大豆科学 Soybean Science
http://ddkx.haasep.cn
2022,41(3):337-344
DOI:10.11861/j.issn.1000-9841.2022.03.0337

四种机器学习算法预测大豆蛋白质定位对比研究

李佳楠[1,2],高兴泉[2],李　卓[1],滕小华[1],黄　斌[1],张继成[3],唐　友[1,2]

(1.吉林农业科技学院 电气与信息工程学院,吉林 吉林 132101;2.吉林化工学院 信息与控制工程学院,吉林 吉林 132000;3.东北农业大学电气与信息工程学院,黑龙江 哈尔滨 150030)

摘　要:为探索不同缺失程度大豆蛋白质亚细胞定位预测的有效方法,提升大豆蛋白质亚细胞定位预测能力,本研究以1万条已知亚细胞定位位置的大豆蛋白质序列数据为研究对象,进行5%、10%、15%、20%和30%不同缺失比例完全随机缺失,分别运用SVM算法、朴素贝叶斯算法和随机森林算法和决策树4种机器学习算法预测缺失序列的亚细胞位置,对原始位置和预测后的位置进行相关性分析,对比分析不同算法的准确性和性能。结果显示:随机森林算法预测的准确率最高;朴素贝叶斯算法的运行速度最快;朴素贝叶斯算法的运行内存最小。在不考虑运行时间和运行内存因素,且对预测的准确率要求较高的情况下,随机森林算法的预测效果要优于另外3种算法;同种情况下,若对运行内存要求较高时,可优先考虑朴素贝叶斯算法。结果说明不同机器学习方法在不同缺失程度的预测需求下的适用性,可应用于大豆蛋白质数据的定位预测。

关键词:支持向量机算法;朴素贝叶斯算法;决策树算法;随机森林算法;大豆蛋白质;完全随机缺失;序列位置预测

Comparative Study of Four Machine Learning Algorithms for Soybean Protein Localization Predicting

LI Jia-nan[1,2], GAO Xing-quan[2], LI Zhuo[1], TENG Xiao-hua[1], HUANG Bin[1], ZHANG Ji-cheng[3], TANG You[1,2]

(1. Electrical and Information Engineering College, Jilin Agricultural Science and Technology University, Jilin 132101, China; 2. School of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin 132000, China; 3. College of Electronic and Information, Northeast Agricultural University, Harbin 150030, China)

Abstract: In order to explore an effective method for predicting the subcellular localization of soybean protein with different degrees of deletion, and improve the prediction ability of soybean protein subcellular localization, this study took 10 000 soybean protein sequence data with known subcellular localization positions as the research object, and carried out 5%, 10%, 15%, 20% and 30% sequences missing at random. Four machine learning methods, namely SVM algorithm, Naive Bayes algorithm, Random Forest algorithm and Decision Tree algorithm, were used to predict the subcellular position of the missing sequence. Correlation analysis was performed between the original position and the predicted position, and the accuracy and performance of different algorithms were compared and analyzed. The results showed that the prediction accuracy of Random Forest algorithm was the highest, the running speed of Naive Bayes algorithm was the fastest, and the running memory of Naive Bayes algorithm was the smallest. When the running time and running memory factors were not considered, and the prediction accuracy was high, the prediction effect of the random forest algorithm was better than the other three algorithms. In the same situation, if the running memory requirements are high, the Naive Bayes algorithm may be preferred. The results show the applicability of different machine learning methods under the prediction requirements of different degrees of missingness, and can be applied to the localization prediction of soybean protein data.

Keywords: Support Vector Machines algorithm; Naive Bayesian algorithm; Decision Tree algorithm; Random Forest algorithm; soybean protein; completely random missing; sequence position prediction

蛋白质参与生命体的繁殖、生长、再发育的整个过程,而蛋白质又是保证整个细胞系统正常运行、高度有序的重要保证,在整个生命过程中有着举足轻重的作用[1]。大多数蛋白质只能在细胞中的一个特定位置(如细胞壁、细胞膜)发挥作用,然而一些其他的蛋白质可以在细胞中的几个位置发挥作用[2]。一个蛋白质如果要正常发挥功能,必须处于细胞中的一个或几个特定的位置上,否则该蛋白质就会失效[3]。在大豆中,液泡膜转运蛋白的非正常定位会导致大豆无法参与并调节细胞的正常代谢,无法维持细胞内环境稳定,大豆适应环境变化和生存能力会显著下降[4];蛋白质在叶绿体内的正常反应直接影响作物的产量[5];与叶绿素有关的蛋白质在大豆种皮颜色的调控过程中起着重要的作用;线粒体蛋白定位异常则与能量代谢缺陷直接相关[6];蛋白质在其他各亚细胞位置上的异常定位

收稿日期:2021-10-22
基金项目:吉林省特色高水平学科新兴交叉学科"数字农业"(2018);吉林省智慧农业工程研究中心项目(2016);国家自然科学基金(31801441)。
第一作者:李佳楠(1995—),男,硕士研究生,主要从事生物信息学研究。E-mail:rate_ljn@163.com。
通讯作者:唐友(1979—),男,博士,教授,高级工程师,主要从事生物信息学及农业信息化研究。E-mail:tangyou@neau.edu.cn。

会产生各种不同的疾病。

现代生物学发展迅速,生物蛋白质数据库中的蛋白质序列数量呈爆炸式增长。面对大量的蛋白质序列,依靠传统的试验方法对蛋白质进行亚细胞定位已经不能满足现阶段低成本、高效率的需求,且试验重复性也较差[7]。机器学习的方法能很好地解决这一需求。利用机器学习算法对不同蛋白质数据集进行亚细胞定位预测的研究有很多,但大多数研究仅局限于细菌、病毒、凋亡蛋白等组成的固定数据集,此类数据集一般由几百条数据组成。查阅相关资料发现,对于数据量大、预测位置多的植物蛋白亚细胞定位的研究较少,现阶段在大豆蛋白质亚细胞定位预测方面尚缺少一种较为适用的方法。支持向量机(Support Vector Machines,SVM)算法是通过将样本集合映射到高维向量中,自动找到对分类集合有较好分类效果的支持向量,使两个样本集合之间的距离达到最大,从而达到比较好的分类效果[8]。朴素贝叶斯(Naive Bayes, NB)算法建立在贝叶斯统计学和贝叶斯决策理论基础上,是一种将先验知识和数据进行综合的理想表达模式,具有模型可解释、精度高等优点[9]。决策树(Decision Tree, DT)算法是一种非参数数据挖掘方法,若目标变量是分类的,则称之为分类树[10]。随机森林(Random Forest, RF)算法能充分利用数据本身的信息,汇总百棵决策树的预测值得到最终结果,对异常值和噪声具有很好的容忍度,且不容易出现过拟合,能有效分析高维复杂的数据。

Top-n-gram 可被看作一个新的基于蛋白质图谱的结构块,包括了在频率图谱中所获得的进化信号信息。由 PSI-BLAST 输出的最多序列经过比对算出的频率分布通过组合,每个氨基酸频率分布中的 n 个最常见氨基酸被转换成 Top-n-gram,并利用每个 Top-n-gram 的出现次数,将蛋白质序列变换为固定维的特征向量[11]。基于距离的 Top-n-gram 是一种基于轮廓的方法,它考虑了 Top-n-gram 对之间的距离。替换蛋白质序列中的所有氨基酸可以表示为 Top-n-gram 序列而不是氨基酸序列。

本研究对多物种蛋白质序列进行不同比例完全随机缺失(Missing Completely At Random,MCAR)处理,采用机器学习的支持向量机算法、朴素贝叶斯算法、随机森林算法和决策树算法,对混合的缺失蛋白质序列在细胞中的位置进行预测,并记录算法运行的时间、内存、位置,对比准确性以及 cor(correlation coefficient)准确性。使用基于距离的 Top-n-gram 方法,通过考虑蛋白质序列中 Top-n-gram 对的相对位置信息,扩展原有的基于 Top-n-gram 的特征向量。研究旨在为实现在不同环境要求以及运算性能的情况下,预测蛋白质序列在细胞中的位置提供科学参考。

1 材料与方法

1.1 材料

本研究的大豆蛋白质序列组数据主要下载自 UniProt(https://www. uniprot. org/uniprot),试验数据主要包含 16 922 条蛋白质序列,包括 11 个大豆亚细胞位置(高尔基体、过氧化物酶体、内质网、细胞壁、细胞核、细胞质、细胞膜、叶绿体、线粒体、液泡和质体)。其中位于高尔基体中的蛋白质 578 条、过氧化物酶体中的蛋白质 397 条、内质网中的蛋白质 621 条、细胞壁中的蛋白质 408 条、细胞核中的蛋白质 10 773 条、细胞膜中的蛋白质 1 559 条、细胞质中的蛋白质 1 257 条、线粒体中的蛋白质 70 条、叶绿体中的蛋白质 644 条、液泡中的蛋白质 153 条、质体中的蛋白质 462 条。

试验硬件环境为 CPU 八核处理器、16 G 内存服务器[12]。

1.2 试验设计

蛋白质序列的特征提取及数据分组:因原始蛋白质序列数据无法被机器学习算法识别,将由氨基酸组成的字母序列转化为数字序列。

试验数据分为 12 组:分别存在于 11 个亚细胞位置的蛋白质序列数据为 11 组,另一组为包含 11 个不同位置的全部蛋白质序列数据。

缺失处理:利用 simFrame 软件包对 12 组数据进行缺失处理,缺失机制为完全随机缺失,缺失比例分别为 5%、10%、15%、20% 和 30%。

缺失处理和预测:使用 R 语言进行数据缺失处理和预测。采用支持向量机算法、贝叶斯算法、随机森林算法和决策树算法预测不同比例缺失的蛋白质位置,并对比分析预测所用的内存、时间、位置对比准确率和 cor 准确率。

具体试验流程图如图 1 所示。

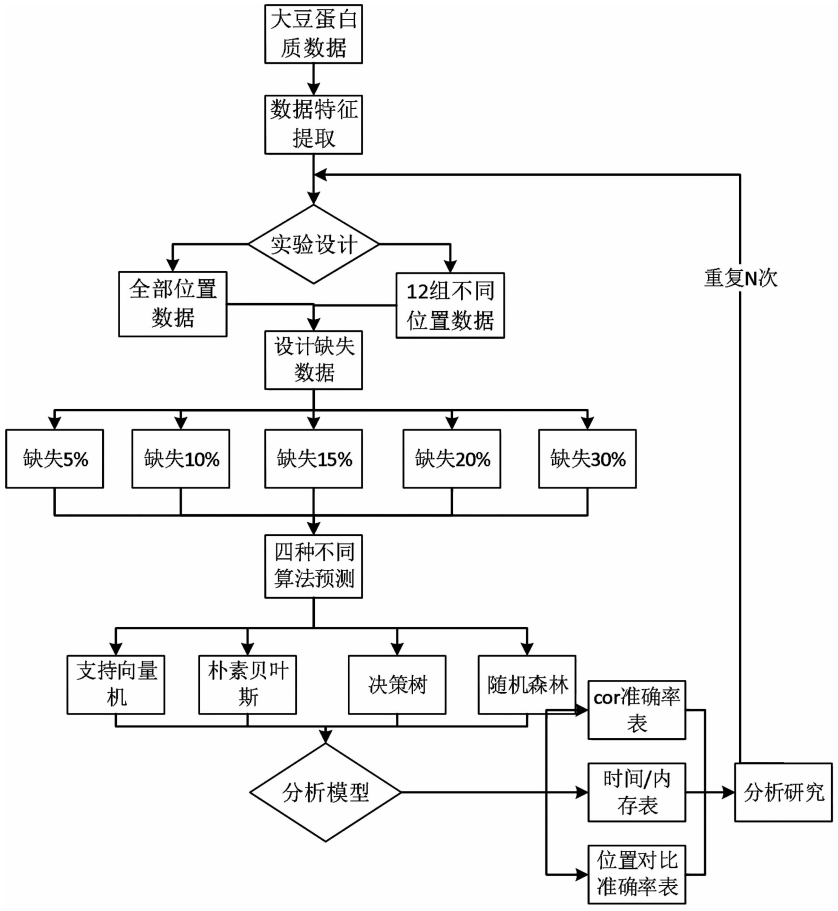


图 1　试验流程图

Fig. 1　The experimental flowchart

1.3　方法

1.3.1　*蛋白质序列特征提取*　选取 Top-1-gram 构造基于距离的 Top-n-gram 特征向量,以降低特征向量的维数和计算量。通过计算一定距离阈值内全部可能的 Top-n-gram 对的出现次数来计算所提出的特征向量[13]。特征向量的维数为:$P = 20 + 20 \times 20 \times DMAX$。其中,20 代表氨基酸字母表的大小,*DMAX* 是距离阈值,表示 Top-1-gram 对之间的最大距离。

1.3.2　*数据随机缺失*　采用 simFrame 软件包和 sample 函数,分别对由大豆的 11 个亚细胞位置(高尔基体、过氧化物酶体、内质网、细胞壁、细胞核、细胞质、细胞膜、叶绿体、线粒体、液泡和质体)组成的特征进行蛋白质数据提取,对提取出的 1 万多条数据进行缺失处理。按照 5%、10%、15%、20%、30% 缺失比例进行完全随机缺失。

1.3.3　*机器学习算法分析*　支持向量机算法:使给出的样本集$(x_i, y_i), i = 1,2,\cdots,n, x \in R^d, y \in \{+1, -1\}$ 满足$y_i[(\omega^t x_i) + b] - 1 \geqslant 0, i = 1,\cdots,n$,找到最大边距的最优超平面分离 y。针对于高维空间的分类问题,引入核函数 $k(x_i, y_i)$,使支持向量机模型的两类样本间分类间隔最大。得到最优分类面决策函数:$f(x) = sgn\{(\omega, x) + b\} = sgn[\sum_{xi \subseteq sv} \alpha_i y_i k(x_i, y_i) + b]$。

朴素贝叶斯算法:根据贝叶斯定理进行推导 $P(y_i \mid x) = \dfrac{P(x \mid y_i)P(y_i)}{P(x)}$,其中各个特征是独立的,得到 $P(a_1 \mid y_1), P(a_2 \mid y_1), \cdots, P(a_m \mid y_1); \cdots; p(a_1 \mid y_n), P(a_2 \mid y_n), \cdots, P(a_m \mid y_n)$。

决策树算法:设置一个训练集 D,假设第 K 类样本所占的比例为$p_k(k = 1,2,\cdots, \mid y \mid)$,其纯度基尼值度量为 $\mathrm{Gini}(D) = \sum_{k=1}^{\mid y \mid} \sum_{k' \neq k} P_k P_{k'} = 1 - \sum_{k=1}^{\mid y \mid} P_k^2$,设属性 a 有 v 个可能值,其中基尼指数定义为 $\mathrm{Gini\_index}(D,a) = \sum_{v=1}^{v} \frac{\mid D^v \mid}{\mid D \mid} \mathrm{Gini}(D^v)$。基尼指数最小的最优划分应该满足:

$a_* = argmin Gini\_index(D,a)$。

随机森林算法:首先对蛋白质亚细胞位置进行 k 轮训练,得到$\{h_1(X), h_2(X), \cdots, h_k(X)\}$[14],最终得到的分类决策 $H(x) = \arg\max_r \sum_{i=1}^{k} I(h_i(x) = Y)$,其中 h_i 表示单个决策树,Y 表示输出变量,根据投票多数的方式决出最终的分类。

1.3.4 预测结果准确性分析 通过对预测后的位置与原位置进行比对,得出位置对比准确率。利用相关系数对原位置和预测位置进行对比分析,判断准确性。

相关系数是用来表示两个变量之间相关密切程度的统计指标,可用公式表述为 $r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var[X]Var[Y]}}$,其中 Cov($X$,$Y$)表示 X 和 Y 的协方差,Var[X]表示 X 的方差,Var[Y]表示 Y 的方差。计算相关系数并采用 R 语言中 cor. test( )函数进行对比。

1.3.5 算法性能分析 统计并对比各算法预测不同位置和比例的数据所用的时间和内存。利用 Python 中的 time 和 psutil 模块,调用 time. clock( )方法计算程序的运行时间,调用 psu-til. process( )方法获取运行程序所占用的内存值。

2 结果与分析

2.1 各算法预测结果分析

本研究中 1 万条大豆蛋白质数据在 4 种不同的机器学习算法下的不同缺失率下预测的相关系数如表 1 所示,对全部 12 个亚细胞位置进行整体预测时,支持向量机算法预测的相关系数高于其他 3 种算法,其次是随机森林算法,朴素贝叶斯算法和决策树算法的相关系数相对较低。对 12 个不同亚细胞位置进行单独预测时,在高尔基体和细胞壁两个位置上预测出的相关系数明显高于其他 10 个亚细胞位置。

表 1 各算法预测相关系数结果

Table 1 The correlation coefficient results of different algorithms

亚细胞位置 Subcellular location	缺失比 Deletion ratio	支持向量机 SVM	朴素贝叶斯 Naive Bayes	决策树 Decision tree	随机森林 Random forest
全部 Global	5%	0.8226	0.3649	0.4280	0.7246
	10%	0.8023	0.3583	0.3799	0.7206
	15%	0.7892	0.4159	0.4472	0.7003
	20%	0.7695	0.3881	0.3914	0.6596
	30%	0.7715	0.3909	0.4019	0.7011
高尔基体 Golgi apparatus	5%	0.7648	0.4443	0.4845	0.6100
	10%	0.7388	0.4276	0.4192	0.7107
	15%	0.6932	0.4486	0.3968	0.6675
	20%	0.7101	0.4619	0.3733	0.6849
	30%	0.7061	0.4187	0.3956	0.6580
过氧化物酶 Peroxidase	5%	0.6164	0.4729	0.3758	0.4237
	10%	0.4738	0.3438	0.3287	0.4478
	15%	0.4474	0.4187	0.4061	0.4482
	20%	0.4141	0.3992	0.2944	0.4497
	30%	0.4314	0.4826	0.4113	0.4237
内质网 Endoplasmic reticulum	5%	0.5656	0.2633	0.3809	0.4042
	10%	0.5830	0.2699	0.1594	0.5929
	15%	0.5715	0.3110	0.0745	0.5080
	20%	0.5414	0.2489	0.2857	0.5031
	30%	0.5341	0.2893	0.0229	0.5028
细胞壁 Cell wall	5%	0.8081	0.7546	0.6351	0.7498
	10%	0.7861	0.7030	0.6580	0.8237
	15%	0.7806	0.7181	0.6537	0.8035
	20%	0.8163	0.7284	0.6443	0.7921
	30%	0.7858	0.7676	0.5320	0.7788
细胞膜 Cell membrane	5%	0.2313	0.1930	0.2005	0.2653
	10%	0.3078	0.2405	0.1534	0.2669
	15%	0.2561	0.2297	0.1765	0.2112
	20%	0.2285	0.2272	0.1102	0.1835
	30%	0.2990	0.2124	0.1494	0.1734
细胞质 Cytoplasm	5%	0.3760	0.3144	0.2837	0.3431
	10%	0.2865	0.2365	0.3115	0.2542
	15%	0.2962	0.3060	0.2466	0.2966
	20%	0.3310	0.2270	0.1029	0.2479
	30%	0.2342	0.2383	0.1457	0.2117

续表 1

亚细胞位置 Subcellular location	缺失比 Deletion ratio	支持向量机 Support Vector Machines	朴素贝叶斯 Naive Bayes	决策树 Decision Tree	随机森林 Random Forest
线粒体 Mitochondria	5%	0.2845	0.2686	0.2102	0.2369
	10%	0.3987	0.1937	0.1220	0.2680
	15%	0.3674	0.2855	0.1608	0.2888
	20%	0.3560	0.2749	0.1168	0.2235
	30%	0.2942	0.3633	0.1069	0.2470
叶绿体 Chloroplast	5%	0.3264	0.2563	0.2111	0.2482
	10%	0.2836	0.3040	0.1442	0.2434
	15%	0.3235	0.2182	0.1020	0.2586
	20%	0.3025	0.2612	0.1310	0.2485
	30%	0.2778	0.2353	0.1040	0.2680
液泡 Vacuole	5%	0.3053	0.2653	0.1823	0.1986
	10%	0.3860	0.2484	0.1351	0.2453
	15%	0.2904	0.2454	0.2168	0.2654
	20%	0.3242	0.2325	0.1992	0.3251
	30%	0.3065	0.2517	0.1349	0.1526
质体 Plastid	5%	0.4530	0.1592	0.2265	0.2106
	10%	0.3877	0.1208	0.1608	0.2435
	15%	0.3099	0.1855	0.1259	0.2142
	20%	0.2735	0.2190	0.1546	0.2174
	30%	0.2363	0.2661	0.1072	0.2337

2.2　不同算法位置准确率对比

2.2.1　*各亚细胞位置蛋白序列整体预测*　经过多次位置对比准确率预测试验,取多次试验结果的平均值以最大程度减小误差。在对 12 个亚细胞位置的蛋白质序列进行整体预测时,5 种不同缺失率下,随机森林算法的预测精准性最高,为 78.7%,均高于其他 3 种算法,其次是支持向量机算法,另外两种算法的预测精准性相对较低。此外,随着缺失率的逐步升高,4 种算法的精准性也在逐步降低(图 2)。

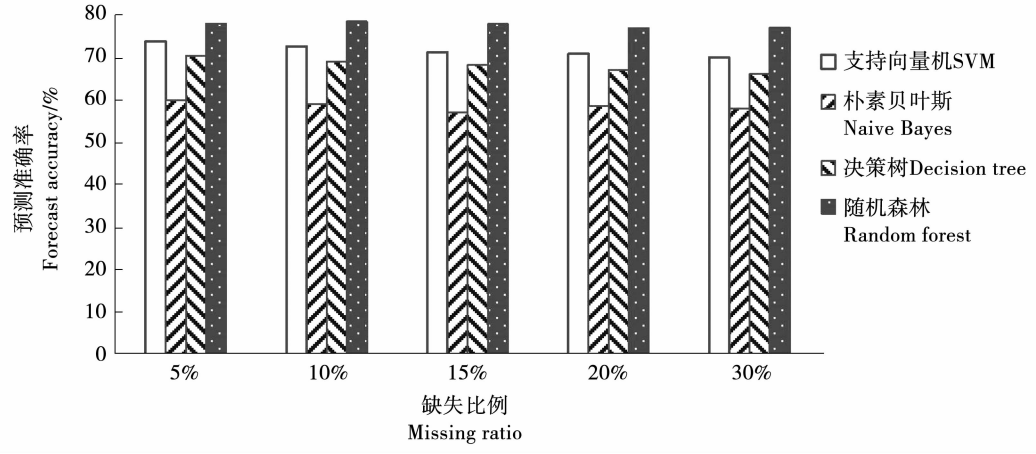
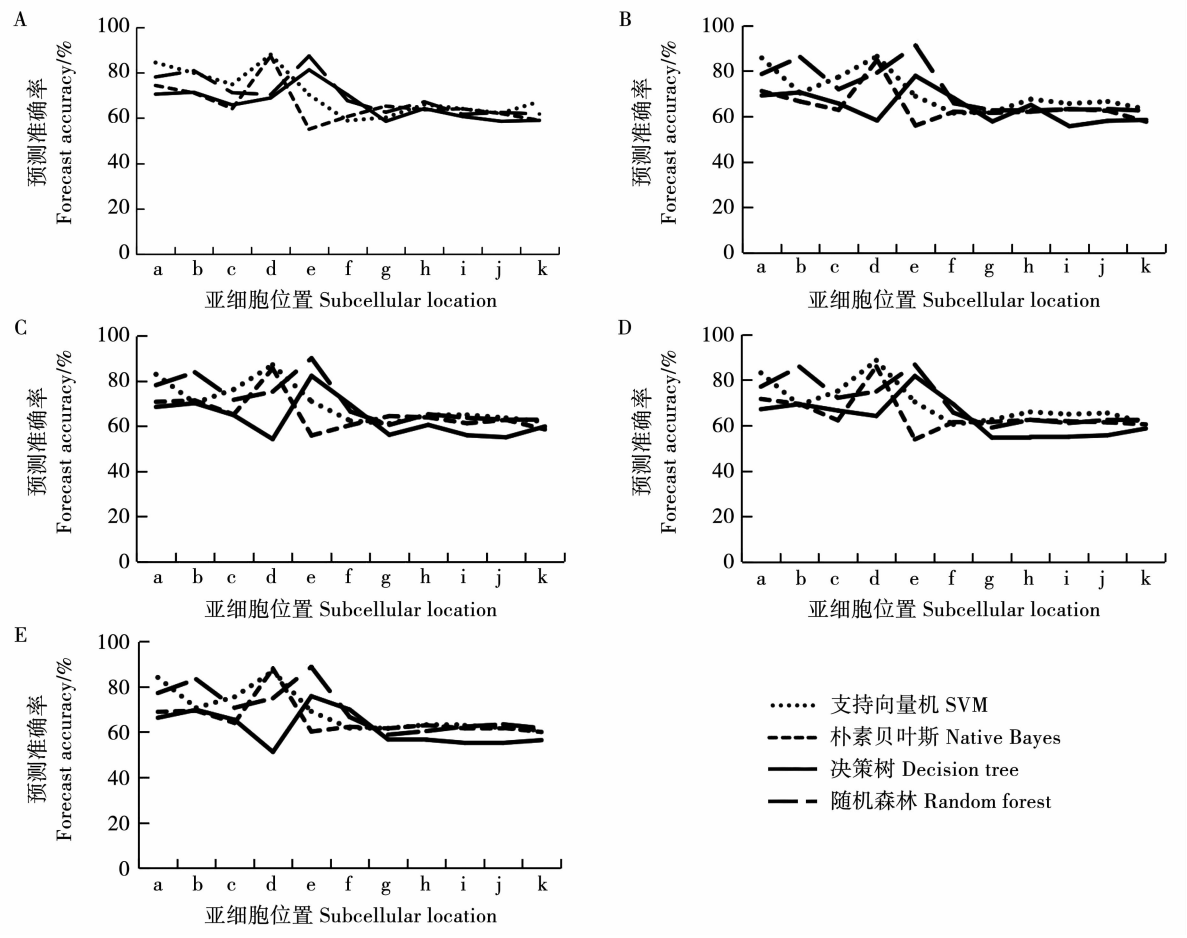


图 2　不同算法整体预测亚细胞位置蛋白序列位置准确率对比

Fig. 2　The accuracy comparison of different algorithms in global predicting subcellular location protein sequences

2.2.2　*不同亚细胞位置蛋白序列单独预测*　使用 4 种不同的算法对 12 个亚细胞位置上的蛋白质序列进行单独预测,4 种不同的算法在细胞壁和高尔基体上的预测精准度相对较高,在其他 10 个亚细胞位置上的精准度相对较低。使用随机森林算法在大豆细胞壁上进行预测时,预测准确率达到 90%,其他 3 种算法在细胞壁上的预测准确率也超出 80%。从整体上来看,随机森林算法和支持向量机算法对各个亚细胞位置进行单独预测的预测精准度依然相对较高(图 3)。



注:a. 高尔基体;b. 过氧化物酶;c. 内质网;d. 细胞壁;e. 细胞核;f. 细胞膜;g. 细胞质;h. 线粒体;i. 叶绿体;j. 液泡;k. 质体。

Note: a. Golgi apparatus; b. Peroxidase; c. Endoplasmic reticulum; d. Cell wall; e. Nucleus; f. Cell membrane; g. Cytoplasm; h. Mitochondria; i. Chloroplast; j. Vacuole; k. Plastid.

图 3 不同算法单独预测各亚细胞位置蛋白序列准确率对比

Fig. 3 The accuracy comparison of different algorithms in predicting protein sequences of each subcellular location

2.3 不同算法性能对比

由图 4 所示,1 万多条大豆蛋白质数据在时间维度上,朴素贝叶斯算法运行时间最快,其次是决策树算法,接着是随机森林算法,SVM 算法的运行时间最慢。由图 5 所示,在内存维度上,随机森林算法占用内存最大,随着缺失率的增加,占用内存逐步降低;SVM 算法和决策树算法占用内存相对较低,缺失率增加后内存占比逐步降低;朴素贝叶斯算法在 4 种算法中内存占比最少。总体来看,除朴素贝叶斯算法外,随着缺失率的增加,其他 3 种算法的运行时间和运行内存都呈线性下降趋势,可见缺失率会对算法的时间和内存维度产生较为明显的影响。

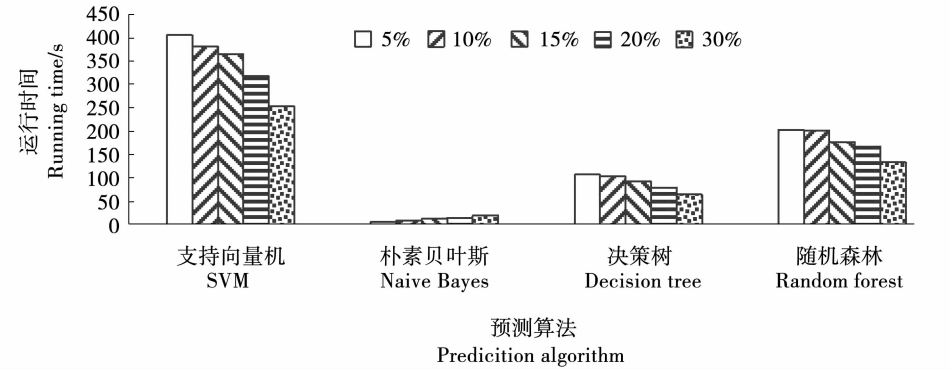


图 4 4 种预测算法在同一运行环境不同缺失率下时间对比

Fig. 4 The time comparison of four prediction algorithms in the same running environment and different missing rate

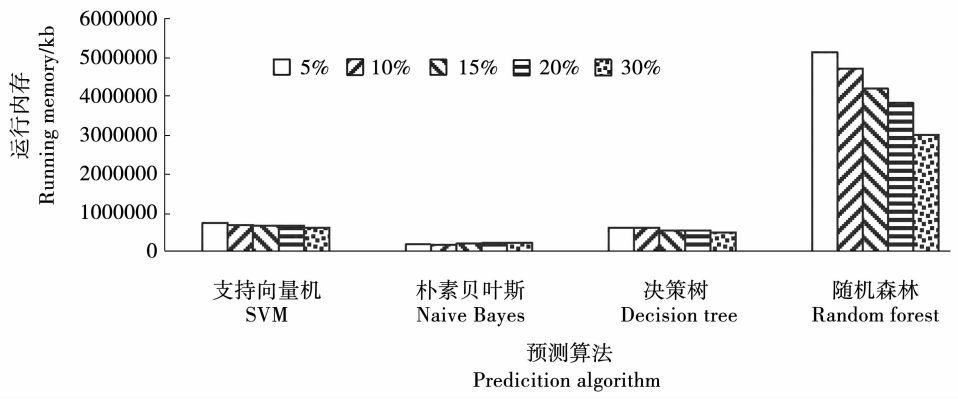


图 5　4 种预测算法在同一运行环境不同缺失率下内存对比

Fig. 5　The memory comparison of four prediction algorithms in the same running environment with different missing rate

3　讨论

经过下载的大豆蛋白质序列数据,其内部结构为字母型数据,不利于机器学习算法进行预测,因此需要将蛋白质序列数据转换为数值型数据,此类方法叫做蛋白质的特征提取[15]。

利用机器学习方法对蛋白质数据进行预测的过程中,属性之间的相关性对朴素贝叶斯算法的性能影响较大,在属性个数较多或属性之间的相关性较大时,会影响算法预测的准确性,因此考虑部分关联性并进行适度的改进是朴素贝叶斯算法的进一步研究方向[16]。构造核函数 $k(x_i,y_i)$ 是支持向量机模型预测的关键,不同的核函数会影响蛋白质亚细胞定位预测的精准性。随机森林算法对蛋白质序列位置预测的准确性较高,但在实际应用中并不普遍,最主要的原因在于随机森林算法运行时间和内存占比相较于其他方法较长、较大,因此选择随机森林预测算法时可选用性能较高的服务器,与此同时改进程序的并行计算,同时可引入 CPU 计算,提升资源利用率[17]。

此外,对特征提取后蛋白质序列数据进行数据预处理操作会对预测的结果产生一定的影响,本研究中使用了一种基于距离的 Top-n-gram 方法,通过考虑蛋白质序列中 Top-n-gram 对的相对位置信息,扩展原有的基于 Top-n-gram 的特征向量,其中包括了在频率图谱中所获得的进化信号信息。对于不同的蛋白质数据,采用多特征融合的方法将几个模型结合起来可能会得到显著的性能,但如果里面包含不合适的模型,结果会适得其反,因为过拟合影响模型的计算。因此如何恰当的进行预处理需要进一步完善和研究。

4　结论

本研究以 1 万条大豆蛋白质数据作为研究对象,将原始蛋白质数据通过特征提取之后由字母型数据转换成数值型数据,再通过对比基于机器学习的支持向量机、朴素贝叶斯、决策树、随机森林 4 种算法的相关系数准确性和位置对比准确性对 4 种算法进行验证。如果不考虑运行时间和运行内存的因素,对预测的准确性要求较高的情况下,随机森林算法是最好的选择。当数据量很大时,如果对算法的运行时间和运行内存的要求较高时,可选择朴素贝叶斯算法。本研究使用的 4 种机器学习算法,对大豆蛋白质序列的预测比较过程对其他农作物的蛋白质序列位置预测具有参考和借鉴意义,为研究者们进行其他农作物蛋白质定位预测研究提供了参考模板,具有一定推广前景。

参考文献

[1] EISENHABER F, BORK P. Wanted: Subcellular localization of proteins based on sequence[J]. Trends in Cell Biology, 1998, 8(4): 169-170.

[2] CHOU K C. Some remarks on predicting multi-label attributes in molecular Biosystems[J]. Molecular BioSystems, 2013, 9(6): 1092-1100.

[3] LUNN E J. Compartmentation in plant metabolism[J]. Journal of Experimental Botany, 2007, 58(1): 35-47.

[4] ENRICO M, MAESHIMA M, EKKEHARD N H. Vacuolar transporters and their essential role in plant metabolism[J]. Journal of Experimental Botany, 2007, 58(1): 83-102.

[5] 白辉, 王宪云, 曹英豪, 等. 水稻叶绿体蛋白质在生长发育过程中的表达研究[J]. 生物化学与生物物理进展, 2010, 37(9): 988-995. (BAI H, WANG X Y, CAO Y H, et al. Expression of

chloroplast proteins in rice during growth and development[J]. Progress in Biochemistry and Biophysics, 2010, 37(9): 988-995.)

[6] 赵丽, 周巧霞, 王拴, 等. 线粒体分裂和融合相关蛋白质的研究进展[J]. 生理学报, 2018, 70(4): 424-432. (ZHAO L, ZHOU Q X, WANG S, et al. Research progress of mitochondrial fission and fusion-related proteins[J]. Acta Physiologica Sinica, 2018,70(4): 424-432.)

[7] CHOU K C, CAI Y D. Using function domain composition and support vector machines for prediction of protein subcellular location[J]. Journal of Biological Chemistry, 2002, 277(48): 45765-45769.

[8] GALAR M, FERNáNDEZ A, BARRENECHEA E, et al. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes[J]. Pattern Recognition, 2011, 44(8): 1761-1776.

[9] MURAKAMI Y, MIZUGUCHI K. Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites[J]. Bioinformatics, 2010, 26(15): 1841-1848.

[10] MABUNI D. A new direct node data splitting technique in decision tree induction[J]. International Journal of Innovative Technology and Exploring Engineering, 2020, 9(7).

[11] MENGTING N, YANJUAN L, CHUNYU W, et al. RFAmyloid: A web server for predicting amyloid proteins[J]. International Journal of Molecular Sciences, 2018, 19(7): 2071.

[12] 唐友, 郑萍, 王嘉博, 等. 对比 Bayesian B 等多种方法的大豆全基因组选择应用研究[J]. 大豆科学, 2018, 37(3): 30-35. (TANG Y, ZHENG P, WANG J B, et al. Application of soybean genome-wide selection by comparing Bayesian B and other methods[J]. Soybean Science, 2018, 37(3): 30-35.)

[13] LIU B, WU H, CHOU K C. Pse-in-One 2.0: An improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences[J]. Natural Science, 2017, 9(4): 67-91.

[14] 李盛. 基于数据挖掘的煤矿微震危害预测实证分析[D]. 昆明: 云南师范大学, 2015. (LI S. Empirical analysis of microseismic hazard prediction in coal mine based on data mining[D]. Kunming: Yunnan Normal University, 2015.)

[15] 未丽, 刘建利. 植物蛋白质亚细胞定位相关研究概述[J]. 植物科学学报, 2021, 39(1): 93-101. (WEI L, LIU J L. An overview of studies related to the subcellular localization of plant proteins[J]. Chinese Journal of Plant Science, 2021, 39(1): 93-101.)

[16] 陈凯. 面向不平衡数据集的朴素贝叶斯文本分类算法改进研究[D]. 哈尔滨: 东北林业大学,2018. (CHEN K. Improvement of Naive Bayesian text classification algorithm for imbalanced data sets[D]. Harbin: Northeast Forestry University, 2018.)

[17] 于合龙, 刘雨帆, 张继成, 等. 基于多种机器学习方法填补大豆基因组缺失的比较研究[J]. 大豆科学, 2021, 40(1): 122-129. (YU H L, LIU Y F, ZHANG J C, et al. A comparative study of filling in the soybean genome deletion based on multiple machine learning methods[J]. Soybean Science, 2021, 40(1): 122-129.)

《大豆科学》正式加入 OSID 开放科学计划

《大豆科学》于 2019 年 8 月 1 日起正式加入 OSID(Open Science Identity)开放科学标识计划。将通过在文章上添加开放科学二维标识码(OSID 码),为读者和作者提供一个与业界同行和专家学术交流的平台,同时提供一系列增值服务,提升论文的科研诚信。

读者可以通过微信扫描论文的 OSID 码,在手机上听论文作者的语音介绍,可以看到论文的重点彩图和实验视频,也可直接与作者进行一对一的交流、关注作者的研究动向等。这些功能有助于读者深入了解该研究的实际状况与实现过程。

作者可以通过专属的 OSID 码对所著论文添加语音,介绍写作背景、动机、趣事以及研究灵感。添加无法在传统印刷出版展示的附加说明,以便更好地展现研究成果,拓展论文的传播方式。同时,通过 OSID 平台每位作者都能拥有所著论文的学术圈和问答,与读者进行交流互动。此外,作者还可以在学术圈发布感兴趣的话题、最新的研究观点、问题征集、学术推荐等,扩大作者自身的影响力,增强与读者的联系。