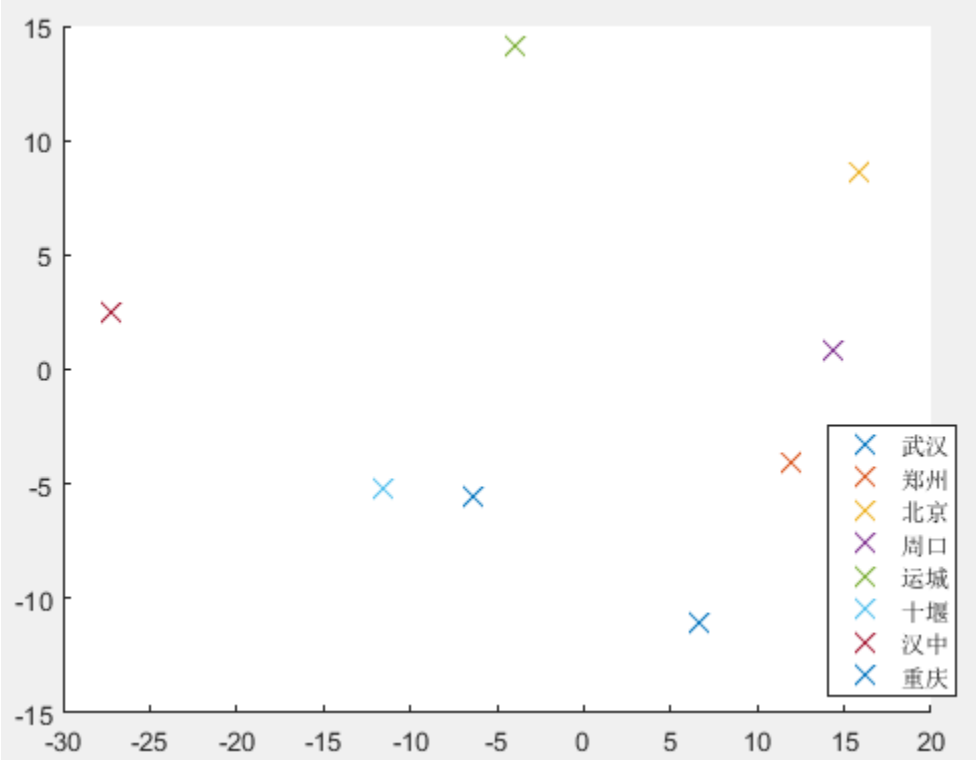
模式识别第八次作业

数 41　李博扬　2014012118

1、代码见压缩包



分析：位置有些许偏差，但仅从交通通勤时间表能推出这样的大致方位已经十分不错了。可以看到，北京与运城的上下偏差较大，重庆位置过于低纬度，周口应在郑州东南方向。造成这些不准确的原因极大的在于，火车通勤时间与距离并不等价，有些火车快，有些火车慢或者有些火车停站时间长等等都会造成很大的误差。

2、

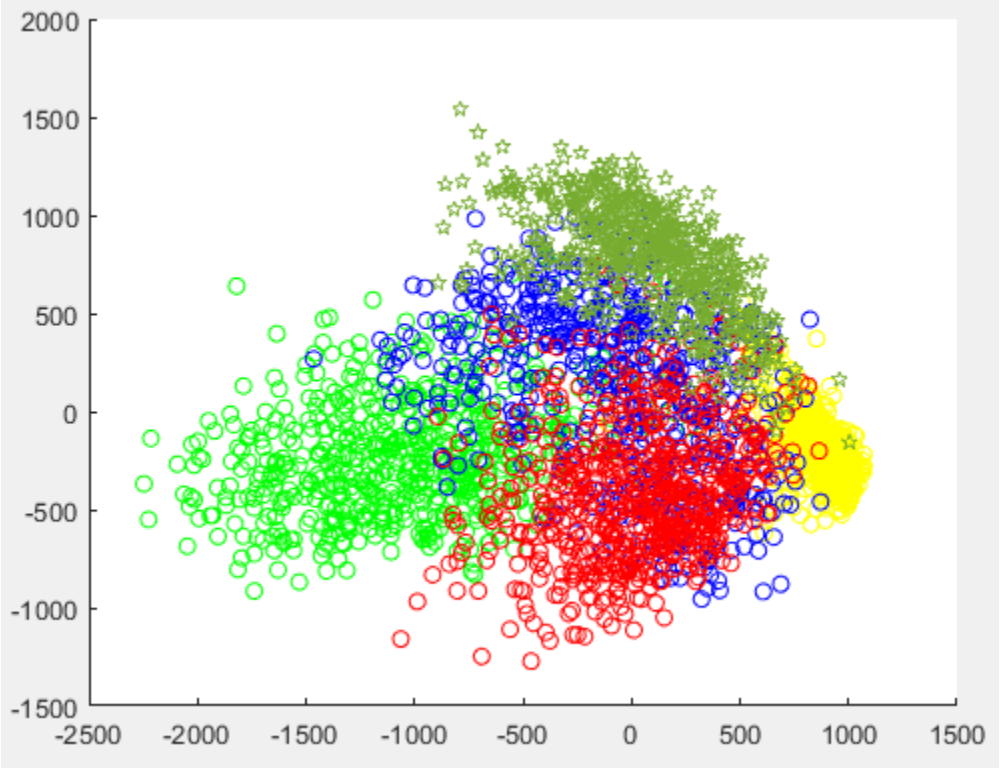
a．PCA：得到数据—计算协方差矩阵—计算协方差矩阵的特征向量和特征值—选择成份组成模式矢量—得到降维后的数据

LLE：寻找每个样本点的 k 个近邻点—由每个样本点的近邻点计算出该样本点的局部重建权值矩阵—由该样本点的局部重建权值矩阵和其近邻点计算出该样本点的输出值

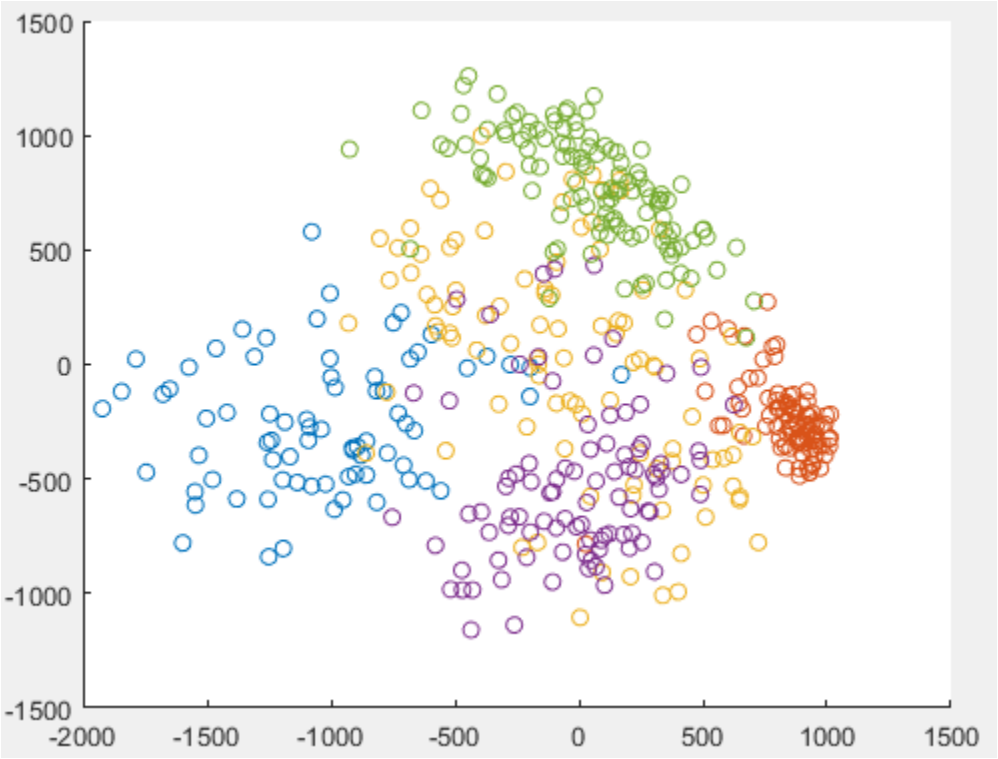
b.

PCA：

trainX 降维：

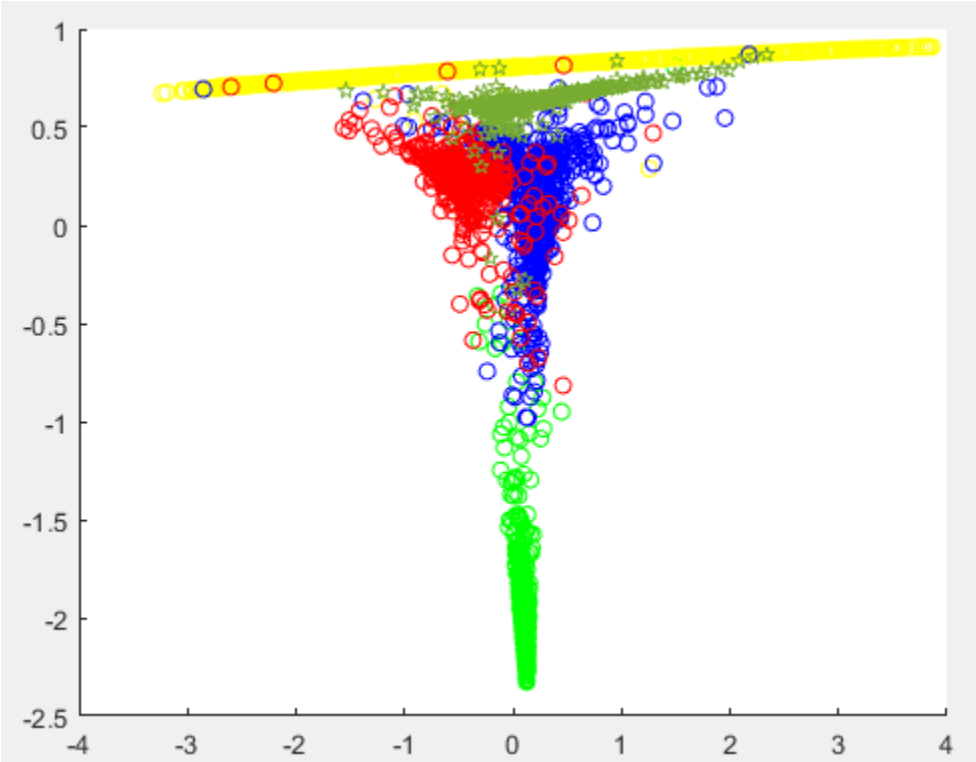


testX 降维：

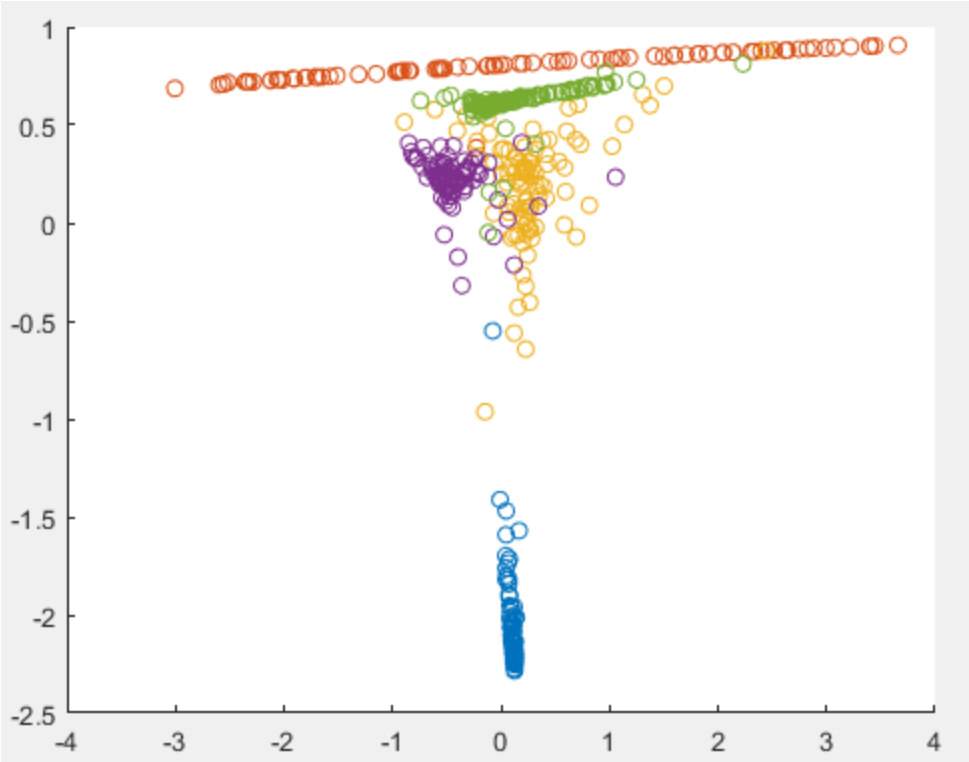


LLE：

trainX 降维



testX 降维



c．选择线性 SVM 分类器，PCA 维数降到 133，LLE 降到 3

PCA：在训练集上正确率 96%

在测试集上正确率 96.61%

```
>> sum(trainedClassifierpca.predictFcn(testXpca)==testY)/472

ans =

    0.9661
```

LLE：在训练集上正确率 95.6%

在测试集上正确率 95.34%

```
>> sum(trainedClassifierlle.predictFcn(testXLLe)==testY)/472

ans =

    0.9534
```

不处理：在训练集上正确率 97.9%

在测试集上正确率 97.88%

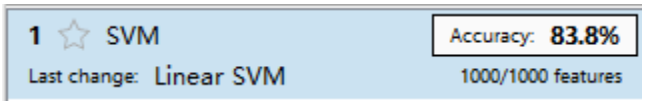
```
>> sum(trainedClassifier.predictFcn(testX)==testY)/472

ans =

    0.9788
```

由上可知，降维之后分类的正确率有所下降，但是下降并不明显。LLE 处理效果略优于 PCA，处理降维之后的数据大大提高了处理效率，可见合理降维是必要的也是可行的。

d. PCA 是最常用的线性降维方法，它的目标是通过某种线性投影，将高维的数据映射到低维的空间中表示，并期望在所投影的维度上数据的方差最大，以此使用较少的数据维度，同时保留住较多的原数据点的特性，但是 PCA 并不试图去探索数据内在结构。LLE 是一种非线性降维算法，它能够使降维后的数据较好地保持原有流形结构。但是 LLE 在有些情况下也并不适用，如果数据分布在整个封闭的球面上，LLE 则不能将它映射到二维空间，且不能保持原有的数据流形。那么我们在处理数据中，首先假设数据不是分布在闭合的球面或者椭球面上。从题中可以看出，LLE 降维幅度比 PCA 更大，测试的效果却更好，说明数据是具有一定流行结构的，PCA 在保持流行结构方面不如 LLE。

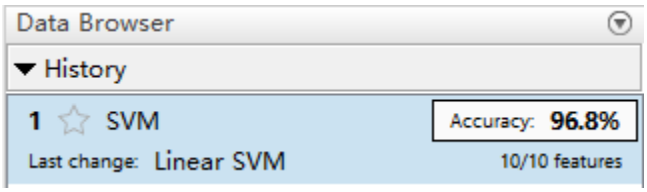
3、

a. 不做特征选择时，用 SVM 分类器选择 10 倍交叉验证分类正确率为：83.8%



b.作特征选择，对每一个特征列，计算与 Y 的相关系数，按相关系数由大到小排序，取前 10 个特征作为新的数据集合。对应的列标号为 48、917、220、5、416、428、386、104、233、1000。（即为代码中 index 前 10 个）

对新数据集用 SVM 分类器，选择 10 倍交叉验证分类正确率为：96.8%



可以看见，虽然特征选取的方法很简单，仅仅依靠相关系数的排序，但是依旧排除了大量与指标集合 Y 无关的特征，大大简化的数据处理难度，提高了效率。特征选择之前，大量无关特征影响了分类的正确性，特征选择之后，正确率提高了 13%，由此可见特征选取是提高学习算法性能的一个重要手段，也是模式识别中关键的数据预处理步骤。