한국 퍼지 및 지능시스템 학회 2003, Vol. 13, No. 4, pp. 386-390

다변량 목표변수를 갖는 의사결정나무의 노드분리에 관한 연구

A Study on the Node Split in Decision Tree with Multivariate Target Variables

김성준
Kim Seong Jun

강릉대학교 산업시스템공학과

요 약

데이터마이닝은 많은 양의 데이터로부터 의사결정에 유용한 패턴을 발견하는 과정으로서 최근 경영 및 공학 분야의 폭넓은 영역에서 많은 관심을 모으고 있다. 어떤 그룹을 여러 하위그룹으로 분류해내는 일은 데이터마이닝의 주요 내용 중 하나이다. 의사결정나무로 알려진 트리기반 기법은 그러한 분류모형을 수립하는 데 효율적인 방안을 제공한다. 트리학습에 있어서 우선적인 관건은 목표변수에 의해 측정되는 노드불순도를 최소화하는 것이다. 하지만 공정관측, 마케팅과학, 임상분석 등과 같은 문제에서는 여러 목표변수를 동시에 고려해야 하는 상황이 쉽게 등장하는 데, 본 논문의 목적은 이처럼 다변량 목표변수를 갖는 데이터셋에서 활용할 수 있는 노드불순도 측정방안을 제시하는 데 있다. 아울러 수치 예를 이용하여 적용결과에 대해 논의한다.

Abstract

Data mining is a process of discovering useful patterns for decision making from an amount of data. It has recently received much attention in a wide range of business and engineering fields. Classifying a group into subgroups is one of the most important subjects in data mining. Tree-based methods, known as decision trees, provide an efficient way to finding the classification model. The primary concern in tree learning is to minimize a node impurity, which is evaluated using a target variable in the data set. However, there are situations where multiple target variables should be taken into account, for example, such as manufacturing process monitoring, marketing science, and clinical and health analysis. The purpose of this article is to present some methods for measuring the node impurity, which are applicable to data sets with multivariate target variables. For illustration, a numerical example is given with discussion.

Key Words : 데이터마이닝, 의사결정나무, 분류, 다변량 목표변수, 노드불순도

1. 서 론

데이터마이닝 (data mining)은 데이터베이스로부터 의사결정에 유용한 패턴을 발견하는 과정으로서 경영, 생의학, 제조 등 다양한 분야에서 활발하게 이용되고 있다 [1]. 데이터마이닝을 수행하는 방법론으로는 크게 기계학습 (machine learning)과 통계분석 (statistical analysis)을 들 수 있다. 기계학습은 상대적으로 대용량의 복잡한 데이터를 다룰 경우 또는 사전지식이 충분치 않을 때 더 효과적으로 활용될 수 있다 [2]. 기계학습의 범주에 속하는 대표적인 것으로는 의사결정나무, 퍼지이론, 신경회로망, 기타 진화연산기법 등을 들 수 있다. 이들은 모두 서로 다른 장단점을 갖고 있지만, 목표변수와 속성변수의 관계를 논리적으로 서술할 수 있다는 점에서 의사결정나무는 분류규칙을 발견하는 데 많이 활용되고 있다. 최근의 한 연구에 따르면 경영 분야의 데이터마이닝 실무 애플리케이션의 절반 정도는 의사결정나무에 기반을 두고 있는 것으로 나타났다 [1].

의사결정나무는 보통 목표변수의 성질에 따라 분류나무 (classification tree) 또는 회귀나무 (regression tree)라 부른다. 이에 관련된 이론적인 내용은 Breiman et al. [3]에 의해 처음으로 종합되었다. 그 이후 지금도 관련 연구가 지속적으로 이루어지고 있지만, 대부분은 목표변수가 하나로 주어지는 상황을 다루고 있다. 하지만 데이터마이닝 실무에서는 다수의 목표변수를 동시에 고려해야 하는 상황을 쉽게 발견할 수 있다. 예컨대 Zhang [4]의 임상보건사례, Ciampi et al. [5]의 영양분석사례, Siciliano and Mola [6]의 재무분석사례 등은 다수의 목표변수를 다루고 있다. 모집단을 더 잘 이해할 수 있을 뿐 아니라 상관구조를 고려할 수 있다는 점에서 다수의 목표변수를 동시에 다루는 것은 충분히 의미가 있다고 판단된다. 그럼에도 불구하고, 노드를 분리하는 방법이나 단일 목표변수를 개별적으로 다룰 때와의 이득비교 등

접수일자 : 2003년 5월 3일
완료일자 : 2003년 6월 5일
감사의 글 : 본 연구는 과학재단 목적기초연구 (과제번호 R05-2001-000-02406-0)의 지원으로 수행되었습니다.

에 대해서 체계적인 연구가 아직 미흡한 실정이다. 다변량 목표변수를 다루는 데 있어 해결해야 할 가장 기초적인 문제는 노드불순도 (node impurity) 의 측정과 그에 따른 노드분리방법에 관련된 것이다. 이에 본 논문에서는 여러 개의 목표변수를 동시에 다룰 수 있는 노드분리방법에 대해 소개하고 수치 예제를 통해 그 적용결과에 대해 논의하고자 한다.

2. 의사결정나무의 노드분리기준

2.1 의사결정나무 개요

어떤 모집단을 속성변수의 값에 따라 계층적으로 분할하는 것을 재귀적 분할 (recursive partitioning)이라고 한다. 재귀적 분할의 과정은 트리 (또는 나무)의 형태로 쉽게 표현할 수 있다. 어떤 목표변수에 대한 동질성 (homogeneity)이 최대화되도록 재귀적 분할작업을 수행하면 동질성이 높은 하부그룹을 발견할 수 있는 데 이 것이 바로 의사결정나무의 목적이다. 의사결정나무에서 뿌리노드 (root node)는 대상 모집단 자체를 의미하는 데 따라서 노드는 속성변수의 해당 조건을 만족하는 모집단의 부분집합을 나타낸다. 각 노드는 적절한 속성변수 및 적절한 기준에 의해 하위노드로 분리된다. 이 때 분리될 하위노드의 수를 둘로 제한하는 경우를 이지분리 (binary split), 제한하지 않으면 다지분리 (multi-way split)라고 부른다. 하위노드를 갖지 않는 노드는 종료노드 (terminal node)라고 부른다. 의사결정나무라는 나무에서 종료노드는 결국 나뭇잎 (leaf)에 해당한다. 노드분리가 완료되어 의사결정나무가 일단 만들어지면 불필요한 가지와 노드를 제거하는 작업이 수행되는 데 이를 가지치기 (pruning)이라고 한다. 어떤 데이터베이스로부터 의사결정나무가 만들어지는 것을 개념적으로 표현하면 다음 그림과 같다.

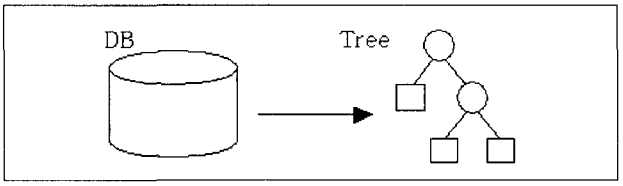

그림 1. DB의 의사결정나무 표현
Fig. 1. Decision tree representation of database

의사결정나무는 여타 기계학습모형에 비해 분류규칙을 발견하는 데 유용한 것으로 알려져 있다. 복잡한 가정이 필요치 않아 적용하기 쉽다는 점과 무엇보다도 결과에 대한 논리적인 해석이 가능하다는 점이 주된 이유라고 하겠다 [7]. 하지만 유용한 의사결정나무를 만들어내기 위해서는 변수선택 및 노드분리 기준, 노드분리 중지규칙, 가지치기 기준 등이 목표변수의 특성에 따라 적절하게 결정되어야 한다. 이 중 본 논문에서는 범주형 목표변수를 위한 노드분리기준에 대해 다루고 그 범위는 이지분리에 국한하기로 한다. 먼저 그 대표적인 내용을 설명하면 다음과 같다.

2.2 노드분리기준

의사결정나무는 트리 전체의 동질성이 최대화되는 방향으로 성장해야 한다. 이 것은 개별 노드의 동질성을 크게 할 수 있는 (또는 불순도를 작게 할 수 있는) 분리기준을 채택함으로써 달성할 수 있다. 다음 그림과 같이 어떤 부모노드 (parent node) t가 두 개의 자식노드 (children nodes) t_L과 t_R로 분리되는 상황을 생각하자. 각각의 노드에 소속된 개체 즉 인스턴스 (instance)의 수는 각각 $n(t)$, $n(t_L)$, $n(t_R)$로 나타내기로 한다. 그러면 $n(t)=n(t_L)+n(t_R)$이 된다. 또 각 노드에서 목표변수 y의 j번째 범주에 속하는 인스턴스의 수를 $n_j(t)$, $n_j(t_L)$, $n_j(t_R)$로 표기한다. 역시 $n_j(t)=n_j(t_L)+n_j(t_R)$가 성립한다. 단 $j=1,2,\cdots,K$.

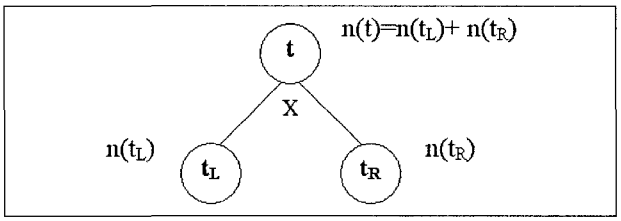

그림 2. 부모노드와 자식노드
Fig. 2. Parent and children nodes

노드 t에서, 범주형 목표변수의 경우 널리 쓰이는 동질성의 척도인 엔트로피 (entropy)와 지니지수 (Gini index)는 다음과 같이 각각 표현된다.

$$h(t)=-\sum_{j=1}^{K}\frac{n_j(t)}{n(t)}\log\frac{n_j(t)}{n(t)} \quad (1)$$

$$h(t)=1-\sum_{j=1}^{K}\frac{n_j^2(t)}{n^2(t)} \quad (2)$$

노드 t의 동질성이 클수록 (또는 불순도가 낮을수록) $h(t)$는 감소하며 0에 가까워진다. 따라서 다음과 같이 정의되는 노드분리이득이 가장 커지는 속성변수와 그 조합을 택함으로써 노드를 분리한다.

$$\eta(t)=h(t)-p(t_L)h(t_L)-p(t_R)h(t_R)$$

여기서 $p(t_L)=n(t_L)/n(t)$이고, $p(t_R)=n(t_R)/n(t)$이다. 이 과정을 모든 노드에 대해 수행하면 결국 트리는 동질성이 최대화하는 방향으로 성장하게 된다. 엔트로피나 지니지수와 동일한 목적으로 카이제곱 통계량 (chi squared statistic)을 사용할 수도 있는 데 잘 알려진 CHAID (chi-squared automated interaction detection) 알고리즘이 그러한 경우이다 [7]. 카이제곱 통계량은 그 자체가 노드분리이득을 뜻하며 다음과 같이 정의된다.

$$\eta(t)=\sum_{j=1}^{K}[n(t)n_j(t_L)-n_j(t)n(t_L)]^2/[n_j(t)n(t_L)]$$
$$+\sum_{j=1}^{K}[n(t)n_j(t_R)-n_j(t)n(t_R)]^2/[n_j(t)n(t_R)]$$

이지분리에서 이 통계량의 자유도는 $(K-1)$이다. 실제로는 이 값에 해당되는 유의확률 (p-value)에 따라 노드분리여부를 결정하게 되므로 결국 통계적인 카이제곱 검정을 수행하는 것과 다를 바가 없다.

노드분리를 위해 이처럼 동질성을 고려하는 대신 비용적인 손실의 개념이 활용될 수 있음이 Kim and Lee [8]에 의해 논의되었다. 노드 t에서 오분류에 의한 기대손실 (Expected Loss)을 가장 단순한 형태로 정의하면 다음과 같다.

$$L(t)=1-\max_{1\le j\le K} p_j(t) \tag{3}$$

식 (3)에서, $p_j(t)$는 노드 t에서 $Y=j$인 인스턴스가 차지하는 비율을 의미한다. 단 $j=1,2,\cdots,K$. 따라서 트리 전체의 손실은 다음 식으로 평가할 수 있다.

$$L(T)=\sum_{t\in T} p(t)L(t)$$

여기서 T는 종료노드의 집합이고 $p(t)=n(t)/N$이다. 단 N은 전체 인스턴스의 수를 뜻한다.

3. 다변량 목표변수를 위한 노드분리기준

M개의 목표변수를 $Y_1, Y_2, \cdots, Y_M$이라 하고 목표변수 Y_i가 K_i개의 범주를 갖는다고 할 때 앞에서 설명한 식 (1)의 엔트로피와 식 (2)의 지니지수는 다음과 같이 일반화될 수 있다 [4, 8].

$$h(t)=-\sum_{j_1=1}^{K_1}\sum_{j_2=1}^{K_2}\Lambda\sum_{j_M=1}^{K_M}\frac{n_{j_1,j_2,\Lambda,j_M}(t)}{n(t)}\log\frac{n_{j_1,j_2,\Lambda,j_M}(t)}{n(t)} \tag{4}$$

$$h(t)=1-\sum_{j_1=1}^{K_1}\sum_{j_2=1}^{K_2}\Lambda\sum_{j_M=1}^{K_M}\frac{n^2_{j_1,j_2,\Lambda,j_M}(t)}{n^2(t)} \tag{5}$$

단, 노드 t에서 $Y_1=j_1, Y_2=j_2, \cdots, Y_M=j_M$에 해당되는 인스턴스의 수를 $n_{j_1,j_2,\cdots,j_M}(t)$로 나타내었다. 식 (3)의 기대손실의 일반화된 표현은 역시 다음과 같다 [8].

$$L(t)=1-\max_{\substack{1\le j_g\le K_g\\ \forall g=1,2,\Lambda,M}} p_{j_1,j_2,\Lambda,j_M}(t) \tag{6}$$

한편 여러 개의 목표변수를 다룰 때 우리는 공분산행렬(covariance matrix)을 얻을 수 있다. 물론 이 것은 범주에 순서가 있는 순서형 데이터인 경우에 의미가 있다. 공분산행렬은 개별 분산 자체 뿐 아니라 상관구조 (correlation structure)에 대한 정보도 함께 제공해 준다. 예를 들어, Zhang [4]은 공분산행렬의 행렬식 (determinant)이 지니지수로 해석될 수 있음을 지적하였으며, 공분산행렬을 이용하여 다음과 같이 정의되는 Hotelling의 T^2를 노드분리기준으로 제안한 바 있다.

$$h(t)=\frac{1}{n(t)}\sum_{i\in I(t)}[\mathbf{y}_i-\bar{\mathbf{y}}(t)]'V^{-1}[\mathbf{y}_i-\bar{\mathbf{y}}(t)] \tag{7}$$

식 (7)에서, V는 뿌리노드에서 얻어진 공분산행렬이고 $I(t)$는 노드 t에 속하는 인스턴스의 집합을 의미한다. 또 $\bar{y}(t)$는 노드 t에서 구한 M개 목표변수의 평균을 나타내는 벡터이다. 원래 Hotelling의 T^2는 통계적 공정관리 분야에서 다변량 품질특성을 모니터링하기 위한 목적으로 이용되어 왔으나, 여기서는 M차원 평균으로부터의 거리를 측정하고 이를 최소화하는 속성변수를 찾기 위해 도입되었다. 역시 다변량 노드분리기준으로서 평균벡터로부터의 거리를 사용하는 방안이 Ciampi et al. [5]에 의해서도 제안되었는 데, 실제로 이들의 알고리즘은 식 (7)과 동일하게 정의되는 Mahalanobis Distance에 근거하고 있다.

지금까지 설명한 노드분리기준은 여러 개의 목표변수를 한꺼번에 고려하고 있다는 점에서 다변량적이라고 말할 수 있다. 이와는 대조적으로, 개별 목표변수마다 노드분리기준을 따로 계산한 후 이들을 가중치로 합산하여 노드분리의 기준으로 삼는 방식을 생각할 수 있다. 목표변수 Y_i에 대해서 평가된 노드 t의 지니지수를 $h(i,t)$라 하고 Y_i에 대한 가중치를 $w(i)$라고 하면 노드 t에서의 종합적인 노드분리기준으로서

$$H(t)=\sum_{i=1}^{M} w(i)h(i,t) \tag{8}$$

를 사용하는 것이다. 단 가중치는

$$\sum_{i=1}^{M} w(i)=1 \text{ and } w(i)\ge 0 \quad i=1,2,\cdots,M$$

의 조건을 만족하는 것으로 한다. 예를 들어 Siciliano and Mola [6]는 각 목표변수에 대해 지니지수를 먼저 구하고 이를 노드마다 재계산된 가중치로 합산하여 노드분리를 결정하는 방법을 사용하였다. 실제로 그들이 노드 t에서 목표변수 Y_g에 적용한 가중치는

$$w(g,t)=h(g,t)/\sum_{i=1}^{M} h(i,t)$$

와 같이 정의되는 데, 이 것은 식 (8)에서 가중치로 $w(i)=1/M$를 대입한 노드분리기준을 사용한 것과 같다는 것을 쉽게 보일 수 있다. 따라서 그들의 가중치는 문제나 목표변수의 특성을 적절하게 수용하기에는 어려울 것으로 판단된다. 또 한편으로 가중치는 보통 분석외적인 요소 예를 들면 손실, 중요도, 선호도 등에 의해서 결정될 때가 많으므로 이 경우 Siciliano and Mola [6]의 방법을 어떻게 적용하는 것이 과연 바람직한가는 재검토할 필요가 있을 것이다.

이 밖에도 식 (8)과 같은 가중합의 형태는 식 (3)에서 기술한 기대손실에 의해서도 나타낼 수 있다 [8]. 즉 목표변수 Y_i에 대해서 평가된 노드 t의 기대손실을 $l(i,t)$라 하면 노드 t에서의 가중 기대손실은

$$L(t)=\sum_{i=1}^{M} w(i)l(i,t) \tag{9}$$

로서 평가할 수 있고 이를 식 (8) 대신 활용할 수 있다는 것이다. 목표변수를 개별적으로 다루면서 식 (8)과 (9)의 형태로 대표되는 가중합에 의해서 노드를 분리하는 방법은 상대적으로 사용하기 쉽고 변수간 중요도를 수용할 수 있다는 장점이 있지만 반대로 상관구조를 반영하기 못한다는 점과 가중치를 부여하는 방법에 따라 분석결과가 민감하다는 문제점을 안고 있다.

이제 간단한 수치예제를 통해 개별 목표변수에 대한 의사결정나무와 다중 목표변수에 의한 의사결정나무를 살펴하고자 한다. 예제로 사용된 데이터는 UCI Repository [9]에 수록되어 있는 간염 데이터의 일부이다. 단지 차이점을 보여주기 위한 것이 예제의 목적이었으므로, 편의상 결측치는 제외하고 이진변수 10가지를 대상으로 하였으며 목표변수는 이들 중 두 가지로 채택하였다. 그림 3과 4는 각각 Y_1과 Y_2에 대한 의사결정나무 결과를 보여주고 있다. 여기서는 노드분리기준으로 식 (2)의 지니지수를 이용하였다.

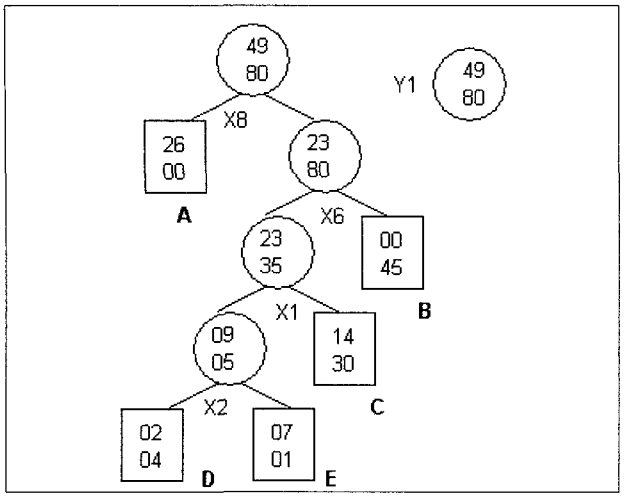

그림 3. Y1에 대한 의사결정나무
Fig. 3. Decision Tree on Y1

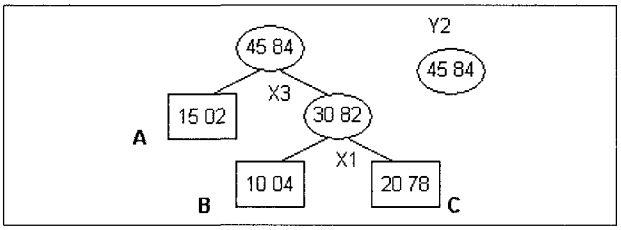

그림 4. Y2에 대한 의사결정나무
Fig. 4. Decision Tree on Y2

그림에서 보는 바와 같이 각 노드 안에는 목표변수에 따른 빈도를 나타내었다. 예를 들어 그림 3의 뿌리노드에는 $Y_1=1$인 인스턴스가 49개, $Y_1=2$인 인스턴스가 80개 포함되어 있다는 의미이며, 그림 4의 뿌리노드에는 $Y_2=1$인 인스턴스가 45개, $Y_2=2$인 인스턴스가 84개 있다는 뜻이다. 먼저 그림 3을 보면 5개의 종료노드 중 노드 A와 E는 $Y_1=1$로 노드 B, C, D는 $Y_1=2$로 분류됨을 알 수 있다. 이 때 종료노드 A, B, C, D, E에서의 오분류 개수는 각각 0, 0, 14, 2, 1이므로 트리 전체의 오분류율은 17÷129=13.2%이다. 반면 그림 4에는 3개의 종료노드가 나타나 있는 데 노드 A와 B는 $Y_2=1$로 노드 C는 $Y_2=2$로 분류할 수 있다. 역시 오분류 개수가 노드 A, B, C에 대해서 각각 2, 4, 20이므로 26÷129=20.2%의 오분류율을 갖게 된다.

한편 다변량 목표변수를 다루기 위해 앞에서 소개한 방법 중 식 (5)의 다변량 지니지수 (multivariate Gini index)를 노드분리기준으로 이용하여 의사결정나무를 작성하면 그림 5의 결과를 얻을 수 있다.

마찬가지로 Y_1과 Y_2에 대한 빈도표는 노드 안에 표시되어 있다. 예를 들어 뿌리노드에는 $(Y_1, Y_2)=(1,1)$인 개체가 28개, $(Y_1, Y_2)=(1,2)$인 개체가 21개, $(Y_1, Y_2)=(2,1)$인 개체가 17개, $(Y_1, Y_2)=(2,2)$인 개체가 각각 63개 포함되어 있다. 그림에서부터 보는 바와 같이 종료노드는 총 6개가 있는 데 그 중 노드 A와 F는 $(Y_1, Y_2)=(2,2)$로, 노드 B는 $(Y_1, Y_2)=(1,1)$로 분류하는 것이 타당함을 알 수 있다. 나머지 경우는 빈도수가 낮아 큰 의미는 없으나 노드 C와 E는 $(Y_1, Y_2)=(2,1)$로, 노드 D는 $(Y_1, Y_2)=(1,1)$로 분류할 수 있을 것이다. 종료노드 A, B, C, D, E, F에 대한 오분류 개수는 각각 5, 9, 3, 4, 3, 17이므로 이 경우 총 오분류율은 41÷129로서 31.8%가 된다. 물론 이 수치는 그림 3과 4에서 구한 개별 목표변수에 대한 오분류율 13.2%와 20.2%에 비해서는 크지만, 그 합인 33.4%보다는 다소 낮다는 것을 알 수 있다.

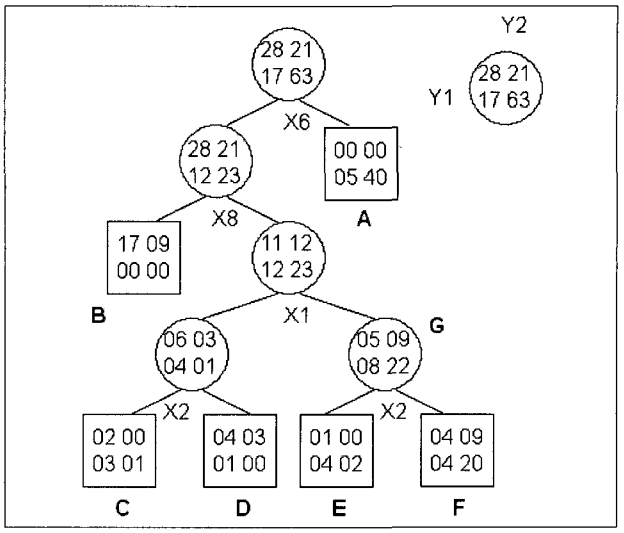

그림 5. 다변량 지니지수에 의한 의사결정나무
Fig. 5. Decision Tree by Multivariate Gini Index

그림 5의 결과는 식 (7)의 Hotelling의 T^2를 이용하였을 때에도 동일하게 얻어졌으며, 식 (8)에 Siciliano and Mola [6]의 가중치 $(w(1), w(2))=(0.5, 0.5)$를 적용하였을 때에도 같은 결과를 얻었다. 따라서 그림 5에 나타난 의사결정나무는 어느 정도 대표성이 있는 결과라고 간주할 수 있을 것이다. 특히 식 (8)의 가중 지니지수에 가중치만 $(w(1), w(2))=(0.25, 0.75)$로 바꿔주었을 때에도 의사결정나무 결과는 그림 6에서 보는 바와 같이 다변량 지니지수를 이용하였을 때와 다소 차이는 있으나 대체로 유사한 형태로 나타났다. 그림 6에 있는 종료노드 C는 그림 5의 종료노드 C와 E를 합쳐 놓은 것에 대응된다. 한편 본 논문에는 수록하지 않았지만 식 (6)의 다변량 기대손실이나 식 (9)의 가중 기대손실을 이용했을 때에는 그림 5나 6과는 상이한 결과가 얻어졌다.

4. 토의

다변량 지니지수에 의해서 얻어진 그림 5의 트리구조를 살펴보면 단순 지니지수로 구한 그림 3의 Y_1에 대한 트리구조와 상당부분 일치하고 있음을 알 수 있다.

이는 Y_2보다는 Y_1이 의사결정나무 결과에 더 큰 영향을 미치고 있음을 시사하는 것이지만 그렇다고 하더라도 Y_2를 함께 고려하는 것이 결과와는 관계가 없다고 말하기는 곤란하다. 예를 들어, Y_1만 고려해서 작성된 그림 3에서는 노드 C가 더 이상 분리되지 않았으나 이 노드에 해당되는 그림 5의 노드 G는 노드 E와 F로 분리되고 있다. 바로 이 부분이 목표변수 Y_2를 같이 고려함으로써 발생된 효과라고 말할 수 있다. 그림 3과 5의 또 다른 차이점이라고 하면, 그림 3에서

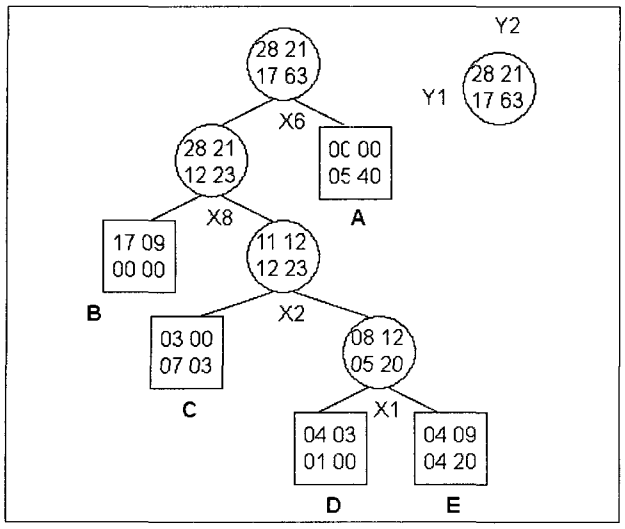

그림 6. 가중 지니지수에 의한 의사결정나무
Fig. 6. Decision Tree by Weighted Gini Index

는 첫 번째 분리변수가 X_8인 반면 그림 5에서는 X_6이라는 것인 데, 두 번째 분리변수를 선택하는 과정에서 이 차이는 상쇄되고 있다.

지금까지 설명한 바와 같이 본 논문에서는 다변량 목표변수를 다룰 수 있는 노드분리방법에 대해 연구하였다. 단일 목표변수를 다룰 때 이용되었던 노드분리기준을 일반화하는 방법과 가중합으로 다루는 방법이 있을 수 있는 데, 전자는 다변량 목표변수의 상관구조를 반영할 수 있다는 장점이 있지만 노드불순도를 어떻게 정의하는 것이 바람직한가가 주요 관건이 되고 후자의 경우에는 변수간 가중치를 수용할 수 있는 장점이 있지만 부여된 가중치에 따라서 분석결과가 민감할 수 있다는 문제점이 있다. 본 논문에서는 간단한 수치예제를 통해 다변량 목표변수를 동시에 다루었을 때와 개별적으로 다루었을 때의 차이점을 살펴보았으며 그 결과, 다변량 목표변수에 의한 의사결정나무는 오분류를 줄일 수 있고 분석대상을 더 잘 이해할 수 있는 가능성을 보여주었다.

비록 의사결정나무 구축결과를 비교하기 위해 3절에서처럼 오분류율 또는 오분류건수를 이용하는 것이 제한적이라 할지라도, 다변량 목표변수를 이용할 때의 잠재성을 설명하는 데에는 충분하였다고 판단된다. 사실 의사결정나무 결과를 제대로 비교하기 위해서는 학습데이터셋 (learning dataset)이 아니라 검증데이터셋 (test dataset)를 이용하여야 한다. 이 부분은 현재 연구가 진행 중에 있어 조만간 다변량 의사결정나무의 타당성을 보다 체계적으로 논의할 수 있을 것으로 기대하고 있다. 아울러 상관구조를 반영하는 방법이나 가중치를 부여하는 알고리즘 등도 함께 비교하는 것도 추진하고 있다.

앞에서 설명한 Hotelling의 T^2는 연속형 목표변수의 경우에도 적용할 수 있지만 본 논문의 내용은 순서가 있는 범주형 (ordered categorical) 목표변수를 주 대상으로 하고 있다. 하지만 데이터마이닝 실무에서는 명목형, 순서형, 척도형 등과 같이 다양한 형태의 목표변수가 혼재해 있는 경우가 많으므로, 이러한 상황을 다룰 수 있는 다변량 노드불순도의 개발도 유익한 연구주제이다.

참 고 문 헌

[1] Indranil Bose and Radha K. Mahapatra, "Business Data Mining A Machine Learning Perspective," Information & Management, Vol. 39, pp. 211-225, 2001.

[2] Katharina D. C. Stark and Dirk U. Pfeiffer, "The Application of Non-parametric Techniques to Solve Classification Problems in Complex Data Sets in Veterinary Epidemiology An Example," Intelligent Data Analysis, Vol. 3, pp. 23-35, 1999.

[3] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone, Classification and Regression Trees, Boca Raton, FL: Chapman & Hall/CRC, 1984.

[4] Heping Zhang, "Classification Trees with Multiple Binary Responses," Journal of the American Statistical Association, Vol. 93, No. 441, pp. 180-193, 1998.

[5] Antonio Ciampi, Djamel A. Zighed, and Jeremy Clech, "Trees and Induction Graphs for Multivariate Response," Lecture Notes in Computer Science, No. 1910, pp. 359-366, 2000.

[6] Roberta Siciliano and Francesco Mola, "Multivariate Data Analysis and Modeling Through Classification and Regression Trees," Computational Statistics & Data Analysis, Vol. 32, pp. 285-301, 2000.

[7] 장남식 외 2인, 데이터마이닝, 대청, 2000.

[8] Seong-Jun Kim and Kang B. Lee, "Constructing Decision Trees with Multiple Response Variables," International Journal of Management and Decision Making, Vol. 6, 2003, to appear.

[9] UCI Repository of Machine Learning Databases, 1998.

저 자 소 개

김성준(Kim Seong-Jun)

1989년 : 연세대학교 응용통계학과 학사 (서울).

1991년 : 한국과학기술원 산업공학과 석사 (서울).

1995년 : 한국과학기술원 산업공학과 박사 (대전).

1995년~현재 : 강릉대학교 산업공학과 부교수.

관심분야 : 지능정보, 다변량분석, 품질관리.

Phone : 033-640-2375

Fax : 033-640-2244

E-mail : sjkim@kangnung.ac.kr