文章编号: 1000-6788(2012)04-0815-11 中图分类号: TP311 文献标志码: A

基于 机森林的潜在 k 近 算法及其在基因表达数据分类中的应用

杨 帆[1], 林 琛[2],

分类器性能, 并据此对样本的分布结构进行深入观测. 丰富的功能使得 迅速成为经典的 数方法之一. 值得一提的是, A , B 都是 B 提出并发展起来的. 正是这些算法的集大成者, 也是 B 学术思想的重要体现 [17].

随机森林因以其下列优点使得其特别适用于基于基因表达数据的癌 诊断 [12,18−24]:

1) 适用于高维小样本数据; 2) 能自动实现特征选择功能且对于无关特征不敏感; 3) 考虑到特征之间的相互作用; 4) 对于二分类问题 多分类问题同样适用; 5) 无需复杂的 数选择过程. 研究者对 包括 在内的其他分类算法在癌 诊断的应用上进行了深入 详尽的比较实验, 其中 [18] 、 [20] 认为, 用于癌 诊断 分类的效果优于包括 在内的一些已知最好的癌 分类方法.

已经成为当前国内外分类 回归方法研究中的前沿 热点, 的作用机制 计学机理是研究者感兴趣的重点之一. [25] 从自适应 近邻 ( ) 的角度分析了 可能的工作机制, 提出了潜在最近邻 (,) 的概念, 指出 可以看作自适应加权的 - : 对于一个待测样本, 随机森林根据输入变量的局部重要性给 - 加权; B [26] 回溯到 ( ) 的概念, 证明了 回归的收敛性, 并通过分析 的联系推导出 的收敛性质 收敛条件; B 等 [27] 又进一步分析了 B ( ) 的关系, 并证明了其收敛性.

上述研究主要从回归的角度研究 , 从理论上部分解释了 的工作机制. 受此启发, 本文从潜在 近邻的角度研究了 算法, 并分析其不足. 第 2 部分首先简要回顾了 A , B 算法, 接着在第 3 部分从分类的角度来分析自适应 近邻 的联系, 给 - 与 进行比较, 分析了 在预测过程中可能存在的信息丢失, 据此提出一种新的投票机制 ( ). 为了对比改进前后的效果, 在第 4 部分选取了国际上公开发表的 6 个癌 基因表达数据集进行了充分的实验, 结果表明改进方法的性能比 有 著提高. 最后, 总结了本文的优缺点 下一步工作.

2 相关研

本文主要研究的是回归 分类问题, 描述如下: 考虑独立同分布的随机变量 (X, Y) 的观测值集合 $S = \{(x_i, y_i), i = 1, 2, \cdots, n\}$, 其中 $X = (X^1, X^2, \cdots, X^d) \in R^d$ 为输入变量, $Y \in R$ 为应变量, R 表示实数集.

2.1 CART

A 是分类树 ( ) 回归树 ( ) 的 称, 由 B 等在 1984 年提出 [13]. A 以基尼指数 ( ) 作为分裂标准, 能够降低数据无序度的属性挑选出来. 在建立 A 时, 分裂属性的选择根据其在不同预测下对样本数据划分的好坏程度来进行的. 根据给 的样本集 S 构建 A 由以下三步组成:

1) 使用 S 构建最大树 h , 确保

他 性

点处, 选择最好的属性分裂方式进行分裂, 直到　子节点; 最后通过剪枝使测试误差最小. 而在随机森林中的单棵树都是未经剪枝的.　　算法描述如下 [12]:

算法 1　Random Forests

输入: 1. 训练集 $S=\ (x_i,\ _i), i=1,2,\quad ,n\ ,(X,Y)\cdot\ R^d\quad R$

2. 待测样本 $x_t\cdot\ R^d$

For $i\ =\ 1,2,\quad ,\ N_{tree}$

(1) 对原始训练集 S Boostrap 抽样, 生成训练集 S_i

(2) 使用 S_i 生成一棵　剪枝的树 h_i:

a. 从 d 个特征中随机选取 M_{try} 个特征

b. 在每个节点上从 M_{try} 个特征依据 Gini 指标选取最优特征

c. 分裂直到树生长到最大

End

输出: 1. 树的集合　$h_i, i=1,2,\quad ,N_{tree}$

2. 对待测样本 x_t, 决策树 h_i 输出 $h_i(x_t)$

回归: $f(x_t)=\frac{1}{N_{tree}}\sum_{i=1}^{N_{tree}}h_i(x_t)$

分类: $f(x_t)=\text{majorit}\quad\text{vote}\ h_i(x_t)\ _{i=1}^{N_{tree}}$

算法 1 中, 用 [illegible] 表示多数投票. 随机森林的泛化误差依赖于以下两个因素:　　中任意两棵树的相关度 ([illegible])　　中单棵树的分类效能 ([illegible]).

定理 1　随机森林泛化误差的上界可以由下式给出 [12]:

$$PE^*\le\rho(1-s^2)/s^2 \tag{1}$$

其中 ρ 是　　中子分类器之间相关度 ρ 的平均值, s 是子分类器 $h(X,\theta_k)$ 的分类效能. 如式 (1) 所示, 为提高　　的预测准确率, 应减小树与树之间的相关度, 同时增大单棵树的分类效能.

随机森林方法采用　A　　算法作为元学习算法, 能同时处理连续属性　类别属性; B [illegible]　随机选择特征分裂的结合, 使该算法能较好地容忍噪声;　　能有效解决不平衡分类问题; 在目前潜那释器输器那尚式票受器权器

(b) (a)

树: (a)

点, 且都是 纯的–, 即仅包含 1 类数据. 在

形区

点上选取使得该节点的 系数达到最小的特征进行分裂, 实现递归划分

个个超矩形区

, 为 Q 义了它在一个单调度量函数 L 下的潜在近邻, 并由

类别. 需要指出的是, 在分类问题中, 与 Q 在决策树 h 同一个 节点上的近

的度量函数 L 原空间中的 式距离是不相同的. 在 2 维 式空间中, 正类样本

分的, 而度量函数 L 却能够使得同类样本之间的相似性达到最大, 异据麻据尽离然理可理离 据看

此在式 (2), (3), (4)　(5) 中存在一　的信息损失, 虽然这部分信息可能很少, 但对分类　回归的效果却有一　的影响, 本文第 4 部分真实　界数据上的实验也验证了这一点.

据此, 本文提出　在预测时应考虑到　B 样本的信息, 对于回归来说, 其计算式 (2)(3) 不改变, 但式子中的 PN_i 应包括　B 样本; 对于分类问题, 见式 (4), 与待测样本　在同一　节点的样本 PN_i 中, 大部分都是　纯的– 训练样本, 其类标签是一致的, 而　B 样本一般仅占少数, 此时如果采取简单多数投票机制, 则　B 样本信息仍然会被训练样本的信息　淹没–, 一些初步的实验也验证了这一点, 因此本文设计了一种新的投票机制, 如式 (6) 所示:

$$f(x_t) = \quad \left(\sum_{x_i \in S} \quad \times Fr(x_i) \right) \tag{6}$$

其中,　$Fr(x_i)$ 表示 x_i 在 N_{tree} 个潜在近邻集合 $\{PN_1, PN_2, \cdots, PN_{N_{tree}}\}$ 中出现的频次.　义频次最高的样本 x 是在该森林下 x_t 的潜在最近邻, 该样本的类标签决　了 x_t 所属类别, 因此简称该算法为　-　(　　　　　　　　) 算法.

算法 2　Random Forests based Potential Nearest Neighbor

输入: 1. 训练集 $S =$　$(x_i,$　$_i), i = 1, 2,$　　$, n$　, $(X, Y) \cdot$　R^d　　R

2. 待测样本 $x_t \cdot$　R^d

(1) 调用算法 1, 生成 RF 模型 $f =$　$h_i, i = 1, 2,$　　$, N_{tree}$

(2) 将全体训练集 S 和待测样本 x_t 同时投入到 RF 中:

For　i　=　$1, 2,$　　, N_{tree}

在决策树 h_i 上寻找与 x_t 同一个叶节点的样本集合 PN_i, 即其潜在近邻

End

输出: 1. 树的集合　$h_i, i = 1, 2,$　　$, N_{tree}$

2. 对待测样本 x_t, 决策树 h_i 输出 PN_i

回归: $f(x_t) = \frac{1}{N_{tree}} \sum_{i=1}^{N_{tree}} \left(\frac{1}{|PN_i|} \sum PN_i \right)$

分类: $f(x_t) = \text{vote}(\underset{x_i \in S}{\text{argma}}\ \text{Fr}(x_i))$

-　　与　　的区别在于以下两点:

1) 在决策树 h_i 上,　　中只有训练样本 S_i　与对待测样本 x_t 本都此成充都比长丢丢第第都此得本到比都不独本到不度

) B 细胞 急性淋巴细胞白血病 细胞 (A - -).

Lung Cancer 数据集来源于美国国立癌 研究所 [8]. 该数据集包含 186 例肺癌病例样本 17 例正常病例样本. 在 186 例病例样本中包含 139 例肺腺癌、21 例鳞状细胞肺癌、20 例肺类癌 6 例小细胞肺癌 .

Tumor 1 数据集来源于美国国立癌 研究所的发展 计划 一. 该计划致力于针对不同的癌 细胞寻找 预测其化学敏感性. 数据集包含 9 种人类常见的癌 肿瘤病例样本, 这 9 种癌 肿瘤包含肺癌肿瘤、乳腺癌肿瘤、结肠癌肿瘤、肾癌肿瘤、骨髓癌肿瘤、中枢神经肿瘤、前列腺肿瘤 卵巢肿瘤.

Tumor 2 数据集来源于弗吉尼亚大学 [10]. 数据集中包含 11 种人类常见的肿瘤病例样本的基因数据. 这 11 种癌 肿瘤的样本包含前列腺癌、膀胱/尿道癌 (移行细胞癌 鳞状细胞癌)、浸润性乳腺导管癌、直肠癌、胃腺癌、透明肾细胞癌、肝癌、卵巢浆液性乳头状腺癌、胰腺癌 肺癌 (肺腺癌 鳞状细胞癌).

表 1 6 个癌症基因表达数据集描述

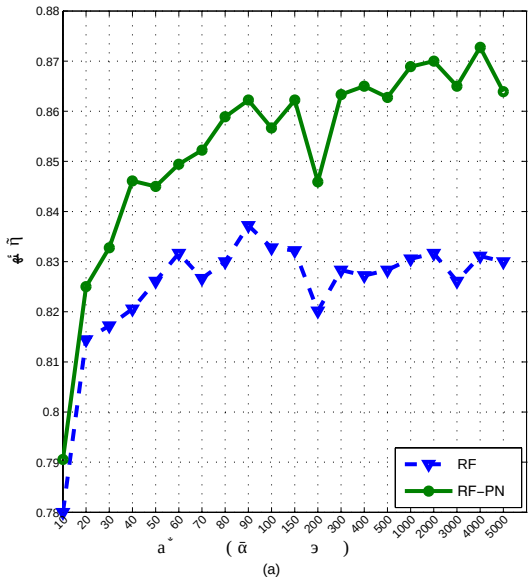
数据集名稳	诊 任务	样本数	基因数	类别	基因/样本之比
B	5 种脑部肿瘤	90	5920	5	66
B	弥漫性大				

为了更细致的展示 100 次重复实验的结果, 图 38 ( ) 采用箱线图的方式, 示 100 次试验中 5 折交叉平均准确率的中位数、75% 25% 分位数等, 用红色箱线代表 的 100 次实验结果, 蓝色箱线代表 - 的 100 次实验结果, 箱中的 圈代表中位数. 横轴标记不同大小的森林下两种不同方法, 例如 -10– 表示包含 10 棵决策树的森林, -10 – 标记包含 10 棵决策树的 - 算法.

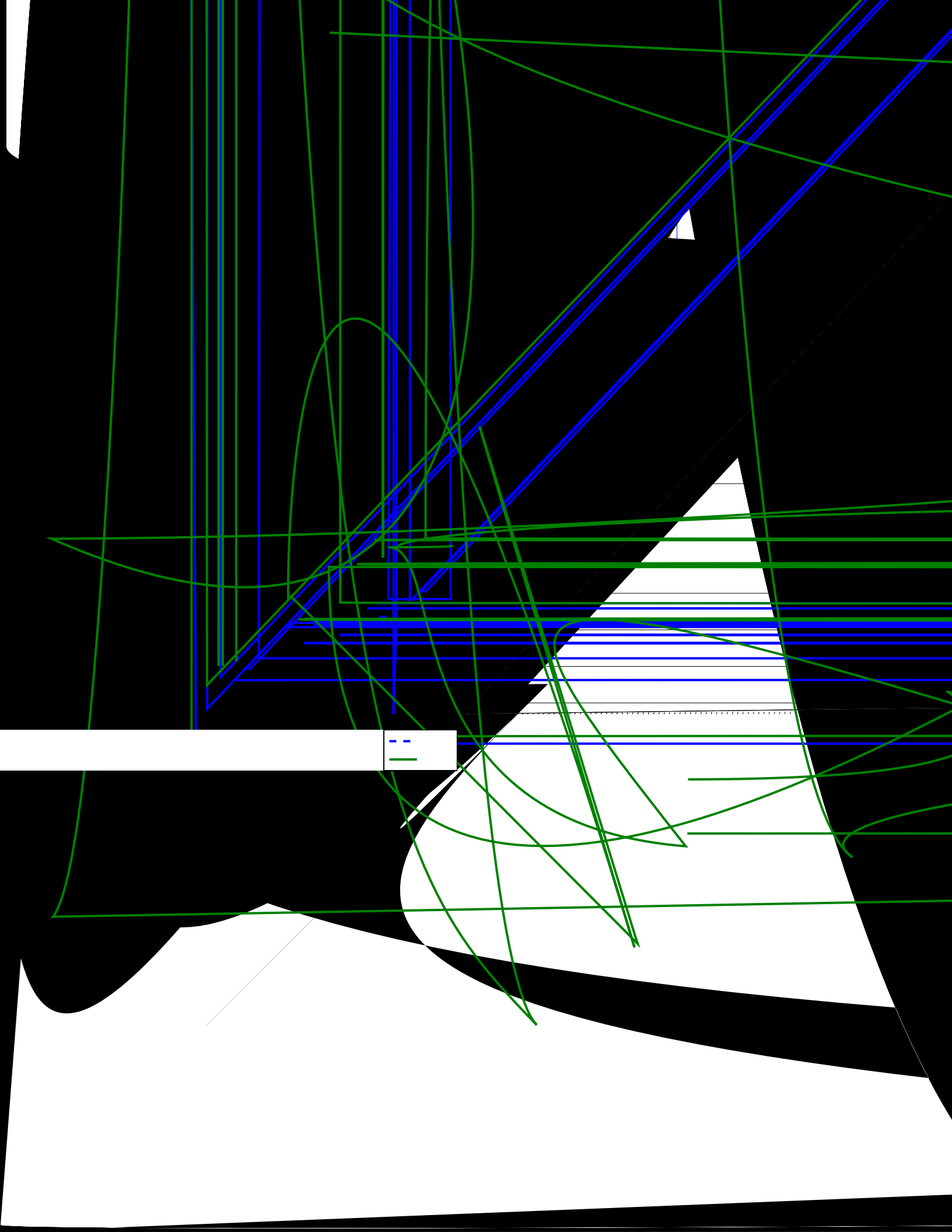
记 Acc_1 为 的平均分类准确率, Acc_2 为 - 的平均分类准确率. 表 2 记录了 6 个数据集上森林规模从 10 增加到 5000 时, - 的平均分类准确率的差值, 即 $Acc_2 - Acc_1$.

表 2　不同规模森林下 6 个数据集上 RF-PN 与 RF 平均分类准确率之差 (%)

森林规模	B	B		-	1	2
10	1.06	0.03	**3.44***	0.60	1.00	2.39
20	1.06	0.03	3.32	0.70	3.17	1.60
30	1.56	1.00	1.72	0.14	1.25	2.21
40	2.56	0.89	2.21	**1.06***	2.33	2.70
50	1.89	0.86	2.10	0.55	3.50	**2.93***
60	1.78	1.30	2.65	0.54	3.75	1.80
70	2.56	1.04	2.67	0.65	4.00	2.00
80	2.89	1.07	2.80	0.52	3.83	2.31
90	2.50	0.71	2.28	0.18	3.42	1.94
100	2.39	1.00	2.42	0.91	4.08	2.52
150	3.00	1.00	2.04	0.64	5.08	2.12
200	2.58	0.95	2.64	0.93	4.63	2.59
300	3.50	**1.39***	2.48	0.94	7.92	2.32
400	3.78	1.00	2.48	0.89	7.25	2.22
500	3.44	0.50	1.66	0.83	8.58	2.36
1000	3.83	1.36	1.95	0.58	**9.50***	1.64
2000	3.83	0.89	2.15	0.63	8.50	1.89
3000	3.89	1.00	2.21	0.64	7.42	2.06
4000	**4.17***	1.11	1.81	0.91	8.58	2.88
5000	3.39	0.86	1.88	0.80	7.00	2.44



(a)




0
0.
0.7
0.72

[2] Van't Veer L J, Dai H, Van de Vijver M J, et al. Gene e pression profiling predicts clinical outcome of breast cancer[J]. Nature, 2002, 415(6871): 530.
[3] Pomero S L, Tama o P, Gaasenbeek M, et al. Prediction of central nervous s stem embr onal tumour outcome based on gene e pression[J]. Nature, 2002, 415(6870): 436 442.
[4] Ali adeh A A, Eisen M B, Davis R E, et al. Distinct t pes of diffuse large B-cell l mphoma identified b gene e pression profiling[J]. Nature, 2000, 403: 503 511.
[5] Valafar F. Pattern recognition techniques in microarra data anal sis a surve [J]. Annals of the New York Academ of Sciences, 2002, 980(1): 41 64.
[6] Brown Grund W N, Lin D, Cristianini N, et al. Knowledge-based anal sis of microarra gene e pression data b using support vector machines[J]. Proceedings of the National Academ of Sciences, 2000, 97(1).
[7] Shipp M A, Ross K N, Tama o P, et al. Diffuse large B-cell l mphoma outcome prediction b gene-e pression profiling and supervised machine learning[J]. Nature Medicine, 2002, 8(1): 68 74.
[8] Bhattacharjee A, Richards W G, Staunton J, et al. Classification of human lung carcinomas b mRNA e pression profiling reveals distinct adenocarcinoma subclasses[C]// Proceedings of the National Academ of Sciences, 2005, 21(15): 3301 3307.
[9] Staunton J E, Slonim D K, Coller H A, et al. Chemosensitivit prediction b transcriptional profiling[C]// Proceedings of the National Academ of Sciences, 2001, 98(19): 10787.
[10] Su A I, Welsh J B, Sapinoso L M, et al. Molecular classification of human carcinomas b use of gene e pression signatures[J]. Cancer Research, 2001, 61(20): 7388.
[11] 刘叶青, 刘三 , 饒明涛. 多项式光滑的半监督 持向量分类机 [J]. 系统工程理论与实践, 2009, 29(7): 113 118.
Liu Y Q, Liu S Y, Gu M T. Pol nomial smooth semi-supervised support vector classifier[J]. S stems Engineering Theor & Practice, 2009, 29(7): 113 118.
[12] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5 32.
[13] Breiman L, Friedman J H, Olshen R A, et al. Classification and Regression Trees[M]. Cole, Pacific Grove, California, USA: Wadsworth & Brooks, 1984.
[14] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123 140.
[15] Ho T K. The random subspace method for constructing decision forests[J]. IEEE Transactions on Pattern Anal sis and Machine Intelligence, 1998, 20(8): 832 844.
[16] Breiman L. Arcing classifiers[J]. Annals of Statistics, 1998: 801 824.
[17] Breiman L. Statistical modeling: The two cultures[J]. Statistical Science, 2001: 199 215.
[18] Wu B, Abbott T, Fishman D, et al. Comparison of statistical methods for classification of ovarian cancer using mass spectrometr data[J]. Bioinformatics, 2003, 19: 1636 1643.
[19] Lee J W, Lee J B, Park M, et al. An e tensive comparison of recent classification tools applied to microarra data[J]. Computational Statistics & Data Anal sis, 2005, 48: 869 885.
[20] Día Uriarte R, de Andrés A. Gene selection and classification of microarra data using random forest[J]. BMC Bioinformatics, 2006, 7(1): 3.
[21] Yang F, Wang H Z, Mi H, et al. Using random forest for reliable classification and cost-sensitive learning for medical diagnosis[J]. BMC Bioinformatics, 2009, 10(Suppl 1): S22.
[22] Statnikov A, Wang L, Aliferis C F. A comprehensive comparison of random forests and support vector machines for microarra -based cancer classification[J]. BMC Bioinformatics, 2008, 9(1): 319.
[23] Strobl C, Boulestei A L, Zeileis A, et al. Bias in random forest variable importance measures: Illustrations, sources and a solution[J]. BMC Bioinformatics, 2007, 8(1): 25.
[24] Strobl C, Boulestei A L, Kneib T, et al. Conditional variable importance for random forests[J]. BMC Bioinformatics, 2008, 9(1): 307.
[25] Lin Y, Jeon Y. Random forests and adaptive nearest neighbors[J]. Journal of the American Statistical Association, 2006, 101(474): 578 590.