湖南农业大学学报(自然科学版)　2015 年 2 月　第 41 卷 第 1 期
Journal of Hunan Agricultural University (Natural Sciences), Feb. 2015, 41(1):062–066

DOI:10.13331/j.cnki.jhau.2015.01.012
投稿网址：http://xb.hunau.edu.cn

一种基于树的蛋白质功能预测算法：KDE–CSSA

陈义明，贺细平，乔波

(湖南农业大学信息科学技术学院，湖南 长沙 410128)

摘　要：针对在每个标签类上直接学习分类模型计算代价高和树层次中低层结点训练数据扭曲的问题，提出了一种基于树层次的蛋白质功能预测算法：核依赖估计–压缩排序选择算法(KDE–CSSA)。该算法先将标签向量投影到标签核的主成分上，仅仅学习少量的回归模型，然后将预测的数值向量投影回原来标签向量空间，利用压缩排序和选择算法获取满足树属性的 0，1 标签向量。在 12 个基因组数据集上使用精确率和召回率作为评测标准的实验结果表明，KDE–CSSA 算法性能优于目前优秀的 CLUS–HMC 算法。

关　键　词：蛋白质；功能预测；主成分分析；核依赖估计；压缩排序与选择算法

中图分类号：TP391　　文献标志码：A　　文章编号：1007–1032(2015)01–0062–05

KDE–CSSA, a tree structure based algorithm for the prediction of protein function

Chen Yiming, He Xiping, Qiao Bo

(College of Information Science and Technology, Hunan Agricultural University, Changsha 410128, China)

Abstract: KDE–CSSD, an tree structure algorithm was proposed for the prediction of protein function based on class hierarchy to solve the issues of high computational cost on label classes through direct learning classification model and of train data skew on class hierarchy among middle or lower level nodes. The algorithm firstly projected label vector onto principle components of label kernel by means of learning less regression models, then, the predicted numeric vector were back projected onto their original vector space, finally, the predicted 0 or 1 label vector meeting tree hierarchy constraint were obtained using compressed sort and selection algorithm. The experiments, adopted precise rate and recall rate as criterion on 12 genomic benchmark data sets, proved that the KDE–CSSA algorithm outperformed the outstanding CLUS–HMC algorithm.

Keywords: protein; function prediction; principle component analysis; kernel dependency estimation(KDE); compressed sort and select algorithm(CSSA)

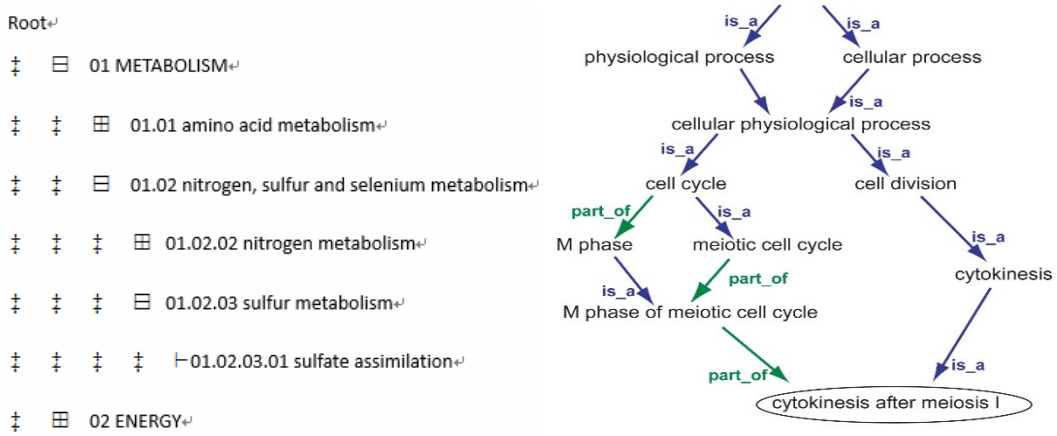
高通量的现代分子生物学实验产生了大量基因和蛋白数据，如基因和蛋白质序列、微阵列和蛋白质互作数据等。利用这些数据和已知蛋白质功能注释推断新蛋白质的功能，为生物学家提供实验参考已成为生物信息学研究的一项重要而紧迫的任务。笔者针对在每个标签类上直接学习分类模型计算代价高和树层次中低层结点训练数据扭曲的问题，提出了一种基于树层次的蛋白质功能预测算法：核依赖估计–压缩排序选择算法(KDE–CSSA)。

1　问题描述

生物学家已将各种蛋白质功能进行了分类整理，它们被组织成 1 种层次结构。典型的有 MIPS (Munich information center for protein sequences)数据库将所有功能组织成树层次结构[1]，而 gene ontology 则组织成有向无环图(DAG directed acyclic gaph)[2]结构。2 种组织结构分别如图 1–a,b 所示。一个蛋白质

收稿日期：2013–12–09　　修回日期：2014–06–25
基金项目：湖南省教育厅项目(11C0653)
作者简介：陈义明(1969—)，男，湖南南县人，博士，副教授，主要从事生物信息学研究，hunauchenym@qq.com

可能同时标记具有多个功能，笔者主要讨论基于树层次结构的蛋白质功能预测问题。



a　Funcat 的树层次　　b　GO 的有向无环图

图1　蛋白质功能的2种组织结构

Fig. 1　Two hierarchy structures of protein function

对每一个蛋白质，可以用实验所得的特征数据将它描述为一个特征向量 x_i，如果使用 d 个功能标签，则它的注释可以表示为一个 d 维 0，1 向量 $y_i \in \{0,1\}^d$。假设需要从 n 个已知蛋白质功能注释预测新的蛋白质功能，从机器学习的角度看，需要从训练数据 $\{(x_i, y_i)\}_{i=1}^n$ 学习分类预测模型，用来对未知蛋白质功能进行预测，机器学习领域称此为多标签分类问题。此外，预测结果还要满足树层次结构约束($\mathcal{T}$ –property)，即如果 1 个标签结点标记为 1，则它的所有祖先结点都应该标记为 1。

机器学习称满足 $\mathcal{T}$ –property 的多标签学习问题为层次多标签分类，是一个很具有挑战性的分类学习问题。蛋白质功能预测便属于这类问题，是生物信息学的一个研究热点。

Barutcuoglu 等[3]在每个标签类学习 1 个分类模型，得到预测结果后进行贝叶斯矫正，大量的标签类使得这种方法具有很高的计算复杂度。同时，低层的标签类具有很少的训练实例，有时正负实例数严重扭曲，影响了学习模型的质量，预测性能较低。CLUS–HMC 是目前性能最好的层次预测方法，是决策树算法的变种，它克服了逐个训练分类器和训练数据集扭曲的弊端[2]，但它使用的算法局限于决策树。

笔者提出一种基于树层次的蛋白质功能预测算法，即 KDE–CSSA，除了可以克服训练大量分类器的巨大计算量和扭曲的训练数据外，和 CLUS–HMC 相比，它能在 2 个阶段使用不同的机器学习模型，具有更好的灵活性。

2　KDE–CSSA 算法

2.1　核依赖估计(kernel dependency estimation，KDE)

在多标签分类问题中，y_i 包含 d 个标签。在蛋白质功能预测问题中，d 很容易达到成百上千个，对大量标签分别训练一个分类器是非常耗时的。使用 KDE 框架[4]可以将标签向量投影到一个低维空间，从而减少训练的分类器数。在多标签学习中，KDE 由以下 3 步组成：

1) 投影。将每个标签向量 y_i 投影到 1 个低维向量 $z_i \in \mathbb{R}^m$。

投影可以使用一些标准的技术来实现，如 Hsu 等[5]使用压缩感知(compressed sensing，CS)方法；Tai 和 Lin 等[6]使用主成分分析(principle component analysis，PCA)方法，并显示 PCA 在精度和效率方面都优于 CS。本研究使用 PCA 方法。同时，为了揭示标签向量间的非线性关系，并且保持标签之间的层次依赖，笔者在标签核上进行主成分分析(kernel principle component analysis，KPCA)，获取标签向量 y_i 的主成分。

定义标签树中结点 i 的特征向量为 $\boldsymbol{l}(i)=[l_1,\cdots,l_d]$，$l_k=1$，当且仅当 k 是 i 的祖先或者 $k=i$，这样的功能类表示方式能够很好地描述应该满足的 $\mathcal{T}$ –property。标签核使用简单的内积核

$\boldsymbol{K}(i,j)=\boldsymbol{l}(i)^{\mathrm{T}}\boldsymbol{l}(j)$，核矩阵 $\boldsymbol{K}$ 是一个 $d\times d$ 的矩阵。中心化核矩阵为 $\boldsymbol{K}'=(\boldsymbol{I}-\frac{1}{d}\boldsymbol{1}_d\boldsymbol{1}_d^{\mathrm{T}})\boldsymbol{K}(\boldsymbol{I}-\frac{1}{d}\boldsymbol{1}_d\boldsymbol{1}_d^{\mathrm{T}})$，其中 $\boldsymbol{I}$ 为 d 维单位矩阵，$\boldsymbol{1}_d$ 为 d 维全 1 列向量。解特征值问题 $\lambda\alpha=\boldsymbol{K}'\alpha$，得到所有特征向量 $\boldsymbol{\alpha}$，以最大的 m 个特征值对应的特征向量为行构成投影矩阵 $\boldsymbol{P}\in\mathbb{R}^{m\times d}$，则按照公式(1)可以计算投影到的 m 维向量 $\boldsymbol{z}_i$。

$$\boldsymbol{z}_i=\boldsymbol{P}y_i \tag{1}$$

2) 学习。令 z_{ij} 为 z_i 的第 j 维分量，$j=1,\cdots,m$，学习从 $\mathcal{X}=\{x\}$ 到 $\{z_{\_j}\}$ 的映射。因为这些投影方向是正交的，笔者独立地训练这 m 个映射模型。采用公式(2)所示的岭回归模型求线性回归的系数 β_j。

$$\beta_j=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}z_{\_j} \tag{2}$$

其中：k 为岭参数；$\boldsymbol{I}$ 是单位矩阵。根据回归模型在验证集上的误差选取合适的岭参数 k。

3) 预测。对于一个测试样本 x，首先使用 m 个学习到的模型获取预测 $\hat{\boldsymbol{z}}=[\hat{z}_1,...,\hat{z}_m]^{\mathrm{T}}$，然后使用公式(3) 将预测结果投影回原始标签空间 $\hat{\boldsymbol{y}}\in\mathbb{R}^d$。

$$\hat{\boldsymbol{y}}=\boldsymbol{P}^{\mathrm{T}}\hat{\boldsymbol{z}} \tag{3}$$

2.2 树结构上的预测(CSSA)

假设测试蛋白质实例 x 注释有 L 个功能标签，并且这些标签是无结构的，在上述预测的结果上，可以将 L 个最大分量的标签标记为 1，其余为 0，从而得到该实例的预测结果。这本质上是求解公式(4)表示的优化问题。

$$\begin{aligned}&\max_{\psi}\ \sum_{i=1}^{d}\widehat{y}_i\psi_i\\&s.t.\ \ \sum_{i=1}^{d}\psi_i=L\end{aligned} \tag{4}$$

其中，$\boldsymbol{\psi}=[\psi_1,\cdots,\psi_d]^{\mathrm{T}}$，且 $\psi_i\in\{0,1\}$。

当功能标签层次是一颗树 $\mathcal{T}$ 时，上述优化问题中的 ψ 还要满足前面定义的 $\mathcal{T}-$property：ψ 中那些位于从树根到树叶路径上的分量应该是非递增的，称这个约束叫 $\mathcal{T}$ –非递增。添加该约束条件后的优化问题如式(5)所示。

$$\begin{aligned}&\max_{\psi}\ \ \sum_{i\in\mathcal{T}}\widehat{y}_i\psi_i\\&s.t.\ \ \psi_i\in\{0,1\},\ \ \forall i\in\mathcal{T},\\&\text{其中，}\sum_{i\in\mathcal{T}}\psi_i=L\text{，}\psi\text{ 满足 }\mathcal{T}\text{–非递增}\end{aligned} \tag{5}$$

Baraniuk 和 Jones 使用贪婪算法高效求解该优化问题，被称为压缩排序选择算法(condensing sort and selecting algorithm CSSA)[7]。算法定义包含至少一个结点的集合为超结点，它的 w 值为所有结点对应 $\widehat{y_i}$ 分量的平均。算法初始设置所有超结点都仅仅包含 1 个结点，它们的 w 值为包含结点的 $\widehat{y_i}$ 分量值。当 $\psi_i\in\{0,1\}$ 时，为了使得 $\sum_{i\in\mathcal{T}}\widehat{y_i}\psi_i$ 最大，每次选取 w 值最大，且将父亲结点已经标记为 1 的超结点标签置为 1。如果 w 值最大的超结点的父亲结点还没有标记为 1，则将父亲结点压缩进该超结点中，形成更大的超结点，重新计算该超结点的 w 值。该过程迭代执行，直到被选取的结点数超过 L 为止。CSSA 算法步骤：

初始化：ψ_0 =1；sum=0。
初始化其他所有结点为超结点，$w_i=\widehat{y_i}$，置 ψ_i =0，i=1,⋯,d，按照 w 值排序
while sum<L do
选取有最大 w 值且还没有标记的超结点 $S^{\star}$
If $\psi(pa(S^{\star}))$ =1 then// $pa(S^{\star})$ 表示 $S^{\star}$ 的父亲结点
$\psi(S^{\star})$ =1 //将 $S^{\star}$ 中所有结点标记为 1
Sum=sum+ $n(S^{\star})$ // $n(S^{\star})$ 表示 $S^{\star}$ 中结点的数目
else
压缩 $S^{\star}$ 和 $pa(S^{\star})$ 作为一个新的超结点，计算它的 w 值
end if
end while

CSSA 算法已经被成功用于小波近似[8]，最近被用于基于模型的压缩感知[9]，它的时间复杂度为 $\mathcal{O}(\mathcal{N}\log\mathcal{N})$。

2.3 蛋白功能的树层次预测算法

在上面的方法中，笔者先将 d 维的标签向量 y_i 投影到低维空间 $\mathbb{R}^m$，在低维空间学习回归模型，减少学习的模型数，同时提高学习模型的质量。然后投影回 d 维空间，使用 CSSA 算法得到预测标签。

蛋白功能的树层次预测算法：

1) 投影。在标签核上执行 KPCA，得到特征值最大的 m 个特征向量，组成投影矩阵 $\boldsymbol{P}$。
2) 学习。For j=1,⋯,m do
使用 $\{(x_i,(\boldsymbol{P}y_i)_j)\}_{i=1}^{N}$，学习第 j 个回归模型
end for
3) 预测。从 m 个回归模型得到 $z\in\mathbb{R}^m$，使用 $\hat{\boldsymbol{y}}=\boldsymbol{P}^{\mathrm{T}}z$ 投影回 d 维空间，由 CSSA 算法得到预测标签。

3　验证实验

3.1　实验设置

笔者使用 Clare[10]发布的 12 个功能基因组数据，每个数据描述酵母基因组的不同方面，有 2 个不同的输出版本，即 MIPS 的树结构标签和 Gene Ontology 的 DAG 结构标签，其中，每个数据集的 2/3 用来训练，剩余的用来测试。在训练数据中，2/3 用来训练，其余 1/3 用于验证确定参数。数据集如表 1 所示。在实验数据集上，比较 KDE–CSSA 和 CLUS–HMC 算法。

表1　实验使用的酵母数据集

Table 1　The summary of yeast data set used in the experiment

数据集	训练实例数	验证实例数	测试实例数	属性数	FunCat 功能数
Seq	1 692	876	1 332	478	500
Phenol	653	352	581	69	456
Struc	1 659	859	1 306	19 629	500
Hom	1 661	876	1 309	47 035	500
Cellcycle	1 625	848	1 278	77	500
Church	1 627	844	1 278	27	500
Derisi	1 605	842	1 272	63	500
Eisen	1 055	528	835	79	462
Gasch1	1 631	846	1 281	173	500
Gasch2	1 635	849	1 288	52	500
Spo	1 597	837	1 263	80	500
expr	1 636	849	1 288	551	500

3.2　性能度量

对于两类分类问题，精确率 *Prec* 和召回率 *Rec* 的计算公式如式(6)、(7)所示。

$$Prec = \frac{TP}{TP + FP} \tag{6}$$

$$Rec = \frac{TP}{TP + FN} \tag{7}$$

TP，*FP* 和 *FN* 分别为真阳性数、假阳性数和假阴性数。在多标签分类问题中，设 TP_i、FP_i和 FN_i 分别表示第 *i* 个标签的真阳性数、假阳性数和假阴性数，精确率和召回率定义如式(8)、(9)所示。

$$Prec = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FP_i} \tag{8}$$

$$Rec = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FN_i} \tag{9}$$

在不同的阈值下，可以得到不同的精确率和召回率。一般来说，召回率越低，则精确率越高；反之，则精确率越低。为了衡量整体性能，通常绘出精确率－召回率曲线，利用曲线下的面积 AUPRC (area under the PR curve)值来比较算法性能，AUPRC 值越大，则模型性能越好[11]。

3.3　验证结果

在 12 个数据集上，KDE–CSSA 对 CLUS–HMC 的 AUPRC 值比较结果如表 2 所示。

由表 2 可以看出，就通常的度量标准 AUPRC 值来说，KDE–CSSA 在 12 个数据集上的性能一致优于 CLUS–HMC。

表2　在 酵母数据集上预测的AUPRC值

Table 2　Comparison the predicted AUPRC between KDE–CSSA and CLUS–HMC using yeast data sets

数据集	AUPRC 值	
	KDE–CSSA	CLUS–HMC
Seq	0.228	0.216
Phenol	0.182	0.167
Struc	0.195	0.187
Hom	0.261	0.252
Cellcycle	0.195	0.181
Church	0.189	0.175
Derisi	0.196	0.182
Eisen	0.224	0.215
Gasch1	0.226	0.212
Gasch2	0.228	0.201
Spo	0.218	0.194
expr	0.226	0.214

4　结　论

本文提出了一种基于树层次的蛋白质功能预测算法(KDE–CSSA)，克服了逐个训练大量分类器带来的巨大计算量和训练数据扭曲影响预测性能的问题，和目前最好的算法 CLUS–HMC 相比，有更好的预测性能，且能够在每个阶段更灵活地运用其他方法。提出的算法可以运用到类似树层次结构

的分类预测中，如文档和图片分类等。将来的研究目标是将这种算法扩展到 DAG 层次结构上，解决在 GO 上蛋白质功能预测的问题。

参考文献：

[1] Ruepp A，Zollner A，Maier D，et al．The funcat，a functional annotation scheme for systematic classification of proteins from whole genomes[J]．Nucleic Acids Research，2004，32(18)：5539–5545．

[2] Schietgat L，Vens C，Struyf J，et al．Predicting gene function using hierarchical multi–label decision tree ensembles[J]．BMC Bioinformatics，2010，11(1)：1–14．

[3] Barutcuoglu Z，Schapire R E，Troyanskaya O G. Hierarchical multi-label prediction of gene function[J]. Bioinformatics，2006，22(7)：830–836．

[4] Weston J，Chapelle O，Vapnik V，et al．Kernel dependency estimation[C]//Advances in Neural Information Processing Systems，2002：873–880．

[5] Hsu D，Kakade S，Langford J，et al．Multi-label prediction via compressed sensing[C]//Advances in Neural Information Processing Systems 22，2009：772–780．

[6] Tai F，Lin H T．Multilabel classification with principal label space transformation[J]．Neural Computation，2012，24(9)：2508–2542．

[7] Baraniuk R，Jones D．A signal-dependent time-frequency representation：fast algorithm for optimal kernel design[J]．Signal Processing，IEEE Transactions on，1994，42(1)：134–146．

[8] Baraniuk R D．Optimal tree pproximation with wavelets [C]//SPIE's International Symposium on Optical Science，Engineering，and Instrumentation，International Society for Optics and Photonics，1999：196–207．

[9] Baraniuk R G，Cevher V，Duarte M F．Model-based compressive sensing[J]．IEEE Transactions on Information Theory，2010，56(4)：1982–2001．

[10] Clare A．Machine learning and data mining for yeast functional genomics[D]．Wales：The University of Wales，2003．

[11] Gene Ontology Consortium．The gene ontology project in 2008[J]．Nucleic Acids Research，2008，36(Suppl.1)：440–444．

责任编辑：苏爱华
英文编辑：王 库

（上接第 52 页）

[15] Banerjee J，Maiti M K．Functional role of rice germin- like protein1 in regulation of plant height and disease resistance[J]．Biochemical and Biophyscial Research Communications，2010(1)：178–183．

[16] 王铖．苔藓植物在园林绿化中的应用[J]．园林，2011(9)：48–53．

[17] 陈俊和，蒋明，张力．苔藓植物园林景观应用浅析 [J]．广东园林，2010(1)：31–34．

[18] 宋晓宏，沙伟，林琳，等．毛尖紫萼藓干旱胁迫 cDNA 文库的构建(英文)[J]．植物研究，2010，30(6)：713–717．

[19] 沙伟，张梅娟，刘博，等．毛尖紫萼藓抗旱相关基因 *Gp–LEA* 的克隆与表达分析[J]．西北植物学报，2013，33(9)：1724–1730．

[20] Woo E J，Dunwell J，Goodenough P W，et al．Germin is a manganese containing homohexamer with oxalate oxidase and superoxide dimutase activities [J]．Natural Structural Biology，2000，7(11)：1036–1040．

[21] Swart S，Logman T J J，Smit G，et al．Purification and partial characterization of a glycoprotein from pea (*Pisum sativum*) with receptor activity for rhicadhesin，an attachment protein of *Rhizobiaceae*[J]．Plant Mol ecular Biology，1994，24(1)：171–183．

[22] 李红丽，刘迪秋，何华，等．类萌发素蛋白在植物防卫反应中的作用 [J]．植物生理学报，2013，49(4)：331–336．

[23] Membré N，Bernier F，Staiger D，et al．Arabidopsis thalianagermin-like protein：Common and specific features point to a variety of functions[J]．Planta，2000，211(3)：345–354．

[24] Komatsu S，Kobayashi Y，Nishizawa K，et al. Comparative proteomics analysis of differentially expressed proteins in soybean cell wall during flooding stress [J]．Amino Acids，2010，39(5)：1435–1449．

[25] Li H Y，Jiang J，Wang S，et al．Expression analysis of *ThGLP*，a new germin-like protein gene，in *Tamarix hispida* [J]．Journal of Forestry Research，2010，21(3)：323–330．

[26] 刘超，张林华，邱红林，等．天山雪莲 *SikGLP* 基因的克隆及表达分析[J]．植物研究，2013，33(4)：461–467．

[27] Lu M，Han Y P，Gao G J，et al．Identification and analysis of the germin-like gene family in soybean[J]．BMC Genomics，2010，11：620–634．

[28] 晋文娟，张少杰，陈双臣，等．过表达 *Mdip1* 基因提高番茄抗低温诱导的氧化胁迫能力[J]．植物生理学报，2013，49(5)：493–500．

责任编辑：苏爱华
英文编辑：王 库