연속형 자료에 대한 나무형 군집화

허명회[1], 양경숙[2]

요 약

본 연구는 반복분할(recursive partitioning)에 의한 군집화 방법을 제안하고 활용 예를 제시한다. 이 방법은 나무 형태의 해석하기 쉬운 단순한 규칙을 제공하면서 동시에 변수선택기능을 제공한다.

주요용어 : 나무형 군집화(tree-structured clustering), 노드 분리, Overall R-Square, K-평균 군집화, 변수선택.

1. 연구배경과 목적

Kass(1980)의 CHAID, Quinlan(1993)의 C4.5, Breiman et al.(1984)의 CART 등은 표적변수(target variable)가 있는 다변량 훈련자료로부터 해석이 쉬운 나무 형태의 분류규칙을 만드는 동시에 주요 변수를 선별해낸다. 때문에 지난 20여년에 걸쳐 많은 연구자들의 관심을 받았고 현업 전문가들로부터도 긍정적인 평가를 받았다.

표적변수가 없는 훈련자료에 대하여도 해석이 쉬운 나무형 규칙을 개발하고자 최근 여러 시도가 있었다. 본 연구의 목적은 반복분할(recursive partitioning)에 근거한 나무형 군집화 규칙을 개발하는 데 있다. 이러한 나무형 군집화(tree-structured clustering)는 군집화에 필요한 일부 변수만을 선별해주므로 해석이 간결하고 적용이 쉽다는 장점을 갖는다.

2. 제안 알고리즘

n개 개체, p개 변량으로 구성된 다변량 자료 $\{x_{ij}; i=1,\cdots,n, j=1,\cdots,p\}$에 대한 군집화를 생각하기로 하자. 다른 언급이 없는 한, p개 변량이 모두 연속형임을 가정하기로 한다. 우리는 이 연구에서 노드 분리 방식으로 2지 분리(binary split)만을 고려할 것이다. 제안 알고리즘은 다음과 같다.

1) 분리 기준과 방법: 개체 수가 n 개인 부모노드(parent node)를 개체 수가 각각 n_1개와 n_2개인 2개의 자식노드(child node)로 분리한다고 하자. 부모노드의 그룹내 제곱합-교차곱 행렬을 W, 자식노드들의 그룹내 제곱합-교차곱 행렬을 각각 W_1과 W_2라고 하자. 그 때 분리 정도를 보여주는 지표로

$$\text{Overall } R^2 = 1 - \text{tr}(W_1 + W_2)/\text{tr}(W) \tag{2.1}$$

를 정의하고 2지 분리의 평가 기준(evaluation criterion)으로 하자. 그러면 부모노드를

자식노드 1: $X_j \le s_j$, 자식노드 2: $X_j > s_j$

---

1) 고려대학교 통계학과 교수. 서울특별시 성북구 안암동 5가 [136-701].
2) 고려대학교 BK21 한국학 교육·연구단 박사후 연구원, 서울특별시 성북구 안암동 5가 [136-701].

로 분리하여 Overall R^2가 최대로 하도록 변수 $X_j\,(j=1,\cdots,p)$를 선택하고 경계값 $s_j \in (-\infty,+\infty)$를 찾는 문제가 된다.

2) 분리 결정 I: 만약 노드내 개체들이 2개 이상의 군집을 형성하고 있는 경우라면 Overall R^2은 '큰' 값을 취할 것이다. 그렇지 않은 경우엔 Overall R^2이 '작은' 값을 취할 것이다. 따라서 노드의 분리 여부를 결정하기 위한 Overall R^2의 임계값이 필요하다. Overall R^2의 영 분포에서 50% 분위수를 임계값으로 사용하면 median unbiased된 노드 분리를 할 수 있을 것이다. 영분포 생성을 위하여 N번의 모의시행을 한다.

3) 분리 결정 II: 분리 결정 I은 매우 많은 계산을 요구한다. 모의 준거자료의 분리에 필요한 계산량이 실제 관측자료의 분리에 요구되는 계산량의 N배이기 때문이다. 따라서 정확성이 일부 결여되더라도 현실적인 대안을 강구할 필요가 있다. 이를 위해 다음 Proposition을 활용한다.

Proposition 1. $N_p(0, C)$ 분포를 최대로 분리하는 변수 X_j는

$$\max_{j=1,\cdots,p} \sum_{k=1}^{p} | c_{jk} | \tag{2.2}$$

로 결정된다. 여기서 c_{jk}는 행렬 C의 (j, k) 요소이다. 그리고 최적 분리 값은 0이다.

3. 간단한 수치 예: Fisher의 붓꽃 자료

Fisher의 붓꽃 자료(iris data)는 3개 품종(1=Setosa, 2=Versicolor, 3=Virginica) 4개 변량(X1: 꽃받침길이, X2:꽃받침폭, X3:꽃잎길이, X4:꽃잎폭)의 150개 개체로 구성되어 있다. 노드의 크기가 50 이하인 경우 분리를 고려하지 않기로 하자. X1-X4로 군집화를 하면 그림 3.1의 나무가 형성된다.

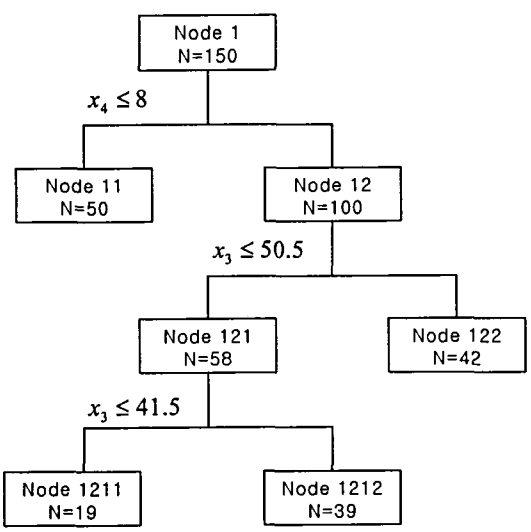

그림 3.1: 붓꽃 자료에 대한 군집화 나무 1

그림 3.1의 군집화 나무에 의한 개체들의 군집과 품종간 교차표는 다음과 같다. 품종 1이 노드 11에 모두 몰리고, 품종 2가 노드 1211, 1212, 122에 나뉘어지고, 품종 3은 노드 1212와 122에 나뉘어진다. 노드 1211과 노드 1212를 품종 2에, 노드 122를 품종 3에 대응시키면 총 10개 개체의 오(誤)대응이 발생한다. 이들은 대부분 노드 1212에서 나왔다.

Species / Node	1	2	3
11	50	0	0
1211	0	19	0
1212	0	30	9
122	0	1	41
합계	50	50	50

4. 결론

이제까지 모든 군집화 변수가 연속형이라고 가정하였다. 이항형인 군집화 변수가 있는 자료에 대한 나무형 군집화는 아마도 2절의 알고리즘을 그대로 적용하여도 될 것이다. 다항형의 군집화 변수가 있는 경우엔 문제가 더욱 복잡해지는데 이에 대하여는 추후의 연구로 넘기기로 하겠다.

참고문헌

강현철, 한상태, 최종후 (2000), 의사결정나무를 활용한 데이터마이닝 예측모형 해석, 한국통계학회 학술발표회 논문집, 2000년 춘계. 39-44.

최대우, 구자용, 최용석 (2004), 배경자료를 이용한 나무군집의 군집분석, 응용통계연구. 17권 3호, 535-545.

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

DeSarbo, W.S., Carrol, J.D., and Clark, L.A., and Green, P.E. (1984), "Synthesized clustering: A method for amalgamating alternative clustering bases with differential weighting of variables," *Psychometrika*, 49. 57-78.

Kass, G. (1980), "An exploratory technique for investigating large quantities of categorical data," *Applied Statistics*, 29(2). 119-219.

Liu, B., Xia, Y. and Yu, P.S. (2000), Clustering through decision tree construction. IBM Research Report RC21695.

Makarenkov, V. and Legendre, P. (2001), "Optimal variable weighting for ultrametric and additive trees and k-means partitioning: methods and software," *Journal of Classification*, 18. 245-271.

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