Classification by Coreset selection

Department of Graduate School of Data Science
KAIST
291, Daehak-ro, Yuseong-gu, Daejeon 34141
gusrn0505@kaist.ac.kr, gusrn0505@naver.com
Cell phone: +82-10-5251-3437

0. Abstract

Semi-supervised learning 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-reliable classification without relying on the performance of the model from a geometric perspective. The problem of lack of label data is solved by pseudo-labeling with this classification method that is not based on the neural network 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 range of 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 unlabeled 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 supervised learning, SSL is vulnerable to errors. If the size of the labeled data is large and it reflects the dataset well then performance of SSL is fine. However, due to the lack of Label data, if label data does not reflect dataset's diversity and is not enough to learn the model, confirmation bias occurs. Failure to prevent the confirmation bias in early stage 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

consistent 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. In this way, confirmation bias is prevented. Meanwhile, pseudo labeling utilizes prediction of learning model. However, if the confirmation bias is not resolved, the performance deteriorates. To solve this problem, set a threshold for reliability, introduce a regulatory term, or pseudo labeled data is used only in the model fine tuning process. 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 about Dataset.

Despite various disadvantages comparing with consistency regularization, there is a need to study Pseudo labeling. In some papers show that the performance is similar to Consistency Regulation when confirmation bias is prevented. 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, propose a new classification method that does not rely on model performance through the Coreset-selection method, which is one of active learning (AL). This will solve the problem of dataset's diversity unreflected and Confirmation bias due to lack of Label Data. Depending on what acquisition strategy AL 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 unlabeled data from a geometric perspective. It will then be shown that high-reliability classification is possible without relying on the performance of the neural network model. The problem of lack of label data is solved by pseudo-labeling data through this classification. Study will also present the range of 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. 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 contribution in the context of the linked research between AL 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. If the two methods are well linked, additional synergy would 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

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 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. And the last assumption means that it is possible to link with representation learning. As a result, distance-based SSL 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 consistent even if deformation is applied to the input of the model. Confirmation bias is effectively removed. This method 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, or using it when reliability of prediction 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 unlabeled 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 is 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 weakness also applies to 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.

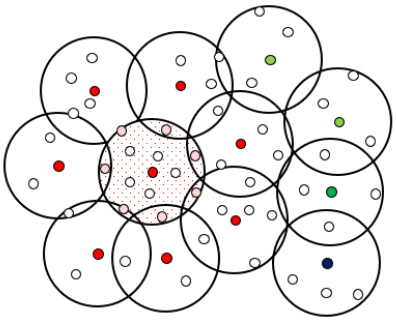
Coreset Selection samples data that maximizes the expected value of model performance. This is the same as sampling data that makes it have a minimum radius δ when constructing subgraphs that can cover the entire data with a given sampling size N{Sener, 2017 #4}. Each sampling data is widely spread regardless of the density of the dataset. For this reason, sampling data set U 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, a reliable classification is conducted by using a subgraph formed through coreset selection. Let dataset $X = \{x_k\}\, s.t.\, k = \{1, \dots,\ K = |X|\}$. Let Sampling size is N and the i-th sampled data is $u_i \in U$. Each u_i is the central point of subgraph G_i. Sample x_k which is minimizing the radius δ of subgraph G_{i+1} as u_{i+1}. Radius δ is $\max_{i\in\{1,\dots,N\}\,\&\,k\in\{1,\dots,K\}} dist(u_i, x_k)$. The density of subgraph G_i 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, x_{ij} s.t. j $\in$ {1,...,n_i | $n_i = number\ of\ unlabel\ data\ in\ G_i$} will have the same class as u_i in high probability. That is, u_i could represent unlabeled data x_{ij} belonging to a subgraph G_i. Furthermore, assumed that subgraphs formed through coreset selection have a radius δ small enough to cover the dataset tightly. Let G_{ik} s.t. k $\in$ {1, ... , m_i | n_i = $\#\ of\ subgraph\ which\ has\ connection\ with\ G_i$} is the subgraph in contact with G_i. Since u_i, u_{ik} is label data, class of u_i, u_{ik} is revealed. At this time, if the class of u_i and u_k are different, it can be inferred that G_i and G_k are in a place where different classes overlap from a geometric point of view. Conversely, subgraph will be located in the center of a certain class when all the classes of u_i and u_{ik} are the same. That is, as shown in Figure 1, it is possible to infer where each subgraph is

located through the geometric relationship between subgraphs and the class of each central point. That is, it is possible to classify the x_{ij} belonging to the subgraph G_i when $u_i = u_{i1} = \cdots = u_{im_i}$ and $n_i > \varphi$. φ is a hyperparameter.



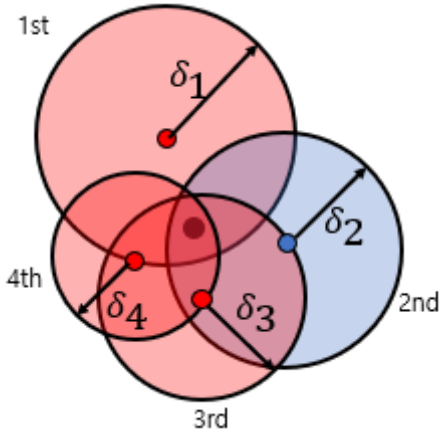
[Figure 1. Geometric relation among subgraphs]

This classification's performance is improved when dataset has small class type, a large number of unlabeled data and the distribution of a specific class is clearly distinguished. If a sufficiently small radius δ and the data distribution is dense for each class is guaranteed, the central part of each class can be classified with high reliability. To avoid confirmation bias, the results of the classification are used only for fine-tuning of the model. This method reflects dataset's diversity through active learning, and at the same time, it is possible to solve the problem of lack of label data.

Classification is possible in another way when the subgraphs are overlapped by performing the coreset selection multiple times. Through one coreset selection, all x_k are included in one or more subgraphs. Therefore, if the coreset selection is performed a total of P times, all x_k belong to at least P subgraphs. The more x_k belongs to a subgraph of a specific class, the higher the probability that x_k is the same class. In addition, the probability will be high in inverse proportion to the size of radius δ^p. For each class, the number of times $s^p = (s^p_{c_1}, \dots s^p_{c_J}\ s.t.\ J\ is\ number\ of\ class\ type)$ and radius δ^p, it is probabilistic through SoftMax.

$$p\left(y = c_j \middle| x_k^p\right) = \frac{\sum_{p=1}^{P} e^{s^p_{c_j}/\delta^p}}{\sum_{j=1}^{J}\sum_{p=1}^{P} e^{s^p_{c_j}/\delta^p}} \ \text{s.t. k} \in \{1 \dots, \text{K}^p\} \quad (3.1)$$

In order to make the radius δ^p smaller according to the coreset selection sequence, extra coreset selection focus on unlabel dataset X^p that is not pseudo-labeled in first time. Confirmation bias may occur when the classification is fixed to the highest probability value. Therefore, mix-up {Zhang, 2017 #8} is applied to soft labeling. In addition, there is a high probability of giving a wrong pseudo label until a sufficient number of subgraph overlaps occur. Therefore, the pseudo label is given only if the probability of a specific class is greater than 1- ε.



[Figure 2. Subgraphs overlapped on x_k through P-times coreset selection]

3-2) Choose Proper δ

A sufficiently small radius δ should be applied to ensure the performance of the classification. As the size of the radius δ is sufficiently small, the peripheral and central portions of the class data may be clearly distinguished. Conversely, if the radius δ is large, the classification performance will be lowered as the subgraphs overlap with the periphery. In addition, the error will increase, such as the area in which minor class included in one subgraph. The size of the radius δ depends on the sampling size N.

Guidelines are needed to select the appropriate sampling size N. Labeling Cost and classification accuracy have a trade-off relationship. When the size of N is small, the labeling cost decreases, but the δ increases so that reliable classification cannot be performed. Conversely, if the size of N is large, δ is reduced, which improves the performance of classification. In this study, it is judged that preventing confirmation bias is a priority. We will prevent confirmation bias by regulating the number of data belonging to the subgraph not to exceed a certain level.

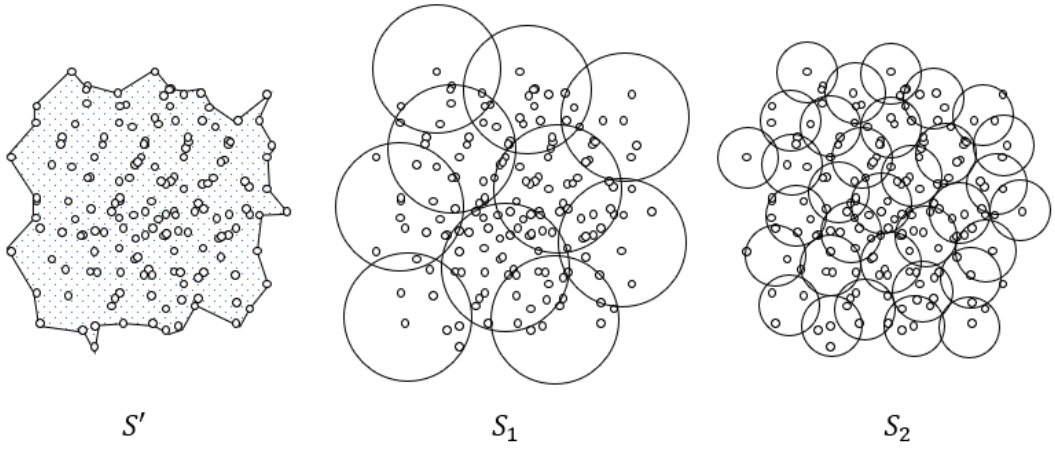
Find the radius δ' to make the amount of data in all subgraphs M or less. Coreset selection is performed with an arbitrary sampling size N^0. Coreset selection is performed again only for data belonging to the densest subgraph. At this time, find a radius δ' that allows the number of data belonging to each subgraph to be less than M. The radius δ' reduces the number of data belonging to the densest subgraph to M. Therefore, even when extended to the entire dataset, all subgraphs contain only M or less data.

We will find a sampling size N' that will make the radius no larger than δ'. A range of sampling size N′ are formed through an area S of the dataset X. However, in order to secure robustness when the outlier deviates significantly from the data distribution, the area S of the dataset is redefined as the area covered by subgraphs. For example, suppose there is two subgraphs with radius δ. At this time, if the center of each subgraph is at a distance of 2δ or more, each subgraph not overlap and S becomes $2\delta^2$. If the two subgraphs overlap, subtract the overlapping portion at $2\delta^2$. The overlapping part can be calculated using the distance between the two central points. A range for the area S will be formed by utilizing the characteristic that the distance between each u_i is at least

δ or more. The characteristic of the distance between the two points are demonstrated by Lemma 1. Our goal is to prove Theorem below.

Theorem 1. Given a pair of arbitrary sample sizes and radius (N, δ), the range of N′ which make radius is not larger than δ' can be calculated.

The whole process is as follows. First, look for δ'. Let δ^0 is radius when coreset selection is performed with an arbitrary sample size N^0. Let S_1 is area of pair (N^0 , δ^0). Also, let S_2 is the area when (N', radius δ'). It is not difficult to assume that N′ is greater than N^0. For each of $S_1\ and\ S_2$ shall be satisfied $\delta_0 , \delta' \geq \max\left(\underset{i,j\in\{1,\dots,K\}}{min} dist(x_i, x_j)\right)$ so that the subgraph can cover the dataset tightly. Or, to robust for the outlier, $\delta_0 , \delta^* \geq average\left(\underset{i,j\in\{1,\dots,K\}}{min} dist(x_i, x_j)\right)$ could be applied. We will empirically verify the difference between the two in the future. Let S' be the area represented by the data on the outermost side of the dataset. The subgraphs of coreset cover all the data, so it is obvious that $S_1\ and\ S_2$ are larger than S'. In addition, it can be confirmed through Figure 3 that S_1 with a larger radius will be greater than S_2. Therefore, $\min(S_2) \leq S_2 \leq S_1$ will be satisfied. Here, S_1 can be calculated, and if the value of $\min(S_2)$ can be obtained by the formulas of N' and δ', an inequality for the target value N' can be created.



[Figure 3. Comparison of area according to area definition and subgraph radius size difference]

In addition, if we can prove $S_2 \leq \max(S_2)$, $S_1 \leq N_0 * \delta_0^2 \pi$ would mean that $S_2 \leq \max(S_2) \leq N_0 * \delta_0^2 \pi$, we can create opposite inequality in N′. We will check this is theoretically or empirically

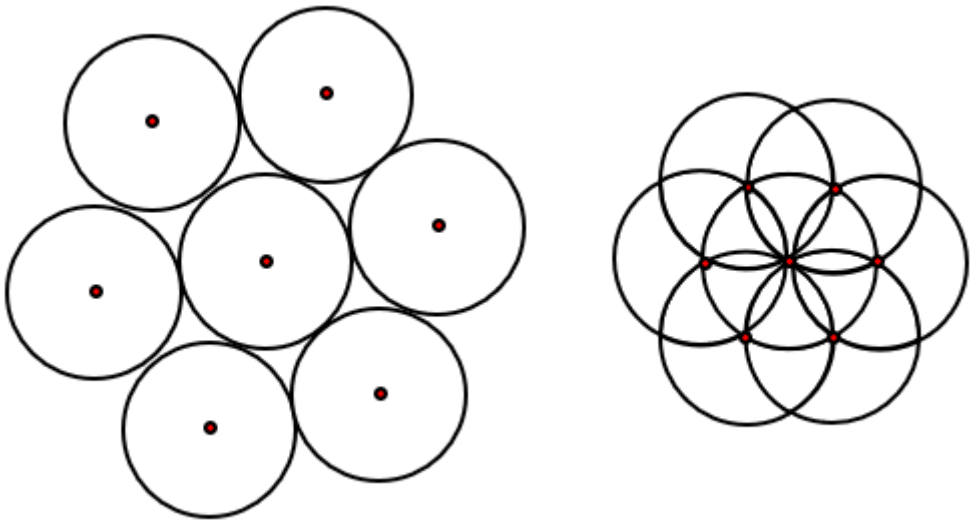
Let's calculate S_1, $\min(S_2)$, $\max(S_2)$. Given a N′ and δ', $\max(S_2)$ is easily calculated. In order to consider about $\min(S_2)$, the characteristic of each subgraph of center point should be used. Let's prove it one by one.

Lemma 1. Each center point of subgraphs is departed at least δ.

Coreset selection is the same as k-center greedy algorithm. The first sampling data is randomly sampled. And define radius δ_i as $\max(dist(x_k, u_i))\ s.t.\ k \in \{1, \dots, K\}$. Sample x_k as u_{i+1} which

satisfy dist$(x_k, u_i) = \delta_i$. Repeat this process until sampling size N is reached. Thus, it satisfies $\max\left(dist(u_i, u_j)\right) = \delta_{\max(i,j)-1}$, $\delta_i \leq \delta_{i-1}$ for all $i, j \in \{1, \dots, N\}$. An equal case is established when there are two or more points having a distance of δ_{i-1}. Finally, $\delta_N \leq \delta_{\max(i,j)-1} = \max(dist(u_i, u_j))$ $for\ all\ i, j \in \{1, \dots, N\}$ is satisfied.

Based on Lemma 1, $\min(S_2)$ can be considered as a situation where each subgraph is overlapped as much as possible at intervals of radius δ as shown in Figure 4. The areas of each of $\min(S_2), \max(S_2)$ can calculate, and will be described in Lemma 2.



[Figure 4. Shape of $Max(S_2)$, $Min(S_2)$]

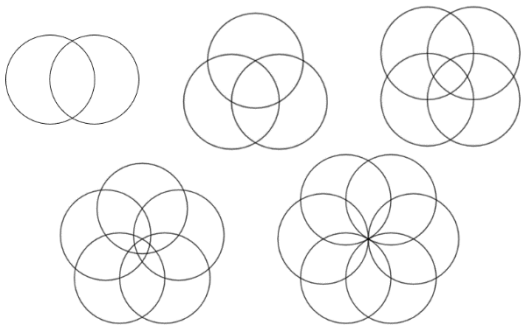
Lemma 2. Let Sample size n and radius 1 are given. Let k = $\max_k(3k^2 - 3k + 1 \leq n)$, $m = n - (3k^2 - 3k + 1)$, $q = quantin(m-1, k-1), d = mod(m-1, k-1)$ $when\ m \geq 1$. Then $Max(S_2) =$ n $* \pi$, $Min(S_2) = \left(2\pi - \frac{9\sqrt{3}}{4}\right)k^2 + \left(-6\pi + \frac{21\sqrt{3}}{4}\right)k + \frac{15\pi}{3} - \frac{\sqrt{3}}{2} + q\left[\left(\frac{2\pi}{3} - \frac{3\sqrt{3}}{4}\right)k - 2\frac{\pi}{3} + \frac{\sqrt{3}}{2}\right] + d\left(\frac{2\pi}{3} - \frac{3\sqrt{3}}{4}\right)$ $when\ m \geq 1$, $\left(2\pi - \frac{9\sqrt{3}}{4}\right)k^2 + \left(-6\pi + \frac{21\sqrt{3}}{4}\right)k + \frac{13\pi}{3}$ when $m = 0$.

*The calculation process is complicated, so process will be described in Appendix later.

There are several things to consider in order to calculate the area S. First, there are only four cases in which each subgraph overlaps, from the case two subgraphs overlap to the case five subgraphs overlap at the same time. This will be demonstrated in Lemma 3. Whether each of the four cases occurs or not is can be distinguished through the adjacency matrix of subgraphs. Finally, the area of the four cases can be calculated through the distance and radius δ between each center.

Lemma 3. When the distance between the centers of circles with the same radius is greater than or equal to the radius, there are only four cases in which multiple circles overlap at the same time.

In order to make that the circles overlap as much as possible, it is assumed that all circles pass through different circles' central points. Figure 5 describes from the case where two circles overlap to the case where six circles overlap. When the six circles overlap, there is only one point overlapping each other. That is, it can be confirmed that the limitation is that the five circles overlap at the same time.



[Figure 5. Case of circle overlap when distance of each central point is same or larger than δ]

Lemma 4. Adjacency Matrix of subgraph distinguish case of overlapping.

Lemma 5. For each case, the area of S can be calculated.

Each Lemma 4 and 5 will be described later through Appendix. Finally, all values related to $\min(S_2) \leq S_1 \ \& \ S_2 \leq \max(S_2) \leq n_0 * \delta_0^2 \pi$ can be changed to expression for N, N′, δ and δ'. The other values except N′ are constants. Therefore, the range for N′ can be obtained. In addition, to ensure the efficiency of the subgraph, it may be considered to include at least M′ point for subgraph.

4. Expected results

This study presents a new classification method using the subgraph of the coreset selection from a geometric perspective. By using this as pseudo labeling, confirmation bias due to lack of label data can be prevented. Furthermore, we present the range for the sample size required to ensure the performance of classification. This provides chance for calculating the labeling cost that requires securing classification performance in the field.

The performance of this method varies according to the characteristics of dataset. The larger the dataset and the smaller the number of classes, the better the classification performance will be. Above all, this method can be applied to all datasets where distance reflects the similarity of characteristics between data. In addition, it is valuable in that it can create synergy if used with representation learning.

<Reference>

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*.