文章编号:1001 - 9081(2005)09 - 2169 - 02

含先验信息的学习机在生物序列分析中的应用

刘 颖[1], 林元烈[2], 覃 征[1,3]
(1. 清华大学 软件学院, 北京 100084; 2. 清华大学 数学科学系, 北京 100084;
3. 清华大学 计算机科学与技术系, 北京 100084)
(yingliu03@ mails. tsinghua. edu. cn)

摘 要:生物序列分析是机器学习和数据挖掘技术一个重要的应用领域。它的特别之处在于,很多有领域背景的先验知识可以在分析过程中得到利用,从而改善分析的效果。在对蛋白质的乙酰化修饰的预测过程中,通过合理地利用先验信息,改进模式提取方法,能够显著地提高支持向量机模型的预测性能。

关键词:先验信息;生物序列分析;机器学习;支持向量机
中图分类号: TP181 **文献标识码**:A

Application of prior-knowledge-bearing learning machine in biological sequence analysis

LIU Ying[1], LIN Yuan-lie[2], QIN Zheng[1,3]
(1. *School of Software, Tsinghua University, Beijing* 100084, *China*;
2. *Department of Mathematical Sciences, Tsinghua University, Beijing* 100084, *China*;
3. *Department of Computer Science and Technology, Tsinghua University, Beijing* 100084, *China*)

Abstract: Biological sequence analysis is an important application domain of data mining technology. Its particularity lies in that a great deal of prior knowledge can be utilized to improve the learning process. In the research of the protein modification of N-acetylation, by properly using prior knowledge and upgrading the pattern extraction method, improvement in performance of the SVM model was achieved.

Key words: prior knowledge; biological sequence analysis; machine learning; support vector machine(SVM)

随着生物信息学的飞速发展,很多机器学习以及数据挖掘领域的思想和方法被引入到这个新兴的领域中。通过对蛋白质分子的氨基酸残基序列进行分析以预测蛋白质分子的结构及功能,是这个领域一个很重要的分支。近年来这方面的工作取得了很大的进展,各种不同的机器学习模型,如贝叶斯网络、人工神经网络、支持向量机等等,都在实际科研中发挥了很大的作用[1]。然而,由于实验数据的限制,往往需要从已有的数据中尽量挖掘出更多的信息来提高预测模型的性能。

特定的问题往往对应了特定的领域知识。如果能在训练机器的同时将更多的领域知识体现在训练样本中,就能使预测模型的性能得到提高。从操作上讲这并没有一定的模式可以遵循。本文以实际项目中的具体问题(蛋白质的乙酰化预测)为例,来展示先验信息挖掘在生物序列分析中的重大作用。

1 问题的数学模型

生物信息学中的问题需要被转化成数学模型之后才能用计算机来进行分析和处理。这里讨论的乙酰化也一样。蛋白质序列由氨基酸残基组成的线性序列,氨基酸残基共有 20 种,每种用一个大写字母表示[2]。例如,某种蛋白质可以用序列的形式表示为 SSVPGESATPQQPGALSESTTQLPG。用数学语言来描述就是:

序列$\{X_n, n \geqslant 1\}$, $X_n \in$ {20 种氨基酸残基}

N-末端乙酰化(N-acetylation)是蛋白质修饰的一种,即在蛋白质分子的某一个氨基酸残基上加上一个乙酰基基团。用上面的数学模型来描述就是序列中的某个 X_i 被作上了标记。又因为这种现象一般发生在 N-末端的前 40 个残基中的某一个,所以问题可以进一步描述为

序列$\{X_n, 1 \leqslant n \leqslant 40\}$ 中,某个 X_i 被标记

一般来说,与被标记的残基相邻的若干个残基排列会呈现出一定的"模式",如果有足够多的样本,我们就能通过训练让机器来学习这种模式,进而用它来预测给出的序列是否含有被标记过的残基。这样生物学问题就转变成了机器学习领域的问题。

2 实验方法

类似这种问题,一般文章所采取的提取模式的方法都是开窗法[3]。即取目标残基周围的连续若干个残基作为一个样本。取出的残基个数叫做窗口的长度。具体的开窗方法不外乎有两种,一种是以目标残基为中心,向两边取相同数目的残基[4];另一种是以目标残基为首,向后取若干个残基[5]。很少有其他的针对特定问题的开窗方式。乙酰化的问题有其特殊的地方。首先,它只会发生在 S,T,A 和 G 这几种特定的残基上[5];其次,它通常发生在序列的 N-端,即 $1 \leqslant i \leqslant 2$ 时,X_i 被标记的概率比较大。

另外,在蛋白质分子合成初始,N-端形成的第一个氨基酸残基是蛋氨酸(20 种氨基酸残基之一,用 M 表示)。但 M 有可能会在接下来的反应中被切除[6,7]。即部分蛋白质分子的 N-端以 M 开头,而其余的不是。文献研究表明,N-端乙酰化

收稿日期:2005 - 03 - 14;**修订日期**:2005 - 05 - 23
基金项目:国家自然科学基金资助项目(10371063);国家科技攻关项目(2004BA711A21);国家 863 计划资助项目(2003AA412020)
作者简介:刘颖(1981 -),男,硕士研究生,主要研究方向:机器学习; 林元烈(1938 -),男,教授,主要研究方向:随机过程、马尔可夫决策论; 覃征(1956 -),男,教授,博士生导师,主要研究方向:软件体系结构、移动计算.

从时间上讲发生在 N-端蛋氨酸切除之后，并且从位置上讲二者也十分接近。因此，我们有理由认为蛋白质分子的 N-端蛋氨酸切除反应对随后发生的乙酰化有一定影响；也就是说，对我们所提取的模式有影响。因此，我们如果将这部分信息在提取的模式中体现出来，就能提高预测模型的性能。通过对训练样本的分析，我们确定了如下的开窗方式，即将目标残基的前一位和后几位残基一起提取出来。如果目标残基在第一位，就用符号“X”来表示它前面的空位。这样，很多样本第一位就是“M”或“X”。如果是“M”，表明该序列 N-端没有发生蛋氨酸切除；如果是“X”，说明该序列发生了 N-端蛋氨酸切除，而且被标记的残基位于序列第一位。这样做实际上是将相关信息编码到了提取出的模式窗口中。

对负样本，也进行类似的操作。把所有的未被标记的 S，T，A 或 G 作为目标残基，以相同的开窗方式进行特征提取。但由于负样本数量庞大，为了使负样本中以 M 或 X 开头的样本提供更多的信息，我们采用了如下的“样本平衡”策略来平衡正样本中添加的信息。具体方法就是在每轮交叉验证中，固定使用以“M”或“X”开头的负样本，再随机选取若干个一般负样本。这样就最大限度地避免了正样本首位的偏向性所带来的副作用。

这种开窗方式带来的新的信息也可以在数学上得到解释：

令事件 S = {N-端发生乙酰化}，M = {N-端发生蛋氨酸切除}则根据全概率公式[8]有

$$P(S) = P(S \mid M) \cdot P(M) + P(S \mid \bar{M}) \cdot P(\bar{M})$$

其中 $P(S \mid M)$ 和 $P(S \mid \bar{M})$ 可以视为从训练中得到的关于蛋氨酸切除对乙酰化的影响信息，即所谓的先验信息。$P(M)$ 和 $P(\bar{M})$ 代表可以从测试集中获得的关于 *N*-端蛋氨酸切除的信息。从上式不难看出，等式右侧的量为模型预测提供了信息。

最后，再通过稀疏编码(sparse coding)[9]就可以把窗口转化成为 0-1 向量输入到训练模型中。具体来说就是先把 20 种残基按字母序排列，每种残基对应一个位置序号。每一种被编码成为一个 21 维的 0-1 向量，这个 21 维向量只有该残基的对应位置为 1，其余位置均为 0。第 21 维用来表示符号“X”。如残基“S”按字母序排第 16，那么它对应的 21-维向量只有第 16 维是 1，其他位全为 0(000000000000000100000)。采用这种编码的目的是避免按数字顺序编码所带来的残基之间产生的偏序关系。

这里采用的预测模型是支持向量机，因为支持向量机在二分类问题中表现出了优良的性能[10]。我们采用的训练集来自相关研究机构在互联网上公布的标准数据集[11]。为了说明问题，我们用两种不同的开窗方式来进行特征提取，都用 SVM 进行 3 重交叉验证(3-fold cross validation)，然后取三次的平均结果作为最终结果。前一种按照经典的方法向后开窗，后一种采用前文所描述的方法，最后比较一下两次实验的结果。为了更好地说明问题，这里只对丝氨酸(S)的数据进行实验，其余三种(T，A，G)方法类似。两次实验的窗口宽度均取为 7。

3 实验结果

为了更好地衡量实验结果，按照生物学中的习惯，我们考察以下三个量：Matthews 相关系数(MCC)[12]、灵敏性 Sensitivity，以及特异性 Specificity。分别定义如下：

$$MCC = \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp + fn)(tp + fp)(tn + fp)(tn + fn)}}$$

$$Sensitivity = \frac{tp}{tp + fn}$$

$$Specificity = \frac{tn}{tn + fp}$$

其中等式右边的参数均来自预测结果，*tp* 表示真阳性，*tn* 表示真阴性，*fp* 表示假阳性，*fn* 表示假阴性。灵敏性反映了模型对乙酰化的敏感程度；特异性反映了模型对非乙酰化的辨别能力。另外，我们还检测了模型对所有负样本的特异性。为了考察模型的泛化性能，我们还用训练集之外的数据作了一次测试。我们从国际权威的蛋白质数据库 UNI-PROT[13]中提取出了 77 个正样本分别让两个模型来识别。实验中我们选择了 SPIDER[14]类库来实现 SVM。其中，SVM 均采用径向基函数(RBF)作为核函数。老方法的优化参数为：$\gamma = 0.13, C = 1.2$；新方法的优化参数为：$\gamma = 0.1, C = 2.1$。实验结果见表 1。

表 1 两种开窗方式的实验结果

开窗方式	后向开	前向加后向开
MCC	0.38	0.94
Sensitivity	71%	100%
Specificity	64%	94%
Specificity on all negative examples	58%	98%
Sensitivity on UNI-PROT data	81%	84%

上面的实验结果表明，改进开窗方式后，实验的各项指标均有大幅度的提高。最后一项只指标提高了 3%，是因为这里统计的是模型从 77 个序列中定性地辨别出发生乙酰化的序列个数，并不能看出对序列中具体位点的判断情况。这个参数只是作为考察模型泛化性的一个参考指标，并不能很精确地反映出两个模型之间的性能差距。

4 结语

实验证明，通过模式提取从而在训练集中加入先验信息可以极大地提高模型预测性能。但是这种操作没有普遍性，必须针对特定的问题。在实际科研中要根据具体的问题背景和样本情况来确定实验方案。

参考文献：

[1] (法)皮埃尔·巴尔迪，(丹麦)索恩·布鲁纳克. 生物信息学：机器学习方法[M]. 张东晖，等译. 北京：中信出版社，2003.

[2] 阎隆飞，张玉麟. 分子生物学[M]. 北京：北京农业大学出版社，1993.

[3] BRUNAK S, ENGELBRECHT J, KNUDSEN S. Prediction of human mRNA donor and acceptor sites from the DNA sequence[J]. Journal of Molecular Biology, 1991, (220): 49 - 65.

[4] GUO J, CHEN H, SUN ZR, *et al.* A Novel Method for Protein Secondary Structure Prediction Using Dual - Layer SVM and Profiles [J]. PROTEINS: Structure, Function, and Bioinformatics, 2004, 54: 738 - 743.

[5] KIEMER L, BENDTSEN JD, BLOM N. NetAcet: prediction of N-terminal acetylation sites[J]. Bioinformatics, 2005, 21(7): 1269 - 1270.

[6] POLEVODA B, SHERMAN F. N^{α}-terminal acetylation of eukaryotic proteins[J]. J. Biol. Chem., 2000, 275: 36479 - 36482.

[7] POLEVODA B, SHERMAN F. N-terminal acetyltransferases and sequence requirements for N-terminal acetylation of eukaryotic proteins [J]. J. Mol. Biol., 2003, 325: 595 - 622.

[8] 林元烈. 随机数学引论[M]. 北京：清华大学出版社，2003.

[9] BLOM N, HANSEN J, BLAAS D, *et al.* Cleavage site analysis in picornaviral polyproteins: discovering cellular targets by neural networks[J]. Protein Sci, 1996, 5: 2203 - 2216.

[10] VAPNIK V. Statistical Learning Theory[M]. Wiley Inter science, 1998.

[11] http://www.cbs.dtu.dk/services/NetAcet/background/dataset.php [EB/OL].

[12] MATTHEWS BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme[J]. Biochim Biophys Acta, 1975, 405: 442 - 451.

[13] APWEILER R, BAIROCH A, WU CH, *et al.* Uniprot: the universal protein knowledgebase[J/DB]. Nucleic Acid Res, 2004, 32 (Database issue): D115 - 119.

[14] http://www.kyb.tuebingen.mpg.de/bs/people/spider/main.html [EB/OL].