基于加权朴素贝叶斯的水质数据分类研究

方志豪[1]，李正权[1,2]，张铭玮[1]
（1.江南大学物联网工程学院，江苏 无锡 214122；2.江苏省未来网络创新研究院，江苏 南京 211111）

摘 要：为更好地实施水环境管理政策，水质评价是基础环节，即根据某一水域多个水质参数，如何将其合理地划分到特定水质类别。针对该问题，提出了一种改进的朴素贝叶斯分类方法，该方法赋予不同属性以不同的权值，削弱了朴素贝叶斯条件独立性的假设，使分类结果更接近实际类别。首先，参考国家地表水水质自动监测站（以下简称国控水站）发布的数据，选取其中 500 条水质数据作为样本，基于溶解氧、高锰酸盐指数、氨氮和总磷 4 个指标建立评价体系；然后，利用改进朴素贝叶斯分类方法对样本进行学习与评价，并采用五折交叉验证法验证其分类性能。结果表明，改进朴素贝叶斯分类方法的准确率、精确率、召回率和 F1 值分别达到 96.0%、95.9%、93.8%和 94.8%，水质数据分类的性能指标相较于其他朴素贝叶斯分类方法更高，可对实际工程中遇到水质数据分类的问题提供一定的参考。

关键词：水质评价；朴素贝叶斯；五折交叉验证；性能指标
中图分类号：X824
文献标志码：A
doi: 10.11959/j.issn.2096−3750.2022.00255

Research on water quality data classification based on weighted Naive Bayes

FANG Zhihao[1], LI Zhengquan[1,2], ZHANG Mingwei[1]
1. School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China
2. Jiangsu Future Networks Innovation Institute, Nanjing 211111, China

Abstract: In order to better implement the water environmental management policies, water quality evaluation is the basic step, that is to reasonably divide it into specific water quality category according to multiple water quality parameters in a certain water area. Aimed at this problem, an improved Naive Bayes classification method was proposed, which endowed different attributes with different weights, weakened the assumption of Naive Bayes conditional independence, and made the classification result closer to the actual category. Firstly, referred to the data released by the national surface water quality automatic monitoring station, 500 water quality data were selected as samples, and an evaluation system with four indicators was established, including dissolved oxygen, permanganate index, ammonia nitrogen and total phosphorus. And then, the improved Naive Bayes classification method was used to learn and evaluate the samples, and its classification performance by the five fold cross validation method was verified. The results show that the accuracy, precision, recall and F1 value of the improved Naive Bayes classification method reach 96.0%, 95.9%, 93.8% and 94.8% respectively, with higher performance index of water quality data classification compared with other Naive Bayes classification method, which can provide some reference for the problem of water quality data classification encountered in actual engineering.

Key words: water quality evaluation, Naive Bayes, five fold cross validation, performance index

收稿日期：2021−10−08；**修回日期**：2021−12−29
通信作者：李正权，lzq722@jiangnan.edu.cn
基金项目：国家自然科学基金资助项目（No.61571108）；无锡市科技发展资金资助项目（No.H20191001，No.G20192010）；未来网络科研基金项目（No.FNSRFP-2021-YB-11）
Foundation Items: The National Natural Science Foundation of China (No.61571108), The Wuxi Science and Technology Development Fund (No.H20191001, No.G20192010), The Future Network Scientific Research Fund Project (No.FNSRFP-2021-YB-11)

0 引言

工业化的发展必然会带来一系列的环境问题，其中水环境的破坏日趋严重。水环境的污染问题严重制约着国家经济发展，影响民众的饮食健康。尤其在发展中国家，缺乏适合当地的水质评估方法，使得水资源管理政策实施起来困难重重，在缺乏有效水质监测与评估的情况下，几乎所有穿过城镇的河流都成为生活、工业和商业所产生废液的倾倒场[1]。因此，积极开展水环境的监控防治刻不容缓。水质数据的分类问题是水环境管理与保护的重要基础[2]，水质数据分类模型的好坏对水环境防治决策起着至关重要的作用。早期的水质数据分类问题依靠人工定期采样水体，然后带回实验室进行水质数据分析，最后做出分类的决策。这种方式虽然能按照指定的标准进行精确的分类，但是针对大量需要监控的场合，需要耗费太多人力、物力，而且不能保证实时性。随着物联网技术的发展，自动水质监测系统将成为主流[3]，如何对采集到的水质数据进行合理且准确的分类，是需要学者研究的课题。

目前，常用的水质评价方法主要有单因子评价法、综合污染指数法、模糊综合评价法、水质指数法等[4-9]。其中，单因子评价法将实际监测值与国家地表水环境质量标准[10]中各污染指标浓度限值进行比较，选取各污染指标所属最差类别作为水质类别，该方法赋予污染最严重的指标 100%的权重，结果表现为“过保护”。而综合污染指数法通过单因子评价法对每一项指标求得相对污染值，然后取其均值，该方法能反映河流综合水质状况[4]，但其评价结果是一个相对值，无法给出水质类别。对于水环境，各污染指标间存在着复杂的相关关系，水质分级标准的界限也比较模糊[5]，因此，基于模糊数学理论的综合评价方法有效契合了水环境中客观存在的不确定性[6]。但该方法无法识别出主要污染指标，且计算过程较为复杂。近几年来，水质指数（WQI, water quality index）法在河流水质分类中得到了充分研究[7]，它将多个水质指标的浓度值进行分级，经过加权组合计算后得到反映水质整体状况的单个值[8]，该方法能同时满足定性评价与定量评价的优点，并且用于不同监测断面之间的比较时，可以确定水质空间变化趋势[8-9]。但 WQI 的计算通常需要监测大量的水质参数，昂贵的监测成本限制了该方法在环境保护经费匮乏的地区使用。随着计算机技术的发展，一些机器学习相关方法被用于水质分类的研究中。文献[11]将随机森林（RF, random forest）算法用于巢湖水质的分类，实验证明该算法在水质分类中的效果优于极限学习机（ELM, extreme learning machine）算法和支持向量机（SVM, support vector machine）算法。文献[12]根据国家地表水环境质量标准建立了反向传播（BP, back propagation）神经网络模型，所得模型在测试数据中达到了 95%的正确率，然后将其用于广西北仑河口红树林自然保护区的水质分类。文献[13]为了更好拟合水质指标与水质等级之间的非线性关系，减少人为干预的因素，尝试将 BP 神经网络模型与模糊综合评价法相结合，然后应用于辽河口湿地不同时期、不同区域的水质评价，结果与实际监测数据基本相同，验证了该方法的可行性。

文献[11-13]考虑了水质指标与类别之间复杂的非线性关系，这导致这些方法复杂度较高，难以应用在实际场合。综合国内外水质评价研究进展，目前尚无通用的评价方法，每种方法各有侧重点，需要因地制宜。近年来一些学者将基于概率与数理统计的贝叶斯理论用于水质评价，有效解决了水质评价中遇到的困难[14]，相较于 BP 神经网络法，贝叶斯分类方法不需要大量水质数据样本且计算简单，分类效果也不低于前者。朴素贝叶斯分类方法是其中应用最广泛的分类方法之一，但朴素贝叶斯引入了条件独立性假设，如果条件成立，相比其他分类方法，该方法的误分类率理论上最小[15]。但实际中的分类问题很难满足该条件，为了减小条件独立性假设带来的影响，一些学者开始探寻各属性之间的相关性，借助其他方法改进朴素贝叶斯。文献[16]提出一种基于分类概率加权的朴素贝叶斯分类方法，该方法对每个特征属性分别做一次分类，然后将分类成功的概率作为各自属性的权重。文献[17]提出了基于属性值频率的实例加权朴素贝叶斯分类方法，该方法将每个训练实例的权重定义为其属性值频率向量与属性值个数向量的内积，大量实验表明，改进的朴素贝叶斯分类效果明显优于标准朴素贝叶斯。文献[18]使用信息增益量化各属性的重要程度，筛选出影响分类结果的主要属性以简化分类模型，然后使用增强训练的方式对分段后的训练集依次进行训练，实验结果取得了不错的分类性能和效率。文献[19]提出一种基于 Bagging 同质特征选择的朴素贝叶斯分类方法，通过选择出最佳特征子集以改进朴素贝叶斯，同时该方法运

行时间较短。文献[20]提出了基于评分搜索的改进树增强朴素贝叶斯分类方法，优化了树形贝叶斯网络结构且去除了冗余属性，实验证明该方法的分类精度要高于其他同类方法。

文献[16-17]分别从属性加权和实例加权的角度改进了朴素贝叶斯。文献[18-19]通过选择最佳特征子集提高分类性能，但该方法不适用于属性较少的分类场合。文献[20]对朴素贝叶斯进行了结构扩展，但该类方法通常计算烦琐、不便于理解，有悖于朴素贝叶斯为简化计算的初衷，其更适用于理论研究。基于此，本文针对所研究的水质分类场合，提出了一种新的实例加权方法，即根据属性取值对分类结果的贡献度为每个实例赋予不同权重。然后，结合文献[16]的属性加权方法，综合考虑属性本身和属性取值对分类结果的影响后，将其定义为加权朴素贝叶斯的最终权值。最后本文以国家地表水水质自动监测站发布数据作为样本，将本文改进的朴素贝叶斯分类方法用于水质数据分类，通过与标准朴素贝叶斯分类方法、文献[16]和文献[17]所提方法进行对比，验证本文所提方法对水质数据的分类性能。实验结果表明，本文所提方法对水质数据的分类性能最优。

1　水质数据样本构建

1.1　数据来源

本文数据来源于国控水站实时监测数据，其发布指标有水温、pH、溶解氧等 9 项。根据 2021 年 5 月全国地表水水质月报，在监测到的 3 573 个国考断面中，水质类别比例如图 1 所示。

水质数据类别分布极不均匀，其中Ⅱ类水质占比 43.3%，Ⅴ类水质和劣Ⅴ类水质仅分别占比 4.7% 和 2.2%，说明全国地表水总体水质良好。考虑到劣Ⅴ类水质数据较少，实验中本文将其划分到Ⅴ类数据集中。现随机选取国控水站 2021 年 7 月发布的 500 条水质数据作为样本集，其数据来源于安徽、江苏、河北等多个省份。经统计，其中水质为Ⅰ类的样本数量为 50 条，Ⅱ类水质数量为 203 条，Ⅲ类水质数量为 135 条，Ⅳ类水质数量为 72 条，Ⅴ类水质数量为 40 条。为直观显示样本类别分布情况，将数量转换为百分比，样本集中水质数据各类别比例如图 2 所示。随机选取的数据样本分布情况与图 1 相差不大，可用于后续实验分析处理。

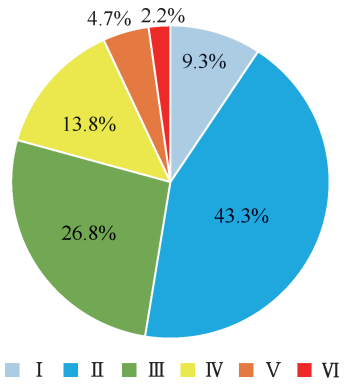

图 1　水质类别比例

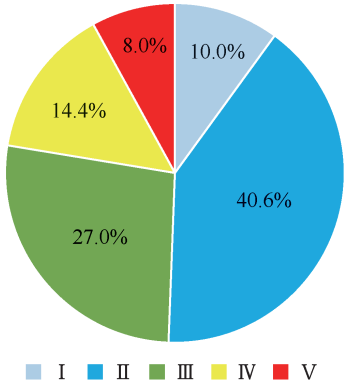

图 2　样本集中水质数据类别比例

1.2　数据预处理

本文以国控水站水质评价指标为准，即 pH、溶解氧、高锰酸盐指数、氨氮、总磷 5 项指标，各指标标准限值参考 GB3838-2002《地表水环境质量标准》[10]，见表 1。据本次收集的样本数据，所有样本的 pH 数值均在 6～9 之间，即不影响水质类别的判定，故剔除该指标后基于剩余的 4 个指标建立评价体系。然后，依据表 1 中各指标的标准限值，对每条样本各属性取值进行离散化处理。其中，落在Ⅰ类区间的属性取值用数值 1 代替，落在Ⅱ类区间的属性取值用数值 2 代替，以此类推。对样本离散化后的 4 个属性而言，其属性值越大，即代表单项指标污染程度越高。最后，将离散化后的 500 条水质样本数据保存在数据库中，作为预处理后的实验数据。

2　水质评价中的加权朴素贝叶斯分类方法

2.1　朴素贝叶斯分类方法

2.1.1　公式推理

朴素贝叶斯分类方法是基于贝叶斯定理的一种分类方法，它假设各属性对结果的影响相互独立，这样就将联合概率密度的计算转化为多个一维

概率密度的计算，降低了计算开销。根据条件独立性假设，朴素贝叶斯分类模型如图 3 所示。

表 1 GB3838—2002《地表水环境质量标准》标准限值节选

指标	Ⅰ类	Ⅱ类	Ⅲ类	Ⅳ类	Ⅴ类
pH 值	6～9	6～9	6～9	6～9	6～9
溶解氧/(mg·L^{-1})	≥7.5	≥6	≥5	≥3	≥2
高锰酸盐指数/(mg·L^{-1})	≤2	≤4	≤6	≤10	≤15
氨氮/(mg·L^{-1})	≤0.15	≤0.5	≤1.0	≤1.5	≤2.0
总磷/(mg·L^{-1})	≤0.02	≤0.1	≤0.2	≤0.3	≤0.4

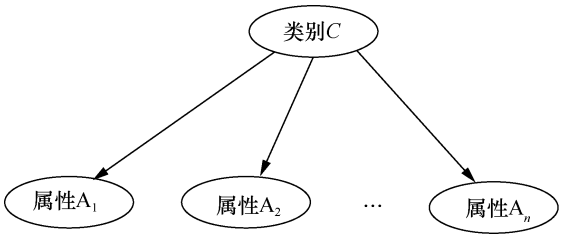

图 3 朴素贝叶斯分类模型

假设 $D=\{A_1,A_2,A_3,A_4,C\}$为预处理后的水质数据样本集，其中，A_1 代表溶解氧属性，A_2 代表高锰酸盐指数属性，A_3 代表氨氮属性，A_4 代表总磷属性，c_i（i=1,2,3,4,5）为类别 C 的取值，表示水质类别的 5 个等级。令 a_j（j=1,2,3,4）为属性 A_j 的具体取值，那么实例 $X=\{a_1,a_2,a_3,a_4\}$属于类 c_i 的概率，由贝叶斯公式可表示为

$$\begin{aligned}&P(c_i\mid a_1,a_2,a_3,a_4)\\&=\frac{P(a_1,a_2,a_3,a_4\mid c_i)P(c_i)}{P(a_1,a_2,a_3,a_4)}\\&=\alpha P(a_1,a_2,a_3,a_4\mid c_i)P(c_i)\end{aligned}\tag{1}$$

其中，$P(a_1,a_2,a_3,a_4)$是全概率公式，对于所有类别为常数，用正则化因子 α 代替。$P(c_i|a_1,a_2,a_3,a_4)$是类 c_i 的后验概率，$P(c_i)$为类 c_i 的先验概率，$P(a_1,a_2,a_3,a_4|c_i)$为实例 X 的后验概率。朴素贝叶斯假设实例 X 各属性间相互独立，根据概率的乘法定理，将式(1)转化为

$$P(c_i\mid a_1,a_2,a_3,a_4)=\alpha P(c_i)\prod_{j=1}^{N}P(a_j\mid c_i)\tag{2}$$

其中，N 表示属性的数量。对于给定某一实例 X，朴素贝叶斯通过计算每个类变量 c_i 在该实例下的后验概率 $P(c_i|X)$，从而找到具有最大后验概率的类，则称该类为极大后验假设，记作 C_{MAP}，如式(3)所示。

$$C_{\mathrm{MAP}}=\arg\max_{c_i\in C}P(c_i\mid a_1,a_2,a_3,a_4)\tag{3}$$

其中，α 为不依赖于 c_i 的常量，对计算结果不造成影响，将式(1)代入式(3)得到

$$C_{\mathrm{MAP}}=\arg\max_{c_i\in C}P(c_i)\prod_{j=1}^{N}P(a_j\mid c_i)\tag{4}$$

2.1.2 参数估计

水质评价中的朴素贝叶斯分类方法如式(4)所示，$P(c_i)$为水质类别的先验概率，其极大似然估计可以根据已有的水质数据样本集计算得到

$$P(c_i)=\frac{|D_{c_i}|}{|D|},1\leqslant i\leqslant 5\tag{5}$$

其中，$|D|$表示训练集中的总样本数，$|D_{c_i}|$表示训练集中类别取值为 c_i 的样本数。

对于条件概率 $P(a_j|c_i)$的计算，一般分为以下两种情况。

1) 如果属性 A_j 的取值 a_j 离散，同样可以根据训练样本的频数得到。

2) 如果属性取值连续，通常有两种方法估计条件概率，第一种方法是将每个连续属性取值离散化，即用相应的离散区间替换连续值；第二种方法是假设连续属性的取值服从某个概率分布，通过样本集估计必要的参数。

对水质评价的 4 个属性来说，其取值都连续，本文依据表 1 中各属性的标准限值对连续值进行离散化。条件概率 $P(a_j|c_i)$的估计可能出现为 0 的情况，从而造成整个后验概率 $P(c_i|a_1,a_2,a_3,a_4)$的计算为 0，此处需要引入拉普拉斯校准，修正后的条件概率估计为

$$P(a_j\mid c_i)=\frac{|D_{c_i,a_j}|+1}{|D_{c_i}|+N_j},1\leqslant j\leqslant 4,1\leqslant i\leqslant 5\tag{6}$$

其中，$|D_{c_i,a_j}|$表示训练集中类别为 c_i 且属性 A_j 取值为 a_j 的样本总数，N_j 表示属性 A_j 可取值的个数。

2.2 加权朴素贝叶斯分类方法

2.2.1 权重

朴素贝叶斯分类方法最大的优点就是计算复杂度低、分类准确度高，但是它的条件独立性假设在实际中往往不成立。同时朴素贝叶斯认为每个属性对分类结果作用都相同，即每个属性的权重都为 1。实际分类问题往往是一部分属性起着决定性作用，不是每个属性对分类决策的贡献程度都相同。

加权朴素贝叶斯分类方法是在式(4)的基础上，计算条件概率 $P(a_j|c_i)$时，为每一项都分配不同的权重，其模型为

$$C_{\mathrm{MAP}}=\arg\max_{c_i\in C} P(c_i)\prod_{j=1}^{N}P(a_j\mid c_i)^{w_j} \tag{7}$$

其中，w_j 为属性 A_j 的权值，N 为条件属性个数，在本文中 N 取值为 4。比较式(4)和式(7)发现，当条件属性的权值都为 1 时，两者完全等同。由此可见，朴素贝叶斯分类方法其实是加权朴素贝叶斯分类方法的一种特例，对后者来说，如何确定各条件属性的权值分配是问题的关键。

2.2.2 权值对分类的影响

加权朴素贝叶斯分类方法中，权值 w_j 代表属性 A_j 对分类决策的影响程度。直观上看，属性的权值越大，对分类结果的影响越大。在水质数据分类中，对给定某一实例 $X=\{a_1,a_2,a_3,a_4\}$，下面具体分析权值 w_j 如何影响概率估计，也就是该实例 X 被判定为某一类别的后验概率。

水质类别共分为 5 个等级，用 c_i 表示。根据式(2)，引入权值 w_j 后实例 X 划分到每个类别 c_i 的后验概率 $P(c_i|X)$可表示为

$$P(c_i\mid X)=\alpha P(c_i)\prod_{j=1}^{N}P(a_j\mid c_i)^{w_j} \tag{8}$$

加权朴素贝叶斯分类方法通过计算式(8)，找到最大后验概率的一项，然后将实例 X 判定为该类别。现将实例 X 属于类别 c_i 的后验概率与所有类别后验概率之和的比值定义为归一化后验概率，其表达式为

$$P'(c_i\mid X)=\frac{P(c_i\mid X)}{\sum_{k=1}^{5}P(c_k\mid X)} \tag{9}$$

通过式(9)看出，后验概率 $P(c_i|X)$越大，则其归一化后验概率也越大。现选取类别 c_1，对其归一化后验概率进行计算，如式(10)所示。

$$P'(c_1\mid X)=\frac{\alpha P(c_1)\prod_{j=1}^{N}P(a_j\mid c_1)^{w_j}}{\sum_{k=1}^{5}\left[\alpha P(c_k)\prod_{j=1}^{N}P(a_j\mid c_k)^{w_j}\right]}=\frac{1}{1+\frac{P(c_2)}{P(c_1)}\prod_{j=1}^{N}\left[\frac{P(a_j\mid c_2)}{P(a_j\mid c_1)}\right]^{w_j}+\cdots+\frac{P(c_5)}{P(c_1)}\prod_{j=1}^{N}\left[\frac{P(a_j\mid c_5)}{P(a_j\mid c_1)}\right]^{w_j}} \tag{10}$$

取其分母的子式 $\frac{P(c_2)}{P(c_1)}\prod_{j=1}^{N}\left[\frac{P(a_j\mid c_2)}{P(a_j\mid c_1)}\right]^{w_j}$ 进行分析，$\frac{P(c_2)}{P(c_1)}$ 取决于样本各类别的先验概率，为常量。对属性 A_j 来说，属性值 a_j 在各类别下的先验概率 $P(a_j|c_i)$取决于训练样本，令 $f_j=\frac{P(a_j\mid c_2)}{P(a_j\mid c_1)}$，则 f_j 也是确定量。当 $f_j<1$ 时，权值 w_j 越大，则分母的子式越小，$P'(c_1\mid X)$ 也就越大；当 $f_j>1$ 时，w_j 越大，则分母的子式越大，$P'(c_1\mid X)$ 也就越小；当 $f_j=1$ 时，$P'(c_1\mid X)$ 不受权值 w_j 的影响。由上述分析，权值 w_j 越大，对实例 X 属于类别 c_i 的归一化后验概率影响越大，也就是说，权值越大的属性对分类决策的贡献程度越大。

2.2.3 权值的计算方法

实际数据分类问题中，条件属性与类别之间通常存在一定的关联。如何量化两者关联程度的关键在于赋予与类别关联程度较大的属性较大的权值，关联程度较小的属性较小的权值。文献[16]提出了基于分类概率加权的朴素贝叶斯分类方法，该方法对每个条件属性分别做一次朴素贝叶斯分类，得到每个属性 A_j 分类正确的概率 P_j，然后将 P_j 归一化后作为属性 A_j 的权值，考虑到该权值较小，需乘以属性的个数 N。本文将该方法应用到水质数据分类中，则每个属性 A_j 的权值定义为

$$w_{1,j}=\frac{N\cdot P_j}{\sum_{k=1}^{N}P_k} \tag{11}$$

朴素贝叶斯的假设就是条件独立性，文献[16]将每个属性单独分类的准确率作为其权值具有一定的合理性。但该方法仅考虑了属性与类别的关联程度，实际上每个属性的不同取值也会对分类结果产生影响。本文针对水质数据的特点，尝试着量化属性取值对分类结果的影响程度。

对于水质监测管理，往往最差的单项指标对分类结果影响最大，以起到及时预警的作用。传统水质评价方法大多赋予污染较高的因子较高权重系数[4]，本文延续该思想，将其量化为实例权值的一部分。在数据离散化处理中，属性取值越大，即代表单项指标评级越差。对实例 $X=\{a_1,a_2,a_3,a_4\}$，属性 A_j 对分类决策的贡献度定义为

$$w_{2,j}=\frac{a_j}{\frac{1}{N}\sum_{k=1}^{N}a_k} \tag{12}$$

综合考虑属性以及属性取值对分类结果的影响程度，将其量化为实例属性的最终权值 w_j。根据式(11)和式(12)，将 w_j 定义为

$$w_j=\frac{w_{1,j}+w_{2,j}}{2} \tag{13}$$

2.3　水质数据分类的实现

本文研究的水质数据有 4 条属性，分别是溶解氧、高锰酸盐指数、氨氮和总磷，其水质类别分为Ⅰ、Ⅱ、Ⅲ、Ⅳ、Ⅴ，共 5 个等级。现利用本文提出的加权朴素贝叶斯分类方法对已有的水质数据样本进行学习，然后对待分类样本进行分类。基于该方法实现水质数据分类的关键在于求解各条件属性的权值，具体分类步骤如下。

步骤 1　数据预处理。将训练样本进行数据离散化。

步骤 2　构造分类器。

步骤 2.1　扫描所有训练样本，统计水质类别为 c_i 的实例数 COUNT(c_i)以及在该类别下属性 A_j 取值为 a_j 的实例数 COUNT(A_j=a_j^c_i)，形成实例统计表。

步骤 2.2　概率参数学习。根据实例统计表，计算所有类别的先验概率 $P(c_i)$以及在类别 c_i 下属性 A_j 取值为 a_j 的先验概率 $P(a_j|c_i)$，将结果保存在先验概率表。

步骤 2.3　权值参数学习。根据先验概率表的结果，对每个属性分别做一次朴素贝叶斯分类，根据各属性分类成功的概率利用式(11)计算得到各属性的权值 $w_{1,j}$，形成权值表 1。

步骤 3　分类。对给定待分类实例 X，将各属性根据其取值进行式(12)的计算，得到各属性的权值 $w_{2,j}$，形成动态的权值表 2，根据式(13)可得到该实例的最终权值表 3，然后调用先验概率表及权值表 3，根据式(7)得到实例 X 的分类结果。

步骤 4　结束。输出类标号。

由以上分类步骤可看出，权值 $w_{2,j}$ 由步骤 3 根据具体实例计算得到，形成的权值表是动态变化的。加权朴素贝叶斯分类方法计算流程如图 4 所示。

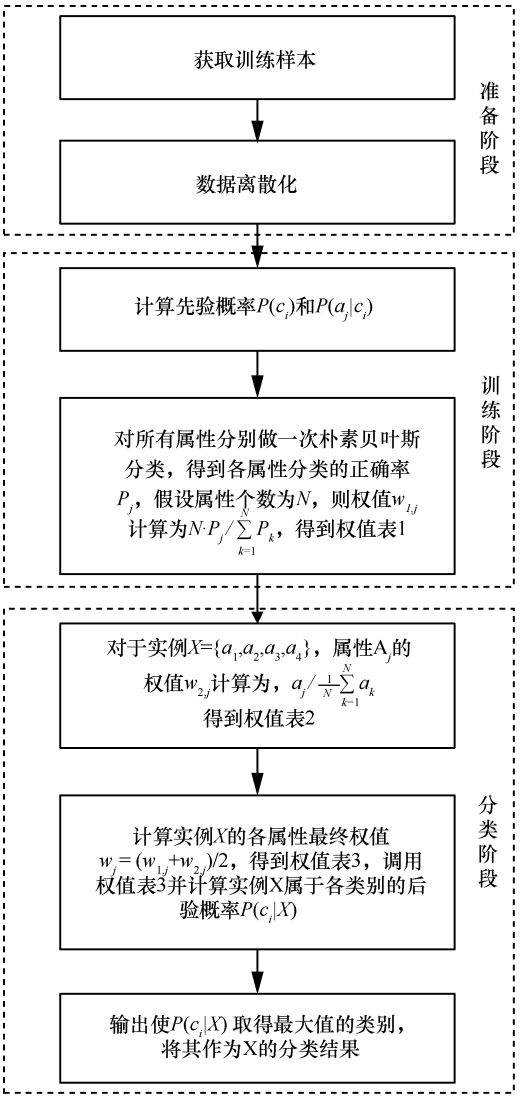

图 4　加权朴素贝叶斯分类方法计算流程

3　实验与结果分析

3.1　实验设计

为测试本文所提方法对水质数据分类的有效性，设计 4 组实验进行对比分析。第 1 组、第 2 组、第 3 组和第 4 组实验分别为标准朴素贝叶斯分类方法、文献[16]所提方法、文献[17]所提方法和本文所提方法对水质数据的分类。

本次实验环境如下：操作系统为 Windows 10，处理器为 3.6 GHz CPU，内存为 8 GB，开发环境是 IDEA 2019 + JDK 1.8，使用 Java 语言进行开发。实验数据为第 1.2 节预处理后的 500 条水质数据样本，使用 MySQL 数据库进行保存。

本文采用五折交叉验证法评估 4 种方法的分类性能。五折交叉验证的步骤如下：将已有水质数据

集平均分为 5 组，尽量保证每组数据类别分布的一致性，轮流将其中 4 组数据的并集作为训练集，剩下 1 组数据作为测试集，综合 5 轮测试结果评估各方法对水质数据的分类性能。五折交叉验证法原理如图 5 所示。

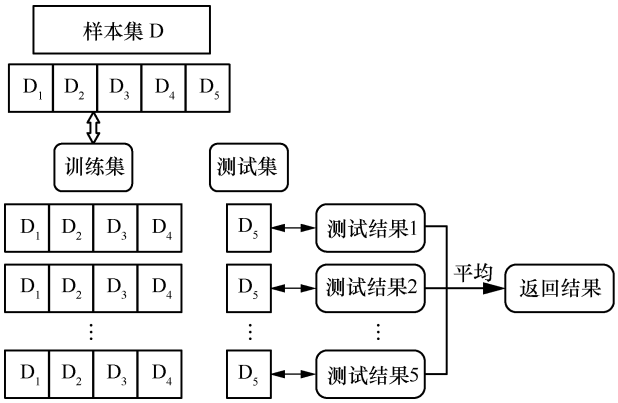

图 5　五折交叉验证法原理

3.2　分类性能评估

3.2.1　评估指标

在机器学习算法中，对于一个二分类问题的性能评估，通常使用准确率、精确率、召回率和 F1 值等指标。为便于分析，二分类混淆矩阵见表 2。

表 2　　二分类混淆矩阵

混淆矩阵	预测为正	预测为负
实际为正	TP	FN
实际为负	FP	TN

其中，TP+FP+FN+TN=总的测试样本数。

准确率表示分类器预测成功的样本数与总的测试样本数的比值，也是最常用的分类性能指标，定义为

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \tag{14}$$

精确率表示分类器对测试集成功预测为正类的样本数与预测为正类的总样本数的比值，定义为

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{15}$$

召回率表示分类器对测试集成功预测为正类的样本数与测试集中实际为正类的样本数的比值，定义为

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{16}$$

精确率越高表示分类器对负样本的区分能力越强，对实际为负类的样本预测为正类的条件越苛刻；召回率越高表示分类器对正样本的区分能力更强，将有可能为正类的样本尽可能都预测为正类。因此，精确率和召回率本身是相互矛盾的，两者在很多情况下不能同时取得提高。为了更好地评估分类器的性能，F1 值则综合考虑了精确率和召回率，定义为两者的调和平均值，计算如下

$$\text{F1}=\frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \tag{17}$$

本文研究的水质数据分类本质上属于多分类问题，可将其拆分为 M 个二分类问题，并使用以上 4 个指标评估 4 种方法对水质数据的分类性能，其中 M 表示水质类别数量。建立的水质分类的混淆矩阵见表 3，用于保存 4 种方法对水质数据测试样本的分类结果。

表 3　　水质分类的混淆矩阵

		预测的类				
		类=1	类=2	类=3	类=4	类=5
实际的类	类=1	d_{11}	d_{12}	d_{13}	d_{14}	d_{15}
	类=2	d_{21}	d_{22}	d_{23}	d_{24}	d_{25}
	类=3	d_{31}	d_{32}	d_{33}	d_{34}	d_{35}
	类=4	d_{41}	d_{42}	d_{43}	d_{44}	d_{45}
	类=5	d_{51}	d_{52}	d_{53}	d_{54}	d_{55}

水质数据级别共分为Ⅰ、Ⅱ、Ⅲ、Ⅳ、Ⅴ，分别用类 1、类 2、类 3、类 4 和类 5 表示，表 3 中 d_{ij} 表示分类方法将实际为类别 i 的样本预测为类别 j 的数量统计。现将水质分类问题拆分为 5 个二分类问题，每次只关注一个类别的分类结果，将该类别视为正类，其他 4 个类别都视为负类，然后分别计算 4 种方法在每个类别下的性能指标，最后综合计算出整体分类性能指标。

根据以上分析，将整体分类性能指标定义为

1) 准确率 $A=\dfrac{\sum_{i=1}^{5}d_{ii}}{\sum_{i=1}^{5}\sum_{j=1}^{5}d_{ij}}$，指分类器分类正确的样本数占总的测试样本数的比值。

2) 精确率 $P=\dfrac{1}{5}\sum_{i=1}^{5}\text{Precision}(i)$，其中 Precision($i$)

代表分类器在类别 i 上的精确率。

3) 召回率 $R=\frac{1}{5}\sum_{i=1}^{5}\text{Recall}(i)$，其中 Recall($i$)代表分类器在类别 i 上的召回率。

4) 得分值 $\text{F1}=\frac{2PR}{P+R}$，F1 值越接近 1，则分类器性能越好。

3.2.2 评估结果

本文根据第 3.1 节的实验方案，将 4 种方法对水质数据样本进行分类，分类结果用混淆矩阵表进行统计。根据式(15)和式(16)，4 种方法在每个类别下的精确率见表 4，4 种方法在每个类别下的召回率见表 5。

表 4　4 种方法在每个类别下的精确率

	类别 1	类别 2	类别 3	类别 4	类别 5
标准朴素贝叶斯分类方法	92.6%	93.5%	95.5%	95.3%	93.9%
文献[16]所提方法	90.9%	94.4%	94.8%	95.5%	96.9%
文献[17]所提方法	96.2%	95.3%	96.3%	95.4%	88.9%
本文所提方法	98.0%	96.2%	95.6%	95.5%	94.4%

表 5　4 种方法在每个类别下的召回率

	类别 1	类别 2	类别 3	类别 4	类别 5
标准朴素贝叶斯分类方法	100.0%	99.5%	94.1%	84.7%	77.5%
文献[16]所提方法	100.0%	99.0%	94.1%	87.5%	77.5%
文献[17]所提方法	100.0%	99.5%	96.3%	86.1%	80.0%
本文所提方法	100.0%	99.5%	97.0%	87.5%	85.0%

根据表 4，4 种方法对每个水质类别分类的精确率几乎都达到了 90%以上，其中本文所提方法在每个类别下的精确率都要高于标准朴素贝叶斯分类方法，只有在类别 5 和类别 3 下的精确率分别低于文献[16]和文献[17]所提方法。由表 5 知，本文所提方法在每个类别下的召回率均高于或等于其他 3 种朴素贝叶斯分类方法。

根据表 4、表 5，计算出各方法对水质样本分类的整体性能指标，4 种方法在水质分类中的性能指标对比如图 6 所示。

本文提出的加权朴素贝叶斯分类方法对水质样本分类的准确率、精确率、召回率和 F1 值分别达到 96.0%、95.9%、93.8%和 94.8%，各性能指标均在 90%以上，取得良好的分类效果。相较于标准朴素贝叶斯分类方法、文献[16]和文献[17]所提方法，本文所提方法对水质样本分类的准确率分别提高 1.8%、1.6%、0.8%，精确率分别提高 1.7%、1.4%、1.5%，召回率分别提高 2.6%、2.2%、1.4%，F1 值分别提高 2.1%、1.8%、1.4%。通过对比，明显看出本文所提方法对水质样本分类的各性能指标相较于其他 3 种方法都有较大幅度的提升，其更加适用于水质分类。

3.3 归一化后验概率分析

对于某个样本实例，朴素贝叶斯分类方法通过计算各属性在每个类别下的后验概率，且以后验概率最大的一项作为该实例的分类结果。因此，当朴素贝叶斯分类方法计算每个测试样本在实际类别下的后验概率时，其值越高，则分类结果可信度越高。为突出加权朴素贝叶斯分类方法对实际类别后验概率的影响，现统计标准朴素贝叶斯分类失误的样本编号，随

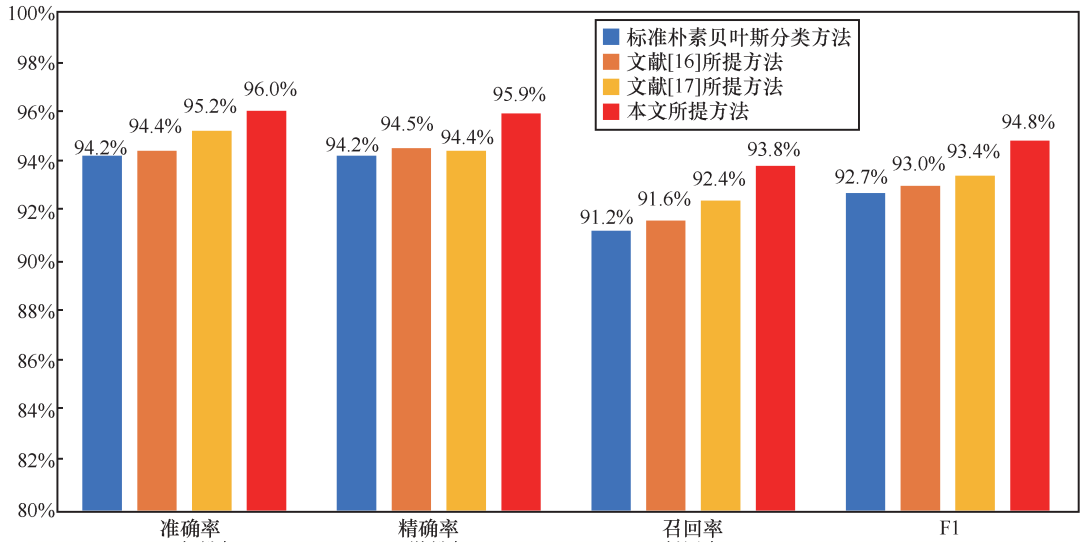

图 6　4 种方法在水质分类中的性能指标对比

机选取 10 个编号，然后收集 4 种方法对这些样本在每个类别下的后验概率。为便于分析，根据式(9)计算出实际类别下的归一化后验概率，实际类别的归一化后验概率对比如图 7 所示。

在选取的 10 个样本中，除了编号为 119 和 437 的样本，本文所提方法对样本实际类别归一化后验概率的计算结果均明显高于其他 3 种方法。实验结果表明，对于水质数据样本的分类，本文所提方法对样本实际类别归一化后验概率的提升最为明显，也就是说其分类结果更接近水质样本实际类别。

4　结束语

针对已有水质数据样本，本文研究了基于朴素贝叶斯的水质分类方法。但朴素贝叶斯的条件独立性假设很大程度上影响了水质数据分类性能，于是本文在该分类方法基础上提出了一种实例加权方法，综合考虑了水质属性及其取值对分类结果的影响程度，用加权后的属性条件概率代替了原有朴素贝叶斯，使分类结果尽可能贴近样本实际类别。本文以国家地表水水质自动监测站发布数据为例，选取了其中 500 条数据作为样本，设计了 4 组对比实验，结果表明本文所提方法对水质数据分类性能最优，其准确率、精确率、召回率和 F1 值分别达到了 96.0%、95.9%、93.8%和 94.8%，可对实际工程中的水质分类问题提供一定的参考。

对于加权朴素贝叶斯水质分类方法，对属性赋予权值虽然提高了分类性能，但没有考虑冗余属性。在加权朴素贝叶斯模型学习时也会赋予冗余属性一个权值，这样不但会影响分类精度，还会影响分类效率。当水质分类中的条件属性增多时，如何约简属性显得格外重要，这也将是下一步研究的方向。

参考文献：

[1] AWOKE A, BEYENE A, KLOOS H, et al. River water pollution status and water policy scenario in Ethiopia: raising awareness for better implementation in developing countries[J]. Environmental Management, 2016, 58(4): 694-706.

[2] 宁阳明, 尹发能, 李香波. 几种水质评价方法在长江干流中的应用[J]. 西南大学学报(自然科学版), 2020, 42(12): 126-133.
NING Y M, YIN F N, LI X B. Application of several evaluation methods for river water quality in the Yangtze River Mainstream[J]. Journal of Southwest University (Natural Science Edition), 2020, 42(12): 126-133.

[3] 乔辉. 基于物联网技术的水质监控系统[J]. 物联网学报, 2017, 1(1): 81-85.
QIAO H. Water quality monitoring system based on IoT[J]. Chinese Journal on Internet of Things, 2017, 1(1): 81-85.

[4] 王竹, 朱士江, 刘扬, 等. 不同水质评价方法在滦河下游段的比较应用[J]. 节水灌溉, 2019(10): 68-72, 77.
WANG Z, ZHU S J, LIU Y, et al. Comparative application of different water quality evaluation methods in the downstream section of Luanhe-River[J]. Water Saving Irrigation, 2019(10): 68-72, 77.

[5] 杨浩, 张国珍, 杨晓妮, 等. 基于模糊综合评判法的洮河水环境质量评价[J]. 环境科学与技术, 2016, 39(S1): 380-386, 392.
YANG H, ZHANG G Z, YANG X N, et al. Comprehensive evaluation on water environment quality of the Tao river based on fuzzy comprehensive Method[J]. Environmental Science & Technology, 2016, 39(S1): 380-386, 392.

[6] 王瑶, 王慧勇, 安丽娟, 等. 黄壁庄水库水质评价及氮污染成因分析[J]. 水电能源科学, 2020, 38(4): 60-63.
WANG Y, WANG H Y, AN L J, et al. Water quality evaluation and cause analysis of nitrogen pollution in huangbizhuang Reservoir[J]. Water Resources and Power, 2020, 38(4): 60-63.

[7] WU Z S, WANG X L, CHEN Y W, et al. Assessing river water quality using water quality index in Lake Taihu Basin, China[J]. Science of the Total Environment, 2018, 612: 914-922.

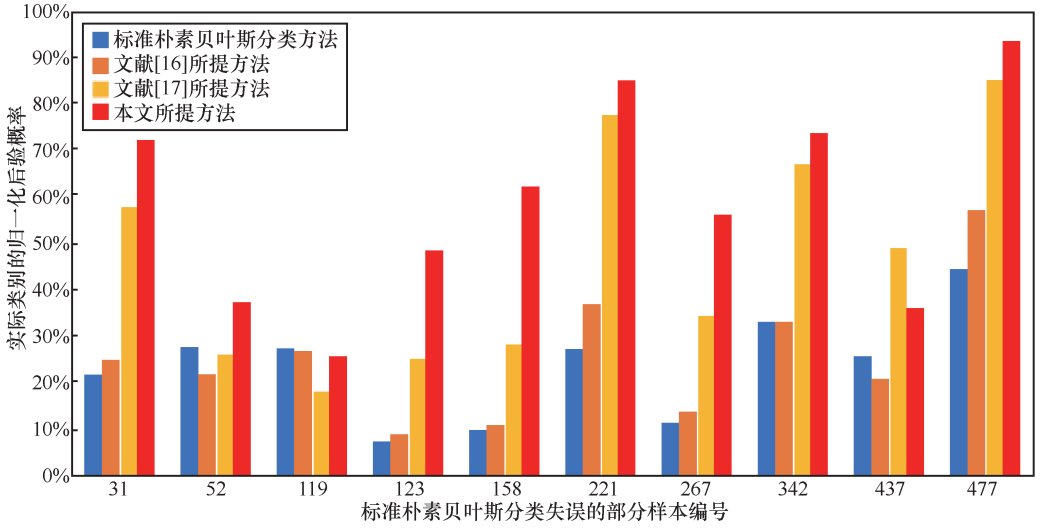

图 7　实际类别的归一化后验概率对比

[8] SUN W, XIA C Y, XU M Y, et al. Application of modified water quality indices as indicators to assess the spatial and temporal trends of water quality in the Dongjiang River[J]. Ecological Indicators, 2016, 66: 306-312.

[9] HOU W, SUN S H, WANG M Q, et al. Assessing water quality of five typical reservoirs in lower reaches of Yellow River, China: using a water quality index method[J]. Ecological Indicators, 2016, 61: 309-316.

[10] 国家环境保护总局, 国家质量监督检验检疫总局. 地表水环境质量标准:GB 3838—2002[S]. 北京: 中国环境科学出版社, 2002.
State Environmental Protection Administration of the People's Republic of China, General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China. Environmental quality standards for surface water.GB 3838—2002[S]. Beijing: China Environment Science Press, 2002.

[11] 张颖, 高倩倩. 基于随机森林分类算法的巢湖水质评价[J]. 环境工程学报, 2016, 10(2): 992-998.
ZHANG Y, GAO Q Q. Water quality evaluation of Chaohu Lake based on random forest method[J]. Chinese Journal of Environmental Engineering, 2016, 10(2): 992-998.

[12] ZHANG R, LIANG S, OU M H, et al. Evaluation of water quality for mangrove ecosystem using artificial neural networks[C]//Proceedings of 2018 International Conference on Advanced Mechatronic Systems (ICAMechS). Piscataway: IEEE Press, 2018: 257-261.

[13] 王铁良, 苏芳莉, 孙迪, 等. 基于模糊BP神经网络的辽河口湿地水质评价[J]. 西北林学院学报, 2020, 35(5): 195-200.
WANG T L, SU F L, SUN D, et al. Water quality evaluation of Liaoheestuary wetland based on back propagation artificial neural Network[J]. Journal of Northwest Forestry University, 2020, 35(5): 195-200.

[14] 万哲慧, 王珅, 冯孙林, 等. 熵权贝叶斯模型在珊溪水库水环境质量评价的应用[J]. 节水灌溉, 2018(3): 55-57, 62.
WAN Z H, WANG S, FENG S L, et al. Application of entropy weight Bayes model in Shanxi reservoir drinking water sources water quality assessment[J]. Water Saving Irrigation, 2018(3): 55-57, 62.

[15] HAN J, KAMBER M, PEI J. Data mining: concepts and techniques, 3rd Edition[M]. San Francisco: Morgan Kaufmann Publishers, 2012.

[16] 张步良. 基于分类概率加权的朴素贝叶斯分类方法[J]. 重庆理工大学学报(自然科学), 2012, 26(7): 81-83.
ZHANG B L. Naive Bayesian classifier method based on the classification of the probability-Weighted[J]. Journal of Chongqing University of Technology (Natural Science), 2012, 26(7): 81-83.

[17] XU W Q, JIANG L X, YU L J. An attribute value frequency-based instance weighting filter for naive Bayes[J]. Journal of Experimental and Theoretical Artificial Intelligence, 2019, 31(2): 225-236.

[18] 陈誉洋. 基于特征选择和增强训练的朴素贝叶斯网络钓鱼检测[D]. 合肥: 安徽大学, 2020.
CHEN Y Y. Naive Bayes phishing detection based on feature selection and reinforcement training[D]. Hefei: Anhui University, 2020.

[19] SUBRAMANIAN R S, PRABHA D. Customer behavior analysis using Naive Bayes with bagging homogeneous feature selection approach[J]. Journal of Ambient Intelligence and Humanized Computing, 2021, 12(5): 5105-5116.

[20] 陈曦, 张坤. 一种基于树增强朴素贝叶斯的分类器学习方法[J]. 电子与信息学报, 2019, 41(8): 2001-2008.
CHEN X, ZHANG K. A classifier learning method based on tree-augmented naive Bayes[J]. Journal of Electronics & Information Technology, 2019, 41(8): 2001-2008.

[作者简介]

方志豪（1996− ），男，江南大学物联网工程学院硕士生，主要研究方向为水质监测系统应用和开发。

李正权（1976− ），男，江南大学物联网工程学院教授，主要研究方向为大规模MIMO技术、协作通信、物联网等。

张铭玮（1998− ），男，江南大学物联网工程学院硕士生，主要研究方向为水质监测系统应用和开发。