文章编号：2095-4980(2021)04-0623-05

基于混合机器学习的电磁功率谱密度预测模型

徐甜甜[1]，韩光洁[*1,2]，邹　岩[3]，朱宏博[4]，王　敏[1]，林　川[5]

(1.大连理工大学 软件学院，辽宁 大连 116620；2.河海大学 物联网工程学院，江苏 常州 213022；
3.沈阳飞机设计研究所，辽宁 沈阳 110035；4.沈阳理工大学 信息科学与工程学院，辽宁 沈阳 110159；
5.东北大学 软件学院，辽宁 沈阳 110819)

摘　要：功率谱密度(PSD)预测是频谱管理中的重要环节。由于功率谱密度具有高度的复杂性、非线性和不确定性，单一的预测模型很难确保预测的准确性和效率。为克服单一预测方法的不足，提出一种混合的机器学习模型，将自组织映射(SOM)网络与回归树(RT)相结合，以预测信号的功率谱密度。使用自组织映射网络将具有相似手工特征的原始样本集聚类成簇；将每一个簇分别构建回归树来预测功率谱密度；最后，使用亚琛工业大学的数据进行实验。结果表明，预测结果的均方根误差比现有方法提高0.824，证明混合模型具有较高的预测精确度和较好的泛化能力。

关键词：功率谱密度；自组织映射；回归树；手工特征

中图分类号：TP183　　**文献标志码：**　　**doi：**10.11805/TKYDA2021084

Electromagnetic Power Spectrum Density prediction model based on hybrid machine learning

XU Tiantian[1]，HAN Guangjie[*1,2]，ZOU Yan[3]，ZHU Hongbo[4]，WANG Min[1]，LIN Chuan[5]

(1.School of Software，Dalian University of Technology，Dalian Liaoning 116620，China；2.School of Internet of Things Engineering，Hohai University，Changzhou Jiangshu 213022，China；3.Shenyang Aircraft Design and Research Institute，Shenyang Liaoning 110035，China；4.School of Information Science and Engineering，Shenyang Ligong University，Shenyang Liaoning 110159，China；5.School of Software，Northeastern University，Shenyang Liaoning 110819，China)

Abstract：Power Spectral Density(PSD) prediction is an important part of spectrum management. Due to the high complexity, nonlinearity and uncertainty of the PSD, it is difficult for a single prediction model to ensure the accuracy and efficiency of the prediction. In order to overcome the disadvantages of a single prediction method, a hybrid machine learning model is proposed to combine a Self-Organizing Map(SOM) network with a Regression Tree(RT) to predict the PSD of the signal. First, the method uses a self-organizing map network to cluster the original sample sets with similar manual features. Then, a RT is constructed for each cluster to predict the PSD. Finally, the data of RWTH from Aachen University are adopted for experiments. The root mean square error of the prediction result is 0.824 higher than that of the existing method, which proves that the hybrid model has higher prediction accuracy and better generalization ability.

Keywords：Power Spectral Density；Self-Organizing Map；Regression Tree；manual feature

在人们的日常生活中，电磁频谱的利用已普遍存在。随着信息技术的发展，对频谱资源的需求迅速增长，频谱资源的日益稀缺一直是人们困扰的问题，但同时也面临着频谱利用率低的问题。频谱预测有助于实现更智能的频谱管理和更高效的频谱使用策略[1-3]。功率谱和频谱都是对信号的研究，功率谱密度(PSD)预测是频谱管理中的重要环节。

自频谱预测概念提出以来，研究者采用了很多算法对此进行研究，包括基于回归分析的预测、基于隐马尔可夫链(Hidden Markov Model，HMM)的预测和基于神经网络的预测。文献[1]基于经验模态分解(Empirical Mode

收稿日期：2021-04-15；**修回日期：**2021-05-10
基金项目：国家重点研发资助项目(2017YFE0125300)；江苏省重点研发资助项目(BE2019648)
*通信作者：韩光洁　email:hanguangjie@gmail.com

Decomposition，EMD)和支持向量机回归(Support Vector Regression，SVR)提出一种频谱预测方法 EMD-SVR。实验结果表明，EMD-SVR 模型在频谱预测中明显优于普通 SVR 模型。文献[4]采用了一种自回归积分滑动平均模型(Autoregressive Integrated Moving Average Model，ARIMA)分析频谱占用率，为认知无线电设备节省了大量扫描时间；文献[5]使用不同的机器学习技术，分析了认知无线电网络的频谱占用情况，包括决策树(Decision Tree，DT)、SVR、线性回归(Linear Regression，LR)和 HMM，以找到分类精确度最高的最佳技术。在计算时间和分类精确度方面对有监督算法和无监督算法进行了详细比较。

近些年来，深度学习的方法已成功用于很多领域，但其在通信领域的应用还处于起步阶段。文献[6]使用多层感知器(Multilayer Perceptron，MLP)神经网络模型设计了频谱预测器，它不需要事先了解授权用户系统的流量特性。通过大量的仿真实验，验证了频谱预测器的性能。文献[7]提出一种基于 LSTM 神经网络的频谱可用性预测系统，使用长短期记忆(Long Short-Term Memory，LSTM)网络发现历史频谱可用性数据之间的频谱时间相关性来进行预测。强化学习是另一种机器学习类型，文献[8]提出一种基于强化学习的微电网能源交易方案，据此预测未来可再生能源的发电量。但对于时间序列的预测，强化学习并不是理想的手段，其在工业界成功的应用很少。

现有的大多数研究都依赖于监督学习技术。在监督学习下，预测结果过分依赖于平均历史模式，无法保证预测的准确性和效率。聚类属于无监督学习算法，可深入了解数据中隐藏的各种关系[9]。自组织映射(SOM)作为一种无监督聚类的学习算法，已广泛用于工程领域[10]。该算法具有网络结构简单，自组织自学习能力强，学习速度快等优点。近年来，基于专业知识得到的手工特征也非常重要[11]。因此，本文提出一种基于手工特征的混合机器学习模型。具体来说，采用 SOM 将具有相似手工特征的 PSD 数据进行聚类，并将回归树(RT)结合使用，以预测信号的 PSD。

1 基本模型

信号的 PSD 预测是非线性时间序列预测的一种特殊情况，数据通常是非线性且有噪声的。分而治之的法则[12]通常用于将一个任务划分为几个较小且简单的任务，以解决复杂问题。基于此，本文采用 SOM 方案将具有相似手工特征的 PSD 数据聚类为几个不相交的簇。然后，在每个簇中构建单独的回归模型进行预测。

1.1 手工特征的选择

受限于低维度数据自动抽取高辨识力表征的难度，本文通过线性相关性选取了部分手工特征附加至自动抽取特征，输入至混合预测模型。选取峰值、均值、均方根值、峰值指标、波形指标、峭度、峭度指标、脉冲指标、裕度指标、歪度指标作为手工特征[11]，特征选择的主要任务是丢弃不提供足够信号特征的无关和冗余特征。本文利用相关性从特征集中选择最敏感的特征[13]。相关性度量标准度量特征和时间之间的线性相关性，如式(1)所示：

$$Corr=\frac{\left|\sum_{t=1}^{T}\left(F_t-\tilde{F}\right)\left(l_t-\tilde{l}\right)\right|}{\sqrt{\sum_{t=1}^{T}\left(F_t-\tilde{F}\right)^2\sum_{t=1}^{T}\left(l_t-\tilde{l}\right)^2}} \tag{1}$$

式中：F_t 和 l_t 为第 t 个样本的特征值和时间值；$\tilde{F}$ 为所有特征的均值；$\tilde{l}$ 为时间的均值；T 为固定时间内样本长度。如果 2 个特征之间的相关性大，则丢掉其中一个特征；2 个特征的相关性较小，则同时保留这 2 个特征。经过筛选后，将峰值指标、均方根值、波形指标作为信号 PSD 的手工特征输入至混合预测模型中进行训练。

1.2 混合预测模型

为准确预测信号的 PSD，提出一种将 SOM 与 RT 相结合的混合机器学习模型。SOM 网络是一种自组织网络，通过自动发现数据样本中的内在规律和基本属性来更改网络的参数或结构。该模型将数据点映射到神经元，并在标准训练过程后用于构建 RT。系统结构如图 1 所示。

如果输入层有 D 个输入向量，可以把输入模式写成 $\boldsymbol{x}=[x_i:i=1,2,\cdots,D]$。输入单元 i 和神经元 j 之间在计算层的连接权重可以写成 $\boldsymbol{m}_j=[m_{ji}:j=1,2,\cdots,L;i=1,2,\cdots,D]$，$L$ 是神经元的总数。所有连接权重都用小的随机值进行初始化。

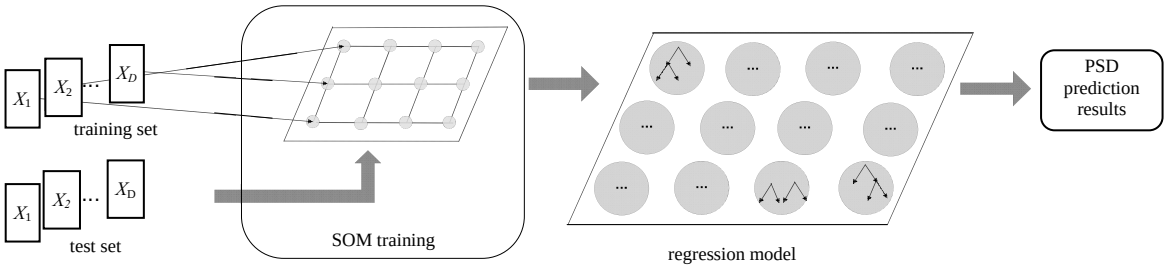

Fig.1 System structure
图 1 系统结构

对于每种输入向量，神经元计算它们各自的判别函数值(欧式 regression model 最小判别函数值的特定神经元为获胜神经元。在数学上，选择具有最大内积 $\boldsymbol{m}_j^{\mathrm{T}}\boldsymbol{x}$ 的神经元等效于最小化输入向量 $\boldsymbol{x}$ 和 $\boldsymbol{m}_j$ 之间的欧式距离。因此，获胜的神经元 n 定义为：$n=\arg\min_{1\leqslant j\leqslant l}\left\{\left\|\boldsymbol{x}-\boldsymbol{m}_j\right\|\right\}$。获胜的神经元决定了兴奋神经元拓扑邻域的空间位置，从而为相邻神经元之间的合作提供了基础。在学习过程中，在地理位置上接近的节点会相互激活，从相同的输入中学习。采用离散时间形式，将 t 时刻的权重向量写成 $\boldsymbol{m}_j(t)$，更新后的权重向量定义为：

$$\boldsymbol{m}_j(t+1)=\boldsymbol{m}_j(t)+h_{jn}(t)\left[\boldsymbol{x}-\boldsymbol{m}_j(t)\right] \tag{2}$$

式中：t 为离散时间坐标；$h_{jn}(t)$ 为邻域函数，通常高斯函数形式为：

$$h_{jn}(t)=\alpha(t)\cdot\exp\left(-\frac{\left\|\boldsymbol{r}_n-\boldsymbol{r}_j\right\|^2}{2\sigma^2(t)}\right) \tag{3}$$

式中：训练速率函数 $\alpha(t)=\alpha_0\exp(-t/\tau_\alpha)$；$\boldsymbol{r}$ 为矩阵中的位置矢量，$\|\boldsymbol{r}_n-\boldsymbol{r}_j\|$为神经元网格上神经元 i 和 j 之间的横向距离。在 SOM 网络中，还有一个特征是拓扑邻域的大小随时间收缩，拓扑邻域的有效宽度 $\sigma(t)$定义为：$\sigma(t)=\sigma_0\exp(-t/\tau_\sigma)$。对该过程的迭代进行会使网络的拓扑有序。

在回归层中，对于每一个聚类后的子集，回归树根据以下步骤生长[14]：在每个节点上，它从手工特征向量中随机选择一些特性作为分割变量。然后选择最佳分割变量 k 与切分点 s：

$$\min_{k,s}\left[\min_{c1}\sum_{x_i\in R_1(k,s)}(y_i-c_1)^2+\min_{c2}\sum_{x_i\in R_2(k,s)}(y_i-c_2)^2\right] \tag{4}$$

遍历变量 k，对固定的切分变量 k 扫描切分点 s，选择使式(4)值最小的对(k,s)。在下一个节点处，重复该过程并将训练集划分为更小的组。直至叶子节点，得到对应的预测值。继续对两个区域调用上述步骤，直至满足停止条件，生成回归树。

2　实验

2.1 数据介绍

使用亚琛工业大学频谱测量活动的频谱数据。整个数据集包含 3 个不同位置测得的 PSD，如表 1 所示。GSM1800 下行链路频带中的时域相关性很强[15]，因此，本文使用 IN 数据集从 1 820.0~1 875.5 MHz 范围内的频谱数据，其中 10 000 条数据用于训练，3 000 条数据用于测试。实验数据来源参考文献[16]。

表 1 数据采集设备放置位置
Table1 Measurement locations

name	short name	short description
Aachen, indoor	IN	modern office building
Netherlands	NE	rooftop location in a mostly residential area in Maastricht, the Netherlands
Aachen, balcony	AB	third floor balcony of a residential building in a rather central housing area of Aachen

2.2 基于相关性系数的手工特征选择

为更加直观地说明各个手工特征对最终结果的影响，本文计算 10 个手工特征的相关性系数，从中选择相关性较小的特征输入到混合模型中。图 2 为 10 个手工特征的相关性热力图。由图 2 相关性系数矩阵看出，颜色越

浅，代表相关性越小。经过筛选后的手工特征为：峰值指标、均方根值、波形指标。

2.3 评估指标

选择 2 个指标来评估模型：平均绝对误差(Mean Absolute Error，MAE)和均方根误差(Root Mean Square Error，RMSE)。计算方法如下：

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|p_i-y_i| \tag{5}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i-y_i)^2} \tag{6}$$

式中：n 为测试数据集中观察的总数；p_i 为预测值；y_i 为实际值。MAE 能更好地反映预测值误差的实际情况。RMSE 评价数据的变化程度，对于异常值比较敏感。

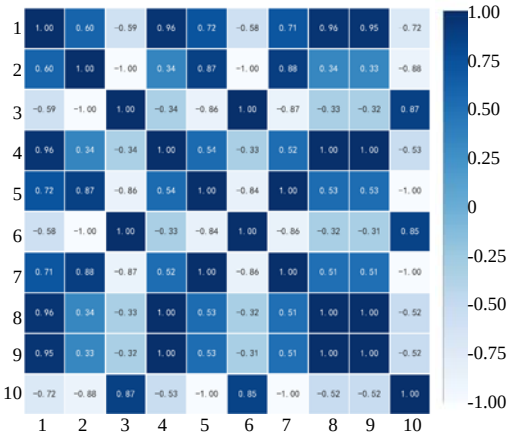

Fig.2 Heat map of correlation coefficient
图 2 相关性系数热力图

2.4 模型训练

输入手工特征后的第一阶段是进行数据预处理。使用 SOM 对具有相似手工特征的数据进行聚类。改变聚类神经元的个数，得到不同的误差，如表 2 所示。

表 2 不同神经元个数的训练误差
Table2 Training error for different numbers of neurons

	2×2	2×3	3×3	3×4	4×4	4×5	5×5
MAE	0.842	0.827	0.619	0.618	0.614	0.624	0.620
RMSE	1.066	1.050	0.811	0.810	0.808	0.816	0.813

由表 2 可知，当神经元个数为 4×4 时，实验获得最佳结果，因此本文 SOM 模型神经元个数为 4×4。对模型参数进行调整，选择最优参数。其中每次迭代期间权重的调整幅度为 0.01，迭代次数为 4 000，SOM 中不同相邻节点的半径为 3。考虑到预测结果可能会受诸多因素影响，每个结果取 10 次重复实验的平均值。在 SOM 训练阶段，将 3 个手工特征输入至网络中进行聚类；在回归阶段，使用每个子类的 PSD 数据和手工特征一起构建回归树。

2.5 评估

将所提出的混合模型与 RT[5]、梯度增强回归树(Gradient Boosted Regression Tree，GBRT)、SVR[1]、LSTM[7]和注意力机制结合 LSTM[17]进行比较，结果如表 3 所示。由表 3 可知，本文提出的模型比 RT 模型、GBRT 模型、SVR 模型、LSTM 模型和 Attention-LSTM 模型具有更好的性能，RMSE 为 0.808。结果表明，聚类将数据集划分为多个子类，增强数据的规律性，提高预测的准确性。

表 3 对比误差
Table3 Contrast error

	RT	GBRT	SVR	LSTM	Attention-LSTM	proposed
MAE	1.952	1.392	1.408	1.707	1.621	0.614
RMSE	2.394	1.632	1.659	2.693	2.576	0.808

3 结论

为实现对信号的 PSD 预测，本文提出了一种混合机器学习模型，将 SOM 与 RT 结合以预测信号的 PSD。首先，将训练数据的手工特征输入到 SOM 中进行聚类。然后，将每一个簇分别构建回归树来预测 PSD。最后，使用亚琛工业大学的数据进行实验。实验结果表明，本文提出的混合模型不仅可以提供更好的信号 PSD 预测，还可以揭示 PSD 数据的固有特性。

参考文献：

[1] DING Guoru,WANG Jinlong,WU Qihui,et al. On the limits of predictability in real-world radio spectrum state dynamics: from entropy theory to 5G spectrum sharing[J]. IEEE Communications Magazine, 2015,53(7):178-183.

[2] 陈国通,刘琪,孙敬. 复杂电磁环境下的高速频谱分析技术[J]. 太赫兹科学与电子信息学报, 2020,18(6):68-72. (CHEN Guotong,LIU Qi,SUN Jing. High-speed spectrum analysis technology in complex electromagnetic environment[J]. Journal of Terahertz Science and Electronic Information Technology, 2020,18(6):68-72.)
[3] JAMALI V,AHMADZADEH A,SCHOBER R. On the design of matched filters for molecule counting receivers[J]. IEEE Communication Letters, 2017,21(8):1711-1714.
[4] WEN Z,TAO L,XIANG W,et al. Autoregressive spectrum hole prediction model for cognitive radiosystems[C]// IEEE International Conference on Communications Workshops. Beijing,China:IEEE, 2008:154-157.
[5] AZMAT F,CHEN Y,STOCKS N. Analysis of spectrum occupancy using machine learning algorithms[J]. IEEE Transactions on Vehicular Technology, 2016,65(9):6853-6860.
[6] URKOWITZ H. Energy detection of unknown deterministic signals[J]. Proceedings of the IEEE, 2005,55(4):523-531.
[7] LI H,DING X,YANG Y,et al. Spectrum occupancy prediction for internet of things via Long Short-Term Memory[C]// 2019 IEEE International Conference on Consumer Electronics. Taiwan,China:IEEE, 2019.
[8] LU X,XIAO X,XIAO L,et al. Reinforcement learning-based microgrid energy trading with a reduced power plant schedule[J]. IEEE Internet of Things Journal, 2019,6(6):10728-10737.
[9] WANG G,JIA R,LIU J,et al. A hybrid wind power forecasting approach based on Bayesian model averaging and ensemble learning[J]. Renewable Energy, 2020(145):2426-2434.
[10] LIU H,BAN X J. Clustering by growing incremental self-organizing neural network[J]. Expert Systems with Applications, 2015,42(11):4965-4981.
[11] CHEN Z,WU M,ZHAO R,et al. Machine remaining useful life prediction via an attention based deep learning approach[J]. IEEE Transactions on Industrial Electronics, 2021,68(3):2521-2531.
[12] LEE J,KIM J,KO W. Day-ahead electric load forecasting for the residential building with a small-size dataset based on a Self-Organizing Map and a stacking ensemble learning method[J]. Applied Sciences, 2019,9(6):1231.
[13] GUO L,LI N,JIA F,et al. A recurrent neural network based health indicator for remaining useful life prediction of bearings[J]. Neurocomputing, 2017(240):98-109.
[14] XU T,HAN G,QI X,et al. A hybrid machine learning model for demand prediction of edge-computing based bike sharing system using internet of things[J]. IEEE Internet of Things Journal, 2020,7(8):7345-7356.
[15] DING G,WU F,WU Q,et al. Robust online spectrum prediction with incomplete and corrupted historical observations[J]. IEEE Transactions on Vehicular Technology, 2017,66(9):8022-8036.
[16] WELLENS M,MAEHOENEN P. Lessons learned from an extensive spectrum occupancy measurement campaign and a stochastic duty cycle model[J]. Mobile Networks and Applications, 2010,15(3):461-474.
[17] LI K,LIU Z,HE S,et al. TF 2 an:a temporal-frequency fusion attention network for spectrum energy level prediction[C]// 2019 16th Annual IEEE International Conference on Sensing,Communication,and Networking. Boston,Massachusetts, USA:IEEE, 2019.

作者简介：

徐甜甜(1995-)，女，在读博士研究生，主要研究方向为机器学习.email:tiantianxu0609@gmail.com.

邹　岩(1983-)，男，工程师，工学博士，主要研究方向为机载射频综合设计、电子战系统设计、多传感器信息融合技术等.

王　敏(1999-)，男，在读硕士研究生，主要研究方向为深度学习.

韩光洁(1972-)，男，教授，博士生导师，主要研究方向为传感器网络、云计算、绿色计算、智能感知技术等.

朱宏博(1986-)，男，副教授，博士，主要研究方向智能计算.

林　川(1988-)，男，副教授，博士，主要研究方向为人工智能、工业物联网.