UMAP&PaCMAP

UMAP (Uniform Manifold Approximation and Projection)

1. 서론

- UMAP은 고차원 데이터를 저차원으로 매핑하는 기법입니다.
- 데이터의 기하학적 구조를 보존하면서 차원을 축소합니다.
- 데이터 시각화, 클러스터링, 이상치 탐지 등 다양한 응용 분야에서 사용됩니다.

2. UMAP의 원리

- **기하학적 구조와 Topological Structure**
  - UMAP은 데이터의 기하학적 구조를 fuzzy simplicial sets의 형태로 모델링합니다.
  - 이를 통해 고차원 데이터의 복잡한 구조를 저차원에서도 유지하려고 합니다.
- **Manifold Learning**
  - UMAP은 manifold learning의 한 종류로, 고차원 데이터가 놓인 manifold를 학습하여 저차원으로 매핑하는 기법입니다.
  - 예를 들어, 3D 공간에 놓인 2D 평면을 학습하여 2D 공간에 표현하는 것과 유사합니다.
    $f : X \subset \mathbb{R}^D \rightarrow Y \subset \mathbb{R}^d$

PaCMAP (Pairwise Controlled Manifold Approximation Projection)

1. 서론

- PaCMAP은 고차원 데이터를 저차원으로 시각화하기 위한 기법입니다.
- UMAP, t-SNE와 같은 기법들의 장점을 결합하여 더 빠르고, 높은 품질의 결과를 목표로 합니다.

2. PaCMAP의 원리

- **Pairwise Distance Control**
  - PaCMAP은 데이터 포인트 간의 거리를 적절하게 제어하여, 국지적 및 전역적 구조를 동시에 보존하려고 합니다.
    $L(y) = \sum_{i,j} w_{ij} \| y_i - y_j \|^2$
    w_{ij}는 데이터 포인트 i와 j 간의 연결 가중치, y_i와 y_j는 저차원 공간에서의 데이터 포인트를 나타냅니다.
- **Stress Minimization**
  - 고차원 공간의 거리 행렬과 저차원 공간의 거리 행렬 사이의 차이를 최소화하는 것을 목표로 합니다.

3. 알고리즘의 구성 요소

- **거리 계산 및 가중치 부여**
  - 고차원 데이터 포인트 간의 거리를 계산하고, 이를 바탕으로 연결 가중치를 부여합니다.

위 수식에서 X는 고차원 데이터, Y는 저차원 데이터, D는 원본 데이터의 차원 수, d는 목표 차원 수를 나타냅니다.

3. UMAP 알고리즘의 구성 요소

- **Fuzzy Simplicial Sets**
  - 각 데이터 포인트의 국지적 이웃 관계를 계산하고, 이를 통해 고차원 공간에서의 연결 관계를 추정합니다.
  - 이웃 선택 및 가중치 부여 방식이 이 부분의 핵심입니다.
- **Cost Function**
  - 차원 축소 후의 데이터 \(Y\)가 원래 고차원 데이터 \(X\)의 구조를 잘 보존하도록 설계됩니다.
  - 이를 위한 목적 함수를 최소화하는 형태로 구성됩니다.
    $\min_Y \sum_{i,j} w_{ij} \| y_i - y_j \|^2$

  w_{ij}는 데이터 포인트 i와 j간의 연결 가중치, y_i와 y_j는 저차원 공간에서의 데이터 포인트를 나타냅니다.
- **Optimization Technique**
  - 일반적으로 SGD (Stochastic Gradient Descent) 등을 사용하여 cost function을 최소화합니다.
  - 이 과정에서 여러 하이퍼파라미터를 튜닝할 필요가 있습니다.

4. UMAP vs. t-SNE & PCA

- **성능 비교**
  - UMAP은 t-SNE보다 계산이 빠르며, PCA보다는 데이터의 복잡한 구조를 더 잘 유지합니다.
- **목적 함수 및 최적화**
  - 목적 함수는 대부분 거리를 보존하는 항과 고차원 및 저차원 거리 사이의 스트레스(stress)를 최소화하는 항으로 구성됩니다.
  - 이 목적 함수를 최소화하기 위해 다양한 최적화 기법을 사용합니다.

4. PaCMAP vs. UMAP & t-SNE

- **성능 비교**
  - PaCMAP은 UMAP, t-SNE와 비교하여 더 빠르고, 국지적 및 전역적 구조를 잘 보존하는 결과를 생성합니다.
- **시간 복잡도 비교**
  - PaCMAP의 계산 복잡도는 일반적으로 $O(N \log N)$으로, 대규모 데이터셋에도 효율적으로 작동합니다.
- **결과 시각화 비교**
  - PaCMAP은 고밀도 및 저밀도 영역을 모두 잘 구분하여 표현합니다.

5. PaCMAP의 장점 및 단점

- **장점**
  - 빠른 계산 속도
  - 국지적 및 전역적 구조의 잘 보존
- **단점**
  - 하이퍼파라미터 튜닝이 필요할 수 있음
  - 결과의 해석이 어려울 수 있음

- **시간 복잡도 비교**
  - UMAP: $O(N \log N)$
  - t-SNE: $O(N^2)$
  - PCA: $O(Nd^2)$, N는 데이터 포인트 수, d는 목표 차원 수입니다.
- **결과 시각화 비교**
  - UMAP은 클러스터 구조를 잘 유지하면서도 더욱 고밀도 데이터에 강한 결과를 보입니다.

5. UMAP의 장점 및 단점

- **장점**
  - 빠른 계산 속도
  - 클러스터 구조 유지
- **단점**
  - 하이퍼파라미터 튜닝이 필요할 수 있음
  - 결과의 해석이 어려울 수 있음