**广州大学学生实验报告**

**开课学院及实验室：**计算机科学与网络工程学院软件实验室 **2020年 11 月20日**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **学院** | **计算机科学与网络工程学院** | **年级/专业/班** | **18级软件工程3班** | **姓名** | 阮逸栋 | **学号** | 1806300142 |
| **实验课程名称** | **机器学习与数据挖掘实验** | | | | | **成绩** |  |
| **实验项目名称** | 多源数据集成、清洗和统计 | | | | | **指导老师** | 彭伟龙 |

**实验一**

1. **组员信息**

阮逸栋（小组共1人）

1. **作业环境（文件说明，函数说明，调用的函数库以及涉及哪些技术）**

**文件说明：**

数据源1：一.数据源1.xlsx

数据源2：一.数据源2-逗号间隔.txt

**调用的数据库以及涉及的技术：**

未调用数据库。

涉及的技术主要包括按要求对数据进行清洗，同时调用了xlrd、copy等模块协助进行处理，最后由于题目要求计算部分不能使用库函数，通过手动对计算函数进行编程的方式对集成和清洗后的数据进行统计。

1. **作业题目和内容**

**题目**：广州大学某班有同学100人，现要从两个数据源汇总学生数据。第一个数据源在数据库中，第二个数据源在txt文件中，两个数据源课程存在缺失、冗余和不一致性，请用C/C++/Java程序实现对两个数据源的一致性合并以及每个学生样本的数值量化。

* 数据库表：ID (int), 姓名(string), 家乡(string:限定为Beijing / Guangzhou / Shenzhen / Shanghai), 性别（string:boy/girl）、身高（float:单位是cm)）、课程1成绩（float）、课程2成绩（float）、...、课程10成绩(float)、体能测试成绩（string：bad/general/good/excellent）；其中课程1-课程5为百分制，课程6-课程10为十分制。
* txt文件：ID(string：6位学号)，性别（string:male/female）、身高（string:单位是m)）、课程1成绩（string）、课程2成绩（string）、...、课程10成绩(string)、体能测试成绩（string：差/一般/良好/优秀）；其中课程1-课程5为百分制，课程6-课程10为十分制。

#### 参考：

数据库中Stu表数据

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Name | City | Gender | Height | C1 | ... | C10 | Constitution |
| 1 | Sun | Beijing | boy | 160 | 87 |  | 9 | good |
| 2 | Zhu | Shenzhen | girl | 177 | 66 |  | 8 | excellent |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |

student.txt中

|  |
| --- |
| ID Name City Gender Height C1 。。。 C10 Constitution  202001 Sun Beijing male 1.80 87 。。。 9 good  202003 Tang Hanghai male 1.56 91 。。。 10 general  ... ... ... .. ... .. ... ... ... |
|  |

两个数据源合并后读入内存，并统计：

#### 1. 学生中家乡在Beijing的所有课程的平均成绩。

#### 2. 学生中家乡在广州，课程1在80分以上，且课程9在9分以上的男同学的数量。(备注：该处做了修正，课程10数据为空，更改为课程9)

#### 3. 比较广州和上海两地女生的平均体能测试成绩，哪个地区的更强些？

#### 4. 学习成绩和体能测试成绩，两者的相关性是多少？（九门课的成绩分别与体能成绩计算相关性）

#### 提示：

参考数据结构：

Student{

int id;

string name;

vector<float> data;

}

可能用到的公式：

|  |  |
| --- | --- |
| 均值公式 |  |
| 协方差公式 |  |
| z-score规范化 |  |
| 数组A和数组B的相关性 | 这里A=[a1, a2,...ak,..., an],  B=[b1, b2,...bk,..., bn],  mean(A)代表A中元素的平均值  std是标准差，即对协方差的开平方。  点乘的定义：IMG_256 |

注意：计算部分不能调用库函数；画图/可视化显示可以用可视化API或工具实现

**具体实验内容：**

**合并文件：**首先定义一个学生类用于存储即将读取的各项参数，由于给定的两个数据一个为.xlsx，另一个为.txt，所以需要定义两个不同的方法分别用于读取对应后缀的文件，然后将读取的数据存入类中。

**数据清洗：**首先，由于读取文件时已将空的值的序号存入一个列表中，当遍历到相同的ID时，只需查看该对象中的空值列表是否存在值，如果有，且遍历到的对象其该值对应的成绩不为空，则将该成绩复制存入。

由于两个文件中的属性单位等并不统一，需要对各列数据进行标准化。两个文件的性别命名并不统一，在合并后的表中，我们将女性统一命名为“female”，男性统一命名为“male”；对ID列数据进行标准化处理，统一为“202\*\*\*”的格式；对身高列数据进行标准化，统一为以cm为单位。还需要注意的是，最后需要将数据转换成float，便于后续去重。紧接着便是对文件进行去重，即去除文件中的重复行，去重完成后还需要判断ID号是否存在相同的，只要检测到ID相同的则判定为数据失效，将所有ID相同的数据项全部删除。

**数据统计：**该部分不详述。

1. **难题与解决**

python相比c++更加容易掌握且应用，所以我这次选择了使用python，但是在编写项目的过程中仍然遇到不少的困难。首先我存储读取的数据的方式是采用了定义一个类来存储，相比起直接用列表存储显得更加复杂，而且后续也因此出现了很多的错误。

其次是计算，在编写计算方法的过程中，由于种种原因导致计算结果与理想结果出现了较大的误差，在修改上通过百度，参照同学的建议，最终得到了想要的结果。