Advances in Applied Mathematics 应用数学进展, 2022, 11(7), 4419-4426
Published Online July 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.117468

Hans汉斯

基于肿瘤组织反卷积和SVM算法的疾病预测分析

周素素，张伟伟*

东华理工大学理学院，江西 南昌

收稿日期：2022年6月6日；录用日期：2022年6月29日；发布日期：2022年7月11日

摘　要

预后和诊断在疾病的预防和治疗中起着至关重要的作用。本文基于临床肿瘤组织的DNA甲基化芯片数据，利用反卷积算法对肿瘤组织进行分解，将估计得到的肿瘤组织中各细胞类型所占比例和细胞类型特异性的甲基化位点作为生物标志物，利用SVM算法构建该疾病的预测模型。TCGA肺腺癌、肾透明细胞癌的数据分析表明，所提方法在预测精确度和鲁棒性上都优于常用算法。

关键词

疾病预测，肿瘤组织异质性，反卷积，SVM

Disease Prediction Analysis Based on Tumor Tissue Deconvolution and SVM Algorithm

Susu Zhou, Weiwei Zhang*

School of Science, East China University of Technology, Nanchang Jiangxi

Received: Jun. 6th, 2022; accepted: Jun. 29th, 2022; published: Jul. 11th, 2022

Abstract

Prognosis and diagnosis play a crucial role in the prevention and treatment of diseases. Based on the DNA methylation microarray data of clinical tumor tissue, this paper decomposes the tumor tissue by using the deconvolution algorithm, takes the estimated cell type proportions and cell type-specific methylation sites as biomarkers, and constructs the disease prediction model by using

*通讯作者。

文章引用: 周素素, 张伟伟. 基于肿瘤组织反卷积和 SVM 算法的疾病预测分析[J]. 应用数学进展, 2022, 11(7): 4419-4426. DOI: 10.12677/aam.2022.117468

SVM algorithm. The data analysis of TCGA lung adenocarcinoma and renal clear cell carcinoma shows that the proposed method is superior to the common algorithms in accuracy and robustness.

Keywords

Disease Prediction, Tumor Tissue Heterogeneity, Deconvolution, SVM

1. 引言

癌症是一类由于细胞分裂和凋亡机制失常而导致的疾病，通常表现为恶性肿瘤[1]。由于其早期诊断困难，复发率和死亡率高，目前已经成为严重威胁人类健康的疾病之一。根据国际癌症研究中心(IARC)的报告，2015 年，全球约有 1410 万例新增癌症病例，死亡人数达到 820 万。据报告估计，20 年后，每年新发癌症病例将增至 2200 万例，与此同时，癌症致死人数将从每年约 820 万人增至每年 1300 万人。而在我国，每年因癌症而死亡的人数高达 280 万，占全球的 34.1%，且呈现出明显的上升趋势。因此构建合适的疾病诊断模型对于癌症的预防和治疗具有重要的实际意义。

传统上，利用各种类型的外科活组织检查(如骨髓活检或针活检)对癌症进行诊断。然而，由于手术的侵入性和肿瘤活检的潜在取样偏差，手术活检往往不是首选[2]。作为替代方法，研究人员和临床医生一直在寻找用于疾病诊断的分子生物标记物。目前的诊断系统都是针对肿瘤组织的基因表达数据进行的，具体做法为：将疾病样本和正常样本进行直接比对，将比对出来的差异表达基因作为预测因子，利用机器学习中的分类算法进行构建。然而，肿瘤组织是高度异质的复杂组织，除癌症细胞外，它还包含一定数量的其它细胞，如正常细胞、免疫细胞、基质细胞、血管细胞等[3]，因此其通过高通量实验所得到的基因组/表观组数据是其所包含细胞类型信号的加权平均值。由于不同的细胞类型具有不同的分子表达谱，并且细胞组成比例在不同的组织样本间也会发生一定的变化。当在不同生物学条件下观察到的组织样本间的差异实际是由于细胞组成比例的变化引起时，这可能导致错误或误导性的发现[4]。乳腺癌数据分析发现，利用此方法进行疾病预测准确度仅达到 80% [5]。

为解决上述问题，本文从肿瘤组织的 DNA 甲基化芯片数据出发，在正态分布假设下构建了肿瘤样本甲基化水平的混合概率模型，利用异质性分解算法 Tsisal [6]识别疾病和正常之间的真实变化信息，将识别出的细胞类型组成比例和细胞类型特异性的甲基化位点作为生物标记物，利用 SVM 算法构建疾病预测模型，并通过 TCGA 两种癌症类型数据分析来说明所提模型的优越性。

2. 基于 DNA 甲基化芯片数据的疾病预测模型构建

模型的训练数据是临床某肿瘤组织样本的 450 K 甲基化芯片数据 Y 和 N 名受试者的疾病状态 Z，其中 N 名受试者中包含 N_0 个病人和 N_1 个健康人。Y 是 $M \times N$ 的矩阵，M 是 CpG 位点的个数。Z 是长度为 N 的二进制向量(1 表示疾病状态，0 表示健康状态)。本文的目标是在训练数据上构建疾病预测模型使用矩阵 Y 来预测疾病状态 Z，并对模型进行评价，继而对未知疾病状态的临床样本甲基化数据进行预测分析。

疾病预测模型构建包含三个步骤：细胞类型特异性甲基化位点识别、利用异质性分解算法估计细胞类型组成比例、利用 SVM 算法进行疾病分类预测。具体实施方案如下：

细胞类型特异性的位点识别是疾病预测模型构建的第一步，本文采取方差分析(选取方差最大的 1000 个位点)、变异系数(选取变异系数最大的 1000 个位点)、几何算法 SISAL (每一个细胞类型特异性排名前 50 的位点)来识别细胞类型特异性的特征。

第二步实施过程中由于肿瘤组织是高度异质的复杂组织，其甲基化矩阵 Y 是所包含细胞类型甲基化谱矩阵 W 与细胞类型所占比例矩阵 H 的乘积，即 $E(Y)=WH$ 。从该式子可知，疾病和正常组之间在 W 或 H 上的差异将会导致 $E(Y)$ 之间的差异。因此，有必要开发反卷积分解算法去识别疾病和正常样本之间真正的变化信息。对矩阵 Y 利用各种分解算法求解 H 的过程叫做反卷积，反卷积的目标是求解出比例矩阵 H，该矩阵是第三步 SVM 的输入。目前常用的反卷积算法主要分为两大类：基于参考基矩阵的方法(reference-based, RB) [7] [8]和不依赖参考基矩阵的方法(reference-free, RF) [9] [10]。RB 方法要求参考基矩阵 W 已知，组份矩阵 H 通过回归模型进行求解。RF 方法不需要参考数据，由于 W 和 H 矩阵均为非负矩阵，因此该方法为经典的非负矩阵分解问题。相对于 RB 方法，RF 方法需要的先验信息少，求解方法也较为成熟，因此有更广泛的应用空间。本文采用 2021 年张伟伟等人[6]开发的 RF 反卷积算法 Tsisal 进行求解比例矩阵 H，它从肿瘤组织的甲基化数据出发，利用几何算法 SISAL 求解甲基化数据构造的单纯形，估计得到细胞类型所占比例矩阵 H 和细胞类型特异性的甲基化位点，当参考数据已知时它可以为估计得到的匿名细胞类型分配类别标签，已验证 Tsisal 在细胞类型估计个数、细胞类型比例估计、细胞类型标签分配上都取得了更加准确的结果。

第三步，当细胞类型特异性的位点和细胞类型比例矩阵得到后，我们可以使用现成的机器学习算法 SVM 来构建疾病预测模型。SVM 算法的求解可以转化为凸二次规划问题：

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K\left(x_i,x_j\right)-\sum_{i=1}^{N}\alpha_i \tag{1}$$

$$s.t.\sum_{i=1}^{N}\alpha_i y_i=0, 0\le\alpha_i\le C, i=1,2,\cdots,N \tag{2}$$

确定分离超平面 $w^*\cdot x+b^*=0$ 和分类决策函数 $f(x)=sign\left(\sum_{i=1}^{N}\alpha_i^* y_i K\left(x_i,x_j\right)+b^*\right)$，其中 x 为训练向量，y 训练向量的标签，$w^*=\sum_{i=1}^{N}\alpha_i^* y_i x_i$ ，$b^*=y_j-\sum_{i=1}^{N}\alpha_i^* y_j K\left(x_i,x_j\right)$ ，α_i^* 为凸二次规划问题最优解的一个分量。如何选取核函数是设计 SVM 分类器的关键步骤，根据 DNA 甲基化芯片数据特点，本文选择灵活性高、决策边界多样、参数少的高斯核函数：

$$K\left(x_i,x_j\right)=\exp\left(-\gamma\left\|x_i-x_j\right\|^2\right),\gamma>0 \tag{3}$$

一方面，我们可以使用细胞类型特异性位点所构成的甲基化谱矩阵 Y' 来预测疾病状态 Z。另一方面，我们可以使用细胞类型比例矩阵 H 来预测疾病状态 Z。由于没有独立的测试集，本文采用十折交叉验证来评价模型的性能，具体来说为：将肿瘤组织样本平均分成十份，每次选取 9 份作为训练集，1 份做测试集，将十次预测结果的均值作为最终的结果，本文选取准确率(准确率 $=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$)来度量模型的性能，高的准确率说明模型有更好的性能。

3. 肺腺癌(LUAD)与肾透明细胞癌(KIRC)样本的实证分析

本文以 TCGA 数据库中 LUAD 和 KIRC 的 Illumina Infinium 450 K 微阵列数据作为研究对象，其中 LUAD 有 466 个肿瘤样本和 32 个正常样本，KIRC 有 325 个肿瘤样本和 160 个正常样本，参考面板数据为 Reinius 等[11]提供的六种细胞类型(CD8T、CD4T、CD56NK、B 细胞、单核细胞(Mono)和粒细胞(Gran))的 450 K 甲基化数据。

首先，我们对 LUAD 和 KIRC 的甲基化数据进行预处理，删除在所有样本上取值全是缺失值的特征，并利用 Combat [12]去除样本之间的批次效应。由于 RefFreeEWAS 和 TOAST 算法需要预先进行特征选择，因此我们对于 RefFreeEWAS 算法选取变异系数(CV)最大的前 1000 个特征，对于 TOAST 选取方差最大的前 1000 个特征。

其次，我们利用 Tsisal 来估计 LUAD 和 KIRC 所包含细胞类型的个数，其估计值均为 9。在细胞类型个数为 9 的假定下分别利用 TOAST 包[13]中的 Tsisal 函数、csDeconv 函数，RefFreeEWAS 包[14]中的 RefFreeCellMix 函数对以上两种癌症类型进行反卷积分解估计细胞类型组成比例矩阵。由于 Tsisal 在有参考数据的前提下可为匿名估计的细胞类型分别类别标签，因此我们在 0.55 的阈值下，为估计得到的匿名细胞类型分配了类别标签。图 1(a)显示了 LUAD 利用 Tsisal 估计出的 9 种细胞类型的比例，其中 Mono 细胞在混合样本中比例最高，占比接近 40%，CD8T 占 18%，Gran 占 7%，该结果与 Lin [13]的结果相一致。图 1(b)显示了 KIRC 利用 Tsisal 估计出的 9 种细胞类型比例，可以看到 KIRC 主要包含 CD8T、CD56NK 这两种免疫细胞，其中 CD56NK 占比接近 30%，这些免疫细胞也被确定为肿瘤中相对更丰富的免疫细胞类型[15]。

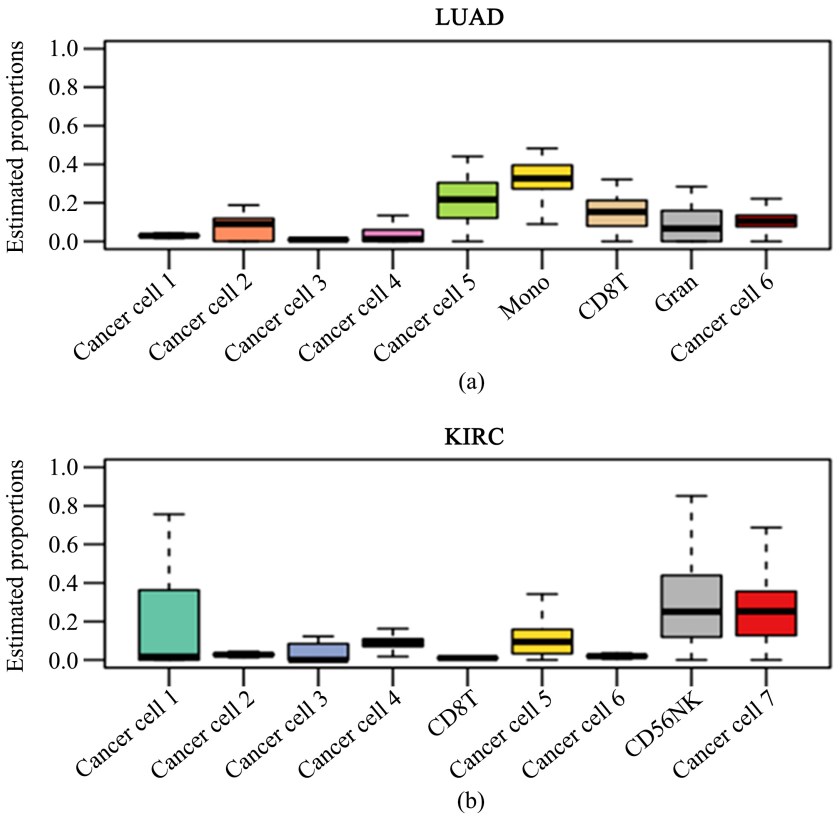

Figure 1. Boxplot of cell type proportions estimated by Tsisal of LUAD and KIRC
图 1. LUAD 和 KIRC 利用 Tsisal 估计的细胞类型组成比例的箱线图

由于肿瘤组织样本中真实的细胞类型组成比例是未知的，因此我们将 Tsisal 估计的细胞类型比例与 TOAST 和 RefFreeEWAS 估计的结果进行一致性分析来说明 Tsisal 估计的准确性。图 2(a)显示了 LUAD 上 Tsisal 估计的 Cancer cell 4 所占比例和 TOAST 估计的 Cancer cell 4 所占比例的散点图，其相关系数高达 0.8，9 种细胞类型上两种方法估计的平均相关系数大于 0.5。图 2(b)显示了 KIRC 上 Tsisal 估计的 Cancer

cell 3 所占比例和 TOAST 估计的 Cancer cell 3 所占比例的散点图，其相关系数高达 0.7，9 种细胞类型上两种方法估计的平均相关系数大于 0.4。这些结果说明了 Tsisal 估计的细胞类型所占比例和 TOAST 的估计结果具有高度的一致性。进一步地，我们将所有免疫细胞所占比例求和作为正常细胞所占比例，将所有癌症细胞所占比例求和作为癌症细胞所占比例，由此计算出每一个肿瘤样本的纯度数据，将该结果与经典的肿瘤纯度估计算法 InfiniumPurify 进行一致性分析。图 2(c)，图 2(d)分别显示了 LUAD、KIRC 上两种算法估计的肿瘤纯度散点图，它们的相关系数分别为 0.75 和 0.64。此结果进一步说明了 Tsisal 估计的细胞类型比例的可靠性。

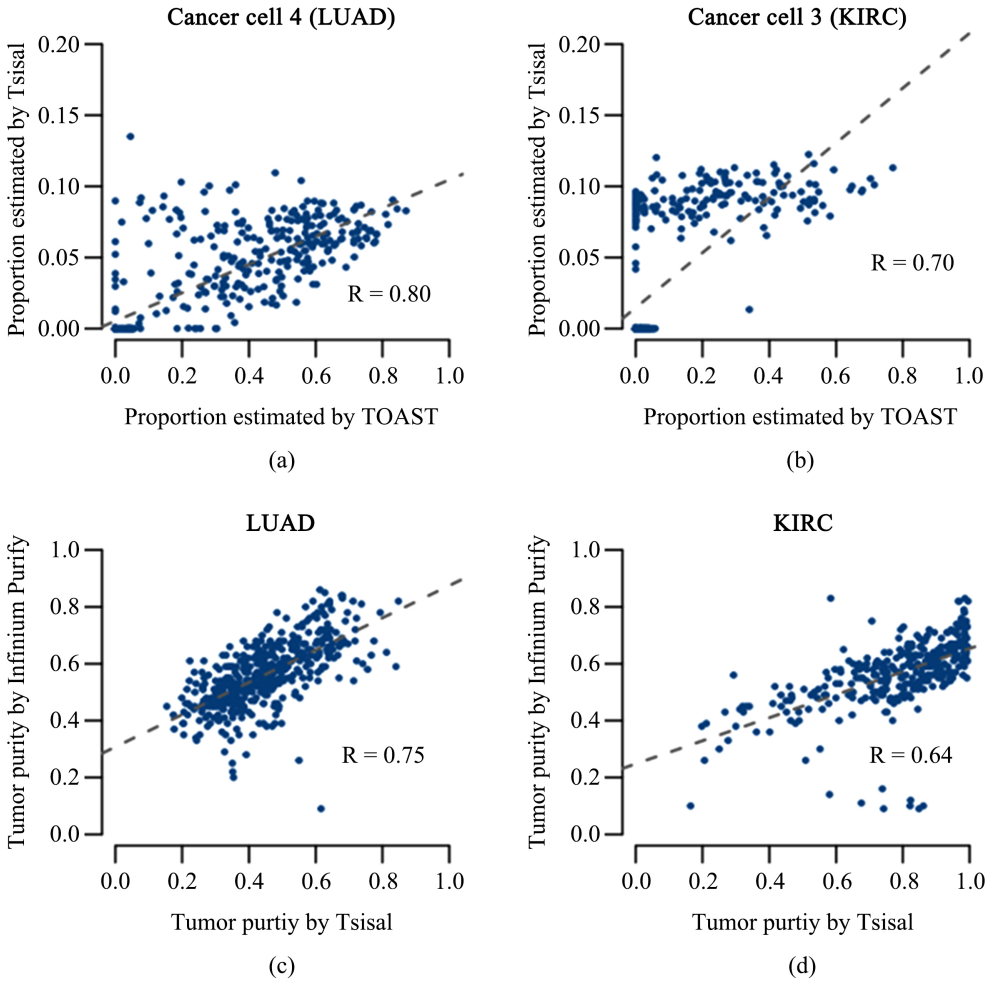

Figure 2. (a) Scatter plot of the estimated cancer cell 4 proportions between Tsisal and TOAST from LUAD; (b) Scatter plot of the estimated cancer cell 3 proportions between Tsisal and TOAST from KIRC; (c) Scatter plot of the estimated tumor purities between Tsisal and TOAST from LUAD; (d) Scatter plot of the estimated tumor purity between Tsisal and TOAST from LUAD

图 2. (a) LUAD 上 Tsisal 估计的 Cancer cell 4 所占比例和 TOAST 估计结果的散点图；(b) KIRC 上 Tsisal 估计的 Cancer cell 3 所占比例和 TOAST 估计结果的散点图；(c) LUAD 上 InfiniumPurify 估计的肿瘤纯度与 Tsisal 估计的肿瘤纯度的散点图；(d) KIRC 上 InfiniumPurify 估计的肿瘤纯度与 Tsisal 估计的肿瘤纯度的散点图

由于 Tsisal 可输出每一个细胞类型特异性的 CpG 位点，因此对于每一个细胞类型我们选取特异性最大的前 3 个 CpG 位点来判断这些位点是否是细胞类型特异性的。图 3(a)显示了 LUAD 中 Cancer cell 1 在

cg00867626 位点高甲基化，Cancer cell 3 在 cg16545496 位点也出现高甲基化。图 3(b)显示 KIRC 中 Cancer cell 2 在 cg11474317 位点表现出高甲基化，CD8T 在 cg13755570 位点也表现出高甲基化。这些位点具有更为重要的生物学意义，也是我们进行疾病诊断预测的重要生物学标志物。

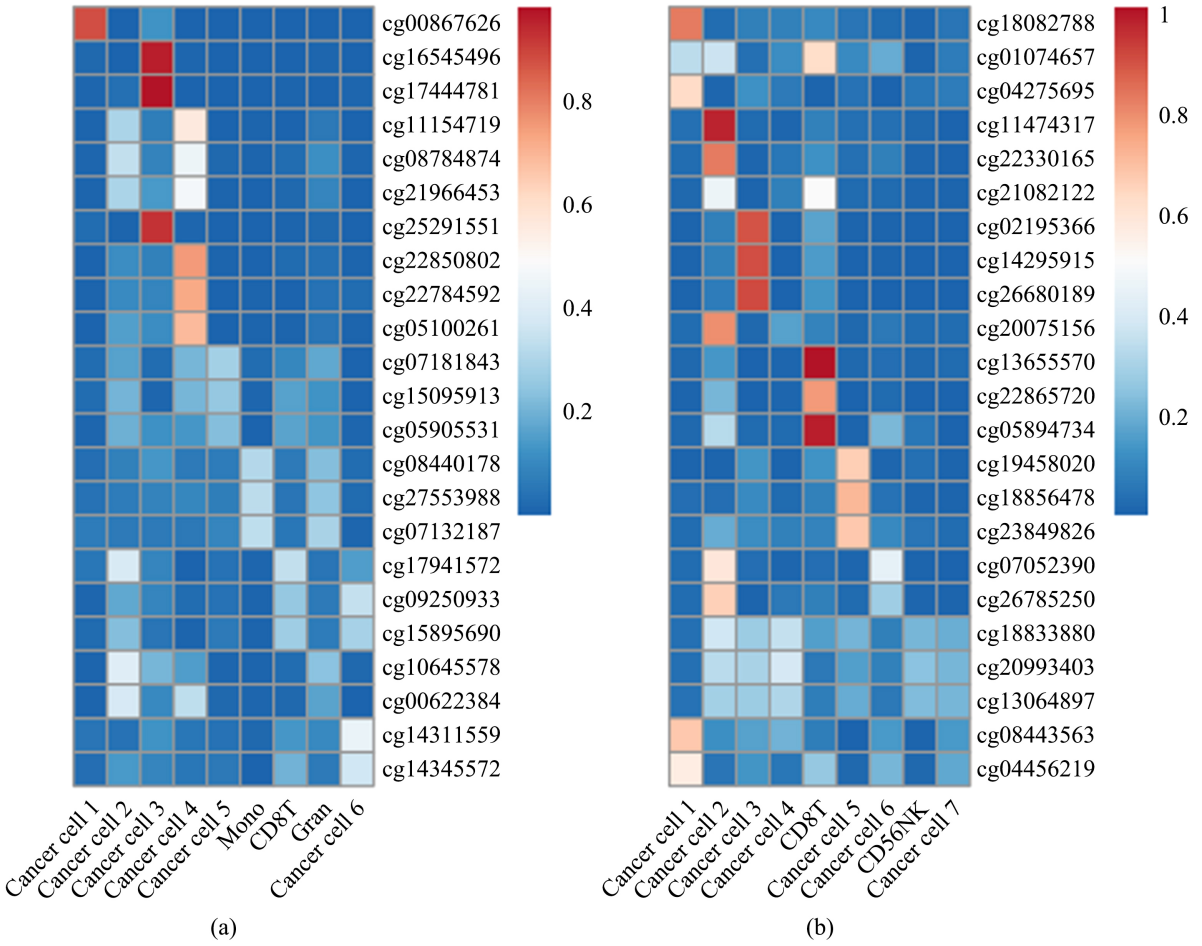

Figure 3. Heatmap of methylation profile data of cell type specific sites identified by Tisal on LUAD and KIRC
图 3. LUAD 和 KIRC 上 Tisal 识别出的细胞类型特异性位点甲基化谱数据的热图

以上分析结果证明了 Tsisal 所估计的细胞类型比例矩阵准确，其识别的位点也是细胞类型特异性的，因此最后我们将 Tsisal 估计得到的肿瘤组织中各细胞类型所占比例和细胞类型特异性的甲基化位点作为生物标志物，利用 SVM 算法进行疾病预测分析。具体做法为：将每一个细胞类型在样本上所占比例矩阵(或细胞类型特异性的甲基化位点在所有样本上的取值矩阵)作为自变量，将每一个样本的疾病状态作为因变量，利用 R 包 el071 中的 svm 函数构建模型，利用 predict 函数进行预测分析。图 4(a)显示了 LUAD 上利用 CV、TOAST 和 Tsisal 选取细胞类型特异性的 CpG 位点(marker)作为 SVM 的输入来预测疾病状态时的准确率，CV 和 TOAST 的准确率分别为 0.92 和 0.96，而 Tsisal 的准确率高达 0.99；当选择细胞类型估计比例(proportion)作为 SVM 的输入来预测疾病状态时，RefFreeEWAS 的准确率为 0.94，TOAST 和 Tsisal 的准确率分别为 0.98 和 0.99。尽管 Tsisal 与 TOAST 准确率较为接近，但 Tsisal 的运行时间是 TOAST 的六分之一。图 4(b)癌症类型 KIRC 上取得了与图 4(a)相类似的结果。这些结果表明利用反卷积算法 Tsisal 和 SVM 算法结合来构建疾病的预测模型具有较高的准确性、鲁棒性和运行效率。

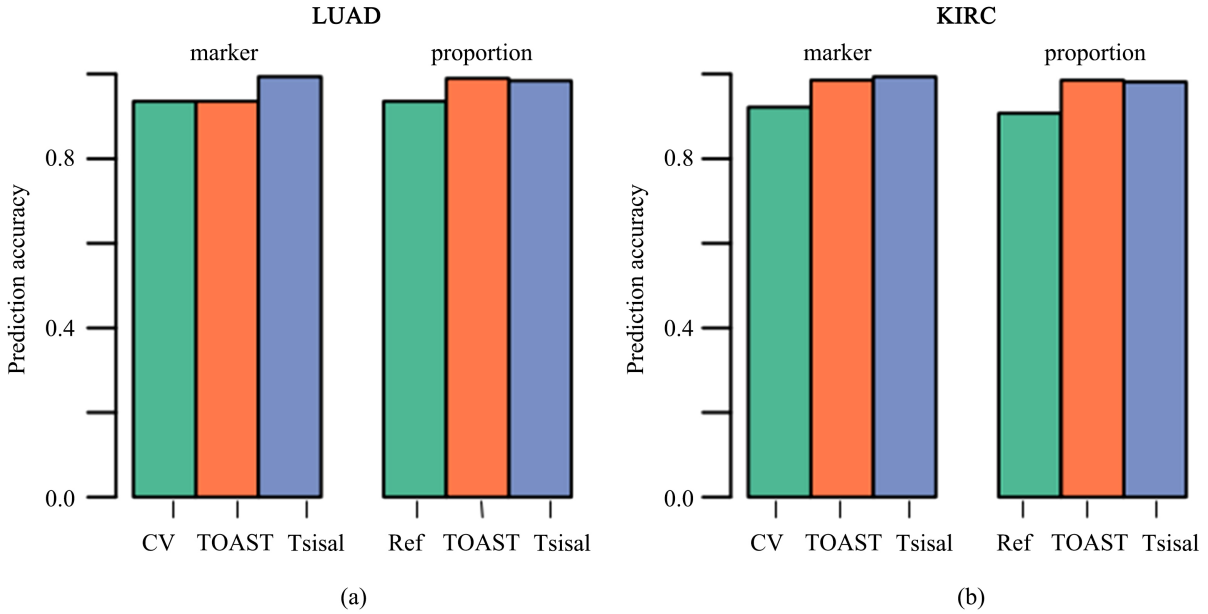

Figure 4. Accuracy of disease prediction under different biomarker selection methods on LUAD and KIRC
图 4. LUAD 和 KIRC 上不同生物标志物选取方法下疾病预测的准确率

4. 结论

本文针对临床肿瘤组织的甲基化数据构建了疾病的预测模型，该模型在正态分布假设下利用不依赖参考基矩阵的反卷积算法 Tsisal 对肿瘤组织进行分解，估计得到肿瘤组织中各细胞类型所占比例和细胞类型特异性的甲基化位点，将识别出的细胞类型组成比例和细胞类型特异性的甲基化位点作为生物标志物，利用机器学习中的 SVM 算法进行疾病预测，并利用十折交叉验证来评价算法的性能。TCGA 数据库中肺腺癌、肾透明细胞癌甲基化数据分析表明，所提模型取得了优于 TOAST 和 RefFreeEWAS 的分析结果。

基金项目

研究得到了国家自然科学基金(61902061)、江西省自然科学基金(20212BAB202001)和江西省科技厅项目(GJJ200704)的支持。

参考文献

[1] Hanahan, D. and Weinberg, R.A. (2011) Hallmarks of Cancer: The Next Generation. *Cell*, **144**, 646-674. https://doi.org/10.1016/j.cell.2011.02.013

[2] Li, Z. and Wu, H. (2019) TOAST: Improving Reference-Free Cell Composition Estimation by Cross-Cell Type Differential Analysis. *Genome Biology*, **20**, 190. https://doi.org/10.1186/s13059-019-1778-0

[3] Joyce, J.A. and Pollard, J.W. (2009) Microenvironmental Regulation of Metastasis. *Nature Reviews Cancer*, **9**, 239-252. https://doi.org/10.1038/nrc2618

[4] Anders, S. and Huber, W. (2010) Differential Expression Analysis for Sequence Count Data. *Genome Biology*, **11**, 1-12. https://doi.org/10.1186/gb-2010-11-10-r106

[5] Onuchic, V., *et al*. (2016) Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types. *Cell Reports*, **17**, 2075-2086. https://doi.org/10.1016/j.celrep.2016.10.057

[6] Zhang, W., Wu, H. and Li, Z. (2021) Complete Deconvolution of DNA Methylation Signals from Complex Tissues: A Geometric Approach. *Bioinformatics*, **37**, 1052-1059. https://doi.org/10.1093/bioinformatics/btaa930

[7] Houseman, E.A., *et al*. (2012) DNA Methylation Arrays as Surrogate Measures of Cell Mixture Distribution. *BMC*

Bioinformatics, **13**, Article No. 86. https://doi.org/10.1186/1471-2105-13-86

[8] Liu, Y., *et al*. (2013) Epigenome-Wide Association Data Implicate DNA Methylation as an Intermediary of Genetic Risk in Rheumatoid Arthritis. *Nature Biotechnology*, **31**, 142-147. https://doi.org/10.1038/nbt.2487

[9] Houseman, E.A., *et al*. (2014) Reference-Free Cell Mixture Adjustments in Analysis of DNA Methylation Data. *Bioinformatics*, **30**, 1431-1439. https://doi.org/10.1093/bioinformatics/btu029

[10] Rahmani, E., *et al*. (2016) Sparse PCA Corrects for Cell Type Heterogeneity in Epigenome-Wide Association Studies. *Nature Methods*, **13**, 443-445. https://doi.org/10.1038/nmeth.3809

[11] Reinius, L.E., Acevedo, N., Joerink, M., *et al*. (2012) Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility. *PLOS ONE*, **7**, e41361. https://doi.org/10.1371/journal.pone.0041361

[12] Johnson, W.E., *et al*. (2007) Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods. *Biostatistics*, **8**, 118-127. https://doi.org/10.1093/biostatistics/kxj037

[13] Lin, Y., Chen, D., Ding, Q., Zhu, X., Zhu, R. and Chen, Y. (2021) Progress in Single-Cell RNA Sequencing of Lung Adenocarcinoma. *Chinese Journal of Lung Cancer*, **24**, 434-440.

[14] Houseman, E.A., Kile, M.L., Christiani, D.C., Ince, T.A., Kelsey, K.T. and Marsit, C.J. (2016) Reference-Free Deconvolution of DNA Methylation Data and Mediation by Cell Composition Effects. *BMC Bioinformatics*, **17**, Article No. 259. https://doi.org/10.1186/s12859-016-1140-4

[15] Whiteside, T.L. (2008) The Tumor Microenvironment and Its Role in Promoting Tumor Growth. *Oncogene*, **27**, 5904-5912. https://doi.org/10.1038/onc.2008.271