Project5

1.Team member

- **m5268101 Liu Jiahe**
- **m5251140 Ken Sato**
- **m5271051 Keisuke Utsumi**

2.Team Project V

- Using SOFM(Self-Organization feature map) algorithm to cluster the Iris dataset (http://www.ics.uci.edu/~mlearn/MLRepository.html).

3.Algorithm

1.About SOFM

This network can reduce data length (data compression). It maps data into lower dimension with preserving neighborhood relation.

It uses Kohonen network, but learning algorithm is changed. Unlike the winner-take-all algorithm, its algorithm updates the weight of all neurons close to winner. The weights of neurons are updated with data x as follows:

The SOFM (Self-Organizing Feature Map) algorithm can reduce the dimensionality of data while preserving the topological structure of the data in space. In this algorithm, the neurons are distributed in a two-dimensional network, each neuron has six neighborhoods, forming a hexagonal grid. The number of weights of each neuron is the same as the dimension of the input data vector. The steps of the algorithm are as follows:

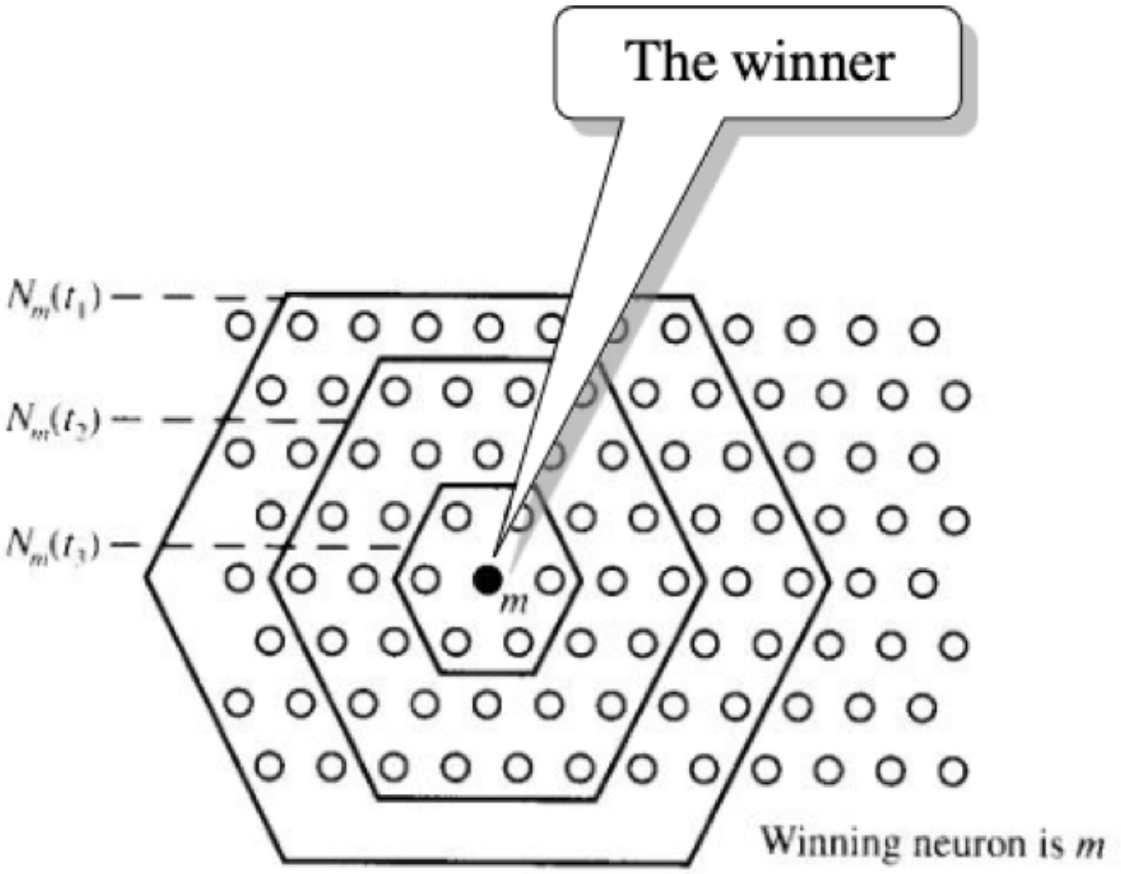
$$W_i = W_i + \alpha(x = W_i), \quad for \quad i \in N_m \tag{1}$$

where N_m is the neighborhood of the m-th neuron, α is the learning rate which decreases with training time and distance between neurons. It is:

$$\alpha = \alpha(t)e^{-r/\sigma^2 t} \tag{2}$$

where $\alpha(t)$ and $\sigma(t)$is the decreasing function, *t* is the trainning time and *r* is the distance between current neuron and winner neuron.

The neurons are usually arranged in a 2- dimensional planar array with hexagonal neighborhoods. During learning, the weights of all neurons in the neighborhood of the winner are updated. The amount of modification is inversely proportional to the distance between the neuron to be updated and the winner. The size of the neighborhood is reduced during learning.



2.Code prpcessing

1. Initialization:
    - Initialize the connection weight matrix w, where $w[n][i]$ represents the connection weight between the nth neuron and the ith component of the input vector.
2. Self-organizing mapping cycle：
   a. Select a training sample vector:
    - Randomly select a sample vector $x[p]$ from the training set.

   b. Find the most similar neuron:
    - Calculate the distance between the sample vector $x[p]$ and each neuron.

- Find the neuron with the smallest distance to the sample vector, i.e., calculate the minimum Euclidean distance:
  $d(m_1, m_2) = \sum_{i=1}^{I}(w[m_1, m_2][i] - x[p][i])^2$, where m_1 and m_2 represent the position of the neuron on the two-dimensional grid.

c. Update the weights of the neurons in the neighborhood:

- Define the learning rate r and the neighborhood radius nc, which gradually decrease with the increase of iterations.
- For each neuron $w[m_1, m_2]$, calculate its position (x_1, x_2) in two-dimensional space. Adding 0.5 to the x-coordinate of even rows is to visually form a hexagonal network, but it does not affect the weight of the neuron.
- If the neuron (m_1, m_2) is in the neighborhood of the closest neuron (m_{10}, m_{20}) (distance less than or equal to the neighborhood radius nc):
  - Update the connection weights:
    $w[m_1, m_2][i] = w[m_1, m_2][i] + r \cdot (x[p][i] - w[m_1, m_2][i])$, where i represents the index of the input vector component.

3. Calibration
   - Assign the label of each input pattern to the most similar (nearest) neuron and update the label of the neuron.
4. Print results:
   - Print the final neuron label matrix to show the classification results of each neuron.

3.The process of SOFM

Use Iris dataset as example

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
....
....
```

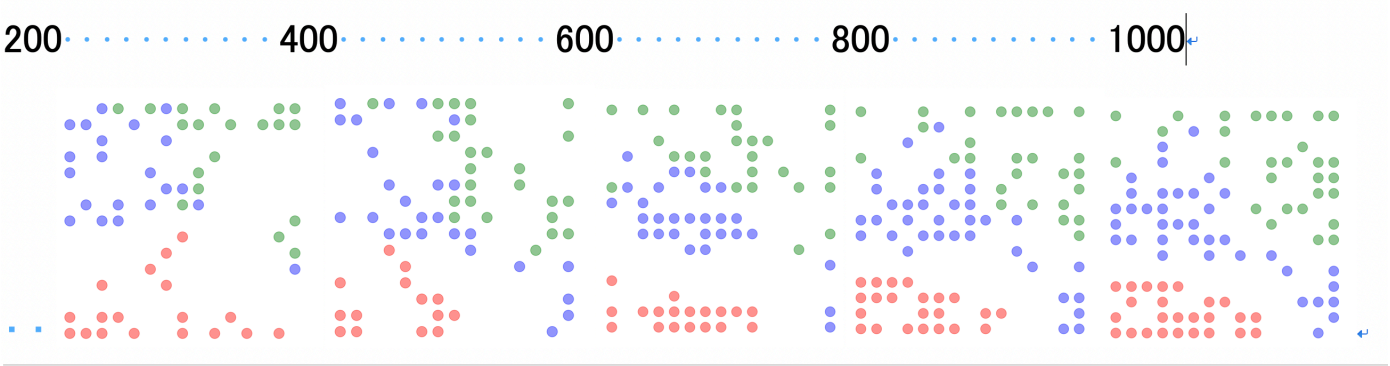
Lots of datasets do not classify the internal data. However, To visually demonstrate the process of the SOFM algorithm, we utilize the provided classification information within the iris dataset.

A----setosa-----red. B----versicolor----blue. C----virginica----green,

learning rate 0.5->0.0

neighborhood size 10->1

project5.c,draw_SOTM.py



As the number of training steps increases, it can be observed that the distances between data of the same type become smaller and they tend to cluster together, while preserving the topological structure.

And different learning rate and neighborhood distance will lead to defferent result.

From above we can konw:

1. A wide range of learning rates in the first 1000 cycles makes classification a mess.
2. Neighborhood size should be set from 6 to 1 with reference to the material for better accuracy.

Next, we will delve deeper into the discussion of the Self-Organizing Feature Map (SOFM).

4.Code

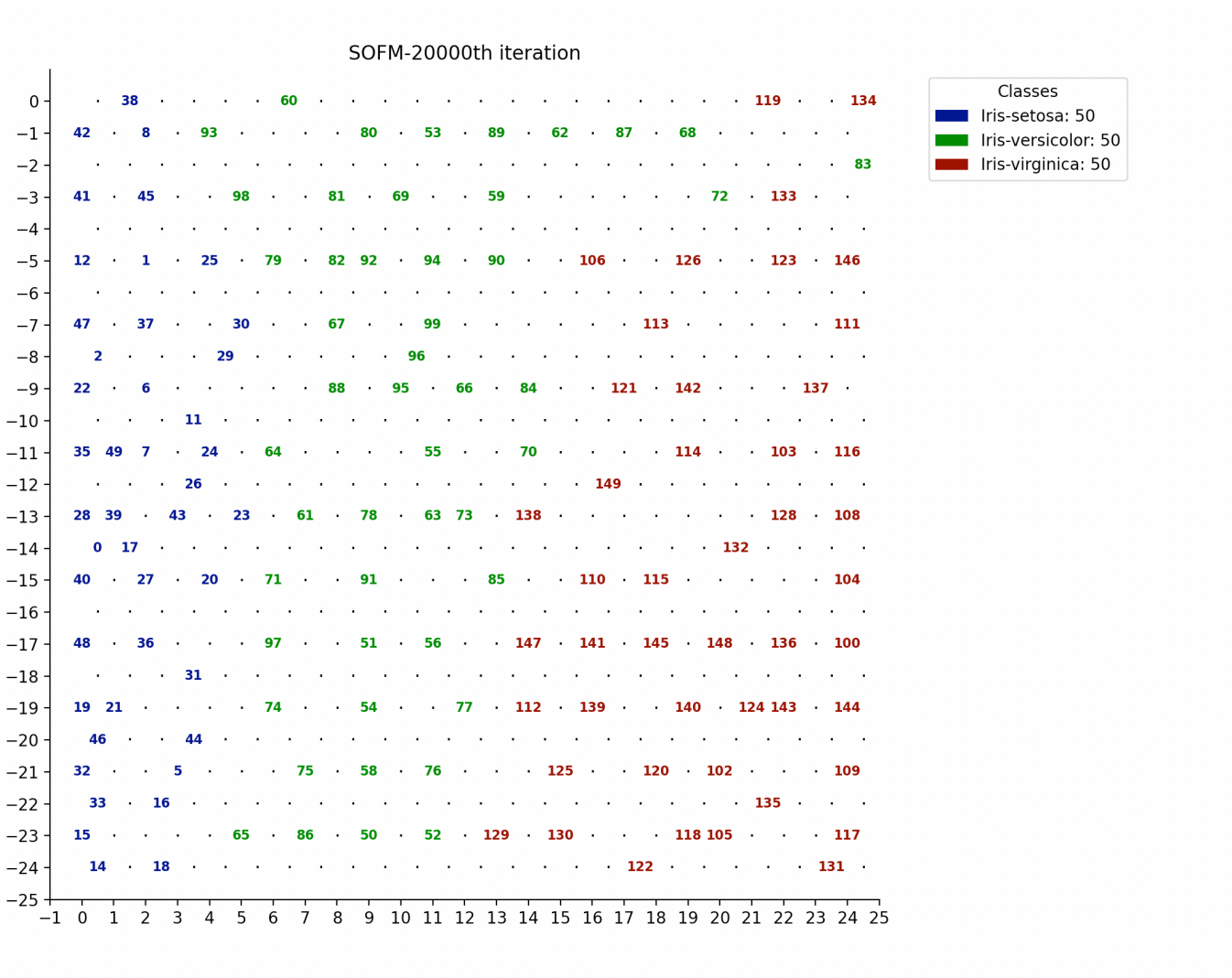
project5.py, use utils.py

```
from utils import *

def main():
    x, label0, label_dict, I, P, color_dict = InputPattern("./iris.data")
    # print(color_dict)
    use_classes_in_iris = True
    use_wta_cluster = True if not use_classes_in_iris else False
    if use_wta_cluster:
        classes = np.loadtxt('./clusters_3.txt', dtype=str)
        label_dict = {label0[i]: classes[i] for i in range(len(classes))}

        unique_classes = np.unique(classes)
        color_list = ['red', 'blue', 'green', 'purple', 'cyan', 'orange', 'magenta', 'lime', 'navy', 'darkred', 'darkgreen', 'gold', 'teal']
        random.shuffle(color_list)  # Randomly shuffle the color list
        color_dict = {unique_classes[i]: color_list[i] for i in range(len(unique_classes))}
        # print(label_dict)
        print(color_dict)

    # print(label_dict)
    w = np.random.random((N, I))
    w = SOFM(3000, 0.5, 0.04, 10, 1, w, x, I, P)
    label = Calibration(w, x, label0, I, P)
    print("\n\nResult after the first 1,000 iterations:\n")
    # PrintResult(label)
    # PrintResult_figure(label, label_dict, 3000, color_dict)
    w = SOFM(17000, 0.04, 0.0, 1, 1, w, x, I, P)
    label = Calibration(w, x, label0, I, P)
    print("\n\nResult after 10,000 iterations:\n")
    PrintResult(label)
    PrintResult_figure(label, label_dict, 20000, color_dict)

if __name__ == "__main__":
    main()


```

5.Results

1.Using given classes

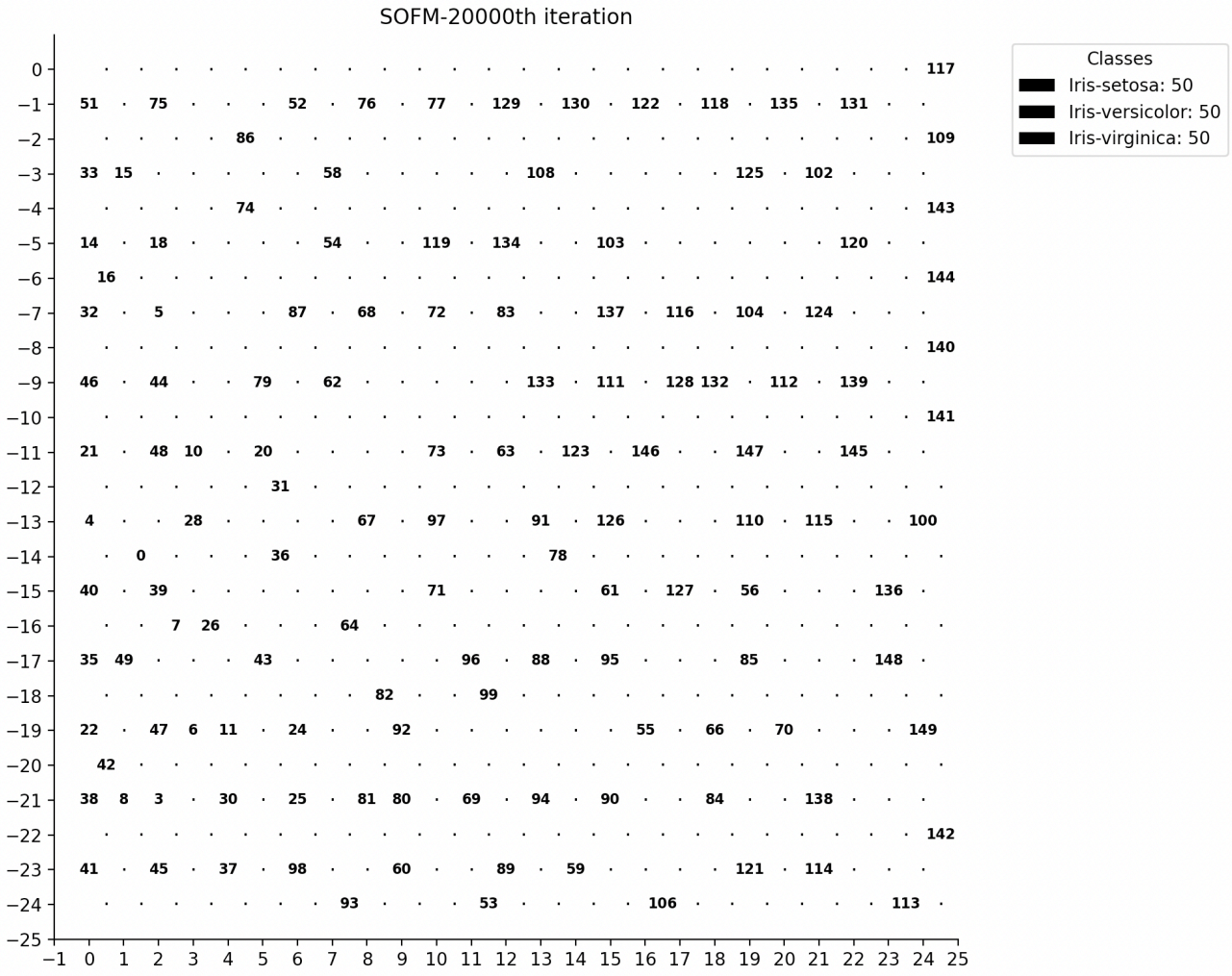
Assign labels ranging from 0 to 149 to the 150 sets of data in the Iris dataset. Classify each label based on the dataset's inherent category information. Then, apply the SOFM algorithm to map the high-dimensional data onto a two-dimensional graph. Color each category separately.



It can be observed that the data retains topological structure after dimensionality reduction. At the same time, the distribution of the data aligns with the given classifications in the dataset. Data belonging to the same class are accurately clustered together.

2.Don’t use given class

In the application scenarios of SOFM (Self-Organizing Feature Map), the dataset itself usually does not have inherent classification information. After applying the algorithm, the result is as shown in the figure.



It can be observed that although SOFM compresses the data while preserving the original data's topological structure, it cannot explicitly divide the data into different clusters. This means that we cannot find explicit clusters from the results of the SOFM algorithm.

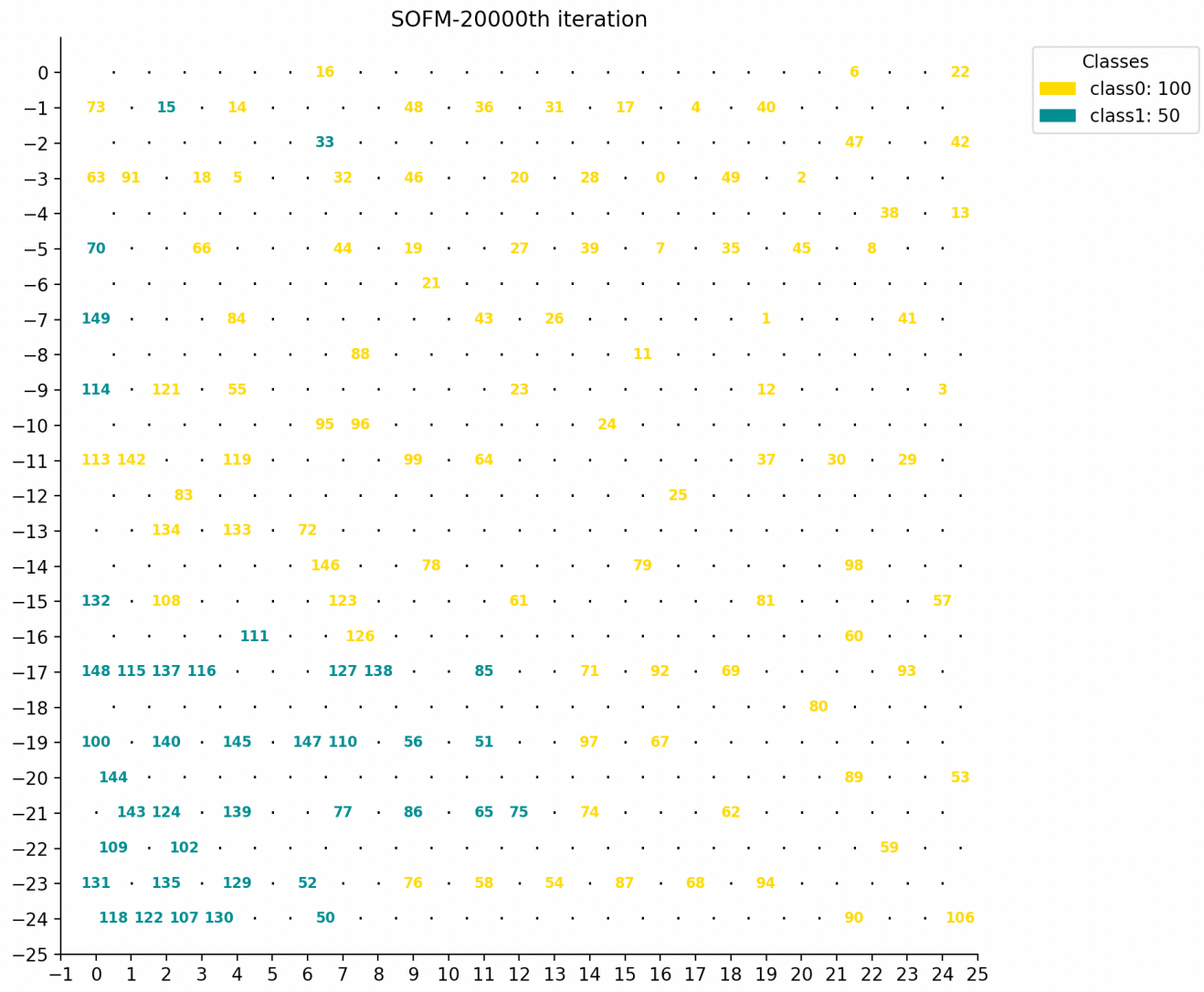
3.Cluster then SOFM

To address the above issue of preserving topological structure while achieving clustering after data compression, we can first use the WTA (Winner-Take-All) algorithm to cluster the dataset. Then, we can apply the SOFM algorithm.

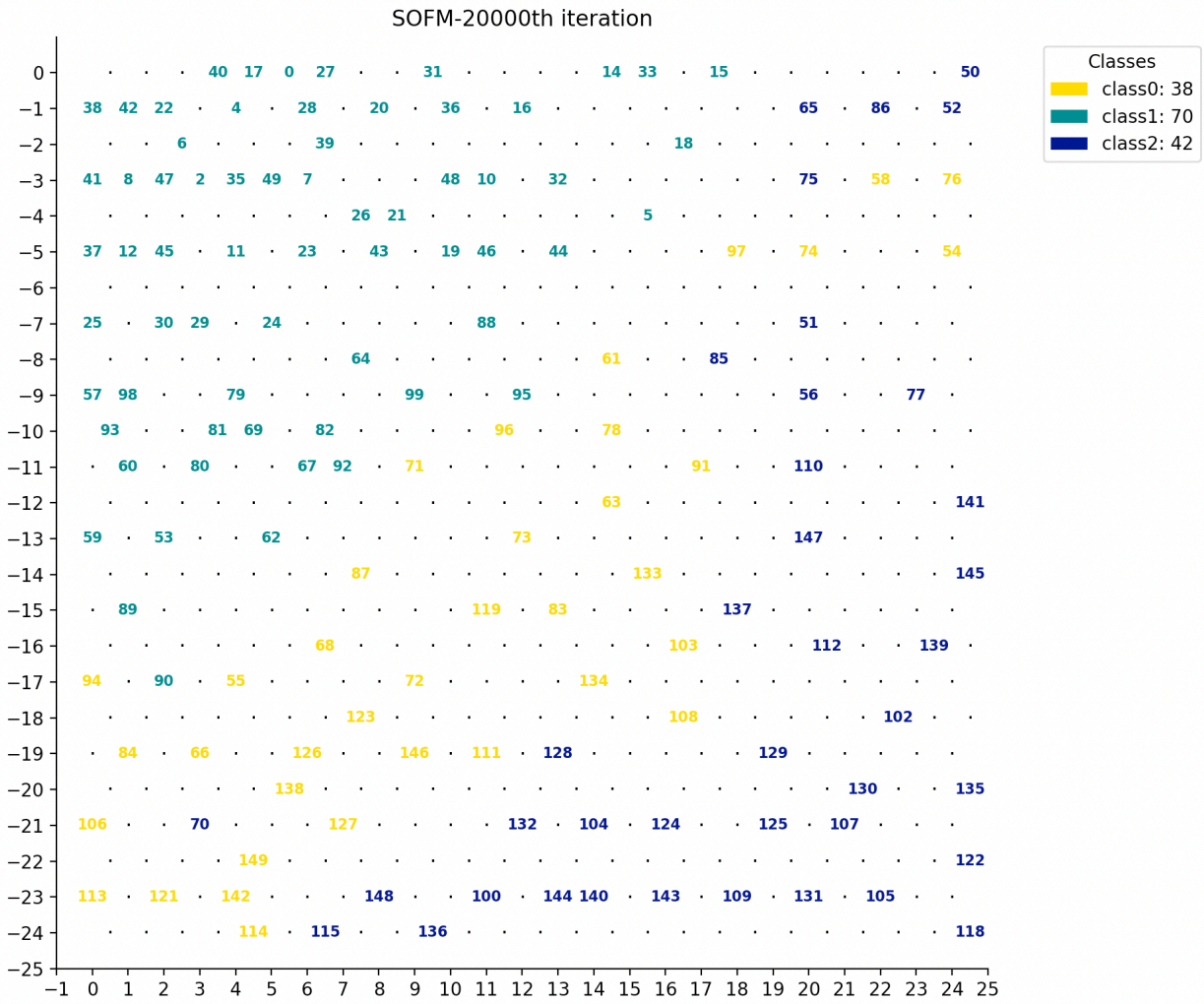
The script output_cluster.py allows customizing the number of clusters and outputs the clustering results to a text file.

The script project5.py reads the text file containing clustering information and implements the SOFM algorithm accordingly.

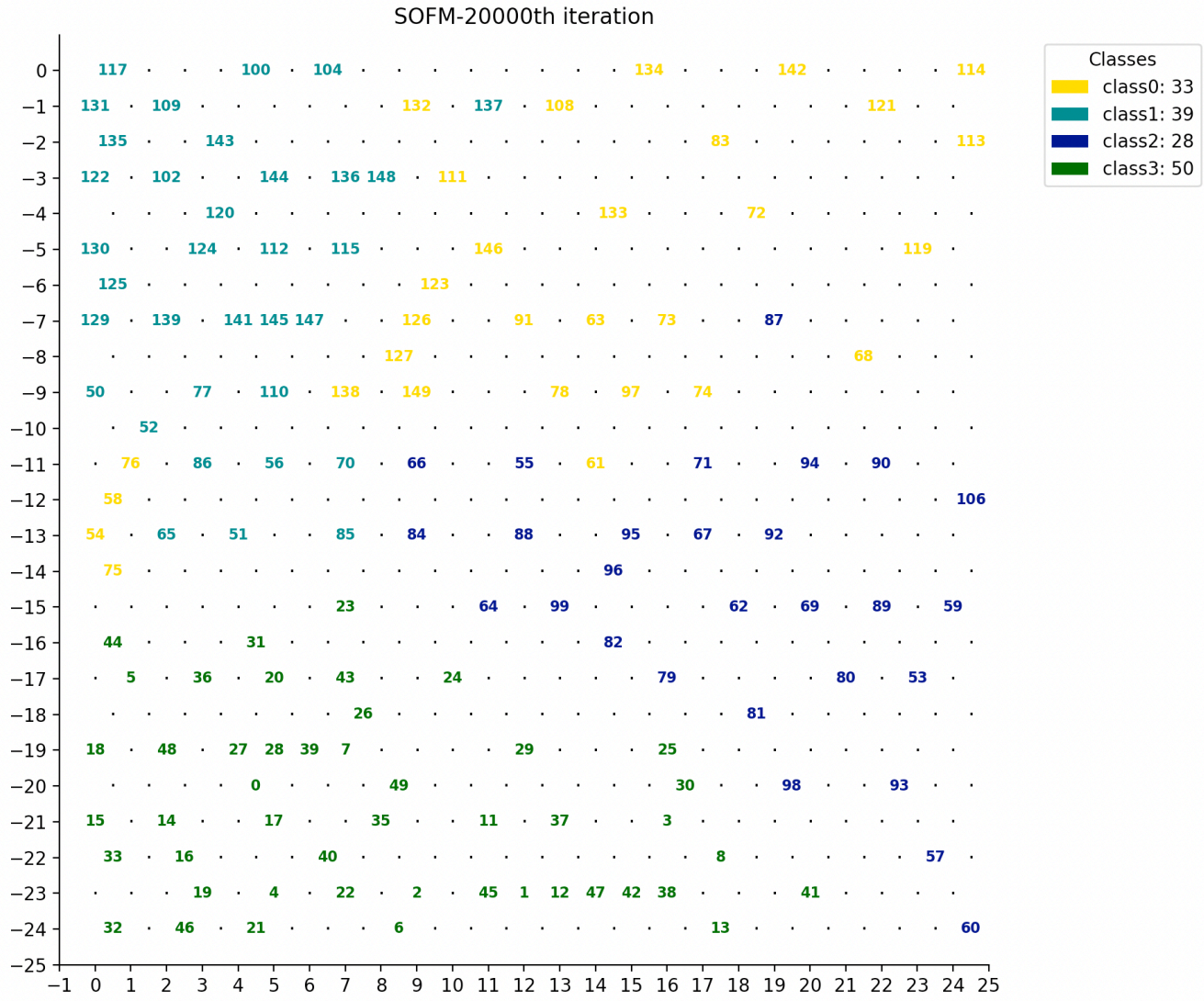
1.n_clusters = 2



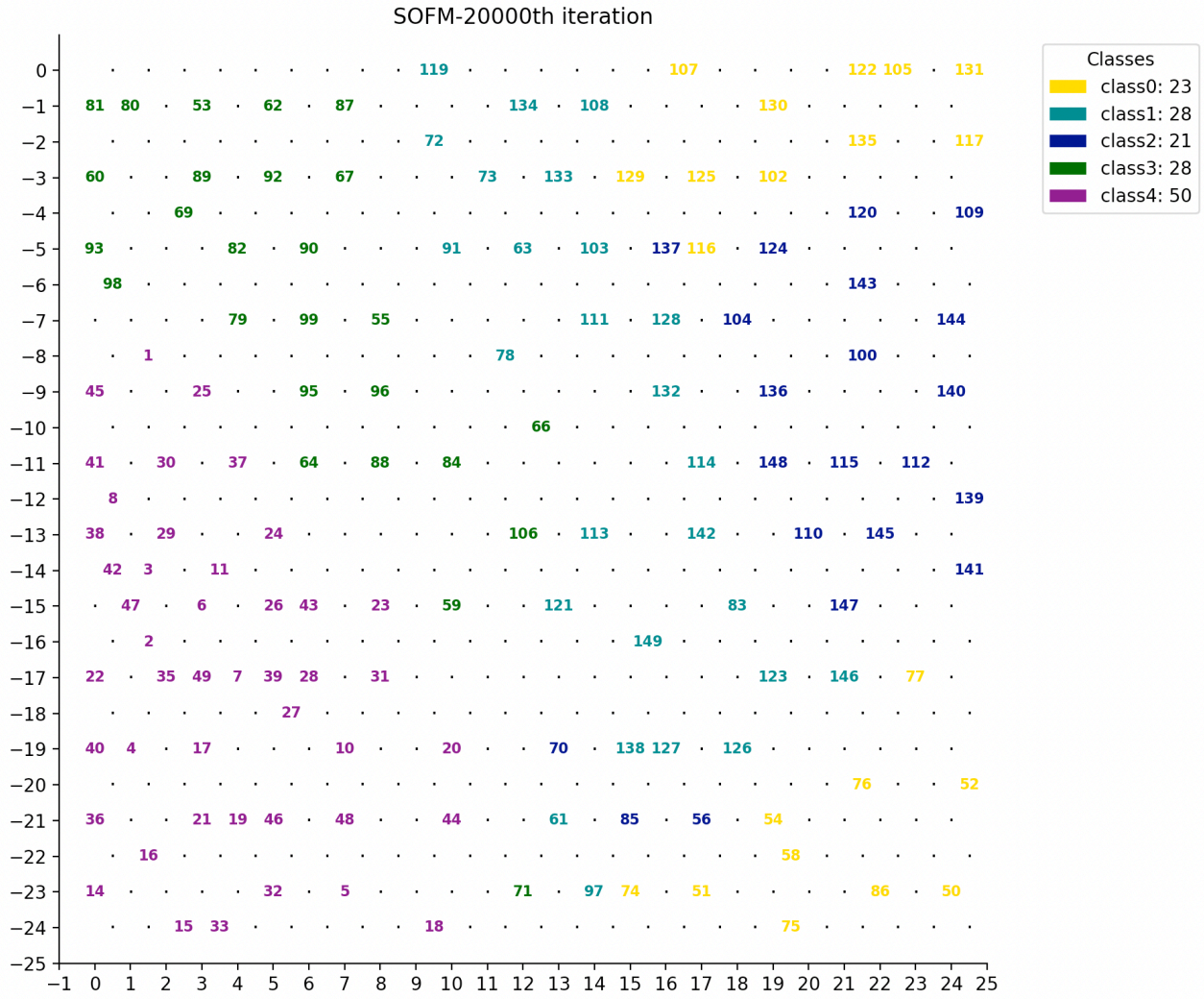
1.n_clusters = 3



3.n_clusters = 4



4.n_clusters = 5



It can be observed that the clustering results are relatively accurate, with data of the same type concentrated together. At the same time, the topological structure is also preserved.

6.Conclusion

The image above demonstrates that the combination of the WTA and SOFM algorithms yields excellent results. It allows for data compression while preserving the original data's topological structure and accurately clustering the data.

Furthermore, by observing the image, it can be noticed that not all 150 labels are fully displayed. This is because:

Each input pattern corresponds to the closest neuron, so it is possible that multiple input patterns correspond to the same neuron.

During the iteration process, if two or more input patterns have the same "closest" neuron, the labels of the later input patterns will overwrite the labels of the previous input patterns in the labeling process, resulting in the omission of some labels in the image display.

Below is the relevant code:

```
def Calibration(w, x, label0, I, P):
    label = ['.'] * N
    for p in range(P):
        d = np.linalg.norm(w - x[p], axis=1)  # calculate the Euclidean
distance
        n0 = np.argmin(d)  # find the index of the smallest distance
        label[n0] = str(label0[p])
    return label
```

It means that different input patterns (p) can correspond to the same neuron (n0).

But it is not a bad thing for different patterns to be assigned to the same neuron.

In practical applications, we don't always need to completely separate all the data. Many times, we want to discover the overall structure of the data rather than （delve into) distinguish every single data point. By increasing the number of neurons, we may obtain a more detailed mapping that separates all the patterns, but this would significantly increase computational complexity.And the model may becoms overly focused on the details of the data while overlooking its overall structure.

I think a reasonable approach is to use a smaller network structure for Self-Organizing Feature Map (SOFM) and then separately list which patterns are assigned to the same neuron. This approach reduces computational complexity while ensuring that no data is overlooked.

For example，we can output the patterns who are assigned the same neuron:

```
(19, 21): class0
(4, 17): class0
(6, 47): class0
(27, 28): class0
(9, 34, 37): class0
(65, 75): class1
(141, 145): class1
(112, 139): class1
(120, 124): class1
(140, 144): class1
(117, 131): class1
(118, 122): class1
(116, 137): class1
(78, 91): class2
(101, 142): class2
```

Use "(9, 34, 37): class0" as example:

In Iris datasets:

label 9:　4.9,3.1,1.5,0.1,Iris-setosa

label 34: 4.9,3.1,1.5,0.1,Iris-setosa

label 36: 5.5,3.5,1.3,0.2,Iris-setosa

We can find that the similarity of these overlapping data is indeed very high.