Hans Journal of Data Mining 数据挖掘, 2026, 16(2), 34-47
Published Online April 2026 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2026.162004

Hans汉斯

基于混合属性模糊相似度的k-近邻分类器

秦少芬，曹梦雪，陈继强*

河北工程大学数理科学与工程学院，河北 邯郸

收稿日期：2026年3月7日；录用日期：2026年3月30日；发布日期：2026年4月8日

摘 要

在医学诊断等实际应用中，广泛存在着数值型、区间型与分类型属性共存的混合数据分类问题。现有方法往往难以充分融合与利用此类异构数据的原始信息，导致分类器性能不佳，无法满足实际应用中对精度与稳健性的要求。为此，文章提出一种基于混合属性模糊相似度的分类器。首先，针对混合数据结构，构建适配属性的模糊相似度；进而基于乘积t-范数，建立一种能够统一处理多类属性的模糊相似度度量。其次，在此基础上设计混合属性模糊相似度分类器，以更有效地利用数据的内在结构与语义信息进行分类。最后，为验证所提分类方法的有效性，将其与最大正区域分类器、线性支持向量机、多层感知机等5种代表性分类器进行对比实验。结果验证了新方法在多个数据集上的优越性能，为混合数据分类问题提供了一种有效的新途径。

关键词

混合数据，模糊相似度，分类器

k-NN Classifier Based on Hybrid-Attribute Fuzzy Similarity

Shaofen Qin, Mengxue Cao, Jiqiang Chen*

School of Mathematics and Physics, Hebei University of Engineering, Handan Hebei

Received: March 7, 2026; accepted: March 30, 2026; published: April 8, 2026

Abstract

In practical applications such as medical diagnosis, there exists a widespread problem of classifying mixed data with coexisting numerical, interval, and categorical attributes. Existing methods often fail to fully fuse and utilize the original information of such heterogeneous data, leading to limited

*通讯作者。

文章引用: 秦少芬, 曹梦雪, 陈继强. 基于混合属性模糊相似度的 k-近邻分类器[J]. 数据挖掘, 2026, 16(2): 34-47.
DOI: 10.12677/hjdm.2026.162004

classification performance, and thus cannot meet the requirements for accuracy and robustness in practical applications. To this end, this paper proposes a classifier based on fuzzy similarity for mixed-attribute data. First, aiming at the mixed data structure, a fuzzy similarity adapted to the attributes is constructed, and then a fuzzy similarity measure capable of uniformly processing multiple types of attributes is built based on the product t-norm. In practical applications such as medical diagnosis, there exists a widespread problem of classifying mixed data with coexisting numerical, interval, and categorical attributes. Existing methods often fail to fully fuse and utilize the original information of such heterogeneous data, which leads to limited classification performance and thus cannot meet the requirements for accuracy and robustness in practical applications. Second, on this basis, a fuzzy similarity classifier for mixed attributes is designed to more effectively utilize the inherent structure and semantic information of the data for classification. Finally, to verify the effectiveness of the proposed classification method, it is compared with five representative classifiers, such as the novel classifier based on maximal positive region, linear support vector machine, and multi-layer perceptron. The results verify the superior performance of the new method on multiple datasets, providing an effective new approach for the mixed data classification problem.

Keywords

Mixed Data, Fuzzy Similarity, Classifier

1. 引言

分类作为机器学习与数据挖掘领域的基础性核心任务，在医疗诊断[1]-[5]、金融风险预警[6]-[8]、工业故障检测[9]-[10]等诸多实际场景中发挥着关键支撑作用，其分类结果的优劣直接决定下游决策的科学性与可靠性。因此，研发适配复杂数据特征、性能优异的分类算法，始终是该领域的研究热点与核心诉求。

在实际应用场景中，数据集往往呈现出属性类型多样化的特征，普遍同时包含数值型属性(如身高、体重、浓度等连续量化指标)、区间型属性(如年龄 20~30 岁、收入 5~8 k 等范围化指标)与分类型属性(如性别、职业、类别标签等离散语义指标)。不同类型属性的特征表达逻辑、数据分布规律存在本质差异：数值型属性的核心特征是量化大小差异，分类型属性的核心特征是类别语义一致性，区间型属性则需兼顾范围边界特征与样本落点的相对位置关系。这种属性异质性对相似度度量提出了双重要求：一是需针对不同属性类型设计适配性度量方法，二是需实现多维度相似度的科学整合，以精准反映样本间的整体关联程度。

在分类问题中，对于分类型数据，早期依赖于手工编码(如有序编码[11]、独热编码[12])后使用汉明距离[13]；随着可学习嵌入[14]与梯度提升树框架的内置处理[15]出现，距离度量逐渐转向能够捕捉语义关系的低维向量表示。对于区间型数据，早期主要采用中心－半径变换[16]、端点距离[17]或 Hausdorff 距离[18]将其映射为数值；后续研究发展出基于区间分布的整体建模方法(如区间专用核函数[19])，以更直接地表达区间所承载的不确定性。对于数值型数据，则经历了从经典欧氏距离[20]、适应相关性的马氏距离[21]，到深度学习时代通过对比损失与度量学习实现任务自适应的可学习距离[22]。

然而，现有分类器在处理包含数值型、区间型以及分类型属性的混合数据时仍存在明显局限。一方面，多数策略依赖将非数值属性(如分类型、区间型)强制转换为数值形式(如独热编码、中心－半径变换)，

这一过程不仅可能引入冗余维度或信息失真，更易破坏属性原有的语义结构与不确定性特征；另一方面，现有方法在整合多源相似度时，通常采用线性加权或基于各属性子空间相似度最小值的聚合策略[23]，这类融合机制难以捕捉不同属性空间之间相似性判断的非线性交互关系，且其参数配置或策略选择多依赖经验设定或后验优化，缺乏坚实的理论支撑。

针对上述问题，本文提出一种面向属性本征特性的模糊相似度建模范式。具体而言，避免对分类型与区间型属性进行数值化转换，而是分别在其原生空间中定义适配的距离度量：对分类型属性，采用归一化汉明距离以刻画类别一致性；对数值型属性，保留欧氏距离以反映连续量值差异。考虑到切片 Wasserstein 距离所采用的投影积分机制，能够从多个方向对分布的整体形态进行刻画，每一个投影方向均反映了原始分布在该视角下的累积分布展宽特性，而这一特性恰好构成区间数据不确定性的数学化表征。因此，对区间型属性，则引入切片 Wasserstein 距离以有效捕捉分布形态与支撑集的不确定性。在各属性类型距离函数的基础上，通过指数映射将各属性子空间的距离转化为相似度，并采用乘积 t-范数融合机制构建全局相似度。该设计不仅尊重各类属性的内在表达逻辑，其乘积形式亦隐含“所有属性均需高度一致才能判定样本相似”的强协同假设，从而在无需参数调优的前提下实现多源相似性的自适应耦合。所提方法为混合属性分类任务提供了一种结构清晰、可解释性强且计算高效的相似性度量框架。

本文结构如下：第 2 节给出了相关理论基础知识；第 3 节建立了面向混合数据的混合属性模糊相似度；第 4 节构建了面向混合数据的基于混合属性模糊相似度的改进 k-近邻分类器；第 5 节结合 UCI 数据集中的 13 个数据集，验证了所提分类方法的可行性和有效性；第 6 节为结论。

2. 理论基础知识

2.1. 模糊相似关系[24]

设U 为非空论域，令 A 为条件属性集，$B\subseteq A$ 为 A 的子集。$R\in F\left(U\times U\right)$ 是 U 上的模糊关系。对 $\forall \mathrm{x},y\in U$，如果 R 满足

1) 自反性：$R_B\left(x,x\right)=1$；
2) 对称性：$R_B\left(x,y\right)=R_B\left(y,x\right)$，

则称 R 是 U 上的一个模糊相似关系。

2.2. *t*-范数[25]

设 $T:\left[0,1\right]\times\left[0,1\right]\to\left[0,1\right]$，对任意 $x,y,z\in\left[0,1\right]$，如果 T 满足：

1) 交换律：$T\left(x,y\right)=T\left(y,x\right)$；
2) 结合律：$T\left(x,T\left(y,z\right)\right)=T\left(T\left(x,y\right),z\right)$；
3) 单调性：如果 $x_1\le x_2,y_1\le y_2$，则 $T\left(x_1,y_1\right)\le T\left(x_2,y_2\right)$；
4) 边界条件：$\forall x\in\left[0,1\right],T\left(x,1\right)=x$，

则称 T 为三角范数，简称为 *t*-范数。

常用的 *t*-范数主要有以下 4 种：

1) Mamdani 算子：$T_M\left(x,y\right)=\min\left\{x,y\right\}$ (最大的三角范数)；
2) 乘积算子：$T_P\left(x,y\right)=x\cdot y$；
3) Lukasiewicz t-范数：$T_L\left(x,y\right)=\max\left\{0,x+y-1\right\}$；
4) $T_{\cos}\left(a,b\right)=\max\left\{ab-\sqrt{1-a^2}\sqrt{1-b^2},0\right\}$。

2.3. 离散测度下的 p-阶切片 W 距离[26]

设两个 d 维离散测度 μ,ν 的支撑集为点云 $X=\{x_1,x_2,\cdots,x_N\}\subset\mathbb{R}^d$，满足 $\sum_{i=1}^{N}\mu_i=1$；$Y=\{y_1,y_2,\cdots,y_M\}\subset\mathbb{R}^d$，满足 $\sum_{j=1}^{M}\nu_j=1$。$\mathbb{S}^{d-1}=\{\theta\in\mathbb{R}^d\mid\|\theta\|=1\}$ 为 d 维单位球面(所有投影方向集合)。从 $\mathbb{S}^{d-1}$ 均匀采样 L 个方向 $\theta_1,\theta_2,\cdots,\theta_L$ 后，各方向下 1D 投影点云的 p 阶 Wasserstein 距离 p 次幂的平均 $\frac{1}{p}$ 次方为

$$\widehat{SW}_P(\mu,\nu)=\left(\frac{1}{L}\sum_{k=1}^{L}W_p^p\left(\mu_{\theta_k},\nu_{\theta_k}\right)\right)^{\frac{1}{p}} \tag{1}$$

其中，$\mu_{\theta_k},\nu_{\theta_k}$ 代表 μ,ν 沿方向 θ_k 的 1D 投影点云，投影值为 $\langle x_i,\theta_k\rangle,\langle y_j,\theta_k\rangle$；$W_p^p\left(\mu_{\theta_k},\nu_{\theta_k}\right)$ 代表 1D 投影点云的最优传输成本。求解过程为先对 $\mu_{\theta_k},\nu_{\theta_k}$ 升序排序，再计算累计权重得到最优传输计划，最后计算运输成本，如算法 1 所示[27]。

算法 1. 基于 1D DOT 问题的离散测度切片 Wasserstein 距离计算伪代码

输入：

$X\in\mathbb{R}^{N\times d}$：离散测度 μ 的支撑集($\sum_{i=1}^{N}\mu_i=1$)；

$Y\in\mathbb{R}^{M\times d}$：离散测度 ν 的支撑集($\sum_{j=1}^{M}\nu_j=1$)；

L：投影方向采样数，p：Wasserstein 距离的阶数(默认 $p=2$)。

输出：

SW：切片 Wasserstein 距离估计值。

1. 从 d 维单位球面均匀采样 L 个方向：$\Theta=[\theta_1,\theta_2,\cdots,\theta_L]\in\mathbb{R}^{d\times L}$，其中 $\theta_k\in\mathbb{S}^{d-1}$ 且 $\|\theta_k\|=1$。

2. For $k=1$ 到 L 做：

2.1. 1D 投影：计算点云沿 θ_k 的 1D 投影：$X_{\theta_k}=X\cdot\theta_k$，第 i 个元素为 $\langle x_i,\theta_k\rangle$；$Y_{\theta_k}=Y\cdot\theta_k$，第 j 个元素为 $\langle y_i,\theta_k\rangle$。

2.2. 1D 离散最优传输问题求解：

a) 排序：对 $X_{\theta_k},Y_{\theta_k}$ 按投影值升序排序，记录排列索引 σ_X,σ_Y，满足 $X_{\sigma_X(1)}\le X_{\sigma_X(2)}\le\cdots\le X_{\sigma_X(N)}$，$Y_{\sigma_Y(1)}\le Y_{\sigma_Y(2)}\le\cdots\le Y_{\sigma_Y(M)}$；排序后投影点 $\hat{X}_i=X_{\sigma_X(i)},\hat{Y}_j=Y_{\sigma_Y(j)}$；排序后的权重 $\hat{\mu}_i=\mu_{\sigma_X(i)},\hat{\nu}_j=\nu_{\sigma_Y(j)}$。

b) 累计权重计算：初始化 $s_0=0,h_0=0$。

对 $i=1,2,\cdots,N$，执行：$s_i=s_{i-1}+\hat{\mu}_i(i)$；

对 $j=1,2,\cdots,M$，执行：$h_j=h_{j-1}+\hat{\nu}_j(j)$；

c) 求解最优传输计划 $\hat{\gamma}_{i,j}$：

初始化运输计划矩阵 $\hat{\gamma}\in\mathbb{R}^{N\times M}$，所有元素都是 0；

令 $i=1,j=1$；

如果 $s_i\le h_{j-1}$ 或者 $h_j\le s_{i-1}$，则 $\hat{\gamma}_{i,j}=0$；

否则如果 $h_{j-1}\le s_{i-1}<s_i\le h_j$，则 $\hat{\gamma}_{i,j}=\hat{\mu}_i$；

否则如果 $s_{j-1}\le h_{j-1}<h_j<s_i$，则 $\hat{\gamma}_{i,j}=\hat{\nu}_j$；

否则如果 $s_{i-1}\le h_{j-1}<s_i\le h_j$，则 $\hat{\gamma}_{i,j}=s_i-h_{j-1}$，

否则如果 $h_{j-1}\le s_{i-1}<h_j\le s_i$，则 $\hat{\gamma}_{i,j}=h_j-s_{i-1}$。

d) 还原运输计划到原索引：$\hat{\gamma}_{i,j}=\hat{\gamma}_{\sigma_X(i),\sigma_Y(j)}$。

e) 计算 1D 离散最优传输成本：$\cos t=\sum_{i=1}^{N}\sum_{j=1}^{M}\gamma_{ij}\cdot\left(X_i-Y_j\right)^p$。

End for

3. 计算切片 Wasserstein 距离估计值：$\widehat{SW}_P=\left(\frac{\cos t}{L}\right)^{\frac{1}{p}}$。

4. 返回 $\widehat{SW}_P$。

3. 混合属性模糊相似度

传统模糊相似度往往采用相同的范式定义不同属性下两个样本间的模糊相似度[28]，没有利用不同类型属性本身包含的信息，一定程度丢失了原始混合数据的信息。鉴于此，本文基于样本属性本身的数据信息，给出了不同类型属性的距离定义以及适配混合属性样本间相似度的定义，创新性地利用乘积 t-范数提出了混合属性模糊相似度，以充分利用原始混合数据的信息提高混合数据的可区分性。

定义 1 [29] 设 $\mathrm{IS}=(U,AT,V,f)$ 为信息系统(Information System, IS)，其中 $U=\{x_1,x_2,\cdots,x_n\}$ 为非空有限对象集(即论域)，$AT=\{a_1,a_2,\cdots,a_m\}$ 为非空有限属性集，$V=\bigcup_{a\in AT}V_a$，V_a 是属性 a 的值域。$f:U\times AT\to V$ 是信息函数，$f(x,a)$ 表示对象 x 在属性 a 上的值。对每个 $x\in U$，$a\in AT$，$f(x,a)\in V_a$。

定义 2 设 $U=\{x_1,x_2,\cdots,x_n\}$ 为非空有限对象集(即论域)，$D=\{d_1,d_2,\cdots,d_r\}$ 为非空有限决策属性集，$A=A^n\cup A^i\cup A^c$ 为信息系统的非空有限属性集，其中 A^n 为数值型属性集，A^i 为区间型属性集，A^c 为分类型属性集。称 $\mathrm{MIS}=(U,AT=A\cup D,V,f)$ 为混合信息系统。对 $\forall x\in U$，$a\in A$，属性值记为 $a(x)$。

混合信息系统中的非空有限属性集为 $A=\{a_1,a_2,\cdots,a_m\}$，设 $B\subseteq A$，属性集 B 中包含 j 个样本，将这 j 个样本按数值型属性、区间型属性和分类型属性排列并重新排序，得到 $B=\{a_1,a_2,\cdots,a_j\},j\le m$。假设 B 属性集中包含 r 个数值型属性 $0\le r\le j$、s 个区间型属性 $0\le s\le j$、t 个分类型属性 $0\le t\le j$，$r+s+t=j\le m$，则在混合信息系统 $\mathrm{MIS}=(U,AT=A\cup D,V,f)$，$B\subseteq A$，$B=B_r^n\cup B_s^i\cup B_t^c$，任意 $x,y\in U$ 在属性集 B 下的混合属性模糊相似度定义如下：

定义 3 设样本 x 在 r 个数值型属性下的值分别为 $x_{r1},x_{r2},\cdots,x_{rr}$，令 $X_r^n=(x_{r1},x_{r2},\cdots,x_{rr})$。样本 y 在 r 个数值型属性下的值分别为 $y_{r1},y_{r2},\cdots,y_{rr}$，令 $Y_r^n=(y_{r1},y_{r2},\cdots,y_{rr})$。任意 $x,y\in U$ 在属性集 B 数值型属性下的模糊相似度定义为：

$$\overline{\overline{R}}_B^n=\exp\left(-p\cdot d_n(x,y)\right) \tag{2}$$

$$d_n(x,y)=\left\|X_r^n-Y_r^n\right\|_2 \tag{3}$$

其中，$p>0$，控制模糊相似度的衰减速度。

定义 4 设样本 x 在 s 个区间型属性下的值分别为 $\left[x_{s1}^L,x_{s1}^R\right],\left[x_{s2}^L,x_{s2}^R\right],\cdots,\left[x_{ss}^L,x_{ss}^R\right]$，令 $X_s^{iL}=\left(x_{s1}^L,x_{s2}^L,\cdots,x_{ss}^L\right)$，$\mu_{\theta_1}=\left(\frac{1}{s},\frac{1}{s},\cdots,\frac{1}{s}\right)$，$X_s^{iR}=\left(x_{s1}^R,x_{s2}^R,\cdots,x_{ss}^R\right)$，$\mu_{\theta_2}=\left(\frac{1}{s},\frac{1}{s},\cdots,\frac{1}{s}\right)$。样本 y 在 s 个区间型属性下的值分别为 $\left[y_{s1}^L,y_{s1}^R\right],\left[y_{s2}^L,y_{s2}^R\right],\cdots,\left[y_{ss}^L,y_{ss}^R\right]$。令 $Y_s^{iL}=\left(y_{s1}^L,y_{s2}^L,\cdots,y_{ss}^L\right)$，$\nu_{\theta_1}=\left(\frac{1}{s},\frac{1}{s},\cdots,\frac{1}{s}\right)$，$Y_s^{iR}=\left(y_{s1}^R,y_{s2}^R,\cdots,y_{ss}^R\right)$，$\nu_{\theta_2}=\left(\frac{1}{s},\frac{1}{s},\cdots,\frac{1}{s}\right)$。任意 $x,y\in U$ 在属性集 B 区间型属性下的模糊相似度定义为：

$$\overline{\overline{R_B^i}} = \exp\left(-p \cdot d_i\left(x,y\right)\right) \tag{4}$$

$$d_i\left(x,y\right) = \widehat{SW_2}\left(\mu,\nu\right) = \left(\frac{1}{2}\sum_{k=1}^{2}\omega_p^p\left(\mu_{\theta_k},\nu_{\theta_k}\right)\right)^{\frac{1}{2}} \tag{5}$$

其中， $p>0$，控制模糊相似度的衰减速度。

定义 5 设样本 x 在 t 个分类型属性下的值分别为 $x_{t1},x_{t2},\cdots,x_{tt}$，令 $X_t^c=\left(x_{t1},x_{t2},\cdots,x_{tt}\right)$。样本 y 在 t 个分类型属性下的值分别为 $y_{t1},y_{t2},\cdots,y_{tt}$，令 $Y_t^c=\left(y_{t1},y_{t2},\cdots,y_{tt}\right)$。任意 $x,y\in U$ 在属性集 B 分类型属性下的模糊相似度定义为：

$$\overline{\overline{R_B^c}} = \exp\left(-p \cdot d_c\left(x,y\right)\right) \tag{6}$$

$$d_c\left(x,y\right) = \frac{H\left(X_t^c,Y_t^c\right)}{t} = \frac{\sum_{t=1}^{t}1_{[x_{th}\neq y_{th}]}}{t} \tag{7}$$

其中，$p>0$，控制模糊相似度的衰减速度。$1\le h\le t$，$1_{[\text{condition}]}$ 是指示函数，当条件为真时值为 1，否则为 0。

定义 6 设 $\text{MIS}=\left(U,AT=A\cup D,V,f\right)$，$B\subseteq A$，$B=B^n\cup B^i\cup B^c$。$\forall x,y\in U$，称

$$\overline{\overline{R_B}}\left(x,y\right) = T\left(\overline{\overline{R_B^n}},T\left(\overline{\overline{R_B^i}},\overline{\overline{R_B^c}}\right)\right) \tag{8}$$

其为 x,y 在属性集 B 下的混合属性模糊相似度，其中 T 为乘积 t-范数。

4. 基于混合属性模糊相似度的 k-近邻分类器

由于大部分分类算法(比如支持向量机[30])只能处理数值型数据，对于包含分类型属性和区间型属性的数据要进行数据转换之后才能进行分类，会损失原始数据的信息。因此，本节在保留分类型属性以及区间型属性原始信息的情况下提出了混合属性模糊相似度分类器(Hybrid Attribute Fuzzy Similarity Classifier, HFSC)。

基于混合属性模糊相似度(式 8)的定义，本节提出了混合属性模糊相似度的 k-近邻分类器，该分类器的伪代码如算法 2 所示。

算法 2. 混合属性模糊相似度分类器(HFSC)

输入：训练集 $Tr=\left\{\left(x_l,y_l\right)\middle|l=1,2,\cdots,n\right\}$ 和测试集 $Te=\left\{x_t\middle|t=1,2,\cdots,m\right\}$，参数 p；

输出：分类准确率。

1. 识别训练集和测试集中每一个属性的类型；
2. 将数值型属性值和区间型属性值进行归一化处理；
3. 对每一个 $x_t\in Te$，
4. 　　对每一个 $x_l\in Tr$，计算
5. 　　数值型属性模糊相似度 $\overline{\overline{R_B^n}}$
6. 　　区间型属性模糊相似度 $\overline{\overline{R_B^i}}$
7. 　　分类型属性模糊相似度 $\overline{\overline{R_B^c}}$
8. 　　混合属性模糊相似度 $\overline{\overline{R_B}}\left(x,y\right)$
9. 　　结束。
10. 　　对 $\overline{\overline{R_B}}\left(x_t,x_l\right)\left(l=1,2,\cdots,n\right)$ 进行排序，找到最大 k 个混合属性模糊相似度对应样本的类别及其决策类，即：$\left\{\left(x_l^i,y_l^i\right)\in Tr\middle|i=1,2,\cdots,k\right\}$，$y_t=\text{mode}\left\{y_l^i\middle|i=1,2,\cdots,k\right\}$，其中 mode 表示取众数。
11. 　　最后计算分类准确率。
12. 结束。

在算法 2 中输入的是数据集的训练集、测试集，输出的是分类准确率。首先，对数据集中的属性类型进行识别，区分数值属性、区间型属性和分类型属性。之后，针对测试集中的每一个样本，分别计算其与训练集中所有样本的混合属性模糊相似度，该相似度综合了不同类型属性的模糊匹配程度。其次，利用基于混合属性模糊相似度的 k-近邻分类规则进行分类：对于每个测试样本，从训练集中筛选出与其混合属性模糊相似度最高的 k 个样本，并根据这些样本的类别标签，通过众数投票确定测试样本的所属类别。最后，将所有测试样本的预测类别与真实类别进行比较，统计正确分类的样本比例，从而得到最终的分类准确率。

5. 实验

5.1. 数据集来源

为验证实验效果，选用来自 UCI 机器学习库(https://archive.ics.uci.edu/)中具有不同属性类型和不同样本数量的 13 个公开数据集进行实验，数据集描述见表 1。区间型数据是采用$\left[(1-\alpha)x_{a_k},(1+\alpha)x_{a_k}\right]$方式生成的，$\alpha\in[0,0.5]$ [31]。

Table 1. Dataset description
表 1. 数据集描述

NO	数据集	样本数	属性数 5	类别数	数据类型
1	Wine	178	13	3	数值型
2	Iris	150	4	3	数值型
3	Cancer	699	9	2	分类型
4	Sonar	208	60	2	数值型
5	Colon	62	2000	2	数值型
6	Yeast	1484	8	10	数值型
7	Wdbc	569	30	2	数值型
NO	数据集	样本数	属性数	类别数	数据类型
8	Pima	768	8	2	数值型
9	Diabetes Risk	520	16	2	分类型、数值型
10	Gall Stone	319	38	2	分类型、数值型
11	Ionosphere	351	33	2	分类型、数值型
12	ILPD	583	10	2	分类型、数值型、区间型
13	HESPE	145	31	8	分类型、区间型

5.2. 数据集归一化

由于大多分类算法基于距离度量进行分类决策，而原始特征(如年龄、收缩压等)具有不同的量纲与取值范围(如年龄通常为 30~80 岁，而收缩压可到 100~200 mmHg)，若直接使用原始数值计算距离，取值范围较大的特征将主导距离计算结果，导致模型对小尺度特征不敏感。为消除量纲差异对相似性度量的干扰，确保各数值型以及区间型特征在距离计算中具有可比性，本文对所有数值型及区间型变量采用如下方法进行尺度统一。

对每个数值型属性值 x_{ij} 进行最大－最小归一化(Min-Max Scaling)：

$$x'_{ij} = \frac{x_{ij} - a_i^{\min}}{a_i^{\max} - a_i^{\min}} \tag{9}$$

设数据集包含 n 个样本，第 i 个属性为区间型属性，记为 a_i。对第 j 个样本 $(j = 1,2,\cdots,n)$，该属性的取值为一个闭区间 $a_i^{(j)} = \left[a_i^{L,(j)}, a_i^{R,(j)}\right]$。首先分别计算第 i 个区间属性下所有样本左右端点的全局最小值 $\left(a_i^{L,\min}, a_i^{R,\min}\right)$ 和最大值 $\left(a_i^{L,\max}, a_i^{R,\max}\right)$，然后分别对第 j 个样本的左右端点进行最大－最小归一化[32]，最后归一化结果为 $a'_i = \left[a_i^{L'}, a_i^{R'}\right]$，其中：

$$a_i^{L'} = \frac{a_i^L - a_i^{L,\min}}{a_i^{L,\max} - a_i^{L,\min}}, a_i^{R'} = \frac{a_i^R - a_i^{R,\min}}{a_i^{R,\max} - a_i^{R,\min}} \tag{10}$$

5.3. 分类性能指标

混淆矩阵[33]是评估分类模型性能的核心工具之一，能够直观呈现模型对各类别样本的分类预测结果。对于二分类任务，其混淆矩阵的具体定义如表 2 所示。其中，真正类(TP)代表正类(少数类)样本被正确预测的数量，真负类(TN)代表负类(多数类)样本被正确预测的数量；假负类(FN)表示正类样本被错误预测为负类的数量，假正类(FP)表示负类样本被错误预测为正类的数量。

Table 2. Confusion matrix
表 2. 混淆矩阵

		预测类别	
		预测正类	预测负类
真实类别	真实正类	真正类(TP)	假负类(FN)
	真实负类	假正类(FP)	真负类(TN)

基于混淆矩阵所提供的分类预测细节，可进一步推导得到反映分类算法综合性能的关键评价指标，包括召回率(Recall)、准确率(Accuracy)、精确率(Precision)及 F1 分数(F1-Score)等。各指标对应的数学计算公式如表 3 所示：

本文选取 F1 分数(F1-Score)、受试者工作特征曲线下面积(Area under Curve, AUC)与准确率(Accuracy)作为分类性能的核心评价指标，三者从差异化维度刻画模型的分类效果：其中 F1 分数是精确率与召回率的调和平均值，可同时兼顾模型对正类样本的“预测可靠性”与“识别覆盖度”，在类别不平衡场景下能更全面地评估少数类的分类表现；AUC 指标基于受试者工作特征曲线(ROC 曲线)计算，其数值关联模型的真正率与假正率，可综合衡量模型对正负两类样本的分类区分能力；而准确率则表征模型正确分类的样本占总样本的比例，是直观反映模型整体分类正确性的基础指标。

Table 3. Confusion matrix formula
表 3. 混淆矩阵公式

度量	公式	直观含义
召回率	$\frac{TP}{TP+FN}$	实际的正例中，被正确预测的比例
准确率	$\frac{TP+TN}{TP+TN+FP+FN}$	预测正确的样本比例

续表

精确率	$\frac{TP}{TP+FP}$	预测为正例的样本中，实际为正例的比例
F1 分数	$\frac{2\times Recall\times Precision}{Recall+Precision}$	精确率和召回率的调和平均
AUC	ROC 曲线下的面积	随机正例得分高于随机负例的概率

5.4. 结果分析

为验证本文所构建的 HFSC 分类器的性能，将其与基于最大正区域的新型分类器(Novel Classifier Based on Maximal Positive Region, MPR) [34]、线性支持向量机(Linear Support Vector Machine, LSVM) [35]、多层感知机(Multi-Layer Perceptron, MLP) [36]、加权 k 近邻(Weighted k-Nearest Neighbor, WKNN) [37]以及半径 k 近邻(Radius k-Nearest Neighbor, RKNN) [38]这 5 种经典基准分类器进行对比实验，所有分类任务均采用统一超参数设置(K = 5)以保证对比公平性。实验运行环境为个人计算机，具体配置如下：操作系统为 64 位 Windows 10，处理器为 AMD Ryzen 5 3500U，内存容量为 10 GB。具体运算结果见表 4。

Table 4. Classification accuracy of each classifier on different datasets
表 4. 各分类器在不同数据集上的分类准确率

数据集	HFSC	MPR	LSVM	MLP	WKNN	RKNN
wine	0.9719 ± 0.0281	0.9549 ± 0.0354	0.9438 ± 0.0505	0.9722 ± 0.0373	0.9611 ± 0.0434	0.9386 ± 0.0631
iris	0.9533 ± 0.0670	0.9400 ± 0.0798	0.9467 ± 0.0499	0.9467 ± 0.0718	0.9533 ± 0.0521	0.8667 ± 0.0789
cancer	0.9671 ± 0.0158	0.9371 ± 0.0295	0.9642 ± 0.0072	0.9613 ± 0.0203	0.7868 ± 0.0549	0.7367 ± 0.0411
sonar	0.8081 ± 0.1046	0.7260 ± 0.1650	0.7788 ± 0.0809	0.8024 ± 0.1051	0.8319 ± 0.0866	0.7600 ± 0.0845
yeast	0.5728 ± 0.0435	0.4724 ± 0.0322	0.5693 ± 0.0428	0.5587 ± 0.0444	0.5862 ± 0.0441	0.5842 ± 0.0367
colon	0.7762 ± 0.1476	0.5190 ± 0.1387	0.7476 ± 0.1518	0.7595 ± 0.0990	0.7786 ± 0.1384	0.7738 ± 0.1514
Pima	0.7265 ± 0.0518	0.6874 ± 0.0509	0.7162 ± 0.0323	0.7227 ± 0.0634	0.7369 ± 0.0597	0.7161 ± 0.0429
WDBC	0.9648 ± 0.0263	0.9298 ± 0.0453	0.9613 ± 0.0246	0.9544 ± 0.0361	0.9666 ± 0.0199	0.8771 ± 0.0398
ionosphere	0.8548 ± 0.0643	0.7863 ± 0.0412	0.8490 ± 0.0480	0.8547 ± 0.0433	0.8462 ± 0.0529	0.6839 ± 0.0709
diabetes	0.8904 ± 0.0385	0.9288 ± 0.0315	0.8865 ± 0.0264	0.9096 ± 0.0385	0.9615 ± 0.0285	0.9442 ± 0.0338
gallstone	0.6398 ± 0.0952	0.5956 ± 0.0562	0.6265 ± 0.0823	0.5860 ± 0.0710	0.6743 ± 0.0889	0.6210 ± 0.0785
HESPE	0.2843 ± 0.1254	0.1571 ± 0.1031	0.2681 ± 0.1105	0.2490 ± 0.1365	0.2402 ± 0.1109	0.2644 ± 0.1531
ILPD	0.6995 ± 0.0755	0.6604 ± 0.0419	0.7136 ± 0.0072	0.7067 ± 0.0198	0.6778 ± 0.0884	0.7079 ± 0.0744
Average	0.7776 ± 0.0680	0.7150 ± 0.0654	0.7670 ± 0.0549	0.7680 ± 0.0605	0.7693 ± 0.0668	0.7288 ± 0.0730

表 4 和图 1 展示了所提出的 HFSC 分类器与 5 种对比模型(MPR、LSVM、MLP、WKNN、RKNN)在 13 个异质数据集上的分类准确率。由分类准确率结果可知，HFSC 以 0.7776 的平均准确率取得了最优的整体性能，明显优于其他分类器。具体而言，在 Iris、Cancer 等数据集上，HFSC 取得了最优分类性能，其准确率与标准差指标表明该方法兼具高判别能力与稳定性；并且在包含区间型属性的混合数据集 HESPE 上，其分类准确率也显著优于所有对比方法。总体来看，HFSC 在多数场景下具备显著竞争优势。

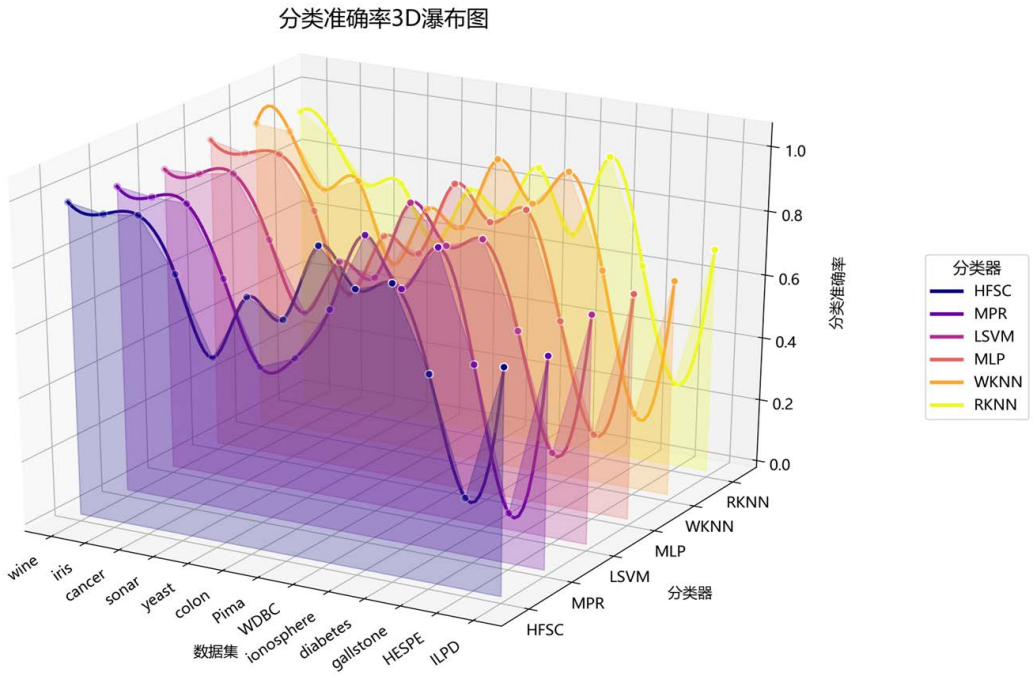
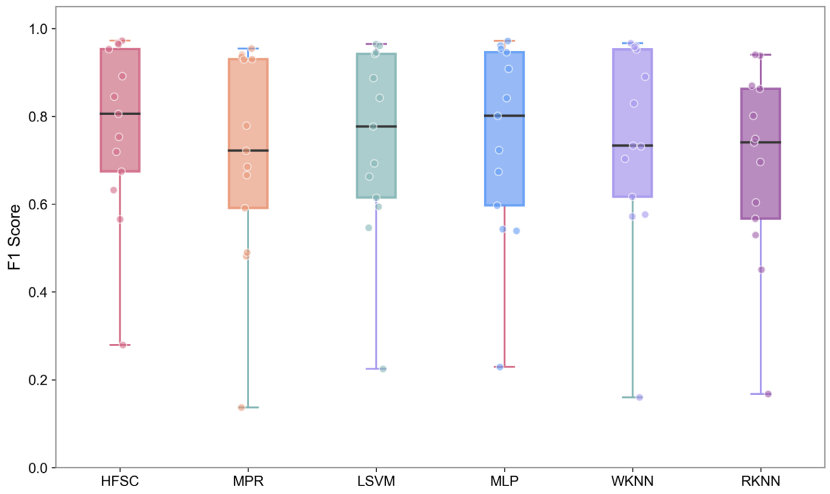


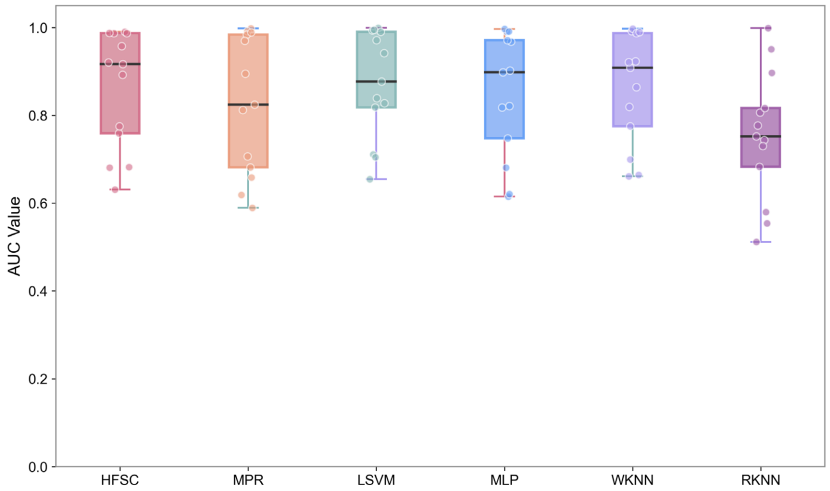
Figure 1. Classification accuracy diagram of each classifier on different datasets
图 1. 各分类器在不同数据集上的分类准确率图

Table 5. F1-score and AUC value of each classifier on different datasets
表 5. 各分类器在不同数据集上的 F1 分数和 AUC 值

数据集	HFSC		MPR		LSVM		MLP		WKNN		RKNN	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
wine	0.9724	0.9870	0.9546	0.9982	0.9422	0.9995	0.9717	0.9902	0.9605	0.9913	0.9381	0.9502
iris	0.9529	0.9902	0.9401	0.9913	0.9453	0.9900	0.9460	0.9967	0.9526	0.9973	0.8629	0.8970
cancer	0.9671	0.9878	0.9359	0.9841	0.9643	0.9936	0.9611	0.9911	0.7032	0.8640	0.6042	0.5797
sonar	0.8055	0.9167	0.7218	0.8122	0.7772	0.8179	0.8012	0.8982	0.8296	0.9229	0.7402	0.7771
yeast	0.5658	0.7592	0.4817	0.8241	0.5463	0.8388	0.5431	0.8180	0.5762	0.8195	0.5669	0.8061
colon	0.7532	0.8922	0.4900	0.6583	0.6930	0.9417	0.7230	0.8208	0.7335	0.9083	0.7486	0.7438
Pima	0.7191	0.7747	0.6854	0.7065	0.6627	0.8275	0.6737	0.7477	0.7315	0.7752	0.6963	0.7302
WDBC	0.9643	0.9870	0.9299	0.9695	0.9606	0.995	0.9537	0.9680	0.9664	0.9867	0.8699	0.8165
ionosphere	0.8447	0.9204	0.7790	0.8947	0.8420	0.8770	0.8417	0.9017	0.8905	0.9213	0.8011	0.5120
diabetes	0.8916	0.9574	0.9294	0.9891	0.8873	0.9709	0.9085	0.9708	0.9583	0.9896	0.9401	0.9987
gallstone	0.6322	0.6820	0.5911	0.6189	0.6149	0.7111	0.5391	0.6150	0.6166	0.6997	0.5299	0.6826
HESPE	0.2791	0.6806	0.1368	0.5895	0.2246	0.6543	0.2294	0.6204	0.1599	0.6609	0.1679	0.5539
ILPD	0.6747	0.6310	0.6660	0.6814	0.5943	0.7047	0.5966	0.6811	0.5717	0.6639	0.4508	0.7521
Average	0.7710	0.8589	0.7109	0.8244	0.7427	0.8709	0.7453	0.8477	0.7423	0.8616	0.6859	0.7538



(a)



(b)

Figure 2. F1-score and AUC diagram of each classifier on different datasets
图 2. 各分类器在不同数据集上的 F1 分数和 AUC 图

针对各分类器在 13 个数据集上的 F1 分数与 AUC 值表现(如表 5 及图 2 所示)，本文进行了综合对比分析。整体而言，本文提出的 HFSC 方法在 F1 分数上表现最优，平均值为 0.7710，优于其他对比方法(MPR: 0.7109, LSVM: 0.7427, MLP: 0.7453, WKNN: 0.7423, RKNN: 0.6859)，表明其在精确率与召回率的综合平衡上具有显著优势。在 AUC 指标上，HFSC 平均值为 0.8589，虽略低于 LSVM (0.8709)与 WKNN (0.8616)，但仍表现稳健，说明其具备良好的类别区分与排序能力。

5.5. 统计检验

为进一步分析 6 种分类器分类准确率的统计差异性，本节采用 Friedman 统计量和 Nemenyi 统计量[39]进行统计检验。这两种统计量分别为

$$\chi_F^2 = \frac{12N}{k(k+1)}\left(\sum_{i=1}^{k} r_i^2 - \frac{k(k+1)^2}{4}\right) \tag{11}$$

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2} \tag{12}$$

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{13}$$

其中 N 和 k 分别表示数据集和算法的个数，r_i 表示第 i 个算法在所有算法中的平均秩次排序，α 表示显著性水平，q_α 是给定 α 的临界值[39]。

基于 Friedman 检验，本研究拒绝了“所有算法性能相同”的原假设($\alpha = 0.05$)，进而采用 Nemenyi 检验进行后续两两比较。在 6 种算法、13 个数据集条件下，计算得到临界距离 CD = 2.0913。CD 图(图 3)显示，HFSC 分类器的平均排名低于其他分类器，表明其性能优于所有对比方法。

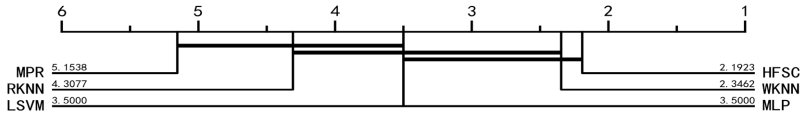


Figure 3. Nemenyi test results for six classifiers (significance level α = 0.05)
图 3. 6 种分类器的 Nemenyi 检验结果(显著性水平 α = 0.05)

5.6. 超参数敏感性分析

本文对 HFSC 分类器中的 k 值进行了敏感性分析，各数据集在不同 k 值下的分类准确率变化如图 4 所示。从图中可以观察到，在所有数据集上，k 为 1 至 3 的准确率曲线基本重合且处于较低位置，表明过小的 k 值会导致模型性能普遍偏低；随着 k 值增大到 4 至 8，各数据集的准确率显著提升，曲线整体上移并趋于平稳，形成性能高位平台区；而当 k 值继续增大至 9 至 10 时，部分数据集的准确率出现回落，曲线下移，表明过大的 k 值可能引入噪声或导致模型过度平滑。整体来看，k 为 4 至 8 是模型性能的稳定区域，其中 k = 5 在各数据集上的表现均处于该平台区内，既避免了过小 k 值带来的波动性，也规避了过大 k 值可能导致的性能衰退，展现出良好的鲁棒性与泛化能力。因此，本文最终选取 k = 5 作为实验参数。

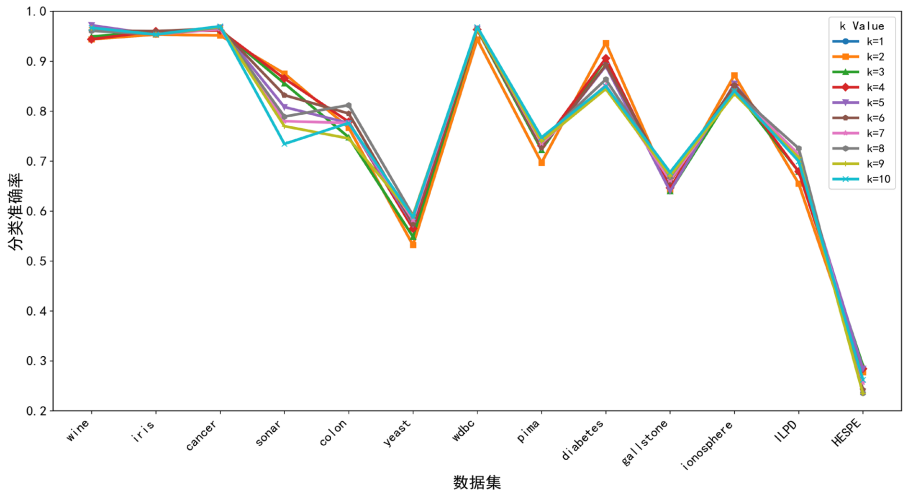


Figure 4. Parameter sensitivity analysis
图 4. 参数敏感性分析

6. 结论

针对包含数值型、区间型与分类型属性的混合数据分类问题，本文提出了一种基于混合属性模糊相似度的分类方法。该方法通过定义融合多类型属性的模糊相似度度量，在有效保留各类数据原始分布与语义信息的基础上，实现了对异构数据的统一相似性评估。基于此构建的混合属性模糊相似度分类器，克服了传统分类模型在处理混合数据时依赖离散化或单一类型假设的局限性。实验结果表明，所提出的HFSC 分类器在多个数据集上表现出稳定且具竞争力的分类性能。最后，对参数 k 进行了敏感性分析，得出 k 为 4 至 8 时是模型性能的稳定区域。

本研究的主要贡献在于提出了一种可解释的混合数据相似度度量框架，并在此基础上构建了高效且稳健的分类模型，为混合属性数据的分类问题提供了新的解决思路。未来工作将集中于相似度度量的自适应优化、面向流数据与半监督场景的拓展，以及与其他深度学习架构的融合研究。

基金项目

河北省中央引导地方科技发展资金项目(246Z1825G)。

参考文献

[1] Sekar, J. and Aruchamy, P. (2025) A Novel Approach for Heart Disease Prediction Using Hybridized AITH^{2}O Algorithm and SANFIS Classifier. *Network*: *Computation in Neural Systems*, **36**, 109-147. https://doi.org/10.1080/0954898x.2024.2404915

[2] Salem, H., Shams, M.Y., Elzeki, O.M., Abd Elfattah, M., F. Al-Amri, J. and Elnazer, S. (2022) Fine-Tuning Fuzzy KNN Classifier Based on Uncertainty Membership for the Medical Diagnosis of Diabetes. *Applied Sciences*, **12**, Article No. 950. https://doi.org/10.3390/app12030950

[3] Höglinger, G.U., Adler, C.H., Berg, D., Klein, C., Outeiro, T.F., Poewe, W., *et al*. (2024) A Biological Classification of Parkinson's Disease: The Synneurge Research Diagnostic Criteria. *The Lancet Neurology*, **23**, 191-204. https://doi.org/10.1016/s1474-4422(23)00404-0

[4] Khan, S.U.R., Bilal, O., Mistry, S., Deb, N., Mahmud, M. and Bhuyan, M. (2025) KDLight: A Lightweight Knowledge Distillation Framework for Medical Image Classification. 2025 *International Joint Conference on Neural Networks* (*IJCNN*), Rome, 30 June-5 July 2025, 1-8. https://doi.org/10.1109/ijcnn64981.2025.11228615

[5] Xing, W. and Bei, Y. (2020) Medical Health Big Data Classification Based on KNN Classification Algorithm. *IEEE Access*, **8**, 28808-28819. https://doi.org/10.1109/access.2019.2955754

[6] Guo, X. (2024) Research on Systemic Financial Risk Early Warning Based on Integrated Classification Algorithm. 2024 *IEEE 2nd International Conference on Electrical, Automation and Computer Engineering* (*ICEACE*), Changchun, 26-28 December 2024, 1586-1591. https://doi.org/10.1109/iceace63551.2024.10898790

[7] Tong, L. and Tong, G. (2022) A Novel Financial Risk Early Warning Strategy Based on Decision Tree Algorithm. *Scientific Programming*, **2022**, Article ID: 4648427. https://doi.org/10.1155/2022/4648427

[8] Hong, S., Wu, H., Xu, X. and Xiong, W. (2022) Early Warning of Enterprise Financial Risk Based on Decision Tree Algorithm. *Computational Intelligence and Neuroscience*, **2022**, Article ID: 9182099. https://doi.org/10.1155/2022/9182099

[9] Wu, H., Triebe, M.J. and Sutherland, J.W. (2023) A Transformer-Based Approach for Novel Fault Detection and Fault Classification/Diagnosis in Manufacturing: A Rotary System Application. *Journal of Manufacturing Systems*, **67**, 439-452. https://doi.org/10.1016/j.jmsy.2023.02.018

[10] Ragab, A., Ghezzaz, H. and Amazouz, M. (2022) Decision Fusion for Reliable Fault Classification in Energy-Intensive Process Industries. *Computers in Industry*, **138**, Article ID: 103640. https://doi.org/10.1016/j.compind.2022.103640

[11] Eye, A.V. and Clogg, C.C. (1996) Categorical Variables in Developmental Research: Methods of Analysis. Elsevier.

[12] Lantz, B. (2015) Machine Learning with R. Packt Publishing.

[13] 张焕炯, 王国胜, 钟义信. 基于汉明距离的文本相似度计算[J]. 计算机工程与应用, 2001(19): 21-22.

[14] Mumtaz, S. and Giese, M. (2021) Hierarchy-Based Semantic Embeddings for Single-Valued & Multi-Valued Categorical Variables. *Journal of Intelligent Information Systems*, **58**, 613-640. https://doi.org/10.1007/s10844-021-00693-2

[15] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. and Gulin, A. (2018) CatBoost: Unbiased Boosting with

Categorical Features. *NIPS*'18: *Proceedings of the* 32*nd International Conference on Neural Information Processing Systems*, Montréal, 3-8 December 2018, 6639-6649.

[16] Billard, L. and Le-Rademacher, J. (2012) Principal Component Analysis for Interval Data. *WIREs Computational Statistics*, **4**, 535-540. https://doi.org/10.1002/wics.1231

[17] Ishibuchi, H., Tanaka, H. and Okada, H. (1993) An Architecture of Neural Networks with Interval Weights and Its Application to Fuzzy Regression Analysis. *Fuzzy Sets and Systems*, **57**, 27-39. https://doi.org/10.1016/0165-0114(93)90118-2

[18] Guo, C. and Liu, Y. (2015) A Feature Selection Method for Symbolic Interval Data. *Operations Research and Management Science*, **24**, 67.

[19] Dai, J., Liu, Y., Chen, J. and Liu, X. (2020) Fast Feature Selection for Interval-Valued Data through Kernel Density Estimation Entropy. *International Journal of Machine Learning and Cybernetics*, **11**, 2607-2624. https://doi.org/10.1007/s13042-020-01131-5

[20] Alencar, G.T., Santos, R.C. and Neves, A. (2022) Euclidean Distance-Based Method for Fault Detection and Classification in Transmission Lines. *Journal of Control, Automation and Electrical Systems*, **33**, 1466-1476. https://doi.org/10.1007/s40313-022-00918-x

[21] Magyar, B., Kenyeres, A., Tóth, S., Hajdu, I. and Horváth, R. (2022) Spatial Outlier Detection on Discrete GNSS Velocity Fields Using Robust Mahalanobis-Distance-Based Unsupervised Classification. *GPS Solutions*, **26**, Article No. 145. https://doi.org/10.1007/s10291-022-01323-2

[22] Cai, B., Xiong, P. and Tian, S. (2023) Center Contrastive Loss for Metric Learning.

[23] Wang, C., Wang, C., Qian, Y. and Leng, Q. (2024) Feature Selection Based on Weighted Fuzzy Rough Sets. *IEEE Transactions on Fuzzy Systems*, **32**, 4027-4037. https://doi.org/10.1109/tfuzz.2024.3387571

[24] Dubois, D. and Prade, H. (1990) Rough Fuzzy Sets and Fuzzy Rough Sets. *International Journal of General Systems*, **17**, 191-209. https://doi.org/10.1080/03081079008935107

[25] 胡宝清. 模糊理论基础[M]. 第 2 版. 武汉: 武汉大学出版社, 2010.

[26] Santambrogio, F. (2015) Optimal Transport for Applied Mathematicians. Springer.

[27] Ma, L., Bian, W. and Xue, X. (2024) Point Clouds Matching Based on Discrete Optimal Transport. *IEEE Transactions on Image Processing*, **33**, 5650-5662. https://doi.org/10.1109/tip.2024.3459594

[28] Liang, P., Lei, D., Chin, K. and Hu, J. (2022) Feature Selection Based on Robust Fuzzy Rough Sets Using Kernel-Based Similarity and Relative Classification Uncertainty Measures. *Knowledge-Based Systems*, **255**, Article ID: 109795. https://doi.org/10.1016/j.knosys.2022.109795

[29] Pawlak, Z. (1982) Rough Sets. *International Journal of Computer & Information Sciences*, **11**, 341-356. https://doi.org/10.1007/bf01001956

[30] Khakbiz, M., Shahmoradi, M.R., Akhlaghi, F. and Soroush, K. (2025) AI-Enhanced Support Vector Machine Framework for Nanoparticle Size and Surface Nanotopography Analysis. *Particuology*, **106**, 156-173. https://doi.org/10.1016/j.partic.2025.07.017

[31] 余建航. 基于粗糙集的几类广义信息系统知识发现与决策方法研究[D]: [博士学位论文]. 哈尔滨: 哈尔滨工业大学, 2020.

[32] Liu, P., Munir, M., Mahmood, T. and Ullah, K. (2019) Some Similarity Measures for Interval-Valued Picture Fuzzy Sets and Their Applications in Decision Making. *Information*, **10**, Article No. 369. https://doi.org/10.3390/info10120369

[33] Yang, X., Huang, P., An, L., Feng, P., Wei, B., He, P., *et al*. (2022) A Growing Model-Based OCSVM for Abnormal Student Activity Detection from Daily Campus Consumption. *New Generation Computing*, **40**, 915-933. https://doi.org/10.1007/s00354-022-00193-z

[34] An, S., Zhao, E., Wang, C., Guo, G., Zhao, S. and Li, P. (2023) Relative Fuzzy Rough Approximations for Feature Selection and Classification. *IEEE Transactions on Cybernetics*, **53**, 2200-2210. https://doi.org/10.1109/tcyb.2021.3112674

[35] Pan, F., Wang, B., Hu, X. and Perrizo, W. (2004) Comprehensive Vertical Sample-Based KNN/LSVM Classification for Gene Expression Analysis. *Journal of Biomedical Informatics*, **37**, 240-248. https://doi.org/10.1016/j.jbi.2004.07.003

[36] Zhang, R., Wang, L., Cheng, S. and Song, S. (2023) Mlp-Based Classification of COVID-19 and Skin Diseases. *Expert Systems with Applications*, **228**, Article ID: 120389. https://doi.org/10.1016/j.eswa.2023.120389

[37] Tarakci, F. and Ozkan, I.A. (2021) Comparison of Classification Performance of kNN and WKNN Algorithms. *Selcuk University Journal of Engineering Sciences*, **20**, 32-37.

[38] 周鹏, 伊静, 朱振方, 等. 面向不平衡分类的固定半径最近邻逐步竞争算法(FRNNPC) [J]. 山东大学学报(理学版), 2019, 54(3): 102-109.

[39] Demšar, J. (2006) Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, **7**, 1-30.