软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(3):822−844 [doi: 10.13328/j.cnki.jos.005681]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

因子分解机模型的宽度和深度扩展研究*

燕彩蓉[1], 周灵杰[1], 张青龙[2], 李晓林[3]

[1](东华大学 计算机科学与技术学院,上海 201620)

[2](计算机软件新技术国家重点实验室(南京大学),江苏 南京 210023)

[3](Department of Electrical and Computer Engineering, University of Florida, Florida 32611, USA)

通讯作者: 燕彩蓉, E-mail: cryan@dhu.edu.cn

摘 要: 因子分解机(factorization machine,简称 FM)模型因为能够有效解决高维数据特征组合的稀疏问题且具有较高的预测精度和计算效率,在广告点击率预测和推荐系统领域被广泛研究和应用.对 FM 及其相关模型的研究进展进行综述,有利于促进该模型的进一步改进和应用.通过比较 FM 模型与多项式回归模型和因子分解模型之间的关联关系,阐述 FM 模型的灵活性和普适性.从特征的高阶交互、特征的场交互、特征的分层交互以及基于特征工程的特征提取、合并、智能选择和提升等角度,总结模型在宽度扩展方面的方法、策略和关键技术.比较和分析了 FM 模型与其他模型的集成方式和特点,尤其是与深度学习模型的集成,为传统模型的深度扩展提供了思路.对 FM 模型的优化学习方法和基于不同并行与分布式计算框架的实现进行概括、比较和分析.最后,对 FM 模型中有待深入研究的难点、热点及发展趋势进行展望.

关键词: 因子分解机;推荐系统;广告点击率预测;特征工程;深度学习;并行与分布式计算

中图法分类号: TP181

中文引用格式: 燕彩蓉,周灵杰,张青龙,李晓林.因子分解机模型的宽度和深度扩展研究.软件学报,2019,30(3):822−844. http://www.jos.org.cn/1000-9825/5681.htm

英文引用格式: Yan CR, Zhou LJ, Zhang QL, Li XL. Research on wide and deep extension of factorization machine. Ruan Jian Xue Bao/Journal of Software, 2019,30(3):822−844 (in Chinese). http://www.jos.org.cn/1000-9825/5681.htm

Research on Wide and Deep Extension of Factorization Machine

YAN Cai-Rong[1], ZHOU Ling-Jie[1], ZHANG Qing-Long[2], LI Xiao-Lin[3]

[1](School of Computer Science and Technology, Donghua University, Shanghai 201620, China)

[2](State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing 210023, China)

[3](Department of Electrical and Computer Engineering, University of Florida, Florida 32611, USA)

Abstract: Since the factorization machine (FM) model can effectively solve the sparsity problem of high-dimensional data feature combination with high prediction accuracy and computational efficiency, it has been widely studied and applied in the field of click-through-rate (CTR) prediction and recommender systems. The review of the progress on the subsequent research on FM and its related models will help to promote the further improvement and application of the model. By comparing the relationship between the FM model and the polynomial regression model and the factorization model, the flexibility and generality of the FM model are described. Considering width extension, the strategies, methods, and key technologies are summarized from the dimensions of high-order feature interaction, field-aware feature interaction and hierarchical feature interaction, as well as feature extraction, combining, intelligent

* 基金项目: 国家自然科学基金(61402100); 中央高校基本科研业务费专项资金(2232016D3-11)

Foundation item: National Natural Science Foundation of China (61402100); Fundamental Research Funds for the Central Universities (2232016D3-11)

本文由智能数据管理与分析技术专刊特约编辑樊文飞教授、王国仁教授、王朝坤副教授推荐.

收稿时间: 2018-07-15; 修改时间: 2018-09-20; 采用时间: 2018-11-01

selection and promotion based on feature engineering. The integration approaches and benefits of FM model with other models, especially the combination with deep learning models are compared and analyzed, which provides insights into the in-depth expansion of traditional models. The learning and optimization methods of FM models and the implementation based on different parallel and distributed computing frameworks are summarized, compared, and analyzed. Finally, the authors forecast the difficult points, hot spots and development trends in the FM model that need to be further studied.

Key words: factorization machine; recommender system; CTR prediction; feature engineering; deep learning; parallel and distributed computing

云计算、移动互联网和社交媒体等技术的迅猛发展,使得网络空间中所蕴含的信息量呈指数级增长.数据量的快速增长造成了严重的信息过载,如何有效地过滤信息,找到最有价值的部分,成为企业乃至国家发展的重要战略目标.作为当前解决信息过载问题及实现个性化信息服务的最有效方法[1],预测/推荐被应用于很多领域,包括在线电子商务(如 Netflix、Amazon、eBay、阿里巴巴和豆瓣[2,3])、信息检索(如 iGoogle、MyYahoo、GroupLens 和百度[4,5])、移动应用(如 Google App[6,7])等.在计算广告领域,常用的过滤手段是点击率(click-through rate,简称 CTR)和转化率(conversion rate,简称 CVR)预测.准确估计 CTR 和 CVR 对提高流量的价值、增加广告收入有重要的指导作用.面对多源、异构、动态的大数据,预测/推荐模型算法在准确性、扩展性和实时性方面面临挑战.

传统的预测/推荐方法主要通过逻辑回归模型和多项式回归模型进行分类、回归或排序.作为一种有效的信息过滤手段,隐语义模型(latent factor model,简称 LFM)在智能信息系统和机器学习领域成为研究热点,并在预测/推荐方面性能表现优异.其中:研究最广泛的矩阵分解(matrix factorization,简称 MF)模型主要用于预测两个分类变量(categorical variable)之间的关系[8];张量分解(tensor factorization,简称 TF)模型是 MF 的扩展,可用于预测多分类变量之间的关系[9];SVD++在 MF 的基础上考虑了偏置因素(如物品流行度和用户个人偏好)对结果的影响,Time SVD++进一步添加了时间因子,使预测结果更加可靠[10,11].

因子分解机(factorization machine,简称 FM)模型由 Rendle 在 2010 年首次提出[12],它综合了矩阵分解和支持向量机(support vector machine,简称 SVM)模型的优势,利用因子分解对变量之间的交互进行建模,尤其适合于数据稀疏的场景.其输入是实数型特征,学习方法与线性回归模型和 SVM 模型类似,内部使用了变量之间的分解交互,在数据稀疏的情况(如 CTR 预测)下展现出非常高的预测质量.FM 模型被提出后,迅速成为学术界和工业界研究和应用的热点[13,14],尤其是近年来深度学习方法和应用的普及,进一步促进了 FM 模型的发展.

本文首先对 FM 模型在宽度扩展方面的相关方法和技术进行了总结,包括特征的低阶和高阶交互所适用的场景、特征与特征所属场的交互方法、特征及其分层关系之间的交互方法以及基于特征工程(feature engineering)进行特征提取、合并、智能选择和提升的相关策略与关键技术;然后,从集成学习角度比较了 FM 模型与其他模型的结合方式和特点,早期主要是与传统模型进行集成,深度学习提出后,FM 与深度学习模型的集成成为研究热点,并在 FM 模型的深度扩展方面展开了深入的研究;接着,从模型学习和实现角度比较和总结了模型的学习优化方法和分布式并行框架的实现策略.通过综述 FM 的最新研究进展,对 FM 未来的可能研究进行展望,一方面促进 FM 在信息系统领域应用的普及,另一方面也为机器学习领域的模型研究提供新的思路.所引用的参考文献来自数据挖掘、信息检索、机器学习等领域的重要期刊和国际顶级会议,包括 SIGMOD, SIGKDD,SIGIR,RecSys,VLDB,WWW,DASFAA,ICML,IJCAI,NIPS 等,文献发表时间为 2010 年 FM 模型提出到本文截稿日期.

本文第 1 节对预测/推荐问题以及独热编码方式和特征表示进行描述,便于后续模型、属性以及特征之间关系的理解.第 2 节根据国内外相关文献阐述 FM 模型的结构,总结 FM 模型在高阶交互、场交互、分层交互与传统模型的集成以及特征工程方面的扩展研究和应用.第 3 节综述 FM 模型与深度学习模型的集成方式及其优缺点分析.第 4 节从 FM 模型的不同学习优化方法和实现框架进行总结.第 5 节对 FM 模型的未来研究进行展望.第 6 节总结全文.

1 预测/推荐问题描述

预测和推荐是两类相关的任务,两者之间的关系体现为:预测的结果可为推荐提供服务,推荐需要以预测为基础.预测任务通常包括 3 种:分类、回归和排序[1].

预测是指通过对训练集$\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$进行学习,建立从输入空间 X 到输出空间 Y 的映射 $f:X\to Y$.若输出空间 Y 离散,那么这类学习任务被称为分类(classification),若分类中只涉及两个类别,即 Y 中只包含两个值,则称为二分类(binary classification),涉及多个类别则称为多分类(multi-class classification).若输出空间 Y 连续,那么这类任务称为回归(regression).对于二分类任务,$Y=\{-1,+1\}$或$\{0,1\}$;对于多分类任务,$Y=\{0,1,\ldots,n\}$,n 为类别数;对于回归任务,Y 通常为实数集$[k_1,k_2]$,其中,k_1 为回归下限,k_2 为回归上限,一般采用相关函数(如 clip)将回归控制在一定范围内.预测的一个典型应用就是 CTR 预测,其结果只有两个,点击或者不点击,可以将其转化为二分类问题.

推荐是指通过分析和挖掘用户(user)与物品(item)之间的二元关系及相关属性,帮助用户从海量数据中发现其感兴趣的有形或无形的物品(如信息、服务、商品等),生成个性化推荐列表[15,16].传统上推荐模型所使用的属性主要包括用户和物品的固有特征、显式的用户评分以及隐式的用户反馈.隐式反馈普遍存在于互联网中,如用户在观看电影、收听音乐、浏览或购买商品等活动时留下的行为痕迹.近几年来,利用社交网络和移动位置信息进行协同过滤推荐成为研究热点[17,18],但由于互联网中用户反馈多以隐式反馈的形式存在,且用户平均浏览商品数一般远小于商品总数,导致设计矩阵(用户物品矩阵)通常是稀疏的.稀疏的数据场景为预测/推荐的研究带来新的挑战.推荐系统和 CTR 预测都可以看做是分类问题在真实场景中的应用.

预测系统通常由 3 部分组成:特征提取、模型建立与训练以及在线服务,所有工作都构建在特征之上.考虑到推荐任务中特征并不总是连续的,多数情况是类别值,所以这些特征数字化将更适合于模型训练.为了应对一些无序的类别特征,可选用独热编码(one-hot encoding)方式对它们进行编码.独热编码又称一位有效编码,是广泛使用的一种特征表示方法.具体操作为:使用 N 位状态寄存器对 N 个状态进行编码,每个状态都有其独立的寄存器位,并且在任意时候,只有一位有效.独热编码解决了分类器不易处理属性数据的问题,在一定程度上扩充了特征.独热编码可将分类特征转化为数值型特征,从而方便进行模型预测.假设某电子商务网站中耳机的品牌包括 Panasonic,Sennheiser,Sony,Sound Intone,Bose,Beats,Audio-Technica 和 Ausdom 共 8 种,采用长度为 8 的向量对耳机品牌进行独热编码,特征 Panasonic 可表示为(1,0,0,0,0,0,0,0),特征 Sennheiser 可以表示为(0,1,0,0,0,0,0,0).图 1 所示描述的是一个推荐问题(第 1 组表示用户特征,第 2 组表示物品特征,第 3 组表示同一用户评分过的其他物品,第 4 组表示时间特征,第 5 组表示本用户最后评分的物品),特征用 x 表示,每个 x_i 表示一个特征向量,对应的目标值为 y_i,图中相关数据来源于 Rendle 等人的论文[12].

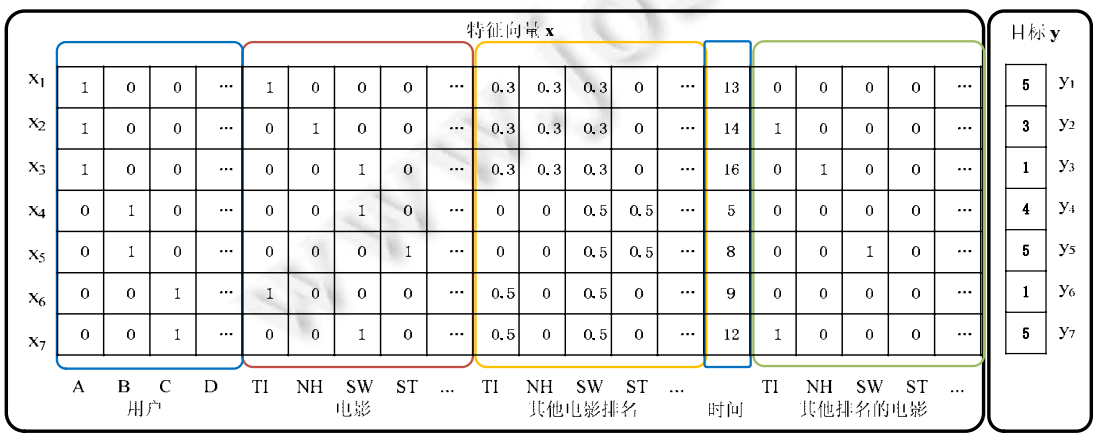

Fig.1 An example of feature vector

图 1 特征向量示例

2 FM 模型及其扩展

FM 模型源于多项式回归模型和因子分解模型,通过特征工程,FM 模型也可以转化为这些模型.本节首先介绍逻辑回归模型、多项式回归模型以及 Poly2 模型之间的关系,分析其特点和不足,阐述因子分解模型的优势,并详细说明因子分解的多项式回归模型在参数学习中的优势,引出本文研究的核心 FM 模型.然后,从特征的高阶交互、场交互、层次交互、与传统模型的集成学习以及特征工程角度讨论 FM 模型的扩展,并对 FM 模型的应用场景进行汇总.

2.1 回归模型与因子分解模型

逻辑回归(logistic regression,简称 LR)模型是应用最广泛的线性模型[19],其最常见的应用场景就是预测概率,即根据输入预测一个值.它可以作为一种分类方法,主要用于二分类问题,模型描述为

$$\hat{y}(x)=\sum_{i=1}^{n}w_ix_i \tag{1}$$

其中,x_i 是输入特征向量的第 i 个分量,向量长度为 n;w_i 是 x_i 对应的权值.

LR 模型的学习与优化通常使用最大似然和梯度下降方法来求解模型的参数.LR 模型的优势是简单、直观,模型求出的系数易于理解,便于解释,不属于黑盒模型.不足之处是模型的输入特征通常依赖于人工方式进行设计,而且 LR 本身是线性模型,无法对特征间非线性关系进行捕捉和自动建模,从而必须依赖于人工方式进行特征组合(feature combination)来实现 2 阶或高阶交互特征的构造.实践中,可利用因子分解机得到的特征交叉系数来选择输入 LR 模型的交叉特征(cross feature)组合,从而避免了繁杂的特征选择工作.

为了充分地利用特征之间的非线性关系,让模型能够学习到 2 阶或高阶交互特征,可以采用多项式回归模型.3 阶多项式回归模型描述为

$$\hat{y}(x)=w_0+\sum_{i=1}^{n}w_ix_i+\sum_{i=1}^{n}\sum_{j\geqslant i}^{n}\hat{w}_{i,j}x_ix_j+\sum_{i=1}^{n}\sum_{j\geqslant i}^{n}\sum_{l\geqslant j}^{n}\tilde{w}_{i,j,l}x_ix_jx_l \tag{2}$$

其中,$X\in\mathbb{R}^n,Y\in\mathbb{R},w_0\in\mathbb{R},W\in\mathbb{R}^n,\hat{W}\in\mathbb{R}^{n\times n},\tilde{W}\in\mathbb{R}^{n\times n\times n}$.

多项式回归模型的优势是能够学习特征之间的非线性交互关系,不足是采用这种方式构造的特征数量与特征值个数的乘积相关.假如某类特征有 1 万个可能的取值,另一类特征也有 1 万个可能的取值,那么理论上这两个特征组合就会产生 1 亿个可能的组合特征项.若加入 3 阶特征组合,则会引入更高的特征维度,导致特征爆炸问题,进而无法有效地学习高阶特征.Poly2 模型的提出充分考虑了多项式回归模型的可行性和实用性,此模型只对 2 阶特征组合进行建模[20].模型描述为

$$\hat{y}(x)=w_0+\sum_{i=1}^{n}w_ix_i+\sum_{i=1}^{n}\sum_{j=i}^{n}\hat{w}_{i,j}x_ix_j \tag{3}$$

其中,$w_0\in\mathbb{R},W\in\mathbb{R}^n,\hat{W}\in\mathbb{R}^{n\times n}$.Poly2 模型中只有 2 阶的交互特征,所以模型的预测和训练复杂度不会太高.但是由于实际应用场景下数据稀疏性问题,使得 Poly2 模型的二次项权重系数的训练学习变得非常困难,从而严重影响模型性能.另外,当每个特征取值较多时,仍然存在计算复杂度过高的问题.

因子分解的多项式回归(factorized polynomial regression)利用每个特征的隐向量进行组合,不需要为每个特征交互产生不同的权值.模型描述为

$$\hat{y}(x\mid w_0,v)=w_0x_0+\sum_{i=1}^{n}\sum_{f=1}^{k_1}v_{i,f}x_i+\sum_{i=1}^{n}\sum_{j\geqslant i}^{n}\sum_{f=1}^{k_2}v_{i,f}v_{j,f}x_ix_j+\sum_{i=1}^{n}\sum_{j\geqslant i}^{n}\sum_{l\geqslant j}^{n}\sum_{f=1}^{k_3}v_{i,f}v_{j,f}v_{l,f}x_ix_jx_l \tag{4}$$

其中,$w_i=\sum_{f=1}^{k_1}v_{i,f}x_i,\hat{w}_{i_1,i_2}=\sum_{f=1}^{k_2}\prod_{j=1}^{2}v_{i_j,f},\tilde{w}_{i_1,i_2,i_3}=\sum_{f=1}^{k_3}\prod_{j=1}^{3}v_{i_j,f}$,1 阶隐向量长度为 k_1,2 阶为 k_2,3 阶为 k_3.

该模型的优势是:可以根据隐向量长度调节参数的数量,通常,隐向量长度远小于特征总数,所以参数大幅减少,学习效率得到提高.但其不足之处在于:公式中存在特征的平方项(2 阶)和立方项(3 阶),不利于参数的学习.

2.2 FM模型及其高阶扩展

FM 旨在解决稀疏数据情况下,特征组合的参数学习不充分的问题.度为 2 的 FM 模型常被称为 basicFM,模型表示为[12]

$$\hat{y}(x)=w_0+\sum_{i=1}^{n}w_ix_i+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\langle v_i,v_j\rangle x_ix_j \tag{5}$$

其中,n 代表样本的特征数量,$w_0\in\mathbb{R}$,$W\in\mathbb{R}^n$,$V\in\mathbb{R}^{n\times k}$,$\langle\cdot,\cdot\rangle$表示大小为 k 的两个向量的点积(inner product),即$\langle v_i,v_j\rangle=\sum_{p=1}^{k}v_{i,p}\cdot v_{j,p}$.通过训练,为每个特征 $O(kn)$学习出唯一对应的隐向量 v_i,$\langle v_i,v_j\rangle$作为特征交叉项 x_ix_j 的权重参数.从模型公式中直观地看,FM 模型的复杂度为 $O(kn^2)$,但是通过下面的等价转换,可以将 FM 的二次项化简,其复杂度可优化到 $O(kn)$,即:

$$\sum_{i=1}^{n}\sum_{j=i+1}^{n}\langle v_i,v_j\rangle x_ix_j=\frac{1}{2}\sum_{p=1}^{k}\left(\left(\sum_{i=1}^{n}v_{i,p}x_i\right)^2-\sum_{i=1}^{n}v_{i,p}^2x_i^2\right) \tag{6}$$

通过随机梯度下降(stochastic gradient descent,简称 SGD)方法对 FM 模型进行训练,模型中各个参数的梯度可以表示如下:

$$\frac{\partial}{\partial\theta}y(x)=\begin{cases}1, & \text{if } \theta \text{ is } w_0\\ x_i, & \text{if } \theta \text{ is } w_i\\ x_i\sum_{j=1}^{n}v_{j,p}x_j-v_{i,p}x_i^2, & \text{if } \theta \text{ is } v_{i,p}\end{cases} \tag{7}$$

根据公式(6)和公式(7),FM 的训练和预测的复杂度均为 $O(kn)$,即 FM 能够在线性时间内进行训练和预测,非常高效,这为 FM 模型的广泛应用打下坚实的理论基础.不过,利用此模型的关键还在于模型输入特征向量的构造.

FM 模型被提出之后,迅速成为研究热点.对 FM 模型的扩展研究工作主要从 3 个方面展开:输入、处理和输出,如图 2 所示.模型输入扩展主要通过特征工程实现;模型处理扩展主要体现为特征交互,分为独立的特征交互和具有关联关系的特征交互,后者进一步把关联关系扩展为场关系和层次关系;输出方面则通过模型集成来实现.以下将重点阐述这些研究工作及其带来的影响.

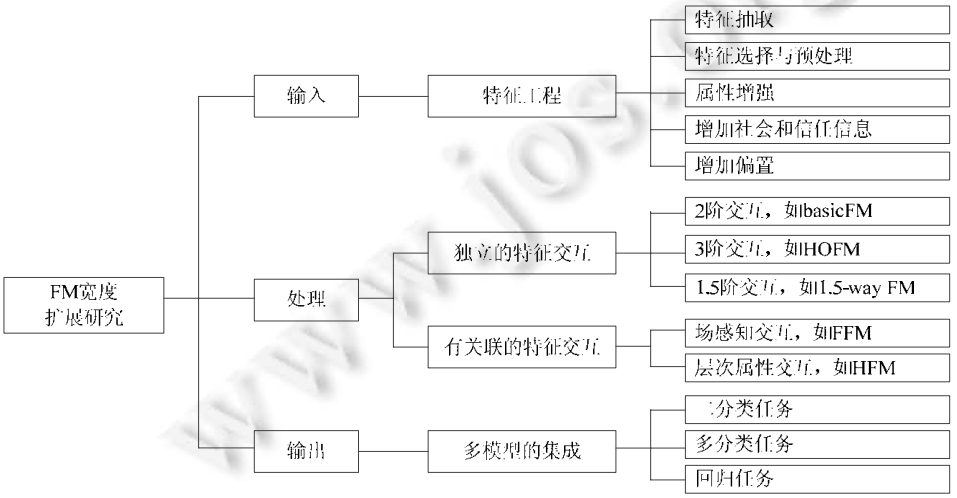

Fig.2 A research graph of wide extension of FM model

图 2 FM 模型的宽度扩展研究图谱

basicFM 中主要描述了 2 阶特征交互,能体现两个特征之间的相互影响.由于 2 阶特征交互关系的捕捉会极

大地提高模型预测和学习的复杂度,所以在实际应用中通常仅使用2阶.而且FM通过特征工程可以转化为矩阵分解模型、张量分解模型以及SVD++模型[21]等.2阶的FM相比于Poly2模型,优势表现为以下两点.

1) FM模型所需要的参数个数远少于Poly2模型.FM模型为每个特征构建一个隐向量,总参数个数为$O(kn)$,其中,k为隐式向量维度,n为特征个数,通常$k<<n$;Poly2模型为每个2阶特征组合设定一个参数来表示这个2阶特征组合的权重系数,总参数个数为$O(n^2)$;

2) 相比于Poly2模型,FM模型能够更有效地进行参数学习.当一个2阶组合特征没有出现在训练集中时,Poly2模型则无法学习该特征组合的权重,但是FM却依然可以学习.因为该特征组合的权重是这2个特征的隐向量的点积,而这2个特征的隐向量可以分别从别的特征组合中学习得到.

总体来说,FM是一种非常有效的能对2阶特征组合进行自动学习的模型.

尽管basicFM已经表现出较高的性能,但是更高阶的特征组合仍然能够引起研究人员的兴趣.Knoll等人提出了FM的高阶扩展模型HOFM(high-order factorization machine)[22],并且推导出了3阶FM模型的线性表示形式,可以用线性复杂度训练此3阶模型.为了把基本的2阶FM模型扩展到3阶,定义了一个矩阵$U\in\mathbb{R}^{p\times m}$,$w_0\in\mathbb{R}$,$W\in\mathbb{R}^n$,$V\in\mathbb{R}^{n\times k}$.3阶的FM模型可以表示为[23]

$$\hat{y}(x)=w_0+\sum_{j=1}^{n}w_ix_i+\sum_{j=1}^{n}\sum_{j'=j+1}^{n}x_jx_{j'}\sum_{f=1}^{k}v_{j,f}v_{j',f}+\sum_{j=1}^{n}\sum_{j'=j+1}^{n}\sum_{j''=j'+1}^{n}x_jx_{j'}x_{j''}\sum_{f=1}^{m}u_{j,f}u_{j',f}u_{j'',f} \tag{8}$$

引入3阶交互后,也不会丢失线性复杂性,因为3阶交互可以推导如下:

$$\sum_{j=1}^{n}\sum_{j'=j+1}^{n}\sum_{j''=j'+1}^{n}x_jx_{j'}x_{j''}\sum_{f=1}^{m}u_{j,f}u_{j',f}u_{j'',f}=\sum_{f=1}^{m}\left(\frac{1}{6}\left(\sum_{j=1}^{n}x_ju_{j,f}\right)^3-\frac{1}{2}\left(\sum_{j=1}^{n}x_j^2u_{j,f}^2\sum_{j=1}^{n}x_ju_{j,f}\right)+\frac{1}{3}\left(\sum_{j=1}^{n}x_j^3u_{j,f}^3\right)\right) \tag{9}$$

通过SGD方法对FM进行训练,3阶FM模型各个参数的梯度表示如下:

$$\frac{\partial}{\partial\theta}\hat{y}(x)=\begin{cases}1, & \text{if } \theta \text{ is } w_0\\ x_j, & \text{if } \theta \text{ is } w_j\\ x_j\left(\sum_{j'=1}^{n}v_{j',f}x_{j'}\right)-v_{j,f}x_j^2, & \text{if } \theta \text{ is } v_{j,f}\\ \frac{1}{2}x_j\left(\sum_{j'=1}^{n}u_{j',f}x_{j'}\right)^2-u_{j,f}x_j^2\left(\sum_{j'=1}^{n}u_{j',f}x_{j'}\right)-\frac{1}{2}x_j\left(\sum_{j'=1}^{n}u_{j',f}^2x_{j'}^2\right)+u_{j,f}^2x_j^3, & \text{if } \theta \text{ is } u_{j,f}\end{cases} \tag{10}$$

Knoll实现的3阶的FM方法与Blondel等人提出的类似[23].Prillo等人观察到:给定$n+2(n>=1)$个向量,一定有两个向量的内积是非负数.所以对于一个三角形关系,两条负数边对应的一定是正数边,这与实际应用会有不符.所以对于某些数据集,FM不能学习到2阶交互参数,这是FM的局限性,所以提出了1.5-way FMs[24].

并非越高阶的特征交互一定能带来更好的预测效果,Knoll等人对高阶的FM和马尔可夫随机游走(Markov random walk,简称MRW)方法进行比较发现:在只有用户对物品的评分数据时,MRW较FM表现得更好;但如果再增加了标签、物品种类等数据特征时,FM会取得更好的效果[25].Yurochkin等人提出了自适应的随机多阶交互模型MiFM(multi-way interactions of arbitrary order)[26],这个交互的选择是由一个基于随机超图的先验分布来决定.

考虑到FM模型中属性交互的必要性和有效性,得出如下结论:当属性数量较少时,如只有评分数据,那么属性的2阶交互无法体现其优越性,FM模型适合处理多属性,当多属性之间具有非线性的约束关系时,预测效果更好;对于通常的应用,2阶FM模型即可,如果在具体应用中发现个别属性之间存在1.5-way FMs所描述的问题,那么要删除这些属性交互,如果应用中多属性之间有强关联性,而且不考虑复杂度,则可以使用3阶的FM模型.更高阶的属性交互在实际应用中将不实用,而且随着深度学习与传统模型的结合,也没有必要在FM模型中进行更高阶的属性交互.

2.3 FM模型的场交互和层次交互扩展

虽然特征之间的非线性交互关系能够为预测提供更多的信息,但是并非所有的特征交互都有效,采用独热

编码后,特征之间的层次结构关系被全部摒弃,这将使得参数评估更复杂,预测准确率不够理想.相关研究主要从两个方面让模型能够体现特征之间的关系:基于场的交互和基于层次关系的交互.

(1) 场交互扩展.

场感知分解机(field-aware factorization machine,简称 FFM)模型是在 FM 的基础上,将相同性质的特征归于同一个场[13].除了特征的一维线性组合,很多数据集的特性表明,对不同场之间的特征交互的捕捉也非常重要.FM 模型的提出,是为了解决特征的交互问题.在多场的类别数据(multi-field categorical data)中,每个场中的特征都会与其余各场中的特征有着不同程度的交互,FFM 模型为每个与其他各场交互的特征都学习一个唯一对应的隐向量,从而充分地利用到数据中的场信息.在 FFM 模型中,每一个特征 x_i,针对其他特征的每一个场 f_j,都会学习产生一个隐向量 w_{i,f_j}.因此,隐向量不仅与特征相关,也与场相关.假设样本的 n 个特征属于 f 个场,那么 FFM 模型的二次项则有 $n\cdot f$ 个隐向量.根据 FFM 模型的场敏感特性,导出其模型方程[13]:

$$\hat{y}(x)=w_0+\sum_{i=1}^{n}w_ix_i+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\langle w_{i,f_j},w_{j,f_i}\rangle x_ix_j \tag{11}$$

其中,f_j 是第 j 个特征所属的场.如果隐向量的长度为 k,那么 FFM 模型的一次项参数有 n 个,二次项参数有 $n\cdot f\cdot k$ 个,模型预测复杂度是 $O(kn^2+kn)$.参数个数的设置,对预测精度会有较大影响.FFM 模型还支持并行化处理,所以计算速度可以进一步提高.而且 FFM 模型以场为基础,在稀疏数据的处理上,比 LR,Poly2,FM 效果要好很多.

虽然 FFM 模型能很好地利用数据中的场信息,然而 FFM 模型中的参数数量与特征数乘以场数量的积成正比,这使得实际应用场景下,参数数量会轻易地达到数以千万计甚至更多.这在现实世界的生产系统中是不可接受的.场加权的分解机 FwFM(field weight factorization machine)可以解决此问题[27].特征 i 和特征 j 的交互表示为 $x_ix_j\langle v_i,v_j\rangle r_{F(i),F(j)}$,其中,$x_i$ 和 x_j 分别为特征 i 和特征 j 对应的嵌入向量,$F(i)$和 $F(j)$分别为特征 i 和 j 所属的场,$r_{F(i),F(j)}$为场 $F(i)$和场 $F(j)$之间的交互强度,FwFM 模型定义为

$$\hat{y}(x)=w_0+\sum_{i=1}^{n}w_ix_i+\sum_{i=1}^{n}\sum_{j=i+1}^{n}x_ix_j\langle v_i,v_j\rangle r_{F(i),F(j)} \tag{12}$$

FwFM 模型相当于 FFM 模型的扩展形式,通过增加权重 $r_{F(i),F(j)}(F_k\neq F(i))$,来显示地捕捉不同场之间的交互程度.FFM 模型通过特征 i 与场 j 中的特征交互时,学习出隐向量 $v_{i,F(j)}$,隐式地捕捉不同场之间的交互程度.n 和 m 分别是特征数与场数量,k 为隐向量的维度,忽略偏置项 w_0,那么 FM 的参数数量为 $n+n\cdot k$,FFM 模型的参数数量为 $n+n\cdot(m-1)\cdot k$,FwFM 模型的参数个数为 $n+n\cdot k+m\cdot(m-1)/2$.一般地,$m<<n$,所以 FwFM 模型相比 FFM 模型参数数量少得多.FwFM 模型以远低于 FFM 模型的参数数量,取得了可以与 FFM 模型相竞争的预测精度,从而能更好地应用于实际的生产系统中.

(2) 层次数据交互扩展.

真实应用场景中,上下文(context)特征之间存在着层次关系.但是 FM 模型在预测过程中很少去挖掘上下文特征的层级特性,因此其预测效果会受到影响.以下从 3 方面对 FM 模型进行层次数据交互扩展.

其一,针对 basicFM 没有充分利用用户和物品中有价值的分类信息(valuable category information),Zhao 等人对用户、物品及其分类关系进行了探索[28]:首先,在考虑用户对某些物品分类的偏好时,提出了 UW-FM(user weight factorization machine)模型.然后,在考虑物品对用户分类的影响时,提出了 IW-FM(item weight factorization machine)模型.最后,合并这两个模型,提出了 CW-FM(category weight factorization machine)模型.CW-FM 使用层级的分类信息(hierarchical category information)来避免带有附属关系(subordinate relations)的特征进行交互.由于物品与其所属的类别特征、用户与其所属的类别特征之间存在附属关系,即每个物品或用户必定属于某个类别,所以模型会确保用户与物品之间的交互,以及用户与物品所属的类别特征进行交互时,能够相互独立,互不影响,从而更好地利用数据中带有的层级的种类信息来提升模型精度.

Wang 等人提出两阶段的建模方法 HFM(hierarchical factorization machine)[29]:第 1 阶段,首先在每个树结构的各节点上局部训练获取 FM 模型的参数,并返回初始输出(粗粒度),通过树结构的马尔可夫模型(tree-structured Markov model)对输出进行全局调整;第 2 阶段,使用广义的卡尔曼滤波算法(eneralized Kalman

filtering algorithm),比如对于数据集 MovieLens-1M,可以分别对用户、物品和时间属性进行 3 层的结构分层,用户为“Root→Gender→Age”,物品为“Root→Release-year→Genre”,时间为“Root→Day-of-week→Year”.

其二,针对特征之间的强层次关系进行交互.强层次关系是指:$w_{i,j}\neq 0\Rightarrow w_i\neq 0$ and $w_j\neq 0$.

Wang 等人提出了 SHA^2(strong hierarchical ANOVA kernel regression)模型及其特殊情况(β=1)下的 SHFM (strong hierarchical factorization machine)模型[30].SHA^2 模型表示为

$$\hat{y}_{SHA^2}(x)=b+\sum_{i=0}^{p}\sum_{j=i+1}^{p}\langle v_i\odot\beta,v_j\rangle x_i x_j \tag{13}$$

其三,为了强调某些特征的重要性,Oentaryo 等人提出了 HIFM(hierarchical importance-aware factorization machine)模型[31],把重要性权重(importance weights)和层次学习(hierarchical learning)引入到模型中.如,对于曝光率高的广告会分配更高的权重,因为如果没能准确预测此类广告,将会带来很大的损失,所以在模型中会赋予它们更高的权重.

实际应用中,各种属性之间的关系错综复杂.如果属性各自独立,那么直接可以作为 FM 模型的输入;如果属性及其特征值比较多,那么可以采用场交互来提高效率;如果属性之间存在包含、依赖等关系,或者为了强调某些属性的重要性,可以在模型中添加层次策略.

2.4 FM模型与传统模型的集成

集成学习(ensemble learning)通常是构建并结合多个弱学习器(weak learner)来完成学习任务.根据个体学习器的生成方式,目前的集成学习方法可分为:1) 个体学习器间存在强依赖关系、必须串行生成的序列化方法,代表技术有 Boosting;2) 个体学习器间不存在强依赖关系、可同时生成的并行化方法,代表技术有 Bagging 和随机森林(random forest).根据多个学习器是否相同,又分为两种:同质集成和异质集成.尽管 FM 在预测/推荐领域相对其他因子分解模型具有更好的通用性且能够取得更好的效果,但是在某些应用场合也会表现出局限性,所以在某些应用场景中,把 FM 模型与其他模型进行集成能够提供更多选择方案.以下对已经研究过的各种集成方案进行总结.

(1) 同质模型集成.

Yuan 等人通过引入 boosting 框架技术提出了 BoostFM 模型[32],即:根据用户与物品的特征信息和用户的隐式反馈(implicit feedback)信息来统一建模,通过加权组合的方式将多个同质的弱学习器转成为一个强学习器.Yan 等人对原始特征的 3 个方面(用户、物品和时间)分别抽取用户特征、物品特征以及时间特征,然后通过 GDBT(gradient boosting decision tree)和 FFM 模型学习到更加抽象的衍生特征;接着,通过这两个不同的特征集训练出两个不同的 FFM 模型;最后,让这两个 FFM 模型做非线性加权集成[33].Hong 等人提出了 CoFM(co-factorization machine)模型[34],采用 FM 模型同时对用户的决定(decisions),即对转发推特和用户转发推特的主题内容分别建模,两方面的数据训练出两个独立的 FM 模型,最终,这两个 FM 模型协同后输出结果.

(2) 异质模型集成.

Leksin 等人将 3 种模型——FM 模型、基于物品的协同过滤模型以及基于内容的主题模型进行线性组合、加权组合,最终结果与使用单一的方法相比有较大精度提升[35].

(3) 把单输出问题转为多输出或者回归问题.

Blondel 等人针对 FM 模型只能输出单值的限制,对模型进行扩展,使其能产生多输出,主要是把基于标量的学习函数扩展为基于向量的学习函数,采用一个 3-路的张量,把二分类问题扩展为多分类问题[36].Wang 等人提出了 RPFM(random partition factorization machine)模型[37],为了有效利用不同的上下文信息,采用随机决策树算法的思想,让样本中相似用户、物品或具有相似上下文的样本分发到树的相同节点中,同一节点中的各样本比处于原始数据集中的各样本之间具有更高的相关度.当要预测一个新样本时,先根据已建立好的树状结构,找到此样本应在的叶节点,然后使用该叶节点上已训练好的 FM 模型,求得此输入样本的输出预测值.若一共建立 N 棵树,对于此样本会有 N 个输出,最后取平均值,便是此新样本的最终输出预测值: $\hat{y}(x_i)=\sum_{t=1}^{N}\hat{y}_t(x_i)\big/N$.模型首

先采用 *K*-means 聚类算法对决策树中的节点进行划分,利用隐因子向量间的相似性作为划分标准,这样划分出来的每个子节点中的样本会具有更高的相关性.Pijnenburg 等人针对 LR 模型不能处理具有大量可能值的分类变量的问题,通过另一种建模技术来解决这个问题,比如朴素贝叶斯.然而,这样处理又失去了回归的一些优点,即模型对变量解释值的显式估计以及对变量到变量依赖的明确洞察和控制.让 LR 去直接处理多层次分类变量(many levels categorical variables),将产生稀疏的设计矩阵(design matrix),这会导致模型过拟合,产生极端系数值,所以预先通过 FM 将多层次分类变量转换成少量的数值变量,然后再应用于 LR 模型进行处理[38].

推荐系统领域中相关研究普遍认为:集成模型相比单一模型将具有更高的精度;如果单一 FM 模型精度已经能够满足应用需求,那么就没有必要集成,因为集成必将带来计算的复杂性;如果单一 FM 模型无法满足应用需求,那么可以考虑同质或异质模型集成;至于 FM 模型与何种模型以及采取何种方式进行集成,那么需要更多的尝试和分析.

2.5 特征工程与FM模型的应用

FM 模型中的特征很多是依赖于人工方式进行选择和设计.特征的多少会影响模型的计算复杂度,有些重要的特征和特征组合无法被专家轻易识别.所以实现特征的预处理以及自动组合挖掘,也成为推荐系统的一个研究热点.相比于其他机器学习系统,推荐系统更依赖于特征工程.根据使用的自动化程度,特征工程又分为人工特征工程和自动化特征工程技术.以下将从 3 个方面来总结特征工程在 FM 中的应用和影响.

(1) 从原始特征中提取更多的信息,再将其与原始特征一起作为 FM 的输入.

Loni 等人并不单独引入额外的信息,而是利用聚类算法,从已有的用户物品评分中充分挖掘信息,并将其作为部分新的特征,提高 FM 模型的预测精度[39].由于用户和物品的簇(cluster)信息隐藏在用户物品矩阵之中,层次交互扩展中采用的策略是由专家指定用户和物品的分类,文中使用 *k*-means 聚类算法,将相近的用户聚为一个簇,最终每个簇都会分配一个在域(domain)中的唯一 id,然后将每个域中各个用户和物品的簇信息,添加到 FM 模型的输入特征向量中,从而扩增了 FM 模型的输入特征向量的维度.即:将隐藏在用户物品评分信息中的特征提取出来,并显式地添加到 FM 的输入特征向量之中,最终起到提高模型预测精度的目的[39].不过,这种方法存在新用户的冷启动问题,因为对于刚注册的用户,由于缺少足够的历史信息,聚成簇会损失掉新用户的个性化问题.可以采用以下解决方法:以电子商品推荐为例,对老用户(比如可以假定购买 5 件以上商品的用户为老用户)和新用户分别进行聚类,这样可以避免减少过多的新用户信息.

(2) 从原始特征中选择部分重要特征,或者进行特征的合并.

并非所有特征的交互都有效,有些交互可能会成为噪声,破坏模型的泛化能力.从原始特征中选择重要特征,或者过滤掉不重要的特征,或者进行特征的合并,会使得减少特征数的同时,进一步提高预测精度和计算效率.Cheng 等人提出了 GBFM 模型[40],把 GBM(gradient boosting machine)模型与 FM 结合,着重于选取好的特征进行交互.提升方法的思路为:对于一个复杂的问题,将多个专家的判断进行适当综合,所得出的结果要比任何一个专家单独判断更加精确.每一步产生一个弱预测模型(如决策树),并加权累加到总模型中,可以用于回归和分类问题;而梯度提升是在此思想下的一种函数或模型的优化方法,如果每一步的弱预测模型都是拟合损失函数的负梯度而得,则称为梯度提升(gradient boosting).通过使得损失函数在梯度上减少的方式进行 *m* 次迭代,最终合并得到一个优秀的提升模型.GBFM 训练的主要迭代公式为

$$\hat{y}_s(x)=\hat{y}_{s-1}(x)+\sum_{i\in C_p}\sum_{j\in C_q}\prod[i,j\in x]\langle V_p^i,V_q^j\rangle \tag{14}$$

GBFM 模型中,如何选取好的特征是重点,采取的特征选择算法是选择尽可能使目标函数降低最快的特征,比如现在有用户(user)、物品(item)和情感(mood)这 3 个上下文特征.对于原始 FM,其模型为

$$\hat{y}(x(u,i,c_3))=w_0+w_u+w_i+\langle V_u,V_i\rangle+\langle V_u,V_{C_3}\rangle+\langle V_i,V_{C_3}\rangle \tag{15}$$

即含有所有特征的交互,而 GBFM 训练完成特征选择后,最终可能去除了 item 和 mood 的交互,产生的模型可能是

$$\hat{y}(x(u,i,c_3))=w_0+w_u+w_i+\langle V_u,V_i\rangle+\langle V_u,V_{C_3}\rangle \tag{16}$$

Xu 等人认为,GBFM 通过基于梯度提升的贪心算法选择特征交互,降低了模型复杂度.但是使用启发式算法并不是最优的特征交互选择方式,并且经过这种方式处理后的交互特征的数量仍然十分庞大,所以提出了稀疏 FM(sparse factorization machine,简称 SFM)模型[41],从学习模型中直接选取交互特征,交互过程中的所有无用特征将会被剔除.为了减少模型的复杂度,识别出有关联的用户特征和物品特征.即仅对有关联的特征进行交互,所以需要学习的参数也大幅减少,模型表示为

$$L_{FM}=(P,Q)=\sum_{p,q}\left(r_{pq}-\begin{bmatrix}x_p^U\\x_q^I\end{bmatrix}^T\begin{bmatrix}P\\0\end{bmatrix}\begin{bmatrix}0\\Q\end{bmatrix}^T\begin{bmatrix}x_p^U\\x_q^I\end{bmatrix}\right)^2=\| R-X_U PQ^T X_I^T \|_F^2 \tag{17}$$

Selsaas 等人根据属性值进行自动的特征合并,然后进行交互,能够有效缓解特征的稀疏性.如在考虑设备属性时,PC 机具有 IP 地址和 Cookie,移动电话(cell phone)则没有 IP 地址,但是它绑定了电话号码,如果按照独热编码方式将产生很多的特征,而且很多特征是稀疏的.所以在采用模型预测之前,先对这些特征进行合并,既能减少计算时间,也能提高预测精度[42].Punjabi 等人提出了鲁棒 FM(robust factorization machine,简称 RFM)和基于场的 FM 变体模型 RFFM.由于数据在收集过程中采用不同的设备,即使相同的设备,因为浏览器不同,都会产生不同的数据,这对数据质量是一个挑战,采用鲁棒优化(robust optimization,简称 RO)框架可以去除噪声[43].Lu 等人针对多视图学习(multi-view learning,简称 MVL)和多任务学习(multi-task learning,简称 MTL)提出多线性 FM (multilinear factorization machine,简称 MFM)模型来应对异构的多源特征[44].Liu 等人提出了 LLFM(locally linear factorization machine)模型来应对局部线性分类器(locally linear classifiers)问题[45].

(3) 对模型属性进行提升,不同属性可以选择不同的提升方法.

实际应用中,用户和物品具有针对性,即属性会表现出一定的特征偏好.以用户为例,80 后以及 90 后初期的男性用户可能会特别喜欢类似于《生化危机 6》、《魔兽》、《刺客信条》等游戏改编的电影,而 90 后女性,特别是 95 后女性可能更中意《从你的全世界路过》、《有一个地方只有我们知道》、《属于你的我的初恋》等爱情片.另外,不同属性之间的影响程度也不同.例如,用户所属职业的重要程度要高于用户家庭所在的地理位置.本项目组在 FFM 模型的基础上提出了智能化场感知模型 iFFM[46],该模型对关键属性进行提升,运用特征工程技术将因子选择智能地嵌入到算法求解过程中,并综合利用吉布斯采样和 SGD 训练模型来提高推荐精度.iFFM 模型中对特征 x_i 进行关系映射表示为

$$\hat{x}_i=x_i+B_i \tag{18}$$

其中,B_i 表示属性 x_i 的提升项,可以进一步分解为 $u_i^T p_u$ (对应用户)或者 $v_i^T q_v$ (对应物品).其中,p_u 和 q_v 分别为用户 u 的偏好、物品与属性的关联程度,u_i 和 v_i 为相应影响权重.假设 x_i 表示“爱情片”,那么 $u_i^T p_u$ 刻画的是用户 u 对“爱情片”的隐含态度.假设 x_i 表示“动作片”,即使用户 u 在数据集上没有观看“动作片”的记录,则依然可以用 $u_i^T p_u$ 模拟用户 u 的偏好.此时,用户的偏好和物品的属性特征可以表示为:

$$y'_{user}(x)=\sum_{i=1,i\notin t(u)}^{t_1} w_{f_i}(x_i+u_i^T p_u)=\hat{y}_{user}(x)+\sum_{i=1,i\notin t(u)}^{t_1} w_{f_i}x_i \tag{19}$$

$$y'_{item}(x)=\sum_{i=1,i\notin t(v)}^{t_2} w_{f_i}(x_i+v_i^T q_v)=\hat{y}_{item}(x)+\sum_{i=1,i\notin t(v)}^{t_2} w_{f_i}x_i \tag{20}$$

其中,w_{f_i} 为特征 x_i 所在的场中 f_i 对 x_i 的影响权重.实际操作中,先判断 x_i 所在的场中 f_i 所属的集合:若是用户属性集合,则按照公式(19)计算;若是物品集合,则按照公式(20)来计算;其余情况下,默认结果为 0.模型表示为

$$\hat{y}(x)=w_0+\sum_{i=1}^n w_i x_i+\varphi_{user}(x)+\varphi_{item}(x)+\sum_{i=1}^n\sum_{j=i+1}^n\langle w_{i,f_j},w_{j,f_i}\rangle x_i x_j \tag{21}$$

(4) 在模型中添加社交或信任等相关信息的特征来提高精度.

随着社会网络和移动技术的发展,社交信息在预测/推荐中起到更大的作用.Ding 等人提出了 SCFM(social and crowdsourcing factorization machine)模型,在 FM 中添加社交和众包信息[47].Zhou 等人提出了 SocialFM 模型,基于社交关系对 FM 进行改进,模型中采用的输入包括用户、物品、用户之间的相似性和信任关系,通过用户之间的关系来提高推荐精度[48].Rendle 等人也提出了采用特征工程,如社会网络信息来提高推荐精度[49].

(5) 在模型中添加偏置进行调整,平衡特征在模型中所占的比重.

Chen 等人指出,大多数的推荐系统都在推荐流行度高的物品,缺乏新颖性,也导致长尾现象的产生[50].其原因是在训练数据集中,频繁出现的物品占据了整个物品集的绝大部分,导致推荐系统给出的推荐列表中流行度高的物品更有优势.为了缓解这种现象,尽可能去推荐一些用户没有接触过但是可能会感兴趣的物品,提出了cost-sensitive FM 模型,其思路是降低流行度偏置(popularity bias),在推荐精度和物品流行度之间找到平衡点.模型构建方法是将代价敏感的学习(cost-sensitive learning,简称 CSL)与 FM 模型进行集成,在不破坏推荐质量的基础上,达到降低流行度偏置的目的.

正如机器学习领域大家所认同的一个观点:数据和特征决定了机器学习的上限,而模型和算法只能逼近这个上限,由此可见特征工程的重要性.所以在应用 FM 模型时,必须要考虑它所应用的领域.FM 被广泛使用于各个领域,表 1 列出 FM 模型被提出后在各个领域的应用.除了表 1 中列出的,FM 模型也在其他领域得到应用,如微博排名[51]、表演艺术市场决策[52]等等.在实际应用中,可根据具体的问题来选择具体特征,通过调节参数来达到理想的预测/推荐效果.

Table 1 Application field of FM model

表 1 FM 模型应用领域

应用领域	预测/推荐任务
电影评分[12]	对电影进行评分预测,为新老用户推荐电影
计算广告领域[13,53]	广告点击率和转化率预测,为用户推荐广告
电子商务领域[34,54,55]	预测用户的购买行为,为用户推荐商品
Web 服务领域[56−58]	Web 服务的 QoS 预测,推荐 Web 服务及组合
股票领域[59]	预测股票走势,为用户推荐股票
基因领域[60]	对基因进行分类
信用领域[61]	预测用户的信用等级
找工作领域[35]	为用户推荐合适的工作机会
移动 APP[62]	为用户推荐可能感兴趣的移动 APP
异常检测[63]	从使用行为中检测异常点

3 FM 模型与深度学习模型的集成

相比逻辑回归模型以及其他因子分解模型,FM 在很多应用领域都展现了其独特的优势.不过,它仍然属于多变量的线性模型,因为在 FM 模型中,对于每个参数$\theta\in\{w_0,\{w_i\},\{v_{if}\}\}$,都可以得到$\hat{y}(x)=g+h\cdot\theta$,其中,$g$ 和 h 与θ无关.而现实世界中的数据关系通常是高度非线性的,无法直接采用线性模型或浅层模型来表达.FM 属于浅层模型,可以有效提取 1 阶特征和 2 阶特征,但是难以挖掘高阶特征.尽管特征工程技术能够帮助 FM 提高精度,但是特征的自动组合非常重要,已成为推荐系统的热点研究方向之一,深度学习作为一种先进的非线性模型技术在特征组合挖掘方面具有很大的优势.

深度学习应用于预测和推荐系统后,激发并产生了很多相关研究和成果[64−66].Zhang 等人把这些模型分为两类:一类是只使用深度学习模型,可以是单一模型,也可以集成多个不同模型来提高推荐的多样性;另一类是把深度学习与传统推荐方法结合起来,可以采用松耦合或者紧耦合的策略[67].实践结果表明:深度学习适用于某些方面,如深度卷积神经网络(deep CNN)非常适合图像特征的提取[68],但是传统模型解释性更强,所以将两者结合将更有效.Wide&Deep 框架自被谷歌提出后被广泛研究和使用[7],此框架是线性模型和深度学习模型有机结合的典范,宽度(wide)部分用于提高记忆性(memorization)能力,深度(deep)部分用于增强泛化性(generalization)能力,Deep&Cross 模型对宽度部分进行进一步扩展[69].由于 FM 模型在预测/推荐领域占有非常重要的地位,因而 FM 模型及深度学习模型的集成方面展开了大量研究,以下将进行详细阐述.

3.1 FNN模型

FM 支持的神经网络(factorization machine supported neural network,简称 FNN)模型于 2016 年被提出[70],其思路为:采用 FM 模型对原始特征的嵌入层进行初始化,将 FM 的输出作为输入放到深度神经网络(dense neural network,简称 DNN)中.图 3 所示为本文对该模型的结构描述.FNN 证明:利用 FM 初始化参数能够使梯度更快地

收敛,最大限度地避免训练过程陷入局部最小,可以获得更好的结果.

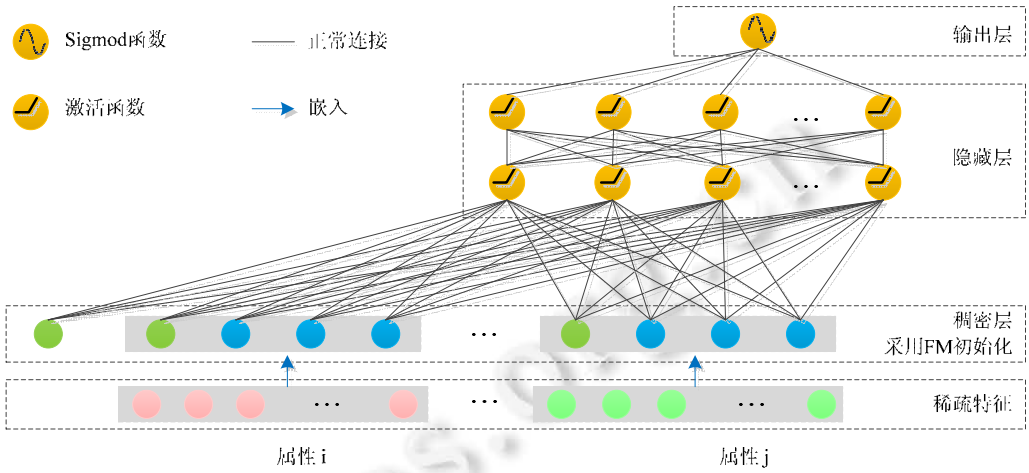

Fig.3 Architecture of FNN model

图 3 FNN 模型结构

模型中最底层是 FM 模型,通过训练后生成一个 z 向量,表示为

$$z=(w_0,z_1,z_2,...,z_n),z_i=(w_i,v_i^1,v_i^2,...,v_i^k) \tag{22}$$

模型隐含层表示为

$$a^{(1)}=f(w^{(1)}z+b^{(1)}),a^{(l+1)}=f(w^{(l)}a^{(l)}+b^{(l)}) \tag{23}$$

其中,$a^{(1)}$是第 1 层(向量 z)的输出;$a^{(l+1)}$是第(l+1)层的输入,也是第 l 层的输出;f(*)为激活函数,如 *sigmoid*,*tanh*, *relu* 等;w 和 b 分别为权重矩阵和偏置.

FNN 可以认为是采用 FM 初始化的 Wide&Deep 模型的深度部分.SNN 模型与 FNN 模型的区别在于底层的训练方法不同,它采用全连接方式,初始化时采用限制玻尔兹曼机(RBM)和自动编码机(DAE).

PNN 模型也采用类似 Wide&Deep 的框架,不过它在嵌入特征时增加了两两交叉的功能,而不是把所有的参数直接输入到隐藏层[71].

3.2 Wide&Deep模型

Wide&Deep 模型是谷歌 2016 年提出[7],最早是用来解决 Google Play 应用商店的 APP 推荐问题.模型中将推荐看做是一个搜索排序问题,输入用户和文本信息的集合,输出经过排序的物品列表.推荐主要解决两类问题:记忆性和泛化性.记忆性由宽度模型主导,根据历史数据学到的信息,比如“百灵鸟会飞”、“老鹰会飞”;泛化性由深度模型主导,推断在历史数据中从未见过的情形,比如“飞机会飞”.具体到推荐系统中,记忆性主要学习用户行为习惯,向用户推荐与历史购买信息相关的物品,解决准确性问题;泛化性是指可以向用户推荐其从未购买过的物品.Wide&Deep 模型将宽度模型(传统的线性模型)和深度模型(深度学习模型)融合在一起进行训练,图 4 所示为本文对该模型的结构描述(宽度部分和深度部分的输入特征不同).

宽度部分采用线性结构,数学公式为 $\varphi_{wide}=w^Tx+b$,其中,x 为特征矩阵,w 为权重矩阵,b 为偏置.与传统逻辑回归不同的是,这里的 x 包含原始特征和少数人工选择的交叉特征.虽然没有使用 FM 模型,但是应用了该模型中特征交叉这一思想.

深度部分采用 DNN 模型.由于人工构建的交叉特征有限,而交叉特征的类别可能包含多个,如 3 个或 4 个,因此需要 DNN 自动构建一些特征,数学公式为

$$a^{(l+1)}=f(w^{(l)}a^{(l)}+b^{(l)}) \tag{24}$$

其中,$a^{(l)}$是第(l+1)层的输入,也是第 l 层的输出;f(*)为激活函数如 *sigmoid*;$w^{(l)}$和 $b^{(l)}$分别为第 l 层权重矩阵和偏置.

从特征角度分析,宽度部分主要学习 1 阶特征和少量 2 阶特征,deep 部分用来学习高阶特征.其优势是同时学习低阶特征和高阶特征.

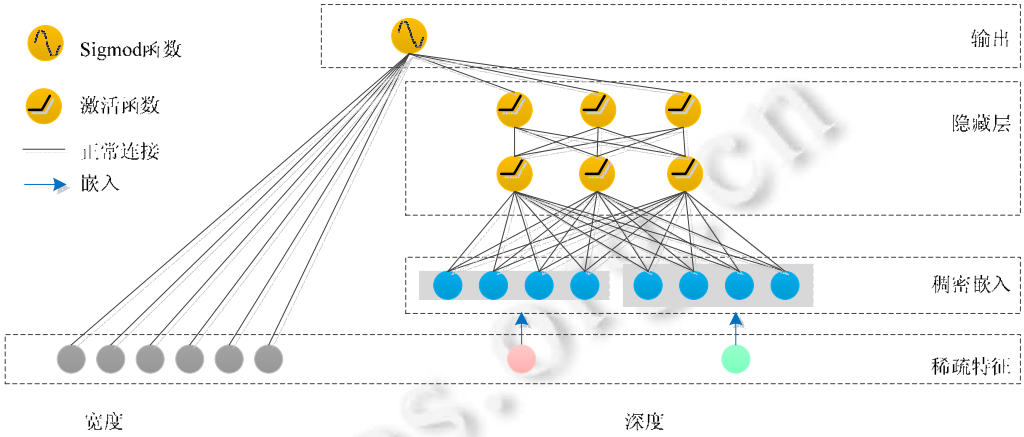

Fig.4 Architecture of Wide&Deep model

图 4 Wide&Deep 模型结构

3.3 Deep&Cross模型

Wide&Deep 模型中,宽度部分采用基于逻辑回归的简单线性模型,它无法获得充分的低阶特征交互信息,为了弥补这一不足,Deep&Cross 提出采用 Cross 网络作为宽度模型来获取低阶交互信息[72],而且 Cross 网络可以变换为 FM 等模型,所以具有非常好的一般性,图 5 所示为本文对该模型的结构描述(宽度模型使用 Cross 网络,可以用 FM 模型替换).

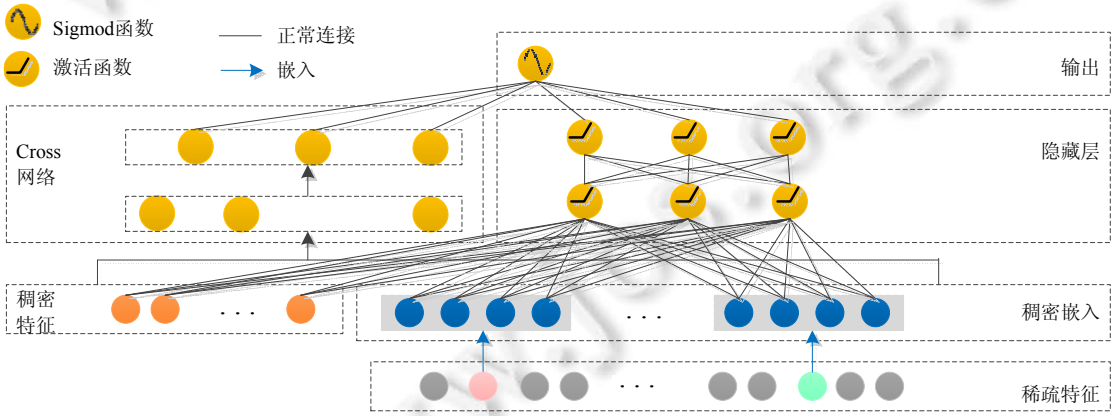

Fig.5 Architecture of Deep&Cross model

图 5 Deep&Cross 模型结构

3.4 DeepFM模型

Wide&Deep 模型以及 Deep&Cross 模型中,宽度和深度部分采用不同的输入,这要求在模型使用中能够判断特征的选择.在 DeepFM 模型[73]中,宽度和深度部分共享原始的输入特征向量,模型更易于使用,图 6 所示为本文对该模型的结构描述(宽度和深度部分共享原始的输入特征向量,使得模型更易于使用).从特征角度来看,DeepFM 的宽度部分采用了 2 阶的 FM 模型结构,主要学习 1 阶特征和 2 阶特征,这也改进了 Wide&Deep 模型中的宽度部分,增加了更多的低阶特征交互学习.考虑到 CNN 模型偏向相邻特征的交互,RNN 模型偏向点击数

据的预测,所以 DeepFM 的深度部分采用 DNN 模型.

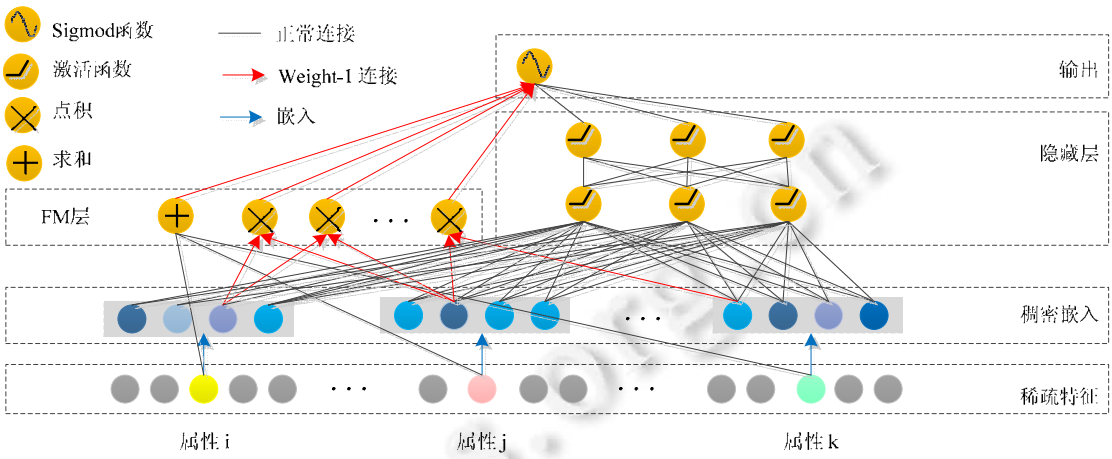

Fig.6 Architecture of DeepFM model

图 6 DeepFM 模型结构

3.5 NFM模型

NFM 模型[74]也采用了类似 Wide&Deep 的宽度和深度学习框架,其输出表示为

$$\hat{y}(x) = w_0 + \sum_{i=1}^{n} w_i x_i + f(x) \tag{25}$$

其中,第 1、2 部分是线性回归模型,与 FM 一致;第 3 部分的 $f(x)$是模型的核心,主要对特征交互进行建模,是一个多层前向神经网络.图 7 所示为本文对该模型的结构描述(深度学习部分的输入是交互特征).

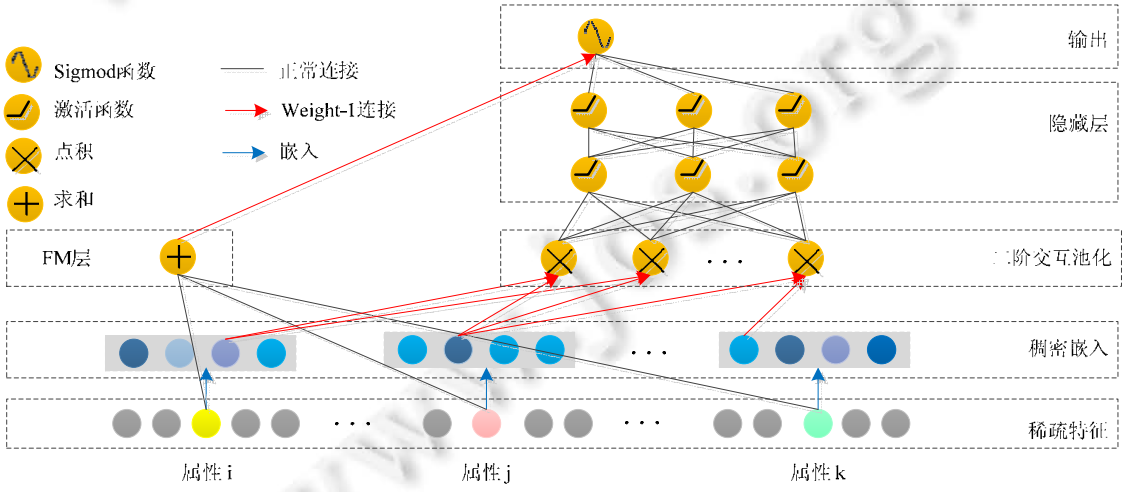

Fig.7 Architecture of NFM model

图 7 NFM 模型结构

3.6 宽度和深度学习模型集成方式分析

FNN 模型实现了 Wide&Deep 框架的深度部分,它采用 FM 对参数进行初始化,然后再深度学习.DeepFM 与 PNN 结构很相似,不同在于 FM 模型处理的属性被单独作为宽度部分.DeepFM 与 Wide&Deep 的不同在于:它把宽度部分的属性替换为 FM 处理后的属性,而且宽度部分与深度部分输入相同.NFM 模型也是基于 Wide&Deep

框架,主要通过输入特征的预处理来提高推荐效果.Deep&Cross 模型则是对文本信息进行处理,并且将 ResNet[75]应用到非图像领域.表 2 对这些模型的特征进行了综合比较.

Table 2 Comparison of depth learning model based on FM model and linear model

表 2 基于 FM 模型的深度学习模型比较

模型	1 阶特征	2 阶特征	高阶特征学习	高阶输入	共享输入
FNN	×	×	deep 网络	嵌入向量	√
Wide&Deep	√	少量,人工	deep 网络	嵌入向量	×
Deep&Cross	cross 网络	cross 网络	cross+deep 网络	嵌入向量	√
DeepFM	√	√	deep 网络	嵌入向量	√
NFM	√	×	deep 网络	2 阶特征	√

这些模型基本包括了传统模型(尤其是 FM)与深度学习模型的不同融合方式,有些是松耦合(两者的最后结果进行合并),有些是紧耦合(一个模型的输入依赖于另一个模型).实际应用中,如果要提高预测或推荐的精度,需要结合应用领域来增加特征工程.

为了深入理解不同模型及其不同结构的优势,本项目组提出了 DGFFM 模型.模型中,宽度部分采用 FFM 模型,深度部分采用 DenseNet 模型,并且在特征中添加了时间动态因子.因为 FNN 和 Wide&Deep 采用两种典型的框架,所以本文实现了两种 DGFFM 模型:DGFFM(W&D)模型采用 Wide&Deep 结构,DGFFM(FNN)采用 FNN 的结构.实验采用的硬件平台为 Inter® Core™ i7-7700 CPU@3.60GHz,65.86GB 内存,976GB 硬盘,64 位 Ubuntu 16.04 操作系统的工作站;编程语言为 Python,框架使用 TensorFlow;数据集为 MovieLens 1M 数据集(简写为 ml-1m,http://grouplens.org/datasets/movielens/1m/)和 Criteo 数据集,ml-1m(softmax 分类)中采用 RMSE 评价指标,Criteo(二分类问题)中采用 AUC 和 LogLoss 评价指标.模型比较结果见表 3,其中,FNN,DeepFM 和 DGFFM 的隐向量维度均设置为 20.DenseNet 部分输出通道设置:ml-1m 为[100,48,32],分别指模块 1 输出通道数、翻译层输出通道数和模块 2 输出通道数,Criteo 为[256,128,64].由于各种深度模型适用于不同的应用场合,所以在实验中没有测试每种模型的效果,只把所提出的 DGFFM 模型与 FNN 和 DeepFM(采用 Wide&Deep 框架)模型进行了对比.

Table 3 Accurace comparison of different models

表 3 模型精度比较

数据集	FNN	DeepFM	DGFFM(FNN)	DGFFM(W&D)
ml-1m	RMSE:0.7608	RMSE:0.7493	RMSE:0.7231	RMSE:0.7028
Criteo	AUC:0.7935 LogLoss:0.4589	AUC:0.7989 LogLoss:0.4503	AUC:0.8167 LogLoss:0.4264	AUC: 0.8185 LogLoss:0.4118

根据表 3 的实验结果,得出以下结论.

1) DeepFM 在两个数据集上均比 FNN 取得了更好的效果.在 ml-1m 上,DeepFM 的 RMSE 减少 1.5%;在 Criteo 上,LogLoss 减少 1.9%,AUC 提高 0.6%.这在一定程度上说明 Wide&Deep 具有结构优势;
2) 3 个模型中,DGFFM 均取得了最好的结果.其原因是 DGFFM 在宽度学习部分基于 FFM 模型,增加了时间因子以及其他的特征工程,而且 DenseNet 相对于标准 DNN 也具有一定优势,宽度和深度两部分优势相结合,进一步提高了模型的预测精度;
3) DGFFM(W&D)结果略好于 DGFFM(FNN),在 ml-1m 上,RMSE 减少 2.8%;在 Criteo 上,LogLoss 减少 3.4%,AUC 提高 0.2%.由于 Wide&Deep 结构中为了保证最终层数不变,深度部分初始输入采用的是偏置+一次项,因此 Wide&Deep 结构中的 DGFFM 比 FNN 结构中的 DGFFM 多了一部分,这可能对最终结果也造成了一些影响.但总体而言,Wide&Deep 结构略好于 FNN 结构.

4 FM 模型学习与分布式并行实现

4.1 FM模型的学习与优化

Rendle 等人采用 3 种学习方法训练 FM 模型:SGD、交替最小二乘法(alternating least-squares,简称 ALS)和马尔可夫蒙特卡洛(Markov chain Monte Carlo,简称 MCMC),这些都可以在 libFM 中找到源码[21].Bayer 等人提出了库 fastFM[76],实现了 FM 的回归、分类和排序,简化了 FM 的使用.从关系型数据设计矩阵会非常庞大,使得学习和预测变得缓慢或者从标准的机器学习算法不可行的角度,Rendle 等人针对关系型数据实现了 FM[77].针对 FM 中包含一个非凸优化问题,导致局部最小化,Blondel 等人提出基于核范式的 FM 的凸形式,采用双块坐标下降算法(two-block coordinate descent algorithm)优化学习[78].Yuan 等人提出两种优化策略——RankingFM(ranking factorization machine)和 LambdaFM(lambda factorization machine)优化 FM 模型[79].

Pan 等人针对广告交易数据的稀疏性,即存在大量零元素,可能会严重影响 FM 模型的性能,提出一种新的稀疏因子分解机(SFM)模型[80],其中使用拉普拉斯分布而不是传统的高斯分布来对参数进行建模,因为拉普拉斯分布可以更好地拟合更高比例的稀疏数据零元素.Saha 等人提出了 NPFM[81],它假设数据服从泊松分布(Poisson distribution),这对于建模和数据训练计算都非常有利;NPFM 作为一个非参数模型,会从数据本身发现最适合的隐因子数量.由于 FM 中用户、物品和上下文变量之间的交互被建模成它们各自隐因子特征的线性组合(linear combination),但是将用户、物品和上下文变量之间的交互限制成线性组合并不现实,为了解决这一限制,Nguyen 等人提出了高斯过程的因子分解机(Gaussian process factorization machine,简称 GPFM)模型,即:使用高斯过程的非线性概率算法来应对上下文感知推荐,可以被应用到隐式反馈数据和显式反馈数据集[82].一般的高斯处理回归的推断和学习算法都是关于数据集样本大小的立方级复杂度(cubic complexity),Huang 等人提出了 GGPFM(grid-based Gaussian processes factorization machine)模型捕捉用户与物品之间的非线性交互[83],将潜在特征(latent features)赋予网格结构(grid structures)降低模型复杂度.通常的学习和训练方法即可满足一般的应用需求,但是不同的应用背景对应不同的特征,模型学习和优化方式的调整可以提高精度,不过都需要通过实验来反复验证.

4.2 FM模型的并行实现

精度和效率是评价预测/推荐模型的两个重要指标.通过从宽度上改进模型以及从深度上与深度学习模型的集成,可以极大地提高模型的精度.FM 提供了线性的计算复杂度和有用的数据嵌入,但是当数据和特征规模增大时,模型的扩展代价非常高.FM 与深度学习集成后,大数据和模型扩展性问题更加严重.机器学习算法的独特性在于:(1) 迭代性,模型的更新需要循环迭代多次;(2) 容错性,每个循环中可能产生的错误不影响模型最终的收敛;(3) 参数收敛的非均匀性,模型中有些参数经过几次循环后不再改变,其他参数可能仍需要很长时间收敛.面对海量的数据加上复杂的数学运算,这些决定了分布式机器学习系统的特殊性.大数据的机器学习存在着许多挑战和机遇[84,85],通常会采用两种方案.

- 首先是数据并行方案.采用经典的主-从服务模式对训练数据进行划分,分布式存储到各个节点上,每个节点都运行着一个或多个模型训练进程,各自完成前向和后向的计算得到梯度;训练结束后,各节点把参数传递给主服务器进行参数的合并与更新,主服务器把更新后的参数再分发到各个节点,再次进行训练.通过多个节点并行训练来提高学习效率;
- 其次是结构并行方案.当模型巨大、单机内存不足时,将计算工作进行划分,即同一个大模型的不同部分交给不同节点负责(如多层网络的各个节点),不过,这样会产生很大的通信开销.结构并行相对数据并行更加复杂,不过开源框架如 TensorFlow 平台直接支持结构并行.

目前,主流的解决方案是使用分布式框架和并行计算模式,硬件方面则使用 GPU 和 TPU 等进行加速.以下将总结 FM 及其变体在提高效率方面的相关研究.

MapReduce 并行计算模式在大数据处理领域应用非常广泛,也可被用于提高 FM 的学习效率.Sun 等人实现了基于 MapReduce 的 SGD 算法用于 FM 模型的学习,主要通过数据并行来提高模型的学习效率.不过,

MapReduce 模式的特征也决定了其更适合处理数据并行[86].Yan 等人基于 spark 平台实现 FM 模型的学习,其核心思想与 MapReduce 类似[47].Knoll 等人采用参数服务器(parameter server,简称 PS)为 FM 提出一种分布式的 SG 算法[22].PS 是一个算法的计算引擎,其计算由两组分开的计算机完成:服务器(server)和工作者(worker). server 用于管理和更新模型的参数,worker 处理训练数据,任务调度器和资源管理器负责控制数据流.Li 等人也采用 PS,通过一个依赖图(dependency graph,简称 DAG)提供了灵活的数据一致性模型[87].Zhong 等人在参数服务器上实现了分布式的 FM,即 DiFacto,采用自适应的内存限制和频度自适应的正则化机制,基于数据和模型统计来执行细粒度的控制,并在多台机器分发 DiFacto[88].Li 等人提出一个新的系统框架,集成了参数服务器和 MapReduce 模式.通过 MapReduce 实现数据并行,通过 PS 实现模型并行,并解决了通信开销问题和参数更新冲突问题[89].机器学习离不开分布式并行计算框架和 GPU 等硬件的支持.

5 FM 模型研究展望

“互联网+”的发展以及大数据技术正在开启一个全新认知的大数据时代,FM 模型是目前预测/推荐领域研究和应用最广泛的模型之一.图 8 所示大数据环境下预测/推荐系统的框架及其所面临的问题.

- 问题 1:多源异构数据带来特征表示的多样性和复杂性.尤其是视觉数据,其特征维度高且数量大,传统推荐主要关注非视觉文本数据及其交互,对非视觉和视觉特征的融合是新型推荐系统建立的基础;
- 问题 2:现有推荐模型对于特征进化趋势缺乏表示.用户偏好与物品特征都会随着时间而发生变化,对这些变化趋势进行合理建模将会提升推荐效果,动态特征的提取与表示是构建动态推荐模型的关键;
- 问题 3:传统模型与深度学习框架适用于不同领域,现有的融合方法不论是共享输入还是独立输入,两部分都是松耦合,两者的结合主要用于增加推荐的多样性.为了发挥共同优势,需要研究模型的集成策略;
- 问题 4:大数据要求模型和训练算法具有高扩展性和高效率,尽管通用分布式计算框架提供了并行处理支持,但两类不同的学习方式具有不同的要求,相关并行处理方法和关键技术还需进一步研究.

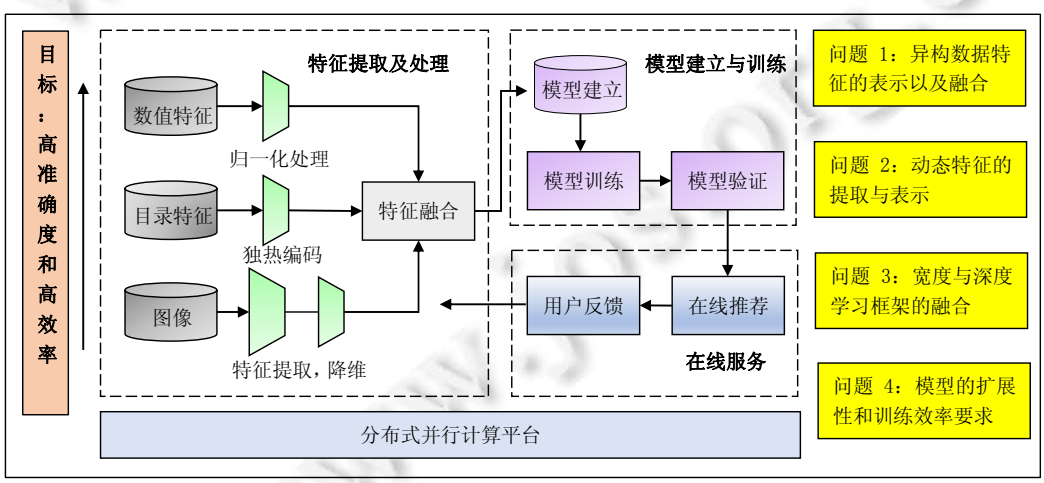

Fig.8 Framework of prediction & recommendation system and the problems faced by big data environment

图 8 预测/推荐系统框架及大数据环境下所面对的问题

通过对已有 FM 的相关研究和应用进行分析,我们认为,目前工作可以从以下两方面进行深入.

5.1 时间动态性

动态建模是推荐系统面临的挑战之一.Koren 将时间动态性应用到矩阵奇异值分解模型 SVD 中,通过提取非视觉特征的时间动态因子,分别对用户偏置和物品偏置进行动态建模,取得了较好的推荐效果[11].He 等人同

时对非视觉特征和视觉特征进行轻量级的时间建模,虽然没有对变化趋势进行细分,但也极大地改善了推荐结果[90].谷歌将流行趋势划分为 6 类:持续上升、季节上升、突然上升、持续下降、季节下降和突然下降,基本囊括时尚物品的所有变化特性.用户行为变化趋势和物品变化趋势有较大差异,通常短期内物品的流行趋势变化不明显,研究不同特征的不同变化趋势对于构建推荐模型具有重要意义[91].相关研究属于特征工程领域,即在原有的属性中添加时间因子.FM 模型中可以归纳为两类时间动态性:偏置动态性和特征动态性.

(1) 偏置动态性.

在推荐系统中,又分为用户的偏置动态性和物品的偏置动态性.设置用户偏置动态性的原因在于:用户对物品的评分习惯可能会随着时间而发生变化.例如,用户 Aphro 过去倾向于给电影《秒速五厘米》评 9 分,现在她对于动画片的狂热减退,只会评 8 分.同样地,设置物品的偏置动态性的原因在于:随着时间的推移,物品的人气会随之变化.以电影《战狼 2》为例,上映 4 小时的票房达 9 741 万,接着,凭借演员精湛的演绎、电影较好的口碑以及网络话题的引燃而迅速火热起来,上映 10 天,票房便突破 31 亿.偏置动态性的表示可以在用户偏置 b_u 中添加如下时间函数:

$$b(t)=b+\tau(t) \tag{26}$$

其中,b 是静态偏置,$\tau(t)$是一个时间函数.对 $\tau(t)$进行建模,常用方式是建立简单的时间线性模型,如 timeSVD++.然而,实际应用场景的表现通常是非线性的,甚至可能是无规则的,很难用确定的公式表示.

(2) 特征动态性.

对于属性 $x_i(1\leqslant i\leqslant n)$,可以细分成随时间变化的动态属性和保持稳定的静态属性,静态属性不需额外处理,动态属性则可添加时间变化函数,调整为

$$x_i(t)=x_i f(t) \tag{27}$$

其中,$f(\cdot)$是非线性激活函数.这些偏置函数和属性函数可以加入到 FM 模型中,对其静态偏置和属性进行调整.

5.2 视觉和非视觉属性融合

传统推荐方法与模型主要关注非视觉文本特征及其交互,如用户与物品的固有特征描述、物品的星级评分、用户的购买历史、书签、浏览日志、查询模式、鼠标活动等.随着机器视觉领域中深度学习的广泛应用,图像特征开始被关注,高维视觉特征也作为预测/推荐模型属性的一部分.视觉特征容易获得且描述准确,因此,如何提取高维视觉特征以及如何把非视觉低维特征和视觉高维特征进行有效融合,成为目前推荐领域的研究热点[92−94],并在 RecSys 2017 会议中作为重要主题列出.

deep CNN 模型最近被成功应用于对象检测、图像匹配等领域,相关研究已证明:基于海量数据训练的 deep CNN 模型可以精确应用于其他数据集,在新的数据集中仍然能够产生很好的效果.假设 f_i 表示物品 i 的原始视觉特征向量,维度为 D 维(可以通过 AlexNet,ResNet 等预先训练好,比如取 AlexNet 的第 2 个全连接层 FC7 的输出作为原始视觉特征向量),那么视觉特征 θ_i 可以按如下方式建模:

$$\theta_i=\boldsymbol{E}^T f_i \tag{28}$$

其中,E 是一个维度为 $D\times F$ 矩阵.此时,θ_i 的特征维度为 $F(F<<D)$,从而达到降维的目的.两个物品之间的视觉关系可以表示为

$$\theta_{i,j}=\boldsymbol{E}^T(f_i-f_j) \tag{29}$$

这种低秩嵌入方法仅仅能捕获两个物品是否关联,关联的原因则不能表达.实际应用中,物品之间的关联关系可能体现为多种原因,如一件 T-恤和一条短裙搭配合适的原因可能是颜色、质地或者款式等.为了解决这个问题,可以考虑采用多重嵌入,两个物品之间的关系可以表示为

$$\theta_{i,j}=\boldsymbol{E}_0^T f_i-\boldsymbol{E}_k^T f_j \tag{30}$$

其中,E_0 把物品 i 对应到一个参考点,$\boldsymbol{E}_0^T f_i$ 对应到一个嵌入空间,$E_k(k=1,2,\ldots,N)$表示与物品 j 的潜在匹配.传统的 FM 主要面向文本等非视觉特征,如果把视觉特征也融入到 FM 模型中,这样 FM 模型的应用将更加广泛.

6 结束语

在预测/推荐系统领域,FM 模型被广泛研究与应用.没有万能的模型,不同业务场景对模型的输入特征、处理逻辑和输出类别会有不同要求.本文从宽度扩展和深度扩展视角对 FM 模型及其变体的研究进行综述,希望能够提供不同的思路,为应用提供不同的选择方案.从对国内外高水平期刊及会议上的文献分析可以看出:将传统 FM 模型与深度学习模型相结合、将视觉与非视觉特征进行融合是研究热点.现有相关工作可以从以下几点进行深入:(1) 研究非视觉特征和视觉特征的融合,目前的相关研究缺乏对用户行为和物品变化趋势的差异化和细粒度处理;(2) 宽度和深度两种学习方式的融合目前主要用于增加推荐的多样性,其耦合方式以及对精度和效率的影响还需进一步研究.

References:

[1] Bobadilla J, Ortega F, Hernando A, Gutierrez A. Recommender systems survey. Knowledge-Based Systems, 2013,46(1):109−132.

[2] Uribe G, Carlos A, Hunt N. The netflix recommender system: algorithms, business value, and innovation. ACM Trans. on Management Information Systems, 2016,6(4):1−19.

[3] Covington P, Adams J, Sargin E. Deep neural networks for youtube recommendations. In: Proc. of the Conf. on Recommender Systems. New York: ACM Press, 2016. 191−198.

[4] Okura S, Tagami Y, Ono S, Tajima A. Embedding-based news recommendation for millions of users. In: Proc. of the 23rd ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2017. 1933−1942.

[5] Kazai G, Yusof I, Clarke D. Personalised news and blog recommendations based on user location, facebook and Twitter user profiling. In: Proc. of the 39th ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2016. 1129−1132.

[6] Huang L, Lin CJ, He J, Liu HY, Du XY. Diversified mobile app recommendation combining topic model and collaborative filtering. Ruan Jian Xue Bao/Journal of Software, 2017,28(3):708−720 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5163.htm [doi: 10.13328/j.cnki.jos.005163]

[7] Cheng HT, Koc L, Harmsen J, Shaked T, Chandra T, Aradhye H, Anderson G, Corrado G, Chai W, Ispir M, Anil R, Haque Z, Hong LC, Jain V, Liu XB, Shah H. Wide & deep learning for recommender systems. In: Proc. of the 1st Workshop on Deep Learning for Recommender Systems. New York: ACM Press, 2016. 7−10.

[8] Yan CR, Zhang QL, Zhao X, Huang YF. Method of bayesian probabilistic matrix factorization based on generalized Gaussian distribution. Journal of Computer Research and Development, 2016,53(12):2793−2800 (in Chinese with English abstract).

[9] Zhao QB, Zhou GX, Zhang LQ, Cichocki A, Amari S. Bayesian robust tensor factorization for incomplete multiway data. IEEE Trans. on Neural Networks and Learning Systems, 2016,27(4):736−748.

[10] Koren Y. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In: Proc. of the 14th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2008. 426−434.

[11] Koren Y. Collaborative filtering with temporal dynamics. In: Proc. of the 15th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2009. 447−456.

[12] Rendle S. Factorization machines. In: Proc. of the 10th IEEE Int'l Conf. on Data Mining. Piscataway: IEEE, 2010. 995−1000.

[13] Juan YC, Zhuang Y, Chin WS, Lin CJ. Field-aware factorization machines for CTR prediction. In: Proc. of the 10th ACM Conf. on Recommender Systems. New York: ACM Press, 2016. 43−50.

[14] Ta A. Factorization machines with follow-the-regularized-leader for CTR prediction in display advertising. In: Proc. of the 2015 IEEE Int'l Conf. on Big Data. Piscataway: IEEE, 2015. 2889−2891.

[15] Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J. GroupLens: An open architecture for collaborative filtering of netnews. In: Proc. of the ACM Conf. on Computer Supported Cooperative Work. New York: ACM Press, 1994. 175−186.

[16] Herlocker JL, Konstan JA, Borchers A. An algorithmic framework for performing collaborative filtering. In: Proc. of the 22nd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 1999. 230−237.

[17] Meng XW, Liu SD, Zhang YJ, Hu X. Research on social recommender systems. Ruan Jian Xue Bao/Journal of Software, 2015, 26(6):1356−1372 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4831.htm [doi: 10.13328/j.cnki.jos.004831]

[18] Meng XW, Chen C, Zhang YJ. A survey of mobile news recommend techniques and applications. Chinese Journal of Computers, 2016,39(4):685−703 (in Chinese with English abstract).

[19] Hong H, Pradhan B, Sameen MI, Chen W, Xu C. Spatial prediction of rotational landslide using geographically weighted regression, logistic regression, and support vector machine models in Xing Guo area (China). Geomatics, Natural Hazards and Risk, 2017,8(2):1997−2022.

[20] Chang YW, Hsieh CJ, Chang KW, Ringgaard M, Lin CJ. Training and testing low-degree polynomial data mappings via linear SVM. Journal of Machine Learning Research, 2010,11(4):1471−1490.

[21] Rendle S. Factorization machines with libFM. ACM Trans. on Intelligent Systems and Technology, 2012,3(3):57.

[22] Knoll J. Recommending with higher-order factorization machines. In: Proc. of the Int'l Conf. on Innovative Techniques and Applications of Artificial Intelligence. Berlin: Springer-Verlag, 2016. 103−116.

[23] Blondel M, Fujino A, Ueda N, Ueda N, Ishihata M. Higher-order factorization machines. In: Proc. of the 30th Conf. on Neural Information Processing Systems. Berkeley: USENIX, 2016. 3351−3359.

[24] Prillo S. An elementary view on factorization machines. In: Proc. of the 11st ACM Conf. on Recommender Systems. New York: ACM Press, 2017. 179−183.

[25] Knoll J, Köckritz D, Groß R. Markov random walk vs. higher-order factorization machines: A comparison of state-of-the-art recommender algorithms. In: Proc. of the Int'l Conf. on Innovations for Community Services. Berlin: Springer-Verlag, 2017. 87−103.

[26] Yurochkin M, Nguyen XL. Multi-way interacting regression via factorization machines. In: Proc. of the 31st Conf. on Neural Information Processing Systems. Berkeley: USENIX, 2017. 2595−2603.

[27] Pan J, Xu J, Ruiz AL, Zhao W, Pan S, Sun Y, Lu Q. Field-weighted factorization machines for click-through rate prediction in display advertising. In: Proc. of the 2018 World Wide Web Conf. on World Wide Web. Berlin: Springer-Verlag, 2018. 1349−1357.

[28] Zhao Y, Mansouri K, Yang Y, Mi ZQ. Rating prediction using category weight factorization machine in bigdata environment. In: Proc. of the IEEE Int'l Conf. on Communication Workshop. Piscataway: IEEE, 2015. 1909−1913.

[29] Wang S, Li C, Zhao K, Chen H. Learning to context-aware recommend with hierarchical factorization machines. Information Sciences, 2017,409:121−138.

[30] Guo R, Alvari H, Shakaria P. Strongly hierarchical factorization machines and anova kernel regression. In: Proc. of the 2018 SIAM Int'l Conf. on Data Mining. Philadelphia: SIAM, 2018. 729−737.

[31] Oentaryo RJ, Lim EP, Low JW, Lo D. Predicting response in mobile advertising with hierarchical importance-aware factorization machine. In: Proc. of the 7th ACM Int'l Conf. on Web Search and Data Mining. New York: ACM Press, 2014. 123−132.

[32] Yuan FJ, Guo GB, Jose JM, Chen L, Yu H, Zhang W. BoostFM: Boosted factorization machines for top-*n* feature-based recommendation. In: Proc. of the 22nd Int'l Conf. on Intelligent User Interfaces. New York: ACM Press, 2017. 45−54.

[33] Yan P, Zhou X, Duan Y. E-Commerce item recommendation based on field-aware factorization machine. In: Proc. of the 2015 Int'l ACM Recommender Systems Challenge. New York: ACM Press, 2015..

[34] Hong L, Doumith AS, Davison BD. Co-Factorization machines: Modeling user interests and predicting individual decisions in Twitter. In: Proc. of the 6th ACM Int'l Conf. on Web Search and Data Mining. New York: ACM Press, 2013. 557−566.

[35] Leksin V, Ostapets A. Job recommendation based on factorization machine and topic modelling. In: Proc. of the Recommender Systems Challenge. New York: ACM Press, 2016..

[36] Blondel M, Niculae V, Otsuka T, Ueda N. Multi-Output polynomial networks and factorization machines. In: Proc. of the 31st Conf. on Neural Information Processing Systems. Berkeley: USENIX, 2017. 3351−3361.

[37] Wang S, Du C, Zhao K, Li C, Li Y, Zheng Y, Wang Z, Chen H. Random partition factorization machines for context-aware recommendations. In: Proc. of the Int'l Conf. on Web-Age Information Management. Berlin: Springer-Verlag, 2016. 219−230.

[38] Pijnenburg M, Kowalczyk W. Extending logistic regression models with factorization machines. In: Proc. of the Int'l Symp. on Methodologies for Intelligent Systems. Berlin: Springer-Verlag, 2017. 323−332.

[39] Loni B, Said A, Larson M, Hanjalic A. 'Free lunch' enhancement for collaborative filtering with factorization machines. In: Proc. of the 8th ACM Conf. on Recommender Systems. New York: ACM Press, 2014. 281−284.

[40] Cheng C, Xia F, Zhang T, King L, Lyu M. Gradient boosting factorization machines. In: Proc. of the 8th ACM Conf. on Recommender Systems. New York: ACM Press, 2014. 265−272.

[41] Xu J, Lin K, Tan PN, Zhou J. Synergies that matter: Efficient interaction selection via sparse factorization machine. In: Proc. of the 2016 SIAM Int'l Conf. on Data Mining. Philadelphia: SIAM, 2016. 108−116.

[42] Selsaas LR, Agrawal B, Rong C, Wiktorski T. AFFM: Auto feature engineering in field-aware factorization machines for predictive analytics. In: Proc. of the IEEE Int'l Conf. on Data Mining Workshop. Piscataway: IEEE, 2015. 1705−1709.

[43] Punjabi S, Bhatt P. Robust factorization machines for user response prediction. In: Proc. of the 2018 World Wide Web Conf. on World Wide Web. Berlin: Springer-Verlag, 2018. 669−678.

[44] Lu CT, He L, Shao W, Cao B, Yu PS. Multilinear factorization machines for multi-task multi-view learning. In: Proc. of the 10th ACM Int'l Conf. on Web Search and Data Mining. New York: ACM Press, 2017. 701−709.

[45] Liu C, Zhang T, Zhao P, Zhou J, Sun JL. Locally linear factorization machines. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. Menlo Park: AAAI, 2017. 2294−2300.

[46] Yan CR, Zhang QL, Zhao X, Huang YF. An intelligent field-aware factorization machine mode. In: Proc. of the Int'l Conf. on Database Systems for Advanced Applications. Berlin: Springer-Verlag, 2017. 309−323.

[47] Ding Y, Wang D, Xin X, Li GQ, Sun D, Zeng XZ. SCFM: Social and crowdsourcing factorization machines for recommendation. Applied Soft Computing, 2018,66:548−556.

[48] Zhou J, Wang D, Ding Y, Yin L. SocialFM: A social recommender system with factorization machines. In: Proc. of the Int'l Conf. on Web-Age Information Management. Berlin: Springer-Verlag, 2016. 286−297.

[49] Rendle S. Social network and click-through prediction with factorization machines. In: Proc. of the KDD-Cup Workshop. New York: ACM Press, 2012. 113.

[50] Chen CM, Chen HP, Tsai MF, Yang YH. Leverage item popularity and recommendation quality via cost-sensitive factorization machines. In: Proc. of the 2014 IEEE Int'l Conf. on Data Mining Workshop. Piscataway: IEEE 2014. 1158−1162.

[51] Qiang R, Liang F, Yang J. Exploiting ranking factorization machines for microblog retrieval. In: Proc. of the 22nd ACM Int'l Conf. on Information & Knowledge Management. New York: ACM Press, 2013. 1783−1788.

[52] Xu Y, Tang Q, Hou LZ, Li M. Decision model for market of performing arts with factorization machine. Journal of Shanghai Jiaotong University (Science), 2018,23(1):74−84.

[53] Juan Y, Lefortier D, Chapelle O. Field-Aware factorization machines in a real-world online advertising system. In: Proc. of the 26th Int'l Conf. on World Wide Web Companion. Berlin: Springer-Verlag, 2017. 680−688.

[54] Chen C, Hou C, Xiao J, Yuan XJ. Purchase behavior prediction in e-commerce with factorization machines. IEICE Trans. on Information and Systems, 2016,99(1):270−274.

[55] Wang Y, Shang W, Li Z. The application of factorization machines in user behavior prediction. In: Proc. of the 15th Int'l Conf. on Computer and Information Science. Piscataway: IEEE, 2016. 1−4.

[56] Cao B, Shi M, Liu XF, Liu JX, Tang Md. Using relational topic model and factorization machines to recommend Web apis for mashup creation. In: Proc. of the Asia-Pacific Services Computing Conf. Berlin: Springer-Verlag, 2016. 391−407.

[57] Wu Y, Xie F, Chen L, Chen C, Zheng Z. An embedding based factorization machine approach for Web service qos prediction. In: Proc. of the Int'l Conf. on Service-Oriented Computing. Berlin: Springer-Verlag, 2017. 272−286.

[58] Tang MD, Zhang TT, Yang YT, Zheng ZB, Cao BQ. QoS-Aware Web service recommendation based on factorization machines. Chinese Journal of Computers, 2018,41(6):1300−1313 (in Chinese with English abstract).

[59] Chen C, Wu D, Hou CY, Yuan XJ. Exploiting social media for stock market prediction with factorization machine. In: Proc. of the 2014 IEEE/WIC/ACM Int'l Joint Conf. on Web Intelligence and Intelligent Agent Technologies. Piscataway: IEEE, 2014. 142−149.

[60] Yamada M, Lian W, Goyal A, Chen JH, Wimalaweane H, Khan SA. Convex factorization machine for toxicogenomics prediction. In: Proc. of the 23rd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2017. 1215−1224.

[61] Zhu G, Li L. Factorization machine based business credit scoring by leveraging internet data. In: Proc. of the Asia-Pacific Web Technologies and Applications. Berlin: Springer-Verlag, 2016. 565−569.

[62] Sun LJ, Fan JF, Yang WQ, Shi YH. Application of factorization machine in mobile App recommendation based on deep packet inspections. Journal of Computer Applications, 2016,36(2):307−310 (in Chinese with English abstract).

[63] Zhu M, Aggarwal CC, Ma S. Outlier detection in sparse data with factorization machines. In: Proc. of the 2017 ACM Conf. on Information and Knowledge Management. New York: ACM Press, 2017. 817−826.

[64] Zheng L, Noroozi V, Yu PS. Joint deep modeling of users and items using reviews for recommendation. In: Proc. of the 10th ACM Int'l Conf. on Web Search and Data Mining. New York: ACM Press, 2017. 425−434.

[65] Chen J, Sun B, Li H, Lu HT, Hua XS. Deep CTR prediction in display advertising. In: Proc. of the 24th ACM Int'l Conf. on Multimedia. New York: ACM Press, 2016. 811−820.

[66] Bracher C, Heinz S, Vollgraf R. Fashion DNA: Merging content and sales data for recommendation and article mapping. In: Proc. of the 22nd ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, Fashion Workshop. New York: ACM Press, 2016.

[67] Zhang S, Yao L, Sun A. Deep learning based recommender system: A survey and new perspectives. ACM Journal on Computing and Cultural Heritage, 2017,1(1):35.

[68] Razavian AS, Azizpour H, Sullivan J, Carlsson S. CNN features off-the-shelf: An astounding baseline for recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2014. 806−813.

[69] Wang R, Fu B, Fu G, Wang ML. Deep & cross network for ad click predictions. In: Proc. of the 17th Knowledge Discovery and Data Mining. New York: ACM Press, 2017.

[70] Zhang W, Du T, Wang J. Deep learning over multi-field categorical data. In: Proc. of the European Conf. on Information Retrieval. Berlin: Springer-Verlag, 2016. 45−57.

[71] Qu Y, Cai H, Ren K, Zhang WN, Yu Y. Product-Based neural networks for user response prediction. In: Proc. of the 2016 IEEE 16th Int'l Conf. on Data Mining. Piscataway: IEEE, 2016. 1149−1154.

[72] Shan Y, Hoens TR, Jiao J, Wang HJ, Yu D, Mao JC. Deep crossing: Web-scale modeling without manually crafted combinatorial features. In: Proc. of the 22nd Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2016. 255−262.

[73] Guo H, Tang R, Ye Y, Li ZG, He XQ. DeepFM: A factorization-machine based neural network for CTR prediction. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. Berlin: Springer-Verlag, 2017. 1725−1731.

[74] He X, Chua TS. Neural factorization machines for sparse predictive analytics. In: Proc. of the 40th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2017. 355−364.

[75] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016. 770−778.

[76] Bayer I. FastFM: A library for factorization machines. Journal of Machine Learning Research, 2016,17:1−5.

[77] Rendle S. Scaling factorization machines to relational data. In: Proc. of the Very Large Data Bases Conf. Endowment. Trondheim: VLDB, 2013,6(5):337−348.

[78] Blondel M, Fujino A, Ueda N. Convex factorization machines. In: Proc. of the Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer-Verlag, 2015. 19−35.

[79] Yuan F, Guo G, Jose JM, Chen L, Yu HT. Optimizing factorization machines for top-*n* context-aware recommendations. In: Proc. of the Web Information Systems Engineering (WISE 2016). Berlin: Springer-Verlag, 2016. 278−293.

[80] Pan Z, Chen E, Liu Q, Xu T, Ma HP, Lin HJ. Sparse factorization machines for click-through rate prediction. In: Proc. of the 2016 IEEE 16th Int'l Conf. on Data Mining. Piscataway: IEEE, 2016. 400−409.

[81] Saha A, Acharya A, Ravindran B, Ghosh J. Nonparametric poisson factorization machine. In: Proc. of the 2015 IEEE Int'l Conf. on Data Mining. Piscataway: IEEE, 2015. 967−972.

[82] Nguyen TV, Karatzoglou A, Baltrunas L. Gaussian process factorization machines for context-aware recommendations. In: Proc. of the 37th Int'l ACM SIGIR Conf. on Research & Development in Information Retrieval. New York: ACM Press, 2014. 63−72.

[83] Huang X, Yang Y, Bao X. Grid-Based Gaussian processes factorization machine for recommender systems. In: Proc. of the 9th Int'l Conf. on Machine Learning and Computing. New York: ACM Press, 2017. 92−97.

[84] Rendle S, Fetterly D, Shekita EJ, Su B. Robust large-scale machine learning in the cloud. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2016. 1125−1134.

[85] Zhou L, Pan S, Wang J, Vasilakos AV. Machine learning on big data: Opportunities and challenges. Neurocomputing. 2017,237: 350−361.

[86] Sun H, Wang W, Shi Z. Parallel factorization machine recommended algorithm based on mapreduce. In: Proc. of the 2014 10th Int'l Conf. on Semantics, Knowledge and Grids. Piscataway: IEEE, 2014. 120−123.

[87] Li M, Liu Z, Smola AJ, Wang YX. Difacto: Distributed factorization machines. In: Proc. of the 9th ACM Int'l Conf. on Web Search and Data Mining. New York: ACM Press, 2016. 377−386.

[88] Zhong E, Shi Y, Liu N, Rajan SJ. Scaling factorization machines with parameter server. In: Proc. of the 25th ACM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2016. 1583−1592.

[89] Li M, Andersen DG, Park JW, Smola AJ, Ahmed A. Scaling distributed machine learning with the parameter server. In: Proc. of the 11th USENIX Symp. on Operating Systems Design and Implementation. Berkeley: USENIX, 2014. 583−598.

[90] He R, Mcauley J. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In: Proc. of the 25th Int'l Conf. on World Wide Web. Berlin: Springer-Verlag, 2016. 507−517.

[91] Wang Y, Ouyang H, Deng H, Chang Y. Learning online trends for interactive query auto-completion. IEEE Trans. on Knowledge and Data Engineering, 2017,29(11):2442−2454.

[92] Saha A, Raykar VC, Khapra M. Joint multi-modal representations for e-commerce catalog search driven by visual attributes. In: Proc. of the 22nd ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2016. 13−17.

[93] He R, McAuley J. VBPR: Visual bayesian personalized ranking from implicit feedback. In: Proc. of the 30th AAAI Conf. on Artificial Intelligence. Menlo Park: AAAI, 2016. 144−150.

[94] He R, Fang C, Wang Z, AcAuley J. Vista: A visually, socially, and temporally-aware model for artistic recommendation. In: Proc. of the 10th ACM Conf. on Recommender Systems. New York: ACM Press, 2016. 309−316.

附中文参考文献:

[6] 黄璐,林川杰,何军,刘红岩,杜小勇.融合主题模型和协同过滤的多样化移动应用推荐.软件学报,2017,28(3):708−720. http://www.jos.org.cn/1000-9825/5163.htm [doi: 10.13328/j.cnki.jos.005163]

[8] 燕彩蓉,张青龙,赵雪,黄永锋.基于广义高斯分布的贝叶斯概率矩阵分解方法.计算机研究与发展,2016,53(12):2793−2800.

[17] 孟祥武,刘树栋,张玉洁,胡勋.社会化推荐系统研究.软件学报,2015,26(6):1356−1372. http://www.jos.org.cn/1000-9825/4831.htm [doi: 10.13328/j.cnki.jos.004831]

[18] 孟祥武,陈诚,张玉洁.移动新闻推荐技术及其应用研究综述.计算机学报,2016,39(4):685−703.

[58] 唐明董,张婷婷,杨亚涛,郑子彬,曹步清.基于因子分解机的质量感知 Web 服务推荐方法.计算机学报,2018,41(6):1300−1313.

[62] 孙良君,范剑锋,杨婉琪,史颖欢.因子分解机算法在基于深度数据包检测的手机应用推荐中的应用.计算机应用,2016,36(2): 307−310.

燕彩蓉(1978－),女,湖北仙桃人,博士,副教授,CCF 专业会员,主要研究领域为云计算,大数据,机器学习.

张青龙(1990－),男,博士生,主要研究领域为推荐系统,机器学习.

周灵杰(1994－),男,学士,主要研究领域为图像处理,推荐算法,深度学习.

李晓林(1973－),男,博士,教授,博士生导师,主要研究领域为深度学习,云计算,大数据.