Scagnostics와 분할 변수 선택 기법을 활용한 점진적인 시각적 분석

신동화 O, 이세희, 서진욱

서울대학교 컴퓨터공학부

dhshin@hcil.snu.ac.kr, shlyi@hcil.snu.ac.kr, jseo@snu.ac.kr

Progressive visual analytics using scagnostics and automatic partitioning variables selection method

DongHwa Shin O, Sehi L'Yi, Jinwook Seo

Department of Computer Science and Engineering, Seoul National University

요 약

이 논문에서는 scagnostics 개념을 기반으로 한 분석 기법 중 하나인 분할 변수 선택 기법과 점진적 시각화를 접목한 시각적 분석 시스템을 제시한다. 대다수의 scagnostics 분석 연구들은 2차원의 산점도를 나누지 않고 있는 그대로 분석하는 것이 주를 이루었다. 그러나 최근에는 분할 변수 하나를 선택해서 산점도 하나를 여러 개로 나누고, 그 중에서 흥미로운 패턴을 갖는 산점도를 발견하려는 연구들이 등장했다. 대표적인 연구로 분할 변수 선택 기법이 있다. 하지만 이를 실제 데이터 분석에 이용하려면 확장성 및 성능의 문제가 생긴다. 우리는 이 분할 변수 선택 기법의 한계를 점진적 시각화로 극복하여 실용적으로 사용할 수 있는 시각적 분석 시스템을 제시한다.

1. 서 론

이 논문에서 시각화 및 분석을 하려는 대상은 일반적인 다차원 데이터이고, 분석 방식은 cognostics 기반 분석이다. Cognostics란 다차원 데이터의 여러 차원들을 조합해 차트들을 만든 뒤, 이러한 여러 차트들에 점수를 매김으로써 개중에 점수가 높은, 즉 흥미로운 차트들을 선별적으로 볼 수 있게 해주는 개념을 의미한다. 어떤 종류의 차트라도 분석의 대상이 될 수 있지만 특히 산점도에 점수를 매기는 기법을 cognostics의 subset인 scagnostics라고 한다.

산점도는 2개의 차원이 각각 X축과 Y축에 연결되어 만들어진다. 그렇다면 N개의 차원이 있는 데이터에서는 N개 중 2개를 뽑는 조합의 개수만큼 산점도를 만들 수 있다. 기존에는 이렇게 만들어질 수 있는 수 많은 산점도 중에 흥미로운 패턴을 갖는 산점도를 발견하는 데에 도움을 주는 방식의 연구가 주를 이루었다. 그러나 최근에는 산점도를 나누었을 때 원래 보이지 않던 패턴이 보일 수 있다는 점에 착안한 연구가 진행되고 있다. 그 중 대표적 연구로는 Trelliscope[1]와 Anand의 분할 변수 선택 기법[2]이 있다.

Trelliscope는 R을 내장하고 있어 다양한 종류의 차트에 대해 매우 높은 자유도를 갖는 분석을 수행할 수 있다. 그러나 아쉬운 점은 나뉘어진 산점도를 또 다시 단순히 하나의 산점도로 보고 점수를 매기는 접근 방식을 취하고 있다는 것이다. 이에 비해 분할 변수 선택 기법은 특정한 (범주형의) 분할 변수를 선택했을 때에 그에 따라 나뉘어지는 산점도 중 '기존 합쳐진 산점도에서는 보이지 않던' 특별한 패턴을 가진 산점도에 대해 높은 점수를 줘서 이를 분석자가 중점적으로 분석할 수 있게 해주는 시스템이다. 나뉘어진 산점도 중에서 단순히 점수가 높은 것을 보여주는 것이 아니라, 나뉘어 졌을 때 기존 합쳐진 산점도와는 확연히 다른 패턴을 보여주는 산점도를 찾는 것이 Trelliscope와는 다른, 이 기법의 핵심이다.

예를 들어, 영화 데이터에서 X축을 '제작비', Y축을 '관객수'로 갖는 산점도를 분석한다고 생각해보자. 상관관계를 점수로 매겨보니 제작비-관객수 간 높은 상관관계가 있음을 알 수 있었다. 그렇다면 이 산점도를 '장르'라는 분할 변수로 나누었을 때, 나뉘어진 산점도들도 모두 높은 상관관계를 갖을까? 혹은 어떤 장르의 경우에는 다른 패턴을 보일까? 우리가 궁금한 지점이 바로 그 지점이다. 만약 액션 영화에 대한 산점도만 따로 떼어서, 제작비-관객수 간 상관관계를 보니 역시 높았다는 사실은 우리의 흥미를 끌지 못한다. 오히려 대부분의 장르가 다 높은 상관관계를 가졌는데, 오로지 멜로만 상관관계를 발견할 수 없었다라고 한다면 그것이 흥미로운 패턴이 될 것이다. 그러면 이는 "왜 멜로 영화만 제작비와 관객수 간 상관관계가 보이지 않는 걸까?" 라는 데이터에 대한 의미 있는 질문을 던질 수 있는 계기를 마련하게 되는 것이다.

이런 분할 변수 선택 기법에도 그 한계는 있다. 분할

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. NRF-2016R1A2B2007153).

Rank	X-Axis	Y-Axis	Partition	Iteration	Max Partition Score	Priority
1.	수익_미국	로튼토마토_평점	장르	202	7.719	High
2.	수익_미국	IMDB_평점	장르	202	7.562	High
3.	수익_전세계	IMDB_평점	장르	199	7.08	High
4.	수익_전세계	로튼토마토_평점	장르	187	6.229	High
5.	제작비	로튼토마토_평점	장르	187	5.282	High
6.	수익_미국	수익_전세계	장르	91	4.395	Normal
7.	로튼토마토_평점	IMDB_평가수	장르	90	3.876	Normal
8.	수익_미국	IMDB_평가수	장르	51	3.864	Normal
9.	제작비	수익_전세계	장르	112	3.687	Normal
10.	수익_전세계	IMDB_평가수	장르	54	3.444	Low
11.	수익_미국	로튼토마토_평점	원작	59	3.385	Low
12.	제작비	로튼토마토_평점	원작	31	3.277	Low
13.	제작비	IMDB_평점	장르	76	3.244	Low
14.	수익_미국	IMDB_평가수	원작	51	3.22	Low
15.	로튼토마토_평점	IMDB_평점	장르	34	3.218	Low
16.	IMDB_평점	IMDB_평가수	장르	30	3.117	Low
17.	수익_미국	수익_전세계	원작	31	3.068	Low

그림 1. 산점도와 분할 변수 항목의 점수와 순위, 진행상황 및 우선순위를 볼 수 있는 리스트 뷰

변수 선택 기법은 기본적으로 임의 순열 (random permutation) 방법을 주로 사용한다. 나뉘어진 하나의 산점도 당 1000번의 임의 순열 방법 적용이 필요하고, 그 후에 동일한 횟수로 scagnostics 점수를 구하는 과정을 거쳐야 한다. 데이터에 수치형 변수가 20개, 범주형 변수가 5개이고, 범주형 범주형 변수들의 범주의 개수는 모두 5개라고 가정하자. 그렇다면 만들어질 수 있는 산점도는 20차원 중 2차원을 뽑아 만들 수 있는 조합의 수 190, 범주형 변수 5개가 5개의 범주로 쪼개어 질 수 있으므로 25, 임의 순열 방법이 쪼개어진 각 산점도당 1000번의 연산을 해야 하므로, 정리하자면 190 x 25 x 1000 = 4,750,000회의 scagnostics 분석을 해야 한다. 그리 크지 않은 25차원의 데이터에 대해서도 아주 많은 횟수의 분석을 해야 최종적인 결과가 나온다는 것이다. 이러한 사실은 이 기법을 상호적 데이터 탐색(interactive data exploration)에 이용하는 데에 큰 걸림돌이 된다. 사람은 10초 이상의 지연 시간이 발생할 경우 인지적으로 집중력을 잃게 된다는 연구[3]가 이를 뒷받침한다.

그래서 이 논문에서는 분할 변수 선택 방법을 실용적으로 시각적 분석에 사용하기 위해 점진적 시각화 기법을 접목시키는 시도를 하였다. Scagnostics 분석을 1000번 모두 수행할 동안 그냥 기다리는 것이 아니라, 1번 수행할 때마다 그 결과를 보여주며, 분석이 돌아감에 따라 계속 결과를 업데이트 하는 것이다. 계속 업데이트되는 결과를 보면 1000번 수행할 때까지 기다리지 않아도 점수가 높아 흥미를 불러일으키는 산점도들을 사전에 발견하고, 그것을 대상으로 더욱 깊이 있는 분석을 진행할 수가 있다.

2. 시각화 및 상호작용 기법

* 산점도 / 분할변수의 점수를 볼 수 있는 리스트 뷰

[그림 1]은 우리가 제안하는 시스템의 주된 화면에 해당하는 리스트뷰이다. 이 리스트의 항목 하나는 산점도와 분할 변수 하나를 나타낸다. 예를 들면

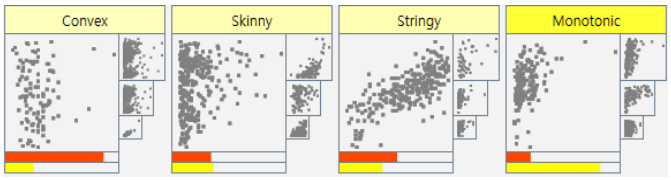

그림 2. Scagnostics 지표 별 상위 점수 산점도 요약 시각화

리스트뷰의 최상단에 위치한 1위 항목은 '수익_미국'을 X축으로 하고 Y축을 '로튼토마토_평점'으로 하는 산점도를 '장르'라는 분할 변수로 나눈 경우를 의미한다. 항목의 우측에는 'Max Partition Score' 라는 항목의 점수가 1000번의 계산을 하는 동안 지속적으로 업데이트된다. 이 점수가 높을수록 지정된 분할 변수로 산점도를 분할을 했을 때, 기존 산점도에서는 보이지 않던 패턴이 보이는 정도가 높다는 의미이다. 1000번 중에서 얼마나 계산을 하였는지는 항목의 진행 막대(progress bar)를 통해서 볼 수 있다.

* 항목의 우선 순위 설정

리스트뷰을 보면 'Priority'라는 항목이 있다. 이는 현재 Low, Mid, High의 3개의 단계로 나뉘어진 우선 순위를 보여준다. 항목들을 돌아가면서 한 번씩 계산을 하는 것을 한 '사이클'이라고 하자. 우선순위가 High인 항목의 경우는, 한 사이클에 원래 한 번 계산이 되어야 할 것을 10번의 scagnostics 점수를 계산하게 된다. Mid와 Low는 각각 5번, 1번씩 계산된다. 이러한 우선순위 개념을 도입한 이유는 항목들 중에서도 애초에 1000번을 돌려보지 않아도 최종적으로 점수가 높을 것 같은 항목들과, 이미 1000번 계산을 하기 전에 점수가 낮아서 다시 점수가 올라오기 힘든 항목들을 선별하여 우선 순위를 다르게 줌으로써, 컴퓨터의 연산 능력을 좀 더 '떡잎이 보이는' 우선 순위가 높은 항목들에 몰아 주자는 취지이다.

* Scagnostics 지표 별 상위 점수 산점도 요약

이 시스템에서 scagnostics 점수를 구하기 위해 사용하는 지표는 9 가지이다. 이 9 가지는 scagnostics 개념을 이용해 데이터 분석을 최초로 시도한 Wilkinson의 연구[4]에서 제시되어 있는 9가지 지표이다. [그림 2]는 그 중에서도 4가지의 지표에 대한 요약 결과를 표시하고 있는 모습이다. 각 지표 별로 가장 큰 산점도는 '나뉘어 졌을 때 기존과 다른' 정도가 가장 큰 산점도이다. 이 점수를 '분할 점수' 라고 하자. 그 다음 크기의 산점도는 분할 점수가 다음으로 큰 산점도이다. 그렇게 각 지표당 대표적으로 분할 점수가 큰 산점도 4개가 표시된다. 분할 점수는 산점도 하단의 노란색 막대로 표시된다. 막대가 길수록 점수가 높은 것이다. 반면 빨간색 막대의 경우는 분할 점수가 아니라 '지표 점수'이다. 지표 점수는 나뉘어지거나 합쳐진 것과는

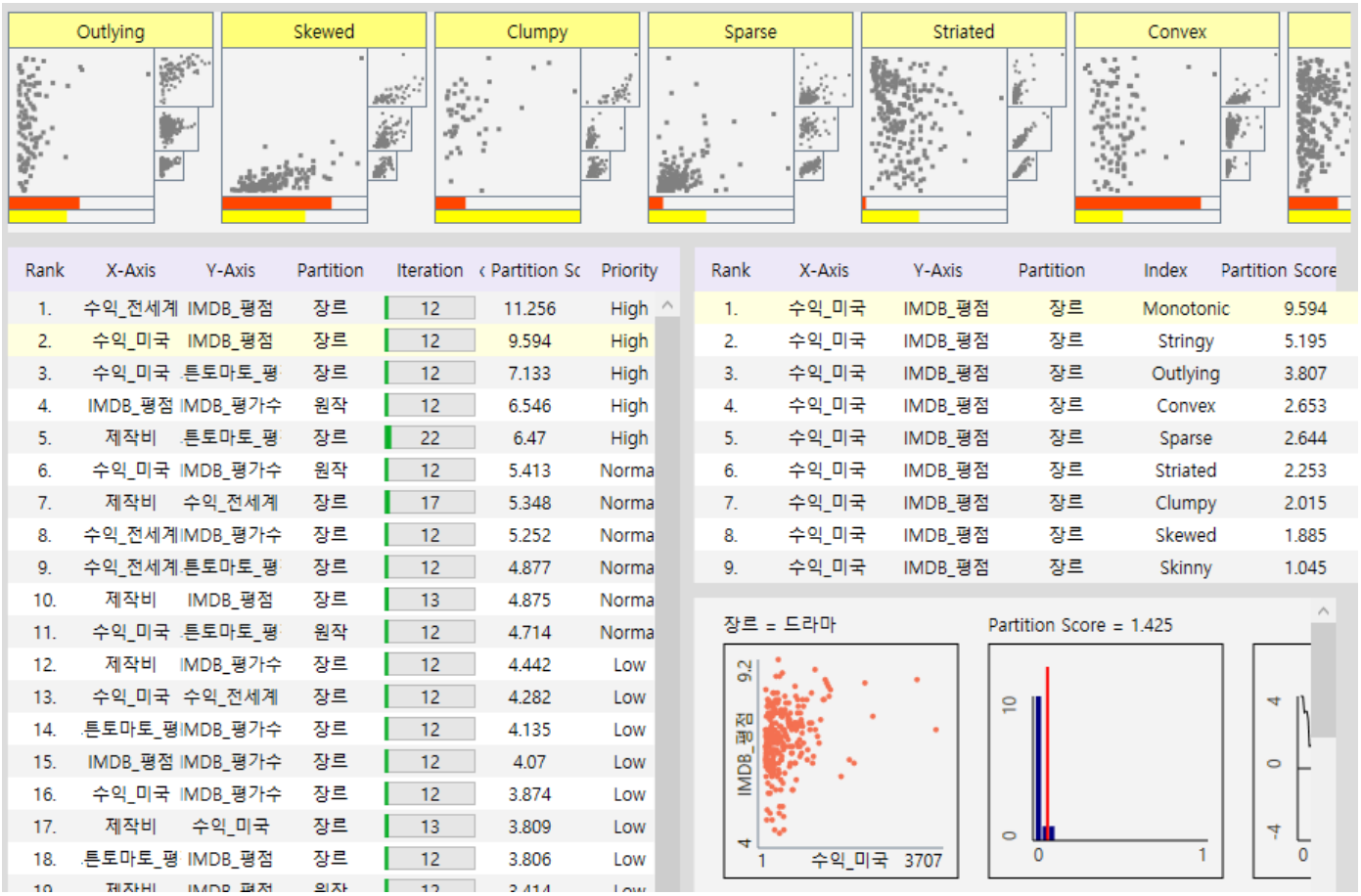

그림 3. 분석 시스템의 전체 구성. 상단에는 [그림 2]에서 소개한 지표 별 요약 시각화가 있으며, 좌측 하단에는 [그림 1]의 리스트 뷰가 있다. 우측의 디테일-리스트 뷰는 좌측 리스트 뷰의 항목 하나를 각 9개의 지표 별로 확장한 항목을 보여준다. 디테일-리스트 뷰의 항목을 클릭하면 우측 하단과 같이 그에 해당하는 분할된 산점도들이 나열된다.

상관없이 순수하게 해당 산점도가 얼마나 'Convex'한지 혹은 'Skinny'한지에 해당하는 점수이다. 즉 지표에 해당하는 특징을 얼마나 강하게 갖고있는 지에 대한 점수라고 다시 표현해볼 수 있겠다. 분할 점수와 지표 점수를 굳이 동시에 표기한 이유가 있다. 예를 들어, 분할 점수는 매우 높은 반면 Convex 지표 점수가 낮은 산점도의 경우는 "나뉘어 졌을 때 기존 합쳐진 산점도와는 매우 다른 패턴을 보이지만, 막상 산점도의 패턴 자체가 그리 볼록(convex)하지는 않다"는 것을 의미한다. 아마 원본 산점도나, 같은 분할 변수 내에서 다른 범주를 갖고 나뉘어진 산점도들은 볼록하지만 이 산점도의 경우만큼은 그렇지 않다는 해석을 유추해 볼 수 있다.

3. 결론 및 향후 연구

이 연구에서는 scagnostics 기법 기반의 분할 변수 선택 기법을 상호적 데이터 분석 시스템에 사용하기 위해 점진적 분석 및 시각화를 접목시키는 시도를 하였다. 이 시스템을 통해서 단순히 2차원의 산점도를 분석하는 것을 넘어서 분할 변수를 통해 한 차원 더 들어간 분석을 함으로써 기존에는 발견하지 못한 데이터의 패턴들을 발견할 수 있을 것으로 기대된다. 향후에는 이 시스템을 이용하여 데이터를 실제로 분석해보는 파일럿 실험을 진행하여, 결과를 시스템 디자인에 반영한 뒤, 좀 더 엄밀한 Case 실험을 통해 시스템의 유효성을 검증할 것이다.

참 고 문 헌

[1] Hafen, R., Gosink, L., McDermott, J., Rodland, K., Dam, K.-V., Cleveland, W. Trelliscope: A system for detailed visualization in the deep analysis of large complex data. In *Proc. LDAV*, 105-112, 2013.

[2] Anand, A., Talbot, J. Automatic Selection of Partitioning Variables for Small Multiple Display. In *IEEE TVCG*, 22(1), 669-677, 2016.

[3] Nielsen, J. Response Times: the Three Important Limits. In *Usability Engineering*, Morgan Kaufmann Publishers Inc., 1993.

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