样本自适应的不平衡分类器

才子昕 王馨月 徐 剑 景丽萍
(北京交通大学交通数据分析与挖掘北京市重点实验室 北京 100044)

摘 要 大数据时代,不平衡数据分类在实际应用场景中频繁出现。以二分类为例,传统分类器由于较难学习少数类数据集内部的本质结构,容易将少数类样本错误分类。针对这一问题,一种有效的解决方法是在传统的方法中引入代价敏感机制,为少数类样本赋予更高的误分代价以提升其预测精度。这类方法同等对待了同类样本集中的数据,然而同一类内的不同样本可能对训练过程有不同程度的贡献。为了提升代价敏感机制的有效性,样本自适应的代价敏感策略为不同的样本赋予不同的权重。首先,通过考察样本局部的类分布情况,判断其距离两类样本边界的远近;然后,根据边界分布理论,即距离决策面越近的样本对决策面位置的影响越大,为距离两类样本边界越近的样本赋予越高的权重。实验过程中,通过将样本自适应代价敏感策略应用于LDM,并在标准数据集上进行一系列对比实验,验证了样本自适应代价敏感策略在处理不平衡数据分类问题上的有效性。

关键词 分类,代价敏感学习,边界样本
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.01.014

Sample Adaptive Classifier for Imbalanced Data

CAI Zi-xin WANG Xin-yue XU Jian JING Li-ping
(Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China)

Abstract In the era of big data, the imbalanced data is ubiquitous and inevitable, which has been a critical classification issue. Taking binary classification as an example, traditional learning algorithms can not sufficiently learn the hidden patterns from the minority class and may be biased towards majority class. To solve this problem, an effective way is using the cost-sensitive learning to improve the performance of prediction for the minority class which assigns a highercost to misclassification of the minority. However, these methods equally treat the instances within one class. Actually, different instances may make different contributions to learning process. In order to make the cost-sensitive learning more effective, this paper proposed a sample-adaptive and cost-sensitive strategy for the classification of imbalanced data, which assigns a different weight to every single instance if misclassification occurs. Firstly, the strategy determines the distances between the boundary and instances according to the local distribution of the instances. Then, it assigns higher weights to the instances nearer to the boundary on the top of the margin theory. In this paper, the proposed strategy was applied to the classical LDM method. And a series of experiments on the UCI datasets prove that the sample-adaptive and cost-sensitive strategy can effectively improve the classifier's performance on imbalanced data classification.

Keywords Classification, Cost-sensitive learning, Boundary sample

1 引言

不平衡数据分类在实际应用场景中频繁出现,比如生物信息学[1-2]、电信或金融风险评估[3-4]、文本分类[5]和医学影像疾病诊断[6-7]等。以二分类为例,通常将样本数量较少的一类作为正类,样本较多的一类作为负类,负类与正类样本数量的比值即为不平衡度(ratio)。由于正类样本数量相对较少,传统分类器为保证训练模型在整体数据集上的预测精度最大化,将分类器决策面 H 向正类倾斜[8]。如图1所示,星形表示正类样本,圆圈表示负类样本。当数据集的不平衡度增加时,决策面的偏移程度会随之加剧,导致大多数甚至所有的正类样本都被误分为负类。在实际应用场景下,提升正类样本的预测精度往往更重要。比如,在对疑似癌症患者的诊断中,将癌症患者误诊为健康人的代价比将健康人误诊为癌症患者的代价高很多[9]。代价敏感是一种常用的解决该类问题的不平衡分类学习方法。其主要思想为:为不同的误分结果赋予不同的误分代价[10],以区别对待重要性不同的样本。因此,现有的针对不平衡数据的代价敏感学习方法通常为正类样本

到稿日期:2018-04-23 返修日期:2018-07-09 本文受国家自然科学基金(61370129,61375062,61632004,61773050)资助。
才子昕(1995-),女,硕士生,主要研究方向为机器学习和不平衡数据分类;**王馨月**(1994-),女,博士生,主要研究方向为不平衡数据分析;**徐 剑**(1994-),男,硕士生,主要研究方向为机器学习和不平衡数据分类;**景丽萍**(1978-),女,博士,教授,CCF会员,主要研究方向为高维数据子空间研究与机器学习,E-mail:lpjing@bjtu.edu.cn(通信作者)。

赋予比负类样本更高的误分代价,通过增加正类样本的误分代价使得分类器的决策面向负类样本的方向偏斜,以此提高正类的预测精度。此类代价敏感机制已经和多种传统方法相结合以用于不平衡数据分类,如代价敏感的 SVM[11] 和代价敏感的 LDM[12] 等。如图 1 所示,传统的分类器在引入基于类别的代价敏感机制后,决策面由 H 调整为 H_1。由于上述代价敏感机制只区别对待不同类样本集之间的误分代价,导致 H_1 虽将大部分样本划分正确,却很难将处于两类边界的样本正确分类。

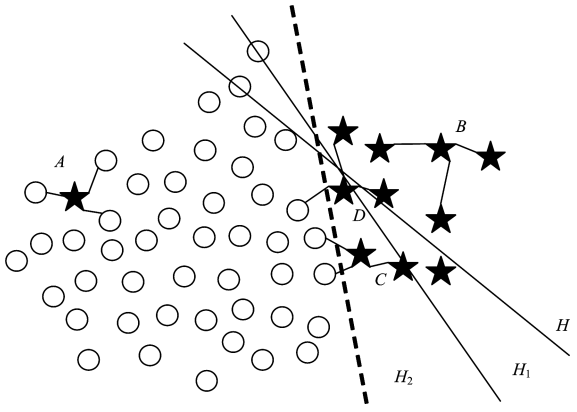

图 1　基于不同算法计算的分类器决策面展示图

Fig. 1　Illustration of separators with different algorithms

为了提升代价敏感机制的有效性,本文提出样本自适应的不平衡代价敏感分类器,在区分不同类别样本重要性的基础上,进一步区分正类样本间的重要程度。根据边界分布理论(margin theory)[13],距离两类样本边界越近的样本,对决策面位置的影响越大。这种处于边界的点相对于类内部点也更容易被预测错误,应该被赋予更高的权重。如图 1 所示,通过强调各类内不同样本的误分代价,基于样本自适应的代价敏感分类器得到了更为准确的决策面 H_2。在加强正类样本的重要性后,分类器可以正确划分所有的样本,不仅提高了正类的预测精度,同时也提高了负类的预测精度,进而更为有效地进行不平衡数据集的分类。

2　相关工作

现有的针对不平衡数据分类的方法大致可划分为数据和算法两个层面[14]。数据层面的方法旨在通过采样的方式调整两类数据的数量,将不平衡数据调整为平衡数据;算法层面的方法通过改进传统算法或提出新的方法来适应不平衡数据分类的学习。

2.1　数据层面的相关工作

数据层面的方法简单且容易实现,主要通过对正类样本过采样或对负类样本降采样来实现数据集的平衡化,随后使用传统分类器在相对平衡的数据集上对模型进行训练。

过采样通过随机复制或生成新样本的方式增加正类样本的数量。在随机复制原始正类样本的方式中,新增加的样本与原始样本的相似性很高,容易引发过拟合的问题。为了解决这一问题,生成新正类样本的过采样方法 SMOTE[15] 被提出,其通过线性插值的方式在任意两个正类样本之间插入新正类样本点。随后,如何选择过采样的参考点以及如何设计新样本生成器成为了过采样研究的核心问题。Borderline-SMOTE[16],ADASYN[17],Kernel ADASYN[18] 等均通过正类样本周边的局部信息找出正类中相对重要的样本作为过采样的参考点,利用这些参考点,使用线性插值法或拟合分布法生成新样本,并通过生成边界正类样本来明确两类样本的边界,从而提高分类器对不平衡数据的分类性能。但是,当不平衡度较高时,正类新样本的方式低效且容易引起样本重叠的问题。

为解决上述问题,许多方法通过对负类数据集进行降采样来平衡两类样本。随机移除样本的降采样策略容易导致重要性信息的丢失,为了更合理地对负类样本进行降采样,许多方法引入了数据清除策略以减少负类样本中的噪音或冗余数据,如 Tomek links,CNN,OSS 等[19]。然而,上述降采样方法都会一定程度地改变负类样本的结构信息。Cieslak 等[20] 提出先对负类样本聚类,然后从聚类后的每个簇中移除一定数量的样本以保证降采样后的负类结构信息与原始数据一致。然而,这种非监督的聚类方法也是机器学习中的一个难点问题。

为了更有效地进行数据的重采样,许多方法将过采样与降采样的方法相结合,例如 SMOTE+Tomek links,SMOTE+ENN 等[19]。此外,Batuwita 等[21] 通过分别衡量两类样本的重要性,对两类数据分别进行重采样,首先利用随机降采样的方法保留距离两类边界较近的正类样本,然后通过随机过采样的方法生成更多的边界负类样本。基于采样策略的处理方法直观且容易实现,但它们在一定程度上改变了原始数据的分布[22]。

2.2　算法层面的相关工作

针对不平衡数据分类,算法层面的方法大致分为集成学习、主动学习和代价敏感学习 3 类。

集成学习方法主要通过训练多个弱分类器来提高整体强分类器的性能。最常用的两种集成学习方法为 Bagging 和 Boosting。通过 Bagging 构造的强分类器由多个权重相等的弱分类器构成,每个弱分类器的数据成员为原始数据的子集,这些子集可通过有(或无)放回地采样得到。弱分类器的融合方式通常有 max,min,product,vote 和 sum。Sun 等对以上融合方式进行了改进,使其更好地用于不平衡数据的集成学习[23]。Chen 等[24] 和 Chan 等[25] 通过对负类数据集进行重新划分获取多个子集,并分别与正类数据集结合以构造相对平衡的数据集,再利用这些平衡的数据集训练多个弱分类器。除此之外,通过 Boosting 构造的强分类器由多个权重不相等的弱分类器构成。在模型迭代的过程中,通过增加上一次分类器误分样本的权重值,使之在下一次迭代中更容易被正确分类。一种最常见的基于 Boosting 思想的算法为 AdaBoost,由 Yoav 等[26] 提出。Wang 等[27] 和 Seiffert 等[28-30] 将 AdaBoost 分别引入传统的不平衡分类学习方法中,前者结合代价敏感 SVM,后者结合数据采样。集成学习方法可以充分利用原始数据,但保证弱分类器的差异性和精确性仍然是集成学习的难点。

主动学习方法通过设计一定的规则,选择既有价值信息又有代表性的样本用于模型的训练。其中,规则的设计既可

以依赖于分类器，也可以独立于分类器。例如，基于贝叶斯分类器的主动选择规则为：选择能使模型更新前后的后验概率提升较大的样本[31]。由于此类规则过于局限，许多专家学者通过设计通用的主动选择规则，使之更有效地处理不平衡数据的分类问题。这类规则在设计时通常会增加对正类选择的倾向性，以保证模型在相对平衡的数据集上进行训练[32]。尽管主动学习模型能很好地处理不平衡数据分类的问题，但学习过程停止标准的设定仍然是一个难点。

代价敏感学习的主要思想为：对不同的分类错误赋予不同的误分代价。由于传统分类器不能直接用于不平衡数据的分类，先后有学者将代价敏感的思想引入 SVM[11] 和 LDM[12] 等。他们均通过增大正类样本在目标函数中惩罚项的系数来增大正类样本的误分代价；同理，通过减小负类样本的相应惩罚系数来降低负类样本的误分代价。实验证明，当正类的惩罚系数为负类惩罚系数的 $ratio$ 倍时分类效果最佳[33]。

基于类别的代价敏感学习方法没有考虑到相同类别样本集中的样本之间的差异性。因此，本文提出样本自适应的代价敏感策略，并将上述策略应用到传统的 LDM 模型中，通过将其与传统的 LDM 以及基于类别代价敏感的 LDM 模型进行对比，验证了样本自适应代价敏感策略的有效性。

3 样本自适应的代价敏感策略

给定二分类数据集 $D=\{(x_i,y_i)\mid x_i\in R^d,y_i\in\{+1,-1\},i=1,\cdots,n\}$，其中 x_i 表示第 i 个样本的特征向量，y_i 表示相应的标签，d 为样本特征向量的维度。记正类和负类的样本数量分别为 m_+ 和 m_-，$ratio$ 表示数据集 D 的不平衡度，即 $ratio=m_-/m_+$。

为了处理不平衡数据分类问题，基于类别的代价敏感策略为正类样本赋予较高的权重 C^+，为负类样本赋予较低的权重 C^-，通常令 $C^+/C^-=ratio$[33]，以使得在提升正类误分代价的同时，保证正类样本的权重总和与负类相等。然而，这种权重的设置只考虑到类间样本重要性的差异，没有考虑到类内样本之间重要性的差异。考虑到不同样本对决策面位置不同程度的影响，本文提出样本自适应的代价敏感策略，目标函数如下：

$$\min_{w,\xi} f(x)+r(w)+\sum_i^n \theta_i L(f(x_i),y_i)$$

其中，w 为 $f(x)$ 中的参数，$r(w)$ 表示参数 w 的正则项，L 为损失函数，θ_i 表示样本 (x_i,y_i) 的权重。

自适应的代价敏感策略根据样本间重要性的差异，自适应地为样本赋予不同的权重，以提升分类器的分类性能。根据边界分布理论，距离决策面越近的样本，对决策面位置的影响越大，信息价值也越高。因此，本文根据样本的 k 近邻（以 $k=3$ 为例）的类分布情况，考察其距离两类边界的远近，以此对其进行重要性估计。具体地，若正类样本的 k 近邻中存在负类样本，则说明其可能处于两类边界，应该被赋予更高的权重。如图 1 所示，点 A,B,C,D 均为正类样本，点 C 的 k 近邻中有 2 个负类样本，点 D 的 k 近邻中有 1 个负类样本，两点都属于正类边界样本，应被赋予较高的权重，并且点 C 的权重应比点 D 高；点 A 的 k 近邻中全部都是负类样本，说明它可能是侵入负类的异常点，应被赋予较低的权重；点 B 的 k 近邻中没有负类样本，说明其是正类内部样本，应被赋予较低的权重。同理，根据负类样本 k 近邻中的正类样本个数，可以对负类样本进行重要性估计。基于这种权重的设置，不仅可以保证正类样本的权重高于负类，同时也为相同类别中更重要的样本赋予了更高的权重，以提升模型对不平衡数据的分类性能。

4 样本自适应的平衡分类器

本节将样本自适应代价敏感策略引入传统的 LDM 分类器，以提升 LDM 处理不平衡数据的性能。

LDM 通过最大化间隔（样本点决策面的距离）分布平均值和最小化间隔分布方差改进 SVM[34]，有效地提升了分类器的性能。其目标函数如下：

$$\begin{aligned}&\min_{w,\xi}\frac{1}{2}w^{\mathrm{T}}w+\lambda_1\overset{\wedge}{\gamma}-\lambda_2\bar{\gamma}+C\sum_{i=1}^n\xi_i\\&\text{s.t. } y_iw^{\mathrm{T}}\Phi(x_i)\geqslant 1-\xi_i\\&\quad\xi_i\geqslant 0,i=1,\cdots,n\end{aligned}$$

其中，λ_1,λ_2 和 C 均为大于 0 的常数，$\bar{\gamma}$ 和 $\overset{\wedge}{\gamma}$ 分别表示样本间隔分布的均值和方差，定义如下：

$$\bar{\gamma}=\sum_{i=1}^n y_iw^{\mathrm{T}}\Phi(x_i)$$

$$\overset{\wedge}{\gamma}=\sum_{i=1}^n\sum_{j=1}^n(y_iw^{\mathrm{T}}\Phi(x_i)-y_jw^{\mathrm{T}}\Phi(x_j))^2$$

为了使 LDM 更有效地应用于不平衡数据的分类，自适应的代价敏感方法通过为每一个数据点引入 θ_i，得到样本自适应的不平衡分类器（SA-LDM）。目标函数如下：

$$\begin{aligned}&\min_{w,\xi}\frac{1}{2}w^{\mathrm{T}}w+\lambda_1\overset{\wedge}{\gamma}-\lambda_2\bar{\bar{\gamma}}+C\sum_i^n\theta_i\xi_i\\&\text{s.t. } y_i\,w^{\mathrm{T}}\Phi(x_i)\geqslant 1-\xi_i\\&\quad\xi_i\geqslant 0,i=1,\cdots,n\end{aligned}$$

其中，$\bar{\bar{\gamma}}$ 是加权后间隔分布的均值：

$$\bar{\bar{\gamma}}=\sum_{i=1}^n\theta_iy_iw^{\mathrm{T}}\Phi(x_i)$$

考虑到正类样本比负类的预测结果更重要，为了简化模型复杂度并降低计算复杂度，SA-LDM 只对正类样本的权重进行细分。θ_i 的定义如下：

$$\theta_i=\begin{cases}h(k_i^-), & y_i=+1\\1, & y_i=-1\end{cases}$$

其中，k_i^- 为权重 θ_i 计算过程中的协变量，表示距离正类样本 (x_i,y_i) 最近的 k 个样本中负类样本的个数，$h(k_i^-)$ 是基于 k_i^- 的函数，形式如下：

$$h(k_i^-)=\begin{cases}\dfrac{k_i^-/k}{Z}\times(m_--m_+^0), & 0<k_i^-<k\\1, & k_i^-=0\vee k_i^-=k\end{cases}$$

其中，m_+^0 表示正类非边界样本的数量，Z 为归一化参数，即 $Z=\sum_{i|0<k_i^-<k}k_i^-/k$。

在基于类别的代价敏感机制的分类方法中，当所有正类样本的惩罚系数总和与负类相等时，分类效果最佳[33]。然而，基于类别的代价敏感机制没有考虑到类内样本间重要性的差异，SA-LDM 在保持两类权重总和相等的前提下，通过考察正类样本 k 近邻中负类样本的个数，为距离边界越近的正类样本赋予越高的权重，为正类内部点、异常点和负类样本

赋予较低的权重，最终实现在较高负类预测精度的基础上，提升正类的预测精度。SA-LDM 区别对待重要性不同的样本，可以更有效地用于不平衡数据的分类。下面给出相关的具体证明。

证明 1　SA-LDM 中两类样本权重总和相等。

正类样本权重总和为所有正类边界样本权重与正类非边界样本权重之和，即：

$$\sum_{i|y_i=+1}^{n}\theta_i=\sum_{i|0<k_i^-<k}h(k_i^-)+m_+^0\times 1 \tag{1}$$

根据 $h(k_i^-)$ 和归一化参数 Z 的定义，正类边界样本权重总和为：

$$\sum_{i|0<k_i^-<k}h(k_i^-)=m_--m_+^0 \tag{2}$$

结合式(1)和式(2)，得出：

$$\sum_{i|y_i=+1}^{n}\theta_i=m_-=\sum_{i|y_i=-1}^{n}\theta_i$$

因此，SA-LDM 中两类样本权重总和相等。

在此基础上，SA-LDM 为拥有较大 k_i^- 值的正类边界样本赋予了较高的权重 θ_i，同时保证了正类边界样本权重高于其他样本(其他样本的权重为 1)，具体证明见证明 2。

证明 2　SA-LDM 处理不平衡数据时，保证正类边界样本权重高于其他样本权重。

正类边界样本 (x_i,y_i) 权重的最低值出现在如下情况中：

$k_i^-=1\wedge k_{j\neq i}^-=k-1$

s.t. $i,j=1,2,\cdots,n$

结合式(2)，若要保证正类边界样本 (x_i,y_i) 权重的最低值大于 1，则应满足以下条件：

$$\min_{i|0<k_i^-<k}h(k_i^-)=\frac{m_--m_+^0}{1+(m_+-m_+^0-1)\cdot(k-1)}>1$$

由上述条件可以推出：

$m_->(k-1)m_+-(k-2)\cdot(m_+^0+1)$

结合 *ratio* 的定义，有：

$$ratio>k-1-\frac{(k-2)\cdot(m_+^0+1)}{m_+} \tag{3}$$

由于 $\frac{m_+^0+1}{m_+}<1$，因此：

$$k-1-\frac{(k-2)\cdot(m_+^0+1)}{m_+}>1 \tag{4}$$

根据不等式的传递性，由式(3)和式(4)得到：

$ratio>1$

因此，SA-LDM 处理不平衡数据时，能保证正类边界样本权重高于其他样本权重。

5　实验与结果

为了验证自适应样本的代价敏感策略的有效性，我们将其应用于 LDM 方法中，并将基于样本权重的代价敏感的方法(SA-LDM)与现有的 LDM 以及基于类别权重的代价敏感方法(CS-LDM)进行实验对比。5.1 节介绍本文使用的针对不平衡数据分类的性能度量指标；5.2 节介绍本文实验中所使用的数据集；5.3 节展示实验结果并对其进行分析。

5.1　性能度量

在不平衡数据分类中，准确度不再适用于评估算法的优劣[14]。当数据集的不平衡度较高时，即使正类样本的预测准确度为 0，分类器仍能获得较高的整体预测准确度。因此，本文采用以下常用于不平衡数据分类的性能度量指标来评价不同分类器的分类性能：正类查全率(recall-P)、负类查准率(precision-N)、宏观 F1 度量(F-macro)、宏观 G 均值(G-macro)和 AUC。其中，宏观 F1 度量以及宏观 G 均值分别为相应指标的算术平均值。

5.2　数据集

实验中使用了 5 个 UCI 的不平衡数据集，不平衡度从 2.78 到 11.76，不平衡度越大的数据集，分类难度越大。数据集的具体信息如表 1 所列。其中，n 表示数据集的样本总数，d 表示样本特征向量的维度，*ratio* 表示数据集的不平衡度。

表 1　5 个 UCI 数据集的详细情况

Table 1　Description of five UCI datasets

Dataset	n	d	*ratio*
Haberman	306	3	2.78
Ecoli	336	7	3.36
CMC	1473	9	3.42
Glass6	214	9	6.38
Balance	625	4	11.76

5.3　实验结果与分析

在实验中，核函数使用 RBF，近邻数 $k\in\{2,3,\cdots,7\}$，SA-LDM 对于每一个数据集的最佳超参数 k 如表 2 所列，其余的参数设置均与文献[12]中相同。

表 2　SA-LDM 处理各个数据集的最优超参数 k

Table 2　Optimal hyperparameters in SA-LDM for each dataset

Dataset	k-optimal
Haberman	3
Ecoli	5
CMC	4
Glass6	6
Balance	3

实验结果取最佳参数下五折交叉验证的平均值，最终结果如表 3－表 7 所列，其中，最佳结果被加粗以突出显示。

表 3　基于不同算法得到的正类查全率

Table 3　Recall-P of different methods

Dataset	LDM	CS-LDM	SA-LDM
Haberman	0.493	0.481	**0.581**
Ecoli	0.829	0.800	**0.886**
CMC	0.167	0.105	**0.219**
Glass6	0.933	**0.967**	**0.967**
Balance	0.164	0.164	**0.207**

表 4　基于不同算法得到的负类查准率

Table 4　Precision-N of different methods

Dataset	LDM	CS-LDM	SA-LDM
Haberman	0.834	0.830	**0.856**
Ecoli	0.980	0.977	**0.987**
CMC	0.796	0.786	**0.798**
Glass6	0.989	**0.995**	**0.995**
Balance	0.931	0.929	**0.932**

表 3 和表 4 分别展示了正类查全率(recall-P)和负类查准率(precision-N)。从表中可以看出，SA-LDM 在这两个指标上均表现最好。此外，图 2 展示了 SA-LDM 相对于另外两种方法在正类查全率上的提升值。提升值(Relative Improve-

ment)的定义如下：

$$Relative\ Improvement=(Rec_1-Rec_2)/Rec_2$$

其中，Rec_1 代表 SA-LDM 方法处理数据集的正类查全率(recall-P)，Rec_2 代表其他方法处理数据集的正类查全率。从图 2 可以看出，SA-LDM 在很大程度上提升了正类的预测结果。

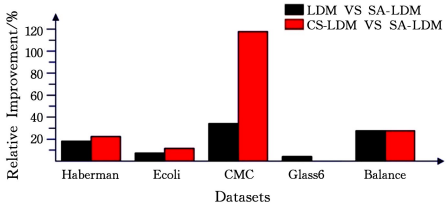

图 2 SA-LDM 方法相对于其他方法基于 5 个 UCI 数据集在正类查全率的相对提升值示意图

Fig. 2 Relative improvement to recall-P made by SA-LDM related to two baselines on five UCI datasets

相对于其他两种方法，SA-LDM 大大提升了正类样本的预测精度；并且 SA-LDM 方法在提升正类样本预测精度的同时，保持了较高的负类样本的预测精度。如表 4 所列，SA-LDM 基于所有数据集的正类查全率(precision-N)的结果都是最好的。以数据集 Ecoli 为例，该数据集共有 336 个样本，其中包括分布较为复杂的正类样本 35 个，负类样本 301 个。SA-LDM 方法对于数据集 Ecoli 的处理结果明显优于其他方法。具体来讲，SA-LDM 方法相对于 LDM 方法在正类查全率上提升了 6.8%，在负类查准率上提升了 0.7%。

宏观 F1 度量(F-macro)、宏观 G 均值(G-macro)和 AUC 是衡量不平衡数据分类的综合性指标，从表 5－表 7 中均可看出，SA-LDM 在这 3 个指标下都表现最佳。

表 5 基于不同算法得到的宏观 F1 度量

Table 5 F-macro of different methods

Dataset	LDM	CS-LDM	SA-LDM
Haberman	0.722	0.713	**0.734**
Ecoli	0.895	0.879	**0.897**
CMC	0.513	0.493	**0.561**
Glass6	0.962	**0.973**	**0.973**
Balance	0.567	0.546	**0.579**

表 6 基于不同算法得到的宏观 G 均值

Table 6 G-macro of different methods

Dataset	LDM	CS-LDM	SA-LDM
Haberman	0.728	0.720	**0.738**
Ecoli	0.896	0.881	**0.899**
CMC	0.543	0.522	**0.569**
Glass6	0.963	**0.974**	**0.974**
Balance	0.585	0.553	**0.595**

表 7 基于不同算法得到的 AUC

Table 7 AUC of different methods

Dataset	LDM	CS-LDM	SA-LDM
Haberman	**0.717**	0.705	0.716
Ecoli	**0.944**	0.943	**0.944**
CMC	0.673	0.593	**0.689**
Glass6	0.948	0.977	**0.978**
Balance	0.532	0.518	**0.551**

以上结果充分证明了样本自适应的不平衡分类器在处理不平衡数据集时的优越性。主要原因为：SA-LDM 方法依据正类样本附近的分布情况，判断其距离两类边界的远近，并根据正类样本距离两类边界的远近为其赋予不同的权重，同时保证两类样本的权重总和相等。LDM 将为所有样本赋予相同的权重，由于正类样本数量远少于负类，因此在整体上正类的权重低于负类，使得决策面向正类偏斜，正类样本的预测精度很低；CS-LDM 将正类赋予较高的权重，同时为负类赋予较低的权重，而没有考虑到同一类内不同样本间的差异。综合看来，SA-LDM 更适用于不平衡数据的分类。

结束语 随着大数据时代的发展，不平衡数据分类问题的研究受到了广泛关注，然而由于数据集在数量上严重失衡，传统的学习方法不能有效地对其进行学习。本文从代价敏感学习入手，提出了样本自适应的代价敏感策略。通过考察正类样本近邻中类别的分布情况，判断其距离两类样本边界的远近，并为距离两类样本边界越近的样本赋予越高的权重，同时保证正类样本的权重总和与负类相同，以此提升正类样本的预测精度。本文将提出的样本自适应的代价敏感策略应用于 LDM，并通过一系列实验验证了样本自适应的不平衡分类器可以有效处理不平衡数据的分类问题。后续的工作拟将样本自适应策略应用于多分类问题。

参 考 文 献

[1] RADIVOJAC P, CHAWLA N V, DUNKER A K, et al. Classification and knowledge discovery in protein databases[J]. Journal of Biomedical Informatics, 2004, 37(4): 224-239.

[2] ZOU Q, GUO M Z, LIU Y, et al. A classification method for class imbalanced data and its application on bioinformatics[J]. Journal of Computer Research and Development, 2010, 47(8): 1407-1414. (in Chinese)
邹权，郭茂祖，刘扬，等. 类别不平衡的分类方法及在生物信息学中的应用[J]. 计算机研究与发展，2010，47(8)：1407-1414.

[3] EZAWA K J, SINGH M, NORTON S W. Learning goal oriented Bayesian networks for telecommunications risk management[C]// Proceedings of the International Conference on Machine Learning. Bari, Italy: Morgan Kauffman, 1996: 139-147.

[4] SANZ JA, BERNARDO D, HERRERA F, et al. A compact evolutionary interval valued fuzzy rule-based classification system for the modeling and prediction of real-world financial applications with imbalanced data[C]// Proceedings of IEEE Trans on Fuzzy Systems, 2015, 23(4): 973-990.

[5] SU J S, ZHANG B F, XU X. Advances in machine learning based text categorization[J]. Journal of Software, 2006, 17(9): 1848-1859. (in Chinese)
苏金树，张博锋，徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报，2006，17(9)：1848-1859.

[6] DEEBA F, MOHAMMED S K, BUI F M, et al. Learning from imbalanced data: a comprehensive comparison of classifier performance for bleeding detection in endoscopic video[C]// Proceedings of International Conference on Informatics, Electronics and Vision. IEEE, 2016: 1006-1009.

[7] RANI K U, RAMADEVI G N, LAVANYA D. Performance of synthetic minority oversampling technique on imbalanced breast cancer data[C]// Proceedings of International Conference on

Computing for Sustainable Global Development. IEEE, 2016: 1623-1627.

[8] PROVOST F. Machine learning from imbalanced data sets 101 [C]// Proceedings of the AAAI'2000 Workshop on Imbalanced Data. IEEE, 2000.

[9] RAO R B. Data mining for improved cardiac care [J]. ACM SIGKDD Explorations Newsletter, 2006, 8(1): 3-10.

[10] DOMINGOS P. MetaCost: A general method for making classifiers cost-sensitive[C]// Proceedings of Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego: CA, ACM, 1999: 155-164.

[11] VEROPOULOS K, CAMPBELL C, CRISTIANINI N. Controlling the sensitivity of support vector machines[C]// Proceedings of the International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 1999: 55-60.

[12] CHENG F Y, ZHANG J, WEN C H. Cost-sensitive large margin distribution machine for classification of imbalanced data[J]. Pattern Recognition Letters, 2016, 80(C): 107-112.

[13] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(5): 273-297.

[14] STEFANOWSKI J. Dealing with data difficulty factors while learning from imbalanced data [OL]. http://www.cs.put.poznan.pl/jstefanowski/pub/jkbook7wersjaWWW.pdf.

[15] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: Synthetic minority oversampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.

[16] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning[C]// Proceedings of International Conference on Intelligent Computing. Springer-Verlag, 2005: 878-887.

[17] HE H, BAI Y, GARCIA E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]// Proceedings of IEEE International Joint Conference on Neural Networks. IEEE, 2008: 1322-1328.

[18] TANG B, HE H. KernelADASYN: Kernel based adaptive synthetic data generation for imbalanced learning[C]// Proceedings of Evolutionary Computation. IEEE, 2015: 664-671.

[19] BATISTA G, PRATI R C, MONARD M C. A study of the behavior of several methods for balancing machine learning training data[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 20-29.

[20] CIESLAK D A, CHAWLA N V, STRIEGEL A. Combating imbalance in network intrusion datasets[C]// Proceedings of IEEE International Conference on Granular Computing. IEEE, 2006: 732-737.

[21] BATUWITA R, PALADE V. Efficient resampling methods for training support vector machines with imbalanced datasets[C]// Proceedings of International Joint Conference on Neural Networks. IEEE, 2010: 1-8.

[22] ZHOU Z H, LIU X Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem[J]. IEEE Trans on Knowledge and Data Engineering, 2006, 18(1): 63-77.

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

[24] CHEN C, BREIMAN L. Using random forest to learn imbalanced data: Technical Report 666 [R]. Berkeley: Department of Statistics, UC Berkeley, 2004.

[25] 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]// International Conference on Knowledge Discovery and Data Mining. AAAI, 1998: 164-168.

[26] YOAV F, SCHAPIRE R E. A desicion-theoretic generalization of online learning and an application to boosting[C]// Proceedings of European Conference on Computational Learning Theory. Heidelberg, Berlin: Springer, 1995: 23-37.

[27] WANG B X, JAPKOWICZ N. Boosting support vector machines for imbalanced data sets[J]. Knowledge and Information Systems, 2010, 25(1): 1-20.

[28] SEIFFERT C, KHOSHGOFTAAR T M, HULSE J V, et al. RUSBoost: A hybrid approach to alleviating class imbalance[J]. IEEE Trans on Systems Man and Cybernetics Part A Systems and Humans, 2010, 40(1): 185-197.

[29] GALAR M, BARRENECHEA E, HERRERA F. EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary under-sampling[J]. Pattern Recognition, 2013, 46(12): 3460-3471.

[30] LIU X Y, WU J X, ZHOU Z H. Exploratory under-sampling for class-imbalance learning[J]. IEEE Trans on System, Man and Cybernetics B, 2009, 39(2): 539-550.

[31] OH S, MIN S L, ZHANG B T. Ensemble learning with active example selection for imbalanced biomedical data classification [J]. IEEE/ACM Trans on Computational Biology and Bioinformatics, 2011, 8(2): 316-325.

[32] ZHANG X X, YANG T B, SRINIVASAN P. Online asymmetric active learning with imbalanced data[C]// Proceedings of ACM SIGKDD International Conferenceon Knowledge Discovery and Data Mining. ACM, 2016: 2055-2064.

[33] AKBANI R, KWEK S, JAPKOWICZ N. Applying support vector machines to imbalanced datasets[C]// Proceedings of the 15th European Conference on Machine Learning. Springer Berlin Heidelberg, 2004: 39-50.

[34] GAO W, ZHOU Z H. On the doubt about margin explanation of boosting[J]. Artificial Intelligence, 2013, 203: 1-18.