

引用格式:杨杰明,高聪,曲朝阳,等.基于代价敏感的随机森林不平衡数据分类算法[J].科学技术与工程,2018,18(6):285—290
Yang Jieming, Gao Cong, Qu Zhaoyang, et al. Random forest classification algorithm based on cost-sensitive for imbalanced data[J]. Science Technology and Engineering, 2018, 18(6): 285—290

基于代价敏感的随机森林不平衡数据分类算法

杨杰明¹ 高聪^{1*} 曲朝阳¹ 阚中锋² 高冶² 常成²

(东北电力大学信息工程学院¹,吉林 132012;国网吉林供电公司²,吉林 132000)

摘要 随机森林在分类不平衡数据时,容易偏向多数类而忽略少数类。可以将代价敏感用于分类器的训练;但在传统代价敏感随机森林算法中,代价函数没有考虑样本集实际分布与特征权重,且在随机森林投票阶段,没有考虑基分类器的性能差异。提出一种改进的代价敏感随机森林算法 ICSRF,该算法首先根据不平衡数据集的实际分布构造代价函数;并将权重距离引入代价函数,然后根据基分类器的性能采取权重投票,提高分类准确率。实验结果表明,ICSRF 算法能有效提高少数类的分类性能,可以较好地处理不平衡数据。

关键词 代价敏感 随机森林 不平衡数据 权重距离

中图分类号 TP391.75; **文献标志码** A

不平衡数据分类是机器学习中的一个重要课题,不平衡数据出现在实际生活的各个领域,如电信行业、医疗诊断、病毒检测等^[1];其中,多数类远远超过少数类,数据呈现不平衡状态。而传统分类算法基于类别平衡假设。一般以提高分类准确率为目标,少数类的性能较差;但在不平衡数据中,少数类通常更重要。例如在医疗诊断问题上,将癌症患者误诊为无病的代价远远大于将无病误诊为癌症患者所付出的代价^[2]。因此基于分类准确率的传统分类算法并不适用在不平衡领域中,如何提升不平衡数据中少数类的识别率成为数据挖掘中迫切需要解决的问题^[3]。

为了解决不平衡数据给传统机器学习算法带来的挑战,多种改进策略被提出来,比如通过调整类别比例来平衡数据、在传统分类算法^[4]中引入误分类代价等。其中,调整类别比例主要是对数据进行过采样与欠采样,使数据集中不同类别的数量达到平衡,在一定程度上可以提高少数类的分类性能。在分类算法上,Adaboost 算法与代价敏感学习结合^[5]。在权重更新公式中引入代价;但是 Adaboost 算法的基分类器的建立都是在整个特征空间创建的,增加了训练时间。随机森林算法与代价敏感学习结合,

在属性分裂度量中引入误分类代价;但是传统代价敏感随机森林算法中的代价函数并没有考虑样本集实际分布;且在构建代价因子过程中,所有特征平等对待,采用欧式距离计算样本之间的距离,而特征空间中所有特征的重要性不相同;并且同一特征对不同类别的重要性也不同,构造代价函数时仅仅计算欧式距离对于重要特征不公平,构建的代价不准确,使代价敏感学习的性能得不到保证,最终造成分类器的性能较差。虽然随机森林在选择训练样本集、属性集过程中引入了随机性,可以较好地避开过拟合问题,但也导致了基分类器在处理不平衡数据时的性能差异;且当数据包含噪声时,随机森林训练的基分类器可能包含噪声,在分类预测阶段,如果进行平等投票,会造成随机森林中包含噪声树,降低分类器的整体性能。

本文提出一种改进的代价敏感随机森林算法 ICSRF;该算法以代价敏感随机森林为基础,首先根据样本实际分布构造代价因子,且在构建过程中引入权重距离。然后在分类器组合过程中利用 AUC 衡量基分类器的整体性能,将每个基分类器的 AUC 值决定在投票阶段所拥有的权重,将基分类器以能力赋予权值进行投票,提升分类器的性能。

1 相关技术

1.1 代价敏感学习

由于类别分布不平衡,分类算法通常会对多数类过度训练而导致少数类的分类性能较差。而代价敏感学习^[6]为不同的分类错误分配不同的代价,避

2017年7月22日收到 国家自然科学基金重点项目(51437003)和吉林省科技计划(20160623004TC)资助
第一作者简介:杨杰明(1972—),博士,教授。研究方向:文本分类和机器学习。E-mail:yjmlzy@sina.com。

*通信作者简介:高聪(1993—),女,硕士研究生。研究方向:大数据环境下不平衡数据学习分类。E-mail:1103878327@qq.com。

免产生高代价的分类错误,以达到最小化分类代价的目标^[7]。

代价敏感一般用代价矩阵(见表1)表示分类器错分时要付出的代价^[8], c_0 为少数类, c_1 为多数类, $F(i, j)$ 表示将 i 错分为 j 要付出的代价。

表1 代价矩阵
Table1 Cost matrix

类别	c_0	c_1
c_0	$F(c_0, c_0)$	$F(c_0, c_1)$
c_1	$F(c_1, c_0)$	$F(c_1, c_1)$

代价矩阵确定后,利用贝叶斯定理构建风险函数,如式(1)所示:

$$R(c_i|x) = \sum P(c_j|x)F(c_j, c_i) \quad (1)$$

式(1)中, $P(c_j|x)$ 表示把样本 x 分为 c_j 类的后验概率。

样本 x 被划分为风险函数值最小的类别,如式(2)所示:

$$c = \underset{1 \leq i \leq l}{\operatorname{argmin}} \{R(c_i|x)\} \quad (2)$$

1.2 决策树

决策树对噪声具有较好的容错性^[9]。决策树首先对样本进行训练,得到决策树模型,然后利用决策树分支节点对测试样本的特征值进行对比,最终在叶子节点完成测试样本的类别判定。

决策树算法有多种,如 ID3, C4.5^[10] 等,算法按照一定的选择原则挑选属性进行分裂, C4.5 采用信息增益率。设 S 是数据集,划分为 n 个子集的信息熵为

$$E(S) = - \sum_{i=1}^n p_i \lg p_i \quad (3)$$

式(3)中, p_i 表示样本中属于第 i 个子集的概率。

假设使用属性 A 进行分裂, A 有 X_A 个不同的属性值, S_v 表示数据集中属性 A 的值为 v 的数据子集。 $E(S_v)$ 表示使用属性 A 进行分裂后,对分支节点的样本集 S_v 分类的熵, $E(S, A)$ 定义为选择属性 A 导致的期望熵,如式(4)所示:

$$E(S, A) = \sum_{v \in X_A} \frac{|S_v|}{|S|} E(S_v) \quad (4)$$

属性 A 对于数据集 S 的信息增益表示为

$$\operatorname{Gain}(S, A) = E(S) - E(S, A) \quad (5)$$

信息增益率定义为

$$\operatorname{Gain_ratio}(S, A) = \frac{\operatorname{Gain}(S, A)}{\operatorname{split_info}_A(D)} \quad (6)$$

式(6)中, $\operatorname{split_info}_A(D)$ 为分裂信息,表示训练集 D 关于 A 的熵,定义为

$$\operatorname{split_info}_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{|D|} \quad (7)$$

1.3 代价敏感随机森林

随机森林是一个包含多棵决策树 $\{h(X, \theta_k), k = 1, 2, \dots, k\}$ 的集成分类器, k 表示随机森林里基分类器的个数,决策树的构建基于两个随机性,通过 Bagging 有放回地随机抽样,在随机子空间进行模型的训练,保证了基分类器的多样性^[11]。由于在选择样本、特征子空间过程中引入了随机性,所以可以较好地避开过拟合问题^[12],提高了分类的精度。最后,随机森林算法组合多棵决策树的分类结果,根据少数服从多数的方式决定样本类别,如式(8)所示:

$$H(x) = \operatorname{argmax}_y \sum_k I[h_k(x) = y] \quad (8)$$

式(8)中, $h_k(x)$ 是决策树模型, y 是决策树的分类结果, $I(\cdot)$ 是指示器函数。

随机森林算法引入代价敏感,一般在决策树属性分裂中引入误分类代价,根据误分类代价的下降值代替传统的基尼指数、信息增益,选择能够使误分类代价下降最快的属性。代价下降值指选择属性 A_i 作为分裂属性与未分裂前的误分类代价的差值,如式(9)所示:

$$\operatorname{Rec} = Mc - \sum_{i=0}^n Mc(A_i) \quad (9)$$

式(9)中, Rec 表示代价下降值, Mc 表示未分裂前的代价值, $\sum_{i=0}^n Mc(A_i)$ 表示选择属性 A_i 分裂后的误分类代价,其中属 A_i 有 n 个值。

2 改进的代价敏感随机森林算法 (ICSRF)

随机森林可以较好地避开拟合,引入代价敏感可以处理不平衡问题。但是代价函数构建的不准确,则达不到处理不平衡数据的目的。传统代价函数的构造没有考虑数据集的实际分布,且采用欧式距离计算样本距离,而特征空间中所有特征的重要性不同,并且同一特征对不同类别的重要性也不同,仅仅计算欧式距离对重要特征不公平,构造的代价函数不准确,导致分类器的整体性能较差。由于随机森林在选择训练样本、特征子空间过程中引入了随机性,导致了基分类器在处理不平衡数据时的性能差异,而传统随机森林算法在最终决策阶段采取平等投票,平等投票会影响分类器的整体性能。

ICSRF 算法重新构造代价函数,在构建过程中考虑样本集实际分布与特征权重,代价函数构造过程在 2.1 节中详细介绍。在基分类器组合阶段,针对不平衡数据,每棵决策树使用 AUC 值进行性能的

评估,利用 AUC 值对数据进行加权投票,权重越大,说明该基分类器性能较好,在最后决策阶段占的权重越大,对于分类性能差的,权重越小,对结果的影响就小。最后随机森林分类器的输出为

$$H_c(x) = \arg \max_y \sum_k \alpha_k I[h_k(x) = y] \quad (10)$$

式(10)中, α_k 表示决策树的投票权重。

2.1 代价函数的构造

根据样本实际分布构造代价因子,将权重距离引入代价函数的计算过程。详细步骤如下。

Step1 分别计算多数类、少数类与整个数据集的数据中心。

计算方法是取每个特征列的算术平均数。数据集表示为如式(11)的矩阵, c 代表类别,每一行代表一个数据样本,每一列代表数据的特征。

$$\begin{bmatrix} x_{11} & \cdots & x_{1m} & c \\ \vdots & & \vdots & c \\ x_{n1} & \cdots & x_{nm} & c \end{bmatrix} \quad (11)$$

多数类的中心计算如下:

$$A_1 = \frac{1}{n} \sum_{i=1}^n x_{i1} \quad (12)$$

$$A_2 = \frac{1}{n} \sum_{i=1}^n x_{i2} \quad (13)$$

\vdots

$$A_m = \frac{1}{n} \sum_{i=1}^n x_{im} \quad (14)$$

得到多数类 c_1 的中心 (A_1, A_2, \dots, A_m) , 以同样的方法计算少数类 c_0 和整个数据集 N 的中心。

Step2 计算各类别中心到整个数据集中心的权重距离。

在数据集中,重要特征相对较少,计算类别中心到整个数据集中心的欧式距离构造代价对重要特征不公平,本算法引入权重距离,利用信息增益衡量每个特征在不同类别中的重要性,如式(15)所示。所有特征在多数类中产生权重向量 $\mathbf{w}(w_1, w_2, \dots, w_m)$, 在少数类中产生向量 $\mathbf{w}'(w'_1, w'_2, \dots, w'_m)$, 计算类别与整个训练集的权重距离时,分别加入各自的权重向量。权重距离如式(16)所示:

$$IG(x_k, c_i) = \sum_{c \in \{c_i, c_j\}} \sum_{x \in \{x_k, x_k\}} P(x, c) \lg \frac{P(x, c)}{P(x)P(c)} \quad (15)$$

式(15)中, $P(c)$ 表示类别 c 在总数据集的概率, $P(x, c)$ 表示在类别 c 中包含特征 x 的概率, $P(x)$ 表示数据集中包含特征 x 的概率。

$$d_i = \sqrt{\sum_{j=1}^m w_j (A_{ij} - \bar{A})^2} \quad (16)$$

式(16)中, w_i 表示特征在多数类的权重值, A_{ij} 是多数类的类别中心,当计算少数类与数据集中心的权重距离时采用特征在少数类中产生的权重向量,类别中心采用少数类的类别中心, \bar{A} 是整个数据集的中心。

Step3 定义 γ 系数。

对于不平衡数据集 N , 包含多数类 c_1 , 少数类 c_0 , 每个类别里有 N_1 和 N_2 个样本,为类别 c_i 定义 γ 系数:

$$\gamma_i = \frac{\sum_{j=0}^1 N_j}{N_i} \quad (17)$$

Step4 构造代价函数。

$$F(c_i, c_j) = \begin{cases} \gamma_i \frac{d'_i}{d''_j}, & d'_i < d''_j \\ \gamma_i \frac{d'_i}{d''_j}, & d''_j < d'_i \\ 0, & i = j \\ 1, & d'_i = d''_j \end{cases} \quad (18)$$

式(18)中, d'_i 表示类别 c_i 中心与整个数据集中心的权重距离, d''_j 表示 c_j 与整个数据集中心的权重距离。

2.2 算法步骤

Step1 按照 2.1 节构造的代价函数计算误分类代价 $F(c_1, c_0)$ 、 $F(c_0, c_1)$ [式(11)~式(17)所示]。

Step2 在原始数据集中通过 Bagging 方法采样,得到 k 个训练子集。

Step3 对于每个训练子集,执行以下步骤。

(1) 从原始数据集的特征中随机抽取 m 个特征。

(2) 在特征子集中计算代价下降值 Rec [式(9)所示]。

(3) 每次选择最大的 Rec 对节点分裂,生成不剪枝的决策树;

Step4 每棵决策树对 Out-of-Bag 样本进行测试,得到每棵决策树的 AUC 值。

Step5 利用 AUC 值对每棵决策树赋予权值。

Step6 对于测试集,加权决策树对样本进行权重投票。

3 实验结果与分析

3.1 实验数据集与评价准则

为了验证 ICSRF 算法的有效性,选择 6 组 UCI 数据集,对多类数据集进行调整,转化为两类数据集,使数据呈现出不平衡状态^[13],如表 2 所示。

表 2 数据集信息

Table 2 The information of data sets

数据集	数据 集数目	特征 个数	少数类 个数	多数类 个数	稀有率/ %
breast-cancer	286	10	85	201	42.3
glass	214	10	70	144	48.6
balance-scale	625	5	49	576	8.5
heart-h	294	14	106	188	56.4
waveform	5 000	21	1 650	3 350	49
diabetes	768	9	268	500	53.5

在不平衡数据中,少数类的分类性能更重要,准确率不能合理的衡量不平衡数据的分类性能^[14],为了更有意义的评估分类算法的性能,通常采用混合矩阵的方法,如表 3 所示。本次实验选择 F-measure、AUC 值作为评价准则衡量算法性能,利用 TPrate 衡量算法对少数类的分类正确率。

表 3 混合矩阵

Table 3 Hybrid matrix

类别	被分为少有类	被分为多数类
实际为少有类	TP	FP
实际为多数类	FN	TN

3.2 实验结果

利用 UCI 数据集对决策树(C4.5)、随机森林分类器(RF)、传统代价敏感随机森林(CSRF)与提出的 ICSRF 算法进行对比。本次实验过程中,传统代价敏感随机森林的误分类代价按照数据类别的不平衡度决定,比如在数据集 breast-cancer 中,多数类与少数类的比例 2.4,如果 $F(c_1, c_0)$ 的值为 1,则 $F(c_0, c_1)$ 为 2.4,其中 c_1 为多数类, c_0 为少数类。实验结果如图 1、图 2、图 3 所示。

图 1 和图 2 分别显示了四种不同分类算法在六组 UCI 数据集上的 AUC, F-measure 性能比较结果,实验记录的是多数类与少数类的 AUC, F-measure 的平均值。可以看出 ICSRF 算法整体性能优于其他

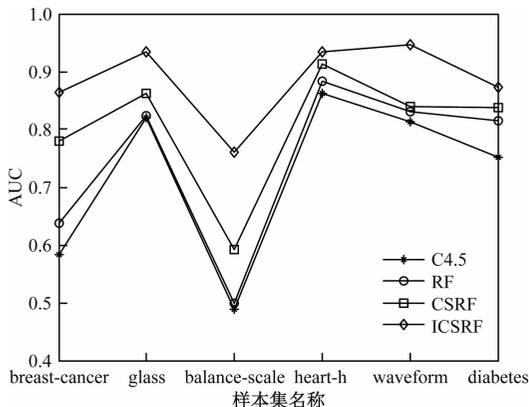


图 1 C4.5、RF、CSRF、ICSRF 的 AUC 值比较

Fig. 1 The AUC information of C4.5, RF, CSRF and ICSRF

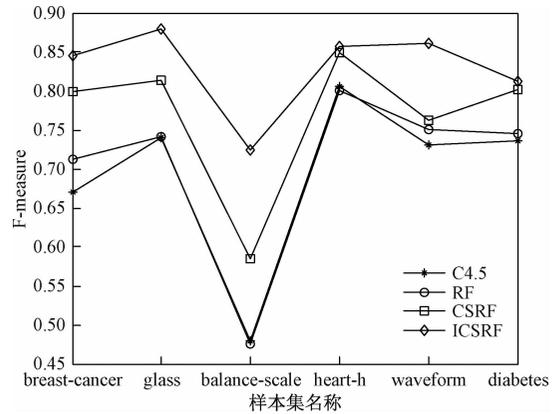


图 2 C4.5、RF、CSRF、ICSRF 的 F-measure 比较

Fig. 2 The F-measure information of C4.5, RF, CSRF and ICSRF

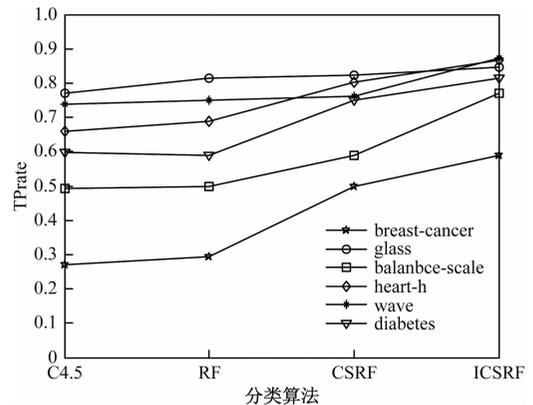


图 3 C4.5、RF、CSRF、ICSRF 的 TPrate 对比曲线

Fig. 3 TPrate contrast curves of C4.5, RF, CSRF, ICSRF

分类算法。其中 C4.5 分类器算法比较直观,实现简单,但是容易对数据造成过拟合,特别当特征较多时,训练的决策树较复杂,且没有考虑不平衡数据的特点,少数类性能很差。RF 分类算法相比决策树有一定的优势,但是没有对少数类采取策略,导致不平衡数据集上性能较低,且 RF 在小量数据集上并不能取得较好的效果,因此 AUC、F-measure 等性能与决策树并没有太大的提高。CSRF 可以较好地处理不平衡数据,在 AUC 性能上有了较大的提高,但是忽略了不同代价类型在分类过程中的重要性。ICSRF 构造代价函数时考虑样本实际与特征权值,避免了 CSRF 代价构造不准确的缺点。并且将不同分类能力的基分类器以 AUC 值赋予权重,有效避免了噪声数据的干扰。实验结果表明,重新构造代价函数、利用 AUC 值对树进行评价,对性能不同的树区别对待对不平衡数据分类是有效的。

从图 1 和图 2 中可以出 ICSRF 算法的性能高于其他分类算法,但是并不能直观地观察到 ICSRF 对少数类的分类性能,观察四种分类算法对少数类的正确率,可以明显观察出算法之间的差别,如图 3 所

示,其中 TPrate 表示实际为少数类被正确分为少数类的概率。

从图 3 中可以看出,相比其他三种分类算法,ICSRF 算法提高了少数类的正确率,其中在不平衡度较小的数据集中,ICSRF 算法在 TPrate 上没有太大的提升,而在不平衡度较大的 balance-scale 与 breast-cancer 数据集中,少数类的正确率得到更明显的提高,说明该算法可以更好地分类不平衡数据。AUC 反应分类器的整体性能,图 1 也可以看出 ICSRF 算法的 AUC 整体分类性得到了一定的提高,说明 ICSRF 算法在一定程度上较小牺牲多数类的正确率,但明显提高了少数类的准确率。

表 4 以 breast-cancer 数据集为例,从时间与不同类别的分类精度上展现了四种不同的分类模型对比。

表 4 breast-cancer 数据集实验对比

Table 4 Comparison results of breast-cancer data set

算法	时间/s	多数类精度/%	少数类精度/%
C4.5	0.09	96	27.1
RF	0.37	86.6	29.4
CSRF	0.43	83.4	49.1
ICSRF	0.49	81.6	59

从表 4 中可以明显观察到,决策树虽然建模时间最短,但是对少数类的分类性能最差。RF、CSRF、ICSRF 建模时间相差不大,多数类精度稍有下降,但是 CSRF、ICSRF 模型的少数类的分类准确率却得到了很大的提高,ICSRF 对少数类的性能最好。

4 结论

不平衡数据给传统机器学习算法带来了挑战,由于传统算法基于类别平衡假设,导致少数类的性能较差^[15]。提出一种改进的代价敏感随机森林算法(ICSRF),ICSRF 算法重新设计代价因子,对基分类器采取权重投票。避免代价函数构造不准确与基分类器性能不同对分类结果产生的影响。利用 AUC 衡量基分类器的性能,每个决策树的 AUC 值决定在投票阶段所拥有的权重。

提出的 ICSRF 算法与传统算法相比,具有更高的 AUC、F-measure、TPrate。说明本方法处理不平衡数据是有效的。

参 考 文 献

- Soda P. A multi-objective optimisation approach for class imbalance learning. *Pattern Recognition*, 2011; 44(8): 1801—1810
- 尹 华, 胡玉平. 一种代价敏感随机森林算法. *武汉大学学报工学版*, 2014; 47(5): 707—711
- Yin Hua, Hu Yuping. A cost-sensitive algorithm based on random

- forest. *Engineering Journal of Wuhan University*, 2014; 47(5): 707—711
- 杨杰明, 闫 欣, 曲朝阳, 等. 基于数据密度分布的欠采样方法研究. *计算机应用研究*, 2016; 33(10): 2997—3000
- Yang Jieming, Yan Xin, Qu Zhaoyang, *et al.* Under-sampling technique based on data density distribution. *Application Research of Computers*, 2016; 33(10): 2997—3000
- 韩 玉, 李美聪, 郭新辰. 基于粗糙集理论的文本分类属性约简算法. *东北电力大学学报*, 2016; 36(5): 92—96
- Han Yu, Li Meicong, Guo Xinchun. The text classification attribute reduction algorithm based on the rough set theory. *Journal of Northeast Dianli University*, 2016; 36(5): 92—96
- Sun Y, Kamel M S, Wong A K C, *et al.* Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 2007; 40(12): 3358—3378
- 孟光胜. 基于关联度和代价敏感学习的决策树生成法. *科学技术与工程*, 2013; 13(5): 1196—1199
- Meng Guangsheng. Decision tree based on correlation degree and cost-sensitive learning. *Science Technology and Engineering*, 2013; 13(5): 1196—1199
- 蒋盛益, 谢照青, 余 雯. 基于代价敏感的朴素贝叶斯不平衡数据分类研究. *计算机研究与发展*, 2011; 48(增刊 1): 387—390
- Jiang Shengyi, Xie Zhaoqing, Yu Wen. Navie bayes classification algorithm based on cost sensitive for imbalanced data distribution. *Journal of Computer Research Development*, 2011; 48(S1): 387—390
- Krawczyk B, Woźniak M, Schaefer G. Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing*, 2013; 14(1): 554—562
- 徐 鹏, 林 森. 基于 C4.5 决策树的流量分类方法. *软件学报*, 2009; 20(10): 2692—2704
- Xu Peng, Lin Sen. Internet traffic classification using C4.5 decision tree. *Journal of Software*, 2009; 20(10): 2692—2704
- 李 勇. 一种基于投票的不平衡数据分类集成算法. *科学技术与工程*, 2014; 14(21): 275—279
- Li Yong. An ensemble algorithms based voting for imbalanced data classification. *Science Technology and Engineering*, 2014; 14(21): 275—279
- 尹 华, 胡玉平. 基于随机森林的不平衡特征选择算法. *中山大学学报(自然科学版)*, 2014; 53(5): 59—65
- Yin Hua, Hu Yuping. An imbalanced feature selection algorithm based on random forest. *Acta Scientiarum Naturalium Universitatis Sunyatseni*, 2014; 53(5): 59—65
- Río S D, López V, Benítez J M, *et al.* On the use of MapReduce for imbalanced big data using random forest. *Information Sciences*, 2014; 285(3): 112—137
- 杨杰明, 乔媛媛, 王 林, 等. 基于流形排序的动态过抽样方法研究. *计算机应用研究*, 2017; 33(6): 1—6
- Yang Jieming, QiaoYuanyuan, Wang Lin, *et al.* Dynamic over sampling method based on manifold ranking. *Application Research of Computers*, 2017; 33(6): 1—6
- 叶志飞, 文益民, 吕宝粮. 不平衡分类问题研究综述. *智能系统学报*, 2009; 4(2): 148—156
- Ye Zhifei, Wen Yimin, Lu Baoliang. A survey of imbalanced pattern classification problems. *CAAI Transactions on Intelligent Systems*, 2009; 4(2): 148—156

15 Peng L, Zhang H, Yang B, *et al.* A new approach for imbalanced data classification based on data gravitation. Information Sciences,

2014; 288(9): 347—373

Random Forest Classification Algorithm Based on Cost-sensitive for Imbalanced Data

YANG Jie-ming¹, GAO Cong^{1*}, QU Zhao-yang¹, KAN Zhong-feng², GAO Ye², CHANG Cheng²

(School of Information Engineering, Northeast Electric Power University¹, Jilin 132012, China;

Jilin Electric Power Supply Company², Jilin 132000, China)

[**Abstract**] The random forest prefers to majority classes rather than minority classes on imbalanced data. The cost sensitive method can be combined with random forest to solve the imbalanced problem. But the traditional cost-sensitive algorithm based on random forest does not consider the actual distribution of data set and feature weight. And in the voting stages of random forest, it does not consider the performance differences of base classifiers. An improved cost-sensitive algorithm was proposed based on random forest ICSRF, which constructs a cost function based on the actual distribution of imbalanced data set and introduced the weight distance, then takes weighted voting according to the performance of the base classifier. It can improve the classification accuracy. The experiment results show that the ICSRF algorithm has higher accuracy rate and can effectively improve the recognition rate of the minority classes.

[**Key words**] cost-sensitive random forest imbalanced data weight distance