LSTM model for identification of essential proteins

Our method

之前我们运用了 LSTM 模型对关键蛋白质进行了预测，所采用的方法与李敏老师在 BIBM 中提出的方法类似，只是我们采用了 LSTM 模型来提取特征。

我们借鉴了李敏老师的思想，通过 Node2vec 算法提取网络拓扑结构特征并与基因表达数据结合用于关键蛋白质的识别，只是这里我们对所有数据采用序列模型进行处理。

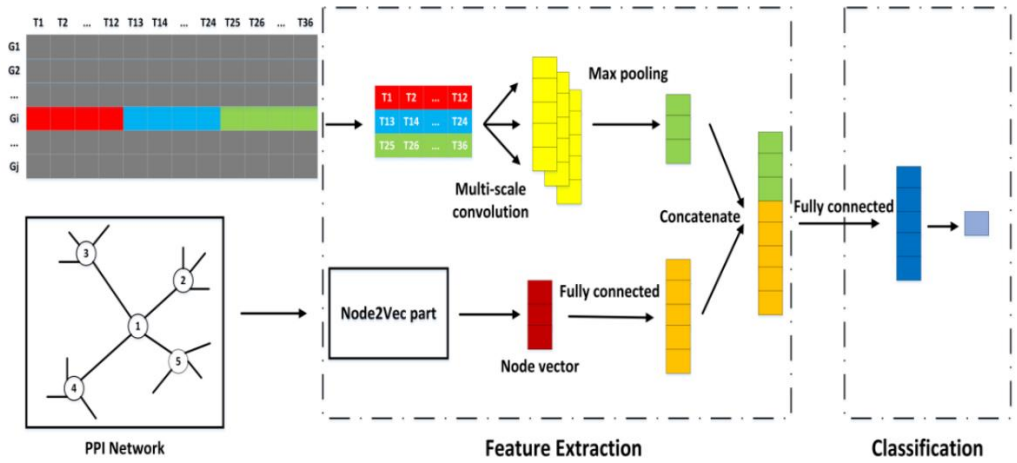

图 1. 李敏老师的网络结构

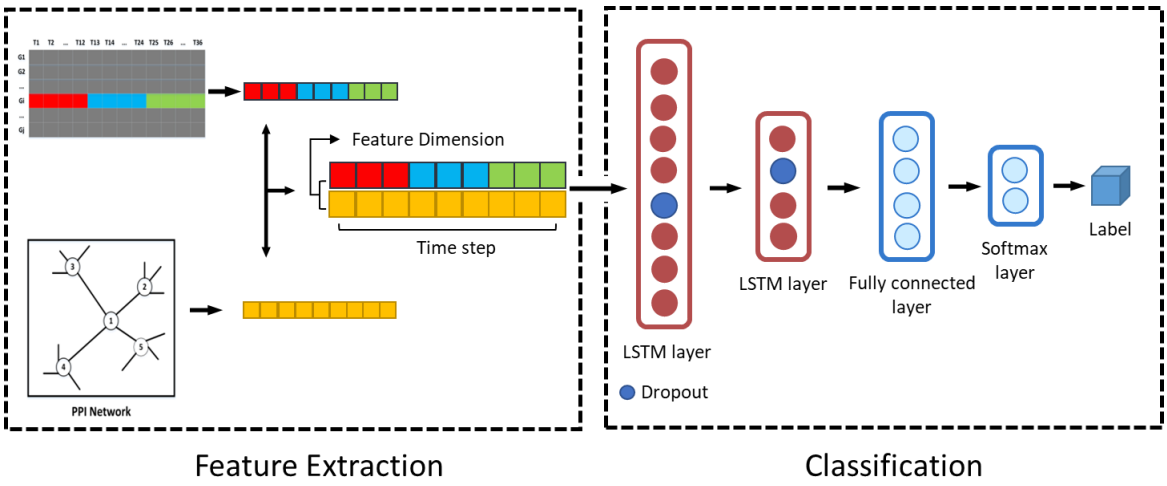

图 2. An overview of our proposed deep learning framework for identifying essential proteins.

在里的序列维度为 2，分别为 node2vec 的网络结构信息和 gene expression data，序列的长度为 36，也即为基因表达序列的长度，这里的网络结构信息长度也只有 36，是为了与基因表达信息长度匹配，而李敏老师的算法里长度是 72，所以我们为了与基因表达数据匹配牺牲了一部分信息。模型对比的结果如下。

Models	Accuracy	Precision	Recall	F-measure	AUC
SVM	0.7053	0.2580	0.1777	0.2105	0.5163
Decision Tree	0.64047	0.1894	0.1911	0.1902	0.4795
Adaboost	0.6748	0.2087	0.1688	0.1867	0.4936
Random Forest	0.6650	0.21	0.1866	0.1976	0.4937
Li et.al	**0.7507**	**0.4186**	0.2317	0.2983	0.5681
Our method	0.71638	0.3402	**0.3933**	**0.3648**	**0.5971**

表 1. compared with machine learning methods

Models	Accuracy	Precision	Recall	F-measure	AUC
DC	0.7329	0.4052	0.3538	0.3778	0.5997
CC	0.7137	0.3572	0.3119	0.333	0.5725
BC	0.7125	0.3542	0.3093	0.3302	0.5708
EC	0.7188	0.3699	0.3230	0.3449	0.5797
LAC	**0.7565**	**0.4641**	**0.4053**	**0.4327**	**0.6331**
Our method	0.71638	0.3402	0.3933	0.3648	0.5971

表 2. compared with traditional methods

结果的话比李敏老师的算法（李敏老师的算法我采用 pytorch 复现的，使用数据与 LSTM model 的一样）表现稍微好一点，但是还是很弱，连 traditional 的方法都比不了。

Now

最近在改这篇 SGC，结合了我们做医学图像的一些经验，我想能不能和用 LSTM 做多模态分割一样把多种特征信息给融合进来，之前的 model 为了维度匹配还牺牲了一部分信息，其他的生物信息，像子细胞位置信息，蛋白质模块信息长度都是 1，和基因表达信息的长度不匹配，所以我们这里采用了分割的思想对特征进行整合。

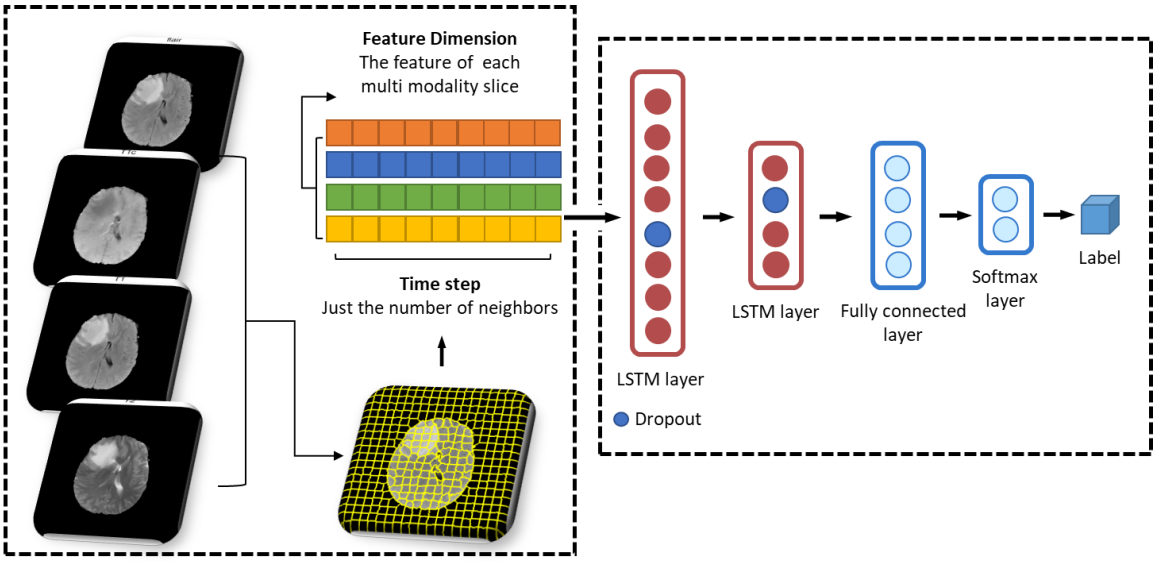

图 3. LSTM model for brain tumor segmentation.

思想是我们用 SGC 中提取得到的特征作为节点特征，然后按照网络信息构建序列，这样序列的维度就是采用的特征的个数（SGC 中采用的是 NC, PeC, PC, NSL），长度就是每个节点的邻居个数。

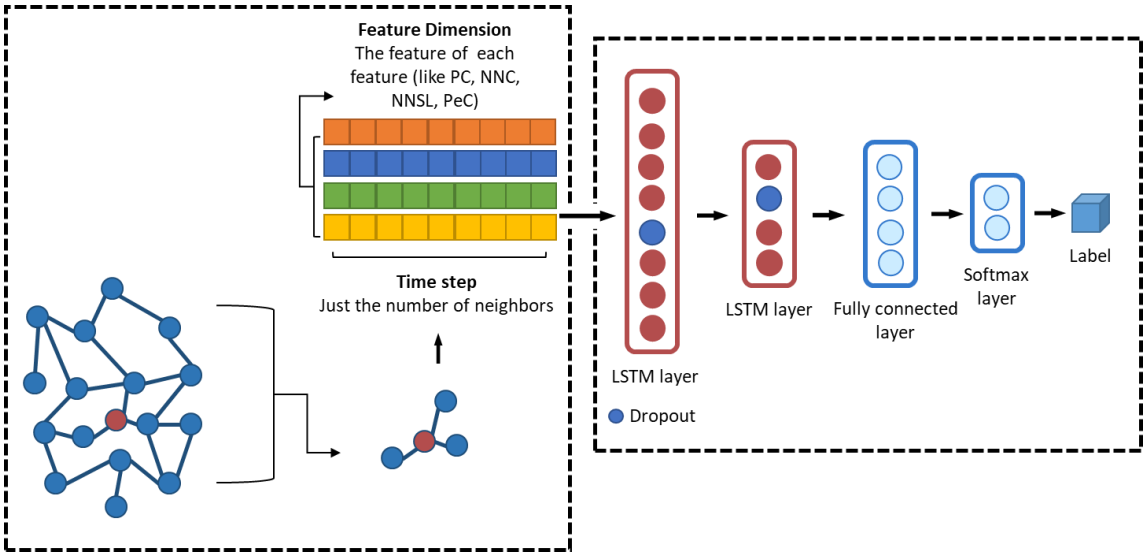

图 4. New LSTM model for brain tumor segmentation.

结果的话取得了较大的识别精度提升，可以看到相比于之前的方法，各个评价方法上都取得了较大的提升。比李敏老师的方法提高了不少，相比于传统的算法也比较能看了。

Models	Accuracy	Precision	Recall	F-measure	AUC
SVM	0.7053	0.2580	0.1777	0.2105	0.5163
Decision Tree	0.64047	0.1894	0.1911	0.1902	0.4795
Adaboost	0.6748	0.2087	0.1688	0.1867	0.4936
Random Forest	0.6650	0.21	0.1866	0.1976	0.4937
Li et.al	0.7507	0.4186	0.2317	0.2983	0.5681
Our method	0.71638	0.3402	0.3933	0.3648	0.5971
Now	**0.7556**	**0.5058**	**0.5159**	**0.5108**	**0.6751**

表 3. compared with machine learning methods

Models	Accuracy	Precision	Recall	F-measure	AUC
DC	0.7329	0.4052	0.3538	0.3778	0.5997
CC	0.7137	0.3572	0.3119	0.333	0.5725
BC	0.7125	0.3542	0.3093	0.3302	0.5708
EC	0.7188	0.3699	0.3230	0.3449	0.5797
LAC	**0.7565**	0.4641	0.4053	0.4327	0.6331
Our method	0.71638	0.3402	0.3933	0.3648	0.5971
Now	**0.7556**	**0.5058**	**0.5159**	**0.5108**	**0.6751**

表 4. compared with traditional methods

这里为了进一步明确各个特征的 1 作用，我们还做了一下消融实验，也就是除掉某个特征，看看整体表现怎么样。这里也可以看出 NNSL 对算法的影响最大。

Models	Accuracy	Precision	Recall	F-measure	AUC
Without PC	0.7409	0.4615	0.4918	0.476	0.6556
Without NNSL	0.7222	0.4111	0.5086	0.4547	0.6469
Without PeC	0.7242	0.3833	**0.5769**	0.4606	0.6694
Without NNC	0.7311	0.4491	0.5637	0.5	0.6736
Now	**0.7556**	**0.5058**	0.5159	**0.5108**	**0.6751**

表 5. Ablation Experiment