Two-class pattern classification problem with RBF and SVM classifier

Chen Yiming
School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
Email: ychen099@e.ntu.edu.sg

Abstract

This report solves a two-class pattern classification problem with Radial Basis Function (RBF) neural network and Support Vector Machines (SVM) neural network, respectively. Classifiers are trained using the given training data, and the performance of two classifiers with different parameters is discussed in this report.

1. Data

The given training data consists of 330 samples, and each piece is represented by a one-dimensional array of length 33. The validation data is selected from training data randomly, and the selected data will not be applied to the training process. There are two kinds of labels in the training label: -1 and 1. The testing data contains 21 samples.

2. RBF Neural Network

In the training process of the Radial Basis Function neural network, centers need to be determined first. There are three typical strategies for neuron center determination:
(1) Random selection from training samples
(2) Prototypes of training samples as neuron centers
(3) Center selection as a model selection problem
In this report, I apply the first two methods to determine the center. I assume the number of neurons in this section is fixed. I will discuss how the performance of the classifier changes with the number of neurons in 4.2.

2.1 Random Center

Parameters:

a. Basis function: Gaussian basis function
b. Center_x = Center_y = 8 (Assume that the centers are distributed in matrix form)
c. Sigma: $\sigma = \mathrm{dmax}/\sqrt{2m}$ = 0.875

Results:

By verifying the model thirty times through the validation set, I calculate the average MSE and accuracy shown in Table 1.

2.2 Prototype Center

There are two typical clustering algorithms: Self-Organizing Map(SOM) and K-Means clustering. Both of them could be applied to prototype center selection.

2.2.1 Self-Organizing Map (SOM)

Parameters:

a. Weight initialization: Randomly take 33 floating-point numbers in the range of -1 to 1 as an initial center sample.
b. Basis function: Gaussian basis function
c. Center_x = Center_y = 8
d. Learning rate η: At iteration n, we have $\eta(n) = \eta_0 * \exp(-n/\tau 2)$. $\eta_0 = 0.1$, $\tau 2 = 1000$.
e. Sigma: At iteration n, we have $\sigma(n)=\sigma_0 * \exp(-n/\tau 1)$. $\sigma_0 = d_{max}/2 = 4.95$, $\tau 1 = 1000/\ln \sigma_0 = 625$
f. Iteration time: 1000

Results:

By verifying the model thirty times through the validation set, I calculate the average MSE and accuracy shown in Table 1.

2.2.2 K-Means

Parameters:

a. Basis function: Gaussian basis function
b. Center number: 64
c. Sigma: 1

Results:

By verifying the model thirty times through the validation set, I calculate the average MSE and accuracy shown in Table 1.

3. SVM Neural Network

Parameters:

a. Kernel Functions: Gaussian kernel function
b. Decision function shape: one v one
c. C: 1.0
d. Gamma: 1.0

Results:

By verifying the model thirty times through the validation set, I calculate the average MSE and accuracy shown in Table 1.

Classifier	Average MSE	Average Accuracy
RBF-Random Center	0.6321	0.8787
RBF-SOM Center	0.5431	0.8453
RBF-K-Means Center	0.4532	0.8820
SVM	0.3510	0.9123

Table 1: Performance of RBF neural network with different strategies for neuron center determination and SVM neural network.

4. Data Visualization

4.1 Performance and stability

Since the validation set is randomly selected and RBF neural network has different centers every time, the performance of these neural networks has a degree of randomness. To observe the performance of these neural networks more intuitively and their stability to random data, I recorded the MSE and accuracy of each time of the thirty calculations and plotted the image with the MSE and accuracy recorded as scatter points. (Show in Figure1)

4.2 Number of RBF Hidden Layer Neurons

In order to compare the performance of different center selection strategies more precisely, the number of neurons in the second part is set to 64 for all designs. To find the suitable number of hidden layer neurons in the RBF classifier, I take the RBF model generated by K-MEANS as an example. I change the number of neurons from 26 to 100 and calculate each digit ten times to reduce the error caused by randomness to get the average of the MSE and accuracy. Scattered points and fitted straight lines are shown in Figure 2.

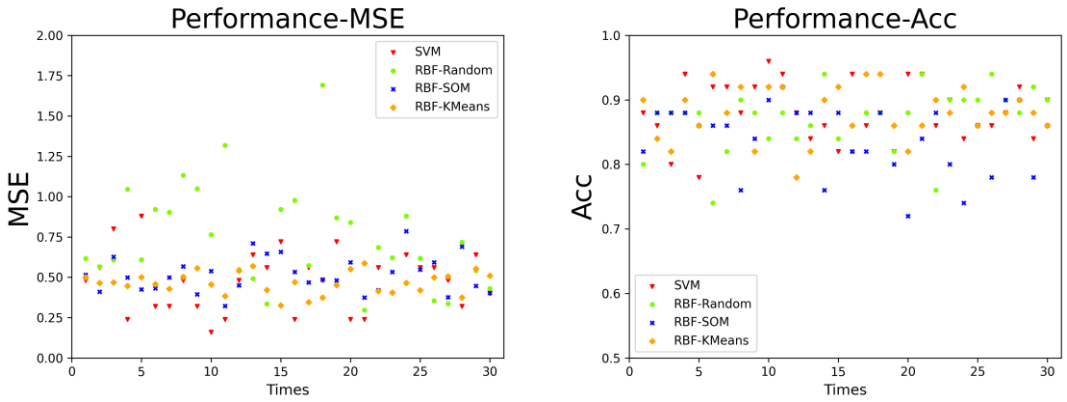

Figure 1: MSE and accuracy performance plot

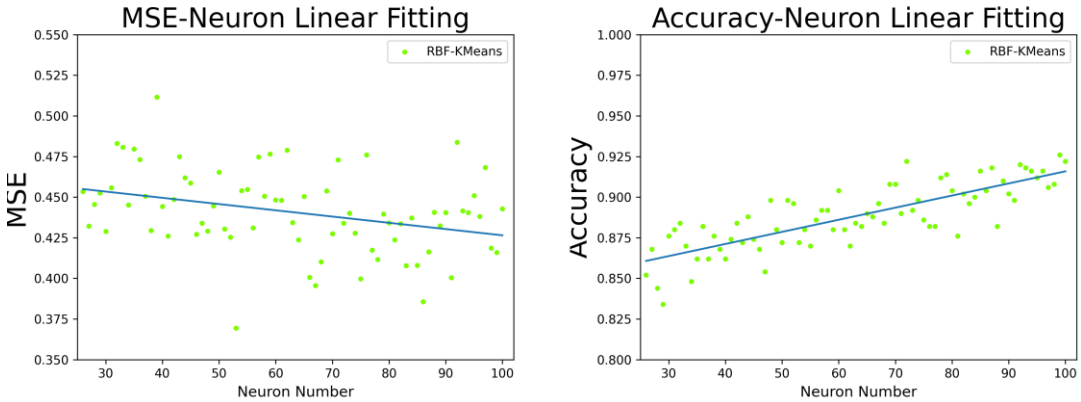

Figure 2: Changes in performance with changes in neuron numbers

4.3 Overfitted Problem

I only focus on the classifier change when the number of neurons increases in a small range in the last part. In this part, I compared the performance of the validation set and training set with the increase of hidden layer neuron numbers in the RBF-KMeans classifier. I use accuracy as the criterion and observe its change as the number of centers increases from 1 to 200. The line chart is shown in Figure 3.

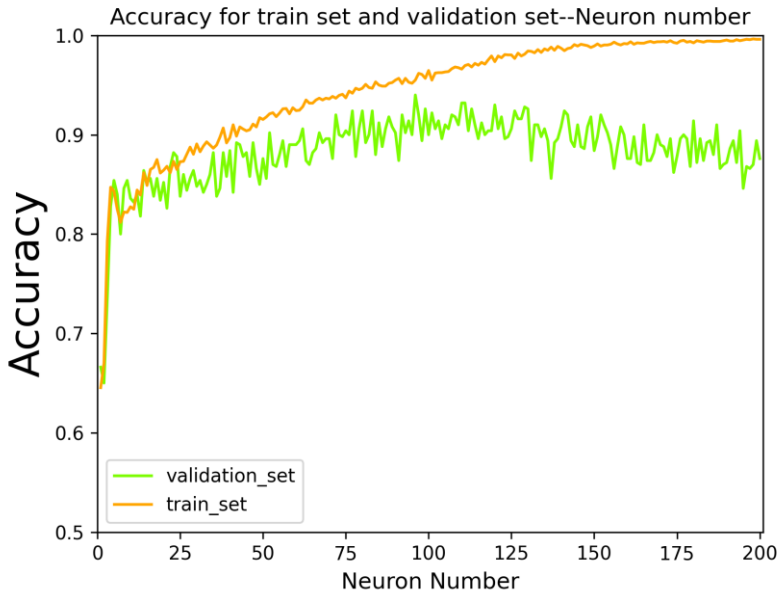

Figure 3: Accuracy for train set and validation set-Neuron number Relationship

5. Test Results

I predict the label of the test data with different classifiers and different center selection strategies, and the predicted label is shown in Table 2.

SVM	1	1	1	-1	1	-1	1
	-1	1	-1	1	1	1	1
	1	1	1	-1	1	-1	1
RBF Random	1	1	1	-1	1	-1	1
	1	1	-1	1	-1	1	1
	1	-1	1	-1	1	-1	1
RBF SOM	1	1	1	1	1	-1	1
	1	1	-1	1	1	1	1
	1	1	1	-1	1	1	1
RBF KMeans	1	-1	1	1	1	-1	1
	-1	1	-1	1	1	1	1
	1	-1	1	-1	1	1	1

Table 2: Predicted label

6. Conclusion

For the Radial Basis Function neural network, both strategies for neuron center determination and the number of hidden layer neurons will affect the performance of the classifier.

1. The random center strategy shows instability. Some MSE values will exceed 1, while some are about 0.3. The reason for the fluctuation is some random centers are more suitable while some are not.
2. The performance of the prototype center is much better than the random one, and the classifier trained using this strategy is relatively stable. In this strategy, the K-Means clustering is more accurate than the SOM clustering as a whole.
3. In the RBF-KMeans model, the MSE gradually decreases, and the accuracy gradually increases with the number of neurons increases from 26 to 100. It indicates that increasing the number of hidden layer neurons within a reasonable range helps to train a classifier with better performance.

   As the neuron increases in a larger range, for example, from 1 to 200, the accuracy of the validation set reached its peak at 95% when the number of neurons was around 100. Then, the accuracy of the validation set begins to decline, while that of the training set is still increasing, and the accuracy reached almost 100% at the end. I think this is a manifestation of data overfitting.

Support Vector Machines neural network showed lower stability but higher overall average performance than RBF neural network (Prototype center strategy).

*Code and prediction labels for testing data are available at:
https://github.com/Accumul/EE7207Assignment.git