心衰风险预测模型

1. 数据源

深圳市医疗大数据项目中针对高血压、糖尿病人群进行基于心血管疾病的多标签分类算法研究。

1. 实验组：高血压糖尿病确诊日期之后并患有指定的心血管疾病的人群提取出来作为实验组数据集,

对照组：高血压人群中未患有实验组的心血管疾病的人群。

多种心血管疾病的研究统计如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 疾病名称 | 高血压组数据量 | 糖尿病组数据量 | 总数据量 |
| 脑卒中 | 5956 | 1933 |  |
| 心衰 | 779 | 300 |  |
| 急性心梗 | 740 | 290 |  |
| 心源性猝死 | 130 | 42 |  |
| 心肌缺血 | 857 | 280 |  |
| 肺心病 | 87 | 29 |  |
| 心脏功能病变 | 2462 | 650 |  |
| 冠心病 | 19761 | 7000 |  |

-arff ../Dataset/output.arff -xml ../Dataset/output.xml -unlabeled ../Dataset/output1.arff

1. 数据分布：

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 疾病名称 | 高血压组 | | | | | | 糖尿病组 | | | | | |
| 心血管病 | 性别分布 | 年龄分布 | | | | | 性别分布 | 年龄分布 | | | | |
| <18 | <29 | <40 | <65 | >=65 | <18 | <29 | <40 | <65 | >=65 |
| 脑卒中 | 男3406 女 2528 | [0, 8, 175, 2896, 2877] | | | | | 男10451 女 7004 | [0,113,1805,12388,3181] | | | | |
| 心衰 | 男 421  女 357 | [0 , 0 , 21, 293 ,465] | | | | | 男10515 女 7032 | [3,157,1782,12412,3220] | | | | |
| 急性心梗 | 男464  女273 | [0, 1, 23, 371, 345] | | | | | 男10441, 女 6928 | [2, 138, 1759, 12319, 3188] | | | | |
| 心源性猝死 | 男54  女76 | [0, 0, 16, 91, 23] | | | | | 男10560 女 7038 | [1, 144, 1759, 12491, 3238] | | | | |
| 心肌缺血 | 男476 女 380 | [0, 3, 46, 441, 367] | | | | | 男10564 女6989 | [3, 149, 1792, 12430, 3205] | | | | |
| 肺心病 | 男54  女32 | [0, 1, 2, 22, 62] | | | | | 男10659 女6890 | [4, 156, 1827, 12436, 3152] | | | | |
| 心脏功能病变 | 男 1268  女 1189 | [0, 17, 225, 1437, 783] | | | | | 男10424  女6986 | [2, 150, 1746, 12365, 3178] | | | | |
| 冠心病 | 男10743  女8976 | [1, 35, 800, 10969, 7956] | | | | | 男9467 女6182 | [2, 149, 1684, 11339, 2498] | | | | |

脑卒中：

-------------抽取前------------------

[[5956, [3406, 2528], [0, 8, 175, 2896, 2877]], [35060, [21046, 13953], [4, 293, 3581, 24783, 6399]]]

--------------抽取后---------------

[[5956, [3406, 2528], [0, 8, 175, 2896, 2877]], [17487, [10451, 7004], [0, 113, 1805, 12388, 3181]]]

心衰：

-------------------抽取前--------------------------

[[779, [421, 357], [0, 0, 21, 293, 465]], [35060, [21046, 13953], [4, 293, 3581, 24783, 6399]]]

--------------------随机抽取对照组后的数据分布---------

[[779, [421, 357], [0, 0, 21, 293, 465]], [17574, [10515, 7032], [3, 157, 1782, 12412, 3220]]]

急性心梗

[[740, [464, 273], [0, 1, 23, 371, 345]], [35060, [21046, 13953], [4, 293, 3581, 24783, 6399]]]

--------------------随机抽取对照组后的数据分布---------

[[740, [464, 273], [0, 1, 23, 371, 345]], [17406, [10441, 6928], [2, 138, 1759, 12319, 3188]]]

心源性猝死

[[130, [54, 76], [0, 0, 16, 91, 23]], [35060, [21046, 13953], [4, 293, 3581, 24783, 6399]]]

--------------------随机抽取对照组后的数据分布---------

[[130, [54, 76], [0, 0, 16, 91, 23]], [17633, [10560, 7038], [1, 144, 1759, 12491, 3238]]]

心肌缺血

[[857, [476, 380], [0, 3, 46, 441, 367]], [35060, [21046, 13953], [4, 293, 3581, 24783, 6399]]]

--------------------随机抽取对照组后的数据分布---------

[[857, [476, 380], [0, 3, 46, 441, 367]], [17579, [10564, 6989], [3, 149, 1792, 12430, 3205]]]

肺心病

[[87, [54, 32], [0, 1, 2, 22, 62]], [35060, [21046, 13953], [4, 293, 3581, 24783, 6399]]]

--------------------随机抽取对照组后的数据分布---------

[[87, [54, 32], [0, 1, 2, 22, 62]], [17575, [10659, 6890], [4, 156, 1827, 12436, 3152]]]

心脏功能病变

[[2462, [1268, 1189], [0, 17, 225, 1437, 783]], [35060, [21046, 13953], [4, 293, 3581, 24783, 6399]]]

--------------------随机抽取对照组后的数据分布---------

[[2462, [1268, 1189], [0, 17, 225, 1437, 783]], [17441, [10424, 6986], [2, 150, 1746, 12365, 3178]]]

冠心病

---------------------原始文件分布-------------------

[[19761, [10743, 8976], [1, 35, 800, 10969, 7956]], [35060, [21046, 13953], [4, 293, 3581, 24783, 6399]]]

--------------------随机抽取对照组后的数据分布---------

[[19761, [10743, 8976], [1, 35, 800, 10969, 7956]], [15672, [9467, 6182], [2, 149, 1684, 11339, 2498]]]

十、缺失值填充

生成相似性矩阵（knn）

命令1：

python knn.py Cj/second\_hf\_codes/new\_features\_dataset2.csv

new\_features\_dataset1.csv文件数据输入格式说明：

标题在第一行，非icd编码特征在前列，icd编码特征紧跟其后，类标号在最后一列。

填充后的数据集：

命令2：

python nullvaluefill.py new\_features\_dataset1.csv（new\_features\_dataset2.csv）

输出数据文件Nullvaluefill.csv（Nullvaluefill3.csv严格保留并发症之前的特征，并不删除特征，仅将并发症之后的特征值置为0）

十一、特征相关性分析

十二、预测模型建立

十三.模型结果分析评价