宏基因组分类问题中的特征提取及其降维研究①

陈 波, 徐 云

(中国科学技术大学 计算机科学与技术学院, 合肥 230027)
(中国科学技术大学 安徽省高性能计算重点实验室, 合肥 230027)

摘 要: 宏基因组测序序列分类问题是宏基因组学研究的一个重点问题. 影响宏基因组分类性能的主要因素是特征向量的提取问题, 如何提取并产生合适的特征向量对于提高宏基因组分类问题的分类精度和运行时间有着重大影响. 因此, 针对宏基因组分类问题的数据特点, 利用三阶马尔可夫模型的性质, 提出了一种基于转移概率矩阵的特征提取方法, 并采用基于互信息的特征选择算法对提取后的特征向量进行降维处理, 最后将新提出的特征向量应用到 SVM 分类算法中, 并与相关算法进行了性能对比. 结果显示, 新提出的特征向量在不同的宏基因组物种之间有着良好的区分度, 特别适用于大规模宏基因组数据的分类问题.

关键词: 宏基因组; 测序序列; 转移概率矩阵; 降维; SVM 算法

Features Extraction and Dimensions Reduction in Metagenomic Binning Problem

CHEN Bo, XU Yun

(Department of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China)
(Anhui Province-MOST Co-Key Laboratory of High Performance Computing and its Application, University of Science and Technology of China, Hefei 230027, China)

Abstract: Metagenomic binning is a fundamental question for metagenomic studies. Features extraction is the main factor which influences the performance of metagenomic binning, and how to extract the appropriate feature vectors will influence the binning accuracy and running time. Therefore, this paper proposes a features extraction method which based on third-order Markov model and transferring probability matrix for metagenomic binning problem. Meanwhile, we employ the features selection method based on mutual information to reduce the dimensions of feature vectors and apply it to support vector machine algorithm for binning as well as making comparisons among similar binning algorithms. The results show that this new features extraction method possesses applicable discriminability among different metagenomic species, which is particularly appropriate for large-scale metagenomic binning problem.

Key words: Metagenome; sequencing reads; transferring probability matrix; dimensions reduction; SVM algorithm

由于生物实验技术的限制, 传统的微生物基因组学研究主要集中在可纯培养的微生物上. 这种研究技术存在两个缺点: (1)自然界中超过 99%的微生物是未知的且不可纯培养的; (2)几乎所有的微生物都依赖于或结构依赖于其他的微生物物种或其宿主微生物[1,2]. 随着微生物研究技术的发展, 把同一环境条件下的微生物物种样本作为一个整体进行研究, 已经成为大势所趋. 宏基因组学(Metagenomics) 就是一种以环境样品中的微生物群体基因组为研究对象, 以功能基因筛选或测序分析为研究手段, 以微生物多样性、种群结构、进化关系、功能活性、相互协作关系及与环境之间的关系为研究目的的新的微生物研究方法[3]. 这种方法使得研究自然界中 99%以上的未知的不可纯培养的微生物物种成为可能, 并且已经在许多不同环境的微生物样本研究中取得一定的进展, 如人类肠道[4]、海生蠕虫[5]、土壤[6]、酸性矿山废水池[7]等. 宏基因组测

① 基金项目:国家自然科学基金(61033009)
收稿时间:2015-02-11;收到修改稿时间:2015-04-26

序序列通常包含多个物种, 每个物种都包括大量的测序序列, 并且它们的分类层次所属和物种数量都是未知的. 因此, 宏基因组学研究的首要问题是在生物学意义上的将不同的测序序列划分到其所属的物种, 即宏基因组分类问题[8].

为研究宏基因组学中关于测序序列的分类问题, 本文主要研究和比较了宏基因组测序数据的特征提取方式和降维方法, 并将SVM算法应用到宏基因组分类问题中, 并与已有算法进行了对比分析. 本文将从以下 4 个部分进行阐述, 第 1 节介绍了宏基因组分类问题的定义及相关研究所取得的成果, 第 2 节提出了宏基因组测序数据的特征提取方法、降维方法和分类模型, 第 3 节将新提出的分类算法与已有的算法进行实验上的对比分析, 第 4 节总结全文的研究工作, 并对下一阶段的工作进行了展望.

1 问题定义及研究回顾

在本小节中, 作者首先给出宏基因组分类问题的定义, 并对这个问题的研究文章和方向进行简短的回顾.

问题定义:

利用已知的参考物种信息建立合适的分类器, 对未知类别的宏基因组测序序列进行分类, 使得同一物种的所有序列尽可能多的分类到同一类中. 例如, 对包含 n 个物种的已知参考序列信息, $\{S_{11}S_{12}...S_{1l}, S_{21}S_{22}...S_{2l},..., S_{i1}S_{i2}...S_{il},..., S_{n1}S_{n2}...S_{nl}\}$, 其中任意序列 $S_i\ (i=1,2,...,n)$ 均是由 A、T、C、G 四种碱基组成的序列, 而 $S_{i1}S_{i2}...S_{il}$ 是属于物种 i 的宏基因组序列. 利用这些已知的参考信息, 提取有效的特征向量, 采用有监督的机器学习方法, 建立合适的分类器, 对大量的未知宏基因组序列数据进行分类, 使得 $S_{i1}S_{i2}...S_{il}$ 所代表的物种 i 尽可能多的包含与之相同或相似的宏基因组物种序列.

宏基因组分类问题——针对不同测序仪产生的宏基因组测序序列, 设计准确而快速的分类算法, 是近来宏基因组学研究的一个热点问题, 已有许多学者对此问题做出了相关研究贡献. 例如, TOCOA[9]使用了KNN 分类思想对长度大于 800bp 的 DNA 片段进行分类, 其最大的优点在于可以在本地进行安装测试, 并能保持后台参考基因组数据库的实时更新. AbundanceBin[10]把这些宏基因组测序产生的 DNA 片段看成含参数的混合 Poisson 分布模型, 并使用 EM(Expectation Maximization)算法对这些参数进行近似估计. 但是这种方法只能处理物种丰度差异比较大的两个物种, 其实际意义相当有限. MetaCluster[11,12]则提出了一种两阶段的基于 Spearman 距离的分类方法, 所利用的特征组成是我们所熟悉的 *k*-mer 子串(即长度为 *k* 的核苷酸序列)频率特征, 而且 MetaCluster 不断发展, 现在已经有多个更新版本.

2 宏基因组序列分析及分类算法设计

2.1 宏基因组数据的特征向量提取

宏基因组测序序列是一组由自动化测序仪产生的DNA序列, 目前主要有以Sanger测序法[13]为代表的第一代测序序列和以罗氏 454 测序法[14]为代表的第二代测序序列. 传统的特征提取方法是基于 k-mer 频率特征作为宏基因组数据的特征, 它针对长度为1000bp(base pair)以上的测序序列分类效果较好. k-mer 频率特征向量通常取 4-mer 或 5-mer 频率信息作为特征向量, 因此 k-mer 频率特征向量通常具有高维性(4-mer 特征向量为 256 维, 5-mer 特征向量为 1024 维), 并且由于测序序列本身的缘故, k-mer 频率特征向量还具有稀疏性, 即特征向量中存在大量的零分量. 高维性和稀疏性都不利于宏基因组数据的分类.

相关研究发现, 宏基因组测序片段的四种碱基 A、T、C、G 的排列分布符合高阶马尔可夫模型的性质, 其中细菌类的宏基因组测序序列最满足三阶马尔可夫模型的性质[15]. 三阶马尔可夫模型是指一个系统在时刻 *t* 的状态 *S* 与其在时刻 *t*-1、*t*-2、*t*-3 的状态都相关, 即状态转移之间存在相关联性. 细菌的宏基因组测序片段四种碱基出现次序存在着显著的生物关联性, 即对于一个长度为 4 的短序列片段, 前三个出现的碱基次序通常决定了最后一个碱基的出现概率, 这正好符合三阶马尔可夫模型的性质. 而对于马尔可夫模型, 描述它概率性质最重要的量是转移概率矩阵. 因此, 我们先利用三阶马尔可夫模型来提取测序序列的特征, 然后用转移概率矩阵来刻画提取后的特征向量. 利用三阶马尔可夫模型需要统计前三个碱基转移到下一个碱基的频率(即由 AAA、AAT、AAC、AAG、…、GGG 分别转移到 A、T、C、G 的频率), 然后需要进行归一化处理, 最后可以得出一个转移概率矩阵. 把转移概率矩阵映射为一维行向量的形式, 这整个过程便是提

取了一条测序序列的特征向量, 待输入到分类器模型进行下一步的学习处理.

由于序列 $S_{i1}S_{i2}...S_{il}$ 中每个位点 S 均由 A、T、C、G 中的任意一种碱基组成, 所以基于三阶马尔可夫模型提取的特征向量为 4^3*4 维, 即转移概率矩阵为 64*4 维. 由于三级马尔可夫模型的转移概率矩阵是高维的, 不适合作为特征向量进行输入, 所以我们对其进行映射处理, 把 64*4 维的高维向量映射为 1*256 维的特征向量 $v=(v_1,v_2,...,v_{256})$. 通过图示论证, 我们可以发现这样提取的特征向量在不同物种之间是有较大的区分度的, 即生物分类层次距离近的物种的测序序列的特征向量是相似的, 生物分类层次距离远的物种的特征向量具有较大的差异性, 如图 1 所示.

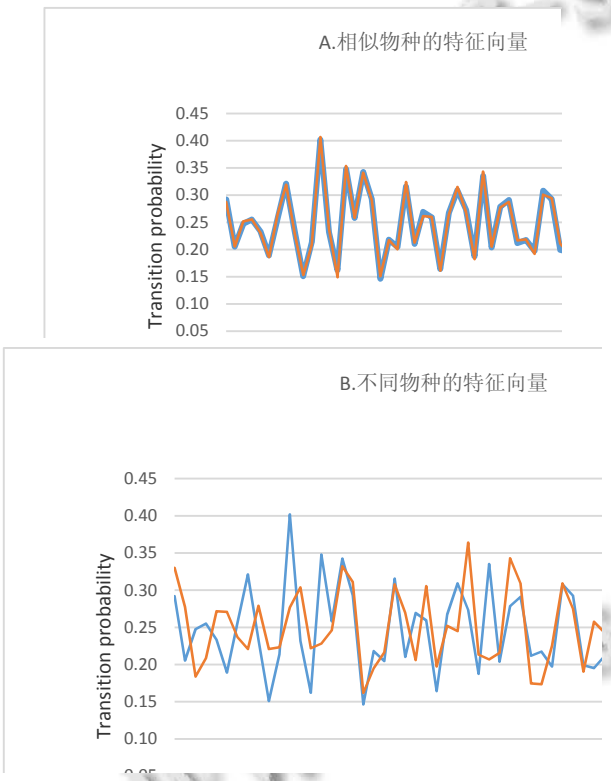


图 1 基于概率转移矩阵的特征向量对比图

2.2 宏基因组特征向量的特征降维

由于初始提取的宏基因组数据的特征向量维度较高, 达到上百维以上, 将此初始提取的特征向量直接用于分类器的训练输入, 是极其耗时的. 此外, 高维的特征向量的各个分量之间是存在相关性的, 即分量存在冗余, 故寻找一种有效的特征向量降维方法, 对于提高分类器的分类精度、降低分类器的训练时间, 是大有裨益的.

特征向量的降维方法包括主成分分析(PCA)、基于信息增益的特征选择、基于互信息的特征选择等. 其中, 主成分分析是采用一种数学意义上的空间降维方法, 找出少量几个综合变量来替换原来的众多变量, 使这些综合变量能尽可能地代表原变量的信息量, 而且彼此之间互不相关. 主成分分析所要做的就是设法将原来众多具有一定相关性的变量, 重新组合为一组新的相互无关的综合变量来代替原来变量, 它主要是一种数学意义上的降维方法. 而特征选择是特征向量降维的重要手段, 将特征选择所获得的特征属性作为数据挖掘的输入属性, 可以有效地加快分类器训练速度、降低分类模型复杂度、提高分类器泛化能力. 常见的特征选择方法包括基于信息增益、基于互信息的特征选择算法, 其中互信息(Mutual Information)一种常见的描述事物之间联系的信息理论中的概念, 它是一类常见的评价函数[16]. 对于两个随机变量 X 和 Y, 它们之间相互依存关系的强弱可以用互信息来表示, 定义如下:

$$I(\mathrm{X};\mathrm{Y})=-\sum_{x,y}p(\mathrm{x},\mathrm{y})\log\frac{p(\mathrm{x},\mathrm{y})}{p(\mathrm{x})p(\mathrm{y})} \quad (1)$$

根据互信息的定义, 可以利用互信息进行宏基因组分类问题的特征选择, 利用互信息 $I(\mathrm{C};\mathrm{F}_i)$ 从全局上衡量特征分量 F_i 和分类类别 C 之间的关系, 那些区分能力较高的特征分量往往具有较高的互信息值. 在处理分类问题提取特征的时候就可以用互信息来衡量某个特征和特定类别的相关性, 如果信息量越大, 那么特征和这个类别的相关性越大. 我们在降低特征向量维数的同时, 尽量减少或去除冗余的信息, 保留甚至增强有意义的信息, 使得特征降维之后的信息损失达到最小.

2.3 宏基因组分类的 SVM 模型及其参数优化

支持向量机(SVM)算法是一种具有较好泛化能力的基于统计学习理论的分类算法, 被认为是有监督分类算法中最好的定式算法[17]. 因此本文采用 SVM 算法作为分类模型. 使用 SVM 算法的过程中, 最重要的一步是关于其参数的调优问题, 相关参数的选择对于 SVM 的分类性能有着重大的影响. 相关研究表明, 惩罚系数 C 和核函数 γ 的参数对于 SVM 的性能有着重大影响[18]. 对于大规模宏基因组数据的分类问题, 快速的选择出最优参数对 (C,γ), 能够有效的减少 SVM 分类器的训练时间, 提高分类器的性能. 目前常

见的参数寻优算法包括网格搜索法、遗传算法、粒子群寻优算法, 其中遗传算法和粒子群寻优算法都是属于启发式算法.

本文采用改进的网格搜索法进行参数调优. 网格搜索是将待搜索的参数范围根据问题所需划分成合适的网格, 通过遍历网格中所有的点对来寻找最优的参数. 在寻找最优参数对 (C,γ) 的过程中, 我们利用交叉验证(Cross Validation)分类准确率来权衡参数对的优劣. 交叉验证是用来验证分类器性能的一种统计方法, 基本思想是将原始数据进行分组, 一部分作为训练集, 另一部分作为验证集. 首先用训练集对分类器进行训练, 再利用验证集来测试训练得到的模型, 以此作为评价分类器的性能指标. 通常采用 *k*-fold 交叉验证, k 的取值一般大于等于 2, 本实验中我们取 $k=5$. 改进的网格搜索过程如下: 首先进行粗略的参数调优, 即参数对呈指数增长模式, 如可以取 $C=2^{-5},2^{-3},...,2^{15},\gamma=2^{-15},2^{-13},...,2^{3}$, 利用 5-fold 交叉验证选择出合适的参数对 (C,γ). 然后在这个所选的区间范围内, 再次利用细致的线性增长式进行搜索, 从而找出全局的最优参数对 (C,γ).

我们将新提出的特征提取及降维方法应用到 SVM 算法中, 并优化 SVM 算法的参数调优过程, 将整个分类模型命名为 MarkovBinning 算法, 其总体流程图如图 2 所示.

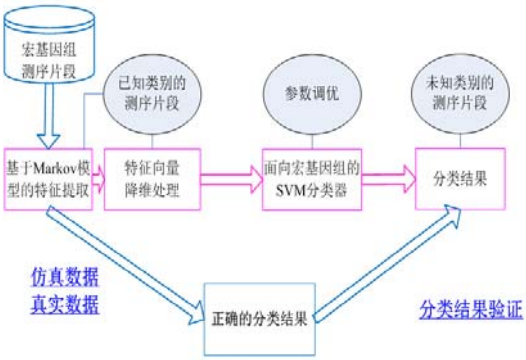


图 2　MarkovBinning 算法总体流程图

3　实验及结果分析

3.1　实验平台与数据集介绍

本实验采用的是英特尔 4 核处理器, 它装配有英特尔 Core(TM)4 核 3.10GHz 处理器, 内存为 4G, 实验平台为 Ubuntu 12.04(64 位)操作系统. 从目前的宏基因组学研究可知, 模拟数据集由于其数据的全面性故被大量的用于宏基因组分类算法的研究中. 在本文的数据集中, 我们考虑了物种的数量、物种的丰度、分类层次上的距离远近以及测序深度等影响因素, 主要使用细菌基因组作为实验样本, 随机地选择了 72 对(共 144 条)来自于 NCBI(National Center of Biotechnology Information)数据库中细菌基因组样本, 利用 MetaSim[19] 软件生成了 54 个模拟测序序列数据集. 根据分类层次的高低不同, 将这 72 对基因组平均的分为 6 个类别: (1)属于同一界(Kingdom)而不同的门(Phylum); (2)属于同一门(Phylum)而不同的纲(Class); (3)属于同一纲(Class)而不同的目(Order); (4)属于同一目(Order)而不同的科(Family); (5)属于同一科(Family)而不同的属(Genus); (6)属于同一属(Genus)而不同的种(Species). 不同的分类层次之间物种的测序序列之间的距离是不同的, 分类层次越低, 距离越小, 因此分类难度越高. 同时, 为了实验数据的全面性, 我们使用 MetaSim 测序序列生成软件不仅生成了 Sanger 测序序列数据集(第一代测序技术代表), 还生成了罗氏 454 测序数据数据集(下一代测序技术代表).

除了模拟数据集, 我们还使用了一个真实的 AMD(Acid Mine Drainage)数据集[2], AMD 数据集的详细信息如图 3 所示. 此数据集包含大约 3000 条 Sanger 测序产生的测序序列, 只有 56%的测序序列可以准确的匹配到 5 条主要的参考基因组上.

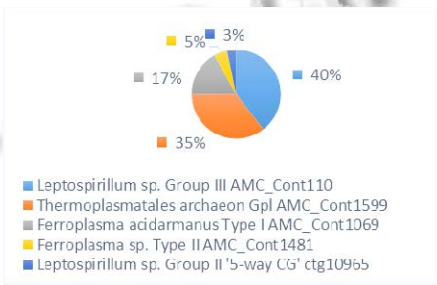


图 3　AMD 真实数据集各物种组成比例

3.2　特征向量提取方法实验结果

本实验中, 我们比较基于三阶马尔可夫模型提取的特征向量和基于 *k*-mer 频率特征向量的宏基因组数据分类性能, 这里的 *k*-mer 频率特征考虑其普遍性我们分别取 $k=4$ 和 $k=5$. 表 1 中为多组宏基因组数据的平均分类结果. 从表 1 中可以看出, 基于三阶马尔可夫模型提取的特征向量应用到 SVM 算法中, 具有良好的分类性能, 不仅分类精度可以超越基于 k-mer 频率

的特征向量, 而且运行时间相对较少, 适合大规模宏基因组数据的分类问题.

表 1 不同特征提取方式的分类性能

特征向量类别	基于转移概率矩阵的特征向量	4-mer 频率特征向量	5-mer 频率特征向量
分类精度	**96.4%**	95.5%	95.8%
运行时间	**27.9s**	26.8s	95.1s

我们对采用不同降维方法而生成的特征向量与不采用降维方法的原始特征向量的分类性能作横向比较, 同时考虑分类精度和运行时间, 定义一个新的综合评价指标 $r=\Delta$运行时间$/\Delta$分类精度. 则根据表 2 有:

$$r_1=1.8/0.3=6.00\,(\text{PCA})$$
$$r_2=23.1/3.8=6.08\,(\text{互信息}) \quad (2)$$
$$r_3=22.8/3.4=6.71\,(\text{条件互信息})$$

可以看出基于条件互信息的特征降维方法的综合性价比最高, 即每损失 1%的分类精度, 可以提高 6.71 个单位运行时间; 基于互信息的特征降维方法每损失 1%的分类精度, 也可以提高 6.08 个单位运行时间; 而基于 PCA 的降维方法效果最差, 这是因为它是一种纯数学意义上的特征降维方法的缘故. 因此可以得出如是结论: 将主成分分析降维方式对于宏基因组数据的分类问题并不适用, 而采用基于互信息或条件互信息的特征选择方法只是略微损失了分类精度, 却得到了 6 倍左右的分类加速效果, 其中基于互信息的特征降维方法更加简单高效, 易于施行. 最终本实验中我们选择基于互信息的特征选择方法, 对基于转移概率矩阵的特征向量进行降维处理.

3.4 模拟数据集实验结果

本实验中的模拟数据集为 Sanger 测序技术产生的宏基因组数据集. Sanger 测序数据集长度都约为 1000bp, 测序错误率为 0.01%, 测序深度为 50X. 由于我们主要考虑的是在不同分类层次下宏基因组分类算法的分类性能, 为了考虑实验的全面性, 所以按照分类类别的数量和物种分类层次的高低将实验分成 2 类宏基因组物种分类(物种丰度 1: 1)、3 类宏基因组物种分类(物种丰度 1: 1: 1)和 4 类宏基因组物种分类(物种丰度 1: 1: 1: 1). 同时我们将 MarkovBinning 算法与 TACOA、AbundanceBin 和 MetaCluster 算法在相同的数据集上作实验性能对比. 图 4 是给出参考物种信息后, 不同的宏基因组分类算法在不同的分类层次和分类数量上的性能对比图.

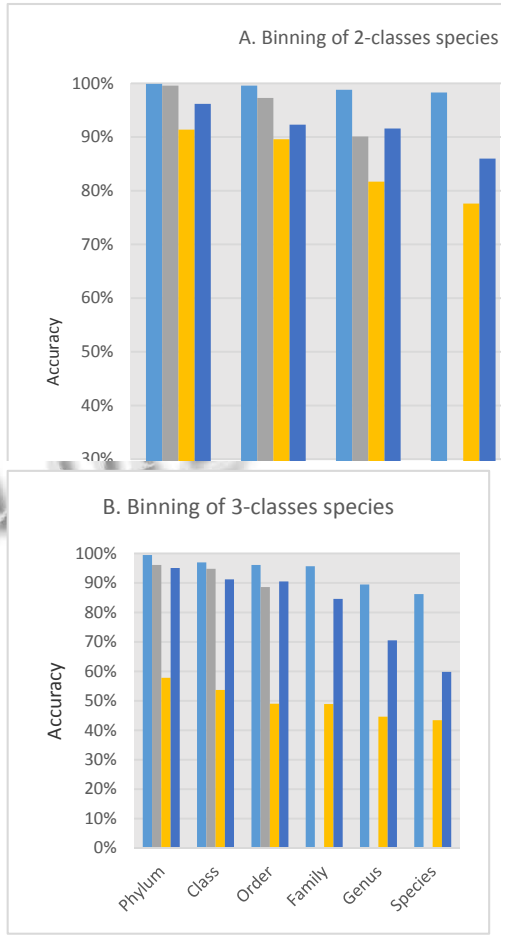


图 4 Sanger 模拟数据集实验结果

从图 4 可以看出, 我们可以看出 MarkovBinning 算法的性能优于 TACOA、AbundanceBin 和 MetaCluster.

随着宏基因组分类物种数量的增加，所有算法的分类准确率有不同程度的下降，但是 MarkovBinning 的下降程度最小. 此外，MarkovBinning 的分类精度在低分类层次上要比 TACOA 高 10%左右，比 AbundanceBin 和 MetaCluster 高 20%~30%左右.

3.5 真实数据集实验结果

我们在真实数据集上同样将 MarkovBinning 算法与 TACOA、AbundanceBin 和 MetaCluster 进行了性能对比. 真实数据集 AMD 上的实验结果如表 3 所示.

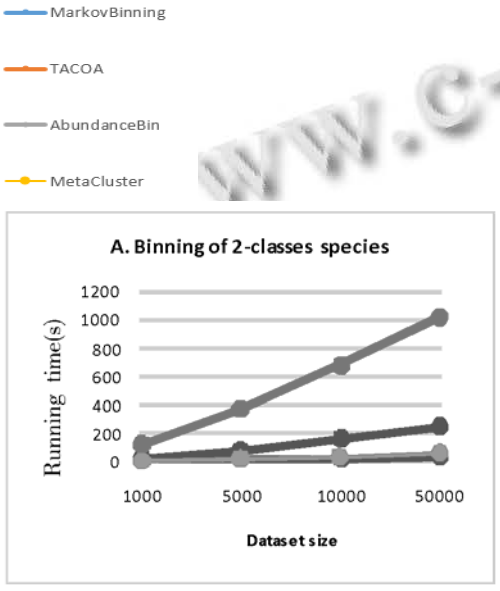
从表 3 中可以看出, MarkovBinning 的分类精度要远优于其他算法.

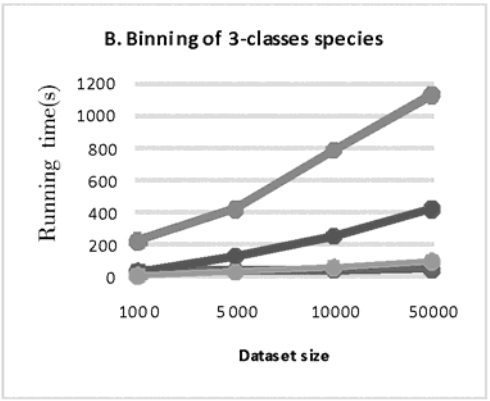
表 3 真实数据集的实验结果

分类精度	Markov Binning	TACOA	Abundance Bin	Meta Cluster
	81.7%	71.4%	46.9%	62.6%

3.6 算法运行时间实验结果

下面我们考虑宏基因组分类算法的时间性能. 随着宏基因组数据规模的日趋增大，可能达到 TB 级别，因此如何降低宏基因组分类的时间复杂度成为决定新算法实际应用性的决定因素. 故我们在本实验中考虑了新算法的运行时间，并与其他 3 个算法进行了对比. 图 5 是不同算法运行时间随着数据规模的增大的变化对比图. 可以看出，随着宏基因组分类数据集的增大，MarkovBinning 的运行时间增加最小，而其他算法均呈不同程度的线性时间增加趋势. 因此可知，MarkovBinning 算法具有良好的应用前景.





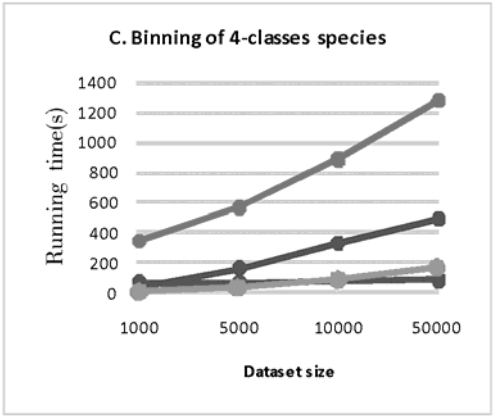


图 5 不同算法在不同规模数据集的运行时间

4 结语

本文提出了一种基于三阶马尔可夫模型的宏基因组数据特征提取方式，并用基于互信息的特征选择方法对提取后的初始特征进行降维处理，将其用于 SVM 分类算法上，且命名为 MarkovBinning 算法. 通过在模拟数据集和真实数据集上比较 MarkovBinning 算法与其他三个已有算法, MarkovBinning 得到了令人满意的分类性能指标，在低层次多物种的宏基因组分类问题上不仅显示出较高的分类精度(平均高出 10%)，而且耗时相对很少(平均减少 3~5 倍). 由此说明，新提出的宏基因组数据特征提取方式具有良好的区分度，可以应用到实际的宏基因组分类问题中去. 此外，由于时间性能上的优越性，该特征提取及降维方式对于日后逐渐增大的宏基因组分类数据集有一定的参考价值. 下一步的研究工作主要是更大规模和更多物种分类的宏基因组分类问题中的特征提取及降维工作.

参考文献

1 Rappe MS, Giovannoni SJ. The uncultured microbial majority.

Annu. Rev. Microbiol, 2003, 57: 369–394.

2 Eisen JA. Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. PLoS Biol, 2007, 5(3): e82.

3 冯美琴.宏基因组学的研究进展.安徽农业科学,2008, 36(2):415–416,479.

4 Gill SR, Pop M, Deboy RT, et al. Metagenomic analysis of the human distal gut microbiome. Science, 2006, 312(5778): 1355–1359.

5 Woyke T, Teeling H, Ivanova NN, et al. Symbiosis insights through metagenomic analysis of a microbial consortium. Nature, 2006, 443(7114): 950–955.

6 Tringe SG, von Mering C, Kobayashi A, et al. Comparative metagenomics of microbial communities. Science, 2005, 308(5721): 554–557.

7 Tyson GW, Chapman J, Hugenholtz P, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature, 2004, 428(6978): 37–43.

8 Mande SS, Mohammed MH, Ghosh TS. Classification of metagenomic sequences: methods and challenges. Brief Bioinform, 2012, 13(6): 669–681.

9 Diaz NN, Krause L, Goesmann A, et al. TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics, 2009, 10: 56.

10 Wu YW, Ye Y. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. J. Comput Biol, 2011, 18(3): 523–534.

11 Wang Y, Leung HC, Yiu SM, et al. MetaCluster 4.0: a novel binning algorithm for NGS reads and huge number of species. J Comput Biol, 2012, 19(2): 241–249.

12 Wang Y, Leung H C, Yiu SM, et al. MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. Bioinformatics, 2012, 28(18): i356–i362.

13 Wu D, Daugherty SC, Van Aken SE, et al. Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. PLoS Biol, 2006, 4(6): e188.

14 Droge J, Mchardy AC. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. Brief Bioinform, 2012, 13(6): 646–655.

15 Zhou F, Olman V, Xu Y. Barcodes for genomes and applications. BMC Bioinformatics, 2008, 9: 546.

16 姚旭,王晓丹,张玉玺,等.特征选择方法综述.控制与决策, 2012,27(2):161–166,192.

17 张学工.关于统计学习理论与支持向量机.自动化学报, 2000,26(1):32–42.

18 Vaapnik Vlaimir N,张学工.统计学习理论的本质.第 2 版. 北京:清华大学出版社,2000.

19 Richter DC, Ott F, Auch AF, et al. MetaSim: a sequencing simulator for genomics and metagenomics. PLoS One, 2008, 3(10): e3373.