



基于 GochiUsa_Faces 数据集分类问题的解决方案

ID 沈运之

052110814

1729990469@qq.com

ID 黄开奕

182111510

1476060622@qq.com

ID 徐行

082120109

1928161381@qq.com

December 21, 2023

Keywords: 图像分类 降维 判别 多因变量线性回归 假设检验

包含 9141 张图片, 初始文件夹里包含 (通道数为 3) 从 26×26 , 到 987×987 尺寸不一的图片, 为了便于处理, 已经经过 python 脚本统一处理为 32×32 . 原数据集来源于 Kaggle: <https://www.kaggle.com/datasets/rignak/gochiusa-faces>.

1. 介绍

1.1. 概要

统计学习中, 分类问题应该算得上是一个相当经典的模型, 大多数方法都可以参与这一问题的解决, 基于此, 用分类问题来应用多元统计分析所学到的知识再合适不过。

分类问题中, 图像分类占据了很大程度的一部分, 然后, 现实中的图片分类问题要经过传感器获取, 以及 Jpeg 压缩一系列退化的过程, 其一般受噪声影响较为严重, 所以我们选择了产生于互联网上的图片, 即动漫人物的图片构建我们的分类问题 (其实单纯是因为兴趣)。

该图片数据集主要由两个文件夹构成, ANIME 文件夹用于训练, DANBOORU 文件夹用于测试, 其中包含 9 个类别, 分别是 Blue Mountain, Chino, Chiya, Cocoa, Maya, Megumi, Mocha, Rize, Sharo 对应数字 0-8; ANIME 包含 59579 张图片, DANBOORU

1.2. 解决方案

首先我们小组成员自行充当分类器, 分类效果非常好, 因此这个学习问题是理论上可以实现。下面我将阐述这份实验提供的解决方案:

Note:

- 首先观察图片数据的特征是否近似满足正态分布, 以及初步构建对于数据认识。
- 然后基于先验, 选择合适的方法进行降维, 并将降至二维进行可视化。
- 对于不同的降维结果, 使用基于模型的多因变量的线性回归, SVM, 以及 model-free 的基于决策树的分类器进行测试, 挑选出最好的结果。

- 基于以上结果进行分析。

1.3. 符号约定

为了便于叙述, 这里规定 N 为数据集样本数, M 为每个样本的特征, 这里定义每个样本的特征为图片张量向量化的结果, X 为 $N \times M$ 的数据矩阵, Y 为 $N \times 1$ 的标签向量, 其中 $y_i \in Z$ and $y_i \in [0, 8]$, 约定每一个样本为 $X_i^\top = [x_{i1} \cdots x_{iM}]$, 对应标签为 y_i , $Y = [y_1 \ y_2 \ \cdots \ y_N]^\top$ 从而有:

$$X = \begin{bmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_N^\top \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{bmatrix}$$

2. 数据属性

在对数据进行进一步分析, 为了尽可能防止出现数值问题 (0-255 以内的数字多次线性组合可能会是很大的值), 首先先将数据通过标准化处理调整为均值为 0, 方差为 1, 设 \bar{x}_i 为数据矩阵 X 第 i 列的样本均值 (也就是随机变量 X_i 的 N 次取样), σ_i 为其标准差, 于是其内的数据 x 的标准化后的值 \tilde{x} 为:

$$\tilde{x} = \frac{x - \bar{x}_i}{\sigma_i}$$

2.1. 类别情况

首先观察最直观的数据属性, 将每个类别在训练集和测试集上的规模画出 (见 Figure 1), 训练集内最少的两个类别为 Mocha 与 Blue Mountain 分别有 1241 个, 1607 个, 而数量最多的类别 Chino 有 12941 个, 倍数达到十倍, 该数据集为长尾数据集, 原数据集作者说, 大部分角色具有明显的特征, 因此我们仍然选择这两个类别作为我们分类任务的一环 (本质上还是因为这个学习问题不太难)。

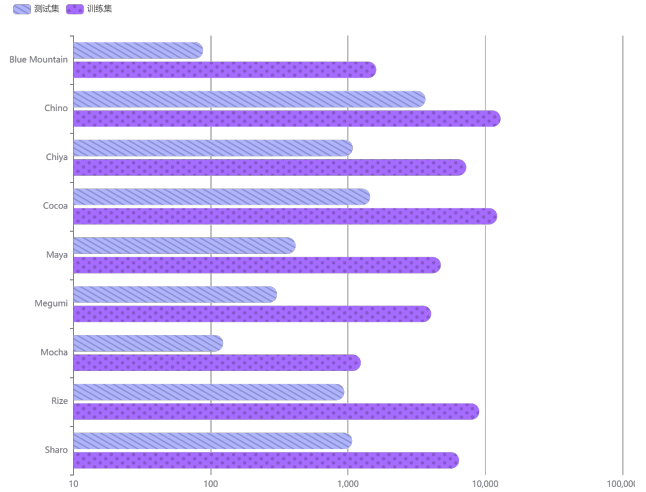


Figure 1: 各类别训练集与测试集的分类情况, 横轴为类别对应的样本数目, 且采用对数刻度

2.1.1. 类别分布

判别分析假设每个总体数据服从正态分布, 此时 FLDA 效果达到最优, 检验一下各个类别是否近似服从 M 元态分布很有必要, 同时也可以为判别法的有效性分析做个铺垫。

一个简单的方式是, 当原假设 $H_0: X \sim N_M(\mu, \Sigma)$ 成立时, 数据总体 X 具有如下性质:

Lema 1

$$D^2 = (X - \mu)^\top \Sigma^{-1} (X - \mu) \sim \chi^2(M)$$

因此, 可以考虑绘制 χ^2 统计量的 Q-Q 图, 设 $D_{(t)}^2$ 为排序后第 t 个样本的马氏距离, 以及 χ_t^2 为 $\chi^2(M)$ 对应的分位数, 通过观察 $(D_{(t)}^2, \chi_t^2) (t = 1, \dots, N)$ 散点图是否近似分布在斜率为 1 的直线上来检验其正态性

2.2. 特征相关性

由于下面要使用线性回归模型, 需要先保证数据特征不存在强相关, 否则严重的多重共线性将导致线性模型 $C^\top C$ 不满秩, 使得线性回归将不存在唯一解, 这可能会影响答案的准确性。注意到样本的特征数为 $M = 3072$, 设样本协方差阵为 S , $V^{1/2} = \text{diag}(\sqrt{S_{11}}, \sqrt{S_{22}}, \dots, \sqrt{S_{MM}})$, 相关系数矩阵 R 由

以下公式给出:

$$R = (V^{1/2})^{-1} S (V^{1/2})^{-1} b$$

实际计算复杂度为 $N \times M^2$, 实际运行却很快, 这可能得归功于 numpy 的矩乘优化, 统计总计 3072×3072 个相关系数, 绘制其频率 (已经划分好分段区间) 直方图 (参考 Figure 2)

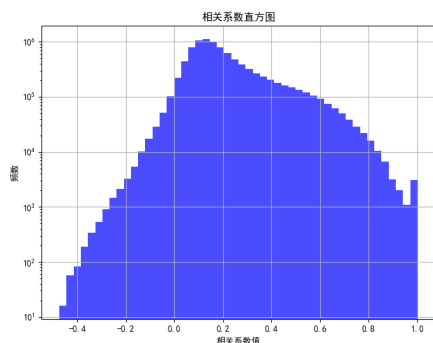


Figure 2: 频率直方图, 已经将纵坐标处理为 log 尺度

计算得到有多达 136496 对特征具有 ≥ 0.7 的相关性 (情理之中), 之后的进一步工作可以尝试使用降维方法减少多重共线性, 那时再做进一步分析。

3. 降维

现在, 我们已经对数据 (训练集) 有了先验认知, 发现数据集存在以下几个问题:

Observación 1

1. 样本特征维度过大, 这无论是对于存储与速度都提出了极大的要求。
2. 特征之间存在不可忽视的相关性。

使用这样的数据来训练传统模型是不可行的, 由于本身该分类问题属于不太难的学习问题, 只是用少量特征来学习大概率可行, 接下来考虑使用降维方法对数据进行简化。

3.1. PCA 降维

最简单的无监督降维是 PCA 方法,

Definición 1 (Nombre de la definición)

Sea $f \dots$

Definición 2

Definición sin nombre ...

Ejemplo para hacer referencia a una definición (teorema, corolario, etc), en la definición 2.

Notación 1

Notación sin nombre ...

Proof. Prueba de teorema

□

Teorema 1

Teorema sin nombre ...

En el teorema ??

Ejemplo 1 (Nombre del ejemplo)

Sea $f \dots$

Ejemplo 2

Ejemplo sin nombre ...

Corolario 1

Sea $f \dots$

Corolario 2

Corolario sin nombre ...

Lema 2 (Nombre del lema)

Sea $f \dots$

Lema 3

Lema si nombre ...

Note: (Nombre de la nota) Sea $f \dots$

Note: Nota sin nombre ...

Vocabulario 1 (Nombre del vocabulario) Sea $f \dots$

Vocabulario 2 Vocabulario sin nombre ...

Algoritmo 1 (Nombre del algoritmo)
Algoritmo con nombre ...

(Nombre de la caja)
Sea $f \dots$

Scaja sin nombre ...



Figure 3: Título de la figura. Decir si es elaboración propia o poner referencia.

Note cómo en la Figura 3 ...

En la Figura 4, en la subfigura 4b se observa que ...

Puede observar en la Tabla 1 ...



(a) Subfigura 1



(b) Subfigura 2

Figure 4: Título para la figura en general. Decir si es elaboración propia o poner referencia.

Table 1: Título de la Tabla. Decir si es elaboración propia o poner referencia.

<i>name</i>	<i>foo</i>			
Models	A	B	C	D
Model X	X1	X2	X3	X4
Model Y	Y1	Y2	Y3	Y4

Ecuación numerada:

$$x = \frac{b \pm \sqrt{b^2 - 4ac}}{2a} \tag{1}$$

En la fórmula 1 ...

Ecuación no numerada:

$$x = \frac{b \pm \sqrt{b^2 - 4ac}}{2a}$$

Ecuación alineada numerada:

$$x = a^2 - b^2 \tag{2}$$

$$= (a - b)(a + b) \tag{3}$$

En las expresiones 2 y 3 ...

Ecuación alineada no numerada:

$$x = a^2 - b^2$$

Ecuación centrada

$$a^2 = b^2 + c^2$$

Ejemplo de código Java:

```

/**
 * This is a doc comment.
 */
package
    com.ocicweb.jnb.lombok;

import java.util.Date;
import lombok.Data;
import
    lombok.EqualsAndHashCode;
import lombok.NonNull;

@Data
@EqualsAndHashCode(exclude={"address", "city", "state", "zip"})
public class Person {
    enum Gender { Male,
        Female }

    // another comment

    @NonNull private
        String firstName;
    @NonNull private
        String lastName;
    @NonNull private
        final Gender
        gender;
    @NonNull private
        final Date
        dateOfBirth;

    private String ssn;
    private String
        address;
    private String city;
    private String state;
    private String zip;
}

```

Este es código en la misma línea `import java.util.Date;`, el símbolo `|` es sólo un delimitador y se puede cambiar por algún otro que no se utilice en el código.

Esta es una cita de la bibliografía: [1]

La bibliografía se prefiere según APA con utilizando biblatex con Biber, también aceptamos el formato IEEE.

4. Bibliography

[1] Cita H

A. Apéndice

Apéndice