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 16 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处理 DNA 序列与基因组学相关实验数据的分类与回归问题。

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[1]。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[2]。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[3,4]。

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[4]。

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)[5]、GRU（Gated Recurrent Unit）[6] 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[7]。

2.3 分类模型

2.3.1 Support Vector Machine

2.3.2 Random Forest

2.4 回归模型

2.4.1 Linear Regression

2.4.2 Lasso 与 Ridge regression

2.4.3 Gradient Boosting Regression Tree

2.5 结果评价

2.5.1 分类结果评价

2.5.2 回归结果评价

2.5.3 交叉检验

3 结果

3.1 分类问题

3.2 回归问题

4 讨论

References

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

[2] 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.

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

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

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

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

[7] 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.