一种基于样本空间的类别不平衡数据采样方法

张永清[1, 2] 卢荣钊[1] 乔少杰[3] 韩 楠[4] GUTIERREZ Louis Alberto[5] 周激流[1]

摘 要 不平衡数据是机器学习中普遍存在的问题并得到广泛研究, 即少数类的样本数量远远小于多数类样本的数量. 传统基于最小化错误率方法的不足在于: 分类结果会倾向于多数类, 造成少数类的精度降低, 通常还存在时间复杂度较高的问题. 为解决上述问题, 提出一种基于样本空间分布的数据采样方法, 伪负样本采样方法. 伪负样本指被标记为负样本 (多数类) 但与正样本 (少数类) 有很大相关性的样本. 算法主要包括 3 个关键步骤: 1) 计算正样本的空间分布中心并得到每个正样本到空间中心的平均距离; 2) 以同样的距离计算方法计算每个负样本到空间分布中心的距离, 并与平均距离进行比较, 将其距离小于平均距离的负样本标记为伪负样本; 3) 将伪负样本从负样本集中删除并加入到正样本集中. 算法的优势在于不改变原始数据集的数量, 因此不会引入噪声样本或导致潜在信息丢失; 在不降低整体分类精度的情况下, 提高少数类的精确度. 此外, 其时间复杂度较低. 经过 13 个数据进行多角度实验, 表明伪负样本采样方法具有较高的预测准确性.

关键词 不平衡数据, 样本空间, 机器学习, 采样方法, 空间中心

引用格式 张永清, 卢荣钊, 乔少杰, 韩楠, Gutierrez Louis Alberto, 周激流. 一种基于样本空间的类别不平衡数据采样方法. 自动化学报, 2022, **48**(10): 2549−2563

DOI 10.16383/j.aas.c200034

A Sampling Method of Imbalanced Data Based on Sample Space

ZHANG Yong-Qing[1, 2] LU Rong-Zhao[1] QIAO Shao-Jie[3] HAN Nan[4]
GUTIERREZ Louis Alberto[5] ZHOU Ji-Liu[1]

Abstract Data imbalance is a very common problem that has been comprehensively studied in machine learning techniques, where the minority class contains very few samples compared with the majority class. The disadvantage of traditional methods based on minimizing the error lies in: they tend to be biased toward the majority class, so these models have low prediction accuracy for the minority class and might have high time complexity. To solve the above problems, a data sampling method based on spatial distribution, Pseudo-negative sampling is proposed. Pseudo-negative samples refer to samples marked as negative samples (majority class) but with a strong correlation with positive samples (minority class). The algorithm mainly includes three key steps:1) calculate the spatial center of the positive samples and figure out the average distance of positive samples to the spatial center; 2) calculate the distance from each negative sample to the spatial center with similar distance calculation approach and compare it with the average distance, and then mark the negative sample as pseudo negative sample whose distance is less than the average distance; 3) delete the pseudo negative samples from the negative samples and add them to the positive sample set. The advantage of the algorithm is that it does not change the number of original data sets, so it does not introduce noise samples or cause potential information loss; the accuracy of a few classes can be improved without decreasing the overall classification accuracy and the time cost is low. Extensive experiments are conducted on thirteen datasets from multiple aspects, and the results show that the pseudo-negative sampling method has high prediction accuracy.

收稿日期 2020-01-16 录用日期 2020-05-03
Manuscript received January 16, 2020; accepted May 3, 2020
国家自然科学基金 (61702058, 61772091, 61802035, 61962006), 四川省科技计划项目 (2021JDJQ0021, 22ZDYF2680, 2021YZD0009, 2021ZYD0033), 成都市技术创新研发项目 (2021-YF05-00491-SN), 成都市重大科技创新项目 (2021-YF08-00156-GX), 成都市 "揭榜挂帅"科技项目 (2021-JB00-00025-GX), 四川音乐学院数字媒体艺术四川省重点实验室资助项目 (21DMAKL02), 广东省基础与应用基础研究基金 (2020B1515120028) 资助
Supported by the National Natural Science Foundation of China (61702058, 61772091, 61802035, 61962006), Sichuan Science and Technolo-gy Program (2021JDJQ0021, 22ZDYF2680, 2021YZD0009, 2021ZYD0033), Chengdu Technology Innovation and Research and Development Project(2021-YF05-00491-SN), Chengdu Major Science and Technology Innovation Project (2021-YF08-00156-GX), Chengdu "Take the lead" Science and Technology Project (2021-JB00-00025-GX), Key Laboratory of Digital Media Art of Sichuan Province, Sichuan Conservatory of Music (21DMAKL02), and Guangdong Basic and Applied Basic Resear-ch Foundation (2020B1515120028)
本文责任编委 董峰
Recommended by Associate Editor DONG Feng

1. 成都信息工程大学计算机学院 成都 610225 中国 2. 电子科技大学计算机科学与工程学院 成都 611731 中国 3. 成都信息工程大学软件工程学院 成都 610225 中国 4. 成都信息工程大学管理学院 成都 610103 中国 5. 伦斯勒理工学院计算机科学系 纽约 12180 美国
1. School of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China 2. School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China 3. School of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China 4. School of Management, Chengdu University of Information Technology, Chengdu 610103, China 5. Department of Computer Science, Rensselaer Polytechnic Institute, New York 12180, USA

Key words Imbalanced data, spatial distribution, machine learning, sampling method, spatial center

Citation Zhang Yong-Qing, Lu Rong-Zhao, Qiao Shao-Jie, Han Nan, Gutierrez Louis Alberto, Zhou Ji-Liu. A sampling method of imbalanced data based on sample space. *Acta Automatica Sinica*, 2022, **48**(10): 2549−2563

不平衡数据广泛存在于实际应用中, 如何有效处理类别不平衡数据已成为目前机器学习领域一个重要的研究热点. 许多生物信息学中的分类问题都面临不平衡数据的问题, 如基因表达数据[1]、蛋白质−DNA 结合数据[2]、mRNA 中的甲基化位点[3]、拼接位置预测[4]、microRNAs 的预测[5]、蛋白质相互作用预测[6] 等. 此外, 不平衡数据还广泛存在于医疗诊断[7−8]、诈骗交易[9] 和网络入侵[10] 等领域. 在数据不平衡问题中, 由于负样本 (多数类) 的数量远远大于正样本 (少数类) 的数量, 使得少数类样本难以被分类器有效学习. 此外, 现有的机器学习算法一般假定类分布均衡或样本错分代价相同. 然而, 真实应用中通常少数类样本比多数类本更为重要, 错分代价更高. 所以对不平衡数据的学习一般无法取得令人满意的结果.

现有方法一般通过数据预处理的方式来重构数据集, 以减少学习过程中样本偏态分布的负面影响, 重采样方法是其中经典的方法. 重采样主要分为欠采样和过采样, 使用欠采样算法可能会移除多数类中潜在的有用信息, 导致分类性能降低, 并且可能破坏样本原始分布. 过采样算法会增加样本量, 这会增加算法的时间成本, 也容易导致过拟合[11]. 此外, 新生成的样本不能保证与原数据有相同的分布. 大多数方法将数据采样到所有类别样本数量一致为止, 采样比例不仅取决于不平衡比例, 还取决于数据的空间分布情况. 因此重采样算法的一个难点在于如何确定采样比例, 即 如何合理地根据数据本身的特点确定具有最佳分类性能的采样比例.

基于上述问题, 亟需提出一种先进的数据采样方法来处理正负样本比例不平衡问题. 本文研究基于以下几点考虑:

1) 在不平衡数据中, 负样本数量占据了绝大多数, 虽然负样本与正样本属于不同的类别, 但是在负样本中可能包括潜在的正样本, 这是之前的研究没有考虑的.

2) 如何根据数据整体的空间分布特点, 自适应地确定采样比例.

3) 基于混合采样方法能很好地避免单独使用欠采样和过采样带来的问题.

为解决上述问题, 本文提出了一种新的基于样本空间的不平衡数据采样方法, 伪负样本采样方法 (Pseudo-negative sampling, PNS), 本文主要贡献有:

1) 提出了伪负样本概念. 在大量的负样本中存在与正样本有类似分布的样本, 因此与正样本具有很高的相似度, 可以将它们定义为被错分了的正样本. 基于这一观察, 本文首次提出伪负样本概念, 将与正样本相似度很高的负样本标记为伪负样本.

2) 根据数据空间分布, 提出一种度量正样本和负样本之间相似性的方法. 算法工作原理为: 使用欧氏距离评价样本之间的相似性, 首先计算正样本的空间中心, 然后将正样本到空间中心的平均距离作为判断是否为伪负样本的阈值, 最后分别计算每个负样本到空间中心的距离. 如果其距离小于阈值, 则将此负样本标记为伪负样本. 将其添加到正样本集中.

3) 通过正负样本之间的相似距离, 自适应地确定不平衡数据采样的比例.

4) 在多个 UCI 数据、KEEL 数据和真实生物信息数据上进行了大量实验, 全面验证了算法的准确率、敏感性、特异性、马修斯相关系数 (Matthews correlation coefficient, MCC)、F-score 和时间效率等性能评价指标. 引入对比算法, 从多角度验证所提出方法的性能优势.

本文结构如下: 第 1 节综述主流的类不平衡数据解决方法; 第 2 节详细说明本文提出的 PNS 采样算法; 第 3 节介绍本文使用的数据集和算法评价指标; 第 4 节对本文提出的采样方法的实验结果进行分析; 第 5 节对本文工作进行总结和展望.

1 相关工作

如何处理类别不平衡数据是分类中的一个关键问题, 并受到广泛关注. 现有方法可分为数据预处理[12−14]、代价敏感学习[15] 和集成学习[16] 三类.

数据预处理是最常用的方法, 因为它独立于分类器, 具有很好的适应性, 主要包括过采样[17] 和欠采样[18]. 过采样是通过创建新的少数类样本来消除偏态分布的危害, 提高少数类的分类性能. 最简单的方法是随机过采样 (Random over-sampling, ROS), 即随机复制少数类样本, 缺点是少数类没有增加任何额外信息, 只是简单复制, 从而增加过拟合的风险, 并且新的数据使训练分类器所需时间增加. 在改进的过采样方法中, Chawla 等[19] 提出了 Synthetic minority oversampling technique

(SMOTE) 算法, 在少数类样本中随机插值邻居样本来生成新样本. 但这种方法容易产生分布边缘化问题, 新生成样本可能会模糊正样本和负样本的边界. 虽然使数据集的平衡性得到了改善, 但加大了分类算法进行分类的难度. Douzas 等[20] 将深度学习模型中的生成对抗网络用于少数类样本的合成, 很好地平衡了数据集, 并取得了较好结果. 欠采样是通过移除多数类样本来消除偏态分布的危害, 从而提高少数类的分类性能. 最简单的方法是随机欠采样 (Random under-sampling, RUS), 即随机地去掉一些多数类样本, 缺点是可能会丢失一些重要信息, 对已有息利用不充分. Wilson[21]提出了一种最近邻规则欠抽样 (Edited nearest neighbor, ENN) 方法, 基本思想是删除其最近的 3 个近邻样本中具有 2 个或者 2 个以上类别不同的样本. 但是大多数的多数类样本附近的样本都是多数类的, 所以该方法所能删除的多数类样本十分有限. 因此, Laurikkala[22] 在 ENN 的基础上提出了邻域清理规则欠采样方法 (Neighborhood cleaning, NCL), 核心思想是找出每个样本的 3 个最近邻样本, 若该样本是多数类样本且其 3 个最近邻中有 2 个以上是少数类样本, 则删除它; 反之, 当该样本是少数类, 并且其 3 个最近邻中有 2 个以上是多数类样本, 则去除近邻中的多数类样本. 但是该方法中未能考虑到在少数类样本中存在的噪声样本, 而且这种方法删除的多数类样本大多属于边界样本, 对后续分类器的分类会产生很大的不良影响.

传统分类器在训练时, 往往以最小化错误率为目标, 这一目标是基于假设: 不同类之间的错误分类具有相同代价, 因此不同类的错分可以被同等对待. 然而在类别不平衡数据集中, 多数类与少数类之间的错分代价往往是不同的, 错分少数类具有更高的代价. 基于这一前提, 代价敏感方法通过引入代价矩阵为不同错分类型赋予不同代价, 然后以最小化代价值为目标来构造分类器. Zhang 等[23] 将代价敏感学习应用于不平衡数据的多分类, 通过一对一分解, 将多分类问题转化成多个二分类子问题并使用代价敏感的反向传播神经网络进行独立学习, 从而减小平均误分代价. Liu 等[24] 提出了一种新的代价敏感的支持向量机 (Support vector machine, SVM) 算法, 该算法首先使用过滤式方法对特征进行挑选, 同时对于代价敏感 SVM 中的参数, 使用元优化算法进行优化. 实验表明, 该方法在对乳腺癌数据的预测上取得了较好结果.

集成学习方法的主要思想是将多个不同的弱学习器组合在一起, 形成一个强学习器. 通过利用每个基学习器之间的差异, 来改善模型的泛化性能. 经典的方法有 Bagging 和 Boosting 等. Breiman[25] 将自采样引入集成学习提出了 Bagging 集成方法, 他通过从原始数据集不断采样产生新的数据子集来训练每个新的分类器, 由于数据子集的不同, 保证了基分类器具有一定的多样性. Schapire[26] 则提出了 Boosting 集成方法, AdaBoost[27]是其中的代表性方法, 它使用整个数据集来不断地训练分类器, 在每一个分类器被训练出来后, 后面的分类器将更多关注被错分的样本, 从而提高少数类的精度. 关注的方法是为样本设置权重, 被前一个分类器正确分类的样本, 权重将降低, 反之将权重提高.

在相似性度量方面, 欧氏距离作为一种简单有效的评价方式被广泛使用, 其计算公式见下:

$$dist(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \tag{1}$$

式中, X 和 Y 表示 2 个被考虑的样本, x_i 与 y_i 表示样本 X 与 Y 的第 i 个特征, n 表示特征数. Elmore 等[28] 提出了基于欧氏距离的主成分分析 (Principal component analysis, PCA) 方法. 该方法使用基于欧氏距离得到的相似度矩阵, 识别彼此接近的参数, 为 PCA 中相似性度量提供了更多选择. Park 等[29] 在对歌曲的相似性识别中, 结合欧氏距离和汉明距离, 提出了一种新的距离度量方法, 称之为条件欧几里得距离.

通过上述工作分析可知, 现有研究工作中存在的突出问题: 1) 采样时, 没有充分考虑数据的空间分布特点, 特别是正样本集的分布, 导致采样具有较大盲目性; 2) 需要人为指定采样比例, 采样比例应该根据数据本身的特点确定, 如何针对不同数据进行采样比例的适应性调整.

2 问题描述

2.1 伪负样本采样方法

算法中使用的主要符号及说明如表 1 所示.

在不平衡数据的负样本集中, 可能存在潜在的正样本, 本文称之为伪负样本. 如果能有效地找出伪负样本, 将其加入到正样本集中同时从负样本集中删除, 便能得到一个数据分布更加合理的数据集. 基于这个数据集训练的分类器可以更好地学习正样本集, 从而提高正样本集的精确度. 基于这一考虑, 本文首次提出了伪负样本采样方法 PNS. 图 1 描述了如何从多数类中找出伪负样本, 图 1 中空心圆代表多数类, 空心五星代表少数类. 首先需要找到少

表 1 符号及说明
Table 1 Symbols and their explanations

名称	解释
D^+, m	正样本集与正样本个数. 包含的样本表示为 $D^+ = \left\{\left(x_1^+, y_1^+\right), \left(x_2^+, y_2^+\right), \cdots, \left(x_m^+, y_m^+\right)\right\}$
D^-, n	负样本集与负样本个数. 包含的样本表示为 $D^- = \left\{\left(x_1^-, y_1^-\right), \left(x_2^-, y_2^-\right), \cdots, \left(x_n^-, y_n^-\right)\right\}$
D^*	伪负样本集. 包含的样本表示为 $D^* = \{(x_1^*, y_1^*), (x_2^*, y_2^*), \cdots, (x_i^*, y_i^*)\}\}$
$Q(x_i)$	样本 x_i 的相似性大小
$dist(x_1, x_2)$	样本 x_1 与样本 x_2 间的欧氏距离
C	正样本空间中心, 是所有正样本的平均值
$meanDist$	将负样本判断为伪负样本的阈值, 其值是所有正样本到空间中心 C 的平均距离

数类的空间中心, 图 1 中用实心五星表示, 并得到所有少数类样本到空间中心的平均欧氏距离, 然后分别计算所有多数类样本到空间中心的欧氏距离. 若某个多数类样本到空间中心的距离越近, 则认为该多数类样本与少数类样本相似性越高. 如果某个多数类样本到空间中心的距离小于平均欧氏距离, 则将此负样本认定为潜在的正样本即伪负样本.

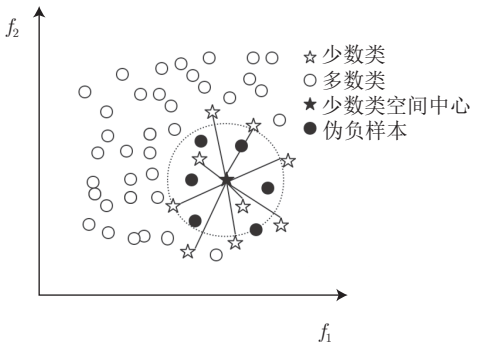


图 1 伪负样本采样方法
Fig.1 Pseudo-negative sampling method

本文使用 $D^+ = (x_1^+, y_1^+), (x_2^+, y_2^+), \cdots, (x_m^+, y_m^+)$ 代表正样本集, $D^- = (x_1^-, y_1^-), (x_2^-, y_2^-), \cdots, (x_n^-, y_n^-)$ 代表负样本集, $D^* = (x_1^*, y_1^*), (x_2^*, y_2^*), \cdots, (x_i^*, y_i^*)$ 代表伪负样本集, 其中 m 表示正样本数量, n 表示负样本数量, i 表示伪负样本数量, 且 $m \ll n$. 伪负样本采样的目的是基于 D^+ 和 D^- 确定伪负样本集 D^*, 其中 $i < n$.

在 PNS 算法中需要首先确定相似性度量方法 $Q(x_i)$, 度量方法作为评价伪负样本的标准起着至关重要的作用. 本文使用样本间的欧氏距离作为标准, $Q(x_i)$ 表示第 i 个样本的相似性大小. 也可以根据数据实际情况进行适应性调整. 然后初始化伪负样本集 $D_0^* = \emptyset$. 在每一次迭代过程中, 算法将根据度量方法 $Q(x_i^-)$ 逐个评价样本 x_i^-, 并将相似性大于某一阈值的负样本加入到伪负样本集 D^* 中:

$$D_k^* = D_{k-1}^* \cup D^{-'} \tag{2}$$

$$D_k^- = D_{k-1}^- - D^{-'} \tag{3}$$

式中, k 表示迭代次数, $D^{-'}$ 表示相似性大于阈值的负样本, D_{k-1}^* 表示上一次迭代后得到的伪负样本集. 同理, D_{k-1}^- 表示上一次迭代后得到的负样本集.

迭代结束之后, 将伪负样本集加入到正样本集当中, 同时得到了平衡后的负样本集. 具体计算过程将在第 2.3 节给出.

2.2 基于欧氏距离的 PNS 采样算法

PNS 算法是基于正样本集空间位置的, 因此, 首先需要找到正样本的空间中心点, 空间中心点 C 是所有正样本的平均值, 计算方法如下:

$$C = \frac{\sum_{i=1}^{m} x_i^+}{|D^+|} \tag{4}$$

式中, x_i^+ 表示正样本集中第 i 个样本. 得到正样本的空间中心后, 需要一个相似性评价阈值来判断是否为伪负样本, 判断阈值由所有正样本到空间中心 C 的欧氏距离的平均值 $meanDist$ 表示, 计算方法如下:

$$meanDist = \frac{\sum_{i=1}^{m} dist\left(x_i^+ C\right)}{|D^+|} \tag{5}$$

式中, $dist(x_i^+, C)$ 表示正样本 x_i^+ 与空间中心 C 之间的欧氏距离. 然后, 计算每个负样本与正样本集的相似性, 正样本集使用空间中心 C 代替, 计算公式如下:

$$Q(x_i^-) = dist\left(x_i^-, C\right) \tag{6}$$

式中, $i = 1, 2, 3, \cdots, n$. $dist(x_i^-, C)$ 表示负样本 x_i^- 与空间中心 C 之间的欧氏距离, 计算结果即为样本 x_i^- 具有的相似性大小. 然后将每个负样本的相似性

$Q(x_i^-)$ 与阈值 $meanDist$ 进行比较, 如果小于阈值, 则认定该负样本为伪负样本, 定义如下:

$$D^* = \left\{x_i^- | Q(x_i^-) < meanDist,\ i = 1, 2, 3, \cdots, n\right\} \tag{7}$$

最终, 将伪负样本集加入到正样本并从负样本集中删除, 最终得到采样后的数据集:

$$D^+ = D^+ \cup D^* \tag{8}$$

$$D^- = D^- - D^* \tag{9}$$

2.3　算法描述

基于上述讨论, 给出本文算法的形式化描述, 如算法 1 所示.

算法基本步骤为: 第 7 ~ 13 步将原始数据集分成正样本集和负样本集; 第 14 ~ 17 步计算正样本的空间中心 C; 第 18 ~ 21 步计算少数类到空间中心的平均距离 $meanDist$; 第 22 ~ 24 步计算每个多数类到平均中心的距离 $Distance_i$; 第 25 ~ 29 步根据多数类样本距离与平局距离判断某个多数类是否为伪负样本, 如果是, 则加入伪负样本; 最后返回采样后的数据集. 其中, $dist(A, B)$ 表示计算 A 点到 B 点的欧氏距离.

算法复杂性分析: 本文提出的算法还具有良好的时间复杂度, 由算法 1 中可以看出, 耗时操作主要集中在 5 个循环操作上: 1) 样本分离操作, 时间复杂度为 $O(k)$, 其中 k 代表样本总数. 2) 计算正样本中心, 时间复杂度为 $O(m)$, m 表示正样本数量. 3) 计算正样本到中心的平均距离, 时间复杂度为 $O(m)$. 4) 计算每个负样本到中心的距离, 时间复杂度为 $O(n)$, n 表示负样本数量. 5) 将每个负样本到中心的距离与平均距离进行比较, 时间复杂度为 $O(n)$. 综上, PNS 算法的总时间复杂度为 $O(k + 2\times m + 2 \times n)n)$, 由于在数据集中 k 等于 m 加上 n, 因此原式可化简为 $O(3 \times k)$. 由此看出, PNS 算法的时间复杂度较低, 是一种高效的算法.

算法 1. 基于伪负样本的采样方法

输入. 原始数据集 $D = (x_1, y_1), (x_2, y_2), \cdots, (x_k, y_k)$

输出. 采样后数据集 D'.

1) $D^- = \emptyset$;
2) $D^+ = \emptyset$;
3) $D^* = \emptyset$;
4) $C = 0$;
5) $meanDist = 0$;
6) $Distance = \emptyset$;
7) for $i = 1$ to k do;
8) if D_i is $PositiveSample$ then;
9) $D^+ = D^+ \cup D_i$;
10) else;
11) $D^- = D^- \cup D_i$;
12) end if;
13) end for;
14) for $j = 1$ to $|D^+|$ do;
15) $C = C + D_j^+$;
16) end for;
17) $C = C / |D^+|$;
18) for $j = 1$ to $|D^+|$ do;
19) $meanDist = meanDist + dist\left(D_j^+, C\right)$;
20) end for;
21) $meanDist = meanDist / |D^+|$;
22) for $i = 1$ to $|D^-|$ do;
23) $Distance_i = dist\left(D_i, C\right)$;
24) end for;
25) for $d = 1$ to $|D^-|$ do;
26) if $Distance_d < meanDist$ then;
27) $D^* = D^* \cup D_d$;
28) end if;
29) end for;
30) $D^+ = D^+ \cup D^*$;
31) $D^- = D^- - D^*$;
32) return $D^{'} = D^- \cup D^+$.

2.4　对比采样方法

本文使用 ROS、RUS、Adaptive synthetic sampling (ADASYN) 和 SMOTE 作为对比采样算法与 PNS 进行比较. 其中, RUS 属于欠采样, 其余方法属于过采样. ROS 与 RUS 均是随机采样, 前者通过随机复制少数类样本对数据进行采样, 后者通过删除多数类样本进行采样. 这两种方法具有实现简单, 采样效果较好的特点.

SMOTE[19] 方法基于少数类间的相似性合成新样本. 对于少数类样本集 S_{min}, 首先计算得到每个样本 $x_i \in S_{min}$ 的 K 近邻. K 近邻被定义为距 x_i 最近的 K 个样本, 距离计算通常是欧氏距离, 整数 K 是人工指定的超参数. 为了合成新样本, 随机从 K 个近邻样本中选择一个求出两者的差, 然后乘以介于 [0, 1] 之间的特征向量差异随机数, 最后加上原始特征 x_i.

$$x_{new} = x_i + (\hat{x}_i - x_i) \times \delta \tag{10}$$

式中, $x_i \in S_{min}$ 是正在被考虑的样本, $\hat{x}_i$ 是 x_i 其中一个 K 近邻样本, 且 $\hat{x}_i \in S_{min}$. $\delta \in [0, 1]$ 是一个随

机数. 因此, 根据式 (10) 得到的合成实例是所考虑的 x_i 与随机选取的 K 近邻的连线线段上的一个点. SMOTE 的提出避免了 ROS 带来的过拟合问题, 同时显著提高分类器性能. 已经在各种领域得到了广泛认可.

He 等[30] 基于对 SMOTE 的改进提出了 ADASYN 采样. ADASYN 的主要思想是根据少数类的分布自适应合成新样本: 在合成新样本过程中, 分类困难的少数类样本会生成更多样本, 反之则会生成较少样本, 以此将决策边界转移到难以学习的样本上. 该方法与 SMOTE 的不同点主要在于对少数类合成样本的控制. 在 SMOTE 中, 对每个少数类都合成相同数量的样本, 而在 ADASYN 中, 处于边界的少数类将合成更多样本. 对边界的检测通过样本的 K 近邻得到, 如果一个少数类的 K 近邻存在越多的多数类, 那么这个少数类被认为离边界越近, 会合成更多样本.

3 数据集及算法评价指标

3.1 数据集

为评价不同样本采样方法在不同数据集上的预测性能, 并与其他常用采样方法进行比较, 本文使用了 7 个 UCI 数据集[31]、4 个 KEEL 数据集[32] 和 2 个真实的生物信息学数据集. 如表 2 所示.

表 2 不平衡数据集信息

Table 2 Information of the imbalanced dataset

来源	数据集	样本数	特征数	比例	特征属性 (连续/离散)
真实数据	SPECT	267	44	4	44/0
	SNP	3074	25	16	25/0
UCI 数据	Ecoli	336	7	8.6	7/0
	SatImage	6435	36	9.3	0/36
	Abalone	4177	8	9.7	6/2
	Balance	625	4	11.7	0/4
	SolarFlare	1389	10	19	0/10
	Yeast_ME2	1484	8	28	8/0
	Abalone_19	4177	8	130	6/2
KEEL 数据	Yeast1289vs7	947	8	30.6	8/0
	Yeast1458vs7	693	8	22.1	8/0
	Yeast4	1484	8	28.1	8/0
	Yeast5	1484	8	32.7	8/0

所有数据集用于二分类问题, 如果出现多分类数据集, 则将其中某一类作为正样本集, 剩下的所有类统一合并为负样本集. 正负样本数据集的不平衡比例从 4 到 130 不等, 较大的不平衡比例表示正样本集和负样本集之间数量差异较大.

3.2 UCI 数据集

Ecoli 数据集包含 35 个少数类和 301 个多数类, 有 7 个特征. 该数据是一组蛋白质定位点数据, 特征包括氨基酸序列和来源信息, 使用这些信息预测蛋白质的定位位点.

SatImage 数据中包含卫星图像 3×3 邻域中的像素的多光谱值, 以及与每个邻域中的中心像素相关联的分类. 通过整合不同类型和分辨率的空间数据 (包括多光谱和雷达数据、地图指示地形、土地利用等) 对场景的解释预计将具有重要意义. 这个数据集中包含 626 个少数类和 5 809 个多数类, 有 36 个特征.

Abalone 是一个通过物理测量来预测鲍鱼年龄的数据集, 物理测量预测鲍鱼年龄是一项既枯燥又耗时的工作, 因此使用已有数据进行预测将是更省时的选择. 这个数据集包含 390 个少数类和 3 787 个多数类, 有 8 个特征.

Balance 数据集是用来模拟心理实验结果的, 每个例子都被分类为天平的左端、右端或是平衡. 属性包括左权重、左距离、右权重和右距离.

SolarFlare 数据集记录了太阳耀斑的数量, 每个属性计算 24 小时内某类太阳耀斑的数量, 每个实例表示太阳上 1 个活动区域内所有种类耀斑数量. 该数据包含 69 个少数类和 1 320 个多数类, 有 10 个特征.

Yest_ME2 数据集是一个酵母菌数据集, 用于预测酵母菌蛋白质的定位位点. 该数据包含 51 个少数类和 1 433 个多数类, 有 8 个特征数.

3.3 真实生物数据

SPECT 数据集是心脏单质子发射计算机断层扫描图像的诊断结果. 每个病人被分为正常和异常两类. 数据包含对 267 个 SPECT 图像集 (患者) 的数据处理结果. 提取总结原始 SPECT 图像的特征, 得到 44 个连续特征. 在 267 个样本中, 包含 55 个正常病人 (少数类) 和 212 个异常病人 (多数类).

SNP 是指在基因组上单个核苷酸的变异, 变异形式包括缺失、颠换、变异和插入. 在人类基因组中大概每 1 000 个碱基就有一个 SNP, 因此 SNP 的数量是相当庞大的. 研究表明, SNP 同人群分类, 遗传疾病都有密切联系. 该数据包含 183 个少数类和 2891 个多数类, 25 个特征.

3.4 KEEL 数据集

本文使用 KEEL 数据集的 4 种酵母菌数据集,

原始数据集是一个多分类数据集. 在 Yeast1289vs7 中, 将属于 VAC 的样本标记为正样本, 属于 NUC、CYT、POX 和 ERL 的标记为负样本. Yeast1458vs7 属于 VAC 的样本标记为正样本, 属于 NUC、ME2、ME3 和 POX 的标记为负样本. 在 Yeast4 和 Yeast5 中, 分别将 ME2、ME1 标记为正样本, 将其他所有样本均标记为负样本. 所有数据集包含 8 个特征.

3.5 评价指标

不平衡数据学习的困难不仅体现在分类器的训练上, 同时还在于如何客观评价不平衡分类器的性能上. 使用总体精度已经不能客观评价不平衡分类器的性能, 因为不平衡数据中多数类与少数类具有不同的重要性, 对少数类的错误将导致更严重的错误. 而总体精度忽略了这一关键因素, 即使将结果全部预测为多数类, 仍能得到较高总体精度, 难以准确反应出分类器在不平衡数据集上的性能. 本节介绍本文使用的评价指标, 并给出计算公式.

分类性能的评估主要基于混淆矩阵, 以二分类为例, 表 3 展示了其混淆矩阵. TP 表示正确预测到的正样本个数, TN 表示正确预测到的负样本个数, FN 表示正样本预测为负样本的个数, FP 表示负样本预测为正样本的个数.

表 3 分类混淆矩阵

Table 3 The confuse matrix of classification

混淆矩阵	预测为正样本	预测为负样本
正样本	TP	FN
负样本	FP	TN

常见的不平衡数据分类问题评价指标有: 准确率 (Accuracy, Acc)、敏感性 (Sensibility, Sen)、特异性 (Specificity, Spe)、MCC、F-score 和 Area under curve (AUC), 计算公式如下:

$$Acc = \frac{TN + TP}{TP + TN + FP + FN} \tag{11}$$

$$Sen = \frac{TP}{TP + FN} \tag{12}$$

$$Spe = \frac{TN}{TN + FP} \tag{13}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{14}$$

$$F\text{-}score = 2 \times \frac{precision \times recall}{precision + recal} \tag{15}$$

F-score 综合考虑了查全率与查准率, 是两者的调和平均数, 其值接近其中较小者. 在不平衡中, 只有当查全率与查准率同时较大时, F-score 才会增大. recall 代表查全率, 表示在原始样本的正样本中, 最后被正确预测为正样本的概率, 计算方法与 Sen 相同; precision 为查准率, 表示预测结果中, 正确预测为正样本的概率如下:

$$precision = \frac{TP}{TP + FP} \tag{16}$$

$$recall = \frac{TP}{TP + FN} \tag{17}$$

AUC 是 Receiver operating characteristic (ROC) 曲线下面积, ROC 图由真阳性率 (TP-rate) 与假阳性率 (FP-rate) 作图而成, ROC 空间中的任意一个点对应分类器在给定分布上的性能, 当真阳性率与假阳性率比值越大时, ROC 就将越接近图形左上角, 此时将得到更大的 AUC 值, 这也意味着分类器结果越理想, AUC 也是评价分类器在不平衡数据上性能的重要指标之一.

4 实验与性能分析

为验证本文方法的有效性, 使用 13 个数据集进行实验. 实验中使用随机森林 (Random forest, RF)[33−34]、SVM[35−36]、逻辑回归 (Logistics regression, LR)[37−38] 和决策树 (Decision tree, DT)[39−40] 作为分类器. RF 属于 Bagging 集成的分类器, 由于使用了多个分类器, 效果通常好于使用单个分类器. SVM 在处理小样本高维度的数据时有其特有的优势, 因为 SVM 最终的决策函数由少数支持向量确定, 复杂性仅仅取决于支持向量数目而不是原始的样本空间. LR 计算代价不高且容易实现, 此外, LR 对数据中小噪声具有一定鲁棒性. DT 算法是一种基于概率的分析方法, 在训练时不需要任何领域的先验知识和参数假设, 计算量相对较小且准确性高, 适合用于高维数据.

在分类器参数选择上, 为了最大化突出采样方法自身的特点, 参数均使用默认参数设置: SVM 的惩罚系数为 1, 核方法为径向基函数核 (Radial basis function, RBF), gamma 值为 1; LR 使用 saga 作为求解器; DT 使用基尼系数评价特征划分质量; RF 使用具有随机属性选择的决策树作为基分类器, 包含 50 个独立的决策树, 每棵决策树同样使用基尼系数评价划分质量.

为保证训练效果, 本文使用 5 折交叉验证的方法, 将数据集随机分成 5 份, 每次将其中 4 份作为训练集, 剩下的 1 份作为测试集, 重复 5 次. 最后将 5 次实验评价结果的平均值作为交叉验证的结果. 所有结果均为 5 次 5 折交叉验证结果. 实验硬件环境为 CPU i5-3230m、操作系统为 Windows10、

开发语言为 Python、集成开发环境为 Pycharm、使用外部库 Numpy、Sklearn 和 Imbalancelearn.

4.1 UCI 和真实数据集上分类性能对比

实验设计如下: 首先使用 PNS 算法对数据进行预处理, 然后分别使用四种不同的分类器对处理后的数据进行训练学习. 实验目的是评价不同分类器对不平衡数据的敏感性并为后面实验选择合适的分类器提供参考.

在 7 个 UCI 数据集和 2 个真实数据集上的结果如表 4 所示. 由表 4 可以看出, SVM 在 SPECT、Abalone、SolarFlare、Yeast_ME2、Ecoli 这 5 个数据集的大多数指标上取得最佳值, RF 在 Abalone_19、SatImage、Balance 和 SNP 这 4 个数据集的多数指标上取得最佳值. 除 Ecoli 数据集的 *Spe* 指标, Abalone 的 *Sen* 指标以及 Abalone_19 的 *AUC* 指标以外, 其余最高值均出自 SVM 与 RF. 因此相比 LR 与 DT, SVM 与 RF 具有更好的分类效果. 这说明 SVM 与 RF 对不平衡数据更具有鲁棒性.

RF 使用了决策树的集成方法, 并且随机森林中每棵决策树的特征选择具有一定随机性, 这增大了决策树间的差异, 从而使集成效果更好, 因此 RF 的结果要优于 DT, 集成方法也是解决不平衡的重要方法之一. SVM 使用核方法将数据映射到高维空间进行划分, 而且 SVM 的超平面只与支持向量有关, 与离决策超平面的数据的多少并不重要, 因此使得 SVM 对不平衡本身并不十分敏感. LR 在预测时会考虑所有样本点到决策平面的距离, 虽然使用了非线性函数进行映射, 但也无法很好消除其影响, 因此, 容易受不平衡影响.

由数据集与分类器的特点可以知道, SVM 趋向于在小样本量的不平衡数据集上具有更好的效果, 而 RF 趋向于在大样本量不平衡数据集上表现更佳. 这也恰恰符合 SVM 与 RF 在平衡数据集上的表现, 这说明 PNS 算法已经将原始的不平衡数据有效采样成了更平衡的数据, 起到了平衡数据集, 提高分类器性能的作用.

4.2 采样方法分类性能比较

根据第 4.1 节实验结果, 本节使用的分类器是支持向量机 (SVM) 和逻辑回归 (LR), 因为它们对不平衡数据集具有不同敏感性, SVM 对不平衡数据不敏感而 LR 对不平衡较为敏感, 如果 PNS 算法在这两个分类器上都表现良好, 那么可以推断出 PNS 算法对大多数分类器均具有较好的提升效果.

实验设计如下: 由于伪负样本采样无需指定采样比例, 会根据数据集自适应确定采样比例, 因此本文对比相同采样比例下各采样方法的性能. 首先使用伪负样本采样方法对原始数据进行采样, 得到平衡后的比例, 然后按照平衡后的比例使用对比算法重新对原始数据进行采样得到采样结果, 最后使用 5 折交叉验证对采样数据集进行评价, 并重复 5 次试验取平均值. 在对比实验中, 将本文提出的

表 4 伪负样本采样在分类器 SVM、LR、DT、RF 上的结果

Table 4 Results of pseudo-negative sampling on classifiers including SVM, LR, DT and RF

数据集	分类算法	*Sen*	*Spe*	*Acc*	*MCC*	*F-score*	*AUC*
Balance	SVM	0.810	**0.967**	0.911	0.804	0.860	0.967
	LR	0.638	0.872	0.789	0.525	0.670	0.868
	DT	0.885	0.950	0.928	0.836	0.889	0.920
	RF	**0.887**	0.956	**0.932**	**0.849**	**0.899**	**0.972**
Ecoli	SVM	**0.826**	0.975	**0.952**	**0.806**	**0.828**	**0.982**
	LR	0.746	**0.975**	0.941	0.755	0.781	0.962
	DT	0.741	0.961	0.932	0.704	0.734	0.865
	RF	0.733	0.975	0.938	0.734	0.756	0.963
SatImage	SVM	**0.924**	0.917	0.919	0.830	0.892	0.980
	LR	0.823	0.827	0.825	0.636	0.772	0.913
	DT	0.847	0.908	0.886	0.754	0.842	0.877
	RF	0.901	**0.950**	**0.933**	**0.854**	**0.906**	**0.984**
Abalone	SVM	0.906	**0.994**	**0.965**	**0.922**	**0.945**	0.966
	LR	0.903	0.978	0.954	0.895	0.928	0.973
	DT	**0.914**	0.949	0.937	0.860	0.906	0.932
	RF	0.904	0.991	0.962	0.916	0.941	**0.981**
SolarFlare	SVM	0.917	**0.976**	**0.954**	**0.901**	**0.936**	0.984
	LR	0.934	0.962	0.951	0.896	0.934	0.973
	DT	0.922	0.956	0.943	0.880	0.924	0.940
	RF	**0.942**	0.957	0.951	0.897	0.935	**0.987**
Yeast_ME2	SVM	**0.757**	**0.982**	**0.946**	**0.791**	**0.818**	**0.976**
	LR	0.573	0.966	0.902	0.608	0.653	0.947
	DT	0.735	0.946	0.911	0.675	0.724	0.843
	RF	0.723	0.976	0.935	0.749	0.782	0.968
Abalone_19	SVM	0.969	0.989	0.982	0.962	0.975	0.996
	LR	0.971	0.984	0.979	0.956	0.971	**0.997**
	DT	0.976	0.982	0.980	0.957	0.972	0.979
	RF	**0.977**	**0.992**	**0.987**	**0.972**	**0.982**	0.997
SPECT	SVM	**0.767**	**0.907**	**0.862**	**0.682**	**0.774**	**0.941**
	LR	0.732	0.862	0.816	0.586	0.707	0.909
	DT	0.627	0.817	0.753	0.440	0.608	0.732
	RF	0.674	0.931	0.846	0.637	0.725	0.929
SNP	SVM	0.677	**0.980**	0.850	0.709	0.795	0.966
	LR	0.692	0.961	0.845	0.693	0.793	0.902
	DT	0.892	0.911	0.903	0.803	0.888	0.902
	RF	**0.900**	0.958	**0.933**	**0.864**	**0.920**	**0.971**

PNS 算法与 4 种数据采样方法进行对比. 对比算法包括 ROS、RUS、SMOTE 和 ADASYN. 结果如表 5 所示.

由表 5 可以看出, PNS 算法具有最好的综合性能. F-score、MCC 和 AUC 被认为是在类别不平衡情况下的综合评价指标. 它们综合了正样本正确率和负样本正确率, 能客观评价不平衡分类器的性能. 在这 3 个指标上使用 SVM 分类器时, 算法在 SPECT、Ecoli、SatImage、Abalone、Balance、SolarFlare、Yeast_ME2 和 Abalone_19 数据集上取得了最好的结果. 而在 SNP 数据集上, 则是 ADASYN 算法取得了较好的结果, 这是因为它们合成的

表 5　伪负样本采样与 ROS, RUS, SMOTE, ADASYN 采样方法对比结果

Table 5　Comparison of pseudo-negative sampling with the methods of ROS、RUS、SMOTE、ADASYN

数据集	评价指标	SVM					LR				
		PNS	ROS	RUS	SMOTE	ADASYN	PNS	ROS	RUS	SMOTE	ADASYN
SPECT	*Sen*	**0.767**	0.746	0.594	0.381	0.438	**0.732**	0.685	0.605	0.643	0.604
	Spe	0.907	0.856	0.860	**0.985**	0.970	**0.862**	0.846	0.828	0.838	0.843
	Acc	**0.862**	0.817	0.760	0.794	0.789	**0.816**	0.793	0.748	0.768	0.751
	MCC	**0.682**	0.590	0.461	0.509	0.531	**0.586**	0.527	0.432	0.507	0.485
	F-score	**0.774**	0.715	0.585	0.535	0.575	**0.707**	0.667	0.594	0.622	0.611
	AUC	**0.941**	0.912	0.861	0.857	0.867	**0.909**	0.889	0.848	0.849	0.824
SNP	*Sen*	0.677	0.842	0.489	0.879	**0.879**	**0.692**	0.614	0.605	0.653	0.637
	Spe	**0.980**	0.908	0.869	0.904	0.897	**0.961**	0.847	0.801	0.852	0.852
	Acc	0.850	0.880	0.705	0.893	**0.889**	**0.845**	0.747	0.713	0.766	0.760
	MCC	0.709	0.754	0.394	0.782	**0.775**	**0.693**	0.479	0.416	0.520	0.505
	F-score	0.795	0.857	0.585	0.876	**0.871**	**0.793**	0.676	0.643	0.706	0.693
	AUC	**0.966**	0.935	0.761	0.949	0.947	**0.902**	0.809	0.765	0.839	0.832
Ecoli	*Sen*	**0.826**	0.715	0.644	0.720	0.661	**0.746**	0.644	0.616	0.610	0.573
	Spe	**0.975**	0.962	0.964	0.963	0.956	**0.975**	0.958	0.954	0.962	0.956
	Acc	**0.952**	0.925	0.916	0.925	0.908	**0.941**	0.908	0.902	0.908	0.900
	MCC	**0.806**	0.693	0.633	0.692	0.623	**0.755**	0.618	0.598	0.612	0.570
	F-score	**0.828**	0.728	0.665	0.727	0.664	**0.781**	0.655	0.634	0.647	0.616
	AUC	**0.982**	0.958	0.949	0.957	0.951	**0.962**	0.936	0.923	0.935	0.930
SatImage	*Sen*	0.924	0.892	0.847	0.915	**0.933**	**0.823**	0.580	0.540	0.595	0.553
	Spe	**0.917**	0.904	0.898	0.907	0.871	**0.827**	0.763	0.747	0.766	0.757
	Acc	**0.919**	0.899	0.879	0.910	0.893	**0.825**	0.697	0.671	0.704	0.683
	MCC	**0.830**	0.786	0.741	0.810	0.784	**0.636**	0.344	0.288	0.361	0.312
	F-score	**0.892**	0.865	0.835	0.880	0.864	**0.772**	0.580	0.539	0.591	0.557
	AUC	**0.980**	0.960	0.946	0.966	0.953	**0.913**	0.778	0.756	0.786	0.768
Abalone	*Sen*	**0.906**	0.721	0.651	0.740	0.703	**0.903**	0.726	0.710	0.735	0.697
	Spe	**0.994**	0.835	0.839	0.830	0.822	**0.978**	0.805	0.802	0.804	0.804
	Acc	**0.965**	0.797	0.776	0.800	0.783	**0.954**	0.779	0.769	0.781	0.769
	MCC	**0.922**	0.549	0.493	0.559	0.515	**0.895**	0.518	0.499	0.525	0.489
	F-score	**0.945**	0.701	0.655	0.709	0.676	**0.928**	0.684	0.669	0.689	0.660
	AUC	**0.966**	0.868	0.840	0.876	0.861	**0.973**	0.850	0.842	0.850	0.836
Balance	*Sen*	0.810	**0.937**	0.619	0.517	0.510	0.638	0.605	0.597	**0.693**	0.518
	Spe	**0.967**	0.775	0.776	0.943	0.940	0.872	0.812	0.778	0.851	**0.962**
	Acc	**0.911**	0.827	0.705	0.798	0.791	0.789	0.740	0.704	0.795	**0.811**
	MCC	**0.804**	0.674	0.385	0.558	0.554	0.525	0.418	0.364	0.549	**0.584**
	F-score	**0.860**	0.783	0.564	0.624	0.627	0.670	0.608	0.565	**0.694**	0.646
	AUC	**0.967**	0.902	0.834	0.884	0.826	0.868	0.831	0.833	**0.902**	0.872

表 5 伪负样本采样与 ROS, RUS, SMOTE, ADASYN 采样方法对比结果 (续表)

Table 5 Comparison of pseudo-negative sampling with the methods of ROS、RUS、SMOTE、ADASYN (continued table)

数据集	评价指标	SVM					LR				
		PNS	ROS	RUS	SMOTE	ADASYN	PNS	ROS	RUS	SMOTE	ADASYN
SolarFlare	*Sen*	**0.917**	0.821	0.528	0.882	0.883	**0.934**	0.599	0.602	0.866	0.860
	Spe	0.976	0.888	0.866	**0.979**	0.973	0.962	0.853	0.824	**0.988**	0.985
	Acc	**0.954**	0.862	0.734	0.943	0.940	**0.951**	0.758	0.734	0.942	0.939
	MCC	**0.901**	0.707	0.418	0.878	0.871	**0.896**	0.470	0.433	0.878	0.870
	F-score	**0.936**	0.815	0.583	0.919	0.915	**0.934**	0.647	0.620	0.917	0.912
	AUC	**0.984**	0.912	0.802	0.969	0.968	**0.973**	0.837	0.790	0.970	0.968
Yeast_ME2	Sen	**0.757**	0.708	0.482	0.721	0.688	0.573	0.548	0.538	**0.633**	0.575
	Spe	**0.982**	0.965	0.970	0.967	0.966	**0.967**	0.958	0.959	0.960	0.960
	Acc	**0.946**	0.923	0.889	0.927	0.920	0.902	0.892	0.884	**0.906**	0.896
	MCC	**0.791**	0.706	0.545	0.720	0.695	0.608	0.566	0.545	**0.634**	0.593
	F-score	**0.818**	0.747	0.575	0.759	0.738	0.653	0.618	0.584	**0.683**	0.643
	AUC	**0.976**	0.955	0.882	0.961	0.955	**0.947**	0.901	0.891	0.910	0.901
Abalone_19	*Sen*	**0.969**	0.885	0.315	0.947	0.948	**0.971**	0.636	0.538	0.725	0.725
	Spe	**0.989**	0.872	0.830	0.877	0.875	**0.984**	0.863	0.829	0.865	0.867
	Acc	**0.982**	0.877	0.613	0.902	0.902	**0.979**	0.780	0.698	0.814	0.815
	MCC	**0.962**	0.743	0.138	0.803	0.802	**0.956**	0.516	0.380	0.595	0.598
	F-score	**0.975**	0.839	0.299	0.876	0.875	**0.971**	0.677	0.539	0.739	0.740
	AUC	**0.996**	0.947	0.715	0.956	0.956	**0.997**	0.877	0.815	0.891	0.893

样本扩充了少数类, 同时未减少多数类样本, 使其有更高的 Sen 值, 但是与 PNS 相比, 它们的 Spe 值更低, 这说明它们是通过牺牲 Spe 来提高其他性能指标的.

当使用 LR 分类器时, PNS 算法在 SPECT、Ecoli、SNP、SatImage、Abalone、SolarFlare、Abalone_19 数据集上取得了最好的结果. 在 Balance 数据集上分别是 SMOTE 和 ADASYN 算法得到较好结果, 这是因为过少的特征数不利于伪负样本的选择, 因此无法准确找到所有伪负样本导致数据没有得到很好的平衡, 同时 LR 分类器对不平衡数据较为敏感. 在 Yeast_ME2 数据集上所得结果与 SVM 分类器 SNP 数据集结果原因类似.

在不平衡数据集的分类当中, 少数类的正确率 (即 Sen) 往往受到更多重视, 因为少数类通常受到更多关注而 Sen 则反映了分类器发现少数类的能力. 在 Sen 指标下, PNS 采样算法在 SVM 的 6 个数据集和 LR 的 7 个数据集上取得最好结果, 这表明本文提出的算法对少数类具有很强的辨别能力. 从侧面也证实了, 分类正确率作为不平衡数据分类的评价指标有时并不能有效地衡量分类器的分类效果.

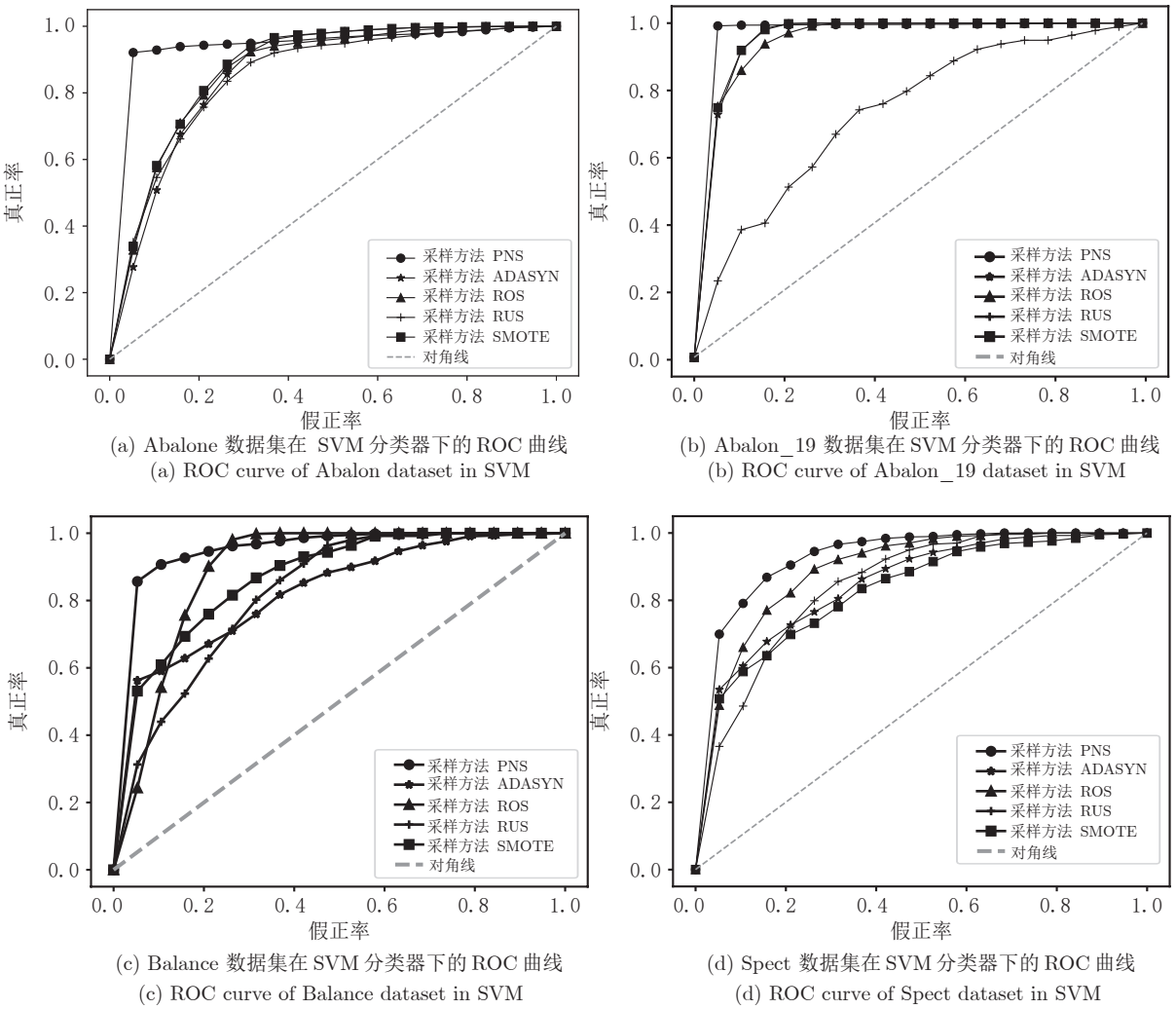
图 2 给出了 4 个数据集在 SVM 分类器下不同采样方法的 ROC 曲线. 由图可知, PNS 采样算法在 4 个数据集上拥有更好的 ROC 曲线, 曲线下面积均大于其他采样方法, 证明了该方法的优越性.

综上所述, PNS 采样算法相比 ADASYN、ROS、SMOTE、RUS 算法, 对数据具有更好的适应性, 因为 PNS 考虑了数据集的样本分布, 从根本上缓解了不平衡数据少数类被忽略的问题, 并且在提高少数类正确率的同时, 其他指标保持不变, 因此从整体上提高了分类器的性能. 此外, 由于 PNS 在对不平衡数据具有不同敏感性的 SVM 与 LR 分类器上取得最好结果, 说明了 PNS 的采样结果可以适用于多数分类器.

4.3 高不平衡比例数据对比分析

本节选择不平衡比例大于 20 的 KEEL 数据, 将所提出的 PNS 方法与 ROS、RUS、SMOTE 和 ADASYN 进行比较, 以验证 PNS 采样方法在处理高不平衡数据时的有效性. 实验设计思路和所用分类器与第 4.2 节相同. 实验结果如表 6 所示.

由表 6 可以看出, PNS 在处理高不平衡数据时, 是具有竞争力的方法. 与其他 4 种采样方法相比, PNS 在 4 个数据的绝大多数评价指标上取得了最好结果. 在只考虑 F-score、MCC 和 AUC 这 3 个指标时, PNS 采样在 SVM 分类器和 LR 分类器的 4 个数据集上获得了最好结果. 以 Yeast1289vs7



(a) Abalone 数据集在 SVM 分类器下的 ROC 曲线
(a) ROC curve of Abalon dataset in SVM

(b) Abalon_19 数据集在 SVM 分类器下的 ROC 曲线
(b) ROC curve of Abalon_19 dataset in SVM

(c) Balance 数据集在 SVM 分类器下的 ROC 曲线
(c) ROC curve of Balance dataset in SVM

(d) Spect 数据集在 SVM 分类器下的 ROC 曲线
(d) ROC curve of Spect dataset in SVM

图 2　4 个 UCI 数据集在 SVM 分类器下的 ROC 曲线
Fig. 2　ROC curve of four UCI datasets in SVM

数据集为例, 在 SVM 分类器上的 F-score、MCC 和 AUC 值分别为 0.909、0.848 和 0.980; 在 LR 分类器上的值分别为 0.780、0.627 和 0.902. 均优于其他采样方法, 这充分说明了 PNS 在处理高不平衡比例数据时具有较好的综合性能. 在考虑 Sen 作为评价指标时, PNS 采样算法在 SVM 的 3 个数据集和 LR 的 2 个数据集上得到最好结果. 说明 PNS 在高不平衡比例数据中依然能很好识别出少数类样本.

此外, 图 3 给出了 Yeast1289vs7 和 Yeast1458vs7 两个数据集在 SVM 分类器下不同采样方法的 ROC 曲线. 由图 3 可知, 相较于对比算法, PNS 的 ROC 曲线拥有更大的曲线下面积, 其次是 SMOTE、ADASYN 和 ROS, 最后是 RUS. 由于 RUS 移除了大量样本, 使得分类器对数据集学习不能很好学习, 从而导致欠拟合. SMOTE、ADASYN 和 ROS 方法生成的样本可能存在噪音或异常值, 导致分类效果不如 PNS. 这说明 PNS 不改变数据集样本数量是一种性能更加优秀的采样方法.

4.4　采样方法时间性能对比

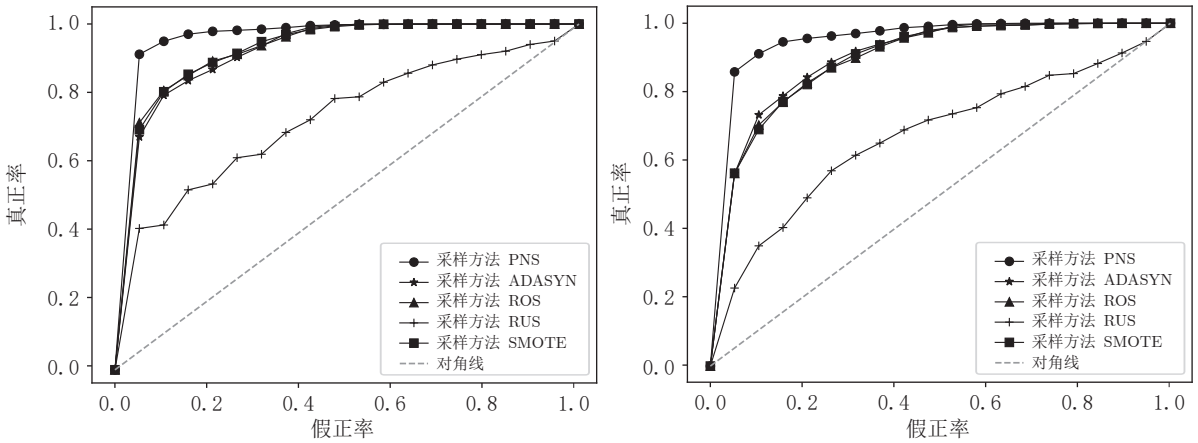
本文算法的另一个优势是相对较少的训练时间. 表 7 展示了不同采样方法在 UCI 数据集上的时间消耗对比.

过采样为了平衡数据集会增加少数类样本数量, 当正负样本比例越大同时需要越平衡的数据集时, 过采样将会生成大量的新样本, 这将显著增加训练所需时间, 并且大量的合成样本可能导致过拟合现象. 同时, 相对于欠采样而言, 欠采样去掉多数类样本, 使训练时间缩短, 但是当少数类样本很少时, 欠采样往往会删除大部分多数类, 这会导致严重的训练不足, 分类器无法很好的学习数据, 从而使训练效果不尽人意.

表 6 高比例不平衡数据采样对比

Table 6 The comparison of high ratio imbalanced data

数据集	评价指标	SVM					LR				
		PNS	ROS	RUS	SMOTE	ADASYN	PNS	ROS	RUS	SMOTE	ADASYN
Yeast1289vs7	*Sen*	**0.892**	0.752	0.533	0.845	0.843	**0.775**	0.691	0.558	0.726	0.719
	Spe	**0.952**	0.919	0.833	0.860	0.844	**0.850**	0.824	0.786	0.815	0.809
	Acc	**0.925**	0.849	0.695	0.853	0.843	**0.817**	0.768	0.668	0.777	0.771
	MCC	**0.848**	0.690	0.392	0.701	0.682	**0.627**	0.521	0.355	0.542	0.529
	F-score	**0.909**	0.806	0.582	0.827	0.817	**0.780**	0.712	0.570	0.731	0.723
	AUC	**0.980**	0.935	0.793	0.930	0.926	**0.902**	0.837	0.793	0.848	0.844
Yeast1458vs7	*Sen*	**0.855**	0.681	0.356	0.713	0.737	0.590	0.503	0.415	0.570	**0.592**
	Spe	**0.934**	0.899	0.879	0.877	0.870	0.835	**0.843**	0.829	0.823	0.820
	Acc	**0.904**	0.820	0.684	0.817	0.821	**0.745**	0.719	0.660	0.731	0.735
	MCC	**0.794**	0.602	0.283	0.599	0.612	**0.437**	0.369	0.265	0.406	0.421
	F-score	**0.866**	0.730	0.431	0.736	0.748	**0.623**	0.562	0.445	0.602	0.617
	AUC	**0.965**	0.904	0.720	0.897	0.899	**0.822**	0.769	0.744	0.792	0.794
Yeast4	*Sen*	**0.770**	0.687	0.543	0.733	0.703	0.574	0.572	0.558	**0.603**	0.566
	Spe	**0.982**	0.969	0.965	0.970	0.966	**0.968**	0.958	0.955	0.959	0.960
	Acc	**0.947**	0.923	0.892	0.930	0.923	**0.904**	0.895	0.886	0.902	0.895
	MCC	**0.798**	0.701	0.571	0.734	0.706	**0.613**	0.582	0.559	0.611	0.584
	F-score	**0.824**	0.741	0.609	0.770	0.747	**0.662**	0.634	0.605	0.656	0.635
	AUC	**0.976**	0.954	0.908	0.961	0.957	**0.946**	0.902	0.881	0.906	0.903
Yeast5	*Sen*	0.704	0.706	0.596	**0.745**	0.721	**0.622**	0.576	0.559	0.590	0.546
	Spe	**0.995**	0.989	0.990	0.991	0.990	0.987	0.987	**0.988**	0.987	0.988
	Acc	**0.980**	0.975	0.970	0.979	0.976	**0.969**	0.966	0.966	0.967	0.967
	MCC	**0.770**	0.714	0.644	0.759	0.728	**0.642**	0.605	0.590	0.614	0.588
	F-score	**0.772**	0.720	0.641	0.765	0.734	**0.647**	0.609	0.587	0.620	0.593
	AUC	**0.994**	0.990	0.986	0.991	0.992	**0.988**	0.988	0.988	0.988	0.988



(a)Yeast1289vs7 数据集在 SVM 分类器下的 ROC 曲线

(a) ROC curve of Yeast1289vs7 dataset in SVM

(b) Yeast1458vs7 数据集在 SVM 分类器下的 ROC 曲线

(b) ROC curve of Yeast1458vs7 dataset in SVM

图 3 2 个 KEEL 数据集在 SVM 分类器下的 ROC 曲线

Fig. 3 ROC curve of two KEEL datasets in SVM classifier

相比于上述采样方法, 本文所提出的采样方法 PNS 则不改变原始样本集的数量, 仅改变了数据分布, 不会因为引入数据而增加时间成本, 也不会删除数据而导致训练不充分, 所以具有较好的结果.

表 7 不同采样方法时间对比
Table 7 Runtime comparison of different sampling methods

数据集	算法	RUS	PNS	SMOTE	ROS	ADASYN
SPECT	SVM	**0.39**	0.53	0.67	0.66	0.71
	LR	**0.56**	0.69	0.80	0.75	0.81
	DT	**0.26**	0.31	0.35	0.32	0.34
	RF	**1.70**	1.77	1.91	1.84	1.98
SNP	SVM	**1.30**	27.92	80.22	92.04	80.74
	LR	**0.70**	1.41	2.16	2.09	2.26
	DT	**0.55**	1.29	2.51	1.55	2.61
	RF	**2.32**	7.32	13.76	9.45	13.91
Ecoli	SVM	**0.31**	0.31	0.36	0.34	0.39
	LR	**0.39**	0.43	0.44	0.44	0.44
	DT	0.23	**0.23**	0.23	0.23	0.24
	RF	1.54	**1.58**	1.56	1.56	1.58
SatImage	SVM	**7.59**	75.68	189.22	201.02	238.91
	LR	**3.00**	6.60	5.94	5.05	6.64
	DT	**1.02**	2.75	4.03	3.47	4.86
	RF	**4.43**	13.48	18.02	16.36	19.92
Abalone	SVM	**3.08**	14.78	62.42	64.35	65.56
	LR	**1.02**	3.58	4.74	4.67	4.81
	DT	**0.52**	0.74	1.31	1.03	1.37
	RF	**2.86**	4.75	9.61	7.73	9.48
Balance	SVM	**0.28**	0.73	1.32	1.58	1.29
	LR	**0.25**	0.35	0.68	0.38	0.68
	DT	**0.22**	0.24	0.27	0.24	0.27
	RF	**1.49**	1.67	1.74	1.73	1.76
SolarFlare	SVM	**0.44**	3.46	9.25	12.31	9.30
	LR	**0.40**	2.00	3.17	2.96	3.17
	DT	**0.29**	0.36	0.46	0.43	0.50
	RF	**1.61**	2.14	2.59	2.57	2.66
Yeast_ME2	SVM	**0.44**	1.84	2.95	3.189	3.161
	LR	**0.44**	0.74	0.86	0.871	0.933
	DT	**0.29**	0.36	0.38	0.361	0.436
	RF	**1.65**	2.24	2.45	2.269	2.452
Abalone_19	SVM	**0.44**	6.81	66.16	75.09	66.20
	LR	**0.46**	3.54	7.06	4.71	4.86
	DT	**0.39**	0.71	1.49	0.86	1.47
	RF	**1.65**	4.45	10.48	5.64	10.18
	总计	**44.69**	197.95	511.77	530.30	567.05

表 7 是各采样方法在不同数据集上使用不同分类器的算法运行时间, 每次实验均为 5 次 5 折交叉验证时间总和, 时间单位为秒.

由表 7 可以看出, RUS 的总计用时最少, ADASYN 的总计用时最多, 分别为 44.692 秒和 567.057 秒. PNS、SMO-TE 和 ROS 的用时分别为 197.954 秒、511.770 秒和 530.303 秒. 由于同属于过采样, 所以 ADASYN、S-MOTE 与 ROS 所用时间处在同一个量级. 使用过采样平衡数据时, 时间成本的增加在所难免, 而随着不平衡比例的增大, 时间成本也会相应增大, 这不利于处理极度不平衡数据. 欠采样虽然减少了时间开销, 但是不能得到满意结果. PNS 方法很好地解决了上述问题, 在不增加时间成本的同时提高分类器性能, 将时间花销控制在可接受范围.

5 结束语

本文提出了一种新型的基于样本空间的不平衡数据采样方法, 即伪负样本采样方法 PNS. 实验结果显示, PNS 采样方法普遍优于其他常用数据采样方法. 在不平衡数据集中由于存在大量负样本, 使有的负样本与正样本具有相似的分布, 与正样本具有很高相似度, 可以将其定义为被错分的正样本, 基于这一考虑本文提出了伪负样本的概念及其采样方法. 具体地, PNS 使用欧几里得距离衡量正负样本间的相似性, 将得到的伪负样本从负样本中删除并加入到正样本中. 本文方法根据样本的空间分布自适应地对数据进行采样, 不需要指定采样比例, 具有较强的适应性, 避免了采样时选择采样比例的困难. 混合采样方法避免了单独使用一种采样方法带来的问题. 此外, 该算法还具有良好的时间复杂性, 采样与训练时间明显少于过采样方法. 因此, PNS 采样方法为处理不平衡数据提供了一种可行的新思路.

未来工作包括: 1) 将本文提出的伪负样本算法与聚类算法结合[41−43], 使用聚类方法获得数据集的更多分布信息, 这将有助于提高采样的精准性; 2) 探索将现有的算法扩展到多分类的任务; 3) 将算法应用于大规模数据集.

References

1 Hou J, Shi X, Chen C, Islam MS, Johnson AF, Kanno T, et al. Global impacts of chromosomal imbalance on gene expression in arabidopsis and other taxa. *Proceedings of the National Academy of Sciences*, 2018, **115**(48): 11321−11330

2 Zhang Y, Qiao S, Ji S, Han N, Liu D, Zhou J. Identification of DNA-protein binding sites by bootstrap multiple convolutional neural networks on sequence information. *Engineering Applications of Artificial Intelligence*, 2019, **79**: 58−66

3 Zhao Z, Peng H, Lan C, Zheng Y, Fang L, Li J. Imbalance learning for the prediction of N 6-methylation sites in mRNAs. *BMC Genomics*, 2018, **19**(1): 574

4 Du X, Yao Y, Diao Y, Zhu H, Zhang Y, Li S. Deepss: Exploring splice site motif through convolutional neural network directly from dna sequence. *IEEE Access*, 2018, **6**: 32958−32978

5 Maji R K, Khatua S, Ghosh Z. A supervised ensemble approach

for sensitive microRNA target prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020, **17**(1): 37−46

6 Zhang X, Lin X, Zhao J, Huang Q, Xu X. Efficiently predicting hot spots in PPIs by combining random forest and synthetic minority over-sampling technique. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2018, **16**(3): 774−781

7 Luo K, Wang G, Li Q, Tao J. An improved SVM-RFE based on *F*-statistic and mPDC for gene selection in cancer classification. *IEEE Access*, 2019, **7**: 147617−147628

8 Fotouhi S, Asadi S, Kattan M W. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of Biomedical Informatics*, 2019, **90**: 103089

9 Soh W W, Yusuf R M. Predicting credit card fraud on a imbalanced data. *International Journal of Data Science and Advanced Analytics*, 2019, **1**(1): 12−17

10 Zhang Hong-Li, Lu Gang. Machine learning algorithms for classifying the imbalanced protocol flows: Evaluation and comparison. *Journal of Software*, 2012, **23**(6): 1500−1516
(张宏莉, 鲁刚. 分类不平衡协议流的机器学习算法评估与比较. 软件学报, 2012, **23**(6): 1500−1516)

11 He H, Garcia E A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 2009, **21**(9): 1263−1284

12 Lin Shu-Yang, Li Cui-Hua, Jiang Yi, Lin Chen, Zou Quan. Under-sampling method research in class-imbalanced data. *Journal of Computer Research Development*, 2011, **48**(S3): 47−53
(林舒杨, 李翠华, 江弋, 林琛, 邹权. 不平衡数据的降采样方法研究. 计算机研究与发展, 2011, **48**(S3): 47−53)

13 Zhang Y Q, Qiao S J, Lu R, Zhao R, Han N, Liu D. How to balance the bioinformatics data: Pseudo-negative sampling. *BMC Bioinformatics*, 2019, **20**(25): 1−13

14 Liu D, Qiao S, Han N, Wu T, Mao R. SOTB: Semi-supervised oversampling approach based on trigonal barycenter theory. *IEEE Access*, 2020, **8**: 50180−50189

15 Jiang Sheng-Yi, Xie Zhao-Qing, Yu Wen. Naive Bayes classification algorithm based on cost sensitive for imbalanced data distribution. *Journal of Computer Research Development*, 2011, **48**(S1): 387−390
(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**(S1): 387−390)

16 Yu L, Zhou R, Tang L, Chen R. A DBN-based resampling SVM ensemble learning paradigm for credit classification with imbalanced data. *Applied Soft Computing*, 2018, **69**: 192−202

17 Castellanos F J, Valero-Mas J J, Calvo-Zaragoza J, Rico-Juan J R. Oversampling imbalanced data in the string space. *Pattern Recognition Letters*, 2018, **103**: 32−38

18 Sun B, Chen H, Wang J, Xie H. Evolutionary under-sampling based bagging ensemble method for imbalanced data classification. *Frontiers of Computer Science*, 2018, **12**(2): 331−350

19 Chawla N V, Bowyer K W, Hall L O, Kegelmeyer W P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 2002, **16**: 321−357

20 Douzas G, Bacao F. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications*, 2018, **91**: 464−471

21 Wilson D L. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, 1972, **SMC-2**(3): 408−421

22 Laurikkala J. Improving identification of difficult small classes by balancing class distribution. In: Proceedings of the 2001 Conference on Artificial Intelligence in Medicine in Europe. Berlin, Ger-many: 2001. 63−66

23 Zhang Z L, Luo X G, García S, Herrera F. Cost-sensitive back-propagation neural networks with binarization techniques in addressing multi-class problems and non-competent classifiers. *Applied Soft Computing*, 2017, **56**: 357−367

24 Liu N, Shen J, Xu M, Gan D, Qi ES, Gao B. Improved cost-sensitive support vector machine classifier for breast cancer diagnosis. *Mathematical Problems in Engineering*, 2018, **4**: 1−13

25 Breiman L. Bagging predictors. *Machine Learning*, 1996, **24**(2): 123−140

26 Schapire R E. The strength of weak learnability. *Machine Learning*, 1990, **5**(2): 197−227

27 Friedman J, Hastie T, Tibshirani R. Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 2000, **28**(2): 337−407

28 Elmore K L, Richman M B. Euclidean distance as a similarity metric for principal component analysis. *Monthly Weather Review*, 2001, **129**(3): 540−549

29 Park M W, Lee E C. Similarity measurement method between two songs by using the conditional Euclidean distance. *Wseas Transaction on Information Science and Applications*, 2013, **10**(12): 381−388

30 He H, Bai Y, Garcia E A, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: Proceedings of the 2008 International Joint Conference on Neural Networks (World Congress on Computational Intelligence). Hong Kong, China: IEEE, 2008. 1322−1328

31 Fernández A, del Río S, Chawla N V, Herrera F. An insight into imbalanced big data classification: Outcomes and challenges. *Complex & Intelligent Systems*, 2017, **3**(2): 105−120

32 Alcalá-Fdez J, Sanchez L, Garcia S, Deljesus M J, Ventura S, Garrell J M, et al. KEEL: A software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, 2009, **13**(3): 307−318

33 Luo Zhen-Zhen, Chen Jing-Ying, Liu Le-Yuan, Zhang Kun. Conditional random forests for spontaneous smile detection in unconstrained environment. *Acta Automatica Sinica*, 2018, **44**(4): 696−706
(罗珍珍, 陈靓影, 刘乐元, 张坤. 基于条件随机森林的非约束环境自然笑脸检测. 自动化学报, 2018, **44**(4): 696−706)

34 Breiman L. Random forests. *Machine Learning*, 2001, **45**(1): 5−32

35 Zhang Xue-Gong. Introduction to statistical learning theory and support vector machines. *Acta Automatica Sinica*, 2000, **26**(1): 32−42
(张学工. 关于统计学习理论与支持向量机. 自动化学报, 2000, **26**(1): 32−42)

36 Cortes C, Vapnik V. Support-vector networks. *Machine Learning*, 1995, **20**(3): 273−297

37 Cox D R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1958, **20**(2): 215−232

38 Mao Yi, Chen Wen-Lin, Guo Bao-Long, Chen Yi-Xin. A novel logistic regression model based on density estimation. *Acta Automatica Sinica*, 2014, **40**(1): 62−72
(毛毅, 陈稳霖, 郭宝龙, 陈一昕. 基于密度估计的逻辑回归模型. 自动化学报, 2014, **40**(1): 62−72)

39 Quinlan J R. Induction of decision trees. *Machine Learning*,

1986, **1**(1): 81−106

40 Wang Xue-Song, Pan Jie, Cheng Yu-Hu, Cao Ge. Self-adaptive transfer for decision trees based on similarity metric. *Acta Automatica Sinica*, 2013, **39**(12): 2186−2192
(王雪松, 潘杰, 程玉虎, 曹戈. 基于相似度衡量的决策树自适应迁移. 自动化学报, 2013, **39**(12): 2186−2192)

41 Qiao Shao-Jie, Jin Kun, Han Nan, Tang Chang-Jie, Gesang Duo-Ji, Gutierrez Louis Alberto. Trajectory prediction algorithm based on Gaussian mixture model. *Journal of Software*, 2015, **26**(5): 1048−1063
(乔少杰, 金琨, 韩楠, 唐常杰, 格桑多吉, Gutierrez Louis Alberto. 一种基于高斯混合模型的轨迹预测算法. 软件学报, 2015, **26**(5): 1048−1063)

42 Qiao Shao-Jie, Han Nan, Ding Zhi-Ming, Jin Che-Qing, Sun Wei-Wei, Shu Hong-Ping. A multiple-motion-pattern trajectory prediction model for uncertain moving objects. *Acta Automatica Sinica*, 2018, **44**(4): 608−618
(乔少杰, 韩楠, 丁治明, 金澈清, 孙未未, 舒红平. 多模式移动对象不确定性轨迹预测模型. 自动化学报, 2018, **44**(4): 608−618)

43 Qiao Shao-Jie, Guo Jun, Han Nan, Zhang Xiao-Song, Yuan Chang-An, Tang Chang-Jie. Parallel algorithm for discovering communities in large-scale complex networks. *Chinese Journal of Computers*, 2017, **40**(3): 687−700
(乔少杰, 郭俊, 韩楠, 张小松, 元昌安, 唐常杰. 大规模复杂网络社区并行发现算法. 计算机学报, 2017, **40**(3): 687−700)

张永清 成都信息工程大学计算机学院副教授. 2016 年获四川大学计算机学院博士学位. 主要研究方向为人工智能和生物信息学.
E-mail: zhangyq@cuit.edu.cn
(**ZHANG Yong-Qing** Associate professor at the School of Computer Science, Chengdu University of Information Technology. He received his Ph.D. degree from the College of Computer Science, Sichuan University in 2016. His research interest covers artificial intelligence and bioinformatics.)

卢荣钊 成都信息工程大学计算机学院硕士研究生. 主要研究方向为机器学习. E-mail: 15928652663@163.com
(**LU Rong-Zhao** Master student at the School of Computer Science, Chengdu University of Information Technology. His main research interest is machine learning.)

乔少杰 成都信息工程大学软件工程学院教授. 2009 年获四川大学博士学位. 主要研究方向为轨迹预测, 移动对象数据库和机器学习. 本文通信作者. E-mail: sjqiao@cuit.edu.cn
(**QIAO Shao-Jie** Professor at the School of Software Engineering, Chengdu University of Information Technology. He received his Ph.D. degree from Sichuan University in 2009. His research interest covers trajectory prediction, moving objects databases, and machine learning. Corresponding author of this paper.)

韩 楠 成都信息工程大学管理学院副教授. 2012 年获成都中医药大学博士学位. 主要研究方向为数据挖掘和人工智能.
E-mail: hannan@cuit.edu.cn
(**HAN Nan** Associate professor at the School of Management, Chengdu University of Information Technology. She received her Ph.D. degree from Chengdu University of Traditional Chinese Medicine in 2012. Her research interest covers data mining and artificial intelligence.)

GUTIERREZ Louis Alberto 伦斯勒理工学院计算机科学系研究员. 主要研究方向为数据挖掘.
E-mail: louisgutierrez2002@gmail.com
(**GUTIERREZ Louis Alberto** Professor in the Department of Computer Science, Rensselaer Polytechnic Institute. His main research interest is data mining.)

周激流 成都信息工程大学计算机学院教授. 主要研究方向为智能计算和图像处理.
E-mail: zhoujl@cuit.edu.cn
(**ZHOU Ji-Liu** Professor at the School of Computer Science, Chengdu University of Information Technology. His research interest covers intelligent computing and image processing.)