# 분류와 회귀나무분석에 관한 소고[1)]

임용빈 · 오만숙
이화여자대학교 통계학과

# Note on classification and regression tree analysis

Yong B. Lim · Man-Suk Oh
Department of Statistics, Ewha Womans University

Key words: Classification and Regression tree, Mutiple trees, Sequential Strategy

## Abstract

The analysis of large data sets with hundreds of thousands observations and thousands of independent variables is a formidable computational task. A less parametric method, capable of identifying important independent variables and their interactions, is a tree structured approach to regression and classification. It gives a graphical and often illuminating way of looking at data in classification and regression problems. In this paper we have reviewed and summarized the methodology used to construct a tree, multiple trees and the sequential strategy for identifying active compounds in large chemical databases.

## 1. 서론

정보기술(Information Technology)의 발전과 더불어서 관심이 있는 반응변수와 관련된 방대한 양의 정보들의 데이터 베이스화가 가능하여져서 대규모의 자료를 처리하고 분석하기 위한 통계적인 방법의 필요성이 대두된다. 자료의 크기가 수만 개에서 수십만 개에 이르고, 각각의 자료점을 설명하는 독립변수 또는 설명변수의 개수가 수백 개에서 수천 개에 이르면서 독립변수들 사이에 상호작용효과까지 기대되는 경우에 기존의 회귀기법에 의한 자료의 분석은 제한적이다. 주요 이유의 하나는 계획행렬의 크기가 너무 커서 최적 모형을 찾기 위한 회귀진단, 주성분 분석법 등을 적용하기가 어렵기 때문이다.

자료를 반응변수들의 값에 따라서 동질적인 그룹으로 가장 쉽게 나누는 방법은 각각의 자료점에 대한 설명변수들의 값에 따라서 예/아니오로 대답할 수 있는 일련의 질문을 통해 자료를 축차적으로 이진(binary) 분방대한 양의 정보들의 데이터 베이스화가 가능하여져서 대규모의 자료를 처리하고 분석하기 위한 통계적인 방법의 필요성이 대두된다. 자료의 크기가 수만 개에서 수십만 개에 이

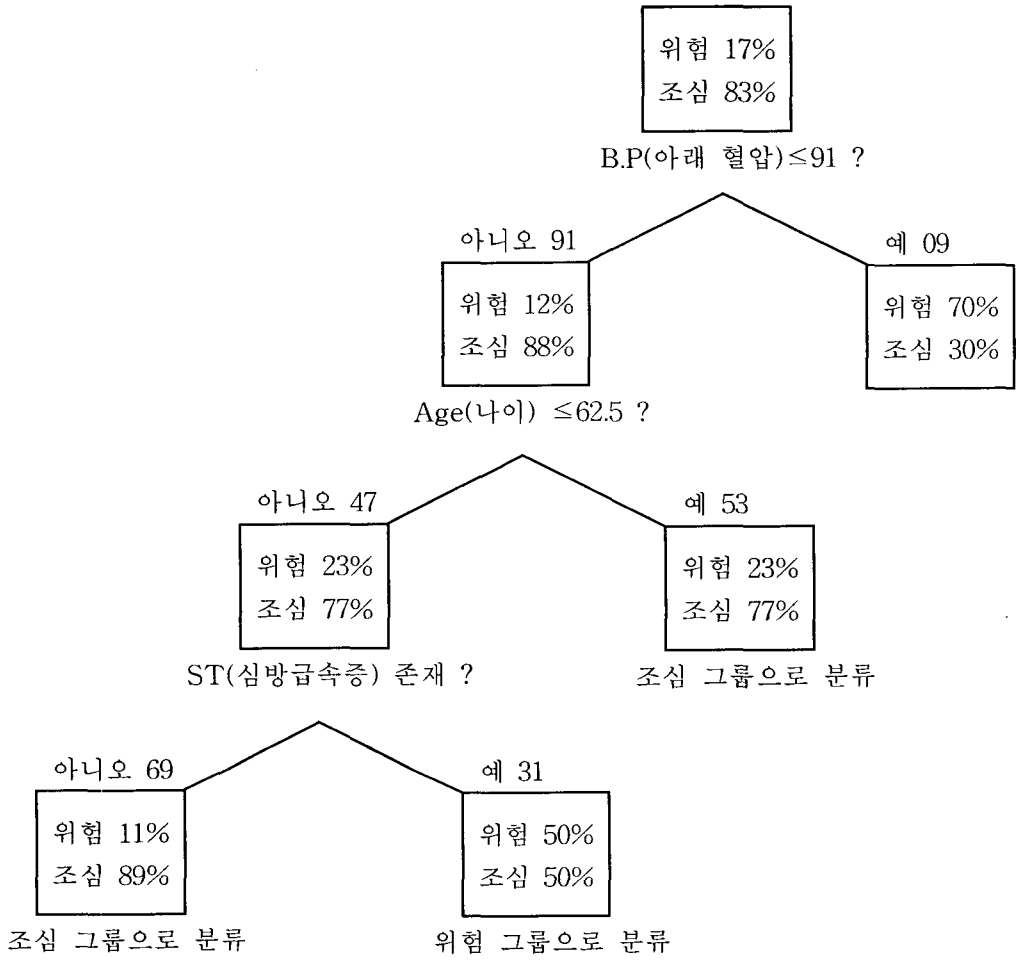
르고, 각각의 자료점을 설명하는 독립변수 또는 설명변수의 개수가 수백 개에서 수천 개에 이르면서 독립변수들 사이에 상호작용 효과까지 기대되는 경우에 기존의 회귀기법에 의한 자료의 분석은 제한적이다. 주요 이유의 하나는 계획행렬의 크기가 너무 커서 최적모형을 찾기 위한 회귀진단, 주성분 분석법 등을 적용하기가 어렵기 때문이다.

자료를 반응변수들의 값에 따라서 동질적인 그룹으로 가장 쉽게 나누는 방법은 각각의 자료점에 대한 설명변수들의 값에 따라서 예/아니오로 대답할 수 있는 일련의 질문을 통해서 자료를 축차적으로 이진(binary) 분할하는 것이다. 물론 이 과정의 각 단계에서 수십, 수백 개의 가능한 이진분리에 대한 평가는 컴퓨터의 활용을 통해서만 가능하다. 의사결정나무는 이 과정을 시각적으로 표현한 나무구조의 그림이다. 반응변수가 범주형 변수인 경우에 이 나무를 분류나무(classification tree)라 하고 반응변수가 계량형 변수인 경우에 회귀나무(regression tree)라 한다.

예를 들어, Breiman et al(1984)에 소개된 샌디에고 의과대학 병원의 응급실에 실려온 심장마비환자들의 자료에 대한 의사결정나무를 생각해 보자. 자료는 215명의 환자들에 대해 처음 24시간 이내에 측정된 나이, 혈압, 심박급속증 등의 진단과 관련된 19개의 설명변수와 처음 30일 이내의 생사에 따른 범주형 반응변수로 구성되어 있다.

그림 1은 사망한 37명의 환자는 '위험'으로, 생존한 178명은 '조심'으로 구분하고 전체 환자들을 진단과 관련된 특정한 설명변수의 값에 따라서 여러 개의 동질적인 소그룹으로 분류하여 얻어진 분류나무이다. 전체 자료는 아래 혈압의 값에 따라서 두 그룹으로 나누어진다. 아래혈압이 91 이하인 환자는 사망자가 70%로 더 많아서 위험그룹으로 분류된다. 아래혈압이 91 보다 높은 나머지 환자 그룹 중에서 나이가 62.5세 이하인 경우는 생존자가 98%로 월등히 많아서 조심그룹으로 분류된다. 마찬가지로 고령에 속하는 나머지 환자들 중에서 심박급속증의 증세가 있으면 위험그룹으로 분류되고 심박급속증의 증세가 없으면 조심그룹으로 분류된다. 정리하면, 3개의 진단변수의 값에 따라서 심장마비환자들이 4 그룹으로 분류되는데, 그 중에서 두 그룹은 사망할 가능성이 높아서 중환자실에 배치되어야 하고, 나머지 두 그룹은 생존자의 비율이 전체환자들 중에서 생존자의 비율인 83% 보다 높아서 입원실에 배치를 고려할 수 있다. 따라서 통계 전문지식이 없는 의사도 응급실에 실려온 신규환자의 3가지 진단 변수의 값을 측정하여 분류나무를 통해서 신규환자가 어느 그룹으로 분류되는 지를 시각적으로 신속히 판단하여 중환자실에 배치여부를 결정하는 것을 도와 줄 수 있게 된다.

의사결정나무의 출발지인 전체 자료 그룹을 뿌리마디(root node)라 한다. 아래혈압 91 이하인 그룹과 같이 분류치가 결정된 그룹을 끝마디(terminal node)라 한다. 아래혈압이 91 보다 높은 그룹과 같이 최종 분류를 위해서 이진 분리(binary split)가 추가로 수행되어야 할 그룹을 중간 마디(internal node)라 한다. 중간마디에서 이진 분리된 두 개의 마디를 자식마디(child node)라 하고 자식마디의 상위마디에 해당되는 중간마디를 부모마디(parent node)라 한다. 중간마디는 축차적인 이진분리를 통해서 끝마디로 분할된다. 뿌리마디에서 끝마디에 도달하는 길(path)이 그 끝마디에 배치된 자료점들의 구조 정보




<그림1> 심장마비환자들의 자료에 대한 분류나무

(structure information)를 제공한다. 반응변수가 범주형 변수(categorical variable)인 분류나무(classification tree)의 경우에 끝마디에서의 예측치는 오분류에 대한 비용을 최소화하는 값으로 결정되는데에 각각의 범주에서 오분류에 대한 비용이 똑 같을 경우에는 그 마디에 배치된 자료점들의 최빈값(mode)이고, 반응변수가 계량변수(quantitative variable)인 회귀나무(regression tree)의 경우에 배치된 자료점들의 평균 또는 중앙값이다.

분류와 회귀나무 예측치(classification and regression tree predictor)들은 구하는 과정이 간단하고 각 자료점은 그 자료점이 속한 끝마디에 도달하는 길(path)에 의해서 설명되어 해석이 쉬운 반면, 예측치가 설명변수

들의 공간의 분할( a partition)인 끝마디에서 같은 값을 갖는 것이 단점이다. 회귀나무의 예측치는 회귀표면의 히스토그램 예측치로 간주되어서 높낮이가 다른 계단함수의 모양이어서, 적절한 모형의 회귀예측치보다 예측력이 떨어질 수 있다. 분류와 회귀나무분석에서는 자동적으로 이진 분리가 이루어져서 각 자료점이 속한 끝마디에 도달하는 길(path)에 의해서 설명변수들로 표현된 구조정보를 시각적으로 얻을 수 있다. 또한 중요한 설명변수와 설명변수들 간의 상호작용효과들을 선별해 낼 수 있게 되어 탐색적 자료분석(exploratory data analysis)의 기법으로 활용할 수 있고, 이 정보를 회귀모형을 결정하는 데에 참조하여서 회귀분석을 하기 위한 보조기법으로도 활용할 수도 있다.

분류와 회귀나무 예측치가 설명변수들의 공간의 분할( a partition)의 경계에서 도약(jump)이 일어날 수 있어서 자료들의 약간의 변화 또는 흔들음(perturbation)이 나무구조나 예측치들의 커다란 변화를 초래할 수 있다는 점에서 나무예측치들이 불안정하다는 사실은 잘 알려져 있다. 불안정성을 개선하기 위한 해결책의 하나는 다중 분류나무와 다중 회귀나무(multiple classification and regression tree)를 생성하는 것이다.

이 논문의 목적은 분류와 회귀나무분석에 관련된 연구결과들을 정리하여 이해하기 쉽게 요약하는 것이다. 2절에서는 분류나무와 회귀나무를 생성하는 알고리즘을 간단하게 소개하고, 분리기준과 나무크기 결정 방법을 정리한다. 3절에서는 다중나무예측치에 대한 알고리즘을 소개하고 4절에서는 적중률을 높이기 위한 축차적인 전략을 설명하고, 신약개발과 관련된 사례의 효율성을 정리한다. 5절에서는 전체 결과를 간략히 요약한다.

## 2. 분리기준과 나무크기 결정

회귀나무와 분류나무를 생성하는 대표적인 알고리즘은 CHi-squared Automatic Interaction Detection (Kass(1980)), Classification And Regression Trees (Breiman 등(1984)), C4.5 (Quinlan(1993))이다. 나무생성원리는 선택된 설명변수의 값에 따라서 각각의 중간마디에서 자료를 그룹내의 원소가 중간마디보다는 더 동질적인 두 그룹(자식마디)으로 분리(split)하는 것이다. 일반적으로 분리의 후보를 결정할 때에 나무가 한쪽 가지로 크게 치우침을 막기 위해서 각각의 자식마디의 크기가 일정조건을 만족하도록 제약조건을 가한다. 나무 생성 알고리즘의 중요 단계는 각 마디에서 두 그룹으로 분리하는 분리기준과 각 마디의 분리를 멈추고, 주어진 마디를 끝마디로 결정하는 정지규칙(stopping rule)에 관한 알고리즘이다. 분리기준은 그룹내의 동질성과 그룹간의 이질적인 특성을 계량적으로 파악하는 기준이고 정지규칙에 따라서 최종 나무가 결정된다.

CHAID는 선택된 설명변수의 값에 따라서 각각의 중간마디에서 자료를 두 그룹으로 분리(split)하는데, 분리기준은 반응변수가 이산형인 경우에는 카이제곱 검정통계량이고, 연속형인 경우에는 F-검정통계량이다. 최적분리는 후보 분리들 중에서 유의확률(p-value)의 값이 가장 작은 분리로 결정되는데 이때의 유의확률의 값이 0.01보다 크면, 분리를 멈추고 해당 중간마디가 끝마디가 된다. 즉, 나무크기를 결정하는 정지규칙은 최적분리의 유의확률의 값에 따라서 결정된다.

CART에서는 분류나무의 경우에 그룹내의 동질성을 불순도 함수(impurity function)에 의해서 측정한다. 불순도 함수는 주어진 그

룹에 대해서 각각의 범주가 나올 확률의 함수로 정의된다. 각각의 범주의 확률이 같은 그룹은 모든 범주를 골고루 포함한 가장 이질적인 그룹이고, 어느 한 범주의 확률이 1인 그룹은 해당 범주만을 포함하는 그룹으로 가장 동질적인 점을 고려하여, 각각의 범주의 확률이 등확률인 경우에 불순도 함수가 최대값을 갖고, 어느 한 범주의 확률이 1인 경우에 최소값을 갖는다. 또한 각 범주의 확률의 순서 바꿈인 순열(permutation)에는 불순도 함수치가 불변하여, 각 범주의 확률에 대해서 대칭(symmetric)인 함수이다. 중간마디 t에서 분리 s가 자료의 비율 $p_R$을 오른쪽 자식 마디 $t_R$로 보내고 자료의 비율 $p_L$을 왼쪽 자식마디 $t_L$로 보내면, 이 분리에 대한 불순도의 차이는

$$\Delta i(s,t) = i(t) - p_R i(t_R) - p_L i(t_L)$$

로 정의되고, 이 값이 분리의 적합도로 간주된다. 분리기준은 불순도의 차이인 $\Delta i(t,s)$를 최대화시키는 분리 $s^*$를 찾는 것이다. 불순도함수의 대표적인 3가지는 다음과 같다.

1. 엔트로피 지수(Entropy index)는 그룹의 정보량(information)의 크기에 관련한 척도로 다항분포의 최우추정량의 극대화를 추구하는 지수로
   $i(t) = -\sum_j p(j|t)\log p(j|t)$로 정의된다.

2. 지니 지수(Gini index)는
   $i(t) = \sum_j p(j|t)(1-p(j|t)) = 1 - \sum_j p^2(j|t)$
   로 불순도 함수를 정의한다. 마디 t에서 범주 j인 원소를 1로 나머지 범주의 원소들을 0으로 할당할 경우에 각각의 범주에 대해서 베르누이 확률변수가 정의되고, 지니지수는 각각의 범주에 대응되는 베르누이 확률변수의 표본분산의 합이어서, 표본분산의 합을 가장 작게 하는 분리를 찾는다고 해석할 수 있다.

3. 투잉 지수(twoing index)는
   $i(t) = [\sum_j |p(j|t_L) - p(j|t_R)|]^2 p_L p_R / 4$ 로 정의되는데 J 개의 범주를 동질적인 두 그룹으로 나누어서 이항 시행 문제로 바꾼 후에 이항 시행의 지니지수를 불순도 함수로 취하여 불순도 함수를 최적화시키는 분리를 찾는 것과 동치인 지수이다.

CART에서는 우선, 훈련용 자료를 가지고 분리기준에 근거한 가장 큰 나무 $T_0$를 생성한 다음에, 말단 가지들 중에서 동질성의 이점이 크지 않은 가지에 대해서 가지치기를 실행하여 부나무(subtrees)들을 생성한다. 이때 가지치기의 기준은 각각의 나무에 대한 비용함수로 나무에 의한 분류예측치의 효율성과 나무구조의 복잡성의 합으로 정의된다. 분류예측치의 효율성을 끝마디에서의 오분류 비용 또는 불순도를 합한 값으로 정의할 수 있고, 복잡성은 나무의 끝마디의 수에 비례하는 선형함수로 표현한다. 복잡성을 결정하는 비례상수 $\alpha$의 값이 작으면 나무구조의 복잡성에 대한 벌칙이 적어서 나무의 크기가 커지고, $\alpha$의 값이 커지면 벌칙이 커짐에 따라서 나무의 크기가 작아진다. 가장 큰 나무인 $\alpha=0$ 에서의 최적나무 $T_0$에서 출발하여 $0<\alpha_1<\alpha_2<\alpha_3<\ldots$에서 바로 앞에 생성된 나무의 가지치기를 통해서 일련의 최적 부나무 $T_1 > T_2 > T_3 > \ldots$들을 생성한다. 생성된 최

적 부나무들 중에서 최종나무를 선택하는 방법은 검증용 자료를 가지고 있는 경우에는 각각의 최적 부나무에서의 검증용 자료에 대한 분류예측치의 효율성을 계산하여서 그 중에서 효율성이 가장 좋은 나무로 결정한다. 총 자료의 크기가 크지 않아서 검증용 자료를 가질 수 없는 경우에는 훈련용 자료를 10개의 그룹으로 랜덤하게 나누어서 각각의 그룹이 검증용 표본이고 나머지 9개의 그룹을 모아서 훈련용 표본으로 취하는 교차타당성 방법을 이용하여 각각의 부나무에서의 효율성을 추정하여 효율성이 가장 좋은 부나무를 최종나무로 취한다.

회귀나무의 경우에 분리기준은 각 마디에서 분류된 관측치와 예측치와의 편차의 제곱합과 절대값의 합으로 이는 최소제곱회귀(least squares regression)와 최소절대회귀(least absolute regression)경우에 대응된다. CART의 가지치기기법과 교차타당성은 수백 개에서 수천 개에 이르는 설명변수를 갖는 대용량 자료에 적용되기에는 계산상의 어려움이 따른다. 설명변수들이 이진(binary) 변수인 경우에 대용량 자료에 효율적인 적용을 위해 Rusinko 등(1999)에 의해 개발된 SCAM(Statistical Classification of Activities of Molecules)은 뿌리마디(root node)를 포함한 모든 중간마디에서 이진분리(a binary split)를 선택하기 위해서 t-검증 기준을 사용한다. 마디마다 각 원자 쌍의 존재 여부에 따라서 두 그룹으로 나누어서 두 그룹간의 모평균을 비교하는 t-검증을 수행한다. 모든 후보 원자 쌍들 중에서 유의확률(p-value)을 최소로 하는 원자 쌍이 자식마디(daughter nodes)로 이진 분리하는 기준의 후보가 된다. 다중비교를 고려한 본페로니 조정 유의확률(Bonferroni adjusted p-value)이 유의하지 않으면, 그 마디에서 더 이상 분리가 일어나지 않고, 그 마디가 끝마디가 된다.

## 3. 다중나무예측치 알고리즘

다중나무예측치 알고리즘의 기본적인 개념은 모평균을 추정할 때에 안정성을 높이기(분산을 작게 하기) 위해서 확률표본을 생성하여 표본평균으로 모평균을 추정하는 원리와 같다. 다중나무를 생성하는 방법으로 Breiman(1996)이 제안한 Bagging은 분석용 자료(training data)의 붓트스랩 표본(bootstrap sample)들을 생성하고 각 표본에 근거하여 나무를 생성한다. 또 다른 방법은 분석용 자료를 가지고 나무를 생성할 때에 각각의 중간마디(internal node)에서 처음 몇 개의 최적분리들 중에서 랜덤하게 하나를 선택하여 나무를 생성하는 것이다. 즉, 분석용 자료를 고정시키고 나무의 중간마디를 결정시에 약간의 흔들음을 준다. 이 과정을 여러 번 반복하면 반복할 때마다 다른 구조를 가진 나무가 생성된다. (Kwok and Carter(1990), Tatsuoka et al(1999)) 분류 다중나무예측치는 각각의 나무예측치들의 빈도가 가장 높은 최빈값이고 회귀 다중나무예측치는 각각의 나무예측치들의 평균이다.

분류의 경우에 arcing 분류예측치는 Freud and Schapire(1996)에 의해 소개되었는데, Bagging과 마찬가지로 다중나무들을 생성하기 위해서 검사용 자료로부터 표본을 복원재추출(resample)한 붓트스랩 표본에 근거하여 나무들을 생성한다. Bagging과의 차이점은 각각의 자료점이 붓트스랩 표본에 뽑힐 확률이 등확률로 출발하여 앞의 단계에서 생

성된 분류나무가 오분류(misclassification)하는 자료점에 대해서는 축차적으로 그 다음 순서의 표본에 뽑힐 확률을 크게 수정하여서 오분류될 자료점의 중요성을 높이고, 다중 분류예측치는 각각의 분류나무 예측치의 가중투표(weighted voting)에 의해서 결정한다. 이때 가중치는 각각의 분류나무 예측치의 오분류률에 따라서 결정된다. Breiman(1997)은 검증용 자료(test data)의 오분류률로 평가할 때에 arcing이 bagging보다 더 효과적이라는 사실을 예증하였다.

다중나무예측치는 안정적이고 효율적인 반면에, 구조정보를 잃어버려서 그림으로 표현할 수 없고, 예측치를 쉽게 해석할 수 없다는 것이 단점이다. Bumping(Bootstrap Umbrella of Model Parameters)은 Bagging에서 생성된 분석용 자료 붓트스랩 표본에 근거한 나무들이나 분석용 자료에 근거한 나무를 생성 시에 각각의 중간마디에서 처음 몇 개의 최적분리들 중에서 랜덤하게 하나를 선택하여 생성되는 나무들로 후보 나무들의 묶음(pool)을 만든 후에 이 중에서 하나의 나무를 선택하는 것이다. 예를 들면, 검증용 자료의 오분류률을 가장 작게 하는 나무를 선택할 수도 있다. 따라서 나무구조를 유지하면서 과대적합(overfitting)을 방지하는 적절한 나무를 선택할 수 있다는 장점이 있다.

## 4. 축차적인 전략

연구 개발 단계에서 대규모의 자료를 처리하고 분석해야 하는 대표적인 경우가 신약개발 (development of new drugs)의 예이다. 항암제나 에이즈 치료를 위한 백신개발 등의 연구 개발의 처음 단계에서의 목표는 질병치료에 도움이 될 생물학적으로 효능 있는 화합물 (potent molecules)들을 찾는 것이다. 조합 화학(combinatorial chemistry)의 기여로 인하여 생물학적으로 효능 있는 화합물(potent molecules)들의 후보가 될 가능성이 있는 수십 만개 혹은 수백만 개의 화합물들의 화학구조들을 수십 개의 조립 블록(building blocs)을 이용하여 생성할 수 있고, 이들 정보로 구축된 데이터 베이스만이 활용가능한 경우를 생각하자. 각 화합물의 화학구조는 비교적 간단한 위상학적 묘사(topological description)인 수천 개의 이진 설명변수로 표현될 수 있다. 데이터 베이스에 구축된 모든 화합물들의 합성과 검사에는 많은 비용과 시간이 소요되어 현실적으로 불가능하다. 따라서 소규모 양의 합성과 검사로 얻어진 자료를 가지고 회귀나무를 이용하여 반응치를 예측하여 검사되지 않은 화합물 중에서 효능이 있으리라 기대되는 화합물들을 선별하여 다음 단계의 검사될 화합물들의 목록을 작성하여 반응치를 측정한 후에 앞에 이미 검사된 화합물 자료에 함께 묶어서 회귀나무를 실행시켜서 반응치의 예측치를 수정하는 과정을 반복하는 축차적인 찾기를 통한 체계적인 접근방법은 실용적인 가치가 매우 높다. 축차적인 전략의 효율성은 축차적인 전략을 통해서 선별된 총 화합물들 중에서 실제로 효능이 있는 화합물들의 비율인 적중률(hit rate)에 의해서 평가될 수 있다. Abt 등(2001)은 사례(case studies)들을 중심으로 적중률이 높은 효율적인 예측방법을 제공하는 축차적인 전략에 대해 연구하였는데, 인용된 자료와 연구결과를 요약하면 다음과 같다.

축차적인 전략을 위해 사용된 자료는 52883개의 화합물 각각의 화학구조들에 대한

정보와 생물학적 활동성(biological activity)을 측정한 반응치를 포함하는 Glaxo Wellcome 자료이다. 반응치는 질병의 치유에 영향을 주리라 기대되는 단백질과의 결합 능력을 측정한 계량값이다. 화합물들의 화학구조를 표현하는 방법으로 Carhart 등(1985)이 제시한 원자 쌍(atom pair)에 근거한 9079 개의 이진(binary) 독립변수들을 사용한다. 한 원자 쌍은 한 쌍의 비수소 원자(non-hydrogen atoms)와 이 원자들을 연결하는 최소 위상학적 거리(minimum topological distance)로 이루어진다. 최소 위상학적 거리는 두 원자 사이를 연결하는 최단 통로(shortest bond path)에 있는 원자들의 수이다. 각각의 원자 쌍은 <원자 1 묘사> - <위상학적 거리> - <원자 2 묘사>의 모양을 갖춘다. 총 52883개의 화합물들이 갖고 있는 원자 쌍들을 모두 찾아보니 총 9079개의 원자 쌍이 발견되었다. 각 원자 쌍이 0, 1의 값을 갖는 이진 독립변수로 간주되어서, 각각의 화합물들은 길이가 9079인 0과 1 숫자들의 배열(bitstrings)로 표시된다. 1은 대응되는 원자 쌍이 그 화합물에 존재하고, 0은 존재하지 않음을 의미한다. 대부분의 화합물들은 총 9079개의 원자 쌍들 중에서 약 200개 미만을 포함한다.

1. 일 단계로 총 52,883 개의 화합물 목록에서 랜덤하게 2,500 개의 화합물을 선택하여 반응치를 얻는다. 일 단계 자료의 크기 2,500 은 전문가가 제시하였고, 랜덤하게 선택된 2,500개의 자료를 가지고 회귀나무 알고리즘인 SCAM을 실행시킨 결과 얻어진 회귀나무가 몇(3) 개 이상의 예측치의 값이 큰 즉, 바람직한 끝마디(good terminal nodes) 를 가지고 있어서 일 단계 표본의 크기로 선택된 값이다.

2. 2 단계에서의 표본의 50%는 검사되지 않은 나머지 화합물들 중에서 예측치의 값이 큰 화합물들을 선택하고, 나머지 50%는 반응치의 값이 큼에도 불구하고 오분류된 화합물들과 화학구조가 유사한 화합물들 중에서 선택하는 것이다. 즉, 나머지 50,383 개의 화합물들을 일 단계에서 구한 나무 예측치의 크기 순서로 나열하여 처음 1,250개를 선택하고, 나머지 1,250개는 일 단계의 회귀나무에서 화합물들의 구조를 인지하지 못한 회귀나무의 가장 왼쪽 끝마디에 속해 있는 일 단계 자료들 중에서 반응치가 큰 화합물들을 찾아내어서 이 화합물들과 화학구조가 가장 비슷한 화합물들을 나머지 선택되지 않은 화합물들 중에서 찾아내어 추가로 1,250개를 선택한다.

3. 1단계의 표본과 2단계의 표본을 함께 묶어서 총 5,000 개의 자료를 가지고 SCAM을 실행시켜서 얻은 회귀나무에 나머지 47,883 개의 검사되지 않은 화합물들을 떨어뜨려서 바람직한 끝마디에 속한(예측치의 값이 큰) 화합물들에 대해서 추가로 검사를 실시하여 반응치를 얻는다.

전문가들은 화합물들의 화학구조에 관한 정보를 활용하여 선택된 표본이 대량 고속 선별의 효율성에 영향을 준다는 견해를 가지고 있었는데, 축차적인 전략에서는 랜덤하게 선택된 일 단계 표본이나 화학 구조에 근거한 체계적인 선택 방법이나 효율성이 비슷

하다는 결과를 얻었다. 그 이유의 하나는 회귀나무의 구조정보를 활용하여 합리적으로 선택된 2단계의 표본이 1단계의 표본을 보완하여 총 5,000 개의 전체 표본에 근거한 회귀나무의 예측성의 정도가 유사하리라고 기대되기 때문이다. 축차전략의 효율성의 측도인 적중율은 2가지 에세이에 대해서 랜덤한 대량 고속 선별에 비해서 각각 약 2.5, 4 배가 높았고, 일 단계 표본이 선택된 후에 나머지 검사된 화합물들 중에서 효능이 있는 화합물들의 비율인 이익률(gain rate)은 약 3.1, 6 배가 높아서 축차전략이 효율적임을 예증하였다.

## 5. 요 약

대용량 자료를 다루는 데이터 마이닝에서 자료분석의 주요 도구인 분류나무와 회귀나무는 모형이 나무구조의 그림으로 시각적으로 표현되어서, 모형의 해석이 쉽다는 이점을 가지고 있다. 자료들의 약간의 변화 또는 흔들음(perturbation)이 나무구조나 예측치들의 커다란 변화를 초래할 수 있다는 점에서 나무예측치들이 불안정하다는 사실은 잘 알려져 있다. 불안정성을 개선하고 예측치의 정밀성을 높이기 위한 해결책의 하나는 다중나무예측치를 사용하는 것이다. 이 때 치루어야할 대가는 예측치가 더 이상 나무구조로 표현되지 않아서 해석의 용이성을 잃는 것이다. 항암제나 에이즈 치료를 위한 백신개발 등의 연구 개발의 처음 단계에서의 목표는 질병치료에 도움이 될 생물학적으로 효능있는 화합물 (potent molecules)들을 합성되지 않은 수십만 개의 화합물들 중에서 찾는 것이다. 데이터 베이스에 구축된 모든 화합물들의 합성과 검사에는 많은 비용과 시간이 소요되어 현실적으로 불가능하다. 따라서 소규모 양의 합성과 검사로 얻어진 자료를 가지고 회귀나무를 이용하여 반응치를 예측하여 검사되지 않은 화합물 중에서 효능이 있으리라 기대되는 화합물들을 선별하는 과정을 반복하는 축차적인 전략은 실용적인 가치가 매우 높은 체계적인 방법이다.

## 참고문헌

[1] 강현철 등(1999), 「데이터마이닝, 방법론 및 활용」, 자유아카데미.

[2] 임용빈, 이소영, 정종희(2001), "대용량 화학 데이터 베이스를 선별하기 위한 결합다중회귀나무 예측치", 「응용통계연구」, 14권(1호), pp. 91-101.

[3] Abt, M., Lim, Y.B., Sacks, J., Xie, M. and Young, S. (2001), A sequential approach for identifying lead compounds in large chemical databases, Accepted for publication in Statistical Sciences.

[4] Breiman, L.(1996). Bagging predictors, *Machine Learning,* vol. 26, No. 2, 123-140.

[5] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and regression trees*, Chapman and Hall, Belmont, CA, Wadsworth.

[6] Breiman L, (1997). Arcing Classifiers. ftp://ftp.stat.berkeley.edu pub/breiman/arc97.ps.

[7] Freund, Y. and Schapire,R. (1996). Experiments with a new boosting

algorithm, *Machine Learning*: Proceedings of the Thirteenth International Conference, July, 1996.

[8] Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data, Applied Statistics, vol. 29, 119-127

[9 Kay Tatsuoka, Chong Gu, Jerome Sacks and S. Stanley Young (1999). Prediction Extreme Values in Large Datasets, Accepted for publication in J. Compt. Graph. Statist.

[10] Kwok, S. and Carter, C. (1990). Multiple decision trees, *Uncertainty in Artifical Intelligence*, vol. 4, 327-335.

[11] Quinlan, J.R. (1993). C4.5 Programs for machine learning. San Mateo: Morgan Kaufmann.

[12] Rusinko, A., Farmen, M., Lambert, C. Brown, P., Yound, S. (1999), Analysis of a large structure/biological activity data set using recursive partitoning, *J. Amer. Chem. Soc.*. vol. 40, 1017-1026