本周工作报告\_5

——2014.11.06

## 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. 利用SVM对Toppgene的结果进行训练，只用到了svm\_learn 模块，然后对学习的结果对452个基因和42个model进行分析
2. 结果\*：结果发现：SVM排序后的结果不太理想，还有toppgene原本的方法比大师兄的TTP的方法还有其它的方法都好很多，或者说明我用SVM的参数或者方法不太对，或者说明SVM的方法在这里并不适用。

## 我的工作（本周的工作）

1. 这周的主要工作是学习并使用RankSVM的方法，这个方法思路稍微比较清晰，首先进行学习，然后进行排序训练：
2. 将数据格式修改成SVM Rank支持的格式，分别保存到toppgene\_svmrank\_train.txt，toppgene\_svmrank\_test.txt，然后利用RankSVM进行打分：核心代码见操作笔记
3. 将我用RankSVM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。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 RankSVM 与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 | 700 |
| TTP | 4 | 7 | 10 | 13 | 15 | 16 | 20 |
| ToppGene | 10 | 26 | 38 | 48 | 61 | 69 | 82 |
| ToppGeneSVMsorted | 2 | 4 | 5 | 6 | 8 | 11 | 11 |
| ToppGeneSVMsorted2 | 1 | 2 | 7 | 11 | 13 | 17 | 18 |
| **TopGeneSVMRranksorted3** | **13** | **24** | **36** | **40** | **48** | **52** | **56** |
| TopGeneSVMRranksorted4 | 1 | 2 | 3 | 4 | 6 | 9 | 10 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 800 | 900 | 1000 | 1100 | 1200 | 1300 | 1400 |
| 24 | 29 | 36 | 42 | 53 | 57 | 61 |
| 86 | 92 | 97 | 103 | 106 | 110 | 118 |
| 13 | 13 | 16 | 18 | 19 | 23 | 25 |
| 22 | 26 | 26 | 34 | 37 | 43 | 46 |
| **61** | **67** | **72** | **85** | **91** | **94** | **96** |
| 12 | 12 | 15 | 15 | 17 | 17 | 17 |

如上表：

**TopGeneSVMRranksorted3**：代表我用RankSVM进行学习还有分类之后的结果，TopGeneSVMRranksorted4：是我对TopGeneSVMRranksorted3的结果进行一个逆序，作为对比来观察我的方法的有效性。

ToppGene：Toppgene网站的结果

ToppGeneSVMsorted：用大师兄的方法训练网站数据得到的结果

ToppGeneSVMsorted2：用大师兄的方法训练网站数据得到的结果的逆序

上表表明：**使用RankSVM比只用svm\_learn进行学习然后排序的方法好很多**（效果是svm\_learn的5倍），不过还是比不上Toppgene这个网站自己的方法，但是差距已经变小了。

所以，我认为，RankSVM还是挺有效的，而且还有改进的空间，可能要通过选择合适的核函数来改进。

**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 | 700 |
| TTP | 1 | 1 | 2 | 3 | 3 | 3 | 3 |
| ToppGene | 7 | 7 | 7 | 10 | 11 | 13 | 14 |
| ToppGeneSVMsorted | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ToppGeneSVMsorted2 | 4 | 4 | 6 | 7 | 7 | 7 | 7 |
| TopGeneSVMRranksorted3 | 2 | 3 | 4 | 4 | 4 | 4 | 6 |
| **TopGeneSVMRranksorted4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

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

对于这42个基因，RankSVM的效果稍微差了一点，TopGeneSVMRranksorted4这个数据说明，RankSVM不会把有用的基因排到后面去

**我还是不太清楚这个452个基因与这42个基因是什么关系？**

## 接下来的工作计划

1. 调整RankSVM的一些参数，比如核函数什么的，看看效果能不能超过Toppgene这个网站的结果
2. 我比较疑惑的地方是：我不知道如何将这个结果与大师兄的结果进行整合
3. 如果确定RankSVM比大师兄的方法好的话，那得考虑重写大师兄的方法了