The Korean Communications
in Statistics Vol. 8, No. 3, 2001
pp. 643-650

Variable Arrangement for Data Visualization

Moon Yul Huh[1], Kwang Ryeol Song[2]

Abstract

Some classical plots like scatterplot matrices and parallel coordinates are valuable tools for data visualization. These tools are extensively used in the modern data mining softwares to explore the inherent data structure, and hence to visually classify or cluster the database into appropriate groups. However, the interpretation of these plots are very sensitive to the arrangement of variables. In this work, we introduce two methods to arrange the variables for data visualization. First method is based on the work of Wegman (1999), and this is to arrange the variables using minimum distance among all the pairwise permutation of the variables. Second method is using the idea of principal components. We investigate the effectiveness of these methods with parallel coordinates using real data sets, and show that each of the two proposed methods has its own strength from different aspects respectively.

Keywords : variable arrangement, parallel coordinates, scatterplot matrix, clustering, principal component

1. 서 론

산점도는 2차원 자료에 대한 모형설정에 있어서 기본적인 그래프로서 폭넓게 사용되고 있다. 3차원 이상의 자료에 대해서는 Becker, Cleveland, Wilks(1987) 들이 산점도 행렬이라고 알려진 사각형 배열에서 모든 변수쌍들의 산점도들을 배열하는 동적 그래프 접근방법을 사용하였다. 또한 산점도 행렬에 대한 동적 그래프 기법을 적용한 방법은 Tierney(1990)에 의해서도 연구되었다. 이러한 연구 후에 S-Plus, SAS, STATISTICA, SPSS 같은 잘 알려진 통계 프로그램의 대부분은 다변량 자료의 탐색에 있어서 이 그래프를 구현하였다. 이 그래프는 변수의 수가 작을 때 효율적이다. 그러나 변수들의 수가 증가할 때 그래프의 강력함은 감소한다. 고차원 자료를 보여주기 위한 이해하기 쉬운 다른 도구는 평행 좌표그래프가 있다. 이 그래프는 많은 변수들을 탐색할 때 산점도 행렬보다 상대적으로 좀더 효율적이다. 최근에 Kensington(2001) 같은 데이타 마이닝 도구들은 자료의 본질적인 구조를 탐색하기 위해 이 방법을 사용하였다. 평행좌표그래프를 효과적으로 사용하기

1 Professor, Department of Statistics, Sungkyunkwan University, Seoul, Korea,
E-mail : myhuh@skku.ac.kr
2 Department of Statistics, Sungkyunkwan University, Seoul, Korea,
E-mail : skr9697@ecostat.skku.ac.kr

위한 연구는 Wegman(1990, 1996, 1999)등에 의해서도 수행되어 왔다.

이들 그래프가 제공하는 시각적인 정보는 변수들의 배열에 많이 의존한다. 그러나 그래프를 실제 그릴 때 변수들의 배열은 데이터베이스에 주어진 자료의 변수배열에 의해 결정되거나 그래프를 그리는 사람의 주관에 의해, 즉 임의의 결정에 의해 이루어진다. Ankerst, et. al.(1999) 는 최적 변수 배열의 문제는 하드 문제 (hard problem)인 것을 보였다. 즉, 최적 배열 문제는 NP-완전인 여행 세일즈맨 문제(Travelling Salesman Problem)와 같다는 것을 보여주었다. 여기서 최적배열이란 여러 가지의 배열 중에서 비유사성 (disimilarity)를 최소화하는 것을 의미한다. Dorgio 등은 (1997) 이 문제를 발견론적으로 해결하는 하나의 방법으로 개미가 목표지점을 찾아가는 논리를 적용하였으며 이를 ant-colony system 이라고 한다.

본 논문에서 우리는 변수들을 배열하는 새로운 방법들을 제시한다. 첫 번째 방법은 각 변수가 다른 모든 변수와 인접되는 변수들의 배열들을 생각하고 여기서 거리가 최소인 배열을 택한다. 예를 들어 5 개의 변수를 (1,2,3,4,5) 라고 표현할 때, (1,2,5,3,4), (2,3,1,5,4), (3,4,2,5,1) 로 주어지는 3 가지 배열은 각 변수가 다른 모든 변수와 인접되도록 배열한 조합이다. 이러한 접근방법은 Wegman(1990) 과 Ankerst (1999) 가 연구한 내용을 확장한 것이다. 이 방법을 PM(Permutation Method)이라 부르기로 한다. 두 번째 방법은 배열의 첫 번째에 자료의 '가장 큰 정보'를 갖는 변수가 나타나는 방식으로 변수들을 재배열하는 것이다. '가장 큰 정보'를 정의하기 위하여 여기에 주성분 개념을 사용하였다. 이 방법을 CM(Component Method)이라 부르기로 한다.

본 논문에서는 이상의 두 가지 방법을 평행좌표그래프에 적용하고 실제 자료를 사용하여 이들의 효율성을 알아보고자 한다. PM 방법의 경우 변수들 간의 구조적 정보에는 관심이 없고 단순히 최소 거리를 갖는 배열을 택하기 때문에 비합리적인 결과를 가져오는 경우가 많았다. 이 방법에 비해 CM을 적용한 결과 매우 합리적인 결과를 보여주었다. 이 들 방법의 효율성을 평가할 때 변수를 각 방법에 의해 배열하고 이를 그래프로 표현한 후, 여기에 k-means 방법과 같은 수학적 군집 기법에 의해 자료를 군집으로 나누고 이 결과를 해당 그래프에 적용시켰다. 다음 두 절에서 우리는 두 가지 방법에 대한 알고리즘을 제시하고 두개의 실제 자료를 사용하여 제시된 방법들의 효율성을 경험적으로 검증하고자 한다.

2. 순열법

평행좌표그래프를 그리기 위해 Wegman(1990)은 각 각의 변수가 다른 모든 변수와 최소 한번 이상 인접해서 만날 수 있게 해 주는 순열을 발생시키는 배열에 대한 간단한 형식을 제공하였다. $v_i^{(j)}$ 를 i 번째 변수에 대한 j 번째 배열이라고 하자. p 개의 변수들에 대한 j 번째 배열에 대한 공식은 다음과 같다.

$$v_i^{(j+1)} = (v_i^{(j)}+1) \bmod p,$$

$$\text{여기서 } j=1,\ldots,[\frac{p+1}{2}]\text{이고 } v_i^{(1)}=v_i,\quad v_1=1,$$

$$v_{i+1}=[v_i+(-1)^{i+1}i] \bmod p,\ i=1,2,\ldots,p-1,$$

$$0 \bmod p=p \bmod p=p,\quad x \bmod p=(p+x) \bmod p,\ \text{if } x<0.$$

이며 [·] 는 · 를 넘지 않는 가장 큰 정수이다. 예를 들어, 변수의 수가 4개 인 경우, 이 방법에 의하면 {1,2,4,3}, {2,3,1,4} 와 같은 2 개의 배열이 만들어진다. 이 배열을 보면, 변수 1은 첫 번째 배열에서 2를 인접해서 만나고, 두 번째 배열에서 3 과 4를 인접해서 만난다. 또 변수 2의 경우, 첫 번째 배열에서 변수 1과 4를 을 인접해서 만나고, 두 번째 배열에서 변수 3을 만난다.

이제 변수 배열 문제를 각 관측값이 모든 변수를 한번씩 거쳐 지나가는데 이렇게 지나가는 데 소요되는 거

리 중에서 가장 작은 것을 고르는 것이라고 하면 이 문제를 다음과 같이 정의할 수 있다.

$$\text{minimize} \quad \sum_{i=1}^{p}\sum_{j=1}^{p} n_{ij} \cdot D_{ij} = 2\sum_{i=1}^{p-1}\sum_{j>i}^{p} n_{ij} \cdot D_{ij} \quad \ldots \ (1)$$

여기서 D_{ij}는 변수 i 와 j 사이의 거리, 즉 비유사성 (dissimilarity)이다. 그리고

$$n_{ij} = \begin{cases} 1 & \text{변수 } i \text{ 와 } j \text{ 가 이웃이면} \\ 0 & \text{그외의경우} \end{cases}$$

비유사성에 대해서는 유클리드 거리함수를 사용한다. 이 연구에서는 다음과 같은 척도 불변 측도(scale invariant measure)를 사용한다.

$$D_{ij} = \sqrt{\sum_{k=1}^{n}(b_{ki} - b_{kj})^2}$$

여기서

$$b_{ki} = \frac{x_{kj} - Min_i(x_{ij})}{Max_i(x_{ij}) - Min_i(x_{ij})}, \quad j = 1, \ldots, n, \quad k = 1, \ldots, p$$

이고 x_{ij}는 j번째 변수에 대한 i번째 관측값이다. 비유사성에 대한 다른 측도는 Kaufman 과 Rousseeuw(1989)에서 찾아볼 수 있다.

순열법 (permutation method, PM)은 p 개의 변수를 Wegman 이 제시한 방법에 의해 $\left[\frac{(p+1)}{2}\right]$ 개의 다른 방법으로 배열하고 이 중에서 (1) 식에 의해 한 개를 고르는 방법이다. 이 방법에 의해 만들어진 배열은 여러 개의 순열 배열 중에 변수 간 거리의 합을 가장 작게 만들어 주는 배열이므로 이는 상관이 깊은 변수들끼리 묶여있는 결과를 가져올 것이다. 다만 고려하는 변수의 배열 종류가 전체 모든 가능한 경우가 아니고 $\left[\frac{(p+1)}{2}\right]$ 개로 한정되어 있지만, 각 변수의 바로 옆에 어떤 변수가 나타나는 것이 가장 효율적인가를 판단할 수 있기 때문에 논리적으로 타당한 알고리즘이다.

3. 주성분법

이 알고리즘의 기본 개념은 자료에서 '가장 큰 정보'를 갖는 변수가 처음 나타나는 방식으로 변수들을 재배열하는 것이다. 그러한 변수가 선택되면 이는 변수 배열의 첫 번째 자리에 위치시키는 것이다. 다음 변수는 첫번째 선택된 변수를 제외한 나머지 변수들 사이에서 가장 큰 정보를 갖는 변수를 선택하는 것이다. 변수의 정보 측도는 자료로부터 획득한 고유값 들 중에서 가장 큰 고유값에 해당되는 주성분을 택하고, 이 주성분에 제공하는 로딩(loading)을 사용한다. 구체적인 알고리즘은 다음과 같다.

$\boldsymbol{X}_{n\times p}$를 n개의 관측값과 p개의 변수를 갖는 자료라고 하자. 이 자료로부터 변수들을 재 배열하는 과정은 다음과 같다.

단계 1. 자료 행렬 $\boldsymbol{X}$로부터 얻어진 유사성 행렬의 가장 큰 고유값에 대응하는 주성분 $\mathbf{u}$를 구한다.

단계 2. $\mathbf{u}$의 가장 큰 절대값에 대응하는 인덱스 k를 얻는다. 이를 가장 큰 정보를 갖는 변수에 해당하는 인덱스로 지정한다.

단계 3. 자료 행렬 $\boldsymbol{X}$에서 k번째 열을 제거함으로서 자료를 축소한다.

단계 4. 변수들이 더 이상 선택되지 않을 때까지 위의 세 단계를 적용한다.

여기서 우리는 유사성의 척도로서 상관행렬을 사용한다. 다른 유사성의 척도들도 생각할 수 있다 (Kaufman 과 Rousseuw, 1989).

R.A. Fisher(1936)에 의한 붓꽃 자료를 예를 들어 설명하기로 한다. 자료는 3개의 품종 Iris Setosa, Iris Versicolor, Iris Virginica에 대해서 50개씩, 총 150개의 관측값으로 이루어져 있다. 자료는 UCI(2001)에서 제공된 것이며 여기서는 품종 변수까지 5개로 이루어져 있으나, 여기서는 품종 변수는 빼고 sepal-length (꽃받침-길이), sepal-width (꽃받침-너비), petal-length (꽃잎-길이), petal-width (꽃잎-길이) 의 4 개 변수만 고려한다.

4. 방법들의 평가

여기서는 두 가지 실제 자료를 사용하여 제안된 방법들의 효율을 평가하고자 한다. 첫 번째 자료는 이미 앞에서 설명한 붓꽃 자료이고, 두 번째 자료는 UCI에서 얻어진 1985년 Auto Imports Database 이다. 이 자료는 205개의 관측값과 26개의 변수들로 구성되어 있다. 본 연구에서는 편의상 이들 26개의 변수들 중에서 make, fuel type, engine location 등 과 같은 11개의 명목변수들은 제외하고, 나머지 15개의 변수들만 선택하였다. 선택된 15개의 변수들은 wheel-base, length, width, height, curb-weight, number-of-cylinders, engine-size, bore, stroke, compression-ratio, horsepower, peak-rpm, city-mpg, highway-mpg, price 이다.

그림 1은 붓꽃 자료를 변수 배열한 결과를 평행좌표계로 나타낸 것이다. 이 그림의 각 선은 하나의 붓꽃을 나타내며 이 붓꽃이 속해있는 품종은 다른 색깔로 표현되었다. 그림 1 (a)는 원래 데이터베이스에 주어진 순서에 따라 변수 배열한 것이고, 그림 1 (b)는 PM 방법에 의해 배열한 것, 그리고 그림 1 (c)는 CM 방법에 의해 배열한 것이다. 붓꽃 자료는 setosa, versicolour, virginica 의 3 품종으로 분류된다. 그림 1(a)을 보면 전반적으로 도형이 체계적으로 작성되지 않은 것을 알 수 있다. PM 방법에 의해 변수를 배열한 결과 앞 절에서 예를 든 두 개의 배열 {1,2,4,3}, {2,3,1,4} 중에서 후자가 선택되었다. 즉, sepal-width, petal-length, sepal-length, petal-width 의 순서가 선택되었으며 결과가 그림 1(b)에 나타나있다. 이 그림을 보면 원래 자료에 비해 매우 조직적으로 배열되어있는 것을 알 수 있다. 즉, 이 배열에는 매우 높은 상관을 갖는 3 개의 변수들 (petal-length, sepal-length, petal-width) 이 근접해 있는 것을 알 수 있다. 그림 1(c)는 CM 에 의한 배열이다. CM에 의한 배열은 데이터의 정보를 가장 많이 갖고있는 변수부터 배치한다고 하였다. 붓꽃 데이터의 경우, petal-length 가 가장 처음 선택된 변수이며, 이 변수는 3 개의 품종을 정확하게 구분해 주는 것을 알 수 있다. 다음에는 petal-width, sepal-length 의 순서로 배치되었으며 변수배치 순서가 붓꽃의 품종을 가르는데 기여하는 정도인 것을 곧 알 수 있다. 이상을 정리하면, PM 방법은 같은 성격을 갖는 변수끼리 한데 묶어 주고, CM 방법에 의해 배열하면 데이터가 갖고 있는 구조를 처음 몇 개의 변수로부터 찾아낼 수 있는 장점이 있다.

붓꽃 자료의 경우 변수의 수가 4개였다. 자동차 자료와 같이 변수의 수가 많은 경우 변수 배열에 따른 시각적 효과를 살펴보기로 한다. 그림 2 는 자동차 자료를 평행좌표계로 표현한 것이다. 그림 2(a)는 UCI 데이터베이스에 주어진 원래 자료에 의한 것이고, (b)와 (c)는 각 각 PM 방법과 CM 방법에 의해 나타낸 것이다. 변수 중 highway-mpg 와 city-mpg 는 각각 $\frac{1}{\text{highway}-\text{mpg}}$ 와 $\frac{1}{\text{city}-\text{mpg}}$ 로 변환하였다. UCI에 주어진 변수들의 순서는 자동차의 특성에 연관되어 있는 것처럼 보인다. 첫 번째 5개의 변수는 자동차의 외부 차원과 관계되어 있고 다음 7개의 변수는 엔진과 관련되어 있는 변수들이고 마지막 3개의 변수는 자동차의 경제성과 관련되어 있다. 이들 순서는 그것 자체로 관심이 있을 수 있다. 그러나 우리의 목적이 시각적 방법에 의해

자동차를 적절한 그룹으로 분류하는 것이라면 원 자료에 주어진 배열을 사용하는 것은 자동차의 구조를 탐색하기에 효율적일 수 없다. 이를 위해 자료를 k-means 방법에 의해 3개의 그룹으로 나누고 이러한 군집 정보를 3개의 다른 색깔을 이용하여 자료에 적용시켰다. 이들 군집 결과를 해석하는 데 본 논문에서 제시한 평행좌표계가 어떻게 사용되는 가를 살펴보기로 한다.

그림 2(a)의 경우 이미 앞에서 설명한 바와 같이 데이터를 3개의 군집으로 나눈 정보가 도형에 산만하게 흩어져 있어 각 그룹이 제공하는 특성을 알아보기 어렵다. 그림 2(b)를 살펴보면, 가운데에 위치한 length, curb-weight,...city-mpg 의 9개의 변수들이 고르게 3개의 그룹에 대한 정보를 제공하고 있는 것을 알 수 있다. 즉, length 가 긴 것은 curb-weight, wheel-base, ..., city-mpg(실제 이 값은 $\frac{1}{\text{city}-\text{mpg}}$ 이다.) 등 모든 변수의 값이 크게 나타나는 것을 알 수 있다. 즉, 자동차 자료를 PM 방법으로 표현한 결과 앞의 붓꽃 자료에서와 같이 상관이 높은 변수들끼리 묶여서 나타나는 현상을 볼 수 있다. 그림 2(c)에 나타난 CM에 의한 평행좌표계는 15개의 변수들 중에 처음 나타난 7개의 변수(curb-weight, engine-size, highway-mpg, width, price, length, city-mpg)들이 자료의 구조를 시각적으로 탐색하는데 있어서 중요한 역할을 하는 것을 알 수 있다. 즉, 이들 7 개 변수들은 수입자동차를 그룹화 하는 데 중요한 변수들이며 k-means 에 의한 군집방법은 자동차의 몸집과 (무게, 길이, 엔진 사이즈) 가격, 그리고 연비에 의해 3 개의 그룹으로 나누는 것을 알 수 있으며 이 외의 다른 변수들 (자동차의 높이, 엔진 rpm 등 기술적인 요소)은 자동차들을 그룹 짓는 데 중요한 역할을 하지 않는 다는 것을 시각적으로 알아볼 수 있다.

5. 결 론

이 연구에서 우리는 시각적 방법에 의해 자료를 적절한 그룹으로 분류할 때 변수의 배열이 중요한 것을 보였으며, 효율적인 변수배열방법을 제안하였다. 여기서 고려한 두 방법 중, Wegman 이 제안한 방법에 기본을 둔 PM 에 의해 변수배열을 하면, 상관이 높은 변수들끼리 묶어서 나타나는 것을 알 수 있었으며, CM에 의한 변수배열 방법은 자료의 구조를 고려하여 가장 설명도가 높은 변수부터 배치하기 때문에 처음 몇 개의 변수만 참고하여 자료의 구조를 파악할 수 있는 것을 알았다. 우리는 이 방법을 평행 좌표계에 적용하였다. 그러나 이 방법은 변수 배열이 필요한 어느 그림에도 적용될 수 있을 것이다. 예를 들어 산점도 행렬 등에도 적용할 수 있으며 이 결과는 본 논문에서 지면상 제외시켰으며 구체적인 내용은 아래 주어진 홈페이지를 참고하면 된다. 다만, CM 방법을 적용하려면 변수의 수가 p 개인 경우 최대 $p!-2$ 회의 고유값과 고유벡터 계산이 필요하게 된다. 앞으로 이 계산을 줄이는 방법을 연구해야 할 것이다. 여기서 사용한 모든 그림은 JAVA로 프로그램이 되어있으며, 이는 다음과 같은 홈페이지에서 다운 받아 사용할 수 있다.
http://stat.skku.ac.kr/~myhuh/software/DAVIS/DAVIS.htm

참 고 문 헌

[1] Ankerst, M., Berchtold, S., Keim, D. (1999), Similarity Clustering of Dimensions for an Enhanced Visualization of Multidiemsnional Data, Visualization, 99, pp. 52-60

[2] Becker, Richard A. Cleveland, Cleveland, William S.,Wilks, Allan R., (1987), Dynamic graphics for Data Analysis, *Statistical Science*, Vol. 2, pp. 355-359

[3] Dorgio, M., Gambardella I.M., (1997), Ant Colony System: A Cooperative Learning Approach to the Travelling Salesman Proble, *IEEE Trans. on Evolutionary*

Computation, Vol 1, No. 1

[4] Fisher,R.A. (1936), The Use of Multiple Measurements in Taxonomic Problems, *Annual Eugenics*, 7, Part II, 179-188

[5] Kaufman, L., Rousseeuw, P., (1989), *Finding Groups in Data*, John Wiley & Sons, Inc., New York

[6] Kensington (2001), http://www.inforsense.com

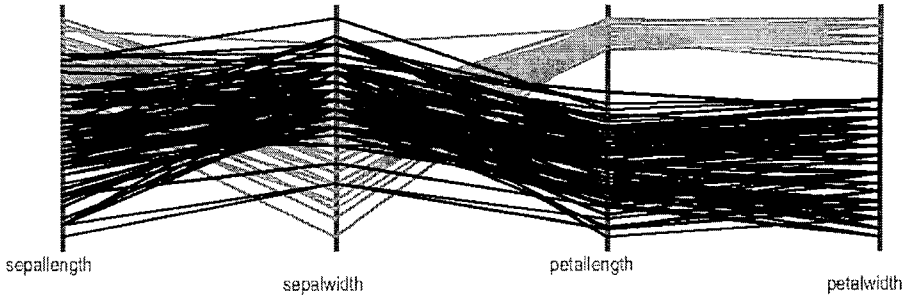
[7] Tierney, Luke (1990), *LISP-STAT* ,John Wiley & Sons, Inc., New York

[8] UCI (2001), http://www.ics.uci.edu/~mlearn/MLRepository.html

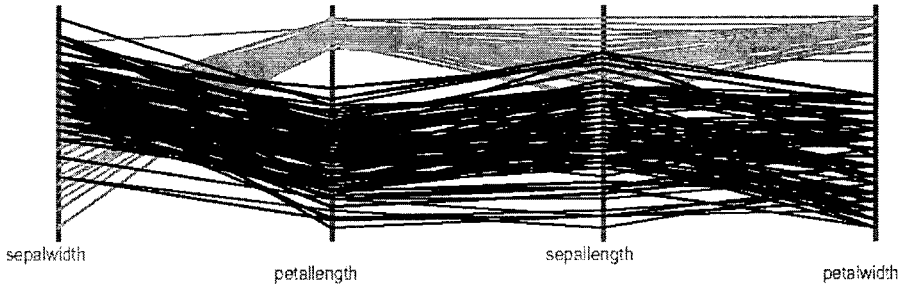
[9] Wegman, Edward J., (1990), Hyperdimensional Data Analysis Using Parallel Coordinate, *Journal of American Statistical Society*, Vol. 85, N0. 411, pp. 664-675

[10] Wegman, Edward J., Luo Qiang, (1996), High Dimensional Clustering Using Parallel Coordinates and the Grand Tour, *Computing Science and Statistics:, Proceedings of the 28th Symposium on the Interface*, Sydney, Australia, pp. 361-368, (Lynne Billard and Nicholas Fisher Ed.) Interface Foundation of North America

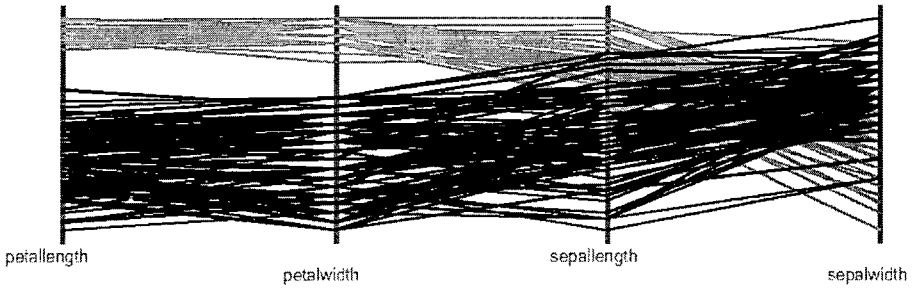
[11] Wegman, Edward J., (1999), Data Mining and Visualization, Bulletin of the International Statistical Institute, ISI 99, Proceedings Book 3, pp 223-226, Helsinki 1999



(a) UCI 데이터베이스에 주어진 변수의 순서

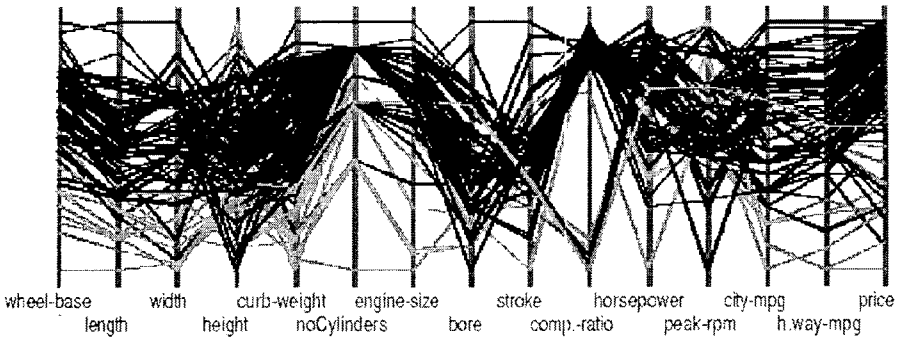


(b) PM 방법에 의한 변수 배열

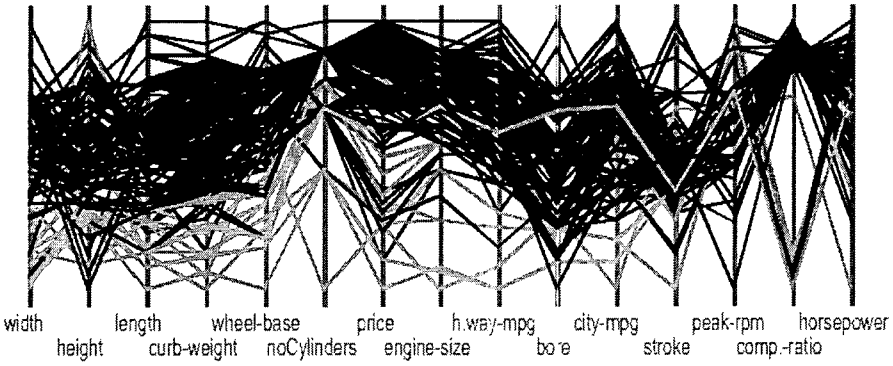


(c) CM 방법에 의한 변수배열

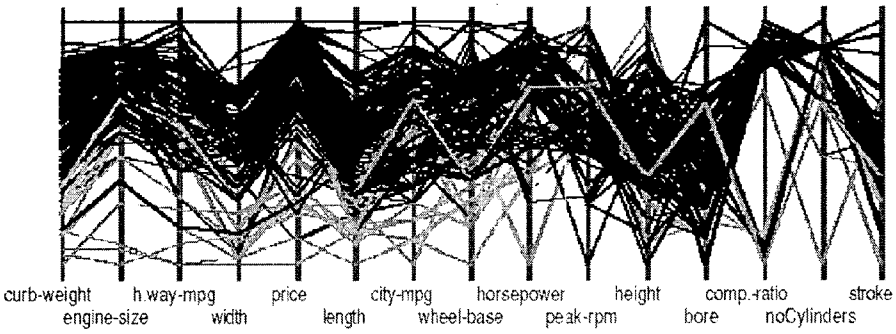
그림 1. UCI 데이터베이스에 주어진 붓꽃자료를 PM 방법과 CM 방법을 적용하여 변수를 재 배열하고 평행좌표계로 나타낸 결과



(a) UCI 데이터 베이스에 주어진 변수의 순서



(b) PM 방법에 의한 변수배열



(c) CM 방법에 의한 변수배열

그림 2. UCI 데이터베이스에 주어진 수입자동차에 대해 PM 방법과 CM 방법을 적용하여 변수를 재 배열하고 평행좌표계로 나타낸 결과