基于特征深度融合的 Web 服务 QoS 联合预测

刘建勋[1,2]，丁领航[1,2]，康国胜[1,2]，曹步清[1,2]，肖勇[1,2]

（1. 湖南科技大学服务计算与软件新技术湖南省重点实验室，湖南 湘潭 411201；
2. 湖南科技大学计算机科学与工程学院，湖南 湘潭 411201）

摘 要：为了解决 Web 服务 QoS 预测准确度不够的问题，针对 QoS 中隐藏的环境偏好信息和多类 QoS 隐藏的共同特征，提出一种基于特征深度融合的 Web 服务 QoS 联合预测方法。考虑 QoS 数据可以建模为用户–服务二部图，采用多组件图卷积神经网络进行特征提取和映射，采用加权融合方法对多类 QoS 特征进行同维映射。使用注意力因子分解机对映射后的特征向量进行一阶特征、二阶交互特征和高阶交互特征的提取，并结合各部分结果实现 QoS 联合预测。实验结果表明，所提方法在均方根误差和平均绝对误差方面优于现有 QoS 预测方法。

关键词：联合预测；服务质量；偏好特征；深度融合

中图分类号：TN92

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2022107

Joint QoS prediction for Web services based on deep fusion of features

LIU Jianxun[1,2], DING Linghang[1,2], KANG Guosheng[1,2], CAO Buqing[1,2], XIAO Yong[1,2]

1. Hunan Provincial Key Lab for Services Computing and Novel Software Technology, Hunan University of Science and Technology, Xiangtan 411201, China
2. School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201, China

Abstract: In order to solve the problem of insufficient accuracy of Web service QoS prediction, a joint QoS prediction method for Web services based on the deep fusion of features was proposed with considering of the hidden environmental preference information in QoS and the common features of multi-class QoS. First, QoS data was modeled as a user-service bipartite graph and multi-component graph convolution neural network was used for feature extraction and mapping, and the weighted fusion method was used for the same dimensional mapping of multi-class of QoS features. Subsequently, the attention factor decomposition machine was used to extract the first-order features, second-order interactive features, and high-order interactive features of the mapped feature vector. Finally, the results of each part were combined to achieve the joint QoS prediction. The experimental results show that the proposed method is superior to the existing QoS prediction methods in terms of root mean square error (RMSE) and average absolute error (MAE).

Keywords: joint prediction, quality of service, preference feature, deep fusion

0 引言

近年来，随着面向服务架构（SOA, software oriented architecture）、云计算、移动计算等技术的广泛应用，大量 Web 服务被创建并发布在互联网上供人们调用，如何从大量的功能相似的 Web 服务中快速、准确地找到高质量的服务是一个挑战性问题。服务质量（QoS, quality of service）描述了服务的非功能属性，是区分功能相似服务的重要参考依据，广泛用于 QoS 感知的服务发现、服务推荐[1-2]、

收稿日期：2022−02−11；**修回日期：**2022−05−07

通信作者：康国胜，guoshengkang@gmail.com

基金项目：国家自然科学基金资助项目（No.61872139）；湖南省教育厅基金资助项目（No.20B244）

Foundation Items: The National Natural Science Foundation of China (No.61872139), Educational Commission of Hunan Province of China (No.20B244)

服务组合等服务管理任务中。常见的 QoS 属性包括响应时间、吞吐量、带宽、丢包率、可靠性等。Web 服务的 QoS 同时依赖于用户和服务双方，且由于动态的网络条件，同一个 Web 服务被不同用户调用的 QoS 可能具有差异性。由于很多服务是收费的，通过调用及监测的方式获取所有用户−服务对的 QoS 是不现实的，因此准确和个性化的 QoS 预测是一种可行的解决方案。

协同过滤（CF, collaborative filtering）技术目前已广泛应用于 Web 服务的 QoS 预测，大致可分为基于邻域的 CF 方法、基于模型的 CF 方法和混合的 CF 方法。基于邻域的 CF 方法的核心思想是依据历史 QoS 数据计算用户或服务之间的相似度并生成相似邻居集，然后依据相似邻居的已有 QoS 估算目标服务的 QoS。基于邻域的 CF 方法较简单，并且一定程度上利用了难以量化的潜在的用户特征或服务特征，但其预测性能受到数据稀疏性问题的影响较大，同时很难利用与目标节点相似度较低的节点所隐含的全局结构信息。基于模型的 CF 方法的核心思想是预定义一个具有适当结构和参数的模型并使用已有的 QoS 数据进行训练，训练后的模型具有较好的 QoS 预测能力，且对整体结构有较好的估计。基于模型的 CF 方法性能较高，在面对数据稀疏性问题时稳健性较强，但传统的基于模型的 CF 方法如矩阵分解方法，难以学习用户和服务的深层特征和隐藏信息，可扩展性有限。

近年来，深度学习技术发展迅猛，并且在 Web 服务的 QoS 预测任务上也得到了一些应用[3]。其中，图卷积神经网络（GCN, graph convolutional neural network）可以通过聚合相邻节点的信息获得目标节点的信息，能缓解数据稀疏性问题；同时，它可以通过神经网络的逐层融合获取图的结构信息和深层特征，能有效解决基于邻域的 CF 方法和基于模型的 CF 方法面临的问题，因而是目前基于深度学习的 QoS 预测方法中性能较好的方法。

然而，现有的基于 GCN 的 CF 方法只考虑用户与服务交互的显式信息，未考虑用户终端的环境特征信息和服务器的环境特征信息。环境特征是指客户端主机或服务器主机的特征，例如网络地址、子网、自治系统、地理位置等。这些因素可以通过不同的组合影响 QoS，可以使用“偏好”来代表用户客户端主机和服务器主机对对方环境的适应程度，对对方环境特征适应程度更高的用户−服务组合可以获得更好的 QoS，可以认为服务满足了用户的“偏好”，用户也满足了服务的“偏好”。因此，如果能够从用户−服务交互信息中挖掘出潜在的环境特征信息，就可以提供更全面和复杂的特征信息来提高 QoS 预测精度。多组件图卷积协同过滤[4]方法是最近提出的一种基于 GCN 的 CF 方法，它考虑了用户−项目的交互信息中潜在的用户对服务的偏好，并将抽象的偏好映射为具体的组件，具有挖掘用户或服务的潜在偏好的能力，因此本文前期工作采用该方法挖掘用户和服务的潜在偏好，并针对 QoS 预测任务提出一种新的方法[5]。然而，该工作依然存在以下 2 个可改进的地方。1) 现有的基于 GCN 的 CF 方法大多只应用于单类 QoS 属性。在真实环境中，多类 QoS 属性分别从不同的角度反映了用户特征或者服务特征，不同类别的 QoS 属性之间存在潜在的共同特征，这些共同特征是单类 QoS 属性的预测模型无法挖掘出来的。例如，一个拥有较优响应时间的用户−服务对可能保持了非常通畅的网络，也说明服务器此时可能负载较小，这些特征使该用户−服务对可能也有较优的吞吐量；同理，拥有较优吞吐量的用户−服务对也很可能具有较优的响应时间。如果能将多类 QoS 属性用合适的方法映射到同一个空间，就能以此建模多类 QoS 属性存在的共同特征及联系，提升 QoS 预测模型的准确度。2) 预测模块应用 DeepFM 对用户和服务的特征向量进行一阶特征、二阶和高阶特征的挖掘，但没有区分不同交互特征的重要性，也没有探究高阶交互特征对预测性能的影响。

基于以上的问题分析，本文提出一种多类 QoS 联合预测（JQSP, joint QoS prediction）方法。首先，引入一个包含多个卷积核的偏好提取模块来提取各单类 QoS 的用户−服务矩阵中隐含的用户偏好特征和服务偏好特征；然后，使用加权融合方法将多类 QoS 的特征提取向量映射到同一个向量空间；最后，使用引入自注意力的因子分解机挖掘融合嵌入向量中的一阶特征和各阶交互特征，并进行多类 QoS 的联合预测。本文的主要贡献总结如下。

1) 分析了提取多类 QoS 数据的共同特征对 QoS 预测精确度提升的有效性，应用偏好提取模块（MGCN 模块）实现了单类 QoS 的环境特征偏好提取，选择加权融合方法将多类 QoS 的提取向量映射到同一空间，实现了多类 QoS 特征融合的目标。

2) 引入带自注意力的因子分解机（ANFM，

attention neural factorization machine）对嵌入向量中的一阶特征、二阶交互特征和高阶交互特征进行深度融合，实现了特征深度融合，并为交互特征赋予注意力权重，提升特征提取的效果，最终实现 QoS 联合预测的目标。

3) 在真实数据集 WS-DREAM 上进行了大量的实验分析，实验结果表明了 JQSP 方法的有效性。

1　相关工作

协同过滤方法是应用最为广泛的 QoS 预测方法，大致可分为基于邻域的 CF 方法、基于模型的 CF 方法和混合的 CF 方法，本文主要介绍前两类方法的相关工作以及基于模型的 CF 方法中发展迅速的基于深度学习的 CF 方法的相关工作。

基于邻域的CF方法基本思想是借助相似用户或服务的历史 QoS 来预测目标服务的 QoS。Shao 等[6]首先提出一种基于用户的 CF 方法，该方法利用 Pearson 相关系数（PCC, Pearson correlation coefficient）计算用户−服务QoS矩阵中所有用户的相似性，然后对目标用户的前 k 个相似用户的历史 QoS 值进行融合，得到预测结果。其后的相关工作大多致力于改进相似性度量办法来增加衡量用户或服务相关性的准确度。例如，Chen 等[7]使用 A-余弦来计算服务之间的余弦相似性，然后减去服务的平均 QoS 向量，以此消除不同 QoS 的尺度影响，有利于相似性计算；任丽芳等[8]在移动边缘计算环境中通过 K-means 聚类确定相似用户和边缘服务器。此外，在 CF 方法中加入用户或服务的时间信息或位置信息也有利于 QoS 预测。例如，Wang 等[9]提出一种基于距离的增强型 Top-K 选择策略，在移动边缘计算任务中利用纬度和经度坐标选择相似边缘服务集；邓璇等[10]引入网络嵌入式学习，提出一种基于信誉感知的 QoS 预测方法，充分挖掘高阶隐式关系。基于邻域的 CF 方法易于实现、效果较好，但它们面临数据稀疏、冷启动等问题，可扩展性也较差。此外，基于邻域的 CF 方法主要利用历史 QoS 值和上下文信息，该特点很好地利用了局部信息，但可能忽略了全局结构。

基于模型的 CF 方法的基本思想是使用历史 QoS 值来训练预定义模型，使模型趋向于真实 QoS 值的分布。最经典的模型是矩阵分解（MF, matrix factorization）模型[11]，其主要思想是将用户−项矩阵分解为用户和项的 2 个潜在因素矩阵的乘积，这 2 个矩阵提取了部分用户或项的特征。大多数基于 MF 的模型采用梯度下降或随机梯度下降方法来寻找目标函数的局部最小值，Luo 等[12]将交替方向法的原理引入基于交替最小二乘法的训练过程中，加速了模型收敛；鲁城华等[13]提出一种基于用户和服务区域信息的 QoS 预测方法，将全局的服务质量信息和局部的区域信息相结合构建预测模型。在改进方法上，Chen 等[14]将用户 ID、服务 ID、服务位置和用户位置信息嵌入向量中，为 MF 模型引入了更多信息；Tang 等[15]通过合并服务用户的位置，改进了经典的因子分解机模型，提升了模型预测的精确度；夏会等[16]分析用户−服务 QoS 矩阵的时空特征，提出一种基于全局和局部结构相似度的稀疏矩阵分解模型；陈蕾等[17]通过将 Web 服务 QoS 预测问题建模为 L2,1 范数正则化矩阵补全问题，提出了一类基于结构化噪声矩阵补全的 Web 服务 QoS 预测方法，有效缓解了 QoS 信息受结构化噪声污染的问题。在引入时间信息时，Luo 等[18]提出了一种有偏的非负张量潜在因子分解模型，有效缓解了 QoS 数据随时间波动的问题。基于模型的 CF 方法使用用户−服务矩阵中的所有 QoS 值来构建全局模型，有效利用了全局信息，因而可以很好地估计整体结构，但传统的基于模型的 CF 方法在挖掘关联性较强的用户组或服务组的局部信息时表现较差，且难以提取高阶特征。

随着神经网络研究的深入，作为基于模型的 CF 方法的一个分支，基于神经网络的 CF 方法得到了较多研究。Kang 等[19]提出一种结合神经网络和注意力的因子分解机模型，能有效捕获非线性特征交互并赋予不同的重要性；Gao 等[20]提出一种能够对上下文信息进行聚类的模糊聚类算法和一种新的组合相似度计算方法，并提出一个新的神经协同过滤（NCF，neural collaborative filtering）模型，可以利用本地和全局特性为预测提供信息；王安迪[21]提出基于自组织映射神经网络与 K-means 两阶段聚类的 QoS 预测方法，对用户和服务分别进行聚类，将基于相似用户的预测值和基于相似服务的预测值结合进行混合预测；Chen 等[22]提出了一种由多个 LSTM 层组成的递归神经网络模型，并使用了多种正则化技术来提升预测性能。在 GCN 相关方法中，Elif 等[23]针对 Wi-Fi6 的 QoS 预测，采用 GCN 对数据进行时间分析，提升了预测的效果。本文的前期工作[5]将用户−服务的历史 QoS 值建模为二部图，采用 GCN 提取和聚合节点的邻居信息，并采用

ANFM 模块挖掘低、高阶交互特征并赋予权值，为预测提供了更多有价值的信息。基于神经网络的 CF 方法能有效提取历史 QoS 值中的高维信息，有能力拟合任何非线性 QoS 分布，但在提取特征的方法上还有较大的改进空间，如果能够从用户−服务交互信息中挖掘出潜在的环境特征信息，就可以提供更全面和复杂的特征信息来提高 QoS 预测精度。

2 方法介绍

本节将详细介绍 JQSP 方法。图 1 展示了 JQSP 方法的工作流程。JQSP 主要由多组件图卷积神经网络（MGCN, multi-component graph convolutional neural network）和 ANFM 这 2 个模块组成。JQSP 方法使用 MGCN 模块提取各单类 QoS 的多维偏好特征，并将多类 QoS 的特征嵌入映射到同一个空间，然后通过 ANFM 模块将拼接的多类 QoS 偏好特征进行深度融合，并实现多类 QoS 的联合预测。作为一个端到端模型，JQSP 以数据预处理后的多个用户−服务 QoS 矩阵作为输入，每个矩阵代表一类 QoS。在整个 JQSP 方法框架中，MGCN 模块包含 3 个子模块。1) 具有节点级注意力的分解器，该子模块可以从服务的特征信息和用户的特征信息中识别和捕获用户−服务交互关系的潜在偏好，并将其映射为具体的组件；2) 具有组件级注意力的组合器，该子模块可以获得上述组件的权重系数，然后通过聚合组件与对应的权重系数得到用户嵌入向量和服务嵌入向量；3) 加权融合器，该子模块负责将多类 QoS 的用户嵌入向量和服务嵌入向量采用加权融合的方式构成融合嵌入向量。ANFM 模块使用线性部分挖掘融合嵌入向量的一阶特征，使用交互部分挖掘其二阶交互特征，使用全连接层部分挖掘其高阶交互特征并应用自注意力层为交互特征分配权重，最后将多个部分的结果相加得到最终的多类 QoS 预测结果。JQSP 方法框架如图 2 所示，

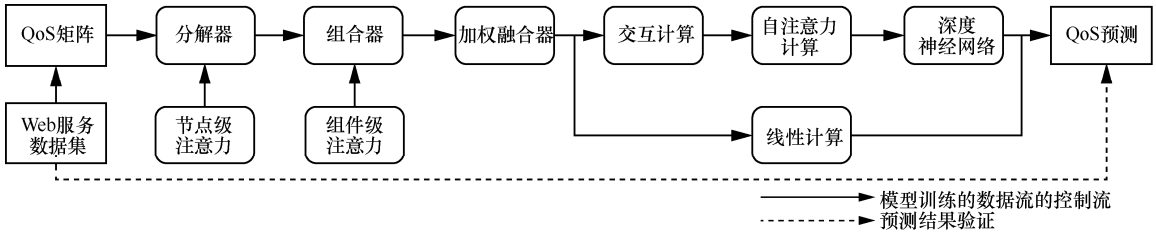


图 1 JQSP 方法的工作流程

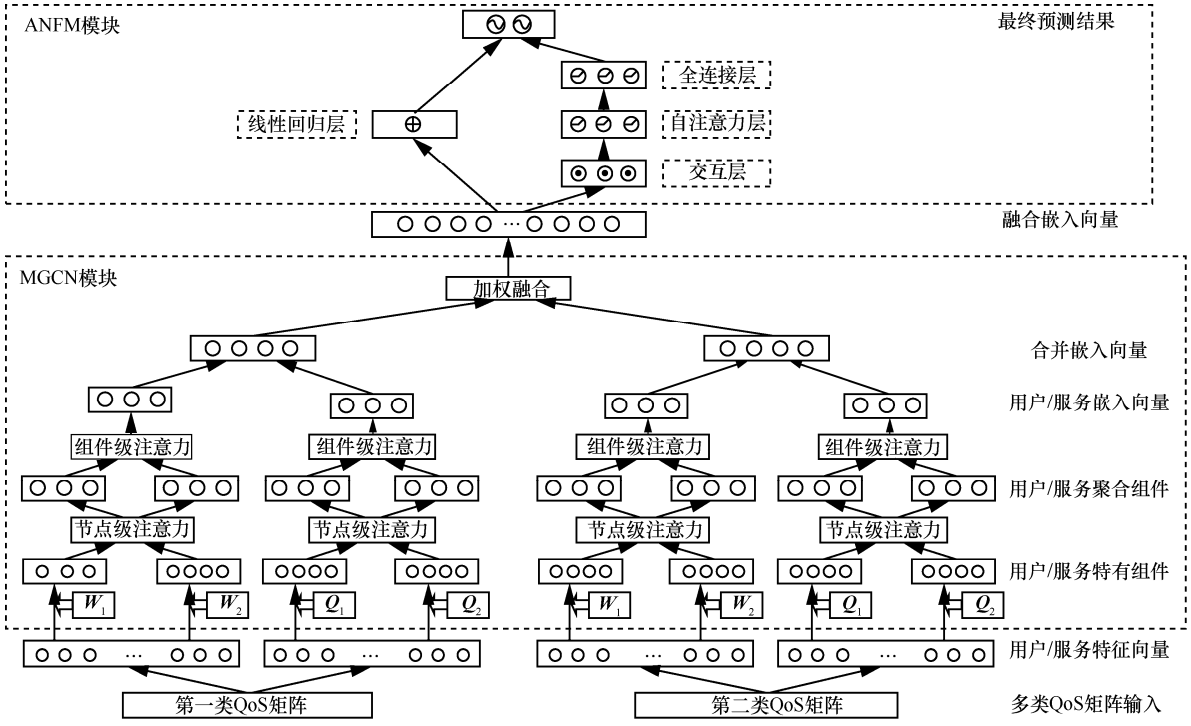


图 2 JQSP 方法框架

考虑两类 QoS 的联合预测情形，其中对每类 QoS 分别为用户和服务提取其偏好特征。

2.1 数据预处理

数据预处理的目的是将 QoS 矩阵转换为模型可以识别的标准格式。在 QoS 预测的背景下，可将用户−服务的历史 QoS 值建模为用户 GCN 服务二部图 $G=\{U,S,R,E\}$，其中 U 和 S 分别为用户 N_u 和服务 N_s 的集合，R 为 QoS 集合，E 为边集合，其元素边 $e=(u,s,r)\in E$ 表示存在用户 u 调用服务 s 的 QoS 值 r。

为便于处理，将 QoS 矩阵分别从用户和服务视角来建模，即用户特征矩阵 $\boldsymbol{U}=[u_1,u_2,\cdots,u_{N_u}]\in R^{L_u\times N_u}$ 和服务特征矩阵 $\boldsymbol{S}=[s_1,s_2,\cdots,s_{N_s}]\in R^{L_s\times N_s}$，其中，$L_u$ 和 L_s 分别为用户特征和服务特征的维度，N_u 和 N_s 分别为用户和服务的数量。

2.2 MGCN 模块

本文将编码器称为 MGCN 模块，它由多个结构相同、参数独立的 Single-MGCN 模块构成。Single-MGCN 模块在本文的前期工作[5]中已经得到实现，其思想主要借鉴文献[4]，可以细分为分解器和组合器 2 个部分，目的是将 QoS 矩阵表征为嵌入向量。Single-MGCN 模块的输入为单类 QoS 的用户特征矩阵和服务特征矩阵，输出为该类 QoS 的用户嵌入向量和服务嵌入向量。

2.2.1 分解器

本文构建分解器从特征信息中识别和捕获交互中的潜在偏好，并将其映射为组件。分解器的输入为用户特征矩阵和服务特征矩阵，输出为用户聚合组件和服务聚合组件。

1) 多组件提取。假定用户−服务二部图 G 受 M 个潜在偏好影响，本文分别为用户和服务设计 M 个独立的转换矩阵作为卷积核，对二部图进行卷积操作：用户转换矩阵组 $\{\boldsymbol{W}_i\}_{i=1}^{M}$ 和服务转换矩阵组 $\{\boldsymbol{Q}_j\}_{j=1}^{M}$，第 m 个转换矩阵捕获第 m 个用户−服务交互潜在偏好。对于服务 i，其第 m 个服务特有组件 h_i^m 可以按式(1)提取；对于用户 j，其第 m 个用户特有组件 p_j^m 可以按式(2)提取。这两组组件包含了 QoS 二部图中环境偏好特征对用户和服务的分量。

$$h_i^m=\boldsymbol{Q}_m s_j, i\in(1,\cdots,N_s) \tag{1}$$

$$p_j^m=\boldsymbol{W}_m u_j, j\in(1,\cdots,N_u) \tag{2}$$

2) 节点级注意力应用。在用户−服务二部图中，每个节点都有若干可用于获取邻域信息的邻居节点。对节点应用节点级注意力，可以学习该节点的各邻居节点的重要性，区分邻居节点之间的差别并为其分配不同的权重，以提升特征提取的效率。

经过多组件提取步骤，用户 j 得到 M 个用户特有组件 $\{p_j^m\}_{m=1}^{M}$，服务 i 得到 M 个服务特有组件 $\{h_i^m\}_{m=1}^{M}$。考虑到用户与不同服务的交互对描述各组件影响不同，分解器应用节点级注意力来凸显对描述组件影响较大的服务。

具体来说，考虑到第 m 个组件对用户 j 调用服务 i 的 QoS 值的影响同时表现在用户特有组件和服务特有组件中，则其影响因子 e_{ij}^m 可以由注意力计算式(3)学习得到，其中，$\mathrm{att}_{\mathrm{node}}$ 表示执行节点级注意力的神经网络，$\boldsymbol{W}_{\mathrm{att},m}$ 表示第 m 个组件的节点级注意力参数矩阵，σ 表示激活函数，$\|$ 表示拼接运算。在获得影响因子 e_{ij}^m 后，将其按 softmax 函数式(4)进行标准化，以获得其权重系数 α_{ij}^m，该权重系数表示该服务特有组件 h_i^m 的节点级注意力权重。

$$e_{ij}^m=\mathrm{att}_{\mathrm{node}}(p_j^m,h_i^m)=\sigma(\boldsymbol{W}_{\mathrm{att},m}^{\mathrm{T}}[p_j^m\,\|\,h_i^m]) \tag{3}$$

$$\alpha_{ij}^m=\mathrm{soft\,max}(e_{ij}^m)=\frac{\exp(e_{ij}^m)}{\sum_{s\in P_u}\exp(e_{ij}^m)} \tag{4}$$

对于用户 j 的调用服务集 P_u 中的所有服务，通过式(5)聚合其服务特有组件 h_i^m 与对应的权重系数 α_{ij}^m，可以得到用户 j 特有的第 m 个服务聚合组件 z_j^m，它描述了用户 j 的客户端主机受第 m 个偏好影响的程度。经过分解器的处理，所有用户都获得了 M 个服务聚合组件 $\{z_j^m\}_{m=1}^{M}$，所有服务也获得了 M 个用户聚合组件 $\{v_i^m\}_{m=1}^{M}$。

$$z_j^m=\sigma\left(\sum_{s\in P_u}\alpha_{ij}^m h_i^m\right) \tag{5}$$

2.2.2 组合器

本文构建组合器学习潜在组件的权重系数并聚合它们，得到嵌入向量。组合器的输入为用户聚合组件和服务聚合组件，输出为用户嵌入向量和服务嵌入向量。

1) 组件级注意力应用。客户端主机对服务环境的偏好可以通过服务聚合组件反映，而服务器对用户环境的偏好可以通过用户聚合组件反映。考虑到不同组件对学习用户嵌入向量或服务嵌入向量有

不同的贡献，组合器应用组件级注意力来凸显对学习嵌入向量影响较大的组件。

具体来说，考虑到第 m 个服务聚合组件的权重系数 β_j^m 同时受到原始的用户特征信息和节点级注意力加权的用户特征信息的影响，本文通过拼接服务聚合组件 z_j^m 和用户特有组件 p_j^m 并通过全连接层，按式(6)得到用户联合向量 $\boldsymbol{d}_j^m$，其中 $\boldsymbol{C}_m$ 是参数矩阵，$\boldsymbol{b}_m$ 是偏置向量。然后，按式(7)学习得到第 m 个服务聚合组件的影响因子 w_m，其中 $\mathrm{att}_{\mathrm{com}}$ 是执行组件级注意力的神经网络，$\boldsymbol{q}$ 是组件级注意力参数矩阵，b 是偏置值，$\boldsymbol{q}$ 和 b 由所有用户聚合组件和服务聚合组件共享，因为这 2 个参数表示客户端主机对不同服务环境和服务器对不同用户环境的共同偏好倾向。然后，将影响因子 w_m 按 soft max 函数式(8)进行标准化，获得第 m 个服务聚合组件的权重系数 β_j^m，该权重系数表示了该服务聚合组件 z_j^m 的组件级注意力权重。

$$\boldsymbol{d}_j^m = \sigma(\boldsymbol{C}_m \times [z_j^m \,\|\, p_j^m] + \boldsymbol{b}_m) \tag{6}$$

$$w_m = \mathrm{att}_{\mathrm{com}}(\boldsymbol{d}_j^m \in (d_j^1, d_j^2, ..., d_j^M)) = \sigma(\boldsymbol{q}^{\mathrm{T}} \times \boldsymbol{d}_j^m + b) \tag{7}$$

$$\beta_j^m = \frac{\exp(w_m)}{\sum_{k=1}^{M} \exp(w_k)} \tag{8}$$

2) 权重聚合。按式(9)聚合服务聚合组件与其对应的权重系数，获得用户 j 的嵌入向量 $\boldsymbol{z}_j$。类似地，可以获得服务 i 的嵌入向量 $\boldsymbol{v}_i$。用户嵌入向量 $\boldsymbol{z}_j$ 和服务嵌入向量 $\boldsymbol{v}_i$ 不仅捕获了低维的用户相似关系和服务相似关系，也捕获了高维的用户–服务交互中隐含的客户端主机对服务环境、服务器对用户环境的偏好信息。

$$\boldsymbol{z}_j = \sum_{m=1}^{M} \beta_j^m \times z_j^m \tag{9}$$

2.2.3 加权融合器

获得用户嵌入向量 $\boldsymbol{z}_j$ 和服务嵌入向量 $\boldsymbol{v}_i$ 后，将它们按式(10)进行拼接，得到合并嵌入向量 $\boldsymbol{e}_{\mathrm{merge}}^{ij}$，所有合并嵌入向量组合获得合并嵌入矩阵 $\boldsymbol{e}_{\mathrm{merge}}$。

$$\boldsymbol{e}_{\mathrm{merge}}^{ij} = [\boldsymbol{z}_j \,\|\, \boldsymbol{v}_i] \tag{10}$$

值得注意的是，合并嵌入矩阵 $\boldsymbol{e}_{\mathrm{merge}}$ 是 Single-MGCN 模块处理单个 QoS 矩阵所获得的单类 QoS 合并嵌入矩阵，仅包含单类 QoS 的用户特征和服务特征。对于多个 Single-MGCN 获得的多类 QoS 合并嵌入矩阵 $\boldsymbol{e}_{\mathrm{merge,k}}$，在对齐用户标识和服务标识后，本文采用加权融合的方式进行特征深度融合，按式(11)获得融合嵌入矩阵 $\boldsymbol{z}_{\mathrm{union}}$ 作为解码器的输入，其中，λ_i 是可训练的权重系数，且 $\sum \lambda_i = 1$。融合嵌入矩阵 $\boldsymbol{z}_{\mathrm{union}}$ 中的每个融合嵌入向量 $\boldsymbol{z}_{ij}$ 都包含了用户 j 和服务 i 在多类 QoS 上的特征。

$$\boldsymbol{z}_{\mathrm{union}} = \lambda_1 \boldsymbol{e}_{\mathrm{merge},1} + \lambda_2 \boldsymbol{e}_{\mathrm{merge},2} + ... + \lambda_k \boldsymbol{e}_{\mathrm{merge},k} \tag{11}$$

2.3 ANFM 模块

本文将解码器称为 ANFM 模块，它主要借鉴 ANFM 模型[24]。ANFM 模块的输入为融合嵌入向量，输出为多类 QoS 预测值。与传统因子分解机相似，ANFM 使用线性部分提取嵌入向量中的一阶特征；而在挖掘嵌入向量交互特征上，ANFM 进一步应用自注意力挖掘交互特征的注意力权重，然后使用深度神经网络挖掘高阶交互特征，这些工作使 ANFM 模块拥有比传统因子分解机 FM 更优的性能。

2.3.1 ANFM 介绍及线性部分计算

ANFM 的核心计算式如式(12)所示，其中，w_0 表示全局偏置，$\boldsymbol{W}_1$ 表示一阶特征提取的参数向量，$h(x)$ 表示可变的高阶特征提取函数。由此，对于输入的混合嵌入向量 $\boldsymbol{z}_{ij}$，首先可以获得线性部分 $y_{\mathrm{linear},ij} = w_0 + \boldsymbol{W}_1 \boldsymbol{z}_{ij}$。

$$y_{\mathrm{ANFM}}(X) = w_0 + \boldsymbol{W}_1 X + h(x) \tag{12}$$

传统 FM 的 $h(x)$ 为 $\sum\sum \boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{v}_j x_i x_j$，该项为二阶因式分解交互项，可以有效提取输入向量中的二阶交互特征，但在处理复杂的现实数据时表达受限[25]。考虑到各组交互对最终预测的贡献不同，本文应用注意力来凸显贡献更大的交互项。文献[24]证明在该任务中，自注意力机制[26]可以减少注意力特征提取对外部信息的依赖，有效捕捉特征的内部相关性，相比其他注意力机制有更优的表现，所以本文应用自注意力来强化高阶特征提取。图 3 展示了适用于本文模型 JQSP 的式(12)中的 ANFM 模块中高阶特征提取函数 $h(x)$ 的框架。

2.3.2 交互层计算

为提取特征之间的交互，对于给定的 d 维输入特征向量 $\boldsymbol{z}_{ij} = \{z_1, z_2, \cdots, z_d\}$，首先为其每个特征元素 z_i 构建交互嵌入向量 $\boldsymbol{e}_i$，按式(13)得到元素嵌入向量 $\boldsymbol{z}_i^e$。所有交互嵌入向量构成交互嵌入矩阵 $\boldsymbol{E}$，

它由所有输入特征向量共享。然后，按式(14)得到二阶交互向量 $\boldsymbol{z}_{\text{pair}}$，其中∘表示哈达玛积。

$$\boldsymbol{z}_i^e = z_i \boldsymbol{e}_i \tag{13}$$

$$\boldsymbol{z}_{\text{pair}} = \sum_{i=1}^{d} \sum_{j=i+1}^{d} \boldsymbol{z}_i^e \circ \boldsymbol{z}_j^e \tag{14}$$

2.3.3 自注意力应用

自注意力机制是注意力机制的变体，它依据输入向量的内部元素相关性计算各元素的自注意力值，减少了对外部信息的依赖，与传统注意力机制相比更加灵活。对二阶交互向量的自注意力机制实现过程如下。对于 e 维二阶交互向量 $\boldsymbol{z}_{\text{pair}} = \{z_1, z_2, \cdots, z_e\}$，首先为每个交互特征元素 z_j 构建注意力嵌入向量 $\mathbf{att}_j$，按式(15)得到元素注意力嵌入向量 $\boldsymbol{z}_j^{\text{att}}$。然后构建 3 个由全部元素嵌入向量共享的自注意力参数矩阵 $\boldsymbol{W}^Q$、$\boldsymbol{W}^K$、$\boldsymbol{W}^V$，按式(16)~式(18)分别计算得到查询向量 $\boldsymbol{Q}_j$、键向量 $\boldsymbol{K}_j$ 和值向量 $\boldsymbol{V}_j$。接着将 $\boldsymbol{Q}_j$ 与 $\boldsymbol{K}_j$ 按式(19)相乘得到 $\boldsymbol{z}_j^{\text{att}}$ 的注意力分数值 score_j，其中⊙表示点乘。

$$\boldsymbol{z}_j^{\text{att}} = z_j \mathbf{att}_j \tag{15}$$

$$\boldsymbol{Q}_j = \boldsymbol{W}^Q \boldsymbol{z}_j^{\text{att}} \tag{16}$$

$$\boldsymbol{K}_j = \boldsymbol{W}^K \boldsymbol{z}_j^{\text{att}} \tag{17}$$

$$\boldsymbol{V}_j = \boldsymbol{W}^V \boldsymbol{z}_j^{\text{att}} \tag{18}$$

$$\text{score}_j = \boldsymbol{Q}_j \odot \boldsymbol{K}_j \tag{19}$$

对于多个交互特征元素 z_j，将它们的注意力分数值按式(20)进行 softmax 归一化，得到对应的权重值 weight_j，该权重值能判断 $\boldsymbol{Q}_j$ 和 $\boldsymbol{K}_j$ 的相似程度，也决定了 $\boldsymbol{V}_j$ 的重要程度。最后，将 weight_j 和 $\boldsymbol{V}_j$ 按式(21)进行加权求和，得到带自注意力的二阶交互特征向量 $\boldsymbol{z}_{ij,\text{att}}$。区别于显式注意力，应用自注意力的二阶交互特征向量可以调整向量内部的元素值，使向量表征更加灵活、准确。

$$\text{weight}_j = \text{soft}\max(\text{score}_j) \tag{20}$$

$$\boldsymbol{z}_{ij,\text{att}} = \sum \{\text{weight}_j \boldsymbol{V}_j\} \tag{21}$$

2.3.4 QoS 联合预测

在获得带自注意力的二阶交互特征向量后，为提取高阶交互特征，本文将 $\boldsymbol{z}_{ij,\text{att}}$ 传入一组全连接层，按式(22)~式(25)计算高阶特征提取向量 $\boldsymbol{h}(\boldsymbol{z}_{ij})$。

$$l_1 = \sigma(\boldsymbol{W}_1 \boldsymbol{z}_{ij,\text{att}} + b_1) \tag{22}$$

$$l_2 = \sigma(\boldsymbol{W}_2 l_1 + b_2) \tag{23}$$

$$l_n = \sigma(\boldsymbol{W}_n l_{n-1} + b_n) \tag{24}$$

$$\boldsymbol{h}(\boldsymbol{z}_{ij}) = \boldsymbol{q}^{\mathrm{T}} l_n \tag{25}$$

其中，$\boldsymbol{W}_i$ 和 b_i 表示第 i 层神经网络的权重矩阵和偏置值，σ 表示激活函数，$\boldsymbol{q}$ 表示预测层权重矩阵。

至此，已获得了适用于本文模型 JQSP 的式(12)中的高阶特征提取函数 $h(x)$。注意，式(25)中 $\boldsymbol{q}$ 表示的是一个权重矩阵而非列向量，其列数等于 MGCN 模块输入的 QoS 矩阵的个数，因此 $\boldsymbol{h}(\boldsymbol{z}_{ij})$ 的输出为一个多维向量，其维度等于 QoS 的类别数。

综合以上结果，按式(26)获得 A 个最终预测结果 $\{y_{ij}^{\text{pred},a}\}_{a=1}^{A}$，其中，$y_{ij}^{\text{pred},a}$ 表示用户 j 调用服务 i 在第 a 类 QoS 上的预测结果，$h_a(\boldsymbol{z}_{ij})$ 表示高阶特征提取向量 $\boldsymbol{h}(\boldsymbol{z}_{ij})$ 的第 a 个分量，线性回归部分 $y_{\text{linear},ij}$ 由所有类的 QoS 预测结果共享。

$$y_{ij}^{\text{pred},a} = y_{\text{linear},ij} + h_a(\boldsymbol{z}_{ij}) \tag{26}$$

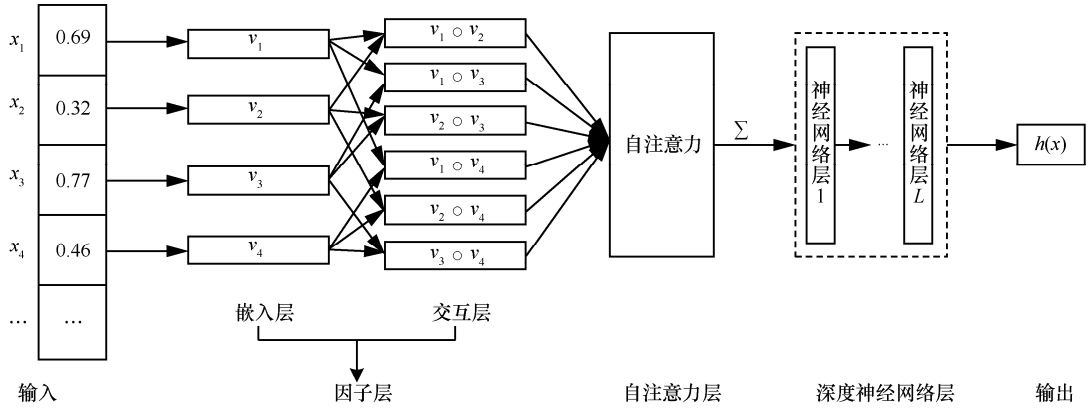


图 3　高阶特征提取函数 $h(x)$框架

2.4 优化

本文方法采用均方误差作为损失函数，具体计算如式(27)所示，其中，O 是已知 QoS 的集合，$|O|$ 是 O 的元素数量，$y_{ij}^{\text{pred},m}$ 是用户 j 对服务 i 的第 m 类预测 QoS 值，$y_{ij}^{\text{true},m}$ 是用户 j 对服务 i 的第 m 类真实 QoS 值，A 是 QoS 的类别数量。此处假定了如果一个用户−服务对获得一类真实 QoS 值，那么就能获得所有类的真实 QoS 值，因此获得的真实 QoS 值的数量为 $A|O|$。

$$\text{loss}_{\text{MSE}}=\frac{1}{A|O|}\sum_{a=1}^{A}\sum_{(i,j)\in O}\left(y_{ij}^{\text{pred},a}-y_{ij}^{\text{true},a}\right)^2 \tag{27}$$

为缓解过度参数化和过拟合问题，本文对损失函数进行 L_0 正则化，对多组件提取矩阵 $\boldsymbol{W}$ 和 $\boldsymbol{Q}$ 进行稀疏化，过滤无关自由度。最终的目标函数为式(28)，其中 $\theta=\{\boldsymbol{W},\boldsymbol{Q}\}$，$\lambda$ 表示用于平衡损失和系数正则化的超参数。

$$\text{loss}=\text{loss}_{\text{MSE}}+\lambda\|\theta\|_0 \tag{28}$$

3 实验评估及分析

本节进行若干对比实验和 JQSP 的消融实验，以期回答以下问题。

问题 1 JQSP 方法是否比其他基线方法表现更优？

问题 2 JQSP 方法模型的各子模块是否产生了预期的作用？具体来说，用于提取环境偏好特征的 MGCN 模块是否对 QoS 预测的准确度有优化作用？引入自注意力的 ANFM 模块是否能更有效地利用特征交互信息来提升 QoS 预测准确度？将多类 QoS 数据的特征映射到同一空间是否能更有效地挖掘隐含特征信息？联合预测是否能提升预测性能？

问题 3 各类超参数是如何影响模型性能？

本文使用如下配置的计算机进行实验。CPU 为 Intel（R）Xeon（R）Silver 4210 CPU @ 2.20 GHz，内存为 64 GB，GPU 为 2 块 GeForce RTX 2080ti。

3.1 数据集描述及处理

为评估模型性能，本文使用公开数据集 WS-DREAM 数据集。该数据集包含了 339 个用户与 5 825 个 Web 服务交互的 1 974 675 个真实 QoS 结果，包括响应时间和吞吐量两类重要 QoS。

为去除无效数据，本文对响应时间数据集进行如下预处理。首先，舍弃响应时间为 0（代表用户未调用该 Web 服务）和响应时间超过 20 s（代表响应时间过长，用户可能放弃调用该服务，所以该响应时间数据没有意义）的元素；然后，为保证联合预测时多类 QoS 的数据在同一个尺度，将响应时间数据集进行 min-max 归一化，使其数据尺度为（0,1）。吞吐量数据集中的元素均为有效数据，故仅进行 min-max 归一化。

现实中，用户通常只会调用少量的服务，从而导致 QoS 数据的用户服务矩阵稀疏。考虑到预处理后的吞吐量矩阵和响应时间矩阵为稠密矩阵，为了在实验中模拟真实情况，本文在训练模型时使用训练集密度（DoT，density of training set）较低的 QoS 矩阵。例如，DoT=5%表示随机选择 5%的 QoS 作为训练集，剩余 95%的 QoS 作为测试集。

3.2 评估标准

为了评价模型的效果，本文采用以下 2 种广泛使用的评价参数：均方根误差（RMSE, root mean square error）和平均绝对误差（MAE, mean absolute error）。

1) 均方根误差。RMSE 表示预测值与真实值偏差的平方与观测次数比值的平方根，反映了样本的分散程度。RMSE 的计算方法为

$$\text{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(y_{\text{pred},i}-y_{\text{true},i})^2}{n}} \tag{29}$$

2) 平均绝对误差。MAE 表示预测值和观测值之间绝对误差的平均值，其所有差值的权重相等。MAE 的计算方法为

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}|y_{\text{pred},i}-y_{\text{true},i}| \tag{30}$$

其中，$y_{\text{pred},i}$ 表示第 i 个预测 QoS 值，$y_{\text{true},i}$ 表示第 i 个真实 QoS 值。RMSE 和 MAE 的值越小，表示模型预测的准确性越高，结果越好。

3.3 对比方法与消融实验

本文将 JQSP 方法与基于邻域、基于因子分解模型和基于神经网络的 CF 方法以及 JQSP 方法的消融实验进行比较，以证明 JQSP 方法的性能。

1) UIPCC[27]。UIPCC 结合了基于用户和基于项目的协作预测方法，采用 PCC 来度量节点之间的相似度，并使用相似用户和相似服务进行 QoS 预测。它属于基于邻域的 CF 方法。

2) PMF（positive matrix factorization）[28]。PMF

采用概率矩阵分解方法对用户−服务 QoS 矩阵进行因子分解来提取隐藏特征，在面对大型稀疏数据集时具有良好的预测效果。它属于基于因子分解模型的 CF 方法。

3) 深度神经模型（DNM，deep neural model）[29]。DNM 是一种基于上下文的 QoS 预测模型，具有较好的预测精度，在面对挖掘异构上下文特征的任务时具有较好的稳健性和可扩展性。它属于基于神经网络的 CF 方法，因其在所有对比方法中表现最优，本文选用该方法作为基准方法。

4) MLP-ANFM（multilayer perceptron-attention neural factorization machine）。MLP-ANFM 使用 MLP 替代 MGCN 模块作为编码器，和 ANFM 模块一起组成完整的端到端模型。该方法的实验结果可论证 MGCN 模块是否对 QoS 预测的准确度有影响。

5) MGCN-MLP（multi-component graph convolutional network multilayer perceptron）。MGCN-MLP 使用 MLP 替代 ANFM 模块作为解码器，和 MGCN 模型一起组成完整的端到端模型。该方法的实验结果可以论证 ANFM 模块是否能有效地利用特征交互信息来提升 QoS 预测准确度。

6) Single-MGCN。Single-MGCN 使用单个 MGCN 模块，同时去掉了加权融合层，使整个模型只训练和预测单类 QoS。该方法的实验结果可以论证联合预测是否比单独预测更精确。

3.4　参数设置

考虑数据稀疏性的影响，本文将 QoS 数据集按如下比例随机分成训练集和测试集：DoT={5%, 10%, 15%, 20%, 25%, 30%}，共 6 个实验组，注意随机拆分时对齐多类 QoS 的用户标识和服务标识。然后，本文对所有方法在所有 DoT 数据集上各进行 5 次实验并取平均值，以评价 QoS 预测性能并进行对比分析。

参考文献[4,30]，JQSP 方法及其消融实验的参数设置如表 1 所示。对于其他对比方法，本文分别按照其参考文献内的最佳参数进行设置。

表 1　　参数设置

参数	值
转换矩阵数量	3
优化器	Adam
神经网络嵌入维度	64
自注意力嵌入维度	64
Dropout	0.5
正则化系数	3.3×10^{-5}
激活函数	ReLU
神经网络层	2
学习率	0.0001

3.5　实验结果与分析

表 2 给出了基于响应时间数据集的所有 QoS 预测评价结果，表 3 给出了基于吞吐量数据集的所有 QoS 预测评价结果，并对每个 DoT 训练集的最优数据进行加粗表示。表 2 和表 3 中 Gains 计算方式如式(31)所示，代表了 JQSP 方法与基准方法 DNM 相比性能提升的程度。

$$\text{Gains}=\frac{\text{eval}_{\text{benchmark}}-\text{eval}_{\text{JQSP}}}{\text{eval}_{\text{benchmark}}} \tag{31}$$

3.5.1　不同模型预测性能比较（问题 1）

由表 2 和表 3 可以得出，对于 RMSE 和 MAE 两项评价指标，PMF 显著优于 UIPCC，在训练集占

表 2　　基于响应时间数据集的所有 QoS 预测评价结果

方法	DoT=5%		DoT=10%		DoT=15%		DoT=20%		DoT=25%		DoT=30%	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
UIPCC	0.197 1	0.086 1	0.184 9	0.079 8	0.177 4	0.073 2	0.166 1	0.068 9	0.158 1	0.066 5	0.152 8	0.058 6
PMF	0.154 8	0.069 7	0.152 4	0.066 7	0.147 5	0.064 8	0.142 6	0.061 9	0.134 1	0.058 5	0.129 3	0.052 1
DNM	0.142 8	0.061 9	0.139 6	0.060 5	0.136 7	0.059 3	0.130 8	0.056 9	0.129 5	0.052 2	0.123 5	0.051 2
MLP-ANFM	0.142 2	0.061 1	0.139 3	0.059 4	0.135 8	0.059 1	0.128 0	0.054 4	0.127 5	0.050 2	0.122 0	0.050 3
MGCN-MLP	0.141 4	0.060 4	0.138 7	0.058 5	0.134 6	0.057 8	0.128 5	0.053 8	0.124 7	0.049 9	0.120 8	0.488 0
Single-MGCN	0.140 8	0.059 9	0.138 8	0.057 7	0.134 4	0.057 2	0.128 2	0.056 5	0.123 0	0.0500	0.118 6	0.047 9
JQSP	**0.135 7**	**0.058 7**	**0.133 5**	**0.055 4**	**0.129 6**	**0.052 5**	**0.122 0**	**0.050 4**	**0.112 6**	**0.047 7**	**0.098 8**	**0.045 2**
Gains	4.97%	5.17%	4.37%	8.43%	5.19%	11.47%	6.73%	11.42%	13.05%	8.62%	20.00%	11.72%

表 3 基于吞吐量数据集的所有 QoS 预测评价结果

方法	DoT=5%		DoT=10%		DoT=15%		DoT=20%		DoT=25%		DoT=30%	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
UIPCC	0.216 2	0.117 8	0.200 5	0.101 6	0.198 4	0.096 9	0.190 3	0.088 3	0.179 2	0.084 7	0.164 1	0.076 4
PMF	0.180 2	0.086	0.176 4	0.075 6	0.168 5	0.069 2	0.160 2	0.061 6	0.158 5	0.059 4	0.154 3	0.058 8
DNM	0.184 4	0.082 2	0.179 8	0.075 8	0.170 9	0.064 3	0.162 9	0.060 5	0.155 7	0.054 7	0.151 0	0.052 9
MLP-ANFM	0.182 3	0.083 6	0.177 0	0.073 8	0.168 8	0.066 3	0.161 1	0.060 0	0.154 8	0.053 8	0.147 7	0.051 1
MGCN-MLP	0.180 3	0.080 5	0.177 2	0.071 6	0.165 8	0.064 2	0.160 3	0.059 9	0.150 2	0.053 8	0.142 8	0.051 5
Single-MGCN	0.179 8	0.080 4	0.175 9	0.072 8	0.162 5	0.061 7	0.158 9	0.059 2	0.148 8	0.052 9	0.137 9	0.051 6
JQSP	**0.172 5**	**0.078 8**	**0.165 9**	**0.070 1**	**0.152 4**	**0.060 2**	**0.146 5**	**0.057 5**	**0.133 3**	**0.051 0**	**0.107 3**	**0.050 2**
Gains	6.45%	4.14%	7.73%	7.52%	10.83%	6.38%	10.07%	4.96%	14.39%	6.76%	28.96%	5.10%

比低的情况下更加明显，这说明了矩阵分解方法在缓解数据稀疏性问题上比基于邻域的方法表现更优。DNM 一定程度上优于 PMF，在少数训练集占比上持平，这说明神经网络方法有比矩阵分解方法更优的建模能力。本文提出的 JQSP 方法及其消融实验（即 MLP-ANFM、MGCN-MLP、Single-MGCN）在 RMSE 和 MAE 上始终优于对比方法，并且在吞吐量数据集、DoT=30%的 RMSE 评价指标上相比基准方法 DNM 有 28.94%的提升率，这说明本文所提方法 JQSP 相比其他基线方法有更优的表现。

3.5.2 JQSP 方法与各消融实验预测性能比较（问题 2）

对比表 2 和表 3 中的 MLP-ANFM 方法和 JQSP 方法可以看到，JQSP 方法性能全面领先 MLP-ANFM 方法，这证明了用于提取环境偏好特征的 MGCN 模块实现了以环境偏好特征为主要目标的细粒度潜在特征的挖掘，对 QoS 预测准确度的提升有帮助。

对比表 2 和表 3 中的 MGCN-MLP 方法和 JQSP 方法可以看到，JQSP 方法依然有较优的表现，这证明 ANFM 模块因为有效提取了输入向量的二阶及高阶交互特征而提升了预测的准确性，实现了特征深度融合。同时注意到 MGCN-MLP 方法和 MLP-ANFM 方法的性能差距较小，说明 MGCN 和 ANFM 这 2 个模块对模型预测性能的影响是接近的，单独增加 2 个模块中的任一个对整个模型性能的提升相差不大。

对比表 2 和表 3 中的 Single-MGCN 方法和 JQSP 方法可以看到，JQSP 方法依然有着较大的领先优势，这证明了与提取单类 QoS 特征相比，将多类相关的 QoS 数据的特征映射到同一空间进行特征提取有更优的表现，QoS 联合预测实现了对多类 QoS 潜在的共同特征的挖掘。同时注意到 Single-MGCN 方法与 MGCN-MLP 和 MLP-ANFM 方法相比更接近 JQSP 方法的表现，这说明与 MGCN 和 ANFM 这 2 个模块相比，单独应用联合预测框架对模型的预测精确度的提升较少，这可能是因为相比于 MGCN 模块挖掘的 QoS 矩阵中的环境偏好特征信息和 ANFM 模块挖掘的二阶和高阶交互信息，Single-MGCN 所挖掘的多类 QoS 潜在的共同特征信息的信息量更少，重要性更低，对模型性能提升的帮助更有限。

综上所述，本文提出的 JQSP 方法中 3 个主要子模块均产生了预期的作用，单独添加任一子模块都能有效提升模型的预测性能。

3.5.3 超参数影响分析（问题 3）

本文针对以下超参数在 DoT=30%的数据集上进行单一变量实验，以探究它们各自对 JQSP 模型性能的影响：转换矩阵数 $m \in \{1,2,3,4,5\}$，神经网络嵌入维度 $d_{\text{neu}} \in \{8,16,32,64,128\}$，自注意力嵌入维度 $d_{\text{att}} \in \{8,16,32,64,128\}$。简便起见，本文仅列出响应时间数据集上的结果。

1) 转换矩阵数。转换矩阵数 m 代表模型捕获潜在环境偏好的数量，增加 m 可以提高模型的捕获能力，但过高的 m 可能超过了真实数据中的潜在环境偏好数，增加了模型复杂度的同时无法提升模型性能。从图 4 的实验结果可以得出，随着 m 的增加，模型性能有所提升，$m=3$ 时模型获得了最优的性能，后续进一步增加 m 无法获得明显性能提升，同时还会大幅增加模型训练的时间。

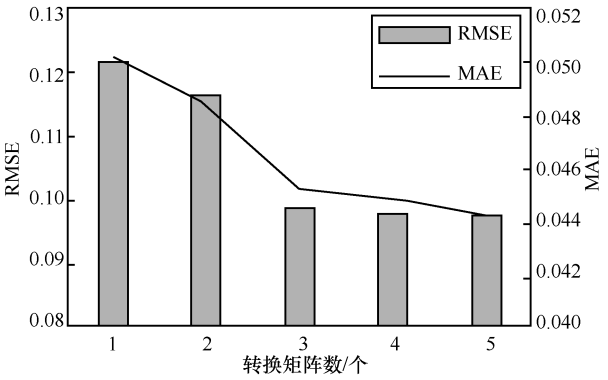


图 4　转换矩阵数对模型性能的影响

2) 神经网络嵌入维度。嵌入维度表示神经网络层使用多少维度来表达特征，维度越高，表达特征越细腻，但过高的维度会引入过多的参数，可能造成过拟合和难以收敛的问题，大幅增加模型训练时间。从图 5 的实验结果可以得出，随着 d_{neu} 增加，模型性能明显提高，$d_{\text{neu}}=64$ 时模型获得最优表达能力，继续增加 d_{neu} 反而导致性能下降。

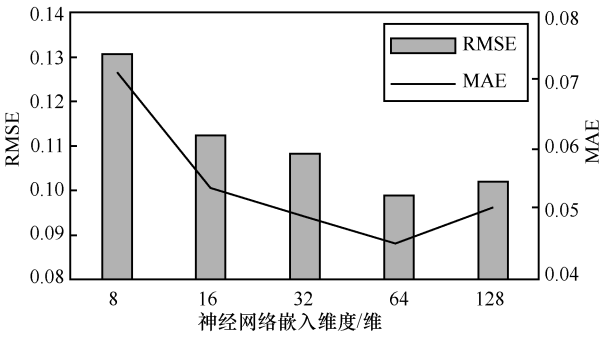


图 5　神经网络嵌入维度对模型性能的影响

3) 自注意力嵌入维度。自注意力嵌入维度表示自注意力参数矩阵使用多少维度来表达自注意力特征。从图 6 的实验结果可以得出，从 8 开始增加 d_{att} 有效提升了模型的性能，$d_{\text{att}}=32$ 时模型获得了最优的表达能力，继续增加 d_{att} 降低了模型的性能。

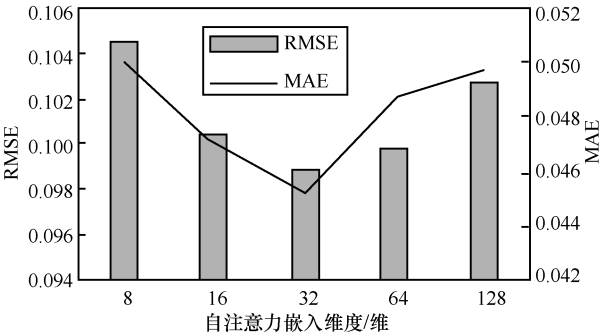


图 6　自注意力嵌入维度对模型性能的影响

4　结束语

本文提出了一种 JQSP 方法用于多类 QoS 联合预测，与现有的 QoS 预测方法相比，所提方法具有以下优点：1) 所提方法能有效识别和挖掘用户偏好信息和服务偏好信息，从而为特征提取提供了更丰富的信息；2) 所提方法将多类相关 QoS 的特征映射到同一个空间进行特征提取，这能获取到处理单类 QoS 无法获取的多类 QoS 相关性特征；3) 所提方法引入带自注意力的因子分解机来挖掘特征提取向量中的一阶特征、二阶和高阶交互特征，并为交互特征赋予注意力权重，有效提升了特征提取的效果，该效果优于传统因子分解机和 MLP。在未来的工作中，将考虑引入更丰富的异构信息来提升模型预测的精确度，同时考虑合理地简化模型结构，降低模型的训练时间，增加其可用性和稳健性。

参考文献：

[1] 方晨, 张恒巍, 张铭, 等. 基于信任扩展和列表级排序学习的服务推荐方法[J]. 通信学报, 2018, 39(1): 147-158.
FANG C, ZHANG H W, ZHANG M, et al. Trust expansion and list-wise learning-to-rank based service recommendation method[J]. Journal on Communications, 2018, 39(1): 147-158.

[2] 赵晨阳, 王俊岭. 基于隐含上下文支持向量机的服务推荐方法[J]. 通信学报, 2019, 40(9): 61-73.
ZHAO C Y, WANG J L. Service recommendation method based on context-embedded support vector machine[J]. Journal on Communications, 2019, 40(9): 61-73.

[3] YIN Y Y, CHEN L, XU Y S, et al. QoS prediction for service recommendation with deep feature learning in edge computing environment[J]. Mobile Networks and Applications, 2020, 25(2): 391-401.

[4] WANG X, WANG R J, SHI C, et al. Multi-component graph convolutional collaborative filtering[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 6267-6274.

[5] DING L, KANG G, LIU J, et al. QoS prediction for Web services via combining multi-component graph convolutional collaborative filtering and deep factorization machine[C]//IEEE International Conference on Web Services. Piscataway: IEEE Press, 202: 551-559.

[6] SHAO L S, ZHANG J, WEI Y, et al. Personalized QoS prediction forWeb services via collaborative filtering[C]//Proceedings of IEEE International Conference on Web Services. Piscataway: IEEE Press, 2007: 439-446.

[7] CHEN L, FENG Y P, WU J, et al. An enhanced QoS prediction approach for service selection[C]//Proceedings of 2011 IEEE International Conference on Services Computing. Piscataway: IEEE Press, 2011: 727-728.

[8] 任丽芳, 王文剑. 一种移动边缘计算环境中服务 QoS 的预测方法[J]. 小型微型计算机系统, 2020, 41(6): 1176-1181.
REN L F, WANG W J. Method for QoS prediction in mobile edge computing environment[J]. Journal of Chinese Computer Systems, 2020, 41(6): 1176-1181.

[9] WANG S G, ZHAO Y L, HUANG L, et al. QoS prediction for service recommendations in mobile edge computing[J]. Journal of Parallel and Distributed Computing, 2019, 127: 134-144.

[10] 邓璇, 吕晟凯. 基于信誉感知与嵌入式学习的 Web 服务 QoS 预测研究 [J]. 物联网技术, 2021, 11(12): 99-103.
DENG X, LYU S K, Research on Web service QoS prediction based on reputation perception and embedded learning [J]. Internet of Things Technologies, 2021, 11(12): 99-103.

[11] SALAKHUTDINOV R, MNIH A, HINTON G. Restricted Boltzmann machines for collaborative filtering[C]//Proceedings of the 24th International Conference on Machine Learning. New York: ACM Press, 2007: 791-798.

[12] LUO X, ZHOU M C, WANG Z D, et al. An effective scheme for QoS estimation via alternating direction method-based matrix factorization[J]. IEEE Transactions on Services Computing, 2019, 12(4): 503-518.

[13] 鲁城华, 寇纪淞. 基于用户和服务区域信息的个性化 Web 服务质量预测[J]. 管理科学, 2020, 33(2): 63-75.
LU C H, KOU J S. Personalized QoS prediction for Web services based on the region information of users and services[J]. Journal of Management Science, 2020, 33(2): 63-75.

[14] CHEN L, XIE F F, ZHENG Z B, et al. Predicting quality of service via leveraging location information[J]. Complexity, 2019, 2019: 4932030.

[15] TANG M D, LIANG W, YANG Y T, et al. A factorization machine-based QoS prediction approach for mobile service selection[J]. IEEE Access, 2019, 7: 32961-32970.

[16] 夏会, 高旻, 邹淑. 时空感知下基于结构相似度的 Web 服务质量预测[J]. 重庆大学学报, 2021, 44(1): 88-96.
XIA H, GAO M, ZOU S. A structure similarity based quality prediction approach for Web service in the spatial-temporal scenario[J]. Journal of Chongqing University, 2021, 44(1): 88-96.

[17] 陈蕾, 杨庚, 陈正宇, 等. 基于结构化噪声矩阵补全的 Web 服务 QoS 预测[J]. 通信学报, 2015, 36(6): 53-63.
CHEN L, YANG G, CHEN Z Y, et al. Web services QoS prediction via matrix completion with structural noise[J]. Journal on Communications, 2015, 36(6): 53-63.

[18] LUO X, WU H, YUAN H Q, et al. Temporal pattern-aware QoS prediction via biased non-negative latent factorization of tensors[J]. IEEE Transactions on Cybernetics, 2020, 50(5): 1798-1809.

[19] KANG G S, LIU J X, XIAO Y, et al. Neural and attentional factorization machine-based Web API recommendation for mashup development[J]. IEEE Transactions on Network and Service Management, 2021, 18(4): 4183-4196.

[20] GAO H H, XU Y S, YIN Y Y, et al. Context-aware QoS prediction with neural collaborative filtering for Internet-of-things services[J]. IEEE Internet of Things Journal, 2020, 7(5): 4532-4542.

[21] 王安迪. 基于 QoS 的 Web 服务质量预测方法研究[D]. 北京: 华北电力大学(北京), 2020.
WANG A D. Research on prediction method of web services quality based on QoS[D]. Beijing: North China Electric Power University, 2020.

[22] CHEN D, GAO M, LIU A, et al. A recurrent neural network based approach for Web service QoS prediction[C]//Proceedings of 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD). Piscataway: IEEE Press, 2019: 350-357.

[23] ELIF A, CANBERK B. Forecasting quality of service for next- generation data-driven WiFi6 campus networks[J]. IEEE Transactions on Network and Service Management, 2021, 18(4): 4744-4755.

[24] ZHANG X W, LI L. Attentional neural factorization machines for knowledge tracing[C]//Knowledge Science, Engineering and Management. Berlin: Springer, 2021: 319-330.

[25] RENDLE S. Factorization machines with libFM[J]. ACM Transactions on Intelligent Systems and Technology, 2012, 3(3): 1-22.

[26] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2017: 6000-6010.

[27] ZHENG Z, MA H, LYU M R, et al. QoS-aware Web service recommendation by collaborative filtering [J]. IEEE Transactions on services computing, 2011, 4(2): 140-52.

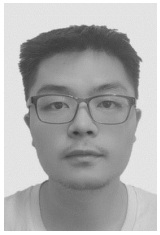
[28] ZHENG Z B, MA H, LYU M R, et al. Collaborative Web service QoS prediction via neighborhood integrated matrix factorization[J]. IEEE Transactions on Services Computing, 2013, 6(3): 289-299.

[29] WU H, ZHANG Z X, LUO J C, et al. Multiple attributes QoS prediction via deep neural model with contexts[J]. IEEE Transactions on Services Computing, 2021, 14(4): 1084-1096.

[30] LOUIZOS C, WELLING M, KINGMA D P. Learning sparse neural networks through L_0 regularization[J]. arXiv Preprint, arXiv: 171201312, 2017.

[作者简介]

刘建勋（1970− ），男，湖南衡阳人，博士，湖南科技大学教授，主要研究方向为工作流管理、服务计算、云计算、语义和知识网格等。



丁领航（1994− ），男，湖南湘潭人，湖南科技大学硕士生，主要研究方向为服务计算与云计算。

康国胜（1985− ），男，湖南郴州人，博士，湖南科技大学讲师，主要研究方向为服务计算和云计算、以数据为中心的业务流程管理、业务流程配置、数据挖掘和社交网络。

曹步清（1979− ），男，湖南湘潭人，博士，湖南科技大学教授，主要研究方向为服务计算、社交网络和软件工程。

肖勇（1995− ），男，湖南湘潭人，湖南科技大学博士生，主要研究方向为服务推荐、服务集群和网络表示学习。

收录声明

本刊对发表的文章，拥有出版电子版、网络版版权，并拥有和其他网站交换信息的权利。本刊支付的稿酬中已经包含上述费用。

Journal on Communications has the copyright to publish electronic edition, online edition of the published articles, and has the right to exchange information with other sites. The expenses have been included in the fee paid by editorial department.

道德声明

本刊发表的论文是作者独立取得的原创性研究成果，无一稿多投；论文内容不涉及国家机密；未曾以任何形式用任何文种在国内外公开发表过；论文内容不侵犯他人著作权和其他权利。若发生一稿多投、侵权、泄密等问题，论文作者将承担全部责任。

The authors of *Journal on Communications* guarantee that their submitted articles are original and contain nothing confidential. The said article is only submitted to *Journal on Communications*. The said article has not been published before and has not been submitted elsewhere for print or electronic publication consideration. The said article is no way whatever a violation or an infringement of any existing copyright or license from the third party. Otherwise, the authors of the said article shall take the blame for the violation or infringement of the related copyright and the leakage of secrets.

通信学报

Journal on Communications

发行代号：国内2-676 / 国外M395

2022年7月25日出版　定价：98.00元

ISSN 1000-436X

CN 11-2102/TN