5월20일~5월26일

∙ 주제로 잡은 PCA Dimensionality Reduction using K-means clustering algorithm for special data sets에서 저번주까지는 PCA에 대하여 공부하였고 이번주부터는 K-means 클러스터링 알고리즘에 대하여 공부하였습니다.

-클러스터링은 전통적으로 데이터 분석을 위한 감독되지 않은 방법으로 간주된다. 그러나 경우에 따라 문제 도메인에 대한 정보를 데이터 인스턴스 자체와 함께 사용할 수 있다.

**1. K-means Clustering**

K-means 클러스터링은 데이터 세트를 k 개의 그룹으로 자동 분할하는 데 일반적으로 사용되는 방법이다. 초기 클러스터 센터를 k 개 선택하고 다음과 같이 반복하여 수정한다 :

1. 각 인스턴스는 가장 가까운 클러스터 센터에 할당됩니다.

2. 각 클러스터 센터 는 구성 인스턴스의 평균으로 업데이트됩니다.

인스턴스를 클러스터에 할당 할 때 더 이상 변경 사항이 없는 경우 알고리즘이 수렴된다.

이 작업에서는 데이터 세트에서 무작위로 선택한 인스턴스를 사용하여 클러스터를 초기화한다. 우리가 사용한 데이터 세트는 숫자 기능만으로 구성된다. 또는 상징적 인 특징들. 숫자 형 feature의 경우 Euclidean distance metric을 사용합니다. 상징적인 특징을 위해 Hamming distance를 계산한다.

마지막으로 k를 선택하는 방법이 있다. k의 최적 값이 이미 알려진 데이터 세트 (즉, 모든 UCI 데이터 세트)에 대해 이용한다; GPS 데이터의 차선을 찾는 현실 세계 문제에 대해 우리는 wrapper search를 사용하여 k의 최상의 가치를 찾는다.

**2. Constrained K-means Clustering**

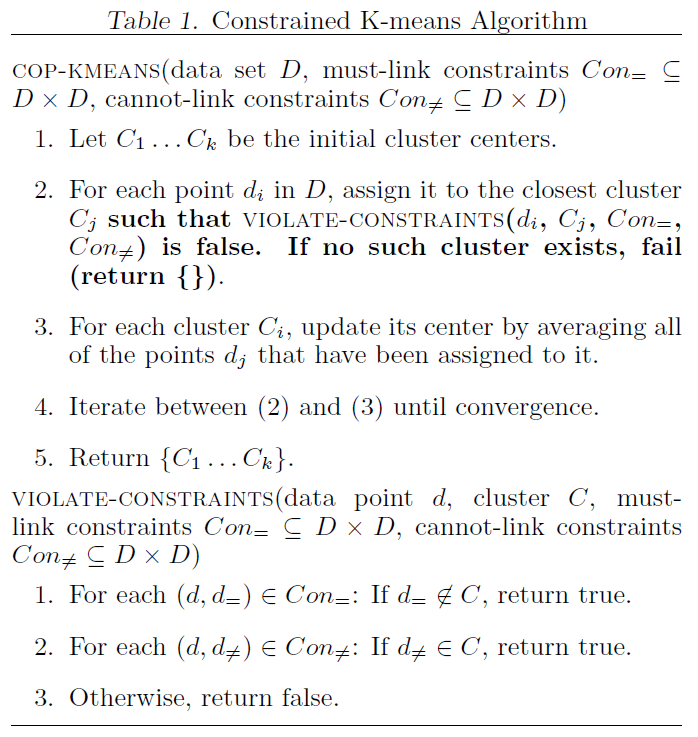
2.1 The Constraints

Partitioning algorithms과 관련하여 인스턴스 수준 제약 조건은 어떤 인스턴스를 함께 그룹화할지 또는 그룹화하지 않을지에 대한 사전 지식을 표현하는 데 유용한 방법이다. 결과적으로, 우리는 pairwise 제약 조건의 두 가지 유형을 고려한다.

∙ 필수 링크 제한 조건은 두 개의 인스턴스가 동일한 클러스터에 있어야 한다는 것을 지정한다.

∙ 제약 조건을 연결할 수 없으므로 두 인스턴스를 동일한 클러스터에 배치하면 안된다.

must-link 제약은 인스턴스에 대한 transitive binary 관계를 정의한다. 결과적으로 (두 종류의) 제약 집합을 사용할 때 제약 조건에 대해 과도적인 종결을 취한다. 전체 집합의 파생 제약 조건이 클러스터링 알고리즘에 표시된다. 일반적으로 제약 조건은 부분적으로 분류 된 데이터 또는 도메인 또는 데이터 세트에 대한 배경 지식에서 파생 될 수 있다.



3.2 The Constrained Algorithm

Table 1에는 수정 된 k-means 알고리즘 (COP-KMEANS)과 굵은 글씨체가 포함 되어있다. 이 알고리즘은 데이터 세트 (D), 필수 링크 제약 세트 () 및 비 결합 제약 세트 ()를 수용한다. 여기 D에서 모든 지정된 제약 조건을 만족하는 인스턴스의 파티션을 반환한다.

주요 변경 사항은 클러스터 할당을 업데이트 할 때 특정 제약 조건을 위반하지 않았음을 보장한다. 여기서는 각 점 를 가장 가까운 클러스터 에 할당하려고 시도한다. 이것은 제약 조건을 위반하지 않는 한 성공한다. 와 동일한 클러스터에 지정되어야 하지만 다른 클러스터에 이미 있는 다른 지점 가 있거나 와 그룹화 할 수 없지만 이미 에있는 다른 지점 가있는 경우 는 불가능하다. 여기서는 합법적으로 를 호스트 할 수 있는 클러스터를 찾을 때까지 클러스터의 정렬 된 목록을 계속한다. 제약 조건은 결코 깨지지 않는다. 그리고 에 대해 유효한 클러스터를 찾을 수 없으면 빈 파티션 ({})이 반환된다.