江苏省人工智能学会
JiangSu Association of Artificial Intelligence

HUAWEI

DIGIX 极客 | 算法精英大赛路演

用户人口属性预测

东风快递团队

1. 赛题理解

根据用户的手机使用习惯（用户属性、手机基本信息及特性、app安装、app使用记录、app所属类别等）构建模型，来预测用户所处的年龄段

2. 解题思路

- 目标是通过用户的特征来预测用户的年龄段,所以关键在于提取用户与年龄相关的特征.
- 主要是通过统计等方式提取用户的特征
- 对于序列化数据,如app激活列表,主要考虑使用word2vec等进行embedding
- 对于时序数据,则主要考虑使用lstm等来进行处理.

3. 特征工程

1. 基本特征

- 用户基本属性 / 手机基本信息 /

2. 统计特征

- 一个月内每个用户所有app使用时长和次数的统计量（最大值/均值/标准差/总和）
- 每个用户一周七天每天所有app使用时长和次数总和
- 一个月内每个用户使用的app总数的统计量（最大值/均值/标准差）
- 每个app类别下用户安装的app数量
- 每个app类别下用户使用的app数量
- 每个类别下安装的app到使用的app的转化率
- 每个类别下安装的app数量所占比例
- 每个类别下使用的app数量所占比例

3. 特征工程

3. Word2Vec 特征

- 对用户安装的app列表和使用的app列表，使用Word2Vec做Embedding
- 对每个用户得到的所有app的embedding进行求和/求平均，作为该用户使用的app列表的向量表示

4. LSTM 时序特征

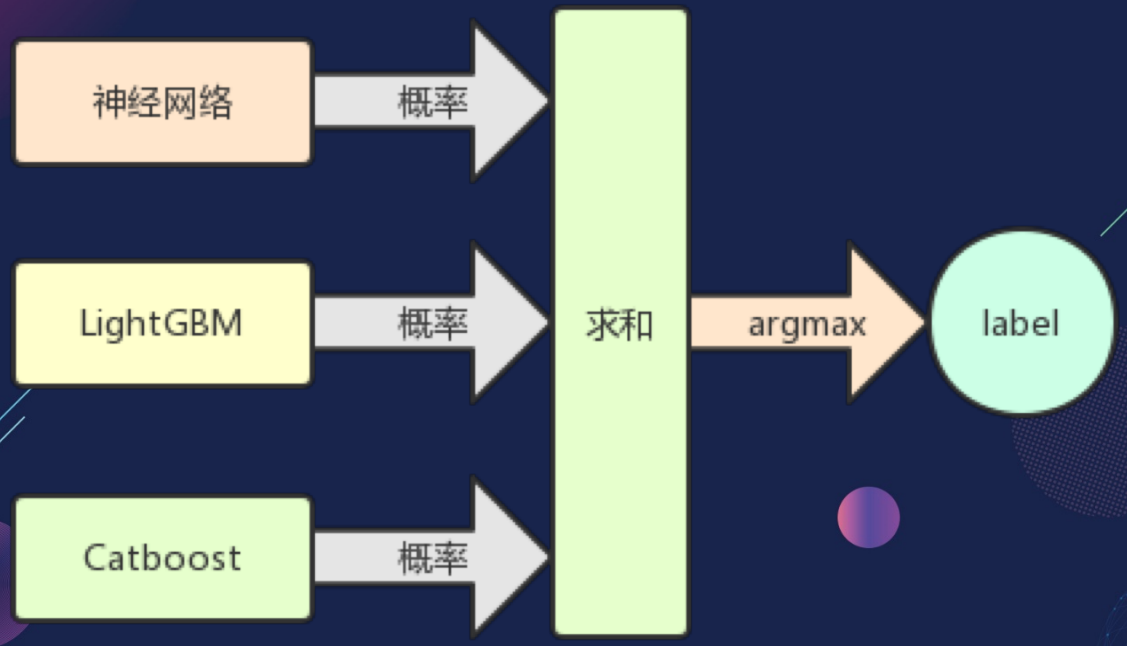
- 对用户每天app的使用日志，统计top100的app的时长和次数，按一周七天对每天的app使用时长和次数进行求和，构造一个（用户数，7，200）的三维矩阵，作为LSTM层的输入
- 对用户每天app的使用日志，统计top50的app的时长和次数，按一个月对每天的app使用时长和次数进行统计，构造一个（用户数，30，100）的三维矩阵，作为LSTM层的输入

5. tfidf 特征

- 对用户安装的app列表和使用的app列表，使用tf-idf和卡方选出特征

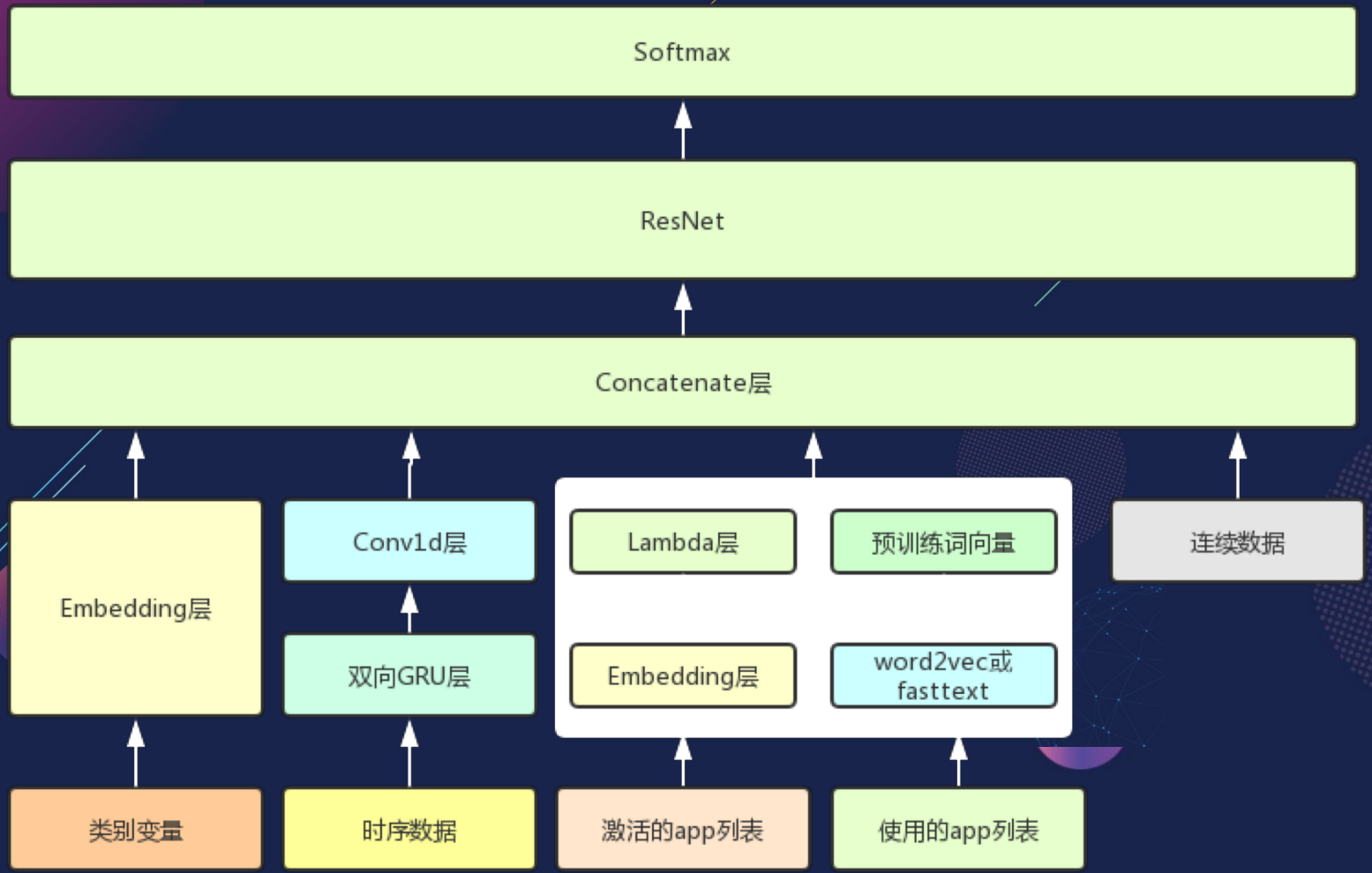
4. 算法实现

使用了神经网络、LightGBM、Catboost 模型，并对最终预测结果进行融合



4. 算法设计

神经网络


Softmax
ResNet
Concatenate层
Embedding层
Conv1d层
双向GRU层
Lambda层
预训练词向量
Embedding层
word2vec或
fasttext
连续数据
类别变量
时序数据
激活的app列表
使用的app列表

4. 算法设计

LightGBM

Catboost

5. 实验过程

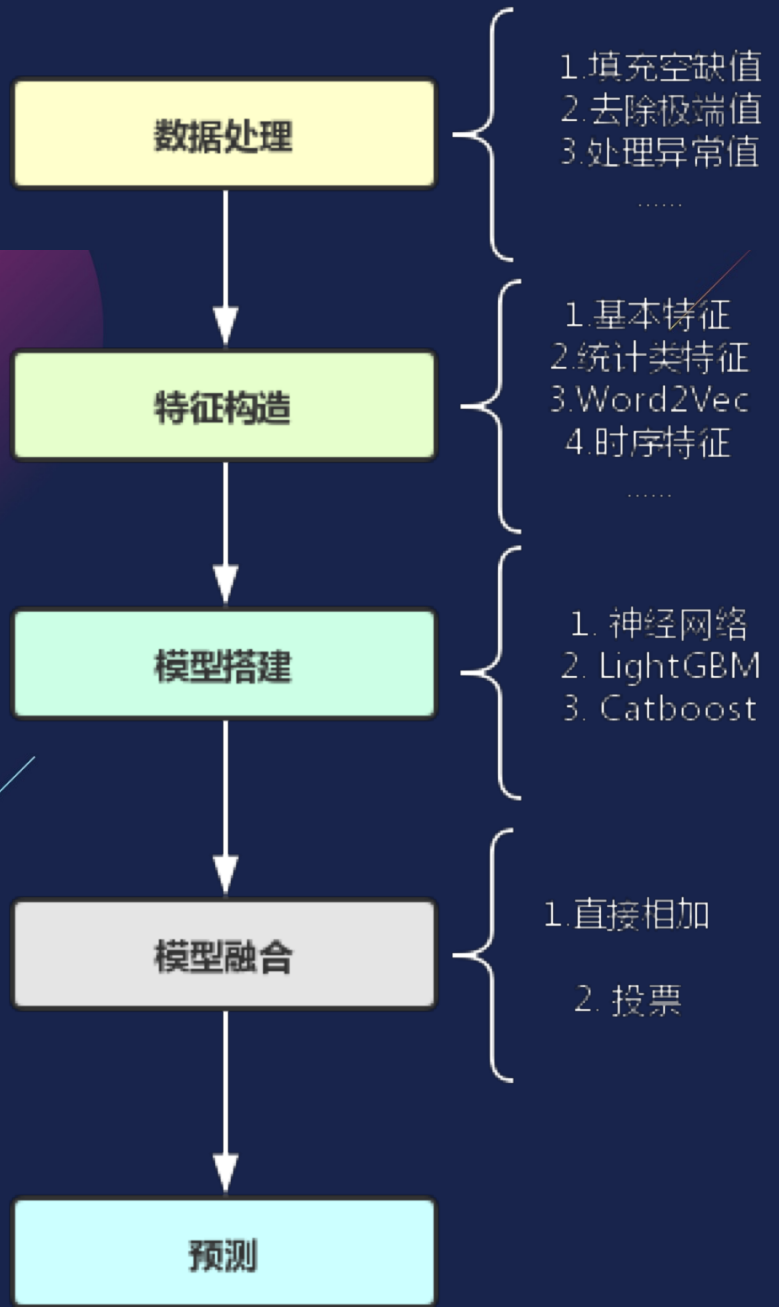
1. 对NN的尝试：

- 对app使用日志记录，取top100的app，按每来天构造一个（用户数，30，200）的三维矩阵，作为LSTM层的输入
- 使用双向LSTM/GRU
- 将DNN替换为残差结构的网络，不断加深网络，

2. 结果调优：

- 五折交叉
- 融合不同类型、不同强弱的模型

6. 整体方案


数据处理
1.填充空缺值
2.去除极端值
3.处理异常值
.....
特征构造
1.基本特征
2.统计类特征
3.Word2Vec
4.时序特征
.....
模型搭建
1. 神经网络
2. LightGBM
3. Catboost
模型融合
1.直接相加
2. 投票
预测

方案优势

- 使用多种方式捕获用户app激活与使用信息.如统计，tf-idf，embedding，lstm等
- 只对热门app进行统计分析，减少计算量
- 使用了残差结构，让网络可以搭得更深，学到更多的信息

HUAWEI

Thank You