不平衡数据分类研究综述

赵 楠 张小芳 张利军
（西北工业大学计算机学院 西安 710000）

摘 要 在很多应用领域中，数据的类别分布不平衡，如何对其正确分类是数据挖掘和机器学习领域中的研究热点。经典的数据分类算法未考虑数据类别的不平衡性，认为类别之间的误分类代价相同，导致不平衡数据分类的效果不理想。针对数据分类的各个步骤，相继提出了不同的不平衡数据分类处理方法。对多年来的相关研究成果进行归类分析，从特征选择、数据分布调整、分类算法、分类结果评估等几个方面系统地介绍了相关方法，并探讨了进一步的探索方向。

关键词 不平衡数据分类，不平衡数据的特征选择，不平衡分类评估，数据分布调整，不平衡数据分类算法
中图法分类号 TP311 **文献标识码** A

Overview of Imbalanced Data Classification

ZHAO Nan ZHANG Xiao-fang ZHANG Li-jun
(School of Computer Science, Northwestern Polytechnical University, Xi'an 710000, China)

Abstract Imbalanced data classification has been drawn significant attention from research community in last decade. Because of the assumption of relatively balanced class distribution and equal misclassification costs, most standard classifiers do not perform well with imbalanced data classification. In view of various phases of data classification, different imbalanced data classification methods have been proposed. The relevant research achievements over the years were analyzed, and various approaches with imbalanced data were introduced from the view of feature selection, adjustment of the data distribution, classification algorithm and classifier evaluation. The future trends and research issues that still need to be faced in imbalanced data classification were discussed in the end.

Keywords Imbalanced data classification, Feature selection for imbalanced data, Imbalanced classification assessment, Adjustment of data distribution, Classification algorithm for imbalanced data

1 引言

分类是一种重要的数据分析技术。数据分类是一个两阶段的过程，包括学习阶段（构建分类器模型）和分类阶段（使用模型预测给定数据的类标号）[1]。其中，学习阶段具体分为 3 个基本过程：数据选择、特征选择、分类模型的构建。目前最常用的分类器有决策树分类器、Logistic 回归、贝叶斯分类器、支持向量机等，这些分类器在不同的分类领域均取得了较好的效果。

不平衡分类问题是指在一个分类问题中某些类的样本数量远多于其他类别的样本数量[2]。拥有大量样本数量的类被称为多数类，拥有少数样本数量的类被称为少数类。类别不平衡的数据普遍存在于现实生活的许多应用中。例如，用于疾病诊断预测的病历数据中，许多少见却非常重要的疾病样本数远小于正常或常见的疾病样本数；用于互联网入侵检测的样本数据中，正常的样本数远多于入侵的样本数。若将传统分类器应用于这些场景而不对类别的不平衡性做任何处理，就会使得多数类淹没少数类（往往是更重要的），得不到好的分类效果。因此，类别分布的不平衡问题是数据分类中很重要的一类问题。

研究人员针对类别不平衡数据的分类问题，相继提出了多个不同的解决办法，按照分类的步骤，可以将其分为特征选择、数据分布调整、模型训练算法几类。特征选择的目的是从全部特征中选择更适合于类别不平衡数据、能反映类别不平衡特点的子集来构建分类器模型，从而使得分类器在类别不平衡的前提下达到较好的性能。数据分布调整主要通过数据重采样或数据分组等手段使得类别在一定程度上达到平衡，从而消除类别不平衡问题。模型训练算法层面主要通过代价敏感学习、集成学习、关联分类等方法来解决类别不平衡问题。由于数据的不平衡性，传统的分类器评估指标并不能准确地评估分类器的性能，对此研究人员相继提出了多种适用于不平衡数据的分类器评估指标。图 1 给出了在分类的不同

本文受中央高校基本科研业务费专项资金（3102015JSJ0004），国家高技术研究发展计划（863）项目（2015AA015307），国家自然科学基金（61402370）资助。

赵 楠（1991－），女，硕士生，主要研究方向为数据挖掘；张小芳（1971－），女，博士，副教授，CCF 会员，主要研究方向为软件工程、数据库技术；张利军（1978－），男，博士，讲师，CCF 会员，主要研究方向为数据挖掘、分布式数据库，E-mail：zhanglijun@nwpu.edu.cn（通信作者）。

机系统应用，2015，24(8)：197-201.
[14] 朱桂英，张瑞林. 基于 Hough 变换的圆检测方法[J]. 计算机工程与设计，2008，29(6)：1462-1464.
[15] ALEGRIA F C，SERRA A C. Computer vision applied to the automatic calibration of measuring instruments[J]. Measurement，2000，28(3)：185-195.
[16] WANG Q，TANG X，DING C，et al. Automatic alignment system based on center point recognition of analog measuring instruments dial[C]// Conference of the IEEE Industrial Electronics Society，IECON 2013. New York：IEEE，2013：5532-5536.
[17] LI B，JIA Z. Some Results On Condition Numbers Of The Scaled Total Least Squares Problem[J]. Linear Algebra & Its Applications，2009，435(3)：674-686.
[18] 张远辉，张鼎，许昌，等. 指针式仪表总体最小二乘图像校验算法[J]. 自动化仪表，2015，36(5)：75-79.
[19] BRESENHAM J. A linear algorithm for incremental digital display of circular arcs[J]. Communications of the Acm，1977，20(2)：100-106.
[20] BELAN P A，ARAUJO S A，LIBRANTZ A F H. Segmentation-free approaches of computer vision for automatic calibration of digital and analog instruments[J]. Measurement，2013，46(1)：177-184.
[21] LIU J，LIU Y，YU L. Novel method of Automatic Recognition for Analog Measuring Instruments[C]// International Conference on Manufacturing Science and Engineering. 2015：67-74.
[22] YANG Z，NIU W，PENG X，et al. An image-based intelligent system for pointer instrument reading[C]// IEEE International Conference on Information Science and Technology. IEEE，2014：780-783.
[23] YUE X F，MIN Z，ZHOU X D，et al. The Research on Auto-recognition Method for Analogy Measuring Instruments[C]// International Conference on Computer，mechatronics，Control and Electronic Engineering. 2010：207- 210.
[24] GONZALEZ R C，WOODS R E. 数字图像处理(第三版)[M]. 阮秋琦，阮宇智，译. 北京：电子工业出版社，2011.
[25] SABLATNIG R，KROPATSCH W G. Automatic reading of analog display instruments[C]// Iapr International Conference on Pattern Recognition. New York：IEEE，1994：794-797.
[26] ALEGRIA E C，SERRA A C. Automatic calibration of analog and digital measuring instruments using computer vision[J]. IEEE Transactions on Instrumentation & Measurement，2000，49(1)：94-99.
[27] HEMMING B，LEHTO H. Calculation of uncertainty of measurement in machine vision case：a system for the calibration of dial indicators[C]// Proceedings of the 18th IEEE Instrumentation and Measurement Technology Conference. Budapest，Hungary：IEEE，2001：665-670.
[28] HEMMING B，FAGERLUND A，LASSILA A. High-accuracy automatic machine vision based calibration of micrometers[J]. Measurement Science & Technology，2007，18(18)：1655-1660.
[29] DATTA A，KIM J S，KANADE T. Accurate camera calibration using iterative refinement of control points[C]// IEEE，International Conference on Computer Vision Workshops. IEEE，2009：1201-1208.
[30] TSAI C Y，TSAO A H，WANG C W. Real-Time Feature Descriptor Matching via a Multi-Resolution Exhaustive Search Method[J]. Journal of Software，2013，8(9)：2197-2201.
[31] HOUGH P V C. Method and means for recognizing complex patterns：U. S. Patent 3069654[P]. 1962.
[32] 陶冰洁，韩佳乐，李恩. 一种实用的指针式仪表读数识别方法[J]. 光电工程，2011，38(4)：145-150.
[33] 颜友福，刘金清，吴庆祥. 基于区域生长的指针式仪表自动识别方法[J]. 计算机系统应用，2015，24(4)：164-170.
[34] 李祖贺，刘嘉，薛冰，等. 面向自动校验系统的指针式压力表读数识别[J]. 计算机工程与应用，2016，52(23)：213-219.
[35] 张春雪. 图像的边缘检测方法研究[D]. 江苏：江南大学，2011.

(上接第 27 页)

[33] 李雄飞，李军，董元方，等. 一种新的不平衡数据学习算法 PC-Boost[J]. 计算机学报，2012，35(2)：202-209.
[34] 袁兴梅，杨明，杨杨. 一种面向不平衡数据的结构化 SVM 集成分类器[J]. 模式识别与人工智能，2013，26(3)：315-320.
[35] ARUNASALAM B，CHAWLA S. CCCS：a top-down associative classifier for imbalanced class distribution[C]// 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2006：517-522.
[36] PATEL H，THAKUR G S. A Hybrid Weighted Nearest Neighbor Approach to Mine Imbalanced Data[C]// International Conference on Data Mining (DMIN). 2016：106.
[37] IMAM T，KAI M T，KAMRUZZAMAN J. z-SVM：An SVM for Improved Classification of Imbalanced Data[C]// Australasian Joint Conference on Artificial Intelligence. Springer Berlin Heidelberg，2006：264-273.
[38] KUBAT M，HOLTE R C，MATWIN S. Machine learning for the detection of oil spills in satellite radar images[J]. Machine Learning，1998，30(2)：195-215.
[39] BRADLEY A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms[M]. Elsevier Science Inc.，1997.
[40] FAWCETT T. An introduction to ROC analysis[J]. Pattern Recognition Letters，2006，27(8)：861-874.
[41] PROVOST F，DOMINGOS P. Tree induction for probability-based ranking[J]. Machine Learning，2003，52(3)：199-215.
[42] HAND D J，TILL R J. A simple generalisation of the area under the ROC curve for multiple class classification problems[J]. Machine Learning，2001，45(2)：171-186.
[43] DAVIS J，GOADRICH M. The relationship between Precision-Recall and ROC curves[C]// 23rd International Conference on Machine Learning. ACM，2006：233-240.
[44] DRUMMOND C，HOLTE R C. Cost curves：An improved method for visualizing classifier performance[J]. Machine learning，2006，65(1)：95-130.

阶段针对不平衡数据所采用的技术。

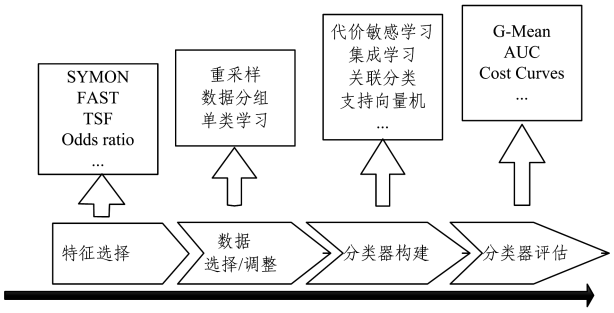

图 1　针对类别不平衡数据的分类技术

本文第 2 节介绍类别不平衡数据中的特征选择问题；第 3 节介绍如何通过调整数据分布来应对类别不平衡问题；第 4 节介绍如何在模型训练阶段处理类别不平衡问题，包括代价敏感学习、集成学习方法、关联分类方法；第 5 节针对类别不平衡问题的分类器进行性能评估；最后总结全文。

2　特征选择

特征选择是数据挖掘算法中一个非常关键的步骤，其目的是基于某种规则选择 j 个特征的子集，使得分类器达到最优的性能，其中，j 是用户自定义的一个参数。由于采样技术和算法层面的方法不足以解决高维的类不平衡问题[5]，而数据集的类别不平衡问题通常都出现在高维数据集上[2]，因此特征选择极为重要。Van 等通过分析 CoIL Challenge 2000 数据集发现，特征选择比分类算法的选择对性能的提升更重要，而且可以最大程度地避免过拟合问题[6]。Forman 在高度不平衡的文本分类问题的研究中发现，在高维数据集中仅仅使用特征选择就可以很大程度地解决数据集中的类别不平衡问题[4]。

但是，Elkan 在研究 CoIL Challenge 2000 数据时发现，特征选择并没有给他的工作带来足够多的好处[7]，相反，特征之间的相互关系应该被考虑在内。他发现，特征选择方法的最大漏洞在于：由于某些特征是多余的，因此没有考虑特征之间高度的相关性。Guyon 等通过很强的理论分析说明了特征选择的局限性[8]。

Moayedikia 等[9]提出了一种利用和声搜索（Harmony-Search）算法解决高维不平衡数据的特征选择方法 SYMON。SYMON 算法使用和声搜索算法将特征选择过程转换成选择最好的特征组合的最优解问题，并且使用对称的不确定加权特征来使其依赖于类标签，从而加强在少数类中识别最频繁的有效特征。经典的特征选择方法都是基于损失函数，即特征对于分类率的贡献大小，但是 Moayedikia 认为基于损失函数的特征选择并非总是最好的，特别是对少数类的预测，特征排序依赖于类标签可以选择到更好的特征子集。同时，SYMON 利用向量调谐操作避免了传统特征选择方法中对相同排名的特征难以选择的问题。在微阵列数据上对多个特征选择方法进行了实验对比，结果表明 SYMON 比当前的基准方法获得了更好的结果。但是，SYMON 算法的计算时间较长，而且特征选择受预设子集大小的限制。

Chen 等[3]提出了一种新的特征选择方法 FAST（Feature Assessment by Sliding Thresholds）。FAST 方法的基本原理是基于 ROC 曲线的面积，对每一个特征训练一个简单的线性分类器，并通过滑动决策边界来得到最优的分类器。大多数简单的特征选择分类器把决策边界设置为两类的中点，但这不一定是决策边界最好的选择。滑动的决策边界可以通过分类更多的 FP（false positive）来增加 TP（true positive）的数量（一般把少数类样本当作正类，把多数类样本当作负类）。相反，可以滑动阈值来减少 TP 的数量，从而避免误分到负类。Chen 提出首先计算出每一个阈值下的真阳性率（true positive rate）和假阳性率（false positive rate），并且构建 ROC 曲线；然后计算 ROC 曲线下的面积 AUC，AUC 的大小表征了特征的预测能力；最后根据 AUC 对特征排序，并进行特征选择。在 5 个不同类别的不平衡数据集上与相关系数方法以及 RELIFT 方法的对比实验表明，FAST 表现最好，尤其在选择的特征数比较少的情况下其优势更明显。

王杰等[10]提出了一种面向非平衡文本情感分类的双边 fisher 特征选择算法 TSF，该算法通过显式地组合正相关和负相关特征，缓解了特征层面的非平衡性。TSF 算法仅仅通过了文本分类的可行性研究，并没有通用性的理论或实验研究，因此其是否具有通用性还有待进一步研究。

Mladenic 等[11]比较研究了 11 种不同的特征选择方法在文本分类中的性能。文本数据有大量的特征，并且有极高的类不平衡性以及不对称的误分代价。不对称的误分代价是指，对于目标类正类（positive）来说，FN（false negative）的代价比 FP（false positive）更大。在雅虎数据集上，使用 11 种不同的方法进行特征选择，然后用朴素贝叶斯进行分类。实验结果表明，11 种方法中，Odds ratio 方法表现出了最好的性能。

针对高维、类别分布不平衡的文本数据集上的 SVM 和二分类问题，Forman 等人研究了 12 种不同的特征选择度量[4]。研究结果表明，在多数情况下，Bi-Normal Separation（BNS）优于其他特征选择度量。BNS 被定义为：

$$BNS(W)=|F^{-1}(tpr)-F^{-1}(fpr)| \tag{1}$$

其中，F^{-1} 是标准正态分布的累积概率分布函数的反函数，tpr 指包含 W 的正类样本数占正类样本总数的比率，fpr 指包含 W 的负类样本数占负类样本总数的比率。

以上研究中，BNS，TSF 和 Odds ratio 在不平衡的文本分类中表现出了良好的性能，但在医疗检测、风险识别等其他场景中是否仍有较好的表现则有待进一步验证。

3　数据分布调整

针对类别不平衡问题，数据分布调整策略是指从数据准备阶段入手，通过对原始不平衡训练数据进行调整，使得不平衡的数据在一定程度上达到平衡状态，从而消除类别不平衡问题。常用的数据分布调整方法包括采样技术、数据分组和单类学习方法。采样是通过对少数类过采样或对多数类欠采样来实现的；数据分组方法是通过把训练集分为多个平衡的子训练集，然后采用集成学习来实现对数据分布的调整；单类学习是仅仅考虑某一类别的数据来训练模型。

3.1　重采样技术

欠采样中最重要的方法就是随机下采样方法，它通过随机地移除多类样本来平衡数据集中的类别分布。该方法的主要缺点是可能会丢失有价值的信息。同样地，随机过采样技术是过采样中最重要的技术，通过复制少数类的样本达到平衡类别分布的目的。这种方法会使数据出现冗余，模型训练复杂度增大，而且容易造成模型过拟合问题。

Chawla 等[12]提出了一种通过创造合成的少数类样本来实现对少数类的过采样的方法，称之为 SMOTE(Synthetic Minority Over-sampling Technique)方法，它是对随机过采样技术的改进，可以在一定程度上避免模型过度拟合的问题。SMOTE 的主要思想是对于每一个少数类样本，通过沿线性加入 k 邻近的少数类样本来引入合成样本，具体实现方法是找到目标样本和其邻近样本之间的特征向量的差异，用 0-1 之间的随机数乘以这个差异，并将结果累加到目标样本的特征向量上。根据过采样的数量需求，k 邻近样本是随机选择的。

Chawla 等[13]提出了一种 SMOTEBoost 方法，该方法把 SMOTE 方法和 Boosting 过程相结合，在每一个 Boosting 过程中运用 SMOTE 方法来合成少数类样本，使得每一个子模型都可以从更多少数类样本中学习，以提高精确度。实验结果表明，SMOTEBoost 比 SMOTE 的性能更好。但是与传统的 Boosting 不同，SMOTEBoost 是从少数类样本中创造合成样本，从而间接地改变更新权重并且补偿倾斜的数据分布。

熊冰妍等[14]提出了一种基于样本权重的欠采样方法 KAcBag(K-means AdaCost Bagging)，该算法以 Bagging 算法为框架，以 AdaCost 权重更新方法为基础，使用 K-means 算法对数据集进行多次聚类，并根据聚类结果更新样本权重，旨在找出位于多数类中心区域的样本；然后通过样本权重对多数类进行欠采样，并与少数类样本组成多个平衡数据集，在多个平衡数据集上应用决策树算法得到多个弱分类器；最后通过弱分类器加权投票来生成最终的分类器。对 UCI 上的 19 组数据和某电信运营客户交换机数据进行实验，结果表明，KAcBag 算法所抽取的样本具有较强的代表性，在一定程度上解决了类别不平衡问题；但算法中 K-means 中的参数 K 的取值对实验影响较大，如何根据数据分布自适应地确定 K 值仍是一个问题。

Kubat 等[15]提出了一种样本的单边选择算法来寻找样本的一致子集，从而实现欠采样。他们使用 Tomek Links 来识别噪音和边缘样本，并且使用 CNN(Condensed Nearest Neighbor)[16]从多数样本中删除那些远离决策边界的实例。

Laurikkala 等[17]提出了 NCR(Neighborhood Cleaning Rule)方法来移除多类样本，从而实现欠采样。NCR 方法计算训练集中每一个样本 E_i 的 3 个最近邻居，如果 E_i 属于多数类，并且被这 3 个邻近样本误分，那么在这 3 个邻近样本中间的这个多数类样本将被移除。

胡小生等[18]提出了一种欠采样与过采样相结合的方法，其目标是在过采样与欠采样之间寻找一个平衡点，使得数据分布达到平衡。具体过程为：使用 SMOTE 方法对少数类进行过采样，并且使用 K-means 对多数类进行聚类，从而得到与少数类数目一致的 K 个多类聚类质心；然后将 K 个聚类质心与所有少数类样本组合，从而形成一个平衡的训练集；最后，在平衡的数据集上使用 K 均值聚类方法把训练集分为与类别数目一致的 K 个簇，并在每个簇上使用 C4.5 算法来构建决策树，通过一序列的基于特征空间上的规则来提炼决策边界。

李克文等[19]提出了一种将采样技术与 Boosting 算法相结合的分类算法 PSBoost，该算法首先应用 SMOTE 算法对少数类实现过采样，然后在不改变数据分布的情况下对所有数据随机欠采样，最后再与 AdaBoot 算法相结合完成数据分类。算法通过对少数类过采样来平衡数据集，再通过整体欠采样来缩小数据集规模，从而减少了模型训练时间。在 6 个不平衡的数据集上与 C4.5、SMOTE 算法、SMOTE＋Boost 算法进行的对比实验表明，PSBoost 算法表现出了优良的性能以及极快的运行速度。

3.2 数据分组

数据分组的主要思想是按一定的规则将不平衡的训练数据集划分成多个平衡数据集，并将在平衡数据集上训练得到的多个分类器按一定的学习方法集成在一起，以此来消除类别不平衡问题。

Chan 等[20]根据少数类的数目把多数类划分成多个子数据集，并将多个多数类的子数据集分别与少数类组合，从而形成多个平衡的子数据集；然后在每一个子数据集上独立地应用学习算法来学习子分类器；最后通过 meta-learning 方法集成子分类器，从而形成最终的分类器。在信用卡使用数据集上的实验表明，该方法可以有效地检测信用卡欺诈使用模式。

Sun 等[21]把不平衡问题转换成多个平衡的学习过程，提出了一种新的基于数据集分组的方法。首先，使用数据平衡方法随机分组或聚类，把原始的不平衡数据集分成多个平衡的数据集；然后，这些分组数据集上构造多个分类器；最后，通过特定规则把多个分类器集成为一个最终分类器。Sun 等基于 Kittler[22]提出的 5 种集成规则，并考虑了新数据与训练数据之间的相关性，提出了 5 种新的集成方法。在 46 个高度不平衡数据集上对朴素贝叶斯、C4.5，PIPPER 等 6 个分类算法进行实验研究，得出了两种最优组合方式，即聚类分组＋最大距离和随机分组＋最大距离。这种数据集再分组方法避免了过采样与欠采样中的有用数据缺失问题或过拟合问题，但是在数据分组过程中，如果使用聚类分组方法，可能会产生新的不平衡数据分组，影响分类结果。

3.3 单类学习

单类学习是指仅使用一个类别的样本来进行学习模型的分类方法。Schölkopf 等[23]提出了单类支持向量机(One-Class Support Vector Machine，OCSVM)，即把数据映射到特征空间，并把原点作为异常点，将寻找到的隔离原点与训练样本的超平面作为决策边界。对于一个新样本，通过其在超平面的边来进行分类决策。后续有很多人用单类学习来解决不平衡数据问题[24-25]。

Cohen 等[24]通过核函数的共形变换，即共行核函数，来提高支持向量机区域边界的分辨率，从而提升单类支持向量机的精确度。共形变换是保持两曲线间夹角和大小不变的变换。他们仅使用未被感染的样本作为训练集，将该方法应用于医院感染的检测过程中，其中被感染样本仅占 11%。实验表明，与传统高斯核函数的单类支持向量机分类器相比，少数类的正确分类数量有一定的提高；与二类支持向量机分类器相比，少数类的识别率明显提高，但是许多未被感染的样例会被误分到已感染中。

Manevitz 等[25]把 Schölkopf 等[23]提出的单类支持向量机算法应用于文本分类中，并且通过改进该算法提出了 outlier-SVM算法，其不仅把原点作为异常点，而且把离原点足够近的点也作为异常点处理。实验结果表明，Schölkopf 等提出的单类支持向量机对参数比较敏感，但是，如果选择恰当，则会产生最好的分类结果。如果仅考虑少数类，则 Schölkopf 算法较优；如果使用所有类别的宏平均，则 outlier-SVM 较优。

由于单类学习只考虑一个类别的样本数据，可能会丢失

大量的有用信息，因此，除非数据类分布极度不平衡，否则不建议使用单类学习。

重采样技术、数据分组、单类学习方法都是通过改变原始数据集的分布来解决分类不平衡问题。重采样技术是解决不平衡问题最简单、最朴素的方法，但易造成重要数据丢失或过拟合。数据分组方法往往需要与集成学习一起使用，复杂度比重采样高，但在一定程度上降低了数据缺失和过拟合的概率。单类学习仅仅考虑某一个类别的样本数据来解决不平衡问题，可能会丢失大量的有用信息，多用于数据极度不平衡的情况。

4　模型训练算法

除了通过特征选择、数据分布调整来降低类别不平衡对分类算法的影响，还可以直接在算法层面，通过设计适用于不平衡数据特征的模型训练算法来解决类别不平衡问题。这方面的研究工作主要有代价敏感学习集成学习以及其他算法，如关联分类算法、K 近邻算法、支持向量机算法的改进等等。

4.1　代价敏感学习

大多数分类算法的设计目标是使得 0-1 损失或错误率最小。这就意味着，大多数的分类算法基于一个假设：所有类别的误分类代价相等。但在类别分布不平衡的数据中，多数类误分为少数类和少数类误分为多数类的代价往往不同。基于这一前提，代价敏感学习通过为不同的类别误分赋予不同的代价（一般少数类误分为多数类的代价高于多数类误分为少数类的代价）来构造分类器。

Elkan 等[26]提出了在不同的误分导致不同的代价情况下的最优学习和决策的基本概念；同时指出，首先给定正确分类和错误分类的代价，并且一个样本应该被预测到拥有最低期望代价的类别中。期望代价可用样本与每一类别的条件概率来计算。设 $C(i,j)$表示实际分类为 j、预测类为 i 时的代价，x 表示一个实例。若 $j=i$，则表示预测正确，否则预测错误。使得 $L(x,i)=\sum_j P(j|x)C(i,j)$最小的类 i，即是实例 x 的最优预测。

Domingos 等[27]提出了一种代价敏感学习方法 MetaCost，该方法可以将原有任意分类算法转换为代价敏感的。MetaCost 首先在训练集中多次取样生成多个模型，得到训练集中每条记录属于各类别的概率；然后计算训练集中每条记录属于每个类别的代价，并修改其类标签为最小代价类；最后在修改过的训练集上学习分类器。实验结果表明，与不使用代价敏感的分类器（cost-blind classifier）相比，该方法可以大幅降低误分代价，且具有良好的可伸缩性。

在代价敏感学习中确定代价时需要足够的先验知识，很难准确设置。对于此问题，蒋盛益等[28]构造了针对不平衡数据分布的自适应代价函数，引进全局代价矩阵，对传统的朴素贝叶斯分类算法进行改进，得到了基于代价敏感的朴素贝叶斯不平衡数据分类算法。实验结果表明，该方法对不平衡数据分类有效且可行。

除了误分代价外，应考虑其他代价。针对分类器分类过程中测试数据可能存在缺失值（missing values）的情况，Chai 等[29]综合考虑了误分代价和测试代价，在测试数据中若存在缺失值，则通过一定的测试策略决定如何选择未知属性，并以最小化误分代价和测试代价之和为目标构建朴素贝叶斯分类器（test-cost sensitive Naive Bayes，csNB）。文中提出了两种测试策略，即顺序测试策略和批量测试策略，不同的策略可能会导致不同的决策。

代价敏感学习通过将样本分类到其期望代价最小的类别中，在一定程度上解决了数据集中类别不平衡的问题，但对代价的研究主要集中于误分代价，对其他代价的研究工作较少。

4.2　集成学习方法

集成学习的主要思想是将多个分类器组合成一个分类器，以提高分类性能。其中，Boosting 可将多个弱分类器组合成一个强分类器，是被广泛使用的集成学习方法。由于 Boosting 简单有效，不少学者用它来处理不平衡数据集的分类问题。

AdaBoost 是采用 Boosting 方法的典型代表，其实质是改变数据分布，根据每次训练集中每个样本分类是否正确以及总体分类的准确率来确定每个样本的权重，并且将更新过权重的新数据发送给下层分类器进行训练，然后把每次训练得到的分类器融合起来作为最终的分类器。Fan 等[30]将 AdaBoost 和代价敏感学习结合起来，提出了 AdaCost 算法。在 Boosting 环节，使用误分代价更新训练集的分布，以达到比 AdaBoost 更少的累积误分代价。与 AdaBoost 最大的不同在于，AdaCost 在权重更新规则中增加了代价调节函数。实验表明，AdaCost 在没有消耗额外计算资源的前提下极大地减少了累积误分代价。

多数关于不平衡数据分类的研究工作都关注二分类问题。Sun 等[31]提出了一种针对多类别的不平衡数据分类方法 AdaC2.M1。该方法将 AdaBoost 与代价敏感学习相结合，并将二分类问题扩展到多类情况。针对多类场景下代价矩阵不便给出的问题，引入遗传算法，用于为每个类搜索最优误分类代价。在 3 个数据集上的实验表明，AdaC2.M1 可显著提高多类别不平衡数据的分类性能。

不少学者采用不同的采样方法进行 Boosting 学习。Chawla 等[13]把 SMOTE 和 Boosting 过程结合起来，提出了 SMOTEBoost 方法。Boosting 算法在迭代过程中对所有误分样本赋予相同的权重，而 SMOTEBoost 算法通过从少数类中创造人工合成样本，间接改变样本的权重，对类别的倾斜分布进行补偿。。文献[32]采用基于权重采样的 Boosting 算法，通过对样本进行权重采样来改变原有数据的分布。其利用采样函数来调整原始 Boosting 损失函数，进一步强调少数类样本的分类损失，使得分类器侧重于有效判别少数类样本，提高了少数类样本的识别率。文献[34]中提出了 3 种不同的采样函数。李雄飞等将过采样技术与 Boosting 相结合，得到 PCBoost 方法，即在每一次迭代中增加合成的少数类样本，并且及时删除被误分的合成样本，防止产生不合适的样本而影响算法性能，再以决策树算法训练出多个弱分类器，将多个弱分类器集成为最终的分类器。PCBoost 不仅拓宽了少数类边界，而且很大程度地避免了噪声合成样本的产生[33]。

袁兴梅等[34]提出了一种基于代价敏感的结构化支持向量机集成分类算法 AdaStASVM，该算法对训练集应用了聚类算法，得到隐含在数据中的结构信息，初始化样本权重，然后通过 AdaBoost 框架对样本权重进行动态调整，适当增大少数类样本权重，使小类的误分代价增大，进而降低类别不平衡带来的影响。

集成学习方法因能极大地提高分类器性能而被广泛应用，通常与采样、代价敏感、最优解方法等结合使用来提高类

别不平衡数据的分类性能。

4.3 其他分类算法

关联分类是一种重要的分类算法，但通常使用支持度-置信度度量挖掘频繁项集，不适用于处理类别不平衡数据。为此，Arunasalam 等使用补类支持度 CCS(Complement Class Support)替代传统关联分类中的支持度-置信度度量，并将 CCS 与自顶向下的行枚举算法相结合，得到适用于类别不平衡数据的分类器 CCCS(Classification by using the measure Complement Class Support)[35]。CCCS 保证了规则的前件与后件的正向相关性，从而保证了少数类样本生成的规则在分类器中能够被有效地显示出来。而且，CCS 的反单调等特性使得其特别适合于不平衡数据。CCCS 的提出给解决类别不平衡数据的分类问题提供了新思路。

Patel 等[36]提出了一种混合加权 K 近邻算法来对不平衡数据进行分类，算法的主要思想是给不同的类设置不同的 K 值和不同的权重，以平衡类别的倾斜。鉴于在不平衡数据中，传统的 K 近邻算法中统一的 K 使得大量的最近邻多为多数类，从而使得少数类很容易被误分为多数类，Patel 等根据类别的大小动态设置 K 值，同时给少数类分配较大的权值，给多数类分配较小的权值，以增加少数类的分类正确性并降低分类器对多数类的偏斜。

Imam 等[37]提出了一种针对不平衡数据的改进支持向量机算法 zSVM。该算法开始使用原始不平衡的数据集来训练 SVM 模型，然后通过去除向多数类的倾斜来修改模型，即通过给正支持向量的系数乘以一个小的正值来修改 SVM 决策函数。正支持向量系数值的修改可以使得决策函数中正支持向量的权重增大，从而减少模型向多数类的倾斜。

5 性能评估

针对分类器的性能评估，最常用的指标有 Accuracy(精度)和 Error Rate(错误率)，但二者均不适用于类别不平衡的场景。如果假设测试样本集中多数类样本占 95%，少数类样本占 5%，某分类器把所有样本都预测为多数类，那么模型的精度即高达 95%，很明显这样的模型是毫无意义的。

在类别不平衡数据分类中，常采用 Precision(准确率/查准率)、Recall(召回率/查全率)以及 F-measure(通常使用 F1)来评估分类器性能。另外，P-R 曲线也常被用来评估分类器的分类性能。P-R 曲线是以 Precision 为纵轴、Recall 为横轴绘制的曲线。

G-Mean 也常被用于不平衡分类的性能评估，当样本分布可能随着时间改变或者训练集和测试集样本分布而不同时，G-Mean 具有很好的鲁棒性[38]。

Precision，Recall，F-Measure 以及 G-Mean 常用于二分类场景，在多分类情况下，往往需要对多个类别上的指标求平均值，这时应该采用宏平均而不是微平均，以避免更小的类被较大的类淹没。

受试者工作特征(Receiver Operating Characteristic，ROC)和 ROC 曲线下面积(Area Under ROC Curve，AUC)也是常用的评估分类器性能的方法[39-40]。ROC 曲线是以真正例率(True Positive Rate，TPR)为纵轴、假正例率(False Positive Rate，FPR)为横轴绘制的曲线。AUC 为 ROC 图中曲线下面覆盖的面积，表达的物理含义是：随机选取正负样本各一个，设分类器预测正样本为正例的概率为 p_1，预测负样本为正例的概率为 p_0，AUC 即为 $p_1>p_0$ 的概率。AUC 对类别是否平衡不敏感，因此可用来对类别不平衡数据分类进行评估。

ROC 和 AUC 也主要用于二分类问题。Fawcett[40]探讨了将 ROC 拓展到多类别场景下的问题，最直接的方法是对每一个类生成一个 ROC 曲线图，把该类作为正类，其他类作为负类。但这种方式会使得使用 ROC 分析不平衡问题的优势减弱，因为它的负类变成了 $n-1$ 个其他类的组合。Provost 等使用加权求和的方式将 AUC 从二分类扩展到多分类场景[41]，这种方法计算简单，但对类别分布和错误代价非常敏感[40]。Hand 等基于 AUC 本身的特性，定义了新的被称为 M 的多类别评估度量[42]，Hand 方法与 Provost 方法的本质区别在于，针对类别集合中的任一类别 C_i，Provost 的方法将 C_i 作为正类，其他所有类作为负类，从而得到 AUC，然后根据类别概率进行加权聚合；而 Hand 的方法将 C_i 作为正类，选取另外一个类别 C_j 作为负类，逐类别地计算 AUC，然后进行聚合，避免了前者的缺点[40]。

Davis 等[43]对 ROC 曲线和 P-R 做了比较，指出当数据集高度不平衡时，ROC 曲线对分类器的性能评估会过分乐观，而 P-R 曲线可以提供更多的关于分类器性能的信息。

以上评估指标均假设类别误分的代价相同，而在很多现实场景中类别误分代价往往不同，在代价不对等的前提下，ROC 曲线不能反映分类器的期望总体代价，为此，Drummond 等[44]提出了一种代价敏感的评估方法——代价曲线(Cost Curves)。代价曲线是以标准化的期望代价为纵轴、以正例概率代价为横轴绘制的曲线。与代价敏感学习类似，代价曲线的绘制需要首先给出类别误分代价。

在类别不平衡度不是很大的情况下，F-measure 和 G-means 作为分类器评测指标均可取得较好的效果，但是当训练集中类别不平衡度增大时，F-measure 的性能更好。ROC 是通过假正率和真正率综合判断分类器的性能，不受类分布的影响，在类别不平衡分类器的评估中应用广泛；但是当数据集高度不平衡时，ROC 曲线会使得对分类器性能的评估过分乐观，这时 P-R 曲线的表现更好。代价曲线弥补了 ROC 曲线不能提供分类器性能的置信区间的缺点，可以提供不同分类器性能的统计意义。

结束语 本文从分类过程的几个不同层面对类别不平衡数据的分类问题做了综述，包括特征选择、数据分布调整、模型训练算法和分类器性能评估。

特征选择面向高维不平衡数据，选择特定的特征子集用于分类，主要方法有基于和声搜索方法的 SYMON、基于 ROC 曲线的 FAST、基于双边 Fisher 的 TSF 以及 BNS 和 OddsRatio 等。有研究人员针对不同的特征选择度量在类别不平衡条件下的表现做了对比。利用特征选择方法解决类别不平衡问题局限于数据维度高的情况，若数据维度空间较小，则特征选择方法对类别不平衡问题的影响有限。

数据分布调整是最直接的解决类别不平衡问题的方法，将不平衡数据转换为平衡数据，以消除类别不平衡带来的影响。转换的主要方法有重采样、数据分组和单类学习。数据分布调整改变了原始数据的分布，有可能会导致有价值信息的丢失或过拟合问题。

模型训练算法直接在构建分类器模型时考虑类别不平衡性，构造适用于类别不平衡数据的分类器模型，主要方法有代

机系统应用，2015，24(8)：197-201.
[14] 朱桂英，张瑞林. 基于 Hough 变换的圆检测方法[J]. 计算机工程与设计，2008，29(6)：1462-1464.
[15] ALEGRIA F C，SERRA A C. Computer vision applied to the automatic calibration of measuring instruments[J]. Measurement，2000，28(3)：185-195.
[16] WANG Q，TANG X，DING C，et al. Automatic alignment system based on center point recognition of analog measuring instruments dial[C]// Conference of the IEEE Industrial Electronics Society，IECON 2013. New York：IEEE，2013：5532-5536.
[17] LI B，JIA Z. Some Results On Condition Numbers Of The Scaled Total Least Squares Problem[J]. Linear Algebra & Its Applications，2009，435(3)：674-686.
[18] 张远辉，张鼎，许昌，等. 指针式仪表总体最小二乘图像校验算法[J]. 自动化仪表，2015，36(5)：75-79.
[19] BRESENHAM J. A linear algorithm for incremental digital display of circular arcs[J]. Communications of the Acm，1977，20(2)：100-106.
[20] BELAN P A，ARAUJO S A，LIBRANTZ A F H. Segmentation-free approaches of computer vision for automatic calibration of digital and analog instruments[J]. Measurement，2013，46(1)：177-184.
[21] LIU J，LIU Y，YU L. Novel method of Automatic Recognition for Analog Measuring Instruments[C]// International Conference on Manufacturing Science and Engineering. 2015：67-74.
[22] YANG Z，NIU W，PENG X，et al. An image-based intelligent system for pointer instrument reading[C]// IEEE International Conference on Information Science and Technology. IEEE，2014：780-783.
[23] YUE X F，MIN Z，ZHOU X D，et al. The Research on Auto-recognition Method for Analogy Measuring Instruments[C]// International Conference on Computer，mechatronics，Control and Electronic Engineering. 2010：207- 210.
[24] GONZALEZ R C，WOODS R E. 数字图像处理(第三版)[M]. 阮秋琦，阮宇智，译. 北京：电子工业出版社，2011.
[25] SABLATNIG R，KROPATSCH W G. Automatic reading of analog display instruments[C]// Iapr International Conference on Pattern Recognition. New York：IEEE，1994：794-797.
[26] ALEGRIA E C，SERRA A C. Automatic calibration of analog and digital measuring instruments using computer vision[J]. IEEE Transactions on Instrumentation & Measurement，2000，49(1)：94-99.
[27] HEMMING B，LEHTO H. Calculation of uncertainty of measurement in machine vision case：a system for the calibration of dial indicators[C]// Proceedings of the 18th IEEE Instrumentation and Measurement Technology Conference. Budapest，Hungary：IEEE，2001：665-670.
[28] HEMMING B，FAGERLUND A，LASSILA A. High-accuracy automatic machine vision based calibration of micrometers[J]. Measurement Science & Technology，2007，18(18)：1655-1660.
[29] DATTA A，KIM J S，KANADE T. Accurate camera calibration using iterative refinement of control points[C]// IEEE，International Conference on Computer Vision Workshops. IEEE，2009：1201-1208.
[30] TSAI C Y，TSAO A H，WANG C W. Real-Time Feature Descriptor Matching via a Multi-Resolution Exhaustive Search Method[J]. Journal of Software，2013，8(9)：2197-2201.
[31] HOUGH P V C. Method and means for recognizing complex patterns：U. S. Patent 3069654[P]. 1962.
[32] 陶冰洁，韩佳乐，李恩. 一种实用的指针式仪表读数识别方法[J]. 光电工程，2011，38(4)：145-150.
[33] 颜友福，刘金清，吴庆祥. 基于区域生长的指针式仪表自动识别方法[J]. 计算机系统应用，2015，24(4)：164-170.
[34] 李祖贺，刘嘉，薛冰，等. 面向自动校验系统的指针式压力表读数识别[J]. 计算机工程与应用，2016，52(23)：213-219.
[35] 张春雪. 图像的边缘检测方法研究[D]. 江苏：江南大学，2011.

(上接第 27 页)

[33] 李雄飞，李军，董元方，等. 一种新的不平衡数据学习算法 PC-Boost[J]. 计算机学报，2012，35(2)：202-209.
[34] 袁兴梅，杨明，杨杨. 一种面向不平衡数据的结构化 SVM 集成分类器[J]. 模式识别与人工智能，2013，26(3)：315-320.
[35] ARUNASALAM B，CHAWLA S. CCCS：a top-down associative classifier for imbalanced class distribution[C]// 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2006：517-522.
[36] PATEL H，THAKUR G S. A Hybrid Weighted Nearest Neighbor Approach to Mine Imbalanced Data[C]// International Conference on Data Mining (DMIN). 2016：106.
[37] IMAM T，KAI M T，KAMRUZZAMAN J. z-SVM：An SVM for Improved Classification of Imbalanced Data[C]// Australasian Joint Conference on Artificial Intelligence. Springer Berlin Heidelberg，2006：264-273.
[38] KUBAT M，HOLTE R C，MATWIN S. Machine learning for the detection of oil spills in satellite radar images[J]. Machine Learning，1998，30(2)：195-215.
[39] BRADLEY A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms[M]. Elsevier Science Inc.，1997.
[40] FAWCETT T. An introduction to ROC analysis[J]. Pattern Recognition Letters，2006，27(8)：861-874.
[41] PROVOST F，DOMINGOS P. Tree induction for probability-based ranking[J]. Machine Learning，2003，52(3)：199-215.
[42] HAND D J，TILL R J. A simple generalisation of the area under the ROC curve for multiple class classification problems[J]. Machine Learning，2001，45(2)：171-186.
[43] DAVIS J，GOADRICH M. The relationship between Precision-Recall and ROC curves[C]// 23rd International Conference on Machine Learning. ACM，2006：233-240.
[44] DRUMMOND C，HOLTE R C. Cost curves：An improved method for visualizing classifier performance[J]. Machine learning，2006，65(1)：95-130.

价敏感学习、集成学习以及其他一些针对不平稳问题进行改进的经典分类算法（如关联分类、SVM、KNN 等）针对不平衡问题的改进算法。需要说明的是，这些方法不是孤立的，很多情况下可融合在一起使用，如集成学习往往与代价敏感学习、重采样等方法一起使用，以提高不平衡条件下的分类性能。

准确率和错误率不适用于类别不平衡数据分类器的评估。在类别不平衡条件下，常用的针对分类器的性能评估指标有 Precision/Recall、F-Measure、P-R 曲线、G-Mean、ROC/AUC、代价曲线等，不同的评估指标各有优缺点，也有研究人员对这些指标做了对比研究。

目前针对类别不平衡问题的研究很多，但大多数研究工作均针对二分类问题。而现实场景中很多问题是多分类问题。虽然有个别研究工作针对多分类问题做了探讨，但对不平衡多分类问题的研究仍然不足，还需要进一步深入研究。

参 考 文 献

[1] HAN J，PEI J，KAMBER M. Data mining：concepts and techniques[M]. Elsevier，2011：162-164.

[2] CHAWLA N，JAPKOWICZ N，KOTCZ A，et al. Special Issue on Learning from Imbalanced Data Sets [J]. ACM SIGKDD Explorations Newsletter，2004，6(1)：1-6.

[3] CHEN X，WASIKOWSKI M. Fast：a roc-based feature selection metric for small samples and imbalanced data classification problems[C]// 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2008：124-132.

[4] FORMAN G. An extensive empirical study of feature selection metrics for text classification[J]. Journal of machine learning research，2003，3(2)：1289-1305.

[5] MEMBER M W，CHEN X W. Combating the Small Sample Class Imbalance Problem Using Feature Selection[J]. IEEE Transactions on Knowledge and Data Engineering，2010，22(10)：1388-1400.

[6] VAN D P P，VAN S M. A bias-variance analysis of a real world learning problem：The CoIL challenge 2000[J]. Machine Learning，2004，57(1)：177-195.

[7] ELKAN C. Magical thinking in data mining：lessons from CoIL challenge 2000[C]// Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2001：426-431.

[8] GUYON I，ELISSEEFF A. An introduction to variable and feature selection[J]. Journal of Machine Learning Research，2003，3(6)：1157-1182.

[9] MOAYEDIKIA A，ONG K L，BOO Y L，et al. Feature selection for high dimensional imbalanced class data using harmony search [J]. Engineering Applications of Artificial Intelligence，2017，57(C)：38-49.

[10] 王杰，李德玉，王素格. 面向非平衡文本情感分类的 TSF 特征选择方法[J]. 计算机科学，2016，43(10)：206-210，224.

[11] MLADENIC D，GROBELNIK M. Feature selection for unbalanced class distribution and naive bayes[C]// ICML. 1999：258-267.

[12] 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.

[13] CHAWLA N V，LAZAREVIC A，HALL L O，et al. SMOTEBoost：Improving prediction of the minority class in boosting[C]// European Conference on Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg，2003：107-119.

[14] 熊冰妍，王国胤，邓维斌. 基于样本权重的不平衡数据欠抽样方法[J]. 计算机研究与发展，2016，53(11)：2613-2622.

[15] KUBAT M，MATWIN S. Addressing the curse of imbalanced training sets：one-sided selection[C]// ICML. 1997：179-186.

[16] HART P E. The Condensed Nearest Neighbor Rule[J]. IEEE Transactions on Information Theory，1968，14：515-516.

[17] LAURIKKALA J. Improving identification of difficult small classes by balancing class distribution[C]// Conference on Artificial Intelligence in Medicine in Europe. Springer Berlin Heidelberg，2001：63-66.

[18] 胡小生，张润晶，钟勇. 两层聚类的类别不平衡数据挖掘算法[J]. 计算机科学，2013，40(11)：271-275.

[19] 李克文，杨磊，刘文英，等. 基于 RSBoost 算法的不平衡数据分类方法[J]. 计算机科学，2015，42(9)：249-252.

[20] CHAN P K，STOLFO S J. Toward Scalable Learning with Non-Uniform Class and Cost Distributions：A Case Study in Credit Card Fraud Detection[C]// KDD. 1998：164-168.

[21] SUN Z，SONG Q，ZHU X，et al. A novel ensemble method for classifying imbalanced data[J]. Pattern Recognition，2015，48(5)：1623-1637.

[22] KITTLER J，HATEF M，DUIN R P W，et al. On combining classifiers[J]. IEEE transactions on pattern analysis and machine intelligence，1998，20(3)：226-239.

[23] SCHÖLKOPF B，PLATT J C，SHAWE-TAYLOR J，et al. Estimating the support of a high-dimensional distribution[J]. Neural computation，2001，13(7)：1443-1471.

[24] COHEN G，HILARIO M，PELLEGRINI C. One-class support vector machines with a conformal kernel. a case study in handling class imbalance[C]// Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Springer Berlin Heidelberg，2004：850-858.

[25] MANEVITZ L M，YOUSEF M. One-class SVMs for document classification[J]. Journal of Machine Learning Research，2001，2(1)：139-154.

[26] ELKAN C. The foundations of cost-sensitive learning[C]// International Joint Conference on Artificial Intelligence. 2001：973-978.

[27] DOMINGOS P. Metacost：A general method for making classifiers cost-sensitive[C]// Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，1999：155-164.

[28] 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[J]. 计算机研究与发展，2011，48(S1)：387-390.

[29] CHAI X，DENG L，YANG Q，et al. Test-cost sensitive naive bayes classification[C]// IEEE International Conference on Data Mining，2004(ICDM'04). IEEE，2004：51-58.

[30] FAN W，STOLFO S J，ZHANG J，et al. AdaCost：misclassification cost-sensitive boosting[C]// ICML. 1999：97-105.

[31] SUN Y，KAMEL M S，WANG Y. Boosting for learning multiple classes with imbalanced class distribution[C]// Sixth International Conference on Data Mining (ICDM'06). IEEE，2006：592-602.

[32] 李秋洁，茅耀斌，王执铨. 基于 Boosting 的不平衡数据分类算法研究[J]. 计算机科学，2011，38(12)：224-228.

（下转第 57 页）

机系统应用，2015，24(8)：197-201.
[14] 朱桂英，张瑞林. 基于 Hough 变换的圆检测方法[J]. 计算机工程与设计，2008，29(6)：1462-1464.
[15] ALEGRIA F C，SERRA A C. Computer vision applied to the automatic calibration of measuring instruments[J]. Measurement，2000，28(3)：185-195.
[16] WANG Q，TANG X，DING C，et al. Automatic alignment system based on center point recognition of analog measuring instruments dial[C]// Conference of the IEEE Industrial Electronics Society，IECON 2013. New York：IEEE，2013：5532-5536.
[17] LI B，JIA Z. Some Results On Condition Numbers Of The Scaled Total Least Squares Problem[J]. Linear Algebra & Its Applications，2009，435(3)：674-686.
[18] 张远辉，张鼎，许昌，等. 指针式仪表总体最小二乘图像校验算法[J]. 自动化仪表，2015，36(5)：75-79.
[19] BRESENHAM J. A linear algorithm for incremental digital display of circular arcs[J]. Communications of the Acm，1977，20(2)：100-106.
[20] BELAN P A，ARAUJO S A，LIBRANTZ A F H. Segmentation-free approaches of computer vision for automatic calibration of digital and analog instruments[J]. Measurement，2013，46(1)：177-184.
[21] LIU J，LIU Y，YU L. Novel method of Automatic Recognition for Analog Measuring Instruments[C]// International Conference on Manufacturing Science and Engineering. 2015：67-74.
[22] YANG Z，NIU W，PENG X，et al. An image-based intelligent system for pointer instrument reading[C]// IEEE International Conference on Information Science and Technology. IEEE，2014：780-783.
[23] YUE X F，MIN Z，ZHOU X D，et al. The Research on Auto-recognition Method for Analogy Measuring Instruments[C]// International Conference on Computer，mechatronics，Control and Electronic Engineering. 2010：207- 210.
[24] GONZALEZ R C，WOODS R E. 数字图像处理(第三版)[M]. 阮秋琦，阮宇智，译. 北京：电子工业出版社，2011.
[25] SABLATNIG R，KROPATSCH W G. Automatic reading of analog display instruments[C]// Iapr International Conference on Pattern Recognition. New York：IEEE，1994：794-797.
[26] ALEGRIA E C，SERRA A C. Automatic calibration of analog and digital measuring instruments using computer vision[J]. IEEE Transactions on Instrumentation & Measurement，2000，49(1)：94-99.
[27] HEMMING B，LEHTO H. Calculation of uncertainty of measurement in machine vision case：a system for the calibration of dial indicators[C]// Proceedings of the 18th IEEE Instrumentation and Measurement Technology Conference. Budapest，Hungary：IEEE，2001：665-670.
[28] HEMMING B，FAGERLUND A，LASSILA A. High-accuracy automatic machine vision based calibration of micrometers[J]. Measurement Science & Technology，2007，18(18)：1655-1660.
[29] DATTA A，KIM J S，KANADE T. Accurate camera calibration using iterative refinement of control points[C]// IEEE，International Conference on Computer Vision Workshops. IEEE，2009：1201-1208.
[30] TSAI C Y，TSAO A H，WANG C W. Real-Time Feature Descriptor Matching via a Multi-Resolution Exhaustive Search Method[J]. Journal of Software，2013，8(9)：2197-2201.
[31] HOUGH P V C. Method and means for recognizing complex patterns：U. S. Patent 3069654[P]. 1962.
[32] 陶冰洁，韩佳乐，李恩. 一种实用的指针式仪表读数识别方法[J]. 光电工程，2011，38(4)：145-150.
[33] 颜友福，刘金清，吴庆祥. 基于区域生长的指针式仪表自动识别方法[J]. 计算机系统应用，2015，24(4)：164-170.
[34] 李祖贺，刘嘉，薛冰，等. 面向自动校验系统的指针式压力表读数识别[J]. 计算机工程与应用，2016，52(23)：213-219.
[35] 张春雪. 图像的边缘检测方法研究[D]. 江苏：江南大学，2011.

(上接第 27 页)

[33] 李雄飞，李军，董元方，等. 一种新的不平衡数据学习算法 PC-Boost[J]. 计算机学报，2012，35(2)：202-209.
[34] 袁兴梅，杨明，杨杨. 一种面向不平衡数据的结构化 SVM 集成分类器[J]. 模式识别与人工智能，2013，26(3)：315-320.
[35] ARUNASALAM B，CHAWLA S. CCCS：a top-down associative classifier for imbalanced class distribution[C]// 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2006：517-522.
[36] PATEL H，THAKUR G S. A Hybrid Weighted Nearest Neighbor Approach to Mine Imbalanced Data[C]// International Conference on Data Mining (DMIN). 2016：106.
[37] IMAM T，KAI M T，KAMRUZZAMAN J. z-SVM：An SVM for Improved Classification of Imbalanced Data[C]// Australasian Joint Conference on Artificial Intelligence. Springer Berlin Heidelberg，2006：264-273.
[38] KUBAT M，HOLTE R C，MATWIN S. Machine learning for the detection of oil spills in satellite radar images[J]. Machine Learning，1998，30(2)：195-215.
[39] BRADLEY A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms[M]. Elsevier Science Inc.，1997.
[40] FAWCETT T. An introduction to ROC analysis[J]. Pattern Recognition Letters，2006，27(8)：861-874.
[41] PROVOST F，DOMINGOS P. Tree induction for probability-based ranking[J]. Machine Learning，2003，52(3)：199-215.
[42] HAND D J，TILL R J. A simple generalisation of the area under the ROC curve for multiple class classification problems[J]. Machine Learning，2001，45(2)：171-186.
[43] DAVIS J，GOADRICH M. The relationship between Precision-Recall and ROC curves[C]// 23rd International Conference on Machine Learning. ACM，2006：233-240.
[44] DRUMMOND C，HOLTE R C. Cost curves：An improved method for visualizing classifier performance[J]. Machine learning，2006，65(1)：95-130.