통계적 그래픽스 도구로서의 정다각기둥평행좌표그림

장대홍[1)]

요약

탐색적자료분석을 위한 도구로서 그래픽 방법들을 자주 쓰게 되는 데 이러한 그래픽 방법 중 평행좌표그림을 대상으로 이 방법을 확장하여 볼 수 있다. 이러한 확장된 그림인 정다각기둥평행좌표그림은 탐색적자료분석 도구로서 유용하게 쓰일 수 있다.

주요용어: 통계적 그래픽스, 정다각기둥평행좌표그림

1. 서론

자료의 크기를 n이라 하고 변수의 개수를 p라 할 때 자료행렬을 $(x_{ij}), i=1,2,\ldots,n;$ $j=1,2,\ldots,p$로 나타낼 수 있다. 이러한 자료행렬에 대한 탐색적자료분석 단계에서 우리는 그래픽 방법들을 자주 이용하게 된다. grand tour, 체르노프얼굴(Chernoff face), glyph, Andrews 곡선, 레이더차트(radar chart, star diagram), 회전(rotation, spin plot), 산포도행렬, 평행좌표그림(parallel coordinate plot), 성좌그림(constellation graph), 원추그림(coneplot) 등이 있다. 이외에도, 수리적이고, 기하학적인 이론이 첨부된 그림인 biplot, 다차원척도법, 대응분석, 주성분그림 등이 있다. 이 들 그림들 중 평행좌표그림은 Inselberg(1985)가 구체적으로 제안한 이후 최근까지도 많은 학문영역에서 다양하게 쓰이고 있다.(Gennings외 3인(1990), Inselberg와 Dimsdale(1990), Wegman(1990), Madhavan외 3인(1991), Miller와 Wegman(1991), Lee외 3인(1995), Bateson과 Curtiss(1996), Keim과 Kriegel(1996), Lee와 Ong(1996), Weber와 Desai(1996), Becker(1997), Inselberg(1998, 2002), Ankerst외 2인(1998), Teppola외 4인(1998), Chou외 2인(1999), Fua외 2인(1999 a, b), Goel외 6인(1999), Groller외 2인(1999), King과 Harris(1999), Hall과 Berthold(2000), Siirtola(2000), Andrienko와 Andrienko(2001), Falkman(2001), Hauser외 2인(2002), Berthold와 Hall(2003), Albazzaz와 Wang(2005)) 'Computational Statistics and Data Analysis'라는 통계관련 저널의 43권 4호(2003년도)는 'Data Visualization'라는 제목으로 특집호로 꾸며졌는데 여기에 평행좌표그림에 대한 여러 편의 논문들이 수록되어 있다. 다음 그림 1은 iris 자료에 대한 평행좌표그림이다. 우리는 대략 두 개의 집락을 확인할 수 있다. 즉 type 1이 첫 번째 집락을 이루고 type 2와 type 3이 두 번째 집락을 이룸을 알 수 있다.

1) (608-737) 부산광역시 남구 대연3동 599-1 부경대학교 자연과학대학 수리과학부 통계학전공 교수

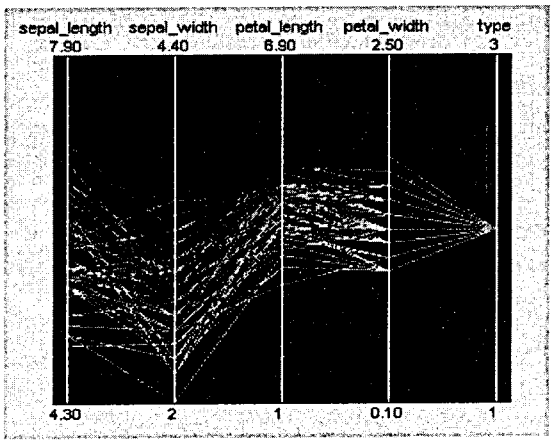


그림 1. iris 자료에 대한 평행좌표그림

그러나 이 평행좌표그림에 대하여 다음과 같은 3가지 문제점이 발생한다.

1. 데이터가 많은 경우 그림의 뭉개짐(over-plotting)
2. 데이터의 식별
3. 변수의 배열

첫 번째 문제점의 해결책으로서 hierarchical parallel coordinate plot를 이용할 수 있고, 두 번째 문제점의 해결책으로서 smooth parallel coordinate plot를 이용할 수 있다. 세 번째 문제점의 해결책으로서 Thom McLean과 Robert McEwen은 평행좌표그림의 확장인 milk carton plot를 제안하고 있는 데 이 들의 홈페이지(www.scs.gmu.edu/~tmclean/parallel3d.html)에 이러한 아이디어가 나타나 있다. 이 3차원 그림을 milk carton plot라는 이름보다는 정다각기둥평행좌표그림(regular polyprism parallel coordinate plot)이라 칭하는 것이 더 적절하다고 생각한다. 본 논문에서는 이러한 정다각기둥평행좌표그림을 수학패키지 Maple을 이용하여 구현하여 보았다.

2. 정다각기둥평행좌표그림

Minitab package에 있는 sample data 중 하나인 wine data(38개의 관측치를 갖고 7개의 변수(clarity, aroma, body, flavor, oakiness, quality, region)를 갖는 자료)에 대하여 Maple package를 이용하여 정다각기둥평행좌표그림을 그리면 다음 그림 2와 같다. 그림 2에서 보는 것처럼 정다각기둥평행좌표그림에서는 정다각기둥의 각 모서리(축)에 각 변수를 대응시킨다. 즉, 7개의 변수 clarity, aroma, body, flavor, oakiness, quality, region를 정칠각기둥의 7개의 모서리에 배열하여 정칠각기둥의 각 면에 변수1-변수2, 변수2-변수3, 변수3-변수4, 변수4-변수5, 변수5-변수6, 변수6-변수7, 변수7-변수1를 대응시켜 평행좌표그림을 그린다. 하나의 데이터는 정칠각기둥의 각 면을 돌아가며 하나의 꺾은선으로 나타내어진다. 인접하여 있지 않은 변수들 사이의 관계는 관심있는 변수들을 가로질러 연결하여 만든 면에 데이터를 나타낸다. 예로 region이라는 변수에 대한 나머지 6개의 변수들 사이의 관계를 나타내면 다음 그림 3과 같다. aroma와 quality 변수에 대략 두 개의 집락이 나타남을 알 수 있다.

1:clarity, 2:aroma, 3:body, 4:flavor, 5:oakiness, 6:quality, 7:region

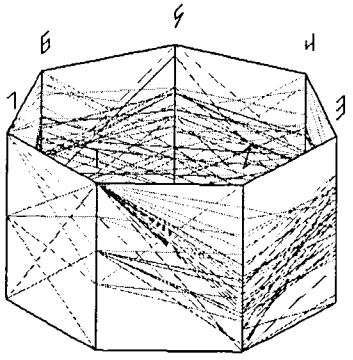


그림 2. wine 자료에 대한 정칠각기둥평행좌표그림

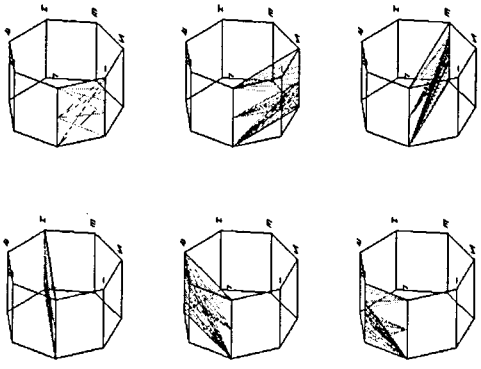


그림 3. region 변수에 대한 나머지 6개의 변수들 사이의 관계를 나타내는
정칠각기둥평행좌표그림

우리는 그림 4처럼 각 축에 점도표(dot plot)를 표시할 수 있다. region은 3개의 값을 갖고 clarity는 5개의 값을 가짐을 알 수 있다.

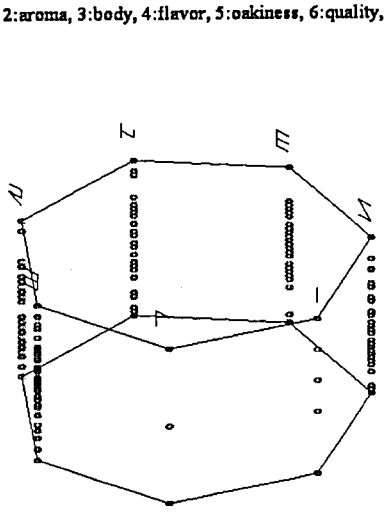
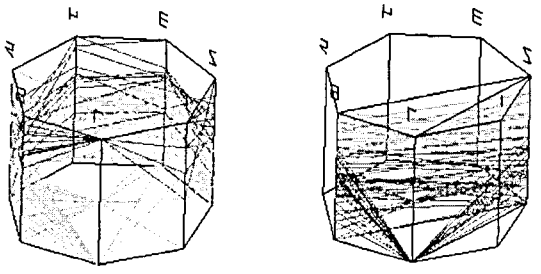


그림 4. 각 변수에 대한 점도표

동적 그래픽스 기능으로서 밝게하기(highlight)와 축선택(axes selection) 작업의 예를 보자. 변수 region 값이 3인 자료에 대하여 밝게하기를 행하면 다음 그림 5(a)와 같다. 두 번째 축(aroma), 여섯 번째 축(quality) 그리고 일곱 번째 축(region) 세 개의 축을 선택하여 이 세 개의 축 만을 이용하여 그림을 그리면 다음 그림 5(b)와 같다. 첫 번째 집락(region 3)은 aroma와 quality 변수에서 두 번째 집락(region 1과 region 2)과 구별되는 것을 알 수 있다.



(a) 밝게하기 (b) 축 선택

그림 5. 동적 그래픽스

3. 결론

탐색적자료분석 단계에서 이용되는 그래픽 방법들 중 평행좌표그림을 확장한 정다각기둥평행좌표그림은 원래의 평행좌표그림을 보완하는 그림으로서 탐색적자료분석시 유용한 그림도구가 될 수 있다.

참고문헌

[1] Albazzaz, H. and Wang, X. Z.(2005), Historical Data Analysis Based on Plots of Independent and Parallel Coordinates and Statistical Control Limits, *Journal of Process Control*, in press.

[2] Andrienko, G. and Andrienko, N.(2001), Exploring Spatial Data with Dominant Attribute Map and Parallel Coordinates, *Computers, Environment and Urban Systems*, 25, 5-15.

[3] Ankerst, M., Berchtold, S. and Keim, D.(1998), Similarity Clustering of Dimensions for an Enhanced Visualization of Multidimensional Data, *Proceedings of IEEE Symposium on Information Visualization*, 52-60.

[4] Bateson, A. and Curtiss, B.(1996), A Method for Manual Endmember Selection and Special Unmixing, *Remote Sensing of Environment*, 55, 229-243.

[5] Becker, O. M.(1997), Representing Protein and Peptide Structures with Parallel-Coordinates, *Journal of Computational Chemistry*, 18, 1893-1902.

[6] Berthold, M. R. and Hall, L. O.(2003), Visualizing Fuzzy Points in Parallel Coordinates, *IEEE Transactions on Fuzzy Systems*, 11, 369-374.

[7] Chou, S., Lin, S. and Yeh, C.(1999), Cluster Identification with Parallel Coordinates, *Pattern Recognition Letters*, 20, 565-572.

[8] Falkman, G.(2001), Information Visualisation in Clinical Odontology: Multidimentional Analysis and Interactive Data Exploration, *Artificial Intelligence in Medicine*, 22, 133-158.

[9] Fua, Y., Ward, M. O. and Rundensteiner, E. A.(1999 a), Navigating Hierarchies with Structure-based Brushes, *Proceedings of IEEE Symposium on Information Visualization*, 58-64.

[10] _____(1999 b), Hierarchical Parallel Coordinates for Exploration of Large Datasets, *Proceedings of Visualization '99*, 43-50.

[11] Gennings, C., Dawson, K. S., Carter, W. H. and Myers, R. H.(1990), Interpreting Plots of Multidimensional Dose-Response Surfaces in a Parallel Coordinate System, *Biometrics*, 46, 719-735.

[12] Goel, A., Baker, C., Shaffer, C. A., Grossman, B., Haftka, R. T., Mason, W. H. and Watson, L. T.(1999), VizCraft: A Multidimensional Visualization Tool for Aircraft Configuration Design, *Proceedings of Visualization '99*, 425-428.

[13] Groller, E., Loffelmann, H. and Wegenkittl, R.(1999), Visualizations of Dynamical Systems, *Future Generation Computer Systems*, 15, 75-86.

[14] Hall, L. O. and Berthold, M. R.(2000), Fuzzy Parallel Coordinates, *Proceedings of 19th International Conference of Fuzzy Information Processing Society*, 74-78.

[15] Hauser, H., Ledermann, F. and Doleisch, H.(2002), Angular Brushing of Extended Parallel Coordinates, *Proceedings of the IEEE Symposium on Information Visualization*, 127-130.

[16] Inselberg, A.(1985), The Plane with Parallel Coordinates, *The Visual Computer*, 1, 69-91.

[17] _____(1998), Visual Data Mining with Parallel Coordinates, *Computational Statistics*, 13, 47-63.

[18] _____(2002), Visualization and Data Mining of High-dimensional Data, *Chemometrics and Intelligent Laboratory Systems*, 60, 147-159.

[19] Inselberg, a. and Dimsdale, B.(1990), Parallel Coordinate: A Tool for Visualizing Multi-dimensional Geometry, *Proceedings of Visualization '90*, 361-378.

[20] Keim, D. A. and Kriegel, H.(1996), Visualization Techniques for Mining Large Databases: A Comparison, *IEEE Transactions on Knowledge and Data Engineering*, 8,923-938.

[21] King, K. and Harris, T.(1999), Parallel-coordinates Visualization of Capillary Transport Model Analysis, *Proceedings of BMES/EMBS Conference on Engineering in Medicine and Biology*, 1193.

[22] Lee, H. and Ong, H.(1996), Visualization Support for Data Mining, *IEEE Expert*, 11, 69-75.

[23] Lee, H., Ong, H., Toh, E. and Chan, S.(1995), A Multi-Dimensional Data Visualization Tool for Knowledge Discovery in Databases, *Proceedings of COMPSAC 95 Conference on computer Software and Applications*, 26-31.

[24] Madhavan, P. G., Xu, B., Penna, M. A. and Low, W. C.(1991), Co-ordinate Transformation in the Hippocampal Place Cell Phenomenon, *Proceedings of IEEE Conference on Systems, Man, and Cybernetics*, 1615-1620.

[25] Miller, J. J. and Wegman, E. J.(1991), Construction of Line Densities for Parallel Coordinate Plots, in *Computing and Graphics in Statistics* edited by A. Buja and P. A. Tukey. Springer-Verlag, New York, NY.

[26] Siirtola, H.(2000), Direct Manipulation of Parallel Coordinates, *Proceedings of IEEE International Conference on Information Visualization*, 373-378.

[27] Teppola, P., Mujunen, S., Minkkinen, P., Puijola, T. and Pursiheimo, P.(1998), Principal Component Analysis, Contribution Plots and Feature Weights in The Monitoring of Process Data From A Paper Machine's Wet End, *Chemometrics and Intelligent Laboratory Systems*, 44, 307-317.

[28] Weber, C. A. and Desai, A.(1996), Determination of Paths to Vendor Market Efficiency Using Parallel Coordinates Representation: A Negotiation Tool for Buyers, *European Journal of Operational Research*, 90, 142-155.

[29] Wegman, E. J.(1990), Hyperdimensional Data Analysis using Parallel Coordinates, *Journal of the American Statistical Association*, 85, 664-675.