计算机科学
COMPUTER SCIENCE

一种基于Meta-learning改进的特征交互算法

白静, 耿新宇, 易流, 穆禹锟, 陈琴, 宋杰

引用本文

白静, 耿新宇, 易流, 穆禹锟, 陈琴, 宋杰. 一种基于Meta-learning改进的特征交互算法[J]. 计算机科学, 2023, 50(11A): 230100087-8.

BAI Jing, GENG Xinyu, YI Liu, MU Yukun, CHEN Qin, SONG Jie. Improved Feature Interaction Algorithm Based on Meta-learning [J]. Computer Science, 2023, 50(11A): 230100087-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230100087

一种基于 Meta-learning 改进的特征交互算法

白 静 耿新宇 易 流 穆禹锟 陈 琴 宋 杰
西南石油大学计算机科学学院 成都 610000
(1255258033@qq.com)

摘 要 特征交互在推荐系统领域的广告点击率(Click-Through Rate,CTR)预测任务中至关重要,当前业界做的特征交互往往是基于内积、外积等矩阵变换,这些操作没有引入额外的信息,可以作为衡量两个向量相似性的手段,但作为特征交互的表示不一定是可靠的,许多特征交互无法有效提高点击率预测性能。首先从改善特征交互方式的角度入手引入额外的参数来学习一个映射,假设这个映射能够将两个向量的表征映射成交互的表征。学习映射的过程能够通过元学习(Meta-learning)来实现,故构建一个学习器以函数的方式表征特征交互。另外,不同的特征对不一定采取相同的方式交互,不能通过同一种交互方式得到所有特征对,因此设计一组元学习器(meta-learner)来学习映射函数,引入门控网络(GateNet)学习模型中元学习器的分布,那么不同的特征嵌入可以由一组元学习器得到表征。基于以上两点提出了一种融合多个元学习器并结合门控网络(Multiple meta-learners combined with GateNet,gate-MML)的特征交互算法,通过学习不同特征的联系和差异提高每个特征交互的质量。为了验证所提算法的性能,在 xDeepFM 模型上采用 gate-MML 做进一步的特征交互,采用 2 个真实广告点击率预测的数据集进行实验,并使用 Logloss 作为损失函数,AUC 作为评价指标。实验结果表明与传统的 CTR 预测模型相比,改进算法提升了广告点击率预测任务的预测性能。

关键词:特征交互;广告点击率预测;元学习;门控网络;推荐系统
中图法分类号 TP181

Improved Feature Interaction Algorithm Based on Meta-learning

BAI Jing,GENG Xinyu,YI Liu,MU Yukun,CHEN Qin and SONG Jie
School of Computer Science,Southwest Petroleum University,Chengdu 610000,China

Abstract Feature interaction is crucial in the field of advertising click-through rate(CTR) prediction in recommendation systems. However,current industry practices for feature interaction often rely on matrix transformations such as inner and outer products,which do not introduce additional information and can only serve as a means of measuring the similarity between two vectors. Therefore,such methods may not reliably represent feature interaction and may not effectively improve the performance of CTR prediction. To address this issue,this paper first introduces additional parameters to learn a mapping from the perspective of improving the feature interaction,assuming that this mapping can map the representation of two vectors to the representation of interaction. The process of learning mapping can be achieved through meta-learning,which constructs a learner to represent feature interactions in a functional manner. Additionally,different features may not adopt the same interaction method,and it is impossible to obtain all feature pairs through a single interaction method. Therefore,a set of meta-lear-ners is designed to learn the mapping function,and a GateNet is introduced to learn the distribution of meta-learners in the model,so that a set of meta-learners can represent different feature embeddings. Based on these two points,a feature interaction algorithm is proposed that combines multiple meta-learners with GateNet(gate-MML),which improves the quality of each feature interaction by learning the connections and differences between different features. To verify the performance of the proposed algorithm,gate-MML is used for further feature interaction in the xDeepFM model,and experiments are conducted on two real advertising click-through rate prediction datasets using Logloss as the loss function and AUC as the evaluation metric. Experimental results show that compared with traditional CTR prediction models,the improved algorithm enhances the prediction performance of advertising click-through rate prediction tasks.

Keywords Feature interaction,Advertising click-through rate prediction,Meta-learning,GateNet,Recommender system

1 引言

近年来,针对广告点击率预测问题最常采用的两种技术是基于特征交互建模和基于用户兴趣挖掘,本文主要讨论的是基于特征交互建模的方法。CTR 预估围绕着如何学到更有用的交叉特征诞生了一系列模型。针对 CTR 预测提出的

基金项目:四川省科技计划项目(2022NSFSC0555)
This work was supported by the Sichuan Science and Technology Program(2022NSFSC0555).
通信作者:耿新宇(gengxy123@126.com)

最早且使用最多的方法是逻辑回归(Logistic Regression, LR)[1],逻辑回归算法利用浅层交互模型融合多种特征进行推荐,它推动了CTR预估的早期研究。为了弥补线性模型不能进行特征交叉的缺陷,许多学者针对特征工程、特征交叉提出了一系列改进的CTR预估模型。由于LR和FTRL[2]等广义线性模型缺乏学习复杂特征交互的能力,Steffen等[3]提出因子分解机(Factorization Machine, FM)模型利用两个特征的Embedding做内积得到二阶特征交叉的权重。FM在2012到2014年前后已经成为业界主流的推荐模型之一,因为FM模型极大地降低了训练的开销,由POLY2[4] $O(n^2)$的复杂度变为$O(kn)$的线性复杂度,k表示隐向量的长度。隐向量的引入使得FM能更好地解决数据稀疏性的问题。但是FM模型通常只在两个特征之间做交叉,一旦超过两个,其复杂度会变得很高。为了得到更高阶的特征组合,Blondel等[5]将二阶的FM扩展到高阶提出HoFM模型并设计了ANOVA核(当高阶大于2时使用),保证了可解释性较强的情况下学习到特征的高阶组合信息。Poly2, FM, HoFM等模型都是在传统LR模型的基础上增加了对特征进行全交叉的自动学习权重部分,除此之外还有通过特征变换,利用GBDT产生高维非线性特征的GBDT+LR[6]组合模型。

为了解决浅层学习模型缺乏深层次高阶特征的问题,深度学习的强大学习能力同样也被运用于CTR预测任务中。许多基于深度神经网络的CTR预测模型被提出并取得了成功[7],其中很多都包含两个常用的组件:嵌入层和MLP隐藏层。随着微软的Deep Crossing模型[8]、谷歌的Wide&Deep模型[9]以及FNN[10]、PNN[11]等一大批优秀的深度学习模型被提出,推荐系统和计算广告领域全面进入了深度学习时代。有用的交互总是稀疏的,DNN很难在大量参数下有效地学习它们。在真实场景中,人工特征能够提高深度模型的性能,但特征工程成本高、需要领域知识,因此有必要对特征空间进行自动扩充。2019年,Liu等[12]提出了一种新的基于卷积神经网络的特征生成模型(FGCNN),通过预先捕获稀疏但重要的特征交互,生成的特征能够降低深度模型的学习难度。引入了Transformer的AutoInt[13]模型可以建模任意高阶特征的交互,并且这种交互是显性的。InterHAt[14]模型能够使用高效的注意力聚合策略提取高阶的特征组合,并且计算复杂度低。InterHAt引入了Transformer做多义的特征交叉,并且使用层次注意力机制提取重要特征组合,能给预测结果提供可解释性。

CTR预测任务的关键挑战是如何有效地对特征交互进行建模,现有的大多数CTR预测模型在特征交互框架中存在以下两方面的问题:(1)业界提出的从原始特征中学习低阶或高阶特征交互的模型往往是基于内积、外积等矩阵变换,单一的交互方式泛化能力弱,且对模型的性能提升也比较小;(2)许多特征交互算法框架广泛采用共用同种交互方法的结构[15],不同特征域间共用底部的隐层,这种结构本质上可以减少过拟合的风险,但是效果上可能会受到特征差异和数据分布的影响,不能获得很好的性能。也有一些模型对每个特征分别学习一套隐层然后学习所有隐层的组合,和共用同种交互结构相比,这些模型增加了针对不同特征的特定参数,在CTR预测任务中对最终效果有提升,但缺点是模型增加了参数量,因此需要更大的数据量来训练模型,不利于在真实生产环境中实际部署使用。

针对以上问题,我们从改善特征交互方式的角度入手引入额外的参数来学习一个映射,假设这个映射能够将两个向量的表征映射成交互的表征。本文设计了一组meta-learner以函数的方式表征特征交互,由于每对特征交互不一定以同样的方式表征,不能通过同一种交互方式得到所有特征对,因此,本文在此基础上采用GateNet从特征层中选择显著潜在信息,那么不同的特征嵌入可以由一组学习得到的meta-learner表征,以此来做CTR预测的可解释性与泛化性更强。

2 相关工作

2.1 Meta Learning的学习过程

Meta-learning也叫做学会学习[16],是机器学习领域一个前沿的研究框架,用于解决模型如何学习的问题。Meta-learning的目的是让模型获得一种学习能力,这种学习能力可以自动学到一些元知识,比如模型的超参数、神经网络的初始参数、神经网络的结构和优化器等[17]。因为在某个任务下由大量数据训练的模型,当切换到另一个任务后需要重新训练,这样非常耗时耗力。Meta-learning的原理如图1所示,它是在任务空间中进行训练,每当模型尝试学习某项任务时,无论成功与否,模型都会获得有用的经验,整合这些经验形成智能体的"价值观",代表一种会学习的能力,可抽象成函数$F(X)$[18]。若出现新的任务,在"价值观"的协助下,模型继续学习新任务的样本,可快速适应和掌握新任务,也就是抽象出对应新任务i的函数$f_i(x)(i=1,2,\cdots,t)$。Meta-Learning可以有效地缓解大量调参和任务切换模型重新训练带来的计算成本问题。

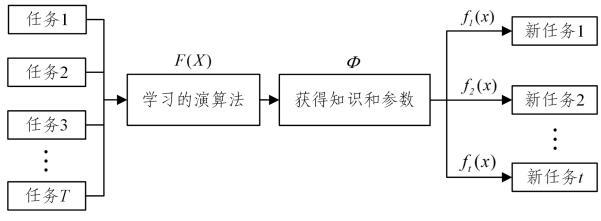


图1 Meta-learning原理框图

Fig. 1 Schematic diagram of meta-learning

Meta-learning分为两个阶段[19],两阶段学习框架如图2所示。第一阶段是训练任务阶段:给定T个子训练任务,每个子训练任务的数据集分为Support set和Query set。首先通过这T个子任务的Support set分别训练出针对各自子任务的模型参数,然后用不同子任务中的Query set分别去测试模型的性能,并由公式$L(\varphi)=l_1+l_2+\cdots+l_T$计算出预测值和真实标签的损失,最后利用梯度下降法更新参数,找到最优的超参设置。第二阶段是测试任务阶段:测试阶段是常规的机器学习过程,将数据集划分为训练集和测试集。利用阶段一中训练得到的超参设置可以对特定的测试任务进行训练。Meta-learning在历史的训练任务中积累了"价值观经验",以实现对模型的优化和调整。比如让Alphago学会下象棋,让一个汽车图片分类器具有对其他物品分类的能力。

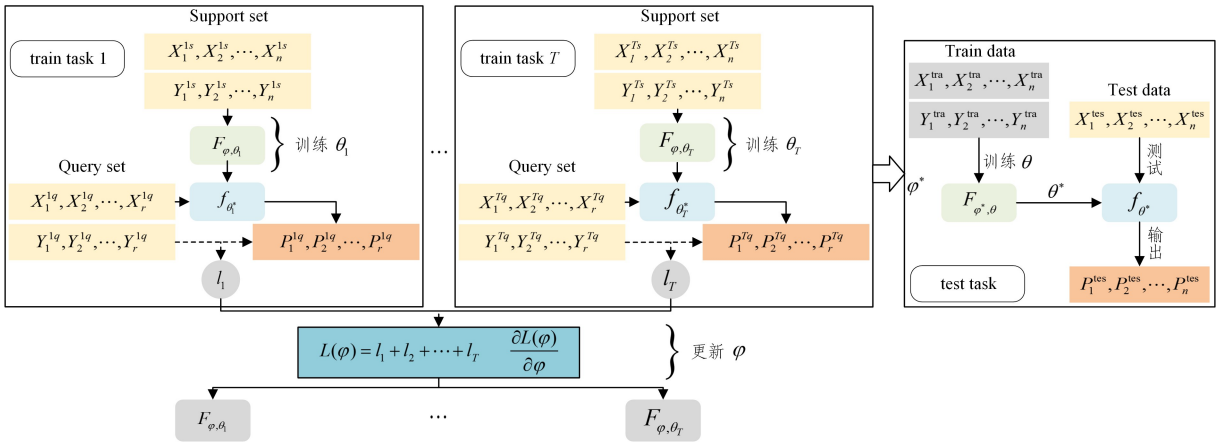


图 2　Meta-learning 的训练任务与测试任务示意图

Fig. 2　Schematic diagram of training and testing tasks of Meta-learning

2.2　门控网络 GateNet

门控机制（Gating mechanism）被广泛应用于计算机视觉[20]、自然语言处理[21]、推荐系统[22]等深度学习领域，用于控制网络中信息流通的路径，是进行特征候选的一个非常有效的手段。已有研究证明，门控机制改善了非凸深度神经网络的训练能力。在推荐系统中，门控机制相当于利用一个调节阀来控制信息流入或流出的程度，常用于 CTR 预测问题的排序模型，如学习各特征的重要程度、不同特征的分布等，进而提升特征交叉效率和模型效果。本文使用 GateNet 在深度神经网络 CTR 预测模型的嵌入层引入特征嵌入门，特征嵌入门提供了一个可学习的特征门控模块，用于从特征级别选择显著的潜在信息，在本文中用于计算 meta-learner 的分布。

3　算法描述

3.1　gate-MML 模型架构

如何对输入的原始特征进行深度交互，使得模型逼近原始特征决定的效果上限，提高点击率预测的精准性是特征交互建模需要重点解决的问题。特征交互建模实际上就是输入原始特征从底层抽取聚合，向网络深层传递，通过训练挖掘更深层次的信息表达。相比而言，模型的 Deep 部分主要聚焦于对交互信息再进行深度抽取。因此，交互建模一般发生在排序模型的底层，用于解决稀疏特征学习不充分、信息利用率低、特征挖掘不够等问题。

理论上，两层神经网络可以拟合任何一种变换（线性或非线性），故将两个嵌入向量作为输入得到一个输出，设计一个函数来描述两个特征之间的交互，相当于引入了额外的权重来学习一种映射。所提算法设计了特征交互元学习器并利用门控网络进行多个元学习器融合，通过学习不同特征的联系和差异提高每个特征交互的质量，并将其命名为融合多个元学习器并结合门控网络（gate-MML）的特征交互算法。gate-MML 算法基于共享表示来学习特定任务的函数，避免了明显增加参数的缺点。gate-MML 算法搭建在模型的输入层之后，模型的结构框架如图 3 所示。

特征嵌入门控层可以将嵌入特征转换为门感知嵌入特征，并有助于从特征层中提取显著的潜在信息。另外，每对特征交互不一定都是用同样的方式来表征，比如年龄和性别之间特征交互的方式、年龄和商品类别之间特征交互的方式可能是不同的。假设在某一个空间可以由多个 meta-learner 来学习不同的表征，比如一共有 k 个 meta-learner 学习不同向量之间的交互，对应了 k 个任务，每个任务学习一个特征交互元学习器，不同的特征交互都可以由不同的元学习器来组合得到。

根据输入特征是只选取一个 meta-learner 还是选取多个 meta-learner 两种情形，特征嵌入门控层的特征交互算法可以分为两种。（1）如果每对特征的 Embedding 向量只通过一种函数映射，也就是只有一个 meta-learner，那么将算法称为含有一个元学习器（One meta-learner，OML）的特征交互算法。因所有的特征交叉只有一种特征交叉方式不利于学习高质量的特征交互，但因需要实现不同的特征组合的特征交叉函数是不同的，如果每对特征都通过不同的特征映射，那么模型复杂度会非常高，所以选取多少个 meta-learner 是一个重要的超参数，本文后续将讨论部分超参数的选取。（2）如果每对特征的 Embedding 向量通过多种函数映射，此时模型能够选取 k 个 meta-learner，如图 3 所示，则将算法称为融合多个元学习器并结合门控网络（gate-MML）的特征交互算法。假设很多特征组合可以共享特征交叉方式，那么学习的则是一组基元，gate-MML 可以表示为式（1）：

$$y=\sum_{i=1}^{k} g(x)_i f_i(x) \tag{1}$$

其中，$g(x)_i$ 表示学习器 f_i 的权重且 $\sum_{i=1}^{k} g(x)_i=1$，$f_i(i=1,2,\cdots,k)$，包括 k 个学习器，g 代表整合了所有学习器结果的门控网络，在 k 个学习器上产生不同的权重。更具体地说，门控网络 g 基于输入产生关于 k 个学习器的分布，最终输出为所有学习器输出的加权和。特征嵌入门的基本步骤是为每个嵌入向量 $\boldsymbol{e}_m$ 计算表示嵌入特征重要性的门控值，将这一步骤形式化公式（2）：

$$g_i=\sigma(W_m \cdot \boldsymbol{e}_m) \tag{2}$$

其中，σ 是门控网络的激活函数，$\boldsymbol{e}_m \in R^D(m=1,2,\cdots,M)$ 是原始嵌入向量，D 表示嵌入向量的维度，W_m 是 GateNet 的学习参数。将门控值赋给相应的特征嵌入并生成门感知嵌入：

$$\boldsymbol{E}_{mi}=\boldsymbol{e}_m \odot \boldsymbol{g}_i \tag{3}$$

其中，$\odot$ 为 Hadamard 积，表示两个矩阵在相同位置上乘积，

$\boldsymbol{e}_m \in R^D (m=1,2,\cdots,M)$是第 m 个原始嵌入向量。一个元学习器的门感知嵌入可以表示为：$\boldsymbol{E}_i=[\boldsymbol{E}_{1i},\boldsymbol{E}_{2i},\cdots,\boldsymbol{E}_{Mi}]$，最后集合所有门感知嵌入并将其视为门特征嵌入：

$$\boldsymbol{X}^k_{\text{gate-MML}}=\sum_{i=1}^{k}\boldsymbol{E}_i \tag{4}$$

将门输出作为标量，代表整个特征嵌入的重要性。

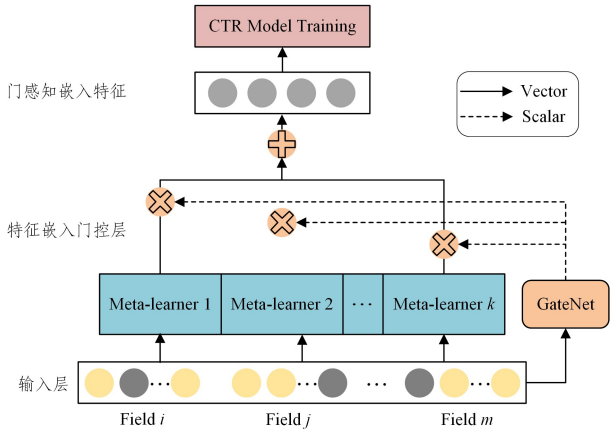


图 3 gate-MML 模型架构图

Fig. 3 Architecture diagram of gate-MML

3.2 gate-MML 在 xDeepFM 上的适配

xDeepFM 模型提出的 CIN 模块在向量级别(vector-wise)进行特征交互，使得该方法在推荐系统领域取得了不错的成绩，为了更好地验证所提算法的性能，本节将 gate-MML 插入到 xDeepFM 模型中，在网络进入 CIN 模块之前先经过特征嵌入门控层得到门感知嵌入特征，gate-MML 是对两个向量经过某种映射变换后达成两个向量的交互，通过引入外来参数来拟合特征交互，这种方式不用定义如何做特征交互，而是使用神经网络学习，基于 xDeepFM 模型的改进可以采用式(5)表示模型中的特征交互部分：

$$\boldsymbol{X}^l_{\text{gate-MML}}=\sum_{i=1}^{H_{l-1}}\sum_{j=1}^{d}\boldsymbol{W}^{l,h}_{ij}F(\boldsymbol{X}^{l-1}_{i,*},\boldsymbol{X}^0_{j,*}) \tag{5}$$

其中，$\boldsymbol{X}^l$ 是由 $\boldsymbol{X}^{l-1}$ 和 $\boldsymbol{X}^0$ 进行交互得到的，l 表示第 l 阶的交互；h 为第 h 个域并且 $1\leqslant h\leqslant H_l$；$\boldsymbol{W}^{l,h}\in\mathbb{R}^{H_{l-1}\times m}$ 是第 h 个特征向量的参数矩阵，d 表示每个 Embedding 的第 d 维；$F(\cdot)$ 则表示由神经网络学习得到的某种映射，反映了两个向量如何去做特征交互。将式(5)得到的特征映射具体化就是构建一组 meta-learner 以函数的方式表征特征交互。Meta-learner 的输入是两个特征的 Embedding 向量，输出是交互后的 Embedding 向量，meta-learner 可以由简单的多层感知机来表示，也可以采用更复杂的方式，相当于以一种可学习的状态来学习它的表征。每个 meta-learner 都能学习到一定的知识，每个任务(每对特征交互)都是由多个 meta-learner 组合求得。

加入 gate-MML 改进的 xDeepFM 模型最终的计算公式如下：

$$\hat{y}=\sigma(\boldsymbol{w}^{\mathrm{T}}_{\text{linear}}\boldsymbol{a}+\boldsymbol{w}^{\mathrm{T}}_{\text{dnn}}f_{\text{dnn}}(\boldsymbol{x}^k_{\text{gete-MML}})+\boldsymbol{w}^{\mathrm{T}}_{\text{cin}}f_{\text{cin}}(\boldsymbol{x}^k_{\text{gete-MML}})+b) \tag{6}$$

其中，σ 是 sigmoid 函数，$\boldsymbol{a}$ 是原始特征，$f_{\text{dnn}}(\boldsymbol{x}^k_{\text{gete-MML}})$ 和 $f_{\text{cin}}(\boldsymbol{x}^k_{\text{gete-MML}})$ 是 DNN 网络和 CIN 网络的输出。$\boldsymbol{W}_*$ 和 b 是可学习的参数。这里采用的损失函数是 logloss 函数：

$$\mathcal{L}=-\frac{1}{N}\sum_{i=1}^{N}y_i\log\hat{y}_i+(1-y_i)\log(1-\hat{y}_i) \tag{7}$$

其中，N 是样本总数，优化过程是如式(8)所示的目标函数：

$$\mathcal{J}=\mathcal{L}+\lambda_*\|\Theta\| \tag{8}$$

其中，λ_* 为正则化项，Θ 表示参数集合，包括线性部分、CIN、DNN 以及 gate-MML 部分的参数。每对特征交互如果只通过一个学习器，则称其为 OML-xDeepFM；如果通过 k 个学习器获得不同的 Embedding，则称其为 gate-MML-xDeepFM。

4 实验结果及分析

本节是实验验证部分，采用 pytorch 深度学习框架进行实验，具体实验环境如表 1 所列。

表 1 实验环境

Table 1 Experimental environment

实验环境	具体信息
CPU	AMD Ryzen 5600X
GPU	GeForce RTX 3060
Memory	32G
Operating System	Windows 10
Python	3.8
CUDA	11.3
PyTorch	1.12.0
Scikit-learn	0.24.1

4.1 实验数据

本文研究使用以下两个 CTR 预测数据集：

(1)Criteo 数据集：该数据集是法国互联网广告公司 Criteo 公布在 Kaggle上的展示广告点击率预测大赛的真实数据。该数据集包含 45840617 条用户点击反馈数据(时间跨度为一周)，大小约为 11 GB。每条记录包括特征和标签两部分，主要特征有用户特征(如性别、年龄、居住城市、浏览历史记录等)、广告自身的特征(如广告 ID、种类、广告的位置等)、场景特征(如用户访问时间、地点、浏览器类型、设备类型等)。标签为用户的点击情况，为 1 表示该条广告被点击，为 0 表示未被点击。数据集包括标签列 label、连续型特征 I1-I13、类别特征 C1-C26。

(2)Ali-CCP 数据集：该数据集是由阿里巴巴在阿里云天池提供的淘宝展示广告点击率预测的数据集，来自用户对淘宝展示广告的浏览和点击日志。特征包括用户域、商品域、组合域以及场景域，包含 38070670 条真实的淘宝用户点击数据(时间跨度为 8 天)。由 4 个数据表组成，分别是原始样本骨架、广告的基本信息、用户的基本信息和用户的行为日志。数据集包括标签列 click 以及 18 个类别型特征。

由于机器设备限制，本文使用随机抽样分别得到 500 万条 Criteo 子数据集和 500 万条 Ali-CCP 子数据集，并通过随机选择 80%的点击率数据作为训练集，其余 20%作为测试集，从训练集中选择 10%作为验证集。

4.2 评估指标

本文研究的是一个二分类问题，分类结果只有“点击”和“不点击”这两个取值。实验的目的是提高广告推荐系统中点击率预估的精确度，期望向特定的用户推荐他点击率高的广告，提高用户和广告的匹配程度，能够做到广告的精准推送。因此，本文选取了 AUC 值和 Logloss 作为评价模型好坏的指标。AUC 衡量的是一个正例的排名高于随机选择的负例的

概率，是 ROC 曲线下的面积，它只考虑预测实例的顺序，对类不平衡问题不敏感。AUC 的上界为 1，值越大越好。Logloss 衡量的是每个实例的预测分数和真实标签之间的距离。Logloss 的下限为 0，表示两个分布完全匹配，越小表示性能越好。

4.3　超参数的学习

本节主要对 gate-MML-xDeepFM 中的部分超参数进行研究，分别是 meta-learner 的个数 k、嵌入向量的维度 D、训练的学习率 η、Dropout 比率以及 MLP 网络的深度。

4.3.1　meta-learner 的个数

gate-MML 算法中最关键的超参数是 meta-learner 的个数 k，因为每两个特征交叉的表示不一定是一样的，所以设计一组 meta-learner 进行学习。实验设置 k 的取值为 0～10，对 gate-MML-xDeepFM 模型预测性能的影响如图 4 所示，可以看出，当 k 设置为 6 时，模型的预测性能最好，故本文将 k 设置为 6。

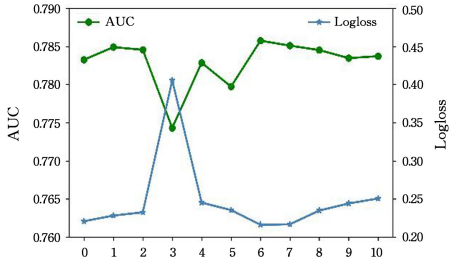


图 4　meta-learner 个数对模型预测性能的影响

Fig. 4　Effect of the number of k on prediction performance

4.3.2　嵌入向量的维度

Embedding 层中嵌入向量的维度对于整个模型的复杂度有较大的影响，特别是对模型的 CIN 网络和 DNN 网络的输入部分有较大影响。在确定 Embedding 层嵌入向量的维度时，CIN 网络层数确定为 3 层，保证模型能够较好地学习特征信息。此处设置的嵌入向量维度选择有{4，8，16，32，64}，实验结果如图 5 所示，可以看出当 D 为 8 时，gate-MML-xDeepFM 的预测性能最优。增加嵌入维度后，模型的性能会出现波动，甚至性能会低于浅层的 CTR 预测模型。这是因为嵌入向量维度大小代表特征信息是否被完全表达，如果维度过小，会导致特征信息无法充分表达；如果维度过大，会导致特征信息表达溢出，超出特征本身所包含的信息。所以 D 增加为 16 到 64 的 AUC 以及 Logloss 都没有明显提升，故本文后续实验选择的嵌入向量维度为 8。

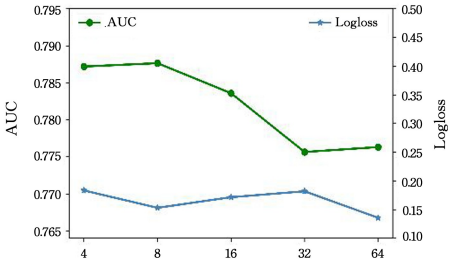


图 5　嵌入向量维度对模型预测性能的影响

Fig. 5　Effect of D on prediction performance

4.3.3　学习率

学习率 η 选择{0.001，0.005，0.01，0.05}，在一定轮数过后逐渐减缓，学习速率的衰减应该在 100 倍以上。实验结果如图 6 所示，可以看出当学习率 η 设置为 0.001 时 gate-MML-xDeepFM 模型的预测性能最优。学习率设置太大会造成网络不能收敛，在最优值附近徘徊，从而忽视找到最优值的位置；学习率设置太小，网络收敛会非常缓慢且可能会进入局部极值点，故本文将学习率 η 设置为 0.001。

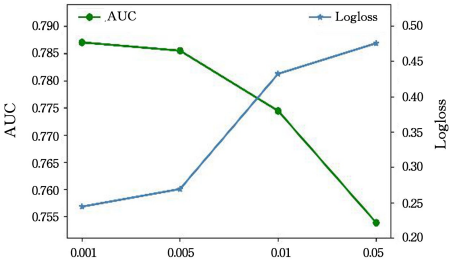


图 6　学习率对模型预测性能的影响

Fig. 6　Effect of learning rate on prediction performance

4.3.4　Dropout 比率

由于数据中的正负样本存在比例不均衡的情况，如果直接训练容易过拟合，故使用不同的 Dropout 比率对 CIN 网络和 gate-MML 网络中隐藏层进行处理可有效降低其过拟合程度，Dropout 比率的选择为 0 到 0.6，其实验结果如图 7 所示。

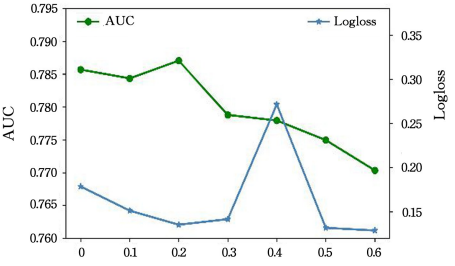


图 7　Dropout 比率对模型预测性能的影响

Fig. 7　Effect of dropout ratio on prediction performance

可以看出当 Dropout 比率为 0.2 时模型预测性能最好，故本文设置 Dropout 比率为 0.2，其含义是在训练时每层有 20%的节点会被随机丢弃，通过这样的方式可以较好地防止模型过拟合。

4.3.5　MLP 网络的深度

选取不同 MLP 层数对模型的预测性能有很大的影响，这里设置的深度为 1 到 6，图 8 展示的是 MLP 网络层数对 gate-MML-xDeepFM 模型的影响，随着 MLP 层数从 1 增加到 6，gate-MML-xDeepFM 模型的性能并没有呈现线性变化。本文认为这是由于神经网络本身的“黑盒”参数训练引起的，神经网络学习参数的不确定性会导致模型性能出现偏差。整体看来，当 MLP 层数为 2 时，gate-MML-xDeepFM 模型的 AUC 和 Logloss 都是最佳的，故本文将 MLP 层数设置为 2 层。

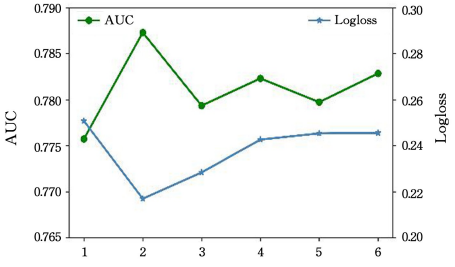


图 8　MLP 网络层数对模型预测性能的影响

Fig. 8　Effect of the number of MLP layers on prediction performance

4.4 消融实验

本节通过消融实验来验证 OML 和 gate-MML 算法是否有利于模型性能的提升。xDeepFM 模型整体分为 3 个子模块，分别是 Linear，DNN 和 CIN 部分，Linear 是为了引入人工设计的特征，DNN 是为了引入隐式特征组合。将 OML 和 gate-MML 算法引入 xDeepFM 中，能够带来 CIN 部分没有涉及的交互方式并产生不同的交互方式组合。因此，将原始 xDeepFM 模型稍作改变，采用 4.3 节选取的超参数进行如表 2 所列的 5 种消融实验。

表 2 5 种组合方式下模型的性能对比

Table 2 Performance comparison in five combination modes

组合方式	Criteo		Ali-CCP	
	AUC	Logloss	AUC	Logloss
L+D+CIN	0.7896	0.4581	0.6293	0.1503
L+D+OML	0.7907	0.4540	0.6177	0.1525
L+D+CIN+OML	0.7923	0.4543	0.6319	0.1509
L+D+gate-MML	0.7951	0.4531	0.6326	0.1498
L+D+CIN+gate-MML	**0.7971**	**0.4522**	**0.6358**	**0.1467**

表 2 中将 Linear 部分表示为 L，将 DNN 部分表示为 D，根据 5 种组合方式下模型在 Criteo 和 Ali-CCP 数据集上的预测结果可以得到如下结论：

(1)L+D+CIN 的组合实际上就是 xDeepFM 模型，而 L+D+OML 组合是将 xDeepFM 的 CIN 层替换为 OML 层，其预测性能比 xDeepFM 更优说明本文提出的采用函数映射表征特征交互的方式是有效的；另外，L+D+CIN+OML 是将 xDeepFM 与 OML 进行融合，由于特征交互更加充分，所以其预测性能也更优。

(2)L+D+gate-MML 组合是将 xDeepFM 的 CIN 层替换为 gate-MML 层，含有 6 个 meta-learner 但是去除 CIN 层的交互方式相比于只含有 1 个 meta-learner 而没有去除 CIN 层的交互方式更有效，反映了增加特定的参数确实提高了特征交互的质量。

(3)L+D+CIN+gate-MML 将所有模块都进行了融合，模型的预测准确率优于其他点击率预测模型。这是因为该模型充分融合了显式与隐式的特征交互。

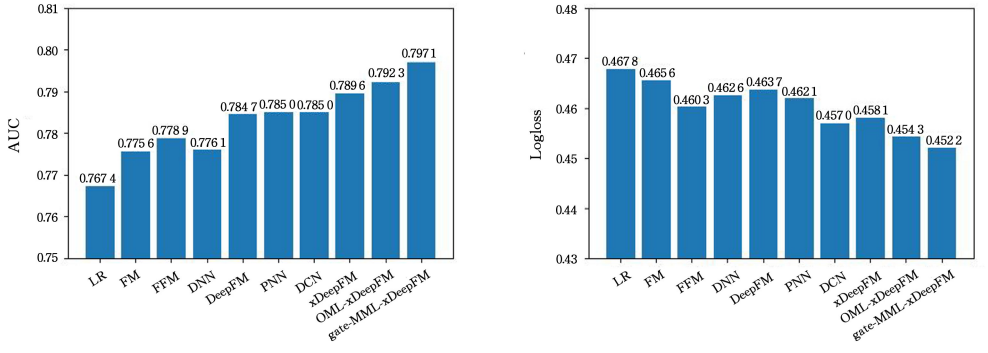
4.5 整体模型性能对比

本节进行对比实验采用的模型包括 LR[1]，FM[3]，FFM[23]，DNN，DeepFM[24]，PNN[11]，DCN[25] 和 xDeepFM[26]，使用 LR 作为性能比较基准模型。参数设置如下：meta-learner 的个数 k 为 6；在需要用到嵌入向量的模型(FM，FFM，DNN，DeepFM，PNN，DCN，xDeepFM，OML-xDeepFM 和 gate-MML-xDeepFM)中，Embedding 向量维度 D 设置为 8；学习率 η 为 0.001，学习率的衰减为 0.0001；Dropout 比率为 0.2；MLP 网络的深度为 2 层；采用 ReLU 作为激活函数，预测节点采用 Sigmoid 作为激活函数。使用 Criteo 测试数据集在多个 CTR 预测模型上的预测结果如表 3 和图 9 所示。

表 3 不同模型在 Criteo 数据集上的性能对比

Table 3 Performance comparison of different models on Criteo

模型	AUC	提升/%	Logloss	提升/%	模型	AUC	提升/%	Logloss	提升/%
LR	0.7674	0	0.4678	0	PNN	0.7850	2.293	0.4621	1.218
FM	0.7756	1.069	0.4656	0.470	DCN	0.7850	2.293	0.4570	2.309
FFM	0.7789	1.450	0.4603	1.603	xDeepFM	0.7896	2.893	0.4581	2.074
DNN	0.7761	1.134	0.4626	1.115	OML-xDeepFM	**0.7923**	**3.245**	**0.4543**	**2.886**
DeepFM	0.7847	2.254	0.4637	0.876	gate-MML-xDeepFM	**0.7971**	**3.870**	**0.4522**	**3.335**



(a)AUC 对比 (b)Logloss 对比

图 9 不同模型在 Criteo 数据集上的性能对比

Fig. 9 Performance comparison of different models on Criteo

表 3 列出了各个模型相比于基准模型 LR 的提升。由实验结果可以看出 LR 模型的 CTR 预测性能最差，因为 LR 是一个线性模型，模型结构比较简单，没有考虑特征交互。有特征交叉的模式得到的点击率预测结果往往比无特征交叉的点击率预测结果准确。FM 和 FFM 相较于 LR 模型的 AUC 值分别提升了 1.069%和 1.450%，FM 和 FFM 在 LR 模型的基础上加入二阶特征交互信息，模型的 CTR 预测性能也得到一定的提高，但是它们都没有挖掘高阶特征交互信息，所以相较于深度学习模型预测性能略差。基于深度学习的模型都比线性模型的预测性能要好，基于深度学习的模型在学习特征的高阶表示比低阶表示更有优势。DNN，DeepFM，PNN 和 DCN 都有学习低阶特征表示的能力，DCN 相较于其他深度学习模型有更高的性能，是因为 DCN 中的 Cross 网络通过显式的特征交互方式能更好地挖掘低阶特征组合信息。xDeepFM 设计了 CIN 模块可以学习显式高阶交互，并且特征交互的阶数往后每一层都会增加，以 vector-wise 方式代替普通 DNN 的元素级别(bit-wise)方式，可以看到 xDeepFM 相较于 LR 模型的 AUC 值和 Logloss 分别提升

了 2.893%和 2.074%。

本文改进的 OML-xDeepFM 模型虽然只含有 1 个 meta-learner，但也获得了不错的性能提升，这说明本文通过设计函数映射来代替乘积、内积、外积等向量之间的线性变换，实现非线性变换，能够在一定程度上提升 CTR 的预测性能。gate-MML-xDeepFM 模型预测性能在 Criteo 数据集上获得最优的性能提升，比 LR 模型在 AUC 和 Logloss 分别提高了 3.870%和 3.335%，说明 gate-MML 算法提取特征组合信息，实现了有效的特征交互，能够获取特征中更多有用的信息，对模型提升 CTR 预测性能有更好的帮助。

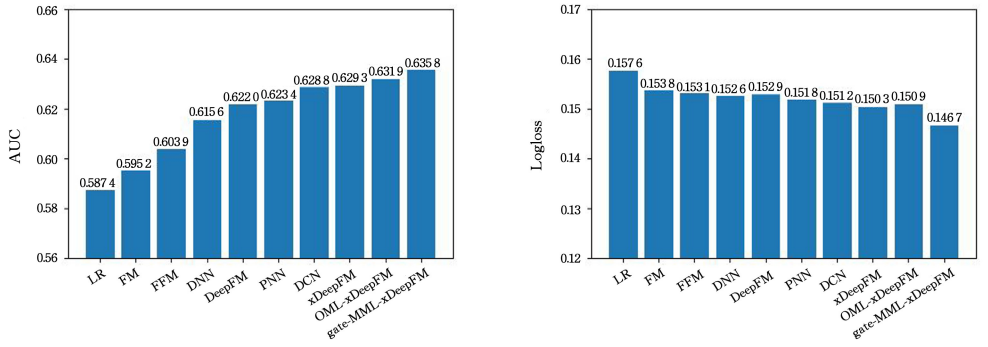
在提取高阶信息、实现精准预测上，特征的深度交叉起到了非常关键的作用。gate-MML 算法做到了两点创新：(1)通过设计学习器来代替乘积、内积、外积等向量之间的线性变换，实现非线性变换，因为理论上两层神经网络可以拟合任何函数，所以我们设计的 meta-learner 能包含所有的特征交叉方式，不论是内积、外积还是其他矩阵操作都可以通过这种形式表示，相当于提出了一种更泛化的特征交互；(2)为了克服不同特征组合都是由相同交互方式得到的不足，假设有 k 个 meta-learner，每个 Embedding 向量都可以由这组 meta-lear-ner 学习得到，既保证了参数量只有这 k 个 meta-learner，又能够使每个特征嵌入表示是不同的，嵌入向量的权重通过 GateNet 计算得到。

利用 Criteo 数据集对模型的超参数进行调整之后，使用 Ali-CCP 数据集验证模型的有效性，实验结果如表 4 和图 10 所示。同样地，表 4 也展示了各个模型相比于基准模型 LR 的提升。

表 4　不同模型在 Ali-CCP 数据集上的性能对比

Table 4　Performance comparison of different models on Ali-CCP

模型	AUC	提升/%	Logloss	提升/%	模型	AUC	提升/%	Logloss	提升/%
LR	0.5874	0	0.1576	0	PNN	0.6234	6.129	0.1518	3.680
FM	0.5952	1.328	0.1538	2.411	DCN	0.6288	7.048	0.1512	4.061
FFM	0.6039	2.809	0.1531	2.855	xDeepFM	0.6293	7.133	0.1503	4.632
DNN	0.6156	4.801	0.1526	3.173	OML-xDeepFM	**0.6319**	**7.576**	0.1509	4.251
DeepFM	0.6220	5.890	0.1529	2.982	gate-MML-xDeepFM	**0.6358**	**8.240**	**0.1467**	**6.916**



(a)AUC 对比　　(b)Logloss 对比

图 10　不同模型在 Ali-CCP 数据集上的性能对比

Fig. 10　Performance comparison of different models on Ali-CCP

在 Ali-CCP 数据集上进行验证可以发现，各种模型实验对比结果与使用 Criteo 数据集的结果类似。本文改进的 OML-xDeepFM 模型与 gate-MML-xDeepFM 模型获得了比其他模型更好的预测性能。OML-xDeepFM 模型比 LR 模型在 AUC 和 Logloss 分别提升了 7.576%和 4.251%，gate-MML-xDeepFM 模型比 LR 模型在 AUC 和 Logloss 分别提升了8.240%和 6.916%。所以在 gate-MML-xDeepFM 模型中我们既保证了不同的特征交互是不同的表示，同时也可以在一个更深的层次来寻找它的共性。

结束语　特征交互对于在推荐系统中实现高精度推荐至关重要。本文针对现有特征交互算法使用同一交互方式共享底层嵌入模式的不足，提出了 gate-MML 特征交互算法。值得一提的是，本文提出的特征交互方式并没有局限在某个模型里面，这是可插拔式的网络结构。因此，本文算法也可以用在其他 CTR 预测模型的嵌入层中，帮助模型更好地学习特征交互。如何丰富特征交叉的方式是未来工作的主要突破点，特征交叉的意义在于提高模型的非线性建模能力，提升模型的效果。然而实际场景往往比实验环境复杂得多，实际获取的待测数据具有多变性，会出现待测数据和训练集分布不一致的情况，对于在底层学习特征映射的元学习有更高的要求。元学习的目标是利用已学习的信息快速适应并掌握未学习的新任务，对未知场景有着较强的适应性和稳健性。针对如何将元学习运用于特征交互中，今后有必要探索更有效的优化方法来增强模型的可解释性并降低计算成本。

参 考 文 献

[1] RICHARDSON M, DOMINOWSKA E, RAGNO R. Predicting Clicks: Estimating the Click-through Rate for New Ads[C]// Proceedings of the 16th International Conference on World Wide Web. Banff, Alberta, Canada: Association for Computing Machinery, 2007: 521-530.

[2] MCMAHAN H B, HOLT G, SCULLEY D, et al. Ad Click Prediction: a View from the Trenches[C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013.

[3] RENDLE S. Factorization Machines[C]// 2010 IEEE International Conference on Data Mining. Sydney, NSW, Australia: IEEE, 2010: 995-1000.

[4] CHANG Y W, HSIEH C J, CHANG K W, et al. Training and

testing low-degree polynomial data mappings via linear SVM [J]. The Journal of Machine Learning Research, 2010, 8(11): 1471-1490.

[5] BLONDEL M, FUJINO A, UEDA N, et al. Higher-order FactorizationMachines[C]// Advances in Neural Information Processing Systems. 2016: 3351-3359.

[6] HE X, PAN J, JIN O, et al. Practical Lessons from Predicting Clicks on Ads at Facebook[C]// Proceedings of the Eighth International Workshop on Data Mining for Online Advertising. New York, NY, USA: Association for Computing Machinery, 2014: 1-9.

[7] ZHANG W, QIN J, GUO W, et al. Deep learning for click-through rate estimation[J]. arXiv: 2104. 10584, 2021.

[8] SHAN Y, HOENS T R, JIAO J, et al. Deep Crossing: Web-Scale Modeling without Manually Crafted Combinatorial Features [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, 2016: 255-262.

[9] CHENG H T, KOC L, HARMSEN J, et al. Wide & Deep Learning for Recommender Systems[C]// Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. New York, USA: Association for Computing Machinery, 2016: 7-10.

[10] ZHANG W, DU T, WANG J. Deep Learning over Multi-field Categorical Data [C] // Advances in Information Retrieval. Cham: Springer International Publishing, 2016: 45-57.

[11] QU Y, CAI H, REN K, et al. Product-Based Neural Networks for User Response Prediction [C] // 2016 IEEE 16th International Conference on Data Mining (ICDM). Barcelona, Spain: IEEE, 2016: 1149-1154.

[12] LIU B, TANG R, CHEN Y, et al. Feature Generation by Convolutional Neural Network for Click-through Rate Prediction [C]// The World Wide Web Conference. 2019: 1119-1129.

[13] SONG W, SHI C, XIAO Z, et al. Autoint: Automatic Feature Interaction Learning via Self-attentive Neural Networks[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019: 1161-1170.

[14] LI Z, CHENG W, CHEN Y, et al. Interpretable Click-through Rate Predictionthrough Hierarchical Attention [C] // Proceedings of the 13th International Conference on Web Search and Data Mining. 2020: 313-321.

[15] YANG Y, ZHAI P. Click-through rate prediction in online advertising: A literature review [J]. Information Processing & Management, 2022, 59(2): 102853.

[16] THRUN S, PRATT L. Learning to learn: Introduction and overview[M]// Springer, Boston, MA, 1998: 3-17.

[17] WANG J X. Meta-learning in natural and artificial intelligence [J]. Current Opinion in Behavioral Sciences, 2021, 38: 90-95.

[18] HOSPEDALES T, ANTONIOU A, MICAELLI P, et al. Meta-learning in neural networks: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(9): 5149-5169.

[19] FINN C, ABBEEL P, LEVINE S. Model-agnostic Meta-learning for Fast Adaptationof Deep Networks[C]// International Conference on Machine Learning. PMLR, 2017: 1126-1135.

[20] TAKIKAWA T, ACUNA D, JAMPANI V, et al. Gated-scnn: Gated Shape Cnns for Semantic Segmentation[C]// Proceedings of the IEEE/CVF International Conference on Computervision. 2019: 5229-5238.

[21] LI C, LI L, QI J. A Self-attentive Model with Gate Mechanism for Spoken Language Understanding[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 3824-3833.

[22] YANG L, ZHONG J, ZHANG Y, et al. An improving faster-RCNN with multi-attention ResNet for small target detection in intelligent autonomous transport with 6G[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 24(7): 7717-7725.

[23] JUAN Y, ZHUANG Y, CHIN W S, et al. Field-aware Factorization Machines for CTR Prediction[C]// Proceedings of the 10th ACM Conference on Recommender Systems. New York, NY, USA: Association for Computing Machinery, 2016: 43-50.

[24] GUO H, TANG R, YE Y, et al. DeepFM: A Factorization-Machine based NeuralNetwork for CTR Prediction[C]// Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. Melbourne, Australia: AAAI Press, 2017: 1725-1731.

[25] WANG R, FU B, FU G, et al. Deep & Cross Network for Ad Click Predictions[C]// Proceedings of the 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, 2017.

[26] LIAN J, ZHOU X, ZHANG F, et al. xdeepfm: Combining Explicit and Implicit Feature Interactions for Recommender Systems[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 1754-1763.

BAI Jing, born in 1996, postgraduate. Her main research interests include machine learning, data mining and recommender system.

GENG Xinyu, born in 1964, professor, master supervisor. His main research interests include data mining and artificial neural networks.