本周工作报告\_4

——2014.10.29

## 之前的思路总结

接上之前的内容，我之前的工作思路大致是：

1. 把ToppGene网站上关于166个疾病预测基因在网站上的数据通过手动的方法下载了下来并进行了一些数据处理；
2. 熟悉SVMlight并将这些Toppgene数据转换成SVMlight支持的格式；
3. 利用SVMlight进行训练，得到一个训练的中间结果。其中遇到了一些问题，比如说会有**498**个随机选择的基因缺失的问题。

## 我的工作（本周的工作）

1. 这周的主要工作是将SVM\_light环境配置好，让svm\_light可以运行在其它文件目录中，我通过查找资料修改了Linux下的默认文件PATH路径，把整个服务器的我的用户命令系统不小心搞坏了，还好后来查资料给恢复过来了；
2. 将我用SVM\_light 排序之后的结果与Toppgene原来的结果还有大师兄的结果进行了比较，核心代码存在了“script\_toppgene\_part3.r”文件中，对比结果保存到结果存入topGeneResult.xlsx，在R空间里存在 topGeneNum 数组中。如下：

**方案一：利用Baylor提供的在452个候选的视网膜疾病致病基因的列表作为参考**，将Toppgene SVM\_light的结果与Toppgene求平均值的结果还有大师兄以前的结果TTP进行比较，再与之前的其它网站运行的结果对比，得到如下结果：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 452 gene | tools | top 100 | 200 | 300 | 400 | 500 | 600 |
| Endeavour | 3 | 6 | 11 | 19 | 23 | 29 |
| gpsy | 4 | 8 | 13 | 17 | 18 | 22 |
| genefriends | 3 | 6 | 10 | 14 | 15 | 19 |
| TTP | 4 | 7 | 10 | 13 | 15 | 16 |
| ToppGene | 10 | 26 | 38 | 48 | 61 | 69 |
| ToppGeneSVMsorted | 2 | 4 | 5 | 6 | 8 | 11 |
| ToppGeneSVMsorted2 | 1 | 2 | 7 | 11 | 13 | 17 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 700 | 800 | 900 | 1000 | 1100 | 1200 | 1300 | 1400 |
| 34 | 37 | 43 | 48 | 53 | 56 | 61 | 65 |
| 26 | 34 | 39 | 44 | 48 | 51 | 56 | 66 |
| 22 | 22 | 23 | 25 | 27 | 31 | 32 | 34 |
| 20 | 24 | 29 | 36 | 42 | 53 | 57 | 61 |
| 82 | 86 | 92 | 97 | 103 | 106 | 110 | 118 |
| 11 | 13 | 13 | 16 | 18 | 19 | 23 | 25 |
| 18 | 22 | 26 | 26 | 34 | 37 | 43 | 46 |

如上表，我对比了Toppgene的方法和大师兄的方法（TTP），还有SVM\_light 训练后的方法，这个SVM\_light 的方法我借鉴了大师兄当初用的思想，但是效果非常差，我不知道为啥会有这么差。

其中，Endeavour，gpsy，genefriends这三个tools是基因预测届比较有名的方法，可以看出来，与TTP（黄色）相比不分伯仲。

但是Toppgene的方法却是遥遥领选与其它的方法。我个人认为这也是应该的，因为TTP只是利用了数据挖掘的方法进行纯数学上的预测，而Toppgene是一个浩大的项目，用了很多已经比较成熟的中间结果。

Toppgene SVM的方法貌似比较不理想，我觉得可能是我的参数选择有问题，我想再仔细研读SVM的文档，再修改一下核函数与训练方法试试。我觉得大师兄的方法可能有很大的问题，因为用大师兄的方法进行训练我不太懂它的理论依据。

**方案二：利用42个TTP\_model\_human基因再进行分析，看在这42个gene中的数量有多少个：**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 42 gene | tools | top 100 | 200 | 300 | 400 | 500 | 600 |
| Endeavour | 0 | 0 | 0 | 0 | 0 | 1 |
| gpsy | 3 | 3 | 5 | 8 | 9 | 12 |
| genefriends | 2 | 2 | 2 | 2 | 4 | 5 |
| TTP | 1 | 1 | 2 | 3 | 3 | 3 |
| ToppGene | 7 | 7 | 7 | 10 | 11 | 13 |
| ToppGeneSVMsorted | 0 | 0 | 0 | 0 | 0 | 0 |
| ToppGeneSVMsorted2 | 4 | 4 | 6 | 7 | 7 | 7 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 700 | 800 | 900 | 1000 | 1100 | 1200 | 1300 | 1400 |
| 1 | 2 | 2 | 2 | 2 | 2 | 3 | 3 |
| 12 | 13 | 16 | 18 | 20 | 21 | 21 | 23 |
| 5 | 6 | 6 | 7 | 7 | 9 | 9 | 9 |
| 3 | 3 | 3 | 3 | 5 | 5 | 6 | 8 |
| 14 | 14 | 16 | 17 | 17 | 17 | 20 | 22 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 7 | 7 | 7 | 7 | 7 | 8 | 10 | 10 |

**结果证明，Toppgene与其它方法好了很多。**

**不过我有个问题，不太清楚这个452个基因与这42个基因是什么关系？**

## 接下来的工作计划

1. 找到svm更合适的参数和合适的方法重新训练，与原来的方法进行对比；我比较疑惑的地方有两处：

a：如果结果真的比原网站简单求p-value和求平均的方法好一些，但是这样可以发论文之类的吗？

b: 我不知道如何将这个结果与大师兄的结果进行整合