응용통계연구 제 8권 1호
1995년 3월 pp. 39-50.

쿨롱네트워크를 이용한 집락분석[1)]

이 석 훈[2)], 박 래 현[2)], 김 응 환[3)]

요 약

기존의 집락분석은 집락화만을 목적으로 하기 때문에 분석이 끝나면 집락분석에 사용된 규칙을 보존하지 못하는 문제를 갖고 있다. 이러한 문제를 인간의 뇌의 성질을 연구하는 신경회로망 분야에서 사용하는 모형중 하나인 쿨롱 에너지 네트워크 모형을 변형 발전 시켜서 해결하여 보았다. 이 모형을 이용한 분석의 실제 예를 보이고 기존의 기법들과의 비교를 통하여 거의 유사한 집락형성을 보여주고 있음을 보였다.

1. 서 론

우리가 관심을 가지고 있는 대상들을 어떤 관점에서 분류한다는 것은 인간의 판단과 인식 체계에서 나타나는 가장 근본적이며 자연적인 현상 중의 하나이다. 이러한 분류에 관한 연구는 탐사적 자료분석을 필요로 하는 모든 과학 분야에서 일반적으로 또는 개별적으로 수행되고 있으며, 따라서 다양한 이름으로 그 연구 분야가 불리어지고 있다. 대표적인 것으로는 패턴인식(pattern recognition), 분류학(classification), 결정분석(decision analysis)의 이름이 공학, 생물학, 경영학, 심리학, 의학 등에서 사용되고 있다. 통계학 분야에서는 주로 집락분석이라는 이름으로 제 분야에서 하고 있는 연구들의 기본적인 관점들에 대하여 토의를 하고 있는데 집락의 대상이 되는 개체들 간의 유사성 측정 방법들과, 개체들의 집락화 방법, 집단 내의 집락의 갯수에 관한 연구가 꾸준히 수행되고 있다.

Jain과 Dubes(1988)가 아주 다양한 분야에서 그 분야의 특성에 따라 독자적인 용어로 집락분석을 연구하는 전문가들이 서로 교통할 수 있도록 하기 위하여 책을 쓴다고 할 정도로 집락분석은 학제간(interdisciplinary)의 연구 영역이라고 할 수 있다. 그들은 집락분석을 일반적으로 잘 알려진 바와 같이 탐사적 자료 분석의 한 도구로서 자료를 효과적으로 조직화하면서 자료 안에 내재된 구조를 발견하는 기법과 방법론을 연구하는 것으로 정의하고, 개체들 자신이 소속되는 집락에 대한 사전 정보가 없다는 점에서 판별분석과 구별하고 있다. 한편, 그들은 사람이 야말로 탁월한 집락 발견자임을 인정하면서 집락분석은 컴퓨터 혁명의 산물로 보고, 따라서 집락분석 기법의 장점을 수작업(manual grouping process)과 비교하여 결과의 일관성(consistency), 신속성(speed) 그리고 신뢰성(reliability)에서 찾고 있다.

이와 같은 Jain과 Dubes의 견해에 전반적으로 동의하면서 본 연구자들은 소위 탁월한 집락 발견자인 사람은 집락화를 하면서 동시에 집락화시킨 후에 자신의 분류 기준이 기억되고 있는

1) 이 논문은 1993년도 학술진흥재단의 자유공모과제 연구비에 의하여 연구되었음.
2) (305-764) 대전시 유성구 궁동 220, 충남대학교 통계학과.
3) (301-130) 대전시 중구 문화동, 충남기계공업고등학교.

반면에, 통계학에서 주로 사용되는 기존 집락분석 기법들은 집락 작업이 끝나면 그 집락 과정에 사용된 기준을 보존할 수 없다는 점을 주목하였다. 집락화 후에 분류기준을 제시하는 데에 관한 연구가 많지 않은 것은 몇가지 이유가 있다고 본다. 그중 하나는 집락분석의 대상이 되는 일반적인 상황이 집락화 기준을 요구하지 않는다고 생각하였기 때문이라고 본다. 그러나 현실적으로는 그렇지 않은 경우도 많다고 생각된다. 예를들면 정신병의 분류같은 분야를 엄밀한 의미에서 보면 집락분석의 결과에서 나타나는 집락을 명명(naming)하여 그 기준을 갖고 환자를 구별하고 있다고 말할 수 있다. 사실 이 논리를 역으로 생각하면 집락화 기준이 제시가 되지 않았었기 때문에 필요로 되는 상황을 찾지 못하였다고도 말할 수 있다고 본다. 다른 하나는 역시 Jain과 Dubes가 말한 바와 같이 컴퓨터 혁명의 산물로서의 한계라고 생각해 볼 수가 있겠다. 따라서 오늘날과 같이 "컴퓨터가 인간의 상식적인 판단능력까지 구현할 수 있는가?"라는 가능성을 향한 연구와 개발이 진행되는 상황에서는 지금까지 받아들여진 집락분석의 정의 또한 컴퓨터 과학의 발전과 함께 새롭게 변화해야 한다고 판단하고, 일반적으로 사람의 뇌가 주어진 자료를 집락화시키는 과정에 대한 하나의 모형과 그 기법을 개발, 제안하고 기존의 다른 집락분석과의 비교를 통하여 특성을 살펴 보는 연구를 수행하게 되었다.

이러한 연구의 흐름은 통계학의 분류(classification)를 신경회로망의 관점에서 토의한 Lippmann(1987)과 Carpenter와 Grossberg(1986), Kohonen(1987), Kosko(1990), Kim(1990), 이석훈 등(1991, 1992), 김응환 (1992a, b, c, d)에서 찾아볼 수 있다.

본 논문의 구성은 2절에서는 본 연구에서 제안하는 모형의 기본 모형인 쿨롱에너지 네트워크 모형에 대하여 검토하고, 3절에서는 쿨롱에너지 네트워크 모형을 변형하여 얻은 집락분석 모형에 대한 토의를 한다. 그리고 4절에서는 집락분석의 절차 및 실례를 제시하고, 5절에서는 다른 기존의 군집분석 방법들과의 비교를 통하여 제안한 분석기법의 특성을 파악하도록 하며, 6절에서 결론을 내리도록 한다.

2 쿨롱에너지 네트워크와 배경

2.1 모형수립의 직관

쿨롱 포텐셜에너지(Coulomb Potential Energy)는 여러 개의 전하(charge)로 형성되는 장(field)에서 결정되는 에너지를 의미한다. 이 에너지는 동류의 부호를 가진 전하는 서로 밀고, 서로 다른 부호를 가진 전하는 서로 잡아당기는 성질로 만들어지는 것이다. Dembo와 Zeitouni(1987)는 M개의 전하가 놓인 시스템을 생각하였다. 그들은 전하 Q_j를 N차원 공간 X_j에 놓았다고 하고, 시험전하(test charge)가 쿨롱 에너지 네트워크 함수인

$$\zeta = -\frac{1}{L}\sum_{j=1}^{M} Q_j|\mu - x_j|^{-L} \tag{2.1}$$

을 가장 낮은 값으로 만드는 곳으로 N차원 공간에서 이동하는 현상을 조사하였는데, 이때 시험전하는 전하의 수 M에 관계없이 적절한 L에 관하여 가장 가까운 이(異)부호 전하로 끌려가는 것을 보임으로써 함수 (2.1)이 N차원 공간의 $X_1,\ldots,X_M$ 각각을 M개의 끌림의 영역(basins of attraction)을 결정하는 고정점(fixed point)들로 해석한 바 있다.

Scofield(1988)는 이 생각을 좀 더 확장하기 위하여 먼저 식 (2.1)에서 시험전하(test charge)가 없으며 M개의 전하가 N차원 공간 $X_1,...,X_M$에 위치할 때의 정전 포텐셜 에너지(electrostatic potential energy)가 다음과 같은 것을 주목하였다.

$$\Psi = \frac{1}{2L}\sum_{i=1}^{M}\sum_{j=1}^{M} Q_i\, Q_j\, |X_i - X_j|^{-L} \qquad (2.2)$$

정전 포텐셜 에너지 함수를 이런 시각으로 바라본 Scofield는 전하가 놓인 위치 X_i를 자신이 소속된 집락을 알고 있는 개체 i를 할당한 점으로 정의하고 Q_iQ_j를 개체 i와 개체 j가 같은 집단에 속하게 되었으면 -1 그렇지 않으면 +1로 정의하여 식 (2.2)를 최소화시키는 $X_1,...,X_M$을 생각하였다. 이 생각을 통하여 같은 집락에 소속된 개체들은 N차원 공간의 가능한 가까운 점으로, 서로 다른 집락에 소속된 개체들은 N차원 공간에서 가능한 서로 멀리 떨어진 점으로 할당되도록 하는 직관을 반영하는 모형을 제시하게 되었다.

구체적인 모형에서 그는 개체 i를 K차원 벡터 $f_i=(f_{1i},..., f_{ki})$로 나타낼 때 개체 i에 할당되는 N차원 점Xi를 다음과 같이 가중치 $\omega_n=(\omega_{n1},...,\omega_{nk})$ $(n=1,...,N)$를 사용하여 비선형으로 정의하였다.

$$X_i = \sum_{n=1}^{N} e_n\, F_n\Big(\sum_{m=1}^{k}\omega_{nm}\cdot f_{mi}\Big) \qquad (2.3)$$

$$F_n\Big(\sum_{m=1}^{k}\omega_{nm}\cdot f_{mi}\Big) = \Big[1 + \exp\Big\{\Big(-\frac{1}{CT}\Big)\Big(\sum_{m=1}^{k}\omega_{nm}\cdot f_{mi}+\theta\Big)\Big\}\Big]^{-1} \qquad (2.4)$$

여기서 e_n, $n=1, ..., N$은 N차원 단위벡터이다.

이때 특이한 점은 (2.4)에서 함수 F_n을 로지스틱 함수로 가정하는 것인데 이는 퍼셉트론(perceptron)을 구현할 수 있으면서 동시에 미분가능한 단조증가함수의 특성 때문에 신경회로망을 이용한 패턴인식 등에 자주 사용되고 있고, C와 T는 시그모이드(sigmoid)함수의 기울기 정도를 조절하는 모수(parameter)이고, θ는 threshold로서 선형결합에서 상수항과 같은 작용을 하게 된다.

다시한번 요약하면 Scofield(1988)의 모형에서 목적함수인 (2.2)를 최소화한다는 것은 K차원 공간에서 서로 같은 집락에 속하는 개체들을 변환된 N차원 공간에서 서로 가까운 위치로 잡아당겨 놓을 수 있고, 서로 다른 집락에 속하는 개체들은 변환된 N차원 공간에서는 서로 멀리 떨어진 위치로 밀어 놓을 수 있도록 하고자 하는 직관을 잘 구현시키는 가중치 ω_{nm} $(n=1,....,N,m=1,...,k)$를 구하는 것으로 받아들일 수 있다.

2.2 목적함수의 최소화와 근사적 방법

목적함수(2.2)의 최소화는 가중치 ω_{nm}(n=1....N, m=1...k)에 관하여 위치에너지 Ψ의 gradient를 계산하여서 수행한다. 요약하면 이 값은 다음과 같이 주어진다.

$$\delta\omega_{nm} = -\left(\frac{\partial\Psi}{\partial\omega_{nm}}\right)\cdot\eta = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} Q_i\, Q_j\, |R_{ij}|^{-(L+2)}\Delta_{nm}(f_i,\, f_j)\cdot\eta \tag{2.5}$$

여기서 $R_{ij} = X_i - X_j$ 이고,

$\partial X_i/\partial\omega_{nm} = e_n F_n(f_i)(1-F_n(f_i))f_{mi}$이므로

$$\Delta_{nm}(f_i,\, f_j) = R_{ij}\frac{\partial}{\partial\omega_{nm}}R_{ij} = [F_n(f_i)-F_n(f_j)]\cdot[F_n(f_i)(1-F_n(f_i))f_{mi}-F_n(f_j)(1-F_n(f_j))f_{mj}]$$

이다.

이때 η는 변환정도를 제한하는 모수로써 상수이거나 또는 역동적으로 변하는 값을 가질 수 있다. 그런데 여기서 $\delta\omega_{nm}$는 모든 가능한 i, j에 대하여 계산한 것에 대한 합으로 구하여지기 때문에 이는 상당한 양의 계산을 요하게 된다. 이에 scofield는 임의의 두 개체 i, j에 대하여

$$\delta\omega_{nm} = (+/-)|X_i - X_j|^{-(L+2)}\,\Delta_{nm}(f_i,\, f_j)\cdot\eta$$

로서 ω_{nm}의 변동량을 결정하여 변화시키고, 이러한 임의의 두 개체를 가능한 한 많은 횟수 추출하도록 하는 방법을 근사적으로 사용하였다. 여기서 양의 부호는 개체 i와 j가 같은 집락에 속한 경우이고 음의 부호는 다른 집락에 속한 경우이다.

3. 집락분석을 위한 CEN모형의 확장

2절에서 scofield가 제안한 모형은 K차원의 한 점으로 표현되는 각 개체가 자신이 속하는 집락을 알고 있는 경우에 이들이 N차원 공간에서 동일 집락의 개체들은 서로 모이고 집락과 집락은 서로 멀리 떨어지는 현상이 나타나도록 하는 직관을 반영하는 것이었다. 본 절에서는 이 모형을 각 개체가 자신의 소속된 집락을 모르는 경우에 사용할 수 있도록 변형확장을 한다.

이 확장의 시도는 우리 인간이 두 눈을 통하여 사물을 관찰 분류하는 경우에 어떤 주관적 척도를 기준으로 일단 가까운 것들은 더 가깝게 생각하고 먼 것은 더 멀게 분류하는 경향을 모방하는 데에서 출발하였다.

이러한 동기로 본 연구에서 개발한 쿨롱집락분석방법에서는 자신이 속한 집락을 전혀 모르는 개체들로 구성된 집단을 몇 개의 집락으로 분류하고자 할 때 우선 가깝고 멀고를 결정하는 주관적인 임계값을 의미하는 모수와 가까운 정도, 먼 정도를 끌어당기고 밀어내는 강도로 변환시키기 위한 모수를 사용하여 다음과 같은 목적함수를 제시하였다.

$$\Psi = \frac{1}{2L}\sum_{i=1}^{M}\sum_{j=1}^{M} Q(d_{ij})\,\|x_i - x_j\|^{-L} \tag{3.1}$$

여기서,

$$Q(d_{ij}) = \begin{cases} -\left(\frac{1}{\beta}\right)^{\alpha}(\beta - d_{ij})^{\alpha}, & d_{ij}\le\beta \\ \left(\frac{1}{\beta-1}\right)^{\alpha}(\beta - d_{ij})^{\alpha}, & d_{ij}>\beta \end{cases}$$

위 (3.1)식에서 d_{ij}는 두 개체 i, j 사이의 상대적인 거리($0 \le d_{ij} \le 1$)를 의미한다. 또한 $\alpha(0 \le \alpha \le 1)$는 두 개체를 서로 당기고 미는 정도를 조절하는 상수이며, $\beta(0 \le \beta \le 1)$은 동일군집의 기준인 임계값 상수이다. 이 때 함수 $Q(d_{ij})$는 두 개체를 서로 당기고, 밀어내는 정도를 설명하며 +1과 -1 사이의 실수값을 가지는데, 이것은 상수 α와 β의 값의 선택에 의존하게 된다. 이들의 관계를 다음 그림 3.1로 나타내었다.

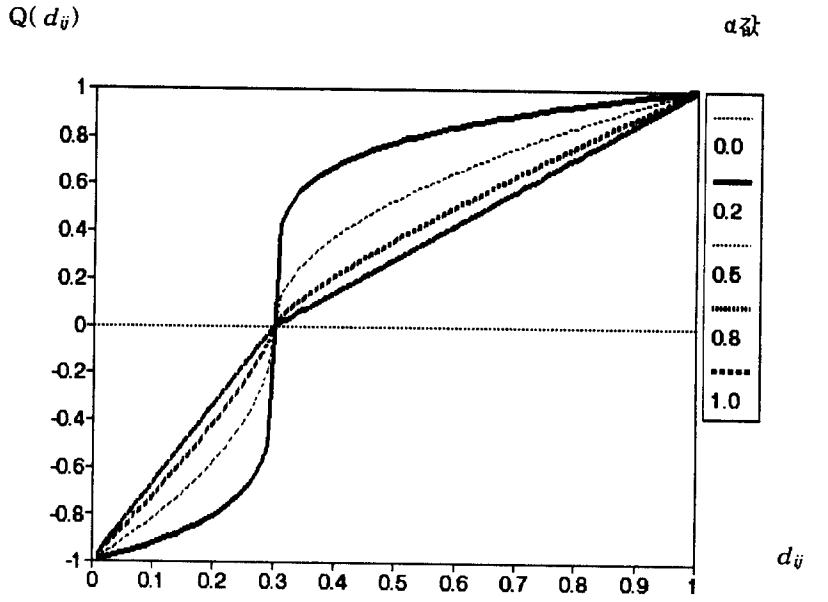

그림 3.1 함수 $Q(d_{ij})$의 그래프 (β=0.3, L=2)

이 함수 $Q(d_{ij})$를 검토해 보면, 두 개체 사이의 상대거리 d_{ij}가 매우 작으면 두 개체는 아주 닮은 것이고, 이에 대응하여 조정된 함수값 $Q(d_{ij})$는 -1 에 거의 가까운 값이 얻어진다. 또한 거리가 커서 두 개체가 아주 다르다면 이 정도에 따라 조정된 함수값 $Q(d_{ij})$는 +1 에 가까운 값이 얻어진다.

그러므로 각 개체들간의 거리를 고려한 이 함수값 $Q(d_{ij})$를 근거로 하여 목적함수값 Ψ을 최소화 한다는 것은 동일집락을 정해 주는 주관적인 임계값의 기준에 의하여 유사하다고 생각되는 가까운 개체끼리는 서로 그 가까운 정도에 따라서 당기게 되어 가까이 모이게 하고, 동시에 임계값 기준으로 서로 거리가 멀어서 다르다고 생각되면 이 개체들 끼리는 먼 정도에 따라서 서로 밀어 내어 더욱 멀어지게 변환한다는 것을 의미한다.

이러한 변환을 임의의 두 개체를 계속적으로 반복하여 관찰하면서 목적함수값을 최소화하는 방향으로 진행하면 Scofield가 보인 바와 같이 변환의 규칙을 결정하는 가중치 ω_{mn}이 어느 한

값으로 수렴하게 된다. 이 수렴된 ω_{nm}의 행렬이 집락형성의 규칙으로 고정되고 이 행렬을 바탕으로 분류가 되지 않는 어떤 새로운 개체를 기존의 자료들에 의하여 분류된 기준으로 분류를 수행할 수 있게 된다.

그리고 상수 α와 β값의 선택에 따라 여러 형태의 집락결과들이 만들어짐을 알 수 있다. 따라서 집락결과가 여러가지 기준에서 판단할 때 적절하지 못할 경우 실험자가 상수를 조정함으로서 자료의 특성에 타당한 새로운 집락결과를 제공할 수 있도록 해 준다.

또한 Scofield(1988)가 다층(multi-layer) 쿨롱에너지네트워크에 적용이 가능한 알고리즘을 제안한 것에 착안하여 필요하다면 각 개체 i를 N차원 공간의 각 점에 대응시키는 지금까지 제안한 방법을 N차원 공간의 Xi를 개체로 보고 다시 P차원의 점으로 변환시키는 과정에 적용하여 일단 N차원상에 집락별로 나누어진 점들을 다시 한번 P차원에 대응시켜서 보다 더 분류가 명확히 될 수 있도록 한다. 대부분의 통계학은 선형변환이 주로 사용되기 때문에 2회 이상의 변환이 의미를 갖지 못하지만 본 논문이 제안하고 있는 방법은 비선형 변환이기 때문에 이러한 시도는 때로는 보다 더 나은 결과를 만들 수가 있다.

4. 집락분석 절차 및 실례

이 절에서는 쿨롱집락분석방법의 기본 알고리즘과 실제로 SPSS/PC+ Manual에 있는 맥주회사별 소비조사 자료의 경우를 소개한다.

4.1 기본 알고리즘

단계 1. 변환 횟수와 각 변환시의 차수를 정하여 네트워크 구조를 세운다.
각 변환마다 ω_{nm}을 초기화한다.
상수 α, β, L, η와 반복회수를 정한다.

단계 2. 각 입력개체 fi에 대하여,

(1) 개체간의 상대 거리 dij와 1단계로 변환된 한 점 Xi를 구한다.

$$X_{\in} = F_n\left(\sum_{m=1}^{k} \omega_{nm} \cdot f_{mi}\right)$$

$$F_n(f_i) = F_n\left(\sum_{m=1}^{k} \omega_{nm} \cdot f_{mi}\right) = \left(1+\exp\left(-(1/T)\cdot\left(\sum_{m=1}^{k} \omega_{nm} \cdot f_{mi}+\theta\right)\right)\right)^{-1}$$

(2) 다음 식을 이용하여 연결강도 ωem를 수정 변화시킨다.

$$\delta\,\omega_{nm} = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} Q(d_{ij}) \parallel x_i - x_j \parallel^{-(L+2)} \Delta_{nm}(f_i, f_j)\cdot\eta$$

$$\Delta_{nm}(f_i, f_j) = [F_n(f_i) - F_n(f_j)]\cdot[F_n(f_i)(1-F_n(f_i))f_{mi} - F_n(f_j)(1-F_n(f_j))f_{mj}]$$

단계 3 반복회수만큼 단계 2를 수행한다.

단계 4. 바로 전층에서 얻어진 출력을 입력으로 사용하여 단계 2에서 단계 3까지 반복한다. 정해진 네트워크 구조상의 모든 층을 수행했으면 단계 5로 간다.

단계 5 .구해진 가중치를 사용하여 집락결과를 확인한다.

단계 6. 결과를 검토한 후 필요하다면 초기치의 상수들을 수정하여

단계 2 에서부터 단계 5까지를 다시 실행한다.
단계 7 새로운 개체의 판별이 필요하면 판별을 실행한다.

4.2 적용 예제

다음의 자료는 1983년 미국의 맥주회사 별 소비조사보고서 자료로서 20개의 맥주제조회사들의 제품을 4가지 변수(칼로리, 소다, 알콜농도, 가격)로서 측정한 것이다 (SPSS, 1986).

표 4.1 미국의 20개 맥주회사별 소비조사 자료

맥 주 회 사	칼로리	소다	알콜	가격	맥 주 회 사	칼로리	소다	알콜	가격
1 budweiser	144	15	4.7	0.43	11 coors	140	18	4.6	0.44
2 schlitz	151	19	4.9	0.43	12 coors light	102	15	4.1	0.46
3 lowenbrau	157	15	4.9	0.48	13 michelob light	135	11	4.2	0.50
4 kronenbourg	170	7	5.2	0.73	14 becks	150	19	4.7	0.76
5 heineken	152	11	5.0	0.77	15 kirin	149	6	5.0	0.79
6 old milwaukee	145	23	4.6	0.28	16 pabst extra light	68	15	2.3	0.38
7 augsberger	175	24	5.5	0.40	17 hamms	136	19	4.4	0.43
8 strohs bohemtan style	149	23	4.7	0.42	18 heilemans old style	144	24	4.9	0.43
9 miller lite	99	10	4.3	0.43	19 olympia gold light	72	6	2.9	0.46
10 budweiser light	113	8	3.7	0.44	20 schlitz light	97	7	4.2	0.47

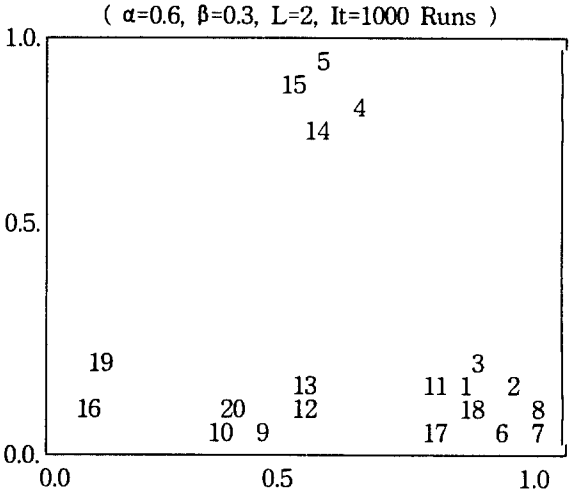

그림 4.1 [예제]에 대한 쿨롱집락분석결과의 2차원 그림

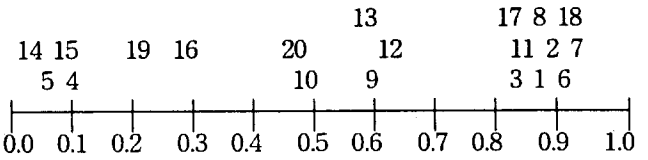

그림 4.2 [예제]에 대한 쿨롱집락분석결과의 1차원 그림

이 4차원 자료(표 4.1)를 쿨롱집락분석방법으로 변환하여 2차원 공간에 표현한 집락결과가 그림 4.1이고, 이를 또다시 1차원 공간에 표현한 집락결과가 그림 4.2이다. 이 공간좌표를 토대로 개체 20개를 집락들로 구분해 본다면, 그림 4.1과 그림 4.2 에서 모두 같은 결과로써 집락 1 (9, 12, 10, 20, 13), 집락 2 (16, 19), 집락 3 (5, 15, 4, 14), 집락 4 (11, 17, 1, 3, 2, 8, 18, 6, 7)를 형성하고 있음을 관찰할 수 있다.

이 결과는 기존의 계보적 집락분석방법인 최장연결법에 의한 분석결과에서 집락의 개수를 4개로 하는 경우와 동일한 것이었다.

쿨롱집락분석방법이 집락을 형성하면서 자동적으로 분류규칙을 기억하여 새로운 개체가 관측이 되었을 경우에 그 개체가 어느 집락에 속할 것인지의 판별을 할 수가 있는 특징을 확인하기 위하여 새로운 자료가 관측되어 표 4.3으로 얻어졌을 경우 이 새로운 개체의 판별에 대한 쿨롱집락분석방법의 결과가 그림 4.3에 있다.

표 4.3 새로운 개체들

개 체	칼로리	소 다	알 콜	가 격
21	0.38	0.62	0.61	-0.46
22	0.81	1.22	0.74	-0.11
23	1.24	-1.21	0.34	0.02
24	-0.60	-0.74	0.90	1.05
25	0.42	0.01	0.21	-0.85

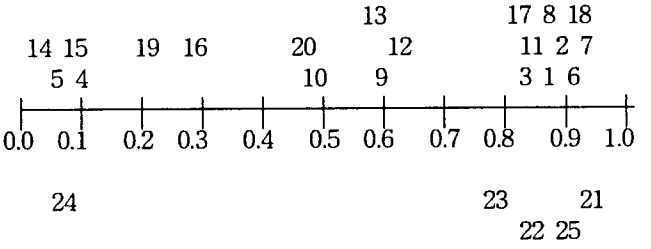

그림 4.3 새로운 개체를 판별한 결과의 1차원 그림

위 표 4.3에서 5개(21, 22, 23, 24, 25)의 자료를 쿨롱집락분석방법을 시행하는 동안 생성하여 기억된 집락형성규칙을 사용하여 어느 소속에 할당을 할 것인가의 판별결과는 앞의 그림 4.3에서 보는 바와같이 새로운 개체 (21, 22, 23, 25)는 기존의 개체 (1, 2, 3, 6, 7, 8, 11, 17, 18)로 형성된 집락으로, 새로운 개체 24는 기존의 개체 (4, 5, 14, 15)로 구성된 집락으로 소속됨을 알 수 있다.

반면에 기존의 계보적 집락분석방법에서는 새로운 개체 5개만을 분류할 수 없기 때문에 원래의 20개의 옛자료와 함께 25개의 자료를 사용하여 다시 계보적집락방법을 시행해야 하며 이때 얻은 결과는 새로운 개체 21, 22, 25는 기존의 개체 11, 17, 21, 1, 25, 3, 2, 8, 18, 22, 6, 7로 구성된 집락에 소속되고 새로운 개체 23, 24는 기존의 집락 (4, 5, 14, 15)에 소속되고 있다. 이 결과는 23번 개체를 제외하고는 쿨롱집락분석방법과 동일한 판별을 하고 있는 것으로 나타났다.

5. K-Means 집락분석기법과의 비교

기존의 집락분석방법과의 비교를 위하여 쿨롱집락분석과 개념이 비슷한 K-Means집락분석기법과의 비교를 위하여 Rand의 C 통계량을 사용하였다. 여기서 자료는 BN(0, 0, 1, 1, 0), BN(0, 4, 1, 1, 0), BN(4, 0, 1, 1, 0)의 이변량 정규분포에서 크기가 각각 10개인 표본을 한 집합으로 IMSL을 사용하여 10개의 집합을 생성하고 이들 자료를 분석하여 얻은 결과의 Rand C 통계량은 표 5.1과 같다. 또한 구형이 아닌 자료에서 나타나는 현상을 보기 위하여 BN(0, 0, 1, 1, 0.5), BN(0, 4, 1, 1, 0.5), BN(4, 0, 1, 1, 0.5)에서 역시 크기가 10인 표본을 한 집합으로 10개의 집합을 생성하여 얻은 Rand C 통계량이 표 5.2와 같다.

표 5.1

set	1	2	3	4	5	6	7	8	9	10	평 균
K-Means	0.917	0.878	1	1	1	1	1	0.878	0.878	1	0.955
Coulumb	0.917	0.844	0.818	1	0.885	0.956	1	0.878	0.717	1	0.902

표5.2

set	1	2	3	4	5	6	7	8	9	10	평 균
K-Means	1	0.814	0.956	0.917	1	0.915	0.956	0.844	0.915	1	0.932
Coulumb	1	0.878	0.956	0.956	0.956	0.818	0.956	0.844	0.956	0.915	0.924

표 5.1과 표 5.2에서 공히 볼 수 있듯이 K-Means방법이 몇 set에서 약간 더 좋은 값을 나타

내고 있는 것으로 나타났으나 대부분에서는 거의 비슷한 값을 보여주고 있어서 쿨롱집락분석기법이 자료를 분리하는 정도에서는 기존의 기법들과 비슷한 수준이라고 할 수 있다고 본다.

본래의 쿨롱집락분석방법은 개체들이 집락화된 상태의 그래픽 결과를 보여주는 것을 목표로 하기 때문에 본 비교 실험에서 기준이 되는 Rand C를 계산할 수가 없는 문제가 있다. 이를 해결하는 방안으로서 위의 실험에서는 각 개체들을 궁극적으로 모두 일차원의 0과 1사이의 점들에 대응시키고 집락의 수가 세개이므로 각 점들 사이의 거리 중 가장 큰 두개를 잡아 그 해당되는 점들의 중간점을 집락의 경계로서 설정하는 알고리즘을 사용하여 정량적인 집락화가 되도록 하였다.

6. 결 과

기존의 집락분석기법들이 집락을 형성하지만 그 집락형성에 사용된 규칙이 보존되지 않는 문제를 해결하여 보고자 신경회로망에서 연구되고 있는 쿨롱에너지 네트워크 모형을 변형 확장시켰다. 이 모형을 통하여 K차원으로 표현된 개체를 적절한 N차원(보통 1, 2차원)의 점에 집락을 형성하도록 하면서 할당시킨다. 이 모형을 사용하여 실제 자료를 분석한 결과는 최장 연결법에 의한 분석 결과와 동일한 것으로 나타났고 또한 시뮬레이션 자료를 이용하여 K-Means 알고리즘과 Rand C를 통한 비교에서도 거의 비슷한 수준의 결과를 나타내었다. 따라서 본 논문에서 제안한 쿨롱집락분석기법은 기존의 기법들과 거의 유사한 분석 결과를 보여주면서 동시에 분석에 사용된 분류 기준이 분석후에 제공되어 이들 형성된 집락들을 토대로 향후 새로운 개체를 판별할 수 있도록 하여 주는 것을 보였다. 제기되는 문제는 주어진 자료의 정보를 평균과 분산-공분산 행렬로 요약하여 사용하는 기존의 통계적 방법이 아닌 임의의 두 개체의 반복적인 비교를 통하여 분류를 강화시켜가는 방법으로 시간이 오래 걸리는 것인데 시간에 관하여는 개체 i가 대응되는 N차원의 점의 각 축(coordinate)을 계산하는데 있어서 개체 i를 표현되는 벡터 fi의 각 성분을 모두 이용하므로 N차원의 각 축이 독립적인 프로세서(processor)로 구성되어 있다면 해결될 수 있는 문제로 병렬형 컴퓨터의 발달 속도를 볼때 곧 극복될 수 있으리라 생각된다.

향후 연구과제 중 하나는 본 연구의 기본 생각이 비선형이라는 사실을 제외하고는 수량화 이론과 많이 유사한데 착안하여 대단히 원시적인 단계에서 두 기법의 비교를 시도하였는데 이 비교를 보다 발전시켜서 수량화 이론의 확장개념으로 쿨롱네트워크 모형을 변형시키는 것을 생각하고 있다.

또 다른 하나는 K차원의 한 점으로 표현된 개체 i를 2차원상의 점으로 변환시키는 과정에서 다단계변화 가능성을 언급하였는데 여기서 중간 단계를 몇 차원으로 하는 것이 좋은지를 결정하는 것과 이들의 영향을 조사하여 보는 것이다.

참고문헌

[1] 김응환, 이경희, 이원돈 (1992a). 평균장이론 신경회로망의 수량화문제에의 응용, 「충남과학연구지」, 제19권 2호.

[2] 김응환, 최희숙, 이원돈 (1992b). Coulomb Energy Network 에서 Temperature 변화에 따른

학습, 「충남과학연구지」, 제19권 2호.

[3] 이석훈, 김응환 (1992). A Clustering Method of Ambiguous Representation, 「Technical Report」, No. 9201, 충남대학교 자연과학대학 계산통계학과.

[4] Carpenter, G. A. and Grossberg, S. (1986). Associative Learning, Adaptive Pattern Recognition and Cooperative Decision Making by Neural Networks, *Proceedings of the International Society for Optical Engineering,* Vol. 218-247.

[5] Dembo, A., and Zeitouni, O. (1987). ARO Technical Report, Brown University, center fer Neural Science, Provcdence, R. I.

[6] Jain, A., and Dubes, R. (1988). *Algorithms for Clustering Data,* Prentice-hall, Englewood Cliffs, New Jerseg.

[7] Kim, D. S. (1990). *Properties and Characteristics of Self-Organizing Neural Networks for Unsupervised Pattern Recognition,* Ph.D Dissertation, Department of Computer Science, University of South Carolina.

[8] Kim, Y. H., Choi, H. S., Lee, K. H. and Lee, W. D. (1992c). Learning of the Coulomb Energy Network on the Variation of the Temperature function, *Proceedings of the International Joint Conference on Neural Networks,* Baltimore, MD, Vol. I, 749-754, July 7-11.

[9] Kim, Y. H., Lee, J. C., Lee ,W. D. and Lee, S. H. (1992d). Pattern classifying Neural Network based on Fisher's Linear Discriminant function, *Proceedings of the International Joint Conference on Neural Networks,* Baltimore, MD, Vol. I, 743-748, July 7-11.

[10] Kohonen, T. (1987). *Self-Organization and Associative Memory,* Springer-Verlag, Berlin, Second Edtion.

[11] Kosko, B. (1990). Unsupervised Learning in Noise, *IEEE Trans. on Neural Networks,* Vol. 1, No. 1, March, 44-57.

[12] Lee, S. H., Lee, J. C., Kim, Y. H. and Lee, W. D. (1991). Pattern classifying Neural Network based on an Entropy Measure, *1991 Japanese Neural Network Society(JNNS),* 72-73.

[13] Lippmann, R. P. (1987). *An introduction to computing with Neural Nets,* IEEE ASSP Magazine, April, 4-22.

[14] Scofield, C. L. (1988). Learning international representations in the coulomb energy network, *International Conference on Neural Network,* Vol I, 271-275.

[15] SPSS/PC+ advanced statistics (1986).

A Clustering Method Using the Coulomb Energy Network[4)]

Sukhoon Lee[5)], Nae-Hyun Park[6)], Yung-Hwan Kim[7)]

Abstract

This article deals with the problem that all the statistical clustering methods do not supply the clustering rule after the analysis. We modify the Coulomb Energy Network model basically developed in physics and suggest one model appropriate for our purpose and show the implementation using an actual data. Finally the method suggested is compared with one of the well known methods, K-means algorithm using Rand C.

4) This paper is supported by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1993.

5) Department of Statistics, Chung Nam University, 220 Gungdong Yousunggu Taejeon, 305-764, KOREA.

6) Department of Statistics, Chung Nam University, 220 Gungdong Yousunggu Taejeon, 305-764, KOREA.

7) Chung Nam Mechanical Technical High School, MunHwa dong Junggu Taejeon, 301-130, KOREA.