高维数据的特征选择研究

杨 杨[1],吕 静[2]

(1. 南京师范大学 强化培养学院,江苏 南京 210046)
(2. 南京师范大学 计算机科学与技术学院,江苏 南京 210046)

[摘要] 特征选择是机器学习的重要研究内容之一. 相对于低维数据的特征选择而言,高维数据的特征选择更具挑战性,尤其是高维小样本的特征选择问题,因而吸引很多研究者的关注. 高维特征选择问题称为稀疏建模问题,其目标是解决现有特征建模方法在高维特征空间失效的问题. 本文对高维数据的特征选择研究成果进行了相应的总结和展望.

[关键词] 高维数据,降维,特征选择
[中图分类号] TP311 [文献标志码] A [文章编号] 1672-1292(2012)01-0057-07

Some Studies on Feature Selection for High Dimensional Data

Yang Yang[1], Lü Jing[2]

(1. Honor School, Nanjing Normal University, Nanjing 210046, China)
(2. School of Computer Science and Technology, Nanjing Normal University, Nanjing 210046, China)

Abstract: Feature selection is a key issue in machine learning field. As compared with feature selection for low dimensional data, feature selection for high dimensional data is a challenging task, especially feature selection issue for high dimensional small size data, so many researcher focus on this problem. In essence, the feature selection problem for high dimensional data is regarded as a sparse modeling issue, whose target is to solve the failure problem of the existing feature modeling methods on high dimensional feature space. Therefore, in this paper, we give a survey of the feature selection methods for high dimensional data, and meanwhile propose some discussions on future work. Our main objective is to provide a reference for readers who are interesting in this research field.

Key words: high dimension data, dimensionality reduction, feature selection

降维是机器学习、数据挖掘和模式识别等研究领域的关键内容之一[1]. 降维技术包括特征抽取技术[2-4]和特征选择技术[1,5,6],其中特征抽取是通过一个变换将原高维空间映射到一个新的低维特征空间,特征选择是从原高维特征空间中选择相关且判别能力强的一个特征子集. 特征抽取和特征选择的共同目标均是将高维空间降低到低维空间,以此来克服因高维数据带来的分类器的过拟合问题,进而增强其泛化性,改进其效率.

已有研究成果表明,特征抽取可有效改进分类器的性能,受到研究者的广泛关注[2-4],尤其是近年来研究者提出的基于流形假设的局部保持投影方法(Local Preserving Projections,LPP),更吸引了机器学习和模式识别专家的重视[3],且基于流形假设的思路还被有效地应用于半监督降维中[4]. 但基于变换的特征抽取是将原特征空间降维到一个低维特征空间,那些不相关和冗余的特征均在降维过程中产生了作用,因而不可避免地影响分类器的性能,且在新的低维特征空间中特征失去原有的物理含义,因此其解释性差.

同时,在某些实际应用中,需要确切知道哪些重要的特征在预测中起到关键作用,显然这就是典型的特征选择问题. 对于这类问题,采用特征抽取策略显然是不合适的,而特征选择方法的可解释性明显增强. 为此,本文在介绍基于流形假设的特征抽取方法 LPP 之后,侧重对高维数据的特征选择研究进行综述,并对未来的研究进行展望.

收稿日期: 2012-01-08.
基金项目: 南京师范大学 2010 年学生科学基金(首批立项).
通讯联系人: 吕 静,讲师,研究方向: 模式识别理论与应用. E-mail: 05275@njnu.edu.cn

1 相关研究工作

1.1 几种经典的特征抽取算法

经典的特征抽取方法包括主成份分析(Principal Component Analysis,PCA)[1,2]、线性判别分析(Linear Discriminant Analysis,LDA)[1,2]、局部保持投影(Local Preserving Projections,LPP)[3]和半监督线性判别分析(Neighorhood Preserving Embedding,NPE)[4]. 下面将简单介绍这几种方法的主要思路.

PCA 的主要思路是: 寻找一个投影矩阵 W 使得重建误差最小,对给定的 l 个样本 $x_1,x_2,\cdots,x_l$,PCA 的目标函数定义如下:

$$\arg\max_{W} tr(W^{\mathrm{T}}S_T W), \tag{1}$$

其中,$S_T=\sum_{i=1}^{i}(x_i-\mu)(x_i-\mu)^{\mathrm{T}}$,$\mu$ 是所有样本的均值.

LDA 的主要思路是: 寻找一个投影矩阵使得在投影后的特征空间中不同类样本分得更开而同类样本更加紧. 对给定的 l 个有标签的样本 $x_1,x_2,\cdots,x_l$,LDA 的目标函数定义如下:

$$W^{*}=\arg\max_{W}\frac{tr(W^{\mathrm{T}}S_b W)}{tr(W^{\mathrm{T}}S_w W)}, \tag{2}$$

其中,$S_b=\sum_{k=1}^{c}l_k(\mu^{(k)}-\mu)(\mu^{(k)}-\mu)^{\mathrm{T}}$,$S_w=\sum_{k=1}^{c}\left(\sum_{i=1}^{l_k}(x_i^{(k)}-\mu^{(k)})(x_i^{(k)}-\mu^{(k)})^{\mathrm{T}}\right)$,$\mu$ 是所有样本的均值,l_k 为第 k 类样本数,$\mu^{(k)}$ 为第 k 类样本的均值,$x_i^{(k)}$ 是第 k 类样本中第 i 个样本,S_b 和 S_w 分别称为类间散布矩阵和类内散布矩阵.

在实际应用中,对小样本问题,直接采用 LDA 或 PCA 均会影响分类器的性能. 通常先进行 PCA 降维后,再进行LDA降维,以避免LDA求解中判别矩阵的奇异性,但该策略会因PCA丢失一些判别信息而影响性能. 为此,研究者采用 LDA + PCA 策略,先 LDA 再 PCA,利用 LDA 零空间的投影信息. 针对高维小样本问题,文献[2]的作者提出了一种适用于小样本问题的基于间隔的特征提取算法,该算法利用高维数据小样本情况下线性可分概率增加以及其低维投影趋于正态分布的特点,定义了新的类别边界,同时考虑了由线性判别分析提出的类内、类间离散度以及类别的方差差异性.

LPP 算法的主要思路是: 基于流形假设,即所谓高维空间的近邻通过降维($y=W^{\mathrm{T}}x$,W 是投影矩阵,W^{T} 是 W 的转置)后在低维空间中仍保持近邻关系. 基于该思路,对给定的 l 个样本 $X=(x_1,x_2,\cdots,x_l)$,LPP 算法的目标函数定义如下.

$$\arg\min_{W}\sum_{i,j}(y_i-y_j)^2 S_{ij}=tr(W^{\mathrm{T}}X(D-S)X^{\mathrm{T}}W), \tag{3}$$

这里,

$$S_{ij}=\begin{cases}\exp\left(\dfrac{-\|x_i-x_j\|^2}{t}\right), x_i\in N_k(x_j) \text{ 或 } x_j\in N_k(x_i),\\ 0, \text{otherwise};\end{cases}$$

或

$$S_{ij}=\begin{cases}\exp\left(\dfrac{-\|x_i-x_j\|^2}{t}\right), \|x_i-x_j\|^2<\varepsilon,\\ 0, \text{otherwise}.\end{cases}$$

其中,x_i 和 x_j 为两个样本,$N_k(x_j)$ 为 x_j 的 k-近邻,$y_i=W^{\mathrm{T}}x_i$,ε 为大于 0 的数. D 是一个对角矩阵,其对角元素为对应行或列的和.

为了有效利用那些容易得到的无标签样本,文献[4]的作者将流形假设的思路运用到半监督线性判别分析中. 设 $X=(x_1,x_2,\cdots,x_l,x_{l+1},\cdots,x_m)$ 为给定的样本集合,$(x_1,x_2,\cdots,x_l)$ 为有标签样本的集合,而 $(x_{l+1},x_{l+2},\cdots,x_m)$ 为无标签样本的集合,得到 NPE 算法的优化函数如下.

$$\arg\max_{w}\frac{tr(w^{\mathrm{T}}S_b w)}{tr(w^{\mathrm{T}}(S_t+\alpha X(D-S)X^{\mathrm{T}})w)}, \tag{4}$$

其中,$S_t = S_b + S_w$ 是定义在有标签样本的集合$(x_1, x_2, \cdots, x_l)$上,S 定义在所有样本的集合$X = (x_1, x_2, \cdots, x_l, x_{l+1}, \cdots, x_m)$上,$\alpha$ 为可调参数.

1.2　经典的高维数据的特征选择方法

特征选择方法大致可分为 Filter、Wrapper 和 Embedded 模型[1,5,6]. Filter 模型通过某个适应函数(Fitness Function)的值来估计某个特征子集的有效性,与具体的分类器无关; Wrapper 模型是用某个特定分类器的性能作为特征子集选择的准则,这种直接优化分类器的策略可改进分类器的泛化性,但计算代价相对较高,且不具有通用性; 而 Embedded 模型为同时进行特征选择和学习器设计. 从是否使用监督信息角度看,特征选择大体上分无监督特征选择、监督特征选择和半监督特征选择方法.

无论采用何种特征选择模型,如何评价特征或特征子集的重要性或判别能力是其关键. 为此,研究者围绕特征或特征子集重要性的判别准则,提出了大量的特征选择算法,尤其是针对高维数据,提出了基于启发式准则[7-11]、间隔[12-14]、信息熵[15]、0 或 1-范数[16,17]、遗传算子[18-21]、集成学习[22,23]、ROC[24]、聚类[25]、代价敏感及投票等其他针对不平衡策略[26-30]的特征选择方法.

1.2.1　启发式特征选择方法

针对高维数据,为提高特征选择的效率,文献[8]提出了 SFS(Sequential Forward Selection)和 SFFS(Sequential Floating Forward Selection),SFS 和 SFFS 均采用重复迭代方法,逐次选择重要的特征加入,形成一个特征排序序列,但这两种方法没有考虑特征之间的相关性. 文献[9]提出了 FS_SFS(Filtered and Supported Sequential Forward Search),通过在动态样本子集上训练单个特征的支持向量机,提出了一种同时考虑单个特征的判别能力和特征之间的相关性策略,但很高维的基因数据,其效率仍然很低. 为提高效率,文献[10]针对单个基因删除方法存在算法效率低,而多个基因删除方法可能会删除一些有用的基因的不足,提出给各基因排序打分,通过一级差分将分数相近的基因分为各个组,并对排序准则分数值最小的基因小组进行递归特征去除,消去噪声基因,而对排序准则分数值最大的基因小组进行 SFS 策略选取具有有效信息的基因. 该方法可有效改进特征选择的有效性,但组数过大或过小会降低分类器的性能,因此,该方法除了如何设定组数外,选取的基因个数也是一件具有挑战性的研究工作. 文献[11]提出了基于 SVM 的迭代选择(Iterative Reduced Forward Selection,IRFS)算法 SVM-IRFS,其主要策略是采用自适应阈值来过滤不相关的基因特征,避免不必要的计算. 该算法显著提高了效率且改进或确保分类精度没有明显下降.

1.2.2　基于假设间隔的加权特征选择方法

文献[12]提出了著名的特征加权方法 Relief,该算法依据下面称为假设间隔的策略,即对给定的样本集合 P 和样本 x,其假设的间隔就是该假设与其他类的最近假设之间的距离.

$$\theta_P(x) = \frac{1}{2}\left(\| x - \text{nearmiss}(x) \| - \| x - \text{nearhit}(x) \| \right), \tag{5}$$

其中,nearhit(x) 和 nearmiss(x) 分别表示在 P 中与 x 最近的同类样本和最近的不同类样本. 依据式(1),希望选择的特征子集对应的假设间隔尽可能大.

Relief 算法的主要步骤如下:

Step 1　初始化权重向量 $W = 0$;

Step 2　for $t = 1$ to T do

(a) 从给定训练样本集合 S 中随机选择一个样本 x;

(b) 依据式(5)计算 calculate nearmiss(x) 和 nearhit(x);

(c) for $i = 1 \cdots N$,

$$w_i = w_i + (x_i - \text{nearmiss}(x)_i)^2 - (x_i - \text{nearhit}(x)_i)^2;$$

Step 3　对 W 向量中的各分量按照权重的大小由大到小排序.

Relief 算法的主要不足是,每次求解样本的同类近邻和异类近邻时没有考虑各特征的贡献,且该算法的效率还需提高. 为此,文献[13]提出了 Simba(Iterative Search Margin Based Algorithm)算法,该算法的思路类似 Relief,但度量假设间隔时巧妙地运用已得到的权向量,即对特征集上的权向量 w,x 的假设间隔定义为:

$$\theta_P^w = \frac{1}{2}\left(\| x - \text{nearmiss}(x) \|_w - \| x - \text{nearhit}(x) \|_w \right), \tag{6}$$

其中,$\| z \|_w = \sqrt{\sum_i w_i^2 z_i^2}$.

以此准则,Simba 算法为提高效率,结合以下的梯度上升策略.

$$(\nabla e(w))_i = \frac{\partial e(w)}{\partial w_i} = \sum_{x \in S} \frac{\partial \theta(x)}{\partial w_i} = \frac{1}{2} \sum_{x \in S} \left(\frac{(x_i - \text{nearmiss}(x)_i)^2}{\| x - \text{nearmiss}(x) \|_w} - \frac{(x_i - \text{nearhit}(x)_i)^2}{\| x - \text{nearhit}(x) \|_w} \right) w_i \tag{7}$$

这里,$e(w) = \sum_{x \in S} \theta_{(S-\{x\})}^w(x)$,$e(w)$ 为特征子集优劣的评价函数.

Simba 算法的主要步骤描述如下:

Step 1 初始权向量 $w = (1,1,\cdots,1)$;

Step 2 for $t = 1$ to T do

(a) 从给定训练样本集合 S 中随机选择一个样本 x;

(b) 依据式(6) 计算 calculate nearmiss(x) 和 nearhit(x);

(c) for $i = 1$ to N do

$$\Delta_i = \frac{1}{2} \left(\frac{(x_i - \text{nearmiss}(x)_i)^2}{\| x - \text{nearmiss}(x) \|_w} - \frac{(x_i - \text{nearhit}(x)_i)^2}{\| x - \text{nearhit}(x) \|_w} \right) w_i;$$

(d) $w = w + \Delta$.

Step 3 $w \leftarrow w^2 / \| w^2 \|_\infty$ when $(w^2)_i = (w_i)^2$.

Step 4 对 W 向量中的各分量按照权重的大小由大到小排序.

随着 $\| w \|$ 不断增加,修正项 Δ 的相对作用逐步减少,因而算法是收敛的. 依据文献[13]的分析,算法 Simba 的计算复杂度为 $O(TNm)$,其中 T 为迭代次数,N 为特征数目,m 为样本集 S 中的样本数目. 可见,Simba 是一种高效的特征选择算法,且文献[13]的大量实验已验证 Simba 胜过 Relief.

文献[14]综合运用已存在的各种算法,如 ReliefF(Relief 算法的扩展,可处理多类)、SBS、SFS、SFBS、聚类算法、遗传算法 GA 和模拟退火算法 SN、Tabu 搜索(TS)和神经网络等,充分利用遗传算法对高维数据的特征选择效果差,而 Tabu 搜索效果好等特点,通过降维将高维数据转化为低维数据的策略,提出 Relief + C 均值聚类、Relief + SFFS、SFFS + C 均值、Relief + SFFS + C 均值聚类特征选择模型,一定程度上提高了特征选择算法的性能.

1.2.3 基于信息熵的特征选择方法

文献[15]的作者采用信息熵和互信息策略,提出了最大依赖、最大相关和最小冗余的特征选择准则 mRMR(minimal-Redundancy-Maximal-Relevance Criterion). 设 $F = \{f_i\}_{i=1}^m$ 为样本的特征集合,c 为目标类特征,mRMR 的主要思路为:寻找到与目标类 c 具有最大相关度且特征之间相关度尽可能小的特征子集. 依据该准则且为快速得到一个特征序列,文献[15]提出了如下模型:

$$\max_{f_j \in F - F_{m-1}} \left[I(f_j, c) - \frac{1}{m-1} \sum_{f_i \in F_{m-1}} I(f_i, f_j) \right], \tag{8}$$

其中,$I(x,y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \mathrm{d}x\mathrm{d}y$,$F_{m-1}$ 为已经得到的特征子集. 式(8)的目标是从其余的特征子集 $F - F_{m-1}$ 选取一个与目标类 c 相关度最大,但与 F_{m-1} 中的特征相关度最小的特征 f_j.

1.2.4 嵌入 0 或 1-范数约束的特征选择方法

除了采用 Filter 模型下的启发式搜索方法外,将分类模型和特征选择(稀疏化)设计在同一个模型下进行求解也是目前有效的方法之一. 尤其对高维数据,嵌入 0 或 1-范数约束到分类模型受到研究者的重视并得到深入的研究. 文献[16,17]在已有的工作基础上,提出了两种有效的模型.

文献[16]针对 0-范数约束方法引入支持向量机使得其目标函数非凸,导致求解困难,而 1-范数约束方法存在目标函数不是严格凸且特征选择能力受限、样本容量等不足,从经验风险最小化角度出发,提出如下一个基于零范数约束的特征选择模型:

$$\hat{\beta} = \arg\min_{\beta} \{ \| Y - X\beta \|^2 + \xi \| \beta \|^2 + \lambda \| \beta \|_0 \}, \tag{9}$$

在式(9)中,当 $\|\beta\|_0$ 被看成常数时(即特征个数确定或同级时),上述模型退化为一个严格凸优化问题,其优化解可容易得到.

进一步,对高维数据,针对 1-范数约束正则化解决方案存在缺乏特征组选能力和特征选择能力受限、样本容量等不足,文献[17]提出一种基于 0-范数约束的特征选择模型 SCOC(Stochastic Complexity Optimization),该模型为通过随机复杂度理论的模型冗余度诱导得到,且容易求解.

1.2.5　基于遗传算法的特征选择方法

基于 Filter 模型的启发式特征选择方法确实可有效地得到一个特征子集,但不能得到多个有效的特征子集,有可能最有效的特征子集未必能够得到. 为此,研究者提出了基于遗传算法的特征选择方法[18-21],这些算法的共同特点是: 发现一个有效的编码方法[20]; 结合领域知识或已有的监督信息设计出有效的适应度函数,例如文献[21]提出遗传算法采用类内类间比作为适应度函数; 寻找高效的求解策略,避免早收敛且确保快速收敛到近似最优解,例如文献[18]针对从高维特征空间中搜寻一个优化的特征子集是一件 NP-complete 问题,提出一个混合算法 SAGA(Search Algorithm and Genetic Algorithms),该算法采用遗传算法的交叉算子,结合模拟退火策略,可避免陷入局部最小,有着很强的局部搜索能力,且计算效率较高. 在人工合成数据集和实际数据集上的实验结果表明,SAGA 相对于其他已存在的算法有较好的性能.

1.2.6　集成特征选择方法

除了采用单个特征选择方法进行有效的特征选择外,采用多个特征选择方法进行特征选择受到研究者的关注. 文献[22]针对单个分类器在高维小样本的微阵列数据上存在分类性能差的不足,分析现有一些针对微阵列数据的集成分类算法存在着分类精度低、计算代价高的原因后,提出了一种基于相关性分析的微阵列数据集成分类算法. 该算法的主要思路为: 改变训练样本集,并采用信噪比(Signal Noise Ratio, SNR)、皮尔森相关系数(Pearson Correlation, PC)、斯皮尔曼相关系数(Spearman Correlation, SC)、欧几里德距离(Euclidean Distance, ED)和余弦相似度(Cosine Coefficient, CC) 5 种不同特征选择策略对相应的训练样本进行特征选择; 对得到的侯选特征集合,通过计算侯选特征子集间的相关性挑选出差异度最大的一组子集来进行训练, 以此来增强子分类器的多样性. 该算法在急性白血病与结肠癌数据集上进行了有效的验证.

类似于现有集成特征选择方法,文献[23]提出了一种可有效避免过拟合且适合微阵列数据的特征选择准则(GoodCF Criterion),该准则还可适用于其他高维小样本数据,其主要思路为: GoodCF 通过组合多个特征选择方法,以避免过拟合.

1.2.7　代价敏感等其他特征选择方法

除了上述主要的方法,将聚类方法嵌入到特征选择方法中也是一种新方法,例如文献[24]提出了一种新的聚类有效性度量准则(Adjusted Rand Index,ARI),将该准则用于特征选择,并与统计测试及 ROC 策略实验比较的结果验证了该准则的有效性.

在实际应用中,样本除了高维外,常常还是不平衡的,即某类样本相对于其他类样本很少,例如文本数据[25-27]和基因数据[28-30]. 针对这类问题,研究者进行了深入的研究,取得了一定的进展.

文献[26]提出了一种新的文本特征选择算法——改进的基尼指数算法,该算法在不同的数据集和不同的分类器上都表现出较好的分类性能. 文献[27]考虑高维不平衡数据的特点,分析了 IG(Information Gain)、CHI(Chi-Square)、CC(Correlation Coefficient) 和 OR(Odds Ratios)的优势和不足,提出了基于多项式朴素贝叶斯和正规化的罗吉斯回归(Logistic Regression)的特征选择方法.

如文献[28]指出的那样,除了对文本分类问题进行特征选择研究外,很少有研究工作考虑其他高维分类问题的特征选择,尤其是高维小样本问题的特征选择. 于是,文献[28]首次用 7 种特征选择方法、3 种不平衡问题的分类算法和 ROC(Receiver Operating Characteristic) 及 PRC(Precision Recall Curve) 评价准则对来自不同应用领域中的不平衡分类问题进行了系统的实验比较,可以观察到采用信号噪声相关系数(Signal-to Noise Correlation Coefficient,S2N) 和滑动阈值特征估计(Feature Assessment by Sliding Thresholds,FAST) 可取得较好的效果. 当然,进一步的研究还需要进一步探索,如特征选择方法是否鲁棒等问题. 而文献[30]采用投票的策略增强了特征选择的有效性.

2 结语

针对高维数据,本文对特征抽取和特征选择的相关研究进行了综述,简介了传统的PCA和LDA后,侧重介绍基于流形假设的LPP方法的主要思路,并对其拓展模型NPE进行了阐述,为半监督降维和分类器设计提供了新的途径.同时,本文重点对高维数据的特征选择方法进行了综述,介绍基于启发式策略、间隔、信息熵、遗传算法、0或1-范数约束、集成学习、代价敏感及聚类等其他策略的特征选择方法,为高维数据的特征选择理清了思路,也指出了未来的研究热点.

针对高维数据的特征选择,以下问题还需要进一步探索研究:

(1) 特征选择方法的稳定性问题;

(2) 两类不平衡数据的特征选择问题;

(3) 多类不平衡数据的特征选择问题;

(4) 高维数据的特征选择效率和有效性问题.

[参考文献](References)

[1] Fukunaga K. Introduction of Statistical Pattern Recognition [M]. 2nd ed. Waltham: Academic Press, 1991.

[2] 黄睿,何明一,杨少军.一种适用于小样本问题的基于边界的特征提取算法[J].计算机学报,2007,30(7):1 173-1 178.
Huang Rui, He Mingyi, Yang Shaojun. A margin based feature extraction algorithm for the small sample size problem [J]. Chinese Journal of Computers, 2007,30(7):1 173-1 178. (in Chinese)

[3] He X F, Niyogi P. Locality preserving projections [C] // Vancouver, Whistler, Eds. Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2003.

[4] Cai D, He X H, Han J W. Semi-supervised discriminant analysis [C] // Eleventh IEEE International Conference on Computer Vision. Brazil: Rio de Janeiro, 2007.

[5] Liu H, Motoda H. Feature Selection for Knowledge Discovery and Data Mining [M]. Boston: Kluwer,1998.

[6] 毛勇,周晓波,夏铮,等.特征选择算法研究综述[J].模式识别与人工智能,2007,20(2):211-218.
Mao Yong, Zhou Xiaobo, Xia Zheng, et al. A survey for study of feature selection algorithms [J]. Pattern Recognition & Artificial Intelligence, 2007,20(2):211-218. (in Chinese)

[7] Yu L, Liu H. Feature selection for high-dimensional data: a fast correlation-based filter solution [C] // Proceedings of the 20th International Conferences on Machine Learning. Washington, DC, 2003: 856-863.

[8] Pudil P, Novovicova J, Kittler J. Floating search methods in feature selection [J]. Pattern Recognition Letters, 1994,15:1 119-1 125.

[9] Liu Y, Zheng Y F. FS_SFS: A novel feature selection method for support vector machines [J]. Pattern Recognition, 2006,39:1 333-1 345.

[10] Zhou X, Mao K Z, Wu X Y, et al. Fast gene selection for microarray data using SVM-Based evaluation criterion [C] // IEEE International Conference on Bioinformatics and Biomedicine. IEEE Computer Society, 2008:386-389.

[11] Kira K, Rendell L. A practical approach to feature selection [C] // Proceedings of 9th International Workshop on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 1992:249-256.

[12] Ran G B, Amir N, Naftali T. Margin based feature selection-theory and algorithms [C] // Proceedings of the 21th International Conference on Machine Learning. Canada: Banff, 2004:43-50.

[13] 王练,李云,汪血焰.高维特征集选择模型研究[J].重庆邮电学院学报:自然科学版,2005,17(1):113-116.
Wang Lian, Li Yun, Wang Xueyan. Study on the model of feature selection from huge feature sets [J]. Journal of Chongqing University of Posts and Telecommunications: Nature Science, 2005,17(1):113-116. (in Chinese)

[14] Peng H C, Long F H, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005,27(8):1 226-1 238.

[15] 刘峤,秦志光,陈伟,等.基于零范数特征选择的支持向量机模型[J].自动化学报,2011,37(2): 252-256.
Liu Qiao, Qin Zhiguang, Chen Wei, et al. Zero-norm penalized feature selection support vector machine [J]. Acta Automatica Sinica, 2011,37(2): 252-256. (in Chinese)

[16] 刘峤,王娟,陈伟,等.基于随机复杂度约束的高维特征自动选择算法[J].电子学报,2011,39(2): 370-374.
Liu Qiao, Wang Juan, Chen Wei, et al. An automatic feature selection algorithm for high dimensional data based on the stochastic complexity regularization [J]. Acta Electronica Sinica, 2011,39(2): 370-374. (in Chinese)

[17] Gheyas I A, Smith L S. Feature subset selection in large dimensionality domains [J]. Pattern Recognition, 2010,(43): 5-13.

[18] 张波涛,刘士荣,吕强. 采用生物信息克隆的免疫算法 [J]. 控制理论与应用, 2010,27(6): 799-803.
Zhang Botao, Liu Shirong, Lü Qiang. Immune algorithm with biologic information clone [J]. Control Theory & Applications, 2010,27(6): 799-803. (in Chinese)

[19] 任江涛,黄焕宇,孙婧昊,等. 基于相关性分析及遗传算法的高维数据特征选择 [J]. 计算机应用,2006, 26(6): 1 403-1 405.
Ren Jiangtao, Huang Huanyu, Sun Jinghao, et al. High-dimensional data feature selection based on relevance analysis and GA [J]. Journal of Computer Applications, 2006, 26(6): 1 403-1 405. (in Chinese)

[20] 吴进文,赵晓翠,陈苗苗. 基于遗传算法的高维特征选择的研究 [J]. 郑州轻工业学院学报: 自然科学版, 2010,25(2): 75-78.
Wu Jinwen, Zhao Xiaocui, Chen Miaomiao. Research on high-dimensional feature selection based on genetic algorithms [J]. Journal of Zhengzhou University of Light Industry: Natural Science, 2010,25(2): 75-78. (in Chinese)

[21] 于化龙,顾国昌,刘海波,等. 基于相关性分析的微阵列数据集成分类研究 [J]. 计算机研究与发展, 2010,47(2): 328-335.
Yu Hualong, Gu Guochang, Liu Haibo, et al. Ensemble classification of microarray data based on correlation analysis [J]. Journal of Computer Research and Development, 2010,47(2): 328-335. (in Chinese)

[22] Byeon B, Rasheed K. Selection of classifier and feature selection method for microarray data [C] // 2010 Ninth International Conference on Machine Learning and Applications (ICMLA). Washington, DC, 2010.

[23] Santos J M, Ramos S. Using a clustering similarity measure for feature selection in high dimensional data sets [C] // Proceedings of ISDA'2010. Cairo, 2010.

[24] 王博,贾焰,杨树强,等. 文本多分类中的特征选择研究 [J]. 计算机工程与科学,2010,32(8): 92-93.
Wang Bo, Jia Yan, Yang Shuqiang, et al. Feature selection for multi-class text categorization [J]. Computer Engineering & Science, 2010,32(8): 92-93. (in Chinese)

[25] 尚文倩,黄厚宽,刘玉玲,等. 文本分类中基于基尼指数的特征选择算法研究 [J]. 计算机研究与发展, 2006, 43(10): 1 688-1 694.
Shang Wenqian, Huang Houkuan, Liu Yuling, et al. Research on the algorithm of feature selection based on gini index for text categorization [J]. Journal of Computer Research and Development, 2006, 43(10): 1 688-1 694. (in Chinese)

[26] Zheng Z, Wu X, Srihari R. Feature selection for text categorization on imbalanced data [J]. ACM SIGKDD Explorations, Newsletter, 2004(6): 80-89.

[27] Wasikowski M, Chen X W. Combating the small sample class imbalance problem using feature selection [J]. IEEE Transaction on Knowledge and Data Engineering, 2010,22(10): 1 388-1 400.

[28] Shahib A A, Breitling R, Gilbert D. Feature selection and the class imbalance problem in predicting protein function from sequence [J]. Applied Bioinformatics, 2005(4): 195-203.

[29] Byeon B, Rasheed K. Selection of classifier and feature selection method for microarray data [C] // 2010 Ninth International Conference on Machine Learning and Applications. Washington, DC: IEEE Computer Society, 2010: 534-539.

责任编辑: 严海琳