信息科学技术专著丛书

基于机器学习的光通信系统物理损伤感知与补偿

陈远祥　编著

JIYU JIQI XUEXI DE GUANGTONGXIN XITONG
WULI SUNSHANG GANZHI YU BUCHANG

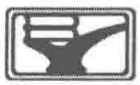
非外借

北京邮电大学出版社
www.buptpress.com

信息科学技术专著丛书

基于机器学习的光通信系统物理损伤感知与补偿

陈远祥　编著



北京邮电大学出版社
www.buptpress.com

内 容 简 介

随着网络系统规模的扩大、网络灵活性的提升，传统的光通信补偿机制由于复杂度较高、补偿效果有限，面临着极大的挑战和升级需求。因此，发展高效智能、低复杂度且具有自适应能力的损伤感知与补偿机制，对构建未来大规模的光通信系统有重要的现实意义。在本书中我们将探索基于机器学习的统计信号处理方法，用于解决光通信系统中灵活多变的物理损伤问题。本书全面系统地介绍了光通信系统中损伤产生的原因及如何应用机器学习进行光通信系统中的参数感知和损伤补偿。此外，本书还介绍了机器学习算法在混沌加密中的具体应用。

图书在版编目(CIP)数据

基于机器学习的光通信系统物理损伤感知与补偿 /陈远祥编著. -- 北京 ：北京邮电大学出版社，2021.10

ISBN 978-7-5635-6491-0

Ⅰ. ①基… Ⅱ. ①陈… Ⅲ. ①光通信系统－研究②机器学习—算法—研究 Ⅳ. ①TN929.1②TP181

中国版本图书馆 CIP 数据核字(2021)第 174393 号

策划编辑：彭 楠 **责任编辑**：刘 颖 **封面设计**：七星博纳

出版发行：北京邮电大学出版社

社　　址：北京市海淀区西土城路 10 号

邮政编码：100876

发 行 部：电话：010-62282185 传真：010-62283578

E-mail：publish@bupt.edu.cn

经　　销：各地新华书店

印　　刷：唐山玺诚印务有限公司

开　　本：720 mm×1 000 mm 1/16

印　　张：14.5

字　　数：321 千字

版　　次：2021 年 10 月第 1 版

印　　次：2021 年 10 月第 1 次印刷

ISBN 978-7-5635-6491-0 定价：69.00 元

前　言

近年来，随着网络用户的持续增加和新型网络数据业务的不断出现，网络中的数据流量急剧增长。在可预见的未来，网络容量将保持每 10 年增长 100 倍的趋势。不断增长的网络数据流量和动态的网络业务对光网络的信息承载能力提出了更高的要求。

目前，基于相干检测的数字信号处理技术(Digital Signal Processing，DSP)是用于补偿光通道物理损伤的主要手段，一般可以实现色散补偿、信道均衡、解偏振复用、载波频偏估计以及载波相位恢复等功能。但是，随着网络系统规模的扩大、网络灵活性的提升，传统的补偿机制由于复杂度较高、补偿效果有限、自主学习能力较差等原因，面临着极大的挑战和升级需求。因此，发展高效智能、低复杂度且具有自适应能力的损伤感知与补偿机制，对构建未来大规模的光通信系统有重要的现实意义。

随着计算机通信技术的飞速发展，人们搜集信息和处理信息的能力得到极大的提升，如何在复杂系统中发掘蕴含的有用信息，成为诸多领域共同追求的目标。正是在这种需求的驱使下，机器学习技术被提出并且受到广泛的关注。机器学习是一系列智能统计算法的统称，这些算法可以通过自主训练和学习来解决各个领域的多种问题，包括计算机视觉、语音识别、自然语言处理、统计学习、数据挖掘和模式识别等，它是实现人工智能的核心技术之一。机器学习的并行分布处理能力以及它所特有的高度容错性、自组织和自学习能力可以让计算机获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。光通信系统可以利用机器学习的智能学习能力，通过自身的学习过程来捕捉信号所遭受的不同损伤特性，进而实现与信号损伤相对应的补偿功能。

因此，在本书中我们将探索基于机器学习的统计信号处理方法，用于解决光通信系统中灵活多变的物理损伤问题。

作　者

目　录

第1章　机器学习基本原理

1.1　机器学习的研究背景以及发展现状

机器学习是人工智能研究发展到一定阶段的必然产物，可解决人工智能发展的瓶颈和局限性，因此成为20世纪80年代之后人工智能的研究重点。

人工智能的发展时期分为推理期、知识期和学习期。推理期是从20世纪50年代初至70年代初，该时期的主流技术是基于符号知识表示的演绎推理技术。知识期是从20世纪70年代中期至80年代，该时期的主流技术是基于符号的知识表示，通过获取和利用领域知识来建立专家系统。学习期是从20世纪80年代至今，该时期的两大主流技术分别是符号主义学习和基于神经网络的连接主义学习。

在人工智能研究处于“推理期”时，人们认为逻辑推理能力是机器具有智能的重点，只要将逻辑推理的能力赋予机器，机器就拥有了智能。这个时期的代表人物是A. Newell和H. Simon，他们后来研发了能够证明著名数学家罗素和怀特海的名著《数学原理》中的第38条定理的程序“逻辑理论家”，此后更是证明了全部的52条定理，由于在此方面的工作成果获得了图灵奖。后来，随着研究的不断发展，人们意识到，仅仅拥有逻辑推理能力实现不了人工智能。

因此，E. A. Feigenbaum等人认为，机器只有拥有了知识，才会拥有智能。因此，从20世纪70年代中期开始，人工智能研究进入了“知识期”。大量专家系统的问世，使得知识期的人工智能蓬勃发展，但是这些系统中的知识，是由人总结出来并且输入到计算机的，计算机能进行多少推理的工作全靠人输入的知识决定。因此，人们逐渐认识到，不仅由人把知识总结出来然后交给计算机是一项困难的工作，而且得到的专家系统也不能在其他领域得到广泛的应用。于是，一些专家让机器自己学习知识，开始了机器学习的历程。

机器学习曾被R. S. Michalski等人划分为“从样例中学习”“在问题中求解和规划中学习”“从指令中学习”等种类。E. A. Feigenbaum把机器学习划分为“机械学习”“示教学习”“类比学习”和“归纳学习”。20世纪80年代以来，应用最广、研究最多的是“从样例中学习”。

机器学习的发展可以分为三个过程：从样例中学习；统计学习；深度学习。

符号主义学习和基于神经网络的连接主义学习是从样例中学习的两大主流技术。早在人工智能发展的“推理期”和“知识期”，基于符号知识表示，人们通过演绎推理以及获取利用领域的知识建立专家系统取得了很大的成果，因此，符号主义学习成为“学习期”的一大主流技术。符号主义学习的代表有决策树学习和基于逻辑的学习，其中决策树学习因简单而至今仍然被广泛使用。

连接主义学习在20世纪50年代的中后期取得了较大的发展，1986年D. E. Rumelhart重新发明的BP算法，一直是被应用得最广泛的机器学习算法之一。连接主义学习产生的是著名的黑箱模型，又称经验模型。黑箱模型指的是一些内部规律少为人知的情况。

20世纪90年代中期，统计学习占据了主流舞台。代表的算法是支持向量机以及核方法。核方法可以把低维空间的非线性不可分问题转换为高维空间的线性可分的问题，通过某种非线性映射把低维空间的原始数据嵌入高维空间然后再使用通用的线性学习器在这个空间分析和处理数据。核方法逐渐成为机器学习的基本内容。

在21世纪初，由于进入了大数据时代，计算设备的性能提升，人们需要高效地处理种类繁多的数据，以连接主义为基本的深度学习满足了人们的需求，在分析语音、图像等问题上，深度学习技术表现出了优越的性能。

目前机器学习的研究工作主要有以下三个方面：面向任务、认知模型和理论分析。面向任务的主要内容是研究和分析一些学习系统。认知模型是研究人类的学习过程，同时进行计算机模拟。理论分析则是探索各种可能的机器学习算法。

今天，机器学习已经走进了人类的日常生活，机器学习在日常生活中的主要应用有模式识别、数据挖掘、图像理解、统计学习、计算机视觉、语音识别、文本情感分析、自然语言处理以及舆情监控等方面。以上只是机器学习应用的一个大的范围，在这些范围中还有诸多分支，比如在自然语言的处理方面又有很多个分支：机器翻译、自动文摘、信息检索、文档分类、问答系统、文字编辑和校对、语言教学以及说话人识别等方面。

此外，机器学习也在生物信息学、生态学、医学、遗传学及地理学等多领域提供了重要的技术支撑，机器学习主要活跃于数据分析的场景，各种各样的机器学习算法丰富了数据分析的内容。

机器学习的研究除可以提升分析数据的能力外，还可以促进我们理解人类是如何思考和学习的。因此，机器学习的研究有着重要的意义。

使用机器学习来解决实际问题的流程如下：抽象问题，把一个实际的问题抽象成数据和数学问题；选择合适的数据集，选择在需要解决的问题中最能够代表该问题的数据集，将该数据集划分为两部分，分别是训练集和测试集；训练模型，选择合适的机器学习算法来训练模型；调整参数，通过模型的评估与选择，选择好模型性能最优、泛化能力最强的模型；测试模型，训练选择好的模型可以用于数据的测试，测试出它的泛

化能力，然后将其推广应用。

由于机器学习的发展，现在很少亲自编程写算法去完成模型的训练预测等功能。现在多调用现有的工具包来实现以上步骤，机器学习库中包含了一个机器学习项目中部分模块的功能，比如 Python 中的 Scikit-Learn 库，Java 中的 JSAT 等机器学习库，能够实现模型训练、数据集的划分、模型的预测以及交叉验证等功能。机器学习库的使用大大减少了算法实现的时间开支，部分库因为其强大的可视化能力、囊括众多功能的强大包含性，能够非常便利地被用户使用。

1.2 机器学习的分类

机器学习是一门研究怎样利用数学手段，通过计算和利用规律性质来改善系统自身的性能的学科。机器学习的主要内容是研究合适的和高效的机器学习算法，有了算法，就能够从我们要研究的数据中训练好模型，从而在面对新的数据或者新的问题时，训练好的模型能够根据形成的规律和经验来给新的数据提供相应的判断。

因此，机器学习的主要目标是学习、策划和改进数学模型。

机器学习的学习问题是通过对数据集的性能以及规律的研究找到它们之间的依赖关系，模拟它们之间的函数关系。在学习过程中，对任意的输入要输出一个特定的数值，使得输出的数值接近训练器的相应输出。学习就是从给定的函数集中找到最能够代表训练数据的函数关系的函数，即在所有对应的函数关系中搜索的过程。

机器学习按照不同的学习方式可以分为监督学习、半监督学习、无监督学习和强化学习。

1.2.1 监督学习

监督学习是抽取数据的特征形成特征向量，通过学习特征向量以及对应的标签可以得到系统模型。将输入的特征作为训练模型的数据，每个训练数据都有一个自己的标签，在建立模型的过程中，机器学习通过不同的算法建立不同的学习过程，然后最终得到训练好的模型。我们将需要预测的数据送到模型进行分析，从而让模型输出预测的结果。将预测结果与实际结果对比，通过交叉验证来为模型选择合适的参数，在每次迭代之后，增加模型的精度，使预测值与期望值之间的差距趋近于零，从而让模型性能达到最佳。总的来讲，监督学习就是学习已有数据，通过学习已有数据得出理想模型。

监督学习的使用场景有两种：第一种是分类问题，它将输入的数据按照学习好的模型进行分类；第二种是回归问题，用来预测一个具体的值，而用逻辑回归进行分类的主要思想是根据现有的数据对分类的边界线建立回归公式。简而言之，如果预测的输出是离散值，则是分类问题；如果预测的输出是连续值，则是回归问题。所有的分类问

题和回归问题都是监督学习。

常见的监督学习的算法有：K-Nearest Neighbors(k 近邻)算法、Decision Tree(决策树)算法、Naive Bayesian(朴素贝叶斯)算法、Logistic Regression(逻辑回归)算法、Support Vector Machine(支持向量机)算法、Ordinary Least Squares Regression(最小二乘法)和 Ensemble Methods(集成方法)。具体的算法原理将在后续的小节中详细展开。

朴素贝叶斯算法是基于贝叶斯定理与特征条件独立假设的分类方法。朴素贝叶斯算法发源于古典数学理论，有着坚实的数学基础和稳定的分类效率；逻辑回归属于判别式模型，实现较为简单，广泛地应用于工业问题，计算量小，速度快，但是容易欠拟合，并且当特征空间很大时，逻辑回归的性能不是很好；最近邻算法的理论成熟，思想简单，可以生成任意形状的分类边界，但是需要大量的内存以及适当的数据预处理方式；决策树的一大优势就是易于解释，它可以毫无压力地处理特征间的交互关系，不用考虑数据是否线性可分的问题，但是容易过拟合；支持向量机 SVM 具有坚实的统计学理论基础，在许多实际应用中展示出卓越的效果，并且可以很好地应用在高维空间中。

1.2.2 无监督学习

无监督学习输入的数据是没有标签的训练数据，我们希望标记的行为是计算机代替人类去进行的。因此，无监督学习的应用主要有三种场景：在无标记的情况下，寻找最好的特征；先按照选取的分类方式，将数据分为不同类别，然后人为地进行标注；从数据中提取具有代表性的数据，标注之后再进行分类器的训练。无监督学习相比监督学习有一个不能忽视的问题：无监督学习的实现过程会更加困难。无监督学习，在本质上是一个统计手段，在没有给定标签的数据中发现数据内在的联系。它有以下三个特点：无监督学习没有明确的目的，无监督学习不需要给数据贴标签，无监督学习无法量化效果。

常见的无监督学习算法有：k-means(聚类)算法、Learning Vector Quantization(学习向量量化)算法、Mixture of Gaussian(高斯混合聚类)算法、Density Based Spatial Clustering of Applications with Noise(基于密度的 DBSCAN)算法以及 AGglomerative NESting(层次聚类 AGNES)算法。

1.2.3 半监督学习

半监督学习是监督学习与无监督学习的结合，半监督学习可以利用没有给定分类类别的数据提高系统模型的学习性能。让学习器在不依赖外界交互的情况下自动地利用未标记样本来提升学习性能，就是半监督学习。

半监督学习使用的数据，既有标记后的数据，也有未标记的数据。做到半监督学

习有一个隐含的假设——“聚类假设”。其核心要义为“相似的样本，相似的输出”，即所谓假设数据存在簇结构，同一个簇的样本属于同一个类别。因为未标记的样本与有标记的样本是从同样的数据源独立同分布采样而来，它们包含着关于数据分布的信息。在对未标记的数据贴标签时可遵循“人以类聚，物以群分”的原则。

半监督学习中包含另一个假设——“流形假设”，即假设数据分布在流形结构上，并且相邻样本具有相似的输出值。流形假设的应用范围比聚类假设更广。这两种假设的实质都是相似的样本具有相似的输出。

半监督学习可以划分为纯半监督学习和直推学习。纯半监督学习假定训练数据中的未标记样本不是待预测的数据，直推学习是待预测数据，学习的目的就是在这些未标记的样本上获得最优泛化性能。

半监督学习多用于标记数据少、未标记数据多的场景，半监督学习可以用来标记数据。

半监督学习的主要算法有：半监督 SVM 算法、图半监督学习、基于分歧的方法、半监督聚类和生成式方法。半监督学习算法是监督学习算法 SVM 在半监督学习上的扩展应用。图半监督学习是将给定的数据集映射成为一个图，把有标记的样本对应的点想象成为染色后的节点，未标记的节点对应为未染色的节点，一个图能够对应一个矩阵，因此可以基于矩阵运算来进行半监督学习算法的推导和分析。

1.2.4 强化学习

强化学习加入了反馈和评价，在输入数据的同时，模型根据输入的数据，立刻进行自我调整。强化学习常见的模型是标准的马尔可夫决策过程。马尔可夫决策过程是一个四元组(X,A,P,R)，其中 X 表示状态空间，表示决策过程中所有的状态集合，表示机器在决策过程中感知到的环境的描述；A 表示动作空间，指的是在决策过程中，机器能够执行的动作；P 是状态之间的转移函数，使得当前的状态能够按照一定的概率转移到另一个状态上去；R 是奖赏函数，是采取动作空间中的某个行为到达下一个状态之后的回报。

强化学习的过程就是不断地尝试，从错误中学习，找到规律，从而学习到达到目标的方法。它会对行为进行打分，记住高分与低分的行为，再次执行时只执行高分的行为，具有分数导向性。AlphaGo 的火热，就是强化学习在现实中最好的一个应用的案例。

1.3 模型的评估与选择

首先，先来明确几个概念。

- 误差：学习器的实际预测输出与真实输出之间的差值。

- 泛化误差：训练好的模型在新样本输入之后输出的误差称为泛化误差，为了模型有一个良好的性能，泛化误差越小越好。
- 过拟合：泛化误差过大，模型的泛化能力过弱的原因常常是，在学习过程中，把训练样本自身的一些特点当作了所有样本都会具有的一般特征，从而导致出现新的未知的输入时，泛化误差过大，出现过拟合现象。
- 欠拟合：与过拟合相对的是欠拟合，由于数据量过小，导致学习过程对数据的一般特征学习有限，无法得到相同训练集呈现的特征。

在模型训练时，对同一个算法，有不同的参数可以选择，不同的参数集合最终训练得到的模型性能也不同。选择出最优的模型是机器学习的中心任务。

通常，在模型的训练中，会将实验数据集分为测试集和训练集，训练集用来训练模型，测试集用来评估选择的模型性能，通过模型在测试集上的误差来近似代替模型的泛化误差。常规的测试集占据整个数据集的 1/3～1/5，剩余的样本用于训练模型。

除此之外，还可以通过交叉验证法来划分数据集，进行训练以及测试。以 10 折交叉验证为例，将数据集均等分成 10 份，每次选择其中 1 份作为测试集进行验证，其余的 9 份作为训练集进行训练，因此一共可以进行 10 次训练和测试，最终返回的是 10 次测试结果的均值。

因此，模型训练的过程就是，先选择算法，然后对所选的算法进行调参，以测试集上的误差（近似于泛化误差）作为评价标准，选择误差最小的算法及对应的参数，然后再将全部的数据集进行训练，得到最终的学习模型。

除开头提过的误差外，模型性能好坏的度量还有 *P-R* 曲线、ROC 曲线以及混淆矩阵。*P-R* 曲线是查准率-查全率曲线，一般应用在二分类问题上，查准率也指准确率，查全率指有正确分类的数据被挑出来的概率。如果一个模型的 *P-R* 曲线能够被另一个模型的 *P-R* 曲线完全包含住，那么后者的性能优于前者的性能。

表 1-1　二分类结果的混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

查准率 P 与查全率 R 分别定义如下：

$$\left.\begin{aligned} P&=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\\ R&=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\end{aligned}\right\} \tag{1-1}$$

一般而言，查准率高时，查全率偏低；查准率低时，查全率偏高。

ROC 曲线的纵轴是 TPR，即真正例率，横轴是 FPR，即假正例率。ROC 曲线与

P-R 曲线类似，通常也应用于二分类问题，当一个模型的 ROC 曲线能够被另一个模型的 ROC 曲线完全包括，说明后者的性能优于前者的性能。但是当两个模型的 ROC 曲线有相交重叠部分，则需要比较 ROC 曲线下的面积。

$$\left.\begin{aligned}\mathrm{TPR}&=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\\ \mathrm{FPR}&=\frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}}\end{aligned}\right\}\tag{1-2}$$

混淆矩阵能够表达出分类的准确程度，同样可以作为模型性能评估的工具。

1.4 支持向量机

SVM 最早在 1995 年由 Cortes 和 Vapnik 正式提出，它在解决小样本、非线性和高维度模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。小样本指的是与问题的复杂度相比，它所需的样本数较少；非线性指的是 SVM 擅长于处理非线性的问题，它能够将低维度的非线性不可分问题通过核方法变换到高维度的线性可分问题，因此它在处理非线性问题上具有很大的优势；高维模式识别指的是样本的维数很高，它能够在文本分类中表现出卓越的优势。

支持向量机 SVM 是一种典型的二分类器。它的基本思想是，找到一个能够正确划分训练集，而且几何间距最大的超平面。它能够在一定程度上克服机器学习中的不可分问题，能够将一维的数据，转化到二维、三维甚至更高维度。支持向量机在处理数据量不大的问题时性能更强更优。总之，SVM 算法是试图找到一个满足分类要求的超平面，不仅能够将两个类别分割开来，而且能够使训练集中的数据点尽量远离该平面，使得该超平面两侧的空白区域面积最大。SVM 支持向量机算法能够同时处理线性和非线性问题，在手写数据识别、人脸识别、数字调制识别、多用户检测以及信道均衡等方面展示出了巨大的优势。

SVM 算法的优点如下：对高维空间适应良好；在维数比样本数多的情况下也表现良好；能够节约内存；可以选择不同的核函数。

SVM 算法的缺点如下：当数据的特征数比样本数多很多时，算法性能较差；选择参数时需要通过交叉验证法来选择，耗费的时间比较多。

1.4.1 超平面

首先明确一下超平面的概念。在数学中，超平面是一个纯粹的代数概念，它是这样定义的：超平面 H 是从 n 维空间到 $n-1$ 维空间的一个映射子空间，它有一个 n 维向量和一个实数定义。它可以把线性空间分割成不相交的两部分。比如，一维空间中，一个点可以把一条线划分成两个部分；二维空间中，一条直线是一维的，它可以把平面分割成两个部分；三维空间中，一个平面是二维的，它可以把一个三维空间分割成