**目 录**

[“中储智运杯”智慧物流大赛货物智能分类模型介绍 2](#_Toc32144)

[背景简介： 2](#_Toc27637)

[建模思路： 2](#_Toc16910)

[建模过程： 3](#_Toc13016)

[数据清洗： 3](#_Toc17001)

[构造线下验证集： 3](#_Toc31828)

[文本向量化： 4](#_Toc10603)

[模型构建： 5](#_Toc5007)

[模型效果分析与改进： 5](#_Toc15128)

[模型改进展望： 6](#_Toc9230)

[预测分析： 6](#_Toc22062)

[总结： 7](#_Toc5558)

[附件内容与含义： 7](#_Toc16022)

# **“中储智运杯”智慧物流大赛货物智能分类模型介绍**

参赛人：台延鑫

南京大学数学系应用统计硕士

# 背景简介：

首先感谢中储智运公司给出的这次比赛机会，让我接触到了真实的数据与业务场景。我选择的是第一题，要求根据提供的货物名称数据、基础分类数据、以及两者的对应关系，建立具有自我学习优化的算法模型。

题目要求货物名称文本进行分类。根据观察，货物名称大部分为专有名词，且货物种类品目繁多。针对此问题，我个人感觉难点有两个：一是货物名词较为专业化，且缺少先验数据（即无法使用语料库来对货物名称的含义进行解析，比如word2vec方法）；二是货物种类数目过多，且有些类别较偏僻，难以直接通过监督学习训练多分类模型。

# 建模思路：

对于此赛题，我的思路如下：

首先，对货物名称文本进行清洗、分词、编码，将每个货物名称向量化；然后，选取相似度指标构建文本两两之间的距离；最后，借鉴K近邻的思想，为每个货物名称匹配最相似的文本，将其归于相似货物所在的类别。

这个想法比较粗糙，下面将介绍我的尝试与建模过程。

# 建模过程：

## 数据清洗：

货物名称中，除了汉字，还包括一些数字与字母。但经过观察，数字与字母多用来区分不同型号，并不会影响具体类别。为了防止其干扰编码，去掉了货物名称中除汉字的其他字符。

去掉了训练数据中类别待定的数据行。

对包含上一级类别，缺失更细粒度类别的数据行进行了填充。

将一级分类中，‘其他’与‘其它’字段进行了统一。

对相同的训练集数据进行了去重。

## 构造线下验证集：

模型需要通过线下的检验来进行调节验证，合理构造线下验证集是一个非常重要的步骤。

经过观察，训练集中有些三级类别只包含一种商品。那么按照K近邻的方法，这些商品一定不会被正确识别，所以未将其包含在验证集中。

对于其他的数据，都作为了验证集，共有12621条。线下评价指标为验证集三级分类的正确率。

## 文本向量化：

在数据经过清洗后，所有文本都变成了纯文字形式。接下来需要对文本进行分词以及编码。

在分词方面，做了一些尝试思考。比如是否只使用分词？使用哪种模式分词？为了提升效果，分词后是否要赋予名词更多的权重？分词是否要加入字符级字段？

经过尝试与比较后，使用了python的jieba库进行分词，选取的分词模式为cut\_all=True（即返回文本中所有可能的词汇，比如‘供电设备’，分词后会返回'供电'、 '电设备'、 '设备'这些所有可能被切割出的词汇）。

由于货物专有名词较多，分词不够精准，故在分词中加入了组成词的所有字符。比如‘供电设备’，最终返回'供电', '电设备', '设备', '供', '电', '设', '备'共七个字段。

编码时，选择的是词袋模型。该模型未考虑字符顺序与出现频率，后续可以进行改进。

最终，每个货物名称文本变成了一个长度为3175的True/False向量。True的含义为该字段在向量中出现，False反之。

## 模型构建：

在文本已经完全编码后，接下来要进行预测分类。这里采用的是K近邻方法，选取该货物名称距离最近的K个货物，然后归于相同的类别。

这里有两个需要确定的地方：一是距离的定义；二是K值的选取与如何按K个最近邻的类别来确定最终分类的类别。

在定义距离时，做了一些尝试。表现较好的有jaccard、sokalsneath距离，最后使用的为sokalsneath距离。sokalsneath距离的计算公式为NNEQ / (NNEQ + 0.5 \* NTT)，NNEQ为两个向量中不相等（一方为False，一方为True）的维度数目，NTT为两个向量中都为True的维度数目。

在尝试K值的过程中，最终还是选择了K=1，即只依据最近一个邻居来归类。

由于直接计算向量两两距离的时间复杂度与空间复杂度过高，最后选取了BallTree的方式来构造K近邻。

# 模型效果分析与改进：

最终，验证集上的正确率为93.075%，可以看出效果较为可靠。

对分类错误的数据进行了分析，发现如下情况：

训练数据本身存在一些问题，存在名称极其接近但类目不同的情况，导致了错判的发生；

缺乏先验知识导致的分类错误：比如‘百合’与‘冷合板’、‘枕头’与‘罐头’匹配到一起；

定语相同但被修饰的名词不同：比如‘保温棉’与‘保温罐’；

# 模型改进展望：

增强训练集数据质量与数量，避免分类不一致的情况，同时增加一些先验的知识；

改进分词质量，可以人工增加分词；

增加名词权重（尝试了但不太奏效，可能方法不太对）；

寻找其他语料库或数据以增加货物信息（比如给的分类代码表）；

# 预测分析：

利用训练集的数据（去除重复与类别未知的数据行），按照之前模型进行训练，来预测每个货物名称的类别。最终结果包含在pre\_result.csv中。

# 总结：

本次建模对文本进行了清洗、分析、编码，利用KNN模型根据向量相似度距离对货物名称进行了分类。可以探索的方法还有很多，这里限于个人能力与时间，只是做了一些简易解法。

# 附件内容与含义：

pre\_result.csv:对5000种货物分类的预测结果

classify.ipynb:ipython脚本，包含建模过程，可直接运行