DOI: 10.11992/tis.201909062

面向不均衡数据的融合谱聚类的自适应过采样法

刘金平[1], 周嘉铭[1], 贺俊宾[1,2], 唐朝晖[3], 徐鹏飞[1], 张国勇[3]

(1. 湖南师范大学 智能计算与语言信息处理湖南省重点实验室, 湖南 长沙 410081; 2. 湖南省计量检测研究院, 湖南 长沙 410014; 3. 中南大学 自动化学院, 湖南 长沙 410082)

摘 要: 分类是模式识别领域中的研究热点, 大多数经典的分类器往往默认数据集是分布均衡的, 而现实中的数据集往往存在类别不均衡问题, 即属于正常/多数类别的数据的数量与属于异常/少数类数据的数量之间的差异很大。若不对数据进行处理往往会导致分类器忽略少数类、偏向多数类, 使得分类结果恶化。针对数据的不均衡分布问题, 本文提出一种融合谱聚类的综合采样算法。首先采用谱聚类方法对不均衡数据集的少数类样本的分布信息进行分析, 再基于分布信息对少数类样本进行过采样, 获得相对均衡的样本, 用于分类模型训练。在多个不均衡数据集上进行了大量实验, 结果表明, 所提方法能有效解决数据的不均衡问题, 使得分类器对于少数类样本的分类精度得到提升。

关键词: 不自适应综合采样法; 不均衡数据集; 谱聚类; 过采样; 模式分类; 数据分布; 有偏分类器; 数据预处理

中图分类号: TP391 **文献标志码:** A **文章编号:** 1673−4785(2020)04−0732−08

中文引用格式: 刘金平, 周嘉铭, 贺俊宾, 等. 面向不均衡数据的融合谱聚类的自适应过采样法 [J]. 智能系统学报, 2020, 15(4): 732–739.

英文引用格式: LIU Jinping, ZHOU Jiaming, HE Junbin, et al. Spectral clustering-fused adaptive synthetic oversampling approach for imbalanced data processing[J]. CAAI transactions on intelligent systems, 2020, 15(4): 732–739.

Spectral clustering-fused adaptive synthetic oversampling approach for imbalanced data processing

LIU Jinping[1], ZHOU Jiaming[1], HE Junbin[1,2], TANG Zhaohui[3], XU Pengfei[1], ZHANG Guoyong[3]

(1. Hu'nan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hu'nan Normal University, Changsha 410081, China; 2. Hu'nan Institute of Metrology and Test, Changsha 410014, China; 3. School of Automation, Central South University, Changsha 410082, China)

Abstract: Classification is a research hotspot in the field of machine learning. Most classic classifiers assume that the distribution of dataset is generally balanced, while the data set in reality often has a problem of class imbalance. Namely, the number of data belonging to the normal/majority category and the amount of anomaly/minority data vary greatly. If the data is not processed, the classifier will ignore the minority and be biased towards the majority, which deteriorates the classification results. Focusing on the problem of data imbalance, this paper proposes a spectral clustering-fused comprehensive sampling algorithm (SCF-ADASYN). First, the spectral clustering method is employed to analyze the distribution information of the minority-type samples in the imbalanced dataset, and the samples of minority class are oversampled to obtain a relatively balanced dataset, used for the classification model training. A large number of experiments have been carried out on multiple unbalanced datasets. The results show that the SCF-ADASYN can effectively improve the imbalance on the data set, and the classification accuracies of the testing classifiers on the unbalanced data set can be significantly improved.

Keywords: adaptive synthetic sampling approach (ADASYN); imbalanced data set; spectral clustering; oversampling; pattern classification; data distribution; biased classifier; data pre-processing

收稿日期: 2019−09−27.

基金项目: 国家自然科学基金项目 (61971188, 61771492); 国家自然科学基金−广东联合基金重点项目 (U1701261); 湖南省自然科学基金项目 (2018JJ3349); 湖南省研究生科研创新项目 (CX20190415).

通信作者: 刘金平. E-mail: ljp202518@163.com.

分类是机器学习的重要一环, 支持向量机、随机森林、K 近邻等[1-2] 分类方法广泛应用于人工智能领域。这些经典的数据分类模型往往假定待

处理数据具有较为均匀的分布特性, 然而, 在实际的工程应用中, 数据往往会出现一类比另一类多的情况, 即分类处理的对象是不均衡数据集[3], 若不对其进行均衡化处理, 那么分类器极有可能忽略少数类数据, 导致所获得的分类模型不精确或者分类性能下降[4-5]。

不均衡数据集在生活生产中十分常见,如何对不均衡数据集的少数类样本进行正确分类是多个领域的重要课题[6]。比如, 在工业过程故障检测与诊断领域[7], 其模式分类的目标是识别出有故障的少数类样本, 而有故障的样本数要远远少于正常 (无故障) 样本数。对这些极度不均衡的数据进行处理, 往往会导致分类器偏向多数类样本, 而难以得到较好的模式分类结果。类似的情况还有医疗诊断[8]、网络入侵监测[9-10] 等领域。并且在实际应用中, 少数类样本的误 (漏) 识别代价往往大于多数类样本的误 (漏) 识别代价。比如, 在癌症筛查和诊断[11] 中, 对少数类类别 (肿瘤) 漏报, 极有可能延误病人的最佳治疗时间, 为病人生命带来不可估量的危害; 在网络入侵监测中, 正常访问与入侵行为存在严重的类别不均衡, 如果不能有效区分入侵与访问, 将严重威胁网络安全。基于这些原因, 不均衡数据的处理方法在国内外受到广泛关注[12]。

现阶段, 从数据层面进行考虑和从算法层面进行考虑是不均衡数据集处理方法中的两大主要分支。其中, 数据层面的处理方法是基于某种规则, 通过删减多数类样本或者增加少数类样本来改善原始数据的不均衡比, 使样本尽可能地均衡化, 方便进行分类模型的训练; 算法层面的处理方法主要包括集成学习[13] 和代价敏感学习[14-15] 方法, 这些方法通过修改分类算法在数据集上的偏置, 使得分类决策向少数类偏移, 从而有效提升分类器在不均衡数据集上的分类精度。

自适应综合过采样算法 (adaptive synthetic sampling approach, ADASYN)[16] 是一种有代表性的数据层面处理方法。ADASYN 基于少数类样本的概率分布对少数类样本进行自适应插值 (过采样), 对少数类样本的扩充, 以实现数据集的均衡化处理。该方法通过设定插值公式进行人工生成样本, 避免了样本的简单随机复制, 有效减弱了模型中可能出现的过拟合现象, 同时顾及了样本的分布信息, 因而在不均衡数据集处理中获得较好的处理结果。然而, 虽然 ADASYN 在对少数类样本进行插值 (过采样) 处理时在一定程度上考虑了少数类样本周围多数类样本的分布情况, 却没有分析和考虑少数类样本间的关联性, 存在少数类样本特征信息利用不充分的问题, 导致所获得过采样样本并不一定满足少数类样本的本质分布特性, 严重时会降低后续分类模型的性能。

本文针对不均衡数据集中少数类样本难以有效分类, 现有过采样方法未能充分利用少数类样本间的特征信息的问题, 提出一种融合谱聚类的自适应综合采样方法 (spectral clustering-fused adaptive synthetic oversampling approach, SCF-ADASYN)。SCF-ADASYN 首先采用谱聚类方法对少数类样本进行分析和处理; 根据少数类的分布结构, 将其聚成若干个簇; 再以少数类样本的聚类簇为单位对少数类进行自适应过采样, 得到均衡数据集, 以用于后续分类器模型训练。最后, 在多个不均衡数据集上进行实验, 通过搭配多种经典模式分类方法进行模式分类实验, 以验证本文所提方法的有效性和性能优越性。

1　相关工作

本节对 ADASYN 和谱聚类方法进行简单介绍, 概述其算法核心思路及主要流程。

1.1　自适应综合过采样

采样是一种常见的数据集预处理方法, 它通过增加少数类样本或减少多数类样本改变其不均衡比, 从而构造出新的训练数据集, 最常见的采样方法包括过采样[17] 和欠采样[18] 方法。

欠采样是一类通过对部分多数类样本进行删减以达到均衡化处理目的的不均衡数据集处理方法, 例如: 压缩最近邻法、随机删除法。研究表明, 欠采样方法在删除样本时会不可避免地丢失信息, 因此并未被广泛采用。

与欠采样相比, 通过增加少数类样本达到均衡化目的的过采样应用更为广泛。综合少数类过采样技术 (synthetic minority oversampling technique, SMOTE)[19] 是一种应用较为广泛的过采样算法。该算法通过线性插值对少数类样本进行过采样, 插值空间位于原数据空间, 因其具有良好的分类效果和简单易于实施的优势而被广泛应用。然而, 研究表明, 该方法会导致类别重叠的问题 (在多数类样本之间线性插值出一个少数类样本而导致类别重叠)。因而, He 等[16] 提出了一种自适应综合过采样方法 (adaptive synthetic sampling approach, ADASYN) 通过预先判定少数类样本周围多数类的分布情况, 对于不同的少数类样本进行自适应插值。

ADASYN 算法流程如下:

不均衡度的计算: $d = m_s/m_l$, 式中 $d \in (0,1]$; 若

$d<d_{\text{th}}$(d_{th} 为设定的不均衡度最大阈值), 则:

1) 应合成样本数计算: $G=\beta(m_l-m_s)$; 其中 $\beta\in[0,1]$ 代表加入合成样本后样本的不均衡度。

2) 少数类样本的 K 近邻查找。找出每个 x_i (少数类样本) 在 n 维空间的 K 近邻, 同时计算其比率 $r_i=\Delta_i/K, i=1,2,\cdots,m$, x_i 的 K 近邻中多数类的数目记作 Δ_i;

3) 对 r 进行正则化处理: $\hat{r}_i=r_i\Big/\sum_{i=1}^{m_s}r_i$, $r_i\Big(\sum\hat{r}_i=1\Big)$ 为概率分布;

4) 计算每个少数类样本应合成的样本数目。每个少数样本 x_i 计算应合成的样本数目为 $g_i=G\times\hat{r}_i$;

5) 对于每个少数类样本 x_i 生成 g_i 个新样本, 步骤如下:

对 1~g_i 个新样本执行循环:

①在每个待合成的少数类样本 x_i 周围 k 个邻居中选择 1 个少数类样本 x_{zi};

②依据式 (1) 插值:

$$s_j=x_i+\lambda\times(x_{zi}-x_i) \tag{1}$$

其中 $(x_{zi}-x_i)$ 是 n 维空间的差异向量, $\lambda\in[0,1]$。

ADASYN 和传统过采样方法 (比如 SMOTE) 相比, 最大优势是能够自适应地决定待合成的少数类样本合成的样本数目, 避免了简单的随机复制带来的过拟合问题。

1.2　谱聚类

谱图划分问题衍生出了谱聚类[20], 它将聚类问题转化为无向图的多路划分问题[21]。样本点用无向图 $G(V,W)$ 中的顶点来指代, 用图中边的权重来指代数据点间的相似性度量, 图 G 中顶点的集合为 K, 图 G 中边权重的集合为 r。以最优化为准则, 在此无向图基础上, 相同类别的点相似性较高, 不同类别的点相似性较低, 流程如图 1 所示。

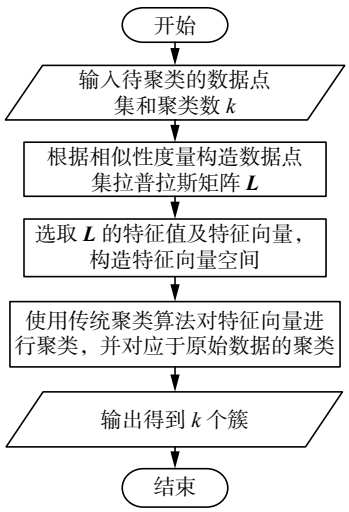

图 1　谱聚类算法的一般过程

Fig. 1　General process of spectral clustering

由 Ng、Jordan 和 Weiss 提出的 NJW 算法[21]是一种经典的谱聚类算法, 给定一批样本维数为 l、样本数量为 n 的样本集 $s=\{s_1,s_2,\cdots,s_n\}\in\Re^l$, NJW 算法流程如下:

1) 构造相似性矩阵 $\boldsymbol{A}\in\Re^{n\times n}$, 矩阵中元素 $\boldsymbol{A}_{ij}=\exp(-\|s_i-s_j\|)/2\sigma^2$, 且当 $i=j$ 时, $\boldsymbol{A}_{ii}=0$;

2) 构造度矩阵 $\boldsymbol{D}$, 相似性矩阵 $\boldsymbol{A}$ 的第 i 行元素值的和是矩阵对角线上的元素 $\boldsymbol{D}(i,i)$; 该矩阵主对角线外的其他值均为 0。由此构造 Laplacian 矩阵 $\boldsymbol{L}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$;

3) 构造矩阵 $\boldsymbol{X}=[x_1,x_2,\cdots,x_k]\in\Re^{n\times k}$: 特征值分解 $\boldsymbol{L}$, 找出 $\boldsymbol{L}$ 前 k 个最大特征值对应的特征向量 $x_1,x_2,\cdots,x_i$, 然后特征向量按列存储;

4) 归一化 X 的行向量得到矩阵 $\boldsymbol{Y}$, $\boldsymbol{Y}_{ij}=\boldsymbol{X}_{ij}\Big/\left(\sum_j\boldsymbol{X}_{ij}^2\right)^{1/2}$;

5) 空间 $\Re$ 中的样本为矩阵 $\boldsymbol{Y}$ 所有行向量 (样本数量为 n, 样本维数为 k), 采用 c-means 进行聚类;

6) 当且仅当矩阵 $\boldsymbol{Y}$ 的第 i 行被划分为第 j 聚类时, 把最初的样本点 s_i 划分为第 j 聚类。

2　融合谱聚类的自适应综合采样算法

2.1　算法的提出

ADASYN 算法[16]在计算插值数目时会计算少数类样本周围的多数类样本的分布情况, 从而在分类边界对少数类样本进行自适应采样。然而在少数类样本之间也存在特征信息的关联, 如果能充分利用少数类样本间的特征关联信息, 再以此决定插值的数目和范围, 将会进一步提高不均衡数据集的分类精度[22]。因此, 本文在 ADASYN 算法的研究现状基础上, 提出一种融合谱聚类的自适应综合采样算法 (SCF-ADASYN)。

2.2　算法设计

SCF-ADASYN 的思路为: 先依据公式 $d=m_s/m_l$(m_s 为多数类样本数, m_l 为少数类样本数) 求出样本的不均衡度 d, 以此计算所需的总插值数 G, 再用谱聚类对少数类数据进行分析, 得到 k 个少数类样本聚类簇, 根据每个簇的少数类样本数目分配其插值数, 最后以簇为单位根据公式 $s_j=x_i+(x_{zi}-x_i)\times\lambda$ 进行样本插值。SCF-ADASYN 算法描述如下。

算法　SCF-ADASYN

输入　含有 m 个样本点 $\{x_i,y_i\}$, $i=1,2,\cdots,m$ 的训练集 A, 最大的不均衡容忍度阈值 (d_{th}), 聚类簇数目 k。

输出　过采样后的数据集 D。

算法实现的主要步骤如下:

1) 不均衡度的计算: $d=m_s/m_l$, 式中 $d\in(0,1]$; 若 $d<d_{\text{th}}$(d_{th} 为一预设阈值), 则执行 2);

2) 求出应合成的少数样本数: $G=(m_l-m_s)\cdot\beta$, 其中 $\beta\in[0,1]$ 表示加入合成样本后的不均衡度;

3) 使用谱聚类方法对少数类样本进行聚类处理, 得到 k 个簇; 计算每个少数类样本簇 C_i(每个簇的少数类样本数记为 n_i, $i=1,2,\cdots,k$) 之间的样本数比值, 并由此比值计算出每个簇的插值数 $G_i=\dfrac{n_i}{m_l}$;

4) 找出每个少数类簇的样本 x_i 在 n 维空间的 K 近邻, 计算其比率 $r_i=\Delta_i/K, i=1,2,\cdots,m$, 其中 Δ_i 是 x_i 的 K 近邻中多数类的数目, $r_i\in(0,1]$;

5) 正则化 r: $\hat{r}_i=r_i\Big/\sum_{i=1}^{m_s}r_i$, r_i 实际上为概率分布;

6) 以少数类样本簇为单位插值:

对 C_1~C_k 执行循环:

①对于少数类簇的每个样本点计算需合成的样本数目, $g_i=\hat{r}_i\times G_i$;

②对于每个少数类样本 x_i 生成 g_i 步骤如下:

对 1~g_i 个新样本执行循环:

(a) 在每个待合成的少数类样本 x_i 周围 k 个邻居中选择 1 个少数类样本 x_{zi};

(b) 依据式 (2) 进行插值:

$$s_j=x_i+\lambda\times(x_{zi}-x_i)\tag{2}$$

式中 $(x_{zi}-x_i)$ 是 n 维空间的差异向量, $\lambda\in[0,1]$。

SCF-ADASYN 算法流程如图 2 所示。由于谱聚类使用数据的相似性矩阵的谱执行降维, 可以在小数据集上产生高质量的聚类, 适用于少数类样本聚类分析, 因此 SCF-ADASYN 在自适应样本插值前利用谱聚类分析少数类样本, SCF-ADASYN 在聚类阶段将少数类样本分为簇, 时间复杂度为 $O(n^3)$; 在插值阶段, 时间复杂度与聚类簇数正相关, 时间复杂度为 $O(Cn)$, 其中 C 是聚类的簇数。因此, 整个 SCF-ADASYN 的时间复杂度为 $O(n^3)$。

3　实验

本文实验包括 2 个部分: 1)验证性实验, 在选用的不均衡数据集上, 对本文提出的 SCF-ADASYN 进行有效性验证, 对比了其相对于未处理以及经 ADASYN 算法处理的评价结果, 分析、讨论本文算法的有效性; 2)在多个常见的不均衡数据集上, 将本文提出的 SCF-ADASYN 与 SMOTE、ADASYN 进行对比, 判断本文算法的优劣。

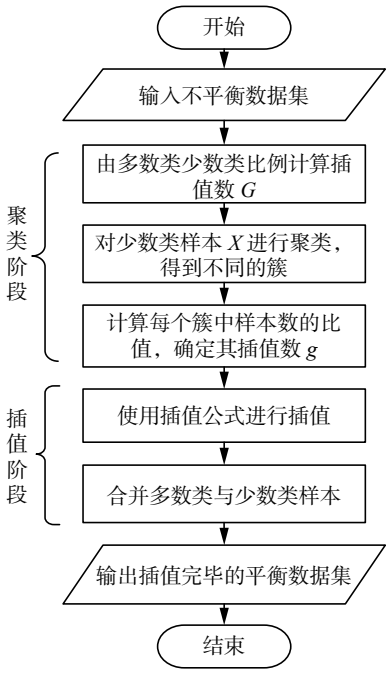

图 2　融合谱聚类的自适应综合采样算法流程图

Fig. 2　Flowchart of SCF-ADASYN

3.1　数据集

表 1 展示了实验中使用的数据集信息, 表中 IR 为不均衡率, 其公式为

$$\text{IR}=\frac{\text{少数类样本数}}{\text{多数类样本数}}$$

表 1　不均衡数据集信息

Table 1　List information of imbalanced datasets

数据集	样本数	属性	少数类	多数类	IR
Blood	671	4	219	452	0.484 5
Pima	768	8	268	500	0.535 0
Abalone	4 177	8	42	689	0.061 0
Haberman	306	3	81	225	0.395 6
Yeast	1 484	8	37	707	0.052 3

数据集 Blood 中的数据是 2007 年某地献血情况统计, 分为献血与没献血 2 类。

数据集 Pima 为凤凰城附近的糖尿病呈阳性的患者分类数据集。

数据集 Abalone 为鲍鱼数据集, 数据集中含有 4 177 个样本, 本文选择其中的“18”作为少数类, 选择其中的“9”作为多数类。

数据集 Haberman 包含病人手术时的多项指标, 以此判断病人的状况。

数据集 Yeast 为酵母数据集, 本文选择数据集标签中的 CYT 和 MIT 作为多数类, 样本数为 707, 选择 EXC 类别作为少数类, 样本数为 37。

Abalone 数据集的不均衡比 0.061 0, 而 Haberman 数据集的不均衡比是 0.395 6, 可以判断出

Abalone 要更加不均衡，选用不同不均衡度的数据集，能直观比较不同不均衡情况下本文方法效果。

本实验将在这 5 个数据集上对 SMOTE 算法、ADASYN 算法以及本文的 SCF-ADASYN 算法进行测试。将 3 种方法处理过的数据集通过支持向量机 (support vector machine, SVM)[23]、随机森林 (random forest, RF)[24]、K 最近邻算法 (k-nearest neighbor, KNN)[25] 等分类器进行分类，按 4∶1 的比率将数据集随机分为训练集和测试集，并运行 5 次取平均值作为结果，比较分析本文方法的优劣。

本文的实验环境为

1) 处理器型号：Inter(R)I5-8300H CPU@ 2.30 GHz;

2) 运行内存：8 GB;

3) 实现语言：Python 3.7;

4) 操作系统：Linux(Ubuntu18.04)。

3.2　评价指标

对于少数类数据的分类评价在不均衡数据分类评价中十分重要[26]。本文用 F-measure、G-mean 以及 AUC[27] 来衡量分类结果。

分类结束后，结果分为 4 种情况：预测正例、预测负例以及真实正例、真实负例，如表 2 所示。

表 2　混淆矩阵

Table 2　Confusion matrix

总样本数	预测正例	预测负例
真实正例	TP	FN
真实负例	FP	TN

将表 2 中的 TP、FP、TN、FN 按照模型的评价需求进行组合就构成了常用的评价标准。本文使用的评价标准包括查准率、召回率、G-mean、F 值以及 AUC。

查准率 (Precision) 为

$$\text{Precision} = \text{TP}/(\text{FP}+\text{TP})$$

该指标表示正确分类的多数类样本与分为多数类的所有样本比值。

召回率 (Recall) 表示被正确分类的少数类样本与实际少数类样本的比值，其计算公式为

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN})$$

G-mean 为查准率和召回率乘积的平方根，它反映出分类器对于多数类和少数类分类的整体能力。因此，采用 G-mean 准则来评价不均衡数据集总体分类性能十分合理。

总体性能指标 G-mean 的计算公式为

$$\text{G-mean} = \sqrt{\frac{\text{TP}\cdot\text{TN}}{(\text{TP}+\text{FN})\cdot(\text{TN}+\text{FP})}}$$

通过 G-mean 的数值来判断分类器的效果，G-mean 数值越大说明召回率和查准率越高，效果越好。

F 值计算公式为

$$\text{F-measure} = \frac{\text{Precision}\times\text{Recall}\times(1+\beta^2)}{\text{Precision}+\beta^2\times\text{Recall}}$$

F-measure 计算公式中包含了查准率和召回率，在实验室中多取 $\beta=1$。当查准率和召回率同时上升时，F-measure 才会提升。因此，本文用 F-measure 衡量对于不均衡数据的分类性能。

因为能独立于数据集的类分布，ROC 曲线对数据集的不均衡性有很好的鲁棒性，本文使用曲线下面积 AUC 来代替 ROC 曲线作为不均衡数据评价方法,值越大代表分类器的性能表现越优秀。

3.3　实验结果及分析

1) 验证性实验

本文验证性实验选用 Pima 数据集进行实验。Pima 数据集为印第安人糖尿病数据集，其中少数类样本 268 例，多数类样本 500 例。比较未进行均衡化处理、采用 ADASYN 处理、SMOTE 处理以及 SCF-ADASYN 算法处理后的 Pima 数据集，在多个分类器下的分类表现。

由表 3 实验结果可知，经 SCF-ADASYN 算法处理后，F-measure, G-mean 以及 AUC 相较于未经采样处理显著提高，说明经本文算法处理后分类器整体的分类性能以及对少数类样本的分类精度都显著提高。也就是说，本文提出的 SCF-ADASYN 算法能有效地处理数据类不均衡的问题，从而提高分类器的性能。

表 3　验证性实验结果

Table 3　Experimental results of validation experiments

分类器	数据层方法	AUC	F值	G-mean
RF	SCF-ADASYN	0.800 5	0.782 1	0.800 4
	ADASYN	0.781 7	0.761 3	0.780 8
	SMOTE	0.788 1	0.741 2	0.774 3
	未处理	0.706 0	0.725 4	0.726 5
KNN	SCF-ADASYN	0.744 6	0.776 2	0.739 3
	ADASYN	0.748 1	0.760 9	0.747 0
	SMOTE	0.723 1	0.712 3	0.776 3
	未处理	0.705 4	0.695 1	0.715 5
SVM	SCF-ADASYN	0.918 1	0.828 3	0.825 2
	ADASYN	0.779 3	0.786 0	0.778 8
	SMOTE	0.795 6	0.775 3	0.741 4
	未处理	0.705 2	0.668 6	0.656 5

将本文方法与 SMOTE 进行比较，3 项指标均有提升，说明本文提出的 SCF-ADASYN 能够有效提升预处理后分类器的分类精度，相较于 SMOTE 具有更好的性能表现。

而经过 SCF-ADASYN 处理后，模型的 AUC、G-mean 以及 F-measure 值相较于搭配 ADASYN 的模型分别提高了 7.19%、4.67% 以及 4.08%，说明 SCF-ADASYN 算法相比于 ADASYN 算法对分类的优化程度更高。

2) 对比性实验

表 4 为 SMOTE 算法，ADASYN 算法以及 SCF-ADASYN 算法在 5 个数据集上搭配不同分类器的 F-measure 值。

表 4　3 种采样方法在不同分类器中的 F-measure
Table 4　F-measure of three comparative sampling methods

数据集	方法	SVM	RF	KNN
Blood	SMOTE	0.732 2	0.698 6	0.735 4
	ADASYN	0.677 6	0.675 4	0.728 7
	SCF-ADASYN	0.700 8	0.739 1	0.734 5
Pima	SMOTE	0.849 3	0.754 0	0.807 5
	ADASYN	0.786 0	0.761 3	0.760 9
	SCF-ADASYN	0.818 1	0.782 1	0.776 2
Abalone	SMOTE	0.391 5	0.385 6	0.397 1
	ADASYN	0.402 5	0.394 6	0.371 5
	SCF-ADASYN	0.412 1	0.406 5	0.404 2
Haberman	SMOTE	0.761 9	0.733 3	0.691 3
	ADASYN	0.666 6	0.719 1	0.702 1
	SCF-ADASYN	0.696 6	0.733 3	0.690 4
Yeast	SMOTE	0.801 2	0.803 5	0.792 3
	ADASYN	0.796 3	0.801 1	0.735 2
	SCF-ADASYN	0.806 2	0.812 4	0.821 5

F-measure 作为召回率和查准率两者的组合，相同的多数类样本查准率下，其数值升高，说明在分类过程中对于少数类样本的分类能力得到提高。

从表 4 可以看出，对比 SVM、RF 以及 KNN 对 5 个数据集进行模式分类，SCF-ADASYN 的 F-measure 值基本高于 ADASYN 算法，其中在 blood 数据集上，使用 RF 作为分类器的分类性能指标 F-measure 提高了 9.43%，这说明经本文方法处理后，在少数类的分类精度上要优于 ADASYN 算法。在 Blood、Haberman 以及 Yeast 数据集上，本文方法的 F-measure 值要高于两个经典算法，这是由于谱聚类在样本数量较少、样本属性较大的数据集上聚类效果更好，能更好地细化出少数类样本之间的特征属性，在这种情况下插值得到的均衡数据集少数类样本的分类精度更高。

表 5 为 SMOTE 算法、ADASYN 算法以及 SCF-ADASYN 算法在 5 个数据集上搭配不同分类器的 G-mean 值。

表 5　3 种采样方法在不同分类器中的 G-mean
Table 5　G-mean of three comparative sampling methods

数据集	方法	SVM	RF	KNN
Blood	SMOTE	0.693 3	0.700 0	0.740 9
	ADASYN	0.660 8	0.677 6	0.703 2
	SCF-ADASYN	0.695 8	0.738 8	0.741 3
Pima	SMOTE	0.831 0	0.769 5	0.792 9
	ADASYN	0.778 8	0.780 8	0.747 0
	SCF-ADASYN	0.815 2	0.800 4	0.739 3
Abalone	SMOTE	0.621 2	0.605 4	0.623 2
	ADASYN	0.632 3	0.633 2	0.647 4
	SCF-ADASYN	0.643 2	0.626 5	0.651 3
Haberman	SMOTE	0.711 3	0.736 2	0.715 9
	ADASYN	0.676 1	0.727 7	0.701 3
	SCF-ADASYN	0.705 9	0.743 8	0.712 7
Yeast	SMOTE	0.702 1	0.712 3	0.695 4
	ADASYN	0.716 4	0.703 3	0.714 5
	SCF-ADASYN	0.723 2	0.726 3	0.723 2

G-mean 为多数类样本分类查准率和少数列样本召回率乘积的平方根，G-mean 提升意味着两者同时提升，因此本文整体的分类性能用它来衡量。

如表 5 所示，经 SCF-ADASYN 算法处理后，分类器的 G-mean 有不同程度的提高。其中，在 Abalone 数据集上，如果采用 RF 作为模型分类器，基于本文所提出的 SCF-ADASYN 进行不均衡数据集处理，其模型分类的 G-mean 要比采用 ADASYN 算法高 3.2%，比 SMOTE 算法高 1.9%。这些结果表明，本文算法与 SMOTE 以及 ADASYN 算法相比，分类效果有较大的提高，能够显著提高不同类别的分类精度，具有良好的适应性。

表 6 为 SMOTE 算法、ADASYN 算法以及 SCF-ADASYN 算法在 5 个数据集上搭配不同分类器的 AUC 值，AUC 值大意味着整体分类效果优秀。

表 6　3 种采样方法在不同分类器中的 AUC
Table 6　AUC of three sampling methods

数据集	方法	KNN	RF	SVM
Abalone	SCF-ADASYN	0.536 4	0.531 4	0.531 4
	ADASYN	0.503 6	0.514 6	0.521 3
	SMOTE	0.516 4	0.521 4	0.512 4
Yeast	SCF-ADASYN	0.801 4	0.812 3	0.801 4
	ADASYN	0.754 5	0.795 6	0.791 4
	SMOTE	0.788 2	0.792 3	0.785 2
Blood	SCF-ADASYN	0.734 0	0.739 4	0.699 8
	ADASYN	0.709 4	0.678 6	0.673 6
	SMOTE	0.741 4	0.703 1	0.697 6
Haberman	SCF-ADASYN	0.722 5	0.743 7	0.707 5
	ADASYN	0.704 8	0.727 8	0.680 9
	SMOTE	0.719 1	0.739 1	0.713 5
Pima	SCF-ADASYN	0.744 6	0.800 5	0.815 4
	ADASYN	0.748 1	0.781 7	0.779 3
	SMOTE	0.798 8	0.769 5	0.832 2

由表 6 可以看出，使用 SVM、RF 以及 KNN 对 5 个数据集进行分类，本文方法的 AUC 值基本高于 SMOTE 算法和 ADASYN 算法，说明经本文方法处理后，模型对于不均衡数据的分类能力有效提升。

上述实验表明，经本文提出的 SCF-ADASYN 方法处理后，各分类器在各不均衡数据集上的模式分类性能显著提升，表明 SCF-ADASYN 方法能够有效地处理数据不均衡的问题。

4　结束语

针对不均衡数据中少数类样本难以分类的问题，本文提出了一种融合谱聚类的自适应综合采样方法。该方法利用谱聚类将少数类样本按照特征信息分成若干个簇，有效获取少数类样本的空间结构，在获得少数类样本空间结构的基础上，再以所获得的聚类簇为单位，对少数类样本进行自适应插值，以此解决数据集的不均衡问题。验证性和对比性实验结果表明，不均衡数据集在经本文算法处理后，在传统分类器上均有更好的少数类分类精度。本文方法融合的谱聚类在处理样本数目较少的数据集时有较好的效果，而当样本数目较大时效果下降，怎样能在大样本数据集上取得更好的效果是进一步研究的方向。

参考文献：

[1] LESSMANN S, BAESENS B, SEOW H V, et al. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research[J]. European journal of operational research, 2015, 247(1): 124–36.

[2] LIU J, HE J, ZHANG W, et al. TCvBsISM: Texture classification via B-splines-based image statistical modeling[J]. IEEE access, 2018, 6(1): 76–93.

[3] 翟云, 杨炳儒, 曲武. 不平衡类数据挖掘研究综述 [J]. 计算机科学, 2010, 37(10): 27–32.
ZHAI Yun, YANG Bingru, QU Wu. Survey of mining imbalanced datasets[J]. Computer science, 2010, 37(10): 27–32.

[4] LIN W C, TSAI C F, HU Y H, et al. Clustering-based undersampling in class-imbalanced data[J]. Information sciences, 2017, 17(2): 409–410.

[5] HE H, GARCIA E A. Learning from imbalanced data[J]. IEEE transactions on knowledge & data engineering, 2009, 21(9): 1263–84.

[6] LIU J, TANG Z, ZHANG J, et al. Visual perception-based statistical modeling of complex grain image for product quality monitoring and supervision on assembly production line[J]. Plos one, 2016, 11(3): 1–25.

[7] 刘天羽, 李国正, 尤鸣宇. 不均衡故障诊断数据上的特征选择 [J]. 小型微型计算机系统, 2009, 30(5): 924–927.
LIU Tianyu, LI Guozheng, YOU Mingyu. Feature selection on unbalanced fault diagnosis data[J]. Journal of Chinese computer systems, 2009, 30(5): 924–927.

[8] YUAN X, XIE L, ABOUELENIEN M. A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data[J]. Pattern recognition, 2018, 77(1): 160–72.

[9] LIU J, HE J, ZHANG W, et al. ANID-SEoKELM: adaptive network intrusion detection based on selective ensemble of kernel ELMs with random features[J]. Knowledge-based systems, 2019, 177(1): 104–16.

[10] 刘金平, 张五霞, 唐朝晖, 等. 基于模糊粗糙集属性约简与 GMM-LDA 最优聚类簇特征学习的自适应网络入侵检测 [J]. 控制与决策, 2019, 34(2): 243–251.
LIU Jinping, ZHANG Wuxia, TANG Zhaohui, et al. Adaptive network intrusion detection based on fuzzy rough set-based attribute reduction and GMM-LDA-based optimal cluster feature learning[J]. Control and decision, 2019, 34(2): 243–251.

[11] FOTOUHI S, ASADI S, KATTAN M W. A comprehensive data level analysis for cancer diagnosis on imbalanced data[J]. Journal of biomedical informatics, 2018, 90(1): 1–29.

[12] ZHOU P, HU X, LI P, et al. Online feature selecton for high dimensional class-imbalanced data[J]. Knowledge-based systems, 2017, 136(15): 187–199.

[13] QIAN Y, LIANG Y, LI M, et al. A resampling ensemble algorithm for classification of imbalance problems[J]. Neurocomputing, 2014, 143(2): 57–67.

[14] LIU M, XU C, LUO Y, et al. Cost-sensitive feature selection by optimizing F-Measures[J]. IEEE transactions on image processing, 2018, 27(3): 1323–35.

[15] 吴雨茜, 王俊丽, 杨丽, 等. 代价敏感深度学习方法研究综述 [J]. 计算机科学, 2019, 46(5): 8–19.
WU Yuqian, WANG Junli, YANG Li, et al. Survey on cost-sensitive deep learning methods[J]. Computer science, 2019, 46(5): 8–19.

[16] HE H, BAI Y, GARCIA E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]// Neural Networks. Hong Kong, China, 2008, 3641−46

[17] AHMAD J, JAVED F, HAYAT M. Intelligent computational model for classification of sub-Golgi protein using oversampling and fisher feature selection methods[J]. Artificial intelligence in medicine, 2017, 78(1): 14–16.

[18] LIN W C, TSAI C F, HU Y H, et al. Clustering-based undersampling in class-imbalanced data[J]. Information sciences, 2017, 17(2): 409–410.

[19] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2011, 16(1): 321–357.

[20] 蔡晓妍, 戴冠中, 杨黎斌. 谱聚类算法综述 [J]. 计算机科学, 2008(7): 14–18.
CAI Xiaoyan, DAI Guanzhong, YANG Libin. Survey on spectral clustering algorithms[J]. Computer science, 2008(7): 14–18.

[21] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: Analysis and an algorithm[C]//Proceedings of the Advances in Neural Information Processing Systems. Berkeley, USA, 2002: 26−34.

[22] 刘金平, 周嘉铭, 刘先锋, 等. 基于聚类簇结构特性的自适应综合采样法在入侵检测中的应用 [J/OL]. 控制与决策: https://doi.org/10.1n3195/j.kzyjc.2019.1672.
LIU Jinping, ZHOU Jiaming, LIU Xianfeng, et al.Toward intrusion detection via cluster-structure characteristics-based adaptive synthetic sampling approach[J/OL]. Control and decision: https://doi.org/10.13195/j.kzyjc.2019.1672.

[23] CHAUHAN V K, DAHIYA K, SHARMA A. Problem-formulations and solvers in linear SVM: a review[J]. Artificial intelligence review, 2018, 6(1): 1–53.

[24] PAUL A, MUKHERJEE D P, DAS P, et al. Improved random forest for classification[J]. IEEE transactionson image processing, 2018, 27(8): 4012–24.

[25] ZHANG S, DENG Z, CHENG D, et al. Efficient KNN classification algorithm for big data[J]. Neurocomputing, 2016, 195(26): 143–8.

[26] 林智勇, 郝志峰, 杨晓伟. 若干评价准则对不平衡数据学习的影响 [J]. 华南理工大学学报(自然科学版), 2010, 38(4): 147–155.
LIN Zhiyong, HAO Zhifeng, YANG Xiaowei. The influence of several evaluation criteria on unbalanced data learning[J]. Journal of South China University of Technology (natural science edition), 2010, 38(4): 147–155.

[27] THARWAT A. Classification assessment methods[J]. Applied computing and informatics, 2018, 12(1): 1–13.

作者简介:

刘金平, 副教授, 博士, 主要研究方向为智能信息处理。

周嘉铭, 硕士研究生, 主要研究方向为数据挖掘、模式识别。

贺俊宾, 硕士研究生, 主要研究方向为模式识别、计算机视觉。