基于隐马尔科夫模型的 DNA 序列分类方法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: DNA 序列分类是生物信息学的一项基础任务, 目的是根据结构或功能的相似性预测 DNA 序列所属的类别. 为进行有效分类, 如何将序列映射到特征向量空间并最大程度地保留序列中蕴含的碱基间顺序关系是一项困难的任务. 为克服现有方法容易导致因 DNA 序列碱基残缺而影响分类精度等问题, 提出一种新的 DNA 序列特征表示方法. 新方法首先为每条序列训练一个隐马尔科夫模型(HMM), 然后将 DNA 序列投影到由 HMM 状态转移概率矩阵的特征向量构成的向量空间中. 基于这种新的特征表示法, 构造了一种 *K*-NN 分类器对 DNA 序列进行分类. 实验结果表明, 新型特征表示方法可以较为完整地保留 DNA 序列中不同碱基间的关系, 充分反映序列的结构信息, 从而有效提高了序列的分类精度.

关键词: DNA 序列; 分类; 特征表示; 隐马尔科夫模型; 特征值分解

DNA Sequence Classification Method Based on Hidden Markov Model

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

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

Abstract: DNA sequence classification is a basic task of bioinformatics, which aims at predicting the category of DNA sequences in terms of their structural or functional similarity. In order to perform an effective classification, how to map the sequences into a feature vector space while retaining the chronological relationships hidden in the sequences as much as possible is currently a difficult task. To address the problems of existing methods, which easily result in affecting the classification accuracy because of incomplete representation of the nucleotides in DNA sequences**,** in this paper, a new feature representation method for DNA sequence is proposed. In the new method, first, each sequence is used to train a Hidden Markov Model (HMM); then, the DNA sequences are projected onto a vector space spanned by the eigenvectors of the HMM state transition probability matrix. Based on the new feature representation, a *K*-Nearest Neighbour classifier is constructed to classify DNA sequences over the vector space. Experimental results show that the new feature representation is able to represent the chronological relationships between different nucleotides in a DNA sequences more integrally. Consequently, the structural information hidden in the sequences can be reflected fully, which in turn improve the classification accuracy of sequences.

Key words: DNA sequence; classification; feature representation; Hidden Markov Models (HMM); eigenvalue decomposition

1 引言

生物信息学(Bioinformatics)是生命科学、计算机科学、信息科学和数学等学科交汇融合所形成的一门交叉学科[1]. DNA 序列数据是生物信息学的主要研究对象之一, 通过分析 DNA 序列, 我们不仅能够理解已有的序列, 而且能够更好地研究新的序列及其功能. 2002年, Tautz 等[2]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, 目前, DNA 序列分

① 基金项目:国家自然科学基金(61175123)

收稿时间:2013-11-24;收到修改稿时间:2013-12-24

类已成为基因研究的一项基础性工作. 当前机器学习领域许多分类方法都已应用到 DNA 序列分类研究[3]. 由于多数成熟的分类模型是面向向量数据类型的, 应用该型分类方法需要从 DNA 序列中提取能够反映碱基间结构特点的特征, 从而将序列映射到由这些特征构成的向量空间中. 然而, 由于 DNA 序列具有由非数值符号构成(因而可以看作是一种事件序列[4])、序列长度差异大、碱基间关联性强同时又存在局部噪声等特点[5], 使得 DNA 序列的有效特征空间表达成为一项困难的任务. 如何从一条DNA序列中提取它的结构信息, 并用适当的数学方法加以描述或表达, 对于DNA序列分类研究是至关重要的, 也是提高DNA序列分类性能的一个关键所在[6].

目前DNA序列特征表示方法主要分为两大类: 基于图形表示法和基于统计表示法. 图形表示法的基本思想是把DNA序列表示成空间的一条曲线, 主要包括G-曲线和H-曲线[7]、Z-曲线[8]等, 具有直观性强的优点, 但不适用于基于数据挖掘的分类应用. 现有的统计特征表示法有显式和隐式两种实现途径. 隐式实现通过核(kernel)隐含地表达序列间的特征关联, 代表性的方法包括string kernel及其衍生方法[9]. 显式实现是在数据预处理过程中提取重要的子结构(包括子序列、子模式等)[10], 并将之作为描述DNA序列的特征, 具有表达直观、可解释性强且具有高度灵活性的优点, 是当前一种被广泛应用和研究的主要方法. 通常, 显式特征提取方法包括两个方面的主题: 特征构成以及衡量特征 “重要性”的统计指标. 业已提出DNA序列的单词频率[10]、二联核苷酸相对丰度特征[11]、碱基对的关联性特征[12]等. 由这些方法提取出来的特征可以反映碱基之间的局部联系, 但忽略了序列整体的结构信息; 此外, 使用简单的频度统计等方法将提取到庞大的特征规模, 容易导致“维数灾难”(the curse of dimensionality[13]), 增加了分类算法的复杂度.

为有效捕捉序列中蕴涵的结构信息, 充分反映DNA序列碱基间的相互关联, 从而降低碱基缺失对序列特征表示的影响, 本文提出一种新的DNA序列特征表示方法用于DNA序列分类, 称之为基于隐马尔科夫模型(Hidden Markov Model, 简称HMM)的特征表示方法. 新方法以HMM状态转移概率矩阵的特征向量来描述DNA序列不同碱基间的整体关系, 与现有方法相比, 能够更为完整地保留整个DNA序列蕴含的结构信息, 且对局部碱基缺失不敏感. 应用于DNA序列分类时, 使用状态转移概率矩阵的特征向量作为序列的特征, 有效压缩了特征的规模. 在多个实际DNA序列数据上的实验结果表明, 基于新特征表示方法的*K*-NN分类器能够取得较好的分类效果.

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

2 背景知识及相关工作

生物学研究表明, DNA序列不是随机的字符串, 它被看作是组成DNA序列的4种核苷酸A(*adenine*)、G(*guanine*)、C(*cytosine*)、T(*thymine*)的线性排列, 其中, 不同排列顺序的DNA区段构成特定的功能单位——基因[14,15]. 不同基因的功能各异, 各自分布在序列的一定区域中. DNA序列分类旨在预测未知类标号DNA序列所属的类别, 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[16,17]. 如前所述, 基于数据挖掘的DNA序列分类通常需要将序列数据映射到某种新的特征空间中, 显然, 序列的特征表示方法将直接影响到分类的性能. 实际上, 特征表示和选择是机器学习和模式识别领域研究的一个重要问题. 本节重点介绍和分析现有若干基于统计的DNA序列特征表示方法.

基于统计的特征表示法从DNA原始序列的结构出发, 根据频率等统计指标提取子结构用以代表该序列. 一种常用的方法便是单词频率(Word Frequency, 简称WF)[10], 这里的单词指序列中出现的一个连续碱基片段, 单词的长度固定为设定的参数值*l*. 这意味着将有4^l种可能的单词组合, 将每个单词看作一个特征时, 任意一条DNA序列就可以表示为4^l维空间的一个向量, 向量每个元素的值为对应单词在序列中出现的频率. 应用此方法将得到为数众多的特征数目, 且基于简单的单词统计并没有考虑到碱基之间的密切关联.

为克服单词频率法的缺点, 一些研究使用了具有更强生物学意义的特征, 例如密码子特征和氨基酸种类特征[18,19]. 设给定一个序列GACCAAGGCAAC, 使用文献[18]的密码子含量特征表示法时提取到的特征为 (GAC)(CAA)(GGC)(AAC) 或者 (ACC)(AAG)(GCA) 或者(CCA)(AGG)(CAA). 显然, 这种方法可以有效压

缩特征的规模，但稳健性较差. 在上述例子中，若丢失一个碱基A，则相应的结果将变为(GAC)(CAG)(GCA)或(ACC)(AGG)(CAA)或(CCA)(GGC)(AAC)，这与原来的特征差距甚远，在某些情况甚至有可能完全不同. Karlin和Ladungal提出的二联核苷酸相对丰度特征(Dinucleotide Relative Abundance，简称DRA)[11]在此基础上增加考虑了相邻两个碱基之间的联系，但局部碱基缺失依然会导致重要特征被忽略，从而影响到DNA序列分类的效果.

应对局部碱基缺失问题的一种思路是考虑DNA序列中相隔h个核苷酸的两个碱基具有的相关性. Liu[12]等提出的基于碱基对关联性(Base-Base correlation，简称BBC)的统计特征表示法使用互信息衡量这种相关性，性，计算公式如下:

$$T_{ij}(h)=\sum_{l=1}^{h}p_{ij}(l)\cdot\log_2(\frac{p_{ij}(l)}{p_i p_j}) \quad (1)$$

其中，p_i为单个碱基出现的频率，$p_{ij}(t)$为相隔h个核苷酸的第i个位置和第j个位置的核苷酸频率. 这里的互信息表示任意位置的核苷酸X相对于相隔h个碱基的另一个核苷酸的信息量，其数值大小衡量了相隔h个位置两核苷酸的相关程度. 文献[13]的分析结果表明BBC方法总体上的稳定性较高. 但是应用BBC统计特征表示方法时，算法效率将随着h值的增大而急剧下降.

上述基于统计的特征表示方法只考虑DNA序列碱基元素或碱基片段间的顺序关联，而却忽略了序列整体碱基间的联系，既序列中蕴含的整体结构信息，这必然导致转换之后序列信息的丢失，从而影响到DNA序列分类的准确性. 此外，这些方法的实现都是建立在序列对比的基础上的，具有较高的算法时间及空间复杂度. 针对这些问题，本文提出对DNA序列进行隐马尔科夫模型(HMM)建模，利用HMM捕捉序列完整的结构信息；提出基于HMM的特征表示方法(HMM-based Feature Representation，简称HMM-FR)，以HMM状态转移概率矩阵的特征向量为表示序列的新特征，特征数目少，有效降低了特征提取算法的复杂度.

3 基于HMM的DNA序列分类

本节详细阐述基于隐马尔科夫模型的DNA序列特征表示方法，结合经典的K-NN分类算法(K-NN)[20]，提出一种新的DNA序列分类方法. 下面，首先描述DNA序列的隐马尔科夫建模方法，接着给出特征表示过程，即HMM状态转移概率矩阵特征值分解；最后分析了新方法的时间复杂度，指出新方法可以在相对于序列长度的线性时间内构建出序列特征集. 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. n条DNA序列组成的集合称为DNA序列集$\Omega=\{S_1,S_2,\ldots,S_n\}$.

3.1 单DNA序列的隐马尔科夫建模

隐马尔科夫模型是一个双重随机过程[21,22]: 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. 由于模型的状态转换过程是不可观察的，因而称之为“隐”马尔可夫模型. 模型基于三个假设: 当前状态只同上一状态相关; 状态之间的转移概率同状态所处具体时间无关; 观察值只与当前状态有关. 这三个假设有效降低了模型的复杂度.

设模型的观察值序列为$O=(o_1,...,o_t,...o_T)$，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}，$q_t\in\{e_1,\ldots,e_i,...,e_d\}$，$d(d>1)$为模型的隐状态数目，$e_i$表示第$i$个隐状态. 序列$O$的隐马尔科夫模型用一个五元组$\mu=(O,Q,A,B,\pi)$来表示:

1) 状态转移概率矩阵 $A=(a_{ij})_{d\times d}$，$a_{ij}=P(q_{t+1}=e_j|q_t=e_i)$，$1\le i,j\le d$;

2) 观察值概率矩阵 $B=(b_{jk})_{d\times d}$，$b_{jk}=P(o_t=v_k|q_t=e_j)$，$1\le j\le d$,$1\le k\le 4$; v_k表示字符集$\mathbf{\Pi}$中的第k个元素;

3) 初始状态概率矢量 $\pi=<\pi_1,\ldots,\pi_i,...,\pi_d>$，$\pi_i=P(q_1=e_i)$,$1\le i\le d$.

给定隐状态数目d和初始状态概率矢量π，为一个DNA序列S建立隐马尔科夫模型(S,Q,A,B,π)的过程如下: 将S作为模型的观察值序列，通过求解

$$w=\arg\max_{\forall A,B}P(S\mid A,B,\pi) \quad (2)$$

学习相对于S的最优模型参数$w=(A,B)$. 由于Q表示的是模型内部的隐状态序列，通过这样的HMM建模，学习得到的w实际上集中描述了S所蕴含的序列结

构信息.

采用经典的Baum-Welch算法[23]求解式(2). 该算法基于EM算法结构[24], 从一个初始的矩阵对(*A*,*B*)出发, 在一个迭代过程中逐步修正*A*和*B*以优化式(2). 令

$$\xi_t(i,j)= P(q_t=\theta_i , q_{t+1}=\theta_j|S,A,B)$$

表示给定*A*和*B*情况下, 模型在时刻*t*(*t*=1,2,…,*T*-1)时处于状态θ_i, 在时刻*t*+1处于状态θ_j的概率; 和表示模型在时刻*t*处于状态θ_i的概率

$$\gamma_t(i) = \sum_{j=1}^{d} \xi_t(i, j)$$

算法每次迭代的E步骤调用前向-后向算法(forward-backward[23])计算给定*A*和*B*下的上述两个概率; 在M步骤使用以下两式更新*A*和*B*中的每个元素a_{ij}和b_{jk}(1≤*i*,*j*≤*d*; 1≤*k*≤4):

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad b_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j) I(o_t = v_k)}{\sum_{t=1}^{T} \gamma_t(j)}$$

这里, $I(\cdot)$是一个指示函数, 即$I(true)$=1和$I(false)$=0. 当*A*和*B*不再发生变化时迭代过程终止, 根据Baum-Welch算法[23]的原理, 此时算法收敛于式(2)的一个局部最优解.

给定长度为*T*的DNA序列*S*, 上述算法E步骤所调用的前向-后向算法[23]的时间复杂度为$O(d^2T)$; M步骤的时间复杂度为$O(T)$. 综上, 对单个序列*S*进行HMM建模的时间复杂度是$O(d^2CT)$, 其中C表示算法的迭代次数. 通常$d<<T$, 且*C*是一个独立于*T*的常数, 因此, 通过上述HMM建模, 可以在(相对于序列长度)的线性时间复杂度内提取到DNA序列*S*的结构信息. 下面, 我们基于模型最优参数(*A*, *B*)中的统计特征表示序列*S*.

3.2 序列特征表示

对于序列集Ω中的*N*个DNA序列$S_1,\ldots,S_i,\ldots,S_N$, 我们使用相同的隐状态数目和初始状态概率矢量分别对它们进行HMM建模, 所建立的*N*个HMM将具有相同的隐状态, 但*N*个模型的状态转移矩阵*A*是有差别的, 这种差别反映了不同DNA序列间的结构差异. 基于这个观点, 本节使用*A*的统计信息来描述每个序列的结构特征, 提出基于状态转移矩阵特征向量的DNA序列特征表示法.

运用特征值分解方法, *d*×*d*的矩阵*A*可以分解成下面的形式:

$$A=X\Sigma X^{-1}$$

其中, Σ是一个对角阵, *d*个对角线上的元素$\lambda_1,\ldots,\lambda_i,\ldots,\lambda_d$为*A*的特征值; *X*是由特征向量构成的矩阵, 与λ_i相对应的特征向量是*X*的第*i*个行向量, 用$\mathbf{x}(\lambda_i)=<x_{i1},\ldots,x_{ij},\ldots x_{id}>$表示. 为使得不同模型(具有相同序号)的隐状态具有对应关系, 我们将*d*个特征值按数值大小由大到小排列, 得到一个新的特征值序列$\lambda'_1\geq\ldots\geq\lambda'_i\geq\ldots\geq\lambda'_d$; 相应地, *X*的第*i*个行向量变为$\mathbf{x}(\lambda'_i)= <x'_{i1},\ldots,x'_{ij},\ldots x'_{id}>$. 最后, 将状态转移矩阵的特征向量作为DNA序列的特征. 形式地, 序列*S*被表达成*d*×*d*维空间的一个向量*V*[*S*]:

$$V[S]=<x'_{11},x'_{12},\ldots,x'_{1d},x'_{21},x'_{22},\ldots,x'_{2d},\ldots,\ldots, x'_{d1},x'_{d2},\ldots,x'_{dd}> \quad (3)$$

式(3)所示的统计特征表示法实质上是将序列*S*投影到一个全新的向量空间, 新空间的特征反映了序列中隐状态变化的主要方向(从矩阵特征值分解的角度看, 特征向量表达了矩阵的 “变化方向”, 按所对应的特征值数值大小排列之后, 依次体现了矩阵主要变化到次要变化), 达到了用较小的特征规模(通常$d^2<<T$)捕捉DNA序列较完整结构特征的目的. 矩阵*A*特征值分解算法的时间复杂度为$O(d^3)$,对特征值进行简单排序(如冒泡排序法)的时间复杂度为$O(d^2)$, 因此, 单个DNA序列统计特征表示法的时间复杂度可以记为$O(d^3)$.

3.3 分类算法

至此, 我们已经将DNA序列转换为d^2维空间的向量, 因而可以使用欧几里得距离等常用度量来衡量序列之间的结构差异(以下简称为序列之间的距离). 形式地, 两个DNA序列S_1和S_2之间的距离用下列$dist(S_1,S_2)$来计算:

$$dist(S_1,S_2)=\|V[S_1]-V[S_2]\|_2 \quad (4)$$

基于公式(4), 现有基于距离的分类算法都可以应用于DNA序列分类, 方法流程如图1所示.

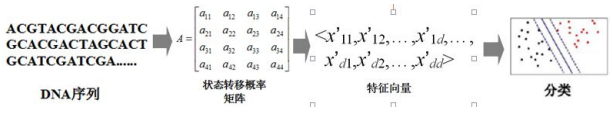

图1 基于隐马尔科夫模型DNA序列分类方法流程图

本文采用*K*-NN分类器[20]对使用HMM-FR特征表示的DNA序列进行分类. *K*-NN是一种 “懒”分类器, 无需训练显式的分类模型, 具有简单、有效且易于实现的优点[20]. 更重要地, *K*-NN分类器可解释性强的特点使它适用于DNA序列分类. 从原理上说, *K*-NN根据

样本间的相似性进行类别预测，而相似性通常基于数据特征之间的“距离”来衡量. 对于DNA序列分类，分类器处理的数据是用HMM-FR表示的特征，这些特征隐含地表达了DNA序列的结构信息，此时使用*K*-NN分类实际上就是依据DNA序列之间的结构相似性进行的，而这正是DNA序列分类的基本出发点[20]. 此外，面对d^2维这样的低维数据时*K*-NN的有效性也已被许多研究所验证[20]. 在本文应用中，给定近邻数目*K*，*K*-NN算法基于公式(4)分别计算测试样本到每个训练样本之间的距离，然后选取*K*个距离最小的样本组成近邻样本集合NN，最后根据NN集中样本的类别标号按多数投票原则确定待测样本的类别. 下面详细描述基于HMM-FR特征表示法和*K*-NN分类原理的DNA序列分类算法.

Input: 由*m*条待分类DNA序列组成的测试样本集合$\Omega_u=\{S_{u1},...,S_{ui},\ldots,S_{um}\}$;
*n*条具类标签的DNA序列组成训练样本集合$\Omega_l=\{S_{l1},...,S_{lj},\ldots,S_{ln}\}$;
隐状态数目*d*和NN数目*K*.

Output: Ω_u种所有序列的的类标签.

Begin
Step 1: 对Ω_u中每条DNA序列进行HMM建模，根据公式(3)提取序列的统计特征，得到代表*m*条序列的向量$V[S_{u1}],\ldots, V[S_{ui}],\ldots, V[S_{um}]$;
Step 2: 对Ω_l中每条DNA序列进行HMM建模，根据公式(3)提取序列的统计特征，得到代表*n*条序列的向量$V[S_{l1}],\ldots, V[S_{lj}],\ldots, V[S_{ln}]$;
Step 3: 对每个待分类序列S_{ui}, i=1,2,…,*m*, 执行以下步骤:
Step 3.1 衡量DNA序列S_{ui}和S_{lj}(j=1,2,…,*n*)之间的结构差异，既用公式(4)计算$V[S_{ui}]$和$V[S_{lj}]$之间的距离;
Step 3.2 选取*K*个距离最小的训练样本构造S_{ui}的*K*NN集合*NN*(*i*)
Step 3.3 输出S_{ui}的类别标签为

$$Label(S_{ui})=\arg\max_{c}\sum_{S_{lj}\in NN(i)} I(c=y_j)$$

其中，y_j表示训练样本S_{lj}的类别标签，*c*表示所有可能的类别标签.
End.

算法步骤1和2的目的是构建DNA序列的HMM-FR表示模型，时间复杂度分别为$O(md^2CT+md^3)$和$O(nd^2CT+nd^3)$. 这里，参数*d*是一个给定的常数且通常数值较小，*C*是独立于*m*和*n*的另一项常数(见3.1节)，因此，这两个步骤的时间复杂度可改写为$O(mT)$和$O(nT)$，这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. 算法的步骤3执行*K*-NN分类，时间复杂度为$O(mn)$.

4 实验与结果分析

本节通过实验验证使用基于隐马尔科夫模型的序列特征表示法HMM-FR及以此为基础的DNA序列分类算法的有效性. 为比较不同特征表示法的性能，选择了WF[10]和简称为64-1、64-2的两种密码子含量特征表示方法[18]作为比较对象. 采用3.3节的算法结构来比较不同特征表示法的性能，也就是首先应用各种表示法对序列进行预处理并表示成向量形式，然后基于*K*-NN算法进行序列分类. 由于*K*-NN的参数NN数目难以确定，本实验测试*K*=1,3,5,7,9五种情况. 不同方法的性能用分类精度来衡量，计算公式如下:

$$Accuracy=\frac{\sum_{i=1}^{m} I(r_{ui}=KNN(S_{ui}))}{m}$$

其中，*K*NN(S_{ui})代表不同算法对待分类DNA序列S_{ui}类别标签的预测结果，r_{ui}是该序列真实的类标签. 实验设置HMM-FR所需的参数*d*=4.

4.1 实验环境及实验数据

在配置为 *Intel (*R*)Core(*TM*)* 2.27 GHz CPU、2 GB 内存、500GB 硬盘，及操作系统为 *Microsoft Windows*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[25]为 PBIL (*http://pbi l.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; 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 3 个类别的 DNA 序列.

表1 实验数据集参数汇总

数据集	类的数目	序列数目	平均序列长度
NETEASE	2	40	111
GENEBANK	2	182	5534
HOVERGEN	6	119	709
BACTERIA	3	134	995

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，得到不同特征表示方法用*K*-NN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. 图3给出了在4个数据集上不同特征表示法使用*K*-NN分类器得到的在不同*K*值中分类精度对比图.

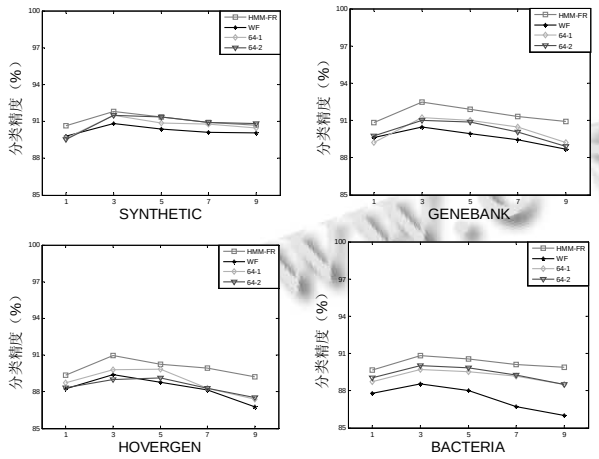

图3 不同数据集上不同特征表示法使用*K*-NN分类器分类精度对比图

表2给出了在4个数据集上的不同特征表示法用*K*-NN分类器得到的平均分类精度对比.

表2 不同数据集上不同特征表示法*K*-NN分类器平均分类精度对比

数据集	HMM-FR	WF	64-1	64-2
SYNTHETIC	91.09	90.24	90.65	90.82
GENEBANK	91.49	88.64	89.22	89.11
HOVERGEN	89.93	88.27	88.95	88.37
BACTERIA	90.21	87.41	89.13	89.33

从图3和表2我们可以看出，在4个数据集上使用本文提出的新特征表示法HMM-FR得到的特征集能够作为已知DNA序列的有效特征，用*K*-NN分类器进行分类，与其他特征表示方法相比较均能取得较高的分类精度. SYNTHETIC 数据集和GENEBANK数据集为二类别分类，HOVERGEN数据集和BACTERIA数据集为多类别分类，由图3可以看出，对于多类别分类HMM-FR的优势更为突出，这是因为新方法完整地保留了DNA序列中不同碱基间的关系，比传统方法包含更多的信息，因此对多类别分类拥有更佳的建模与分类效果. 同时，从图3和表2中可以看出在GENEBANK数据集上HMM-FR得到的DNA序列特征集用于分类效果明显优于其他三种特征表达，从表1中可以看到，GENEBANK数据集的序列长度明显比HOVERGEN数据集相对较长，从实验结果我们可以得到，HMM-FR对于序列长度较长的DNA序列的特征表示优于其他三种统计特征表示法. 另外，由图3中新特征表示方法在不同数据集上的分类精度可以看出HMM-FR对*K*-NN分类器中*K*值的选取相对不敏感.

下面通过实验验证HMM-FR用于构建DNA序列新表示模型的时间效率. 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HMM-FR方法构建这20条DNA序列的平均时间消耗. 图4给出了基于隐马尔科夫模型的特征表示法HMM-FR在GENEBANK数据集上构造特征集所需要的平均时间(秒)，从图4可以看出新方法可以在相对于序列长度的近似线性时间内构建出DNA序列特征集，具有良好的算法可伸缩性.

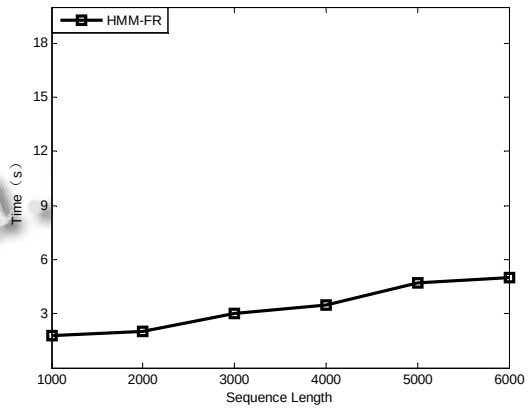

图4 HMM-FR在GENEBANK数据集上构造特征集的平均时间消耗

5 小结及进一步研究方向

本文针对现有DNA序列统计特征表示方法的不足，提出一种新的DNA序列特征表示方法，基于隐马尔科夫模型的特征表示法HMM-FR，以HMM状态转移概率矩阵作为序列的有效特征，该特征充分体现了DNA序列不同碱基间的关联，反映了序列的结构信息. 在序列

特征表示部分, 我们用矩阵特征值分解对其向量化, 用较小的特征规模反映了DNA序列较完整结构特征. 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, 与现有若干代表性的DNA序列统计特征表示方法相比, 新方法对序列特征的表达能力更强, 可以有效提高DNA序列近邻分类的分类精度. 下一步的工作将尝试在HMM-FR基础上应用其它类型的分类器, 同时将该方法推广到蛋白质序列、语音序列以及时间序列等更多应用领域的序列数据挖掘中.

参考文献

1 Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics? A proposed definition and overview of the field. Methods Information in Medicine, 2001, 40(4): 346−358.

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

3 杨旸.基于机器学习方法的生物序列分类研究[博士学位论文].上海:上海交通大学,2009.

4 王清毅,刘洁,蔡庆生.事件序列中的知识发现研究.小型微型计算机系统,1999,20(1):16–19.

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

6 林珠,邢延.数据挖掘中适用于分类的时序数据特征提取方法.计算机系统应用,2012,21(10):224–229.

7 Hamori E, Ruskin J. A novel method of representation of nucleotide series especially suited for long DNA sequences. Journal of Biological Chemistry, 1983, 258: 1318–1327.

8 Zhang R, Zhang CT. Z curves: An intuitive tool for visualizing and analyzing DNA sequences. Journal of Biomolecular Structure & Dynamics, 1994, 11: 767–782.

9 Leslie C, Eskin E, Noble WS. The spectrum kernel: A string kernel for SVM protein classification. Pacific Symposium on Biocomputing, 2002, 7: 566–575.

10 Sinha S, Tompa M. A statistical method for finding transcription factor binding sites. Proc. of the International Conference on Intelligent Systems for Molecular Biology. 2000, 8: 344–354.

11 Karlin S, Burge C. Dinucleotide relative abundance extremes: A genomic signature. Trends Genet, 1995, 11(7): 283–290.

12 Liu ZH, Liu HD, Li JR, Sun X, Jiao D. Base-Base Correlation:A novel sequence feature and its application. The 1st International Conference on Bioinformatics and Biomedical Engineering. 2007. 370–373.

13 Liu Z, Meng J, Sun X. A novel feature-based method for whole genome phylogenetic analysis without alignment: Application to HEV genotyping and subtyping. Biochemical and Biophysical Research Communiication, 2008, 368(2):223–230.

14 Chen X, Kwong S, Li M. A compression algorithm for DNA sequences and its applications in genome comparison. Genome Inform Ser Workshop Genome Information, 1999, 10: 51–61.

15 Churchill GA. Stochastic models for heterogeneous DNA sequences. Bulletin of Mathematical Biology, 1989, 51:79–94.

16 Gelfand MS, Mironov AA, Pevzner PA. Gene recognition via spliced alignment. Proc. of the National Academy Sciences of the United States of America (PNAS), 1996, 93: 9061–9066.

17 Sze SH, Roytberg MA, Gelfand MS, Mironov AA, Astakhova TV, Pevzner PA. Algorithms and software for support of gene identification experiments. Bioinformatics, 1998, 14(1): 14−19.

18 周玉元,周铁军.DNA 序列分类的 Fisher 判别法.湖南农业大学学报(自然科学版),2003,(5):437–440.

19 蔡春,苗立峰,邓乃扬.DNA 序列特征提取方法研究.北京联合大学学报(自然科学版),2008,4:70–72,79.

20 Cover TM, Hart PE. Nearest Neighbor Pattern Classification. IEEE Trans. on Information Theory, 1967, 13(1): 21–27.

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

22 朱明,郭春生.隐马尔可夫模型及其最新应用与发展.计算机系统应用,2010,19(7):255–259.

23 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.

24 Alpaydin E.范明等译.机器学习导论.北京:机械工业出版社,2009.

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