数据挖掘课程项目报告

精准营销中的搜狗用户画像挖掘

赵文天 2120171105　赵赫 2120171103　朱佳琪 2120171107　张映雪 2120171101

一、 问题描述

1.1 问题研究背景

“物以类聚，人以群分”这句古语不仅揭示了物与人的自组织趋向，更隐含了“聚类”和“人群”之间的内在联系。在现代数字广告投放系统中，以物拟人，以物窥人，才是比任何大数据都要更大的前提。如何把广告投放给需要的人，是大数据在精准营销中最核心的问题，如何越来越精确的挖掘人群属性，也一直是技术上的天花板。对于企业主来说，了解自身产品的受众有助于进行产品定位，并设计营销解决方案。

在现代广告投放系统中，多层级成体系的用户画像构建算法是实现精准广告投放的基础技术之一。其中，基于人口属性的广告定向技术是普遍适用于品牌展示广告和精准竞价广告的关键性技术。人口属性包括自然人的性别、年龄、学历等基本属性。在搜索竞价广告系统中，用户通过在搜索引擎输入具体的查询词来获取相关信息。因此，用户的历史查询词与用户的基本属性及潜在需求有密切的关系。

如以下几个例子：

- 年龄在 19 岁至 23 岁区间的自然人会有较多的搜索行为与大学生活、社交等主题有关
- 男性相比女性会在军事、汽车等主题有更多的搜索行为
- 高学历人群会更加倾向于获取社会、经济等主题的信息

搜狗提供了用户历史一个月的查询词与用户的人口属性标签（包括性别、年龄、学历）做为训练数据。因此，本问题目标为通过机器学习、数据挖掘技术构建分类算法，根据用户的历史查询词来对新增用户的人口属性进行判定。

1.2 任务定义

基于以上问题研究背景，本课程项目解决的任务有以下形式化定义：

- 给出包含有N个用户历史搜索 query 文本的数据集。
- 对于任意用户$user_i$，其搜索记录中包含n个 query，每个 query 为中文短文本。

- 数据集中同时给出用户$user_i$对应的人口属性标签（性别标签$gender_i$，年龄标签age_i，学历标签$education_i$）。
- 利用训练集中的用户 query 文本和人口属性标签数据进行训练分类模型。要求该模型能够根据测试集中任意用户的历史搜素 query 文本，预测出该用户对应的人口属性标签。

二、 数据预处理

2.1 数据集描述

本课程项目使用的数据来自于 2016 年 CCF 大数据竞赛搜狗用户画像提供的搜狗搜索数据。该数据集共包含训练数据 100000 条、测试数据 20000 条。

数据各个字段描述如下：

- ID 加密后的 ID
- Age 0：未知年龄;
  1：0-18 岁;
  2：19-23 岁;
  3：24-30 岁;
  4：31-40 岁;
  5：41-50 岁;
  6： 51-999 岁
- Gender 0：未知;
  1：男性;
  2：女性
- Education 0：未知;
  1：博士;
  2：硕士;
  3：大学生;
  4：高中;
  5：初中;
  6：小学
- Query List 搜索历史记录中的 query 列表

2.2 数据分析与可视化

经过对数据集的初步分析，我们可以得到在训练数据上，性别、年龄、学历的分布分别如图 2-1、图 2-2、图 2-3 所示。

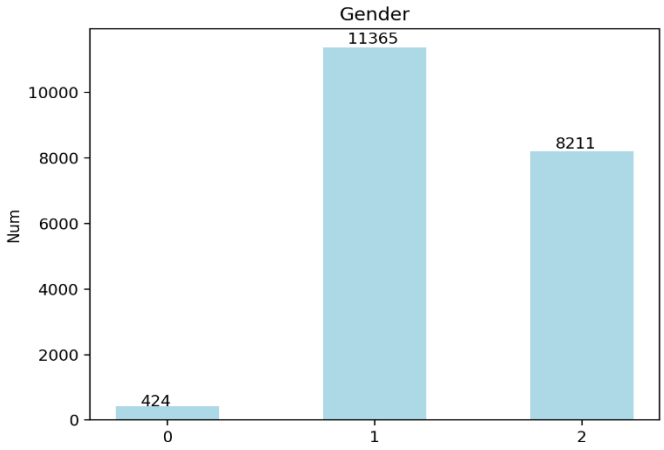

图 2-1 训练数据上的用户性别分布

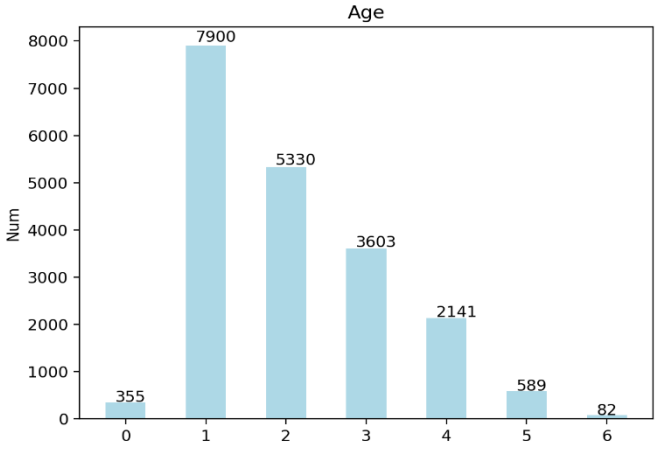

图 2-2 训练数据上的用户年龄分布

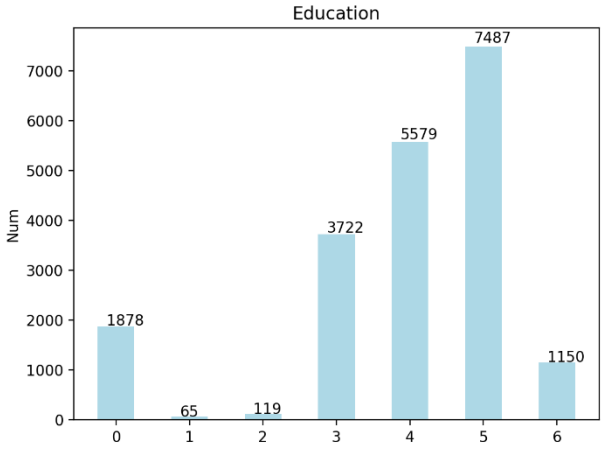

图 2-3 训练数据上的用户学历分布

通过对数据的可视化观察我们有以下发现：

- 三个人口属性标签数据都存在着部分缺失。

- 在性别分布上，男性略多于女性，整体分布较为均衡，缺失标签较少。
- 在年龄分布上，由 1-6 依次减少，说明年轻用户群体占比较大。
- 在学历分布上，1、2 类别较少，多数集中在 3、4、5 类，说明高学历用户群体占比较大。且缺失标签数据较多。
- 存在较严重的数据不平衡问题。

此外，我们还对 query 短文本的情况进行了统计分析，得到的结果如下：

- 单个用户 query 条数最大值：747
- 单个用户 query 条数最小值：50
- 单条 query 长度最大值：338
- 单条 query 长度最小值：1

单个用户 query 条数的分布情况如图 2-4 所示，单条 query 的长度分布情况如图 2-5 所示。

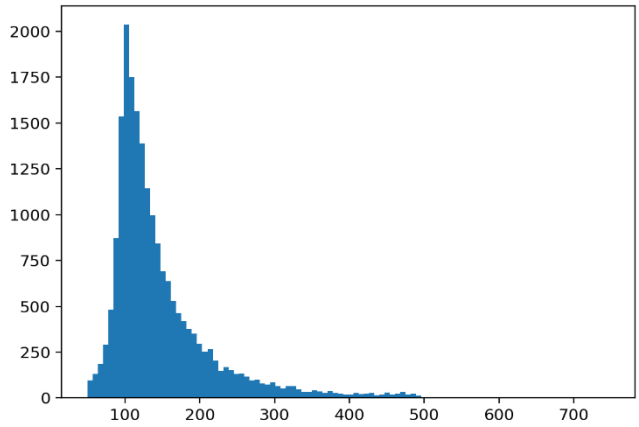

图 2-4 单个用户 query 条数的分布情况

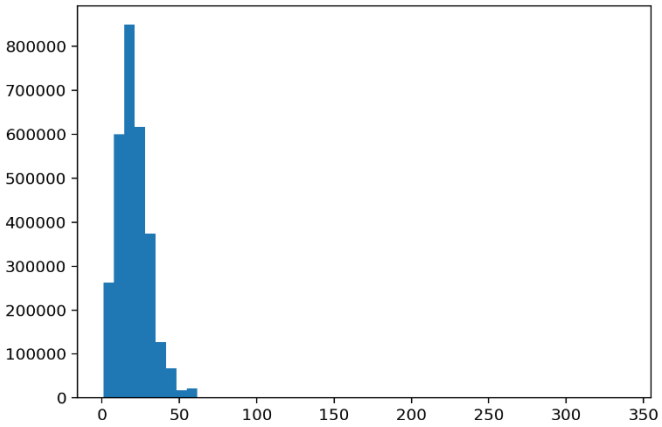

图 2-5 单条 query 长度的分布情况

通过可视化观察我们可以发现，两个分布都基本服从正态分布，因此在实验的过程中，我们在对 query 进行向量化表示时，剔除掉长度大于 60 的

query 文本。

2.3 数据预处理

- **对 query 短文本的预处理：**对每个用户的 query 短文本，考虑到部分 query 中含有的英文单词和数字、链接等会对语义特征的提取产生影响，因此我们去除了 query 中全部的非中文字符，并对剩余中文文本进行去停用词处理。此外，我们去除了 query 中全部以 http://为开头的网页链接字符串。将经过处理的 query 文本，我们使用结巴分词工具对短文本进行分词。
- **对人口属性类别标签的预处理：**对于性别属性标签缺失的数据，我们将其直接从训练数据中剔除。对于年龄属性标签缺失的数据，我们按照训练集上的年龄分布从 1-6 标签中选择一个标签进行缺失值填补。对于学历属性标签缺失的数据，我们同样按照训练集上的学历分布从 1-6 标签中选择一个标签进行缺失值填补。

三、 用户画像分类模型

3.1 query 文本的特征提取与向量表示

对于经过预处理、分词的 query 短文本，我们进一步提取特征，并将每个用户的全部 query 转化为一个 300 维的向量表示。

首先，我们使用 Word2Vec 在中文维基百科数据集上训练得到一个 300 维的词向量模型。基于该模型，我们可以得到 query 中的每个中文单词所对应的 300 维词向量。每个 query 的向量由该 query 中全部单词的词向量加和取平均值得到。既：

$$vec(query_i) = \frac{\sum_{j=1}^{n} vec(word_j)}{n}$$

进一步地，每个用户的向量表示为该用户全部 query 向量的加和取平均值。既：

$$vec(user_k) = \frac{\sum_{i=1}^{q} vec(query_i)}{q}$$

因此，我们将每个用户表示为一个 300 维的向量，该向量作为下一步神经网络的输入。

3.2 神经网络分类模型

我们采用基于深度学习的神经网络来分别构造在三个人口属性标签上的分类器模型。该神经网络分类模型具有以下结构：

- 输入层：300 维的用户 query 向量
- 隐含层 1：150 维全连接层，tanh 激活函数

- 隐含层 2：80 维全连接层，ReLU 激活函数
- 隐含层 3：40 维全连接层，ReLU 激活函数
- 隐含层 4：20 维全连接层，ReLU 激活函数
- 输出层：softmax，对应性别、年龄、学历的输出维数分别为 3、7、7

模型的整体架构如图 3-1 所示。

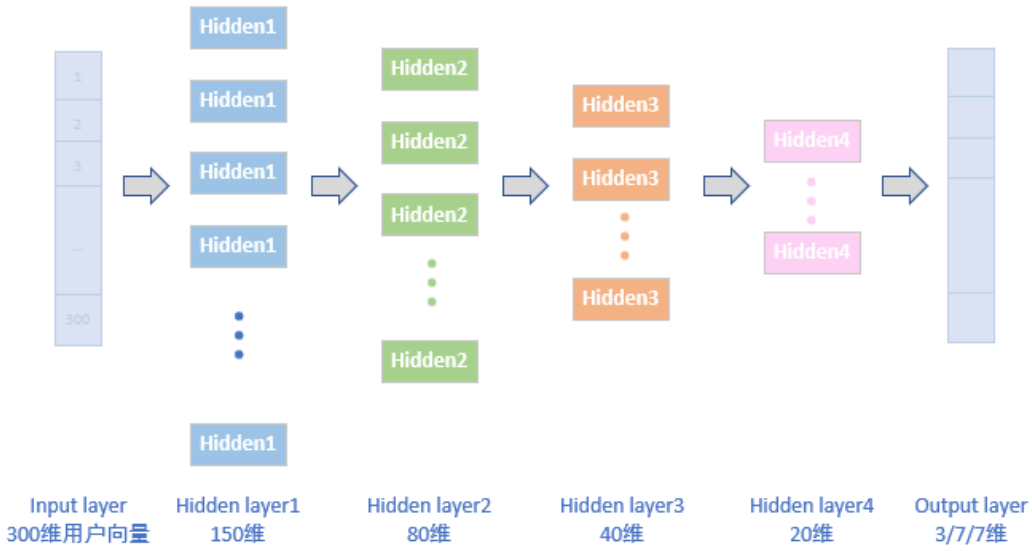

图 3-1 用户画像预测模型整体架构

四、 实验与结果

4.1 实验环境与配置

实验代码使用 python3.6.2 编写，神经网络搭建借助于 Tensorflow 深度学习框架。

模型采用交叉熵作为损失函数，AdamOptimizer 进行梯度下降训练，对 4 个隐含层参数进行随机 dropout，训练时 dropout 比例值为 0.3，测试时不进行 dropout。全局学习率设置为 1e-4，训练迭代 200 轮。

4.2 实验结果

我们将预测属性标签结果与正确属性标签进行对比，以准确率（Precision）、召回率（Recall）、F1-score 作为结果性能指标。对于三个属性标签上的分类器，得到的结果分别如表 4-1、4-2、4-3 所示。

表 4-1 性别属性分类器结果

	Precision	Recall	F1-score
1	0.754482	0.768725	0.761537
2	0.814923	0.834213	0.824455
All	0.784703	0.801469	0.792996

表 4-2 年龄属性分类器结果

	Precision	Recall	F1-score
1	0.650048	0.778987	0.708700
2	0.524043	0.441734	0.479381
3	0.378397	0.606716	0.466098
4	0.361216	0.044371	0.079034
5	0.0	0.0	0.0
6	0.0	0.0	0.0
All	0.318951	0.311968	0.288869

表 4-3 学历属性分类器结果

	Precision	Recall	F1-score
1	0.0	0.0	0.0
2	0.0	0.0	0.0
3	0.446264	0.585864	0.506623
4	0.451802	0.330166	0.381524
5	0.569811	0.839989	0.679011
6	0.0	0.0	0.0
All	0.244646	0.292667	0.261193

4.3 结果分析

- 观察表 4-1 性别属性分类器的结果，我们可以发现在 1、2 属性类别上，该分类器都取得了较为不错的结果，准确率和召回率相对平均，都达到了 0.80 左右。这可能是因为在性别属性上，数据缺失较少，且数据分布较为均衡，利于模型的构建与训练。这一结果也同时证明了我们的方法对用户性别属性进行画像的有效性。
- 观察表 4-2 年龄属性分类器的结果，我们可以发现在 1、2、3、4 类别上，分类器表现较好，且总体效果呈现 1>2>3>4，这与各个分类在数据集上的分布完全一致，而我们训练得到的分类器无法对 5、6 进行分类，可能是由于 5、6 类别的数据量相比其他类别过少，使得数

据量悬殊的差距使得训练出的分类器丧失了对该小部分数据分类的能力，因此也使得在总体上分类器的表现差强人意。

- 观察表 4-3 学历属性分类器的结果，我们可以发现其与年龄属性出现了同样的问题：在数据量占比较多的 3、4、5 类别上分类性能较好，而在数据量较少的 1、2、6 类别上丧失了分类能力，这一进一步证明了我们的猜想：分类器的性能与数据量与数据分布有很大的关系。因此，为了提升分类器的性能，我们必须设法提升分类器对在训练集上分布较少的样例的判别能力。

五、 总结

本课程项目主要面向基于用户的历史搜索 query 短文本，构建用户人口属性标签（性别、年龄、学历）预测模型，从而实现对用户的画像。我们首先分别对 query 文本、属性标签进行数据预处理，包括本文内容清洗、短文本分词、标签缺失值处理等；然后利用 Word2Vec 词向量对 query 进行特征提取，并将用户转化为向量表示；最后通过搭建神经网络模型构造属性标签分类器。我们在搜狗提供的数据集上进行了模型测试，结果显示，我们的预测模型对于三类属性标签都取得了不错的效果，证明了我们方法的有效性。

在我们完成本课程项目的过程中，遇到了以下几个问题：

- 数据样本分布不均衡问题。通过对实验结果的观察我们可以发现，由于属性标签类别在数据集上分布的不均衡，对实验结果造成较大的影响，使得我们训练的分类器在分类的过程中不具备对稀有属性类别的分辨能力，而是倾向于将数据分至占比较大的标签类别。而这种数据分布的不均衡造成的结果是，即使我们的分类模型在部分类别上表现较好，但因为缺少对其他几个稀有类别的判别能力，使得整个多分类器的整体性能较差。
- Query 特征提取方法问题。在设计 Query 特征提取方法的过程中，我们还尝试了将用户全部 query 中的词向量加和直接求平均的方法来作为用户向量进行实验，实验结果不如当前所使用的先求 query 向量表示，再加和求平均作为用户向量表示的方法。此外，我们起初还尝试了使用 TF-IDF 作为 query 的特征构造 3000 维向量，并使用 CH2 卡方检验来选择特征词表，但可能由于该高维特征向量过于稀疏，将其作为神经网络的输入时没有取得很好的效果。
- 神经网络结构的设计问题。起初我们尝试了将 word2vec 用户向量和 TF-IDF 用户向量作为神经网络的输入，并在输入层上搭建 CNN 来构造神经网络分类器。但由于输入向量较为稀疏，且用户的 query 条数以及 query 长度分布较为分散，需要对短文本进行大量的填补，此外还有数据量较大，网络参数较多训练难度大等问题，没有取得预想的效果。

针对以上问题，我们在未来的工作中有以下设想的解决方案：

- 针对数据样本分布不均衡问题，首先，可以考虑尝试一些简单的数据均衡的方法，例如上采样/下采样方法，使训练数据集中各个属性标签类别上的分布保持均匀，再在调整过的数据集上训练分类模型。
- 针对 query 特征提取方法问题，和神经网络结构设计的问题，由于我们

初期尝试过的基于 word bag 的方法对 query 的特征提取能力较差，在未来的研究中，我们还可以尝试将当前用户画像的问题抽象为短文本分类的问题，其中一些近几年出现的算法如 FastText、Text-CNN 等都可以进行尝试。将当前的对每个用户建模改为对短文本建模，然后将对每个短文本分类的结果采用投票机制作为对用户属性的分类。