www.kiise.or.kr

ISSN 1229-6821

제34권 제11호 통권 제330호

정보과학회지

Communications of the Korean Institute of Information Scientists and Engineers

2016. 11

데이터 시각화와 시각적 분석

한국정보과학회
Korean Institute of Information Scientists and Engineers

정보과학회지

2016. 11
제34권 제11호
통권 제330호

목차

Communications of KIISE

November 2016

Vol. 34 No. 11
Serial No. 330

Contents

특집원고

다차원 데이터의 의미적 군집 분석을 위한 시각화 방법에 관한 연구

아주대학교 ▌하효지·한현우·배성윤·이지혜·손상준·홍창형·신현정·이경원

1. 서 론

다차원 데이터(Multidemensional data)는 많은 양의 변수(Variables)가 포함된 데이터를 말하며, 이를 효율적으로 정제하고 분석하기 위한 연구가 다양하게 이루어지고 있다[1,2]. 그 중 다차원 데이터를 이용한 시각화(Visualization)는 데이터의 차원을 분석하기 쉽도록 축소하고, 동시에 데이터가 가지고 있는 특성을 최대한 보여주는 것이 핵심이다[3]. 때문에 다차원 데이터를 시각화할 때는 데이터의 특성을 제시하기 위해 군집단위의 분석이 주로 이루어지며, 이를 위해 데이터마이닝 기법 중 '군집화(Clustering)' 기법이 자주 사용된다[4,5,6]. 군집화는 데이터의 속성과 유사도에 따라 데이터를 분류함으로써 다차원 데이터를 특성에 맞게 정제하는 기법이다[7].

그러나 기존에 쓰이는 군집화 기법은 데이터를 군집하는 과정에 있어서 군집의 일부가 유의미한 결과를 보여주지 못하는 경우가 있다[8]. 예를 들어 나누고자 하는 집단의 수를 정한 뒤 군집화 분석을 진행한다고 가정하면, 그 중 일부 군집은 개체 수가 현저히 작거나 데이터의 특성을 의미적으로 해석할 수 없는 상황이 발생한다. 이러한 경우에는 최적의 군집 결과물을 얻기 위해 사용자가 군집의 수를 다시 지정할 수밖에 없다.

군집 분석을 위한 시각화 기법은 군집 결과를 네트워크로 표현하거나 2D radvis 기법, Parallel coordinate 기법을 활용하는 경우가 있다[9,10]. 그러나 이상의 시각화 기법은 군집화의 결과물을 의미적으로 시각화되었는지에 대해 즉각적으로 파악하기 쉽지 않다[8]. 또한, 데이터의 차원을 축소하는 과정에서 변수의 개수에 제한을 두는 경우가 있다[10].

† 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2015R1A5A7037630)

따라서 군집 분석을 위한 시각화 시스템을 제작하기 위해서는 군집단위로 보는 것 이외에도 각 군집의 세부적인 특성도 함께 보는 것이 필요하다. 더불어서 시각화를 분석하는 사용자가 원하는 대로 변수를 조합하여 다차원적 분석이 가능한 시스템을 제안하는 것이 필요하다. 본 논문에서는 다차원 데이터를 의미적으로 군집할 수 있는 시각화를 제안하는 것을 목적으로 하며, 아래와 같은 연구 절차를 거쳤다.

첫째, 다차원 데이터 군집 시각화의 기존 사례 중 Radvis 및 Parallel coordinate를 기반으로, 각각의 시각화 기법이 다차원 데이터의 군집 구조를 어떻게 나타내고자 했는지 동향을 파악하였다. 그 과정에서 시각화 기법에 대해 정리하고 그에 대한 장점과 단점을 정리하였다.

둘째, 다차원 데이터를 활용한 군집 분석 시각화를 위해 3D radvis 기법과 Parallel coordinate 기법을 함께 활용하였다. 이 때, 3D radvis 기법은 각 군집에 대한 의미적 해석을 위해 활용할 수 있고, Parallel coordinate 기법은 다차원 데이터가 가지는 세부적인 특성을 파악하기 위해 활용할 수 있다는 점을 보여주었다.

셋째, 다차원 데이터를 기반으로 군집 분석을 시행할 때 일부 집단을 선정하여 세분화할 수 있는 방법을 제안하였다. 그리고 선택된 군집을 세분화하기 위해 사용된 개발 과정을 정리했다.

넷째, 시각화 시스템의 전체적인 인터렉션 기능을 소개하였다. 또한, 케이스 스터디를 수행하여 군집 분석 과정을 설명하고, 세분화 결과의 적절성에 대해 검증하였다.

다섯째, 본 논문의 시각화가 실제 전문 분야에 있어 활용가치가 있는지를 살펴보기 위해, 전문가를 대상으로 인터뷰를 진행하였다. 그 결과 우리의 연구는 해당 전문 분야에서 요구하는 군집 분석에 대해 유의미한 집단 분석 결과를 제공해 줄 수 있음을 확인하였다.

본 논문의 2절에서는 다차원 데이터를 활용한 군집

분석 시각화 기법의 연구 동향을 서술한다. 그리고 3절에서는 본 논문의 시각화 기법 설명과 개발과정에 대해 언급하고 군집의 세분화 과정에 대해서도 언급한다. 4절에서는 시스템에서 제공하는 인터랙션 기능에 대해 서술한다. 그리고 5절에서는 케이스 스터디 및 전문가 집단의 검증과정을 설명한다. 마지막으로 6절에서는 연구에 대한 결론 및 향후 연구 계획에 대해 서술한다.

2. 다차원 데이터의 군집 분석 시각화에 대한 연구 동향

2절에서는 다차원 데이터를 시각화하는 여러 방법 중, 군집 분석에 주로 사용되는 Radvis 기법과 Parallel coordinate 기법의 연구 동향을 서술하고자 한다. 우선, 각 시각화 기법에 대한 이해를 돕고자 특징 및 구성원리를 서술하였다. 그리고 두 기법에 대한 기존 연구 사례를 소개하였고, 그 과정에서 데이터를 의미적으로 군집할 수 있는 기능이 어느 수준으로 구현되고 있는지 연구의 트렌드를 살펴보았다.

2.1 Parallel coordinate 시각화 및 Radvis 시각화의 특징

Parallel coordinate는 n차원 공간 안에 있는 데이터들의 집합을 효과적으로 보여주기 위해 고안된 시각화 방법이다[11]. 일반적으로 변수의 개수를 N개라고 가정했을 때, 그림 1의 왼쪽을 보면, Parallel coordinate를 이루고 있는 축(Axis)은 n개의 등간격 평행 라인으로 이루어진다[11,12]. 그리고 하나의 라인은 하나의 데이터가 보유한 각 변수들의 값에 따라 각 축을 이은 결과물이다. 그림 1의 오른쪽을 보면, Parallel coordinate는 각 변수의 대부분 선이 평행일 때 두 차원 사이에 유사한 관계라고 해석할 수 있다. 또한, 대부분의 선이 교차할 때는 상이한 관계라고 해석할 수 있다[11].

Radvis는 n차원의 지점을 평면으로 맵핑하기 위해 훅의 법칙(Hook's law)을 이용하여 데이터의 차원을 줄여, 다양한 변수가 표현된 평면 안에서 노드(Node)들의 분포를 볼 수 있는 시각화 방법이다[13,14]. 그림 2의 왼쪽을 보면, 원의 둘레에 위치한 S1~S5 지점은

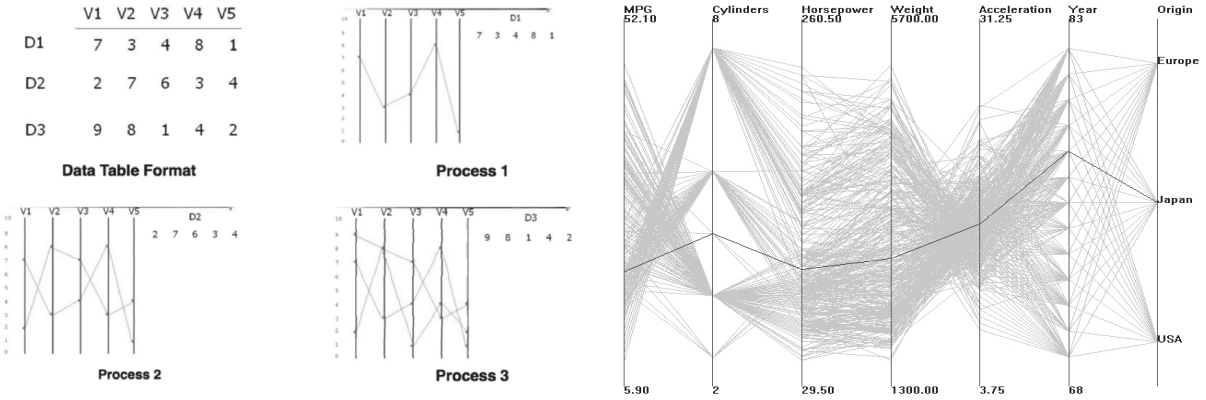

그림 1 (왼쪽) Parallel coordinate의 형성 원리; (오른쪽) Parallel coordinate 시각화의 예시

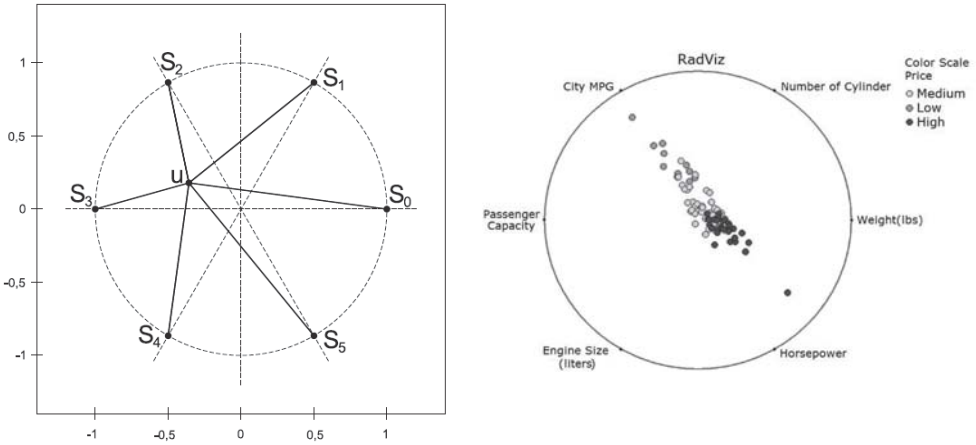

그림 2 Radvis 시각화 – (왼쪽) 노드 U가 변수 S1~S5에 의해 힘을 받아 배치되는 모습 (오른쪽) 노드들이 각 변수들에 의해 분포되는 예시

'Radvis의 차원 앵커(Dimension anchor)'라고 불리는 지점으로[15] 데이터의 각 변수들이 위치하게 된다. 그리고 U지점은 원 안에 표현된 노드이며, 노드의 위치는 각 변수에 대한 값이 장력으로 정의된다. 그러므로 변수 값이 클수록, 원에 배치되어 있는 각 변수의 차원 앵커의 위치에 가깝게 위치한다. 따라서 노드 U는 S2와 S3 변수에 대해 높은 값을 가진다는 점을 알 수 있다. Radvis 시각화는 노드 간의 밀집 정도에 따라 데이터 간 관계를 식별하는 데 도움을 줄 수 있다. 또한, 데이터의 특성을 패턴으로 제공한다는 점이 특징이다.

2.2 다차원 데이터 시각화의 군집 분석 연구 동향

Parallel coordinate 기법 및 Radvis 기법을 기반으로 다차원 데이터의 군집 분석을 시행한 사례들을 정리하면 아래와 같다.

우선 Parallel coordinate의 연구 동향을 살펴보면, 그림 3의 왼쪽은 Ying-Huey Fua의 연구로, 계층에 따른 군집 데이터 정보를 전달하기 위해 계층적 군집화(Hierarchical clustering)를 시행한 뒤 그 결과를 시각화로 나타내었다[16]. 이 연구는 계층 구조가 복잡해질 때 그래프 패턴이 복잡해지는 문제를 해결하기 위해 시각화를 군집 단위로 묶어서 표현했다는 점이 특징이다.

그리고 그림 3의 오른쪽은 Hong Zhou의 연구로, Parallel coordinate의 라인 그래프가 서로 복잡하게 얽혀 있을 때 나타나는 시각적 클러터를 최소화하기 위해 Curved Line 기법을 사용한 사례이다[9]. Curved Line은 Parallel coordinate 라인에 시각적으로 라인이 겹쳐 보이는 번들(Visual bundle)을 형성하여, 데이터를 군집 단위로 보여줄 수 있는 기법이다. 이러한 기능은 Parallel coordinate를 이루는 라인 그래프의 흐름을 군집 단위로 정리할 수 있다는 것이 특징이다. 그러나 이와 같은 연구들은 데이터의 군집 분포와 데이터의 자세한 내용을 함께 봐야하는 상황에서 적용하기는 어렵다는 단점이 있다.

다음으로 Radvis에 대한 연구 동향을 살펴보면, 그림 4의 왼쪽은 John Sharko의 연구로[15] 데이터의 군집 분포를 명확히 분류하기 위해 Radvis 시각화에서 원 안의 면적을 변수의 개수에 따라 나눈 뒤 그 면적을 기반으로 노드를 당기는 시각화 기법을 소개하고 있다. 이는 노드가 차원 앵커의 지점으로 당겨지는 기존의 사례와 달리,

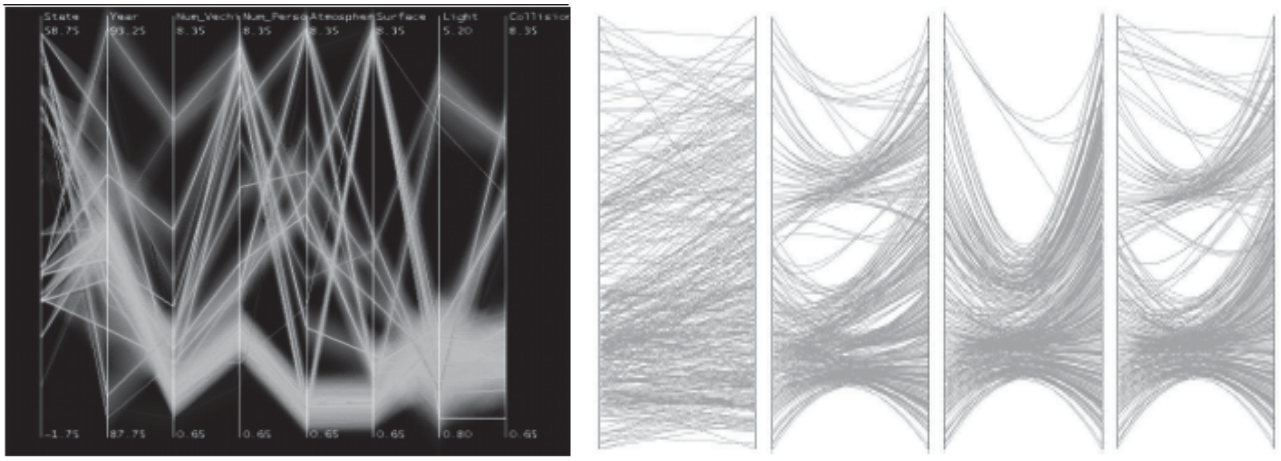

그림 3 (왼쪽) 계층적 군집화 Parallel coordinate (오른쪽) Curved Line 기법을 사용한 Parallel coordinate

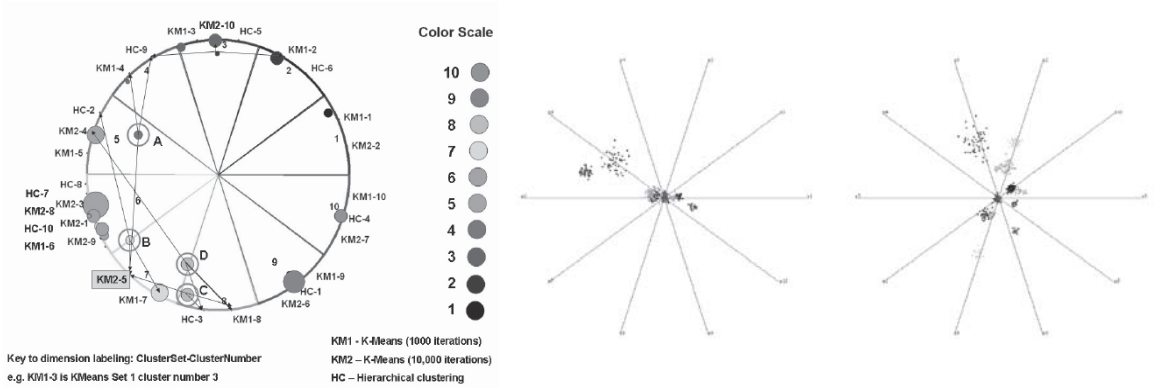

그림 4 (왼쪽) 면적을 기반으로 노드를 당기는 Radvis (오른쪽) 변수의 재배열을 이용하여 최적의 군집 수를 찾아내는 Radvis

면적에 따른 장력이 더 작용되므로 데이터의 군집이 더욱 잘 이루어지는 효과가 나타난다.

그리고 그림 4의 오른쪽은 Huynh Thi Thanh Binh의 연구로 Radvis의 변수 위치를 재배열하는 연구를 수행하였다[17]. 이 경우 변수를 재배열하는 과정에서 최적의 군집 개수를 찾아내고 각 군집에 대한 의미 해석을 부여한다는 점이 특징이다. 하지만 위의 연구들은 Radvis를 이루는 변수의 개수가 증가하여 여러 방향에서 노드를 당기게 되면, 대부분의 노드가 중앙에 뭉친다는 문제를 해결하지 못한다. 또한, 각 군집에 최적의 의미를 부여하기 위해서는 군집화를 여러 차례 실행해야 한다는 번거로움이 있다.

마지막으로 Radvis와 Parallel coordinate를 결합한 연구 사례를 살펴보면, 군집에 따라 Radvis의 노드 색을 부여하였다. 그리고 Parallel coordinate의 필터링 기능을 소개하였다.

그러나 Enrico 연구에서도 Radvis를 형성하는 변수의 개수가 많아지면, 노드들이 가운데로 뭉친다는 문제점이 있다. 또한, Parallel coordinate에서 군집 내의 세부 속성을 파악하기 어렵다.

본 논문에서는 군집을 분류하는 기능 외에도 군집 일부를 세분화하는 시스템 개발을 위해서 Radvis 기법과 Parallel coordinate의 기법을 조합하였으며, Radvis 기법을 적용할 때는 변수의 개수가 많아질 때 노드들이 가운데로 뭉치는 문제점을 해결하고자 3D radvis 형태의 기법을 제안하고자 한다.

3. 시각화 기법에 대한 소개 및 개발과정, 군집의 세분화 과정

3절에서는 시각화 기법에 대한 원리와 개발과정에 대해 언급하고, 군집을 세분화하는 원리에 대해 설명한다. 본 논문에서는 2D radvis의 시각적 기능을 개선한 3D radvis와 Parallel coordinate를 결합하였다. 시각화 기법에 대한 원리 및 개발과정, 시스템의 인터랙션 기능을 서술하면 아래와 같다.

3.1 3D radvis

본 논문에서는 3D radvis 기법을 이용하여 다차원 데이터를 군집화한 후, 원하는 군집을 별도로 선택하여 세분화하는 방법을 제시하였다. 기존의 2D radvis에서는 배치된 값들이 서로 달라도 중복되는 공간에 위치하는 데이터가 많아, 같은 공간 위치에 표시되더라도 실제 데이터는 서로 다른 값을 가지고 있는 경우가 있었다. 따라서 2D radvis의 한계점을 보완하기 위해 3D 개념을 시각화에 도입하였으며, 3차원 도형의 회전과 Z축의 깊이감을 활용하여 데이터의 변수들이 가지는 수치값을 고려하여 데이터를 분석하고자 하였다[19].

3D radvis는 Z축의 높이를 이용하여 각 변인의 범위값 크기를 표현한다는 특징이 있다. 각각의 노드는 폴리곤 안에서 힘을 받게 되는데, 힘을 받는 원리를 서술하면 다음과 같다.

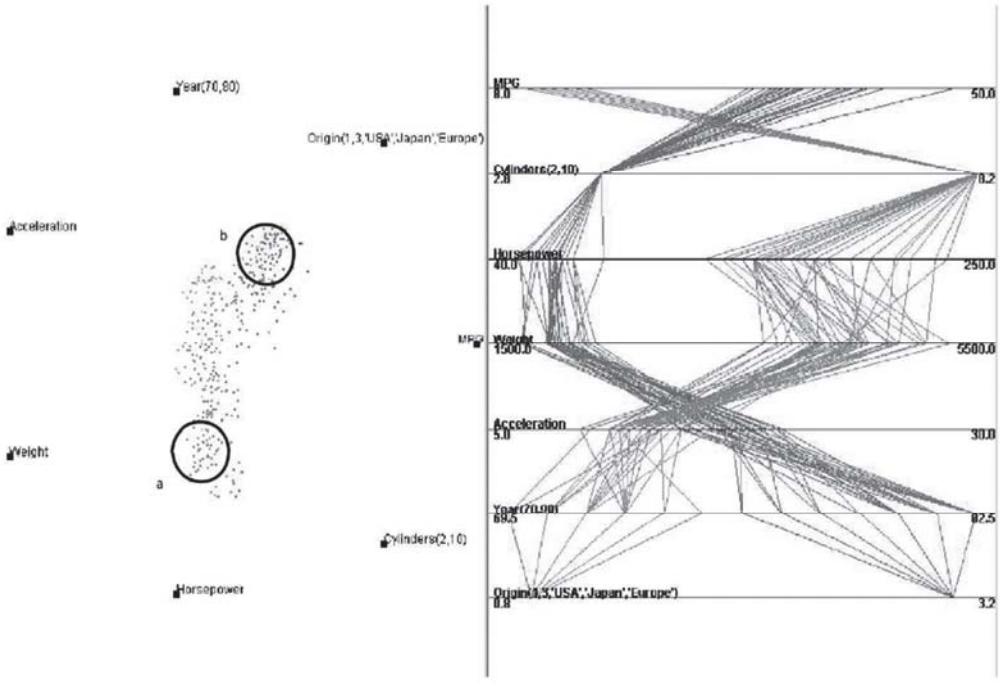

그림 5 Radvis 와 Parallel coordinate를 결합한 사례

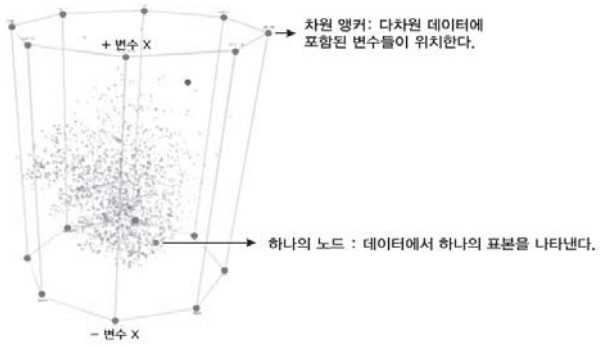

그림 6 3D radvis의 구성원리

노드가 받는 힘은 노드 하나가 가지는 각각의 변수값이 최솟값인지 최댓값인지, 또는 중간값인지 여부에 따라서 작용하는 힘이 각각 다르다. 따라서 변수값이 최댓값에 가까워지면 최댓값이 위치한 차원 앵커로 노드가 힘을 받게 된다. 반대로 변수값이 최솟값에 가까워지면 폴리곤 아랫면의 중점에서 노드가 힘을 받는다. 그리고 변수값이 최대와 최소 사이의 값일 때는 최솟값과 최댓값을 잇는 곡선(또는 직선)을 따라 선형적으로 힘이 작용한다. 이러한 힘의 작용은 노드를 분포시킬 때 중앙 지점에 중첩되지 않고 3D 폴리곤에 펼쳐지는 결과를 갖게 된다. 그림 6은 3D radvis 시각화에서 차원앵커가 위치하는 부분과 노드가 위치하는 부분을 보여준다.

3.2 Parallel coordinate

본 논문에서 활용된 Parallel coordinate 기법은 데이터의 특성을 보다 자세하게 분석하기 위해 다양한 조건을 부여할 수 있도록, 멀티 필터링 기능을 고안하였다. 멀티 필터링 기능은 한 변수에 대해서 2가지 이상의 조건을 적용할 수 있는 방법으로써 보고자 하는 데이터의 범위값을 다양하게 지정할 수 있다. 그림 7은 본 연구의 시스템에서 멀티 필터링이 작동되었을 때 선택되는 라인을 보여준다. 멀티 필터링 기능 이외에 Parallel coordinate를 구성하기 위한 개발과정은 Inselberg. A.의 연구를 참고하였다[11].

3.3 데이터의 군집과정 및 세분화과정

본 논문에서는 군집의 초기 중점을 설정하는 방법으로 Random과 Forgy 알고리즘을 적용하였으며, 사용자는 둘 중에 원하는 방법을 선택하여 군집화를 진행할 수 있다. Random 알고리즘은 3D radvis 내부에 임의로 군집의 중심점을 생성하기 때문에, 군집화를 진행할 때마다 다른 형태의 군집이 생성된다. 반면에 Forgy 알고리즘은 군집의 중점을 특정 노드를 선택하여 진행함으로써 같은 조건의 군집화에서는 같은 결과가 도출된다[20].

Random 또는 Forgy의 알고리즘을 선정하고 군집화를 시작하면, 나누고자 하는 군집의 개수에 따라 군집의 중심값이 형성된다. 그리고 그 중심값과 각 노드들 간의 의미거리(유클리디안 거리)를 계산하여 노드들은 가장 가까운 군집에 포함된다. 이후 군집의 중심은 군집에 속한 노드들의 중점으로 이동한다. 이러한 과정이 군집의 중심점이 변하지 않을 때까지 반복이 된다. 아래는 클러스터의 중심값 및 3D radvis의 물리적인 위치를 이용하여 도출하는 의미거리를 계산하는 식이다. 클러스터의 중심값 P를 구하는 식에서 P_i는 i 번째 노드 위치, n은 클러스터의 노드 개수를 뜻한다.

$$P = \frac{\sum_{i=1}^{n} P_i}{n} \quad (1)$$

그리고 3D radvis의 물리적인 위치를 이용하여 도출하는 의미거리 D를 구하는 식에서 P_x, P_y, P_z는 노드 P의 x, y, z 좌표값이고 Q_x, Q_y, Q_z는 각각 노드 Q의 x, y, z 좌표값이다.

$$D = \sqrt{(P_x - Q_x)^2 + (P_y - Q_y)^2 + (P_z - Q_z)^2} \quad (2)$$

군집화 후 각 군집의 상태를 확인하는 과정에서 물리적으로 넓은 공간에 분포하는 군집이 있거나, 유의한 해석을 이끌어내기 어려운 군집이 있을 수 있다. 이러한 경우에는 다른 특징을 기반으로 세분화를 더 진행할 수 있다는 점을 뜻한다. 따라서 본 논문에서는 넓은 공간에 분포하는 군집을 세분화하고 상대적으로 유의하지 않은 군집을 해석하기 위해 세분화 할 수 있는 기능을 제공하였다. 이에 대한 방법은 아래와 같다.

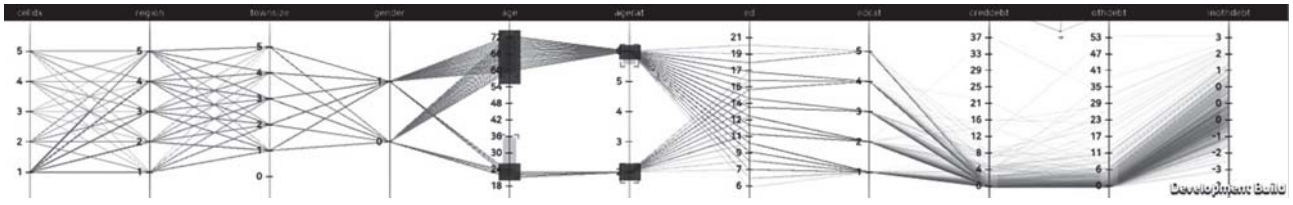

그림 7 Parallel coordinate의 멀티 필터링 기능: 그림에서 age 와 agecat 변수에 대해서 멀티 필터링이 적용됨

우선 세분화가 필요한 군집을 택한 후 세분화하면 이후의 군집화 과정은 세분화가 필요한 군집이 위치하는 물리적인 공간 내에서만 군집화를 진행한다. 본 논문이 제시하는 세분화 방법은 클러스터에서 상대적으로 물리적 공간이 넓은 군집만을 세분화한 후, 이전에 진행한 전체 군집 결과와 세분화 결과를 함께 시각화에 나타낸다. 따라서 전체적인 군집과 세분화된 군집을 비교 분석할 수 있다.

4. 시각화 인터랙션

4절에서는 본 논문에서 개발한 시각화 시스템이 제공하는 기능을 서술하고 각 기능이 대해 어떤 인터랙션을 제공하는지 소개하고자 한다. 시스템의 대표적인 기능을 정리하면 아래와 같다.

4.1 변수의 선택

사용자는 시각화 분석에 사용할 변수를 원하는 갯수만큼 직접 선택할 수 있다. 변수를 선택하기 위해서는 각 폴리곤의 왼쪽에 위치하는 변수 선택 메뉴를 참고할 수 있다. 변수 선택 시, 변수 이름 앞에 위치한 라디오버튼이 체크되고 폴리곤의 Z축 모서리와 Parallel coordinate의 축이 추가된다.

4.2 군집 세분화 인터랙션

사용자는 군집화 된 결과를 바탕으로 일부 군집을 세분화하여 쪼갤 수 있다. 우선 세분화할 군집을 선택한다. 그 다음 '군집 개수 설정창'에 세분화하려는 군집의 수를 입력한다. 마지막으로 왼쪽 폴리곤 하단의 'Focus Clustering(군집 세분화) 버튼'을 선택하면 군집이 세분화 된다. 세분화 할 수 있는 알고리즘의 종류로는 Random과 Forgy 중 하나를 선택할 수 있다. 사용자는 각각의 클러스터가 모두 해석에 있어 유의미한 결과가 나올 때까지 세분화 작업을 반복할 수 있다.

4.3 3D radvis의 노드 필터링 및 Parallel coordinate의 필터링 기능

사용자는 자신이 보고자 하는 표본을 필터링하기 위해 마우스 드래그 기능을 활용한다. 본 연구가 제시하는 3D radvis에서는 화면에 보이는 영역을 마우스로 드래그하여 필터링 할 수 있다. 그리고 Parallel coordinate에서는 각 변수가 표시된 축을 마우스로 드래그하면 사용자가 원하는 값 범위에 해당하는 데이터만을 볼 수 있다. 본 논문의 3절에서 이미 언급했듯이, Parallel coordinate는 변수 내의 멀티 필터링이 가능하다.

4.4 모드 전환

사용자는 군집 분석을 하기 전의 모습과 군집 분석

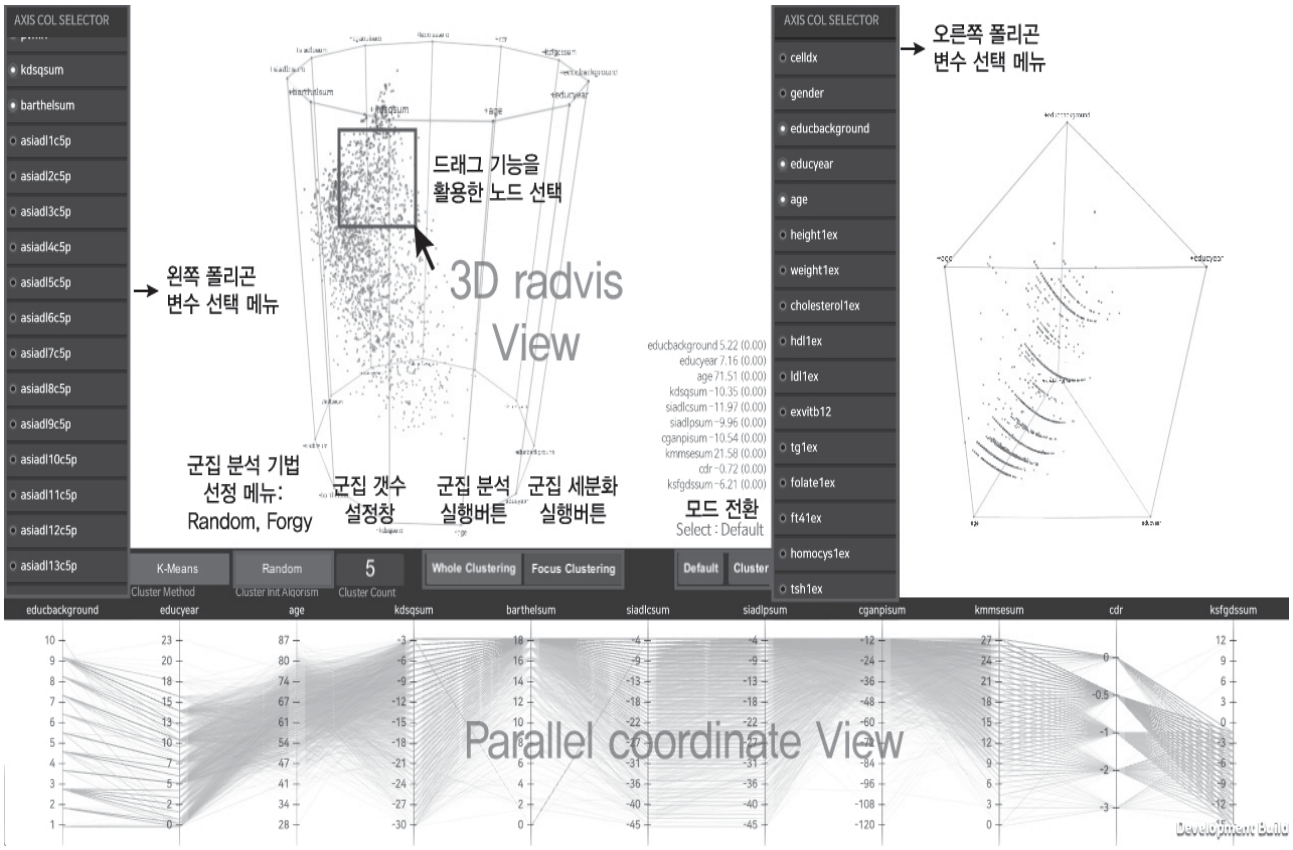

그림 8 시각화 시스템 인터페이스

을 하고 난 뒤의 모습을 모드 전환 기능을 이용하여 비교가 가능하다. 왼쪽의 폴리곤이 위치한 부분에서 오른편의 하단을 보면 'Default'와 'Cluster'버튼이 있는데. Default를 선택하면 군집 분석 전의 색상 정보로 노드가 표시되고, Cluster를 선택하면 군집 분석 및 세분화 분석이 된 이후의 색상 정보로 노드가 표시된다.

5. 케이스 스터디 및 토론

5절에서는 본 논문에서 제시한 시각화 시스템이 다차원 데이터의 군집 분석 및 세분화 과정을 잘 보여주는지 검증하기 위해 케이스 스터디를 제안하였다. 또한, 실제 전문 분야에서 본 논문의 시각화가 활용 가치가 있는지 알아보기 위해 전문가 인터뷰를 진행한 뒤, 그 결과를 정리하였다.

5.1 케이스 스터디: CREDOS 코호트 데이터을 활용한 분석

여기에서는 CREDOS(Clinical Research Center for Dementia of South Korea)의 치매환자 진단 코호트의 데이터를 활용하였다[21]. CREDOS 코호트 데이터는 약 14,000여명 치매환자의 21,000여건에 달하는 진단 기록을 포함한다. 그리고 데이터의 변수로는 치매의 최종진단에 관여하는 신경심리 검사, 행동심리 검사 및 신체검사 정보와 개인정보를 포함하고 있다. 변수에 대한 정보는 아래의 표 1과 같다.

본 논문에서는 CREDOS 코호트 데이터 중 각 심리검사(KDSQ, CGA-NPI, Barthel-ADL, S-IADL, K-MMSE, SNSB, GDSSF-K, CDR)의 총점 변수들을 사용하여 군집의 세분화 분석을 진행하였다. 또한, CREDOS 코호트 데이터의 표본은 치매의 질병단계에 따라서 5단계로 분류되며 명칭은 각각 SMI(Subjective Memory Impairment), MCI(Mild Cognitive Impairment), VCI (Vascular Cognitive Impairment), SVD(Subcortical Vascular Dementia), AD(Alzheimer's Disease)로 나뉜다. 그리고 SMI에서 AD로 갈수록 만성도가 중증임을 뜻한다. 시각화 분석을 위해 변수를 선정한 뒤에 나타나는 3D radvis 시각화 및 Parallel coordinate 시각화의 모습은 그림 9와 같다.

표 1 CREDOS Cohort Data의 변수 정보

변수 유형		변수의 설명
환자정보	개인정보	나이, 성별, 교육연한, 학력
	신체검사	콜레스테롤(일반, HDL, LDL), Apoe 유전자, 비타민 B12
심리측정	심리검사	KDSQ(한국 치매 스크리닝 질문 모음), CGA-NPI(간병인 관리를 위한 신경심리학 인벤토리), Barthel-ADL(일상 활동을 위한 Barthel index), S-IADL(서울형 일상 생활기능 검사), K-MMSE(한국형 미니 정신 상태 검사), SNSB(서울 신경심리 검사-Dementia Version), GDSSF-K(단축형 노인우울 척도), CDR(치매 임상평가 척도)

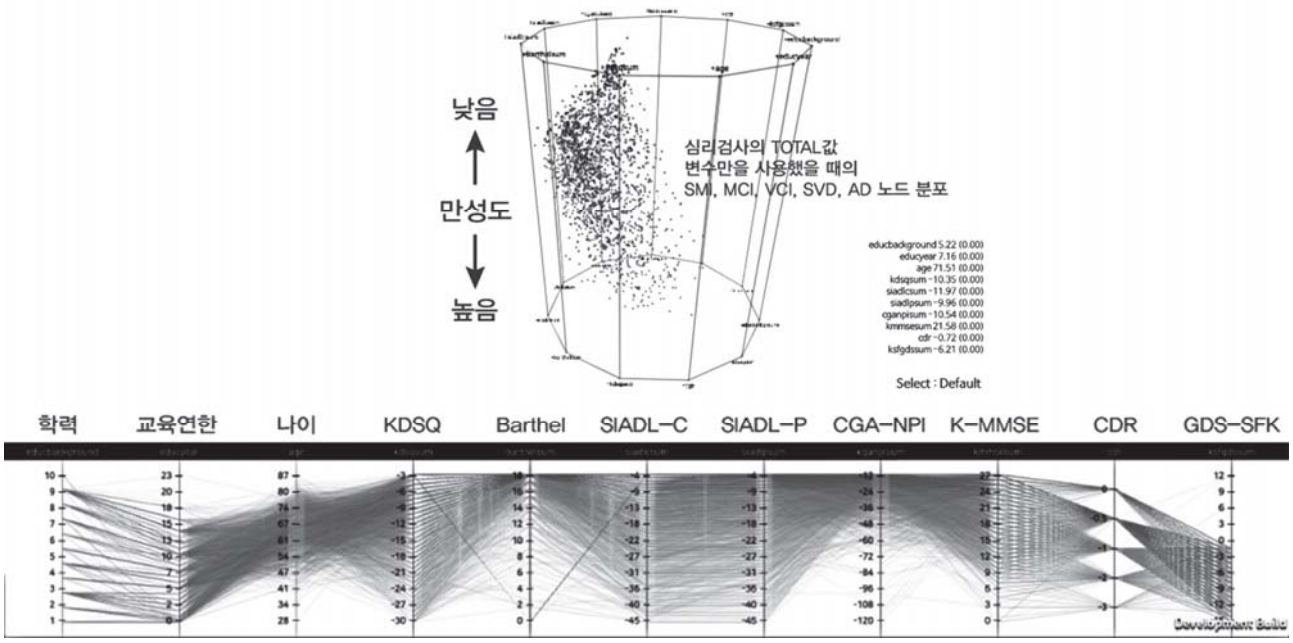

그림 9 CREDOS 코호트 데이터의 시각화

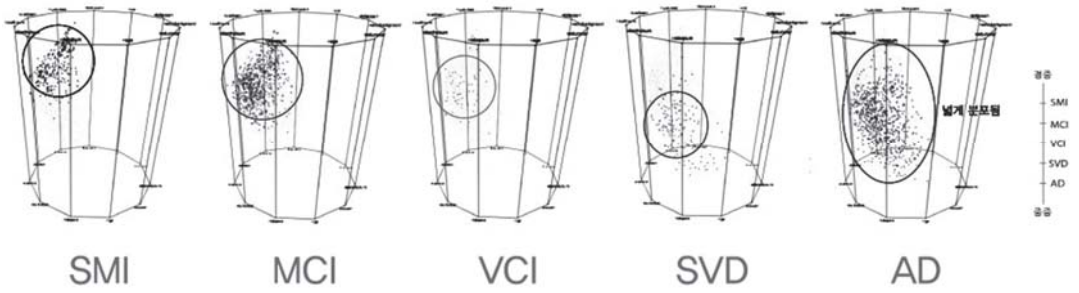

그림 10 치매환자들의 질병단계에 따른 3D radvis의 노드 분포 비교

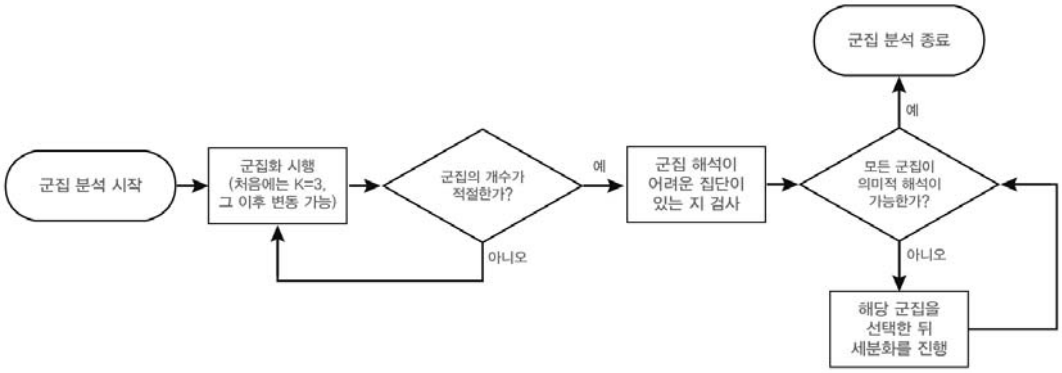

그림 11 케이스 스터디를 위한 군집화 및 세분화 과정 순서도

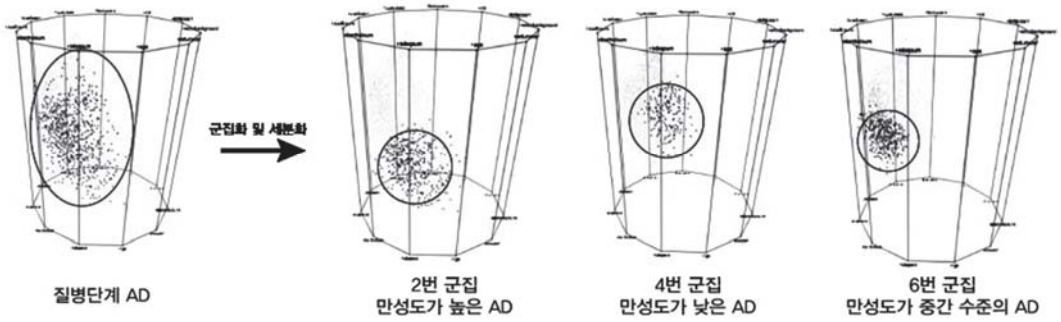

그림 12 AD 집단이 2번 군집, 4번 군집, 6번 군집으로 세분화된 결과

그림 9에서 3D radvis의 Z축의 높이가 낮을수록 만성도가 높고, 높을수록 낮은 만성도를 의미한다. 만성도가 높은 사람들의 경우, 심리검사 응답값이 낮기 때문이다. 따라서 3D radvis는 만성도에 따라 다른 높이를 가지고 있으며, 이를 통해 환자들의 상태를 알 수 있다. 또한, 3D radvis에서 하나의 노드는 환자 한 명이 응답한 치매검사 정보를 담고 있다. 각 노드 색상은 치매 진행 단계에 따라, 진단결과가 SMI일 경우 SMI면 파랑(양호), MCI는 하늘(다소 양호), VCI는 연두(중간), SVD는 주황(다소 위험), AD는 빨강(매우 위험)으로 표시했다. Parallel coordinate는 사용자 선택한 변수들의 검사 수치를 라인 그래프로 보여준다.

그림 10은 각 질병단계에 따른 3D radvis의 노드 분포를 나타낸다. 그림 10을 보면 만성도가 낮은 SMI와 MCI, 중간 수준인 VCI, 다소 높은 수준인 SVD에 대한 분포는 군집을 이루고 있지만, 높은 수준인 AD에 대해서는 노드들이 넓게 분포한다는 것을 알 수 있다. 이는 AD 단계의 집단이 낮은 만성도와 높은 만성도 집단 모두를 포함하고 있음을 의미한다. 따라서 본 연구에서는 AD 집단을 세분화하는 작업을 수행하였다. 그림 11은 케이스 스터디를 위해 시행된 군집화 및 세분화의 순서도를 보여주며, 그림 12는 AD 집단이 세분화된 모습을 보여준다.

그림 12를 보면, AD 단계 집단은 3개의 군집으로 세분화되었다. 각 군집을 확인한 결과, 2번 군집은 3D radvis 시각화에서 높이가 낮으므로 높은 만성도의 AD 집단이 군집되었다고 볼 수 있다. 그리고 4번 군집은 2, 6번 군집에 비해 높은 곳에 위치하므로, 만성도가 낮은 AD 집단이

표 2 전체 군집의 세분화 결과 나타나는 심리검사 점수의 평균값 및 해석

변수 (범위값)	2번 군집의 평균값 및 해석		4번 군집의 평균값 및 해석		6번 군집의 평균값 및 해석	
KDSQ (3~30)	20.78	간병인 없이 생활이 어려움	10.47	혼자서 어느 정도 일상생활	4.33	혼자 일상생활 가능
Barthel (0~18)	17.02	보행, 씻기, 목욕 행동에 어려움	19.49	가끔 실수해도 씻기, 목욕 가능	19.49	가끔 실수해도 씻기, 목욕 가능
SIADL-C (4~45)	28.55	요리, 약 챙겨먹기 등에서 이상이 발생	11.98	정상적인 일상생활이 가능	3.20	정상적인 일상생활이 가능
K-MMSE (0~27)	14.80	시간, 장소인지 및 기억 회상에 이상	18.54	기억 회상에 가끔 이상을 보임	27.11	정상
CDR (0~3)	1.40	심한 기억장애, 집안생활 수행이 불가	0.62	기억장애가 전혀 없거나 경미한 건망증	0.42	기억장애가 전혀 없거나 경미한 건망증

군집되었다고 볼 수 있다. 마지막으로 6번 군집은 폴리곤의 중간 높이에 위치하고 있어, 만성도가 중간인 AD 집단이 군집되었다고 볼 수 있다.

위의 결과를 토대로 AD의 성격을 가지는 3개의 군집이 적절하게 세분화 되었는지를 확인하기 위해 각 군집이 보유한 변수들의 평균값을 비교하였다. 그리고 진단척도에 따라 검사 결과값을 해석하여 정리하였다. 평균값을 정리하여 해석한 표는 표 2와 같다.

표 2에 제시한 변수 중, 치매 진단에 대표적으로 사용하는 변수인 KDSQ와 S-IADL를 이용한 평균 점수 해석은 다음과 같다. 우선 치매 진단 척도인 KDSQ의 경우, 각 군집에서 5점 이상의 차이를 보이고 있었다. 이는 곧 KDSQ의 점수가 가장 높은 2번 군집은 간병인 없이 식사와 보행이 어려운 것을 의미하며, 4, 6번 군집은 부분적으로는 혼자 일상생활이 가능함을 의미한다. 이외에도 SIADL-C의 경우, 2번 군집에 속하는 환자는 전화, 쇼핑, 요리 등의 일상생활에서 이상이 발생한다고 볼 수 있다. 하지만 상대적으로 점수가 낮은 4, 6번 군집의 경우, 2번 군집에 비해 정상적인 일상생활이 가능하다. 따라서 검사의 평균값 비교를 통해, 군집의 세분화가 적절하게 이루어졌다는 사실을 확인할 수 있었다.

5.2 전문가의 인터뷰 및 토론

본 연구에서는 3D radvis와 Parallel coordinate 기반의 데이터 군집 결과가 사용자에게 어떠한 활용 가치를 주는지에 대해 조사했다. 이를 위해 정신과 의사 및 임상심리 전문가 2명을 대상으로 인터뷰를 진행하였다. 인터뷰 진행 방법은 참가자들이 직접 시각화 프로그램을 사용하면서 주어진 질문들을 푸는 방식이었다. 질문들은 군집 세분화의 적절성을 묻는 문항으로 구성되었다. 더불어 3D radvis 시각화의 군집화 결과를 해석할 때, 참가자들의 임상 경험을 토대로 결과를 해석하도록 하였다. 이 과정에서 나오는 시각화 기능에 대한 피드백도 함께 받았다.

인터뷰 결과, 두 전문가는 CREDOS 코호트 데이터에서 치매환자의 일정한 분포 패턴이 발견되는 점을 긍정적으로 평가하였다. 특히 질병 단계가 혼재되어 있는 곳에서 3D radvis를 통해 군집을 세부적으로 나눌 수 있는 기능에 대해서 긍정적인 반응을 보였다. 덧붙여 K-means 알고리즘을 기반으로 분류되는 군집에 대해서는, 치매환자 세분화를 위해 질병 단계를 구체적으로 분류하면 좋을 것 같다는 반응도 있었다.

CREDOS 코호트 데이터가 3D radvis에서 5개로 클러스터링이 되는 경우에 대해서는, 질병의 이름이 다를 뿐 치매를 크게 경증, 중증으로 분류할 수 있다는 점에서 군집의 세분화 결과가 충족됨을 확인할 수 있었다. 그리고 같은 질병 단계를 가지고 있는 집단이 세분화 과정을 통해 나뉘는 결과는 향후 치매환자를 진단하는데 좋은 정보가 될 것임을 확인하였다. 또한, 실제 의학 분야에서 데이터의 분포와 검사값을 다변인으로 분석하는 연구가 많지 않기 때문에, 치매 연구자들에게 새로운 연구 아이디어를 제시해 줄 수 있을 것으로 기대하였다.

6. 결론 및 향후 과제

본 논문은 다차원 데이터를 대상으로 군집을 세분화하기 위한 시각화 시스템을 소개하였다. 우선, 이번 연구를 위해 관련 연구 사례의 동향을 파악하였다. 또한, 최적의 다차원 군집 분석 시각화가 되기 위해서는 3D radvis 기법 및 Parallel coordinate 기법을 서로 결합하는 것이 적절하다는 점을 확인하였다. 그리고 두 시각화 기법에 대한 구성 원리와 시각적 기능에 대해 설명한 뒤 군집을 세분화하는 방법에 대해 설명하였다. 또한, CREDOS 코호트 데이터를 활용하여 케이스

스터디를 수행하였다. 마지막으로, CREDOS 코호트 데이터의 군집 세분화 결과를 확인하고, 전문가 인터뷰를 통해 의학적 해석이 가능하다는 점을 검증하였다.

본 논문의 시스템은 특정 집단을 선택한 뒤 그것을 세분화 할 수 있는 기능이 제공된다. 이는 곧 기존의 군집 분석 방법에서 일부 집단의 표본 수가 0인 오류를 최소화하고, 군집의 일부를 해석할 수 없었던 상황을 해결할 수 있음을 의미한다.

또한, 2D radvis 시각화에서 나타나는 노드의 중첩으로 인한 문제는 3D radvis에서 Z축을 기준으로 노드가 가지는 변수들의 값에 따라 분포를 펼치는 방법을 통해 해결될 수 있을 것이라 예상한다. 그리고 Parallel coordinate 기법은 3D radvis로 구성된 노드 군집을 세분화하는 과정에서 라인 그래프의 패턴을 각각 나누어준다. 이는 각 변수들의 값 차이를 보는데 도움이 될 것으로 예상한다. 마지막으로 본 연구의 시스템은 변수를 자유롭게 선택하고 위치시킬 수 있어, 사용자가 원하는 맞춤 분석 도구를 제공해줄 수 있을 것이라 기대한다.

본 논문은 향후 K-means 군집화 방법 중 Random, Forgy 이외의 기법을 사용하여 보다 다양한 세분화 기능을 제공할 계획이다. 또한, 사용자 인터뷰 과정에서 변수선택 및 시각화의 조작 기능의 전반적인 User Interface 개선이 필요하다는 의견이 있었기 때문에 각 전문 분야에서 원하는 기능이 무엇인지 조사하여 기능을 추가할 계획이다. 마지막으로 CREDOS 코호트 데이터 이외의 데이터를 적용하여 본 논문의 시스템이 다양한 분야에서 활용될 수 있다는 것을 검증할 계획이다.

참고문헌

[1] Etemadpour, R., Motta, R., de Souza Paiva, J. G., Minghim, R., de Oliveira, M. C. F., and Linsen, L., “Perception-based evaluation of projection methods for multidimensional data visualization", IEEE transactions on visualization and computer graphics, Vol. 21, No.1, pp. 81-94, 2015.

[2] Ankerst, M., Berchtold, S., and Keim, D. A., “Similarity clustering of dimensions for an enhanced visualization of multidimensional data", IEEE Information Visualization, pp. 52-60, 1998.

[3] Gorban, A. N., Kégl, B., Wunsch, D. C., and Zinovyev, A. Y., Principal manifolds for data visualization and dimension reduction, pp. 96-130, Springer, Berlin-Heidelberg, 2008.

[4] Linsen, L., Van Long, T., Rosenthal, P., Rosswog, S., “Surface extraction from multi-field particle volume data using multi-dimensional cluster visualization", IEEE Transactions on Visualization and Computer Graphics, Vol.14, No.6, pp.1483-1490, 2008.

[5] Wu, J., Liu, H., Xiong, H., Cao, J., Chen, J., “K-means-based consensus clustering: A unified view", IEEE Transactions on Knowledge and Data Engineering, Vol.27, No.1, pp.155-169, 2015.

[6] Reza, M. S. and Ruhi, S, “Study of Multivariate Data Clustering Based on K-Means and Independent Component Analysis", American Journal of Theoretical and Applied Statistics, Vol.4, No.5, pp. 317-321, 2015.

[7] Ding, C. and He, X., “Cluster merging and splitting in hierarchical clustering algorithms", In Data Mining IEEE International Conference, pp. 139-146, 2002.

[8] Cao, N., Gotz, D., Sun, J., and Qu, H., “Dicon: Interactive visual analysis of multidimensional clusters", IEEE transactions on visualization and computer graphics, Vol.17, No.12, pp.2581-2590, 2011.

[9] Zhou, H., Yuan, X., Qu, H., Cui, W., and Chen, B., “Visual clustering in parallel coordinates", In Computer Graphics Forum, Vol. 27, No. 3, pp.1047-1054. 2008.

[10] Wang, C., Xiao, Z., Liu, Y., Xu, Y., Zhou, A., & Zhang, K., “SentiView: Sentiment analysis and visualization for internet popular topics", IEEE transactions on human-machine systems, Vol.43, No.6, pp.620-630, 2013.

[11] Inselberg, A., “The plane with parallel coordinates', The visual computer, Vol.1, No.2, pp. 69-91, 1985.

[12] Inselberg, A., and Dimsdale, B., “Parallel coordinates", In Human-Machine Interactive Systems, pp. 199-233, 1991.

[13] Di Caro, L., Frias-Martinez, V., and Frias-Martinez, E., Analyzing the role of dimension arrangement for data visualization in radviz, pp. 125-132, Springer, Berlin-Heidelberg, 2010.

[14] Rubio-Sánchez, M., Raya, L., Diaz, F., and Sanchez, A., “A comparative study between RadViz and Star Coordinates", IEEE transactions on visualization and computer graphics, Vol.22, No.1, pp.619-628, 2016.

[15] Sharko, J., Grinstein, G., and Marx, K. A., “Vectorized radviz and its application to multiple cluster datasets". IEEE transactions on Visualization and Computer Graphics, Vol.14, No.6, pp.1444-1427, 2008.

[16] Fua, Y. H., Ward, M. O., and Rundensteiner, E. A., “Hierarchical parallel coordinates for exploration of large datasets", In Proceedings of the conference on Visualization'99: celebrating ten years, pp. 43-50, 1999.
[17] Binh, H. T. T., Van Long, T., Hoai, N. X., Anh, N. D., and Truong, P. M., “Reordering dimensions for Radial Visualization of multidimensional data－A Genetic Algorithms approach", In 2014 IEEE Congress on Evolutionary Computation (CEC), pp. 951-958, 2014.
[18] Bertini, E., Dell'Aquila, L., and Santucci, G., “Springview: Cooperation of radviz and parallel coordinates for view optimization and clutter reduction", In Coordinated and Multiple Views in Exploratory Visualization (CMV'05), pp. 22-29, 2005.
[19] 성정환, 이대영, 김형구., “2D 와 3D Graphic 기반으로 구성된 GUI 의 효율성의 차이", 한국콘텐츠학회논문지, 제7권, 제7호, pp.87-95, 2007.
[20] Yadav, A. K., Tomar, D., and Agarwal, S., “Clustering of lung cancer data using Foggy K-means", In Recent Trends in Information Technology (ICRTIT), pp. 13-18, 2013.
[21] Choi, S. H. and et al., "Driving in Patients with Dementia: A CREDOS (Clinical Research Center for Dementia of South Korea) Study", Dementia and Neurocognitive Disorders, Vol.13, No.4, pp. 83-88, 2014.

약 력

하 효 지

2013 아주대학교 미디어학부 졸업 (학사)
2014～현재 아주대학교 일반대학원 라이프미디어협동과정 석사과 통합과정 재학
관심분야 : 다차원 데이터 분석, 데이터 시각화, 시각화 분석, UX디자인
Email : hjha0508@ajou.ac.kr

한 현 우

2014 아주대학교 미디어학부 졸업 (학사)
2014～현재 아주대학교 일반대학원 라이프미디어협동과정 석사과 통합과정 재학
관심분야 : 데이터 시각화, 컴퓨터 그래픽스, 웹 공학, 소프트웨어 공학
Email : ainatsumi@ajou.ac.kr

배 성 윤

2015 아주대학교 미디어학부 졸업 (학사)
2015～현재 아주대학교 일반대학원 라이프미디어협동과정 석사과정
관심분야 : 데이터 시각화, 기계학습
Email : roah@ajou.ac.kr

이 지 혜

2015 아주대학교 심리학과 졸업 (학사)
2015～현재 아주대학교 일반대학원 라이프미디어협동과정 석사과정
관심분야: 정보 시각화, 사용성 실험 설계 및 평가, 데이터 분석, UX디자인
Email: alice0428@ajou.ac.kr

손 상 준

2002 연세대학교 의과대학 졸업
2013 연세대학교 대학원 의학박사
2014 명지병원 임상교수
2014～현재 아주대학교 의과대학 정신건강의학과 진료 및 연구조교수
관심분야: 노인정신건강 (불안, 불면, 화병, 치매)
Email: sjsonpsy@ajou.ac.kr

홍 창 형

2005 연세대 노화과학연구소 연구교수
2006 광주시 정신보건센터장
2007 세브란스병원 정신과 전임의
2008 연세대 노화과학협동과정 박사
2015～현재 아주대학교 의과대학 정신건강의학과 교수
관심분야: 치매, 인지장애, 노인성우울증
Email: antiagint@ajou.ac.kr

신 현 정

2005 서울대학교 공과대학 [산업공학/데이터마이닝] 공학박사
2005 Max-Planck-Institute for Biological Cybernetics (독일) 연구원
2006 Friedrich-Mierscher-Laboratory, Max-Planck-Institute(독일) 수석연구원
2006 서울대학교 의과대학 연구교수
2006～현재 아주대학교 공과대학 산업공학과 교수
관심분야: 기계학습, 데이터마이닝, 바이오메디컬 인포매틱스
Email: shin@ajou.ac.kr

이 경 원

1996 국민대학교 시각디자인학과 (학사)
2002 Computer Graphics & Interactive Media, Pratt Institute (석사)
2003～현재 아주대학교 정보통신대학 미디어학과 교수
2009～2010 Visualization and Interface Design Innovation Lab, University of California at Davis 방문교수
관심분야: 정보시각화, 인간-컴퓨터 상호 작용, 미디어아트
Email: kwlee@ajou.ac.kr