基于概率采样和集成学习的不平衡数据分类算法

曹雅茜 黄海燕

(华东理工大学化工过程先进控制和优化技术教育部重点实验室 上海 200237)

摘 要 集成学习由于泛化能力强,被广泛应用于信息检索、图像处理、生物学等类别不平衡的场景。为了提高算法在不平衡数据上的分类效果,文中提出一种基于采样平衡和特征选择的集成学习算法OBPD-EFSBoost。该算法主要包括3个步骤:首先,依据少数类高斯混合分布得到的概率模型,进行过采样构造平衡数集,扩大少数类的潜在决策域;其次,每轮训练个体分类器时,根据上一轮的错分样本综合考虑样本和特征的加权,过滤冗余噪声特征;最后,通过个体分类器的加权投票得到最终的集成分类器。8组UCI数据分类结果表明,该算法不仅有效提高了少数类的分类精度,同时还弥补了Boosting类算法对噪声特征敏感的缺陷,具有较强的鲁棒性。

关键词 不平衡数据分类,集成学习,特征选择,概率分布

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.05.031

Imbalanced Data Classification Algorithm Based on Probability Sampling and Ensemble Learning

CAO Ya-xi HUANG Hai-yan

(Key Laboratory of Advanced Process Control and Optimization for Chemical Processes (East China University of Science and Technology), Ministry of Education, Shanghai 200237, China)

Abstract Ensemble learning has attracted wide attention in imbalanced category circumstances such as information retrieval, image processing, and biology due to its generalization ability. To improve the performance of classification algorithm on imbalanced data, this paper proposed an ensemble learning algorithm, namely Oversampling Based on Probability Distribution—Embedding Feature Selection in Boosting (OBPD-EFSBoost). This algorithm mainly includes three steps. Firstly, the original data are oversampled based on probability distribution estimation to construct a balanced dataset. Secondly, when training base classifiers in each round, OBPD-EFSBoost increases the weight of misclassified samples, and considers the effect of noise feature on classification results, thus filtering the redundant noise feature. Finally, the eventual ensemble classifier is obtained through weighted voting on different base classifiers. Experimental results show that the algorithm not only improves the classification accuracy for minority class, but also eliminates the sensitivity of Boosting to noise features, and it has strong robustness.

Keywords Imbalanced data classification, Ensemble learning, Feature selection, Probability distribution

1 引言

在机器学习中,类别分布不均衡的现象被称为不平衡问题。将常规算法直接应用于该问题时,分类结果往往会偏向多数类,造成少数类无法被正确识别。而且,传统算法多数是基于整体准确率最大化来训练分类器,因此会忽略少数类样本的误分,从而影响传统分类器的分类结果。但在许多实际应用中,少数类样本往往比多数类样本更具有价值,例如,银行欺诈用户识别[1]、医学癌症诊断[2]和网络黑客入侵[3]等。

为解决上述问题,诸多学者主要从数据层面和算法层面入手提出方案。

数据层面主要是对训练集进行欠采样或过采样,从而使数据集达到平衡状态。但这两种数据处理方式均没有遵循数据自身的分布规律,当添加或删除的样本与真实分布不完全一致时,将会不可避免地引入噪声,扭曲了数据的空间分布[4-5]。

算法层面主要是依据类别分布不平衡的特点对传统算法进行改进。例如,集成学习算法通过每轮关注上一轮的错分样本,训练多个弱分类器构建强分类器,从而提高最终学习效果。Boosting技术是一种常见的集成学习算法,Adaboost[6],SMOTEBoost[7],RUSBoost[8],Boundary-Boost[9]和BNU-SVM[10]等均是基于Boosting的改进。其中,SMOTEBoost算法通过在部分少数类样本与其近邻样本的连线上随机取点,生成无重复的少数类样本,从而使样本集的类别达到平衡,最后再利用集成学习算法对分类器进行训练。但该算法由于在插入新样本时并未考虑少数类的分布,因此会带来很

到稿日期:2018-04-26 返修日期:2018-08-17

曹雅茜(1993—),女,硕士生,主要研究方向为机器学习、数据挖掘,E-mail:yaxi_cao@163.com;**黄海燕**(1972—),女,博士,副教授,主要研究方向为控制与优化、复杂工业过程建模,E-mail:huanghong@ecust.edu.cn(通信作者)。

多噪音样本，影响个体分类器精度，造成过泛化。RUSBoost 算法是通过随机选取多数类样本与少数类组成平衡的样本子集，然后利用多组差异化的样本子集训练个体分类器，但该方法在移除多数类样本时可能会丢失有助于提高分类效果的重要信息。Boundary-Boost 算法采用少数类的边界样本进行过采样，并在每轮训练中删除上一轮错分的合成样本，但该方法容易忽视其他区块的样本分布，导致小区块样本集的分类正确率较低。

此外，选择不同特征子集构建子空间，训练得到不同的个体分类器也是集成学习思想中一种行之有效的途径[11]。一般而言，特征集中或多或少都包含着一些对分类结果无贡献的冗余噪声特征，对学习问题有很大的负面影响[12]。利用特征选择不仅能有效降低特征空间的维数，加快算法的运行效率，而且有助于寻找更精简、更易理解的算法模型。为了提高特征选择算法对不平衡数据的处理能力，Yin 等[12]利用 Hellinger 距离指标来度量在不同类别中特征的分布情况，并 根据分布进行特征选择，但该方法对类别不平衡的敏感度不足。Alibeigi 等[13]在无监督环境下，利用概率密度分析各个特征中数据的分布状况，并通过特征之间的数据分布关系来进行特征选择。但是上述方法均是根据度量指标来排除相关性低的特征，选择适配度最高的特征集合作为选择结果，对于特征较多的高维数据集来说，可能会丢掉有用信息。

鉴于上述问题，本文从采样平衡和集成学习入手，针对 SMOTEBoost 易改变数据集原始分布的缺陷，提出利用概率分布估计进行过采样的方法；针对 SMOTEBoost 没有考虑噪声特征影响的缺陷，提出在 AdaBoost 的每轮迭代中，根据错分样本消除噪声特征干扰的方法。多组实验结果表明，本文所提算法可以有效改善不平衡数据中少数类分类精度低的问题。

2 算法

Hansen 和 Salamon 以神经网络为个体分类器的集成学习算法为例，分析指出集成学习比个体分类器性能优越的充分必要条件是：1）个体分类器平均准确率高；2）个体分类器之间存在较大的差异[14]。其中，条件 1）是指个体分类器可以取得较好的分类性能；条件 2）是指个体分类器需要在不同的子集上进行训练，提高多样性。使得样本空间的错误分布不同。如果分布相同，那么集成后还是得到相同的错误分布，性能改善相对有限。

考虑上述两个条件，本文提出了 OBPD-EFSBoost（Oversampling Based on Probability Distribution-Embedding Feature Selection in Boosting）算法。首先，对少数类建立高斯混合模型，依据概率分布估计结果进行过采样，构造平衡训练集。该方法能够合成更加准确的真实样本，挖掘少数类潜在的特征空间，同时不影响数据的真实分布，保证个体分类器的准确性与多样性。其次，对于每轮的错分样本，除增加错分样本自身权重之外，还需计算特征对错分样本的影响权重，排除权重低的无关特征，以减少对下一轮个体分类器的影响。最后，构成泛化能力强的集成分类器。

2.1 基于概率分布估计的过采样算法

基于概率分布估计的过采样算法（Oversampling Based on Probability Distribution，OBPD）是基于高斯理论构建平衡样本集的一种手段。其中，高斯混合模型（Gaussian Mixture Model，GMM）是对数据真实分布进行表示的参数模型，它是对单一高斯密度函数的扩展，可以用来逼近任意形状的概率密度。参数模型由 L 个高斯模型的混合加权得到，其表达式如下：

$$p(x)=\sum_{l=1}^{L}p(l)p(x|l)=\sum_{l=1}^{L}\pi_l N(x|u_l,\sigma_l) \tag{1}$$

其中，x 是一个 D 维向量；$w_l=1,2,\cdots,L$ 是加权权重，且满足 $\sum_{l=1}^{L}\pi_L=1$；$N(x|u_l,\sigma_l)$ 为第 l 个高斯概率分布，可表示为：

$$N(x|u_l,\sigma_l)=\frac{1}{(2\pi)|\sigma_l|^{1/2}}\exp\{-\frac{1}{2}(x-u_l)^{\mathrm{T}}\sigma_l^{-1}(x-u_l)\} \tag{2}$$

其中，u_l 和 σ_l 分别是均值向量和协方差矩阵。

对于连续特征，常用 EM 算法估计高斯混合分布的均值向量和协方差矩阵，但该算法容易陷入局部最优，且需要人为设定高斯模型的个数。因此本文首先选择 Figueiredo 等[15]提出的方法求解模型个数，该方法的优势在于可以自动确定高斯模型的个数；其次，利用 AIC 准则选择拟合真实分布最优的模型；最后，通过 Gibbs 抽样生成连续特征。对于离散型特征来说，由于无法直接使用高斯混合分布建模，因此本文采用轮盘赌选择法来合成新值。首先统计少数类中每个离散特征不同值的出现频率，然后根据相应的频率随机产生离散特征下的值。

OBPD-Sampling 算法的具体步骤如算法 1 所示。

算法 1 OBPD-Sampling

输入：D^+ 为不平衡训练集 D 中的少数类样本，n^+ 表示少数类样本数量，令 $d_i=\{x_1^i,\cdots x_{ac}^i,x_{ac+1}^i,\cdots x_{ac+ad+1}^i\}$ 表示数据集中有 ac 个连续型特征值和 ad 个离散型特征值；T 为 Gibbs 抽样迭代次数；过采样率为 200%

输出：s 个合成的少数类样本

1. 针对 D^+ 中样本的连续特征值，根据 EM 算法和 AIC 准则确定其高斯混合联合分布 P。
2. 针对 D^+ 中样本的离散特征值，分别计算 x_{ac+i}^i（其中，i 满足 i>=1&i<=ad+1）中每个特征值出现的频率。假设 x_{ac+i}^i 的特征值域为 $\{a_1,a_2,\cdots,a_k\}$，则每个特征值出现的频率为 $\{p_1,p_2,\cdots,p_v\}$。
3. 根据估计到的概率分布模型，对 D^+ 中的每个样本进行如下操作：
   1）对于 $\{x_1^i,x_2^i,\cdots,x_{ac}^i\}$，采用 Gibbs 进行随机抽样。计算其中一个特征在其他特征下的条件分布函数 $x_i^{(t+1)}\sim P(x_i|x_1^{(t+1)},\cdots,x_{i-1}^{(t)},x_{i+1}^{(t)},\cdots,x_{ac}^{(t)})$，循环执行 T 次生成连续特征值 $(\tilde{x}_1^i,\tilde{x}_1^i,\cdots,\tilde{x}_{ac}^i)$；
   2）对于 $(x_{ac+1}^i,\cdots,x_{ac+ad+1}^i)$ 根据每个特征下不同值出现的频率利用轮盘赌选择随机产生值，即 $(\tilde{x}_{ac+1}^i,\cdots,\tilde{x}_{ac+ad+1}^i)$；
   3）将步骤 1）和步骤 2）生成的值合并产生一个合成样本 $(\tilde{x}_1^i,\cdots,\tilde{x}_{ac}^i,\tilde{x}_{ac+1}^i,\cdots,\tilde{x}_{ac+ad+1}^i)$；
   4）循环执行步骤 1）－步骤 3）直到产生 s 个合成样本。

图 1 是 SMOTE，RWO 和 OBPD 3 种不同过采样方法的图示，其中圆点表示少数类，三角形表示多数类，五角星表示合成的少数类样本。图 1（a）是原数据的分布，图 1（b）为 SMOTE 采样后的数据分布，可以发现 SMOTE 在近邻点之间的线性插值忽略了数据的分布规律，可能会破坏原始数据的分布。图 1（c）为 RWO-Sampling[16] 采样后的数据分布，

RWO-Sampling 根据每个特征的概率分布，以随机游走的方式合成新样本，在一定程度上扩展了分类边界。但是这两类过采样方法并没有考虑到数据中普遍存在的小区块问题[17]，容易产生噪声数据。图 1(d)为 OBPD-Sampling 采样后的数据分布，通过对原数据进行高斯混合建模，能够发现潜在的子特征空间。与 SMOTE 和 RWO-Sampling 相比，OBPD-Sampling 能够通过更加准确的概率分布模型，合理增加少数类样本。

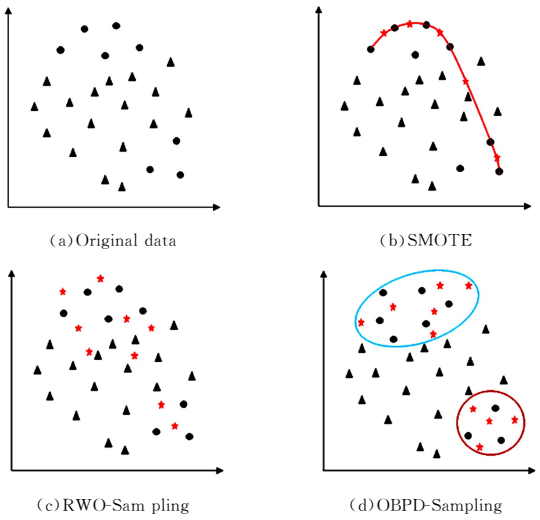

图 1　不同过采样方法的图示

Fig. 1　Illustrations of different over-sampling approaches

2.2　影响样本分类的两个要素

集成学习算法的原理是：通过在不同的数据子集上训练，得到多个有差异的个体分类器，然后将这些分类器加权结合，最终得到一个强分类器。

本文以朴素贝叶斯作为个体分类器。假设利用原始数据集的离散型特征集合 A_1 到 A_f（连续型特征可转化为离散型特征）判断类别 C，给定一个样本，特征值为 a_1 到 a_f，最优决策 $c=\arg_c \max P_r(C=c|A_1=a_1\wedge\cdots\wedge A_f=a_f)$。根据贝叶斯理论：

$$P_r(C=c|A_1=a_1\wedge\cdots\wedge A_f=a_f)=\frac{P_r(A_1=a_1\wedge\cdots\wedge A_f=a_f|C=c)}{P_r(A_1=a_1\wedge\cdots\wedge A_f=a_f)}P_r(C=c) \tag{3}$$

先验概率 $P_r(C=c)$可以从训练集中按式(4)估计得到：

$$\overline{P}_r(C=c)=\frac{count(C=c)}{count(trainset)} \tag{4}$$

根据贝叶斯原理可得：

$$\begin{aligned}P_r(A_1=a_1\wedge\cdots\wedge A_f=a_f|C=c)&=P_r(A_1=a_1|A_2=a_2\wedge\cdots\wedge A_f=a_f,C=c)\times\\&P_r(A_2=a_2|A_3=a_3\wedge\cdots\wedge A_f=a_f,C=c)\times\cdots\times\\&P_r(A_f=a_f|C=c)\end{aligned} \tag{5}$$

若假设特征之间相互独立，则可得：

$$\begin{aligned}P_r(A_1=a_1|A_2=a_2\wedge\cdots\wedge A_f=a_f,C=c)&=P_r(A_1=a_1|C=c)\end{aligned} \tag{6}$$

对 $A_2,\cdots,A_f$ 也做如上简化，可以得到：

$$\begin{aligned}P_r(A_1=a_1\wedge\cdots\wedge A_f=a_f,C=c)&=P_r(A_1=a_1|C=c)\times P_r(A_2=a_2|C=c)\times\cdots\times\\&P_r(A_f=a_f|C=c)\end{aligned} \tag{7}$$

其中，$P_r(A_i=a_i|C=c)$可由式(8)估计得到，所给出的估计值可以使样本出现的概率最大：

$$\overline{P}_r(A_i=a_i|C=c)=\frac{count(A_j=a_j\wedge C=c)}{count(C=c)} \tag{8}$$

以二分类为例，0 和 1 表示类别，其中一个测试样本的特征值表示为 $a_1,a_2,\cdots,a_f$，$b_0=P_r(C=0)$，$b_1=P_r(C=1)=1-b_0$，$b_{i0}=P_r(A_i=a_i|C=0)$，$b_{i1}=P_r(A_i=a_i|C=1)$。由上述推理可知，类别 0 和类别 1 的后验概率满足式(9)和式(10)：

$$p=P_r(C=1|A_1=a_1\wedge\cdots\wedge A_f=a_f)=(\prod_{i=1}^{f}p_{i1})b_1/z \tag{9}$$

$$q=P_r(C=0|A_1=a_1\wedge\cdots\wedge A_f=a_f)=(\prod_{i=1}^{f}p_{i0})b_0/z \tag{10}$$

其中，z 是用来归一化的常数。对式(9)和式(10)取对数并相减，可得：

$$\log p-\log q=\sum_{i=1}^{f}(\log p_{i1}-\log p_{i0})+(\log b_1-\log b_0) \tag{11}$$

设 $w_i=\log p_{i1}-\log p_{i0}$，$b=\log b_1-\log b_0$，可得：

$$\log\frac{p}{q}=\sum_{i=1}^{f}w_i+b \tag{12}$$

由式(12)可知，样本的分类同时由 b 和 w_i 决定。从特征的角度考虑，如果一个样本类别为 1，则朴素贝叶斯分类器得出 p 小于 q，那么该样本会被错分为类别 0。对于 $w_i(i=1,\cdots,f)$，若 $w_i<0$，说明特征 i 对样本的正确分类起了负面作用，如果将特征 i 删除，则样本被正确分类的可能性会增大。以上从理论上证明了排除噪声特征和选择最佳特征子集有助于提高分类正确率。将特征选择嵌入到每轮个体分类器的训练过程中，使得新一轮不仅可以使用有偏向的样本集，也可以使用有偏向的特征子集，进一步提高个体分类器的准确率和多样性，从而提高集成分类器的分类效果。

2.3　OBPD-EFSBoost 算法

OBPD-EFSBoost 算法首先通过 OBPD-Sampling 算法进行样本合成，然后结合 AdaBoost 和 Relief 算法的思想，得到新的样本集和特征集，训练每轮个体分类器，直至设定的迭代次数，最后采用加权投票法组合得到集成分类器。将特征选择嵌入到 AdaBoost 算法的思路，不仅可以弥补 AdaBoost 易受噪声特征影响的缺陷，同时还保留了 AdaBoost 将注意力集中于错分样本的特性。OBPD-EFSBoost 算法流程图如图 2 所示。

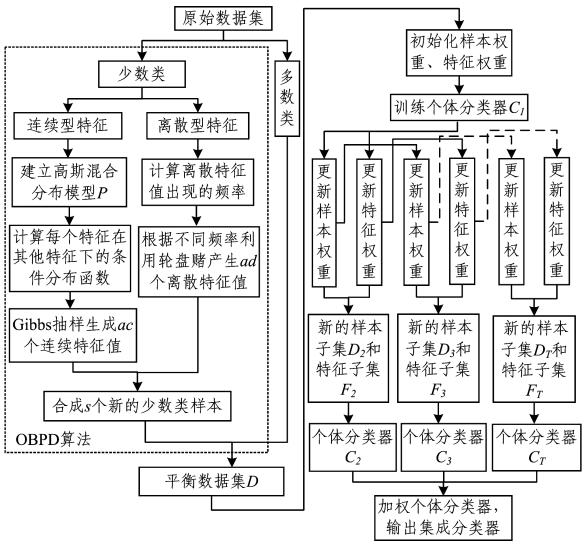

图 2　OBPD-EFSBoost 算法流程图

Fig. 2　Flow chart of OBPD-EFSBoost algorithm

Relief 是通过特征和类别的相关性来计算特征权重的一种特征选择算法。如果样本与异类最近邻 M 在某个特征上的距离小于与同类最近邻 H 的距离，则说明该特征对区分类别是有益的，应增加该特征的权重；反之，则降低该特征的权重。特征权重的更新公式如式(13)所示：

$$w(j)=w(j)+\frac{d(x_i(j),M_i(j))}{num}-\frac{d(x_i(j),H_i(j))}{num} \tag{13}$$

其中，$x_i(j)$表示所选样本 x_i 第 j 维特征的值；$d(A(j),B(j))$表示样本 A 和样本 B 在第 j 维特征上的距离，num 表示抽样次数。

Relief 算法运行效率高，相对于很多特征评估算法需要特征之间相互独立的假设，它对噪声和特征的相互作用并不敏感。因此，本文借鉴 Relief 算法的思想，根据每个特征对错分样本的影响程度来调整特征权重，目的是在训练下一轮个体分类器之前，排除干扰特征，使得所选特征子集可以在下一轮提高上轮错分样本被正确分类的概率。

设在第 t 轮迭代时，采用 $t-1$ 轮选择的特征子集构造分类器 C_t；对每个被 C_t 错分的样本 x_n，设 x_n 的类别为 y_n，分类器 C_t 对 x_n 的分类结果为 $C_t(x_n)$。如果 $\Pr(x_{nm}|y_n)-\Pr(x_{nm}|C_t(x_n))<0$，说明在计算概率 $\Pr(x_n,C_t)$时，特征 f_m 起负作用，应降低该特征的权值；反之，如果 $\Pr(x_{nm}|y_n)-\Pr(x_{nm}|C_t(x_n))>0$，则该特征起了正作用，应该增加 f_m 的权值。特征权值按式(14)更新。

$$w(f_m)=w(f_m)+\Pr(x_{nm}|y_n)-\Pr(x_{nm}|C_t(x_n)) \tag{14}$$

在特征权值确定之后，认为特征权值为负的特征对错分样本起到负作用，因此将其删除，这样就得到下一轮的特征子集 F'。在新的样本子集 D'和新的特征集 F'上训练下一轮个体分类器。最后通过多数投票法加权得到集成分类器。

OBPD-EFSBoost 算法的主要步骤如算法 2 所示。

算法 2 OBPD-EFSBoost

输入：不平衡训练集 D，特征集合 F=$\{f_1,f_2,\cdots,f_M\}$，分类算法 I，个体分类器数量为 T

输出：集成分类器 C(X)

1. 调用算法 1，生成平衡样本集 D′=$\{d_1,d_2,\cdots,d_N\}$。
2. 初始化样本权重为 $w(d_n)=1/N$，特征权重为 $w(f_m)=1/M$，F′=F。
3. 使用具有权值分布 D′的训练集学习，得到个体分类器 $C_t=I(D',F')$。
4. 计算个体分类器 C_t 的分类误差 $\varepsilon_t=\frac{1}{n}\sum_{x_n\in D';C_t(x_n)\neq y_n}w(x_n)$。
5. 根据分类误差计算：$\beta_t=\frac{\varepsilon_t}{1-\varepsilon_t}$。
6. 计算个体分类器 C_t 的系数：$\alpha_t=\frac{1}{2}\log\frac{1}{\beta_t}$。
7. 更新训练数据集的权值分布。对每个 $d_n\in D'$，如果 $C_t(x_n)=y_n$，则 $w(d_n)=w(d_n)\beta_t$。
8. 根据式(14)计算特征权重。
9. 如果 $w(f_m)<0$，则 $F'=\{F'-f_m\}$。
10. t++，如果 t<T，返回步骤 3；
11. 加权得到集成分类器：

$$C(X)=\arg\max_{y\in Y}\sum_{t:C_t(X)\neq y}\alpha_t$$

3 分类性能评价

在评价分类性能和指导分类器建模时，评估度量起着至关重要的作用。以二分类为例，学习结果可以由表 1 的混淆矩阵表示，本文将关注的少数类定义为正类，多数类定义为负类。

表 1 二分类问题的混淆矩阵

Table 1 Confusion matrix for two-class problem

	预测为正类	预测为负类
实际为正类	*TP*	*FN*
实际为负类	*FP*	*TN*

相关评估指标如下。

查准率(Precision)：

$$Precision=\frac{TP}{TP+FP} \tag{15}$$

查全率(Recall)：

$$Recall=\frac{TP}{TP+FN} \tag{16}$$

特效性(Specificity)：

$$Specificity=\frac{TN}{TN+FP} \tag{17}$$

F 值(F-value)：

$$F\text{-}value=\frac{(1+\beta^2)\times recall\times precision}{\beta^2\times recall+precision} \tag{18}$$

G 均值(G-mean)：

$$G\text{-}mean=\sqrt{recall\times specificity} \tag{19}$$

查准率表示正类样本被正确分类的比例；查全率表示正类样本被正确分类的完整度；特效性表示负类样本被正确分类的完整度；F-value 是查全率和查准率的调和均值，参数 β 表示 Recall 与 Precision 的相对重要性，通常取 $\beta=1$；G-mean 同时关注两个类的性能，表示正类和负类分类准确率的均衡值。

本文采用 G-mean 和 F-value 作为分类器性能的评价指标。G-mean 的目标是保持正负两类分类精度平衡条件下的总体精度的最大化，若正类的分类精度较高但负类分类精度较低，则 G-mean 较低，只有当两者都较高时，G-mean 才会较高。F-value 值高则表示查全率和查准率都比较高。

4 实验设计与结果分析

4.1 数据集

为检验和评估本文算法的性能，实验选取了 8 组 UCI 不平衡数据集进行测试，相应的类别标签取值分别为{0，+1}，数据集详细信息如表 2 所列。

表 2 实验数据集信息

Table 2 Information of experimental datasets

数据集	特征数	样本数	不均衡比
Ionosphere	34	351	1.79
Credit-g	24	1 000	2.33
Sonar	60	208	1.14
Breat-cancer	9	277	2.42
Segment	19	1 500	6.37
Letter	16	20 000	25
Vowel	13	990	10
Vehicle	18	846	3.25
Yeast	8	1 484	28.1

4.2 实验设置

在 Matlab2016b 实验环境下对本文算法进行验证，根据

3 种分类指标来评价本文提出的算法。本文将 OBPD-EFSBoost 算法与 AdaBoost，EFSBoost，SMOTEBoost，RWOBoost 和 OBPDBoost 进行比较。其中，EFSBoost 算法是在 AdaBoost 的每轮训练中对噪声特征进行过滤；OBPDBoost 是基于本文所提算法 1，将 SMOTEBoost 算法的 SMOTE 阶段替换为 OBPD。

为了更加全面地评估该算法，采用五折交叉验证，将算法运行 10 次的平均值作为最终的结果。其中，4 种算法的个体分类器均采用朴素贝叶斯模型（Naive Bayesian Model，NBC），个体分类器设为 50 个。SMOTEBoost 和 RWOBoost 算法的过采样率设为 200%，OBPD-EFSBoost 中 OBPD 阶段的 Gibbs 采样步长设为 100。所有实验均在 4 * 1.60 GHz，8 GB 内存，Matlab2016b 的环境下完成。

4.3　实验结果分析

从图 3 和图 4 可以看出，本文提出的算法 OBPD-EFSBoost 在除 Breat-Cancer 和 Vowel 之外的 7 组数据集的两个分类指标上都得到了不同程度的提高，与 SMOTEBoost 算法相比，9 组数据在 F-value 和 G-mean 两个指标的提高平均值达到了 6.68%和 1.92%。

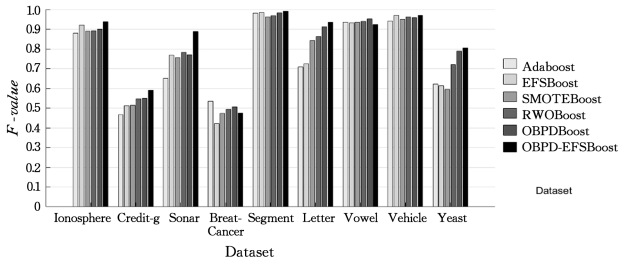

图 3　几种算法的 *F-value* 值对比

Fig. 3　Comparison of *F-value* for different algorithms

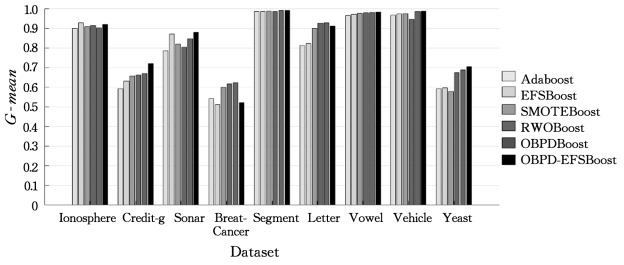

图 4　几种算法的 *G-mean* 值对比

Fig. 4　Comparison of *G-mean* for different algorithms

为了进一步分析采样平衡和特征选择这两种策略，我们将 OBPDBoost 与 SMOTEBoost，RWOBoost 两个算法的实验结果进行对比，以验证 OBPD 算法的优势；设计一个新的实验，将 OBPD-EFSBoost 与 AdaBoost 算法进行对比，研究两种算法在噪声特征影响下的鲁棒性。

（1）基于概率分布估计的过采样效果分析

对比图 3 和图 4 可以发现，RWO 和 OBPD 两种基于概率的过采样方法所合成的样本集普遍优于 SMOTE 这种基于近邻的过采样方法，这说明了采样时考虑数据中不同类别的概率分布的重要性。RWO 的过采样策略虽然也是基于数据分布合成新样本，但其随机游走的方式容易引入干扰样本，未能很好地保证过采样的质量。

以 Letter 为例，OBPDBoost 相比于 SMOTEBoost 和 RWOBoost，*F-value* 值的提高率分别为 8.17%和 5.56%，*G-mean* 值的提高率分别为 3.20%和 1.22%。这进一步说明了利用 OBPD 算法合成的样本更加符合少数类的分布规律，有助于后期对分类模型的训练。

（2）OBPD-EFSBoost 算法对噪声特征的鲁棒性分析

为了研究 OBPD-EFSBoost 算法依据每轮错分样本对噪声特征的过滤效果，本文对实验中最后一轮所剩的特征进行统计，统计结果如表 3 所列。

表 3　最后一轮所选特征信息

Table 3　Selected feature information in last round

数据集	所选特征序号
Ionosphere	5,7,13,24,3,8,12,15
Credit-g	2,10,1,4,5,8,3
Sonar	11,10,36,21,12,37,20,45,44,46
Breat-cancer	4,6,5
Segment	10,11,12,17,2,13,14
Letter	11,15,7,14,9,12,16,10
Vowel	5,7,9,4,6
Vehicle	12,11,7,4,13,3,10,1,8,17,18,14,6
Yeast	4,6,3,8,2

结合分类效果，OBPD-EFSBoost 每轮过滤噪声特征的优势在 Ionosphere，Credit-g，Vehicle 和 Sonar 4 组数据集上得到一定的体现，尤其对于高维数据集来说，OBPD-EFSBoost 算法过滤冗余噪声特征的优势更为明显。相比于 AdaBoost 算法，本文方法在指标 F-value 和 G-mean 上的平均提高率为 8.94%和 5.29%。

需要说明的是，Credit-g 数据集在多组算法上表现不佳的原因是其中有多个特征存在值丢失，且丢失情况在两类间不平衡。为验证算法受噪声特征的影响程度，我们手动将 Credit-g 丢失值最多的一维特征删除，构造数据集 Credit-g1。

Ionosphere 是判断雷达信号是否正常的数据集，第一维是雷达的编号，其余为天线收集到的电磁波信号。为加强实验可靠性，将 Ionosphere 的第一维特征删除，形成样本集 Ionosphere1；给 Ionosphere 每个样本添加一组 0 到 1 之间的随机数作为新的一维特征，构造样本集 Ionosphere2。表 4 列出了 OBPD-EFSBoost 和 SMOTEBoost 算法在这 5 组数据集上的分类效果。

表 4　噪声特征存在时两种算法的分类效果对比

Table 4　Comparison of two algorithms when noise features exist

样本集	SMOTEBoost		OBPD-EFSBoost	
	F-value	*G-mean*	*F-value*	*G-mean*
Credit-g	0.5136	0.6582	**0.5909**	**0.7205**
Credit-g1	0.5751	0.7000	**0.5913**	**0.7193**
Ionosphere	0.8887	0.9075	**0.9374**	**0.9187**
Ionosphere1	0.8845	0.9144	**0.9296**	**0.9173**
Ionosphere2	0.6472	0.6724	**0.9381**	**0.9200**

由表 4 可以看出，相比于 Credit-g 数据集，SMOTEBoost 算法在删除掉噪声特征后的 Credit-g1 上的分类准确率均有提升，可见 Boosting 类算法对此类噪声比较敏感。类似的情况也发生在 Ionosphere2 上，增加干扰噪声特征后，两个指标值均出现下降。而 OBPD-EFSBoost 算法因为每一轮个体分类器都是基于特征选择得到的，所以对噪声特征有较强的鲁

棒性。观察实验过程可知,我们在 Ionosphere2 数据集上手动增加的特征在第 4 轮之前就会被过滤掉,进一步验证了 OBPD-EFSBoost 算法的稳定性和可靠性。

(3)其他实验效果分析

从图 3 和图 4 可以看出,在 Breat-cancer 数据集上,指标 F-value 和 G-mean 出现了下降,相比于 AdaBoost 算法分别下降了 21.24%,10.64%。结合表 3 可知,Breat-cancer 最后一轮所选特征只剩下 3 维,所以可能是由于 Breat-cancer 作为低维数据集,可抽取的特征有限,在进行多轮特征选择后,会丢掉一些相对重要的特征,导致分类精度下降。

在 Letter 和 Vowel 数据集上,F-value 和 G-mean 两个指标值出现了一个上升一个下降的情况。我们将两组数据的 Recall 和 Specificity 指标在表 5 中进行对比。可以发现,两组数据集因样本数量多,不均衡比高,因此在提升少数类分类精度时,会以牺牲多数类分类精度作为代价。但总体来说,多数类的错分会维持在一个可控的范围内。由表 5 可以看出,相比于多数类准确率(Specificity)的降低,OBPD-EFSBoost 算法对少数类正确率(Recall)的提升程度更高。

表 5 两组数据集的查全率和特效性比较

Table 5 Comparison of recall and specificityon two datasets

数据集	算法	*Recall*	*Specificity*
Letter	SMOTEBoost	0.8727	**0.9372**
	OBPD-EFSBoost	**0.9291**	0.9315
Vowel	SMOTEBoost	0.9135	**0.9748**
	OBPD-EFSBoost	**0.9827**	0.9413

从算法的时间复杂度考虑,OBPD-EFSBoost 由于比 SMOTEBoost 算法在第一阶段多了高斯建模一项,在第二阶段多了特征评估一项,因此运行时间比 SMOTEBoost 算法长。但是,由于每轮会减少多个噪声特征,所以缩短训练个体分类器的时间会缩短。因此,OBPD-EFSBoost 的运行效率和 SMOTEBoost 是同等数量级的,对于高维数据集来说,运行效率甚至更高。

结束语 本文提出了一种有效改善不平衡数据分类效果的集成学习算法。首先,通过估计原始数据集中少数类的概率分布,合理进行过采样;然后,根据上一轮的错分样本,对样本权值和特征权值进行调整,过滤掉噪声特征,训练下一轮个体分类器;最后,加权集成得到最终的分类器。该方法由于对原始数据集分布影响小,在训练中通过更新特征权值排除影响分类效果的噪声特征,使得分类精度取得了较为明显的提升,尤其在高维数据集上提升更为显著。同时,所选的最佳特征也对实际应用有一定的指导意义。

参 考 文 献

[1] POZZOLO A D,CAELEN O,BORGNE Y A L,et al. Learned Lessons in Credit Card Fraud Detection from A Practitioner Perspective[J]. Expert Systems with Applications,2014,41(10):4915-4928.

[2] PARVIN H,MINAEIBIDGOLI B,ALINEJADROKNY H. A New Imbalanced Learning and Dictions Tree Method for Breast Cancer Diagnosis[J]. Journal of Bionanoscience,2013,7(6):673-678.

[3] LARADJI I H,ALSHAYEB M,GHOUTI L. Software Defect Prediction Using Ensemble Learning on Selected Features[J]. Information & Software Technology,2015,58:388-402.

[4] ZHANG C,WANG G,ZHOU Y,et al. A new approach for imbalanced data classification based on minimize loss learning[C]// IEEE Second International Conference on Data Science in Cyberspace. IEEE,2017:82-87.

[5] CAO P,YANG J,LI W,et al. Hybrid Sampling Algorithm Based on Probability Distribution Estimation[J]. Control and Decision,2014,29(5):815-520. (in Chinese)
曹鹏,李博,栗伟,等. 基于概率分布估计的混合采样算法[J]. 控制与决策,2014,29(5):815-520.

[6] FREUND,YOAV,SCHAPIRE,et al. A Decision-theoretic Generalization of On-line Learning and An Application to Boosting [C]// European Conference on Computational Learning Theory. Springer,Berlin,Heidelberg,1995:23-37.

[7] CHAWLA N V,BOWYER K W,Hall L O,et al. SMOTE:synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research,2002,16(1):321-357.

[8] SEIFFERT C,KHOSHGOFTAAR T M,HULSE J V,et al. RUSBoost:A Hybrid Approach to Alleviating Class Imbalance [J]. IEEE Transactions on Systems Man & Cybernetics Part A Systems & Humans,2009,40(1):185-197.

[9] LI K,FANG X,ZHAI J,et al. An Imbalanced Data Classification Method Driven by Boundary Samples-Boundary-Boost[C]// International Conference on Information Science and Control Engineering. IEEE,2016:194-199.

[10] BAO L,CAO J,LI J,et al. Boosted Near-miss Under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets[J]. Neurocomputing,2016,172(C):198-206.

[11] YIN H,HUY P. An Imbalanced Feature Selection Algorithm Based on Random Forest[J]. Acta Scientiarum Naturalium Universitatis Sunyatseni,2014,53(5):59-65.

[12] YIN L,GE Y,XIAO K,et al. Feature Selection for High-dimensional Imbalanced Data[J]. Neurocomputing,2013,105(3):3-11.

[13] ALIBEIGI M,HASHEMI S,HAMZEH A. Unsupervised Feature Selection Based on the Distribution of Features Attributed to Imbalanced Data Sets[J]. International Journal of Artificial Intelligence & Expert Systems,2011,2(1):2011-2014.

[14] HANSEN L K,SALAMON P. Neural Network Ensembles [M]. IEEE Computer Society,1990,12(10):993-1001.

[15] FIGUEIREDO M A T,JAIN A K. Unsupervised Learning of Finite Mixture Models[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2002,24(3):381-396.

[16] ZHANG H,LI M. RWO-Sampling:A random walk over-sampling approach to imbalanced data classification[J]. Information Fusion,2014,20(1):99-116.

[17] WEISS G M. The Impact of Small Disjuncts on Classifier Learning[M]. Data Mining,2009:193-226.