Overseas English Testing: Pedagogy and Research 国外英语考试教学与研究, 2020, 2(4), 167-174
Published Online November 2020 in Hans. http://www.hanspub.org/journal/oetpr
https://doi.org/10.12677/oetpr.2020.24016

Hans汉斯

数据挖掘技术在语言测试研究中的应用

王　萍[1]，辜向东[2*]

[1]西安欧尔意信息科技有限公司，陕西　西安
[2]重庆大学，重庆
Email: cristine_jobmail@163.com, *xdgu@cqu.edu.cn

收稿日期：2020年11月3日；录用日期：2020年11月16日；发布日期：2020年11月30日

摘　要

信息技术的发展给语言测试带来了新变化，也对语言测试研究方法提出了新要求。在大数据背景下，越来越多的语言测试学者尝试运用数据挖掘技术研究语言测试问题。为方便读者了解数据挖掘技术应用于语言测试研究的现状，本文首先介绍数据挖掘的基本概念、主要方法以及数据挖掘过程，然后重点介绍数据挖掘技术在语言测试研究中的应用现状，并按研究主题对相关文献进行分类讨论。最后，对数据挖掘技术应用于语言测试研究的启示、不足和未来的研究方向进行阐述。

关键词

数据挖掘技术，语言测试，跨学科研究，研究方法

Application of Data Mining Techniques in Language Assessment Research

Ping Wang[1], Xiangdong Gu[2*]

[1]OAE Publishing Incorporation, Xi'an Shaanxi
[2]Chongqing University, Chongqing
Email: cristine_jobmail@163.com, *xdgu@cqu.edu.cn

Received: Nov. 3rd, 2020; accepted: Nov. 16th, 2020; published: Nov. 30th, 2020

Abstract

The rapid development of information technology has brought changes to language assessment, which calls for incorporating new research methodologies in language assessment research. In the

*通讯作者。

文章引用: 王萍, 辜向东. 数据挖掘技术在语言测试研究中的应用[J]. 国外英语考试教学与研究, 2020, 2(4): 167-174.
DOI: 10.12677/oetpr.2020.24016

big data era, more and more researchers are trying to apply data mining techniques to language assessment research. For readers' better understanding of this interdisciplinary research area, this review first introduces what data mining is, including its concept, main methods and procedure; then reviews the literature classified in themes. In the end, implications and future research directions are discussed.

Keywords

Data Mining Techniques, Language Assessment, Interdisciplinary Research, Research Methodology

1. 引言

语言测试是语言教学中不可或缺的组成部分[1]，也是外语教育工作者一直关心的重要议题。近年来，在教育信息化、远程教育等背景下，信息技术对外语教育的影响越发显著[2]。信息技术在改变传统的课程体系和教学模式的同时，也对语言测试的形式和方法产生了诸多影响。从最初的基于计算机的考试、基于网络的考试到后来的作文自动评分系统的开发和应用，如文本自动评分工具 E-rater、文本智能改编、题库建设、自适应考试系统的设计与开发等。除了改变语言测试的形式和方法外，信息技术尤其是数据挖掘技术对语言测试领域的研究方法也有显著影响，越来越多的研究者开始尝试用数据挖掘技术来解决语言测试研究中的难题[3]。

本文将首先介绍数据挖掘的概念、任务以及数据挖掘过程，然后按主题分类讨论数据挖掘技术在语言测试研究中的应用，最后对大数据时代背景下数据挖掘技术应用于语言测试研究的未来进行展望。

2. 数据挖掘概述

2.1. 数据挖掘的概念与任务

数据挖掘是指从大量的数据中发现隐藏的模式和知识的过程[4]。数据挖掘兴起于上世纪 90 年代，融合了统计学、人工智能，尤其是机器学习技术和数据库技术[5]，目前已广泛应用于金融、医疗及其他科学领域[2]。

数据挖掘有两类任务，描述和预测。描述是指发现数据中存在的模式，使用户更容易理解数据[6]，预测是指用已知变量预测未知变量[7] [8]。上述两类数据挖掘任务常用的方法主要包括分类、聚类、关联规则挖掘等。分类的目的在于为对象划分一个类别[2] [5]，如将试题难度划分为高、中、低三个类别。常用的分类算法有决策树、贝叶斯网络、逻辑回归、人工神经网络和支持向量机等。聚类是将相似的对象划分到同一个簇，使得同一簇中对象之间的相似性最大，而不同簇中对象之间的相似性最小[6]。与分类不同，聚类所要划入的类别通常是未知的，即聚类过程是探索性的，对要划入的类别没有预设，常用的聚类算法有 K-means 算法[2]。关联规则挖掘旨在发现数据之间隐含的关联[8]，例如挖掘超市售货记录数据后，发现啤酒和尿不湿的销售额之间存在关联。常用的关联规则挖掘算法有 Apriori 算法、Brute Force 算法、Enumeration-Tree 算法等。

此外，对于不同的数据类型，还有其他的数据挖掘方法，如数据流挖掘、文本挖掘、时间序列挖掘、空间数据挖掘、网页挖掘和图数据挖掘等。

2.2. 数据挖掘过程

数据挖掘主要有数据准备、建模和模型评估三个步骤(见图 1)。

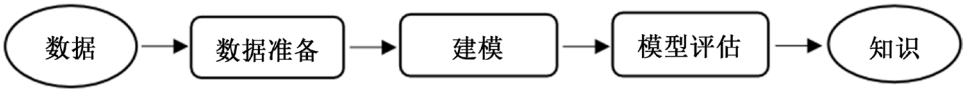

Figure 1. Data mining procedure
图 1. 数据挖掘流程

数据准备是指将原始数据转换为符合数据挖掘要求的数据。因为原始数据可能包含噪声、缺失值和不一致的数据，因此需要对数据进行转换以满足数据挖掘要求[2]。数据准备主要包括：数据清洗，即处理数据中的缺失值、噪声和不一致数据等的过程；数据归约，指压缩原始数据的大小以提高数据挖掘效率，常见的数据归约类型有数据取样、降维等；数据转换，即将数据转换成数据挖掘算法需要的类型[6]。

建模是指用模型展示对象之间的关系。数据挖掘模型主要分为描述类模型和预测类模型。前者旨在发现数据中已有的关系和规律，如近年来部分高校通过挖掘学生的校园卡消费数据，发现贫困学生，进而实现贫困补助的“精准”发放。后者则是通过已知变量预测和推断未知变量，如通过分析学生的学习行为数据，预测学生是否存在挂科的风险，进而对其进行教学干预。

模型评估是指评估模型的性能，选出最优模型[6] [9]。常用的模型评估指标有：准确率(accuracy)、召回率(recall)、精确度(precision)。此外，其他指标如 F-measure，AUC 等也可以用作模型评估的参考。

因数据挖掘技术存在诸多优点，如方法多样，大多数算法已相当成熟，可直接应用；能够高效处理大规模数据，并从中发现隐藏的规律和知识；以数据为驱动，无需对变量之间的关系做出预设，对数据的分布也没有严格要求等等，越来越多学者尝试将数据挖掘技术应用于语言测试研究。

3. 数据挖掘技术在语言测试研究中的应用

语言测试研究中对数据挖掘技术的应用尚处于探索阶段，研究文献数量相对较少，且主要集中于近几年。笔者以“数据挖掘 + 语言测试”、“数据挖掘 + 题目难度/文本难度”、“算法 + 语言测试/语言教学”、“人工智能 + 教育/教育测量/语言测试”等为关键词进行大范围搜索，然后逐篇筛选，最终得到 26 篇最相关的符合要求的文献。文献来源主要有：语言测试期刊和教育测量期刊、大规模考试研究报告、数据库等。

按研究主题可以将文献分为五类，分别是成绩预测、语言知识与语言技能的关系、题目难度预测、评分员行为研究和写作测试。

3.1. 成绩预测

成绩预测，即预测学生的考试成绩和考试表现，主要有两种类型:一是根据一项考试的成绩预测其他考试的成绩，如 Gurlitz [10]用托福口语成绩预测学生的 SPEAK (Speaking Proficiency English Assessment Kit)成绩；二是根据学生的背景信息和学习行为等，预测其能否通过某项考试，如 Meng *et al.* [11]根据学生的人口学信息、学习日志和绩点预测其专四考试成绩。成绩预测研究中用到的数据挖掘技术主要有回归分析和贝叶斯算法。

现有研究多集中于用托福考试成绩预测学生在 ITA (International Teaching Assistant)考核中的表现，如 Xi [12]为了验证国外高校将托福口语成绩作为考核 ITA 申请者的参考标准之一是否理据充分，用逻辑回归分析托福口语成绩对学生 SPEAK 成绩的预测力，结果显示托福口语成绩能够有效预测考生的 SPEAK 成绩，且能够对考生的能力水平作显著区分。

尽管托福口语成绩作为选拔 ITA 的参考具有一定的理据，然而由于考生群体多样性，考生的个体差异对选拔结果可能产生的影响有待研究。Gurlitz [10]用考生的托福口语成绩预测其 SPEAK 考试成绩，根据生源地将考生分为东亚组、印度半岛组和其他组，通过建立贝叶斯成绩预测模型。研究发现当托福口语成绩为 24 分时，无论考生来自哪个地区，均不能通过 SPEAK 考试的几率为 0.56，说明决策者在筛选 ITA 时应要求考生的托福口语成绩高于 24 分；当托福口语成绩高于 24 分时，东亚考生不能通过 SPEAK 考试的几率为 0.79，而印度半岛的考生则为 0.43，其他地区的考生为 0.15。有趣的是，该研究发现当托福口语成绩 ≥28 分时，印度半岛的学生 SPEAK 考试失败的几率最大。以上研究表明，将托福考试成绩尤其是口语成绩作为选拔 ITA 的参考时，应考虑学生的个体差异，如地域差异、文化和语言环境等。

除了总成绩以及口语成绩，有学者也研究了托福其他单项成绩对于ITA选拔的参考价值。如Mercado [13]研究了托福考试各单项成绩对于学生 SPEAK 考试成绩的预测作用。通过建立亚高斯贝叶斯成绩预测模型，该研究发现，在托福考试各单项成绩以及所有单项成绩的组合中，口语成绩的预测准确率最高，其次是听力成绩，说明在选拔 ITA 时，除参考总成绩外，还应综合考虑各单项成绩，尤其是口语成绩和听力成绩。

还有学者根据学生的课业成绩预测其某项考试的成绩，如 Meng *et al.* [11]以国内某高校英语专业本科生为研究对象，提取学生大学四年每门课的成绩、家庭经济状况、父母文化水平以及从学生大学四年记录的学习日志中反映学生学习状态的关键词，用以预测学生专四表现。该研究收集了问卷调查和学校教务管理系统中学生的成绩数据，分别用朴素贝叶斯算法、神经网络算法和逻辑回归这三种数据挖掘技术建立成绩预测模型，用预测准确率、精确度、召回率、F 值等指标评估模型，最终发现在所采用的三种数据挖掘技术中，朴素贝叶斯模型能够最准确地预测学生的专四成绩。该研究的创新性在于综合考虑了多种因素(如家庭经济条件、父母文化程度和学业成绩等)对专四考试成绩可能产生的影响。研究意义在于能及早发现可能无法通过考试的学生，并从多方面着手，如家庭经济条件、学生的心理状况(压力、焦虑)等对学生进行教学干预。

数据挖掘中的预测任务应用于成绩预测时，一方面，对于学业成绩的预测能够及早发现可能会挂科或学习困难的学生，从而建立预警机制，对学生进行有针对性的指导和帮助；另一方面，对于考试成绩的预测能够为相关政策的制定和人才选拔提供参考。

3.2. 语言知识与语言技能的关系

语言知识与语言技能的关系研究主要是分析语言知识，如词汇、语法等对听力、口语、阅读、写作以及翻译等语言技能的影响。语言测试中此类研究多使用回归分析、基于进化算法的符号回归分析、神经网络算法等数据挖掘技术，集中于分析学习者的词汇知识和语法知识对其阅读理解和听力理解能力的影响，如 Han [14]采用回归分析的方法研究词汇提取能力和工作记忆容量对阅读理解的影响。

然而有学者(Perkins *et al.* [15], Aryadoust [16])指出，人文社科领域往往涉及众多变量，且变量之间关系错综复杂，而线性回归方法预设自变量和因变量呈线性关系且自变量之间不存在交互影响，因此用线性回归的方法分析数据可能会得出不够准确的结果。此外，在人文社科领域，异常的数据本身往往包含重要信息，而线性回归方法往往忽略了这些异常数据[17] [18]。例如某学生的成绩远远低于平均值，在线性回归中这样的异常数据可能由于不具代表性而被忽略，然而对于人文社科领域的学者而言，进一步挖掘其异常成绩背后的社会学原因，如家庭和教育环境等，进而更全面地了解并帮助学生，可能更有价值和意义。因此，在人文社科研究中，仅仅用线性回归的方法是不够的，某些时候得出的结果甚至不准确，那么探索更适合用来解决问题的非线性的方法就显得尤为重要。

Aryadoust [16]将进化算法的思想融入回归分析中，用基于进化算法的符号回归研究认知策略的使用和词汇语法知识对听力理解的影响。结果显示，词汇知识和语法知识以及元认知策略中的计划和评估策

略、问题解决策略对听力理解有显著的预测作用。而使用线性回归得出的结论则显示语法知识和问题解决策略对听力理解有预测作用，但预测作用较弱。上述两种方法得出的结果差异表明，在建模之前应先检验数据，根据数据的实际情况采取合适的数据分析方法，即如果自变量和因变量呈非线性关系，则应采用非线性模型。

Aryadoust & Baghaei [19]用神经网络算法研究学生的词汇知识和语法知识与其阅读水平之间的关系。结果显示，训练的神经网络模型准确预测出了78%的学生阅读水平，表明词汇知识和语法知识对阅读理解有着重要影响。

将数据挖掘技术应用于分析语言知识和语言技能的关系，有助于深入理解语言学习和语言技能发展的本质，揭示语言加工过程，且数据挖掘技术中的很多非线性模型如神经网路模型、逻辑回归模型等因其自身的特点，在人文社科领域包括教育测量研究中具有很强的适用性。

3.3. 题目难度预测

题目难度预测主要在于发现影响题目难度的因素有哪些，进而通过操纵这些因素实现题目难度调控的目的。此类研究用到的数据挖掘技术主要有回归分析、分类回归树、神经网络算法、路径分析、自适应神经模糊推理系统等。Freedle & Kostin 在系列研究中[20] [21]，运用多元回归来考查影响题目难度的因素。主要思路是，对阅读/听力理解文本、多项选择题题干和选项中可能影响题目难度的变量进行编码和提取，用多元回归分析研究这些变量对题目难度变异的解释力，解释的难度变异越多，表明该组变量对难度的影响越大。

Perkins *et al.* [15]尝试用神经网络算法研究题目难度，考查了阅读文本特征，如所谈论的话题是否属于人文学科，段落数、平均每段的词数、阅读文本和多项选择题题干及选项的论元/修饰语/谓词数。结果显示，经过训练的神经网络模型能够较准确地预测题目难度。该研究探索了神经网络算法对题目难度预测的适用性，对后来的研究具有重要的启示和借鉴意义，如在 Perkins *et al.* [15]研究的基础上，韩菡[22]和付佩宣[23]用神经网络算法预测汉语考试阅读理解题目难度，发现经过训练的神经网络模型对题目难度预测准确率较高。然而，值得指出的是，Perkins *et al.* [15]的研究仍存在不足之处，例如其关于阅读理解所涉及的认知要求的界定过于简单，而且理论依据不足。Gao & Rogers [24]基于广泛的文献梳理，提出了一个相对全面且有充分理论依据的阅读理解认知加工模型，用基于决策树的回归分析方法探究影响阅读理解题目难度的认知因素有哪些，加深了学界对阅读理解认知加工过程的理解。

还有学者对比分析多种数据挖掘技术对题目难度预测的适用性，如 Aryadoust [25]对比分析了自适应神经模糊推理系统和路径分析两种方法对听力理解题目难度的预测。结果表明，自适应神经模糊推理系统在教育测量领域有很强的适用性，可以与路径分析方法相结合使用。Aryadoust & Goh [3]对比分析了回归分析、分类回归树、神经网络算法在听力理解多项选择题题目难度预测中的差异，发现人工神经网络算法的预测效果最佳。Rupp *et al.* [26]将多元线性回归与分类回归树结合起来研究影响听力理解和阅读理解多项选择题题目难度的文本因素、题目因素以及文本和题目的交互因素。研究表明，将上述两种方法结合使用有助于得到更全面更准确的结果。

题目难度调控是试题开发的重要环节，因为只有难度适宜的题目才能有效地测量出考生的真实水平，而要确保题目难度适宜，试题开发者需要不断试测和调整题目难度。数据挖掘技术能够从多方面、多角度入手，研究影响题目难度的多种因素。而用数据挖掘技术探索影响题目难度的因素，进而实现题目难度调控，将极大地减轻试题开发者的工作量，节省人力物力资源。

3.4. 评分员行为研究

评分员行为研究主要是通过问卷调查和访谈等了解评分员在评分过程中所侧重的方面，进而识别出

不同类型的评分员。此外，也有研究通过调查评分员的认识和实际评分行为，分析两者之间有没有偏差，研究评分员的认识如何影响其实际评分行为。

在此类研究中，聚类分析技术比较常用。如 Eckes [27]收集了评分员对作文进行评分的数据，先用 Rasch 模型检验评分员的评分行为是否存在显著差异，然后对评分员做聚类分析，结果显示共有六类评分行为，分别侧重词汇句法成熟度和写作任务完成度、单词拼写和语法运用准确规范、文章结构清晰完整、文章流畅通顺、句法多样且行文流畅、思路清晰且个人观点鲜明。Eckes [28]进一步研究了评分员的认识与其实际评分行为之间的关系，聚类分析结果显示，共有四种不同类型的评分员，且评分员的认识在很大程度上决定其评分行为，例如对于他们认为比较重要的语言点，在评分过程中会比较严格，相反，对于他们认为不太重要的语言点，在评分过程中就会适当宽松。

评分是测试的重要环节，与考生、考试研发者、用人单位等多方利益相关者密切相关，因此对于评分行为的研究意义重大。由于评分一定程度上属于个体化的行为，且不同的人对评分规则的解读可能存在差异，那么能否确保评分员在评分之前对评分规则有正确的理解，对评分规则的解读出现分歧时该如何协调。在实际评分过程中，评分员是否会严格按照评分规则执行，亦或会加入自己的评分经验、主观感受、个人观点等对评分规则进行“过滤”等等，都值得深入研究。数据挖掘中的聚类分析技术可以从大量评分行为数据中发现特有的评分规律和模式，具有一定的客观性和科学性，将极大地便利对评分结果的解读，对评分员进行培训，以提升评分准确性与可靠性。

3.5. 写作测试

数据挖掘技术应用于写作测试研究主要分为两类：一是研究写作的文本特征(词汇、语法、语篇)与写作质量之间的关系；二是关于作文自动评分系统的优化问题。写作测试研究中用到的数据挖掘技术主要有支持向量机、自动线性建模、基于遗传算法的符号回归分析。如 Aryadoust [29]用自动线性建模技术研究随着 ESL 写作教学的推进，作文的文本特征(词汇复杂度、词汇多样性、句法复杂性、文本的衔接性等)和写作质量之间的关系呈现怎样的动态变化。研究发现，上述文本特征对写作质量的预测力随着时间的推进(从学期初到学期末)有所下降。同样是关于 ESL 写作质量的预测，Aryadoust [30]用基于遗传算法的符号回归研究词汇、句法、语篇层面的文本特征对写作质量的预测，发现词汇多样性、实词数量、词汇熟悉度、潜在语义分析指标对写作质量具有显著的预测作用。

除了写作质量预测，数据挖掘技术也被用来研究写作自动评分系统的优化和改进，Jin & He [31]提出假设，认为词的潜在语义特征或可提高作文自动评分系统的准确率。该研究首先基于一部分人工批改的作文样例提取了作文的语法、篇章流畅性、写作任务要求等特征，同时提取了三个潜在语义特征，分别是切题度、CBOW (continuous bag-of-words model)文章连贯性、RAE (recursive autoencoder)文章连贯性。用支持向量机算法构建作文自动评分模型，通过对比人工评分与模型评分之间的差异，发现增加上述三个潜在语义特征后，作文自动评分系统的准确率有所提高。

数据挖掘技术能高效地处理大规模写作数据，自动挖掘出作文的各项文本特征与写作质量之间的关系，有助于教师更详细地了解学生写作能力的发展。数据挖掘技术将助力对作文文本特征进行全面纵深的研究，不仅限于文章的表层或部分特征，这将有望不断提高作文自动评分系统的准确性，促进写作评分的客观性和公平性。

4. 数据挖掘技术应用于语言测试研究的启示

上述五个方面的研究，对语言测试具有理论、方法和实践意义。理论上，上述研究在某种程度上均能为语言测试的效度验证，尤其是为构念效度和评分效度提供证据。方法上，研究中多种数据挖掘技术

的应用，尤其是非线性模型如神经网络模型、决策树模型、逻辑回归模型等，表明在语言测试领域，应根据数据的真实情况选择合适的数据分析方法，当自变量与因变量呈非线性关系且自变量之间存在交互影响时，采用非线性的模型有助于得出更准确的结论。实践上，成绩预测能够及早发现可能挂科或学习困难的学生，进行教学干预；题目难度预测能发现与题目难度有关的因素，在开发试题时可以通过调控这些因素合理控制试题难度等。

值得注意的是，数据挖掘技术应用于语言测试研究时仍存在不足。数据挖掘只是对数据进行描述和预测，即揭示变量间隐藏的关系，它不能发现变量间的因果关系。以题目难度预测为例，数据挖掘技术只是挖掘出了和题目难度有关联的因素，而不能证明正是这些因素决定了题目难度。数据挖掘技术属于工程领域，其关注的是模型的最优化，而语言测试研究需要探究问题背后的教育学和社会学因素，最终服务于语言测试效度的提升，更好地满足考生、用人单位以及其他利益相关者的需求。因此，将数据挖掘技术应用于语言测试研究时，需要做一些适应性的改变，使之更好地服务于语言测试研究，如减少特征个数使模型更容易解释。

综上可知，数据挖掘技术已经应用于研究诸多语言测试问题，并取得了一定成果。然而除上述五个方面的研究外，更多主题，如测试的社会公平性、教师测评素养等尚待开展。更多语言测试数据有待挖掘，如论坛或贴吧上关于备考和考试的讨论以及经验贴等数据、相关的考试政策和文件、历年考试真题、考生关于备考的日志或学习记录等。其他数据挖掘技术有待尝试，如时间序列分析、关联规则挖掘、异常检测等。数据挖掘技术以高效处理海量数据，并从中发现隐藏的规律和模式见长，在未来的语言测试研究中，将发挥更大的作用和价值。

基金项目

国家社科基金重点项目“基于证据的四六级、雅思、托福考试效度对比研究”(项目编号：14AYY010)。

参考文献

[1] 杨惠中, 桂诗春. 语言测试的社会学思考[J]. 现代外语, 2007, 30(4): 368-374.

[2] 周庆, 牟超, 杨丹. 教育数据挖掘研究进展综述[J]. 软件学报, 2015, 26(11): 3026-3042.

[3] Aryadoust, V. and Goh, C. (2014) Predicting Listening Item Difficulty with Language Complexity Measures: A Comparative Data Mining Study. CaMLA Working Papers. https://www.researchgate.net/profile/Vahid_Aryadoust/publication/265249052_CaMLA_Working_Papers_Predicting_Listening_Item_Difficulty_with_Language_Complexity_Measures_A_Comparative_Data_Mining_Study_Predicting_Listening_Item_Difficulty_with_Language_Complexity_Measures_A_Com/links/5405efab0cf2c48563b1f95b.pdf

[4] Witten, I. and Frank, E. (2005) Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edition, Morgan Kaufmann Publishers, San Francisco.

[5] Cios, K., Pedrycz, W., Swiniarski, R. and Kurgan, L. (2007) Data Mining: A Knowledge Discovery Approach. Springer, New York.

[6] Aggarwal, C. (2015) Data Mining: The Textbook. Springer, New York. https://doi.org/10.1007/978-3-319-14142-8

[7] Hand, D., Mannila, H. and Smyth, P. (2001) Principles of Data Mining. The MIT Press, Massachusetts.

[8] Yin, Y., Kaku, I., Tang, J. and Zhu, J. (2011) Data Mining: Concepts, Methods and Applications in Management and Engineering Design. Springer, London. https://doi.org/10.1007/978-1-84996-338-1

[9] Gorunesco, F. (2011) Data Mining: Concepts, Models and Techniques. Springer, Berlin.

[10] Gurlitz, M. (2015) Forecasting Speak Test Score from TOEFL Score: A Bayesian Model for Screening International Teaching Assistants. *Systems and Information Engineering Design Symposium*, Charlottesville, 24 April 2015. https://doi.org/10.1109/SIEDS.2015.7116971

[11] Meng, Y., Gu, X., Zhou, Q. and Zhong, Y. (2017) Analyzing and Predicting the TEM-4 Performance of English Majors in China. *Proceedings of the 9th International Conference on Computer Supported Education*, **1**, 256-261. https://doi.org/10.5220/0006263102560261

[12] Xi, X. (2008) Investigating the Criterion-Related Validity of TOEFL Speaking Scores for IAT Screening and Setting Standards for ITA (TOEFL iBT Research Reports PR-08-02). Educational Testing Service, Princeton. https://doi.org/10.1002/j.2333-8504.2008.tb02088.x

[13] Mercado, V. (2017) Using TOEFL Sub-Scores to Predict SPEARK Test Outcome: A Multivariate Bayesian Model. *Systems and Information Engineering Design Symposium* (*SIEDS*), Charlottesville, 28 April 2017. https://doi.org/10.1109/SIEDS.2017.7937714

[14] Han, F. (2012) The Contribution of Lower-Level Processing to Foreign Language Reading Comprehension with Chinese FEL Learners. In: Zhang, Q. and Yang, H., Eds., *Proceedings of Pacific Rim Objective Measurement Symposium* (*PROMS*) 2012 *Conference*, Springer, Berlin, Heidelberg, 229-239. https://doi.org/10.1007/978-3-642-37592-7_16

[15] Perkins, K., Gupta, L. and Tammana, R. (1995) Predicting Item Difficulty in a Reading Comprehension Test with an Artificial Neural Network. *Language Testing*, **12**, 34-53. https://doi.org/10.1177/026553229501200103

[16] Aryadoust, V. (2015) Application of Evolutionary Algorithm-Based Symbolic Regression to Language Assessment: Toward a Nonlinear Modelling. *Psychological Test and Assessment Modelling*, **3**, 301-337.

[17] Alamir, M. (1999) Optimization Based Nonlinear Observers Revisited. *International Journal of Control*, **72**, 1204-1217. https://doi.org/10.1080/002071799220353

[18] Keith, T. (2006) Multiple Regression and Beyond. Pearson, Boston.

[19] Aryadoust, V. and Baghaei, P. (2016) Does EFL Readers' Lexical and Grammatical Knowledge Predict Their Reading Ability? Insights from a Perceptron Artificial Neural Network Study. *Educational Assessment*, **21**, 135-156. https://doi.org/10.1080/10627197.2016.1166343

[20] Freedle, R. and Kostin, I. (1991) The Prediction of SAT Reading Comprehension Item Difficulty for Expository Prose Passages (ETS Research Report RR-91-29). Educational Testing Service, Princeton. https://doi.org/10.1002/j.2333-8504.1991.tb01396.x

[21] Freedle, R. and Kostin, I. (1999) Does the Text Matter in a Multiple-Choice Test of Comprehension? The Case for the Construct Validity of TOEFL's Mini-Talks. *Language Testing*, **16**, 2-32. https://doi.org/10.1177/026553229901600102

[22] 韩菡. 基于人工神经网络预测汉语阅读理解测验题目难易度的研究[D]: [硕士毕业论文]. 北京: 北京语言大学, 2005.

[23] 付佩宣. 基于人工神经网络的 C.TEST 阅读理解题目难度的预测研究[J]. 华文教学与研究, 2014(4): 71-78.

[24] Gao, L. and Rogers, W. (2011) Use of Tree-Based Regression in the Analyses of L2 Reading Test Items. *Language Testing*, **28**, 77-104. https://doi.org/10.1177/0265532210364380

[25] Aryadoust, V. (2013) Predicting Item Difficulty in a Language Test with an Adaptive Neural Fuzzy Inference System. *IEEE Workshop on Hybrid Intelligent Models and Applications* (*HIMA*), Singapore, 16-19 April 2013. https://doi.org/10.1109/HIMA.2013.6615021

[26] Rupp, A., Garcia, P. and Jamieson, J. (2001) Combine Multiple Regression and Cart to Understand Difficulty in Second Language Reading and Listening Comprehension Test Items. *International Journal of Testing*, **1**, 185-216. https://doi.org/10.1207/S15327574IJT013&4_2

[27] Eckes, T. (2008) Rater Types in Writing Performance Assessment: A Classification Approach to Rater Variability. *Language Testing*, **25**, 155-185. https://doi.org/10.1177/0265532207086780

[28] Eckes, T. (2012) Operational Rater Types in Writing Assessment: Linking Rater Cognition to Rater Behavior. *Language Assessment Quarterly*, **9**, 270-292. https://doi.org/10.1080/15434303.2011.649381

[29] Aryadoust, V. (2013) Can Computer-Generated Linguistic Features Predict Second Language Students' Writing Score across Time? *Proceedings of* 2013 *Technology Enhanced Learning Symposium*, Singapore, 7 October 2013, 16-19.

[30] Aryadoust, V. (2016) Application of Genetic Algorithm-Based Symbolic Regression in ESL Writing Research. In: Aryadoust, V. and Fox, J., Eds., *Current Trends in Language Testing in the Pacific Rim and the Middle East*: *Policies*, *Analysis and Diagnosis*, Cambridge Scholars Publishing, Newcastle, 35-46.

[31] Jin, C. and He, B. (2015) Utilizing Latent Semantic Word Representations for Automated Essay Scoring. 2015 *IEEE* 12*th International Conference on Ubiquitous Intelligence and Computing* & 2015 *IEEE* 12*th International Conference on Autonomic and Trusted Computing* & 2015 *IEEE* 15*th International Conference on Scalable Computing and Communications and Its Associated Workshops* (*UIC-ATC-ScalCom*), Beijing, 10-14 August 2015. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.202