基于动态规划的可扩展特征识别

陈金桥，董 珉
（厦门大学信息科学与技术学院自动化系 福建 厦门 361005）

【摘 要】：本文基于动态规划算法提出了可扩展的特征识别方法及其实现程序 TerminusMiner，能够有效识别出序列的高低质量区和 5TSS，3TSS，5TNS，3TNS 等特征。方法对序列特征进行分段识别，并且针对每一分段都能进行扩展性的识别，从而能够很好的适应不定特征的识别，达到较好的识别效果和识别速度。通过对来自 www.conifergdb.org 的 30 多万条松树 EST 序列进行特征识别，验证了该方法。

【关键词】：特征识别 分段识别 EST 序列

0、引言

序列比对是生物信息学领域中的一项基础而重要的任务，它是生物信息学领域其他任务的基础。特征识别作为序列比对的应用，对生物信息学中污染去除、EST 序列注释等等任务的基础。特别是在 EST 的 cDNA insert 的提取中起到重要的作用，因为它能够标识出 cDNA insert 提取的依据，并且能够识别出高低质量区，从而提取出较好的 cDNA inset[1]。

序列比对作为生物信息学中重要的基础，目前已经取得了较好的成果，出现了较多的算法和程序[2][3][4][5][6]。然而作为其重要的应用，特征识别并没有取得好的方法，只是序列比对方法的应用，从而导致了特征识别的准确度不高或不能进行识别。

针对以上情况，本文提出了基于动态规划算法，对特征进行分段识别，使用错配个数替代得分矩阵进行特征识别，使用可扩展识别方法替代全部识别的特征识别方法。该方法能够较好的应用于特征识别，特别是不定长特征的识别。通过对 30 多万条松树序列的特征识别，验证了本文方法。

1、方法

本文提出的特征识别方法首先利用动态规划算法识别的准确性，但是使用分段扩展的方法提高动态规划算法的识别速度，然后使用滑动窗口方法对动态规划算法得到的结果进行扩展，从而得到方法最终得到的特征。

方法包括使用动态规划方法对特征进行分段识别和使用滑动窗口方法进行扩展识别（如图 1）。通过动态规划算法对分段特征进行识别，得到符合错配条件的最短匹配单元。在得到最短匹配单元的基础上，利用滑动窗口方法进行扩展，从而得到最终的特征。

图 1 特征识别框架

1.1 使用动态规划得到最短识别单元

动态规划算法是序列比对的基础算法，它能够得到全局最优解，而不会陷入局部最优解。本文方法基于动态规划方法的全局最优解优点，使其能够得到所有符合条件的最短识别单元。方法首先对特征利用其特征进行分段，然后对各段采用简单的得分矩阵进行匹配，从而得到最短识别单元。

1）特征分段

针对特征的特点进行分段，以下仅以 3TSS 为例进行分析。3TSS 包含 polyA、限制酶以及载体三部分，其中 polyA 是变长的，并且可以两端进行扩展，载体部分也是变长，但是只能针对一边进行扩展，所以本特征分成了两部分识别，polyA 作为一部分，首先以 8 个长度作为最短长短进行识别；限制酶和载体作为一部分识别，首先取 15 个碱基长度进行识别。

图 2

2）各段识别

对于已经分段的特征，针对每段利用动态规划进行识别。由于方法是针对特征的错配个数，所以方法提出了简单的得分矩阵（表 1），从而能够很多的体现错配个数，并得到最后的结果。

	匹配正确	匹配错误	插入	删除
得分	0	-1	-1	-1

表 1 得分矩阵

1.2 扩展最短识别单元

由于特征是不定长度的，所以在识别出最短匹配单元的基础上，需要进行扩展。本文方法采用滑动窗口的方法进行扩展。方法通过每次取固定长度的窗口，并判断窗口中碱基的正确匹配个数是否达到阈值，来决定是否继续扩展。

图 3 展示了 polyA 的扩展过程，其中窗口的长度是 10，窗口正确匹配个数阈值为 8，即窗口中正确匹配的碱基个数大于或等于 8，则继续扩展，否则停止扩展，当前的 polyA 就是最终识别的 polyA。

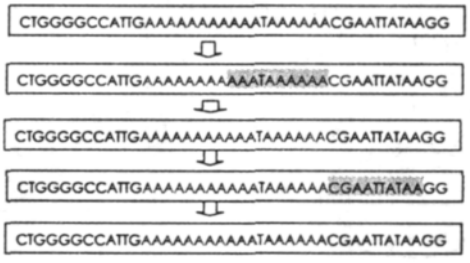

图 3 polyA 的扩展识别过程。其中红色部分表示识别出来的 polyA，黄色背景表示当前扩展部分。

图 4 COLD1_13_C08.b1_A029

2、实验与分析

实验的程序是由 Java 实现的 TerminusMiner，数据来自 www.conifergdb.org 的 30 多万条松树 EST 序列[1][2][7][8]。程序使用大约 10 个小时的时间进行特征识别，共识别出大约 29 万条序列包含特征，具有较高的识别率。

（下转第 84 页）

典大小相同的一个数组。

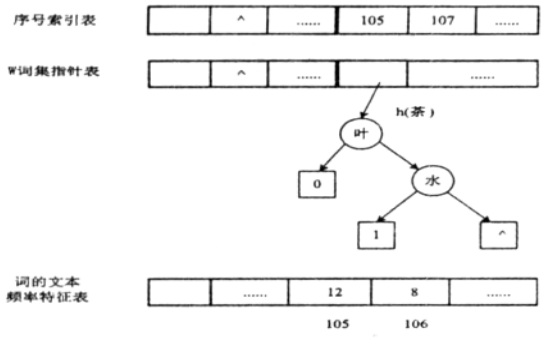

图 2 词典的数据结构示意图

(5) 特征项选取:代表该特征项的词的文本频率大于等于 1。

(6) 特征项权值计算公式:目前使用较多的是 *tf–idf* 方法,其计算公式为:

$$w_{ik}=\frac{tf_{ik}\times\log(N/df_k)}{\sqrt{\sum_{T_k\in D_i}[tf_{ik}\times\log(N/df_k)]^2}}$$

其中,tf_{ik} 表示特征项 T_k 在文本 D_i 中出现的次数,即特征项 T_k 相对于文本 D_i 的文本频率,tf_{ik} 越高, 意味着特征项 T_k 对于文本 D_i 越重要。df_k 表示含有特征项 T_k 的文本数量,df_k 越高,意味着特征项 T_k 在衡量文本相似性方面的作用越低。$N=|D|$,即全部文本的数量,分母归一化因子。$idf_k=log(N/df_k)$ 称为特征项 T_k 逆向文本频率,idf_k 越高,意味着特征项 T_k 对于文本的区别作用越大。

(7) 特征项权值表: 用于记录向量空间中每个特征项的权值,规模与特征向量空间的大小相同。当某一文本被向量化后,即构成一张赋了值的特征向量权值表。

(8) 特征项文本频率特征表:为一规模与向量空间大小相同的数组,数组元素记录了该特征的文本频率。

4.实验分析

本实验所用的语料库来自网站" 中文自然语言处理开放平台",由复旦大学李荣陆博士提供,本文选取了其中的 2690 篇文档,采用台湾大学林智仁(Chih–Jen Lin)博士等开发设计的 LIBSVM 工具,选择分类问题(C–SVC:选择惩罚参数为 C 的 Support Vector Classification),核函数类型设置为 RBF 核。为了获取最优的训练参数, 本文使用 LibSVM 提供的交叉验证工具对训练集进行了 10 重交叉验证, 得到的结果为,C=5,g=0.05。在此情况下,对 SVM 进行模拟实验获得的平均实验数据如表 1 所示:

类别	计算机	体育	农业	政治	文学	经济
查准率	0.832	0.825	0.763	0.698	0.738	0.786
查全率	0.855	0.815	0.687	0.754	0.672	0.815
F1 值	0.842	0.835	0.782	0.712	0.679	0.821

表 1 在 SVM 下的分类效果

根据以上实验结果进行分析:

(1)SVM 有着良好的分类性能、泛化能力,同时,SVM 进行分类决策的时间相对较快。这些优点使得 SVM 技术应用在信息过滤系统中,有着一定的优势。

(2)由于没有一个标准化的向量空间词典,虽然分类结果在一定程度上有所提高,但还不是完全精确。

5.小结

本章重点讨论了普遍认为性能较好的 SVM 分类方法中的词典结构问题,SVM 在特征独立性的基础上进行了模拟实验,实验证明,基于 SVM 的分类算法精确度高,分类速度快,但由于没有一个标准化的向量空间词典在文本分类处理时也有分错的词。

参考文献:

1.李荣陆.文本分类及其相关技术研究[D].上海:复旦大学,2005.
2.林鸿飞,李业丽,姚天顺.中文文本过滤的信息分流机制[J].计算机研究与发展,2000,37(4):470–476.
3.林鸿飞.中文文本过滤的逻辑模型.东北大学博士论文,2000, 6–7
4.林鸿飞,战学钢,姚天顺.基于概念扩充的中文文本过滤模型[J].计算机科学,2000,27(2):88–90.
5.邓乃扬,田英杰.数据挖掘中的新方法----支持向量机.北京:科学出版社,2004.
6.Nello Cristianini, John Shawe–Taylor.支持向量机导论.李国正等译.北京:电子工业出版社.2004,3
7.黄萱菁,吴立德.独立于语种的文本分类方法[D].中文信息学报,2000,14(6):1–7.

(上接第 78 页)

本文方法不仅就有较高的识别率, 并且具有较高的识别准确率。图 4 展示了 COLD1_13_C08.b1_A029 序列的特征识别结果,图中序列前端识别出来的特征是 5TNS,后端是 3TNS,两者都得到了精确的识别。识别出来的端非常好的符合了 EST 序列的期望结构,从而验证了特征识别的准确性。

3、结论与展望

通过对 30 多万条序列的实验表明,本文提出的方法能够较好的适应序列的特征识别,具有较高的识别率和准确率。方法基于动态规划算法,能够很好的得到全局最优特征,而不会陷入局部最优解; 同时采用扩展识别方法来弥补动态规划算法的速度较慢,并且能够很好的适应不定长度的特征识别。

本文采用动态规划方法进行最短单元识别, 能够避免陷入局部最优解, 并且采用扩展方法缓解动态规划算法的速度慢缺陷,但如果能够采用其他的智能识别方法[9][10],则能够进一步的加速方法的识别速度,这也是本文方法的进一步研究任务。

参考文献:

1.Chun Liang, Yuansheng Liu, Lin Liu, Adam C. Davis, Yingjia Shen, Qingshun Quinn Li.. ESTs with cDNA termini: Previously Overlooked Resources for Gene Annotation and Transcriptome Exploration in Chlamydomonas reinhardtii[J]. Genetics, 2008, 179:83–93.
2.Chun Liang, Gang Wang, Lin Liu Guoli Ji, Lin Fang, Yuansheng Liu, Kikia Carter, Jason S Webb and Jeffrey FD Dean. ConiferEST: an integrated bioinformatics system for data reprocessing and minig of conifer expressed sequence tags(ESTs)[J], BMC Genomics 2007, 8:134.83–93
3.BLAST http://blast.ncbi.nlm.nih.gov/Blast.cgi[Z].
[4.Andress Wilm,Desmond G. Higgins and Cedric Notredame R–Coffee: a method for multiple alignment of non–coding [J] RNA Nucleic Acids Research, 2008, Vol. 36, No. 9
5.James Robert White, Michael Roberts, James A. Yorke and Mihai Pop. Figaro: a novel statistical method for vector sequence removal [J]. BMC Bioinformatics 2008 Vol. 24 no. 4.
6.Arthur L. Delcher, Adam Phillippy, Jane Carlton and Steven L. Salzberg Fast algorithms for large–scale genome alignment and comparison[J]. Nucleic Acids Research 2002 Vol. 30 No. 11.
7.Liang C, Sun F, Wang H, Qu J, Freeman RM Jr, Pratt LH, Cordonnier–Pratt MM: MAGIC –SPP: a database –driven DNA sequence processing package with associated management tools[J]. BMC Bioinformatics 2006, 7: 115.
8.Liang C, Wang G, Liu L, Ji G, Liu Y, Chen J, Webb JS, Reese G, Dean JF. WebTraceMiner: a web service for processing and mining EST sequence trace files[J]. Nucleic Acids Res (2007) 35:W137–W142
9.M Dorigo, V Maniezzo and A Colorni. The Ant System: Optimization by a colorry of cooperation agnets [J]. IEEE Transactions on Systems, 1996, 26(1):1–13.
10.Viharos Z J, etal. Traning and Application of Artificial Neural Networks with Incomplete Data[J]. In:LNAI 2358,2002.649–659.