基于传统机器学习方法应用于“智慧政务”中的文本挖掘

摘要

在此题中，问题 1 对群众留言内容分类构建一级标签模型，通过该模型减少工作人员的工作量，提升工作效率，该问题的特点是多标签分类。问题 2 是定义合理的热度评价指标，对热点问题进行挖掘，并罗列出排名前五的热点问题和相对应的留言信息，该问题的特点是根据指标来划分热点程度。问题 3 针对答复意见，并从相关性、完整性、可解释性对答复意见的质量给出评价方案，该问题的特点是要从相关性、完整性、可解释性等程度进行求解。

本文针对问题 1，我们认为这是多标签分类的问题，通过特征选择，降低特征空间维数，提高分类精度和效率，根据传统机器学习方法的随机森林模型构建模型 I 。在对 SVM 分类模型改进的基础上建立模型 II。对模型进行了合理的理论证明和推导，然后借助于蒙特卡洛方法(Monte Carlo method)和 Python 软件，对附件中所提供的数据进行了筛选，去除异常数据，对残缺数据进行适当补充，并从中随机抽取了 3 组数据（每组 8 个采样）对理论结果进行了数据模拟，结果显示，理论结果与数据模拟结果基本吻合。最后我们分别对这两个模型使用了 F-Score 评价方法，结果显示，SVM 模型更适合于问题一。

对于问题 2，我们建立了定性指标的数学模型，引入虚拟变量，根据集中于某段时间反映特定地点和特定时间，对指标的定性状态进行离散化。运用 TF-IDF 权值技术提取关键词，并且通过数据预处理、去重、结巴分词，过滤停用词等操作，获取样本的 TF-IDF 权值向量，定位出排名前五的热度问题（见文件“热点问题表.xls”）并保存热点问题对应的留言信息（见文件“热点问题留言明细表.xls”）。

对于问题 3，我们设定了评价指标，并根据回复的相关性、完整性、可解释性针对答复意见的质量建立了加权平均模型，结合模型给出了对于答复质量的评价方案。

关键词：传统机器学习方法；随机森林模型（Random Forest）；SVM 分类模型；TF-IDF 技术； K-means 聚类分析； Python

Text mining in "smart government" based on traditional machine learning method

Abstract

In this question, question 1 constructs a one-level label model for the classification of the content of the public message. Through this model, the workload of the staff is reduced and the work efficiency is improved. The characteristic of this problem is multi label classification. Problem 2 is to define a reasonable heat evaluation index, mine the hot issues, and list the top five hot issues and corresponding message information. The feature of this problem is to classify the hot issues according to the index. Question 3 gives an evaluation plan for the quality of the reply from the aspects of relevance, integrity and interpretability. The characteristic of this question is to solve it from the aspects of relevance, integrity and interpretability.

In this paper, for problem 1, we think it is a multi label classification problem. Through feature selection, we can reduce the dimension of feature space, improve the accuracy and efficiency of classification, and build a model based on the random forest model of traditional machine learning method. Based on the improvement of SVM classification model, model II is established. The model is proved and deduced reasonably, and then the Monte Carlo method is used Method) and python software, the data provided in the attachment is screened, the abnormal data is removed, the incomplete data is supplemented appropriately, and three groups of data (8 samples in each group) are randomly selected from them to simulate the theoretical results. The results show that the theoretical results are basically consistent with the data simulation results. Finally, we use the F-score evaluation method for these two models, and the results show that SVM model is more suitable for problem one.

For problem 2, we establish a mathematical model of qualitative indicators, introduce virtual variables, and discretize the qualitative status of indicators according to the focus on a certain period of time to reflect a specific place and a specific time. Using TF-IDF weight technology to extract keywords, and through data preprocessing, de duplication, word segmentation, filtering stop words and other operations, obtain TF-IDF weight vector of samples, locate the top five heat problems (see the file "heat problems table. XLS") and save the message information corresponding to the heat problems (see the file "heat problems message table. XLS").

For question 3, we set up the evaluation index, and according to the relevance, integrity and interpretability of the reply, we set up a weighted average model for the quality of the reply. Combined with the model, we give the evaluation scheme for the quality of the reply.
Keywords: traditional machine learning method; random forest; SVM classification model; TF-IDF technology; K-means clustering analysis; Python

一、问题重述

近年来，随着网络问政平台的兴起，政府从网络平台了解的民情越来越多，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。同时，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，对提升政府的管理水平和施政效率具有极大的推动作用。而现在，我们可以利用自然语言处理和文本挖掘的方法解决上述所说的难题。。在此题中，我们需要完成群众留言分类，热点问题挖掘，答复意见的评价三个问题。

对于问题一，需按照一定的划分体系对留言进行分类，所以建立关于留言内容的一级标签分类模型。导入附件 2 给出的数据，进行分类。对于问题二，根据附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果。对于问题三，针对附件 4 相关部门对留言的答

复意见，从答复的相关性、完整性、可解释性等角度对 答复意见的质量给出一套评价方案，并尝试实现。

二、问题分析

（一）问题 1 的分析

大数据时代下，智慧政务即通过“互联网+政务服务”构建智慧型政府，利用云计算、移动物联网、人工智能、数据挖掘、知识管理等技术，提高政府在办公、监管、服务、决策的智能水平，形成高效、敏捷、公开、便民的新型政府，实现由“电子政务”向“智慧政务”的转变。运用互联网、大数据等现代信息技术，加快推进部门间信息共享和业务协同，简化群众办事环节、提升政府行政效能、畅通政务服务渠道，解决群众“办证多、办事难“等问题【1】。研究问题 1 有助于减少工作人员的工作量，提高工作效率，降低出错率，并且不需要凭借人工经验对问题留言进行分类。

问题 1 属于多标签分类数学问题，而针对此类问题，我们经常是将其视为基于单个优化问题的多标签分类问题，其基本思想为只建立一个最优化问题直接处理数据集中的所有样本。而多标签数据集中地样本往往拥有多个标签，该算法实现虽有一定难度，但其优点是没有改变数据集的结构，也没有破坏类别之间的关联关系，并且还能反映多标签分类的特殊性质。【2】

问题 1 仅要求构建一级标签分类模型，我们首先想到要对留言中的特征进行分类选择，因而打算采用决策树方法，但考虑到分类留言的数据集过大，可能需要构建多个决策树，而如果有个别决策树因为异常值的影响便会导致预测不准确，且决策树需要采用所有的特征集样本，容易出现对训练集具有很好的效果，但对数据集效果很差，经过权衡利弊，我们选择以随机森林模型（Random Forest）为基础建立模型Ⅰ。而经过进一步的深入随机森林模型我们发现，它的计算量过大，即应用于智慧政务的时候，往往会因为计算量大，而导致无法做到对实时数据的预测。于是我们又在模型 I 的基础上以建立模型Ⅱ，并向 SVM 分类模型靠近。我们使用 Python3 对结果分别进行预测，并将结果进行比较。

（二）问题 2 的分析

在某一时间段内群众集中反映的某一问题留言称为热点问题，而及时发现热点问题并解决，能够提升工作效率且提升群众满意度。而针对这一问题问题 2，考虑到这是热度问题，出现的关键词次数会比较多，我们选取了适度指标并对其正向化，然后对指标进行均值化，消除指标间两极不同的影响，使各个指标转化成可以直接加减的数值，再使用层次分析法（AHP），将主客观结合在一起得出指标权重。最后运用综合指数法，将实际值与标准值进行对比后再使用线性综合汇总得到综合评分。成功定义指标后，我们认为，某个词在文本中出现的越多，则热度越高，于是我们获取了 TF-IDF 权值向量。根据指标在 Python 上对附件 3 进行去重、结巴分词、过滤停用词，最终得到排名前五的热度问题（见文件“热点问题表.xls”）和热点问题对应的留言信息（见文件“热点问题留言明细表.xls”）。

（三）问题 3 的分析

我们发现问题 3 中相关部门对留言的回复时间大多在半个月内，但系统显示的时间却会出现延迟，从一个月到两个月不等，这对群众满意度会大打折扣，这也是当前许多投诉回复慢的原因之一，根据可行性、相关性以及可解释性我们在加权平均模型上建立了模型III。

三、模型假设

1. 假设题目所给的数据真实可靠
2. 假设问题 1 中的留言信息都是有效的
3. 假设问题 2 中的权值向量是正确的

四、定义与符号说明

符号	定义与说明
$I(x)$	表示随机变量的信息
$p(x_i)$	表示当 x_i 发生时的概率
P_i	表示第 i 类的查准率
R_i	表示第 i 类的查全率
M	超平面

五、模型建立与求解

问题 1 的数学模型

（一）模型 I （随机森林的数学模型）

$$I(X = xi) = -\log_2 p(x_i)$$

模型II（SVM分类模型）

$$M = \arg_{\alpha,\beta} \max(\frac{1}{\|\alpha\|})$$

问题2

问题2中我们使用了TF-IDF权值向量，TF-IDF是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF加权的各种形式常被搜索引擎应用，作为文件与用户查询之间相关程度的度量或评级。除了TF-IDF以外，因特网上的搜索引擎还会使用基于链接分析的评级方法，以确定文件在搜寻结果中出现的顺序。[3]然后使用了Python的结巴分词以及去除停用词，部分停用词表如下：

说	人	元	hellip	&	，	？	、	。	"	"	《	》	！	，	：	；	？	""
的	了	人民	末	啊	阿	哎	哎呀	哎哟	唉	俺	俺们	按	按照	吧				
吧哒	把	罢了	被	本	本着	比	比方	比如	鄙人	彼	彼此	边						
别	别的	别说	并	并且	不比	不成	不单	不但	不独	不管								
不光	不过	不仅	不拘	不论	不怕	不然	不如	不特	不惟									
不问	不只	朝	朝着	趁	趁着	乘	冲	除	除此之外	除非	除了							

六、模型评价与推广

根据问题1所提的要求，通常使用F-Score评价方法对分类模型进行评价：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i}$$

七、结论

此次比赛，我们是抱着尝试的心态来参与，由于自身能力不足，以及对数学建模的了解不够深入，所以导致模型没有构造成功，还请主办方原谅。我们秉着以赛促学的心态进行学习，虽然此次比赛没有取得成绩，但是下一次比赛我们力

取建造出模型。

七、参考文献

[1] 中国安防行业网．智慧政务的基本概念及其发展概述 http://news.21csp.com.cn/c16/201805/11370033.html
[2] 中国农业大学数学专业多标签分类问题的解法综述.https://wenku.baidu.com/view/26d0873267ec102de2bd8991.html
[3]https://baike.baidu.com/item/tf-idf/8816134?fr=aladdin