Review of Dimension Reduction Methods 요약

2022.01.20

Abstract

각 차원축소 방법론들의 특징, 장단점, 적용 분야, 데이터 타입 리뷰

선형/비선형

선형 방법론 – PCA, LDA, SVD, LSA, SPP, ICA, PP

비선형 방법론 – KPCA, MDS, Isomap, LLE, SOM, SVQ, t-SNE, UMAP

지도/비지도

지도 – LDA, LVQ (위에 없는 것들로는, supervised variants of PCA, LPP, KPCA, and MDS)

비지도 – 위의 방법론 중 나머지 다

준지도 – semi-supervised variants of PP, t-SNE, LDA

1. Introduction

많은 분야에서 빅데이터가 급격히 생성되면서 이를 다루기 위한 방법들이 연구되었다. 특히, 분류 정확도, 패턴 인식, 시각화에서는 고차원 데이터가 차원의 저주 등의 문제를 발생시켰고, 이를 위해 차원을 낮춰주는 방법론인 Dimension reduction이 연구되어왔다.

차원축소는 feature selection, feature extraction 두 가지 방식으로 실행될 수 있는데, feature selection은 주요한 몇 가지 변수들만 선택하는 것이고, feature extraction은 많은 변수들의 선형 결합으로 구성된 몇 가지 독립 변수들을 찾는 것이다.

각 방법론마다 적용할 수 있는 데이터 타입은 제한적일 수 있으며, 이 논문에서는 텍스트, 이미지, 오디오, 비디오, 시계열, 구조화된 데이터들을 고려한다.

3.1. Linear Dimension Reduction Techniques

3.1.1. Principal Component Analysis (PCA)

- 가장 오래됐고 널리 쓰이는 방법으로, 선형 비지도 차원축소 방법론
- 가장 많은 변동을 설명하는 독립된 선형 결합들을 찾아낸다.

- 분포 가정이 필요 없어 다양한 데이터 타입을 다룰 수 있다.

- 적용 분야: 머신러닝, 이미지/음성 processing, 컴퓨터 비전, 텍스트 마이닝, 시각화, biometrics, robotic sensor data, 얼굴 인식

- 한계점:

The PCA transformation, despite its widespread use, relies on second-order statistics. The principal components can be highly statistically dependent though uncorrelated and this can lead to PCA failing to find the most compact description of the data. PCA geometrically models the data as a hyperplane embedded in a space that is ambient and requires a larger dimensional representation than would be found by a non-linear technique if the data components have non-linear dependencies. This has prompted the development of non-linear alternatives to PCA [29]. PCA methods also fail to account for outliers which are common in realistic training sets because they employ least squares estimation techniques [30].

- extension:

RPCA – robust해서 다양한 이미지 데이터에 적용, 이상치에 영향 잘 안 받음
ERPCA – suitable for data of different sizes, outliers ok
LPCA – image, speech data
ROBCA – more accurate and faster than the traditional PCA
…

- 적용 분야: 텍스트, 이미지, 오디오, 비디오, 시계열, 구조화된 데이터

3.1.2. Singular Value Decomposition (SVD)

- 행렬방정식 계산과 data reduction 형태의 문제에 사용, 행렬 분해

- 그 외) digital image processing, taxonomic classification of biological sequences, pattern recognition, gene expression data, signal processing, Natural Language Processing (NLP), bio-informatics, and text summarization

- 단점:

계산 비용이 높지만 랜덤 샘플링이 적용되면 개선 가능
데이터의 비선형성과 이상치에 민감

- 적용 분야: 텍스트, 이미지, 오디오, 비디오, 시계열, 구조화된 데이터

3.1.3. Latent Semantic Analysis (LSA)

- unsupervised LDR mapping

- text data 특화. 검색 시스템 개선에 사용(연관된 문서들을 같은 군집으로)

- PCA, SVD의 computation에서 발전된 버전. 벡터 기반.

- HD 코퍼스를 낮은 차원으로 축소.

- cognitive functions of LSA는 단어의 의미를 학습, 이해한다. 또한 단어-단어, 어절-어절 간의 measure, relationship을 생성. 동의어 문제도 다룰 수 있다.

- 한계점:
  같은 형태의 단어에 대해 다양한 의미를 학습하여, 전혀 다른 의미의 평균치로 해석된다는 문제가 있다. 하지만 보통 우세하게 쓰이는 뜻이 있기 때문에 괜찮다.
  LSI는 정렬되지 않은 단어 모음으로 텍스트가 표현된다는 BOW 단점이 있지만, multi-gram 사전으로 이를 다룰 수 있다.

- 적용 분야: 텍스트

3.1.4. Locality Preserving Projections (LPP)

- projective maps that solve problems that are variational in nature and preserve optimally the neighborhood structure of the data set.

- 분산을 최대화하도록 데이터를 투영하는 고전적인 선형 방법이기 때문에 PCA의 보이기도 한다.

- LPP의 투영행렬이 서로 직교하지 않아 reconstruction이 어렵고, 이를 해결하기 위해 OLPP가 제안되었다. 또한 OLPP의 높은 계산 비용을 해결하고자 FOLPP가 제안되었다.

- 적용 분야: 텍스트, 이미지, 오디오, 비디오, 시계열, 구조화된 데이터

3.1.5. Independent Component Analysis (ICA)

- statistical signal processing technique used for the exploration of multi-channel data

- independent source들의 선형 결합을 모델링. PCA에 independence를 추가한 확장 버전.

- 독립 조건을 통해 PCA보다 더 의미있는 component를 찾을 수 있다.

- 과적합 위험이 줄어들고, data reconstruction 가능

- 한계점: 대부분의 ICA 알고리즘이 Gradient descent 방법에 기반하기 때문에 global minimum을 찾았다고 확신하기 어렵고, 따라서 해석에 주의하여야 한다.

- 적용 분야: 텍스트, 이미지, 오디오/신호, 비디오, 시계열, 구조화된 데이터

3.1.6. Linear Discriminant Analysis (LDA)

- 분류 문제에서 많이 쓰이고 있는 지도 기법
- 그룹간 분산은 크게 하고 그룹내 분산은 작게 하여, 변수들의 선형 결합을 linear classifier로 사용한다.
- 한계점: small sample problem => *해결책 3가지*
- 적용 분야: 텍스트, 이미지, 오디오, 비디오, 시계열, 구조화된 데이터

3.1.7. Project Pursuit (PP)

- 데이터 탐색에 주로 쓰이는 비지도 기법
- 낮은 차원의 선형 투영을 찾고 흥미로운 패턴을 찾는다.
- interestingness가 PP index
- 장점: 다른 패턴 적합에도 유연하게 적용할 수 있으며, 샘플에 없는 점도 매핑할 수 있어 투영 공간에서 새로운 예시를 보여줄 수 있다.
- 지도 학습에도 적용이 되었다. (projection pursuit indices)
- computational difficulty가 단점이며, gradient 방법, Newton-Raphson method, GA, SA 등의 최적화 방법이 사용된다.
- 적용 분야: 텍스트, 이미지 데이터

3.2. Non-Linear Dimensionality Reduction Techniques

3.2.1. Kernel Principal Component Analysis (KPCA)

- PCA에 커널을 적용한 비선형 방법으로, PCA는 공분산행렬의 고유벡터를 계산한다면, KPCA는 커널 매트릭스의 고유벡터를 계산한다.
- 커널 매트릭스 K는 새로운 feature의 내적으로 계산되며, 가우시안, polynomial, hyperbolic tangent, radial 등이 많이 쓰인다.

- computational difficulty가 단점이라 EM 알고리즘을 적용하는 방법도 제안되었다.

- 적용 분야: 이미지, 오디오, 비디오, 시계열 데이터

3.2.2. Multidimensional Scaling (MDS)

- 데이터 포인트 쌍 간의 similarity나 dissimilarity를 보존하는 것이 목적

- 주어진 데이터를 더 쉽게 이해하고 해석할 수 있도록 표현하는 것이 주요 목적. 보통 2차원 또는 3차원 맵에 similarity나 dissimilarity에 따른 거리로 표현된다.

- 요인 분석과 비슷하지만 MDS는 선형성과 정규성이라는 엄격한 가정을 필요로 하지 않는다는 장점이 있다.

- MDS의 가정은 차원의 수가 점의 수보다 하나 적어야 한다..?

- MDS는 데이터 탐색이나 다변량 분석에 많이 쓰인다.

- 단점: 이상치에 민감하다.

- 적용 분야: 텍스트, 이미지, 오디오, 비디오, 시계열, 구조화된 데이터

3.2.3. ISOMAP

- 내재적인 데이터 구조를 찾는 것이 목적이다.

- 고차원에서의 직선 거리는 유지하는 저차원 공간을 찾는 것?

- 장점: 계산 효율성, 점근적 수렴, global 최적화 등 PCA와 MDS의 주요 특성을 결합함.

- MDS와 비슷하지만 manifold 거리를 유지하려 한다는 특성이 있다.

- 실시간 비디오 이상치 탐지(도시 도로 교통 상황에 적용), 스피치 요약, 크랙 탐지, 얼굴 인식 등에 사용

- 지도 학습 버전도 제안되었음

- 단점: 계산 비용이 높고, manifold 샘플링이 잘 안 되면 성능 떨어짐

- 적용 분야: 텍스트, 이미지, 오디오, 비디오 데이터

3.2.4. Locally Linear Embedding (LLE)

- 데이터의 로컬 특성만 보존하는 것이 목적이다.

- 얼굴 인식과 원격 감지 등에 적용. 최근에는 MRI, 알츠하이머의 해마 형태 분석 등 의료 분야

- 단점: 노이즈에 민감, 새로운 데이터를 잘 처리 못함

- 지도, 준지도 버전도 제안되었다.

- 적용 분야: 이미지, 오디오, 비디오 데이터

3.2.5. Self-Organizing Map (SOM)

- 비지도 인지 학습. 인공신경망을 위해 고안된 방법

- *create effectively spatially organized internal representations of many input signals of features and their abstractions*

- 문장의 의미적 관계를 식별할 수 있다.

- 음성 인식, 데이터 마이닝 탐색,

- 복잡한 문제를 쉽게 해석할 수 있는 데이터 매핑으로 단순화 가능

- 크고 복잡한 데이터셋을 군집화 가능

- 한계점:
  의미 있는 군집화를 위해 충분한 데이터 필요
  벡터의 가중치는 성공적인 데이터 그룹핑과 식별가능한 입력에 기반해야 한다.
  그룹화가 unique할 경우 완벽한 매핑을 하기 어렵다.

- 적용 분야: 텍스트, 이미지, 오디오, 비디오, 시계열, 구조화된 데이터

3.2.6. Learning Vector Quantization (LVQ)

- SOM과 비슷한 경쟁학습, 지도 학습이며 인공신경망

- 그룹의 영역을 나타내는 프로토타입 학습이 목표. 그룹 영역은 프로토타입간 하이퍼플레인으로 정의된다.

- SOM처럼 모든 출력 노드가 경쟁하여 입력 패턴과의 유사성에 따라 우승 노드가 결정되지만, SOM과 달리 LVQ는 우승 노드만 업데이트하여 출력 공간이 topologically ordered 되진 않는다.

- 클러스터링을 위해 비지도 학습으로도 사용될 수 있고, 이해하기 직관적이고 간단하다는 장점이 있다.

- 한계점: 느린 수렴 속도, 불안정한 작동

- 적용 분야: 텍스트, 이미지, 오디오, 비디오, 시계열, 구조화된 데이터

3.2.7. t-Stochastic Neighbor Embedding (t-SNE)

- SNE의 변형. SNE는 데이터 쌍의 거리로부터 확률 분포를 구성하는 것이 목표이다.

- 단일 세포 분석에 일반적으로 사용

- 단점: 느린 계산 속도, 너무 큰 데이터는 잘 못 다룸, 대규모 정보 손실.

- 적용 분야: 텍스트, 이미지, 오디오, 비디오, 시계열, 구조화된 데이터

3.2.8. Uniform Manifold Approximation and Projection (UMAP)

- Riemannian geometry와 algebraic topology 이론을 기반으로 구성된 비지도 학습

- issue of uniform data distribution on manifolds를 다룸

- *이론적 개념*

- 적용 분야: 이미지, 오디오, 비디오, 구조화된 데이터

3.3. Overview of Sufficient Dimension Reduction 충분 차원 축소

- SIR, SAVE, PHD, MAVE, SCR, IR, DR 등

- 한계점:
  선형 조건이나 등분산 조건이라는 증명하기 어려운 가정을 필요로 한다.
  central subspace를 철저히 추정하지 못하는 경우가 많다.

4. Conclusions

- 각 방법론의 목표

  PCA: the preservation of variance

SVD: optimal dimension reduction

LSI/LVQ: classification accuracy

LPP, KPCA, MDS, LLE, Isomap: the extraction of manifolds

SOM: prediction accuracy

t-SNE and UMAP: preservation of neighborhood