# 基于混合SVM方法的蛋白质二级结构预测算法

隋海峰 曲 武 钱文彬 杨炳儒
(北京科技大学信息工程学院 北京 100083)

**摘 要** 预测蛋白质二级结构,是当今生物信息学中一个难以解决的问题。由于预测蛋白质二级结构的精度在蛋白质结构研究中起到非常重要的作用,因此在基于KDTICM理论基础上,提出一种基于混合SVM方法的蛋白质二级结构预测算法。该算法有效地利用蛋白质的物化属性和PSI-SEARCH生成的位置特异性打分矩阵作为双层SVM的输入,从而大大地提高了蛋白质二级结构预测的精度。实验比较分析表明,新算法的预测精度和普适性明显优于目前其他典型的预测方法。

**关键词** 蛋白质二级结构预测,混合SVM方法,复合金字塔模型
**中图法分类号** TP391 **文献标识码** A

## Protein Secondary Structure Prediction Algorithm Based on Mixed-SVM Method

SUI Hai-feng QU Wu QIAN Wen-bin YANG Bing-ru
(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, China)

**Abstract** Protein secondary structure prediction is one of the most important problems in bioinformatics. The protein secondary structure prediction accuracy plays an important role in the field of protein structure research. In this paper, using a Knowledge Discovery Theory based on the Inner Cognitive Mechanism (KDTICM), an efficient protein secondary structure prediction algorithm based on mixed-SVM (support vector machine) approach was proposed. The algorithm makes full use of the evolutionary information contained in the physicochemical properties of each amino acid and a position- specific scoring matrix generated by a PSI-SEARCH multiple sequence alignment, secondary structure can be predicted at significantly increased accuracy. At last, the experiments were used to show the superior accuracy and generality of the new algorithm than other classical algorithm.

**Keywords** Protein secondary structure prediction, Mixed-SVM method, Compound pyramid model

## 1 引言

随着后基因组时代的发展,已知蛋白质序列的数量正在迅速增长。然而,不断增长的蛋白质序列的数量要远大于已测定的蛋白质结构,从而使得蛋白质序列和结构之间的差距越来越大。因此,需要利用有效的预测方法来进一步缩小这方面的差距。而且研究蛋白质的二级结构意义重大,分析蛋白质二级结构、功能及其关系是蛋白质组计划中的一个重要组成部分。研究蛋白质二级结构,有助于了解蛋白质的作用,了解蛋白质如何行使其生物功能,认识蛋白质与蛋白质(或其它分子)之间的相互作用,这无论是对于生物学还是对于医学和药学,都是非常重要的。

蛋白质二级结构在生物化学及结构生物学中是指一个生物大分子,如蛋白质及核酸(DNA或RNA)、局部区段的三维排布方式等。在蛋白质中,二级结构则是以主链中氨基之间的氢键模式来定义,即DSSP[1]所定义的氢键。由于蛋白质三级结构[2]还不能直接从序列中用生物学计算方法准确预测,为此利用二级结构预测的结果预测三级结构,通过它使复杂的3D结构映射到一维平面。同时,提高二级结构预测精度将会在蛋白质三级结构预测中起到非常重要的作用[3]。

目前,研究人员提出了许多有关蛋白质二级结构的预测方法,例如神经网络[4,5]、支持向量机[6]、隐马尔科夫模型[7]、数据挖掘[8]等。本文在基于KDTICM理论的基础上,首先提出了一种基于混合SVM方法(MSVM)的蛋白质二级结构预测算法,然后构造一个新的复合金字塔模型,以获得更高的预测精度。新算法有效地利用内含丰富的进化信息的蛋白质物化属性和PSI-SEARCH[9]生成的位置特异性打分矩阵作为双层SVM的输入,从而极大地提高了二级结构预测的精度。通过算法实验,在CB513测试数据集上,所有残基的$Q_3$精度达到85.58%,而SOV99精度达到81.15%。在CASP394盲测数据集上,所有残基的$Q_3$精度达到86.05%,而SOV99精度达到84.46%。实验结果表明,新算法用于预测CB513和CASP394序列时其$Q_3$和SOV99精度优于目前其他的方法,包括流行的Psipred[10]方法,同时表明新算法具有较好的普适性。

到稿日期:2010-11-07 返修日期:2011-02-23 本文受国家自然科学基金项目(60875029)资助。
隋海峰(1976-),男,博士生,主要研究方向为数据挖掘、金融数据分析、生物信息学;曲 武(1982-),男,博士生,主要研究方向为数据挖掘、生物信息学,E-mail:quwu.ustb@gmail.com(通信作者);钱文彬(1984-),男,博士生,主要研究方向为数据挖掘、生物信息学;杨炳儒(1943-)男,教授,博士生导师,主要研究方向为人工智能、数据挖掘和柔性建模。

## 2 相关知识

### 2.1 多序列比对图谱和编码方式

通过蛋白质序列的多序列比对图谱而不是仅依靠序列来预测蛋白质结构，是公认可提高预测精度的方法，因为已知结构的蛋白质同源性是预测蛋白质二级结构最可靠的方法。序列的多重比对反映了蛋白质家族的共同特征，提取了结构保守信息及蛋白质家族中特定的残基替换模式，同时多序列比对所携带的进化信息也表明了蛋白质进化过程中的相互作用。PSI-SEARCH 采用迭代搜索策略生成多序列比对图谱，其基本过程为：先用目标序列扫描数据库，找到一组序列，该序列又产生一组新的搜索序列图谱，然后再用这组搜索序列图谱查找新的序列。这种方式生成的位置特异性评分矩阵(PSSM)[11]包含丰富的进化信息，使二级结构预测精度有了较大的提高，尤其是对 β-折叠的预测精度。此外，氨基酸的物化属性对蛋白质的二级结构影响较大，因此在进行结构预测时需要考虑氨基酸的物化属性，如疏水性、带电性、氢键等，可根据氨基酸各方面的性质及残基之间的组合预测可能形成的二级结构。

因此，混合 SVM 方法的输入向量由 PSI-SEARCH 搜索蛋白质结构序列数据库(Protein Structure Sequences)输出的位置特异性评分矩阵(PSSM)以及蛋白质 4 个重要的物化属性(疏水、带电、氢键和巨大残基含量)构成。对于 PSI-SEARCH，期望值 $E$ 的阈值设置为 0.005，迭代次数是 5 次，搜索的目标数据库为蛋白质结构序列数据库。我们用 SEG[12]软件来覆盖低相似度序列区域、无规则卷曲区域和跨膜区域。对于冗余度在 75%以上的序列所有的比对内容都将被过滤，在前期的处理中发现此时可取得最优的结果。位置特异性打分矩阵有 $20*N$ 个元素，其中 $N$ 是目标序列的长度，每个元素代表某个特定残基被替换的可能性，该替换是对模板中给定的比对位置利用基于 BLOSUM80[13]打分矩阵的加权平均数进行替换。图谱矩阵中元素的取值范围为[−7，7]，利用函数 $x'=(x-m_i)/(M_i-m_i)$ 进行归一化操作，其中 $M_i$ 和 $m_i$ 分别代表第 $i$ 个属性的最大值和最小值。

与 PHD[14]编码方式一样，我们使用滑动窗口的方法。为了使窗口可以由 $N$ 端滑到 $C$ 端，为每个残基编码增加了第 21 个元素，用来标识滑动窗口是否在 $N$ 端或 $C$ 端。混合 SVM 的编码方式包括 PSSM 和上面提及的 4 个重要的物化属性，因此每个输入向量有 $25*w$ 个元素，其中 $w$ 代表滑动窗口的大小，这个窗口沿着蛋白质序列的残基滑动。

### 2.2 预测精度的评估

我们采用最广泛使用的 DSSP[1]来定义蛋白质二级结构。DSSP 标准将二级结构分为 8 类：$H$(α 螺旋)、$G$(3-螺旋)、$I$(π-螺旋)、$E$(β 折叠)、$B$(β 桥)、$T$(转弯)、$S$(弯曲)以及其他结构(一)。这 8 种结构可以合并为 3 类，本文采用 DSSP 中最严格的定义方法，即将 $H$、$G$ 合并为 $H$，$E$、$B$ 合并为 $E$，其他状态合并为 $C$。蛋白质二级结构预测的方法是以 α 螺旋($H$)、β 折叠($E$)和无规则卷曲($C$)这 3 种状态进行评估，其余的状态都归入这几个状态之一。对于二级结构预测精度的评估有很多方法，最常用的评价指标是 $Q_3$[15]和 SOV99[16]。

$Q_3$ 定义为整个预测的蛋白质中被正确预测的数量百分比，定义如下：

$$Q_3=\frac{\sum_{i\in\{H,E,C\}}\#\text{ of residues correctly predicted}_i}{\sum_{i\in\{H,E,C\}}\#\text{ of residues in class }i}\times 100 \quad (1)$$

评估标准只依赖于 3 种状态(α 螺旋、β 折叠和无规则卷曲)，因此被称为 $Q_3$。本文中采用每个残基的三态预测精度来衡量预测性能。

每个残基对于二级结构每种类型的预测精度($Q^H$，$Q^E$，$Q^C$，$Q_H^{pre}$，$Q_E^{pre}$，$Q_E^{pre}$)计算如下：

$$Q_i(\%)=\frac{\#\text{ of residues correctly predicted}_i}{\#\text{ of residues in class }i}\times 100 \quad (2)$$

$$Q_i^{pre}(\%)=\frac{\#\text{ of residues correctly predicted}_i}{\backslash\#\text{ of residues predicted}_i}\times 100 \quad (3)$$

式中，$i$ 的状态可能是 $H$、$E$、$C$。

片段交叠准确率评估(SOV99)是二级结构片段交叠而不是单个残基的二级结构预测结果评估方法。SOV99 计算如下：

$$SOV=100\times\left[\frac{1}{N}\sum_{i\in\{H,E,C\}}\sum_{S(i)}\frac{\min ov(s_1,s_2)+\delta(s_1,s_2)}{\max ov(s_1,s_2)}\times len(s_1)\right] \quad (4)$$

对于每种状态 $i$ 的片段交叠精度定义如下：

$$SOV(i)=100\times\frac{1}{N(i)}\left[\sum_{S(i)}\frac{\min ov(s_1,s_2)+\delta(s_1,s_2)}{\max ov(s_1,s_2)}\times len(s_1)\right] \quad (5)$$

式中，$s_1$ 和 $s_2$ 分别为观察到的二级结构片段和预测的二级结构片段；$len(s_1)$为片段 $s_1$ 中残基的数量；$\min ov(s_1,s_2)$为 $s_1$ 和 $s_2$ 实际交叠的长度，比如两个片段中都有状态为 $i$ 的残基组成的长度；$\max ov(s_1,s_2)$为在 $s_1$ 或 $s_2$ 中状态为 $i$ 的残基总的长度；$\delta(s_1,s_2)$是一个整数值，定义如下：

$$\delta(s_1,s_2)=\min\begin{Bmatrix}\max ov(s_1,s_2)-\min ov(s_1,s_2)\\ \min ov(s_1,s_2)\\ \text{int}(len(s_1)/2\\ \text{int}(len(s_2)/2)\end{Bmatrix} \quad (6)$$

式中，$N(i)$是状态为 $i$ 的残基个数，定义如下：

$$N(i)=\sum_{S(i)}len(s_1)+\sum_{S'(i)}len(s_1) \quad (7)$$

式中，正常的 $N$ 值为 3 种状态的 $N(i)$之和，即：

$$N=\sum_{i\in\{H,E,C\}}N(i) \quad (8)$$

$S(i)=\{(s_1,s_2):s_1\cap s_2\neq\emptyset\}$，$s_1$ 和 $s_2$ 都在状态 $i$ 的构象中。$S'(i)=\{s_1:\forall s_2,s_1\cap s_2=\emptyset\}$，$S'(i)$是 $s_1$ 和 $s_2$ 都在状态 $i$ 中的所有片段集合。

由于 SOV99 对偏 α 螺旋、β 折叠和无规则卷曲非常敏感，当蛋白质类型是偏 α 螺旋、β 折叠或无规则卷曲时，即使 $Q_3$ 精度比较高，SOV99 的精度也可能很低。

## 3 混合 SVM 方法

### 3.1 双层 SVM 结构

在预测子系统中，我们用到了双层 SVM 结构，结构如图 1 所示。第一层是一个 SVM 三分类器，将序列中的氨基酸分为 3 种类型之一($H$、$E$、$C$)。输出结果表示该残基属于该类的概率。因为连续一致的模式是相关联的(例如一个螺旋至少包含 4 个连续一致的结构，一个折叠至少包含 3 个连续一致的结构)，第二层 SVM 三分类器会过滤掉第一层中连续的输出。第二层的目标输出是与第一层输出相同的。与第一层 SVM 一样，第二层也使用 SVM 三分类器，使得每个残基分到具有最大输出值的那一类。

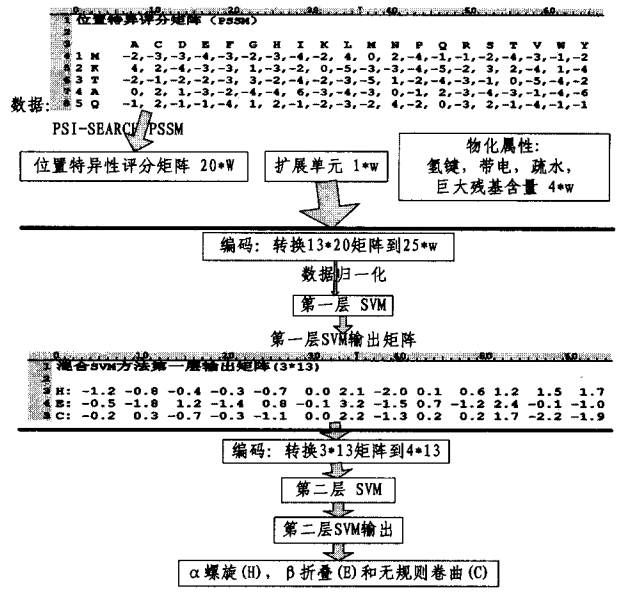
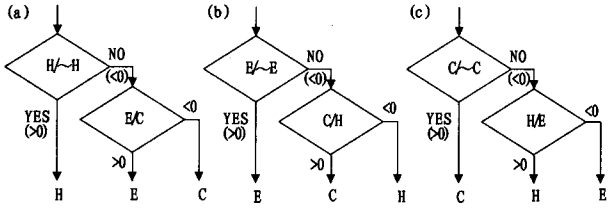



图1 混合SVM模型结构($w$=13)

MSVM系统包括3个部分:系统输入向量(氨基酸物化属性和位置特异性打分矩阵(PSSM))、第一层SVM、第二层SVM。系统输入用滑动窗口的方法转化为25 * 13矩阵。这些向量输入第一层SVM输出的一个3 * 13矩阵,表示残基属于该类的概率。应用滑动窗口,第一层的输出可以转化为4 * 13的矩阵,作为第二层SVM的输入。最终的输出决定于第二层SVM的输出。

### 3.2 三分类器的设计

在构造三分类器之前,先构造几个SVM二分类器,包括3个一类对余类二分类器(一是正类,余类为负类)$H/\sim H$,$E/\sim E$和$C/\sim C$,以及3个一对一二分类器$H/E$,$E/C$和$C/H$。例如,分类器$H/E$对$\alpha$螺旋和$\beta$折叠的样本进行训练,就可以对测试样本$\alpha$螺旋和$\beta$折叠进行分类。




(a) SVM_TREE1,(b) SVM_TREE2,(c) SVM_TREE3,每个都是由两个二分类器串联而成。以SVM_TREE1为例,如果第一个二分类器的输出$H/\sim H$大于0,该样本则判为$\alpha$螺旋($H$),否则将用到第二个分类器$E/C$。如果$E/C$的输出大于0,则样本被判为$\beta$折叠($E$),否则将被判为卷曲($C$)。

图2 三分类器的结构

机器学习方法的目标,就二级结构预测来说是构造一个具有良好预测性能的三分类器。因此,关键是用以上训练好的二分类器来设计三分类器。基于多个二分类器有多种方法来设计三分类器。我们在复合金字塔模型中实现并改进了Hua and Sun[17]提出的方法。该方法基于3个一类对余类的二分类器($H/\sim H$,$E/\sim E$,$C/\sim C$)和3个一对一的二分类器($E/C$,$C/H$,$H/E$)。3个级联的三分类器,SVM-TREE1($H/\sim H$,$E/C$),SVM-TREE2($E/\sim E$,$C/H$)和SVM-TREE3($C/\sim C$,$H/E$),分别由两个二分类器组成,如图2所示。在三分类器中,测试样本的类别定义为到SVM-TREE1,SVM-TREE2和SVM-TREE3分类器最优超平面的最大正类距离SVM-MAX-D。MSVM结合6个二分类器进行投票,使用的投票原则是:如果6个二分类器中多数判为状态$i$,则预测样本判为$i$。MSVM决策函数使用智能决策技术来合并以上所有三分类器的结果。

### 3.3 预测参数的优化

SVM二分类器的构造比神经网络要简单得多,SVM只需要选择一个适当的核函数和正则化参数进行训练。大量的测试表明,径向基核函数(RBF)定义为:

$$K(x_i,x_j)=\exp(-\gamma\| x_i-x_j \|^2) \quad (9)$$

一旦核函数确定,核函数参数$\gamma$和正则化常数$C$将使用文献[16,17]中提到的方法进行优化,$\gamma$=0.05,$C$=1.0。我们用以上参数分别构造SVM分类器。

本文分析了分类器的输入窗口大小对预测精度的影响,并进行反复测试。一个合适的窗口大小可以具有良好的预测性能,因为太短的残基片段可能会遗漏一些重要的分类信息,而太长的片段可能会引入过量的噪声。如表1所列,当窗口大小超过13时,二分类器的预测精度变化不大,这说明SVM方法可以有效地处理噪声。而且,当窗口大小为13时,分类器的分类精度总体较高。因此,本文将窗口大小选为13。

表1 二分类器预测精度对窗口大小依赖性的预测

| Classifier | L=9 | L=11 | L=13 | L=15 | L=17 | L* |
|---|---|---|---|---|---|---|
| H/~H | 86.21 | 86.46 | 87.24 | 87.00 | 86.74 | 13 |
| E/~E | 83.64 | 81.22 | 85.63 | 85.44 | 85.10 | 13 |
| C/~C | 81.84 | 80.62 | 87.53 | 84.02 | 85.21 | 13 |
| E/C | 82.36 | 82.27 | 80.03 | 83.14 | 82.32 | 15 |
| C/H | 80.42 | 83.08 | 85.10 | 83.84 | 82.59 | 13 |
| H/E | 89.12 | 91.32 | 91.50 | 90.78 | 87.93 | 13 |

注:L* 的值是对二分类器最优化窗口的大小。以上为7倍交叉验证的结果。

### 3.4 大样本集的学习策略

尽管SVM已被证实是一种很有效的学习方法,研究人员在分类和寻优速度方面也做了很深入的研究,但对于大规模样本集,特别是在支持向量较多的情况下,寻优和分类速度还是不能满足实际的需求。显然,对于小规模的训练集,训练的速度会快得多。问题的关键在于如何把大规模训练集的规模降低而又不影响分类的正确率。为了解决这一问题,本文采用如下的学习策略:首先用一个小规模的样本集训练,得到一个初始的分类器,然后使用这个分类器对大规模训练集进行修剪,修剪后得到一个较小规模的约减集,再用这个约减集进行训练,得到最终的分类器。算法具体的步骤如图3所示。

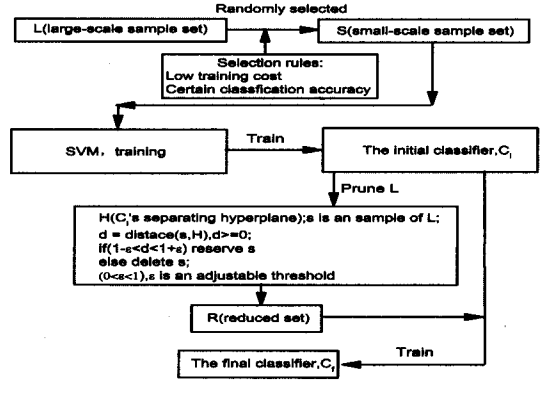


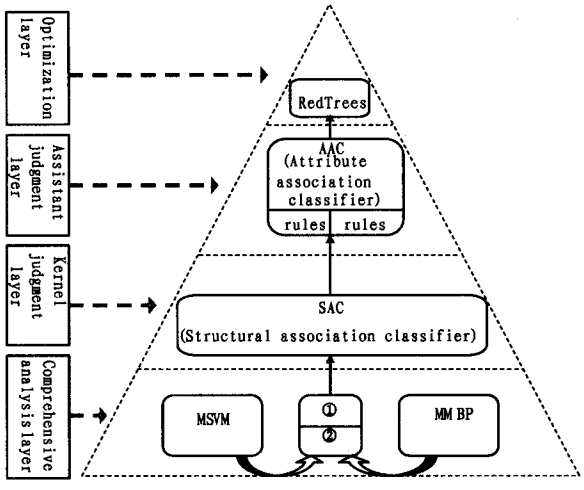

图3 SVM对于大样本集的学习策略

这种修剪策略有效利用了SVM的本质,即分类器仅与支持向量有关,与其它向量(样本)无关。通过采用这种修剪

策略，留下的样本将对分类是有贡献的，而删除的样本对分类贡献较少，甚至起反作用(如导致过学习反而降低了分类器的精度)。实验结果也证实，采用这种学习策略不仅很大程度上降低了学习的代价，而且以这种方式获得的分类器的分类精度完全可以与直接通过大规模样本集训练得到的分类器的分类精度相媲美，甚至更优(这说明这种策略还具有一定的抑制过学习的作用)，同时使得分类速度得到大幅度提高。

## 4 复合金字塔模型

对于蛋白质二级结构预测这类非平凡的复杂问题，一般单一的预测模型或通过将多个单一模型简单组合构成的预测模型，是难以获得令人满意的预测结果的。复合金字塔模型[18]采取了逐步求精、多层递阶的架构，各个层次各有侧重、功能独立且通过智能接口紧密衔接，因此合成金字塔模型具有比传统预测方法更高的预测精度。基于 KDTICM[19] 理论的复合金字塔模型的架构如图 4 所示。




综合分析层包括 MSVM 方法和 MMBP 方法，产生两种结果：一种表示为①，直接送到结果库；另一种表示为②，需要进一步使用 SAC[20]，AAC[20] 和 RedTrees[21] 方法来判定。

图 4 复合金字塔模型

复合金字塔模型由 4 层构成，各层功能独立、紧密协同。这 4 层分别是综合分析层、核心判定层、辅助判定层和优化层，综合分析层使用智能决策合并了同源性分析方法(MMBP)与 MSVM 类化分析结果。MSVM 分类和同源分析都是基于序列结构和氨基酸物化属性的，因此本层合并了物化属性分析与结构序列分析结果。SAC(结构化关联分类)模型处于分类核心层，承担综合分析层中难以判定的数据分类。AAC(属性关联分类)模型位于辅助判定层，通过对氨基酸物化属性的关联分析，建立精化规则库，然后利用改进的 CBA 算法，预测下两层无法判断的数据。最后一层为优化层，使用 RedTrees 方法处理上面 3 层难以预测的结构。

## 5 实验及性能分析

### 5.1 测试集

1)CB513[22]：是由 Cuff 和 Barton 创建的含有 513 条非同源蛋白质序列的数据集，有 84119 个氨基酸，其中 α 螺旋占 30.87%、β 折叠占 21.28%、$c$ 卷曲占 47.85%。其主要目的是评估和改进蛋白质二级结构的预测方法。CB513 数据集包含了几乎整个 RS126[15] 数据集，除了 9 条同源性评分大于 5SD 的蛋白质。该数据集是蛋白质二级结构预测中最常用的数据集之一。

2)CASP394：蛋白质结构预测技术评估(CASP)是一个国际性的蛋白质结构预测技术评比活动，每两年举行一次。这个评估比赛给预测蛋白质结构的研究人员和预测服务器提供了一个公平的、评估其方法的盲测平台，同时也为研究人员展示了结构预测的当前技术水平。鉴于此，使用 CASP 测试集进行方法评估和比较是非常重要的。为了测试 MSVM 方法，本文从 CASP 的官网 http://predictioncenter.org 抽取了 CASP4－CASP8 的蛋白质比赛数据，构造了 CASP394 蛋白质测试数据集，包含 394 条蛋白质序列、87565 个氨基酸，其中 α 螺旋占 32.52%，β 折叠占 20.72%、$c$ 卷曲占 46.76%。

### 5.2 训练集

为了验证 MSVM 方法的可靠性和稳定性，交叉验证是十分必要的，这种方法可以最小化测试数据集和训练数据集部分重叠的概率。本文中，在训练和测试蛋白质数据集 CB513[22] 上使用 7 次交叉验证方法。在 CB513 数据集中，序列长度低于 30 的蛋白质将会被过滤掉，得到 480 条蛋白质序列，因为序列长度较短的蛋白质不能够产生有效的 PSI-SEARCH 比对图谱。这 480 条蛋白质被分割成 7 个子样本，其中一个单独的子样本被保留作为验证模型的数据，其他 6 个样本用来训练(一个子样本为 72 条序列，其他 6 个子样本为 68 条序列)。交叉验证重复 7 次，每个子样本验证一次，平均 7 次的结果得到一个单一估测。对于每一次的分割，计算每个子样本残基的数目和二级结构类型(H，E 和 C)的比例，最后选择带有最小偏倚的子样本。这个过程可以避免因选择那些严重偏倚的子集而导致预测结果不精确。

### 5.3 实验对比

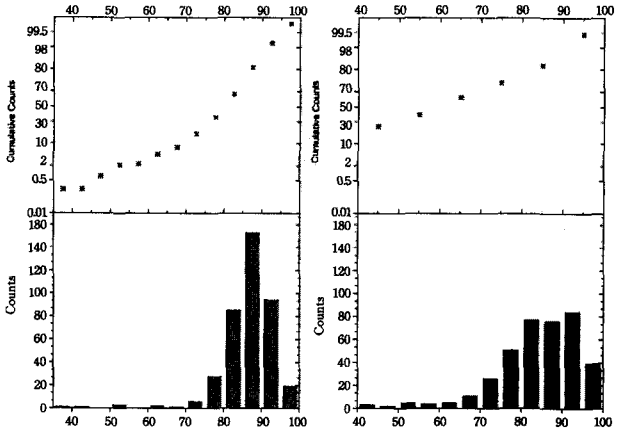
我们选择 CB513 和 CASP394 测试集进行实验，并使用 $Q_3$，SOV99 作为评价指标。在 CB513 数据集上，CPM 的各层预测精度及其处理规模如表 2 所列。在 CASP394 数据集上使用 8 到 3 归类方法获得的结果统计如图 5 所示。所有的结果都是用 $Q_3$ 和 SOV99 进行评分。同时选择了 5 个国际上主流的蛋白质二级结构预测服务器 Psipred，Jpred，PHD expert，SAM-T08 和 Proteus2。预测服务器的详细信息见表 3。在 CASP394 数据集上与本文的 CPM 方法进行的比较如图 6 和图 7 所示。

表 2 CB513 数据集上，预测精度和处理规模

| Module | Accuracy | | Percent |
|---|---|---|---|
| | MSVM | MMBP | |
| Comprehensive Analysis Layer | 118862/146233= 81.28% | 123324/146233= 84.33% | 113638/146233= 77.71% |
| | 105106/113638= 92.49% | | |
| Kernel Judgment Layer | 19961/32194= 62.00% | | 32194/146233= 22.02% |
| Assistant Judgment Layer | 86/ 401= 21.45% | | 401/146233=0.27% |
| Total | 125153/146233=85.58% | | |

表 3 国际上最常用的 5 个蛋白质二级结构预测服务器

| Server | Location | Method |
|---|---|---|
| Psipred[10,25] | University College London, UK | Neural network |
| Jpred[23,26] | The University of Dundee | Consensus-deriving |
| PHD[14,27] | Columbia University, USA | Neural network |
| SAM[7,28] | University of California, USA | HMM |
| Proteus2[24,29] | University of Albert | Consensus-deriving |



(a) $Q_3$ Scores(%) on the CASP394 dataset　(b) SOV99 Scores(%) on the CASP394 dataset

注:SOV99 对于偏螺旋、折叠和卷曲的蛋白质非常敏感。例如当蛋白质类型是偏 α 螺旋、β 折叠和 c 卷曲时,尽管 $Q_3$ 分数比较高,SOV99 分数也可能非常低。

图 5　在 CASP394 数据集上使用 8 到 3 归类方法的结果统计

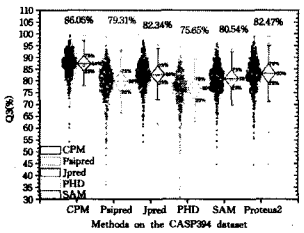



图 6　CPM, Psipred, Jpred, PHD expert, SAM-T08, Protesu2 在 CASP394 数据集上使用 8 到 3 的归类方法取得的 $Q_3$ 精度分布

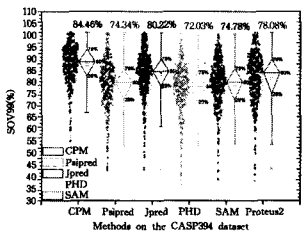



图 7　CPM, Psipred, Jpred, PHD expert, SAM-T08, Protesu2 在 CASP394 数据集上使用 8 到 3 的归类方法取得的 SOV99 精度分布

从以上结果可以得出结论:新算法在盲测数据集 CASP394 上也可以取得很好的分类效果。同时,我们的方法比其他 5 个主流的蛋白质二级结构预测服务器有更好的预测精度以及更强的泛化能力。

**结束语**　本文在基于 KDTICM 理论基础的复合金字塔模型上,提出了一种混合 SVM 方法的蛋白质二级结构预测算法。该算法有效地利用蛋白质进化信息的物化属性和 PSI-SEARCH 生成的位置特异性打分矩阵作为双层 SVM 的输入,使得预测的精度得到了显著的提高。实验结果表明,该算法的预测精度优于目前的其他主流方法。为此,我们开通了蛋白质二级结构预测的在线自动服务网站(http://kdd.ustb.edu.cn/protein_Web/)。如何把蛋白质二级结构预测的结果应用到三级结构的预测,将是我们下一步研究的重点。

## 参考文献


[1] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features [J]. Biopolymers, 1983, 22 (12): 2577-637

[2] 张晓龙,李婷婷,芦进. 基于 Toy 模型蛋白质折叠预测的多种群微粒群优化算法研究[J]. 计算机科学, 2008, 35(10): 230-235

[3] Yang A S, Wang L Y. Local structure prediction with local structure-based sequence profiles[J]. Bioinformatics, 2003, 19: 1267-1274

[4] Qian N, Sejnowski T J. Predicting the secondary structure of globular proteins using neural network models[J]. Journal of Molecular Biology, 1988, 202: 865-884

[5] 闫化军,傅彦,章毅,等. 神经网络方法预测蛋白质二级结构[J]. 计算机科学, 2003, 30(11): 48-52

[6] Hua S J, Sun Z R. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach[J]. J Mol Biol, 2001, 308(2): 397-407

[7] Karplus K, Karchin R, Draper J, et al. Combining local - structure, fold-recognition, and new fold methods for protein structure prediction[J]. Proteins, 2003, 53(6): 491-496

[8] Li J Y, Wong L S, Yang Q. Data mining in Bioinformatics[J]. IEEE Intelligent Systems, 2005, 20(6): 16-18

[9] Smith T F, Waterman M S. Identification of common molecular subsequences[J]. J Mol Biol, 1981, 147(1): 195-197

[10] Jones D T. Protein secondary structure prediction based on position-specific scoring matrices[J]. J Mol Biol, 1999, 292(2): 195-202

[11] Ben-Gal I, Shani A, Gohr A, et al. Identification of transcription factor binding sites with variable-order bayesian networks[J]. Bioinformatics, 2005, 21(11): 2657-2666

[12] Wootton J C, Federhen S. Analysis of compositionally biased regions in sequence databases[J]. Methods Enzymol, 1996, 266: 554- 571

[13] Heniko S, Heniko J G. Amino acid substitution matrices from protein blocks[J]. Proceedings of the National Academy of Sciences, 1992, 89(22): 10915- 10919

[14] Rost B, Sander C, Schneider R. Phd-an automatic mail server for protein secondary structure prediction[J]. Comput Appl Biosci, 1994, 10(1): 53-60

[15] Rost B, Sander C. Prediction of secondary structure at better than 70% accuracy[J]. J Mol Biol, 1993, 232(2): 584-99

[16] Zemla, et al. A modified definition of SOV, a segment-based measure for protein secondary structure prediction assessment [J]. Proteins, 1999, 34(2): 220-223

[17] Hua S J, Sun Z R. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach[J]. J Mol Biol, 2001, 308(2): 397-407

[18] 杨炳儒,谢永红,侯伟,等. 基于复合金字塔模型的蛋白质二级结构预测系统[J]. 科学通报, 2009, 21(54): 3311-3319

[19] 杨炳儒. 基于内在机理的知识发现理论及其应用[M]. 北京:电子工业出版社, 2004

[20] Yang B R, Hou W, Zhou Z. Kaapro: An approach of protein secondary structure prediction based on kdd* in the compound pyramid prediction model[J]. Expert Systems with Applications, 2009, 36(5): 9000-9006

[21] Hou Wei, Yang Bing-ru, Wu Chen-sheng, et al. RedTrees: A relational decision tree algorithm in streams[J]. Expert Systems with Applications, 2010, 37(9): 6265-6269


(下转第 188 页)

集群环境能同时处理的块数，并行性才会表现得比较好。

**结束语** 我们在 Hadoop 这一并行环境运行下实现了一种经典的基于向量空间模型的文本分类算法——TFIDF 分类算法。这种并行化的结果在与集群环境能处理的相对合适的数据集上取得了不错的效果，与理想的线性加速比相当，能够很快地训练出模型和取得分类结果。实验表明，Hadoop 环境在大数据集上的分类远远优于单机算法，这一结果正好弥补了在单机上无法完成的海量数据的挖掘。

下一步我们将从文本分类的角度探索更加实际的应用，情感分析是文本分类的一种特殊应用，并且有一定的挑战性，我们将从这个方面出发，找到更加合理的应用。另外，海量数据集的处理是我们研究的重点，未来我们立志于把海量数据处理做成一种服务，以满足广大用户的需求。

## 参 考 文 献


[1] Sebastiani F. Text Categorization[Z]. Encyclopedia of Database Technologies and Applications. 2005:683-687

[2] Joachims T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization[C]// Proceedings of the Fourteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1997

[3] Yang Y. An Evaluation of Statistical Approaches to Text Categorization[J]. Journal of Information Retrieval, 1999, 1(1/2): 67-88

[4] Rocchio J J Jr. Relevance Feedback in Information Retrieval [M]. Salton G, ed. The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971: 313-323

[5] Tzeras K, Hartmann S. Automatic Indexing Based on Bayesian Inference Networks[C]// Proc. 16th ACM Int. SIGIR Conference. 1993: 22-34

[6] Masand B, Lino G, Waltz D. Classifying News Stories Using Memory Based Reasoning[C]// 15th ACM SIGIR Conference. 1992: 59-65

[7] Apte C, Damerau F, Weiss S. Automated Learning of Decision Rules for Text Categorization[J]. ACM Trans. on Information Systems, 1994, 12(3): 233-251

[8] Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features[C]// Proc. 10th European Conference on Machine Learning (ECML). 1998: 137-142

[9] Salton G, Buckley C. Term Weighting Approaches in Automatic Text Retrieval[J]. Information Processing and Management, 1988, 24(5): 513-523

[10] Kruengkrai C, Jaruskulchai C. A Parallel Learning Algorithm for Text Classification[C]// Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. 2002: 201-206

[11] Lertnattee V, Theeramunkong T. Parallel Text Categorization for Multi-dimensional Data[C]// K. M. Liew, et al., eds. PDCAT 2004, LNCS 3320. 2004: 38-41

[12] Hadoop T W. The Definitive Guide[M]. YAHOO! Press, 2009

[13] Gil-García R, Badía-Contelles J M, Pons-Porrata A. Parallel Nearest Neighbour algorithms for Text Categorization[C]// EURO-PAR 2007 Parallel Processing, Lecture Notes in Computer Science. 2007: 328-337

[14] Khan D. CAKE-Classifying, Associating & Knowledge Discovery An Approach for Distributed Data Mining (DDM) Using PArallel Data Mining Agents (PADMAs)[Z]. Web Intelligence and Intelligent Agent Technology, 2008

[15] Paul S, Saravanan D V. Hash Partitioned Apriori in Parallel and Distributed Data Mining Environment with Dynamic Data Allocation Approach[C]// ICCSIT '08 Proceedings of the 2008 International Conference on Computer Science and Information Technology. 2008

[16] Qiu Xiao-hong, Fox G, Yuan Hua-peng, et al. Parallel Data Mining on Multicore Clusters[C]// Seventh International Conference on GCC'08. 2008: 41-49

[17] Wang Jin-lin, Chen Xi, Zhou Ke-fa, et al. Parallel Research of Sequential Pattern Data Mining Algorithm[C]// 2008 International Conference on Computer Science and Software Engineering. 2008

[18] Wang Jin-lin, Chen Xi, Zhou Ke-fa. Research on a Scalable Parallel Data Mining Algorithm[C]// 2009 Fifth International Joint Conference on INC, IMS and IDC. 2009: 888-893

[19] 王鄂，李铭. 云计算下的海量数据挖掘研究[J]. 现代计算机，2009

[20] 高洁，吉根林. 文本分类技术研究[J]. 计算机应用研究，2004

[21] 谭松波，王月粉. 中文文本分类语料库-TanCorpV1.0[OL]. http://www.searchforum.org.cn/tansongbo/corpus.htm

[22] 搜狐新闻数据(SogouCS)[OL]. http://www.sogou.com/labs/dl/cs.html

[23] http://mahout.apache.org/


(上接第 173 页)


[22] Cuff J A, Barton G J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction [J]. Proteins, 1999, 34(4): 508-19

[23] Cuff J, Clamp M, Siddiqui A, et al. JPRED: A consensus secondary structure prediction server[J]. Bioinformatics, 1998, 14: 892-893

[24] Montgomerie S, Cruz J A, Shrivastava S, et al. Proteus2: a web server for comprehensive protein structure prediction and structure-based annotation[J]. Nucleic Acids Res, 2008, 36: 202-209

[25] http://bioinf.cs.ucl.ac.uk/psipred/psiform.html

[26] http://www.compbio.dundee.ac.uk/www-jpred/

[27] http://cubic.bioc.columbia.edu/predictprotein/

[28] http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html

[29] http://wks16338.biology.ualberta.ca/proteus2/