합동적 Cognostics 프레임워크를 이용한 다차원 데이터의 시각적 분석

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

A Visual Exploratory Data Analytics Framework for Combined-Cognostics of Multidimensional Data

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

요 약

이 논문에서는 다차원 데이터에 대한 시각적 탐색 분석 방법론으로서 합동적 Cognostics 프레임워크를 제시한다. 이는 Cognostics라는 데이터 분석 방법을 크게 세 가지 측면(차원, 지표, 데이터 부분집합)에서 확장한 개념이다. 이를 통해 사용자는 분석에 유익할만한 시각화를 찾을 수 있고, 여러 지표들을 한꺼번에 적용한 결과를 봄으로써 분석을 기존에 비해서 효율적으로 수행할 수 있다. 또한 데이터에서 따로 관심이 있는 부분들을 차트 등의 시각화에서 손쉽게 추출하여 동일한 분석을 수행할 수 있다.

1. 서 론

다차원 데이터의 시각적 분석이란, 일반적으로 3차원 이상의 데이터를 인간이 인지하기에 효과적인 차트나 시각화 기법을 사용하여 보여줌으로써 데이터에 대한 통찰을 얻는 것을 말한다. 3차원 이상의 데이터를 2차원에 해당하는 모니터에 나타내는 방식으로 분석을 돕는 기술 중에 사영 추적 (Projection Pursuit)[1] 방법이 있다. 이 방법은 많은 변수들 중 가장 데이터 분석에 유익할 법한 2개의 변수 쌍을 선별하여, 인간이 자연스럽게 인지할 수 있는 2차원 시각화로 보여준다. 위의 설명에서 유추할 수 있듯이, 2개의 변수의 쌍, 즉 좋은 사영(Projection)을 어떠한 기준으로 선별하는가의 문제가 이 방법의 핵심이라고 볼 수 있다. 이렇게 변수 혹은 변수의 쌍을 고를 때 1985년에 Tukey가 고안한 Cognostics[2]라는 개념이 사용될 수 있다. 이는 인간이 관심을 가질 만한 시각화를 "컴퓨터의 도움을 받아 진단 한다"는 의미로 해석할 수 있다. 그림 1을 보면 여러 산점도들이 나오는데 기준을 이상치(Outlying) 으로 했을 경우, 그에 따른 점수가 0점부터 1점까지 매겨질 수 있다. 기준을 줄무늬가 있는(Striated)로 두었을 때에도 마찬가지로 산점도에 대해 점수를 매길 수 있다. 이러한 방식으로 인간은 특정 시각화에 매겨진 점수를 통해, 해당 시각화가 어떠한 측면에서(보통은 선택한 지표가 기준) 어떤 성질을(점수를) 가지고 있는지를 대략적으로 파악할 수 있고, 그 때문에 좀 더 관심 있게 볼 시각화를 선택하는 데에 큰 도움을 받을 수 있다.

이 연구에서는 기존 Cognostics 개념을 이용한 데이터 분석 연구들을 확장하는 시각적 탐색 분석 방법으로서, 합동적 Cognostics 프레임워크(Combined-Cognostics Framework. CCF) 라는 개념을 제시한다.

이 논문은 2016년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임 (No. NRF-2011-0030813).

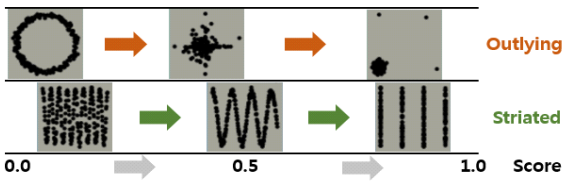

그림 1. 여러 산점도에 대해 두 지표로 Cognostics를 수행한 결과이다. ScagExplorer[3]의 그림을 재구성하였다.

2. 합동적 Cognostics 프레임워크

이 연구에서 제시하는 프레임워크는 크게 차원간, 지표(Index)간, 데이터 부분집합(Subset)간 합동적 Cognostics로 분류될 수 있다.

- **차원간 합동적 Cognostics**

시각화의 종류나 방식에 따라서 그것이 표현하고자 하는 변수의 개수, 즉 차원은 다를 수 있다. 히스토그램의 경우는 1개의 변수를, 산점도의 경우는 2개의 변수를 시각화한다. Cognostics의 개념을 Feature라는 표현으로 구현한 기존 연구인 Rank-by-Feature[4]의 경우는 먼저 각각의 변수들의 특징을 파악하고, 여기에서 관심을 갖게 된 변수들을 위주로 해당 변수와 다른 변수간의 2차원 관계에 대해 파악하는 순으로 분석을 진행한다. 이러한 프로세스를 조금 더 발전시킨 형태로서 CCF에서는 1차원 Cognostics의 결과를 보여주는 한편 선택적으로 2차원 Cognostics의 결과를 동시에 시각화할 수 있게 하였다. 이러한 '차원간의 합동적 Cognostics'라는 개념을 통해 사용자는 1차원과 2차원의 Cognostics의 결과를 비교하고 그 관계를 효과적으로 파악할 수 있다.

- **지표간 합동적 Cognostics**

같은 차원에 대한 Cognostics 지표의 경우에도 시각화의 종류별로 다른 여러 가지가 존재할 수 있다. 어떠한 지표를 사용

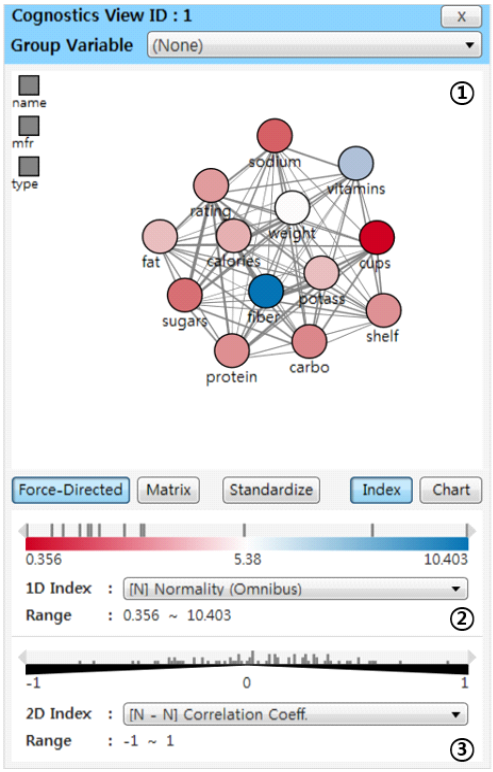

그림 2. Cognostics 유닛
Cognostics를 수행하는 분석의 기본 단위 인터페이스

하느냐에 따라 Cognostics의 결과가 달라지고, 그에 따라서 사용자가 관심을 갖고 분석할 변수 또는 시각화가 달라질 수 있다. 이 때문에 사용자는 다양한 지표를 사용하여 결과를 다각적으로 분석, 비교해볼 요인이 생긴다. CCF에서는 이러한 사용자의 요구를 충족시키기 위해서 작업 공간상에 다양한 지표를 활용한 분석 결과를 동시에 파악할 수 있게 해주는 '지표간 합동적 Cognostics'라는 개념을 제시한다.

- **데이터 부분집합간 합동적 Cognostics**

일반적으로 사용자는 처음에 전체 데이터 레코드(Record)를 모두 포함하여 분석을 수행한다. 하지만 분석의 과정에서 데이터의 특정 레코드들만을 추출하여 따로 분석을 해보고, 이를 전체 데이터를 대상으로 분석한 결과와 비교함으로써 선택한 데이터의 부분집합이 어떤 특징을 가지고 있는지 파악하고자 하는 사용자의 요구가 있을 수 있다. 그에 따라 CCF에서는 분석 결과로 나온 차트들에서 자유롭게 마우스 드래그(Drag)를 통해 관심이 가는 데이터의 특정 부분집합만을 따로 추출할 수 있게 하였다. 그를 이용해 다시금 Cognostics를 수행하고 그 결과를 원본 데이터의 분석 결과와 한 공간에서 보기 쉽게 비교해볼 수 있는 개념으로 우리는 '데이터 부분집합간 합동적 Cognostics'라는 개념을 제시한다.

3. 시각화 및 상호작용 기법

본 연구에서 제시하는 세 가지의 합동적 Cognostics 개념을 효과적으로 수행하기 위해서 사용자 인터페이스인 Cognostics 유닛(Unit)을 설계하였다(그림 2). Cognostics 유닛은 크게 세 부분으로 구성되는데, 가장 위의 부분은 데이터 변수들의 Cognostics 수행 점수를 한 눈에 볼 수 있는 Overview(그림 2-1)이다. 그 다음으로는 1D와 2D 지표를 설정하는 부분(그림 2-2, 2-3)이 있다. 다음에서는 Cognostics 유닛의 주요 시각화 및 상호작용 기법들에 대해 소개한다.

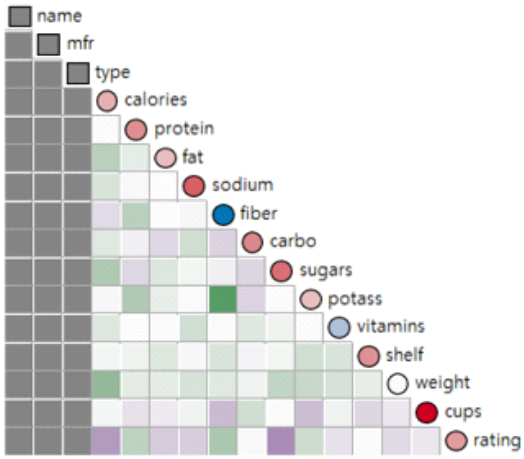

그림 3. 행렬 시각화를 이용한 Overview

- **결과를 한 눈에 파악할 수 있는 Overview**

그림 2-1의 Overview는 노드-링크 다이어그램 (Node-Link Diagram) 혹은 행렬 형태의 (Matrix View) 시각화를 통해 변수들의 Cognostics 결과를 보여준다. 노드-링크 다이어그램과 행렬 시각화를 통해 Overview를 표현한 이유는 두 시각화 모두 1D와 2D의 정보를 한꺼번에 표현 가능하며 상호보완적인 성격을 갖고 있기 때문이다[5]. 노드-링크 다이어그램의 경우는 노드(Node)의 색깔로 1D 점수를, 링크(Link) 혹은 에지(Edge)의 굵기로 2D 점수를 나타낼 수 있다. 행렬 시각화의 경우는 각 열의 위에 있는 노드 형태의 인터페이스의 색깔로 1D 점수를, 행렬 상의 각 칸(Cell)의 색깔로 2D 점수를 나타낼 수 있다.

그림 2-2, 와 그림 2-3은 각각 1D, 2D에 지표를 설정하고 그에 따른 결과 점수를 볼 수 있는 인터페이스이다. 각각에는 작은 회색 막대들로 이루어진 히스토그램(Histogram)이 있어서 변수들(1D) 혹은 변수들 간의 관계(2D)가 갖는 전반적 점수 분포를 알 수 있다.

- **필요한 부분만을 집중할 수 있는 Filtering**

데이터를 탐색할 때, 사용자들은 모든 정보를 한 눈에 보고 싶어 하기도 하지만 조금 더 깊은 수준의 분석을 위해 원하는 부분들 외의 정보는 모두 Filtering을 함으로써 중요 부분들만을 좀 더 집중적으로 보고자 하는 요구를 가질 수 있다. 이를 위해 그림 2-2와 2-3의 히스토그램 양 끝단에 삼각형 모양의 인터페이스를 만들어 이를 좌우로 드래그함으로써 원하는 점수대를 갖는 노드와 링크만을 Overview에서 표현해줄 수 있다. 이는 행렬 시각화에서도 마찬가지로 수행이 된다.

- **하나 혹은 두 변수의 분포를 나타내는 차트 시각화**

분석을 통해서 관심이 생기는 변수 혹은 변수들이 생기면 사용자는 이것들이 갖는 분포를 직접 확인해보고자 한다. 그를 위해 Overview에서 각 변수, 혹은 변수 간 관계에 해당하는 인터페이스를 선택 시, 하단에 차트가 띄워진다. 지원하는 차트는 히스토그램과 산점도이다.

- **지표간 합동적 Cognostics**

Cognostics 유닛 하나에는 각각 한 번에 하나의 1D, 그리고 2D 지표만을 사용할 수밖에 없다. 그러나 여러 개의 유닛을 사용한다면 한 번에 여러 지표가 적용된 결과를 확인할 수 있다. 관심이 가는 지표를 각각의 유닛에 적용 후, 그에 대한 결과를 교집합, 혹은 합집합 시킨 결과를 하나의 유닛에 표현할 수 있는 인터페이스 및 상호작용을 제공한다. 그림 4는 2개의 유닛

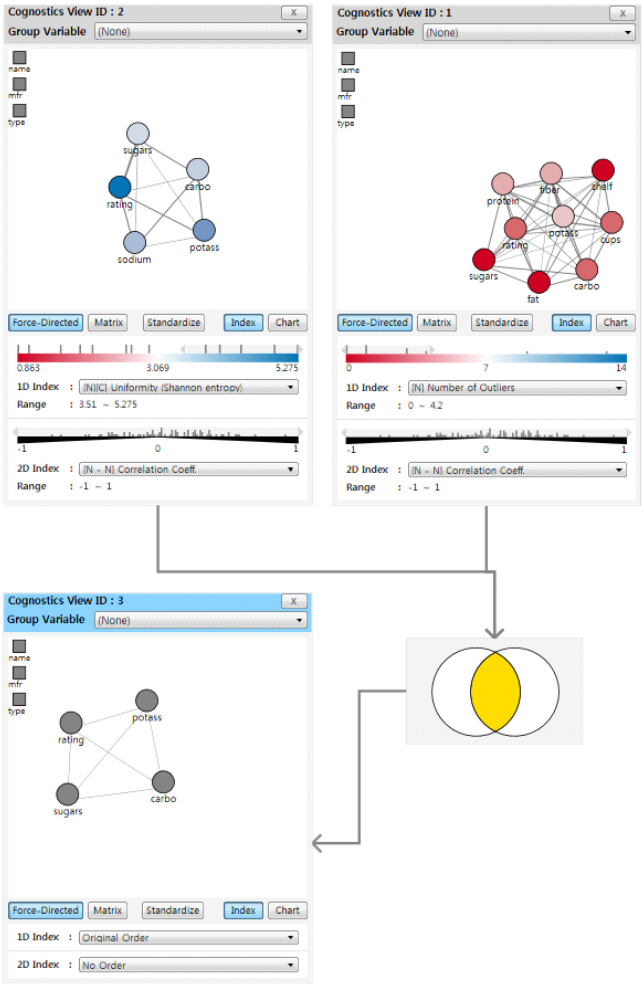

그림 4. 두 개의 유닛의 결과의 교집합에 해당하는 결과가 하나의 유닛에 표시된다. '지표간 합동적 Cognostics' 개념을 실제 인터페이스 상으로 수행하는 모습이다.

내에서 지표를 다르게 설정 후, 사용자가 원하는 대로 Filtering을 수행한 뒤, 두 유닛에 공통적으로 나타나는 변수, 혹은 변수간의 관계의 교집합을 하나의 유닛에 표시하는 모습이다. 이렇게 하면 가령 'Uniformity는 높고, Outlier는 적은 변수들을 찾기'와 같은 작업을 효과적으로 수행할 수 있게 되는 것이다.

- **데이터 부분집합간 합동적 Cognostics**

사용자는 위에서 기술한 대로 원하는 변수 혹은 변수간의 관계에 대해 히스토그램이나 산점도를 띄울 수 있는데, 해당 차트에서 원하는 부분의 데이터 부분집합만을 선택 후, 빈 공간에 드래그를 함으로써 그 데이터만을 갖는 유닛을 추가적으로 만들 수 있다. 그림 5의 경우는 변수 'calories'가 비교적 높은 데이터만을 추출한 뒤, 기존 유닛과 같은 지표를 설정하여 그 결과의 차이를 확인하고 있는 모습이다. 그림에서 볼 수 있듯이 같은 지표를 설정하였는데도 Overview의 전반적인 Node 색깔이 다름을 알 수 있다.

4. 데이터 분석 및 논의

2개의 범주형 변수와 13개의 연속형 변수를 가진 Cereals 데이터[6]를 이 연구에서 제시한 프레임워크를 통해 분석했다. 각각의 변수들에 대해 먼저 정규성(Normality)을 측정해보았다. 그 결과, 'fiber'라는 변수의 정규성이 가장 낮음을 노드-링크 다이어그램을 통해 한 번에 파악할 수 있었다. 이 변수에 대해 더 자세

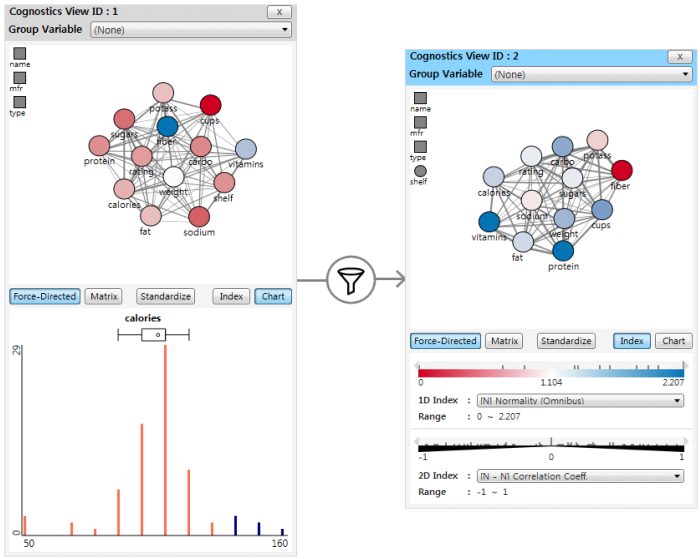

그림 5. 변수 히스토그램(좌하단)에서 원하는 부분(남색 막대)을 선택 후, 그 데이터만의 Cognostics 유닛(우측)을 생성하여 분석한다. '데이터 부분집합간 합동적 Cognostics' 개념을 실제로 수행하는 모습이다.

히 알아보고자, 해당 변수의 분포를 자세히 볼 수 있는 히스토그램을 띄웠고, 거기에서 수치가 1.5 * IQR을 초과하는 높은 수치를 지닌 데이터들만을 추출하여 새로운 Cognostics 유닛을 생성하였다. 그 유닛에서 각 변수들 간의 상관계수(Pearson's Correlation Coefficient)를 측정하자, 기존 전체 데이터에서는 발견할 수 없었던 변수들 간의 상관관계를 발견할 수 있었다. 이처럼 분석의 시작점이 막연한 상황에서 Cognostics를 통해 계속적인 분석의 실마리를 제공한다는 점에서 CCF의 의의를 찾을 수 있었다.

5. 결론 및 향후 연구

본 연구에서는 기존 시각적 탐색 분석 연구에서 사용되는 Cognostics 라는 개념을 확장하여 좀 더 효과적인 데이터 분석을 수행할 수 있는 세 가지 방법을 합동적 Cognostics라는 새로운 프레임워크로서 소개하였다. 향후에는 이 프레임워크를 실제 유저에게 배포하여 이를 정량적, 정성적으로 평가하고 그것을 바탕으로 더욱 개선시키는 연구를 진행할 예정이다.

참 고 문 헌

[1] Huber, P. J. Projection pursuit, The Annals of Statistics, 13(2), 435-475, 1985.

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

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

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

[5] Ghoniem, M., Fekete J.-D., Castagliola, P. A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations, In Proc. Information Visualization, 17-24, 2004

[6] Cereals Datafile. Retrieved from http://lib.stat.cmu.edu/DASL/Datafiles/Cereals.html