**Classification by Coreset selection**

GSDS 석사과정 20224314

강현구

1. **Abstract**

Semi-supervised learning(이하 SSL)은 Labeling Cost 문제를 해결할 수 있는 주요 방안 중 하나이다. 하지만 모델을 학습하거나, Dataset의 Diversity를 반영하지 못할 정도로 Label Data가 부족할 경우 Confirmation bias가 발생할 수 있다. Pseudo labeling간 이를 막기 위해 클래스별 Prior 확률을 활용하나, 비현실적이다. 따라서 Label 데이터 선정부터 Dataset의 Diversity를 고려하며, Dataset에 대한 사전 지식 없이 Label Data 문제를 해결할 필요가 있다.

본 연구에서는 Active learning의 coreset Selection을 활용한 새로운 Classification 방법을 제안한다. Coreset Selection을 통해 Dataset의 Diversity를 보장한다. 또한 Coreset Selection 간 형성한 subgraph들은 기하적 관점에서 모델의 성능에 의존하지 않고 높은 신뢰도의 Classification을 가능하게 한다. 즉, SSL 모델에 의거하지 않은 Classification으로 Pseudo labelling 함으로써 Label Data 부족 문제를 해결한다. 본 방법은 클래스의 수가 적고 데이터 분포가 밀접해 있는 데이터셋에 대해선 높은 효율 및 정확도를 갖출 것이다. Distance base method가 적용될 수 있는 모든 분야에 적용할 수 있다. 또한 Represent learning과의 시너지가 높다. 또한 적절한 Sampling size를 제안할 것이다. 이로써 현장에서 Labeling cost를 측정할 여지를 준다.

**1. Introduction**

최근Deep learning(이하 DL)모델은 많은 양의 Labeled Data를 기반으로 다양한 분야에서 성과를 내고 있다. 하지만 모델이 필요로 하는 데이터가 증가함에 따라, Labeling Cost를 어떻게 해결할 것인가 중요한 화두가 되었다. 이와 관련하여 Semi-supervised learning(이하 SSL)의 중요성이 강조된다. SSL은Label Data과 Unlabel Data모두를 활용한다. 그 결과 SSL은 적은 비중의 Label Data로도 높은 성능을 낼 수 있다. SSL은 Labeling Cost 로 인해 supervised learning이 제한되는 분야에 폭 넓게 사용되고 있다.

SSL 은 SL 방법과 달리 오류에 취약하다. Labeled 데이터의 규모가 커 Dataset을 잘 반영하며 학습 모델의 성능을 이끌어 낼 수 있다면 괜찮다. 하지만 Label 데이터가 적어 Dataset의 Diversity을 반영하지 못하며, 또한 모델을 학습하기에 충분하지 않다면 Confirmation bias가 발생한다. 초기에 발생한 Confirmation bias을 해결하지 못한다면 모델 전체의 성능을 크게 악화시킬 수 있다.

SSL은 크게 Consistency 방법과 Pseudo Labeling 방법 2가지로 나눠진다. Consistency 방법은 모델의 Input에 변형이 가해져도 모델의 결과는 일정하게 유지돼야 한다는 가정 하에 이뤄진다. Label Data에 변형을 가한 데이터를 생산(augmentation) 함으로써 동일한 Class의 Label Data를 생성한다. 이로써 Confirmation bias를 피할 수 있다. 한편 Pseudo labeling은 학습된 모델을 통해 신뢰도가 높게 나온 Unlabel Data에 대해 Pseudo labeling을 부여한다. Pseudo labeling 한 Unlabel data를 모델 학습에 활용한다. 단, Confirmation bias을 해소하지 못했을 경우 오히려 성능이 악화된다. 이를 해소하기 위해 모델 Fine tuning 과정에만 한정적으로 활용하거나, 신뢰도에 대해Threshold를 설정, 또는 규제항을 도입한다. 그러나 Pseudo labeling은 Unlabel에 대한 판단을 학습 모델 성능에 의존한다. 즉, Label Data가 부족하여 모델 자체의 성능을 보장할 수 없을 때 Confirmation bias는 필연적이다. 초기 모델 오류를 방지하고자 각 클래스별 비율에 대한 Prior을 규제항으로 도입하지만, Dataset에 대한 사전 확률을 안다는 것은 비현실적이다.

여러 단점에도 불구하고 Pseudo labeling을 연구할 필요성이 있다. 일부 몇 논문에서는 Pseudo labeling에서 Confirmation bias를 방지하였을 경우 Consistency regularization 만큼의 성능을 확보한 연구가 있다. 무엇보다 Pseudo labeling은 Consistency regularization과 상반된 방법이 아닌 병행 가능한 방법이다. Pseudo labeling의 성과를 향상시키는 것은 추후 Consistency regularization 방법과의 연계에서 좋은 결과를 가져올 것이다.

본 연구에서는 Active learning의 일환인 Coreset-selection 방법을 통해 모델 성능에 의존하지 않는 Classification 방법을 제안한다. 이로써 Label Data 부족으로 인한 Dataset의 Diversity 미반영 및 SSL 모델 초기 학습 Confirmation bias 발생 문제를 해결할 것이다. Active learning은 어떤 Acquisition Strategy을 활용하냐에 따라 Sampling Data가 Dataset의 특정 정보를 반영시킬 수 있다. Coreset selection은 Dataset의 Diversity를 보장하는 Data를 Sampling 한다. 더 나아가 본 연구에서 기하적 관점에서 각 Label Data는 Unlabel Data에 대해 대표성을 띄고 있음을 증명할 것이다. 이후 모델의 성능에 의존하지 않고 높은 신뢰도의 Classification이 가능함을 보일 것이다. Classification을 통해 Pseudo label data를 확보하여 Label Data 부족 문제를 해결한다. 또한 신뢰도 높은 Classification을 위해 필요로 하는 Sampling size를 제시할 것이다.

본 연구는 새로운 Classification 방법이라는 점에서 Contribution이 있다. 또한 SSL에 적용하여 Label Data 부족 문제를 해소할 수 있다. 또한 Subgraph를 활용한 기하적 관점에서의 Classification은 클래스의 수가 적고 데이터 분포가 밀접해 있는 데이터셋에 대해선 높은 효율 및 정확도를 갖출 것으로 기대한다. Distance base method가 적용될 수 있는 모든 분야에 적용할 수 있다. Dataset의 주요 feature들을 추출할 수 있고, 데이터의 분포에 영향을 줄 수 있는 Representation learning과의 시너지가 높다. 또한 본 연구는 Classification의 성능을 보장하기 위해서 필요로 하는 Sampling size의 범위를 계산할 것이다. 이는 현장에서 성능 확보를 위해 필요한 Labeling Cost를 측정할 수 있도록 한다.

본 연구는 DL 모델과 연계 가능한 Neural network를 활용하지 않는 distance based classification이라는 점에서 contribution이 있다. 또한 Active learning과 SSL의 연계 연구 맥락에서 contribution이 있다. Active learning과 SSL의 연계에 대한 연구는 드물다. 일부 있는 연구조차 Active learning과 SSL을 병렬적으로 활용한다는 한계가 있다. AL은 모델 학습에 가장 도움될 것으로 여겨지는 Data를 Sampling 하는 방법이며, SSL은 주어진 Label Data를 Leverage 하는 방법이다. 즉, 두 방법을 잘 연계한다면 추가적인 시너지를 낼 것이라 판단된다. 본 측면에서 AL이 제공하는 Data의 특성을 적극적으로 활용한 본 SSL 방법론은 좋은 시작점이 될 것으로 생각한다.

**2. Related works**

2-1). Semi supervised learning

SSL은 Labeling Cost 문제로 일부의 Label Data만을 구할 수 있는 현장에서 널리 활용된다. SSL은 크게 3가지 가정을 기반으로 한다. 첫째, 확률 밀도가 높은 지역의 입력값 x1, x2가 가깝다면 각각의 연관된 Label y1, y2도 가깝다. 둘, 모델의 결정 경계가 데이터의 확률 밀도가 높은 곳을 지나지 않는다. 마지막으로 고차원의 입력 데이터가 저차원의 Manifold를 따라 놓여 있다. 즉, SSL은 Data 간의 거리가 Data 특성의 유사성을 반영하며, 각 Class 별로 밀도를 통해 경계 지점을 파악할 수 있을 때 활용할 수 있다. 그 결과 Distance-based 방법들이 적용될 수 있는 다양한 영역에 폭 넓게 활용되고 있다.

SSL의 주요 방법 중 하나는 Consistency regularization 방법이다. 모델의 Input에 변형이 가해져도 모델의 결과는 일정하게 유지돼야 한다는 가정 하에 적용된다. Confirmation bias를 효과적으로 제거하여 Pseudo labeling 보다 적극적으로 활용되고 있다. Label Data을 Permutation하거나 Noise를 적용한 경우에도 Class는 변하지 않는 다는 점을 활용하여 Label Data를 Augment 한다. 한편 Unlabeled Data에 대해서도 정확한 Class를 몰라도 Data augmentation한 것과의 예측은 동일해야 함을 활용한다.

또 다른 SSL의 방식은 Pseudo labeling이며, 본 연구에서 주로 다룬다. Pseudo labeling은 Unlabel Data에 대한 예측을 leverage함으로써 모델의 성능을 향상시킨다. 단, 초기의 오류가 확대 재 생산되는 Confirmation bias가 발생할 수 있다. 이 문제를 해결하기 위해 Pseudo labeling을 모델 Fine Tuning에만 한정적으로 사용한 후 제거하거나, Threshold를 설정하여 일정 이상의 신뢰도를 가진 값에만 Pseudo labeling을 부여하는 방안을 쓰고 있다. Confirmation bias는 Label Data가 모델을 학습하기에도 부족할 때 심화된다. Label Data가 적은 학습 초기의 모델의 성능은 좋지 않다. 따라서 Pseudo labeling 한 결과가 실제와 달라, 실제 Dataset 에서의 클래스 비율과 상이하게 달라질 수 있다. 이를 막기 위해 Dataset의 클래스의 사전 확률을 규제항으로 도입하나, Dataset의 사전 정보를 규제항으로 사용한다는 것은 비현실적이다.

2-2) Active learning

Active learning(이하 AL) 또한 Labeling cost를 줄일 수 있는 방법이다. AL은 적은 수의 Labeling으로도 전체 데이터셋이 Labeling 되었을 때와 유사한 성능을 가져올 정보 가치가 높은 Unlabel Data를 선별한다. AL의 유형은 크게 3가지가 있다. Membership query Synthesis은 Labeling을 요청할 데이터를 생성한다. Stream-based Selective sampling은 새로운 sample Data에 대해서 레이블이 필요한지를 판단한다. 마지막으로 Pool-based sampling은 주어진 Dataset에서 중요한 데이터를 선별한다. 대부분의 AL 연구는 Pool-based sampling 연구에 치중되어 있으며, 본 연구에서도 Pool-based sampling에 초점을 둔다.

Pool-based sampling AL에서 Acquisition strategy에 따라 Unlabel Data를 선별한다. 크게 Uncertainty-based approach, Expected-based approach, Diversity-based approach가 있다. Uncertainty-based approach는 Bayesian network Model에 의거하여 각 Data별 Uncertainty를 측정한다. 이후 Uncertainty score가 높은 순으로 데이터를 Sampling 하여 Labeling을 한다. Uncertainty 가 높은 데이터들이 각 Class 별 경계 값 주변에 놓여 있는 것을 고려할 때, Uncertainty가 높은 데이터들은 서로 비슷한 상황에 놓여 있는 경우가 많다. 즉, 각 Class 별로 경계 영역을 학습하는 데에는 도움이 되나 Dataset의 Diversity를 반영하지 못한다. Expected-based approach는 모델 성능을 가장 향상시킬 것으로 기대되는 Data들을 선별한다. 단, Uncertainty based approach와 동일하게 모델의 성능에 의거하여 판단한다는 점에서 Label Data가 부족할 때엔 적용하기 어렵다. 마지막으로 Diversity-based approach는 데이터셋의 Diversity를 보장할 수 있는 데이터를 선별하는 방법이다. 대표적으론 Coreset selection이 있다.

Coreset Selection은 모델 성능의 기댓값을 최대화하는 Data를 Sampling 한다. 이는 주어진 Sampling size b개로 전체 데이터를 덮을 수 있는 subgraph들을 구성할 때, 최소한의 반지름 를 가지게 만드는 데이터를 Sampling 하는 것과 동일하다{Sener, 2017 #4}. 각 Sampling Data들은 Dataset의 밀집 정도에 무관하게 넓게 퍼져 있다. 이로 인해 Sampling point 는 특정 클래스의 값 만을 포함하는 것이 아닌 전반적인 데이터 셋을 반영한다. 또한 Data 간의 거리를 기반으로 Sampling 하여 SSL 모델의 성능에 의존하지 않는다. 따라서 Label Data가 부족한 상황에서도 동일한 Sampling 성능을 낼 수 있다.

**3. Method**

3-1) Classification

모델의 성능에 의존하지 않고 신뢰도 높은 Classification을 통해 Pseudo labeling을 부여하겠다. Label Data가 부족하기 때문에 모델의 성능을 신뢰할 수 없다. 따라서 본 연구에서는 Coreset selection 간 형성된 subgraph를 활용하여 신뢰도 높은 Classification을 진행한다. Dataset을 이라 하자. Sampling size를 N이라 할 때, i번째로 Sampling 된 data는 라 한다. 각 들은 subgraph 의 중점이다. subgraph 의 반지름 을 최소화시키는 를 로 Sampling한다. 반지름 은 이다. 각 subgraph의 밀도는 해당 subgraph에 포함된 Data의 개수로 측정할 수 있다. 확률 밀도가 높은 지역의 입력값 x1, x2가 가깝다면 각각의 연관된 Label y1, y2도 가깝다는 SSL의 가정을 통해, 밀도가 높은 subgraph 에 속한 Data s.t. j ∈ {1,…, | }들은 높은 확률로 중점인 와 클래스가 동일할 것이다. 즉, 은 subgraph에 속한 unlabel Data를 대표할 수 있다.

더 나아가 Active selection을 통해 형성한 subgraph들이 Dataset을 촘촘하게 덮을 수 있을 정도로 충분히 작은 반지름 을 가졌다고 가정한다. s.t. k ∈ {1,…, | } 를 와 접하고 있는 subgraph라고 하자. 각 subgraph의 중점 , 들의 클래스는 알고 있다. 이때 기하적 관점에서 의 Class가 다를 경우 와 가 서로 다른 클래스들이 겹치는 곳에 있다고 유추해볼 수 있다. 반대로 모든 와 의 클래스가 동일할 때 특정 클래스의 중심부에 위치할 것이다. 즉, Figure1과 같이 subgraph 간의 기하적 관계와 각 중점의 클래스를 통해서 각 subgraph들이 클래스별 데이터 분포간 어디에 위치해 있는지 유추할 수 있다. 즉, 이며, 을 성립하는 subgraph 에 속한 unlabel Data 에 대해 와 동일한 Class로 Classification을 할 수 있다. 는 hyperparameter이다.

[Figure 1. Subgraph 간 기하적 관계]

본 Classification 방법은 클래스의 종류가 적고, Unlabeled data가 많아 특정 클래스의 분포가 명확히 구분할 때 성능이 향상된다. 충분히 작은 반지름 , 클래스별 데이터 밀집이 보장된다면 각 클래스별 중심부에 해당하는 데이터들을 높은 신뢰도로 Classification을 할 수 있다. Confirmation bias를 피하기 위해 Classification의 결과는 모델의 Fine-tuning에만 활용한다. 이로써 Active learning을 통해 Dataset의 Diversity를 반영하며 동시에 Label Data 부족 문제를 해소할 수 있다.

Coreset selection을 다회차 진행하여 subgraph들이 중첩되었을 때 또 다른 방식으로 Classification이 가능하다. 1번의 Coreset selection을 통해 모든 들은 1개 이상의 subgraph에 포함된다. 따라서 Coreset selection이 총 P번 진행된다면 모든 들은 최소 P개의 subgraph안에 속하게 된다. 은 특정 Class의 중점을 가진 subgraph에 많이 속할수록 동일한 클래스일 확률이 높을 것이다. 또한 각 subgraph들의 반지름 에 반비례하여 확률이 높을 것이다. 각 클래스 별로 subgraph에 속하는 횟수 와, 반지름을 기반으로 softmax를 통해 확률화한다.

(3.1)

Coreset selection 회차에 따라 반지름 을 작게 만들기 위해 첫번째 방법으로 Pseudo labeling되지 않은 Unlabel Dataset 에 한해 진행한다. 데이터의 클래스를 가장 높은 확률 값으로 고정할 경우 Confirmation bias가 발생할 수 있다. 따라서 Mix up{Zhang, 2017 #8}을 적용하여 soft labeling한다. 또한 충분한 횟수의 subgraph의 중첩이 생기기 전엔 틀린 pseudo label을 부여할 확률이 높다. 따라서 허용 오차율 을 도입하여 특정 class일 확률이 1-보다 클 경우만 pseudo label을 부여한다.

[Figure 2. P 회차 Coreset selection을 통해 에 중첩된 subgraph]

3-2) Choose Proper

Classification의 신뢰도를 보장하기 위해 충분히 작은 반지름 을 적용해야 한다. 반지름 의 크기가 충분히 작을수록 클래스 데이터의 주변부, 중심부를 명확히 구분할 수 있다. 반대로 반지름의 크기가 크다면 subgraph들이 주변부와 겹쳐 Classification 성능이 낮아질 것이다. 더불어 Minor Class 데이터가 분포한 영역이 한 subgraph에 포함되는 등 오차가 커질 것이다.

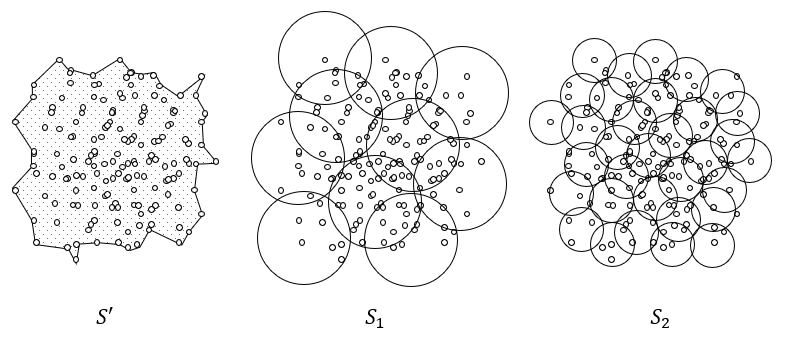
적절한 Sampling size N을 고를 가이드라인이 필요하다. Labeling Cost와 classification 정확도는 Trade off 관계를 띈다. 반지름 의 크기는 Sampling size N에 달려 있다. 즉, N의 크기가 작을 땐 Labelling Cost는 줄어들지만 이 커져 신뢰도 높은 Classification을 할 수 없다. 반대로 N의 크기가 크다면 이 줄어들어 Classification의 성능이 향상된다. 본 연구에선 Confirmation bias을 방지하는 것이 우선이라 판단한다. subgraph의 속하는 데이터의 개수가 일정 이상 넘지 않도록 규제함으로써 Confirmation bias를 방지하겠다.

모든 subgraph에 속한 데이터 개수를 M개 이하로 만들 반지름 을 찾는다. 임의의 Sampling size 으로 Coreset selection을 수행한다. 가장 밀도가 높은 subgraph에 속하는 Data에 한정하여 다시 Coreset selection을 수행한다. 이때 각 subgraph에 속한 Data의 수가 M개 보다 적도록 하는 반지름 을 찾는다. 반지름 은 가장 밀도가 높았던 subgraph에 속하는 데이터의 개수를 M개로 줄인다. 따라서 전체 dataset에 확장했을 때에도 모든 subgraph는 M개 이하의 Data만을 포함한다.

반지름의 크기를 보다 크지 않게 만들 Sampling size N’을 찾을 것이다. Dataset X의 넓이 S를 통해서 Sampling size N’의 범위를 형성할 수 있다. 단, Outlier가 데이터 분포에서 크게 벗어나 있는 경우에 대해 robust을 확보하기 위해, Dataset의 넓이 S를 subgraph 들로 cover되는 넓이로 재정의한다. 예시로 반지름 의 subgraph을 2개를 만들었다고 하자. 이때 각 subgraph의 중점이 2이상의 거리를 띄고 있다면, 서로 겹치지 않아 S는 2\* 가 된다. 만약 두 subgraph가 겹친다면 2 에서 겹친 부분만큼 뺀다. 겹치는 부분은 두 중점 사이의 거리를 활용하여 구할 수 있다. 이후 각 간의 거리는 최소 이상의 간격을 두고 있다는 특성을 활용하여 넓이 S에 대한 Bound을 형성할 것이다. 두 중점 사이의 거리에 대한 특성은 Lemma 1을 통해서 증명한다. 우리의 목표는 아래 Theorem을 증명하는 것이다.

**Theorem 1.** 임의의 Sample size과 반지름의 쌍 (N,)이 주어졌을 때, 반지름의 크기가 보다 크지 않도록 하는 N’의 범위를 계산할 수 있다.

전체 과정은 다음과 같다. 먼저 을 찾는다. 임의의 Sample size 으로 Coreset selection 했을 때의 반지름의 크기를 이라고 한다. 이때의 넓이를라고 하자. 또한 Sample size N’, 반지름 반지름의 경우의 넓이를 라고 하자. N’은 임의의 보다 크다 가정하는 것은 어렵지 않다. 각 모두 subgraph가 촘촘히 Dataset를 Cover 할 수 있도록 을 충족시켜야 한다. 또는 Outlier에 대해 Robust 하도록 을 적용한다. 추후 둘의 차이를 경험적으로 검증하겠다. Dataset의 가장 바깥에 놓인 Data들로 표현되는 넓이를 S’ 라고 하자. coreset의 subgraph들은 모든 Data들을 cover한다. 따라서 S’보다 의 넓이가 큰 건은 자명하다. 또한 Figure 3을 통해서 반지름이 더 큰 가 보다 클 것임을 확인할 수 있다. 따라서 을 만족할 것이다. 여기서 을 계산할 수 있고, N’과 의 수식으로 의 값을 구할 수 있다면, 목표 값인 N’에 대한 부등식을 만들 수 있다.

[Figure 3. 넓이 정의 및 subgraph 반지름 크기 차이에 따른 넓이 비교]

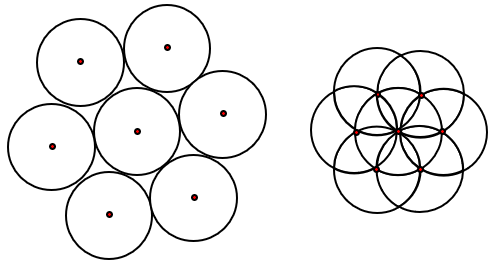
또한 , 에서 을 증명해줄 수 있다면 N’에 다른 반대 방향의 부등식을 만들 수 있다. 이에 대해 이론적 또는 경험적으로 성립하는지 확인하겠다.

, , 에 대해 계산하겠다. 은 Sample size N’와 반지름 이 주어졌으니 각 subgraph들이 모두 겹치는 부분이 없을 때임을 쉽게 생각할 수 있다. 을 유추하기 위해선 각 subgraph 중점에 대한 특성을 활용해야 한다. 하나씩 증명하겠다.

**Lemma 1.** 각 subgraph의 중점들은 최소 이상의 간격을 두고 있다.

Coreset selection은 k-center greedy 알고리즘과 방법이 동일하다. 첫 Sampling Data는 Random하게 Sampling 한다. 그리고 은 Sampled Data 와 가장 거리가 먼 거리로 정의한다. 이후 의 거리를 가진 를 로 Sampling 한다. 이 과정을 Sampling size N에 도달할 때까지 반복한다. 이로써 이며, for all 이 성립한다. 의 거리를 가지는 점이 2개 이상일 때 등호가 성립한다. 따라서 을 성립한다.

Lemma 1에 의거하여 은 Figure 4와 같이 각 subgraph들이 모두 반지름 만큼의 간격으로 최대한 겹쳐 있는 상황임을 생각할 수 있다. 이로써 각각의 넓이를 구할 수 있으며, Lemma 2에서 밝히도록 하겠다.

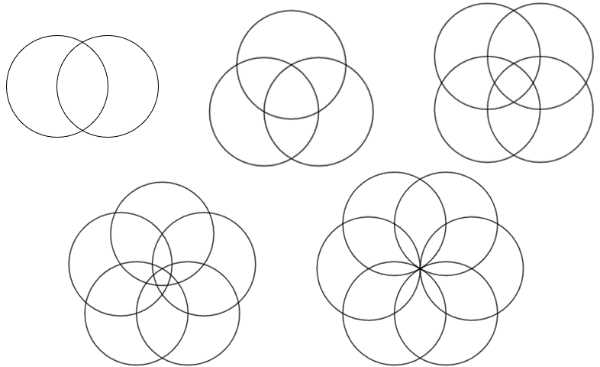
[Figure 4. ), ) 예상도]

**Lemma 2.** Sample size가 n이고 반지름이 이라 하자. Let k = , 이때 ) = , 이다.

\*계산과정이 복잡하여 추후 Appendix에 기술하도록 하겠음.

넓이 S를 계산하기 위해서 고려해줘야 할 사항들은 여러가지가 있다. 먼저 각 subgraph들이 겹치는 경우의 수는 두 subgraph가 겹칠 때부터 5개의 subgraph가 동시에 겹칠 때까지 총 4가지가 있다. 이는 Lemma 3에서 증명하겠다. 각 4개의 경우가 발생하는 경우는 subgraph 간의 Adjacency Matrix를 통해 구분할 수 있다. 마지막으로 4개의 경우의 넓이를 각 중점 간의 거리, 반지름 을 통해 구할 수 있다.

**Lemma 3.** 동일한 반지름을 가지는 원들의 중점 간의 거리가 반지름 이상일 때, 다수의 원이 동시에 겹치는 경우의 수는 총 4가지다.

 원들이 최대한 겹치는 경우를 가정하기 위해 모든 원들이 서로 다른 원의 중점을 지난다고 가정한다. Figure 5는 원이 동시에 2개 겹치는 경우부터 6개의 원이 겹치는 경우까지 묘사한다. 이때 6개의 원이 겹치는 순간 서로 겹치는 영역은 한 점에 불과하다. 즉, 최대한 겹치는 상황을 가정했을 때에도 5개의 원이 동시에 겹치는 것이 한계임을 확인할 수 있다.

[Figure 5. 원의 중점 간의 거리가 이상일 때의 경우의 수]

**Lemma 4**. Adjacency Matrix를 통해 각 경우의 수를 구분할 수 있다.

**Lemma 5**. 각 경우에 대해서 중점 간의 거리, 반지름 을 통해 넓이 S를 구할 수 있다.

각 Lemma 4,5는 추후 Appendix를 통해서 서술하도록 하겠다. 이로써 와 과 관련된 모든 값들을 N, N’, 에 대한 식으로 바꿀 수 있다. 이때 N’을 제외한 나머지 값들은 상수이다. 따라서 식을 정리함으로써 N’에 대한 부등식을 구할 수 있다. 이외에도 subgraph의 효율을 보장하기 위해 각 subgraph 마다 최소 M’ 개의 점을 포함하도록 고려할 수 있겠다.

**4. 예상 기대효과**

본 연구는 Coreset selection의 Subgraph를 기하적인 관점에서 활용한 Classification 방법을 제시한다. 이를 Pseudo labeling으로 활용함으로써 Label Data 부족으로 인한 confirmation bias를 해소할 수 있다. 더 나아가 Classification의 성능을 보장하기 위해 필요로 하는 Sample size에 대한 bound를 제시한다. 이는 현장에서 Classification 성능 확보까지 필요로 하는 Labeling Cost를 계산할 수 있는 여지를 제공한다.

본 방식은 Dataset의 특성에 따라 성능이 변한다. 데이터셋의 크기가 크고 클래스가 적어 밀도가 높은 Dataset일수록 Classification 성능이 뛰어날 것이다. 한편 본 방법은 Distance을 측정할 수 있으며, distance가 데이터 간의 특성의 유사성을 반영하는 모든 Dataset에 적용할 수 있다. 더불어 representation learning과 함께 활용한다면 시너지를 낼 수 있다는 점에서 가치가 있다.

**<참고문헌>**

Liu, P., Zhang, H., & Eom, K. B. (2016). Active deep learning for classification of hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10*(2), 712-724.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., . . . Wang, X. (2021). A survey of deep active learning. *ACM computing surveys (CSUR), 54*(9), 1-40.

Sener, O., & Savarese, S. (2017). Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

Siméoni, O., Budnik, M., Avrithis, Y., & Gravier, G. (2021). *Rethinking deep active learning: Using unlabeled data at model training.* Paper presented at the 2020 25th International Conference on Pattern Recognition (ICPR).

Wang, K., Zhang, D., Li, Y., Zhang, R., & Lin, L. (2016). Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology, 27*(12), 2591-2600.

Zhu, J.-J., & Bento, J. (2017). Generative adversarial active learning. *arXiv preprint arXiv:1702.07956*.