Solve Lack of Label Data : Partial Classification by Geometric Attribute of Coreset selection

Hyeongu Kang
Knowledge Innovation Research, GSDS, KAIST

Research Purpose

Due to the labeling cost, it is difficult to secure enough label data to learn the model.

Problem 1

Most of pseudo labeling rely on the performance of the NN model. Or it is unrealistic, they require prior knowledge Of the dataset.

Problem 2

Lack of label data cannot guarantee the performance of neural network models.

Even a way to solve the labeling cost cannot be expected if label data is insufficient. Therefore, we need new a highly reliable pseudo labeling method that does not require a learning process through label data and can be applied without prior knowledge of the dataset is needed.

Research Question

RQ1. How high-accuracy pseudo labeling without NN model and prior knowledge?

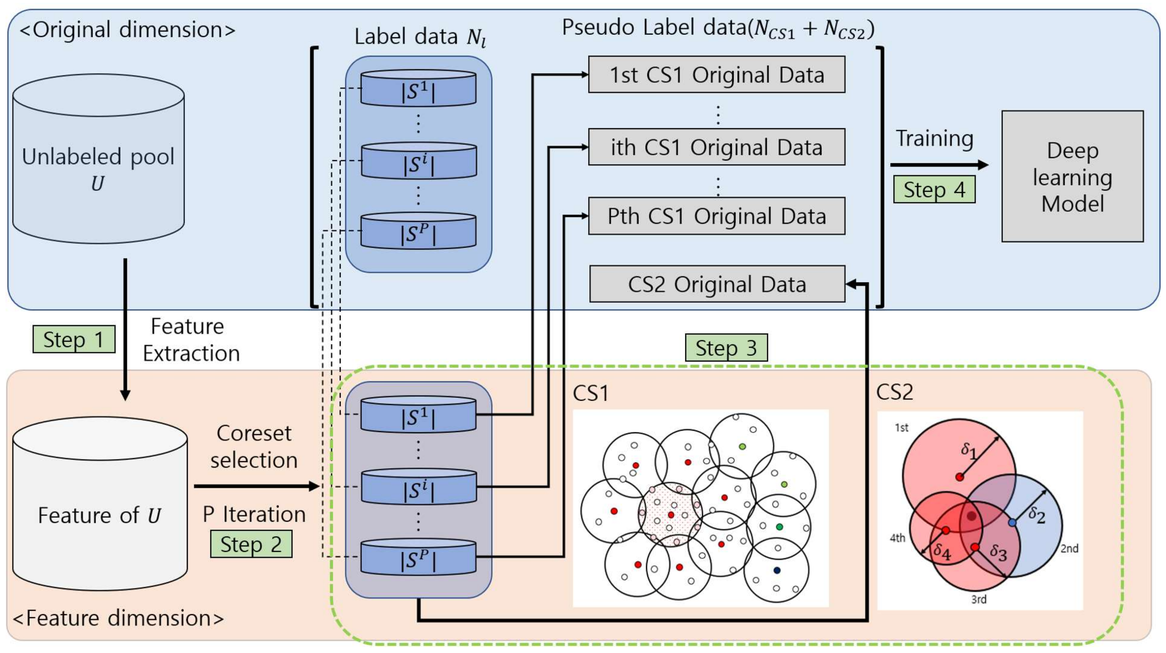
A. Propose a new classification method utilizing Coreset-selection.

- Each sampled data through the Coreset selection could be representative of the unlabeled data from a geometric perspective. Using this, we will perform high-accuracy classification for a large number of unlabeled data.

RQ2. How prevent confirmation bias of pseudo labeling?

A1. Prevention of over-trust in DL Model : Soft labeling / Mix-up data augmentation
A2. Ensuring high-accuracy of classification: Hyper-parameter setting

Method



Step 01 - Feature extraction

Convolution autoencoder is applied to reduce image data to low-dimensional features

- AL has difficult in applying to high-dimensional data such as Image data(Tong, 2001). DL with feature extraction capability can address limitation of AL(Ren et al., 2021)

Assume that CAE not only dimension reduce but also cluster data in low dimension

- Auto encoder can capture the repetitive structure of data with dimension reduction(Y. Wang, Yao, & Zhao, 2016). In addition, when classification was performed in the k-nearest neighbor method after feature extraction, CAE was one of most accurate methods with an accuracy of 85%(Hurtik, Molek, & Perfilieva, 2020).

Step 02 - Coreset selection

Coreset selection is performed through a low-dimensional features

- When forming a subgraph that covers all data with a given sampling size N_l, we sample the data u_i that makes it have a minimum radius δ(Sener & Savarese, 2017).
- Through the smoothness assumption of SSL(Chapelle, 2009), the labeled center u_i of each subgraph can represent data belonging to the subgraph if the radius δ is small and subgraph G_i is dense.

Step 03 - CS1 & CS2

CS1 : If the subgraph G_i is at the center of a specific class, the class of u_i would be the same as the unlabeled data x belonging to the subgraph G_i with a high probability. When $u_i = u_{i1} = \cdots = u_{im_i}$ (m_i : num of encountered subgraph) and density of subgraph $G_i > \alpha$ is satisfied, G_i would belong to the center of a specific class.

CS2 : Probability of unlabel data x in a particular class is proportional to the number of times it belongs to the subgraph of the class, and will be inversely proportional to the radius δ of each subgraph. Consider ratio of radius δ and iteration-specific weights.

$$p(y = c_j | x_k^p) = \frac{\sum_{p=1}^{P} e^{s_{c_j}^p / \delta^{p\prime}}}{\sum_{j=1}^{J} \sum_{p=1}^{P} e^{s_{c_j}^p / \delta^{p\prime}}} \text{ s.t. } \delta^{p\prime} = \frac{\delta^p}{\min\{\delta^p\}} * \frac{\mathrm{i}}{P} \;\&\; \mathrm{k} \in \{1 \ldots, \mathrm{K}^p\}$$

Step 04 - Training

Train DL model with methods to help prevent confirmation bias

- Soft labeling / Warm training
- Mix-up data augmentation & small batch size setting(Arazo, 2020)

Experiment

01 Datasets and training parameter

- Dataset : MNIST, Fashion-MNIST, EMNIST-letter, CIFAR 10/100
- Lr = 0.001 / Optimizer = Adam / Mix-up augmentation $\alpha' = 4$
- Training epoch / batch size : 10, 5, 5 / 8, 100, 100 *(label / CS1 / CS2 pseudo label)
- Model : 13-CNN-based CAE(representation learning) / 13-CNN (DL model)

02 Performance of CS1 & CS2

[Table 1]

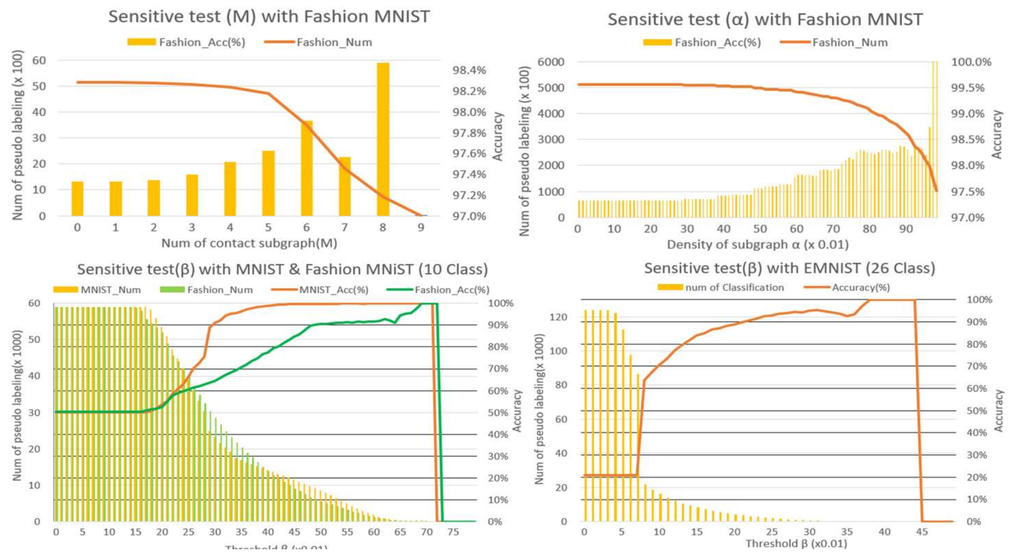
Dataset	N_l ($\|S^i\|$, P)	CS1 Only (N_{CS1}, Acc)		CS2 Only (N_{CS2}, Acc)	
MNIST	1000 (1000, 1)	11992(x12)	99.59(%)	X	X
(1 x 28 x 28)	1000 (100, 10)	15305(x15)	99.97(%)	14237(x14)	94.19(%)
FashionMNIST	1000 (1000, 1)	4217(x4)	96.84(%)	X	X
(1 x 28 x 28)	1000 (100, 10)	6492(x6)	95.56(%)	5641(x5)	90.48(%)
EMNIST-Letter	1000 (1000, 1)	4697(x4)	93.52(%)	X	X
(1 x 28 x 28)	1000 (100, 10)	7732(x7)	95.16(%)	0	0(%)
CIFAR10	1000 (1000, 1)	15(x0.01)	66.67(%)	X	X
(3 x 32 x 32)	1000 (100, 10)	46(x0.05)	73.91(%)	2250(x2)	20.56(%)
CIFAR100	1000 (1000, 1)	9(x0.01)	33.33(%)	X	X
(3 x 32 x 32)	1000 (100, 10)	51(x0.05)	76.46(%)	0	0(%)

[Table 2] * MNIST dataset

N_l ($\|S^i\|$, P)	CS1 Only (N_{CS1}, Acc)		CS2 Only (N_{CS2}, Acc)		Both ($N_{CS1} + N_{CS2}$, Acc)	
50 (50,1)	738(x14)	99.04(%)	X	X	X	X
100 (100, 1)	3886(x38)	99.84(%)	X	X	X	X
250 (250, 1)	4024(x16)	98.03(%)	X	X	X	X
500 (500, 1)	9354(x18)	99.66(%)	X	X	X	X
750 (750, 1)	10430(x14)	99.75(%)	X	X	X	X
1000 (1000, 1)	11992(x12)	99.59(%)	X	X	X	X
100 (10, 10)	4505(x45)	100(%)	4268(x42)	99.5(%)	4971(+466)	99.02(%)
250 (25, 10)	8874(x35)	99.92(%)	8244(x33)	99.68(%)	9961(+1087)	99.56(%)
500 (50, 10)	11486(x22)	99.85(%)	10672(x21)	98.2(%)	12200(+714)	98.97(%)
750 (75, 10)	13012(x17)	99.84(%)	13810(x18)	99.17(%)	13012(+260)	99.84(%)
1000 (100, 10)	15305(x15)	99.78(%)	14237(x14)	94.19(%)	15305(+252)	99.47(%)

- CS1 and CS2 for small image datasets have an accuracy of 95% or more for unlabeled train data 10 to 40 times that of a given sampling size N_l. Due to the limitations of CAE's performance, it is not applicable in high-resolution images.
- The accuracy and CS2 and N_{CS1} increases when by dividing sampling size N_l it into P iteractions.

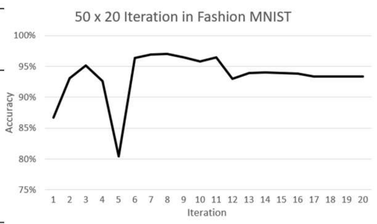
03 Sensitive test for hyperparameter M, α, β



- Introducing M, α of CS1 improved accuracy but decreased N_{CS1} and N_{CS2}. N_{CS1} and N_{CS2} are rapidly reduced compared to improved accuracy.
- β of CS2 appears more conservative than the actual probability of class. The appropriate β changes according to the dataset and the number of Iterations.

04 Limitation

N_l ($\|S^i\|$, P)	CS1 N_{CS1}	CS2 + N_{CS2}	Total Acc	CNN Acc (N_l only)	CNN Acc (+N_{CS1})	CNN Acc (+N_{CS2})
100 (10, 10)	4505	466	99.02	49.84	32.87	31.65
250 (25, 10)	8874	1087	99.56	54.45	39.23	44.52
500 (50, 10)	11486	714	98.97	68.26	52.35	58.03
750 (75, 10)	13012	260	99.84	63.03	54.77	55.42
1000 (100, 10)	15305	252	99.47	67.60	58.17	53.12



When pseudo label data with 99% accuracy is applied to learning, performance down.

- Similar data are not effective to training DL model.
- A pseudo labeled data is imbalanced. Imbalanced labeled data causes overfitting, not generalization of the model.
- Misclassified data would have a significant adverse effect on model learning.

When the sampling size is small, the accuracy may suddenly drop.

- There may be cases in which the central class of the subgraph that encountered is the same even though it is not representative for subgraph.
- Fundamental problem is the lack of performance of the representation learning. This case can be prevented by increasing the sampling size and assigning M conditions.

Conclusion and future work

Classification can be performed with 95% accuracy for unlabeled data, which is several times the given label data size N_l. When suitable representation learning for datasets is applied, the efficiency is 10 to 40 times higher, and the accuracy is also close to 99%. It could be applied test dataset as the classification method .

Future work

01 Various types of representation learning methods such as ResNet 18 and DGI will be applied.

01 Apply data augmentation about unclassified class of CS1 & CS2 for solving imbalance of pseudo-labeled data

01 Introduce threshold on loss value for mislabeled data detection

01 Improve CS2 method and generalize the β by maximizing the number of Iterations. The dimension of feature can be increased to 3 and 4 dimensions.