# 大数据期末 project

11510606 朱涵枫

## 摘要

本文主要围绕已知的乳腺癌数据展开分析，由于数据量较大且种类较多首先对数据进行了初步的分析，然后按照题目要求进行了数据的预处理，在问题二中对 PAM50mRNA 所用到的 43 个蛋白质基因用 K-Means 进行聚类，并于测试结果进行了比较。问题三中以 PAM50mRNA 所用到的 43 个蛋白质基因作为输入值，已知测试结果作为标签，分别利用了朴素贝叶斯、随机森林、K 近邻三种分类方法进行分类，之后对数据进行修改完善，重新计算并于原计算结果进行比较。问题四中用 PAM50mRNA 的 43 个蛋白质基因来对其他已知结果进行分类（CN Clusters、miRNA Clusters 等）通过与问题三中所得结果进行比较来判断这些测试是否也用到这 43 个基因。

## 一、数据分析与处理

**1）题目一共给出了三个文件，对文件中具体数据分析如下：**

| 数据类别 | 所属位置 |
|---|---|
| 83 个测试者（3 个正常者，3 个测量两次） | 文件 1 |
| 全部的蛋白质基因（可视作基因库） | 文件 1 |
| 105 个乳腺癌患者 | 文件 2 |
| 他们的具体测试结果 | 文件 2 |
| PAM50mRNA 检测所用到的 100 个蛋白质基因 | 文件 3 |

**表 1：三文件包含信息**

各数据间的主要关系是：

①在文件 2 的 105 个乳腺癌患者中与文件 1 中测试者编号能够对应的只有 77 个，需要对文件 2 中的患者进行筛选。

②在文件 3 中所出现的 100 个蛋白质基因与文件一中对应的只有 43 个，需要对文件 3 中的蛋白质基因进行筛选。

**2）之后开始对数据进行处理**

①首先对文件 1 中重复的三列进行筛选

在 Excel 中仅对出现的测试者编号进行重复值筛选，得到筛选结果如下：

| 列名称 | 位置 |
|---|---|
| A0-A12D.01TCGA | 第 0 列 |
| A0-A12D.05TCGA | 第 9 列 |
| C8-A131.01TCGA | 第 1 列 |
| C8-A131.32TCGA | 第 67 列 |

| | |
|---|---|
| AO-A12B.01TCGA | 第 2 列 |
| AO-A12B.34TCGA | 第 73 列 |

**表 2：重复测试者信息**

因此在数据预处理前删除了数据的前三列以及最后三列健康人的数据并且去除了无用的 gene_symbol 和 gene_name 两列，并把删除后的数据作为初始数据进行处理。

②对数据进行预处理

首先通过缺失值查找发现，数据中存在缺失值，选择采用中位数来进行数据的填充。

之后通过数据分析发现，数据间差别较大，为了使数据波动较小因此选择 Z-scale 标准化；

为了防止过拟合，使假设更好拟合训练数据，选择对数据进行正则化；

最后容易看出数据中存在少量的异常值，在对异常值处理时，直接选用 Z-scale 标准化后的结果，将结果中大于 3 小于-3 的值直接删去然后利用中位数进行填充。

经过一系列数据预处理得到最后数据结果如下：

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 0 | -0.100876 | -0.087080 | -0.086764 | -0.093288 | -0.087027 | -0.086847 | -0.086287 |
| 1 | -0.917057 | -0.931697 | -0.927725 | -0.936935 | -0.930814 | -0.931783 | -0.926554 |
| 2 | 2.417365 | 2.418979 | 2.419895 | 2.444497 | 2.423793 | 2.420506 | 2.422360 |
| 3 | 0.507856 | 0.508338 | 0.508790 | 0.508893 | 0.507820 | 0.508870 | 0.508878 |
| 4 | 1.226838 | 1.217054 | 1.220779 | 1.222975 | 1.221941 | 1.218033 | 1.220525 |
| 5 | 0.995087 | 0.994363 | 0.990783 | 0.991498 | 1.001830 | 0.991013 | 0.990690 |
| 6 | 0.773747 | 0.766199 | 0.766732 | 0.770845 | 0.758340 | 0.763334 | 0.763138 |
| 7 | 0.083390 | 0.063939 | 0.064335 | 0.074425 | 0.067268 | 0.071135 | 0.068203 |
| 8 | 0.048630 | 0.060513 | 0.060912 | 0.057125 | 0.063959 | 0.060881 | 0.054327 |
| 9 | -0.443432 | -0.409834 | -0.413283 | -0.426016 | -0.413145 | -0.413459 | -0.412621 |

**表 3：预处理后的数据**

## 二、对 PAM50mRNA 所用基因用 K-Means 进行聚类

**1）首先对 PAM50mRNA 用到的基因进行筛选，对真正参与测试的患者进行筛选**

通过分析发现，第二个文件中只有 77 个真正参与了第一个文件中的蛋白质测试，因此将第二个文件与第一个问件的测试者编号进行对应，在筛选后最终找到真正参加测试的 77 个病人编号。

对第三个文件进行分析过后发现 PAM50mRNA 的检测一共用到了 100 个蛋白质基因，但是与有一些基因在第一个文件中是无法找到的，因此将第三个文件中出现的蛋白质基因编码与第一个文件对应，在筛选后最终找到了真正有用的 43 个基因。

最后将 43 个蛋白质基因与 77 个测试者数据进行整合，最后得到了如下的数据：（数据较多，这里仅以片段展示）

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| TCGA-A2-A0EY | -1.015330 | -0.840446 | -3.059805 | -3.638587 | -2.547645 | 5.800973 | -3.571964 | -6.465874 |
| TCGA-AO-A0JC | -2.308412 | -2.479849 | -4.512165 | -3.012865 | -3.736020 | -1.931249 | -2.872597 | 3.458127 |
| TCGA-AO-A0JM | -0.488477 | 2.120020 | -5.818289 | -7.698595 | -5.387528 | 4.567836 | -2.249127 | -6.635367 |
| TCGA-AO-A12B | -1.651943 | -3.573312 | -5.760191 | -5.337661 | -4.632492 | -4.404098 | -4.855177 | 2.073744 |
| TCGA-A8-A06N | -0.358817 | 0.269542 | -3.695040 | -3.438625 | -3.427354 | -1.209779 | -3.385087 | -2.331246 |
| TCGA-A8-A06Z | 0.717278 | -0.059650 | -1.185145 | -1.605106 | -0.926869 | -1.672300 | 0.215425 | 3.650705 |
| TCGA-A8-A079 | 2.430691 | 0.705999 | -5.562909 | -4.316382 | -2.624667 | -2.766468 | -5.051767 | 2.038266 |
| TCGA-AN-A0AJ | 1.480831 | -0.471788 | -4.539342 | -3.936113 | -3.754418 | -3.112428 | -2.790221 | -1.099243 |
| TCGA-AN-A0AM | -0.505383 | -0.731681 | -5.408514 | -6.045502 | -5.777296 | -4.251878 | -3.962719 | -7.730167 |
| TCGA-AN-A0FK | -1.381394 | -4.620440 | -3.169129 | -3.341252 | -2.187244 | -4.037569 | -3.266926 | 2.413140 |
| TCGA-AR-A0TT | -1.349875 | 0.455181 | -7.449721 | -5.494549 | -5.139396 | -4.773258 | -5.293173 | -3.345323 |
| TCGA-AR-A0TV | -0.040732 | 0.620403 | -5.539917 | -4.749917 | -4.077577 | -2.057754 | -2.971016 | -3.422045 |
| TCGA-AR-A1AV | 1.882713 | 0.999190 | -3.354167 | -4.738353 | -4.157370 | -1.038267 | -1.951240 | 3.475731 |
| TCGA-BH-A0C7 | 0.944990 | -0.918145 | -1.125656 | -0.244293 | -0.474116 | -2.779049 | -2.540300 | 1.812966 |
| TCGA-BH-A0DD | 1.189354 | -1.680075 | -4.047413 | -2.822685 | -2.782799 | 0.992271 | -1.098212 | 0.689609 |
| TCGA-C8-A12U | 0.126063 | 0.437190 | -4.983742 | -4.189453 | -4.200434 | -4.562806 | -3.113319 | -1.528401 |
| TCGA-E2-A15A | 0.229658 | -0.441373 | -1.183981 | -2.951031 | -1.649230 | -4.682292 | -5.152014 | -4.158888 |

77 rows × 43 columns

**表 4：整合后的数据**

2）利用 KMeans 进行聚类分析

首先还是对得到 77*43 的数据进行了预处理(包括 Z-scale、缺失值填补、数据正则化)，对将处理过后的数据进行 KMeans 聚类操作。

在操作时人为的将聚类数分成了 2、3、4、20、76 等，利用 for 循环对结果进行打印，并以轮廓系数（Silhouette）、均分（Homogeneity score）两项作为评价指标，通过调用之前得到的特定 43 个蛋白质基因进行聚类得到结果如下：（数据较多，仅以部分展示）

```
Silhouette Coefficient for k == 2: 0.1524
Homogeneity score for k == 2: 0.1597
------------------------
Silhouette Coefficient for k == 3: 0.1452
Homogeneity score for k == 3: 0.3607
------------------------
Silhouette Coefficient for k == 4: 0.1496
Homogeneity score for k == 4: 0.4999
------------------------
Silhouette Coefficient for k == 5: 0.1122
Homogeneity score for k == 5: 0.4718
------------------------
Silhouette Coefficient for k == 6: 0.0813
Homogeneity score for k == 6: 0.5563
------------------------
Silhouette Coefficient for k == 7: 0.0795
Homogeneity score for k == 7: 0.5594
------------------------
```

**图 1：特定 43 个蛋白质基因聚类结果**

结果分析：若仅从结果看，分成 76 类是最好的但考虑到现实中只有 77 个患者，因此这一结果是没有意义。在综合考虑后发现将 PAM50mRNA 分成四类的轮廓系数以及均分都较高，因此判定将诊断结果分成四类效果较好，与真实的 PAM50mRNA 的诊断结果（4 类）进行对照，也可以很明显的发现，这一聚类是有一定可靠性的。但是结果相对较差，个人认为这在很大程度上是因为数据量太少，在较低维度数据量较小的情况下进行聚类分析往往得到的效果是不好的。

为了证实选择特定的 43 个基因的重要性，在之后还进行了一个对照实验，在实验中不再选取特定基因，而是从基因库中随机选取了 43 个进行聚类，得到结果如下：（数据较多，仅以部分数据展示）

```
Silhouette Coefficient for k == 2: 0.1155
Homogeneity score for k == 2: 0.0048
------------------------
Silhouette Coefficient for k == 3: 0.1031
Homogeneity score for k == 3: 0.0522
------------------------
Silhouette Coefficient for k == 4: 0.0774
Homogeneity score for k == 4: 0.1062
------------------------
Silhouette Coefficient for k == 5: 0.0765
Homogeneity score for k == 5: 0.1097
------------------------
Silhouette Coefficient for k == 6: 0.0699
Homogeneity score for k == 6: 0.1541
------------------------
Silhouette Coefficient for k == 7: 0.0584
Homogeneity score for k == 7: 0.2324
------------------------
```

**图 2：随机 43 个蛋白质基因聚类结果**

通过对照可以很明显的发现，随机选取 43 个基因得到的结果非常差，这也在一定程度上说明了对特定基因筛选的重要性。最后通过数据可视化，用热力图画出了 77 个病人与 43 个基因的关联图：

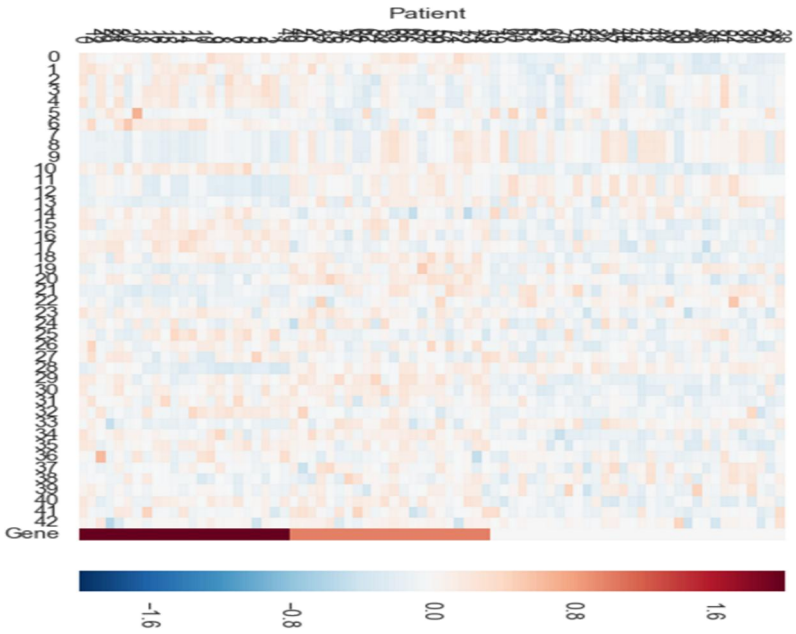

图 3：77 个病人与 43 个基因的

# 三、对患者进行分类，并以 PAM50mRNA 诊断结果作标签

## 1）对患者分类的实现

在这一问中依据数据是否处理，设计了一组对照实验。

①不对数据进行任何处理

首先对不做任何处理的 77*43 个数据进行分类，在分类时随机选择 77 个患者中的 70%作为训练集，剩余的 30%作为测试集，分别利用了朴素贝叶斯、随机森林、K 近邻三种不同的分类模型来实现，以准确率、召回率、F1 值 作为评价指标，得到第一次计算的结果如下：

| 模型 | 准确率 | 召回率 | F1 值 |
|---|---|---|---|
| 朴素贝叶斯 | 0.68 | 0.62 | 0.63 |
| 随机森林 | 0.77 | 0.75 | 0.75 |
| K 近邻 | 0.78 | 0.75 | 0.75 |

表 5：未处理数据的分类结果

②对数据进行处理后

之后对数据进行了一些预处理操作（Z-scale 标准化、正则化、异常值处理），依旧以随机的 70%作训练集，剩下的 30%作测试集，再次计算得到结果如下：

| 模型 | 准确率 | 召回率 | F1 值 |
|---|---|---|---|
| 朴素贝叶斯 | 0.72 | 0.67 | 0.68 |
| 随机森林 | 0.81 | 0.79 | 0.79 |
| K 近邻 | 0.78 | 0.75 | 0.75 |

表 6：处理过数据后的分类结果

对结果的分析：通过两次对照实验发现，在对数据进行一些处理后，朴素贝叶斯和随机森林的分类效果得到了明显的提升，但是 K 近邻的结果基本没有变化。同时使用各种模型分类的结果都不太理想，个人感觉是因为 PAM50mRNA 的检测结果需要用到 100 个蛋白质基因，而在分类时我只用到了其中的 43 个，这就直接导致了数据流失，这在一定程度上可以对结果不好给出很好的解释。

## 2）对各个基因重要性的评判

首先对以上结果比较发现，随机森林结果最好，因此选用这一结果来分析 43 个基因对 PAM50mRNA 诊断结果的影响，通过提取特征重要性来对 43 个基因的影响力进行排序，得到结果如下：（数据太多，仅以片段表示）

```
Feature ranking:
1. feature 21 (0.078069)
2. feature 25 (0.076630)
3. feature 28 (0.056536)
4. feature 5 (0.047741)
5. feature 7 (0.041535)
6. feature 15 (0.037517)
7. feature 12 (0.037188)
8. feature 20 (0.031022)
9. feature 11 (0.030953)
10. feature 19 (0.029999)
11. feature 0 (0.027369)
12. feature 8 (0.026951)
13. feature 9 (0.025823)
14. feature 6 (0.025797)
15. feature 3 (0.023397)
```

**图 4：基因影响力排序**

之后为了方便观察做出了各因素重要性的直观图：

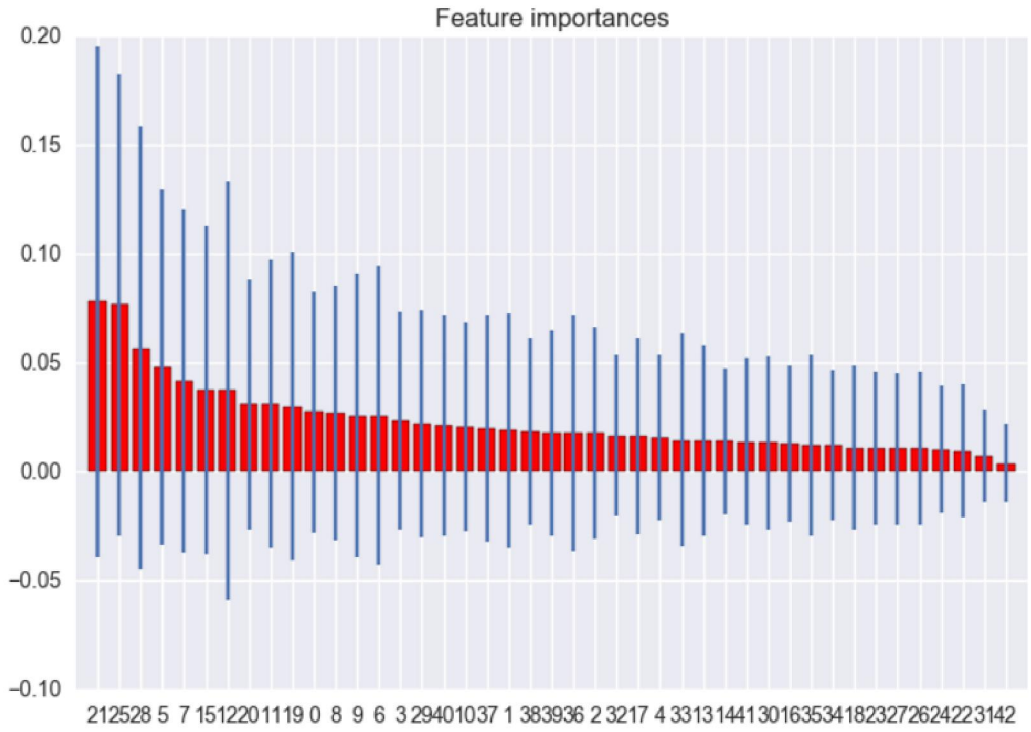


**图 5：43 个基因影响力直方图**

结果分析：通过排序和直方图可以很容易的看出第 21、25、28、7、5 号基因对于 PAM50mRNA 的诊断结果影响较大，汇总得如下表格：

| 蛋白质基因编码 | 对结果影响大小 |
| --- | --- |
| NP_000517 | 0.078069 |
| NP_005219 | 0.076630 |
| NP_058519 | 0.056536 |
| NP_001116539 | 0.047741 |
| NP_058518 | 0.041535 |

**表 7：主要基因影响**

# 四、如何判断 Clinicaly（第二个文件）中已知的其他诊断结果是否也用了 PAM50 mRNA 所用到的 43 个蛋白质基因？

**1）问题重述**

在第二个文件中可以看到与 PAM50mRNA 的分类结果相似还有了一些其他的诊断分类结果，从第三个文件发现 PAM50mRNA 的诊断用了了 100 个蛋白质基因，但是只能从第一个文件中找到 43 个对应的基因，于是在第二问、第三问中利用这 43 个基因来对 PAM50mRNA 进行了聚类和分类的探讨与分析，那么是否可以继续用这 43 个基因来对其他结果进行分析呢，以及其他诊断结果是否也像 PAM50mRNA 一样用到了这 43 个基因呢？

**2）问题的分析**

在这一问中可以通过利用这 43 个基因来对其他的诊断结果进行分类分析，但是对结果直接进行分析是不现实的，但是由于第三问中已经利用这 43 个基因对 PAM50mRNA 进行了分类，因此可以将两次的结果进行比较，通过比较来得出最终的评判结果。

**3）问题的解决**

像第三问中一样，利用随机森林来对 77*43 个数据进行分类，在分类时随机选择 77 个患者中的 70%作为训练集，剩余的 30%作为测试集，与第三问不同的是这里的 train_y 和 test_y 不再是 PAM50mRNA，而是其他的诊断结果以准确率、召回率、F1 值 作为评价指标，得到计算的结果如下：

| 诊断结果类别 | 准确率 | 召回率 | F1 值 |
| --- | --- | --- | --- |
| CN Clusters | 0.43 | 0.25 | 0.24 |
| SigClust Unsupervised mRNA | 0.60 | 0.67 | 0.61 |
| miRNA Clusters | 0.27 | 0.42 | 0.32 |
| RPPA Clusters | 0.31 | 0.38 | 0.33 |
| methylation Clusters | 0.65 | 0.58 | 0.61 |

| | | | |
|---|---|---|---|
| PAM50 mRNA | 0.81 | 0.79 | 0.79 |

表 8：各种诊断结果分类结果

4）结果的分析及扩展

可以很明显的看出利用 PAM50 mRNA 的 43 个基因来对其他诊断结果进行分类的效果并不理想，这也在一定程度上说明了其他的诊断结果并没有完全用到这 43 个基因，但是通过比较不同结果可以发现差别较大，这也说明不同的诊断结果之间也是可以进行分类的。

**个人仅以结果为数据参考，进行了如下操作:**

将准确率、召回率、F1 值全都在 50%以上的视作可以利用这 43 个基因进行分类，并把这一类称作 43 基因可测类。

将准确率、召回率、F1 值全都在 50%以下的视作不可以利用这 43 个基因进行分类的，并把这一类称作非 43 基因可测类。

按照这一标准我有将诊断结果进行了分类得到具体两大类如下：

| 43 基因可测类 | 非 43 基因可测类 |
|---|---|
| SigClust Unsupervised mRNA | miRNA Clusters |
| methylation Clusters | CN Clusters |
| PAM50 mRNA | RPPA Clusters |

表 9：对各个诊断结果分类

按照这个思路，可以在一定程度上将诊断结果的类别再次进行分类，以 43 个基因为例，若一个人已经进行了这 43 个基因的检测，那么就可以直接得到 SigClust Unsupervised mRNA、methylation Clusters、PAM50 mRNA 这三个诊断结果，然后以这三个结果对患者是否患乳腺癌进行评判，这样一来可以大大提高诊断效率，减少诊断的复杂度。

# 总结

通过本次 Project 个人从几个方面对大数据有了更深刻的认识主要有如下几点：

1. 对于复杂的数据在进行处理前一定要对数据进行充分的认识，防止对数据含糊不清、盲目处理情况的发生。以本题为例，3 个文件数据种类较多一定要有充分认识。
2. 对于较大量的数据一定要学会进行筛选，提取有用信息，以在本题为例必须要对 43 个特定基因进行筛选，如果仅是随机选取效果会很差。
3. 数据的预处理对于计算结果有很大影响，很重要。以本题为例，在进行数据预处理后结果在一定程度上有了提高。