基于半监督学习的蛋白质相互作用预测模型

安计勇[1,2],闫子骥[1,2]
(1. 中国矿业大学 矿山数字化教育部工程研究中心,江苏 徐州 221000;
2. 中国矿业大学 计算机科学与技术学院,江苏 徐州 221000)

摘 要:基于有监督学习的预测模型在预测过程中存在以下缺陷:一是过分依赖训练集中有标签样本的数量,导致分类精度受有标签样本数量多少的制约;二是其预测分类一次完成,导致大量的无标签样本无法用来修正分类器的预测精度,大量数据信息被浪费,从而影响分类性能。针对以上问题,该文提出一种基于AP聚类与Renyi熵融合的自训练半监督相关向量机分类预测模型。该模型通过AP聚类分析与Renyi熵来共同标记无标签样本的标签类别,筛选置信度高的无标签样本扩充原有训练集进行自训练迭代分类,降低噪声数据对分类器预测精度的影响,构造出了性能最优的基于半监督学习的蛋白质相互作用预测模型。通过在M. musculus、H. pylori和H. sapiens蛋白质相互作用数据集上的实验验证,证明了提出的半监督分类预测模型的有效性。

关键词:相关向量机;半监督学习;自训练;AP聚类;Renyi熵;分类预测
中图分类号:TP311.13 **文献标识码**:A **文章编号**:1673-629X(2021)07-0007-06
doi:10.3969/j.issn.1673-629X.2021.07.002

Prediction Model of Protein-protein Interactions Based on Semi-supervised Learning

AN Ji-yong[1,2], YAN Zi-ji[1,2]
(1. Engineering Research Center of Mine Digitalization of Ministry of Education,
China University of Mining and Technology, Xuzhou 221000, China;
2. School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221000, China)

Abstract: In the prediction process, the disadvantage of the prediction model based on supervised learning are as follows: firstly, due to over-dependence on the number of labeled samples in the training set, the classification accuracy is limited by the number of labeled samples. Secondly, its prediction classification is completed once, resulting in a large number of unlabeled samples that cannot be used to amend the prediction accuracy of the classifier, and a large amount of data information is wasted, thus affecting the classification performance. In view of the above problems, we propose a new self-training semi-supervised classification prediction model of RVM based on AP clustering and Renyi entropy fusion. This model can greatly reduce the influence of noise data on the prediction accuracy of classifier by using AP clustering and Renyi entropy fusion to assign labels for unlabeled samples. The semi-supervised classifier with optimal performance was constructed through adding the unlabeled samples with high degree of confidence to the training set and executing the self-training iteration classification with the expanded training set. It is demonstrated that the proposed prediction model is effective by experimenting validation on M. musculus, H. pylori and H. sapiens datasets.

Key words: relevance vector machine; semi-supervised learning; self-training; AP clustering; Renyi entropy; classification prediction

0 引 言

在机器学习领域中,根据训练集中有标签和无标签样本的数量,可以将机器学习分为有监督学习[1]、无监督学习[2]和半监督学习[3-4]。有监督学习训练集只包含有标签样本,根据有标签样本集训练模型,用训练好的模型预测无标签样本的标签类别;无监督学习训练集只包含无标签样本,根据样本间的内在联系,判定样本的标签类别。有监督学习要得到好的训练模型,通常需要足够多的有标签样本数据,但实际上有标签样本数据的获取通常会耗费大量的人力、物力及财力,需要付出昂贵的成本。但现实中大量存在的无标签样本数据,则相对容易获取。将有标签样本和无标签样

收稿日期:2020-07-03 **修回日期**:2020-11-05
基金项目:国家自然科学基金面上项目(61572506);中央高校基本科研业务费(学科前沿专项)(2019XKQYMS88)
作者简介:安计勇(1975-),男,博士研究生,副教授,研究方向为机器学习、数据挖掘。

本有效结合来获取更好的分类效果,是当前机器学习领域迫切的研究内容。由于半监督学习的训练集不仅包含有标签样本,而且包含无标签样本,学习过程中能够同时利用少量的有标签样本与大量的无标签样本,能够有效地融合两者所蕴含的信息,因此现实中,针对有标签样本较少,无标签样本大量存在的数据集的分类,通常采用半监督分类算法。

该文提出的半监督预测模型主要针对如下问题:

(1)基于有监督学习的预测模型在预测过程中存在以下缺陷:一是过分依赖训练集中有标签样本的数量,导致分类精度受有标签样本数量多少的制约;二是其预测分类一次完成,导致大量的无标签样本无法用来修正分类器的预测精度,大量数据信息被浪费,从而影响分类性能。

(2)采用生物实验方法获取有标签的蛋白质相互作用样本既耗时、费力且成本较高,因此,在蛋白质相互作用预测领域同样存在有标签数据少且获取难,无标签数据容易获取的现实问题,如 M. musculus、H. pylori 和 H. sapiens 三个数据集。

基于以上分析,该文提出一种基于 AP 聚类与 Renyi 熵融合的自训练半监督相关向量机分类预测模型。该模型通过 AP 聚类分析与 Renyi 熵来共同标记无标签样本的标签类别,筛选置信度高的无标签样本扩充原有训练集进行自训练迭代分类,降低了噪声数据对分类器预测精度的影响,构造出了性能最优的基于半监督学习的蛋白质相互作用分类预测模型。

1 模型相关理论介绍

1.1 相关向量机

相关向量机(relevance vector machine,RVM)以贝叶斯概率为框架,是一种基于稀疏贝叶斯理论的核函数学习方法[5-6],其训练是在贝叶斯框架下进行的,在先验参数的结构下基于主动相关决策理论(automatic relevance determination,ARD)来移除不相关的点,从而获得稀疏化的模型[7-8]。

RVM 分类算法的数学模型基本形式如下:

$$y(x;w)=\sum_{i=1}^{M}w_i k(x,x_i)+w=\varnothing(x)w \tag{1}$$

在分类算法中,将公式(1)中的 $y(x)$ 通过 Logistic Sigmoid 函数 $\sigma(y)=\frac{1}{1+e^{-y}}$ 转换为线性模型,则似然估计概率分布为:

$$P(t\mid w)=\prod_{i=1}^{m}\sigma[y(x_i)]^{t_i}\{1-\sigma[y(x_i)]\}^{1-t_i} \tag{2}$$

$$P(w\mid t,\mu)\propto P(t\mid w)P(w\mid a) \tag{3}$$

由于分类算法中 $P(t\mid w)$ 不是标准的正态分布,所以无法求解定积分,但是可以用拉普拉斯方法近似地逼近:

固定 μ,求出 w 的最大值:

$$\begin{aligned}w_{MP}&=\arg\max_w P(w\mid t,\mu)=\\&\arg\max_w\frac{P(t\mid w)P(w\mid\mu)P(\mu)}{P(\mu,t)}=\\&\arg\max_w P(t\mid w)P(w\mid\mu)=\\&\arg\max_w\log\{P(t\mid w)P(w\mid\mu)\}\end{aligned} \tag{4}$$

$$\log\{P(t\mid w)P(w\mid\mu)\}=\sum_{i=1}^{N}[t_i\log y_i+(1-t_i)\log(1-y_i)]-\frac{1}{2}\boldsymbol{w}^{T}\boldsymbol{A}\boldsymbol{w} \tag{5}$$

上式中,$y_i=\sigma\{y(x_n;w)\}$,$\boldsymbol{A}=\mathrm{diag}(\mu_0,\mu_1,\cdots,\mu_N)$。

(1)采用 Laplace 方法,对公式(5)两次求导可以得到如下公式:

$$g=\nabla_w\log\{P(t\mid w)P(w\mid\mu)\}=\varnothing^{T}(t-y)-\boldsymbol{A}\boldsymbol{w} \tag{6}$$

$$H=\nabla_w\nabla_w\log\{P(t\mid w)P(w\mid\mu)\}=(-\varnothing^{T}\boldsymbol{B}\varnothing-\boldsymbol{A})^{-1} \tag{7}$$

$$\nabla_w=-H^{-1}g \tag{8}$$

$$w_{MP}=w_{MP}+\nabla_w \tag{9}$$

(2)计算权重 w 的后验概率:

$$P(w\mid t,\mu)=\frac{1}{\sqrt{(2\pi)^{m+1}\det(\Sigma)}}\exp\{-\frac{1}{2}(w-w_{MP})^{T}\Sigma^{-1}(w-w_{MP})\} \tag{10}$$

其中,

$$\boldsymbol{\Sigma}=(\varnothing^{T}\boldsymbol{B}\varnothing+\boldsymbol{A})^{-1},\ w_{MP}=\sum\varnothing^{T}\boldsymbol{B}tv \tag{11}$$

公式中,

$$\boldsymbol{B}=\mathrm{diag}(\beta_1,\beta_2,\cdots,\beta_m),\quad \beta_i=\sigma[y(x_i)]\{1-\sigma[y(x_i)]\} \tag{12}$$

(3)联合公式(11)、(12)和公式(13)更新超参数 μ。

$$\mu_i^{new}=\frac{Y_i}{w_{MP}^2} \tag{13}$$

其中,迭代公式如公式(14)所示。

$$\mu_i^{new}=\frac{\gamma_i}{\mu_i^2} \tag{14}$$

训练预测完成后,RVM 分类器会得到一系列取值为 0 到 1 之间的概率预测值,根据这些预测值对全部数据可进行识别判断。

1.2 AP 聚类

Affinity Propagation (AP)[9-11]聚类是一种根据数

据对象之间的相似度自动进行聚类的方法,隶属于划分聚类方法的一种。AP 算法有两个重要的消息 Responsibility 和 Availability。$R(i,k)$ 描述了数据对象 k 适合作为数据对象 i 聚类中心的程度,表示的是从 i 到 k 的消息;$A(i,k)$ 描述了数据对象 i 选择数据对象 k 作为它聚类中心的适合程度,表示从 k 到 i 的消息。$R(i,k)$ 与 $A(i,k)$ 越大,那么数据对象 k 就越有可能作为聚类的中心。AP 算法就是不断迭代更新每一个数据对象的吸引度和归属度,直到迭代一定的次数,产生 m 个高质量的聚类中心,同时将其余数据对象分配到相应的聚类中。

AP 聚类算法在数据点的相似度矩阵上进行聚类。因为聚类的目标是使数据点与其类代表点之间的距离达到最小化,因此选用欧氏距离作为相似度的测量标准,即任意两个点 x_i 和 x_j 之间的相似度为:

$$s(i,k) = -d^2(x_i,x_j) = -\|x_i - x_j\|^2, i = k \quad (15)$$

AP 算法执行步骤如下:

Step1:计算相似度矩阵 $\boldsymbol{S}$;Preference 赋值;

Step2:计算数据对象之间的 Responsibility 值:

$$r(i,k) \leftarrow s(i,k) - \max_{j \neq k}(s(i,j) + a(i,j)) \quad (16)$$

Step3:计算数据对象之间的 Availability 值:

$$a(i,k) \leftarrow \min\{0, r(k,k) + \sum_{j \neq i,k} \max(0, r(j,k))\} \quad (17)$$

$$a(k,k) \leftarrow \sum_{j \neq k} \max(0, r(j,k)) \quad (18)$$

Step4:基于如下数学描述更新 Responsibility 和 Availability 的值:

$$r_{i+1}(i,k) = \lambda * r_i(i,k) + (1-\lambda) * r_{i+1}^{ok1},$$
$$\lambda \in [0.5,1] \quad (19)$$

$$a_{i+1}(i,k) = \lambda * a_i(i,k) + (1-\lambda) * a_{i+1}^{ok1}(i,k),$$
$$\lambda \in [0.5,1] \quad (20)$$

$$a_{i+1}(k,k) = p(k) - \max[a_{i+1}(k,j) + s_{i+1}(k,j)],$$
$$j \in [1,2,\cdots,N], j \neq k \quad (21)$$

Step5:当迭代次数超过最大值或聚类中心不再发生改变时算法结束,输出类中心和每个类包含的数据点;否则返回 Step2。

1.3　Renyi 熵(Entropy of Information and Renyi)

在信息论中,熵用来表示平均信息量,Shannon 提出的熵定义为 Shannon 熵,如下式所示:

$$H(A) = -\sum_{i=1}^{n} P(i)\log P_i \quad (22)$$

式中,$P(i)$ 是概率密度函数,作为熵的一种,Shannon 熵满足如下性质:

(1) H 是连续的;

(2)如果 $P(i)$ 都相同,则有 $P(i) = \frac{1}{N}$;

(3) H 是递增的。

而 Renyi 熵满足以上条件中的第 1 条和第 2 条,所以 Renyi 熵是 Shannon 的广义形式[12-13],如下式所示:

$$R(A) = \frac{1}{1-\alpha}\ln \sum P(i)^{\alpha} \quad (23)$$

与 Shannon 熵相比较可以得知,由于 Renyi 熵具有一个可调节参数 α,因此它通常能够灵活地度量信息量,并且当 $\alpha \to 1$ 的时候,$R(A) \to H(A)$。

2　基于 AP 聚类与 Renyi 熵融合的自训练半监督相关向量机分类预测模型

该文提出的基于 AP 聚类与 Renyi 熵融合的自训练半监督相关向量机分类预测模型技术路线如图 1 所示。

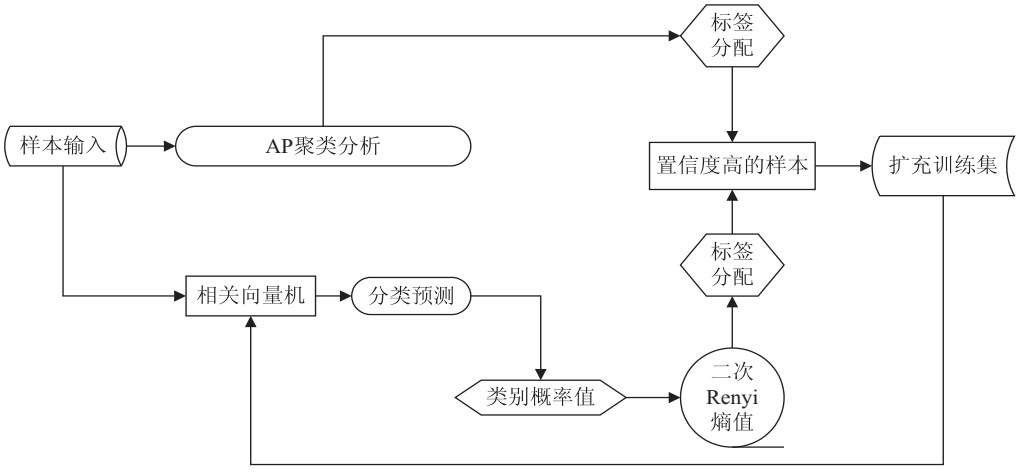

图 1　基于 AP 聚类与 Renyi 熵的自训练半监督 RVM 分类预测模型技术路线图

模型算法执行步骤如下:

Setp1:将数据集的有标签样本和无标签样本一起进行 AP 聚类分析,根据 AP 聚类分类结果初步确定无标签样本的标签类别。确定无标签样本的标签类别采用如下方法:

当 AP 聚类分析完成后,所有样本被分为 N 个类

别,在某一类别中,如果有标签样本的数量占本类别数量的比重 NS 满足如下条件:

$$令:\ \mathrm{MS} = \frac{样本总数}{聚类个数}$$

$$\mathrm{LMS} = \frac{有标记样本总数}{聚类个数}$$

$$\mathrm{NS} \geqslant \tau \frac{\mathrm{LMS}}{\mathrm{MS}},\ 0 < \tau < 1 \tag{24}$$

则该类别中的无标签样本分配与有标签样本一样的标签类别。这里 τ 是调节因子,为了确定最佳的 τ 值,将全部有标签样本作为实验数据集,即有标签样本数据集的 20% 作为有标记样本,剩余的 80% 假定为无标记样本。全部数据集进行 AP 聚类,基于公式(24)判断无标签样本的所属类别,从而得出 AP 聚类的预测准确率。实验中,三个数据集 M. musculus、H. pylori 和 H. sapiens 的最佳 τ 值分别为 0.82、0.63 和 0.58。

Step2:将有标签样本作为训练集,采用相关向量机作为预测分类器,进行无标签样本的标签识别,得到无标签样本所属类别的概率值。

Step3:判断是否满足迭代结束条件,是,转到 Step6,否,转到 Step4。

Step4:根据 Step2 得到的类别概率值,通常将概率值最大的类别标记为该样本的最终识别类别。但是,许多无标签样本预测出的类别概率值几乎相同,差别很小,如果单从概率值来判定无标签样本的最终类别,往往会造成错判和漏判,从而生成噪声数据,影响自训练半监督分类器的预测性能。该文通过采用 AP 聚类与 Renyi 熵融合的方法来共同决定无标签样本的标签类别。由于蛋白质相互作用数据样本存在较大的类别不确定性,因此标签的分配一定程度上就是对不确定性的度量。而 Renyi 熵是一种稳定的熵度量方法,对混杂或具有不规则碎片形状的非可加性系统提供更佳的解释,而这一点能够满足蛋白质相互作用样本数据的特征需要,所以采用 Renyi 熵能够更好地对蛋白质相互作用样本进行度量。由于二次 Renyi 熵比较稳定,而且计算量小,容易实现,因此该文采用二次 Renyi 熵作为样本类别不确定性的度量。二次 Renyi 熵数学描述如下:

$$\mathrm{RR}_2(X) = -\ln \sum_{i=1}^{n} p(x_i)^2 \tag{25}$$

其中,$p(x_i)$ 是蛋白质序列对的预测概率值。为了防止线性回归算法在计算概率过程中出现无穷大的数值,该文对公式(25)进行归一化处理,从而有:

$$\mathrm{RR}_2(X) = -p(x_i)\ln \sum_{i=1}^{n} p(x_i)^2 \tag{26}$$

显然,K 个样本中的最大 Renyi 熵为:

$$\mathrm{RS}(U) = \max_k \{\mathrm{RR}_2(X)\} \tag{27}$$

式中,RS(U) 表示蛋白质序列对样本中最大 Renyi 熵的若干个样本,熵越大的样本不确定性越大,信息量也越大,也是无法确定分类信息的样本,根据有标签样本的标签信息,将这些熵值最大的样本分配相应的类别标签。

Step5:将 Renyi 熵与 AP 聚类分析标签类别判定一致的无标签样本添加到现有的训练集中,用扩充后的训练集继续迭代训练分类器,转到 Step2。

Step6:输出分类结果,算法结束。

3 实 验

3.1 实验数据集

为了验证提出的分类预测模型的有效性,该文在三个蛋白质相互作用数据集 M. musculus、H. pylori 和 H. sapiens 上进行了实验验证,表 1 列出了实验数据集的样本数量。

表 1 实验数据集样本数量

数据集	蛋白质序列	有标记样本	无标记样本	样本总数
M. musculus	355	313	3 130	3 443
H. pylori	706	1 420	14 200	15 620
H. sapiens	1 083	1 412	14 120	15 532

3.2 实验结果及分析

为了描述方便,表 2 列出了基于不同自训练方法的半监督相关向量机中文名称及英文简称,其中 ST 表示自训练,SSRVM 表示半监督相关向量机。

表 2 基于不同半监督相关向量机英文简称

中文名称	英文简称
基于 AP 聚类的自训练半监督 RVM	AP-ST-SSRVM
基于 Renyi 熵的自训练半监督 RVM	Renyi-ST-SSSRVM
基于 AP 聚类与 Renyi 熵融合的自训练半监督 RVM	AP-Renyi-ST-SSS-RVM

实验中,针对蛋白质序列特征向量的生成,该文采用文献[14]提出的基于位置特异性打分矩阵(PSSM)的串行多特征融合的蛋白质序列特征提取方法,该方法通过局域蛋白质序列 PSSM 矩阵编码捕获序列上连续的和间断的蛋白质相互作用信息;通过串行多特征融合实现序列中蕴含的多种关键特征信息的整合;针对样本测试集与训练集的构建,该文分别从三个数据集中随机抽取有标签样本的 20% 作为测试集,80% 作为初始预测模型训练集。当模型每次迭代结束后针对每个数据集的测试集样本进行预测分类,得出当前模型的预测准确率,从而了解当前模型的预测性能。

下面列出了不同的预测模型在 M. musculus、H. pylori 和 H. sapiens 数据集上的实验结果,如表 3 ~ 表 5 所示。

表 3　M. musculus 数据集不同预测模型预测结果
(准确率%)

迭代次数	AP-ST-SSRVM	Renyi-ST-SSSRVM	AP-Renyi-ST-SSSRVM
1	71.12	73.22	74.52
2	73.35	76.46	76.65
3	75.98	78.42	78.62
4	78.68	79.38	80.28
5	78.93	80.01	81.32
6	78.91	80.21	82.17
7	78.88	80.20	83.35
8	78.92	80.18	83.98
9	#	80.22	83.96
10	#	#	83.97
11	#	#	83.98

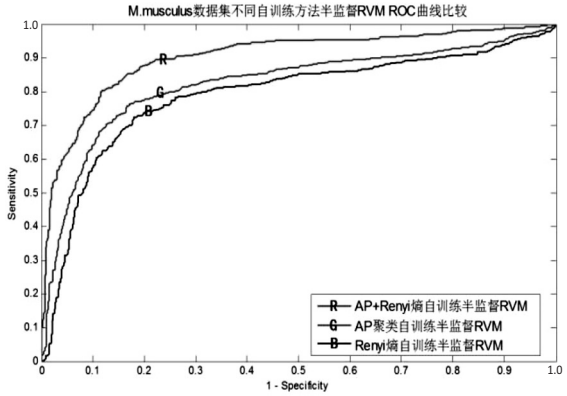

图 2　M. musculus 数据集不同预测模型 ROC 曲线比较

表 4　H. pyloris 数据集不同预测模型预测结果
(准确率%)

迭代次数	AP-ST-SSRVM	Renyi-ST-SSSRVM	AP-Renyi-ST-SSSRVM
1	73.35	74.32	75.69
2	74.56	76.65	77.12
3	76.98	78.92	78.41
4	78.62	80.28	79.32
5	79.98	81.32	80.69
6	80.32	82.17	81.98
7	81.11	83.35	83.26
8	81.35	83.72	85.27
9	81.67	83.7	87.96
10	81.65	83.73	88.68
11	81.68	83.71	88.66
12	81.66	83.69	#
13	81.67	#	#

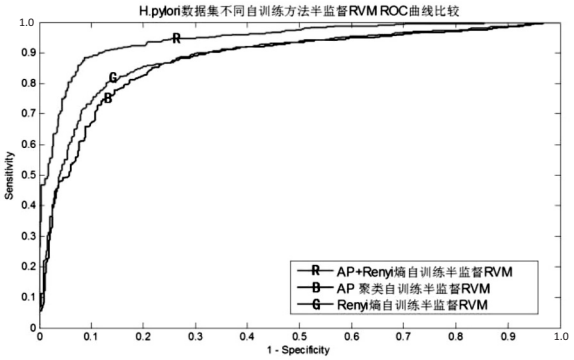

图 3　H. pylori 数据集不同分类算法 ROC 曲线比较

表 5　H. sapienss 数据集不同预测模型预测结果
(准确率%)

迭代次数	AP-ST-SSRVM	Renyi-ST-SSSRVM	AP-Renyi-ST-SSSRVM
1	72.91	73.86	74.73
2	73.58	76.62	76.97
3	75.28	78.68	78.57
4	76.53	79.26	79.23
5	77.34	80.18	80.56
6	78.76	81.95	81.58
7	79.98	82.11	82.55
8	80.87	82.88	83.39
9	81.31	83.68	85.27
10	81.29	83.66	86.32
11	81.32	83.67	88.75
12	81.30	83.69	88.72
13	81.71	#	#
14	81.69	#	#

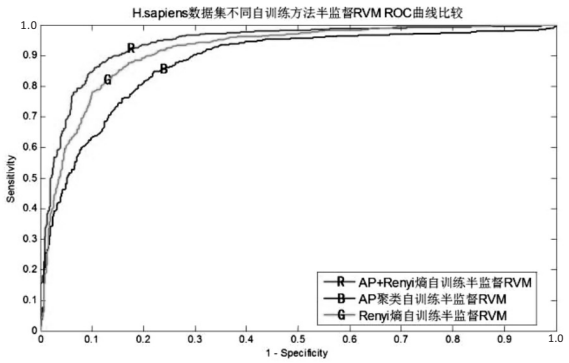

图 4　H. sapiens 数据集不同分类算法 ROC 曲线比较

从表 3 ~ 表 5 中可以看出,AP-ST-SSRVM、Renyi-ST-SSRVM 及 AP-Renyi-ST-SSRVM 分类算法针对 M. musculus、H. pylori 和 H. sapiens 三个数据集的初始预测准确率都相对较低,分别为 71.12%、73.35% 及 72.91%;73.12%、74.32% 及 73.86% 和 74.52%、75.69% 及 74.73%。但随着训练样本数的增加,三种

分类算法针对三个数据集的预测准确率都有了明显提升;AP-ST-SSRVM 分别迭代 5 次、10 次和 9 次后;Renyi-ST-SSRVM 分别迭代 6 次、8 次和 9 次后以及 AP-Renyi-ST-SSRVM 分别迭代 8 次、10 次和 12 次后它们的预测准确率曲线趋于平直。AP-Renyi-ST-SSRVM 的迭代次数多于其他两种分类算法,但它的预测准确率是最高的。迭代次数多是因为 AP-Renyi-ST-SSRVM 相比 AP-ST-SSRVM 增加了二次 Renyi 熵的验证,相比 Renyi-ST-SSRVM 增加了 AP 聚类分析,从而增加了计算开销。但相对于能够得到较高的预测准确率,这种开销成本的增加是可以忽略的。同样的,通过图 2 ~ 图 4 分别展示的针对 M. musculus、H. pylori 和 H. sapiens 三个数据集三种分类算法的 ROC 曲线对比,进一步证明了 AP-Renyi-ST-SSRVM 分类算法在预测性能上优于其他两种分类算法。

此外,该文提出的分类预测模型分别与其他研究学者提出的预测模型在 M. musculus、H. pylori 和 H. sapien 数据集上进行了比较分析,如表 6 所示。

表 6 M. musculus、H. pylori 和 H. sapient 数据集不同预测模型预测结果比较(准确率%)

预测模型	M. musculus	H. pylori	H. sapien
Huang' work[15]	79.87	82.18	82.22
You' work[16]	83.39	85.77	88.81
Gao' work[17]	N/A	84.84	N/A
Nanni' work[18]	N/A	86.60	N/A
文中预测模型	83.98	88.67	88.69

从表 6 可以看出,文中构建的预测模型在 M. musculus 和 H. pylori 数据集上的预测准确率都高于其他预测模型,在 H. sapien 数据集上预测准确率也高于 Huang' work[15] 的预测模型,同 You' work[16] 的预测准确率基本相同。这进一步验证了提出的基于半监督学习的蛋白质相互作用预测模型的有效性。

AP-Renyi-ST-SSRVM 分类算法的主要优势在于:通过 AP 聚类与 Renyi 熵融合的方法将置信度高的无标签样本,即 AP 聚类分析与二次 Renyi 熵判定类别一致的样本,标记为有标签样本,加入到原有训练集中,用扩充后的训练集进行自训练迭代分类,构造出了性能最优的半监督分类器。通过以上处理可以大大减少由于误判而生成噪声数据的数量,从而能够降低噪声数据对分类器预测性能的影响,提高预测准确率。

4 结束语

通过实验结果还发现,基于不同自训练方法的半监督相关向量机模型的预测准确率与训练集有标签样本数的多少密切相关,初始有标签样本数越多,分类准确率越高,并且随着训练样本的不断增加,预测准确率有明显提升;但当训练样本集到一定规模后,即使再添加更多的有标签样本,预测准确率也无明显变化,达到一种饱和状态。因此,基于以上分析可以得出如下结论:

(1)提出的基于 AP 聚类和 Renyi 熵融合的自训练半监督相关向量机分类预测模型极大降低了噪声数据对分类器预测性能的影响。初始训练只需选择较少量的有标签样本,通过自训练识别无标签样本并添加到当前训练集,预测模型通过多次迭代学习和纠错,能够获得好的预测性能。模型预测准确率较高,预测分类效果良好,可以应用到多种类型的蛋白质相互作用预测分类中;

(2)有标签样本的数量影响半监督分类算法的预测性能。随着新的有标签样本不断添加到训练集,预测模型的分类准确率和分类效果都有较大提高,但当训练集达到一定规模时,预测性能又趋于平稳。因此,半监督学习中有标签样本数量的合适选择是一个值得研究的问题,要充分平衡半监督学习的优点和有标签训练样本数量之间的关系,使最终的分类结果能够达到最优。

参考文献:

[1] CHEN X, GUPTA A. Webly supervised learning of convolutional networks[C]//IEEE international conference on computer vision (ICCV). Santiago: IEEE, 2015: 1431-1439.

[2] WANG X, SONTAG D, WANG F. Unsupervised learning of disease progression models[C]//Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'14). New York, NY, USA: ACM, 2014: 85-94.

[3] BLUM A. Semi-supervised learning[M]. Berlin: Springer, 2016.

[4] RASMUS A, VALPOLA H, HONKALA M, et al. Semi-supervised learning with ladder networks[J]. arXiv: 1507.02672, 2015: 1-9.

[5] FEI S, HE Y. A multiple-kernel relevance vector machine with nonlinear decreasing inertia weight PSO for state prediction of bearing[M]//Advances in data analysis. Berlin, Heidelberg: Springer, 2015: 585-592.

[6] HOANG N D, BUI D T. A novel relevance vector machine classifier with cuckoo search optimization for spatial prediction of landslides[J]. Journal of Computing in Civil Engineering, 2016, 30(5): 04016001.

[7] KALTWANG S, TODOROVIC S, PANTIC M. Doubly sparse relevance vector machine for continuous facial behavior estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(9): 1748-1761.

(下转第 27 页)