南京大学

本 科 毕 业 设 计

院　　系 计算机科学与技术系

专　　业 计算机科学与技术

题　　目 基于搜索关键词的用户属性分析预测

年　　级 2013 级 学号 131220167

学生姓名 禤宝琼

指导老师 黄宜华 职称 教授

论文提交日期 2017.06

南京大学本科生毕业论文（设计、作品）中文摘要

题目：基于搜索关键词的用户属性分析预测

计算机科学与技术 院系 计算机科学与技术 专业 2013 级

本科生姓名：禤宝琼

指导教师（姓名、职称）：黄宜华 教授

摘要：

在广告的精准投放中，根据用户的历史行为来反推用户的属性是一项基础技术。用户在搜索引擎中查询的内容与用户的性别、年龄、学历等有着密切的关系。例如人群男性在军事、汽车主题上有更多的搜索行为，19～23 岁搜索行为中较多与大学生活、社交生活有关，高学历人群更倾向于获取社会、经济方面的信息。

本文研究以用户历史的查询关键词与用户的人口属性标签（性别、年龄、学历）做为训练数据集，利用搜索关键词与用户属性的关联性，通过机器学习、数据挖掘技术构建分类算法来对新增搜索用户的人口属性进行判定。

关键词：文本分类；朴素贝叶斯；支持向量机；KNN；用户画像

南京大学本科生毕业论文（设计、作品）英文摘要

THESIS： Analysis and Prediction of User Attributes based on User's Searching Keywords

DEPARTMENT：Computer Science and Technology

SPECIALIZATION：Computer Science and Technology

UNDERGRADUATE：BaoQiong Xuan

MENTOR：YiHua Huang

ABSTRACT：

In the accurate delivery of advertising, it is a basic technology to derive user's attributes according to the user's historical behavior.

The content that the user inquires in the search engine is closely related to the user's sex, age, educational background and so on. For example, crowds of men have more searches for military and car themes. Youth's search behavior is more related to college life and social life. Highly educated people tend to gain social and economic information.

In this study, user history query key words and user demographic attribute tags (gender, age, educational background) are used as training data sets. By using the relevance of search key and user attributes, a classification algorithm is constructed by machine learning and data mining technology to determine the population attributes of new search users.

KEY WORDS: text classification; Naive Bayesian; SVM; KNN; Persona

目录

第 1 章 绪论

1.1 研究背景和意义

在互联网高速发展的大背景下，随着网络广告的兴起，广告行业迎来了更多的机会和更多的挑战。网络广告不只是传统意义上的“广而告之”，而是有目标性的“精准投放”，广告的展示渠道以及内容创意也进行了深度更迭。尤其现在网络信息过载给人们生活带来很不好的体验，用户更是对狂轰滥炸式的广告非常反感。要将用户对网络广告的反感度降到最低，那就必须把广告投放给真正对它有需求的人，这就是所谓的“精准投放”。

近几年来，大数据技术发展迅猛，通过大数据技术来实现大规模数据场景下广告的精准投放，能显著提升广告投放效率、提高经济效益，并且这种方式下，用户能得到更佳的互联网体验。

广告精准投放中，关键的问题是如何找到合适的投放人群。现实生活中，每个人都有自己的住址、姓名、爱好、学历、经历、性别等特征，这些特征将人与人区分开来。网络用户与真实用户一一对应，同样有着各自的属性特征，通过这些特征能够勾勒出相应虚拟用户的人物原型。这就是所谓的用户画像。

用户画像由各种用户属性组成，用户画像挖掘技术中，根据用户的浏览、搜索等行为来反推获取用户属性是一项非常基础且非常重要的技术。本文将根据用户搜索行为与用户属性的相关性对用户属性进行分析预测。

用户在搜索引擎中查询的内容与用户的学历、年龄、性别等有着密切的关系，例如人群男性在军事、汽车主题上有更多的搜索行为，19～23 岁搜索行为中较多与大学生活、社交生活有关，学历高的人群具有更多的社会、经济方面的搜索行为。本文研究以用户搜索关键词与用户的性别、年龄、学历标签（用户属性）等历史数据作为训练数据集，利用搜索关键词与用户属性的关联性，基于数据挖掘、机器学习方面的技术来构建分类模型，对新增用户进行用户画像，完成对年龄、性别、学历的分析预测。

1.2 研究现状

本文根据用户搜索关键词来分析预测用户属性，也就是根据用户的历史搜索

行为对用户进行分类，属于分类预测的范畴。

分类预测是对各种类型的数据进行分类分析的一种关键技术方法，也是数据挖掘领域的主要研究问题之一。

数据集中用户的历史搜索行为是由用户过去一个月在搜索引擎中输入的关键词组成的文本，由此看来，本文的分析预测任务本质上是一个文本分类任务。

文献[1]中指出，在众多文本分类算法中，贝叶斯，K-最近邻，支持向量机以及神经网络等算法有较为优秀的表现。

朴素贝叶斯（Naive Bayesian）分类的核心流程是计算一个待分类样本属于各个分类的概率，这些概率是由贝叶斯定理计算得来的，并根据概率大小来判定该样本所属类别，即将其判定为具有最大概率的那一类。

支持向量机（SVM，Support Vector Machine）算法寻找最佳超平面分割两类线性可分的样本，对于待分类项，看它与这个超平面的相对位置就可以判断它属于哪一类。

K-最近邻算法根据某种空间距离，如欧式距离或曼哈顿距离等，挑选出最靠近未分类项的 K 个样本，看这些样本中哪个类别占的数目最多，就把该待分类项判定为属于那一类。

通过观察训练数据集，发现一个问题是，各类别的样本数量可能出现分布不均衡的情况。例如，在样本数据集中，年龄分布中，随着年龄的增加样本数据量逐渐、甚至急剧减少，0-18 岁有 7900 个样本，19-23 岁有 5330 个样本，24-30 岁有 3603 个样本，31-40 岁有 2141 个样本，41-50 有 589 个样本，51-999 岁仅有 82 个样本。对于这个情况，文献[2]中介绍了一种单分类支持向量机（One-class SVM）的方法，该算法构造一个高维超球面，使得一类数据样本全部位于该超球面内，那么当新的数据出现时，若新数据处于该超球面内，则其属于这个类别否则不属于。该方法适用于有两种类型样本，但其中一类型样本数目缺失或远少于另一类型样本数目的情形。推而广之，该方法也能用于多种类别的场景。

1.3 本文研究内容

本文研究以用户查询关键词文本与用户的人口属性标签（性别、年龄、学历）历史数据做为训练数据集，进行用户分类预测分析。

用户人口属性标签包括年龄、性别、学历，其中，性别包括有男、女 2 类标

签，年龄包括有 0-18 岁、19-23 岁、24-30 岁、31-40 岁、41-50 岁、51-999 岁 6 类标签，学历包括有博士、硕士、大学、高中、初中、小学 6 类标签。

这里，采用交叉验证的方法，对数据集进行了不同的测试集和训练集的划分。

利用向量空间模型对文本做数值化的表达，在向量空间模型中，文本可以被表示成一组向量：$(w_1, w_2, \ldots w_n)$，w_i为第 i 个特征项的权重，以词作为特征项，本文选取词 TF-IDF 值表示权重。为此，首先对用户的搜索关键词文本做了分词、统计词频等处理。

然后，对特征做降维处理，原始的特征向量高达数十万维，极大地降低了分类的效率和准确度。本文分别采取了 CHI 统计、互信息（MI）、信息增益（IG）三种特征选择方法来选择保留更具代表性的特征，去除其他特征。

最后利用 SVM、K-最近邻、朴素贝叶斯等算法构造相应的分类模型，对测试数据进行分析分类，并做参数的调优等改进，对各分类模型的准确度、时间效率等做分析比较。

1.4 论文结构安排

第一章为绪论，主要介绍了在现代精准广告投放的背景下用户画像挖掘工作的研究背景与意义。同时简要地对本文所属的文本分类研究范畴所用的基本方法进行了描述，并介绍了其核心部分即分类分析的经典算法和研究现状。

第二章简单地介绍了本文有关的工作背景以及概念定义。首先介绍了应用于文本表示的向量空间模型。然后介绍了文本分类常用的特征选择方法。并分别对基于超平面划分的，基于距离的，基于概率统计理论的、基于神经网络的分类算法进行了较为详细的介绍。

第三章分三部分介绍本文根据用户搜索关键词预测分析用户属性的方案，即中文分词、特征选择、基于不同分类算法构造不同的分类模型；进一步分析了相关部分的缺失数据处理，最后给出了该分类算法的整体框架。

第四章对实验结果进行分析，介绍了实验的相关环境和数据集情况，统计了测试结果，对不同分类模型的预测准确度和时间成本做了对比。

第五章总结评价了本文主要的研究内容。对本文提出的解决方案的优劣进行了评价，并提出进行优化的可能方案，做出了下一步努力方向的建议。

第 2 章 相关工作与背景介绍

2.1 文本表示

未经处理的原始文本不能直接作为文本分类算法的输入，因此文本分类需要进行预处理。文本预处理阶段的目的是将文本转化为某种可计算的数值化表示形式，以便后续进行分析计算处理。

2.1.1 向量空间模型

本节介绍利用向量空间模型对文本做数值化的表示方法。

在向量空间模型中，文本可以被量化成一组线性无关的向量，它的每个元素是带有权重的特征项[4]。特征项一般是词或短语，为组成文本的原始词汇流，特征权重度量一个特征项在文档中的地位，即较大权重的特征项更重要，更能代表文本。

根据特征项和特征权重，文档可以被表示为：$(w_1, w_2, \ldots w_n)$，其中w_i为第i个特征项的权重。

那么，在向量空间模型中，文本可以以多个词权重组成的向量表示，词也可以由其在多篇文本的权重组成的向量表示，这种对偶关系能在文本向量矩阵中得到很好的体现，在文本向量矩阵中每列代表一个词，每行代表一篇文本[3]：

$$\begin{bmatrix} w_{11} & \cdots & w_{1j} \\ \vdots & \ddots & \vdots \\ w_{i1} & \cdots & w_{ij} \end{bmatrix}$$

在上述矩阵中，w_{ij}表示第 i 份文档中的第 j 项特征词的权重，权重的取值依据可以是词的出现次数，也可以是出现频率，tf-idf 亦是本文采用的词语加权方法。接下来的章节将对 tf-idf 加权方法进行解释介绍。

2.1.2 TF-IDF

TF-IDF 基于词频与逆向文件频率来评价一个词在一份文档中的重要性，这个重要程度是相对该文档所在文档集或语料库中的其他文档而言的。

在 TF-IDF 方法中，一个适宜用来分类的词语应当具备这样的性质：其在某

篇文档中具有较高的词频（出现频率），且在别的文档中出现频率较低。也就是说，满足以上性质的词语有着良好的类别区分能力。在这种方法中，词语的重要程度与它在文档集中的出现频率是反比下降的关系，但同时也与它在文档中的出现频率是正比增长的关系。

一个词语在一篇文档中的 TF-IDF 值是其 TF 值与 IDF 值的乘积，IDF 是逆向文件频率（Inverse Document Frequency），TF 是词频（Term Frequency）。

下面给出对于文档d_j中某个词t_i的 TF-IDF 权重的计算步骤：

（1）步骤 1：计算 $tf_{i,j}$

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

上式中，分母是在文档d_j中全部词汇的出现频数总和，而分子$n_{i,j}$则是该词在文件d_j中的出现频数。

（2）步骤 2：计算idf_i

$$idf_i = \log \frac{|D|}{|\{j:t_i \in d_j\}|} \tag{2}$$

其中|D|表示文件集的总文件数，分母表示包含词t_i的文档数。

（3）步骤 3：将（1）（2）两式计算出的结果相乘得到$tfidf_{i,j}$

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{3}$$

由以上公式及步骤可以看出，对于某个特定文档，若文档中的某个词在该文档中的出现频率高，且该词在整个文件集中的出现频率低，则可以得到较高的 TF-IDF 权重。由此可以看出，TF-IDF 偏向于过滤掉掉如“的”、“怎么”之类的常用词语，留下真正重要的词语。

2.2 特征选择

一篇中文文档中会出现成百上千甚至上万个词语，整个文档集的词语总数能高达数十万甚至百万，这导致了文本表示向量的稀疏性和文本特征空间的高维性，这是文本分类所需面对的困难之一。常见的解决思路是通过某种方法衡量词语的重要性，从中挑选出具有代表性的词汇，并过滤掉其他词汇，从而达到降维的目的。

采用一种有效且合适的特征选择方法，对提高分类的准确度和降低分类的时间成本具有重要意义。如何量化特征的重要程度，是各种特征选择算法的核心内容，也是它们的主要区别。下文将介绍四种常见的特征选择方法。

（1）文档频率

一个词语，它的文档频率（Document Frequency）表示的是训练文档集中有多少文档包含该词语。文档频率方法用于提取在文档集中具有一定出现次数的词语，目的是剔除出现次数过少的罕见词汇。文档频率是最简单的特征选择方法[4]。

（2）CHI 统计，即卡方校验

CHI 统计方法，又称卡方校验，用于度量词与类别的相关性[4]。一个词对于某个特定类别的 CHI 值计算方式如下:

$$\chi^2(t,c)=\frac{N\times(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{4}$$

其中，c 为某一特定的类别，t 表示某个特征词， N 表示训练文档集中的文档总数，A 表示包含 t 的 c 类文档数目，B 表示包含 t 但不属于 c 类的文档数目，C 表示不包含 t 但属于 c 类的文档数目，D 表示不包含 t 也不属于 c 类的文档数目。

（3）互信息

为了衡量一个特征对于某类别所含有的信息量，我们常用互信息法。互信息值（词t与类别c）计算方法如下：

$$MI(t,c)=\log\frac{P(c|t)}{P(t)}=\log P(c|t)-\log P(t) \tag{5}$$

其中，P(c|t)表示包含词 t 的文档有多大可能属于 c 类，P(t)表示训练文档集中的某篇文档包含词 t 的概率。

（4）信息增益

信息增益度量一个特征为分类系统带来的信息量[5]。词 t 的信息增益值计算方式如下：

$$IG(t)=-\sum_{i=1}^{m}P(c_i)\log P(c_i)+P(t)\sum_{i=1}^{m}P(c_i|t)\log(c_i|t)+P(\bar{t})\sum_{i=1}^{m}P(c_i|\bar{t})\log(c_i|\bar{t}) \tag{6}$$

可以看出，信息增益实际上在求解这个词存在和不存在时系统的信息量差值。其中，$P(c_i)$表示c_i类文档在训练文档集中出现的可能性大小，P(t)表示文档集

的一篇文档中有多大可能包含词t，$P(c_i|t)$表示包含词t的文档属于c_i类的概率，$P(\bar{t})$表示不包含词t的文档属于c_i类的条件概率，m表示类别数。

2.3 分类算法

本文根据用户历史搜索关键词组成的文本，对用户属性标签进行判断，也就是根据用户搜索文本对用户进行分类。因此，本文所要解决的问题，属于文本分类的范畴。下面介绍一些经典的文本分类和挖掘算法。

2.3.1 基于贝叶斯决策理论的分类算法

将概率统计知识应用到分类任务中的算法有很多种，其中使用范围最广的是基于贝叶斯决策理论的分类算法，包括贝叶斯网络、朴素贝叶斯等。由于在由于基于朴素贝叶斯的分类模型构建是本文的重要研究内容，本节将着重介绍该算法。

贝叶斯定理是朴素贝叶斯算法的理论基石。下面直接给出贝叶斯定理：

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{7}$$

其中，$P(A)$、$P(B)$分别表示事件 A、B 发生的概率，$P(B|A)$表示在事件 A 已经发生的条件下事件 B 发生的概率，$P(A|B)$表示事件 B 已经发生的前提下事件 A 发生的概率。

朴素贝叶斯（Naive Bayesian）分类的核心是计算一个待分类样本属于各类的概率，这些概率均由贝叶斯定理计算得来，并根据概率大小来判定该样本所属类别，即将其判定为具有最大概率的那一类。下面给出朴素贝叶斯的标准分类过程：

（1）有一个待分类样本$x = (w_1, w_2, \ldots w_n)$，每个$w_i$为x的一个特征属性；

（2）有类别集合$C = (y_1, y_2, \ldots y_n)$；

（3）计算$P(y_1|x), P(y_2|x), \ldots, P(y_n|x)$；

（4）如果$P(y_k|x) = \max\{P(y_1|x), P(y_2|x), \ldots, P(y_n|x)\}$，则x属于类别$y_k$。

其中，第 3 步中的$P(y_i|x) = P(x|y_i) * P(y_i)/P(x)$，由于$P(x)$对于所有类别为常数，我们只需比较分子大小，分子大小由下列公式求得：

$$P(x|y_i) * P(y_i) = P(w_1|y_i) \times P(w_2|y_i) \times \ldots \times P(w_n|y_i) \times P(y_i)$$

那么第 4 步最后转化为，求 $\underset{k}{\operatorname{argmax}} P(y_k)\prod_{i=1}^{n} P(w_i|y_k)$。

2.3.2 基于超平面划分的分类算法

空间的点可以被一个超平面划分为两部分，在分类问题中，两类线性可分的样本可被一个超平面分割开来。那么，对于未知的待定项，看它与这个超平面的位置关系就可以判断出它属于哪一类。由这个数据分类问题的处理理论出发，发展出了许多基于超平面划分的分类算法，主要包括感知机、支持向量机等。由于本文主要利用支持向量机构造分类模型，本节着重对其相关理论进行介绍。

支持向量机（SVM，Support Vector Machine）的基础概念为试图找到一个“最优”超平面作为分类界线，保证分类间隔（Margin）尽可能最大化，从而满足分类要求并获得较高的分类精度。分别找到被超平面分割开的两类中距离超平面最近的样本，用这两类样本分别建立平行于超平面的平面，这两个平面间的距离为分类间隔。

下图 2-1 中红叉与蓝圈是两类样本，从左至右是三种超平面分割方法。如图 2-2 所示可以直观地看出最右边的划分方式下，超平面两侧有最大的空白区域，即最大化了分类间隔，容忍误差的能力最强，即支持向量机方法所要找到的最优分类平面。而感知机算法是找到随意一个能分割两类样本的超平面即可。

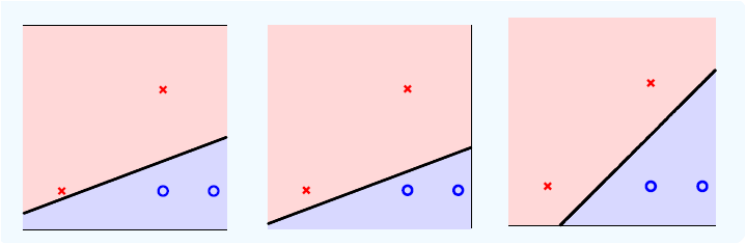

图 2-1 三种超平面分割方式

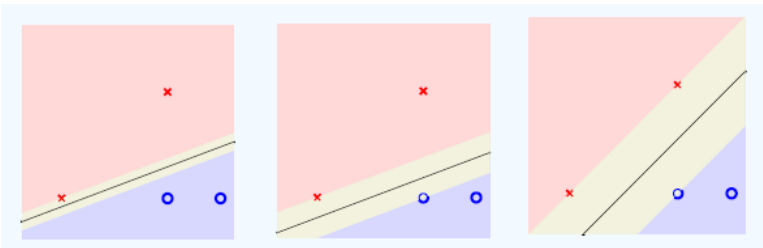

图 2-2 超平面两侧空白区域

对于线性可分的分类任务，若训练集为$T=\{(x_1,y_1),\ldots,(x_i,y_i)\}\in(X\times Y)$，其中$x_i\in X=R^n$，$x_i\in Y=\{1,-1\},i=1,\ldots,i$;设超平面$(w\cdot x)+b=0$，则$1/\|w\|$为超平面关于训练集的集合间隔。寻找最大间隔超平面的任务可转化为求解下以最优化问题[4]：

$$\min_{wb}\tau(\omega)=1/2\|w\|^2\ ,\ y_i((\omega\cdot x_i)+b)\geq 1,i=1,\ldots,i \tag{8}$$

支持向量机使用核函数来解决线性不可分问题，非线性可分的原空间可以被映射到更高维的空间，使其在高维空间中线性可分。支持向量机核函数就是某个高维空间的内积，其在支持向量机中起着至关重要的作用。

2.3.3 基于距离的分类算法

基于距离的分类算法的主要代表是 K 最近邻(KNN，k-Nearest Neighbor)算法。同时，它是本文采用的主要算法之一，因此接下来着重介绍 KNN 算法。

同时，KNN 也是基于向量空间模型的优秀的文本分类算法。该方法的思路是：根据某种空间距离，如欧几里得距离，挑选出 K 个最靠近待分类项的样本，统计这些样本中哪个类别的样本数最多，就把该待分类项判定为属于那一类。

KNN 通过计算新数据与训练数据特征值之间的距离，然后选取 K(K ≥1)个距离最近(即特征空间中最邻近)的邻居进行分类判断。图 2-3 是 KNN 对新数据X_u进行分类的过程示意图，如图所示，样本空间有w_1，w_2，w_3三类样本，在计算出新数据与其他训练数据的距离后，选取出了 5(K 为 5)个距离最近的样本，即黑色箭头指向的 5 个点，而这 5 个样本中w_1类的样本最多，最后判定新数据属于w_1类。

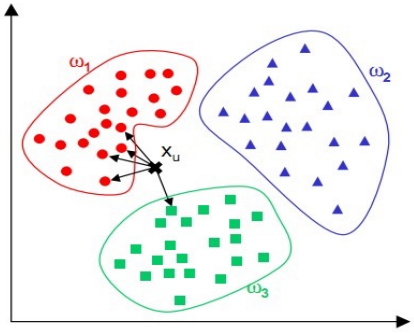

图 2-3 KNN 分类过程示意图

由于 K 值对文本分类效果影响很大，所以 K 值的选择与调整十分重要。但是目前除了根据试验结果来反馈调节 K 值外没有其他更好的确定 K 值的方法[6]。

关于样本间距离的计算，由于本文采用了向量空间模型来表示文本，我们只要计算文本向量间的距离即可。欧几里得距离与曼哈顿距离是 KNN 算法常用的两种距离。

曼哈顿距离公式为：

$$\mathrm{Dist}(x_i, x_j) = \sum_{l=1}^{n} \left| x_i^l - x_j^l \right| \tag{9}$$

欧几里得距离公式为：

$$\mathrm{Dist}(x_i, x_j) = \left(\sum_{l=1}^{n} \left| x_i^l - x_j^l \right| \right)^{\frac{1}{2}} \tag{10}$$

2.4 本章小结

分类模型的构造是本文的关键，本章介绍的几种分类算法各有优缺点。朴素贝叶斯有稳定的分类效率，算法也比较简单，常用于文本分类，但贝叶斯定理成立的假设前提是各属性间互相独立，现实情况中此前提不大容易成立，因此容易损失分类性能。k-近邻方法简单，易于理解，易于实现，但其没有训练过程，需要分类时与先前存放的样本做距离对比来判定类别归属，若样本集较大，需要较多的计算开销，且其还有个不足是，当样本空间不平衡时，在 K 个邻居中占多数的类别更容易是更大容量类。SVM 最小化结构风险，泛化能力优秀，但本质上是二分类算法，处理多分类问题效率会有所下降。

本章先介绍了向量空间模型，并介绍了用于降低特征空间的维数的一些常见的特征选择方法，接着分别介绍了基于贝叶斯理论的、基于超平面划分的和基于距离的分类算法的主要思想、代表算法及其优劣，为下文描述的根据用户搜索文本预测用户属性标签的方法提供理论依据。

第 3 章 用户属性分类预测方法

3.1 用户属性预测方法简介

本章将介绍根据用户搜索关键词预测分析用户属性的方法及其具体的算法设计与实现。本文提出的解决方法主要分成三个步骤。

首先，对每个用户初始的搜索关键词文本数据集进行处理，利用优秀的分词工具“结巴分词”对文本进行分词，并针对数据集文本的特点引入相应的网络词汇，以及做停用词处理。

接着对分词后的文本做向量化处理，这里以词作为特征项，用单词的 TF-IDF 表示权重，并对文本特征向量做降维处理，剔除出现频率极低的词汇，最终保留十万左右的特征项。

最后分别训练用于预测年龄、性别、学历的三个分类器，对于每个分类器的分类算法选择，本文分别设计实现了朴素贝叶斯、SVM、KNN 算法，并对比分析了它们在这个分类任务上的效果。

此外，由于所使用的数据集中还存在缺失的用户标签数据，本文分别采用的处理方法是，根据其他非缺失数据训练出一个分类模型，对这些标签缺失数据文本进行分类，再将这些数据纳入训练数据集中重新训练出新的分类模型。

通过以上的算法过程描述可知, 本文关于基于用户搜索关键词预测用户属性的主要处理流程为：

搜索文本预处理 -> 特征抽取和转换 -> 训练各种分类模型 -> 测试 -> 实验结果分析。

3.2 中文分词

中文分词是中文文本处理的一个基础性工作，本文利用优秀的开源中文分词工具“结巴分词”对用户搜索文本进行中文分词，根据切分出来的词汇数量和质量的差异，划分成不同的分词模式。常见的分词模式分两种，一种是精确模式，试图将句子最精确地切开使得分词结果更符合原始语义；另一种是全模式，全模式会找出句子中全部可以组成词语的连续字串。

下面举例说明两种分词模式的差异：对句子“南京大学计算机系”进行分词

操作，全模式下的结果是“南京/南京大学/计算/计算机/计算机系/算机/系”，精确模式下的结果是“南京大学/计算机系”。

两种模式各有优劣，对于本文的数据集而言，全模式的分词一定程度上对样本数据做了补充增添，并且不用担心有些句子中的关键短词没有被识别出来，但是也存在增添了很多垃圾无意义词汇的缺点；而精确模式下虽然有时一些重要短语没被识别出来，但却不会对训练过程造成额外的“噪音污染”。本文通过实验发现，在本文的任务中，使用精确模式的分类器的准确率比使用全模式的高 1%。

3.3 特征工程

分词过后的文本仍然不能作为分类算法的输入，需要对其做进一步的特征处理。

在第 2.1 小节中介绍了文本表示的方法，根据特征项和特征权重，文档可以被向量化为：$(w_1, w_2, \ldots w_n)$其中w_i为第 i 个特征项的权重。一般选取词作为特征项，常用特征项权重有多种，包括出现次数、在文档中出现频率 tf 以及 tf-idf 等。本文分别采用 tf 和 tf-idf 方法对特征项进行加权并对两者做对比，tf-idf 方法在第 2.1.2 节中已经做过详细介绍，在此不再赘述。

经过以上处理之后，我们发现文本特征向量的维度高达三十多万，极大影响了分类的效率和准确率，因此必须采取降维措施，也就是进行特征选择。

本文第 2.2 小节中介绍了常见的特征选择方式，在本文的任务中，虽然 DF 方法是最简单的特征选择方法，但是 DF 仅考虑了频率因素而没有考虑类别因素，导致 DF 算法非常容易引入一些高频却没有意义的词[11]。如在本文的数据集中，“图片”、“下载”、 “电影”等，DF 值排名前列，然而，在各类别中都是高频词，对分类并没有多大的意义。因此本文主要使用了信息增益（IG）、互信息（MI）、CHI 统计这三种特征选择方法，并且，CHI 统计有最佳的表现。

3.4 分类模型

前面已经叙述了对用户原始搜索文本进行分词处理和特征处理的方案，处理过后的用户搜索文本向量和用户标签已经可以作为分类器的输入了，本章将继续描述分类模型部分的工作。

本文的任务需要对用户的年龄、性别、学历进行分析预测，也就是给用户贴上相应的标签，其中，性别包括有男、女 2 类标签，年龄包括有 0-18 岁、19-23 岁、24-30 岁、31-40 岁、41-50 岁、51-999 岁 6 类标签，学历包括有小学、初中、高中、大学、硕士、博士 6 类标签。前文已经说过，这相当于对用户的搜索文本进行分类，属于哪个年龄类、哪个性别类、哪个学历类。

针对这个情况，本文分别为这三种用户属性构造三个分类器，把数据集划分成训练集和测试机进行交叉验证，并对比分析基于不同的分类算法的分类模型在此任务上的效果差异。

3.4.1 基于朴素贝叶斯的分类模型

在本文第 2.3.1 小节中对朴素贝叶斯算法进行过描述，其理论基础是贝叶斯定理，但条件的独立性即特征属性间互相独立是贝叶斯定理成立的假设前提。为了将朴素贝叶斯算法应用到文本分类任务中，本文采用词袋（Bag of Words）模型，在这种模型中，文本中词语顺序和语法都会被忽略，相当于无序的词汇集合。

假设$x = (w_1, w_2, \ldots w_n)$为一个待分类项，类别集合$C = (y_1, y_2, \ldots y_n)$，从本文第 2.3.1 小节的推导可知，对 x 的分类任务最后转化为求$\underset{k}{\operatorname{argmax}} P\left(y_k\right) \prod_{i=1}^{n} P(w_i|y_k)$。那么，如何求得$P(w_i|y_k)$成了朴素贝叶斯算法的关键问题，不同的$P(w_i|y_k)$的计算方法区分了不同的朴素贝叶斯分类模型，文本分类领域中常见的朴素贝叶斯模型有以下两种：

（1）多项式模型（Multinomial Naive Bayes）

多项式朴素贝叶斯，会为类别集合训练出一个特征分布，该分布由每个类别 y 的θ_y向量组成，$\theta_y = (\theta_{y_1}, \ldots, \theta_{y_n})$，其中，n 是特征项总数（在本文任务中 n 去重后的所有单词的数量），θ_{y_i}是第i项特征(第i个词)出现在属于类y的样本的概率，即$P(w_i|y_k)$。$P(w_i|y_k)$的计算方法在下文会介绍。

多项式朴素贝叶斯在本文任务上的应用过程如下：

1) 假设所预测的用户属性的标签类别集合为$C = (y_1, y_2, \ldots y_m)$, m 为类别总数，对于年龄属性，m=2，对于年龄和学历，m=6。

2) 构造训练数据集$T = (X_1, X_2, \ldots X_n)$，其中$X_i$是类别$y_i$的样本集合，

$X_i = (x_1, x_2, \ldots x_n)^T$，每个$x_i$都是一个属于类别$y_i$的用户搜索文本经过分词和特征处理后的向量表示，即$x_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,n})$。

3) 训练过程中，对于集合 C 中的每个类别y，假设其样本空间为x_k，先计算 P(y)，$P(y) = |X_k| / |T|$；再计算θ_y＝（$\theta_{y1}, \ldots, \theta_{yn}$），由下列公式计算，也就是$P(w_i|y)$：

$$\theta_{y_i} = \frac{N_{y_i} + \alpha}{N_y + n\alpha} \tag{11}$$

其中，$N_{y_i} = \sum_{j=1}^{|x_i|} w_{ji}$，$N_y = \sum_{i=1}^{n} N_{y_i}$，α是一个平滑参数（下称 alpha），取值在[0, 1]区间，α参数的设置是为了防止当$N_{y_i} = 0$时$P(w_i|y) = 0$的情况，一旦$P(w_i|y) = 0$，$\prod_{i=1}^{n} w_i P(w_i|y)$也等于 0，干扰了最终概率的计算，同时 alpha 参数的设置不会引起额外的精度损失。

若文本向量的特征权重是词在文档中的出现次数，上式就可以理解为N_{y_i}表示第i个词在y类的所有文档中共出现过多少次，N_y表示在y类的所有文档中每个词出现次数之和。实践证明，若特征权重选取 tf-idf 能达到相同甚至更优的表达效果。

4) 测试过程，对于待分类项$x = (w_1, \ldots, w_n)$，计算它属于每个类别的概率并取值最大的作为预测结果：

$$\hat{y} = \underset{y}{\operatorname{argmax}} P(y|x) = P(y) \prod_{i=1}^{n} w_i P(w_i|y) \tag{12}$$

其中，P(y)和$P(w_i|y)$已在训练过程中根据训练集数据算出，$\hat{y}$为该待测项的分类结果。

（2）伯努利模型（Bernoulli Naive Bayes）

伯努利模型与多项式模型最大的不同是，伯努利模型输入的样本向量中的每个元素只有两个值：1/0（出现/没出现）。也就是即使输入的文本向量是以 tf-idf 为权重的，伯努利模型依然只是根据该词 tf-idf 的值大于 0 还是等于 0 将其转化成 1 或 0，表示这篇文档包含或不包含这个词。

伯努利贝叶斯在本文任务上的应用过程与多项式贝叶斯基本一致，除了在第 3 步中对$P(w_i|y)$的计算有所不同，伯努利贝叶斯的$P(w_i|y)$的计算如下：

当$w_i = 1$时，$P(w_i|y) = P(w_i = 1|y)$ (13)

当$w_i = 0$时，$P(w_i|y) = 1 - P(w_i = 1|y)$ (14)

$$P(w_i = 1|y) = (N_{y_i} + \alpha)/(|X_k| + 2\alpha) \quad (15)$$

N_{y_i}的计算同多项式贝叶斯中，由于在伯努利模型中，w_i只有 0/1 两种值，表示该词是否出现在某篇文档中，因此，w_i实际上是类 y 下包含该词的文档总数；X_k是类 y 的样本集合，$|X_k|$是类 y 下的文档（样本）总数。这里，α同样是一个为了避免$P(w_i|y)$为 0 的情况而设置的平滑参数，取值在[0, 1]区间。

关于两类朴素贝叶斯模型中 alpha 的取值，本文实验从 0.01 开始以 0.01 为步长试验不同的 alpha 值下贝叶斯分类模型的性能表现。

3.4.2 基于支持向量机的分类模型

在第 2.3.2 节中对支持向量机（SVM）算法的介绍中可以看出超平面的划分方法最初是为了解决两类分类问题，而在本文任务中，除了性别只有男/女两类外，年龄和学历都属于多类分类问题，都分别有 6 个类别。虽然 SVM 分类器的本质是两类分类器，但目前已有不少方法可以将 SVM 扩展到多类分类问题[12]，这些方法大致分为两大类:

（1） 可以通过求解一个最优化问题来实现多类分类，该最优化问题合并了多个分类面的参数求解问题。

（2） 构造并组合一系列的两类分类器来实现多分类器的构造。

相比于较常用的第二类算法，第一类方法计算流程看似简单却难以实现，因为其计算复杂度、参数求解中涉及的变量、运行时间远多于第二类算法，而且并未得到更优的分类结果。本文使用第二类算法，通过 One-against-Rest 策略构造组合多个两类分类器，One-against-Rest 策略依次用一个两类 SVM 分类器将每一类与其它所有类别区分开来，对于 n 类问题，将得到 n 个两类分类器，分类时根据分类函数值大小将其分类。

本文第 2.3.2 节中介绍了对于线性不可分的问题，支持向量机通过引入核函数，可将样本从原始空间映射到一个更高维的特征空间，使得样本在这个特征空间内线性可分。特征空间的好坏对 SVM 的性能至关重要，因此选择哪种核函数是构造 SVM 分类器中的关键问题，若采用了不合适的核函数，会直接导致样本

被映射到一个不合适的特征空间，从而导致构造出来的 SVM 模型性能不佳。

多项式核、线性核、高斯核等是常用的核函数，根据前人的经验，对文本数据通常采用线性核，情况不明时可先尝试高斯核。本文主要的数据类型是文本，因此采用的是线性核。

3.4.3 基于 K-最邻近算法的分类模型

在前面第 2.3.3 节中介绍过，KNN 算法的基本思想是：根据某种空间距离，如欧式距离或曼哈顿距离等，挑选出最靠近待分类项的 K 个样本，看这 K 个样本中哪个类别的样本数最多，就把该待分类项判定为属于那一类。KNN 最大的缺点在于它是一种懒惰学习方法，分类时需要对整个无序的训练集进行对比搜索，而在本文的分类任务中，文本向量维数高，训练样本集数量大，会导致很大的计算成本。

为了解决上述问题，本文使用了改进的 KNN 分类模型：基于 KD-Tree 的 KNN 文本分类算法，下面对其进行介绍。

KD-Tree 是 K 维的二叉查找树，KD-Tree 的每个节点代表 K 维空间的一个点，利用 KD-Tree 可以快速查找给定 k 维数据的最邻近点。并且，树的每一层的分枝决策都以这一层的分辨器为依据[13]。第 i 层的分辨器定义为：i mod k。

KD-Tree 的存储规则为：对第i 层的任意一个节点n，若它的右子树非空，那么节点 n 的i mod k值小于其右子树上的每个节点；若它的左子树非空，那么节点n的第 i mod k维值大于其左子树上的每个节点；并且它的左右子树皆为 KD-Tree。

有了以上对 KD-Tree 的定义介绍，下面给出算法步骤：

（1） 建立一个空的 KD-Tree, 依次将训练集中每个用户的搜索文本向量插入到 KD-Tree 中。

（2） 对于一个待分类文本向量，在构建好的 KD-Tree 中查找该待分类文本向量，获取它的祖先节点集。

（3） 分别计算待分类文本向量与祖先节点集中每个祖先节点的欧几里得距离，距离最近的祖先节点的文本类别就是该待分类文本的文本类别

（4） 对测试集中的每个用户搜索文本，重复第 2、3 步，直至测试集中的每个用户相关属性都计算完毕。

以上算法步骤适用于每个用户属性（学历、年龄、性别），在此就不重复叙述了。

3.5 缺失数据处理

实际的数据集中会存在缺失的标签数据。例如在本文的数据集中，有 355 个样本的年龄标签缺失，424 个样本的性别标签缺失，1878 个样本学历标签缺失。对于这个问题，本文尝试了两种处理方法。

一种做法是认为它是脏数据，认为这条用户数据记录是不可靠的，把它从类别样本数据集中剔除，这样做的好处是能够完全避免这些数据的干扰，但缺点也是明显的，由于这些缺失数据的数目并不在少数，一定程度上减少了用于训练的样本量，降低了分类模型可达到的最大精度。

另一种做法是先利用其他非缺失数据训练出一个分类模型，对这些缺失用户标签的样本进行分类预测，从而对缺失的标签数据进行填补，最后将填补后的数据样本纳入训练数据集中重新训练出新的分类模型。这样做的好处是能够利用起这些缺失的数据，扩大训练集规模，坏处是分类模型的预测准确度是有限的，填补的用户标签并非完全准确，会在一定程度上对最终分类模型的训练造成噪音干扰。

实验发现，对于性别属性而言忽略缺失标签的做法比填补缺失标签的做法有 1%的增益，但对于年龄和学历而言，填补缺失标签的做法比忽略缺失标签的做法提高了 1%的准确率。原因是，性别标签只有两种，样本数量也比较平衡，所以填补缺失标签增加样本的益处小于其噪音干扰带来的坏处。而年龄和学历各有 6 种标签，样本数目也非常不平衡，所以填补缺失标签增加样本的益处大于其噪音干扰的坏处。

3.6 用户属性分类预测算法总体处理框架

经过上文的介绍，本文基于用户搜索关键词的用户属性分析预测算法与总体处理框架如下图所示：

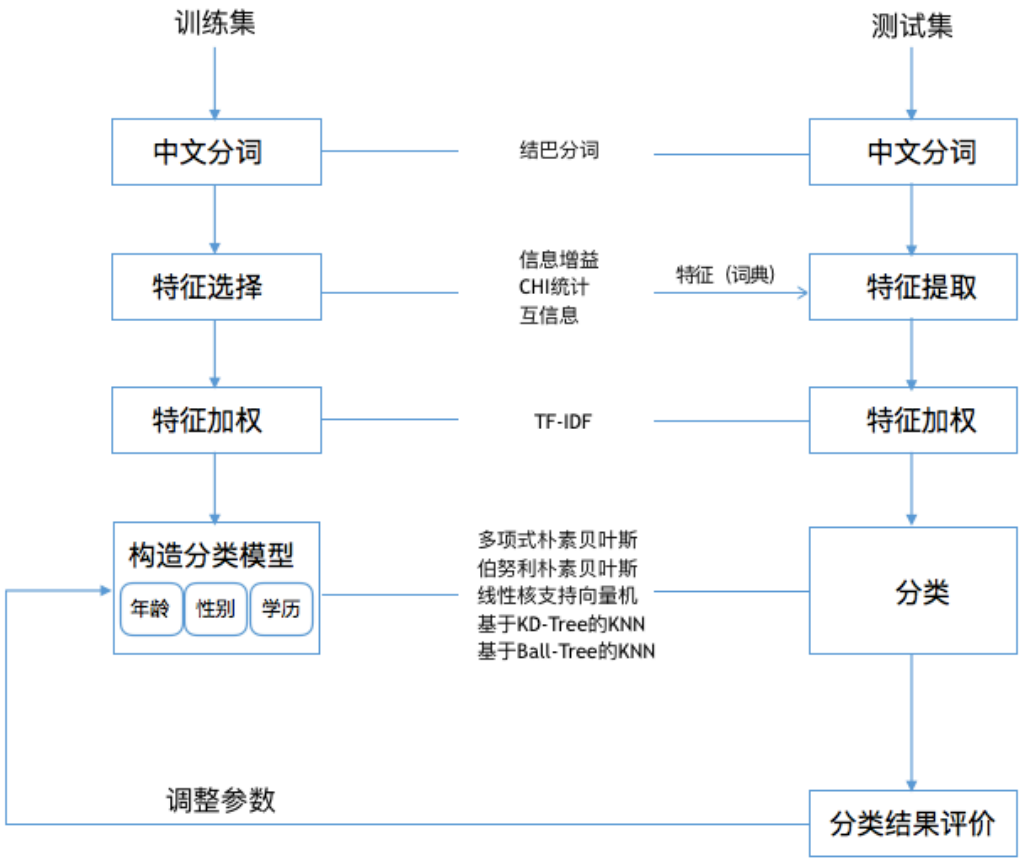

图 3-1 本文算法与总体处理流程框架图

第 4 章 实验与评估

4.1 实验环境

本文实验中的计算环境如下所示：

（1）硬件环境

CPU： 1.6 GHz Intel Core i5

内存： 8 GB 1600 MHz DDR3

SSD 闪存空间： 256GB

（2）软件环境

操作系统：Mac OS 10.11.6

开发语言：python

开源工具：Scikit-learn, Liblinear, Libsvm, jieba

4.2 数据集

本文实验中所使用的数据集来源于搜狗，数据集共有 20000 条记录，其中每条记录都是由 5 个字段组成的，前四个字段是用户的信息，分别是用户的 ID、年龄（Age）、性别（Gender）、学历（Education），第 5 个字段是该用户历史一个月的搜索记录（Query List），字段的详细情况见表 1。数据格式示例如下：

C98C47F106	1	1	5	空格符号复制临沂是哪里lol镇魂街拉棺图片纹身视频合成app 长颜草
9D8C36B494	2	2	4	选中不连续的段落并为其套用样式微微一笑很倾城你是我学生又怎样

经过观察发现，数据集中存在较为严重的样本空间分布不平衡情况。在性别属性中，男性样本有 11365 个，女性样本有 8211 个，男性样本比女性样本多了三千多个样本。在年龄类别分布中，随着年龄的增加样本数据量逐渐甚至急剧减少，0-18 岁有 7900 个样本，19-23 岁有 5330 个样本，24-30 岁有 3603 个样本，31-40 岁有 2141 个样本，41-50 岁有 589 个样本，51-999 岁仅有 82 个样本。学历样本的分布与年龄一样严重，甚至更差，初中学历的有 7487 个样本，高中有 5579 个样本，大学有 3722 个样本，小学有 1150 个样本，硕士有 119 个样本，

博士仅有 65 个样本。这种情况是相应分类器性能不佳的直接原因。

表 1 字段详细情况

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

4.3 分类模型性能分析与比较

本文对分类结果的评估涉及到准确率与召回率两个指标，并以准确率为主要评价手段，下面给出它们的计算方式。

$$准确率=\frac{被正确分类到类别\ y\ 的样本数}{实际上属于类别\ y\ 的样本数}\times 100\%$$

$$召回率=\frac{被分类到类别\ y\ 的样本数}{所有被分类到类别\ y\ 的样本数}\times 100\%$$

$$平均准确率=\frac{被正确分类的样本数}{所有测试样本数}\times 100\%$$

为了不复杂化结果，这里给出当训练集样本数量与测试集样本数量的比例是 6:1 时的分类结果比较分析。

4.3.1 特征选择对分类模型性能的影响

这里，选取的性别属性作为预测对象，分类模型是多项式朴素贝叶斯（平滑参数取0.01），分类模型是“定量”，以特征选择算法以及特征数量作为“变量”，观察分析不同特征选择算法、不同特征数量下分类模型在预测用户性别时的平均准确率趋势。

图4-1中，横轴是特征数量，纵轴是平均准确率，蓝色折线代表CHI2统计，绿色折线代表IG统计，红色折线代表MI。

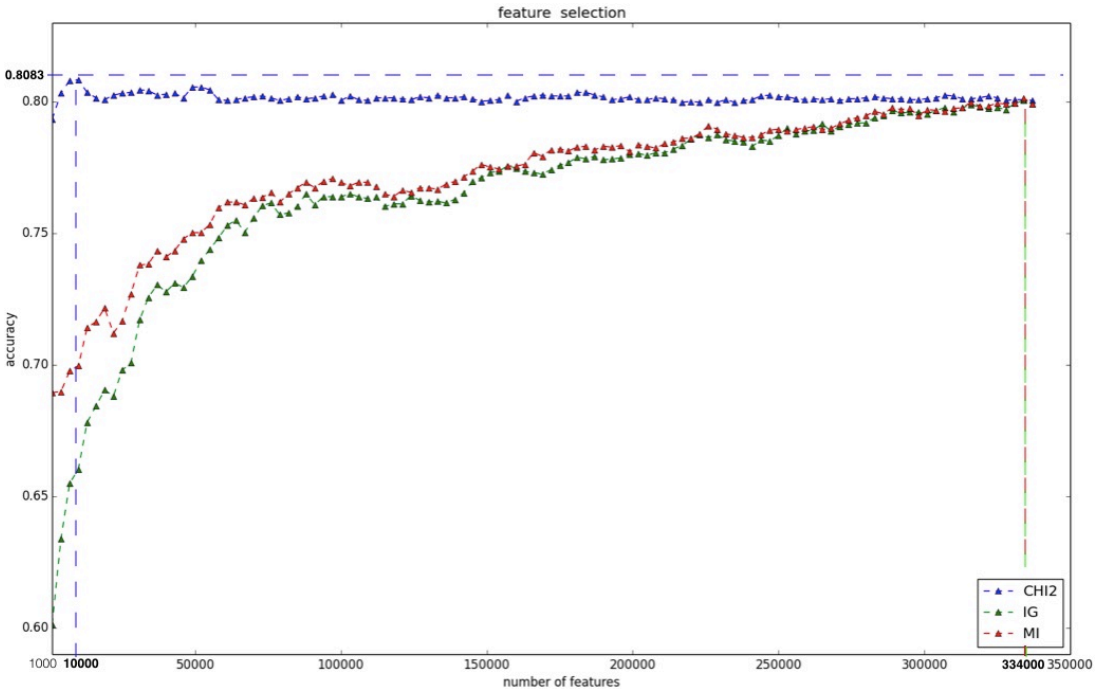

图 4-1 特征选择算法、特征数量对性别分类平均准确率的影响

由图4-1可以看出，分类器的性能随着特征选择的数量的增加，先大幅增长，随后到达顶峰，之后略微下降，总体趋势是：

（1）在特征数量较少的情况下，不断增加特征的数量，会显著提高分类器的性能，呈现“上升”趋势；

（2）随着特征数量的不断增加，将会引入一些不重要的特征，甚至是噪声，此时再增加特征数量已难以提高分类器性能，因此，分类器的性能将会呈现略微“下降”并趋于平稳的趋势。

CHI统计达到的平均分类准确率峰值是80.83%，MI和IG都是80.04%。虽然这三种特征选择算法下平均分类准确率的峰值只是相差0.79%，但是CHI统计在特征数量为10000时，平均分类准确率就已升到最高值，但IG和MI要在特征数量高达33000左右时分类平均准确率才能达到最高值，而原始的特征数量是349000。这说明了，在本文数据集上，与互信息以及信息增益方法相比，CHI统计方法具有巨大的优势，CHI统计方法更能提取出最重要的特征。

综合来看，图4-1的分类器性能走向趋势体现出了特征选择的重要性：选择出具有代表性的特征，并降低噪声，可提高分类算法的泛化能力。

4.3.2 朴素贝叶斯的伯努利模型和多项式模型

这里对朴素贝叶斯的两种模型做实验分析来预测用户属性，分别是伯努利模型与多项式模型，图4-2、图4-3、图4-4分别是年龄、性别、学历的平均分类准确率折线图，特征集是由CHI2统计法选出来的10000个特征。横轴是平滑参数(称为alpha)，纵轴是平均准确率，蓝色折线是多项式贝叶斯的预测结果，绿色折线是伯努利贝叶斯预测结果。

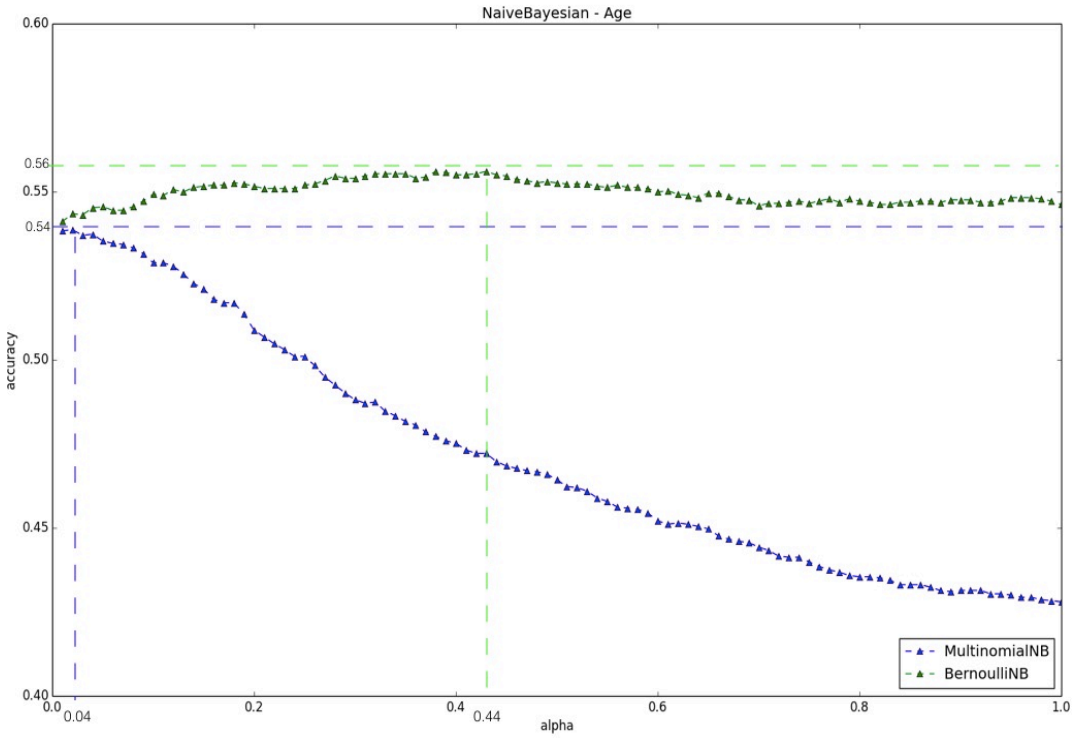

图 4-2 alpha 值对两种 Naïve Bayesian 模型的性别平均分类准确率的影响

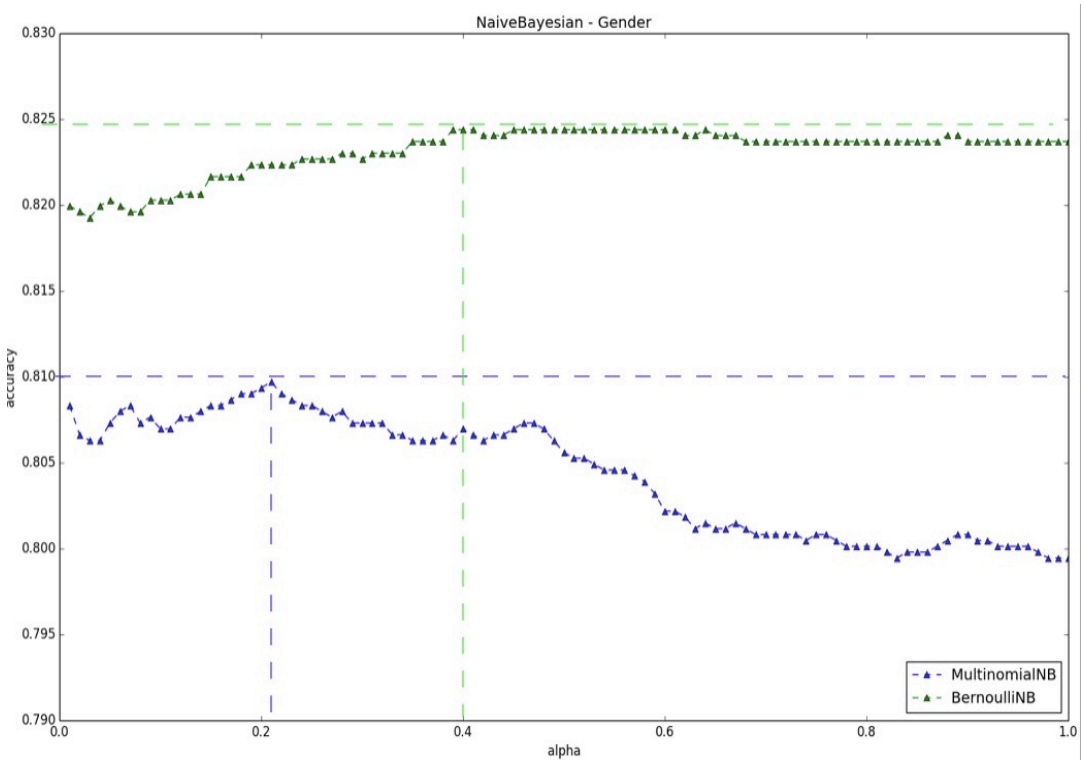

图 4-3 alpha 值对两种 Naïve Bayesian 模型的性别平均分类准确率的影响

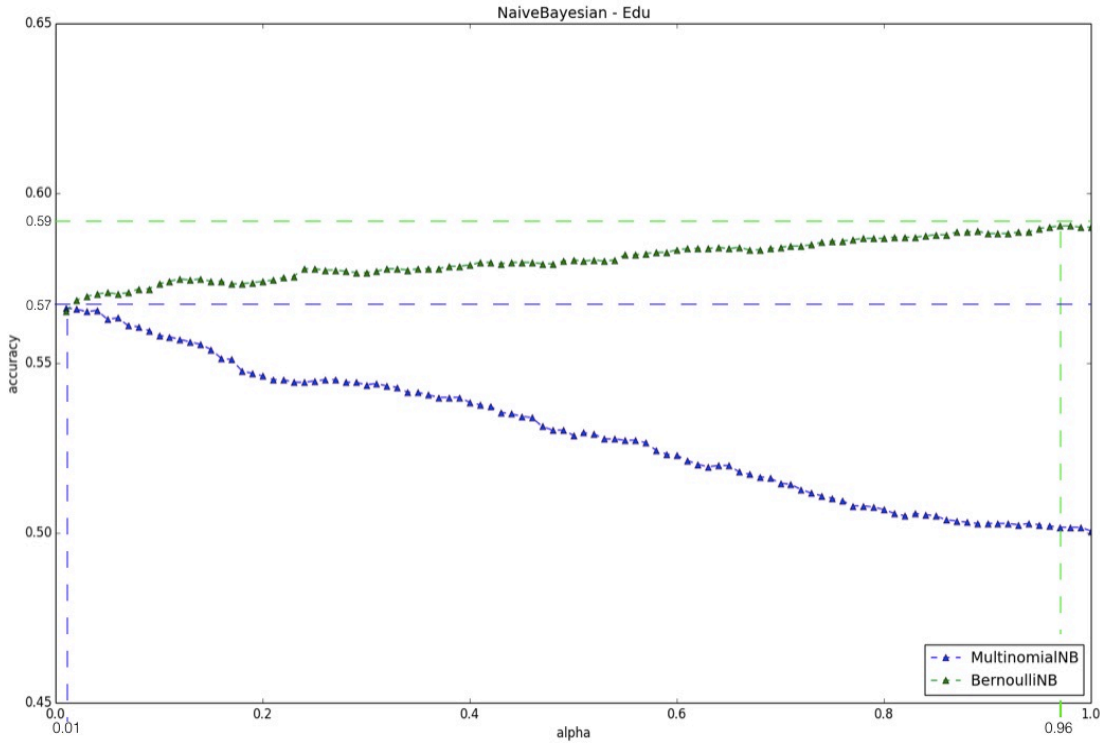

图 4-4 alpha 值对两种 Naïve Bayesian 模型的学历平均分类准确率的影响

综合以上三张实验结果图，可以看出：

（1）在本文分类任务中，伯努利贝叶斯的性能比多项式贝叶斯更优；

（2）alpha的大小能影响两种贝叶斯分类模型的预测精度，对多项式贝叶斯的影响尤其大。

（3）预测精度一开始随着alpha值的增大而波动上升，到达峰值后，随着alpha值得增大而波动下降，选取最优的alpha值对构造贝叶斯分类器意义重大。

4.3.3 KNN 分类器中 K 值的选取

下图4-5是使用KNN分类器来来预测用户的年龄、性别、学历得到的分类平均准确率折线图，特征集是由CHI2统计法选出来的10000个特征。横轴是K值（KNN算法选取K个最邻近邻居来做类别统计），纵轴是平均准确率，蓝线代表性别，红线代表学历、绿线代表年龄。

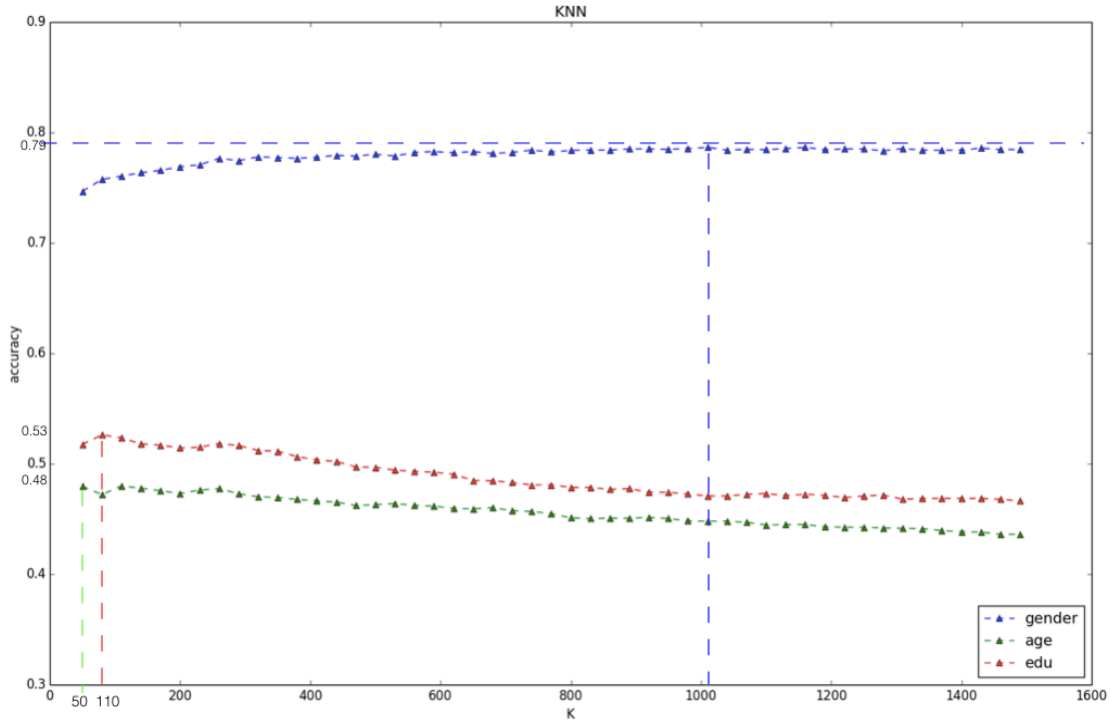

图 4-5 K 值对 KNN 分类模型的影响

由图4-5可以直观地看出：

（1）KNN对性别的预测准确度远高于年龄和学历；

（2）对学历的预测准确度在K为80时达到最大，对年龄的预测准确度在K为110时达到最大，之后随着K值的增加，学历和年龄的预测准确度有着明显的

下降趋势。而性别的预测准确度在K在1160时达到最大，之后随着K值的增加虽有略微下降，但并不明显。

首先，性别只有2类，这两类样本数量分布并不悬殊且数量较大，但性别和学历都各有6类且分布较为悬殊。性别和学历这种有限的样本数量以及样本空间的极不平衡的数据集情况是造成它们的平均准确率不佳的根本原因。

此外，上述样本空间不平衡情况导致了性别和学历的准确率达到峰值后随着K值的增加有非常明显的下降，因为当K值变大时待测样本的K个邻居中占多数的更容易是更大容量类的样本。而性别的样本空间分布相对均衡，所以其预测准确率并未随着K值的增加产生明显的下降趋势。

4.3.4 各分类模型间的对比

下表2中记录的是：在三种构建好的分类模型下，三种用户属性里每种类别的分类准确率和召回率。

表 2 各类别在不同分类模型下的准确率和召回率

			朴素贝叶斯		支持向量机		K-最近邻	
属性	类别	训练样本数	准确率	召回率	准确率	召回率	准确率	召回率
性别	男	9670	86%	84%	84%	85%	79%	85%
	女	6979	78%	81%	79%	77%	77%	70%
学历	小学	987	31%	21%	36%	7%	33%	1%
	初中	6371	68%	75%	67%	79%	55%	84%
	高中	4754	51%	45%	49%	48%	42%	39%
	大学	3161	56%	62%	57%	55%	58%	23%
	研究生	98	0	0	0	0	0	0
	博士	58	0	0	0	0	0	0
年龄	0-18	6746	66%	74%	65%	80%	52%	83%
	19-23	4492	58%	47%	55%	51%	43%	29%
	24-30	3083	41%	53%	43%	42%	37%	32%
	31-40	1822	33%	29%	31%	21%	28%	8%
	41-50	492	14%	2%	18%	2%	17%	2%
	51-99	66	0	0	0	0	0	0

其中，分类模型方面，朴素贝叶斯分类模型选取的代表是表现更为优秀的伯努利朴素贝叶斯分类模型，SVM分类器使用线性核，K-最近邻分类器选取的是基于KD-Tree的改进分类方法。

参数方面，朴素贝爷斯的alpha值以及K-最近邻分类模型的K值也已调整到最佳值。由于各种原因，有些分类器完全没有预测出某种类别的样本，如研究生和博士，准确率为0，召回率的分母为0无法计算，这部分召回率在表中登记仍然登记为0。

表2中数值为0的准确率和召回率，基本上是来自学历中的研究生和博士类别，以及年龄中的51-99岁类别。它们的共同点是样本数量极低，17000条训练记录中，硕士、博士仅分别占98、58条，51-99岁的记录也仅有66条。

此外，年龄中41-50岁类别的准确率和召回率也普遍比较低，虽然41-50岁类别在数据集中分别有492，比51-99岁要多出许多，但比起其他大容量类别，41-50岁的样本数目还是相当少。

性别属性中，分类模型在男性类别的表现均优于女性类别，原因还是在于样本数量不均，男性类别以及女性类别的样本数量都非常大，这是它们取得较高预测精度的原因，但同时存在男性类别的样本数比女性类别多出三千左右，这种样本不均衡使得女性类别的分类精度差于男性类别。

总体来看，在本文的分类任务中，伯努利朴素贝叶斯与支持向量机分类模型表现较为优秀，K-最近邻分类模型稍微逊色。并且，若一个类别训练样本数越多，它与其他类别的训练样本数量越均衡，那么该分类器分类结果越好。

通过以上的实验结果表明，针对根据用户搜索关键词预测分析用户属性这一课题，本文所研究提出的基于文本分类技术的解决方案是可行且有效的。特征数量、特征选择算法、参数的设置都对分类模型的性能有较大影响。学历及年龄的样本空间分布极其不均匀，导致了相应分类模型性能不佳，训练样本越多、各类样本数量越均匀，分类结果越好。

第 5 章 总结与展望

5.1 本文工作总结

现代互联网广告有着迫切的“精准投放”需求，其中，关键的问题是如何找到合适的投放人群。用户画像挖掘技术中，根据用户的浏览、搜索等行为来反推获取用户属性是一项非常基础且重要的技术。本文的研究内容是基于用户搜索行为与用户属性的相关性对用户属性进行分析预测，由于用户搜索记录是文本数据，用户各属性的标签都是离散值，因此本文采用文本分类的相关技术来分析和预测用户属性。

本文从中文分词、特征选择、特征加权、构建分类模型等多个方面着手，至下而上搭建了一套分类算法和处理过程来对用户属性进行判别。通过对分类算法的实验结果分析，证明了本文采取的分析预测方法是有效可行的，同时对比了采用不同的特征选择方法、不同的分类模型对分类器性能的影响。

本文的主要工作内容有以下几点：

（1）调查并研究了几种经典的分类算法。由于本文是根据用户搜索记录文本对用户进行分类，故而主要考虑了在短文本分类领域表现良好的分类算法。主要是基于概率统计知识的朴素贝叶斯算法，基于最优超平面划分的支持向量机算法以及基于向量距离比较的 K 最邻近算法。

（2）研究了中文文本预处理的基本流程和方法，以及相关特征选择和特征加权的算法。原始的文本不能直接应用到分类模型中，需要做相应的中文分词、特征选择、特征加权等工作。本文采用“结巴分词”工具对原始文本做分词处理，并采用了 TF-IDF 文本特征加权方法。特征选择方面，采用了 CHI 统计、信息增益、以及互信息四种特征选择方法，同时在实验中采用并比较了多种特征选择方法对分类模型性能的影响。

（3）为了提高各分类模型的性能，研究了各分类算法的改进方法，比如，基于 KD-Tree 的改进的 KNN 算法，使用线性核的支持向量机等。

（4）针对构建分类模型时的参数设置问题，研究了参数值设置方法，同时会依据分类结果的反馈适当调整参数的值。

（5）设计实现了完整的基于用户搜索关键词的用户属性分析预测算法，给

出了分类算法的总体框架，并通过实验分析比较了各种情况下的分类效果，包括不同特征数量、不同特征选择方法、不同分类算法、不同参数值下的分类模型性能优劣。

5.2 进一步的工作与展望

本文的工作在取得一定成果的同时依然有许多方面需要得到进一步的改善:

（1）中文分词步骤中，由于中文词汇系统较为复杂，没有做同义词的转换和错别字的纠正，可能会导致分类性能上的一些损失。

（2）样本空间不平衡造成的分类模型性能不佳的问题尚未得到很好解决，虽然分类器在具有大数量类别上取得了较高的准确率和召回率，但小数量类别的分类结果较差。过采样的方法虽然能增加样本数，但却会破坏原本样本空间的结构，可能会降低分类器的性能。下一步工作应该尽量采取手段，增加小数量类别的样本数，亦可尝试文献[2]中提到的 one-class svm 的方法。

（3）分类模型参数的设置上，如朴素贝叶斯中的平滑参数和 KNN 中的 K 值，都是基于经验和逐步的观察来调整的，可以采用更具理论基础的方法进行最优参数搜索与设置。

（4）本文使用的数据集质量存在一定问题，因为存在多人共用一个搜狗账号的现象，如一个家庭的成员都通过同一个搜狗账号进行搜索。针对这个问题，可以提前对数据集进行清洗，人工过滤掉不合理的样本。

参考文献

[1] 高洁, 吉根林. 文本分类技术研究[J]. 计算机应用研究, 2004, 21(7):28-30.

[2] David M.J.Tax,Robert P.W.Duin. Support vector domain description[J].Pattern Recognition Letters,1999,20:1191-1199.

[3] 牛玲. 一种基于向量空间模型的改进文本分类算法[J]. 情报杂志, 2006, 25(6):63-64.

[4] 张浩, 汪楠. 文本分类技术研究进展[J]. 科技信息:科学·教研, 2007(23):99-100.

[5] 张少宏, 李继巧, 罗嘉怡,等. 基于信息融合的网页文本聚类距离选择方法[J]. 广州大学学报(自然科学版), 2016, 15(1):80-89.

[6] 刘辉, 应培培. 一种改进的 KNN 文本分类算法[J]. 信息安全与技术, 2011(7):25-27.

[7] 张冬生. 支持向量机在分类问题中的应用研究[J]. 黑龙江科技信息, 2010(35):64-64.

[8] Vapnik V N. An overview of statistical learning theory [J].IEEE Trans Neural Network,1999,10(5):988-999.

[9] Guo G, Wang H, Bell D, et al. KNN Model-Based Approach in Classification[J]. Lecture Notes in Computer Science, 2003, 2888:986-996.

[10] Y.Yang.A Comparative Study on Feature Selection in Text Categorization. In: Proceeding of the Fourteenth International Conference on Machine Learning (ICML'97),412- 420,1997.

[11] 周茜, 赵明生, 扈旻. 中文文本分类中的特征选择研究[J]. 中文信息学报, 2004, 18(3):17-23.

[12] 刘志刚, 李德仁, 秦前清,等. 支持向量机在多类分类问题中的推广[J]. 计算机工程与应用, 2004, 40(7):10-13.

[13] 刘忠, 刘洋, 建晓. 基于 KD-Tree 的 KNN 文本分类算法[J]. 网络安全技术与应用, 2012(5):38-40.

致谢

本文研究工作的顺利完成离不开导师黄宜华老师和学长朱光辉对我的悉心指导与监督。首先，非常感谢黄宜华老师帮助我确定毕业设计研究课题，并指导我进行论文编写，提出大量修改意见。其次，非常感谢朱光辉学长在刚开始进行毕业设计时向我提供大致研究思路，并及时解答我在完成毕业设计过程中遇到的各种问题。

此外，感谢所有给予我鼓励和帮助的同学，感谢肖鹏、陆蓓蓓等同学在我论文撰写过程中，帮助我解决许多编辑工具使用、文档格式转换、论文内容规范等方面的问题。

通过本次毕业设计，我得到了许多成长，自己独立完成工作的能力、调研资料的能力、代码编写的能力等等，都得到大大提高。所以非常感谢学校开设毕业设计这样的课题来进一步检验和锻炼我们的研究和创新能力。同时，也要感谢大学四年教导我的各位老师，您们传授的知识是我完成毕业设计的基础，您们渊博的知识和严谨的学术研究作风都给我留下深刻的印象，今后也会继续向您们学习。

最后，再次向所有在大学四年里给予我关心和帮助的人们表示由衷的感谢！