# 基于成对约束降维的 MicroRNA预测

魏 爽, 杨 明

(南京师范大学计算机科学与技术学院, 江苏 南京 210046)
(江苏省信息安全保密技术工程研究中心, 江苏 南京 210046)

[摘要] MicroRNA是一类内源、单链非编码小 RNA, 在生物体内发挥着重要的调控作用. 对 microRNA的预测有助于研究和理解它们的生物学功能. 目前, 针对成对约束的 microRNA预测方法还报道不多. 为此, 本文提出了一个基于成对约束的降维算法, 该算法并入数据局部结构保持策略, 以此有效改进 microRNA的预测性能. 在 microRNA 数据集和 UCI数据集上的实验结果表明, 新提出的基于成对约束的降维方法是有效可行的.

[关键词] miRNA, 成对约束, 降维, 预测
[中图分类号] TP181 [文献标识码] A [文章编号] 1001-4616(2010)04-0166-06

# The Prediction of MiRNA Based on Pairwise Constrains Dimensionality Reduction

Wei Shuang, Yang Ming

(School of Computer Science and Technology, Nanjing Normal University, Nanjing 210046, China)
(Jiangsu Research Center of Information Security & Privacy Technology, Nanjing 210046, China)

**Abstract** MicroRNAs are a class of non-coding RNAs of single-stranded, endogenous, which play an important role in gene regulation. The prediction of MicroRNAs will help a lot to study and understand their biological function. At present, very little work has been done for the prediction of microRNAs by using pairwise constrains. Therefore, after adding local structure preserving strategy, a dimensionality reduction algorithm based on pairwise constrains is introduced to improve the prediction of microRNAs effectively. Experimental results on microRNA and UCI data sets validate the performance of the proposed algorithm.

**Key words** miRNA, pairwise constrains, dimensionality reduction, prediction

MicroRNA (miRNAs)是近年来发现的一类长度为 21~ 23 nt的内源、单链非编码小 RNA, 在后转录过程中指导基因的表达调控[1]. 目前的研究表明, 其形成可分为 3个阶段: 首先, miRNA由 DNA转录成初级转录物 (pri-mRNA); 而后经 Drosha酶剪切形成长度约为 70 nt的带茎环结构的 mRNA前体 (pre-mRNA); 最后在转运蛋白 exportin-5的作用下由细胞核内转到细胞质中, 经 Dicer酶进一步切割产生成熟的 miRNA[2-5]. MRNA在真核有机体内是保守的, 且被认为是基因调控中的重要组成部分[6-10]. 因此, 对 mRNA的预测有着重要的生物学意义.

层次计算方法和判别识别技术已被用于 miRNA的预测, 但预测的效果并不理想. 自 2005年以来, 机器学习方法, 如支持向量机 (SVM)[11]和随机森林[12]被广泛用于识别 miRNA, 并得到了很好的效果. Xue的 3SVM[13], 从 mRNA的序列和二级结构中提取出 32个三联特征, 以其简单易于理解而受到生物学家的重视. MiPred[14]加入了两个热力学特性, 采用随机森林方法进行预测, 进一步改进了预测性能. 而 yasMR[15]加入了 miRNA与编码序列的比对作为新特征, 采用碱基对概率方法对 mRNA预测, 也获得较高的预测性能. 然而, 上述这些方法只从 miRNA的序列和二级结构能提取其固有特征, 并没有对这些特征的可

收稿日期: 2010-06-10
基金项目: 国家自然科学基金 (60873176)、江苏省自然科学基金 (BK2008430).
通讯联系人: 杨 明, 博士, 教授, 博士生导师, 研究方向: 数据挖掘、机器学习、粗集理论与应用. E-mail: m.yang@njnu.edu.cn

取性进行处理, 即没有考虑那些分类贡献不大的特征对分类精度的影响.

目前, 研究者提出了大量的降维方法, 其中包括成对约束[16]. 所谓成对约束, 其包含两个约束: must-link约束和 cannot-link约束. 这两个约束都只考虑一对样本是否属于同一类而不考虑样本的标签信息. 成对约束是比标签信息更一般的信息, 因为从标签信息可以获得成对约束信息, 但反之则不行. 近年来, 成对约束信息已被作为先验知识来对数据进行降维. Shental等提出了一种相关成分分析算法 (RCA)[17], 该算法只利用了 must-link约束关系, 但是忽略了 cannot-link约束关系. Tang和 Zhong[18] 同时利用了 must-link和 canno+ link约束信息来指导降维, 但是他们没有利用大量的有用的无标签数据.

基于对以上的分析, 本文提出了一种基于成对约束的降维算法. 为了避免因样本个数过少或样本维数过高出现的过拟合问题, 本文引入了 Tikhonov正则项来修正[19]. 运用该降维算法对 miRNA 原有的特征进行了降维处理, 去除那些对分类贡献不大的特征或影响分类精度的特征, 从而提高 mRNA 的预测精度.

## 1　基于成对约束的 LDA (Linear Discriminant Analysis)

与样本标签集不同, 成对约束集合并不给出样本的标签信息, 它只提供成对约束的信息. 成对约束具体可分为 must-link约束和 cannot-link约束两种形式[16]. 从有标签的数据集中随机选取 2个样本, 根据其类标是否相同, 把它们分到 must-link约束或 cannot-link约束. Must-link约束表示一个样本对属于同一类, 但是不知道确切的标签信息; cannot-link约束表示该样本对不属于同一类. 成对约束信息是一种比标签信息更一般的信息, 可以从标签信息中得到成对约束信息, 反之则不行, 因为我们不能通过成对约束关系得到样本的标签信息.

若给定的高维数据为 $\boldsymbol{X} = \{x_1, x_2, \cdots, x_n\} \subset \mathbf{R}^D$, 降维后的数据为 $\boldsymbol{Y} = [y_1, y_2, \cdots, y_n] \subset \mathbf{R}^d$, 成对约束集合分别为 $M$ 和 $C$, 其中 $M$ 对应 must-link约束, $C$ 对应 cannot-link约束, $M$ 和 $C$ 表示如下:

$$M = \{(x_i, x_j) \mid x_i \text{和} x_j \text{属同一类}\};$$

$$C = \{(x_i, x_j) \mid x_i \text{和} x_j \text{不属同一类}\}.$$

基于 $M$ 和 $C$, 定义 $S_w$ 和 $S_b$ 如下:

$$S_w = \frac{1}{N_1} \sum_{(x_i, x_j) \in M} (x_i - x_j)(x_i - x_j)^T, \tag{1}$$

$$S_b = \frac{1}{N_2} \sum_{(x_i, x_j) \in C} (x_i - x_j)(x_i - x_j)^T. \tag{2}$$

其中, $N_1$ 和 $N_2$ 分别为 $M$ 和 $C$ 中约束对的个数. 为了使降维后 $M$ 中的数据更加紧凑, 而 $C$ 中的数据更加分散, 定义目标函数 (类似 LDA) 如下:

$$J(w) = \arg\max_w \frac{w^T S_b w}{w^T S_w w}. \tag{3}$$

由 (3) 可知, 它反映了数据的判别信息, 但大量的无标记的样本并没有用到, 因此它不能完全反映出数据的内部结构信息. 为了使得降维算法能更全面地体现出数据的结构信息, 本文提出了基于成对约束的半监督降维算法.

## 2　基于成对约束的半监督降维算法

为有效利用局部结构信息, 使数据结构更加紧凑, 算法引入分层聚类思想. 首先, 在整个训练数据集上进行分层聚类, 把数据聚成几个簇, 使得每个簇中的数据结构更相似. 考虑到数据的距离信息, 这里通过 PCA[20] 的目标函数来保持每个簇中数据的局部结构信息.

$$J(w) = \arg\max_w w^T S_{pca} w, \tag{4}$$

其中, $S_{pca} = S_{pca_1} + S_{pca_2} + \cdots + S_{pca_n}$, $n$ 为聚类的个数.

将 (4) 并入到 (3) 中, 可得新的降维策略如下:

$$J(w) = \arg\max_w \frac{w^T (S_b + \alpha S_{pca}) w}{w^T S_w w}, \tag{5}$$

其中, α为调节参数.

从(5)可见,它的主要思想为:希望利用数据的成对约束信息和数据的局部结构信息来有效进行降维,使得降维后的样本保持局部结构和判别能力.

进一步,为解决因样本维数过高或样本个数过少出现的过拟合问题,借助文献[19]的策略,并入Tikhonov正则项($w^Tw = I$),可得基于成对约束的局部保持半监督降维模型(local semi-supervised linear discriminant analysis, LSLDA),其目标函数如下:

$$J(w) = \arg\max_{w} \frac{w^T(S_b + \alpha S_{pca})w}{w^T(S_w + \sigma I)w}, \tag{6}$$

其中 σ表示正则项系数.

## 3 实验结果

为验证新提出的降维模型的有效性,实验中采用 miRNA数据集和 UCI(University of California, Irvine)数据集,并用不同的分类器和不同的降维方法分别测试新提出的降维方法的有效性. UCI数据集的URL是 http://archive.ics.uci.edu/ml

在所有的实验中,成对约束信息可这样得到:从训练集中随机选取一对样本,如果它们属于同一类,就把它们放到must-link集合;反之,则放到 cannot-link集合. 在 SVM分类器上,用 yasMIR使用的 LibSVM包,惩罚参数 $C$和 RBF核参数 γ用网格搜索法得到,搜索范围在 $(2^{-5}, 2^5)$,并先归一化数据到 $(-1, 1)$.

### 3.1 实验数据

MiRNA数据集的训练集为 TR-C,由 163个人类 mRNA前体正类样本和 168个有类似 mRNA前体结构的负类样本组成.在训练集上,用 5折叠交叉验证和网格搜索法找到最优参数和最好的训练精度.测试集包括 TE-CH、TE-CP、UPDATE和 CONSERVED-HAIRPIN 4个数据集,正负类样本均取自于人类的基因.其中,TE-CH有 30个正类样本无负类样本,TE-CP有 1 000个负类样本无正类样本;UPDATE有 39个正类样本;CONSERVED-HAIRPIN有 3个正类样本和 2 441个负类样本.UCI数据集包括 Heart-disease, WPBCdata, Ionosphere和 Sonar 4个数据集,每个数据集的样本都分为两类.其中,Heart-disease数据集共有 270个样本,13个特征;WPBCdata数据集共有 194个样本,32个特征;Ionosphere数据集共有 351个样本,34个特征;Sonar数据集共有 208个样本,60个特征.随机选取一半的样本作为训练集,剩下的样本作为测试集.所得到的实验结果取 10次不同约束信息下的平均值.模型(6)中的参数取值分别设为 $\alpha = 10$, $\sigma = 0.1$.

### 3.2 实验结果分析

对 miRNA 4个数据集,实验结果如图 1所示.由图 1可见,随着成对约束个数的增加,识别精度开始逐渐提高(如成对约束个数取 50~70时,识别精度取到最优),但随着成对约束个数的进一步增加,分类精度反而有下降的趋势.这说明不是成对约束个数取得越多,分类精度就越高.此外,在 SVM、1NN和 C4.5 3个分类器中, SVM分类器的分类效果最好,在 TE-CH数据集上与 yasMIR一样达到了 100%的正确率;在 TE-CP和 UPDATE 2个数据集上均比 yasMIR的正确率高,尤其是在 UPDATE数据集上.而在 CONSERVED-HAIRPIN数据集上,在某些约束个数上也取得了比 yasMIR好的结果.

为进一步验证提出的降维算法的有效性,论文用不同的降维方法对 mRNA数据集进行了降维,采用 SVM对其分类,分类结果如图 2所示.相对于经典的 PCA和 LDA降维方法而言,新提出的降维算法在 4个数据集上均取得了较好的效果,这进一步证实了结果与理论分析的一致性.

同时,由图 3和图 4可见,新提出的降维算法在 UCI数据集同样取得较好的性能.总之,无论在不同分类器的实验上还是在不同降维方法的实验上,分类精度都基本随着成对约束个数的增加而提高,这说明新提出的算法可适用于其它数据集.

## 4 结语

本文提出了基于成对约束的降维算法,该算法有效利用了比标签信息更一般化的成对约束信息和局部结构信息,在 mRNA 4个数据集和 UCI数据集上,由新的降维算法所诱导出来的分类器均有良好的性能.下一步研究的目标:利用降维子空间进行分类器的集成.

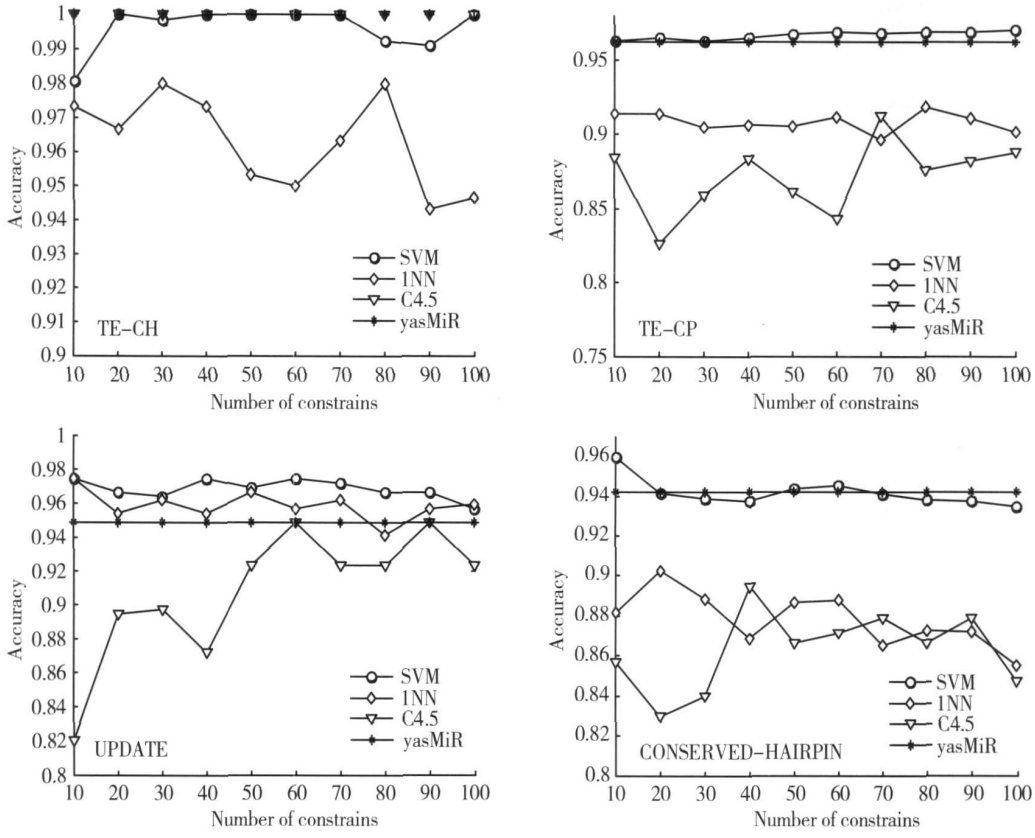


图 1　MiRNA 数据集上分类精度随约束个数变化的结果比较(对不同分类器)

**Fig.1　Classification accuracy on MiRNA data set with different number of constraints (different classifier)**

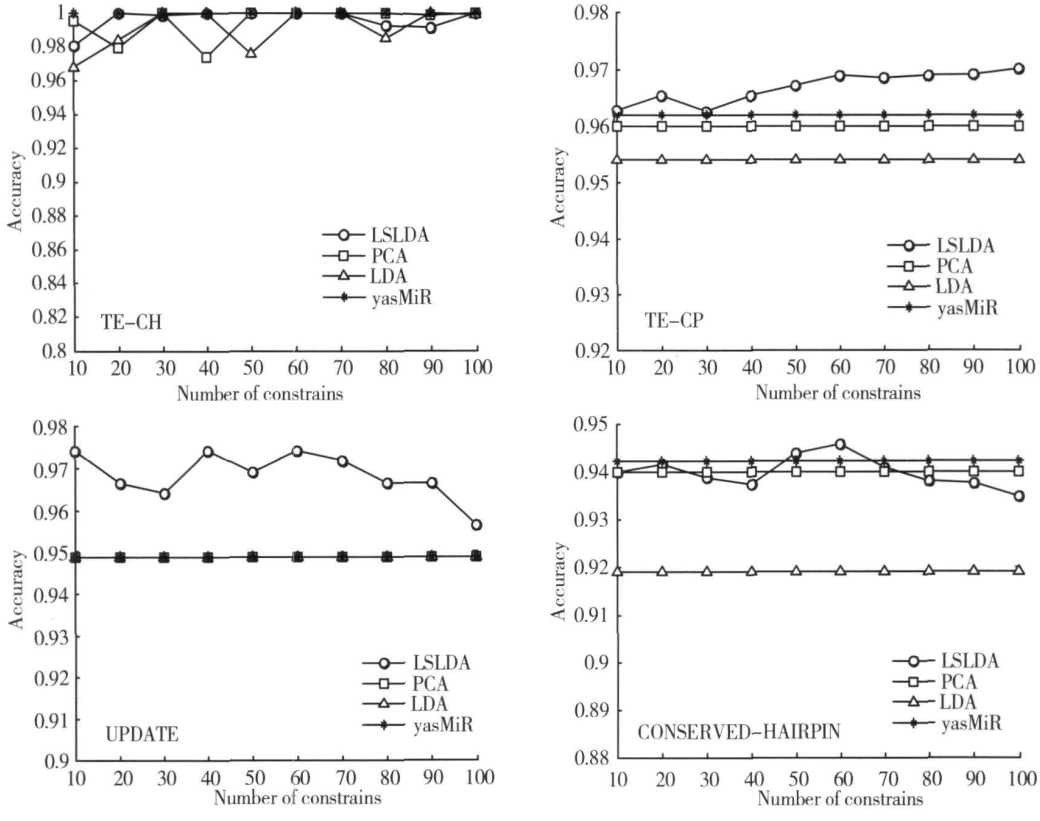


图 2　MiRNA 数据集上分类精度随约束个数变化的结果比较(对不同降维方法)

**Fig.2　Classification accuracy on MiRNA data set with different number of constraints (different dimension reduction method)**

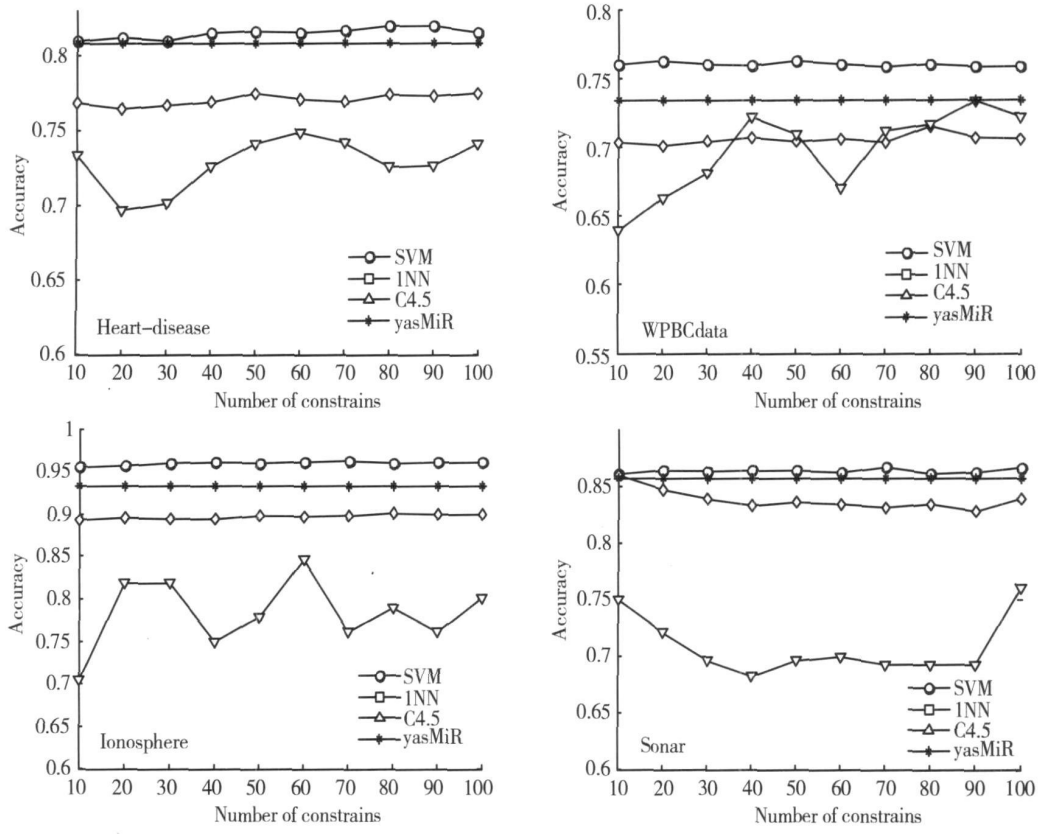


图 3　UCI 数据集上分类精度随约束个数变化的结果比较(对不同分类器)

**Fig.3　Classification accuracy on UCI data set with different number of constraints (different classifier)**

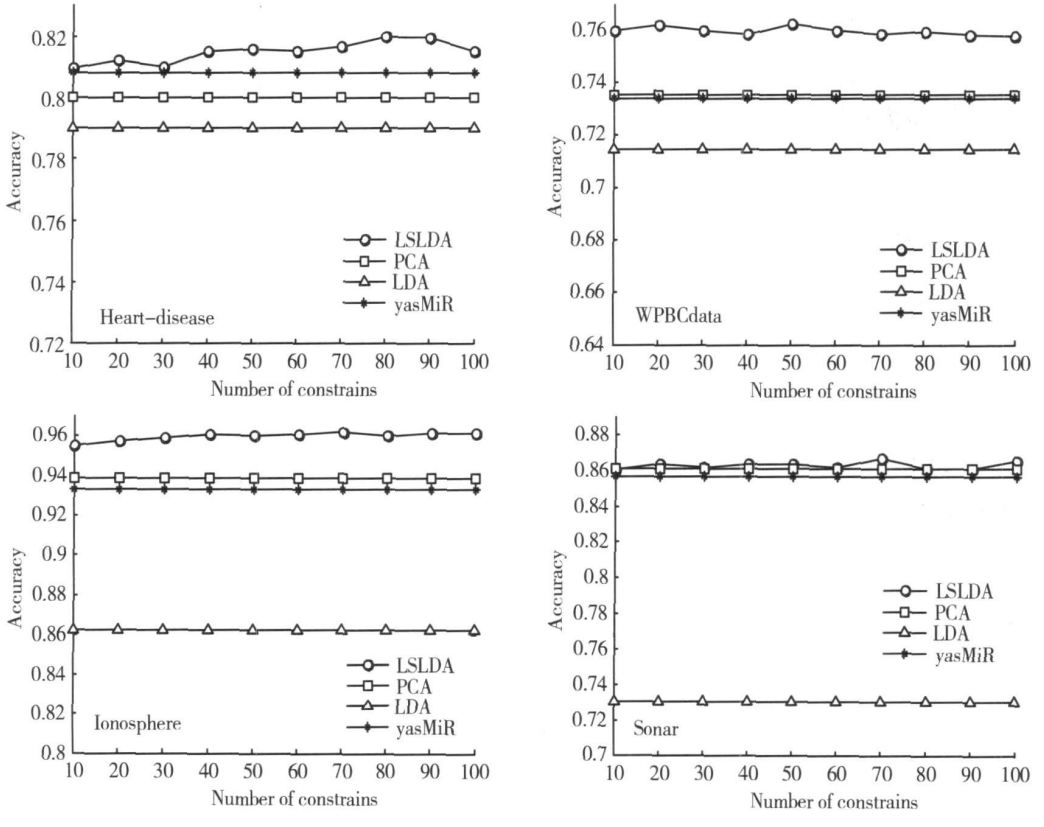


图 4　UCI 数据集上分类精度随约束个数变化的结果比较(对不同降维方法)

**Fig.4　Classification accuracy on UCI data set with different number of constraints (different dimension reduction method)**

[参考文献]


[1] Fire S Xu, Montgomery M, Kastas S, et al. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans[J]. Nature, 1998, 391(6669): 806-811.

[2] Lee Y. MicroRNA maturation: stepwise processing and subcellular localization[J]. EMBO J, 2002, 21(17): 4663-4670.

[3] Bartel D P. MicroRNAs: genomics, biogenesis, mechanism, and function[J]. Cell, 2004, 116(2): 281-297.

[4] Kurihara Y, Watanabe Y. Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions[J]. Proc Natl Acad Sci, 2004, 101(3): 12 753-12 758.

[5] Zhang B. MicroRNAs and their regulatory roles in animals and plants[J]. J Cell Physiol, 2007, 210(2): 279-289.

[6] Tanzer A, Stadler P F. Molecular evolution of a microRNA cluster[J]. J Mol Biol, 2004, 339(2): 327-335.

[7] Molnar A, Schwach F, Studholme D J, et al. miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii[J]. Nature, 2007, 447(7148): 1126-1129.

[8] Kren B T, Wong P Y, Sarver A, et al. microRNAs identified in highly purified liver-derived mitochondria may play a role in apoptosis[J]. RNA Biol, 2009, 6(1): 65-72.

[9] Grosshans H, Slack F J. Micro-RNAs: small is plentiful[J]. J Cell Biol, 2002, 156(1): 17-21.

[10] Lee C T, Risom T, Strauss W M. Evolutionary conservation of microRNA regulatory circuits: an examination of microRNA gene complexity and conserved microRNA-target interactions through metazoan phylogeny[J]. DNA Cell Biol, 2007, 26(4): 209-218.

[11] Vapnik V N. The Nature of Statistical Learning Theory[M]. 2nd ed. New York: Springer-Verlag, 1999.

[12] Breiman L. Random forests[J]. Mach Learn, 2001, 45(1): 5-32.

[13] Xue C. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machines[J]. BMC Bioinformatics, 2005, 6: 310.

[14] Ng K L S, Mishra S K. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures[J]. Bioinformatics, 2007, 23(11): 1321-1330.

[15] Pasaila D, Mohorianu I, Ciortuz L. Using base pairing probabilities for MiRNA recognition[C]// Proceeding of the 10th International Symposium on Symbolic and Numeric Algolithm for Scientific Computing. Timisoara: IEEE Conference Publishing, 2008: 519-525.

[16] Wang Xiumei, Gao Xinbo, Yuan Yuan, et al. Semi-supervised Gaussian process latent variable model based on pairwise constraints[J]. Neurocomputing, 2010, 73: 2186-2195.

[17] Bar-hillel A, Hertz T, Shental N, et al. Learning a mahalanobis metric from equivalence constraints[J]. Journal of Machine Learning Research, 2005, 6: 937-965.

[18] Tang Wei, Zhong Shi. Pairwise constraints-guided dimensinality reduction[C]// SDM Workshop on Feature Selection for Data Mining. Bethesda, 2006.

[19] Cai Deng, He Xiaofei, Han Jiawei. Semi-supervised discriminant analysis[C]// Proceedings of the 11th IEEE International Conference on Computer Vision. Rio de Janeiro, 2007.

[20] Hotelling H. Analysis of a complex of statistical variables into principal components[J]. Journal of Educational Psychology, 1933, 24(1933): 417-441.


[责任编辑: 丁　蓉]