文章编号:1671-6833(2023)04-0022-07

复合可靠性分析下的不平衡数据证据分类

田鸿朋,张 震,张思源,肖宗荣,董佳兵

(郑州大学 电气与信息工程学院,河南 郑州 450001)

摘 要:针对传统分类模型在处理不平衡数据时会侧重于大类而忽略小类的问题,提出了一种复合可靠性分析下的不平衡数据证据分类方法,通过评估分类模型的全局可靠性和局部可靠性来提升模型对每个不平衡测试样本的分类能力。首先,对大类多次降采样,采样后的数据与小类组成多个训练子集,用这些子集训练得到多个分类模型,通过最大均值差异度量采样前后数据分布的差异性得到不同分类模型的全局可靠性。其次,利用待测样本在训练集中的近邻来评估其分类结果的局部可靠性,待测样本与其近邻具有相似的数据分布和空间结构,分类模型对近邻的分类结果与真实类别偏差越小,其局部可靠性就越大。最后,在证据推理框架下,全局可靠性与局部可靠性组合为复合可靠性因子对不同分类模型得到的分类结果进行折扣,将部分概率值分配给完全未知类来表征数据类别的不确定性,用 Dempster-Shafer (DS)规则融合多个折扣后的分类结果做决策分析。实验结果表明:所提方法对 KEEL 和 UCI 数据库的 12 个不平衡数据分类结果的平均 *FM* 为 80.18%,*GM* 为 87.24%,相较于其他不平衡数据分类方法中最优结果分别高出 8.1%和 4.99%。所提方法的有效性在不平衡数据分类中得到了证实。

关键词:不平衡数据; 分类; 全局可靠性; 局部可靠性; 证据推理

中图分类号:TP181 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2023.04.012

不平衡数据是指数据集不同类间的数据分布不均衡,其中某一类或多类的样本数量远远超过其他类[1]。例如,在网络安全中,异常行为通常只占一小部分,而大多数行为都是正常的。一般来说,相较于大类中的样本,小类中的样本可能携带罕见但更有用的信息[2]。然而,一些基础分类模型,如支持向量机(SVM)分类器[3],致力于最大限度提高整体分类精度,大多数小类样本被分配到大类,因此对不平衡数据分类的性能较差[4]。

近年来,不平衡数据分类问题备受关注,研究成果丰硕,大致可以分为 3 类:采样方法[5]、代价敏感学习方法[6]和集成学习方法[7]。采样方法侧重于对不平衡数据进行预处理,使类间数据达到均衡,然后就可以用基础分类模型分类。代价敏感的学习方法是对小类样本分配较高权重,这样可以减小对其分类的错误率。集成学习方法结合了不同子集训练的多个分类模型,相互间提供了互补性信息,解决了单一分类模型对不平衡数据分类性能较差的问题。

虽然以上的不平衡数据分类方法在一些场景中具有一定的有效性,但是这些方法只考虑全局最优,并不一定适合每个测试数据。例如,位于不同类别重叠区域的数据就难以准确划分,这些数据分布的特殊性导致其类别存在一定的不确定性[8],因此容易被错误分类。

针对上述问题,提出了一种复合可靠性分析下的不平衡数据证据分类方法。首先,为了评估分类模型的整体性能,本文设计了一种全局可靠性评估策略,在对大类多次降采样的基础上,分别训练多个分类模型,通过计算采样前后数据分布的差异性来评估不同分类模型的全局可靠性。其次,为了提升分类模型对每个测试数据的局部分类性能,本文设计了一种局部可靠性评估策略,通过评估分类模型对测试数据邻域分类结果的性能

收稿日期:2023-02-21;**修订日期**:2023-03-23

基金项目:河南省重大公益专项(201300311200)

作者简介:田鸿朋(1996—),男,河南南阳人,郑州大学博士研究生,主要从事人工智能、模式识别的研究,E-mail:tianhongpeng1220@163.com。

通信作者:张震(1966—),男,河南郑州人,郑州大学教授,博士,博士生导师,主要从事多媒体信息安全、图像处理、模式识别的研究,E-mail:zhangzhen66@126.com。

引用本文:田鸿朋,张震,张思源,等. 复合可靠性分析下的不平衡数据证据分类[J]. 郑州大学学报(工学版),2023,44(4):22-28. (TIAN H P, ZHANG Z, ZHANG S Y, et al. Imbalanced data evidential classification with composite reliability [J]. Journal of Zhengzhou University (Engineering Science),2023,44(4):22-28.)

来量化该测试数据被正确分类的可靠性,这样能够使分类模型对每个测试数据实现局部最优。最后,在证据推理框架下[9],结合全局可靠性因子和局部可靠性因子对不同分类模型的结果做折扣融合并决策分析。

1　相关工作

1.1　基于采样的方法

采样方法侧重于对输入数据进行预处理,以解决不平衡的问题,大致可以分为过采样和降采样两种方法。过采样方法[10-11]致力于生成数据来增加小类中的样本数量。例如,随机过采样方法(random over-sampling approach, ROS)[10]通过随机抽样的方式选取少量数据,然后将所选样本添加到原始小类中。降采样方法[12-13]致力于去除一部分样本来减少大类样本的数量。例如,随机欠采样方法(random undersampling approach, RUS)[12]随机选取一些与小类样本数量相同的大类样本,去掉其他样本,可能会丢失一些重要信息。

1.2　基于集成学习的方法

与采样方法不同,集成学习方法是将多个互补分类器组合在一起,以提高分类模型的整体性能。EasyensEmble 是 Liu 等[14]提出的一种带双层结构的集成学习算法,先将大类进行欠采样,采样出多个子集分别与小类结合成多个子训练集,再以 AdaBoost 算法对每个子训练集进行训练从而生成基础分类模型。

2　本文方法

2.1　全局可靠性分析

假设测试集 $X=\{x_1,x_2,\cdots,x_H\}$ 在辨识框架 $\Omega=\{\omega_{\min},\omega_{\mathrm{maj}}\}$ 下被训练集 $Y=\{y_1,y_2,\cdots,y_G\}$ 训练得到的模型分类, Y_{maj}、$Y_{\min}$ 分别表示数据量多的大类样本集合和数据量少的小类样本集合。

为了均衡数据的不同类别,本文使用一种多重降采样方法对 Y_{maj} 的数据降采样,从而得到 T 个子集 $\{Y^1_{\mathrm{maj}},Y^2_{\mathrm{maj}},\cdots,Y^T_{\mathrm{maj}}\}$,$T$ 值计算如下:

$$T=[IR]; \tag{1}$$

$$IR=\frac{|Y_{\mathrm{maj}}|}{|Y_{\min}|}。 \tag{2}$$

式中:IR 为可用于测量不平衡数据不平衡程度的不平衡比率;$|\cdot|$ 表示势,用于统计集合中元素数量;$[\cdot]$ 为四舍五入符号。

每个子集的样本数与 $Y_{\min}$ 的样本数相同,并与 Y_{maj} 组合生成新的训练子集 $\{Y^1,Y^2,\cdots,Y^T\}$。每个训练子集训练一个基本分类模型可用于对待测样本 x_i 分类,定义为

$$\boldsymbol{P}^t_i=\Gamma^t(x_i\mid Y^t)。 \tag{3}$$

式中:$\Gamma^t(\cdot)$ 表示基础分类模型;Y^t 表示第 t 个训练子集;$\boldsymbol{P}^t_i=[p^t_i(\omega_{\min}),p^t_i(\omega_{\mathrm{maj}})]$ 表示 x_i 的分类结果。

由于分类任务的一个基本假设是训练数据和测试数据独立同分布,而对不平衡的训练数据多次降采样后得到多个训练子集的数据分布与原始分布不可避免地会存在一定的偏差,且不同的训练子集相对应的分布偏差也不同,因此由这些训练子集训练得到的分类模型也有不同的全局可靠性。本文利用最大均值差异 MMD 来计算采样前后数据分布差异性,从而评估不同训练模型得到的分类结果的可靠性。Y^t 训练模型得到分类结果 P^t_i 的全局可靠性 α_{it} 定义为

$$\alpha_{it}=\mathrm{e}^{-MMD(Y^t,Y^t_{\mathrm{maj}})}\Big/\sum_{t=1}^{T}\mathrm{e}^{-MMD(Y^t,Y^t_{\mathrm{maj}})}; \tag{4}$$

$$MMD(Y,Y^t_{\mathrm{maj}})=\left\|\frac{1}{N}\sum_{i=1}^{N}\phi(y_i)-\frac{1}{M}\sum_{j=1}^{M}\phi(y_j)\right\|^2。 \tag{5}$$

式中:N 为 Y_{maj} 中训练样本数量;M 为训练子集 Y^t_{maj} 的样本数量;$\phi(\cdot)$ 表示一个映射函数,通常为高斯核函数,主要是将 Y_{maj} 和 Y^t_{maj} 映射到同一空间,然后计算两个数据分布间的差异性。

可以看出,训练子集与原始数据分布间的差异性越大,对应分类结果的全局可靠性 α_{it} 越小。

全局可靠性反映的是分类模型的全局最优化性能,但是由于不同模型对于特定测试样本的分类结果不同,因此还需要评估模型对每个测试样本的局部性能,这将在下一节中详细介绍。

2.2　局部可靠性分析

如图 1 所示,一个两类的不平衡数据,每个样本有两维属性,对应于 x 轴坐标和 y 轴坐标。$Y_{\min}$ 和 Y_{maj} 的训练样本分别用正方形和圆形标记。两个测试样本(x_1、x_2)用五角星标记,并假设它们属于大类。从大类中采样两个子集 Y^1_{maj} 和 Y^2_{maj},每个子集样本数量与小类相同。将它们与 $Y_{\min}$ 相结合,生成平衡的训练集,从而分别训练得到不同的分类器 Γ^1 和 Γ^2。可以看出,Y^2_{maj} 对应的分类器 Γ^2 可以正确地将样本 x_2 分配给大类,但是该样本可能会被 Y^1_{maj} 对应的 Γ^1 错误分类。尽管 Γ^1 可能比 Γ^2 有更优的全局分类性能,但可能不适合某些样本(例如 x_2)的正确分类。

由于测试样本与其近邻具有相似的空间结构和数据分布,因此它们的分类结果也是相似的。本文通过评估不同分类模型对待测样本近邻的分类性能

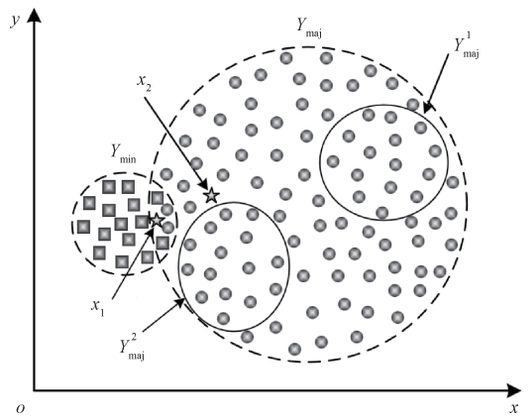

图1 不同分类模型的可靠性示意图

Figure 1 Illustration of the reliability of different classifiers

来量化模型对该样本分类的局部可靠性。基于上述分析,假设样本 x_i 在训练集 Y 中的 K 个近邻为 $\{y_1, y_2, \cdots, y_K\}$,分类模型 Γ^t 分类 x_i 得到的分类结果 P_i^t 的局部可靠性 β_{it} 定义为

$$\beta_{it} = \mathrm{e}^{-\lambda_{it}} \Big/ \sum_{t=1}^{T} \mathrm{e}^{-\lambda_{it}}; \tag{6}$$

$$\lambda_{it} = \sum_{k=1}^{K} \sqrt{\sum_{\omega_c \in \Omega} [p_k^t(\omega_c) - l_k(\omega_c)]^2}。 \tag{7}$$

式中:$p_k^t(\omega_c)$ 为近邻 y_k 的分类结果中数据类 ω_c 的概率值;向量 $\boldsymbol{L}_k = [l_k(\omega_{\min}), l_k(\omega_{\mathrm{maj}})]$ 表示 y_k 的类别标签。

可以看出,分类模型 Γ^t 分类 x_i 近邻的分类性能越好,其分类 x_i 得到分类结果的局部可靠性也就越大。

至此,样本 x_i 分类结果的全局可靠性以及局部可靠性都已计算得到,在下一节中会详细介绍如果有效地利用这些可靠性因子并融合多个分类结果。

2.3 复合可靠性折扣融合

样本 x_i 有 T 个分类结果 $\{P_i^1, P_i^2, \cdots, P_i^T\}$,在证据推理[9](Dempster-Shafer, DS)框架下每个分类结果都可以看作是一个证据源,每个证据源都有一个全局可靠性和局部可靠性。为了全面评估不同分类结果的可靠性,本文定义了一种复合可靠性因子 γ_{it}:

$$\gamma_{it} = \alpha_{it}\beta_{it} \Big/ \sum_{t=1}^{T} \alpha_{it}\beta_{it}。 \tag{8}$$

式中:α_{it} 和 β_{it} 分别为 x_i 第 t 个分类结果的全局可靠性和局部可靠性。

在获得复合可靠性因子后,用折扣融合方法将 x_i 的 T 个分类结果进行折扣融合,被折扣的分类结果定义如下:

$$\begin{cases} m_i^t(\omega_c) = \lambda_{it} p_i^t(\omega_c); \\ m_i^t(\Omega) = 1 - \sum\limits_{\omega_c \in \Omega} \lambda_{it} p_i^t(\omega_c)。 \end{cases} \tag{9}$$

式中:$m_i^t(\cdot)$ 表示基本信任值,类似于概率框架下的分类概率,它表示待测数据属于某一类别的置信度;Ω 表示全集,也被称为完全未知类。

通过折扣融合可以将 x_i 分类结果的部分概率值分配给完全未知类来表征数据类别的不确定性。复合可靠性因子越高,说明该分类结果越可靠,不确定性越小,被折扣掉的信息也就越少。经过折扣后的证据之间的冲突程度变小,在这种情况下就可以用基础的 DS 融合规则[15]结合多个折扣后的结果:

$$(m_i^1 \oplus m_i^2)(\omega_c) = \frac{\sum\limits_{\omega_u \cap \omega_v = \omega_c} m_i^1(\omega_u) m_i^2(\omega_v)}{1 - \sum\limits_{\omega_u \cap \omega_v = \phi} m_i^1(\omega_u) m_i^2(\omega_v)}。 \tag{10}$$

经 DS 融合后的结果 m_i 中还包含有完全未知类的基本信任值,为了便于决策,本文将 m_i 转换成 pignistic 概率[16] $BetP(\cdot)$ 进行最终决策,定义如下:

$$BetP(\omega_c) = \sum_{\omega_c \in \Omega} m_i(\omega_c) + \frac{1}{|\Omega|} m_i(\Omega)。 \tag{11}$$

式中:$|\Omega|$ 为 Ω 中元素的数量。最后,测试样本 x_i 可以分配到概率最大的类别。

本文所提出算法的计算步骤如下。

输入:测试集 $X = \{x_1, x_2, \cdots, x_H\}$、训练集 $Y = \{y_1, y_2, \cdots, y_G\}$、辨识框架 $\Omega = \{\omega_{\min}, \omega_{\mathrm{maj}}\}$、小类样本集合 $Y_{\min}$ 和大类样本集合 Y_{maj};

输出:测试集类别。

① for t = 1 to T

对训练集的大类数据 Y_{maj} 降采样;

用 Y_{maj} 训练分类模型测试样本分类;

依据式(4)、(5)计算分类的全局可靠性;

② endfor

③ for i = 1 to H

样本 x_i 在训练集中搜索近邻并分类;

④ for t = 1 to T

分类模型 Γ^t 对 x_i 的近邻分类;

依据近邻分类结果计算局部可靠性;

根据式(8)计算复合可靠性因子;

折扣分类结果 P_i^t;

⑤ endfor

根据式(10)融合多个分类结果;

根据式(11)做最终决策;

⑥ endfor

3　实验

将本文方法与其他几种相关的典型方法进行比较。本文中的所有实验都是在配备英特尔酷睿 i7-9750H 芯片和 16 G 运行内存的计算机上进行。主要用两个常用的不平衡分类指标 *FM* 和 *GM* 评价不同方法的性能。*FM* 和 *GM* 的值越高,该方法的性能就越好。

3.1　数据集

使用 KEEL 库和 UCI 数据库的 12 个不平衡数据集来测试和评估不同方法的性能。每个数据集使用五折交叉验证进行实验,能有效避免偶然性对结果的影响。这些数据集的基本信息如表 1 所示,包括数据集、数据库、样本数量、属性个数和不平衡比例 *IR*。

表 1　不平衡数据集基本信息

Table 1　Basic information of imbalanced datasets

数据集	数据库	样本数量	属性个数	*IR*
caesarian	UCI	80	5	1.35
ecoli2	KEEL	336	7	5.46
fertility_Diagnosis	UCI	100	9	7.33
glass1	KEEL	214	9	1.82
glass0123vs456	KEEL	214	9	3.20
Lymphography	UCI	142	18	1.33
new-thyroid1	KEEL	215	5	5.14
new-thyroid2	KEEL	215	5	5.14
shuttle-2vs5	KEEL	3 316	9	66.17
vehicle0	KEEL	846	18	3.25
vowel0	KEEL	988	13	9.98
page-blocks0	KEEL	5 472	10	8.79

3.2　对比方法

采用一些典型的不平衡数据分类方法做对比实验。ROS 和 SMOTE 是最具有代表性的过采样方法,被用来生成数据集来增加小类样本的数量,然后利用平衡训练集对基本分类模型训练。RUS 和 CBU 是欠采样方法,采用随机策略和聚类方法从大类样本中抽取一些样本来平衡数据集。RUSBOOST 和 EasyEnsemble 是集成学习方法,对原始数据集进行多次采样,并训练不同的分类器做出决策。此外,EATWSVM[17] 在 TWSVM 公式中引入了内核增强,从而对不平衡数据进行分类。

3.3　本文方法对比相关不平衡数据分类方法

实验中将本文方法与其他对比方法相比较。不同方法分类结果的 *FM* 值和 *GM* 值分别见表 2 和表 3,其中最后一行分别代表不同方法在不同数据集上得到的 *FM* 和 *GM* 的平均值。从实验结果可以看出,在大多数情况下,本文方法性能比其他不平衡数据分类方法更好;从平均值来看,本文方法的整体性能要明显优于其他方法,*FM* 和 *GM* 值相较于其他最优的不平衡数据分类方法分别高出 8.1% 和 4.99%。此外,不同方法在分类部分数据的 ROC 曲线和 *PR* 曲线如图 2 和图 3 所示。结果表明,本文方法的 ROC 曲线[18] 和 *P-R* 曲线下方面积要大于其他方法,这进一步证实本文方法在分类不平衡数据的有效性。分析原因是本文方法在处理数据不均衡问题时不仅能够评估不同分类模型的全局性能而且量化了这些模型对每个测试样本的局部分类性能,从而提升了模型的分类性能,并在一定程度上避

表 2　不同方法分类不同数据集的 *FM* 值

Table 2　*FM* values of different datasets classified by various methods

数据集	*FM*/%							
	ROS	RUS	SMOTE	CBU	RUSBOOST	EasyEnsemble	EATWSVM	本文方法
caesarian	61.21	59.50	65.85	64.35	62.08	61.35	55.71	67.23
ecoli2	71.92	70.75	70.92	71.90	71.93	71.23	59.64	73.09
fertility_Diagnosis	24.13	29.91	28.16	25.02	24.68	20.19	25.00	35.00
glass1	58.58	57.36	55.61	54.47	55.87	55.30	41.76	62.39
glass0123vs456	83.98	85.76	82.98	86.15	83.20	84.64	60.72	86.16
Lymphography	77.92	80.05	79.82	79.96	77.30	79.92	63.50	82.13
new-thyroid1	97.13	96.17	97.33	93.67	91.57	97.33	71.64	100.00
new-thyroid2	97.13	95.14	98.67	90.14	88.30	96.00	70.05	100.00
shuttle-2vs5	93.61	75.60	93.61	72.34	83.64	78.30	93.02	98.00
vehicle0	91.64	92.45	91.91	89.06	77.94	93.48	83.15	94.49
vowel0	80.11	73.07	81.74	70.87	75.87	71.01	64.82	82.69
page-blocks0	22.87	22.59	18.38	31.54	66.00	26.64	46.84	80.96
FM 平均值	71.69	69.86	72.08	69.12	71.53	69.62	61.32	80.18

表3 不同方法分类不同数据集的 *GM* 值

Table 3 *GM* values of different datasets classified by various methods

数据集	*GM*/%							
	ROS	RUS	SMOTE	CBU	RUSBOOST	EasyEnsemble	EATWSVM	本文方法
caesarian	60.60	58.97	63.72	63.85	60.67	59.75	52.28	66.31
ecoli2	90.58	89.77	90.21	90.15	88.16	89.98	79.89	90.93
fertility_Diagnosis	58.16	62.46	61.16	56.52	56.61	50.32	45.07	63.72
glass1	61.26	57.62	54.08	52.34	64.92	55.64	51.36	63.71
glass0123vs456	90.96	91.26	90.01	90.77	91.23	91.28	70.61	91.94
Lymphography	80.16	82.32	81.80	82.07	80.15	82.13	67.82	83.58
new-thyroid1	98.25	99.16	99.44	98.59	94.73	99.44	77.44	100.00
new-thyroid2	98.24	98.87	99.72	97.74	93.89	99.16	76.07	100.00
shuttle-2vs5	96.69	99.49	96.69	99.39	99.69	99.57	97.76	99.97
vehicle0	94.91	96.64	95.97	95.02	87.22	96.67	90.64	97.68
vowel0	95.50	95.71	96.75	94.43	86.19	94.86	70.41	97.80
page-blocks0	50.26	46.02	38.21	48.85	83.51	43.70	57.36	91.20
GM 平均值	81.29	81.52	80.65	80.81	82.25	80.21	69.74	87.24

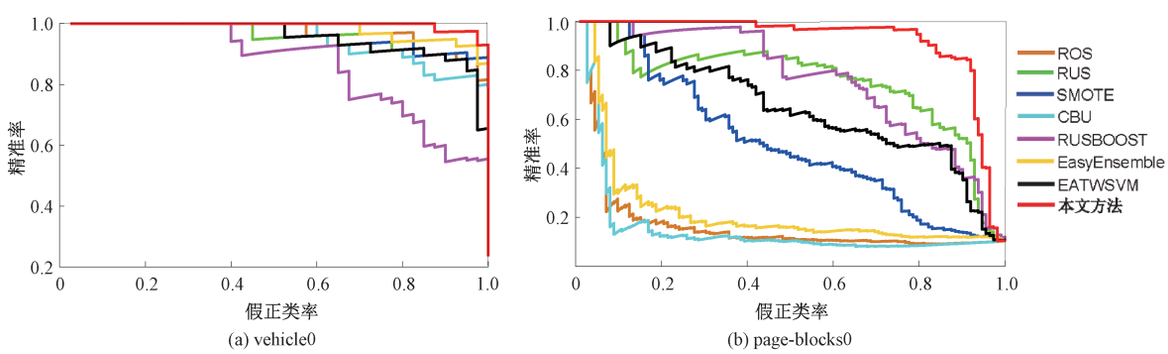

图2 不同方法分类不同数据集的 ROC 曲线

Figure 2 ROC curves of different methods in classifying various datasets

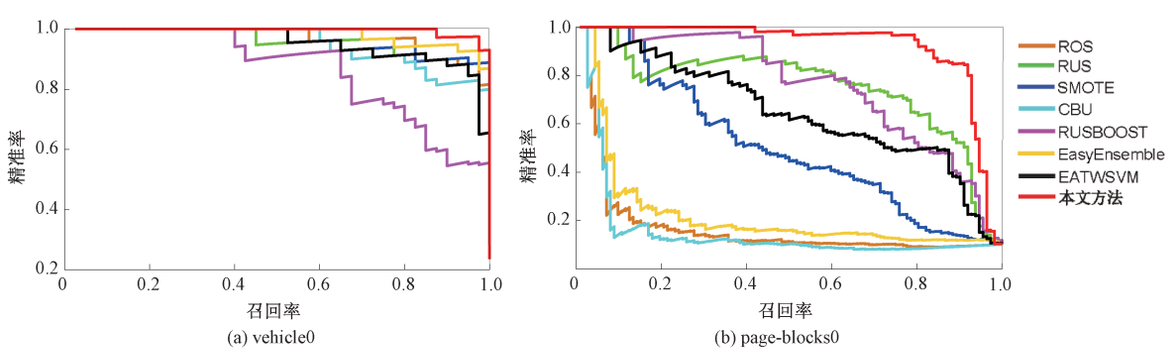

图3 不同方法分类不同数据集的 *P-R* 曲线

Figure 3 *P-R* curves of different methods in classifying various datasets

免了错误分类的风险。因此,本文方法相较于其他方法获得了较好的分类性能。

3.4 参数讨论

K 值是影响本文方法性能的关键参数,K 表示本文方法在评估模型局部可靠性时所需要找的近邻个数。图4所示为 K 值变化对本文方法在不同不平衡数据分类中性能的影响。其中横坐标为 K 的值,取值为3~15。实验结果表明,随着 K 的增加,本文方法在不同不平衡数据上分类性能的变化较小,这表明本文方法对 K 值具有一定的鲁棒性。在实际应用中推荐 $K\in[3,15]$ 作为该参数的取值范围,默认值为7。

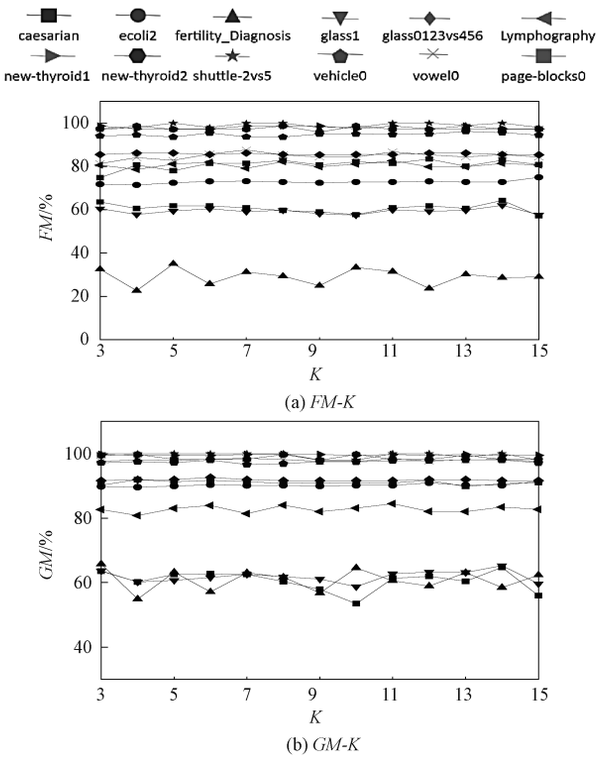

图 4　本文方法在 K 取不同值时的分类结果

Figure 4　Classification results of the proposed method with different K

4　结论

(1)提出了一种复合可靠性分析下的不平衡数据证据分类方法,该方法通过评估分类模型的全局可靠性和局部可靠性来提升模型对不平衡数据的分类性能。

(2)实验用公开的不平衡数据集验证了本文算法的有效性,FM 值和 GM 值相较于其他不平衡数据分类方法分别高出 8.1%和 4.99%。

(3)在未来的研究中,将尝试利用集成学习的思想结合不同采样方法的优势,研究更高性能的不平衡数据分类方法。

参考文献:

[1] 李艳霞, 柴毅, 胡友强, 等. 不平衡数据分类方法综述[J]. 控制与决策, 2019, 34(4): 673-688.
LI Y X, CHAI Y, HU Y Q, et al. Review of imbalanced data classification methods[J]. Control and Decision, 2019, 34(4): 673-688.

[2] 胡峰, 王蕾, 周耀. 基于三支决策的不平衡数据过采样方法[J]. 电子学报, 2018, 46(1): 135-144.
HU F, WANG L, ZHOU Y. An oversampling method for imbalance data based on three-way decision model[J]. Acta Electronica Sinica, 2018, 46(1): 135-144.

[3] 张震, 张英杰. 基于支持向量机与 Hamming 距离的虹膜识别方法[J]. 郑州大学学报(工学版), 2015, 36(3): 25-29.
ZHANG Z, ZHANG Y J. Iris recognition method based on support vector machine and Hamming distance[J]. Journal of Zhengzhou University (Engineering Science), 2015, 36(3): 25-29.

[4] 韩敏, 朱新荣. 不平衡数据分类的混合算法[J]. 控制理论与应用, 2011, 28(10): 1485-1489.
HAN M, ZHU X R. Hybrid algorithm for classification of unbalanced datasets[J]. Control Theory & Applications, 2011, 28(10): 1485-1489.

[5] 刘定祥, 乔少杰, 张永清, 等. 不平衡分类的数据采样方法综述[J]. 重庆理工大学学报(自然科学), 2019, 33(7): 102-112.
LIU D X, QIAO S J, ZHANG Y Q, et al. A survey on data sampling methods in imbalance classification[J]. Journal of Chongqing University of Technology (Natural Science), 2019, 33(7): 102-112.

[6] 孙艳歌, 邵罕, 杨艳聪. 基于代价敏感不平衡数据流分类算法[J]. 信阳师范学院学报(自然科学版), 2019, 32(4): 670-674.
SUN Y G, SHAO H, YANG Y C. Classification for imbalanced data streams based on cost-sensitive[J]. Journal of Xinyang Normal University (Natural Science Edition), 2019, 32(4): 670-674.

[7] 王乐, 韩萌, 李小娟, 等. 不平衡数据集分类方法综述[J]. 计算机工程与应用, 2021, 57(22): 42-52.
WANG L, HAN M, LI X J, et al. Review of classification methods for unbalanced data sets[J]. Computer Engineering and Applications, 2021, 57(22): 42-52.

[8] ZHANG Z W, TIAN H P, YAN L Z, et al. Learning a credal classifier with optimized and adaptive multiestimation for missing data imputation[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(7): 4092-4104.

[9] SHAFER G. A mathematical theory of evidence[M]. Princeton: Princeton University Press, 1976.

[10] HE H B, GARCIA E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[11] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[12] GUO H X, LI Y J, SHANG J, et al. Learning from class-imbalanced data: review of methods and applications[J]. Expert Systems With Applications, 2017, 73: 220-239.

[13] LIN W C, TSAI C F, HU Y H, et al. Clustering-based undersampling in class-imbalanced data[J]. Information Sciences, 2017, 409-410: 17-26.
[14] LIU X Y, WU J X, ZHOU Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems, Man, and Cybernetics Part B, Cybernetics: a Publication of the IEEE Systems, Man, and Cybernetics Society, 2009, 39(2): 539-550.
[15] CHALLA S, KOKS D. Bayesian and Dempster-Shafer fusion[J]. Sādhanā, 2004, 29: 145-174.
[16] SMETS P. Decision making in the TBM: the necessity of the pignistic transformation[J]. International Journal of Approximate Reasoning, 2005, 38(2): 133-147.
[17] JIMENEZ-CASTAÑO C, ALVAREZ-MEZA A, OROZCO-GUTIERREZ A. Enhanced automatic twin support vector machine for imbalanced data classification [J]. Pattern Recognition, 2020, 107: 107442.
[18] 逯鹏, 李奇航, 尚莉伽, 等. 基于优化极限学习机的CVD预测模型研究[J]. 郑州大学学报(工学版), 2019, 40(2): 1-5.
LU P, LI Q H, SHANG L J, et al. A CVD prediction model based on optimized extreme learning machine[J]. Journal of Zhengzhou University (Engineering Science), 2019, 40(2): 1-5.

Imbalanced Data Evidential Classification with Composite Reliability

TIAN Hongpeng, ZHANG Zhen, ZHANG Siyuan, XIAO Zongrong, DONG Jiabing

(School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract: To address the problem that traditional classification models mainly focused on majority class while ignoring minority class for classifying imbalanced data, an imbalanced data evidential classification method with composite reliability was proposed. This method could improve the classification ability of the model for each imbalanced test sample by evaluating the global reliability and local reliability of the classification model. Firstly, the method implemented under-sampling for majority class multiple times. The sampled subsets combined with minority class to form multiple training subsets. Multiple classification models were trained using these subsets. The maximum mean discrepancy measured the difference of data distribution before and after sampling, which could measure global reliability of the classification results obtained by classification models. Then, the local reliability of the classification result of the test sample in the training set was evaluated by using its nearest neighbors. The test sample and its nearest neighbors had similar data distribution and spatial structure. The smaller the deviation between the classification result of the classification model and the ground truth, the greater the local reliability of the classification result obtained by the classification model. Finally, under the framework of evidential reasoning, the global reliability and local reliability were combined as composite reliability factors to discount the classification results obtained from different classification models. Partial probability values were assigned to completely unknown classes to represent the uncertainty of classes. Dempster-Shafer (DS) rule was employed to fuse the classification results after multiple discounts for decision analysis. The experimental results showed that the average *FM* and *GM* of the proposed method for the classification of 12 imbalanced data in KEEL and UCI database were 80.18% and 87.24%, respectively, which were 8.1% and 4.99% higher than those of other best imbalanced data classification methods, respectively. This proved the effectiveness of the proposed method in classifying imbalanced data.

Keywords: imbalanced data; classification; global reliability; local reliability; evidential reasoning