응용통계연구 제 **14**권 *1*호, *2001*년, *pp. 91-101*

대용량 화학 데이터 베이스를 선별하기 위한 결합다중회귀나무 예측치 *

임용빈[1] 이소영[2] 정종희[3]

요 약

다중나무예측치들이 한 개의 나무 예측치 보다 검증용 자료 오분류률을 줄이는데 있어서 더 정확하다 라는 것은 잘 알려져 있는 사실이다. 다중나무를 생성하는 두 가지 방법이 있다. 하나는 원래의 훈련용 자료를 재 추출하여 수정된 훈련용자료들을 만든 다음에 각각의 수정된 훈련용 자료에 근거하여 나무를 만드는 것이다. arcing 알고리즘이 효율적이라고 알려져있다. 다른 방법은 각각의 마디에서 최적 분리의 후보들 중에서 랜덤하게 하나를 선택하여 나무를 생성하는 데에, 이 과정을 반복하면 원래의 훈련용 자료에 대해서 비교적 좋은 나무들을 생성하리라 기대된다. 우리는 arcing의 각 단계에서 후자의 다중회귀나무예측치들을 사용하는 결합다중회귀나무예측치를 제안하고, 효능 있는 화합물들을 찾기 위한 고속의 대량 선별 자료 분석의 예를 통해서 예측방법들의 효율성을 비교한다.

주요용어: 결합다중회귀나무예측치, 대량 고속선별.

1. 서론

컴퓨터의 발전과 더불어서 관련된 많은 정보들의 데이터 베이스화가 가능하여져서 대규모의 자료를 처리하고 분석하기 위한 통계적인 방법의 필요성이 대두된다. 연구 개발 단계에서 대규모의 자료를 다루어야 하는 대표적인 경우가 신약 개발 (development of new drugs)의 예이다. 항암제나 에이즈 치료를 위한 백신개발 등의 연구 개발의 처음단계에서는 합성 가능한 화합물(compounds)들의 화학구조들만의 정보로 구축된 데이터 베이스만이 활용 가능하다. combinatorial chemistry의 기여로 인하여 질병치료에 도움이 되는 생물학적으로 효능 있는 화합물 (potent molecules)들의 후보가 될 가능성이 있는 수십 만개 혹은 수백만 개의 화합물들의 화학구조들을 수십 개의 building blocs을 이용하여 생성할 수 있다. 각 화합물의 화학구조는 비교적 간단한 위상학적 묘사(topological description)인 수천 개의 설명변수로 표현될 수 있다. 데이터 베이스에 구축된 모든 화합물들의 합성과 검사에는 많은 비용과 시간이 소요된다. 따라서 소규모 양의 합성과 검사로 얻어진 자료를 가지고 반응치를 예측하여 검사되지 않은 화합물 중에서 효능이 있으리라 기대되는 화합물들

* 이 논문은 1998년도 한국과학재단 핵심연구비 지원(과제번호 KOSEF 981-0105-024-2)에 의한 연구결과임.
1) (120-750) 서울 서대문구 대현동 11-1, 이화여자대학교 자연과학대학 통계학과, 교수
E-mail: yblim@mm.ewha.ac.kr
2) (120-750) 서울 서대문구 대현동 11-1, 이화여자대학교 자연과학대학 통계학과, 석사과정
3) (120-750) 서울 서대문구 대현동 11-1, 이화여자대학교 자연과학대학 통계학과, 석사과정

을 선별하고, 선별된 화합물들을 합성하고 반응치를 측정하여 효능 있는 화합물인지를 판단한다. 예측치(predictors)들의 효율성은 선별된 총 화합물들 중에서 실제로 효능이 있는 화합물들의 비율인 적중률(hit rate)에 의해서 평가될 수 있다. 적중률이 높은 효율적인 예측방법에 대한 실용적인 가치는 매우 크다.

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 최적모형을 찾기 위한 회귀진단, 주성분 분석법 등을 적용하기가 어렵기 때문이다. 중요 독립변수와 그들의 상호작용효과를 판별할 수 있는 쉽고 자동적인 분석방법이 회귀나무구조의 접근방법 (Regression Tree-structured approach)인 축차 분할 (Recursive Partitioning)이다. 축차분할에서는 독립변수들의 공간(space of independent variables)이 축차적인 분리에 의해서 끝마디(terminal nodes)로 분할된다. 끝마디에 도달하는 길(path)이 그 마디에 배치된 자료점들의 구조 정보(structure information)를 제공한다. 반응변수가 계량변수(quantitative variable)인 회귀나무(regression tree)의 경우에 끝마디에서의 예측치는 그 마디에 배치된 자료점들의 평균이고, 반응변수가 범주형 변수(categorical variable)인 분류나무(classification tree)인 경우에 끝마디에서의 예측치는 배치된 자료점들의 최빈값(mode)이다.

분류와 회귀나무 예측치(classification and regression tree predictor)들은 구하는 과정이 간단하고 각 자료점은 그 자료점이 속한 끝마디에 도달하는 길에 의해서 설명되어 해석이 쉬운 반면 자료들의 약간의 변화 또는 흔들음(perturbation)이 나무구조나 예측치들의 커다란 변화를 초래할 수 있다는 점에서 불안정하다는 사실은 잘 알려져 있다. 불안정성을 개선하기 위해서 다중 분류나무와 회귀나무(multiple classification and regression tree)를 생성한다. 분류의 경우에 주어진 자료점에서의 다중예측치는 각각의 나무에서의 예측치들의 최빈값이고, 회귀의 경우에는 평균이다. 다중나무를 생성하는 방법으로 Breiman(1996)이 제안한 bagging은 분석용 자료(training data)의 붓트스랩 표본(bootstrap sample)들을 생성하고 각 표본에 대해서 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 결정 시에 약간의 흔들음을 준다. 이 과정을 여러 번 반복하면 반복할 때마다 다른 구조를 가진 나무가 생성된다. (Kwok 와 Carter(1990), Tatsuoka 등(1999)) 분류의 경우에 검사용 자료로부터 표본을 복원 재추출(resample)할 때에, 각 자료점이 뽑힐 확률이 등확률로 출발하여 생성된 분류나무가 오분류(misclassification)하는 자료점에 대해서는 축차적으로 그 다음 순서의 표본에 뽑힐 확률을 크게 수정하면서 축차적으로 분류나무들를 생성하고, 다중 분류예측치는 각 분류나무 예측치의 가중투표(weighted voting)에 의해서 결정하는 arcing 분류예측치는 Freud 와 Schapire(1996)에 의해 소개되었다. Breiman(1997)은 검증용 자료(test data)의 오분류률로 평가할 때에 arcing이 bagging보다 더 효과적이라는 사실을 예증하였다.

이 논문에서는 분석용 자료의 재추출에 의한 다중나무예측치인 arcing과 주어진 분석용

자료를 고정시키고 구해지는 Tatsuoka 등(1999)의 다중나무예측치를 결합한 결합다중 회귀나무 예측치 (combined multiple regression tree predictor)를 제안한다. 먼저 분류에 적용되는 arcing을 회귀에 적용시키기 위해서 각 끝마디의 자료점들의 반응치가 평균보다 3σ를 벗어난 상위 반응치이면, 자료점이 오분류되었다고 정의한다. 우리는 반응치가 큰 화합물의 선별에만 관심이 있기에 오분류를 상위 반응치로만 한정한다. 그 다음에 arcing의 각 단계에서 추출된 표본에 근거한 예측치로는 각 마디에서 처음 몇 개의 최적분리들을 랜덤하게 선택하여 다중나무를 생성하여 구해지는 Tatsuoka 등(1999)이 제안한 다중예측치를 사용한다. 2절에서는 다중나무예측치들의 효율을 비교하기 위해서 적용된 자료를 소개하고 3절에서는 다중나무예측치들을 간략하게 설명하고 다중나무예측치들의 성능 평가기준으로 상대 적중률(relative hit rate)을 정의한다. 4절에서는 2절에 소개한 자료를 통해서 회귀나무예측치, arcing, Tatsuoka 등(1999)의 다중나무예측치와 이 논문에서 제안한 결합다중나무예측치들의 성능을 비교한다.

2. 자료

예측치들의 성능비교를 위해 사용된 자료는 53203개의 화합물 각각의 화학구조들에 대한 정보와 생물학적 활동성(biological activity)을 측정한 반응치를 포함하는 Glaxo Wellcome 자료이다. 반응치는 질병의 치유와 관련이 있다고 연구된 단백질과의 결합능력을 측정한 계량값이다. 화합물들의 화학구조를 표현하는 방법으로 Carhart 등(1985)이 제시한 원자 쌍(atom pair)에 근거한 위상학적인 묘사의 결과인 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. 원자의 묘사는 원자에 부착된 본드의 수와 이 본드들에 있는 $\pi-$전자들의 수로 구성된다. 예를 들면 두 개의 비수소 원자에 부착되고 이웃과 한 개의 $\pi-$전자를 공유하는 탄소원자는 C(2,1)로 표시된다.

총 53203개의 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. 이 자료의 특징은 대부분의 원자 쌍들이 아주 드물게 나타난다는 사실이다. 약 40% 의 원자 쌍들이 10개 이하의 화합물에서만 나타난다.

3. 다중나무예측치 알고리즘

축차분할(recursive partitioning)의 잘 알려진 알고리즘은 Breiman 등(1984)이 제시한 CART(Classification And Regression Trees)이다. CART는 최종 나무의 크기와 끝마디들을 결정하기 위해서 교차타당성(cross-validation)과 가지치기 기법(pruning technique)을 사용

하는데, 2절에서 소개된 대용량 자료에 적용되기에는 계산상의 어려움이 따른다. 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 두그룹으로 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.

arcing(adaptively resample and combine)은 분류에 적용되는 다중분류예측치이다. arcing의 개념을 회귀에 적용시키기 위해서, 각 끝마디의 자료점들의 반응치가 끝마디에서의 예측치인 평균보다 3σ를 벗어난 상위 반응치이면, 그 자료점이 오분류되었다고 정의한다. 분석용 자료의 복원 추출을 통하여 각 단계의 분석용 자료가 결정되고, 회귀나무예측치가 얻어진다. 그런데 (k+1)번째 분석용 자료를 추출할 때, 최초의 분석용 자료의 각각의 자료점들이 뽑힐 확률은 앞의 k개의 단계에서 회귀나무예측치들이 그 자료점을 제대로 분류했는지에 따라서 뽑힐 확률이 수정된다. 그 다음에 arcing 회귀예측치는 각 단계에서 얻어진 회귀나무예측치들의 가중평균(weighted average)에 의해 결정된다. arcing의 구체적인 알고리즘은 다음과 같다.(Freud 와 Schapire(1996), Breiman(1997))

분석용 자료를 T라 하자.

1) k-번째 단계에서 T의 각각의 자료점이 뽑힐 확률 $\{p^{(k)}(n)\}$ 에 따라서 복원 추출된 분석용 자료를 $T^{(k)}$를 구하고 분석용 자료$T^{(k)}$를 사용하여 회귀나무 $R_{(k)}$를 구한다.

2) T의 n-번째 자료점이 회귀나무 $R_{(k)}$에서 오분류되면 d(n)=1, 제대로 분류되면 0이라 한다.

3) $\varepsilon_k = \sum_n p^{(k)}(n)d(n)$, $\beta_k = (1-\varepsilon_k)/\varepsilon_k$ 로 정의하고 (k+1)-번째 단계에서 n-번째 자료점이 뽑힐 확률은 $p^{(k+1)}(n) = p^{(k)}(n)\beta_k^{d(n)} / \sum_i p^{(k)}(i)\beta_k^{d(i)}$로 수정된다. 총 K개의 회귀나무가 생성된 후, arcing 회귀예측치는 각 단계의 회귀예측치들에 가중치 $\log(\beta_k)$를 주어 가중평균을 구한다.

Tatsuoka 등(1999)은 분석용 자료를 고정시키고, 각각의 중간마디마다 t-검증의 유의확률을 가장 작게 하는 유의한 b개의 최적분리들 중에서 한 개를 랜덤하게 선택하여 회귀나무를 생성하는 과정을 K번 반복하여 K개의 회귀나무를 생성한다. 이 과정을 반복할 때마다 다른 구조를 가진 나무가 생성되리라 기대된다. Tatsuoka 등(1999)의 다중회귀예측치인 averaging은 K개의 회귀나무예측치들의 평균으로 구해진다.

우리가 새로이 제시하는 결합다중회귀예측치는 분석용 자료의 재추출에 의한 다중나무예측치인 arcing과 주어진 분석용 자료를 고정시키고 구해지는 Tatsuoka 등(1999)의 averaging을 결합하여 구해진다. 앞의 arcing 알고리즘에서 2)를 다음과 같이 수정한다. arcing의 k-번째 단계에서 분석용 자료$T^{(k)}$를 고정시키고 Tatsuoka 등(1999)의 방법에 따라서 적당한 수의 회귀나무를 생성한다. T의 n-번째 자료점에서의 예측치는 각각의 회귀나무에서의

끝마디의 예측치들의 평균이고, σ는 SCAM을 실행시켜서 얻은 회귀나무에서 n-번째 자료점이 배치된 끝마디의 표본표준편차로 추정하여서 n-번째 자료점의 오분류를 결정한다.

4. 평가기준과 성능비교

일반적으로 예측방법들을 비교하는 기준은 평균제곱오차(mean squared error)이다. 평균제곱오차 기준은 전체 자료점에서의 모형의 적합성을 평가한다. 그런데 2절에서 소개된 바와 같은 신약 개발과 관련된 자료에서는 전반적인 자료의 적합성보다는 평균 반응치가 상위 극단에 속하는 화합물들의 선별에만 관심이 있다. 따라서 분석용 자료에 근거하여 얻어진 예측방법들을 검사용 자료(test data)에 있는 화합물들에 적용시켜서 검사용 자료의 화합물들을 예측치의 크기 순서로 나열하여 예측치가 가장 큰 화합물부터 차례차례 검사한다. 검사한 자료들 중에서 실제로 반응치가 상위 극단에 속하는 화합물들의 비율을 적중률(hit rate)이라 하고, 상대적중률(relative hit rate)은 이 적중률과 랜덤하게 검사할 때에 기대되는 적중률과의 비율로 정의된다. 이 논문에서는 상대적중률에 의해서 예측방법들을 비교하려 한다.

데이터 베이스에 있는 화합물들 중에서 반응치가 대략 상위 0.5%이내에 속하는 화합물들이 화학자가 선별해 내고 싶어하는 효능 있는 화합물들이라고 한다. 2절에서 소개된 Glaxo Wellcome자료는 앞에서 소개한 바와 같이 반응치의 측정결과가 이미 알려진 자료이다. 먼저 총 자료 중에서 반응치의 값이 상위 약 0.5%에 해당되는 250개의 화합물들을 구별해 낸다. 예측방법들의 효율성을 비교하기 위해서 53203개의 화합물들 중에서 랜덤하게 2500개를 추출하여 분석용 자료를 얻는다. 예측방법의 효율성이 랜덤하게 선택된 분석용 자료에 따라서 영향을 받을 수도 있기에 두 번째 분석용 자료를 50703개의 화합물들 중에서 랜덤하게 선택하고, 나머지 48203개의 화합물을 검사용 자료로 취한다. 각각의 분석용 자료에 근거하여 검사용 자료의 화합물들에 대한 예측치를 구하고, 예측치가 가장 큰 화합물부터 차례차례 검사하여, 그 중에서 실제로 선별해낸 상위 250개 화합물들의 적중률과 상대적중률을 계산한다. 동일 과정을 2회 반복하여 총 4개의 분석용 자료로부터 얻어진 상대적중률의 평균을 구한다.

그림 4.1은 2500개의 분석용 자료에 근거하여 얻어진 SCAM 회귀나무이다. 중간마디의 이진 분리를 위한 유의수준은 1%이고, 자식마디의 크기를 적어도 10으로 하는 원자 쌍들만이 분리의 후보로 정하였다. 회귀나무의 가장 상단의 마디(top node)가 뿌리마디(root node)이다. 뿌리마디는 2500개의 분석용 자료로 구성된다. 각 후보 원자 쌍들의 유(1), 무(0)에 따라서 2500개의 분석용 자료가 두 그룹으로 나누어진다. 두 그룹간의 모평균의 차이를 비교하는 t-검증의 유의확률을 최소로 하는 원자 쌍이 265번 원자 쌍이다. 뿌리마디의 2500개 분석용 자료들 중에서 265번 원자 쌍이 존재하는 화합물들은 오른쪽 마디로 분류되고, 265번 원자 쌍을 갖지 않는 화합물은 왼쪽의 마디로 분류된다. 이 정보가 뿌리마디에서 오른쪽 마디를 연결하는 직선 상에 265:1로, 왼쪽마디를 연결하는 직선 상에 265:0으로 각각 표시된다. 마디의 모양이 타원형인 것은 중간 마디(internal node)를 뜻하며, 사각형인 것은 끝마디(terminal node)를 나타낸다. 마디 안의 숫자들은 그 마디에 속한 반응치

들의 평균이며, 중간마디 아래에는 그 마디에 속한 화합물들의 반응치들로부터 계산된 표준편차와 화합물들의 개수가 표시되어 있다. 끝마디 아래에는 표준편차, 화합물들의 개수 외에 그 마디에 속하는 화합물 중에서 상위 250개에 해당되는 화합물들의 개수가 나타나 있다. 이 회귀나무에 48203개의 검사용 자료를 떨어뜨려서 중간마디를 연결하는 원자쌍들의 유, 무에 따라서 끝마디에 배치한다. 끝마디를 예측치의 크기순서로 나열하고, 예측치가 가장 큰 끝마디에 배치된 화합물부터 랜덤하게 하나씩 선택하여 검사하고, 검사가 모두 끝나면, 그 다음 순서의 끝마디에 대해서 동일한 과정을 반복한 후에 상대적중률을 계산한다.

3절에 소개된 arcing 알고리즘에 따라서 1000개의 회귀나무를 생성하여 arcing 회귀예측치를 구한다. 각 단계의 분석용 자료 에 근거한 회귀나무를 구하는 알고리즘으로 SCAM을 사용한다.

그림 4.2는 동일한 분석용 자료를 고정시키고, 각각의 중간마디마다 유의한 5개의 최적분리들 중에서 랜덤하게 하나를 선택하여 생성된 회귀나무이다. 이와 같이 1000개의 회귀나무를 생성하고, 각각의 회귀나무에 48203개의 검사용 자료를 떨어뜨려 한 화합물마다 1000개의 예측치를 얻게 한다. 그 1000개의 예측치들의 평균이 averaging에 의한 다중예측치이다.

약 1000개의 회귀나무들의 결합에 의한 결합다중예측치를 구하기 위해서 arcing 알고리즘에서 단계의 수를 33 단계로 한다. 각각의 단계에서 훈련용 자료를 재추출하고, 고정된 분석용 자료에 근거하여 30개의 회귀나무를 생성하여 계산된 averaging에 의한 예측치가 arcing의 각 단계에서의 회귀예측치이다. 따라서 결합다중예측치는 총 990개의 회귀나무들을 생성하여 구해진다.

그림 4.3은 앞에서 소개한 4가지 방법인 SCAM 예측치, averaging 다중회귀나무예측치, arcing 다중회귀나무예측치, 그리고 결합다중회귀나무 예측치를 이용하여 계산되어진 상대적중률을 그래프로 나타낸 것이다. 검사용 자료에 있는 화합물들 중에서 검사를 위해서 선별되는 화합물들의 순서는 각각의 예측방법에 따른 예측치들의 크기순서로 검사용 자료의 화합물들을 나열하여 검사 순서가 결정된다. 검사용 자료의 극히 일부분을 선별하여 실제로 효능 있는 화합물들을 얼마나 찾았는가, 즉 상대적중률에 의해서 예측방법의 실용적인 가치가 평가된다. 따라서 반응치가 상위 0.5%에 속한 화합물들이 효능 있는 화합물이란 사실을 고려할 때에, 각각의 예측방법에 대해서 약 1% 이내의 선별된 검사용 자료에 있는 화합물들에 대한 상대적중률의 크기 변화가 우선적으로 관심대상이 된다. 그림 3을 보면, 처음에는 4가지 방법에 따라 상대적중률의 차이가 있지만, 점점 검사비율이 증가할수록 상대적중률은 차이가 별로 없어지고, 검사비율이 1이면 모두 다 검사하였기에 랜덤적중률과 같아져서 1로 수렴하리라 기대된다. 검사비율 0.01 이하에서의 4가지 방법의 상대적중률을 비교하면 결합다중예측치가 averaging이나 arcing보다 검사비율 0.008 까지는 대체로 우월하고, 0.008~0.01사이에서는 averaging과 비슷하여, 우리가 기대한 대로 결합다중예측치인 arc-averaging이 arcing과 averaging에 의한 예측방법들 보다 다소 우월함을 알 수 있다. 이는 본 논문에서 새로 제시한 결합다중회귀나무 예측방법이 앞으로의 대용량자료분석에서 예측방법들의 하나로 고려될 수 있음을 예증한다.

감사의 글

본 논문의 연구분야에 관심을 갖게 해준 Jerry Sacks, Stan Young과 averaging에 의한 다중나무예측치의 code를 제공해준 Kay Tatsuoka에 감사를 드린다.

참고문헌

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

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

[3] Breiman, L. (1997). Arcing Classifiers.

ftp://ftp.stat.berkeley.ed/pub/breiman/arc97.ps.

[4] Freund, Y. and Schapire, R. (1996). Experiments with a newboosting algorithm, Machine Learning: Proceedings of the Thirteenth International Conference, July, 1996.

[5] Tatsuoka, K., Gu, G., Sacks, J. and Young, S.S. (1999). Prediction Extreme Values in Large Datasets, Accepted for publication in J. Compt. Graph. Statist.

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

[7] Rusinko, A., Farmen, M., Lambert, C. Brown, P. and Yound, S. (1997). Analysis of a large structure/biological activity data set using recursive partitoning, submitted to J. Amer. Chem. Soc..

[8] Venables, W.N. and Ripley, B.D. (1994). Modern Applied Statistics with Splus, Springer, New York.

[2000년 6월 접수, 2000년 12월 채택]

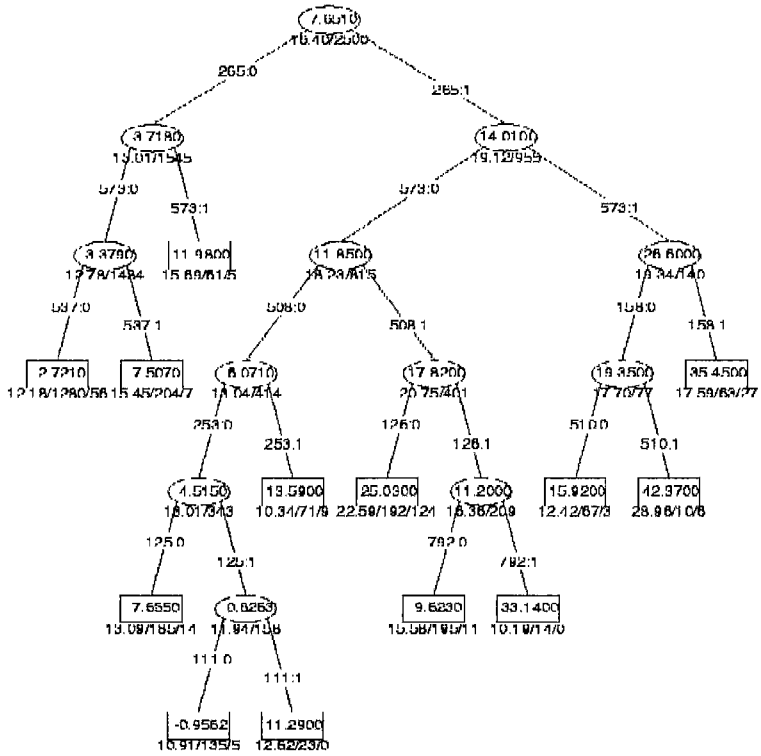

그림 4.1: *SCAM* 회귀 나무

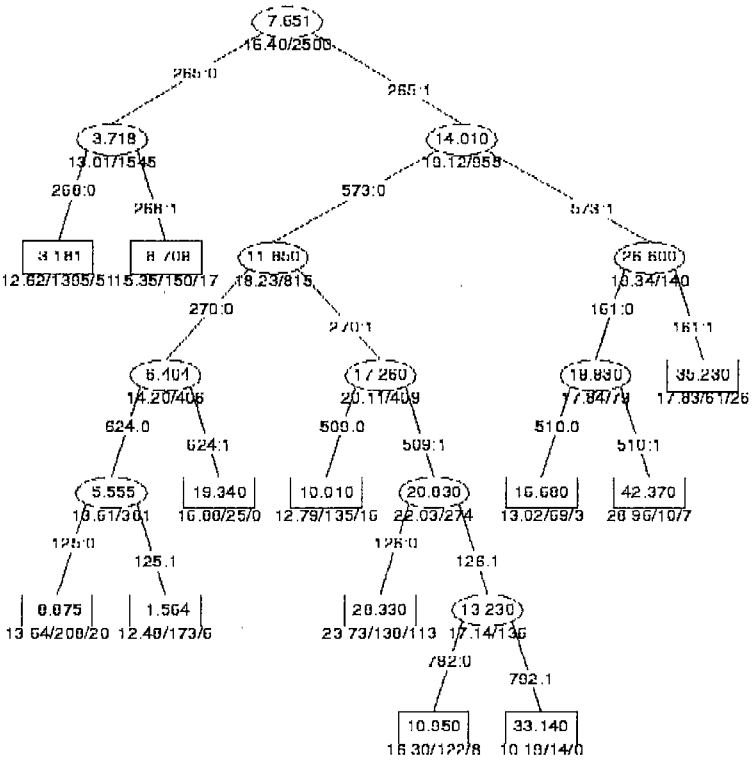

그림 4.2: 각 마디의 최적 분리들 중에서 랜덤하게 선택된 대체 $SCAM$ 회귀 나무

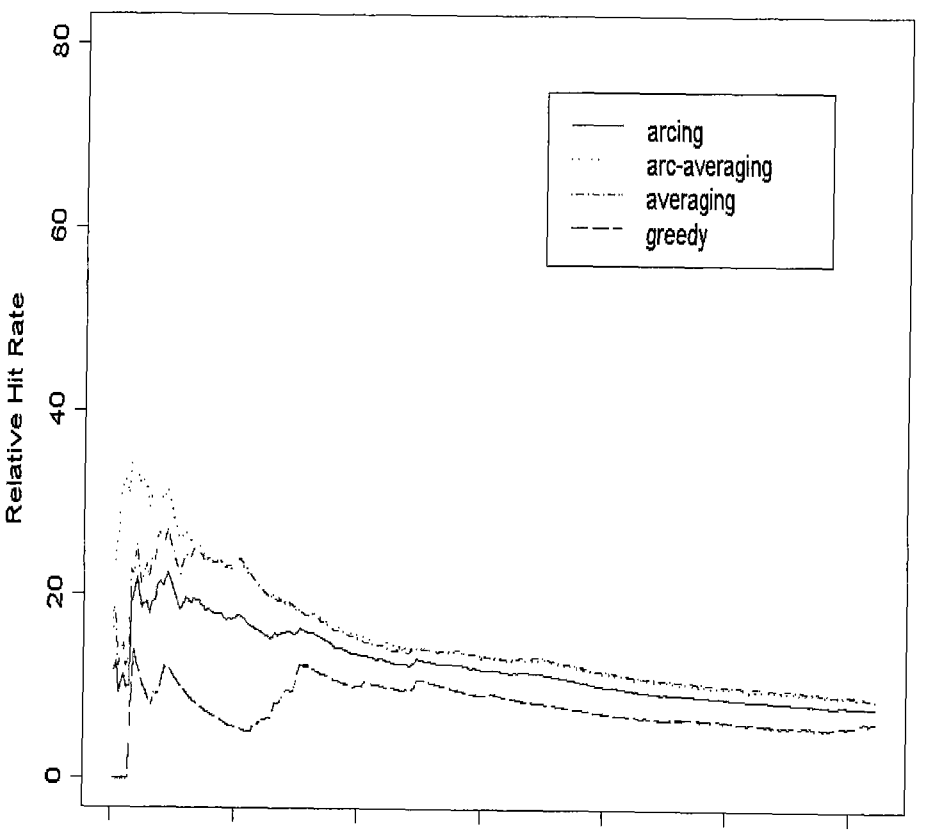

그림 4.3: 4가지 예측 방법에 대한 상대 적중률

A Combined Multiple Regression Trees Predictor for Screening Large Chemical Databases *

Yong B. Lim[1] S. Y. Lee[2] J. H. Chung[3]

ABSTRACT

It has been shown that the multiple trees predictors are more accurate in reducing test set error than a single tree predictor. There are two ways of generating multiple trees. One is to generate modified training sets by resampling the original training set, and then construct trees. It is known that arcing algorithm is efficient. The other is to perturb randomly the working split at each node from a list of best splits, which is expected to generate reasonably good trees for the original training set. We propose a new combined multiple regression trees predictor which uses the latter multiple regression tree predictor as a predictor based on a modified training set at each stage of arcing. The efficiency of those prediction methods are compared by applying to high throughput screening of chemical compounds for biological effects.

Keywords: Combined Multiple Regression Trees Predictor; High Throughput Screening.

* This paper was supported by the Korea Science & Engineering Foundation Grant 981-0105-024-2.

1) Professor, Department of statistics, Ewha Womans University. E-mail: yblim@mm.ewha.ac.kr

2) Graduate Student, Department of statistics, Ewha Womans University.

3) Graduate Student, Department of statistics, Ewha Womans University.