Historia Mathematica Vol. 10. No. 2, 1997, 48-52

군집방법의 역사와 응용사례에 관한 고찰

서경대학교 응용통계학과 **이승우**

Abstract

통계학이란 미래에 대한 예측을 하고 이에 대비하여 합리적인 의사결정을 내리는데 도움을 받을 수 있는 학문이다. 최근 다변량 통계분석은 관찰이나 실험의 대상이 되는 하나 이상의 변수들을 동시에 분석할 수 있는 매우 실제성이 높은 분석방법으로 통계학, 경영학, 사회학, 심리학, 생물학 등 여러 전공 분야에서 복잡하고 다양한 자료 분석에 폭넓게 활용되고 있다.

이 논문에서는 다변량 분석 방법 중 컴퓨터와 통계 분석 소프트웨어의 발전으로 인하여 최근에 활발히 연구되고 있는 군집방법의 역사와 여러 연구분야의 실제자료분석에 응용할 수 있도록 군집분석을 6가지로 나누어서 분류하였고 그 방법론을 제시하였다.

0. 서론

우리는 어떤 자연적인 현상이나 사회적인 현상을 과학적으로 연구하기 위해서는 자료를 수집하고 분석하여 그 타당성을 검정하는 과정을 거쳐 인정받은 이론을 도출해야 하는데, 이러한 검정과정에서의 복잡성을 단순화하기 위하여 개발된 통계적 분석 기법이 다변량분석이다. 다변량분석은 하나 이상의 변수를 동시에 분석하는 것을 뜻하며, 그 목적은 여러 변수들의 연관성을 측정, 설명, 예측하는 데 있다. 다변량 분석 방법은 다변량 자료 연구에 대한 분석적인 분야이며 의사결정을 위한 자료의 수집과 분석, 설계에 영향을 미치는 통계학의 한 분야이다. 다변량 자료에 대한 통계분석은 자료의 성질 또는 연구 목적에 따라 다양한 방법으로 시도하여 자료를 비교적 적은 모수(parameter), 즉 자료를 간략화 하여 쉽게 파악하는데 있으며 자료의 복잡성으로 인하여 컴퓨터를 이용한 SAS, SPSS, S-plus 같은 통계처리 프로그램을 이용한 분석법이 개발되었다.

다변량 모집단의 모수에 대한 가설을 검정하는 기법으로서 다변량 분산 분석(multivariate analysis of variance)과 다변량 공분산 분석(multivariate analysis of covariance) 등이 있으며 변수들간의 상호 독립성 혹은 종속성 조사에 관한 기법으로서 회귀분석(regression

analysis), 상관분석(multiple correlation analysis), 정준상관분석(cannonical correlation analysis) 등이 있다. 변수사이의 상호관련성에 기초한 자료의 축소 또는 구조의 단순화를 목적으로 가치있는 정보를 손실하지 않으면서 연구 대상을 가능한 한 적은 수의 변수로 표현하는 기법으로서 주성분 분석(principal component analysis), 요인분석(factor analysis) 등이 있다. 집단사이의 상호관련성에 기초한 분류 또는 집단화를 목적으로 측정된 자료의 특성을 기초로 비슷한 것들끼리 집단으로 묶는 기법으로서 판별분석(discriminant analysis), 군집분석(clustering), 다차원 척도 분석(multi-dimensional scaling) 등이 있다.

1. 군집분석

1970년대 이후 컴퓨터의 급진적인 발달에 의하여 염두조차 낼 수 없었던 새로운 이론들이 개발하여 왔으며 그 하나의 예로서 자료를 같은 유형의 그룹으로 분류하고자하는 방법중 가장 일반적인 통계적인 방법 중에서 군집방법(cluster method)을 들 수 있다. 군집방법이란 관심의 대상이 되는 다수의 개체(individual)를 몇 가지의 군집으로 분류하되, 같은 군집에 속하는 개체들 간에는 서로 유사하고 다른 군집에 속하는 개체들 간에는 서로 상이하도록 하고자 하는 통계적 방법으로서 일반적으로 어떤 연구 하에서 개체들의 집단화를 위하여 사용되었다. 여기서 말하는 개체는 경우에 따라서는 사람일 수도 있고, 동물일 수도 있는 관심의 대상이 되는 것들의 낱낱을 통칭한다. 군집분석은 변수를 분류하여 구룹화 하기 위한 여러 가지 방법으로서 통계학자인 Kendall과 Stuart는 군집분석이라는 용어를 사용하였다.

고대에는 사람들을 어떤 유형별로 각각 분류할 수 있었다. 힌두교인 들은 사람들을 성별, 신체적 그리고 행동상의 특성에 따라 6가지 유형으로 분류하였고 각각 동물의 이름을 부여하였다. 초기 그리스와 로마의 의사들은 4가지의 영(humour)의 혼합으로부터 유래됐다고 여겨지는 신체적인 특성의 변이에 근거한 여러 가지 종류의 인쇄학을 발전시켰다. 이러한 인쇄학중에서 가장 탁월한 것으로 여겨지는 것은 그리스의 의사인 Galen(A. D. 129-199)의 인쇄술이다. 그는 여러 가지 질병중 특히 어떤 병원체에 쉽게 감염되는 체질과 개개인의 행동의 차이는 서로 밀접한 관련이 있다고 가정했으며 이에 따라 9가지의 유형별로 인간들을 구분하였다. 18세기에 접어들어, 스웨덴의 식물학자인 Linnaeus(A. D. 1707-1778)는 식물과 동물의 세계를 분류하는데 관심을 기울였다. 앞서 언급했듯이, 우리가 현실에서 소유하고 있는 모든 지식들은 유사한 종류와 유사하지 않은 종류의 모임으로 구분할 수 있는 방법을 찾고자 하고 또, 그 얻어진 방법에 따라 모든 사물을 구분 짓는다.

대부분 군집분석의 초기의 업적은 생물학과 동물학의 분야에서 사용되었으며 이 분석은 일반적으로 분류학으로 더 알려졌다. 초기에 가장 광범위한 의미에서 분류학은 과학적인 방법보다는 예술의 의미가 강조되었다. 그러나 점차 이러한 기술의 목적은 초기 Adanson(18세기)의 착상(idea)에 근거한 수치적인 분류학의 방법에 의하여 궁극적으로 개발되었다. 최근에 자연과학을 제외한 모든 분야에서 수리적인 분류 방법에 사용하기 위하여 Zubin(1938)과 Thorndike(1953)에 의하여 제시된 탁월한 아이디어에 의한 시도가 있었다. 그러나 일반

적으로 그들의 착안 한 사용법은 매우 복잡한 계산을 필요로 하기 때문에 컴퓨터에 의하여 계산 가능한 지난 15년동안에 비로서 보급되었다. 또한 이시기에 여러 다른 분야에서 종사하는 연구가에 의하여 군집방법외에 새로운 분류방법을 개발하였다.

지난 10년동안 수리통계학자와 수학자는 군집분석의 보다 더 형식적인 접근방법을 위한 필요성을 인식했고 정확한 통계적 모형과 수학적 모형을 공식화하기 위한 시도와 더불어 분류문제를 더 엄격하게 채택하기 위하여 새로운 시도를 하였다. 이러한 더 형식적인 접근방법의 예는, Wolfe(1970)와 Jardine & Sibson(1971)의 업적에서 엿볼 수 있다.

기본적인 군집방법 이론에 대하여 간략히 소개하면 다음과 같다. 군집분석에서는 집단의 수 혹은 집단 구조에 대한 가정이 없으며, 어떤 개체나 대상들을 밀접한 상사성(similarity) 또는 거리(distance)에 의하여 특성을 지닌 개체들을 몇 개의 군집으로, 동질성을 지닌 군집으로 집단화하는 다변량기법이다. 만약 n개 개체를 c개의 군집으로 분류할 수 있는 가능한 모든 경우를 나열하고 그 중에서 가장 만족스러운 것을 찾아내면 되지만 불가능하다. 따라서 가장 만족스러운(optimal) 해답을 찾는 대신에 어느 정도 만족스러운 해답을 체계적으로 찾을 수 있는 한 가지 방법으로 계보적 군집방법(hierarchical clustering method)이 있으며 알고리즘은 다음과 같이 요약된다.

<계보적 군집방법의 알고리즘>

1단계 : n개 개체 각각을 하나의 군집으로 간주한다.
2단계 : 가장 유사한 두 군집을 찾아서 합친다.
3단계 : 합쳐진 두 군집과 나머지 군집간의 상사성을 계산하다.
4단계 : 전체가 하나의 군집이 될 때까지 2단계와 3단계를 $n-1$번 반복한다.

군집분석을 수행하기 위해서 상사성이 높은 개체들은 같은 군집에 포함시키고, 상대적으로 상사성이 낮은 개체들은 서로 다른 군집에 포함시킬 수 있도록 해주는 상사성 혹은 비상사성(similarity/dissimilarity)의 정도를 측정하는 기준 척도로서 그 객체들간의 거리 개념을 이용하여 Minkowski 거리, Euclidean 거리, Mahalanobis 거리등으로 정의할 수 있다.

계보적 군집 방법에는 상사성이 밀접한 개체들을 군집으로 형성하는 병합적(agglomerative) 방법과 비상사성이 큰 개체를 단계적으로 분리해나가는 분할적(divisive) 방법으로 구분할 수 있으며 그중에서 두 집단간의 최단거리로 측정할 수 있는 방법을 최단 연결법(single linkage)이라고 하고 두 집단간의 최장거리로 측정하는 것을 최장 연결법(complete linkage)이라고 하며, 두 집단간의 평균거리로 측정하는 것을 평균 연결법(average linkage)이라고 하며, 두 집단 사이의 거리는 두 집단의 중심간의 거리로 계산되는 중심 연결법(centroid linkage)등이 있으며, 중위수 연결법(median linkage), Ward의 방법, Lance와 Williams의 방법 등이 있다.

최적 분리 군집방법은 어떤 개체가 특정한 군집에 할당되면 다른 군집에 다시 할당될 수 없는 단점을 극복한 방법으로 K-평균 군집방법과 트레이스에 근거한 방법 등으로 정의된

다. 또한 중복군집과 Fuzzy군집 등이 있다.

2. 군집분석으로 적용 가능한 응용분야

군집분석은 복잡한 자료들의 구조를 발견하는 다변량 분석법으로서 많은 다양한 기법들을 포함하고 있다. 오늘날, 군집분석의 필요성이 많은 연구 분야에서 자연 발생적으로 부각되고 있으며 중요한 연구분야로서는 다음과 같이 세분화 할 수 있다.

1. 생명과학 : 생물학, 생태학, 동물학, 식물학, 고생물학, 화석학, 미생물학, 곤충학
2. 행동 및 사회과학(Behavioral and social science) : 심리학, 사회학, 인류학, 언어학, 고고학, 범죄학, 형사학
3. 지구과학 : 지질학, 지리학, 토질연구(soil science), 지역연구(regional study)
4. 의학 : 정신의학, 세포학, 병리학, 임상진단
5. 공학 : 인공지능, 전기공학, 인공 두뇌학, systems science, pattern recognition
6. 정보과학, policy science : 경제학, operations research, information retrieval, political science, marketing research

위에서 언급한 생명과학 분야로서, 이 군집 방법은 식물, 동물, 곤충, 세포, 박테리아 등의 미생물, 상호 의존적 관계로 공생하는 식물의 군락, 구시대의 생존한 생명체의 화석 기록물과 같은 생명체를 연구의 목적에 따라 유사한 특성을 지닌 어떤 개체나 대상들을 몇 개의 군집으로 집단화하는 기법이다. 이 분석을 사용하는 목적으로는 독특하나 잡다한 종(species) 중에서 변종(subspecies)을 완벽하게 제거하는 분류법을 개발하는 것이다.

행동 및 사회과학 분야에서는 군집분석의 적용이 예외적으로 다양성 있게 적용된다. 군집분석의 대부분의 개체는 훈련 방법, 행동 모형(behavior pattern), 인간행동의 요인, 인간의 판단, 조직 통제가 있는 단체, 가족집단, 장기적으로 약을 복용하는 환자, 검사항목(test item), 어떤 지역의 이웃 사람들, 클럽 및 협회, 범죄 및 범죄자, 학생, 학과 과목, 문화, 언어, 교습법(teaching techniques), 공예품, 인공유물, 유적지 등으로 이루어진다. 전통적으로 이러한 적용분야는 다변량 통계분석의 요인분석에서도 적용 가능하다.

지구과학 분야는 대륙 및 암반형성, 토양, 하천의 생성, 도시, 국토, 세계각지의 지역, 그리고 대륙을 유용하게 사용하는 방법 등이 군집분석의 응용분야에 포함된다.

생명과학과 밀접한 관계를 이루고 있는 의학 분야에서 군집분석의 개체는 질병, 환자, 병의 증상 그리고 실험실에서의 검사(laboratory test) 등이다. 이 분야에서의 분석의 목적은 보다더 효과적이고 경제적인 수단으로서 환자의 진료를 정확한 진단으로 유도하여 개발을 도모하는 것이다.

일부 공학 분야에서도 군집 분석이 적용되고 있다. 이에 대한 전형적인 예로서, 펜 또는 연필 등을 사용하여 손으로 쓴 문자, 언어의 종류, 지문 채취, 영상 및 무대 배경, 전자 심장

계, 파도의 형태(waveform), 전파 탐지기의 신호 그리고 전자회로 디자인 등에 적용된다. 분석 방법에 의한 공학분야의 응용 범위는 비교적 자료의 수가 많이 필요하지 않기 때문에 이 분야에서의 군집분석의 필요성은 나날이 급증할 것이며 뿐만 아니라, 일부 극단적인 새로운 문제의 발생도 예견된다.

정보, 정책 및 결정과학(decision) 분야에서는 나라 및 주(states), 입법부 의원, 정치적인 논쟁에 대한 투표, 생산자, 소비자, 제작용품, 상품의 매매, 판매 계획, 연구 및 개발 계획, 선거구, 투자, 개인의 양도 증서, 신용카드 지불 유예 기간, 공장 위치 그리고 건물 계획 도안과 같은 문제를 다룬다. 이러한 분석의 연구 분야는 지난 수년동안 군집분석의 가장 혁신적인 발전을 거듭했다.

3. 맺는 말

통계학자 하디트와 페트리노비치(1976)에 의하면, 다변량분석법은 앞으로 지배적으로 등장할 것이며 문제를 제거하는 방법과 연구를 설계하는 방법에 커다란 변혁을 가져올 것이라고 예견하였다. 위에서 언급한 이 분석법은 많은 변수들의 상호 관련성을 이해하고 그에 따른 해석을 요구하기 때문에 어려운 점이 많았으나 근래에 와서 우수한 통계패키지 프로그램의 출현과 기타 컴퓨터 환경의 개선에 힘입어 비교적 복잡한 문제들에 대한 의문점들에 관하여 현상에 영향을 주는 복합적인 변수들간의 연관상태를 있는 그대로 이해할 수 있으며, 나아가서 변수들을 분리시킨 후 개별적인 효과를 이론적인 접근방법으로 통계학을 연구할 수 있게 해 준다

군집분석은 자료의 구조에 관한 정보가 없고 분명한 분류기준이 없거나 알려져 있지 않은 상태에서 활용할 수 있도록, 객체들에 관한 정보를 보다 적은 개수의 동질성 집단으로 분류함으로써, 최소의 정보 손실을 도출한다는 점에서 자료의 탐구, 자료의 단순화 및 집약 과정으로 인식될 수 있다. 군집화의 결과는 모집단의 구조적 특성에 관한 정보를 도출함으로서 자료의 질적 수준 및 관리에 대한 평가의 올바른 이해를 통해 통계 전문가의 역할을 수행하는데 혁신적으로 진전을 도모할 수 있다. 이런 관점에서 군집분석은 다변량 분석기법의 그 어느 것보다 선행되어야 할 연구 분야로 고려된다.

참고 문헌

1. Richard A. Johnson and Dean W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall, 1992.
2. 강병서, *다변량 통계학*, 법문사, 1988.
3. 김기영, 전명식, *SAS 군집분석*, 자유아카데미, 1990.