Olivetti人脸数据的分析

PB11000333 黄庆

2014 年 6 月 17 日

摘要

本文是对灰度脸图数据的聚类的分析，由于数据是400x4096 的矩阵，我们首先要做的是通过高维数据可视化来观察数据是否具有明显的类。然后尝试层次聚类、k-means算法、k-medoids算法和谱聚类的聚类方法,由于可以直观地认定每个人的10张照片就是一类，称之为标准分类，从而可以有效的比较各种聚类质量。其次分析聚类产生错误的因素和通过主成分的方法去除噪声的影响来和之前的结果比较，最后我们出于好奇，考虑单独对人的眼睛进行研究，意外地得出比较好的分类，即眼睛是人脸非常重要的特征

1 简介

Olivetti Faces Dataset(http://cs.nyu.edu/ roweis/ data.html)包含了一些人的脸图, 该数据是AT&T剑桥实验室于1992 年4 月到1994年4月期间采集的. 该数据集包含了40位人的灰度脸图, 每张图的分辨率为64x64, 每个人共有10张不同姿态下的脸图. 有些人的脸图是在不同时间点,不同光照,以及不同脸部表情(睁/闭眼睛, 微笑/不笑)下拍摄的. 所有图片都是在黑色均匀背景下拍摄. 我们从该数据集提取出来变量faces 的值并存储为*faces.txt*, 其为为400x4096矩阵. 本报告以分析该数据为目的.脸图如图1。

该数据集被广泛用于验证统计学习方法的性能, 由于每个人仅有10次重复观测, 因此更适合用于无监督的或者半监督的统计学习方法中. 例如Bien, J., and Tibshirani, R. (2011), Verma et al. (2013), Calandriello et al. (2013)等等.

2 高维数据可视化

2.1 通过选取主成分可视化

主成分前3维占的比重是47.5% ，我们可以尝试去观察这样的可视化效果。3维主成分拟合图片如图2。尽管这样的图片失真较为严重，但作为前期的定性估计，不失为一种方式，我们可以观察每个人的数据点是否都较为集中，这样比较适合去估计聚类的效果。3维的主成分数据点如图3.虽然我们无法很好区别分类，但是图上同样颜色的点通常都是成团的，所以我们可以认为，聚类处理是合理的。

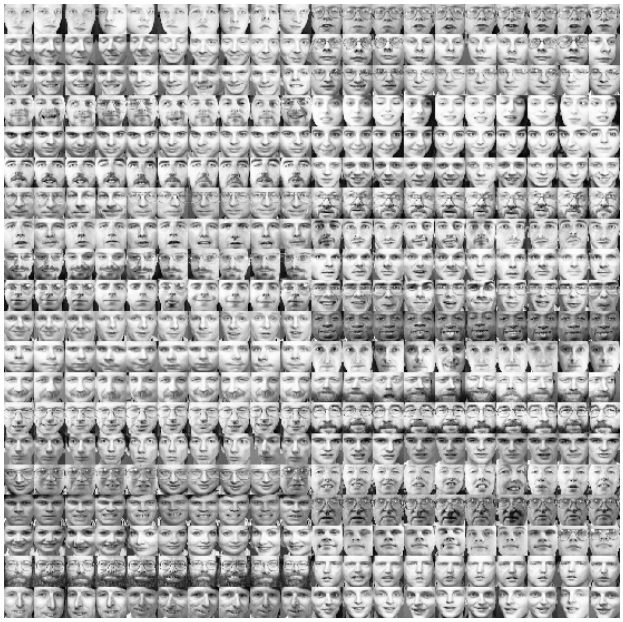

图 1: 脸图图像

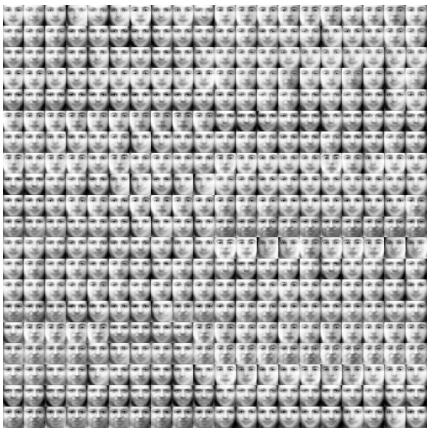

图 2: 取三个主成分的拟合图片

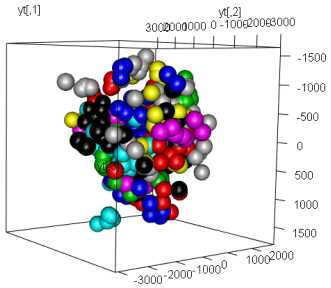

图 3: 3维主成分的数据点图

2.2 平行坐标图

在之前的学习中有一种更加直观地方式去观察这样的问题。R中的parallelplot函数可以观察较高维的数据，为了方便观察我们提取脸图数据的前20个主成分（可解释方差的百分比为76.3%，对图片还原度很高），做出前5个人的所有图片的数据集的平行坐标图如图4：

那么，在这里我们可以看到的是对于有些人的数据点而言10张照片具有非常明显的相像，10张照片波动性质一致，而有些人的数据则和别的人数据容易混合，体现在中部有些曲线的杂乱无章，这也解释了后面的聚类分析中同一个人的数据错位。

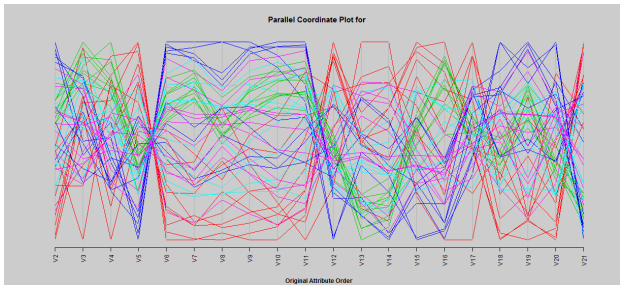

图 4: 前5个人的平行坐标图

2.3 自主制映射聚类

为了更好地可视化聚类，在属性数据分析中有一种自主制映射的办法，即Self-Organizing Maps(SOM),它通过“无监督学习”将高纬度的数据进行处理后再以低维视图表达分析结果，而映射图上保留原输入样本空间的拓扑性质。

算法如下

(1) 初始化：随机选择一个向量作为W_j的初始值

(2) 抽样：在输入空间中随机选择一个点x

(3)匹配：找到胜出神经元$l(x)$, 其权值向量最靠近输入点x

(4) 更新：对权值进行更新

(5) 循环：返回到第二部直至特征空间不再变化

通过这种方式实现的可视化如图5 相比于之前两张图，这张图更加清楚地看到数据集呈现明显

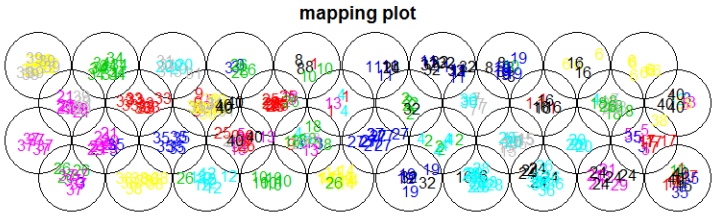

图 5: 自主制映射聚类的结果图

的抱团，所以我们对它进行聚类分析是合理的。

3 聚类方法的比较

在这里提出四种聚类的方法，分别是层次聚类、K-means，K-medoids和谱聚类的方法,我们将后三种方法放在一起比较。

3.1 层次聚类

类间联系程度度量采用欧式距离和average linkage.对全部数据的层次聚类结果如图6.

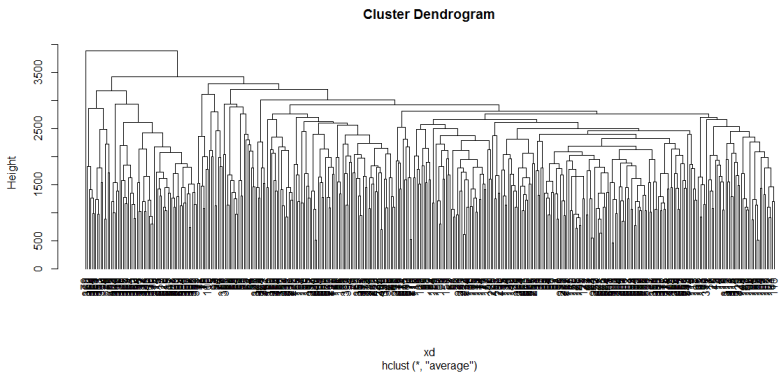

图 6: 全局层次聚类结果图

由于层次聚类分支较多，数据较大，图6的观察显得较为困难。这里只能将一些小样本展示出来，图7和图8是对第一个人和第二个人的层次聚类分析

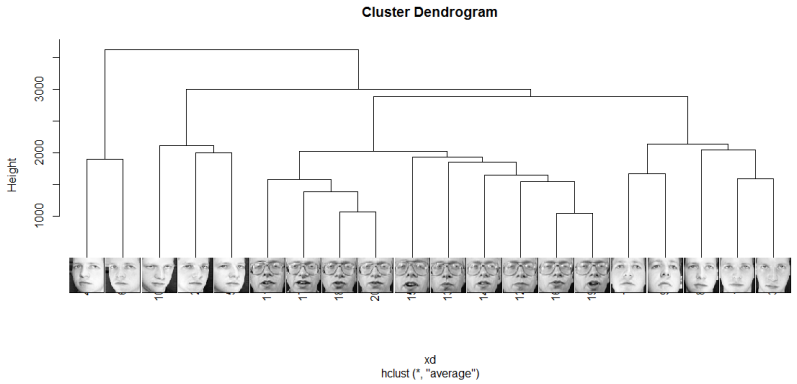

图 7: 1,2两人层次聚类结果图

图7和图8的结果基本一致，并由此可以知道，由于同一个人的照片具有很强的相似性，所以同一个人的照片通常就在同一个分支下面，但是也有例外的情况，这里的人1就是会出现不同的分支，因为该人在不同的照片中脸的朝向不同，导致1人的照片被放在不同的分支下，不过也可以看出相同分支下1 人的脸的朝向是相同的。这里很好的解释了下面的聚类结果中，1这个人总是会被

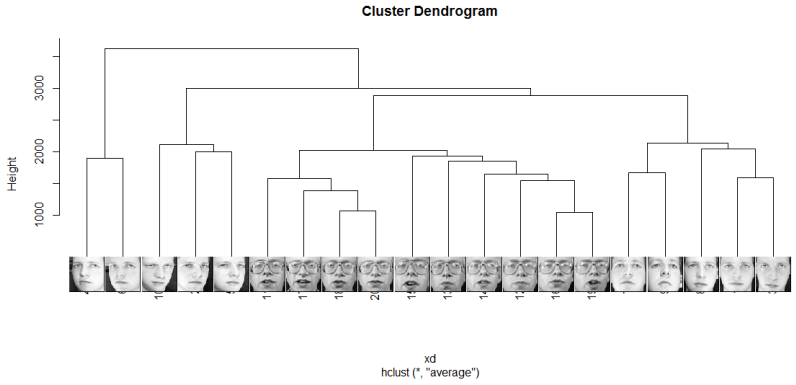

图 8: methond by bien,J.,and Tibshirani, R.(2011)

拆开，而无论是什么方法，2人的图片始终都是在一起的。

3.2 K-means和K-medoids聚类

3.2.1 k-means聚类

对数据集用K-means进行聚类，根据人数为40个人，我们把数据分成40 类，分类结果如图9。虽然从图中感觉看不出明显的分类，但是这也与数据样本太大有关。所以我们不妨采用量化的办法。

利用Rand——Index统计量的办法去检验分类效果，由于最理想的分类就是每个人的照片分到了一起，称之为标准分类，即是一个40*10的分类。我们将这样的分类和k-means 得到的结果相比较便可以得到Rand-Index统计量为0.974，由于非常接近1，所以说k-means的结果是非常好的。

我们看看k-means对1、2两个人的分类：

1：7 6 7 25 19 25 7 7 7 19

2：37 37 37 37 37 37 37 37 37 37

这是一二两个人所属的类别，结合上面层次聚类的图7 和图8，可以看出我们并不能说k-means把1人分的不好，而是同一个人照片之间差异的确很大。计算做出了正确的应对。

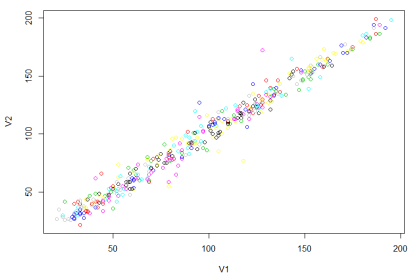

图 9: k-means聚类结果图

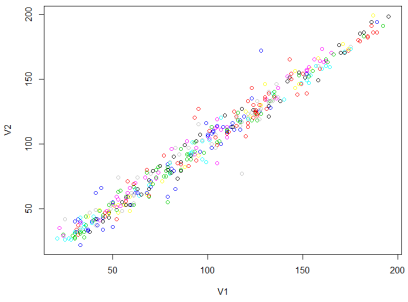

图 10: k-medoids聚类的结果图

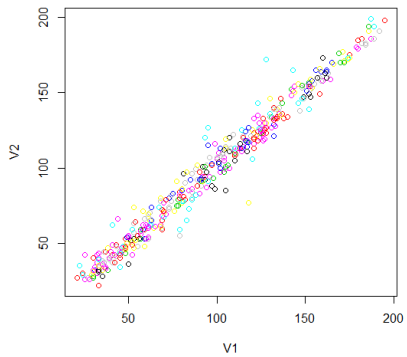

图 11: 谱聚类的结果图

3.2.2 K-medoids

同上我们可以把参数设成40，观察结果。图像如图10：不过跟K-means一样，图像是难以区分分类的，计算Rand-Index统计量结果为：0.969可见跟真实照片归属很接近。
同样我们比较1、2,两个人的聚类类别：
1：1 2 1 3 2 3 1 4 1 5
2：6 6 6 6 7 6 6 6 6 6
可见我们可以得出类似的结论。

3.3 谱聚类

谱聚类的图像结果如图11，聚类大小分别为：Cluster size:
8 15 4 21 4 8 14 5 11 3 5 20 4 5 14 10 2 31 10 6 3 19 5 11 6 4 2 11 27 5 8 5 10 15 12 5 20 12 12 8
其Rand-index统计量为：0.968,我们也可以认为谱聚类也是非常好的聚类方法。

3.4 总结与比较

由此可以知道后三种聚类方法结果有差异，但从聚类来说都有各自的有效性，并且与真实的聚类都较为接近。考虑czekanowski_dice距离，这三者分别为k-means:0.459,k-medoids:0.424,谱聚类：0.430，通过这个统计量的计算，我们可以认为在这个问题中k-means的方法更合适些，下面的一些问题的讨论便基于k-means方法。

4 聚类问题的拓展

4.1 与主成分的方法结合

我们知道照片是有很大的噪音的，经过计算前80维的主成分的百分比就高达92%,那么我们去花大量的时间和精力去计算4096维的数据是缺乏时效性的。同时我们不能保证后面的噪音是否会影响最终的聚类，在此做简单的主成分的分析。对比K-means聚类的结果差异。

利用R中的system.time命令来计算运行时间

```
> system.time(c1 <- kmeans(x, 40,100,25))
 用户  系统  流逝
49.64  0.17 50.00
> system.time(c11<- kmeans(yt,40,100,25))
用户 系统 流逝
0.20 0.00 0.21
>
```

图 12: 直接计算（上）和主成分方法（下）的运行时间

算法的效率体现在用户时间，比较两种方法的用户时间，我们看到这种方式带来非常明显的运算效率的提升，如果最后的分类结果是良好的，那么我们完全有理由用主成分的方式去分析这个问题。

计算这两种聚类的rand-index比较统计量为0.978，也是非常接近1的，czekanowski_dice距离为0.606。同时它与原来标准的40*10的每个人的照片归属分类相比计算出来的该统计量结果为0.972，czekanowski_dice距离为0.461，这与k-means方法非常接近。所以我们完全有理由接受在允许的条件下选择主成分分析的方式去处理该问题。

4.2 对每个人的局部特征进行分类

由于在现实中我们可能会关注每个人的脸型、鼻梁、发型等等的特征，那么局部特征的分类也是非常值得参考的。

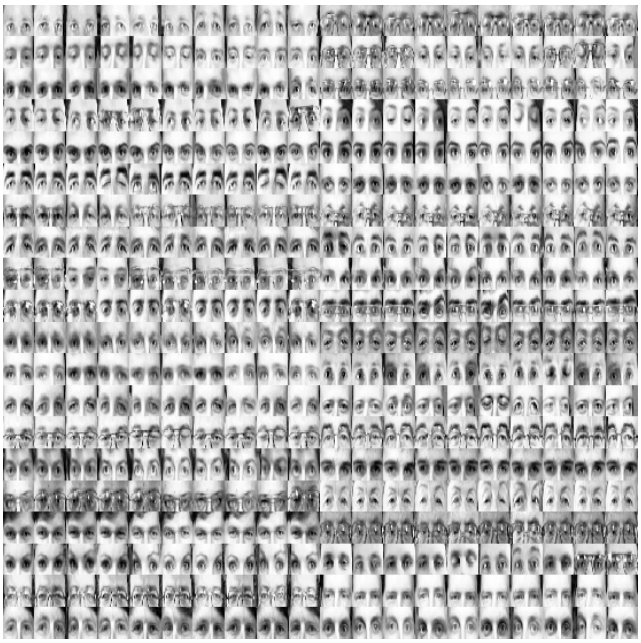

图 13: 人的眼睛图

在这里我们考虑对人的眼睛进行聚类，由于是单独对眼睛进行聚类我们就很难说具体的类别个数

有多少，所以我们考虑som方法和kmeans 方法结合的方式去解决这个问题。眼睛的图片如13
首先用som的方法去投影这样一个数据集，粗略估计聚类的个数。结果如图14

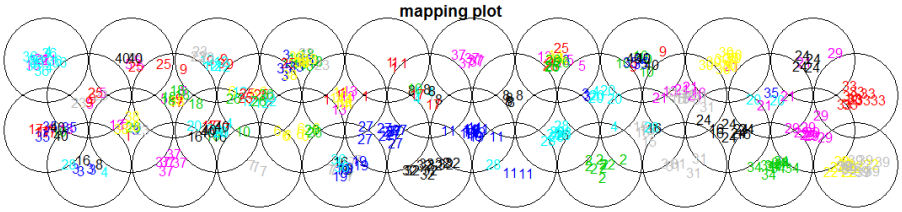

图 14: som分析结果图

所以我们可以估计这些数据的类也在30-50之间。
下面我们用Gap Statistic的方法去观察最好的k
计算结果如图15

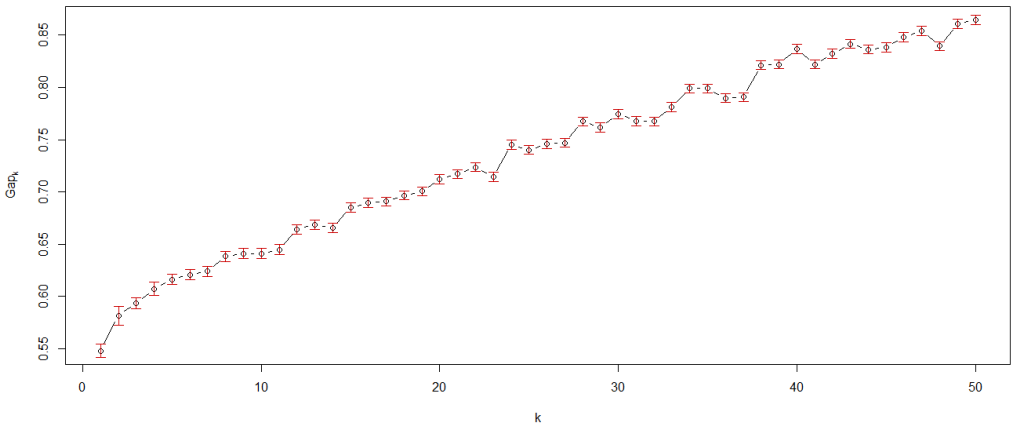

图 15: Gap-Statistic分析结果图

由图15，我们不妨继续尝试使用k=40 去聚类，考虑眼睛聚类的结果和真实的照片归属差异有多大，经过计算我们发现rand-index统计量为0.970，czekanowski_dice距离也为0.428，虽然比kmeans在czekanowski_dice 距离上稍微差了一点，但是你能发现眼睛也可以做一个主要的特征帮助人们快速分类。这也应证中国常说的“眼睛是心灵的窗户”，每个人的眼睛都是与众不同的。

5 结论与展望

我们在上面的讨论中可以看出，k-means，k-medoids和谱聚类是比较方便而且较为准确的聚类方式。在这个问题中相比较之下k-means是一种更加好的聚类方式，而这种方法的确在统计学中广泛应用。在讨论k-means聚类的基础之上，我们对主成分的方法进行分析，我们可以见到主成分能有效的排除噪音，而且加快运算速度，对于维数特别高的数据是一种非常有效的缩短运算时间的方式。此外，关于把眼睛作为识别人的特征是一种非常有趣的方式，同时在现实生活中期待会有相应的应用。

6 R code文件

可在home.ustc.edu.cn/∼passues/Hqface.txt 访问，看到源代码。

参考文献

[1] Bien, J., and Tibshirani, R. (2011), Hierarchical Clustering with Prototypes via Minimax Linkage, The Journal of the American Statistical Association

[2] Verma, T.; Sahu, R.K., (2013) PCA-LDA based face recognition system & results comparison by various classification techniques, Green High Performance Computing (ICGHPC), 2013 IEEE International Conference on , vol., no., pp.1,7, 14-15

[3] Calandriello D., Niu G., Sugiyama M. (2013), Semi-Supervised Information-Maximization Clustering