Homework-2 (Chapter 7. Ensemble Learning)

110321006 蔡秉翰

110321007 劉竑毅

在 MNIST dataset 中，將 Random Forest、Extra Trees 和 SVM 這三種不同的機器學習模型結合在一起，以提高對手寫數字辨識的準確性。這個應用的核心是使用 Soft Voting 策略，綜合各個模型的預測結果，從而獲得更穩健且準確的整體模型。

使用 MNIST 數據集，其中包含了手寫數字的 28x28 像素圖像。這是一個常用的基準測試數據集，用於評估機器學習模型對手寫數字的識別能力。

使用的模型:

1.Random Forest： 這是一種基於決策樹的集成方法，通過多棵樹的投票結果來進行預測。

2.Extra Trees： 與隨機森林相似，但在構建每棵樹時使用了更多的隨機性，從而提高了多樹集成的多樣性。

3.SVM： 這是一種二元分類和回歸分析的機器學習模型，被廣泛應用於圖像分類任務。

第一部分:

一、沒有對數據做處理、也沒有調整超參數(後面會有)

1.數據準備： 將 MNIST 數據集分為訓練集 50000、驗證集 10000 和測試集 10000。

2.模型訓練： 分別使用三個模型訓練模型。

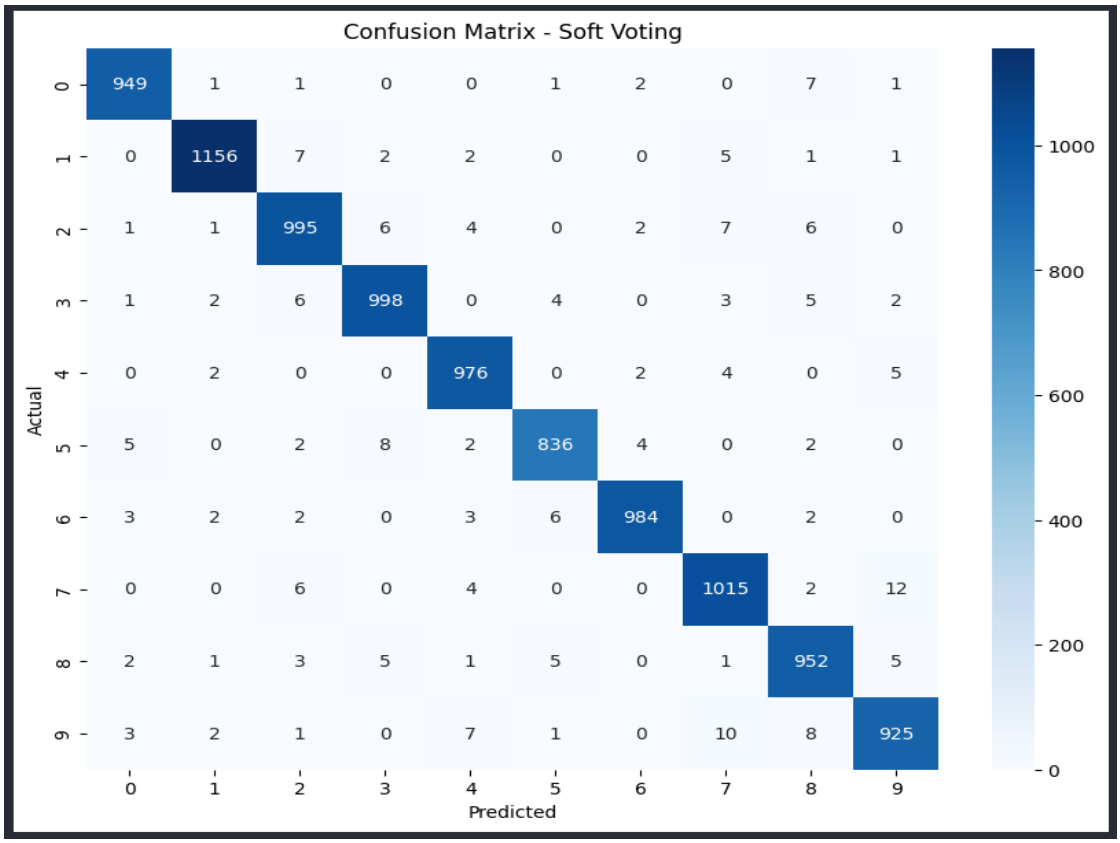
3.個別模型評估： 輸出各個模型在驗證集上的準確率。

```
RandomForest 驗證集準確率: 0.9693
ExtraTrees   驗證集準確率: 0.9715
SVM          驗證集準確率: 0.9775
```

4.集成模型構建： 使用 Soft Voting 構建集成模型，將三個模型的預測結果組合。

5.集成模型評估： 輸出集成模型在驗證集上的準確率。

```
Soft Voting 驗證集準確率: 0.9779
```



二、對數據做處理(max-min)、沒有調整超參數(後面會有)

1.數據準備： 將 MNIST 數據集分為訓練集 50000、驗證集

10000 和測試集 10000，X 的值除以 255.0。

2.模型訓練： 分別使用三個模型訓練模型。

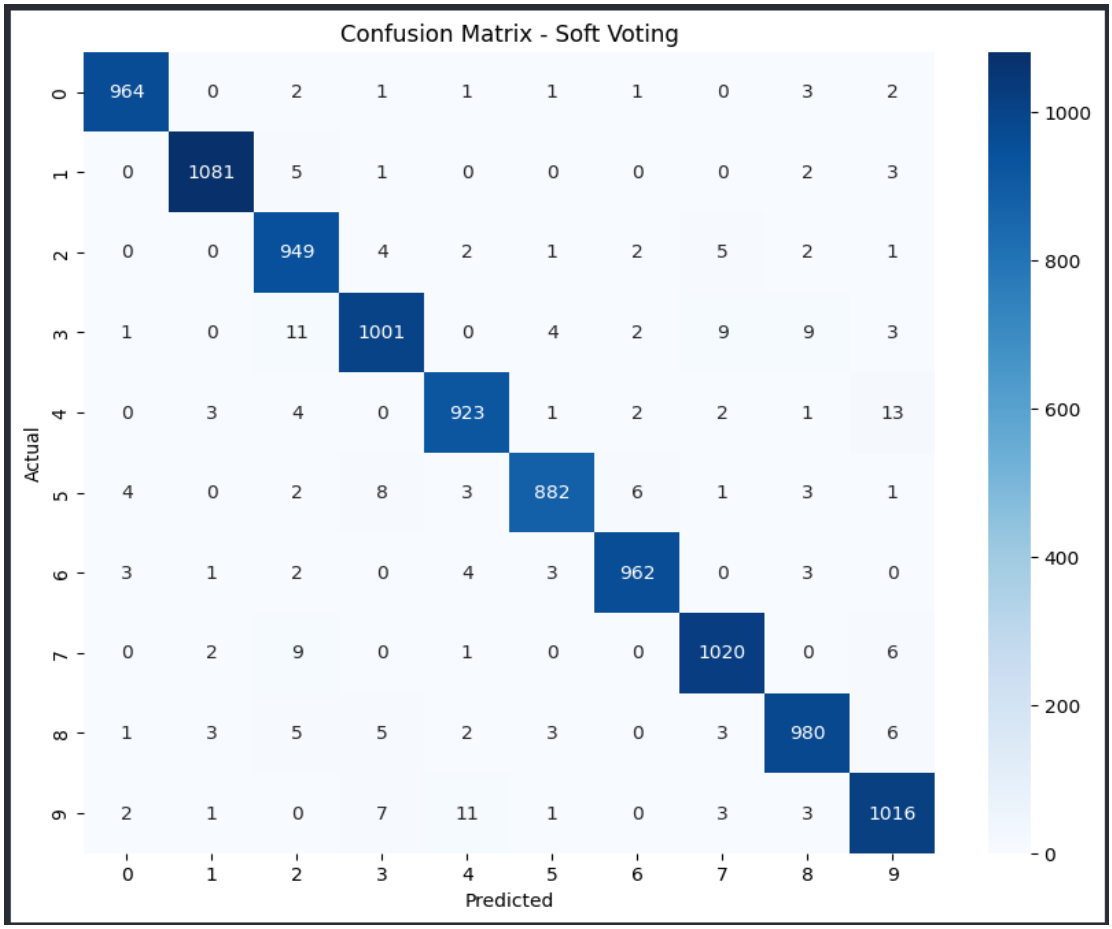
3.個別模型評估：輸出各個模型在驗證集上的準確率。

```
RandomForest 驗證集準確率: 0.9685
ExtraTrees   驗證集準確率: 0.9711
SVM          驗證集準確率: 0.9778
```

4.集成模型構建： 使用 Soft Voting 構建集成模型，將三個模型的預測結果組合。

5.集成模型評估： 輸出集成模型在驗證集上的準確率。

```
Soft Voting 驗證集準確率: 0.9786
```



三、對數據做處理(max-min)、調整超參數

1.數據準備： 將 MNIST 數據集分為訓練集 50000、驗證集 10000 和測試集 10000，X 的值除以 255.0。

2.設定超參數:使用 Grid Search CV，在我預設的超參數範圍，找出最佳參數

```
bestRandomForest:  {'max_depth': None, 'min_samples_split': 3, 'n_estimators': 200}
bestExtraTrees:    {'max_depth': None, 'min_samples_split': 2, 'n_estimators': 200}
bestSVM:           {'C': 0.1, 'gamma': 0.01, 'kernel': 'linear'}
```

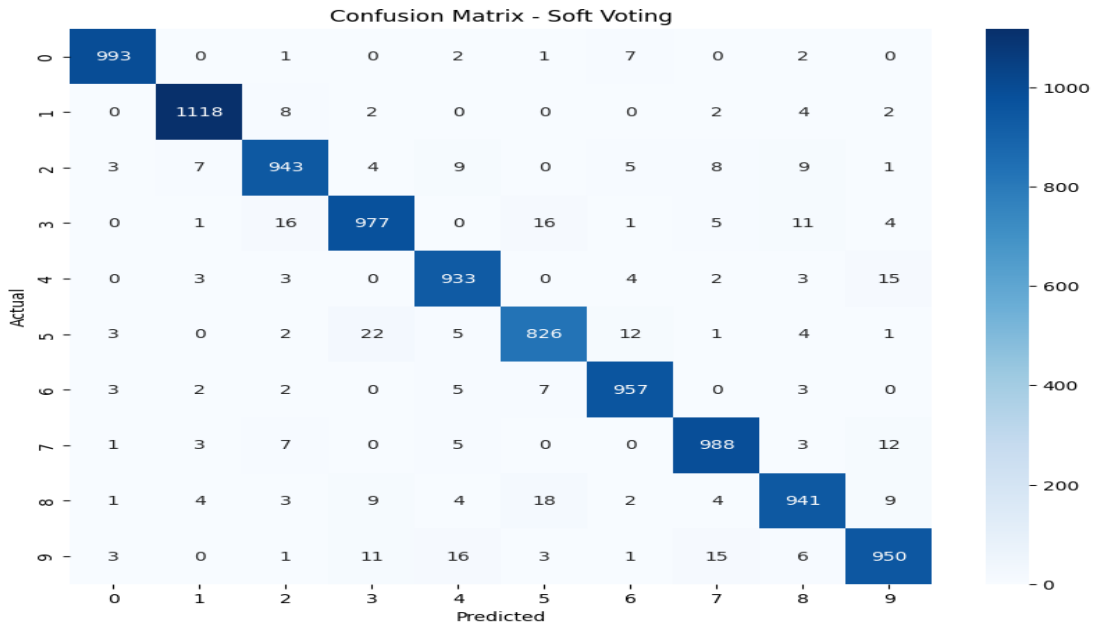
3.模型訓練：分別使用三個模型訓練模型。

4.個別模型評估：輸出各個模型在驗證集上的準確率。

```
RandomForest 驗證集準確率: 0.9696
ExtraTrees   驗證集準確率: 0.9722
SVM          驗證集準確率: 0.9451
```

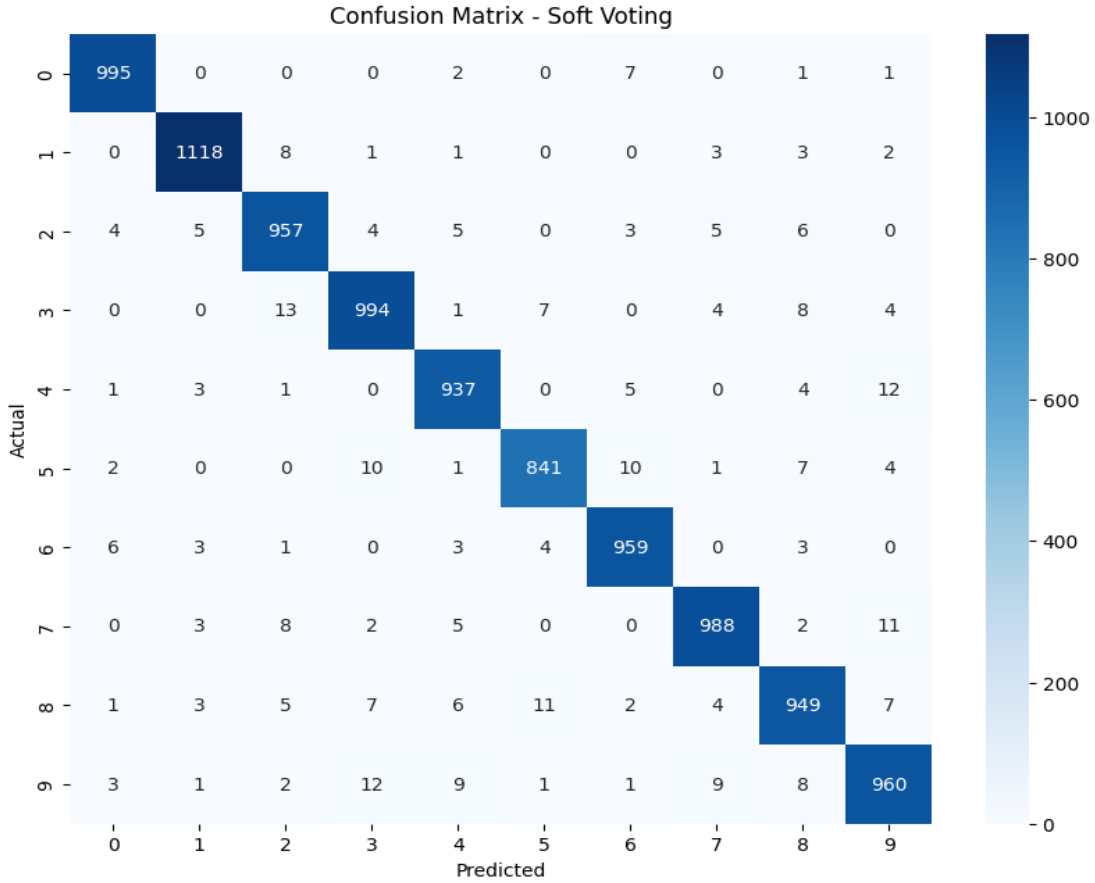
5.集成模型評估： 輸出集成模型在驗證集上的準確率。

```
Soft Voting 驗證集準確率: 0.9655
```



因為 SVM 的預設 RBF 訓練太慢了所以只用 LINEAR，又準確率只有 0.94 太差了所以我考慮 VOTING 的時候剔除 SVM，以下是剔除後的 VOTING

```
Soft Voting 驗證集準確率: 0.9711
```



以三個實驗來說，0.9779、0.9786、0.9711，以第二個實驗最高，第三個實驗搞了半天超參數，SVM 跑太久了弄不出來，可能準確率較低的伏筆就在這裡。

以第二個 SOFT VOTING 模型作為最後考試:

```
Soft Voting 測試集準確率: 0.9783
```

準確率為 0.9783

第二部分 FASHION MNIST:

一、沒有對數據做處理、也沒有調整超參數(後面會有)

1.數據準備： 將 FASHION MNIST 數據集分為訓練集 50000、驗證集 10000 和測試集 10000。

2.模型訓練： 分別使用三個模型訓練模型。

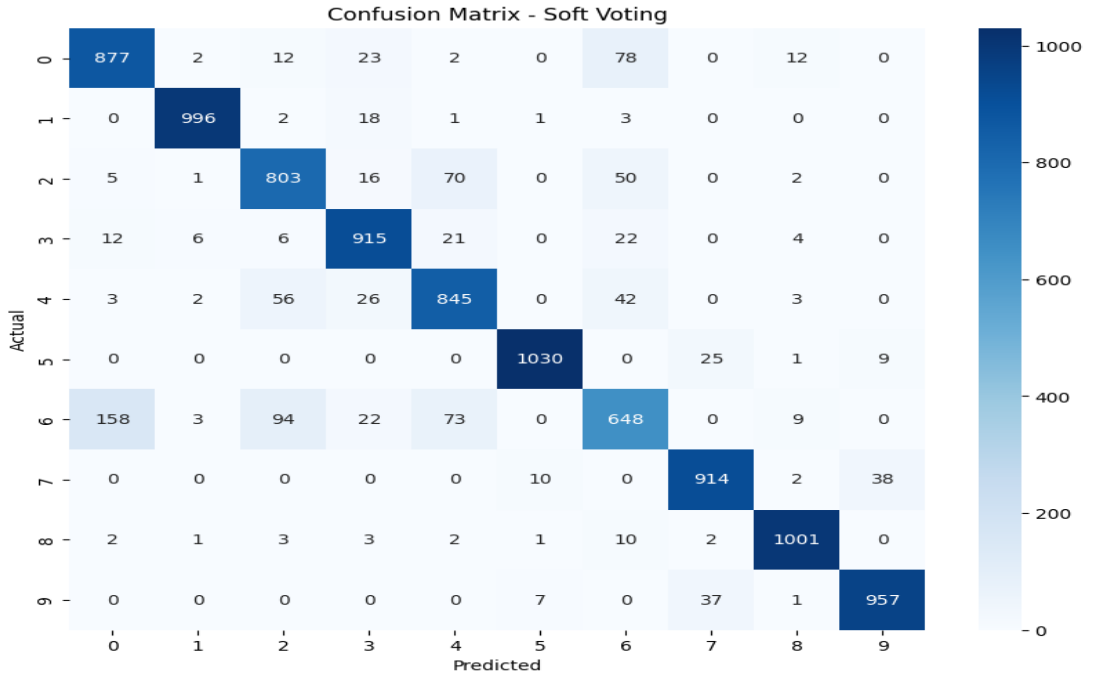
3.個別模型評估： 輸出各個模型在驗證集上的準確率。

```
RandomForest 驗證集準確率: 0.8783
ExtraTrees   驗證集準確率: 0.8812
SVM          驗證集準確率: 0.8898
```

4.集成模型構建： 使用 Soft Voting 構建集成模型，將三個模型的預測結果組合。

5.集成模型評估： 輸出集成模型在驗證集上的準確率。

```
Soft Voting 驗證集準確率: 0.8913
```



二、對數據做處理(max-min)、沒有調整超參數(後面會有)

1.數據準備： 將 FASHION MNIST 數據集分為訓練集 50000、驗證集 10000 和測試集 10000，X 的值除以 255.0。

2.模型訓練： 分別使用三個模型訓練模型。

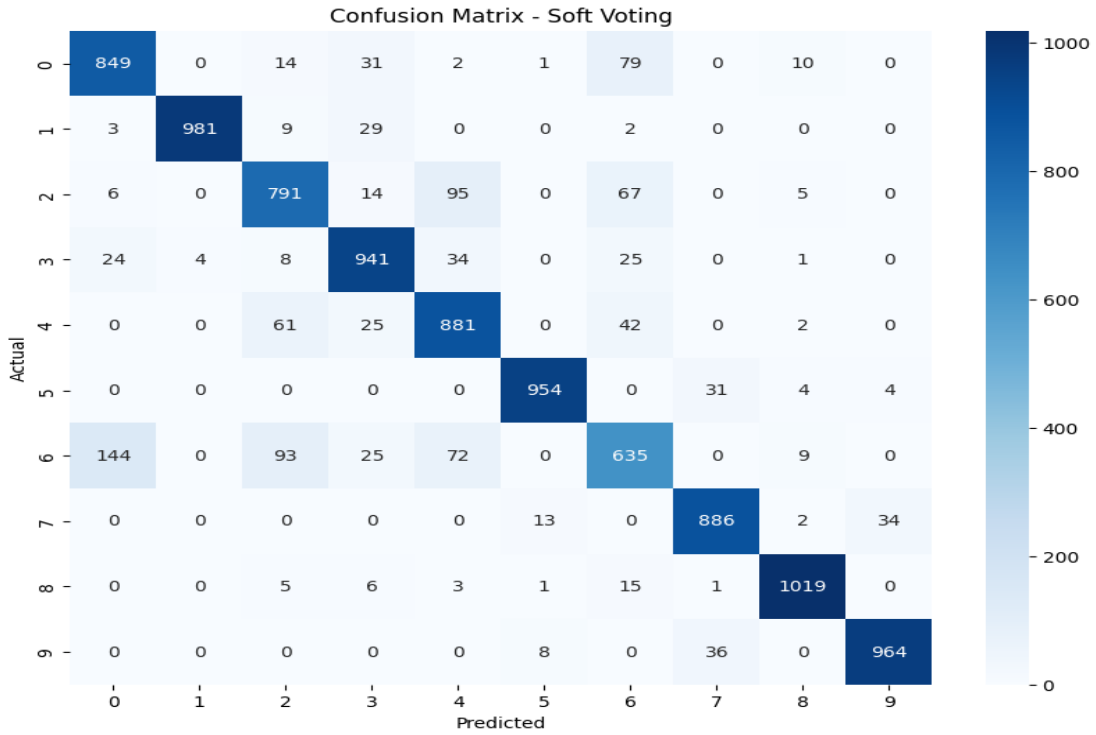
3.個別模型評估：輸出各個模型在驗證集上的準確率。

```
RandomForest 驗證集準確率: 0.8784
ExtraTrees   驗證集準確率: 0.8774
SVM          驗證集準確率: 0.8903
```

4.集成模型構建： 使用 Soft Voting 構建集成模型，將三個模型的預測結果組合。

5.集成模型評估： 輸出集成模型在驗證集上的準確率。

```
Soft Voting 驗證集準確率: 0.8931
```



三、對數據做處理(max-min)、調整超參數

1.數據準備： 將 FASHION MNIST 數據集分為訓練集 50000、驗證集 10000 和測試集 10000，X 的值除以 255.0。

2.設定超參數:使用 Grid Search CV，在我預設的超參數範圍，找出最佳參數

```
bestRandomForest:  {'max_depth': 30, 'min_samples_split': 2, 'n_estimators': 150}
bestExtraTrees:    {'max_depth': 30, 'min_samples_split': 3, 'n_estimators': 200}
bestSVM:           {'C': 0.1, 'gamma': 0.01, 'kernel': 'linear'}
```

3.模型訓練：分別使用三個模型訓練模型。

4.個別模型評估：輸出各個模型在驗證集上的準確率。

```
RandomForest 驗證集準確率: 0.88
ExtraTrees   驗證集準確率: 0.877
SVM          驗證集準確率: 0.8643
```

5.集成模型評估： 輸出集成模型在驗證集和測試集上的準確率。

```
Soft Voting 驗證集準確率: 0.8824
```

以下是剔除 SVM，同樣原因

```
Soft Voting 驗證集準確率: 0.8802
```

然而勝率並沒有提高

以三個實驗來說最強的是第二個 0.8931，做最後考試:

```
Soft Voting 測試集準確率: 0.8901
```

這次的實驗讓我深入了解了在 MNIST 和 Fashion MNIST 數據集上應用集成學習的效果。透過結合 Random Forest、Extra Trees 和 SVM 三種不同的機器學習模型，我們成功建構了一個效能優越的集成模型，並在不同的數據處理和超參數調整情境下進行了評估。

我們嘗試了不同的數據處理方式，包括未處理、標準化，以及同時進行超參數調整。結果顯示，在某些情境下，對數據進行標準化能夠提升模型的性能。同時，透過超參數的調整，我們進一步優化了模型的表現。

然而，我們也遇到了一些挑戰，例如 SVM 模型的訓練時間過長，因此我們最終選擇在 Soft Voting 時剔除了 SVM。這種權衡和嘗試的過程使我更深入地理解了不同模型和方法之間的平衡。

最後，我選擇以第二次 Soft Voting 實驗的結果作為最終結論。這次實驗在 MNIST 數據集上達到了 0.9783 的準確率，顯示了我們建構的集成模型在手寫數字辨識上有著優異的性能。對於 Fashion MNIST，最佳實驗達到了 0.8901 的準確率，儘管相對較低，但仍然顯示了集成學習在不同類型的數據上的通用性。

這次的學習經驗讓我更加熟悉機器學習的實際應用，並提升了我在數據處理、模型選擇和評估方面的能力。期待未來進一步深入探討機器學習領域，挑戰更複雜的任務和模型。