《金融技术与实践》课题报告

通过身体数据匹配提供穿衣决策的交互界面

信息学院 陈冠华 2017202095

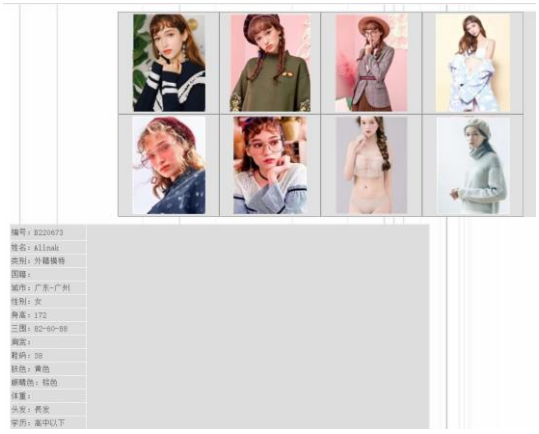
目录

绪论

本设计意图通过数据分析的方式，为线上购买衣物提供决策帮助。具体地说，对模特身体数据进行聚类分析，分析用户的身体数据落在哪个聚落后，为用户推荐与其位于相同聚落的明星以及其代言的产品。这个过程通过交互界面窗口实现。

数据爬取

我的数据源选择了国内最大的模特资源网站中国模特网（http://www.chinaeve.com/ModelList.aspx）。在网页的爬取过程中我发现该网站的模特信息网页有两种不同的布局模式。分别为下两图所示。



因此我针对两种布局设计了不同的爬虫程序，对网站进行了两次爬取之后获得了所有模特的数据（约 4800 条左右）。

以下是我在爬虫过程中总结的一些问题与解决办法：

遇到的问题	遇到问题的原因	解决办法
程序无法找到元素	网页没有加载完成	调用Sleep函数
	窗口没有切换	切换窗口句柄
	查找的元素在iframe里面	切换到iframe里后再查找元素
程序找到元素后无法执行点击操作	要点击的元素被其他元素覆盖	不要使用.click（），使用browser.execute_script('arguments[0].click();', element)

数据清洗

3.1 对不规范的值的处理

不规范的数据类型有（0，00，-1，-）这四类。数据案例如下三图所示。

```
200,2,1,B26074,广告模特,2,青岛,192,0,0,0,00,45,黄色,黑色,中发,,0
200,3,4,K115315,外籍童模,2,境外,104,-1,-1,-1,-,-,白色,褐色,中发,,0
200,3,4,K115315,外籍童模,2,境外,104,-1,-1,-1,-,-,白色,褐色,中发,,0
```

这些不合规范的数据项被全部设置成 0。

3.2 对带单位数据的处理

```
99,2,3,N1420140,,2,成都,63,40,40,37,22,8cm,黄色,黑色,短发,,0
```

某些元素后带单位，如 cm，码。这些数据的单位后缀被直接去除。

3.3 语言转化

```
34,4,4,L124995,日韩港台,3,境外,173,89,65,89,,41,黄色,淺啡,長髮,,0
```

某些文字数据是用繁体字书写，在独热编码的时候妨碍了类别的转换。因此所有的繁体字全部被转换成了简体中文。

3.4 数据的提取

爬取的数据并不全是分析所需要的。我从爬取的数据中抽取了其中的（Height，Chest，Waist，Hips，Eyes，Skin，Suits，Shoes）这几种属性进行数据分析。

数据描述

属性名称	含义
Height	身高（cm）
Chest	胸围（cm）
Waist	腰围（cm）
Hips	臀围（cm）
Eyes	眼睛颜色
Skin	皮肤颜色
Suits	西装尺码
Shoes	鞋子尺码

数据预处理

4.1 对类别数据的处理

类别数据包括眼睛颜色与皮肤颜色，我对这两种属性进行了独热编码。

4.2 对异常值的探查与处理

4.2.1 统计学方法

根据达伊拉原则，假设每列数据服从正态分布，探测那些位于两侧尾部的数据。然而结果并不是很理想，程序识别的的异常值都很“正常“。达伊拉原则不适用于此设计的数据集的原因应该在与本数据并不服从简单的正态分布，他们是取自不同类别的人群的人体数据，至少是三个正态分布的叠加的混合分布。

4.2.1 机器学习方法

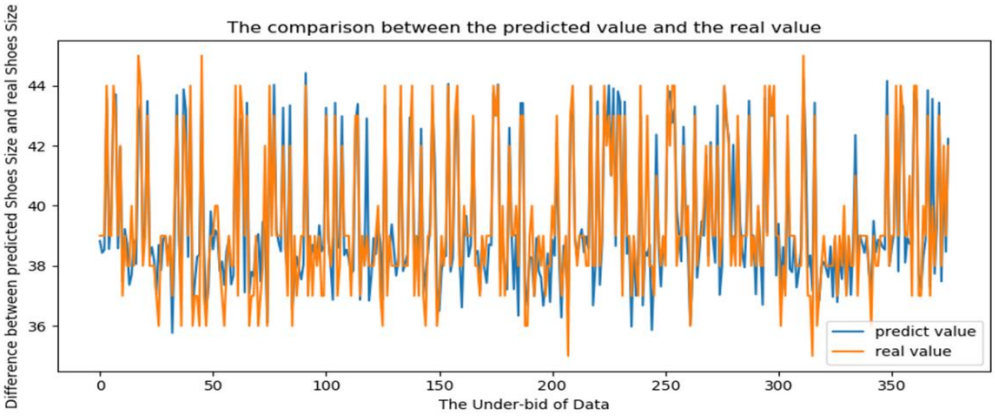
随后采用了机器学习中的 KNN 算法。KNN 算法检测异常值的原理是：对于异常值而言，距离他们最近的 N 个节点中（在 N 合适的情况下），离异常点最远的点与异常值的欧氏距离会特别大。由于 Suits 与 Shoes 属性有缺失值，数据集被切成四份（如下图所示），并且设置每份数据中异常值的比例为 0.05。通过 KNN 算法检测出的异常值结果符合预期，因此被从原数据中剔除出去。

	Suit值为空	Suit值不为空	Shoes值为空	Shoes值不为空
1	√		√	
2	√			√
3		√	√	
4		√		√

4.3 对缺失值的处理

4.3.1 随机森林算法

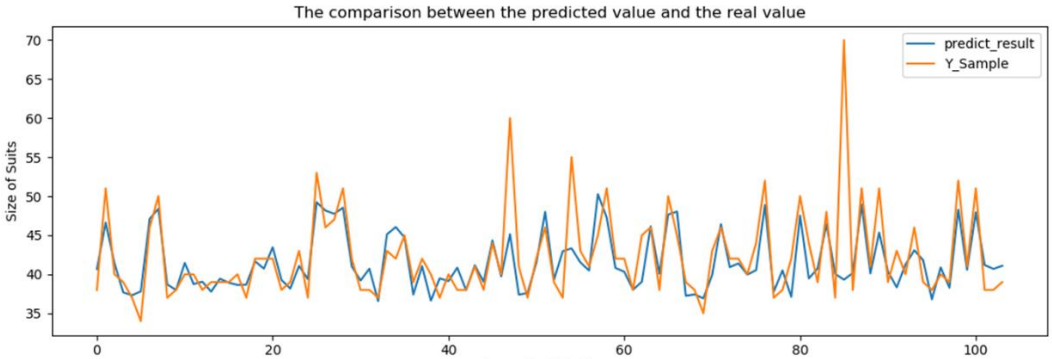
采用随机森林算法填充缺失值较少的 Shoes 值。标签集被分切成训练集（约 3900 条）与测试集（约 300 条）。经过训练后预测值与测试集的比较如下图所示：



About 400 pieces of data were extracted to test the results of random forest fill missing values

4.3.2 多元线性回归

采用多元线性回归算法填充缺失值较多的 Suits 值。标签集被分切成训练集（约 40000 条）与测试集（约 200 条）。经过训练后预测值与测试集的比较如下图所示：

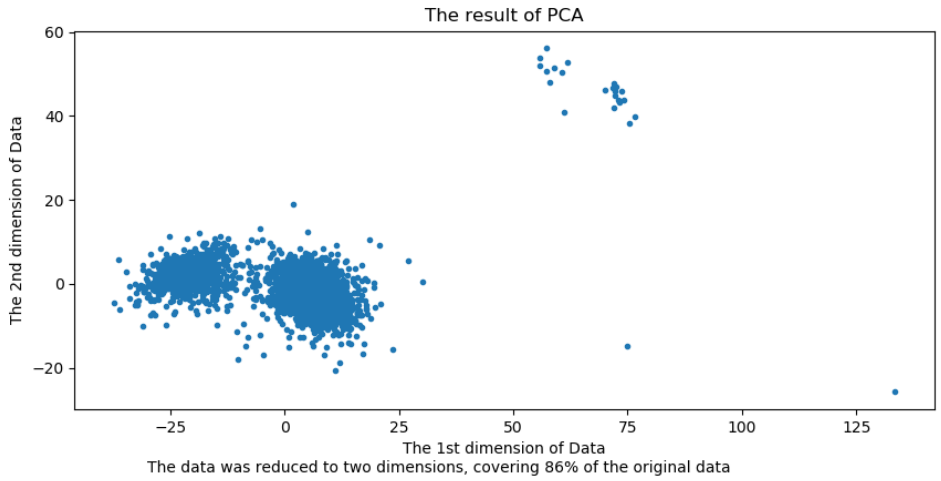


About 400 pieces of data were extracted to test the results of Multiple linear regression filling missing values

数据聚类分析

5.1 降至两维并可视化的尝试

采用 PCA 降维方法，保留 86%数据信息的基础上将数据降至两维并可视化。结果如下图所示：

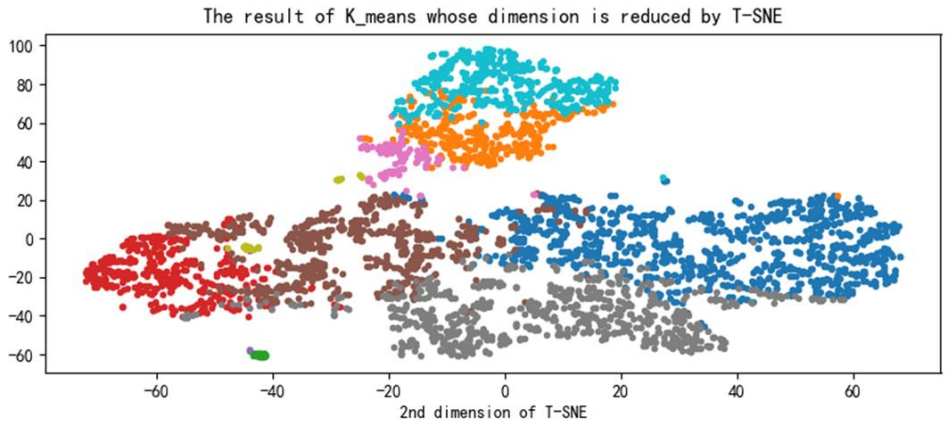


5.2 降至四维

考虑到 86%的信息保留率是个比较低的数字。最终数据被降维至四维，保留原数据 95%的信息。

数据结果降维可视化

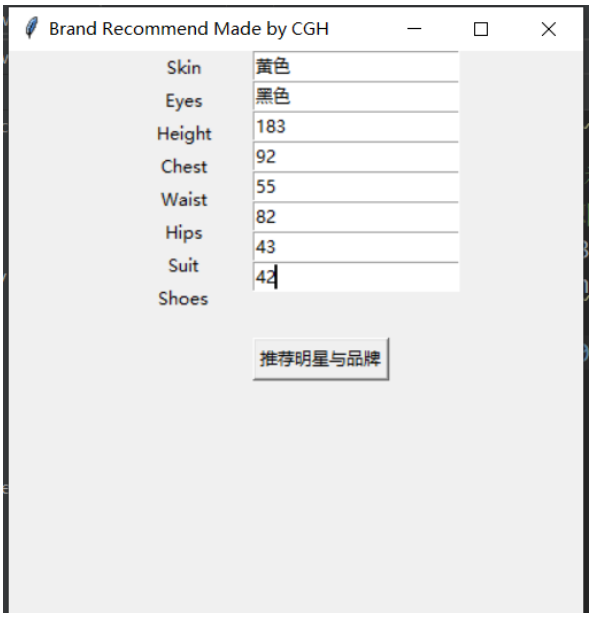
对四维数据进行 K-means 聚类分析，并采用 T-SNE 算法对降维后的数据进行可视化。结果如下图所示：

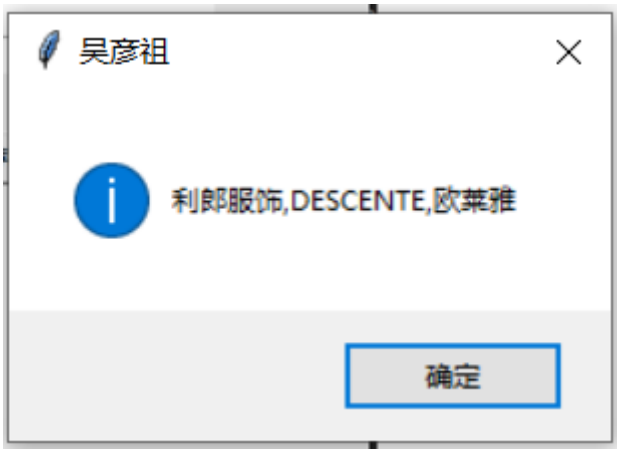


The result of PCA is put into a K-Means machine(cluster=10), the result of which is shown after being processed by T-SNE

图形化交互界面的设计

采用 python 的 tkinter 包设计了一个包括输入身体数据与输出窗口的交互界面。运行图片如下图所示：





输入身体数据点击按钮后，会有一个弹窗显示推荐的明星与明星所代言的服装产品。