利用基本信息和行为数据发现高校贫困学生

聂 敏，张 杨，邓 辉，王 伟，夏 虎，周 涛*

(电子科技大学大数据研究中心 成都 611731)

【摘要】高校学生的扶贫助困工作一直是教育各界关注的重点，如何利用有效的大数据分析手段减轻评审工作量和公平化评审流程，从而实现高校精准扶贫的目标，是一项值得深入研究的问题。该文以高校学生行为数据为基础，结合高校数据的时序性特点，抽取学生基本信息和行为数据的多维特征，提出基于深度学习理论的 CW-LSTM 算法进行预测。最后使用真实数据对模型进行验证，结果显示，该方法优于朴素贝叶斯算法和决策树算法。

关 键 词 大数据； 数据挖掘； 家庭贫困学生； 学生行为数据

中图分类号 TP391 **文献标志码** A doi:10.12178/1001-0548.2020139

Identifying Poor Students in Universities by Using Basic Information and Behavioral Data

NIE Min, ZHANG Yang, DENG Hui, WANG Wei, XIA Hu, and ZHOU Tao*

(Big Data Research Center, University of Electronic Science and Technology of China Chengdu 611731)

Abstract The poverty alleviation work for college students has always been the focus of attention in education. How to use effective big data analysis methods to reduce the workload of review and fair review process and achieve the goal of targeted poverty alleviation in colleges and universities is a question worthy of further study. Based on the behavioral data of college students, this paper combines the time-series characteristics of college data, extracts the basic information and multi-dimensional features of behavioral data, and proposes a clockwork long short-term memory (CW-LSTM) algorithm based on deep learning theory for prediction. Finally, the model is verified using real data, and the results show that our method is better than the Naive Bayes algorithm and decision tree algorithm.

Key words big data; data mining; poor family students; student behavioral data

近年来，高校领域的大数据应用研究工作越来越受到各方关注[1-16]。为了评判学生在校期间的表现，文献 [5] 在 2012 年率先将数据挖掘技术应用于高校数据。2014 年，文献 [6] 继续深入研究了这个方向，将更多的数据用于评判学生的学业。后续，学者利用大数据分析手段，继续深入研究了学生行为对成绩或职业的影响[7-15]。这些研究都将目的定位于学生学业或职业选择，未关注学生家庭的经济情况。高校学生的培养，一直是国家和社会高度关注的。在培养高校人才的战略中，每年的教育支出也在逐步上涨。其中，相当一部分的支出会用于家庭贫困的学生，以帮助其顺利完成学业。目前高校对于家庭贫困学生的认定工作存在着不少漏洞，过程也非常繁琐低效，没有达到精准资助的要求。在当下的大数据时代，如何利用多维学生数据分析学生的家庭贫困信息是非常有必要的。

本文以学生行为数据为基础，利用大数据挖掘的相关技术，构建了家庭贫困学生挖掘算法，为高校扶贫工作提供支持。所谓家庭贫困学生挖掘，即基于学生在学校中的消费数据和其他行为数据，预测其家庭经济条件：是否存在困难。根据高校学生数据的维度丰富和时序性特点，本文抽取了学生基本信息的统计特征和行为数据的时序性特征，提出了深度学习算法 (clockwork recurrent neural network, CW-RNN) 的改进方法 CW-LSTM，用于评估学生的各维度特征，综合判定其经济条件。最后，本文

收稿日期：2020－02－19；修回日期：2020－05－15
基金项目：国家自然科学基金 (11975071，61433014)
作者简介：聂敏 (1989－)，男，博士，主要从事教育大数据方面的研究.
通信作者：周涛，E-mail：zhutou@ustc.edu

利用某高校 2011～2014 级学生在 2012 年−2015 年产生的数据进行分析，验证了本文方法的有效性。

1 CW-LSTM 算法框架

神经网络结构已经应用在 AI 领域的各个方面，在研究之初，为了将以往的信息连接到当前的任务中，研究者在网络结构中引入了循环结构，即 RNN。其计算方式为：

$$s_t = f_s(\boldsymbol{W}s_{t-1} + \boldsymbol{W}_{\text{in}}x_t) \tag{1}$$

$$o_t = f_{\text{o}}(\boldsymbol{W}_{\text{out}}s_t) \tag{2}$$

式中，x是输入；$\boldsymbol{W}_{\text{in}}$为输入层矩阵；$\boldsymbol{W}$是隐藏层矩阵；$\boldsymbol{W}_{\text{out}}$为输出层矩阵；$s$是隐藏层输出；$o$是输出层输出；$f_s$为隐藏层激活函数；$f_{\text{o}}$为输出层激活函数。通过$s_{t-1}$～$s_t$的循环结构实现信息的复用。但是 RNN 网络仅能记忆短期信息，对于长时间序列，会造成信息丢失。为了解决这样的信息丢失，文献 [17] 提出了改进的算法——CW-RNN。CW-RNN 将隐含层分为多个模块，并对每个模块设定时间频率，以便每个模块的单独管理。在每个模块内部进行全连接，在模块间进行高时钟频率模块向低时钟频率模块的连接，如图 1 所示。Hidden 表示隐藏层。在隐藏层中，多个模块的时间频率为$T_1,T_2\cdots,T_g$。体现在公式中为：将$\boldsymbol{W}$与$\boldsymbol{W}_{\text{in}}$分为$g$块。

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_1 \\ \boldsymbol{W}_2 \\ \vdots \\ \boldsymbol{W}_g \end{bmatrix}, \quad \boldsymbol{W}_{\text{in}} = \begin{bmatrix} \boldsymbol{W}_{\text{in}_1} \\ \boldsymbol{W}_{\text{in}_2} \\ \vdots \\ \boldsymbol{W}_{\text{in}_g} \end{bmatrix} \tag{3}$$

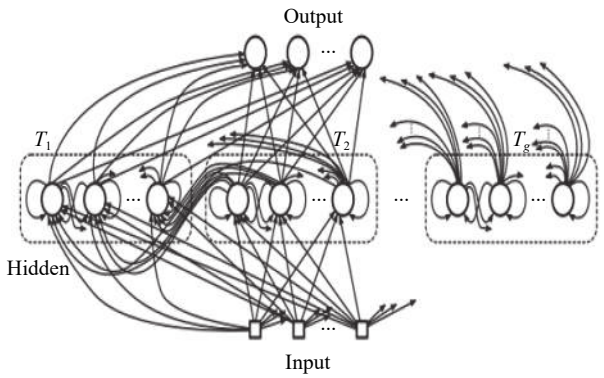

图 1 CW-RNNs 网络结构

在运算的时候，只会有部分模块参与运算，不参与运算的模块就置为 0，实现了对长短时间的处理。

LSTM 也可以部分解决 RNN 的长时间序列信息丢失问题[18]。在 LSTM中，每个神经元都是一个细胞，在每个细胞中，都包含存储器和 3 个门：输入门、输出门和遗忘门。输入门决定了哪些新的输入信息加入到存储器，遗忘门决定了从存储器中丢失哪些信息，输出门决定了每个状态的输出值。其单一神经元的结构如图 2所示。其中，x_t表示t时刻的输入，s_{t-1}表示$t-1$时刻的输出，h_{t-1}表示$t-1$时刻的细胞状态，s_t表示t时刻的输出，h_t表示t时刻的细胞状态。

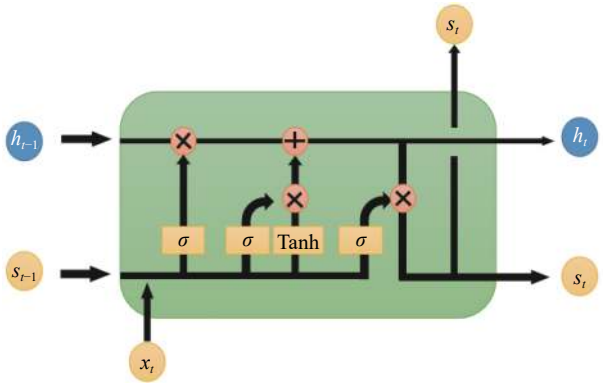

图 2 LSTM 的细胞结构示意图

在每个细胞中，首先计算遗忘门：

$$f_t = \sigma(\boldsymbol{W}_f[s_{t-1}, x_t] + b_f) \tag{4}$$

式中，σ是 sigmoid 激活函数，具体表示为：

$$\sigma(x) = \frac{1}{1+\text{e}^{-x}} \tag{5}$$

$\boldsymbol{W}_f$是遗忘门的权重矩阵；b_f是遗忘门的偏置。然后计算输入门：

$$i_t = \sigma(\boldsymbol{W}_i[s_{t-1}, x_t] + b_i) \tag{6}$$

$$g_t = \tanh(\boldsymbol{W}_g[s_{t-1}, x_t] + b_g) \tag{7}$$

式中，tanh是tanh激活函数，具体表示为：

$$\tanh(x) = \frac{\text{e}^{x}-\text{e}^{-x}}{\text{e}^{x}+\text{e}^{-x}} \tag{8}$$

$\boldsymbol{W}_i$、$\boldsymbol{W}_q$都是权重矩阵；b_i、b_q都是偏置。通过式 (4)～式 (6) 可以更新细胞状态为：

$$h_t = f_t h_t - 1 + i_t g_t \tag{9}$$

最后计算输出门：

$$o_t = \sigma(\boldsymbol{W}_{\text{o}}[s_{t-1}, x_t] + b_{\text{o}}) \tag{10}$$

$$s_t = o_t \tanh(h_t) \tag{11}$$

式中，$\boldsymbol{W}_{\text{o}}$是输出门权重矩阵；$b_{\text{o}}$是输出门偏置。模型最终训练的就是所有的权重矩阵和偏置。

CW-RNN 网络的设计简单，层次清晰，但其表达能力不强，容易出现高偏差的情况。而 LSTM

算法结构复杂，表征能力强，但是其参数多，训练复杂度高，有些超参数 (即不能通过训练得到的参数值，如网络隐藏层数、迭代轮数等) 需要人工提前配置，如果超参数设置不合理，其性能也会受到较大影响。为了结合两种算法各自的优点，本文提出两种算法的融合算法——CW-LSTM。CW-LSTM 算法保留 LSTM 中的输入门和输出门，而对于其处理长时间依赖的遗忘门，使用 CW-RNN 网络的多模块管理和高时钟频率模块向低时钟频率模块里的连接来实现。

在 CW-LSTM 算法中，每个存储块中包含存储器、输入门和输出门。对每个存储块内部按照 CW-RNN 网络的方式进行构建，将存储器设置为多个，并且配置不同的时钟频率，然后进行分组管理，不同存储器之间由高时钟频率向低时钟频率进行连接。图 3 展示了单个存储块的结构，其构建了一个 4 个周期的 CW-LSTM 存储块。利用多个这样的存储块，就可以构建 CW-LSTM 网络。对于 CW-LSTM 的计算，输入门和输出门的计算方式与 LSTM 一样，对于状态的管理，和 CW-RNN 一样，将状态权重矩阵分为 g 个模块，运算的时候只有高时钟频率向低时钟频率的连接模块才会进行计算。

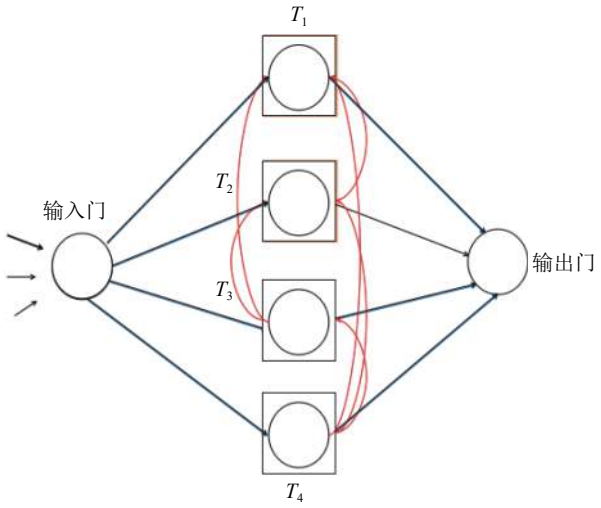

图 3　CW-LSTM 存储块结构

本文也对 3 种网络结构的训练参数个数和效率进行了计算。假设 CW-RNN、LSTM 和 CW-LSTM 3 种网络的隐藏层数都为 M，对于 CW-LSTM 和 CW-RNN，周期为 R，每个分组内节点数量为 N，则有 $M=RN$。用 O 表示网络中需要训练的参数个数，3 种网络表示为：

$$\begin{cases} O_{\text{LSTM}} = M^2 \\ O_{\text{CW-RNN}} = \dfrac{3MN-N}{2} \\ O_{\text{CW-LSTM}} = M\dfrac{R+1}{2} + N^2 \end{cases} \tag{12}$$

可以看出，3 种网络结构的时间复杂度都为 $O(M^2)$，CW-LSTM 网络计算效率介于两者之间。

2　基于 CW-LSTM 的家庭贫困学生挖掘模型

针对高校学生统计数据的丰富维度和行为数据的时序性特点，本文针对性地抽取了多个特征进行研究。最后将处理好的特征输入到 CW-LSTM 模型中进行贫困预测。所有的数据均是在匿名的条件下采集和试用。

2.1　特征提取

基本统计特征是利用数理统计技术获取的一些基本特征。在学生基本信息上，本文考虑性别、生源地、民族和年级 4 个维度的特征。在消费数据中，根据获得的数据分布以及学生在校期间的消费范围，将消费数据分为食堂消费数据和其他消费数据，食堂消费数据包含早餐、中餐、晚餐和宵夜，其他消费数据包含超市、洗澡、洗衣等非食堂消费。另外，再将消费数据细分为消费次数、消费平均值和最大值。还提取了其他数据特征，如图书馆门禁、寝室门禁、成绩和寒暑假留校情况。

抽象特征的构建是结合家庭贫困学生挖掘的目标和相关业务人员的工作经验所提出的。主要包括规律性和朋友圈经济水平。规律性可以通过一个人特定时段间隔行为发生的熵来描述。假设时间间隔为 n，即 $T=\{t_1,t_2\cdots,t_n\}$，任何一个学生的行为在 t_i 时间间隔发生的概率的计算公式为：

$$P_v(T=t_i)=\frac{n_v(t_i)}{\sum n_v(t_i)} \tag{13}$$

式中，$n_v(t_i)$是行为v在时间间隔t_i内发生的频率。则行为v的熵为：

$$E_v=-\sum_{i=1} P_v(T=t_i)\lg P_v(T=t_i) \tag{14}$$

一种行为的熵越高，那么该行为在不同时间段内发生的概率越不均匀，也就是这个行为的规律性较低。在本文的研究中，考虑了食堂就餐、非食堂消费和去图书馆这 3 种行为的熵。

对于现在的高校学生，朋友圈能够反应相当多

的信息，而一个人的经济水平可能会与其朋友圈平均经济水平相关。首先，引入亲密度的概念，其表示两个人的关系密切程度。然后计算任意两个学生的亲密度$R_A(B)$，设置阈值 H，认为与 A 亲密度大于 H 的同学 $B(R_A(B)>H)$ 就是 A 的朋友。以此构建朋友圈。对于亲密度，可以通过两个人在某一时间段内同时出现在相同地点的次数来计算，并且不同的刷卡场景需要有不同的权重。学生 A 与学生 B 在时间周期 T 内的亲密度计算公式为：

$$R_A(B)=\sum_{i\in L}\left[\frac{R_A^i(B)}{C_A(i)}\frac{|S|}{S_A(i)}\right] \tag{15}$$

式中，L表示所有的刷卡地点；$C_A(i)$表示在时间周期T以内，学生 A 在地点i的总刷卡次数；$R_A(B)$表示在时间周期 T 内，学生 A 与学生 B 在地点 i 的共同出现次数；$|S|$表示学生总数；$S_A(i)$表示与学生 A 在地点 i 共同出现的总人数。可以看出，亲密度是有向的，A 对于 B 的亲密度很高并不意味着 B 对于 A 的亲密度就一定很高，即在式 (15) 中$R_A(B)\neq R_B(A)$。基于式 (15)，可以计算任意两个学生 A 和 B 的亲密度 $R_A(B)$，并设定阈值 H，认为满足 $R_A(B)>H$要求的学生 B 是 A 的朋友——这样就可以得到学生 A 的朋友圈。接下来通过学生朋友圈中获得过助学金的学生数量，以及该学生的朋友数量来定义朋友圈经济水平F_A，有：

$$F_A=\frac{P_A^2}{N_A} \tag{16}$$

式中，N_A代表学生 A 的朋友总数；P_A代表 A 的朋友中家庭贫困的朋友数。

2.2 特征选择和模型结构

本文进行了特征的提取，但是提取出的特征并不都是有用的，这主要是因为，有些特征非常稀疏，不利于后序的计算。还有些特征之间具有很强的关联性，导致多种特征只需要其中一种或几种就能够达到想要的结果。因此有些特征就变得冗余了，需要进行特征选择。本文采用后剪枝的 C4.5 算法进行特征选择，即首先将数据划分为训练集和验证集，在训练集上用 C4.5 算法生成决策树，然后进行剪枝。具体操作为：对每一个非叶子节点来说，删除以此节点为根节点的子树，让这个节点变为叶子结点，该叶子节点对应的类别为相应训练数据中占优的类别。如果这样操作在验证集上的准确率没有比原来的差，就将此节点设置为叶子节点，删除此节点以下的所有特征。

在经过特征抽取和特征选择后，得到了高校学生数据的一系列特征。将得到的特征按照{月, 学期}的时间周期进行分组，然后将其输入到 CW-LSTM 的不同分组中，完成算法的输入层构建。在隐藏层中，构建全连接网络，并且网络的神经元数量与输入层相同。最后，在输出层设置一个输出神经元，其内部不同周期的存储器表示不同时间周期的预测结果，然后通过不同的权重连向输出门，得到最终的预测结果。

3 实验结果

以某高校 2011～2014 级学生为例，对家庭贫困学生挖掘模型进行验证。获取到学生基本信息数据 32 318 条，消费数据约 1.6 亿条，图书馆门禁数据 1 400 余万条，寝室门禁数据 2 800 余万条，成绩数据近 200 万条，助学金信息数据 8 889 条。在式 (15) 中，时间周期 T 统一取为一月。在亲密度计算中，将阈值 H 设置为 0.35。本文将所有的数据随机划分为训练集 (80%) 和测试集 (20%)，在特征选择阶段，选用后剪枝的 C4.5 算法获得的有用特征如表 1 所示。

表 1 剪枝后的特征

特征	类型	内容
基本统计量特征	基本信息	性别
		生源地
		民族
	消费信息	食堂消费最大值
		食堂消费平均值
		食堂消费次数
		其他消费次数(如洗衣、洗澡、体育场、图书馆等)
		其他消费最大值
		其他消费平均值
	图书馆门禁	泡馆次数
		泡馆时长
抽象特征	行为熵	食堂消费
		非食堂消费
		去图书馆
	朋友圈	朋友圈经济水平指数

得到最终的特征后，将助学金信息作为训练标签，即认为获得助学金的学生为家庭贫困学生，没有获得助学金的学生为家庭非贫困学生，共有家庭贫困学生 20 070 名，家庭非贫困学生 18 731 名。对于测试数据的所有特征，将其输入到 2.2 节所述的 CW-LSTM 模型结构中，设置迭代轮次为

1 000。在模型对比中，本文选择朴素贝叶斯算法和 C4.5 决策树算法。

由于本文采用的是回归算法，最后模型的输出结果是一个连续值，表示这个学生属于家庭贫困学生的概率。本文将这个概率从大到小排序，取前 f 的样本，作为预先设定为家庭贫困学生的人数占比。准确率即为前 f 样本中的确是家庭贫困的学生比例。当 f 较小时，表示仅取预测为家庭贫困学生概率较大的样本，因此其准确率比较高。从图 4 可以看出，当 f>0.1 时，CW-LSTM 算法的准确率优于朴素贝叶斯算法和决策树算法。

对于分类问题，AUC 值也是一个常见的评价指标，即 ROC(receiver operator curve) 曲线下的面积[18-19]。本文也对家庭贫困学生分类问题的 AUC 值进行了计算。结果显示，朴素贝叶斯算法的 AUC 值为 0.64，决策树算法的 AUC 值为 0.652，而本文提出的 CW-LSTM 算法的准确率为 0.659，同样也说明 CW-LSTM 算法的效果是要优于决策树算法和朴素贝叶斯算法的。

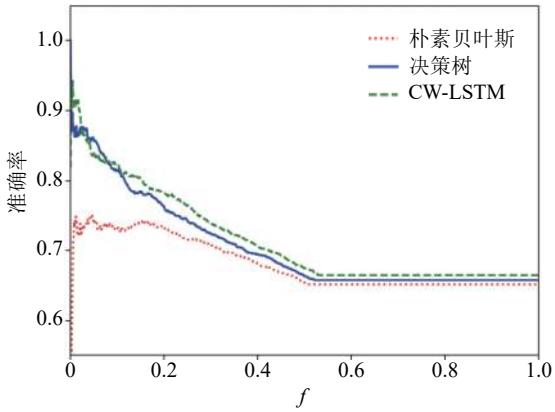

图 4　准确率随结果集比例 f 的变化

另外，通过决策树模型，本文对特征的重要性进行分析，如图 5 所示。从结果可以看出，与消费有关的数据在预测中有着至关重要的作用，靠前的特征都与消费有关，这主要是因为预测目标就是学生经济水平。在消费数据中，食堂消费数据更加重要，其平均值、最大值和次数的重要性都要高于其他消费数据。另外，提出的抽象数据特征也有着非常重要的作用，消费行为的熵和朋友圈经济水平两者的重要性之和超过了 30%。其他统计特征重要性要远低于消费数据。

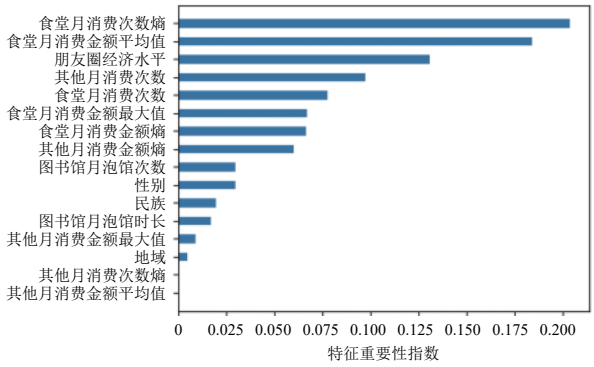

图 5　特征重要性指数

4　结 束 语

高校学生一直是国家和社会关注的焦点，本文利用大数据分析技术，对高校的家庭贫困学生进行挖掘。针对于高校学生的数据特点，抽取了学生的统计特征和行为时序性特征。然后根据学生数据的时序性特点，综合 CW-RNN 和 LSTM 的优点，提出了 CW-LSTM 算法来处理高校学生数据。最后利用某高校的真实学生数据，对模型进行了验证。模型整体结果比朴素贝叶斯算法、决策树回归算法好。本文的研究方法可用于其他教育大数据分析，如毕业去向预测。本文的研究结果为实现高校更加精准扶贫工作提供了理论依据和确实可行的实验方法，还可以在保证准确性的同时提高评审效率。

参 考 文 献

[1] BAUM S, SCHWARTZ S. Student aid, student behavior and educational attainment[M]. London: Routledge, 2015.

[2] LIM V K G, TEO T S H. Sex, money and financial hardship: An empirical study of attitudes towards money among undergraduates in Singapore[J]. J Econ Psychol, 1997, 18(4): 369-386.

[3] ANDREWS B, WILDING J M. The relation of depression and anxiety to life-stress and achievement in students[J]. Br J Psychol, 2004, 95(4): 509-521.

[4] ROBERTS R, GOLDING J, TOWELL T, et al. The effects of economic circumstances on British students' mental and physical health[J]. J Am Coll Health, 1999, 48(3): 103-109.

[5] YADAV S K, BHARADWAJ B, Pal S. Data mining applications: A comparative study for predicting student's performance[EB/OL]. (2012-02-22). https://arxiv.org/abs/1202.4815.

[6] VEERAMUTHU P, PERIASAMY R. Application of higher education system for predicting student using data mining techniques[J]. Int J Inn Re Adv En, 2014, 1(5): 36-38.

[7] YAO H, LIAN D, CAO Y, et al. Predicting academic performance for college students: A campus behavior perspective[J]. ACM TIST, 2019, 10(3): 1-21.

[8] CAO Yi, GAO Jian, ZHOU Tao. Orderliness of campus

lifestyle predicts academic performance: A case study in Chinese university[M]//BAUMEISTER H, MONTAG C. Digital Phenotyping and Mobile Sensing: Studies in Neuroscience, Psychology and Behavioral Economics. Cham: Springer, 2019.

[9] CAO Y, GAO J, LIAN D, et al. Orderliness predicts academic performance: Behavioural analysis on campus lifestyle[J]. J R Soc Interface, 2018, 15(146): 20180210.

[10] LIAN D F, LIU Q. Jointly recommending library books and predicting academic performance: A mutual reinforcement perspective[J]. Journal of Computer Science and Technology, 2018, 33(4): 654-667.

[11] NIE M, YANG L, SUN J, et al. Advanced forecasting of career choices for college students based on campus big data[J]. Front Comput Sci, 2018, 12(3): 494-503.

[12] ZHU T, LIU Q, HUANG Z, et al. MT-MCD: A multi-task cognitive diagnosis framework for student assessment[C]// Inter Conf Data Sys Adv App. Cham: Springer, 2018: 318-335.

[13] YAO H, NIE M, SU H, et al. Predicting academic performance via semi-supervised learning with constructed campus social network[C]//Inter Conf Data Sys Adv App. Cham: Springer, 2017: 597-609.

[14] LIAN D, YE Y, ZHU W, et al. Mutual reinforcement of academic performance prediction and library book recommendation[C]//2016 IEEE 16th International Conference on Data Mining (ICDM). [S.l.]: IEEE, 2016: 1023-1028.

[15] 罗清红. 数据, 大数据与教育大数据[J]. 教育科学论坛, 2016(10): 7-9.
LUO Qing-hong. Data, big data and education big data[J]. Education Science Forum, 2016(10): 7-9.

[16] KOUTNIK J, GREFF K, GOMEZ F, et al. A clockwork rnn[EB/OL]. (2014-02-14). https://arxiv.org/abs/1402.3511.

[17] GREFF K, SRIVASTAVA R K, KOUTNÍK J, et al. LSTM: A search space odyssey[J]. IEEE Trans Neu Net Learn Sys, 2016, 28(10): 2222-2232.

[18] 刘建国, 周涛, 郭强, 等. 个性化推荐系统评价方法综述[J]. 复杂系统与复杂性科学, 2009(3): 5-14.
LIU Jian-guo, ZHOU Tao, GUO Qiang, et al. Overview of the evaluated algorithms for the personal recommendation systems[J]. Complex Systems and Complexity Science, 2009(3): 5-14.

[19] 朱郁筱, 吕琳媛. 推荐系统评价指标综述[J]. 电子科技大学学报, 2012, 41(2): 163-175.
ZHU Yu-xiao, LÜ Lin-yuan. Evaluation matrics for recommender systems[J]. Journal of University of Electronic Science and Technology of China, 2012, 41(2): 163-175.

编 辑 蒋 晓