

文章编号: 1000-8055(2023)05-1250-09

doi: 10.13224/j.cnki.jasp.20210304

基于近似粗糙集分辨力分类算法的飞机类型预测

国 强^{1,2}, 赵佳琦¹

- (1. 中国航空无线电电子研究所 航空电子系统综合技术重点实验室, 上海 200040;
2. 哈尔滨工程大学 信息与通信工程学院, 哈尔滨 150001)

摘 要: 传统的 C4.5 分类决策树作为数据分类算法具有计算简单、准确率高的优势, 由于飞机具有参数多和数据量大的因素, C4.5 算法需要对连续属性值进行多次顺序扫描, 分类时间效率较低。针对此问题, 提出近似粗糙集和决策分辨力分类算法, 利用粗糙集近似度来判断属性划分样本数据能力, 并将其代入到决策分辨力算法中, 以决策分辨力最大的属性作为分裂特征建立分类决策树。算法在保证分类决策准确率的同时, 提高计算效率并减少过拟合问题的产生。通过对 UCI (University of California, Irvine) 数据集上多组数据样本的对比实验分析, 验证了本文提出 PSRP (rough set and resolving power) 的算法在保证相同准确率的情况下, 平均计算时间效率提升约 10%, 可靠性提升 2%。

关 键 词: 近似粗糙集; 决策分辨力; 分类决策树; 数据融合; 飞机类型预测

中图分类号: V247; TP273

文献标志码: A

Aircraft type prediction based on approximate rough set resolution classification algorithm

GUO Qiang^{1,2}, ZHAO Jiaqi¹

- (1. Science and Technology on Avionics Integration Laboratory,
China Aviation Radio Electronics Research Institute, Shanghai 200040, China;
2. College of Information and Communication Engineering,
Harbin Engineering University, Harbin 150001, China)

Abstract: As a traditional data classification algorithm, C4.5 Decision Tree has the advantages of simple calculation and high accuracy. Due to the fact that aircraft has many parameters and large amount of data, C4.5 Decision Tree requires multiple sequential scanning of continuous attribute values, but the classification time efficiency is low. To solve this problem, an approximate rough set resolution classification algorithm was proposed. Rough set approximation was used to determine the ability of attributes to divide sample data. According to the probability acquired, the attribute with the largest resolution was used as the split character to establish the classification decision tree. The algorithm improved the computational efficiency and reduced the generation of over-fitting problem while ensuring the accuracy of classification decision. Through comparative experimental analysis of multiple sets of data samples on UCI (University of California, Irvine) databases, it proved that the PSRP (rough set and resolving power) algorithm improved the average computational time efficiency by about 10% and reliability by 2% while ensuring the same accuracy.

收稿日期: 2021-06-16

基金项目: 航空科学基金(202000550P6001)

作者简介: 国强(1972—), 男, 教授, 博士, 主要从事雷达通信抗干扰及电子对抗的研究。E-mail: guoqiang@hrbeu.edu.cn

通信作者: 赵佳琦(1997—), 女, 硕士, 主要从事粗糙集理论及飞机类型预测的研究。E-mail: zhaojq97@hrbeu.edu.cn

引用格式: 国强, 赵佳琦. 基于近似粗糙集分辨力分类算法的飞机类型预测[J]. 航空动力学报, 2023, 38(5): 1250-1258. GUO Qiang, ZHAO Jiaqi. Aircraft type prediction based on approximate rough set resolution classification algorithm[J]. Journal of Aerospace Power, 2023, 38(5): 1250-1258.

Keywords: approximate rough set; decision resolution power; classification decision tree; data fusion; aircraft type prediction

随着信息技术的发展,掌握信息战是现代战争的必然要求,是陆海空联合作战的必要联络体系。信息技术逐渐成为民用和军事领域的重要组成部分,广泛应用于远程预警、目标检测^[1]、导弹制导、目标跟踪^[2]、遥感测控^[3]、抗密集欺骗式干扰^[4]及低分辨雷达目标识别^[5]等军事行动中。信息技术通过对海量信息处理以及数据挖掘可以预测战场发展趋势,在实际作战中信息技术更是扮演着重要角色。通过分析数据信息,灵活应对战局的变化从而占得先机。进一步加快作战节奏,提高识别目标的准确率与精确度,使指挥效益更加准确可靠。

数据分类是预测飞机类型的重要环节,由于雷达组网系统采集的目标特征信息存在数据量大、不确定性强的弊端^[6],建立决策树形成预测分类模型尤为重要,并以此完成对未知目标的识别预测功能。欧阳瑞麒等^[7]使用卷积神经网络对飞机图像进行识别,目标识别能力较好。刘相云等^[8]提出在遥感图像下飞机目标识别算法,通过提取显著目标并定位连通区域来确定目标对象。现有的飞机类型预测大多通过分析图像来识别,具有一定的距离局限性,因此分类决策模型成为飞机类型预测的关键性问题。

粗糙集可以有效针对未知目标的不确定、不完备、不一致信息进行分析和处理。与证据理论、模糊推理^[9]、随机集理论、可能性理论等相比,粗糙集^[10]运用传统集合理论精确计算数据上界和下界,以此判断未知目标类型。王越等^[11]提出基于变精度粗糙集的决策树构造方法,该算法只能处理离散型数据,对于阈值参数的选取没有理论依据。刘军^[12]提出基于叶枝比率的决策树构建算法,该算法提高了决策树预测的准确率,但计算量较大,对于信息量大的数据缺乏快速分类能力。针对上述问题,本文采用近似粗糙集算法给出相对最优的提取规则。

分类决策树的核心算法是基于信息熵的ID3(iterative dichotomiser 3)和C4.5算法^[13],此算法分类规则准确率高且便于理解。为了保证分类决策树一直沿着信息熵最大的方向生长,Fisher等利用ID4(iterative dichotomiser 4)算法规则提出了增量式学习系统,Utoff进而提出了增量决策树归纳算法ID5R,解决了条件属性连续分类的问题。尚朝轩等^[14]提出了基于类决策树分类的特征层融

合识别算法,克服数据量大的问题并提高了信息利用率和抗干扰能力。但仍然存在计算效率较低、信息增益率计算不稳定等问题。

基于上述两种算法的描述,本文提出基于近似粗糙集和决策分辨力的分类算法(rough set and resolving power, RSRP),决策分辨力是表征不确定信息的方法,其与信息熵^[15]不同点是可以直接度量信息量大小,该方法简洁高效便于自适应实现。本文使用近似度来判断属性划分不确定性数据的概率,代入到决策分辨力中,将分辨力最大的属性作为分裂特征构造分类决策树。本方法在C4.5决策树算法模型的基础上减少了对信息增益率的计算步骤,在保证分类准确率的基础上,提高计算效率和降低决策树复杂度。通过多组来自不同领域的真实数据集的分类数据对比分析,验证了本算法的准确可靠性。

1 近似粗糙集

1.1 粗糙集基本理论

粗糙集理论是波兰数学家Pawlak提出的一种处理解决不一致性和不确定性问题的典型数学方法^[16]。运用集合二元关系的不可分辨性来划分论域,得到不可分辨关系的等价类,近似空间由全体等价类建立构造而成^[17]。

为了更加直观的理解粗糙集理论,引入一个信息系统 $S = (U, A, V, f)$,非空样本集 U 称为论域,属性集合 $A = C \cup D$, C 和 D 分别表示条件属性集和决策属性集, V 表示属性的值域,信息函数集 f 是 U 和 A 的关系集。

设任意属性集 r 是 A 上的等价关系,论域 U 上的不可分辨关系定义为

$$\begin{aligned} N(r) &= \{(x, y) | (x, y) \in U \times U, \\ &f(x, a) = f(y, a), \forall a \in r\} \end{aligned} \quad (1)$$

不可分辨关系 $N(r)$ 是一个等价关系,记 $[x]_r = \{y \in U | y \in N(r)\}$ 为不可分辨关系 $N(r)$ 的包含对象 x 的等价类, U 中所有等价类构成 U 的划分,记为 $U/N(r) = \{[x]_r | x \in U\}$ 。

粗糙集的另一重要概念是上、下近似集,设集合 X 属于 U , r 是 A 上的等价关系,定义两个子集为

$$\bar{r}(X) = \{x | x \in U \wedge [x]_r \cap X \neq \emptyset\} \quad (2)$$

$$\underline{r}(X) = \{x | x \in U \wedge [x]_r \subset X\} \quad (3)$$

分别称 $\bar{r}(X)$ 和 $\underline{r}(X)$ 为 X 的 r 上近似集和 r 下近似集。

通过对数据上界和下界的刻画构造目标集 X 近似空间,下近似集表示在 (U,r) 知识基下精确分类,上近似集表示存在等价关系属于目标集的粗糙分类。由于用两个精确集合对目标集分类过于绝对,在实际应用中,存在信息采集错误、数据不一致、噪声数据比重较大等情况^[18],对实际分类结果造成影响,需要解决的问题是如何利用现有的知识粒度对目标集 X 进行划分。

现有的扩展粗糙集理论有基于等价关系泛化,基于包含算子的扩展,基于粗糙集的效能评估^[19]等。最经典的扩展粗糙集模型有概率粗糙集和变精度粗糙集^[20],两种方法针对传统理论判别条件过于绝对的问题,具有较强的容错能力。但仍存在明显不足,其模型中的阈值参数由专家经验给出,本文针对阈值选取提出了相对合理的最优近似粗糙集模型。

1.2 近似粗糙集模型

在等价关系情况下,构建目标集 X 的最优近似粗糙集,要先获得 x 属于目标集合的隶属函数,设 U 是非空样本集,对象目标集 $X \in U$,对于任意 $x \in U$,论域 U 关于任意属性集 r 上的等价关系 $U/N(r)$,隶属函数定义为

$$\mu_x^r(x) = \frac{|[x]_r \cap X|}{|[x]_r|} \quad (4)$$

$\mu_x^r(x) \in [0,1]$,它表示任意元素 x 属于集合 X 的概率,隶属度等于1时表示粗糙集下近似集,隶属度在 $0 \sim 1$ 之间表示介于上近似集和下近似集之间的元素。因此通过描述隶属度大小,可以模糊处理粗糙集的上下界,从而构建介于上下近似集之间的近似粗糙集 $r_\lambda(X)$,其定义为

$$r_\lambda(X) = \{x \in U | \mu_x^r(x) \geq \lambda\} \quad (0 < \lambda \leq 1) \quad (5)$$

λ 是隶属度的取值范围,将 λ 作为限定隶属函数大小的标准,相当于扩大了下近似集的包容程度,从一定程度上提高了粗糙集模型的容错能力和系统稳健性,这样一方面完善了粗糙集近似空间的概念,另一方面提高粗糙集理论从不确定数据中得到确定性信息的能力。

在不同数据信息系统中,存在来源不同、分布不均、不确定性强的数据信息,如何找到相对最优的近似集取值是问题的关键,由张清华等^[21]提出的粗糙集最优近似集的概念可知:以 $r_\lambda(X)$

作为目标集 X 的近似粗糙集,则相似度可定义为

$$S(X, r_\lambda(X)) = \frac{|X \cap r_\lambda(X)|}{|X \cup r_\lambda(X)|} \quad (6)$$

根据文章中的定理,设论域 U 是一个非空样本集,对象目标集 $X \in U$, r 是 U 上的等价关系,对任意的 $0.5 < \lambda \leq 1$,有

$$S(X, \underline{r}(X)) \leq S(X, r_\lambda(X)) \leq S(X, r_{0.5}(X)) \quad (7)$$

设论域 U 是一个非空样本集,对象目标集 $X \in U$, R 是属性集 A 上的等价关系,对任意的 $0.5 \leq \lambda_1 \leq \lambda_2 \leq 1$,则

$$S(X, R_{\lambda_2}(X)) \leq S(X, R_{\lambda_1}(X)) \quad (8)$$

式(7)、式(8)表明,当近似度在区间 $[0.5,1]$ 上时, $r_\lambda(X)$ 的近似度要比 $\underline{r}(X)$ 更接近目标集合,且 $r_{0.5}(X)$ 是目标集合的最优近似集。在区间 $[0,0.5]$ 上并不一定最优,当满足一定条件时,可作为 X 的最优近似集。

通过使用UCI(University of California, Irvine)数据集中多组不同类型数据集代入近似粗糙集模型中,得到适应性强的最优近似粗糙集。为了更清晰的说明本节列举其中一个公开数据集实例fertility diagnosis进行分析。在数据集决策信息表中, $U = \{x_1, x_2, \dots, x_{100}\}$ 为非空样本集,条件属性集 $C = \{a_1, a_2, \dots, a_9\}$,决策属性集 $D = \{d\}$,该目标集包含两个离散决策值 $d = \{1, 2\}$, λ 取 $[0,1]$ 。根据粗糙集的基本理论知识,可得到两个等价关系 $U/N(C)$, $U/N(D)$ 。由决策属性集中两个取值形成决策目标集 X_1 和 X_2 ,两个决策集在近似空间的不同划分值 λ 的条件下,得到如图1所示近似粗糙集分类的正确率,错误率和未识别率。

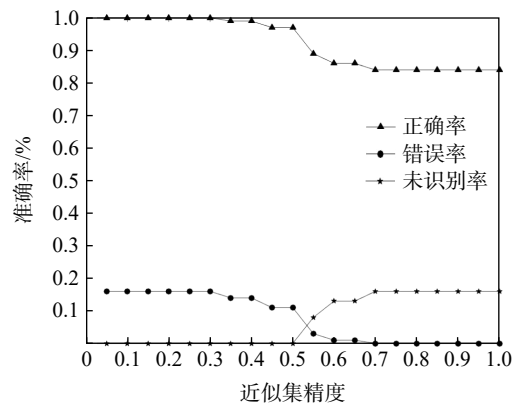


图1 近似粗糙集正确率,错误率和未识别率

Fig. 1 Approximate rough set accuracy, error rate and unrecognized rate

从图1可以得出, $\lambda \approx 0$ 时, 即决策规则越接近上近似集正确率提高, 但错误率也是最高。 $\lambda = 1$ 时, 即决策规则越接近下近似集错误率为0, 但随之正确率降低, 未识别率提高。 $\lambda = 0.5$ 时, 决策规则正确率相对较高, 错误率和未识别率也相对较低, 相对于前面两种情况, 将近似度0.5作为提取决策规则分类效果更好。经过多组实验的验证, 在多数分类决策情况下, $r_{0.5}(X)$ 可以作为最优近似粗糙集决策规则。

由于粗糙集理论的变量都以离散值的形式存在, 不能对连续属性值处理, 扩展的粗糙集仍在粒度水平下用对属性值离散化的方法解决对样本集分类等问题, 离散化处理不能普遍适用于对整体样本集的特征描述, 在实际应用中有明显不足。因此引入C4.5决策树模型, 解决了对连续值处理的问题, 提高算法模型的适用范围, 产生规则易于理解和实现。

2 决策分辨力模型

2.1 传统决策算法缺点

分类决策树是一种将数据集内属性特征作为决策依据并进行分类的树形结构, 通过对一组无规则、无次序的样本集递归构建决策树^[22], 来对未知属性类别的样本进行分析预测。它分为内部节点和叶子节点两类, 内部节点由条件属性和划分条件构成, 叶子节点由决策属性构成, 即样本决策所属类别。信息熵是分类决策树最重要的数学理论知识, 传统的决策算法ID3和C4.5都是以信息熵的增益或增益率作为分类决策规则。

设 S 是一个非空数据样本集合, 决策属性集 $D = \{d_1, d_2, \dots, d_m\}$ 有个 m 取值, 则决策属性集对样本集划分的信息熵为

$$I(D) = - \sum_{i=1}^m p_i \log_2(p_i) \quad (9)$$

p_i 表示每个决策属性占样本总数的概率, 条件属性集 $C = \{a_1, a_2, \dots, a_v\}$, 按照 v 个条件属性集对样本集划分的信息熵为

$$I_c(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} I(D_j) \quad (10)$$

可得每个条件属性 C 的信息增益为

$$G(C) = I(D) - I_c(D) \quad (11)$$

传统ID3决策算法把信息增益最大的属性作

为分类属性, 把分类属性值相同的样本分类到同一个子样本集中, 递归调用本算法。当子样本集内条件属性全部相同或样本个数唯一时, 将决策属性值作为该分支的叶子节点并停止递归调用。其健壮性好且搜索空间完整, 但分类时倾向于取值样本较多的属性, 分类结果会因样本集决策属性值个数不平衡而导致分类准确率降低。

C4.5决策树作为ID3决策树算法的改进, 用信息增益率作为判决条件, 能解决不平衡数据样本分类问题。可以直接处理包含连续数据特征的样本集, 但由于算法对样本数据多次排序和顺序扫描, 导致计算效率低, 在处理大规模数据集时耗时较大^[23]。因此本文引入决策分辨力理论, 将其与近似粗糙集模型相结合提高计算效率, 并简化决策树构建复杂度。

2.2 决策分辨力理论

在决策信息系统中, 考虑的不仅是某一情况发生的不确定度大小, 而是考虑完整描述信息系统所有可能发生情况的不确定度。信息熵是经典的不确定性度量方法, 其扩展方法还有互信息函数^[24]和分辨力信息函数等, 这些方法可以用于解决目标检测和识别等信息不确定性强问题^[25]。

决策分辨力与信息熵都是通过信息传递体现出信息量大小的, 而决策分辨力主要通过两个概率分布共同描述不确定信息的差异程度和可信度。与传统信息熵相比, 通过整体概率分布和条件概率分布共同描述信息的不确定性, 对样本信息描述更加可靠, 对不一致信息的辨识和去除能力更加准确, 对处理复杂关系下的数据样本集能力较强。

设 U 为包含 L 个样本的数据集, C 为包含 v 个条件的属性集, 决策属性集 D 有个 m 取值, 决策分辨力定义为

$$H(p, q) = - \sum_{j=1}^v p_j \log_2(q_j) \quad (12)$$

其中 p_j 表示条件 a_j 情况下属于决策值 d_i 的条件概率, q_j 表示决策值 d_i 下所有条件属性取值个数在样本集中的占比, 也称为不确定度, 即 $q_j = |d_i|/L$ 。

当条件属性在样本集中的不确定度 q_j 越大, 同时条件概率 p_j 越大, 该条件属性决策分辨力越大, 表示该条件属性将样本均衡分类的能力越强。决策分辨力可以作为信息熵的扩展算法, 当只有两个决策属性取值时, 决策分辨力效果等同于传

统信息熵算法;当有两个以上决策属性取值时,时间效率相对提高,并降低决策树结构复杂度,使算法简洁高效。本文实验数据集针对不同决策属性取值个数进行分析研究。

3 近似粗糙集与分辨力模型

3.1 算法优化模型建立

近似粗糙集可以处理数据信息不确定性强的样本,不足之处在于无法处理包含连续属性值的样本集;决策分辨力算法的引入可以解决该问题,提高算法适用范围和泛化能力,而使用近似粗糙集和决策分辨力计算概率值,可以降低决策树的复杂度,并提高计算效率。

在分类决策算法之前通常要对样本集进行数据预处理,其中几个重要的部分是数据约简^[26]、样本抽样和决策树剪枝^[27],这些问题有很多种处理方案和策略,本模型使用近似粗糙集进行决策分类。RSRP 算法模型流程图如图2所示。其中,设 N 为每次迭代循环结点,经过一次迭代循环将得到的属性类赋值给 N 。

本文 RSRP 算法模型主要分为两个部分,具体模型建立方案如下:

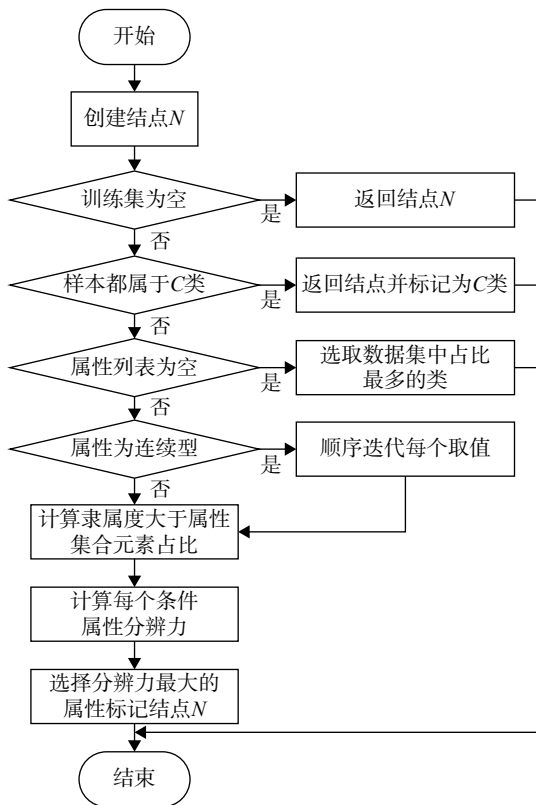


图2 RSRP 算法流程图

Fig. 2 RSRP algorithm flow chart

1) 近似粗糙集概率计算

本算法模型可以用本文第1.2节的隶属函数表示,其中 $[x]$ 表示决策属性取值个数, X 表示条件属性取值属于决策属性取值的个数,由于隶属函数大于0.5时,该条件属性划分目标正确率较高,因此把隶属函数大于0.5的 X 划分到支持近似集,并将近似集元素个数占样本总数比作为决策分辨力的条件概率 p_i 进行计算。

2) 决策分辨力分类规则

最大决策分辨力用于构建分类决策树的分裂特征,选取最优分裂特征可以将无规则、无次序的数据样本集充分分类,使树每个枝干的样本特征性更明显。决策分辨力的条件概率和不确定度由近似粗糙集划分,能统一属性的离散取值和连续取值以判别标准进行计算,使分类规则更加简便可靠,样本分布更加均匀。

3.2 算法具体步骤

输入:训练信息系统 $S = (U, A, V, f)$,非空样本训练集 U 包含 L 个样本, $A = C \cup D$, C 为条件属性集, D 为决策属性集,粗糙集近似度为 λ 。

输出:tree 一个分类决策树,每一节点包含四个元素,tree.feature 为节点分裂特征;tree.location 分裂特征为连续属性时,location 为分裂区间;分裂特征为离散属性时,location 为属性的离散取值;tree.child 为存放子节点的结构体,tree.value 为属性样本值。

步骤1 判断训练样本条件属性集 C 中 v 个属性取值的离散性,令取值种类大于10个为连续属性;循环调用条件属性 a_j ,若属性 a_j 是离散的,则跳转到步骤2;若属性 a_j 是连续的,转到步骤3。

步骤2 计算样本在 a_j 条件下的各个取值属于决策值的隶属度 μ_j 。

步骤3 将 a_j 条件取值排序并取唯一,顺序迭代每个取值,计算样本在 a_j 条件下大于或小于等于条件取值属于决策属性值的隶属度 μ_j 。

步骤4 计算 $\mu_j > \lambda$ 的条件属性集合包含的样本个数占总样本的比例为 p_j 。

步骤5 计算 a_j 条件下条件属性集合包含的样本个数占总样本的比例为 q_j 。

步骤6 根据决策分辨力算法获得各条件属性的 $H(p, q)$,以最大决策分辨力的属性为该节点的tree.feature 分裂特征,按照该属性值继续分类。

步骤7 递归调用步骤1构建分类决策树的内部节点,直到集合所有样本决策值完全相同或

只包含单一条件属性, 停止递归。

输出构造决策树后, 将新采集到的测试样本集代入到分类决策树中, 即得到预测决策结果。

3.3 算法模型性能评价指标

在分类决策算法模型中, 有很多算法性能的评估标准, 传统评价指标是分类的准确率, 即预测对了多少样本。本节引入二分类问题中的查准率 e (precision)和查全率 c (recall)组合成的评价指标 F_1 , 由于决策取值可能为多个, 因此将样本决策取值 d_i 实际和预测扩展定义, T 为实际是 d_i 预测也是 d_i 的样本个数; F 为实际不是 d_i 预测是 d_i 的样本个数; P 为实际是 d_i 预测不是 d_i 的样本个数; N 为实际不是 d_i 预测也不是 d_i 的样本个数。

$$e = \frac{1}{m} \sum_{i=1}^m \frac{T}{T+F} \quad (13)$$

$$c = \frac{1}{m} \sum_{i=1}^m \frac{T}{T+N} \quad (14)$$

$$F_1 = \frac{2 \times e \times c}{e + c} \quad (15)$$

除了 F_1 评价指标, 本节还引入两个重要的评价指标时间效率(runtime efficiency)和树形复杂度(tree complexity)。时间效率依赖循环迭代次数的多少, 在分类决策中迭代次数越少计算效率越高。树形复杂度即决策树高度, 其大小依赖分类算法的优劣, 样本分布越均匀, 建立的分类决策树就稳定, 分类的稳定性和泛化能力越好。

4 实验结果与分析

4.1 实验数据

本节通过选取UCI数据库中公开的9组不同领域数据和间谍飞机类型预测的实验验证近似粗糙集决策分辨力分类算法的有效性。各样本数据集的样本个数、条件属性和决策类个数如表1所示。

9组不同领域数据集数据特征根据仪器采集数据或样本固有属性特征组成, 数据集分别为心力衰竭临床记录数据集、丙肝数据集、用户知识建模数据集、肝病患者数据集、电离层数据集、疾病类型检测数据集、输血服务数据集、批发客户类型数据集、病变数据集筛查数据集。spy plane finder为间谍飞机类型预测的数据集, 该数据是从飞机跟踪网站Flightradar24获得的4个多月的飞机应答机检测结果。通过训练普通民航飞

表1 数据集描述

Table 1 Data set description

序号	数据集名称	样本个数	条件属性	决策类
1	heart failure	299	12	2
2	hcvdat	615	12	5
3	user model	258	5	4
4	liver patient	583	10	2
5	ionosphere	351	33	2
6	exasens	399	3	4
7	transfusion	748	4	2
8	wholesale customer	440	7	2
9	messidor	1 151	19	2
10	spy plane finder	597	31	2

机的各项参数, 区分普通飞机和间谍飞机。研究表明, 对飞机的各项参数分类决策有利于提高目标跟踪传感器管理和精准打击的准确率, 因此本文采用Flightradar24网站上飞机的真实飞行参数进行实验分析。

4.2 实验结果

实验分为3个部分: 第1部分是分析RSRP算法近似度 λ 取值[0,1]对样本集分类决策正确率和 F_1 评价指标的影响, 第2部分对比分析C4.5、RoughSets-CART(RS-CART)^[28]和RSRP三种算法的正确率、时间效率和树形复杂度, 第3部分预测间谍飞机类型。

4.2.1 RSRP算法的近似度取值

本节实验主要分析粗糙集近似度取值对构建分类决策树的影响, 由于不同数据集之间的特征不尽相同, 每个条件属性对应决策值占比也不同, 因此大部分实验研究的粗糙集模型都要根据实际情况确定近似度的取值。在实际应用中, 确定最优近似度要多次迭代循环时间效率较低, 不能满足紧急情况下对样本集的分类决策。

使用RSRP算法对数据集wholesale customer分类决策, 由图3可知, 近似度在[0.5,1]范围内分类正确率和 F_1 指标相对较高, 波动相对平稳。对于取值在[0,0.5]范围时, 其分类规则接近上近似集, 由于符合条件的属性类别太多, 无法准确区分条件属性对样本不确定性的分类能力。因此分类时间不足无法准确找到更优近似度的情况下, 取0.5为分类算法最优近似度比较合理。

图4记录了数据集heart failure分类决策的结果, 近似度取值为0.8时比取0.5的准确率高, 由

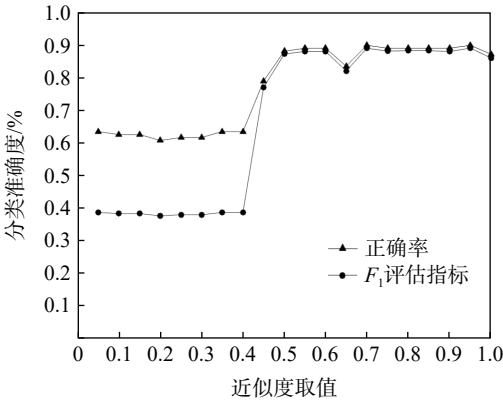


图 3 批发客户数据集的近似度取值
Fig. 3 Approximation value of data set wholesale customer

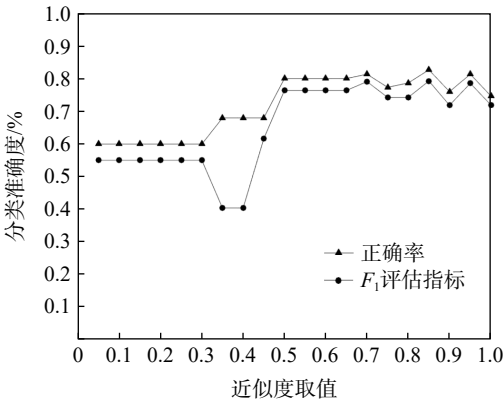


图 4 心力衰竭数据集的近似度取值
Fig. 4 Approximation value of data set heart failure

此表明 0.5 近似度是相对最优取值, 不排除其他取值时有更高的准确率。综合各方面因素采用 0.5 近似度进行分类决策。

4.2.2 分类决策算法比较

本节对 UCI 数据集上的 9 组数据样本集进行

分类决策, 将 C4.5、RS-CART 与本文 RSRP 算法进行对比分析。近似度取 $\lambda = 0.5$, 对比实验采用相同抽样方法进行预测, 预测结果评估指标为: F_1 准确率、时间效率和树形复杂度。实验结果为 50 次随机取值下的平均结果。

C4.5、RS-CART 与本文 RSRP 决策算法实验对比结果如表 2 所示, 从表 2 中得到以下结论:

1) 在 heart failure、hcvdat、ionosphere 样本集中, 3 种算法预测 F_1 准确率大致相同, RSRP 算法预测的时间效率和树形复杂度明显优于 C4.5 算法。

2) 在 liver patient、exasens、messidor 样本集中, RSRP 算法预测 F_1 准确率明显优于 C4.5 和 RS-CART 算法, 3 种算法准确率相对较低, C4.5 和 RS-CART 算法过拟合问题严重, 且由于过拟合时间效率和树形复杂度明显不足, RSRP 算法的各项评估指标没有明显变化, 算法相对稳定。

3) 从整体效果表明, RSRP 算法 F_1 准确率平均值优于 C4.5 算法和 RS-CART 算法, 而在算法稳定性和时间效率上, RSRP 算法明显好于 C4.5 算法, 略优于 RS-CART 算法; 当样本集数据量大的情况下, 为了保证分类的准确率和时间效率, RSRP 算法适应性和可靠性更强。

4.2.3 预测间谍飞机类型

在飞机类型预测问题中由于数据量庞大, 首先对数据进行筛选, 在包含 19 799 架飞机样本的数据集中抽取 500 架普通飞机和 97 架间谍飞机的航行参数组成表 1 中 spy plane finder 数据集, 参数分别为: duration 每个飞行段的飞行时间,

表 2 数据集分类评估指标对比
Table 2 Comparison of evaluation indexes for data set classification

数据集名称	F_1			时间效率			树形复杂度		
	C4.5	RS-CART	RSRP	C4.5	RS-CART	RSRP	C4.5	RS-CART	RSRP
heart failure	0.799 7	0.771 5	0.815 2	0.282 9	0.186 2	0.242 1	6.78	4.42	3.98
hcvdat	0.908 4	0.861 6	0.892 5	1.843 0	1.044 3	1.024 3	7.90	4.82	4.00
user model	0.895 4	0.829 2	0.906 5	0.204 4	0.176 7	0.191 3	5.56	4.54	4.24
liver patient	0.669 3	0.694 2	0.702 0	1.651 7	0.866 1	0.687 5	44.09	6.44	6.39
ionosphere	0.882 3	0.892 0	0.896 4	1.910 0	1.058 4	1.129 2	4.52	3.64	3.46
exasens	0.518 8	0.521 4	0.541 6	0.078 9	0.066 2	0.073 4	8.02	6.12	6.54
transfusion	0.759 1	0.777 5	0.769 5	0.250 5	0.190 5	0.173 6	17.54	7.96	5.02
wholesale customer	0.889 5	0.899 6	0.906 5	0.874 3	0.631 7	0.641 7	8.62	5.20	4.40
messidor	0.592 4	0.555 1	0.641 0	39.241 8	7.753 8	6.821 7	104.2	10.24	6.6
平均值	0.768 3	0.755 8	0.785 7	5.148 6	1.330 4	1.220 5	23.025 6	5.931 1	4.958 9

box 围绕每个飞行段绘制的矩形边界框的面积, speed 飞行速度, altitude 飞行垂直高度, steer 指南针方位变化, squawk 应答器发送代码, flight 飞行总段数, observation 应答器检测总数。实际训练集中的参数 duration、box、speed、altitude 按照 5 个飞行段计算各自比例, 参数 steer 按照 8 个划分的角度区间计算比例, 构成 31 个条件属性集合。

如图 5 所示, 近似度取 [0.5, 1] 区间的预测准确率大致相同, 因此取 $\lambda = 0.5$ 对间谍飞机类型预测进行对比分析实验。

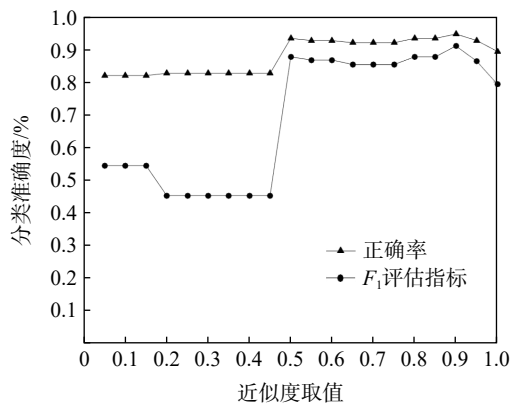


图 5 飞机类型预测近似度取值

Fig. 5 Approximation value of aircraft type prediction

通过表 3 可以看出三种算法 F_1 准确率大致相同, 而 RSRP 算法的时间效率和树形复杂度均优于 C4.5 和 RS-CART 算法, 与 UCI 数据集上其他样本集实验结果一样, RSRP 在保证相同准确率的情况下, 提升了计算时间效率和可靠性。

表 3 飞机类型预测评估指标对比

Table 3 Comparison of aircraft type prediction evaluation indexes

评价指标	C4.5	RS-CART	RSRP
F_1	0.871 0	0.865 5	0.874 4
时间效率	6.069 1	5.573 8	4.717 8
树形复杂度	7.64	4.58	4.94

5 结 论

飞机类型预测在实际作战过程中有至关重要的作用, 通过对样本数据信息挖掘, 不仅能提高目标识别的准确率, 而且能减轻飞行员的工作负担。传统的粗糙集理论只能处理离散数据, 在实际应用中不够灵活, 而 C4.5 算法由于对连续值的多次排序和顺序扫描, 导致计算效率较低。因此

本文提出了基于近似粗糙集和决策分辨力的分类算法, 本算法通过以近似度判断属性划分不确定性数据的概率, 代入到决策分辨力中, 将分辨力最大的属性作为分裂特征构造分类决策树。通过对多组真实数据和飞机类型预测的实验分析, 可以得出本文 RSRP 算法, 在略优于 C4.5 算法准确率的基础上, 提高了算法的时间效率和降低了树形复杂度, 避免了分类决策树由于树形复杂度过高导致的过拟合问题的产生。在军事作战领域中, 提高了目标识别和精准预测的效率, 增强算法的可靠性和稳定性, 在飞机类型预测数据中得到较好结果。

参考文献:

- [1] COLUCCIA A, PARISI G, FASCISTA A. Detection and classification of multirotor drones in radar sensor networks: a review[J]. *Sensors*, 2020, 20(15): 1-22.
- [2] LEI P, WANG J, GUO P, et al. Automatic classification of radar targets with micro-motions using entropy segmentation and time-frequency features[J]. *Electron Commun*, 2011, 65(10): 806-813.
- [3] ZHOU G Q, ZHANG R T, ZHANG D J. Manifold learning co-location decision tree for remotely sensed imagery classification[J]. *Remote Sens*, 2016, 8(855): 1-31.
- [4] 同亚龙, 王彤, 代保全, 等. 一种抗密集欺骗式干扰的机载雷达动目标检测方法[J]. *电子与信息学报*, 2015, 37(3): 658-664.
- [5] TONG Yalong, WANG Tong, DAI Baoquan, et al. A moving target detection method to suppress dense deception jamming for airborne radar[J]. *Journal of Electronics & Information Technology*, 2015, 37(3): 658-664. (in Chinese)
- [6] 王福友, 罗钉, 刘宏伟. 低分辨率机载雷达飞机目标分类识别技术研究[J]. *雷达学报*, 2014, 3(4): 444-449.
- [7] WANG Fuyou, LUO Ding, LIU Hongwei. Low-resolution airborne radar aircraft target classification[J]. *Journal of Radars*, 2014, 3(4): 444-449. (in Chinese)
- [8] 雷涛, 旷生玉, 杨玲. 基于深度学习的雷达辐射源型号识别方法[J]. *电子信息对抗技术*, 2019, 34(4): 29-34.
- [9] LEI Tao, KUANG Shengyu, YANG Ling. Radar emitter type identification algorithm based on deep learning[J]. *Electronic Information Warfare Technology*, 2019, 34(4): 29-34. (in Chinese)
- [10] 欧阳瑞麒, 雍杨, 王兵学. 卷积神经网络在飞机类型识别中的应用[J]. *兵工自动化*, 2017, 36(12): 71-75.
- [11] OUYANG Ruilin, YONG Yang, WANG Bingxue. Application of convolution neural network in aircraft type recognition[J]. *Ordinance Industry Automation*, 2017, 36(12): 71-75. (in Chinese)
- [12] 刘相云, 龚志辉, 金飞, 等. 结合显著图和深度学习的遥感影像飞机目标识别[J]. *测绘通报*, 2019(3): 27-31.
- [13] LIU Xiangyun, GONG Zhihui, JIN Fei, et al. Aircraft target recognition in remote sensing image combined saliency map with deep learning[J]. *Bulletin of Surveying and Mapping*, 2019(3): 27-31. (in Chinese)
- [14] ZHANG K, ZHAN J M, WU W Z. Novel fuzzy rough set models and corresponding applications to multi-criteria decision-making[J]. *Fuzzy Sets and Systems*, 2020(383): 92-126.
- [15] YAO Y Y, SHE Y H. Rough set models in multigranulation spaces[J]. *Information Sciences*, 2016(327): 40-56.

- [11] 王越, 万洪. 一种新的应用变精度粗糙集的决策树构造方法[J]. 重庆理工大学学报(自然科学), 2013, 27(11): 58-64.
WANG Yue, WAN Hong. A new method for constructing decision tree based on variable precision rough set[J]. Journal of Chongqing University of Technology (Natural Science), 2013, 27(11): 58-64. (in Chinese)
- [12] 刘军. 基于叶枝比率的决策树构建算法[J]. 信息安全, 2013(2): 9-12.
LIU Jun. Algorithm of constructing decision tree based on the leaf and branch ratio[J]. Netinfo Security, 2013(2): 9-12. (in Chinese)
- [13] JIANG J F, ZHU X Y, HAN G J, et al. A dynamic trust evaluation and update mechanism based on C4.5 decision tree in underwater wireless sensor networks[J]. IEEE Transactions on Vehicular Technology, 2020, 69(8): 31-40.
- [14] 尚朝轩, 王品, 韩壮志, 等. 基于类决策树分类的特征层融合识别算法[J]. 控制与决策, 2016, 31(6): 1009-1014.
SHANG Chaoxuan, WANG Pin, HAN Zhuangzhi, et al. Feature-level fusion recognition algorithm based on analogy decision tree classification[J]. Control and Decision, 2016, 31(6): 1009-1014. (in Chinese)
- [15] XIAO F Y. Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy[J]. Information Fusion, 2019(46): 23-32.
- [16] 陈泽华, 张裕, 谢刚. 不一致决策表规则获取的粒计算方法[J]. 控制与决策, 2015, 30(4): 709-714.
CHEN Zehua, ZHANG Yu, XIE Gang. GrC method of rule acquisition for inconsistent decision table[J]. Control and Decision, 2015, 30(4): 709-714. (in Chinese)
- [17] LI J H, REN Y, MEI C L, et al. A comparative study of multigranulation rough sets and concept lattices via rule acquisition[J]. Knowledge-Based Systems, 2016, 91: 152-164.
- [18] 孙长嵩, 董西国, 张健沛. 一个基于粗糙集和决策树的最简分类规则集生成算法[J]. 哈尔滨工程大学学报, 2002, 23(5): 87-91.
SUN Changsong, DONG Xiguo, ZHANG Jianpei. Algorithm based on rough set and decision tree to gain minimal classing rule set[J]. Journal of Harbin Engineering University, 2002, 23(5): 87-91. (in Chinese)
- [19] 王礼沅, 张恒喜, 徐浩军. 基于粗糙集的空战效能多指标综合评估模型[J]. 航空学报, 2008, 29(4): 880-885.
WANG Liruan, ZHANG Hengxi, XU Haojun. Multi-index synthesis evaluation model based on rough set theory for air combat efficiency[J]. Acta Aeronautica et Astronautica Sinica, 2008, 29(4): 880-885. (in Chinese)
- [20] JIAN L R, LI M Y. A variable precision fuzzy rough set approach to a fuzzy-rough decision table[J]. Science and Engineering Research Center, 2016: 250-254.
- [21] 张清华, 薛玉斌, 王国胤. 粗糙集的最优近似集[J]. 软件学报, 2016, 27(2): 295-308.
ZHANG Qinghua, XUE Yubin, WANG Guoyin. Optimal approximation sets of rough sets[J]. Journal of Software, 2016, 27(2): 295-308. (in Chinese)
- [22] ZHANG X, MEI C L, CHEN D G, et al. Feature selection in mixed data: a method using a novel fuzzy rough set-based information entropy[J]. Pattern Recognition, 2016(56): 1-15.
- [23] ZHAN J M, ALI M I, MEHMOOD N. On a novel uncertain soft set model: Z-soft fuzzy rough set model and corresponding decision making methods[J]. Applied Soft Computing, 2017(56): 446-457.
- [24] ZHAN J M, LIU Q, HERAWAN T. A novel soft rough set: soft rough hemirings and corresponding multicriteria group decision making[J]. Applied Soft Computing, 2017(54): 393-402.
- [25] 张铁柱, 蒋宏. 机载雷达和红外数据融合的智能目标识别[J]. 红外与激光工程, 2010, 39(4): 756-760.
ZHANG Tiezhu, JIANG Hong. Intelligent target recognition based on the data fusion of radar and infrared imaging sensor[J]. Infrared and Laser Engineering, 2010, 39(4): 756-760. (in Chinese)
- [26] 李雪岩, 李学伟, 蒋君. 基于知识粒度特征的多目标粗糙集属性约简算法[J]. 控制与决策, 2021, 36(1): 196-205.
LI Xueyan, LI Xuewei, JIANG Jun. Multi objective rough set attribute reduction algorithm based on the characteristics of knowledge granularity[J]. Control and Decision, 2021, 36(1): 196-205. (in Chinese)
- [27] BINH T P, ABOLFAZL J, INDRA P. A novel hybrid support vector machine with decision tree for data classification[J]. International Journal of Advanced and Applied Sciences, 2017, 4(9): 138-143.
- [28] 庞梦洋, 索中英, 郑万泽, 等. 基于RS-CART决策树的航空发动机小样本故障诊断[J]. 航空动力学报, 2020, 35(7): 1559-1568.
PANG Mengyang, SUO Zhongying, ZHENG Wanze, et al. Small sample fault diagnosis of aeroengine based on RS-CART decision tree[J]. Journal of Aerospace Power, 2020, 35(7): 1559-1568. (in Chinese)

(编辑: 张 雪)