다차원 데이터의 데이터 공간 탐색을 위한 Cognostics 기반 시각적 분석 시스템

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

A Visual Analytics System for Exploring Data Space of Multidimensional Data Using Cognostics

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

요 약

이 논문에서는 Cognostics 기반의 분석 기법들이 차원 공간의 탐색에 치우쳐져 있는 점을 주목하여, 그동안 잘 다뤄지지 않았던 데이터 공간의 탐색을 중점적으로 다룬다. 다차원 데이터 분석을 위해 산점도를 만든다면 X, Y축에 여러 변수를 지정해서 만들어 볼 수 있을 것이다. 이렇게 만들 수 있는 수많은 산점도들 가운데 유의미한 패턴을 갖는 것을 우선적으로 보자는 개념이 Cognostics 기반 분석이다. 그러나 데이터의 레코드 전체가 아닌, 특정 레코드들만을 뽑아 산점도를 만든다면 기존 산점도에서 보이지 않았던 패턴들이 보일 수 있다. 이러한 특성에 초점을 맞춰 우리는 사용자가 Cognostics를 이용하여 데이터 공간을 효과적으로 탐색할 수 있는 시각적 분석 시스템을 제안한다.

1. 서 론

다차원 데이터는 일반적으로 3차원 이상으로 구성된 데이터를 일컫는다. 만약 데이터가 2차원으로서, 두 개의 변수로만 구성이 되어 있다면 각각을 X축, Y축으로 하는 산점도(scatter plot)를 만들어 살펴보면 그 데이터의 패턴을 쉽게 확인할 수 있다. 그러나 다차원 데이터의 경우는 변수가 많기 때문에 한꺼번에 많은 차원을 확인할 수 있는 산점도보다 복잡한 시각화를 사용하거나, 혹은 여러 변수의 조합으로 만들 수 있는 다수의 산점도들을 모두 살펴보아야 한다. Cognostics[1] 기반 기법이란 다차원 데이터를 통해 만들어 볼 수 있는 수많은 시각화들에 대해 특정한 기준으로 점수를 부여한다. 사용자는 그 점수를 기준으로 유의미한 패턴을 갖는 시각화를 선별적으로 관찰할 수 있다.

기존에 이러한 Cognostics를 바탕으로 한 여러 시각적 분석 도구가 고안되어왔다[2,3,4,5]. 그러나 대부분의 경우 어떠한 특정한 차원을 선택하여 시각화를 만들 것 인지에 대한 문제, 즉 차원 공간(dimension space)의 탐색에 치우쳐져 있다. 주어진 데이터 전체가 아니라 특정 부분만을 대상으로 Cognostics 분석을 할 경우, 시각화들이 갖는 패턴 양상은 매우 달라질 수 있다. 예를 들어, 학생들의 시험 성적 데이터를 Cognostics로 분석할 경우, X축을 "결석일수", Y축을 "중간고사 성적" 으로 갖는 산점도의 점수를 내어 볼 수 있다. Cognostics 점수 지표는 피어슨 상관 계수(Pearson's correlation coefficient) 등으로 설정할 수 있다. [그림 1]은 여기에 범주형 변수인 "자유 시간"이 "아주 적음"과 "아주 많음"인 경우만을 뽑아 점수를 부여한 결과이다. 기존에는 별로

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

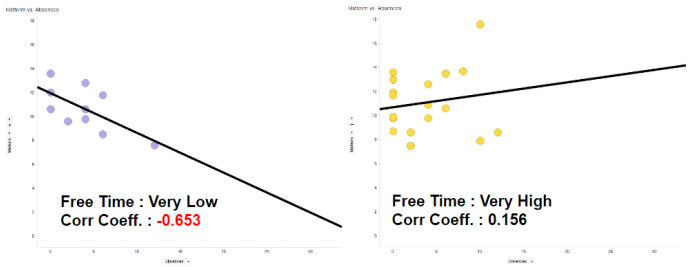

그림 1. 전체 데이터를 "자유 시간" 이라는 범주형 변수로 나눈 결과이다. Cognostics 점수의 차이를 발견할 수 있다.

유의미한 점수를 갖지 못했던 산점도 일지라도 이처럼 자유 시간을 적게 갖는 그룹의 경우 상관계수가 -0.653으로 상당히 의미 있는 음의 상관계수를 갖게 된 것을 알 수 있다. 이렇게 Cognostics 기반의 점수를 부여하더라도, 데이터 전체가 아닌 특정 부분만을 추출하여 같은 시각화를 나타낼 경우, 그 패턴이나 점수는 상당히 달라 질 수 있음을 볼 수 있다. 우리는 이러한 점에 착안하여 데이터의 차원 공간 뿐 아니라 데이터 공간(data space)을 탐색할 수 있는, 즉 데이터의 부분집합들 간의 Cognostics 분석 결과의 차이를 보여주고 이를 사용자가 상호작용을 통해 주도적으로 검토할 수 있는 시각적 분석 시스템을 제안한다.

2. 시각화 및 상호작용 기법

이 연구에서 제시하는 시각적 분석 시스템은 수치형(Numerical) 변수들에 대한 시각화를 평가한다. 분석의 대상이 되는 시각화는 히스토그램(Histogram)과 산점도이며, 각각은 변수 한 개(1D)와 두 개(2D)로 구성할 수 있는 시각화이다. 범

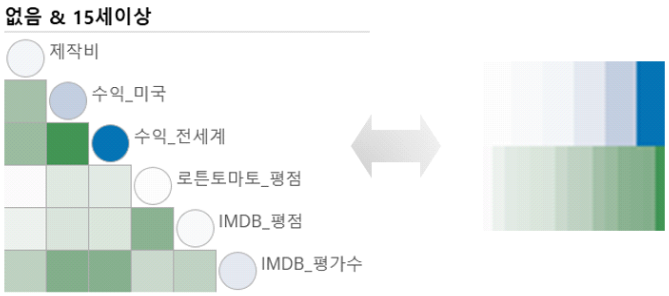

그림 2. 부분집합 하나의 Cognostics 결과를 나타내는 두 인터페이스이다. 좌측은 결과를 자세히, 우측은 이를 축약하여 Overview 상에서 보여준다.

주형(Categorical) 변수들은 데이터의 부분집합을 나누는 기준으로 사용한다.

인터페이스는 크게 부분집합들에 대한 분석 결과 간 유사도 등을 한눈에 파악할 수 있는 Overview, 자세한 수치를 볼 수 있는 Detail, 그리고 유사도를 어느 정도까지 보여줄 것인지에 대한 기준인 임계값을 설정하는 Control Panel로 나뉜다.

- **Cognostics 분석 결과 시각화**

[그림 2]의 좌측 대각 행렬 모양의 인터페이스는 데이터의 Cognostics 분석 수행결과를 나타낸다. 대각선상의 원들은 각 변수들로 나타낼 수 있는 히스토그램의 지표 계산 결과를 나타낸다. 산점도의 경우, 예컨대 "제작비"와 "수익_미국" 이라는 두 원이 만나는 사각형 셀(Cell)로 나타내며, 이는 위 두 변수가 X, Y축이 되어 구성된 산점도이다. 원과 셀의 색깔이 곧 해당 히스토그램, 혹은 산점도의 지표점수를 나타낸다. 따라서 만들어 볼 수 있는 여러 산점도의 지표 점수 분포를 한눈에 파악 할 수 있다. 예를 들면, [그림 2]에서 산점도에 대한 지표가 피어슨 상관계수라고 한다면 "수익_미국"과 "수익_전세계"로 만든 산점도가 가장 색이 진하므로, 상관계수가 가장 높음을 알 수 있는 것이다.

데이터의 부분집합마다 이렇게 Cognostics 분석을 한다면, 행렬 인터페이스가 부분집합의 개수만큼 있어야 할 것이다. 그러나 이는 비교적 공간을 많이 차지하기에 이러한 행렬들 여러 개를 한 눈에 보여주기는 힘들다. [그림 2]의 우측에 있는 줄무늬 사각형은 Cognostics 결과 행렬을 축약하여 놓은 인터페이스이다. 이를 앞으로 노드(Node)라고 칭할 것이다. 각 셀이 갖는 값이 좌에서 우로 오름차순으로 정렬되어, 히스토그램 및 산점도들이 갖는 점수 분포 파악이 가능하다. 중심부를 기준으로 상단은 히스토그램, 하단은 산점도에 대한 점수를 나타낸다.

- **Overview – 부분집합간 관계를 전반적으로 파악**

[그림 3-1]는 위에서 설명한 부분집합 인터페이스를 이용하여 분석의 전반적인 결과를 파악할 수 있는 Overview 역할을 한다. 현재 분석 중인 데이터는 영화 데이터이며, 부분집합을 나누는 기준을 "등급", 그리고 "원작" 이라는 두 개의 범주형 변수로 설정해놓은 상황이다. 이에 대해 히스토그램(변수 한 개)의 경우 정규성(Normality)을, 산점도의 경우는 피어슨 상관

그림 3. 시각 분석 시스템의 전체 화면.
두 범주형 변수 "등급"과 "원작"을 기준으로 격자형으로 부분집합을 나눈 뒤, 각각에 대해 Cognostics 분석을 수행한 결과이다. 격자 사이의 에지를 통해 각 부분집합들 간의 결과의 차이를 확인할 수 있다.

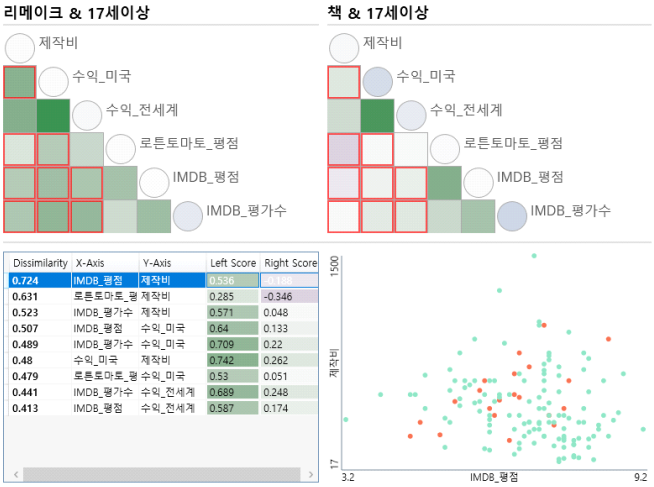

그림 4. 두 부분집합 간의 결과를 비교할 수 있는 Detail 인터페이스이다.

계수를 지표로 하여 Cognostics 분석을 한 모습을 나타내고 있다. 각 노드들 사이에는 빨간, 혹은 초록 에지(Edge)가 존재하는데 이는 각각 얼마나 차이가 나는지 혹은 유사한지를 나타낸다. 예를 들면, '17세 이상 & 리메이크'에 해당하는 부분집합과 '17세 이상 & 책'에 해당하는 부분집합은 빨간 에지가 매우 굵고 선명하므로 다른 부분집합들에 비해 그 Cognostics 분석 결과가 많이 차이가 난다는 것을 의미한다.

- **Detail – 두 부분집합 간의 결과 비교**

Overview 인터페이스 상의 노드와 그 사이의 에지의 굵기 및 색깔로는 전반적인 결과 파악은 가능하지만, 부분집합 간 실제로 어떤 차이가 있는지에 대해서는 알 수가 없다. 이를 보기위해 에지를 클릭하면 [그림 4]와 같이 두 노드를 비교하기 위한 팝업(popup) 인터페이스가 나타난다. 인터페이스 상단의 좌, 우측은 두 노드가 갖는 행렬 인터페이스이다. 행렬 인터페이스 내에 존재하는 산점도들 가운데서, 사용자가 지정한 지표 결과 값(임계값) 간의 차이를 넘는 산점도의 경우는 빨간 테두리가 쳐져있어서 어느 부분에서 차이가 나는지를 알 수 있다. 예를 들면 좌측 부분집합의 "제작비"와 "수익_미국"으로 이루어진 산점도의 지표 값은 0.742, 우측 부분집합의 해당 산점도는 0.262라면, 차이 값은 0.742 – 0.262 = 0.48 이 된다. 이 때 사용자가 정해놓은 임계값이 0.4라면, 0.48은 그 이상이 되므로, 두 산점도의 셀에는 빨간 테두리가 쳐지게 되는 것이다. 하단부의 좌측에는 그 점수가 실제로 얼마인지, 또한 얼마만큼의 차이가 나는지를 표 형태로 볼 수 있으며, 하단부의 우측에는 실제 산점도 상에서 어떻게 차이가 나는지 까지 확인할 수 있다.

- **Control – 동적 쿼리 수행이 가능한 유사도 및 비유사도 임계값 설정**

[그림 3-2]는 사용자가 직접 유사도 혹은 비유사도의 임계값을 설정함으로써 노드들 간에 나타나는 에지 들을 조절할 수 있는 인터페이스이다. 슬라이더를 좌우로 움직임으로서, 실시간으로 그 결과의 변화를 확인할 수 있는 동적 쿼리(dynamic querying)[6] 상호작용이 가능하다. 현재는 유사도의 임계값이 0.2, 비유사도의 임계값이 0.4로 설정되어 있는 모습이다. 이는 같은 변수로 만들어진 산점도간에 점수 차이가 0.2 이하이면 유사한 것으로 보겠다는 것이고, 마찬가지로 0.4 이상 차이가 날 경우 유사하지 않은 것으로 간주하겠다는 의미이다.

3. 결론 및 향후 연구

본 연구에서는 기존의 Cognostics 개념을 이용한 분석들이 데이터의 부분집합, 즉 데이터 공간을 탐색하는 데에는 미흡했다는 점을 보완하기 위해 시각적 분석 인터페이스를 제시하였다. 또한 분석의 용이성을 위한 상호작용 기법을 소개하였으며 향후에는 이 분석 기법을 실제 유저에게 배포하여 이를 정량적, 정성적으로 평가하고 그 결과를 바탕으로 더욱 깊이 있는 분석을 가능케 하는 연구를 진행할 예정이다.

참 고 문 헌

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

[2] Seo, J., Shneiderman, B. A Rank-by-Feature Framework for Interactive Exploration of Multidimensional Data, Information Visualization, 4(2), 96-113, 2005.

[3] Wills, G., Wilkinson, L. AutoVis: automatic visualization, Information Visualization, 9(1), 47-69, 2010.

[4] 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 Large-Scale Data Analysis and Visualization (LDAV), 2013 IEEE Symposium on, 105-112, 2013.

[5] Dang, T., N., Wilkinson, L. ScagExplorer: Exploring Scatterplots by Their Scagnostics, In Proc. IEEE Pacific Visualization Symposium, 73-80, 2014.

[6] Shneiderman, B. Dynamic Queries for Visual Information Seeking, in IEEE Software, 11(6), 70-77, 1994.