SVM与主动学习方法相结合的蛋白质相互作用预测

史文丽[1] 郭茂祖[1] 李 晋[1,2] 刘晓燕[1]
(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)[1]
(哈尔滨医科大学生物信息科学与技术学院 哈尔滨 150081)[2]

摘 要 提出了基于SVM的主动学习算法,用来解决蛋白质相互作用的预测问题。细胞中的生物过程是通过蛋白质相互作用实现的。但是通过实验验证蛋白质之间是否具有相互作用的代价非常大,而且数据很难获取。为了在有限的阳性样本情况下更加快速准确地预测蛋白质之间是否具有相互作用,引入了主动学习方法。主动学习算法可以用来构造有效训练集,其目标是通过迭代抽样,每次寻找最富有信息量的数据点,找到最有利于提升分类效果的样本,进而减小分类训练集的大小。比较了5种不同的主动学习算法,以寻找在有限资源前提下提高分类算法效率的最佳途径。实验表明,主动学习方法与SVM算法相结合,能够在保证SVM分类性能的前提下,有效减少学习所需的样本数量。

关键词 支持向量机,主动学习,蛋白质相互作用
中图法分类号 TP18 **文献标识码** A

Protein-protein Interaction Prediction Combining Active Learning with SVM

SHI Wen-li[1] GUO Mao-zu[1] LI Jin[1,2] LIU Xiao-yan[1]
(School of Computer Sciences and Technology, Harbin Institute of Technology, Harbin 150001, China)[1]
(College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China)[2]

Abstract An active learning method using SVM was introduced in this paper to solve the problem of protein-protein interaction prediction task. Biological processes in cells are carried out through protein-protein interactions. Since determining whether a pair of genes interacts by wet-lab experiments is resource-intensive, we proposed a support vector machine active learning algorithm for interaction prediction. Active machine learning can guide the selection of pairs of genes for future experimental characterization in order to accelerate accurate prediction of the human gene interactome. As a method of constructing an effective training set, the goal of active learning algorithm is to find informative sample which can enhance the classification results of the model during the iteration, thereby reducing the size of the training set and improving the efficiency of the model within limited time and resources. The experiment shows that compared with the general SVM, active learning with SVM can reduce the number of examples effectively on the premise of keeping correctness of the classifier.

Keywords Support vector machine, Active learning, Protein-protein interaction

1 引言

蛋白质相互作用(protein-protein interaction,PPI)是生物体中所有生物过程的核心,在细胞的职能中发挥关键的作用,使细胞可以传导信号、完成代谢途径和促进生物体的组织功能。构建蛋白互作网络可以使我们对蛋白质的功能有更加全面的认识,加快生物医学的发展,了解疾病的机制,并可能从中发现新的药物[1]。目前实验中大都采用高通量的方法,如酵母双杂交(Y2H)和质谱分析,来检测蛋白质是否互作。然而,这些方法都有较高的假阳性率,很多情况下会出现其中一种方法预测出某蛋白质对具有相互作用而另一种方法却预测出相反的结果。通过验证得知,通过酵母双杂交方法预测出来的相互作用的误报率在70%左右[2],在多种预测酵母蛋白质相互作用的高通量方法中,结果相同的仅占3%左右[3]。在像人类这样的复杂生物体中应用高通量的方法,预测每一个可能互作的蛋白质对,其成本和精力都将是非常昂贵的。据估计,人体中约有15到60万对的互作蛋白,但参考人类蛋白质数据库[3],目前已知的或可能的只有3.8万。因此为了尽快完成相互作用组的预测,非常有必要提出计算机预测方法。

机器学习方法作为高通量方法的重要补充,可以加快重建相互作用组。经过生物学家在个别蛋白质上几十年的研究,以及高通量技术的进步,一些计算系统已被开发出来用于预测蛋白质相互作用[4]。新的算法最近开始不断涌现,尤其是基于机器学习的蛋白质相互作用预测模型。贝叶斯分类

到稿日期:2013-05-20 返修日期:2013-07-16 本文受国家自然科学基金(60932008,61172098,61271346),高等学校博士学科点专项科研基金(20112302110040)资助。

史文丽(1989—),女,硕士生,主要研究方向为生物信息学;郭茂祖(1966—),男,教授,博士生导师,主要研究方向为生物信息学与机器学习,E-mail:maozuguo@hit.edu.cn(通信作者);李 晋(1983—),男,博士生,讲师,主要研究方向为生物信息学。

器[5,6]、随机森林[6,7]、逻辑回归[6,7]、支持向量机(support vector machine,SVM)[6]、决策树[8]均已被应用于蛋白质相互作用(PPI)预测。它们通过学习已知PPI的信息(标记好的训练数据)与其他一些间接信息,例如,基因本体注释、基因表达相关、序列的同源性等,来预测未知的PPI。本文提出在支持向量机(SVM)中应用主动学习的方法训练分类器。支持向量机是基于统计学习理论的一种通用有效的机器学习方法,与主动学习相结合能够使之充分利用未标记样本建立最有价值的训练集,得到的分类器具有较高的泛化能力。

通过实验验证蛋白质之间是否具有相互作用不仅费时、费力,而且训练集可能包含大量的冗余样本,因此,在监督学习任务中最小化分类训练集的大小,可以减小标注成本,有效提高训练效果。主动学习是监督学习类型的一种,其主要目标是高效地寻找训练数据集中高信息量的样本,即选择出在实验室中进行过验证的蛋白质相互作用对进行标注。

现有的用于分类问题的主动学习算法一般有以下3种常用形式[9]:(1)基于评委的启发式方法[10],选择一定数量的分类器,利用初始训练集训练这些分类器,并选择分类结果中最不一致的样本;(2)基于置信度的方法[11],选择当前分类器最不能确定类别的样本,一般认为这些样本最具有价值;(3)基于后验概率的启发式方法,根据预测所得样本后验概率值的大小对候选样本集进行排序,选择精确度较高的样本。本文采用了SVM主动学习策略用于蛋白质互作的预测。一般情况下,监督机器学习方法使用特定决策方法避免过度拟合,尽量减小归一错误率(或误差函数),即决策函数 f^* 可表征为:

$$f^* = \arg\min_{f_i \in F}\left[\sum_{i \in D} \| \gamma_i - f_j(\vec{x_i}) \| + \beta(f)\right]$$

式中,$\vec{x_i}$ 是实例 i 的特征值,γ_i 是真实的标签(或值),$f(\vec{x_i})$ 是预测的标签(或值),D 是训练数据的集合,f 是一组可能的预测函数(如SVM)。主动学习就是基于这个标准,从所有可能的实例中选择下一个 $\vec{x}_{i+1}$,使得如果知道它的真实标签 γ_{i+1},可以最大限度地提高我们的估计函数 f^*。

本文中的主动学习算法可以表述为以下模型:

$$A=(C,L,S,Q,U)$$

其中,C 为 SVM 分类器,L 为已经标注好的训练样本集,Q 为所选的查询函数,用于在未标注的样本中选择数据,U 为整个蛋白质对数据集,S 为监督者,对未标注样本进行标注。该算法从处理所有未标记的数据开始,由监督者选择初始训练集建立初始分类器模型。然后 S 从未标注样本集 U 中,按照查询标准 Q 选取一定的未标注样本进行标注,并加入到训练集 L 中,重新训练分类器,直至达到停止标准。SVM通过学习能够代表整个样本集的特征子集(SV),移除非 SV 样本,为实现主动学习提供了可能。

聚类是普遍用来选择有代表性数据点的前处理步骤。应用于主动学习中的聚类算法有K-means[12]和K-medoids[8,13]。基于置信度的选择策略包括选择最接近决策超平面[14]的支持向量机分类器,它选择与支持向量机分类超平面最接近的数据点用于标记。聚类方法与主动学习方法结合,能够有效改善分类器的性能。

2 方法

2.1 评价指标

准确率(precison)用来计算在所有分类器预测出具有相互作用的蛋白质对中正确分类的蛋白质对的数量。召回率(recall)表示分类器能够识别出的具有相互作用的蛋白质对与所有相互作用蛋白质对的比率。F-score是准确率和召回率的调和平均值。F-score通过结合准确率和召回率来度量算法的精确度。因此,它可以被用作评价方法好坏的指标。

为了更详尽地说明,我们假设样本集 $D=\{(x_1,y_1),(x_2,y_2)\cdots(x_m,y_m)\}$,$y_i \in \{0,1\}$,$i=1,2,\cdots,m$,分类器预测的结果为 z_i,$i=1,2,\cdots,m$,则

$$\text{准确率}=\frac{\sum_{i=1}^{m}(z_i=y_i=1)}{\sum_{i=1}^{m}(z_i=1)}$$

$$\text{召回率}=\frac{\sum_{i=1}^{m}(z_i=y_i=1)}{\sum_{i=1}^{m}(y_i=1)}$$

$$F\text{-}score=2*\frac{\text{准确率}*\text{召回率}}{\text{准确率}+\text{召回率}}$$

支持向量机分类:

支持向量机是基于统计学习理论的一种通用有效的机器学习方法,具有较强的泛化能力。支持向量机(SVM)在特征子集上训练数据。该方法从样本集中选择一组特征子集,使得对于特征子集的划分等价于对整个样本集的分割,这组特征子集称为支持向量(SV)。基于主动学习的分类器将SVM分类器的训练过程看作一个迭代过程,首先每次迭代都从未标记样本中通过某种度量方法找到最有价值的样本进行人工标注,然后加入SVM的训练集,循环直到分类器的精度或循环次数达到某一阈值时停止。SVM的实现采用林智仁开发的libsvm软件包[15]。

用SVM实现主动学习具体的算法描述如下:

输入:未带类别标注的候选样本集D,每次从候选样本中采样的个数n

输出:分类器f

步骤:

1)从候选样本集D中选择n个样本并标注类别,构造初始训练样本集 I_0,使 I_0 中至少包含有一个正例样本和一个负例样本,执行 $D_0=D-I_0$ 操作。

2)进行第i次采样学习,在样本集 I_{i-1} 基础上寻找最优分类超平面 f_i,从样本集 D_{i-1} 中按照某种策略选择n个样本,将这n个样本组成的集合记为 B_i。

3)正确标注这n个样本的样本类别。

4)执行 $I_i=I_{i-1}\cup B_i$,$D_i=D-I_i$,当准确率满足某种指标时终止学习,否则返回到第2)步。

5)返回 $f=f_i$。

2.2 主动学习数据选择策略

为了完成主动学习训练方法,所有的数据都是未标记的,主动学习方法根据实例(蛋白质对)的分布和在每次迭代中优化学习到的决策函数依据某一准则标记数据。重复这一过程,直到达到标注数据的最大值。原始主动学习方法中每次选择一个数据点的标签,迭代次数等于总共需要数据点的个数,本文中为了减少分类器的重新训练的次数,每次选择多个数据点。在下面描述的几种不同类型的数据选择方法中,每次迭代均选择250个数据点进行标注,迭代共进行12次,所以最终的训练集中有3000个数据点。

A. 基准-随机数据选择策略

为了便于比较,首先采用每次随机从数据集中选择特定数量的数据的主动学习方法。此方法中使用2组训练数据来构造SVM,它们所包含正例的比率不同,分别是20%和45%。训练集的大小逐步递增,蛋白质对从250增加至3000,增量为250。每次迭代中的250对数据是从整体的

15000 个数据点中随机选择的。SVM 在每次迭代中被优化，然后使用统一的测试数据对其性能进行评估。

B. 基于密度的选择策略

随机选择数据不能确定所选样本的特征，很可能每次只选择了相同类别的数据，不利于主动方法的学习，而聚类可以根据特征的值来识别同类的对象，将蛋白质对根据表达特征值的相似程度聚为不同的类。每次在聚完之后的不同类别中按照特定比例选择数据，可以使主动学习方法选择到的数据分布更均匀，信息量更大，更有代表性。

在这种主动学习技术中，数据首先根据 K-means 算法聚为 K 类。选定的点根据簇分布，并与之所在群集的大小成比例。在每次迭代中需要数据点的总标签数是固定的。假设 n_i 是 C_i 簇中数据点的个数，N 为总数据量的大小。那么从簇 C_i 中选取数据的数量为 s_i：

$$s_i = S * n_i / N$$

式中，$S=\sum_{i=1}^{K} s_i$。

在每个集群 C_i 中，选择最接近中心的 s_i 个未标记数据点并得到它们的类别，加入到训练集中。

C. 基于置信度的选择策略(随机种子)

根据 SVM 的特点，当使用一组初始的数据样本训练好支持向量后，以后的每次分类中都可以计算出其余样本与分类超平面之间的距离，即置信度。

在这种主动学习策略中，首次迭代需要标签的数据是随机选择的，使用这些数据建立一个初始的 SVM。之后每次迭代选取离分类超平面最近的样本。这些样本因为类别最不确定，也最有可能被分错，所以信息量最大，也最有可能改变超平面的位置，而远离超平面的样本对其位置的改善影响不大。

在每次迭代中，选择离超平面最近的 250 个数据点并得到它们的标签。将之添加到现有的已被标号的数据，再利用这些新形成的数据训练一个新的 SVM。这个新的 SVM 用于在下一次迭代中选择数据。

D. 以密度为种子的置信度选择策略

此方法与 C 方法基本相同，不同点在于在第一次迭代中，根据密度选择数据(通过执行 K-means 聚类方法)，而不是随机选择，这样就可以保证初始分类器的性能不至于太差。

E. 基于先验知识的方法

以上几种方法都是根据聚类或置信度等外界方法或分类器本身所学习到的知识来进行主动学习的，这些方法都具有依赖性，若所学知识不准确，很可能对分类器的性能影响极大。本文提出一种基于先验知识的主动学习数据选择方法，方法中引入 Confusion 这一概念，表示前 m 次预测的差异。计算公式如下：

$P_{A0}(x)=$ 前 m 个分类器将蛋白质对 x 分类为 0 的平均概率

$P_{A1}(x)=$ 前 m 个分类器将蛋白质对 x 分类为 1 的平均概率

其中，0 代表没有相互作用，1 代表有相互作用。

Confusion 表示平均预测值相对熵的总和。

$$Confusion(x)=\sum_{j\in(0,1)} p_{Aj} * \log(p_{Aj})$$

这种方法要求首次使用的 m 个分类器使用其他机制生成，后续的迭代中使用上述 Confusion 的值选择数据点。由于“以密度为种子的不确定性方法”与其他方法相比效果更好(见结果部分)，首次使用的 m 个分类器使用此种方法。

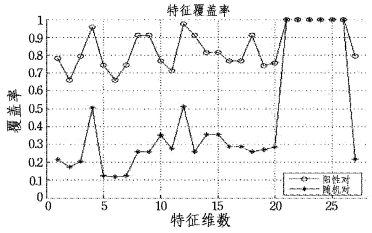
3 数据

3.1 数据集和特征描述

本文中使用 Qi Yanjun 等创建的数据集[16]，这组数据集是专门为了评价主动学习算法而创建的。数据集中有 14600 对已知的互作蛋白，这些蛋白质对被称为阳性对。随机产生 40 多万对与阳性对不重叠的样本，称其为随机对，并认为它们之间是无相互作用的。研究表明，随机产生的 1000 对蛋白质对中具有相互作用的概率小于 1 对[5,17]。

PPI 的预测是一个二分类任务：每个特征向量对应于一对蛋白质，被分类为具有相互作用或无相互作用两类。Qi Yanjun 为阳性对和随机对都计算出了特征向量。向量共具有 27 维，包括基因本体论(GO)的细胞主分(1)、GO 分子功能(1)、GO 生物过程(1)、组织中发生的共现(1)、基因表达(16)、序列相似性(1)、同源性(5)和域相互作用(1)(括号中的数字对应相应向量的维数)。两个基因的 GO 功能相似性取决于它们在 GO 数据库中共享的 term 之间的相似性，分别对应于生物过程、分子功能和细胞组分生成 3 个 GO 特征。16 个基因表达的特征是从 NCBI 数据库中 16 个基因表达数据集里蛋白对的相关系数计算得来的。“组织功能”是一个二值特征，表示两个蛋白质是否属于同一组织。序列相似性特征使用 BlastP 序列比对计算蛋白质对的 E-value。“域相互作用特征”中，蛋白质对相互作用的概率由两种蛋白质中存在的域的相互作用的概率决定。“PPI 同源性特征”由与给定的人类蛋白质同源的蛋白质是否在其他物种中交互(如酵母等)而估计出。每种特征的详细信息可以在他们的网站[16]上得到。

为了理解特征空间，首先需要研究每个特征的覆盖率，即蛋白质对每个可用特征值所占的百分比。图 1 分别显示出阳性对和随机对每个特征值可用部分所占的百分比。可以看出，阳性对中所有特征都可用的蛋白质对占很大比例，但随机对中只有约 20%～40%是可用的。基因表达特征是个特例，对于所有的蛋白质对它都是可用的。阳性对比随机对数据的缺失值要少，有些蛋白质对的特征向量包含多个缺失值，甚至有些特征向量具有 80%的缺失值，相反，有些蛋白质对的特征覆盖率为 100%，即其所有特征都可用。为了保持阳性对与随机对特征覆盖的平衡(详见结果部分)，我们建立了一个同质的数据子集，以保证每对蛋白质有 80%以上的特征覆盖率。该同质子集总共具有 55950 对蛋白质，在它们中随机选择 10000 对蛋白质对进行训练和 10000 对进行测试，该子集用于算法实现和结果评估。



x 轴显示了人类蛋白质互作对的 27 种特征，y 轴显示了每种特征的覆盖率

图 1 互作对与随机对的特征覆盖率

该子集中，阳性对和阴性对按 20%～80%的比例相结合(详见结果部分)。

3.2 蛋白质对特征空间的覆盖

阳性对中大多数特征向量的覆盖率都达到了 60%，而且近 20%的数据具有所有 27 个特征，覆盖率达到 100%。然

而，大部分随机对都只有 20％的特征可用，达到 100％覆盖率的几乎没有。因此，分类器必须可以学习缺失值的模型，而不只是学习现有特征的生物学规律。

本文进行了一个小实验来评估两类数据的特征覆盖率是否明显影响实验结果。对于特征向量中的元素，用 1 代替存在的特征，0 代替缺失的特征。换句话说，如果基因本体位点的值是已知的，那么不管该位点的值是多少，该特征都用 1 替换。用这些新的特征向量训练支持向量机。我们把这种新的特征向量称为“覆盖向量”。对应于原有特征向量的 27 维，新的向量也具有 27 维。如图 2 所示，支持向量机在这些二值覆盖向量上训练的结果是：准确度达到 72％、召回率达到 53％、F-score 达到 61％；而在原始特征向量上的结果为精度 87％、召回率 24％和 F-score 39％；覆盖向量比原始向量的训练效果反而更好(Fscore 值更高)。这表示，在阳性对和随机对的蛋白质对分类中，覆盖向量要优于特征向量。而特征向量由于对阳性样本的特征覆盖率更高，所以得到的准确率更高。

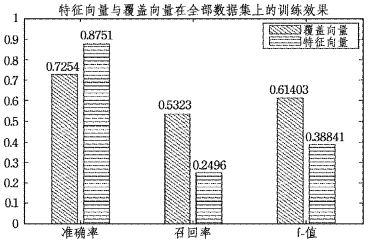


图 2　覆盖特征向量和全部数据的训练结果

出现这种情况的原因可能是，因为通过实验验证的具有相互作用的蛋白质对非常重要，它将极有可能被其它实验当作特征，从而在蛋白质对的特征向量中表现为多个都存在的特征。

3.3　缺失值的处理

为了评估学习算法在没有特征覆盖带来的间接偏置影响下预测相互作用的真正能力，采用统计分析软件 SPSS 来处理缺值[18]。SPSS 软件中提供了 4 种缺值处理方法，分别是列表、配对、EM 和回归算法。本文中使用 EM 算法来对蛋白质对数据进行缺值处理，生成新的数据集。在这个新的数据集中，所有的特征向量都认为是存在的。

4　实验

本文选择了包含 20％阳性对的训练数据集来评价所有的主动学习算法。这是因为自然界中具有相互作用的蛋白质对比非相互作用的要低。但是正如在结果部分所述，如果训练数据(包含几千对)中阳性对的比例非常低(比如只有 1％)，召回率将非常之小。为了比较主动学习与非主动学习方法的学习能力，我们选择具有 20％阳性对的训练数据集。

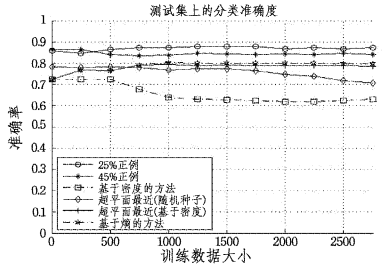
每种方法都以 250 个标记好的蛋白质对进行初始化。在 K-means 聚类中，选择集群数 K 为 2。5 种算法中每次迭代都要求 250 个数据点的标号。随着被标记数据的更新，训练出一个 SVM，并在每次迭代中使用测试数据评估其性能。

以上每种算法被执行 5 次，结果取其平均值。这样做是因为以上方法中有两个初始数据的选择是随机的，因此其性能可能会因为初始值的选择而有所不同。多次运行求平均值可以提供更可靠的性能比较。

本文在上述的训练和测试数据集上对 5 种算法进行了评价，并分别计算出了其准确率、召回率和 F-score 值。

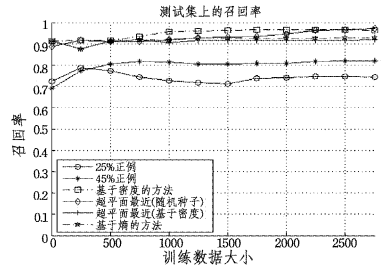
5　结果与分析

图 3 显示了 4 种方法在每次迭代中的准确率。图 4 显示的是召回率，图 5 是 F-score 值。



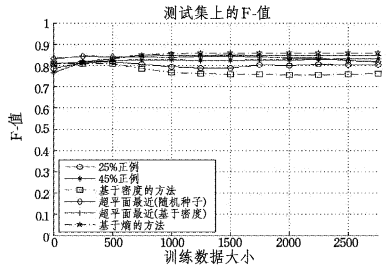
x 轴表示标注数据点的个数，y 轴表示各种算法的准确度

图 3　准确率



x 轴表示标注数据点的个数，y 轴表示各种算法的召回率

图 4　召回率



x 轴表示标注数据点的个数，y 轴表示各种算法的 F-score

图 5　F-值

从图中可以看出，4 种主动学习方法在具有 500 个数据点时比随机方法在具有 3000 个数据点时的 F-score 值还要高。这表明了主动学习策略在标记数据较少的情况下可以获得更准确的结果。另外，将训练数据中阳性对的百分比从 20％提高至 45％，召回率及 F-score 值可以得到很大提升，它虽然比随机方法的精确度低，但具有较高的召回率，权衡两者之后的 F-score 值得到了大的提高。

对基于密度的主动学习方法精确度较低的内在原因进行进一步分析，可知，因为非相互作用的蛋白质对占训练数据的绝大多数，所以在每次迭代中，基于密度的方法比随机方法更偏向于选择较大比例的非相互作用的蛋白质对进行标记，分类器对于相互作用的蛋白质对的特征还没有完全学会，因此分类相互作用的蛋白质对有一定的难度。训练数据中的阳性对越多，分类器对阳性对的学习就更全面，从而可以达到更高的召回率(见图 4)，但是训练数据中非相互作用蛋白质对的减少使得对于阴性对的学习不足，从而非相互作用对会被归类为相互作用对，导致精度降低(见图 3)。为了检验这一推理，我们将随机方法应用到含 45％相互作用对的数据集上。结果表明，在训练数据中随机选择 45％的阳性对将导致精确度的下降，但是召回率增加(见图 3 和图 4)。主动学习方法

因为选择更富有信息量的数据点，其性能还是优于随机选择方法。

基于密度的方法在约1500个数据点时达到F-score的最大值。此方法在1500个标记样本时召回率达到95%左右，但之后没有进一步的提高。这可能是因为在聚类中心选取的1250～1500个数据点足以代表数据的分布。更多的样本似乎并不会再提供更多的信息，这种方法与随机和其他主动学习方法相比精度较低。经过分析发现，其对数据的聚类是不精确的，在训练数据集中，两类样本的平均纯度为83.4%，而由相互作用对主导的集群的平均纯度为77.5%。另外由于训练数据主要由非相互作用对主导，实际上相互作用对的64.55%存在于由非相互作用对主导的集群中。这些问题都限制了用纯粹基于聚类的方法分类蛋白质对的相互作用。

在基于随机种子的超平面最近方法中，主动学习首次迭代的召回率增加的幅度最大，此时数据点的个数从250增加到500(见图4)，这导致F-score从0.8左右上升到0.85以上。然而第一次迭代后精度下降。正如前面所述，这是由于基于超平面的方法趋向于选择大量的相互作用对。它首次迭代中选择的数据点(该方法首次选择的250个点)的65%是相互作用对，比数据集中的比例还要高得多。然而，在接下来的迭代中精确度逐渐增加，在3000个数据点时达到70.5%(见图3)。

基于密度种子的超平面最近方法比基于随机种子的超平面最近方法所得到的F-score更高。从图中可以看出，基于密度选择种子比随机选择种子更能增加召回率(见图4)(因为它可以更好地表示相关数据分布)，从而导致更高的F-score。

基于先验知识的方法可以得到最高的召回率和F-score值。因为它考虑了数据点的前3次预测值($M=3$)，而不像其他主动学习方法中F-score在前几次迭代后就没有明显的改进了，基于先验知识的方法的F-score能够保持持续上升。在3000个标记数据点时它达到了91%的召回率、80%的精确度和85%的F-score。

从图2与图3－图5的对比中可以看出，主动学习方法比在原始数据集上训练SVM可以达到更好的效果，这可能是因为原始训练集中非相互作用的蛋白质对数量非常庞大(达到四十万之多)，导致对阳性对的学习不足，召回率降低。

结束语 本文对蛋白质相互作用预测任务中的5种不同的主动学习算法进行了评价。结果表明，主动学习在标记训练数据较少的情况下比单独使用SVM能够更好地进行学习。基于密度的方法通过选择最能够代表未标记对的数据来提高召回率。在混合方法中应用基于密度的种子比随机种子更能提高性能。值得注意的是，计算之前预测的差异性(基于先验知识的方法)比在当前分类器中预测单一样本的值更准确(基于随机/密度种子的置信度选择策略)。事实上相互作用的蛋白质对在整体未标记的集合中所占的比例很低，这导致了对于定义阳性对的规则/特征的更快学习，定义正相互作用的蛋白质对表现出了这些方法对于蛋白质相互作用预测问题在阳性对明显低于阴性对的情况下的适用性。

许多人类的蛋白质相互作用仍未被发现，了解人类蛋白互作组可以在疾病研究和药物发现中发挥重要的作用。这里描述的主动学习方法通过选择最富有信息量的蛋白质对进行标注来达到更高的准确度。该算法可以应用于选择候选蛋白质对，这些候选蛋白质对的相互作用的状态由实验确定，其有助于准确地预测其他的相互作用。这种方法可以帮助降低建立人类蛋白互作组的成本和精力，大大减少了用来确定蛋白质互作的一些新的体外实验。

参 考 文 献

[1] Mohamed T, Tarun S, Madhavi K G. An efficient heuristic method for active feature acquisition and its application to protein-protein interaction prediction[J]. BMC Proceedings, 2012, 6(Suppl 7): S2

[2] Deane C M, Salwinski L, Xenarios I, et al. Protein interactions: two methods for assessment of the reliability of high throughput observations[J]. Mol Cell Proteomics, 2002, 1(5): 349-356

[3] von Mering C, Krause R, Snel B, et al. Comparative assessment of large-scale data sets of protein-protein interactions[J]. Nature, 2002, 417(6887): 399-403

[4] Ito T, Tashiro K, Muta S, et al. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins[J]. PNAS, 2000, 97(3): 1143-1147

[5] Jansen R, Yu H, Greenbaum D et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data[J]. Science, 2003, 302(5644): 449-453

[6] Qi Y, Bar-Joseph Z, Klein-Seetharaman J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction[J]. Proteins, 2006, 63(3): 490-500

[7] Lin N, Wu B, Jansen R, et al. Information assessment on predicting protein-protein interactions[J]. BMC Bioinformatics, 2004, 5: 154

[8] DeBarr D, Wechsler H. Spam Detection using Clustering, Random Forests, and Active Learning[C]// Sixth Conference on Email and Anti-Spam. Mountain View, California, 2009

[9] Tuia D, Ratle F, Pacifici F, et al. Active learning methods for remote sensing image classification[J]. IEEE Trans. Geosci. Remote Sens., 2009, 47(7): 2218-2232

[10] Dagan I, Engelson S. Committee-based sampling for training probabilistic classifiers[C]// Proceedings of the 12th International Conference on Machine learning. 1995: 150-157

[11] 韩光，赵春霞，胡雪蕾. 一种新的SVM主动学习算法及其在障碍物检测中的应用[J]. 计算机研究与发展，2009，46(11)：15-20

[12] Tang M, Luo X, Roukos S. Active learning for statistical natural language parsing[C]// ACL 2002. Philadelphia, PA, USA 2002

[13] Shen X, Zhai C. Active Feedback in Ad Hoc Information Retrieval[C]// 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05). 2005: 59-66

[14] Campbell C, Cristianini N, Smola A. Query Learning with Large Margin Classifiers[C]// Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000). Morgan Kaufman, 2000

[15] Chang C C, Lin C J. LIBSVM. A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2: 1-27

[16] Qi Y, Klein-Seetharaman J, Bar-Joseph Z. A mixture of feature experts approach for protein-protein interaction prediction[J]. BMC Bioinformatics, 2007, 8(Suppl 10): S6

[17] Tong A H, Lesage G, Bader G D, et al. Global mapping of the yeast genetic interaction network[J]. Science, 2004, 303(5659): 808-813

[18] SPSS Inc. IBM SPSS Statistics 20 Brief Guide. pdf5