# 一种新的模糊决策树算法
## ——基于加权毕达哥拉斯模糊熵

刘 帅[1],吴 涛[1,2],方 越[1],胡皓玮[1]
1. 安徽大学 数学科学学院,合肥 230031
2. 安徽大学 计算机智能与信号处理教育部重点实验室,合肥 230039

**摘 要**:传统的模糊决策树虽然可以从模糊数据中抽取模糊分类规则,但只能获取节点的隶属度信息,无法得出样本数据对于节点的非隶属度和犹豫度信息,导致数据分类的准确率不高。针对此,基于毕达哥拉斯模糊集理论,提出了一种新的加权毕达哥拉斯模糊决策树算法(Weighted Pythagorean Fuzzy Decision Tree,WPFDT)。首先,通过改进的 K-means 聚类算法得到连续属性数据的聚类中心,并结合三角模糊数对连续数据进行模糊处理;其次,定义并计算每一个属性的加权毕达哥拉斯模糊熵,选择加权毕达哥拉斯模糊熵最小的属性作为决策树根节点,在根节点下递归选择模糊熵最小的属性作为分裂节点,同时通过阈值控制树的规模,得到从根节点到叶子节点路径的模糊规则以及模糊规则的隶属度、非隶属度以及犹豫度,并完成预测分类,直至生成 WPFDT 模型;最后,选取 UCI 上的 3 个医学数据集(Haberman、Breast Cancer、Parkinson)进行实验,在分类准确率和得出模糊规则的数量与 3 种传统决策树算法(模糊 ID3 算法、C4.5 算法、CART 算法)比较,实验结果表明:WPFDT 在分类精度和树大小上都优于其他传统决策树算法,并且有较高的召回率和精确率。

**关键词**:加权毕达哥拉斯模糊熵;模糊决策树算法;数据分类;模糊规则
**中图分类号**:TP18 **文献标识码**:A doi:10.16055/j.issn.1672-058X.2023.0001.014

**A New Fuzzy Decision Tree Algorithm: Based on Weighted Pythagorean Fuzzy Entropy**

LIU Shuai[1], WU Tao[1,2], FANG Yue[1], HU Haowei[1]
1. School of Mathematical Sciences, Anhui University, Hefei 230031, China
2. Key Laboratory of Intelligent Computing & Signal Processing of Ministry of Education, Anhui University, Hefei 230039, China

**Abstract:** Although the traditional fuzzy decision tree can extract fuzzy classification rules from fuzzy data, it can only obtain the membership information of nodes. It cannot obtain the non-membership and hesitation information of sample data for nodes, resulting in low accuracy of data classification. In order to solve this problem, a new Weighted Pythagorean Fuzzy Decision Tree (WPFDT) algorithm was proposed based on the Pythagorean fuzzy set theory. Firstly, the cluster center of continuous attribute data was obtained by the improved K-means clustering algorithm, and the continuous data was fuzzily processed by combining with triangular fuzzy numbers. Secondly, the weighted Pythagorean fuzzy entropy of each attribute was defined and calculated. The attribute with the lowest weighted Pythagorean fuzzy

**收稿日期**:2022-01-08 **修回日期**:2022-05-18 **文章编号**:1672-058X(2023)01-0085-06
**基金项目**:国家自然科学基金(61806001);国家级大学生创新创业训练项目(202110357006);安徽大学研究生创新项目资助.
**作者简介**:刘帅(1995—),男,山西大同人,硕士研究生,从事智能计算和统计决策相关研究.
**通讯作者**:吴涛(1970—),男,安徽太和人,教授,博士,从事智能分析与决策、信息粒计算与智能计算研究. Email: Wutao@ahu.edu.cn.
**引用格式**:刘帅,吴涛,方越,等.一种新的模糊决策树算法——基于加权毕达哥拉斯模糊熵[J].重庆工商大学学报(自然科学版),2023,40(1):85—90.
LIU Shuai, WU Tao, FANG Yue, et al. A new fuzzy decision tree algorithm: based on weighted pythagorean fuzzy entropy[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2023, 40(1): 85—90.

entropy was selected as the decision-making root node, and the attribute with the lowest fuzzy entropy was recursively selected as the splitting node under the root node. At the same time, the size of the tree was controlled by the threshold. The fuzzy rules of the path from the root node to the leaf node as well as the membership degree, non-membership degree and hesitation degree of the fuzzy rules were obtained, and the prediction classification was completed until the WPFDT model was generated. Finally, three medical data sets (Haberman, Breast Cancer, and Parkinson) on UCI were selected for the experiment, and the classification accuracy and the number of fuzzy rules were compared with three traditional decision tree algorithms (fuzzy ID3 algorithm, C4.5 algorithm, and CART algorithm). The experimental results show that WPFDT is superior to other traditional decision tree algorithms in classification accuracy and tree size, and has higher recall rate and accuracy.

**Keywords:** weighted Pythagorean fuzzy entropy; fuzzy decision tree algorithm; data classification; fuzzy rules

## 1 引 言

随着决策树在数据挖掘中的普遍运用,其在图像识别、机器学习、数据分类等方面均取得了显著效果。传统决策树算法如ID3算法和C4.5算法在处理不平衡数据时,分类效果不稳定,尤其在处理更复杂的模糊数据时不能实现更好的分类,所以将模糊理论和决策树相结合,模糊决策树(Fuzzy Decision Tree, FDT)被提出。

目前很多专家提出多种FDT算法, Wang等[1]提出基于模糊规则的模糊决策树算法,该模糊决策树的节点能够涉及多个特征模糊规则,证明其有更好的分类性能;翟俊海等[2]提出一种模糊粗糙决策树算法,结合了知识的粗糙度和数据的模糊性,该算法比模糊ID3算法有更高的分类精度;Zheng等[3]将隶属函数模型扩展到模糊随机森林中应用于风险识别和预测,结果表明该方法生成的决策树比经典决策树更准确;Wang[4]等提出决策树与模糊粗糙集中与属性约简融合的方法,该算法性能明显优于其他使用优势粗糙集的融合方法;Idris等[5]提出FID3算法,将模糊系统与ID3算法相结合,在乳腺癌数据集分类上有较高的准确率;Li等[6]将分类问题转化为模糊粒度空间来求解,将数据进行模糊粒度化,提出一种自适应全局随机聚类算法,认为选择扩展属性的标准是信息增益比,结果可知有较高的准确率和鲁棒性;Farna等[7]提出基于卡方值的多柔性模糊决策树,将传统卡方统计扩展到模糊卡方统计用于数据分类,实验结果表明该算法比传统卡方统计算法有更优的分类效果。

上述FDT算法把模糊粗糙领域概念与传统决策树结合构建模糊决策树,只是选择分裂属性的标准不同,每个样本属于节点的程度都是用隶属度来表示,但是无法获取样本数据对于节点非隶属度和犹豫度的信息,显然这些算法在实际中不能更好地全面获取数据信息,导致数据分类准确率不高。针对此,基于毕达哥拉斯模糊集(Pythagorean Fuzzy Set, PFS)理论,定义一种新的加权毕达哥拉斯模糊熵,并提出一种新的加权毕达哥拉斯模糊决策树算法(WPFDT),将PFS与FDT相结合,推导了WPFDT完整的构建过程,相较于传统FDT算法,WPFDT可以同时包含样本与节点之间的隶属度、非隶属度和犹豫度,可以更全面地描述节点中的模糊信息,从而提升数据分类准确性,该算法在处理具有模糊信息的实际问题中有更好的分类效果。

## 2 预备知识

在本节中,首先给出毕达哥拉斯模糊集的基本知识,然后介绍FDT的基本知识,最后介绍如何将连续数据转换成PFS的主要手段。

**定义1**[8] 设 $X$ 为论域,则称 $A=\{\langle x,\mu_A(x),\nu_A(x)\rangle \mid 0\leqslant\mu_A^2(x)+v_A^2(x)\leqslant 1, x\in X\}$,$\mu_A(x)$,$\nu_A(x)$,$\pi_A(x)\in[0,1]$ 为PFS,其中 $\mu_A(x)$,$\nu_A(x)$,$\pi_A(x)$ 分别是 $A$ 的隶属度、非隶属度、犹豫度。$A_i=\{\langle x,\mu_{A_i}(x),\nu_{A_i}(x)\rangle \mid x\in X\}$,$A_1\subset A_2(i=1,2)$,当且仅当 $\mu_{A_1}(x)\leqslant\mu_{A_2}(x)$,$\nu_{A_1}(x)\geqslant\nu_{A_2}(x)$ $(\forall x\in X)$。

**定义2**[9-10] 设 $\Omega_i\subseteq F(U)$ $(1\leqslant i\leqslant m)$ 是 $m$ 维的模糊子集,如果 $T$ 是FDT,则满足以下条件:

(1) $T$ 的节点属于 $F(U)$;

(2) 记 $F(U)$ 由所有子节点构成,若有非叶子节点 $N$,则存在 $i(1\leqslant i\leqslant m)$,满足 $\Gamma=\Omega_i\cap N$;

(3) $T$ 的每一个叶子节点对应一个或多个决策分类属性值。

其中,$\Omega_i$ 为分裂属性的分裂子集,$\Omega_i$ 中的元素是分裂属性中的取值。

**定义3**[10] 假设 $P_i(i=1,2,\cdots,n)$ 是属性 $A_i$ 的模糊分裂,且有 $f_s\in P_i(s=1,2,\cdots,l)$ 表示属性 $A_i$ 模糊分裂后的取值,则定义以下集合:

$$B\equiv P_{i1}\wedge P_{i2}\cdots\wedge P_{ik}(i=1,2,\cdots,N;k=1,2,\cdots,n)$$

$$w=\sum_{i=1}^{N}\prod_{r=1}^{k}P(f_s\mid o_{t,s}) \tag{1}$$

通过上面的定义,可以计算模糊频率:

$$P(A_{i1}\text{is}\,f_1\,|A_{i2}\text{is}\,f_2\wedge\cdots\wedge A_{ik}is\,f_k)=\frac{w(\,|A_{i1}\text{is}\,f_1\wedge\cdots\wedge A_{ik}\text{is}\,f_k)}{\sum_{f\in P_s}w(\,|A_{i1}\text{is}\,f_1\wedge\cdots\wedge A_{ik}\text{is}\,f_k)}\tag{2}$$

**定义 4**[11]　设 $A$ 是 $X$ 上的模糊集,$\{\mu_1,\cdots,\mu_n\}$,$\mu_i>\mu_{i+1}$ 为 $A$ 的隶属度,$A$ 的质量分配函数 $m_A$ 的概率分布为

$$\begin{aligned}m_A(F_i)&=\mu_i-\mu_{i+1}\\ m_A(F_1)&=1-\mu_2\end{aligned}\tag{3}$$

其中,$F_i=\{x\in\Omega\,|\mu(x)\geqslant\mu_i\}$,$i=1,\cdots,n$;集合 $F_1,\cdots,F_n$ 是 $m_A$ 的焦点元素。

**定义 5**[11-12]　设 $P$ 是有限论域 $X$ 上的概率分布,取值为$\{p_1,\cdots,p_n\}$,满足 $0\leqslant p_{i+1}\leqslant p_i$ 和 $\sum_{i=1}^{n}p_i=1$,则如果 $P$ 是模糊集 $A$ 的最小偏差分布时,当且仅当以下条件成立:

$$\begin{aligned}m_A(G_i)&=y_i-y_{i+1},i=1,\cdots,n-1\\ m_A(G_n)&=y_n,m_A(G_1)=1-y_2\end{aligned}$$

其中,

$$G_i=\{x\in\Omega\,|p(x)\geqslant p_i\}$$

$$y_i=|G_i|p_i+\sum_{j=i+1}^{n}(\,|G_j|-|G_{j-1}|)\,p_j\tag{4}$$

**定义 6**[11-12]　在 PFS 中,$p_{os}^{+}$ 表示最大支持的比例,$p_{os}^{-}$ 表示最大反对的比例,则有下面公式:

$$\mu=1-p_{os}^{-};\nu=1-p_{os}^{+}\tag{5}$$

由于 $\mu^2+\nu^2+\pi^2=1$ ,所以有:

$$\mu=1-p_{os}^{-},\nu=1-p_{os}^{+}$$

$$\pi=\sqrt{1-(1-p_{os}^{+})^2-(1-p_{os}^{+})^2}\tag{6}$$

## 3　一种新的加权毕达哥拉斯模糊熵

在文献[14]中,作者定义的某些毕达哥拉斯模糊熵没有考虑犹豫度,所以当模糊性和犹豫性都最大时取到最大值。当 $\mu=\nu$ 时,模糊度最大,这使得毕达哥拉斯熵值与模糊集定义的熵有矛盾,这是因为毕达哥拉斯模糊集的熵值受到模糊性和犹豫性的影响,当模糊性越大时,熵越大;犹豫性越大时,熵也越大。针对这种情况,对文献[14]定义做以下修改:即当 $\mu=\nu=0$ 时,熵值取最大,此时模糊性和犹豫性都最大,所以能使熵值最大。接下来在考虑犹豫性和模糊性的基础上,对犹豫度和模糊度赋予权重,定义新的加权 Pythagorean 模糊熵。

**定义 7**　$A$ 是 $X$ 上的 PFS,$A$ 的模糊度为 $h_A(x)=1-|\mu_A^{\,2}(x)-v_A^{\,2}(x)\,|$,其中 $h_A(x)\in[0,1]$。

**定义 8**　设 $A$ 和 $B$ 是 $X$ 上的 PFS,有以下条件:

(1) $A$ 为清晰集,等价于 $E(A)=0$;

(2) $\forall x_i\in X,\mu_A(x_i)=\nu_A(x_i)=0$,等价于 $E(A)=1$;

(3) $E(A)=E(A^c)$;

(4) $E(A)$ 是关于模糊度 $h_A(x)$ 的单调增函数,是关于犹豫度 $\pi_A(x)$ 的单调增函数。

若 $E$ 满足以上条件,称 $E$ 为 PFS 上的毕达哥拉斯模糊熵。

条件(4)的说明:当 $\pi_A(x)=\pi_B(x)$ 且 $h_A(x)\leqslant h_B(x)$ 时,$E(A)\leqslant E(B)$;当 $h_A(x)=h_B(x)$ 且 $\pi_A(x)\leqslant\pi_B(x)$ 时,$E(A)\leqslant E(B)$。

**定义 9**　设论域 $X=\{x_1,x_2,x_3,\cdots,x_n\}$,$A$ 是 $X$ 上的 PFS,定义新的加权 Pythagorean 模糊熵如下:

$$\begin{aligned}E(A)&=\frac{1}{n}\sum_{x\in X}\omega_1 f_A(x_i)+\omega_2\pi_A(x_i)=\\ &\frac{1}{n}\sum_{x\in X}w_1(1-|\mu_A^2(x)-\nu_A^2(x)\,|)+\\ &w_2\sqrt{1-\mu_A^2(x)-\nu_A^2(x)}\end{aligned}\tag{7}$$

其中,$\omega_1+\omega_2=1$,$0\leqslant\omega_1\leqslant1$,$0\leqslant\omega_2\leqslant1$。

**证明**

1) $E(A)=0\Leftrightarrow\forall x_i\in X,\omega_1 f_A(x_i)+\omega_2\pi_A(x_i)=0\Leftrightarrow$

$$\forall x_i\in X,\omega_1(1-|\mu_A^{\,2}(x)-v_A^{\,2}(x)\,|)+\omega_2\sqrt{1-\mu_A^{\,2}(x)-v_A^{\,2}(x)}=0$$

$0\leqslant|\mu_A^{\,2}(x)-v_A^{\,2}(x)\,|\leqslant1$ 且 $0\leqslant\mu_A^{\,2}(x)+v_A^{\,2}(x)\leqslant1\Leftrightarrow$ $\forall x_i\in X,|\mu_A^{\,2}(x)-v_A^{\,2}(x)\,|=1$ 且 $\mu_A^{\,2}(x)+v_A^{\,2}(x)=1\Leftrightarrow$ $\forall x_i\in X,u_A(x_i)=0,v_A(x_i)=1$ 且 $u_A(x_i)=1,v_A(x_i)=0$ $\Leftrightarrow A\in P(X)$。

2) $E(A)=1\Leftrightarrow\forall x_i\in X,\omega_1 f_A(x_i)+\omega_2\pi_A(x_i)=1\Leftrightarrow$

$$\forall x_i\in X,\omega_1(1-|\mu_A^{\,2}(x)-v_A^{\,2}(x)\,|)+\omega_2\sqrt{1-\mu_A^{\,2}(x)-v_A^{\,2}(x)}=1\Leftrightarrow$$

$$\forall x_i\in X,1-|\mu_A^{\,2}(x)-v_A^{\,2}(x)\,|=1$$

$$\sqrt{1-\mu_A^{\,2}(x)-v_A^{\,2}(x)}=1\Leftrightarrow\forall x_i\in X,|\mu_A^{\,2}(x)-v_A^{\,2}(x)\,|=0$$

$$\mu_A^{\,2}(x)+v_A^{\,2}(x)=0\Leftrightarrow\forall x_i\in X,u_A(x_i)=v_A(x_i)=0$$

3) $E(A)=\frac{1}{n}\sum_{x\in X}\omega_1(1-|\mu_A^{\,2}(x)-v_A^{\,2}(x)\,|)+$

$$\omega_2\sqrt{1-\mu_A^{\,2}(x)-v_A^{\,2}(x)}=E(A^C)$$

4) 令 $h_A(x)=1-|\mu_A^{\,2}(x)-v_A^{\,2}(x)\,|=a$,$\pi_A(x)=\sqrt{1-\mu_A^{\,2}(x)-v_A^{\,2}(x)}=b$,满足 $0\leqslant a\leqslant1,0\leqslant b\leqslant1$,$E(A)=\frac{1}{n}\sum_{x\in X}\omega_1 a+\omega_2 b$,记 $E_i(A)=\omega_1 a+\omega_2 b$,分别关于 $a,b$ 求偏导 $\frac{\partial E_i(A)}{\partial a}=\omega_1\geqslant0$,$\frac{\partial E_i(A)}{\partial b}=\omega_2\geqslant0$,则 $E(A)$ 是关于

$\pi_A(x)$ 的单调增函数,关于 $h_A(x)$ 的单调增函数。

新的加权毕达哥拉斯模糊熵对模糊性和犹豫性赋予权重,符合毕达哥拉斯模糊熵的条件,同时考虑了模糊度和犹豫度对新的加权毕达哥拉斯模糊熵所起的作用,所以更符合客观实际。

## 4 WPFDT 的生成过程

### 4.1 属性的模糊化

对于属性的模糊处理主要有两种,如果是清晰属性,则直接根据语言术语语义将隶属度标记为 0 或者 1;如果是连续属性,则首先需要通过使用改进的 K-means 算法[14]求出各属性的聚类中心,得出各属性中心点就是三角隶属度函数的参数,从而将属性划分为对应数量的模糊集,然后用相应的模糊语言术语来描述连续值属性,最后使用上述所得的三角隶属度函数把数据转换成模糊数据。

### 4.2 分裂属性的选择标准

以属性 $A$ 为例计算新的加权毕达哥拉斯模糊熵的过程如下:假设属性 $A$ 有 $l$ 个模糊子集 $A_s(s=1,\cdots,l)$,两个决策属性 $C^+$ 和 $C^-$,则由式(1)得:

$$C^+:w(C^+\wedge A_s),C^-:w(C^-\wedge A_s),s=1,\cdots,l$$

由式(2)计算每一个子节点 $A_s$ 的相对频率:

$$P(C^+|A_s)=\frac{w(C^+\wedge A_s)}{w(C^+\wedge A)},P(C^-|A_s)=\frac{w(C^-\wedge A_s)}{w(C^-\wedge A)}$$

接下来计算属性 $A$ 分别支持 $C^+,C^-$ 的最大可能性:由定义 4,将 $p(C^+|A_s),p(C^-|A_s)$ 分别按照降序排列得到集合 $\{p_1,\cdots,p_l\}$ 和 $\{p_1,\cdots,p_l\}$,由定义 5,将 $\{\mu_1,\cdots,\mu_l\}$ 按照降序前的顺序分别得到 $p_{os}^+(A_s),s=1,2,\cdots,l$ 和 $p_{os}^-(A_s),s=1,2,\cdots,l$;由式(6)计算 PFS 的隶属度、非隶属度和犹豫度即 $\{\mu(A_s),\upsilon(A_s),\pi(A_s)\}$,$s=1,2,\cdots,l$。每个 $A_s$ 由以下 PFS 描述:$\{A_s,\mu(A_s),\upsilon(A_s),\pi(A_s)\},s=1,2,\cdots,l$,其中 $\mu$ 表示属于 $C^+$ 类,$\upsilon$ 表示属于 $C^-$ 类,$\pi$ 表示犹豫信息,最后根据式(7)计算属性 $A$ 的加权毕达哥拉斯模糊熵 $E(A)$。

通过上述,计算所有条件属性的新的加权毕达哥拉斯模糊熵,选择熵最小的属性为决策树的根节点,递归地计算加权毕达哥拉斯模糊熵,从而选择分裂节点。

### 4.3 WPFDT 的剪枝处理

WPFDT 在生成过程中需要限制树的深度,如果生成的 WPFDT 深度太深,就会增加复杂的计算量,而且不一定有好的分类效果,因此要限制 WPFDT 的生长。在 FDT 中常用的限制树规模的方法有预剪枝、后剪枝等,都通过控制显著水平、真实度来控制树的规模,显著水平 $\alpha$ 和真实度 $\beta$ 的定义见参考文献[15]。对于 WPFDT 的每一个待分裂节点都可以知道其隶属度、非隶属度,通过比较 $\mu_A(x)$、$\nu_A(x)$ 与设定阈值的大小来决定是否对节点进行分割;设定的阈值为 0.96(一般 $\beta_0>0.75$)[9],当 $\mu_A(x)$、$\nu_A(x)$ 有一个值大于阈值时,则停止对该节点进行分割,当 $\mu_A(x)$、$\nu_A(x)$ 都小于阈值时,则继续对该节点进行分割。

### 4.4 模糊规则抽取和分类预测

生成的 WPFDT 转换成模糊规则的过程如下:如果有 $n$ 个叶子节点就会有 $n$ 条模糊分类规则,$n$ 个模糊规则,WPFDT 在抽取规则时不仅会得出叶子节点的分类结果,也会得出属于不同决策属性的 $\mu$、$\nu$,同时也会得出模糊分类规则的犹豫度。WPFDT 在预测过程中会有多个模糊规则可以适合匹配,选择所有隶属度中最大值的模糊规则结果作为预测结果,并得出规则的 $\mu$、$\nu$、$\pi$。

## 5 WPFDT 算法流程

**步骤 1** 输入训练集,运用改进的 K-means 算法[14]和三角隶属度函数把数据转换成模糊数据。

**步骤 2** 对于每一个 $A_i$ 的每一个模糊子集值 $A_{ij}$ $(1\leqslant i\leqslant n;1\leqslant j\leqslant k_i)$,根据式(1)和式(2)计算 $A_{ij}$ 相对于 $C^+$ 或 $C^-$ 的模糊频率 $p_{ij}$。

**步骤 3** 根据式(4)和式(5)得到属性 $A_{ij}$ 分别支持 $C^+$ 和 $C^-$ 的最大可能性 $p_{os}^+(A_{ij})$ 和 $p_{os}^-(A_{ij})$。

**步骤 4** 根据式(6)得每个子节点 $A_{ij}$ 由以下毕达哥拉斯模糊集描述:$\{A_{ij},\mu(A_{ij}),\upsilon(A_{ij}),\pi(A_{ij})\}$。

**步骤 5** 根据式(7)计算每一个属性 $A_i$ 的加权毕达哥拉斯模糊熵 $E(A_i)$,选取最小的 $E(A_i)$ 作为根节点。

**步骤 6** 设定阈值 $\beta_0$,当每个子节点 $\mu(A_{ij})$、$\nu(A_{ij})$ 都小于 $\beta_0$ 时,继续对该节点 $A_i$ 划分,否则标记成为叶子节点。

**步骤 7** 在根节点下递归选取最小的加权毕达哥拉斯模糊熵对应的属性作为分裂节点,直至最终生成 WPFDT 模型。

## 6 WPFDT 实例分析

假设有 $A$ 和 $B$ 两个条件属性,决策属性为 $C$,其中 $A$ 和 $B$ 分别有两个条件属性 $A_1,A_2$ 和 $B_1,B_2$,$C$ 有 $C^+$ 和 $C^-$ 两个决策属性,假设有 4 个样本隶属度数据如下:$\mu(A_1)=(1,1,0.59,0.01)$,$\mu(A_2)=(0,0,0.41,0.99)$,$\mu(B_1)=(1,0,1,0.88)$,$\mu(B_2)=(0,1,0,0.12)$,$\mu(C_1)=(0,1,0,1)$,$\mu(C_2)=(1,0,1,0)$。

下面进行 WPFDT 的构建:$A$ 有属性值 $A_1$ 和 $A_2$,由式(2)计算相对频率 $p(C^+ \mid A_1)=0.505$, $p(C^+ \mid A_2)=0.495$, $p(C^- \mid A_1)=0.795$, $p(C^- \mid A_2)=0.205$,根据式(3)和式(4)计算得 $\mu_2=2*0.495=0.99$, $\mu_1=1$, $P_{os}^+(A_1)=1$, $P_{os}^+(A_2)=0.99$,同理 $P_{os}^-(A_1)=1$, $P_{os}^-(A_2)=0.41$,由式(6)得 $\mu(A_1)=0$, $\nu(A_1)=0$, $\pi(A_1)=1$; $\mu(A_2)=0.59$, $\nu(A_2)=0.01$, $\pi(A_2)=0.807$。同理,按照上述步骤计算可得: $\mu(B_1)=0$, $\nu(B_1)=0.124$, $\pi(B_1)=0.992$; $\mu(B_2)=1$, $\nu(B_2)=0$, $\pi(B_2)=0$。在式(7)中,考虑犹豫度和模糊度的权重均为 $\frac{1}{2}$,即 $\omega_1=\omega_2=\frac{1}{2}$,此时式(7)变为 $E=\frac{1}{2}\sum_{i=1}^{2}\frac{\pi^2(x)+1-|\mu^2(x)-v^2(x)|}{2}$,则 $E(A)=0.749\ 95$, $E(B)=0.492\ 3$, $E(B)<E(A)$,所以选择属性 $B$ 作为决策树的根节点。取阈值 $\beta_0=0.96$,前面计算得到了 $B_1$、$B_2$ 的隶属度、非隶属度 $\mu(B_1)=0$, $\nu(B_1)=0.124$,显然 $\mu(B_1)$、$\nu(B_1)$ 都小于 $\beta_0=0.75$,所以继续对该节点 $B_1$ 划分, $\mu(B_2)=1$, $\nu(B_2)=0$,因为 $\mu(B_2)=1>0.75$,所以将 $B_2$ 标记成为叶子节点。

接下来在属性 $B_1$ 的条件下计算模糊子集 $A_1$、$A_2$ 的新的加权 Pythagorean 模糊熵。按照上面步骤,计算 $A_1$、$A_2$ 分别相对于 $B_1$ 的相对频率,用式(4)和式(5)计算属性 $A$ 在 $B_1$ 这个条件属性下支持 $C^+$的最大可能性。由式(7)可得 $\mu(A_1)=0$, $\nu(A_1)=0.98$, $\pi(A_1)=0.199$; $\mu(A_2)=0.59$, $\nu(A_2)=0$, $\pi(A_2)=0.807$,由于只有两个属性,所以直接确定叶子节点,最终生成的 WPFDT 如图 1 所示。

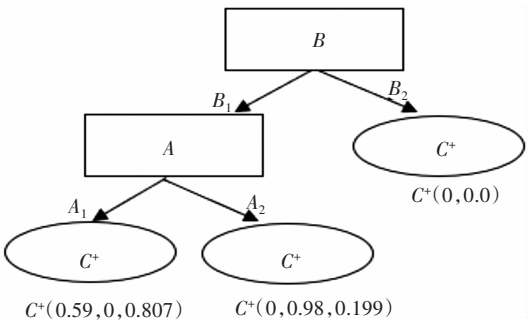


**图 1　加权毕达哥拉斯模糊决策树**

**Fig. 1　Weighted Pythagorean fuzzy decision tree**

由图 1 可知,规则 1:如果 $B$ 是 $B_1$ 并且 $A$ 是 $A_1$,则分类为 $C^+$,规则的 $\mu$、$\nu$、$\pi$ 分别为 0.59、0、0.807。规则 2:如果 $B$ 是 $B_1$ 并且 $A$ 是 $A_2$,则分类为 $C^-$,规则的 $\mu$、$\nu$、$\pi$ 分别为 0、0.98、0.199。规则 3:如果 $B$ 是 $B_2$,则分类为 $C^+$,规则的 $\mu$、$\nu$、$\pi$ 分别为 1、1、0。

接下来用 $\mu(A_1)=0.37$, $\mu(A_2)=0.63$, $\mu(B_1)=1$, $\mu(B_2)=0$ 进行预测。与规则 1 适合的隶属度 $\min\{1.00,0.63\}=0.63$,该样本属于 $C^+$类的可能性为 0.59,属于 $C^-$类的可能性为 0,犹豫度为 0.807。与规则 2 适合的隶属度为 $\min\{1.00,0.37\}=0.37$,该样本属于 $C^+$类的可能性为 0,属于 $C^-$类的可能性为 0.98,犹豫度为 0.199。与规则 3 适合的隶属度为 $\min\{0.00\}=0$,表示该样本属于 $C^+$类的可能性为 1,属于 $C^-$类的可能性为 0,犹豫度为 0。最终选择与所有规则适合最高的条件隶属度结果作为分类结果,即选择规则 1,分类为 $C^+$。

## 7　WPFDT 实验分析

### 7.1　分类准确率对比

为进一步说明加权毕达哥拉斯模糊决策树的优越性,选取 UCI 上的 Haberman、Breast Cancer、Parkinson 3 个医学数据集,将其与 3 种传统决策树算法(CART 算法、C4.5 算法、模糊 ID3 算法)进行实验比较,得到的分类准确率如表 1,并得出 WPFDT 算法 的准确率、精确率、召回率、$F_1$ 值如表 2 所示。

**表 1　分类准确率的对比**

**Table 1　Comparison of classification accuracy**

| 算　法 | Parkinson/% | Breast Cancer/% | Haberman/% |
|---|---|---|---|
| 模糊 ID3 | 80.86 | 82.47 | 83.14 |
| C4.5 | 78.55 | 73.25 | 74.53 |
| CART | 68.56 | 83.62 | 70.03 |
| WPFDT | 89.70 | 88.24 | 86.76 |

**表 2　WPFDT 算法评价指标**

**Table 2　WPFDT algorithm evaluation index**

| 数据集 | 准确率/% | 精确率/% | 召回率/% | $F_1$ 值 |
|---|---|---|---|---|
| Parkinson | 89.70 | 93.10 | 90.00 | 91.52 |
| Breast Cancer | 88.24 | 96.55 | 84.85 | 90.32 |
| Haberman | 86.76 | 87.50 | 90.32 | 88.88 |

由表 1 可知:本文的 WPFDT 算法在 3 个医学数据集上,分类准确率是最高的,表明对分裂节点过程中获取的模糊信息越全面,分类准确率越高。由表 2 知:WPFDT 算法在较高的分类准确率情况下,有较高的召回率和精确率。

### 7.2　规则数量比较

叶子节点的数量越多,抽取规则的数量就越多,在分类预测时匹配规则就越多,得出的模糊决策树规模就越大,分类的准确性也越高,但当规则数量过多的时候,不仅带来过多的复杂计算,而且容易出现过拟合。下面比较 4 种算法在 3 个医学数据集(Haberman、

Breast Cancer、Parkinson)上得出模糊规则的数量,得到的结果如表3所示。

**表3 IF THEN规则的数量**

**Table 3 Number of IF THEN rules**

| 数据集 | 模糊ID3 | C4.5 | CART | WPFDT |
|---|---|---|---|---|
| Parkinson | 16 | 15 | 7 | 13 |
| Breast Cancer | 13 | 14 | 8 | 10 |
| Haberman | 8 | 11 | 6 | 9 |

由表3可知:在Haberman、Breast Cancer、Parkinson 3个医学数据集上,本文的WPFDT算法得出的规则数更加适中,说明生成的模糊决策树更容易理解,用较合适的规则就能进行更好地分类预测,分类性能更好。

## 8 结论与展望

提出一种新的加权毕达哥拉斯模糊决策树算法(WPFDT)。首先,使用改进的K-means聚类算法得到连续属性聚类中心,并结合三角模糊数对连续数据进行模糊处理;其次,定义并计算每一个属性的加权毕达哥拉斯模糊熵,选择加权毕达哥拉斯模糊熵最小的属性作为决策树根节点,在根节点下递归选择模糊熵最小的属性作为分裂节点,直至生成WPFDT模型,同时通过阈值控制树的规模,得到从根节点到叶子节点路径的模糊规则以及模糊规则的$\mu$、$\nu$、$\pi$,并完成预测分类;最后,将UCI上的数据集与3种传统决策树算法进行实验比较,结果表明:WPFDT在分类精度和树大小方面都优于其他决策树。在进一步工作中,将结合卷积神经网络[16],模糊逻辑优化加权毕达哥拉斯模糊熵,生成具有更优分类性能的模糊决策树。

**参考文献**(References):


[1] WANG X C, LIU X D. Fuzzy rule based decision trees[J]. Pattern Recognition, 2015, 48(1): 50—59.

[2] 翟俊海,侯少星,王熙照.粗糙模糊决策树归纳算法[J].南京大学学报(自然科学版),2016,52(2):306—312.
ZHAI Jun-hai, HOU Shao-xing, WANG Xi-zhao. Rough fuzzy decision tree induction algorithm [J]. Journal of Nanjing University (Natural Science Edition), 2016, 52(2): 306—312.

[3] ZHENG H, HE J, ZHANG Y C, et al. A general model for fuzzy decision tree and fuzzy random forest[J]. Computational Intelligence, 2019, 35(2): 310—335.

[4] WANG J, QIAN Y, LI F, et al. Fusing fuzzy monotonic decision trees [J]. IEEE Transactions on Fuzzy Systems, 2020, 28(5): 887—900.

[5] IDRIS N F, ISMAIL M A. Breast cancer disease classification using fuzzy-ID3 algorithm with FUZZYDBD method: automatic fuzzy database definition[J]. PeerJ Computer Science, 2021, 7(2): 427—448.

[6] LI W, MA X Y, CHEN Y M, et al. Random fuzzy granular decision tree[J]. Mathematical Problems in Engineering, 2021 (10): 1—17.

[7] FARNAZ M, MARYAM M, SEYYED M R, et al. Chi-MFlexDT: Chi-square-based multi flexible fuzzy decision tree for data stream classification [J]. Applied Soft Computing, 2021, 105(7): 107—118.

[8] YAGER R R, ABBASOV A M. Pythagorean membership grades, complex numbers and decision making [J]. International Jounal of Intelligent Systems, 2013, 28(5): 436—452.

[9] 王熙照.模糊示例学习研究[D].哈尔滨:哈尔滨工业大学,1998.
WANG XI-zhao. Study on the fuzzy learning from examples[D]. Harbin: Harbin Institute of Technology, 1998.

[10] 李艳凤.基于直觉模糊集的决策树算法研究及应用[D].北京,北京交通大学,2019.
LI Yan-feng. Research and application of decision tree algorithm based on intuitionistic fuzzy sets [D]. Beijing: Beijing Jiaotong University, 2019.

[11] BALDWIN J F, LAWRY J, MARTIN T P. The application of generalized fuzzy rules to machine learning and automate knowledge discovery [J]. International Journal of Uncertainty, Fuzzyness and Knowledge-Based Systems, 1998, 6(5): 459—487.

[12] SZMIDT E, BALDWIN J F. Intuitionistic fuzzy set functions, mass assignment theory, possibility theory and histograms [C]//IEEE International Conference on Fuzzy Systems, Vancouver. CNADA: Institute of Electrical and Electronics Engineers ZnC. 2006: 35—41.

[13] PENG X D, YUAN H Y, YANG Y. Pythagorean fuzzy information measures and their applictions [J]. International Journal of Intelligent Sytems, 2017, 32(10): 991—1029.

[14] YU S S, CHU S W, WANG C M, et al. Two improved k-means algorithms [J]. Applied Soft Computing, 2017, 68: 747—755.

[15] ZHAI T H. Fuzzy decision tree based on fuzzy-rough technique [J]. Soft Computing, A Fusion of Foundations, Methodologies and Applications, 2011, 15(6): 1087—1096.

[16] 吴宇雳,李渊强.基于卷积神经网络的病理细胞核分割[J].重庆工商大学学报(自然科学版),2019,36(3):67—71.
WU Yu-li, LI Yuan-qiang. Convolutional network based pathological nucleus segmentation [J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2019, 36(3): 67—71.


**责任编辑:李翠薇**