



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 20, 2023

Keywords: 图像分类 降维 判别 多因变量线性回归 假设检验

DANBOORU 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,

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:

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

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)。

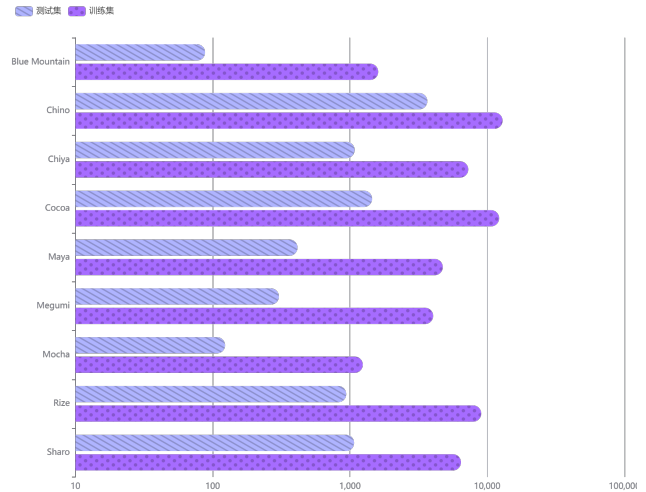


Figure 1: 各类别训练集与测试集的分类情况, 横轴为类别对应的样本数目, 且采用对数刻度

2.2. Subsección

2.2.1. Subsubsección

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 ...

Teorema 1 (Nombre del teorema)

Sea $f \dots$

Proof. Prueba de teorema

□

Teorema 2

Teorema sin nombre ...

En el teorema 1

Ejemplo 1 (Nombre del ejemplo)

Sea f ...

Ejemplo 2

Ejemplo sin nombre ...

Corolario 1

Sea f ...

Corolario 2

Corolario sin nombre ...

Lema 1 (Nombre del lema)

Sea f ...

Lema 2

Lema si nombre ...

Note:

(Nombre de la nota) Sea f ...

Note:

Nota sin nombre ...

Vocabulario 1 (Nombre del vocabulario)

Sea f ...

Vocabulario 2

Vocabulario sin nombre ...

Algoritmo 1 (Nombre del algoritmo)

Algoritmo con nombre ...

Observación 1

Observación sin nombre ...

(Nombre de la caja)

Sea f ...

Scaja sin nombre ...



Figure 2: Título de la figura. Decir si es elaboración propia o poner referencia.

Note cómo en la Figura 2 ...



(a) Subfigura 1



(b) Subfigura 2

Figure 3: Título para la figura en general. Decir si es elaboración propia o poner referencia.

En la Figura 3, en la subfigura 3b se observa que ...

Puede observar en la Tabla 1 ...

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:

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

(1)

En la fórmula 1 ...

Ecuación no numerada:

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

Ecuación alineada numerada:

$$\begin{aligned} x &= a^2 - b^2 & (2) \\ &= (a - b)(a + b) & (3) \end{aligned}$$

En las expresiones 2 y 3 ...

Ecuación alineada no numerada:

$$\begin{aligned} x &= a^2 - b^2 \\ &= (a - b)(a + b) \end{aligned}$$

Ecuación centrada

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

Ejemplo de código Java:

```
/**
 * This is a doc comment.
 */
package
    com.ociweb.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.

3. Bibliography

[1] Cita H

A. Apéndice

Apéndice