Scagnostics를 이용한 다차원 데이터 탐색에 관한 사례연구

신동화O, 이세희, 서진욱
서울대학교 컴퓨터공학부
dhshin@hcil.snu.ac.kr, shlyi@hcil.snu.ac.kr, jseo@snu.ac.kr

A case study on exploration of multidimensional data using scagnostics

DongHwa ShinO, Sehi L'Yi, Jinwook Seo
Department of Computer Science and Engineering, Seoul National University

요 약

기존에 다차원 데이터를 탐색하기 위한 여러 시각적 분석 방법들이 고안되어왔다. Cognostics 기반 방식은 다차원 데이터를 통해 만들 수 있는 수많은 시각화를 특정한 지표를 기준으로 계산하여 점수를 낸 뒤, 그 점수를 기반으로 시각화 탐색의 우선순위를 정하는 방식이다. 본 논문에서는 그 중에서도 산점도를 중점적으로 분석하는 방식인 Scagnostics를 활용한 점진적 다차원 데이터 탐색이라는 기존 연구를 바탕으로한 사례 연구를 제시함으로써 연구의 활용 가능성을 검증하고자 한다.

1. 서 론

다차원 데이터의 시각적 분석 방법으로 Cognostics[1] 기반의 분석 방식이 있다. Cognostics 기반 분석 방식이란 다차원 데이터를 통해 만들 수 있는 수많은 시각화를 특정한 지표[2]를 (Monotonicity, Skewed, etc...) 기준으로 계산하여 점수를 낸 뒤, 그 점수를 기반으로 시각화 탐색의 우선순위를 정하는 방식이다. 다차원 데이터의 특성상 다 살펴보기 힘들 정도로 수많은 차트를 그려볼 수 있다는 점에서 Cognostics 방식은 데이터 분석가에게 효율적인 분석을 가능케 한다는 데에 의의가 있다.

Cognostics 분석 방식 가운데에서도 주로 산점도를 중점적으로 분석 대상으로 취급하는 방식을 Scagnostics(Scatterplot + Diagnostics)[2]라고 한다. 산점도는 두 개의 변수를 각각 X축과 Y축에 연결시켜서 두 변수에 대한 분포를 볼 수 있게 해주는 대표적인 시각화이다. 만약 Scagnostics를 이용하여 20차원(데이터의 변수 혹은 열이 10개)의 데이터를 분석하고자 한다면 20개의 변수 중 2개를 고르는 경우의 수인 190개의 산점도가 대상이 된다. 그래서 190개의 산점도를 특정 지표들을 기준으로 계산하고, 그 결과에 해당하는 점수와 순위가 나오면 순위가 높은 순서대로, 혹은 조금 특이한 점수 분포를 보이는 산점도 부터 확인해 볼 수 있다.

기존에 우리는 이 Scagnostics 분석 방식에 범주형 변수를 분석 대상으로 포함 시킴과 동시에, 이 때문에 급격하게 늘어나는 산점도의 수를 효과적으로 처리하기 위해 점진적인 시각적 분석을 분석에 접목시키는 연구를 진행하였다[3]. 그 결과 범주형 변수를 분석에 포함시킨 뒤, 특정한 범주로 산점도를 분할하였을 때, 눈에 띄는 흥미로운 패턴을 갖는 산점도들의 범주를 확인할 수 있었다. 본 논문에서 우리는 이 기존 연구에 대한 확장으로서, 제공되는 시각화 및 상호작용 방식을 이용한 사례 연구를 제시한다.

이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. NRF-2016R1A2B2007153).

2. 사례 연구

이번 사례 연구에서 사용하는 데이터는 대학생들의 학점 및 인적성 검사 결과 점수에 해당하는 데이터이다. 학점은 총 두 개학기의 평균평점이 존재하며, 인적성 검사 결과는 100점 만점에 해당하는 18개 항목으로 이루어져 있다. 즉 수치형 변수가 20개인 20차원 데이터인 것이다. 산점도를 분할하는 범주형 변수로 쓰일 변수는 2개이며 그 두 변수는 각각 '입학 전형', '학적 상태'이다. 입학 전형은 '수시 1학기', '수시 2학기', '정시'로 세 가지 범주가 있으며, 학적 상태는 '재학생'과 '휴학생' 두 가지 범주를 가지고 있다.

가장 눈에 띄는 지표는 이상치의 많고 적음을 파악하는 지표인 'Outlying' 이었다. 수많은 Index들의 점수가 표시되는 overview에서 먼저 Outlying 지표만을 보기 위해 overview의 우측 체크박스에서 다른 모든 지표들을 체크 해제하고 Outlying 지표만을 체크하였다(그림 1, 좌측 상단). 그 후에 가장 눈에 띄었던 점은 산점도의 중앙에서 조금 좌측의 상단에 위치하는 점이었다. 해당 점의 경우는 Index Score(지표 점수, 얼마나 Outlying 한지)가 높아서 분석자가 관심 있게 볼 이상치들을 다수 포함하고 있을 가능성이 큰 점이다. 중요한 점은 이 점이 Partition Score(분할 점수, 다른 범주들에 비해 해당 범주가 얼마나 다른 패턴을 띄고 있는가)가 꽤나 높다는

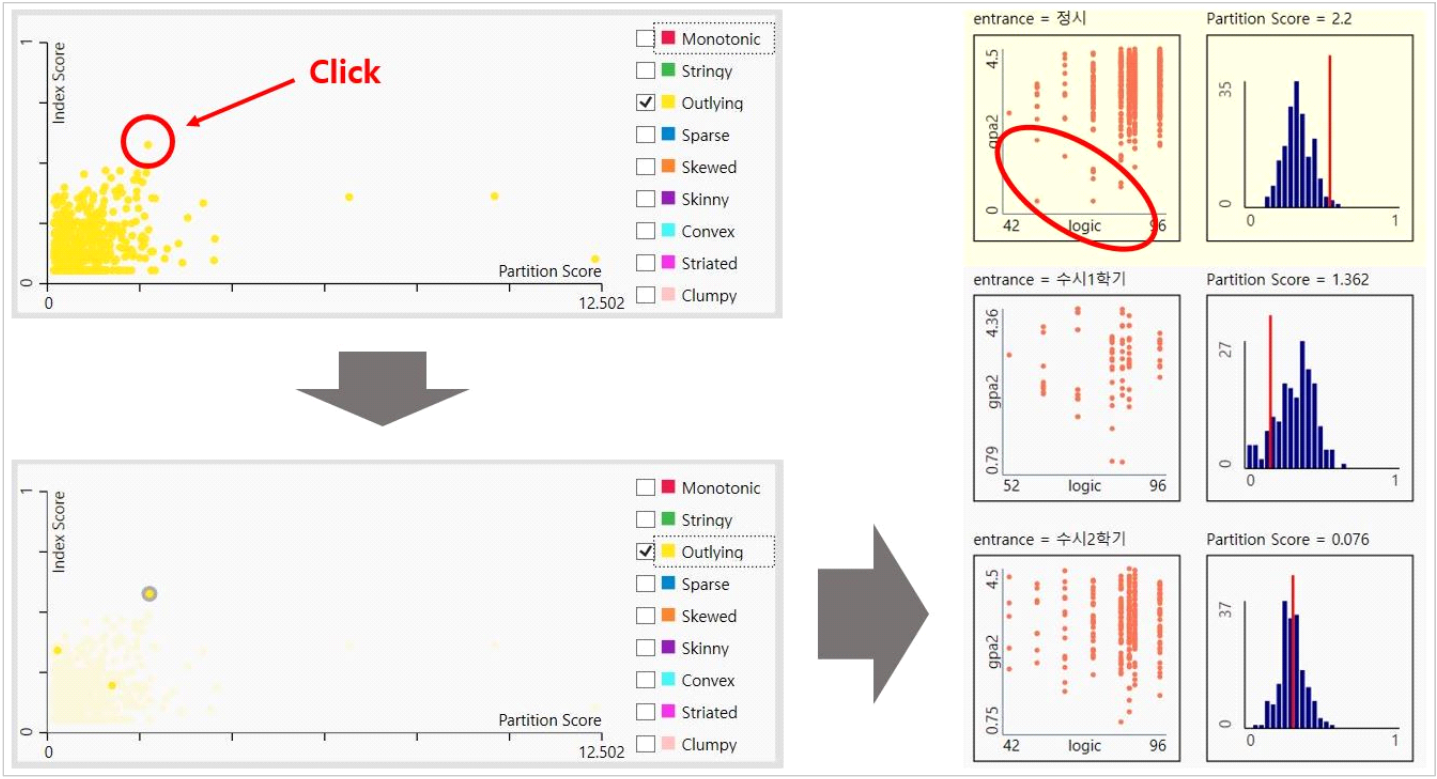

그림 1 Outlying 지표를 활용하여 overview에서 눈에 띄는 점을 파악한 뒤, 이를 클릭하여 자세한 정보를 확인하고 있다. 자세한 정보란, 분할(범주형) 변수로 분할된 산점도 3개이며 그중 Outlying 점수가 가장 높은 '정시' 범주를 중점적으로 보고 있는 모습이다.

점이었다. 해당 점을 클릭해보니, 분할 변수는 입학 전형이었고, X축은 논리력, 그리고 Y축은 평점평균을 나타내는 산점도였다. 우리가 클릭한 점은 그 중에서도 '정시'로 들어온 학생들의 논리력 및 평점평균의 분포를 알 수 있는 산점도였다. 분할 점수가 꽤 높은 편인 것으로 볼 때, 정시가 수시 1, 2학기에 해당하는 산점도보다 이상치를 더욱 많이 포함하고 있는 것으로 보이며, 실제로도 그러했다(그림 1, 우측). 특히나 정시 산점도를 살펴보면 이상치로 잡힌 것으로 추정되는 점(학생)들이 논리력은 꽤 높은 수준을 보이지만 평점평균이 떨어져서 이상치로 잡혔음을 추정해볼 수 있다. 이를 좀 더 자연스럽고 이해하기 쉬운 언어로 해석해보자면, "정시로 들어온 학생들은 논리력 수준에 비해 평점평균이 낮은 학생들이 다른 입학 전형에 비해 많이 발견된다" 정도로 해석해 볼 수 있다.

3. 한계 및 개선방안

이번 사례 연구를 통해서 기존 분석 방식의 한계와 그 개선방안에 대해서도 주목하였다. 가장 큰 한계로 보이는 점은 각 지표들의 점수를 인간이 해석하기에 다소 어려움이 있다는 것이다. 따라서 각 Index Score에 대한 이해를 돕기 위해 차트 위에 시각적 요소들을 추가로 보여주려고 한다. 예를 들어 'Convex' 지표의 경우 Scagnostics 알고리즘에 의해 계산되는 점들의 외곽선을 산점도 위에 시각적으로 그려줌으로써 사용자는 각 산점도가 얼마나 convex(볼록)한지에 대해 좀 더 쉽게 판단할 수 있다. 추가로 특정 지표만 체크박스로 선택했을 때(이번 사례 연구에서 Outyling만을 선택한 것과 같이), 해당 점들을 기준으로 overview 산점도의 X, Y축 스케일을 새로 지정함으로써, 현재 다른 지표 때문에 나머지 점들이 모두 축에 가까이 붙어서 패턴이 보이지 않는 문제를 해결해 볼 수도 있을 것이다.

4. 결 론

본 연구에서는 기존 Scagnostics를 활용한 점진적 다차원 시각적 데이터 분석 연구의 일환으로서 사례 연구를 제시하였다. 사례 연구 결과 기존 연구를 통해 데이터가 내포하고 있는 정보를 효과적으로 발견할 수 있었고, 기존 연구의 제한점 및 개선방안에 대해 고찰해볼 수 있는 계기가 되었다. 이 연구를 계기로 아직 진행하지 못한 정량적 데이터 분석 또한 설계할 계획이다.

참 고 문 헌

[1] Cleveland, W. The Collected Works of John W. Tukey: Graphics 1965-1985, Chapman & Hall/CRC, 5, 1988.

[2] Wilkinson, L., Anand, A., Grossman, R. Graph-theoretic scagnostics. In Proc. INFOVIS, 157-164, 2005.

[3] 신동화, 이세희, 서진욱. Scagnostics와 분할 변수 선택 기법을 활용한 점진적인 시각적 분석. 한국정보과학회 학술발표논문집, Vol.2017 No.12, 1349-1351, 2017.