MachineLearningHomework1_KMeansPictureSegment

201250068 陈骏

环境配置

```
Python 3.7
Package:
    pandas
    skimage
    sklearn
```

项目代码

代码结构

```
D:.
|  PicRead.py
|  README.md
|  SVM.py
|
├─.idea
|  |  .gitignore
|  |  misc.xml
|  |  modules.xml
|  |  SVM.iml
|  |  workspace.xml
|  |
|  └─inspectionProfiles
|          profiles_settings.xml
|          Project_Default.xml
|
├─data
|      t10k-images-idx3-ubyte
|      t10k-labels-idx1-ubyte
|      train-images-idx3-ubyte
|      train-labels-idx1-ubyte
|
└─__pycache__
        PicRead.cpython-37.pyc
```

数据集

采用Fashion MNIST Dataset作为训练数据集，该数据集是一个使用很广泛的用于图片分类核图像识别的数据集。与MNIST手写数据集相同，该数据集的图片为28*28的灰度图片，也拥有十种分类结果。

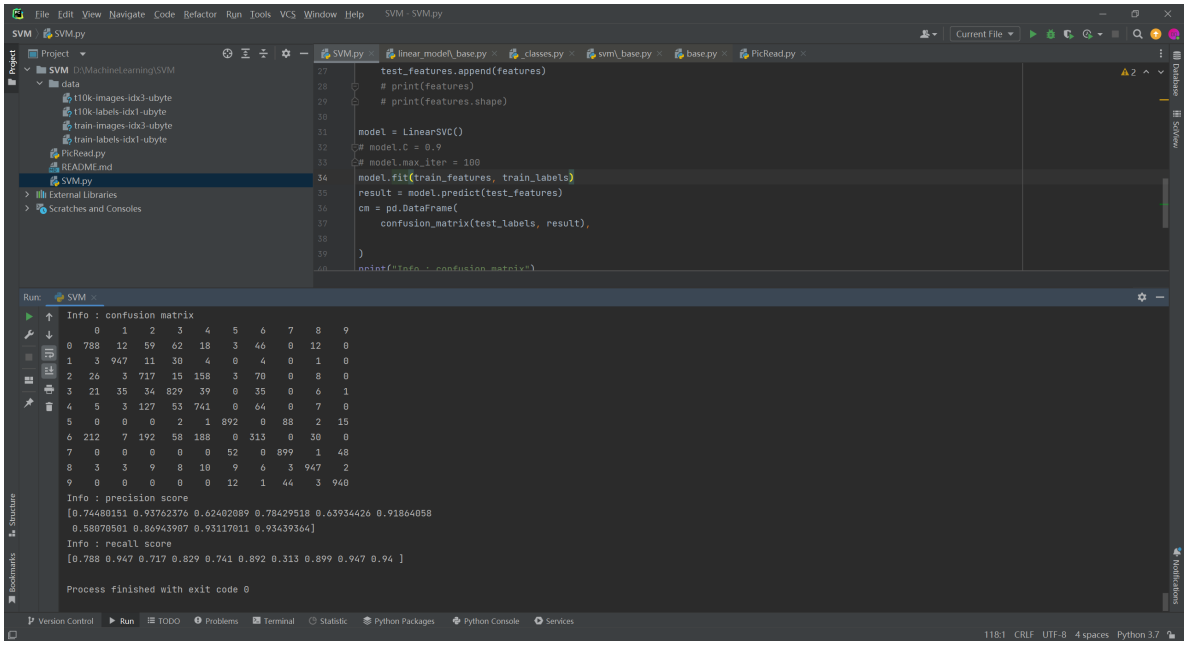
SVM实现

SVM是一个分离超平面，在二维向量的分类中，这种平面表现为多个直线将二维平面分为多个区域，在高维向量中，表现为超平面。在线性支持向量机中，一个样例会计算其到分界超平面的垂直距离，而SVM则是将所有训练样本的其离最近的超平面的距离最小化。在sklearn的LinearSVC中，由函数_fit_liblinear实现迭代过程。

代码内容

```
from PicRead import read_idx3_ubyte_pixel_file
from PicRead import read_idx3_ubyte_label_file
import pandas as pd
from skimage.feature import hog
from sklearn.metrics import confusion_matrix, precision_score, recall_score
from sklearn.svm import SVC, LinearSVC, NuSVC

train_data, (height, width) = read_idx3_ubyte_pixel_file("data/train-images-
idx3-ubyte")
train_labels = read_idx3_ubyte_label_file("data/train-labels-idx1-ubyte")
test_data = read_idx3_ubyte_pixel_file("data/t10k-images-idx3-ubyte")[0]
test_labels = read_idx3_ubyte_label_file("data/t10k-labels-idx1-ubyte")

# print(train_data)
# print(train_labels)

train_features = []
test_features = []
for i in train_data:
    patches = hog(i, orientations=10)
    patches = patches.flatten()
    train_features.append(patches)
    # print(patches)
    # print(patches.shape)
for i in test_data:
    features = hog(i, orientations=10)
    features = features.flatten()
    test_features.append(features)
    # print(features)
    # print(features.shape)

model = LinearSVC()
# model.C = 0.9
# model.max_iter = 100
model.fit(train_features, train_labels)
result = model.predict(test_features)
cm = pd.DataFrame(
    confusion_matrix(test_labels, result),

)
print("Info : confusion matrix")
print(cm)
print("Info : precision score")
print(precision_score(test_labels, result, average=None))
print("Info : recall score")
print(recall_score(test_labels, result, average=None))
```

训练结果



优化尝试

特征提取

SVM模型是对向量进行分类，因此需要将图片特征提取为向量，可以采用卷积等方法提取。此处采用skimage.feature.hog进行方向梯度直方图提取作为向量。

算法优化

选取不同的模型与参数

Linear

线性分类，惩罚力度C，为0-1之间，C越大分类效果越好，但会导致过拟合。

Rbf

高斯模型分类，Gamma值越小，分类越连续，越大越分散。

Poly

多项式分类，维度值degree，参数效果类似于前两种。