

人脸识别与主成分降维

贾文鑫 51184407128

2019 年 1 月 15 日

摘要

本文试图对 Olivetti Faces Dataset 人脸数据集进行判别分类，并使用主成分分析方法对数据降维，最后根据准确率、kappa 值和运行时间比较降维前后的各种分类方法。得出结论，LDA 方法的分类效果最优，KNN 方法的分类效果次之，但是运行时间显著快于其他几种方法；数据降维之后，各种分类方法的运行时间有显著降低，且分类效果没有受到较大的影响。

本文包含两张图片，两张表格。

关键词：人脸识别；LDA；KNN；主成分分析

第一章 引言

1.1 研究背景

人脸识别是当前的一个热门的研究方向，对人脸图像进行分类是其中一个较大的分支，但是图像数据一般都是维数特别高的矩阵形式，所以在人脸图像分类中，常常会使用降维方法。本文尝试运用不同的分类器对人脸数据进行判别分析，并对降维前后的效果进行对比。

1.2 研究方法

本文采用了 LDA、决策树、KNN 和朴素贝叶斯四种方法对 Olivetti Faces Dataset 人脸数据进行分类，采用主成分分析来对数据降维，采用准确率、kappa 值和运行时间作为评价指标。

LDA 方法为最大似然判别法中采用线性法则的一种分类方法，最大似然判

别法通过确定 X 的最大似然来预测观测 x 的类。

决策树方法是一种简单且广泛使用的分类技术，因为这种方法可以直观地展现分类的过程，可视化效果特别好，但是需要权衡分类正确率和树的大小。

KNN 方法基于测量两个观测之间的距离，KNN 法则分类一个观测 x 的方法为在训练集中找到 k 个与 x 最近的训练点，采取最多投票制分类 x ：即将 x 分类到包含其 k 个最近邻训练点最多的类里，最优的 k 采用交叉验证方法选择。

朴素贝叶斯方法是基于贝叶斯定理与特征条件独立假设的分类方法，发源于古典数学理论，有着坚实的数学基础，以及稳定的分类效率。同时，朴素贝叶斯方法所需估计的参数很少，对缺失数据不太敏感，算法也比较简单。但是朴素贝叶斯方法基于属性之间相互独立的假设在实际应用中往往是不成立的。

主成分分析（PCA）通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量，转换后的这组变量叫主成分，选取使累计方差贡献率达到一定程度的前 n 个主成分达到数据降维的目的。

准确率即为正确分类的观测占总观测的比例； $kappa$ 值是一种衡量分类精度的指标， $k = (p_0 - p_e) / (1 - p_e)$ ， p_0 是总体分类精度， p_e 随机猜测的分类精度；运行时间为每种分类方法在 R 语言中的运行时间。

1.3 论文结构安排

本文结合本课程学习的各种分类方法，对人脸数据进行判别分类和降维，并且选取适当的指标进行评估。本文共分为三个部分，第一章介绍了选题背景，研究方法。第二章是对降维前后的各种分类方法进行比较。第三章对本文各种方法进行总结。

第二章 数据分析

2.1 数据读取

将 Olivetti Faces Dataset 人脸数据读取到 R 中，查看人脸数据中的第一行，将其转换为 64×64 的图片矩阵，图片如下：

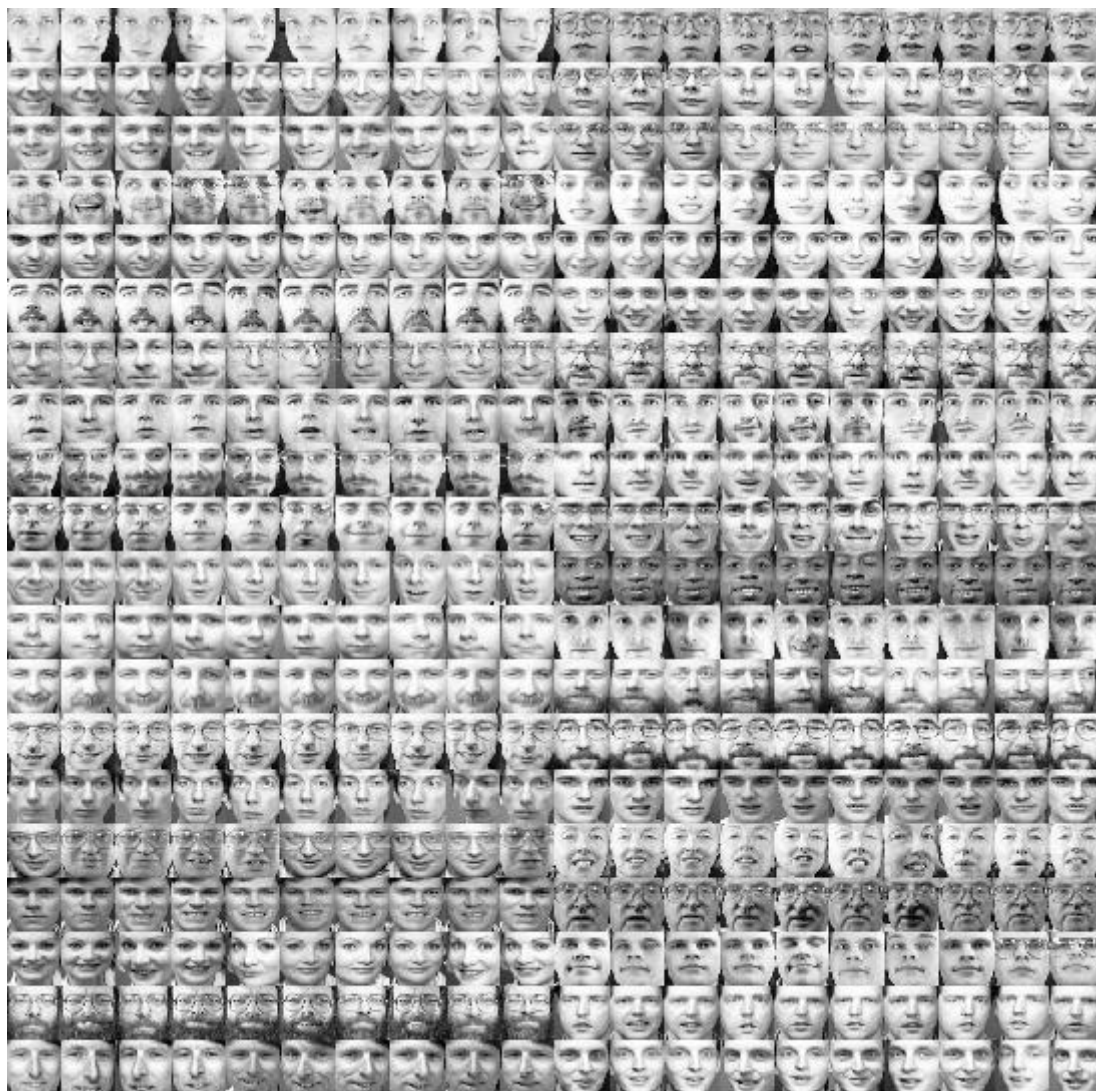


图 1 人脸图

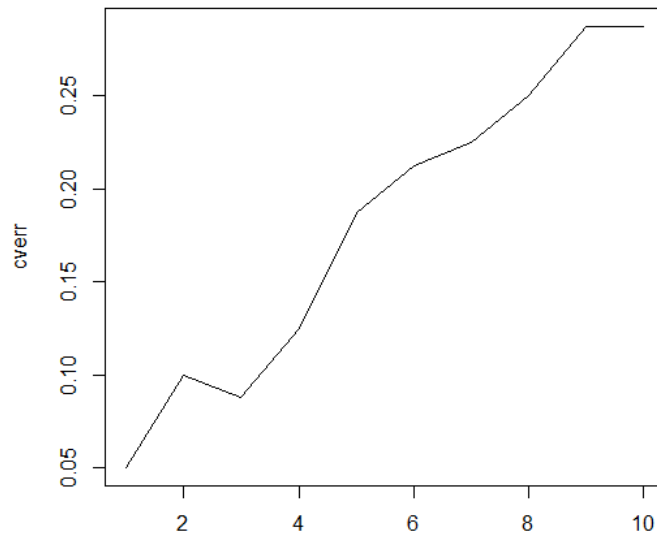
将数据按照不同的人分为 40 类，每类共有 10 张每个人不同姿态下的脸图。随机选择每类数据中的 80% 建立训练集，20% 建立测试集，训练集用于建立模型，测试集用于检验模型效果。

2.2 模型效果评估

本文选取了四种常用的方法进行分类，用正确率、kappa 值和运行时间这三个指标来评价分类方法的效果。

其中 KNN 方法的最优参数 $k=1$ ，参数 k 在测试集中的错误率如下图所示。

图2 KNN最优参数k



四种分类方法的分类效果如下表所示：

表 1 分类方法效果评价表

分类方法	正确率	kappa 值	运行时间(秒)
LDA	1	1	8.02
决策树	0.4625	0.4487	20.81
KNN(k=1)	0.9375	0.9359	0.72
朴素贝叶斯	0.925	0.9231	7.48

从上表可以看出 LDA 方法的正确和 kappa 值最高，达到了 1 的完美程度，但是运行时间较长，KNN 方法的正确率和 kappa 值都排在第二位，但是运行时间特别快，决策树方法的效果最差，每项指标都排在最后。

2.3 主成分降维

因为每一张图片都是 64×64 的矩阵，拉直之后有 4096 维，所以有必要对其进行降维处理，选取前 70 个主成分使得累计方差贡献率达到 90%，将原始数据降维，按照原先的训练集和测试集的分类方法再次建立训练集和测试集。与降维之前的分类效果进行对比，结果如下：

表 2 降维前后比较表

分类方法	未降维			降维后		
	正确率	kappa 值	运行时间(秒)	正确率	kappa 值	运行时间(秒)
LDA	1	1	8.02	1	1	0.06
决策树	0.4625	0.4487	20.81	0.4	0.3846	0.35
KNN(k=1)	0.9375	0.9359	0.72	0.95	0.9487	0.01
朴素贝叶斯	0.925	0.9231	7.48	0.9875	0.9872	0.17

从上表可以看出，经过降维之后，每一种分类方法的运行速度都有显著的降低，而且 KNN 方法和朴素贝叶斯方法的正确率和 kappa 值也有所提升，降维对分类的效果有显著的提升。

第三章 总结

本文所使用的四种分类方法的效果具有明显的差异，LDA 方法的正确率和 kappa 值达到了 1 的完美程度，但是运行时间并不占优；决策树方法各类指标都是最差的，可以认为决策树方法不适合处理人脸数据；KNN 方法的正确率和 kappa 值都处于较高的水平，而且运行时间明显快于其他几种方法；朴素贝叶斯方法正确率和 kappa 值与 KNN 方法处于同一水平，但是运行时间相比于 KNN 方法没有优势。经过降维处理之后，每一种分类方法的运行速度都有显著的降低，尤其是 LDA 方法在降维之后运行时间降低到了和 KNN 相近的水平，降维的效果非常好。

综合来看，LDA 方法是最优的。虽然 KNN 方法的正确率和 kappa 值很高，运行时间也很短，但是人脸识别对准确率的要求特别高，而且经过降维处理之后，LDA 方法的运行时间下降的非常明显，而准确率并没有受到影响，所以 LDA 方法更适合处理此类数据。

本文尝试使用 K 均值方法对 Olivetti Faces Dataset 人脸数据进行聚类分析，但是最终效果特别不理想，故没有在本文中展示，查看相关文献发现改进的 K 均值方法聚类效果有显著提升，课后将会继续研究。

参考文献

- [1] Robert I. Kabacoff. R 语言实战[M].北京:人民邮电出版社,2016.
- [2] 周志华. 机器学习[M].北京:清华大学出版社,2016.
- [3] 美团算法团队. 美团机器学习实践[M].北京:人民邮电出版社,2018.
- [4] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning [M]. Springer,2009.
- [5] 谢娟英,高红超,谢维信.K 近邻优化的密度峰值快速搜索聚类算法[J].中国科学:信息科学,2016,46(02):258-280.

- [6] Bien, J., and Tibshirani, R. (2011), Hierarchical Clustering with Prototypes via MinimaxLinkage, The Journal of the American Statistical Association.
- [7] Verma, T.; Sahu, R.K., (2013) PCA-LDA based face recognition system & results comparison by various classification techniques, Green High Performance Computing (ICGHPC), 2013IEEE International Conference on , vol., no., pp.1,7, 14-15.
- [8] Calandriello D., Niu G., Sugiyama M. (2013), Semi-Supervised Information-MaximizationClustering.