文章编号:2096-1472(2023)10-0011-04
DOI:10.19644/j.cnki.issn2096-1472.2023.010.003

利用自适应半监督决策树的溢油自动识别模型

刘宏宇,周 慧

(大连东软信息学院大数据科学系,辽宁 大连 116023)
✉ liuhongyu@nou.cn; zhouhui@neusoft.edu.cn

摘 要:目前,海面溢油检测分类器大多为监督学习分类器,但是针对特定油进行分类时,标签数据较少,因此监督学习分类器难以获得较好的识别效果。为提升识别准确率,采用了最大相关-最小冗余(mRMR)特征选择方法,同时为了解决标签样本较少的问题,选择自适应的半监督决策树学习模型,对公开的海面溢油样本集在不同的标签样本比例下进行分类实验,在仅有5.0%、7.5%、10.0%、15.0%和20.0%标签样本的情况下,自适应的半监督决策树学习模型的识别准确率,相比监督学习分类模型SVM和决策树的识别准确率,分别平均提升了26.22%和16.22%。实验结果表明,该方法在标签样本较少的情况下实用性较强。

关键词:溢油自动识别;半监督决策树;自适应置信度
中图分类号:TP753 **文献标志码**:A

Oil Spill Automatic Identification Model Using Adaptive Semi-Supervised Decision Tree

LIU Hongyu, ZHOU Hui

(*Department of Big Data Science, Dalian Neusoft Information University, Dalian* 116023, *China*)
✉ liuhongyu@nou.cn; zhouhui@neusoft.edu.cn

Abstract: Most classifiers currently used in sea surface oil spill detection are supervised learning classifiers. But when classifying specific oil, there is limited label data, making it difficult for supervised learning classifiers to achieve good recognition results. To improve the recognition accuracy, Max-Relevance and Min-Redundancy (mRMR) feature selection method is adopted. In order to solve the problem of fewer label samples, an adaptive semi-supervised decision tree learning model is selected to conduct classification experiments on the publicly available sea surface oil spill sample set at different label sample ratios. Compared to the supervised learning classification model SVM and decision tree, the recognition accuracy of the adaptive semi-supervised decision tree learning model is improved by an average of 26.22% and 16.22%, respectively, when 5.0%, 7.5%, 10.0%, 15.0%, and 20.0% label samples are only available. The experimental results show that the proposed method has strong practicality in the case of fewer label samples.

Key words: automatic identification for oil spill; semi-supervised decision tree; adaptive confidence

0 引言(Introduction)

目前,全球海运业快速发展,因船只碰撞导致的溢油事故频发,而溢油监测是溢油应急管理和决策支持的主要部分[1-3]。

随着机器学习算法的广泛应用,溢油的分类和检测也越来越多地应用机器学习方法[4-5]。例如,FINGAS等[6]使用马尔可夫随机场算法的油膜识别准确率为86.5%。XU等[7]采用局部自适应阈值和SVM分类器进行溢油识别。在带标签油膜样本充足的情况下,上述方法是有效的。但在实际应用中,油膜标签数据较少而无标签数据容易获取。通过大量未标记数据提升学习器性能,是目前应用最为广泛的半监督学习方法[8]。

本文首先采用最大相关-最小冗余(mRMR)算法提取更具鉴别力的油膜特征,其次在标签数据较少的情况下,利用自适应置信度的半监督决策树进行溢油识别,最后采用公开数据集进行实验证明本文所提分类器具有较好的泛化能力。

收稿日期:2023-01-19

1 溢油特征选择(Oil spill characteristics selection)

在溢油识别的研究过程中,需要结合不同类别的图像特征,而不是依赖于单一特征,主要包括几何特征、统计特征、纹理特征等[9]。通常依靠经验选择溢油特征,但是仅凭经验难以得到合适的特征集,尤其是特定油种。特征选择算法是解决上述问题的有效手段之一[10]。利用 mRMR 算法对候选特征进行有效选择,计算特征与目标变量之间的最大相关性,以及特征间的最小冗余性,利用互信息度量对特征进行评价,从特征集合中筛选出合适的特征子集。溢油特征选择的具体过程如下。

(1)计算最大相关性与最小冗余度。最大相关性体现特征对类别的区分能力,具体是找到一个包含$|S|$个特征的特征集S,使得S中的所有特征与类别的相关性最大化;最小冗余度考虑特征之间最小相似性,具体是找到一个包含$|S|$个特征的特征集S,使得S中的每个特征之间是相互最小相似。

$$\max D(S,C)=\frac{1}{|S|}\sum_{x_i\in S}I(x_i,C) \tag{1}$$

$$\min R(S)=\frac{1}{|S|^2}\sum_{x_i,x_j\in S}I(x_i,x_j) \tag{2}$$

其中,S为特征子集,$C=\{c_1,c_2,c_3,\cdots,c_n\}$为类别变量,$x_i$、$x_j$为第$i$、$j$个特征,$I$为$c(x_i,x_j)$的相关函数,即互信息,公式如下:

$$I(x;y)=\iint P(x,y)\log\frac{P(x,y)}{P(x)P(y)}\mathrm{d}x\mathrm{d}y \tag{3}$$

(2)利用增量搜索方法寻找近似最优的特征。假设已有特征集S_{m-1},并且$x_i\in X-S_{m-1}$,$\max\Phi(D,R)$作为特征评价标准,其中$\Phi=D(S,C)/R(S)$,则 mRMR 评价条件如下:

$$mRMR_m(x_i)=\max_{x_i\in X-S_{m-1}}\left[\frac{I(x_i,C)}{\frac{1}{|S|}\sum_{x_j\in S_{m-1}}I(x_i,x_j)}\right] \tag{4}$$

(3)根据 mRMR 评价和排序结果,从初始的特征集中提取最具有分辨性的特征子集,组成输入特征向量$\boldsymbol{X}$,$\boldsymbol{X}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$。

2 自适应半监督溢油检测模型(Adaptive semi-supervised oil spill detection model)

在众多的监督学习分类算法中,决策树是非常有效且应用较为广泛的经典算法之一,具有参数少、容易解释、适合集成等优点。决策树中的内部结点称为决策结点,每个决策结点包含1个测试条件,并根据测试结果发射2个或2个以上的分支;树中的叶子结点被称为预测结点,每个预测结点对应1个类别(分类树)。待预测的样本从树根出发,依据自身输入属性值和当前决策结点的测试条件决定流向哪个分支,当到达预测结点后,便得到预测结果。同时,决策树从根到叶子结点路径上的测试条件的提取可以认为是一条规则,一棵训练好的决策树容易转化为一系列预测规则。

自适应的半监督决策树采用模糊聚类的方法划分无标签样本,该划分方法将样本到各个分类目标(即簇心)的距离作为优化目标。自适应的半监督决策树完成一次划分后,样本到各个簇心的距离同样能反映该样本隶属于各个相应簇的程度。对于一个样本来说,它与哪个分类目标的距离越近,则它对相应类别的隶属程度就越高。同时,一个样本对不同簇隶属程度间的差异,也能反映该样本划分的模糊程度。如图1中μ_1、μ_2和μ_3分别表示三个簇的中心,无标签样本x_1、x_2与μ_1的距离分别小于它们与μ_2、μ_3的距离。从聚类的角度看,x_1、x_2都属于簇$C1$;但从模糊聚类的角度看,x_1与μ_1的距离远远小于其与μ_2、μ_3的距离,x_2与μ_1的距离仅略小于其与μ_2的距离。说明x_2分簇模糊程度要高于x_1的分簇模糊程度。

当使用自适应半监督决策树预测一个无标签样本时,样本从根结点出发,需要经过多次划分直至到达叶子结点。在每次划分中,需要计算该样本与当前结点所有子结点表示簇的中心的距离,该样本坠入具有最近距离的相应分支。与此同时,最近距离与次近距离的比值可以同时被计算求得,这个比值在这里被定义为该样本在本轮划分中的模糊度。假设图1中的样本x_1与μ_1、μ_2和μ_3的距离分别为0.1、1和2,样本x_1本轮的划分模糊度为0.1/1=0.1,假设图1中样本x_2与μ_1、μ_2和μ_3的距离分别为0.4、0.5和2.3,样本x_2本轮的划分模糊度为0.4/0.5=0.8。

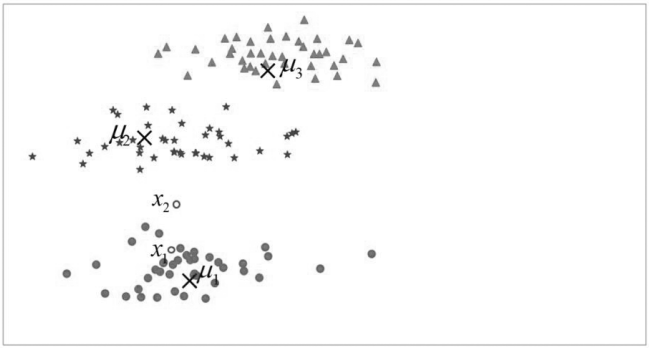


图1 样本划分的模糊程度

Fig. 1 The fuzziness degree of sample division

预测时,当无标签样本经历多次划分到达叶子结点后,该样本的各轮划分模糊度已经被计算,这些模糊度的平均值就是关于这个样本的预测模糊度,可用于衡量对该样本预测的置信程度。预测模糊度越低,表示预测置信度越高。公式(5)和公式(6)给出了样本$\boldsymbol{x}$预测模糊度的计算方法。

$$fuzziness(\boldsymbol{x})=\frac{1}{n_s}\sum_{i=1}^{n_s}\frac{dis(i,\boldsymbol{x},\boldsymbol{\mu}_{\text{nearest1}})}{dis(i,\boldsymbol{x},\boldsymbol{\mu}_{\text{nearest2}})} \tag{5}$$

$$dis(i,\boldsymbol{x},\boldsymbol{\mu})=(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \tag{6}$$

公式(5)和公式(6)中,$\boldsymbol{x}$ 为 mRMR 特征选择方法提取溢油图片的灰度统计特征和纹理统计特征,$fuzziness(\boldsymbol{x})$表示样本 $\boldsymbol{x}$ 获得的预测模糊度,n_s 表示样本 $\boldsymbol{x}$ 被划分的次数,$dis(i,\boldsymbol{x},\boldsymbol{\mu}_{\text{nearest1}})$和 $dis(i,\boldsymbol{x},\boldsymbol{\mu}_{\text{nearest2}})$分别表示第 i 轮划分中样本 $\boldsymbol{x}$ 与最近簇心和次近簇心的距离,$\boldsymbol{C}$ 为特征向量 $\boldsymbol{x}$ 协方差矩阵,$\boldsymbol{C}=E\{[\boldsymbol{x}-E(\boldsymbol{x})][\boldsymbol{x}-E(\boldsymbol{x})^{\mathrm{T}}]\}$。

3　实验分析(Experimental analysis)

3.1 mMRM 特征选择分析

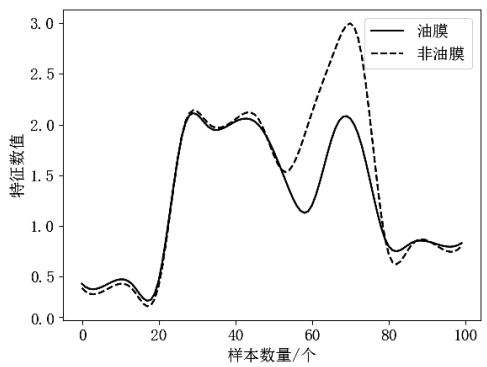
实验利用训练集提取特征向量并根据训练模型识别遥感图像中的油膜与非油膜。获取溢油特征后,利用 mRMR 算法进行特征提取,并且分别利用决策树和 SVM 两个分类模型验证特征的有效性。不同的特征提取数量对应的油膜识别准确率如表 1 所示。

表1　mRMR 特征选择后的识别准确率

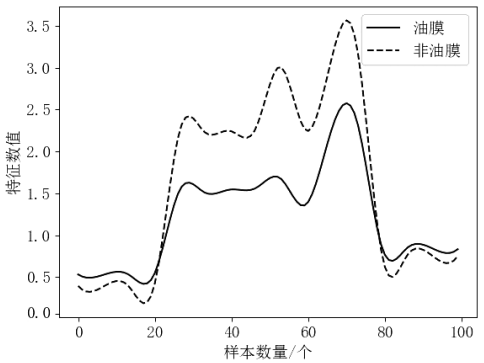
Tab.1 Recognition accuracy after mRMR feature selection

mRMR 特征选择个数/个	SVM 识别准确率/%	决策树识别准确率/%
5	79.1	87.2
6	76.9	89.6
7	76.9	90.1
8	78.1	91.7
9	78.1	89.4
10	77.5	86.2
11	77.5	85.2
12	76.3	84.0

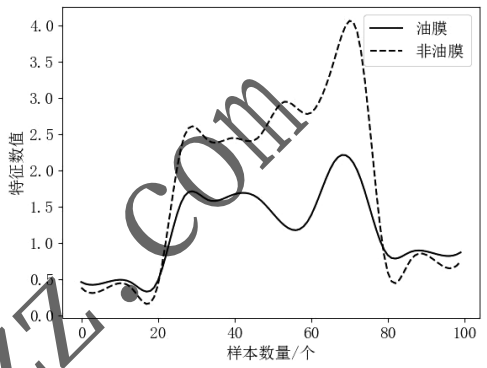
通过实验结果可以看出,利用 mRMR 计算特征之间的相关性,并从候选特征中筛选合适的特征子集,当特征选择个数为 8 个时,在不同的分类算法中,溢油识别效果都能达到最好。最终提取的 8 个特征在油膜和非油膜的对比结果如图 2 所示。



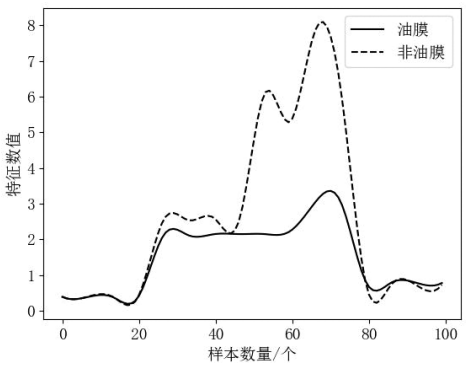
(a)角二阶矩



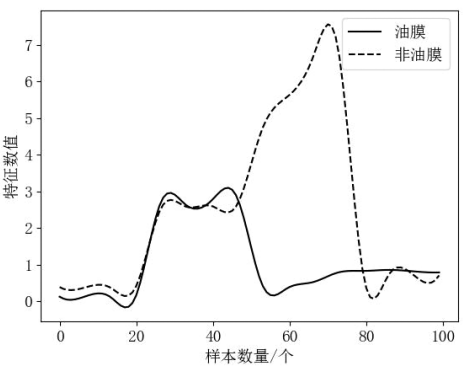
(b)熵



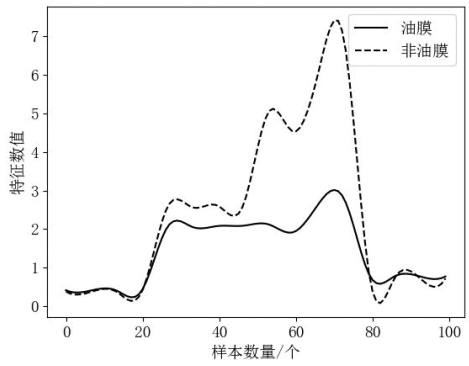
(c)纹理局部相似性



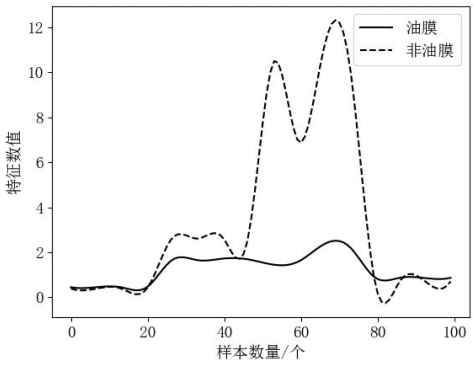
(d)细长度



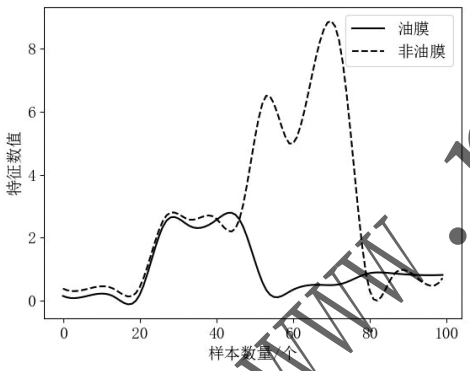
(e)灰度能量



(f)角能量



(g)灰度方差



(h)灰度均值

图 2 mRMR 特征选择结果

Fig. 2 mRMR feature selection results

3.2 自适应半监督决策树实验

利用 SAR 图像进行海面溢油区域检测过程中，无标签的雷达图像样本大量存在，而有标签样本却很难获得，因此需要采用半监督学习算法。为了模拟标签数据较少的场景，仅使用 5.0%、7.5%、10.0%、15.0%和 20.0%的标签样本比例作为训练样本，不同标签样本比例下自适应置信度半监督决策树模型的识别准确率比决策树模型分别高 13.6%、7.5%、5.8%、5.6%和 5.4%。不同标签样本比例下的识别准确率如表 2 所示。

表2 不同标签样本比例下的识别准确率

Tab.2 Recognition accuracy under different label sample ratios

标签样本比例/%	决策树识别准确率/%	本文模型识别准确率/%
5.0	64.3	77.9
7.5	71.7	79.2
10.0	73.6	79.4
15.0	74.7	80.3
20.0	78.5	83.9

在 5.0%、7.5%、10.0%、15.0%和 20.0%的标签样本比例下，自适应半监督决策树的最终模型识别准确率分别为 77.9%、79.2%、79.4%、80.3%和 83.9%，说明利用自适应半监督决策树多次自训练后，基本能够学习到油膜的特征并识别出油膜图像。

表 3 中，经过 mRMR 特征选择后，本文模型在不同标签样本比例下都有最好的表现，并且识别准确率进一步提升。在 5.0%标签样本比例下，本文模型的识别准确率比 SVM 高 30.8%，比决策树提高 22.4%；在 7.5%标签样本比例下，本文模型的识别准确率比 SVM 和决策树分别提升了 31.1%和 15.7%；在 10.0%标签样本比例下，本文模型的识别准确率比 SVM 和决策树分别提升了 22.7%和 15.1%；在 15.0%标签样本比例下，本文模型的识别准确率比 SVM 和决策树分别提升了 24.1%和 15.6%；在 20.0%标签样本比例下，本文模型的识别准确率比 SVM 和决策树分别提升了 22.4%和 12.3%。以上数据说明，在标签数据较少的情况下，自适应半监督决策树通过挖掘无标签数据中的信息，获得了性能更好的分类模型。相比监督学习分类模型 SVM 和决策树的识别准确率，本文模型在不同标签样本比例下平均提升了 26.22%和 16.22%。

表3 mRMR 特征选择后不同标签样本比例下的识别准确率

Tab.3 Recognition accuracy under different label sample ratios after mRMR feature selection

mRMR 特征选择	监督学习模型		半监督	不同标签样本比例的识别准确率/%					
	SVM	决策树	本文模型	5.0%	7.5%	10.0%	15.0%	20.0%	100.0%
√	√			55.9	56.3	66.0	66.2	68.4	78.1
√		√		64.3	71.7	73.6	74.7	78.5	91.7
√			√	**86.7**	**87.4**	**88.7**	**90.3**	**90.8**	91.7

(下转第 62 页)