序列构造神经网络与多维数据分析

王仁武·著

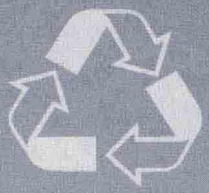
Sequence Constructive Neural Network and Multi-Dim Data Analysis

SASSP 上海社会科学院出版社

序列构造神经网络与多维数据分析

王仁武·著

Sequence Constructive Neural Network and Multi-Dim Data Analysis



上海社会科学院出版社

图书在版编目(CIP)数据

序列构造神经网络与多维数据分析/王仁武著. —上海：上海社会科学院出版社，2008
ISBN 978-7-80745-315-4

Ⅰ. 序… Ⅱ. 王… Ⅲ. ①人工神经元网络②数据库系统-系统分析 Ⅳ. TP183 TP311.13

中国版本图书馆 CIP 数据核字(2008)第 170918 号

序列构造神经网络与多维数据分析

作　　者：王仁武
责任编辑：陆　峥
封面设计：闵　敏
出版发行：上海社会科学院出版社
上海淮海中路 622 弄 7 号　电话 63875741　邮编 200020
http://www.sassp.com　E-mail：sassp@sass.org.cn
经　　销：新华书店
印　　刷：上海宝山杨中印刷厂
开　　本：787×1092 毫米　1/16 开
印　　张：10
插　　页：2
字　　数：180 千字
版　　次：2008 年 11 月第 1 版　2008 年 11 月第 1 次印刷

ISBN 978-7-80745-315-4/TP·001　　定价：25.00 元

前　言

随着社会和科学技术的不断进步，知识逐渐成为创新的核心，知识创新成为知识经济发展的最主要的动力源泉。由于信息技术和网络技术的发展，人们在生活和工作中产生数量浩瀚、种类繁杂的数据资源。成千上万的计算资源、数据资源、软件资源与各种数字化设备和控制系统共同构成了生产、传播和使用知识的重要载体。然而，信息技术和网络技术的迅速发展并未使得用户在处理信息、定位感兴趣的信息资源时变得异常方便。据统计，全世界的数据量每18个月增长一倍，而人们却越来越感到知识十分贫乏，也就是出现“数据爆炸，知识贫乏”的窘境。如何在浩瀚繁杂的数据中分析、获得有益的知识，并且指导人们利用这些知识进行科学正确的决策，是当前智能信息处理研究者的重要任务。在此背景下，数据多维分析以及知识发现(Knowledge Discovery in Database, KDD)应运而生，并迅速发展成为当前的研究热点。

知识发现就需要有能使从大量数据中发现潜在规律、提取有用知识的方法和技术。知识发现不但能够学习已有的知识，而且能够发现未知的知识，得到的知识是“显式”的，既能为人所理解，又便于存储和应用，因此一出现就得到广泛的重视。因此，可以看出，研究如何从数据中分析相关信息的技术和方法将成为一个研究重点。

截止到目前，基于已有数据的知识发现已在市场营销、金融风险评估、生物工程、Internet信息挖掘、电子商务、天文和空间科学、工业制造和控制系统等许多领域得到了广泛的应用。

尽管已有越来越多的利用数据进行信息知识发现的成功应用，但在多维数据分析的研究上仍然遇到许多的困难和挑战，总结起来大致具有以下的特点：

(1) 要处理的对象规模比较大，具有数以亿计的资源、而且伴随着较多的

属性特征；

(2) 可计算性难度高，因为算法需要从大规模数据中分析相关知识，因此算法技术本身要能够在计算速度和计算精度上折衷；

(3) 具有高度的“自动化”的需求；

(4) 技术方法应能处理不同的数据类型，如多媒体数据，文本和视频等数据，以及 web 数据和各类数据的混合体等，并能处理部分数据间的稀疏关系；

(5) 可以对不同类型的数据进行样本化、数据减少、降低数据维的操作。尽管大的样本允许我们管理高维数据，但人类对高维空间的理解和评估仍处于探索状态，这使得我们对分析结果的理解有一定局限。

综上所述，从数据中发现知识的技术研究与应用发展在过去的几十年来已经取得了很大的进展，但对于当前信息爆炸情况下的多维数据分析的相关技术研究，还有许多技术上的问题和限制尚待解决和完善。

研究背景

显然，任何一个没有学习能力的系统都很难被认为是一个真正的智能系统，但随着机器学习研究及应用的不断发展，尽管“学习机制”还是研究动力之一，然而，“烦恼的网络”危机使得更为重要的推动力来自“有效利用”信息。当前，很多传统领域借用机器学习来提高研究水平，应用驱动的机器学习方法层出不穷，特别是基于机器学习的数据分析方法已成为解决复杂问题的关键技术之一。因此，当前机器学习的角色也逐渐发生了转变，已发展到一个新阶段。

近年来，文本与图像占信息的绝大多数，在文本分析与自然语言理解上，数据资源建设逐渐完善，人们关注的焦点是机器学习及其在这些数据资源上的深层次应用。当前，针对信息的复杂多样性，涌现出很多新的机器学习方法。比如：可用于特征抽取的流形机器学习，即稀疏数据的非线性处理方法；改善机器人适应环境变化性能的增强学习；可用于药物设计的多实例学习和半监督学习；广泛用于搜索引擎的 Ranking 学习；能够快速过滤海量数据的数据流(Data stream)学习，等等。虽然这些新的机器学习方法仍处于探索和实验观察阶段，但是，它们已充分表明，基于机器学习的数据分析方法越来越成为解决复杂问题的关键。

因此，现阶段机器学习研究不应再过多地强调模拟人的学习能力，应该把机器学习真正当成一种支持技术(也就是说，它是一种重要手段而非目的)，考虑不同领域甚至不同学科对机器学习的需求，找出其中迫切需要解决的共性问

题，并进行深入研究。有人把这种视角下的机器学习称为“普适机器学习”(Pervasive Machine Learning，简称 PML)。

当前数据信息发展极为迅速，考虑到经验非线性方法，如人工神经网络(ANN)。这种方法利用已知样本建立非线性模型，克服了传统参数估计方法的困难。但是，这种方法缺乏一种统一的数学理论，且网络学习是“黑盒”模型，还存在神经网络结构选择问题、局部极小点问题等。另外也有一些神经网络的动态方法，但这些方法其网络构造主要还是进行误差反馈方法，处理的对象主要还是低维的小容量数据，对使用神经网络方法处理多维的海量数据则没有系统性地进行。因此本书研究侧重于具体的应用模型，从网络构造的基本方法出发，进行神经网络的构造，部分地克服传统神经网络方法中存在的一些缺陷，解决传统方法在海量数据分析中遇到的困难。

研究目的与意义

本书涉及的研究领域有：智能科学、计算机科学、数学、神经生理学等领域的交叉研究，且与认知、数学等领域有潜在的关系。本书的主要研究目的是针对大容量多维数据的特征，研究动态序列构造神经网络并且把其应用到多维数据分析的应用中。

本书在多维数据分析的需求基础上，结合序列构造神经网络，给出了具体的多维数据分析方法，在序列构造神经网络的研究基础上，针对具体的网络构造模型、算法进行了研究，重点研究网络构造过程的基本模型和构造算法，并结合不同的神经网络元结构，提出了具体的网络构造算法。提出了一些新的算法，并通过实验证明算法的可行性，通过这些算法能够实现序列神经网络的构造，从而为数据的多维分析创造条件，进而为知识发现提供高效途径，同时能改进大数据量数据分析的能力，从整体上提高信息知识发现的水平。

本书选择了基于序列构造神经网络的多维数据分析来进行研究，目前，国内可见的多数信息分析与数据挖掘的文章集中于对相关规则、粗糙集等技术的研究，神经网络的研究集中在已有的难以训练的神经网络模型，例如：结合遗传算法的前馈神经网络、SVM 等。对于小容量数据，这些模型都可以完成相关的数据分析任务；但是，对于数据量大、维数高的实际数据而言，可计算能力就变差。因此本书研究的一个主要内容则是序列神经网络的构造，通过序列构造神经网络对大数据量数据进行剖面映射，然后在构造好的神经网络上进行数据内容分析，其分析过程就是对原始数据映射数据的信息融合过程，把实际数据的

关系转换为序列构造神经元之间的关系分析，通过这种神经网络的非线性映射，大大减少了原始样本点的数据量，具有较大的实际意义。本书首次提出了利用序列构造神经网络模型对多维数据结构进行分析研究。为了保证一定的完整性，本书首先讨论序列构造神经网络模型、算法，然后研究如何在此基础上进行多维数据知识发现，最后给出实验结果和结论，并通过实际应用进行了说明。实验表明，本书的研究方法和研究结果是可行的和合理的。

本研究的意义在于：

(1) 为多维数据分析提供一个可行的理论模型和解决思路。多维数据的分析处理适应了当前信息发展的趋势，在许多相关领域都有着广泛的应用前景。本书试图研究一种新型数据分析模式，并希望在关键技术上有所突破；

(2) 序列构造神经网络模型、理论算法的研究，在智能神经研究领域进行理论创新和应用创新尝试；

(3) 通过序列构造神经网络实现对原始样本数据的映射，序列构造神经网络具有构造简单、快捷的特点，在网络构造速度和数据精度上做了折中，为神经网络用于多维数据的研究提供了可能；

(4) 在基于序列构造神经网络的基础上，提出了通过神经元关系进行数据融合分析的方法，为多维数据的实际应用提供了思路；

(5) 随着相关研究的深入，可通过分析序列构造神经网络映射关系，来揭示样本数据的内在联系，进而促进对外部世界事物的认识。

主要研究内容和创新点

作为多维数据分析的方法，传统神经网络学习方法因具有难以训练的“黑盒”特性，在实际数据量爆炸的情况下，网络难以训练，在数据具有多维特征的情况下更加有困难。前馈网络使用梯度下降算法，因而遇到了存在局部极小、收敛速度慢等问题。本书提出了动态构造判别序列的神经网络方法，把静态神经元网络结构转换为动态神经网络构造问题；同时，把各个不同剖面神经元的构造网络与多维数据分析相结合，并且相应地提出了相关算法、模型。动态构造神经网络使得对信息爆炸情况下多维数据分析提供了一种方法，使得我们对复杂的、信息不完备的对象的分析提供了解决问题的方案。通过特定神经元类型，使得神经网络的构造更加直观、方便，而且有许多传统固定网络所不具备的性能。

对于多维数据的分析，动态构造方式适应这种灵活数据变化的需求，把不

同角度构造的序列神经网络进行信息合成，这样可以设计出学习速度快，在精确度和速度之间取得折中。由于多维数据分析是在本书提出的模型上进行的，并且在实际信息系统中得到了应用，因此有着很强的应用效果，总结起来创新点大致如下：

(1) 为适应当前信息技术高速发展过程中数据具有复杂多变和多维的特点，同时考虑到固定结构的神经网络模型在进行网络训练过程中难以适应这种类型数据的特点，本书提出了神经网络的动态构造方法，即序列构造神经网络的网络模型。该模型通过使用动态构造方法把神经网络的构造转换为判断规则序列，采用“分而治之”的思想对原有数据进行映射，通过序列构造神经网络的不同神经元序列，对训练数据集合进行映射。

(2) 针对序列构造神经网络的理论和构造方法进行研究，分析了序列构造神经网络构造的动态特点，提出了相应的训练方法和通用构造规则；说明了序列构造神经网络对复杂多变的多维数据分析的能力，并且从理论上对其进行了探讨。另外，通过实验验证了提出算法的实际可操作性，同时，把实验效果与SVM等其他方法进行了对比，进而说明提出的神经网络模型和相关算法的实用性。

(3) 针对多维数据的分析，提出了神经网络对不同数据剖面规则的映射，通过不同规则映射的神经网络，分析其数据点之间的关系，从而对多维数据点进行分析；提出了相应的分析方法，这些分析方法通过分析数据点之间的关系，通过序列神经网络神经元对其进行映射，为数据分析任务的客观性提供了基础，并且把这种方法应用于实际房产租赁指数的分析中，为此类系统的多维分析提供了新的思路和处理方法。

在研究过程中，本书得到了博士生导师陈家训教授的悉心指导。复旦大学的高传善教授、上海交通大学的白英彩教授、华东师范大学的王能教授、东华大学的邵世煌教授和曹奇英教授也曾给予了至今难以忘怀的帮助和鼓励。在研究过程中，笔者也参考了大量的专业文献，在此表示谢忱！

在撰写过程中，笔者努力反映专业的最新成果和研究动态，但是由于技术研究的日新月异，加上笔者才疏学浅和时间仓促，不足甚至错误之处在所难免，请广大读者批评指正，提出宝贵意见，以便笔者能提高和修正。

内容提要

序列构造神经网络是一种新的神经网络的方法模型，是基于神经网络中神经元的动态构造技术来构建的，它符合机器对数据学习的本质要求：即动态、主动学习的过程，而非静止、被动的学习过程。在本书中，作者研究了序列构造神经网络的原理、构造方法和实现方法，并用序列构造神经网络对多维数据分析进行研究和应用探索。

我们知道，在机器学习的研究中，创立新的理论模型和算法一直被认为是智能机器学习领域急需解决的问题。神经网络技术在过去的几十年里得到了长足的发展，神经网络在模式识别和数据分析等应用领域得到了广泛的应用，其中前馈网络扮演了重要角色。当前信息技术飞速发展，海量复杂的数据应运而生，而且这些数据大都具有多维的特点；新的技术发展也为神经网络提出了新的问题，由于网络内部神经元连结结构固定，基于内部神经元全连结条件下的传统神经网络或其改进方法，在进行网络训练之前需预先确定其所使用的神经元数量，此种模式难以适应复杂变化的多维数据分析的应用需求。

针对传统的前馈网络模型存在隐含层难以确定、算法收敛速度慢、网络结构难以训练、难以适应复杂变化的多维数据分析的应用需求等诸多问题，以及在工程中的实际应用和使用效果不甚理想等现状，作者提出并探讨了序列构造神经网络模型。

序列构造神经网络模型并不是从固定网络结构出发来进行网络内部神经元的训练，而是通过使用动态构造的内部神经元对训练数据的不同子集进行空间映射，动态地构造内部神经元进而构造神经网络的结构，使得网络结构适应外部训练的数据要求及其后续的变化。该训练过程具有动态伸缩和可以适应外部数据变化和多维的优点。本书中在相关内容的研究过程中穿插实验效果

的比对，进而说明序列构造神经网络的特点。本书就此理论的确立和开展进行了以下诸多方面的研究工作：

首先对序列构造神经网络基础理论、方法进行了分析和研究。本书对序列构造神经网络与传统方法的不同之处进行了比较，对序列构造神经网络的构造方法和学习过程进行了阐述，描述了网络构造过程中的基本原理，讨论了其网络构造的方法；同时，也通过实验比对了序列构造神经网络和传统全连接方法在训练速度上的差别。实验结果显示，在相同的识别误差情况下，序列构造神经网络拥有较快的训练速度，比较适合对大数据量的分析，为后续章节的研究奠定了理论基础。

其次，针对序列构造神经网络的构造算法的一般过程进行了讨论，并提出了若干具体的、可行的神经网络内部神经元构造算法。对这些算法进行了分析验证，大多经实验证实是有效的，并进一步说明了这些算法对于多维数据分析的有效性。而且，序列构造神经网络算法不寻求空间的最优超平面（事实上在多维情况下是 NP 难问题），而是通过序列神经元对数据进行描述。通过这种方法，可以非常容易地构造出对原始多维数据空间序列神经元的映射。这种映射过程，适用于经常变换的、时序性强的、多维数据的应用场合。序列构造神经网络模型相对于传统固定结构的网络而言，能够满足对多维数据进行剖面剖分的数据分析的要求。

本书结合序列构造神经网络的优点，提出了基于序列构造神经网络的多维数据分析方法，针对多维数据的特点通过对数据不同侧面的序列构造神经网络分析，给出序列构造神经网络对于数据点的描述，进而对不同神经网络进行剖面分析，通过不同序列构造神经网络分析，从而完成对已有数据点集的多维数据分析。另外，本书给出了在序列构造神经网络过程中的数据预处理方法，并且给出数据的多侧面序列构造神经网络的分解，以及序列构造神经网络的多维数据分析算法。

最后，文章结合所研究的序列构造神经网络及多维数据分析的理论和算法，在当前国家推行廉租房的时代背景下，把研究成果应用到现实房产租赁指数系统中。通过神经网络对房产租赁信息的不同剖面分析，从中发现有价值的租赁指数信息，为廉租房政策的制定和实施提供参考依据，有着很强的实际应用价值。

通过上述一系列有价值的研究工作，作者论证了序列构造神经网络对于多维数据分析处理的优势，并给出了现实房产租赁指数多维数据分析的应用探

索。序列构造神经网络对于多维数据的分析应是一种新颖而又有着广泛应用前景的神经网络构造模型。

本书根据其内容分为七章，安排如下：

第一章：主要介绍多维数据分析的发展概况，应用背景，以及相关问题，对当前各种多维数据分析方法进行了探讨。

第二章：主要介绍人工神经网络的发展概况，应用背景，以及相关问题，对当前固定神经网络训练过程中的主要存在的问题进行了描述。

第三章：从动态构造的思想出发，分析样本数据，给出了序列构造神经网络的训练过程、识别过程，讨论了内部神经元的构造和相关算法的关键技术，并对网络序列构造的机理进行了分析。

第四章：提出了实际多类样本的网络内部神经元构造方法，给出了通用的动态构造原理和方法，并且对网络训练过程中的收敛性进行了分析，给出了网络构造的一般通用过程。

第五章：结合具体不同神经元类型，对网络隐含层神经元进行实现，对其神经元构造的机理进行了分析，并且对不同类型的神经元给出了具体的算法；最后通过仿真实验和比对分析，验证了所提出的相关算法，并对不同神经元的构造进行了评估。

第六章：结合已经提出的神经网络，对多维分析的方法进行了研究，把不同侧面的信息与神经网络构造相结合，使用序列构造神经网络大大减少原始样本的信息，起到信息约简的作用；然后提出了对构造好的神经网络进行分析的方法，近而来解决数据多维分析的问题；通过事例验证了该方法能够明显地降低多维分析的计算时间，并通过对不同侧面信息的神经网络构造和信息合成，使之具备了多维数据分析处理的能力。

第七章：本章以房产行业中的多维数据分析应用为研究内容，考虑了房产信息数据大量的特点，在序列构造神经网络的基础上建立了一个基本的分析模型用于房产信息的多维分析，并且通过具体的数据来验证算法的可行性。

目　录