文章编号:1000－2367(2019)03－0028－07 DOI:10.16366/j.cnki.1000－2367.2019.03.005

基于三支决策的二阶段分类模型研究

徐久成,徐战威,李梦凡,王楠

(河南师范大学 计算机与信息工程学院;河南省高校计算智能与数据挖掘工程技术研究中心,河南 新乡 453007)

摘 要:当对三支决策边界域进一步划分时,边界域知识存在划分信息不足,从而导致分类精度不高,针对上述问题提出一种新的基于三支决策的二阶段分类模型(TWD-TP).第一阶段根据贝叶斯规则构建三支决策中样本的条件概率,通过求解最优化损失函数得到所需阈值,然后按照三支决策规则对数据集进行划分.三支决策是基于最小风险贝叶斯决策理论的划分,在其正域、负域中包含一定的误分类样本;在第二阶段通过类标签索引分别将正域、负域中误分样本作为增量信息引入延迟决策域,形成重构边界域,最后对重构边界域进行划分.实验结果表明:所提出的 TWD-TP 模型不仅能在三支决策划分中筛选出高误分类特征的样本,同时其重构边界域中不能被划分的样本得到正确划分,分类精度进一步提高.

关键词:三支决策;二阶段;增量信息;边界域

中图分类号:TP181 **文献标志码**:A

分类是当前数据挖掘过程中的一个基础性问题,其从训练数据或历史经验中构造出相应的分类模型,并根据该模型将未知类别样本划分为正类或负类(以二分类问题为例),达到样本准确分类的目的.这同决策问题中依据历史经验和当前已知的决策信息做出接受或拒绝的决策类似.传统分类模型,实质是通过对样本属性或特征知识的学习,将其分成正类或负类,或是对其采取接受或拒绝的二支决策.实际情况中,决策者通常因为掌握的决策信息不足,导致特征相似度较高的样本难以被正确划分.比如,在依据概率判定样本属于正类或负类的问题中,某一样本 M 在决策条件下判定其为正类的概率为 0.51,按照分类概率值大小原则,样本 M 被划分到正类中,但同时意味着该样本可能有高达 49% 的误分类率,这在实际分类任务中,存在较高的误分类代价.针对该问题,三支决策理论[1]依据最小风险代价原则,通过引入损失函数,将决策信息不足而难以划分的样本暂时划分到边界域中,待决策信息充足时再进行决策,从而降低误分类率.

决策粗糙集理论[2]以其更接近人类认知与决策模式的优势,已广泛应用于多个学科和领域,包括医疗诊断、统计学中的假设性检验、管理学和论文评审等.近年来将三支决策应用到数据挖掘过程中的研究逐渐流行[3－5].Yao 等人[6－7]通过比较三支决策与标准粗糙集概率模型,分析了三支决策在概率粗糙集模型理论前景上的优势.Huang 等人[8]通过估计用户对项目的偏好程度将三支决策边界域作为可能推荐给用户的项目引入用户推荐系统,进而提高了推荐质量.仇等人[9]将三支决策理论应用到医院分级诊疗决策中,给出了分级诊疗的定量决策方法.三支决策中,边界域是待决策区域,不直接对其做出决策,等待信息充分时再进一步处理.目前,如何对边界域知识进行准确划分已成为三支决策模型需要解决的一个重要问题[10].在三支决策边界域处理上,典型的数据挖掘和机器学习算法偏重于追求较高的分类精度,大部分研究仍采用传统分类算法对其作进一步划分,而边界域划分面临决策信息不足的问题依然凸显,分类结果仍然具有一定的代价风险.针对边界域对象的处理是三支决策在分类任务中的一个研究热点,Zhang 等人[11]在三支决策划分正、负域的基础上构建了两个边界向量,提出了一种基于粗糙集方法和质心解来处理不确定边界的三支决策模型;

收稿日期:2018-05-22;修回日期:2018-12-17.

基金项目:国家自然科学基金(61370169;61402153;60873104);中国博士后科学基金项目(2016M602247);河南省科技攻关重点项目(162102210261).

作者简介:徐久成(1964－),男,河南洛阳人,河南师范大学教授,博士生导师,研究方向为粒计算、数据挖掘和生物信息.

通信作者:徐战威,E-mail:xzw.htu@foxmail.com

徐等人[12－13]基于三支决策改进支持向量机学习方法，实现边界域的决策划分；Li 等人[14－15]将三支决策划分后的负域和边界域作为未知对象，采用集成学习方法对其进行处理，并将其应用到软件冲突预测中，一定程度上弥补了边界域决策信息不足的问题.但在实际的分类任务中，通常会存在一些特征相似度较高的样本，该部分样本极易被划分到错误的类别当中.

综上，本文针对三支决策边界域对象进一步划分时，边界域信息不足导致难以划分，通过引入三支决策中划分错误的样本，对原始边界域进行适度扩充.首先对三支决策规则中条件概率与阈值大小进行比较，将数据集划分为正类域、负类域和延迟决策域，形成三支决策划分.然后利用类别标签索引将正类域、负类域中划分错误的样本作为增量信息添加到边界域中构建新的边界域，并对其进行训练，最后在标准 UCI 数据集中，采用常用分类算法验证本文所提模型的有效性.实验结果表明，TWD-TP 模型能在三支决策划分中有效提高边界域样本的分类精度.

1　基本概念

三支决策理论是在决策粗糙集的基础上对传统二支决策语义的拓展，更合理地解释了决策粗糙集语义，即认为边界域也是可以做出决策的.文献[16－17]对三支决策理论进行了详细的介绍，本节对三支决策理论简要概述.

设 $\Omega=\{X,\neg X\}$ 为正、负两状态集合，X 和 $\neg X$ 为具有互补关系的两种状态.设 $A=\{a_P,a_N,a_B\}$ 为决策状态集，a_P，a_N 和 a_B 分别表示将对象划分到 $POS(X)X$，$NEG(X)$ 和 $BND(X)3$ 种决策动作.当对象 x 真实属于状态 X 时，分别做出，a_P，a_N 和 a_B3 种决策所对应的损失函数值即为 λ_{PP}，λ_{NP} 和 λ_{BP}；当 x 真实属于状态 $\neg X$ 时，分别做出 a_P，a_N 和 a_B3 种决策所对应的代价函数值为 λ_{PN}，λ_{NN} 和 λ_{BN}，如表 1 所示：

表 1　代价矩阵

Tab.1　Cost matrix

Action	Cost Function	
	X	$\neg X$
a_P	λ_{PP}	λ_{PN}
a_B	λ_{BP}	λ_{BN}
a_N	λ_{NP}	λ_{NN}

因此，做出 a_P，a_N 和 a_B 3 种决策的期望损失可分别表示为：

$$\begin{cases}R(a_P\mid[x]_R)=\lambda_{PP}P(X\mid[X]_R)+\lambda_{PN}P(\neg X\mid[x]_R),\\ R(a_N\mid[x]_R)=\lambda_{NP}P(X\mid[X]_R)+\lambda_{NN}P(\neg X\mid[x]_R),\\ R(a_B\mid[x]_R)=\lambda_{BP}P(X\mid[X]_R)+\lambda_{BN}P(\neg X\mid[x]_R),\end{cases}\tag{1}$$

其中，$[x]$ 为样本在属性集下的等价类，$P(X\mid[x])$ 和 $P(\neg X\mid[x])$ 分别表示将等价类$[x]$ 划分为 X 和 $\neg X$ 的概率.按照贝叶斯决策准则，决策时通常会选择期望损失最小的决策行为作为最佳行动方案.因此，可得到如下形式的三支决策模型的决策规则：

(1)若 $R(a_P\mid[x]_R)\leq R(a_N\mid[x]R)$ 和 $R(a_P\mid[x]R)\leq R(a_B\mid[x]R)$ 同时成立，那么 $x\in POS(X)$，

(2) 若 $R(a_N\mid[x]R)\leq R(a_P\mid[x]R)$ 和 $R(a_N\mid[x]R)\leq R(a_B\mid[x]R)$ 同时成立，那么 $x\in NEG(X)$，

(3) 若 $R(a_B\mid[x]R)\leq R(a_P\mid[x]R)$ 和 $R(a_B\mid[x]R)\leq R(a_N\mid[x]R)$ 同时成立，那么 $x\in BND(X)$.

由 X 和 $\neg X$ 关系互补可知 $P(X\mid[x]_R)+P(\neg X\mid[x]_R)=1$，所以上面规则只与概率 $P(X\mid[x]_R)$ 和损失函数 λ 有关.假设：$0\leq\lambda_{PP}\leq\lambda_{BP}\leq\lambda_{NP}$，$0\leq\lambda_{NN}\leq\lambda_{BN}\leq\lambda_{PN}$，根据以上 3 条决策规则，$\alpha$、$\beta$ 和 γ 可分别表示为：

$$\alpha=\frac{\lambda_{PN}-\lambda_{BN}}{(\lambda_{PN}-\lambda_{BN})+(\lambda_{BP}-\lambda_{PP})},\gamma=\frac{\lambda_{PN}-\lambda_{NN}}{(\lambda_{PN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{PP})},\beta=\frac{\lambda_{BN}-\lambda_{NN}}{(\lambda_{BN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{BP})},\tag{2}$$

通过阈值(α,β)的引入，可得到三支决策规则，分别为：

接受决策规则(P)：若 $P(X\mid[x]_R)\geq\alpha$，则 $x\in POS(X)$，

拒绝决策规则(N):若 $P(X \mid [x]_R) \leq \beta$,则 $x \in NEG(X)$,

延迟决策规则(B):若 $\beta < P(X \mid [x]_R) < \alpha$,则 $x \in BND(X)$.

上述三支决策规则描述了基于决策粗糙集的三支决策语义,给出了一种贝叶斯最小风险下的三支决策语义解释[18].

2 三支决策二阶段分类模型

2.1 TWD-TP 模型条件概率的计算

为解决三支决策模型中条件概率计算问题,根据最小风险贝叶斯决策,采用朴素贝叶斯策略来度量三支决策中的条件概率,该方法可避开贝叶斯公式中类条件概率难以从有限的训练样本直接估计得到的障碍.因等价类$[x]_R$ 可表示成具有相同属性值的元素集合,则基于条件属性划分下等价类$[x]_R$ 的概率可表示为:

$$P([x]_R) = P(\nu_1, \nu_2, \cdots, \nu_d) = P(\nu_1)(\nu_2)\cdots(\nu_d) = \prod_{i=1}^{d} P(\nu_i), \tag{3}$$

其中 d 为条件属性的个数,ν_i 为$[x]_R$ 在第 i 个条件属性下的取值,同理可得:

$$P([x]_R \mid X) = P(\nu_1, \nu_2, \cdots, v_d \mid X) = P(\nu_1 \mid X)P(\nu_2 \mid X)\cdots P(\nu_d \mid X) = \prod_{i=1}^{d} P(\nu_i \mid X), \tag{4}$$

因此,样本 x 在 X 类别上的三支决策条件概率为:

$$P(X \mid [x]_R) = P(X)\frac{P([x]_R \mid X)}{P([x]_R)} = \prod_{i=1}^{d} \frac{P(\nu_i \mid X)}{P(\nu_i)} P(X). \tag{5}$$

2.2 TWD-TP 模型阈值的计算

对于给定的决策信息表,其决策代价表示为:

$$COST = COST_{POS} + COST_{BND} + COST_{NEG}. \tag{6}$$

(6)式可进一步表示为:

$$COST = \sum_{p_j \geq \alpha} (1 - p_j) \cdot \lambda_{PN} + \sum_{\beta < p_k < \alpha} (p_k \cdot \lambda_{BP} + (1 - p_k) \cdot \lambda_{BN}) + \sum_{p_i \leq \beta} p_t \cdot \lambda_{NP}, \tag{7}$$

其中,$p_j = P(X \mid [x_j]_R)$,$p_k = P(X \mid [x_k]_R)$,$p_t = P(X \mid [x_t]_R)$ 表示每个对象 x 属于类 X 的概率,根据(7)式,以最小化损失为目标函数的最优化问题可表示为:

$$\alpha, \beta, \gamma = \min_{\alpha, \beta, \gamma} \sum_{p_j \geq \alpha} (1 - p_j) \cdot \lambda_{PN} + \sum_{\beta < p_k < \alpha} (p_k \cdot \lambda_{BP} + (1 - p_k) \cdot \lambda_{BN}) + \sum_{p_i \leq \beta} p_t \cdot \lambda_{NP}, \tag{8}$$

其中,$0 < \beta < \gamma < \alpha < 1$,(1)式中,阈值 α,β 和 γ 可以被 6 个损失函数表示,通常在做出正确决策时不会带来任何损失,即 $\lambda_{PP} = \lambda_{NN} = 0$,则剩余 4 个损失函数可表示为:

$$\lambda_{PN} = \lambda_{PN}, \lambda_{NP} = \frac{1 - \gamma}{\gamma} \cdot \lambda_{PN}, \lambda_{BN} = \frac{\beta \cdot (\alpha - \gamma)}{\gamma \cdot (\alpha - \beta)} \cdot \lambda_{PN}, \lambda_{BP} = \frac{(1 - \alpha) \cdot (\gamma - \beta)}{\gamma \cdot (\alpha - \beta)} \cdot \lambda_{PN}. \tag{9}$$

根据(9)式,假设 λ_{PN} 值为 1,(8)式的最优化问题可表示为:

$$\alpha, \beta, \gamma = \min_{\alpha, \beta, \gamma} \sum_{p_j \geq \alpha} (1 - p_j) + \sum_{p_i \leq \beta} p_t \cdot \frac{1 - \gamma}{\gamma} + \varepsilon \cdot \sum_{\beta < p_k < \alpha} \left(\frac{\beta \cdot (\alpha - \gamma)}{\gamma \cdot (\alpha - \beta)} \cdot (1 - p_k) + \frac{(1 - \alpha) \cdot (\gamma - \beta)}{\gamma \cdot (\alpha - \beta)} \cdot p_k \right), \tag{10}$$

其中,$0 < \beta < \gamma < \alpha < 1$,$\varepsilon \geq 1$,$\varepsilon$ 为惩罚因子,在 TWD-TP 模型第一阶段划分中当误分类较多时,用以避免把过多的对象划分到边界域,因此 3 个阈值可以通过(10)式的最优化问题求解得到.

2.3 三支决策二阶段分类算法的构建

传统二支决策方法追求单一的分类精度,如图 1 所示,当阈值 $\alpha = 0.4$,条件概率 $P \geq \alpha$ 时,样本被划分到正域 $POS(X)$;当 $p < \alpha$ 时,样本被划分到负域 $NEG(X)$;当样本之间概率非常接近,采用这种分类机制可能带来较大误分类代价.相比二支决策,三支决策优势通过引入延迟决策区域 $BND(X)$,弥补二支决策中缺乏误分类容忍机制问题.如图 2 所示,当条件概率 $P \geq \alpha$ 时,样本被划分到正域 $POS(X)$;$P \leq \beta$ 时,样本被划分到负域 $NEG(X)$;当 $\beta < P < \alpha$ 时,样本被划分到延迟区域 $BND(X)$,不直接对其进行划分,待信息充

分时作进一步处理[19]．基于此，本文提出一种新的基于三支决策的二阶段分类模型，如图 3 所示，该方法在三支决策延迟区域 $BND(X)$基础上通过引入误分类样本来扩展原始边界域，形成重构边界域．第一阶段被错误划分的样本与第二阶段边界域样本一起进行训练，训练后的分类器能对重构边界域进一步划分．根据三支决策二阶段分类思想，构建基于三支决策的二阶段分类算法模型 TWD-TP，如图 4 所示，重构后的边界域为：

$$\Delta BND(X)=BND(X)+(\Delta POS(X)+\Delta NEG(X)), \tag{11}$$

其中 $\Delta POS(X)$ 和 $\Delta NEG(X)$ 分别为第一阶段划分到正域和负域的错分样本，$BND(X)$ 为延迟区域．

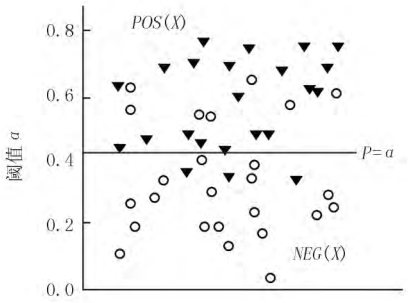

图 1　二支决策方法

Fig. 1　Two way decisions

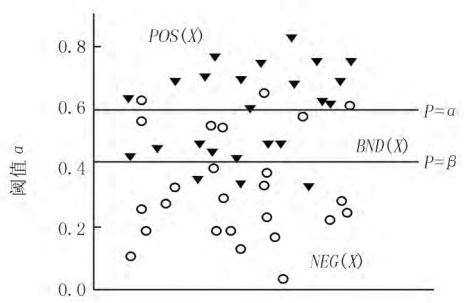

图 2　三支决策方法

Fig. 2　Three-way decisions

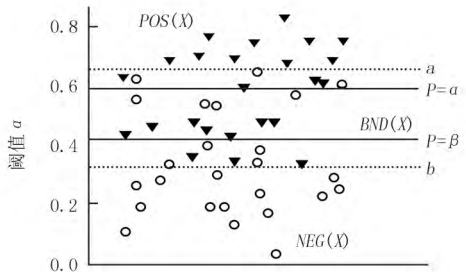

图 3　三支决策二阶段方法

Fig. 3　Two stage based on three-way decisions

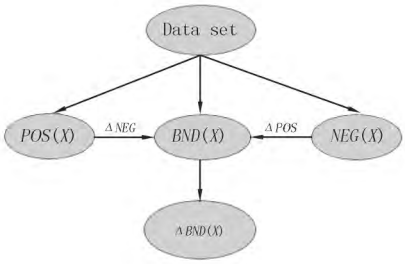

图 4　TWD-TP 模型

Fig. 4　TWD-TP Model

注：图 4 中 Data set 为原始数据集，$POS(X)$，$BND(X)$ 和 $NEG(X)$ 分别代表经过三支决策划分后的正类域、边界域和负类域；ΔNEG，ΔPOS 分别代表正类域、负类域中错分的样本集；$\Delta BND(X)$ 代表重构后的边界域．

算法　基于三支决策的二阶段分类算法

输入：数据集 $D=\{x_1,x_2,\cdots,x_m\}$，共包含 m 个样本，样本类别标签($x=1,\neg x=0$)；

输出：重构边界域 B；

步骤 1　由公式(3)和(4)计算数据集中每个对象 x 在类别 X 上的条件概率；

步骤 2　根据公式(10)求解最优化损失函数得到阈值参数 α,β；

步骤 3　依据三支决策规则，if $p_j\geq\alpha$，则将 x_j 划分为正域 $POS(X)$，if $\beta<p_k<\alpha$ 则将 x_k 划分为边界域 $BND(X)$，if $p_t\leq\beta$，则将 x_i 划分为负域 $NEG(X)$；

步骤 4　正域 $POS(X)$：$for\ j=1\ to\ n, n<m$，如果 $x_i\neq 1$，则把 x_j 存入 $\Delta POS(X)$；

步骤 5　负域 $NEG(X)$：$for\ t=1\ to\ n, n<m$，如果 $\neg x_i\neq 0$，则把 x_t 存入 $\Delta NEG(X)$；

步骤 6　根据(11)式重构边界域，即 $\Delta BND(X)$；

步骤 7　$B=\Delta BND(X)$；

步骤 8　采用 Weka 工具中 NB、SVM、J48 和 RF 等分类器对重构边界域 B 进行分类验证；

步骤 9 结束.

3 实验分析

3.1 实验数据

为验证本文 TWD-TP 模型的有效性,在 4 种公开数据集上进行仿真实验,其下载来源为:http://archive.ics.uci.edu/ml/datasets.html,4 种数据集均为二分类数据集,详细介绍如表 2 所示.

表 2 数据集信息

Tab.2 Data set

Number	Dataset Name	Attribute	Classification	Number
1	Breast-cancer	10	Benign/Maligant	458/241
2	Haberman's Surival	4	Long/Short	225/81
3	Tic-Tac-Toe	10	Win/Fail	626/332
4	Mammographic Mass	6	Benign/Maligant	516/445

3.2 实验结果

为检验本文模型对数据处理的有效性,从分类性能和与传统三支决策分类方法对比这 2 个方面进行验证.

3.2.1 分类性能

为验证本文模型的分类性能,实验使用 Weka 中几种常用的分类器对样本数据进行分类验证,并将结果与等量原始数据的分类性能进行对比,实验均采用十折交叉方法,结果对比如图 5～8 所示.

从图 5～8 中可看出,本文 TWD-TP 模型在分类能力上表现出良好的分类性能,除 Haberman's Survival 数据在 NaiveBayes 分类器上准确率略低外,其他数据集的分类准确率均高于原始数据样本的准确率,同时也高于随机抽取 6 组与重构样本数等量样本准确率的均值.为保证实验样本数目的一致,更合理地验证所提方法的有效性,实验除了同原始数据集在分类性能上做对比之外,还分别在每个数据集中随机抽取 6 组与重构样本数等量的样本,对其分类精度求均值进行比较,具体结果如表 3 所示.

从表 3 可以看出,本文算法在数据分类能力方面表现出良好的分类性能,针对边界域中部分难以划分的样本具有较优的区分度.比如 Haberman's Survival 数据采用 Random Forest 分类器,分类精度从原始数据集的 67.32%提高到 82.55%,准确率增加 15.23%,同时随机抽取与重构样本数等量的 6 组数据的分类准确率均值从 71.39%提高到 82.55%,准确率增加 11.16%;Mammographic Mass 数据集采用 Lib-SVM 分类器,分类精度从原始数据集的 79.61%提高到 91.57%,准确率增加 11.96%,随机抽取与重构样本数等量的 6 组数据的分类准确率均值从 78.91%提高到 91.57%,准确率增加 12.66%,从而说明本文 TWD-TP 模型可行有效.

表 3 不同的数据集在各个分类器上的分类精度对比

Tab.3 Comparison of classification accuracy of different data sets on each classifier %

Classifier	Breast-cancer	Haberman's Surival	Tic-Tac-Toe	Mammographic Mass
Naive Bayes	95.85/96.32/97.05	74.83/73.60/72.67	71.60/70.14/74.13	80.54/80.14/89.71
Lib-SVM	95.85/95.51/97.79	73.52/74.18/75.58	87.78/80.68//84.68	79.61/78.91/91.57
J48	95.27/93.67/96.32	71.89/73.36/75.00	88.51/90.04/91.90	81.89/82.47/90.63
Random Forest	96.56/96.46/97.79	67.32/71.39/82.55	94.67/92.56/95.95t	79.71/91.57/80.55

注:表中斜线左侧数据为原始数据集的分类精度,斜线右侧为重构样本集的分类精度,中间部分为从原始数据集随机抽取与重构样本数目等量的 6 组数据分类精度结果的均值.

3.2.2 与传统三支决策分类方法对比

为保证实验的科学性和合理性,本次实验选用相同的数据集,阈值设置与本文 TWD-TP 模型一致,将本文方法与传统三支决策分类方法在分类精度上作比较,实验结果如表 4 所示.

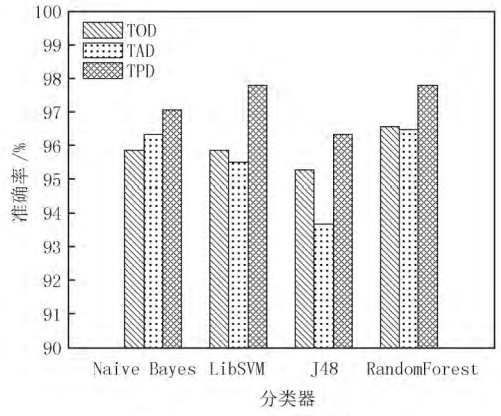

图 5　Breast-cancer 数据集

Fig. 5　Breast-cancer data sets

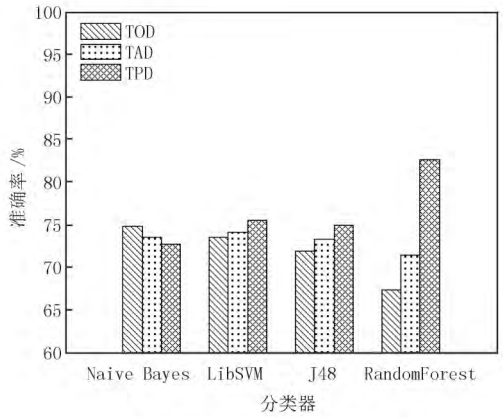

图 6　Haberman's Suriva 数据集

Fig. 6　Haberman's Surival data sets

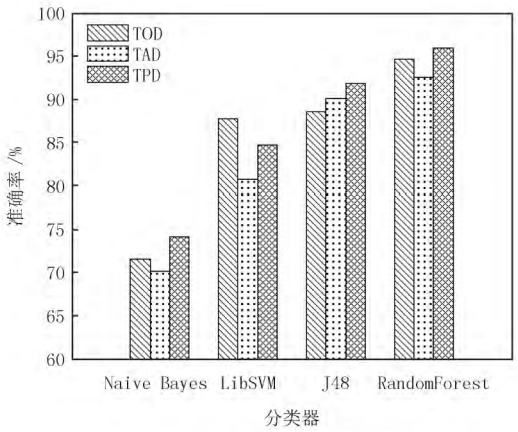

图 7　Tic-Tac-Toe 数据集

Fig. 7　Tic-Tac-Toe data sets

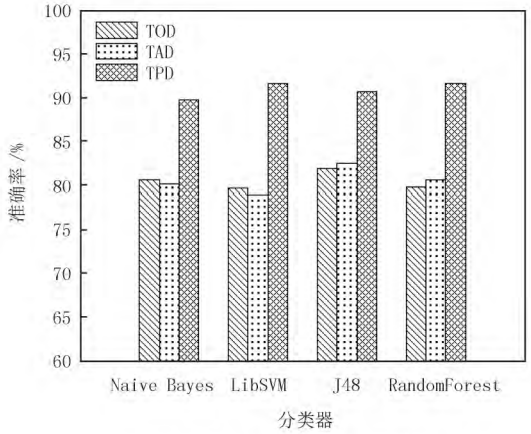

图 8　Mammographic Mass 数据集

Fig. 8　Mammographic Mass data sets

表 4　与传统三支决策分类精度对比

Tab.4　Comparison with classification accuracy of traditional three decision　　%

Data set	Breast-cancer	Haberman's Surival	Tic-Tac-Toe	Mammographic Mass
3WD(TOD)	95.27	70.26	90.02	90.21
3WD(TAD)	76.81	68.21	84.61	90.94
3WD(TPD)	96.89	80.41	92.91	95.31

注：表中 TOD 为原始数据集，TPD 为本文重构样本集，TAD 为从原始数据集随机抽取与重构样本数目等量的 6 组数据分类精度的均值.

从表 4 中可看出，对于不同数据集，基于三支决策的二阶段分类方法的分类精度均高于传统三支决策方法，其中在 Haberman 数据集上的分类精度高达 80.41%，比随机抽取与重构样本数目等量的 6 组数据分类精度的均值高 12.20%，同时也比原始数据的分类精度高 10.15%.针对实验过程中，重构边界域和原始数据集经过三支决策划分后的边界域大小不一致问题，本文从两个边界域中能被正确划分的样本数目进行对比说明.经过实验分析，原始数据集经过三支决策划分后边界域中不能被直接划分的样本，在对其进行重构划分后，该部分样本数目减少，即出现一部分样本被正确划分.从表 5 可看出，对每个数据集，其重构边界域中不能被正确划分的样本数目比原始数据集经过三支决策划分后的边界域样本数目都有所减少，进而说明本文 TWD-TP 模型合理有效.

4 结束语

三支决策在分类任务中考虑误分类代价,能使误分类风险尽可能减少.基于样本特征的机器学习算法在分类任务中通常追求较高分类精度.本文研究两者在分类问题上的优势,提出一种新的基于三支决策的二阶段分类模型.其中对三支决策边界域划分,突破原有边界域的限制.通过误分类样本引入,能使重构边界域对象在被划分时,训练过程更具针对性,能使原始边界域中一些不能被划分的样本正确划分,分类准确率得到提升,同时 TWD-TP 模型也能在三支决策分类中筛选出部分具有较高误分类特征的样本.下一步工作将重点对该模型进行适当扩展,并将其应用到实际的问题中去.

表 5 原始边界与重构边界不能被划分样本数目对比

Tab.5 Comparison of the number of samples that cannot be divided %

Number	Breast-cancer	Haberman's Surival	Tic-Tac-Toe	Mammographic Mass
TOB	33	91	95	94
TPB	8	28	49	25

注:TOB 为原始数据集经过三支决策划分后边界域的样本数,TPB 为重构边界域中不能被正确划分的样本数.

参 考 文 献

[1] Yao Y Y.An outline of a theory of three-way decisions[C]//Rough Sets and Current Trends in Computing 8th international conference.Berlin:Springer,2012:1-17.

[2] 王思华,杨桐,段启凡,等.基于 DT 法和粗糙集理论的接地网安全性状态评定[J].电力系统保护与控制,2017,45(2):48-54.

[3] Jia X Y,Liao W H,Tang Z M,et al.Minimum cost attribute reduction in decision-theoretic rough set models[J].Information Sciences,2013,219(1):151-167.

[4] Li W,Miao D Q,Wang W L,et al.Hierarchical rough decision theoretic framework for text classification[C]//Proceedings of The 9th International Conference on Cognitive Informatics.piscataway:IEEE Press,2010:484-489.

[5] Li H X,Zhou X Z,Zhao J B,et al.Cost-sensitive classification based on decision-theoretic rough set model[C]//Proceedings of The 7th International Conference on Rough Sets and Knowledge Technology.Berlin:Springer,2012:379-388.

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

[7] Yao Y Y.The superiority of three-way decisions in probabilistic rough set models[J].Information Sciences,2011,181(6):1080-1096.

[8] Huang J J,Wang J,Yao Y Y,et al.The cost sensitive three-way recommendations by learning pair wise preferences[J].International Journal of Approximate Reasoning,2017,86:28-40.

[9] 仇国芳,王小宁.基于三支决策的医院分级诊疗决策研究[J].河南师范大学学报(自然科学版),2018,46(3):106-111.

[10] 陈夏艳,陈洁.基于代价敏感边界域处理的社团发现算法[J].数码设计,2017,6(3):1672-9129.

[11] Li Y F,Zhang L B,Xu Y,et al.Enhancing binary classification by modeling uncertain boundary in three-way decisions[J].IEEE Transactions on Knowledge and Data Engineer,2017,29 (7):1438-1451.

[12] 徐久成,刘洋洋.基于三支决策的支持向量机增量学习方法[J].计算机科学,2015,42(6):82-87.

[13] 徐存东,张锐,王荣荣,等.基于改进支持向量机的盐碱地信息精确提取方法研究[J].灌溉排水学报,2018,37(9):62-68.

[14] Li W W,Huang Z Q,Jia X Y.Two-phase classification based on three-way decisions[C]//International Conference on Rough Sets & Knowledge Technology.Berlin:Spinger,2013:338-345.

[15] Li W W,Huang Z Q,Li Q.Three-way decisions based software defect prediction[J].Knowledge-Based Syetems,2016,91:263-274.

[16] 刘盾,姚一豫,李天瑞.三支决策粗糙集[J].计算机科学,2011,38(1):246-250.

[17] Liu D,Li T R,Liang D C.Incorporating logistic regression to decision-theoretic rough sets for classification[J].International Journal of Approximate Reasoning,2013,559(1):197-210.

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

[19] Jia X Y,Shang L.Three-way decisions versus two-way decisions on filtering spam email[M].London:Transctions on Rough Sets,2014:69-91.

(下转第 124 页)

[12] 查思静,周浩祥,游永豪,等.2014 年太仓竞走世界杯我国女子 10 km 队员竞走技术的运动学特征[J].南京体育学院学报(自然科学版),2015(5):22-26.

[13] 郎雪梅,纪仲秋.我国优秀女子竞走运动员竞走技术的生物力学分析[J].中国体育科技,2003(4):32-33.

Estimating walking athlete's step length and step frequency based on the acceleration sensor

Tang Jianjun, Zhou Yang

(College of Education, Beijing Sport University, Beijing 100084, China)

Abstract: Sports capture technology is a powerful assistant for sports training and physical education teaching, which can capture the motion parameters and assist the coaches to evaluate and adjust the training effect. Based on the three-axis acceleration sensor, an algorithm is designed to calculate the walking step length and step frequency. The three-axis motion information of the athlete is captured, the noise interference of the sampled data is eliminated by the low-pass filtering method, the step frequency is solved based on the acceleration curve in the gait cycle, and the calculation model of the walking step is established. The test data show that the calculated step size and step frequency are high and the error is small. Obtaining the step size and step frequency of each gait can detect the stability of the technical structure of the walking race in the training and improve the training efficiency.

Keywords: tri-axial accelerometer; motion capture technology; step length; step frequency; walking

[责任编校 杨浦 王凤产]

(上接第 34 页)

Research on two-stage classification model based on three-way decisions

Xu Jiucheng, Xu Zhanwei, Li Mengfan, Wang Nan

(College of Computer and Information Engineering; Henan Technology Research Center for Computational Intelligence and Data Mining, Henan Normal University, Xinxiang 453007, China)

Abstract: Aiming at the further division of the three-way decisions boundary domains, the problem of insufficient classification accuracy of the boundary knowledge of the three-way decisions caused by insufficient information, this paper proposes a new two-stage classification model based on three-way decisions(TWD-TP). In the first stage, the conditional probabilities of samples in three-way decisions are constructed by Bayesian rule, the required thresholds are obtained by solving the optimal loss function. Then the data sets are divided according to the three decision rules. However, the three-way decisions are based on the division of least-risk Bayes decision theory, including some misclassified samples in positive and negative domains. In the second phase, the samples of misclassification in positive domain and negative domain are introduced into the delayed decision domain as incremental information by class label index to construct new boundary domain, that is, reconstruction boundary domain. Finally, the classifier is used to perform classification verification on the reconstruction boundary domain objects. The experimental results show that the TWD-TP model proposed in this paper can not only filter out the samples with high misclassification features in the three-way decisions division, but also can correctly divide the previously undivided samples in the reconstruction boundary and improve the classification accuracy.

Keywords: three-way decisions; two-stage; incremental information; boundary domain

[责任编校 陈留院 赵晓华]