**Solve lack of label : Partial classification by geometric attribute of coreset selection**

GSDS 석사과정 20224314

강현구

1. **Abstract**

The performance of the neural network model cannot be guaranteed if label data is insufficient. There are studies that have been solved without relying on the neural network model in the lack of label data, but there are unrealistic assumptions that they already know about prior knowledge of the ratio by class. In this study, pseudo-labelling the unlabeled data through high-accuracy classification that does not require a learning process. We propose a new classification method using coreset selection’s geometric attribute. This method classified 10 to 40 times unlabeled data with 95% or more accuracy by labeled data for small image datasets.

**1. Introduction**

Recently, deep learning (DL) model has been achieving results in various fields based on a large amount of labeled data. However, as the DL model increases in the data required for learning, how to solve the labeling cost has become an important topic. Semi-supervised learning (SSL) is one of the ways to solve labeling cost, which assumes insufficient label data and many unlabeled data situations. This study focuses on image classification through SSL.

SSL is largely divided into consistency regularization and pseudo labeling(Yang, Song, King, & Xu, 2022).

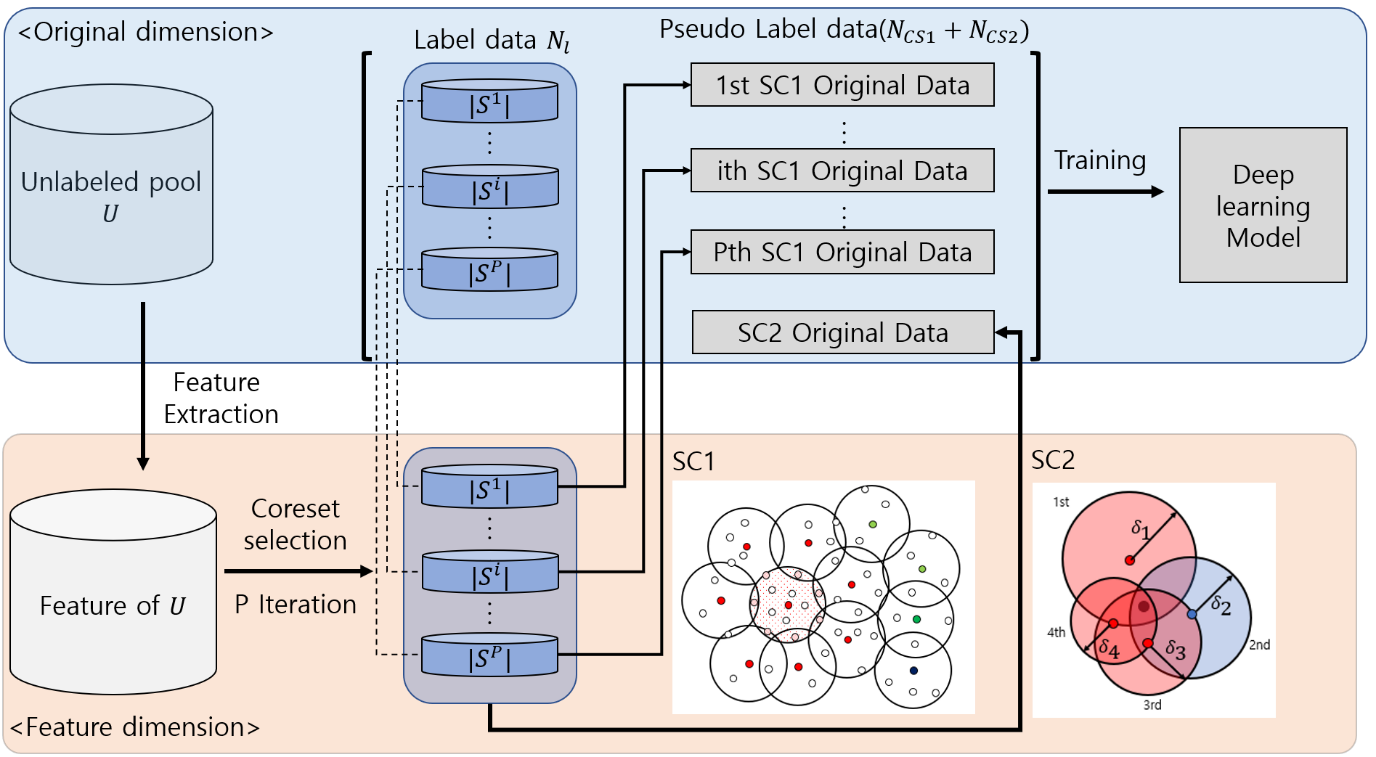
In this paper, the pseudo labeling method is focused. Consistency regularization is preferred due to its high performance in recent image classification studies. However, both are necessary because dataset's diversity characteristic is different that can be obtained through each method. Consistency regularization has limitations in directly utilizing unlabeled data, even though there are various augmentation method of label data. On the other hand, in the case of pseudo labeling, there is a confirmation bias, but the information of the unlabeled data can be directly used. In other words, it is not a contradictory method, but a parallel method. The potential of pseudo labeling can be seen through a study(Arazo, Ortego, Albert, O’Connor, & McGuinness, 2020) that solved confirmation bias well and performed better than consistency regularization.

In pseudo labeling, it is important to resolve the confirmation bias well. Confirmation bias means that the incorrect prediction value for the unlabeled sample in the process of learning the model increases the performance deterioration of the neural network model. Existing studies introduce a threshold setting for reliability for the prediction of DL models(Cascante-Bonilla, Tan, Qi, & Ordonez, 2021), and a regulatory term for soft labeling and initial convergence of model.(Zhang, Cisse, Dauphin, & Lopez-Paz, 2017) Most measures assume that the neural network has sufficient accuracy. However, if the label data is insufficient, the accuracy of the neural network model is greatly reduced, making it difficult to trust. As an example, when 13 CNNs are applied to MNIST datasets, an accuracy of model is just 30% when 100 Label data is given. Measures to prevent confirmation bias are also meaningless when the performance of the model itself cannot be guaranteed. One study successfully prevents confirmation bias by using class-specific dictionary ratios as regulatory terms by assuming that the dataset will be class-balance dataset (Arazo et al., 2020). However, it is difficult to expect to know whether the dataset is class-balance, or prior knowledge of class distribution. Therefore, more realistic measures are needed.

This study proposes a classification method that does not require a learning process through the coreset-selection method, which is part of active learning. Active learning is a method of sampling data that is useful for learning a model. Coreset selection does not rely on neural network-based models by utilizing distance information from data. Furthermore, it will be shown that each sampled data through the coreset selection is representative of the unlabeled data from a geometric perspective. Additionally, classification with high reliability will be conducted based on geometric relationships. At this time, the unsupervised dimension reduction method will be applied to simplify the geometric relationship of the subgraph and because active learning is difficult to apply to high-dimensional data. In this study, the convolution autoencoder (CAE) is applied to reflect the structural information of image data. The whole process is shown in Figure 1. First, feature extraction is performed with CAE for each image data. Coreset selection is performed through the feature, and classification will be performed through the geometric relationship of subgraphs. After that, the DL model will be trained through data from coreset selection and new classification. In other words, the problem of label data shortage is solved by increasing the label data to be used for learning through classification of unlabeled data with high accuracy.

**2. Related works**

2-1). Pseudo labeling

Pseudo labeling utilizes predictions about unlabeled data for model learning. In the early stage when confirmation bias was not solved, pseudo labeling was limited to fine-tuning of the model (Lee, 2013). Since then, various measures have been taken to solve the confirmation bias, such as granting uncertainty weight(Shi, Gong, Ding, Tao, & Zheng, 2018), and soft labeling. There is a limitation that each measure relies on a neural network-based model that requires learning. The 

[Figure 1. Whole process]

performance of the model is unreliable when label data is insufficient. Existing methods are meaningful when the model is sufficiently learned and performance is guaranteed.

There is a way to resolve the confirmation bias without relying on the neural network model. In addition to soft labeling, Arazo's study effectively prevented confirmation bias by using mix-up data augmentation, minimum batch, and utilize dataset's class ratio information by regularization term(Arazo et al., 2020). This study performed better than consistency regulation on CIFAR 10/100, SHVN datasets. However, there are limitations in Arazo's research. In this study, by assuming a class-balance scenario, the class ratio information of the dataset is applied as a regulatory term. In other words, there is a limitation in that it cannot be applied if prior knowledge of the class ratio of the dataset cannot be obtained

2-2) Active learning

Active learning(Tong, 2001)(AL) is also a method of reducing the labeling cost. AL samples the most useful data from the unlabeled dataset to reduce the labeling cost as much as possible while maintaining performance. There are three main types of AL(Ren et al., 2021). This study focuses on pool-based sampling that selects important data from a given dataset. Pool-based sampling AL selects data according to the acquisition strategy. Acquisition strategy is divided by uncertainty-based, expected-based and diversity-based approach Uncertainty-based approaches measure uncertainty for each data according to the neural network model(Gal, Islam, & Ghahramani, 2017; K. Wang, Zhang, Li, Zhang, & Lin, 2016). The expect-based approach also selects the data that is expected to improve the model performance the most (Sener & Savarese, 2017), and relies on the neural network model. Finally, diversity-based approach is a method of screening data that can guarantee the diversity of a dataset. Representatively, there is a coreset selection(Sener & Savarese, 2017). Coreset Selection showed that selecting subgraphs with minimum radius covering all data is in the end most useful for model learning. While most ALs rely on the performance of neural network models, corset selection relies only on distance information between data. In this study, a subgraph formed between coreset selection will be used.

AL has a limitation that it is difficult to apply to high-dimensional data, and does not have the ability to calculate low-dimensional features from high-dimensional data(Tong, 2001). There are studies that apply low-dimensional feature extraction through DL to address this limitation. DL models can be easily applied to high-dimensional data and feature engineering without human intervention in the data. In the case of DBL, high-dimensional image data is derived from an autoencoder with a low-dimensional feature, and then sampling data which is representing dataset and or have uncertainty (Liu, Zhang, & Eom, 2016). In the case of CEAL, active learning is performed on the unreliable data of CNN's model, and pseudo-labeling is performed only when the reliability is above a certain level(K. Wang et al., 2016). For BCNNs, the bayesian network is utilized to calculate uncertainty and use it as an acquisition strategy(Gal et al., 2017). However, the above studies become unreliable as performance of the neural network model unreliable. Acquisition strategy based on a neural network model that has not been sufficiently learned is not reliable.

On the other hand, there are studies that extract the feature of high-dimensional data through the unsupervised deep learning method and then apply AL. Prior works mainly extract low-dimensional image data through autoencoder(Liu et al., 2016; Sun, Li, Wang, Plaza, & Chen, 2016). VAL samples data focusing on diversity in latent space through a variational autoencoder(Sinha, Ebrahimi, & Darrell, 2019). This method can produce its original performance even if label data is insufficient due to unsupervised learning. Therefore, in this study, we will also apply feature extraction through unsupervised learning to apply high-dimensional image data.

**3. Method**

**3.1) Model structure**

Notation refers to Arazo’s research (Arazo et al., 2020). Let Dataset D as unlabeled set and labeled set s.t. . The two new classification methods presented in this study will be called CS (Coreset selection-based classification) 1 and CS2, respectively. CS1 performs classification in the form of a hard label by utilizing the geometric relationship between subgraphs in contact. CS2 calculates a class- probability vector according to the degree of overlap of subgraphs for each data when performing multiple iterations of the coreset selection. A specific method will be described in detail later.

To prevent Confirmation bias, we apply Mix-up data augmentation and setting a minimum number of labeled samples per mini-batch. Mix-up data augmentation is a powerful regulatory method that combines data augmentation and label smoothing (Zhang et al., 2017). Confirmation bias is well prevented when Mix-up data augmentation and setting a minimum number of labeled samples for mini-batch k simultaneously(Arazo et al., 2020). We will make a difference in the number of epoch and batch size between label data and pseudo labeled data by CS1 and CS2 during CNN model learning.

**3-2) Coreset selection**

Coreset selection samples data that maximizes the expected model performance. This is the same as sampling data that makes it have a minimum radius when constructing subgraphs that can cover the entire data with a given sampling size (Sener & Savarese, 2017). Each sampling data is widely spread, regardless of the density of the dataset. For this reason, sampling point not only includes a specific class, but also reflects an overall dataset. It also samples based on the distance between the data. Therefore, the same sampling performance can be achieved even when label data is insufficient.

AL has difficulty expanding to high-dimensional data such as images and text(Tong, 2001). For this reason, most studies on AL focus on low-dimensional problems or utilize features of pre-extracted high-dimensional data. To apply AL in the latter case, since AL does not have the ability to extract features of data, so it is necessary to apply additional measures. Accordingly, synergy can be improved by using DL and AL together, which have the ability to process high-dimensional data and perform feature extraction without human intervention(Ren et al., 2021).

**3-3) Dimension reduction**

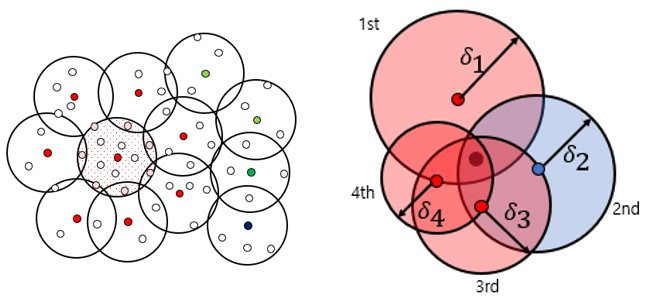
Dimension reduction through feature extraction is required to apply coreset selection to high-dimensional image data. The CS1 method requires two-dimensional or three-dimensional features in that it utilizes the adjacency between subgraphs. CS1 can be applied when the central classes of the subgraphs encountered are all the same. The detailed process will be described later. The number of being able to classified have a trade-off relationship with accuracy of CS1. The higher the number of subgraphs encountered, the stronger the conditions and the higher the precision. On the other hand, decreases as the conditions are strengthened. The dimension of the data controls the number of subgraphs encountered. In the case of two dimensions, the number of subgraphs available to contact for each subgraph is 7 to 8 at most, but in the case of three dimensions, the number increases to 20 to 30. As the dimension increases, the number of subgraphs encountered increases exponentially. Even if MNIST data of 784 dimensions is reduced to 10 dimensions, it is still high. In this study, the dimension of the feature is reduced to two dimensions.

In this study, convolution autoencoder (CAE) is used to reduce dimensions. In this study, it is assumed that when dimension of data reduced through CAE, the information for each class will be well-preserved and clustered for each class. Autoencoder not only capture the repetitive structure but also play the role of dimension reduction(Y. Wang, Yao, & Zhao, 2016). In addition, when classification was performed by the k-nearest neighbor method using the unsupervised dimension reduction methods, CAE was one of the most accurate methods with an accuracy of about 85 percent(Hurtik, Molek, & Perfilieva, 2020). From this, it can be inferred that there is an effect of clustering similar classes of data when dimension reduction is performed through CAE.

**3-4) Coreset selection based classification (CS1 & CS2)**

Extract the two-dimensional feature using CAE from the given dataset. After that, the coreset selection is performed based on the distance between features. At this time, the euclian distance is applied. labeled data are the results of sampling through coreset selection. When a total of p coreset selections are performed, is called a data set sampled between p-th iterations, and be each sampling point. In this case, is satisfied. Let the radius of subgraph centered on each be . Coreset selection is the same as K-centers algorithm. It samples the unlabeled data as , which minimizes the radius of subgraph . Thus, is . The density of each subgraph would be measured by the number of data .

The CS1 method is classified through a geometric relationship between subgraphs. For convenience, we will simplify the notation to . SSL assumes that the data x1, x2 in the probability-dense region are close, then the associated Label y1, y2 are similar (Chapelle, Scholkopf, & Zien, 2009). Data s.t. j ∈ {1,…, | } belonging to the dense subgraph can be assumed that label is same as ’s label. Thus, could represent unlabeled data belonging to the subgraph. Furthermore, it is assumed that subgraphs formed through active selection have a radius small enough to cover the dataset tightly. Let s.t. k ∈ {1,…, | } be the subgraph which has connection with . We know the classes of the central and of each subgraph. At this time, if the classes of are different, it can be inferred that and are in the boundary area of different classes from a geometric point of view. Conversely, it will be located in the center of a particular class when all the classes of and are the same. That is, when and density of subgraph is satisfied, the unlabeled data belonging to subgraph can be classified as same class as . is the hyperparameter for the subgraph density. In addition, when a few subgraphs overlap, it is necessary to prevent accidental misclassification by having

[Figure 2. Concept of SC1(Left), SC2(right)]

different classes even though it is not the center of a specific class. The number of of subgraphs which contact to other subgraphs is also added as a condition to be greater than or equal to the hyperparameter M.

CS2 can be applied when subgraphs overlap by performing the coreset selection multiple times. All unlabeled data between one coreset selection are included in one or more subgraphs. If the coreset selection is performed a total of P times, all belong to at least P subgraphs. The more belongs to a subgraph whose center is a specific class, the higher the probability that it is the same class. In addition, the probability will be high in inverse proportion to the radius of each subgraph. For each class, we calculate probability of class via SoftMax based on the number of times and. In addition, the length of each compared to the smallest radius in p cycles is used to prevent the size of the radius from being very large and the probability value of each become low. In addition, to give weight to the recent iteration, we will add the weight of for the i-th Iteration. Hard label is given only when the most likely value is greater than threshold

(3.1)

4. Experiment

4-1) Dataset and training

We will verify the performance of CS1 and CS2 with five datasets: MNIST, Fashion-MNIST, EMNIST-letter, and CIFAR10/100. Each dataset has 10 classes, except EMNIST-letter and CIFAR 100. EMNIST-letter and CIFAR 100 have 26 and 100 classes, respectively. MNIST and Fashion-MNIST consist of a total of 60K training images and 10K test images, respectively. EMNIST-letter consists of 124.8K and 20.8K. Images of three datasets have a resolution of 28x28. CIFAR 10/100 consists of 50K and 50K training images and 10K and 10K test data, respectively. Image of dataset are color images and have a resolution of 32x32. The case of sampling at once for =1K and the case of sampling 10 times for 100 are tested. The result of each dataset may partially vary depending on the performance of CAE, so the number of detailed cases will be checked only with MNIST. The CAE model utilizes a well-known 13-CNN model structure, and the decoder part is configured in reverse order. The DL model is 13 CNN.

The same hyperparameters were applied for each dataset. In the case of the convolution autoencoder, lr = 0.001 and Adam is applied as an optimizer to learn train dataset for 100 epochs. After that, in learning the 13-CNN model, label dataset and pseudo labeled dataset by CS1 and CS2 were separately learned. Between 13-CNN model learning, we applied the same hyperparameters as when learning CAE. According to the research of {Arazo, 2020 #9}, the value of hyperparameter of mix-up data augmentation was set to 4. In addition, the train batch size of label dataset was set to 8 and the train batch size of SC1 and SC2 was set to 100, respectively. In addition, for warm training for each, label data learned 10 epochs, and then 5 epochs were learned through each SC1 and SC2 label data.

4-2) Performance of CS1 & CS2

The performance of CS1 and CS2 for each dataset is shown in Table 1. There are no subgraph restrictions in CS1 with density, and the threshold of CS2 was set to 0.5.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | (||, P) | CS1 Only ( | | SC2 Only ( | |
| MNIST  (1 x 28 x 28) | 1000 (1000, 1) | 11992(x12) | 99.59(%) | X | X |
| 1000 (100, 10) | 15305(x15) | 99.97(%) | 14237(x14) | 94.19(%) |
| FashionMNIST  (1 x 28 x 28) | 1000 (1000, 1) | 4217(x4) | 96.84(%) | X | X |
| 1000 (100, 10) | 6492(x6) | 95.56(%) | 5641(x5) | 90.48(%) |
| EMNIST-Letter  (1 x 28 x 28) | 1000 (1000, 1) | 4697(x4) | 93.52(%) | X | X |
| 1000 (100, 10) | 7732(x7) | 95.16(%) | 0 | 0(%) |
| CIFAR10  (3 x 32 x 32) | 1000 (1000, 1) | 15(x0.01) | 66.67(%) | X | X |
| 1000 (100, 10) | 46(x0.05) | 73.91(%) | 2250(x2) | 20.56(%) |
| CIFAR100  (3 x 32 x 32) | 1000 (1000, 1) | 9(x0.01) | 33.33(%) | X | X |
| 1000 (100, 10) | 51(x0.05) | 76.46(%) | 0 | 0(%) |

[Table1. Performance of CS1 and CS2 for each datasets]

CS1 has at least accuracy of 95% in MNIST and Fashion EMNIST, which are low-resolution datasets. On the other hand, in the case of CS2, performance is not good except for MNIST and Fashion MNIST datasets. At this time, the accuracy and of the CS2 method change depending on how Threshold is set. Details will be covered in 4-3). However, it can be seen that performance deteriorates significantly on the high-resolution dataset, CIFAR 10/100. This is considered to have lost class information during dimension reduction due to limitations in the performance of the CAE model for high-resolution images. In other words, MNIST, Fashion-MNIST, and EMNIST-Digit with low resolution contained class information even when two-dimensional features were extracted through 13-CNN-based CAE, but not for CIFAR 10 and CIFAR 100. CIFAR 10 /100 needs to be reviewed for ResNet 18 or another dimension reduction and clustering application.

We can check the relationship between various sampling sizes and sampling and Iteration when we have sufficient CAE performance. Considering that the performance of CS1 and CS2 is currently good in MNIST Dataset, it is considered to satisfy the assumption. At this time, the threshold of CS2 was set to 0.5.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| (||, P) | CS1 Only ( | | SC2 Only ( | | Both ( | |
| 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(%) |

[Table 2. Performance of CS1 and CS2 in MNIST]

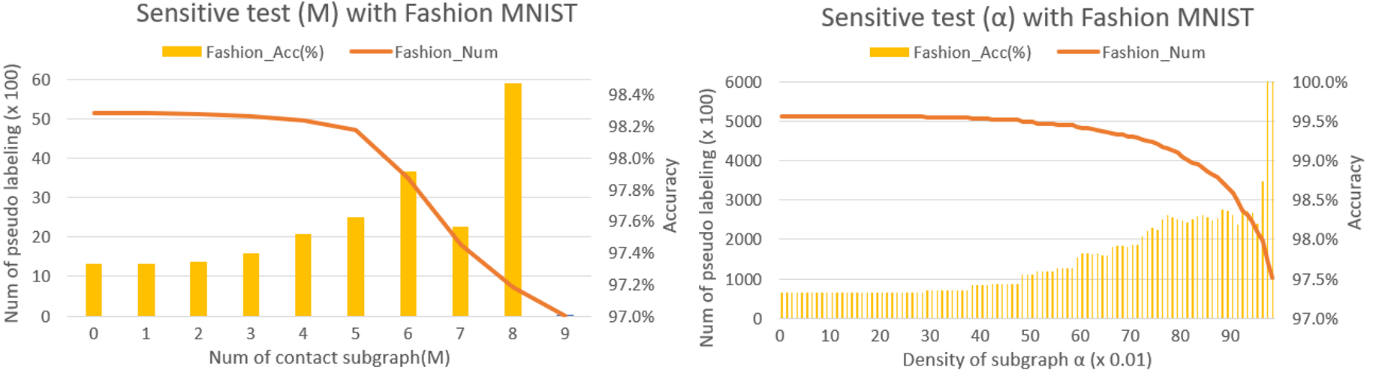
The performance of CS1 and CS2 is similar. In most cases, the accuracy reaches 99%. In terms of accuracy, the size of and whether Iteration exist or not does not affect significantly. Both CS1 and CS2 methods can be classified for unlabeled train data 10 to 40 times that of a given sampling size . After CS1 is applied, CS2 can be additionally applied to the unlabeled data. In this case, as the is increased, the size of the additional is reduced. From this, CS1 and CS2 perform classification for unlabeled data under similar conditions, but some areas are different. As increases, the difference in classification between the two methods decreases. In addition, it can be seen that the efficiency of the CS1 and CS2 methods gradually decreases as increases. This is because classification is impossible even if the sampling size is increased for the boundary or where different class data are overlapped.

increases when dividing as P iterations. This can be explained through the snapshot of coreset selection process. There is no difference between sampling 100 data and dividing it into 10 times and sampling 10 pieces each from the perspective of coreset selection. Only the number of times SC1 is applied for each moment of coreset selection is different. That is, it is the same as taking a snapshot about p times of the relationship of subgraph during a total of coreset selection. The SC2 method can also be applied when there are a sufficient number of iterations. Accordingly, as the number of Iterations increases, more data would be classified.

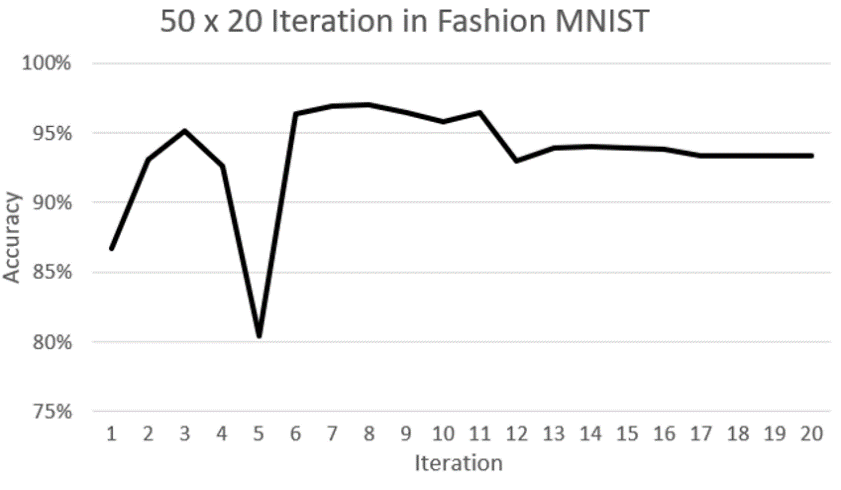
However, the sampling size for each Iteration should not be lowered than a certain amount. Even though it is at the boundary between classes, there would be cases where the classes of subgraphs encountered by chance are all the same. At this time, the accuracy of CS1 and CS2 deteriorates significantly. To prevent this, increase the sampling size so that the number of subgraphs encountered as much as possible. In addition, the larger the sampling size, the smaller the Radius , increasing the probability that only the same class belongs in the subgraph. However, for fixed , it is necessary to appropriately adjust the sampling size with the iteration. In addition, we will check whether it is possible to prevent this case by limiting the number of subgraphs M encountered or limiting the density α of the subgraphs.

4-3) Sensitivity test about hyperparameter

The sensitivity test for the number of subgraphs contacted in SC1, the subgraph density α, and the threshold in SC2 is conducted. The accuracy and the number of classifications according to the value of each hyperparameter in the situation where = 1000. In order to visualize the difference in performance, the Fashion MNIST dataset, which has lower accuracy than MNIST was checked.

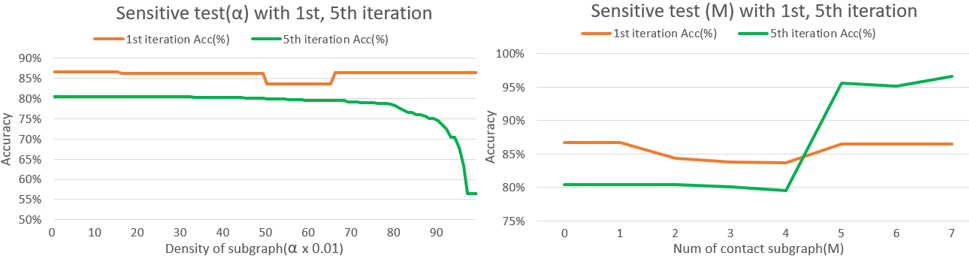


[Figure 3. SC1 Sensitive test when = 1000, = 1]

 When SC1 was applied with = 1000 and P = 1 on the Fashion MNIST dataset without additional constraints, was 5131 and the accuracy reached 97.3%. As M was added as a limiting condition, decreased, and accuracy showed an increasing trend. However, when M=6, reduced by 1,616, but the accuracy increased just 0.4%. The M-limiting condition does not have a significant effect on accuracy compared to the decrease in . Subgraph's Density α is similar. Accuracy improves with the addition of the α condition, but the reduction amount of increases exponentially. Even until =0.8 where is gradually reduced, decreases about 800 in increasing accuracy by 1%.

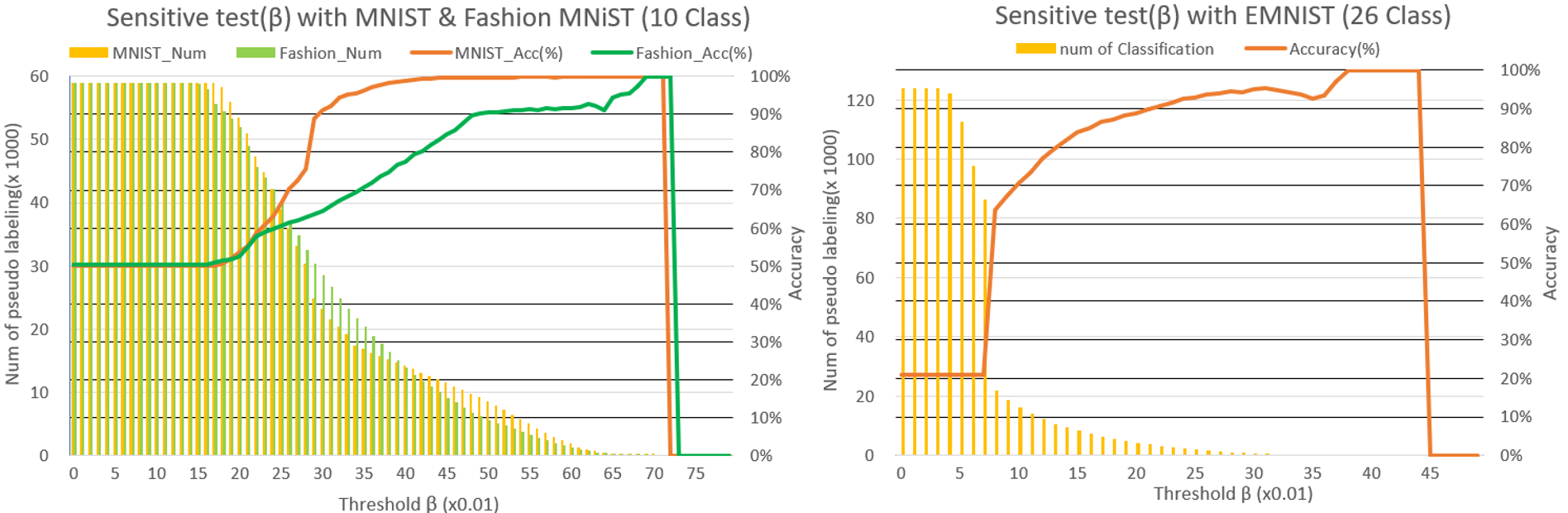
[Figure 4. Accurancy of SC1 when = 1000, = 50, = 20]

Furthermore, the conditions for M and do not completely prevent CS1 errors that occur with low probability. In CS1, when sampling size for each session is set small, there is a case where the central class of the subgraph encountered is the same even though it is not the center of a specific class by chance. Through Figure 4, it can be confirmed that the accuracy between 1st and 5th period suddenly decreases. When the mi α condition is added, the effect depends on the situation. The fundamental problem of the above phenomenon is the insufficient performance of the representation learning method. If clustering, which is completely divided according to each class, is not possible, multiple classes may be mapped to the same low-dimensional area. A place where several class data are aggregated may have a high density even at the boundary. That is, even if the α condition is strengthened as shown in Figure 5, the accuracy may decrease. On the one hand, even if several classes are united, it is difficult to satisfy the conditions of CS1. However, the number of subgraphs encountered is at most eight on two dimensions. That is, there is a possibility that all center of subgraphs ui which is not representing the subgraph's data sampled by chance are the same. The possibility can be reduced by adding the number M of subgraphs encountered as a condition, but it is not a perfect solution. This can be confirmed through the trend of accuracy change for 1st and 5th period in Figure 5. In addition to improving the performance of representation learning, it is also a method to reduce the size of radius δ by increasing the size of |Si|. As the size of δ decreases, the gap between subgraphs decreases, enabling more detailed class distinction. In addition, the number of subgraphs encountered increases, reducing the possibility of overlapping classes of subgraphs due to chance.



[Figure 5. Sensitive test of bad case in CS1]

In addition, sensitive tests were conducted on MNIST, Fashion MNIST, and EMNIST to analyze the limiting condition of threshold in CS2. In all datasets, and accuracy have a tradeoff relationship. In MNIST and Fashion MNIST datasets with the same number of classes, the trend of change in accuracy with according to is similar. Each part painted in red is a point with an accuracy of more than 95%. From this, it can be seen that is stricter than reflecting the probability that each specific value belongs to. However, the appropriate value of varies for each data set. Considering the results of Table 1, it seems to depend on how well the CAE reflects the information of each dataset.



[Figure 6. SC2 Sensitive test in ]

The appropriate is influenced by the number of iterations and the number of classes. As the number of Iterations increases, the probability of the class to which it belongs will increase as the subgraph information of the periphery continues to be reflected. On the other hand, in the case of EMNIST, it can be confirmed that the appropriate decreases significantly as the number of classes increases. This is because classes that have never been counted in obtaining the large probability proportion of SoftMax. For example, in a dataset with a total of 10 classes, there is a data belonging to class 1. When the data belongs to class 1 subgraph only once, the probability of class 1 according to the CS2 method is just 23%. As the number of classes increases, the probability measured by the CS2 method become decrease further. It is necessary to take measures to generalize the for various datasets in the future

4-4) Limitation

The accuracy for test dataset become poor when using pseudo labeled data with higher accuracy than 99% via CS1 and CS2 for learning 13 CNN models on the MNIST.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| (||, P) | CS1 | CS2 | Total Acc | CNN Acc  ( only) | CNN Acc  (+) | CNN Acc  (+) |
| 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 |

[Table 3. Model performance when using pseudo labeled data by CS1, CS2]

There are three main interpretations of this. First, the data classified through the CS1 and CS2 methods are data belonging to the center of the class, and all have similar characteristics. Therefore, using the data for learning does not have a significant impact on the 13-CNN model. Second, the balance between label data is changed. The methods of CS1 and CS2 are effective when the feature extracted by CAE is well distinguished from other classes. In the case of MNIST between this study, the classification was good for class of 0, 4, and 6 data, but in the case of other classes, just a few cases were applicated. More label data but imbalanced causes overfitting rather than generalization of the model. In particular, although not included in this paper, even when CS1 and CS2 had 100% accuracy, their performance was the same or rather degraded. Third, the data of the misclassified minority class has a great adverse effect on model learning. Due to the lack of performance of CAE, the features were similar, but the original data can differ greatly. As the different class is incorrectly classified same as major of pseudo labeling data, it may have a confirmation bias.

**5. Future work**

The biggest limitation is that the CS1 and CS2 methods can be applied only to datasets with low resolution. It is necessary to find other methods to map high dimension data to a low level and to clustering for each class simultaneously. In this paper, we applied to the 13 CNN model. In the future, it is necessary to apply the unsupervised presentation learning method for various image data such as ResNet 18 and DGI. Next problem is that the high-accuracy pseudo-labeled data obtained through the CS1 and CS2 methods does not lead to improved performance of the DL model. In the future, it is necessary to solve imbalance classification about unlabeled data and confirmation bias due to the misclassification. The former may be partially resolved by applying a better representation learning method. If the feature extraction is clustered to be well distinguished for each class, CS1 and CS2 can be applied to various class data. In other way, by assuming that the class ratio of the label data selected through the Coreset selection is similar to the original dataset, the influence of the unbalanced pseudo labeling data can be regularized. In addition, data augmentation may be performed only for classes that the CS1 and CS2 methods have not been applied. Confirmation bias could be solved by applying outlier detection to a small amount of misclassification data. In the case of incorrectly labeled data, it was confirmed that the loss value jumped significantly. Therefore, if the value of loss is limited to below a certain value, misclassification may be prevented.

In addition, there are several ways for improvement of CS2. When the number of iterations was increased for a given , the number of classifiable data increased. It could be expanded to Iterated for every coreset selection. Above all, since a subgraph is formed for each Iteration, each point is included in at least subgraphs. This will allow us to maximize the number of Iterations to generalize the Threshold and ensure great performance for various datasets. In addition, if only the CS2 method is considered, there is no need to lower the dimension of the feature to two dimensions. The methods of CS1 and CS2 focused on classification within train dataset in this paper. However, if the size of the train dataset is large enough to serve as a population, the representation learning method learned in the train dataset can also be applied in the test dataset. It means that CS1 and CS2 can be conducted for train dataset.

**6. Conclusion**

This study presents a new classification CS1 and CS2 using the subgraph of coreset selection from a geometric perspective. On the small Image dataset MNIST, Fashion MNIST, and EMNIST datasets, we can perform partial classification with 95% accuracy for unlabeled data, which is several times the given Label data l. When representation learning method suitable for datasets is applied, the efficiency become 10 to 40 times higher, and the accuracy is also close to 99%. To ensure the performance of SC1 and SC2, the tradeoff between sampling size |S| and the number of iterations P should be considered. The hyperparameter M should be applied to prevent misclassification in the SC1 situation. In two dimensions, M=6 is appropriate. On the other hand, the density α of the subgraph is not very useful. Threshold is more conservative hyperparameter than the probability that the actual class belongs to. should be set lower as the number of classes increases and the number of Iterations decreases. There are still many things to improve, but it is meaningful in that it is a new classification method to solve the lack of label data.

**<Reference>**

Arazo, E., Ortego, D., Albert, P., O’Connor, N. E., & McGuinness, K. (2020). *Pseudo-labeling and confirmation bias in deep semi-supervised learning.* Paper presented at the 2020 International Joint Conference on Neural Networks (IJCNN).

Argote, L., & Ingram, P. (2000). Knowledge transfer: A basis for competitive advantage in firms. *Organizational behavior and human decision processes, 82*(1), 150-169.

Cascante-Bonilla, P., Tan, F., Qi, Y., & Ordonez, V. (2021). *Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning.* Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.

Chapelle, O., Scholkopf, B., & Zien, A. (2009). Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks, 20*(3), 542-542.

Gal, Y., Islam, R., & Ghahramani, Z. (2017). *Deep bayesian active learning with image data.* Paper presented at the International Conference on Machine Learning.

Grandvalet, Y., & Bengio, Y. (2004). Semi-supervised learning by entropy minimization. *Advances in neural information processing systems, 17*.

Hurtik, P., Molek, V., & Perfilieva, I. (2020). Novel dimensionality reduction approach for unsupervised learning on small datasets. *Pattern Recognition, 103*, 107291.

Lee, D.-H. (2013). *Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks.* Paper presented at the Workshop on challenges in representation learning, ICML.

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.

Luo, Y., Zhu, J., Li, M., Ren, Y., & Zhang, B. (2018). *Smooth neighbors on teacher graphs for semi-supervised learning.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Qiao, S., Shen, W., Zhang, Z., Wang, B., & Yuille, A. (2018). *Deep co-training for semi-supervised image recognition.* Paper presented at the Proceedings of the european conference on computer vision (eccv).

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.

Sajjadi, M., Javanmardi, M., & Tasdizen, T. (2016). Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems, 29*.

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

Shi, W., Gong, Y., Ding, C., Tao, Z. M., & Zheng, N. (2018). *Transductive semi-supervised deep learning using min-max features.* Paper presented at the Proceedings of the European Conference on Computer Vision (ECCV).

Sinha, S., Ebrahimi, S., & Darrell, T. (2019). *Variational adversarial active learning.* Paper presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision.

Sun, Y., Li, J., Wang, W., Plaza, A., & Chen, Z. (2016). *Active learning based autoencoder for hyperspectral imagery classification.* Paper presented at the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS).

Tong, S. (2001). *Active learning: theory and applications*: Stanford University.

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.

Wang, Y., Yao, H., & Zhao, S. (2016). Auto-encoder based dimensionality reduction. *Neurocomputing, 184*, 232-242.

Yang, X., Song, Z., King, I., & Xu, Z. (2022). A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.