)$와 $p(B \mid t)=1-p(A \mid t)$ 를 초 계급의 사전확률이라고 하자.
3. 이차판별함수 ax^2+bx+c 의 계수는 방정식에 log를 취하여 정리하면

$a=s_A^2-s_B^2$

$b=2(\bar{x}_A s_B^2-\bar{x}_B s_A^2)$

$c=(\bar{x}_A s_B)^2-(\bar{x}_B s_A)^2+2s_A^2 s_B^2 \log[\{p(A \mid t)s_B/p(B \mid t)s_A\}]$

와 같이 구할 수 있다.

4. d가 다음과 같을 때 노드는 $X=d$ 에서 분할한다.

(a)만약 $a=0$ 이면,

$$d=\begin{cases} (\bar{x}_A+\bar{x}_B)/2-2(\bar{x}_A-\bar{x}_B)s_A^2\log\dfrac{p(A\mid t)}{p(B\mid t)}, & \bar{x}_A\neq\bar{x}_B, \\ \bar{x}_A, & \bar{x}_A=\bar{x}_B. \end{cases}$$

(b) 만약 $a\neq 0$ 이면,

(i) 만약 $b^2-4ac<0$ 이면, $d=(\bar{x}_A+\bar{x}_B)/2$ 라고 정의한다.

(ii) 만약 $b^2-4ac\geq 0$ 이면,

A. 두 개의 근 $(2a)^{-1}\{-b\pm\sqrt{b^2-4ac}\}$ 중 $\bar{x}_A$ 에 가까운 근을 d로 정의한다. 단, 이때 갈라지는 가지는 반드시 자료가 포함되어 있어야 한다.

B. 그 외의 경우에는 $d=(\bar{x}_A+\bar{x}_B)/2$ 로 정의한다.

QUEST 알고리즘은 범주형 변수를 분류에 적용하지 못하는 단점을 보완하여 범주형 변수를 가변수로 바꾸어 주어 판별분석에 사영시키는 방법을 이용하여 분류하게 된다. 이진분할을 하고, 범주형 변수를 사용할 뿐만 아니라 모집단의 분포를 주어야 하는 판별분석의 단점을 개선하였다.

7. 결 론

앞에서 언급한 각 알고리즘의 효율성을 비교하고자 한다. 통계학 분야에서 R. A. Fisher가 1930년대 붓꽃에 대한 자료를 분석할 때에 판별분석 기법을 처음으로 적용한 iris 자료와 Breiman의 숫자인식 자료, 파동형태 자료를 이용하여 오분류율을 구하였다. 이때의 오분류율은 전체 자료에서 실제의 계급을 제대로 분류하지 못한 관찰치의 비율이다.

7.1. 자료

7.1.1. 숫자인식자료

숫자인식은 보통 전자시계나 계산기에서처럼 7개의 수평과 수직 불이 켜지고 꺼짐의 조합에 따라 표현된다. 이 자료는 고장난 계산기로부터 얻어진 자료이다. 각 7개의 자리에 불이 들어 올 확률은 0.1이다. 좀더 세밀하게 말하자면, 그 자료는 랜덤 벡터 $(x_1,\ldots,x_7,Y)$ 로부터의 결과이고 이때의 Y는 계급이고, 숫자 1,...,10 은 동등한 확률을 갖는다고 가정하자. 또한 주어진 Y에 대하여 $x_1,\ldots,x_7$ 은 각각 독립이고 0.9의 확률과 에러 확률이 0.1로서 Y에 대응하는 값에 같은 확률 값을 갖는다.

7.1.2. 파동형태 자료

다음으로 파동형태자료의 인식문제에 대해서 알아보고자 한다. 세 개의 파동형태인 $h_1(t), h_2(t), h_3(t)$ 에 기초로 한다. 각 계급은 이들 파동 형태 중에 두 개의 랜덤 볼록한 모형의 결합으로 고려되어진다. 측정 벡터는 $X = \{x_1, \ldots, x_{21}\}$ 즉, 21-차원이다. 계급 1의 벡터 x를 생성하기 위해 독립적으로 일양분포 난수 u를 생성하고 또한 21개의 평균이 0 분산이 1인 정규분포의 난수 $\varepsilon_1, \ldots, \varepsilon_{21}$ 를 생성한다. 다음으로

$$x_m = uh_1(m) + (1-u)h_2(m) + \varepsilon_m, \quad m = 1,2,\ldots,21 \tag{7.1}$$

이라고 하자. 또한 계급 2 벡터 x는

$$x_m = uh_1(m) + (1-u)h_3(m) + \varepsilon_m, \quad m = 1,2,\ldots,21 \tag{7.2}$$

이라고 가정하자. 또한 계급 3 벡터 x는

$$x_m = uh_2(m) + (1-u)h_3(m) + \varepsilon_m, \quad m = 1,2,\ldots,21 \tag{7.3}$$

이라고 가정하자. 이러한 자료의 형태를 파동형태(Breiman et. al., 1984) 자료라 하고 5000개를 생성하여 이 자료를 가지고 각 인식률을 비교하고자 한다. 또한 $h_1(t)$, $h_2(t), h_3(t)$의 형태는 다음 그림7.1, 그림7.2, 그림7.3과 같다.

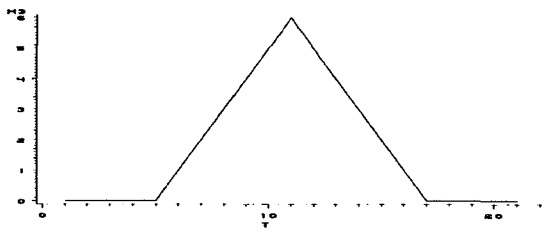

그림 7.1. $h_1(t)$ 함수의 형태

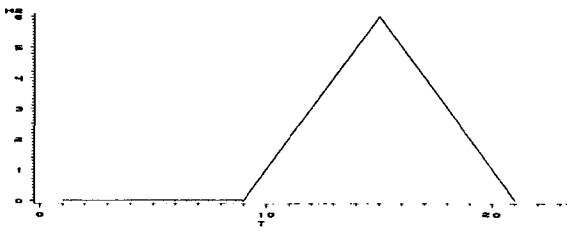

그림 7.2. $h_2(t)$ 함수의 형태

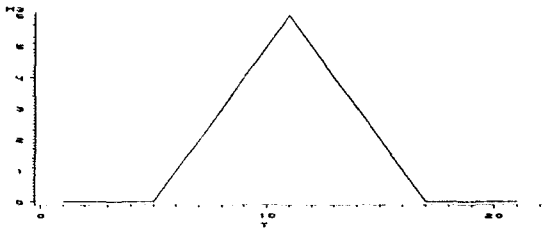

그림 7.3. $h_3(t)$ 함수의 형태

7.2. 분석결과

iris자료, 숫자인식 자료, 파동형태의 자료를 가지고 오분류율을 비교하면 다음 표 7.1와 같은 결과를 얻을 수 있다.

표7.1. 자료에 따른 오분류율 비교 (*:불가)

	iris자료		숫자인식자료		파동형태자료	
	학습자료 (150)	테스트자료 (30)	학습자료 (200)	테스트자료 (48)	학습자료 (4500)	테스트자료 (500)
판별분석	0.027	0.000	*	*	0.027	0.000
C4.5	0.064	0.073	0.220	0.365	0.064	0.073
CAL5	0.027	0.033	0.350	0.313	0.027	0.033
QUEST	0.002	0.000	0.191	0.375	0.000	0.375

위의 결과를 살펴보면, 판별분석의 경우에는 자료가 iris자료인 경우나 파동형태의 자료 모두 비슷한 오분류율을 보여 수치적 자료에 대해서는 자료의 수에 비해 안정적임을 입증하였다. 기계학습 알고리즘에서는 C4.5와 CAL5가 비슷한 iris자료와 파동형태의 자료에서는 낮은 오분류율을 보였으나, 범주형 변수들로만 이루어진 숫자인식 자료에서는 C4.5가 0.22로 더 좋은 오분류율을 보였다. 전체적으로 C4.5가 자료에 변화에 따른 오분류율이 안정적인 것으로 나타났다. 모든 알고리즘이 판별분석을 제외하고는 숫자인식자료에서 오분류율이 높아졌으며 iris 자료의 경우는 QUEST가 테스트자료를 다 분류해 내었으며 다음으로는 판별분석의 오분류율이 낮게 나온 것을 알 수 있다.

참 고 문 헌

[1] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984), *Classification and Regression Trees*, Wadsworth, Belmont.

[2] Fukunaga, K., and Narendra, P. M. (1975), "A Branch and Bound Algorithm for Computing k- Nearest Neighbours," *IEEE Transaction on Computers*, C-25, 917-922.

[3] Lim, Tjen-Sien. and Loh, Wei-Yin (1998), "An Empirical Comparison of Decision Trees and Other Classification Methods", *University of Wisconsin, Madison Technical Report 979.*

[4] Loh, Wei-Yin. and Vanichsetakul., N. (1988), "Tree-structured Classification via Generalized Discriminant Analysis (with discussion)," *Journal of the American Association* 83, 715-728.

[5] Mingers, J.(1989), "An empirical comparison of selection measure for decision -tree induction", *Machine Learning*, 3, 4, 319-342.

[6] Nakhaeizadeh, G., and Taylor, C. C. (1997), *Machine Learning and Statistics, the interface,* John Wiley & Sons. New York.

[7] Quinlan, J. R.(1979), *Discovering rules by induction from large collections of examples,* In D. Michie(ed.), *Expert Systems in the Micro Electric Age*, Edinburgh, UK: Edinburgh Press.

[8] Quinlan, J. R.(1983a), *Learning efficient classification procedures*, In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell(ed.), *Machine Learning: An Artificial intelligence Approach,* Palo Alto, CA: Tioga Press.

[9] Quinlan, J. R.(1986b), "Induction of decision trees", *Machine Learning* 1, 1, 81 - 106, Reprinted in J.W. Shavlik and T. G. Dietterich(ed.), *Reading in Machine Learning*, San Mateo, CA: Morgan Kaufmann.

[10] Quinlan, J. R.(1992), "Themes and issue in empirical learning", *Proceedings of the Sixth National Conference of the Japanese Society for Artficial Intelligence, Tokyo.*

[11] Quinlan, J. R.(1993), *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kaufmann.

[12] Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press, New York.

저자소개

전홍석 : 서울대학교 수학과를 졸업했으며, 미국 위스콘신 대학교에서 통계학 박사를 취득하였다. 현재 인하대학교 이과대학 통계학과 교수로 재직중이다. 관심분야는 통계계산, 데이터 마이닝등 이다.

이주영 : 한국 외국어대학교 수학과를 졸업했으며, 인하대학교에서 박사를 취득하였다. 현재 국립보건원에서 Post Doc. 으로 재직중이다. 관심분야는 통계계산, 데이터 마이닝등 이다.