面向大规模认知诊断的 DINA 模型快速计算方法研究

王 超[1],刘 淇[1],陈恩红[1],黄振亚[1],朱天宇[1],苏 喻[2],胡国平[3]
(1. 中国科学技术大学计算机科学与技术学院,安徽合肥 230027;
2. 安徽大学计算机科学与技术学院,安徽合肥 230039;
3. 科大讯飞股份有限公司,安徽合肥 230088)

摘 要: 在教育教学的过程中,如何诊断学生的知识水平是一个重要的问题.传统方法大多由教师根据学生的表现和成绩进行人工判断,存在效率低、主观性强的问题,且难以做到针对大量学生的个性化诊断.近年来,认知诊断模型中的 DINA 模型被广泛应用于诊断学生个性化知识掌握程度.然而传统 DINA 模型大多基于小样本数据,当面对在线教育带来的大规模数据处理需求时,存在收敛速度慢的问题,难以实际应用.针对 DINA 模型计算时间过长的问题,本文首先给出了 DINA 模型的收敛性证明,并提出了三种能够加速 DINA 求解的算法:(1)增量算法,它将学生数据划分为多个学生块,每次迭代只访问其中一个学生块;(2)最大熵方法,它只访问在极大化模型熵的过程中影响较大的学生数据;(3)基于前两者的混合方法.最后,本文通过真实数据和模拟数据上的实验,分析证明了三种方法均能在保证 DINA 模型有效性的情况下,达到几倍至几十倍的加速效果,有效地改善了 DINA 模型的计算效率.

关键词: 教育数据挖掘; 认知诊断; DINA 模型; EM 算法; 加速收敛
中图分类号: TP301 **文献标识码:** A **文章编号:** 0372-2112 (2018) 05-1047-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.05.004

The Rapid Calculation Method of DINA Model for Large Scale Cognitive Diagnosis

WANG Chao[1],LIU Qi[1],CHEN En-hong[1],HUANG Zhen-ya[1],ZHU Tian-yu[1],SU Yu[2],HU Guo-ping[3]
(1. *School of Computer Science and Technology,University of Science and Technology of China,Hefei,Anhui* 230027,*China*;
2. *School of Computer Science and Technology,Anhui University,Hefei,Anhui* 230039,*China*;
3. *USTC iFLYTEK Co,Ltd,Hefei,Anhui* 230088,*China*)

Abstract: How to assess students' cognitive structure is an important problem in the process of education and teaching. Traditionally,teachers evaluate a student based on their classroom performance and scores,which is lack of efficiency,objectivity,and it is hard to treat anyone equally. To solve this problem,DINA model,which is able to evaluate knowledge proficiency of students,has become a popular Cognitive Diagnosis model with a good interpretation. However,traditional DINA models are all based on small samples. When it comes to large-scale online learning scenario,the calculation will be significantly time-consuming. To address these issues,we first give proof of the convergence of DINA model,and then propose three acceleration methods. To be specific,the first one,called Incremental DINA(I-DINA),can partition the student data into blocks and iterate through the blocks. Then the second one,Maximum-Entropy DINA(ME-DINA),can choose and only access the most powerful students. At last,we combine the advantages of these two methods and propose the last model called Incremental Maximum Entropy DINA(IME-DINA). Extensive experiments on both a real-world dataset and simulation data demonstrate that our models can achieve dozens of acceleration without reducing the effectiveness of DINA Model.

Key words: educational data mining; cognitive diagnosis; DINA model; convergence acceleration; expectation maximization algorithm

收稿日期: 2016-12-20; **修回日期:** 2017-03-07; **责任编辑:** 蓝红杰
基金项目: 国家 863 高技术研究发展计划(No. 2015AA015409);国家杰出青年科学基金(No. 61325010);国家自然科学基金(No. 61672483,No. U1605251);中科院青年创新促进会会员专项基金(会员编号 2014299)

1 引言

随着互联网教育的迅猛发展,大量在线智能教育系统进入了公众的视线①,为学生实现自主学习提供了可能[1,2]. 然而,在线教育系统在提供便利的同时,由于其具有庞大的学习资源库,往往也会给平台提供自主学习服务带来诸多困难[3,4]. 因此,基于学生学习数据,借助技术手段准确地对学生进行学习分析,从而为学生进行个性化的学习推荐,让在线教育系统做到因材施教,已成为当前面向教育数据挖掘分析的重要研究问题[5].

要想达到因材施教的教育目标,首先要清楚地鉴别每个学生的潜在学习状态[6]. 传统教学方法依靠教师经验判断,不仅耗费大量的时间和精力,结果也不够准确,更不适合在线教育需要面对的学生规模. 为此,教育心理学家提出了认知诊断评估方法. 认知诊断评估基于学生对试题的作答情况,通过对学生进行个性化的认知诊断[7]建模,从而得到学生潜在知识水平的掌握情况,进而为学生个性化学习、资源推荐提供了基础[8]. 具体地,它假设题目与知识点之间存在显示关联,可以用 Q 矩阵[9]来表示(表1展示了三道数学题目与四则运算五个知识点之间的 Q 矩阵). 进一步,它认为学生对于试题的作答表现受到学生对知识点的掌握程度(学生潜在学习状态)的影响[9].

表1 试题知识点关联 Q 矩阵

题目 \ 知识点	加法	减法	乘法	除法	括号
11 +3 -5	1	0	1	0	0
21* 4/3	0	1	1	1	0
(5 +2) * 12	0	0	0	1	1

在认知诊断评估中,Deterministic Inputs,Noisy "And" gate model(DINA 模型)旨在对学生多维知识点掌握程度进行建模分析[12],它能够在精准建模学生学习状态的同时保证了较好的可解释性,近年来广受学者关注和研究[13]. 具体地,DINA 结合了 Q 矩阵作为试题的先验知识,将学生的潜在学习状态描述成一个多维知识点掌握向量,同时引入题目的猜测和失误参数,以准确地在多维知识层面诊断学生的认知学习状态. 以表1中题目为例,某学生根据 DINA 模型得到的认知诊断结果如图1所示,由图可知该学生掌握了加减乘除,但没有掌握括号知识点.

然而,受技术和环境所限,传统的 DINA 模型多基于小样本[14],如一次考试结果. 而在线教育系统所提供的学生数据规模远超传统应用场景,此时 DINA 模型的诊断速度将大大降低,限制了 DINA 模型的应用[15]. 近年来,有学者提出通过增加超参和采样训练的方式改善 DINA 模型的计算效率,如 HO-DINA[16]和 FuzzyC-DF[17]模型. 但其建模过程均需引入超参,从而破坏了 DINA 模型在参数解释性上的优势. 因此,如何在不增加参数的情况下且不破坏模型收敛性的情况下,优化 DINA 模型的计算效率,使其能够更好的适应大规模学生知识点掌握评估,仍是近年来学者们研究的重点,也是本文关注的核心问题.

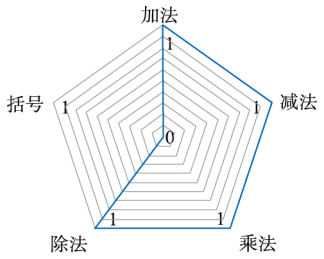

图1 认知诊断结果雷达图

针对上述问题,本文结合 DINA 模型中 EM 算法的数据划分特性及收敛性[18],首先从两种角度提出了增量 DINA 模型(I-DINA)与最大熵 DINA(ME-DINA)模型. 其中,增量 DINA 模型通过将原始数据分成多个学生块增量式地处理迭代,而最大熵 DINA 模型则是筛选掉极大化模型熵的过程中影响较小的学生. 最后,本文结合前两种方法,进一步提出了增量最大熵 DINA 模型(IME-DINA). 相比于前文提到的改进方法,本文给出的三种方法一方面保留了 DINA 模型原有的所有参数,保证了可解释性这一优点;另一方面,由于本文所提出的三种优化算法均是基于 EM 步骤的改进,故仍然可以迭代求解,并保证其参数收敛性. 总而言之,本文的主要贡献总结如下:

(1)据作者调研所知,本文首次给出了 DINA 模型收敛性的证明,完善了 DINA 模型的研究.

(2)据作者调研所知,本文首次对 DINA 模型中数据采用熵解释的办法给出了合理的数据划分方法,并给出了理论证明.

(3)本文所提出的三种 DINA 模型优化方法,均不改变 DINA 模型的参数,使得其在提高 DINA 模型的运行效率的同时仍可保证 DINA 模型的可解释性,具有更好的可扩展性.

2 相关工作

本章节将从认知诊断模型、EM 算法优化研究两个方面介绍相关工作.

2.1 认知诊断模型

认知诊断模型旨在通过学生的做题历史记录,诊

① 猿题库:http://www.yuantiku.com/;全通教育:http://www.qtone.cn/;智学网:http://www.zhixue.com/

断学生的潜在能力,可以分为单维连续模型(以 IRT 模型[19]为代表)和多维离散模型(以 DINA 模型为代表)[6]. 相比于类 IRT 模型[19],DINA 模型具有更优秀的参数可解释性,近年来受到学者的关注并被广泛应用[15,20],与之相关的研究如 Templi(2006)提出的补偿型知识掌握模式 DINO 模型[21]和 De La Torre J(2011)提出的一般化 G-DINA 模型[22]. 然而,上述关于 DINA 的研究大多仍基于小样本测试数据,且局限于 DINA 模型的诊断效果,却忽视了由于收敛速度慢而导致难以大规模应用的现实问题[14,15]. 随着教育数据的增长,传统 DINA 模型不再适应现有的数据规模.

针对 DINA 模型计算效率的问题,有学者开始研究相关改进模型[13,16,17]. 其中,最为有效的改进模型有 HO-DINA 模型和 FuzzyCDF 模型[16,17]. HO-DINA[16]模型为了降低知识点的维度,假设知识点掌握程度受到一个代表综合能力的超参所控制. FuzzyCDF[17]模型则利用模糊理论将 DINA 模型中的知识点掌握程度由 0-1 离散分布改为在 [0,1] 区间上的连续分布,为此同样需要引入超参. 这两种方法虽然降低了计算复杂度,但同时引入了两个问题: 一是失去了原本参数优秀的现实可解释性,二是破坏了 EM 算法的适用性.

2.2　EM 算法

在 DINA 模型中,存在无法直接观察的隐变量,即学生的知识点掌握程度,故需要采用 EM 算法[23]来解决不完全数据的参数估计问题. 但当缺失的信息量很大时,EM 算法的收敛速度将会很缓慢[24],在 DINA 模型中这一点尤为明显.

为了改进 EM 算法的收敛速度,不少学者提出了诸多改进的算法,均可分为基于 E 步或 M 步的加速方法. 如基于 E 步的加速方法中,MCEM 算法[25]在 E 步难以得到期望的显示表达式时,用 Monte Carlo 模拟来完成 E 步. 在基于 M 步的加速方法中,ECM 算法[26]和 ECME 算法[27]在 M 步没有显示表达式时,通过用有约束条件情况下完整数据或观察数据对数似然函数的极大化来代替 M 步中极大化步骤,实现计算的简化; 此外,A-ECM 算法[28]根据 EM 算法初始迭代速率很快与 Aitken 法末尾迭代速率很快的特点,混合两种方法来加快收敛速度等. 由于 DINA 模型中 M 步有显示表达式,而 E 步的计算时间却很长,故本文将介绍三种基于 E 步的 DINA 加速方法.

3　DINA 模型

DINA 模型通过利用 Q 矩阵的信息来建模学生对题目的作答变量,借此诊断学生的知识点掌握程度. 一般假设各题作答相互独立,满足伯努利分布. 当有 I 个学生,J 道题目和 K 个知识点时,DINA 模型中的潜在作答变量 η_{ij} 可以表示为:

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} \tag{1}$$

其中 $\alpha_{ik}=1$ 或 0 表示学生 i 掌握或没有掌握知识点 k,学生的潜在能力矩阵 $\boldsymbol{\alpha}=\{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_I\}'$,$Q$ 矩阵 $Q=\{q_{jk}\}_{J*K}$. 故 η_{ij} 反映了学生能力是否足够答对该题. 在引入失误率 s 和猜测率 g 两种题目参数后,实际响应矩阵 $X\{X_{ij}\}_{I*J}$ 的概率模型为:

$$P_j(\boldsymbol{\alpha}_i) = P(X_{ij}=1+\boldsymbol{\alpha}_i) = g_j^{1-\eta_i}(1-s_j)^{\eta_i} \tag{2}$$

表 2 给出了 DINA 模型的符号及对应描述. 图 2 展示了 DINA 模型的图模型.

表 2　DINA 模型所涉及的符号及描述

符号	描述
i	学生
j	题目
k	知识点
$\boldsymbol{X}$	得分矩阵
X_{ij}	学生 i 在题目 j 上得分情况
$\boldsymbol{\alpha}$	潜在能力矩阵
α_{ik}	学生 i 对知识点 k 掌握情况
$\boldsymbol{\eta}$	潜在作答矩阵
η_{ij}	学生 i 在题目 j 上潜在作答情况
$\boldsymbol{Q}$	题目知识点关联矩阵
q_{jk}	题目 j 对知识点 k 考查情况
s_j	题目 j 的失误率
g_j	题目 j 的猜测率

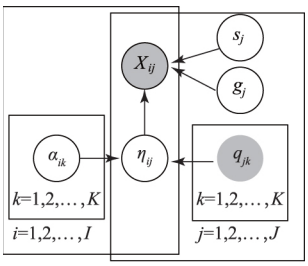

图2　DINA图模型

DINA 模型的总似然函数为:

$$L(\boldsymbol{X}) = \prod_{i=1}^{I} L(\boldsymbol{X}_i) = \prod_{i=1}^{I}\sum_{l=1}^{L} P(\boldsymbol{X}_i|\boldsymbol{\alpha}_l)\,P(\boldsymbol{\alpha}_l) \tag{3}$$

其中,$L=2^k$ 由于式(3)中含有隐变量 $\boldsymbol{\alpha}_l$,无法直接进行极大似然估计,DINA 模型引入 EM 算法,采用极大边缘似然估计的方法[12]求解:

E 步: 利用上一轮得到的 s_j 与 g_j 估计计算矩阵 $P(\boldsymbol{X}|\boldsymbol{\alpha}) = [P(\boldsymbol{X}_i|\boldsymbol{\alpha}_l)]_{I\times L}$ 的值,并利用 $p(\boldsymbol{X}|\boldsymbol{\alpha})$ 计算矩阵 $P(\boldsymbol{\alpha}|\boldsymbol{X}) = [P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i)]_{I\times L}$ 的值,其中 $i=1,2,\cdots,I$,$l=1,2,\cdots,L$.

M 步: 分别令 $\frac{\partial \log L(\boldsymbol{X})}{\partial s_j}$ 和 $\frac{\partial \log L(\boldsymbol{X})}{\partial g_j}=0$,可得:

$$\hat{s}_J = \frac{I_{jl}^1 - R_{jl}^1}{I_{jl}^1} \tag{4}$$

$$\hat{g}_j = \frac{R_{jl}^0}{I_{jl}^0} \tag{5}$$

其中 I_{jl}^0 表示属于第 l 种知识点掌握模式的学生中缺乏至少一个第 j 题所需知识点的人数期望，R_{jl}^0 表示 I_{jl}^0 中回答正确第 j 题的人数期望，I_{jl}^1 和 R_{jl}^1 的含义与 I_{jl}^0 和 R_{jl}^0 相似，不同之处在于 I_{jl}^1 与 R_{jl}^1 是在学生掌握所有第 j 题所需知识点的情形下的期望. 故可由 E 步中得到的估计，计算 I_{jl}^0，R_{jl}^0，I_{jl}^1 和 R_{jl}^1 的值，并由此得到新的 s_j 与 g_j 估计. 由于 E 步中需多次计算大规模矩阵乘积，DINA 模型的计算效率较低，且此前没有工作给出 DINA 模型收敛性的相关证明.

4 DINA 模型的快速实现算法

鉴于 DINA 模型计算效率较低，本章将介绍三种加速 DINA 模型的方法——增量方法、最大熵方法和结合前两种方法的混合方法及其收敛性证明. 三种方法的共同点是在 DINA 模型的 E 步中每次迭代只遍历部分学生，对不访问的学生保留上一次迭代得到的结果[28]. 整体流程如图 3.

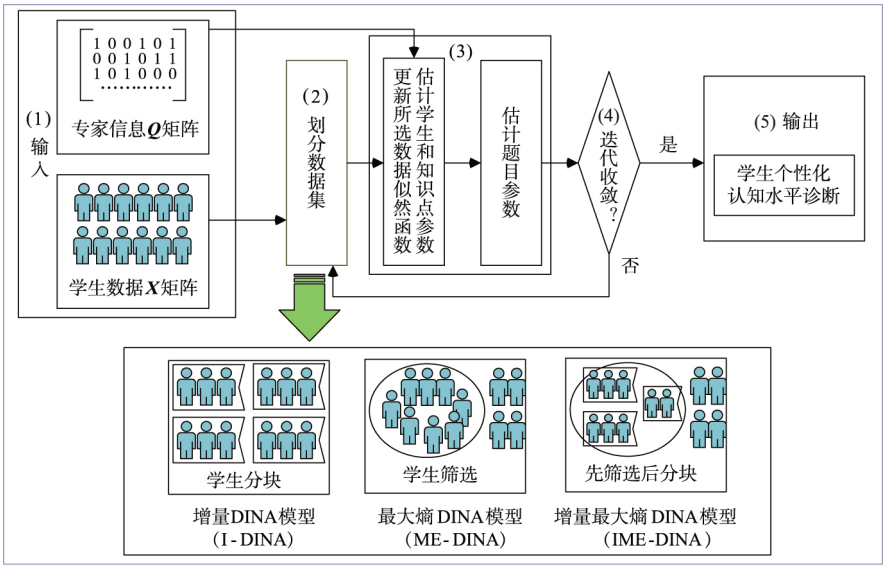

图3 加速DINA模型框架

(1) 输入: 学生对一组题目的实际响应矩阵 X 和专家提供的试题知识点关联矩阵 Q.

(2) 划分学生数据集: 按照所选择的加速方法，将学生矩阵 X 划分为需要在此次迭代中访问的学生矩阵 X_Y 与不需要访问的学生矩阵 X_N.

(3) 迭代求解 DINA: 按照第 3 章中介绍的步骤进行迭代求解，在更新似然函数时只访问学生矩阵 X_Y，保留上次迭代得到的 X_N 中学生的结果.

(4) 判断迭代是否收敛: 若不收敛，返回第(2)步.

(5) 利用加速 DINA 模型的迭代结果，对学生知识点掌握程度进行个性化分析.

为确保模型的实用性，必须保证其可以收敛. 本文先给出 DINA 模型的 EM 收敛性证明，并在之后给出改进算法的收敛性证明. 据作者调研所知，本文首次给出了 DINA 模型收敛性的证明.

定理 1 DINA 模型中的 EM 步骤收敛.

证明 给定学生得分矩阵 X 和题目知识点关联矩阵 Q，问题是 DINA 模型是否可以保证达到模型总对数似然函数 $l(X)$ 的极大化从而停止迭代.

$$\begin{aligned} l(X) &= \prod_{i=1}^{n} \log L(X_i) \\ &= \sum_{i=1}^{n} \log \sum_{l=1}^{L} P(X_i \mid \alpha_l) P(\alpha_l) \\ &= \sum_{i=1}^{n} \log \sum_{l=1}^{L} \mu_l \frac{P(X_i \mid \alpha_l) P(\alpha_l)}{\mu_l} \\ &\geqslant \sum_{i=1}^{n} \sum_{l=1}^{L} \mu_l \log \frac{P(X_i \mid \alpha_l) P(\alpha_l)}{\mu_l} \end{aligned}$$

令 $l(X)^t$ 表示第 t 次迭代时的总似然函数值，则说明算法收敛只需证明 $l(X)^{t+1} > l(X)^t$ 即可.

$$\begin{aligned} l(X)^{t+1} &= \sum_{i=1}^{n} \sum_{l=1}^{L} \mu_l^{t+1} \log \frac{P(X_i \mid \alpha_l)^{t+1} P(\alpha_l)}{\mu_l^{t+1}} \\ &> \sum_{i=1}^{n} \sum_{l=1}^{L} \mu_l^{t} \log \frac{P(X_i \mid \alpha_l)^{t+1} P(\alpha_l)}{\mu_l^{t}} \\ &\geqslant \sum_{i=1}^{n} \sum_{l=1}^{L} \mu_l^{t} \log \frac{P(X_i \mid \alpha_l)^{t} P(\alpha_l)}{\mu_l^{t}} \\ &= l(X)^t \end{aligned} \tag{6}$$

其中第一个不等号成立是因为只有当 μ_l 取值为 $\mu_l^{t+1} = P(\alpha_l \mid X_i)^{t+1}$ 时才满足琴生不等式等号成立条件; 第二

个不等号成立正是 M 步的工作. 故 DINA 模型中的 EM 步骤是可以保证收敛的.

证毕.

在 4.2 节到 4.4 节将分别介绍本文给出的三种基于 E 步的 DINA 模型加速方法.

4.1　增量 DINA 模型

当 DINA 模型的 E 步需要耗费大量时间时,增量式的迭代参数是一种很直观的想法,增量 DINA 模型(Incremental DINA,I-DINA) 通过将 DINA 模型中的完整 E 步(complete E-step) 改变为部分 E 步(partial E-step) 的方法来减少计算花费的时间[29]. 即将数据集 X 划分成 N 个不相交的学生块$\{X^1,X^2,\cdots,X^N\}$,在每次进行 I-DINA 模型的部分 E 步时,只遍历其中一个学生块 X^j 来更新似然函数,而对于其他学生块,则保留上一次迭代得到的似然函数值. 算法 1 展示了 I-DINA 模型的具体步骤.

算法 1　I-DINA 模型

输入: 学生得分矩阵 X,知识点关联矩阵 Q
输出: 猜测率 g,粗心率 s,似然函数 $P(\boldsymbol{\alpha}|X)$
1. 算法初始化,取初值 $s^0,g^0,j=0$;
2. 将数据集 X 划分为 N 个学生块$\{X^1,X^2,\cdots,X^N\}$;
3. WHILE 算法未收敛 do
　　Select j // partial E-step
　　IF $X_i\in X^j$,THEN 由 s^{t-1},g^{t-1} 计算 $P(X_i|\boldsymbol{\alpha}_l)^t$
　　　　ELSE $P(X_i|\boldsymbol{\alpha}_l)^t=P(X_i|\boldsymbol{\alpha}_l)^{t-1}$
　　由 $P(X|\boldsymbol{\alpha})^t$ 计算 $P(\boldsymbol{\alpha}|X)^t$;
　　由 $P(\boldsymbol{\alpha}|X)^t$ 计算 s^t,g^t//M-step
　END
4. RETURN $s,g,P(\boldsymbol{\alpha}|X)$

I-DINA 模型与 DINA 模型一样具有理论上的收敛保证. 证明如下.

定理 2　I-DINA 模型中的 EM 步骤收敛.

证明　与 DINA 模型的证明步骤是类似的,仍然是用式(6) 说明 $l(X)^{t+1}\geqslant l(X)^t$ 即可. 注意到式(6) 中只要 $P(X_i|\boldsymbol{\alpha}_l)^t$ 与 $P(X_i|\boldsymbol{\alpha}_l)^{t+1}$ 不完全相等,第一个不等号就是成立的. 对于 I-DINA 模型,虽然当 $X_i\notin X^j$ 时,$P(X_i|\boldsymbol{\alpha}_l)^t=P(X_i|\boldsymbol{\alpha}_l)^{t+1}$,但对于 $X_i\in X^j$,$P(X_i|\boldsymbol{\alpha}_l)^t$ 是会更新的,除非已极大化总似然函数,迭代停止了,否则 $P(X_i|\boldsymbol{\alpha}_l)^t$ 与 $P(X_i|\boldsymbol{\alpha}_l)^{t+1}$ 不可能完全相等,所以第一个不等号成立. 而第二个不等号成立是由 M 步所保证的,I-DINA 模型和 DINA 模型的 M 步是相同的. 综上,I-DINA 模型也可以保证 EM 步骤是收敛的.

证毕.

I-DINA 模型相比 DINA 模型,划分学生块的数目对于模型计算时间的影响是明显的,最极端的情况,当只有一个学生块时,I-DINA 模型实际上就是 DINA 模型,这一点将在实验部分进一步讨论.

4.2　最大熵 DINA 模型

I-DINA 模型由于缺乏对学生数据的先验信息,导致每次模型的计算时间差异很大,此时一种对学生数据的筛选机制会有很大的帮助,为此本文提出了最大熵 DINA 模型(Maximum Entropy DINA,ME-DINA) . 它基于这样的经验事实: 在每次迭代过程中,不是所有的数据都有相同的地位. 有些数据对于迭代过程中参数的步长贡献很大,而另一些数据可能起不到太大作用,但二者遍历花费的时间是相同的,这就拖慢了迭代收敛的速度.

具体的说,ME-DINA 模型通过将学生数据集 X 重新划分为对参数更新贡献较大的变化集 X^C 和对贡献较小的懒惰集 X^l. 在每次迭代过程中,将会选择进行完整 E 步或者懒惰 E 步(lazy E-step) . 在完整 E 步之后会按照某种规则对学生数据集 X 进行一次筛选,划分为懒惰集 X^l 与变化集 X^C 两个集合; 当进行懒惰 E 步的时候,只访问属于变化集 X^C 的学生数据来更新似然函数,而对于懒惰集 X^l,则保留上一次迭代得到的似然函数值. 算法 2 展示了 ME-DINA 模型的具体步骤.

ME-DINA 模型的核心就是用来划分懒惰集与变化集的筛选条件. 根据最大熵原理,模型应当选用具有最大熵时的概率分布,ME-DINA 模型就是从这个角度入手,制定的筛选规则. 令 $\mathrm{ST}_{\max}(X_i)$ 和 $\mathrm{ST}_{\min}(X_i)$ 分别表示学生 i 属于第 l 种知识点掌握模式概率的上下界,即:

$$\mathrm{ST}_{\max}(X_i)=\max_{l\in\{1,\cdots,L\}}[p(\boldsymbol{\alpha}_l|X_i)] \tag{7}$$

$$\mathrm{ST}_{\min}(X_i)=\min_{l\in\{1,\cdots,L\}}[p(\boldsymbol{\alpha}_l|X_i)] \tag{8}$$

则当 $\mathrm{ST}_1<\mathrm{ST}_{\min}(X_i)$ 且 $\mathrm{ST}_{\max}(X_i)<\mathrm{ST}_2$ 时,将学生数据 X_i 划分至变化集 X^C,否则将学生数据 X_i 划分至懒惰集 X^l,其中 ST_1 与 ST_2 是先验阈值. 由于 $H(\tilde{p}_i)=-E_{\tilde{p}_i}[\log\tilde{p}_i]$ 是 $\tilde{p}_i=p(\boldsymbol{\alpha}_l|X_i)$ 的信息熵,且 $\sum_{l=1}^{L}p(\boldsymbol{\alpha}_l|X_i)=1$,故上述方法实际上筛选掉了当前熵比较大的学生数据,而保留熵比较小的学生数据迭代增加熵值,以此极大化模型总的熵值.

算法 2　ME-DINA 模型

输入: 学生得分矩阵 X,知识点关联矩阵 Q
输出: 猜测率 g,粗心率 s,似然函数 $P(\boldsymbol{\alpha}|X)$
1. 算法初始化,取初值 s^0,g^0;
2. WHILE 算法未收敛 do
　　IF 执行 complete E-step THEN
　　　　由 s^{t-1},g^{t-1} 计算 $P(X_i|\boldsymbol{\alpha}_l)^t$,并计算 $P(\boldsymbol{\alpha}|X)^t$
　　　　划分变化集 X^C 与懒惰集 X^l

END
ELSE 执行 lazy E-step
If $X_i \in X^C$,then 由 s^{t-1},g^{t-1} 算 $P(X_i|\alpha_l)^t$
else $P(X_i|\alpha_l)^t = P(X_i|\alpha_l)^{t-1}$
由 $P(X|\alpha)^t$ 计算 $P(\alpha|X)^t$
END
由 $P(\alpha|X)^t$ 计算 s^t,g^t// M-step
END
3. RETURN s,g,$P(\alpha|X)$

注意到本文对 I-DINA 模型的证明其实适用于所有不完整 E 步方法的 DINA 模型,即有:

定理 3 ME-DINA 模型中的 EM 步骤收敛.

证明 完全类同于 I-DINA 模型的证明,当 $X_i \in X^l$ 时,$P(X_i|\alpha_l)^t = P(X_i|\alpha_l)^{t+1}$,但对于 $X_i \in X^C$,$P(X_i|\alpha_l)^t$是变化的,故式(7)仍然成立. ME-DINA 模型的 EM 步骤是收敛的.

证毕.

4.3 增量最大熵 DINA 模型

ME-DINA 模型相比 I-DINA 模型,由于阈值确定后,学生数据集的划分方法也随之确定,故不会出现时间上的不稳定性,但加速效果往往不如 I-DINA 模型好. 于是,本文结合了前两种方法的优点,即 I-DINA 模型的快速性与 ME-DINA 模型的稳定性,进一步提出了增量最大熵 DINA 模型(Incremental Maximum Entropy DINA,IME-DINA). 即进行 E 步计算时划分懒惰集 X^l 与变化集 X^C,再在变化集 X^C 中划分学生快,同样可以保证收敛. 算法 3 展示了 IME – DINA 模型的具体步骤.

算法 3 IME-DINA 模型

输入: 学生得分矩阵 X,知识点关联矩阵 Q
输出: 猜测率 g,粗心率 s,似然函数 $P(\alpha|X)$
1. 算法初始化,取初值 s^0,g^0;
2. WHILE 算法未收敛 do
IF 执行 complete E-step THEN
由 s^{t-1},g^{t-1} 计算 $P(X_i|\alpha_l)^t$,并计算 $P(\alpha|X)^t$
划分懒惰集 X^l 与变化集 $X^C = \{X^1, X^2, \cdots, X^N\}$;
END
ELSE 执行 partial lazy E-step,Select j
IF $X_i \in X^j$,THEN 由 s^{t-1},g^{t-1} 算 $P(X_i|\alpha_l)^t$
ELSE $P(X_i|\alpha_l)^t = P(X_i|\alpha_l)^{t-1}$
由 $P(X|\alpha)^t$ 计算 $P(\alpha|X)^t$
END
由 $P(\alpha|X)^t$ 计算 s^t,g^t// M-step
END
3. RETURN s,g,$P(\alpha|X)$

5 实验

由于 DINA 模型非凸,不同的初值可能会造成不同的局部解,讨论理论收敛速度的意义不大,本文着重通过实验,从加速能力和参数影响两方面来测试三种加速方法的性能.

5.1 数据集介绍

数据集 1 和数据集 2 来自某市部分高中生的数学考试记录,最早由文献[17]提出并使用. 此外,本文额外生成了一个包括十万学生的模拟数据集(数据集 3)进行实验,从而考查在面对大规模数据的情况下三种加速方法的加速能力. 表 3 给出了数据集的简单统计. 图 4 展示了两个数据集的 Q 矩阵,黑色方块表示题目和该知识点相关. 实验均在 intel 4 核处理器环境下使用 R 语言运行.

表 3 数据集的简单统计

数据集	学生数	题目数	知识点
数据集 1	4,209	15	11
数据集 2	3,911	16	16
数据集 3	100,000	16	16

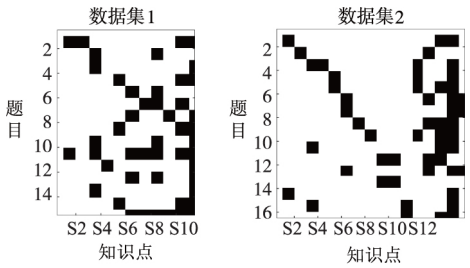

图4 两个数据集的Q矩阵

5.2 模型效果对比实验

本实验考察三种加速方法能否在不降低 DINA 诊断效果的同时提高计算速度. 由于 DINA 模型输出的学生认知诊断结果是一个不可实际测量的变量,本文将其转化为学生在试题上的作答预测来进行评估. 在计算时间性能方面,本文通过对比 4 种模型达到收敛时所用的时间加以验证评估.

实验中,对每个数据集,抽取其中 20% 的学生和题目作为测试集. 剩余部分作为训练集. 本文采用精准度(accuracy)来评估学生答题预测实验的结果,精准度定义如下:

$$\text{accuracy} = \frac{\text{succ}}{\text{total}} \tag{9}$$

其中 succ 表示预测结果和实际结果相符的记录个数,total 表示测试集中所有的记录个数,则精准度表示测试集中预测结果和实际相符的记录个数所占的比例.

为了保证实验结果的可靠性,实验均采用 10 次 5 折交叉验证法. 实验结果如表 4 和图 5 所示.

表 4　模型精准度(accuracy) 记录表

数据集 \ 模型	DINA	I-DINA	ME-DINA	IME-DINA
数据集 1	0.8640	0.8609	0.8658	0.8670
数据集 2	0.8636	0.8636	0.8595	0.8619

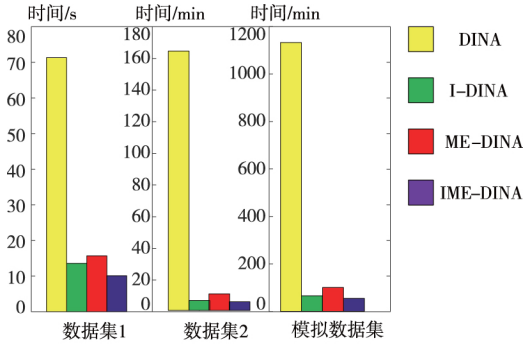

图5　模型计算时间对比图

由实验结果可知:

(1) 四种 DINA 模型在学生试题答题预测结果上很接近,不会影响到诊断效果. 同时,三种加速方法对比 DINA 模型,计算速度都有极大地提升. 且面对十万级别的学生数据表现同样出色.

(2) 对于不同的数据集,三种加速方法的加速能力也不同. 其中 IME-DINA 模型由于结合了其它两种方法的优点,可以达到最快的计算速度.

5.3　三种加速方法的参数设置实验

本节将讨论参数对三种方法加速能力的影响.

I-DINA 模型的计算速度受学生块数量的影响. 学生块数量太少时每次迭代时间仍较长,太多又会导致每次迭代改变太小. 本实验中采用确定每个学生块内的人数的方法进行学生块划分,实验指标为计算时间. 实验结果如图 6 所示. 从图 6 可知,数据集 1(数据集 2) 大约在每个学生块内 12(17) 人时达到最低计算时间.

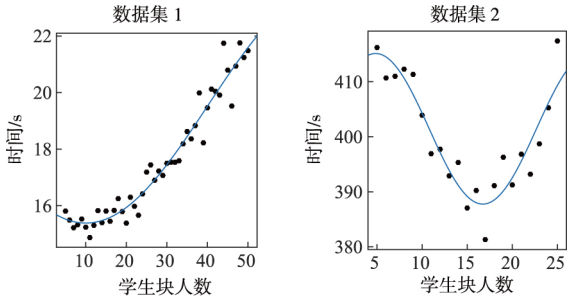

图6　I-DINA模型中学生块的人数对时间的影响

实验中 I-DINA 模型在时间上的波动性非常明显,同样设置下最慢记录可以达到最快记录 3 倍以上的时间,这可能会限制 I-DINA 模型的实际应用.

ME-DINA 模型的计算速度则受阈值 ST_1 和 ST_2 的影响. 如果 ST_1 取值太低或 ST_2 取值太高,就无法对数据做出有效地筛选,反之则会造成过度筛选. 在实验中,本文令 $ST_2 = 1 - ST_1$,故只需调节 ST_1 的值即可,图 7 为实验结果,其中横坐标为 ST_1 的值. 从图 7 可知,数据集 1(数据集 2) 大约在 $ST_1 = 0.25$(0.18) ,$ST_2 = 0.75$(0.72) 时达到最低计算时间.

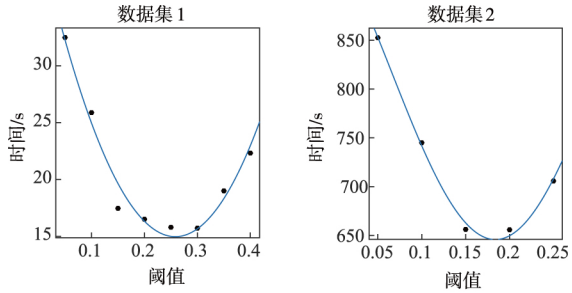

图7　ME-DINA模型中阈值对时间的影响

IME-DINA 模型的计算速度会同时受学生块人数和阈值影响,因此需要同时讨论这两个参数. 图 8 为固定 $ST_1 = 0.1$,$ST_2 = 0.9$ 时,模型计算时间与每个学生块内人数之间的关系图. 数据集 1 和数据集 2 大约都在每个学生块 8 个人时计算时间最少. 图 9 为固定每个学生块 10 个人时,模型计算时间与每个阈值的关系图,横坐标为 ST_1 的值. 数据集 1 和数据集 2 大约都在 $ST_1 = 0.1$ 到 0.15,即 $ST_2 = 0.85$ 到 0.9 之间计算时间最少. 相比 I-DINA 模型,IME-DINA 模型由于结合了 ME-DINA 模型的筛选机制,在计算时间上的稳定性大大提高,不会出现像增量加速算法中出现的大量偏差,在 90% 的实验次数里,时间的波动幅度均小于平均值的 10% .

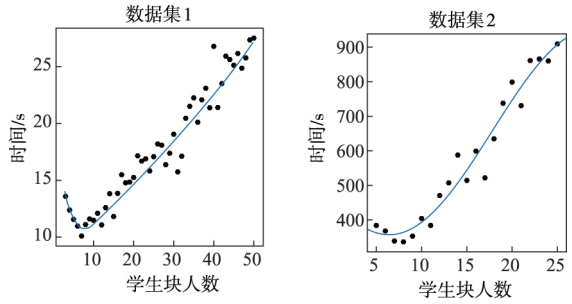

图8　IME-DINA模型中学生块的人数对时间的影响

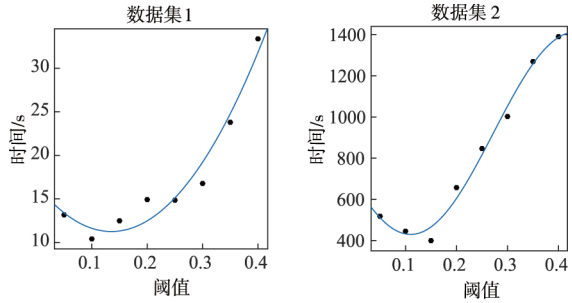

图9　IME-DINA模型中阈值对时间的影响

5.4 实验结果讨论

从诊断准确性的对比实验可以看到,本文所提出的三种加速方法均没有降低 DINA 模型的诊断效果.而在计算时间性能的对比实验中,三种加速方法均在三个数据集上达到了出色的效果.说明了本文所提出的三种加速方法均可以适用于实际应用

中可能会出现的数据集规模,如此大大提高了 DINA 模型的实际应用价值.

本文提出的三种加速方法中,IME-DINA 模型由于结合了其它两种方法的优点,能够达到最快的加速效果,且 IME-DINA 模型的参数选择更加稳定.故而在实际应用中,推荐采用 IME-DINA 模型进行学生的个性化认知诊断,阈值可以设置为下阈值在 0.1 到 0.15 之间,上阈值在 0.85 到 0.9 之间,而学生块的数目需要视数据集的具体规模而定.

6 结论和展望

DINA 模型在教育领域上应用价值很大,然而模型计算时间过长的问题阻碍了其实际应用.本文针对 DINA 模型中 EM 步骤里 E 步计算量大,耗费时间长的问题,提出了三种基于 E 步的加速方法.第一种是 I-DINA 模型,将学生数据划分为多个学生块并且循环访问;第二种是 ME-DINA 模型,只访问在极大化模型熵的过程中对迭代影响较大的学生数据;第三种是 IME-DINA 模型,这是一种基于前两种方法的混合方法.接着本文通过对比实验证明了三种加速方法均可以在不降低传统 DINA 模型诊断效果的情况下,提升几倍至几十倍的计算速度,同时首次给出了 DINA 模型及三种加速方法的理论收敛性证明.最后通过实验讨论了不同参数对三种加速方法造成的影响.

为了继续提高认知诊断模型在知识点层面上个性化评估学生学习状态的效果,未来研究工作可以从以下几个方面进行:(1)随着在线教育数据量的不断增大,未来可以考虑设计分布式的 DINA 算法,以进一步保证 DINA 模型的效率;(2)本文给出的三种方法,均不局限于 DINA 模型,未来的工作里,还可以尝试应用于其它基于 EM 算法的认知诊断模型.

参考文献

[1] Premchaiswadi W, Porouhan P. Process modeling and decision mining in a collaborative distance learning environment [J]. Decision Analytics, 2015, 2(1): 1-34.

[2] 康叶钦. 在线教育的"后 MOOC 时代"——SPOC 解析 [J]. 清华大学教育研究, 2014, 35(1): 85-93.

[3] Anderson A, Huttenlocher D, Kleinberg J, et al. Engaging with massive online courses [A]. Proceedings of the 23rd International Conference on World Wide Web [C]. ACM, 2014. 687-698.

[4] Vukicevic M, Jovanovic M Z, Delibasic B, et al. Recommender System for Selection of the Right Study Program for Higher Education Students [M]. RapidMiner: Data Mining Use Cases and Business Analytics Applications, 2013.

[5] Baker R S, Inventado P S. Educational Data Mining and Learning Analytics [M]. Learning Analytics, 2014. 61-75.

[6] Leighton J P, Gierl M J. Cognitive diagnostic assessment for education: Theory and applications [J]. Journal of Qingdao Technical College, 2007, 45(4): 407-411.

[7] DiBello L V, Roussos L A, Stout W. 31A Review of cognitively diagnostic assessment and a summary of psychometric models [J]. Handbook of Statistics, 2006, 26: 979-1030.

[8] Barnes T. The Q-matrix method: Mining student response data for knowledge [A]. Proceedings of the AAAI-2005 Workshop on Educational Data Mining [C]. Pittsburgh, PA, 2005. 1-8.

[9] Junker B, Sijtsma K. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory [J]. Applied Psychological Measurement, 2001, 25(3): 258-272.

[10] Fan X. Item response theory and classical test theory: an empirical comparison of their item/person statistics [J]. Educational & Psychological Measurement, 1998, 58(58): 357-381.

[11] An X, Yung Y F. Item Response Theory: What it is and how you can use the IRT procedure to apply it [A]. Proceedings of the SAS Global Forum 2014 Conference [C]. SAS Institute, 2014. 364-2014.

[12] De La Torre J. DINA Model and parameter estimation: A didactic [J]. Journal of Educational and Behavioral Statistics, 2009, 34(1): 115-130.

[13] 张潇,沙如雪. 认知诊断 DINA 模型研究进展 [J]. 中国考试, 2013(1): 32-37.

[14] de la Torre J. Application of the DINA Model Framework to Enhance Assessment and Learning [M]. Self-Directed Learning Oriented Assessments in the Asia-Pacific, New York: Springer, 2012. 92-110.

[15] Torre J D L, Minchen N. Cognitively diagnostic assessments and the cognitive diagnosis model framework [J]. Psicología Educativa, 2014, 20(2): 89-97.

[16] Torre J D L, Douglas J A. Higher-order latent trait models for cognitive diagnosis [J]. Psychometrika, 2004, 69(3): 333-353.

[17] Wu R, Liu Q, Liu Y, et al. Cognitive modelling for predicting examinee performance [A]. Proceedings of the 24th

International Conference on Artificial Intelligence [C]. AAAI Press,2015. 1017 – 1024.

[18] Neal R M,Hinton G E. A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants [M]. Learning in Graphical Models. Springer Netherlands, 1998. 355 – 368.

[19] Mu J. Handbook of modern item response theory [J]. Journal of the American Statistical Association, 1997, (92): 245 – 256.

[20] 朱天宇,黄振亚,陈恩红,等. 基于认知诊断的个性化试题推荐方法 [J]. 计算机学报,2017,40(1): 176 – 191.

[21] Templin J L,Henson R A. Measurement of psychological disorders using cognitive diagnosis models [J]. Psychological Methods,2006,11(3): 287 – 305.

[22] De La Torre J. The Generalized DINA model framework [J]. Psychometrika,2011,76(2): 179 – 199.

[23] Dempster A. Maximum likelihood from incomplete data via the EM algorithm [J]. Journal of the Royal Statistical Society,1977,39(1): 1 – 38.

[24] Wu C F J. On the convergence properties of the em algorithm [J]. Annals of Statistics,1982,11(1): 95 – 103.

[25] Booth J G, Hobert J P. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm [J]. Journal of the Royal Statistical Society,1999,61(1): 265 – 285.

[26] Meng X L,Rubin D B. Maximum likelihood estimation via the ECM algorithm: a general framework [J]. Biometrika,1993,80(2): 267 – 278.

[27] Liu C,Rubin D B. The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence [J]. Biometrika,1994,81(4): 633 – 648.

[28] Mclachlan G J. On Aitken's Method and other approaches for accelerating convergence of the em algorithm [A]. Proceedings of the AC Aitken Centenary Conference [C]. Dunedin: University of Otago Press (1995), 1998. 201 – 209.

[29] Bo T,Meek C,Heckerman D. Accelerating EM for large databases [J]. Machine Learning, 2001, 45(3): 279 – 299.

作者简介

王　超　男,1995 年生于安徽淮南. 中国科学技术大学计算机科学与技术学院硕士研究生,研究方向为机器学习、推荐系统.
E-mail: wdyx2012@ mail. ustc. edu. cn

刘　淇　男,1986 年生于山东临沂,博士,副教授,研究方向为数据挖掘与知识发现、机器学习方法及其应用.
E-mail: qiliuql@ ustc. edu. cn

陈恩红(通信作者)　男,1968 年生于安徽宣城,博士,教授,博导,国家杰出青年基金获得者,IEEE 高级会员,研究方向为机器学习、数据挖掘、个性化推荐系统.
E-mail: cheneh@ ustc. edu. cn