응용통계연구 제 18권 3호, 2005년, pp. 661-671

연속형 자료에 대한 나무형 군집화*

허명회[1] 양경숙[2]

요 약

본 연구는 반복분할(recursive partitioning)에 의한 군집화 방법을 개발하고 활용 예를 보인다. 노드 분리 기준으로는 Overall R-Square를 채택하였고 실용적인 노드 분리결정 방법을 제안하였다. 이 방법은 연속형 자료에 대하여 나무 형태의 해석하기 쉬운 단순한 규칙을 제공하면서 동시에 변수선택기능을 제공한다. 활용 예로서 Fisher의 붓꽃 데이터와 Telecom 사례에 적용해 보았다. K-평균 군집화와 다른 몇 가지 사항이 관측되었다.

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

1. 연구 배경과 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. 첫 번째는 K-평균 군집화와 같은 기존의 군집화 방법을 시행하여 군집 코드를 확보한 후에 군집 코드를 표적으로 나무형 분류를 적용하는 방법이다 (강현철외 2인, 2000). 이 2-단계 방법은 군집화 형성은 기존의 군집화로 하고 군집화 해석은 나무형 분류규칙으로 한다는 것인데 양자가 꼭 맞지는 않으므로 본격적인 나무형 군집화 기법으로 보기 어렵다. 두 번째는 실제자료와 대비되는 가상적인 준거자료(배경자료)를 생성시킨 후 실제자료와 준거자료를 구분하는 나무형 규칙을 만드는 방법이다 (Liu et al., 2000; 최대우외 2인, 2004). 좋은 아이디어이기는 하지만 준거자료의 개념 정의에 따라 군집화 결과가 달라질 수 있다.

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. 다만, 첫 연구에서는 대상 자료의 속성을 연속형으로 국한하기로 한다. 몇몇 사례에 이 기법을 적용해보고 토의하기로 한다.

* 이 연구는 고려대학교 특별연구비에 의하여 수행되었음.

1) (136-701) 서울특별시 성북구 안암동 5가. 고려대학교 통계학과 교수.
E-mail: stat420@korea.ac.kr

2) (136-701) 서울특별시 성북구 안암동 5가. 고려대학교 BK21 한국학 교육·연구단 박사후 연구원.
E-mail: myksyang@naver.com (교신저자)

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. 일반적으로 나무형 군집화 기법에서 요구되는 사항은 다음 두 가지일 것이다.

1) 노드를 분리한다면 어느 변수, 어느 값을 경계로 할 것인가?

2) 노드를 분리할 것인가, 아니면 분리하지 말 것인가?

우리는 이 연구에서 노드 분리 방식으로 CART에서처럼 2지 분리(binary split)만을 고려할 것이다.

2.1. 분리 기준과 방법

일반적으로 나무형 기법에서 요구되는 사항중 1)은 평가 기준으로부터 파생된다. 개체수가 n개인 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

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

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

$$\text{자식노드 1: } X_j \le s_j \text{ , 자식노드 2: } X_j > s_j$$

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. 다시 말하여, p 개 변수 각각에 대하여 $(n-1)$개 중간 자료값(mid-values)을 경계로 나누어 (2.1)의 지표를 산출하여 비교하면 된다. Overall R^2의 최대값을 그냥 Overall R^2로 하여도 혼동이 되지 않을 것이기 때문에 Overall R^2을 Maximum Overall R^2를 대신해 사용하기로 하겠다.

한편 Overall R^2는 각 변수의 척도에 의존하므로, 어떤 특별한 이유가 없는 한 각 변수의 척도를 군집화 이전에 균일하게 조정할 필요가 있다. 이에 따라, 본 연구에서는 모든 노드 분리에 앞서 평균 0, 표준편차 1의 표준화 처리를 할 것이다. 자식노드가 다음 단계에서 부모노드가 되면 평균 0, 표준편차 1의 표준화 처리를 새로 한다.

2.2. 분리 결정

나무형 기법에서 요구되는 2)의 문제를 해결하기 위한 방법으로 다음과 같이 분리결정 I, 분리결정 II 의 2가지 방법을 제안한다.

A) 분리결정 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 영분포(null distribution) 생성을 위해 다음 체계를 상정하기로 한다.

1) 현재 노드내 자료들의 분산-공분산 행렬을 구한다. 그것을 C라고 하자. 변수표준화에 의하여 C의 모든 대각요소는 1이다.

2) 독립적인 n개의 $p \times 1$ 임의벡터 $x_1, \cdots, x_n$을 $N_p(0, C)$ 로부터 생성시킨다. 따라서 준거개체들은 관측개체들과 유사한 1차 및 2차 모멘트를 가지며 단일 군집을 이룬다.

3) Overall R^2가 최대가 되도록 준거개체들을 2개의 군집으로 분리한다.

4) 단계 2부터 단계 3을 N(예컨대 100)번 반복함으로써 Overall R^2의 영 분포를 만든다.

Overall R^2의 영 분포에서 50 % 분위수를 임계값으로 사용하면 중위적 비편향된(median unbiased) 노드 분리를 할 수 있을 것이다. 여기서 중위적 비편향된 노드분리란 불필요한 노드 분리를 할 확률과 안할 확률을 같게 함으로써 군집화 나무의 크기를 중간수준이 되도록 제어한다는 뜻이다.

B) 분리 결정 II: 분리 결정 I은 매우 많은 계산을 요구한다. 모의 준거자료의 분리에 필요한 계산량이 실제 관측자료의 분리에 요구되는 계산량의 N(예컨대 100)배이기 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

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

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

그 이유는 다음과 같다. 식 (2.2)에서 노드내 표준화를 하는 경우 $\mathrm{tr}(W)$는 항상 p이다. 또한 변수 $X_j (j = 1, \cdots, p)$로 분리한 2개 그룹으로부터 얻는 총 설명력은

$$\text{Overall } R^2 = 1 - \textstyle\sum_{j=1}^{p}(1 - R_j^2)/p = \sum_{j=1}^{p} R_j^2/p$$

로 표현된다 (여기서 R_j^2는 변수 X_j에서 나온 결정계수). Overall R^2가 R_j^2의 합(sum) 형태로 표현되므로 $p = 2$인 경우로 문제를 축소하는 것이 가능하다. $N_2(0, C)$, $C = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$에서 일반성을 잃지 않고 $\rho \geq 0$ 을 가정하고 $X_1 = a \geq 0$ 에서 2개 그룹으로 분할하면 이로부터 결과되는 두 그룹 중심간 유클리드 제곱 거리의 크기(=BetweenSS)는 a(=Cut)과 ρ(=Rho)의 함수로 표현된다. 그림 2.1은 a와 ρ의 각 경우에서 10,000개의 개체를 모의생성시켜 얻은 BetweenSS를 플롯한 결과이다. 직관적으로 예상할 수 있듯이, BetweenSS는 Cut가 작아짐에 따라 증가하고 Rho가 커짐에 따라 증가한다. 따라서, Rho의 각 경우에서 BetweenSS는 Cut=0에서 가장 큰 값을 취하므로 Cut=0에서 Rho 값이 큰 변수와 결합하여야 총 설명력이 커진다. 그림 2.2는 Cut=0의 경우에서 Rho와 BetweenSS의 관계를 보여준다. BetweenSS가 Rho가 커짐에 따라 증가함을 볼 수 있다. 따라서 일반적으로 2인 경우 Proposition 1에 의하여 변수를 선택하면 Overall R^2가 최대가 된다.

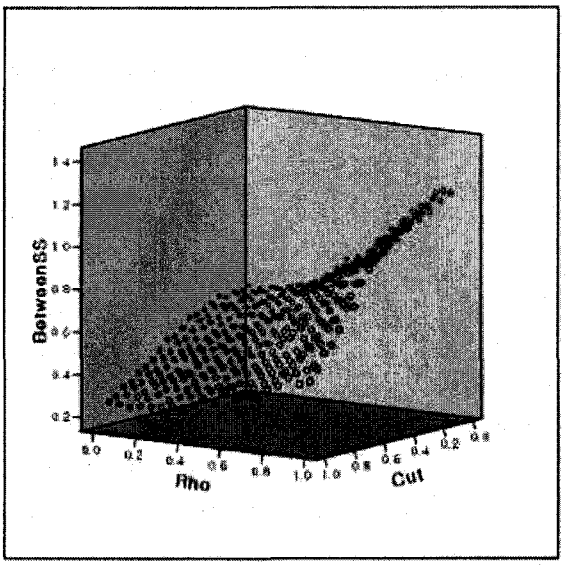

그림 2.1: Rho, Cut의 함수로서의 BetweenSS

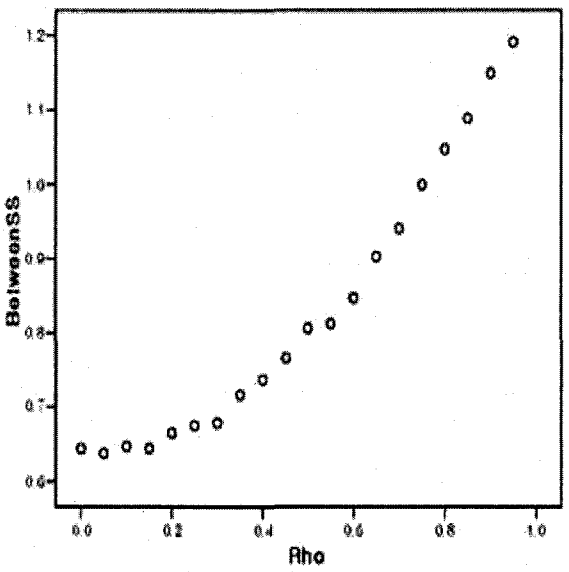

그림 2.2: Cut=0에서 Rho의 함수로서의 BetweenSS

3. 피셔의 붓꽃 데이터

Fisher의 붓꽃 자료(iris data)는 3개 품종(1=Setosa, 2=Versicolor, 3=Virginica) 4개 변량(X_1:꽃받침길이, X_2:꽃받침폭, X_3:꽃잎길이, X_4:꽃잎폭)의 150개 개체로 구성되어 있다. 노드의 크기가 50 이하인 경우 분리를 고려하지 않기로 하자. $X_1 - X_4$로 군집화를 하면 그림 3.1의 나무가 형성된다. 여기서 분리경계값은 원자료 척도로 다시 역변환시킨 결과이다.

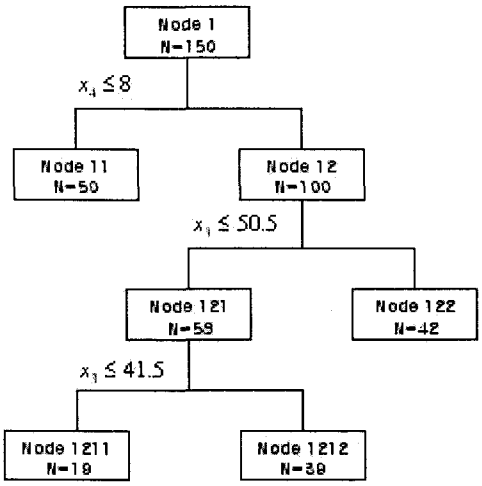

그림 3.1: 붓꽃 자료에 대한 군집화 나무 1

뿌리노드에서 X_4(=꽃잎폭)가 선택되었고 표준변량 값 $x_4 = 8$을 경계로 나눌 수 있었다. 이 때의 Overall R^2는 0.6294였고 분리 결정 II에 의한 노드 1의 대응 영 분포의 임계값은 0.4586이었다 (분리 결정 I에서 N=100번의 모의시행으로 구한 임계값은 0.464로 별 차이가 없었다). 따라서 노드 1이 노드 11과 노드 12로 나뉘어졌다. 노드 11은 크기가 50으로 최저크기 기준의 적용을 받아 재분리가 고려되지 않았고 노드 12는 크기가 100으로 재분리 대상이 되었다. 노드 12의 분리에서는 X_3(=꽃잎길이)가 선택되었고 노드내에서 재산출한 $x_3 = 50.5$를 경계로 노드 121과 노드 122로 나뉘어졌다. 이 때의 Overall R^2는 0.4459였고 대응하는 영 분포의 임계값은 0.4258이었다 (분리 결정 I에 의해 구한 임계값은 0.4524였다). 노드 121은 크기가 58로 재분리 대상이 되었고 노드 122는 크기가 42로 재분리 대상이 되지 않았다. 노드 121의 분리에서는 X_3(=꽃잎길이)가 선택되었고 $x_3 = 41.5$를 경계로 노드 1211과 노드 1212로 분리되었다. 이 때의 Overall R^2는 0.4186였고 대응하는 영 분포의 임계값은 0.3683이었다 (분리 결정 I에 의한 임계값은 0.316이었다). 노드 1211과 노드 1212의 크기는 19와 39로 최소크기 제한에 걸려 더 이상 재분리가 고려되지 않았다.

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

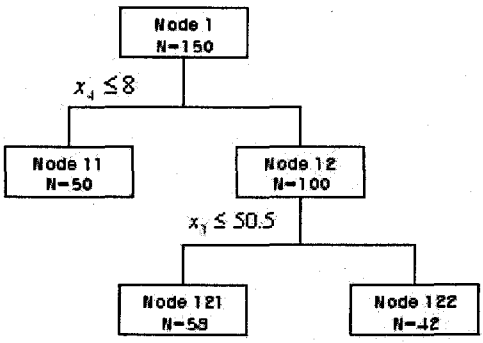

그림 3.2: 붓꽃 자료에 대한 군집화 나무 2

10개 개체의 오(誤)대응이 발생한다. 이들은 대부분 노드 1212에서 나왔다.

Species / Node	Setosa	Versicolor	Virginica
11	50	0	0
1211	0	19	0
1212	0	30	9
122	0	1	41
합계	50	50	50

나무의 깊이를 2로 제한하게 되면 노드 1211과 노드 1212가 합쳐져 그림 3.2와 같이 된다. 개체들의 군집과 품종간 교차표는 다음과 같다. 총 10개의 오대응이 있음을 볼 수 있다.

Species / Node	Setosa	Versicolor	Virginica
11	50	0	0
121	0	49	9
122	0	1	41
합계	50	50	50

참고로 군집 수 3의 K-평균 군집화의 결과는 다음과 같다 (SAS Proc Fastclus, Version 9). 이 때의 오(誤)대응 개체 수는 16개로 군집화 나무의 10개보다 다소 많다.

Cluster \ Species	Setosa	Versicolor	Virginica
1	50	0	0
2	0	48	14
3	0	2	36
합계	50	50	50

그림 3.1과 그림 3.2의 군집화 나무에는 4개 변수 중 2개 변수만 나타나지만 K-평균 군집화는 모든 변수가 포함된다는 점에서도 두 군집화 방법이 차이가 있다.

4. Telecom 사례분석

Telecom 회사의 1000명 고객의 통화기록 자료를 분석하기로 한다. Telecom 데이터는 통신분야의 데이터마이닝을 위한 SPSS사의 클레멘타인 응용 템플릿인 Teco CAT에 들어 있는 사례 데이터이다. 원래 데이터 크기는 이보다 규모가 훨씬 크고 변수들도 더 많지만 본 연구에서는 1000 케이스로 제한하였다. 본 연구에서 사용한 군집화 변수는 주간통화시간(X_1), 야간통화시간(X_2), 주말통화시간(X_3), 국제통화시간(X_4), 평균주간통화시간(X_5, 1통화당), 평균야간통화시간(X_6, 1통화당), 평균주말통화시간(X_7, 1통화당) 등 7개로 모두 연속형이다. 노드 분리를 위한 최소 크기를 250으로 하여 2절의 군집화 알고리즘을 적용해 보았다.

그림 4.1에서 보듯이 4개의 변수가 선발되었고 6개의 군집이 산출되었다. 이들 최종 군집의 특징은 다음과 같다.

최종 6개 노드의 평균값

Node	11	1211	1212	12211	12212	1222
n	202	167	228	229	166	8
X_1	49.70	518.37	702.00	456.22	599.14	483.98
X_2	39.21	126.55	99.34	301.47	274.33	254.36
X_3	20.92	60.01	46.37	48.34	41.89	51.74
X_4	18.62	206.75	228.73	221.45	241.87	212.30
X_5	5.85	5.70	5.98	6.13	6.21	4.98
X_6	3.71	8.42	7.60	6.66	4.97	178.56
X_7	3.96	6.79	4.16	4.89	6.30	3.69

- Node 11 (202명): 통신활동이 전반적으로 저조한 그룹, X_4(=국제통화시간)가 작다.

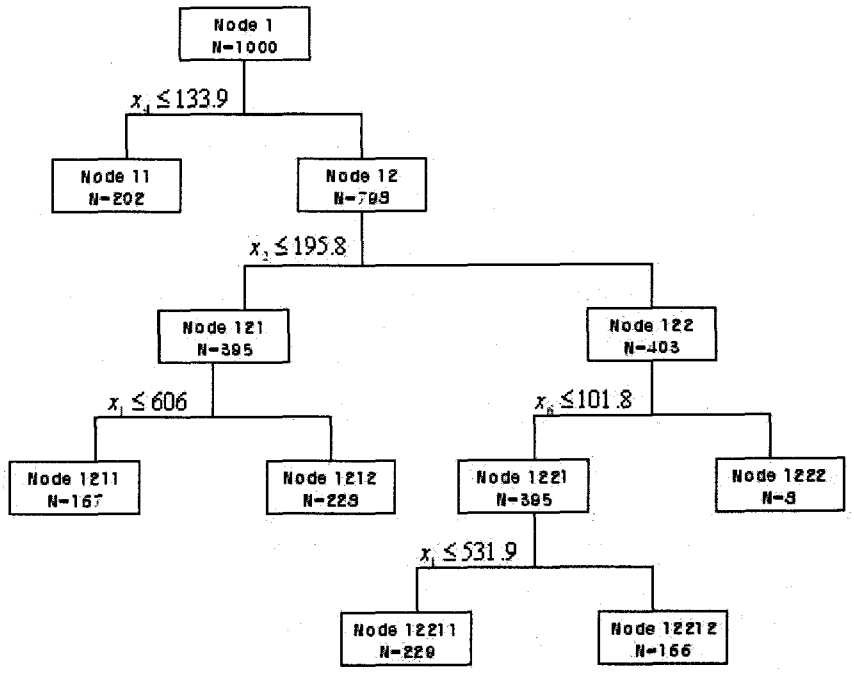

그림 4.1: Telecom 사례의 군집화 나무

- Node 1211 (167명): X_4(=국제통화시간)가 작지 않은 편이나 X_2(=야간통화시간)와 X_1(=주간통화시간)이 작은 편이다.
- Node 1212 (228명): X_4(=국제통화시간)가 작지 않은 편이나 X_2(=야간통화시간)는 작고 X_1(=주간통화시간)이 크다.
- Node 12211 (229명): X_4(=국제통화시간)가 작지 않은 편이나 X_2(=야간통화시간)가 큰 편이고 X_6(=평균야간통화시간)와 X_1(=주간통화시간)이 작다.
- Node 12212 (166명): X_4(=국제통화시간)가 작지 않은 편이나 X_2(=야간통화시간)가 큰 편, X_6(=평균야간통화시간)이 작고 X_1(=주간통화시간)이 크다.
- Node 1222 (8명): X_4(=국제통화시간)가 작지 않은 편이나 X_2(=야간통화시간)가 크다. X_6(=평균야간통화시간)가 특이하게 크다. 군집크기가 매우 작다.

군집 수를 6으로 하여 K-평균 군집화를 해보았다. 그 결과는 다음과 같다.

6개 군집의 중심

Cluster	1	2	3	4	5	6
n	10	50	9	15	197	719
X_1	462.18	518.34	528.00	550.40	48.61	572.10
X_2	154.38	194.26	171.13	207.34	39.23	201.22
X_3	14.40	64.00	79.00	55.99	20.80	47.66
X_4	167.52	212.05	217.61	217.39	18.18	225.21
X_5	80.55	9.63	2.13	5.21	4.52	5.15
X_6	3.69	5.13	3.12	155.09	3.72	5.96
X_7	2.37	22.04	72.77	3.76	3.54	3.54

- Cluster 1 (10명): X_3(=주말통화시간)은 작으나 X_5(=주간통화시간)가 가장 크다.
- Cluster 2 (50명): X_7(=평균주말통화시간)과 X_5(=평균주간통화시간)가 대체로 크다.
- Cluster 3 (9명): X_3(=주말통화시간)과 X_7(=평균주말통화시간)이 상대적으로 크다.
- Cluster 4 (15명): 다른 그룹에 비해 X_6(=평균야간통화시간)이 가장 길다.
- Cluster 5 (197명): 전반적으로 통신활동이 낮은데 특히 X_4(=국제통화시간)가 작다.
- Cluster 6 (719명): 전반적으로 통신활동이 활발한 그룹으로 특히 X_1(=주간통화시간)과 X_4(=국제통화시간)가 다른 그룹에 비해 상대적으로 길다.

Cluster 5와 Cluster 6에 대다수의 케이스가 속하고 나머지는 소수 군집이라고 하겠다. 다음은 나무형 군집화와 K-평균 군집화 사이의 교차분류표이다. K-평균 군집화의 Cluster 5와 나무형 군집화의 Node 11은 거의 대응하는 관계에 있지만, 전체의 70 % 정도를 점유하는 K-평균 군집화의 Cluster 6은 나무형 군집화의 Node 1211, 1212, 12211, 12212 등으로 나뉘는 모습을 볼 수 있다. 이렇게 되는 근본적인 이유는 K-평균 군집화가 1회의 사전 척도화를 하여 개체간 거리 측도를 고정시키는 반면 나무형 군집화는 노드별로 매번 사전 척도화를 하여 개체간 거리 측도를 가변화한다는 데 있는 것으로 생각된다.

Node \ Cluster	1	2	3	4	5	6	Total
11	3	1	1	0	197	0	202
1211	1	17	3	3	0	143	167
1212	2	6	2	4	0	214	228
12211	1	15	1	0	0	212	229
12212	3	11	2	0	0	150	166
1222	0	0	0	8	0	0	8
Total	10	50	9	15	197	719	1000

5. 맺음말

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.

군집화에서 각 변수의 척도 결정은 변수 가중(variable weighting)의 문제로 알려져 있고(DeSarbo, Carrol, Clark and Green, 1984) K-평균 군집화에서의 변수 가중 문제도 일부 연구된 바 있다(Makarenkov and Legendre, 2001). 나무형 군집화는 4절에서 언급한 바와 같이 변수 가중화를 1회로 확정짓지 않으므로 융통성 있게 내재하는 군집을 찾아낼 것으로 기대한다.

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>, **17**, 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**, 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.

[2005년 3월 접수, 2005년 4월 채택]

Tree-structured Clustering for Continuous Data*

Myung-Hoe Huh[1] Kyung-Sook Yang[2]

ABSTRACT

The aim of this study is to propose a clustering method, called tree-structured clustering, by recursively partitioning continuous multivariate data based on overall R^2 criterion with a practical node-splitting decision rule. The clustering method produces easily interpretable clustering rules of tree types with the variable selection function. In numerical examples (Fisher's iris data and a Telecom case), we note several differences between tree-structured clustering and K-means clustering.

Keywords: Tree-structured clustering, Node splitting, Overall R-Square, K-means clustering, Variable selection.

* This work was supported by a Korea University Grant.

1) Professor, Dept. of Statistics, Korea University. Anam-Dong 5-Ga, Sungbuk-Gu, Seoul 136-701, Korea
E-mail: stat420@korea.ac.kr

2) Post Doctoral Researcher, Brain Korea 21 The Education and Research Group for Korean Studies. Korea University. Anam-Dong 5-Ga, Sungbuk-Gu, Seoul 136-701, Korea.
E-mail: myksyang@naver.com (Communication author).