基于线性回归和属性集成的分类算法

强保华[1] 唐 波[1] 王玉峰[2] 邹显春[3] 柳正利[1] 孙忠旭[1] 谢 武[1]

(桂林电子科技大学广西可信软件重点实验室,广西云计算与大数据协同创新中心 桂林 541004)[1]

(中国电子科技集团公司第54研究所 石家庄 050081)[2] (西南大学计算机与信息科学学院 重庆 400715)[3]

摘 要 对于高维度小样本数据的分类问题,高维属性的复杂性限制了分类模型预测的准确率。为了进一步提高准确率,提出了基于线性回归和属性集成的分类算法。首先,采用线性回归为每一个属性构建属性线性分类器(Attribute Linear Classifier,ALC);其次,为了避免因ALC数量过多而导致准确率下降,利用经验风险最小化策略中的经验损失值作为评估标准来优选ALC;最后,应用多数投票法来集成被筛选的ALC。采用高维度小样本的基因表达数据集进行实验,结果显示该算法具有比逻辑回归、支持向量机和随机森林算法更高的准确率。

关键词 线性回归,单属性分类,经验损失,属性集成,多数投票法

中图法分类号 TP18 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.06.035

Classification Algorithm Using Linear Regression and Attribute Ensemble

QIANG Bao-hua[1] TANG Bo[1] WANG Yu-feng[2] ZOU Xian-chun[3] LIU Zheng-li[1] SUN Zhong-xu[1] XIE Wu[1]

(Guangxi Cooperative Innovation Center of Cloud Computing and Big Data, Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China)[1]

(The 54th Research Institute, China Electronics Technology Group Corporation, Shijiazhuang 050081, China)[2]

(College of Computer and Information Science, Southwest University, Chongqing 400715, China)[3]

Abstract For the classification problems of high-dimensionality and small-sample data, the predictive accuracy of the classification model is restricted by the complexity of the high dimensional attributes. To further improve the accuracy, a classification algorithm using linear regression and attributes ensemble (LRAE) was proposed. The linear regression is utilized to construct an attribute linear classifier (ALC) for each attribute. To avoid the decrease of accuracy caused by too many ALCs, empirical loss value in the empirical risk minimization strategy is used as the evaluation criteria to select ALCs. The majority voting method is adopted to integrate ALCs. The results of experiments using gene expression data demonstrate that the accuracy of LRAE algorithm is relatively higher than that of logistic regression, support vector machine and random forest algorithms.

Keywords Linear regression, Single attribute classification, Empirical loss, Attribute ensemble, Majority voting method

1 引言

信息技术的发展带来了数据的多样性,高维度、少样本数据的预测分类问题给我们带来了挑战。分类(Classification)是数据挖掘和机器学习中应用十分广泛的重要技术之一,其通过分析已知所属类的对象数据集构建分类模型(也称为分类器,Classifier)来预测新数据的所属类。当前分类算法可根据分类的模型分为两类:线性分类(Linear Classification)和非线性分类(Nonlinear Classification)。面对高维数据分类问题,在训练分类模型的时间方面,线性分类算法比非线性分类算法更有优势[1]。由于具有简单、易于分析和实现以及容易扩展为非线性分类方法的特点,线性分类一直是模式分类的研究热点,且在语音识别、图像处理、信息检索和数据挖掘等领域得到了广泛应用[2]。目前经典的线性分类算法有支持向量机(Support Vector Machine,SVM)[3-4]和逻辑回归(Logistic Regression,LR)[5],它们的特点在于分别使用了SMO[6](Sequential Minimal Optimization)和梯度下降法[7](Gradient Descent Method)求解分类模型。这种逐步优化模型的迭代方法的缺陷在于计算资源消耗较大,因此,如何在节省计算资源的情况下建立线性分类模型并获得较高的预测分类准确率

到稿日期:2017-01-17 返修日期:2017-03-27 本文受国家海洋技术公共福利项目(201505002),国家自然科学基金(61462020),广西可信软件重点实验室开放项目(KX201510),广西云计算与大数据协同创新项目(YD16E04),研究生创新项目(YJCXS201538)资助。

强保华(1972-),男,教授,主要研究方向为云计算、大数据,E-mail:qiangbh@guet.edu.cn;唐 波(1990-),男,硕士生,主要研究方向为数据挖掘,E-mail:tangbocqhc@163.com;王玉峰(1965-),男,主要研究方向为服务计算和信息集成;邹显春(1965-),男,副教授,主要研究方向为智能信息处理;柳正利(1989-),男,硕士生,主要研究方向为服务计算;孙忠旭(1987-),男,工程师,主要研究方向为数据分析;谢 武(1979-),副教授,主要研究方向为大数据和数据挖掘,E-mail:xiesixchannels@126.com(通信作者)。

具有很高的研究价值。集成学习(Ensemble Learning)在提高模型的泛化能力上有优势，因此被 Dietterich[8] 在《AI Magazine》上列为机器学习领域的四大研究方向之首。周志华等人[9]于 2002 年首先提出了“选择性集成”的概念，肯定回答了“使用少量的基学习机是否可以达到更好的性能”的疑问，并在国内外集成学习领域引起了强烈反响[10]。目前的集成学习方法可分为两类：1)个体学习机之间存在强依赖关系、必须串行生成的序列化方法，如 Boosting[11]；2)个体学习机之间不存在强依赖关系、可同时生成的并行化方法，如 Bagging[12] 和“随机森林”(Random Forest，RF)[13]。其中 RF 算法因在很多现实任务中展现出强大的性能而被誉为“代表集成学习技术水平的方法”[14]。

本文根据线性分类和集成方法的优势提出了一种新的线性分类算法，即基于线性回归和属性集成的分类算法(Classification Algorithm using Linear Regression and Attribute Ensemble，LRAE)。该算法主要有 3 部分：1)根据属性的区分类的能力，提出为每个属性构建单属性线性分类器(Attribute Linear Classifier，ALC)；2)为了避免过多的 ALC 引起的准确率下降，筛选出以经验损失值为基准的性能较好的 ALC；3)考虑到单个 ALC 分类能力的局限性，使用多数投票法集成筛选出的 ALC 并确定最终分类结果。最后，通过在高维度、少样本的基因表达数据集上的实验来验证算法的准确率和有效性。

2　单属性线性分类器

线性模型通常由多个输入变量组成。对于给定的数据集，数据对象的属性会被作为输入变量，类标记会被作为输出变量。属性以数据对象的方式在一定程度上反映类的特性，因此，类可以由这些属性的特性来进行识别。这说明每个属性都有区分类的能力。为此，本文提出为每个属性建立属性线性分类器 ALC。由于单个 ALC 的分类能力容易受到自身的限制，为了提高预测分类的准确率，集成多个 ALC 就显得很有必要。然而 ALC 之间的分类能力不一致，而且集成大量的 ALC 并不意味着分类预测的准确率会被提高，因此我们需要从大量的 ALC 中筛选出最适合的 ALC 作为集成方法的基分类器。经验风险最小化的经验损失值能在一定程度上反映模型的预测正确率，因而它可以成为评估模型好坏的基准。因此，本文将经验损失值作为评估标准来筛选 ALC。

对于给定的数据集 $D=\{x_i,y_i\}_{i=1}^n$，x_i 为第 i 个样本的属性值，y_i 为其对应的类标记值。属性线性分类器 ALC 的线性模型被定义为 $y'=l(x)$，它表示属性值 x 通过函数 $l(x)$ 映射到预测值 y'。ALC 的线性函数定义为：

$$l(x)=k*x+b \tag{1}$$

其中，k,b 为未知参数。函数 $l(x)$ 作为分类预测模型，当预测值 y' 越接近类的实际标记值 y 时，说明该模型的预测准确率越高，即 y' 和 y 的距离越小，ALC 的分类预测准确率越高。因此，求解 k 和 b 的关键在于如何衡量 y' 和 y 之间的差别。依据经验风险最小化策略[15](Empirical Risk Minimization，ERM)，经验损失函数计算 $l(x)$ 的预测值 y' 与实际值 y 在训练数据集 D 中的平均损失，记为 $L_D(y_i,l(x_i))$。损失函数计算 y' 和 y 的差值，记作 $f(y_i,l(x_i))$。函数 $L_D(y_i,l(x_i))$ 的值称为经验损失(Empirical Loss)值，记作 R_{emp}。

$$R_{emp}=L_D(y_i,l(x_i))=\frac{1}{n}\sum_{i=1}^{n}f(y_i,l(x_i)) \tag{2}$$

经验风险最小化的策略认为，经验风险最小化的模型就是最优的模型。据此，怎样求解最优化模型就变成了怎样寻找 k^* 和 b^* 来使得经验损失值最小化。

$$(k^*,b^*)=\arg\min_{(k,b)} L_D(y_i,l(x_i)) \tag{3}$$

周志华[14]提到，均方误差是回归任务中最常用的性能度量。基于均方误差最小化来进行模型求解的方法称为“最小二乘法”(Least Square Method)。在线性回归中，最小二乘法就是试图找到一条直线，使所有样本到直线上的欧氏距离之和最小。损失函数定义如下：

$$f(y_i,l(x_i))=(y_i-l(x_i))^2 \tag{4}$$

因此，k 和 b 的求解变成了最小化 $L_D(y_i,l(x_i))$。这被称为线性回归模型的最小二乘“参数估计”(Parameter Estimation)。将 $L_D(y_i,l(x_i))$ 分别对 k 和 b 求导，得到如下公式：

$$\frac{\partial L_D(y_i,l(x_i))}{\partial k}=\frac{2}{n}\sum_{i=1}^{n}[kx_i^2+(b-y_i)x_i] \tag{5}$$

$$\frac{\partial L_D(y_i,l(x_i))}{\partial b}=\frac{2}{n}\sum_{i=1}^{n}[kx_i+(b-y_i)] \tag{6}$$

然后，令式(5)和式(6)等于 0，可得到 k 和 b 的最优的闭式(closed-form)解如下：

$$k=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2} \tag{7}$$

$$b=\frac{1}{n}(\sum_{i=1}^{n}y_i-k\sum_{i=1}^{n}x_i) \tag{8}$$

将 k 和 b 代入式(2)可计算出相应的经验损失值 R_{emp}。求解模型参数和经验损失值后，ALC 就被成功创建。

3　LRAE 算法

3.1　LRAE 的结构

给定数据集 $D=\{x_i,y_i\}_{i=1}^n$ 中包含了样本集 $S=\{s_1,s_2,\cdots,s_n\}$，其中 $x_i=\{x_i^{(1)},x_i^{(2)},\cdots,x_i^{(m)}\}$ 表示含有 m 个属性的第 i 个样本；y_i 表示 x_i 对应的类标记值。ALC 是以属性为对象创建的，所以需要将数据集 D 转换为属性数据集 A。属性数据集 A 定义为 $A=\{A_j\}_{j=1}^m$ 且 $A_j=\{x_i^{(j)},y_i\}_{i=1}^n$。$x_i^{(j)}$ 表示为第 i 个样本的属性 j 的值。A_j 表示为第 j 个属性的数据集，它包含了所有样本的属性 j 的值和对应的类标记值。

如图 1 所示，LRAE 类算法的整体流程如下。

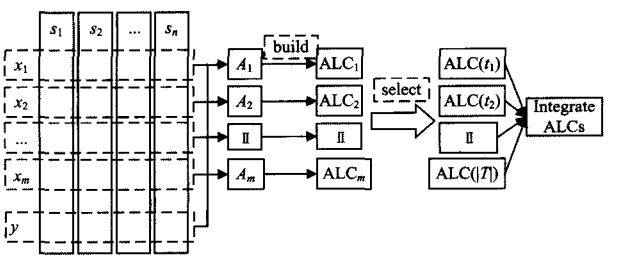

图 1　LRAE 算法的整体结构图

首先，将数据集 D 转换为属性数据集 A；其次，在每个数据集 A_j 中创建一个 ALC；然后，选择拥有最小 R_{emp} 的 ALC 作为基分类器；最后，采用多数投票法集成基分类器的预测值。在图 1 中，T 表示被选中的 ALC 的索引号的集合，$|T|$ 表示集合 T 的基数。

3.2 LRAE 的实现

ALC 由属性标号 j、线性模型参数 k 和 b 以及经验损失值 R_{emp} 构成。模型参数 k 和 b 可以通过式(7)和式(8)直接计算求解。接着，每个 ALC 的经验损失值可通过式(4)计算求解。依据 R_{emp} 和需要被筛选 ALC 的最大数量值 t，LRAE 算法可筛选出 ALC 作为基分类器。算法 1 实现创建和筛选 ALC。

算法 1

输入：数据集 D 和基分类器的最大数量 t
输出：ALC 模型集合 M

```
1. A=transform(D);
2. Initialize M;
3. for j=1,2,…,m
4.     k,b,Remp=calculate(Aj);
5.     model(j)=new Model(j,k,b,Remp);
6.     if (model(j). isNeed())
7.     then M. add(model(j));
8. end for
9. return M
```

对于二分类来说，ALC 的线性模型的预测值 y' 并不直接与类标记值 y_0 或 y_1 相等，因而还需要判断 y' 是代表 y_0 还是 y_1。预测值与某个类标记值的距离越近，说明预测值代表它的可能性越大。因此，将距 y' 近的类标记值作为预测的类标记值。具体实现如下：如果 $|y'-y_0|<|y'-y_1|$，那么 $y'=y_0$，否则 $y'=y_1$。最后，采用多数投票法选择预测最多的类标记作为最终的预测结果。算法 2 实现预测分类的。

算法 2

输入：测试样本 s，ALC 集合 M，类标记值 y_0 和 y_1
输出：类预测标记值 p

```
1. Initialize R;//初始化每个 ALC 的预测值
2. For each M(j) in M
3.     value=s. get(j);
4.     predict=M(j). predict(value);
5.     If |predict-y0|<|predict-y1|
6.       then predict=y0;
7.     else predict=y1;
8.     R. add(predict);
9. End for
10. p=countByMajorityVoting(R);
11. return p;
```

4 实验

4.1 实验数据集

LRAE 算法是针对高维度、少样本数据的二分类问题设计的，所以实验中选择的基因表达数据集中[16]只有两个类。实验数据集下载的网址：http://featureselection. asu. edu/datasets. php[17]。基因表达数据通过 DNA 微阵列杂交实验获得，它们在上述网址的名称分别为“ALLAML”，“GLI_85”，“Prostate_GE”和“SMK_CAN_187”。“ALLAML”(AML)为急性骨髓性白血病和急性淋巴细胞白血病。“GLI_85”(GLI)为神经胶质瘤，它是成年人中最常见的原发性恶性肿瘤。“Prostate_GE”(Prostate)指前列腺癌，是危害男性健康最常见的肿瘤之一，其发病率居美国男性恶性肿瘤之首。“SMK_CAN_187”(SMK)是对有肺癌的吸烟者和没有肺癌的吸烟者进行基因分析的结果。它们的样本和属性的数量如表 1 所列。

表 1 数据集的样本和属性的数量

	AML	GLI	Prostate	SMK
样本数量	72	85	102	187
属性数量	7129	22285	5966	19993

这些数据集有一个共同的特点，即样本数量少(72，85，102 和 187)，属性数量多(7129，22285，5966 和 19993)，它们代表着维度高、样本少的数据集。

4.2 实验方案

本文使用简单交叉验证法(Hold-out Cross Validation)[16]进行实验，该方法将给定的数据集随机分为两部分，一部分作为训练集，另一部分作为测试集。通常他们在整个数据集中所占的百分比为 70%和 30%。由于随机划分数据集 D 存在偶然性，为了使实验结果更具说服力，实验被重复运行 20 次，并取 LRAE 算法分类预测准确率的平均值作为最终实验结果。实验步骤如下：

步骤 1 按照 70%和 30%的比例随机划分数据集 D 为训练集 $D1$ 和测试集 $D2$；

步骤 2 将数据集 $D1$ 作为输入，使用算法 1 构建和筛选 ALC 模型集；

步骤 3 将数据集 $D2$ 作为输入，使用算法 2 完成分类的预测和验证；

步骤 4 对步骤 1－步骤 3 重复执行 20 次，并统计算法的平均准确率。

算法 1 中有两个输入：$D1$ 和配置参数 t。具有不同 t 值的 LRAE 算法可以在 $D2$ 中评估算法的特性并寻找最优参数模型。为了验证 LRAE 算法在面对高维度小样本的二分类问题时的有效性，采用不同的算法进行比较。逻辑回归(LR)和支持向量机(SVM)都使用线性模型来预测分类，且 LRAE 中的 ALC 也有线性模型，所以它们之间具有可比性。此外，随机森林(RF)的基分类器的预测结果是由多数投票法集成的，LRAE 也使用了多数投票法，所以它们之间也具有可比性。在实验中，LR，SVM 和 RF 算法以同样的实验方案与 LRAE 算法进行分类预测准确率的比较。为了统一实验环境，所有的算法都在 spark 平台中实现，LR，SVM 和 RF 算法的代码都来自 spark 中的 MLlib 包。

4.3 实验结果及分析

LRAE 算法分别使用不同数量的 ALC 进行预测分类，实验结果如表 2 所列(其中“ALC_t”表示 ALC 的最大个数为 t)。

表 2　LRAE 的准确率/%

	AML	GLI	Prostate	SMK
ALC_1	87.4901	78.7288	87.8064	59.3477
ALC_3	90.6632	86.3221	90.8972	61.7379
ALC_5	89.2800	84.3015	91.4967	62.8451
ALC_8	90.5609	83.5610	90.1293	63.6213
ALC_10	90.2565	84.0349	90.3427	64.6560
ALC_20	89.3715	85.3191	90.2409	65.3930
ALC_30	88.0113	83.9677	89.3396	66.0831
ALC_40	88.0085	84.1031	88.4017	66.4820
ALC_50	88.2726	84.1479	87.0730	65.6183
ALC_70	87.9467	84.7026	86.4747	65.7647
ALC_100	86.9935	84.3039	85.8395	64.8016

从表 2 中可以看出，随着 ALC 个数的增加，LRAE 算法的准确率出现先上升后下降的现象。对于 ALC 作为基分类器的 LRAE 算法，通过实验结果证明大量的基分类器并不等于预测分类的高准确率，这符合周志华[10]提出的选择性集成的概念。当 $t=1$ 时，被选择的 ALC 的经验损失值最小，对于单个 ALC 来说，它的预测分类准确率应该是最高的。但当 t 的数量逐渐增加时，LRAE 的预测分类准确率在逐渐增加。例如，从表 2 中看出，在 Prostate 数据集中 $t=5$ 时，LRAE 算法的预测分类准确率最高。随着 t 继续增加，LRAE 算法的准确率逐渐下降。我们发现，随着 t 的增加，选择的 ALC 的经验损失值也在增加，这也意味着它的预测分类准确率较低。在第一个 ALC 之后添加的 ALC 的预测分类准确率越来越低，对于 LRAE 算法，先添加的 ALC 起到了辅助的作用，而后添加的 ALC 却成了负担，所以 LRAE 算法的准确率随着 ALC 个数的增加出现先上升后下降的现象。

LR，SVM 和 RF 预测分类的准确率如表 3－表 5 所列。在表 3 中，“LR_5”表示逻辑回归算法迭代 5 次。在表 4 中，“SVM_100”表示支持向量机算法迭代 100 次。在表 5 中，“RF_5”表示随机森林树的棵数为 5。

表 3　LR 算法的准确率/%

	AML	GLI	PROSTATE	SMK
LR_10	86.5061	71.8691	55.1428	50.0015
LR_50	86.0466	86.7604	63.3424	49.8383
LR_300	86.0466	85.4419	84.0584	50.2243
LR_3000	86.0466	85.4419	84.0584	64.3098

表 4　SVM 算法的准确率/%

	AML	GLI	PROSTATE	SMK
SVM_100	77.8393	70.7550	43.5047	47.2537
SVM_1000	82.8143	82.5746	88.2339	52.7107
SVM_10000	85.1070	84.5329	85.7056	61.6332
SVM_20000	85.0175	86.2823	82.2734	70.4007

表 5　RF 算法的准确率/%

	AML	GLI	PROSTATE	SMK
RF_5	90.2182	75.7470	75.9714	62.0647
RF_10	92.6329	76.2887	80.4541	63.5182
RF_15	91.4670	80.2406	84.4014	62.4338
RF_20	93.0794	82.2800	82.1187	64.3277

图 2 分别选择了 LR，SVM，RF 和 LRAE 算法在 4 个数据集中预测分类的最大准确率进行比较。为了使 LRAE 与另外 3 个算法的比较更加直观，创建了表 6。在表 6 中，如果 LRAE 算法比另一个算法在给定数据集中有更高的预测分类准确率，则标注“H”，否则标注“L”。

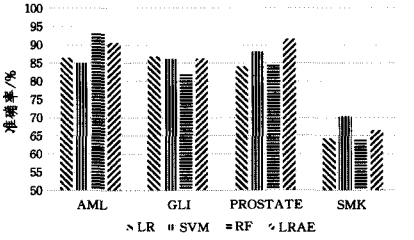

图 2　LR，SVM，RF 和 LRAE 算法的最大准确率

表 6　LRAE 与 LR，SVM 和 RF 的准确率比较结果

	AML	GLI	PROSTATE	SMK
LR	H	L	H	H
SVM	H	H	H	L
RF	L	H	H	H

LR 和 SVM 的分类模型是线性的，而且只有一个由多个属性作为输入变量的线性模型。与 LR 和 SVM 不同，LRAE 有多个线性模型，而且每个线性模型都是由一个属性数据集训练得到的。如表 6 所列，LRAE 有比 LR 和 SVM 更高的准确率，这说明 LRAE 比 LR 和 SVM 更好地利用了属性的多样性。

RF 将决策树作为基分类器，并通过多数投票法集成基分类器的预测结果。RF 的每个基分类器的预测分类都是由多个属性来完成的。与 RF 不同，LRAE 的每个基分类器 ALC 都是由一个属性来完成预测分类的。如表 6 所列，LRAE 有比 RF 更高的分类预测准确率，这说明 LRAE 在利用多数投票法集成基分类器时比 RF 更加高效。

通过分析以上数据集中各算法预测分类的实验结果可知，在处理高维度、小样本数据的二分类问题中，LRAE 具有比 LR，SVM 和 RF 更高的分类预测准确率，这说明了 LRAE 应对分类问题时的有效性和高准确率。

结束语　在应对高维度、小样本数据的二分类问题中，本文提出的基于线性模型与属性集成的分类算法通过实验展现出了相对较高的分类预测准确率。然而，ALC 却难以完整地表示属性的特性，所以还需要对 LRAE 算法进行进一步研究。未来的工作将围绕 ALC 的结构和集成方法展开研究。

参 考 文 献

[1] YUAN G X, HO C H, LIN C J. Recent Advances of Large-Scale Linear Classification [J]. Proceedings of the IEEE, 2012, 100 (9): 2584-2603.

[2] LIU Z W. Research on Linear Classification Algorithm Based on Combination and Optimization [D]. Xi'an: Xidian University, 2013. (in Chinese)
刘志伟. 基于组合优化的线性分类算法研究[D]. 西安：西安电子科技大学，2013.

[3] JOACHIMS T. Training linear SVMs in linear time [C]// Twelfth ACM Sigkdd International Conference on Knowledge Discovery&Data Mining. Philadelphia, USA: ACM press, 2006: 217-226.

（下转第 244 页）

Method Using Bat Algorithm [J]. Neurocomputing, 2015, 177 (c): 612-619.

[11] WANG G G, CHU H C E, MIRJALILI S. Three-dimensional path planning for UCAV using an improved bat algorithm[J]. Aerospace Science & Technology, 2016, 49: 231-238.

[12] WANG J, FAN X, ZHAO A, et al. A Hybrid Bat Algorithm for Process Planning Problem[J]. IFAC-Papersonline, 2015, 48(3): 1708-1713.

[13] LUO J, LIU L, WU X. A double-subpopulation variant of the bat algorithm [J]. Applied Mathematics & Computation, 2015, 263(C): 361-377.

[14] YIN J T, LIU Y L, LIU L, et al. Efficient hybrid bat algorithm [J]. Computer Engineering and Applications, 2014, 50(7): 62-66. (in Chinese)
尹进田，刘云连，刘丽，等. 一种高效的混合蝙蝠算法[J]. 计算机工程与应用，2014，50(7)：62-66.

[15] HE X, DING W J, YANG X S. Bat algorithm based on simulated annealing and Gaussian perturbations[J]. Neural Computing & Applications, 2013, 25(2): 459-468.

[16] WANG X, WANG W, WANG Y. An Adaptive Bat Algorithm [M]// Intelligent Computing Theories and Technology. Springer Berlin Heidelberg, 2013: 216-223.

[17] XIAO H H, DUAN Y M. Research and Application of Improve Bat Algorithm Based on DE Algorithm [J]. Computer Simulation, 2014, 31(1): 272-277. (in Chinese)
肖辉辉，段艳明. 基于 DE 算法改进的蝙蝠算法的研究及应用[J]. 计算机仿真，2014，31(1)：272-277.

[18] LIU C P, YE C M, LIU M C. Optimization strategy from nature: perceive as bat [J]. Application Research of Computers, 2013, 30(5): 1320-1322, 1356. (in Chinese)
刘长平，叶春明，刘满成. 来自大自然的寻优策略：像蝙蝠一样感知[J]. 计算机应用研究，2013，30(5)：1320-1322，1356.

[19] WANG W, WANG Y, WANG X W. An Improved Bat Algorithm with Memory Characteristic [J]. Computer Application and Software, 2014, 31(11): 257-259, 329. (in Chinese)
王文，王勇，王晓伟. 一种具有记忆特征的改进蝙蝠算法[J]. 计算机应用与软件，2014，31(11)：257-259，329.

[20] CHEN Z, YONG Q Z, LU M D. A Simplified-Adaptive Bat Algorithm Based on Frequency [J]. Journal of Computational Information Systems, 2013, 9(16): 6451-6458.

[21] PANT M, THANGARAJ R, ABRAHAM A. Particle swarm optimization using adaptive mutation [C]// Proc of 19th International Workshop on Database and Expert Systems Application. Turin: IEEE, 2008: 519-523.

[22] LI Y, MA L. Bat-inspired Algorithm: A Novel Approach for Global Optimization [J]. Computer Science, 2013, 40(9): 225-229. (in Chinese)
李煜，马良. 新型全局优化蝙蝠算法[J]. 计算机科学，2013，40(9)：225-229.

[23] JORDEHI A R. Chaotic bat swarm optimization (CBSO)[J]. Applied Soft Computing, 2014, 26(c): 523-530.

[24] MENG X B, GAO X Z, LIU Y, et al. A novel bat algorithm with habitat selection and Doppler effect in echoes for optimization [J]. Expert Systems with Applications, 2015, 42(17/18): 6350-6364.

[25] YANG X S. Bat algorithm for multi-objective optimization [J]. International Journal of Bio-Inspired Computation, 2011, 3(5): 267-274.

(上接第 215 页)

[4] HSIEH C J, CHANG K W, LIN C J, et al. A dual coordinate descent method for large-scale linear SVM[C]// International Conference on Machine Learning. Helsinki, Finland: IEEE press, 2008: 1369-1398.

[5] CRAMER J S. The origins of logistic regression: 02-119/4[R]. Uinkeveren: Tinbergen Institute, 2002.

[6] PLATT J. Sequential minimal optimization: A fast algorithm for training support vector marchines [J]. Journal of Information Technology, 1998, 2(5): 1-28.

[7] BOYDSL, VANDENBERGHE. Convex Optimization[M]. Cambridge, UK: Cambridge University Press, 2004.

[8] DIETTERICH T G. Machine learning research: Four current directions [J]. AI Magazine, 1997, 18(4): 97-136.

[9] ZHOU Z H, WU J X, TANG W. Ensembling neural networks: Many could be better than all[J]. Artificial Intelligence, 2002, 13 (1/2): 239-263.

[10] ZHANG C X, ZHANG J S. A Survey of Selective Ensemble Learning Algorithms [J]. Chinese Journal of Computer, 2011, 34 (8): 1399-1410. (in Chinese)
张春霞，张讲社. 选择性集成学习算法综述[J]. 计算机学报，2011，34(8)：1399-1410.

[11] FREUND Y, ROBERT E S. A decision-theoretic generalization of on-line learning and an application to boosting [J]. Journal of Computer and System Sciences, 1997, 55(1): 119-139.

[12] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996, 24(2): 123-140.

[13] BREIMAN L. Random forests [J]. Machine Learning, 2001, 45 (1): 5-32.

[14] ZHOU Z H. Machine Learning[M]. Beijing: Tsinghua University Press, 2016. (in Chinese)
周志华. 机器学习[M]. 北京：清华大学出版社，2016.

[15] LI H. Statistical Learning Method [M]. Beijing: Tsinghua University Press, 2012. (in Chinese)
李航. 统计学习方法[M]. 北京：清华大学出版社，2012.

[16] LI Y, SI J, ZHOU G J, et al. FREL: A Stable Feature Selection Algorithm [J]. IEEE Trains. Neural Netw, 2015, 26(7): 1388-1402.

[17] LU H J, AN C L. Disagreement Measure Based Ensemble of Extreme Learning Machine for Gene Expression Data Classification [J]. Chinese Journal of Computer, 2013, 36(2): 341-348. (in Chinese)
陆慧娟，安春霖. 基于输出不一致测度的极限学习机集成的基因表达数据分类[J]. 计算机学报，2013，36(2)：341-348.