응용통계연구 제 19권 2호, 2006년, pp. 271-282

혼합형 데이터에 대한 나무형 군집화*

양경숙[1] 허명회[2]

요 약

본 논문에서는 범주형과 연속형 변수들이 혼합된 데이터에 적용할 수 있는 나무형 군집화 알고리즘을 제안하였다. 특히 혼합된 변수들이 공통의 의미를 갖도록 하기 위해 범주형 변수들을 전처리하는 방법을 고안하였다. 수치 예로서 SPSS의 신용(credit) 데이터와 독일신용자료(German credit data)에 알고리즘을 적용하고 그 결과를 검토하였다.

주요용어: 혼합형 데이터, 나무형 군집화, 노드분리, 변수 선택.

1. 연구배경 및 목적

본 연구의 목적은 범주형 변수와 연속형 변수가 혼합되어 있는 데이터에 적용할 수 있는 나무형 군집화 알고리즘을 개발하는 것이다. 선행연구(허명회·양경숙, 2005)에서 제안했던 연속형 데이터에 대한 나무형 군집화는 군집분리에 중요한 변수선택을 다루며 시각적으로 결과를 나타내기 때문에 군집에 대한 해석이 용이한 장점이 있다.

현재까지 혼합형 데이터에 대한 나무형 군집화 연구는 찾아보기 어렵다. Liu et al.(2000)이 제안한 나무구조의 군집분석 알고리즘이나 최대우 외 2인(2004)이 제안한 배경자료를 이용한 나무구조의 군집분석은 연속형 데이터, 혹은 연속형과 이산형 데이터에 대한 것이다. 엄격히 범주형 데이터에 대한, 혹은 혼합된 데이터에 대한 연구는 아니다. 따라서 나무형 군집화 알고리즘 제안에 앞서 범주형 변수만 있는 경우 혹은 연속형 변수와 혼합되어 있는 경우 적용하도록 제안된 군집화 알고리즘(clustering algorithm)을 간략히 살펴보자. 군집화 알고리즘에 대해 정리된 문헌은 많지만 최근의 자료로 Zhang et al.(1997)과 Berkhin(2002)의 논문을 주로 검토하기로 한다.

범주형 자료에 대한 군집화 알고리즘으로 1987년에 Fisher가 제안한 COBWEB은 이산형 데이터에 대한 군집화 알고리즘이다. 이는 범주형 변수의 개별적인 속성들을 이산화 시킬 때 범주의 수가 클 경우 비효율적이라는 단점을 가지고 있다. 또한 연속형 변수가 있을 경우, 사전에 '이산화' 작업이 요구되며 변수들 사이의 상관관계를 고려하지 않는다는

* 이 연구는 고려대학교 특별연구비에 의하여 수행되었음.

1) (136-701) 서울특별시 성북구 안암동 5가, BK21 한국학 교육·연구단, 박사후 연구원.
E-mail: myksyang@dreamwiz.com
2) (136-701) 서울특별시 성북구 안암동 5가, 고려대학교 통계학과, 교수.
E-mail: stat420@korea.ac.kr

한계점을 가지고 있다. 1989년에 Gennari, Langley, Fisher가 제안한 CLASSIT 알고리즘은 COBWEB과 유사하나 연속형 데이터에만 적용하며 정규분포를 가정한다(Zhang et al., 1997).

한편 확률모형에 근거하지 않고 거리에 기초한 군집분석은 분할 군집화(partitioning clustering)와 위계적 군집화(hierarchical clustering)로 나눌 수 있다. 분할 군집화는 1973년에 Duda 와 Hart에 의해서, 그리고 1990년에 Kaufman과 Rousseeuw에 의해 연구되었다. 이들이 제안한 방법은 모든 가능한 데이터 값을 반복적으로 시도하여 적절하게 나누는 것으로 각 군집은 K-평균군집과 같이 군집의 중심으로 대표된다. 그러나 K-평균군집과 마찬가지로 초기값에 민감하다. 위계적 군집화는 1973년 Duda와 Hart, 1983년 Murtagh에 의해 연구되었는데 최적의 군집화를 달성하는 것이 목적이 아니라 좀 더 유사한 군집들을 묶는데 연구 초점이 있었다(Zhang et al., 1997).

Zhang et al.(1997)은 대용량 데이터에 효율적인 군집화 방법으로 BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)라는 알고리즘을 제안하였다. 이는 대용량 데이터를 1차적으로 몇 개의 부그룹(subgroup)에 대한 기술통계량으로 요약한다. 이때의 기술통계량을 군집특성(CF-Tree: clustering feature)이라고 부르는데 2차적으로는 원데이터 대신 요약된 통계량들을 이용해서 군집화를 하는 방법이다. 따라서 대용량 데이터에 대해 컴퓨터 메모리의 한계나 알고리즘 수행시간을 줄일 수 있는 장점이 있다. 이 연구도 계량형 데이터에만 적용 가능하지만 기존의 COBWEB, CLASSIT과 달리 변수들간 독립을 전제하지 않는다는 차이점이 있다.

TomChiu et al.(2001)은 우도함수를 이용하여 Zhang이 제안한 BIRCH 알고리즘의 단점을 극복한 군집화 방법을 제안하였다. 현재 이 알고리즘은 SPSS에서 2단계 군집분석방법으로 제공되고 있다.

국내에서는 김보화·김규성(2002)이 대용량 범주형 데이터에 대한 K-모드 알고리즘과 ROCK 알고리즘을 비교 연구하였다.

이들 연구에서는 범주형 변수들을 포함한 데이터에 대한 군집화 알고리즘 개발이 주된 목적이었다. 따라서 의사결정나무와 같이 분리된 노드들의 특성을 선택된 일부 변수들에 의해 쉽게 해석할 수 있는 방법은 제안되지 않았다.

본 연구에서는 혼합형 데이터에 대한 이러한 문제점을 감안하여 시각적으로 해석이 용이한 나무형 군집화 알고리즘을 제안하고 두 가지 수치 예에 적용하여 그 결과를 검토하도록 한다.

2. 범주형 변수의 코딩과 전처리

연속형·범주형 혼합 자료의 군집화에서 서로 다른 종류의 변수들이더라도 공통의 의미를 갖도록 자료를 전처리할 필요가 있다. 예컨대 데이터마이닝 소프트웨어인 SPSS 클레멘타인은 연속형 변수를 최소값 0, 최대값 1이 되도록 척도화 하면서 동시에 범주형 변수에 대하여는 1차로 각 범주를 더미(dummy) 코드로 표현한 다음 $\sqrt{2}$ 로 나눠 준다 (그 이유는 그 변수가 취하는 범주가 상이한 2개 개체간 거리가 1이 되도록 하기 위해서이다). 그러나

이와 같은 범위 표준화는, 1) 범주형 변수에서는 값이 다르기만 하면 1의 차이가 나지만 2) 연속형 변수에서는 한 개체가 최소값을 취하고 다른 개체가 최대값을 취하는 경우에만 1의 차이가 나기 때문에, 통계인들이 받아들이기 쉽지 않다.

우리는 선행연구(허명회·양경숙, 2005)에서 연속형 자료에 대한 전처리로서 각 변수의 표준편차가 1이 되는 표준화변환을 한 바 있으므로 여기서 그에 맞추어 범주형 변수를 코딩하고 척도화하는 방향으로 진행하고자 한다.

연속형 변수 X에 대하여는 $X = x_1, \cdots, x_n$을

$$z_i \leftarrow (x_i - \overline{x})/s, \quad i = 1, \cdots, n$$

로 바꾼다. 이것의 의미를 다음과 같은 맥락에서 해석하고자 한다. 모든 개체 쌍 차이 합은

$$\sum_{i=1}^{n}\sum_{i'=1}^{n} d^2(x_i, x_{i'}) = 2n(n-1)s^2$$

가 되는데 여기서 s는 변수 X의 표준편차이다. 따라서

$$\sum_{i=1}^{n}\sum_{i'=1}^{n} d^2(x_i, x_{i'})/s^2 = 2n(n-1)$$

즉,

$$\sum_{i=1}^{n}\sum_{i'=1}^{n} d^2(x_i/s, x_{i'}/s) = 2n(n-1)$$

이다. 따라서 변수 X를 s로 나누어 척도화해 줌으로써 임의의 2개 개체 사이 평균 제곱거리는 2로 동일하게 된다.

이제부터 범주형 변수 X의 경우를 생각하기로 하자. $X = x_1, \cdots, x_n$이 k개 범주 중 1개를 취한다고 하자: $x_i = 1, \cdots, k$ for $i = 1, \cdots, n$. 일차적으로, 변수 X를 더미변수 $D^{(1)}, \cdots, D^{(k)}$로 나타낸다. 즉,

$$x_i = j \quad (j = 1, \cdots, k) \quad \Longleftrightarrow \quad d_i^{(j)} = 1, \quad d_i^{(j')} = 0, \quad for \quad j' \neq j.$$

그러면

$$\sum_{i=1}^{n}\sum_{i'=1}^{n} d^2(x_i, x_{i'}) = \sum_{j=1}^{k} 2n_j(n - n_j) = 2(n^2 - \sum_{j=1}^{k} n_j^2)$$

이 된다. 여기서 $n_j = \#\{i : x_i = j\}$. 따라서 더미 변수에 대한 척도화 인수(scaling factor) s는 다음 식을 만족해야 한다.

$$2n(n-1)s^2 = 2(n^2 - \sum_{j=1}^{k} n_j^2).$$

즉,

$$s^2 = (n^2 - \sum_{j=1}^{k} n_j^2)/n(n-1) \simeq 1 - \sum_{j=1}^{k} p_j^2, \quad \text{여기서} \quad p_j = n_j/n.$$

수치 예로서 변수 X가 범주값 a, b, c를 5, 3, 2의 빈도로 취하는 경우 ($n = 10$), 척도화 인수 s는

$$s = \sqrt{1 - 0.5^2 - 0.3^2 - 0.2^2} = 0.787$$

이므로 더미 변수 값 1은 1.27(=1/0.787)로 재코딩(recoding)되어야 한다.

특수한 경우로서 이항형 변수, 즉 $k = 2$인 경우에서는 $p_1 + p_2 = 1$이므로

$$s = \sqrt{1 - p_1^2 - p_2^2} = \sqrt{2p_1p_2}$$

이다. 따라서 $X = 1$이면 $(1/\sqrt{2p_1p_2}, 0)$으로, $X = 2$이면 $(0, 1/\sqrt{2p_1p_2})$으로 코딩한다.

이항형 변수 X를 1개 변수로 코딩해야 한다면 0 또는 $1/\sqrt{p_1p_2}$여야 할 것이다. 왜냐하면 X의 표준편차가 $\sqrt{p_1p_2}$이기 때문이다. 그렇기 때문에 이항형 변수 X를 2개의 더미 코드로 표현한 경우와 개체간 거리는 같게 된다. 다시 말하여, 이항형 변수는 어떤 방식으로 코딩하더라도 마찬가지이다.

이상과 같은 범주형 변수의 코드화와 전처리는 다음과 같은 일반적 성질을 갖는다: k개 범주를 갖는 변수 X를 앞의 재코딩 방식으로 k개의 표준화 더미변수 $Z_1, \cdots, Z_k$의 합으로 표현하면

$$\sum_{j=1}^{k} var(Z_j) = s^{-2} \sum_{j=1}^{k} p_j(1-p_j) = (1 - \sum_{j=1}^{k} p_j^2)/s^2 = 1$$

이 된다. 즉 각 표준화 더미 변수 분산들의 총합은 항상 1이 된다. 이것은 연속형 변수에 대한 표준화 변환이 분산 1을 취하는 것과 맥을 같이 한다.

3. 제안 알고리즘

n개 개체, p_c개의 연속형 변수와 p_d개의 범주형 변수로 구성된 혼합형 다변량 자료에 대한 군집화를 생각하기로 한다. 일반적으로 나무형 군집화 기법에서 요구되는 사항은 다음 두 가지이다.

1) 노드를 분리한다면 어느 변수, 어느 값을 경계로 할 것인가?

2) 노드를 분리할 것인가, 아니면 분리하지 말 것인가?

CART에서와 같이 노드 분리 방식으로 2지 분리(binary split)만을 고려할 것이다.

노드 분리 : 변수 X_j가 연속형인 경우는 부모노드를

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

로 분리하면 되고 이 때 X_j의 순서화 자료값들의 $(n-1)$개 중간값(mid-values)이 분리점 s_j의 후보가 된다. 문제는 변수 X_j가 범주형인 경우인데 범주형을 다시 명목형과 순서형으로 구분하여 분리 방법을 달리할 필요가 있다.

1) 변수 X_j 가 k개의 범주를 갖는 명목형인 경우는 일차적으로 2절에서 제안한 더미변수 코딩과 척도화를 거친다. 2지형의 노드 분리에 있어서는 k개의 범주 $\alpha_1, \cdots, \alpha_k$를 임의의 2개 소그룹으로 나누는 모든 경우를 고려해야 할 것이다. 그 경우의 수는 $2^{k-1}-1$이다. 예컨대 $k=4$인 경우 다음의 7가지 범주 결합을 고려해야 한다.

범주	경우1	경우2	경우3	경우4	경우5	경우6	경우7
1	좌	좌	좌	좌	좌	좌	좌
2	우	좌	우	우	좌	좌	우
3	우	우	좌	우	좌	우	좌
4	우	우	우	좌	우	좌	좌

2) 변수 X_j가 k개의 범주를 갖는 순서형인 경우는 일차적으로 2절에서 제안한 더미변수 코딩과 척도화를 거친다. 그러나 범주간 순서가 있기 때문에 k개의 범주 $\alpha_1, \cdots, \alpha_k$를 2개 소그룹으로 나누는 경우 수는 $k-1$이다. 예컨대 $k=4$인 경우 다음의 3가지 범주 결합한다.

범주	경우1	경우2	경우3
1	좌	좌	좌
2	우	좌	좌
3	우	우	좌
4	우	우	우

변수 선택 : 연속형 변수와 범주형 변수에 대한 전처리(척도화와 중심화 포함)를 거쳐

$$Overall \quad R^2 = 1 - trace(W_1 + W_2)/trace(W) \tag{3.1}$$

를 최대화시키는 변수 X_j와 분리경계 값 s_j 또는 범주 묶음을 찾아 부모 노드를 2개 자식 노드로 분리한다. 여기서, W는 부모노드의 그룹내 제곱합-교차곱 행렬이고 W_1과 W_2는 자식노드들의 그룹내 제곱합-교차곱 행렬을 나타낸다.

분리 결정 : 선행연구인 (허명회·양경숙, 2005)에서 연속형 자료에 대한 분리 결정 방식을 원용한다. 그것을 간단히 정리하면 다음과 같다. Overall R^2의 임계값을 정하기 위해 아래 방식으로 Overall R^2의 영 분포(null distribution)를 생성한다.

1) 현재 노드내 자료들의 분산-공분산 행렬을 구한다. 그것을 C라고 하자. 변수 전처리(척도화와 중심화 포함)를 거쳐 C의 모든 대각요소는 1이 된다.

2) 독립적인 n개의 $p_0 \times 1$ 임의벡터 $x_1, \cdots, x_n$을 $N_{p_0}(0, C)$로부터 생성시킨다. 여기서 $p_0 = p_c + \sum_{j=1}^{p_d} k_j$. 따라서 준거개체들은 관측개체들과 유사한 1차 및 2차 모멘트를 가지며 단일 군집을 이룬다.

3) $\sum_{k=1}^{p_0} |c_{jk}|$를 최대로 하는 변수 X_j를 찾는다. 여기서 $c_{j,k}$는 행렬 C의 (j,k) 요소이다. 분리경계 값을 0으로 한다.

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

Overall R^2의 영 분포에서 50 % 분위수 Overall $R^2_{0.50}$를 임계값으로 사용하면 중위적 비편향된(median unbiased) 노드 분리를 할 수 있다. Overall R^2_{max}가 임계값 보다 크지 않으면 더 이상 노드 분리를 하지 않는다. 노드의 크기가 일정 크기 이하일 경우에도 더 이상의 분리를 고려하지 않는다. 더불어 나무의 깊이(depth)에 제한을 둘 수도 있을 것이다.

4. 수치 예 : SPSS 신용(credit) 데이터

SPSS사의 AnswerTree에 포함되어 있는 신용(credit) 데이터는 총 323케이스로 직업(X_1:class), 급여 지급형태(X_2:pay), 연령(X_3:age), Amex카드 소지여부(X_4:amex)와 그들의 신용여부를 나타내는 종속변수(Y)로 구성되어 있다. 본 연구에서는 종속변수를 제외한 4개 변수들을 가지고 개체들에 대한 나무형 군집화를 시도하도록 한다. 변수들의 각 범주는 다음과 같다.

X_1 : 관리직(1), 전문직(2), 사무직(3), 숙련 노동직(4), 비숙련노동직(5)
X_2 : 주급(0), 월급(1)
X_3 : 25세 미만(1), 25세 이상 35세 이하(2), 36세 이상(3)
X_4 : 무소유(0), 소유(1)

여기서 변수 X_2와 X_4는 이항형이므로 2절에서 언급한 것처럼 2개의 더미변수를 활용하였다. X_1은 명목형이므로 5개의 더미변수를 생성하고 노드 분리시 15가지 경우를 고려하였다. X_3에 대해서는 순서형으로 처리하여 3개의 더미변수를 생성시킨 후 2가지 경우를 고려하여 군집화 알고리즘을 적용하였다. 이때 노드분리 중지 기준으로 노드의 표본크기가 전체 표본크기의 25 %이하가 되면 중지하도록 하였다.

그림 4.1은 나무형 군집화 알고리즘을 수행시킨 결과이다. 그리고 각 노드분리시 계산한 Overall R^2_{max}와 분리기준 II에 의해 계산된 임계값 Overall $R^2_{0.50}$은 다음과 같다.

Node	Overall R^2_{max}	Overall $R^2_{0.50}$
Node1	0.3359	0.2168
Node11	0.3366	0.1864
Node12	0.3361	0.1909
Node112	0.3008	0.2897

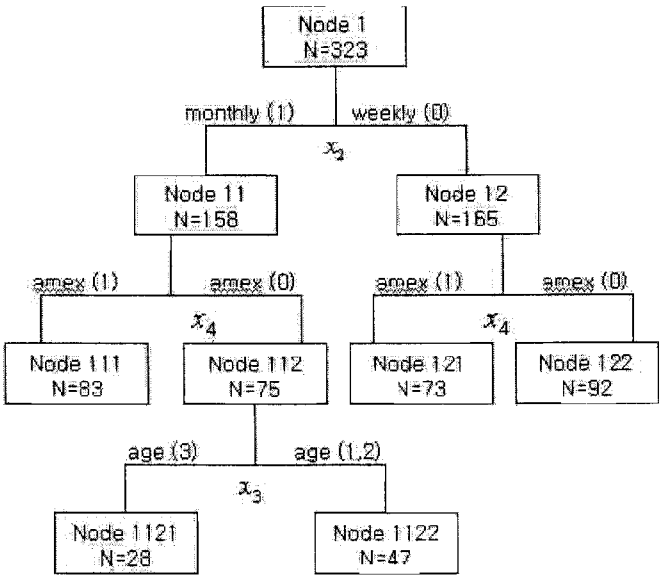

그림 4.1: SPSS 신용(credit) 데이터에 대한 나무형 군집화 결과

첫 번째 노드분리에서 발생한 2개의 군집을 원래의 종속변수(Y)와 대비시켜보면 약 85 %의 정확률을 나타낸다. 최종 선택된 5개의 노드와 신용유무와의 관계는 다음과 같다.

Y	Node					Total
	Node 111	Node 1121	Node 1122	Node 121	Node 122	
Good	68 (81.9 %)	28 (100.0 %)	37 (78.7%)	9 (12.3 %)	13 (14.1%)	155 (48.0 %)
Bad	15 (18.1 %)	0 (0.0 %)	10 (21.3%)	64 (87.7 %)	79 (85.9%)	168 (52.0 %)
Total	83	28	47	73	92	323

이 테이블에서 각 노드별 'Good'인 비율과 데이터 전체에서 'Good'이 차지하는 비율 48.0 %간의 절대차이를 계산한 후 각 노드별 개체수를 가중치로 하여 가중평균을 계산하면 35.29 %로 계산된다. 따라서 3개 변수를 이용한 나무형 군집화의 효과는 35.29 %라고 할 수 있겠다.

5. 독일 신용 데이터

독일 신용 데이터(German credit data)는 1994년 독일의 한스 호프만(Hans Hofmann) 교수에 의해 만들어진 것으로 신용자 700명, 신용불량자 300명에 대해 관측한 당좌예금잔

고, 신탁기간, 신용전력, 신용금액, 거주기간 등 21개 변수로 구성되어 있다.

본 연구에서는 그룹분리 코드(목표변수)를 제외한 20개의 변수를 가지고 나무형 군집화를 시도하였다. 그중 범주가 11개인 X_4(대출목적) 변수는 5개의 범주로 재결합시켰다. 즉 범주를 (40,41), (42,43,44,45), 46, (48,49), 410(범주 47은 해당 데이터가 없으므로 제외됨)으로 재분류하였다. 나무형 군집화를 적용하기 위해 고려한 명목형 변수는 X_4 외에 X_9(personal status and sex), X_{14}(other installment plans), X_{15}(housing) 등이다. 순서형 변수로는 X_1(checking account), X_3(credit history), X_6(savings account), X_7(employment years), X_{10}(other debtors), X_{12}(property), X_{17}(job category)을 고려하였고 X_{19}(telephone)과 X_{20}(foreign worker)를 포함한 그 밖의 9개 변수들은 이항형 변수들로 2절에서 제안한 방법으로 다루어졌다.

나무형 군집화 결과는 그림 5.1과 같다. 1000명의 고객을 세분화하는데 가장 중요한 변수는 고객이름으로 전화번호가 등록되었는지 그렇지 않은지를 나타내는 X_{19}(telephone)임을 알 수 있다. 전화번호가 등록된 404명은 다시 그들의 신탁금액(X_5)에 의해 분리되고 전화번호가 등록되어 있지 않은 나머지 596명은 현재 은행에 개설된 계좌개수(X_{16})에 의해 분리되고 있다. 나무깊이 3에서는 모두 채무공동의무자(people being liable) 수(X_{18})에 의해 노드 분리가 이루어지고 있다.

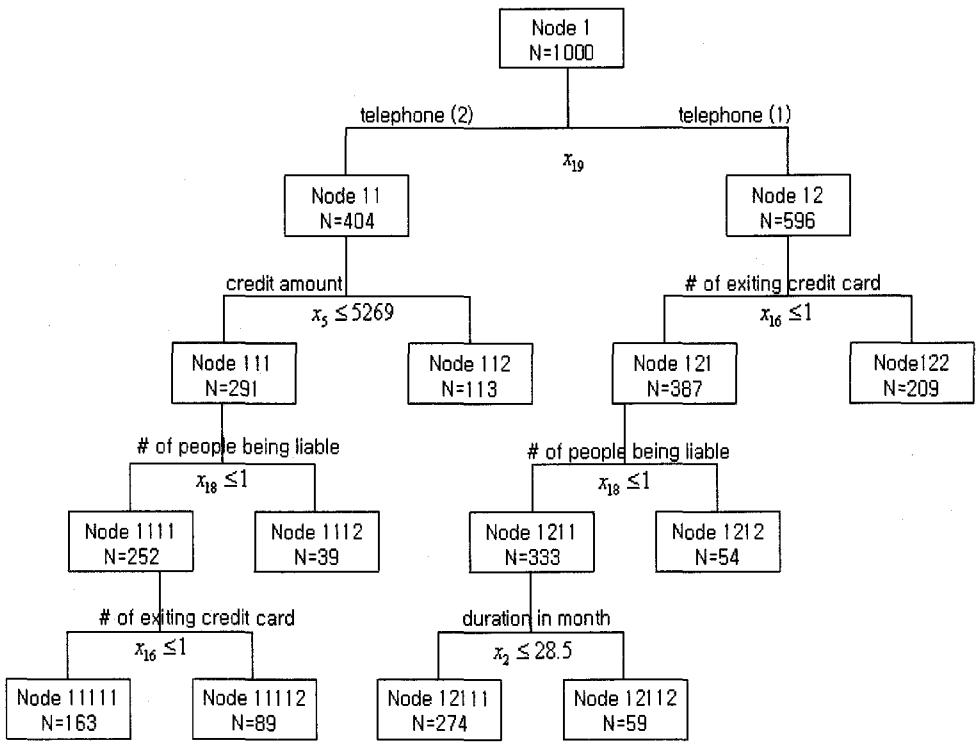

그림 5.1: 독일 신용 데이터에 대한 나무형 군집화 결과

다음은 분리기준이 된 변수 X_{19}(telephone), X_5(credit amount), X_{18}(# of people being liable), X_{16}(# of existing credits), X_2(duration in month)에 대하여 최종 8개 노드별 분포를 살펴본 결과이다.

Node	Size	X_{19}		X_5		X_2		X_{18}		X_{16}			
		1	2	mean	stddev	mean	stddev	1	2	1	2	3	4
11111	163	0	163	2455.8	1158.0	18.39	9.0	163	0	163	0	0	0
11112	89	0	89	2242.1	1240.4	18.5	8.3	89	0	0	77	11	1
1112	39	0	39	2675.4	1231.5	21.2	10.6	0	39	20	14	3	2
12	113	0	113	8857.2	2973.9	33.4	13.7	92	21	63	46	3	1
12111	274	274	0	2011.1	1289.3	14.7	5.9	274	0	274	0	0	0
12112	59	59	0	4869.9	2649.6	40.5	9.7	59	0	59	0	0	0
1212	54	54	0	2607.2	1997.8	18.3	13.1	0	54	54	0	0	0
122	209	209	0	2808.8	2288.6	19.3	10.6	168	41	0	196	11	2
Total	1000	596	404	3271.2	2822.7	20.9	12.0	596	404	633	333	28	6

각 노드분리에서 계산한 Overall R^2_{max}와 분리기준 II에 의해 계산된 임계값 Overall $R^2_{0.50}$은 다음과 같다. 이 임계값은 분석데이터와 동일 공분산 구조를 갖는 난수를 표본수만큼 100번 발생시켜 만든 임계값 분포로부터 구한 50분위수이다.

Node	Overall R^2_{max}	Overall $R^2_{0.50}$	Node	Overall R^2_{max}	Overall $R^2_{0.50}$
Node1	0.063	0.054	Node11	0.064	0.058
Node12	0.065	0.054	Node111	0.060	0.048
Node121	0.065	0.059	Node1111	0.065	0.051
Node1211	0.066	0.065			

다음은 원 데이터에 포함된 종속변수와 나무형 군집분석에서의 8개 최종 노드를 비교한 결과이다. 각 노드별 'Bad'가 차지하는 비율과 전체 'Bad' 비율 30.0 %간의 절대차이를 계산한 후 군집별 크기를 가중치로 가중평균을 계산하면 6.4 %가 된다. 또한 각 노드별로 계산된 지니지수를 이용하여 Gini지수의 감소량을 계산하면 0.015 이다.

	나무형 군집화 결과								Total
	11111	11112	1112	12	12111	12112	1212	122	
good	125	70	34	62	194	20	34	151	700
(%)	(76.7)	(78.7)	(87.2)	(54.9)	(70.8)	(50.8)	(63.0)	(72.7)	(70.0)
bad	38	19	5	51	80	29	20	58	300
(%)	(23.3)	(21.3)	(12.8)	(45.1)	(29.2)	(49.2)	(37.0)	(27.8)	(30.0)
Total	163	89	39	113	274	59	54	209	1000
Gini	0.357	0.335	0.223	0.495	0.413	0.499	0.466	0.400	
Entropy	0.235	0.225	0.166	0.298	0.262	0.300	0.286	0.256	

6. 맺음 말

100개 이상의 변수들을 가진 대용량 데이터에 대한 군집분석을 시도할 때 통계전공자들은 보통 주성분분석을 하여 차원을 축소시키고 나서 군집분석을 하려고 한다. 그러나 주성분 분석에 의한 군집분석 결과를 실무자들은 별로 선호하지 않는다. 왜냐하면 주성분 축의 해석을 각 변수들을 보면서 또 해야 하기 때문이다.

본 연구에서 제시한 방법은 변수들이 많을 경우 군집분석에 매우 중요한 변수들을 선별해준다. 차원축약 없이 중요한 일부 변수를 가지고 군집의 특성을 쉽고 빠르게 이해할 수 있는 것은 물론이다. 특히 연속형 변수들이나 범주형 변수들에 관계없이 두 가지 유형의 데이터가 혼합되어 있을 경우 적용할 수 있다는 장점을 지닌다.

한편 혼합형 데이터에 대한 지도학습(supervised learning)에서는 CART, C4.5 등 나무형 분류기법들이 변수선택 편의 문제를 가지고 있다 (Loh and Shih,1997). 즉, 범주의 개수가 많을수록 중요 변수로 선택될 가능성이 높다. 분류나무의 편의문제에 대해서는 송문섭·윤영주(2001), Lee and Song(2002), Song et al.(2004), 정성석·김순영·임한필(2004) 등이 연구한 바 있다. 본 연구가 제안하는 군집화 방법론에서도 범주의 수가 많은 변수일수록 채택되는 편향이 존재할 수 있지만 이 문제에 대하여는 아직 본격적인 탐구를 하지 못하였다.

참고문헌

김보화, 김규성 (2002). K-모드 알고리즘과 ROCK 알고리즘의 개선, <응용통계연구 >, **15**, 381-393.

송문섭, 윤영주 (2001). 데이터마이닝 패키지에서 변수선택 편의에 관한 연구, <응용통계연구>, **14**, 475-486.

정성석, 김순영, 임한필 (2004). 의사결정나무에서 분리변수 선택에 관한 연구, <응용통계연구>, **17**, 347-357.

최대우, 구자용, 최용석 (2004). 배경자료를 이용한 나무구조의 군집분석, <응용통계연구>, **17**, 535-545.

허명회, 양경숙 (2005). 연속형 자료에 대한 나무형 군집화, <응용통계연구>, **18**, 661-671.

Berkhin, P. (2002). *Survey of Clustering Data Mining Technique.* Technical Report, Accrue Software.

Boley, D. (1998). Principal directions divisive partitioning, *Data Mining and Knowledge Discovery*, **2**, 325-344.

Ganti, V., Gehrke, J., Ramakrishnan, R. (1999). CACTUS-clustering categorical data using summary, *In Proceedings of the ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, USA. 73-83.

He Z., Xu X., Deng S., and Song Y. (2002). dNumber: A fast clustering algorithm for very large categorical data sets, 1-13. (http://citeseer.ist.psu.edu/ he02dnumber.html).

Lee, Y. M. and Song, M. S. (2002). A study on unbiased methods in constructing classification trees, *The Korean Communications in Statistics*, **9**, 809-824.

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

Loh, W. Y. and Shih,Y. S. (1997). Split selection methods for classification trees, *Statistica Sinica*,**7**, 815-840.

Song, H. I., Song, E. T. and Song, M. S. (2004). A study on the bias reduction in split variable selection in CART, *The Korean Communications in Statistics*, **11**, 553-562.

TomChiu, DongPing Fang, John Chen, Yao Wang, Christopher Jeris. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment, *Proceedings of the seventh ACM SIG KDD international conference on knowledge discovering and data mining.* 263-268.

Zhang, T., Ramakrishnan, R., and Livny, M. (1997). BIRCH: A new data clustering algorithms and its applications, *Data Mining and Knowledge Discovery*, **1**, 141-182.

[2005년 8월 접수, 2006년 3월 채택]

Tree-structured Clustering for Mixed Data*

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

ABSTRACT

The aim of this study is to propose a tree-structured clustering for mixed data. We suggest a scaling method to reduce the variable selection bias among categorical variables. In numerical examples such as credit data, German credit data, we note several differences between tree-structured clustering and K-means clustering.

Keywords: Mixed data, Tree-structured clustering, Node splitting, Variable selection.

* This work was supported by a Korea University Grant.

1) 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@dreamwiz.com

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