

带特征染色体的遗传算法与支持向量机结合 进行生物水质预警的研究

杜秋菊, 徐建瑜*, 葛英辉

(宁波大学 信息科学与工程学院, 浙江 宁波 315211)

摘要: 近年来水污染事件频发, 给人类饮用水安全带来巨大隐患, 而生物监测法能够从生物学角度对水质状况做出综合分析. 选取斑马鱼作为受试生物, 当水质发生变化时, 鱼群在毒性物质作用下产生应激反应, 此时运用图像处理技术量化鱼类群体行为参数; 并用带特征染色体的遗传算法与支持向量机相融合的方法进行特征选择 and SVM 参数优化, 分析影响水质预警的关键量化参数. 结果显示, 新方法 with 网格搜索法、不带特征染色体的遗传算法与 SVM 结合的方法相比较, 有更高的准确率.

关键词: 水质监测; 斑马鱼; 特征染色体; 遗传算法; 支持向量机

中图分类号: TP183

文献标志码: A

文章编号: 1001-5132 (2016) 01-0042-05

随着我国经济的持续高速发展, 水环境日益恶化, 近年来频繁发生的水污染事件, 不仅严重阻碍了社会经济的发展, 而且给人们的生命安全造成极大威胁. 如 2005 年松花江硝基苯污染事件, 给沿岸百万居民的生活带来不便; 2008 年云南昆明阳宗海砷污染事件, 严重影响了当地居民的生活, 并且要 10 年甚至更长的时间才能恢复水质; 2009 年江苏盐城自来水厂受到酚类化合物污染, 使市民的生活及工业生产受到严重影响.

面对频繁发生的水污染事件, 目前应用比较广泛和成熟的水质监测方法是理化分析法^[1], 但是这种方法对水质的分析测定需要一定时间, 不能及时监测水质, 并且不能直接反映多种污染因素联合作用对饮用水安全的影响. 另一种方法是生物监测法^[2], 其利用水生生物个体或群体的健康状况、生理特征和行为反应等的变化来表征水体质量的变化, 进而揭示水环境污染的状况, 从生物学角度为水质预警提供科学依据和信息基础. 生物监测法弥补了理化分析法的缺点, 能够对水质做出

实时综合的评价.

近年来有人将一些机器学习算法应用于水质监测的数据分析, 如自组织映射、小波变换^[3]、马尔科夫模型^[4]等, 但由于水生生物运动的随机性、复杂性和非线性, 导致监测到的数据量很大且数据离散度大, 使得算法存在耗时长、准确率低等缺点. 廖悦等^[5]提出基于支持向量机(Support Vector Machine, SVM)和遗传算法结合的水质分类预警方法, 但是该算法只考虑到 SVM 参数对准确率的影响, 没有考虑有效分类量化参数的选取, 同时 10% 的误警率对于实现在线预警还有困难.

笔者采用带特征染色体的遗传算法与 SVM 结合, 提取出对分类预警起关键作用的特征子集, 然后对水质量化参数进行分类, 进而提高了预警的准确率.

1 实验材料和实验过程

实验选取健康的成年红斑马鱼(*Danio rerio*)作为受试生物, 实验用水为除氯并曝气 48 h 以上的自

收稿日期: 2014-03-20.

宁波大学学报(理工版) 网址: <http://journalg.nbu.edu.cn/>

基金项目: 国家自然科学基金(51208270); 浙江省新苗人才计划项目(2012R405054); 宁波市自然科学基金(2013A610158, 2013A610147); 王宽诚教育基金.

第一作者: 杜秋菊(1987-), 女, 河北衡水人, 在读硕士研究生, 主要研究方向: 水质预警和数字图像处理. E-mail: duqiuju1234567@sina.com

*通信作者: 徐建瑜(1973-), 女, 黑龙江牡丹江人, 博士/副教授, 主要研究方向: 水质预警和数字图像处理. E-mail: xujianyu@nbu.edu.cn

来水,水温为 $(23.5 \pm 0.5)^{\circ}\text{C}$, pH 值为 7.2 ± 0.2 , 光照周期为 12 h:12 h(亮:暗). 斑马鱼在实验室驯养 2 周,实验前 24 h 停止喂食.

根据 Cu^{2+} 对斑马鱼的 24h- LC_{50} 和实验容器储水容量计算得实验所需的 $\text{CuSO}_4 \cdot 5\text{H}_2\text{O}$ 质量(精确到 0.1 mg), 配置 24h- LC_{50} 受试溶液.

在实验前 2 h, 放置 10 条斑马鱼在定制的实验装置中^[5]. 当鱼群适应实验环境 2 h 后, 用摄像头实时捕捉鱼群在加入 Cu^{2+} 前后各 1 h 内的行为反应和体色变化的视频序列. 视频图像传入计算机并进行实时预处理、目标识别以及行为和位置参数的量化, 量化后得到的数据记录到 SQL 数据库中. 最后从数据库导出数据进行离线分析和处理, 完成 SVM 分类预警参数的选择和特征子集的选取.

2 特征提取

文中选取 9 个量化参数来分析鱼群的运动, 其中主要的几个量化参数介绍如下:

(1)鱼群重心横、纵坐标(pixels): 根据文献[6]方法计算鱼群在水平和竖直的重心. 它们分别反映了鱼群在水平和竖直方向的平均位置和鱼群的分布情况.

(2)鱼群水平、竖直方向标准差^[6](pixels): 根据鱼群重心和各条鱼的投影计算鱼群在水平、竖直方向上的空间标准差, 2 个参数反映出鱼群体的密集

程度.

(3)平均游速(body length speed): 用鱼游动的体长倍数来衡量^[7], 反映鱼群体的活跃程度.

(4)体色: 红斑马鱼在受到污染物刺激时, 鱼的体色会逐渐发生变化(如变浅或变暗), 根据文献[8]方法量化体色.

(5)水面、水底鱼的数量: 鱼在遇到毒性物质时会出现缺氧或游动无力等状况, 严重时鱼会浮于水面或沉于水底, 所以量化水面与水底鱼的数量可以从一方面反映水质状况.

(6)速度方差: 某时刻 10 条鱼速度的方差, 反映了鱼群运动的一致性.

SQL 数据库中记录的 9 个特征值见表 1.

表 1 鱼群 9 个量化特征值

序号	特征值	含义
1	Shoal MidX	群心横坐标
2	Shoal MidY	群心纵坐标
3	Shoal DensityX	水平方向方差
4	Shoal DensityY	竖直方向方差
5	Shoal Activity	平均游速
6	ShoalHSI Intensity	体色
7	Variance	速度方差
8	NumTop	水面鱼条数
9	NumLow	水底鱼条数

在实验过程中以每秒 5 帧的速率进行采集, 并

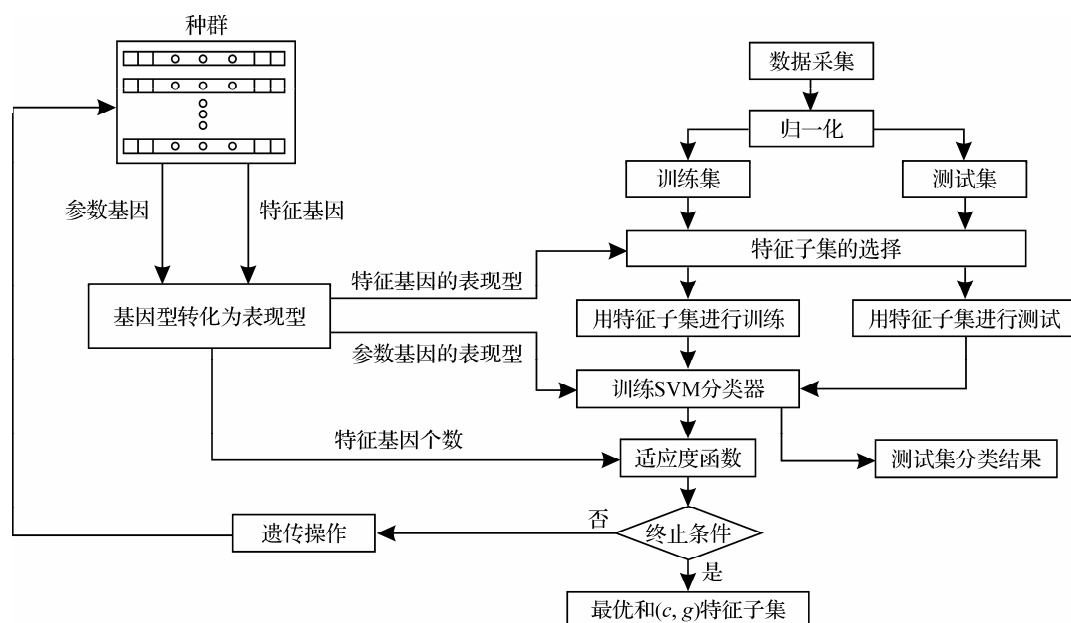


图 1 带特征染色体遗传算法流程

同时处理视频图像,每秒5帧图像计算得到的参数取平均值作为该秒的参数值,同时将提取的9个特征参数记录到SQL2000数据库中.鱼群中毒前后各1h的数据全部记录到数据库中,为后续的数据处理做准备.

3 带特征染色体的遗传算法

遗传算法^[9]是由 John Holland 在 20 世纪 60 年代提出的一种基于模拟达尔文自然选择和生物系统遗传的自适应优化搜索方法,它采用适应度函数和概率变换规则来指导搜索方向.遗传操作包括选择、交叉、变异操作.将遗传算法应用于支持向量机特征选择和参数优化,所得到的结果是优化的参数和特征子集.

带特征染色体的遗传算法^[10]以遗传算法为基础,它与不带特征染色体遗传算法的根本区别在于遗传算法不仅用于支持向量基参数(惩罚参数 c 和核参数 g)的优化,同时找到最优的特征子集使分类准确率最高.新生成的带特征子集的染色体经过适应度评价,选择其中适应度高的染色体构成下一代种群,继续进行下一代的遗传操作.带特征染色体的遗传算法流程如图 1 所示.

带特征染色体的遗传算法与 SVM 结合进行特征选取和参数优化的具体步骤如下.

3.1 染色体编码及初始种群的产生

带特征染色体的遗传算法与 SVM 结合的方法编码^[11-12]分为两部分(图 2),其中前一部分 $b_c^1 - b_c^n$ 和 $b_g^1 - b_g^n$ 分别是对参数 c 和参数 g 的二进制编码;而后一部分,设含有 N_f 个特征的原始特征集 $F = \{f_1, f_2, \dots, f_{N_f}\}$,对每个特征子集进行二进制编码,得到长度为 N_f 的二进制码串 $X = x_1 x_2 \dots x_{N_f}$,若 $x_i = 1$,表示特征子集包含第 i 个特征 f_i ,否则不包含.因此, $H = \{x_1 x_2 \dots x_{N_f} \mid x_i \in (0,1), i \in [1, N_f]\}$ 可表示所有特征子集的集合,即个体空间.

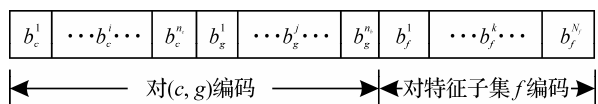


图2 带特征染色体的染色体编码

3.2 个体适应度值

支持向量机的分类准确率和被选择的特征数用于构造适应度函数^[13].根据高的适应度值取

决于高的分类准确率和少的特征数的特点,适应度函数定义为:

$$R_{\text{fitness}} = w_1 \times R_{\text{accuracy}} + w_2 \times 1 / \sum_{i=1}^{N_f} x_i,$$

式中, N_f 为总的特征数, w_1 和 w_2 分别是分类准确率和被选特征数的权值($w_1 + w_2 = 1$).

3.3 根据特征排序

首先,利用遗传算法本身的随机性将遗传算法重复运行 N 次,每次选取末代优良个体 n 个,共得到 $N \times n$ 个优良个体.然后,统计每个特征在较优特征子集中出现的频度,并按照频度值对特征进行降序排列,得到排序后的特征序列 $G = \{g_1, g_2, \dots, g_{N_f}\}$.最后,假设关键特征集 R_0 为空,按照特征的排序进行试探,第 i 次操作时将特征 g_i 加入已测试特征集 $R_i = \{g_1, g_2, \dots, g_i\}$,观察当前特征集 R_i 对于测试样本的分类能力,在达到较高准确率后即得到较优的特征集.

4 实验结果及分析

4.1 实验设计

实验采用的平台是 Intel Core i3-2130 3.4 GHz CPU, 4GB RAM, Win7 操作系统,开发环境为 Matlab R2008a.

算法参数设置:种群大小为 20,交叉率 0.4,变异率 0.02.染色体采用二进制编码,总长度为 22,其中参数 c 由前 8 位编码构成,搜索范围为 (0,100);参数 g 由中间 5 位编码构成,搜索范围为 (0,30);最后 9 位为特征染色体的编码.根据多次试验经验选取 $w_1 = 0.8$, $w_2 = 0.2$.初始种群个体随机生成.

从数据库中选取 4 484 个数据,其中 970 个数据为加入 1 倍 24h-LC₅₀ Cu²⁺前的数据,其余 3 514 个数据是加入 Cu²⁺后的数据.每次试验随机选取 2 242 个作为训练集,另外 2 242 个数据作为测试集.在特征子集的查找、模型的训练和测试中都使用 k 次折叠交叉检验(其中 $k = 5$),以提高测试结果的准确性,减少偶然性.

4.2 特征子集选取

采用带特征染色体的遗传算法与 SVM 结合的方法,训练 10 次,每次提取末代准确率最高的前 10 个个体(共 100 个个体),保存这 100 个个体的 100 个特征子集,构成 100×9 的矩阵,其中为 0 表

示没有被选中, 不为 0 表示作为特征子集. 统计在 100×9 的矩阵中, 以 1~9 的频度得到结果(表 2).

表 2 100 个个体中各个特征的频度

序号	特征值	频度
1	Shoal MidX	14.41
2	Shoal MidY	10.01
3	Shoal DensityX	11.61
4	Shoal DensityY	13.75
5	Shoal Activity	13.36
6	ShoalHSI Intensity	11.61
7	Variance	10.14
8	Num Top	4.67
9	Num Low	10.41

根据 3.3 节中的方法, 当加入 1、4、5、6、3、9、7 特征参数以后, 分类的准确率已达到最大, 再加入其它特征分类的准确率下降. 由此可以看出, 对生物水质预警起关键作用的斑马鱼量化参数是 Shoal MidX、Shoal DensityY、Shoal Activity、Shoal HSI Intensity、Shoal DensityX、Num Low 和 Variance. 而 ShoalMidY 和 Num Top 参数对分类准确率起副作用, 使得分类准确率下降. 用得到的 7 个特征进行基于遗传算法的 SVM 分类预警, 因为训练集和测试集是由 4 484 个数据随机划分, 重复 10 次实验得到的测试集和训练集的准确率和参数 (c, g)(表 3).

表 3 测试集和训练集准确率

序号	测试集/%	训练集/%	最优 c	最优 g
1	93.53	94.64	13.42	2.84
2	92.73	95.53	22.65	3.81
3	92.75	94.73	52.27	1.16
4	92.82	96.20	1.19	15.7
5	92.9	95.45	16.67	3.24
6	93.22	94.69	98.94	1.42
7	92.67	95.49	3.22	7.01
8	93.32	94.96	26.05	3.67
9	92.6	96.03	20.71	4.92
10	92.93	97.41	1.75	22.16

4.3 3 种方法的比较

为了进行对比试验, 分别采用传统网格搜索法(Grid Algorithm)、不带特征染色体的遗传算法与 SVM 结合(GASVM)以及带特征染色体的遗传算法

与 SVM 结合的方法(Feature Subset + GASVM)进行水质量化参数的分类, 比较各自的性能.

前 2 种方法中, 原始数据先用 Principal Components Analysis(PCA)降维, 然后用网格搜索法或不带特征染色体的遗传算法与 SVM 结合对训练集进行模型训练, 测试集进行分类, 重复 10 次得到训练模型平均所需时间和测试集平均准确率.

采用带特征染色体的遗传算法与 SVM 结合的方法获得最优特征子集(Feature Subset), 即频度最高的 7 个特征, 然后用 SVM 对 7 维数据进行分类, 训练 10 次得到平均所需时间和平均准确率.

表 4 为上述 3 种方法对有毒和无毒水质环境下提取的鱼群量化参数进行分类的结果. 由数据可见, 网格搜索法虽然能在较短的时间内寻找到(c, g), 但是其分类准确率明显低于后 2 种方法, 且会导致误警率升高. 在训练 SVM 模型过程中, 带特征染色体的遗传算法与 SVM 结合的方法比不带特征染色体的遗传算法所需时间短, 且分类准确率较高.

表 4 3 种方法分类结果的比较

算法	所需时间/s	准确率/%
Grid Algorithm	383	86.56±0.36
GASVM	529	89.31±0.43
Feature Subset+GASVM	474.8	92.95±0.31

5 结语

综上所述, 运用带特征染色体的遗传算法与 SVM 结合的方法先提取出对水质预警起关键作用的量化参数, 再进行水质有毒和无毒的分类, 效果比前两种方法好, 它使得在降低误警率的同时对提高水质预警的实时性有很大帮助.

参考文献:

- [1] 汪红军, 周连凤. 水环境安全预警系统研究进展[J]. 环境科学与管理, 2011, 36(3):32-35.
- [2] STOREY M V, VANDER G B, BURNS B P. Advances in on-line drinking water quality monitoring and early warning systems[J]. Water Research, 2011, 45(2):741-747.
- [3] KIM C K, KWAK I S, CHA E Y, et al. Implementation of wavelets and artificial neural networks to detection of toxic response behavior of chironomids (*Chironomidae: Diptera*) for water quality monitoring[J]. Ecological Modelling, 2006, 195(1/2):61-71.

- [4] LIU Y, LEE S, CHON T. Analysis of behavioral changes of zebrafish (*Danio rerio*) in response to formaldehyde using Self-organizing map and a hidden Markov model [J]. Ecological Modelling, 2011, 222(14):2191-2201.
- [5] LIAO Y, XU J Y, WANG Z W. Application of biomonitoring and support vector machine in water quality assessment [J]. Zhejiang Univ Sci B, 2012, 13(4):327-334.
- [6] ISRAELI D, KIMMEL E. Monitoring the behavior of hypoxia-stressed *Carassius auratus* using computer Vision[J]. Aquacultural Engineering, 1996, 15:423-440.
- [7] XU J Y, LIU Y, CUI S R, et al. Behavioral responses of tilapia (*Oreochromis niloticus*) to acute fluctuations in dissolved oxygen levels as monitored by computer vision[J]. Aquacultural Engineering, 2006, 35(3):207-217.
- [8] 徐建瑜, 姜雄晖, 刘鹰. 基于计算机视觉的鱼体色明暗程度量化方法[J]. 农机化研究, 2006(6):140-142.
- [9] HOLLAND J H. Adaptation in natural and artificial systems[M]. Ann Arbor: University of Michigan Press, 1975:105-121.
- [10] 赵明渊, 唐勇. 基于带特征染色体遗传算法的支持向量机特征选择和参数优化[J]. 控制与决策, 2010, 25(8):1133-1138.
- [11] 杨洁, 郑宁. 基于遗传算法的 SVM 带权特征和模型参数优化[J]. 计算机仿真, 2008, 25(9):115-117.
- [12] HUANG C L, WANG C J. A GA-based feature selection and parameters optimization for support vector machines [J]. Expert Systems with Applications, 2006, 31(2):231-240.
- [13] SEO K K. A GA-based feature subset selection and parameter optimization of support vector machine for content based image retrieval[J]. In Advanced Data Mining and Application, 2007, 4632:594-604.

Implementation of GA-based Feature Subset Selection Combined with SVM for Biological Water Quality Monitoring

DU Qiu-ju, XU Jian-yu*, GE Ying-hui

(Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China)

Abstract: In recent years, the water pollution incidents have been reported more frequently than ever before, imposing hazards to the drinking water. Bio-monitoring method can be used to perform a comprehensive check and analysis of water quality. In this article, *Danio rerio* is selected as a testing organism, the fish responding to stress under the action of toxic substances is observed. Image processing technique is used to quantify the fish population parameters. The support vector machine (SVM) and genetic algorithm with characteristic chromosome are applied for feature selection and SVM parameter optimization simultaneously, analyzing the key quantitative parameters issuing the water quality warning. The method reported in this work achieves higher accuracy than both Grid algorithm and the genetic algorithm without characteristic chromosome.

Key words: water quality monitoring; *Danio rerio*; feature subset; genetic algorithm; support vector machine (SVM)

(责任编辑 章踐立)