Advances in Applied Mathematics 应用数学进展, 2024, 13(9), 4108-4118
Published Online September 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.139392

Hans汉斯

基于KPCA的不平衡数据欠抽样算法

王晓玲[1]，金永超[1,2]，刘威伟[1]，王希胤[1,2*]

[1]华北理工大学理学院，河北 唐山
[2]河北省数据科学与应用重点实验室，河北 唐山

收稿日期：2024年8月10日；录用日期：2024年9月2日；发布日期：2024年9月12日

摘　要

在现实世界的分类任务中，不平衡数据通常呈现非线性分布的特点，而传统的抽样方法难以有效处理这些非线性，导致分类效果不佳。为了解决这个问题，本文提出了一种基于核主成分分析(KPCA)的欠抽样方法。该方法通过使用非线性核函数将原始数据映射到适当的高维空间使其线性化，然后根据每个样本在核主成分上的得分来选择性地删除多数类样本，从而实现欠抽样。在9组具有不同平衡率的数据集上，采用本文提出的方法进行了欠抽样预处理，并使用逻辑回归(Logistic Regression)分类器进行分类。实验结果表明，在Accuracy、F1-measure和AUC值三个指标中，本文方法分别在7组、8组和9组数据集上取得了最高评分。这表明该方法在不平衡数据集上具有良好的分类性能。

关键词

不平衡数据，欠抽样，核主成分分析，分类

KPCA-Based Under-Sampling Algorithm for Unbalanced Data

Xiaoling Wang[1], Yongchao Jin[1,2], Weiwei Liu[1], Xiyin Wang[1,2*]

[1]School of Science, North China University of Science and Technology, Tangshan Hebei
[2]Key Laboratory of Data Science and Application of Hebei Province, Tangshan Hebei

Received: Aug. 10th, 2024; accepted: Sep. 2nd, 2024; published: Sep. 12th, 2024

Abstract

The unbalanced data in the real classification task are mostly characterized by nonlinear distribution, and the traditional sampling method is not good at dealing with this kind of nonlinearity resulting in unsatisfactory sample classification effect. Aiming at this problem, an under-sampling method

*通讯作者。

文章引用: 王晓玲, 金永超, 刘威伟, 王希胤. 基于 KPCA 的不平衡数据欠抽样算法[J]. 应用数学进展, 2024, 13(9): 4108-4118. DOI: 10.12677/aam.2024.139392

based on KPCA is proposed. The method maps the original data to a suitable high-dimensional space to make it linearly divisible by nonlinearly transforming the kernel function, and de-redundantly removes the majority class by calculating the scores of individual samples on the kernel principal components in order to achieve the purpose of under-sampling. After the under-sampling preprocessing of nine datasets with different balance rates, the classification is performed using Logistic Regression classifier model. The experimental results show that the algorithm of this paper obtains the highest evaluation metrics under Accuracy, F1-measure and AUC value scores under 7, 8 and 9 groups of datasets, respectively, which shows that the method has a good classification performance on unbalanced datasets.

Keywords

Unbalanced Data, Under-Sampling, Kernel Principal Component Analysis, Classification

1. 引言

随着大数据技术的发展，数据的体量在飞速地增长。如何在海量的数据中精准、高效地挖掘有价值的信息是非常重要的。由于数据的多样性和复杂性，会出现许多不平衡数据，即在统计样本或数据集中不同类别的样本数量存在严重的不均衡情况。在二分类数据中，一种类别的样本数量远远大于其他类别的样本数量一般称为“多数类”，而其他类别则称为“少数类”。在实际应用中普遍存在不平衡数据分类问题，如：信贷评估[1]、医疗诊断[2]、诈骗检测[3]等。

目前对不平衡数据进行分类的处理包括两个层面：算法层面和数据层面[4]。算法上主要是通过发掘或改进算法使其对不平衡数据分类有效；数据上主要是在具体分类前对数据进行预处理使数据达到相对平衡，主要有对少数类进行过抽样、对多数类进行欠抽样，以及过、欠结合的混合抽样。本文主要关注欠抽样算法。

欠抽样是通过剔除部分多数类样本来保持数据平衡的。文献[5]提出，利用欠抽样算法对数据进行预处理是比利用过抽样算法效果要好。文献[6]提出了随机欠抽样方法，可随机地减少多数类样本来缩小少数类与多数类的样本差，进而得到与少数类样本数量相同的多数类样本达到样本相对平衡。文献[7]提出了基于聚类的欠抽样，使每一簇原始数据被其相应的聚类质心来代替从而减少多数类中数据点数使之与少数类数据点数达到相同。文献[8]提出了一种基于聚类分布 K-means 欠抽样算法，依据聚类分布和从多数类聚类到少数类聚类中心的距离来选择样本，保持原始分布的同时提高边界样本的采样率。这些方法旨在对原始不平衡数据进行预处理使样本类别更加均衡，从而提高模型的性能和泛化能力。但现实任务中数据大都呈现非线性、非正态的分布特点，传统的方法在处理不平衡数据的非线性问题时仍有不足。

1992 年核函数成功应用到了 SVM 中，核函数的重要性被人们所了解。1998 年，B. Schokopf 等人[9]基于核函数与主成分分析的有机结合，提出核主成分分析(KPCA)。核主成分分析能够捕捉到数据中的非线性结构，并将其转化为线性可分的形式，从而更好地实现数据的降维和特征提取等工作。在诸如模式识别、图像处理和数据挖掘等领域[10]-[15]，核主成分分析已被广泛应用，并在处理复杂数据集和提高分类性能方面发挥了重要作用。

基于不平衡数据的非线性问题，本文从数据层面出发提出一种基于 KPCA 的欠抽样算法。将原始数

据在非线性核函数映射的作用下变成高维空间中线性可分的数据，通过计算核矩阵并求解其特征值和相应的特征向量以及核主成分，进而将高维空间中的数据在核主成分上进行投影得出样本的综合评分，最后按照评分对多数类样本进行排名并删除排名靠前且相似度高的多数类样本从而达到欠抽样的目的，余下的多数类样本与少数类样本相组合可达到相对平衡的状态进而最大化的降低非平衡数据集的不平衡率。

2. 理论知识

2.1. 核函数

核方法的基本原理[14]为：当需要处理的变量为非线性时，引入核函数，它是一个非线性变换，用来将原始空间 R 中的线性不可分问题转化为高维特征空间 H 中的线性可分问题。图 1 展示了数据从输入空间映射到高维特征空间的过程。

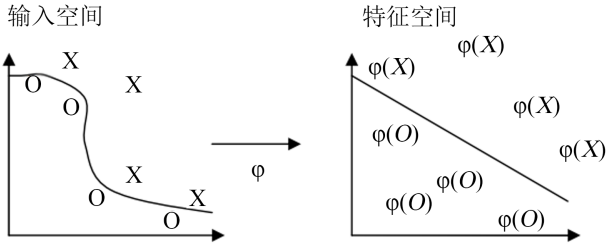

Figure 1. Nonlinear mapping
图 1. 非线性映射

核函数定义[14]：假设存在映射 φ，能将二元函数 $X^n \times X^n$，映射到特征空间 H 中，并且满足 $K(x,y)=(\varphi(x)\cdot\varphi(y))$，称 K 是核函数。常见的核函数包括多项式核函数、高斯径向核函数、多层感知机核函数等。

在本文的实际应用中采用了以下三种核函数，具体形式如下：

1) 多项式核函数

$$k(x,y)=(x\cdot y+c)^d \tag{1}$$

其中 $c\geq 0$，d 是整数，均为自定义参数。当参数 d 较小时，此函数的外推能力较强，但随着 d 的增大，外推能力则逐渐变弱。

2) 高斯径向核函数

$$k(x,y)=\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{2}$$

其中 $\sigma>0$ 是自定义的高斯核函数的带宽参数。它决定了高斯分布的标准差，即控制了核函数的平滑程度和波动范围。

3) 多层感知机核函数

$$k(x,y)=\tanh(-b(x\cdot y)-c) \tag{3}$$

其中 b，c 是自定义参数。

2.2. 核主成分分析原理

核主成分分析的核心思想[16] [17]是将核函数与主成分分析相结合，在非线性变换函数 φ 的作用下将

原始空间 R 中数据映射到特征空间 H，即 R 中的样本点 $X_1, X_2, \cdots, X_n$ 转变为 H 中的样本点 $\varphi(X_1), \varphi(X_2), \cdots, \varphi(X_n)$，并在 H 中进行主成分分析。

实现过程为：若样本集合中有 n 个样本点，用两者的内积 $\varphi(X_i)\cdot\varphi(X_j)$ 表示 R 中的两个数据 X_i 和 X_j 在 H 空间的距离。假设 H 空间中的样本 $\varphi(X_1), \varphi(X_2), \cdots, \varphi(X_n)$ 已经中心化，即

$$\sum_{k=1}^{n}\varphi(X_k)=0 \tag{4}$$

则在 H 空间中进行主成分分析，通过计算协方差矩阵

$$C=\frac{1}{n}\sum_{i=1}^{n}\varphi(X_i)\cdot\varphi(X_i)^{\mathrm{T}} \tag{5}$$

求得满足 $\lambda V=CV$ 的特征值 $\lambda(\geq 0)$ 以及相应的特征向量 $V(\in H)$。

由于 $V\in span\{\varphi(X_1), \varphi(X_2), \cdots, \varphi(X_n)\}$ [17]，则存在一组参数 $\alpha_1, \alpha_2, \cdots, \alpha_n$ 使

$$V=\sum_{j=1}^{n}\alpha_j\varphi(X_j),\ j=1,2,\cdots,n \tag{6}$$

故 $\lambda V=CV$ 就等价于 $\lambda(\varphi(X_k)\cdot V)=(\varphi(X_k)\cdot CV)\quad k=1,2,\cdots,n$。

通过定义 $n\times n$ 的核矩阵 K [17]，$K(X_i, X_j)=\varphi(X_i)\cdot\varphi(X_j)$，上述 $\lambda V=CV$ 等价为 $n\lambda\alpha=K\alpha$ 对核矩阵 K 进行特征分解得到其特征值 $\overline{\lambda}_1\geq\overline{\lambda}_2\geq\cdots\geq\overline{\lambda}_n(\overline{\lambda}=n\lambda)$，$\alpha_1, \alpha_2, \cdots, \alpha_n$ 为相应特征向量。则 $\varphi(X)$ 在 H 空间向量 V^k 上的投影为：

$$(V^k\cdot\varphi(X))=\sum_{i=1}^{n}\alpha_i^k(\varphi(X_i)\cdot\varphi(X))=\sum_{i=1}^{n}\alpha_i^k K(X_i, X) \tag{7}$$

此外，当假设 $\sum_{k=1}^{n}\varphi(X_k)=0$ 不成立时，只需将矩阵 K 调整为

$$\hat{K}=K-I_nK-KI_n+I_nKI_n \tag{8}$$

其中 $(I_n)_{ij}=1/n$。

3. 基于 KPCA 欠抽样的不平衡数据分类算法

本文采用基于 KPCA 的欠抽样算法对不平衡数据进行预处理，以便可以初步帮助解决传统方法不善于解决现实任务中的非线性问题。经此算法预处理后的数据最大可能达到多数类与少数类相对平衡的状态。基于 KPCA 欠抽样算法的实现步骤如下：

1) 建立初始样本矩阵

设有 m 个待评价样本，n 个属性特征。假设第 i 个样本在第 j 个属性下的取值为 x_{ij}，建立初始矩阵 S 为：

$$S=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

初始样本矩阵归一化处理：

$$S' = \begin{pmatrix} x'_{11} & x'_{12} & \cdots & x'_{1n} \\ x'_{21} & x'_{22} & \cdots & x'_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x'_{i1} & \cdots & x'_{ij} & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ x'_{m1} & x'_{m2} & \cdots & x'_{mn} \end{pmatrix}$$

上述式子中：

$$x'_{ij} = \frac{x_{ij} - \min\left(x_j\right)}{\max\left(x_j\right) - \min\left(x_j\right)}, i = 1,2,\cdots,m; j = 1,2,\cdots,n. \tag{9}$$

2) 依据数据特点选取核函数及核参数[18]

选择与样本数据特点贴切的核函数并设置相对最优的参数值将 R 空间数据映射到 H 空间中，可得到核矩阵 K，其中 $K_{ij} = \varphi\left(S'_i\right)\cdot\varphi\left(S'_j\right)$， S'_i 和 S'_j 代表矩阵 S' 的第 i 行和 j 列所对应的数。

3) 中心化核矩阵

中心化后的核矩阵记为 K_c，其中元素表达式为 $K_{cij} = K_{ij} - \bar{K}_i - \bar{K}_j + \bar{K}$， $\bar{K}$ 为核矩阵中所有元素是均值。

4) 求解特征值和特征向量

对中心化的核矩阵 K_c 利用 $K_c V_k = \lambda_k V_k$ 进行特征分解，得到核矩阵 K_c 的特征值 λ_k 和相应的特征向量 V_k。

5) 确定特征空间的维数 t

t 满足下式的最小值

$$\frac{\sum_{k=1}^{t}\lambda_k}{\sum_{k=1}^{n}\lambda_k} \geq 0.8 \tag{10}$$

6) 计算样本点 X 的主成分

计算投影：

$$\left(V_k \cdot \varphi\left(X\right)\right) = \sum_{j=1}^{n}\alpha_j^k K\left(X_j, X\right) \tag{11}$$

其中， α_j^k 为特征向量 V_k 的第 j 个元素， $\left(V_k \cdot \varphi\left(X\right)\right)$ 为数据点 X 的第 k 个主成分。

7) 计算样本主成分因子得分 Y_{ij}

其中， $i = 1,2,\cdots,m$， $j = 1,2,\cdots,t$ 将原始样本数据点投影到主成分上，得到相应的投影坐标值即为样本主成分因子得分。 Y_{ij} 表示为对于第 i 个样本数据点在第 j 个主成分上的因子得分。

8) 最终得出每个样本的得分 $F_i, i = 1,2,\cdots,m$

样本得分为主成分因子得分的线性组合， $F_i = \omega_1 Y_{i1} + \omega_2 Y_{i2} + \cdots + \omega_t Y_{it}$，其中， $Y_{i1}, Y_{i2}, \cdots, Y_{it}$ 是主成分因子得分， $\omega_1, \omega_2, \cdots, \omega_t$ 是每个主成分的权重系数，用来表达主成分在样本的综合得分中的贡献程度。

9) 根据每个样本评分 F_i 对样本进行降序排列并删除前 r 个多样本数据来达成欠抽样的目的。其中，r 为多数类与少数类个数差。

算法流程图如图 2 所示：

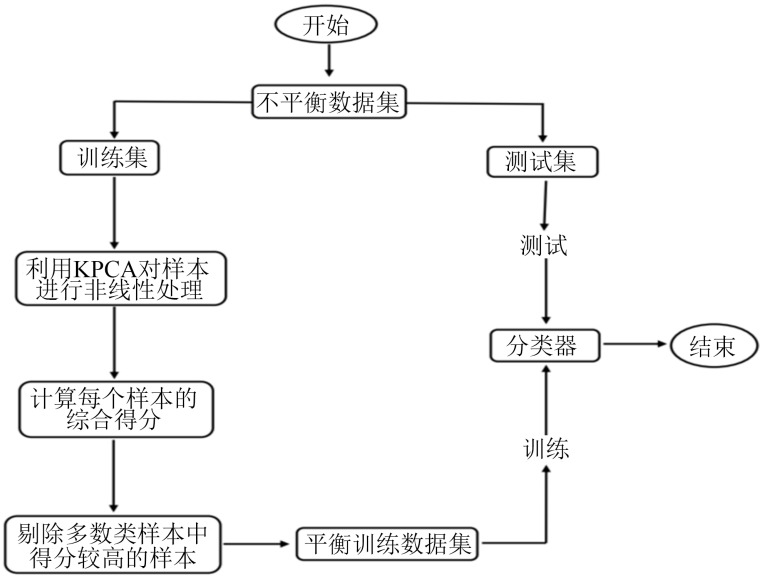

Figure 2. Flow chart based on KPCA under-sampling algorithm
图 2. 基于 KPCA 欠抽样算法流程图

4. 实验结果与分析

4.1. 实验数据

本文为验证基于 KPCA 欠抽样算法在不平衡数据集上的分类效果，使用 UCI 数据库中 9 组数据集进行实验，这 9 组数据集的不平衡度呈递增形式且范围由 1.79%~33.74%，实验数据集信息如表 1。此外部分数据是多分类数据集，本实验运用一种应用广泛的转化方法[19]将多分类数据集中的某些类别合并使之成为二分类数据集。例如一个 7 分类数据集 steel-plates，steel-plates_0 表示将数据集中标签为 0 的样本看作少数类，那么多数类则是由标签 1 到 6 这六类样本合并而成。

Table 1. Experimental datasets description
表 1. 实验数据集描述

编号	数据集	样本总数	属性	多数类样本数	少数类样本数	不平衡度/%
1	Ionosphere	351	34	225	126	1.79
2	QSAR	1055	41	699	356	1.96
3	German Credit	1000	24	700	300	2.33
4	Wall-following_1	5456	24	4630	826	5.61
5	Steel-plates_1	1941	27	1751	190	9.22
6	Steel-plates_0	1941	27	1783	158	11.28
7	Wall-following_3	5456	24	5128	328	15.63
8	Thyroid_1	3772	21	3581	191	18.74
9	Ozone	2536	72	2463	73	33.74

4.2. 实验设计

本文采用五折交叉验证方法进行实验，即将每个数据集划分为 5 份，选取 4 份作为训练集，1 份作

为测试集，每个数据集重复实验 5 次，实验 5 次并以平均值作为最终的评价结果。分类器采用线性回归模型中的 Logistic Regression 模型。本文实验包括 3 种情况，具体如下：

1) 原始数据不进行欠抽样预处理，仅使用 Logistic Regression 分类器模型对不平衡数据集进行分类。

2) 采用随机欠抽样算法对原始数据进行预处理，再利用 Logistic Regression 分类器模型对预处理后的不平衡数据集进行分类。

3) 采用本文提出的基于 KPCA 的欠抽样方法对不平衡数据集进行预处理，再利用 Logistic Regression 分类器模型对预处理后的不平衡数据集进行分类。

4.3. 评价指标

本文采用基于混淆矩阵分析的方法来对不平衡数据集分类器性能进行评价。此方法在当少数类样本远小于多数类样本时，可以适当减少分类器把所有的样本都分类为多数类。因此，本文采用准确率(Accuracy)、F1-measure 和 AUC 值作为评价指标测试 3 种实验情况的分类结果。

Table 2. Confusion matrix
表 2. 混淆矩阵

		预测值		计数
		Positive	Negative	
真实值	Positive	*TP*	*FN*	*P*
		(True positive)	(False Negative)	
	Negative	*FP*	*TN*	*N*
		(False Positive)	(True Negative)	

由表 2 可看出数据的预测值与真实值的组合有四种：分别为：预测和真实同为正例、预测和真实同为负例、预测为正例真实为负例、预测为负例真实为正例。*TP*、*TN*、*FP*、*FN* 分别表示其对应样本个数，有 $TP+TN+FP+FN=$ 样本总数。

1) Accuracy：所有结果中，预测正确的结果所占的比例，计算公式为：

$$\frac{TP+TN}{TP+FP+TN+FN} \tag{12}$$

2) F1-measure：综合了查准率(Precision)和召回率(Recall)，是他们的调和平均，用来反应模型的稳健性，计算公式为：

$$F1=\frac{2TP}{2TP+FN+FP} \tag{13}$$

3) AUC：AUC [20] (Area Under Curve)是 ROC 曲线下的面积，用来衡量算法的性能。

ROC 曲线是受试者特征曲线(Receiver Operating Characteristic Curve)是一种用于反映诊断检查方法敏感性和特异性特征的曲线图。ROC 曲线的纵轴是真正例率(Ture Positive Rate，简称 *TPR*)公式为：

$$TPR=\frac{TP}{TP+FN} \tag{14}$$

横轴是假正例率(False Positive Rate，简称 *FPR*)公式为：

$$FPR=\frac{FP}{FP+TN} \tag{15}$$

4.4. 结果分析

实验分析了三种情况的 Accuracy、F1-measure、AUC。表 3 列出了三种算法在各个数据集下的评价指标值，其中原数据不进行预处理直接分类的方法缩写为 DTC；随机欠抽样缩写为 RUS；加黑数字为同组最大值。

由表 3 可以看出基于 KPCA 的欠抽样方法与其他算法相比，整体来看在绝大多数数据集上都表现的不错。在 AUC 值指标方面：相比其他方法本文方法体现出其稳定的优势，在各个数据集中都有不同的提升。在指标 Accuracy 和 F1-measure 值上：本文方法在大部分数据集上表现依旧不错，但在极个别数据集上表现略微不足。

为了更好的观察对比表 3 中的三种方法在这 9 组数据集上三种指标上的表现，下面采用 3 个柱状图形式直观的展示 3 个评分的对比结果。

Table 3. Classification of Accuracy, F1-measure and AUC of different classification algorithms on unbalanced datasets
表 3. 不同分类算法在不平衡数据集上的 Accuracy、F1-measure 值、AUC 值对比

DataSet	Accuracy			F1-measure			AUC		
	DTC	RUS	KPCA	DTC	RUS	KPCA	DTC	RUS	KPCA
Ionosphere	0.8592	0.8553	0.8868	0.8958	0.8608	0.9118	0.8945	0.8968	0.9446
QSAR	0.8107	0.7804	0.8364	0.6591	0.7707	0.8548	0.8931	0.8632	0.8987
German Credit	0.7600	0.7611	0.7667	0.5200	0.7485	0.7879	0.8130	0.8359	0.8430
Wall-following_1	0.8638	0.8327	0.8617	0.3890	0.8421	0.8431	0.8681	0.8848	0.8879
Steel-plates_1	0.8823	0.8684	0.8947	0.8167	0.8673	0.900	0.9449	0.9458	0.9495
Steel-plates_0	0.9451	0.7895	0.8438	0.9713	0.7872	0.8333	0.9317	0.8754	0.9326
Wall-following_3	0.9224	0.9087	0.9316	0.4800	0.9062	0.9384	0.9565	0.9589	0.9643
Thyroid_1	0.6468	0.6522	0.6609	0.6593	0.6721	0.6777	0.7158	0.7112	0.7264
Ozone	0.8630	0.8636	0.8636	0.8523	0.8421	0.8636	0.8920	0.9146	0.9250

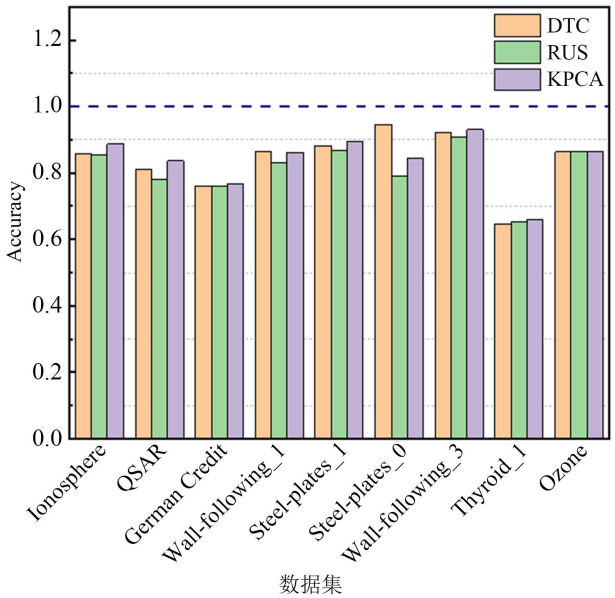

Figure 3. Accuracy comparison of the three algorithms
图 3. 三种算法的 Accuracy 对比

图 3 给出了三种方法的 Accuracy 的对比结果，可以看出本文算法在其中 6 组数据集中相对其他两种算法均有小提升，但在 Wall-following_1 和 Steel-plates_0 数据集中，本文算法的表现力不如 DTC 方法且略有降低。但由图 3 知在 Wall-following_1 数据中虽然其 Accuracy 略有下降但与 F1-measure 的进步相比，Accuracy 下降的不明显。而在 Steel-plates_0 数据集中图 3 和图 4 结果表明本文算法与原始数据直接进行分类的方法相比在利用 KPCA 进行预处理时对于特征的提取时可能导致了一些关键信息丢失，减低了多数类的分类准确率以及召回率使得模型整体 Accuracy 与 F1-measure 值均比不进行预处理直接分类的方法略低一些。

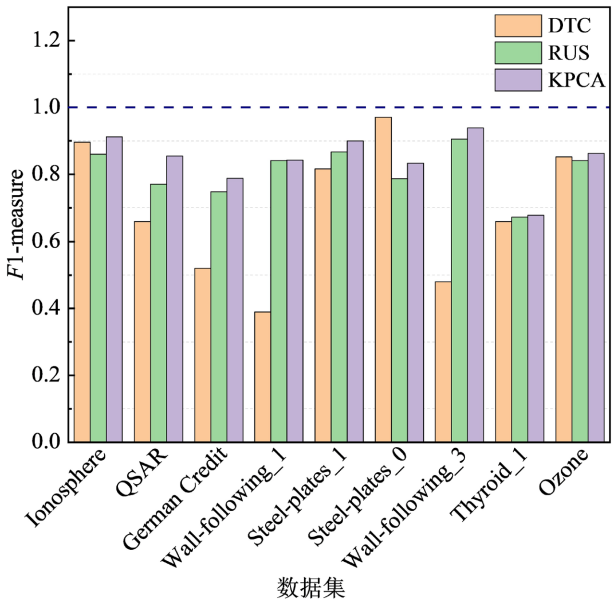

Figure 4. F1-measure comparison of the three algorithms
图 4. 三种算法的 F1-measure 对比

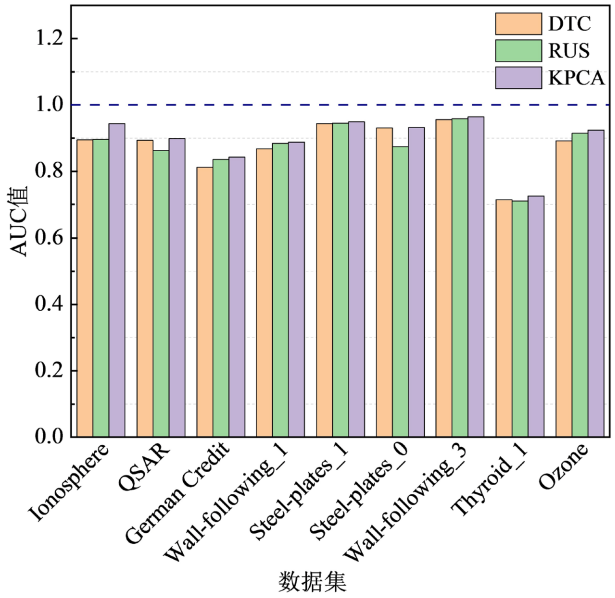

Figure 5. AUC comparison of the three algorithms
图 5. 三种算法的 AUC 值对比

图 4 给出了三种算法在 F1-measure 上的对比结果，可以看出除了上述提到的第六个数据集外，与其他两种方法相比，基于 KPCA 欠抽样算法在各个数据集上均有所提高。尤其在第 2、3、4 以及第 7 个数据集上相对于其他两种方法表现优秀，相对于最低评分平均提升了 68%。因此，在 F1-measure 方面有显著进步，进而进一步验证了本文方法的有效性。

图 5 给出了三种方法的 AUC 值对比结果，清楚的观察到相对于其他两种方法，本文方法在这 9 组数据集上的表现一贯优秀，在每个数据集上均有不同程度的提升。其稳定性可以得出此方法在处理不平衡数据集时能力更强，具有一定的鲁棒性。同时也验证了此方法有效性和稳定性。

5. 结论

本文通过对传统的欠抽样算法不擅长处理的不平衡数据具有的非线性问题进行研究：提出一种基于 KPCA 的欠抽样算法，根据得到的样本评分进行剔除评分较高的多数类样本，使不平衡数据集的不平衡率尽可能的降低。在此基础上，将 9 组不同非平衡率的数据集进行欠抽样处理后，采用 Logistic Regression 分类器模对预处理后的不平衡数据集进行分类预测。实验结果显示，与现有的欠抽样方法相比，本文方法在 Accuracy、F1-measure 值和 AUC 值三个方面都有显著提升。然而，在某些数据集中，本文方法在特征提取阶段可能会导致多数类样本的关键信息被删除，从而降低了分类器的整体分类准确度。因此，解决这一问题将成为下一步研究的重点。

基金项目

本文得到了国家自然科学基金项目(32070669)和唐山市人才资助项目(16013601)的支持。

参考文献

[1] Ileberi, E., Sun, Y. and Wang, Z. (2022) A Machine Learning Based Credit Card Fraud Detection Using the GA Algorithm for Feature Selection. *Journal of Big Data*, **9**, Article No. 24. https://doi.org/10.1186/s40537-022-00573-8

[2] Shilaskar, S., Ghatol, A. and Chatur, P. (2017) Medical Decision Support System for Extremely Imbalanced Datasets. *Information Sciences*, **384**, 205-219. https://doi.org/10.1016/j.ins.2016.08.077

[3] Zakaryazad, A. and Duman, E. (2016) A Profit-Driven Artificial Neural Network (ANN) with Applications to Fraud Detection and Direct Marketing. *Neurocomputing*, **175**, 121-131. https://doi.org/10.1016/j.neucom.2015.10.042

[4] 李昂, 韩萌, 穆栋梁, 等. 多类不平衡数据分类方法综述[J]. 计算机应用研究, 2022, 39(12): 3534-3545.

[5] Kubat, M., Hotle, R. and Matwin, S. (1997) Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In: Fisher, D.H., Ed., *International Conference on Machine Learning*, Morgan Kaufmann Publishers, 179-186.

[6] Sowah, R.A., Agebure, M.A., Mills, G.A., Koumadi, K.M. and Fiawoo, S.Y. (2016) New Cluster Undersampling Technique for Class Imbalance Learning. *International Journal of Machine Learning and Computing*, **6**, 205-214. https://doi.org/10.18178/ijmlc.2016.6.3.599

[7] Lin, W., Tsai, C., Hu, Y. and Jhang, J. (2017) Clustering-Based Undersampling in Class-Imbalanced Data. *Information Sciences*, **409**, 17-26. https://doi.org/10.1016/j.ins.2017.05.008

[8] Song, A. and Xu, Q. (2018) Imbalanced Data Classification Based on MBCDK-Means Undersampling and GA-ANN. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L. and Maglogiannis, I., Eds., *Artificial Neural Networks and Machine Learning—ICANN* 2018, 349-358. https://doi.org/10.1007/978-3-030-01421-6_34

[9] Twining, C.J. and Taylor, C.J. (2003) The Use of Kernel Principal Component Analysis to Model Data Distributions. *Pattern Recognition*, **36**, 217-227. https://doi.org/10.1016/s0031-3203(02)00051-1

[10] 陈祥涛, 张前进. 基于核主成分分析的步态识别方法[J]. 计算机应用, 2011, 31(5): 1239.

[11] Rosipal, R. and Girolami, M. (2001) An Expectation-Maximization Approach to Nonlinear Component Analysis. *Neural Computation*, **13**, 505-510. https://doi.org/10.1162/089976601300014439

[12] 赵丽红, 孙宇舸, 蔡玉, 等. 基于核主成分分析的人脸识别[J]. 东北大学学报: 自然科学版, 2006, 27(8): 847-850.

[13] Dachapak, C., Kanae, S., Yang, Z. and Wada, K. (2003) Kernel Principal Component Regression in Reproducing Kernel Hilbert Space. *Proceedings of the ISCIE International Symposium on Stochastic Systems Theory and Its Applications*,

2003, 213-218. https://doi.org/10.5687/sss.2003.213

[14] Zelias, A.J. (19921) Multicollinearity of Variables an Embarrassing Problem of Econometrics. Krakow Academy of Economics.

[15] 雷银香, 熊科云. 中医药领域不平衡数据的特征选择和分类方法研究[J]. 信息与电脑, 2023, 35(24): 55-57.

[16] 潘继斌. 核函数的概念、性质及其应用[J]. 湖北师范学院学报(自然科学版), 2007, 27(1): 10-12.

[17] 吴今培. 基于核函数的主成分分析及应用[J]. 系统工程, 2005, 23(2): 117-120.

[18] 陈将宏, 张渊渊. 核主成分分析中核参数选择的遗传算法[J]. 计算机与现代化, 2011(11): 1-2, 14.

[19] Rivera, W.A. (2017) Noise Reduction a Priori Synthetic Over-Sampling for Class Imbalanced Data Sets. *Information Sciences*, **408**, 146-161. https://doi.org/10.1016/j.ins.2017.04.046

[20] Ling, C.X., Huang, J. and Zhang, H. (2003) AUC: A Statistically Consistent and More Discriminating Measure than Accuracy. *International Joint Conference on Artificial Intelligence* 2003, Acapulco, 9-15 August 2003, 519-524.