Advances in Applied Mathematics 应用数学进展, 2023, 12(9), 4105-4111
Published Online September 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.129402

Hans汉斯

基于图卷积神经网络的胃癌和结直肠癌的生存预测

路晓雯

绍兴文理学院数理信息学院，浙江 绍兴

收稿日期：2023年8月21日；录用日期：2023年9月15日；发布日期：2023年9月21日

摘　要

胃癌和结直肠癌是我国恶性肿瘤中比较常见的两类，同时也是难以治愈的两种癌症。医学界为了统计癌症病人的生存率，提出了5年生存率作为一个有效指标，以此统计癌症病人的存活率。本文将胃癌和结直肠癌的全组织病理学图像(WSI)进行了切片，将切片后的图像进行特征提取后，进行了患者层面的图的构造；将构造好的图形放入构造好的4层图卷积神经网络(GCN)中进行训练，结合每个患者的总生存时间和生存状态，得到了胃癌和结直肠癌的C-index值分别为0.58和0.65，二者结果均高于之前提出的卷积神经网络模型。

关键词

胃癌，结直肠癌，图卷积神经网络，生存概率

Survival Prediction of Gastric Cancer and Colorectal Cancer Based on Graph Convolutional Neural Networks

Xiaowen Lu

School of Mathematical Information, Shaoxing University, Shaoxing Zhejiang

Received: Aug. 21st, 2023; accepted: Sep. 15th, 2023; published: Sep. 21st, 2023

Abstract

Gastric cancer and colorectal cancer are two common types of malignant tumors in China, and they are also two cancers that are difficult to cure. In order to calculate the survival rate of cancer pa-

文章引用: 路晓雯. 基于图卷积神经网络的胃癌和结直肠癌的生存预测[J]. 应用数学进展, 2023, 12(9): 4105-4111.
DOI: 10.12677/aam.2023.129402

tients, the medical community proposed the 5-year survival rate as an effective indicator to calculate the survival rate of cancer patients. In this paper, the whole histopathological images (WSI) of gastric cancer and colorectal cancer were sectioned, and the patient-level map was constructed after feature extraction of the sliced images. The constructed graph was put into the constructed 4-layer graph convolutional neural network (GCN) for training, and combined with the total survival time and survival state of each patient, the C-index values of gastric cancer and colorectal cancer were obtained to be 0.58 and 0.65, respectively, which were higher than the previously proposed convolutional neural network model.

Keywords

Gastric Cancer, Colorectal Cancer, Graph Convolutional Neural Network, Survival Probability

Copyright © 2023 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

癌症的致死率非常高，是世界范围内的主要死因之一。而癌症预后可以帮助医生对患者进行诊断，针对不同癌症患者制定相应的治疗方案。近年来，利用深度学习进行癌症预后取得了较大进展。癌症预后是一项极具挑战性的工作，不仅要考虑肿瘤组织的特征，还需要考虑肿瘤组织周围的特征，以评估患者的死亡风险[1] [2]。

近年来，很多深度学习的方法被用来用于全视野数字切片(WSI)的生存分析。比如，在神经网络的最后一层使用 COX 比例风险函数作为输出，进行预测[3]；或者使用 K-Means 方法，在 WSI 水平上进行聚类，作为卷积神经网络的输入进行预测[4]。

本论文使用了图卷积神经网络进行主要模型，用于预测胃癌和结直肠癌患者的生存预测，将 WSI 切片抽象为图，作为模型的输入。与其他深度学习的方法相比，图卷积神经网络不仅学习 WSI 中每个切片的特征，还将此切片及相邻切片的特征进行组合学习，具有较好的周围环境感知能力[5]。

2. 数据选择与预处理

2.1. 数据选择

本实验用到的数据集为胃癌(STAD)和结直肠癌(COAD)的 WSI 数据集。以上数据均来自 https://portal.gdc.cancer.gov/网站。

WSI 的全称为 Whole Slide Image，名称为全视野数字切片。这种图像的像素很大，通常以万为级别。本实验所用到的数据是胃癌和结直肠癌的 WSI 图片。由于本文研究的问题是胃癌和结直肠癌的 1 年期生存率，所以用到了癌症病人的 WSI 图片以及生存信息。每个癌症病人可能具有多个 WSI 图片，而每个癌症病人具有各自的生存信息，信息包括：样本生存时间(OS.time)，是指从癌症病人确诊到最后一次随访的时间；生存状态(OS)，这里的 OS 取 1 或 0，1 代表该病人为死亡状态，0 代表病人为存活状态；生存时间(survival_months)指的是病人从确诊到最后一次随访的时间，以月为单位。根据 OS.time 和 OS 规定该病人的最终标签(label)，判断标准见图 1 所示：

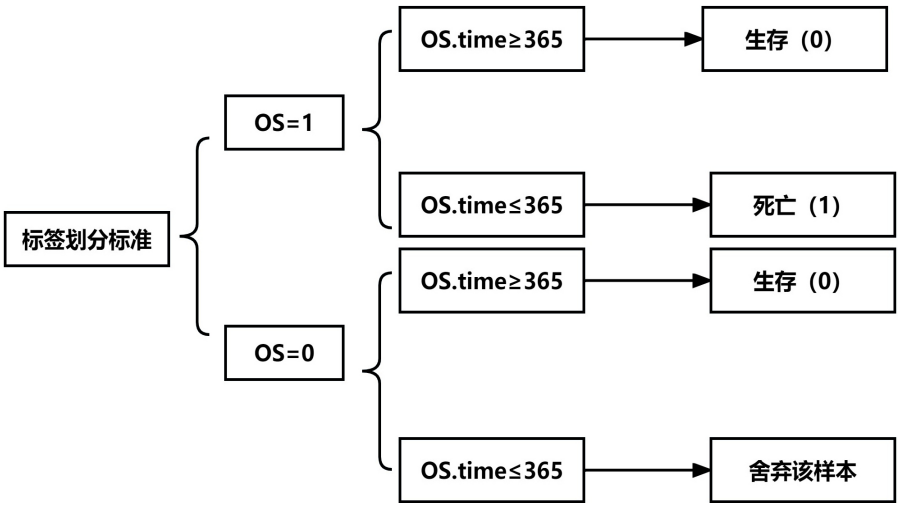

Figure 1. Criteria for dividing patient survival
图 1. 病人生存划分标准

2.2. 数据预处理

对于胃癌和结直肠癌的 WSI 图片，如图 2 所示，我们采用以下流程进行预处理：

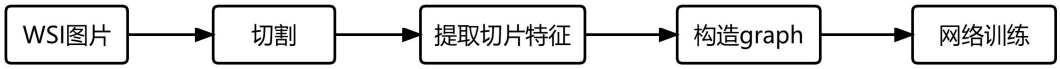

Figure 2. Preprocessing flowchart
图 2. 预处理流程图

2.2.1. WSI 图像切割及特征提取

WSI **(**whole slide images**)**，名称为全视野数字切片，一般是通过对癌肿病理图像进行扫描得到。WSI 图像的像素非常大，通常以万为单位，一张 WSI 图像通常会占用几十 MB 到几 G 不等。一张 WSI 图像包含了很多病理学信息。如图 3 所示，这是一张结直肠癌的 WSI 图像。

Figure 3. WSI image of colorectal cancer
图 3. 结直肠癌 WSI 图像

对于像素巨大的 WSI 图像，我们采用 OTSU 算法进行切割。OTSU 算法主要是通过将上述类型的 WSI 图像进行灰度级划分，再采取划定阈值的方法将图像中病例部分分割出来，以实现图像的分割，分割之后的切片用来进行特征提取。

将 WSI 进行分割之后，我们使用在 ImageNet 数据集上预训练好的 Resnet50 网络对 WSI 图像切片提取出了 1024 维的特征。

2.2.2. Graph 的构造

我们在患者层面上构造 WSI 的 Graph。对于一张 WSI 的若干切片，将这些切片定义为在 WSI 二维坐标上的点，这些点构成了 WSI 图的节点；Resnet 网络提取的 1024 维特征作为每个节点的特征；将每个切片与其在二维坐标平面上周围 8 个点进行连接，作为 WSI 图的边。

如下图所示，以红色节点为例，将周围的 8 个节点进行连接，作为整张 WSI 图片的边。一个患者可能具有多张 WSI 图片，将患者的多个 WSI 图片分别进行如下图的构造，之后再将这几张 WSI 图片构成的图作为子图，形成该患者的 Graph。图 4 为 WSI 图片节点和边的构造示例。

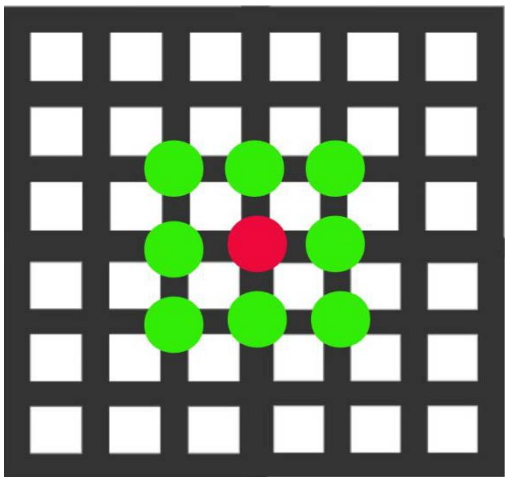

Figure 4. Illustration of the WSI diagram
图 4. WSI 图示意

2.2.3. 网络训练

我们采用 Patch-GCN [5]网络进行训练。由于本文构造的患者的每张 WSI 图片上的每个节点与其周围的 8 个节点相连，连边密度比较大，网络每增加一层，每个节点之间的邻居会增加很多，从而加重训练负担。另外，由于 GCN 的传播机制，每个节点在网络层数增加的过程中，聚合道德信息也会重复性增加，从而导致训练资源浪费。根据以上考虑，我们将图卷积神经网络的层数设置为 4 层，并在网络的最后一层使用 Cox 函数进行回归，进行生存分析。

3. 实验结果

我们使用了胃癌和结直肠癌的数据进行 Graph 的构造和图卷积神经网络的训练。使用数据为胃癌和结直肠癌的 WSI 图像数据以及两种癌症患者 5 年期生存数据，将患者层面的数据按照训练集和测试集 4:1 的比例进行 5 折交叉验证，训练的到的 5 个癌肿 C-index 值进行平均，作为该癌肿最后的 C-index 值，并绘制了两个癌肿的 AUC 曲线，如图 1 所示。

另外，为了进一步说明本文使用的图卷积神经网络(GCN)的性能，我们将原始数据(WSI)同样进行卷积神经网络的训练，模型为 Ruoyu Li 等人提出的 graph CNN [6]网络。如表 1 所示，最终得出使用图卷积神经网络进行预测的胃癌和结直肠癌的 C-index 值分别为 0.65 和 0.58。实验结果证明本文使用的图卷积神经网络(GCN)在预测胃癌和结直肠癌的生存概率的性能上的确优于卷积神经网络。

Table 1. County level planning schedule
表 1. 县域等级规划一览表

	GCN	CNN
胃癌(STAD)	0.65	0.63
结直肠癌(COAD)	0.58	0.54

此外，我们将两个癌肿的预测结果绘制了 AUC 曲线，如图 5 和图 6 所示。

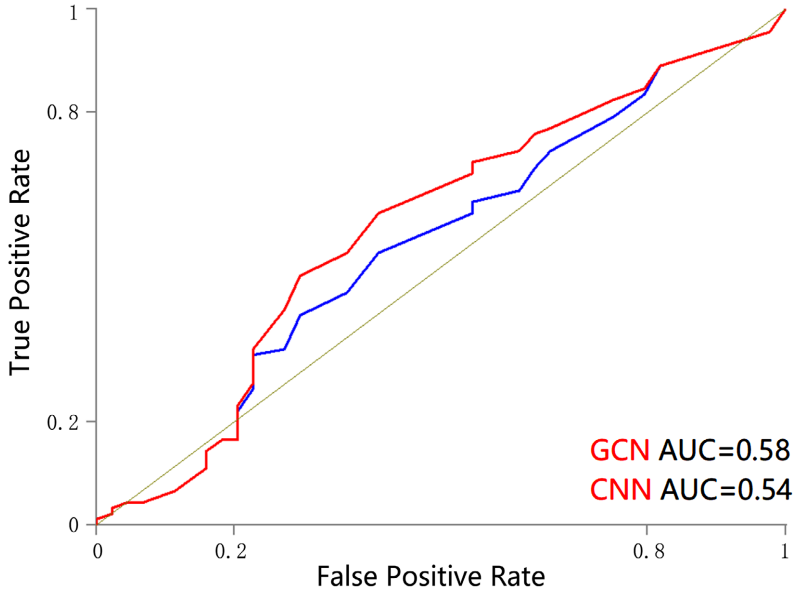

Figure 5. AUC curve for predicting survival probability of colorectal cancer (COAD)
图 5. 结直肠癌(COAD)生存概率预测 AUC 曲线

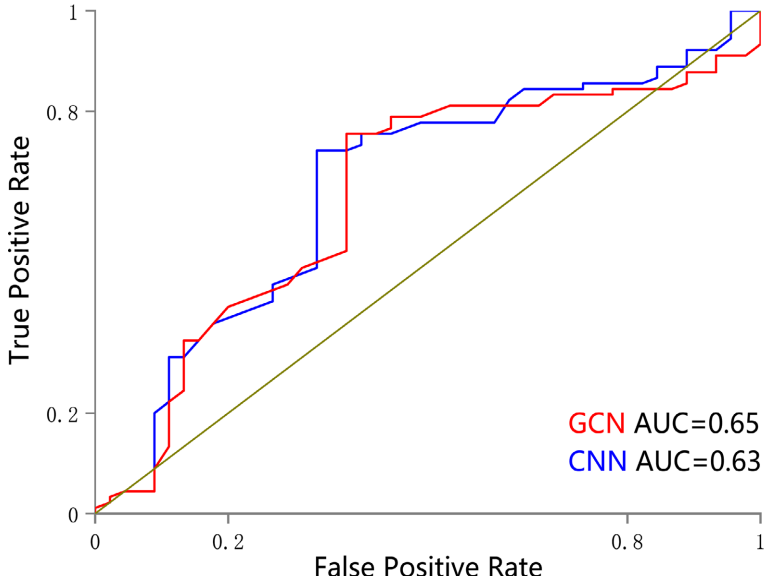

Figure 6. AUC curve for predicting survival probability of gastric cancer (STAD)
图 6. 胃癌(STAD)生存概率预测 AUC 曲线

4. 方法原理

4.1. 图卷积神经网络

在图卷积神经网络(GCN)提出之前，深度学习一直都是以卷积神经网络(CNN)为主。CNN 的核心在于卷积核通过在输入图片上平移进行类似点乘的操作来进行图片的特征提取。而 GCN 的核心则在于通过图上的节点和边的连接来进行节点之间的特征融合。GCN 在一定程度上达到了各个节点特征之间的特征融合。

GCN 模型的输入是图数据，假设每张图片具有 N 个节点和 M 条边，每个节点都有 D 维特征，那么每个节点都会有 D × D 维的特征矩阵 X，N 个节点构成的邻接矩阵 A 和特征矩阵 X 就是 GCN 的输入。本文使用的 GCN 网络共有 4 层，每一层都相当于一轮特征学习过程。

4.2. C-index

C-index 全称为一致性指数(Concordance Index) [7]，最早于 1996 年在范德堡大学的 Frank E Harrell Jr. 教授提出，通常被用来评估生存模型的预测结果。C-index 的计算方法如下：将 n 个病人两两结对，数量为 C(n, 2)，将预测生存概率高低与病人实际生存状态相一致的对数除以 C(n, 2)，得到的比例就是一致性指数，它实质上计算出了预测结果与实际状态相一致的概率。根据以上计算方法，一致性指数的值应该在 0.5 至 1 之间。

5. 结论

本文对胃癌(STAD)和结直肠癌(COAD)的全视野组织图像(WSI)进行了切割、特征提取等预处理操作；预处理结束后，对每一张 WSI 进行了图的节点和边的构造；最后采用图卷积神经网络，对两个癌肿(胃癌和结直肠癌)的五年生存信息进行了生存分析，最终得到胃癌的 C-index 值为 0.65，结直肠癌的 C-index 值为 0.58，较卷积神经网络的结果有所提升。

6. 讨论

在以往关于 WSI 图像的生存分析中，大多数都是使用卷积神经网络(CNN)进行训练，从而得到结果。比较常见的有多实例学习(MIL)方法、弱监督方法，这次方法虽然能解决 WSI 上很多分类、回归任务，但并未达到 WSI“全局”学习。也就是说，之前的方法并未将 WSI 中不同切块之间的特征进行整合学习；而图卷积神经网络(GCN)却很好地改进了这一点，其利用 WSI 层面的图结构，将每个切片看作图中的节点，通过节点之间的边的连接，完成了节点之间特征学习，从而具有全局性，在一定程度上也可以看作“多尺度”的学习。

参考文献

[1] Balkwill, F.R., Capasso, M. and Hagemann, T. (2012) The Tumor Microenvironment at a Glance. *Journal of Cell Science*, **125**, 5591-5596. https://doi.org/10.1242/jcs.116392

[2] Saltz, J., Gupta, R., Hou, L., Kurc, T., Singh, P., Nguyen, V., Samaras, D., Shroyer, K.R., Zhao, T., Batiste, R., *et al.* (2018) Spatial Organization and Molecular Correlation of Tumor-Infifiltrating Lymphocytes Using Deep Learning on Pathology Images. *Cell Reports*, **23**, 181-193.E7. https://doi.org/10.1016/j.celrep.2018.03.086

[3] Lu, C., Romo-Bucheli, D., Wang, X., Janowczyk, A., Ganesan, S., Gilmore, H., Rimm, D. and Madabhushi, A. (2018) Nuclear Shape and Orientation Features from H&E Images Predict Survival in Early-Stage Estrogen Receptor-Positive Breast Cancers. *Laboratory Investigation*, **98**, 1438-1448. https://doi.org/10.1038/s41374-018-0095-7

[4] Zhu, X.L., Yao, J.W., Zhu, F.Y. and Huang, J.Z. (2017) WSISA: Making Survival Prediction from Whole Slide Histopathological Images. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26

July 2017, 6855-6863. https://doi.org/10.1109/CVPR.2017.725

[5] Chen, R.J., Lu, M.Y., Shaban, M., *et al*. (2021) Whole Slide Images Are 2D Point Clouds: Context-Aware Survival Prediction Using Patch-Based Graph Convolutional Networks. In: de Bruijne, M., *et al*., Eds., *Medical Image Computing and Computer Assisted Intervention—MICCAI* 2021, Springer, Cham, 339-349. https://doi.org/10.1007/978-3-030-87237-3_33

[6] Li, R.Y., Yao, J.W., Zhu, X.L., Li, Y.Q. and Huang, J.Z. (2018) Graph CNN for Survival Analysis on Whole Slide Pathological Images. In: Frangi, A., Schnabel, J., Davatzikos, C., Alberola-López, C. and Fichtinger, G., Eds., *Medical Image Computing and Computer Assisted Intervention—MICCAI* 2018, Springer, Cham, 174-182. https://doi.org/10.1007/978-3-030-00934-2_20

[7] Harrell Jr., F.E., Lee, K.L. and Mark, D.B. (1996) Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine*, **15**, 361-387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4