Inductive matrix completion for predicting gene–disease associations

赵连敏

2018 年 8 月 15 日

厦门大学信息科学与技术学院

摘要

- 创新点
  - 传统方法依赖于特定的症状，由此限制了其应用
  - IMC 方法结合了多种有关于疾病与基因的特征去了解潜在的因素以解释被发现的疾病与基因间的联系
  - 这些特征包括了微阵列所表达出的信息与同疾病相关的文本数据
- 结论
  - 预测成功率达到 25%，远超第二名 15% 的结果
  - 对未在训练集见过的疾病泛化能力也比较强

绪论

- 重要步骤：对致病基因的计算机排序
- 研究问题：类似于对推荐系统的研究
- 研究目的：补全偏好矩阵
- 存在问题：联系矩阵极端稀疏以及缺乏消极联系
- 研究现状
  - 大部分方法无法预测未被研究过基因联系的疾病
  - 少部分通过对与已研究过疾病的相似度计算来实现预测
  - 目前这些方法无法做到对不同数据类型数据的综合应用
- 研究成果
  - 解决矩阵稀疏的问题
  - 对于未在训练集中遇到的样本也可以进行预测
  - 对于已研究过的疾病也能发现其与新基因的联系
  - 性能卓越

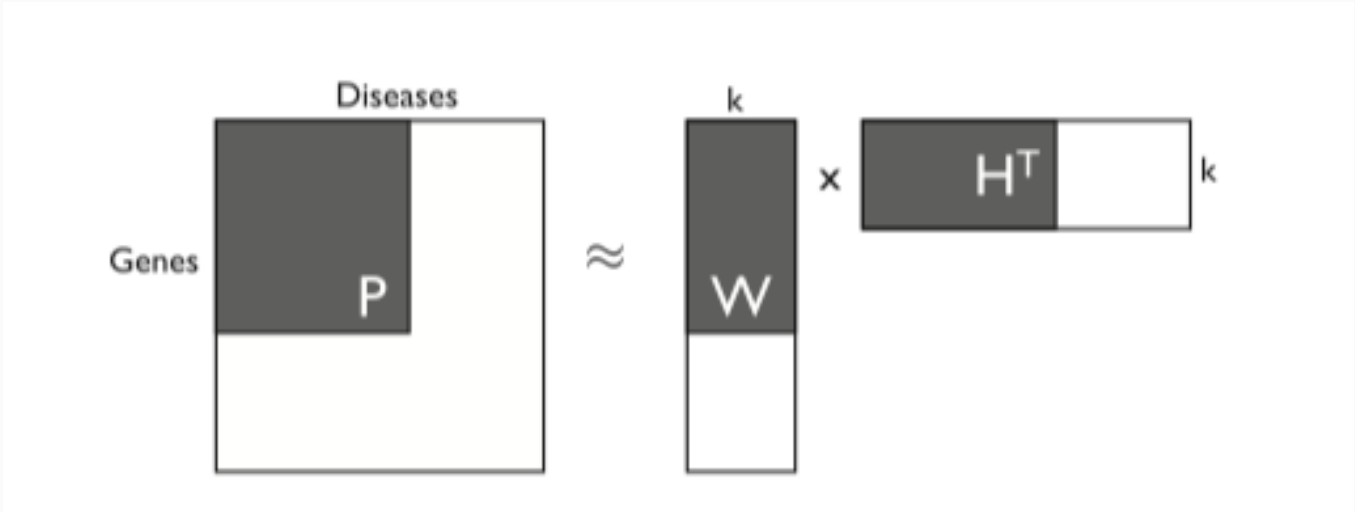
传统矩阵补全推导方法

联系矩阵 $P \in \mathbb{R}^{N_g \times N_d}$

$$P_{ij} = \begin{cases} 1 & \text{基因 i 与疾病 j 有关} \\ 0 & \text{二者关系未被观察到} \end{cases}$$

优化目标

$$\min_{W \in \mathbb{R}^{N_g \times k}, H \in \mathbb{R}^{N_d \times k}} \sum_{(i,j) \in \Omega} (P_{ij} - W_i^T H_j)^2 + \frac{1}{2}\lambda(\|W\|_F^2 + \|H\|_F^2)$$



本文的研究思路

为了使用多个辅助来源的数据，可以将每个基因转化为一个特征向量，其特征由各个数据来源共同确定，以其为样本，将疾病作为标签，这样，这一问题就转化为为每一个样本进行多标签标注的问题。

传统的多标签学习任务还有着一定的缺陷（如标签一经确定，就已经固定下来，无法扩展），同时也为了对各个数据源都进行使用。最终选取了一种可迁移的多标签学习方法。

符号约定

符号	意义
$x_i \in \mathbb{R}^{f_g}$	基因 i 的特征向量
$yj \in \mathbb{R}^{f_d}$	疾病 j 的特征向量
$X \in \mathbb{R}^{N_g \times f_g}$	全体基因特征矩阵，每一行代表一个基因
$Y \in \mathbb{R}^{N_d \times f_d}$	全体疾病特征矩阵，每一行代表一个疾病
$Z \in \mathbb{R}^{f_g \times f_d}$	要求得的低秩矩阵，用以求解关联矩阵 P，$Z = WH^T$
$P \in \mathbb{R}^{N_g \times N_d}$	关联矩阵，前文已定义，此处 $P_{ij} = x_i^T Z y_j$
Ω	全体训练集
$W \in \mathbb{R}^{f_g \times k}$	前文已定义
$H \in \mathbb{R}^{f_d \times k}$	前文已定义
λ	正则系数

优化目标

$$\min_{W\in\mathbb{R}^{f_g\times k},H\in\mathbb{R}^{f_d\times k}} \sum_{(i,j)\in\Omega} \ell(P_{ij}, x_i^T W H^T y_j) + \frac{1}{2}\lambda(\|W\|_F^2 + \|H\|_F^2)$$

这就是本文要优化的目标函数，其中损失函数 ℓ 可以选择平方损失函数等，用以控制结果不会偏离太远，而后一项正则项则是用来控制 W、H，使之不会过大。在 IMC 这一方法之中，即使出现了新的疾病 j'，只要给出其特征向量 $y_{j'}$，也可以依据此来计算出其与各基因的关系。同时，新的待优化参数 H、W 数量从与基因及疾病个数相关变为了与二者特征数相关，大大减少了参数个数。

数据集与样本特征

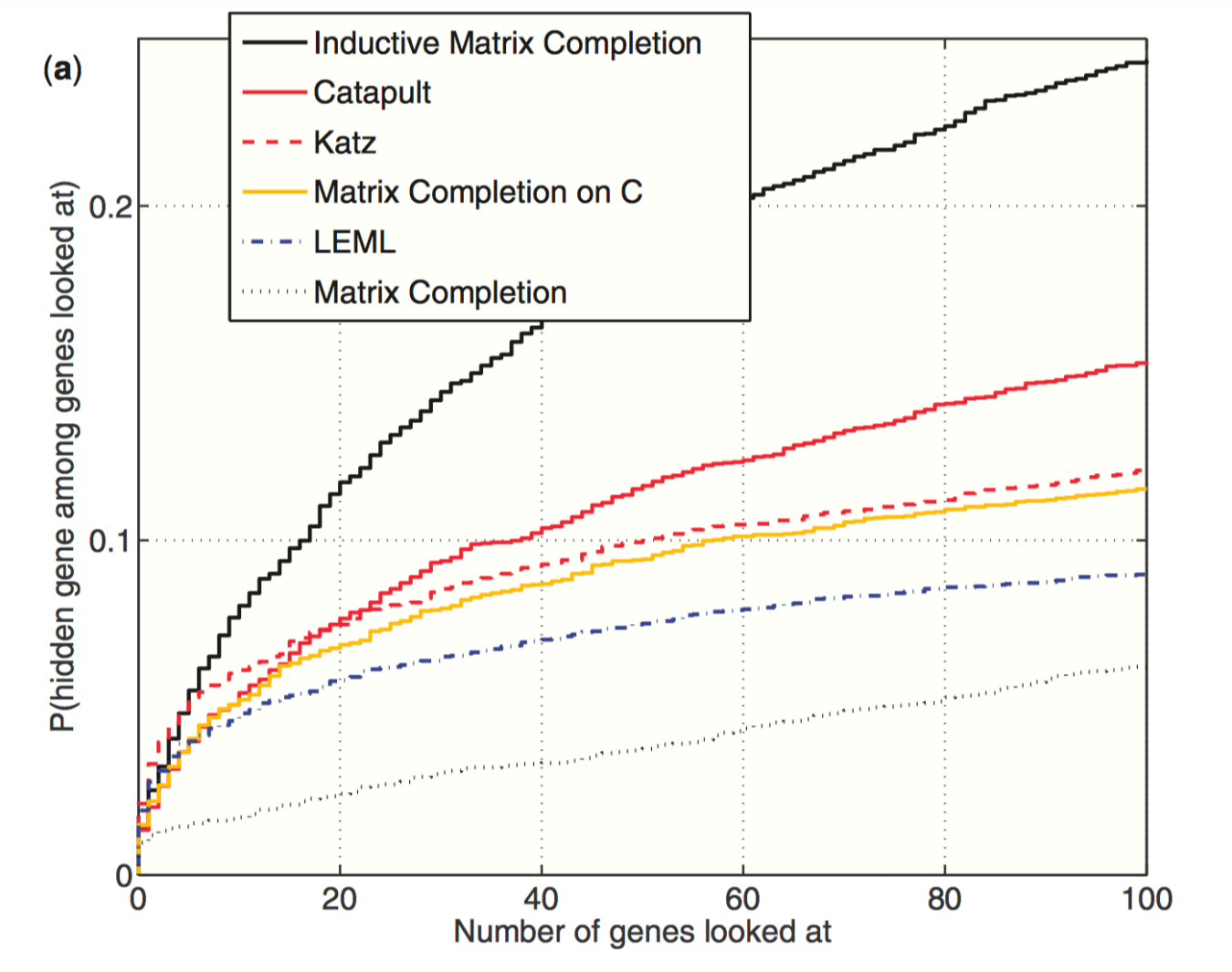
- 关联矩阵 P 数据来源：OMIM associations
- 基因特征 X 数据来源：
  - BioGPS
  - HumanNet
  - gene orthologygene–phenotype associations of other species
- 疾病特征 Y 数据来源：MimMiner

其余方法及比较标准

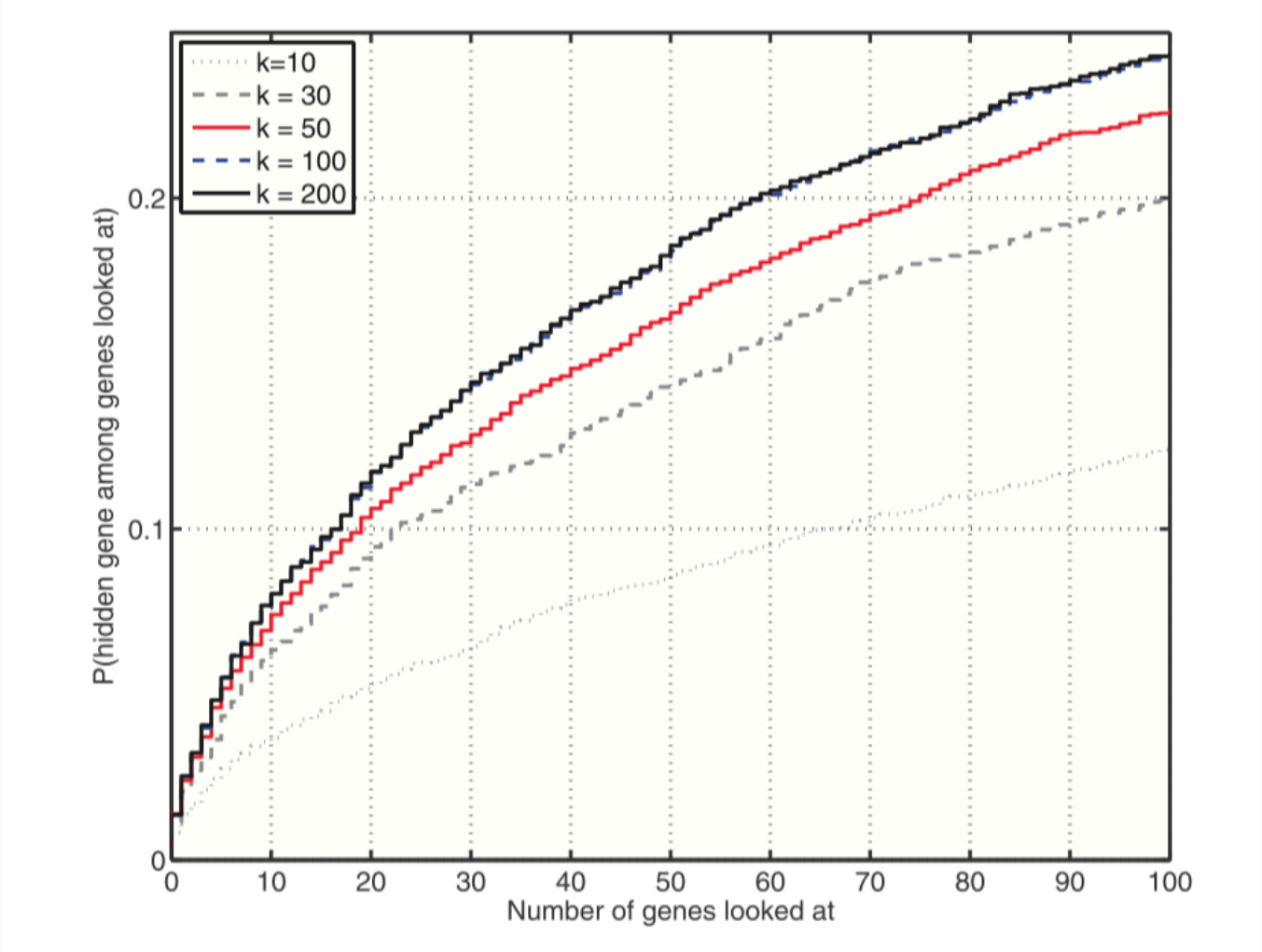
- *CATAPULT*
- Katz on the combined network
- Matrix completion on the combined network
- LEML

在比较试验之中，选取了特征数 $f_g = 300, f_d = 200$，损失函数选取了平方损失函数，对于缺失数据，以零补全，正则化系数选取为 0.2。

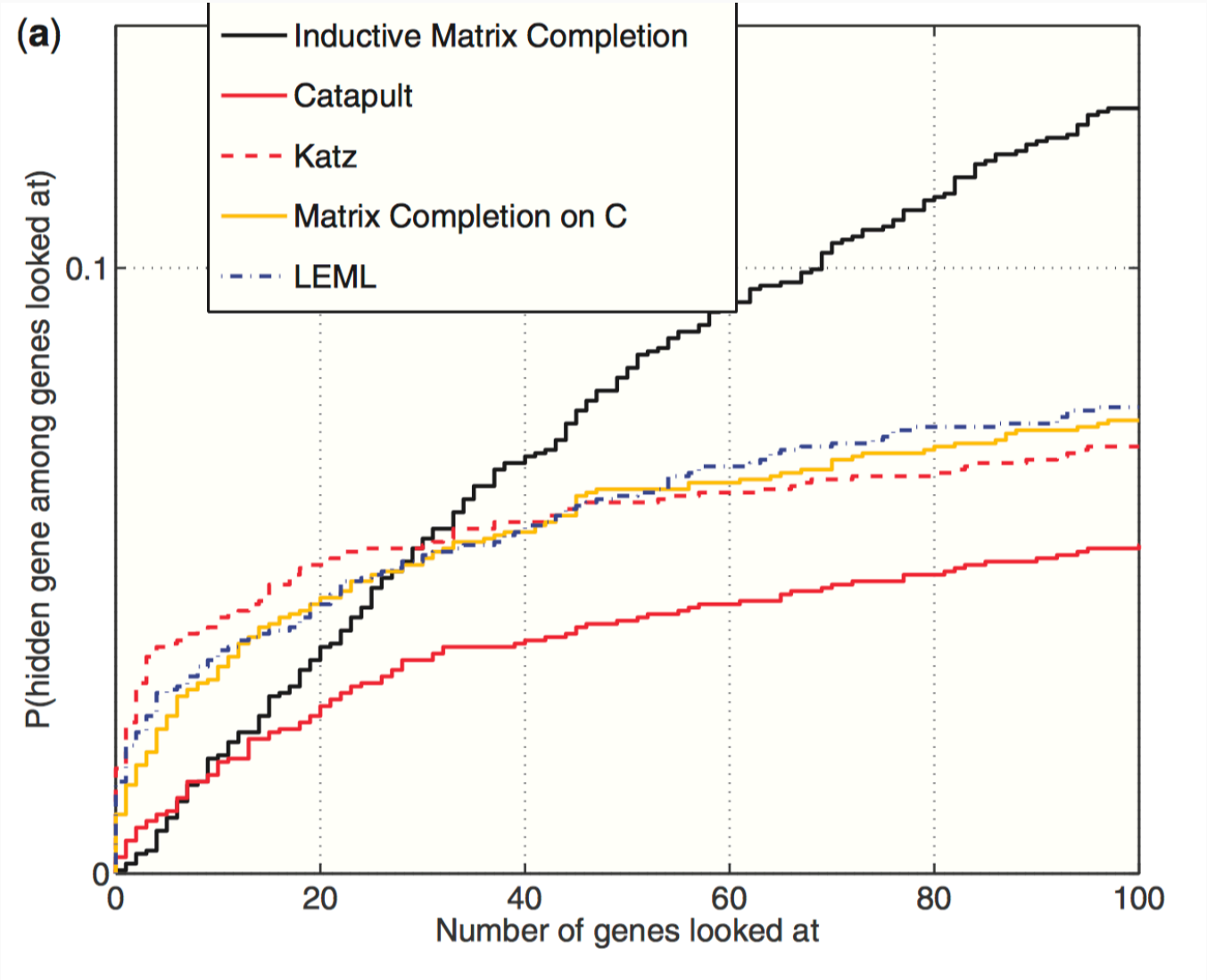
比对结果——不同方法影响


(a)
Inductive Matrix Completion
Catapult
Katz
Matrix Completion on C
LEML
Matrix Completion
P(hidden gene among genes looked at)
0
0.1
0.2
0
20
40
60
80
100
Number of genes looked at

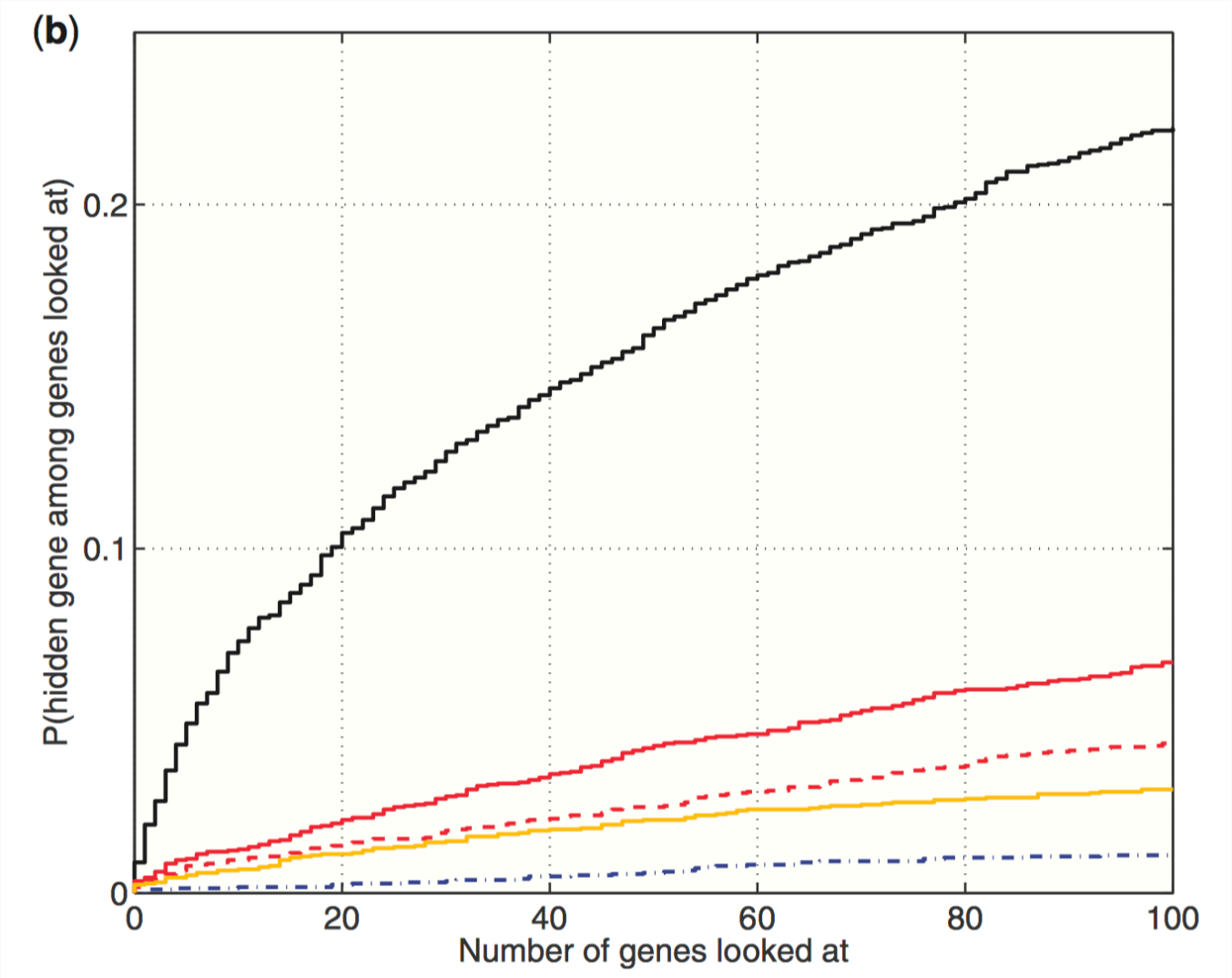
比对结果——k 值的影响


k=10
k = 30
k = 50
k = 100
k = 200
P(hidden gene among genes looked at)
0
0.1
0.2
0
10
20
30
40
50
60
70
80
90
100
Number of genes looked at

比对结果——隐去基因间联系



比对结果——隐去疾病信息


(b)
P(hidden gene among genes looked at)
0
0.1
0.2
0
20
40
60
80
100
Number of genes looked at

结论

- 使用多方面特征数据
- 将原始的矩阵补全方法进行扩展，实现与训练数据的解耦合