文章编号:1007-5321(2021)05-0127-06 DOI:10.13190/j.jbupt.2021-017

多头自注意力在双曲空间下的点击率预测

韩越林, 王小玉
(哈尔滨理工大学 计算机科学与技术学院, 哈尔滨 150080)

摘要: 在推荐系统中,了解用户行为背后的复杂功能交互,对预测用户点击广告或商品的概率至关重要. 人们一直努力寻找稀疏和高维原始特征的低维表示形式及有意义的组合. 其中深度交叉网络可以显式地在每一层进行特征交叉,但其"一视同仁"地对待所有交叉特征,未考虑不同特征对结果的影响,造成一些有用信息被消除. 因此提出了多头自注意力神经网络在双曲空间下的点击率预测模型. 在双曲空间下,模型不再使用内积而使用洛伦兹距离违背三角不等式程度来度量特征之间的相似性与相关性,从而避免了维度灾难. 实验表明,就模型准确性而言,其在点击率预测数据集上均优于深度交叉网络.

关 键 词: 双曲空间; 多头自注意力; 洛伦兹距离
中图分类号: TP181 **文献标志码**: A

Click-Through Rate Prediction of Multi-Head Self-Attention in Hyperbolic Space

HAN Yue-lin, WANG Xiao-yu
(School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China)

Abstract: In recommendation systems, understanding the complex functional interactions behind user behaviors is crucial to predict the clicking probability of users on advertisements or commodities. Efforts have been made to find low-dimensional representations and meaningful combinations of sparse and high-dimensional original features. Among them, the deep & cross network can explicitly cross features at each layer. However it treats all crossing features "equally" and does not consider the influence of different features on the results, which may eliminate some useful information. Therefore, a prediction model of click-through rate of multi-head self-attention neural network in hyperbolic space is proposed. In hyperbolic space, the model uses Lorentzian distance instead of inner product, to measure the similarity and correlation between features, which can avoid dimension disaster. Experimental results show that the model is superior to the deep & cross network on predicting click-through rate data sets in terms of accuracy.

Key words: hyperbolic space; multi-head self-attention; Lorentzian distance

在推荐系统中,点击率(CTR, click-through-rate)预测是一个关键问题. 随着数据量和数据维度的迅速增多,深度神经网络成为推荐系统的主流.

1 研究现状与创新点

1.1 研究现状

对于 CTR 预测所依赖的数据而言,有以下 2 个

收稿日期: 2021-01-27
基金项目: 国家自然科学基金项目(60572153,60972127)
作者简介: 韩越林(1995—), 男, 硕士生.
通信作者: 王小玉(1971—), 女, 教授, 硕士生导师, E-mail: wangxiaoyu@hrbust.edu.cn.

问题：① 输入数据具有稀疏和高维[1]的特性. 用户浏览数据和商品属性通常是离散的；②正如文献[2]所示,高阶特征相互作用对于预测性能至关重要. 枚举所有的高阶特征,让机器去学习是不可能实现的. 需要寻找稀疏和高维输入特征的低维表示,以及对特征组合的不同顺序进行建模. 例如,因式分解机(FM,factorization machines)[3]将多项式回归模型与因式分解相结合,但其受多项式拟合时间的限制,仅能获取低阶特征.

针对上述问题,近年来已有基于深度神经网络的模型[2]来对高阶特征交互进行建模. 其中,全连接前馈神经网络(属于深度神经网络的分支)通常用来捕获高阶特征交互. 但是,这种方法在学习乘法特征交互时效率低下[4]；而且,由于其采用隐式方式学习特征交互,缺乏特征组合的合理解释.

Vaswani 等[5]提出了自注意力(包括单头自注意力和多头自注意力),使自注意力成为热点. 传统的自注意力,使用内积来度量特征间的相似性与相关性. Hsieh 等[6]提出,在度量特征相似性与相关性方面,使用欧式空间距离比使用内积更合理,研究特征交互也可以从几何意义出发.

1.2 创新点

由于欧式空间表达能力较弱,提出了采用表达能力更强的双曲空间. 这是因为双曲空间的表达能力强于欧式空间,相同的上界在双曲空间中嵌入的容量大于欧式空间.

根据 Hsieh 等[6]的观点,对多头自注意力机制进行了改进并做出了创新:在多头自注意力中首次使用洛伦兹距离；利用违背三角不等式的程度来度量特征间的相似性与相关性. 在相似性与相关性度量中,采用洛伦兹距离,借助双曲空间下该距离可以违背三角不等式的特殊性(使用不等式符号来刻画双曲空间中点之间的成对关系),避免了使用向量内积可能出现的维度灾难.

2 模型结构

双曲空间下多头自注意力深度神经网络模型的结构如图1所示. 数据通过输入层进入嵌入层,在嵌入层中数据编码映射到低维空间中进行降维. 降维后的数据进入并行的深度网络和多头自注意力网络,进行筛选. 高阶组合特征[5]由注意力筛选. 筛选规则:在同一自注意力层,任意2个特征都能与双曲空间的原点构成三角形,用违背三角不等式的程度来度量2个特征之间的相似性与相关性. 将单个注意力层扩展到多个,相同子空间中的不同特征进行交互,不同子空间交互是随机的. 通过对不同特征所占权重的重新分配,突出强相关特征,弱化不相关特征. 多个交互层之间进行堆叠,即可捕捉更高阶的组合特征. 组合层将2个网络合并后输出.

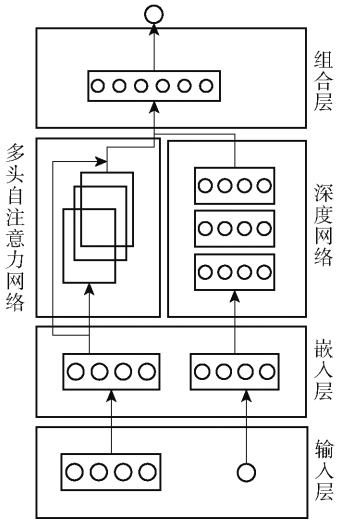


图1 双曲空间下多头自注意力深度神经网络模型

2.1 嵌入层

在CTR预测中,输入特征多为分类特征,例如“颜色”“形状”. 这样的特征通常被编码为独热向量,直接使用会导致数据稀疏和空间利用率极低. 嵌入层是一个查找表,用于将稀疏特征投影到双曲空间中转换成低维密集向量(即输入特征).

2.2 多头自注意力网络

2.2.1 双曲空间与洛伦兹距离

双曲几何旨在研究具有恒定负曲率的非欧式空间. HyperML[7](双曲度量学习)中提到由于双曲空间的指数扩展性质,在双曲空间中移动一点经过一定距离所需要的力比欧式空间中小得多. 在双曲几何中,平面的曲率是负数,与欧式空间有一些不同的性质.

双曲空间中向量 $\boldsymbol{a}=[a_0,a_1,\cdots,a_n]$,$\boldsymbol{b}=[b_0,b_1,\cdots,b_n]$,其中 n 为双曲空间维度,$\boldsymbol{a}$ 和 $\boldsymbol{b}$ 的洛伦兹内积定义为

$$\langle \boldsymbol{a},\boldsymbol{b}\rangle_{\mathrm{L}}=-a_0b_0+\sum_{i=1}^{n}a_ib_i \tag{1}$$

双曲空间 $H^{n,\beta}$ 中的向量符合:

$$H^{n,\beta}=\{\boldsymbol{z}\in\mathbf{R}^{n+1}:\|\boldsymbol{z}\|_{\mathrm{L}}^{2}=-\beta,z_0\geqslant\beta\} \tag{2}$$

其中:原点 G 的向量定义为 $\boldsymbol{g}=(\beta,0,0,\cdots,0)$,当

$\beta=1$ 时，该空间称为单位双曲面空间. 原点向量 $\boldsymbol{g}$ 与向量 $\boldsymbol{y}$ 的关系 $\langle \boldsymbol{g},\boldsymbol{y}\rangle = -z_0 \leqslant \beta$，$\| \boldsymbol{z} \|_{\mathrm{L}}^2$ 为向量 $\boldsymbol{z}$ 的洛伦兹范数，对于每个向量 $\boldsymbol{z}=[z_0,z_1,\cdots,z_n]$ 中的第 1 个分量为

$$z_0 = \sqrt{\beta + \sum_{i=1}^{n} z_i^2} \tag{3}$$

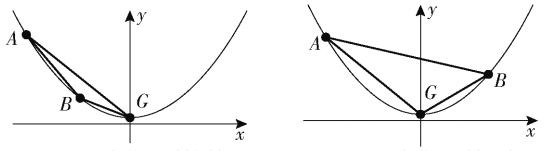
$\boldsymbol{a},\boldsymbol{b} \in H^{n,1}$ 间的洛伦兹距离为

$$d_{\mathrm{L}}^2(\boldsymbol{a},\boldsymbol{b}) = \| \boldsymbol{a}-\boldsymbol{b} \|_{\mathrm{L}}^2 = -2-2\langle \boldsymbol{a},\boldsymbol{b}\rangle_{\mathrm{L}} \tag{4}$$

三角不等式是具有正定黎曼度量的最关键的几何性质之一. 它指出，对于任意 3 个点 X,Y 和 Z 能形成的三角形中，任何两条边的长度之和应大于或等于第 3 条边的长度：

$$d(X,Y) \leqslant d(X,Z) + d(Z,Y) \tag{5}$$

在双曲空间中，因黎曼度量为负数，洛伦兹距离在一定情况下会违背三角不等式. 例如原点 G 和任意两点 A 和 B 形成的三角形，A,B 两点在 y 轴的同一侧时，如图 2(a)所示，洛伦兹距离遵守三角不等式；A,B 两点在 y 轴的两侧且相距很远时，如图 2(b)所示，则洛伦兹距离违背三角不等式.



(a) A,B两点位于y轴同侧　　(b) A,B两点位于y轴两侧

图 2　曲面模型遵守三角不等式情况

2.2.2　自注意力结构

引入查询-键值自注意力[8]来确定哪些特征组合是有意义的. 查询-键值自注意力结构如图 3 所示，其中 $\boldsymbol{Q}$ 为查询矩阵，$\boldsymbol{K}$ 为键矩阵，$\boldsymbol{V}$ 为值矩阵.

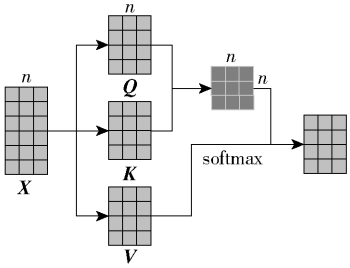


图 3　查询-键值自注意力结构

1) 输入特征矩阵 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]$，$\boldsymbol{X}$ 通过矩阵乘法线性映射为注意力空间内的矩阵：

$$\boldsymbol{Q} = \boldsymbol{W}_{\mathrm{q}}\boldsymbol{X} \tag{6}$$

$$\boldsymbol{K} = \boldsymbol{W}_{\mathrm{k}}\boldsymbol{X} \tag{7}$$

$$\boldsymbol{V} = \boldsymbol{W}_{\mathrm{v}}\boldsymbol{X} \tag{8}$$

其中 $\boldsymbol{W}_{\mathrm{q}},\boldsymbol{W}_{\mathrm{k}},\boldsymbol{W}_{\mathrm{v}}$ 分别为 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 的转换矩阵.

以特征 m 为例，将特征 m 的查询矩阵 $\boldsymbol{Q}_m$ 和所有特征的键矩阵 $\boldsymbol{K}$ 的相似性与相关性作为权重(特征 m 和特征 p 投影到双曲空间的 S,T 两点与双曲空间原点 G，三点共同组成的三角形，该三角形违背三角不等式的程度作为权重)，归一化权重后与所有特征的值矩阵 $\boldsymbol{V}$ 进行加权求和，得出与特征 m 相关的高阶特征.

2) 特征 m 和特征 p 相似性与相关性计算公式为

$$\phi(\boldsymbol{Q}_m,\boldsymbol{K}_p) = \frac{d_{\mathrm{L}}^2(\boldsymbol{Q}_m,\boldsymbol{K}_p) - d_{\mathrm{L}}^2(\boldsymbol{Q}_m,\boldsymbol{g}) - d_{\mathrm{L}}^2(\boldsymbol{K}_p,\boldsymbol{g})}{\langle \boldsymbol{g},\boldsymbol{Q}_m\rangle_{\mathrm{L}}\langle \boldsymbol{g},\boldsymbol{K}_p\rangle_{\mathrm{L}}} \tag{9}$$

其中：$\boldsymbol{g}=(1,0,0,\cdots,0)$，为双曲空间的原点向量；分子表示双曲空间下洛伦兹距离违背三角不等式的程度；分母用于标度目标函数，使得函数的区间被限制在一定的范围内，从而避免维度灾难.

3) 使用函数 softmax 进行归一化，特征 m 和特征 p 归一化后的相似性与相关性为

$$a_{m,p} = \frac{\exp(\phi(\boldsymbol{Q}_m,\boldsymbol{K}_p))}{\sum_{i=1}^{n}\exp(\phi(\boldsymbol{Q}_m,\boldsymbol{K}_i))} \tag{10}$$

4) 加权求和产生与特征 m 相关的高阶特征为

$$\hat{\boldsymbol{x}}_m = \sum_{k=1}^{n} a_{m,i}\boldsymbol{V}_k \tag{11}$$

其中 $\boldsymbol{V}_k$ 为第 $k(k\leqslant n)$ 个特征的值矩阵，由式(8)定义.

2.2.3　多头自注意力

将原来单个自注意力结构扩展为 h 个，使用时需要从小到大调试 h 的值，使其达到最优值.

特征 m 的最终高阶特征 $\widetilde{\boldsymbol{X}}_m$ 定义为

$$\widetilde{\boldsymbol{X}}_m = \hat{\boldsymbol{x}}_m^{(1)} \oplus \cdots \oplus \hat{\boldsymbol{x}}_m^{(i)} \oplus \cdots \oplus \hat{\boldsymbol{x}}_m^{(h)} \tag{12}$$

其中：i 为第 $i(i\leqslant h)$ 个注意力结构，$\oplus$ 为合并操作.

残差结构可以保留部分初始特征，减缓过拟合. $\widetilde{\boldsymbol{X}}_m$ 经过残差处理得出最终残差高阶特征为

$$\boldsymbol{X}_m^{\mathrm{res}} = \mathrm{ReLU}(\widetilde{\boldsymbol{X}}_m + \boldsymbol{W}^{\mathrm{res}}\boldsymbol{X}_m) \tag{13}$$

其中：$\boldsymbol{X}_m$ 为特征 m 的输入特征，$\boldsymbol{W}^{\mathrm{res}}$ 为权重矩阵，ReLU 为激活函数.

将所有 n 个特征的结果合并，得出最高阶组合特征：

$$\boldsymbol{X}^{\mathrm{res}} = \boldsymbol{X}_1^{\mathrm{res}} \oplus \boldsymbol{X}_2^{\mathrm{res}} \oplus \cdots \oplus \boldsymbol{X}_n^{\mathrm{res}} \tag{14}$$

2.3 深度网络

深度网络是一个全连接的前馈神经网络,共有 k 层隐藏层,其中每个隐藏层的特征计算公式为

$$\boldsymbol{h}_{i+1} = \mathrm{ReLU}(\boldsymbol{W}_i \boldsymbol{h}_i + \boldsymbol{b}_i) \tag{15}$$

其中 $\boldsymbol{W}_i$ 和 $\boldsymbol{b}_i$ 为第 i 个隐藏层的参数.

2.4 组合层

组合层将多头自注意力网络与深度网络的输出连接起来,得出预测值为

$$p_i = \sigma(\boldsymbol{W}_x \boldsymbol{X}^{\mathrm{res}} + \boldsymbol{W}_h \boldsymbol{h}_k) \tag{16}$$

其中:$\boldsymbol{W}_x$ 和 $\boldsymbol{W}_h$ 为组合层的权重向量;$\sigma(\cdot)$ 为 sigmoid 函数,$\sigma(x) = 1/[1 + \exp(-x)]$.

3 模型分析及高阶组合特征

对所提模型中高阶特征进行分析. 假设 y_1, y_2, y_3 和 y_4 分别表示4个特征. 在任意一个自注意力层中,每个特征都能与任意其他特征进行交互.

在第一个自注意力层中,与 y_1 相关的特征权重高. 所有二阶组合特征(如 $f(y_1, y_2), f(y_1, y_3)$)可以在第一个自注意力层得到,即可提取出权重最高的二阶组合特征 $e_1^{\mathrm{res}}(f(y_1, y_2))$.

证明可以在多个自注意力层中对高阶特征进行进一步交互. 第一个自注意力层生成的组合特征 e_1^{res} 和第二个自注意力层生成的组合特征 $e_2^{\mathrm{res}}(f(y_2, y_3))$,通过允许 e_1^{res} 参加 e_2^{res} 建模,就涉及到 y_1, y_2, y_3 的三阶组合特征,e_1^{res} 包含特征 y_1 和 y_2,e_2^{res} 包含特征 y_2 和 y_3. 同理,第一个自注意力层生成的组合特征 e_1^{res} 和第三个自注意力层生成的组合特征 $e_3^{\mathrm{res}}(f(y_3, y_4))$,可通过允许 e_1^{res} 参加 e_3^{res} 建模,就涉及到 y_1, y_2, y_3 和 y_4 的四阶组合特征. 依此类推,多头自注意力就能获取高阶组合特征.

4 模型评估与结论

4.1 数据集

Criteo 数据集[9]用来预测广告的 CTR,含 26 个分类特征和 13 个整数特征,数据按时间排序. KDD12 数据集[10]由 KDDCup 2012 发布. 由于工作重点是 CTR 预测,将是否点击视为一个二分类问题.

4.2 对比模型

4.2.1 仅使用单个特征的线性方法

逻辑回归[11](LR, logistic regression)模型仅模拟原始特征的线性组合.

4.2.2 基于因式分解机的方法——浅层模型

因式分解机[3](FM, factorization machine)模型使用分解技术来建模二阶组合特征相互作用. 注意力因式分解机[12](AFM, attentional factorization machine)模型使用注意力机制来区分二阶组合特征的不同重要性来扩展 FM.

4.2.3 高阶特征交互的方法——深度模型

深交叉[13](DC, deepcrossing)模型有残差结构的全连接神经网络通过隐式学习非线性交互信息. 神经网络因式分解机[1](NFM, neural factorization machine)模型在二阶组合特征交互层之上堆叠了深度神经网络. 深度交叉网络[14](DCN, deep & cross network)模型中的跨网络交叉是该模型的核心,在层次上采用级联特征向量的外部乘积来显式建模特征交互. 基于自注意力神经网络的自动特征交互学习[15](AutoInt, automatic feature interaction learning via self-attentive neural networks)模型,通过多头自注意力机制显式构造高阶特征.

4.3 模型的研究与使用

4.3.1 在 Criteo 数据集下超参数对模型的影响

使用交叉熵和接受者操作特性曲线下面积(AUC, area under curve)2个评估指标来评估模型. 交叉熵损失函数(以下简称交叉熵)定义为

$$L = -\frac{1}{N}\sum_{i=1}^{N} u_i \mathrm{lb}(p_i) + (1-u_i)\mathrm{lb}(1-p_i) \tag{17}$$

其中:p_i 为根据式(16)计算出的标签,u_i 为真实标签,N 是批次数据的数量. 交叉熵数值越小模型越优秀. AUC 在0.5和1.0之间. AUC 越接近1.0,模型越优秀.

1) 激活函数

根据文献[16]可知,激活函数 ReLU 和 tanh 比激活函数 sigmoid 更适合于深度模型. 6种模型采用2种激活函数条件下的交叉熵和 AUC 对比,如图4所示. 由图可知,对于所有深度模型,采用 ReLU 激活函数均比使用 tanh 函数有更好的表现.

2) 数据批量标准化与 Dropout 比较

Dropout 通过忽略部分神经单元减少过拟合现象. 在使用数据批量标准化(BN, batch normalization)之后,Dropout[17]并不能改善所有模型的性能. 比较使用 BN 的模型和 Dropout 为50%的模型的交叉熵和 AUC,结果如图5所示. 由图5可知,对于所提出的模型,使用 BN 比 Dropout 更有效.

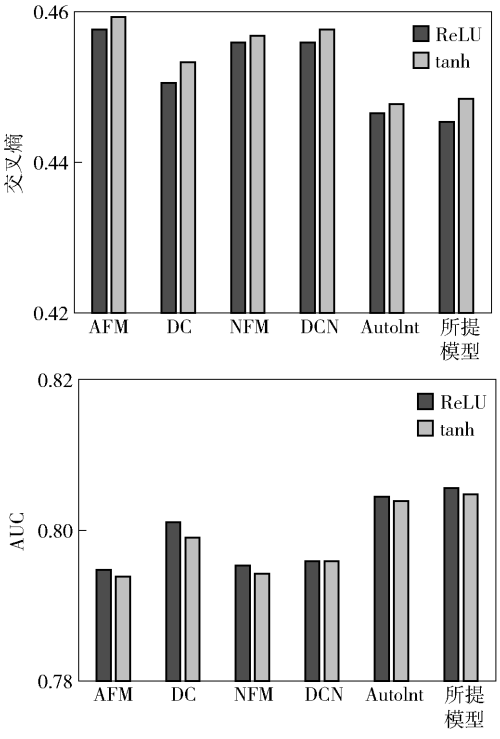


图 4　采用不同激活函数对模型性能的影响

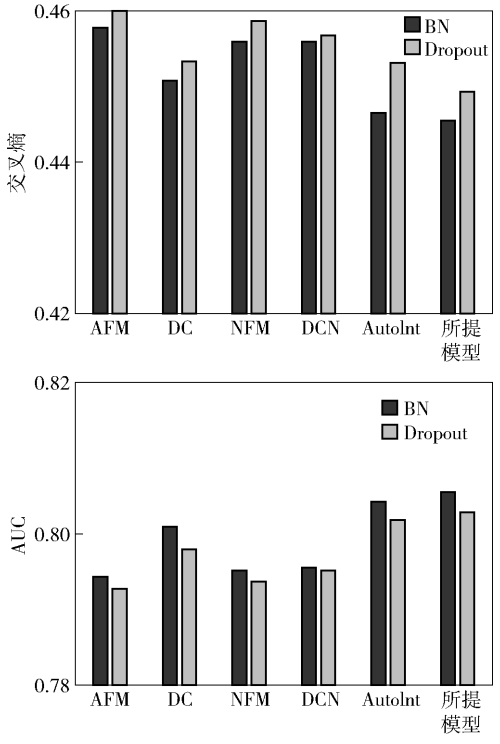


图 5　使用 BN 和 Dropout 对模型性能的影响

4.3.2　超参数的使用

显卡中统一计算设备架构的一个并行单元有 32 个线程，设置 32 的倍数效率高. 当嵌入层尺寸过小时，会损失过多的原始信息；过大时，训练时间指数增加，综合考虑，设置为 16 较为合适. 使用 BN，激活函数使用 ReLU. 优化器 Adam 对批次数据随机优化，批此数据大小默认为 512.

4.4　模型评估

4.4.1　不同数据集上的性能对比

使用 3 个评估指标来评估模型：AUC、交叉熵和均方根误差（RMSE，root mean squared error）. 均方根误差通过模型的估计值与真实值的偏差来计算，偏差数值越小，模型性能越好.

由表 1 和表 2 可知，所提模型在 2 个数据集上交叉熵和 RMSE 最小，AUC 最大，性能最好.

表 1　Criteo 数据集下不同模型的性能对比

模型	交叉熵	AUC	RMSE
LR	0.470 5	0.781 2	0.388 3
FM	0.459 1	0.790 1	0.383 9
AFM	0.457 8	0.794 8	0.382 6
DC	0.450 9	0.801 2	0.375 6
NFM	0.456 2	0.795 7	0.381 5
DCN	0.455 7	0.796 3	0.381 7
AutoInt	0.446 6	0.804 7	0.375 3
所提模型	0.445 6	0.805 9	0.373 6

表 2　KDD12 数据集下不同模型的性能对比

模型	交叉熵	AUC	RMSE
LR	0.168 4	0.736 2	0.400 7
FM	0.157 3	0.775 8	0.397 4
AFM	0.156 9	0.776 5	0.397 0
DC	0.158 7	0.772 0	0.397 2
NFM	0.163 0	0.751 7	0.398 9
DCN	0.156 3	0.777 3	0.396 6
AutoInt	0.155 4	0.784 7	0.387 9
所提模型	0.155 0	0.786 2	0.387 1

4.4.2　消融实验

修改或删除网络模型的部分结构来评估网络模型是否有效. a 代表双曲空间下使用点积来度量特征间相似性与相关性的网络；b 代表欧式空间下使用点积来度量特征间相似性与相关性的网络. 消融实验结果如表 3 所示. 由表 3 可知，提出的网络模型为有效模型.

表3 消融实验对模型的评估

数据集	模型	交叉熵	AUC
Criteo	a	0.447 0	0.804 2
	b	0.446 4	0.804 9
	所提模型	0.445 6	0.805 9
KDD12	a	0.155 2	0.785 3
	b	0.155 3	0.785 0
	所提模型	0.155 0	0.786 2

5 结束语

主要研究了双曲空间下多头自注意力模型,结合了双曲空间、洛伦兹距离,其中双曲空间扩展了表达能力,在考虑几何意义下使用洛伦兹距离的性质.避免了使用内积造成的维度灾难.

模型评估表明,所提模型优于其他对比模型.在模型评估中,添加了注意力机制的AFM性能要优于NFM,所提模型性能优于DCN和AFM,表明多头自注意力在推荐系统中行之有效.

参考文献:

[1] He Xiangnan, Chua Tat-Seng. Neural factorization machines for sparse predictive analytics[C]// SIGIR'17: The 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Tokyo: ACM, 2017: 355-364.

[2] Cheng Heng-Tze, Koc L, Harmsen J, et al. Wide & deep learning for recommender systems[C]// DLRS 2016: The 1st Workshop on Deep Learning for Recommendder Systems. Boston: ACM, 2016: 7-10.

[3] Rendle S. Factorization machines[C]//2010 IEEE International Conference on Data Mining. Sydney: IEEE, 2010: 995-1000.

[4] Beutel A, Covington P, Jain S, et al. Latent cross: making use of context in recurrent recommender systems[C]// WSDM'18: The Eleventh ACM International Conference on Web Search and Data Mining. Los Angeles: ACM, 2018: 46-54.

[5] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]// NIPS'17: The 31st International Conference on Neural Information Processing Systems. California: NIPS, 2017: 5998-6008.

[6] Hsieh C, Yang Longqi, Yin Cui, et al. Collaborative metric learning[C]// WWW'17: The 26th International Conference on World Wide Web. Switzerland: Geneva, 2017: 193-201.

[7] Tran L, Tay Y, Zhang Shuai, et al. HyperML: a boosting metric learning approach in hyperbolic space for recommender systems[C]// WSDM'20: The 13th International Conference on Web Search and Data Mining. New York: ACM, 2020: 609-617.

[8] Miller A, Fisch A, Dodge J, et al. Key-value memory networks for directly reading documents[C]// EMNLP 2016: Conference on Empirical Methods in Natural Language Processing. Austin: ACL, 2016: 1400-1409.

[9] Criteo. Criteo_dataset[EB/OL]. (2019-10-09)[2021-10-25]. https://www.kaggle.com/mrkmakr/criteo-dataset.

[10] Special Interest Group Knowledge Discovery and Data Mining. KDD cup 2012, track 2[EB/OL]. (2012-02-20)[2021-10-25]. https://www.kaggle.com/c/kdd-cup2012-track2/data.

[11] Kleinbaum D, Dietz K, Gail M, et al. Logistic regression[M]. New York: Springer-Verlag, 2002: 536-537.

[12] Xiao Jun, Hao Ye, He Xiangnan, et al. Attentional factorization machines: learning the weight of feature interactions via attention networks[C]// IJCAI'17: The 26th International Joint Conference on Artificial Intelligence. Melbourne: IJCAI, 2017: 3119-3125.

[13] Ying Shan, Hoens T, Jiao Jian, et al. Deep crossing: web-scale modeling without manually crafted cominatorial features[C]// KDD'16: The 22nd ACM SIGKDD International Conference on Know-ledge Discovery and Data Mining. San Francisco: ACM, 2016: 255-262.

[14] Wang Ruoxi, Fu Bin, Fu Gang, et al. Deep & cross network for ad click predictions[C]// ADKDD'17: The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2017: 1-7.

[15] Song Weiping, Shi Chence, Xiao Zhiping, et al. Autoint: automatic feature interaction learning via selfattentive neural networks[C]// CIKM'19: The 28th ACM International Conference on Information and Knowledge Management. New York: ACM, 2019: 1161-1170.

[16] Qu Yanru, Cai Han, Ren Kan, et al. Product-based neural networks for user response prediction[C]//2016 IEEE 16th International Conference on Data Mining. Barcelona: IEEE, 2016: 1149-1154.

[17] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Machine Learning Research, 2014, 15(1): 1929-1958.