计算机科学
COMPUTER SCIENCE

基于组合结构的逻辑回归点击预测算法

郭尚志, 廖晓峰, 鲜开义

引用本文

郭尚志, 廖晓峰, 鲜开义. 基于组合结构的逻辑回归点击预测算法[J]. 计算机科学, 2024, 51(2): 73-78.

GUO Shangzhi, LIAO Xiaofeng, XIAN Kaiyi. Logical Regression Click Prediction Algorithm Based on Combination Structure [J]. Computer Science, 2024, 51(2): 73-78.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230100052

基于组合结构的逻辑回归点击预测算法

郭尚志　廖晓峰　鲜开义
重庆大学计算机学院　重庆 400030

摘　要　随着互联网和广告平台的飞速发展,面对海量的广告信息,为了提升用户点击率,提出一种改进的基于组合结构的逻辑回归点击预测算法 LRCS(Logical Regression of Combination Structure)。该算法基于不同类别特征广告受众可能不同的特点,首先,采用 FM 进行特征组合,产生两类组合特征;其次,将一类特征组合作为聚类算法的输入进行聚类;最后,将另一类特征组合输入由聚类产生的分段 GBDT+逻辑回归组合的模型中进行预测。在两个公开数据集中进行了多角度验证,结果表明与其他几类常用的点击预测算法相比,LRCS 在点击预测上有一定的性能提升。

关键词:逻辑回归;特征组合;聚类;组合推荐;人工智能;智能制造
中图分类号　TP391

Logical Regression Click Prediction Algorithm Based on Combination Structure

GUO Shangzhi, LIAO Xiaofeng and XIAN Kaiyi
College of Computer Science, Chongqing University, Chongqing 400030, China

Abstract　With the rapid development of the Internet and advertising platforms, in the face of massive advertising information, in order to improve the user click rate, an improved logical regression click prediction algorithm, logical regression of combination structure(LRCS) based on composite structure is proposed. The algorithm is based on different types of features, which may have different audiences. First, FM is used to combine features to generate two types of combined features. Secondly, a kind of feature combination is used as clustering algorithm for clustering. Finally, another type of feature combination is input into the segmented GBDT+logical regression combination model generated by clustering for prediction. Through multi angle verification in two public datasets, and compared with other commonly used click prediction algorithms, it shows that LRCS has a certain performance improvement in click prediction.

Keywords　Logical regression, Feature combination, Clustering, Combination recommendation, Artificial intelligence, Intelligent manufacturing

随着互联网的飞速发展,人们的生活越来越便利,广告平台也有了长足的进步。为了提升用户的点击率,相关算法经历了贝叶斯分类器[1]、SVM[2]、FM、逻辑回归算法[3]、GBDT+逻辑回归[4],深度学习[5]、强化学习[6]等逐步演变的过程。广告平台点击数据,可以直接表达为曝光和点击,经过简单的归约,把点击作为正样本,把未点击作为负样本,就可以根据一段时间的数据样本训练一个分类器。形式上,假设广告为 A,上下文为 C,非点击和点击分类为 E,则每个样本均可以看作一个$\langle A,C|E\rangle$的元组,其中 E 的值只有 0 和 1 两种。对于每一个 Advert 和 Context 的组合$\langle A,C\rangle$,需要一个模型来对其进行分类,即可区分点击与非点击。如果对于每个$\langle A,C\rangle$可以预测一个 CTR(0,1)[7],即可考虑为一个逻辑回归问题。但是,提升点击预测的能力也需要考虑多个方面,主要原因有:用户群体的差异性,其对广告的喜好不同;广告本身正样本占比很低,用户点击的概率也很低;当前的大部分算法模型都需要对原始数据进行处理,一般采用独热编码、连续数据分箱、归一化等处理才能使用,而处理后的数据变得极其稀疏。处理这些数据一般采用关联分析、降维、组合特征、特征嵌入等方式[8],把稀疏的高维特征转化为较低维的低维稠密向量来表示,以提升算法的能力[9]。组合特征主要是对特征值进行充分的融合,为了提升特征的综合能力,因子分解机(Factorization Machine,FM)[10]通过对特征进行两两组合,提升特征的表达能力;梯度提升树(Gradient Boosting Decision Tree,GBDT)通过联合多个决策树来混合特征,让特征充分混合;深度学习兴起以后,深度因子分解机(Deep Factorization Machine,DeepFM)[11]被提出,但是 DeepFM 只考虑了特征的深度融合,没有考虑一些重要特征单独对算法的影响;随后,深度交叉网络(Deep&Cross,D&C)[12]和深宽网络(Deep & Wide)[13]等模型相继被提出。提升模型的能力可以从两个方面入手:一方面是提升特征的组合能力,对于重要的特征需要特别考虑,将其无差别地直接输入特征组合,会忽视一些重要特征的表达能力,如对男性和女性无差别地推送广告,是无法提升用户的点击率的;另一方面是提升算法本身的能力,通过组合不同类型算法来提升算法的预测性能。

到稿日期:2023-01-10　返修日期:2023-05-22
通信作者:郭尚志(921919001@qq.com)

近年来，广告越来越向个性化方向发展。为了更好地迎合客户喜好，提升点击率，出现了划分用户与广告域的方法。本文基于广告的受众可能不同的特点，希望通过组合特征进行聚类[14]，发现不同用户或广告的分类特点，再进行点击预测，提升用户的点击率和粘合度。为此，提出一种基于组合结构的逻辑回归点击预测算法(LRCS)。

本文的主要贡献如下：

1)提出 LRCS 模型。通过 FM 对数据特征进行混合，原始特征经独热编码后，特征维度极高，数据变得更加稀疏，该方式一方面能使原始特征充分融合，另一方面可以起到一定的降维作用。

2)采用分类的特征处理方式与自动聚类。通过 FM 计算两类嵌入特征集合，一类较粗粒度地输入聚类算法进行分类，另一类细粒度地输入至根据聚类算法产生的聚类数决定的分段 GBDT 与逻辑回归组成算法进行预测。FM 是两两特征融合，在 GBDT 中可以让特征再次充分融合，最终由逻辑回归给出预测值。

3)验证模型性能。通过在公开的数据集上与其他算法模型进行对比，验证了所提模型的性能。

本文第 1 章是相关工作介绍；第 2 章给出 LRCS 模型说明；第 3 章介绍 LRCS 算法的具体步骤；第 4 章进行实验对比分析；最后总结全文。

1 相关工作

为了提升模型的预测能力，发现特征的组合关系至关重要。但特征的人工组合费时费力，人们更希望特征可以自动相互组合，并能够发现特征组合的最佳表达[15]。例如，FM 模型及后来的 FFM，都采用自动组合的方式；为了区分前面提到的各特征在预测模型中的重要性，提出基于注意力因子分解机(Attentional Factorization Machine，AFM)[16]、基于积的神经网络(Product based Neural Networks，PNN)[17]等模型；随着深度学习网络的快速发展，尤其是卷积神经网在图片识别中大放异彩，文献[18]提出基于卷积神经网络的特征生成模型，通过卷积层组合成新特征；为了考虑一阶特征的重要性，将深度神经网络与一阶常规算法相结合，提出神经因子分解机(Neural Factorization Machine，NFM)[19]、深宽模型(Deep&Wide)等模型。

LRCS 采用基于聚类的方法来产生特征组合，主要是基于历史行为信息，为物或人找到偏好相似的最近邻对象，以“人以群分”或“物以类聚”的思想，对用户或物进行分类[20]。聚类可以从不同维度展开研究，对于点击广告来说，一是从用户角度进行聚类，把相似的用户分类到一起；二是从广告本身考虑，将不同的广告分类到一起；三是同时从用户和广告考虑，把相似用户喜好的广告类型聚合到一起；四是基于子空间的方法[21]。文献[22]提出一种可扩展的子空间聚类算法，该算法以得分阀值决定用户的分类，对用户感兴趣的内容进行分类，忽略用户的低评分内容，在用户的子空间中寻找相关用户并提供推荐。这种方式一定程度上会提供精准的推荐，但是很难为一些冷门的广告或者刚上架的广告提供推荐。文献[23]提出了一种基于用户喜好寻找近邻用户的算法，该算法把用户的喜好划分为喜欢和不喜欢两类，通过构造用户近邻树来提供推荐能力，通过分层来建立用户的相关兴趣远近，但该算法同样也没有考虑上面提到的问题。

模型集成是提升性能的常用方法之一，即合并多个机器学习模型来构建更强大模型的方法[24]。已证明有两种集成模型对大量分类和回归的数据集都是有效的，二者都以决策树为基础，分别是随机森林(Random Forest)[25]和梯度提升决策树(Gradient Boosted Decision Tree，GBDT)。随机森林主要解决单棵决策树过拟合的问题，通过训练多决策树来减少这种过拟合，同时又保证预测的性能。在回归的场景一下，求多棵树的平均值来决定最终的预测结果，在分类的情况下采用投票机制，所有树的结果中哪个类型得票多则预测为该类型。梯度提升回归树，通过合并多棵决策树来构建一个更为强大的模型。梯度上升树同样可以用于回归，也可以用于分类分析。与随机森林生成决策树的方式不同，梯度提升树采用逐步提高性能的方式来创建新的树，每生成一棵新树都是对前一棵树的残差进行拟合。梯度提升树可以控制树的多少和拟合树的深度，方便进行计算，性能较好。

LRCS 采用不同类型的横型混合集成。模型集成学习也可以采用不同类型的算法进行混合，如 Meta 公司提出的 GBDT+逻辑回归的方法，对特征充分混合，也可以灵活地调整树的深度、棵数，并解决逻辑回归只考虑各个独立特征的问题。而逻辑回归作为最终的输出，由于算法实现简单，被广泛应用于工业界，其分类时计算量极小，速度快，存储资源少，可解释性好，方便观察样本概率或分类[26]。

2 LRCS 模型

2.1 问题描述

给定一个样本数据集，每个样本(U,A,C)表示用户 U 对广告 A 的点击，样本数据中包含标签特征 $[L_1,L_2,\cdots,L_n]$，n 表示标签特征向量的维数。本文考虑个性化的特征分类，划分聚类数，通过组合算法提升预测点击率。

2.2 系统模型

如图 1 所示，本文提出的算法模型包括 3 个部分：第一部分为嵌入层，利用 FM 处理原始特征域，通过特征组合形成组合特征，将其提供至上层使用；第二部分利用 K-medoids 聚类算法对数据分簇聚类；第三部分利用 GBDT+逻辑回归算法进行点击预测。

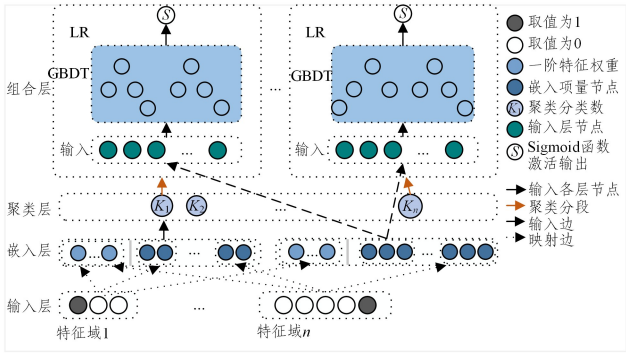


图 1 系统模型

Fig. 1 System model

2.3 嵌入层

在点击预测的方法中，可以从用户角度来考虑，以用户的年龄、性别、爱好、职业等特征来划分，也可以从广告内容、类别、位置等特征来划分，或者同时从两个角度来划分特征。每一个特征即为一个域，单个域常采用独热编码。由于编码后的数据极其稀疏，同时逻辑回归无差别地考虑所有项，没有考虑各特征组合，因此本文采用 FM。其基本思想是：将式(1)中的 w_{ij} 改写为式(2)的隐向量方式[27]，与 SVD 有异曲同工之妙。在可控的精度条件下，通过有限的 K 维向量即可表达原始高维数据，同时起到降维的作用；把二阶的权重组合分解为两个矩阵乘法，算法复杂度降至 $O(Kn)$。同时生成稠密向量集合，对没有出现的特征集合也可以进行计算和预测。嵌入层采用 FM 来组合特征进行预处理，对于输入聚类层，向量维度值 $C_i=5$，对于输入组合层，向量维度值 $D_i=25$。

$$f(x)=\boldsymbol{w}_0+\sum_{i=1}^{n}\boldsymbol{w}_i\boldsymbol{x}_i+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\boldsymbol{w}_{ij}\boldsymbol{x}_i\boldsymbol{x}_j \tag{1}$$

其中，n 代表样本的特征数量；$\boldsymbol{x}_i$ 是第 i 个特征的值；$\boldsymbol{w}_0$，$\boldsymbol{w}_i$，$\boldsymbol{w}_{ij}$ 是模型参数；只有当 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 都不为 0 时，交叉才有意义。在数据稀疏的情况下，满足交叉项不为 0 的样本非常少，当训练样本不足时，很容易导致参数训练不充分而不准确，最终影响模型的效果。常用的方式是交叉项采用矩阵分解来近似解决问题，最终公式如下：

$$f(x)=\boldsymbol{w}_0+\sum_{i=1}^{n}\boldsymbol{w}_i\boldsymbol{x}_i+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\langle v_i\cdot v_j\rangle\boldsymbol{x}_i\boldsymbol{x}_j \tag{2}$$

其中，$\langle v_i\cdot v_j\rangle=\sum_{c=1}^{C}v_i^c v_j^c$。

2.4 *K*-medoids 聚类层

K-medoids 聚类是一种改进的 K-means 算法。K-means 算法的主要问题是受到离群点或噪声的影响比较大，这是因为 K-means 更新中心的方式是计算簇内均值向量，离群点会极大地影响某属性列的均值计算，从而导致中心点偏离[28]。K-medoids 中心点算法解决了这个问题，其思想为：尝试在簇内再次计算到各点的距离，若替代后总代价减小，用非中心点替代中心点。更新的中心点总是簇内某样本，且用总代价衡量，中心点不会偏离簇[29]。聚类函数如下：

$$E=\sum_{i=1}^{k}\sum_{x\in\boldsymbol{C}_i}\|\boldsymbol{x}-\boldsymbol{m}_i\|_2^2 \tag{3}$$

其中，$\boldsymbol{C}_i$ 为第 i 类的集合，$\boldsymbol{m}_i$ 为第 i 类的簇心。聚类算法的最终目的是实现 E 代价最小，即 $\min(E)$。

2.5 基于 GBDT 算法与逻辑回归算法组合层

提升算法的预测性能，常用的方法之一就是集成多个弱的学习器。GDBT 采用先训练一棵决策树，后继的决策树根据前一棵决策树的残差来提升拟合能力，直到达到树的阈值 K，然后对 K 棵决策树进行加权计算，最终给出整个集成学习器的预测值。GBDT 的函数如下：

$$Y=\sum_{i=1}^{m}f_m(x) \tag{4}$$

GDBT 算法模型是个迭代多轮的加法模型。每轮迭代输出一个基模型 $f_m(x)$，其中 m 表示第 m 轮迭代。最终，每轮迭代输出的模型经过加法求和，得到最终 GBDT 模型的输出。整个模型采用平方误差的损失函数：

$$L(w)=\sum_{i=1}^{n}(\hat{y}_i-y_i)^2 \tag{5}$$

参数 w 为每个基模型的每个分裂特征和每个分裂阈值。对于损失函数中的预测值 y_i 有：

$$y_i=\sum_{m=1}^{M-1}f_m(x_i)+f_M(x_i) \tag{6}$$

将式(6)代入损失函数得：

$$L(w)=\sum_{i=1}^{n}(\hat{y}_i-\sum_{m=1}^{M-1}f_m(x_i)-f_M(x_i))^2 \tag{7}$$

若令：$r_i=y_i-\sum_{m=1}^{M-1}f_m(x_i)$，则损失函数最终表示为：

$$L(w)=\sum_{i=1}^{n}(r_i-f_M(x_i))^2 \tag{8}$$

在训练基模型时，其训练目标是去拟合残差 r_i，使损失函数达到最小。

逻辑回归使用 Sigmoid 函数来建模。Sigmoid 函数是一个 S 形的曲线，输入值远离 0 值时函数值会很快接近 0 或者 1，该特征对于解决二分类问题十分重要。该函数也具有很强的鲁棒性，同时函数的输入值范围$(\infty,-\infty)$映射到输出值$(0,1)$之间具有概率意义。例如：将一个样本输入到的函数中，输出 0.7，意味着该样本有 70%的概率为正例，30%的概率为负例。逻辑回归的函数形式如下：

$$f(x_i)=\frac{1}{1+e^{-(\boldsymbol{w}^{\mathrm{T}}x_i+\boldsymbol{b})}} \tag{9}$$

其中，$\boldsymbol{w}$ 和 $\boldsymbol{b}$ 为待求参数，$\boldsymbol{w}$ 为 n 维特征向量。

对于式(9)中的 $\boldsymbol{w}$ 参数，统计学中以极大似然估算来求解，即在这组参数下，x 的似然度最大。似然函数如下：

$$L(\boldsymbol{w})=\Pi g(x_i)^{y_i}(1-g(x_i))^{1-y_i} \tag{10}$$

由于式(10)中连乘函数不易求导，可对两边同时取对数将其转换为加法计算；为了防止采用梯度下降法求导累加值太大而造成梯度爆炸，可以除以样本总数 n，最终得到损失函数如下：

$$L(\boldsymbol{w})=\min\left(-\frac{1}{n}\sum_{i=1}^{n}y_i(\boldsymbol{w}^{\mathrm{T}}x_i+\boldsymbol{b})-\ln(e^{\boldsymbol{w}^{\mathrm{T}}x_i+b}+1)\right) \tag{11}$$

3 LRCS 算法步骤

基于第 2 章中提出的系统模型，给定任意目标数据 $u_i\in U$，若要为 u_i 预测其是否点击，首先通过 FM 进行向量嵌入，然后进行聚类，最后把 u_i 的高维向量输入至组合逻辑回归算法。最终公式如下：

$$f(x_i)=\lambda_i\frac{1}{1+e^{-(\boldsymbol{w}^{\mathrm{T}}\sum_{i=1}^{m}f_m(x_i)+\boldsymbol{b})}} \tag{12}$$

其中，$\lambda_i\in[k_1,k_2,\cdots,k_i]$，每条数据集只有一个 k_i 值为 1，其他值为 0，根据簇类算法训练 k_i 个组合逻辑回归算法。考虑特征充分融合，并降低特征维度，同时考虑点击用户个性化分类，提出基于组合结构的逻辑回归算法 LRCS。该算法由特征嵌入算法、特征聚类算法和组合逻辑回归算法组成。

3.1 嵌入层算法进行特征混合

特征嵌入：把原始的特征处理为独热编码，然后通过 FM 进行特征的两两融合，为聚类层和组合层提供指定维度的特征输出 $\boldsymbol{U}_{Ci}$ 和 $\boldsymbol{U}_{Di}$，为后面两部分算法提供输入。算法 1 给出了嵌入算法的具体过程。

算法 1 嵌入层算法

输入：特征集合 U_i，聚类层每个征维数 C_i，组合层每个特征维数 D_i，迭代次数 S_i

输出：聚类层特征向量集合 $\boldsymbol{U}_{Ci}$，组合层特征向量集合 $\boldsymbol{U}_{Di}$

1. $\boldsymbol{U}_i$集合处理为 one-hot 编码

2. M_i, N_i = shape($\mathbf{U}_i$)
3. For S_i do
4. For M_i do
5. 计算损失函数,更新参数
6. For N_i do
7. 更新参数
8. For C_i do 这里取 C_i 或者 D_i,隐向量维度
9. 更新 $\mathbf{U}_{ci}$ 或者 $\mathbf{U}_{Di}$
10. 输出特征向量集合 $\mathbf{U}_{ci}, \mathbf{U}_{Di}$
11. Return $\mathbf{U}_{ci}, \mathbf{U}_{Di}$

3.2 基于 *K*-medoids 聚类算法进行聚类

聚类层进行数据聚类。考虑到不同特征的用户或广告的受众可能不同,通过指定的聚类数 K_i 有效地把初始数据划分成不同的簇,最终得到划分好的特征簇 C_i。算法 2 给出了特征聚类算法的具体过程。

算法 2 用户特征聚类算法

输入:聚类层特征集合 $\mathbf{U}_{ci}$,聚类数 K_i
输出:特征划分簇 $\mathbf{C}_i$
1. 从 $\mathbf{U}_i$ 中随机选择 K_i 个用户作为初始均值向量 $\mathbf{M}_i$
2. Repeat
3. 初始化 $\mathbf{U}_i = \emptyset$
4. For shape($\mathbf{U}_i$) do
5. 计算样本 $\mathbf{U}_i$ 到各 $\mathbf{M}_i$ 的距离
6. 将样本 $\mathbf{U}_i$ 划入相应的簇 $\mathbf{C}_i$
7. 完成所有 $\mathbf{U}_i$ 归类
8. For K_i do
9. 计算 $\mathbf{M}_i$ 到本簇 $\mathbf{C}_i$ 中所有点的距离和 $dist_1$
10. For shape($\mathbf{C}_i$) do
11. 计算该簇中所有点至其他点的距离和 $dist_2$
12. If $dist_1 > dist_2$
13. 更新 $\mathbf{M}_i$ 为新聚类点
14. End if
15. Until $\mathbf{M}_i$ 不再变化
16. Return $\mathbf{C}_i$

3.3 组合层算法进行预测

组合层将 LGBM 算法[30]与逻辑回归算法组合,最终给出预测结果。LGBM 算法对嵌入输入的特征再次进行融合。该算法属于 GBDT 算法的一种,是由微软发布的轻量梯度提升树算法,最主要的特点是计算较快,支持回归和分类树的模型。LGBM 的训练结果输入至逻辑回归,以训练逻辑回归算法。算法 3 给出了组合逻辑回归算法的具体过程。

算法 3 组合逻辑回归算法

输入:特征向量 $\mathbf{U}_{Di}$
输出:用户点击概率集合 P_i
1. 划分集合为测试集与验证集
2. 调用 LGBM 算法
3. 将 LGMB 算法的输出做归一化处理,得到特征集合 $\mathbf{R}_{ci}$,
4. 输入 $\mathbf{R}_{ci}$ 至逻辑回归函数
5. 输出预测概率 P_i
6. Return P_i

4 实验评估

4.1 实验环境与数据集

实验环境如表 1 所列。

表 1 配置实验环境

Table 1 Experiment environment configuration

分类	配置内容	备注
操作系统	64 位 Windows11 操作系统	
CPU	Inter(R) Xeon(R) CPU E7-88674 core	
内存(RAM)	32GB	
存储	512GB 固态硬盘	
编程语言	Python3.9	
机器学习框架	Scikit-learn 1.1.0	默认

本文基于公开数据集 Criteo 和 iPinYou。首先说明实验数据集的设置,然后介绍对比算法。

Criteo 数据集是 2014 年 Kaggle 平台上发起的广告点击率预测大赛数据集。该数据集包含 4584 万条真实展示广告的点击反馈,为了保证正负样本的比例平衡,对样本进行了负采样,点击率约为 26%,数据集包含 13 个数值特征与 26 个分类特征。考虑到数据量较大,本数据在使用时采用 5 折交叉验证。

iPinYou 是 2013 年公开的真实广告数据集。样本包含广告曝光机会和点击日志,其中的每一个样本表示信息和用户的点击反馈,可以用来训练预测算法。特征包含用户类别标签、所在区域、城市、IP 地址、浏览器信息、广告宽度、高度、可见性、所属网站、广告商 ID 等特征。数据集包含 1950 万条数据记录,按日期倒排序。两个数据集均按 0.8∶0.1∶0.1 的比例拆分为训练集、测试集、验证集。

数据集统计信息如表 2 所列。

表 2 数据集信息

Table 2 Dataset information

		样本总数	正样本数	特征数	聚类层特征输入数	组合层特征输入数
Criteo	5 折交叉	45840616	11745438	39	234	1014
iPinYou	训练集	12316228	2463245	16	96	256
	测试集	1033085	615812	16	96	256

为了对 LRCS 进行评估,使用交叉熵损失(Logloss)和 AUC(Area Under the ROC Curve)作为模型的验证标准。AUC 和交叉熵损失是常用的二分类问题的评估方法,它们是两个相对的指标,AUC 相对越高、交叉熵损失越低,说明模型符合预期。AUC 的计算如式(13)所示:

$$AUC = \frac{\sum_{i=1}^{m} r_i - \frac{m(1+m)}{2}}{mn} \tag{13}$$

其中,r_i 代表第 i 个样本;m 和 n 代表正样本和负样本个数。Logloss 函数如下:

$$L(w) = \min\left(-\frac{1}{n}\sum_{i=1}^{n} y_i p_i - \ln(p_i + 1)\right) \tag{14}$$

其中,p_i 为第 i 个预测值。

为了验证 LRCS 的性能指标,采用以下几种算法与其进行对比。

1)LR:逻辑回归算法,利用原始特征进行逻辑回归分类,能够较好地学习到全局的偏差,使用简单,运算速度快。

2)FM:采用两两配对的方式,因子分解求得两个隐向量矩阵,可以有效地进行特征组合。

3)GBDT:可以有效融合数据各个特征,通过不断集成新的决策树拟合数据。这里采用 LGBM 算法,参数 $max\_depth = 4$, $num\_leaves = 40$, $n\_estimators = 200$。

4）GBDT＋逻辑回归（LR）：通过组合多种算法，提升预测效果。

5）Deep Network：包含多个隐层的前馈神经网络，隐层的节点与层数可以根据需要进行调整。层数为 2，节点数为 100。

4.2 实验对比

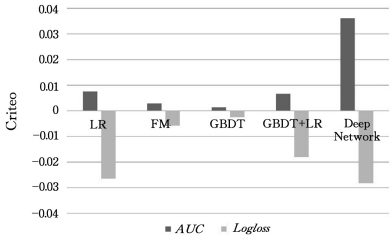
模型的实验性能对比如表 3 所列。从中可以得到如下结论：1）LR 由于只考虑各特征的独立性，没有进行特征融合，与其他算法相比性能表现相对较差；2）FM 模型和 GBDT 模型增强特征组合，效果比 GBDT＋LR 相对要好，这可能是组合过程中原始特征维度太过稀疏造成的；3）在 2 个数据集上，LRCS 与对比算法的 AUC 和 Logloss 得分都有提升，说明 LRCS 算法采用特征处理方法与算法组合可以提升模型的性能。

表 3 实验对比

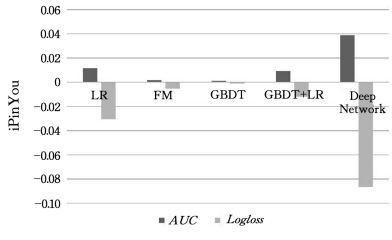
Table 3 Experimental comparison

模型	Criteo		iPinYou(10^{-2})	
	AUC	*Logloss*	*AUC*	*Logloss*
LR	0.7787	0.5694	0.7283	0.6194
FM	0.7823	0.5579	0.7356	0.6043
GBDT	0.7835	0.5561	0.7361	0.6108
GBDT＋LR	0.7794	0.5546	0.7301	0.6082
DeepNetwork	0.7563	0.5704	0.7082	0.6531
LRCS	**0.7846**	**0.5547**	**0.7369**	**0.6011**

图 2 给出了 LRCS 算法与其他算法的 AUC 提升和 Logloss 降低的对比。可以看出，LRCS 与效果最好的 GBDT 算法相比，在数据集 Criteo 上 AUC 提升了 0.14%，Logloss 下降了 0.25%；在数据集 iPinYou 上 AUC 提升了 0.11%，Logloss 下降了 0.12%。实验结果说明采用特征嵌入与聚类分类再组合的 LRCS 算法可以提升点击性能。



(a) Criteo dataset promotion



(b) iPinYou dataset promotion

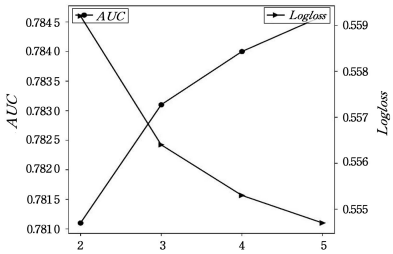
图 2 LRCS 与其他算法的对比

Fig. 2 Comparison between LRCS and other algorithms

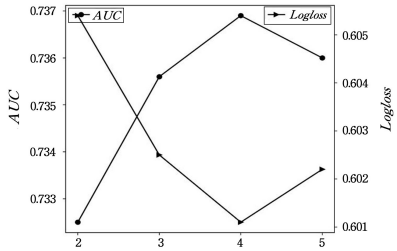
4.3 超参数分析

LRCS 算法中有两个非常重要的超参数：聚类层向量维度值 C_i 和输入组合层向量维度值 D_i。

聚类层向量维度值 C_i 如图 3 所示，从中可以看出在两个数据集中的性能变化。性能随着聚类数的增加而提升，在数据集 Criteo 上，当聚类数值为 5 时，取得最好成绩；在数据集 iPinYou 上，当聚类数值为 4 时，取得最好成绩。



(a) Number of Criteo clusters

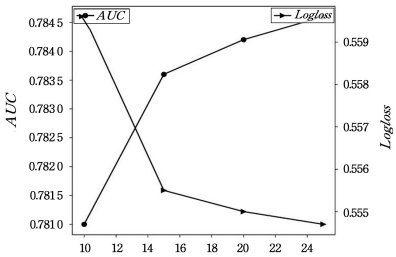


(b) Number of iPinYou clusters

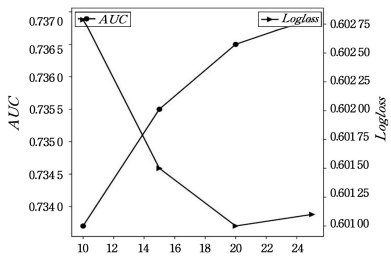
图 3 聚类层向量维度值 C_i

Fig. 3 Cluster layer vector dimension value C_i

输入组合层向量维度值 D_i 如图 4 所示。可以看出，性能随着维度数的增加而提升，在据集 Criteo 聚类数值为 25 时，取得最好成绩；在数据集 iPinYou 聚类数值为 25 时，AUC 取得好成绩，但是其 Logloss 有所上升，可能的原因是 iPinYou 正样本数据比例较低，原始特征数据比较稀疏。



(a) Number of Criteo dimensions



(b) Number of iPinYou dimensions

图 4 组合层向量维度值 D_i

Fig. 4 Combination layer vector dimension value D_i

结束语 本文提出一种基于组合结构的逻辑回归算法来预测点击，首先通过 FM 对数据特征进行混合，原始特征经独热编码后，特征维度极高，数据变得更加稀疏，通过 FM 融合后，特征得到较好的表示，同进也降低了特征的维度。通过 FM 计算两类嵌入特征集合，一类输入聚类算法进行分类，另一类高维嵌入特征输入至根据聚类个数而产生分段 GBDT 与逻辑回归组成的算法，在 GBDT 中可以让特征再次充分融合，最终由逻辑回归给出预测值。实验结果显示，预测性能得到提升，另外对嵌入特征维度也进了对比。

未来，在特征处理方面可以采用图嵌入等，也可以在框架上从深度学习及注意机制[31]方向进行研究。

参 考 文 献

[1] FOO L K, CHUA S L, IBRAHIM N. Attribute weighted naïve bayes classifier[J]. Computers, Materials & Continua, 2022, 71(1): 1945-1957.

[2] HU R, ZHU X, ZHU Y, et al. Robust SVM with adaptive graph learning[J]. World Wide Web, 2020, 23(3): 1945-1968.

[3] SHERWIN J S, CHARTIER J. Parameter optimization of logistic regression classifiers[J]. BMC Neuroscience, 2013, 14(1): 1-2.

[4] TIAN X, WANG J, WEN Y, et al. Multi-attribute scientific documents retrieval and ranking model based on GBDT and LR[J]. Math. Biosci. Eng., 2022, 19: 3748-3766.

[5] GHARIBSHAH Z, ZHU X, HAINLINE M. Deep learning for user interest and response prediction in online display advertising[J]. Data Science and Engineering, 2020, 5(1): 12-26.

[6] PI Q, BIAN W, ZHOU G, et al. Practice on long sequential user behavior modeling for click-through rate prediction[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 2671-2679.

[7] ZHOU G, ZHU X, SONG C, et al. Deep interest network for click-through rate prediction[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 1059-1068.

[8] HUANG Q, XU Y Y, CHEN Y, et al. An Adaptive Mechanism for Recommendation Algorithm Ensemble[J]. IEEE ACCESS, 2019, 7: 10331-10342.

[9] SEDLMAIR M, MUNZNER T, TORY M. Empirical guidance on scatterplot and dimension reduction technique choices[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12): 2634-2643.

[10] RENDLES. Factorization machines [C]// 2010 IEEE International Conference on Data Mining. IEEE, 2010: 995-1000.

[11] 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. 2016: 43-50.

[12] GUO H, TANG R, YE Y. DeepFM: A Factorization-Machine-based Neural Network for CTR Prediction[J]. Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 1725-1731.

[13] WANG R, FU B, FU G, et al. Deep & cross network for ad click predictions[C]// Proceedings of the ADKDD'17. 2017: 1-7.

[14] GÜNER S, CODAL K S, GECER H S, et al. Using k-means clustering algorithm in the identification of traffic accident patterns: the application of Sakarya province[J]. Journal of Business Science, 2018, 6(3): 89-105.

[15] ISHIKAWA T, YATA N, NAGAO T. Automatic Classification of Paper Using Combinational Optimization of Image Features [J]. Japan Tappi Journal, 2011, 65(6): 595-604.

[16] XIAO J, YE H, HE X, et al. Attentional factorization machines: Learning the weight of feature interactions via attention networks[J]. arXiv: 1708.04617, 2017.

[17] QU Y, CAIH R. Product-based neural networks for user response prediction[C]// Proceedings of the IEEE International Conference on DataMining. Barcelona, Spain, 2016: 6.

[18] 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.

[19] HE X, CHUA T S. Neural factorization machines for sparse predictive analytics[C]// Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017: 355-364.

[20] MOSKVICHEV O, NIKISHCHENKOV S, MOSKVICHEVA E. Optimization of production and transport infrastructure based on cluster analysis methods[C]// E3S Web of Conferences. EDP Sciences, 2020, 164: 03008.

[21] MARTÍNEZ-CEVALLOS D, PROAÑO-GRIJALVA A, ALGUACIL M, et al. Segmentation of participants in a sports event using cluster analysis[J]. Sustainability, 2020, 12(14): 5641.

[22] AGARWAL N, HAQUE E, LIUH, et al. Research paper recommender systems: A subspace clustering approach[C]// International Conference on Web-Age Information Management. Berlin, Heidelberg: Springer, 2005: 475-491.

[23] SUN X H, ZHANG L. Collaborative filtering recommendation algorithm based on scoring region subspace [J]. Computer Science, 2022, 49(7): 50-56.

[24] RISHICKESH R, SHAHINA A, NAYEEMULLA KHAN A. Predicting forest fires using supervised and ensemble machine learning algorithms[J]. International Journal Recent Technology Engineering, 2019, 8: 3697-3705.

[25] DÉSIR C, BERNARD S, PETITJEAN C, et al. One class random forests[J]. Pattern Recognition, 2013, 46(12): 3490-3506.

[26] VANI M S, RAJASHREE S. Forecast of Mobile Ad Click Through Logistic Regression Algorithm[J]. Journal of Innovation in Computer Science and Engineering, 2016, 6(1): 29-32.

[27] WANG S, SUN G, LI Y. SVD++ recommendation algorithm based on backtracking[J]. Information, 2020, 11(7): 369.

[28] JUNG H G. Medoid selection from sub-tree leaf nodes for k-medoid clustering-based hierarchical template tree construction [J]. Electronics Letters, 2013, 49(2): 108-109.

[29] BIJALWAN A, PUROHIT K C, MALIK P, et al. A Self-Adaptable Angular Based K-Medoid Clustering Scheme(SAACS) for Dynamic VANETs[J]. Electronics, 2022, 11(19): 3071.

[30] LI J, HUANG Y, QIAO M, et al. Effects of water soaked height on the deformation and crushing characteristics of loose gangue backfill material in solid backfill coal mining[J]. Processes, 2018, 6(6): 64.

[31] WANG Q, LIU F, ZHAO X, et al. Session interest model for CTR prediction based on self-attention mechanism[J]. Scientific Reports, 2022, 12(1): 1-13.

GUO Shangzhi, born in 1982, Ph.D, senior engineer. His main research interests include artificial intelligence and intelligent manufacturing, intelligent recommendation, machine learning, and big data applications.

（责任编辑：柯颖）