Proceedings of KFIS Fall Conference, 2003. 5. 3

의사결정나무에서 다중 목표변수를 고려한

Splitting Decision Tree Nodes with Multiple Target Variables

김성준
강릉대학교 산업시스템공학과

Seong-Jun Kim
Department of Industrial and Systems Engineering, Kangnung National University
E-mail : sjkim@kangnung.ac.kr

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 classification models. 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 several methods for measuring the node impurity, which are applicable to data sets with multiple target variables. For illustrations, numerical examples are given with discussion.

Key words : Decision Tree, Classification, Node Impurity Measures, Multiple Target Variables

I. 서 론

데이터마이닝 (Data Mining)은 데이터베이스로부터 의사결정에 유용한 패턴을 발견하는 과정으로서[6] 경영, 의학, 제조 등 다양한 분야에서 활발하게 이용되고 있다. 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[3]. 기계학습의 범주에 속하는 대표적인 기법으로는 의사결정나무, 퍼지이론, 신경회로망 등 각종 진화연산모형을 들 수 있다. 이들은 모두 서로 다른 장단점을 갖고 있지만, 목표변수와 속성변수의 관계를 논리적으로 서술할 수 있다는 점에서 의사결정나무는 분류규칙을 발견하는 데 많이 활용된다. 최근의 한 연구에 따르면 경영 분야의 데이터마이닝 애플리케이션의 절반 정도가 의사결정나무에 기반을 두고 있다[6].

의사결정나무는 목표변수의 성질에 따라 보통 분류나무 (classification Tree) 또는 회귀나무 (Regression Tree)라 부른다. 이에 관련된 체계적인 내용은 Breiman et al.[1]에 의해 처음으로 종합되었다. 그 이후 지금도 관련 연구가 지속적으로 이루어지고 있지만, 대부분은 목표변수가 하나로 주어지는 상황을 다루고 있다. 데이터마이닝 실무에서는 다수의 목표변수를 고려해야 하는 상황을 쉽게 찾을 수 있다. 그와 같은 구체적인 예는 Zhang[2], Ciampi et al.[4], Siciliano and Mola[5] 등에서 소개

하고 있다. 모집단을 더 잘 이해할 수 있을 뿐 아니라 상관관계를 파악할 수 있다는 점에서 다수의 목표변수를 동시에 다루는 것은 충분한 의미를 갖는다고 판단된다. 그럼에도 불구하고, 노드를 분리하는 방법이나 단일 목표변수를 개별적으로 다룰 때와의 성능비교 등에 대해서는 체계적인 연구가 아직 미흡한 실정이다. 본 논문에서는 여러 개의 목표변수를 동시에 다룰 수 있는 노드분리방법에 대해 소개하고 수치예제를 통해 그 적용결과를 논의하고자 한다.

II. 의사결정나무와 노드분리기준

2.1 의사결정나무의 개요

어떤 모집단을 속성변수의 값에 따라 계층적으로 분할하는 것을 Recursive Partitioning (RP)이라고 한다. RP의 결과는 트리 (또는 나무)의 형태로 쉽게 표현할 수 있다. 어떤 목표변수에 대한 동질성 (Homogeneity)이 최대화되도록 RP 작업을 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 나뭇잎에 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.

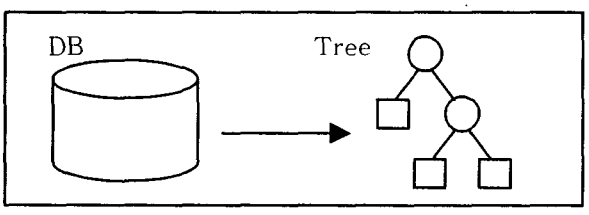

그림 1. DB의 의사결정나무 표현

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. 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 다룬다. 먼저 그 대표적인 내용을 설명하면 다음과 같다.

2.2 노드분리기준

의사결정나무는 트리 전체의 동질성이 최대화되는 방향으로 성장해야 한다. 이 것은 개별노드의 동질성을 크게 할 수 있는 (또는 불순도를 작게 할 수 있는) 분리기준을 채택함으로써 달성할 수 있다. 어떤 노드 t가 두 개의 하위노드 t_L과 t_R로 분리되는 상황을 생각하자. 노드에 속하는 데이터 즉 인스턴스의 수는 각각 $n(t)$, $n(t_L)$, $n(t_R)$로 나타낸다. 각 노드에서 목표변수 y의 j번째 범주에 속하는 인스턴스의 수는 $n_j(t)$, $n_j(t_L)$, $n_j(t_R)$로 각각 나타낸다. 단 j=1, 2, …, K. 노드 t에서, 범주형 목표변수의 경우 널리 쓰이는 동질성의 척도인 엔트로피 (Entropy)와 지니지수 (Gini Index)는 다음과 같이 각각 표현된다.

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

따라서 다음과 같이 정의되는 노드분리이득 (Node Splitting Gain)이 가장 커지는 속성변수와 그 조합을 택함으로써 노드가 분리된다.

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

같은 목적으로 카이제곱 통계량 (Chi-Square Statistic)을 사용하는 경우도 있는 데 CHAID (Chi-Squared Automated Interaction Detection) 알고리즘이 바로 그러하다. 카이제곱 통계량은 그 자체가 노드분리이득을 의미하며 다음과 같이 정의된다.

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

이지분리로 수행할 경우 이 통계량의 자유도는 (K-1)이다. 실제로는 이 값에 해당되는 유의확률에 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[7]에 의해 논의되었다. 노드 t에서 오분류에 의한 기대손실 (Expected Loss)을 가장 간단하게 정의하면 다음과 같다.

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

트리 전체의 손실은 다음과 같이 구할 수 있다.

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

단 T는 종료노드의 집합이고 p(t)=n(t)/N이다. 여기서 N은 인스턴스의 수이다.

2.3 다중 목표변수를 위한 노드분리기준

M개의 목표변수를 $Y_1, Y_2, \cdots, Y_M$이라 하고 목표변수 Y_i는 K_i개의 범주를 갖는다고 할 때 2.2절의 엔트로피와 지니지수는 다음과 같이 일반화된다[2, 7].

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

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

마찬가지로 기대손실 역시 다음과 같다[7].

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

한편 여러 개의 목표변수를 다룰 때 우리는 공분산행렬 (Covariance Matrix)을 추정할 수 있다. 물론 이 것은 범주에 순서가 있을 경우에 의미가 있다. 공분산행렬은 개별 분산 자체뿐 아니라 상관구조 (Correlation Structure)에 대한 정보도 함께 제공해 준다. Zhang[2]은 공분산행렬의 행렬식이 지니지수로 해석될 수 있음을 지적하였으며, Hotelling의 T^2 통계량도 노드분리기준으로 사용될 수 있음을 기술하였고 Ciampi et al.[4]도 Mahalanobis의 거리척도의 채택을 제안하였다.

지금까지 설명된 방법은 노드분리에 여러 개의 목표변수를 동시에 고려하고 있는 것이 특징이다. 이와는 대조적으로, 개별 목표변수마다 노드분리기준을 구한 후 이들을 가중치로 합산하는 방식을 생각할 수 있다. 예를 들어 Siciliano and Mola[5]는 각 목표변수에 대해 지니지수를 구하고 이를 노드마다 재계산된 가중치로 합산하여 노드분리를 결정하는 방법을 사용하였다. 이 방법은 상대적으로 사용하기 쉽다는 장점이 있지만 상관구조를 반영하기 못한다는 점과 가중치를 어떻게 부여하는가에 분석결과가 민감하다는 문제점을 안고 있다.

2.3. 수치예제

간단한 수치예제를 통해 개별 목표변수에 대한 의사결정나무와 다중 목표변수에 의한 의사결정나무를 비교한다. 예제로 사용된 데이터는 UCI Repository[8]에 수록되어 있는 간염 데이터의 일부로, 이진변수 10가지를 대상으로 하였으며 목표변수는 이들 중 적절하게 두 가지로 잡았다. 그림 2와 3은 각각 Y_1과 Y_2에 대한 의사결정나무 결과를 보여주고 있다. 노드분리기준으로는 지니지수를 이용하였다.

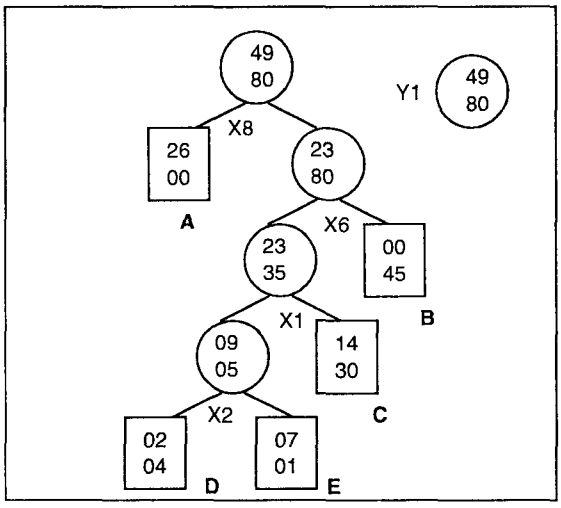

그림 2. Y_1에 의한 의사결정나무

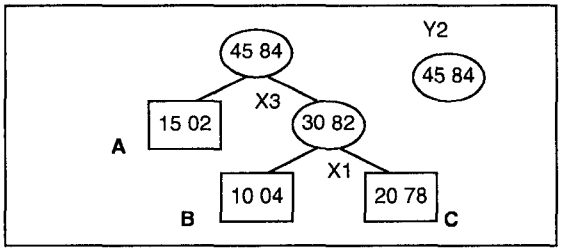

그림 3. Y_2에 의한 의사결정나무

노드 안에는 각 목표변수에 대한 빈도를 나타내었다. 먼저 그림 2를 보면 6개의 종료노드 중 노드 A와 E는 Class 1로 노드 B, C, D는 Class 2로 분류할 수 있음을 알 수 있다. 이 경우 오분류율은 17/129로서 13.2%이다. 반면 그림 3에는 3개의 종료노드가 나타나 있는데 노드 A와 B는 Class 1로 노드 C는 Class 2로 분류하면 26/129=20.2%의 오분류율을 갖게 된다. 이제 2.3절에서 소개한 노드분리기준 중 다변량 지니지수를 이용하여 Y_1과 Y_2를 동시에 고려한 의사결정나무를 만들어 보자. 결과는 다음 그림 4에 나타난 바와 같다.

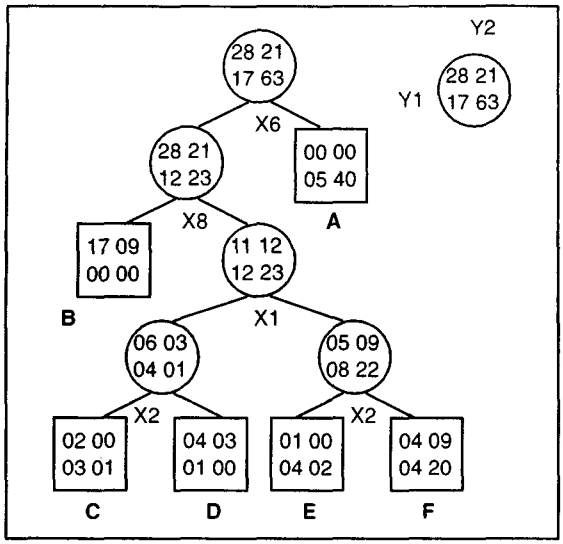

그림 4. Y_1과 Y_2에 의한 의사결정나무

마찬가지로 Y_1과 Y_2에 대한 빈도표는 노드 안에 표시되어 있다. 그림에서부터 노드 A와 F는 Class (2, 2)로, 노드 B는 (1, 1)로 분류하는 것이 타당함을 알 수 있다. 나머지 경우는 빈도수가 낮아 큰 의미는 없으나 노드 C와 E는 Class (2, 1)로, 노드 D는 Class (1, 1)로 분류할 수 있다. 이 경우 오분류율은 41/129로서 31.8%가 된다. 이 수치는 개별 목표변수에 대한 트리의 오분류율 합인 33.4%보다 다소 낮다는 것을 알 수 있는 데, 이 차이는 Y_1과 Y_2의 상관관계가 강할수록 더욱 두드러지게 된다. 본 논문에는 수록하지 않았으나 추가로 실시된 수치실험에서는 다중 목표변수에 의한 오분류율이 개별 트리의 오분류율 합보다 두 배 가까이 낮게 나타났다. 그러므로 목표변수 간에 상관관계가 존재할 경우 본 논문의 내용은 오분류율을 개선하는 데 도움이 될 것으로 기대된다.

III. 결 론

다중 목표변수가 등장할 때의 의사결정나무에 관련해서 크게 두 가지의 연구방향을 생각할 수 있다. 하나는 목표변수를 별개로 다룰 때에 비해 어떤 차이가 있는가이며, 다른 하나는 노드분리기준에 관한 것이다. 본 논문에서는 먼저 다중 목표변수를 노드분리기준에 대해 소개하였으며, 수치예제를 통해 그 적용결과를 예시하였다. 또한 개별 목표변수에 의한 결과와 비교하고 그 차이점을 설명하였다. 분석결과, 다중 목표변수 의사결정나무는 오분류율을 개선할 수 있는 잠재력을 보여주었다. 또한 상관관계가 높은 그룹을 식별하는 것은 개별 의사결정나무로는 얻을 수 없는 정보이다.

본 논문에서는 다중 목표변수를 다룰 수 있는 노드분리기준을 크게 두 가지로 대별하여 소개하였다. 어느 방식이 바람직한가에 대해서는 보다 체계적인 연구가 필요할 전망이며, 특히 상관구조를 반영하는 방법 및 가중치를 부여는 알고리즘 등이 함께 비교되어야 할 것으로 판단된다. 데이터마이닝 실무에서는 목표변수가 명목형, 순서형, 척도형 등으로 혼재할 가능성이 높은 데 이러한 상황을 다룰 수 있는 방안도 제시되어야 할 것이다.

감사의 글 : 본 연구는 과학재단 목적기초연구의 지원으로 수행되었습니다. (과제번호: R05-2001-000-02406-0)

IV. 참고문헌

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

[2] Heping Zhang, "Classification Trees with Multiple Binary Responses", *Journal of the American Statistical Association*, Volume: 93, Number: 441, 1998, Page(s): 180-193

[3] 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*, Volume: 3, 1999, Page(s): 23-35

[4] Antonio Ciampi, Djamel A. Zighed, and Jeremy Clech, "Trees and Induction Graphs for Multivariate Response", *Lecture Notes in Computer Science*, Number: 1910, 2000, Page(s): 359-366

[5] Roberta Siciliano and Francesco Mola, "Multivariate Data Analysis and Modeling Through Classification and Regression Trees", *Computational Statistics & Data Analysis*, Volume: 32, 2000, Page(s): 285-301

[6] Indranil Bose and Radha K. Mahapatra, "Business Data Mining - A Machine Learning Perspective", *Information & Management*, Volume: 39, 2001, Page(s): 211-225

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

[8] UCI Repository of Machine Learning Databases, 1998