信息检索大作业

大数据精准营销中搜狗用户画像挖掘

常海浪,彭泽军,王文安,周世宇

目录

一、 赛题描述 4

竞赛背景 4

竞赛简介 4

二、 数据描述 5

数据集： 5

数据介绍： 5

任务描述 6

三、 数据预处理 6

数据分布分析 7

不确定类别样本的比例 7

样本类别分布 7

搜索词分布 8

数据清洗 9

特征构建 10

特征选择 10

特征提取 11

离群点检测 12

四、 模型构建 13

模型框架 13

数据集划分 14

模型选择 14
假设空间 14
评价标准 14
比赛结果 15
五、 结论 15
SVM 分类器的效果相当不错 15
TF 在分类中作用并不大 15
权重构建对分类性能影响巨大 16

一、 赛题描述

竞赛背景

"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。本题目以精准广告中一个具体问题为例，希望发掘到数据挖掘的优秀人才。

竞赛简介

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。

举例如下：

1、 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

2、 男性相比女性会在军事、汽车等主题有更多的搜索行为

3、 高学历人群会更加倾向于获取社会、经济等主题的信息

本题目提供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，要求参赛人员通过机器学习、数据挖掘技术构建分类算法来对新增用户的人口属性进行判定。

二、 数据描述

数据集：

数据文件 备注

Train.csv 带标注的训练集

Test.csv 测试集

数据介绍：

本数据来源于搜狗搜索数据，ID 经过加密，训练集中人口属性数据存在部分未知的情况（该情况为竞赛题目特定设置，需要参赛人员的解决方案能够考虑数据缺失对算法性能的影响）。数据所有字段如下表所示：

字段	说明
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：未知学历; 1：博士; 2：硕士; 3：大学生; 4：高中; 5：初中; 6：小学
Query List	搜索词列表

数据示例：

对于 train.csv 中的数据记录：

00627779E16E7C09B975B2CE13C088CB 4 2 0 钢琴曲欣赏 100 首 一个月的宝宝眼睫毛那么是黄色 宝宝右眼有眼屎 小儿抽搐怎么办 剖腹产后刀口上有线头 属羊和属鸡的配吗

表明该用户是一个 31-40 岁之间，女性，学历未知。

任务描述

本题目提供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，要求参赛人员通过机器学习、数据挖掘技术构建分类算法来对新增用户的人口属性进行判定。即对 test.csv 文件中的每条记录进行年龄、性别、学历的判断。

三、 数据预处理

为方便描述，引入标记

X	样本特征
A	年龄标签 age，值为 1,2,3,4,5,6
G	性别标签 gender，值为 1,2
E	教育程度标签 education，值为 1,2,3,4,5,6
0	不确定类别标签填充值

数据分布分析

不确定类别样本的比例

竞赛要求输出不能包含 0 类即不确定类别信息。考虑到不确定类别信息的样本可能来源于任意类别的样本，所以预测和训练过程中不包含 0 类样本是很有必要的。如果把不确定类别单独作为一个类，会很大程度上干扰分类器的决策边界，所以需要对 0 类样本进行一些必要的处理

以 20 000 条初赛数据数据为例，其中缺失年龄的用户有 355 条，缺失性别的用户有 424 条，缺失学历的数据有 1879 条。所有缺失标签的用户有 2337 条。

样本类别分布

我们在研究样本类别分布时，假设样本三个类别间是相互独立的，即

$$\mathrm{p}(\mathrm{A}, G, E|X) = \mathrm{p}(\mathrm{A}|X)p(E|X)p(G|X)$$

这样，预测标签（A,G,E）就可以归约为分别预测 A，E, 和 G。

这个假设十分粗糙，因为事实上年龄和教育有很强的相关性，预测年龄和预测教育程度应该有很高的一致性，而性别基本上是和其他两个属性是独立的。即：

$$\mathrm{p}(\mathrm{A}, \mathrm{G}, \mathrm{E}|\mathrm{X}) = \mathrm{p}(\mathrm{A}, \mathrm{E}|\mathrm{X})\mathrm{p}(\mathrm{G}|\mathrm{X})$$

我们分析认为，当每一个分类器效果足够好的时候，虽然两个模型概率值得出的结果是不同的，但他们从属与某个类别的相对顺序是不变的。

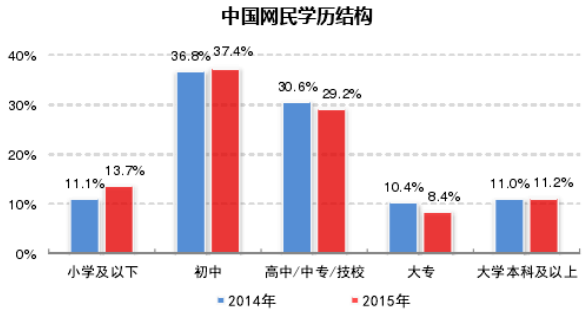
三个类别的比例如下：

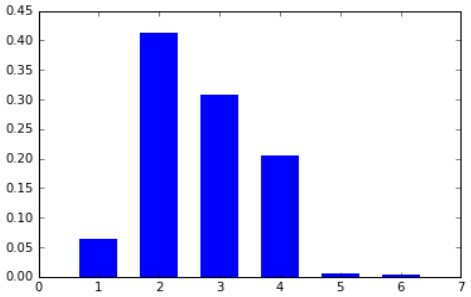
类别	1	2	3	4	5	6
A	7900	5330	3603	2141	589	82
G	11365	8211	--	--	--	--
E	65	119	3722	5579	7487	1150

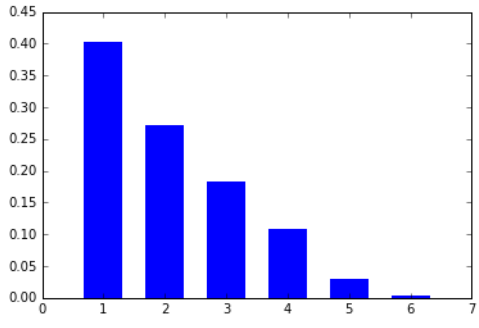
我们发现搜狗这一搜索引擎的用户中，0-18 岁的用户最多，而 0-23 岁年龄段用户占据了所有用户的绝大部分。用户的性别上男性用户略多于女性用户。在受教

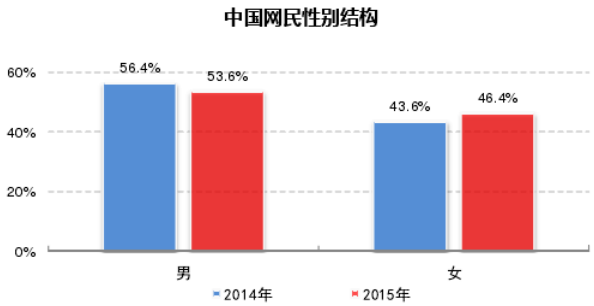
育程度上，搜狗搜索引擎的主要用户为初高中生。

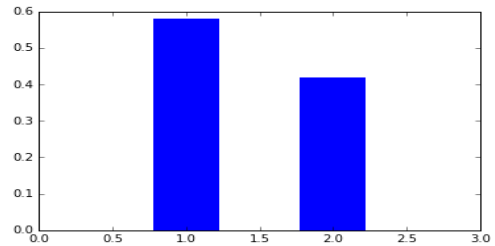
对比中国网民年龄，性别，学历结构图（图片来自于 CNNIC）











注意到，在学历结构上，CNNIC 的大专，大学本科及以上应该合并，搜狗用户的 4,5,6 项应该合并。在年龄结构上，搜狗数据的年龄划分与 CNNIC 的年龄划分的区间并不一致。

这样我们发现，搜狗用户的用户结构和中国网民的整体分布大致是一致的，这也就意味着搜狗用户群体并没有很大的特殊性。

搜索词分布

对于用户查询语句，我们使用中文分词工具将其转化为词项。分别统计每个用户每个词项出现的次数以及出现各个词项的用户数。

以初赛 20 000 条数据为例，经过使用分词工具 NLPIR 分出词项共计 208304 条，

词的 DF 分布如下：

DF	1	1-10	10-100	100-1000	1000-10000	10000-+
Count	113259	61285	25620	7088	1025	27

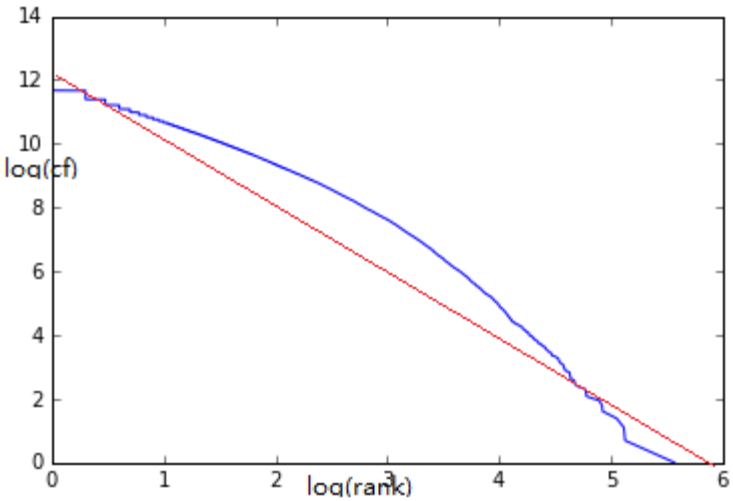
明显发现的规律是，高 DF 的词语数量相对低 DF 词语的数量是十分稀少的。一半以上的词语只出现过一次，这对于分类是毫无意义的，去除这部分词可以极大程度上减少将来训练所需的时间和空间。DF 小于 10 的词语占据超过 80%。

统计 DF 的前六名，分别是：

的：19871，是：18539，什么：18076，怎么：17081，有：16567，\blank：15488

出现最多的词为助词“的”。同时“什么”，“怎么”大量出现，说明用户习惯以提问的方式使用搜索引擎而非关键词查询。空格符出现在超过 75%的用户中，说明用户习惯使用空格符作为查询关键词的分割符。

同时，统计每个词项在所有文档中出现的次数，回执 log(rank)-log(cf)图像：



大致服从 Zips 法则。

数据清洗

通过上述分析发现，整个数据集的质量还是相当高的，基本符合官方给的统计数据以及过去的一些经验分布。唯一需要注意的一点是未知类别信息样本。这种样

本往往从各个类别中产生，如果保留这个类别，会对分类边界造成不可估计的影响，当然可以通过一些机器学习的方法对标签进行预先预测，但是这样的精度本身不高，不如直接放弃这些量并不算大的样本。

特征构建

将每个用户作为一个样本，用户标签作为预测标签，希望从用户的所有查询记录中构建出可用的特征。

我们将用户的所有的查询作为一篇文档，这主要考虑到用户的单词搜索记录太短，无法包含足够多的有用信息。

采用 tf-idf 权重构建特征。

tf 计算方式为$\mathrm{tf} = \log_{10} tf + 1$

idf 计算方式为$\mathrm{idf} = \log_{10}\left(\frac{N}{df} + 1\right)$

我们考虑到，idf 作为全局特征，与单个用户无关。如果所有的用户来自同一个分布产生的样本，那么样本数量越多，idf 的计算就越精准。比赛官方给了 20 000 条有标签的数据作为训练集，20 000 条无标签的数据作为测试集，我们认为充分利用到这 20 000 条无标签数据对于特征构建是十分有比较的。所以我们在实际计算 idf 中，使用的 N=40 000，df 为全部 4 万条记录的 df。

得到 df 后，我们构建了用户-词项矩阵，由于矩阵的大部分为 0，我们采用稀疏矩阵进行空间的压缩。

特征选择

tf-idf 权重实际上已经对特征进行了一次选择，它将一些不重要的词语的权重降的很低，但是仍然需要进一步选取有用的特征。

从前面对数据的分析我们可以知道，DF=1 的词项在所有词项中的比例超过 50%，而 DF=1 对于分类来说是毫无帮助的，所以我们毫不犹豫的把这部分词项给剔除。类似的，DF 较低的词我们认为在分类中作用不大，我们也需要去除。经过反复测试，我们发现将 DF 小于 5 的词语去除是比较合适的。

DF 特别大的词语一般都是停用词，或者由于所有类别的用户的部分查询记录里都包含该词项，所以用处也不大，我们剔除了 DF 大于 68%*N（N 为样本数量）的词项。

我们分别从底部剔除 DF 较小的词和从顶部剔除 DF 较大的词时发现，中频度的词对于分类的影响是十分巨大的，而低频词即使减去 10 万个左右，分类的精度都不会很快的下降，而高频度词语本来就很少。

在上述的特征选择过程中，仅仅是通过一些简单的规则去除信息量较少的词项，还需要进一步根据类别信息。

卡方χ^2统计量是在文本分类中最常用的特征选择标准，我们计算所有词项的χ^2值，按照 tok-K 规则选取出来部分词项作为预测的特征。在初赛中，我们选取的 K 为五万，在复赛中，我们选取的 K 为三十万（复赛中使用 Jieba 分词工具，并且由于样本数量大大增加，导致分词自后 DF>1 的词项约 60 万个作用，DF>5 的词项也有约 50 万个）。

特征提取

无论是 10 万维还是 30 万维的向量，使用任何分类器进行训练都是一个不可能的任务，就样本规模而言，即使分类器可以训练下去，过拟合问题也是不可避免的。我们使用线性代数上的一些方法对特征空间进行降维。

奇异值分解 SVD 是一种基于线性变换的降维方法，它将根据样本在空间上的分

布将每个样本投影到一个低维空间上，同时尽可能的保持了样本的信息。

任何一个矩阵 S，都可以写成它自身的奇异值分解形式：

$$S = U * \Lambda * V$$

如果分别取 U 的前 n 行和 V 的前 m 列，那么就得到了 S 的底秩逼近。

可以根据需要选择保持特征的维度，我们在初赛中将样本投影到了一个 300 维的空间上，在复赛中选取了 450 维的空间。

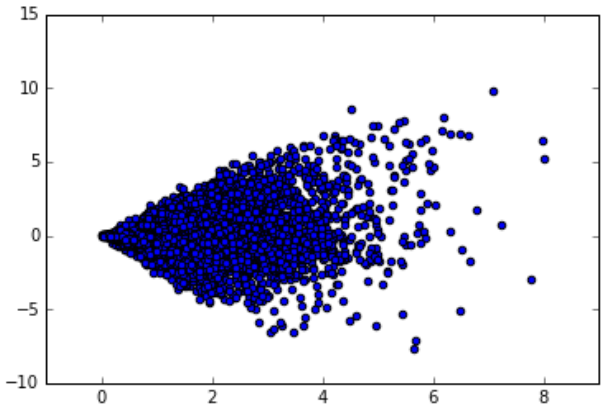
同样 SVD 是根据样本在空间的分布来决定投影的方向，增加样本量有助于提高精度，所以我们在特征提取的过程中将训练集的 20 000 个样本也加了进去。

离群点检测

离群点会很大程度上影响分类器的性能，希望能通过一定的方法去除离群点。

离群点去除的方法很多，比如各种聚类方法原则上都可以作为去离群点的方法。

我们简单的采取均值方差原则进行离群点检测



我们采用6σ原则去除离群点。如果样本服从正太分布，该原则保证了正常点只有百万分之三个几率被排除在外。虽然我们的数据不服从正太分布，但是 6sigma 原则也将极大的保证正常样本不会被这个方法剔除。

同时考虑到这样一个问题，经过 SVD 后的赝本是按照重要程度排序的，我们不希望由于不重要的特征导致样本的损失，所以离群点检测实际上只进行了前 6 个

维度。

四、 模型构建

模型框架

如同上文介绍的那样，我们将分别对三个属性单独进行预测，认为任意两个属性是相互无关的。

基本预测过程如下：

构建用户词项矩阵

获得该矩阵的一个拷贝

去除 0 类样本对应的行

根据 DF 剔除词项对应的列

分裂数据集为训练集，开发

用卡方进行特征选择

用 svd 进行降维

训练分类器

评估分类器

模型选择

合并训练集，开发集，测试集

构建用户词项矩阵

获得该矩阵的一个拷贝

去除 0 类样本对应的行

特征选择

用 svd 进行降维

将测试集分离出来

使用训练集，开发集训练已选择出的分类器

使用训练好的分类器预测

数据集划分

比赛过程中虽然可以提交结果通过精度来对模型的好坏进行反馈，但是每天有提交限制，所以自己评估模型很有必要。

初赛数据较少，我们采用了交叉验证的方法对数据集进行划分，每次随机的将数据集划分成 5 份，4 份训练，1 份评价，通过这样的方式选取模型和参数。复赛数据量较大，一次交叉验证时间过多，我们将训练数据划分为训练集和开发集，虽然效果不如交叉验证，但是也可以选出不错的参数。

模型选择

假设空间

我们选取了过去普遍认为较好的一些分类器进行选择，分别有 KNN，朴素贝叶斯（伯努利模型和多项式模型），logistics 回归，支持向量机（线性核，多项式核，径向基核）。

当然，特征选择过程中选择的特征个数，降维后保留的维数，去低频词的阈值都是需要选择的参数。

评价标准

比赛的评价标准是精度，为了能提高比赛成绩，我们也采用精度作为评价标准。同时，还要考虑到训练时间和所需空间。太长的时间开销和太大空间开销的模型都将被放弃。

KNN 由于计算时间过长，尝试之后就放弃了。

其中径向基核的支持向量机总是会有很好的表现，而 logistics 回归效果比支持向量机略低但训练时间比较短。训练最快的是朴素贝叶斯，它的效果总体上来说也不算太差。

比赛结果

初赛

5	InitZero	0.68975

复赛

16	InitZero	0.70443

五、 结论

我们在比赛测试过程中发现以下一些经验性的结论

SVM 分类器的效果相当不错

基于 SVM 分类器根据查询词对用户属性进行判断具有不错的可行性，尤其是对性别标签上，SVM 分类器能够达到 83%以上的准确率。根据少量数据的测试，此时机器的判断效果甚至超过小组成员的判断。

TF 在分类中作用并不大

tf-idf 权重虽然思想简单但是往往很有用。在使用过程中发现，idf 的影响要远远超过 tf，线性的 tf 表现相当的差，强制令 tf 等于 1 并不会严重的导致分类效果的下降，对 tf 进行对数的限制是在二者中取得的较好的平衡。这可能是由于查询过程中经常由于一次查询结果不理想而反复查询包含某个关键词的语句导致 tf

激增但是语义特征并没有明显变化的缘故。

权重构建对分类性能影响巨大

在同一个分类器下，对分类器参数进行调整并不能有效的提高分类效果，但是我们尝试改变 tf-idf 构建方式时，分类器效果发生了巨大的变化，如上文所述，线性的 tf 并不能具有很好的效果，而对数限制的 tf-idf 中，我们测试的底数为 10 时达到最好的分类效果。另外，有论文表示在权重加入类别信息会有效的提高分类效果，但遗憾的是，在我们已经尝试的方法中，并没有找到如各种论文所述的修正方式，能够对我们的比赛结果进行提高。