Pure Mathematics 理论数学, 2019, 9(10), 1115-1122
Published Online December 2019 in Hans. http://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2019.910137

Hans汉斯

Clustering of Cancer Subtypes Based on Similarity Network Fusion of Normalized Euclidean Distance and Spectral Neighbor

Kuan Yang[1], Yaping Zhao[2]

[1]School of Mathematical Sciences, Ocean University of China, Qingdao Shandong
[2]Qingdao Foreign Affairs Service Vocational School, Qingdao Shandong
Email: yangkuan.51@163.com

Received: Nov. 14th, 2019; accepted: Nov. 27th, 2019; published: Dec. 4th, 2019

Abstract

Similarity network fusion (SNF) is an effective clustering method to identify cancer subtypes. By using SNF, patient similarity networks for each of their data types are integrated into a single similarity network, which contains all the information of patients. In this paper, a similarity network fusion based on normalized Euclidean distance and spectral clustering (NSSNF) is proposed by using normalized Euclidean distance and redefined neighbor of the patient to reduce the noise of data analysis in similarity networks and increase the complementarity between the data from different similarity networks. Finally, for the five cancer data types from the TCGA database, the data analysis was performed by the NSSNF method, and the results of the evaluation indexes DB and CH showed that the NSSNF method is superior to the SNF method, NSNF method and CSNF method.

Keywords

Subtype, Network Fusion, Normalization, Euclidean Distance, Neighbor, Spectral Clustering, Similarity Network

基于归一化欧氏距离和谱分邻的融合网络对癌症亚型聚类

杨　宽[1]，赵亚萍[2]

[1]中国海洋大学数学科学学院，山东　青岛
[2]青岛外事服务职业学校，山东　青岛
Email: yangkuan.51@163.com

收稿日期：2019年11月14日；录用日期：2019年11月27日；发布日期：2019年12月4日

文章引用：杨宽，赵亚萍．基于归一化欧氏距离和谱分邻的融合网络对癌症亚型聚类[J]．理论数学，2019，9(10): 1115-1122. DOI: 10.12677/pm.2019.910137

摘　要

相似网络融合(SNF)是一种确定癌症亚型的有效聚类方法，可以将病人不同数据类型的相似网络融合成一个相似网络，融合后的相似网络包含病人的所有信息。本文使用归一化的欧氏距离和重新定义的病人的邻居，提出了基于归一化欧氏距离和谱聚类分邻的相似网络融合(NSSNF)，减少了相似网络中数据分析时产生的噪声，并增加了不同相似网络的数据间的互补性，最后，针对数据库TCGA中的五种癌症数据，利用NSSNF方法进行数据分析，评价指标DB和CH的结果表明NSSNF方法优于传统SNF方法、NSNF方法和CSNF方法。

关键词

亚型，网络融合，归一化，欧氏距离，邻居，谱聚类，相似网络

1. 引言

2018 年 4 月 5 日，《Cell》公布了泛癌症图谱(Pan-Cancer Altas)，其分子特征是 20,000 多种原发性癌症和匹配的 33 种癌症类型的正常样本，其中，居于癌症发病率首位的肺癌是一种高度错综复杂的异质性疾病，目前肺癌的精确治疗已发展到新的阶段，已确定 EGFR 突变、ALK 融合等七种不同的肺癌亚型，这七种不同的肺癌亚型需要不同的靶向药物进行精确治疗，然而，肺癌以及其它癌症的治疗也存在着许多问题，如何寻找未知的靶基因，确定未知的癌症亚型还需要更多地努力[1] [2]。相似网络融合是一种有效确定癌症亚型的聚类方法，可以将病人不同的数据类型整合到一个网络中，通过相似网络融合，病人之间的弱相似性减少，降低了数据分析时的噪声，强相似性增加，增加了不同相似网络的数据间的互补性。2014 年由 Bo Wang [3]等人首次提出相似网络融合方法((Similarity Network Fusion) SNF)，利用 TCGA 计划中病人的三种特征 DNA 甲基化、mRNA 表达以及 microRNA 表达三种数据构建病人的相似融合网络，最后对融合后的网络进行聚类，对比于其它利用单一种数据进行聚类的方法提高了聚类的性能，但也存在着不足，少量的数据样本对应着高维数据，降低数据噪声，扩大数据间的互补性还需要提高。2014 年，Jiang Xingpeng [4]改进了邻居的定义，提出了一致相似网络融合(CSNF)，用一致 k 近邻代替原来的 k 近邻方法，以此来推断微生物之间的关系，提高了聚类性能。2017 年，Zhang Yong [5]等人提出一种新的基于相似网络融合的多视图聚类算法(RSNF)，这种方法结合了随机森林的强度优势以及 SNF 聚类方法的鲁棒性优势，并将其应用于分析人体微生物数据，改善了 SNF 的聚类性能。2017 年，Zhao Yaping [6]提出了包含邻居信息的相似网络融合(NSNF)，用包含邻居信息的多重紧密 k 近邻方法代替原来的 k 近邻方法来构建邻居间的相似网络，将其运用于癌症亚型聚类，与原有方法相比，聚类效果提高很多。2018 年，张月[7]等人，提出了一种基于一致相似度网络融合的极化 SAR 图像非监督地物分类方法，修正了误分像素的类别标签，提高了地物分类的精度，和传统极化 SAR 图像地物分类方法相比，聚类性能有了很大的提高。2018 年，Ning Chen [8]提出了基于利用上下文信息改进 SNF 的信息检索(CI-SNF)，对 SNF 融合相似性采用 Jaccard 距离，增强了样本的局部一致性，引入倒排索引技术，提高了聚类的效率，并且将其应用于歌曲识别、图像分类、癌症亚型识别以及药物识别中。

相似网络融合经过这几年有了很大的发展，但是，如何更精确地区分亚型，提高聚类的有效性一直是一个发展中的难题。本文提出了基于归一化欧氏距离和谱分邻相似网络融合方法(NSSNF)，用两种方式改进了相似网络融合方法(SNF)，一方面，用归一化欧氏距离对病人相似性中的距离重新定义，这样求得的更具有全局性，减少了病人相似数据的噪声；另一方面，用谱聚类分邻的方法重新定义病人邻居间相似网络中的邻居，分邻方法更加科学，同一组邻居强相似性增加，不同组邻居弱相似性减少，增加了数据间的互补性，融合网络更具有全局性，聚类效果优于 SNF 方法、NSNF 方法和 CSNF 方法。

2. 方法

2.1. 数据来源

数据来源于 TCGA 网站，本文下载了五种癌症数据：肺癌(LSCC)、肾癌(KRCCC)、肠癌(COAD)、乳腺癌(BIC)和胶质瘤(GBM)。每种癌症数据都包含 3 种数据类型：DNA 甲基化、mRNA 表达、miRNA 表达。每种癌症详细信息见表 1。

Table 1. Detailed data for each cancer
表 1. 每种癌症的详细数据

癌症	病人数量(人)	DNA 甲基化(种)	mRNA 表达(种)	miRNA 表达(种)
肺癌(LSCC)	80	23,074	12,042	352
肾癌(KRCCC)	92	24,960	17,899	329
肠癌(COAD)	70	23,088	17,814	312
乳腺癌(BIC)	80	23,094	17,814	354
胶质瘤(GBM)	215	1305	12,042	534

2.2. 相似网络融合

本文提出的方法归一化欧氏距离和谱分邻的相似网络融合(NSSNF)一共有五个步骤：构建病人相似网络、病人相似网络标准化、构建邻居间的相似网络、相似网络融合和对融合网络进行谱聚类。

2.2.1. 构建病人相似网络

假设有n个病人和m种数据类型(比如：DNA 甲基化，mRNA 表达 miRNA 表达)。病人的相似网络可以表示成$G=(V,E)$，其中顶点集 V 表示病人$\{x_1,x_2,\cdots,x_i,\cdots,x_n\}$，边集 E 表示病人之间相似性程度，病人之间的相似网络可以用 $n\times n$ 的相似矩阵来表示，其中$W(i,j)$表示病人与病人间的相似性，受 Danfeng Qin [9]等人的启发，病人之间的相似性定义如下

$$W(i,j)=\exp\left(-\frac{d^2(x_i,x_j)}{\mu\varepsilon_{i,j}}\right), \tag{1}$$

其中，$W(i,j)$的值越大，代表着病人 x_i 和病人 x_j 相似性越高，μ 表示超参数并且可以经验性的给出，$\varepsilon_{i,j}$ [3]是消除缩放的尺度参数，定义如下

$$\varepsilon_{i,j}=\frac{mean\left(d(x_i,N_i)\right)+mean\left(d(x_j,N_j)\right)+d(x_i,x_j)}{3}, \tag{2}$$

其中，$d(x_i,x_j)$表示病人 x_i 病人 x_j 病人之间归一化欧氏距离，定义为

$$d\left(x_i,x_j\right)=\frac{\rho\left(x_i,x_j\right)}{\frac{1}{n-1}\sum_{l\neq i}^{n}\rho\left(x_i,x_l\right)}, \tag{3}$$

$\rho\left(x_i,x_j\right)$表示病人x_i和x_j之间的欧氏距离，$mean\left(d\left(x_i,N_i\right)\right)$表示距病人$x_i$最近的$k$个邻居归一化欧氏距离的平均值，$N_i$是指包含病人$x_i$在内距它归一化欧氏距离最近的$k$个病人组成的邻居。

本文用归一化欧氏距离为基础描述这种相似关系，构建病人之间的相似网络更具有全局性，和SNF方法相比，减少了数据噪声。

2.2.2. 病人相似网络标准化

因为要把不同数据类型的相似矩阵融合成一个矩阵，所以要对不同数据类型的相似网络进行标准化[3]。病人相似网络被重新定义之后，病人相似性具有全局性，病人x_i和病人x_j之间的相似性$W\left(i,j\right)\neq W\left(j,i\right)$，标准化过程定义如下

$$P\left(i,j\right)=\begin{cases}\dfrac{W\left(i,j\right)}{\sum_{j\neq l}W\left(i,l\right)+\sum_{i\neq h}W\left(j,h\right)}, & j\neq i\\ \dfrac{1}{2}, & j=i\end{cases} \tag{4}$$

从而，可以求出标准化后的病人相似网络$P=\left(P\left(i,j\right)\right)$。

2.2.3. 构建邻居间的相似网络

下面通过谱分邻的方法构建邻居间的相似网络，谱分邻就是用谱聚类[10]算法定义病人邻居间相似网络中的邻居，步骤如下

1) 矩阵W作为相似性矩阵。

2) 求度矩阵D，其中$D\left(i,j\right)=\begin{cases}\sum_{j=1}W\left(i,j\right), & i=j\\ 0, & i\neq j\end{cases}$。

3) 求拉普拉斯矩阵$L=I-D^{-1/2}WD^{-1/2}$。

4) 求L最小的k个特征值对应的特征向量V。

5) 对特征向量V进行*K-means*聚类。

通过谱分邻方法，n个病人被聚成C组，一般情况下，病人邻居间的聚类数目为$C=\left[\frac{n}{k}\right]$。病人邻居间的相似网络定义如下

$$S\left(i,j\right)=\begin{cases}\dfrac{2W\left(i,j\right)}{\sum_{l\in U_r}W\left(i,l\right)+\sum_{h\in U_r}W\left(j,h\right)}, & i,j\in U_r\\ 0, & \text{otherwise}\end{cases} \tag{5}$$

其中，$U_r\left(r=1,2,\cdots,C\right)$表示每一组邻居，每一组内的病人之间的相似性很高，不同组病人间的相似性为零，从而，可以求出病人邻居间的$S=\left(S\left(i,j\right)\right)$。

本文通过谱分邻方法构建病人邻居间的相似网络，和SNF方法相比，同一组邻居间的相似性更高，不同组邻居间的相似性更低，提高了聚类效果。

2.2.4. 相似网络融合

相似网络融合采用了一种基于消息传递理论的非线性方法[11]，矩阵P包含所有病人的相似信息，而矩阵S只包含病人邻居间的相似信息，将矩阵P作为初始矩阵，矩阵S作为核矩阵进行迭代融合，进

行一定次数的迭代，每种数据类型的网络逐渐收敛。

对于一种癌症，有 n 个病人和 3 种数据类型(DNA 甲基化，mRNA 表达和 miRNA 表达)，通过公式(1)、(4)、(5)，可以分别计算出病人 DNA 甲基化的相似网络 $W^{(1)}$、标准化矩阵 $P^{(1)}$ 和邻居间相似矩阵 $S^{(1)}$，同理，计算出 mRNA 表达矩阵 $W^{(2)}$、$P^{(2)}$ 和，miRNA 表达矩阵 $W^{(3)}$、$P^{(3)}$ 和 $S^{(3)}$。令 $P_0^{(1)}=P^{(1)}$，$P_0^{(2)}=P^{(2)}$，$P_0^{(3)}=P^{(3)}$，$S_0^{(1)}=S^{(1)}$，$S_0^{(2)}=S^{(2)}$，$S_0^{(3)}=S^{(3)}$。迭代步骤如下

$$P_{t+1}^{(1)}=S^{(1)}\frac{P_t^{(2)}+P_t^{(3)}}{2}\left(S^{(1)}\right)^{\mathrm{T}}, \tag{6}$$

$$P_{t+1}^{(2)}=S^{(2)}\frac{P_t^{(1)}+P_t^{(3)}}{2}\left(S^{(2)}\right)^{\mathrm{T}}, \tag{7}$$

$$P_{t+1}^{(3)}=S^{(3)}\frac{P_t^{(1)}+P_t^{(2)}}{2}\left(S^{(3)}\right)^{\mathrm{T}}, \tag{8}$$

其中，t 是迭代的次数。经过 t 次迭代之后，网络融合矩阵可用下列矩阵表示

$$P=\frac{P_t^{(1)}+P_t^{(2)}+P_t^{(3)}}{3}。 \tag{9}$$

从而，可以求出包含病人全局信息的融合网络 $P=\left(P(i,j)\right)$。

2.2.5. 对融合网络进行谱聚类

包含全局信息的融合矩阵 P 作为相似性矩阵进行谱聚类[10]，谱聚类的步骤如下：

1) 融合矩阵 P 作为相似性矩阵。
2) 求度矩阵 D，矩阵 D 是相似矩阵 P 的度矩阵。
3) 求拉普拉斯矩阵 L，其中 $L=I-D^{-1/2}PD^{-1/2}$。
4) 求出 L 的最小的 k 个特征值和对应的特征向量 V。
5) 将特征向量 V 进行 *K-means* 聚类。

从而，n 个病人被聚成了 k 个不同的类别。

3. 结果

评价聚类的有效性指标主要有两种。第一种是 Davies Bouldin (DB) [12]指标，DB 指标主要描述病人的类内散度与各聚类中心的间距，定义为

$$\mathrm{DB}=\frac{1}{k}\sum_{i=1}^{k}\max_{i=1,2,\cdots,j;i\neq j}\left\{\frac{\left(\sqrt{\frac{1}{n_i}\sum_{x\in C_i}d^2\left(x,z_i\right)}+\sqrt{\frac{1}{n_j}\sum_{x\in C_j}d^2\left(x,z_j\right)}\right)}{d\left(z_i,z_j\right)}\right\}, \tag{10}$$

其中，z_i 和 z_j 分别是第 i 类和第 j 类的中心点，n_i 和 n_j 分别是第 i 类和第 j 类的类内数量。从 DB 指标可以看出 DB 越小表示类与类之间的相似性越低，从而对应最佳的聚类结果。

第二种是 Calinski Harabasz (CH) [13]指标，CH 中类内离差矩阵描述紧密度，类间离差矩阵描述分离度，定义为

$$\mathrm{CH}=\frac{\sum_{i=1}^{k}n_i\cdot d^2\left(z_i,z_{tot}\right)}{k-1}\cdot\frac{n-k}{\sum_{i=1}^{k}\sum_{x\in C_i}d^2\left(x,z_i\right)}, \tag{11}$$

其中，z_{tot} 是整个数据集的中心。从 CH 指标可以看出 CH 越大代表着类自身越紧密，类与类之间越分散，

即聚类结果更好。

通过 SNF、CSNF、NSNF 和 NSSNF 方法，分别对 5 种癌症进行了相似网络融合，并计算了 4 种方法的聚类性能。详细结果见表 2，表 3。

Table 2. Detailed data for CH Value
表 2. 每种癌症 CH 指标的详细数据

癌症 / 方法	GBM	COAD	BIC	LSCC	KRCCC
SNF	31.6507	20.3564	31.1949	68.4477	41.0086
NSSNF	39.8457	23.1632	44.4084	93.1285	55.4177
NSNF	34.3276	21.8923	38.5398	88.2387	54.7843
CSNF	37.2583	22.4897	32.5632	77.2634	50.7786

Table 3. Detailed data for DB Value
表 3. 每种癌症 DB 指标的详细数据

癌症 / 方法	GBM	COAD	BIC	LSCC	KRCCC
SNF	1.1371	1.0084	0.5090	0.5480	0.7586
NSSNF	1.0986	0.8075	0.3514	0.3871	0.5913
NSNF	1.1028	0.9376	0.4562	0.4956	0.7289
CSNF	1.1267	0.9146	0.4034	0.4840	0.6547

对比四种 SNF、NSSNF、NSNF 和 CSNF 方法，数据表明 NSSNF 方法的 DB 值小于其它三种方法的 DB 值，NSSNF 方法的 CH 值大于其它三种方法的 CH 值，因此本文改进的 NSSNF 方法聚类性能优于其它三种方法。具体对比结果见图 1，图 2。

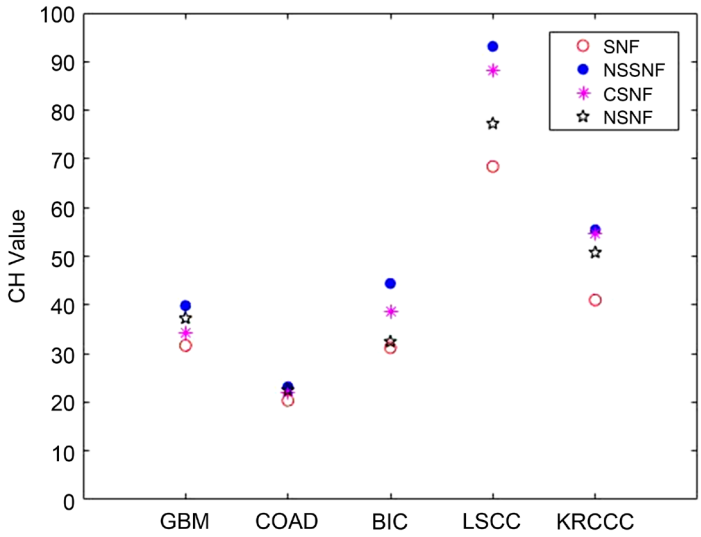

Figure 1. Comparison of CH values of five cancers data
图 1. 五种癌症数据 CH 值的比较

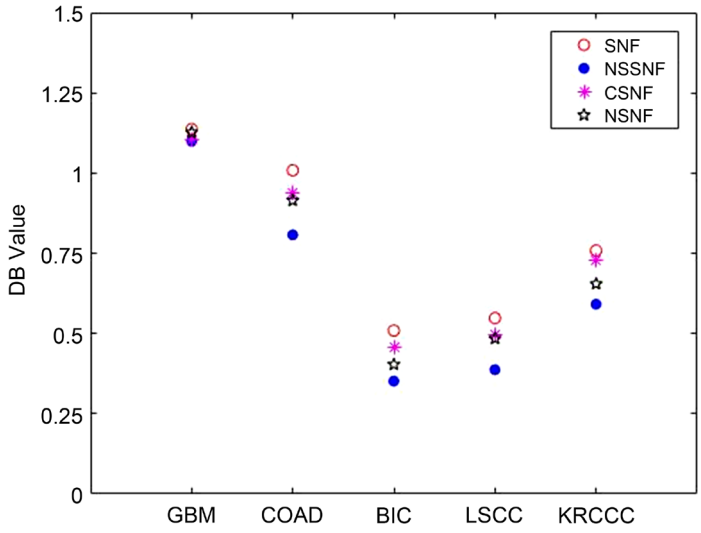

Figure 2. Comparison of DB values of five cancers data
图 2. 五种癌症数据 DB 值的比较

4. 结论

本文采用基于归一化欧氏距离的 NSSNF 方法，通过谱聚类改进邻居定义，构建患者相似度网络，将不同数据类型的相似性网络融合到一个相似性网络中，然后进行谱聚类。最后，通过 DB 值和 CH 值来评价谱聚类的有效性，数据分析结果均表明，NSSNF 聚类方法明显优于 SNF 方法、CSNF 方法和 NSNF 方法。

基金项目

国家自然科学基金项目(11271341)资助。

参考文献

[1] Wu, D., Wang, D.C., Cheng, Y., *et al*. (2016) Roles of Tumor Heterogeneity in the Development of Drug Resistance: A Call for Precision Therapy. *Seminars in Cancer Biology*, **42**, 13-19. https://doi.org/10.1016/j.semcancer.2016.11.006

[2] 周彩存, 刘桑田. 肺癌的靶向治疗与精准医学[J]. 医学研究生学报, 2017, 30(11): 1132-1139.

[3] Wang, B., Mezlini, A., Demir, F., *et al*. (2014) Similarity Network Fusion for Aggregating Data Types on a Genomic Scale. *Nature Methods*, **11**, 333-337. https://doi.org/10.1038/nmeth.2810

[4] Jiang, X.P. and Hu, X.H. (2014) Inferring Microbial Interaction Networks Based on Consensus Similarity Network Fusion. *College of Computing and Informatics*, **57**, 1115-1120. https://doi.org/10.1007/s11427-014-4735-x

[5] Zhang, Y., Hua, X. and Jiang, X.P. (2017) Multi-View Clustering of Microbiome Samples by Robust Similarity Network Fusion and Spectral Clustering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **14**, 264-271. https://doi.org/10.1109/TCBB.2015.2474387

[6] Zhao, Y.P., Zhao, X.Q., Wang, K.D., *et al*. (2017) Clustering Cancer Subtypes Based on Neighborhood-Information Embedded Similarity Network Fusion. *Periodical of Ocean University of China*, A1, 155-160.

[7] 张月, 邹焕新, 邵宁远, 等. 基于相似度网络融合的极化 SAR 图像地物分类[J]. 系统工程与电子技术, 2018, 40(2): 295-302.

[8] Chen, N. (2019) CI-SNF: Exploiting Contextual Information to Improve SNF Based Information Retrieval. *Information Fusion*, **52**, 175-186. https://doi.org/10.1016/j.inffus.2018.08.004

[9] Qin, D.F., Wengert, C., van Gool, L., *et al*. (2013) Query Adaptive Similarity for Scale Object Retrieval. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, 23-28 June 2013, 1610-1617.

[10] Von Luxburg, U. (2007) A Tutorial on Spectral Clustering. *Statistics and Computing*, **17**, 395-416.

https://doi.org/10.1007/s11222-007-9033-z

[11] Nigro, J.M., *et al*. (2005) Integrated Array-Comparative Genomic Hybridization and Expression Array Profiles Identify Clinically Relevant Molecular Subtypes of Glioblastoma. *Cancer Research*, **65**, 1678-1686. https://doi.org/10.1158/0008-5472.CAN-04-2921

[12] Davies, D.L. and Bouldin, D.W. (1979) A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis Machine Intelligence*, **1**, 224. https://doi.org/10.1109/TPAMI.1979.4766909

[13] Caliski, T. and Harabasz, J. (1974) A Dendrite Method for Cluster Analysis. *Communications in Statistics*, **3**, 1-27. https://doi.org/10.1080/03610917408548446