软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(10):3732−3753 [doi: 10.13328/j.cnki.jos.006365]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

机器学习中原型学习研究进展*

张幸幸[1,2], 朱振峰[1,3], 赵亚威[4], 赵 耀[1,3]

[1](北京交通大学 信息科学研究所, 北京 100044)

[2](清华大学 计算机科学与技术系, 北京 100084)

[3](现代信息科学与网络技术北京市重点实验室(北京交通大学), 北京 100044)

[4](国防科技大学 计算机系, 湖南 长沙 410073)

通信作者: 赵耀, E-mail: yzhao@bjtu.edu.cn

摘 要: 随着信息技术在社会各领域的深入渗透, 人类社会所拥有的数据总量达到了一个前所未有的高度. 一方面, 海量数据为基于数据驱动的机器学习方法获取有价值的信息提供了充分的空间; 另一方面, 高维度、过冗余以及高噪声也是上述繁多、复杂数据的固有特性. 为消除数据冗余、发现数据结构、提高数据质量, 原型学习是一种行之有效的方式. 通过寻找一个原型集来表示目标集, 以从样本空间进行数据约简, 在增强数据可用性的同时, 提升机器学习算法的执行效率. 其可行性在众多应用领域中已得到证明. 因此, 原型学习相关理论与方法的研究是当前机器学习领域的一个研究热点与重点. 主要介绍了原型学习的研究背景和应用价值, 概括介绍了各类原型学习相关方法的基本特性、原型的质量评估以及典型应用; 接着, 从原型学习的监督方式及模型设计两个视角重点介绍了原型学习的研究进展, 其中, 前者主要涉及无监督、半监督和全监督方式, 后者包括基于相似度、行列式点过程、数据重构和低秩逼近这四大类原型学习方法; 最后, 对原型学习的未来发展方向进行了展望.

关键词: 原型学习; 数据约简; 度量学习; 模型优化; 机器学习

中图法分类号: TP181

中文引用格式: 张幸幸, 朱振峰, 赵亚威, 赵耀. 机器学习中原型学习研究进展. 软件学报, 2022, 33(10): 3732–3753. http://www.jos.org.cn/1000-9825/6365.htm

英文引用格式: Zhang XX, Zhu ZF, Zhao YW, Zhao Y. Prototype Learning in Machine Learning: A Literature Review. Ruan Jian Xue Bao/Journal of Software, 2022, 33(10): 3732–3753 (in Chinese). http://www.jos.org.cn/1000-9825/6365.htm

Prototype Learning in Machine Learning: A Literature Review

ZHANG Xing-Xing[1,2], ZHU Zhen-Feng[1,3], ZHAO Ya-Wei[4], ZHAO Yao[1,3]

[1](Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China)

[2](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

[3](Beijing Key Laboratory of Advanced Information Science and Network Technology (Beijing Jiaotong University), Beijing 100044, China)

[4](Department of Computer Science, National University of Defense Technology, Changsha 410073, China)

Abstract: With the in-depth penetration of information technology in various fields, there are many data in the real world. This can help data-driven algorithms in machine learning obtain valuable knowledge. Meanwhile, high-dimension, excessive redundancy, and strong noise are inherent characteristics of these various and complex data. In order to eliminate redundancy, discover data structure, and improve data quality, prototype learning is developed. By finding a prototype set from the target set, the data in the sample space can be reduced, and then the efficiency and effectiveness of machine learning algorithms can be improved. Its feasibility has been proven in many applications. Thus, the research on prototype learning has been one of the hot and key research topics in the field of machine

* 基金项目: 科技创新 2030——“新一代人工智能”重大项目(2018AAA0102101); 国家自然科学基金(U1936212, 61976018)
收稿时间: 2020-08-26; 修改时间: 2021-01-22; 采用时间: 2021-04-17; jos 在线出版时间: 2021-05-20

learning recently. This study mainly introduces the research background and application value of prototype learning. Meanwhile, it also provides an overview of specialties of various related methods in prototype learning, quality evaluation of prototypes, and typical applications. Then, the research progress of prototype learning with respect to supervision mode and model design is presented. In particular, the former involves unsupervision, semi-supervision, and full supervision mode, and the latter compares four kinds of prototype learning methods based on similarity, determinantal point process, data reconstruction, and low-rank approximation, respectively. Finally, this study looks forward to the future development of prototype learning.

Key words: prototype learning; data reduction; metric learning; model optimization; machine learning

在当今信息爆炸时代, 信息的种类和数量空前激增. 面对如此海量的数据, 以机器学习尤其是深度学习为核心的人工智能技术得到了长足的发展. 然而需要指出的是, 数据在量上的膨胀未必能带来在质上的提高. 如何有效地选择“用的了”且“用的好”的数据、如何从数据中获取最有用的信息, 成为摆在机器学习研究中的重要问题. 诚如《大趋势》的作者奈斯比特所说: “我们被数据淹没, 但却渴求着知识”[1]. 一方面, 海量数据为基于数据驱动的机器学习方法获取有价值的信息提供了充分的空间; 另一方面, 高维度、过冗余以及高噪声也是上述繁多、复杂数据的固有特性. 这不但造成存储资源的巨大浪费, 而且还会显著提升学习算法的复杂度. 更严重的是: 它们还会将真正有价值的信息湮没, 从而恶化学习算法的性能. 为消除数据冗余、发现数据结构、提高数据质量, 从特征空间与样本空间进行数据约简是两种行之有效的方式, 在增强数据可用性的同时, 提升机器学习算法的执行效率. 其中, 前者涉及到的技术包括特征降维(dimensionality reduction)[2,3]和特征选择(feature selection)[4,5]; 而后者则涉及样本空间的原型生成(prototype generation)[6]和原型选择(prototype selection)[7]. 本文将样本空间的原型生成与选择, 统称为原型学习(prototype learning).

实质上, 原型学习问题涉及到众多领域的应用场景, 因而作为机器学习的研究重点之一, 与原型学习相关的理论与方法的研究得到了国际上众多学者的普遍关注. 在国际有关机器学习的主流会议, 如 Advances in Neural Information Processing Systems (NIPS)、Int’l Conf. on Machine Learning (ICML)、Int’l Joint Conf. on Artificial Intelligence (IJCAI)和 AAAI Conf. on Artificial Intelligence (AAAI)等, 以及 IEEE Trans. on Pattern Analysis and Machine Intelligence (IEEE TPAMI)、Journal of Machine Learning Research (JMLR)等重要国际杂志上, 每年都有大量的关于原型学习的最新工作发表. 此外, 来自美国东北大学的 Ehsan Elhamifar 教授、耶鲁大学的 Amin Karbasi 教授、IBM Research AI 的 Rameswar Panda 研究科学家等人在 Computer Vision and Pattern Recognition (CVPR 2016、CVPR 2018、CVPR2019)国际会议上, 专门组织了关于原型选择中的算法与优化的专题讲座[8,9]. 通过对以上大量文献的梳理可以看出: 原型学习的研究成果有助于挖掘出数据中最具价值的信息, 提高用于机器学习的数据质量, 降低机器学习算法的计算复杂度、节约目标数据的存储成本、实现机器学习模型的轻量化(模型压缩)等. 同时, 这也为大数据时代下的计算机视觉、图像与自然语言处理、生物医学、信息推荐等领域提供理论基础与技术支撑, 满足与原型学习有关的应用需求.

鉴于原型学习问题在机器学习中的重要性, 国内一些研究机构近些年也对此开展了相关研究, 诸如南京理工大学的杨静宇教授课题组[10]、西安电子科技大学的焦李成教授课题组[11]、清华大学的张长水教授课题组[12]、南京大学的周志华教授课题组[13]、北京大学的张志华教授课题组[14]、中国科学院自动化研究所的刘成林研究员课题组[15]等. 这些课题组的工作主要围绕监督条件下的核学习、主动学习以及示例学习中的原型选择、矩阵列选择问题、图像分类中的原型学习等进行研究. 此外, 国内的一些研究学者还基于粗糙集理论从数据的不确定性角度开展数据约简研究[16], 这类方法虽然能够有效去除数据冗余进而发现数据结构, 但是对获得的原型的代表性缺乏直观物理解释, 并且原型的质量还不足以满足众多应用的需求. 更为重要的是: 尽管目前国内外学者已经发表了大量关于原型学习的研究成果, 但是关于原型学习的综述性文献却很稀少, 对于原型的定义与解释也不够清晰. 因此, 本文梳理了原型学习领域的相关文献, 对不同文献所采用的方案、面向的应用以及存在的问题进行了归纳总结. 通过对前人工作的学习与理解, 我们能够发现原型学习领域研究近几十年来的理论与应用发展脉络. 同时, 通过分析近几年原型学习领域的最新研究成果, 我们可以把握当前主流的研究兴趣与方向, 探究诸多应用背景对原型学习的具体需求, 从而对未来原型学习研究的理论与应

用发展方向进行一定的预测, 进而更高效地地服务实际应用.

具体来说, 本文首先赋予原型学习明确的数学定义与物理概念, 并介绍原型学习的研究背景和应用价值. 接下来依据目前相关文献的内容, 概括介绍了各类原型学习方法的基本特性、原型的质量评估标准以及原型学习的典型应用. 在此基础上, 我们进一步挑选不同类型原型学习方法中具有代表性的文献, 对其解决的问题以及方法进行深入介绍. 从原型学习的监督方式及模型设计两个视角重点介绍了原型学习的研究进展, 其中, 前者主要涉及无监督、半监督和全监督方式, 后者包括基于相似度、行列式点过程、数据重构和低秩逼近这四大类原型学习方法. 最后, 重新梳理了原型学习领域研究的发展脉络, 综合文献研究成果及应用实例, 根据目前研究中存在的问题及原型学习研究领域发展趋势, 探讨未来可能的发展方向.

1 原型学习的相关概念与研究意义

为了能够更好地理解原型学习的物理意义, 本节首先从广义上明确地阐述原型学习的定义及重要组成部分, 这也是首次赋予原型学习明确的概念. 进而介绍机器学习中原型学习的研究背景, 尤其是在计算机视觉、模式识别、数据可视化、资源配置、信息推荐等领域的应用价值与研究意义.

1.1 原型学习的相关概念

原型学习, 通常也被称为数据/示例选择(data/instance selection)[17,18]、子集选择(subset selection)[19]、样例选择(exemplar/representative selection)[20,21]、核心集构造(core-set construction)[22,23]. 原型学习问题可以定义为[7,24]: 假设有一个源集(source set) $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_m\}$和一个目标集(target set) $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,...,\boldsymbol{y}_n\}$, 分别包含 m 和 n 个元素, 原型学习旨在从源集 $\boldsymbol{X}$ 中寻找一个原型集$\boldsymbol{\Omega}\subseteq\boldsymbol{X}$, 使得$\boldsymbol{\Omega}$能够最大程度地保持目标集 $\boldsymbol{Y}$ 所蕴含的信息; 同时, $\boldsymbol{\Omega}$中所有元素具有最少的重叠信息. 图 1 为一个原型学习示意图(从一个源集 $\boldsymbol{X}$ 中学习 k 个原型, 从而使得源集 $\boldsymbol{X}$ 到目标集 $\boldsymbol{Y}$ 的全连接, 约简为从原型集$\boldsymbol{\Omega}$到目标集 $\boldsymbol{Y}$ 的全连接). 源集和目标集之间关系的不同, 对应了不同的实际场景. 例如: 当源集 $\boldsymbol{X}$ 为与目标数据集 $\boldsymbol{Y}$ 有关的 m 个模型组成的集合时(如为描述生成目标数据集的先验分布集合), 该问题演化为模型原型学习(model prototype learning); 当源集 $\boldsymbol{X}$ 为与目标数据集 $\boldsymbol{Y}$ 有关的 m 个样本组成的集合时, 其可演化为样本原型学习(sample prototype learning); 特别地, 当源集 $\boldsymbol{X}$ 与目标集 $\boldsymbol{Y}$ 一致时(即 $\boldsymbol{X}=\boldsymbol{Y}$), 此即为典型的原型学习.

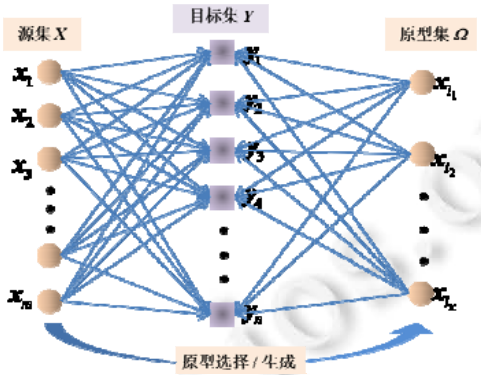


图 1 原型学习示意图

需指出的一点是: 原型学习和聚类技术有一定的相关性, 但在定义、目标、实施和结果上又不尽相同.

- 就定义和目标而言: 原型学习旨在利用源集学习出目标集的一个原型集, 来表示目标集的结构或容量等信息(如图 2 所示); 而聚类旨在对目标集进行分组, 从而获得目标集的语义信息. 当源集 $\boldsymbol{X}$ 与目标集 $\boldsymbol{Y}$ 是同一个数据集时, 原型学习可以利用聚类技术实现, 其中, 聚类中心担任目标集的原型. 因此在某种意义上, 原型学习可以视作一种细粒度的聚类问题;
- 在实施和结果上, 不同于经典的聚类问题, 原型学习不仅可以在无监督条件下完成, 还可以在半监督

或者全监督下实施. 最终, 原型学习算法获得的原型的数量是灵活变化的, 而聚类技术获得的类中心的个数却是通过先验知识预先确定的. 原型选择机制不必输出每一个元素的语义信息, 而是更加精准地捕获数据的内蕴结构.

图2直观展示了一个合成数据集的模型与样本原型选择结果. 当选择样本(数据)原型时, 源集 $\boldsymbol{X}$ 和目标集 $\boldsymbol{Y}$ 一致, 均为红色的样本点; 当选择模型原型时, 源集 $\boldsymbol{X}$ 是绿色的三角形、四边形、五边形等各种多边形, 目标集 $\boldsymbol{Y}$ 是所有红色的数据点. 该原型选择的目的是保持目标集的容量与结构信息.

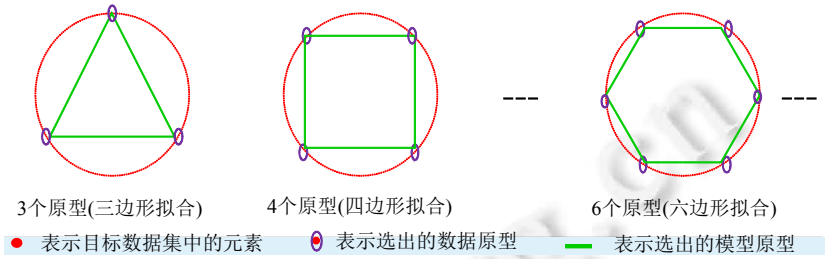
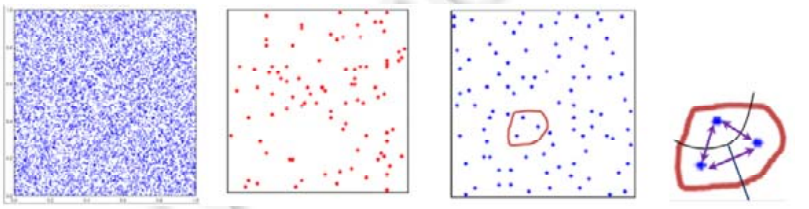


图2 合成数据的模型原型与数据原型选择结果, 该原型选择的目的是保持目标集的容量与结构信息

为保证能够从目标集中选择或者生成最具信息量的元素, 从而实现对目标集中所有元素的精准刻画, 本文把机器学习中原型学习理论与方法研究的重要问题凝炼为:

- 原型学习的多样性问题——如何保证学习出的原型能够在样本空间具有很好的覆盖性. 面对繁多而复杂的数据, 从全局来说, 选择出来的原型应能够“窥一斑而知全豹”, 即对于固定规模的原型集, 其能够在样本空间中具有充分的覆盖性, 从而反映数据的总体特性. 而如何考虑目标集的全局结构, 是保持原型学习多样性的关键;
- 原型学习的可解释性问题——如何保证学习出的原型能够在样本空间具有很好的代表性. 从局部来看, 原型集中的每个元素应具有很好的代表性, 即能够最大程度地蕴含该原型邻域内的目标元素的共性信息, 从而代替该邻域执行其他任务. 本质上, 这涉及到如何定量评价目标集中潜在原型的显著程度;
- 原型学习的相容性问题——如何保证学习出的原型能够在样本空间具有很好的互斥性. 对于源集中的某个元素, 当其被选为原型时, 与其相关性强的元素被选为原型的概率将大幅度降低, 这也意味着原型集中的元素彼此间应具有互斥性, 即具有最少的冗余. 而如何引入互斥性约束, 是解决该问题的核心.

对于上述原型学习中涉及的核心问题, 可借助图 3 进一步理解. 对于一个在二维空间中均匀分布的目标集/源集(如图3(a)所示), 与图3(b)中通过随机选择得到的原型集相比, 实际应用更希望得到如图3(c)所示的保有多样性、可解释性以及相容性的原型选择结果, 并将依据任务需求着重强化其中某些特性, 比如: 场景分类任务需要强化原型的多样性, 以识别尽可能多的场景; 视频摘要任务需要着重提高原型的可解释性, 以概括整段视频所表达的内容; 资源配置任务需要优化原型的相容性, 以最小化资源分配的成本.

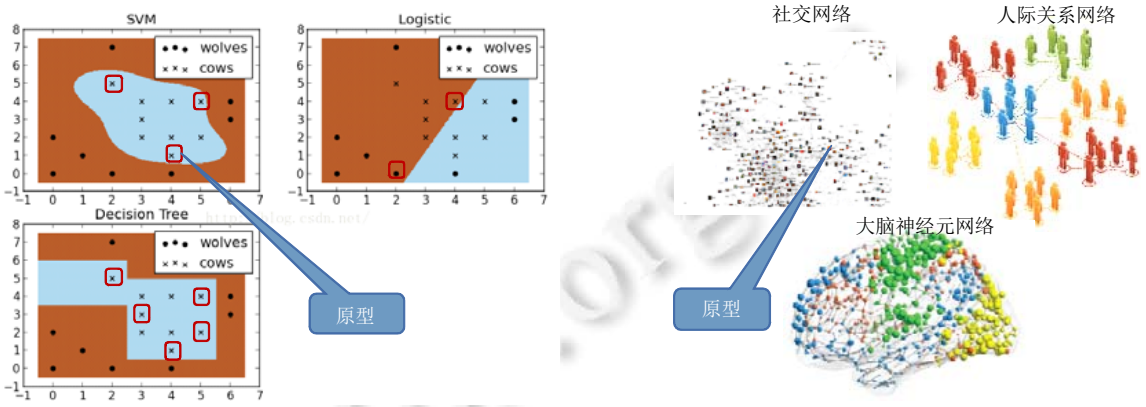


(a) 均匀分布的目标集 (b) 随机选择 (c) 保有多样性、可解释性以及相容性的选择 (d) 局部放大的原型互斥性

图3 原型学习中核心问题示意图

1.2 原型学习的应用场景

如何从海量数据中合理地选择原型, 在机器学习和数据科学中占有重要地位. 目前, 原型学习问题已经是机器学习中主动学习、自步学习、生成对抗网络、支持向量机、模型压缩等算法的核心. 如图 4(a)所示, 通过学习目标集的原型, 来重构支持向量机(support vector machine, SVM)、Logistic 回归、决策树(decision tree)等机器学习算法[25−29]的分类面; 如图 4(b)所示, 除了图像等视觉数据, 通过原型学习, 还可以对社交网络、人际关系网络、大脑神经元网络等图数据进行采样. 因此, 原型学习不仅面向图像、文本等视觉数据类型, 还可以面向图数据等多元化数据类型[30], 实现其原型集的选择.



(a) 基于原型学习的机器学习算法 (b) 原型学习诱导的图数据采样任务

图 4 机器学习中的原型学习问题

实际上, 原型学习问题还同时涉及了计算机视觉、模式识别、图像和自然语言处理、生物医学、传感网络、信息推荐等领域的众多应用. 部分基于原型学习的应用示例如图 5 所示. 其中, 对于相簿更新系统[31], 摄影集担任源集, 而用户的初选照片担任目标集, 使用原型选择方法能够从源集中选出最符合用户原始收藏习惯的照片; 对于视频摘要系统[32], 源集和目标集均是目标视频, 通过对目标集的原型选择来获取视频序列的关键帧, 从而形成对目标视频的故事概述; 在电影推荐系统中[33], 客户未观看的电影担任源集, 而所有已观看的电影担任目标集, 使用原型选择方法可以从源集中选出最迎合客户喜好的电影并加以推荐; 在摘要时间表生成系统中[19,34,35], 语料库同时担任源集和目标集, 通过对目标集的原型选择来提取该语料库的摘要时间表, 以方便人们快速获取其重要信息; 对于运动分割系统[36], 目标视频序列同时担任源集和目标集, 通过对目标集的原型学习来定位运动目标的主要轨迹; 在人脸提取系统中[37], 人脸集合同时担任源集和目标集, 通过对目标集的原型选择来消除身份重复的人脸图像, 从而节约数据存储成本; 对于设施选址系统, 设施担任目标集, 全部候选位置担任源集, 通过从源集中选择原型去代表目标集来实现设施选址.

互联网信息爆炸时代, 膨胀的数据规模只有在有效的数据处理下才能够进行高效的机器学习. 其中, 原型学习是数据处理的重要手段, 它可以完成对数据的快速存储、压缩、生成、清洗、可视化和标注. 下面给出原型学习问题的一些典型应用[38,39].

- 数据存储、压缩: 挑选出的原型能够代替目标集描述复杂事件, 化繁为简, 从而用于数据压缩或者模型蒸馏[40−42]; 此外, 基于原型所获得的学习、推断算法在取得与基于原始数据同等性能的同时, 还可以大幅度降低计算时间和内存需求, 诸如 k 最近邻(k-NN)[43]、支持向量机(support vector machine, SVM)[44]、卷积神经网络(convolutional neural networks, CNN)[45]、零样本学习(zero-shot learning, ZSL)[46,47]、小样本学习(few-shot learning, FSL)[48]等算法;
- 数据可视化: 挑选原型能够有效地可视化文本/网页、语音、图像和视频信息, 如视频摘要、文本摘要、相册主题等, 从而在分析数据时增强其可解释性, 同时也可满足用户在日常生活中快速浏览存储的

需求[19,22,49];

- 数据清洗、合成: 原型可以被用来合成服从目标集分布的新的数据. 机器学习经常面临有效数据不足和缺失问题(如不平衡的类分布和多视角数据缺失), 该问题在异常检测中至关重要, 诸如电力盗窃、银行的欺诈交易、罕见疾病识别等. 而使用原型选择或者生成技术不仅可以净化目标集, 还可以合成新的数据, 以解决数据不足、缺失问题[50];
- 数据呈现: 在商品推荐系统中, 对排序模型得到的商品推荐列表进行原型选择, 使得最终呈现给用户的是一个最具代表性和多样性的商品子集. 这样既能推荐给用户所倾心的商品, 同时, "多样性"也保证了商家能够推荐虽然目前不流行但很有前景的潜在商品[51];
- 数据标注: 众多基于机器学习的模型都涉及到数据标注问题, 以便进行全监督/半监督学习. 然而, 随着数据标注成本的提高, 在海量数据中有选择地标注样本变得至关重要. 而原型选择能够提供一个极具代表性和多样性的样本子集用于标注, 从而极大地缓解了标注成本与工作量. 同时, 还可以消除原始数据中的异常, 进而提高下游算法的效率[52];
- 资源配置: 原型选择可以协助有关部门实现资源的合理分配. 例如: 在对传感网中的传感器进行布置时, 在可供选择的 n 个位置中, 选择一个由 k 个位置组成的子集进行配置, 可以达到有限资源最大化利用的目的[53,54].

(a) 相簿更新系统(信息推荐); (b) 电影推荐系统(信息推荐); (c) 视频摘要系统(数据可视化);
(d) 时间表生成系统(数据可视化); (e) 运动分割系统(模式识别);
(f) 人脸提取系统(数据存储与压缩); (g) 设施选址系统(资源配置)

图 5 基于原型学习的应用示例

2 原型学习的研究概况

本节从原型学习的相关方法以及原型质量的定性/定量评估两个层面, 对原型学习的国内外研究现状进

行梳理和概述.

2.1 原型学习的相关方法

从技术角度, 原型的获取可以通过两种方式: 一类是直接从源集中选取代表性的数据来构成原型集, 称为选择法; 另外一类则是在源集中通过融合方式重新生成一组具有代表性的点, 称为生成法. 实际应用中, 若源集未被特定给出, 一般默认目标集同时担任源集. 目前, 根据是否有语义信息约束, 原型学习方法可以分为无监督[52,55,56]、半监督[57,58]和全监督[45,58−61]的原型学习. 其中,

- 全监督下的原型学习一般根据样本对分类结果的贡献度来选择对模型最具可解释性的样本原型, 诸如最为经典的学习矢量量化(learning vector quantization, LVQ)方法[62,63]以及第一个基于深度学习的原型学习方法——卷积原型学习(convolutional prototype learning, CPL)方法[15];
- 半监督下的原型学习需要通过带标签样本向不带标签样本的信息传递来推理出所有样本的权重及相关性, 进而令权重较高的样本担任目标集的原型. 文献[57]提出了一种集成原型选择与特征学习的算法, 这也是在半监督场景下学习原型的最新代表作;
- 最后, 作为最常见、研究频率最高的场景, 无监督原型学习通常根据目标集的固有分布与结构来选择对数据最具代表性的样本原型, 诸如 Elhamifar 教授在文献[36]中提出的基于非相似性的稀疏子集选择(dissimilarity-based sparse subset selection, DS3)方法以及在文献[64]中提出的稀疏建模样例选择(sparse modeling representative selection, SMRS)方法.

最近, 文献[65]提出一种面向视频分析的原型学习方法, 能够灵活应用于无监督、半监督和全监督原型学习场景中.

事实上, 原型学习主要通过优化一个特定的目标函数, 诸如设施选址(facility location)[38,66]、 最大割(maximum cut)[55,56]、最大边缘相关度(maximum marginal relevance)[67]、稀疏编码(sparse coding)[68,69]或行列式点过程(determinantal point process, DPP)[70]来选择数据集中最具有代表性或多样性的一个子集作为原型集. 然而, 几乎所有的原型学习准则都存在非凸(non-convex)和 NP-hard 问题[36]. 因此, 原型学习从优化角度可以视作一个次模优化(submodular optimization)问题[71−73]. 根据对原型学习模型设计, 尤其是优化准则的不同, 目前的原型学习方法主要可以分为 4 类: (1) 基于相似度/非相似度的原型学习[38,55,56,58,66,67,74]; (2) 基于行列式点过程的原型选择[75−77]; (3) 基于数据重构的原型学习[64,78−81]; (4) 基于低秩逼近的原型选择[82,83]. 本质上, 基于相似度/非相似度的原型学习方法主要通过最小化目标集与原型集之间的全局非相似性, 来选出满足需求的最具代表性的原型集. 较为典型的有 K 中心点(K-medoids)方法[84]、仿射传播聚类(affinity propagation clustering, AP)方法[39]以及最新发表的 DS3 方法[36]. 基于行列式点过程的原型选择方法, 诸如 DPP[76]和 k-DPP[75], 旨在通过最大化行列式的取值来解决数据间的相容性问题, 从而完成原型的多样化选择. 基于数据重构的原型学习方法通常将原型选择视作字典选择问题, 并最小化原型集重构目标字典的误差, 以选择出最具代表性的原型. SMRS[64]是此类方法的典型代表, 而后出现结构化稀疏字典选择(structured sparse dictionary selection, SSDS)方法[79]等, 以利用不同的正则约束从不同角度来着重提高原型的多样性、代表性或互斥性. 不同于上述 3 类方法, 基于低秩逼近的原型选择方法(诸如 CUR[81]和 Nyström 方法[83,85,86])通常利用矩阵分解找到目标集矩阵的几行(列)来近似整个低秩矩阵, 而这几行(列)即目标集的原型集. 最近, 文献[38]首次尝试将原型选择与度量学习相结合, 以显著提高原型的代表性和多样性. 本质上, 该方法融合了基于相似度/非相似度的原型学习以及基于数据重构的原型学习方法的双重优势, 因此也得到广泛的关注.

图 6 利用现有文献举例阐述了原型学习的 3 种监督方式和 4 类原型学习模型. 从中发现: 除基于数据重构的原型学习模型之外, 基于相似度、行列式点过程以及低秩逼近的三大类原型学习模型均能在不同的监督方式下工作. 原因在于: 现有的基于数据重构的原型学习方法还未充分考虑如何有效利用少量数据的语义信息, 来进一步提高原型的代表性.

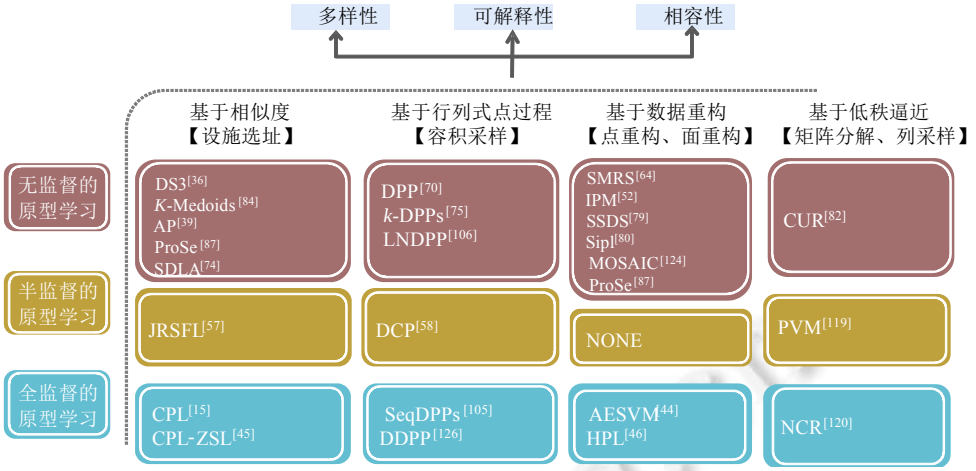


图 6 原型学习相关方法分类图与文献索引

2.2 原型的质量评估

在原型学习中, 所学原型的质量, 尤其是原型的多样性、可解释性和相容性, 直接决定了机器学习、数据分析与处理、计算机视觉等原型学习相关应用的效率. 因而, 如何评价原型质量的好坏至关重要. 现有的原型学习方法涉及定性和定量两种评价方式, 但并不存在一个统一的评价标准. 总的来说, 定性评价通常观察并可视化原型在目标集中的位置, 以证明所学原型完全覆盖了目标集的分布, 并揭示了它的结构信息[79,87]. 而定量评价则需要依赖具体任务与算法. 对于全监督原型学习方法, 分类、检索等任务的性能间接反映出原型的质量. 例如: CPL[15]在 CNN 中额外引入原型学习损失函数, 从而不仅提高了在大数据集上的图像识别性能, 还能够识别新出现的图像种类. 而识别精度的提高, 主要源于所学原型的强判别性和代表性. 与全监督场景类似, 半监督原型学习方法通常也借用图像分类、检索、视频分析等任务的性能来反映原型的质量[65]. 此外, k-NN、SVM、CNN 等分类算法通常也被用于评价原型的判别性和多样性. 其中, 从原始训练集中选择出的原型充当分类算法的约简训练集, 用于执行测试集的分类任务[36,63,74,79].

不同于上述两类有监督原型学习场景, 无监督原型学习方法通常利用原型学习最直观的应用, 诸如视频摘要[64,83]、推荐系统[51]和运动分割[36,87], 以评价原型的代表性和相容性. 特别地, 无监督下的原型一般被赋予明确的物理意义, 比如视频摘要中的关键帧和运动分割中的运动目标, 因此, 原型的可解释性也得到了充分保证.

3 原型学习的监督方式

原型学习中, 语义信息约束了原型学习的监督方式. 目前, 原型学习方法依据监督方式可以分为 3 类: (1) 无监督原型学习[52,55,56]; (2) 半监督原型学习[57,58]; (3) 全监督原型学习[15,58−61].

3.1 无监督原型学习

目前, 有关原型学习的大部分工作都属于无监督范畴, 它们主要通过优化一个特定的准则, 诸如设施选址[38,66]、最大割[55,56]、最大边缘相关度[67]、稀疏编码[68]或行列式点过程[70], 来选择目标数据集中最具有代表性的一个子集. 无监督原型选择方法最终需要使得该子集满足覆盖性、代表性和互斥性, 并根据真实原型对其进行定性或定量评估.

若要选择覆盖面积最大的子集, 实现原型的多样化选择, 一种常见的方法是应用 DPP[70]. 近年来, DPP 的许多变体, 如 k-DPPs[75]被提了出来, 并应用于计算机视觉和模式识别等多个领域[49]. 例如: 一个用于选择 k 个原型的代表性方法 k-DPPs[75], 通常被用来选择视频序列的关键帧.

为了显著提高原型集的代表性, 现有文献通常采样稀疏编码准则和设施选址准则. 其中, 前者通常假定

目标数据位于一个或多个子空间中[64,87,88], 这样便于将原型选择问题转换为稀疏字典选择问题, 并用字典重构误差衡量原型集在目标集中的重要性. 因此, 字典选择方法一般可被用于原型选择. Elhamifar 等人最先尝试该选择方案, 并验证了其有效性[64]. 此后, 文献[68]在字典选择中加入局部先验以抑制异常值, 提高原型选择的鲁棒性. 文献[79]进一步为字典选择施加结构稀疏约束, 以同时增强原型的多样性. 设施选址准则[71,83,89]一般则是基于给定的成对相似性或相异性, 选择编码损失(服务成本)最小的数据点作为原型. 其中, 原型集编码目标集的损失与原型的重要性成反比关系. 该准则实质上与聚类方法[39,90,91]密切相关, 因此部分聚类方法也可用于原型选择.

原型集内元素间的关联性对于面向时序数据的视觉应用十分重要, 因此, 最大割准则和最大边缘相关准则被用于原型学习. 具体来说, 最大割准则即求一个任意图的最大割, 是 Karp 的 21 个组合问题(Karp-Cook 列表)之一[55]. 将目标集中的元素视作这个图的顶点, 该准则使得原型集和互补子集之间的边数尽可能地大, 而后借用多项式时间算法或者近似算法得到顶点集的一个子集, 即原型集. 而最大边缘相关准则善于保持数据重排列后的查询相关性, 同时力求减少原型集的冗余度, 因而经常应用于文本/视频摘要任务中[67]. 综上所述, 尽管没有语义信息可以利用, 但是无监督原型学习能够最大程度地利用目标集的固有结构, 并从中选出信息密度最高、最能够代表目标集的原型集.

3.2 半监督原型学习

尽管文献[39,49,70,89,92–94]已经对原型学习问题进行了比较完备的研究, 但是大多数以前的工作都采用了无监督方式. 也就是说, 在没有监督的情况下, 从给定的目标数据中找到最具代表性的一个子集. 事实上, 在许多应用中, 用户不仅希望找到一个有代表性的子集(即哪些样本是原型), 而且希望了解它们是什么(即识别它们的类别). 换句话说, 尽管人们可以基于相似性度量找到目标集的原型并将剩余的数据样本与之关联, 但除非提供标签, 否则根本不知道每个原型的确切类别. 近来, Elhamifar 等人引入一个附加的已知各元素是什么的源集, 并建议选择该源集中的元素作为原型来表示目标集[34,47]. 这样, 通过传递源集中原型的类别标签, 可以很容易地识别目标项. 但是, 这个方案是在一个封闭的假设下运行, 即源集合知道目标中所有可能出现的类别. 在许多实际应用中可能并不是这样, 例如, 最近发表的文献[57]采用半监督方法, 同样引入一个源集, 并将原型选择建模成设施选址问题, 但是利用它从目标集中而不是源集中找出代表性样本. 更重要的是, 该方法并不认为源集覆盖了目标数据中的所有类别. 此外, 该方法将标记的源数据合并到目标函数中, 以便它们可以监督新目标数据的原型选择. 最后, 通过选定的原型来形成源数据和目标数据之间的连接, 进而将标签从源集转移到目标集. 如图 7 所示, 利用文献[57]提出的方法可以学习出目标集的原型, 进而将源集中“马”“狮子”和“小鹿”的标签转移到目标集中相应的分组内.

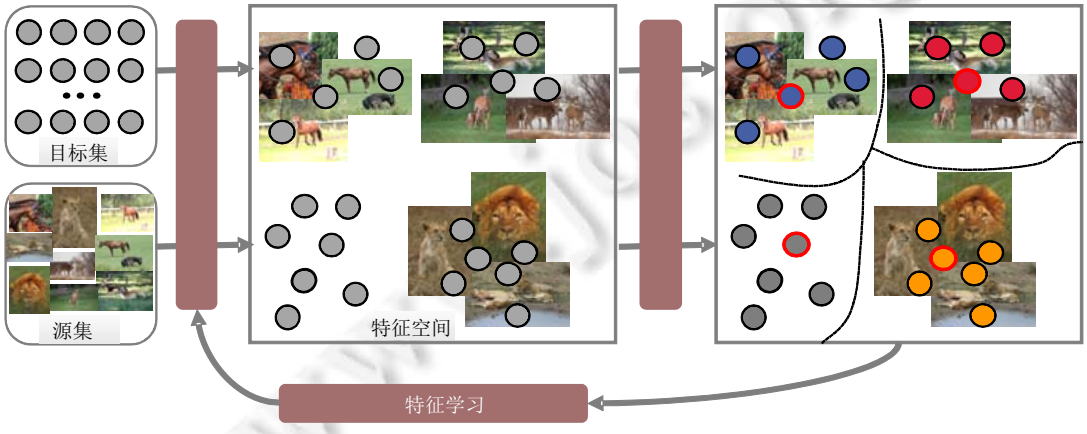


图 7 基于 CNN 的半监督原型学习方法示意图

3.3 全监督原型学习

实际应用中, 可通过诸多渠道获取数据的语义信息, 并用其监督原型学习[34,49,95]. 目前, 有监督原型学习有基于浅层模型和深度学习的原型学习两种.

基于浅层模型的原型学习方法通常用来约简 *k*-NN、SVM 等推理算法的训练数据. 通过选择原始训练集的原型并将其作为新的训练集, 来减少 *k*-NN、SVM 等算法的存储空间和计算时间. 同时,利用原型集去除原始训练集中的冗余和噪声, 进而提高 *k*-NN、SVM 等算法的分类性能[43,44,63,96−99]. 例如, 文献[98]提出一种约简支持向量机(reduced support vector machines, RSVM), 其核心思想是: 从整个训练集的每一类中随机挑选一个小容量子集, 来构造一个约简的核矩阵, 并用它代替原本全稠密的核矩阵, 以快速学习一个非线性 SVM 模型. 针对 RSVM 中随机选择原型致使分类性能不够稳定的缺陷, 一系列新的原型选择策略被提出. 文献[44]提出一种基于凸包和极值点的线性时间算法, 以在每一类训练数据的核空间中选择最具代表性的原型. 实验结果表明: 用该原型集训练的 SVM 可比标准的 LIBSVM 快 10^3 倍, 同时取得与 LIBSVM 相当的分类精度. 同样地, 为了提高 *k*-NN 分类的效率, 文献[97]提出一种寻找 *k*-NN 分类器原型的新方法. 它通过梯度下降和确定性退火过程优化原型在训练集中的位置, 并设计一个原型初始化策略, 旨在以最少的原型数目为训练集提供最大的分类精度, 从而实现机器学习模型(*k*-NN)的轻量化.

由 Kohonen 提出的 LVQ[62]是另一种最近邻原型分类器, 它可以被视为一个两层神经网络模型, 在输出层的每个节点有一个权值向量(即目标集的原型)与它连接. 通过使不同类别权向量之间的边界逐步收敛至贝叶斯分类边界, 来寻找权值即原型的最优值. 目前, LVQ 已经产生很多变种[63], 可以进一步分为两大类: 第 1 类旨在设计收敛速度更快的权值更新条件和规则来学习原型[99]; 另一类则通过定义与原型相关的损失函数, 并通过优化损失函数来系统地学习原型[97]. 需要注意的是: 该类原型学习方法都是基于手工设计的特征, 在 CNN 到来之前曾被广泛应用于各种模式识别和计算机视觉任务中.

最近, 得益于深度学习的快速发展, 原型学习也开始借助神经网络强大的表征学习能力, 训练基于深度学习的原型学习模型. 文献[15]提出的 CPL 方法是第一个尝试该原型学习模式的工作, 如图 8 所示, 它将 CNN 与基于分类器的原型学习相结合. 假设所有类服从高斯分布, CPL 通过最大化类内聚集度和类间散度, 设计一个基于距离度量的原型损失函数, 从而学习目标集中每一类别的数据原型. 进而, 将测试数据与所有原型做匹配, 可以对测试集做出判决, 最终实现高精度和强鲁棒的模式分类. 同样地, Elhamifar 教授也提出一个有监督原型选择框架, 它将深度表征学习与基于设施选址准则的原型选择相串联, 从而学习一个有监督代表性原型选择模型, 并有效应用于任意的视频摘要任务中[61]. 可以发现: 引入深度学习的数据原型学习通常是将目标集的表征学习与原型学习相融合, 利用目标集更具表示能力的 CNN 特征来显著提高原型的代表性. 有监督模式下的大量实验结果也表明: 相比传统的基于手工特征的原型学习, 基于深度学习的原型学习方法虽然也是通过优化目标函数来学习原型, 但其在算法求解效率和原型学习性能上具有更大的优势, 因此得到更广泛的关注.

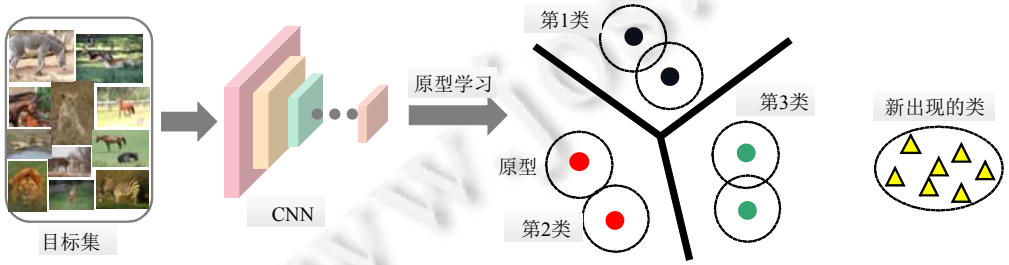


图 8 基于 CNN 的全监督原型学习方法示意图

4 原型学习的相关模型

为了获取满足机器学习算法、数据分析与处理、计算机视觉、模式识别、信息推荐、资源配置等后续任

务需求的高质量原型, 现有的原型学习方法通常设计各种各样的目标优化函数, 以约束原型的可解释性、多样性以及相容性. 依据模型的优化目标, 目前, 原型学习方法主要包括 4 类: (1) 基于相似度/非相似度的原型学习; (2) 基于行列式点过程(DPP)的原型选择; (3) 基于数据重构的原型学习; (4) 基于低秩逼近的原型选择.

本节从原型学习的模型设计视角, 详细描述目前原型学习的常用方法, 并分析该类方法所面临的各种问题及诸多挑战.

4.1 基于相似度的原型学习

原型学习可以在目标集的特征空间或者度量(核)空间中进行, 其中, 后者促使一系列基于相似度/非相似度的原型学习模型被提出. 目前, 该类方法主要涉及以下几种原型选择准则: 最大割目标(maximum cut objective)[55,56]、限量的/不限量的设施选址目标(capacitated/uncapacitated facility location objectives)[38,66]以及最大边缘相关度(maximum marginal relevance)[67]. 本质上, 基于相似度/非相似度的原型学习方法主要通过最小化目标集与原型集之间的全局差异来选出极具可解释性(即代表性)的原型集.

较为典型的方法是 *K*-Medoids[84], 不同于 *K*-Means 方法[100]在连续空间寻找最优解, *K*-Medoids 的取值只能是目标集 $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,...,\boldsymbol{y}_n\}$中的元素. 但是二者的目标一致, 都希望原型到其余所有(当前类别中的)元素的距离之和最小. 它可以用如下数学形式来描述:

$$\min_{\mu_j\in Y,r_{ij}}\sum_{i=1}^{n}\sum_{j=1}^{k}r_{ij}D(y_i,\mu_j) \tag{1}$$

其中, $D(y_i,\mu_j)$表示目标集中的元素 $\boldsymbol{y}_i$ 和当前参考原型μ_j之间的差异值, 这个值是人为提前设定的, 诸如欧氏距离、余弦距离等; r_{ij} 是 $\boldsymbol{y}_i$ 与当前参考原型μ_j之间的隶属度. 求解公式(1)则可以得到最优解 $\{\mu_j\}_{j=1}^{k}$, 而这正是目标集的 k 个原型.

基于 *K*-Medoids 和 *K*-Means 提出的 *K*-prototypes[101]和 AP 方法[39], 虽然能够基于(非)相似度量选出原型, 但是以迭代的方式逐一选择原型容易使最终选择结果陷入局部最优. 考虑到上述不足, 近来, Elhamifar 等人在给定非相似度的基础上, 通过一个行稀疏正则化的迹最小化模型, 称为 DS3[36], 将目标集合中的每一个元素都指派给一个原型. 并且, 在此基础上, 对于时序数据, Elhamifar 等人将转移概率考虑到 DS3 模型[94]中, 使得所选原型能够精准捕获序列数据的分布特点, 且在原型序列上呈现出较高的转移概率; 进一步地, 提出采用次模优化算法实现时序数据的在线原型选择[71]; 为完成同样的任务, Elhamifar 教授又开发了一个新的效用函数(utility function)及一个快速优化算法, 用于序列数据中的子集选择, 其中, 边际收益不再基于每一个条目单独计算, 而是利用任务的序列结构和动态规划精确计算[102]. 最近, Elhamifar 教授提出一种协作序列式子集选择框架, 以解决来自无约束教学视频中的无监督程序学习(procedure learning)问题. 该框架通过学习同一个任务下多个视频的状态与过渡, 建立一个动态模型, 找到所有状态中最能够代表所有输入视频的一个状态子集, 从而提取出全部视频所描述的该任务的关键步骤及次序[103]. 同时, 为实现有监督原型学习, 基于设施选址效用函数, Elhamifar 教授进一步提出一个有监督子集选择框架. 本质上, 该框架旨在学习输入数据的 CNN 特征, 使其到“设施选址”能够有效恢复出数据给定的原型[61]. 然而: (1) 此类方法虽然主流, 但是由于源集和目标集分布未知, 通常不在一个域, 且容易被异常点污染, 选择何种(非)相似性度量也众说纷纭; (2) 基于相似度选择原型需要工作在全局的(非)相似性之上, 不能依据样本规模进行尺度伸缩变化, 所以不具有普适性, 不能够有效地应用于大规模数据集的原型学习.

4.2 基于行列式点过程的原型选择

行列式点过程(DPP)首次由 Macchi 在 1975 年定义[104], 一个在目标集 $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,...,\boldsymbol{y}_n\}$上的点过程 P 是目标集 $\boldsymbol{Y}$ 上的一种点模式的概率测度, 是 $\boldsymbol{Y}$ 的有限子集. 为了从 $\boldsymbol{Y}$ 中找到一个能够代表 $\boldsymbol{Y}$ 的子集$\boldsymbol{\Omega}$, DPPs 定义了一种在 $\boldsymbol{Y}$ 上的概率测度如下:

$$P(\Omega;L)=\frac{\det(L_\Omega)}{\det(L+I)} \tag{2}$$

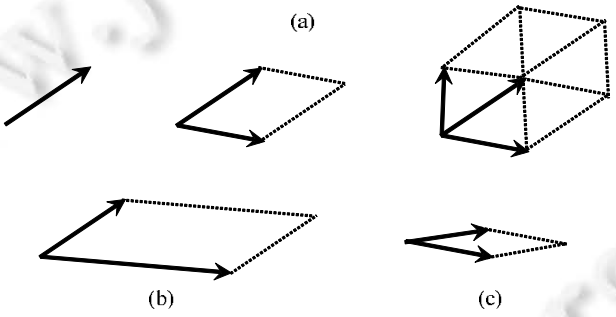
$$\sum_{\Omega\subseteq Y}\det(L_\Omega)=\det(L+I) \tag{3}$$

其中, det(·)表示计算一个矩阵的行列式; $\boldsymbol{I}$ 是单位矩阵; $L\in\mathbb{R}^{n\times n}$ 为在目标集 $\boldsymbol{Y}$ 上定义的半正定矩阵, 也被称为 $\boldsymbol{L}$-因子($\boldsymbol{L}$-ensemble), L_{ij} 度量了目标集中第 i 个和第 j 个元素之间的关联, $\boldsymbol{L}_{\boldsymbol{\Omega}}$是由$\boldsymbol{\Omega}$中元素在 $\boldsymbol{L}$ 中索引的子块矩阵. 因为需要计算 $\boldsymbol{L}$, 基于 DPP 的原型选择实质上也是在度量空间进行的. 最终, 子集$\boldsymbol{\Omega}$作为目标集中原型的概率与矩阵 $\boldsymbol{L}_{\boldsymbol{\Omega}}$的行列式(即容量)成正比关系. 通过最大化容量子集(maximum volume subset), 来找到一个最具多样性的原型集$\boldsymbol{\Omega}$. 它的数学形式表达如下:

$$\max_{\Omega\subseteq Y}P(\Omega;L) \tag{4}$$

本质上, DPP 可以视作一个抽样方法, 如图 9 所示, 目标集中两个元素作为原型被抽取的概率不仅与单一元素被抽取的概率相关, 还与这两个元素的相关性有关. 如果单一元素被选择的概率越大, 同时元素之间的相似度越低, 这个集合被选择的概率就会越高. 作为一种衡量观测子集元素间关系的概率统计模型, DPP 很好地解决了原型集内元素间的相容性(即互斥性)问题, 有利于形成原型集的多样性(覆盖性), 因而在与原型选择有关的推理、摘要生成、重要性采样、传感器布置、推荐系统等方面得到了广泛的应用[54,105]. 由于 DPP 通常由半正定核矩阵参数化, 根据该矩阵的对称性, 基于 DPP 的原型选择方法可以被分成两大类.

- 一类是对称 DPPs, 它能够在上述应用中取得可观的性能. 但是由于对称核只能编码元素间的互斥性或负相关, 这使其在实际应用中存在很大的局限性. 以购物网站的商品选择为例, 其任务是在结账前给用户购物篮提供良好的商品推荐, 对于只能编码负相关性的对称 DPP 模型, 不可能直接编码正交互性, 例如, 购买的包含视频游戏机的购物篮更可能也包含游戏控制器;
- 非对称 DPPs 能够有效地解决该缺陷. 然而, 由于非对称核分解的复杂性, 目前只有极少数文献[106]关注这一类算法.



(a) 选择集合$\boldsymbol{\Omega}$的概率是$\boldsymbol{\Omega}$内元素张成的空间容积
(b) 若$\boldsymbol{\Omega}$内一个元素的幅度变大, 选择$\boldsymbol{\Omega}$的概率就随之增大
(c) 若$\boldsymbol{\Omega}$内两个元素相似度增大, 选择$\boldsymbol{\Omega}$的概率将会降低

图 9 DPP 的几何物理意义

为消除标准 DPP 中采样子集容量的不确定性, Kulesza 等人在文献[75]中提出了固定子集规模的行列式点过程 k-DPP, 以实现采样过程的可控性. 针对 DPP 中参数似然估计的非凸性以及 NP-hard 问题, Gillenwater 等人提出了一种基于期望最大化算法的 DPP 参数估计方法, 用于获得边缘核矩阵的稳健估计[107]. 考虑视频数据的时序结构特性, Gong 等人[49]提出了一种序贯行列式点过程 SeqDPP, 以序贯方式实现视频的具有时序关联的多样性采样. 图像的自动标注也可看作是一个标注标签列表中的原型挑选问题, 为此, Wu 等人[108]提出了一种基于 DPP 的图像自动标注方法, 能够使得标注的标签具有很好的语义覆盖面. 此外, 类似于 DPP 的原理, 容积采样(volume sampling)方法[77,108]同样利用矩阵行列式对样本空间容积进行度量, 从而使得选择的原型具备良好的物理解释性. 尽管基于 DPP 的原型选择不再存在计算复杂度的问题, 但是仍有几个开放问题值得探究: (1) 是否可以在更复杂的约束下对容许集进行 DPP 推理? (2) 对于 DPP 编码的条件独立性关系, 是否存在一个更易实现的描述? (3) 如何从一个有标签的训练集中学习一个 DPP 的相似核?

4.3 基于数据重构的原型学习

与基于相似度的原型学习不同，基于数据重构的原型学习通常是在目标集的特征空间学习原型. 目前，这类方法主要涉及多线性编码(multi-linear coding)[64]和稀疏编码两种评价准则，其核心思想是：通过最小化原型集重构目标集的残差，来保证原型的可解释性(即代表性). 总体而言，基于重构的方法包含稀疏字典选择(sparse dictionary selection, SDS)[78]、SMRS[64]、SSDS[79]等点重构方法，以及 RSVM[98]等面重构方法.

对于基于点重构的原型学习方法，其主要思想是，目标集中每一个元素都能够表示成源集中元素的一个线性组合[12]. 较为典型的方法是利用稀疏线性编码准则的 SDS[78]和 SMRS[64]，二者的区别在于对选择矩阵的稀疏性做不同的正则约束，以满足异常检测、视频摘要等不同的任务需求. 然而，SMRS[64]等方法只考虑了原型的代表性. 为了同时提高原型的多样性(覆盖性), Wang 等人在 2017 年提出了一个结构化稀疏字典选择模型(SSDS)[79]，它通过构造 3 项正则系数约束来实现原型的多样化选择. 一般地，基于点重构的原型选择方法可以用如下数学形式表达:

$$\min_{C\subseteq \Pi} \| Y - XC \|_F^2 + f(\boldsymbol{C}) \tag{5}$$

其中，$C\in \mathfrak{R}^{m\times n}$ 是重构系数矩阵，$f(\boldsymbol{C})$是施加在 $\boldsymbol{C}$ 上的稀疏约束，Π是 $\boldsymbol{C}$ 的可行域，一般需要满足等式约束 $\mathbf{1}^{\mathrm{T}}\boldsymbol{C}=\mathbf{1}^{\mathrm{T}}$. 本质上，稀疏约束还能够约束所选原型的数量，最常见的 $f(\boldsymbol{C})=\|C\|_{0,q}$且 $q>1$. 一旦获得稀疏表示系数矩阵 $\boldsymbol{C}$，原型索引可以通过 $\boldsymbol{C}$ 的行稀疏值来确定，从而从源集 $\boldsymbol{X}$ 中选出一个最具代表性的子集来表示目标集 $\boldsymbol{Y}$. 图 10 展示了基于数据重构的原型学习方法框架，依托选出的原型，还能够将目标集分组.

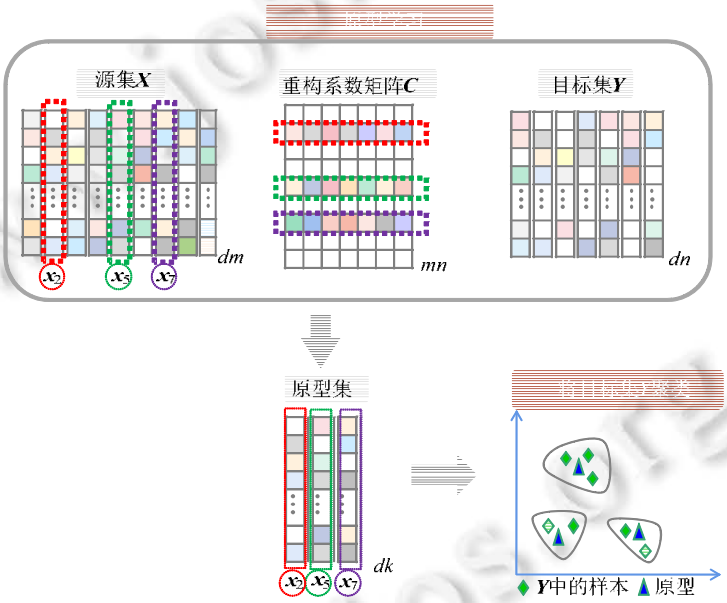


图 10 基于数据重构的原型学习方法示意图

有时，数据不能仅仅用点重构表示，所以，基于平面重构的方法恰好弥补了点重构的缺陷. 其主要针对分类决策与训练样本凸包有关的一类分类器，包括 SVM、k-NN 等，提出了一系列原型选择方法，从而构造最优分类面，提高分类算法的泛化能力[43,44,96–98]. 例如，在 SVM 模型中，权重系数准确地量化了每个数据在线性/非线性 SVM 分类面重构时的重要性；文献[44]提出根据权重系数快速选出能够重构分类面的训练数据(即原型)；RSVM[98]则利用从训练集中随机选出的一个子集(即原型集)来训练 SVM 分类面，并使其能够逼近原始训练集学习到的非线性 SVM 分类面.

然而，已有的基于重构的原型选择工作仍有若干严重不足，包括：(1) 重构函数对原型的代表性的影响非常明显，而目前几乎所有的函数都是线性重构，并没有探索源集与目标集的结构信息；(2) 该大类原型学习方

法在最终选择原型时的标准不够统一, 常用的基于重构系数权重的方案过于主观; (3) 此外, 原型学习本质上是针对海量数据的优化问题, 现有技术对相关理论性问题(如最佳原型数量的上下界)考虑不足; (4) 基于面重构的原型选择模型需要监督信息参与, 这在实际应用中具有一定的局限性.

4.4 基于低秩逼近的原型选择

不同于上述 3 类方法, 基于低秩逼近选择原型能够在特征空间或者度量(核)空间进行. 在特征空间, 较为典型的方法为排序显示 QR (rank revealing QR, RRQR)[109]、列子集选择(column subset selection, CSS)[110−112]和 CUR 分解[82], 其主要思想是: 通过矩阵分解, 并利用随机或贪婪算法来寻找低秩矩阵列的子集, 使得目标集矩阵的几行(列)能够近似整个低秩矩阵, 而这几行(列)即目标集的原型集. 例如, CUR 分解本质上可以看作是如下优化问题:

$$\min_{C,U,R} \| Y - CUR \|_F^2 + f(\boldsymbol{C}, \boldsymbol{R}) \tag{6}$$

其中, $f(\boldsymbol{C},\boldsymbol{R})$表示施加在矩阵 $\boldsymbol{C}$ 和 $\boldsymbol{R}$ 上的约束, 且矩阵 $\boldsymbol{C}$ 由目标集 $\boldsymbol{Y}$ 中的几列组成, 而矩阵 $\boldsymbol{R}$ 由目标集 $\boldsymbol{Y}$ 中的几行组成.

为了能够实施低秩矩阵分解, 目标集通常需要位于一个或多个低维子空间中. 该要求与基于数据重构的原型学习方法类同, 并且最小化目标集的重构误差也保证了原型的可解释性(即代表性). 然而, 这类基于矩阵分解的操作方法会带来巨大的计算代价, 也不适合海量数据的原型选择问题. 近年来, 核方法已成功地应用于各种复杂和非线性结构的实际问题中. Nyström 方法被认为是最具代表性的核矩阵低秩逼近方法[113−120], 其主要思想是: 从核矩阵中选择某几列, 然后使用采样列之间的相关性以及剩下的列形成一个对原始核矩阵的低秩逼近. 具体来说, 给定目标集 $\boldsymbol{Y}$ 的半正定核矩阵 $K \in \mathfrak{R}^{n\times n}$, 从中采样 $c<<n$ 列来产生 $\boldsymbol{K}$ 的近似矩阵, 然后将 $\boldsymbol{K}$ 的行和列根据采样后的结果重新排列如下:

$$\boldsymbol{K} = \begin{bmatrix} W & K_{21}^{\mathrm{T}} \\ K_{21} & K_{22} \end{bmatrix}, \ \boldsymbol{C} = \begin{bmatrix} W \\ K_{21} \end{bmatrix} \tag{7}$$

Nyström 方法旨在使用公式(7)中的 $\boldsymbol{W}$ 和 $\boldsymbol{C}$, 并依照公式(8)中的规则来获取 $\boldsymbol{K}$ 的 $k \leqslant n$ 秩逼近矩阵 $\tilde{\boldsymbol{K}}$:

$$\tilde{\boldsymbol{K}} = CW_k^{+}C^{\mathrm{T}} \approx \boldsymbol{K} \tag{8}$$

其中, $\boldsymbol{W}_k$ 是 $\boldsymbol{W}$ 关于谱范数或者 Frobenius 范数的最佳 k 秩逼近, 而 $\boldsymbol{W}_k^{+}$ 是 $\boldsymbol{W}_k$ 的伪逆矩阵.

所以, 在使用 Nyström 方法时, 一个重要的问题是如何采样获取原型. 最简单、最常见的方案是均匀采样[112], 此外还有贪婪式[83,113]和序贯式采样[114]. 而为了获得先验的采样概率, 各种杠杆分数(leverage score)指标也被用来定义采样概率[115]. 文献[116]给出了不同采样方案的实验比较. 最近, Wang 等人[14]提出了一种自适应采样方法, 并对采样误差进行了分析. 不同于一般的 Nyström 方法只使用一个列的子集, Li 等人[86]提出了一种集成 Nyström 方法, 通过重复列采样并进行组合来实现矩阵逼近. 这将生成一个更精确的近似值, 但同时也使得计算的复杂度增加数倍.

基于以上描述, 通过矩阵逼近的方式选择出的样本子集(原型), 只是从重构核矩阵的角度近似最优, 并且缺乏良好的几何与物理解释意义, 也不具有多样性和相容性. 此外, 为了达到最优逼近, 往往还需要借助其他方法, 如 K-Means[100]等来获得相应的子集, 这无疑为原型选择增加了额外的计算量.

5 总结与展望

本节将对原型学习的现有工作加以概括和总结, 并针对目前研究内容的缺陷, 对原型学习未来的工作给予规划和展望.

5.1 原型学习的工作总结

本文通过对现有文献的梳理和总结, 分别从原型学习的监督方式与模型设计两个层面, 介绍了已有的原型学习方法.

目前, 原型学习方法依据监督方式可以分为无监督、半监督和全监督原型学习. 其中, 无监督场景最为常见, 而半监督研究成果最为稀少. 然而, 由于语义信息的约束, 全监督下学习的原型具有更强的判别性, 而无监督学习出的原型由于充分挖掘出目标集的分布与结构信息, 因此具有更强的代表性. 另外, 依据模型设计的目标, 现有的原型学习方法主要有基于相似度/非相似度、行列式点过程、数据重构和低秩逼近的原型学习这 4 种. 其中, 基于数据重构的原型学习对服从子空间分布的目标集更有效; 而基于相似度/非相似度的原型学习虽然不限制数据特质, 但是需要事先提供一个极具判别力的度量空间; 基于行列式点过程的原型学习方法能够利用行列式约束原型的多样性, 但是优化目标单一, 忽略了原型的可解释性等其他质量指标; 最后, 基于低秩逼近的原型学习原理最为直观, 同时也容易实现, 但是关于矩阵运算的复杂度在一定程度上限制了该类方法在大规模数据集上的实施. 表 1 选取了无监督下常用的四大类原型学习方法作为对比参考, 因为其他监督方式下的代表性方法较少, 所以在此不多加讨论. 由表1可以看出, 目前绝大部分原型学习方法都没有灵活应对噪声的能力. 此外, 在原型学习时引入的约束越多, 原型的质量也越高, 分类效果越好. 但同时, 该算法的复杂度随之有所提高.

表 1 原型学习方法性能对比

方法	对噪声鲁棒性	时间复杂度(假设 $n>>k$)	原型质量		
			原型多样性	原型可解释性	原型相容性
DS3[36]	√	$O(n^3)$	√	√	–
K-Medoids[84]	×	$O(n^2)$	√	√	–
AP[39]	×	$O(n^2)$	√	√	–
ProSe[87]	√	$O(n^3)$	√	√	√
SDLA[74]	√	$O(n^3)$	√	√	√
DPP[70]	×	$O(n^3)$	√	–	√
k-DPPs[75]	×	$O(n^2)$	√	–	√
SMRS[64]	×	$O(n^3)$	–	√	–
IPM[52]	×	$O(n^2)$	–	√	–
SSDS[79]	×	$O(n^3)$	√	√	√
Sipl[80]	×	$O(n^3)$	√	√	–
MOSAIC[124]	√	$O(n^3)$	√	√	–
CUR[82]	×	$O(n^2)$	–	√	–

注: “√”表示该项目存在; “×”表示不存在; “–”表示目标函数未明确指出是否存在该特性, 因而不加以讨论

5.2 原型学习的未来方向

通过分析近几年原型学习领域的最新研究成果, 探究诸多应用背景对原型学习的具体需求, 本文对未来原型学习研究的理论与应用发展方向进行了一定的预测, 具体包括:

- 知识迁移驱动的原型生成

原型质量与原型学习模型可利用的信息量成正比关系. 显然, 提供的信息越多, 获取的原型越能更好地捕获目标数据的分布. 然而, 现有的原型学习只利用有限的目标数据且无任何监督信息, 这必然会弱化原型的可解释性. 为了充分利用大数据时代已标注的海量数据资源, 未来工作应沿着跨域引入辅助集的方向探索, 而非传统的监督式原型学习[119,120]. 但同时, 尽可能约简与目标集相关的假设和先验知识, 以增加其适用范围. 例如, 为了将训练好的复杂模型的“知识”迁移到一个结构更为简单的小型网络中, 进而方便模型部署, Hinton 等人曾提出了蒸馏网络模型[121]. 受该模型启发, 文献[122]提出了数据蒸馏的概念: 保持模型固定, 尝试将大型训练数据集中的知识提炼成小数据. 其中, 这些合成的少量数据不需要一定来自正确的数据分布, 但当作为模型的训练数据学习时, 能达到近似在原始数据上训练的效果. 未来, 是否可以将模型蒸馏和数据蒸馏进行融合以进行原型学习十分值得探讨. 其核心思想是: 利用一个额外的有标记的辅助数据集, 即使其语义标签集合与目标数据集的语义标签集合没有任何交集, 但通过模型知识蒸馏并迁移到目标数据集, 可以生成一个小规模的原型集. 这样不仅可以描述目标数据集的分布, 还可以泛化到其他数据集, 并对其进行分类、预测等多种推理任务. 以一个小样本学习的任务为例: 为了识别“犀牛”和“斑马”这两个物种, 专家分别标注了 5 个样本. 从如此极少量样本中快速学习一个分类模型, *k* 近邻分类器固然可行, 但是如若可以利用庞大

的已标注的数据集以及预训练模型(比如 ImageNet 数据集中的“黄牛”和“白马”)进行知识迁移, 以更加全面地学习“犀牛”和“斑马”的原型, 进而利用原型进行 k 近邻分类, 那么性能在一定程度会得以提升.

- 有缺陷数据的原型生成

对于给定的目标集数据, 通常在底层特征空间和高层语义空间容易出现缺陷. 对于前者, 现有的大多数原型学习算法通常假设数据点可以用低维子空间很好地逼近. 然而, 实际应用中的很多数据具有显著的噪声点、离群值和缺失值, 因此低维子空间(或者它们的并集)可能无法很好地拟合数据[36,123,124]. 这一事实要求原型学习算法同时具有鲁棒性, 能够在所有这些缺陷存在的情况下识别极具信息量的数据点. 然而, 现有的有缺陷数据的原型生成算法主要针对特定类型的噪声和异常值. 比如文献[123]提出一种可伸缩的列/行子空间追踪算法, 用于从含有任意幅度的稀疏噪声的数据中选择原型. 对于后者, 现有的原型学习方法基本上都是以无、自、半监督和全监督的方式实施, 并未考虑标签也容易存在缺陷. 比如, 数据只给出了粗粒度标签或者有噪声的标签. 对此, 弱监督的原型学习则需要引起极大的重视, 包括不确切监督和不精确监督, 以迎合实际应用场景的迫切需求.

- 原型分布式学习

鉴于隐私保护等安全考量, 目标数据极有可能来自不同的工作站且无法集中受理, 这也为原型学习带来极大的挑战. 随着目前硬件设备计算能力的快速升级以及分步式计算框架的逐渐成熟, 针对分布式环境下的原型学习研究是十分有潜力的一个方向, 具体的研究内容包括将经典的原型学习算法在分布式框架下实施, 亦或是设计全新的面向分步式系统的原型学习算法. 这一方向的优势在于: 通过最大化利用分步式计算的效能, 不仅保护各个工作站的数据信息, 同时还可以有效解决传统原型学习算法的准确率与效率无法同时满足的问题.

- 面向深度学习的原型学习

原型学习方法通过识别信息量最大的训练实例来提高大规模数据集机器学习的数据效率, 已经受到越来越多的关注. 然而, 由于它们依赖于需要学习的特征表示, 因此在深度学习中应用它们的成本可能高得令人望而却步. 随着深度学习算法的普及, 面向深度学习的原型学习研究显得越来越重要. 目前, 在相关领域内, 已经存在一些研究工作[17]着眼于如何通过一个小代理模型来执行原型学习, 从而大大提高计算效率. 因此, 设计一个与下游任务有关且泛化能力极强的代理模型也是十分有潜力的研究方向. 例如, 通过从目标模型中移除隐藏层, 使用更小的体系结构, 并训练更少的时间段, 这样创建的代理模型可以将训练速度提高一个数量级. 此外, 得益于深度学习已经取得的出色性能, 将原型学习模型与深度学习框架相结合, 也是未来发展的必然趋势[17,122,125,126].

- 原型的质量评价体系

目前, 几乎所有的原型学习算法都需要依赖其他各种应用(诸如视频摘要、聚类和分类检索)来证明它们的有效性. 换言之, 原型学习问题迫切需要一个统一的数据驱动的评价标准, 而非任务驱动的评价标准, 以精准度量原型的质量, 尤其是可解释性、代表性和多样性. 事实上, 原型选择的目的是通过采集最具信息量的样本子集, 形成对目标集的有效刻画. 而相对于非原型来说, 原型之所以具有很好的可解释性(代表性)的原因在于其具有很强的显著性. 目前, 得到广泛研究的通过自下而上的(bottom-up)方式来获得原型集的方法, 难以对选择出的原型的代表性给出定量的解释. 针对该问题, 未来拟通过图传递模型, 建立一种基于显著性采样的自上而下(top-down)原型选择机制, 从而有效地获得评价潜在原型显著性的置信度. 此外, 对于选择出来的原型, 未来还可以通过 Nyström 方法来计算样本矩阵恢复的质量, 进而以此定量评价选出原型的质量.

References:

[1] Naisbitt J, Wrote; Mei Y, Trans. Megatrends: Ten New Directions Transforming Our Lives. Beijing: China Federation of industry and Commerce Press, 2009. 91−91 (in Chinese).

[2] Gao Y, Ma JY, Zhao MB, Liu W, Yuille AL. NDDR-CNN: Layerwise feature fusing in multi-task CNNS by neural discriminative dimensionality reduction. In: Proc. of the Computer Vision and Pattern Recognition. 2019. 3205−3214.

[3] Esser E, Moller M, Osher S, Sapiro G, Xin J. A convex model for non-negative matrix factorization and dimensionality reduction on physical space. IEEE Trans. on Image Processing, 2012, 21(7): 3239−3252.

[4] Liu H, Motoda H. Computational Methods of Feature Selection. CRC Press, 2007.

[5] Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research, 2003, 3(Mar.): 1157−1182.

[6] Zhang XX, Zhu ZF, Zhao Y. Sparsity induced prototype learning via l_1-norm grouping. Journal of Visual Communication and Image Representation, 2018, 57: 192−201.

[7] Bien J, Tibshirani R. Prototype selection for interpretable classification. The Annals of Applied Statistics, 2011, 5: 2403−2424.

[8] http://khoury.neu.edu/home/eelhami/tutorial_cvpr2018.htm

[9] http://khoury.neu.edu/home/eelhami/tutorial_cvpr2016.htm

[10] Jiang WH. Research on sample selection and its applications in pattern recognition [Ph.D. Thesis]. Nanjing: Nanjing University of Science and Technology, 2008 (in Chinese with English abstract).

[11] Xiong L. Research on the data selection and learning algorithms under big data [Ph.D. Thesis]. Xi'an: Xidian University, 2014 (in Chinese with English abstract).

[12] Chen S, Zhang CS. Selecting informative universum sample for semi-supervised learning. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence. 2009. 1016−1021.

[13] Qian C, Shi JC, Yu Y, Tang K, Zhou ZH. Parallel Pareto optimization for subset selection. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence. 2016. 1939−1945.

[14] Wang SS, Zhang ZH. Efficient algorithms and error analysis for the modified Nyström method. In: Proc. of the Artificial Intelligence and Statistics. 2014. 996−1004.

[15] Yang HM, Zhang XY, Yin F, Liu CL. Robust classification with convolutional prototype learning. In: Computer Vision and Pattern Recognition. 2018. 3474−3482.

[16] Deng DY. Research on data reduction based on rough sets and extension of rough set models [Ph.D. Thesis]. Beijing: Beijing Jiaotong University, 2007 (in Chinese with English abstract).

[17] Coleman C, Yeh C, Mussmann S, Mirzasoleiman B, Bailis P, Liang P, Zaharia M. Selection via proxy: Efficient data selection for deep learning. In: Proc. of the Int'l Conf. on Learning Representations. 2020.

[18] Li YF, Zhou ZH. Improving semi-supervised support vector machines through unlabeled instances selection. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2011.

[19] Wei K, Liu YZ, Kirchhoff K, Bilmes J. Using document summarization techniques for speech data subset selection. In: Proc. of the 2013 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013. 721−726.

[20] Liu HP, Liu YH, Sun FC. Robust exemplar extraction using structured sparse coding. IEEE Trans. on Neural Networks and Learning Systems, 2014, 26(8): 1816−1821.

[21] Dornaika F, Aldine IK. Decremental sparse modeling representative selection for prototype selection. Pattern Recognition, 2015, 48(11): 3714−3727.

[22] Har-Peled S, Kushal A. Smaller coresets for *k*-median and *k*-means clustering. Discrete & Computational Geometry, 2007, 37(1): 3−19.

[23] Liu YN, Li JZ, Gao H. Research on core-sets selection on massive incomplete data. Chinese Journal of Computers, 2018, 41(4): 915−930 (in Chinese with English abstract).

[24] Takigawa I, Kudo M, Nakamura A. Convex sets as prototypes for classifying patterns. Engineering Applications of Artificial Intelligence, 2009, 22(1): 101−108.

[25] Rahmani M, Atia GK. Spatial random sampling: A structure-preserving data sketching tool. IEEE Signal Processing Letters, 2017, 24(9): 1398−1402.

[26] Paul S, Bappy JH, Roy-Chowdhury AK. Non-uniform subset selection for active learning in structured data. In: Computer Vision and Pattern Recognition. 2017. 6846−6855.

[27] Loosli G, Canu S, Bottou L. Training invariant support vector machines using selective sampling. Large Scale Kernel Machines, 2007, 2.

[28] Kaushal V, Iyer R, Kothawade S, Mahadev R, Doctor K, Ramakrishnan G. Learning from less data: A unified data subset selection and active learning framework for computer vision. In: Proc. of the IEEE Winter Conf. on Applications of Computer Vision. 2019. 1289−1299.

[29] Fu TF, Gao T, Xiao C, Ma TF, Sun JM. Pearl: Prototype learning via rule learning. In: Proc. of the ACM Int'l Conf. on Bioinformatics, Computational Biology and Health Informatics. 2019. 223−232.

[30] Zheng S, Zhu ZF, Zhang XX, Cheng J, Zhao Y. Distribution-induced bidirectional generative adversarial network for graph representation learning. In: Computer Vision and Pattern Recognition. 2020. 7224−7233.

[31] Cong Y, Liu J, Sun G, You QZ, Li YC, Luo JB. Adaptive greedy dictionary selection for Web media summarization. IEEE Trans. on Image Processing, 2016, 26(1): 185−195.

[32] Singh A, Virmani L, Subramanyam AV. Image corpus representative summarization. In: Proc. of the 5th IEEE Int'l Conf. on Multimedia Big Data. 2019. 21−29.

[33] Gillenwater J, Kulesza A, Fox E, Taskar B. Expectation maximization for learning determinantal point processes. In: Advances in Neural Information Processing Systems. 2014. 3149−3157.

[34] Elhamifar E, Kaluza MCDP. Subset selection and summarization in sequential data. In: Advances in Neural Information Processing Systems. 2017. 1035−1045.

[35] Wang DD, Zhu SH, Li T, Gong YH. Comparative document summarization via discriminative sentence selection. ACM Trans. on Knowledge Discovery from Data, 2013, 7(1): 1−18.

[36] Elhamifar E, Sapiro G, Sastry SS. Dissimilarity-based sparse subset selection. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2016, 38(11): 2182−2197.

[37] Wang HX, Yuan JS. Representative selection on a hypersphere. IEEE Signal Processing Letters, 2018, 25(11): 1660−1664.

[38] Cornuejols G, Fisher ML, Nemhauser GL. Exceptional paper-location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. Management Science, 1977, 23(8): 789−810.

[39] Frey B, Dueck D. Clustering by passing messages between data points. Science, 2007, 315(5814): 972−976.

[40] Misra I, Shrivastava A, Hebert M. Data-driven exemplar model selection. In: Proc. of the IEEE Winter Conf. on Applications of Computer Vision. 2014. 339−346.

[41] Gan M, Chen GY, Chen L, Chen CLP. Term selection for a class of separable nonlinear models. IEEE Trans. on Neural Networks and Learning Systems, 2020, 31(2): 445−451.

[42] Zhang XX, Zhu ZF, Zhao Y, Zhao YW. ProLFA: Representative prototype selection for local feature aggregation. Neurocomputing, 2020, 381: 336−347.

[43] Garcia S, Derrac J, Cano JR, Herrera F. Prototype selection for nearest neighbor classification: Taxonomy and empirical study. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000, 34(3): 417−435.

[44] Nandan M, Khargonekar PP, Talathi SS. Fast SVM training using approximate extreme points. Journal of Machine Learning Research, 2014, 15(1): 59−98.

[45] Liu ZZ, Zhang XX, Zhu ZF, Zheng S, Zhao Y. Convolutional prototype learning for zero-shot recognition. Image and Vision Computing, 2020, 98: 103924.

[46] Zhang XX, Gui SP, Zhu ZF, Zhao Y, Liu J. Hierarchical prototype learning for zero-shot recognition. IEEE Trans. on Multimedia, 2020, 22(7): 1692−1703.

[47] Pekalska E, Duin RP, Paclık P. Prototype selection for dissimilarity-based classifiers. Pattern Recognition, 2006, 39: 189−208.

[48] Dong NQ, Xing EP. Few-shot semantic segmentation with prototype learning. British Machine Vision Conf., 2018, 3(4).

[49] Gong BQ, Chao WL, Grauman K, Sha F. Diverse sequential subset selection for supervised video summarization. In: Proc. of the Advances in Neural Information Processing Systems. 2014. 2069−2077.

[50] Xia GY, Chen BJ, Sun HJ, Liu QS. Nonconvex low-rank kernel sparse subspace learning for keyframe extraction and motion segmentation. IEEE Trans. on Neural Networks and Learning Systems, 2020.

[51] Hartline J, Mirrokni V, Sundararajan M. Optimal marketing strategies over social networks. In: Proc. of the Int'l Conf. on World Wide Web. 2008. 189−198.

[52] Zaeemzadeh A, Joneidi M, Rahnavard N, Shah M. Iterative projection and matching: Finding structure-preserving representatives and its application to computer vision. In: Computer Vision and Pattern Recognition. 2019. 5414−5423.

[53] Joshi S, Boyd S. Sensor selection via convex optimization. IEEE Trans. on Signal Processing, 2008, 57(2): 451−462.

[54] Krause A, Singh A, Guestrin C. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. Journal of Machine Learning Research, 2008, 9(Feb.): 235−284.

[55] Hadlock F. Finding a maximum cut of a planar graph in polynomial time. SIAM Journal on Computing, 1975, 4(3): 221−225.

[56] Motwani R, Raghavan P. Randomized Algorithms. Cambridge: Cambridge University Press, 1995.

[57] Wang SC, Meng JJ, Yuan JS, Tan YP. Joint representative selection and feature learning: A semi-supervised approach. In: Computer Vision and Pattern Recognition. 2019. 6005−6013.

[58] Shah A, Ghahramani Z. Determinantal clustering process—A nonparametric Bayesian approach to kernel based semisupervised clustering. In: Proc. of the Uncertainty in Artificial Intelligence. 2013. 566−575.

[59] Sener O, Savarese S. Active learning for convolutional neural networks: A core-set approach. In: Proc. of the Int'l Conf. on Learning Representations. 2018.

[60] Yeh CK, Kim J, Yen IEH, Ravikumar PK. Representer point selection for explaining deep neural networks. In: Proc. of the Advances in Neural Information Processing Systems. 2018. 9291−9301.

[61] Xu C, Elhamifar E. Deep supervised summarization: Algorithm and application to learning instructions. In: Advances in Neural Information Processing Systems. 2019. 1107−1118.

[62] Kohonen T. The self-organizing map. Proc. of the IEEE, 1990, 78(9): 1464−1480.

[63] Liu CL, Nakagawa M. Evaluation of prototype learning algorithms for nearest-neighbor classifier in application to handwritten character recognition. Pattern Recognition, 2001, 34(3): 601−615.

[64] Elhamifar E, Sapiro G, Vidal R. See all by looking at a few: Sparse modeling for finding representative objects. In: Proc. of the Computer Vision and Pattern Recognition. 2012. 1600−1607.

[65] Zhang XX, Zhu Z, Zhao Y, Chang DX. Learning a general assignment model for video analytics. IEEE Trans. on Circuits and Systems for Video Technology, 2017, 28(10): 3066−3076.

[66] Charikar M, Guha S, Tardos E, Shmoys DB. A constant-factor approximation algorithm for the k-median problem. Journal of Computer and System Science, 2002, 65(1): 129−149.

[67] Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proc. of the Research and Development in Information Retrieval. 1998. 335−336.

[68] Meng JJ, Wang HX, Yuan JS, Tan YP. From keyframes to key objects: Video summarization by representative object proposal selection. In: Computer Vision and Pattern Recognition. 2016. 1039−1048.

[69] Cong Y, Yuan JS, Luo JB. Towards scalable summarization of consumer videos via sparse dictionary selection. IEEE Trans. on Multimedia, 2011, 14(1): 66−75.

[70] Kulesza A, Taskar B. Structured determinantal point processes. In: Advances in Neural Information Processing Systems. 2010. 1171−1179.

[71] Elhamifar E, Kaluza MCDP. Online summarization via submodular and convex optimization. In: Computer Vision and Pattern Recognition. 2017. 1783−1791.

[72] Das A, Kempe D. Approximate submodularity and its applications: Subset selection, sparse approximation and dictionary selection. Journal of Machine Learning Research, 2018, 19(1): 74−107.

[73] Schlegel M, Pan YC, Chen JC, White M. Adapting kernel representations online using submodular maximization. In: Proc. of the Int'l Conf. on Machine Learning. 2017. 3037−3046.

[74] Zhang XX, Zhu ZF, Zhao Y, Kong DQ. Self-supervised deep low-rank assignment model for prototype selection. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence. 2018. 3141−3147.

[75] Kulesza A, Taskar B. *k*-DPPS: Fixed-size determinantal point processes. In: Proc. of the Int'l Conf. on Machine Learning. 2011. 1193−1200.

[76] Kulesza A, Taskar B, *et al*. Determinantal point processes for machine learning. Foundations and Trends® in Machine Learning, 2012, 5(2−3): 123−286.

[77] Derezinski M, Warmuth MK, Hsu DJ. Leveraged volume sampling for linear regression. In: Advances in Neural Information Processing Systems. 2018. 2505−2514.

[78] Cong Y, Yuan JS, Liu J. Sparse reconstruction cost for abnormal event detection. In: Computer Vision and Pattern Recognition. 2011. 3449−3456.

[79] Wang HX, Kawahara Y, Weng CQ, Yuan JS. Representative selection with structured sparsity. Pattern Recognition, 2017, 63: 268−278.

[80] Zhang XX, Zhu ZF, Zhao Y. Sparsity induced prototype learning via l_1-norm grouping. Journal of Visual Communication and Image Representation, 2018, 57: 192−201.

[81] Yang CL, Shen JL, Peng JY, Fan JP. Image collection summarization via dictionary learning for sparse representation. Pattern Recognition, 2013, 46(3): 948−961.

[82] Bien J, Xu Y, Mahoney MW. CUR from a sparse optimization viewpoint. In: Advances in Neural Information Processing Systems. 2010. 217−225.

[83] Williams CK, Seeger M. Using the nyström method to speed up kernel machines. In: Advances in Neural Information Processing Systems. 2001. 682−688.

[84] Kaufman L, Rousseeuw P. Clustering by means of medoids. In: Proc. of the Statistical Data Analysis Based on the l_1-norm and Related Methods. 1987. 405−416.

[85] Drineas P, Mahoney MW. On the Nyström method for approximating a gram matrix for improved kernel-based learning. Journal of Machine Learning Research, 2005, 6(Dec.): 2153−2175.

[86] Li M, Bi W, Kwok JT, Lu BL. Large-scale Nyström kernel matrix approximation using randomized SVD. IEEE Trans. on Neural Networks and Learning Systems, 2014, 26(1): 152−164.

[87] Zhang XX, Zhu ZF, Zhao Y, Chang DX, Liu J. Seeing all from a few: l_1-norm-induced discriminative prototype selection. IEEE Trans. on Neural Networks and Learning Systems, 2019, 30(7): 1954−1966.

[88] Drineas P, Mahoney MW, Muthukrishnan S. Subspace sampling and relative-error matrix approximation: Column-based methods. In: Proc. of the Int'l Workshop on Approximation Algorithms for Combinatorial Optimization and the Int'l Workshop on Randomization and Approximation Techniques in Computer Science. Berlin, Heidelberg: Springer, 2006. 316−326.

[89] Boutsidis C, Sun J, Anerousis N. Clustered subset selection and its applications on it service metrics. In: Proc. of the ACM Conf. on Information and Knowledge Management. 2008. 599−608.

[90] Elhamifar E, Sapiro G, Vidal R. Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery. In: Advances in Neural Information Processing Systems. 2012. 19−27.

[91] Frey BJ, Dueck D. Mixture modeling by affinity propagation. In: Advances in Neural Information Processing Systems. 2006. 379−386.

[92] Lashkari D, Golland P. Convex clustering with exemplar-based models. In: Advances in Neural Information Processing Systems. 2008. 825−832.

[93] Yue Y, Joachims T. Predicting diverse subsets using structural SVMs. In: Proc. of the Int'l Conf. on Machine Learning. 2008. 1224−1231.

[94] Prasad A, Jegelka S, Batra D. Submodular meets structured: Finding diverse subsets in exponentially-large structured item sets. In: Advances in Neural Information Processing Systems. 2014. 2645−2653.

[95] Sánchez JS, Pla F, Ferri FJ. Prototype selection for the nearest neighbour rule through proximity graphs. Pattern Recognition Letters, 1997, 18(6): 507−513.

[96] Zhang K, Chao WL, Sha F, Grauman K. Summary transfer: Exemplar-based subset selection for video summarization. In: Computer Vision and Pattern Recognition. 2016. 1059−1067.

[97] Decaestecker C. Finding prototypes for nearest neighbour classification by means of gradient descent and deterministic annealing. Pattern Recognition, 1997, 30(2): 281−288.

[98] Lee YJ, Mangasarian OL. RSVM: Reduced support vector machines. In: Proc. of the Int'l Conf. on Data Mining. 2001. 1−17.

[99] Lee SW, Song HH. Optimal design of reference models for large-set handwritten character recognition. Pattern Recognition, 1994, 27(9): 1267−1274.

[100] Duda RO, Hart PE, Stork DG. Pattern Classification. John Wiley & Sons, 2012.

[101] Huang Z. Extensions to the *k*-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 1998, 2(3): 283−304.

[102] Elhamifar E. Sequential facility location: Approximate submodularity and greedy algorithm. In: Proc. of the Int'l Conf. on Machine Learning. 2019. 1784−1793.

[103] Elhamifar E, Naing Z. Unsupervised procedure learning via joint dynamic summarization. In: Proc. of the Int'l Conf. on Computer Vision. 2019. 6341−6350.

[104] Macchi O. The coincidence approach to stochastic point processes. Advances in Applied Probability, 1975, 7(1): 83−122.

[105] Sharghi A, Borji A, Li C, Yang TB, Gong BQ. Improving sequential determinantal point processes for supervised video summarization. In: Proc. of the European Conf. on Computer Vision. 2018. 517−533.

[106] Gartrell M, Brunel VE, Dohmatob E, Krichene S. Learning nonsymmetric determinantal point processes. In: Advances in Neural Information Processing Systems. 2019. 6715−6725.

[107] Gillenwater J, Kulesza A, Taskar B. Discovering diverse and salient threads in document collections. In: Proc. of the Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012. 710−720.

[108] Wu BY, Jia F, Liu W, Ghanem B. Diverse image annotation. In: Computer Vision and Pattern Recognition. 2017. 2559−2567.

[109] Chan TF. Rank revealing QR factorizations. Linear Algebra and Its Applications, 1987, 88: 67−82.

[110] Deshpande A, Rademacher L. Efficient volume sampling for row/column subset selection. In: Proc. of the Annual Symp. on Foundations of Computer Science. 2010. 329−338.

[111] Farahat AK, Elgohary A, Ghodsi A, Kamel MS. Greedy column subset selection for large-scale data sets. Knowledge and Information Systems, 2015, 45(1): 1−34.

[112] Wang YN, Singh A. Provably correct algorithms for matrix column subset selection with selectively sampled data. Journal of Machine Learning Research, 2017, 18(1): 5699−5740.

[113] Smola AJ, Schökopf B. Sparse greedy matrix approximation for machine learning. In: Proc. of the Int'l Conf. on Machine Learning. 2000. 911−918.

[114] Ouimet M, Bengio Y. Greedy spectral embedding. In: Proc. of the Int'l Conf. on Artificial Intelligence and Statistics. 2005.

[115] Farahat A, Ghodsi A, Kamel M. A novel greedy algorithm for Nyström approximation. In: Proc. of the Int'l Conf. on Artificial Intelligence and Statistics. 2011. 269−277.

[116] Gittens A, Mahoney MW. Revisiting the Nyström method for improved large-scale machine learning. Journal of Machine Learning Research. 2016, 17(1): 3977−4041.

[117] Kumar S, Mohri M, Talwalkar A. Sampling methods for the Nyström method. Journal of Machine Learning Research, 2012, 13(Apr.): 981−1006.

[118] Kumar S, Mohri M, Talwalkar A. Sampling techniques for the nystrom method. In: Proc. of the Artificial Intelligence and Statistics. 2009. 304−311.

[119] Zhang K, Kwok JT, Parvin B. Prototype vector machine for large scale semi-supervised learning. In: Proc. of the Int'l Conf. on Machine Learning. 2009. 1233−1240.

[120] Rudi A, Camoriano R, Rosasco L. Less is more: Nyström computational regularization. In: Advances in Neural Information Processing Systems. 2015. 1657−1665.

[121] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv: 1503.02531, 2015.

[122] Wang T, Zhu JY, Torralba A, Efros AA. Dataset distillation. arXiv preprint arXiv: 1811.10959, 2018.

[123] Rahmani M, Atia G. Robust and scalable column/row sampling from corrupted big data. In: Proc. of the Int'l Conf. on Computer Vision Workshops. 2017. 1818−1826.

[124] Sedghi M, Geo M, Atia G. A multi-criteria approach for fast and robust representative selection from manifolds. IEEE Trans. on Knowledge and Data Engineering, 2020. [doi: 10.1109/TKDE.2020.3024099]

[125] Kirsch A, van Amersfoort J, Gal Y. Batchbald: Efficient and diverse batch acquisition for deep Bayesian active learning. In: Advances in Neural Information Processing Systems. 2019. 7026−7037.

[126] Xie PT, Salakhutdinov R, Mou LT, Xing EP. Deep determinantal point process for large-scale multi-label classification. In: Proc. of the Int'l Conf. on Computer Vision. 2017. 473−482.

附中文参考文献:

[1] Naisbitt J, 著; 梅艳, 译. 大趋势: 改变我们生活的十个新方向. 北京: 中华工商联合出版社, 2009. 91.

[10] 姜文瀚. 模式识别中的样本选择研究及其应用 [博士学位论文]. 南京: 南京理工大学, 2008.

[11] 熊霖. 大数据下的数据选择与学习算法研究 [博士学位论文]. 西安: 西安电子科技大学, 2014.

[16] 邓大勇. 基于粗糙集的数据约简及粗糙集扩展模型的研究 [博士学位论文]. 北京: 北京交通大学, 2007.

[23] 刘永楠, 李建中, 高宏. 海量不完整数据的核心数据选择问题的研究. 计算机学报, 2018, 41(4): 915−930.

张幸幸(1993－), 女, 博士, 主要研究领域为机器学习, 包括原型学习和小样本学习.

赵亚威(1991－), 男, 博士, CCF 专业会员, 主要研究领域为机器学习与最优化理论分析.

朱振峰(1974－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为图像/视频分析与理解, 机器学习, 网络信息挖掘与分析.

赵耀(1967－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为数字媒体信息处理, 包括图像与视频编码、数字水印与数字取证、图像/视频分析与内容理解.