计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.180601099

基于多类邻域三支决策模型的不平衡数据分类

向 伟[1] 王新维[2]
1 四川文理学院智能制造学院 四川 达州 635000
2 四川大学计算机学院 成都 610065

摘 要 不平衡数据分类是一种重要的数据分类问题。对于不平衡数据中规模较小的类,传统的分类算法的分类效果较差。对此,提出一种多类邻域三支决策模型的不平衡数据分类算法。首先,将传统的三支决策在混合数据和多个类的情形下进行推广,提出了混合数据的多类邻域三支决策模型;然后,在该模型中给出一种自适应代价函数的设定方法,并基于该方法提出了多类邻域三支决策模型的不平衡数据分类算法。仿真实验的结果表明,所提出的分类算法对于不平衡数据具有更好的分类性能。

关键词:不平衡数据;分类;三支决策;代价函数;自适应
中图法分类号 TP18

Imbalance Data Classification Based on Model of Multi-class Neighbourhood Three-way Decision

XIANG Wei[1] and WANG Xin-wei[2]
1 School of Intelligent Manufacturing, Sichuan University of Arts and Science, Dazhou, Sichuan 635000, China
2 College of Computer Science, Sichuan University, Chengdu 610065, China

Abstract Imbalance data classification is an important data classification problem, traditional classification algorithm does not have better classification effect for smaller class in imbalance data. Therefore, this paper proposed an algorithm of imbalance data classification based on multi-class neighbourhood three-way decision. In the case of mixed data and multiple classes, traditional three-way decision is firstly generalized, and the multi-class neighbourhood three-way decision model of mixed data is presented. Then, a setting method of self-adaption cost function is given in the model, and based on this method, the algorithm of imbalance data classification of multi-class neighbourhood three-way decision model is proposed. Simulation experiment results show that the proposed classification algorithm has better classification performance for imbalance data.

Keywords Imbalance data, Classification, Three-way decision, Cost function, Self-adaption

不平衡数据指数据集类别、大小、规模的不平衡。对于这类分类问题,传统的机器学习算法的分类结果都倾向于规模较大的类,使得规模较小的类的分类效果很差[1]。然而,对于现实中的不平衡数据集,规模较小的类往往具有更重要的价值,因此不平衡数据的分类问题是目前机器学习和数据挖掘等领域的一个研究热点[2-3]。

为了对不平衡数据取得更好的分类效果,目前学者们提出了很多解决方法。已有解决方案主要分为两个方面:1)从数据集本身出发来解决问题,即通过重采样的技术来平衡数据集,从而提高分类算法的分类性能[4-6];2)从算法角度来解决问题,即在不平衡数据分类方面对传统算法进行改进,使其对不平衡数据达到较好的分类效果。对于后者,具体的改进算法有:基于改进 k 近邻的分类算法[7]、基于改进 SVM 的分类算法[8-9]、基于集成学习的分类算法[10]、基于代价敏感学习的分类算法[11-12]。其中,基于代价敏感的改进算法也是一种较为常用的方法,其主要思想是对不平衡数据中规模较小的类设定较高的误分类代价,从而提高小类的分类准确度。代价敏感学习在不平衡数据分类方面具有较好的分类效果[12]。

三支决策是加拿大著名学者 Yao[13] 提出的一种新型的决策模型,它在决策方法中融入了代价敏感学习,并且将决策的结果分成 3 个部分,即接受决策、延迟决策和拒绝决策。对于有充分置信信息的对象,直接采用接受决策或拒绝决策;对于所给信息不充分和模棱两可的对象,采用延迟决策。这种决策方法更加符合实际应用的决策模式,目前已被成功运用在分类学习、决策提取和不确定性推理等领域[14-16]。

然而,传统的三支决策模型仅适用于离散型的数据,并且只能处理二分类问题[13]。现实中数据类型的复杂性以及类别的多样性,使得传统的三支决策模型面临一定的挑战。虽然多分类问题可以转化为二分类问题[17],但是 Yao 对每个类设定了相同的代价[13],对于不平衡数据,这样的假设是不合

到稿日期:2018-06-19 返修日期:2018-11-18 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:四川省教育厅重点项目(16ZB0360)
This work was supported by the Major Project of Sichuan Education Department (16ZB0360).
通信作者:向伟(xiangwei19766@163.com)

理的。因此，本文提出一种基于多类邻域三支决策模型的不平衡数据分类算法。首先，在数值型邻域三支决策模型[18]的基础上提出混合数据类型的多类邻域三支决策，对数据集中的每个类设定不同的代价函数矩阵；然后，根据多类邻域三支决策模型，提出一种不平衡数据的分类算法，给出一种自适应代价函数矩阵的设定方法，并从理论层面分析了该方法的有效性；最后，通过仿真实验验证了所提算法在不平衡数据分类中的有效性与优越性。

1 三支决策模型

分类学习中的标准数据集可表示为信息系统的形式。设信息系统 $IS=(U,AT)$，U 表示信息系统的对象集，也被称为样本集；AT 表示信息系统的属性集，也称为特征集。

Yao[13]提出的三支决策模型建立在决策理论粗糙集的基础上。在决策理论粗糙集模型中，对象 x 所处的状态可表示为集合 $\Omega=\{X,\sim X\}$，即对象隶属于 X 或 $\sim X$。对象 x 所要进行的行为可表示为集合 $\Gamma=\{a_P,a_B,a_N\}$，即对象 x 被分入 X 的正区域、边界域和负区域 3 种情况。对象 x 在所处状态采取相应的行为时，将产生一定的代价。表 1 是决策理论粗糙集模型的代价函数列表，也称为代价函数矩阵。

表 1 代价函数矩阵

Table 1 Cost function matrix

	X	$\sim X$
a_P	λ_{PP}	λ_{PN}
a_B	λ_{BP}	λ_{BN}
a_N	λ_{NP}	λ_{NN}

其中，λ_{PP}，λ_{BP} 和 λ_{NP} 分别表示对象 x 隶属于 X 时被分入 X 的正区域、边界域和负区域所付出的代价；λ_{PN}，λ_{BN} 和 λ_{NN} 分别表示对象 x 隶属于 $\sim X$ 时被分入 $\sim X$ 的正区域、边界域和负区域所付出的代价。

在实际的决策过程中，由于数据是不确定的，因此任意对象 x 对 X 和 $\sim X$ 都有一定的隶属程度。决策理论粗糙集模型通过条件概率的方式来衡量对象与集合之间的隶属度，即对象 x 关于集合 X 的隶属度为：

$$P(X|[x])=\frac{|X\cap[x]|}{|[x]|}$$

其中，$[x]$ 表示对象 x 的等价类[13-14]。

根据表 1 的代价函数矩阵，可以得到对象 x 采取 3 种行为时的代价：

$$L(a_P|[x])=\lambda_{PP}\cdot P(X|[x])+\lambda_{PN}\cdot P(\sim X|[x])$$

$$L(a_B|[x])=\lambda_{BP}\cdot P(X|[x])+\lambda_{BN}\cdot P(\sim X|[x])$$

$$L(a_N|[x])=\lambda_{NP}\cdot P(X|[x])+\lambda_{NN}\cdot P(\sim X|[x])$$

其中，$L(a_P|[x])$，$L(a_B|[x])$ 和 $L(a_N|[x])$ 表示对象 x 被分入 X 的正区域、边界域和负区域的代价结果。

根据最小代价原则，可以得到如下 3 条决策规则：

(1)如果 $L(a_P|[x])<R(a_B|[x])$，并且 $L(a_P|[x])<R(a_N|[x])$，那么 $x\in POS(X)$；

(2)如果 $R(a_B|[x])<R(a_P|[x])$，并且 $R(a_B|[x])<R(a_N|[x])$，那么 $x\in BUN(X)$；

(3)如果 $R(a_N|[x])<R(a_P|[x])$，并且 $R(a_N|[x])<R(a_B|[x])$，那么 $x\in NEG(X)$。

其中，$POS(X)$，$BUN(X)$ 和 $NEG(X)$ 分别表示集合 X 的正区域、边界域和负区域[13]。

由于 $P(X|[x])+P(\sim X|[x])=1$，并且通常代价函数满足 $0\leqslant\lambda_{PP}\leqslant\lambda_{BP}\leqslant\lambda_{NP}$，$0\leqslant\lambda_{NN}\leqslant\lambda_{BN}\leqslant\lambda_{PN}$，因此：

(1)如果 $P(X|[x])\geqslant\alpha$ 且 $P(X|[x])\geqslant\gamma$，那么 $x\in POS(X)$；

(2)如果 $P(X|[x])<\alpha$ 且 $P(X|[x])<\beta$，那么 $x\in BUN(X)$；

(3)如果 $P(X|[x])<\gamma$ 且 $P(X|[x])\leqslant\beta$，那么 $x\in NEG(X)$。

其中，α，β 和 γ 是决策理论粗糙集模型中一组重要的阈值，它对三支决策的诱导起着关键的作用。

$$\alpha=\frac{\lambda_{PN}-\lambda_{BN}}{(\lambda_{PN}-\lambda_{BN})+(\lambda_{BP}-\lambda_{PP})}$$

$$\beta=\frac{\lambda_{BN}-\lambda_{NN}}{(\lambda_{BN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{BP})}$$

$$\gamma=\frac{\lambda_{PN}-\lambda_{NN}}{(\lambda_{PN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{PP})}$$

若 $0\leqslant\beta<\gamma<\alpha\leqslant1$，则：

(1)如果 $P(X|[x])\geqslant\alpha$，那么 $x\in POS(X)$；

(2)如果 $\beta<P(X|[x])<\alpha$，那么 $x\in BUN(X)$；

(3)如果 $P(X|[x])\leqslant\beta$，那么 $x\in NEG(X)$。

上述 3 条决策规则即 Yao 提出的三支决策模型。若对象 x 关于集合 X 的隶属度大于阈值 α，此时 $x\in POS(X)$，即接受决策；若对象 x 关于集合 X 的隶属度介于阈值 α 与 β 之间，此时 $x\in BUN(X)$，即延迟决策；若对象 x 关于集合 X 的隶属度小于阈值 γ，此时 $x\in NEG(X)$，即拒绝决策。

2 多类邻域三支决策模型的不平衡数据分类

现实中的许多分类问题都表现出了类别的不平衡性。对于一个不平衡数据，传统的分类方法对规模较小的类别不具有较好的分类效果，如何对不平衡数据达到较好的分类效果是目前分类问题研究中的一个热点。

本节将提出一种针对混合型数据的多类邻域三支决策模型；同时针对类别不平衡的问题，设定一种自适应的代价函数矩阵来满足类别的不平衡要求；最后提出基于多类邻域三支决策模型的不平衡数据分类算法。

2.1 混合数据的多类邻域三支决策模型

邻域关系是处理连续型数据的一种常用二元关系。Hu 等[19-20]将邻域关系进一步推广用于处理离散和连续混合类型的数据；同时为了适应多分类情形，Liu 等[17]提出了改进的三支决策模型。因此，本节在前人研究的基础上提出一种混合数据的多类邻域三支决策模型。

定义 1[19-20] 设信息系统 $IS=(U,AT)$，属性子集 $A\subseteq AT$ 是一个离散型和连续型并存的集合，$A=A_C\cup A_N$，其中 A_C 表示离散型属性集，A_N 表示连续型属性集。定义 A 上混合数据类型的邻域关系 MN_A 为：

$$MN_A^{\delta}=\{(x,y)\,|\,d_{A_N}(x,y)\leqslant\delta\wedge a(x)=a(y),\forall a\in A_C\}$$

其中，$x,y\in U$，$d_{A_N}(x,y)$ 表示对象 x 和 y 的距离度量，常用的

距离函数采用闵可夫斯基距离;δ 称为邻域半径,是一个非负常数;$a(x)$和 $a(y)$分别表示对象 x 和 y 在属性 a 下的属性值。

定义 2[19-20]　设信息系统 $IS=(U,AT)$,属性子集 $A\subseteq AT$ 是一个离散型和连续型并存的属性集,A 上确定的混合数据类型的邻域关系为 MN_A^δ。那么,$x\in U$ 在 MN_A^δ 下的邻域类定义为:

$$mn_A^\delta(x)=\{y\in U\mid (x,y)\in MN_A^\delta\}$$

根据定义 1,对象 $x\in U$ 的邻域类是与 x 距离相近的对象集合。由于连续型属性的存在,使得等价关系无法建立,因此对于对象集 X,包含连续属性的信息系统中对象 $x\in U$ 关于 X 的隶属度[18]可表示为:

$$P(X\mid mn^\delta(x))=\frac{|X\cap mn^\delta(x)|}{|mn^\delta(x)|}$$

通过邻域类的方法来代表信息系统中相近的对象,从而表示出混合类型数据下对象与集合之间的隶属度量后,就可以给出混合数据类型的三支决策模型。同时,为了适应多类别的数据,下面将提出混合数据类型的多类邻域三支决策模型。

对于包含 m 个类的信息系统,m 个类可表示为 $Class=\{C_1,C_2,\cdots,C_m\}$,其中 C_i 表示第 i 个类的对象集,因此 $C_i\cap C_j=\varnothing(i\neq j)$,$\bigcup_{i=1}^{m}C_i=U$,同时 $\sim C_i=\bigcup_{j=1,j\neq i}^{m}C_j$,所以 $x\in U$ 将有 m 种状态,即状态集为 $\Omega=\{C_1,C_2,\cdots,C_m\}$,对象 x 可进行的行为仍然为 $\Gamma=\{a_P,a_B,a_N\}$,即表示对象 x 被分入类别 C_i 的正区域、边界域和负区域 3 种情况。表 2 是对象 x 处于类别 C_i 时采取相应行为的代价函数矩阵。

表 2　类别 C_i 的代价函数矩阵

Table 2　Cost function matrix of class C_i

	C_i	$\sim C_i$
a_P	$\lambda_{PP}^{C_i}$	$\lambda_{PN}^{C_i}$
a_B	$\lambda_{BP}^{C_i}$	$\lambda_{BN}^{C_i}$
a_N	$\lambda_{NP}^{C_i}$	$\lambda_{NN}^{C_i}$

其中,$\lambda_{PP}^{C_i}$,$\lambda_{BP}^{C_i}$和 $\lambda_{NP}^{C_i}$分别表示对象 x 隶属于类别 C_i 时被分入 C_i 的正区域、边界域和负区域所付出的代价;$\lambda_{PN}^{C_i}$,$\lambda_{BN}^{C_i}$和 $\lambda_{NN}^{C_i}$分别表示对象 x 隶属于类别 $\sim C_i$ 时被分入 $\sim C_i$ 的正区域、边界域和负区域所付出的代价。

由于信息系统包含 m 个类,因此需要对 $Class=\{C_1,C_2,\cdots,C_m\}$中的每个类设定如表 2 所列的代价函数矩阵,即总共有 m 个代价函数矩阵。

根据表 2 中类别 C_i 的代价函数矩阵,可以得到对象 x 采取 3 种行为时的代价:

$$L_{C_i}(a_P\mid mn(x))=\lambda_{PP}^{C_i}\cdot P(C_i\mid mn(x))+\lambda_{PN}^{C_i}\cdot P(\sim C_i\mid mn(x))$$

$$L_{C_i}(a_B\mid mn(x))=\lambda_{BP}^{C_i}\cdot P(C_i\mid mn(x))+\lambda_{BN}^{C_i}\cdot P(\sim C_i\mid mn(x))$$

$$L_{C_i}(a_N\mid mn(x))=\lambda_{NP}^{C_i}\cdot P(C_i\mid mn(x))+\lambda_{NN}^{C_i}\cdot P(\sim C_i\mid mn(x))$$

其中,$L_{C_i}(a_P\mid[x])$,$L_{C_i}(a_B\mid[x])$和 $L_{C_i}(a_N\mid[x])$表示对象 x 被分入 C_i 的正区域、边界域和负区域的代价结果。

类似于经典三支决策模型,由于 $P(C_i\mid mn(x))+P(\sim C_i\mid mn(x))=1$,并且代价函数满足 $0\leqslant\lambda_{PP}^{C_i}\leqslant\lambda_{BP}^{C_i}\leqslant\lambda_{NP}^{C_i}$,$0\leqslant\lambda_{NN}^{C_i}\leqslant\lambda_{BN}^{C_i}\leqslant\lambda_{PN}^{C_i}$,因此有:

(1)如果 $P(C_i\mid mn(x))\geqslant\alpha$ 且 $P(C_i\mid mn(x))\geqslant\gamma$,那么 $x\in POS(C_i)$;

(2)如果 $P(C_i\mid mn(x))<\alpha$ 且 $P(C_i\mid mn(x))<\beta$,那么 $x\in BUN(C_i)$;

(3)如果 $P(C_i\mid mn(x))<\gamma$ 且 $P(C_i\mid mn(x))\leqslant\beta$,那么 $x\in NEG(C_i)$。

其中,$(\alpha^{C_i},\beta^{C_i},\gamma^{C_i})$是类别 C_i 下对应的一组阈值,它只对类别 C_i 的决策有效。

$$\alpha^{C_i}=\frac{\lambda_{PN}^{C_i}-\lambda_{BN}^{C_i}}{(\lambda_{PN}^{C_i}-\lambda_{BN}^{C_i})+(\lambda_{BP}^{C_i}-\lambda_{PP}^{C_i})}$$

$$\beta^{C_i}=\frac{\lambda_{BN}^{C_i}-\lambda_{NN}^{C_i}}{(\lambda_{BN}^{C_i}-\lambda_{NN}^{C_i})+(\lambda_{NP}^{C_i}-\lambda_{BP}^{C_i})}$$

$$\gamma^{C_i}=\frac{\lambda_{PN}^{C_i}-\lambda_{NN}^{C_i}}{(\lambda_{PN}^{C_i}-\lambda_{NN}^{C_i})+(\lambda_{NP}^{C_i}-\lambda_{PP}^{C_i})}$$

在上述推导分析中,我们得出了类别 C_i 的三支决策。对于信息系统中的其他类,按照相同的方法进行推导,可以得出该类别的一组阈值,因此 m 个类将会得到 m 组阈值,在进行决策的过程中,每组阈值之间是相互独立的。

假设存在如下情形:类别 C_p 对应的一组阈值为 $(\alpha^{C_p},\beta^{C_p},\gamma^{C_p})=(0.50,0.14,0.35)$,类别 C_q 对应的一组阈值为 $(\alpha^{C_q},\beta^{C_q},\gamma^{C_q})=(0.39,0.13,0.36)$,对于对象 x,有 $P(C_p\mid mn(x))=0.5$,$P(C_q\mid mn(x))=0.4$,即满足 $P(C_p\mid mn(x))\geqslant\alpha^{C_p}$,$P(C_q\mid mn(x))\geqslant\alpha^{C_p}$。那么根据所提出的多类邻域三支决策模型,对象 x 既被判定为类 C_p 又被判定为类 C_q,这显然是不合理的。出现这类情形主要是由于,对象的取值比较模糊,介于类与类之间的边缘,从而出现了其被判定为多个类的情形。本文将采用最小代价的方法来解决此类问题。

定义 3　设信息系统 $IS=(U,AT)$包含 m 个类别,表示为 $Class=\{C_1,C_2,\cdots,C_m\}$,类别 C_i 对应的一组阈值为$(\alpha^{C_i},\beta^{C_i},\gamma^{C_i})$。对象 $x\in U$,若 $x\in POS(C_{t1})$,$x\in POS(C_{t2})$,$\cdots$,$x\in POS(C_{tk})(C_{t1},C_{t2},\cdots,C_{tk}\in Class)$,那么,对象 x 在多类邻域三支决策模型下的决策为:

$$x\in POS(C_\pi)$$

其中,$C_\pi=\arg\max\limits_{1\leqslant j\leqslant k}L_{C_{tj}}(C_{tj}\mid mn(x))$。

定义 3 表明,当对象决策为多个类的正区域时,最终决策的类别为所有类别中决策代价最小的,这样就使得所提出的多类邻域三支决策模型具有最小的决策代价。

2.2　不平衡数据分类算法

现实应用中存在着大量的不平衡数据集,由于类别不平衡,一般的分类器倾向于将样本预测为规模较大的类,从而使得规模较小的类别的分类效果很差。对分类器融入代价敏感学习是解决不平衡数据分类的一种重要方法,而三支决策正是一种基于代价敏感学习的决策模型,并且对于不确定性较大的样本给出了一种延迟决策的判定模式,更加适用于现实

环境下的决策应用。因此，本节在 2.1 节提出的多类邻域三支决策模型的基础上给出一种不平衡数据的分类算法。

针对混合数据的分类，Hu 等[20]基于邻域粗糙集模型提出一种邻域的二分类算法，具体如算法 1 所示。

算法 1[20] 邻域分类器

输入：训练集 IS=(U,AT)，测试对象 t，邻域半径 δ，特定的距离函数

输出：测试对象 t 的类

1. 根据距离函数计算测试对象 t 与 x∈U 之间的距离；
2. 计算出测试对象 t 在训练集 IS 中的邻域类 $mn^{\delta}(x)$；
3. 找出邻域类 $mn^{\delta}(x)$中包含对象最多的类别，记为 C_{max}；
4. 将测试对象 t 的类别标记为 C_{max}。

事实上，Hu 等[20]提出的邻域分类器是一种二支决策模型，即最终的分类结果只有两种情况，即属于类别 C_{max}和不属于类别 C_{max}；并且当训练集的类别不平衡时，测试对象 t 的邻域类 $mn^{\delta}(x)$中包含较多的大类，这使得测试对象 t 的类判别倾向于大类。因此，Hu 等提出的邻域分类器对不平衡数据的分类效果不理想。本文在多类邻域三支决策模型下将邻域分类器进行推广，提出多类邻域三支决策的邻域分类算法。

多类邻域三支决策邻域分类算法的分类思想类似于邻域分类器，主要是在邻域分类器的基础上加入了代价敏感学习。首先对训练集中的每个类设定代价函数矩阵，然后得出每个类的一组阈值(α,β,γ)；对于测试对象 t，根据邻域类对每个类别的隶属程度来判别类标记，当出现被多个类别标记的情形时，根据定义 3 的方法选择决策代价最小的类别作为最终的判别类。多类邻域三支决策邻域分类算法的具体过程如算法 2 所示。

算法 2 多类邻域三支决策邻域分类算法(Multiple Three-way Decision Classification，MTWDC)

输入：训练集 IS=(U,AT)，训练集的类别划分 Class=$\{C_1,C_2,\cdots,C_m\}$，类别 C_i 的代价函数矩阵，测试对象 t，邻域半径 δ，特定的距离函数

输出：测试对象 t 的类

1. 根据类别 C_i 的代价函数矩阵计算对应的一组阈值$(\alpha^{C_i},\beta^{C_i},\gamma^{C_i})$；
2. 根据距离函数计算测试对象 t 与 x∈U 之间的距离；
3. 计算出测试对象 t 在训练集 IS 中的邻域类 $mn^{\delta}(x)$；
4. 记候选类集合为 Φ=∅，计算测试对象 t 与每个类 C_i 的隶属度 $P(C_i|mn^{\delta}(t))$，若 $P(C_i|mn^{\delta}(t))\geqslant\alpha^{C_i}$，那么 $\Phi=\Phi\cup\{C_i\}$；
5. 若 Φ 中只包含一个类，那么将测试对象 t 的类别判定为 Φ 中的类，如果 Φ 中包含多个类，那么进入第 6 步；
6. 对 Φ 中的每个类别 C_i 计算决策代价 $L_{C_i}(a_P|mn^{\delta}(t))$，将测试对象 t 的类别判定为决策代价最小的类，即测试对象 t 的最终类别为 $C_t=\arg\min_{C_i\in\Phi} L_{C_i}(a_P|mn^{\delta}(t))$。

将所提出的多类邻域三支决策邻域分类算法用于不平衡数据分类时，对每个类别的代价函数矩阵的设定至关重要。由于不平衡数据中类别之间的规模差异较大，因此针对不同的类别将设定不同的代价函数矩阵，并且设定的代价函数矩阵与对应类别的大小相适应，这样就能够对不平衡数据进行更好的分类。下面通过一个例子来具体体现。

例 1 设训练集 $IS=(U,AT)$，训练集的类别划分 $Class=\{C_1,C_2\}$，其中$|C_1|=12$，$|C_2|=4$，显然训练集类别具有较大的不平衡性。设 C_1 的代价函数矩阵如表 3 所列。

表 3 类别 C_1 的代价函数矩阵

Table 3 Cost function matrix of class C_1

	C_1	$\sim C_1$
a_P	$\lambda_{PP}^{C_1}=0$	$\lambda_{PN}^{C_1}=10$
a_B	$\lambda_{BP}^{C_1}=2$	$\lambda_{BN}^{C_1}=5$
a_N	$\lambda_{NP}^{C_1}=4$	$\lambda_{NN}^{C_1}=0$

由于类别 C_1 的规模较大，类别 C_2($\sim C_1$)的规模较小，一般分类器很容易将原本属于类别 C_2 的对象判为类别 C_1，因此这里对 C_2 设定很高的误分类代价，令 $\lambda_{PN}^{C_1}=10$；原本属于类别 C_1 的对象被判为类别 C_2 具有较小的代价结果，这里令 $\lambda_{NP}^{C_1}=4$。同时，分类正确的代价为 0，即 $\lambda_{PP}^{C_1}=0$，$\lambda_{NN}^{C_1}=0$。将代价 $\lambda_{BP}^{C_1}$取 $\lambda_{PP}^{C_1}$和 $\lambda_{NP}^{C_1}$的中间值，$\lambda_{BN}^{C_1}$取 $\lambda_{PN}^{C_1}$和 $\lambda_{NN}^{C_1}$的中间值。根据多类邻域三支决策模型，可以得到类别 C_1 的阈值为 $\alpha^{C_1}=0.71$。

类似于 C_1 的代价函数矩阵结果，可以得到 C_2 的代价函数矩阵如表 4 所列。

表 4 类别 C_2 的代价函数矩阵

Table 4 Cost function matrix of class C_2

	C_2	$\sim C_2$
a_P	$\lambda_{PP}^{C_2}=0$	$\lambda_{PN}^{C_2}=4$
a_B	$\lambda_{BP}^{C_2}=5$	$\lambda_{BN}^{C_2}=2$
a_N	$\lambda_{NP}^{C_2}=10$	$\lambda_{NN}^{C_2}=0$

根据多类邻域三支决策模型，可以得到类别 C_2 的阈值为 $\alpha^{C_2}=0.29$。

设测试对象 t 属于类别 C_2，对应的邻域类为 $mn(t)$，由于类大小不平衡，设 $mn(t)$包含 C_1 类的 4 个对象，包含 C_2 类的 3 个对象，那么有：

$$P(C_1|mn(t))=0.57<\alpha^{C_1}$$

$$P(C_2|mn(t))=0.43>\alpha^{C_2}$$

根据三支决策模型，测试对象 t 被判定为类别 C_2。由于 $P(C_1|mn(t))>P(C_2|mn(t))$，那么邻域分类器会将测试对象 t 判定为类别 C_1，因此邻域分类器出现了误分类。

通过例 1 的分析可以看出，在多类邻域三支决策模型中设定与类别相适应的代价函数，可以更好地分类不平衡数据，因此多类邻域三支决策模型运用于不平衡数据的分类是合适的。如何准确地设定能够反映出类别误分类的代价将对不平衡数据的分类性能产生重要影响。表 3 和表 4 所列的代价函数矩阵只是根据数据类别情况给出的大致取值。为了便于实际分类工作的开展，本文提出一种自动确定类别代价函数矩阵的方法。

在不平衡数据集中，类别的不平衡主要体现在类大小的不平衡方面，因此可以考虑建立类别的代价函数矩阵与类别大小之间的关系，通过类大小来确定类的代价函数矩阵，这样就可以使得代价函数矩阵自适应地对应不平衡数据。

定义 4 设训练集 $IS=(U,AT)$，训练集的类别划分为 $Class=\{C_1,C_2,\cdots,C_m\}$，定义类别 $C_i(1\leqslant i\leqslant m)$的自适应代价函数矩阵如表 5 所列。

表 5　类别 C_i 的自适应代价函数矩阵

Table 5　Adaptive cost function matrix of class C_i

	C_i	$\sim C_i$
a_P	$\lambda_{PP}^{C_i}=0$	$\lambda_{PN}^{C_i}=\frac{\lvert C_i\rvert}{\lvert U\rvert}$
a_B	$\lambda_{BP}^{C_i}=\frac{\lvert\sim C_i\rvert}{2\cdot\lvert U\rvert}$	$\lambda_{BN}^{C_i}=\frac{\lvert C_i\rvert}{2\cdot\lvert U\rvert}$
a_N	$\lambda_{NP}^{C_i}=\frac{\lvert\sim C_i\rvert}{\lvert U\rvert}$	$\lambda_{NN}^{C_i}=0$

观察表 5 可以看出，类别 C_i 的误分类代价 $\lambda_{NP}^{C_i}$ 根据 $\sim C_i$ 的对象集占整个训练集 U 的比值来定义，因此类别 $\sim C_i$ 的误分类代价 $\lambda_{PN}^{C_i}$ 通过 C_i 的对象集占整个训练集 U 的比值来定义。当类别 C_i 较小时，$\lambda_{NP}^{C_i}=\frac{\lvert\sim C_i\rvert}{\lvert U\rvert}$ 的值较大，当类别 C_i 较大时，$\lambda_{NP}^{C_i}=\frac{\lvert\sim C_i\rvert}{\lvert U\rvert}$ 的值则较小，这刚好与例 1 的分析结果相吻合，因此这种代价函数的确定方式是合理的，可以随着数据集来自动确定。

3　仿真实验

本节将在现实数据集上对本文提出的多类邻域三支决策分类算法与目前常用的分类算法进行分类比较，然后通过分类结果来验证所提算法的有效性。

3.1　实验数据集

从 UCI 机器学习数据集中选择 6 个标准数据集，具体如表 6 所列。其中每个数据集都存在一定的类别不平衡情形，数据 car 为离散型的数据集，其余都为数值型或混合类型的数据集。

表 6　实验数据集

Table 6　Experimental data set

数据集	对象	属性	类大小分布
car	1728	6	{1210,384,69,65}
ecoli	336	7	{143,77,52,35,20,5,2,2}
glass	214	9	{70,76,29,17,13,9}
yeast	1484	8	{463,424,244,163,51,44,35,30,20,5,2,2,1}
wall	5456	4	{2205,2097,826,328}
fog	151987	9	{59185,87655,5147}

3.2　评价指标

在分类学习中，一般通过数据集的分类精度来评估分类算法的性能；但是对于不平衡数据，分类精度这一指标不具有较好的评价效果，这主要是由于分类精度反映的是数据集整体的分类准确率，而不平衡数据中的小类发生错误分类并不会对整体分类精度产生较大的影响。针对不平衡数据的分类问题，一些新的评价指标被提出，如 G-mean[21]，F-measure[22] 以及 AUC[23] 等。

分类算法对不平衡数据的具体分类结果可以通过一个混淆矩阵来表示，具体如表 7 所列。

表 7　不平衡数据的分类混淆矩阵

Table 7　Classification obfuscation matrix of imbalance data

	判定为小类	判定为大类
实际为小类	TP	FN
实际为大类	FP	TN

根据混淆矩阵，分类算法的 G-mean 评价指标表示为：

$$G\text{-}mean=\sqrt{SE\cdot SP}$$

其中，$SE=\frac{TP}{TP+FN}$，$SP=\frac{TN}{TN+FP}$。G-mean 是评价分类算法对小类和大类两方面分类性能的综合评估指标。

F-measure 是一个更加注重小类分类效果的评价指标，因此通过它对不平衡数据进行分类评估具有很重要的价值。评价指标 F-measure 表示为：

$$F\text{-}measure=\frac{2\cdot precision\cdot recall}{precision+recall}$$

其中，$precision=\frac{TP}{TP+FP}$，$recall=\frac{TP}{TP+FN}$。

AUC 也是一种重要的分类评估指标，详细的解释与描述可见文献[23]。

3.3　参数设置

本文提出的不平衡数据分类算法包含两大类参数：1）训练集中类别的代价函数矩阵，本实验中按照定义 4 所提出的方法进行设定；2）邻域半径 δ，δ 的大小直接影响混合数据类型中对象与集合之间隶属度的度量，因此选择合适的 δ 对分类算法的分类性能至关重要。由于 δ 的大小与隶属度之间的关系是不确定的，而 δ 的大小又受到训练集中数值型属性数目的影响，因此一般情况下很难直接对邻域半径 δ 进行设定。

Hu 等[20] 在邻域分类器模型中给出了一种间接的邻域半径选取方法。对于训练集 $IS=(U,AT)$，测试对象为 t，设 t 与训练集对象 $\forall x\in U$ 的距离的最大值为 $\max_t$，最小值为 $\min_t$，那么可以通过下式来选择邻域半径 δ：

$$\delta=\min_t+\omega\cdot(\max_t-\min_t)$$

其中，$\omega\in[0,1]$。这样就可以通过 0 与 1 之间的数值来确定 δ 的取值。为了获取最优实验参数，我们选择不同的 ω 分别进行实验，然后通过实验结果来确定最终的参数结果。设定 ω 在 $[0.05,0.3]$ 上以 0.05 为间隔进行实验，各数据集在不同 ω 下分类结果的 *G-mean*，*F-measure* 以及 *AUC* 的平均值如表 8 所列。

表 8　不同 ω 下的实验结果

Table 8　Experimental results under different ω

ω 值	car	ecoli	glass	yeast	wall	fog
0.05	0.8934±0.0085	0.8528±0.0097	0.7841±0.0073	0.7456±0.0055	0.7463±0.0086	0.8217±0.0059
0.10	**0.9433±0.0062**	**0.9167±0.0152**	0.8348±0.0087	0.7729±0.0034	**0.7812±0.0060**	**0.8642±0.0068**
0.15	0.9257±0.0103	0.9058±0.0134	**0.8436±0.0121**	**0.7972±0.0074**	0.7747±0.0076	0.8547±0.0084
0.20	0.9174±0.0083	0.8945±0.0085	0.8251±0.0094	0.7682±0.0057	0.7573±0.0062	0.8512±0.0067
0.25	0.8853±0.0115	0.8646±0.0125	0.8084±0.0075	0.7563±0.0145	0.7395±0.0079	0.8324±0.0081
0.30	0.8528±0.0163	0.8214±0.1020	0.7668±0.0116	0.7265±0.0138	0.7137±0.0097	0.8158±0.0117

通过表 8 的实验结果可以看出，当 ω 取值为 0.1 时，大部分数据集都取得了较高的评价指标值；当 ω 取值为 0.15 时，部分数据集也有较优的结果。综合考虑，将 ω 设定为 0.1 进行实验。

3.4 实验结果及分析

为了验证所提 MTWDC 算法在不平衡数据上的分类性能，选取 3 种不同方法的不平衡数据分类算法进行对比：基于代价敏感学习的朴素贝叶斯分类算法（算法 1）[24]、基于集成学习方法的 AdaBoost 算法（算法 2）、基于类别加权的 k 近邻分类算法（$k=5$）（算法 3）[25]。实验中采用 5 折交叉验证的方法，对于部分数据集的类别对象数目不足 5 的情况，可根据实际情况的数目进行实验。根据分类结果计算出每种分类算法的 G-mean，*F-measure* 以及 *AUC* 值，并且最终的结果用“均值±标准差”来表示。

表 9 列出了 4 种分类算法在各个数据集上分类结果的 *G-mean* 值。表 10 列出了 4 种分类算法在各个数据集上分类结果的 *F-measure* 值。表 11 列出了 4 种分类算法在各个数据集上分类结果的 *AUC* 值。表中加粗的结果表示每行的最大值。观察这 3 个表可以发现，在 G-mean 评价指标中，除 glass 数据集外，本文所提出的 MTWDC 算法在其余数据集中具有最好的结果；在 F-measure 评价指标中，除 car 数据集外，MTWDC 算法在其余数据集中具有最好的结果；在 AUC 评价指标中，除 ecoli 和 wall 数据集，MTWDC 算法在其余数据集中同样具有最好的结果。整体比较可以得出，MTWDC 算法具有较好的评价结果，算法 1 次之，而算法 2 具有最低的评价结果，说明算法 2 对不平衡数据较为敏感。

表 9 4 种分类算法的 *G-mean* 结果

Table 9 *G-mean* of 4 classification algorithms

数据集	MTWDC	算法 1	算法 2	算法 3
car	**0.9427±0.0036**	0.9376±0.0068	0.9087±0.0124	0.9136±0.0058
ecoli	**0.8924±0.0027**	0.8637±0.0067	0.8219±0.0087	0.8454±0.0076
glass	0.8255±0.0062	**0.8355±0.0055**	0.7745±0.0072	0.7973±0.0125
yeast	**0.7429±0.0014**	0.7348±0.0085	0.6874±0.0107	0.6928±0.0059
wall	**0.7583±0.0077**	0.7254±0.0047	0.6953±0.0074	0.7164±0.0038
fog	**0.8363±0.0053**	0.8163±0.0065	0.7886±0.0086	0.7782±0.0029

表 10 4 种分类算法的 *F-measure* 结果

Table 10 *F-measure* of 4 classification algorithms

数据集	MTWDC	算法 1	算法 2	算法 3
car	0.9276±0.007	**0.9323±0.0096**	0.8589±0.0126	0.8775±0.0072
ecoli	**0.9367±0.0018**	0.9163±0.0057	0.8835±0.0055	0.9023±0.0087
glass	**0.8068±0.0051**	0.7937±0.0088	0.7845±0.0074	0.8043±0.0039
yeast	**0.7746±0.0016**	0.7446±0.0047	0.7544±0.0121	0.7364±0.0059
wall	**0.7952±0.0048**	0.7776±0.0069	0.7255±0.0127	0.7448±0.0075
fog	**0.8735±0.0027**	0.8557±0.0116	0.8272±0.0075	0.8326±0.0082

表 11 4 种分类算法的 *AUC* 结果

Table 11 *AUC* of 4 classification algorithms

数据集	MTWDC	算法 1	算法 2	算法 3
car	**0.9714±0.0081**	0.9553±0.0140	0.9167±0.0163	0.9327±0.0078
ecoli	0.9318±0.0107	**0.9428±0.0130**	0.9082±0.0152	0.8858±0.0156
glass	**0.8725±0.0122**	0.8356±0.0095	0.7930±0.0132	0.8173±0.0133
yeast	**0.8016±0.0074**	0.7679±0.0125	0.7342±0.0147	0.7549±0.0119
wall	0.8026±0.0057	0.7726±0.0106	0.7664±0.0153	**0.8139±0.0078**
fog	**0.9017±0.0113**	0.8716±0.0133	0.8442±0.0170	0.8520±0.0149

本文所提出的 MTWDC 算法是建立在三支决策模型基础上的一种分类算法，通过最小化分类代价来学习阈值，从而对数据进行分类，因此算法的误分类代价也是评估算法分类性能的一种重要指标。本实验根据这 4 种算法的分类结果计算出对应的误分类代价，具体如表 12 所列。其中，MTWDC 算法的误分类代价 *cost* 采用下式进行计算：

$$cost=\sum_{i=1}^{m} cost_{C_i}$$

$$cost_{C_i}=n_{\mathrm{BP}}^{C_i}\cdot\lambda_{\mathrm{BP}}^{C_i}+n_{\mathrm{NP}}^{C_i}\cdot\lambda_{\mathrm{NP}}^{C_i}$$

其中，$Class=\{C_1,C_2,\cdots,C_m\}$，$cost_{C_i}$ 表示类别 C_i 的误分类代价，$n_{\mathrm{BP}}^{C_i}$ 表示 C_i 类中标记入边界域中的对象，$n_{\mathrm{NP}}^{C_i}$ 表示 C_i 类中标记入负区域中的对象。观察表 12 可以看出，MTWDC 算法在所有数据集中具有最低的误分类代价，算法 1 次之，算法 2 具有最高的误分类代价。这主要是由两个方面的因素造成的：1）MTWDC 算法对不平衡数据具有分类的优越性，通过对不同大小的类别设定不同的代价达到更高的分类准确度；2）MTWDC 算法是一种三支决策的分类方法，即决策结果还存在一种处于延迟决策的形式，这类对象是否被接受，需要进一步考虑分析，而其余 3 种算法是直接进行分类处理，因而误分类的代价会更高。

表 12 4 种分类算法的误分类代价

Table 12 Misclassification cost of 4 classification algorithms

数据集	MTWDC	算法 1	算法 2	算法 3
car	64.4676±2.6253	68.7563±2.8765	84.7653±4.2243	76.8742±5.7821
ecoli	12.8274±5.8737	14.1754±7.5324	17.5626±9.8412	22.2364±6.4318
glass	24.1464±4.8347	27.8352±8.4613	36.8256±8.3215	32.2465±6.3107
yeast	82.0535±1.6378	94.2954±3.5140	100.5274±2.6454	95.6835±5.2445
wall	284.3251±3.6548	343.2491±5.2195	358.7547±8.8162	310.3572±6.3321
fog	2956.9271±3.8542	3355.5368±6.1526	3716.8351±4.5146	3574.7396±7.2453

综合比较 4 种算法的实验结果可以发现，本文提出的 MTWDC 算法对于不平衡数据的分类具有更好的分类性能。

结束语 不平衡数据中类别的规模差异较大，使得传统的分类算法对规模较小的类不具有较好的分类效果。三支决策模型是一种新型的决策模型，它通过在决策方法中融入代价敏感学习，并且将决策的结果分成接受决策、延迟决策和拒

绝决策3种形式，得到了更优越的决策结果。本文将传统的三支决策在混合数据和多个类的情形下进行推广，提出了一种混合数据的多类三支决策模型，并对数据中的每个类设定一种与数据集类规模相适应的代价函数，使其对不平衡数据具有更好的分类决策效果，进而提出了基于多类三支决策模型的不平衡数据分类算法。实验分析中，通过对不平衡数据进行分类，证明了所提算法比其他算法具有更好的分类性能。

参 考 文 献

[1] ZHANG S, SADAOUI S, MOUHOUB M. An empirical analysis of imbalanced data classification[J]. Computer & Information Science, 2015, 8(1): 151-162.

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

[3] HE H L, ZHANG W Y, ZHANG S. A novel ensemble method for credit scoring: Adaption of different imbalance ratios[J]. Expert Systems with Applications, 2018, 98(15): 105-117.

[4] RIVERA W A. Noise reduction a priori synthetic over-sampling for class imbalanced data sets[J]. Information Sciences, 2017, 408: 146-161.

[5] DOUZAS G, BACAO F, LAST F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE[J]. Information Sciences, 2018, 465: 1-20.

[6] CORDÓN I, GARCÍA S, FERNÁNDEZ A, et al. Imbalance: Oversampling algorithms for imbalanced classification in R[J]. Knowledge-Based Systems, 2018, 161: 329-341.

[7] ZHU Y J, WANG Z, GAO D Q. Gravitational fixed radius nearest neighbor for imbalanced problem[J]. Knowledge-Based Systems, 2015, 90: 224-238.

[8] WU G, CHANG E. KBA: Kernel boundary alignment considering imbalanced data distribution[J]. IEEE Transactions on Knowledge & Data Engineering, 2005, 17(6): 786-795.

[9] GUPTA D, RICHHARIYA B, BORAH P. A fuzzy twin support vector machine based on information entropy for class imbalance learning[J]. Neural Computing & Applications, 2018(3): 1-12.

[10] DÍEZ-PASTOR J F, RODRÍGUEZ J J, GARCÍA-OSORIO C, et al. Random Balance: Ensembles of variable priors classifiers for imbalanced data[J]. Knowledge-Based Systems, 2015, 85(2/3): 96-111.

[11] KHAN S H, HAYAT M, BENNAMOUN M, et al. Cost-sensitive learning of deep feature representations from imbalanced data[J]. IEEE Transactions on Neural Networks & Learning Systems, 2018, 29(8): 3573-3587.

[12] CAO C J, WANG Z. IMCStacking: Cost-sensitive stacking learning with feature inverse mapping for imbalanced problems[J]. Knowledge-Based Systems, 2018, 150(15): 27-37.

[13] YAO Y Y. Three-way decisions with probabilistic rough sets[J]. Information Sciences, 2010, 180(3): 341-353.

[14] ZHOU B. Multi-class decision-theoretic rough sets[J]. International Journal of Approximate Reasoning, 2014, 55(1): 211-224.

[15] LIANG D C, LIU D, KOBINA A. Three-way group decisions with decision-theoretic rough sets[J]. Information Sciences, 2016, 345: 46-64.

[16] CHEN Y F, YUE X D, FUJITA H, et al. Three-way decision support for diagnosis on focal liver lesions[J]. Knowledge-Based Systems, 2017, 127: 85-99.

[17] LIU D, LI T R, LI H X. A multiple-category classification approach with decision-theoretic rough sets[J]. Fundamenta Informaticae, 2012, 115(2/3): 173-188.

[18] LI W W, HUANG Z Q, JIA X Y, et al. Neighborhood based decision-theoretic rough set models[J]. International Journal of Approximate Reasoning, 2016, 69: 1-17.

[19] HU Q H, YU D R, LIU J F, et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Sciences, 2008, 178(18): 3577-3594.

[20] HU Q H, YU D R, XIE Z X. Neighborhood classifiers[J]. Expert Systems with Applications, 2008, 34(2): 866-876.

[21] KUBAT M, HOLTE R, MATWIN S. Learning when negative examples abound[C]// European Conference on Machine Learning. Springer Berlin Heidelberg, 1997: 146-153.

[22] DAVIS J, GOADRICH M. The relationship between Precision-Recall and ROC curves[C]// Proceedings of the, International Conference on Machine Learning (ICML 2006). New York, USA: ACM Press, 2006: 233-240.

[23] FAWCETT T. An introduction to ROC analysis[J]. Pattern Recognition Letters, 2006, 27(8): 861-874.

[24] JIANG S Y, XIE Z Q, YU W. Classification of naive Bayes imbalanced data based on cost sensitive[J]. Journal of Computer Research and Development, 2011, 48(S1): 387-390.

[25] PATEL H, THAKUR G S. A hybrid weighted nearest neighbor approach to mine imbalanced data[C]// International Conference on Data Mining. Las Vegas: IEEE, 2016: 106-112.

XIANG Wei, born in 1976, associate professor. His main research interests include computer-based information processing & intelligent algorithm.