Modern Engineering and Application 现代工程与应用, 2026, 4(1), 44-51
Published Online August 2026.
https://ufascience.com/mea/index

基于可解释无监督机器学习的信息物理系统维护

胡宇航，罗浩铭*

山西大学数学科学学院，太原 030006，山西，中国

摘要：信息物理系统（CPS）在当代技术格局中扮演着关键角色，对稳健、透明的机器学习（ML）模型的需求也因此变得迫切。本文探讨如何把可解释人工智能（XAI）的原则融入无监督机器学习（UML）方法，以增强对信息物理系统内部复杂关系的可解释性与理解。研究的重点包括：把自组织映射（SOM）作为具有代表性的无监督学习算法加以应用，以及引入可解释的机器学习方法。信息物理系统的数据由物理过程与数字部件融合而成，本身就极为复杂，本文对由此带来的挑战作了深入讨论。无监督学习中传统的黑箱做法往往妨碍人们理解模型给出的结论，因而并不适合关键的信息物理系统应用。为此，本文提出一个新的框架：在利用自组织映射这一强大的无监督方法的同时，借助可解释人工智能技术保证其可解释性。本文全面梳理了现有的可解释人工智能方法及其向无监督学习范式的迁移，并着重讨论如何为信息物理系统领域中习得的模式建立透明的表示。所提方法力求通过生成人类可理解的可视化结果与解释来提升模型的可解释性，从而弥合先进机器学习模型与领域专家之间的鸿沟。

关键词：信息物理系统；可解释人工智能；可解释机器学习；自组织映射；无监督机器学习

Cyber Physical Systems Maintenance with Explainable Unsupervised Machine Learning

YuHang Hu, HaoMing Luo*

School of Mathematical Sciences, Shanxi University, Taiyuan 030006, Shanxi, China

Abstract: As cyber-physical systems (CPS) continue to play a pivotal role in modern technological landscapes, the need for robust and transparent machine learning (ML) models becomes imperative. This research paper explores the integration of explainable artificial intelligence (XAI) principles into unsupervised machine learning (UML) techniques for enhancing the interpretability and understanding of complex relationships within CPS. The key focus areas include the application of self-organizing maps (SOMs) as a representative unsupervised learning algorithm and the incorporation of interpretable ML methodologies. The study delves into the challenges posed by the inherently intricate nature of CPS data, characterized by the fusion of physical processes and digital components. Traditional black-box approaches in unsupervised learning often hinder the comprehension of model-generated insights, making them less suitable for critical CPS applications. In response, this research introduces a novel framework that leverages SOMs, a

文章引用: 胡宇航，罗浩铭. 基于可解释无监督机器学习的信息物理系统维护[J]. 现代工程与应用, 2026, 4(1): 44-51.
DOI: https://doi.org/10.61784/mea2009

powerful unsupervised technique, while concurrently ensuring interpretability through XAI techniques. The paper provides a comprehensive overview of existing XAI methods and their adaptation to unsupervised learning paradigms. Special emphasis is placed on developing transparent representations of learned patterns within the CPS domain. The proposed approach aims to enhance model interpretability through the generation of human-understandable visualizations and explanations, bridging the gap between advanced ML models and domain experts.

Keywords: Cyber-physical systems; Explainable artificial intelligence; Interpretable machine learning; Self-organizing maps; Unsupervised machine learning

一、引言

在信息物理系统（CPS）领域，机器学习（ML）算法的融合正日益普遍。然而，许多机器学习模型的黑箱特性使人难以理解其决策，而这一点对信息物理系统的安全性与可靠性至关重要。本文探讨可解释无监督机器学习（EUML）方法的应用，重点关注可解释人工智能（XAI）原则、自组织映射（SOM）、可解释机器学习以及无监督机器学习（UML）[1-3]。机器学习方法向信息物理系统的快速渗透，已显著提升了这类复杂系统的能力。但许多机器学习模型的不透明性带来了理解与信任其决策上的困难，在安全攸关的领域尤其如此[4-5]。本文以可解释无监督机器学习方法为着眼点，回应信息物理系统对透明性与可解释性的需求。这一探讨的关键组成部分包括可解释人工智能的原则、自组织映射、可解释机器学习与无监督机器学习[6-7]。

随着对可解释无监督机器学习认识的深入，有必要思考无监督学习在信息物理系统中意味着什么——在这类系统里，物理部分与信息部分彼此关联，对开放程度的要求也更高。本文的引言为更深入地考察可解释无监督机器学习奠定基础，目标是弥合两方面的落差：一方面是无监督学习模型并不总是清晰可读，另一方面是信息物理系统需要清晰、可理解的决策过程[8]。后续各节将逐层展开有监督学习与无监督学习，考察可解释性的原则，并最终聚焦于无监督学习与可解释性在信息物理系统中的交汇之处。整个考察将以对可解释自组织映射的详细讨论收束，把它作为应对传统黑箱模型在信息物理系统应用中所带来挑战的一条有前景的路径[9]、。通过这一探讨，本文希望为信息物理系统这一复杂领域中可解释、可信赖的机器学习方案的演进有所贡献。

有监督机器学习（SML）是机器学习领域的基石，它在带标签的数据集上训练模型以完成预测或分类。有监督机器学习在各个领域的有效性早已得到确认[10]，因而被广泛采用。然而随着模型复杂度上升，其可解释性往往随之下降，决策过程也变得难以理解。在安全攸关的应用中，弄清模型为何作出某个具体预测至关重要，这促使人们去探索别的路径。与有监督机器学习不同，无监督机器学习处理的是无标签数据，力图在没有显式指导的情况下揭示数据背后的模式与结构。聚类与降维是无监督机器学习中的常见任务，而带标签样本的缺失使习得表示的可解释性面临挑战。由于信息物理系统涉及物理部件与信息部件之间的复杂交互，能否解读数据内部的潜在关系，就成为有效决策的关键[11]。

可解释机器学习（XML）作为一个关键方向出现，正是为了回应许多机器学习模型的黑箱特性。特征重要性与模型无关方法在提升可解释性上已见成效，但这些技术主要是为有监督学习场景设计的。随着机器学习在信息物理系统中的应用不断加深，无监督学习的可解释性需求日益凸显，促使人们去探索可解释无监督机器学习方法。有了这一背景，就更容易理解标准的有监督与无监督机器学习模型会带来哪些问题[12]，在信息物理系统中尤其如此。在随后几节中，本文将从可解释无监督机器学习的不足入手，把已有的可解释机器学习术语映射到无监督情形，梳理近期研究，并最终讨论它如何用于解决信息物理系统所特有的问题。把可解释人工智能纳入人工智能工作流的思路见图 1，其目标是在人工智能生命周期的各个阶段都使用可解释的方法。

二、可解释的无监督机器学习

可解释无监督机器学习包含一组关键诉求，正是这些诉求把它与传统的无监督学习方法区别开来。首要的诉求包括透明性、可解释性，以及对无监督模型决策过程给出洞察的能力。在信息物理系统的语境下，错误决策的后果可能十分严重，因此这些诉求对于保证所部署的机器学习模型可信、可靠就变得不可或缺[13]。

要开发出有效的可解释无监督机器学习算法，就必须把可解释机器学习的概念调整并延伸到无监督的情形。可解释性、开放性与可问责性都需要在无监督学习的框架下重新加以思考。把可解释机器学习的术语映射到可解释无监督机器学习，为评估和改进无监督机器学习模型的可解释性提供了一个框架。本文对可解释无监督机器学习方法的现状作了全面梳理，指出其进展、挑战与潜在应用。这一文献回顾说明了在解决无监督学习模型可解释性问题上已经取得的进展，并为随后在信息物理系统这一具体语境中考察可解释无监督机器学习作了铺垫。本节聚焦可解释无监督机器学习在信息物理系统领域中的应用[14]，讨论信息物理系统所特有的挑战，例如物理部件与信息部件之间的动态互动、实时决策的需要，以及复杂互联系统对透明性的要求，同时也讨论在信息物理系统应用中部署可解释无监督机器学习的潜在用例、收益与注意事项。

本节对可解释无监督机器学习的讨论，为后续几节更详细地考察一种具体方法——自组织映射——作了准备。通过陈述诉求、映射已有术语、回顾文献并把可解释无监督机器学习置于信息物理系统的语境之中，本文希望对复杂动态系统中可解释无监督学习的潜力与挑战给出一个较为全面的认识[15]。

（一）可解释自组织映射

自组织映射已成为无监督学习中的一种有力方法，在聚类与降维任务中尤为突出。自组织映射由 Kohonen 提出，它把高维输入数据映射到低维的神经元栅格上，并保持输入空间的拓扑关系。自组织映射在捕捉数据内部复杂结构上表现出色，但由于所习得表示本身的复杂性，其可解释性一直有限。

图 1 给出了二维自组织映射在输出空间与输入空间中的结构。图 1（a）表示自组织映射的输出空间，其中神经元排布在固定的二维栅格上，栅格上预先设定的拓扑连接保持了邻域关系；图 1（b）则是同一个自组织映射在训练之后映射到输入空间的样子，此时神经元调整自身位置以贴合输入数据的分布与簇结构。自组织映射的学习行为由若干关键超参数支配，即学习率与最佳匹配单元（BMU）的邻域半径，二者在每一轮迭代中逐步减小，以保证平滑收敛与准确的数据表示。

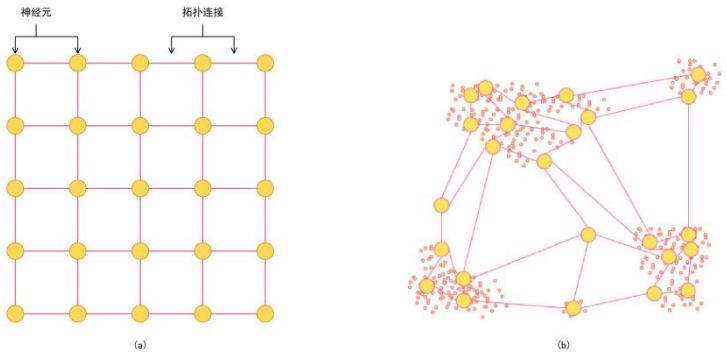

Figure 1 SOMs Shown in (a) the Output Space and (b) the Input Space Changed to Fit the 2D Spread of the Points that were Entered

图 1 自组织映射在（a）输出空间与（b）输入空间中的表现，输入空间中的映射已调整为贴合所输入各点的二维分布

为应对传统自组织映射在可解释性上的困难，人们提出了各种修改与扩展，由此产生了可解释自组织映射。

本节讨论对自组织映射所作的这些增强，重点在于这些修改如何使模型更易解释。神经元重要性打分、特征归因以及习得表示的可视化等技术都被纳入考察，以便更深入地理解自组织映射框架内部的决策过程。可解释性在信息物理系统中很重要，因为理解复杂数据集之间的相关关系在此至为关键。可解释自组织映射把自组织映射对拓扑结构的捕捉能力与可解释性所带来的透明性结合起来，有望为信息物理系统应用中的无监督学习提供一个稳健的方案。

随后几节将进一步考察可解释自组织映射的实验设置与结果，评估其模型保真度、局部与全局可解释性，以及在信息物理系统中的可用性。这一实证分析旨在验证可解释自组织映射在应对信息物理系统所提出的具体挑战时是否有效，并评估其在复杂动态环境的真实应用中的潜力。

三、实验设置与结果

（一）模型保真度

在实验设置中，本文对可解释自组织映射的保真度作了严格评估：以传统自组织映射为对照开展比较分析，考察可解释自组织映射能否准确表示给定信息物理系统数据集中的复杂关系。评估采用聚类准确率、拓扑结构保持度与重构误差等指标来量化模型的保真度。

（二）局部可解释性

本文考察可解释自组织映射的局部可解释性，以了解模型在多大程度上能对单个数据点给出洞察。研究运用神经元重要性打分与特征归因技术，逐样本地识别影响模型决策的关键因素。这一分析旨在揭示可解释自组织映射在信息物理系统语境下所能达到的可解释性粒度。

（三）全局可解释性

本文进一步从更宏观的视角评估可解释自组织映射的全局可解释性：通过在系统整体层面考察习得的表示，评估模型揭示信息物理系统数据集中总体模式、异常与关联关系的能力。研究采用热力图、簇摘要等可视化手段，以便全面理解可解释自组织映射所达到的全局可解释性。

（四）在信息物理系统中的可用性

本文开展了真实场景实验，评估可解释自组织映射在信息物理系统中的可用性：把模型部署到信息物理系统环境中，在模拟这类系统动态、互联特性的场景下考察其表现。分析中一并考虑了实时决策、对变化条件的适应能力，以及对系统可靠性与安全性的总体影响。

图 2 给出了本研究在 bank_marketing 数据集上所用聚类质量指标的变化情况，说明簇数与自组织映射维度如何影响这些指标的演化。对给定的自组织映射规模，聚类质量指标分别在自组织映射神经元权重（蓝色）与训练数据集（橙色）上计算。该分析的目的是确定理想的自组织映射维度与簇数。当训练好的自组织映射神经元能够准确代表整个数据集时，可以预期对自组织映射权重所作的聚类分析会呈现出与整个数据集相当的模式。图 2 表明，随着簇数增加，两者遵循同样的模式。Davies-Bouldin 指数值与轮廓系数表明，对所考察的自组织映射维度而言，三到五个簇是最优的。

本文按标准差对特征排序，并对随机特征或无关紧要的特征加以改动，以检验所提假设。也就是说，在给定 p% 基数的前提下，分别对最重要的特征（标准差最低者）、随机选取的特征和最不重要的特征改动 p%。对测试集中的每一条数据记录，本文都检查在上述三种情形下改动特征取值是否会改变其簇标签。每种情形又考察两种可能：一是可以被另一个簇替换簇标签的测试数据记录所占比例，二是可以被其余所有簇替换的比例。对某些数据点而言，仅仅按相近簇的判据扰动特征取值，未必足以把它移出原来的簇；本文因而确保至少来自

另外一个簇的特征取值有可能影响某个数据点的簇归属。设想一个四簇的场景，数据点 j 属于簇 2，本文尝试把它的簇标签由 2 改为其他，并用簇 1、簇 2 与簇 4 的均值替换其特征取值。可以预期，基数 n% 越小、被替换的比例就越大。各数据集的替换比例见图 3。除 KDD 数据集的第二种情形（即“可被其余所有簇替换簇标签的测试数据记录占多大比例？”）之外，蓝色柱代表相关特征，棕色柱代表随机特征，各数据集皆是如此。KDD 数据集类别极不均衡、训练与测试之间差距显著，可能正是该情形下表现不佳的原因。不过在第一种情形（有多少测试数据记录的簇标签可能被改变？）下，KDD 的表现符合预期。这些实际结果印证了本文的预判：所提出的自组织映射方法是依据关键判据来确定数据记录的簇标签的。

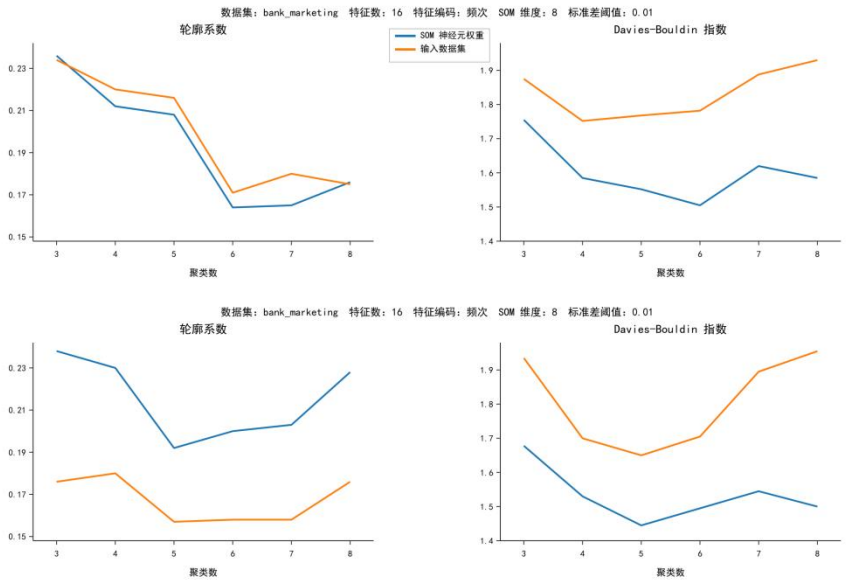

Figure 2 A Way to Judge the Quality of K Clusters Using the Silhouette Coefficient and the Davies-Bouldin Index for Various SOM Map Sizes

图 2 用轮廓系数与 Davies-Bouldin 指数判定不同自组织映射规模下 K 个簇质量的一种方法

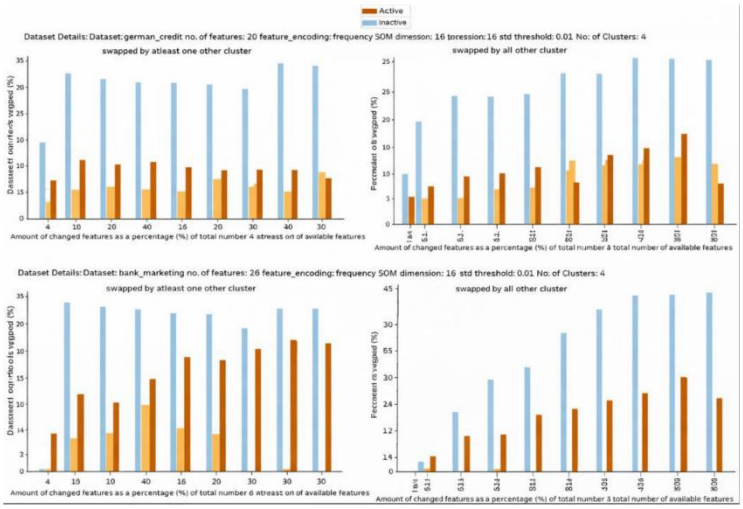

Figure 3 Clustering Performance Comparison Using SOM and Input Dataset

图 3 使用自组织映射与输入数据集的聚类性能对比

本文还做了一项补充实验，以验证所选的 K 个特征中有多大比例进入了某个最佳匹配单元的最重要特征

表（算法 IV）。对测试集中的每一条数据记录，先逐特征计算该记录与其最佳匹配单元之间的 l1 距离，再按 l1 距离由小到大对特征排序。本文的推断是：一个数据点最突出的特征应当是在空间上离它最近的那些特征，而这些特征也应出现在其最佳匹配单元的优先特征表之中。在特征距离按升序排好之后，用最近、随机、最远三种策略之一挑选 K 个特征，进而确定有多大比例的 K 个特征进入了最佳匹配单元的关键特征表。结果见图 4，其中横轴为特征总数 K，纵轴为被判定为相关（即进入特征表）的那部分特征所占的比例（%）。蓝色代表最近的特征，黄色代表随机特征，绿色代表最远的特征。对所有的 K 而言，蓝色柱都给出最高的比例，黄色柱次之。这说明各最佳匹配单元所确定的关键特征表中，确实包含了邻近的特征。

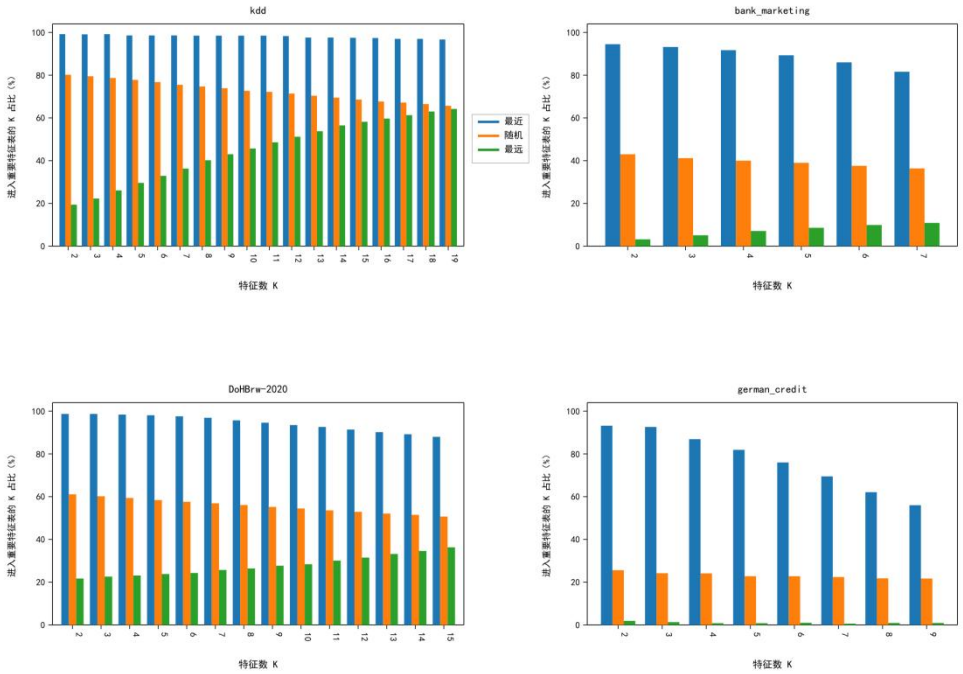

Figure 4 The Percentage of Nearest K Features in the BMU's most Significant Feature List

图 4 最近的 K 个特征进入最佳匹配单元最重要特征表的比例

图 5 展示了 KDD 数据集中 flag 特征在各簇之间变化时的全局可解释性。自组织映射的神经元被归为三组，U 矩阵给出了各簇之间的距离与分离程度，关键特征的取值区间则与簇归属对照呈现；此外还借助 u-map 核验了簇的离散情况。KDD 数据集的 flag 特征如图 5 所示：第一幅图是原始数据，描绘自组织映射神经元的簇分离情况；第一行第二幅图给出 flag 特征在三个簇（分量图）上的取值，可见其取值在三组之间各不相同；第一行第三幅图中三个簇界限分明，颜色较浅的区域表示神经元之间的距离，区域越大则簇之间越可区分。图 5 第二行给出各簇特征取值尺度的细粒度表示。需要注意的是，即使在同一个簇内部，不同神经元在同一特征上的取值也可能不同；领域专家若想了解某个特征在簇内部的表现，正需要这类数据。簇 0 的 flag 特征取值较大（0.7—1.0），簇 1 呈现中间区间（0.45—0.52），另一个簇则呈现很低的区间（0.15）。此外，图中还给出了某一特征取值出现在簇内的可能性。以簇 0 为例，特征取值在各簇之间存在差异。

初始实验的完整性检验见图 4。本文对 p 取了多个不同的值，并分别处理活跃特征、随机选取的特征与最不重要的特征，计算在全部特征中改动 p 个之后簇标签发生变化的数据点所占的百分比。考察的两种情形是：簇标签可以被另一个簇的标签替换的测试数据记录所占比例（左），以及没有任何簇的标签可以将其替换的比例（右）。

本文对实验设置所得的结果作了批判性讨论，强调可解释自组织映射在信息物理系统语境下的长处与潜在局限，并就可扩展性、计算效率与泛化能力作了讨论；此外还讨论了在真实的信息物理系统应用中部署可解释自组织映射的实际意义，从而对其在提升模型透明性与可解释性上的效果给出较为全面的认识。本节说明了可解释自组织映射如何解决信息物理系统中无监督机器学习的问题。下一节将从这些数据出发得出推论，并就这一快速发展的方向提出进一步研究与开发的建议。

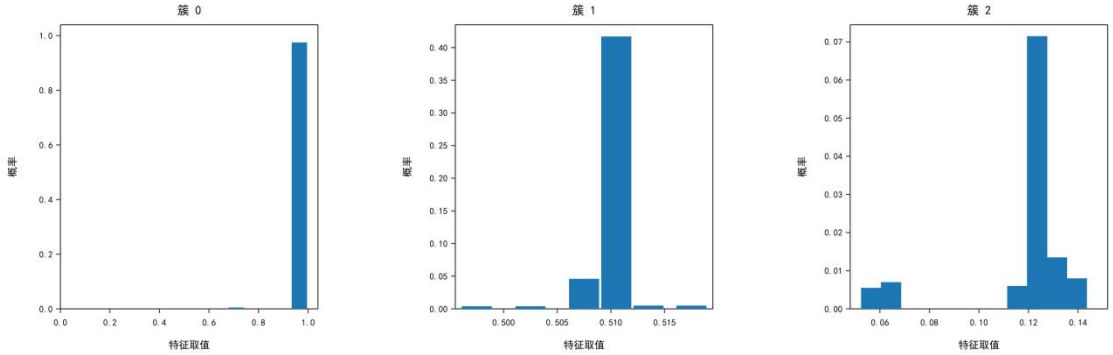

Figure 5 Feature behavior for the 'flag' feature of the KDD data set across clusters

图 5 KDD 数据集 flag 特征在各簇之间的表现

四、结论

本文对可解释无监督机器学习作了考察，讨论了它的诉求、已有可解释机器学习术语向无监督情形的迁移、当前文献的梳理，以及可解释无监督机器学习原则在信息物理系统这一复杂领域中的应用。作为一种有前景的可解释无监督机器学习方法，可解释自组织映射的引入回应了无监督学习模型对可解释性的需求。随后各节详述了实验设置与结果，对模型保真度、局部与全局可解释性以及可解释自组织映射在信息物理系统环境中的可用性作了细致考察。通过开展真实场景实验并分析性能指标，本文为可解释自组织映射在提升信息物理系统应用透明性与可解释性上的有效性提供了实证依据。讨论部分把理论洞察与实证发现结合起来，指出可解释自组织映射在信息物理系统中的长处、局限与实际意义，并就可扩展性、计算效率与实时决策作了讨论，从而对可解释无监督机器学习可能带来的影响给出一个整体的图景。展望未来，围绕可解释无监督机器学习——尤其是在信息物理系统语境下——继续开展研究大有可为。可解释无监督学习技术的进步，将为复杂动态系统中安全、可靠、透明的机器学习方案的持续发展作出重要贡献。本文只是一块垫脚石，期待更多人在可解释无监督机器学习与信息物理系统的交汇处继续探索与创新。

参考文献

[1] 李清泉. 车载激光雷达高精度定位技术. 测绘学报, 2023, 51(04): 437-446.

[2] 赵金洲. 页岩油体积压裂增产技术. 石油学报, 2022, 43(04): 511-520.

[3] Malik M I, Ibrahim A, Hannay P, et al. Developing resilient cyber-physical systems: a review of state-of-the-art malware detection approaches, gaps, and future directions. Computers, 2023, 12(4): 79.

[4] 苑世剑. 大型构件液态模锻成型工艺. 塑性工程学报, 2023, 29(04): 1-8.

[5] 周绪红. 钢混组合桥梁受力优化设计. 中国公路学报, 2022, 35(04): 1-14.

[6] 戴琼海. 工业视觉缺陷检测深度学习模型. 自动化学报, 2020, 48(06): 1105-1116.

[7] 林君. 深部矿产电磁探测装备研发. 地球物理学报, 2022, 65(05): 1921-1932.

[8] Jabeen R, Singh Y, Sheikh Z A. Machine learning for security of cyber-physical systems and security of machine learning: attacks, defences, and current approaches//The International Conference on Recent Innovations in Computing, 2022: 813-841.

[9] Mozaffari F S, Karimipour H, Parizi R M. Learning based anomaly detection in critical cyber-physical systems//Security of Cyber-Physical Systems. Cham: Springer International Publishing, 2020: 107-130.

[10] 徐惠彬. 航空热障涂层制备技术. 材料研究学报, 2022, 36(04): 241-256.

[11] 刘清友. 海洋钻井平台防腐防护技术. 海洋工程, 2020, 40(02): 1-9.

[12] 王建华. 基坑工程承压水突涌防控. 岩土工程学报, 2020, 44(06): 1001-1010.

[13] 谭久彬. 超精密测量仪器误差抑制技术. 光学精密工程, 2022, 30(06): 685-698.

[14] Hasan M K, Abdulkadir R A, Islam S, et al. A review on machine learning techniques for secured cyber-physical systems in smart grid networks. Energy Reports, 2024, 11: 1268-1290.

[15] Ahmed R S, Ahmed E S A, Saeed R A. Machine learning in cyber-physical systems in industry 4.0//Artificial Intelligence Paradigms for Smart Cyber-Physical Systems, 2021: 20-41.