Simple Outline

3 独立数据集上交叉验证

1. 实验对比
    - **表格1** 已经完成
2. 各个数据集的相关信息
    - **表格2**:

据	**疾病名称**	**GEO数据-标题**	**原来选中的基因特征**	**GEO数据中对应的基因特征**
stric	Gastric cancer	GSE66229 (Molecular analysis of gastric cancer identifies discrete subtypes ...)	...	...

3. 细致分析
    - 挑几个实验细致说怎么做的，效果怎样，实验结果说明了什么。

4 理论分析与SVM-RFE对比

1. SVM-RFE 理论背景与弱点分析
2. GA分析
    - 模式定理
    - 马尔科夫模型
    - 实验角度的评估和分析是另一条思路
3. 实验
    - 对比实验
        - GA-RFE | SVM-RFE | (xgBoost-RFE)
    - 图和表格

More about Q4

- GA的理论分析(**弱点**)，两种：
    - 模式定理。

    种群中好的模式 会随迭代成指数增长，但不能提供收敛到全局最优点的分析。

    - 马尔可夫链。

    证明了经典GA每次迭代时在有精英保留策略下，最终可以收敛到全局最优点，否则不行。精英保留策略现在我们算法里有。

    - 问题

现在算法相比于原来的GA做了很多细节调整，然后原来的两种理论分析模型套用过来有些地方就比较难对应上了，这里稍难回复。

  - **Experimental analysis**是理论分析外的另一重要思路，可能更适合模拟退火和遗传算法这种启发式的有很多参数的优化算法[1]。

- 回到这个问题开始，为什么筛选特征：

  1. Structural Risk Minimization (SRM): In the spirit of Structural Risk Minimization (see e.g. Vapnik, 1998; Guyon, 1992) it is possible to use the ranking to define nested subsets of features F1 \in F2 \in ··· \in F.
  2. Disease related genes are potential bio-markers useful in further biological analysis

SVM-RFE遵循了SRM，是经典的特征筛选算法，有很多优势： 1.2.3.

Meanwhile, it should be noted that there are some shortcomings of SVM-RFE.

- SVM-RFE中的SVM

  - Insights: "To test the idea of using the weights of a classifier to produce a feature ranking, we used a state-of-the-art classification technique: Support Vector Machines (SVMs)."

  - STATE-OF-THE-ART in 2002, out-of-date NOW.

  - Replaced by RF, xgboost, LightGBT

    - more effective in evaluating the feature importance

- 使用SVM-RFE一些弱点

  1. SVM-RFE这一框架判断特征权重仅基于训练数据集，无交叉验证，**天生缺陷**
     - SVM-RFE select the lowest weighted feature in the **training dataset**, without any validation in the validation/test dataset, overfitting problem might appear.
  2. 对线性核SVM，特征需要预处理
     - 减均值除方差。这样才能同等数量级考虑进行筛选。 From each feature, we subtract its mean and divide the result by its standard deviation. This ensures that feature scales are comparable.
  3. 特征被删掉后不再会出现
  4. 多分类问题 不直接支持
  5. 使用的分类器种类受限。SVM-线性核，推广到别的核需要更多讨论。

上述5个问题在GA-SVM框架中都能被解决掉。

- 本质上这不是一个排名的问题，直接**筛选基因组合**更合适。

  - 一次筛选，长久使用
  - However, the features that are top ranked (eliminated last) are not necessarily the ones that are individually most relevant. Only taken together the features of a subset Fm are optimal in some sense.
  - Subset selection manner can preserve complementary genes.

- Both GA-RFE, SVM-RFE sub-optimal

  - NO theoretical guarantee for a good feature subset found by SVM-RFE
  - The criteria DJ(i) or (wi)2 only **estimate** the effect of removing one feature at a time on the objective function.
  - SVM-RFE probably weaker than the GA based RFE in this problem --- powerful swarm intelligence based method
  - GA-RFE could find more compact feature subset

- 对比**实验**
  - 原先SVM-RFE的实验数据集并不完善(Only simple Leuk and Colon)。
    - Test on more complex datasets and do comparison.
  - 比SVM对特征打分更好的分类器来做RFE过程
    - RF 随机森林
    - xgboost / lightGBM
  - 以上的RFE过程实验结果弱于MGRFE/GA-RFE吗

[1]. Johnson D S. A theoretician's guide to the experimental analysis of algorithms[J]. Data structures, near neighbor searches, and methodology: fifth and sixth DIMACS implementation challenges, 2002, 59: 215-250.