Advances in Applied Mathematics 应用数学进展, 2020, 9(4), 478-484
Published Online April 2020 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2020.94058

Hans汉斯

Diagnosing the State of Individuals by Relative Entropy Score

Junxia Wang, Rui Liu

School of Mathematics, South China University of Technology, Guangzhou Guangdong
Email: 2336334759@qq.com, scliurui@scut.edu.cn

Received: Mar. 19th, 2020; accepted: Apr. 2nd, 2020; published: Apr. 9th, 2020

Abstract

The development of complex disease can be divided as three states as follows: a normal state, a pre-disease state, and a disease state. Furthermore, there are a lot of significant differences between the network of the normal state and the disease state. If we can obtain the features of the network in normal and the disease state, we achieve the purpose of disease early warning. In this study, we proposed an algorithm to learn the features of different states, and thus distinguish different states. We certificate the effectiveness of the method by applying this method to the data set Lung squamous cell carcinoma.

Keywords

Relative Entropy Score (RES), Network Feature, Disease Diagnosis

通过相对熵得分诊断个体状态

王俊霞，刘　锐

华南理工大学数学学院，广东 广州
Email: 2336334759@qq.com, scliurui@scut.edu.cn

收稿日期：2020年3月19日；录用日期：2020年4月2日；发布日期：2020年4月9日

摘　要

复杂疾病的发展需要经历三种状态：正常状态、前疾病状态和疾病状态。此外，生物网络的正常状态与疾病状态之间存在着很大的差异。如果能获得网络在不同状态下的特征，便可以达到疾病预警的目的。在本研究中，我们提出一个算法学习网络在不同状态下的特征，从而判断个体状态。我们将该方法应用到真实数据集肺鳞状细胞癌，从而验证了该方法的有效性。

文章引用: 王俊霞, 刘锐. 通过相对熵得分诊断个体状态[J]. 应用数学进展, 2020, 9(4): 478-484.
DOI: 10.12677/aam.2020.94058

关键词

相对熵得分(RES)，网络特征，疾病诊断

1. 引言

随着经济的发展，人们对生活的方方面面提出了越来越多的要求，包括生活质量、身体健康等。近年来，生物医学越来越受到人们的重视，随之而来的是相关理论研究的蓬勃发展。相关研究表明，病情恶化不一定是平稳的，而是突然的[1]。这意味着在病情转变之前，存在一个临界状态。一般来说，从健康状态发展到疾病状态要经历以下三个阶段：正常状态、疾病前状态(或临界状态)和疾病状态。正常状态是一个稳定状态，代表一个相对健康的阶段。疾病前状态通常定义为在达到临界点之前的正常状态的极限[1]。众所周知，用网络或边生物标志物来刻画疾病是一种比较科学而且令人信服的方法，但是要获得个体分子之间的边或相关性是很困难的。因此边生物标志物应运而生[2]。

很多复杂疾病的治愈率低或治疗后生活质量差，一部分是由于疾病本身较难治愈，而更多的是由于发现的不及时从而错过最佳治疗时间而变得难以治愈。因此，如果能提前捕捉到疾病的信号，就可以提升病人的生活质量，甚至挽救病人的生命。复杂疾病的网络在正常状态和疾病状态下具有很大的差别，这意味着不同状态的网络所具有的特征不尽相同。这为探测疾病信号提供了可能。在本文中，我们提出一个称为相对熵得分(RES)的指标。根据样本的 RES 值，确定样本所处的状态，从而给出不同的建议。如果一个人处于疾病状态，手术是一个合理的建议；反之，建议服药或注射治疗。

2. 方法

2.1. 理论基础

给定多个正常样本(m 个样本)和疾病样本(n 个样本)，带有 k 个基因的基因表达数据可以表示为图 1(A)所示的数据矩阵表达，其中有 k 个基因，对照组和疾病组的样本大小分别为m和n。设 $x_i \in R^m$ 和 $y_i \in R^n$ 分别表示对照组和疾病组第 i 个基因的表达向量，即 x_{ij} 为对照组第 i 个基因的第 j 个样本表达，y_{ij} 为疾病组第 i 个基因的第 j 个基因表达。

每个基因对的特征 RES_N 和 RES_D 如图 1(B)所示进行构造。基因对 u 和 v 左侧正常状态的行向量与基因对 v 和 u 左侧正常状态的行向量之和的一半即为正常状态下的基因对 u 和 v 的相对熵得分，即 RES_N；同理，基因对 u 和 v 右侧疾病状态的行向量和基因对 v 和 u 右侧疾病状态的行向量之和的一半是在疾病状态下基因对 u 和 v 的相对熵得分，即 RES_D。

2.2. 算法

以下为学习网络的特征的步骤：

- 选择差异表达基因

人体内众多基因中只有一部分对本研究起着关键作用，而另一些对我们的研究影响甚微。因此，我们只需要选择差异表达基因。

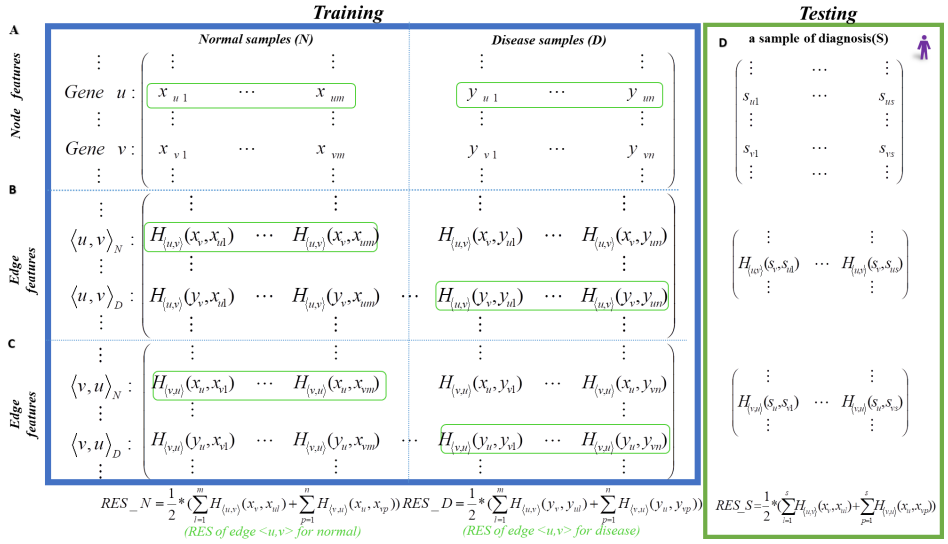

Figure 1. Data matrices for node features and edge features. (A) Gene expression data of *k* genes, where are *m* samples on the normal state and *n* samples on the disease state. (B) The data matrix of edge features of gene-pair *u* and *v*, which aims to find the RES of gene-pair *u* and *v*. Each column is one sample. One-half of the sum of the row vector on the left of gene-pair *u* and *v* and the row vector on the left of gene-pair *v* and *u* on the normal state is the RES of gene-pair *u* and *v* on the normal state, namely, RES_N. Similarly, one-half of the sum of the row vector on the right of gene-pair *u* and *v* and the row vector on the right of gene-pair *v* and *u* on the disease state is the RES of gene-pair *u* and *v* on the disease state, namely, RES_D. (C) When a sample comes to diagnose, we can utilize his past and present RES values to judge physical condition

图 1. 节点特征和边特征的数据矩阵。(A) *k* 个基因的基因表达数据，其中 *m* 个样本为正常状态，*n* 个样本为疾病状态。(B) 基因对 *u* 和 *v* 的边的特征数据矩阵，其目的是寻找基因对 *u* 和 *v* 的相对熵得分，每列为一个样本。正常状态下，基因对 *u* 和 *v* 左侧的行向量与基因对 *v* 和 *u* 左侧的行向量之和的一半即为正常状态下的基因对 *u* 和 *v* 的相对熵得分，即 RES_N；同理，疾病状态下，基因对 *u* 和 *v* 右边的行向量和基因对 *v* 和 *u* 右边疾病状态的行向量之和的一半是基因对 *u* 和 *v* 对疾病状态的相对熵得分，即 RES_D。(C) 当对一个样本进行诊断时，利用样本过去和现在的相对熵得分来判断此时的身体状态

- 选择差异相关基因对

在研究由边连接而成的网络时，通常用 PCC (皮尔逊相关系数)来表征两个基因之间的相关性。差异边定义如下：

$$\left\{\langle i,j\rangle \,|\, \left|r_{ij}^{N}-r_{ij}^{D}\right|>\delta\right\} \tag{1}$$

其中，i 和 j 分别代表研究中的基因 i 和基因 j。当 $\left|r_{ij}^{N}-r_{ij}^{D}\right|>\delta$，基因 i 和基因 j 之间存在差异边，反之则不存在。r_{ij}^{N} 和 r_{ij}^{D} 分别代表基因 i 和基因 j 在正常状态和疾病状态下的 PCC。在本研究中，肺鳞状细胞癌(LUSC)的阈值 δ 分别为 1.97。

- 计算相对熵得分(RES)

对于被挑选出来的差异边，计算相对熵得分，为下一步的研究做准备。基因对 u 和 v 在正常状态下的相关性为

$$H\left(u,v\right)=\sum_{l=1}^{m}H_{\langle u,v\rangle}\left(x_v,x_{ul}\right) \tag{2}$$

其中

$$H\left(x_v,x_{ul}\right)=p\left(x_{v1}\right)*\log\left(\frac{p\left(x_{v1}\right)}{p\left(x_{vl}\right)}\right)+p\left(x_{v2}\right)*\log\left(\frac{p\left(x_{v2}\right)}{p\left(x_{vl}\right)}\right)+\cdots+p\left(x_{vm}\right)*\log\left(\frac{p\left(x_{vm}\right)}{p\left(x_{vl}\right)}\right) \tag{3}$$

因此，基因对 v 和 u 在正常状态的相关性为

$$H(v,u)=\sum_{p=1}^{n}H_{\langle v,u\rangle}\left(x_u,x_{vp}\right) \tag{4}$$

显然，基因对 u 和 v 在疾病状态下的相关性为

$$H(u,v)=\sum_{l=1}^{m}H_{\langle u,v\rangle}\left(y_v,y_{ul}\right) \tag{5}$$

因此，基因对 v 和 u 在正常状态的相关性为

$$H(v,u)=\sum_{p=1}^{n}H_{\langle v,u\rangle}\left(x_u,x_{vp}\right) \tag{6}$$

而且

$$\mathrm{RES\_N}=\frac{1}{2}*\left(\sum_{l=1}^{m}H_{\langle u,v\rangle}\left(x_v,x_{ul}\right)+\sum_{p=1}^{n}H_{\langle v,u\rangle}\left(x_u,x_{vp}\right)\right) \tag{7}$$

代表了边 $\langle u,v\rangle$ 在正常状态下的相对熵得分。

$$\mathrm{RES\_D}=\frac{1}{2}*\left(\sum_{l=1}^{m}H_{\langle u,v\rangle}\left(y_v,y_{ul}\right)+\sum_{p=1}^{n}H_{\langle v,u\rangle}\left(y_u,y_{vp}\right)\right) \tag{8}$$

代表了边 $\langle u,v\rangle$ 在疾病状态下的相对熵得分。

2.3. 真实数据集的数据存取与处理

肺鳞状细胞癌(LUSC)数据集来自 TCGA 数据库(http://cancergenome.nih.gov)。LUSC 数据集有 178 个疾病样本。在临床上，肺鳞状细胞癌分为 7 期(IA、IB、IIA、IIB、IIIA、IIIB、IV)，即有 6 种划分方式。例如，分为对照期(包括临床 IA、IB、IIA 期)和疾病期(包括临床 IIB、IIIA、IIIB、IV 期)。我们将该算法应用于数据集。

首先，挑选具有差异表达的基因(LUSC 挑选了 70 个基因)。这些基因之间通过边(即相关性)进行连接。

接下来，选择有显著相关性的基因对，得到差异网络，其性能在不同的状态下有显著的差异。

紧接着，通过上述算法计算出各自状态下的相对熵得分。

最后，观察网络在不同状态下呈现出的差异，研究其各自的特点。

3. 结果

肺鳞状细胞癌是一种非小细胞肺癌，起源于肺气道中的鳞状细胞，因为薄而扁平的细胞在显微镜下看起来像鱼鳞，所以被称为肺鳞状细胞癌[3] [4]。

对于 LUSC 数据集，临床上分为七个阶段(IA, IB, IIA, IIB, IIIA, IIIB, IV)，共 178 个样本。所以，对于控制组和疾病组的样本数据，一共有六种划分方式。在每种划分方式下，我们应用上述算法得到结果如图 2 所示。显然，对于相对熵得分的方法来说，第三种划分方式下得到的效果最好($p=0.042$)，而基因表达的方法在最后一种划分方式下的效果最好，$p=0.021$。

显然，基因表达的方法的结果与相对熵得分的方法的结果存在很大的差异，基因表达的存活曲线在最后一种划分方式表现良好($p=0.021$)，而相对熵得分的存活曲线在第三种划分方式下表现良好($p=0.042$)。LUSC 的最后一种划分方式为：IA、IB、IIA、IIB、IIIA、IIIB 和 IV。LUSC 的第三种划分方式是：IA，IB，IIA 和 IIB，IIIA，IIIB，IV。通过查询大量的资料，我们发现相对熵得分的结果更加合理。理由如下：

- I 期和 II 期癌症的生存率和治愈率都很高。
- III 期癌症在某些情况下可以治愈。
- IV 期复发癌几乎无法治愈。治疗的目的是延长和提高生活质量[5]。

通过相关的文献我们发现相对熵得分的结果更符合临床实验。

之后，在第三种划分方式下，我们将这些样本的相对熵得分放入分类器中，结果如图 3 所示，这种划分方式下得到的平均 auc 为 0.9768532。然而，将基因表达放入同一分类器，结果却大相径庭，通过基因表达的方法得到的平均 auc 仅为 0.5658102。

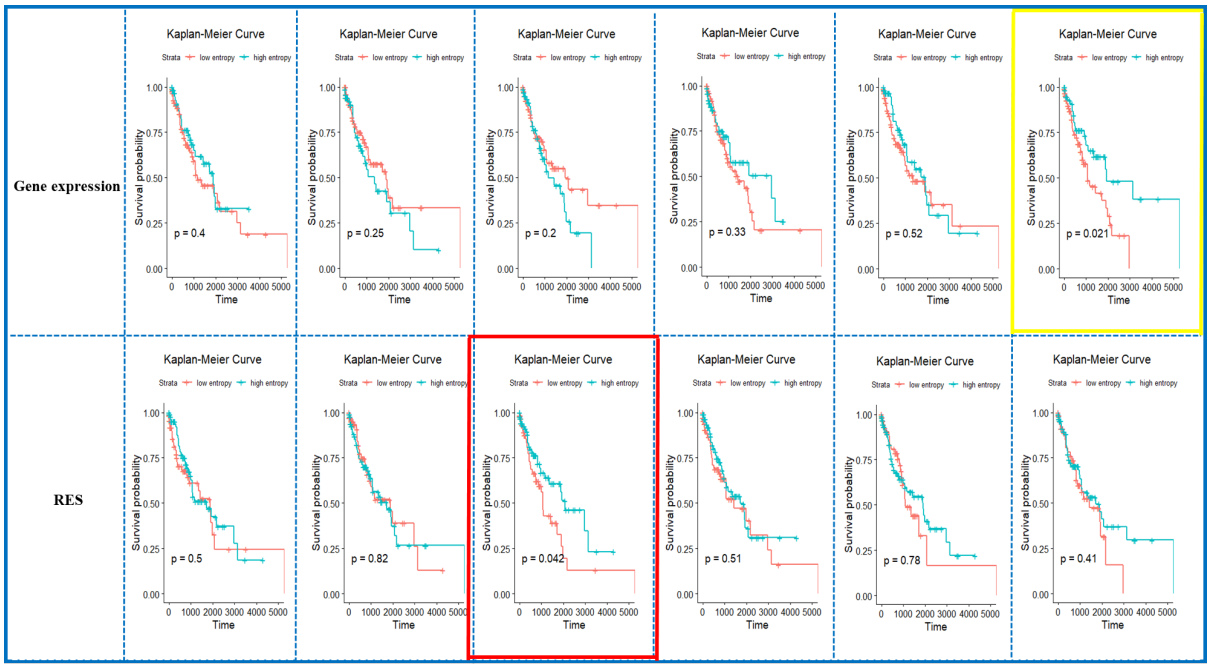

Figure 2. Comparison of gene expression and RES on the LUSC data set, respectively. There are 7 stages and 178 samples for LUSC. There are 6 divisions. Survival analysis under different partition modes are shown respectively in Figure 2. Obviously, the third division is most accord with our expectations

图 2. LUSC 数据集上基因表达和相对熵得分结果的比较。178 个样本，7 个阶段，6 种划分方式。不同划分方式下的生存分析如图 2 所示。显然，第三种划分方式最符合我们的期望

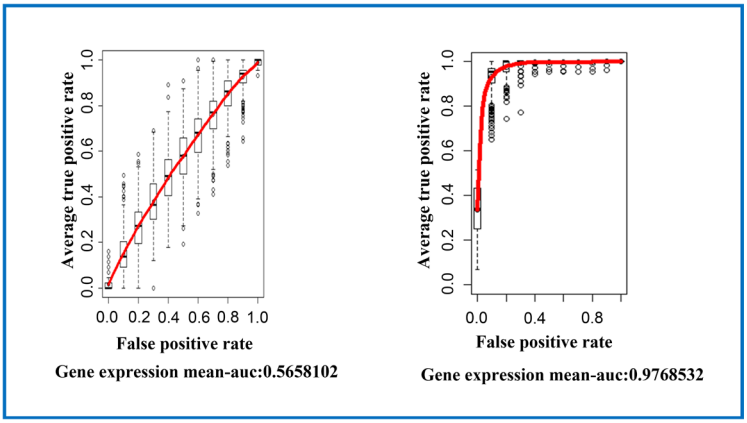

Figure 3. The classification of the third partition. And the mean-auc of the classification is 0.9768532, while the gene expression mean-auc is only 0.5658102

图 3. 第三种划分方式下的分类结果。RES 平均 auc 为 0.9768532，基因表达得到的平均 auc 只有 0.5658102

为了演示 LUSC 差异网络的演化，图 4 分别显示了正常状态和疾病状态的网络。显然，这两个网络之间存在着非常显著的差异。

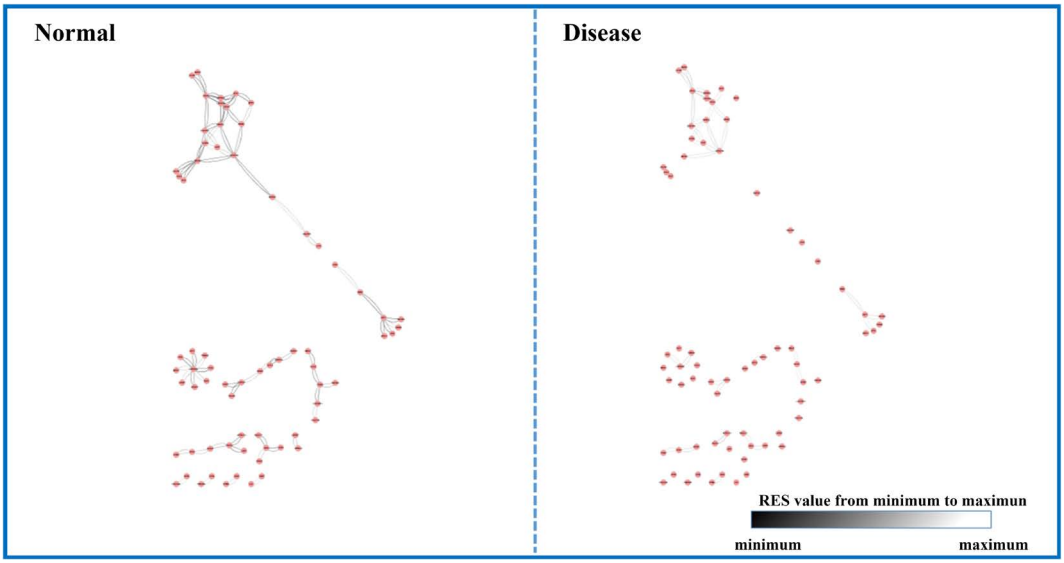

Figure 4. The dynamical evolution of differential network of LUSC shows the difference of the network in the normal state and the disease state, respectively
图 4. LUSC 差异网络的动力学演化。分别显示了在正常状态和疾病状态下的网络差异

4. 讨论

相比于临床上将疾病的发展划分为多个阶段，本文提出的方法为大多数人而言提供了更多的便利。毕竟大多数人真正关心的是身体的健康与否。在这项工作中，我们提出了一个区分正常状态和疾病状态的算法。通过两个实际数据集验证了该方法的有效性。

这种方法有以下几个优点：

首先，该方法依赖于生物网络，结果是可令人信赖的。

其次，它具有较高的实用性，能很好地适应复杂的生物系统。

第三，它可以提前探测到疾病信号，从而达到疾病预警的目的。

最后，虽然该算法在某些方面取得了一定的进展，但算法的灵敏度和准确度仍有提高的空间。

致　谢

本文受广东省基础与应用基础研究基金资助(No. 2019B151502062)。

基金项目

广东省基础与应用基础研究基金资助(No. 2019B151502062)。

参考文献

[1] Chen, P., Li, Y., Liu, X., Liu, R. and Chen, L. (2017) Detecting the Tipping Points in a Three-State Model of Complex Diseases by Temporal Differential Networks. *Journal of Translational Medicine*, **15**, 217.

[2] Zhang, W.W. and Tao, Z. (2015) Diagnosing Phenotypes of Single-Sample Individuals by Edge Biomarkers. *Journal of Molecular Cell Biology*, **7**, 231-241.

[3] Non-Small Cell Lung Cancer Treatment (PDQ®): General Information about Non-Small Cell Lung Cancer. National

Cancer Institute Website. http://www.cancer.gov/types/lung/patient/non-small-cell-lung-treatment-pdq

[4] NCI Dictionary of Cancer Terms. National Cancer Institute Website. http://www.cancer.gov/dictionary

[5] Non-Small Cell Lung Cancer. https://medlineplus.gov/ency/article/007194.htm