

doi:10.3969/j.issn.1001-893x.2016.08.003

引用格式:贾云峰,邱琳,魏鸿浩.基于 k 最近邻回归的频谱占用度预测[J].电讯技术,2016,56(8):844-849. [JIA Yunfeng, QIU Lin, WEI Honghao. Spectrum occupancy prediction based on k-nearest neighbor regression[J]. Telecommunication Engineering, 2016, 56(8):844-849.]

基于 k 最近邻回归的频谱占用度预测*

贾云峰,邱琳**,魏鸿浩

(北京航空航天大学 电子信息工程学院,北京 100191)

摘要:认知无线电技术可以在授权用户和非授权用户间进行频谱分配,预测模型的建立可帮助非授权用户推断频谱空洞是否可用,不仅能提升频谱利用率而且还能降低冲突率。采用理论分析、监测实验、数学建模、数据实证等方法,对频谱占用度建模理论进行了研究。针对频谱的可预测性问题,通过对数据集的分析,使用 k 最近邻(kNN)回归模型预测频谱的信道-场强值。基于观测数据呈现出的周期性,提出了一种针对周期性数据进行优化的 kNN 模型,并用其进行预测。比较了原始 kNN 回归模型和优化后的周期性 kNN 模型在测试数据上的预测精度,结果表明优化后的模型比原始的 kNN 模型有着更好的预测精度。

关键词:认知无线电;频谱分配;频谱占用度;场强预测;k 最近邻回归

中图分类号:TN98 文献标志码:A 文章编号:1001-893X(2016)08-0844-06

Spectrum Occupancy Prediction Based on
k-Nearest Neighbor Regression

JIA Yunfeng, QIU Lin, WEI Honghao

(School of Electronic and Information Engineering, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract: Cognitive radio technology can conduct spectrum allocation between the authorized users and secondary users. The establishment of predication model can help secondary users infer whether the spectrum hole is available, which can both improve spectral efficiency and reduce collision rate. By means of theoretical analysis, experiment monitoring, mathematical modeling and data demonstration, spectrum occupation modeling theory is researched. For the predictable problems of spectrum, through the analysis of data group, k-Nearest Neighbour (kNN) regression model is used to predict the channel-field value of spectrum. At the same time, based on the periodicity shown by the observation data, a kNN model is proposed to optimize periodical data and offers predication. Then the predication accuracy is compared in test data of original kNN regression model and optimized periodical kNN. The result shows the optimized model is of better predication accuracy than the original kNN model.

Key words: cognitive radio; spectrum allocation; spectrum occupancy; field strength prediction; kNN regression

1 引言

频谱占用度是描述电磁频谱使用规律和利用程

度的重要物理量,是无线电监测中用来衡量无线电业务频段拥挤程度和信道可利用性的主要统计量。

* 收稿日期:2015-12-30;修回日期:2016-03-28 Received date:2015-12-30;Revised date:2016-03-28
基金项目:国家自然科学基金资助项目(61371007)
Foundation Item: The National Natural Science Foundation of China(No. 61371007)

** 通信作者:qiulin@buaa.edu.cn Corresponding author:qiulin@buaa.edu.cn

频谱占用度建模研究,是在一定时段内覆盖特定频段的电磁环境监测数据支持下,研究不同无线电业务频段和信道占用度的统计规律,建立能够准确表征频谱占用度数学特性的理论模型。频谱占用度能够综合反映电磁频谱的频域-时域变化规律,定量描述频谱状态变化趋势,深入揭示频谱使用行为信息,在战场电磁环境评估、频谱管理工程和认知无线电技术等领域具有重要应用价值。

传统上电磁环境监测中频谱占用度分析主要是基于监测设备和人工经验相结合的方法,即通过使用监测接收机对特定的频段进行扫描,或者用频谱分析仪对之前设定的信道进行监测,获取特定无线电频段(典型为 30 MHz ~ 3 GHz)范围内频谱场强与时间的对应数据集,通过人工选取电磁背景噪声电平值,记录任何高于噪声电平或预设门限电平的信号^[1]。这种以测量、搜集、记录和显示为主的频谱占用度测量方法缺乏定量数学模型的支撑,无法深入描述频谱历史使用数据的统计规律,不能精确预测频谱未来状态和波动信息。因此,传统频谱占用度测量和分析模式越来越无法适应当前频谱资源精确和动态共享的发展需求。

近年来,国内外学者对频谱占用度进行了研究。文献[2-4]提出用马尔科夫链建立频谱占用度时域统计模型,并通过多节点同时检测实验表明主用户发射信号场强在空域服从高斯分布。文献[5-7]通过对大量电磁环境的测试,采用指数分布和几何分布函数来拟合频谱占用度时变特性,并基于随机场理论建立频谱利用空域分布模型。上述研究进展对于认识无线电监测频谱的时变统计规律以及量化表征方法具有重要意义。文献[8]提出了频谱占用度的测量与分析方法,获得了监测点的现场实测数据,并经过分析得到了相应的时间-频率-占用度对照图,其研究相对比较浅显,未能对频谱的统计特性进行进一步分析。文献[9]通过递归图和递归定量分析,定性和定量研究了基于某 GSM 基站流量的时间序列可预测性特征,但缺乏对模型的具体构建。本文对频谱进行分析与预测,为之后的异常检测提供依据。

由于传统电磁环境监测中频谱占用度分析无法精确预测频谱未来状态,同时,从监测到的频谱数据来看,数据分布并不满足正态性,甚至不同频段的信号分布形状之间差异非常大。因此,本文首先根据

其数据特征选择 k 最近邻(k-Nearest Neighbour, kNN)回归算法对频谱场强值进行预测;然后,根据频谱数据呈现出的周期性对模型进行改进;最后,根据国际电联的 ITU-R 建议 S M. 1536——频道占用度测量,得到频谱占用度。

2 频谱测量数据的统计评估

2.1 认知无线电中的频谱预测技术

频谱预测技术即对频谱的历史数据进行预测。在实际频谱中,不同的无线频段分配给不同的服务,导致其频谱使用状态呈现一定的相关特性,即在一个授权频段内,频谱使用呈现有规律的特征。因此,可以利用这些规律、特性对频谱未来的使用情况进行预测。通过频谱预测可以减小对主用户的干扰的同时寻找到更多的频谱接入机会,保证主用户及认知用户 QoS 的同时提高了认知用户的吞吐量。

对于频谱预测,需要做的工作主要分为三步:一是采集频谱资料;二是对频谱数据进行分析;最后是根据相应的预测方法对频谱进行预测。

2.2 频谱数据的采集

本研究中的监测数据来自对北京市海淀区北京航空航天大学校园内进行连续约 64 h(2015 年 3 月 13 日 9 时 52 分~2015 年 3 月 16 日 1 时 55 分)频段为 88 ~ 108 MHz,即 FM 广播业务的实地频谱监测。监测设备包括 Agilent N9340B 频谱分析仪、数据记录和存储计算机以及 CS-AOS30-3000V 有源全向天线。该实验系统能满足本研究工作的需要。

数据采集软件采用北京航空航天大学电磁兼容研究所开发的电磁环境监测系统 V1.0,监测频段设置为 30 ~ 3 000 MHz,频率扫描分辨率为 50 kHz,每次扫描采样的频率点数为 144 360 个,在监测时间内连续测量得到 64 个时间点的“频谱-场强”数据样本。由于在该监测系统中每一次扫描频谱数据被记录为一个文本文件,因此原始数据存储格式为 64 个文本文件。除了第一个小时的数据包含 361 个数据集外,其余 63 个小时中每一个小时都包含了 360 个数据集。将该数据集进行三维(频率-时间-场强)可视化处理得到如图 1 所示的监测数据分布图。在频谱占用度统计和分析中,可以根据实际需求抽取不同的无线电业务频段或信道频率,从而得到相应的监测频谱场强值。

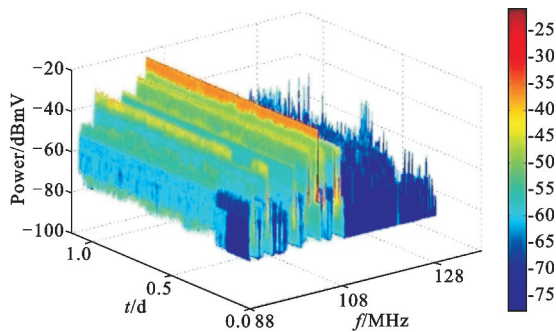


图 1 频谱监测数据频率-时间-场强分布图
Fig. 1 Frequency-time-field intensity distribution of spectrum monitoring data

2.3 数据的预处理

本文使用 R(版本号:3.2.2)做数据分析和建模。原始数据为 63 个文本文件,每个文本记录了 400 个信道(88~108 MHz)在 1 h 中每 10 s 测量一次的观测值。一次典型的观测值如图 2 所示,此数据来自于第 2 个小时中的第 5 次测量。把 63 个记录封装成矩阵,最终得到维数为 22 680×401 的矩阵。

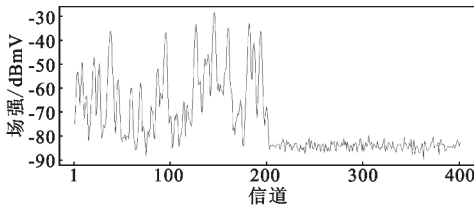


图 2 第 2 个小时中的第 5 次测量得到的 400 个信道的观测值
Fig. 2 Observed value of 400 channels during the fifth time measurement in the 2nd hour

2.4 模型优劣的评价指标

本文选取拟合优度和均方误差作为信道-场强预测的评价指标。

拟合优度 $R_{squared}$ 被定义为

$$R_{squared} = 1 - SSE / SST, \\ SST = SSR + SSE。$$

式中: SST 、 SSE 、 SSR 分别是离差平方和、残差平方和和回归平方和。拟合优度越大,说明数据相似度越好。

均方误差 (Mean Squared Error, MSE) 是衡量“平均误差”的一种较为简便的方法,它可用来评价数据的变化程度。数学定义如下:

$$MSE = \frac{1}{n} \sum_{i=1}^n (observe\ d_1 - predict\ d_1)^2。$$

在本研究中,我们分别用 $R_{squared}$ 和 MSE 来评

价两条曲线的相似度。

3 频谱预测算法

数据挖掘技术是一个内容广泛的理论体系和算法集。数据挖掘所发现的知识主要包括广义知识、关联知识、分类知识、预测性知识和偏差型知识等。采用数据挖掘等理论和方法分析处理频谱监测数据,求解设定参数和目标函数,获取有价值的频谱资源使用信息和知识是当前对频谱监测和频谱占用度数据研究的重点方向。

3.1 kNN 最近邻分类算法

kNN 方法是指从训练集找出 k 个最接近测试对象的训练对象,再从这 k 个训练对象中找出居于主导的类别,将其赋值给测试对象。从普遍意义上来说,kNN 是一种基于实例的学习方法^[10],同时,kNN 也是一种惰性学习方法。

kNN 分类方法很容易理解和实现,而且在许多情况下表现良好。虽然数据在每个小周期内很有规律,但是在一个小时内,变化很复杂,很难写出简单的函数表达式,所以普通的回归类型的方法都无法使用。改进后的 kNN 回归是基于局部数据点的,而这个“局部”在时间上是不连续的,所以选择 kNN 方法能达到较好的预测效果。

kNN 方法需要考虑几个关键要素:一是用于决策一个测试对象类别的已被标记对象集合;二是用来计算对象间邻近程度的距离或其他相似性指标;三是最近邻的个数 k ;四是基于 k 个最近邻及其类别来判定目标对象类别的方法。

算法 1 是最近邻算法的高层描述。对于测试样例 $z = (x', y')$,算法计算它和所有训练样例 $(x', y') \in D$ 之间的距离,以确定其最近邻列表 D_z 。

算法 1 k -最近邻分类算法

- (1) 令 k 是最近邻数目, D 是训练样例的集合;
- (2) for 循环每一个测试样例 $z = (x', y')$ do;
- (3) 计算 z 和每个样例 $(x', y') \in D$ 之间的距离 $d(x', x)$;
- (4) 选择离 z 最近的 k 个训练样例的集合 $D_z \subseteq D$;
- (5) $y' = \operatorname{argmax}_v \sum_{(x_i, y_i) \in D_z} I(v = y_i)$;
- (6) end for。

一旦得到最近邻列表,测试样例就会根据最近邻中的多数类进行分类:

$$y' = \operatorname{argmax}_v \sum_{(x_i, y_i) \in D_z} I(v = y_i)。$$

式中: v 是类标号; y_i 是一个最近邻的类变换; $I(\cdot)$ 是指示函数,当参数为真时,返回 1;为假,返回 0。

以上 kNN 算法针对的是分类问题,目标值是离散变量,而我们需要的是测量值,是一个连续的变量,所以要用到 kNN 回归。

kNN 回归的思想是找到离待测样本最近的 k 个已知样本,将它们的均值作为待测样本的值。例如:设 $a_1, a_2, \dots, a_k$ 分别是 x 的 k 个最近邻样本,它们的值为 $v_1, v_2, \dots, v_k$,则目标 x 的值是

$$V_x = \frac{1}{k} \sum V_i。$$

图 3 和图 4 表明每小时的均值和每分钟的均值有明显的周期性,所以为了预测未来的某一个时间 t 的值,可以根据历史数据中跟 t 处在相同周期位置的那些值来预测,因此我们将不同周期中相同位置附近的值也作为“邻居”。由此,我们提出改进的 kNN 模型。

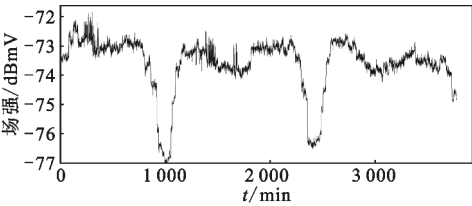


图 3 每分钟的均值
Fig.3 Mean value per minute

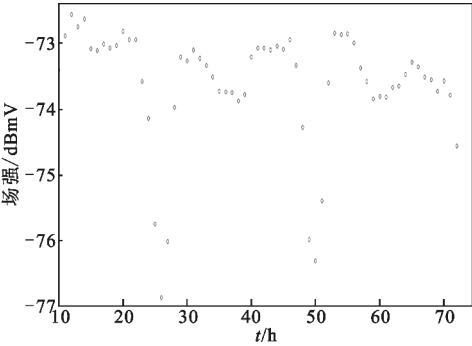


图 4 每小时的均值
Fig.4 Mean value per hour

3.2 优化的 kNN 回归模型

优化的 kNN 回归模型主要针对周期性数据而修改。

设:待测样本 X , 所在的时间是 T , 由于周期性的存在,总的来说,最接近 T 的值是跟它相差 24 h 的整数倍的那些时间点 X_i 的值 V_i 。找到这些值之后,使用 kNN 算法:选择时间上最近的 k 的观测值求均

值来得出每一个待测点的值。算法 2 是对优化的 kNN 的描述。

算法 2 优化的 k -最近邻分类算法

- (1)通过待测点 X 的时间 T ,找到 $x_1, x_2, \dots, x_m$, 这些点是与 X 在时间上距离 24 h 的整数倍的点;
- (2)计算 $x_1, x_2, \dots, x_m$ 这些点对应的观测值 $xv_1, xv_2, \dots, xv_m$;
- (3)对于每一个 x_i ,计算出 x_i 的 k 个最近邻的均值 v_i ,得到 m 个均值 $v_1, v_2, \dots, v_m$;
- (4)对于我们的目标值 X ,它的值为 $V_x = \frac{1}{m} \sum V_i$ 。

4 预测结果

图 5 是以 $Rsquared$ 作为评价实际数据值和预测数据值的相似度的标准,使用优化的 kNN 回归模型的信道-场强预测结果,从直方图上看,预测的正确率大约在 88% 以上。

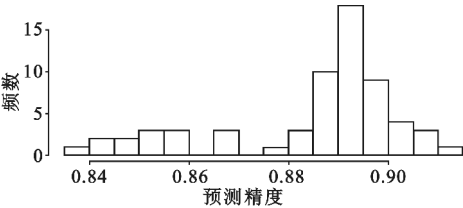


图 5 优化的 kNN 算法对频谱预测准确率的直方图
Fig.5 Spectrum prediction accuracy histogram based on the optimized kNN algorithm

由于用 $Rsquared$ 考察频谱预测准确率会出现不同小时之间两两的相似度很接近于 0 的情况,所以我们用均方误差直接比较预测值和真实值在绝对值上的差异。

将 3 月 13 日 10:52:19 ~ 3 月 15 日 02:54:10 之间采集的数据集作为基础,通过它们预测剩余时间上的数据值,测试了原始 kNN (ori_kNN) 和改进 kNN (m_kNN) 的结果。讨论 $k=2$ 和 $k=3$ 时的预测 MSE 值,如表 1 所列:第 2、3 列是 $k=2$ 的情况下 MSE 值,第 4、5 列是 $k=3$ 的情况下 MSE 值;第 2、4 列是 ori_kNN 预测误差列,第 3、5 列是 m_kNN 预测误差列。从数值上看,取 $k=3$ 较好。在 $k=3$ 时,比较 ori_kNN 和 m_kNN 的误差大小可以发现,m_kNN 误差更小的情况出现了 15 次,ori_kNN 误差更小的情况出现了 8 次。所以,改进的 kNN 回归模型在预测上有更高的精度。

表 1 原始 kNN 回归与改进的 kNN 回归在 23 个测试样本上的误差

Tab. 1 The original kNN regression and the improved kNN regression on the error of the 23 test samples

样本	模型均方差			
	原始 $k=2$	改进 $k=2$	原始 $k=3$	改进 $k=3$
1	23.18	22.78	31.72	31.58
2	25.76	25.55	23.82	23.80
⋮	⋮	⋮	⋮	⋮
23	26.98	26.32	28.47	27.81
误差更小次数 (百分比)	11 (47.83%)	12 (52.17%)	8 (34.78%)	15 (65.22%)

5 频谱占用度的获得

电磁环境监测和频谱感知是频谱管理部门和用户获取空中无线电波信息的基本手段,衡量一定无线电业务或特定无线信道利用程度的物理量称为频谱占用度,它是频谱管理部门掌握频谱资源利用情况、预测频谱资源变化趋势以及高效配置频谱资源的基本依据,也是认知无线电用户确定动态频谱接入策略的必备条件。

频道占用度是频谱占用度的一种^[8],它能度量和描述频谱资源的使用情况。频道占用度被定义为占用时间与总测量时间的百分比。判定占用与否的标准是接收信号是否大于门限电平;当接收信号大于门限电平时判为占用,否则为不占用。

图 6 为对测量的数据经过合并周期,占用度、门限计算得到的频率-占用度二维图。图内显示的是一天内各频点的总体占用度值。横轴代表的是被测频段的频率 88~108 MHz,包含 400 个信道;纵轴代表的是时间占用度,它是信道门限值之上的数目除以该信道总的数目,用百分比表示。图中选择的判决门限值是-80 dBmV/m。

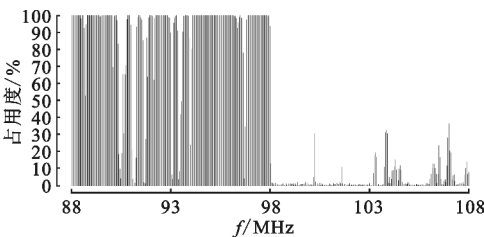


图 6 频道占用度

Fig. 6 Channel occupancy rate

从图 6 中能看到一些未被使用的空白频谱,而且被占用频段集中在 88~98 MHz;在 98~108 MHz 范围内,频道占用度较低,即频谱的实际利用不高。

由此看来,我国分配的 FM 广播业务频段仍有较大的使用空间。

6 结束语

频谱占用度信息是认知无线电系统的重要构成要素,准确的频谱占用度信息能为认知无线电用户提供频谱接入的最佳时机和方式,为其科学动态选频和系统间电磁兼容提供关键数据支持。电磁频谱高效利用的需求日趋紧迫,可靠的频谱预测有利于提高频谱利用率。

本文构建了基于 kNN 回归的频谱占用度预测模型。因数据集展现出明显的周期性,因此,我们在传统的 kNN 算法上进行适当的改进。当以 *Rsquared* 作为评价实际数据值和预测数据值的相似度的标准时,得出改进的 kNN 回归预测准确率在 88% 以上。在误差度量方法为均方误差 (MSE) 的条件下,比较不同 k 取值的原始 kNN 回归 (ori_kNN) 和改进后的 kNN 回归 (m_kNN) 的预测结果,得出 m_kNN 比 ori_kNN 在预测效果上更好的结论。最后通过分析,获得频道占用度。通过对不同时间的信道占用度的区别进行统计监测来得出无线电监测中的频谱异常,是下一步的研究方向。

参考文献:

[1] 周鸿顺. 频谱监测手册[M]. 北京: 人民邮电出版社, 2006: 9-10.
ZHOU Hongshun. Spectrum monitoring handbook [M]. Beijing: The People's Posts and Telecommunications Press, 2006: 9-10. (in Chinese)

[2] LOPEZ-BENITEZ M, CASADEVALL F. Empirical time-dimension model of spectrum use based on a discrete-time Markov chain with deterministic and stochastic duty cycle models[J]. IEEE Transactions on Vehicular Technology, 2011, 60(6): 2519-2533.

[3] LOPEZ-BENITEZ M, CASADEVALL F. Discrete-time spectrum occupancy model based on Markov chain and duty cycle models[C]//Proceedings of 2011 IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN). Aachen, Germany: IEEE, 2011: 90-99.

[4] LOPEZ-BENITEZ M, CASADEVALL F. Spatial duty cycle model for cognitive [C]//Proceedings of 2010 IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC). Istanbul, Turkey: IEEE, 2010: 1631-1636.

[5] WELLENS M, RIIHIJARVI J, MAEHOENEN P. Spatial statistics of spectrum usage: from measurements to spectrum models [C]//Proceedings of 2009 IEEE Interna-

tional Conference on Communications. Dresden, Germany: IEEE, 2009: 1–6.

[6] WELLENS M, RIIHIJARVI J, MAEHOENEN P. Evaluation of cooperative spectrum sensing based on large scale measurements[C]//Proceedings of 3rd IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks. Chicago, IL: IEEE, 2008: 1–12.

[7] WELLENS M, RIIHIJARVI J, MAEHOENEN P. Empirical time and frequency domain models of spectrum use[J]. Physical Communication, 2009, 10(4): 10–32.

[8] 王江舟, 李军芳, 范若璐, 等. 频谱占用度测量与分析[J]. 西安邮电大学学报, 2014, 19(5): 10–15.
WANG Jiangzhou, LI Junfang, FAN Ruolu, et al. Measurements and analysis of spectrum occupancy[J]. Journal of Xi'an University of Posts and Telecommunications, 2014, 19(5): 10–15. (in Chinese)

[9] 李红岩. 认知无线电系统中频谱可预测性的递归定量分析[J]. 电讯技术, 2015, 55(2): 124–128.
LI Hongyan. Recurrence quantification analysis of spectrum predictability in cognitive radio system[J]. Telecommunication Engineering, 2015, 55(2): 124–128. (in Chinese)

[10] AHA D W, KIBLER D, ALBERT M K. Instance-based learning algorithms[J]. Machine Learning, 1991, 6(1): 37–66.

作者简介:



贾云峰(1975—),男,湖北人,2002 年于国防科技大学获博士学位,现为副教授、硕士生导师,主要研究方向为微波电磁场、电磁兼容技术与直升机系统级电磁兼容设计;

JIA Yunfeng was born in Hubei Province, in 1975. He received the Ph. D. degree from National University of Defense Technology in 2002.

He is now an associate professor and also the instructor of graduate students. His research concerns microwave electromagnetic field, electromagnetic compatibility technology and the helicopter system-level electromagnetic compatibility design.

邱琳(1993—),女,江西人,硕士研究生,主要研究方向为电磁场与电磁兼容;

QIU Lin was born in Jiangxi Province, in 1993. She is now a graduate student. Her research concerns electromagnetic field and electromagnetic compatibility.

Email: qiulin@ buaa. edu. cn

魏鸿浩(1991—),男,辽宁人,硕士研究生,主要研究方向为电磁场与电磁兼容。

WEI Honghao was born in Liaoning Province, in 1991. He is now a graduate student. His research concerns electromagnetic field and electromagnetic compatibility.