**Classification by Coreset selection**

GSDS 석사과정 20224314

강현구

1. **Abstract**

Semi-supervised learning (SSL) is one of the main ways to solve the labeling cost problem. However, confirmation bias can occur if the number of label data is insufficient enough to reflect dataset's diversity or for learning model. To prevent this in pseudo labeling, the prior probability for each class is used, but it is unrealistic. Therefore, it is necessary to consider dataset's diversity from the selection of label data, and solve the scarce of label data problem without prior knowledge of dataset.

In this study, propose a new classification method using coreset selection of active learning. It ensures dataset's diversity through Coreset Selection. In addition, the subgraphs formed by coreset selection enable high-reliability classification without relying on the performance of the model from a geometric perspective. In other words, the problem of lack of label data is solved by pseudo-labeling with this classification method that is not based on the SSL model. This classification will have a wide range of applications where distance base method can be applied. It also has high synergy with representation learning. Also suggest an appropriate sampling size. This gives chance to measure the labeling cost in the field.

**1. Introduction**

Recently, the deep Learning (DL) model has been achieving results in various fields based on a large amount of labeled data. However, as the data required by the model increases, how to solve the labeling cost has become an important topic. In this regard, the importance of Semi-supervised learning (SSL) is emphasized. SSL utilizes both label data and unlabel data. As a result, SSL can achieve high performance even with a small proportion of label data. SSL is widely used in areas where supervised learning is restricted due to labeling costs.

Unlike the SL method, SSL is vulnerable to errors. If the size of the labeled data is large and it reflects the dataset well and can lead to the performance of the learning model, it is fine. However, due to the lack of Label data, it does not reflect dataset's diversity, and if it is not enough to learn the model, confirmation bias occurs. Failure to solve the initial confirmation bias can significantly deteriorate the overall performance of the model.

SSL is largely divided into two methods: the consistency regularization and the pseudo labeling. The consistency regularization is based on the assumption that the results of the model must remain consistant even if the input of the model is deformed. Label data of the same class is generated by augmenting data that has been deformed to label data of the same class is generated. In this way, confirmation bias can be avoided. Meanwhile, pseudo labeling utilize classification result of learning model which has high reliability. However, if the confirmation bias is not resolved, the performance deteriorates. To solve this problem, pseudo labling data is used only in the model fine tuning process, set a threshold for reliability, or introduce a regulatory term. However, Pseudo labeling relies on the performance of the learning model. In other words, confirmation bias is inevitable when the performance of the model itself cannot be guaranteed due to lack of label data. To avoid initial model errors, prior study introduces Prior for each class-specific ratio as a regulatory term, but it is unrealistic to know the prior probability for Dataset.

Despite various disadvantages comparing with consistency regularization, there is a need to study Pseudo labeling. In some papers secured the performance similar to Consistency Regulation when confirmation bias is prevented in pseudo labeling. Above all, pseudo labeling is not a method contrary to consistency regularization, but a method that can be parallel. Improving the performance of pseudo labeling will bring good results in connection with the consistency regulation method in the future.

In this work, we propose a new classification method that does not rely on model performance through the Coreset-selection method, which is one of active learning. This will solve the problem of dataset's diversity unreflected and SSL model initial learning Confirmation bias due to lack of Label Data. Depending on what acquisition strategy Active learning utilizes, sampling data can reflect certain information in dataset. Coreset selection samples data that guarantees dataset's diversity. Furthermore, in this study, it will be proved that each sampled data is representative of unlabel data from a geometric perspective. It will then be shown that high-reliability classification is possible without relying on the performance of the model. The problem of lack of label data is solved by pseudo-labeling data through this classification. Study will also present the sampling size required for reliable classification.

This study has a contribution in that it is a new classification method which is not depend on neural network model. It can also be applied to SSL to resolve the label data shortage problem. In addition, classification from a geometric perspective using subgraph is expected to have high efficiency and accuracy for datasets with fewer classes and dense dataset. It is judged that the scope of application will be wide in that it is distance based method. It also has high synergy with representation learning, which can affect the distribution of data and extract main feature of dataset. And study will suggest required sampling size for classification performance. It provides chance for calculating the labeling cost for achieving performance in the field.

This study has a contribution in that it is a distance-based classification that does not utilize the neural network that can be linked to the DL model. In addition, there is a constribution in the context of the linked research between Active Learning and SSL. Research on the link between AL and SSL is rare. Even some studies did, but they have limitations in using AL and SSL in parallel. AL is a method of sampling data that is considered to be most helpful for model learning, and SSL is a method of leveraging a given label data. In other words, if the two methods are well linked, it is judged that additional synergy will be created. In this respect, this SSL methodology, which actively utilizes the characteristics of the data provided by AL, will be a good starting point.

**2. Related works**

2-1). Semi supervised learning

SSL is widely used in the field where only some label data is available due to the labeling cost problem. SSL is largely based on three assumptions. First, if the input data x1 and x2 of the region with a high probability density are close, the respective associated label y1 and y2 are also close. Second, the decision boundary of the model doesn't exist where the probability density of the data is high. Finally, high-dimensional input data is placed along the low-dimensional manifold. In other words, SSL can be used when the distance between data reflects the similarity of the data characteristics and the boundary point can be identified through density for each class. As a result, distance-based methods are widely used in various areas to which they can be applied.

One of the main methods of SSL is the consistency regulation method. It is applied under the assumption that the results of the model must remain consistant even if deformation is applied to the input of the model. Confirmation bias is effectively removed and is being used more actively than pseudo labeling. It utilizes that even if label data is permuted or noise is applied, label data is augmented by taking advantage of the fact that the class does not change. Even if the exact class is not known for unlabeled data, it is utilized that the prediction with the data augmentation must be the same.

Another method of SSL is pseudo labeling, which is mainly covered in this study. Pseudo labeling improves the performance of the model by leveraging predictions about unlabeled Data. However, a confirmation bias in which an initial error is enlarged and reproduced may occur. To solve this problem, pseudo labeling is limitedly used for model fine-tuning and then removed, or using it when reliability of classification is above a certain level by setting a threshold. Confirmation bias is intensified when label data is insufficient to learn the model. The performance of the model in the early stages of learning with less label data is not good. Therefore, the result of pseudo-labeling is different from the actual one, and may be different from the class ratio in the actual dataset. To prevent this, the prior probability of dataset's class is introduced as a regulatory term, but it is unrealistic to know dataset's prior class ratio information as a regulatory term.

2-2) Active learning

Active learning (AL) is also a method of reducing the labeling cost. AL selects unlabel data with high information value that will bring similar performance as when the entire dataset is labeled. There are three categories of AL. Membership query synthesis generates data to request labeling. Stream-based Selective sampling determines whether a label is required for the new sample data. Finally, pool-based sampling selects important data from a given Dataset. Most AL studies focus on pool-based sampling studies, and this study also focuses on pool-based sampling.

In Pool-based sampling AL, unlabeled data is selected according to the acquisition strategy. There are uncertainty-based approach, expected-based approach, diversity-based approach. The uncertainty-based approach measures the uncertainty of each data according to the bayesian network model. After that, data is sampled and labeled in the order that the uncertainty score is high. Considering that data with high uncertainty are placed around the boundary area of each class, sampled data are similar. In other words, it is helpful to learn the boundary area for each class, but it does not reflect dataset's diversity. The expect-based approach selects data that is expected to improve model performance the most. However, it is difficult to apply it when performance of the model is bad. This point includes uncertainty-based approach. Finally, diversity-based approach is a method of screening data that can guarantee the diversity of a dataset. A typical example is the corset selection.

Coreset Selection samples data that maximizes the expected value of model performance. This is the same as sampling data that makes it have a minimum radius δ when constructing subgraphs that can cover the entire data with a given sampling size N{Sener, 2017 #4}. Each sampling data is widely spread regardless of the density of the dataset. For this reason, sampling data does not include only values of a specific class, but reflects the overall dataset. Moreover, it does not rely on the performance of the SSL model by sampling based on the distance. Therefore, the same sampling performance can be achieved even when label data is insufficient.

**3. Method**

3-1) Classification

We will pseudo-labeling through reliable classification without relying on the performance of the model. Due to the lack of label data, the performance of the model is unreliable. Therefore, in this study, a reliable classification is conducted by using a subgraph formed through coreset selection. Let Dataset . Let Sampling size is N, the i-th sampled data is . Each is the central point of subgraph . minimizing the radius of subgraph is sampled as Radius is . The density of subgraph would be measured by the number of data included in the corresponding subgraph. By assuming that the input values x1, x2 in the region with high probability density are close and the respective associated Labely1, y2 are also close, s.t. j ∈ {1,…, | } will have the same class as in high probability. That is, could represent unlabeled data belonging to a subgraph .

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 , 들은 Labeling 하여 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 자연수로 hyperparameter이다.

본 방법은 Classification 방법은 라벨 클래스의 수가 적거나 Unlabeled data가 많아 특정 클래스의 분포가 명확히 구분할 때 부각된다. 충분히 작은 반지름 , 클래스별 데이터 밀집이 보장된다면 각 클래스별로 중심부에 해당하는 데이터들을 한번에 높은 신뢰도로 Pseudo labeling을 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.



[Figure 1. Subgraph 간 기하적 관계]

Coreset selection을 다회차 진행하여 subgraph들이 중첩되었을 때 또 다른 방식으로 Pseudo labeling이 가능하다. 각 Coreset selection 간 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 높아질 것이다. 또한 각 subgraph들의 반지름 에 반비례하여 확률이 높아질 것이다. p번째 Coreset selection 간 각 클래스 별로 subgraph에 속하는 횟수 와, 반지름을 기반으로 softmax를 통해 확률화한다.

(3.1)

계산상 효율을 높이기 위해 Pseudo labeling되지 않은 Unlabel Dataset 에 한해 진행한다. 데이터의 클래스를 가장 높은 확률 값으로 고정할 경우 Confirmation bias가 발생할 수 있다. 따라서 Confirmation bias를 방지하기 위해 Mix up{Zhang, 2017 #8}을 적용하여 soft labeling한다. 한편으로 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]

4-2) 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 조건을 만족하는 경우가 줄 것이다. 더불어 Minor Class 데이터가 분포한 영역이 한 subgraph에 포함되는 등 오차가 커질 것이다.

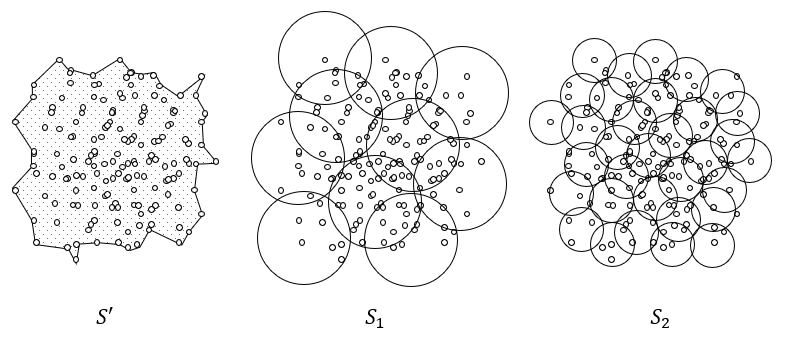
적절한 Sampling size n을 고를 가이드라인이 필요하다. Labeling Cost와 classification 정확도 사이에 Trade off 관계를 띈다. 반지름 의 크기는 Sampling size 에 달려 있다. 즉, N의 크기가 작을 땐 Labelling Cost는 줄어들지만 이 커져 신뢰도 높은 Classification을 할 수 없다. 반대로 N의 크기가 크다면 이 줄어들어 subgraph 간 간격이 줄어들어 높은 정확도의 Classification이 가능하다. 그럼에도 Pseudo labeling 간 가장 피해야할 문제는 Confirmation bias라 판단한다. 따라서 Labelling cost가 있음에도, Confirmation bias을 피하는 것에 우선순위를 둔다. 각 subgraph의 속하는 데이터의 개수가 일정 이상 넘지 않도록 규제함으로써 Confirmation bias를 방지하겠다.

모든 subgraph에 속한 데이터 수를 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 Bound를 형성할 수 있다. 단, Outlier가 데이터 분포에서 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. 따라서 sample size, 반지름 , 그리고 각 중점 간의 거리를 구할 수 있다면 넓이 S를 구할 수 있다. 각 subgraph 중점 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 찾는 과정을 진행한다. 임의의 Sample size 을 Coreset selection 알고리즘에 input으로 넣어 의 값을 얻을 수 있다. 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. 본래의 Dataset의 넓이를 S’ 라고 할 때, 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]

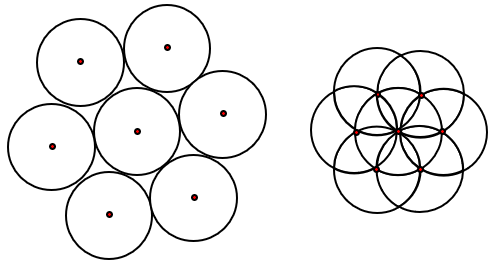
또한 , 에서 을 증명해줄 수 있다면 N’에 다른 반대 방향의 부등식을 만들 수 있다. 이에 대해 이론적 또는 경험적으로 성립하는지 확인하겠다.

, , 에 대해 계산하겠다. 은 이미 Sample size n’와 반지름 이 주어졌으니 각 subgraph들이 모두 겹치는 부분이 없을 때이다. 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 쉽게 생각할 수 있다. 이로써 각각의 넓이를 구할 수 있으며, Lemma 2에서 밝히도록 하겠다.

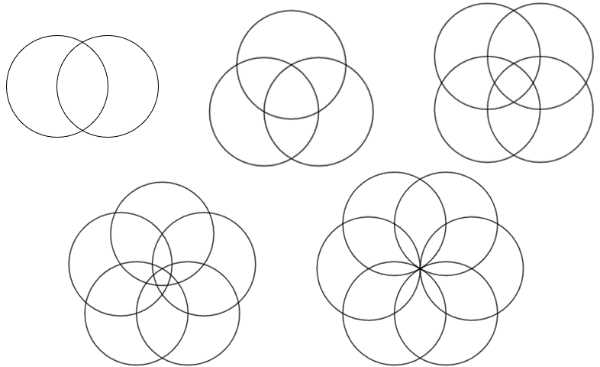
[Figure 4. ), ) 예상도]

**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 4가지 다. 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.

 최대한 원들이 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.

**5. 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 신뢰도 보장을 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*.