[빅데이터 HW2]

# Non-exhaustive, Overlapping k-means

2014314856 소프트웨어학과 오예린

**Abstract**

k-means와 같은 기존의 클러스터링 알고리즘은 disjoint하고 exhaustive한(모든 single 데이터 포인트가 정확히 하나의 클러스터에 속해야 한다는 성질) 클러스터를 만드는 결과를 내었다. 그러나 실제 데이터 셋에서는 클러스터가 겹칠 수 있으며 어떤 클러스터에도 속하지 않는 outliers가 존재할 수 있다. 때문에 이 문제를 해결하기 위한(overlapping clustering) 알고리즘( ex) fuzzy k-means ) 들이 제기되었으나 대부분의 알고리즘들은 overlap문제와 outlier detection을 별개의 문제로 다룰 뿐 통합적인 측면에서 보지 않았다. 하지만 이 논문에서는 overlap과 non-exhaustiveness의 문제를 통합적인 측면에서 바라보며 간단하고 직관적인 objective function을 제시하고자 한다. NEO-K-Means 라 불리는 알고리즘은 이해하기 쉬운 parameters(overlap과 non-exhaustiveness 의 정도 나타냄)를 함께 표시하여 기존의 k-means objective를 재구성한다.

## 1. Introduction

이 논문에서는 groups은 존재하지만 깔끔한 분리가 부족하고 data가 outlier를 포함하고 있는 real-world data의 관점에서 clustering 문제를 다룬다. 선행 연구들과의 가장 중요한 차이는 non-exhaustiveness와 overlap 문제를 하나의 framework에서 처리한다는 것이다. 그래프 클러스터링의 맥락에서, 기존의 normalized cut 기반의 그래프 클러스터링 objective를 non-exhaustive, overlapping한 환경으로 확장했고, 이렇게 확장된 그래프 클러스터링 objective는 특정한 weight와 kernel하에서 수학적으로 weighted kernel NEO-K-Means objective와 동일하다. 실험적인 결과도 NEO-K-Means 알고리즘이 다른 최신의 overlapping community detection 문제보다 월등함을 보여준다.

## 2. Non-exhaustive, Overlapping k-means

k-means objective function을 확장하기 위해, Assignment matrix인 U를 적용했다. $x_i$가 클러스터 j에 속할 때 $u_{ij}$ = 1 이고 그렇지 않은 경우는 0이다. non-exhaustive, overlapping clustering에서 U의 row에는 1이 여러 개 있을 수 있다. (하나의 data point가 여러 개의 클러스터에 속할 수 있으므로) 또한 row의 모든 요소가 0일 수도 있다.(outlier) 정사각 행렬인 $U^TU$에서 대각성분은 클러스터의 사이즈와 같으며 trace($U^TU$)는 클러스터 사이즈의 합과 같다. U 안에서 추가적인 할당을 얼마나 할지 조절하기 위해(U안에서 1의 숫자 control) U안의 전체 assignments가 n+αn과 같아지도록 제한을 둔다. (α는 overlap을 얼마나 허용할지에 대한 인자) 따라서 k-means의 첫 번째 확장은 다음과 같다. 각 데이터 포인트를 모든 클러스터에 할당하는 상황은 제외하기 위해 0≤α≤(k-1)의 제한을 둔다. 이제 여기에 non-exhaustivenss constraint를 추가하면 위와 같다. β는 outlier를 몇 개까지 허용할지에 대한 정보이다. (βn개의 데이터는 어떠한 클러스터에도 속하지 않는다. 즉, n개의 데이

$$\min_{U} \quad \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ij}\|\mathbf{x}_i - \mathbf{m}_j\|^2, \text{ where } \mathbf{m}_j = \frac{\sum_{i=1}^{n} u_{ij}\mathbf{x}_i}{\sum_{i=1}^{n} u_{ij}}$$
$$\text{s.t.} \quad \text{trace}(U^TU) = (1+\alpha)n.$$

(2.3)

$$\min_U \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ij}\|\mathbf{x}_i - \mathbf{m}_j\|^2, \text{ where } \mathbf{m}_j = \frac{\sum_{i=1}^{n} u_{ij}\mathbf{x}_i}{\sum_{i=1}^{n} u_{ij}}$$

$$\text{s.t. trace}(U^TU) = (1+\alpha)n,\ \textstyle\sum_{i=1}^{n} \mathbb{I}\{(U\mathbf{1})_i = 0\} \le \beta n.$$

터 포인트 중 적어도 n-βn개의 데이터 포인트는 반드시 클러스터에 속해야 한다.)

NEO-K-Means의 알고리즘은 다음과 같다.

· 클러스터의 중심을 초기화 한다.

· 클러스터의 중심을 계산한 후, 모든 데이터 포인트와 클러스터 간의 거리를 구한 후, 각 데이터 포인트에 가장 가까운 클러스터와 그 거리를 저장한다. 그리고는 각 데이터 포인트의 가장 가까운 클러스터까지의 거리를 오름차순으로 정렬한다.

· n+αn개 데이터 포인트를 클러스터에 할당하는 데, n-βn개의 데이터는 반드시 어떤 클러스터에 속해야 한다. (클러스터 간 중복을 허용하지 않음)

· n+αn개의 데이터 중 n-βn개를 제외한 나머지 데이터, αn+βn개의 데이터에 대해서는 같은 클러스터에 여러 번 들어가지 않도록 한다.

· 모든 할당이 이루어지고 난 다음에는 각 클러스터에 대해 중심을 다시 계산한다.

· 이 절차를 objective function이 충분히 작거나 maximum number of iterations에 닿을 때까지 반복한다.

**Weighted Kernel NEO-K-Means**

$$\min_U \quad \sum_{c=1}^{k}\sum_{i=1}^{n} u_{ic}w_i\|\phi(\mathbf{x}_i) - \mathbf{m}_c\|^2,$$

$$\text{where } \mathbf{m}_c = \frac{\sum_{i=1}^{n} u_{ic}w_i\phi(\mathbf{x}_i)}{\sum_{i=1}^{n} u_{ic}w_i}$$

$$\text{s.t.} \quad \text{trace}(U^TU) = (1+\alpha)n,\ \textstyle\sum_{i=1}^{n} \mathbb{I}\{(U\mathbf{1})_i = 0\} \le \beta n.$$

NEO-K-Means를 weighted kernel case로 확장하는 것은, non-exhaustive, overlapping 한 그래프 클러스터링이 가능하게 만든다.

**3. Graph Clustering using NEO-K-Means**

$$\max_Y \sum_{j=1}^{k} \frac{\mathbf{y}_j^T A \mathbf{y}_j}{\mathbf{y}_j^T D \mathbf{y}_j}$$

$$\text{s.t. trace}(Y^TY) = (1+\alpha)n,\ \textstyle\sum_{i=1}^{n} \mathbb{I}\{(Y\mathbf{1})_i = 0\} \le \beta n.$$

NC은 그래프 클러스터링의 성능을 확인할 수 있게 한다. non-exhaustive, overlapping 한 그래프 클러스터링을 위해 assignment matrix Y와 α, β를 적용한다. 그리고 나면, weighted kernel NEO-K-Means의 objective와 같아짐을 알 수 있다.

**4. Experimental Results**

NEO-K-Means와 fuzzy, MOC, OKM, explicit/implicit sparsity constrained clustering을 비교했고 Synthetic Data와 'yeast','music','scene' 의 real world dataset을 이용해 실험했다. NEO-K-Means는 다른 최신의 알고리즘과 비교했을 때, 모든 데이터 셋에서 가장 우수한 성능을 보였으며, weighted kernel k-means variation은 large scale 네트워크에서 overlapping communities를 찾게 한다.