REPORT

: Spectral clustering 과 KPCA 비교 분석

과목명 지식발견

담당교수 차광호

학과 산업정보시스템전공

학번 17101937

이름 김도현

제출일 20.06.07

컴퓨터와 인터넷이 발전함에 따라 정보의 양은 방대해지고, 이에 따라 데이터의 특성을 분류하거나 어떻게 구성되었는지를 알아내는 알고리즘이 요구되었다. 데이터를 이용하는 방법 중 하나로, 라벨링이 되어 있지 않은 데이터를 이용하는 비지도학습(Unsupervised) 기법이 있다. 라벨링이 되어 있지 않은 데이터는 입력 값에 대한 목표치가 주어지지 않았음을 의미하고, 해당 기법으로 스펙트럴 클러스터링 (Spectral clustering)과 커널 주성분 분석(Kernel Principal Component Analysis)가 있다.

클러스터링 알고리즘은 고차원에 존재하는 데이터들을 k개의 클러스터로 나눠주는 것을 목표로 한다. 데이터가 밀집되어 있거나 선형 부분공간에 존재하고 있는 경우에 대해서는 k-means 클러스터링과 같은 간단한 방법을 이용하면 된다. 그러나 우리가 다루는 대부분의 데이터는 비선형 적이면서 유클리드 (Euclidean) 거리로 측정하기 어렵기 때문에, 서로 연결되어 있는 대상을 같은 그룹으로 포함시키는 스펙트럴 클러스터링 기법을 사용하게 되었다.

스펙트럴 클러스터링은 관련도가 높은 대상을 하나의 그룹으로, 관련도가 떨어지는 대상은 다른 그룹 으로 포함시킨다. 스펙트럴 클러스터링 알고리즘의 가장 큰 장점은 non-convex 경계를 가진 데이터 집합에서 클러스터를 찾아낼 수 있다는 점으로, 기존의 문제점들을 해결할 수 있게 되었다.

또한, 방대한 데이터에 대해 조금 더 분류하기 쉽거나 처리하기 쉽도록 차원을 낮춰주는 방법으로 PCA(Principal Component Analysis)가 존재한다. PCA알고리즘은 차원축소를 위해서 선형 변환 기법을 사용하기 때문에 선형으로 분리를 못하는 경우에 대해서는 적용하기 어려워지는데, 커널 주성분 분석은 이런 문제점에 대해서 PCA에 커널 트릭을 적용함으로써 해결해준다.

• Spectral Clustering

1. 데이터 간의 거리를 기반으로 유사도 행렬(K)을 계산한다.
   * 두 점 간의 거리가 멀수록 유사도는 떨어지고, 거리가 짧을수록 유사도는 높아진다.

K=

두 점 간의 거리

1. 대각행렬(D)를 통해, 정규화 된(normalized) 유사도 행렬을 계산한다.

D=

=

(D는 대각 행렬이므로, 대각성분을 제외하고 모두 0이다.)

1. 행렬 의 고유 벡터() 를 작은 것부터 구한다.
2. k개의 고유 벡터를 열(column)로 사용하는 행렬(V)를 생성한다.

V=

1. m개의 새로운 데이터 포인트를 K-means 알고리즘을 사용하여 K개의 클러스터로 분할한다.

new data point = row(i) of matrix V

•KPCA

1. kernel 을 선택한다.
2. NN 그람행렬(Gram matrix) K를 생성한다.
3. 아래의 식을 통해 그람행렬 K를 로 만들어준다.

모든 성분이 1인 NN행렬

1. 내적 행렬(dot product matrix)의 N개의 고유 벡터를 계산한다.

(고유 벡터의 각 요소는 데이터 포인트를 변형된 높은 차원으로 사영 시킨 것과 같다.)

스펙트럴 클러스터링과 커널 주성분 분석은 임베딩 고유의 특성을 변화한 것이기 때문에 의 값에 따라 결과가 매우 달라진다. 또한 두 알고리즘 모두 고유 벡터를 구하는 계산을 하게 되는데, 어떤 고유벡터가 가장 차별된 결과를 가져오는지 알 수 없으므로 최대한 많은 고유 벡터를 사용해야 한다.

스펙트럴 클러스터링의 가장 큰 장점은 희소 데이터(sparse data)에 대해서도 합리적인 속도로 계산 할 수 있다는 점이다. 그러나 모든 데이터셋에 대해 이루어지는 과정이므로 몇 개의 그룹으로 나눠지는 가에 따라 결과가 달라질 수 있다는 문제점이 있다.

커널 주성분 분석의 가장 큰 장점은 데이터 셋에 포함되지 않은 점에 대해서도 사영 시킬 수 있다는 것이다. 새로운 샘플에 대해서 행해질 수 있는 이런 방법을 ‘학습자의 일반화 능력(generalization ability of the learner)’ 이라 한다. 또한 커널 함수는 무한 차수의 비선형 함수를 통해 비선형 데이터를 매핑 시키는데, 이로 인해 조금 더 선형 데이터로 이루어질 수 있게 된다. 스펙트럴 클러스터링과 커널 주성분 분석으로 인해 비선형 적인 데이터에 대해서도 학습이 이루어질 수 있게 되었고, 방대해지는 학습 데이터에 대해서 다양한 처리 과정이 이루어질 수 있게 되었다.

참고자료

* 지식발견[kpca.pdf]
* 지식발견[SpectralClustering.pdf]
* 논문 [*SPECTRAL CLUSTERING AND KERNEL PRINCIPAL COMPONENT ANALYSIS ARE PURSUING GOOD PROJECTIONS* by VIKAS CHANDRAKANT RAYKAR] <http://www.umiacs.umd.edu/labs/cvl/pirl/vikas/projects/spectral_clustering.pdf>
* 논문 [*A Tutorial on Spectral Clustering* by Ulrike von Luxburg Max Planck Institute for Biological Cybernetics Spemannstr] <https://arxiv.org/pdf/0711.0189.pdf>
* 논문 [커널주성분분석을 이용한 회분 공정모니터링 기술 by 유창규] <https://www.cheric.org/files/research/ip/p200303/p200303-701.pdf>
* 논문 [*Spectral clustering: summary and recent research issues* by Jeong Sanghun, Bae Suhyeon, Kim Choongrak] <https://www.koreascience.or.kr/article/JAKO202014151761079.page>