生物统计课程论文

徐伟泽
2018302110174
预防兽医学，动物医学院
华中农业大学

2019 年 6 月 17 日

摘要：本文探讨了几种机器学习模型在 DNA 序列分类与回归问题上的应用，结合具体的数据集对几种模型的相应问题上的表现进行了评估。其中分类问题数据来自于 DAP-Seq 数据得到的转录因子结合位点序列数据，回归数据来自于 CRISPR-Cas9 敲除实验对癌细胞生长影响数据。通过比较各种方法应用在相应数据上得到的结果，讨论了机器学习模型在核酸序列数据的应用方法以及存在的问题。

1 前言

目前机器学习、统计学习方法已经被广泛应用于 DNA 序列与基因组学相关实验数据的处理、挖掘。机器学习技术大体上可分为需要标签数据的有监督学习技术与不需要标签的无监督学习技术。对于有监督学习，两类常见的问题是分类问题和回归问题。在基因组数据的机器学习任务上，分类模型常被应用于序列功能预测，比如 TF binding site 预测[1,2]、染色质开放区域预测[3]、非编码 DNA 功能预测[4] 等问题。对于序列数据，由于其本身是非结构化的，无法直接作为特征，在训练模型前需要对其进行特征提取, 图 1 (a)。根据之前的研究，特征提取根据对信息的利用情况大致上可以分为两类：1. 不考虑上下文的特征，比如 k-mer 计数，或者其他在此之上衍生出的特征提取技术，比如 gapped k-mer[1]。2. 考虑序列上下文信息的特征提取，比如 sequence embedding[5]、RNN（递归神经网络）[6] 等等。一般来说后者由于考虑了更丰富的信息，在数据量充足、问题复杂时往往表现出更加优异的结果。但相应的对数据量的要求较高而且计算效率上不如前者，所以在实践中还需要结合具体的问题与数据进行选择。

2 材料与方法

2.1 数据

2.1.1 分类问题

分类问题数据来自于 2016 年发表于 Nature Biotechnology 的 DAP-Seq 数据[7]。数据包括正样本与负样本共 5834 条长度为 201bp 的 DNA 序列。其中正样本为实验得出的 TF（Transcription Factor，译：转录因子）结合位点附近的 DNA 序列，负样本为染色体上随机抽取的相等长度的序列。

2.1.2 回归问题

回归问题的数据集来自于 CRISPR-Cas9 敲除 p53 enhancer 筛选实验[8]。预测的根据为敲除位点附近的核酸序列，预测目标为敲除后癌细胞生长的 Enrichment Z-Score。

2.2 序列特征提取

2.2.1 k-mer 计数

k-mer 计数是一种常用的较为朴素的序列特征提取方法。k-mer 指的是序列中长度为 k 的子序列，当序列为 DNA 时，所有可能的 k-mer 种类数量为 4^k，所以 k-mer 计数特征可表示为一个长度为 4^k 的向量 $F_k(s) = [c_1, ...c_i, ...c_{4_k}]$ 其中 c_i 为第 i 个 k-mer 再序列 s 中出现的次数。对于 k-mer 计数特征，k 是唯一的参数，在之前的研究中一般将 k 设置为 6 左右[1,5]。

2.2.2 k-mer sentence 与 Seq2Vec

除了将 k-mer 计数作为序列特征，之前的研究中还有研究者借鉴自然语言处理中的方法，将序列视为由 k-mer 作为词的句子。然后将句子嵌入到欧式空间中，将嵌入后得到的向量作为特征[5]。具体的嵌入方法为对 sentence 中每一个 k-mer word 的 embedding 进行优化，使得 softmax 函数输入前 n 个 words 预测接下来出现的 word 的准确率最大。具体细节可参照 2018 年发表于 BMC Genomics 上的文献[5]。最终的实现中采用了 Python 包 Gensim[9] 中的 Doc2Vec 实现。

2.2.3 Recurrent Neural Network

既然将 DNA 序列特征提取能够类比于自然语言序列的特征提取，那么可以进一步的借鉴自然语言处理中的其他技术来进行 DNA 序列特征提取。比如使用 RNN (Recurrent Neural Network, 译：递归神经网络) 或者 LSTM (Long Short-Term Memory)[10]、GRU（Gated Recurrent Unit）[11] 等基于神经网络的技术来做特征提取。当然之前已经有研究者使用这类技术提取特征用于 TF 结合位点预测[6]。

本文中使用了双向 LSTM 与注意力机制 相结合的神经网络架构，见图 1。它处理以 $k = 6$ 的 k-mer word 为单位的 k-mer sentence，提取特征用于序列分类问题。双向的 LSTM 单元用于考虑 DNA 序列中隐含的上下文信息，注意力单元负责为序列中不同的 k-mer 输入分配不同的权重。网络的参数通过 SGD（Stochastic gradient descent，译：随机梯度下降）方式进行更新。

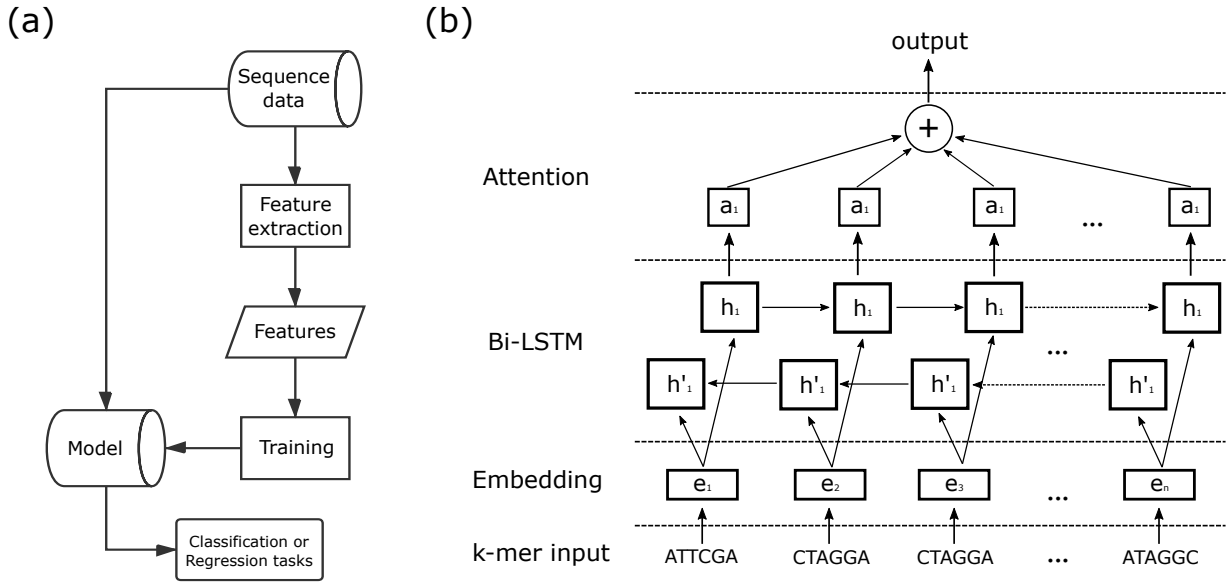

图 1: (a) 序列数据相关机器学习任务处理流程图。(b) 本文中使用的双向 LSTM 网络架构示意图。

2.3 分类与回归模型

本文中试验的分类器模型有两种：SVM（Support Vector Machine，译：支持向量机）与 Random Forest（译：随机森林）。试验的回归模型选择了一种线性模型 LASSO（least absolute shrinkage and selection operator）与一种基于树的方法 Gradient Boosting Regression Tree。超参数选取使用 Grid Search 策略搜索交叉检验中的最优参数。由于均为常见统计学习模型，原理此处不再赘述。代码实现均来自于 Python 包 Scikit-Learn。

2.4 模型选择与结果评价

2.4.1 交叉检验

CV（Cross Validation，译：交叉检验）策略被用作模型超参数选择与结果评价。在模型选择过程中，首先在样本中划分出 test set（测试集），然后剩下的部分进行 k-Fold 交叉检验用于模型选择。将数据划分为 k 份，然后一次将其中的一份作为 validation set，其他作为 training set。进行 k 次训练与测试，选出使得整体结果最优的超参数作为最终模型。然后使用 test set 以外的所有数据进行模型训练，然后再 test set 上进行测试，对模型进行评价。

2.4.2 分类结果评价

在二分类试验中，分类结果可划分为四类：TP（True Positive，真阳性）、FP（False Positive，假阳性）、TN（True Negative，真阴性）、FN（False Negative，假阴性）。由此，可以计算以下指标作为分类结果好坏的评价：1. ACC（准确率）用于衡量总体分类正确样本所占比例，$ACC = \frac{TP+TF}{TP+FP+TN+FN}$。2. Precision（精确度）衡量被分类为阳性的结果中 TP 所占比例，$precision = \frac{TP}{TP+FP}$。3. Recall（召回率）衡量被分类为阴性的结果中 TN 所占比例，$recall = \frac{TN}{TN+FN}$。4. F_1 为 Precision 与 Recall 的调和平均数，对二者权衡考虑，$F_1 = (\frac{recall^{-1}+precision^{-1}}{2})^{-1} = \frac{2TP}{2TP+FP+FN}$。5. ROC AUC （Receiver Operating Characteristic Area Under Curve）ROC 曲线为通过改变判断阈值，使得假阳性率与真阳性率不断发生变化而得到的曲线。该曲线下的面积可作为二分类结果好坏的一个不受判定阈值影响的衡量标准。

2.4.3 回归结果评价

对于回归试验的结果，主要考察的指标有两个。1. MSE（Mean Square Error）为真实结果与模型预测结果之差的平方均值，$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$。数值越小代表回归模型对数据的拟合越小。2. R^2 用来衡量模型所解释的变异占数据总变异的比例，越接近于 1 代表回归模型对数据变异的解释越好。定义为：$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$，其中 $SS_{tot} = \sum_i y_i - \overline{y}$ 为总体变异量，$SS_{res} = \sum_i y_i - \hat{y}_i$ 为模型解释后剩下的变异量。

3 结果

3.1 分类问题

分类问题的评价结果见表 1，从中可以看出，基于 k-mer 计数特征训练出的分类器可以很好的解决该问题，AUC 可以达到接近 99% ，而 Seq2Vec 提取到的特征训练出的分类器分类的结果反而不如 k-mer 计数，这可能是由于样本数量不能满足需要导致的。而双向 LSTM 在这个问题上也可以得到很好的结果。在测试数据集上的 ROC 曲线见图 2(a)-(c)。

3.2 回归问题

回归问题的评价结果见表 2，从目前的结果来看，R^2 小于 0 且波动较大，模型无法对数据趋势进行拟合。即使尝试不同的特征序列长度，该回归问题还是无法得到很好的求解。分析原因可能是数据中含有

Model	ACC	Precision	Recall	F1	AUC
k-mer + SVM	0.9657±0.0043	0.9654±0.0077	0.9646±0.0059	0.9650±0.0043	0.9906±0.0031
k-mer + Random Forest	0.9650±0.0086	0.9621±0.0080	0.9668±0.0134	0.9644±0.0088	0.9909±0.0037
Seq2Vec + SVM	0.8120±0.0177	0.8077±0.0212	0.8094±0.0211	0.8084±0.0177	0.8958±0.0112
Seq2Vec + Random Forest	0.7520±0.0151	0.7632±0.0115	0.7160±0.0313	0.7386±0.0194	0.8330±0.0130
Bi-LSTM + Softmax	0.9358±0.0915	0.9317±0.1067	0.9519±0.0480	0.9402±0.0795	0.9602±0.0624

表 1: 分类问题各模型 10Fold 交叉检验各检验指标的均值与标准差。

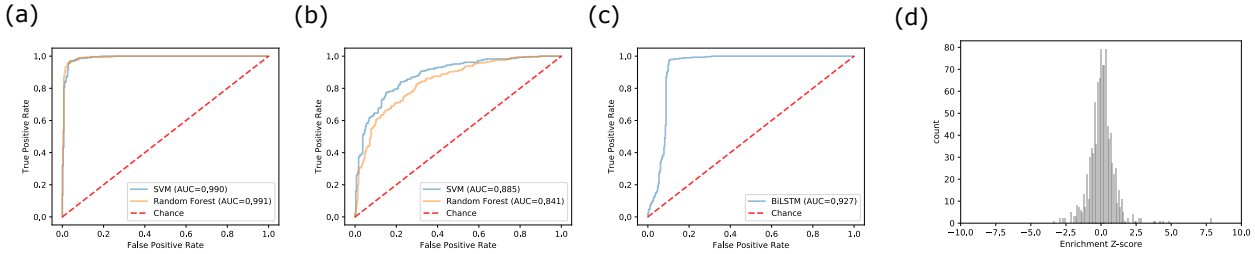

图 2: (a)-(c) 分类问题在测试数据集上的 ROC 曲线，从左至右对应 k-mer 计数、Seq2Vec、LSTM。(d) 回归问题中 Enrichment-Zscore 的分布。

的信息量不足，从对该数据集的统计中（图 2(d)）可以看出，大多数样本的 Enrichment score 较低，分布在 $[-2.5, 2.5]$ 之内，仅有少量（约 1.8%）的样本在此范围外。此外，样本的数量很少只有 1000 左右。而 k-mer 特征具有较高的维度，导致样本在空间中过于稀疏。

Model	seq length: 51	seq length: 101	seq length: 201
k-mer + Lasso	0.9278±1.4552	0.9278±1.4552	0.9278±1.4552
	-30.4784±23.1083	-30.4784±23.1083	-30.4784±23.1083
k-mer + GBTree	0.9986±1.3768	1.0926±1.4189	1.0668±1.3994
	-58.5660±38.4751	-73.7684±60.0907	-71.9895±48.3185
Seq2Vec + Lasso	0.9278±1.4552	0.9278±1.4552	0.9278±1.4552
	-30.4784±23.1083	-30.4784±23.1083	-30.4784±23.1083
Seq2Vec + GBTree	0.9921±1.3925	1.0401±1.4636	1.0500±1.4468
	-55.4674±27.9595	-59.5471±42.9350	-63.9507±30.7841
BiLSTM	1.0753±0.5401	1.0792±0.2081	1.0808±0.3321
	-0.0713±0.0512	-0.0638±0.0333	-0.0610±0.0219

表 2: 表格中列举了三种不同长度特征序列在几种不同模型上的回归效果。格中第一行为 MSE 的均值与标准差，第二行为 R^2 的均值与标准差。

4 讨论

本文在两个序列相关的统计学习问题数据集上进行了研究。讨论了三种对 DNA 序列数据进行特征提取的方法：k-mer 计数、Sequence embedding 与递归神经网络。并对在相应的数据上进行了实际测试。

其中分类数据使用 k-mer 计数方法进行特征提取后给分类器进行学习能够取得很好的效果。而后两种特征提取方法的结果反而不如 k-mer 计数，可能的原因是数据量不够。而对于回归问题，本文中试验的方法无法很好的对其进行拟合。可能的原因有：1. 数据量不够，由于序列的可变性较高，对于长度为 n

的序列，多样性有 4^n 种，且 k-mer 特征的维度较高而样本数目过少很可能使得数据过于稀疏，难以对其进行拟合。2. 数据中的信息量不足，通过对样本数据的统计，可以看到大多数样本的 y 值（Enrichment Z-score）在很小的范围内波动，仅有少量样本具有较大的 y 值。3. 序列敲除后的效应无法简单通过序列信息进行预测，也许仅仅是 DNA 序列信息不足以推测敲除后带来的表型。需要结合其他种类数据，比如 ChIP-Seq 试验数据等等。

针对以上分析，几种可能的改进：1. 收集更多的可靠的实验数据。2. 进行数据增强，在训练集中将具有较高 y 值的数据对应的序列做单碱基的改变，增加其在训练集中所占的比例，提升信息量。3. 以公开数据库中的数据为基础（如 ENCODE 数据库），综合敲除位点附近该细胞系的 ChIP-Seq 数据（如组蛋白修饰、转录因子结合等等），将其作为输入，增加信息量。4. 在增加数据量的同时，试验其他特征提取方法与回归模型。

References

[1] Mahmoud Ghandi et al. “Enhanced regulatory sequence prediction using gapped k-mer features”. In: *PLoS computational biology* 10.7 (2014), e1003711.

[2] Babak Alipanahi et al. “Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning”. In: *Nature biotechnology* 33.8 (2015), p. 831.

[3] David R Kelley, Jasper Snoek, and John L Rinn. “Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks”. In: *Genome research* 26.7 (2016), pp. 990–999.

[4] Jian Zhou and Olga G Troyanskaya. “Predicting effects of noncoding variants with deep learning–based sequence model”. In: *Nature methods* 12.10 (2015), p. 931.

[5] Wanwen Zeng, Mengmeng Wu, and Rui Jiang. “Prediction of enhancer-promoter interactions via natural language processing”. In: *BMC genomics* 19.2 (2018), p. 84.

[6] Zhen Shen, Wenzheng Bao, and De-Shuang Huang. “Recurrent Neural Network for Predicting Transcription Factor Binding Sites”. In: *Scientific reports* 8.1 (2018), p. 15270.

[7] Ronan C O’ Malley et al. “Cistrome and epicistrome features shape the regulatory DNA landscape”. In: *Cell* 165.5 (2016), pp. 1280–1292.

[8] Gozde Korkmaz et al. “Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9”. In: *Nature biotechnology* 34.2 (2016), p. 192.

[9] Radim Řehůřek and Petr Sojka. “Software Framework for Topic Modelling with Large Corpora”. English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. `http://is.muni.cz/publication/884893/en`. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[10] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[11] Kyunghyun Cho et al. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078* (2014).