心理学报 2026, Vol. 58, No. 4, 755−772 https://doi.org/10.3724/SP.J.1041.2026.0755
Acta Psychologica Sinica https://cstr.cn/32110.14.2026.0755

基于迁移学习与 Q 矩阵约束的神经网络认知诊断方法*

陶金洪[1] 赵 蔚[1] 程 诺[1] 乔丽方[2] 姜 强[1]

([1]东北师范大学信息科学与技术学院, 长春 130117) ([2]河北师范大学教育学院, 石家庄 050000)

摘 要 神经网络作为最重要的机器学习方法已被广泛地用于认知诊断, 但目前仍没有一种简单通用的神经网络认知诊断方法。因此, 提出一种 Q 矩阵约束的神经网络认知诊断方法(Bi-QNN), 并基于迁移学习进行训练。新模型的优势在于:(1)使用人员无需专门设计网络结构, 新模型可以根据Q矩阵与交互式Q矩阵自适应任意数据集; (2)网络结构的设计原理源于 GDINA 模型, 使其能够较好地表达属性的主效应与交互效应; (3)基于迁移学习的模型训练方案能有效地解决标记数据稀缺问题, 提高模型的易用性与适用范围。实验结果表明: Bi-QNN 在模拟数据集上的预测误差整体上比参数化方法 GDINA 与 DINA 的表现更好; 在一定的范围内, 模型对属性数量敏感性相对较低, 当属性数量增加时在一定程度上仍能保持较好的分类准确率; 基于迁移学习训练的 Bi-QNN 方法能更好地适应不同样本量的数据集, 在模拟数据与实证数据的多种条件下保持对其它模型的领先; 模型性能的进一步提升受到基于参数模型的模拟数据的限制, 对试题质量仍有一定的敏感性。

关键词 认知诊断, Q 矩阵, 人工神经网络, 迁移学习

分类号 B841

1 引言

认知诊断评估(Cognitive Diagnostic Assessment, CDA)旨在根据被试的作答反应精细化地评估他们的潜在特质或特定知识技能的熟练程度, 为教学或心理干预提供个性化的指导, 这使它被广泛应用于心理评估和个性化学习(Xin et al., 2022)。研究者们针对主观题与客观题已经提出大量的认知诊断模型, 可分为二级计分模型与多级计分模型(高旭亮等, 2021); 从模型的计算上分为参数模型与非参数模型(Liu et al., 2022)。其中, 被广泛应用的参数化多级计分模型有部分信用 DINA 模型 PC-DINA (Partial Credit Diagnostic Model; de la Torre & Minchen, 2019)、顺序 GDINA 模型 seqGDINA (The sequential G-DINA; Ma & de la Torre, 2016)等。本研究主要聚焦于二级计分模型, 其中最具代表性的非补偿模型包括 NIDA (noisy inputs, deterministic "and" gate; Junker & Sijtsma, 2001) 和 DINA (Deterministic input, noisy "and" gate model; de la Torre, 2009)等, 而补偿模型则包括 Templin 和 Henson (2006)提出的 DINO (Deterministic input, noisy "or" gate model) 与 NIDO (Noisy input, deterministic-"or"-gate model)等。除此之外, 还有更一般化的模型如 GDINA (The generalized DINA; de la Torre, 2011)等。参数化认知诊断模型以概率统计为基础, 通过各种参数估计方法完成模型的参数估计, 如最大似然估计等(Sorrel et al., 2023)。这类方法在样本量不足时通常会出现参数估计的边界问题, 从而造成模型的可信度被高估或低估的情况(Yamaguchi, 2023)。针对这类问题, 非参数认知诊断方法被提出, 作为在小班化教学场景的替代方案(汪文义 等, 2021)。非参数方法通常被分为基于向

收稿日期: 2024-03-11

* 国家自然科学基金面上项目(62577018)。

通信作者: 赵蔚, E-mail: zhaow577@nenu.edu.cn

量相似度距离的方法与基于机器学习的方法(郭磊, 周文杰, 2021)。在基于向量距离的方法中最具有代表性的方法包括 NPC (Nonparametric classification method; Chiu & Douglas, 2013), GNPC (Generalized nonparametric classification method; Chiu et al., 2018)等。除此之外, 还有基于各种向量距离改进的方法, 如汉明距离、马氏距离和曼哈顿距离等各类基于向量相似度计算的判别方法(徐慧颖 等, 2023)。

机器学习在认知诊断中被应用最广泛的是基于聚类的方法(Guo et al, 2020)。在早期 Chiu 等(2009)使用层次凝聚聚类和 K 均值方法将考生分为具有相同属性模式的不同簇。康春花等(2015)利用属性复合得分向量与 K 均值聚类提出等级反应聚类诊断方法(GRCDM), 之后为弥补样本量较小时 K 均值聚类精度较低的问题又提出了 KNN CDM (康春花 等, 2019)。Guo 等(2020)为弥补 K 均值聚类的不足提出基于谱聚类的方法。Zhang 等(2022)融入理想反应模式数量作为 K 值, 提出非参数多级评分认知诊断方法(EW-KNN)。除了基于聚类的方法之外, 也有研究人员将其它的机器学习方法应用到认知诊断中, 比如 Liu 和 Cheng (2018)提出将支持向量机(Support Vector Machine, SVM)应用于认知诊断, 并在小样本条件下取得与之前的研究相当的性能。随着人工智能的发展, 深度神经网络方法已成为最具吸引力的方法之一(Liu, 2021)。如 Cui 等(2016)结合 DINA 模型的理想反应数据来训练基于神经网络的认知诊断模型, 其实验结果表明基于理想数据训练的神经网络诊断结果不尽人意。汪文义等(2016)结合概率神经网络(PNN)与支持向量机(SVM)用于认知诊断, 表明在独立属性结构下 PNN 的表现较好。Chen 和 Yan (2021)针对不同的属性层级关系, 使用基于 DINA 的模拟数据训练神经网络认知诊断模型, 其结果表明神经网络能有效对属性掌握模式进行分类, 并且还表明属性结构的复杂性越高分类精度越低。聂畅等(2021)从属性数量与层级、题目质量、样本量等多个因素分析神经网络认知诊断的性能, 其结果表明样本量对神经网络认知诊断性能的影响比题目质量和属性数量小。另外, Wen 等(2020)利用人工神经网络作为隐马尔可夫模型(HMM)的测量模型, 实现了纵向认知诊断, 以监测学生认知属性的发展。Wang 等(2020)提出一种可扩展神经认知诊断框架(NeuralCD), 该框架可以较好地利用学习者与试题之间的交互信息, 以获得具有一定可解释性的诊断结果。Xue 和 Bradshaw (2021)为了解决标记数据的稀缺问题将神经网络与 DINA 和 DINO 模型集成, 实现半监督神经网络认知诊断模型, 其结果表明他们的方法是有效的, 同时当项目的区分度降低时, 诊断准确性也会跟着降低。

综上所述, 除了传统的参数、非参数方法和基于聚类的方法, 众多研究人员已将神经网络应用于认知诊断并证明它的有效性。然而, 这些研究都是在特定的实验条件下完成, 具有一定的场景特殊性, 这使得目前仍然没有一种简单且普遍通用的神经网络认知诊断方法。在实际测验情境中, 每一场测验可能都会存在不同数量的被试、知识或属性与题目等。神经网络方法被用于认知诊断时, 其网络结构, 比如网络的深度和每层神经元的数量等应该如何设计是一个令人比较困惑的问题。尤其对于没有深厚人工智能专业知识的心理学者与教育者们来说尤为困难。在过去这个问题一直没有较好的解决方法。此外, 神经网络的参数求解属于监督学习的范畴, 通常用梯度下降方法进行求解。如何在只有少量标注数据情况下训练出一个具有较好泛化能力的神经网络认知诊断模型一直是研究者们面临的挑战(Xue & Bradshaw, 2021)。因此, 尽管神经网络方法在各个领域比如自然语言处理、计算机视觉等都取得了巨大的成功, 但在认知诊断上, 不像基于概率统计的参数化方法与基于向量距离的非参数方法被广泛的应用和推广。

因此, 我们从 GDINA 模型的假设出发, 提出一种被称为 Bi-QNN 的神经网络认知诊断方法, 以解决当前神经网络方法应用于认知诊断中面临的模型设计与训练困难和通用性挑战等问题。具体来说, Bi-QNN 的网络结构由 Q 矩阵与交互式 Q 矩阵共同约束, 并以此来表达属性的主效应与交互效应, 这样的设计能够使 Bi-QNN 模型的网络结构可以根据 Q 矩阵自动适应不同的数据。更重要的是教学人员在使用 Bi-QNN 模型时, 只需提供 Q 矩阵与属性间的交互关系矩阵即可自动完成 Bi-QNN 模型的构建, 省去了亲自设计神经网络模型的困难。另外, 为了应对各种场景下 Bi-QNN 的训练, 我们基于迁移学习设计了对 Bi-QNN 的训练方案, 并在大量模拟实验和实证数据分析中评估了 Bi-QNN 模型的性能。最后, 本文的结构如下：首先介绍 Q 矩阵与 GDINA 模型的相关概念和理论, 其次阐述 Bi-QNN 认知诊断模型的构建过程和基于迁移学习对它的训练方法, 之后通过模拟和实证研究对 Bi-QNN 的性能进行评估, 最后对结果进行讨论与展望。

2　相关概念与定义

2.1　Q 矩阵

Q 矩阵为描述测验题目与属性之间关联关系的二值矩阵(Tatsuoka, 1995), 具体为每一行代表一道题目, 每一列表示一个属性, 其定义如公式(1)所示：

$$Q=(\boldsymbol{q_{jk}})_{\boldsymbol{J\times K}} \tag{1}$$

公式(1)中的 **Q** 矩阵表示了 J 道题目与 K 个属性之间的关系, 其中属性在不同的情境下可代指知识、技能和潜在特质等, 为方便描述后续统一用属性一词进行表述。在 **Q** 矩阵中, 若 $q_{jk}=1$ 表示题目 j 考察了属性 k, 否则为没有考察属性 k。属性之间的关系可以表示为独立关系与相互关系, 其中独立关系表现为属性之间互相独立, 而相互关系表现为属性之间存在直接或者间接的依赖或影响。属性之间的这些关系通常可以用可达矩阵 **R** 表示, 其定义如公式(2)所示：

$$\mathbf{R}=(\boldsymbol{r_{K_iK_j}})_{\boldsymbol{K\times K}} \tag{2}$$

其中, 可达矩阵的第 K 行表示属性 K_i 与其它属性之间的直接或间接关系, 如果 $r_{K_iK_j}=1$ 表示属性 K_i 与属性 K_j 之间存在相互关系, 否则表示属性 K_i 与属性 K_j 互相独立。

2.2　GDINA 认知诊断模型

本研究中神经网络的设计灵感来自 GDINA 模型, 其项目反应计算如公式(3)所示：

$$P(\boldsymbol{\alpha}_{\ell j}^{*})=\delta_{j0}+\sum_{k=1}^{K_j^{*}}\delta_{jk}\alpha_{\ell k}+\sum_{k'=k+1}^{K_j^{*}}\sum_{k=1}^{k_j^{*}-1}\delta_{jkk'},\alpha_{\ell k}\alpha_{\ell k'}+\cdots+\delta_{j12\cdots K_j^{*}}\prod_{k=1}^{K_j^{*}}\alpha_{\ell k} \tag{3}$$

GDINA 模型根据每道题目真实考察的属性情况, 将被试在该题目上的属性掌握模式缩减为 $L=2^{K_j^{*}}$ 种, 其中 K_j^{*} 表示题目 j 实际考察的属性数量, 其计算如公式(4)所示:

$$K_j^{*}=\sum_{k=1}^{K}q_{jk} \tag{4}$$

其中, $\boldsymbol{\alpha}_{\ell j}^{*}$ 为一个 $L\times K^{*}$ 的二值矩阵, 它表示题目 j 实际考察的属性构成的属性模式集合。如公式(4) GDINA 模型将各个属性对正确回答一道题目的作用分为三个部分。首先第一项截距 δ_{j0}, 表示被试在未掌握任何属性的情况下正确回答题目 j 的概率; 第二项主效应 δ_{jk}, 表示被试掌握属性 k 对正确回答题目 j 的直接贡献; 剩下的为交互效应, 其中 $\delta_{jkk'}$ 表示被试同时掌握属性 k 与属性 k' 对正确回答题目 j 的间接贡献; $\delta_{j1,\cdots,K_j^{*}}$ 表示掌握题目 j 考察的所有属性对正确回答题目 j 的贡献。

3　融入 Q 矩阵的神经网络认知诊断模型

3.1　交互关系矩阵与交互式 Q 矩阵

如公式(1)所示的 **Q** 矩阵, 在未指明属性之间的层级和依赖关系时, 根据二项式定理, 排除属性自身与自身的关系, K 个属性之间存在 2^K-K-1 种交互关系。这显然与实际测验中属性之间真实存在的相互关系不相符合。为了符合实际并以一种更简洁的方式表示属性之间的相互关系, 在本文中用一种被称为交互关系矩阵的二值矩阵 $\boldsymbol{Q}^{\#}$ 来表示多个属性之间的相互关系, 如公式(5)所示：

$$\boldsymbol{Q}^{\#}=(q_{mk}^{\#})_{M\times K} \tag{5}$$

交互关系矩阵 $\boldsymbol{Q}^{\#}$ 由 M 行 K 列组成, 其中每一行表示一种属性之间的交互关系。比如, 若 $\boldsymbol{Q}^{\#}$ 中有一个包含 4 个属性的行 $\mathbf{q}^{\#}=(1,0,1,1)$, 其表示第一、第三和第四个属性之间存在交互关系。

交互关系矩阵 $\boldsymbol{Q}^{\#}$ 仅表征属性之间存在的交互关系, 并没有表示每一道题目中存在的交互关系。因此, 在本文中用一个被称为交互式 **Q** 矩阵的二值矩阵 $\boldsymbol{Q}^{*}$ 来表示 **Q** 矩阵中每一道题目中存在的交互关系, 可由 **Q** 矩阵与交互关系矩阵 $\boldsymbol{Q}^{\#}$ 计算获得, 如公式(6)所示：

$$\boldsymbol{Q}^{*}=(q_{jm}^{*})_{J\times M},q_{jm}^{*}=\prod_{k=1}^{K}q_{jk}^{q_{mk}^{\#}},\quad j=1,2,\cdots,J,m=1,2,\cdots,M \tag{6}$$

其中, M 为交互关系矩阵 $\boldsymbol{Q}^{\#}$ 的行数, $q_{jm}^{*}=1$ 表示题目 j 中存在交互关系矩阵 $\boldsymbol{Q}^{\#}$ 中的第 m 种交互关系, 否则表示不存在这种交互关系。

3.2　基于 Q 矩阵与交互式 Q 矩阵约束的神经网络认知诊断模型

GPT4 的发布标志着深度神经网络在图像和自然语言处理等多个领域都取得了更进一步的重大突破。最早的深度前馈神经网络由 Rosenblatt 等(1958)提出, 并因 Rumelhart 等(1986)提出 BP 算法解决了神经网络参数计算问题, 使得深度神经网络在各个领域得到了长足的发展, 并取得了令人瞩目的成就。因此, 本文从 GDINA 模型的原理出发, 结合深度神经网络提出一种基于 Q 矩阵与交互式 Q 矩阵约束的神经网络认知诊断模型, 并命名为 Bi-QNN。如公式(5)GDINA 模型将正确回答题目的贡献分为三大部分, 分别为题目的基准概率即截距, 单个属性对题目的主效应和多个属性之间对题目

的交互效应。受 GDINA 模型的启发，我们设计的 Bi-QNN 模型的网络架构如图 1 所示，Bi-QNN 模型的网络结构包含 2 条计算流。首先是代表主效应的绿色计算流，其次是代表交互效应的橙色计算流，并在隐藏层神经元的计算中保留了偏置项。这使得 Bi-QNN 模型的计算与 GDINA 模型保持高度的一致性。

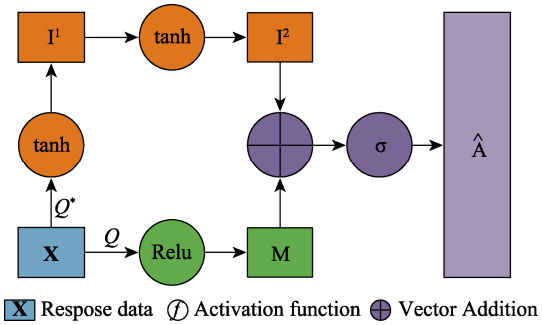

图 1 Bi-QNN 网络架构图

注：绿色表示主效应，橙色表示交互效应，紫色为综合结果，彩图见电子版

Bi-QNN 模型的详细计算过程如下，首先是表示主效应的计算流，其每一层的维度由 **Q** 矩阵决定。在该计算流中，网络的输入层为学生的作答反应模式，由 J 个神经元构成，每个神经元表示一个题目；隐藏层由 K 个神经元构成，每一个神经元表示一种属性。其具体计算如公式(7)所示：

$$M = Relu(\mathbf{X}(\mathbf{Q} \odot \mathbf{W}^m) + \mathbf{b}^m) \tag{7}$$

其中，$\mathbf{M} \in \mathbb{R}^{N \times K}$ 表示主效应隐藏层的输出，$\mathbf{X} \in \mathbb{R}^{N \times J}$ 表示被试们的作答响应数据，$\mathbf{W}^m \in \mathbb{R}^{J \times K}$ 表示该隐藏层的权重，$\mathbf{b}^m \in \mathbb{R}^K$ 为每个隐藏层神经元的偏置项。在 GDINA 模型中，主效应通常被认为是非负的，因此 $Relu(\cdot)$ 函数被作为主效应的激活函数。需要特别说明的是 $\odot$ 为矩阵的哈达马积(Hadamard product)，通过计算 **Q** 矩阵与该隐藏层的权重的哈达马积，实现对该隐藏层与输入层神经元之间链接的约束。

橙色的计算流表示属性之间的交互效应。在 GDINA 模型中对于任意一个题目存在的交互效应是以该题目中实际考察的属性数 K^* 来计算的，具体为 $2^{K^*} - K^* - 1$ 种交互效应。而实际并不是任何组合的属性之间都是真实存在并相互影响的，同时由于二项式展开项为指数级，这导致交互效应的数量会随着属性数量的增加而急剧增加，也会导致 GDINA 模型在属性数量增加后，模型的参数量也会急剧增加，因此在估计其模型参数时就要求更大的样本量才能保证模型的可靠性。考虑到如果完全依据 GDINA 模型对交互效应的表征方式来设计神经网络模型，会导致网络结构变得非常宽，同时会导致网络结构变得非常稀疏。过宽的网络会增加模型的复杂度，导致计算成本和训练时间上升；而过于稀疏的网络会影响梯度传播，使得训练过程不稳定，模型难以收敛，降低模型在实际应用中的泛化能力。为了克服这些问题，在本研究中，我们参照 GDINA 模型的理念，但是并不完全依照 GDINA 模型去表征交互效应，而是引入了交互式 **Q** 矩阵来优化网络结构，减少不必要的节点和连接，从而实现在保证模型性能的前提下确保模型的计算效率。具体为对于 K 个属性之间的交互关系，我们希望由专家根据专业知识和实际情况明确给出属性间的交互关系，即交互关系矩阵 $\boldsymbol{Q}^{\#}$。依据公式(6)计算得到交互式 **Q** 矩阵 $\boldsymbol{Q}^*$，并依据它设计交互效应计算流的网络结构。在 Bi-QNN 模型中交互效应的计算具有两个隐藏层，其中第一个隐藏层的计算如公式(8)所示：

$$\mathbf{I}^1 = tanh[\mathbf{X}(\mathbf{Q}^* \odot \mathbf{W}^{i1}) + \mathbf{b}^{i1}] \tag{8}$$

其中，$\mathbf{I}^1 \in \mathbb{R}^{N \times M}$ 表示第一个隐藏层的计算结果，$\mathbf{X} \in \mathbb{R}^{N \times J}$ 表示被试的作答响应数据，$\mathbf{Q}^* \in \mathbb{R}^{J \times M}$ 为交互式 **Q** 矩阵，$\mathbf{W}^{i1} \in \mathbb{R}^{J \times M}$ 表示该隐藏层的权重，$\mathbf{b}^{i1} \in \mathbb{R}^M$ 为该层中每个隐藏层神经元的偏置项。$tanh(\cdot)$ 为激活函数，其取值范围为 $(-1,1)$。与前面的方式相同，通过交互式 **Q** 矩阵 $\mathbf{Q}^*$ 对网络结构进行约束。第二个隐藏层的计算如公式(9)所示：

$$\mathbf{I}^2 = tanh(\mathbf{I}^1\mathbf{W}^{i2} + \mathbf{b}^{i2}) \tag{9}$$

其中，$\mathbf{I}^2 \in \mathbb{R}^{N \times K}$ 表示第二个隐藏层的输出，$\mathbf{W}^{i2} \in \mathbb{R}^{M \times K}$ 表示该隐藏层的权重，$\mathbf{b}^{i2} \in \mathbb{R}^K$ 为该层中每个隐藏层神经元的偏置项。

通过 Sigmoid 激活函数对主效应和交互效应的和进行仿射变换，即可得到 Bi-QNN 模型对被试属性掌握概率的预测结果 $\hat{\mathbf{A}}$，如公式(10)所示：

$$\hat{\mathbf{A}} = \sigma((\mathbf{M} + \mathbf{I}^2)\mathbf{W} + b) \tag{10}$$

其中，$\hat{\mathbf{A}} \in \mathbb{R}^{N \times K}$ 表示模型对被试属性掌握情况的预测结果，$\sigma(\cdot)$ 为 Sigmoid 激活函数。参照参数模型预测属性概率截断的设置，将 $\hat{\mathbf{A}}$ 中大于 0.5 的值置 1，否则置 0，即可得到离散化后的属性掌握情况。

通过公式(7)-(10)完成 Bi-QNN 模型的前向计算。通过反向传播算法更新模型参数，Bi-QNN 模型的损失函数为均方误差函数，如公式(11)所示。

$$\mathcal{L}(\theta) = \frac{1}{2N}(\hat{\mathbf{A}} - \mathbf{A})^{\mathrm{T}}(\hat{\mathbf{A}} - \mathbf{A}) \tag{11}$$

其中，θ 表示模型的参数，包括该网络中各层的权

重和偏差，**A** 是被试们的真实属性掌握模式，N 为被试人数。

4　基于迁移学习的神经网络认知诊断模型训练方法

4.1　迁移学习及其方法选择

在机器学习中，迁移学习是一种特定的学习模式，旨在将模型在某项任务中学到的知识，迁移并应用到另一相关任务中，从而提升模型在新任务上的学习效果与泛化能力(Pan & Yang, 2010)。迁移学习能够解决在新任务中缺乏足够标注数据的问题。通过将模型在相关任务中学到的知识迁移到新任务中，可以有效提升模型在新任务上的学习效果和泛化能力，从而避免了在新任务上重新收集数据并进行人工标注的高成本和实现难度。这种特性正好可以解决认知诊断任务中场景多变与标注数据稀少的问题。

为了方便描述，对迁移学习进行形式化的定义。首先是领域，它是模型学习的主体，由数据和生成这些数据的概率分布构成。一个领域可以被形式化地表示为公式(12)所示：

$$\mathcal{D}=\{\mathcal{X},\mathcal{Y},P(\boldsymbol{x},y)\} \tag{12}$$

其中，$\mathcal{D}$ 表示一个领域，$\mathcal{X}$ 和 $\mathcal{Y}$ 分别表示数据所处的特征空间和类别标签空间。领域上的任意一个样本 $(\boldsymbol{x_i},y_i)$ 都包含样本特征和对应的标签，即满足 $\boldsymbol{x}_i\in\mathcal{X},y_i\in\mathcal{Y}$。同时，领域中的样本服从概率分布 $P(\boldsymbol{x},y)$，即 $(\boldsymbol{x},y)\sim P(\boldsymbol{x},y)$。迁移学习中通常会包含源领域 $\mathcal{D}_s$ 和目标领域 $\mathcal{D}_t$。源领域中包含了大量的标注数据，模型可以从中学习到知识，它是要被迁移的对象，目标领域是我们最终要赋予其知识的对象，迁移学习需要把模型在源领域学习到的知识传递到目标领域。结合领域的概念迁移学习被形式化地定义为：

给定一个源域 $\mathcal{D}_s=\{\boldsymbol{x}_i,y_i\}_{i=1}^{N_s}$ 和目标域 $\mathcal{D}_t=\{x_j,y_j\}_{j=1}^{N_t}$，其中 $\boldsymbol{x}\in\mathcal{X},y\in\mathcal{Y}$。迁移学习的目标是当以下三种情形：

(1)目标域与源域的特征空间不同，即 $\mathcal{X}_t\neq\mathcal{X}_s$；

(2)目标域与源域的类别标签空间不同，即 $\mathcal{Y}_t\neq\mathcal{Y}_s$；

(3)目标域与源域的特征和类别标签空间相同，但概率分布不同，即 $P_t(\boldsymbol{x},y)\neq P_s(\boldsymbol{x},y)$

至少有一种成立时，利用源域数据去学习一个目标域上的预测函数 $f:\boldsymbol{x}_t\mapsto y_t$，使得 f 在目标域上的预测误差 ℓ 最小，如公式(13)所示：

$$f^*=\arg\min_f\ \mathbb{E}_{(\boldsymbol{x},y)\in\mathcal{D}_t}\mathcal{L}(f(\boldsymbol{x}),y) \tag{13}$$

其中，f^* 为通过迁移学习得到的在目标域上的最优预测函数，$\mathcal{L}(\cdot)$ 为代价函数。

迁移学习中根据迁移的知识类型，可分为基于实例、特征、关系和模型的迁移学习(Zhuang et al., 2021)。其中，尽管基于实例、特征和关系的迁移学习方法在早期得到广泛的关注，但在实际应用中存在一定的局限性。其中，基于实例的方法高度依赖源域与目标域数据分布的相似性，当源域和目标域存在较大的领域差异时，容易导致负迁移(Pan & Yang, 2010)。基于特征的方法在实现特征空间对齐的同时需兼顾判别性，面临实现难度较高的问题(Wang & Deng, 2018)；基于关系的方法则常常依赖于对源目标任务之间复杂关系的建模，限制了其适用范围(Zhuang et al., 2021)。相比之下，基于模型的迁移学习方法可以共享模型在源域上学习到的参数，尤其是在神经网络方法中，只需要在目标域上微调模型，模型在源域上学习到的参数就可以在目标域上获得很好的迁移效果(Tan et al., 2018)。这使得基于模型的迁移学习方法在神经网络的训练上具有较好的通用性和普遍性。本研究中，我们聚焦于用神经网络方法实现认知诊断，会面临样本量少、数据标注成本高和属性掌握模式类别不平衡等问题，因此本文选择基于模型的迁移学习方法来实现对 Bi-QNN 模型的优化。

4.2　基于预训练−微调的神经网络认知诊断模型训练方法

神经网络的训练是以模型在数据集上的预测值与真实值之间的损失最小为目标，所以训练神经网络需要人工提前标注真实的结果，即在认知诊断中需要用专家对学生作答反应的诊断结果作为模型训练的基准。而认知诊断的目的是希望机器去自动诊断学生的知识掌握情况，而不是人为的去诊断。但是神经网络的训练本身需要人工诊断的数据作为训练的依据，这就导致了二者之间的巨大矛盾。另外，测验通常都是一次性的，每一场考试几乎都需要重新出题，即考卷通常不会被重复使用在相同的被试身上。也就是说即使训练出一个很好的神经网络模型，但是也只适用于某一场特定的考试，而通常的考试又几乎都是一次性的，从而使基于神经网络的认知诊断模型可复用性差。这就导致了神经网络的方法虽然在各个领域都取得令人瞩目的成就，但在认知诊断上仍然没有被广泛使用。

本文中我们使用基于模型中预训练−微调的迁移学习方法来解决上面提到的困境。具体地，通过

用一种或者多种参数化认知诊断模型生成大量的模拟数据集来预训练神经网络模型, 然后将预训练好的模型迁移到较小的数据集上进行微调, 从而完成神经网络认知诊断方法的训练。这个过程包含预训练与微调两个步骤, 首先是预训练, 即对于一个给定的Q矩阵, 基于参数化认知诊断方法生成预训练模拟数据源域 $\mathcal{D}_s$, 在源域 $\mathcal{D}_s$ 上学习一个神经网络认知诊断目标函数 f, 使得 f 在源域 $\mathcal{D}_s$ 上具有最小代价 ℓ, 如公式(14)所示:

$$\theta_s^* = \arg\min_{\theta} \mathcal{L}(\mathcal{D}_s; \theta) \tag{14}$$

其中, θ 表示函数 f 的参数, θ_s^* 表示函数 f 在源域 $\mathcal{D}_s$ 上学到的最优参数, $\mathcal{L}$ 为代价函数。第二步为微调, 即在真正需要被预测的数据集目标域 $\mathcal{D}_t$ 上对认知诊断目标函数 f 进行再训练。具体为利用预训练的参数 θ_s^* 在目标域 $\mathcal{D}_t$ 上学习一个由 θ_t^* 所表征的被更新后的认知诊断目标函数 f, 如公式(15)所示:

$$\theta_t^* = \arg\min_{\theta} \mathcal{L}(\theta \mid \theta_s^*, \mathcal{D}_t) \tag{15}$$

其中, θ_t^* 为认知诊断目标函数 f 在目标域 $\mathcal{D}_t$ 上具有最小代价的最优参数。本研究中以迁移学习的方式对 Bi-QNN 模型训练的具体过程如图 2 所示。

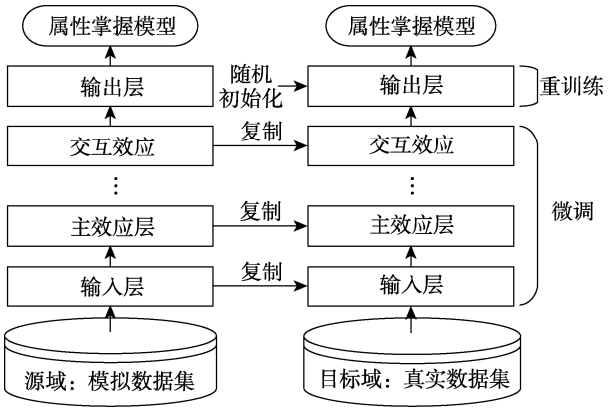

图 2 基于微调−预训练的 Bi-QNN 训练示意图

首先是在模拟数据集源域 $\mathcal{D}_s$ 上训练源域网络, 然后把在源域数据集上学习到的关于主效应和交互效应的知识迁移到目标域中, 即复制源域网络中主效应层和交互效应层的参数到目标域网络中进行微调, 并在目标网络中重新训练模型的网络参数。

5 模拟研究

5.1 研究目的

模拟研究的目是为了在更广泛的条件下评估基于迁移学习训练的Bi-QNN模型在分类误差与分类准确性上的性能和表现, 同时将其与神经网络方法(ANN)、参数化方法(GDINA, DINA)和非参数化方法(NPC, GNPC)进行性能对比, 以讨论各方法在不同条件下的优劣。

5.2 研究设计

为了在模拟实验中更全面地评估 Bi-QNN 模型, 参考以往的研究, 被试的属性分布采用多元正态分布生成; 在属性数量上, 已有的研究表明属性数量范围通常在 3 到 8 (Qin & Guo, 2024), 3 个属性在样本量较少的情况比较常见, 而在模拟研究中 5 个属性是应用最广泛的(Nájera et al., 2021); 题目数量一般由题目数与属性数的比率决定, 通常要求题目数量为属性数量的 3 倍以上(宋丽红 等, 2024); 为了更好的评估模型对属性数量与题目数量的敏感性, 我们设计了 3 种规格的 Q 矩阵, 其(属性数量, 题目数量)分别为(3, 15)、(5, 15)和(5, 25)。对于题目质量分别考虑生成模拟数据时题目的高质量与低质量两个因素, 参考过去的研究, 分别将题目的猜测与失误参数属于均匀分布 0.05 到 0.15, 0.15 到 0.30 定义为高质量与低质量题目, 即 P_0 和 $1-P_1 \in \mathbf{U}(0.05, 0.15)$ 为高质量题目, P_0 和 $1-P_1 \in \mathbf{U}(0.15, 0.30)$ 为低质量题目(郭磊, 周文杰, 2021; 聂畅 等, 2021; Cui et al., 2016)。另外, 在每个条件下都分别获取样本量为 $N(N = 50,100,200,300,500)$ 的子数据集。根据以上的设定构成 $2\times2\times2\times5 = 40$ 种条件来生成多样化的实验数据。

5.2.1 题目模拟

模拟实验中属性为 3 题、题目数量为 15 的 Q 矩阵被设计如公式 16 所示, $\boldsymbol{Q}_1$ 由两个单位矩阵构成以确保其完备性和参数可识别性(Chiu & Douglas, 2013; Xu & Zhang, 2016), 剩余题目则包含至少 2 个以上的属性构成更复杂的属性结构。

$$\mathbf{Q}_1^{\top} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \tag{16}$$

属性数量为5、题目数量为15的Q矩阵与Zhan等(2022)研究中图二的 Q 矩阵保持一致, 具体如公式(17)所示:

$$\mathbf{Q}_2^{\top} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \end{pmatrix} \tag{17}$$

属性数量为 5, 题目数量为 25 的 Q 矩阵, 参考 Zhan 等(2022)研究中属性数量为 5、题目数量为 30 的 Q 矩阵, 在本研究中被设计如公式(18)所示：

$$\mathbf{Q}_3^{\top}=\begin{pmatrix}1&0&0&0&0&1&1&1&1&0&0&0&0&0&0&1&1&1&1&1&1&0&0&0&0\\0&1&0&0&0&1&0&0&0&1&1&1&0&0&0&1&1&1&0&0&0&1&1&1&0\\0&0&1&0&0&0&1&0&0&1&0&0&1&1&0&1&0&0&1&1&0&1&1&0&1\\0&0&0&1&0&0&0&1&0&0&1&0&1&0&1&0&1&0&1&0&1&1&0&1&1\\0&0&0&0&1&0&0&0&1&0&0&1&0&1&1&0&0&1&0&1&1&0&1&1&1\end{pmatrix} \tag{18}$$

交互关系矩阵的生成方式为每次实验从 K 个属性间存在的 2^K-K-1 种交互关系中随机的选取指定数量的交互关系来构成本次实验的交互关系矩阵, 并保证交互关系矩阵中不存在重复的关系。其中, 在实验中对于 3 个属性和 5 个属性的交互式 Q 矩阵, 在确保交互关系最少包含两个属性且不重复的前提下, 设定的交互关系数量分别为 4 和 10, 即 $\boldsymbol{Q}_1$ 的交互关系矩阵为 4×3 的二值矩阵, $\boldsymbol{Q}_2$ 和 $\boldsymbol{Q}_3$ 为 10×5 的二值矩阵, 它们分别表示对于 3 个属性的数据选择了 4 种交互关系, 对于 5 个属性的数据从潜在的 26 种交互关系中随机地选取 10 种交互关系。

5.2.2　被试模拟

为了使模拟研究中属性的分布更符合实际测评中被试属性状态的分布, 任一被试 i 在每个属性上的连续能力值向量 $\boldsymbol{\theta}_i=(\theta_{i1},\theta_{i2},\cdots,\theta_{iK})$ 都由多元正态分布 $MVN(0,\boldsymbol{\Sigma})$ 产生, 其中表示属性间相关性协方差矩阵 $\boldsymbol{\Sigma}$ 的非对角线元素的值被设置为 0.5。被试的属性掌握模式 $\boldsymbol{\alpha}_i=(\alpha_{i1},\alpha_{i2},\cdots,\alpha_{iK})$, 其计算方法如公式(19)所示：

$$\alpha_{ik}=\begin{cases}1, & \text{if} \quad \theta_{ik}\geqslant\Phi^{-1}\left(\dfrac{k}{K+1}\right)\\ 0, & \text{otherwise}\end{cases} \tag{19}$$

其中, K 为属性数量, $\Phi(\cdot)$ 为正态分布概率密度函数。源域模拟数据是从由 GDINA 与 DINA 分别在高质量与低质量条件下各生成的数据集中分层采样出 1000 条数据, 即不同 Q 矩阵下的源域数据集都包含 4000 条数据, 分别包含 DINA 模型生成高质量与低质量数据各 1000 条和 GDINA 模型生成的高质量与低质量的数据各 1000 条。目标域数据用相同的方式得到每个条件下源域样本量四分之一的数据, 即目标域数据集共包含 1000 条数据。然后根据不同的样本量 N 从中采样出用于实验的子数据集和测试集。源域数据只生成一次, 模型在源域上被预训练 10 次, 选择平均 Loss 最小和最稳定的模型作为预训练模型。Bi-QNN 的再训练、ANN 训练、参数模型和非参数模型的拟合, 每一次都重新生成目标域数据, 并根据子样本量 N 进行采样。为了减少随机实验误差, 参照聂畅等(2021)实验重复次数设计, 每个不同样本量的实验被重复 30 次。

5.2.3　实验步骤

(1)设置神经网络的超参数, 其中 ANN 与 Bi-QNN 的超参数：训练迭代次数为 100 次, 批量大小在样本量小于等于 100 时设置为 16, 其它情况下设置为 32, 在 $\boldsymbol{Q}_1$ 与 $\boldsymbol{Q}_2$ 生成的数据集上的学习率分别为 0.001, $\boldsymbol{Q}_3$ 生成的数据集上为 0.002。其中, ANN 的网络结构参考 Cui 和聂畅等人的设计(聂畅等, 2021; Cui et al., 2016), ANN 由输入层、隐藏层和输出层 3 层结构组成, 其中隐藏层与 Cui 等(2016)的设计相同。

(2)用每一个条件下的源域模拟数据集对 Bi-QNN 进行预训练, 其中表示 Bi-QNN 网络中交互效应神经元数量的交互式 Q 矩阵利用公式(6)计算获得。

(3)完成 Bi-QNN 预训练模型在目标域上的迁移和对 ANN 的训练, 即从目标域数据集中采样出子训练集和测试集, 并在训练集上对 Bi-QNN 进行微调和对 ANN 的训练。

(4)用(3)中训练得到的 Bi-QNN 和 ANN 模型在测试集上进行预测, 同时在测试集上拟合参数模型和非参数模型, 最后统计各个模型在不同性能指标上的得分。

5.3　评价指标

为了评价模型预测值与真实值的误差, 均方根误差(RMSE)被选作衡量模型整体预测误差的评价指标, 其定义如公式(20)所示：

$$\mathrm{RMSE}=\sqrt{\frac{1}{N\times K}\sum_{i=1}^{N}\sum_{k=1}^{K}(\alpha_{ik}-\hat{p}_{ik})^2} \tag{20}$$

其中, N 表示样本量, K 为属性数, $\alpha_{ik}\in\{0,1\}$, 它别表示被试 i 在属性 k 上真实的属性掌握情况。$\hat{p}_{ik}=p(\hat{\alpha}_{ik}=1)$ 表示模型预测的被试 i 在属性 k 上的掌握概率, 其中 $\hat{p}_{ik}\in(0,1)$。

为了全面的评估 Bi-QNN 方法对被试属性掌握

情况预测的反真率，属性判准率(Attribute Match Ratio, AMR)和模式判准率(Pattern Match Ration, PMR)被作为模型性能的评价指标(Wang et al., 2015)。其中, AMR 和 PMR 的定义如公式(21)和(22)所示：

$$\mathrm{AMR}=\frac{1}{N\times K}\sum_{i=1}^{N}\sum_{k=1}^{K}\mathrm{I}(\alpha_{ik}=\hat{\alpha}_{ik}) \tag{21}$$

其中，AMR 表示模型预测的属性与真实属性的一致程度，I(·) 为指示函数，$\hat{\alpha}_{ik}\in\{0,1\}$，表示模型预测的被试 i 在属性 k 上的属性掌握情况。

$$\mathrm{PMR}=\frac{1}{N}\sum_{i=1}^{N}\mathrm{I}(\boldsymbol{\alpha}_i=\hat{\boldsymbol{\alpha}}_i) \tag{22}$$

其中, PMR 表示模型预测的被试属性掌握模式与被试真实的属性掌握模式的一致程度，$\boldsymbol{\alpha}_i=(\alpha_{i1},\alpha_{i2},\cdots,\alpha_{iK})$ 与 $\hat{\boldsymbol{\alpha}}_i=(\hat{\alpha}_{i1},\hat{\alpha}_{i2},\cdots,\hat{\alpha}_{iK})$ 分别表示被试 i 真实的属性掌握模式与模型预测的属性掌握模式。

5.4 实验结果

5.4.1 均方根误差结果

表 1 呈现的是各个模型在不同条件下模拟数据集上实验的 RMSE 平均值结果。从项目质量的角度

表 1 各个模型在不同条件下的均方根误差(RMSE)

模拟数据	质量	N	DINA	GDINA	NPC	GNPC	ANN	Bi-QNN
SD1	高	50	0.214	0.248	0.239	0.178	0.411	0.172
		100	0.185	0.200	0.239	0.174	0.384	0.159
		200	0.176	0.176	0.233	0.171	0.327	0.155
		300	0.169	0.171	0.237	0.166	0.276	0.145
		500	0.159	0.163	0.232	0.162	0.213	0.141
		Mean	0.181	0.192	0.236	0.170	0.322	0.154
	低	50	0.374	0.413	0.399	0.365	0.434	0.296
		100	0.362	0.385	0.396	0.364	0.416	0.285
		200	0.343	0.358	0.394	0.351	0.375	0.268
		300	0.340	0.340	0.393	0.342	0.338	0.261
		500	0.338	0.324	0.395	0.345	0.296	0.258
		Mean	0.351	0.364	0.395	0.353	0.372	0.274
SD2	高	50	0.325	0.368	0.338	0.326	0.387	0.273
		100	0.322	0.356	0.332	0.308	0.372	0.270
		200	0.300	0.333	0.320	0.293	0.331	0.258
		300	0.300	0.312	0.318	0.282	0.303	0.251
		500	0.286	0.293	0.313	0.280	0.265	0.243
		Mean	0.307	0.332	0.324	0.298	0.332	0.259
	低	50	0.475	0.512	0.471	0.490	0.405	0.335
		100	0.438	0.499	0.464	0.477	0.405	0.339
		200	0.396	0.460	0.464	0.473	0.372	0.323
		300	0.384	0.433	0.457	0.466	0.353	0.330
		500	0.360	0.396	0.453	0.461	0.328	0.321
		Mean	0.411	0.46	0.462	0.473	0.373	0.330
SD3	高	50	0.350	0.374	0.333	0.284	0.379	0.216
		100	0.327	0.347	0.324	0.261	0.370	0.207
		200	0.319	0.347	0.322	0.253	0.321	0.206
		300	0.306	0.336	0.318	0.249	0.296	0.203
		500	0.298	0.332	0.315	0.247	0.249	0.197
		Mean	0.320	0.347	0.322	0.259	0.323	0.206
	低	50	0.441	0.476	0.473	0.450	0.393	0.298
		100	0.409	0.461	0.451	0.433	0.404	0.305
		200	0.382	0.447	0.443	0.427	0.367	0.298
		300	0.375	0.426	0.441	0.422	0.346	0.298
		500	0.358	0.407	0.438	0.420	0.315	0.287
		Mean	0.393	0.443	0.449	0.430	0.365	0.297

注：SD1、SD2 和 SD3 分别表示 Q 矩阵 $\mathbf{Q}_1$、$\mathbf{Q}_2$ 和 $\mathbf{Q}_3$ 生成的模拟数据，质量高低分别表示试题参数 P_0 和 $1-P_1\in\mathrm{U}(0.05,0.15)$ 与 P_0 和 $1-P_1\in\mathrm{U}(0.15,0.30)$, N 表示样本量。

来看, 在相同条件下, 所有模型在低质量数据集上的均方根误差(RMSE)得分均高于高质量数据集。这一结果表明, 在相同条件下所有模型的预测性能均受到数据质量的影响。其中, 多数情况下 Bi-QNN 模型在相同条件下的受影响程度更小。这表明其在相同条件下对数据质量的敏感性会更低。从样本量的视角分析, 在相同条件下, 所有模型的 RMSE 得分整体上都随样本量的增加而减小。这表明在相同条件下样本量会影响各模型的性能。然而, Bi-QNN 模型的 RMSE 得分随样本量变化的趋势相对缓和, 表明其在相同条件下对样本量的鲁棒性更好。

此外, Bi-QNN 模型在多数条件下的 RMSE 得分均低于其他模型, 这意味着其在相同条件下实现了更小的整体预测误差, 从而表现出更好的系统稳定性。进一步地, 在相同条件下, 各个模型在模拟数据集 SD1、SD2 和 SD3 上的 RMSE 得分整体上呈现出 SD2 > SD3 > SD1 的趋势。这表明各个模型在复杂度较低的数据集上具有更好的表现。具体而言, 在题目数量相同且属性数量较少与属性数量相同题目数量较多时, 模型展现出更佳的性能和稳定性。前者可以对比各个模型在数据集 SD1 与 SD2 上的 RMSE 得分表现, 后者可对比各模型在数据集 SD2 与 SD3 上的 RMSE 得分表现。

综上所述, 在相同条件下, 所有模型在高质量、较大样本量以及复杂度较低的数据集上均表现出更小的整体预测误差。其中, Bi-QNN 模型在所有测试条件下的 RMSE 得分均优于其他模型, 验证了该方法的预测可靠性和稳定性。

5.4.2　属性一致性分类结果

表 2 呈现的是各个方法在不同条实验件下的属

表 2　各个模型在不同条件下的属性判准率(AMR)

模拟数据	质量	*N*	DINA	GDINA	NPC	GNPC	ANN	Bi-QNN
SD1	高	50	0.945	0.929	0.942	0.966	0.816	0.963
		100	0.957	0.953	0.942	0.968	0.814	0.968
		200	0.959	0.962	0.944	0.970	0.892	0.970
		300	0.962	0.964	0.944	0.972	0.929	0.972
		500	0.965	0.967	0.946	0.973	0.966	0.975
		Mean	0.958	0.955	0.944	0.970	0.883	0.970
	低	50	0.835	0.818	0.839	0.865	0.764	0.886
		100	0.840	0.830	0.843	0.867	0.776	0.892
		200	0.85	0.844	0.844	0.876	0.810	0.904
		300	0.852	0.854	0.845	0.882	0.860	0.908
		500	0.853	0.864	0.846	0.883	0.896	0.911
		Mean	0.846	0.842	0.843	0.875	0.821	0.900
SD2	高	50	0.878	0.852	0.885	0.898	0.805	0.906
		100	0.877	0.858	0.890	0.907	0.825	0.907
		200	0.887	0.867	0.898	0.915	0.868	0.917
		300	0.889	0.880	0.899	0.920	0.886	0.921
		500	0.899	0.892	0.902	0.921	0.917	0.926
		Mean	0.886	0.870	0.895	0.912	0.860	0.915
	低	50	0.748	0.723	0.777	0.759	0.773	0.846
		100	0.765	0.726	0.785	0.772	0.771	0.845
		200	0.792	0.748	0.785	0.776	0.815	0.855
		300	0.802	0.767	0.790	0.782	0.831	0.858
		500	0.823	0.794	0.796	0.788	0.854	0.858
		Mean	0.786	0.752	0.787	0.775	0.809	0.852
SD3	高	50	0.865	0.847	0.889	0.918	0.825	0.940
		100	0.877	0.861	0.895	0.933	0.825	0.938
		200	0.882	0.866	0.896	0.936	0.876	0.945
		300	0.888	0.873	0.899	0.938	0.897	0.947
		500	0.894	0.875	0.901	0.940	0.931	0.950
		Mean	0.881	0.864	0.896	0.933	0.871	0.944
	低	50	0.772	0.766	0.775	0.797	0.793	0.882
		100	0.794	0.770	0.797	0.813	0.775	0.866
		200	0.811	0.771	0.803	0.817	0.825	0.882
		300	0.818	0.783	0.806	0.822	0.840	0.882
		500	0.834	0.795	0.808	0.823	0.869	0.888
		Mean	0.806	0.777	0.798	0.814	0.820	0.880

性分类准确率(AMR), 表中的值为实验的平均值。观察表中各模型在模拟数据集 SD1、SD2 和 SD3 上的属性分类准确率(AMR)得分表现。从题目质量的角度来看, 在相同条件下, 各模型在高质量数据集上的 AMR 得分显著高于低质量数据集。具体而言, 参数化模型与非参数化模型的 AMR 得分在低质量数据集上平均下降约 0.1, 而 Bi-QNN 与 ANN 模型的下降幅度相对较小, 仅约为 0.05, 这表明神经网络模型对题目质量的鲁棒性更强, 并且 Bi-QNN 模型在整体上都取得了更高的 AMR 得分, 体现出其在处理不同题目质量场景下的优势。

从样本量角度分析, 在相同条件下, 各模型在样本量较大的数据集上普遍取得了更高的 AMR 得分, 说明更大的样本量有助于提升模型的稳定性与准确性。值得注意的是, ANN 模型在样本量较低的数据集上其性能随样本量的波动趋势更大。例如在低质量的 SD3 数据集上, 当样本量为 100 时, ANN 与 Bi-QNN 的 AMR 得分均出现下滑, 表明在题目数量较多且样本量较小时, 神经网络模型因输入的神经元数量增多并在样本量较小时会增加模型训练的不稳定性。

进一步比较各模型在不同数据集上的表现, 可以发现各模型在数据集 SD1 上的表现整体优于 SD3, 而 SD3 又优于 SD2。对比各模型在数据集 SD1 与 SD2 中的表现, 所有模型在 SD1 上都获得了更高的 AMR 得分。表明在题目数量相同时, 属性数量的减少有助于提升模型性能。同时也说明属性数量是影响模型性能的重要因素, 且在这方面 Bi-QNN 展现出更强的适应能力与性能稳定性。对比各模型在数据集 SD2 与 SD3 的表现, 即在属性数量相同时, 各个模型在题目数量较多的 SD3 数据集上表现更佳。说明在属性数量一定时, 增加题目数量能够提升模型预测的可靠性与表现力。此外, 对比各个模型在数据集 SD1 与 SD3 上的表现, 各个模型在 SD1 数据集上取得的 AMR 得分都更高, 即在题目数与属性数之比(Ratio of number of items to attribute , JK)相同的情况下, 各模型在复杂程度更低的数据集上具有更好的表现。另外, 进一步分析各个模型的表现从数据 SD2 到 SD1 与 SD3 表现的变化来看, 在 SD1 上的表现更好, 说明模型受属性数量的影响会比题目数量的影响更大, 即属性数量相比题目数量对模型性能的影响更大。从影响程度上看, 参数化模型受属性数量的影响更大, Bi-QNN 在属性数量变化时整体的表现更平稳, 显示出其对属性数量更高的适应性和稳定性。

综上所述, 模型性能受到样本量、题目质量、属性数量以及题目数量的综合影响。其中, 属性数量对模型的影响相较于题目数量更显著, 而高质量数据和大样本量则有助于提升预测精度。从平均表现来看, 除在样本量为 50 的高质量 SD1 数据集上 GNPC 略优于 Bi-QNN 外, 在其它条件下, Bi-QNN 均取得了更优的 AMR 得分。总体而言, Bi-QNN 模型在多种模拟条件下都表现出更优秀的性能和稳定性, 验证了其在认知诊断任务中的有效性。

5.4.3 模式一致性分类结果

表 3 展示了各模型在不同实验条件下的模式分类准确率(PMR), 其呈现形式与表 2 相似, 不同之处在于 PMR 相比 AMR 采用了更为苛刻的评估标准。

整体来看, 表 3 中的结果表明, 在相同条件下, 各模型在 PMR 上的得分普遍低于在 AMR 上的得分, 这一趋势与过去研究结论一致(Chen & Yan, 2021; Nájera et al., 2021)。从整体趋势看, 各模型在 PMR 上的表现与其在 AMR 上的表现具有高度一致性。具体而言, 各模型在高质量、样本量较大、属性数量较少的数据集上普遍获得更高的 PMR 得分。例如, 在相同条件下, 模型在数据集 SD1 上的 PMR 得分高于 SD3, 在 SD3 上的得分又高于 SD2, 再次表明数据质量、样本量和属性数量是影响模型性能的重要因素。

另外, 需要注意的是, 在高质量的 SD1 数据集上, GNPC 模型在样本量小于或等于 200 的子数据集上取得了全方位领先其它模型的 PMR 表现, 而在该条件下当样本量超过 200 时, Bi-QNN 模型则在 PMR 上实现了对所有模型的超越。这一现象说明, 在高质量、属性较少的小样本场景中, 非参数模型仍然具有一定优势。但随着样本量的增加, Bi-QNN 更能发挥其良好的性能, 从而取得更好的整体表现。此外, Bi-QNN 模型在绝大多数实验条件下在 PMR 指标上均表现出全面领先的优势, 尤其在低质量数据集中的表现尤为突出。具体而言, 在低质量的 SD1、SD2 和 SD3 数据集上, Bi-QNN 模型的 PMR 得分平均领先其他模型约 15%至 20%, 而在 AMR 指标上的平均领先幅度为 5%至 10%。这也进一步说 Bi-QNN 在应对题目质量较低的数据时能够更有效地捕捉潜在的认知模式, 从而表现出具有更强的稳健性和可靠性。

表 3　各个模型在不同条件下的属性模式分类准确率(PMR)

模拟数据	质量	N	DINA	GDINA	NPC	GNPC	ANN	Bi-QNN
SD1	高	50	0.847	0.797	0.836	0.907	0.498	0.883
		100	0.878	0.864	0.835	0.912	0.497	0.904
		200	0.883	0.891	0.840	0.917	0.716	0.908
		300	0.891	0.897	0.841	0.923	0.811	0.923
		500	0.899	0.904	0.847	0.925	0.904	0.925
		Mean	0.880	0.871	0.840	0.917	0.685	0.909
	低	50	0.591	0.539	0.604	0.652	0.378	0.694
		100	0.601	0.563	0.606	0.654	0.424	0.711
		200	0.622	0.594	0.610	0.679	0.505	0.742
		300	0.624	0.615	0.606	0.690	0.636	0.751
		500	0.629	0.637	0.607	0.689	0.726	0.757
		Mean	0.613	0.590	0.607	0.673	0.534	0.731
SD2	高	50	0.550	0.460	0.554	0.593	0.321	0.615
		100	0.536	0.466	0.574	0.619	0.370	0.621
		200	0.568	0.483	0.593	0.649	0.480	0.657
		300	0.571	0.525	0.602	0.673	0.537	0.669
		500	0.604	0.551	0.609	0.675	0.653	0.689
		Mean	0.566	0.497	0.586	0.642	0.472	0.650
	低	50	0.269	0.193	0.318	0.247	0.299	0.436
		100	0.293	0.201	0.334	0.282	0.247	0.422
		200	0.337	0.234	0.339	0.288	0.344	0.447
		300	0.362	0.274	0.350	0.298	0.386	0.457
		500	0.407	0.328	0.362	0.309	0.457	0.461
		Mean	0.334	0.246	0.341	0.285	0.347	0.445
SD3	高	50	0.527	0.459	0.570	0.680	0.349	0.728
		100	0.544	0.468	0.587	0.716	0.349	0.729
		200	0.572	0.477	0.593	0.729	0.496	0.754
		300	0.587	0.481	0.599	0.738	0.583	0.765
		500	0.595	0.479	0.601	0.739	0.715	0.779
		Mean	0.565	0.473	0.590	0.720	0.498	0.751
	低	50	0.325	0.261	0.331	0.323	0.320	0.527
		100	0.356	0.262	0.363	0.371	0.307	0.507
		200	0.393	0.272	0.376	0.376	0.353	0.538
		300	0.402	0.286	0.380	0.376	0.402	0.538
		500	0.435	0.313	0.382	0.386	0.488	0.558
		Mean	0.382	0.279	0.366	0.366	0.374	0.534

综上所述, PMR 指标不仅验证了各模型在更加严格分类标准下的适应能力, 也进一步突出 Bi-QNN 模型在复杂数据情境下的表现。其在高质量数据和低质量数据中均能取得较好的成绩, 尤其在低质量数据集的属性掌握模式识别上表现出较强的稳定性与泛化能力。

6　实证数据分析

6.1　数据与实验说明

为了进一步验证基于迁移学习训练 Bi-QNN 模型的有效性, 认知诊断领域中被广泛使用的分数减法数据集被用来作为实证分析的实验数据集。该数据集包含 536 个作答样本, 在本研究中, 分别使用了包含 20 个题项与 8 个属性的原始数据集(FRAC)以及缩减后包含 15 个题项与 5 个属性的子数据集(Sub FRAC)。二者可从 R 软件包 CDM 中获取, 其中前者的Q矩阵同DeCarlo (2011)研究中的表1, 后者的 Q 矩阵同 DeCarlo (2012)研究中的表 7, 具体的题目详情请参见原始文献。

依据数据集 FRAC 和 Sub FRAC 各自的 Q 矩阵和分数减法专业知识, 构建它们各自的属性交互关系矩阵如公式(23)和(24)所示：

$$\boldsymbol{Q}_1^{\#\mathrm{T}} = \begin{pmatrix} 1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 1&1&1&1&1&1&1&1&1&1&0&1&1&0&0&0&1 \\ 0&0&1&1&0&0&0&0&1&1&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&1&1 \\ 0&0&0&0&1&1&1&1&1&1&1&0&0&0&1&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&1&0&1&1&0&0 \\ 0&1&0&1&0&1&0&1&0&1&1&0&1&1&1&1&1 \\ 0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0 \end{pmatrix} \tag{23}$$

$$\boldsymbol{Q}_2^{\#} = \begin{pmatrix} 1&1&0&0&0 \\ 1&0&1&0&0 \\ 0&1&1&0&0 \\ 0&0&1&1&0 \\ 1&1&1&0&0 \\ 1&0&1&1&0 \\ 0&1&1&1&0 \\ 0&0&1&1&1 \end{pmatrix} \tag{24}$$

其中，$\boldsymbol{Q}_1^{\#}$ 为 FRAC 的交互关系矩阵，其维度为 17×8 表示依据分数减法专业知识从数据集 FRAC 包含的 8 个属中存在的 $2^8-9=247$ 种交互关系中选择 17 种典型的属性交互类型；$\boldsymbol{Q}_2^{\#}$ 为 Sub FRAC 的交互关系矩阵，其维度为 8×5，表示从 $2^5-6=26$ 中潜在的交互关系中，选择了 8 种典型的属性交互类型。为了方便理解以 $\boldsymbol{Q}_2^{\#}$ 的第一行为例对属性之间的交互关系进行说明，即 $\boldsymbol{q}_{2_1}^{\#}=(11000)$，它表示在 Sub FRAC 数据集中的属性 a 与属性 b 之间存在相互关系。具体地，根据 Sub FRAC 数据集的 Q 矩阵，表示在化解分数计算结果为最简分式前，需要先完成基本的分数减法运算。其它的，即第 1 行以后表示的交互关系以此类推。

Bi-QNN 的预训练数据与模拟研究生成预训练数据的方式相同，分别基于 FRAC 和 Sub FRAC 各自的 Q 矩阵，用 GDINA 和 DINA 模型分别生成各自的预训练模拟数据集，二者的模拟数据被我们分别设置为包含 800 和 1600 条高质量和低质量的数据。由于原始的实证数据集 FRAC 与 Sub FRAC 都只包含了 Q 矩阵与被试的作答反应数据，并没有提供专家诊断的被试属性掌握模式数据。过去的研究者们通过比较模型在完整数据集与其子数据集上的分类一致率来评估模型的性能(Chiu et al., 2018; Wang et al., 2024)。因此，我们参照汪大勋等(2023)的实验设计，用 GDINA 和 DINA 模型在 FRAC 与 Sub FRAC 的完整数据集上的预测结果作为基线，分别以 N = (50,100,200,300,500)分层采样出用于微调的训练集与测试集。另外，在实证数据分析实验中，Bi-QNN 的预训练重复次数与实验结果的重复次数与模拟研究相同。在 FRAC 与 Sub FRAC 两个数据集上的实验中，Bi-QNN 模型使用了相同的超参数配置：迭代次数设置为 200 次，批量大小为 128，学习率均为 0.01。

6.2 实验结果

6.2.1 分数减法子数据集实验结果

表 4 呈现了各个模型在 Sub FRAC 数据集上不同样本量下的属性与模式的分类准确性结果。与模拟实验一样，样本量 N 仍然为 50 至 500 共 5 个子集。除此之外，对于模式分类一致性增加 PRM(K-1)指标，即放宽了属性掌握模式分类一致率有一个属性分类错误情况的考虑。首先是聚焦于各个模型在 Sub FRAC 数据集上的 AMR 的性能表现。从整体上来看，基于神经网络的方法中 Bi-QNN 模型在各个子数据集上的 AMR 表现都优于其它模型，ANN 在较大样本量(比如 $N\geqslant300$)的子数据集上的表现仅次于 Bi-QNN，参数化的模型(DINA 和 GDINA)的整体表现次之，非参数化的模型(NPC 和 GNPC)整体表现相对较差。

其次，样本量与模型表现的趋势方面与模拟研究相似，各模型在 AMR 上的得分整体上都呈现出随样本量增大而增大的趋势。其中，非参数模型中 NPC 方法表现相对平稳，GNPC 模型的在样本量 N 大于 100 后出现略微下降的趋势，这可能与 GNPC 模型在计算分类模式时同时考虑连接机制与非连接机制的权重引起的。最后是各个模型在 PMR 与 PMR(K-1)上的表现。从整体上看，因为 PMR(K-1)比 PMR 增加了一个属性分类错误的容错，整体上各个模型在 PMR(K-1)上的得分比 PMR 高。另外，各个模型在两个属性一致性指标上的表现与在 AMR 的趋势保持一致，Bi-QNN 模型在 Sub FRAC 的各个子数据集上的表现整体上保持优势。其中，样本量为 50 时，DINA 模型在 PMR 上的表现略优于 Bi-QNN 模型。

6.2.2 分数减法原始数据集结果

表 5 与表 4 具有相同的结构，它呈现的是各个模型在 FRAC 数据集上的属性与模式的分类准确性。从整体上看，各个模型在 FRAC 数据集上的表现与在 Sub FRAC 上具有相似的趋势，Bi-QNN 模型的表现仍然全面的保持对其它模型的领先，其次是 ANN 模型。另外，各个模型在 FRAC 数据集上的表现都弱于在 Sub FRAC 数据集上的表现，特别地，

表 4　各个模型在 Sub FRAC 数据集上的属性与模式分类一致性

真实数据	性能指标	N	DINA	GDINA	NPC	GNPC	ANN	Bi-QNN
SFRAC	AMR	50	0.920	0.906	0.840	0.854	0.887	0.923
		100	0.931	0.896	0.841	0.860	0.927	0.956
		200	0.942	0.920	0.848	0.812	0.951	0.959
		300	0.953	0.919	0.851	0.786	0.953	0.963
		500	0.953	0.929	0.848	0.773	0.950	0.960
		Mean	0.940	0.914	0.846	0.817	0.934	0.952
	PMR	50	0.706	0.664	0.508	0.500	0.572	0.674
		100	0.742	0.630	0.530	0.541	0.701	0.803
		200	0.761	0.710	0.533	0.448	0.777	0.820
		300	0.807	0.700	0.550	0.421	0.795	0.836
		500	0.799	0.736	0.548	0.392	0.785	0.829
		Mean	0.763	0.688	0.534	0.460	0.726	0.792
	PMR(K–1)	50	0.908	0.872	0.756	0.798	0.890	0.946
		100	0.926	0.874	0.754	0.800	0.943	0.979
		200	0.951	0.905	0.768	0.709	0.978	0.976
		300	0.960	0.903	0.779	0.661	0.973	0.980
		500	0.966	0.917	0.772	0.613	0.964	0.974
		Mean	0.942	0.894	0.766	0.716	0.950	0.971

表 5　各个模型在 FRAC 数据集上的属性与模式分类一致性

真实数据	性能指标	N	DINA	GDINA	NPC	GNPC	ANN	Bi-QNN
FRAC	AMR	50	0.919	0.887	0.837	0.854	0.901	0.946
		100	0.926	0.889	0.829	0.841	0.947	0.956
		200	0.940	0.908	0.835	0.827	0.964	0.964
		300	0.949	0.921	0.842	0.829	0.960	0.964
		500	0.955	0.932	0.837	0.823	0.958	0.962
		Mean	0.938	0.907	0.836	0.835	0.946	0.958
	PMR	50	0.529	0.450	0.346	0.340	0.460	0.652
		100	0.575	0.465	0.315	0.314	0.663	0.707
		200	0.639	0.508	0.351	0.311	0.741	0.752
		300	0.690	0.575	0.384	0.327	0.722	0.755
		500	0.720	0.599	0.374	0.308	0.716	0.738
		Mean	0.631	0.519	0.354	0.320	0.660	0.721
	PMR(K–1)	50	0.866	0.754	0.637	0.666	0.787	0.931
		100	0.867	0.753	0.619	0.602	0.919	0.954
		200	0.901	0.827	0.638	0.598	0.97	0.965
		300	0.924	0.850	0.647	0.591	0.958	0.961
		500	0.930	0.886	0.641	0.575	0.952	0.962
		Mean	0.898	0.814	0.636	0.606	0.917	0.955

Bi-QNN 模型在 FRAC 数据集上的 AMR 得分与 Sub FRAC 没有明显的区别, 而在 PMR 与 PMR(K–1)的得分略有降低, 相比其它模型, 其下降程度最低。

除此之外, 基于神经网络的方法(Bi-QNN 和 ANN)相对参数化方法(DINA 和 GDINA)与非参数方法(NPC 和 GNPC)在 FRAC 数据上在 PMR 与 PMR(K-1)的整体得分比在 Sub FRAC 数据上的分差更大, 这表明在 FRAC 数据集上基于神经网络的方法具有更大的优势。特别地, 当样本量 N 为 50 时, Bi-QNN 模型在 FRAC 数据集上的 AMR、PMR 与 PMR(K-1)三个指标上都有明显的领先, 比在 Sub FRAC 数据上的优势更突出。

7 讨论与结论

7.1 讨论与局限

7.1.1 Bi-QNN 的模型设计

Bi-QNN的设计来源于GDINA模型, 包含了表示主效应与交互效应两条计算流, 并分别通过Q矩阵与交互式Q矩阵进行约束, 这使得其网络结构具有较好的语义与可解释性。并且这种约束类似于Dropout方法, 有助于提高模型的鲁棒性(Srivastava et al., 2014)。除此之外, Bi-QNN 网络的宽度也由 Q 矩阵和交互式Q矩阵约束, 让其网络结构可以跟随任意数据集自适应地调整, 对于使用人员来说可以不用亲自去思考网络结构的深度和宽度等专业的问题, 可以极大地提高 Bi-QNN 的易用性与可推广性。这是Bi-QNN与Cui等(2016)、Chen和Yan (2021)等人提出的基于固定结构的神经网络认知诊断方法最大的不同。

这样的设计, 使 Bi-QNN 模型在一定程度上降低了对属性数量的敏感性, 从而在一定程度上克服传统的参数模型与非参数模型对属性数量较为敏感, 无法在具有较多属性的数据上取得较好可靠性的问题(Sen & Cohen, 2021)。我们可以从前面的实验分析中找到证据, 以数据集 SD2 为基准, 所有模型在数据集 SD1 和 SD3 上的表现均优于 SD2, 且 SD1 的表现又优于 SD3。从实验结果看, 观察各个模型从数据 SD2 到数据 SD1 与 SD3 的性能提升表现, 在 SD1 上获得了更高的性能提升, 表明在相同条件下, 降低相同程度的属性数量与增加相同程度的题目数量, 模型在前者上取得更好的表现, 说明模型更容易受到属性数量的影响, 这与以往的研究结论是一致的(Nájera et al., 2021)。而 Bi-QNN 模在数据集 SD2 和 SD3 都取得了更好的表现。尤其是在低质量的 SD2 和 SD3 数据集上, Bi-QNN 模型在 AMR 上的得分整体上领先约 10%, 而 PMR 得分整体领先约 15%到 20%, 并且实证数据中 Bi-QNN 模型在 FRAC 数据集上也表现出更小的性能变化, 进一步说 Bi-QNN 在一定程度上确实可以缓解属性数量增长对模型性能的影响, 并在应对题目质量较低的数据时具有更强的稳健性和可靠性, 能够更有效地捕捉潜在的认知模式。

7.1.2 Bi-QNN 模型的训练方法

神经网络方法应用于认知诊断, 除了在网络结构的设计上存在困难, 还面临难以保障潜变量标签的正确性问题。即在实际测验中, 大多数题目和试卷都是一次性的, 题目的重复利用率较低, 给属性标注带来极大的挑战。因此, 会经常存在缺乏大量标注有被试作答模式与属性掌握模式的数据集, 导致神经网络模型的训练变得极为困难。对于神经网络模型的训练 Cui 等(2016)使用理想的反应模式和理想的掌握模式来训练, Xue 和 Bradshaw (2021)基于 DINA 模型以半监督协同训练方式进行训练。然而, 对于给定的Q矩阵其对应的理想反应模式和理想掌握模式非常有限(de la Torre, 2009), 尤其是理想反应模式通常难以覆盖被试的真实作答模式。同时在实际测验中, 被试实际属性掌握模式的类别通常远少于理想掌握模式的类别, 这使得基于理想数据训练的神经网络模型在测试集上的泛化能力较差。虽然基于 DINA 的半监督训练方法可以在一定程度上利用非专家标记的学生反应数据, 但 DINA 模型难以表达属性间的交互效应, 同时该方法对初始标签和标签噪声敏感, 限制了模型性能的进一步提高。本文采用迁移学习中预训练−微调的方式来对 Bi-QNN 模型进行训练。首先让 Bi-QNN 在较大样本量的模拟数据集上进行预训练, 然后在到真正要诊断的小数据集上进行微调。从而实现在较少样本(比如 $N = 100$)的情况下完成对 Bi-QNN 的训练。这样不仅使 Bi-QNN 在不同样本量的数据集上保持良好的性能, 还可以提升 Bi-QNN 的普适性与可推广性, 这在教学测验实践中显得尤为重要。

以上结论可以从实验中找到证据, 在模拟实验中, 相同条件下 Bi-QNN 模型在不同样本量数据集上的 AMR 和 PMR 得分相比其它模型变化更平稳, 即在样本量较小的情况下 Bi-QNN 模型仍然可以保持较好的性能, 尤其是在低质量的数据集上更为突出。比如在低质量的 SD2 和 SD3 数据集上 Bi-QNN 在保持 PMR 得分领先的同时, 在不同样本量上的 PMR 得分以 0.016 和 0.019 的标准差保持最平稳。另外, Bi-QNN 模型在低质量的 SD1、SD2 和 SD3 数据集上, 当样本量为 50 时在 AMR 和 PMR 得分上的领先程度相比其它样本量具有更明显的优势。从模型的整体系统误差上来看, Bi-QNN 也在更小的样本量下取得了更突出的领先优势。而在真实的数据集 Sub FRAC 与 FRAC 上, 因试题质量的分布范围更小, Bi-QNN 在 AMR 与 PMR 上的得分几乎都取得了在任意样本量下的领先。

7.1.3 局限与展望

尽管 Bi-QNN 在模拟数据集和真实数据集上都表现良好, 但它仍然存在一些局限性。从实验结果

中可以发现, Bi-QNN 虽然在属性数量和样本量上都有相对较好的鲁棒性, 尤其在属性数量的泛化能力上, 可以在一定程度上克服模型对属性数量敏感的问题。实验结果表明, 所有模型都在高质量的数据上取得更好的表现, 这与先前的研究一致(Chen & Yan, 2021; Cui et al., 2016)。由于模拟数据受到参数模型的影响, 使得 Bi-QNN 在一定程度上也会受到题目质量的影响但是其被影响的程度相对其它模型更小, 其性能的平均下降程度约为参数模型和非参数模型的一半, 保持在 0.05 左右。另外虽然基于迁移学习的方法可以让 Bi-QNN 从模拟数据中学习到属性与题目之间的关系。但是当题目质量较高且样本量非常小时($N \leqslant 50$), 神经网络的参数求解仍然面临困难, 此时非参数模型仍然具有优势, 尤其是在属性数量也较少的情况下。比如在样本量为 50 的 SD1 的高质量数据集中 GNPC 分别以 0.966 与 0.907 的 AMR 与 PMR 得分领先所有模型。另外, 尽管基于迁移学习的方法在一定程度上可以缓解数据标注的成本, 但同时也带来了额外的计算开销。即在模型预训练阶段, 由于需要生成大量模拟数据并在其上完成预训练, 因此该阶段通常会消耗更多的计算资源与时间。其中预训练时间与模拟数据集的规模密切相关。在本文实验中, 预训练过程在 CPU 上的平均耗时约为 20 秒, GPU 上约为 30 秒; 微调阶段的耗时随样本量增加, 在 CPU 上约为从 1 到 12 秒不等, GPU 上约为从 1 到 30 秒不等; 对于预测的时间花费皆不超过 1 秒。相比之下, 其它模型的整体运行时间约在 1 秒至 5 秒之间。实验中具体耗费的时间详情见网络版附录。

除此之外, 模拟实验数据主要基于非补偿模型的假设生成, 因此, Bi-QNN 在符合补偿模型假设的数据上的性能和可靠性是否依然优异, 仍需进一步验证。同时, 由于模拟实验中的属性不具备具体含义, 因此我们没有深入探讨交互关系矩阵的设定对模型性能的影响, 后续这将是一个重要的研究焦点。另外, 本研究仅关注二级计分模型, 而在认知诊断的现实需求中, 多级计分模型的应用场景更为丰富, 未来将深度学习的方法应用到此类任务上更需要进一步探索。由于目前的方法只关注到了被试的作答结果, 没有利用被试的作答过程数据, 而神经网络非常善于利用多模态数据进行学习, 在未来可以考虑融入学习者的作答行为数据以降低模型对猜测和失误的敏感性(田亚淑 等, 2023)。对于属性之间的交互关系在 Bi-QNN 中通过基于 Q 矩阵和交互式 Q 矩阵来实现, 未来可以考虑融入特定学科的知识图谱以更好的表达知识之间的相关关系来帮助提升模型的性能。

7.2　研究结论

本研究提出了融入 Q 矩阵约束的神经网络认知诊断模型 Bi-QNN, 并基于迁移学习方法完成对其训练, 从研究结果可以得出以下结论:

(1) Bi-QNN 模型除了可以较好地表达属性与题目之间的关系, 其网络结构能自动适应数据, 用户在使用它时仅需提供 Q 矩阵与属性交互关系矩阵即可, 具有较好的易用性与可推广性。另外, 实验结果表明, Bi-QNN 能在一定程度上缓解属性数量增加模型性能降低的影响, 从而在一定范围内, 当属性数量增加时模型仍然能具有较好性能。

(2)基于迁移学习训练 Bi-QNN 模型可以有效的解决认知诊断场景中样本量不足的问题, 同时在大量的模拟数据上学习到的属性与题目之间的关系, 迁移到特定的诊断数据集中使 Bi-QNN 在不同的样本量上都具有较好的稳健性。尤其在数据质量较低和属性数量相对较多的情况下, 这一表现更为突出。

(3)模型的选择建议, 从模拟研究结果看, 在属性数量较少($N < 5$)且样本量也较少($N < 50$)时非参数模型具有更突出的优势。但是, 从模拟数据与真实数据分析的结果表明, Bi-QNN 模型在大多情况下的整体表现都更好, 尤其在实证数据上 Bi-QNN 的表现领先所有模型, 因此, 可以看出 Bi-QNN 是一个可以胜任多种条件下的通用模型。

(4)模型的适用场景, 综合 Bi-QNN 的易用性、通用性以及在一定的范围内对属性数量的耐受性可以得出 Bi-QNN 模型的使用场景。首先, Bi-QNN 非常适合用在心理专家开发好心理测评工具的测评场景, 即一旦根据测评量表训练好 Bi-QNN 模型, 则模型即可被反复使用。其次, 在教学中, Bi-QNN 非常适合作为在线学习系统中的通用测评模型, 尤其是具有知识图谱支持的学习系统中, 模型可根据组卷算法自动完成训练。而在课堂教学环境中, 当属性数量相对较多($K > 5$)且样本量大于 50 的情境下, Bi-QNN 仍然是一个综合而言较好的选择。

参　考　文　献

Chen, D., & Yan, C. (2021). Classification of attribute mastery patterns using deep learning. *Open Journal of Modelling and Simulation*, *9*(2), 198−210.

Chiu, C.-Y., & Douglas, J. (2013). A nonparametric approach

to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, *30*(2), 225−250.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*(4), 633−665.

Chiu, C.-Y., Sun, Y., & Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika*, *83*(2), 355−375.

Cui, Y., Gierl, M., & Guo, Q. (2016). Statistical classification for cognitive diagnostic assessment: An artificial neural network approach. *Educational Psychology*, *36*(6), 1065−1082.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the q-matrix. *Applied Psychological Measurement*, *35*(1), 8−26.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement*, *36*(6), 447−468.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115−130.

de la Torre, J., & Minchen, N. D. (2019). The G-DINA model framework. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models: Methodology of educational measurement and assessment* (pp. 155–169). Springer, Cham.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179−199.

Gao, X. L., Gong, Y., & Wang, F. (2021). Research progress in polytomous cognitive diagnosis model. *Journal of Psychological Science*, *44*(2), 457−464.

[高旭亮, 龚毅, 王芳. (2021). 多级评分认知诊断模型述评. *心理科学*, *44*(2), 457−464.]

Guo, L., Yang, J., & Song, N. (2020). Spectral clustering algorithm for cognitive diagnostic assessment. *Frontiers in Psychology*, *11,* 944.

Guo, L., & Zhou, W. J. (2021). Nonparametric methods for cognitive diagnosis to multiple-choice test items. *Acta Psychologica Sinica*, *53*(9), 1032−1043.

[郭磊, 周文杰. (2021). 基于选项层面的认知诊断非参数方法. *心理学报*, *53*(9), 1032−1043.]

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3)*,* 258−272.

Kang, C. H., Ren, P., & Zeng, P. F. (2015). Nonparametric cognitive diagnosis: A cluster diagnostic method based on grade response items. *Acta Psychologica Sinica*, *47*(8), 1077−1088.

[康春花, 任平, 曾平飞. (2015). 非参数认知诊断方法: 多级评分的聚类分析. *心理学报*, *47*(8), 1077−1088.]

Kang, C. H., Zhang, S. J., Li, Y. B., & Zeng, P. F. (2019). The cognitive diagnosis of k-nearest neighbor algorithm and its application. *Journal of Jiangxi Normal University* (*Natural Science*), *43*(2), 135−141+159.

[康春花, 张淑君, 李元白, 曾平飞. (2019). KNN 认知诊断法及其应用. *江西师范大学学报*(*自然科学版*), *43*(2), 135−141+159.]

Liu, C., & Cheng, Y. (2018). An application of the support vector machine for attribute-by-attribute classification in cognitive diagnosis. *Applied Psychological Measurement*, *42*(1), 58−72.

Liu, Q. (2021, August). Towards a new generation of cognitive diagnosis. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21* (pp.4961−4964). Montreal, Canada.

Liu, Y., Zhang, T., Wang, X., Yu, G., & Li, T. (2022). New development of cognitive diagnosis models. *Frontiers of Computer Science*, *17*(1), 171604.

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *The British Journal of Mathematical and Statistical Psychology*, *69*(3), 253−275.

Nájera, P., Abad, F. J., & Sorrel, M. A. (2021). Determining the number of attributes in cognitive diagnosis modeling. *Frontier of Psychology, 12,* 614470.

Nie, C., Sun, X. J., & Xin, T. (2021). Factors affecting the classification accuracy in cognitive diagnosis assessment based on BP neural network. *Journal of China Examinations*, (3), 28−35.

[聂畅, 孙小坚, 辛涛. (2021). 基于 BP 神经网络的认知诊断评估分类准确率影响因素分析. *中国考试*, (3), 28−35.]

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345−1359.

Qin, H., & Guo, L. (2024). Using machine learning to improve Q-matrix validation. *Behavior Research Methods*, *56*(3), 1916–1935.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386−408.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1: Foundations, pp. 318−362). MIT Press.

Sen, S., & Cohen, A. S. (2021). Sample size requirements for applying diagnostic classification models. *Frontiers in Psychology*, *11,* 621251.

Song, L. H., Wang, W. Y., & Ding, S. L. (2024). Q-matrix theory and its applications in cognitive diagnostic assessment. *Advances in Psychological Science*, *32*(6), 1010−1033.

[宋丽红, 汪文义, 丁树良. (2024). 认知诊断评估中 Q 矩阵理论及应用. *心理科学进展*, *32*(6), 1010−1033.]

Sorrel, M. A., Escudero, S., Nájera, P., Kreitchmann, R. S., & Vázquez-Lira, R. (2023). Exploring approaches for estimating parameters in cognitive diagnosis models with small sample sizes. *Psych*, *5*(2), 336−349.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929−1958.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A survey on deep transfer learning. In V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, & I. Maglogiannis (Eds.), *Artificial neural networks and machine learning − ICANN 2018* (Vol. 11141, pp. 270−279). Springer.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327−359). Lawrence Erlbaum Associates, Inc.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287−305.

Tian, Y. S., Zhan, P. D., & Wang, L. J. (2023). Joint cognitive diagnostic modeling for probabilistic attributes incorporating item responses and response times. *Acta Psychologica*

Sinica, *55*(9), 1573−1586.
[田亚淑, 詹沛达, 王立君. (2023). 联合作答精度和作答时间的概率态认知诊断模型. *心理学报*, *55*(9), 1573−1586.]
Wang, D., Ma, W., Cai, Y., & Tu, D. (2024). A general nonparametric classification method for multiple strategies in cognitive diagnostic assessment. *Behavior Research Methods*, *56*(2), 723−735.
Wang, D. X., Xiao, Q. W., Tan, Q. R., Cai, Y., & Tu, D. B. (2023). A non-parametric multi-strategy cognitive diagnosis method. *Journal of Psychological Science, 46*(4), 971−979.
[汪大勋, 肖清文, 谭青蓉, 蔡艳, 涂冬波. (2023). 非补偿的非参数化多策略认知诊断方法:NCNPMSC 方法开发. *心理科学, 46*(4), 971−979.]
Wang, F., Liu, Q., Chen, E., Huang, Z., Chen, Y., Yin, Y., … Wang, S. (2020, April). Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the 34th AAAI conference on artificial intelligence* (Vol. 34, No. 04, pp. 6153–6161). New York, NA.
Wang, M., & Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing*, *312*, 135−153.
Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, *52*(4), 457−476.
Wang, W. Y., Ding, S. L., Song, L. H., Kuang, Z., & Gao, H, Y. (2016). Application of neural networks and support vector machines to cognitive diagnosis. *Journal of Psychological Science*, *39*(4), 777−782.
[汪文义, 丁树良, 宋丽红, 邝铮, 曹慧媛. (2016). 神经网络和支持向量机在认知诊断中的应用. *心理科学*, *39*(4), 777−782.]
Wang, W. Y., Song, L. H., Ding, S. L., Wang, T., & Xiong, J. (2021). A probabilistic representation approach for the nonparametric classification method to cognitive diagnosis. *Journal of Psychological Science*, *44*(5), 1249−1258.
[汪文义, 宋丽红, 丁树良, 汪腾, 熊建. (2021). 非参数认知诊断方法下诊断结果的概率化表征. *心理科学*, *44*(5), 1249−1258.]
Wen, H., Liu, Y., & Zhao, N. (2020). Longitudinal cognitive diagnostic assessment based on the HMM/ANN model. *Frontiers in Psychology*, *11,* 2145.
Xin, T., Wang, C., Chen, P., & Liu, Y. (2022). Editorial: Cognitive diagnostic models: Methods for practical applications. *Frontiers in Psychology*, *13,* 895399.
Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika*, *81*(3), 625−649.
Xu, H. Y., Chen, Q. P., Liu, Y. H., & Zhan, P. D. (2023). Nonparametric diagnostic classification for polytomous attributes: A comparison of 18 distance discriminant methods. *Journal of Psychological Science*, *46*(6), 1486−1494.
[徐慧颖, 陈琦鹏, 刘耀辉, 詹沛达. (2023). 多分属性的非参数诊断分类:18 种距离判别法的对比. *心理科学*, *46*(6), 1486−1494.]
Xue, K., & Bradshaw, L. P. (2021). A semi-supervised learning-based diagnostic classification method using artificial neural networks. *Frontiers in Psychology*, *11,* 618336.
Yamaguchi, K. (2023). On the boundary problems in diagnostic classification models. *Behaviormetrika*, *50*(1), 399−429.
Zhan, P., Man, K., Wind, S. A., & Malone, J. (2022). Cognitive diagnosis modeling incorporating response times and fixation counts: Providing comprehensive feedback and accurate diagnosis. *Journal of Educational and Behavioral Statistics*, *47*(6), 736−776.
Zhang, W., Meng, L., & Liang, B. (2022). EW-KNN: Evaluating information technology courses in high school with a non-parametric cognitive diagnosis method. *Interactive Learning Environments*, *31*(10), 6783−6798.
Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43−76.

Cognitive diagnosis method via neural networks with transfer learning and Q-matrix constraints

TAO Jinhong[1], ZHAO Wei[1], CHENG Nuo[1], QIAO Lifang[2], JIANG Qiang[1]

([1] *School of Information Science and Technology, Northeast Normal University, Changchun* 130117, *China*)

([2] *College of Education, Hebei Normal University, Shijiazhuang* 050000, *China*)

Abstract

Cognitive diagnostic assessment (CDA) is an important educational assessment method that identifies the strengths and weaknesses of students in specific cognitive skills or attributes. Artificial neural networks (ANNs) can learn complex, nonlinear relationships from data and have become one of the most widely used machine learning methods in CDA. However, most existing ANN-based CDA methods require users to design the network structure manually, which is a challenging task for education professionals without AI expertise. Moreover, neural network training often encounters scarce labeled data, which limits their usability and applicability in cognitive diagnostic practice. Therefore, a simple and easy-to-use general neural network cognitive diagnosis method that can automatically adapt to different datasets and learning tasks is still lacking.

In this paper, we propose a neural network cognitive diagnosis method (Bi-QNN) that is constrained by the Q-matrix and an attribute interaction matrix and uses transfer learning for training. Our method has the

following advantages: (1) Its network structure can be automatically constructed according to the Q-matrix and interaction matrix corresponding to any dataset, eliminating the need for manual design of the neural network. (2) The network structure design of the new model is inspired by the GDINA model, which can better express and capture the main and interaction effects of attributes. (3) The model training scheme based on transfer learning helps address the scarcity of labeled data, thereby improving the usability and wider applicability of the model.

To evaluate the performance of Bi-QNN, we conduct extensive experiments on simulated and real datasets covering various scenarios of CDA. Experimental results show that Bi-QNN has lower prediction errors on the simulated datasets than the parametric methods GDINA and DINA, indicating a better fit to the data. Our model is robust to the number of attributes and maintains high classification accuracy as this number increases, demonstrating that Bi-QNN can handle complex problems with more attributes in CDA. The training method based on transfer learning enables Bi-QNN to adapt effectively to datasets with varying sample sizes, maintaining superior performance compared with other models across multiple conditions in simulated and empirical datasets. Bi-QNN generally outperforms other models, suggesting that it can benefit from knowledge transfer and can generalize to new domains.

Bi-QNN is a simple, easy-to-use general neural network cognitive diagnosis method with good expressiveness and adaptability. It can provide more accurate and reliable diagnostic feedback for students and teachers and facilitate personalized and adaptive learning. The improvement in model performance is limited by the reliance on simulated data, and the model remains slightly sensitive to the quality of the test items. These issues need to be verified and improved on more datasets.

Keywords cognitive diagnostic assessment, Q-matrix, artificial neural network, transfer learning

附录:

附表 1　Bi-QNN 模型在各数据集上的训练时间统计

数据集	质量	预训练		样本量	微调	
		CPU	GPU		CPU	GPU
SD1	高	18.685	30.479	50	1.258	3.103
				100	2.004	5.145
				200	3.613	9.206
				300	5.158	14.330
				500	10.017	23.457
	低			50	1.739	3.131
				100	2.610	5.198
				200	4.100	9.666
				300	5.942	14.693
				500	10.814	27.488
SD2	高	19.504	30.229	50	1.783	3.478
				100	2.260	5.654
				200	4.103	10.063
				300	5.213	14.456
				500	10.960	27.522
	低			50	1.823	3.491
				100	2.284	5.856
				200	4.130	10.375
				300	5.418	14.502
				500	11.811	27.588
SD3	高	21.764	31.407	50	1.890	3.534
				100	2.519	5.931
				200	4.293	10.697
				300	5.433	14.481
				500	11.891	27.578
	低			50	1.906	3.525
				100	2.730	5.905
				200	4.360	10.702
				300	5.782	14.503
				500	11.935	27.612
SFRAC		20.875	30.717	50	1.794	3.513
				100	2.249	5.897
				200	4.233	10.319
				300	6.126	14.527
				500	11.742	27.582
FRAC		22.332	32.942	50	1.914	3.649
				100	2.741	5.908
				200	4.541	10.423
				300	5.785	15.832
				500	11.995	28.182

注：CPU 型号: Intel(R)Core(TM) i7; GPU 型号：RTX3070; 时间单位为秒; 表中任意条件下模型的预测时间皆小于 1 秒。