DOI: 10.3969/j.issn.1009-6868.2015.05.009
网络出版地址：http://www.cnki.net/kcms/detail/34.1228.tn.20150923.0923.002.html

半监督多视图学习在大数据分析中的应用探讨

Semi-Supervised Multi-View Learning in Big Data

中图分类号：TN929.1 文献标志码：A 文章编号：1009-6868 (2015) 05-0032-003

摘要： 半监督多视图学习是机器学习领域一种极具潜力的大数据处理和分析方法，该方法能有效处理异构和半监督数据，并能方便地在线化和并行化，适合处理海量数据。该方法在大数据时代的应用前景值得研究人员和业界关注。指出未来需要通过引入其他领域新的研究技术和成果，不断丰富和完善半监督多视图学习的理论体系和算法设计，并在实验和实践中不断检验和探索。

关键词： 半监督；多视图；大数据；并行化

Abstract: This paper introduces a promising machine-learning paradigm called semi-supervised multi-view learning. With this paradigm, information is extracted from heterogeneous and semi-supervised data sets. Lately, multi-view learning has been scaled up online and through parallelization to deal with emerging big data challenges. Due to its successful application in many research domains and the fact that it has been explored and used by leading companies, multi-view learning may have a future in the big-data era as a major data analytic technique. New research techniques should be introduced into this area to improve the theoretical system and algorithm design of semi-supervised multi-view learning.

Key words: semi-supervised; multi-view; big data; parallelization

蓝超/LAN Chao[1]
饶泓/RAO Hong[2]
浣军/HUAN Jun[1]

(1.堪萨斯大学工程学院，堪萨斯 66045；
2. 南昌大学信息工程学院，江西 南昌 330031)
(1.College of Engineering, University of Kansas, Kansas 66045, USA;
2. College of Information Engineering, Nanchang University, Nanchang 330031, China)

- 如何有效利用无监督数据，是半监督学习研究的核心问题
- 提高无监督数据的样本复杂度的效率是值得关注的问题
- 多视图学习的样本复杂度分析主要集中于主动学习

机器学习已成为产业界大数据分析的主流工具。在2015年北京全球软件开发大会上，基于大数据的机器学习和数据挖掘专题讨论得到了业界广泛参与，如百度、搜狗、阿里巴巴、京东、美团、猿题库等著名企业参与了该专题的讨论。当前，机器学习在产业界的推广正处于白热化阶段。与此同时，如何结合大数据的特点，选择恰当的学习方法仍是值得探讨的问题。

从数据角度而言，大数据之“大”并不仅限于容量，也指数据类型多和收发速度快，三者统称为大数据的“3V”特征[1]。因此，若想迅速有效地处理和分析大数据，不仅需要从数据库和计算机体系结构等入手（如IBM的Hadoop，微软的大数据生态系统），更需选择合适的数据分析方法，才能充分挖掘海量数据中潜在的信息和价值。

除了3V，大数据普遍存在却鲜有讨论的另一特征是半监督性。在传统机器学习中，用于训练分类系统的数据大多是有监督的，即数据的实际分类已知。然而，许多应用问题如自然语言处理，大部分数据都不知道实际分类，这类数据称为无监督数据。如何有效利用无监督数据，使其帮助有监督数据一起训练分类系统，是半监督学习[2]研究的核心问题。随着大数据时代的到来，大量无监督数据将被采集，导致越来越多应用问题演变成大规模的半监督学习问题。

大规模半监督学习在学术界的探讨可追述到2005年Delalleau提出的高效无参归纳法[3]。在该论文中，作者提出将用于构造核矩阵的（无监督）数据集进行下采样，减小核矩阵大小，从而提高算法的存储和计算效率。在此思想基础上衍生出来的工作有很多，其中以大规模化基于核或基于图的半监督学习方法为主。除了数据下采样，也有研究员通过构造混合分布模型以减小参数估计的计算复杂度。此外，半监督学习的大规

收稿日期：2015-08-27
网络出版时间：2015-09-23

模化也可通过提高优化算法效率或并行程度来实现。

尽管有着近十年的研究历史，目前大部分的大规模半监督学习方法仍具有局限性。一方面，它们大都基于传统的单视图机器学习方法，忽略了大数据的异构性，容易导致信息丢失；另一方面，规模化算法的策略较为传统，如下采样或加速优化，无法有效满足大数据带来的新的挑战，如数据的在线化和分布化等。

文章结合大数据的特点介绍机器学习领域的一种新兴的半监督学习方法——多视图学习。它不仅在许多研究领域已得到成功应用，更因具备有效利用无监督异构数据的能力和迅速发展的大规模化算法，有望成为大数据时代最具潜力的数据分析工具之一。

1 多视图学习优点多

1.1 多视图学习及其优点

多视图学习[4]是指专门针对多视图数据而进行建模和学习。其中，多视图数据是指由多组（往往具有不同意义的）特征进行描述的数据，而每一组特征称为一个视图。多视图学习的主要思想是基于无监督数据的视图一致性，即分类器在同一无监督数据不同视图下的分类结果应基本一致。将此约束加入学习法则，多视图学习便能巧妙利用无监督数据帮助分类器的训练。

多视图学习的一个显著优点是缓解过学习问题，即由于模型过于复杂而将数据噪声也学入分类法则的现象。传统的机器学习方法大多将数据的所有视图堆砌成一个高维的单视图数据，进行建模和学习。此时，如果视图间存在冗余信息，为高维数据所建的模型将比实际需求更为复杂，容易导致过学习。而多视图学习则为每个视图分别建模，有效降低了模型复杂度。

多视图学习的另一优点是提升数据的总体分类能力。当数据特征所蕴含的分类信息总体较弱时，可将特征集拆分成多个视图进行多视图学习[5]。通过各视图的弱分类器协同训练，达到各视图“单独学习弱，集成学习强”的目的。

另外，多视图学习还能有效处理异构数据[6]。大数据时代，数据的异构性越来越强。比如，客服中心为了提高工作效率，需要根据来电客户的个人信息和语音信息对其来电目的进行快速预测。这里，用户的个人信息是静态文本数据，而语音信息则是动态时序数据，两者不仅数据类型不同，也往往服从不同的后验（预测）概率分布。强行统一或堆砌两类特征不仅给建模带来困难，也容易导致信息丢失，失去大数据分析的优势。多视图学习则允许各视图分别选择合适的分类器，再通过协同训练提高视图整体的分类能力。

1.2 多视图学习的理论研究

多视图学习的理论分析主要基于其分类器的泛化误差。一个分类器的泛化误差是其在指定数据分布下，错分数据的概率。许多理论研究证明：多视图中的协同学习方法在满足条件时，分类器的泛化误差将以极大的概率收敛到极小的范围之内。

多视图学习的早期理论分析基于两个假设[7]：视图充分性，即每个视图能分别提供充分的（但不必很强）分类信息。该假设在大数据中较容易被满足；各视图间条件独立，即给定数据分类，其各视图间统计独立。

多视图学习另一个理论研究是其样本复杂度。样本复杂度指通过多少数据的训练就能保证分类器收敛到预期的泛化误差。目前，多视图学习的样本复杂度分析主要集中于主动学习，即选取哪些无监督数据进行人工分类，才能使分类器的训练最快收敛。在此问题中，有监督数据的样本复杂度被证明与泛化误差的倒数成 log 比[8]，而无监督数据样本复杂度则与泛化误差的倒数成正比[9]。在大数据时代，无监督数据的容量常常远大于有监督数据，是计算机的存储和计算的主要负担。因此，提高无监督数据的样本复杂度的效率是值得关注的问题。

1.3 多视图学习的常见算法及大规模化算法

大部分多视图算法为每个视图分别建立分类器，基于协同训练方式的不同，有两种主要的多视图算法：一是基于迭代的协同训练算法[7]，另一个是基于协同正则化的算法[10]。基于迭代的协同训练算法的每个回合由一个视图的分类器标注一部分无监督数据的类别，并将它们加入有监督数据集，一起重新训练其他视图的分类器，以达到视图一致，提高视图总体的分类能力。这类算法直观有效，也较容易处理异构数据，但因其迭代的训练方式，学习效率相对较低；基于协同正则化的算法为所有视图的分类器统一建立一个目标函数，通过特定约束条件达到视图一致，并一次性地训练完所有分类器。该类方法避免了迭代训练，计算效率往往更高。

为各视图分别建模能最大程度保留视图的个性特征，但对视图一致性的要求也更高。当一致性不能被较好满足时，该类算法的效果会有所下降[11]。另一类多视图学习算法则侧重于将多个视图进行有机融合成单视图，再进行单视图学习[12]。这类算法有利于提高特征的鲁棒性和泛化能力，不同于简单的堆砌视图，但也具有丢失视图个性特征的风险。为此，有学者提出同时学习视图的共性特征和个性特征[13]。

除了改良传统算法，多视图学习也正沿着在线化和并行化的方向发展，以迎接大数据的新挑战。在线多视图学习假设训练数据分批，依次呈现给分类器进行学习。此时，如何为分类器设计高效的更新算法，是在线

多视图学习的研究问题。目前研究的主流算法是基于协同正则化的算法，而其在线化的主要策略是通过引入随机梯度下降法，实现目标函数的在线优化[14]。该策略的优点是计算速度快，存储空间小，但对学习参数的敏感性较高。

目前的并行多视图学习研究主要基于MapReduce的框架。如爱立信研究院提出[15]将不同数据的不同视图被分配到不同计算单元进行计算，以提高计算效率。但由于视图一致性的要求，不同计算单元间往往需要频繁通信，这成为系统效率的决定性瓶颈，也是当前分布式机器学习的一个研究热点。并行或分布式多视图学习中通信量如何增长？有哪些影响因素？如何设计低通信量的高效算法？这些都是大规模多视图学习需要探索的问题。

2 多视图学习应用广

多视图学习在许多研究领域都已获得成功应用，包括网页分类、自然语言处理，计算机视觉、医疗诊断、药物分析和化学分析。在网页分类中，每个网页的内容和链表可表示为两个视图；在文本分类中，同一文本的不同语言版本可表示为不同视图；在用户决策识别系统中，用户的语音和姿势可表示成两个视图；在图像检索和标识中，图像本身和周边的文本信息可表示成两种视图。即使在一些数据并不自然展示出多个视图的问题中，也可通过从单视图中提取多个视图进行多视图学习，以充分发挥多视图学习的优势。

多视图学习在工业界也被积极探索和推广，涉及领域包括机器翻译、情感预测、图像检索和推荐系统等。在微软研究的机器翻译中[16]，测试文章在不同解码器下的译文被视作不同视图下的准参考译文加入训练集参与翻译器的特征权重的训练，从而使翻译器更全面的适应测试集领域的文章的特点，达到领域自适应的目的。

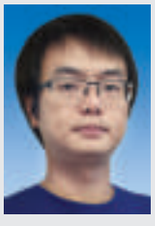
3 结束语

半监督多视图学习是大数据时代极具潜力的分析工具。其在许多研究领域已获得成功应用，并在工业界被积极推广。但同时也需指出，大数据的大容量、分布式和在线化等特征为半监督多视图学习带来了新的挑战。比如，如何有效降低无监督数据的样本复杂度，如何降低分布式视图间的通信量等。要解决这些问题，需要通过引入其它领域新的研究技术和成果，不断丰富和完善半监督多视图学习的理论体系和算法设计，并在实验和实践中不断检验和探索。

参考文献

[1] LABRINIDIS A, and JAGADISH H V. Challenges and Opportunities with Big Data: A community white paper developed by leading researchers across the United States [R]. 2012
[2] ZHU X J. Semi–supervised learning literature survey [EB/OL]. http://pages.cs.wisc.edu/~jerryzhu/research/ssl/semireview.html
[3] DELALLEAU O, YOSHUA B and ROUX N. Efficient non–parametric function induction in semi– supervised learning [C]// Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, Bridgetown, Barbados, 2005: 96–103
[4] XU C, TAO D C, and CHAO C. A Survey on Multi–view Learning [Z]. arXiv preprint: 1304.5634, 2013
[5] CHEN M M, CHEN Y X and WEINBERGER K Q. Automatic feature decomposition for single view co–training [C]//Proceedings of the 28th International Conference on Machine Learning, Washington, USA, 2011: 953–960
[6] LIAN W Z, RAI P, SALAZAR E and CARIN L. Integrating Features and Similarities: Flexible Models for Heterogeneous Multiview Data [C]//Twenty–Ninth AAAI Conference on Artificial Intelligence, Austin Texas, USA, 2015
[7] BLUN A and MITCHELL T. Combining labeled and unlabeled data with co–training [C]//Proceedings of the eleventh annual conference on Computational learning theory, Madison, USA, 1998: 92–100
[8] WANG W and ZHOU Z H. On multi–view active learning and the combination with semi–supervised learning[C]//Proceedings of the 25th international conference on Machine learning, Helsinki, Finland, 2008: 1152–1159
[9] BALCAN M F and BLUM A. A Discriminative Model for Semi–Supervised Learning [J]. Journal of the ACM (JACM), 2010,57(3):19. doi: 10.1145/1706591.1706599
[10] SINDHWANI V, NIYOGI P and BELKIN M. A co–regularization approach to semi–supervised learning with multiple views[C]// Proceedings of ICML workshop on learning with multiple views, Bonn, Germany, 2005: 74–79
[11] CHRISTOUDIAS C, URTASUN R and DARRELL T, Multiview leanringin the presence of view disagreement [C]// Conference on Uncertainty in Artificial Intelligence, California, USA, 2012
[12] CHEN N, ZHU J and XING E P, Predictive subspace learning for multi–view data: a large margin approach[C]//Advances in neural information processing systems, Hyatt Regency, Vancouver Canada, 2010
[13] JING X Y, HU R M, ZHU Y P, Wu S S, LIANG C. and YANG J Y. Intra–View and Inter–View Supervised Correlation Analysis for Multi–View Feature Learning[C]// Twenty–Eighth AAAI Conference on Artificial Intelligence, Qu é bec City, Qu é bec, Canada, 2014
[14] Ruijter T D, Tsivtsivadze E and Heskes T. Online co–regularized algorithms [M]. Germany: Springer Berlin Heidelberg, 2012
[15] HARIHARAN C. and SUBRAMANIAN S. Large scale multiview learning on mapreduce[C]// 19th International Conference on Advanced Computing and Communications, Chennai, India, 2013
[16] MAEIREZO B, LITMAN D and HWA R. Co–training for predicting emotions with spoken dialogue data[C]// Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, Barcelona, Spain, 2004

作者简介



蓝超，美国堪萨斯大学计算机系在读博士；研究方向为机器学习和模式识别；在国际会议及期刊上发表论文10余篇。

饶泓，南昌大学信息工程学院教授；主要研究方向为机器学习、数据挖掘、智能信息处理；主持多项国家及省市科技项目；在国内外学术刊物及国际会议发表论文30余篇。

浣军，美国堪萨斯大学计算机系教授，美国国家卫生局、航空航天局以及自然科学基金评审委员会、委员，Elisver及Springer大数据期刊编委等；研究方向为机器学习和大数据挖掘，以及在生物信息学中的应用；于2009年获美国国家科学基金会教师早期职业发展奖；在《自然–生物技术》等国际学术期刊及会议上发表论文100余篇。