第 36 卷第 1 期 系统工程理论与实践 Vol.36, No.1
2016 年 1 月 Systems Engineering — Theory & Practice Jan., 2016

doi: 10.12011/1000-6788(2016)01-0189-11 中图分类号: TP301.6 文献标志码: A

一种基于 Boosting 的集成学习算法在不均衡数据中的分类

李诒靖 [1,2], 郭海湘 [1,2], 李亚楠 [1,2], 刘 晓 [1,2]

(1. 中国地质大学 经济管理学院, 武汉 430074; 2. 中国地质大学 数字化商务与智能管理研究中心, 武汉 430074)

摘 要 针对多类别不均衡数据的分类问题, 从数据集的特征选择和集成学习两个角度出发, 提出了一种新的针对不均衡数据的分类方法 — BPSO-Adaboost-KNN 算法, 算法采用基于多分类问题的可视化的 AUCarea 作为分类评价指标. 为了测试算法的性能, 本文选取了 10 组 UCI 和 KEEL 选取的测试数据集进行测试, 结果表明本算法在有效提取关键特征后提高了 Adaboost 的稳定性, 在十组数据的分类精度上相比单纯使用 KNN 分类器有 20%~40% 不等的提高. 在本算法和其他 state-of-the-art 集成分类算法对比中, BPSO-Adaboost-KNN 能够取得较优或相当的结果. 最后, 本文将该算法应用到石油储层含油性的识别中, 成功提取了声波、孔隙度和含油饱和度三个关键属性, 在分类精度上相比传统分类算法有了大幅度提高, 在江汉油田五口油井 oilsk81~oilsk85 上的分类精度均达到 98% 以上, 比单纯使用 KNN 的精度高出了 20%, 尤其在最易错分的油层和差油层中有良好的分类效果.

关键词 不均衡数据; 特征提取; 分类; 石油储层

A boosting based ensemble learning algorithm in imbalanced data classification

LI Yijing[1,2], GUO Haixiang[1,2], LI Yanan[1,2], LIU Xiao[1,2]

(1. College of Economics and Management, China University of Geosciences, Wuhan 430074, China; 2. Research Center for Digital Business Management, China University of Geosciences, Wuhan 430074, China)

Abstract This paper focused on multi-class imbalanced data classification, proposed a BPSO-Adaboost-KNN ensemble learning algorithm based on feature selection and ensemble learning. What's more, the algorithm used a visual AUCarea metric to evaluate the performance of classifier when dealing with multi-class classification problems. Then the paper used 10 groups of UCI and KEEL data sets to test the proposed algorithm. The results show that the proposed algorithm improves the stability of the Adaboost after extract the key features, and the classification accuracy for ten groups of data are 20%~40% higher than the KNN classifier. When comparing BPSO-Adaboost-KNN with other three state-of-the-art ensemble algorithms, BPSO-Adaboost-KNN can obtain equal or better results. At last, the proposed algorithm is used in oil-bearing of reservoir recognition, three key attributes are selected (acoustic wave, porosity and oil saturation) successfully. The classification precision reaches more than 98% in oilsk81~oilsk85

收稿日期: 2014-06-26

作者简介: 李诒靖 (1992–), 女, 汉, 江西赣州人, 硕士研究生, 主要从事软计算、机器学习研究, E-mail: liyijing024@gmail.com; 郭海湘 (1978–), 男, 汉, 湖南湘乡人, 教授, 博士, 博士生导师, 主要从事软计算、复杂系统模拟与决策研究, E-mail: faterdumk0732@sina.com.

基金项目: 国家自然科学基金 (71103163, 71103164, 71301153, 71573237); 教育部新世纪优秀人才支持计划 (NCET-13-1012); 中央高校基本科研业务费专项资金资助 (CUG120111, CUG110411, G2012002A, CUG140604); 构造与油气资源教育部重点实验室开放课题 (TPR-2011-11)

Foundation item: National Natural Science Foundation of China (71103163, 71103164, 71301153, 71573237); Program for New Century Excellent Talents in University of Ministry of Education of China (NCET-13-1012); Special Funding for Basic Scientific Research of Chinese Central University (CUG120111, CUG110411, G2012002A, CUG140604); Structure and Oil Resources Key Laboratory Open Project of China (TPR-2011-11)

中文引用格式: 李诒靖, 郭海湘, 李亚楠, 等. 一种基于 Boosting 的集成学习算法在不均衡数据中的分类 [J]. 系统工程理论与实践, 2016, 36(1): 189–199.

英文引用格式: Li Y J, Guo H X, Li Y N, et al. A boosting based ensemble learning algorithm in imbalanced data classification[J]. Systems Engineering — Theory & Practice, 2016, 36(1): 189–199.

Jianghan well logging data, which is 20% higher than KNN classifier. Particularly, the proposed algorithm has significant superiority when distinguishing the oil layer from other oil layers.

Keywords imbalanced data; feature selection; classification; oil reservoir

1 引言

分类问题是机器学习领域的重点研究内容, 目前相关的分类方法已经非常成熟, 如传统的决策树、贝叶斯、人工神经网络、K- 近邻、支持向量机等. 但这些分类方法在处理不均衡数据分类时的表现往往比在类别均衡的数据差, 不能达到理想的分类效果.

所谓不均衡数据是指在数据集中某个或某些类的样本远多于其他类, 而某些类的样本数量相对很少. 通常人们把样本数远多于其他类的类别称为多数类, 样本数相对较少的类称为少数类 [1]. 在不均衡的数据中, 少数类往往是更受关注的对象, 并且少数类的错分代价相对较大. 例如癌症的诊断, 癌症患者属于少数类, 如果癌症病人错诊为健康人, 会耽误病人的最佳治疗时机, 从而对病人造成生命威胁 [2]. 本文所研究的石油储层含油性识别亦是如此, 石油储层的含油性识别概括起来就是综合运用测井数据以揭示含油性的展布情况, 含油性识别包括了储层含油性预测和关键测井属性提取 [3]. 由于测井数据各类别的样本数量差异较大, 且在错分代价上, 油层被错分的代价远大于其它层, 因此, 储层识别是一个典型的多分类不均衡数据的分类问题.

处理不均衡数据的分类问题的策略通常有三种: 第一就是找出新的分类评价标准, 使得分类器不像多数类偏倚. 第二种策略又可分为数据层面的方法和算法层面的方法. 数据层面上主要采用抽样的方法平衡各类别的样本数, 如 Chawla 等人提出的 SMOTE 算法 [4] 和 Dehmeshki 等人提出的基于规则的数据过滤技术 [5] 都是典型的利用抽样来平衡各类别样本数的方法. 算法层面常用的方法有集成学习 (ensemble learning)、代价敏感学习 (cost sensitive learning)、单类分类器 (one-class) 等. 集成学习包括了 1990 年 Schapire 提出的 Boosting 算法 [6] 和 1996 年 Breiman 提出的 Bagging 算法 [7]. 近几年来也提出了许多基于上述算法的改进算法. 集成学习方面人们通常是在 Adaboost 算法和 Bagging 算法的基础上做一些改进 [8−9], 如文献 [10] 提出了一种多分类代敏感 AdaBoost 算法, 文献 [11] 提出的在 Boosting 加入 SMOTE 算法以增加少数类的数量. 代敏感学习方法在应用中更多的作为一种优化策略加入到其他算法中, 如文献 [12] 提出的代价敏感的神经网络算法, 文献 [13] 提出的 CS-SVM 算法定义了代价敏感最优决策超平面 (CS-ODH) 并在 SVM 的最小化目标函数中加入代价敏感因子. 单类分类器只对单一类的样本进行训练并对该类样本定义一个边界, 根据边界求取方法的不同又可分为基于密度、基于神经网络、基于聚类和基于支持向量机的方法 [14]. 最后一个策略是对数据中不相关或是冗余的特征进行约简, 从而达到减少特征个数, 提高模型的分类精确度, 并减少时间复杂度的目的.

以上三个策略立足从不同角度来解决不均衡数据分类问题, 但单一的策略难以满足不同不均衡分布的数据的要求. 因此本文采取了一种融合了以上三种策略的集成学习方法, 提出了一种针对多分类不均衡数据的 BPSO-Adaboost-KNN 集成学习算法. 本算法首先利用 BPSO 对数据集进行了特征选择, 之后采用 Adaboost 集成学习框架, 提出基于 KNN 分类器的 Adaboost 算法, 利用集成学习的思想提高了 KNN 这一传统分类器在不均衡数据中的学习能力, 获得了良好的预测效果. 最后, 本文采用了一种基于多分类问题的 AUC 值 —— AUCarea 作为分类器的评价标准, 更加客观地评价不均衡数据的分类效果的优劣.

2 BPSO-Adaboost-KNN 集成学习算法

2.1 Adaboost-KNN 分类算法

2.1.1 Adaboost 系列算法

Boosting 的基本思想是采用组合学习的方法, 可以将预测精确度很低的弱学习器提升为能达到人们要求的、预测精度高的强学习器 [15]. 现如今人们常用的 Boosting 算法更多的是指 1996 年 Schapire 和 Freund 共同提出的 Adaboost (adaptive boosting) 算法 [16], 它被评为数据挖掘十大算法之一 [17].

基本 AdaBoost 分类算法是针对二分类问题的, 其基本思想是: 首先选定一弱分类器 I 和训练集 $S = \{x_i, y_i\}, i = 1, 2, \cdots, m$, 表示对应样本的类标签, 之后为训练集设定一个分布 D, 即为每个训练样本设定一个权重并初始化为 $1/m$. 然后, 调用弱分类器进行 T 次迭代, 每次迭代都根据训练结果更新样本上的权重, 基本规则是对训练失败的样本赋予较大的权重, 这样下次迭代时分类器将重点学习那些失败的样本. 每次迭代后都会得到一个预测函数 h, 每个预测函数根据其预测精度也赋予一个权重. 经过 T 次迭代后产生一个预

测序列 $h^1, h^2, \cdots, h^T$. 最后, 对预测序列采用带权重的投票法得到最终预测函数 H. 经过 AdaBoost 算法后, 便可以将分类准确率不高的弱分类器转化为一个分类效果好的强分类器.

针对不同的分类问题, Schapire 先后提出了 AdaBoost.M1、AdaBoost.M2、AdaBoost.MR、AdaBoost.MH 等衍生算法 [18]. 其中 AdaBoost.M1、AdaBoost.M2 用于解决多类单标签问题, 而后两者用于解决多类多标签问题. AdaBoost.M1 是 AdaBoost 算法的最简单直接的多类扩展, 但它对于弱学习器的性能要求很高, 并不适用于所有的弱学习器 [19]. 由于本文主要研究多类单标签问题, 且所选取的弱分类器 KNN 是一种相对来说精度比较高的分类器, 应此采取的 Adaboost 系列中的 Adaboost.M1 算法. AdaBoost.M1 算法的具体流程如表 1 所示.

表 1　AdaBoost.M1 算法描述

算法 1: AdaBoost.M1 分类算法
输入: 训练集 $S=\{x^i, y^i\}$, $i=1,2,\cdots,m$, $y_i \in \{Y\}$, $Y=\{c^1, c^2, \cdots, c^k\}$, c^k 为类标签; 迭代次数 T; 弱分类器 I.
1　初始化分布权重: $D(i)=1/m$
2　for $t=1$ to T
3　调用 $I(S, D_t)$ 训练分类器, 得到弱假设 $h_t = X \to \{c_1, c_2, \cdots, c_k\}$
4　计算分类错误率 $\varepsilon_t \leftarrow \sum_{i=1}^{m} D(i)[y_i \neq h_t(x_i)]$
5　if $\varepsilon_t > 0.5$ then
6　$T \leftarrow t-1$
7　continue
8　end if
9　令参数 $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$
10　for $i=1$ to m
11　更新权重 $D_{t+1}(i) = D_t(i)\beta_t^{1-[y_i \neq h_t(x_i)]}$
12　end for i
13　end for t
输出: 最终结果 $H(x) = \arg\max_{y \in Y}(\sum_{t=1}^{T} \ln(\frac{1}{\beta_t})[y_p \neq h_t(s_p)])$

2.1.2　KNN 分类器

KNN (K-nearest neighbor) 是一种非常简单有效的分类方法, 本文在 Adaboost 中选取的弱分类器就是 KNN 分类器. KNN 算法的基本思想可参照文献 [20].

我们选择 KNN 作为基分类器主要基于以下两点考虑:

第一, 算法的计算复杂度: 在本文的算法框架中, 我们采用了 Adaboost 和 BPSO 两种迭代算法, 这意味着本算法的时间复杂度相对较高. 在每次迭代过程中, 算法都将调用一次基分类器进行训练, 因此在选择基分类器时分类器的时间复杂度应当尽量小. KNN 是传统分类算法中时间复杂度较低的算法, 其时间复杂度仅为 $O(N)$. 而大多数 Adaboost 算法倾向于选择决策树或 SVM 等分类效果更好但时间复杂度相对较高的算法, 例如 C4.5 算法的时间复杂度为 $O(N*|D|*\log|D|)$, SVM 更是达到了 $O(N_{sv}^3 + N*N_{sv}^2 + D*N*N_{sv})$ (这里 D 为样本的维度, N_{sv} 为支持向量的个数). 在本算法框架中并不合适.

第二, 适用于多分类问题: 虽然所有二分类算法都可以通过 OVO 或者 OVA[29] 策略扩展到多分类问题, 但这样一来算法的复杂度将大大提高. 因此在选择基分类器中我们将选择可直接用于多分类问题的算法, KNN 正好满足我们的要求.

2.2　分类评价指标

在传统的分类方法中, 常用训练精度作为评价指标. 然而对于不均衡数据分类问题来说, 训练精度并不能真正反映分类器的性能. 例如之前提到的病人识别, 假设正常人的样本占了 99%, 若某分类器将所有的就诊人都归为正常人, 那么这个分类器的分类精度高达 99%. 但是事实上如果没有把病人识别出来, 这样的分类器是没有实用价值的. 在不均衡数据分类评价标准中, ROC (receiver operating characteristic) 曲线是公认的一种全面地分类器标价准则 [21]. ROC 曲线的绘制是基于机器学习方法的基础评价指标 —— 混淆矩阵得来的, 表 2 就是一个二分类问题的混淆矩阵.

表 2　两类混淆矩阵

	预测正类	预测负类
实际正类	TP	FN
实际负类	FP	TN

ROC 曲线取 $\frac{FP}{TN+FP}$ (FPR) 为轴, 取 $\frac{TP}{TP+FN}$ (TPR) 为轴. 对于某个分类器, 通过调整分类器的阈值在上述坐标轴中得到一组 (FPR, TPR), 连接这些点便形成了 ROC 曲线 [22], 这里分类器的阈值取自分类器的

概率输出. 由于 ROC 曲线不能定量评价分类器, 因此人们通常采用 ROC 曲线下方的面积 AUC (area under ROC curve) 作为评价标准, AUC 越大越好. 传统的 ROC 曲线只能应用于二分类问题, 具有很大的局限性.

要将二分类问题的 AUC 值扩展到多分类问题, 可以考虑以下方法: 对于一个包含 $r_m+\delta$ 个类的数据集, 将所有类两两组合分别求其 AUC 值, 最后得到 C_c^2 个 AUC 值, 再进行整合得出最终结果 [23].

先考虑三个类的情况, 假设三个类的类标签分别为 1、2、3, 首先考虑第 1 类和第 2 类, 假设为其设定一个决策阈值为 P_1 (若分类器判定某样本属于第 1 类的概率超过阈值 P_1, 则认为该样本属于第 1 类, 反之属于第 2 类), 一个阈值可以得到一组 TPR 和 FPR 的值, 通过调整 P_1 的大小可以得到一系列 TPR 和 FPR, 根据这些 TPR 和 FPR 就可以画出第 1 类和第 2 类对应的 ROC 曲线, 进而求得 $AUC_{1.2}$, 假设为 0.85; 类似的, 可以求得第 2 类和第 3 类的 $AUC_{2.3}$ (假设为 0.8) 以及第 3 类和第 1 类的 $AUC_{1.3}$ 值 (假设为 0.75).

在求得 C_c^2 个 AUC 值之后, 通常的做法是将所有 AUC 值求均值作为最终的结果, 但这种方法有一个缺陷, 就是对于不同的分类器, 可能两两求得的 AUC 各不相同, 但最后求均值得到的结果却是相同的, 从而不能很好地比较分类器的好坏. 例如设 q 个 AUC 值分别为 $r_1, r_2, \cdots, r_q$, 当某两个 AUC 值同时变为 $r_n-\delta$ 和 $r_m+\delta$, 此时对所有 AUC 求平均值可以得到 $AUC_{\text{avg}} = r_1+r_2+\cdots+(r_m+\delta)+(r_n-\delta)+\cdots+r_q = \sum_{i=1}^{q} r_i$, 与变化前得到的值相同. 本文采用的是一种将所有 AUC 值放在极坐标 (polar) 中, 以一种可视化的方式呈现出来, 并以所有 AUC 值所围成图形的面积作为定量的指标, 面积越大说明分类效果越好. 这里我们统一取所有 AUC 所能围成的面积的最大值, 设该面积为 AUCarea, 如图 1 所示.

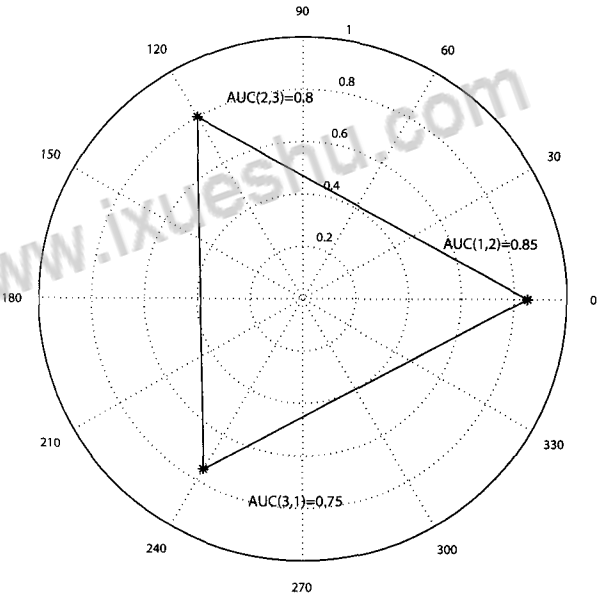

图 1 三类问题的 AUC polar 图 ($AUC_{1.2}=0.85$, $AUC_{2.3}=0.8$, $AUC_{1.3}=0.75$)

根据面积的计算公式, $AUCarea$ 的计算公式如公式 (1) 所示:

$$AUCarea = \frac{1}{2}\sin\left(\frac{2\pi}{q}\right)\left(\left(\sum_{i=1}^{q-1} r_i \times r_{i+1}\right) + (r_q \times r_1)\right) \tag{1}$$

其中 $q = C_c^2$ 为 AUC 总数, c 为类别数.

考虑到大多数评价标准的取值在 [0, 1] 内, 我们用 $AUCarea/Maximum\ AUCarea$ 对 $AUCarea$ 进行了归一化, 这里的 $Maximum\ AUCarea$ 是通过对 (1) 中所有 r_i 取 1 得到的. 计算归一化的 $AUCarea$ 的公式如 (2) 所示. 在本文中将采用归一化的 $AUCarea$:

$$AUCarea = \frac{\frac{1}{2}\sin(\frac{2\pi}{q})((\sum_{i=1}^{q-1} r_i \times r_{i+1}) + (r_q \times r_1))}{\frac{1}{2}\sin(\frac{2\pi}{q}) \times q} = \frac{(\sum_{i=1}^{q-1} r_i \times r_{i+1}) + (r_q \times r_1)}{q} \tag{2}$$

前文提到绘制 ROC 曲线时需要分类器给出一个输出概率, 这个概率表示分类器预测一个样本对于某一类的隶属程度. 因此在本算法中也需要为 Adaboost 和 KNN 分别设定一个概率输出矩阵. 假设一个包含 S 个特征、m 个样本、C 个类别的数据集 $S=\{s_1, s_2, \cdots, s_m\}$ 类标签为 $C_1, C_2, \cdots, C_m$, 为 Adaboost 和 KNN 分别建立一个 $m*C$ 维的概率输出矩阵 $Ada\_score_{m\times C}$、$Knn\_score_{m\times C}$. 其中 $Knn\_score(i,j)$、$Ada\_score(i,j)$ 分别表示在 KNN 和 Adaboost 在分类时所判定的样本 s_j 属于第 j 类的概率. 对于样本 s_j, 统计其 k 个近

邻中属于每一个类的近邻数 $n_1, n_2, \cdots, n_c$, $\sum_{j=1}^{c} n_j = k$, 样本 s_j 属于第 s_j 类的概率 $Knn\_score(i,j) = n_j/k$. 假设 Adaboost 算法的迭代次数为 T, 在 Adaboost 算法的每次迭代中计算 KNN 的分类错误率 ε_t, 则 $Knn\_score(i,j)$ 相对应的 $Ada\_score(i,j)$ 就取 T 次得到的带权重的 $Knn\_score(i,j)$, 即 $Ada\_score(i,j) = \sum_{t=1}^{T}(\frac{1-\varepsilon_t}{\sum_{t=1}^{T}1-\varepsilon_t} \times Knn\_score_t(i,j))$, 其中 $\frac{1-\varepsilon_t}{\sum_{t=1}^{T}1-\varepsilon_t}$ 代表了第 t 迭代采用的 KNN 分类器的权重. 计算完分类器在所有类别上的概率输出后, 就可以求出任意两个类的在 Adaboost 和 KNN 分类器中的 ROC 曲线和 AUC 值, 在求出 C_c^2 个 AUC 值后, 根据公式 (1) 即求得面积.

2.3 基于 BPSO 的特征选择

利用智能算法来进行特征选择是一种基于随机搜索策略的特征提取办法. 粒子群优化算法 (particle swarm optimization, PSO) 是 1995 年由 Kennedy 和 Eberhart 通过观察鸟类捕食行为而提出的一种基于群体的智能演化算法 [24]. 标准 PSO 算法只能用于连续实数空间中的问题, 但是在现实生活中很多问题是建立在离散二进制空间上的, 基于此, 1997 年, Kennedy 和 Eberhart 又进一步提出了二进制粒子群优化算法 (binary particle swarm optimization, BPSO)[25].

基于 BPSO 的特征选择算法沿用了 [25] 的基本算法流程, 只在粒子表示和适应度值选择上进行了设计. 本文将二进制粒子编码为 D 维 0-1 向量, 粒子的位置向量的每个字符位都代表一个特征, 如字符位的值为 1 表示该特征被选择, 为 0 表示不被选择. 粒子的适应度值通常为粒子所表示的样本子集的分类精度, 但本算法选取的是 $AUCarea$ 作为粒子的适应度值.

在算法中加入特征选择是基于了以下考虑: 特征选择是一种降低样本空间噪声的有效方法 [25]. 在不均衡数据中, 因为少数类样本的不足, 分类器常常将少数类样本作视作噪声数据 [10,16]. 通过特征选择剔除特征集中不相关和冗余的特征, 降低样本空间的噪声 [26], 可以降低少数类样本被视为噪声数据的风险, 从而提高少数类的训练效果.

2.4 BPSO-Adaboost-KNN 集成学习算法

BPSO-Adaboost-KNN 集成学习算法首先采用 BPSO 算法进行特征提取, 然后将粒子所选择的样本子集代入 Adaboost-KNN 中进行分类, 得到分类结果后计算 $AUCarea$ 值作为适应度值. BPSO-Adaboost-KNN 集成学习算法的具体流程表 3 所示.

表 3 BPSO-AdaBoost-KNN 集成学习算法

算法 3: BPSO-AdaBoost-KNN 集成学习算法	
1	begin
2	随机初始化粒子
3	while(未达到最大迭代次数或未达到停止准则)
4	for $i = 1$ to N(种群数)
5	根据粒子 i 提取的特征子集抽取样本子集 $data_i$
6	初始化分布权重: $D_1(i) \leftarrow 1/m$ // 开始 Adaboost.M1
7	for $t = 1$ to T
8	调用 KNN 分类器, 得到弱假设 $h_t : data_i \rightarrow \{c_1, c_2, \cdots, c_k\}$ 和概率输出矩阵 Knn_score_t
9	计算分类错误率 $\varepsilon_t \leftarrow \sum_{i=1}^{N} D(i)[y_i \neq h_t(s_i)]$
10	令参数 $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$
11	更新权重 $D_{t+1}(i) = D_t(i)\beta_t^{1-[y_i \neq h_t(x_i)]}$
12	end for t
13	计算分类序列 $H(x) = \arg\max_{y \in Y}(\sum_{t=1}^{T} \ln(\frac{1}{\beta_t})[y_p \neq h_t(s_p)])$
14	计算 Adaboost 的概率输出矩阵 $Ada\_score = \sum_{t=1}^{T}(\frac{1-\varepsilon_t}{\sum_{t=1}^{T}1-\varepsilon_t} Knn\_score_t)$
15	计算所有 AUC 值, 绘制 polar 图并计算面积 $AUCarea_i$, $fitness(x_i) = AUCarea_i$.
16	if $fitness(x_i) > fitness(pbest_i)$ //更新粒子群体和个体最优解
17	then $pbest_i = x_i$
18	if $fitness(x_i) > fitness(gbest)$
19	then $gbest = x_i$
20	更新粒子 i 的速度和位移
21	end for i
22	end while
输出: 最优特征子集 S_{best} 和分类结果, 对应的最优 polar 图、每个 AUC 值和 $AUCarea_{\text{best}}$	

3 仿真实验及性能分析

3.1 实验的设计与数据

本文选用了 10 组不同的数据集进行实验, 10 组数据全部来自于 UCI 数据库和 KEEL 数据库. 各个数据集的特征信息如表 4 所示. 本文设计了两组实验来检验 BPSO-Adaboost-KNN 的性能. 在 3.2 中我们分别验证了 BPSO 和 Adaboost 的有效性. 3.3 中我们将 BPSO-Adaboost-KNN 与另外三种集成学习算法进行比较.

表 4 测试数据集详细特征信息

数据集	样本数	样本分布	IR	类别数	特征数	选择特征数
Balance	625	(49/288/288)	1:6	3	4	3
Glass	214	(70/76/17/13/9/29)	1:8	6	9	6
Landsat	2000	(461/224/397/211/237/410)	1:2	6	36	24
New_thyroid	215	(150/35/30)	1:5	3	4	2
Zoo	101	(41/20/5/13/4/8/9)	1:10	7	16	9
Ecoli	358	(143/77/2/2/35/20/5/42)	1:72	8	7	5
Page-Blocks	5472	(4913/329/87/115/28)	1:175	5	10	7
Wine-Quality-White	4898	(2198/1457/880/175/163)	1:14	5	11	5
Autos	159	(2/14/33/32/20/9)	1:16	6	25	8
Shuttle	2175	(1194/1/4/236/86)	1:1194	7	9	6

实验中我们采用十字交叉验证 (10-fold cross-validation) 的测试方法. 该方法将数据集等分为 10 份, 轮流将其中 9 份作为训练集, 1 份作为测试集进行实验, 将 10 次实验得出的指标值的均值作为对本算法最终评价结果. 实验参数方面, Adaboost 的迭代次数 T、种群个数 N 和 BPSO 最大迭代次数 $Max\_iteration$ 都设为 50, BPSO 中的相关参数都选择标准 PSO 的参数设置, 即惯性权重设为 0.729, 学习因子 $c1$, $c2$ 均设为 1.49445. KNN 中 k 设为 5.

3.2 BPSO 和 Adaboost 的有效性验证

为了同时验证本文提出的 BPSO-Adaboost-KNN 算法中 Adaboost 的性能和特征提取的有效性, 实验分别测试了不经过特征提取单独使用 KNN 算法、不经过特征提取使用 Adaboost-KNN 算法、采用特征提取的 BPSO-KNN 算法、采用特征提取的 BPSO-Adaboost-KNN 算法四种算法所得出的 AUCarea 和分类精度. 在表 4 的最后一栏里我们列出了 BPSO 所提取的特征数情况, 表 5 为以上四种算法的 AUCarea 和分类精度的对比情况.

表 5 KNN, Adaboost-KNN, BPSO-KNN 和 BPSO-Adaboost-KNN 的对比情况

Dataset	KNN		Adaboost-KNN		BPSO-KNN		BPSO-Adaboost-KNN	
	Accuracy	AUCarea	Accuracy	AUCarea	Accuracy	AUCarea	Accuracy	AUCarea
Balance	0.5873	0.5216	0.7421	0.5838	0.7698	0.6332	0.9563	0.7875
Glass	0.6477	0.7896	0.9773	0.8345	0.6705	0.7155	0.9886	0.8721
Landsat	0.7519	0.7016	0.9239	0.8412	0.8030	0.7726	0.9551	0.8930
New_thyroid	0.9186	0.7860	0.9884	0.9172	0.9186	0.8768	1.0000	1.0000
Zoo	0.7907	0.8362	0.9302	0.9822	0.8140	0.9431	1.0000	1.0000
Ecoli	0.8222	0.7398	0.9111	0.8858	0.8296	0.8137	0.9556	0.8996
Page-Blocks	0.9316	0.8057	0.9936	0.9364	0.9462	0.9066	0.9954	0.9465
Wine-Quality-White	0.5250	0.6520	0.8909	0.7854	0.5712	0.6893	0.8962	0.8454
Autos	0.5694	0.6046	0.8388	0.8096	0.6714	0.7443	0.9796	0.9270
Shuttle	0.9193	0.9161	0.9285	0.9244	0.9754	0.9691	0.9985	0.9890

从表 5 中可以看出, 采用两种评价指标所得的结果是一致的, 即正确率越高, AUCarea 越大. 从结果上看, Adaboost-KNN 的分类效果要远好于 KNN, 说明经过 Adaboost 集成学习后, 确实把 KNN 这一在不均衡数据中分类效果较差的弱分类器转化为了强分类器, 在分类精度上有了很大的提高. 其中 BPSO-Adaboost-KNN 算法对 10 个数据集的分类效果最好, 特别是对数据集 New_thyroid 和数据集 Zoo 进行分类时正确率达到了 100%. 另外经过特征提取的两类算法和不经过特征提取的两类算法的比较中, 显然经过特征提取的算法分类效果较好, 说明 BPSO 算法有效地提取出了关键属性, 剔除了不相关或冗余的特征, 达到了减少特

征个数, 提高模型精确度, 并减少运行时间的目的.

为了更直观地看出四种算法在这两种评价指标下的分类效果, 图 2 给出了四种算法的 AUC 值的 polar 图 (因为空间有限这里只列出了其中 5 个数据集的情况). 可以明显看出 BPSO-Adaboost-KNN 在 polar 图中面积都是最大的.

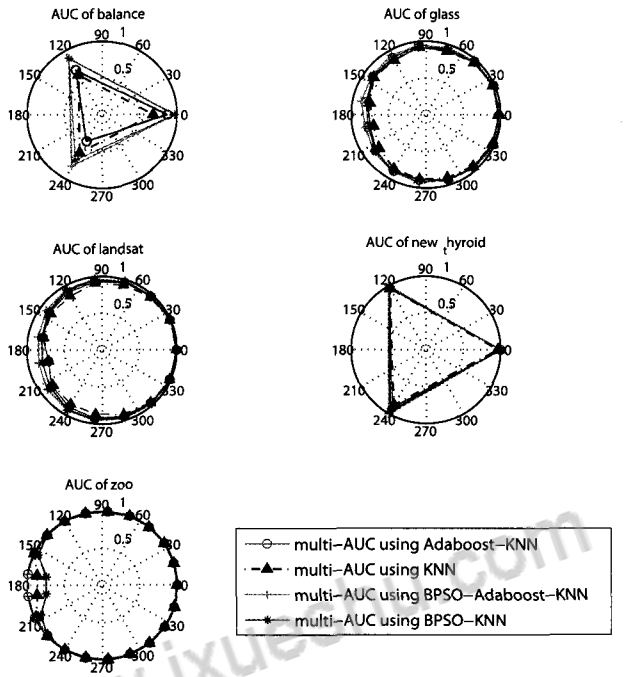

图 2　5 个数据集的 AUC polar 图对比

3.3 BPSO-Adaboost-KNN 与其他集成算法的对比

为了进一步测试 BPSO-Adaboost-KNN 的性能, 本文将其与 SMOTEBoost[12], CS-MCS[27], PUSBE[28] 三种集成分类算法进行了对比. 三种集成分类算法的集成策略可见表 6, 实验结果如表 7 所示其中各个数据集最好的结果已经加粗.

表 6　三种集成分类算法的集成策略描述

算法	策略
SMOTEBoost	过采样算法 SMOTE + Adaboost + Ripper 基分类器
PUSBE	Filter 特征选择算法 + boosting + SVM 基分类器 + 基于多样性的基分类器淘汰策略 基于 BP 神经网络的分类器融合策略
CS-MCS	欠采样算法 RUS + 代价敏感决策树 + 基于 GA 算法的基分类器融合策略

表 7　BPSO-Adaboost-KNN 与其他三种集成学习算法的对比情况

Dataset	SMOTEBoost		CS-MCS		PUSBE		BPSO-Adaboost-KNN	
	Accuracy	AUCarea	Accuracy	AUCarea	Accuracy	AUCarea	Accuracy	AUCarea
Balance	0.8803	0.6792	0.7869	0.6512	0.7738	0.6291	**0.9563**	**0.7875**
Glass	0.8915	0.7589	0.9255	0.8078	0.9432	0.8580	**0.9886**	**0.8721**
Landsat	0.9426	0.8824	0.7919	0.7581	0.9551	0.7788	**0.9551**	**0.8930**
New_thyroid	**1.0000**	**1.0000**	0.9863	0.9226	**1.0000**	**1.0000**	**1.0000**	**1.0000**
Zoo	**1.0000**	**1.0000**	0.7442	0.8310	**1.0000**	**1.0000**	**1.0000**	**1.0000**
Ecoli	0.9381	0.7846	**0.9825**	**0.9355**	0.8667	0.7514	0.9556	0.8996
Page-Blocks	0.9072	0.8287	0.8974	0.7900	**0.9954**	**0.9752**	0.8962	0.8454
Wine-Quality-White	0.8125	0.7040	0.7627	0.7000	**0.9818**	0.9200	0.9796	**0.9270**
Autos	0.8742	0.7970	0.8365	0.7627	**0.9937**	**0.9818**	0.9796	0.9270
Shuttle	**0.9991**	**0.9960**	0.9862	0.9168	0.9816	0.9142	0.9985	0.9890

从表 7 中可以看出, 四种算法的表现是相当的, 它们分别在不同的数据集中取得了优于其他三种算法的结果. 总体来看, PUSBE 和 BPSO-Adaboost-KNN 的表现略比其他两种算法好. 但是, 相比较 PUSBE,

BPSO-Adaboost-KNN 在每个数据集中的表现与其他三种算法的差距很小, 而 PUSBE 在如 Balance, Ecoli 等数据集的表现稍有欠缺. 另外, PUSBE 和 BPSO-Adaboost-KNN 都采取了特征选择技术, 这也说明了特征选择在处理不均衡数据分类问题上的有效性.

4 BPSO-Adaboost-KNN 算法在石油储层含油性识别中的应用

4.1 数据准备和预处理

石油储层含油性识别的测井属性包括: 声波 (AC)、中子 (CNL)、深测向电阻率 (RT)、孔隙度 (POR)、含油饱和度 (So)、渗透率 (PERM), 共六个属性. 测井的识别结论解释有油层, 差油层, 水层和差油层四种. 本实验数据来自于江汉油田某区块的五口井 Oilsk81、Oilsk82、Oilsk83、Oilsk84、Oilsk85 的测井数据. 本实验选用 Oilsk81 井的数据为训练集, 首先以 Oilsk82~Oilsk85 单独作为测试集实验四次, 然后将四个测试集合并为一个合集再进行一次实验. 表 8 为 Oilsk81 井的部分测井属性和对应的测井解释结论 (由于版面有限, 其余井的数据略), 表 9 为 5 个测井数据的样本分布情况.

表 8 Oilsk81 井测井解释成果表

层号	声波/(μs/m)	中子/%	深测向电阻率/(Ω.m)	孔隙度/%	含油饱和度/%	渗透率/(mμm^2)	结论
1	195	7.5	13.0	6.0	0	0	干层
2	225	10.0	7.3	11.0	0	0	水层
3	230	14.0	5.5	12.0	0	0	水层
4	220	9.0	25.0	9.0	56	1.3	油层
5	225	8.0	30.0	9.0	58	2.3	油层
6	210	7.0	26.0	6.0	0	0	干层
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	201	6.0	16.0	7.0	40	0.4	差油层
31	213	9.5	12.0	9.0	61	2	油层

表 9 各井样本分布情况

井号	各类样本分布情况 (干层, 水层, 油层, 差油层)
Oilsk81	(14, 2, 12, 3)
Oilsk82	(28, 2, 7, 11)
Oilsk83	(28, 3, 12, 7)
Oilsk84	(32, 6, 5, 9)
Oilsk85	(44, 4, 6, 11)

表 10 五次实验最优特征子集

测试集	最优特征子集
Oilsk82	(1, 0, 0, 0, 1, 0)
Oilsk83	(0, 0, 0, 1, 0, 1)
Oilsk84	(1, 0, 0, 1, 1, 0)
Oilsk85	(1, 0, 0, 0, 1, 0)
Oilsk82~Oilsk85 合集	(1, 0, 0, 1, 1, 0)

注: 特征集从左到右依次为: (声波、中子、深测向电阻率、孔隙度、含油饱和度、渗透率)

注意到测井数据集中声波属性的数据与其他属性之间有数量级的差异, 也即该数据集存在奇异样本数据. 为了后面数据处理方便, 保证程序运行时收敛加快, 因此在实验前先对数据进行归一化处理, 将所有数据限制到 [0, 1] 范围内. 实验参数方面的选取与仿真实验参数一致, 实验对比仍采用 3.2 的对比方式.

4.2 实验结果及性能分析

实验结果中, 五次实验 BPSO 选择的最优特征子集如表 10 所示, 其中 1 代表被选择, 0 代表未被选择. 从表中可以看出, 每个测井所得的最优特征子集不完全相同, 这体现了算法的自适应性. 此外, 声波、孔隙度、含油饱和度、渗透率均被不同的测试集选择过, 其中, 声波、孔隙度、含油饱和度在四个单独井测试集里被选择的次数较多, 且在合集中也被选择为最优特征子集, 因此可以判定, 声波、孔隙度、含油饱和度为测井数据中的参考关键属性.

表 11 含油性识别中四种算法性能对比

Dataset	KNN		Adaboost-KNN		BPSO-KNN		BPSO-Adaboost-KNN	
	Accuracy	AUCarea	Accuracy	AUCarea	Accuracy	AUCarea	Accuracy	AUCarea
Oilsk82	0.5786	0.7708	0.8996	0.9117	0.9082	0.8750	0.9743	0.9883
Oilsk83	0.8639	0.9400	0.8791	0.9628	0.9082	0.9600	1.0000	0.9976
Oilsk84	0.8822	0.7500	0.9280	0.9615	0.9030	0.9231	0.9853	0.9938
Oilsk85	0.6629	0.7846	0.7311	0.9252	0.7779	0.9231	0.9408	0.9882
Oilsk82~85	0.7844	0.8093	0.8005	0.9377	0.8944	0.8977	0.9470	0.9839

表 11 统计了五次实验中四种算法所得到的分类正确率和 AUCarea 值, 表中值均为 50 次重复实验的平

均值, 图 3 为 50 次实验的错误率统计图, 图 4 为 5 个测井数据的 AUC polar 图. 从表 11 中可看出五次实验中, 四种算法正确率从大到小排名依次为 BPSO-Adaboost-KNN、Adaboost-KNN、BPSO-KNN、KNN, 说明经过特征提取后的两个分类算法要优于未经过特征提取的两个分类算法, 且 Adaboost-KNN 算法的分类错误率均明显低于 KNN 算法. 另外从图 3 中可以发现, Adabost-KNN 是一种不稳定的分类器, 每次试验的分类结果可能不同, 但是在经过特征提取后, BPSO-Adaboost-KNN 分类的稳定性要好于不经过特征提取时的 Adaboost-KNN 算法, 也说明特征提取剔除了一些噪声数据, 从而增强了分类器的稳定性.

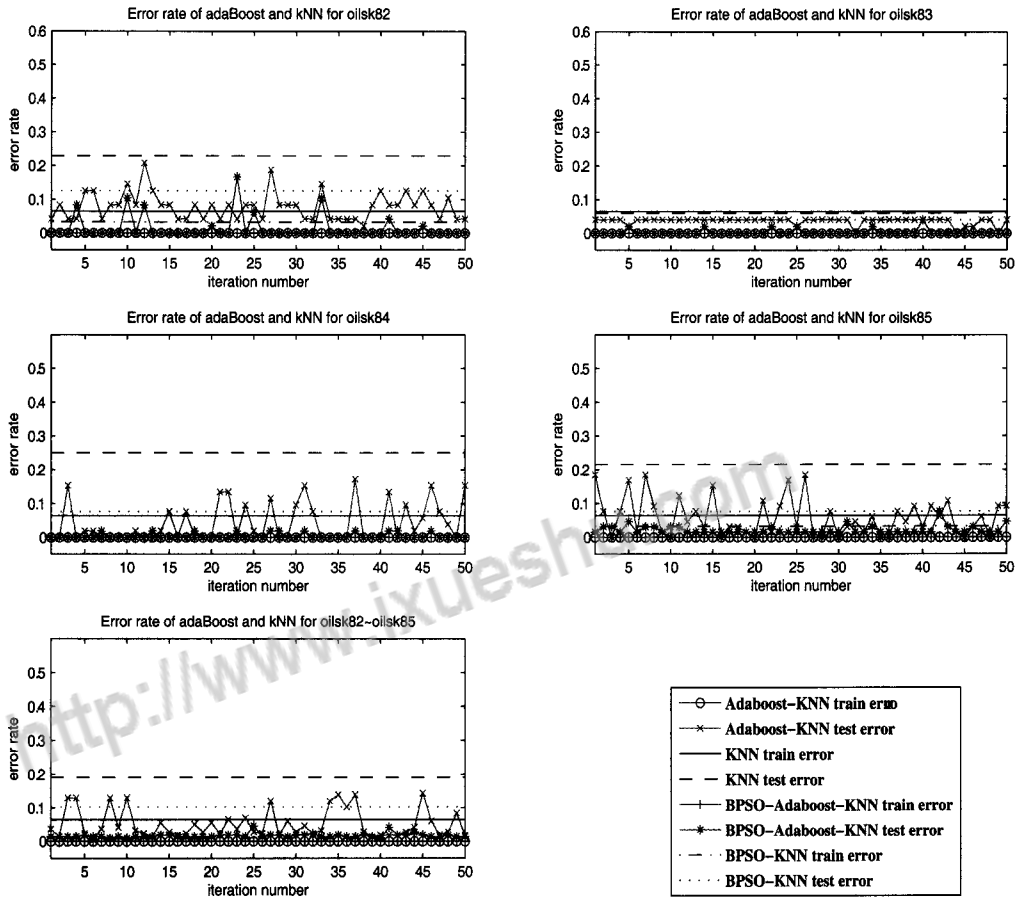

图 3　50 次试验错误率对比

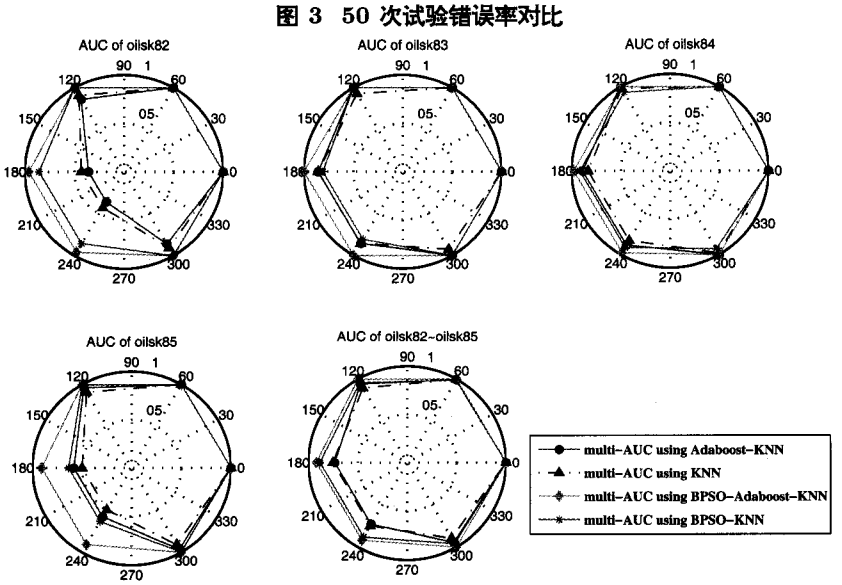

图 4　5 个测井数据的 AUC polar 图对比

另外, 采用 AUC 作为评价标准还能很清晰地看出数据集中任意两类的差异性. 若两个类的 AUC 值为

1, 说明分类器在对这两类的分类中没有产生相互错分的情况, 一定程度上说明了这两个类的差异性大. AUC 值在 [0.5, 1] 之间, 说明分类器对这两类产生了相互错分的情况, 越接近 1, 错分样本数越少. 当 AUC 值低于 0.5 时, 说明分类器在这两类的分类中采取的是随机猜测的方法, 也就是说分类器对于这两类的分类是失效的. Oilsk82 和 Oilsk85 井在采用未经特征提取的 KNN 算法时, 在某些类中的 AUC 值低于 0.5, 说明 KNN 对于这两类的分类是失效的. 表 12 分别为 BPSO-Adaboost-KNN 和 KNN 算法中任意两类的 AUC 值.

表 12 含油性识别中任意两类 AUC 值对比

Oilsk82		Oilsk83		Oilsk84		Oilsk85		Oilsk82~85	
Our algorithm	KNN algorithm	Our algorithm	KNN algorithm	Our algorithm	KNN algorithm	Our algorithm	KNN algorithm	Our algorithm	KNN algorithm
1	1	1	1	1	1	1	1	1	0.9995
1	1	1	1	1	1	1	1	1	1
1	0.8636	1	1	1	1	1	1	1	0.9596
1	0.8636	1	1	1	0.9167	1	0.9090	1	0.8965
1	0.7987	1	1	1	0.9815	0.9090	0.9090	0.9793	0.8491
0.9610	0.3571	1	0.8571	0.9778	0.8333	0.9090	0.5000	0.9189	0.7522

表 12 中最易错分的两个类为油层和差油层类, 说明这两类的差异性最小, 也最难区分. 另外水层与其它层也有一定错分的情况, 这主要是由于测井数据中, 水层样本的数量最少, 学习起来最困难. 而对比 BPSO-Adaboost-KNN 和 KNN 可以看出, BPSO-Adaboost-KNN 任意两类的 AUC 值均大于等于 KNN 算法, 且在对水层和油层、水层和差油层以及油层和差油层的区分中, 有了大幅度的提高. 这也可以说明, 通过 Adaboost 集成学习, 分类器对少数类样本有了更强的学习能力.

5 结论

本文提出了一种基于集成学习的 BPSO-Adaboost-KNN 分类算法, 本算法首先利用 BPSO 对数据集进行特征提取, 选取最优特征子集后采用 Adaboost-KNN 算法进行分类. 在分类评价标准上, 本文将传统二分类评价指标 AUC 扩展到了多分类问题中, 提出了一种基于面积的 AUCarea 评价指标, 并将其作为 BPSO 的适应度值和最终分类的结果呈现. 在对算法进行详细描述后, 选取了 10 组 UCI 测试数据集对本算法进行了测试, 测试结果证实了本算法的优越性. 最后将 BPSO-Adaboost-KNN 算法运用到石油储层含油性识别中, 对测井数据进行了分类和关键属性提取, 最后利用 AUC 值分析了不同储层之间的相似程度, 其中相似度最高的为油层和差油层, 利用 BPSO-Adaboost-KNN 算法能很好地对这两类进行区分而不易将油层错分, 对石油储层识别具有很大的现实意义.

参考文献

[1] Searle S R. Linear models for unbalanced data[M]. Wiley, 1987.

[2] 王和勇, 樊泓坤, 姚正安, 等. 不平衡数据集的分类方法研究 [J]. 计算机应用研究, 2008, 25(5): 1301–1308.
Wang H Y, Fan H K, Yao Z A, et al. Research of imbalanced data classification[J]. Application Research of Computers, 2008, 25(5): 1301–1308.

[3] 涂承胜, 刁力力, 鲁明羽, 等. Boosting 家族 AdaBoost 系列代表算法 [J]. 计算机科学, 2003, 30(3): 30–34.
Tu C S, Diao L L, Lu M Y, et al. The typical algorithm of AdaBoost series in Boosting family[J]. Computer Science, 2003, 30(3): 30–34.

[4] Chawla N V, Bowyer K W, Hall L O. SMOTE: Synthetic minority over-sampling technique[J]. Artificial Intelligence Research, 2002, 16(3): 321–357.

[5] Lee C Y, Lee Z J. A novel algorithm applied to classify unbalanced data[J]. Applied Soft Computing, 2012(12): 2481–2485.

[6] Schapire R E. The strength of weak learnability[J]. Machine Learning, 1990, 5(2): 197–227.

[7] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123–140.

[8] 杨明, 尹军梅, 吉根林. 不平衡数据分类方法综述 [J]. 南京师范大学学报 (工程技术版), 2008, 8(4): 7–12.
Yang M, Yi J M, Ji G L. Classification methods on imbalanced data: A survey[J]. Journal of Nanjing Normal University (Engineering and Technology Edition), 2008, 8(4): 7–12.

[9] Galar M, Fernández A, Barrenechea E. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 2012,

42(4): 463–484.

[10] 林智勇, 郝志峰, 杨晓伟. 不平衡数据分类的研究现状 [J]. 计算机应用研究, 2008, 25(2): 332–336.
Lin Z Y, Hao Z F, Yang X W. Current state of research on imbalanced data sets classification learning[J]. Application Research of Computers, 2008, 25(2): 332–336.

[11] 付忠良. 多分类问题代价敏感 AdaBoost 算法 [J]. 自动化学报, 2011, 37(8): 973–983.
Fu Z L. Cost-sensitive AdaBoost algorithm for multi-class classification problems[J]. Acta Automatica Sinica, 2011, 37(8): 973–983.

[12] Nitesh V C, Aleksandar L, Lawrence O H, et al. SMOTEBoost: Improving prediction of the minority class in boosting[C]// 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, 2003, 107–119.

[13] Zhou Z H, Liu X Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem[J]. IEEE Transactions on Knowledge and Data Engineering, 2006: 63–77.

[14] 郑恩辉, 李平, 宋执环. 代价敏感支持向量机 [J]. 控制与决策, 2006, 21(4): 473–476.
Zheng E H, Li P, Song Z H. Cost sensitive support vector machines[J]. Control and Decision, 2006, 21(4): 473–476.

[15] Naimul M K, Riadh K, Imran S A, et al. Covariance-guided one-class support vector machine[J]. Pattern Recognition, 2014, 47(6): 2165–2177.

[16] Sun Y M, Kamel M S, Wong A K. Cost-sensitive boosting for classification of imbalanced data[J]. Pattern Recognition, 2007, 40: 3358–3378.

[17] Freund Y, Schapire R E. Experiments with a new boosting algorithm[C]// Proceedings of the Thirteenth International Conference on Machine Learning, 1996: 148–156.

[18] 曹莹, 苗启广, 刘家辰, 等. AdaBoost 算法研究进展与展望 [J]. 自动化学报, 2013, 39(6): 745–758.
Cao Y, Miao Q G, Liu J C, et al. Advance and prospects of AdaBoost algorithm[J]. Acta Automatica Sinica, 2013, 39(6): 745–758.

[19] Thomas G D. Ensemble learning[M]// Handbook of Brain Theory and Neural Networks, 2002.

[20] Cover T, Hart P. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967: 21–27.

[21] 张顺, 张化祥. 用于多标记学习的 K 近邻改进算法 [J]. 计算机应用研究, 2011, 28(12): 4445–4450.
Zhang S, Zhang H X. Modified KNN algorithm for multi-label learning[J]. Application Research of Computers, 2011, 28(12): 4445–4450.

[22] 汪云云, 陈松灿. 基于 AUC 的分类器评价和设计综述 [J]. 模式识别与人工智能, 2011, 24(1): 64–71.
Wang Y Y, Chen S C. A survey of evaluation and design for AUC based classifier[J]. PR&AI, 2011, 24(1): 64–71.

[23] Richard M E, Jonathan E F. Multi-class ROC analysis from a multi-objective optimisation perspective[J]. Pattern Recognition Letters, 2006(27): 916–927.

[24] Hassan M R, Ramamohanarao K, Karmakar C K, et al. A novel scalable multi-class ROC for effective visualization and computation[C]// Advances in Knowledge Discovery and Data Mining, 14th Pacific-Asia Conference, 2010: 107–120.

[25] 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): 211–217.
Mao Y, Zhou X B, Xia Z, et al. A survey for study of feature selection algorithms[J]. PR&AI, 2007, 20(2): 211–217.

[26] Chuang L Y, Chang H W, Tu C J, et al. Improved binary PSO for feature selection using gene expression data[J]. Computational Biology and Chemistry, 2008, 32(1): 29–38.

[27] Krawczyk B, Wozniak M, Schaefer G. Cost-sensitive decision tree ensembles for effective imbalanced classification[J]. Applied Soft Computing, 2014(14): 554–562.

[28] Krawczyk B, Schaefer G. An improved ensemble approach for imbalanced classification problems[C]// 8th IEEE International Symposium on Applied Computational Intelligence and Informatics, 2013.

[29] Fernández Á, López V, Galar M, et al. Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches[J]. Knowledge-Based Systems, 2013(42): 97–110.

阅读此文的还阅读了：

1. 一种不均衡数据的改进蚁群分类算法
2. 数据采集算法的研究
3. 面向不均衡小样本训练集的改进Boosting算法
4. Boosting算法及其在化学数据挖掘中的应用
5. 一种基于Boosting的油田水淹层识别算法
6. 基于Gradient Boosting的车载LiDAR点云分类
7. 基于ODR和BSMOTE结合的不均衡数据SVM分类算法
8. 基于Boosting机制的决策树集成分类器识别嗜热和常温蛋白
9. 基于形状无关纹理和Boosting学习的人 口统计学分类1
10. 集成学习算法在不平衡分类中的应用研究
11. 基于集成的非均衡数据分类主动学习算法
12. 基于半监督学习的数据流混合集成分类算法
13. 不均衡数据集中KNN分类器样本裁剪算法
14. 基于C4．5和Boosting算法的数据库负载自动识别
15. 基于流形学习的单类分类算法及其在不均衡声目标识别中的应用
16. 一种基于核的模糊多球分类算法及其集成

17. Boosting算法在文本自动分类中的应用

18. 一种结合半监督Boosting方法的迁移学习算法

19. 浅议企业应用集成

20. 赋石水库数字智能化管理系统探讨

21. 不均衡数据集中基于Adaboost的过抽样算法

22. 一种鲁棒的基于在线boosting目标跟踪算法研究

23. 集成学习:Boosting算法综述

24. 基于数据关系的svm多分类学习算法

25. 一种适用于不均衡数据集分类的KNN算法

26. 电力企业信息集成技术应用

27. 一种基于抽样与约简的集成学习算法

28. 关于警务管理系统的体系架构的研究

29. Online blind source separation based on joint diagonalization

30. 一种基于Web的分类体系学习算法

31. 基于MAS的异构多数据源集成

32. 一种用于非平衡数据分类的集成学习模型

33. 一种改进的基于概念格的数据挖掘算法

34. 数据网格建设与系统集成

35. DTN中一种基于内容分类的数据分发算法

36. 一种鲁棒的基于在线boosting目标跟踪算法研究

37. 基于Boosting的BAN组合分类器

38. 基于Ranking Loss的多标签分类集成学习算法

39. 一种不均衡数据集的决策树改进算法

40. 一种基于聚类的不平衡数据分类算法

41. 一种基于免疫聚类算法的数据分类

42. 不均衡数据下基于SVM的故障检测新算法

43. 基于谱聚类欠取样的不均衡数据SVM分类算法

44. PDM系统在船舶设计业务中的集成研究和应用

45. Boosting算法在基因表达谱样本分类中的应用

46. 不均衡数据下基于SVM的故障检测新算法

47. 一种基于遗传算法的受限制的分类器学习算法

48. DTN中一种基于内容分类的数据分发算法

49. 基于分类的集成学习算法研究

50. 基于PDM平台现代集成线束制造系统研究