备注：本文已被《哈尔滨商业大学学报(自然科学版)》2022年10月刊录用。

一种基于机器学习和 D-S 证据理论的饮用水质量预测方法

马晓剑，张家绪，林煜华，王奥

（东北林业大学 理学院，哈尔滨 150040）

摘　要：传统机器学习算法在饮用水质量预测问题中，由于特征空间高冲突，导致判别准确率低、判别结果不稳定，本文创新性地提出了适用于饮用水质量预测的 D-S 证据理论集成学习算法，该算法通过引入证据理论对传统机器学习算法的判别结果进行集成，将分类器的判别结果转换为基本概率指派，针对分类器判别结果产生冲突的问题，通过引入 BJS 散度定量描述基本概率指派之间的冲突，结合邓熵对分类器判别结果的支持度、置信度进行建模，利用加权重构的基本概率指派重新进行 Dempster 证据融合，再利用证据融合后的决策形成集成判别结果. 实验结果表明，本文算法相较于两种单一基分类器在准确率指标下分别提高 6.06%、1.58%，精确度指标下提高 82.65%、18.65%，表明了本文算法在饮用水质量预测问题中的有效性，具有一定的推广价值.

关键词：机器学习；集成学习；证据理论；BJS 散度；水体质量预测

中图分类号：TP391.41　　**文献标识码**：A

A novel classification algorithm of drinking water quality based on machine learning and D-S evidence theory

MA Xiao-jian, ZHANG Jia-xu, LIN Yu-hua, WANG Ao

(School of Science，Northeast Forestry University，Harbin 150040，China)

Abstract: In order to solve the problem of low accuracy and unstable classification results in the high conflicts of feature spaces in the drinking water quality classification problem, an ensemble learning algorithm utilizing D-S evidence theory for drinking water quality prediction is Innovatively proposed. The prediction results of machine learning classifiers are transferred into basic probability assignments, and the conflicts are quantitatively measured by BJS divergence, combined with Deng Entropy to model the support and confidence of the classifiers, then the ensemble prediction after evidence fusion is formed. Experiments show that compared with single base classifiers, the accuracy of the proposed algorithm is improved by 6.06% and 1.58% under the accuracy criterion, and 82.65% and 18.65% under the precision criterion, showing the effectiveness of proposed algorithm in the prediction of drinking water quality.

Key Words: Machine learning; Ensemble learning; Evidence theory; BJS divergence; Water quality prediction

引言

近年来，随着人类对生态文明建设的逐渐重视，自然水体资源是否适用于饮用也日渐受到广泛关注.如何利用简单高效的方法对饮用水资源质量进行分析与预测，是当下研究的重要课题.在水体质量预测与机器学习方法相结合的领域，李雪清等人选取多种气象指标和经济指标，提出了一种基于多源时空数据和机器学习的区域水质预测模型[1]，但由于该方法主要选取宏观评价指标进行预测建模，因此仅适用于时空意义下的水体环境质量预测.戴青松等人利用 LWCA-SVM 模型提出了一种基于机器学习和狼群搜索的饮用水质量预测模型[2]，但该方法由于引入了启发式智能优化，因此收敛于全局最优的速度较慢，算法效率仍有待提高.

D-S 证据理论是一种基于不确定理论的信息融合方法，该方法是贝叶斯理论的推广，常用于解决多源信息下的信息融合问题，是一种简单高效的人工智能决策技术，现已广泛应用于图像处理、计算机视觉、专家系统等领域[3-5].证据理论在证据高冲突情况下的判别精度不高，因此常常因受到噪声信号的干扰而丧失决策的可信度，如何改善证据理论在高冲突情况下的判别精确度是当前证据理论主要待解决的问题[6].在证据理论中，高冲突抑制方法主要分为两种，一种是修改证据融合规则，另一种是修改证据，但第一种方法

往往会失去 Dempster 组合规则具有的良好数学性质，而第二种方法往往通过考虑来自证据的可信度、信息量等信息对证据进行修正，具有更好的可解释性[7].在机器学习领域，证据理论可作为一种集成学习的方法改善机器学习性能，此时证据理论将多分类器给出的预测结果视为信号源，对信号源提供的信息量及可信度进行建模，可以改善多分类器架构下的机器学习分类与预测性能[8].

虽然利用 D-S 证据理论改善高冲突下机器学习判别精确率的方法已广泛应用于故障诊断、多时空数据融合、异常检测等邻域[9-11]，但现有的应用证据理论在饮用水质量预测问题中所做的工作仍然较少，而该问题下的特征空间信息冲突常常导致机器学习算法识别精确度不高，因此本文创新性地引入证据理论基于高冲突对机器学习算法进行集成，应用证据理论对多分类器预测结果进行优化，提出了基于线性核支持向量机、随机森林和证据理论的饮用水质量集成预测模型.针对特征空间高冲突从而导致分类器结果高冲突的问题，本文引入 BJS 散度[12]对来自分类器信息源的高冲突进行抑制，决策时综合考虑信息源的支持度、置信度，实验结果表明本文算法相较于单一机器学习模型具有更高的分类准确率和精确度，显著改善了基分类器的分类效果.

1 D-S 证据理论相关概念

定义 1 （基本概率指派（mass 函数））假设集合$I=\{i_1,i_2,\dots,i_n\}$是辨识框架，其中$i_s\neq i_t, s\neq t$是两两互异的元事件，则构造辨识框架的幂集2^I到实数区间$[0,1]$的映射$m:2^I\to[0,1]$，如果映射m满足下列条件：

$$\begin{cases} m(\emptyset)=0 \\ \sum_{A\subset 2^I} m(A)=1 \end{cases}$$

则称该映射m为基本概率指派 s.在证据理论中，A被称为焦元.此时$m(A)$被视为m对应的信息源分配给焦元A的信度.

定义 2 （Dempster 组合规则）假设$m_1,\dots,m_c$是辨识框架I下的c组基本概率指派，则 Dempster 组合规则表示如下：

$$\begin{cases} m(\emptyset)=0 \\ m(A)=\dfrac{1}{1-k}\sum_{\cap A_j=A}\prod_{1\le r\le c} m_r(A_j) \end{cases}$$

其中$A_1,\dots,\mathrm{A_s}$为$s=|2^I|$组焦元，$k=\sum_{\cap A_j=\emptyset}\prod_{1\le r\le c} m_r(A_j)$为冲突系数，是 Dempster 组合规则中的证据冲突度量.

定义 3 （BJS 散度[12]）假设A_j是基本概率指派m的焦元，m_1,m_2是辨识框架I下的两组基本概率指派，则m_1和m_2之间的 BJS 散度定义如下：

$$BJS(m_1,m_2)=\frac{1}{2}\left[S\left(m_1,\frac{m_1+m_2}{2}\right)+S\left(m_2,\frac{m_1+m_2}{2}\right)\right],$$

其中$S(m_1,m_2)=\sum_i m_1(A_i)log\frac{m_1(A_i)}{m_2(A_i)}$. 按照定义，BJS 散度还可以按照下式进行展开：

$$\begin{aligned} BJS(m_1,m_2) &= E\left(\frac{m_1+m_2}{2}\right)-\frac{1}{2}E(m_1)-\frac{1}{2}E(m_2) \\ &= \frac{1}{2}\left[\sum_i m_1(A_i)log\left(\frac{2m_1(A_i)}{m_1(A_i)+m_2(A_i)}\right)+\sum_i m_2(A_i)log\left(\frac{2m_2(A_i)}{m_1(A_i)+m_2(A_i)}\right)\right], \end{aligned}$$

其中$E(m_j)=-\sum_i m_j(A_i)\,log\,m_j(A_i)(i=1,2,\dots,n;j=1,2)$为来自基本概率指派$m_j$的 Shannon 熵，在实际计算时采用该式可简化计算.BJS 散度作为证据距离的度量，可用于定量估计基本概率指派之间的差异性.

2 基于机器学习和 D-S 证据理论的饮用水质量预测方法

在不同水体的饮用水质量预测问题中，基于机器学习方法的分类误差主要来源于不同可饮用水体的样本分布之间存在较大冲突，即可饮用水和非可饮用水在特征指标下的分布不存在显著差异，因此分类器在学习时因为受到冲突的干扰，难以学习到有助于精确分类的有效信息并形成具有高置信度的判别结果，给

准确识别可饮用水体样本造成了较大困难.本文创新性地引入证据理论对分类器的判别可信度进行建模，同时考虑来自分类器判别结果的信息量，应用证据理论和 BJS 散度对来自分类器信息源的判别结果进行信息融合，从而提升集成学习的效果，实验结果表明本文算法显著提高了算法的分类精确度.

2.1 探索性数据分析及证据理论的引入

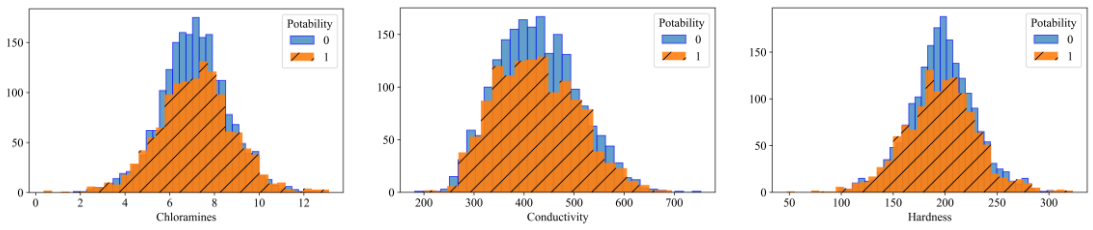
本文算法采用的数据集来自 Kaggle 数据平台的 Water Quality 数据集[13].该数据集采集了 3276 个来自不同水体的水质指标，用于评估水体样本的可饮用类型，本文选取该数据集 9 个指标作为特征指标,用于构造机器学习算法的特征空间，9 个水质评价指标的基本描述如表 1 所示：

表 **1** 本文选取的 **9** 个水质评价特征指标及单位

Table 1 9 features with units of water quality evaluation in this paper

编号	评价指标
1	酸碱度(pH)
2	硬度(Hardness)，mg/L
3	总溶解固体(TDS)，ppm
4	氯胺(Chloramines)，ppm
5	硫酸盐(Sulfate)，mg/L
6	电导率(EC)，μS/cm
7	总有机碳(TOC)，ppm
8	三卤甲烷(THMs)，μg/L
9	浑浊度(Turbidity)，NTU

绘制可安全饮用、不可安全饮用水体的样本分布直方图如图 1 所示，并设置高斯函数为核函数，基于核密度估计给出两类总体的概率密度如图 2 所示.由图 1、图 2 可知，两类总体在 9 个特征指标下的分布密度基本相似，差异仅表现为两类样本的分布量不均衡，因此当分类器基于上述先验对待测样本进行判别时，会由于两类样本之间的特征相似度过高而产生难以精确区分的情形.在证据理论中，若将一个特征指标视为一个信号源，则上述样本在 9 个特征指标下相似性过高的情形被称为来自多个信号源的信息产生了冲突，正是这种冲突使得机器学习算法不能生成有把握的分类结果，例如分类器可能在样本的输入特征不具有显著区分度时给出 P(正样本)=P(负样本)=0.5 的判别结果，此时可认为分类器判别失效.由于在这种特征空间中分类器形成精确判别结果的把握降低，机器学习算法判别的准确率也易受到样本特征随机扰动的影响，从而丧失稳定性.由于证据理论可以在多源信息存在较大冲突的情况下较稳定地形成决策，因此本文提出使用证据理论对多分类器进行集成学习，利用证据理论对结果进行信息融合，利用融合修正后的结果提升多分类器的判别性能，使得算法能在特征高冲突的前提下产生精确的决策.



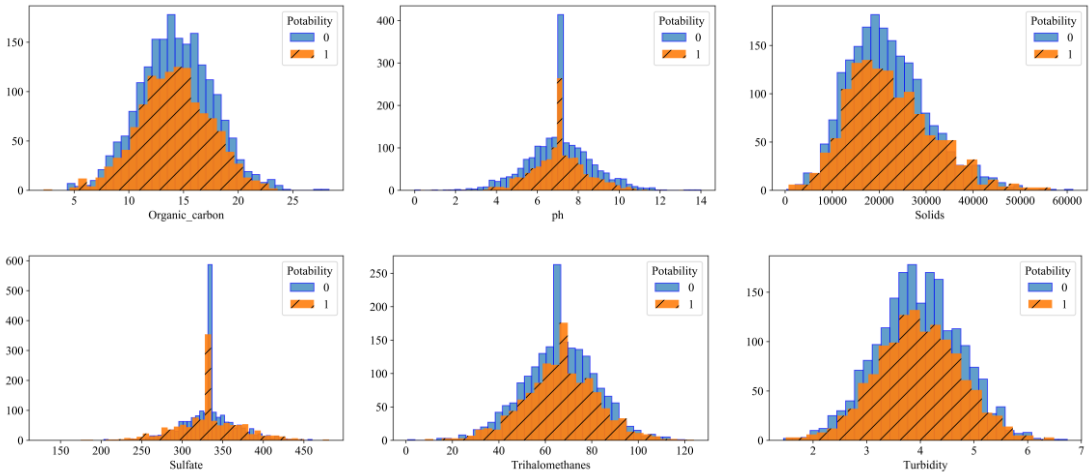


图 **1** 两类总体在 **9** 个特征指标下的样本分布直方图

Figure 1 Sample distribution histogram of two kinds of population under 9 features

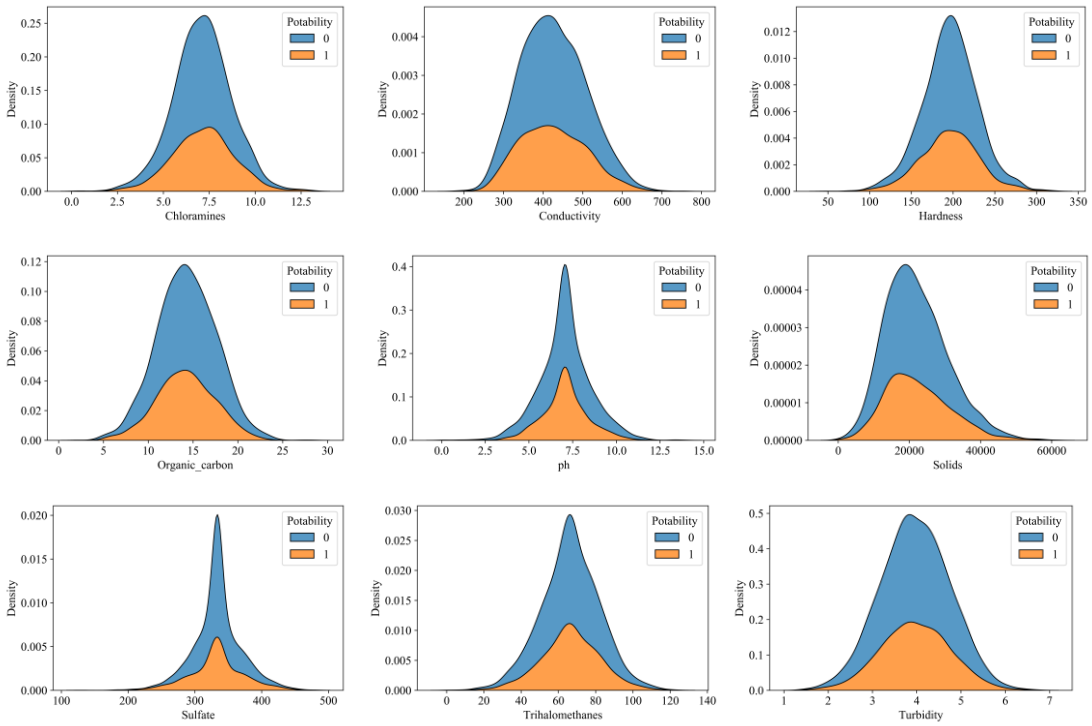


图 **2** 两类总体在 **9** 个特征指标下的核密度估计

Figure 2 Kernel density estimation of two kinds of population under 9 features

2. 2 集成学习和基于证据理论的信息融合

本文算法基于 9 个水体特征指标训练分类器，随后对待测水体样本是否可安全饮用进行预测.本文选取线性核支持向量机和随机森林作为基分类器，在测试集上输出分类器的判别结果，并通过证据理论的后处理改善集成学习的准确率.假设线性核支持向量机和随机森林可被视为信号源，那么 2 种不同的机器学习算法在训练集上给出的预测结果可以被视为信号源提供的信息，利用这些信息证据理论可以给出信息融合后的决策.本文对符号做如下约定：若记测试集为T，那么用$|T|$表示测试集的基数，即待测水体样本总数，不妨记t_i为测试集中第i个待测水体样本，并记标签集为$L=\{l_0,l_1\}$，其中l_0表示该样本属于不可安全饮用水体，l_1表示该样本属于可安全饮用水体.依据证据理论，可构造辨识框架$I=\{\emptyset,l_0,l_1,l_0\cup l_1\}$，由于在第$i$个待测

水体样本上线性核支持向量机和随机森林可以分别给出该待测样本属于可安全饮用和不可安全饮用水体的概率$p_{SVM,i}(l_0)$，$p_{SVM,i}(l_1)$，$p_{RF,i}(l_0)$，$p_{RF,i}(l_1)$，若记两种分类器信息源给出的初步基本概率指派为$\widetilde{m}_{SVM}$，$\widetilde{m}_{RF}$，同时对分类器给出的待测水体样本属于l_0、l_1、l_0或l_1、不属于l_0和l_1的概率进行建模，则两种初步基本概率指派可分别按照下式构造：

$$\begin{cases}\widetilde{m}_j(i,l_0)=p_{j,i}(l_0)\\ \widetilde{m}_j(i,l_1)=p_{j,i}(l_1)\\ \widetilde{m}_j(i,l_0\cup l_1)=1-\left|p_{j,i}(l_0)-p_{j,i}(l_1)\right|\\ \widetilde{m}_j(i,\emptyset)=0\end{cases}$$

其中$j\in\{SVM,RF\}$.记$S_j=\sum_l\widetilde{m}_j(i,l)$，其中$l\in I$，经归一化后形成基本概率指派$m_{SVM}$，$m_{RF}$：

$$m_j(i,l)=\frac{\widetilde{m}_j(i,l)}{S_j},l\in I,j\in\{SVM,RF\}$$

显然$\sum_l m_j(i,l)=1,l\in I,j\in\{SVM,RF\}$.由上式分析可知，按上述方法构造的基本概率指派满足以下性质：若分类器j给出的待测样本属于某一焦元的概率越大，则对应该焦元的初步基本概率指派也就越大，基本概率指派也越大；如果记$CF_{j,i}=\left|p_{j,i}(l_0)-p_{j,i}(l_1)\right|$，则$CF_{j,i}$可用于表征第$j$个分类器对第$i$个待测水体样本判别的困惑度，$CF_{j,i}$越小，表明分类器给出待测水体样本属于两个总体的可能性越接近，该样本属于两个总体的概率差异越小，可以认为分类器在判别待测水体样本类别时的困惑度越大，$\widetilde{m}_j(i,l_0\cup l_1)$和$m_j(i,l_0\cup l_1)$作为$CF_{j,i}$的反单调函数可以很好地把握分类器判别结果之间的差异，从而表达分类器对判别结果的把握程度，可以充分发挥证据理论表达不知道的能力，然而直接基于基本概率指派进行信息融合，可能由证据之间的冲突产生错误的分类结果，因此获得基本概率指派后，本文继续利用 BJS 散度定量估计基本概率指派之间的距离，利用加权修正后的基本概率指派进行信息融合，进而对待测样本的所属类别进行判别.

基于 BJS 散度的基本概率指派修正方法主要分为三步.第一步，首先依据 BJS 散度的定义计算第i个待测水体样本下线性核支持向量机和随机森林基本概率指派之间的证据距离：

$$BJS_i(m_{SVM},m_{RF})=\frac{1}{2}\left[\sum_l m_{SVM}(i,l)log\left(\frac{2m_{SVM}(i,l)}{m_{SVM}(i,l)+m_{RF}(i,l)}\right)+\sum_l m_{RF}(i,l)log\left(\frac{2m_{RF}(i,l)}{m_{SVM}(i,l)+m_{RF}(i,l)}\right)\right]$$

其中$l\in I$.随后，依据两组基本概率指派的 BJS 散度构建差异度量矩阵：

$$DMM_i=\begin{bmatrix}0 & BJS_i(m_{SVM},m_{RF})\\ BJS_i(m_{RF},m_{SVM}) & 0\end{bmatrix}$$

然后计算差异度量矩阵中单一基本概率指派与其它基本概率指派的平均 BJS 散度，该散度作为这一基本概率指派与其它基本概率指派之间的平均距离，可以表征该基本概率指派与其它基本概率指派之间的平均异化程度，如果这一平均距离越大，可以认为该基本概率指派与其它基本概率指派的偏差越大，因此有把握认为该基本概率指派给出的可信信息越小，在后续算法中应当减少该基本概率指派的信息表达.显然，由于本文只有来自线性核支持向量机和随机森林的两组基本概率指派，依据 BJS 散度的对称性，需要求解的 BJS 散度只有一个，因此平均 BJS 散度$\overline{BJS}_{i,SVM}=\overline{BJS}_{i,RF}=BJS_i(m_{SVM},m_{RF})$.

随后计算来自各个基本概率指派的支持度.由上述分析可知，单一基本概率指派的支持度越高，表明该基本概率指派与其它基本概率指派的差异程度越小，应具有更好的信息支持能力：

$$Support_{i,j}=\frac{1}{\overline{BJS}_{i,j}},j\in\{SVM,RF\}$$

最后计算来自支持度的证据置信度：

$$CV_{i,j}=\frac{Support_{i,j}}{\sum_k Support_{i,k}},j,k\in\{SVM,RF\}$$

第二步，对来自基本概率指派的信息量进行建模.在证据理论中，邓熵[14]作为一种定量计算基本概率指派不确定信息量的度量方法，常用于估计基本概率指派可供决策的信息量大小，若基本概率指派对应的邓熵越大，则该基本概率指派提供的不确定信息量越大.计算第j个基本概率指派的邓熵如下：

$$DE_{i,j} = -\sum_{l} m_j(i,l) log \frac{m_j(i,l)}{2^{|l|}-1}$$

其中$l \in I$，$j \in \{SVM, RF\}$.进一步计算第j个基本概率指派的信息量：

$$IV_{i,j} = \exp\{DE_{i,j}\} = \exp\left\{-\sum_{l} m_j(i,l) log \frac{m_j(i,l)}{2^{|l|}-1}\right\}$$

其中$l \in I$，$j \in \{SVM, RF\}$.针对信息量指标进行归一化处理，可得归一化信息量如下：

$$\overline{IV}_{i,j} = \frac{IV_{i,j}}{\sum_k IV_{i,k}}$$

其中$j, k \in \{SVM, RF\}$.归一化信息量可以保证信息量指标控制在[0,1]之间.

第三步，在获得了第j个基本概率指派的证据置信度和归一化信息量后，综合考虑上述两个指标提供的决策信息量，计算第j个基本概率指派的权重：

$$W_{i,j} = CV_{i,j} \times \overline{IV}_{i,j}, j \in \{SVM, RF\}$$

对权重进行归一化，获得归一化权重：

$$\overline{W}_{i,j} = \frac{W_{i,j}}{\sum_k W_{i,k}}$$

其中$j, k \in \{SVM, RF\}$.最后计算加权后的基本概率指派：

$$Wm_i = \sum_{k} \overline{W}_{i,k} \times m_k, k \in \{SVM, RF\}$$

最后依据 Dempster 组合规则对来自加权基本概率指派的信息进行融合，获得最终判别结果，至此基于证据理论和 BJS 散度的集成学习算法判别结束，选取概率最大的类别作为该待测水体样本的判别结果，本文算法的计算流程如表 2 所示.

表 **2** 本文算法计算流程

Table 2 Calculation flow of proposed algorithm

算法：基于机器学习和证据理论的水体质量预测算法
Step1. 对原始数据集按照比例系数δ随机分割，获得训练集和测试集T
Step2. 基于训练集训练线性核支持向量机、随机森林两种分类器
Step3. 向训练后的分类器输入测试集T，获得预测概率$p_{SVM,i}(l_0)$，$p_{SVM,i}(l_1)$，$p_{RF,i}(l_0)$，$p_{RF,i}(l_1)$，$1 \le i \le \|T\|$
Step4. 构建初步基本概率指派$\widetilde{m}_{SVM}$，$\widetilde{m}_{RF}$，归一化后构建基本概率指派m_{SVM}，m_{RF}
Setp5. 依据m_{SVM}，m_{RF}计算 BJS 散度的差异度量矩阵DMM，计算$CV_{i,j}$和$\overline{IV}_{i,j}$，并计算归一化权重$\overline{W}_{i,j}$
Step6. 计算重新加权后的基本概率指派Wm_i，基于 Dempster 组合规则进行信息融合，输出判别结果

3 实验设计与结果分析

本文算法基于 Python 3.7 开发，设置比例系数$\delta = 0.35$将数据集分割为训练集和测试集，对比算法选取线性核支持向量机、朴素贝叶斯、决策树、随机森林开展对比实验，选取准确率(Accuracy)和精确度(Precision)作为分类器性能的评价指标，其中准确率的计算公式如下：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

其中TP，TN是正确被识别为可安全饮用水体和不可安全饮用水体的样本总数，FP，FN是被错分为可安全饮用水体和不可安全饮用水体的样本总数，精确度指标可以衡量不同分类器正确分类的能力.精确度的计算公式如下式所示：

$$Precision = \frac{TP}{TP + FP}$$

精确度可以衡量分类器识别真正正样本的能力[15].本文算法与 4 种对比算法在准确率、精确度 2 个指标下的分类结果如表 3 所示：

表 **3** 本文算法和 **4** 种对比算法的准确率、精确度指标

Table 3 Accuracy and precision of proposed algorithm and four comparison algorithms

	SVM	朴素贝叶斯	决策树	随机森林	本文算法
Accuracy	0.6253	0.6277	0.6033	0.6701	**0.6859**
Precision	0.0000	0.4689	0.4476	0.6488	**0.8265**

由表 3 的结果可知，本文算法在准确率、精确度两个指标下均取得了最好效果，其中准确率指标相较于线性核支持向量机提高了 6.06%、随机森林提高了 1.58%，精确度指标相较于线性核支持向量机提高了 82.65%、随机森林提高了 18.65%.由表 3 的结果还可知，线性核支持向量机在饮用水质量预测问题中的精确度得分为 0，表明该算法未能成功识别任何可安全饮用的水体样本，因此可认为该算法在该问题下完全失效，但在引入随机森林和证据理论对分类结果进行集成后，该指标相较于支持向量机和随机森林均获得了显著提高，这一结果说明本文算法能显著降低将不可安全饮用水体错分为可安全饮用水体的风险，进一步表明了本文算法在饮用水质量预测问题中提高机器学习算法的有效性.

4 结论

由于现有的饮用水质量预测数据集中存在的不同群体特征分布过于接近、区分度不显著的高冲突问题，本文从提高机器学习算法在饮用水质量预测问题中的性能出发，提出了一种使用证据理论和 BJS 散度加权的集成学习方法，该方法通过证据理论将分类器的输出转换为基本概率指派，对基本概率指派的冲突性、支持度、置信度进行建模，将来自基本概率指派的信息重新加权，成功抑制了来自分类器输出的冲突，提高了模型的分类效果.本文创新性地将基于证据理论和BJS散度的集成学习方法应用于饮用水质量预测问题，实验结果表明本文算法相较于对比算法在准确率、精确度两个指标下均获得了显著提高，进一步说明了本文算法改善机器学习算法在饮用水质量预测问题中性能的稳定性和有效性.

参考文献：

[1]李雪清,郑航,刘悦忆等. 基于多源数据机器学习的区域水质预测方法研究[J].水利水电技术(中英文), 2021:1-14.

[2]戴青松,王沛芳,王超等.基于 LWCA-SVM 模型对洪泽湖饮用水源地二河闸断面水质的预测分析[J].中国农村水利水电,2017,07:62-66+71.

[3]Wen J U, Liu J, Zhang M, et al. An optical image quality evaluation method based on evidence theory[A].2018 IEEE CSAA Guidance, Navigation and Control Conference (GNCC)[C].Xiamen: IEEE, 2018, 1-5.

[4]Zhao J, Liu S, Wan J, et al. Change Detection Method of High Resolution Remote Sensing Image Based on D-S Evidence Theory Feature Fusion[J].IEEE Access, 2021, 9: 4673-4687.

[5]Pan Y, Zhang L, Li Z W, et al. Improved Fuzzy Bayesian Network-Based Risk Analysis With Interval-Valued Fuzzy Sets and D-S Evidence Theory[J].IEEE Transactions on Fuzzy Systems, 2019, 28, 9: 2063-2077.

[6]蒋雯,邓鑫洋. D-S 证据理论信息建模与应用[M]. 北京:科学出版社,2018:16-17.

[7]Chaosheng Zhu, Fuyuan Xiao, A belief Hellinger distance for D–S evidence theory and its application in pattern recognition[J]. Engineering Applications of Artificial Intelligence, 2021, 106:104452-104462.

[8]Zheng Y, Li G, Zhang W, et al. Feature Selection with Ensemble Learning Based on Improved Dempster-Shafer Evidence Fusion[J]. IEEE Access, 2019, 7 : 9032-9045.

[9]Wang Z, Rongxi W, Gao J, et al. Fault recognition using an ensemble classifier based on Dempster–Shafer

Theory[J]. Pattern Recognition, 2020, 99:107079-107094.
[10]路军,王梓耀,余涛.基于朴素贝叶斯和 D-S 证据理论的多时空数据融合[J].电气技术, 2019, 20(11): 27-32+45.
[11]Shu Xiaosong et al. Dam anomaly assessment based on sequential variational autoencoder and evidence theory[J]. Applied Mathematical Modelling, 2021, 98: 576-594.
[12]Fuyuan Xiao. Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy[J]. Information Fusion, 2019, 46 : 23-32.
[13] Aditya Kadiwal. Water Quality | Kaggle[EB/OL]. (2021-04-25)[2021-04-25]. https://www.kaggle.com/adityakadiwal/water-potability.
[14]Yong Deng. Deng entropy[J]. Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena, 2016, 91 : 549-553.
[15]崔巍,贾晓琳,樊帅帅等.一种新的不均衡关联分类算法[J].计算机科学,2020,47:488-493.

作者简介：

马晓剑（1977 年---），女，黑龙江省哈尔滨市人，硕士，副教授，研究方向为数据挖掘和图像处理。CCF 会员（会员号：），Email:

张家绪（2001 年—），男，山东省临沂市人，本科，研究方向为数据挖掘、证据理论和机器学习。

王奥（2001 年—），男，内蒙古通辽人，本科，研究方向为证据理论、数据挖掘和图像处理。

林煜华（2000 年—），男，福建省莆田市人，本科，研究方向为证据理论和机器学习。