Prototype based hierarchical clustering

제안할 알고리즘 동작 시 관측되는 중간 결과 정리 및 알고리즘 설명

Key point: soft clustering, bottom-up clustering, KL-Divergence, merging strategy

Considerations

1. 계층 구조가 만들어지는 tau와 min sample을 결정
   1. Tau 값만의 변화는 prototype의 수의 감소에 영향을 미치지 않는다.
      1. 최소 샘플 수는 변화가 없기 때문에 tau 값의 증가는 HR의 크기만 증가하고 프로토타입의 수는 변화가 없다.
   2. Tau 값을 고정하고 min sample 값을 증가시킬 때 프로토타입의 수는 감소
      1. Tau 값을 고정 후 최소 샘플 수의 증가에 따라 프로토타입의 수는 감소하며 병합되는 효과를 가지지만 검증이 필요
         1. 2차원 데이터로 검증해본다면 이해에 도움을 줄 가능성이 있다.
            1. Consideration 4를 참고
   3. Tau 값의 범위 제한이 필요
      1. 실험에 사용되는 데이터셋 내 임의의 데이터 의 범위는 을 가지므로 tau 값은 으로 제한하여 iteration을 진행하여야 할 것으로 예상
2. Sub cluster 간의 병합이 이루어질 때 생성되야 할 새로운 프로토타입에 대한 정의
   1. 실험에 사용되는 데이터셋 내 임의의 데이터 만 프로토타입으로 선정해야만 하는 것으로 알고 있었음
      1. 교수님께 문의하여 프로토타입에 대한 재정의를 병합되는 sub cluster들의 min값과 max 값을 이용하여 tau값을 계산하고 새로운 프로토타입 을 정의한다.
3. Initial data point 선정 방식의 영향
   1. 프로토타입 결정과 클러스터링 결과에 대한 영향
      1. 반복을 데이터 읽는 순서에 따라서 진행하고 만약 앞쪽의 데이터가 새로운 HR cluster로 선정된다면 뒤쪽의 데이터들은 중복을 제거하는 규칙에 의해 새로운 클러스터로 고려가 안됨
4. 클러스터의 구성 결과에 대한 검증할 지표가 필요
   1. 지도학습 방식과 유사하게 정답 라벨과 클러스터링 결과 간의 비교 방식
      1. 분류 문제로 클러스터링 결과를 비교
      2. 클러스터링의 클러스터 인덱스는 무작위로 선정되기 때문에 다른 방식을 고려해야 할 것 같다.
5. 상향식 클러스터링은 일종의 트리의 형태로 구성되어 계층 정보를 가지고 있어야한다.

Experiment design

1. Prototype based hierarchical clustering with HR distance

실험에 대한 고찰

* 1. tau값과 최소 샘플 수를 변화시키면서 적절한 tau와 최소 샘플 수 선정 방법에 대한 이해를 한다.
     1. 초기 최소 샘플 수 변화에 따른 실험
        1. 초기 최소 샘플 수는 3으로 설정하며 점차 증가시키고 tau값은 고정
           1. 초기 최소 샘플 수가 증가할수록 첫 단계의 프로토타입의 수는 감소
     2. Tau 값 변화에 따른 실험
        1. Tau 값 변화가 미치는 영향을 실험하기 위해 최소 샘플 수를 고정
           1. Tau값의 증가는 프로토타입의 수 감소(병합)을 야기하지 못하며 프로토타입이 포함하는 초월 사각형의 크기만 증가하여 초월 사각형들 간의 겹치는 영역이 증가한다.
     3. Tau값과 최소 샘플 수의 변화에 따른 실험
        1. 두 파라미터 모두 변화, 변화 방식은 추후에 A와 B가 끝나면 결정
     4. Tau 값이 증가할수록 프로토타입의 수는 증가하며 최소 샘플 수의 증가는 프로토타입의 수가 감소한다.
  2. 적절한 tau값과 최소 샘플 수는 무엇일까?
     1. ‘적절하다’의 정의를 내려야 tau값의 증가와 최소 샘플 수의 증가에 대한 기준을 설정할 수 있다.
        1. 예를 들면, 데이터의 수는 500개이고 알고리즘의 첫 단계에서 초월 사각형의 수는 500이다. 다음 단계에서 초월 사각형의 수를 조정할 때 이전 단계의 반으로 설정하여 파라미터를 찾는 방식
        2. 두 번째 방법:
           1. n개의 데이터를 가지는 데이터셋에서 초기 tau와 최소 샘플 수로 구성된 k개의 중복을 허용하지 않는 초월 사각형 집합을 생성한다(0<k<=n). 각 초월 사각형들은 초월 사각형 내 데이터들을 대표하는 하나의 프로토타입을 가진다.
           2. k개의 프로토타입에 대해서 병합을 진행하고 새로운 중심점(프로토타입)과 초월 사각형의 대각선 길이(tau)를 정의한다.
           3. 모든 점이 하나의 초월 사각형으로 병합될 때까지 (b) 과정을 반복하거나 tau와 최소 샘플 수의 변화에도 병합이 이루어지지 않을 경우까지 반복한다. 이전 단계에서 병합을 시도할 때보다 현 단계에서 병합을 시도할 경우 프로토타입의 수는 감소하고 프로토타입들 간의 거리는 증가하므로 tau값은 증가시키고 최소 샘플 수는 이전 단계에서 생성된 새로운 프로토타입의 수까지 반복
     2. 다음 단계에서 초월 사각형을 찾을 때 tau값을 증감 여부?

프로그램 설계

1. Find\_noise

모델에 주어진 초기 tau값과 최소 샘플 수에 기반하여 임의의 데이터 에서 tau의 거리 내 최소 샘플 수 이상의 데이터가 존재하는 조건을 만족하는 영역을 추출하고 조건에 부합하지 않는 임의의 데이터 는 노이즈로 처리

1. Get\_prototypes

각 단계에서 추출한 초월 사각형들의 프로토타입과 해당하는 인덱스를 저장

1. Merging\_prototypes

초월 사각형들 간의 병합이 발생할 경우 병합이 발생한 초월 사각형들에 대해 새로운 프로토타입과 tau값을 계산하여 새로운 초월 사각형을 정의

1. Make\_hierarchy\_hr

종료 조건인 두 가지 경우를 만족할 때까지 반복하여 새로운 초월 사각형들을 찾는다. 첫 번째 조건, 모든 점이 하나의 초월 사각형으로 병합

두 번째 조건, tau와 최소 샘플 수의 변화에도 초월 사각형들의 상태 변동이 없는 경우

1. Fit\_predict
   1. 종료 조건을 만족할 때까지 make\_hierarchy\_hr 함수를 실행
      1. Find\_noise함수를 통해 입력 데이터에 대한 초월 사각형들을 생성
      2. Get\_prototypes 함수로 생성된 초월 사각형 정보인 프로토타입과 초월 사각형이 가지는 데이터의 인덱스를 저장
2. Predict
3. Prototype based hierarchical clustering with KL-Divergence
   1. HR 클러스터들 간의 유사도 또는 거리를 측정하기 위한 KL-Divergence 계산은 두 가지 방식으로 진행한다.
      1. 각 HR 클러스터가 포함하는 데이터들을 각 축에 대해 1변량 정규분포로 확률밀도함수를 생성하여 계산하는 방식
      2. 모든 축을 포함하는 다변량 정규분포로 확률밀도함수를 생성하여 계산하는 방식
   2. KLD 값에 따른 순위 비교를 통해 사용자가 정한 k개의 sub cluster를 병합하는 방식을 선정한다.
      1. K의 수에 따른 클러스터 구성 결과 비교

가제: Prototype based hierarchical clustering with HR distance

1. Assumption
   1. 최소 샘플 수는 프로토타입의 수만큼 반복을 진행
   2. 초기 모델의 프로토타입의 수는 데이터의 수와 동일
   3. 다음 단계의 프로토타입의 수는 이전 단계의 프로토타입의 수의 절반으로 설정
2. Algorithm
   1. 데이터의 수만큼 프로토타입을 생성하기 위한 초기 tau값과 최소 샘플 수를 반복을 통해 설정한다.
   2. 데이터의 수만큼 프로토타입을 생성
   3. Tau값을 고정하고

가제: Prototype based hierarchical clustering with KL-Divergence

1. Assumption
   1. 두 sub cluster간의 분포 유사도 측정을 위해 KL-Divergence를 계산하고 함수에 대입하여 0과 1의 값 범위를 가지는 sub cluster 유사도 지표 M을 생성한다.
   2. HR 클러스터 생성 알고리즘을 통해 생성된 HR 클러스터들 중 어떤 하나의 HR 클러스터 HR1은 지표 M을 이용하여 1~k번 째로 높은 유사도를 보이는 HR 클러스터들과 병합한다. K는 사용자가 정하는 파라미터이다.
   3. 병합을 진행하면서 HR 클러스터가 주어진 모든 데이터를 포함할 때 병합은 종료된다.
2. Algorithm
   1. 데이터의 수만큼 프로토타입을 생성하기 위한 초기 tau값과 최소 샘플 수를 반복을 통해 설정한다.
   2. 데이터의 수만큼 프로토타입을 생성
   3. Tau값을 고정하고