응용통계연구 제 **14**권 1호, *2001*년, *pp. 177-190*

시각적 군집분석에 대한 전략 *

허문열 [1)]

요 약

전통적으로 많이 사용하는 군집분석의 방법들은 개체간의 거리를 고려하여 이들을 분류해 내는 것이며, 따라서 거리 측정 방법에 따라 여러 형태의 군집분석 방법이 나타나게 된다. 어떤 방법을 적용하던 간에 그 결과는 고정된 수치로써 나타난다. 다차원 자료의 구조파악이 몇 개의 수치로 나타나게 되면 어쩔 수 없이 정보의 손실이 발생하게 된다. 이를 보완하기 위해 시각적 매체를 동원하여 다차원 자료의 구조를 파악하는 연구가 있었으며, 이를 시각적 군집분석이라고 명명하고 있다. 본 연구에서는 시각적 군집분석에 대한 기본적 개념과 이를 위한 통계 도형의 활용, 구현방법 등에 대해 살펴보기로 한다.

주요용어: 데이터 시각화 , 군집분석, 산점도 행렬, 평행좌표계, 덴드로그램.

1. 개요

능력이 있는 사람이라면 복잡하게 얽혀있는 사건 (또는 물건, 대상, objects) 들을 조직적으로 정리하고 여기서 적절한 정보를 획득할 수 있어야한다. 이러한 과정은 이들 사건들이 가지고 있는 성격들을 파악하고 여기서 공통적인 내용을 갖는 것과 그렇지 않은 것들을 분류함으로써 이루어진다. 다행히 분류하고자 하는 대상들이 눈에 보이는 것이라면, 이를 위한 과정도 쉬울 뿐 아니라 결과의 유용성도 쉽게 판단할 수 있다. 그러나 이 대상들이나 대상들의 성격을 시각적으로 직접 관측할 수 없거나 객관적으로 수량화하기 어려운 경우, 분류의 과정과 결과의 유효성을 판단하는 것이 어려워지고 객관성이 결여되기 쉽다. 이러한 경우, 어떤 성격을 택하는 가, 그리고 이를 어떻게 수량화하고 어떤 방법을 사용하여 분류하는 가가 논란의 대상이 될 수 있다.

시각적 군집분석은 시각화 도구를 사용하여 자료의 구조를 파악함으로서 자료에 대한 군집을 파악하려는 연구라고 할 수 있다. 즉, 시각적 군집분석(Visual Clustering, VC)은 Visualization + Clustering 이라고 할 수 있다. 여기서 시각화(Visualization)라고 하면 분석자(혹은 시스템 사용자)가 자료를 도형으로 표현할 때 자료의 변화에 따른 도형의 조정이 실시간에 이루어지는 경우를 말한다. 기존의 통계적 도형은 자료가 주어지고 도형의 형식이 주어지면 하나의 정적인 결과가 주어졌으나, 시각화에서는 어느 순간에라도 자료의 조정이 일어날 수 있고, 도형의 형식이 변화할 수도 있으며 이 두 가지 과정이 실시간에 나타나게 된다. 이러한 과정은 반드시 컴퓨터에 의해서만 이루어지는 것은 아니지만, 컴퓨터

* 본 논문은 한국학술진흥재단 연구비 (19980364-000)에 의해 이루어 졌음.

1) (110-745) 서울특별시 종로구 명륜동 3가 53번지, 성균관대학교 통계학과, 교수
E-mail: myhuh@skku.ac.kr

그림 1.1: 강원도 더덕과 북한산 더덕을 시각적인 방법에 의해 분류해 놓았다. 이러한 분류과정은 더덕에 대한 지식을 가진 사람이면 누구든 쉽게 처리할 수 있고, 분류 결과도 쉽게 파악할 수 있다. 이를 수학적 방법에 의해 분류하려면, 더덕의 각종 성격을 수량화 해야하고 이 결과를 적절한 분류방법에 적용시켜야 하며, 결과를 파악하는 것도 용이하지 않다.

를 이용하는 것이 편리하기 때문에 시각화라고 하면 특별한 언급이 없는 한 사용 도구는 컴퓨터가 된다. 시각화 방법에 의해 데이터의 구조를 파악하고, 데이터에 대한 추론(Visual Inference)을 하는 과정에 대해 많은 학자들은 객관성이 결여되었다고 비판적인 의견을 갖고 있다. 그러나 VC는 의식적이던 무의식적이던 우리가 항상 수행하고 있으며, 주위에서도 항상 나타나고 있다.

그림 1.1을 살펴보자. 이는 노점상의 "더덕 좌판"에 여러 종류의 더덕을 각 산지별로 분류해 놓은 것이다. 즉, 여기서는 여러 종류의 "더덕"이라는 대상들이 있으며 이 대상에 대한 "전문가"가 이들을 VC에 의해 같은 성격을 갖는 것들끼리 분류하여 놓았다. 그러나 통계학자들이 다루는 다차원 자료는 더덕의 경우와 같이 항상 명쾌하게 대상을 볼 수 있는 것이 아니다. 예를 들어 유명한 Fisher의 분꽃 자료의 경우, 3종류의 꽃에 대해 각각 50개씩 전체 150개의 관측값이 있고 각 관측값에 대해 4가지를 측정하였다. 만약 이 자료로부터 150개의 분꽃을 실제 모양 그대로 재생할 수 있다면, 분꽃에 대한 전문가는 VC에 의해 그림 1.1과 같은 형식으로 3가지 분꽃으로 분류할 수 있을 것이다. 그러나 우리에게 주어진 4차원 자료만 이용하여서는 원래의 분꽃을 재생할 수 없기 때문에 주어진 자료를 이용한 통계적 도형을 통해 VC를 수행하게 된다. 따라서 이 결과는 주어진 자료에 의해 매우 달라질 수 있다. 그러나 해당 자료에 대해 전문적인 지식이 있는 사람이 이 자료를 이용하여 VC를 수행한다면 매우 유용한 결과를 획득할 수 있을 것이다. 통계학자가 여기서 기여

할 수 있는 일은 과학적이고 효율적인 VC의 방법을 개발하는 것이다. 이렇게 함으로서 이를 사용하는 사람이 효율적이고 객관적인 결론을 유도할 수 있도록 도와주어야 한다.

VC의 연구에 대한 고전적인 참고자료는 Cleveland와 McGill이 편집한 *Dynamic Graphics for Statistics*(1988) 가 있으며, *Journal of Computational and Graphical Statistics*, IEEE 의 *Visualization Proceedings* 등에 많은 논문이 발표되었다. VC를 위한 소프트웨어로는 1998년 Swayne, Cook, Buja 가 발표한 *XGobi*(1998)가 있고, Unwin, Hawkins, Hofmann, Siegel 의 *MANET*(1996), Wegman 등에 의한 *ExplorN*(1997, 1999), Lee, Kim, Huh, Jeong 의 *MUVIS-1*(1997), 본 연구팀이 개발하 고 있는 *SIVON*(2000a, 2000b, 2000c) 등이 있다.

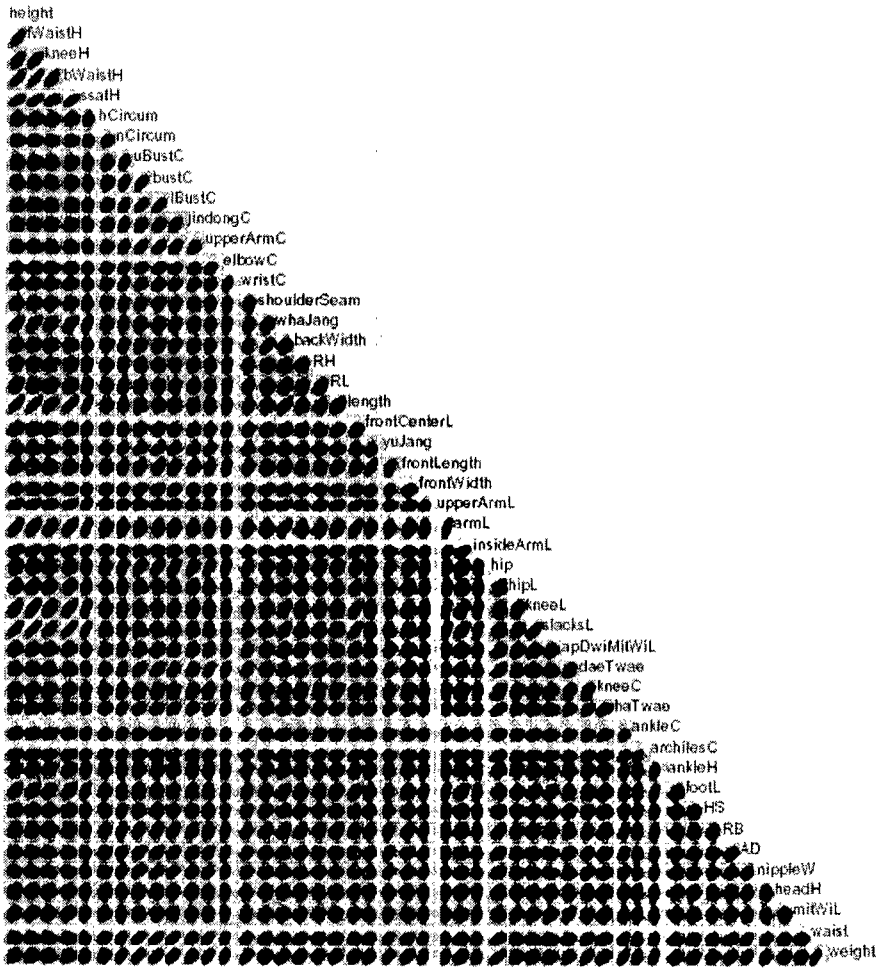

그림 1.2: 인체측정자료에 대한 산점도행렬. 변수의 수가 많아지면 산점도행렬이 제공하는 정보는 매우 빈약해 지는 것을 알 수 있다.

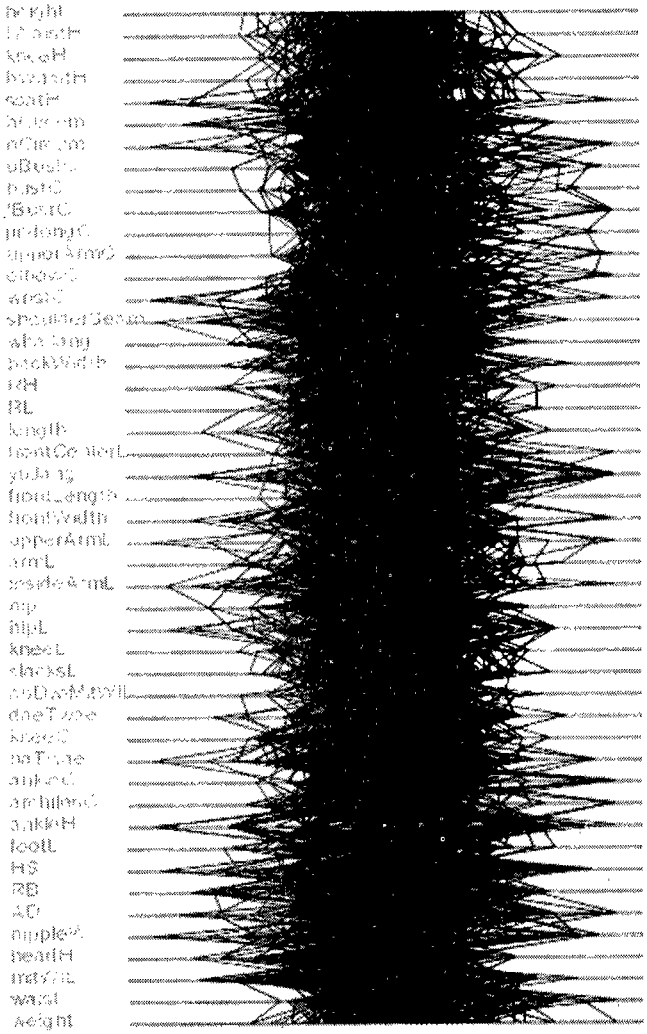

그림 1.3: 인체 측정자료에 대한 평행좌표계. 이 도형은 자료를 k-means 방법에 의해 2 개의 그룹으로 나누고, 각 그룹을 다른 색깔로 나타낸 것이다. 평행좌표계는 각 선들의 움직임에 대한 추이로부터 자료의 구조를 파악하는 것이 목적이기 때문에 변수의 수가 커지는 것에 대해 산점도 행렬만큼 영향을 받지 않으나, 역시 변수의 수가 많아지면, 유용성이 떨어진다. 이 도형을 활용하면, 오류 때문에 데이터베이스에서 특이하게 작거나 큰 값이 있는 것을 찾아낼 수 있다. 예를 들어 ssatH (샅 높이)와 bWaistH (등허리높이) 두 변수를 비교해 볼 때, 두 변수가 모두 길이에 해당하는 변수이므로 하나가 크면 다른 것도 커야한다. 그러나 ssatH에서 아주 작거나 큰 값이 bWaistH에서는 동일한 양상을 보이지 않는 것을 알 수 있다.

2. VC를 위한 통계적 도형과 특성

VC를 위한 통계적 도형은 기존의 모든 방법이 동원될 수 있다. 예를 들어 일차원 도형의 경우 막대그래프, 히스토그램, 상자수염도(box-whisker plot), 라인차트 (line chart), 파이차트, 점 도형(dot plot), FEDF(Huh, 1995) 등이 있다. 이차원 도형에서는 가장 많이 사용되는 산점도가 있다. 3차원 이상의 자료를 도형으로 표현하는 방법에 대해서도 산점도 행렬, 평행 좌표계 (parallel coordinates), 별 도형 (star plot)등 여러 가지가 있으나 가장 많이 사용되는 도형은 산점도 행렬이라고 할 수 있다. 산점도가 많이 활용되는 이유는 3차원 공간에 살고있는 사람이 가장 잘 파악할 수 있는 도형이 2차원이기 때문인 것으로 파악되고 있다 (실 제 사물의 형상은 3차원이지만 우리가 눈으로 보는 부분은 2차원 상에 투영된 형상을 보는 것으로 생각할 수 있다. 이는 카메라 렌즈를 통해 사물을 보는 경우를 생각하면 된다). 그러나 산점도는 자료의 크기가 커질 때, 중복되는 점들이 많이 나타나기 때문에 유용성에 한계를 갖고 있으며, 산점도 행렬은 변수의 수가 많아질 때 표현의 한계를 갖고 있다 (그림 1.2 참조). 또한 필자는 많은 사람들이 산점도 행렬로부터 자료의 성격을 파악하는 과정에 대해 어려움을 갖고있으며, 이를 파악하는 데는 훈련이 필요한 것을 경험하였다.

자료의 크기가 작으면 VC를 운영하는 데 별 문제가 없다. 반대로 자료의 크기가 커지면 컴퓨터 주기억용량의 문제와 계산시간이 많이 소요되기 때문에 자료의 변화에 따른 도형의 변화가 실시간에 이루어지기 어렵다. 그러면, 어떤 경우 자료의 크기가 작고, 또 크다고 할 수 있는가? 다음 두 가지를 생각해 보자.

1. 관측값의 수가 많은 경우
2. 변수의 수가 많은 경우

일차원 도형의 경우에는 관측값의 수에 따라 계산 속도가 결정된다. 그러나 막대그래프나 복스플롯과 같은 경우, 관측값의 수가 아무리 많다고 하더라도 수행시간이 많이 소요될 뿐이지, 도형을 그리고 이를 이해하는 데는 아무런 지장이 없다.

고차원 자료를 위한 도형의 경우, 관측값의 수 보다는 변수의 수에 민감해지는 도형이 많다. 특히 산점도 행렬의 경우 자료의 크기보다는 변수의 수에 따라 계산시간이 많이 소모되고, 또한 변수의 수가 많아지면 이 도형의 효용성이 없어진다. 그림 1.2의 산점도 행렬은 변수의 수가 47개이고 관측값의 크기가 234개인 인체자료에 대한 것이다. 관측값의 수가 n개이고 변수의 수가 p 개인 경우 산점도 행렬을 그리려면, $p(p-1)/2$ 개의 산점도에 n개 점의 위치를 계산하고 이를 찍어야 한다. 각 점의 위치는 (x, y) 한 쌍으로 나타나기 때문에 점의 위치만을 계산하는 데 $np(p-1)$ 개의 계산이 필요하다. 따라서 $n = 1,000$이고 $p = 10$이면 $90,000$ 개의 점을 위한 계산이 필요하여 $n = 1,000$이고 $p = 50$이라면 약 $2,500,000$ 개의 계산이 필요하다. 즉, 산점도 행렬을 만드는 데 자료의 수는 1차의 영향이 있지만, 변수의 수는 2차의 영향을 미친다.

평행좌표계는 변수의 수가 많아지거나 자료의 크기가 커지더라도 이를 모두 표현할 수 있다 (그림 1.3 참조). 물론 여기서도 자료의 크기가 커지면 중복된 선들이 나타나고 변수

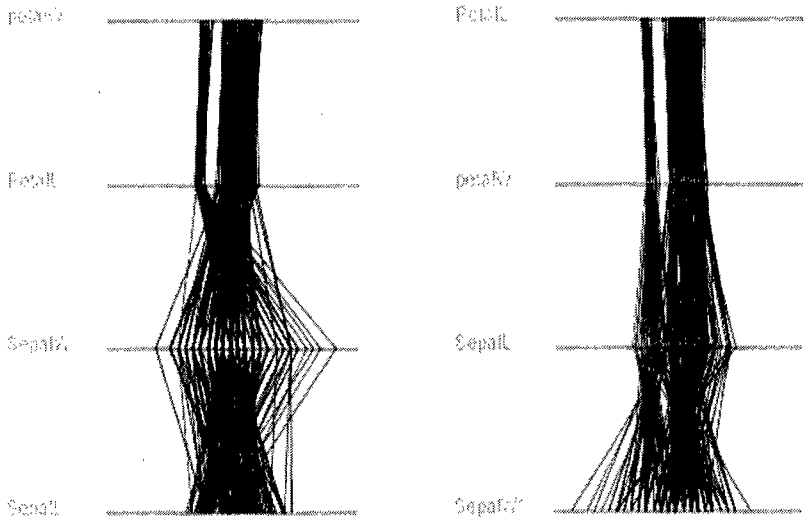

그림 2.1: Fisher 의 Iris 데이터를 평행좌표계로 표현하였다. 왼쪽은 원래 주어진 변수의 순서대로 그린 것이고, 오른쪽은 Huh (2000c)가 제시한 방법에 의해 변수를 재배열하여 그린 것이다. 오른편 그림을 참고하면, 자료가 두 개의 그룹으로 구분되는 것을 뚜렷이 알 수있다. 이 도형은 자료를 k-means 방법에 의해 두 개의 그룹으로 나누고 이를 두 가지 색깔로 표현한 것이다.

가 많아질수록 도형의 효용성이 떨어진다. 그러나 개개 관측값의 상대적 위치로부터 정보를 파악 하는 산점도와 달리, 평행좌표계는 연결된 선들의 집단적인 추이로부터 자료의 구조를 파악하는 데 그 목적이 있기 때문에 자료의 크기나 변수의 수에 대해 산점도 행렬만큼 민감하지는 않다. 평행좌표계에서 표현된 인근 두 변수의 연결로부터 산점도와 유사한 정보를 획득할 수 있다. 즉, 인근 변수들을 잇는 직선의 추이가 평행하다면 두 변수들은 매우 유사한 성격을 갖고 있으며, 이 추이가 엇갈린다면 음의 상관을 보인다고 할 수 있고, 아무런 추이를 찾을 수 없다면, 두 변수의 상관계수는 0에 가깝다고 할 수 있다.

평행좌표계나 산점도행렬에서 간과해서는 안 될 중요한 점은 변수의 배열 순서이다. 변수 를 어떻게 배열하는 가에 따라 이들 도형이 제공하는 정보는 매우 다를 수 있다 (그림 2.1). 통계패키지들이 제공하는 도형은 대이터베이스에 주어진 변수의 순서를 그대로 따르는 경우가 대부분이다. 이에 대해 Ankerst 등(1999)의 연구와 Huh(2001b)의 연구가 있다.

고차원 자료의 표현에서 특이할 만한 기술은 Asimov, Buja 등이 제안한 Touring 으로써 1985년 발표된 이후 지속적인 연구가 있었다 (Buja, Cook, Swayne, 1996; Buja, Cook, Asimov, Hurley, 1997). IBM Almaden 연구소의 Modha 등은 k-means에 의해 주어지는 각 군집의 중심을 이용하여 이차원 평면을 만들어 여기에 각 관측값을 투영하는 과정을

touring 함으로서 k-means의 효율성을 판단하는 cluster-guided tour를 제안하였다(1998).

Huh등은 Grand Tour를 변형시킨 Tracking Grand Tour를 제안하였다(2000, 2001a). Grand Tour는 기본적으로 고차원 자료를 2차원 공간상에 산점도의 형식으로 각 점들을 찍어주는 것이기 때문에 관측값의 수가 많아지면, 그만큼 수행시간이 많이 요구된다. 그러나 변수의 수가 많아지면 자료의 구조를 파악하는 데 소요되는 touring의 횟수는 급격히 증가한다. 예를 들어 3차원 구(sphere)의 구조를 파악하는 데 10회의 회전이 필요하였다고 하자. 이와 유사한 4차원 초구(hyper sphere)의 구조를 파악하려면, 약 40회의 회전이 필요하다고 할 수 있다. 이는 4차원 초구의 구조를 파악하기 위해 한 차원을 고정시켜놓고, 다른 3차원으로 이루어진 구를 회전한다고 생각하면 상상할 수 있다. 또한 10차원 초구의 경우 $10! = 3,629,800$회의 회전이 필요하다. 그러나 이는 구조가 가장 간단한 초구의 경우이지만, 일반적인 자료형태의 경우, 이 자료의 구조를 파악하기 위해서는 상상할 수 없을 만큼 많은 회전이 필요할 것이다. touring에 의한 다차원 구조를 파악하는 데 몇 회의 회전이 필요한가에 대해서는 이론적인 근거가 없지만, 변수가 47개인 인체 자료를 사용하여 실험하여본 결과 touring에 의해 자료의 구조를 파악한다는 것은 거의 불가능하다는 사실을 발견하였다.

자료 자체를 도형으로 표현하는 방법 이외에 덴드로그램 (dendrogram)과 같이 다차원 자료의 각 점들의 거리를 2차원 행렬로 변환하고, 이를 2차원 상에 표현하는 방법도 있다. 덴 드로그램을 사용하는 경우 관측값의 수가 최종적인 가지의 수가 된다. 여기서 변수의 수는 관측값들 사이의 거리를 측정하는 데 사용될 뿐이다. 따라서 최종적인 표현에 있어서는 변수의 수와 관계없이 관측값의 수에 따라 표현의 효용성이 결정된다. 덴드로그램으로 표현할 수 있는 자료의 한계는 이를 그리는 매체에 따라 결정된다 (그림 2.2). 물론 덴드로그램을 scrollbar 등을 이용하면 이론적으로는 크기가 아무리 큰 자료라도 그려낼 수 있다. JMP(1999)의 덴드로그램 등은 이런 방법을 이용한다. 그러나 덴드로그램을 한 화면에 표현하지 않고 scroll을 하면서 보게되면, 덴드로그램의 본래 목적을 달성할 수 없게된다. 이는 가지가 많은 나무의 구조를 파악하기 위해서 이를 한눈에 들어오게끔 관측하여야 하나, 나 무가 너무 큰 경우 일부분씩 잘라서 관측하면 가지들의 상대적인 위치를 판단하기 어렵기 때문에, 이 나무의 구조를 파악하기 어려운 것과 마찬가지이다. 덴드로그램을 그리는 데, 관측값의 크기가 n 이면, 거리행렬의 크기는 $n \times n$ 이다. 행렬의 계산에서 각 원 소를 배정도(double precision)로 하면, 하나의 숫자에 대해 8byte가 필요하고, 따라서 거리 행렬만을 위해 $8n^2$ byte가 필요하다. 따라서 1,000개의 관측값이 있다면, 이를 위한 거리행렬에 $8M$byte 의 메모리가 요구된다. 덴드로그램을 그리는 데는 주기억장치의 용량만이 문제가 되는 것은 아니다. 거리행렬에서 성격이 가장 가까운 두 객체를 골라야 하고, 이것이 선택되면 이를 이용하여 객체를 다시 정돈하고, 이렇게 정리된 객체들을 이용하여 다시 가장 가까운 두 객체를 고르게 된다. 이러한 과정은 모든 객체가 하나의 집단이 될 때 까지 평균적으로 $log_2 n + 1$회 동안 계속해서 진행된다. 따라서, 덴드로그램을 그리는 데는 대량의 메모리뿐만 아니라 엄청난 계산량이 필요하게 된다.

고차원 자료를 표현하는 또 하나의 방법으로 Becker 등이 제안한 Trellis Display 가 있으며 (1996), Cleveland 는 Trellis 도형을 그의 저서 *Visualizing Data*(1993)에서 매우 폭 넓

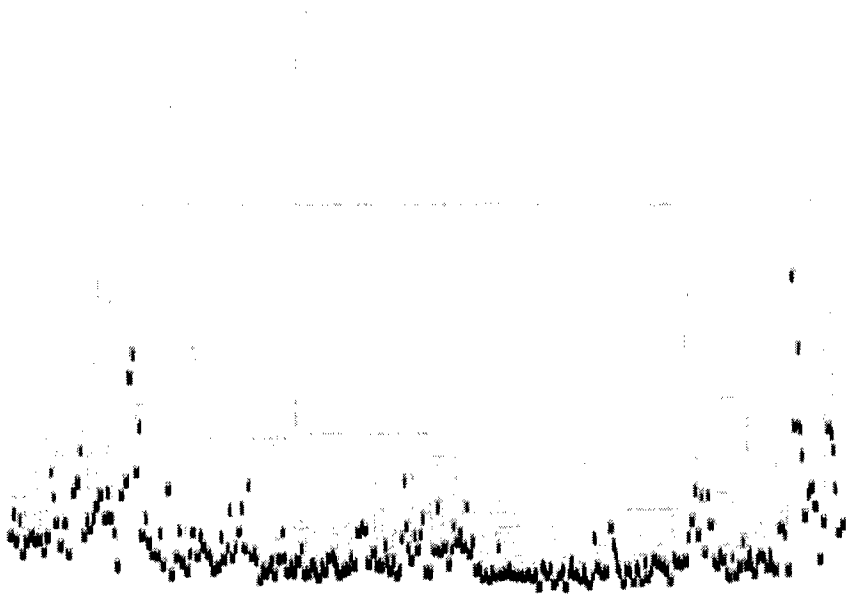

그림 2.2: 통계청의 95년 Census 자료 중에서 변수 5개, 400개 동을 랜덤하게 추출하여 덴드로그램으로 표현하였다. 거리행렬은 최단거리를, 군집방법은 평균거리 방법을 적용하였다. 여기서 볼 수 있는 바와 같이 관측값의 수가 이미 400개만 되더라도 각 관측값을 식별하는 것이 어려워진다.

게 활용하고 있다. 방대한 고차원자료의 시각화를 위해서는 Keim이 1996년 발표한 Pixel-Oriented 방법이 독특하다.

3. VC의 구현을 위한 전략

VC에서 가장 핵심적인 요소는 사용자와 시스템간의 통신이 실시간에 이루어지는 데 있다. 사용자와 시스템 간의 대화를 실시간에 처리하는 방법은 다음과 같다.

1. 데이터베이스에서 사용자가 직접 자료를 변화시키고, 이 변화가 도형에 미치는 영향을 조사한다. 예를 들어 데이터베이스에서 관심이 있는 변수들만 선택한 후 덴드로그램을 다시 그려봄으로써 새로운 변수의 선택이 군집분류에 미치는 영향을 탐색할 수 있다. 또는 데이터베이스에서 데이터의 일부분을 택한 후 (예: 남, 여 자료 중, 여자만 선택) 도형에서 이들의 상대적인 패턴이 어떻게 나타나는 가를 살펴볼 수 있다.

2. 도형에 나타나 있는 관측값 중 관심이 있는 부분을 사용자가 선택하고, 이렇게 선택된 부분이 다른 도형이나 데이터베이스에서 어떤 부분에 속하는 가를 조사한다. 이

방법은 주로 산점도나 touring에서 나타난 관측값들 중에서 관심이 있는 부분을 마우스를 이용하여 brushing하여 처리한다. 이러한 과정은 FEDF를 이용하면 더욱 효율적으로 처리할 수 있다. 예를 들어 한 변수의 중앙 50%에 해당하는 관측값들만 선택하는 과정이 그림 3.1에 나타나 있다.

3. 수학적 분석방법의 결과를 도형에 연결시킴으로써 수학적 분석결과의 유용성을 조사한다. 예를 들면, k-means 방법에 의해 자료를 2개의 그룹으로 나누고 이들 2개 그룹이 평 행좌표에 나타나는 결과를 살펴봄으로써 k-means에 의한 군집방법이 잘 되었는가를 조사 할 수 있다 (그림 1.3, 2.1 참조.)

사용자와 시스템 사이의 통신방법을 아무리 잘 설계 하더라도, 이를 처리해주는 컴퓨터의 속도가 느린 경우 시각적 효과가 없어진다. 그러나 이미 2절에서 언급한 바와 같이 VC를 위한 도형을 그리는 과정은 엄청난 양의 계산이 필요하므로 이를 효율적으로 처리할 수 있는 적절한 하드웨어와 소프트웨어의 선택이 매우 중요하다. 하드웨어의 경우 매우 빠른 속 도로 계산할 수 있는 컴퓨터가 필요하고, 정교한 도형을 표현할 수 있는 화면이 필요하다. 최근 컴퓨터 하드웨어 기술의 발달로 미루어 볼 때 VC를 위한 하드웨어 환경은 문제가 되지 않는 것으로 판단된다.

소프트웨어의 경우 통계계산과 그래픽을 효율적으로 처리해 줄 수 있는 도구를 찾아야 한다. 최근에는 많은 소프트웨어들이 훌륭한 GUI(Graphic User Interface) 기능을 갖고 있기 때문에 선택 폭이 넓다고 할 수 있다. 통계인들이 가장 먼저 생각할 수 있는 소프트웨어는 SAS, S-Plus, SPSS 등의 기존 패키지이다. 그러나 이들 패키지는 문자그대로 패키지이기 때문에 시각화를 위한 새로운 시스템을 구현하는 데 적절하지 못하다. C++ 등과 같은 언어 를 사용하는 것도 생각할 수 있다. XLISP-STAT 과 같은 언어는 기종에 무관한 언어이며 풍부한 통계계산 기능과 그래픽 기능이 있기 때문에 매우 유용한 개발도구가 될 수 있다. 그러나 이 언어는 사용자 그룹이 한정되어 있으며, 상용이 아니기 때문에 다량 데이터의 처 리, 결측값의 처리, 그래픽 정도 (high quality graphics) 등에서 한계를 갖고 있다. TCL/TK 도 생각할 수 있다. 이 도구는 기종 무관하고 훌륭한 그래픽 기능이 있으며 네트워크 상에서 applet으로 구현할 수 있는 장점이 있다. 그러나 결정적인 단점은 수행속도가 떨어지는 것이다. 이들을 종합해 보면 개발도구는 다음의 기능들을 갖추고 있는 것이 바람직하다.

1. 데이터 관리능력
2. 기종 무관할 것
3. 풍부한 그래픽기능이 있을 것
4. 풍부한 라이브러리가 있을 것
5. 많은 사람이 사용할 것
6. 수행속도가 빠를 것

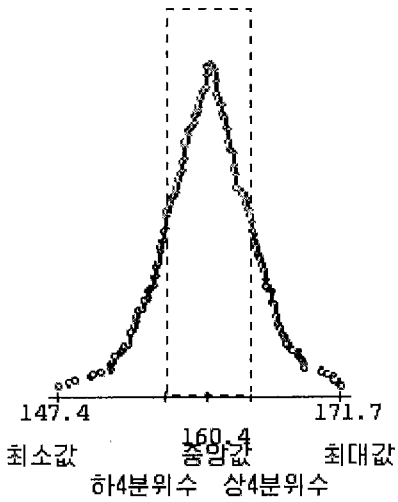

그림 3.1: 인체측정 데이터에서 234명 키의 FEDF. 중앙 50%에 해당하는 127명만 선택하려면 하 4분위수와 상 4분위수에 해당하는 부분을 그림과 같이 마우스로 선택하면 된다.

7. 네트워크 상에서 사용할 수 있고 인터넷 관련 소프트웨어들과 호환 가능 할 것

이상의 내용을 감안할 때 가장 바람직한 것은 JAVA라 할 수 있다. 아직도 JAVA는 개발 단계에 있기 때문에 빠른 속도로 언어의 기능이 변하고 있다. 또한 많은 사람들이 이를 이용하고 있으며, 훌륭한 라이브러리가 무료로 제공되고 있고, 기종에 무관할 뿐만 아니라 네트워크 상에서 applet으로 실행 가능하기 때문에 특히 통계교육을 위한 소프트웨어의 개발을 위해서는 훌륭한 도구라고 할 수 있다.

4. 맺는 말

우리는 요즈음 주위에 수많은 소프트웨어들의 홍수 속에서 살고 있다. 이들의 유용성은 결국 얼마나 많은 사용자가 이들을 사용하고 있는가에 있다. 이는 단지 상업용 소프트웨어뿐만 아니라 연구용 소프트웨어에서도 마찬가지이다. 시각적 군집분석도 그 결과가 결국 소프트웨어로 나타나기 때문에 예외가 아니다. 소프트웨어가 성공하려면 다음과 같은 3가지 요소가 유기적으로 고려되어야 한다.

대부분 통계 소프트웨어를 작성하는 사람들은 나름대로 뚜렷한 목표를 갖고 있다. 그러나 사용자에 대한 고려를 하지 않는 경우가 많다. 이는 특히 우리 나라와 같이 시장이 작은 환경에서 심각히 고려해야 할 요소라고 생각한다. 따라서 가능하면 세계 전체를 대상으로

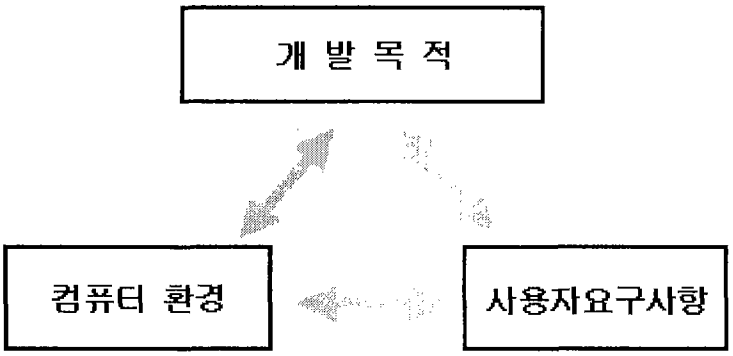

그림 4.1:

하 는 것이 바람직하다고 할 수 있다. 그리고 무엇보다 개발환경도 중요하다. 날로 발전해 가는 컴퓨터 하드웨어와 소프트웨어의 환경을 고려해야 한다. 예를 들어 몇 년 전만 해도 DOS가 시장을 점령하였다. 만약 DOS를 기준으로 하여 소프트웨어를 작성하였다면 이제는 쓸모없는 것이 되고 말았을 것이다. 구현 도구는 가능한 한 기종에 무관한 것을 택해야 한다. 미래를 고려할 때, 네트워크 환경에서 운용될 수 있는 소프트웨어가 바람직하다.

본 논문에 나타난 모든 도형(그림 1.1, 3.1은 제외)은 본인과 관련 연구팀이 개발한 SIVON (System for Information Visualization On the Network)을 이용하여 그린 것이다. 이 시스 템은 JAVA로 작성하였으며, 다음 web site에서 down 받아 실험할 수 있다.

http://stat.skku.ac.kr/~myhuh/software/Sivon/applet/SIVON.html

참고문헌

[1] Ankerst, M., Berchtold, S. and Kleim, D. (1999). "Similarity Clustering of Dimensions for an Enhanced Visualization of Multidimensional Data". *Visualization*, Vol. 99, 52-60

[2] Asimov, D. (1985). "The grand tour: A tool for viewing multidimensional data". *SIAM Journal on Scientific and Statistical Computing*, Vol. 6, No. 1, 128-143

[3] Asimov, D. and Buja, A. (1985). "Grand tour methods: an outline". *Proceedings of the 17th Symposium on Interface of Computer Science and Statistics.*

[4] Becker, R.A., Cleveland, W.S. and Shyu, M. (1996). "The Visual Display and Control of Trellis Display", *Journal of Computational and Graphical Statistics*, Vol. 5, No. 2, 123-155

[5] Buja, A., Cook, D. and Swayne, D.F. (1996). "Interactive High-Dimensional Data Visualization", *Journal of Computational and Graphical Statistics*, Vol.. 5, No. 1, 78-99.

[6] Carr, D.B., Wegman, E.J. and Luo, Q. (1997a). "ExplorN: Design Considerations Past and Present," Technical Report 137, *Center for Computational Statistics*, George Mason University, Fairfax, VA.

[7] Cleveland, W.S. and McGill, M.E., eds (1988). *Dynamic Graphics for Statistics*, Wadsworth & Brooks/Cole, Belmont, CA.

[8] Cleveland, W.S. (1993). *Visualizing Data*, Hobart Press, Summit, New Jersey

[9] Cook, D., Buja, A. and Cabrera, J. (1993). "Projection Pursuit Indices Based on Expansions with Orthonormal Functions", *Journal of Computational and Graphical Statistics,* Vol. 2, No. 3, 225-250.

[10] Dharmendra S. Modha, Scott Spangler and Shivakumar Vaithyanathan (1998). "Multidimensional Cluster Visualization using Guided Tours ", *net.Mining*, GBIS/NA, IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120-6099

[11] Huh, Moon Yul (1995). "Exploring Multidimensional Data with FEDF", *Journal of Computational and Graphical Statistics*, Vol. 4, No. 4, 335-343

[12] Huh, Moon Yul (2000). "Developing a Software Runnable on the Internet: Is This Necessary?", *Japanese Society of Computational Statistics*, Kakoshima,Japan, May 25-26, 111-114

[13] Huh, Moon Yul, Kim, K.Y. and Buja, Andreas (2001a). "Visualization of Multidimensional Data Modifications of the Grand Tour", 게재예정, Journal of the Royal Statistical Society: Series C (Applied Statistics)

[14] Huh, Moon Yul (2000c). "Variable Arrangement for Visual Clustering", 발표예정, Workshop on Statistical Techniques in Data Mining with Applications, The Fifth Pacific-Asia Conference on Knowledge Discovery and data Mining (PAKDD 2001), 2001년 4월 16일, 홍콩

[15] *JMP* (1999). *Statistical Discovery Software from SAS*, 유니컨설팅

[16] Lee, K.M., Kim, K.Y., Huh, M.Y. and Jeong, N.C. (1997). "Dynamic Graphics Approach for Cluster Analysis", *Computing Science and Statistics*, 28, 385-388.

[17] Kleim (1996). "Pixcel-Oriented Visualization Techniques for Exploring Very Large Data Bases", *Journal of Computational and Graphical Statistics,* Vol 5, Number 1, March 1996.

[18] Swayne, D.F., Cook, D. and Buja, A. (1998). "XGobi: Interactive Dynamic Graphics in

the X Window System", *Journal of Computational and Graphical Statistics*, Vol. 7, No. 1, 113-130.

[19] Unwin, A.R., Hawkins, G., Hofmann, H. and Siegl, B. (1996). "MANET: Missings Are Now Equally Treated", *Journal of Computational and Graphical Statistics*, Vol. 5, No. 2, 113-122.

[20] Wegman, E.J. and Luo, Q. (1997). "High Dimensional Clustering using Parallel Coordinates and the Grand Tour", *Computing Science and Statistics*, 28, 361-368.

[21] Wegman, Edward J. (1999). "Data Mining and Visualization", it Bulletin of the International Statistical Institute, ISI 99, Proceedings Book 3, pp 223-226, Helsinki, Finland

[2000년 7월 접수, 2000년 11월 채택]

Strategy for Visual Clustering *

Moon Yul Huh[1)]

ABSTRACT

Conventional clustering methods are to group data objects according to a selected agglomerative method based on a specific distance measure. The results depend much on the methods we are applying, hence it becomes inevitable that we lose much of the information underlying the data when we are using a specific index of distance measure to apply a selected aglomerative method. Visual clustering tries to use all the information of the data to understand the underlying structure. In this paper, we will investigate the motivation, the use of statistical graphics, and the implementation strategy for visual clustering.

Keywords: Data visualization, Clustering, Scatterplot matrix, Parallel coordinates, Dendrogram.

* The research is partially supported by Korea Research Foundation 19980364-000.

1) Professor, Dept. of Statistics, Sungkyunkwan University.
E-mail: myhuh@skku.ac.kr