응용통계연구 제 17권 3호, 2004년, pp. 535-545

배경자료를 이용한 나무구조의 군집분석 *

최대우 [1] 구자용 [2] 최용석 [3]

요 약

이 논문에서 제안하는 군집분석방법은 분석자료와 동일한 구조의 배경자료를 생성하고 이를 나무모형의 분류기법을 이용하여 분리해 냄으로써 변수들의 규칙으로 정의되는 군집을 형성한다. 배경자료는 reverse-arcing 알고리즘을 통하여 분석자료와 공간상에서 대비되도록 생성되며 군집이 효과적으로 식별되도록 돕는다. 이 방법은 분석자료에 이산형 변수가 혼합된 경우에도 적용할 수 있으며 모의실험자료와 실제 자료를 이용하여 제안된 알고리즘의 성능을 규명하였다.

주요용어: 주요용어 : 군집분석, 나무모형, 배경자료, reverse-arcing

1. 서론

군집분석은 목표 변수 없이 설정된 유사성(similarity)에 기준하여 유사한 자료로 이루어진 군집을 형성하는 다변량 통계분석 방법으로, 군집분석의 사용자는 대량의 자료에서 군집들을 탐색하고 이들의 분석을 통하여 의미 있는 정보를 발굴할 수 있는 것이다. 군집분석의 적용 분야를 설명하면, 시장조사(market research)의 경우 군집분석을 활용하여 수익성 있는 틈새시장(niche market)을 찾아내거나 구매자의 구매 패턴을 알아낼 수 있다. 보험의 경우 군집별로 보험사고 발생 집단의 특성 파악을 통하여 군집별 보험원가의 산출과 다양한 위험의 인수 및 평가기준(underwriting rule)의 설정에 필요한 정보를 제공해 줄 수 있다. 일반적으로 사용되는 군집분석 방법으로는 유사성 측도에 의해 군집 대상을 순차적으로 통합 또는 분리해 나아가는 계층적(hierarchical) 군집방법과 군집의 수를 정하고 해당군집의 중심(centroid)에 개체를 편입시켜 군집을 형성하는 분할적(partitioning) 군집방법으로 구별할 수 있다. 군집 대상은 자료의 개체나 변수가 될 수 있으며 유사성 측도는 개체인 경우 유클리드 거리, 맨하탄 거리 등이 사용될 수 있으며 변수인 경우에는 상관계수를 사용할 수 있다. 계층적 군집방법은 군집대상을 순차적으로 통합하여 나가는 응집분석(agglomerative analysis)과 유사성이 낮은 군집대상을 분리해 나가는 분할분석(divisive analysis)으로 구별하며, 분할적 군집방법으로는 k-means 방법과 k-medoids 방법을 들 수 있다. 군집분석에 대

* 구자용의 연구는 2002년도 인하대학교의 지원에 의하여 연구되었음 (INHA-22433)

1) (449-791) 경기도 용인시 모현면 왕산리 산 89, 한국외국어대학교 정보통계학과, 부교수
E-mail: dachoi@dreamwiz.com
2) (136-701) 서울 성북구 안암동 고려대학교 통계학과, 교수
E-mail: jykoo@korea.ac.kr
3) (449-791) 경기도 용인시 모현면 왕산리 산 89, 한국외국어대학교 수학과 통계전공, 박사과정
E-mail: yschoikdw@dreamwiz.com

한 최근 연구 동향에 대해서는 Berkhin(2002)를, k-medoids 방법에 대한 설명은 Rousseeuw *et al.*(1990)를 참조할 수 있다.

본 논문에서 제안하는 방법은 분석자료와 같은 구조의 가상의 관찰치인 "배경자료"를 분석자료의 분포와 대비되도록 생성하고 이렇게 만들어진 통합된 자료에 이미 알려진 나무모형 분석을 시행하여 분석자료의 군집을 규칙 형태로 도출하는 것이다. 이는 앞서 설명한 고전적 분석방법이 자료간의 유사성에 기반한 방법이라면 제안방법은 자료의 공간상 분포에 기반한 방법이다. 구체적으로 살펴보면, 본 알고리즘에서는 연속형과 이산형 값을 가진 벡터의 개체들로 구성된 분석자료에 같은 자료구조의 배경자료를 분석자료가 충분히 포함하도록 생성한다. 그 다음 이 두 자료의 영역이 겹치는 공간의 자료들을 제거한 후 이들을 하나의 자료로 통합하고 이 자료를 배경자료 또는 분석자료 여부를 종속변수로 하여 나무모형기법을 이용하여 두 자료들을 분류 한다. 통합된 자료에 나무모형이 적용되면 분석자료 또는 배경자료만으로 구성될 수 있도록 그 순수도가 높아지는 방향으로 분할이 이루어진다. 최종적으로 생성된 노드(terminal node)는 분할에 사용된 변수들의 규칙으로 구성되며 다차원 공간상의 박스(box) 형태를 갖게 된다. 분석자료의 주변에 근접하여 배경자료가 위치한다면 자연스럽게 최소크기의 분석자료 공간을 정의하는 박스들을 만들어 낼 수 있게 되는데 이들 박스들은 각각 설명변수들의 규칙들로 정의되는 군집으로 간주될 수 있다.

본 연구에서 제안한 알고리즘은 나무모형(Breiman,1984)과 다수의 분류모형들을 결합하여 그 성능을 향상시키는 방법론인 bagging(Breiman, 1996) 및 arcing(Breiman,1998)을 근간으로 하였다. 특히 reverse-arcing은 arcing을 역으로 착안된 방법으로 분석자료의 군집을 식별하는데 기여할 뿐 아니라 나무모형 적용시 자연스러운 가지치기(pruning) 효과를 나타내어 의미 있는 규모의 군집을 형성하는데 결정적인 역할을 하는 방법으로 기존에 시도되지 않은 새로운 방법이다. 유사한 연구로 Liu *et al.*(2000)는 연속형으로만 구성된 분석자료에 배경자료를 실제 생성하지 않고 나무모형 분석시 불순도(impurity) 계산에만 단순히 그 존재를 가정하여 사용하였다.

본 논문의 구성은 다음과 같다. 제2장에서는 새로운 알고리즘과 그 특성을 기술하였다. 여기서는 배경자료를 생성하고 이 중 효과적인 배경자료만을 걸러내는 과정과 배경자료를 이용하여 분석자료의 군집을 형성하는 방법을 설명한다. 제3장에서는 모의실험자료와 실제자료에 제안된 알고리즘을 적용하고 그 성능을 살펴본다.

2. 제안 알고리즘

본 알고리즘에서는 분석자료와 배경자료가 공간상에서 혼재되지 않고 두 자료의 경계를 뚜렷하게 하기 위하여 arcing의 기본 알고리즘을 역으로 적용하여 분석자료로 오분류되는 배경자료 개체의 출현을 억제하는, 소위 reverse-arcing 방법을 사용하였다. Arcing은 오분류된 관찰치가 보다 자주 표본추출 되도록 추출확률을 조정하여 붓스트랩 표본을 추출하고 각 표본을 이용하여 특정 판별 모형을 생성한 후 예측결과는 가중 평균으로 결정하는 방법이다. 전체적인 알고리즘은 다음의 다섯 단계로 구성된다.

[알고리즘]

- STEP 0: 분석자료의 변수 값 범위를 충분히 포함하는 공간에 균등(uniform)하게 초기 배경자료를 생성한다.
- STEP 1: bagging을 통하여 배경자료의 각 관찰치에 대하여 분석자료로 오분류되는지 여부를 결정한다.
- STEP 2: reverse-arcing을 적용하여 배경자료의 각 관찰치가 재추출될 확률을 계산한다.
- STEP 3: 앞 단계의 추출확률로 배경자료를 재추출 한다.
- STEP 4: 정해진 횟수만큼 STEP 1-3을 반복한다.
- STEP 5: 분석자료와 최종적으로 추출된 배경자료에 나무모형을 이용하여 군집분석을 한다.

2.1. Reverse-arcing 알고리즘을 이용한 배경자료의 생성

초기 배경자료는 분석자료를 충분히 포함할 수 있도록 생성한다. 즉, 분석자료 중 연속형 변수의 경우 최대값 및 최소값을 포함하는 범위 내에서 균등하게(uniform) 자료 값을 생성하고, 이산형 변수의 경우 관측된 수준(level) 값들에서 균등하게 자료를 생성하여 초기 배경자료를 준비한다. 구체적으로 살펴보면, 분석자료의 관찰치가 N개의 연속형 속성과 M개의 이산형 속성으로 구성된 (N+M)×1의 벡터인 경우 배경자료도 같은 구조로 생성한다. 여기에서 연속형 변수에 대해서는 각 변수별로 출현 가능한 값의 범위를 충분히 포함하도록 범위를 설정하고 그 범위 내의 균등하게 무작위로 추출하며 이산형 변수는 각 변수별로 수준값들 중 균등하게 무작위로 추출한다. 이렇게 생성된 배경자료는 분석자료의 전 영역을 커버하기에 충분하도록 생성한다. 그 후 배경자료와 분석자료가 혼재한 영역에서는 reverse-arcing알고리즘을 이용해서 분석자료 영역에 있는 배경자료를 제거하고 대신 배경자료의 영역에서만 원래의 배경자료 수 만큼을 초기 배경자료에서 재추출 하게 된다. reverse-arcing은 다음과 같은 알고리즘을 적용한다.

[알고리즘]

- STEP 0: 초기 배경자료 R을 생성한 후 각 관찰치에 종속변수로 R에는 Y=0을, 분석자료 T에는 Y=1을 부여한다.
- STEP 1 : R의 각 관찰치가 동일한 확률로 추출 될 수 있도록 추출확률벡터 P_0를 다음과 같이 구성한다. P_0=$(p_0(1), p_0(2), \ldots, p_0(N_R))$, $p_0(i)$=$1/N_R$ (N_R은 배경자료의 수)

- STEP 2 : 각 k번째 반복에서의 추출확률벡터 $P_k=(p_k(1), p_k(2), \ldots, p_k(N_R))$를 이용하여 배경자료 R로부터 새로운 배경자료 R_k를 생성하고 T와 R_k를 통합한 S_k에 대하여 bagging을 적용하여 분류모형 C_k를 생성한다.
- STEP 3 : C_k가 R의 i 번째 자료에 대하여 정분류 되어있으면 $d(i)$=0을, 오분류 되어 있으면 $d(i)$=1을 부여한다.
- STEP 4: $k+1$번째 추출 확률벡터 P_{k+1}의 i 번째 자료추출확률 $p_{k+1}(i)$는 $p_{k+1}(i)=p_k(i)\beta_k^{d(i)q}/\sum_i p_k(i)\beta_k^{d(i)q}$로 계산되며 $\varepsilon_k=\sum p_k(i)d(i)$이고, $\beta_k=\varepsilon_k/(1-\varepsilon_k)$, 그리고 q는 고정된 양수이다.
- STEP 5: 앞의 STEP 2-4 과정을 일정횟수 반복한다.

Arcing과의 차이점은 β_k 를 역수로 하여 $\beta_k=\varepsilon_k/(1-\varepsilon_k)$를 적용한 부분이다. 또한 생성된 각 분류자를 $log(\beta_k)$로 가중평균하지 않고 다만 재추출 확률을 산출하는 과정으로만 사용하였다. 위의 알고리즘 STEP4 에서 q는 초기 배경자료중 오분류되는 관찰치가 새로운 배경자료로 추출될 가능성의 속도를 조절하는 역할을 한다. 이 과정을 거치게 되면 배경자료는 분석자료의 (N+M)차원 공간상 주변에만 위치하게 된다. 이러한 알고리즘을 적용하게 되면 연속형과 이산형의 경우 다음과 같은 결과를 얻게 된다. 그림 2.1은 Hastie *et al.* (2001)의 figure 14.3을 참조하여 생성한 것으로 2개의 이변량 정규분포 $N_2\left(\begin{bmatrix}0.9\\0\end{bmatrix}, \begin{bmatrix}0.7 & -0.65\\-0.65 & 2\end{bmatrix}\right)$와 $N_2\left(\begin{bmatrix}0.9\\0\end{bmatrix}, \begin{bmatrix}0.7 & 0.65\\0.65 & 2\end{bmatrix}\right)$ 로 부터 각각 200개씩 생성한 분석자료 (■) 400개와 이 자료에 대한 초기배경자료(□) 400개의 산포도를 보여준다.

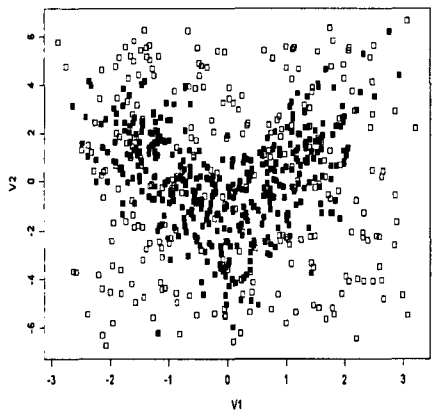

그림 2.1: 초기배경자료와 분석자료

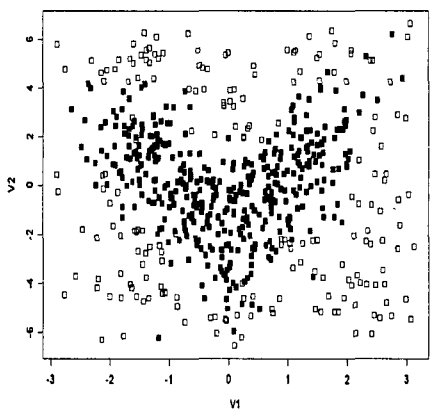

그림 2.2: reverse-arcing 10회 적용후

이 자료에 reverse-arcing 알고리즘을 10회 적용한 결과 그림 2.2와 같이 분석자료의 영역에서 초기 배경자료의 일부가 사라지고 분석자료의 주변에 배경자료 추출이 많아지는 것

을 볼 수 있다. 즉 분석자료의 영역에 있는 배경자료의 재추출 확률이 reverse-arcing 과정이 반복될수록 재추출 확률이 현저히 작아져서 결국 분석자료가 밀집한 부분에 위치한 초기 배경자료 개체는 더 이상 추출되지 않기 때문이다.

이산형의 경우, 존재하는 수준값 내에서 초기 배경자료가 생성되므로 연속형 경우와는 달리 모든 수준 값이 분석자료와 겹치게 된다. 하지만 reverse-arcing 알고리즘을 적용하게 되면 배경자료의 수준별 비율은 분석자료의 역의 형태로 나타난다. 그림 2.3과 같이 분석자료에 초기 배경자료를 생성 후 reverse-arcing 알고리즘을 적용하게 되면 그림 2.4와 같은 분포의 배경자료가 형성된다. 즉, 분석자료의 비율이 높은 수준에서는 낮은 비율의 배경자료가 생성되고 그 반대의 경우 높은 비율의 배경자료가 생성되는 현상이 나타난다.

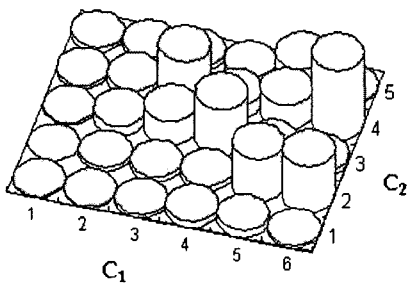

그림 2.3: 분석자료

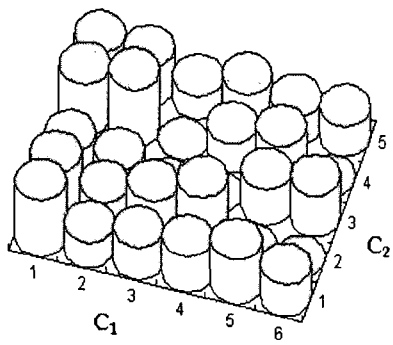

그림 2.4: reverse-arcing 10회 적용후

2.2. 나무모형을 이용한 군집 형성

Reverse-arcing알고리즘 적용 후 최종 생성된 배경자료와 분석자료를 통합하고 두 자료를 분류하기 위하여 나무모형을 적용하게 되면 각 최종 노드는 minimal cost-complexity 에 의해 불순도가 최소화되도록 형성되어 배경자료 또는 분석자료로 분류되는 최종 노드를 형성한다. 분석결과에서 분석자료로 분류하는 말단노드를 군집으로 간주한다. 나무모형 분석은 별도의 가지치기를 하지 않으나 reverse-arcing은 자연스럽게 가지치기(pruning) 효과를 나타내었다. Reverse- arcing 적용 횟수를 늘리면 분석자료 내의 배경자료가 사라지는 것 뿐 아니라 분석자료 주변에 나타나는 배경자료가 조금씩 후퇴하게 되어 나무모형 분석시 분석자료에 대한 변별력이 높아져 생성될 군집의 크기가 커지게 된다. 따라서 자연스러운 나무모형의 가지치기 결과를 얻게 되는 것이다. 아래의 그림 2.5 및 그림 2.6은 앞의 그림 2.1 및 그림 2.2에서 사용한 자료에 reverse-arcing 알고리즘 적용 후 나무모형 분석을 실시한 결과를 S-PLUS©의 partition plot을 이용하여 도식화한 것이다. 그림 중, 분석자료는 "0"으로 배경자료는 "r"로 식별 표기하였다. 초기 배경자료와 분석자료를 합하여 나무모형을 적용한 결과인 그림 2.5와 reverse-arcing 30회 적용후의 분석결과인 그림 2.6을 비교해보면 위에서 언급한 가지치기 효과가 존재함을 확연히 알 수 있다.

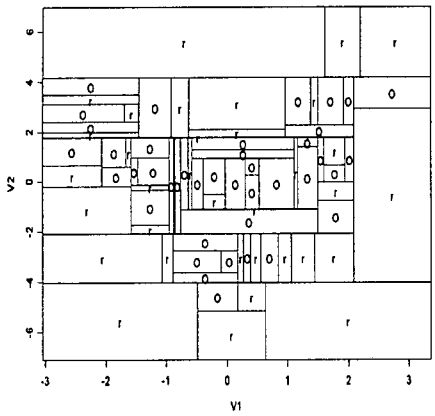

그림 2.5: reverse-arcing 적용전

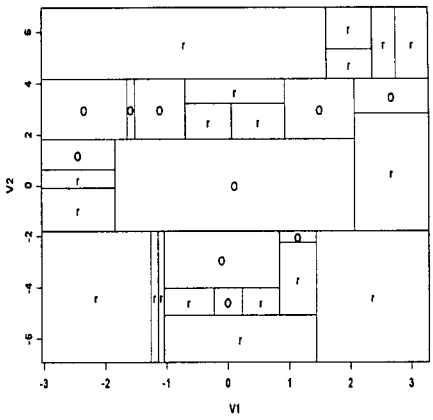

그림 2.6: reverse-arcing 적용후

3. 예제분석

본 절에서는 앞서 제안한 알고리즘이 실제 자료에서 어떠한 결과를 산출하는지 연속형 자료만으로 구성된 실제 자료와 연속형과 이산형이 혼합된 모의실험 자료를 사용하여 살펴보았다. 군집방법의 효율성을 평가하기 위하여 이미 분류 결과가 있는 자료에 군집분석을 시행한 후 각 군집에 포함된 개체가 얼마만큼 동질적으로 군집에 포함되었는지 비교하는 방법을 사용하였다.

3.1. 모의실험 자료분석

이산형과 연속형 자료가 혼합된 자료에서 본 알고리즘의 효과를 평가하기 위하여 다음과 같이 모의자료를 생성하였다. 이 분석자료는 0과 1사이에서 균일하게 생성된 변수 $X1$, $X2$와 이산형 변수 $C1$,$C2$의 4개 변수로 구성되어 있다.

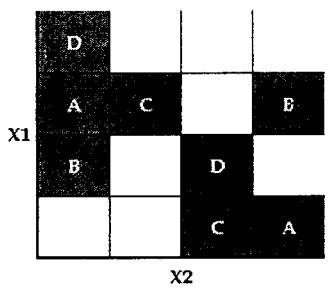

그림 3.1: $X1$,$X2$ 생성위치

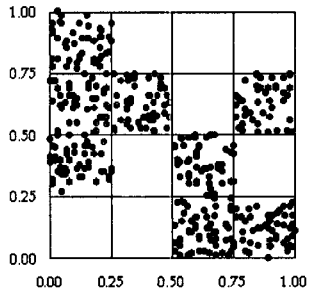

그림 3.2: $X1$,$X2$ 생성결과

	C1	C2	계
D	1	1	97
C	1	2	114
B	2	1	96
A	2	2	93

그림 3.3: $C1$,$C2$ 부여내역

표 3.1: 모의자료 분석결과

[배경자료 800개 reverse-arcing 20회]

군집	구성비	군집규칙	A	B	C	D
1	26.0%	[0.285<X1<0.75812]∩[-0.00401<X2<0.25208]∩[C1=1]	47	57	0	0
2	14.8%	[X1>0.74597]∩[-0.00401<X2<0.25208]∩[C2=2]∩[C1=2]	0	0	0	59
3	13.0%	[-0.00037<X1<0.25145]∩[X2>0.49956]∩[C2=1]	27	0	25	0
4	7.8%	[0.49593<X1<0.72041]∩[X2>0.80862]∩[C2=2]	0	31	0	0
5	7.3%	[0.50676<X1<0.72041]∩[0.33470<X2<0.49956]∩[C1=2]	0	0	29	0
6	3.5%	[0.17848<X1<0.21006]∩[X2>0.49956]∩[C2=1]	5	0	9	0
7	3.5%	[0.25145<X1<0.72041]∩[0.51803<X2<0.56412]∩[C2=2]	0	0	0	14
8	3.3%	[0.56412<X2<0.715]∩[0.25145<X1<0.39851]∩[C1=2]∩[C2=2]	0	0	0	13
9	3.0%	[0.25208<X2<0.33470]∩[0.50676<X1<0.72041]∩[C1=2]	0	0	12	0
10	3.0%	[0.46002<X1<0.72041]∩[0.75818<X2<0.80862]∩[C2=2]∩[C1=1]	0	12	0	0

[배경자료 1600개 reverse-arcing 20회]

군집	구성비	군집규칙	A	B	C	D
1	15.3%	[0.27970<X1<0.50337]∩[-0.00021<X2<0.25038]∩[C2=2]∩[C1=1]	0	61	0	0
2	14.8%	[X1>0.75248]∩[-0.00021<X2<0.25038]∩[C2=2]∩[C1=2]	0	0	0	59
3	11.8%	[0.48583<X1<0.75432]∩[-0.00021<X2<0.25038]∩[C2=1]∩[C1=1]	47	0	0	0
4	10.8%	[0.50275<X1<0.72024>]∩[X2>0.75343]∩[C2=2]∩[C1=1]	0	43	0	0
5	10.3%	[0.515<X2<0.75059]∩[0.25942<X1<0.495]∩[C1=2]∩[C2=2]	0	0	0	41
6	6.8%	[0.25038<X2<0.43926]∩[0.51962<X1<0.72024]∩[C1=2]∩[C2=1]	0	0	27	0
7	6.8%	[0.545<X2<0.81331]∩[0.08751<X1<0.72024]∩[C2=1]	4	0	23	0
8	5.5%	[0.18177<X1<0.26930]∩[0.43926<X2<0.75059]∩[C2=1]∩[C1=2]	0	0	22	0
9	3.8%	[0.00943<X1<0.05120]∩[X2>0.43926]∩[C2=1]	11	0	4	0
10	3.5%	[X2>0.81331]∩[0.08751<X1<0.18177>]∩[C1=1]]∩[C2=1]	14	0	0	0

구체적인 자료 생성방법은 다음과 같다. 그림 3.1에 정한 A,B,C,D 구간에서 그림 3.2와 같이 균일하게 난수 400개를 생성하고 각 구간에 상응하는 $C1$, $C2$를 그림 3.3의 기준으로 부여하였다. 분석에서는 배경자료를 800개 및 1600개 생성하고 reverse-arcing 20회 실시하였으며 그 결과 표 3.1의 결과를 얻었다.

배경자료 800개의 경우 그림 3.1의 좌측 A,B와 하단의 A,C가 구분되지 않은 것으로 나타났다. 분석 배경자료 1600개의 경우에는 그림 3.4에서 확인되는 바와 같이 그림 3.1 하단의 C,A를 제외하고는 대부분의 경우 뚜렷하게 판별해내고 있다. C,A 부분은 군집 9,10에서 보다 작은 군집형태로 추가적으로 판별해내고 있다.

이상과 같이 본 제안 알고리즘은 혼합형 자료에서 적절한 결과를 산출할 수 있음을 보여주고 있다.

3.2. Iris 자료 분석

연속형 자료 분석의 예로 R.A. Fisher의 Iris 자료를 사용하였다. 이 자료는 붓꽃(iris)의 세 부 종인 Setosa(ST), Versicolor(VC), Virginica(VG)의 각각 50개체에 대해 Sepal length(sl), Sepal.width(sw), Petal.length(pl), Petal.width(pw)를 측정한 관찰치로 구성되어있다. 분석을 위하여 배경자료는 300개와 600개를 각각 생성하였다. 분석결과 표 3.2와 같이 규칙들에 의해서 군집을 정의할 수 있게 된다.

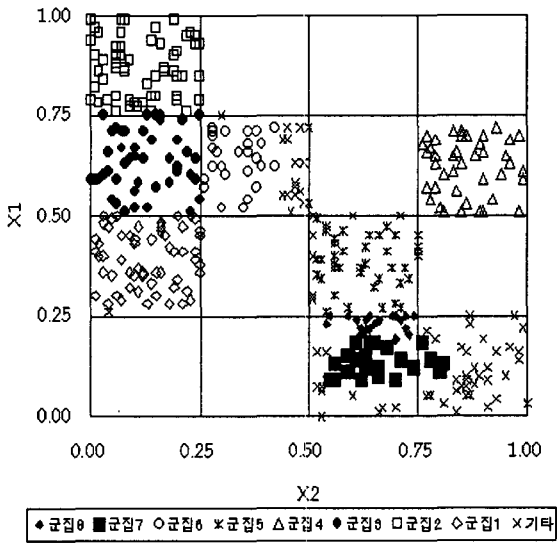

그림 3.4: 배경자료 1600개 경우 군집 도해

표 3.2: Iris 자료에 대한 분석결과

[배경자료 300개 reverse-arcing 20회]

군집	구성비	군집규칙	ST	VC	VG
1	33.3%	[4.2804<sl<6.1499]∩[sw>2.1998]∩[pw<0.5149]	50	0	0
2	6.0%	[sl>4.2804]∩[2.1998<sw<2.8117]∩[pl>1.9505]∩[0.9940<pw<1.15]	0	9	0
3	2.7%	[sl>4.2804]∩[2.1998<sw<2.8117]∩[pl>1.9505]∩[1.15<pw<1.2960]	0	4	0
4	2.7%	[sl>4.2804]∩[2.4269<sw<3.8013]∩[pw>1.2960]∩[3.9768<pl<4.15]	0	4	0
5	50.0%	[sl>4.2804]∩[2.4269<sw<3.8013]∩[pl>4.15]∩[pw>1.2960]	0	26	49

[배경자료 600개 reverse-arcing 20회]

군집	구성비	군집규칙	ST	VC	VG
1	2.7%	[4.7514<sl<6.2987]∩[1.9887<sw<2.29803>]∩[0.9829<pw<1.5168]	0	3	1
2	22.0%	[sl>4.3829]∩[sw>2.2980]∩[pl<2.0595]∩[pw<0.2170]	33	0	0
3	9.3%	[4.3829<sl<5.9043]∩[sw>2.2980]∩[pl<3.8964]∩[0.2170<pw<0.4014]	14	0	0
4	4.0%	[sl>4.3829]∩[2.2980<sw<2.6002]∩[2.7748<pl<3.8964]∩[0.4014<pw<1.1964]	0	6	0
5	4.0%	[4.3829<sl<5.5744]∩[2.2980<sw<3.1959]∩ [3.8964<pl<4.6276]∩[pw>1.1894]	0	5	1
6	34.7%	[5.5744<sl<6.9061]∩[sw>2.2980]∩[3.8964<pl<5.4120]∩[pw>1.1894]	0	32	20
7	2.7%	[5.5744< sl<6.3905]∩[sw>2.29803]∩[5.4120<pl<6.03231]∩[pw>1.3733]	0	0	4
8	12.0%	[sl>6.3905]∩[sw>2.2980∩ sw<3.25]∩[pl>5.4120]∩[pw>1.7]	0	0	18
9	3.3%	[3.25<sw<3.8220]∩[sl>6.3905]∩[pl>5.4120]∩[pw>1.7]	0	0	5

표 3.3: Iris 자료에 대한 k-means 분석 결과

k=2	군집1	군집2				전체
Setosa	50	0				50
Versicolor	3	47				50
Virginica	0	50				50
계	50	97				150
k=3	군집1	군집2	군집3			계
Setosa	0	50	0			50
Versicolor	2	0	48			50
Virginica	36	0	14			50
계	38	50	62			150
k=5	군집1	군집2	군집3	군집4	군집5	계
Setosa	0	0	0	28	22	50
Versicolor	23	27	0	0	0	0
Virginica	22	1	27	0	0	50
계	45	28	27	28	22	150

배경자료 300개를 사용한 경우 Setosa 전부는 하나의 군집을 이루었다. 군집-2와 3은 3개 변수(sl, sw와 pl)의 범위가 일치하여 형성된 군집이 거의 유사하다 할 수 있으므로 13개의 Versicola 13개체로 이루어졌다고 할 수 있다. 군집-5의 경우 26개의 Versicolor를 포함하고 있으나 Virginica가 49개로 군집 내 개체 전체의 65%를 차지하고 있다. k-means 군집분석의 결과인 표 3.3과 비교할 때, Versicolor와 Virginaca가 동일한 군집에 섞여있어 두 방법다 이 종류들을 구별하여 군집을 형성하는데 어려움이 있음을 보여주고 있다. 배경자료가 600개인 경우, 군집-2, 3은 Setosa만으로 구성되나 각 군집의 해당 변수 범위가 서로 연결되어 있어 하나의 군집으로 통합할 수 있을 것이다. 군집-7, 8, 9도 각 변수 범위의 교집합을 정리하는 경우 (5.5744<sl<6.3905)∩(2.2980<sw<3.8220)∩(5.4120<pl<6.0323)∩(1.3733<pw)인 군집으로 표현되며 27개의 Virginica만으로 구성된다. 여기서 군집-7,8,9는 k=5일때 k-means결과 (표 3.3참조) 군집-3과 그 구성개체가 비슷한 것으로(85%이상이 동일한 개체임) 나타났다

분석 결과 측면에서 두 방법중 특별한 비교우위를 가지고 있다고 말하기 어렵지만 k-means의 경우 군집개체의 공간상 위치나 그 속성의 특성을 알기 위해서는 별도의 분석이 필요한 반면 본 알고리즘에서는 속성을 규칙으로 표기하여줌으로써 군집간 특성을 파악할 수 있게 해준다.

4. 결 론

기존의 군집분석이 거리 개념에 기반한 방법이라면 본 논문에서 제안한 알고리즘은 분석자료의 공간상 분포 형태 및 밀도에 기반한 군집방법이라 할 수 있다. 본 연구의 의의를 정리해 본다면, 첫째, 군집분석에 reverse-arcing이라는 새로운 개념의 배경자료 생성 알고

리즘과 나무모형을 통하여 분석자료의 군집을 공간상에서 찾아내고 형성된 군집은 각 변수별 규칙으로 표현된다는 점과, 두 번째로는 변수가 이산형, 연속형 또는 두 속성이 혼재한 경우에도 군집을 형성할 수 있다는 것이다. 셋째, 분석자료와 배경자료의 경계선 상에 있어 모호한 위치에 있는 분석자료의 일부 관찰치는 배경자료의 군집으로 분류되어 군집형성에서 제외되고 보다 명확히 분류되는 자료만을 군집으로 형성하고 있다는 것이다. 본 알고리즘이 제공하는 군집이 박스 형태로 형성되는 점을 고려해 볼 때 Friedman *et al.* (1998)의 interbox dissimilarity 와 같은 비유사성을 이용한 군집간 통합도 추가 연구의 여지가 있다.

참고문헌

Berkhin, P. (2002). *Survey of Clustering Data Mining Techniques*, Accrue Software.

Breiman, L.(1996). Bagging predictors, *Machine Learning*, **24**, 123-140.

Breiman, L.(1998). Arcing classifiers, *Annals of Statistics*, **26**, 801-849.

Breiman, L., Friedman, J. H., Olshen R. A., Stone, C. J.(1984). *Classification and Regression Tree*, Champman & Hall, New York.

Friedman, J. H. and Fisher, N. I. (1999). Bump hunting in high-dimensional data, *Statistics and Computing*, **9**, 123-143.

Hastie, T. and Tibshirani, R. and Friedman, J.H.(2001). *The Elements of Statistical Learning*, Springer, New York.

Johnson, R. A. and Wichern, D. W.(1992). *Applied Multivariate Statistical Analysis*, Prentice Hall, Englewood Cliffs.

Kaufman, L., Rousseeuw, P. J. (1990). *Finding Groups in Data*, Wiley Interscience, New York.

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

[2004년 2월 접수, 2004년 6월 채택]

Tree Based Cluster Analysis Using Reference Data*

Dae Woo Choi [1] Ja Yong Koo [2] Yong Seok Choi [3]

ABSTRACT

The clustering method suggested in this paper produces clusters based on the "rules of variables" by merging the "training" and the identically structured reference data and then by filtering it to obtain the clusters of the "training data" through the use of the "tree classification model". The reference dataset is generated by spatially contrasting it to the "training data" through the "reverse arcing" algorithm to effectively identify the clusters. The strength of this method is that it can be applied even to the mixture of continuous and discrete types of "training data" and the performance of this algorithm is illustrated by applying it to the simulated data as well as to the actual data.

Keywords: Cluster analysis; Tree model; Reference data; Reverse-arcing

* Research of Ja-Yong Koo was supported by INHA UNIVERSITY Research Grant. (INHA-22433)

1) Professor, Department of Statistics, Hankook University of Foreign Studies, 89 Wangsanri Mohyunmyon, Yonginsi, Kyongkido, 449-791, Korea.
E-mail: dachoi@dreamwiz.com

2) Professor, Department of Statistics, Korea University, Anam-Dong Sungbuk-Ku, Seoul 136-701, Korea.
E-mail: jykoo@korea.ac.kr

3) Graduate Student, Department of Mathematics, Hankook University of Foreign Studies, 89 Wangsanri Mohyunmyon, Yonginsi, Kyongkido 449-791, Korea.
E-mail: yschoikdw@dreamwiz.com