基于无监督机器学习算法有效改善癌症基因的识别和优化

第二组第二十六号

大纲

1. 研究背景
2. 研究的目的和意义
3. 数据与材料
4. 技术路线
5. 结果
6. 结论

研究背景

- 使用统计模型对RNA-Seq数据处理得到的差异表达的基因一般被人们认为是和癌症相关的基因。但是，在使用统计方法筛选相关基因的过程中面临着许多的缺陷。

    1. **相关性缺陷**：由于癌症发生的复杂性，并不能完全断定在统计学上 RNA 差异表达的基因就一定和癌症相关，而没有差异表达显著性的基因就一定和癌症不相关。
    2. **信息量不足**：肿瘤的形成与发生是一个复杂的过程，仅仅使用单一的基因表达 RNA-Seq 数据并不能够完整的涵盖肿瘤细胞中绝大部分的生物信息。
    3. **样本不平衡**：肿瘤样本的数量要明显多于正常样本的数量。在如此严重不平衡样本之间做统计检验可能会降低统计检验的效能，提高错误率。
    4. **统计显著性模糊**：在做统计检验的过程中，如果一个基因的统计显著性 P 值为 0.0101，但是统计显著性阈值是 0.01。在这种情况下很难将这个基因划分进差异表达基因类或者非差异表达基因类。
    5. **强烈的人为主观性**：确定统计显著性阈值时带有很强的人为主观性，选取不同的统计显著性阈值可以得到不同的相关基因，这样给最终的结果带来了不稳定性也给客观的研究肿瘤发生与形成带来了困难。

研究目的和意义

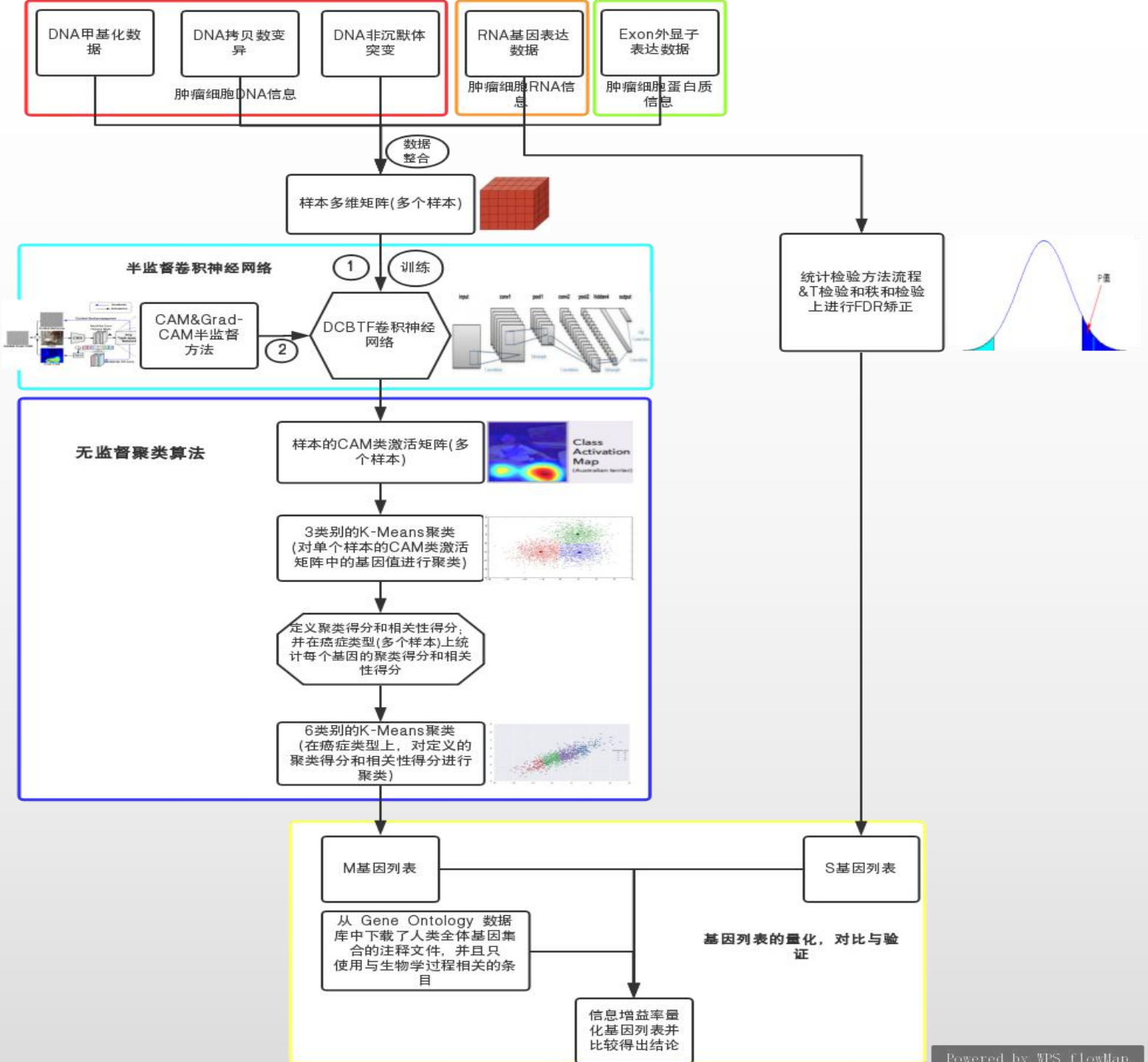
- 使用最新的**半监督卷积神经网络算法，无监督聚类的机器学习方法来克服上述统计检验中的缺陷。**
  - 为了得到更准确的结果，本人设计了一个适应这个任务的**卷积神经网络**来对癌症样本分类，在对癌症样本分类之后使用Grad-CAM方法得到**类激活矩阵(ClassActivationMapping)**，最后使用**无监督聚类算法**得到与癌症相关的基因
- 通过相关的半监督与无监督机器学习算法，可以较为完美的绕过现阶段统计检验面临的缺陷，减少人为主观性，提高结果可靠性。

数据与材料

TCGA Set	Copy number	DNA Methylation	Somatic Non-Silent Mutation	Gene Expression RNA-seq	Exon Expression RNA-seq
TCGA Liver Cancer	gistic2	Methylation450k	bcm automated	IlluminaHiSeq	IlluminaHiSeq
TCGA Lung Cancer	gistic2	Methylation450k	PANCANAWG	IlluminaHiSeq	IlluminaHiSe
TCGA Stomach Cancer	gistic2	Methylation450k	wustl_hiseq automated	IlluminaHiSeq BC	IlluminaHiSeq-UNC
TCGA Breast Cancer	gistic2	Methylation450k	wustl curated	IlluminaHiSeq	IlluminaHiSeq

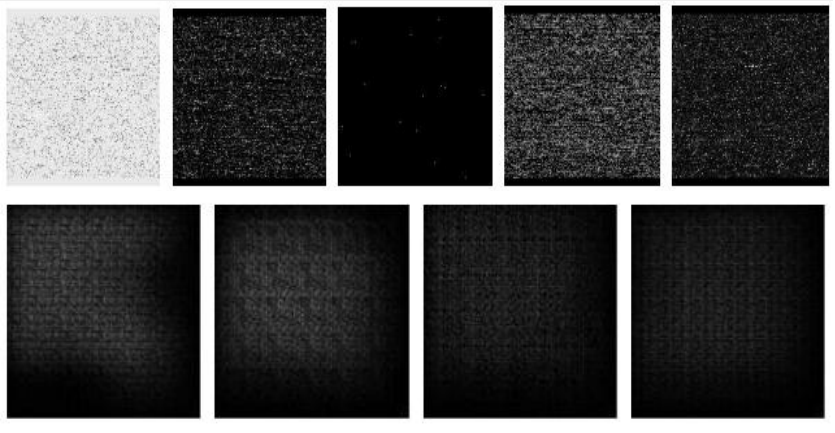
注：第一列至第六列分别为，TCGA 数据集合、拷贝数数据、甲基化数据、非沉默体突变数据、基因表达数据和外显子表达数据。

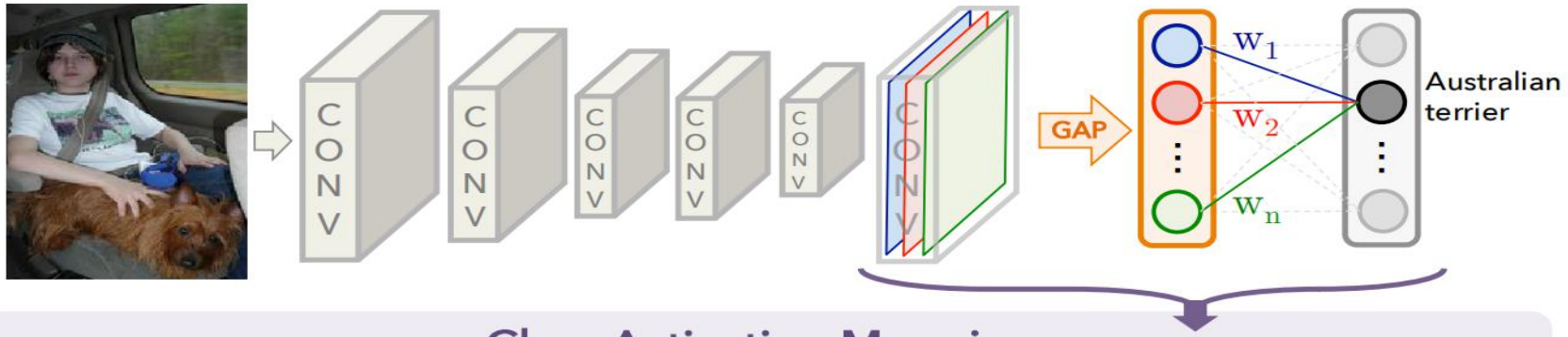
技术路线

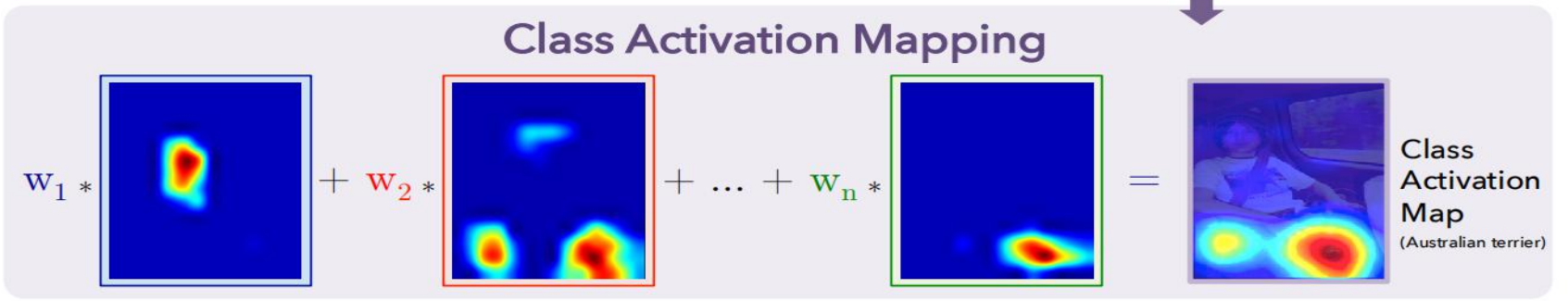


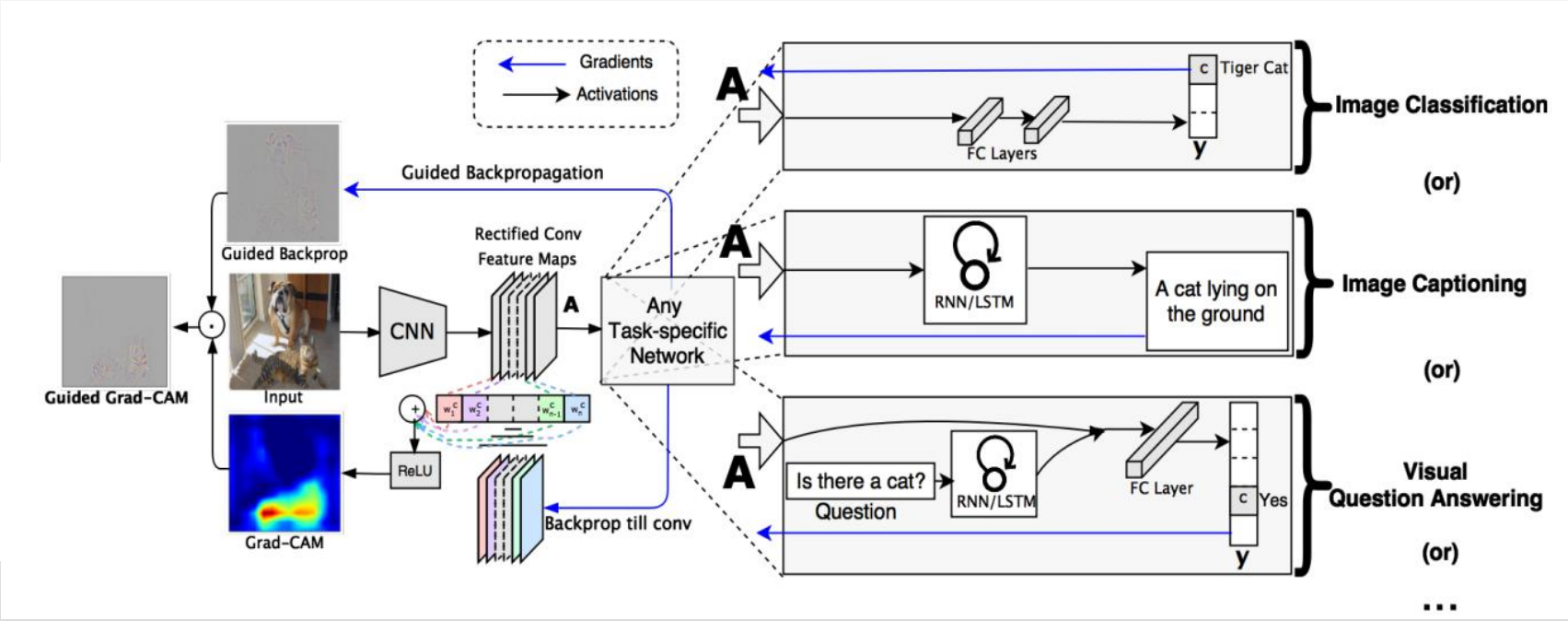
技术路线


Brushing teeth
Cutting trees




CONV
CONV
CONV
CONV
CONV
CONV
GAP
w_1
w_2
w_n
Australian terrier


Class Activation Mapping
$w_1 *$
+ $w_2 *$
+ ... + $w_n *$
=
Class Activation Map
(Australian terrier)


Gradients
Activations
Guided Backpropagation
Guided Backprop
Guided Grad-CAM
Input
CNN
Rectified Conv Feature Maps
A
Any Task-specific Network
ReLU
Grad-CAM
Backprop till conv
C
Tiger Cat
FC Layers
y
Image Classification
(or)
RNN/LSTM
A cat lying on the ground
Image Captioning
(or)
Is there a cat?
Question
RNN/LSTM
FC Layer
C
Yes
y
Visual Question Answering
(or)
...

结果

1. 乳腺癌：

- 在统计显著性阈值为 0.01 的情况下，使用上述用统计检验方法从乳腺癌的RNA-Seq 数据上发现了 12194 个差异表达基因。在使用机器学习方法得到的最终集合内，不相关类中有 10718 个基因，可能不相关类有 432 个基因，可能相关类有 479 个基因，相关类有 5164 个基因，可能强相关的有 457 个基因，强相关的有 2462 个基因。
- 全体人类基因的信息熵为 11.7022，使用统计检验方法得到的信息熵为11.5953，信息增益率为 0.1262。使用机器学习方法得到的信息熵为 11.4543，信息增益率为0.1391。相比于统计检验的方法，机器学习的方法提升率为10.1913%。

2. 肝癌：

- 用统计检验方法从肝癌的 RNA-Seq 数据中发现了 9839 个差异表达基因。在使用机器学习方法得到的最终集合内，不相关类中有 10697 个基因，可能不相关类有 784 个基因，可能相关类有 844 个基因，相关类有 4794 个基因，可能强相关的有 535 个基因，强相关的有 2058 个基因。
- 全体人类基因的信息熵为 11.7022，使用统计检验方法得到的信息熵为11.5977，信息增益率为 0.1044。使用机器学习方法得到的信息熵为 11.4357，信息增益率为0.1423。相比于统计检验的方法，机器学习的方法提升率为33.3034%。

结果

1. 肺癌：

- 用统计检验方法从肺癌的 RNA-Seq 数据中发现了12704 个差异表达基因。在使用机器学习方法得到的最终集合内，不相关类中有11493 个基因，可能不相关类有 796 个基因，可能相关类有 783 个基因，相关类有 4214 个基因，可能强相关的有 563 个基因，强相关的有 1683 个基因。
- 全体人类基因的信息熵为 11.7022，使用统计检验方法得到的信息熵为11.5963，信息增益率为 0.1299。使用机器学习方法得到的信息熵为 11.4441，信息增益率为 0.1418。相比于统计检验的方法，机器学习的方法提升率为 9.1836%。

2. 胃癌：

- 用统计检验方法从胃癌的 RNA-Seq 数据中发现了 9379 个差异表达基因。在使用机器学习方法的到的最终集合内，不相关类中有 10851 个基因，可能不相关类有 273 个基因，可能相关类有 292 个基因，相关类有 5538 个基因，可能强相关的有 252 个基因，强相关的有 2506 个基因。
- 全体人类基因的信息熵为 11.7022，使用统计检验方法得到的信息熵为11.5847，信息增益率为 0.1179。使用机器学习方法得到的信息熵为 11.4772，信息增益率为0.1350。相比于统计检验的方法，机器学习的方法提升率为 14.5508%。

结论

- 本研究中，使用半监督和无监督的机器学习算法在很大程度上克服了传统统计检验的不足之处。在很大程度上克服了人为主观性对最后结果的影响并且克服了统计检验的显著性模糊等问题。在最后的结果比较中，模仿决策树算法对全体数据分类的过程，使用信息增益率这一评估标准给出量化结果。量化结果表明，在乳腺癌中，机器学习算法相较统计检验方法提升了 10.1913%；在肝癌中提升了 33.3034%；肺癌中提升了 9.1836%；胃癌中提升了 14.5508%。在四种不同的癌症中，平均提升 16.8073%。因此，本研究的机器学习的方法可以有效的减少人为主观性并且最后识别的癌症基因相较于统计检验来说也有相当的提升，奠定了进一步对癌症研究的基础。