한국통계학회논문집
2010, 17권, 2호, 153-164

선 모자이크 도표를 이용한 동적 그래픽스

차운옥[1,a], 이경미[b], 최병수[a]

[a]한성대학교 멀티미디어공학과, [b]서일대학교 인터넷정보과

요 약

본 논문에서는 이산형과 연속형 데이터가 혼합되어 있는 데이터 구조를 탐색하기 위하여 동적 기법을 사용하였다. 이산형 변수들간의 관계를 표현하는 선 모자이크 도표와 연속형 변수들의 관계를 위한 산점도, 일변량 변수 관점에서의 데이터의 분포를 파악할 수 있는 상자도표를 동시에 사용하면서, 동적인 기법들을 적용하여 다차원 데이터에 대한 구조를 좀 더 쉽게 파악할 수 있음을 보였다.

주요용어: 선 모자이크 도표, 동적 그래픽스.

1. 서론

최근 많이 연구하고 있는 데이터마이닝의 통계적 기법 중 데이터 시각화(data visualization)는 데이터 속에 내재되어 있는 정보를 용이하게 파악할 수 있는 방법이다. 데이터 시각화는 데이터로부터 유용한 지식을 추출하기 위해 데이터에 자동화 기술을 적용하는 과정과 정보를 발견하는 과정이라고 정의할 수 있다 (Nicholas, 1999). 이 방법에서는 데이터를 컴퓨터 그래픽으로 표현해 주며 데이터와 분석가 사이의 상호작용이 가능하기 때문에, 데이터의 크기나 변수의 수가 많은 대용량 데이터를 탐색하는 경우 기존의 전통적 데이터 분석기법보다 훨씬 우수한 방법이라 할 수 있다.

데이터 시각화에서 동적 그래픽스(dynamic graphics)의 기본적인 원리는 데이터 컨디셔닝(data conditioning)으로서, 데이터의 일부를 선택(select)하거나 삭제(delete), 집중(focus)하여 이 결과가 통계적 모형이나 도형에 미치는 영향을 분석하는 것이다. 데이터 컨디셔닝은 통계적 도형 위에서 컴퓨터의 마우스를 이용하여 데이터의 일부를 하이라이트(highlight) 또는 브러싱(brushing)하고 이를 다른 도형과 링크함으로서 이루어진다. 데이터 탐색 소프트웨어 DAVIS(DAta VISualization system) (Huh와 Song, 2002)에서는 동적 기능을 사용하여 데이터 시각화를 하는 것이 가능하다.

Hartingan과 Kleiner (1981)는 다차원 분할표를 그림으로 표현하는 방법으로서 모자이크 도표(mosaic plot)를 소개하였고, 이 도표는 Friendly (1994, 1999)에 의해 독립성 검정에 대한 시각적 추론을 하는 작업으로 확장되었다. 모자이크 도표는 분할표를 셀의 도수와 비례적인 크기를 갖는 타일로 표현하는 기법이다. 그러나 차원이 높아질수록 그림이 혼란스러워지는 단점이 있다.

Huh (2004)는 모자이크 도표의 개념을 이용하지만 모자이크 도표에 비해 많은 장점을 가지고 있는 선 모자이크 도표(line mosaic plot)를 소개하였다. 이 도표에서는 분할표의 각 셀을 같은 크기를 갖는 사각형으로 그리고 모자이크 도표의 타일의 크기만큼의 선으로 대체하는 것이다. 그러므로 차원이 높아지더라도 그림의 식별력이 현저하게 향상된다. 모자이크 도표에는 동적 기능을 연동시키는 것이 어렵지만, 선 모자이크 도표의 경우 특정부분에 해당하는 데이터를 쉽게 찾을 수 있으므로 동적 기능 연동이 수월하게 이루어질 수 있다.

본 연구는 2009년도 한성대학교 교내연구비 지원과제임.

[1] 교신저자: (136-792) 서울시 성북구 삼선동 2가 389, 한성대학교 멀티미디어공학과, 교수.
E-mail: wcha@hansung.ac.kr

본 논문에서는 선 모자이크 도표에 데이터 시각화 기능을 연동시켜 동적 그래픽스를 생성시키는 과정을 설명하고, 동적 그래픽스를 통한 선 모자이크 도표가 실제 데이터 탐색과정에 어떻게 이용될 수 있는지 알아보고자 한다. 본 연구를 위해서는 8개의 이산변수와 6개의 연속변수가 혼합되어 있는 데이터로부터 이산변수에 대한 선 모자이크 도표를 작성한다. 그 다음 선 모자이크 도표의 특정 부분에 대응하는 연속변수의 산점도(scatter plot)와 상자도형(box plot)을 그려, 선 모자이크 도표와 이들 사이를 동적으로 연결하여 데이터를 탐색하는 과정을 살펴보고 이 방법의 효율성과 유용성을 분석한다. 본 연구의 의의는 선 모자이크 도표에 동적기능을 연동시켜 데이터를 탐색함으로써 기존의 방법으로는 분석할 수 없었던 다양한 정보를 얻을 수 있음을 직접적으로 보인 것이다.

본 논문의 구성은 다음과 같다. 2장에서는 모자이크 도표와 선 모자이크 도표를 소개하고, 3장에서는 선 모자이크 도표로부터의 동적 그래픽스에 대하여 기술하였다. 4장에서는 다른 도형으로부터 선 모자이크 도표로의 동적 그래픽스를, 5장에는 결론을 기술하였다.

2. 모자이크 도표와 선 모자이크 도표

이장에서는 2차원 이상의 이산형 변수들 간의 관계를 알아보기 위한 시각화 방법인 모자이크 도표와 선 모자이크 도표를 그리는 알고리즘에 대하여 소개하고, 기존의 모자이크 도표에 비해 선 모자이크 도표의 장점에 대해 기술한다.

2.1. 모자이크 도표 알고리즘

모자이크 도표를 그리기 위한 알고리즘은 크게 두 가지로 제시되어 있다. 첫 번째는 특정한 차원의 분할표에 맞는 모자이크 도표를 그리는 것이고, 두 번째는 재귀적인 구조를 사용하여 그리는 것이다. Wang (1985)과 Friendly (1994)는 4개의 변수를 갖는 4차원 분할표에 대하여 모자이크 도표를 그리는 알고리즘을 제시하였으며, Emerson (1998)은 재귀적인 구조로 모자이크 도표를 그리는 알고리즘을 *S*로 구현하였다. 이 알고리즘들은 모두 범주형 자료에 대한 분할표가 입력값으로 주어져야만 그릴 수 있다. 본 절에서는 Wang (1985)과 Friendly (1994)가 소개한 알고리즘을 기초로 하여 n원 분할표에 대한 모자이크 도표를 그리는 알고리즘을 소개한다.

[모자이크 도표 알고리즘]

Input: n가지 변수를 갖는 n원 분할표

Output: 모자이크 도표

Step 1. 모자이크 도표가 그려질 영역은 하나의 사각형(타일)으로부터 출발한다.

Step 2. For $s = 1$ to n번째 변수

Step a. s번째 변수에 대한 주변빈도 $m_{s|1,2,\ldots,(s-1)}$를 구한다.

(여기서, $m_{s|1,2,\ldots,(s-1)}$는 첫 번째 변수에서 $(s-1)$번째 변수까지에 대한 주변빈도가 주어졌을 때 이에 대한 s번째 변수에 대한 빈도)

Step b. if s번째 변수가 홀수번째

then 이전 단계에서 생성된 타일들에 대하여 주변빈도에 비례하는 너비를 갖도록 이전 단계의 타일들을 수직 방향(세로)으로 분할한다.

if s번째 변수가 짝수 번째

then 이전 단계에서 생성된 타일들에 대하여 주변빈도에 비례하는 높이를 갖도록 이전 단계의 타일들을 수평 방향(가로)으로 분할한다.

2.2. 선 모자이크 도표 알고리즘

선 모자이크 도표를 그리기 위해서는 먼저 모자이크 행렬을 작성해야 한다.

(1) 모자이크 행렬

변수의 수가 k이고 데이터의 개수가 n이라 하면 이것으로부터 크기가 $n \times k$인 데이터행렬 $\boldsymbol{X}$를 구성할 수 있다. 여기서 k개의 변수 중 p개의 이산변수를 모자이크 변수라 하고 모자이크 변수로 이루어진 $n \times p$ 데이터행렬 $\boldsymbol{V}$를 구성한다. $\boldsymbol{V}$의 한 인스턴스를 벡터 $\boldsymbol{v}$로 나타내면 $\boldsymbol{v}$의 성분은 v_j, $j = 1, \ldots, p$들이다. p개의 이산변수들의 범주 수로 이루어진 벡터를 $\boldsymbol{n} = (n_1, n_2, \ldots, n_p)$라 하면 데이터행렬 $\boldsymbol{V}$로부터의 2차원 모자이크 행렬 $\boldsymbol{F}$의 행의 크기는 $\prod_{i=1}^{[p/2]} n_{2i}$이고 열의 크기는 $\prod_{i=0}^{[(p-1)/2]} n_{2i+1}$이 된다. 여기서 $[x]$는 x를 넘지 않는 정수를 의미한다. 모자이크 행렬 $\boldsymbol{F}$를 구하는 알고리즘과 선 모자이크 도표 알고리즘은 다음과 같다.

[모자이크 행렬 알고리즘]

Input: 데이터행렬 $\boldsymbol{X}$

Output: 모자이크 행렬 $\boldsymbol{F}$

Step 1. 데이터행렬 $\boldsymbol{X}$로부터 모자이크 변수 p개로 이루어진 데이터행렬 $\boldsymbol{V}$를 구성한다.

Step 2. 다음 식에 의해 $\boldsymbol{V}$의 한 인스턴스 $\boldsymbol{v}$로부터 $\boldsymbol{F}$ 행렬의 $\boldsymbol{v}$가 속하는 행과 열을 나타내는 I와 J를 계산한다.

$$I = \sum_{i=1}^{[p/2]} (v_{2i} - 1) \prod_{j=i+1}^{[p/2]} n_{2j} + v_{2[p/2]}$$

$$J = \sum_{i=0}^{[(p-1)/2]} (v_{2i+1} - 1) \prod_{j=i+1}^{[(p-1)/2]} n_{2j+1} + v_{2[(p-1)/2]+1}$$

여기서, $[x]$는 x를 넘지않는 정수.

Step 3. $F_{I,J}$의 요소를 1 증가 시킨다.

Step 4. Step 2와 Step 3을 자료의 개수 n만큼 반복한다.

[선 모자이크 도표 알고리즘]

Step 1. 모자이크 행렬의 (행의 수 × 열의 수) 만큼의 셀을 같은 크기의 사각형으로 그린다.

Step 2. 모자이크 행렬의 각 성분에 해당되는 크기를 선으로 나타낸다.

2.3. 실제 데이터에 대한 모자이크 도표와 선 모자이크 도표

이 절에서 사용한 실제 데이터 Heart Disease는 심장병(협심증) 진단을 위한 14개의 변수(이산형 8가지, 연속형 6가지)에 대해 270명의 성인남녀를 대상으로 조사한 것으로서 UCI 데이터베이스에서 제공하는 것이다 (Merz와 Murphy, 1996). 각 변수에 대한 설명과 기술통계량 및 빈도를 표 1과 2에 나타내었다. Heart Disease 데이터의 이산형 변수에 대한 모자이크 행렬은 다음 표 3, 4와 같다.

Heart Disease 데이터에서 4개의 이산변수로 성별, 가슴통증유형, 심전도결과, 조형술에 의한 협심증 진단을 택해서 그린 모자이크 도표와 선 모자이크 도표는 그림 1, 2와 같다. 이 데이터에서는 $p = 4$, $\boldsymbol{n} = (2, 4, 3, 2)$가 되며, sex = female, painType = asymptomatic, restecg = abnormal, diagnosis = no인 자료 값들에 대한 벡터는 $\boldsymbol{v} = (1, 1, 1, 1)$로서 모자이크 행렬 알고리즘에서 $\{I = 1, J = 1\}$로 구해진다.

표 1: 이산형 변수의 범주 및 빈도

변수설명(변수명)	범주(빈도)
성별(sex)	female(87), male(183)
가슴통증유형(painType)	asymptomatic(129), non-anginal(79), atypical(42), typical(20)
공복시혈당(fbs)	FALSE(230) : <= 120 mg/dl, TRUE(40) : > 120mg/dl
심전도결과(restecg)	abnormal(2), normal(131), probable(137)
운동부하검사(exang)	no(181), yes(89)
운동부하검사의 ST요소 기울기상태(slope)	downsloping(18), flat(122), upsloping(130)
thal	fixed_defect(14), normal(152), reversable(104)
조형술에 의한 협심증진단(diagnosis)	No(150) : < 50% diameter narrowing, Yes(120) : > 50% diameter narrowing

표 2: 연속형 변수의 기술통계량

변수명	평균	표준편차	최소값	최대값	Q1	중위수	Q3	결측치
연령(age)	54.433	9.109	29	48	55	61	77	0
휴식시 혈압(blood_pressure)	131.344	17.862	94	120	130	140	200	0
콜레스테롤양(cholesterol)	249.659	51.686	126	213	245	282	564	0
최대심장박동수(thalach)	149.678	23.166	71	133	154	167	202	0
ST절(oldpeak)	1.050	1.145	0	0	0.8	1.8	6.2	0
관상동맥조형술에 타겟이 되는 관상동맥수(ca)	0.670	0.944	0	0	0	1	3	0

표 3: Heart Disease 데이터에 대한 $\boldsymbol{F}$

0	8	9	0	13	8
1	6	11	0	26	47
0	9	6	0	14	6
0	0	1	0	3	3
1	17	13	0	19	12
0	1	0	0	8	8
0	3	1	0	2	9
0	0	0	0	2	3

표 4: diagnosis를 목적변수로 하였을 때의 Heart Disease 데이터에 대한 $\boldsymbol{F}$

diagnosis = no 또는 $k = 1$일 때						diagnosis = yes 또는 $k = 2$일 때					
0	8	9	0	13	8	1	6	11	0	26	47
0	9	6	0	14	6	0	0	1	0	3	3
1	17	13	0	19	12	0	1	0	0	8	8
0	3	1	0	2	9	0	0	0	0	2	3

2.4. 모자이크 도표와 선 모자이크 도표의 비교

2.3절에서 실제 데이터에 대해 모자이크 도표와 선 모자이크 도표를 그려본 결과 선 모자이크 도표에서 데이터의 분포를 훨씬 쉽게 파악할 수 있었다. 모자이크 도표에 비해 선 모자이크 도표의 장점은 다음과 같다.

첫째로, 인간의 지각 능력은 2차원적인 면 보다는 1차원적인 선에서 더 뛰어나다는 것이다. 기존의 모자이크 도표가 분할표의 한 셀을 비례적인 면에 의해 표현한 것에 비해 선 모자이크 도표에서는 각 셀의 크기만큼의 선분으로 표현하기 때문에 분할표의 독립성 검정을 파악하는데 훨씬 더 효율적이다. 따라서 차원이 높아져도 그림의 식별력이 모자이크 도표에 비해 현저히 향상된다.

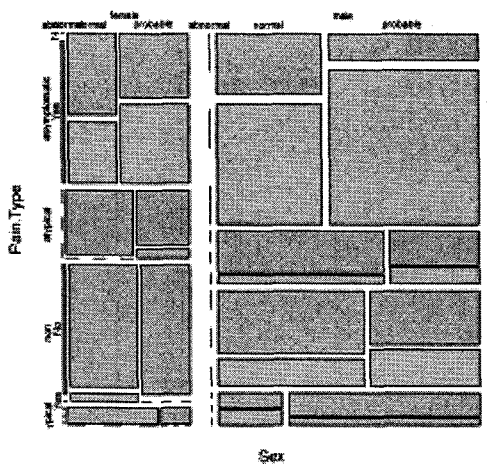

그림 1: *Heart Disease* 데이터에 대한 모자이크 도표

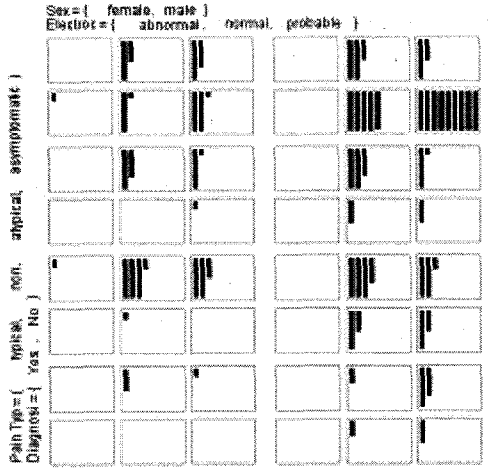

그림 2: *Heart Disease* 데이터에 대한 선 모자이크 도표

둘째로, 일반화된 모자이크 도표의 알고리즘을 구현하기 위해서는 재귀호출(recursion)을 이용해야 하나, 선 모자이크 도표의 알고리즘은 직접적이다. 즉, 선 모자이크 도표의 알고리즘은 차원의 증가나 셀 요인의 증가가 전혀 문제되지 않는다. 이것은 컴퓨터의 연산 수행 시간이나 기억장소 측면에서 훨씬 더 유리하다는 것을 의미한다.

마지막으로, 모자이크 도표는 그림을 그리기 위해 일단 자료로부터 분할표를 작성하고 그 분할표로부터 해당하는 셀들의 크기에 비례해서 사각형을 그려나가는 데 비해, 선 모자이크 도표는 자료로부터 직접 해당하는 셀에 선분을 추가하는 방법으로 그릴 수 있다. 이는 더욱 다양화되어 가고 있는 동적 그래픽스 기법을 적용하여 자료의 구조를 탐색하는데 있어 매우 중요하고 효율적인 방법이다.

3. 선 모자이크 도표로부터의 동적 그래픽스

동적 그래픽스는 마우스와 같은 입력장치들을 사용하여 컴퓨터 스크린 상의 그래픽 요소들을 '직접 조작'할 수 있고, 그 결과 그래픽에 '즉각적인 변화'를 줄 수 있는 기법이다. 대표적인 동적 그래픽스 기법은 마우스에 의한 브러싱(brushing), 포커싱(focusing), 연결(linking), 삭제(deleting), 표식화(labeling), 색입히기(coloring) 등이 있다. 이러한 기법들은 정적으로 그려진 다양한 도표들을 동적으로 연결하고 다양한 동적인 기법을 적용함으로써, 각각의 통계 도표를 통해 정적으로 데이터 구조를 탐색하는 경우보다 좀 더 효율적이고 강력한 데이터 탐색을 할 수 있게 된다.

본 논문에서는 이산형 데이터의 관계를 효율적으로 파악해 주는 선 모자이크 도표와 연속형 데이터 구조를 탐색하기 위해 사용하는 산점도와 상자도형에 대하여, 동적 그래픽스 기법을 적용함으로서 이산형 데이터와 연속형 데이터가 혼합되어 있는 데이터들의 구조를 효율적으로 탐색하기 위한 방법을 제시하고자 한다.

먼저 선 모자이크 도표로부터 다른 도표로의 동적 그래픽스를 적용하여 데이터의 구조를 탐색하는 것에 대하여 논의한다. 여기서, 선 모자이크 도표로부터라는 의미는 선 모자이크 도표의 특정 타일을 클릭했을 때, 그 마우스 이벤트에 반응하여 다른 도표들이 변하는 것을 의미한다. 즉, 선 모자이크 도표에서 클릭하거나 포커싱된 타일의 도표상의 위치를 계산하고 위치 값을 데이터 값으로 변환시킨 후, 다른 도표상에 이 데이터들을 다른 색으로 표시하거나 이 데이터만을 가지고 다른 도표를 다시 그려 데

이터의 구조를 분석하는 과정을 말한다. 선 모자이크 도표의 타일들은 모두 같은 크기를 가지므로, 마우스 위치에 따른 타일의 도표상의 위치(즉, 가로위치 I, 세로 위치 J)를 쉽게 알 수 있다. 이것으로부터 모자이크 행렬 요소에 해당하는 데이터 벡터 v들을 다음 알고리즘에 의해 쉽게 구할 수 있다. 그러나 모자이크 도표 상에서 계산하려면 각 타일의 크기에 대한 정보 등을 기억하고 있어야 한다.

[모자이크 행렬의 요소 $F_{I,J}$에 해당하는 데이터 벡터 v들을 구하는 알고리즘]

Input: 모자이크 도표 상의 위치 I, J

Output: 데이터 벡터들 v[]

Step1. J로부터 홀수 행의 벡터들 $v_1, v_3, \ldots, v_{2[(p-1)/2]+1}$을 계산한다.

$$v_i = \begin{cases} 1 + \dfrac{J-1}{\prod_{j=[(i-1)/2]+1}^{[(p-1)/2]} n_{2j+1}}, & i = 1,3,\ldots,2[(p-1)/2]-1, \\ 1 + \mathrm{Mod}\left(J-1, n_{2[(p-1)/2]+1}\right), & i = 2[(p-1)/2]+1. \end{cases}$$

Step2. I로 부터 짝수 행의 벡터들 $v_2, v_4, \ldots, v_{2[p/2]}$를 계산

$$v_i = \begin{cases} 1 + \dfrac{I}{\prod_{j=[i/2]+1}^{[p/2]} n_{2j}}, & i = 2,4,\ldots,2[p/2]-1, \\ 1 + \mathrm{Mod}\left(I-1, n_{2[p/2]}\right), & i = 2[p/2], \end{cases}$$

여기서, $\mathrm{Mod}(x,y) = x - x \times [x/y]$

이 알고리즘에서는 도표상의 위치와 변수 값의 크기 벡터 n을 이용하여 데이터 값 벡터 v를 계산하는데, 이 과정은 10진수를 2진수로 변환하는 과정과 유사하게 이루어진다. 이 과정에 대한 함수 프로그램이 알고리즘을 기초로 하여 다음에 주어져 있다. 이와 같은 방법으로 데이터 값이 계산되면, 마우스 클릭으로 선택된 타일에 해당하는 데이터들이 동적으로 다른 도표에 반응하도록 데이터 정보를 전달할 수 있다.

[선 모자이크 도표의 각 타일의 위치 I와 J로부터 자료 값 벡터 v를 구하는 함수]

```
/*
  Input : int I, int J, int n[], int p
 Output : int v[]
*/
void vFrom_I_J(int p, int n[], int I, int J, int v[])
{
    I = I - 1;
    J = J - 1;
    for(int i=p ; i >= 1 ; i--)
    {
        if(i % 2 == 0)
        {
```

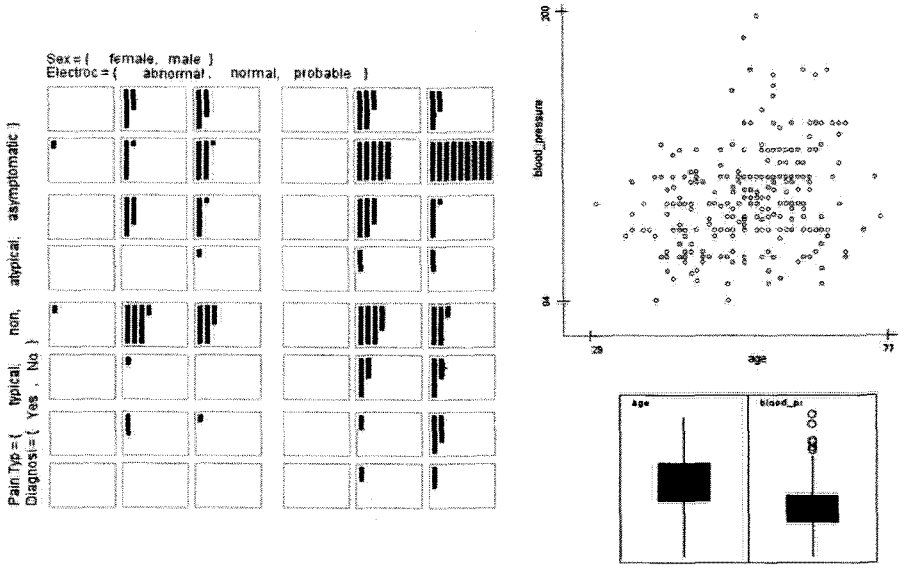

그림 3: *Heart Disease 데이터의 구조를 탐색하기 위한 도표들*

```
                    v[i] = I % n[i] + 1;
                    I = I / n[i];
                } else {
                    v[i] = j % n[i] + 1;
                    J = J / n[i];
                }
        }
}
```

그림 3은 앞 절에서 소개한 Heart Disease 데이터에 대한 선 모자이크 도표와 연속형 변수 연령, 혈압에 대한 산점도와 상자도표를 동시에 표시한 것이다. 그림 3의 선 모자이크 도표에서 sex = male, painType = asymptomatic, restecg = probable, diagnosis = yes에 해당하는 $\{I = 2, J = 6\}$ 타일(2행6열의 타일)을 마우스로 클릭한 하면 그림 4와 같이 동적인 변화들이 각 도표에 발생하게 된다. 동적인 변화의 과정을 살펴보면 먼저 선 모자이크 도표에서 타일을 마우스로 클릭하면 선택된 타일에 그려져 있는 선들이 다른 색으로 표시된다. 이 때 선택된 타일에 속하는 데이터 벡터를 계산하여 해당하는 데이터 정보를 다른 도표에 전달한다. 이 정보는 데이터 집합에서 관측 값들을 구분하는 자료 점의 레이블이나 아이디와 같은 정보가 된다. 전달된 데이터 정보는 동적 그래픽스 기법인 색입히기(coloring)을 적용하여 산점도와 상자도표 등 모든 도표에서 해당하는 자료 점들을 선택되지 않는 자료 점과 구분하기 위해 다른 색을 지정하여 화면에 나타낸다. 상자도표의 경우에는 해당 데이터와 선택되지 않는 다른 데이터들을 두 그룹으로 나누어 각각에 대한 상자도표를 그리게 된다. 이 때 새로 그려진 상자도표는 산점도에서 선택된 데이터를 나타내는 색과 동일한 색으로 각 변수에 대한 상자도표가 그려지므로 선택된 데이터와 선택되지 않는 데이터, 두 그룹의 데이터 분포를 비교할 수 있다. 그림 4의 선 모자이크 도표로부터의 동적인 변화의 결과를 보면, 선 모자이크 도표의 $\{I = 2, J = 6\}$에 해당하는 데이터들

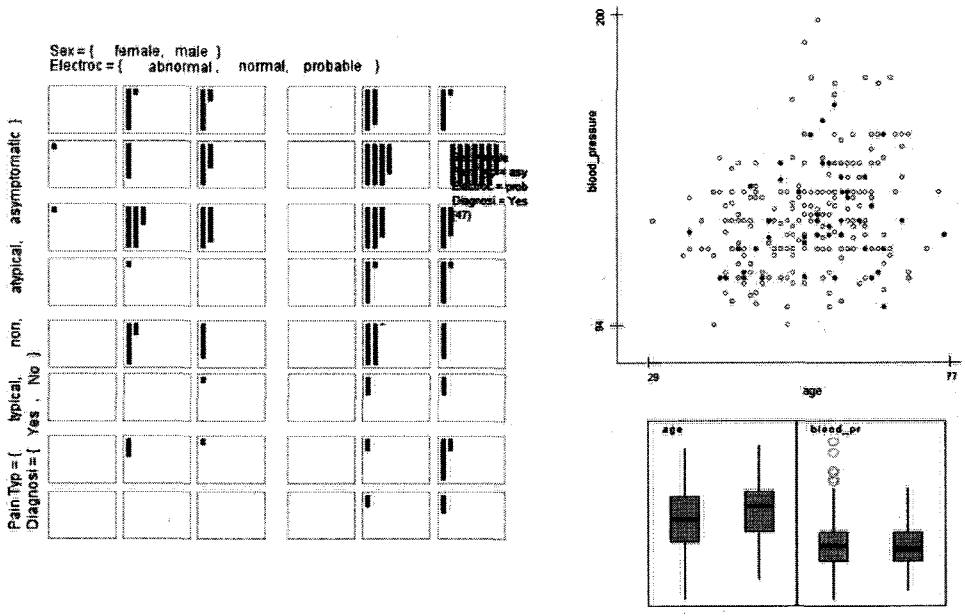

그림 4: *선 모자이크 도표의 타일을 클릭한 경우 다른 도표들로의 동적인 변화*

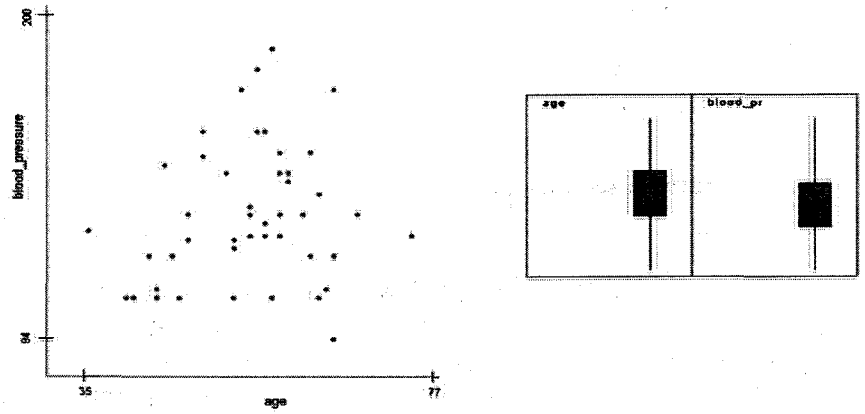

그림 5: *선 모자이크 도표의 타일을 클릭한 후 해당 타일을 포커싱한 경우 다시 그려진 산점도와 상자 도표*

은 연령(age)과 혈압(blood_pressure)에 대한 산점도에서 빨강색으로 표시되며, 연령과 혈압 두 변수 모두 데이터 값 범위 내에서 고르게 분포하고 있는 것으로 보인다. 상자도표의 경우에는 연령은 선택된 데이터 폭이 좀 더 넓고, 혈압은 좁은 것으로 나타나고 있다.

그림 5는 그림 4의 선 모자이크 도표에서 선택한 타일에 대하여 포커싱(focusing)을 적용한 경우의 결과 그림이다. 포커싱이란 선택된 데이터에 대해서만 모든 도표 내에 다시 그려주는 기능으로서 선택된 데이터에 대한 다차원 구조를 탐색하고자 할 때 사용할 수 있다.

4. 선 모자이크 도표로부터의 동적 그래픽스

이 장에서는 선 모자이크 도표로의 동적 그래픽스에 대해 논의한다. 여기서 선 모자이크 도표로란 의미는 다른 도표의 일부 데이터를 마우스로 클릭하거나 드래깅하여 발생한 마우스 이벤트에 반응하여 선 모자이크 도표가 동적으로 변하는 것을 의미한다. 그림 6의 산점도의 일부에 포커싱을 하면 해당 데이터 값을 읽어 선 모자이크 도표에 해당하는 위치를 계산하고 그 타일에 해당하는 개수 만큼에 변화를 주는 과정을 말한다. 즉, 자료값 벡터 v와 변수값의 크기 벡터 n을 이용하여, 도표상의 위치 I와 J를 계산한다. 이 과정은 2진수로부터 10진수를 계산하는 과정과 유사하게 진행된다. 이 과정에 대한 함수 프로그램을 다음에 기술하였다.

[자료 값 벡터 v로부터 선 모자이크 도표상의 타일 위치 I와 J를 구하는 함수]

```
/*
  Input :  int n[], int v[], int p
Output :  T[], T[0] -> I, T[1] -> J
*/
void I_J_From_v(int p, int n[], int v[], int T[])
{
        int I, J;
        int s;

        I = v[2]-1;
        for(int i=2 ; i < p-1 ; i+=2) {
                I = I*n[i+2] + v[i+2] - 1;
        }
        I++;

        J = v[1]-1;
        for(int i=1 ; i < p-1 ; i+=2) {
                J = J*n[i+2] + v[i+2] - 1;
        }
        J++;

        T[0] = I;
        T[1] = J;
}
```

그림 6은 산점도에서 혈압이 높은 자료 점들을 마우스로 드래그하여 선택한 경우 선택된 데이터에 대하여 빨강색으로 표시되면서 다른 도표에 해당하는 자료 점에 대한 정보를 같은 색으로 표시하고 있다. 혈압이 높은 경우 선 모자이크 도표를 보면 성별, 가슴통증유형, 심전도결과 및 조형술에 의한 진단 변수의 모든 범주에 고르게 분포하고 있음을 알 수 있다. 또한 상자도표에서는 혈압이 높은 자료가 연령층도 높게 분포하는 것으로 나타나고 있다. 그림 7은 산점도에서 혈압이 높은 데이터를 선택하여 포커싱한 경우의 동적인 변화를 나타낸 것이다.

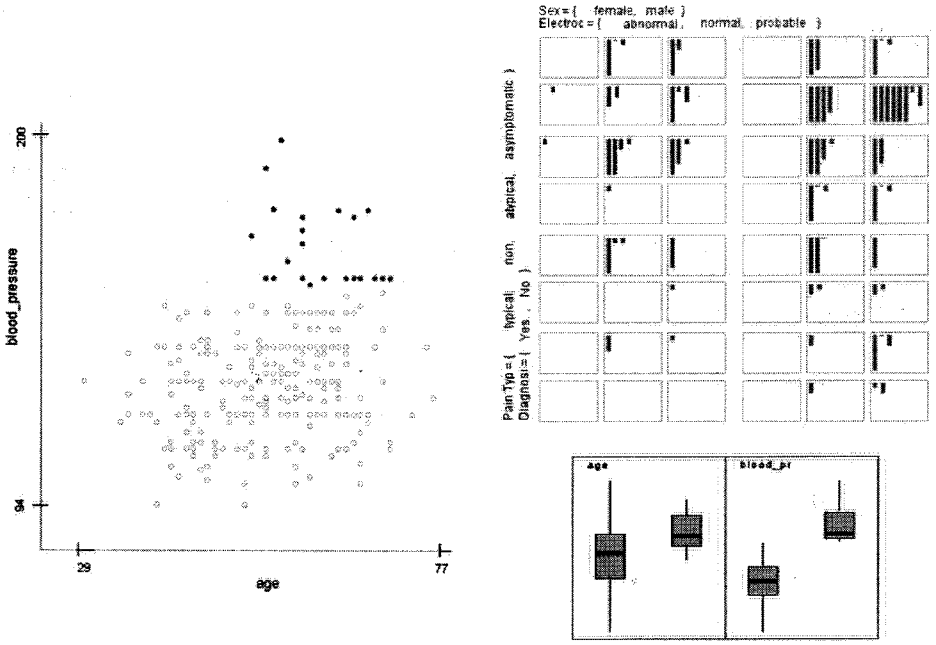

그림 6: *산점도에서 혈압이 매우 높은 그룹을 드래깅하여 선택한 경우*

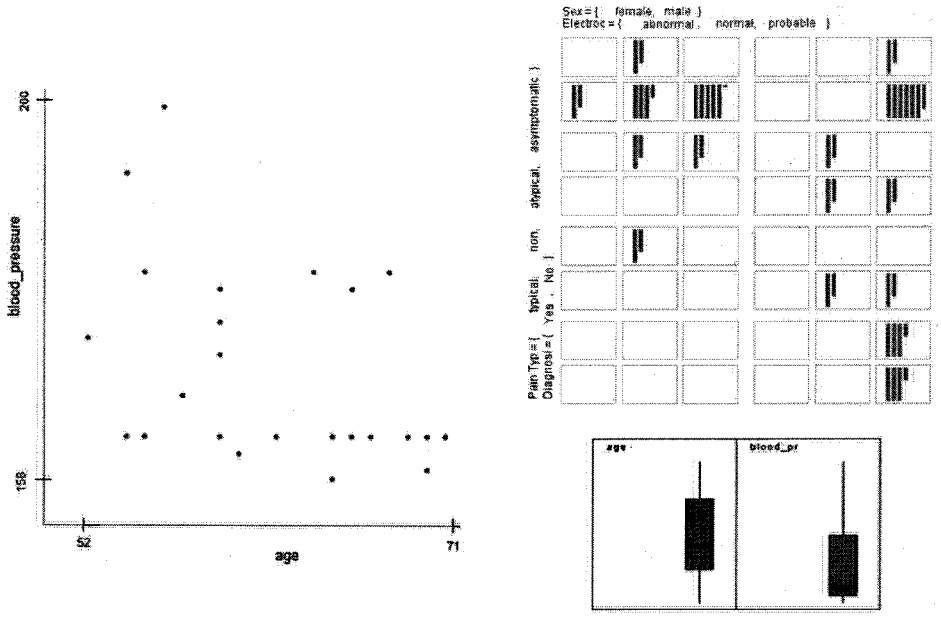

그림 7: *산점도에서 혈압이 매우 높은 그룹을 포커싱한 경우의 동적인 변화*

5. 결론

본 논문에서는 이산형과 연속형 데이터가 혼합되어 있는 데이터 구조를 탐색하기 위하여 동적 기법을 사용하였다. 이산형 변수들 간의 관계를 표현하는 선 모자이크 도표와 연속형 변수들의 관계를 위한 산점도, 일변량 변수 관점에서의 데이터의 분포를 파악할 수 있는 상자도표를 동시에 사용하면서, 동적인 기법들을 적용하여 다차원 데이터에 대한 구조를 좀 더 쉽게 파악할 수 있음을 보였다. 선 모자이크 도표로부터 다른 도표로의 동적 그래픽스는 선 모자이크 도표의 타일들이 모두 같은 크기를 가지므로 클릭하거나 포커싱한 타일에 대응하는 데이터들을 쉽게 파악하여 다른 도표로 전달할 수 있기 때문에 가능하다. 또, 다른 도표로부터의 선 모자이크 도표로의 동적 그래픽스는 다른 도표의 일부 데이터를 클릭하거나 포커싱 하면 해당 데이터 값으로부터 선 모자이크 도표에 해당하는 위치를 쉽게 계산할 수 있기 때문에 가능하다. 본 논문에서는 연속형 변수에 대해 산점도와 상자도표를 작성하였는데 선 모자이크 도표와 다른 도표간의 동적 그래픽스도 쉽게 이루어질 수 있다.

참고 문헌

Friendly, M. (1994). Mosaic displays for multi-way contingency tables, *Journal of the American Statistical Association*, **89**, 190–200.

Friendly, M. (1999). Extending mosaic displays: Marginal, partial, and conditional views of categorical data, *Journal of Computational and Graphical Statistics*, **8**, 373–395.

Hartigan, J. A. and Kleiner, B. (1981). *Mosaics for Contingency Tables*, Eddy, W.F., editor, Computer Science and Statistics: Proceeding of the 13th Symposium on the Interface, 268–273, Springer-Verlag, New York.

Huh, M. Y. and Song, K. Y. (2002), DAVIS: A Java-based Data Visualization system, *Computational Statistics*, **17**, 411–423.

Huh, M. Y. (2004). Line Mosaic Plot: Algorithm and Implementation. invited paper, Proceedings in COMPSTAT, 277–285.

Huh, M. Y. (2009). http://stat.skku.ac.kr/myhuh/

Emerson, J. W. (1998). Mosaic displays in S-PLUS: A general implementation and a case study, *Statistical Computing and Graphics Newsletter*, **9**, 17–23.

Merz, C. J. and Murphy, P. M. (1996). UCI Repository of Machine Learning Databases, Department of Information and Computer Science, University of California, Irvine, CA(http://www.ics.edu/~mlearn/MLRepository.html)

Nicholas, C. J. (1999). The Emergence of Data Visualization and Prospects for its Business Application, Masters of Information Systems Management Professional Seminar.

Wang, C. M. (1985). Application and computing of mosaics, *Computational Statistical and Data Analysis*, **3**, 89–97.

2009년 12월 접수; 2010년 1월 채택

Dynamic Graphics Using Line Mosaic Plot

Woon Ock Cha[1,a], Kyung Mi Lee[b], Byong Su Choi[a]

[a]Department of Multimedia Engineering, Hansung University
[b]Department of Internet Information, Seoil University

Abstract

This study is about the dynamic graphics which can be used for the exploration of the characteristics of data comprising discrete and continuous variables. Simultaneously using line mosaic plot for the relation of discrete variables and box plot together with scatter plot for the relation of continuous variables, we have applied dynamic methods among these plots to demonstrate that the structure and characteristics of the multivariate data could be easily analyzed.

Keywords: Line mosaic plot, dynamic graphics.

This research was financially supported by Hansung University in the year of 2009.

[1] Corresponding author: Professor, Department of Multimedia Engineering, Hansung University, 389, 2-Ga, Samsun-Dong, Sungbuk-Gu, Seoul 136-792, Korea. E-mail: wcha@hansung.ac.kr