Using Projection-Based Clustering to Find Distance- and Density-Based Clusters in High-Dimensional Data 요약

2022.03.23. 예지혜

Abstract
고차원의 데이터는 거리와 밀도 구조 (DDS) 모두에 의해 클러스터가 형성되기 때문에 대부분의 알고리즘은 클러스터링을 정확히 잡아내지 못한다. 이를 해결하기 위해 projection-based clustering (PBC) 방법이 소개되었다. 투영과 클러스터링을 동시에 진행하기 때문에 지형도 (topographic map)를 통해 클러스터 경향과 클러스터의 개수를 모두 잘 찾아낸다.

1. Introduction

많은 데이터마이닝 방법론들이 정보의 유사성에 집중하는데, 이는 data-driven과 need-driven으로 나눌 수 있다. 여기서는 데이터가 가지고 있는 정보를 발견해내고 클러스터 내 유사성과 클러스터간 비유사성에 집중할 것이므로 data-driven에 집중한다.

클러스터링은 거리 기반(compact structures) 방법과 밀도 기반(connected structures) 방법으로 나눌 수 있고, 데이터에 따라 알맞은 방법을 적용해야 한다.

한 가지 접근 방법은 차원 축소 방법을 이용해 고차원 데이터를 저차원 데이터로 투영시키는 것이다. 이렇게 투영시키면 실제로는 가까운 데이터가 멀게 투영되는 문제가 발생하기도 하지만 여전히 기본 방법으로 쓰인다. 이러한 문제를 해결하기 위해 저자는 지형도를 통한 클러스터링 시각화 방법을 제안한다.

2. Common Projection Methods as an Approach for Dimensionality Reduction

일반적으로 차원 축소는 Manifold learning과 Projection으로 나눌 수 있다. Manifold learning은 고차원 데이터의 거리가 보존되는 subspace를 찾기 때문에 보통 2차원 이상의 결과가 나타나며, 시각화에 좋은 방법은 아니다.
Projection은 데이터를 2차원으로 투영하여, 거리 또는 밀도 기반의 데이터 구조를 찾는 시각화에 좋은 방법이다. 선형과 비선형 방법으로 나눌 수 있고, 선형 방법은 데이터 좌표를 정규 직교 회전시키는 방법으로 PCA, ICA, PP 등이 있다.

2.1 Combining Dimensionality Reduction with Clustering

일부 변수가 중요하지 않아 보이면 보통 PCA나 FA를 적용하는데, 여기에 k-means 방법을 통한 클러스터링을 진행하는 방법을 Tandem Clustering이라 한다.
이 경우, 클러스터링이 결국 비슷한 정보들의 그룹으로 정의되는 경우가 있어 클러스터링과 차원 축소를 동시에 진행하는 방법인 RKM (reduced k-means)이 제안되었다. subspace에 직교하는 방향으로 분산이 커서, RKM이 잘 작동하지 않을 때는 FKM (factorial k-means)을 사용한다.

subspace clustering은 충분차원축소를 기반으로 하는 방법으로, output 변수에 대한 최대한 많은 정보를 갖는 선형 subspace를 찾는 것을 목표로 한다. 따라서 보통 2차원보다 높다.

이와는 다르게, Projection-Based Clustering (PBC)는 "관련있는" 이웃한 정보들만을 유지하도록 2차원으로 데이터를 투영시킨다. 또한, 이러한 투영을 위해 비선형 투영을 사용한다. 비선형 투영 방법으로는 CCA, t-SNE, NeRV, Pswarm 등이 있다.

2.2 Projection-Based Clustering

[PBC의 3단계]

1) 비선형 투영 방법으로 고차원 데이터를 투영시킨다.
2) 투영된 데이터 포인트에 simplified emergent self-organizing map (ESOM) 알고리즘을 사용하여 generalized U-matrix를 적용한다.
3) 지형도(topographic map)를 결과물로 얻으며, 클러스터링은 이 지형도에 표현된다.

* best-matching units (BMUs) : 투영된 데이터 포인트 p가 discrete lattice 위로 변형된 것
* $M = \{m_1, m_2, ..., m_n\}$: 2D lattice 위의 뉴런의 위치
* D : high-dimensional distance
* $W = \{w(m_i) = w_i | i = 1, 2, ..., n\}$: the corresponding set of weights or prototypes of neurons

⇒ simplified SOM 알고리즘은 각 데이터에 대해 적합한 bmu를 찾는다.

$$\mathrm{bmu}(l) = \mathrm{argmin}_{m_i \in M} \{D(l, w_i)\}, i \in \{1, \ldots, n.\}$$

* generalized U-matrix

$N(j)$를 m_j와 인접한 8개의 이웃이라 하면 generalized U-matrix의 높이는 다음과 같이 정의된다. 즉, 인접한 프로토타입과의 평균 거리를 의미한다.

$$u(j) = \frac{1}{n} \sum_{i \in N(j)} D\left(w(m_i), w\left(m_j\right)\right), n = |N(j)|$$

* Voronoi region of bmu L_i

$$L_i = \left\{x | x \in \mathbb{R}^n \wedge D(x, \mathrm{bmu}_i) < D\left(x, \mathrm{bmu}_j\right) \forall i \neq j\right\}$$

* Delaunay graph

bmu_i and bmu_j connected ⇔

$$\exists x \in \mathbb{R}^n | D(x, \mathrm{bmu}_i) = D\left(x, \mathrm{bmu}_j\right) \wedge D(x_{\mathrm{bmu}_i}) < D(x, \mathrm{bmu}_k) \forall k \neq i, j$$

Two types of clusters
- Compact structures : 그룹간 vs. 그룹내 거리 비교 기반 (유클리드 그래프)
- Connected structures : 데이터의 이웃 및 밀도 기반 (다양한 그래프)

color scale of topographic map = distance
: blue (sea level) < green and brown (low hills) < white (high mountains with snow)
색깔이 같으면 같은 클러스터로 간주