机器学习简介

俞扬

南京大学计算机软件新技术国家重点实验室

学习能力是人类智能的一种体现，对于一个计算机系统，如果缺乏类似的“学习”能力，很难被认为是具有智能的。机器学习是人工智能的核心研究领域之一，一个广泛接受的定义是，机器学习是研究如何“利用经验来改善计算机系统自身的性能”的学科[17]。机器学习可以被划分为“机械学习”、“示教学习”、“类比学习”和“归纳学习”四种类型[2]。其中，归纳学习的目标是从个例数据中进行抽象、发现个例背后的规律。随着计算机计算能力、通讯能力和存储能力的快速发展，人们收集数据的能力得到了显著的增强。人们对利用数据需求的增加与归纳学习的发展相互促进，使得自20世纪80年代以来，归纳学习成为机器学习中被研究得最多、应用最广的分支。

在归纳学习中处理的数据，通常是对象的特征的描述。例如苹果可以用大小、质量、色泽等特征描述为一个特征向量。常见的归纳学习任务是要从示例样本的特征以及给出的对应概念标记数据中进行学习，发现特征与概念标记之间的关系。例如对于苹果，我们可以指定概念标记为“成熟”与“不成熟”，并给出一些示例样本，包含一部分成熟苹果的描述和一部分不成熟苹果的描述，并且每一条描述与“成熟”或“不成熟”概念关联。归纳学习从有限的示例样本中学习，得到的特征向量与概念标记之间的关系并不是随意的，而是对于尚未观察到的示例上，这个关系也要尽可能的成立，也就是说归纳学习的目标是得到可以泛化的特征与概念标记间的关系。也正是由于归纳学习的泛化能力，使得学习的结果可以用于未见示例的预测，从而一定程度上满足了人们对于数据利用的需求。本文对归纳学习的框架和应用进行一个简单的介绍。

根据样本与概念标记之间的关系不同，可以将归纳学习进一步划分为不同的设定框架。传统对归纳学习设定框架的划分包括监督学习、非监督学习和强化学习三种[23]。他们之间的区别被认为是概念标记的显示程度上的区别：在监督学习中概念标记对应于每一个特征向量，在非监督学习中只有特征向量而没有概念标记（或者可认为概念标记不可见），强化学习则介于他们两者之间，即概念标记仅在一系列的行为后才能获得。而在本世纪内，源于实际应用环境的需求，新兴的更贴近应用需求的归纳学习框架得到了发展，包括半监督学习、多示例学习、多标记学习等。

半监督学习问题来源于现实应用中，收集特征数据的代价很低，而将特征向量关联到概念标记的代价却很高这一情况[22][26]。例如，可以很容易收集上百万条互联网网页的内容，但是要标记网页的类别，却需要花费大量人力逐条标记。在这样的情况下，我们面对的数据有一小部分包含了概念标记，即监督信息，而绝大部分则没有监督信息，因此这样的问题框

架被称为半监督学习。如果使用传统的监督学习方法，那么只能利用少量的有监督信息的数据，却无法利用大量没有监督信息的样本。使用半监督学习方法，我们可以在标记代价保持在可以接受的范围下，利用唾手可得的未标记数据来获得更好泛化能力的学习器。

半监督学习环境示例

多示例学习问题，是研究者们在对药物活性预测问题的研究时，为了更好的表达对象而提出的[5][27]。在多示例框架中，对个例对象不在只使用一个特征向量来描述，而是使用一组特征向量来描述。例如，在药物活性预测问题上，一个药物分子可以有很多个“低能量形态”，每一个形态用一个特征向量来描述，因此一个药物分子用一“包”特征向量来描述。同时，一个包对应了一个概念标记，例如药物分子的标记是“有效”或者“无效”。与使用单个特征向量的描述相比，多示例描述具有更好的表达能力，能够将对象的性质更有效的展现出来，同时概念标记不再对应到单个特征向量，而是提升到了包的层次。使用多示例学习方法，可以更好的表达对象丰富的语义信息，有助于获得更好泛化能力的学习器。

一副图像的多示例表示

多标记学习问题源于一个对象往往同时具有多个方面的概念标记[16][25]，例如一篇报道奥运会开幕式新闻文章可能同时具有奥运、运动员、国家元首等等标记。与传统的监督学习相比，多标记学习面临的是标记过剩的情况。多标记学习问题的一个最简单的处理方式是将标记分离，对每一个标记，使用传统的单标记学习方法来处理。然而这样的简单做法，并没有充分利用这些标记信息，同时也会面临逐一预测标记时计算开销过大的问题。多标记学习方法利用标记之间的关系，能够获得更好的泛化能力[20]，也有利于减少计算开销[8]。

一幅图像的多标记表示

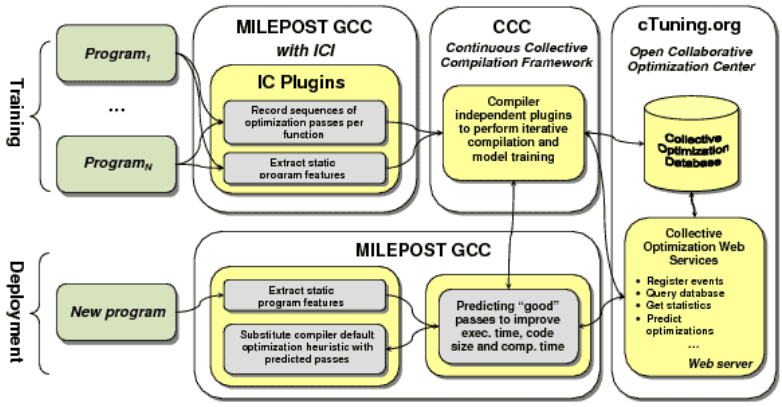
除了学习框架外向更贴近应用需求发展外，学习结果的评价也更趋向符合应用需求。学习器在处理真实世界中的数据时，由于噪音等因素，往往难以做到完美无缺。传统对分类器的性能评价是以分类错误为指标，即有多少比例的示例被分错类别。在具体问题中，人们通常都会评估一个错误的代价，而往往不同的错误具有不同的代价。例如，在诊断病人时，将癌症诊断为正常的代价，会比将正常诊断为癌症的代价高许多。**代价敏感学习**的目标并不是简单的使得错误的数量最小，而是使得总的误分类代价最小，这样的决策更符合人们的需求[21]。

机器学习除了用在通常被认为需要“智能”的领域，例如数据挖掘[24]、自然语言处理[3]、信贷决策[4]、医学诊断[13]、生物信息学[12]、电力监控[15]、网络入侵检测[19]、天气预报[11]、工业控制[18]、飞行控制[1]，近年来也渗透到计算机的“底层”领域中。下面我们简单介绍机器学习在编译器、软件工程、和处理器设计三个方面的应用。

在计算机芯片的设计中，由于芯片设计包含了大量的参数，例如缓存大小、队列长度等等，使得设计方案的空间非常大，而对每一个设计方案，需要通过多达数周的时间来进行评价。因此几乎不可能通过穷举每一种可能的设计方案来找到最佳方案。为了解决这个问题，机器学习方法被用来对方案—性能关系建模[9]，然后对于一个设计方案，可以直接预测出其性能，从而帮助寻找更好的方案，免去传统评价方法的耗时。归纳学习方法离不开示例样本，由于传统评价方式极其耗时，使得在有限时间内收集到的标记样例数量有限，使得学习结果的精度难以提高。最近的一项芯片设计空间探索工作[7]使用了半监督学习方法，在标记样本有限的情况下利用未标记样本来提高预测准确度，较以往方法获得了30%到84%的性能提升。

编译器是计算机软件的基础部件。随着芯片的不断升级换代，不同的程序在不同芯片上的性能要求对编译器配置提出了挑战。为了处理各种可能的需求，编译器提供了许多配置参数，例如常用的GCC编译器提供了近100个参数，不同的参数对于编译出的程序的运行时间、程序大小等会有影响，但是在一种处理器上对每一个程序手工调配参数是几乎不可能的。为了优化编译器配置，机器学习方法被用来预测给定程序所对应的配置。实际上，Milepost GCC项目就是一个使用机器学习方法来优化编译器配置的开源编译器[6]。对于一个程序源码，提取例如一个方法中的基本块数量等特征向量。训练集是一些已知最优配置的程序的特征向

量。对于一个新的程序，Milepost GCC预测出需要使用的编译配置。报告称在ARC处理器上经过Milepost GCC的编译，MiBench标准程序集中的程序运行时间平均提高了11% [6]。



Milepost GCC结构图（来自ctuning.org）

在软件工程领域，软件的缺陷发现是一个重要问题。通常软件的缺陷是通过软件测试以及软件验证等方法来发现。通过从代码中提取特征向量，并给出一定的软件缺陷标记，研究者们已经尝试使用机器学习方法可以对新的代码缺陷进行预测，一方面可以快速的发现缺陷，另一方面还可以帮助定位缺陷的位置。研究者们发现，即使对于软件陆续升级改进的版本，基于机器学习的方法都能有效的发现缺陷[10]。最近的一项软件缺陷预测工作中，利用主动学习和半监督学习方法，研究者减少了对标记样例数量的需求，对于一个新的软件项目，只需要人工检测少数几个模块，就可以构建较好的学习器进行缺陷预测[14]。

机器学习是目前计算机科学中最活跃的研究分支之一，值得一提的是，2010 年图灵奖得主美国 Harvard 大学 L. Valiant 教授和 2011 年图灵奖得主美国 UCLA 大学 J. Pearl 教授都是机器学习领域的学者。可以预见，机器学习技术将被应用到越来越多的领域，为研究者提供新的思路，还将给应用者带来更多的回报。

参考文献

[1] P. Abbeel, A. Coates, A. Y. Ng, and M. Quigley. An application of reinforcement learning to aerobatic helicopter flight. In B. Schölkopf, J. C. Platt, T. Hoffman eds., *Advances in Neural Information Processing Systems 19*, 2005, 1－8.

[2] P. R. Cohen and E. A. Feigenbaum, eds. The Handbook of Artificial Intelligence, vol.3, New York, NY: William Kaufmann, 1983.

[3] W. Daelemans and W. Hoste. Evaluation of machine learning methods for natural language processing tasks. In: *Proceedings of the 3rd International Conference on*

Language Resources and Evaluation, Canary Islands, Spain, 2002, 755–760.

[4] R. H. Davis, D. B. Edelman, and A. J. Gammerman. Machine-learning algorithms for credit-card applications. *Journal of Mathematics Applied in Business and Industry*, 1992, 4, 43–51.

[5] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple-instance problem with axis-parallel rectangles. Artificial Intelligence, 1997, 89(1–2): 31–71.

[6] G.Fursin, Y. Kashnikov, A. Wahid Memon, Z. Chamski, and O. Temam, et al. Milepost GCC: Machine learning enabled self-tuning compiler. *International Journal of Parallel Programming*. 2011, 39(3):296–327.

[7] Q. Guo, T. Chen, Y. Chen, Z.-H. Zhou, W. Hu, and Z. Xu. Effective and efficient microprocessor design space exploration using unlabeled design configurations. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, Barcelona, Spain, 2011, pages 1559–1564.

[8] D. Hsu, S. Kakade, J. Langford, and T. Zhang: Multi-Label Prediction via Compressed Sensing. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. Williams, and A. Culotta eds., Advances in Neural Information Processing Systems 22, 2009, pp. 772–780

[9] P.J. Joseph, V. Kapil, and M.J. Thazhuthaveetil. Construction and use of linear regression models for processor performance analysis. In *Proceedings of the 12th International Symposium on High Performance Computer Architecture*, 2006, pages 99–108.

[10] T. M. Khoshgoftaar, E. B. Allen, W. D. Jones, and J. P. Hudepohl. Classification-tree models of software-quality over multiple releases. *IEEE Transactions on Reliability*, 2000. 49(1):4–11.

[11] V. M. Krasnopolsky, and M. S. Fox-Rabinovitz. Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. *Neural Networks*, 2006, 19(2): 122–134.

[12] Y.-X. Li, S. Ji, S. Kumar, J. Ye, and Z.-H. Zhou. Drosophila gene expression pattern annotation through multi-instance multi-label learning. *ACM/IEEE Transactions on Computational Biology and Bioinformatics*, 2012, 9(1): 98–112.

[13] M. Li and Z.-H. Zhou. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. *IEEE Transactions on Systems, Man and Cybernetics – Part A: Systems and Humans*, 2007, 37(6): 1088–1098.

[14] M. Li, H. Zhang, R. Wu, and Z.-H. Zhou. Sample-based software defect prediction with active and semi-supervised learning. Automated Software Engineering, 2012, 19(2): 201－230.

[15] Y. Lu, M. A. Masrur, Z. H. Chen, and B. F. Zhang. Model-based fault diagnosis in electric drives using machine learning. *Transactions On Mechatronics*, 2006, 11(3): 290－303.

[16] A. McCallum. Multi-label text classification with a mixture model trained by EM. In *Working Notes of the AAAI'99 Workshop on Text Learning*.

[17] T. M. Mitchell TM. *Machine Learning*. New York: McGraw-Hill, 1997.

[18] C. Sammut. Automatic construction of reactive control systems using symbolic machine learning. *Knowledge Engineering Review*, 1996, 11, 27－42.

[19] C. Sinclair, L. Pierce, and S. Matzner. An application of machine learning to network intrusion detection. In: *Proceedings of 15th Computer Security Applications Conference*, 1999: 371－377.

[20] M.-L. Zhang and Z.-H. Zhou. Multi-label learning by instance differentiation. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, Vancouver, Canada, 2007, pp.669－674.

[21] Z.-H. Zhou and X.-Y. Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 2006, 18(1): 63-77.

[22] X. Zhu, and A. Goldberg. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3, Morgan & Claypool Publishers, 2009, pp. 1－130.

[23] 王珏. 关于机器学习的讨论. 见: 王珏, 周志华, 周傲英 主编, *机器学习及其应用*, 北京: 清华大学出版社, 2006. 1－31.

[24] 张敏灵, 周志华. 多标记学习. 见: 周志华, 杨强 主编, *机器学习及其应用* 2011, 北京: 清华大学出版社, 2011, 179－199.

[25] 周志华. 机器学习与数据挖掘. *中国计算机学会通讯*, 2007, 3(12): 35－44.

[26] 周志华. 半监督学习中的协同训练风范. 见: 周志华, 王珏 主编, *机器学习及其应用* 2007, 北京: 清华大学出版社, 2007, 259－275.

[27] 周志华. 多示例学习. 见: 刘大有 主编, *知识科学中的基本问题研究*, 北京: 清华大学出版社, 2006, 322－336.