DOI 编码：10.13646/j.cnki.42-1395/u.2020.05.028

航标运行状态模式识别和数值预测

陈麒龙[1]，陆一军[2]

（1. 中国人民大学，北京 100872；2. 交通运输部东海航海保障中心，上海 200086）

摘　要：针对航标运行状态模式识别依赖经验阈值的现状，为检验经验阈值是否具有普适性，提出基于概率的阈值模式识别效率度量算法。实验结果表明：该算法能准确度量阈值的模式识别效率；经检验，经验阈值不具备普适性。因而，提出基于概率的模式识别模型。实验结果表明：以概率作为阈值具有普适性，该模型能准确识别频繁模式和异常模式，且性能更好。为实现数值预测，提出基于概率密度的加权平均算法。实验结果表明：该算法的预测精度较高。本文为航标运行状态模式识别和数值预测提出了新的解决方案。

关键词：水路运输；航标；概率；模式识别；数值预测

中图分类号：U644.8　　**文献标识码**：A　　**文章编号**：1006—7973（2020）05-0069-05

航标遥测数据是反映航标运行状态的数值信息，包括：数据采集时间（Time）、电压（Voltage）、电流（Current）、航标位置（Longitude、Latitude）、离位距离（Distance）。频繁模式表示航标的“常态”，异常模式表示航标的“非常态”。对频繁模式和异常模式的识别，传统方法是依据经验阈值进行分类，存在主观臆断的问题。对航标运行状态的数值预测，目前仍处于研究阶段。如何检验经验阈值是否具有普适性，如何实现航标运行状态的数值预测，是亟待解决的问题。

对数据的频繁模式和异常模式的模式识别，已有不少算法和模型，如：基于相关性度量算法、基于频繁子树算法、基于最大熵隐马尔科夫模型，以及基于统计特征的支持向量机[1-4]。移动对象位置预测的模型有：马尔科夫模型、高斯混合模型、卷积神经网络模型[5-7]。核密度估计（kernel density estimation，KDE）是一种估计数据的概率密度函数（probability density function，PDF）的算法，利用概率密度函数可以计算出给定数值区间的概率。概率可以用来度量经验阈值的模式识别效率，以此来检验经验阈值是否有效，判定经验阈值是否具有普适性。概率反映随机事件发生的可能性，是客观的，以概率作为阈值进行分类，就是将“大概率”的数据作为“常态”，将“小概率”的数据作为“非常态”，从而使阈值成

后再加密传输，期间要采用不同的密钥。对于网络数据传输安全，节点加密效果非常好，但是攻击者对于通信业务的分析采用节点解密形式予以防范却是存在缺陷的。

（3）端端加密。可以将其定义为包加密或者脱线加密，从源点到终点，在传输数据过程中允许以密文形式存在。在到达终点之前，不进行端端加密消息的解密，主要就是由于整个传输过程中，这些消息均受到保护。通常，在对敏感信息进行传输过程中必须应用端端加密形式。

最有效的网络安全技术之一就是密码技术。依托网络进行加密操作，对于非授权用户的入网或者搭线窃听可以有效地防止，这也是一种对恶意软件进行防范的有效措施。

4.3 数据备份策略

施桥船闸目前对信息化的依赖比较高，服务器及数据的稳定性被给予很高的要求，实现这样的目的，除了采购质量良好的硬件设施外，还可以有效地利用私有云实现信息备份和灾难冗余。对于需要高度可靠性的用户，这样的方式可以使文件信息的安全有保障，正常情况本机存储，私有云自动备份，当工作设备的系统出现故障时，备份设备可以随时异地提供数据读取，保证单位整体的正常运转。同时使用 UPS 对重要设备进行不间断的供电，保证硬件设施的良好运行。

5 结语

近几年，互联网技术获得跨越式发展，在这一过程中伴随而生的巨大问题就是网络安全问题。这是一个普遍存在、覆盖面广的复杂性问题，同时还会涉及到违法犯罪等活动。而在涉及到以下简单的网络安全问题时，仅仅确保无关人员无法完成读取信息，或者不能对传输的信息进行修改。网络安全问题中，部分对象无权使用网络，但是却试图借助一些软件来实现远程服务，窃取一些信息。对于合法消息重播和截获问题也是安全性处理的对象。

本文从建设智慧船闸角度对于解决基础网络设施安全进行了描述，旨在为单位提供信息的完整性、认证性、保密性的保护机制，避免网络系统、数据、服务遭到破坏或者侵扰。现在较为普遍应用的方法有加密技术、认证技术、防火墙等，由于越来越大的运行规模，使得单位网络涉及的安全性问题呈现复杂化特征。因此，维护网络安全将是一件关键任务。在此过程中要对安全因素综合考虑，将有关的安全防范技术进行相互结合，采取科学有效的安全防范措施，保证网络安全。

参考文献：

[1] 王加雪，钱江 . 智慧船闸 [M]. 东南大学出版社 ,2018

[2] [美] 本・斯派维 乔伊・爱彻利维亚 . Hadoop 安全大数据平台的隐私保护 [M]. 人民邮电出版社，2017

[3] 刘化君 . 网络安全与管理 [M]. 电子工业出版社 ,2019.

[4]（美）Saadat Malik 网络安全原理与实践 [M]. 人民邮电出版社 ,2019.

为一种客观的指标，而具有普适性。概率密度与概率是正相关的，将概率密度转化为权重，以加权平均数作为预测值，既消减了极端值的影响，又使预测值趋于“大概率”。相对于相关性度量算法、频繁子树算法、马尔科夫模型、支持向量机、高斯混合模型、卷积神经网络模型等，核密度估计和概率的计算过程更为简单，算法和模型易于解释，且性能良好，适合航标运行状态模式识别和数值预测。

1 经验阈值检验

1.1 核密度估计原理

电压、电流和离位距离都是连续型随机变量。假设连续型随机变量 x 的值 x_1，x_2，x_3，…，x_n 取自连续分布 $p(x)$，在任意点 x 处的核密度估计定义为：$\hat{p}(x)=\frac{1}{nh}\Sigma_{i=1}^{n}k(u)$；$\hat{p}(x)$ 称为概率密度函数，$\hat{p}(x)$ 非负且积分为 1；$k(u)$ 称为核函数，$u=\frac{x-x_i}{h}$，h 称为带宽；本文的实验使用的是高斯核，即：$k(u)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$，最优带宽 $h=1.06\sigma n^{-1/5}$。由于连续型随机变量某一确定值的概率必然为零，只能积分计算随机变量某一区间的概率。因此，对于连续型随机变量 x，x 取值范围内的任意 a 和 b，有：$P(a<x<b)=\int_a^b p(x)dx$，$P(a<x<b)$ 表示随机变量介于区间 [a,b] 所对应的随机事件的概率，$p(x)$ 为连续型随机变量 x 的概率密度函数。

1.2 实例分析

已知经验阈值：电压 10.8 V，电流 0.09 A，离位距离 150 m。以洋山港主航道的 Y4# 灯浮标连续 60 天凌晨 3 时的航标遥测数据为例（如表 1），检验经验阈值是否有效，是否具有普适性。

电压的概率密度分布如图 1 所示。对电压经验阈值构造区间为（0,10.8]，计算出电压小于或等于 10.8 V 的概率为 0，表明在凌晨 3 时，以“10.8 V”作为电压阈值无法有效识别异常模式，应当增大阈值。当阈值为“13.2 V”时，区间（0,13.2]的概率为 0.0651，表明在该时段，以“13.2 V”作为阈值识别异常模式的效率为 6.51%，识别频繁模式的效率为 93.49%。

电流的概率密度分布如图 2 所示。对电流经验阈值构造区间为 [0,0.09]，计算出电流小于或等于 0.09 A 的概率为 0.0506，表明在凌晨 3 时，以“0.09 A”作为电流阈值，识别异常模式的效率为 5.06%，识别频繁模式的效率为 94.94%，电流经验阈值有效。

离位距离的概率密度分布如图 3 所示。对离位距离经验阈值构造区间为 [150,+ ∞)，计算出离位距离大于或等于 150 m 的概率为 0，表明在凌晨 3 时，以“150 m”作为离位距离阈值，无法有效识别异常模式，应当减小阈值。当阈值为“75 m”时，区间 [75,+ ∞) 的概率为 0.0436，表明在该时段，以“75 m”作为阈值识别异常模式的效率为 4.36%，识别频繁模式的效率为 95.64%。

以上实验表明：

（1）概率可以准确度量阈值的模式识别效率，可以用来检验经验阈值是否有效；

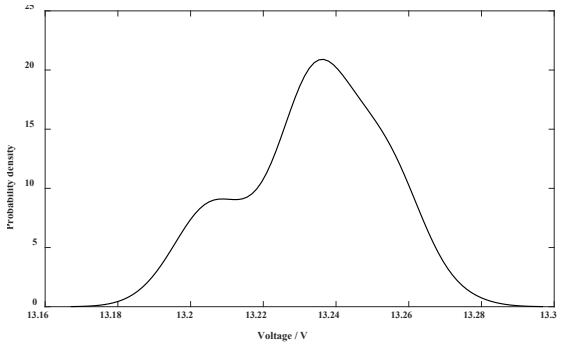

图 1 电压概率密度分布图

表 1 Y4# 灯浮标连续 60 天凌晨 3 时航标遥测数据

Time	Voltage / V	Current / A	Longitude / °	Latitude / °	Distance / m
11/1 3:08	13.272	0.096	122.28241670	30.54305556	8.0
11/2 3:08	13.256	0.096	122.28261110	30.54302778	11.1
11/3 3:08	13.260	0.096	122.28261110	30.54308333	11.1
11/4 3:08	13.256	0.096	122.28258330	30.54300000	10.1
… …	… …	… …	… …	… …	… …
12/27 3:08	13.192	0.100	122.28311110	30.54286111	62.5
12/28 3:08	13.216	0.100	122.28286110	30.54302778	34.8
12/29 3:08	13.244	0.100	122.28250000	30.54308333	3.1
12/30 3:08	13.232	0.092	122.28252780	30.54291667	15.7

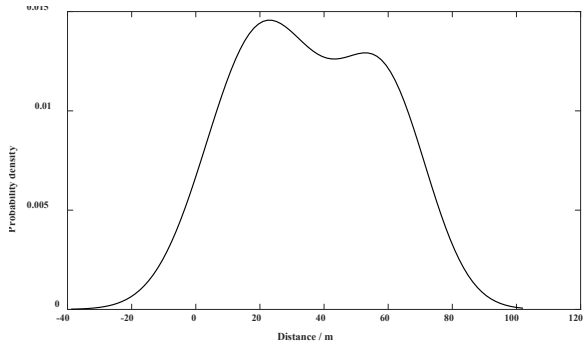

图 2 电流概率密度分布图

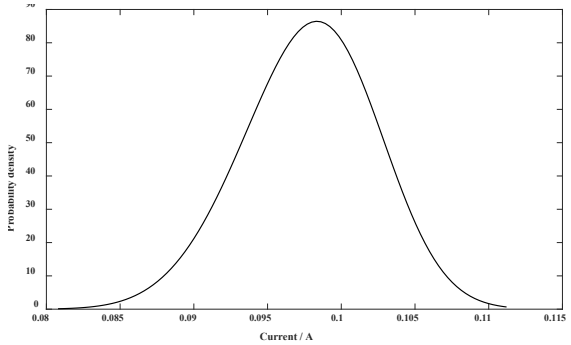

图 3 离位距离概率密度分布图

（2）经验阈值不具有普适性；

（3）利用概率可以找到合适的阈值。

2 模式识别

2.1 基于概率的模式识别原理

基于概率的模式识别的思路是：以理论概率作为阈值，将概率小于理论概率的样本单元作为异常模式，而概率大于理论概率的样本单元作为频繁模式。模式识别流程是：第一步，对样本容量为 n 的样本估计概率密度函数；第二步，以新观测值为中心构造区间；第三步，积分计算区间的概率；第四步，计算理论概率作为阈值，将区间的概率与阈值进行比较和分类。

区间长度应当根据样本数据精度来设置，假设新观测值为 x_i，样本数据的精度为 b，那么区间为：$[x_i-(b/2), x_i+(b/2)]$。阈值 a 的计算公式为：a=b/R，R 表示样本数据的极差，即：R=max(x)− min(x)。阈值的本质是：将样本的值域等间隔划分为 m 个区间，区间长度为 b，样本单元落入某一区间的理论概率，即：m=R/b，a=1/m=b/R。

2.2 实例分析

以洋山港主航道 Y4# 灯浮标“12/31 3:08”的航标遥测数据为例（电压 13.228 V，电流 0.098 A，离位距离 43.6 m）。

电压的数据精度为 0.001，样本数据的极差为 0.08。因此，阈值为 0.0125。新观测值 13.228 的区间为 [13.2275,13.2285]，区间的概率为 0.0171，大于阈值，为频繁模式。

电流的数据精度为 0.001，样本数据的极差为 0.08。因此，阈值为 0.0125。新观测值 0.098 的区间为 [0.0975, 0.0985]，区间的概率为 0.0860，大于阈值，为频繁模式。

离位距离的数据精度为 0.1，样本数据的极差为 63.2。因此，阈值为 0.0016。新观测值 43.6 的区间为 [43.55, 43.65]，区间的概率为 0.0013，小于阈值，为异常模式。

以上实验可以得出结论：

（1）以概率作为阈值，使阈值成为一种客观的指标，具备普适性；

（2）基于概率的模式识别模型能够有效识别频繁模式和异常模式。

2.3 与传统方法比较

传统方法的优点是：直接进行数值对比，计算量小。缺点是：①阈值不具备普适性，如果阈值设置不合理就无法识别异常模式；②阈值设置过程繁琐，为保证阈值有效，需要先度量阈值的模式识别效率，找出合适的阈值；③当灯器设备的规格型号改变时，就必须重新设置电压和电流的阈值；④阈值的模式识别效率需要定期评估，需要定期调整阈值。

新模型的优点是：①以概率作为阈值，具有普适性；②阈值设置简单、灵活可控，可以使用理论概率，也可以使用其他概率；③灯器的型号规格改变时，无需重新设置电压和电流的阈值；④模型易于解释，阈值就是模式识别的效率，对于给定的观测值，阈值越小，分类结果越偏向频繁模式，阈值越大，分类结果越偏向异常模式。缺点是：需要计算概率密度函数和概率，比传统方法的计算量大。

综上所述，新模型的性能比传统方法更好，但是计算量更大。在航标管理上，总是希望发现航标潜在的异常，而且现在的服务器性能完全能够满足新模型的计算需求。因此，推荐使用新模型。

3 数值预测

3.1 基于概率密度的加权平均算法

第一步：以近 60 天同时段的数据为样本，估计样本的概率密度函数，电压和电流的概率密度函数估计详见上文 1.1 节，此处不再赘述。航标位置包括了经度和纬度，是二维向量。二维向量的核密度估计定义为：$\hat{p}(x)=\frac{1}{nh^2}\sum_{i=1}^{n}k(u)$，k(u) 是二维空间的核函数，本文的实验使用的是二维标准正态密度函数，即：$k(u)=(2\pi)^{-1}\exp(-u^{T}u/2)$，$u=\frac{x-x_i}{h}$；h 为带宽，最优带宽 $h=n^{-1/6}$；二维向量 x 的取值落在区域 D 内的概率 $P(x\in D)=\int_D p(x)\,dx$，p(x) 为二维向量 x 的概率密度函数。

第二步：取概率密度峰值和对应的变量值，计算权重。假设概率密度峰值 p_1，p_2，p_3，…，p_n 所对应的变量值为 x_1，x_2，x_3，…，x_n，第 i 个变量值 x_i 的权重 w_i 的计算公式为：

$$w_i=p_i/\sum_i^n p_i$$

第三步：计算加权平均数作为预测值。

计算公式为：$\bar{x}=\sum_i^n x_i w_i$。

3.2 实例分析

已知“12 月 31 日凌晨 3 时”的实测数据：电压 13.228 V、电流 0.098 A、航标位置 (122.28244440°，30.54266667°)。以表 1 的数据为样本，计算“12 月 31 日凌晨 3 时”的预测值及误差，过程数据如表 2 所示。

电压的概率密度是双峰分布（如图 1），预测值为 13.2282，误差为 0.0002；电流的概率密度是单峰分布（如图 2），因此权重为 1，预测值为 0.0983，误差为 0.0003；航标位置的概率密度是多峰分布（如图 4），分别对经度和纬度计算加权平均数，预测值为 (122.28278039°，30.54292107°)，以欧氏距离表示的误差为 0.00042。

3.3 数值预测精度评估

以洋山港主航道 Y4# 灯浮标 12 月 1 日至 12 月 7 日各时段的数值预测为例。实验组：新算法，对照组：中位数。度量指标：均方误差，$MSE=\frac{\sum_{i=1}^{n}(x_i-y_i)^2}{n}$，$x_i$ 是预测值，y_i 是实测值。如表 3 所示，各时段的实验组 MSE 都比较小，表明新算法的预测精度较高；从各时段的 MSE 看，大多数时段的实验组比对照组小，且 MSE 之和，实验组也比对照组小，表明新算法的预测精度优于中位数。

3.4 统计性质分析

样本数据的特性对预测精度的影响体现在：样本数据的方差越小，则 MSE 越小；反之，样本数据的方差越大，则

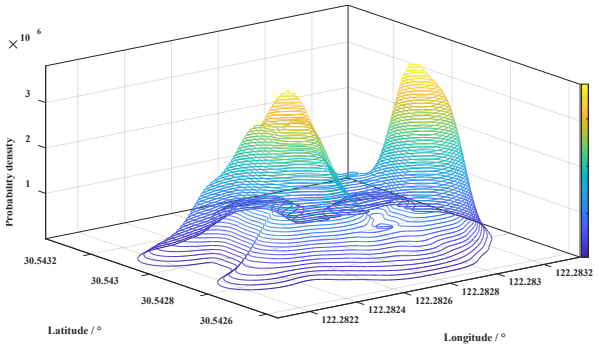

图 4 航标位置概率密度分布图

表 2 数值预测过程数据

	Variable value	Peak of probability density	Weight
Voltage	13.2090	9.1027	0.3034
	13.2366	20.8998	0.6966
Current	0.0983	86.4490	1.0000
Position	122.28270582 , 30.54273635	961004.7709	0.0630
	122.28255386 , 30.54287606	3229882.1765	0.2117
	122.28282738 , 30.54289779	1336222.8824	0.0876
	122.28308065 , 30.54291021	3808913.3070	0.2497
	122.28259438 , 30.54295367	3589009.1924	0.2353
	122.28288817 , 30.54297230	1443149.9071	0.0946
	122.28313637 , 30.54311201	370066.8596	0.0243
	122.28273621 , 30.54318031	515886.1850	0.0338

表 3 实验组和对照组的 MSE

Period of time	Voltage Experimental group	Voltage Control group	Current Experimental group	Current Control group	Position Experimental group	Position Control group
0	0.0021	0.0030	1.5271E-05	1.6000E-05	9.9048E-08	1.0846E-07
1	0.0010	0.0015	1.3096E-05	1.6000E-05	8.4066E-08	7.9585E-08
2	0.0015	0.0018	1.3817E-05	1.6000E-05	6.0923E-08	5.8755E-08
3	0.0013	0.0011	1.3817E-05	1.6000E-05	8.7965E-08	8.1999E-08
4	0.0005	0.0004	1.3492E-05	1.6000E-05	9.9120E-08	8.4775E-08
5	0.0004	0.0004	1.1922E-05	1.6000E-05	7.4348E-08	1.0370E-07
6	0.0001	0.0001	0	0	7.4745E-08	8.7206E-08
7	0.0008	0.0012	0	0	6.7818E-08	8.8624E-08
8	0.0027	0.0033	0	0	1.1039E-07	1.3719E-07
9	0.5773	0.0427	0	0	1.7008E-07	1.5212E-07
10	8.1230	7.6014	0	0	1.7498E-07	2.7032E-07
11	7.0842	6.4278	0	0	1.8240E-07	2.0613E-07
12	8.3942	10.2773	0	0	1.4970E-07	1.5824E-07
13	4.1005	5.6803	0	0	9.0220E-08	9.9529E-08
14	5.7730	9.3327	0	0	6.8452E-08	7.8311E-08
15	6.3783	8.5873	0	0	7.3521E-08	8.2001E-08
16	2.4888	2.6739	0	0	8.1679E-08	1.2225E-07
17	0.1224	0.1423	0	0	6.9014E-08	9.4824E-08
18	0.0518	0.0723	1.8286E-05	2.5946E-05	7.4974E-08	8.1833E-08
19	0.0235	0.0317	1.8609E-05	2.2857E-05	7.4926E-08	9.3476E-08
20	0.0125	0.0179	2.1288E-05	3.2571E-05	9.0715E-08	1.1359E-07
21	0.0067	0.0093	2.0292E-05	1.3714E-05	1.5903E-07	1.9696E-07
22	0.0036	0.0056	1.7566E-05	1.6000E-05	2.1182E-07	2.7989E-07
23	0.0022	0.0037	1.4140E-05	1.6000E-05	2.2072E-07	2.3225E-07

MSE 越大。将概率密度峰值转化为权重，以加权平均数作为预测值，消减了极端值的影响，使预测值趋于“大概率”。概率密度峰值反映的是“常态”情况下的数值水平，未来偶然出现的“非常态”的实测值，将导致短期内的 MSE 变大，但是对长期的 MSE 影响不大。

4 结论

针对航标运行状态模式识别依赖经验阈值的现状，为检验经验阈值的普适性，提出基于概率的阈值模式识别效率度量算法，并用于检验经验阈值。经检验，经验阈值不具备普适性。因而，提出基于概率的模式识别模型，该模型能够有效识别频繁模式和异常模式，而且比传统方法的性能更好。为实现数值预测，提出基于概率密度的加权平均算法，该算法的数值预测精度较高。本文为航标运行状态模式识别和数值预测提供了新的解决方案。下一步，将研究航标漂移、灯器设备故障导致的“持续非常态”情况下的航标运行状态数值预测，拟从短期观测数据着手，分析数值变化趋势，比较和分析线性回归模型、非线性回归模型、时间序列模型的拟合效果和预测精度，寻找合适的模型。

参考文献：

[1] 任永功，高鹏，张志鹏 . 一种利用相关性度量的不确定数据频繁模式挖掘 [J]. 小型微型计算机系统，2019, 40(03):623-627.

[2] 吉小洪，徐爱萍 . 基于 TrieMerging 机制数据流滑动窗口模型的频繁模式挖掘 [J/OL]. 计算机应用研究 :1-7[2020-02-20]. https://doi.org/10.19734/j.issn.1001-3695.2019.01.0006.

[3] 胡江，赵冬梅，张旭，等 . 基于最大熵隐马尔科夫模型的电网故障诊断方法 [J]. 电网技术，2019, 43(09):3368-3375.

[4] 刘玉敏，刘莉 . 基于统计特征的动态过程质量异常模式识别 [J]. 统计与决策，2017(19):32-36.

[5] 宋路杰，孟凡荣，袁冠 . 基于 Markov 模型与轨迹相似度的移动对象位置预测算法 [J]. 计算机应用，2016, 36(01):39-43+65.

[6] 乔少杰，金琨，韩楠，等 . 一种基于高斯混合模型的轨迹预测算法 [J]. 软件学报，2015, 26(05):1048-1063.

[7] 肖延辉，王欣，冯文刚，等 . 基于长短记忆型卷积神经网络的犯罪地理位置预测方法 [J]. 数据分析与知识发现，2018, 2(10):15-20.

[8] 关绍云，郑丽坤，金一宁，等 . 基于高斯核函数的局部离群点检测算法 [J]. 哈尔滨商业大学学报 (自然科学版), 2019, 35(02):185-190+203.

[9] Andrew Harvey, Vitaliy Oryshchenko. Kernel density estimation for time series data[J]. International Journal of Forecasting, 2012, 28(01):3-14.

[10] Moses Charikar, Paris Siminelakis. Hashing-Based-Estimators for Kernel Density in High Dimensions[C]// 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS). IEEE, 2017.

[11] 马梦知，范厚明，黄莒森，等 . 基于非参数核密度估计的集装箱码头交通需求预测模型 [J]. 大连海事大学学报 (自然科学版), 2019, 45(01):77-84.

[12] 程媛，迟荣华，黄少滨，等 . 基于非参数密度估计的不确定轨迹预测方法 [J]. 自动化学报，2019, 45(04):153-164.