纸质作业数据智能采集分析与个性化学习系统

□文 / 陈李江、张东祥

陈李江

阿凡题创始人兼 CEO。北京大学及耶鲁大学计算机联合培养博士、北京市特聘专家，并荣获中关村高端领军人才、腾讯年度教育企业风云人物、全球十大杰出新潮商等十余项荣誉。2013 年 10 月创办人工智能教育企业阿凡题，是首家拥有人工智能研究院的教育科技企业。

新加坡国立大学博士，现为浙江大学“百人计划”研究员、博士生导师、中国计算机学会数据库专委，已发表高水平论文 70 余篇，引用 1500 余次，获新加坡国立大学研究成果奖、2014 数据库 A 类国际会议 VLDB 最佳论文候选、2018 ACM 中国新星提名奖。

张东祥

“因材施教”的个性化教学一直以来都是教育领域中的一个重要研究方向，随着深度学习等前沿技术的不断发展进步，人工智能与教育的结合为智能化的个性化教学提供了新的研究方法。本文以阿凡题发布的基于纸质化作业的个性化学习系统为例，分析该系统是如何在不额外增加老师负担的前提下，通过人工智能和大数据分析等一系列技术手段来自动收集和分析学生的作业数据，实现个性化的教学指导方案。

一、背景介绍

个性化教学长久以来一直是智能教育的一个重要发展方向，它以获得学生平时的学习数据为前提，通过大数据分析和人工智能技术来诊断学习情况，并提出行之有效的改进方法。教育部在《教育信息化 2.0 行动计划》中也强调以人工智能、大数据、物联网等新兴技术为基础，依托各类智能设备及网络，积极开展智慧教育创新研究和示范，推动新技术支持下教育的模式变革和生态重构。

现阶段在中小学教学中，有两个主流方案可以大规模地收集学生数据，为个性化教学实现数据支撑。第一个方案是针对考试环节的阅卷机系统。阅卷机系统要求学生在固定的答题纸上作答，老师批改答题纸，再把答题纸通过扫描仪采集到系统中并进行数据分析。阅卷机系统可以较好地收集并分析学生考试数据。然而，在实际教学过程中，要使用阅卷机系统，老师需要提前把考试试题录入系统，并重新整理题目格式，再生成标准的答题模板供学生作答，这个步骤需要耗费老师大量时间。目前阅卷机系统一般在期中、期末、月考等正式考试中使用，在日常作业中使用较少。第二个方案是收集课堂教学环节数据的智慧教室系统。智慧教室可以实现教学环节全过程的数据记录，包括老师备课、讲课、课堂互动环节的数据，有效提升教学环节的互动效率。但是，智慧教室系统需要为每一位学生配置一台平板电脑，部署成本非常高，短期内无法覆盖所有师生。另外，在上课环节，学生听课过程中能够采集到的学情数据量较少，仅仅依赖于课堂互动数据无法实现有效的个性化教学。

除了考试和上课环节的学习数据，学生的另一大学习数据来源是平时的纸质作业，具备频率高、覆盖广等特点，能够更好地追踪学生的学习情况。2018 年 8 月，教育部等八部门联合制定了《综合防控儿童青少年近视实施方案》，倡导学校原则上使用纸质作业，使用电子产品开展教学时长原则上不超过教学总时长的 30%，以便更好地保护学生的视力。在此背景下，如何把学生的日常纸质作业全面收集并有效分析是一个技术难题，也是智能教育在学校落地的重要挑战。

为解决这一难题，阿凡题率先发布了基于纸质作业数据的个性化学习系统，这是一个端到端的一体化解决方案，通过学校打印机终端来采集学生的日常纸质作业，并基于人工智能技术加以分析，准确把握每个学生的学习进展，从而辅助制定个性化学习方案。该系统的优点在于能够贯彻教育部纸质化作业的倡导，并且不额外增加老师的教学负担。此外，阿凡题自行研发的自动化版面分析、智能标签、手写和批改痕迹识别等多项核心技术也能够有效地整合到该应用场景，极大地便利了教师的教学与学生的学习。目前该系统已经在近 40 所公立学校和 200 家线下辅导机构实施部署，有效达到了个性化学习的目的。

二、系统应用场景

学生纸质作业的采集有两种主流的手段：一是让学生自行通过手机拍照上传，二是学校统一部署扫描仪或者高拍仪进行收集。前者的优势是数据采集的时间和场所较为灵活，但仍存在两个严重的不足：首先是学生需要拥有使用手机的权限，这容易受到家长的抵制，也有悖教育部的倡导精神；其次是学生拍照的质量无法保障，拍摄的照片可能存在模糊、阴影和倾斜等问题，对后期的自动化识别算法带来很大的技术挑战。因此，本系统采用直接在学校部署高清扫描仪的集中式一体化管理方式来保障图片数据采集质量，通过自动识别作业版面和文本内容、自动标注题目标签，并根据批改痕迹自动识别正确的题目和错误的题目，最终生成一份学生的学情分析报告和个性化学习方案。

为了应对不同学校的教学需求，该系统提供两种部署方案：一种是建立集中的数据采集中心，所有班级的学生统一到数据中心进行作业扫描（如图 1 所示），能够实现资源的有效利用和设备维护，但需要占用额外的办公场所；另一个方案是在每个班级部署数据采集和打印输出一体机设备，好处是无需占用额外场所并且能够减少学生的步行距离（如图 2 所示）。

图 1：基于数据采集中心的部署方案

图 2：基于班级的部署方案

无论哪种部署方案，数据采集、分析和输出学情报告的流程是一致的，都包含如下三个部分：

第一，数据采集：这套系统是在不改变教师传统教学习惯的前提下，为满足学校常态化应用的需求，针对不同的教学场景，将日常的作业、周练、考试等过程性和结果性数据进行伴随式的采集，采用高拍仪和平板相结合的数据采集方式。同时，为了满足学生采集时的效率问题，我们支持整页数据采集（有别于传统的拍照搜题，只支持单道题拍照）。

第二，数据分析：整页作业采集到服务器端后，我们首先通过智能模板识别技术，把一页作业进行自动切题；然后，根据老师在作业中的批改记录（一般是红色笔批改的对号和错号），把错误的题目加入学生的错题集；再调用 OCR 识别技术和标签识别技术，这是分析学生的学情信息的核心环节；最后，根据学生作答的基本情况，生成学情报告和个性化学习方案。

第三，打印输出：为了响应教育部关于纸质作业的相关要求，我们支持学生的学情报告和个性化学习方案打印输出。在个性化学习方案中，我们提供了错题的举一反三强化练习，学生可以通过这些题目进一步巩固薄弱知识点。

图 3：个性化学情分析报告

三、AI 技术实现

基于纸质作业的个性化学习系统需要有足够强大的后台 AI 技术支持，涉及到版面分析、OCR 识别、手写识别、智能题目标注、举一反三等技术，目标是提升系统自动化程度从而达到尽可能降低人工成本的目的。下面我们将简要介绍每个核心技术模块的基本原理。

版面分析

在获取用户的纸质作业扫描图片之后，智能版面分析技术会对任意排版的作业进行语义识别和分析，识别出试卷科目和年级，并对试卷中所有题目进行自动切分，获取试卷中所有题目的题号、题型、题干文本以及学生作答情况等信息。简而言之，它将基于页的图片数据分解成基于题目为单元数据的信息，通过结合识别技术和标签算法，自动对学生错题进行整理和分析。因此，智能版面切割与学情报告相结合的自动化模式可以极大提高老师的生产力。以前基于老师手动生成的模式需要清楚了解学生的实际学习情

况，并耗费一位一线教师数个小时的时间来设计一份优质的个性化学情报告，而本文设计的系统可以实现全自动化，几秒钟便可以生成个性化学情报告。

OCR 识别

OCR 识别的基本任务是从图片中获取题目的文本信息。目前 OCR 识别属于较为成熟的技术，已被阿凡题、作业帮和猿辅导等 AI 教育公司广泛应用到基于手机的拍照搜题 APP 产品当中，能够支持千万级别学生的日常作业拍搜需求。由于题目的信息主要以打印体形式呈现，字与字之间有空白间隔，常用的识别策略是利用图片的连通域进行切割，一个字符或者汉字通常对应一个矩形连通区域，之后再通过训练一个卷积神经网络来对每个矩形框的内容进行识别。为了进一步提升准确率，可引入自然语言处理常用的语言模型（language model）用于对识别结果进行修正。同时，为了能够处理公式结构或者一个汉字分为多个连通区域的情况，可以设置基于规则的算法来进行识别矫正。

手写识别

手写识别的难点在于数据可能涵盖多种科目（英语、语文、数学、物理、化学等），混合了多种语言（英语、汉语）和专业符号（数学、物理、化学等），而且不同用户的书写习惯差异性大，这使得设计一种通用的手写识别方法成为非常具有挑战的问题。在中文、英语作文批改的业务场景中，由于用户的书写习惯呈行结构，因此阿凡题设计了基于行字符的识别模型。具体而言，其采用主流的基于卷积神经网络和双向长短记忆模型的时序分类模型，如图 4 所示，通过卷积神经网络提取行图片的视觉特征，然后通过序列数据分类的模型，对行图片的字符进行连续预测。对于复杂排版的书写结构（如数学公式、理科表达式等），则充分考虑拍照作业图片中手写数据的结构化特点，进而采用基于序列翻译和注意力机制的模型。

智能标注

该系统使用的是阿凡题自主研发的题库智能标注系统来自动获取每道题目对应的知识点标签，其基本原理是通过构建细粒度知识点地图，并采用深度学习模型和自然语言处理技术，对题目进行智能标注。经专业教师的效果验证，标注能力相当于高级教师的专业水准（准确率达到 95%），也成为国内首个对外开放的智能题库标注服务。除了知识点的掌握情况，一份高质量的学情报告还能因地制宜，通过海量题库查找该地区五年内的各大考试试卷，总结出学生这些错题考点在考试当中出现的频次和分值，题目和考点的重要性可以通过考试频次和所占分值来确定。此外，如何对学生的能力进行综合量化分析也是生成一份高质量个性化诊断报告的必要环节。本系统采用的是自主研发的

三维衡量体系，聚焦三个重要维度：知识（knowledge）、技能（skills）和能力（ability）。其中，知识（K）代表具有事实、概念或程序性质的有组织的信息体，是教学中传授与传承的内容；技能（S）指用身体、语言或智力操作信息或事物的熟练程度，是教育或教学应完成的目标，表现的是学生练习与训练后的结果；能力（A）表示执行当前可观察到的活动的优秀程度，体现的是教育作用相同的情况下，学生之间存在的个体差异。

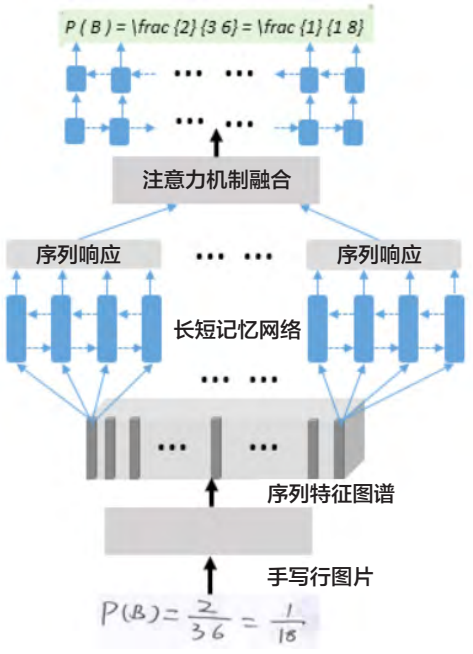

图 4：基于深度学习的手写识别模型

举一反三

举一反三的本质是针对学生错题进行相似题目推荐，目的是让学生能够对薄弱的知识点进行反复练习直至熟练掌握。传统基于编辑距离（edit distance）的字符串相似性的算法并不奏效，原因是它没有区分不同单词的重要程度，容易造成推荐的题目重复或者跟考察的知识点不是特别相关。本系统采用的策略是将题目映射到细粒度的题型，通过监督机器学习的方式来预测题目文本的每个单词跟知识点和题型的相关程度，并基于此权重来计算候选题目和学生错题的相似性。为了进一步提升举一反三的题目推荐速度，我们还引入了基于 gram 的索引技术来快速过滤不相关的题目。

四、总结

本文针对目前学生学习数据收集的困境，遵从教育部的有关指示精神，提出了基于纸质化作业的个性化学习系统。目标是在不额外增加老师负担的前提下，通过人工智能和大数据分析等一系列技术手段来实现一个低成本、可持续的软硬件结合的系统，自动收集和分析学生作业，并实现真正意义上的个性化指导方案。该系统有效地集成了多项 AI 技术，包括版面分析、OCR 识别、手写识别、题目标注和举一反三等核心模块，有望成为“AI+ 教育”的一个成功的代表性案例。

查看内容精选