生物统计课程论文

Runze li

2020 年 6 月 19 日

摘要：本文利用分类和回归这两类统计学习的方法，分别对 DNA 序列进行分类或回归学习，并且结合各类模型的评价标准对模型进行评估。在分类问题中，本文利用 SVM 和 randomForest 对 DAP-seq 得到的转录因子结合位点数据进行分类；而回归数据则是来自于 CRISPR-Cas9 敲除实验对癌细胞生长影响数据。本文分别在这两个数据集上进行建模，并且对模型进行评价。

关键词：分类　回归　SVM　randomForest　lasso

1. 前言

在生物信息邻域，统计学习已被较多的应用到了基因组学或者高通量测序中，将统计学习的方法和思想融入到组学数据分析中俨然成为了一种趋势。对于统计学习，主体分为聚类，分类和回归。目前，利用统计学习的思想可以预测基因结构，预测 lncRNA，甚至应用于序列比对当中。而对新的物种相关结构或者功能的预测，需要基于已挖掘的物种数据的基础上，并把这一部分称之为训练集，统计学习的模型需要利用训练集进行参数训练，选取合适的参数进行新物种的预测。基于上述思想，在训练集上选取合适的特征，将对模型的训练起到关键作用。当训练集数据量充足的时候，模型得到充分训练，预测的效果较好，否则，模型预测的假阳性率，假阴性率会上升。

所以，针对不同数据类型的数据，有不同的特征提取方法，灵活运用数据的特征，将有利于模型的预测，并且提高模型预测的准确度。

2. 材料与方法

2.1 数据

2.1.1 分类问题

分类问题的数据集来自于 2016 年在 Nature Biotechnology 的 DAP-seq 数据(Ronan,Malley et al,2016)。其中，正负样本分别为 2987 和 2847 条长度为 201bp 的 DNA 序列。其中正样本为实验得到的转录因子 TF 结合位点附近的 DNA 序列，负样本为在基因组上碎金抽取的等长 DNA 序列。

2.1.2 回归问题

回归问题的数据集来自于 CRISPR-Cas9 敲除基因 p53 增强子的筛选实验。其中，预测变量是敲除位点附近的 DNA 核酸序列，预测响应变量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 计数是在 DNA 序列特征提取中比较常用的方法，k-mer 是指在一段 DNA 序列中，长度为 k 的子序列所构成的集合(Mahmoud et al,2014)。通常对于一段序列，我们取相邻的 k-mer，并且统计 k-mer 子序列的种类个数(陈丽萍等,2003）。一般来说，k 值取 6 左右。

本文采用另一种基于 k-mer 的统计方法，即统计相邻 3 个碱基的组合，那么

一共有 64 中组合，分别是 AAA，AAC，AAG，AAT，ACA，ACC，ACG，ACT，AGA，AGC，AGG，AGT，ATA，ATC，ATG，ATT，CAA，CAC，CAG，CAT，CCA，CCC，CCG，CCT，CGA，CGC，CGG，CGT，CTA，CTC，CTG，CTT，GAA，GAC，GAG，GAT，GCA，GCC，GCG，GCT，GGA，GGC，GGG，GGT，GTA，GTC，GTG，GTT，TAA，TAC，TAG，TAT，TCA，TCC，TCG，TCT，TGA，TGC，TGG，TGT，TTA，TTC，TTG，TTT。

分别统计每三个相邻碱基各组合出现的次数，选取次数出现较多的碱基组合(蔡春等,2008)。

aaa	aac	aag	aat	aca	acc	acg	act	aga	agc	agg	agt	ata	atc	atg	att
4	2	1	1	1	0	7	1	0	0	7	0	1	0	0	1
caa	cac	cag	cat	cca	ccc	ccg	cct	cga	cgc	cgg	cgt	cta	ctc	ctg	ctt
0	3	0	1	0	0	0	0	1	0	5	0	1	0	0	2
gaa	gac	gag	gat	gca	gcc	gcg	gct	gga	ggc	ggg	ggt	gta	gtc	gtg	gtt
2	2	5	0	3	0	0	1	8	4	1	1	1	1	0	0
taa	tac	tag	tat	tca	tcc	tcg	tct	tga	tgc	tgg	tgt	tta	ttc	ttg	ttt
2	2	0	0	0	0	0	1	0	0	1	1	1	0	2	0

表 1 三碱基组合示意图

2.3 模型选择

2.3.1 分类问题

在分类问题中，本文采用两种分类模型，即 SVM（支持向量机）和 randomForest（随机森林）。其中 SVM 基于线性分类模型进行分类，而 randomForest 基于树模型进行分类，并且两个模型均采用 10X 交叉验证对模型进行评价。

2.3.2 回归问题

在回归问题中，为防止数据回归中的维数灾难，本文采取引入惩罚系数的回

归，即岭回归和 lasso 回归。并且预测模型采用 10X 交叉验证进行模型评估。

2.4 模型评估

2.4.1 交叉验证

交叉验证是指在给定的建模样本中，拿出大部分样本进行建模型，留小部分样本用刚建立的模型进行预测，并求这小部分样本的预测误差，记录它们的平方加和。这个过程一直进行，直到所有的样本都被预报了一次而且仅被预报一次。常用于当数据集较小的建模当中。

交叉验证的基本思想是把在某种意义下将原始数据(dataset)进行分组,一部分做为训练集(train set),另一部分做为验证集(validation set or test set),首先用训练集对分类器进行训练,再利用验证集来测试训练得到的模型(model),以此来做为评价分类器的性能指标。

2.4.2 分类模型评估

ROC 曲线的全称是 Receiver Operating Characteristic Curve，中文名字叫“受试者工作特征曲线”，该曲线的横坐标为假阳性率（False Positive Rate, FPR），N 是真实负样本的个数，FP 是 N 个负样本中被分类器预测为正样本的个数。纵坐标为真阳性率（True Positive Rate, TPR），P 是真实正样本的个数，TP 是 P 个正样本中被分类器预测为正样本的个数。

AUC 定义为 ROC 曲线下围成的面积，通常根据 AUC 数值的大小来判断分类模型的好坏：

- AUC = 1，是完美分类器。
- AUC = [0.85, 0.95], 效果很好

- AUC = [0.7, 0.85], 效果一般
- AUC = [0.5, 0.7],效果较低，但用于预测股票已经很不错了
- AUC = 0.5，跟随机猜测一样，模型没有预测价值。
- AUC < 0.5，比随机猜测还差；但只要总是反预测而行，就优于随机猜测。

2.4.3 回归模型评价

回归问题，本文采用传统 k-mer 的方式，统计每一个子序列种类的个数，并且把出现频率最高的（包括最高的 8 个，6 个，4 个）子序列提取出来作为特征值。

回归问题的评价主要有两个评价标准：1.R^2，用于评价模型拟合程度；2.MES，即真实值与预测值的误差，用于评价模型的预测效果。

3. 结果

3.1 分类问题

在分类问题中，本文再用两个二分类分类器，即 SVM 和 randomForest（图 1，图 2 分别为 SVM 和 randomForest 的 ROC 曲线），并且利用 10X 交叉验证评价模型。在该问题中，利用 CTA,GAC,GTC,TGG 这四个指标进行预测的结果为：SVM 模型进行预测的平均 AUC 达到 0.817，randomForest 模型进行预测的平均 AUC 达到 0.821。

而利用 AAA, AAC, AAG, AAT 这四个分类指标的结果为：SVM 模型进行预测的平均 AUC 达到 0.725，randomForest 模型进行预测的平均 AUC 达到 0.819。

最后利用 TAA,AGA,CTC,GGA 作为分类指标的结果为：SVM 模型进行预测的平均 AUC 达到 0.817，randomForest 模型进行预测的平均 AUC 达到 0.821。

Model	CTA GAC GTC TGG	AAA AAC AAG AAT	TAA AGA CTC GAA
SVM + AUC	0.817 ± 0.004	0.725 ± 0.042	0.873 ± 0.088
RandomForest + AUC	0.821 ± 0.005	0.819 ± 0.086	0.873 ± 0.125

表 2 各类指标基于 SVM 和 randomForest 分类器的 AUC 值

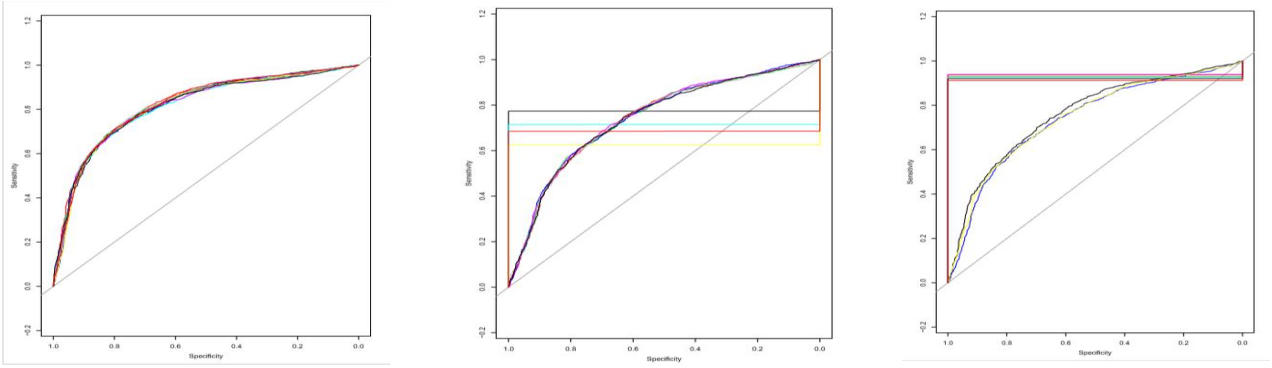

图 1　从左到右分别是指标 CTA GAC GTC TGG，AAA AAC AAG AAT 和 TAA AGA CTC GAA 的 SVM 分类模型 ROC 曲线

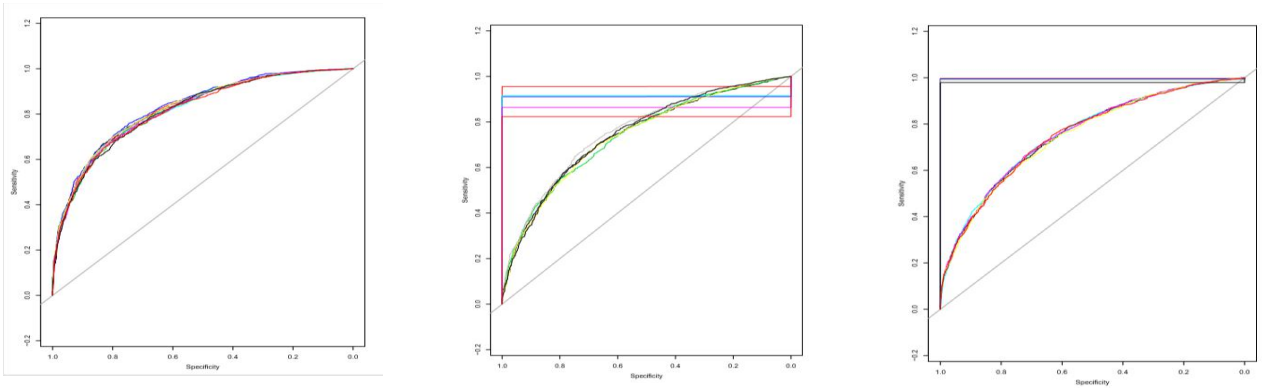

图 2 从左到右分别是指标 CTA GAC GTC TGG，AAA AAC AAG AAT 和 TAA AGA CTC GAA 的 randomForest 分类模型 ROC 曲线

3.2 回归问题

在回归问题中，本文采用岭回归和 lasso 来进行建模，并且利用 r 和 MSE 来进行模型的评价。但是从建模效果来看，利用岭回归和 lasso 都不是太好（表 3）。由图 3 可以看到，建模时的响应变量大部分都位于-2.5 和 2.5 之间，不利于进行线性回归计算，因此造成的误差也比较大，其中 r 值波动较大，MSE 数值较大，预测不太准确。

Model	4 指标	6 指标	8 指标
Lasso + r	0.005 ± 0.005	-0.003 ± 0.004	-0.011 ± 0.015
Lasso + MSE	1.105 ± 0.153	1.085 ± 0.151	1.078 ± 0.224
Ridge regression + r	-0.004 ± 0.004	-0.005 ± 0.005	-0.002 ± 0.003
Ridge regression + MSE	1.070 ± 0.161	0.916 ± 0.161	1.065 ± 0.139

表 3 各指标在岭回归和 lasso 中 r 和 MSE

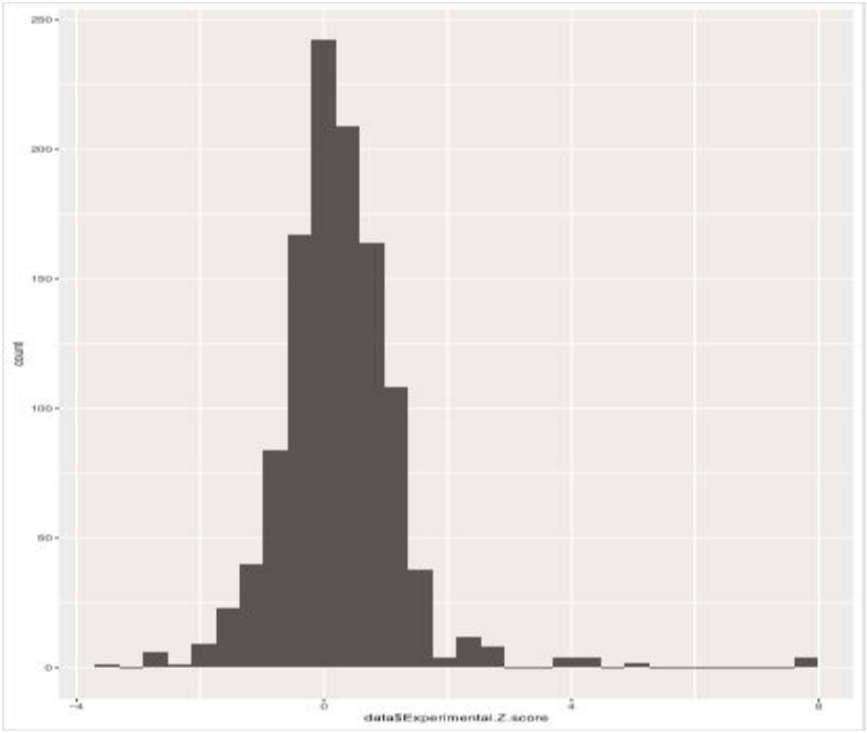

图 3 z-score 分布直方图

4. 讨论

本文利用相邻三碱基的组合来作为分类问题的特征值，从分类模型的效果来看，两类模型的 AUC 值都在 0.8 左右，预测分类比较好。

而在回归问题中，利用 6-mer 的方式进行特征值提取，分别用子序列种类最

高的 8 个（6 个，4 个）子序列作为特征值，采用岭回归和 lasso 来进行建模，并且利用 r 和 MSE 来进行模型的评价。但是从建模效果来看，利用岭回归和 lasso 都不是太好，r 值较小，且负数较多，预测值与真实值的误差较大，原因可能是响应变量分布集中于-2.5 到 2.5 之间，而 R 包 glmnet 是基于线性模型进行建模，因此大部分数据落在-2.5 到 2.5 之间会影响建模，所以需要补充数据，或者重新实验来达到最终的目的。

参考文献

[1] 蔡春,苗丽峰,邓乃扬“DNA 序列特征提取方法探究”北京联合大　学学报 第 22 卷 74 页.

[2] 陈丽萍,乔元华“SVM 方法在 DNA 序列识别中的应用”中国现场统计研究会第十一届学术年会论文集 卷 1.

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

[4] 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.

[5] 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.

[6] Friedman, J., Hastie, T. and Tibshirani, R. (2008) Regularization Paths for Generalized Linear Models via Coordinate DescentJournal of Statistical Software, Vol. 33(1), 1-22 Feb 2010.