基于多分类系统的文本挖掘应用

摘要：网络问政平台逐步增长的情况下，一方面，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战；另一方面，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势。

本文围绕智慧政务系统及其若干关键模型与挖掘算法进行了一系列研究，主要研究内容和研究成果如下：

1、针对附件二留言信息的分类，我们将它定义为文本多分类问题。首先我们对附件二的留言数据进行了清洗，其次创建了停用词词库进行去停用词，同时计算了每条留言信息的 TF-IDF 向量，然后查找与每种类别最为相关的词条。完成上述所有的数据转换后，我们对逻辑回归、朴素贝叶斯、线性支持向量机及随机森林分类器对文本分类的准确性进行了评估，最终提出一种有效的融合模式。

2、针对附件三热点问题的挖掘，我们将它定义为文本聚类的问题。首先我们对附件三的留言数据进行了清洗，然后对留言时间进行了切分处理，并对留言主题使用了 jieba 进行分词。其次创建了停用词词库进行去停用词，生成 TF-IDF 矩阵文档并将其转为了数组形式。最后我们对 K-means 聚类和 DBSCAN 聚类进行了比较，最终选取了效果较好的 K-means 聚类算法得出了热点排名前五的问题。

3、对于第三题，我们将它定义为自动评分体系的实现。我们首先对群众留言详情进行特征提取和权重选取，构建一个类似考试的标准答案的框架，然后再根据、分词、特征及其权重和匹配规则来构建一个知识库，将知识库运用到评价意见数据集上，使得评价意见可以自动分词，并基于构建的框架进行推理，实现评价意见与留言详情的匹配评分，以此来对评价意见进行评价。

本文采用文本多分类方法实现留言信息的分类，采用均值聚类算法实现热点问题的挖掘以及对刘艳信息进行特征提取和权重选取来实现对评价意见进行评价，采用的都是传统的方法，在交叉信息和无效文本处理上有待提高，但都达到了预期目标，能应用于多分类系统的文本挖掘。

关键词： 智慧服务；文本多分类；文本聚类；分类器；K-means 算法；LDA 模型

Application of text mining based on multi classification system

Abstract: With the gradual growth of the online questioning platform, on the one hand, the amount of text data related to various social conditions and public opinion has continued to rise, which has brought great challenges to the work of relevant departments that used to manually divide messages and organize hotspots. On the other hand, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of smart government systems based on natural language processing technology has become a new trend of social governance innovation and development.

This article has conducted a series of studies around the smart government system and its several key models and mining algorithms. The main research contents and research results are as follows:

（i） For the classification of message information in Annex II, we define it as a text multi-classification problem. First of all, we cleaned the message data of Annex II. Secondly, we created a stopword vocabulary to remove stopwords. At the same time, we calculated the TF-IDF vector of each message information, and then looked for the most relevant entry for each category. . After completing all the above data conversions, we evaluated the accuracy of text classification by logistic regression, naive Bayes, linear support vector machine and random forest classifier, and finally proposed an effective fusion model.

（ii） For the mining of the hot issues in Annex III, we define it as the problem of text clustering. First, we cleaned the message data of Annex III, and then divided the message time, and used jieba to segment the message topic. Secondly, a stopword vocabulary was created to remove stopwords, generate a TF-IDF matrix document and convert it into an array form. Finally, we compared K-means clustering and DBSCAN clustering, and finally selected the better-performing K-means clustering algorithm to get the top five hot spots.

(iii) For the third question, we define it as the realization of an automatic scoring system. We use the message details of the masses as the standard answer, let the computer understand the natural language in the sentiment of the message, and then use the evaluation opinion as the answer to the reply, that is, the answer equivalent to the candidate’s answer, and match the evaluation opinion with the message details to score To evaluate the relevance, completeness and interpretability of the evaluation opinions.

In this paper, the text multi classification method is used to realize the classification of message information, the mean clustering algorithm is used to realize the mining of hot issues, and the feature extraction and weight selection of Liu Yan's information are used to realize the evaluation of the evaluation opinions. The traditional methods are used, which need to be improved in the cross information and invalid text processing, but all achieve the expected goal, and can be applied to multi classification system Unified text mining.

Keywords: smart service; text multi-classification; text clustering; classifier; K-means algorithm; LDA model

目录

1. 问题重述

随着网络的普及和高速发展，互联网成为在我们的日常生活中越来越重要的角色，公民可以在各式各样的平台上行使各项权利。随着微信、微博、市长信箱、阳光热线等网络问政平台上线的环境下，群众可以通过问政平台反映涉及民生的各类问题，也可以对政府各项工作提出意见和建议，网络成为了民众参政议政的新渠道。然而面对几乎呈指数级增长的留言,纯人工归类会带来巨大的工作量，从而影响留言的时效和解决问题的效率。因此在大数据、云计算、人工智能等技术高速发展的背景下，建立基于自然语言处理技术的文本多分类已经是进行文本分类的新趋势。

本研究的主要内容为：结合已有的文本分类标签和问政平台的群众留言信息数据，利用算法挖掘与建立模型对以下三方面问题进行解决：

(1)、留言分类问题。根据已有的留言体系，使用各类模型以及它的算法对问政平台上的群众留言信息进行类别归属，并给留言贴上分类标签。

(2)、热点挖掘问题。对留言信息相似度高的文本进行分类打包，再结合点赞和反对信息设计算法，给出排名前五的热点问题。

(3)、评价答复信息问题。对留言详情进行特征选择后构建框架和知识库，再对评价意见基于框架进行推理，对评价意见进行评分，以此来解决对评价意见的评价。

2. 模型假设

针对实际的留言分类和热点挖掘而言，为了获得精准的分类和挖掘，我们可以做出如下合理的假设：

(1)假设所给的数据集是完备可靠的；

(2)假设所给的数据具有适用性和实效性。

3. 数据预处理

数据预处理是指在进行主要的算法及构建模型以前对数据进行的一些处理。原始的数据存在数据缺失、数据冗余等问题，不进行处理会严重影响数据的挖掘

质量以及模型的预测效果。因此，必须先对给定的原始数据进行数据预处理操作，以此提高数据集的质量，使得建立的模型更为准确。

3.1 数据选取

在题目给出的数据中，包含了多个文件，每个文件中又包含了群众留言的许多信息。例如留言编号，留言用户，留言主题，留言详情等，我们需要从中选择出我们所需的数据来进行处理。在处理留言分类问题上，我们选取了附件二中的留言详情和标签分类这两个数据集；在处理热点挖掘问题上，我们选取了附件三中的留言主题和留言时间这两个数据集；在处理答复评价问题上，我们选取了附件四中的留言详情和答复意见这两个数据集。

3.2 数据清洗

3.2.1 缺失数据的分析与处理

缺失的概率是随机的，比如群众可能点击错误而导致留言信息为空。数据缺失也与自身的值有关，比如群众可能对每个主题所要填写的信息并不明确，导致信息填制空缺等。在本题中我们首先对选取的留言信息是否存在缺失数据进行了分析，分析得到在本次所给出的留言信息数据集中我们所需要用到的数据集信息完备，没有缺失值的存在。

3.2.2 异常数据的分析和处理

异常数据也称之为离群点。离群点是远离数据集中部分的分散数据，它在很大程度上影响着对结果的准确性。因此，我们需要对异常的数据进行处理。我们对留言信息进行了正则化处理。发现对于留言分类问题中的附件二，多条留言详情中包含冗余的无效信息。对于热点挖掘问题中的附件三，因为我们选取的是留言主题的数据集，因此发现的异常点可以通过留言详情中的信息来进行修正，修正前后见表 3-1。

表 3-1 留言详情处理对比示例表

留言详情特殊字符处理对比	
处理前	处理后
\n\t\t\t\t\t\n\t\t\t\t\A3 区大道西行便道，未管所路口至加油站路段，...	A3 区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房..

\n\t\t\t\t\t\n\t\t\t\t\t 位于书院路主干道的在水-方大厦--楼至四楼人为..	位于书院路主干道的在水--方大厦一楼至四楼人为拆除水、电等设施后,烂尾多年,用护栏围着，不但占..
\n\t\t\t\t\t\n\t\t\t\t\t 尊敬的领导:肚区苑小区位于 AI 区火炬路，小...	尊敬的领导: A1 区苑小区位于 A1 区火炬路，小区物业 A 市程明物业管理有限公司，经小区业主同意...
\n\t\t\t\t\t\n\t\t\t\t\tA1 区 A2 区华庭小区高层为二次供水，楼顶水箱..	A1 区 A2 区华庭小区高层为二次供水，楼顶水箱长年不洗，现在自来水龙头的水严重霉味，大家都知道..

3.3 分词

无论是有关留言分类的文本分类问题还是有关热点问题挖掘的文本聚类问题，或者是评价答复意见的问题，我们都首先需要对文本做分词处理，例如附件二中我们所要用到的留言详情“A3 区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房项目施工围墙内。每天尤其上下班期间这条路上人流车流极多，安全隐患非常大。强烈请求文明城市 A 市，尽快整改这个极不文明的路段。”我们希望将其切分为“A3 区 大道 西行 便道 未管所 路口 至 加油站 路段 人行道 包括 路灯杆 被圈 西湖 建筑 集团 燕子山 安置房 项目 施工 围墙内 每天 尤其 上下班 期间 这条路上 人流 车流 极多 安全隐患 非常 大 强烈 请求 文明城市 A 市 尽快 整改 这个 极 不文明 的 路段。”在进行分词时，我们选用了jiaba中适合文本分析的精确模式来进行分词，它会试图对文本信息进行最精确的分词。部分分词结果如图 3-1 所示：

A3	区	一米阳光	婚纱	艺术摄影	是否	合法	纳税
咨询	A6	区	道路	命名	规划	初步	成果
反映	A7	县	春华	镇金鼎村	水泥路	、	自来水
A2	区	黄兴路	步行街	大	古道	巷	住户
A	市	A3	区	中海	国际	社区	三期
A3	区麓	泉	社区	单方面	改变	麓	谷
A2	区富	绿	新村	房产	性质	是	什么
对	A	市	地铁	违规	用工	问题	质疑
A	市	6	路	公交车	随意	变道	通行
A3	区	保利	麓	谷林语	桐梓	坡路	与麓
A7	县	特立	路	与	东四	路口	晚
A3	区	青青	家园	小区	乐果	果	零食
关于	拆除	聚美龙楚	在	西地省	商学院	宿舍	旁
A	市利保	壹号	公馆	项目	夜间	噪声	扰民
A	市	地铁	3	号线	星沙	大道	站
A4	区	北辰	小区	非法	住	改商	问题
请	给	K3	县	乡村	医生	发	卫生室
A7	县	春华	镇	石塘	铺村	有	党员
咨询	异地	办理	出国	签证	问题		

图 3-1 分词情况示例图

3.4 去停用词

在留言信息中，如果不加处理就对数据集进行算法计算或构建模型，那么会因为很难单独表达文档相关信息的停用词，降低运行的速度和结果的准确性。因为每条留言信息中的停用词很难单独表现出有关留言分类标签的特征。停用词包括语气助词、副词、介词、连接词等，他们通常单独看来并没有较为明确的意义，只有将其放入一个完整的句子中才有一定作用，如常见的“的”、“这”、“那”之类。因此为了增加后续处理文档数据的准确性，减少对留言信息中的有效信息形成的干扰，使留言信息中有关分类标签的特征更集中突出，我们必须对留言信息数据集进行消除噪音的处理，即去停用词。本文采用的停用词表是网络课程中所借用的停用词表，该表中部分停用词表见表 3-2 所示，然后根据停用词表进行了去停用词。

表 3-2 部分停用词表

部分停用词表													
】	8	exp	② c	⑥	——	〈	」	〔	一.	一何	一定	一样	一致
【	9	sub	③	⑦	■	〉	『	〕	一一	一切	一方面	一次	一般
,	:	sup	③]	⑧	▲	《	』	〕〔	一下	一则	一旦	一片	一起
!	://	|	④	⑨	。	》	【	㈧	一个	一则通过	一时	一番	一转眼
会	::	}	⑤	⑩	、	》),	】	一	一些	一天	一来	一直	一边

3.5 创建字典对象保存类标签

在有关本题的留言分类中，以每一条留言都作为主体，作为主体的留言详细信息中具有很多种属性，每种属性又有不同的取值，并且属性的数量和属性取值的数量都是在不断变化的，特别是当这些数量的变化很快时，就应该考虑引入数据字典的设计方法。数据字典是各类数据描述的集合 ，它是进行详细的数据收集和数据分析所获得的主要结果。所以在本题中基于已有的附件一所有的留言信息的分类标签，我们需要创建字典对象并保存分类标签，为后续的留言分类提供可行性。

3.6 判断类别是否平衡

如果每个留言类别中包含的留言信息条数有较大差异，那么在进行算法运算和构建模型时会产生较大的误差，甚至极个别的信息会作为异常值被忽视，因此我们需要对类别的平衡事先做出判断，观察数据是否需要进行修正，每个留言标签类别下的留言条数如图 3-2 所示，可以看出虽然留言类别有一定的差距，但并不影响整体的运算。

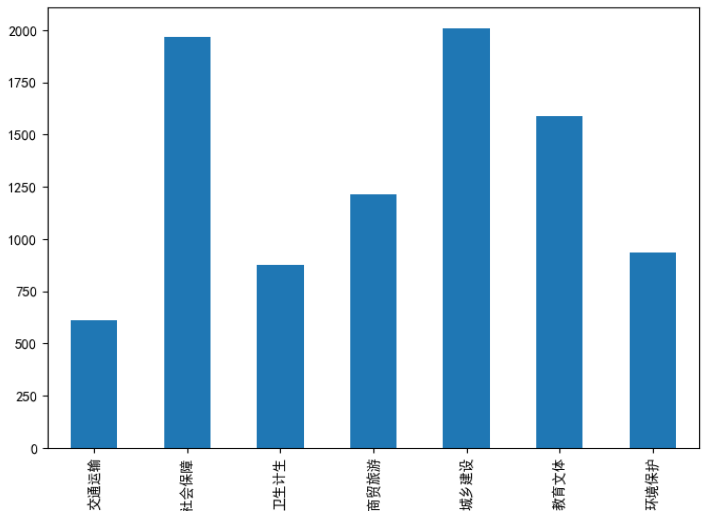


图 3-2 主题占比图

4. 文本表示

文字是人类在认知过程中产生的高层认知抽象实体，文本表示是指将实际的文本内容转变为机器可读的表示结果。文本表示包括表示和计算两个方面，表示是指特征的提取，计算是指权重的定义和语义相似度的定义。挖掘算法与构建模型无法直接对文本的原始形式做处理，因为它们期望的输入是长度固定且为数值型的特征向量，而不是具有可变长度的原始文本。因此，在进行挖掘算法和构建模型之前，我们要对数据进行文本表示处理。为提取留言信息中的特征，我们构建了向量空间模型，选取了模型中较为适合进行文本分类的常用的构造 TF-IDF 词袋模型的方法。其中 TF 是term frequency的缩写，在本次研究中表示的是在某条留言信息中含有的某一个词语在词条留言信息中出现的次数，这个数字通常都会用词频除以文章总词数来进行归一化，以防止它偏向过长的文件。而 IDF 逆向文件频率是inverse document frequency的缩写，它反应了某条留言信息中的某个词在所有留言详情这列信息中出现的频率。如果一个词在很多的文本中出

现，那么它的 IDF 值就低。那么反过来如果一个词在比较少的文本中出现，那么它的 IDF 值就高。

TF 的计算公式如下：

$$TF_{\omega} = \frac{N_{\omega}}{N} \tag{4.1}$$

其中N_{ω}是在某一文本中 ω 词条出现的次数，N是该文本总词条数。

IDF 的计算公式：

$$IDF_{\omega} = log\left(\frac{Y}{Y_{\omega}+1}\right) \tag{4.2}$$

其中Y是语料库的文档总数，Y_{ω}是包含词条 ω 的文档数，分母加一是为了避免 ω 未出现在任何文档中从而导致分母为 0 的情况。

TF 判断的是某条留言信息中该字或词语是否是当前留言信息中的重要词语，但是如果只用词语出现频率来判断，那么可能会出现一个问题，就是有些无意义的通用词可能也会出现很多次，如：的、我们、在等。当然我们对文本进行预处理的时候一般会去掉这些停用词，但仍然会有很多通用词无法避免地出现在很多文档，而它们并不是那么重要。IDF 判断的是在多条留言信息中是否都出现了此词语，即多条留言信息中都出现的就是通用词，它实质上是抑制通用词的重要性。那么将上述求出的 TF 和 IDF 相乘就得到该词语在当前留言信息和整个留言信息数据集的相对重要性。TF − IDF的公式如示：

$$\mathrm{TF}-IDF_{\omega} = TF_{\omega} * IDF_{\omega} \tag{4.3}$$

从以上计算公式便可以看出，某一特定留言信息中的高词语频率，以及该词语在整个留言信息数据集的低文件频率，可以产生出高权重的TF − IDF。TF − IDF与一个词在当前文档中的出现次数成正比，与该词在整个语料库中的出现次数成反比。因此，TF − IDF倾向于过滤掉常见的词语，保留重要的词语。

4.1 特征选择与加权

4.1.1 特征选择

经过构建向量空间模型，使用TF − IDF对留言信息进行了处理，但并不是所有得到的特征都对文本分类有影响，现在所得到的原始特征空间非常大，为了提高留言分类的性能，我们需要从原始特征空间中剔除那些对分类贡献度较小的特征，选取那些贡献度大的特征，从中选取出能够充分体现留言分类标签的特征子空间。我们采用了文档频率

与χ^2统计两种较为常用的特征选择方法，并与之比对，最终选取了效果较好的χ^2统计特征选择方法。

4.1.2 文档频率

文档频率是最简单的特征抽取技术，它在本次研究中是指在已有的留言分类标签体系下，只关注每条留言信息中拥有某个词项的文本的个数，而忽略这个字词在所有留言信息中出现的次数。它的本质就是将文档频率低于设定阈值的特征项作为低频词，把它从原始特征空间中移除，以此来降低特征空间的维数。但对于本次研究来说，每条留言信息中有较多的出现次数较少的关键特征，运用此方法会使有效特征大量缺失，使得最后的分类效果并不好，因此我们在与χ^2统计的特征选择方法进行对比后，摒弃了此方法。利用文档频率归类出各一级分类的特征如图 4-1 和图 4-2 所示。

```
# '交通运输':
  . Most correlated unigrams:
. 我是一名出租车司机
. 中通
  . Most correlated bigrams:
. 无网约车资格证 希望相关部门上路设卡排查
. 无网约车驾驶证 无网约车资格证
# '劳动和社会保障':
  . Most correlated unigrams:
. 胡厅长
. 尊敬的胡厅长
  . Most correlated bigrams:
. 胡厅长 您好
. 尊敬的胡厅长 您好
# '卫生计生':
  . Most correlated unigrams:
. 尊敬的张厅长
. 张厅长
  . Most correlated bigrams:
. 尊敬的张厅长 您好
. 张厅长 您好
```

图 4-1 特征归类示例图 1

```
# '商贸旅游':
  . Most correlated unigrams:
. 妻离子散
. 洗脑
  . Most correlated bigrams:
. 尊敬的李局长 您好
. 手机0000 00000000
# '城乡建设':
  . Most correlated unigrams:
. 采光
. 通风
  . Most correlated bigrams:
. 县长 您好
. 住宅质量保证书 住宅使用说明书
# '教育文体':
  . Most correlated unigrams:
. 教育局
. 尊敬的教育局领导
  . Most correlated bigrams:
. 陈局长 您好
. 王厅长 您好
# '环境保护':
  . Most correlated unigrams:
. 严重污染环境
. 刘厅长
  . Most correlated bigrams:
. 蒋厅长 您好
. 尊敬的刘厅长 您好
```

图 4-2 特征归类示例图 2

4.1.3 χ^2统计

卡方检验是数理统计中一种常用的检验两个变量独立性的方法，最基本的思想就是通过观察实际值与理论值的偏差来确定理论的正确与否。普通做法是先做出“两个变量是独立的”的原假设，然后观察值与如果两者确实独立的情况下应有的理论值的偏差程

度，如果偏差足够小，就认为误差是很自然的样本误差，是测量手段不够精确导致或者偶然发生的，两者确确实实是独立的，此时就接受原假设；如果偏差大到一定程度，使得这样的误差不太可能是偶然产生或者测量不精确所致，就认为两者实际上是相关的，即否定原假设，而接受备择假设[1]。在进行特征选择时，很难说明某一词条和某一类别关联到什么程度才算是有表征作用，因此我们只需要借用卡方统计来选取最为相关的即可。所以在进行原假设时，我们需要做出“某一词条与某一类别不相关”的原假设，然后通过计算词条与类别的卡方值，对卡方值进行从大到小的排序，某一词条对于类别之间卡方统计值越高，那么说明这个词条与类别的相关性越大。计算公式如下：

$$\chi^2(t, C_i) = \frac{N(AD - CB)^2}{(A + C)(B + D)(C + D)} \tag{4.4}$$

其中N表示数据集中留言总数，A表示属于C_i类且包含词条t的留言数，B表示不属于C_i但包含词条t的留言数，C表示属于C_i类且不包含词条t的留言数，D表示不属于C_i类且不包含词条t的留言数。

4.2 加权

特征加权是对文本特征空间中的每个特征项赋予权重的过程，权重大小由特经过特征选择对分类的贡献度决定，特征所具有的类别区分能力越强，赋予的权重就越大。经过特征选择的特征空间中的特征项不仅具备反映文本内容的能力，而且能够区分不同的文本。因此，特征权重需要满足与该特征在某文本中出现的概率成正比与其他文本中出现该特征的文本频率成反比两个条件。在加权处理上我们使用了 TF-IDF 算法来对特征进行加权。

5. 留言分类算法模型设计

完成上述所有数据转换后，现在我们已经拥有了留言信息所有的特征和标签，现在需要来训练分类器，我们可以使用许多算法来解决这类问题。

5.1 逻辑回归算法

逻辑回归是一个分类的算法，把我们的输入值在线性回归中转化为预测值，然后映射到Sigmoid 函数中，将值作为x轴的变量，y轴作为一个概率，预测值对应的Y值越接近于 1 说明完全符合预测结果，它实际上是属于一种分类方法[2]。使用逻辑回归的因变量一般都是二分类的，当然也可以是多分类的，但二分类更为简便、常用，可解释性也

更强。因此进行在本次研究中我们首先将原本的多分类问题，拆解成多个符合逻辑回归便于计算的二分类问题，再分别训练二分类模型，之后将多个二分类模型进行集成，选取概率值最大的那个作为最终的分类结果。对于本题中留言分类的多文本问题，转换后的各个二分类的因变量就为是否为此类别，值为“是”或“否”。在此之后我们需要假设因变量服从伯努利分布，使用Sigmoid函数映射函数引入非线性因素，此函数的定义域为$(-\infty,+\infty)$，而值域为(0,1)，函数形式为：

$$g(z)=\frac{1}{1+e^{-z}} \tag{5.1}$$

当x趋于负无穷时，y趋于 0；在x趋于正无穷时，y趋于 1；当x为 0 时，y为 0.5。函数的输出映射在(0,1)之间，因此可以轻松处理两分类问题。函数图像如下：

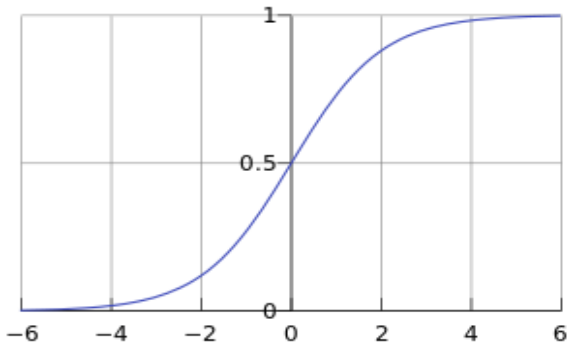


图 5-1 Sigmoid 函数图像

我们将留言信息作为输入x，以此来求取它的归属类别，我们要求取的参数记为 θ，那么这个模型就是将留言信息特征的线性组合作为自变量，使用Sigmoid函数映射到(0,1)上。因此，我们就是要使用模型来求解以归属的类别作为权值的解，那么我们就可以代入Sigmoid函数以此来做出预测函数：

$$h_\theta(x)=g(\theta^T x)=\frac{1}{1+e^{-\theta^T x}} \tag{5.2}$$

函数得到的结果是在给定x和 θ 的条件下，特征属于$y=1$的概率。因此将留言信息作为输入得到的分类结果值为“是”或“否”，即 0 或 1 的概率为：

$$P(y=1|\,x;\theta)=h_\theta(x) \tag{5.3}$$

$$P(y=0|\,x;\theta)=1-h_\theta(x) \tag{5.4}$$

此时决策边界为$\theta^T x = 0$，进一步则可以通过决策函数$y^* = 1, if P(y = 1|x) > 0.5$来判断样本的类别。所以，当我们要判别留言中蕴含的特征归属哪一类时，只需要求出一个函数中的z值：

$$\theta_0 + \theta_1 x_1 + \cdots . + \theta_n x_n = \sum_{i=0}^{n} \theta_i x_i = \theta^T \tag{5.5}$$

是某样本数据的各个特征，维度为 n。进而求出$h_\theta(x)$来计算它属于哪一类别，最终可以通过决策函数来判断它属于哪一类别。最终我们使用了代价函数来衡量我们对留言分类的预测类别与它实际归属的类别的差异。逻辑回归的代价函数如下：

$$\mathrm{J}(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\left(y^{(i)} log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right)\right] \tag{5.6}$$

我们使用了sklearn模块来实现逻辑回归的运算，我们选用默认的OvR改进方法使它支持多分类。OvR是One Vs Rest的缩写，它的本质就是首先选取一类作为类别 1，将剩下的所有的分类作为类别二，这样就得到了多个分类任务，对于每个分类任务，我们都可以使用逻辑回归来计算一条留言信息属于这里单个类别的概率，概率最大的就是这条留言信息最终的分类。多分类的OvR图像如图所示：

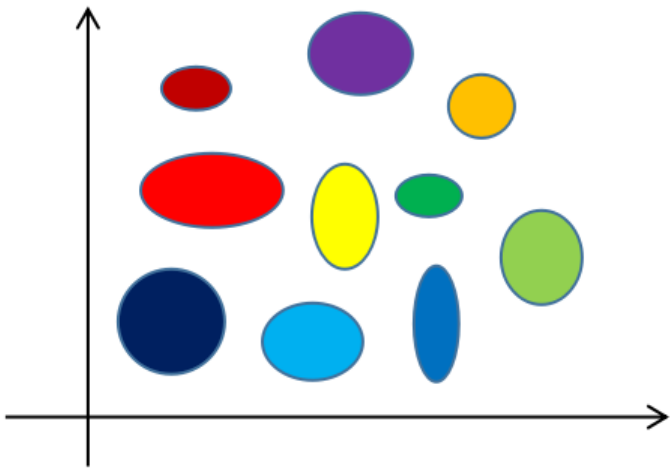


图 5-2 多分类 OVR 图像

sklearn中的Logistic回归是加入了正则化的，在sklearn中，加入了正则项的损失函数表达式为：

$$\mathrm{Loss} = \mathrm{C}.\mathrm{J}(\theta) + L_i \tag{5.7}$$

其中L_i是超参数，可以指定使用L_1正则还是L_2正则，我们采取了默认的使用L_2来进行正则，然后绘制了决策边界，并对真实的决策边界和预测的决策边界进行了对比。

5.2 朴素贝叶斯算法

贝叶斯分类是以贝叶斯定理为基础的一类算法分类的总称，朴素贝叶斯分类是贝叶斯分类中最简单常见的一种方法。朴素贝叶斯分类器会单独计算每一特征被分类的条件概率，进而综合这些概率并对其所在的特征向量做出分类预测[3]。在此题中我们将留言信息的标签作为类别 Y，将留言信息中所蕴含的信息作为特征向量 X，贝叶斯的公式如下：

$$\mathrm{P}(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{5.8}$$

朴素贝叶斯是一个基于贝叶斯理论的分类器。它会对特征属于哪一分类的概率进行比较，然后集合这些概率对特征向量所属的类别做出分类预测。

在使用朴素贝叶斯之前，我们需要对于留言分类问题做出的基本假设是：

（1）留言中所出现的属于分类标签的特征出现的可能性与它和其他单词相邻没有关系，特征间相互独立，如属于城乡建设类中留言信息出现的加油站等特征与属于卫生计生类中留言信息出现的小孩户口等特征相互独立。

（2）每个类别中的特征同等重要。

那么为求出所需要的留言特征属于的类别，用到的朴素贝叶斯公式为：

$$\mathrm{P}(Y|X) = \frac{P(x_1|Y)P(x_2|Y)\dots P(x_n|Y)}{P(X)} \tag{5.9}$$

算法流程为：

（1）在数据处理时，我们就已经基于已有的标签类别对留言信息所属的标签的特征属性进行了划分，得到了m具有n个特征的样本，这些样本分别属已有的留言信息数据集中的类别。

（2）将给出的留言信息数据集作为训练样本，通过数据分析可以得到每个特征的在某一类下的条件概率$\mathrm{P}(x_i|Y)$，再通过全概率公式求得$\mathrm{P}(x)$，全概率公式为：

$$\mathrm{P}(X) = P(X|Y_1)P(Y_1) + P(X|Y_1)P(Y_2) + P(X|Y_3)P(Y_3) + \cdots P(X|Y_K)P(Y_K) \tag{5.10}$$

其中$\mathrm{P}(X|Y_k)$可根据特征独立性展开。

（3）将求得的全概率和条件概率带入朴素贝叶斯公式，求得这条留言中所蕴涵的特征属性归属于某一类别的概率。

在留言分类问题中，我们使用$\mathrm{TF}-\mathrm{IDF}$加权技术对留言信息进行了单词统计，它是具有多元离散特征的分类，因此我们选用多项式朴素贝叶斯分类器。多项式模型在计算先验概率$\mathrm{P}(Y_k)P(Y_k)$和条件概率$\mathrm{P}(x_i|Y_k)P(x_i|Y_k)$时，会做一些平滑处理，具体公式为：

$$\mathrm{P}(Y_k)=\frac{N_{Y_k}+\alpha}{N+K_\alpha} \tag{5.11}$$

其中 N 是样本数，N_{Y_k}是类别为Y_k的样本数，K 为总的类别个数，α 为平滑值。

$$\mathrm{P}(x_i|Y_k)=\frac{N_{Y_k,x_i}+\alpha}{N_{Y_k}+n\alpha} \tag{5.12}$$

N_{Y_k,x_i}表示的含义为类比为Y_k，且特征为x_1的样本数，n表示特征x_1可以选择的数量。理论上来说，朴素贝叶斯模型与其他分类方法相比具有最小的误差率。然而实际上，留言信息数据集属性的独立性往往是很难满足的。例如如果在一条留言信息中出现了两种类别的属性特征，往往就会判断失误，造成误差，因此我们需要对模型进行改进。我们通过使用较为简便的卡方检验来构造留言信息属性间的相关系数，来以此作为权值属性加权。属性加权其本质是为了解决各属性对于类别决策影响相同的问题，对各特征属性赋予一个权值。卡方检验是在本题中主要是比较一条留言信息中所拥有的两个或两个以上的特征的样本率以及两个分类变量的关联性分析。其本质就是比较预测频数和实际频数的吻合程度问题，然后选取显著的那个类别作为标签归属。卡方检验的计算公式为：

$$\chi^2=\sum\frac{(A-T)^2}{T} \tag{5.13}$$

其中，在本次研究中 A 为留言信息实际归属的类别，T 为使用朴素贝叶斯分类器后留言信息预测归属的类别。

利用多类标分类器朴素贝叶斯算法进行留言分类的特征与设计，并对此模型进行评估，找出此算法模型将留言归类错误的一些实例，如图 5-3 和图 5-4 所示。

'环境保护' predicted as '城乡建设' : 157 examples.
'交通运输' predicted as '城乡建设' : 126 examples.
'教育文体' predicted as '城乡建设' : 242 examples.
'劳动和社会保障' predicted as '城乡建设' : 275 examples.
'商贸旅游' predicted as '城乡建设' : 238 examples.
'卫生计生' predicted as '城乡建设' : 123 examples.
'城乡建设' predicted as '环境保护' : 25 examples.
'教育文体' predicted as '环境保护' : 15 examples.
'劳动和社会保障' predicted as '环境保护' : 13 examples.
'城乡建设' predicted as '交通运输' : 22 examples.
'教育文体' predicted as '交通运输' : 14 examples.
'城乡建设' predicted as '教育文体' : 48 examples.
'环境保护' predicted as '教育文体' : 29 examples.
'交通运输' predicted as '教育文体' : 10 examples.
'劳动和社会保障' predicted as '教育文体' : 49 examples.
'商贸旅游' predicted as '教育文体' : 39 examples.

图 5-3 留言归类错误示例图 1

'卫生计生' predicted as '教育文体' : 19 examples.
'城乡建设' predicted as '劳动和社会保障' : 91 examples.
'环境保护' predicted as '劳动和社会保障' : 37 examples.
'交通运输' predicted as '劳动和社会保障' : 23 examples.
'教育文体' predicted as '劳动和社会保障' : 78 examples.
'商贸旅游' predicted as '劳动和社会保障' : 57 examples.
'卫生计生' predicted as '劳动和社会保障' : 48 examples.
'城乡建设' predicted as '商贸旅游' : 26 examples.
'环境保护' predicted as '商贸旅游' : 15 examples.
'教育文体' predicted as '商贸旅游' : 20 examples.
'劳动和社会保障' predicted as '商贸旅游' : 25 examples.
'城乡建设' predicted as '卫生计生' : 17 examples.
'环境保护' predicted as '卫生计生' : 10 examples.
'教育文体' predicted as '卫生计生' : 15 examples.
'劳动和社会保障' predicted as '卫生计生' : 26 examples.
'商贸旅游' predicted as '卫生计生' : 13 examples.

图 5-4 留言归类错误示例图 2

再使用卡方检验来查找与每个类别最相关的词条后得到了较好的效果，如图 5-5 和图 5-6 所示。

图 5-5 留言归类示例图 1

图 5-6 留言归类示例图 2

5.3 线性支持向量机算法

支持向量机是一种经典的二元分类算法，如果不考虑集成学习算法，不考虑特定的训练集，在分类算法中支持向量机的效果非常好。它的基本思想是构建一个超平面作为决策平面，使得两个类别之间的间隔最大化。理论上我们可以找到多个超平面将数据分开，并且优化时希望所有的点都离超平面尽可能的远，但是实际上离超平面足够远的点基本上都是被正确分类的，所以多度在意那些离超平面远的点是没有意义的，我们反而更应该关心那些离超平面很近的点，这些点比较容易分错。所以说我们只要让离超平面比较近的点尽可能的远离这个超平面，那么分类效果就会较好。对于线性支持向量机算法，只要找到离超平面的最近点，通过其约束条件求出最优解即可实现分类[4]。

在本次留言分类的研究中，我们设线性可分样本集为

$$(\bar{x}_i, y_i), (\bar{x}_i, y_i) \in R^d \times \{\pm 1\}, i = 1,2, \dots, k$$

那么 d 维空间中线性判别函数的一般形式为：

$$g(\bar{x}) = \bar{w} \cdot \bar{x} + b$$

分类面方程为：

$$\bar{w} \cdot \bar{x} + b = 0$$

将判别函数进行归一化，使两类所有样本都满足$|g(\bar{x})| \geq 1$，如果分了面对所有样本正确分类，则应满足约束$y_i[\bar{w} \cdot x_i + b] - 1 \geq 0 \quad i = 1,2,\ldots,k$

最优分类面应使两类样本到决策面的最小距离$\frac{1}{\|\bar{w}\|}$尽量大。即在下述条件下的约束最小化：

$$\Phi(\bar{w}) = \frac{1}{2}\|w\|^2 = \frac{1}{2}\bar{w} \cdot \bar{w}$$

可以看出这是一个二次优化问题，采用拉格朗日乘子法，可将原问题转化为优化问题的对偶形式：在约束

$$\sum_{i=1}^{k} a_i y_i = 0 \qquad a_i \geq 0, i = 1,2,\ldots,k$$

下对拉格朗日系数a_i求解下列函数：

$$Q(\bar{a}) = \sum_{i=1}^{k} a_i - \frac{1}{2}\sum_{i=1}^{k} a_i a_j y_i y_j (\bar{x}_i \cdot \bar{x}_j) \tag{5.14}$$

的最大值。这是一个不等式约束下二次函数极值问题，存在唯一解，若a_i^*为最优解，则

$$\bar{w}^* = \sum_{i=0}^{k} a_i^* y_i \bar{x}_i$$

根据 Kuhn-Tucker 条件，这个优化问题的解须满足：

$$a_i[y_i(\bar{w} \cdot \bar{x}_i + b) - 1] = 0 \qquad i = 1,2,\ldots,k$$

由上式可知，远离分类面的样本所对应的a_i^*必定为零，非零a_i^*所对应的样本必定位于分类面上，称之为位于分类面上的样本为支持向量。b^*值可由任意支持向量带入最优分类面方程得到。求解上述问题后得到的最优分类函数为：

$$\mathrm{f}(\bar{x}) = sgn\left[\sum_{i=1}^{n} \bar{w}^* \cdot \bar{x} + b^*\right] = sgn\left[\sum_{i=1}^{n} a_i^* y_i (\bar{x}_i \cdot \bar{x}) + b^*\right] \tag{5.15}$$

n 为支持向量数

根据上式可以确定输入的文本是否属于该分类面所对应的类别，对于$\mathrm{n}(n > 2)$类别的，对于$\mathrm{n}(n > 2)$类别的情况，可以某一类样本为正例，其他类样本为反例训练，分别得到 n 个二元分类器，再由 n 个二元分类器构成一个 n 元分类器[5]。

支持向量机的学习能力是独立于特征空间维数的，决定分类面性质的只是训练样本中的支持向量部分，这样，分类器也可以很好地在高维空间中得到应用，适合于解决特征空间维数较高(一般大于 10000 个特征)的文本分类问题。

5.4 随机森林算法

随机森林就是通过集成学习的思想将决策树集成的一种算法，它的基本单元是决策树，决策树往往会产生过拟合问题，随即森林阻止了这类问题的发生。

针对研究中的留言分类问题，我们已知共有n 个样本，M 个特征维度，指定一个常数$m \ll M$，随机地从M 个特征中选取m 个特征子集，每次树进行分裂时，从这m 个特征中选择最优的。一棵留言分类的决策树的构建方式如图 5-7 所示：

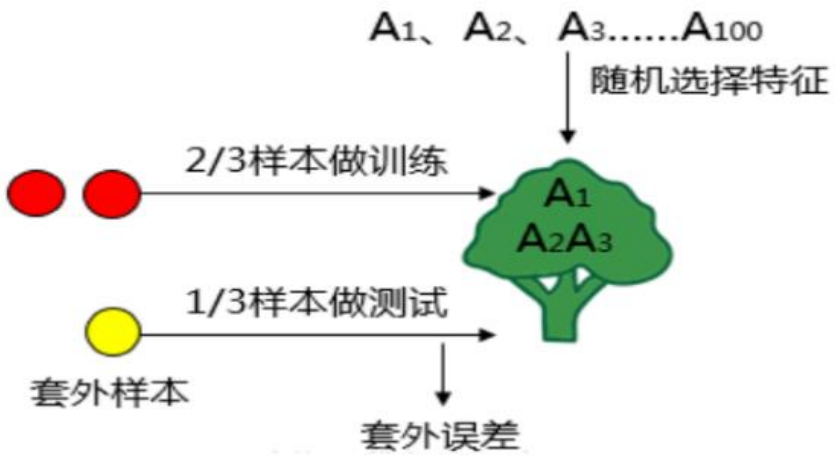


图 5-7 一颗决策树构建图

按照这种方法，可以构建出很多棵树，并且不对它进行剪枝，使它包括较多的特征。可是这么多棵树综合评判的结果不可以作为最后的结果，它会通过迭代会选用更好的特征进行分枝。对于每一棵树都有m 个特征，要知道某个特征在这个树中是否起到了作用，可以随机改变这个特征的值，使得“这棵树中有没有这个特征都无所谓”，之后比较改变前后的测试集误差率，误差率的差距作为该特征在该树中的重要程度，测试集即为该树抽取样本之后剩余的袋外样本。在一棵树中对于每个特征都计算一次，就可以算每个特征在该树中的重要程度。我们可以计算出所有树中的特征在各自树中的重要程度。但这只能代表这些特征在树中的重要程度不能代表特征在整个森林中的重要程度。

每个特征在多棵数中出现，取这个特征值在多棵树中的重要程度的均值即为该特征在森林中的重要程度。如下式：

$$MAD(A_i) = \frac{1}{n}\sum_{t=1}^{n}(errOOB_{t1} - errOOB_{t2}) \tag{5.16}$$

其中 n 表示特征 Ai 在森林中出现的次数，$errOOB_{t1}$ 表示第 t 棵树中 A_i 属性值改变之后的袋外误差，$errOOB_{t2}$ 表示第 t 棵树中正常 A_i 的袋外误差。可以用下图图 5-8 来表示：

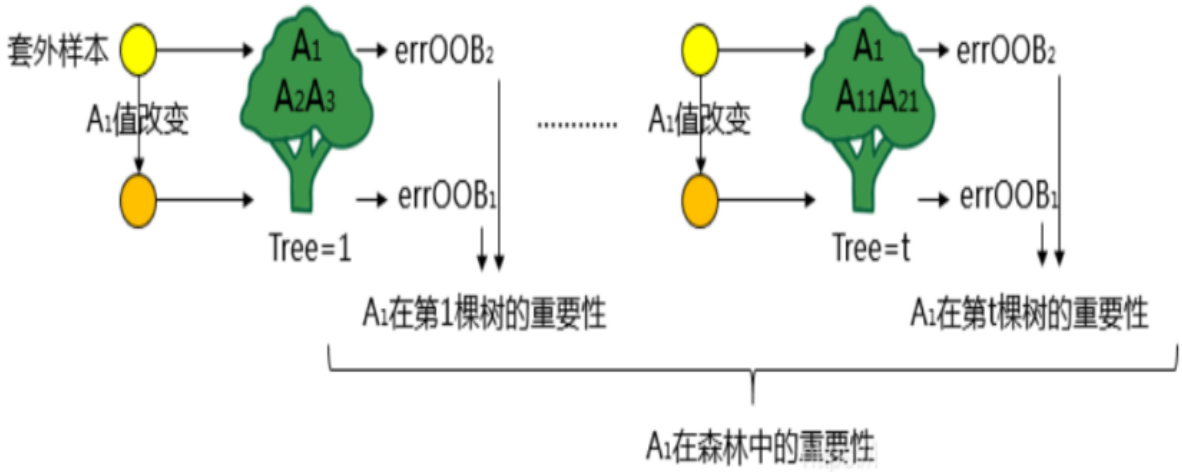


图 5-8 特征的重要性示例图

这样就得到了所有特征在森林中的重要程度。将所有的特征按照重要程度排序，去除森林中重要程度低的部分特征，得到新的特征集。这时相当于我们回到了原点，这算是真正意义上完成了一次迭代。按照上面的步骤迭代多次，逐步去除相对较差的特征，每次都会生成新的森林，直到剩余的特征数为m为止。最后再从所有迭代的森林中选出最好的森林。迭代的过程如图 5-9 所示：

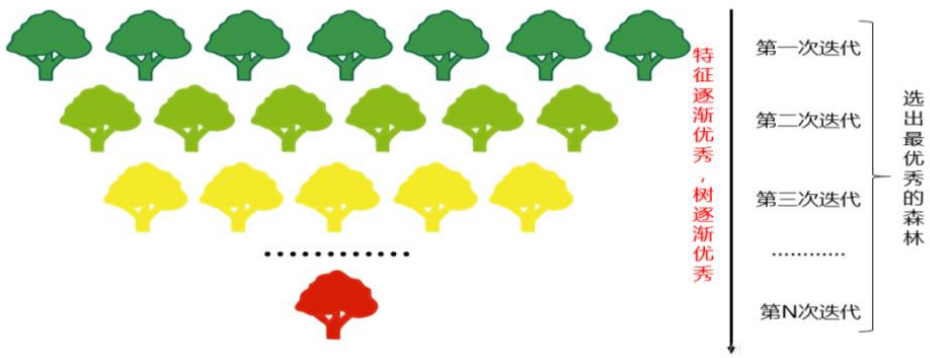


图 5-9 随机森林迭代过程示意图

得到了每次迭代出的森林之后，因为随机森林是一种集成算法，所以最后的森林不一定是最优的，但我们需要选择出最优秀的森林，所以引入一个指标 OOB 来评价一个森林的好坏，OOB 是 out-of-bag 的缩写，在上文中所用到的 OOB 用于评价套外样本在树中的误差率，而这里的 OOB 评价套外样本在森林中的误差率。由于在构建每棵树时，我们对训练集使用了不同的随机且有放回地抽取，所以对于第 k 棵树而言，大约有 1/3 的训练实例没有参与第 k 棵树的生成，所以在一棵树的生成过程并不会使用所有的样本，未使用的样本就叫(Out_of_bag)袋外样本，它们称为第 k 棵树的oob样本。而这样的采样特点就允许我们进行oob估计。

通过袋外样本，可以评估这个树的准确度，其他子树叶按这个原理评估，最后可以取平均值，即是随机森林算法的性能。

特征选择原理：因为袋外样本的存在，因此为了节省时间并不需要进行十字交叉测试，只需要通过依次对每个特征赋予一个随机数，观察算法性能的变化，倘若变化大，则说明该特征重要，使用sklern模块会对每个特征赋予一个分数，分数越大，特征越重要，因此，可以根据特征重要性排序，然后选择最佳特征组合。

每个样本在多棵树中是套外样本，通过多棵树的预测这个样本的结果。预测方式如下图 5-10 所示：

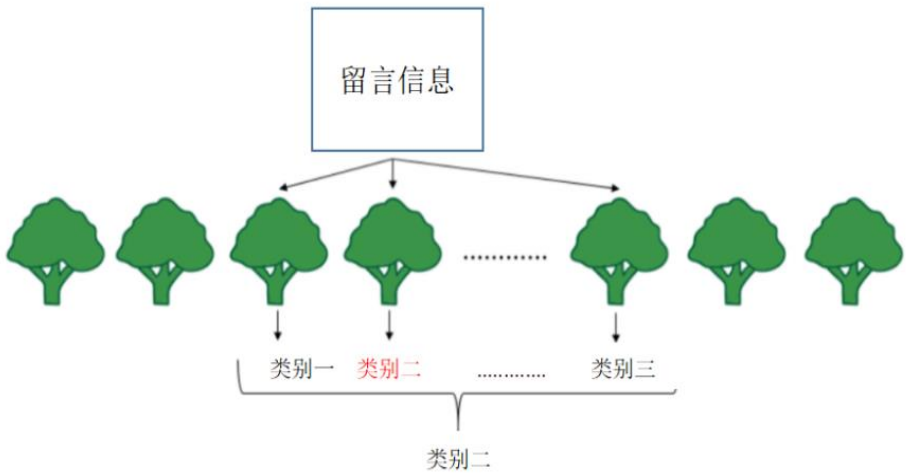


图 5-10 多棵树预测示意图

预测出所有样本的结果之后与真实值进行比较，就可以得到这个森林的套外误差率。只需要选择套外误差率最小的森林作为最终的随机森林模型，以此来得到归属分类。

5.5 算法的推荐及其改进

在上述中每种算法都有一定的误差，所以并不存在特别准确的模型，因此我们对模型进行了组合，充分利用各个模型的优点，改进缺点，取长补短。组合形成一个精准的留言分类模型，如图 5-11 所示：

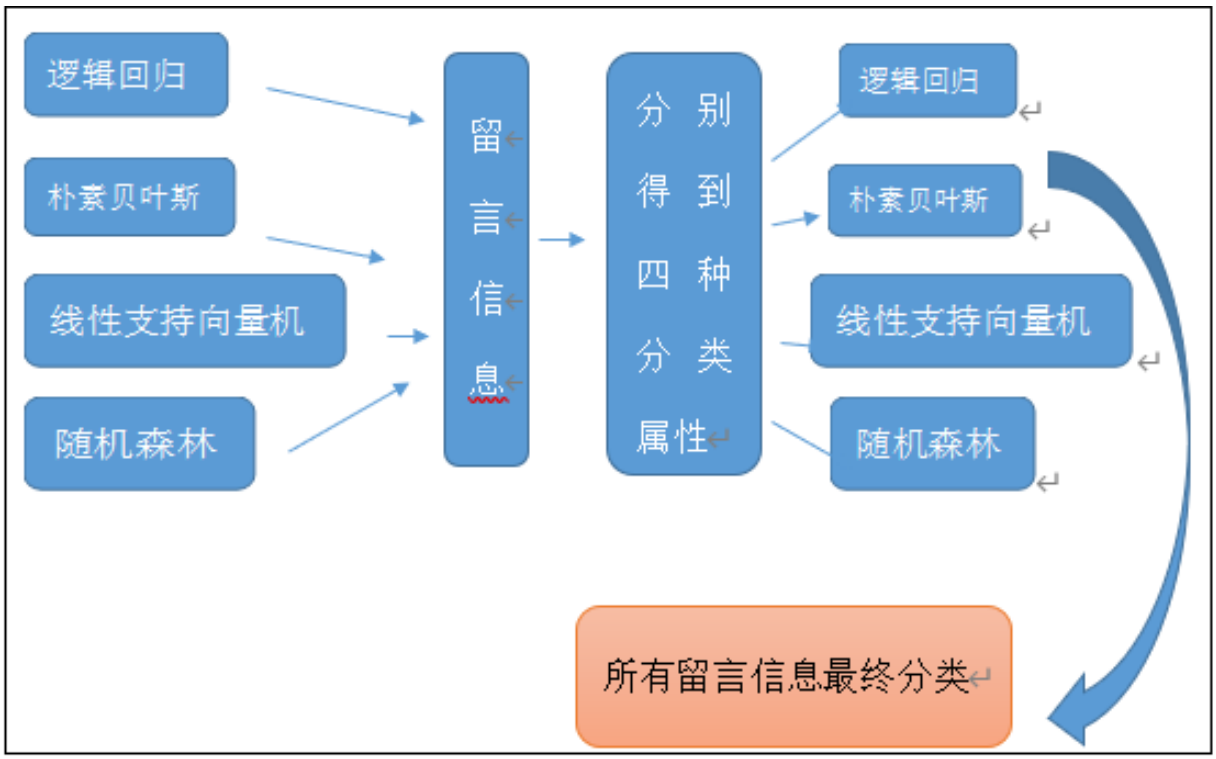


图 5-11 组合分类模型图

为了提升留言分类所涉及的算法的性能，本文将上述所提出的算法进行加权融合，即针对每一名群众，各算法均对每条留言信息进行分类，均推断出每条留言信息的分类。针对每条留言信息，各推荐算法均用于预测信息所在的类别，进行线性加权取均值作为对该留言信息的最终类别，这样得出的归属类别更具一般性。当算法获得所有留言的最终归属类别后，选择多个算法分类较多的那一类为该留言进行分类，各留言信息最终得分占评分最大值的比重作为该留言的归属分类，从而实现算法之间的融合推荐。

6. 归类相似留言，挖掘热点问题

6.1 留言时间切分处理

为了保证留言的实效性，在某一时间段内更好的反映热点问题，我们对时间进行了切分。在数据处理时发现 2017 年至 2018 年的留言条数较少，所以对 2017 年至 2018 年包含的所有留言进行了一次切分，2019 年至 2020 年留言条数较多。如图 6-1 所示：

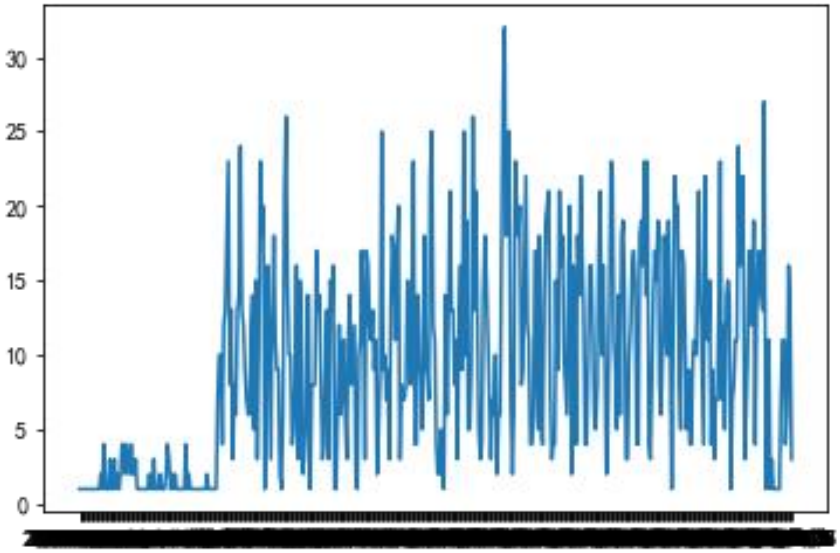


图 6-1 留言统计图

我们对 2019 和 2020 年以三个月为一个时间段进行切分。部分时间切分数据如图 6-2 所示：

2019.1.1-2019.3.30

	留言编号	留言用户	...	反对数	点赞数
4	285107	A00024520	...	0	0
5	283135	A00057709	...	0	0
6	199492	A00027830	...	0	1
7	238266	A00063188	...	0	0
8	215087	A00051311	...	0	3
..	...	...	...	..	..
975	208479	A00010328	...	0	0
976	265242	A00054222	...	0	0
977	201867	A000106252	...	0	0
978	200748	A00097226	...	0	0
979	266563	A00078073	...	0	0

[976 rows x 7 columns]

4	A市电建星湖湾强制要求业主收房
5	A市A1区万国城moma未经业主同意强建养老院
6	反映A市美麓阳光楼盘质量问题和停工问题
7	冰雪天气，A7县校车停运合理吗
8	职能部门相互推诿，A5区植物园社区山水熙园小区业主苦不堪言
...	...
975	A5区学而思培优华盛花园培训点噪音扰民，影响居民生活
976	请依法处理好A7县星沙商业乐园的历史遗留问题
977	A市梅溪湖金茂悦小区周边乱停车依然严重
978	A市古曲南路广益实验中学门口摊贩脏乱差
979	请解决A2区坡子街办事处居民房屋居住问题

2020.1.1-2020.3.30

	留言编号	留言用户	...	反对数	点赞数
4246	203012	A00074795	...	0	0
4247	278563	A00097910	...	0	0
4248	208195	A00028561	...	0	0
4249	189294	A00083527	...	0	0
4250	237799	A00090621	...	0	0
...	...	...	...	..	..
4321	222105	A00046446	...	0	0
4322	272224	A909224	...	0	1
4323	287331	A909217	...	0	0
4324	214282	A909209	...	0	0
4325	215842	A909210	...	0	0

[80 rows x 7 columns]

4246	质疑A1区老地方美食广场无证无照经营
4247	A2区长大建设集团福满新城物夜施工扰民严重
4248	A8县市中梁首府业主请求维权
4249	A市南站里面的12306客服虚设
4250	建议A2区暮云街道月塘路增设限时停车位
...	...
4321	A7县高桥镇百录村村民盼望已久的水泥路在哪？
4322	丽发新城小区噪音大粉尘大,求撤走搅拌站
4323	A2区李丽发新城附近无资质混凝土搅拌站为何禁而不止？
4324	A市丽发新城小区附近搅拌站噪音扰民和污染环境
4325	A2区丽发新城小区附近太吵了

图 6-2 时间划分情况

6.2 K-means 聚类算法

聚类是基于数据中的模式将整个数据划分为组(也称为簇)的过程,它是没有任何固定目标变量的无监督学习问题。K-means 算法是一种基于质心的算法，或基于距离的算法，我们计算将点分配给一个簇的距离，算法的主要目的是最小化点与它们各自的簇质心之间的距离之和。质心是每个簇的均值向量，即向量各维取平均即可。在 K-means 中，每个聚类都与一个质心相关联。在研究热点挖掘问题时，使用这种算法之前，我们需要对已经处理好的留言主题数据集进行假设，假设条件如下：

1.每条留言主题中的所有数据点应该彼此相似。

2.来自不同留言主题中的数据点应尽可能不同。

假设完毕后我们首先确定我们希望将留言主题数据集经过聚类得到 k 个集合的 k 值，然后从留言主题数据集中随机选择 k 个数据点作为质心对数据集中每一个点，计算其与每一个质心的距离，我们选取的是更适合计算文本相似度的余弦距离，将文本已转换为权值向量的基础上，通过计算两个向量的夹角余弦值，就可以评估他们的相似度。余弦值的范围在[-1,1]之间，值越趋近于 1，代表两个向量方向越接近；越趋近于-1，代表他们的方向越相反。为了方便聚类分析，我们将余弦值做归一化处理，将其转换到[0,1]之间，夹角余弦越接近于 1，表示样本越相似；越接近于 0，表示样本越不相似。余弦距离计算公式如下：

$$\cos\theta = \frac{x_1 x_2 + y_1 y_2}{\sqrt{{x_1}^2 + {y_1}^2}\sqrt{{x_2}^2 + {y_2}^2}} \tag{6.1}$$

然后重复将所有点分配给到某个质心距离最近的簇和重新计算新形成的簇的质心，计算质心并基于它们与质心的距离将所有点分配给簇的步骤是单次迭代。如果新形成的簇的质心不会改变数据点保留在同一个簇中达到最大迭代次数如果新形成的簇的质心没有变化，我们就可以停止算法。即使在多次迭代之后，所有簇都还是相同的质心，我们可以说该算法没有学习任何新模式，并且它是停止训练的标志。另一个明显的迹象表明，在多次迭代训练的之后，如果数据点仍然都在同一簇中，我们应该停止训练过程。最后，如果达到我们设置的最大迭代次数，也可以停止训练。在热点挖掘问题中，为了使数据更容易可视化并决定哪些簇更好。我们在对留言时间进行切分后的基础上，选取了每条留言主题中的地点和事件这两个特征,只利用留言主题中的地点和时间来进行细分，其流程为：

输入:训练数据集$D = x^{(1)}, x^{(2)} \ldots\ldots x^{(m)}$聚类簇数 k:

过程: 函数kMeana$(D, k, moxIter)$从 D 中随机选择 k 个样本作为初始"簇中心"向量：$u^{(1)}, u^{(2)} \ldots\ldots u^{(k)}$

Repeat

令$C_i = 0(1 \le i \le k)$

for j $= 1,2, \ldots,$ m

计算样本$x^{(j)}$与各"簇中心"向量$u^{(i)}(1 \le i \le k)$的欧式距离

根据距离最近的"簇中心"向量确定$x^{(j)}$的簇标记:$\lambda_j = argmin_{i\in\{1,2,\ldots.k\}} d_{ji}$

将样本$x^{(j)}$划入相应的族: $C_{\lambda_j} = C_{\lambda_j} \cup \{x^{(j)}\}$

end for

for i $= 1,2, \ldots,$ k

计算新族中心向量: $(u^{(i)})' = \frac{1}{|C_i|}\sum_{x\in C_i} x$

$(u^{(i)})' = u^{(i)}$ $t\hbar en$

将当的放中心向量$u^{(i)}$更新为$(u^{(i)})'$

保持当前均值向量不变

end if

end for

else

until 当的放中心向量均末更新

输出：簇划分$C = C_1, C_2 \ldots\ldots C_k$

K 均值算法中有两个十分棘手的问题，第一个问题就是 k 值的确定，可以看到，K 均值算法需要提前告知样本中类的个数。但是对于一般的问题，我们并不知道会有多少个类，因此很难选择具体的 k 值。第二个问题就是收敛性问题，K 均值聚类属于启发式方法，它是随机选取的质心，因此可能会收敛到局部最小值，而非全局最小值，所以不能保证收敛到全局最优，初始中心的选择会直接影响到聚类结果，那么我们使用一种二

分 K—means 算法。在二分 k-means 聚类中，使用一种用于度量聚类效果的指标 SSE(Sum of Squared Error)，即对于第 i 簇，其 SSE 为各个样本点到“簇中心”点的距离的平方的和，SSE 值越小表示数据点越接近于它们的“簇中心”点，聚类效果也就越好，以此作为划分簇的标准。SSE 算法的思想是将整个样本集作为一个簇，该“簇中心”点向量为所有样本点的均值，计算此时的 SSE。若此时簇个数小于 kk ，对每一个簇进行 kmeans聚类(k=2k=2) ，计算将每一个簇一分为二后的总误差 SSE，选择 SSE 最小的那个簇进行划分操作。其实整个过程和决策树相似，都是利用一个指标，寻找最好划分的方式。算法流程如下：

输入:训练数据集$D = x^{(1)}, x^{(2)} \ldots\ldots x^{(m)}$聚类簇数 k:

过程: 函数kMeans$(D, k, maxIter)$

将所有点看作一个簇，计算此时“簇中心”向量: $\mathrm{u}^{(\mathrm{i})} = \frac{1}{\mathrm{m}} \sum_{\mathrm{x} \in \mathrm{D}} \mathrm{x}$

while “簇中心”个数 h <k:

for i = 1,2,...,h do

将第 i 簇使用kmeans算法进行划分，其中 k= 2

计算划分后的误差平方和 SSE:

比较 k 种划分的 SSE 值，选择 SSE 值最小的那簇划分进行划分

更新簇的分配结果

添加新的“簇中心”

until 当前“簇中心个数达到 k:

输出:簇划分$\mathrm{C} = \mathrm{C}_1, \mathrm{C}_2 \ldots\ldots \mathrm{C}_\mathrm{k}$

该算法需要确定簇的个数，而我们需求中分类的个数是未知的。因此，希望通过观察性能度量指标 DI 和 DBI 的变化趋势来确定一个合适 k 值。

考虑聚类结果的簇划分 C={C1,C2,...,Ck}，定义簇 C 内样本间的平均距离

$$\mathrm{avg}(\mathrm{C}) = \frac{2}{|\mathrm{C}|(|\mathrm{C}| - 1)} \sum_{1 \le \mathrm{i} \le \mathrm{j} \le |\mathrm{C}|} \mathrm{dist}(\mathrm{x_i}, \mathrm{x_j}) \tag{6.2}$$

簇 C 内样本间的最远距离：$\mathrm{diam}(\mathrm{C}) = \max_{1 \le \mathrm{i} \le \mathrm{j} \le |\mathrm{C}|} \mathrm{dist}(\mathrm{x_i}, \mathrm{x_j})$

簇 C_i与簇C_j最近样本间的距离：$d_{min}(C) = min_{x_i \in Ci, x_j \in Cj} dist(x_i, x_j)$

簇 C_i与簇C_j中心点间的距离：$d_{cen}(C) = dist(u_i, u_j)$

DB 指数($Davies-Bouldin\ Index, DBI$):

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{j\neq i}(\frac{avg(Ci)+avg(Cj)}{d_{cen}(u_i,u_j)}) \tag{6.3}$$

Dunn 指数(Dunn Index, DI):

$$DI = \min_{1\leq i\leq k}\left\{\min_{j\neq i}\left(\frac{d_{min}(Ci,Cj)}{max_{1\leq l\leq k}diam(Cl)}\right)\right\} \tag{6.4}$$

DBI 值越小越好，而 DI 则相反，值越大越好

6.3DBSCAN 密度聚类算法

DBSCAN 是一种基于密度的聚类算法，这类密度聚类算法一般假定类别可以通过样本分布的紧密程度决定。DBSCAN 由密度可达关系导出的最大密度相连的样本集合，即为我们最终聚类的一个类别，或者说一个簇。这个 DBSCAN 的簇里面可以有一个或者多个核心对象。如果只有一个核心对象，则簇里其他的非核心对象样本都在这个核心对象的 $\epsilon\epsilon$-邻域里；如果有多个核心对象，则簇里的任意一个核心对象的 $\epsilon\epsilon$-邻域中一定有一个其他的核心对象，否则这两个核心对象无法密度可达。这些核心对象的 $\epsilon\epsilon$-邻域里所有的样本的集合组成的一个 DBSCAN 聚类簇。任意选择一个没有类别的核心对象作为种子，然后找到所有这个核心对象能够密度可达的样本集合，即为一个聚类簇。接着继续选择另一个没有类别的核心对象去寻找密度可达的样本集合，这样就得到另一个聚类簇。一直运行到所有核心对象都有类别为止[6]。对于一些异常样本点或者说是少量游离于簇外的样本点，我们将这些样本点标记为噪音点。对于距离的度量问题，即如何计算某样本和核心对象样本的距离。在 DBSCAN 中，一般采用最近邻思想，采用某一种距离度量来衡量样本距离，比如欧式距离。这和 KNN 分类算法的最近邻思想完全相同。对应少量的样本，寻找最近邻可以直接去计算所有样本的距离，如果样本量较大，则一般采用 KD 树或者球树来快速的搜索最近邻。如果某些样本可能到两个核心对象的距离都小于 $\epsilon\epsilon$，但是这两个核心对象由于不是密度直达，又不属于同一个聚类簇，那么就采用先来后到的方法，先进行聚类的类别簇会标记这个样本为它的类别[7]。

它的算法流程如下：

输入：样本集 D=$(x_1,x_2,\dots,x_m)(x_1,x_2,\dots,x_m)$，邻域参数$(\varepsilon,MinPts)(\varepsilon,MaxPts)$，样本距离度量方式

输出：簇划分 C.

（1）初始化核心对象集合Ω=∅Ω=∅，初始化聚类簇数k=0，初始化未访问样本集合ΓΓ = D, 簇划分C = ∅∅

（2）对于j=1,2,...m，按下面的步骤找出所有的核心对象：

通过距离度量方式，找到样本$x_i x_j$的ϵϵ-邻域子样本集$N\epsilon x_i N\epsilon x_j$

如果子样本集样本个数满足|Nϵ(xj)|≥MinPts|Nϵ(xj)|≥MinPts， 将样本$x_j x_j$加入核心对象样本集合：Ω=Ω∪{xj}Ω=Ω∪{xj}

（3）如果核心对象集合Ω=∅Ω=∅，则算法结束，否则转入步骤4.

（4）在核心对象集合ΩΩ中，随机选择一个核心对象oo，初始化当前簇核心对象队列Ωcur={o}Ωcur={o}，初始化类别序号k=k+1，初始化当前簇样本集合Ck={o}Ck={o}，更新未访问样本集合Γ=Γ−{o}Γ=Γ−{o}

（5）如果当前簇核心对象队列Ωcur=∅Ωcur=∅，则当前聚类簇CkCk生成完毕，更新簇划分C={C1,C2,...,Ck}{C1,C2,...,Ck}，更新核心对象集合Ω=Ω−CkΩ=Ω−Ck， 转入步骤3。

（6）在当前簇核心对象队列ΩcurΩcur中取出一个核心对象o′ o′，通过邻域距离阈值ϵϵ找出所有的ϵϵ-邻域子样本集Nϵ(o′)Nϵ(o′)，令Δ=Nϵ(o′)∩ΓΔ=Nϵ(o′)∩Γ，更新当前簇样本集合Ck=Ck∪ΔCk=Ck∪Δ，更新未访问样本集合Γ=Γ−ΔΓ=Γ−Δ，转入步骤5.

输出结果为： 簇划分C={C1,C2,...,Ck}{C1,C2,...,Ck}

6.4 算法的选择

对于热点挖掘问题，我们使用了K-means聚类和DBSCAN聚类两种算法，因为留言信息数据集中聚类间距差相差较大，使用DBSCAN聚类质量较差，而且调参相对于K-Means聚类算法显得较为复杂，不仅需要对距离阈值ϵϵ进行调参，还需要对邻域样本数阈值MinPts也进行调参，使用不同的参数组合对最后的聚类效果有较大影响，目前我们还未找到一种较为理想的参数组合，所以我们选取了K-means聚类算法。

7. 意见答复评价

对于对答复意见的评价问题，我们将它看作与主观题自动评分相似的原理。在此次研究中，将留言详情看作标准答案，将答复意见看作考生回答的答案，对两者进行匹配即可得到关于答复意见的评分。答复意见的自动评分原理见图 7-1。

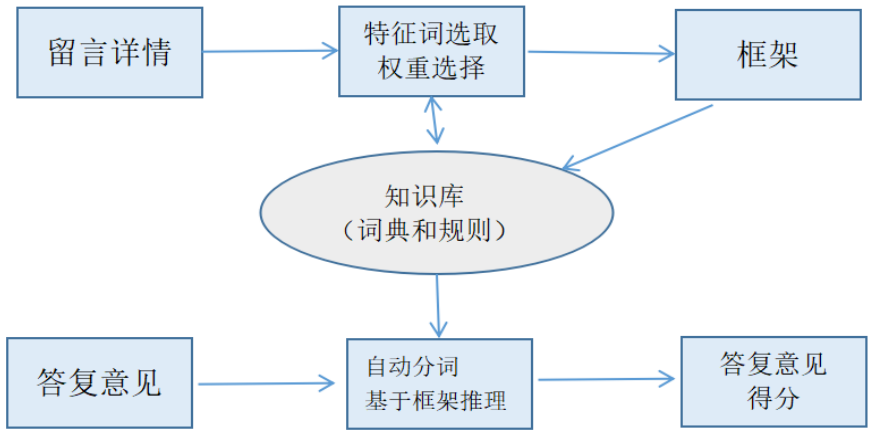


图 7-1 答复意见的自动评分原理图

留言详情与答复意见都是以文本方式存储，我们首先对文本进行了分词，使得分词后的文本更容易地被计算机处理。然后我们使用了矢量空间模型的算法思想。在矢量空间模型中，文本被看作由一组正交词条所生成的矢量空间。根据这个思想，同时考虑到评价答复意见主要是探索是否包含留言详情中的要点，因此提出答复意见评判的步骤如下：

（1）留言详情是由一些要点组成的，如果把留言详情 A 看成一个由 n 个要点 P_i 组成的集合，则可以这样表示留言详情：

$$A = \{P_1, P_2 \ldots\ldots P_i \ldots. P_n\}$$

（2）设每个要点 P_i 的分值为 M_i，则该留言的总分 M 为：

$$M = \sum_{i=1}^{n} M_i \tag{7.1}$$

（3）按照矢量空间模型的思想，将留言详情每一个要点 P_i 被看成是由 K_i 个特征词 w_i 组成的向量 P：

$$P = \{W_1, W_2 \ldots\ldots W_i \ldots. W_n\}$$

（4）由 TF-IDF 算出每个特征词的权重 w_j，则其归一化权重为：

$$\overline{W_j} = \frac{W_j}{\sqrt{\sum_{j=1}^{k}(W_j)^2}} \tag{7.2}$$

（5）设答复意见的每一个要点 P'i 也被看成是由 K'i 个特征词 w'j 组成的向量 P

$$P' = \{W'_1, W'_2 \ldots\ldots W'_i \ldots. W'_n$$

(6) 如果答复意见和留言详情的向量间的距离为零，则说明答复意见和留言详情完全匹配，答复意见可以拿到该点所有的分值，即：

$$M'_i = (1 - \sqrt{\sum_{j=1}^{k}((W_j - W'_j) \times \overline{W_j})^2}) \times M_i \tag{7.3}$$

(7) 根据上式，则答复意见所得总分 M' 为：

$$M'_i = \left(1 - \sqrt{\sum_{j=1}^{k}\left(\left(W_j - W'_j\right) \times \overline{W_j}\right)^2}\right) \times M_i \tag{7.4}$$

8. 结果对比与分析

8.1 问题一结果分析

问题一需提交建立关于留言内容的一级标签分类模型，来解决政务系统依靠人工处理留言存在效率低，工作量大，错误率高等一系列问题。

本小题给出 4 种推荐算法：逻辑回归、朴素贝叶斯、线性支持向量机(SVM)、随机森林的测试结果，其中也给出了不同算法在准确率(Precision)、召回率(Recall)和 F1 指标的对比，四种算法的详细对比结果见下列各图。

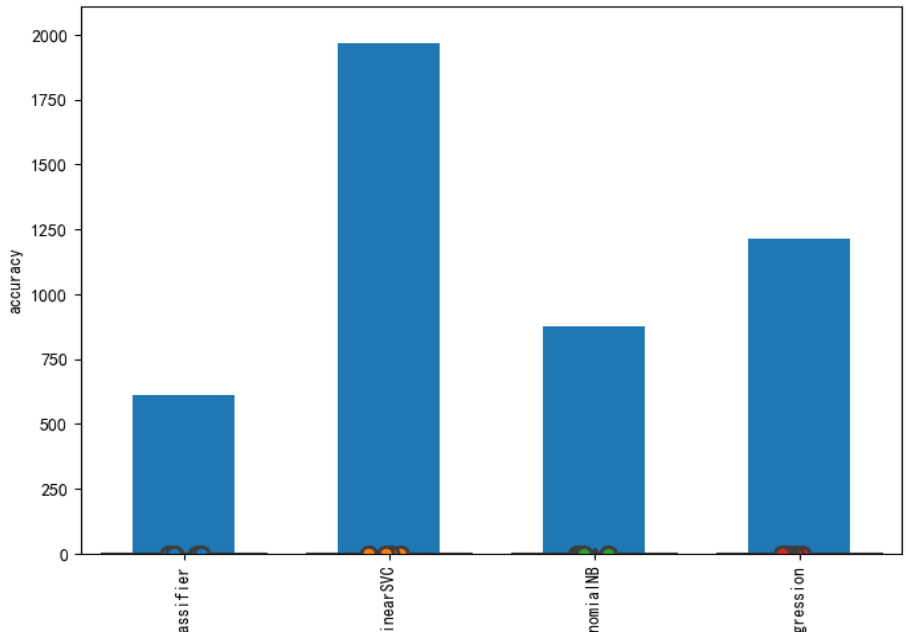


图 8-1 各算法精确率直方图

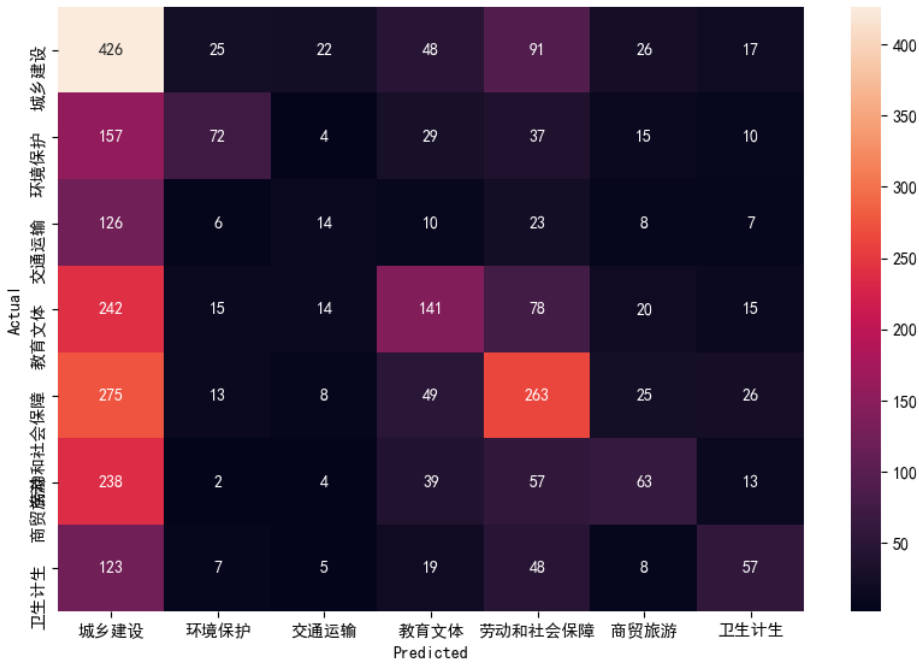
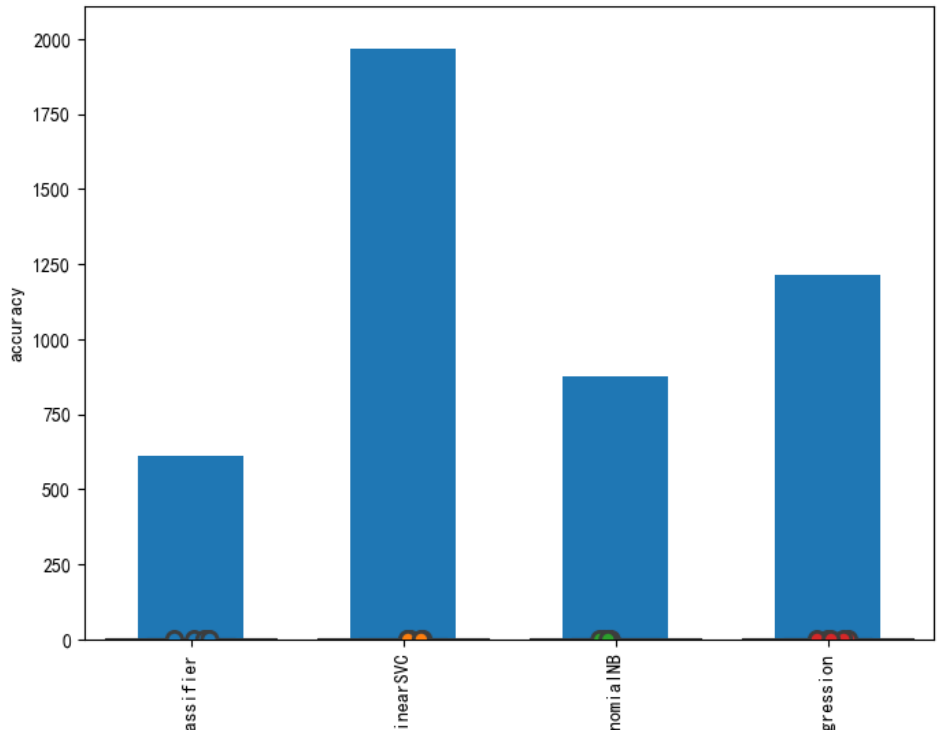
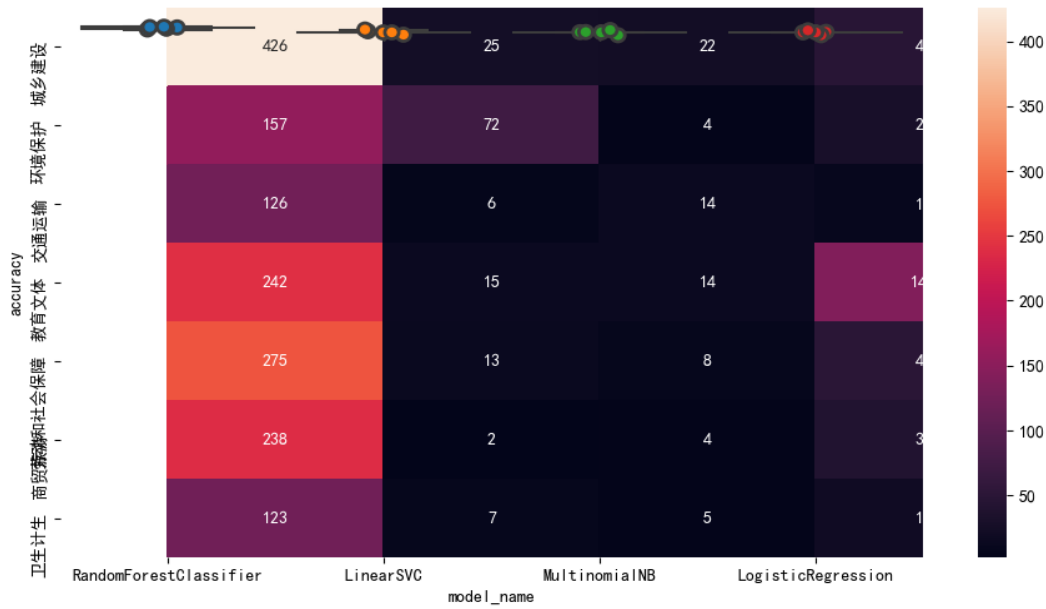


图 8-2 实际的类别与预测的对比图

可看出，使用线性支持向量机构造的模型对留言分类的效果最好，为使模型的精确度更高，我们使用卡方检验来查找与每个类别最相关的词条来对留言进行分类，将模型优化，优化后的模型预测结果如下列图 8-3 和图 8-4 所示。



8-3 优化后的直方图



8-4 优化后的对比图

对模型进行优化后，计算相对较好的线性支持向量机模型的 fit 值，如表 8-1 所示。

表 8-1 问题 1 预测结果评估结果表

问题 1 预测结果评估结果				
	Precision	Recall	F1-score	support
城乡建设	0.27	0.65	0.38	655
环境保护	0.51	0.22	0.31	324
交通运输	0.20	0.07	0.11	194
教育文体	0.42	0.27	0.33	525
劳动和社会保障	0.44	0.40	0.42	659
商贸旅游	0.38	0.15	0.22	416
卫生计生	0.39	0.21	0.28	267
Macro avg (宏平均值)	0.37	0.28	0.29	3040
Weighted avg (加权平均数)	0.38	0.34	0.32	3040
Accuracy (精确)			0.34	3040

8.2 问题二结果分析

对于问题二，我们需要将附件 3 中众多留言识别出相似的留言，并给出排名前 5 的热点问题，同时将特定地点或人群的数据归并，即把相似的留言进行归类。也就是说，我们要将某一时段内群众反映次数较多的某一问题的主题挖掘出来，并对这些主题根据频数进行排序处理。并将排名前 5 的热点问题对应的留言信息整理出来。如表 8-2 所示。

表 8-2 热点词提取结果表

热点词提取结果	
Topic #0	问题 反映 咨询 西地省 幼儿园 a7 开发商 投诉 中学 质量 学生 a4 收费 a2 学院 关于 a6 购房 办理 安置
Topic #1	a7 何时 车位 请问 投诉 公交车 县星沙 大道 小学 a3 是否 销售 物业 违规 为什么 不能 景园 可以 滨河 路口
Topic #2	建议 业主 地铁 解决 a8 请求 规划 拖欠 中心 号线 希望 设置 a9 县星 县泉塘 何时能 公交 县市 怎么
Topic #3	建设 a6 什么 加快 有限公司 公司 时候 涉嫌 公园 为何 拆迁 项目 非法 举报 a3 街道 没有 a2 a5 政策
Topic #4	小区 a3 扰民 严重 a1 a2 a4 噪音 施工 附近 街道 a7 居民 a5 国际 社区 房屋 新城 影响 安全隐患

9.总结

9.1 结论

本文首先对数据进行预处理，选取数据，判断缺失值，修正异常值，对文本进行分词、去停用词，创建字典对象保存类别并对类别之间是否平衡作出判断。基于已有的标签分类，对留言信息数据集进行处理。

为了对留言信息进行分类，我们基于已有的标签分类，利用 TF-IDF 构造词袋模型，并使用卡方统计进行特征选择，然后通过使用逻辑回归算法、朴素贝叶斯算法、线性支持向量机算法以及随机森林四种算法来构建模型，最终实现对留言信息的分类。

为了实现对热点问题的挖掘，我们对留言时间进行了切分处理，并对留言主题使用了 jieba 进行分词。其次创建了停用词词库进行去停用词，生成 TF-IDF 矩阵文档并将其转为了数组形式。最后我们对 K-means 聚类和 DBSCAN 聚类进行了比较，最终选取了效果较好的 K-means 聚类算法得出了热点排名前五的问题，得到附件：“热点问题表.xls”，基于已有的热点问题，我们给出相应热点问题对应的留言信息，得到附件“热点问题留言明细表.xls”。

对了评价答复意见，我们首先对群众留言详情进行特征提取和权重选取，构建一个类似考试的标准答案的框架，然后再根据、分词、特征及其权重和匹配规则来构建一个知识库，将知识库运用到评价意见数据集上，使得评价意见可以自动分词，并基于构建的框架进行推理，实现评价意见与留言详情的匹配评分，以此来对评价意见进行评价。

9.2 回顾与展望

在本次实验的过程中，发现了一些问题，也得到了一些启发，主要可以概括为以下几点：

（1）文本特征的选取是进行文本处理的重要步骤，我们虽然考虑了较为复杂的特征提取方法，但本小组能力有限，最终选取了传统的 TF-IDF 算法，得出的结果有一定的误差。

（2）留言信息具有交叉性，一些含有多个类别特征的留言信息对于留言标签有重大的影响，而我们在分类的过程中没有将这种差别考虑在内。

（3）在数据处理的过程中没有完全考虑到所有无效性的文本，虽然简单剔除了某些无效的词条，但对于未剔除的无效的词条均默认为对文本是有重大贡献的。

（4）对于留言信息的分类和聚类我们均使用的传统算法，虽然目前已经具有更好的深度学习算法，但对于传统算法来说，如果将参数调制适度并予以改进，那么也能取得较好的效果。本小组能力有限，在实践有限的情况下并没有能实现所有的想法。虽说是传统的算法，但并没有我们想象中的那么简单。本次研究主要通过看网上课程视频、找网页研究学习等学习方式，我们的能力得到了锻炼，知识库更加丰富。

数据挖掘是一项有趣的研究，思路和方法很重要。小组想到了很多思路，但是很多时候都没有能力去实现思路，这让我们深刻认识到自身的局限性，感受到需要学习的东西还有很多。值得庆幸的是，虽然由于今年的特殊情况小组的成员虽然各在一方，但小组的氛围一直很好，即使感受到困难做不下去，也一直互相鼓励，团队的分工协作与互帮互助使我们完成了这一问题。团结合作与坚持使成功达成目标成为可能。

致谢

在这篇报告完成之际十分感谢泰迪杯官方给出关于智慧服务的构建赛题，感谢大赛组委会给我们一次锻炼的机会，给予我们创新与展示的平台。

其次感谢我们的指导老师和大赛组委会为我们指明方向，给我们提供了指路的明灯。

最后感谢小组的每一位成员对此次比赛的付出，以及成员的亲友对成员的支持，我们将继续努力。

参考文献

[1] 丁霄云, 刘功申, 孟魁. 基于一类 SVM 的不良信息过滤算法改进[J]. 计算机科学, 2013, 040(0z2):86-90,114.

[2]Keating, Kim a, Cherry, Steve. Use and Interpretation of Logistic Regression in Habitat-Selection[J]. Journal of Wildlife Management, 2011, 68(4):774-789.

[3] I. Rish. An empirical study of the naive bayes classifier[J]. Journal of Universal Computer Science, 2001, 1(2):127.

[4] 都云琪, 肖诗斌. 基于支持向量机的中文文本自动分类研究[J]. 计算机工程, 2002, 28(11):137-138.

[5] 胡睿康 《大学生论文联合对比库》 基于线性无关性的无监督学习算法的应用

[6] 唐龙峰 《大学生论文联合对比库》 基于科技文献的聚类算法设计与实现

[7] 唐龙峰 《大学生论文联合对比库》 基于并行集群的文献聚类算法