半监督多标记学习的基因功能分析

陈晓峰[1] ,王士同[1] ,曹苏群[1,2]
(1. 江南大学 信息工程学院 ,江苏 无锡 214122 ;2. 淮阴工学院 机械系 ,江苏 淮安 223001)

摘 要:传统的机器学习主要解决单标记学习 ,即一个样本仅有一个标记. 在生物信息学中 ,一个基因通常至少具有一个功能 ,即至少具有一个标记 ,与传统学习方法相比 ,多标记学习能更有效地识别生物相关基因组的功能. 目前的研究主要集中在监督多标记学习算法. 然而 ,研究半监督多标记学习算法 ,从已标记和未标记的基因表达数据中学习 ,仍然是未解决问题. 提出一种有效的基因功能分析的半监督多标记学习算法 SML_SVM. 首先 ,SML_SVM 根据 PT4 方法 ,将半监督多标记学习问题转化为半监督单标记学习问题 ,然后根据最大后验概率原则 (MAP) 和 K 近邻方法估计未标记样本的标记 ,最后 ,用 SVM 求解单标记学习问题. 在 yeast 基因数据和 genbase 蛋白质数据上的实验表明 ,SML_SVM 性能比基于 PT4 方法的 MLSVM 和自训练 MLSVM 更优.

关键词:半监督 ;多标记 ;自训练 ;支持向量机
中图分类号: TP181 **文献标识码**: A **文章编号**:1673-4785(2008)01-0083-08

Gene function analysis of semi-supervised multi-label learning

CHEN Xiao-feng[1] , WANG Shi-tong[1] , CAO Su-qun[1,2]
(1. School of Information Technology , Jiangnan University , Wuxi 214122 , China ; 2. Department of Mechanical Engineering , Huaiyin Institute of Technology , Huai'an 223001 ,China)

Abstract: Conventional machine learning is used only for single label learning , implying that every sample has only one label. However , in bioinformatics , a gene has more than one function , so it needs more than one label. Therefore , multi-label learning is more effective for identifying gene groups than conventional learning approach. Current research mainly focuses on supervised multi-label learning. The problem of effective semi-supervised multi-label learning strategies for labeled examples and unlabeled examples of gene expression datasets still remains unsolved. In this paper , a semi-supervised multi-label learning algorithm , named SML_SVM , is presented as an effective multi-label learner for analysis of gene expressions with at least one function. First , the proposed SML_SVM algorithm transforms the semi-supervised multi-label learning into corresponding semi-supervised single-label learning by the PT4 method , then it labels unlabeled examples using the maximum a posteriori (MAP) principle in combination with the K-nearest neighbor method , and finally , it solves the corresponding single-label learning problem using SVM. The distinctive characteristic of the proposed algorithm is its efficient integration of SVM-based single-label learning with MAP and K-nearest neighbor methods. Experimental results with a real Yeast gene expression dataset and a Genbase protein dataset show that the proposed SML_SVM algorithm outperforms the PT4-based MLSVM method and self-training MLSVM.

Keywords: semi-supervised ; multi-label ; self-training ; support vector machine

基因功能预测是生物学的重要任务 ,它有助于理解细胞的分子生物机制. 随着 DNA 微序列技术的发展 ,生物学家可以同时监测成千上万的基因. 微序列技术的使用 ,产生了大量的基因表达数据. 早期研究中 ,通常用无监督聚类方法分析基因表达数据 ,如层次聚类[1] 、自组织映射[2] 和基于 SSMCL 的 OPTOC[3] 等. 这些聚类算法假定相似的基因表达数

收稿日期:2007-04-13.
基金项目:国家"863"基金资助项目(2006AA10Z313) ;国家自然科学基金资助项目(60773206/ F020106 ,60704047/ F030304) ;国防应用基础研究基金资助项目(A1420461266) ;教育部跨世纪优秀人才支持计划基金资助项目(NCET-04-0496) ;教育部科学研究重点基金资助项目(105087).
通讯作者:王士同. wxwangst @yahoo. com. cn.

据仅有一个相似的功能. 非监督的聚类算法的优点是在训练中不需要先验知识. 然而,如果获取到基因表达数据的先验信息或功能信息,且有些基因同时具有若干种功能,在这种情况下,非监督聚类算法不是基因功能预测的最好选择.

如果将基因的功能视为标记,则在传统学习中,每个基因表达数据样本属于一个类,即一个样本有且仅有一个标记. 在很多真实世界问题中,一个样本可能同时属于多个类,即样本有多个标记. 例如,在文本分类中,每篇文档会同时属于多个主题,文档的内容同时涉及多个方面,如"政府"和"健康"[4-5]. 在生物信息学中,一个基因序列具有若干个功能,如"新陈代谢"和"蛋白质合成"[6]. 在语义场景分类中,场景图片会属于多个类别,如"沙滩"和"日出"[7]. 在音乐分类中,乐曲同时属于"摇滚"和"民谣"[8]. 研究人员提出了多标记学习算法来解决上述问题.

在基因功能预测方面,获得已标记样本的代价比较高,一方面是因为需要较多的人力参与,另一方面是因为样本数量急剧增长,大规模的标定样本非常困难. 由于 DNA 测序的自动化,使得生物序列数据库的数量以指数方式急剧增长,而基因功能分析的速度没有大的变化,不能满足应用需求. 在这种情况下,用于基因序列功能预测的已标记样本远小于未标记样本. 近年来,研究人员提出了监督的多标记学习算法,在监督多标记学习中,不考虑未标记样本的内在信息. 与监督学习相比,半监督学习同时使用已标记数据和未标记数据,提高了学习器的性能,在性能上具有一定优势[9]. 将半监督学习引入多标记学习,可以降低多标记学习在应用中的成本,使得在仅需人力处理少量样本的情况下,得到比监督多标记学习更好的效果.

针对上述问题,文中提出一种半监督多标记支持向量算法(semi-supervised multi-label support vector machine,SML_SVM),并给出解决该问题的策略. SML_SVM 先用 PT4 策略把半监督多标记学习问题转化为半监督单标记学习问题,然后用基于后验概率最大原则对未标记样本进行标记,再用 SVM 求解每个单标记学习问题.

1 多标记学习

设训练集为 $T=\{(x_1,Y_1),\cdots,(x_i,Y_i),\cdots,(x_m,Y_m)\}$ $(x_i\in X, Y_i\subseteq Y)$,其中 X 为输入空间,$Y=\{1,2,\cdots,Q\}$ 为有限个标记的集合. 多标记学习从训练集中构造多标记学习器 $h: X\to 2^Y$. 在一般情况下,多标记学习器学习实值函数 $f: X\times Y\to R$,如果样本 x_i 的标记集为 Y_i,对于属于 Y_i 的标记,实值函数 f 的输出值较大,如 $y_1\in Y_i$, $y_2\in Y_i$,则 $f(x_i,y_1)>f(x_i,y_2)$[10].

实值函数 $f(\cdot)$ 可以转化为排列函数 $\mathrm{rank}_f(\cdot)$,它将函数 $f(x_i,y)$ 的输出映射到 $\{1,2,\cdots,Q\}$,如果 $f(x_i,y_1)>f(x_i,y_2)$,则 $\mathrm{rank}_f(x_i,y_1)<\mathrm{rank}_j(x_i,y_2)$.

求解多标记学习问题的策略分两类:1)将多标记学习转化为单标记学习;2)将传统算法改造为能处理多标记问题的算法[11].

将多标记学习转化为单标记学习,有 6 种策略[11]:1)主观地或随机地选择多标记样本的某一个标记为训练标记,而丢弃该样本的其他标记,记为 PT1;2)丢弃训练集的所有多标记样本,仅保留单标记样本,记为 PT2;3)将具有相同标记的多标记样本组成一个新单标记类,记为 PT3;4)训练 $|Y|$ 个分类器 $H_{l_{ab}}: X\to\{l_{ab},\neg l_{ab}\}$,其中每个分类器 $H_{l_{ab}}$ 将样本分为 $\{l_{ab},\neg l_{ab}\}$ $(l_{ab}\in Y)$,记为 PT4;5)根据样本的标记,将所有样本分为 Q 类单标记数据集,即将样本 (x_i,Y_i) 分解为 $|Y_i|$ 个样本 (x,l_{ab}) $(l_{ab}\in Y_i)$,然后学习 Q 个基于覆盖的单标记分类器,记为 PT5;6)将样本 (x_i,Y_i) 分解为 $|Y|$ 个样本 $(x_i,l_{ab},Y[l_{ab}])$,如果 $l_{ab}\in Y_i$,则 $Y[l_{ab}]=1$,否则 $Y[l_{ab}]=-1$,记为 PT6. 其中,策略 PT1 和 PT2 会丢失多标记信息,很少使用.

研究人员还将传统算法做一定的修改,使之能处理多标记问题,如变体的 C4.5 算法[12]、Adaboost.MH 和 Adaboost.MR[5]、ML-kNN[10,13]、SVM[6,14-15]、BP_MLL[16]、Boosting[17]等. 在半监督学习方面,文献[18]提出一种半监督多标记学习框架,根据输入空间和输出空间的相似性,把半监督多标记学习转化为 NMF 问题来求解. 文献[19]用 Bayes 语义模型和 EM 算法解决 Web 文本挖掘问题,文献[20]提出文本分类的 TFIDF-NB 协同训练算法,这 2 种算法实质上可以被认为是半监督多标记算法.

衡量多标记数据集性质的 2 个指标,标记均值和标记密度定义如下:

标记均值为数据集的样本平均标记数,按下式计算:

$$\mathrm{LC}(T)=\frac{1}{m}\sum_{i}^{m}|Y_i|. \tag{1}$$

标记密度为数据集样本的标记数除以 $|Y|$ 的平

均值,按下式计算:

$$\mathrm{LD}(T) = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i|}{Y}. \tag{2}$$

设测试集为 $Z=\{(x_1,Y_1),(x_2,Y_2),\cdots,(x_r,Y_r)\}$,多标记学习算法的性能衡量指标如下:

1)汉明损失:测试样本的全体误分率,即不属于 i 类的样本被预测为 i 类,或者属于 i 类但没有被标记为 i 类. 汉明损失越小越好,计算公式如下:

$$h_{\mathrm{loss}}(h) = \frac{1}{r}\sum_{i=1}^{r}\frac{1}{Q}|h(x_i)\Delta Y_i|. \tag{3}$$

式中:Δ 表示 2 个集合差异,h 为多标记学习器.

2)一类错误:假如多标记分类器对样本 x_i 求出的排序最高的标记不在 x_i 对应的 Y_i 中,则称为一类错误,表示为 one-error(f),该值越小越好,计算公式如下:

$$\text{one-error}(f) = \frac{1}{r}\sum_{i=1}^{r}[[\operatorname{argmax}_{y\in Y_i} f(x_i,y)]]\notin Y_i]. \tag{4}$$

3)覆盖率:平均需要将标记序列下降多少可以覆盖样本对应的所有标记,表示为 coverage(f). 该值越小越好,其计算公式如下:

$$\mathrm{coverage}(f) = \frac{1}{r}\sum_{i=1}^{r}\max_{y}\operatorname{rank}_f(x_i,y) - 1. \tag{5}$$

4)排列损失:样本标记排列次序的平均错误,表示为 $r_{\mathrm{loss}}(f)$,该值越小越好,计算公式如下:

$$r_{\mathrm{loss}}(f) = \frac{1}{r}\sum_{i=1}^{r}\frac{1}{|Y_i||\bar{Y}_i|}|\{(y_1,y_2)\,|\,f(x_i,y_i)\le f(x_i,y_2),(y_1,y_2)\in Y_i\times\bar{Y}_i\}|. \tag{6}$$

式中:$\bar{Y}_i$ 表示 Y 的补集.

5)平均精度:多标记学习器预测样本的多标记是正确标记的平均比例,表示为 avgprecc(f),该值越大越好,计算公式如下:

$$\mathrm{avgprecc}(f) = \frac{1}{r}\sum_{i=1}^{r}\frac{1}{|Y_i|}\sum_{y\in Y_i}\frac{|\{y'\,|\,\operatorname{rank}_f(x_i,y')\le\operatorname{rank}_f(x_i,y),y'\in Y_i\}|}{\operatorname{rank}_f(x_i,y)}. \tag{7}$$

2　半监督多标记学习算法 SML_SVM

2.1　SML_SVM

设 $L=\{(x_1,Y_1),(x_2,Y_2),\cdots,(x_l,Y_l)\}$ 为已标记数据集,$U=\{x_{l+1},x_{l+2},\cdots,x_{l+u}\}$ 为未标记数据集. 其中,$x_i\in X(i=1,\cdots,l+u)$,X 为输入空间,$Y_i\subseteq Y(i=1,\cdots,l)$,$Y=\{1,2,\cdots,Q\}$ 为有限个标记的集合,l 为已标记数据集的样本数量,u 为未标记数据集的样本数量,一般来说,在半监督学习中,有 $l\ll u$. 为简便起见,文中假定在已标记数据集 L 中不存在标记缺失的情况,标记集 Y 中的所有成员都在 L 中出现,即 $\bigcup_{i=1}^{m}Y_i = Y$.

设 (x_i,Y_i) 为已标记样本,$Y_i=\{y_{i,1},\cdots,y_{i,t_i}\}$ $(t_i\ge 1)$,Y_i 为样本 x_i 对应的标记集,y_{i,t_1} 为 x_i 的第 t 个标记. 与传统的单标记学习不同的是,在多标记学习中,$t_i\ge 1$,样本 x_i 至少有一个标记. 多标记学习的一种解决方法是把多标记学习转化为单标记学习. 如前所述,有 6 种转化策略,其中,PT1 和 PT2 两种转化策略会丢失较多的多标记信息,不考虑使用,PT3 方法是通过把具有相同标记的多标记样本组成单标记数据集的方法转化的,往往会使新单标记数据集样本数量较少,在半监督多标记学习中,由于已标记样本远小于未标记样本,这种转化对学习是不利的,因此 PT3 策略不适合半监督多标记学习. 文中用 PT4 策略把多标记学习转化为单标记学习.

将 L 按照 PT4 策略转化,对于样本 (x_i,Y_i),首先根据标记集 Y_i 把 (x_i,Y_i) 分解成单标记样本集 $\Omega_i=\{(x_i,y_{i,1}),\cdots,(x_i,y_{i,t_i})\}$,$\Omega_i$ 中有 t_i 个单标记样本,这样多标记数据集 L 转化为 $L'=\bigcup_{i=1}^{l}\Omega_i=\{(x_1,y_{1,1}),\cdots,(x_1,y_{1,t_1}),\cdots,(x_i,y_{l,1}),\cdots,(x_l,y_{l,t_1})\}=\bigcup_{l_{ab}=1}^{Q}S_{l_{ab}}$,$S_{l_{ab}}$ 为 L 数据集中具有标记 l_{ab} 的样本的集合.$|Y|$ 表示标记集 Y 的样本数目. 根据 L 学习 $|Y|$ 个二分的分类器 $H_{lab}:X\to\{l_{ab},\neg l_{ab}\}$,其中每个分类器 $H_{l_{ab}}$ 按照样本是否具有标记 l_{ab} 将数据集 L 分为 $\{l_{ab},\neg l_{ab}\}$ $(l_{ab}=1,\cdots,Q)$ 两类.

通过上述的转化方式,半监督多标记学习问题被转化成了若干个半监督单标记学习问题. 根据近邻最大后验概率方法计算未标记样本被正确标记的概率,其方法叙述如下:

设在对标记 l_{ab} 分类的半监督单标记学习中,$L_{l_{ab}}=\{(x_1,y_1),\cdots,(x_l,y_l)\}$ 为已标记数据集,$U=\{x_{l+1},x_{l+2},\cdots,x_{l+u}\}$ 为未标记数据集. 在 $L_{l_{ab}}$ 中,如果 x_i 具有标记 l_{ab},则 $y_i=+1$,否则 $y_i=-1$. $N^+(x_i)$ 表示 x_i 在训练集的 K 个近邻里具有 +1 标记样本的集合,$N^-(x_i)$ 表示 x_i 的在训练集的 K 个近邻里具有 -1 标记样本的集合.

设未标记样本为 $\tilde{u}$,首先分别计算它在训练集的 K 个近邻中具有 +1 和 -1 标记的样本,记为

$N^+(\tilde{u})$和$N^-(\tilde{u})$,$|N^+(\tilde{u})|$和$|N^-(\tilde{u})|$表示$N^+(\tilde{u})$和$N^-(\tilde{u})$的样本数目.用H_1^-表示$\tilde{u}$具有标记-1,H_0^-表示$\tilde{u}$不具有标记-1,用H_1^+表示$\tilde{u}$具有标记+1,H_0^+表示$\tilde{u}$不具有标记+1.根据$N^+(\tilde{u})$和$N^-(\tilde{u})$计算$\tilde{u}$标记的最大后验概率的公式如下:

$$y_{\tilde{u}} = \operatorname{argmax}_{b\in\{0,1\},q\in\{+,-\}}\{P(H_b^q \| N^q(\tilde{u})|)\} = \operatorname{argmax}_{b\in\{0,1\},q\in\{+,-\}}\{\frac{P(H_b^q)P(|N^q(\tilde{u})\| H_b^q)}{P(|N^q(\tilde{u})|)}\} = \operatorname{argmax}_{b\in\{0,1\},q\in\{+,-\}}\{P(H_b^q)P(|N^q(\tilde{u})\| H_b^q)\}. \tag{8}$$

为计算$y_{\tilde{u}}$,需要从训练集中计算先验概率$P(H_p^q)(q\in\{+,-\}, b\in\{0,1\})$和后验概率$P(|N^q(\tilde{u})\| H_b^q)$.

在半监督单标记学习中,先用已标记数据集$L_{l_{ab}}$训练一个SVM分类器,然后对未标记数据集U进行标记,根据SVM分类器的分类结果,与分类超平面距离越远的点,越可能被正确分类,而在分类超平面附近的点,错分的可能比较大,也就是说,与分类超平面距离远的点分类结果的信任度高.从未标记数据集中选择具有最高信任度的样本组成候选未标记数据集U',再根据最大后验概率原则,把U中被正确标记的概率最大样本取出,组成已标记集$L'_{l_{ab}}$.下一轮迭代的训练集为$L\cup L'$.循环该过程,直到达到算法的最大迭代次数.SML_SVM算法步骤如表1所示.

2.2 计算复杂度

SML_SVM算法的复杂度与SVM的复杂度紧密相关,然而,SVM的不同优化求解方法的时间复杂度和空间复杂度差异较大,因此直接计算SML_SVM的复杂度并不方便.根据文献[21],文中通过计算SML_SVM训练中所有样本的总数来衡量算法的复杂度.定理1表明,SML_SVM算法与标记样本与未标记的数量是线性关系而不是指数关系.

定理1 SML_SVM算法的训练中的样本复杂度为$O(\text{MaxIter}\cdot|Y|\cdot(l+u))$,其中MaxIter是最大迭代次数,$|Y|$为训练集中的标记类别数,$l$和$u$分别是已标记样本数量和未标记样本数量.

证明 SML_SVM将半监督多标记学习问题转化为$|Y|$个半监督单标记学习问题.首先计算任意一个半监督单标记迭代学习中的样本总数.设l_{ab}为标记集Y中的任意1个标记,在第1次迭代训练中,l和u分别为已标记样本和未标记样本的数量.在第1次迭代后,up_1p_2个未标记样本被标记,且加入到已标记样本中,且up_1个样本将从未标记样本集U中删除.因此,在第2次迭代中,已标记样本数为$l+up_1p_2$,未标记样本数为$u(1-p_1)$.以此类推,在第i次迭代中,已标记样本数和未标记样本数分别为$l+up_1p_2+up_1p_2(1-p_1)+\cdots+up_1p_2(1-p_1)^{i-2}=l+up_2(1-(1-p_1)^{i-1})$和$u(1-p_1)^{i-1}$.由此可知,任意一个半监督单标记迭代学习中,训练样本总数为$\sum_{i=1}^{\text{MaxIter}}(l+up_2(1-(1-p_1)^{i-1}))=\text{MaxIter}(l+up_2)-up_2\frac{1-(1-p_1)^{\text{MaxIter}}}{p_1}$.由于SML_SVM将半监督多标记学习问题转化为$|Y|$个半监督单标记学习问题,在SML_SVM中,参与训练的样本总数为$|Y|(\text{MaxIter}(l+up_2)-up_2\frac{1-(1-p_1)^{\text{MaxIter}}}{p_1})$.由于$|Y|(\text{MaxIter}(l+up_2)-up_2\frac{1-(1-p_1)^{\text{MaxIter}}}{p_1})<|Y|\text{MaxIter}(l+up_2)<|Y|\text{MaxIter}(l+u)$,故SML_SVM算法的训练样本复杂度为$O(\text{MaxIter})\cdot|Y|(l+u)$.

表1 SML_SVM算法步骤

Table 1 Steps of SML_SVM algorithm

SML_SVM(L, U, K, MaxIter, p_1, p_2)

输入参数:

L:已标记数据集;

U:未标记数据集;

K:近邻数;

MaxIter:半监督训练最大迭代次数;

p_1:从U中选择构建U'的未标记样本比例;

p_2:从U中选择构建L'的未标记样本比例.

步骤:

1)把多标记数据集L根据PT4策略转化为单标记数据集L';

2)根据训练若干个单标记SVM分类器,构成多标记分类器SML_SVM;

3) Iter←1;

4) while (Iter <= MaxIter)

a)用SML_SVM对U标记;

b)从U计算标记信任度最高的p_1比例的样本,组成U';

c)计算U中样本标记的最大后验概率,取p_2比例的后验概率最高的样本组成L';

d)跟据$L\cup L'$,训练新的SML_SVM;

e) Iter←Iter+1;

5) end while.

输出:

半监督多标记学习器SML_SVM

3　基因功能分析

3.1　实验设置

文中实验主要是在酵母菌基因数据集和 genbase 蛋白质功能数据集上进行的. ML SVM (multi-label support vector machine) 是基于 PT4 策略的多标记支撑向量机算法,它是监督学习算法[11]. 在 ML SVM 中,仅使用已标记数据训练多标记分类器,未标记数据不参与训练,这就意味着未标记数据不能用于提高分类器的精度. 在半监督学习中,自训练策略是一种常用的半监督学习方法[11]. 在自训练的半监督学习中,首先根据少量的已标记数据训练出分类器,然后用该分类器对未标记数据进行分类,再把符合预定准则(如信任度最高)的分类结果视为已标记数据加入到训练集中,根据更新后的训练集重新训练分类器,直到达到训练停止条件为止. 自训练 ML SVM 是基于自训练策略的半监督多标记支持向量算法.

ML SVM 是只根据已标记的多标记样本训练多标记支持向量机的算法,未标记的多标记样本并没有被利用. Self-training ML SVM 是基于自训练策略的半监督多标记支持向量算法,它同时使用已标记的多标记样本和未标记的多标记样本训练多标记支持向量机,在每轮训练中,具有最高信任度的样本会加入到训练集,反复训练,直到达到停止条件. 与 ML SVM 相比,SML_SVM 可以根据未标记样本提高分类器性能,而与 Self-training ML SVM 相比,SML_SVM 选择未标记样本加入训练集的策略是不同的,性能更优. 文中的实验中,将 SML_SVM 分别与 ML SVM 和 Self-training ML SVM 对比,考察 SML_SVM 的性能.

3.2　酵母菌基因功能分析

酵母菌基因数据集 (yeast saccharomyces cerevisiae) 是常用的多标记学习算法性能测试数据集[6,11,22]. 它分为训练集和测试集 2 个部分,其中,训练集的样本为 1 500 个,测试集的样本为 917 个,特征为 103 维,均为数值型,样本标记为 14 种,标记均值为 4.25,标记密度为 0.3. 图 1 给出了 yeast 基因功能分类第 1 层次及基因 YAL041W 的 4 个功能.

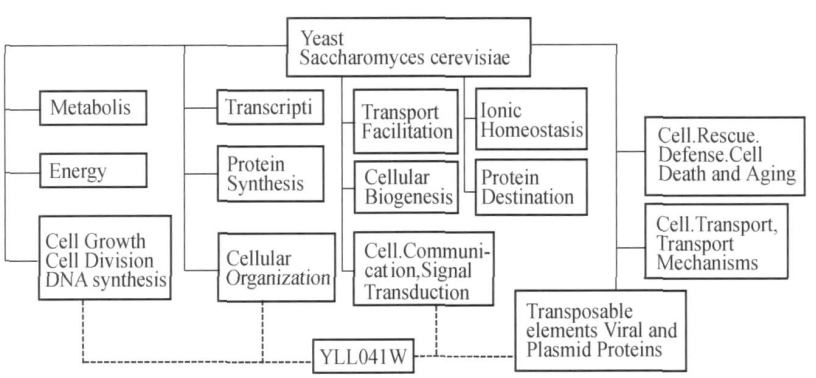


图 1　Yeast 基因功能分类第 1 层次,基因 YAL041W 有 4 个功能(粗框表示)

Fig. 1　The first level of hierarchy of the yeast gene functional classes, and gene YAL041W with four labels (shown with bold borders)

实验 1　对比 SML_SVM 与 ML SVM 和 Self-training ML SVM 的性能. 实验方法是从训练集中随机抽取 10 %,即 150 个样本,组成已标记数据集 L,即 $l = 150$,将训练集余下的 1 350 个样本和测试集 917 个样本组成未标记数据集 U,即 $u = 2\,267$. 实验结果的性能指标均是在 yeast 数据集的测试集上取得的. 在实验中,3 种算法的核函数均为高斯核,$C = 100$,$\sigma = 0.1$. SML_SVM 和 Self-training ML SVM 的最大均迭代次数均为 10 次,每轮训练均选取信任度最高的 10 %未标记样本组成候选未标记数据集 U'. SML_SVM 算法的近邻数 K 取 3. 实验重复 10 次,平均结果如表 2 所示,其中"↓"表示该性能指标越小越好,"↑"表示该性能指标越大越好.

实验 2　考察了 SML_SVM 算法的近邻数 K 对算法的影响. 实验中,SML_SVM 算法的已标记数据集、未标记数据集、核函数参数和最大迭代次数与实验 1 相同. 表 3 给出 K 在不同取值时 SML_SVM 算法性能指标的均值,其中 $K = 3$ 时的实验结

果已经在表2中给出,故不再重复.

表2 Yeast数据集的实验结果

Table 2 Experimental results of yeast dataset

	SML_SVM	SVM	Self-training MLSVM
Hamming Loss ↓	0.247 78 ±0.011	0.270 99 ±0.010	0.499 38 ±0.015
Ranking Loss ↓	0.238 9 ±0.015	0.261 46 ±0.017	0.497 29 ±0.024
One-error ↓	0.285 71 ±0.029	0.370 77 ±0.034	0.656 49 ±0.037
Coverage ↓	7.777 5 ±0.40	7.882 2 ±0.411	10.690 3 ±0.512
Average Precision ↑	0.700 81 ±0.022	0.661 4 ±0.021	0.435 96 ±0.023

表3 K的不同取值时的实验结果

Table 3 Experimental results with different K

K	2	4	5	6	7
Hamming Loss ↓	0.246 22	0.244 35	0.244 43	0.246 22	0.244 45
Ranking Loss ↓	0.238 76	0.246 8	0.241 26	0.238 78	0.241 26
One-error ↓	0.287 9	0.276 99	0.278 08	0.287 91	0.278 09
Coverage ↓	7.661 9	7.907 3	7.707 7	7.661 91	7.707 82
Average Precision ↑	0.700 84	0.696 47	0.700 8	0.700 85	0.700 80

实验3考察最大迭代次数对算法的影响.实验中其他参数均与实验1相同.为简洁起见,仅在表4给出 $K=3$,MaxIter取不同值时,SML_SVM算法性能指标的均值.

表4 K=3,MaxIter取不同值时的实验结果

Table 4 Experimental results with different MaxIter when K=3

Iteration	2	3	4	5	10
Hamming Loss ↓	0.247 55	0.247 7	0.246 61	0.246 24	0.247 78
Ranking Loss ↓	0.241 25	0.240 26	0.239 18	0.238 77	0.238 9
One-error ↓	0.303 16	0.283 53	0.288 99	0.287 91	0.285 71
Coverage ↓	7.659 8	7.681 6	7.657 6	7.662 1	7.777 5
Average Precision ↑	0.694 54	0.699 67	0.700 13	0.700 83	0.700 81

在表2中可以看出,在5个性能指标上,SML_SVM算法都有一定的提高.Self-training MLSVM 5个性能指标上都是最低的,在平均精度上甚至低于随机猜测.显然,Self-training MLSVM会把每次训练的错误累计到分类器中,不但不能提高性能,反而使性能严重下降.

表3的实验结果显示,近邻数 K 对SML_SVM的影响不大,较好的结果是在 $K=2$ 和 $K=5$ 时得到的,然而,在 K 取其他值时,SML_SVM变化仍然比较小.这说明SML_SVM对 K 是鲁棒的.

最大迭代次数对SML_SVM的影响比较大,从表4可以看出,当MaxIter较小的时候,算法的性能是差的,随着MaxIter的增加,性能逐渐变好,当达到一定限度后,增加最大迭代次数就不起作用了.事实上,最多迭代10次就可以达到最好的性能.这说明,SML_SVM对未标记样本数据集的内在信息的利用是有限度的,在最大后验概率准则下,避免由于引入未标记样本参与训练而带来的累计训练误差.

3.3 Genbase蛋白质功能预测分析

Genbase是生物蛋白质结构数据集[11,23].训练集有463个蛋白质样本,测试集有199个蛋白质样本.特征为1 185维,所有的属性均为离散的.在genbase蛋白质数据集中,共有27种标记,标记均值为1.35,标记密度为0.05.表5列出了若干蛋白质族和它们对应的功能,其中PDOC××××表示蛋白质族.

表5 蛋白质族及其对应功能

Table 5 Protein family and its functions

蛋白质族	功能
PDOC00064	氧还原酶
PDOC00154	异构酶
PDOC00224	细胞活素类和增长因子
PDOC00343	结构蛋白质
PDOC00561	受体
PDOC00662	DNA或RNA关联蛋白质
PDOC00670	转移酶
PDOC00791	蛋白质分泌和衍生物
PDOC50007	水解酶

在该实验中,参数设置同3.2节酵母菌基因功能分析实验相同.表6～8给出实验结果.表6表明,

SML_SVM 的性能比 MLSVM 和 Self-training MLSVM 更优,在 5 个指标上均达到最好. SML_SVM 在 K,MaxIter 参数上的实验结论与 3.2 节酵母菌基因功能分析实验相似,且最大迭代次数 MaxIter 对 SML_SVM 的影响比较大.

表 6　Genbase 数据集实验结果

Table 6　Experimental results of genbase dataset

	SML_SVM	MLSVM	Self-training MLSVM
Hamming Loss $\times 10^{-3}$ ↓	5.4545 ±2.4	6.4554 ±2.6	46.86 ±25.9
Ranking Loss $\times 10^{-3}$ ↓	4.454 ±3.6	8.4249 ±3.9	40.682 ±25.4
One-error $\times 10^{-2}$ ↓	3.1818 ±2.26	3.7771 ±2.28	20.503 ±8.44
Coverage ↓	0.54091 ±0.0854	0.59187 ±0.09441	0.07159 ±0.34
Average Precision ↑	0.95989 ±0.0379	0.928996 ±0.4520	0.59565 ±0.043

表 7　K取不同值时的实验结果

Table 7　Experimental results with different K

K	2	4	5	6	7	8
Hamming Loss $\times 10^{-3}$ ↓	6.1818	6.0455	6.1818	5.4545	6.0445	6.6364
Ranking Loss $\times 10^{-3}$ ↓	6.8832	6.3241	6.8832	4.454	6.3241	7.1352
One-error $\times 10^{-2}$ ↓	4.4318	4.43045	4.4318	3.1818	2.7045	2.8409
Coverage ↓	0.59773	0.57841	0.59773	0.58091	0.57841	0.60341
Average Precision ↑	0.93665	0.94361	0.93665	0.93989	0.94361	0.93365

表 8　K=3 时不同的实验结果

Table 8　Experimental results with different when K=3

Iteration	2	3	4	5	10
Hamming Loss $\times 10^{-3}$ ↓	5.6364	5.6182	5.5501	5.4876	5.4545
Ranking Loss $\times 10^{-3}$ ↓	6.8354	6.0274	5.2113	4.4421	4.454
One-error $\times 10^{-2}$ ↓	5.012	4.7126	4.6654	4.5211	4.4315
Coverage ↓	0.59091	0.58341	0.5673	0.5488	0.54091
Average Precision ↑	0.93078	0.93869	0.9425	0.9434	0.95989

4　结束语

文中提出了基因表达数据的半监督多标记学习问题,实现了半监督多标记支撑向量算法 SML_SVM. SML_SVM 首先使用 PT4 策略把半监督多标记学习问题转化为半监督单标记问题,然后用基于后验概率最大原则对未标记样本分类,通过迭代的方式求解每个半监督单标记学习问题. 实验表明,SML_SVM 比自训练 MLSVM 和 MLSVM 性能更好,提高多标记学习的性能. 在 yeast 基因功能分析和 genbase 蛋白质数据上的实验表明,SML_SVM 能利用未标记样本的信息,提高多标记学习的性能. SML_SVM 算法的不利之处是由于将多标记问题转化为若干个不相关的单标记问题,所以,各标记间的信息在算法中没有得到充分的利用,未来的工作是研究标记间信息对半监督多标记学习的影响.

参考文献:

[1] EISEN M B, SPELLMAN P T, BROWN P O, et al. Cluster analysis and display of genome-wide expression patterns[C]// Proceedings of the National Academy of Science of the United States of America. Washington, D. C, USA, 1998.

[2] TAMAYO P, SLONIM D, MESIROV J, et al. Interpreting patterns of gene expression with self-organizing maps[C]// Proceedings of the National Academy of Sciences of the United States of America. Washington, D. C, USA, 1999.

[3] WU S, LIEW A W C, YAN H, et al. Cluster analysis of gene expression data based on self-splitting and merging competitive learning[J]. IEEE Transactions on Information Technology in Biomedicine, 2004, 8(1): 5-15.

[4] MCCALLUM A K. Multi-label text classification with a mixture model trained by EM[C]// Working Notes of the AAAI'99 Workshop on Text Learning. Orlando, USA, 1999.

[5] SCHAPIRE R E, SINGER Y. Boostexter: a boosting-based system for text categorization[J]. Machine Learning, 2000, 39(2-3): 135-168.

[6] ELISSEEFF A, WESTON J. A kernel method for multi-labeled classification[C]// Advances in Neural Information Processing Systems 14. Cambridge: MIT Press, 2002.

[7] BOUTELL M R, LUO J, SHEN X, et al. Learning multi-label scene classification[J]. Pattern Recognition, 2004, 37(9): 1757-1771.

[8] OGIHARA L I T M. Detecting emotion in music[C]// Proceedings of the International Symposium on Music Information Retrieval. Maryland, USA: ISMIR Press, 2003.

[9] ZHU X J. Semi-supervised learning literature survey [R]. Department of Computer Sciences, University of Wisconsin, Madison, 2005.

[10]ZHANG M L, ZHOU Z H. ML-KNN: a lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7): 2038-2048.

[11]TSOUMAKAS G, KATAKIS I. Multi-label classification: an overview[J]. International Journal of Data Warehousing and Mining, 2007, 3(3):1-13.

[12]CLARE A, KING R D. Knowledge discovery in multi-label phenotype data[C]// Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2001). Freiberg, Germany: Springer, 2001.

[13]LUO X, ZINCIR H. Evaluation of two systems on multi-class multi-label document classification[C]// Lecture Notes in Computer Science. Freiberg, Germany: Springer, 2005.

[14]GODBOLE S, SARAWAGI S. Discriminative methods for multi-labeled classification[C]// Lecture Notes in Computer Science. Germany: Springer, 2004.

[15]ZHOU Z H, ZHANG M L. Multi-instance multi-label learning with application to scene classification[C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2007.

[16]ZHANG M L, ZHOU Z H. Multilabel neural networks with applications to functional genomics and text categorization[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(10): 1338-1351.

[17]施彤年, 卢忠良, 荣 融, 等. 多类多标签汉语文本自动分类的研究[J]. 情报学报, 2003, 22(3): 306-309.
SHI Tongnian, LU Zhongliang, RONG Rong, et al. Research on the Chinese text categorization of multi-classification and multi-label[J]. Journal of the China Society for Scientific and Technical Information, 2003, 22(3): 306-309.

[18]LIU Y, JIN R, LIU Y. Semi-supervised multi-label learning by constrained non-negative matrix factorization[C]// Proceeding of the Twenty-First National Conference on Artificial Intelligence, Eighteenth Conference on Innovative Applications of Artificial Intelligence. Boston: AAAI Press, 2006.

[19]宫秀军, 史忠植. 基于 Bayes 潜在语义模型的半监督 Web 挖掘[J]. 软件学报, 2002, 12(8):1508-1514.
GONG Xiujun, SHI Zhongzhi. Semi-supervised web mining based on bayes latent semantic model[J]. Journal of Software, 2002, 12(8): 1508-1514.

[20]彭 雅, 林亚平, 陈治平. TFIDF_NB 协同训练算法[J]. 小型微型计算机, 2004, 25(12): 2243-2246.
PENG Ya, LIN Yaping, CHEN Zhiping. TFIDF-NB co-operative training algorithm[J]. Mini-micro Systems, 2004, 25(12): 2243-2246.

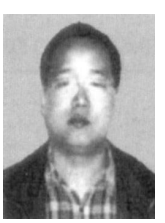
[21]KLAUS B, JOHANNS F, EYKE H. A unified model for multilabel classification and ranking[C]// Proceeding of the 15th Eureopean Conference on Artificial Intelligence. Riva del Garda, Italy: IOS Press, 2006.

[22]PAVLIDIS P, WESTON J, CAI J, et al. Combining microarray expression data and phylogenetic protellles to learn functional categories using support vector machines[R]. CUCS-011-000, Department of Computer Science, Columbia University, Columbia, 2000.

[23]DIPLARIS S, TSOUMAKAS G, MITKAS P, et al. Protein classification with multiple algorithms[C]// Lecture Notes in Computer Science. Volos, Greece: Springer, 2005.

作者简介:

陈晓峰,男,1977 年生,博士研究生,主要研究方向为机器学习、模式识别.



王士同,男,1964 年生,教授,博士生导师,主要研究方向为模糊人工智能、模式识别、图像处理和生物信息学等,先后十多次留学英国、日本和香港地区,在国内外重要杂志上发表学术论文数十篇.

曹苏群,男,1976 年生,博士研究生,主要研究方向为模式识别,图像处理、软件工程等.