DNA 序列的二阶隐马尔科夫模型分类①

郭彦明, 陈黎飞, 郭躬德

(福建师范大学 数学与计算机科学学院, 福州 350007)

摘 要: 隐马尔可夫模型是对 DNA 序列建模的一种简单且有效的模型, 实际应用中通常采用一阶隐马尔可夫模型. 然而, 由于其一阶无后效性的特点, 一阶隐马尔科夫模型无法表示非相邻碱基间的依赖关系, 从而导致序列中一些有用统计特征的丢失. 本文在分析DNA序列特有的生物学构造的基础上, 提出一种用于DNA序列分类的二阶隐马尔可夫模型, 该模型继承了一阶隐马尔可夫模型的优点, 充分表达了蕴涵在 DNA 序列中的生物学统计特征, 使得新模型具有明确的生物学意义. 基于新模型, 提出一种 DNA 序列的贝叶斯分类新方法, 并在实际DNA序列上进行了实验验证. 实验结果表明, 由于二阶隐马尔可夫模型充分反映了DNA序列碱基间的结构信息, 新方法有效地提高了序列的分类精度.

关键词: 隐马尔可夫模型; 二阶隐马尔科夫模型; DNA 序列; 贝叶斯分类算法; 分类;

Second-Order Hidden Markov Model for DNA Sequence Classification

GUO Yan-Ming, CHEN Li-Fei, GUO Gong-De

(School of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350007, China)

Abstract: Hidden Markov Model (HMM) is one of the simple but effective models for DNA sequence modeling, and the first-order HMM has been popularly used in practice. However, due to the non-aftereffect property, a first-order HMM cannot describe the dependencies between adjacent bases. This generally results in the loss of useful statistics in sequences. In this paper, based on the analysis of the specific biological structure for DNA sequences, a second-order HMM for DNA sequence classification is proposed. The new model inherits the advantages of the first-order model, while fully expresses the biological statistics contained in the DNA sequences, which makes the model more meaningful in biology. Based on the new model, a new Bayesian method is proposed for DNA sequence classification, which is experimentally evaluated on the real DNA sequences. The experimental results show that the new method is able to obtain high classification accuracy, as the structure information hidden in bases of DNA sequences can be captured more adequately by the new second-order HMM.

Key words: Hidden Markov Models (HMM); second-order Hidden Markov Model; DNA sequence; Bayesian classification; classification

1 引言

生物信息学作为一门新的学科领域, 它是把基因组DNA序列信息分析作为源头, 在获得蛋白质编码区的信息后进行蛋白质空间结构模拟和预测, 然后依据特定蛋白质的功能进行必要的生物学研究[1], 这些研究有助于我们了解生命本质和进化过程, 这些工作都需要建立在深刻理解 DNA 和蛋白质构造的基础之上. DNA 序列分析是生物信息学最主要的研究内容之一, 它可以分为两个主要部分: 一是 DNA 序列组成(特别是涉及到基因组层次上)分析, 二是 DNA 序列之间的比较分析[1]. 2002 年, Tautz 等首次提出 DNA 序列分类的概念, 并将之作为生物分类系统的主要平台[2], 目前,

① 基金项目:国家自然科学基金(61175123)

收稿时间:2014-12-25;收到修改稿时间:2015-03-18

DNA 序列分类已成为基因研究的一项基础性工作. 如何对这些DNA序列正确分类, 是生物数据挖掘研究的一项重要内容. 当前机器学习领域的许多分类方法都已成熟地应用到 DNA 序列分类研究中[3].

研究者们针对 DNA 序列提出了相应的分类方法, 其大致可以归纳为三类: 一类是基于特征表示的分类, 另一类是基于序列间距离的分类, 第三类是基于统计概率模型的分类; 其中基于统计概率模型的分类以模型简单, 易于理解及具有较低时间复杂度等优点成为一种比较有优势的DNA序列分类方法, 由于其能够有效挖掘序列中潜在的统计特性而备受关注. 该方法是建立在使序列模型化基础之上的, 最简单的生成统计概率模型是朴素贝叶斯分类器模型[4], 因其模型简单, 已广泛应用于DNA序列分类中, 但其属性独立性假设常与实际应用相违背, 难以准确刻画复杂序列. 已提出的马尔科夫模型与隐马尔科夫模型在一定程度上克服了朴素贝叶斯分类器的不足, 其可以较好地刻画序列元素间的顺序依赖关系[5]. 已有的马尔科夫模型, 隐马尔可夫模型(简称 HMM)、变阶马尔可夫模型(简称 VLMM)等统计学模型都已应用于 DNA 序列建模, 其中又以一阶模型最受关注[5]. 然而, 一阶隐马尔科夫模型存在一阶无后效性的缺陷, 忽略了在数值输出中非相邻状态间的依赖关系, 丢失了一些有用统计特征, 从而影响了分类精度.

针对上述缺陷, 为了能够有效捕捉DNA序列中蕴涵的复杂结构信息, 充分反映DNA序列碱基间的相互关联性, 本文在充分考虑DNA序列的生物学构造的基础上提出一种基于二阶隐马尔可夫模型(HMM2)的贝叶斯新分类方法, 新方法在一阶隐马尔可夫模型基础上综合考虑了相邻以及非相邻碱基间的相互影响, 充分表达了DNA序列的碱基间的密切关联, 使构建的新模型具有更丰富的统计特征, 与现有统计模型相比, 能够完整地保留整个DNA序列所蕴含的结构信息, 在多个实际DNA序列数据集上的实验结果表明, 新方法可以对 DNA 序列数据进行有效分类.

本文组织结构如下: 第 2 章介绍背景知识与相关工作; 第 3 章详细论述基于二阶隐马尔科夫模型的DNA 序列分类方法; 第 4 章给出实验环境和实验结果分析; 第 5 章总结全文, 并给出未来的研究方向.

2 背景知识及相关工作

DNA 序列分类挖掘的目的是将具有相似特点的序列划分到相同类中, 这样的类中的DNA序列具有共同的特性(相同的结构或功能等), 从而可以预测未知序列的功能, 进行 DNA 分子中的基因辅助识别等[3]. 数据挖掘是目前最有效的数据分析手段, 用于发现大量数据所隐含的各种规律. 在 DNA 序列分析中, 数据挖掘技术有着非常广阔的前景, 对于提高数据处理能力、产生有价值的生物学知识起着重要作用. 自 DNA 序列数据库建立以来, 研究者开始采用不同方式分析 DNA 序列, 随着 DNA 序列数据规模的不断增大, 关于DNA序列分类技术的研究也一直在发展, 本节根据目前的研究现状, 重点介绍和分析现有若干DNA序列数据分类挖掘方法.

基于序列特征表示的分类, 是将DNA序列数据映射到合适的算法形式, 使其适合传统的分类方法(如决策树[6]、神经网络[7]、支持向量机 SVM(support vector machine)[8]等), 因为这些传统分类方法都是面向特征向量的, 难以对具有 DNA 序列特性(由非数值符号构成、序列长度差异大、碱基间关联性强同时又存在局部噪声等)的数据集直接建立分类模型. 目前 DNA 序列特征表示方法主要分为两大类: 基于图形表示法[9]和基于统计表示法[10]. 基于序列特征表示的分类方法对分类结果具有很好的可解释性, 然而, 这类方法仅将 DNA 序列的单个碱基作为属性, 在构造分类器时, 单独考虑每个属性特征, 认为碱基在连续位置间是无依赖关系的、独立的. 但事实上, 只考虑属性值而忽略属性间相互关联性的方法, 将造成这些分类方法在 DNA 序列分类上缺乏准确性. 因此, 一个能够有效分类 DNA 序列数据的分类器应该是综合考虑子序列的统计特性和属性间相对位置信息的.

基于序列距离的分类方法是定义一个距离函数来度量两两序列之间的距离, 当距离函数确定后就可以用传统分类算法如 k 近邻[11]和支持向量机[8]等进行分类. 常用的一类距离度量算法是基于序列对齐的距离度量, 在给定相似矩阵的前提下, Needleman 和 Wunsch 于 1970 年提出了双序列全局比对 Needleman-Wunsch 动态规划比对算法[12], 通过动态规划计算出两两序列之间的最佳全局对齐距离; 而相对于全局比对算法, 局部对比算法通过比较序列间对相似区域来度量两两序列的距离, 1975 年 Smith 和 Waterman 在 Needleman-Wunsch 算法的基础上, 提出了两序列局部对比 Smith-Waterman 算法[13], 目前, Smith-Waterman

算法和 BLAST 算法[14]是两种使用最为广泛的序列局部对比算法. 基于序列距离的分类算法的优点是相对简单, 但存在一个较大的缺陷就是两两序列间的距离计算量会随数据规模的增大呈几何式增长[15].

序列分类的另一类方法是基于生成统计概率模型, 假设每个类别中的序列是由一个潜在的统计概率模型 *Model* 生成, 给定一个确定类别的 DNA 序列, 模型 *Model* 刻画了此类别序列中各属性与子序列的概率分布. 在模型训练阶段, 通过 DNA 序列数据学习模型 Model 的参数; 在分类阶段, 计算未知 DNA 序列与各个模型 *Model* 的相似度, 最后将未知 DNA 序列判别为相似度最高的类别. 此过程可以形式化描述为: 对于每个类别 c_j, 训练一个相应的统计模型 $Model_{cj}$, 然后对新的未知序列 $S=(o_1...o_t...o_T)$, 计算 S 与每个模型 $Model_{cj}$ 的相似度, 根据最大相似原则判定 S 的类别, 如下公式所示

$$c \leftarrow \arg\max_{c_j \in C} \prod_{i=1}^{M} P(S_i \mid S_1,...,S_{i-1},\ Model_{c_j})$$

在基于模型的DNA序列分类方法中, 隐马尔可夫模型(HMM)是由马尔可夫链发展扩充而来的一种随机模型, 因其能够对 DNA 序列数据进行有效建模, 成为 DNA 序列分类分析中应用广泛的模型之一.

虽然现有 DNA 序列分类方法在分类精度方面取得了一些成果, 但都是基于一阶隐马尔科夫模型, 而因为一阶无后效性的特点所丢失的统计特征对于 DNA 序列分类是非常重要的. 为了准确的描述 DNA 序列数据, 构建符合 DNA 序列特性的分类模型, 本文提出将二阶隐马尔科夫模型用于DNA序列分类, 基于新模型, 提出一种 DNA 序列的贝叶斯分类新方法. 新分类方法克服了一阶隐马尔科夫模型在 DNA 序列模型构建中一阶无后效性的不足, 有效地提高了序列的分类精度.

3　基于二阶隐马尔可夫模型的DNA序列分类

本节详细阐述基于二阶隐马尔科夫模型的 DNA 序列分类. 下面, 首先描述 DNA 序列的生物学特点, 接着给出模型构建过程, 即用于DNA序列的二阶隐马尔科夫模型的构建, 最后, 详细阐述了基于二阶隐马尔可夫模型的 DNA 序列分类方法.

约定使用的记号如下:

定义1. 记一条DNA序列为$S=(o_1...o_t...o_T)$, 其中:

- T称为序列S长度;
- t=1,2,...,T表示S中的位置;
- $o_t \in \mathbf{\Pi}$, 是字符集$\mathbf{\Pi}$={A,C,T,G}的一个元素.

定义2. 给定W个观察序列集合: $O=\{S^{(1)}, S^{(2)},\ldots, S^{(w)}\}$

其中, $S^{(w)}=o_1^{(w)}\ldots o_t^{(w)}\ldots o_{Tw}^{(w)}, 1\le w\le W$

3.1 DNA 序列的生物学构造分析

由生物学和生物化学知识[16], 组成DNA的4种碱基有腺嘌呤(A)、鸟嘌呤(G)、胞嘧啶(C)及胸腺嘧啶(T), 组成 RNA 的四种碱基为腺嘌呤(A)、鸟嘌呤(G)、胞嘧啶(C)和尿嘧啶(U).

现代生物学已充分证明, DNA 是遗传的主要物质基础, 生物机体的遗传信息以密码的形式编码在 DNA 分子上, 表现为特定的碱基排列顺序, 通过 DNA 的复制由亲代传递给子代. 如图 1 所示, 根据遗传信息传递的中心法则, 遗传信息从DNA转录给RNA, 然后在 RNA 的控制下, 根据每三个碱基决定一个氨基酸的三联体密码规则, 翻译合成具有特定氨基酸顺序的特定蛋白质, 使后代表现出与亲代相似的遗传性状. 这个过程是遵循碱基互补配对原则, RNA 相邻 3 个碱基为一个密码子, 64 种密码子中的 61 种决定 20 种氨基酸, 密码子序列决定氨基酸排列合成蛋白质的序列, 其他 3 种为合成的终止信号, 由此可知, DNA 上的 3 个相邻碱基与一个氨基酸对应. 例如, RNA 序列 UAGCAAUCC 包含了三个密码子: UAG, CAA 和 UCC. 这段RNA编码代表了长度为3个氨基酸的一段蛋白质序列. 因此, 我们有理由认为, DNA 序列中相互关联的三个碱基的关系更为密切.

图 1　DNA 序列遗传信息传递的中心法则

隐马尔科夫模型已在 DNA 序列建模方面被广泛应用. 然而, 其应用主要局限在一阶隐马尔科夫过程, 其具有一阶无后效性的特点, 它的两个基本假设在实际应用研究中并不十分合理, 其中关于状态转移的假设认为: 在 t+1 时刻的状态转移只与该时刻的状态有关, 而与之前的时刻没有关系; 输出值的马尔科夫假设认为: 在 t 时刻输出观察值的概率, 只取决于 $t_i \le t$ 的时刻, 这些假设是为了简化模型而提出的, 然而对于 DNA 序列并不是完全合理的, 因为它忽略了非相邻碱

基间的依赖关系. 虽然一阶隐马尔可夫模型及其衍生模型均有很好的序列建模效果, 但现有模型的共同特点是序列的隐状态仅根据相邻碱基间的信息判定, 未考虑非相邻碱基的影响, 然而DNA序列又是碱基间相互依赖关系很强的字符序列, 针对这一特性本文提出将二阶隐马尔可夫模型用于 DNA 序列分类.

3.2 DNA 序列的二阶隐马尔可夫建模

为模型化 3.1 节提出的 DNA 序列生物学特点, 我们使用二阶隐马尔可夫模型为DNA序列建模. 由于实际问题比马尔可夫链模型所描述的更为复杂, 观察到的事件并不与状态一一对应, 而是通过一组概率分布相联系的, 这样的模型我们称之为隐马尔科夫模型, 其是一个双重随机过程[7]: 一个随机过程是具有一定状态数的马尔可夫链, 这是描述状态转移的基本随机过程, 另一个过程描述状态和观察值之间的统计对应关系. 二阶隐马尔科夫模型基于这样的假设: 隐藏的状态序列是一个二阶 Markov 链:时刻 t +1 的状态不仅与时刻 t 的状态有关, 而且与时刻 t-1 的状态有关; 同样假设任一时刻出现的观测矢量的概率不仅依赖于系统当前时刻所处的状态, 而且依赖于系统前一时刻所处的状态.

设模型的观察值序列为$S=(o_1...o_t...o_T)$, 相应的状态序列为$Q=(q_1,...,q_t,...,q_T)$, 其中, $o_t\in\mathbf{\Pi}=\{A,C,T,G\}$, $q_t\in\{\theta_1,\ldots,\theta_i,\ldots,\theta_d\}$, $d(d>1)$为模型的隐状态数目, θ_i表示第i个隐状态. 序列S的隐马尔科夫模型用一个五元组 $\mu=(A_1,B_1,A_2,B_2,\pi)$来表示:

①初始状态概率分布: $\pi=\{\pi_i\}$, $\pi_i=P(q_1=\theta_i),1\le i\le d$;

②状态转移概率分布: $A_1=\{a_{ij}\}$,$A_2=\{a_{ijk}\}$,

其中: $a_{ij}=P(q_{t+1}=\theta_j|q_t=\theta_i)$, $1\le i,j,k\le d$,

$a_{ijk}=P(q_{t+1}=\theta_j|q_t=\theta_i,q_{t-1}=\theta_k)$, $1\le i,j,k\le d$;

③观察值概率分布: $B_1=\{b_i(l)\}$, $B_2=\{b_{ij}(l)\}$;

其中, $b_i(l)=P(o_t=v_l|q_t=\theta_i)$, $1\le i,j\le d,1\le l\le 4$;

$b_{ij}(l)=P(o_t=v_l|q_t=\theta_i,q_{t-1}=\theta_j)$, $1\le i,j\le d,1\le l\le 4$;

v_l表示字符集$\mathbf{\Pi}$中的第l个元素;

给定隐状态数目d和初始状态概率矢量π, 为一类DNA序列集$O=\{S^{(1)},S^{(2)},\ldots,S^{(w)}\}$建立二阶隐马尔科夫模型 $\mu=(A_1,B_1,A_2,B_2,\pi)$的过程如下: 将$O$作为模型的观察值序列, 学习相对于序列集$O=\{S^{(1)},S^{(2)},\ldots,S^{(w)}\}$的最优模型参数 $\mu=(A_1,B_1,A_2,B_2,\pi)$. 由于$Q$表示的是模型内部的隐状态序列, 通过这样的HMM2建模, 学习得到的μ实际上集中描述了序列集O所蕴含的序列结构信息.

3.3 多观测序列的 HMM2的训练算法

根据3.2节提出的 DNA 序列的二阶隐马尔可夫模型, 本节描述由 Baum-Welch 算法[17]来实现多观测序列的 HMM2的训练.

给定 W 个观察序列集合: $O=\{S^{(1)},S^{(2)},\ldots,S^{(w)}\}$

其中, $S^{(w)}=o_1^{(w)}\ldots o_t^{(w)}\ldots o_{T_w}^{(w)}$,$1\le w\le W$

采用经典的Baum-Welch算法[17]来训练模型, 该算法基于EM算法结构, 在给定一组观察值序列$O=\{S^{(1)},S^{(2)},\ldots,S^{(w)}\}$的情况下, 该算法能够确定一个最优的二阶隐马尔科夫模型 $\mu=(A_1,B_1,A_2,B_2,\pi)$, 使$P(O|\mu)$最大.

令 $\xi_t^{(w)}(i,j,k)=P(q_{t-1}=\theta_i,q_t=\theta_j,q_{t+1}=\theta_k|O^{(w)},\mu)$

表示给定模型μ和第w个观察序列$O^{(w)}$的情况下, 时刻t-1(t=1,2,…,T-1)时处于状态θ_i, 时刻t时处于状态θ_j, 在时刻t+1处于状态θ_k的概率; $r_t^{(w)}(i,j)$表示给定模型μ和第w个观察序列$O^{(w)}$的情况下, t-1时处于状态θ_i, 时刻t时处于状态θ_j的概率: $r_t^{(w)}(i,j)=P(q_{t-1}=\theta_i,q_t=\theta_j|O^{(w)},\mu)$

算法每次迭代的E步骤调用前向-后向算法(forward-backward)[5]计算给定矩阵下的上述两个概率; 在M步骤使用以下两式更新矩阵中每个元素, 文献[18]等推导并证明了多观察序列HMM2的Baum-Welch重估计计算公式, 算法如下:

$$\pi_i=\frac{\sum_{w=1}^{W}\sum_{j=1}^{d}\sum_{k=1}^{d}\xi_2^{(w)}(i,j,k)}{W}\qquad 1\le i\le d$$

$$a_{ij}=\frac{\sum_{w=1}^{W}\sum_{k=1}^{N}\xi_2^{(w)}(i,j,k)}{\sum_{w=1}^{W}\sum_{j=1}^{N}\sum_{k=1}^{N}\xi_2^{(w)}(i,j,k)}\qquad 1\le i,j\le d$$

$$a_{ijk}=\frac{\sum_{w=1}^{W}\sum_{t=2}^{T_w-1}\xi_2^{(w)}(i,j,k)}{\sum_{w=1}^{W}\sum_{t=2}^{T_w-1}\sum_{k=1}^{N}\xi_2^{(w)}(i,j,k)}\qquad 1\le i,j,k\le d$$

$$b_i(l)=\frac{\sum_{j=1}^{N}\sum_{k=1}^{N}\xi_2^{(w)}(i,j,k)\bullet I_{o_t^{(w)},v_l}}{\sum_{w=1}^{W}\sum_{j=1}^{N}\sum_{k=1}^{N}\xi_2^{(w)}(i,j,k)}\qquad 1\le i\le d,1\le l\le 4$$

$$b_{ij}(l)=\frac{\sum_{w=1}^{W}\sum_{t=2}^{T_w-1}\sum_{k=1}^{N}\xi_2^{(w)}(i,j,k)\bullet I_{o_t^{(w)},v_l}}{\sum_{w=1}^{W}\sum_{t=2}^{T_w-1}\sum_{k=1}^{N}\xi_2^{(w)}(i,j,k)}\qquad 1\le i,j\le d,1\le l\le 4$$

这里, $I(\cdot)$是一个指示函数, 即$I(true)=1$和$I(false)=0$. 根据改进的Baum-Welch算法的原理, 此时算法收敛于一个局部最优解.

3.4 基于 HMM2 的贝叶斯分类方法

图2给出了基于二阶隐马尔可夫模型的DNA序列分类步骤, 基于 $P(O|\mu)$的模型类框架, 假设每一个类别 C_i都存在一个对应的模型 μ_i, 这个模型 μ_i是从数据

集 C_i 的训练样本中学习得到的. 在这一模型框架下, 分类器的任务就是找到一条最适合待分类序列 O 的模型 μ, 记作 $P(\mu|O)$.

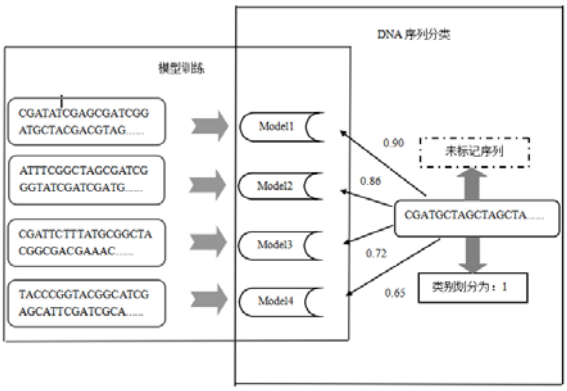

图2　基于二阶隐马尔可夫模型的DNA序列分类步骤

显然, 给定$P(\mu)$和$P(O)$, $P(\mu|O)$可以通过贝叶斯公式计算:

$$P(\mu \mid O)=\frac{P(\mu \mid O)P(\mu)}{P(O)}$$

我们从训练样本数据中学习得到C_i的HMM2模型μ_i, 假设所有的模型都拥有同样的概率$P(\mu)$, 同时, $P(O)$也是一个固定值, 因此, 我们的DNA序列分类问题可以转化成为:

$$class(O)=argmax_i P(\mu_i|O)=argmax_i \frac{P(O|\mu_i)P(\mu)}{P(O)}=argmax_i P(O|\mu_i)$$

对于给定的待分类DNA序列O, 利用前向或后向算法[5]依次计算每个类别C_i的$P(O|\mu_i)$, 最后得到待分类DNA序列的类别标号.

对于经典一阶隐马尔科夫模型的时间复杂度$O(m(T\text{-}1)d^2)$来说, 二阶隐马尔科夫模型的时间复杂度$O(m(T\text{-}1)d^3)$是有所提高, 但当隐状态数目一定时, 也就意味着算法用于构建序列新表示模型的时间与序列数目和序列长度 T 之间均呈线性关系.

4　实验与结果分析

本节通过实验验证二阶隐马尔科夫模型在 DNA 序列分类应用中的有效性. 选择经典一阶隐马尔科夫模型, 朴素贝叶斯分类器模型作为对比对象, 并通过它们在 DNA 序列分类中的应用来检验模型的有效性, 实验采用 Macro-F1(宏平均)指标作为评价标准. Macro-F1 先计算每个类别的准确率、召回率和 F1 值然后取算术平均值, 可以衡量大多数类别的分类效果. 根据文献[19]的模型选择部分的理论与实验结果表明, 隐马尔科夫模型隐状态个数越多, 模型性能趋于越高, 但是隐状态个数超过一定数量时, 模型性能会趋于稳定. 同时我们也发现继续增加隐状态个数只会增加时间消耗, 使模型效率降低. 在实验中我们通过设置隐状态个数分别为 2,4,6,8,10 来测试隐状态对模型分类准确率的影响, 发现当 d≧4 时分类准确率趋于稳定. 综合以上因素本实验设定模型的隐状态数目 d=4.

4.1 实验环境及实验数据

在配置为 *Intel*(R)*Core*(TM)2.27GHzCPU、2GB 内存、500GB 硬盘, 及操作系统为 *MicrosoftWindows*7 的计算机上进行实验, 并使用 *Java* 语言编写的程序实现算法. 实验采用 4 个数据集, 详细参数如表 1 所示. NETEASE 数据集为 2000 年网易杯全国大学生数学建模竞赛题目(*http://www.mem.edu.cn/*)提供的数据文件 *Art—model—data*; GENEBANK 数据集为从原始的 GENEBANK 中抽取出的 182 自然序列; HOVERGEN 数据集[20]为 PBIL(*http://pbil.univ-lyon1.fr/*)的一个同源脊椎动物基因库 HOVERGEN 中抽取出的 6 个类别的 DNA 序列, 去除重复序列, 共有 119 个序列; 第 4 个数据集 BACTERIA 为 NCBI(http://www. ncbi.nlm. nih.gov/)的一个细菌基因库抽取的 4 个类别的 DNA 序列.

表 1　实验数据集参数汇总

数据集	类数目	序列数目	平均长度
SYNTHETIC	2	40	111
GENEBANK	2	182	5534
HOVERGEN	6	119	709
BACTERIA	4	169	2125

4.2 实验结果

对于 SYNTHETIC 数据集, 由于数据集较小, 故将数据集 1-20 号序列作为对 DNA 序列分类的训练集, 然后对数据集 21-40 号序列进行测试, 获得测试集的分类精度, 得到不同模型分类器的分类精度对比结果; 对另外 3 个自然数据集, 实验采用 5-折验证法. 通过随机抽样将每个数据集均分为 5 个子集, 每次选择其中的 4 个子集为训练数据, 剩余的第 5 个子集为测试数据. 每个算法都在这 5 对训练集与测试集上运行一遍并计算分类效果. 实验结果中 3 个模型分类效果以 Macro-F1 指标对比, 具体数据见表 2.

表 2 3 个模型分类效果的 Macro-F1 指标对比

算法	SYNTHETIC	GENEBANK	HOVERGEN	BACTERIA
BAYES	0.781	0.810	0.833	0.825
HMM1	0.850	0.874	0.862	0.881
HMM2	0.936	0.917	0.891	0.908

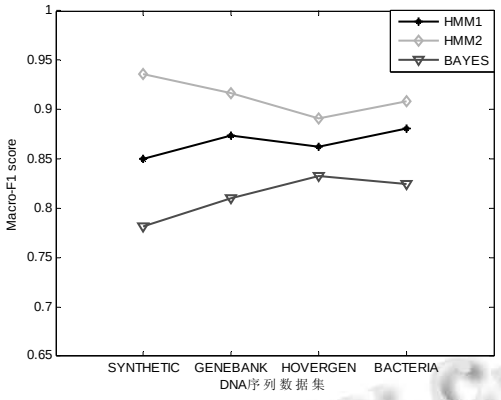

图 3 不同数据集上使用三种不同分类模型分类效果的 Ma cro-F1 指标对比图

如表 2 与图 3 所示, 在 4 个数据集上 HMM1 与 HMM2 对序列的预测效果优皆于 BAYES 算法. 朴素贝叶斯模型因其属性独立假设, 所以不能很好地描述 DNA 序列碱基之间的依赖关系, 因而其模型分类效果较差. 同时, 表 2 与图 3 表明 HMM1 与 HMM2 总的预测能力相当, 但后者算法对 DNA 序列数据集拥有更佳的建模与分类效果. HMM2 算法在 SYNTHETIC 数据集上的分类效果明显优于 HMM1. 从表 1 中的数据集参数信息中可以看出, SYNTHETIC 数据集数据集序列长度相对较短、序列数目较少. 由于 HMM2 算法对 DNA 序列构建的模型中更加充分的考虑了碱基间的相互关联, 因而在序列较短且序列数据不多的情况下(如 SYNTHETIC 数据集), 模型也能有效地表达数据的特征, 即 HMM2 在 DNA 序列数据集较小情况下的也能有效地构建模型以表达序列特征. 一个明显的特点是, 模型的分类精度与数据集规模呈相反的方向发展, 也就是说 HMM2 模型对 DNA 序列元素间关系可以在很短的序列依赖上就刻画出序列的特点, HMM2 模型对 HMM1 模型分类改进效果比较突出.

下面通过实验验证二阶隐马尔科夫模型用于构建 DNA 序列分类模型的时间效率. 为了使实验简单有效, 我们记录单条 DNA 序列二阶隐马尔科夫模型的构建时间. 由于 GENEBANK 数据集包含序列较多且序列长度较长, 因此, 选择 GENEBANK 数据集中序列为实验数据. 我们首先将数据集按序列长度等分为 6 部分, 在每部分随机抽取 20 条序列, 计算 HMM2 方法构建这 20 条 DNA 序列的平均时间消耗. 图 4 给出了二阶隐马尔科夫模型在 GENEBANK 数据集上构建模型的平均时间消耗, 从图 4 可以看出二阶隐马尔科夫模型可以在相对于序列长度的近似线性时间内构建出 DNA 序列模型, 具有良好的算法可伸缩性.

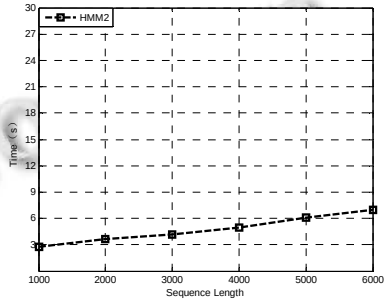

图 4 HMM2 在 GENEBANK 数据集上构建模型的平均时间消耗

5 小结及进一步研究方向

本文针对 DNA 序列的特殊生物学构造及一阶隐马尔可夫模型的一阶无后效性, 提出一种用于 DNA 序列分类的二阶隐马尔可夫模型, 基于新模型, 提出一种 DNA 序列的贝叶斯分类新方法, 新方法克服了一阶隐马尔科夫模型在 DNA 序列分类模型构建中的一阶无后效性, 充分表达了蕴涵在 DNA 序列中的生物学统计特征, 系统描绘了复杂 DNA 序列的生物特性, 在多个 DNA 序列公开数据集上验证了所提方法的可行性和有效性, 实验结果表明, 改进的分类方法对 DNA 序列特性的表达能力增强, 充分反映了 DNA 序列碱基间的结构信息, 能够有效地进行 DNA 序列分类. 下一步工作重点是深入分析各参数对模型构建的影响, 以及研究模型的在线学习能力.

参考文献

1 梁艳春,张琛,杜伟,吴春国,曹忠波.生物信息学中的数据挖掘方法与应用.北京:科学出版社,2011.

2 窦向梅,肖晖,黄大卫.DNA 分类概述.生物学通报,2008,6: 23–26.

3 朱扬勇,熊赟.DNA 序列数据挖掘技术.软件学报,2007,18 (11):2766–2781.

4 Kim SB, Rim HC, Yook D, et al. Effective methods for improving Naive Bayes text classifiers. PRICAI 2002: Trends in Artificial Intelligence. Springer. 2002. 414–423.

5 Lawrence R. A Tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE, 1989, 77(2): 257–286.

6 Quinlan JR. Induction of decision trees. Machine Learning, 1986, 1(1): 81–106.

7 Hassoun MH. Fundamentals of artificial neural networks. MIT Press, 1995.

8 Steinwart I, Christmann A. Support vector machines. Springerverlag New York, 2008.

9 白凤兰.生物序列的图形表示及其应用[博士学位论文].大连:大连理工大学,2005.

10 孙啸,傅静,焦典,等.利用序列统计特征分析基因组序列.北京:中国科学技术大学出版社,2004.

11 Cover TM, Hart PE. Nearest neighbor pattern classification. IEEE Trans. on Information Theory, 1967, 13(1): 21–27.

12 Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular biology, 1970, 48(3): 443–453.

13 Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of Molecular Biology, 1981, 147(1): 195–197.

14 Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. Journal of Molecular Biology, 1990, 215(3): 403–410.

15 蒋红敬.隐马氏模型在生物信息学中的应用及其算法的改进[硕士学位论文].长沙:中南大学,2009.

16 Atiyah M. Mathematics: Frontiers and perspectives. Providence: AMS, 2000.

17 Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden Markov models in computational biology: applications to protein modeling. Journal of Molecular Biology, 1994, 235: 150l–1531.

18 杜世平.多观察序列 HMM2 的 Baum-Welch 算法.生物数学学报,2007,22(4):685–690.

19 Zhong S, Ghosh J. A unified framework for model-based clustering and its application to clustering sequences. Journal of Machine Learning Research, 2003.

20 Wei D, Jiang Q, Wei Y, Wang S. A novel hierarchical clustering algorithm for gene sequences. BMC Bioinformatics, 2012, 13(1): 174–188.