002 | 聊聊2017年KDD大会的时间检验奖

2017-10-10 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 10:59 大小 5.04M

国际数据挖掘与知识发现大会**ACM SIGKDD**（ACM SIGKDD Conference on Knowledge Discovery and Data Mining），简称**KDD**，是由美国计算机协会**ACM**（The Association for Computing Machinery）的数据挖掘与知识发现专委会**SIGKDD**（Special Interest Group on Knowledge Discovery and Data Mining）主办，堪称数据挖掘研究领域的顶级会议。

KDD 最早是从 1989 年开始的 KDD 研讨班（Workshop）发展而来，当时的研讨班依托于人工智能顶级会议 IJCAI 大会或者 AAAI 大会，而后在 1995 年升级成为会议的模式，到现在已经有 20 多年的历史。今年的 KDD 大会于 8 月 13 日至 17 日在加拿大哈利法克斯成功召开。

SIGKDD 每年都会奖励一篇论文，这篇论文要在过去十年间对研究、方法论以及实践产生重大影响，这就是所谓的**时间检验奖**（Test of Time Award），引用次数以及对一个领域的影响力度是评选这个奖项的重要指标。

2017 年的 KDD 时间检验奖授予了美国康奈尔大学信息科学系主任、计算机科学系教授索斯藤·乔基姆斯（Thorsten Joachims）。这次授予是为了表彰他的论文《线性时间内训练线性支持向量机》（Training Linear SVMs in Linear Time），这篇论文也是 2006 年的 KDD 最佳论文，引用数超过 1600 多次。

Thorsten 的学术贡献

Thorsten 是一位机器学习界享有盛誉的学者，也是 ACM 和 AAAI 的双料院士，他所有论文的引用数加起来超过了 4 万次。2001 年从德国多特蒙德大学博士毕业后，他正式加入康奈尔大学从事机器学习研究。

获得这个奖项之前，Thorsten 曾多次获得重要奖项，比如 2017 年 ACM WSDM 的最佳论文奖（Best Paper Award）、2016 年 ACM SIGIR 的时间检验奖、2015 年 ACM KDD 的时间检验奖、2009 年 ECML 的最佳论文奖、2009 年 ICML 的 10 年最佳论文奖（Best 10-Year Paper Award）、2006 年 ACM KDD 的最佳论文奖、2005 年 ICML 的最佳论文奖、2005 年 ICML 的优秀学生论文奖、2005 年 ACM KDD 的最佳学生论文奖等。

Thorsten 在机器学习领域一直有着非常特殊的贡献。首先，他在支持向量机（SVM）的应用上做出了诸多努力。比如这次的时间检验奖，**就是奖励他如何把支持向量机的训练达到线性复杂度，从而使支持向量机在大规模数据上的应用成为可能。**

Thorsten 还致力于把支持向量机的基本算法，也就是仅仅支持分类问题和回归问题的算法，应用到更加复杂的有结构的输出结果上，俗称结构化的支持向量机算法。得益于这项工作，支持向量机可以对信息检索中很多复杂的、非二分的评估指标进行直接优化，如 F1 值（F-score）、平均精度均值（Mean Average Precision），从而让支持向量机的应用变得更加广阔。

在让支持向量机能够顺利应用到信息检索的过程中，Thorsten 还发现了另外一个问题，那就是如何利用搜索引擎的间接用户反馈（Implicit Feedback）来训练排序算法（经常是一个结构化的支持向量机模型）。具体来说，传统的搜索系统和信息检索系统主要是依靠人工标注的训练数据来进行优化和评估。这里所说的人工标注训练数据，主要是指人为地评价目标查询关键字和所对应的网页是否相关。

早期大家发现，虽然搜索引擎可以利用这样的数据来优化排序算法，但是搜索引擎在使用过程中会产生很多用户数据。这些数据可以是用户点击搜索页面结果产生的信息，也可以是其他的信息（比如用户在搜索页面的驻留时间等等）。早期这些信息并没有用于优化搜索引擎。以 Thorsten 为主的一批学者意识到点击信息的重要性，然后开始利用这些数据来训练和评估排序算法。这是 Thorsten 的第二个主要学术贡献。

Thorsten 第三个主要学术贡献，也是他最近几年的学术成功，那就是把**因果推论（Causal Inference）**和机器学习相结合，从而能够更加无偏差地训练模型。可以说这部分工作开创了一个新领域。

长期以来，如何有效地应用用户产生的交互数据来进行模型训练，都是大规模机器学习特别是工业界机器学习的难点。一方面，工业系统能够产生很多用户数据；另一方面，这些用户数据又受到当前部署系统的影响，一般都有一定的偏差。

因此工业级机器学习系统面临一个长期挑战，那就是，如何能够在评估模型以及训练模型的时候考虑到这样的偏差，从而去除这样的偏差。

Thorsten 利用因果推论中的倾向评分（Propensity Scoring）技术以及多臂赌博机（Multi-armed Bandit）思想，把这样的方法成功地引入到机器学习中，使得无偏差地训练模型成为可能。目前，这方面的新研究和新思想正在机器学习以及应用界产生越来越多的共鸣。

线性大规模支持向量机

回到这篇时间检验奖的论文，它解决的是大规模优化支持向量机的问题，特别是线性支持向量机。这篇文章**第一次提出了简单易行的线性支持向量机实现**，包括对有序回归（Ordinal Regression）的支持。算法对于分类问题达到了 O(sn)（其中 s 是非 0 的特征数目而 n 是数据点的个数），也就是实现了线性复杂度，而对有序回归的问题达到了 O(snlog(n)) 的复杂度。算法本身简单、高效、易于实现，并且理论上可以扩展到核函数（Kernel）的情况。

在此之前，很多线性支持向量机的实现都无法达到线性复杂度 。比如当时的 LibSVM（台湾国立大学的学者发明）、SVM-Torch、以及早期的 SVM-Light 中采用的分解算法（Decomposition Method）都只能比较有效地处理大规模的特征。而对于大规模的数据 (n)，则是超线性（Super-Linear）的复杂度。

另外的一些方法，能够训练复杂度线性地随着训练数据的增长而增长，但是却对于特征数 N 呈现了二次方 (N^2) 的复杂度。因此之前的这些方法无法应用到大规模的数据上。这样的情况对于有序回归支持向量机更加麻烦。从德国学者拉尔夫·赫布里希（Ralf Herbrich）提出有序回归支持向量机以来，一直需要通过转化为普通的支持向量机的分类问题而求解。这个转换过程需要产生 O(n^2) 的训练数据，使得整个问题的求解也在这个量级的复杂度。

这篇文章里，Thorsten 首先做的是对普通的支持向量机算法的模型形式（Formalism）进行了变形。他把传统的分类支持向量机（Classification SVM）写成了**结构化分类支持向量机（Structural Classification SVM）**，并且提供了一个定理来证明两者之间的等价性。粗一看，这个等价的结构化分类支持向量机并没有提供更多有价值的信息。然而这个新的优化目标函数的对偶（Dual）形式，由于它特殊的稀疏性，使它能够被用来进行大规模训练。紧接着，Thorsten 又把传统的有序回归支持向量机的优化函数，写成了结构化支持向量机的形式，并且证明了两者的等价性。

把两种模型表达成结构化向量机的特例之后，Thorsten 开始把解决结构化向量机的一种算法——**切割平面算法（Cutting-Plane）**，以下称 CP 算法，运用到了这两种特例上。首先，他展示了 CP 算法在分类问题上的应用。简单说来，这个算法就是保持一个工作集合（Working Set），来存放当前循环时依然被违反的约束条件（Constraints），然后在下一轮中集中优化这部分工作集合的约束条件。

整个流程开始于一个空的工作集合，每一轮优化的是一个基于当前工作集合的支持向量机子问题，算法直到所有的约束条件的误差小于一个全局的参数误差为止。Thorsten 在文章中详细证明了这个算法的有效性和时间复杂度。相同的方法也使得有序回归支持向量机的算法能够转换成为更加计算有效的优化过程。

Thorsten 在文章中做了详尽的实验来展现新算法的有效性。从数据的角度，他使用了 5 个不同的数据集，分别是路透社 RCV1 数据集的好几个子集。数据的大小从 6 万多数据点到 80 多万数据点不等，特征数也从几十到四万多特征不等，这几种不同的数据集还是比较有代表性的。从方法的比较上来说，Thorsten 主要比较了传统的分解方法。

有两个方面是重点比较的，第一就是训练时间。在所有的数据集上，这篇文章提出的算法都比传统算法快几个数量级，提速达到近 100 倍。而有序回归的例子中，传统算法在所有数据集上都无法得到最后结果。Thorsten 进一步展示了训练时间和数据集大小的线性关系，从而验证了提出算法在真实数据上的表现。

第二个重要的比较指标是算法的准确度是否有所牺牲。因为有时候算法的提速是在牺牲算法精度的基础上做到的，因此验证算法的准确度就很有意义。在这篇文章里，Thorsten 展示，提出的算法精度，也就是分类准确度并没有统计意义上的区分度，也让这个算法的有效性有了保证。

Thorsten 在他的软件包 SVM-Perf 中实现了这个算法。这个软件包一度成了支持向量机研究和开发的标准工具。

小结

今天我和你分享了 Thorsten 的这篇论文，堪称支持向量机文献史上的经典。一起来回顾下要点：第一，Thorsten 在机器学习领域有三大主要学术贡献；第二，这篇论文理论论证非常扎实，算法清晰，而且之后通过有效的实验完全验证了提出算法的有效性。文章开启了支持向量机在搜索领域的广泛应用，不愧为 2006 年的 KDD 最佳论文以及今年的时间检验奖论文。

最后，给你留一个思考题，在什么应用场景下，线性大规模支持向量机可以有比较好的效果?

欢迎你给我留言，和我一起讨论。

扩展阅读：Training Linear SVMs in Linear Time

上一篇　001 | 如何组建一个数据科学团队？

下一篇　003 | 数据科学家基础能力之概率统计

精选留言 (6)

写留言

JIA
2017-09-25
11

真是开眼界了，原来大牛是这么读经典论文的，学习的榜样啊！

展开

Zolynn
2017-09-29
5

收获很多

展开

yaolixu
2018-06-18

3

洪老师好,您说:“Thorsten 利用因果推论中的倾向评分（Propensity Scoring）技术以及（Multi-armed Bandit）思想，把这样的方法成功地引入到机器学习中，使得无偏差地训练模型成为可能。”
我对这方面的研究感兴趣,查看Thorsten教授的主页, 找到下面这篇论文: T. Joachims, A. Swaminathan, T. Schnabel, Unbiased Learning-to-Rank with Biased Feedback,...

展开

2018-01-12

3

实时相关的任务，比如实时推荐，实时分类等

展开

谢贵阳Garr...
2019-04-08

谁能解释一下什么是有序回归?

展开

登高
2018-05-13

文章没怎么(˚o˚;看懂，希望随着学习的深入可以明白

展开