统计学前沿论文最新成果

2018.11.02 方建勇

提示：采用手机 safari 微软翻译技术

[1] **arXiv: 1811.00007 [pdf,其他]**

深潜变量模型的介入鲁棒性

拉斐尔祖特尔, Đorđe Miladinović,斯蒂芬. 鲍尔, Schölkopf

主题:**机器学习 (ML)**;机器学习 (cs。LG

学习迎刃而解表示的能力，在高维、非结构化数据中分割底层的变化源对于数据高效和稳健地使用神经网络至关重要。在最近的时间里提出了旨在实现这一目标的各种办法，因此，确认现有工作是指导进一步发展的一项关键任务。以前的验证方法侧重于生成因子和学习特征之间的共享信息。然而，稀有事件或多重因素对编码的累积影响仍然 uncaptured。我们的实验表明，这已经成为一个简单的，无噪声数据集明显。这就是为什么我们引入了介入性鲁棒性评分，它提供了对学习表现的稳健性的定量评估，在对生成因子和变化的有害因素的干预方面。我们展示了如何从标记的观测数据 (可能混淆) 中估算出此分数，并进一步提供了一种在数据集大小中线性扩展的高效算法。广泛的实验说明了我们因果动机框架的好处。

[2] **arXiv: 1811.00062 [pdf, ps,其他]**

下一元预测序列建模方法的跨学科比较

尼克税,艾琳 Teinemaa, Sebastiaan, 面包车 Zelst

主题:**机器学习 (ML)**;计算和语言 (cs。CL);机器学习 (cs。LG

序列性质的数据在许多应用程序域中出现，如文本数据、DNA 序列和软件执行跟踪。不同的研究学科已经开发出了从这些数据集学习序列模型的方法: (i) 在机器学习领域中，如 (隐) 马尔可夫模型和递归神经网络已经开发并成功地应用于广范围的任务, (ii) 在过程挖掘过程中发现技术旨在生成人-解释描述模型，而 (iii) 在语法推理领域中，重点是查找形式语法的描述性模型。尽管它们具有不同的焦点，但这些字段共享一个共同的目标-学习一个准确描述基础数据中行为的

模型。这些序列模型是生成的，我. e，它们可以预测给定未完成序列后可能发生的元素。到目前为止，这些领域的发展主要是孤立的，不存在比较。本文提出了一种跨学科的实验评估方法，将序列建模技术与四实际序列数据集的下一元预测任务进行比较。研究结果表明，在精度方面，通常没有目标可解释性的机器学习技术优于过程挖掘和语法推理领域的技术，目的是产生解释模型。

[3] **arXiv: 1811.00074 [pdf,其他]**

通过精确保证高效收集互联车辆数据

涅金 Alemazkoor,哈迪总统

主题:**应用 (统计。AP)**;信号处理 (eess。SP

互联车辆以非常高的频率传播详细数据，包括其位置和速度。这些数据可用于交通系统的精确实时分析、预测和控制。这种分析的突出挑战是如何不断收集和处理海量数据。为了应对这一挑战，高效收集数据对于防止负担过重通信系统和过度计算和内存容量至关重要。在这项工作中，我们提出了一个高效的数据收集方案，它只选择和传输一小部分数据，以减轻数据传输负担。作为演示，我们使用了建议的方法来选择从安全试验模型部署数据集中提供的 1 万个互联车辆行程中传输的数据点。结果表明，根据所需精度，采集比可小至 0.05。同时，利用所提出的数据收集方法，对行程时间估计精度进行了仿真研究。结果表明，所提出的数据收集方法可以显著提高出行时间估计的准确性。

[4] **arXiv: 1811.00097 [pdf,其他]**

基于模型聚类的交叉变异进化算法

沙龙 m. 麦克尼古拉斯,保罗. 麦克尼古拉斯,丹尼尔 a. Ashlock

主题:**计算 (统计)**;机器学习 (统计 ML)

在基于模型的聚类问题中，期望最大化 (EM) 算法几乎无处不在，用于参数估计;然而，由于其单一的路径，单调的性质，它可能会陷入局部极大值。而不是使用 EM 算法，进化算法 (EA) 开发。此 EA 有助于不同的搜索的健身景观，即，可能性表面，利用交叉和突变。此外，该 EA 是一种有效的 "硬" 基于模型的聚类方法，因此它可以被看作是 k-均值算法的一种泛化，它本身等同于一种高斯混合模型的分类 EM 算法，具有球形组件协方差。在多个数据集上对 EA 进行了说明，并将其性能与 k-均值聚类以及基于模型的聚类与 EM 算法进行了比较。

[5] **arXiv: 1811.00115 [pdf,其他]**

降维具有可量化的缺陷：两个几何边界

Kry 益洲吕,加文蓝田伟丁,瑞通黄,罗伯特 J. 麦肯

评论:第三十二次神经信息处理系统会议 (NIPS 2018)，蒙特利尔，加拿大

期刊编号:神经信息处理系统 (NIPS 2018)

主题:**机器学习 (ML)**;机器学习 (cs。LG

本文从定量拓扑的角度研究了信息检索中的维数约简 (DR) 映射。特别是，我们表明，没有 DR 地图可以实现完美的精度和完美的召回同时。因此，一个连续的 DR 映射必须有不完美的精度。我们进一步证明了李氏连续 DR 地图的精度上的上限。虽然精度是信息检索设置中的自然度量，但它并不测量检索到的数据的"错误"。因此，我们提出了一种新的基于沃瑟斯坦距离的测量方法，具有类似的理论保证。在我们的证明的关键技术步骤是一个特定的优化问题的 $我$ 2-在一组约束的分布沃瑟斯坦距离。我们为这个优化问题提供了一个完整的解决方案，在技术方面可以有独立的利益。

[6] **arXiv: 1811.00153 [pdf,其他]**

动态种群增长模型的序列设计方法

汝张,专场德文郡,普里塔姆 Ranjan

评论: 36 页

主题:**方法 (统计)。我)**;计算 (统计 CO)

全面了解各种害虫的种群增长，对于有效的作物管理往往是至关重要的。我们的激励应用来自于校准两延迟丝光 (全爪) 模型，用于模拟螨 (科赫) 或欧洲红螨在苹果树叶上的生长，并降低产量。我们将重点放在反问题上，即估算出一个与现场观测相匹配的、可产生计算机模型输出的该模型的参数/输入集。场观测和贸变输出的时间序列性质使得反问题比标量值模拟器案例更具挑战性。
在精神上，我们遵循计算机实验的流行顺序设计框架。然而，由于时间序列的响应，基于奇异值分解的高斯过程模型用于代理模型，并对后续点的选择提出了新的期望改进准则。本文还提出了从最终代理提取最优逆解的新判据。与现有技术相比，三模拟实例和现实生活中的贸标校准问题已被用来证明所提方法的准确性更高。

[7] **arXiv: 1811.00183 [pdf,其他]**

扬声器 Diarization 的有效度量学习流水线设计

维韦克西瓦拉曼 Narayanaswamy,廖四辉 j. 德博契亚格拉杰,欢歌,列斯 Spanias

主题:**机器学习 (ML)**;机器学习 (cs。LG);声音 (cs。SD);音频和语音处理 (eess)

最先进的扬声器 diarization 系统利用来自外部数据的知识，以预先训练的距离度量的形式，有效地确定相对的扬声器标识到看不见的数据。然而，最近的许多关注点是选择合适的特征提取器，从预训练的*我*一通过不同的序列建模体系结构 (例如 1 维 CNNs、LSTMs、注意模型) 了解到制图表达的向量，同时采用现成的公制学习解决方案。本文认为，无论特征提取器如何，都必须仔细设计一个度量学习管道，即损失函数、采样策略和 discrimnative 边距参数，用于构建稳健的 diarization 系统。此外，我们还建议采用细粒度验证过程，以获得对公制学习管道泛化能力的综合评价。为此，我们测量不同语言扬声器的 diarization 性能，以及录制中扬声器数量的变化。通过实证研究，我们为不同设计选择的有效性提供了有趣的见解，并提出了建议。

[8] **arXiv: 1811.00203 [pdf,其他]**

潜伏高斯计数时间序列建模

翼速贾,圣斯特凡诺斯 Kechagias,詹姆斯李伍斯,罗伯特隆德,阿达姆库斯与 Pipiras

主题:**方法 (统计)。我)**

本文提出了固定计数时间序列 copula 建模的理论和方法。这些技术使用潜伏的高斯过程和分布变换来构造具有非常灵活的相关特征的静止序列，它们可以具有任何预先指定的边际分布，包括经典泊松、广义泊松、负二项式和二项式计数结构。基于计数序列的均值和相关函数的高斯伪似然估计范式是通过一些新的厄米扩展来开发的。研究了粒子滤波方法，以近似计数序列的真似然。在这里，与隐马尔可夫模型和其他 copula 似然逼近的连接进行。该方法的有效性得到了证明，并利用这些方法来分析一个计数序列，其中包含 1893 年以来在大联盟棒球比赛中未打棒球的年度数字。

[9] **arXiv: 1811.00255 [pdf,其他]**

HMLasso：用于高维度和高度缺失数据的套索

普宪藤泽,威一郎川

主题:**机器学习 (ML)**;机器学习 (cs。LG

稀疏回归 (如套索) 在处理高维度数据数十年中取得了巨大成功。但是，很少有适用于缺失数据的方法，这通常发生在高维数据中。最近, CoCoLasso 被建议处理高维缺失数据，但它仍然遭受高度缺失的数据。本文提出了一种新的用于高度缺失数据的套索式回归技术，称为 "HMLasso"。我们使用平均估算协方差矩阵，这是臭名昭著的一般由于其估计偏差的缺失数据。但是，通过使用与成对协方差矩阵的有用连接，我们有效地将其合并到套索中。由此产生的优化问题可以看作是 CoCoLasso 与缺失比率的加权修改，对于高度缺失的数据非常有效。根据我们的知识，这是第一种能够有效处理高维度和高度缺失数据的方法。表明该方法有利于协方差矩阵的非渐近性质。数值实验表明，该方法在估计误差和泛化误差方面具有较高的优越性。

[10] **arXiv: 1811.00293 [pdf,其他]**

噪声整流神经网络中深信号传播的临界初始化

奥亚普里托里厄斯, Biljon,史蒂夫克朗,赫尔曼乔纳斯坎佩尔

评论: 20 页, 11 个数字，在第三十二会议上接受的神经信息处理系统 (NIPS 2018)

主题:**机器学习 (ML)**;机器学习 (cs。LG

随机正规化是一个重要的武器，在一个深度学习从业者的阿森纳。然而，尽管最近的理论进展，我们对噪声如何影响深神经网络中信号传播的理解仍然有限。通过在均值场理论的基础上扩展最近的工作，建立了随机转正神经网络信号传播的新框架。我们的噪声信号传播理论可以包含几种常见的噪声分布，包括加法和乘法高斯噪声以及差。利用该框架研究了噪声 ReLU 网络的初始化策略。我们表明，没有关键的初始化策略存在使用加性噪声，随着信号传播爆炸，无论选择的噪声分布。对于乘法噪声 (例如辍学)，我们确定依赖于噪声分布的第二时刻的替代关键初始化策略。实际数据的模拟和实验证实了我们所建议的初始化能够在深网络中稳定地传播信号，而使用初始化无视噪声也无法做到这一点。此外，我们还分析了输入之间的相关动态。更强的噪声正规化被显示为减少对噪声 ReLU 网络的输入的歧视性信息能够传播的深度，即使在初始化处于临界时也是如此。我们支持我们对这些可训练深度的理论预测以及模拟，以及 MNIST 和 CIFAR-10 的实验。

[11] **arXiv: 1811.00306 [pdf,其他]**

因子数过高估计时的高维因子模型的一致估计

Haeran 町 Barigozzi

主题:**方法 (统计)。我)**

n 维向量时间序列的一个高维 r 因子模型的特点是在 r th 和 (r + 1) 之间存在较大的 eigengap (增加 n)-协方差矩阵的最大特征值。因此，主成分分析法 (PCA) 是一种常用的因子模型估计方法，它的一致性，在正确估计时，在文献中得到了很好的确立。然而，在有限样本中，各种因子数估计往往缺乏明显的 eigengap。我们在经验证明，他们倾向于过度估计系数的存在中的相关性 (非因子驱动) 组件，并进一步证明, r 的过度估计会导致 PCA 中的不可忽略的错误估计。为了解决这个问题，我们提出了两个新的估计值，基于对样本特征向量的项进行上限或缩放，而不知道真正的因子数，它比 PCA 估计对 r 的过度估计更敏感。我们在理论上和经验上表明，两个估计值成功地控制了过度估计误差，并证明了它们在宏观经济和财务时序数据集上的良好表现。

[12] **arXiv: 1811.00314 [pdf,其他]**

空间函数线性模型及其估计方法

婷婷黄,吉尔伯特萨波塔,汇文王,姗姗

主题:**计算 (统计 CO)**

经典的函数线性回归模型 (FLM) 及其扩展，基于所有个体相互独立的假设，得到了很好的研究，并被许多研究人员使用。这种独立性假设有时在实践中被违反，特别是在科学学科中收集网络结构数据，包括市场营销、社会学和空间经济学。然而，相对较少的研究已经审查了 FLM 在数据与网络结构的应用。提出了一种新的空间函数线性模型 (SFLM)，它将空间自回归参数和空间权重矩阵融合到 FLM 中，以适应个体间的空间依赖性。该模型是相对灵活的，因为它利用 FLM 处理高维协变量和空间自回归 (SAR) 模型捕获网络依赖关系。开发了一种基于功能主成分分析 (针对) 和最大似然估计的估计方法。仿真研究表明，当网络结构存在时，我们的方法与 FLM 的针对方法同样具有良好的性能，并优于后者。还使用了真实的天气数据来证明 SFLM 的效用。

[13] **arXiv: 1811.00410 [pdf,其他]**

关系推理的扩张 DenseNets

Antreas 安东尼奥,阿格涅斯卡佳,埃利奥特 j. 克劳利,阿莫斯 Storkey

评论:扩展抽象

主题:**机器学习 (ML)**;机器学习 (cs。LG

尽管它们在许多任务中表现出色，但深度神经网络常常在关系推理中挣扎。最近，引入了一个用于考虑对象对之间关系的插件关系模块，从而纠正了这一问题。不幸的是，这是搜寻昂贵。在这个扩展的抽象，我们表明, DenseNet 合并扩张卷积擅长于 CLEVR 数据集的关系推理，允许我们放弃这个关系模块及其相关的费用。

[14] **arXiv: 1811.00423 [pdf, ps,其他]**

乘性潜力模型

丹尼尔 j. 泰特,布鲁斯 j. Worton

主题:**机器学习 (ML)**;机器学习 (cs。LG

动态系统的贝叶斯建模必须在提供过程的完整机械规范之间达成妥协，同时保持灵活处理数据相对于模型复杂性而言稀疏的情况，或完全规范是很难激励的。潜力模型通过指定一个加性潜伏高斯过程 (GP) 强制期限的简洁线性演化方程来实现这一双重目标。在这项工作中，我们扩展了潜伏力框架，允许 GP 与潜在状态之间的乘法交互，从而对轨迹的几何形状进行更多的控制。遗憾的是，推断不再简单，因此我们引入了基于逐次逼近方法的逼近，并利用仿真研究对其性能进行了研究。

[15] **arXiv: 1811.00439 [pdf,其他]**

二元介质非稀有二元结果的精确参数因果中介分析

马可多雷蒂,拉吉,埃琳娜 Stanghellini

评论: 23 页, 2 数字

主题:**方法 (统计)。我)**

本文利用二进制介质的设置，推导出自然直接和间接效应在赔率比尺度上的精确参数表达式。我们建议的效果分解不要求结果是罕见的，并概括现有的，允许曝光和中介之间的相互作用和混淆协变量。此外，它还概述了因果效应与对应路径特定的逻辑回归参数之间的解释关系。我们的研究结果适用于来自波斯尼亚-黑塞哥维那的小额供资实验的数据。同时，还实现了一种基于稀有结果假设的估计比较的仿真研究。

[16] **arXiv: 1811.00450 [pdf,其他]**

C + + 中的 R 友好多线程

托马斯 Nagler

主题:**计算 (统计 CO)**

从 R 调用多线程 c++ 代码有其危险。由于 R 解释器是单线程的，因此不能检查用户是否中断或从多个线程打印到 r 控制台。但是，可以从主线程与 R 同步。r 包 RcppThread (当前版本 0.5.0) 包含一个头仅 c++ 库，用于线程安全通信与利用此事实的 R。它包括线程的 c++ 类、线程池和常规与 R 同步的并行循环。本文介绍了包的功能，并给出了它的用法示例。同步机制也可能适用于其他线程框架。基准表明，尽管同步会导致开销，但 RcppThread 的并行抽象与在统计计算中遇到的典型场景中的其他常用库具有竞争力。

[17] **arXiv: 1811.00457 [pdf,其他]**

利润-最大程度的 a/b 测试

埃里亚，罗恩.伯曼

主题:**应用 (统计。AP)**;方法 (统计)。我

营销人员通常使用 a/b 测试作为战术工具，在测试阶段比较市场治疗，然后将更好的治疗方法部署到其余的消费者群体中。尽管这些测试传统上使用假设测试进行了分析，但我们重新将此类战术测试视为在测试的机会成本 (某些客户获得次优处理) 和潜在损失之间的明确权衡。与将次最优处理部署到其余的人口有关。

我们推导出利润最大化测试大小的闭合形式表达式，并显示它比通常建议的假设测试小得多，特别是当响应是嘈杂或总人口较少时。使用小的维持组的常见做法可以通过不对称先验来合理化。建议的测试设计实现了几乎相同的预期遗憾作为灵活，但难以实施的多武装强盗。

我们在三种不同的营销环境中展示了该方法的优势--网站设计、展示广告和目录测试--我们从过去的数据中估算先验信息。在所有三种情况下，最佳样品尺寸比传统的假设测试要小得多，从而带来更高的利润。

[18] **arXiv: 1811.00462 [pdf, ps,其他]**

广义线性模型大数据分析的分数匹配代表性方法

可人李,洁阳

评论: 26 页, 4 数字

主题:**方法 (统计)。我)**

我们提出了一种快速有效的策略，称为代表性方法，用于大数据分析的线性模型和广义线性模型。对于大数据集的给定分区，此方法为每个数据块构造一个具

有代表性的数据点，并使用代表数据集来匹配目标模型。在时间复杂性方面，它与文献中的次像素采样方法一样快。在效率方面，其参数估计精度优于分治法。通过全面的仿真研究和理论论证，提出了两种代表性的方法。对于线性模型或具有平坦逆链函数的广义线性模型和连续变量的适中系数，我们建议平均代表(MR)。对于其他情况，我们建议得分匹配代表（SMR）。作为航空公司实时性能数据的说明性应用，MR 和 SMR 在可用时与完整数据估计一样好。此外，建议的代表性策略是分析分散在互联计算机网络上的海量数据的理想选择。

[19] **arXiv: 1811.00465 [pdf,其他]**

通过主次要分配问题学习签名行列式点过程

维克多-伊曼纽尔布鲁内尔

评论:NIPS 接受的较短版本（神经信息处理系统）2018

主题:**统计学理论（数学。ST)**

对称行列式点过程（DPP）是一类概率模型，用于对表现出排斥行为的项目随机选择进行编码。他们在机器学习中吸引了大量的注意力，当需要返回各种各样的项目时。抽样和学习这些对称的民进党是相当清楚的。在这项工作中，我们考虑了民进党的新类，我们称之为民进党，在那里我们打破了对称性，并允许有吸引力的行为。我们通过片刻的方法为学习签名的 DPP 设置了基础，解决了一类矩阵的所谓主分配问题 K，满足 $K_{我,J}=\pm K_{J,我}$, $我\neq J$，在多项式时间。

[20] **arXiv: 1811.00488 [pdf,其他]**

超高维加法部分线性模型的稀疏模型辨识与学习

信义李,李王,丹内特尔顿

主题:**方法（统计)。我)**

加法部分线性模型（APLM）将非参数回归的灵活性与回归模型的吝啬结合起来，广泛应用于多元非参数回归中的一种常用工具，以缓解 "维度 "。在实践中提出的一个自然问题是在非参数部分中结构的选择，即连续协变量是否以线性和非参数形式进入模型。本文提出了一种用于超高维杀伤人员地雷的同时稀疏模型辨识和学习的综合框架，其中线性和非参数分量可能大于样本大小。我们提出一个快速有效的两阶段程序。在第一阶段，我们将非参数函数分解为线性部分和非线性部分。用恒样条基逼近非线性函数，提出了用自适应群套索选择非零分量的三重惩罚方法。在第二阶段，利用高阶多项式样条曲线对所选协变量进行了backfitted，并应用样条曲线局部线性平滑法求出估计的渐近正态性。该过程显示为模型结构识别的一致性。它能正确有效地识别零、线性和非线性分量。可以对线性系数和非参数函数进行推断。我们进行仿真研究以评估该方法的性能，并将该方法应用于玉米基因型的茎尖分生（SAM）的数据集。

[21] **arXiv: 1811.00512 [pdf, ps,其他]**

通过模拟学习学习光束搜索策略

雷纳托 Negrinho,马修 r. 安东尼·葛姆雷,杰弗里 j. 戈登

评论:发布于 NIPS 2018

主题:**机器学习 (ML)**;人工智能 (cs。AI);机器学习 (cs。LG

在结构预测问题中，波束搜索被广泛用于近似译码。模型通常在测试时使用光束，但忽略它在火车时的存在，因此不明确地学习如何使用光束。我们开发了一个统一的元算法学习光束搜索策略使用模仿学习。在我们的设置中，光束是模型的一部分，而不仅仅是近似解码的伪影。我们的元算法捕获现有的学习算法，并建议新的。它还让我们展示了学习光束搜索策略的新颖无悔保证。

[22] **arXiv: 1811.00535 [pdf, ps,其他]**

Cox 回归模型的高维鲁棒推理

圣春,朱青宇,咸阳张,广诚

主题:**统计学理论 (数学。ST)**

我们考虑了基于 Lin 和魏的低维结果的潜在被错误指定 Cox 比例危险模型的高维推断 [1989]。提出了一种基于对数部分似然函数的 sparsified 套索估计算法，并将其收敛到伪真参数向量。有趣的是，真参数的稀疏性可以从上面的限制参数推断出来。此外，上述 (非稀疏) 估计器的每个分量都显示为渐近法线，方差可在模型 misspecifications 下一致估计。在某些情况下，这种渐近分布导致了有效的统计推断程序，通过数值算例说明了其经验性表现。

11 月 18 日 (星期五) 的交叉列表, 2

[23] **arXiv: 1811.00002 (来自 cs 的交叉列表。SD) [pdf,其他]**

WaveGlow: 一种基于流的语音合成生成网络

瑞安 Prenger,拉斐尔山谷,布莱恩卡坦扎罗

评论: 5 页, 1 图, 1 表, 13 方程式

主题:**声音 (cs。SD)**;人工智能 (cs。AI);机器学习 (cs。LG);音频和语音处理 (eess);机器学习 (统计 ML)

在本文中，我们提出了 WaveGlow：一个基于流的网络，能够从 mel 图谱生成高质量的语音。WaveGlow 结合了光晕和 WaveNet 的见解，以提供快速、高效和高质量的音频合成，而无需自动回归。WaveGlow 仅使用单个网络进行实施，仅使用单一成本函数进行培训：最大限度地提高培训数据的可能性，使培训过程简单而稳定。我们的 PyTorch 实现在 NVIDIA V100 GPU 上以超过 500 kHz 的速率产生音频采样。平均的意见得分表明，它提供的音频质量和最好的公开可用的 WaveNet 实现一样好。所有代码将在网上公开提供。

[24] **arXiv: 1811.00003 (来自 cs 的交叉列表。SD) [pdf,其他]**

复杂情感识别的深层网络特征

Bhalaji Nagarajan, V 拉玛纳穆尔蒂 Oruganti

主题:**声音 (cs。SD)**;机器学习 (cs。LG);音频和语音处理 (eess);机器学习 (统计 ML)

本文研究了不同的声学特征、基于音频事件的特征和基于语音的自动翻译的词汇特征在复杂情感识别中的影响，如好奇心。预先训练网络，即 AudioSet 网络、VoxCeleb 网络和深度语音网络，针对不同的语音应用进行了广泛的培训。这些网络的深层信息被视为描述符，并被编码成特征向量。对由 8 个复杂情感组成的 EmoReact 数据集的实验结果表明了其有效性，在文献中对 0.69 的基线产生了最高的 F1 分数 0.85。

[25] **arXiv: 1811.00006 (eess 的交叉列表) [pdf,其他]**

用于设备连续语音识别的低维瓶颈特性

大卫 b. 拉姆齐,凯文麦塔斯,多米尼克 Roblek,马修沙里菲

评论:提交给 ICASSP 2019

主题:**音频和语音处理 (eess)**;机器学习 (cs。LG);声音 (cs。SD);机器学习 (统计 ML)

低功耗数字信号处理器 (dsp) 通常具有非常有限的内存，用于缓存数据。在本文中，我们开发了可在 DSP 上运行的高效瓶颈功能 (BNF) 提取器，并重新训练一个基线大词汇连续语音识别 (语音) 系统，以使用这些 BNFs 的准确性损失极小。小型 BNFs 允许 DSP 芯片在主应用处理器暂停时缓存更多音频功能，从而减少电池的整体使用量。我们所提供的系统能够将标准、定点 DSP 频谱特性的足迹减少 10，而不会在 word 错误率 (wer) 中造成任何损失，且系数仅为 64，而 WER 的相对增加仅为 5.8%。

[26] **arXiv: 1811.00052 (来自 cs 的交叉列表。LG) [pdf,其他]**

图 CNN 的一些新的层架构

梅 Gadiya, Sethi

评论: 5 页数, 1 图，提交给 ICASSP 2019 复杂和超复数领域学习方法特别会议，英国布莱顿, 12-17, 2019

主题:**机器学习 (cs。LG)**;机器学习 (统计 ML)

卷积神经网络 (CNNs) 最近在一个网格 (例如由像素网格组成的图像) 的监督分类方面取得了长足进展，但在几个有趣的数据集中，要素之间的关系可以更好地表示为常规图而不是常规网格。虽然最近的算法，使 CNNs 适应图表已经显示了有希望的结果，他们大多忽视学习显式操作的边缘功能，而专注于顶点要素单独。我们提出了用于神经网络的卷积、池和完全连接层的新配方，可更全面地利用多维图中的信息。使用这些图层可提高基准图形数据集的最先进方法的分类准确性。

[27] **arXiv: 1811.00073 (来自 cs 的交叉列表。LG) [pdf,其他]**

用于建模和学习混杂因子的β伯努利过程的深层生成模型

Prashnna K Gyawali,卡梅伦骑士,信使 Ghimire, b. 米兰霍拉切克,约翰·沙巴进步党,林薇王

主题:**机器学习 (cs。LG)**;机器学习 (统计 ML)

虽然深入的制图表达学习已经越来越能够将任务相关的表述与数据中的其他混淆因素分离开来，但仍然存在两个重大挑战。首先，数据中经常会出现一个未知且可能无限的混淆因子。其次，并非所有这些因素都是显而易见的。在本文中，我们提出了一种深层条件生成模型，它学会了将任务相关表示与未知数量的混杂因素 (可能无限增长) 分开。这是通过与贝叶斯非参数因子模型结婚的深生成

模型的表征力实现的，其中监督确定性编码器学习与任务相关的表示法和带有印度自助餐的概率编码器过程（IBP）学习不可观测混杂因素的未知数量。我们在两个数据集中测试了所呈现的模型：手写数字数据集（MNIST）增加了彩色数字和临床心电图数据集，具有显著的主题间差异，并增强了信号伪影。这些不同的数据集突出显示了模型在数据复杂性的情况下增长的能力，并确定了未观测到的混杂因素的缺失或存在。

[28] **arXiv: 1811.00075 (来自 cs 的交叉列表。LG) [pdf,其他]**

2018 多变量时间序列分类存档

安东尼巴格诺尔,晃映,杰森线,迈克尔弗林,詹姆斯大,亚伦博斯特罗姆,保罗索瑟姆,埃蒙· Keogh

主题:**机器学习（cs。LG)**;机器学习（统计 ML)

在 2002 年, UCR 时序分类存档首次发布十六数据集。它逐渐扩展，直到 2015，当它的大小从 45 数据集增加到 85 数据集。在 2018 年 10 月添加了更多的数据集，使总数增加到 128。新的存档包含一系列问题，包括可变长度序列，但它仍然只包含单变量时间序列分类问题。引入档案的动机之一是鼓励研究人员对新提出的时序分类（TSC）算法进行更严格的评估。它已经奏效：最新的 TSC 研究使用所有 85 数据集来评估算法的进展。对多元时间序列分类的研究，其中多个系列与每个类标签相关联，是在十年前单变量 TSC 研究的一个位置。使用很少的数据集对算法进行评估，而改进声明不是基于统计比较。我们的目标是通过形成移动电话归档的第一个迭代来解决这个问题，并在网站 www.timeseriesclassification.com 上托管。与单变量档案一样，这一提法是东英吉利大学和加州大学河滨分校（UCR）的研究人员之间的合作努力。2018 年份包含 30 个数据集，范围广泛，尺寸和系列长度。对于此存档的第一个迭代，我们将所有数据的格式设置为相等的长度，不包括缺少数据的系列，并提供火车/测试拆分。

[29] **arXiv: 1811.00080 (eess 的交叉列表) [pdf]**

四维扫描透射电镜的流形学习

新李,昂德莱伊 e. 戴克,马克. 奥克斯利,安德鲁 r. 卢皮尼,麦克因尼斯,约翰·杰西,谢尔盖. 加里宁

主题:**图像和视频处理 (eess)**;材料科学（mtrl);数据分析、统计和概率（物理数据 a);机器学习（统计 ML)

四维扫描透射电镜（4 d-杆）的局部原子衍射模式正在兴起，作为探测原子结构和原子电场复杂细节的强大技术。然而，对大量数据的高效处理和解释仍然具有挑战性，特别是对于二维或轻型材料，因为在像素化阵列上记录的衍射信号较弱。在这里，我们采用数据驱动的歧管倾斜方法，实现对 4 维杆数据集的直观可视化和勘探分析，从单层石墨烯中提取对原子解析的偏转模式的实际空间邻近影响,在像素检测器上记录的单个掺杂原子。这些提取的模式与单个 atom 站点和子晶格结构相关，通过多模式视图有效识别单个掺杂异常。我们相信多方面的学习分析将加速物理发现耦合数据丰富的成像机制和材料，如铁电，拓扑旋转和 van der 异质结构。

[30] **arXiv: 1811.00102 (来自 cs 的交叉列表。LG) [pdf,其他]**

数据集中的实际簇数

琥珀斯利瓦斯塔瓦,启东 Baranwal,斯里尼瓦沙 Salapaka

主题:**机器学习 (cs。LG)**;人工智能 (cs。AI);机器学习 (统计 ML)

群集分析中的主要挑战之一是估计数据集中的实际簇数。本文将聚类解决方案的持久性概念量化为一系列分辨率比例，用于表征自然簇并估计数据集中的实际簇数。我们表明，这种持久性的量化与评估底层集群协方差矩阵的最大特征值有关。各种标准和合成数据集的详细实验表明，所建议的基于持久性的指标优于现有方法，如差距统计方法、X 意味着 G 意味着 PG-手段、dip 算法和信息理论方法，准确预测集群的真实数量。有趣的是，我们的方法可以解释在确定性退火算法的相变现象，其中集群中心的数量变化 (分叉) 相对于退火参数。然而，本文提出的方法与聚类算法的选择无关;并可与任何合适的聚类算法结合使用。

[31] **arXiv: 1811.00103 (来自 cs 的交叉列表。LG) [pdf,其他]**

公平 PCA 的价格：一个额外的维度

萨米拉 Samadi, Uthaipon Tantipongpipat,杰米摩根斯坦,莫希特辛格,桑托斯 Vempala

主题:**机器学习 (cs。LG)**;机器学习 (统计 ML)

研究 PCA 的标准降维技术是否无意中产生了两个不同种群的不同保真度的数据表示。我们在几个真实世界的数据集上显示, PCA 在人口 A 上的重建误差高于 B (例如，女性对男性或低相对于高学历的个体)。即使数据集与 a 和 B 的

样本数量相似，也可能发生这种情况。这就推动了我们对维数还原技术的研究，它保持了 A 和 B 的相似保真度。我们定义了公平 PCA 的概念，并给出了一项多项式时间算法，用于查找数据的低维表示，这是近最优的这一措施。最后，我们在真实世界的数据集上显示，我们的算法可用于高效生成数据的公平低维表示。

[32] **arXiv: 1811.00112 (来自 cs 的交叉列表。CV) [pdf,其他]**

生成照片逼真的训练数据，提高人脸识别精度

丹尼尔 Sáez 特里格罗斯旅馆,李孟,玛格丽特 Hartnett

主题:**计算机视觉和模式识别 (cs。CV)**;机器学习 (cs。LG);机器学习 (统计 ML)

本文研究了利用合成数据扩充人脸数据集的可行性。特别是，我们提出了一种新的生成对抗网络 (GAN)，它可以将与身份相关的属性从非身份相关属性中解脱出来。这是通过训练一个嵌入网络，将离散标识标签映射到一个简单的先验分布之后的身份潜伏空间，并训练 GAN 对该分布的样本进行调理。我们建议的 GAN 允许我们通过在训练集中生成主题的合成图像和不在训练集中的新科目的合成图像来增加面部数据集。通过使用 GAN 训练的最新进展，我们发现我们的模型生成的合成图像是照片逼真的，与增强数据集的训练确实可以提高人脸识别模型的准确性，与实际训练的模型相比图像。

[33] **arXiv: 1811.00121 (来自 cs 的交叉列表。CR) [pdf,其他]**

基于混合模型的防御方法对朴素贝叶斯垃圾邮件过滤器的数据中毒攻击

大卫 j. 米勒,信义湖,镇翔,乔治 Kesidis

主题:**密码学和安全性 (cs。CR)**;机器学习 (cs。LG);机器学习 (统计 ML)

朴素的贝叶斯垃圾邮件过滤器非常容易受到数据中毒攻击。在这里，已知的垃圾邮件来源/黑名单 IPs 利用事实，他们收到的电子邮件将被视为 (地面真相) 标记垃圾邮件示例，并用于分类器培训 (或重新培训)。攻击源因此生成电子邮件，将扭曲垃圾邮件模型，可能导致分类器准确性的极大退化。这种攻击之所以成功，主要是因为幼稚贝叶斯 (NB) 模型的表示能力较差，只有一个 (组件) 密度代表垃圾邮件 (加上可能的攻击)。我们提出一种基于 NB 模型混合使用的防御措

施。我们证明，已学会的混合物几乎完全隔离攻击在第二个 NB 组件，与原始垃圾邮件组件基本上不变的攻击。我们的方法解决了在新数据的情况下重新训练分类器的方案，以及在原始垃圾邮件训练集中嵌入攻击的更具挑战性的场景。即使对于较弱的攻击强度，基于 BIC 的模型顺序选择选择一个双组件解决方案，它调用基于混合的防御。TREC 2005 垃圾邮件语料库提供了有希望的结果。

[34] **arXiv: 1811.00123 (从物理学的交叉列表. 化学-ph) [pdf,其他]**

压缩原子的物理特性以改善预测化学

约翰.柏氵 乔,凯文岛,昆瑶,约翰。

评论: 6 页, 5 数字

主题:**化学物理学 (物理学)**;机器学习 (统计 ML)

许多尚未解决的问题的答案在于分子和材料的顽固化学空间。机器学习技术正迅速发展成为一种有效压缩和探索化学空间的方法。机器学习技术的一个最重要的方面是通过特征向量来表示，它应该包含做出准确预测所必需的最重要的描述符，其中至少包括分子中的原子物种或材料。在这项工作中，我们引入了原子物种的物理特性的压缩表示，我们称之为元素模式。元素模式通过捕捉周期表的许多细微差别和原子物种的相似性提供了极好的表征。我们将元素模式应用于机器学习算法的几个不同任务，并表明它们使我们能够对这些任务进行改进，甚至超越了更高精度的预测。

[35] **arXiv: 1811.00128 (来自 cs 的交叉列表。LG) [pdf,其他]**

面向多步模型强化学习的一种简单方法

Kavosh 阿萨迪,埃文迎合,迪彭德拉 Misra,迈克尔 l. 利特曼

主题:**机器学习 (cs。LG)**;人工智能 (cs。AI);机器学习 (统计 ML)

当环境交互成本高昂时，基于模型的强化学习通过提前规划和避免代价高昂的错误来提供解决方案。基于模型的代理通常学习单步转换模型。本文提出了一种多步模型，用于预测具有可变长度的动作序列的结果。我们表明该模型易于学习，模型可以进行策略条件预测。我们报告的初步结果，显示了一个明显的优势，多步骤模型相比，它的一步对应。

[36] **arXiv: 1811.00145 (来自 cs 的交叉列表。LG) [pdf, ps,其他]**

通过稀有事件仿真实现可扩展的端到端自主车辆测试

马修 O ' 凯利, 安 Sinha, Hongseok Namkoong,约翰迪杜齐,罗斯 Tedrake

评论:NIPS 2018

主题:**机器学习 (cs。LG)**;机器人 (cs);机器学习 (统计 ML)

虽然自主车辆 (AV) 技术的最新发展突出了重大进展，但我们缺乏严格和可扩展测试的工具。现实世界的测试,事实上评估环境，使公众处于危险之中，由于事故的罕见性质，将需要数以亿计的里程来统计验证绩效索赔。我们实施了一个模拟框架，可以测试整个现代自主驾驶系统，特别是使用深度学习感知和控制算法的系统。利用自适应重要性抽样方法加速稀有事件概率评估，估计了基于标准交通行为的基础分布下的事故概率。我们在高速公路上展示了我们的框架，加速了系统评估 2-20 时间过天真的蒙特卡洛抽样方法和 10-300 P 次(其中P是处理器数量在实际测试。

[37] **arXiv: 1811.00148 (来自 cs 的交叉列表。LG) [pdf,其他]**

有限观测的二次张量的恢复保证

红阳,领夏朗,摩西恰里卡尔,盈余梁

主题:**机器学习 (cs。LG)**;数据结构和算法 (cs。DS);机器学习 (统计 ML)

我们考虑了预测张量缺失项的张量完成问题。常用的 CP 模型有三种产品形式，但是一个二次模型的备用系列，它们是成对产品而不是三重产品的总和，从推荐系统等应用中应运而生。非凸方法是学习二次模型的选择方法，本文研究了它们的样本复杂性和误差保证。我们的主要结果是，随着样本数量在维度中只是线性的，均方根误差目标的所有局部极小值均为全局极小，并准确恢复原始张量。这些技术导致简单的证明，表明凸松弛可以恢复二次张量提供的线性样本数。我们通过对合成和真实世界数据的实验来证实我们的理论结果，表明二次模型在可用观测量有限的情况下比 CP 模型具有更好的性能。

[38] **arXiv: 1811.00152 (来自 cs 的交叉列表。LG) [pdf,其他]**

混合密度生成对抗网络

哈米德 Eghbal-扎德,沃纳 Zellinger,威德默

评论: 13 页, 3 数字

主题:**机器学习 (cs。LG)**;机器学习 (统计 ML)

生成的敌对网络有惊人的能力产生尖锐和逼真的图像，虽然他们已知遭受所谓的模式崩溃问题。本文提出一种新的 gan 变体称为混合密度 gan，同时能够生成高质量的图像，通过鼓励鉴别器在其嵌入空间中形成簇来克服这一问题，进而导致发生器利用这些信息并发现数据中的不同模式。这是通过将高斯密度函数定位在单纯形的角部，使用所产生的高斯混合物作为一个似然函数超过鉴别器嵌入，并制定基于这些的 GAN 训练的客观功能可能性.我们表明，只有在生成的和实际的分布完全匹配时，才能达到最佳的训练目标。我们进一步支持我们的理论结果，并对一个合成和几个真实图像数据集 (CIFAR-10、CelebA、MNIST 和 FashionMNIST) 进行实证评估。我们以经验证明 (1) 在混合密度 GAN 中生成的图像的质量和它们与真实图像的强烈相似性，由 Fr echet 起始距离 (FID) 测量，与最先进的方法相比非常有利, (2)能够避免模式崩溃并发现所有数据模式。

[39] **arXiv: 1811.00155 (来自 cs 的交叉列表。LG) [pdf,其他]**

内存约束核逼近的低精度随机傅里叶特征

剑张,艾夫纳,三刀,克里斯托弗雷岛

主题:**机器学习 (cs。LG)**;人工智能 (cs。AI);机器学习 (统计 ML)

我们研究如何训练在内存预算下很好地泛化的内核逼近方法。在最近的理论研究基础上，我们定义了一个核逼近误差测度，它比常规指标更能预测核逼近方法的经验泛化性能。此定义的一个重要结果是，内核逼近矩阵必须是高秩才能达到接近逼近。由于存储高秩逼近是内存密集型的，因此我们建议使用随机傅立叶特征的低精度量化 (LP RFFs) 在内存预算下构建高秩逼近。从理论上讲，量化在重要设置中对泛化性能的影响微乎其微。从经验上说，我们在四个基准数据集上演示了 RFFs 可以匹配全精度 RFFs 和 Nystr \"{o} m 方法的性能，分别为 3 x 10 x 和 50 x 460 x 的内存。

[40] **arXiv: 1811.00159 (来自 cs 的交叉列表。IR) [pdf,其他]**

分级分解的聚类单调变换

Gaurush 喜拉安达尼,拉加夫

Somani, OluwasanmiKoyejo, Sreangsu Acharyya

评论:前两位作者同样对该文件作出了贡献。WSDM 2019 中出现的纸张

主题:**信息检索 (cs。IR)**;机器学习 (cs。LG);机器学习 (统计 ML)

利用用户项评级矩阵的低级结构是许多推荐引擎的关键。但是，现有的推荐引擎会强制评分员使用异构行为配置文件将其内部评级比例映射到通用评级比例(例如 1-5)。这种非线性变换的评级等级打破了评级矩阵的低级结构，因此导致了较差的拟合和相应的建议。本文提出了分级分解 (CMTRF) 的聚类单调变换，这是一种新的方法，可以在未知的种群段上对未知单调变换进行回归。从根本上讲，对于推荐系统，该技术搜索分级秤的单调变换，从而使其更适合。这与基础矩阵分解回归模型相结合，使用户的评分与共享的低维结构一起利用。可以为每个用户、一组用户或为所有用户生成分级比例转换，这是本文提出的三种简单高效算法的基础，所有这两种方法都可在分级秤转换之间交替使用。和矩阵分解回归。尽管非凸性，CMTRF 在理论上表明在温和条件下恢复一个独特的解决方案。两个合成和七实际数据集的实验结果表明，CMTRF 优于其他最先进的基线。

[41] **arXiv: 1811.00170 (来自 cs 的交叉列表。LG) [pdf]**

PerceptionNet: 一种用于晚期传感器融合的深度卷积神经网络

卢森堡 Kasnesis,土耳其 z Patrikakis,雅科沃斯 s Venieris

评论:本文已被接受出版的智能系统会议 (IntelliSys) 2018

主题:**机器学习 (cs。LG)**;机器学习 (统计 ML)

基于运动传感器的人类活动识别 (HAR) 在过去几年中引起了很多关注，因为感知人的状态使上下文感知应用程序能够根据用户的需求调整其服务。然而，运动传感器融合和特征提取尚未达到其全部潜能，仍然是一个开放性问题。本文介绍了一种深度卷积神经网络 (CNN)，它将 2D 后期卷积应用到多模式时序传感器数据中，从而自动提取 PerceptionNet 的高效特性。我们对两个公共可用的 HAR 数据集进行评估，以证明所提出的模型能够有效地融合多模式传感器，提

高 HAR 的性能。特别是, PerceptionNet 超越了最先进的 HAR 方法的性能基于: (i) 从人类提取的特征, (ii) 深度 CNNs 开发早期融合方法，和 (iii) 长短期记忆 (LSTM)，平均准确度超过 3%.

[42] **arXiv: 1811.00178 (来自 cs 的交叉列表。LG) [pdf, ps,其他]**

使用多次权重更新的在线学习

Charanjeet,展开

主题:**机器学习 (cs。LG)**;机器学习 (统计 ML)

在线学习使部分数据到达时决定序列，而数据的下一次移动是未知的。本文提出了一种新的思想，即多次权重更新，以迭代的方式为同一实例更新权重。本文采用常用的文献分析方法，利用既定的工具进行实验研究。结果表明，不同数据集和算法的错误率降低为零或接近零。架空运行成本不太高，实现了接近零的误率，进一步加强了所提出的技术。所提出的技术有助于应对现实生活中的挑战。

[43] **arXiv: 1811.00181 (来自 cs 的交叉列表。LG) [pdf,其他]**

提高图形注意模型的鲁棒性

乌代尚卡尔 Shanthamallu,廖四辉 j. 德博契亚格拉杰,列斯 Spanias

主题:**机器学习 (cs。LG)**;机器学习 (统计 ML)

可以利用数据固有结构的机器学习模型得到了突出的重视。特别是，由于它在多个领域的广泛应用，在图形结构化数据的深度学习解决方案中出现了激增。图形注意网络是图中的一类特征学习模型的新补充，它利用注意机制有效地学习了半监督学习问题的连续向量表示。本文对该模型进行了详细的分析，并对其行为提出了有趣的见解。具体来说，我们表明模型很容易受到对手 (流氓节点) 的攻击，因此提出了一种新的正则化策略来提高模型的鲁棒性。使用基准数据集，我们使用所建议的 "双文" 稳健变体，演示半监督学习的性能改进。

[44] **arXiv: 1811.00200 (来自 cs 的交叉列表。LG) [pdf,其他]**

统计套利的在线学习算法

克里斯托弗馆长

主题:**机器学习 (cs。LG)**;机器学习 (统计 ML)

统计套利是使用均值回归模型的一类金融交易策略。相应的技术依赖于一些假设，它们可能无法容纳一般的非平稳随机过程。本文提出了一种基于在线学习的统计套利的替代技术，它不需要这样的假设，并且受益于强大的学习保证。

[45] **arXiv: 1811.00208 (来自 cs 的交叉列表。LG) [pdf,其他]**

多标签鲁棒分解自动编码器及其在药物-药物相互作用预测中的应用

许楚,杨林,净月高,江涛王,雅夏王,乐业王

主题:**机器学习 (cs。LG)**;人工智能 (cs。AI);机器学习 (统计 ML)

药物-药物相互作用 (DDIs) 是可预防住院和死亡的主要原因。预测 DDIs 的发生有助于药物安全专业人员分配调查资源，并及时采取适当的监管措施。传统的 DDI 预测方法基于药物的相似性预测 DDIs。最近，研究人员发现，通过更好地模拟双线性形式的药物对之间的相互作用，可以改善预测性能。然而，利用双线性形式的浅模型在捕获药物对之间复杂的非线性相互作用时受到限制。为此，我们提出了用于 DDI 预测的多标签鲁棒分解自动编码器 (缩写为 MuLFA)，它学习了药物对之间相互作用的表征，并具有表征复杂非线性相互作用的能力。正是。此外，还设计了一种新的 CuXCov 损耗，可以有效地学习 MuLFA 的参数。此外，该解码器能够为特定 DDIs 生成药物对的高风险化学结构，帮助药剂师更好地了解药物化学与 DDI 之间的关系。真实世界数据集的实验结果表明，MuLFA 始终优于最先进的方法;特别是，它提高了 21:3% 的预测性能，与前 50 个常见 DDIs 的最佳基线相比. 我们还说明了各种案例研究，以证明 MuLFA 在 DDI 诊断中产生的化学结构的有效性。

[46] **arXiv: 1811.00210 (来自 cs 的交叉列表。AI) [pdf,其他]**

利用图神经网络进行自适应规划调度

腾飞马,帕特里克费伯,思玉火,杰陈,迈克尔。

主题:**人工智能 (cs。AI)**;机器学习 (cs。LG);机器学习 (统计 ML)

自动化规划是人工智能的基础领域之一。由于单个计划器不可能在所有任务和域中发挥良好的作用，因此基于组合的技术在最近变得越来越流行。特别是，深度学习作为一种有前景的在线规划师选择方法应运而生。针对规划任务结构图表示的最新进展，提出了一种图形神经网络 (器) 方法来选择候选计划者。拟订是有利的，在一个简单的替代，卷积神经网络，因为它们是不变的节点排列和它们合

并节点标签，以更好的推断。
另外，针对成本优化规划，提出了一种两阶段自适应调度方法，进一步提高了给定任务及时解决的可能性。计划程序可能会在中场休息时切换到不同的计划者，这取决于第一个观察到的性能。实验结果验证了该方法在强基线、深度学习和非深度学习基础上的有效性。

[47] **arXiv: 1811.00217 (来自 cs 的交叉列表。LG) [pdf,其他]**

元 DES。Oracle: 集合选择的元学习和特征选择

拉斐尔·克鲁兹,罗伯特萨柏林,乔治华盛顿埃德蒙多·卡瓦尔康蒂

评论:关于信息融合的论文发表

期刊编号:卷 2017 年 11 月 38 日，页 84-103

主题:**机器学习 (cs。LG)**;机器学习 (统计 ML)

动态集合选择 (DES) 中的关键问题是确定分类器能力计算的合适判据。有几个标准可用于衡量基本分类器的能力水平，如本地准确性估计和排名。但是，仅使用一个标准可能会导致对分类器能力的估计较差。为了解决这一问题，我们提出了一种新的动态集合选择框架，它采用元学习，称为元 DES。元 DES 框架的一个重要方面是，可以将多个标准嵌入到编码为不同元特征集的系统中。但是，某些 DES 标准不适合于每个分类问题。例如，当类之间存在高度重叠时，局部精度估计可能会产生较差的结果。此外，如果对相应数据优化了元分类器的性能，则可以获得更高的分类精度。本文提出了一种基于 Oracle 形式化定义的元 des 框架的新版本，称为元 des。甲骨文。Oracle 是一种抽象的方法，表示理想的分类器选择方案。为了提高元分类器的性能，提出了一种采用拟合谨慎二元粒子群优化 (基本粒子) 的元特征选择方案。元分类器和 Oracle 提供的输出之间的差异最小化。因此，元分类器有望获得类似于 Oracle 的结果。使用 30 分类问题进行的实验表明，基于 Oracle 定义的优化过程与元 DES 框架的早期版本相比，可显著提高分类精度，并其他最先进的 DES 技术。

[48] **arXiv: 1811.00223 (来自 cs 的交叉列表。SD) [pdf,其他]**

用于柔性音色控制的神经音乐合成

金正日,雷切尔比特纳,阿帕娜,胡安·贝洛

主题:**声音 (cs。SD)**;音频和语音处理 (eess);机器学习 (统计 ML)

像 WaveNet 这样的 raw 音频波形合成模型最近取得的成功激发了一种新的音乐合成方法，其中整个过程---从乐谱和仪器信息创建音频样本---使用生成神经网络进行建模。本文介绍了一种具有灵活音色控制的神经音乐合成模型，它由一个反复的神经网络所组成，其条件是在一个博学的仪器嵌入后 WaveNet 声码器。所学的嵌入空间成功地捕获了大型数据集中音色中的各种变体，并通过在嵌入空间中的仪器之间插值来实现音色控制和变形。对合成质量进行了数值计算和感知，并给出了交互式 web 演示。

[49] **arXiv: 1811.00246 (来自 cs 的交叉列表。LG) [pdf,其他]**

萨恩：顺序关注的关系推理

Jinwon, Sungwon 吕, Sungzoon 町

主题:**机器学习 (cs。LG)**;机器学习 (统计 ML)

提出了一种名为萨恩（顺序注意关系网络）的注意模块增强关系网络，通过提取参考对象并在对象之间进行有效的配对来实现关系推理。萨恩极大地减少了关系网络的计算和内存需求，从而计算所有对象对。与其他模型相比，它还显示了 CLEVR 数据集的高精度，尤其是在关系问题上。

[50] **arXiv: 1811.00247 (来自 cs 的交叉列表。LG) [pdf,其他]**

公平分类器的神经网络框架

P 邓, Sujit 古加尔

主题:**机器学习 (cs。LG)**;机器学习 (统计 ML)

机器学习模型广泛应用于决策，尤其是预测任务。这些模型可能对特定种族、性别或年龄的特定敏感群体有偏见或不公平。研究人员已经将努力定性为一种特定的公平定义，并将其强制实施到模型中。在这项工作中，我们主要关注以下三定义、不同的影响、人口均等和均衡赔率。研究人员已经表明，除非分类器是完美的，否则校准的量词不能满足均衡赔率。因此，主要的挑战是确保一定程度的公平，同时保证尽可能多的准确性。
公平约束是复杂的，不需要凸。将它们集成到机器学习算法中是一项重大挑战。因此，许多研究人员试图想出一个替代的损失，这是凸的，以建立公平分类器。此外，某些文件尝试通过预处理数据来构建公平表示，而不管使用的是哪种分类器。这种方法不仅需要很多不切实际的假设，还需要人工设计的分析解决方案来构建机器学习模型。相反，我们提出了一个自动解决方案，推广任何公平约束。我们使用的是分批训练的神经网络，直接强制将公平约束作为损失函数，而无需进一步修改。我们还试验了其他复杂的性能指标，如 H 均值损耗、Q 平均损耗、f-测量;不需要任何代理损失功能。我们的实验证明，网络的性能与艺术状态相似。

这样，就可以根据所需的分类器的公平性约束和性能度量来插入适当的损耗函数，并训练神经网络来实现这一点。

[51] **arXiv: 1811.00260 (来自 cs 的交叉列表。LG) [pdf, ps,其他]**

地平线: Facebook 开源应用强化学习平台

杰森戈西,爱德华多,曾溢滔梁, Kittipat Virochsiri,豫辰他,扎克 Kaden,维韦克纳拉亚南,惠叶

评论: 6 页

主题:**机器学习 (cs。LG)**;人工智能 (cs。AI);机器学习 (统计 ML)

本文介绍了 Facebook 开源应用强化学习 (RL) 平台的前景。Horizon 是一个端到端平台，旨在解决工业应用的 RL 问题，其中数据集很大 (数以亿计的观测值)，反馈环路很慢 (与模拟器)，并且实验必须小心完成，因为它们不在模拟器中运行.与其他 RL 平台 (通常设计用于快速原型和实验) 不同的是, Horizon 设计的是以生产用例为首要考虑的。该平台包含用于培训流行的深度 RL 算法的工作流，包括数据预处理、特征转换、分布式培训、反事实策略评估和优化服务。我们还展示了与地平线一起训练的模型在 Facebook 上显著优于和取代监督学习系统的真实示例。

[52] **arXiv: 1811.00264 (来自 cs 的交叉列表。LG) [pdf,其他]**

多个内核 K-通过选择代表性内核来进行聚类

吴亚强姚

评论: 8 页, 7 数字

主题:**机器学习 (cs。LG)**;机器学习 (统计 ML)

要在原始要素空间中对非线性可分离的数据进行聚类,K-意味着群集扩展到内核版本。但是，内核的性能 K-意味着聚类在很大程度上取决于内核函数的选择。为了缓解这一问题，已将多个内核学习引入到 K-意味着聚类，以获得最佳的内核组合的聚类。尽管多个内核的成功 K-意味着聚类在各种情况下，现有的工作很少更新基于内核多样性的组合系数，从而导致所选内核包含高冗余，并会降

低聚类性能和效率。本文提出了一种简单而有效的策略，从预先指定的内核中选择一个不同的子集作为代表内核，然后将子集选择过程纳入多个 K-意味着群集。代表内核可以表示为重要的组合权重。由于得到的目标函数的非凸性，我们开发了一种交替最小化方法来优化选定内核和集群成员的组合系数。我们对若干基准和真实世界数据集的建议方法进行评估。实验结果表明，与先进的方法相比，我们的方法具有竞争力。

[53] **arXiv: 1811.00321 (来自 cs 的交叉列表。LG) [pdf, ps,其他]**

液体时恒递归神经网络作为通用逼近器

拉敏 m 哈萨尼,马蒂亚斯加斯莱希纳,亚历山大阿米尼,丹妮拉, Radu Grosu

评论:本文简要介绍了液体时常数 (LTC) 递归神经网络的通用逼近能力，为其动力学提供了理论界

主题:**机器学习 (cs。LG)**;神经和进化计算 (cs);机器学习 (统计 ML)

本文介绍了液体时间常数 (LTC) 递归神经网络 (RNN) 的概念，即连续时间 RNNs 的子类，其非线性突触传输模型实现了不同的神经元时间常数。这一特性受到小物种神经系统的传播原理的启发。它使模型能够用少量的计算单元来近似连续映射。我们表明，任何有限的弹道 n 维连续动力系统可以近似于隐藏单元的内部状态和 nLTC 网络的输出单位。在这里，我们也理论上找到他们的神经元状态和不同的时间常数的界限。

[54] **arXiv: 1811.00338 (来自 cs 的交叉列表。LG) [pdf,其他]**

利用智能手机在野外进行深度学习的步态识别

秦邹、陵王、羿赵、前王、朝沈、清泉李

主题:**机器学习 (cs。LG)**;信号处理 (eess。SP);机器学习 (统计 ML)

与其他生物识别相比，步态具有不显眼、不易隐蔽的优点。惯性传感器 (如加速度计和陀螺仪) 通常用于捕获步态动力学。目前，这些惯性传感器已普遍集成在

智能手机中，一般人广泛使用，这使得收集步态数据非常方便和廉价。本文研究了在野外使用智能手机进行步态识别的方法。与传统的方法通常要求人走在指定的道路和/或以正常步行速度，建议的方法收集惯性步态数据在自愿的情况下不知道什么时候，在哪里，和如何用户走。为了获得较高的人员识别和认证能力，本文提出了从步行数据中学习和建模步态生物特征的深度学习技术。针对鲁棒步态特征表示，提出了一种基于卷积神经网络和递归神经网络连续提取空间域和时域特征的混合深神经网络。在实验中，智能手机收集的两个数据集总共有 118 主题用于评估。实验结果表明，该方法在人的身份识别和认证方面分别达到了 93.5% 和 93.7% 的精度。

[55] **arXiv: 1811.00353 (从数学的交叉列表。PR) [pdf, ps,其他]**

Banach 空间中的汉森-赖特不等式

Radosław 亚当恰克,律师 Latała,律师 Meller

主题:**概率 (数学。公关)**;功能分析 (数学. FA);统计学理论 (数学。圣

我们讨论了高斯随机变量中二次形式的矩和尾的双边边界，其值在 Banach 空间中。我们陈述一个自然的猜想，并表明它持有额外的对数系数。此外，在某类巴拿巴空间 (包括 $我 R$-空格) 这些对数因子可能被消除。作为推论，我们推导出 subgaussian 随机变量中二次形式的尾部和力矩的上界，从而扩展了汉森-赖特不等式。

[56] **arXiv: 1811.00401 (来自 cs 的交叉列表。LG) [pdf,其他]**

过度不变性导致敌对漏洞

杜兰戈-亨利雅各布森,贝赫曼,理查德泽梅尔,马加什陆慈

主题:**机器学习 (cs。LG)**;人工智能 (cs。AI);计算机视觉和模式识别 (cs。CV);机器学习 (统计 ML)

尽管其性能令人印象深刻，但深度神经网络在分布外输入方面表现出惊人的失败。对抗实例研究的一个核心思想是揭示这种分布变化下的神经网络误差。我们将这些错误分解为两个互补源：灵敏度和不变性。我们显示深度网络不仅对其输入的任务无关的变化太敏感，而且是众所周知的，从小量对抗的例子，但也太不固定的范围广泛的任务相关的变化，从而使巨大的区域在输入空间易受敌对攻击。在确定了这种过度不变性后，我们建议使用双射深度网络来实现对所有变体

的访问。我们引入 metameric 采样作为对这些网络的分析攻击，无需进行优化，并表明它揭示了误分类输入的大子空间。然后，我们将这些网络应用到 MNIST 和 ImageNet，并表明一个人可以在不改变隐藏激活的情况下操纵几乎任何图像的类特定内容。此外，我们通过信息理论分析扩展了标准的交叉熵损失，从而加强了对这种操作的模型，为克服基于不变性的脆弱性提供了明确的第一种方法。最后，我们通过实证的方法来说明其控制不受欢迎的类特定不变性的能力，并展示了克服敌对事例的一个主要原因的承诺。

[57] **arXiv: 1811.00416 (来自 cs 的交叉列表。LG) [pdf,其他]**

MoDISco v0.4 4.2-α：技术说明

文狄 Shrikumar,凯瑟琳·天,安娜谢尔比纳, Žiga Avsec,苏哈塔加戈班纳吉, Mahfuza 作者,苏拉克奈, 春妮 Kundaje

评论:此 https URL 中可用的实现

主题:**机器学习 (cs。LG)**;基因组学 (q 生物。GN);机器学习 (统计 ML)

MoDISco (转录因子从重要性分数中发现) 是一种从基因组序列数据计算的 basepair 级重要性分数中识别图案的算法。本文介绍了 MoDISco 版本 0.4.4.2-alpha (在 https://github.com/kundajelab/tfmodisco/tree/v0.4.2.2-alpha 中可用) 的方法。

[58] **arXiv: 1811.00424 (来自 cs 的交叉列表。LG) [pdf, ps,其他]**

基于分布式 ReliefF 的火花特征选择

劳尔-何塞-帕尔马-门多萨,丹尼尔罗德里格斯,路易斯-马科斯

主题:**机器学习 (cs。LG)**;分布式、并行和集群计算 (cs。DC);机器学习 (统计 ML)

功能选择 (FS) 是机器学习和数据挖掘领域中的一个重要研究领域，删除不相关的冗余功能通常有助于减少处理数据集所需的工作量，同时保持甚至改进处理算法的精度。但是，为在单机上执行而设计的传统算法缺乏可扩展性，无法应对当前大数据时代所提供的日益增长的数据量。ReliefF 是在许多 FS 应用程序

中成功实现的最重要的算法之一。在本文中，我们提出了一个完全重新设计的分布式版本的流行 ReliefF 算法基于新的火花集群计算模型，我们已经称为 DiReliefF。与 Hadoop 的 MapReduce 模型实现相比，Spark 的处理速度快得多，因此其受欢迎程度越来越高。我们的建议的有效性在四个公开可用的数据集上进行了测试，它们都有大量实例，其中两个也有大量的功能。这些数据集的子集还用于将结果与算法的非分布式实现进行比较。结果表明，非分布式实现在没有专用硬件的情况下无法处理如此大的数据量，而我们的设计可以以可扩展的方式处理它们，并具有更好的处理时间和内存使用率。

[59] **arXiv: 1811.00429 (来自 cs 的交叉列表。LG) [pdf,其他]**

马尔可夫决策过程中的时间正则化

皮埃尔 Thodoroff,奥黛丽杜兰德,乔艾皮诺,多依娜 Precup

评论:作为会议文件发表在 NIPS 2018

主题:**机器学习 (cs。LG)**;机器学习 (统计 ML)

强化学习的几个应用由于高方差而受到不稳定性的影响。这在高维域中尤其普遍。正则化是机器学习中一种常用的技术，以减少方差，以引入一些偏置的代价。大多数现有的正则化技术侧重于空间 (感性) 正则化。然而在强化学习中，由于行李员方程的性质，还有机会利用基于轨迹的价值估计平滑度的时态正则化。本文探讨了一类时态正则化方法。通过马尔可夫链概念，我们正式地描述了这种技术引起的偏差。通过一系列简单的离散和连续 mdp 以，说明了时态正则化的各种特征，表明该技术在高维雅达利游戏中也能提供改进。

[60] **arXiv: 1811.00458 (来自 cs 的交叉列表。LG) [pdf,其他]**

通过端到端移位学习减少偏倚：在公民科学中的应用

陈,卡拉 p. 戈麦斯

主题:**机器学习 (cs。LG)**;人工智能 (cs。AI);机器学习 (统计 ML)

公民科学项目成功地收集各种应用程序的丰富数据集。然而，公民科学家收集的数据往往是有偏见的，更符合公民的喜好，而不是科学目标。提出了一种从科学目标向有偏数据转变的端到端学习方案，并通过对训练数据进行重加权来补偿移位。应用于公民科学项目 \ textit {eBird} 的鸟类观测数据，我们演示了 SCN

如何量化数据分布变化，以及优于没有解决数据偏差的监督学习模型。与协变量转移背景下的其他竞争模型相比，我们进一步展示了 SCN 在处理海量高维数据的有效性和能力方面的优势。

[61] **arXiv: 1811.00464 (来自 cs 的交叉列表。LG) [pdf, ps,其他]**

一种挖掘异构非随机电子健康记录数据的潜在主题模型

悦里,马诺利斯 Kellis

主题:**机器学习 (cs。LG)**;机器学习 (统计 ML)

电子健康记录 (EHR) 是丰富的异构患者健康信息收集，其广泛的采用为系统的健康数据挖掘提供了巨大的机会。然而，异构的 EHR 数据类型和偏置确定施加计算挑战。在这里，我们提出了一种集成协作过滤和潜伏主题模型的无监督生成模型 mixEHR，它利用潜伏疾病-主题分布共同模拟数据观测偏差和实际数据的离散分布。我们将 mixEHR 应用于模拟数据集的 1280 万表型观测，并使用它揭示潜在疾病的主题，解释 EHR 结果，将缺失数据归咎于，并预测重症监护病房的死亡率。通过模拟和真实数据，我们表明 mixEHR 优于以往的方法，并揭示有意义的多疾病洞察。

[62] **arXiv: 1811.00513 (来自 cs 的交叉列表。CR) [pdf,其他]**

自然审计师：如何判断某人是否用你的话来训练他们的模型

张从正歌,维塔利 Shmatikov

主题:**密码学和安全性 (cs。CR)**;计算和语言 (cs。CL);机器学习 (cs。LG);机器学习 (统计 ML)

为了帮助强制实施诸如 GDPR 等数据保护法规和检测未经授权的个人数据使用，我们提出了一种新的公众 {模型审核} 技术，使用户能够检查其数据是否用于培训机器学习模型。我们专注于审计生成自然语言文本的深度学习模型，包括单词预测和对话生成。这些模型是许多流行的在线服务的核心。此外，他们经常接受非常敏感的个人数据的培训，例如用户的信息、搜索、聊天和评论。
我们设计和评估一个有效的黑箱审核方法，它可以检测到一个模型的查询，如果使用特定用户的文本来对其进行培训 (在数以千计的其他用户中)。与之前对 ML 模型的成员推断的工作相反，我们不假设模型产生数值置信值。我们通过实证的方式证明，我们可以成功地审核模型，它们的通用性很好，而不是一来培训

数据。我们还分析了文本生成模型是如何记忆单词序列的，并解释了为什么这种记忆会使其易于审计。

[63] **arXiv: 1811.00521 (来自 cs 的交叉列表。LG) [pdf,其他]**

最小化闭合 k 总损耗提高分类

布莱恩,詹姆斯. 邹

主题:**机器学习 (cs。LG)**;人工智能 (cs。AI);机器学习 (统计 ML)

在分类中，个体损失的实际方法是平均损失。当实际的利息指标为 0-1 损失时，通常会将某些表现良好的 (例如凸) 代理项的平均代理损失降至最低。最近，其他一些合计损失，如最大损失和平均顶部 K 建议将损失作为替代目标，以解决平均损失的缺点。但是，我们确定常见的分类设置，如数据不平衡，有太多的简单或含糊的例子等，当平均，最大和平均顶部 K 即使是在无限大的训练集上，所有的决策边界都会受到不理想的影响。为了解决这个问题，我们提出了一个新的分类目标，称为 "近 K 聚合损失，在这种情况下，我们自适应地尽量减少接近决策边界点的损失。在优化近距离时，我们提供了 0-1 精度的理论保证。K 总损失。我们还在 PMLB 和 OpenML 基准数据集上进行系统实验。关闭 K 在 0-1 测试精度方面取得显著成就，改善$\geq 2\%$ 和 $p<0.05$，在超过 25% 的数据集与平均、最大和平均顶部相比,K.相比之下，以前的总损失比近 K 不到 2% 的数据集。

[64] **arXiv: 1811.00525 (来自 cs 的交叉列表。LG) [pdf,其他]**

论对抗性例子的几何学

马克扈利,迪伦哈德菲尔德-Menell

主题:**机器学习 (cs。LG)**;机器学习 (统计 ML)

对抗性的例子是机器学习模型普遍存在的现象，在这种情况下看似难以察觉的扰动对输入导致误分类的统计准确模型。我们提出了一个几何框架，从多方面的重建文献中汲取工具，来分析对抗性实例的高维几何。特别是，我们强调了余维数的重要性：对于嵌入在高维空间中的低维数据流形，有许多方向可以用来构

造对抗性实例。对抗性的例子是学习决策边界的自然结果，它可以很好地分类低维数据流形，但不正确地对歧管附近的点进行分类。使用我们的几何框架，我们证明 (1) 不同规范下的稳健性之间的权衡，(2) 在数据周围的球的对抗性训练是样本效率低下，和 (3) 足够的取样条件下，最近邻分类器和基于球的对抗训练是健壮的。

[65] **arXiv: 1811.00539 (来自 cs 的交叉列表。LG) [pdf,其他]**

非线性输出变换的深层结构预测

科林格雷勃, Ofer 饭,亚历山大施维英

评论:出现在 NIPS 2018

主题:**机器学习 (cs。LG)**;机器学习 (统计 ML)

深层结构模型广泛用于诸如语义分割这样的任务，其中变量之间的显式相关性提供了重要的先验信息，通常有助于减少深度网的数据需求。然而，当前的深层结构模型受到通常非常局部的邻域结构的限制，由于计算复杂度的原因不能增加，而且输出配置或其表示形式不能进一步转变。最近处理这些问题的方法包括深层网中的图形模型推断，以便允许后续的非线性输出空间转换。然而，这些配方的优化是有挑战性的，不太清楚。在这里，我们开发了一种新的模型，它概括了现有的方法，如结构预测能量网络，并讨论了一种保持现有推理技术适用性的公式。