https://doi.org/10.1051/jnwpu/20203830471

一种水下目标识别的最大信息系数特征选择方法

张牧行[1], 申晓红[1], 何磊[1], 王海燕[1,2]

(1.西北工业大学 航海学院, 陕西 西安 710072; 2.陕西科技大学 电子信息与人工智能学院, 陕西 西安 710021)

摘 要:由于未经选择的特征集合中包含的无关特征和冗余特征会导致识别性能和识别效率的下降,特征选择是识别任务中的重要步骤。然而,基于辐射噪声识别水下目标时,由于目标的多样性和水声信道的复杂性,提取的声学特征之间存在多种线性相关之外的复杂关系。针对此问题,以归一化最大信息系数度量特征与类别之间的相关度以及特征之间的冗余度,提出了基于归一化最大信息系数的特征选择方法(NMIC-FS),并在实测数据集上以随机森林和支持向量机等模型估计的平均分类精度评估其性能。水下目标数据分析结果表明,与未选择前相比,NMIC-FS 所得特征子集性能在更短的分类时间得到更高的分类正确率。与相关特征选择法、拉普拉斯分数法和套索法等方法相比,NMIC-FS 在特征选择过程中能更迅速地提升分类正确率,可用更少的特征得到与使用特征全集时相当的分类正确率。

关 键 词:特征选择;舰船辐射噪声;最大相关系数

中图分类号:TP391.4 **文献标志码**:A **文章编号**:1000-2758(2020)03-0471-07

近年来,研究人员提出了多种声学特征提取技术,以提高由目标辐射噪声特性识别舰船的能力。然而,不同应用场景下这些特征具有不同的有效性,甚至存在无效和冗余的特征[1]。直接使用所有特征进行识别分类,不仅会消耗额外的存储空间,增加模型训练成本[2],还会因模型复杂度过高而降低识别正确率。因此,特征降维是识别分类系统中的重要环节。

特征降维可分为特征变换与特征选择 2 类[4]。特征变换的目的是去除特征间的冗余,如采用主成分分析(principal components analysis, PCA)等[5]。基于相关分析,特征变换能够在保留原特征主要信息的前提下得到其低维表示形式。然而,特征降维方法仅考虑了特征之间的冗余,忽略了特征与类别间的关联,因此无法量化各个特征的有效性。此外,变换后的低维特征丢失了原始特征的物理含义,不利于研究人员的理解和分析[4]。

特征选择的关键是构建有效的特征子集评价指标,包括基于分类性能的评价指标以及基于数据本身统计特性的评价指标等。据此,特征选择可以分为封装模型、嵌入模型和过滤模型 3 类[6]。封装模型基于特征子集的分类性能进行特征选择。例如,文献[7]利用支持向量分类器的结果,得到了具有较高分类正确率的特征子集。但受限于分类器的训练过程,封装模型存在运算量大,计算时间长的问题。为了兼顾分类性能和特征选择效率,嵌入模型在构建分类器的过程中进行特征选择。文献[8]提出了最小绝对值收敛和选择算子法(least absolute shrinkage and selection operator, LASSO),通过在线性分类器中引入范数正则化项对特征权重进行稀疏化,然后选择权重系数大于 0 的特征。在此基础上,研究人员提出了多种基于 LASSO 的特征选择方法[9-10],能够同时进行分类器构建和特征选择,具有良好的稳定性。但是,LASSO 方法在处理高维数据时计算量较大,容易产生过拟合[11]。此外,由于 LASSO 方法所选特征子集依赖于分类器,泛化能力较差[4]。为了得到通用性更强的特征子集,过滤模型采用独立于分类器的方法,仅依据数据本身特性评估特征或特征子集。典型过滤模型包括 Fisher 得分法[12-13]、拉普拉斯分数法[14-15](Laplacian score,

收稿日期:2019-07-01 **基金项目**:国家重点研发计划(2016YFC1400204)资助

作者简介:张牧行(1985—),女,西北工业大学博士研究生,主要从事特征选择和目标识别研究。

LS) 以及相关特征选择法[16-17] (correlation-based feature selection, CFS)。Fisher 得分法对每个特征打分,使得在高分特征构建的空间里,同类别样本点间距尽可能小,而异类样本点间距尽可能大。LS 能够依据样本空间的局部几何结构对特征子集进行评估。然而,上述 2 种方法均无法去除特征集合中的冗余特征[18],即与类别无关的特征或与已选特征提供相似信息的特征[19]。基于特征与类别高度相关且特征间低冗余的原则,Hall 提出了基于相关度量的 CFS。该方法利用 Pearson 系数,从特征与类别以及特征相互之间的相关性对特征进行综合评价。但由于使用线性相关度量准则,此方法不适用于分析特征间的非线性关系。

在对水下目标辐射噪声进行特征选择时,由于目标的多样性和水声信道的复杂性,其声学特征之间可能存在多种线性相关之外的复杂关系。针对此问题,本文提出一种基于归一化最大信息系数的特征选择方法 (normalized maximal information coefficient feature selection, NMIC-FS)。通过最大信息系数度量特征间的线性及非线性关系,该方法能够全面评估特征子集的相关度、冗余度;同时,结合前向序列特征搜寻策略,该方法能够快速搜寻特征子集空间,具有较高的计算效率。本文将所提方法应用于实测 ShipsEar[20] 舰船辐射噪声数据集。实验结果表明所提方法可在保持较高分类性能的前提下去除无关、冗余的特征;同时,与其他特征选择方法的比较结果表明,该方法能够以较小规模的特征子集得到更高的分类正确率。

1 最大信息系数理论

最大信息系数[21-22] (maximal information coefficient, MIC) 以 2 个随机变量间的联合概率密度度量其相关程度。MIC 不仅可以度量随机变量之间的线性关系,还可以度量随机变量之间的非线性关系以及广义的非函数关系,从而可以挖掘出随机变量之间的深层联系。

由于实际应用中通常无法直接获得随机变量的联合密度函数,需要由样本估计其经验联合概率密度函数。对于二维联合随机变量 (X,Y),其样本集合记为 $D=\{(x_i,y_i)\mid i=1,2\cdots N\}$。其中,$N$ 为样本容量。通过分别将 X 和 Y 的值域划分为 m 和 n 个不同的区间,可将样本空间离散化为 $m\times n$ 的网格 G。在指定的网格 G 下,经验联合概率密度和经验边缘概率密度可分别由各格子中的样本数目和区间内的样本数目在样本容量中的占比估计,并可进一步估计出互信息 $I(D\mid_G)$

$$I(D\mid_G)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (1)$$

式中:$D\mid_G$ 表示使用网格 G 划分样本集合 D 时引入的概率分布;$p(x)$ 和 $p(y)$ 分别是 X 和 Y 的经验边缘概率密度;$p(x,y)$ 是 X 和 Y 的经验联合概率密度。

在离散化样本集合 D 时,同样的网络规模 $m\times n$ 下可能存在多种不同的值域划分方式。所有可能的网格 G 上的最大互信息记为

$$I^*(D,m,n)=\max_G I(D\mid_G) \quad (2)$$

为了比较不同规模网格 G 上的最大互信息,将 $I^*(D,m,n)$ 进一步标准化如下

$$M(D)_{m,n}=\frac{I^*(D,m,n)}{\log_2\min\{m,n\}} \quad (3)$$

基于最大互信息,随机变量 X 和 Y 之间的 MIC 定义为

$$\mathrm{MIC}(X,Y)=\max_{mn<B(N)}\{M(D)_{m,n}\} \quad (4)$$

式中,$B(N)$ 为样本个数的函数,通常设置 $B(N)=N^{0.6}$[21]。$\mathrm{MIC}(X,Y)$ 的取值范围在 $[0,1]$。

2 归一化最大信息系数特征选择方法

归一化最大信息系数特征选择方法 (NMIC-FS) 主要由特征子集评价指标与特征子集搜索策略两部分构成。本节分别介绍该方法中基于归一化 MIC 的特征子集评价指标以及启发式的前向序列特征子集搜索策略。

2.1 基于归一化 MIC 的特征子集评价

特征选择的目的是寻找一个最优特征子集,使该集合中的特征与类别高度相关,同时特征间互不相关。记 $\{x_i,c_i\mid i=1,2\cdots N\}$ 是样本容量为 N,特征全集为 $F=\{f_1,f_2\cdots f_M\}$ 的数据集,其中 x_i 为具有 M 维特征的样本,c_i 为对应的类别,将 N 个样本的类别记作类别变量 c。本文分别使用 $\mathrm{MIC}(f_i,c)$ 和 $\mathrm{MIC}(f_i,f_j)$ 估计特征 f_i 与类别 c 的相关度以及特征 f_i 和 f_j 之间的冗余度。$\mathrm{MIC}(f_i,c)$ 的取值越接近于 1 表明该特征与类别的关联度越高,反之两者相关程度越弱。同样,$\mathrm{MIC}(f_i,f_j)$ 取值越高说明 f_i 和 f_j 之间的冗余度越高,相互可替代性越强。

为了提高特征选择性能,本文依据相关度以及冗余度这 2 个指标优化特征子集,使得该子集中的特征与类别高度相关,同时各个特征之间互不相关。然而,对于不同的数据集,$\mathrm{MIC}(f_i,c)$ 和 $\mathrm{MIC}(f_i,f_j)$ 的取值区间不同。若直接使用 $\mathrm{MIC}(f_i,c)$ 和 $\mathrm{MIC}(f_i,f_j)$ 估计相关度和冗余度,则取值较小者会被淹没,从而使特征子集的评价指标出现较大的估计偏差。为了消除这一影响,本文定义归一化缩放的度量特征 f_i 与类别 c 相关度

$$R_a(f_i,c)=\frac{\mathrm{MIC}(f_i,c)-\min\limits_{f_k\in F}\mathrm{MIC}(f_k,c)}{\max\limits_{f_k\in F}\mathrm{MIC}(f_k,c)-\min\limits_{f_k\in F}\mathrm{MIC}(f_k,c)} \tag{5}$$

和度量特征 f_i 与 f_j 之间冗余度

$$R_b(f_i,f_j)=\frac{\mathrm{MIC}(f_i,f_j)-\min\limits_{f_k,f_l\in F}\mathrm{MIC}(f_k,f_l)}{\max\limits_{f_k,f_l\in F}\mathrm{MIC}(f_k,f_l)-\min\limits_{f_k,f_l\in F}\mathrm{MIC}(f_k,f_l)} \tag{6}$$

式中,$k\neq l$。

采用归一化 MIC 对特征子集 $S\subset F$ 进行评价,则特征与类别之间的归一化相关度、2 个任意特征 f_i 和 f_j 之间的归一化冗余度可分别表示为

$$R_{a'}(S)=\frac{1}{s}\sum_{f_i\in S}R_a(f_i,c) \tag{7}$$

$$R_{b'}(S)=\frac{1}{s(s-1)}\sum_{\substack{f_i,f_j\in S,\\ i\neq j}}R_b(f_i,f_j) \tag{8}$$

式中,$s=|S|$,$f_i,f_j\in S$ 且 $i\neq j$。

为了使所选特征子集中特征与类别高度相关,而各特征之间互不相关,需要综合考虑相关度和冗余度这 2 个指标。据此,综合(7)式和(8)式,可得特征子集 S 的整体评价指标

$$J(S)=R_{a'}(S)-R_{b'}(S) \tag{9}$$

该指标对特征与类别间的相关程度和备选特征子集冗余度进行整体评估。所选特征子集保留了与类别相关度高的特征,降低了特征间的冗余度,能够在提高分类正确率的同时降低计算复杂度。

2.2　NMIC-FS 特征选择策略

特征选择策略是特征选择的另一个重要构成,即如何从特征全集中找出可使性能评价指标最大化的特征子集。特征选择策略可采用遍历搜索、随机搜索和启发式搜索等。为了能够以较低的运算成本获得较高的性能评价指标,本文采用启发式的前向序列搜索策略。

该策略的基本思想是从剩余特征集中逐个选取特征并与已选特征组合成备选特征子集,然后选择可使备选特征子集性能指标最大化的特征。记 S 为已选特征子集,$R=F\backslash S$ 为剩余特征集。算法开始时 S 为空集,无需考虑特征间的冗余度。此时可选择与类别相关度最高的特征,并将其加入 S,即

$$S=\varnothing\cup\operatorname*{argmax}_{f_i\in F}(\mathrm{MIC}(f_i,c)) \tag{10}$$

之后,逐个从 R 中选出使(9)式最大的特征,并将其加入 S,即

$$S=S\cup\operatorname*{argmax}_{f_i\in R}J(S') \tag{11}$$

式中,备选特征子集 $S'=S\cup f_i$。

NMIC-FS 整体流程如下:

输入:特征全集 $F=\{f_1,f_2\cdots f_M\}$,设已选特征子集 $S=\varnothing$,剩余特征集为 $R=F$

输出:优化特征子集 $S=\{f'_1,f'_2\cdots f'_k\}$

1)　$\forall f\in F$,由(10)式更新已选特征子集 S,$R=F\backslash S$;

2)　$\forall f\in R$,$S'=S\cup f$,由(11)式更新已选特征子集 S,$R=F\backslash S$;

重复步骤 2 直到 $|S|=M$ 或 $J(S)$ 达到最大值。

3　实验数据分析

3.1　数据及其分析流程

1)目标辐射噪声数据

本节将所提 NMIC-FS 应用于 ShipsEar 数据集[20]。该数据集采集于西班牙大西洋沿岸的 Vigo 港口附近区域。水听器的采样频率为 52 734 Hz,并前置了一个截止频率为 100 Hz 的高通滤波器,以降低浅水环境噪声的影响。录音中包括了 11 种不同的船舶类型和背景噪声。船只类型有渔船、远洋班轮、渡轮、拖船、摩托艇、游艇、小帆船等。

由于舰船辐射噪声的多样性以及水声环境的复杂性,该数据集有效的声学特征难以直接确定。因此,本实验提取尽可能多的声学构成原始特征集合,然后使用 NMIC-FS 进行特征选择。舰船辐射噪声的主要特征集中在 3 kHz 以下的低频部分。因此,特征提取前需要将原信号降采样至 6 kHz,以使提取的特征更有代表性。将每段信号提取的声学特征以向量形式构成样本,则从该数据集中可得样本数 2 785。每个样本中各个特征的物理含义如表 1 所示。

表 1 水下目标样本特征说明

特征名称	特征编号
过零率	1
能量	2
能量熵	3
谱质心	4
谱延展	5
谱熵	6
谱通量	7
谱滚降	8
Mel 频率倒谱系数	9~21
谐波比	22
基频	23
音色向量	24~35
一阶差分 Mel 频率倒谱系数	36~48
二阶差分 Mel 频率倒谱系数	49~61

2) 数据分析流程

为了验证 NMIC-FS 在水下目标识别中的有效性和优越性,本文分别用 LS、CFS、LASSO 以及 NMIC-FS 对提取的特征集合进行选择,并分别将支持向量机(support vector machine,SVM)和随机森林(random forest,RF)作为分类模型,以评估各方法所得特征子集的分类能力。

实验包括两部分:①NMIC-FS 选择前后 SVM 和 RF 2 种分类模型所得分类性能的对比,目的是验证 NMIC-FS 所得的特征子集是否能够提升实测水下目标辐射噪声的分类性能。②LS、CFS、LASSO 和 NMIC-FS 4 种特征选择方法的比较,目的是验证 NMIC-FS 是否能更有效地进行特征选择,即更快获得具有更高分类正确率、更小规模的特征子集。

3.2 NMIC-FS 特征选择前后分类效果比较

表 2 和表 3 分别展示了以 SVM 和 RF 作为分类模型时特征选择前后的分类正确率和分类时间。其中,为了准确评估各特征子集的分类能力,本文采用 10 次十折交叉验证估计平均分类正确率,以降低因训练/测试样本划分不同而引入偏差。具体步骤是:在 1 次十折交叉验证中,首先将所有样本随机划分为 10 等份,然后分别将其中 1 份作为测试集,剩余 9 份作为训练集,可得 10 个不同的测试分类正确率。重复 10 次上述过程,将所有测试分类正确率的均值作为分类正确率的最终估计值。

依据 NMIC-FS 的特征选择结果,当特征子集规模为 26 维时,以 SVM 为分类模型的分类正确率达到最大值 81.2%,相应的分类时间为 3.13 s。与选择前 61 维特征集合相比,NMIC-FS 得到的特征子集分类时间更短,且能够达到更高的分类正确率,在保持较高分类性能的前提下去除无关、冗余的特征,如表 2 所示。

表 2 特征选择前后的 SVM 分类结果

状态	特征集规模	分类正确率/%	分类时间/s
选择前	61 维	78.3	6.38
选择后	26 维	81.2	3.13

当特征子集规模为 30 维时,以 RF 为分类模型的分类正确率达到最大值 82.4%,相应的分类时间为 6.61 s。与特征选择前 61 维特征集合相比,NMIC-FS 同样显著缩短了分类时间、提高了分类正确率,如表 3 所示。

表 3 特征选择前后的 RF 分类结果

状态	特征集规模	分类正确率/%	分类时间/s
选择前	61 维	79.0	9.02
选择后	30 维	82.4	6.61

3.3 NMIC-FS、LS、CFS 和 LASSO 特征选择比较实验

使用 NMIC-FS 对上述舰船辐射噪声的 61 维声学特征进行选择,并从样本空间分布和特征选择过程 2 个方面与 LS、CFS 和 LASSO 进行对比。

3.3.1 最佳二维特征的样本空间分布

为了比较 LS、CFS、LASSO 和 NMIC-FS 的特征选择效果,由各方法得到的最佳二维特征构成样本空间,并以可视化方式观察样本的空间分布,如图 1 所示。图中特征 1 和特征 2 分别表示各方法所得的第 1 最佳特征和第 2 最佳特征,2 类样本分别来自于滚装船和摩托艇。由 2 类样本的空间分布可看到,NMIC-FS 获得的最佳二维特征可使 2 类样本间的重叠面积最小,具有更强的可分性,如图 1d)所示。

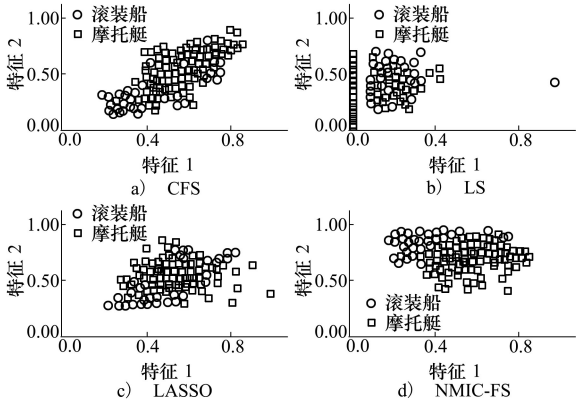

图 1　4 种方法最佳特征滚装船和摩托艇样本空间分布图

3.3.2　特征选择过程对比

为了比较 LS、CFS、LASSO 和 NMIC-FS 搜寻优化特征子集的能力,本文基于各特征选择方法估计的特征重要性,按由高至低的次序将特征逐个加入特征子集,然后使用分类正确率评估特征子集的分类性能。依据分类正确率随特征子集规模增加的变化趋势,对各特征选择方法进行分析比较。以分类正确率作为特征子集性能指标时,不同的训练/测试样本划分方式以及分类模型的选择均会影响分类正确率的估计。在本实验中,同样以 10 次十折交叉验证的方式估计特征子集的平均分类正确率。同时,实验中分别采用 SVM 和 RF 作为分类模型评估特征子集的分类能力。

图 2 展示了以 SVM 为分类模型时 4 种方法的特征选择过程。由分类正确率的变化趋势可以看到,当特征子集规模逐渐增加时,NMIC-FS 能够以最快的速度提升分类正确率。当特征子集规模为 19 时,其平均分类正确率为 78.0%,与使用全部特征能够获得的分类正确率相当。当特征子集规模达到 26 时,平均分类正确率达到 81.2%并趋于稳定,高于使用全部特征时的分类正确率。此时,所选特征子集中已包含可用于舰船辐射噪声分类的所有相关信息。当特征子集规模继续增大时,平均分类正确率有所下降,这是因为分类模型的复杂度随特征维数增加,但并没有引入新的有效信息。总体来讲,同 CFS、LASSO 以及 LS 相比,NMIC-FS 能够以更快的速度提升特征子集的分类正确率,表明其能够从原有特征集合中快速搜寻到更优的特征子集。同时,NMIC-FS 能够以更少的特征获得等于甚至优于使用全部特征时的分类性能。

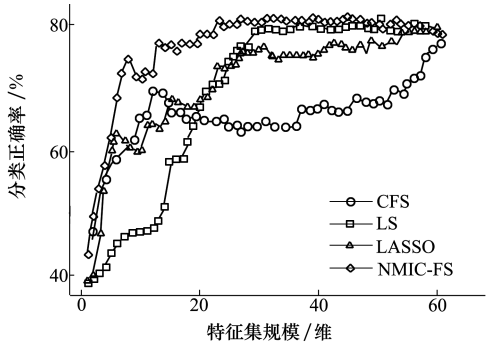

图 2　基于 SVM 分类性能的特征选择过程

图 3 中展示了以 RF 为分类模型时 4 种方法的特征选择过程。NMIC-FS 可在特征子集规模为 11 时达到与使用全部特征时相当的分类正确率 79%。此外,在特征子集规模为 30 时达到最高分类正确率 82.4%。同 CFS、LASSO、LS 相比,在特征子集规模相同的情况下,NMIC-FS 的特征子集可得到更高的分类正确率;在分类性能相同的情况下,NMIC-FS 所需特征子集规模更小。

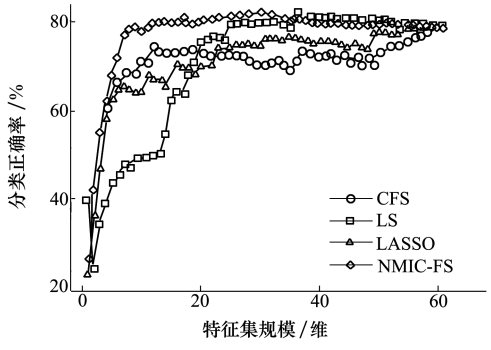

图 3　基于 RF 分类性能的特征选择过程

比较图 2 和图 3,可以看到 4 种方法在 SVM 模型和 RF 模型下的特征选择过程具有明显差异。不同分类模型下特征子集的分类性能随子集规模增加时的变化趋势不同,且 RF 模型获得的分类正确率高于 SVM 模型获得的分类正确率。这种差异表明特征子集的分类性能与分类模型有关。此外,在 SVM 模型和 RF 模型下,NMIC-FS 均能找到优于其他 3 种方法的特征子集,对这 2 种分类模型具有良好的适应性。

4　结　论

在基于辐射噪声进行目标识别时,声学特征中

的无关、冗余数据不仅会降低系统的分类性能,还会增加系统的运算负担和存储成本。基于归一化的最大信息系数,本文提出的 NMIC-FS 能够综合度量特征与类别的相关性和特征间的冗余性,并结合前向顺序搜索策略实现了快速特征选择。NMIC-FS 在 ShipsEar 实测数据集上的应用结果表明,该方法能在保持较高分类正确率的前提下大幅减少分类所需的特征数目,提高识别系统的效率。与 CFS、LASSO 和 LS 的比较表明,NMIC-FS 在 SVM、RF 两种不同的分类模型下的均能够以更少的特征表征原有特征集,且特征子集的分类正确率随着子集规模上升最快,证实了方法的有效性和实用性。

参考文献:

[1] 杨宏晖, 戴健, 孙进才, 等. 用于水声目标识别的自适应免疫特征选择算法[J]. 西安交通大学学报, 2011, 45(12): 28-33
YANG Honghui, DAI Jian, SUN Jincai, et al. A New Adaptive Immune Feature Selection Algorithm for Underwater Acoustic Target Classification[J]. Journal of Xi'an Jiaotong University, 2011, 45(12): 28-33 (in Chinese)

[2] ALELYANI S, TANG J, LIU H. Feature Selection for Clustering: a Review[M]. New York: CRC Press, 2014

[3] YU L, LIU H. Efficient Feature Selection via Analysis of Relevance and Redundancy[J]. Journal of Machine Learning Research, 2004, 5: 1205-1224

[4] TANG J, ALELYANI S, LIU H. Feature Selection for Classification: a Review[M]. New York: CRC Press, 2014

[5] ABDI H, WILLIAMS L J. Principal Component Analysis[J]. WIREs Computational Statistics, 2010, 2(4): 433-459

[6] GUYON I, ELISSEEFF A. An Introduction to Variable and Feature Selection[J]. Journal of Machine Learning Research, 2003, 3: 1157-1182

[7] 杨宏晖, 孙进才, 袁骏. 基于支持向量机和遗传算法的水下目标特征选择算法[J]. 西北工业大学学报, 2005, 23(4): 512-515
YANG Honghui, SUN Jincai, YUAN Hong. A New Method for Feature Selection for Underwater Acoustic Targets[J]. Journal of Northwestern Polytechnical University, 2005, 23(4): 512-515 (in Chinese)

[8] TIBSHIRANI R. Regression Shrinkage and Selection via the Lasso: a Retrospective[J]. Journal of the Royal Statistical Society: Series B(Statistical Methodology), 2011, 73(3): 273-282

[9] ZOU H. The Adaptive Lasso and Its Oracle Properties[J]. Journal of the American Statistical Association, 2006, 101(476): 1418-1429

[10] ZOU H, HASTIE T. Regularization and Variable Selection via the Elastic Net[J]. Journal of the Royal Statistical Society: Series B(Statistical Methodology), 2005, 67(2): 301-320

[11] CAI J, LUO J, WANG S, et al. Feature Selection in Machine Learning: a New Perspective[J]. Neurocomputing, 2018, 300: 70-79

[12] GU Q, LI Z, HAN J. Generalized Fisher Score for Feature Selection[C]//Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Arlington, Virginia, 2011: 266-273

[13] ZHOU M. A Hybrid Feature Selection Method Based on Fisher Score and Genetic Algorithm[J]. Journal of Mathematical Sciences: Advances and Applications, 2016, 37(1): 51-78

[14] HE X, CAI D, NIYOGI P. Laplacian Score for Feature Selection[C]//Proceedings of the 18th International Conference on Neural Information Processing Systems, Cambridge, MA, 2005: 507-514

[15] HUANG R, JIANG W, SUN G. Manifold-Based Constraint Laplacian Score for Multi-Label Feature Selection[J]. Pattern Recognition Letters, 2018, 112: 346-352

[16] HALL M A. Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning[C]//Proceedings of the Seventeenth International Conference on Machine Learning, San Francisco, CA, 2000: 359-366

[17] MURSALIN M, ZHANG Y, CHEN Y, et al. Automated Epileptic Seizure Detection Using Improved Correlation-Based Feature Selection with Random Forest Classifier[J]. Neurocomputing, 2017, 241: 204-214

[18] ZHAO Z, WANG L, LIU H, et al. On Similarity Preserving Feature Selection[J]. IEEE Trans on Knowledge and Data Engineering, 2013, 25(3): 619-632

[19] HU L, GAO W, ZHAO K, et al. Feature Selection Considering Two Types of Feature Relevancy and Feature Interdependency [J]. Expert Systems with Applications, 2018, 93: 423-434

[20] SANTOS-DOMíNGUEZ D, TORRES-GUIJARRO S, CARDENAL-LóPEZ A, et al. ShipsEar: an Underwater Vessel Noise Database[J]. Applied Acoustics, 2016, 113: 64-69

[21] RESHEF D N, RESHEF Y A, FINUCANE H K, et al. Detecting Novel Associations in Large Data Sets[J]. Science, 2011, 334 (6062): 1518-1524

[22] RESHEF Y A, RESHEF D N, FINUCANE H K. Measuring Dependence Powerfully and Equitably[J]. Journal of Machine Learning Research, 2016, 17(211): 1-63

Feature Selection on Maximum Information Coefficient for Underwater Target Recognition

ZHANG Muhang[1], SHEN Xiaohong[1], HE Lei[1], WANG Haiyan[1,2]

(1.School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China;
2.School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China)

Abstract: Feature selection is an essential process in the identification task because the irrelevant and redundant features contained in the unselected feature set can reduce both the performance and efficiency of recognition. However, when identifying the underwater targets based on their radiated noise, the diversity of targets, and the complexity of underwater acoustic channels introduce various complex relationships among the extracted acoustic features. For this problem, this paper employs the normalized maximum information coefficient (NMIC) to measure the correlations between features and categories and the redundancy among different features and further proposes an NMIC based feature selection method (NMIC-FS). Then, on the real-world dataset, the average classification accuracy estimated by models such as random forest and support vector machine is used to evaluate the performance of the NMIC-FS. The analysis results show that the feature subset obtained by NMIC-FS can achieve higher classification accuracy in a shorter time than that without selection. Compared with correlation-based feature selection, laplacian score, and lasso methods, the NMIC-FS improves the classification accuracy faster in the process of feature selection and requires the least acoustic features to obtain classification accuracy comparable to that of the full feature set.

Keywords: feature selection; ship-radiated noise; maximum correlation coefficient