算法整体流程

将数据集提取 sift 特征，转化为 15 个特征集
对每个特征集进行 kmeans 聚类
根据聚类结果生成每张训练集的特征向量
根据特征向量训练线性核 SVM
将测试集转换为特征向量，输入 SVM 得到分类结果

函数功能说明

def trainSet2featureSet():
对训练集中每张图片提取 sift 特征，分类别保存在 15 个文件内

def feature2vector(features, centers):
features 特征集
centers kmeans 聚类得到的中心集合
将特征集转换为特征向量

def learnVocabulary():
对 trainSet2featureSet()输出的特征集进行 kmeans 聚类，聚类结果保存在 15 个文件内

def trainSVM():
使用 feature2vector(features, centers)生成特征向量，根据特征向量训练线性 SVM，训练结果保存在 svm.clf 中

def train():
整合所有训练步骤

def test():
读取 svm.clf，将训练集转换为特征向量输入 svm 得到分类结果，分类结果输出为混淆矩阵，并保存为文件

def check_result():
读取并输出 test()保存的混淆矩阵文件

输入参数说明

```
#提取的 sift 特征数目
sift_num=200
#k
wordCnt = 50
#kmeans 结束精度要求
eps=0.1
#kmeans 最大迭代次数
max_iter=20
#kmeans 重复次数
re_kmeans=3
```

混淆矩阵（参数如上）

```
[54, 1, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, 1, 1, 5]
[0, 75, 5, 2, 1, 0, 0, 1, 2, 0, 0, 0, 1, 1, 3]
[6, 3, 121, 3, 9, 0, 0, 0, 3, 0, 4, 6, 3, 1, 2]
[2, 2, 1, 47, 2, 0, 0, 0, 0, 2, 0, 2, 1, 1, 0]
[2, 2, 5, 8, 95, 4, 0, 5, 5, 1, 1, 2, 2, 2, 5]
[0, 1, 1, 1, 0, 205, 0, 0, 0, 0, 1, 0, 1, 0, 0]
[2, 0, 0, 1, 1, 0, 169, 0, 0, 0, 0, 0, 3, 1, 1]
[2, 1, 1, 4, 2, 2, 0, 89, 1, 1, 1, 3, 1, 1, 1]
[2, 3, 4, 0, 7, 1, 0, 2, 135, 0, 0, 1, 3, 0, 0]
[1, 0, 0, 1, 0, 0, 2, 3, 1, 212, 1, 1, 1, 1, 0]
[15, 0, 3, 2, 9, 1, 0, 0, 1, 2, 223, 2, 1, 1, 0]
[5, 7, 1, 6, 16, 2, 1, 1, 1, 0, 0, 96, 2, 1, 3]
[3, 5, 3, 9, 5, 0, 4, 8, 2, 1, 0, 3, 151, 1, 11]
[1, 0, 4, 5, 1, 0, 0, 2, 0, 0, 0, 0, 1, 51, 0]
[5, 6, 17, 2, 9, 0, 2, 6, 3, 0, 0, 6, 12, 2, 95]
```

性能影响分析

kmeans 环节耗时在训练中占比最大

re_kmeans 显著影响训练时间,取 20 时训练时间约 30min，取 3 时训练时间约 5min

以下是 re_kmeans=20，其他参数不变时输出的混淆矩阵

[53, 1, 2, 2, 1, 0, 0, 2, 1, 0, 0, 0, 1, 3, 0]
[4, 77, 3, 0, 0, 0, 0, 0, 0, 1, 0, 0, 2, 0, 4]
[1, 6, 128, 3, 4, 0, 0, 0, 5, 1, 1, 2, 0, 2, 8]
[1, 4, 2, 40, 6, 0, 0, 0, 0, 1, 0, 0, 1, 3, 2]
[5, 2, 6, 25, 74, 0, 0, 3, 1, 1, 5, 6, 2, 3, 6]
[3, 0, 0, 0, 0, 201, 0, 2, 0, 0, 0, 0, 1, 0, 3]
[0, 1, 1, 3, 0, 0, 168, 0, 2, 0, 0, 0, 0, 0, 3]
[0, 1, 1, 5, 2, 6, 0, 90, 0, 2, 0, 0, 1, 1, 1]
[2, 3, 6, 4, 4, 1, 1, 0, 127, 0, 1, 2, 0, 2, 5]
[1, 0, 5, 1, 3, 0, 0, 2, 0, 203, 0, 1, 5, 1, 2]
[1, 0, 5, 10, 5, 1, 0, 0, 0, 0, 231, 3, 2, 0, 2]
[1, 0, 3, 6, 7, 0, 0, 3, 2, 0, 0, 113, 0, 2, 5]
[18, 12, 8, 13, 5, 2, 3, 2, 6, 6, 4, 5, 107, 12, 3]
[0, 1, 2, 3, 4, 1, 0, 0, 0, 0, 0, 1, 1, 51, 1]
[2, 6, 20, 4, 6, 1, 3, 1, 0, 3, 2, 7, 5, 6, 99]

值得注意的是,kmeans 重复次数的增加对某些分类的分类准确度产生了负面影响。

max_iter 也会显著影响训练时间，取 30 时训练时间约 10min，取 20 时训练时间约 5min

以下是 max_iter=30 时的混淆矩阵

[50, 2, 4, 2, 4, 0, 0, 1, 1, 0, 0, 2, 0, 0, 0]
[2, 77, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 7, 1]
[8, 1, 134, 5, 1, 1, 0, 2, 2, 3, 0, 0, 1, 2, 1]
[6, 3, 2, 38, 2, 0, 1, 1, 0, 1, 0, 1, 0, 4, 1]
[7, 4, 7, 5, 84, 4, 0, 0, 7, 0, 4, 3, 6, 5, 3]
[5, 0, 0, 0, 0, 204, 0, 0, 0, 0, 0, 0, 0, 0, 1]
[0, 0, 0, 0, 1, 0, 174, 0, 1, 0, 0, 0, 1, 0, 1]
[3, 3, 0, 4, 4, 2, 0, 88, 2, 0, 0, 1, 1, 1, 1]
[2, 1, 5, 0, 8, 1, 0, 1, 135, 0, 1, 2, 2, 0, 0]
[0, 1, 2, 0, 1, 1, 0, 0, 1, 214, 1, 0, 3, 0, 0]
[5, 1, 0, 2, 9, 1, 0, 4, 3, 1, 225, 2, 4, 1, 2]
[16, 12, 8, 8, 15, 0, 1, 1, 2, 0, 3, 67, 5, 0, 4]
[10, 5, 8, 6, 18, 2, 1, 5, 6, 2, 3, 8, 115, 4, 13]
[1, 4, 0, 6, 1, 0, 0, 1, 0, 0, 0, 0, 3, 47, 2]
[15, 4, 8, 3, 4, 0, 3, 1, 12, 0, 1, 6, 9, 3, 96]

调参结果和 re_kmeans 增加的效果很相似，对某些类别的分类结果产生了负面影响

wordCnt 影响训练时间的同时比较明确的的影响分类准确率
以下是 wordCnt=30 时的混淆矩阵
[44, 2, 2, 1, 3, 2, 0, 3, 0, 1, 0, 2, 2, 0, 4]
[2, 60, 5, 0, 1, 2, 0, 2, 5, 2, 0, 6, 2, 2, 2]
[14, 5, 81, 26, 5, 0, 0, 8, 5, 6, 0, 6, 3, 1, 1]
[4, 0, 1, 47, 1, 0, 0, 0, 1, 1, 0, 1, 2, 0, 2]
[30, 3, 17, 4, 65, 1, 1, 0, 2, 1, 0, 9, 0, 2, 4]
[1, 2, 1, 0, 0, 199, 0, 2, 0, 1, 0, 1, 1, 2, 0]
[1, 0, 0, 1, 1, 0, 164, 0, 0, 0, 0, 0, 1, 2, 8]
[6, 1, 6, 4, 5, 1, 0, 82, 0, 2, 1, 0, 1, 0, 1]
[2, 6, 0, 1, 2, 1, 0, 4, 138, 0, 1, 0, 0, 0, 3]
[3, 2, 24, 0, 5, 3, 1, 13, 9, 155, 2, 3, 3, 0, 1]
[1, 4, 5, 0, 3, 10, 1, 1, 1, 0, 224, 1, 3, 1, 5]
[10, 14, 6, 0, 23, 0, 0, 0, 1, 0, 0, 77, 4, 6, 1]
[24, 16, 34, 13, 8, 7, 6, 7, 5, 5, 6, 8, 59, 4, 4]
[1, 0, 0, 4, 2, 4, 0, 0, 1, 0, 2, 1, 1, 48, 1]
[11, 0, 0, 7, 17, 2, 1, 8, 12, 4, 9, 3, 2, 3, 86]
wordCnt 下降时可以明显观察到大部分类别分类准确率的下降

sift_num 显著影响分类准确率
以下是 sift_num=100 产生的混淆矩阵
[35, 0, 0, 12, 5, 0, 0, 1, 2, 1, 2, 1, 5, 2, 0]
[3, 66, 2, 1, 4, 0, 0, 1, 2, 0, 1, 7, 2, 0, 2]
[6, 5, 111, 10, 5, 0, 0, 2, 2, 0, 0, 2, 5, 2, 11]
[9, 1, 0, 32, 2, 1, 0, 0, 4, 1, 1, 2, 3, 2, 2]
[15, 9, 4, 14, 71, 1, 2, 2, 2, 1, 2, 5, 7, 2, 2]
[5, 1, 2, 3, 5, 188, 1, 1, 0, 0, 0, 0, 1, 2, 1]
[3, 2, 1, 1, 12, 0, 154, 1, 3, 0, 0, 1, 0, 0, 0]
[3, 1, 1, 1, 3, 0, 0, 90, 0, 1, 1, 5, 0, 2, 2]
[1, 7, 2, 12, 15, 3, 1, 3, 87, 3, 0, 7, 7, 7, 3]
[1, 0, 2, 4, 4, 0, 0, 3, 2, 202, 0, 0, 2, 3, 1]
[4, 1, 9, 8, 13, 1, 1, 2, 0, 1, 203, 5, 5, 3, 4]
[1, 4, 1, 10, 4, 1, 2, 2, 7, 2, 4, 93, 3, 3, 5]
[11, 7, 14, 11, 12, 6, 4, 3, 7, 8, 3, 7, 97, 7, 9]
[3, 3, 1, 3, 2, 3, 0, 0, 3, 2, 1, 1, 6, 36, 1]
[5, 1, 1, 11, 14, 0, 0, 4, 3, 4, 13, 12, 12, 2, 83]
明显观察到各类别分类准确率大幅下降