中南大學
CENTRAL SOUTH UNIVERSITY

文本分类步骤
学习报告

学生姓名	maybeLocalhost
学　　号	
专业班级	
指导教师	
学　　院	计算机学院
完成时间	2021.04

目录

一、自然语言处理

自然语言处理的流程大致可以分为以下五个步骤：

(1) 通过网络爬虫或本地导入等方式获取文本。

(2) 对文本进行预处理，将文本进行分词并取出其中的语气词和停用词。

(3) 对文本进行特征化处理，使用独热编码或词嵌入技术，将词语映射成其对应的词向量形式。独热编码又称之为一位有效编码，即 N 位寄存器与 N 个状态一一对应，寄存器中只有一位有效；词嵌入技术将词语量化到低维度的稠密向量空间，维数固定，相比于独热编码，效率更高。

(4) 针对模型进行训练，可以使用基于支持向量机、决策树、临近算法或逻辑回归等机器学习模型算法，也可以使用基于卷积神经网络或循环神经网络等基于深度学习的模型算法。

(5) 使用测试集对训练模型进行验证评估模型算法的优劣。

针对上述五个步骤进行简单的介绍：语料库的收集大多采用网络爬虫或本地文本数据集。语料预处理阶段主要包括对收集来的语料库进行语料清理、分词、词性标注和去停顿词等操作。在特征化环节，需要对完成预处理的文本进行向量化，将完成分词的词语表示成向量形式，以便计算机能够对其进行计算。这样的操作有助于通过向量的表达方式，发现不同词语之间的相似关系。在模型训练环节，使用的训练方法包括传统的有监督、无监督和半监督学习模型等，具体使用的模型需要根据不同的应用场景进行选择。针对建模后的效果进行评价，常用的效果评估指标有准确率、召回率等。

二、文本分类

2.1 文本分类的原理

文本分类的原理是在已有数据基础上，总结归纳出一个分类函数或者是构建出一个分类模型（即分类器），通过分类函数或分类器，可以将数据库中的数据根据其特征，相应地分到具体的某一个类别中，并最终将其应用于数据的预测。目前常见的分类器算法有：朴素贝叶斯分类器（NBC）、临近算法（KNN）、支持向量机（SVM）和基于卷积神经网络的分类模型。

2.2 文本分类的步骤

文本分类属于有监督的学习，其流程图和具体步骤如下：

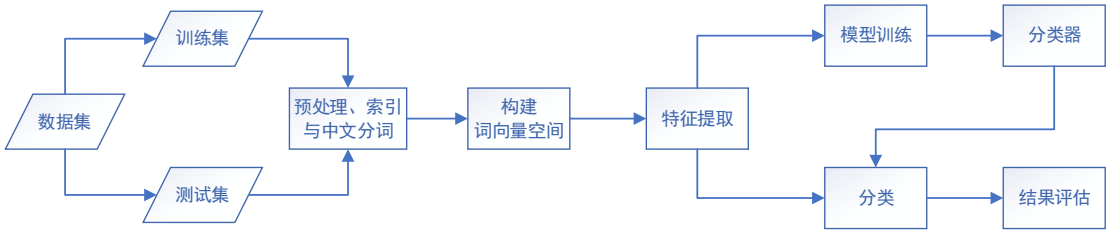

图 1 文本分类流程图

(1) **数据集分类**：将数据集分为训练集和测试集两部分，训练集负责模型训练，测试集负责进行验证评估。

(2) **预处理**：对数据集文本进行预处理操作，去除文本噪声信息，如 HTML 标签、文本的格式转换、检测句子边界等。

(3) **索引**：将文档分解为基本处理单元，同时降低后续处理的开销。

(4) **中文分词**：使用中文分词器将文本进行分词并去除其中的语气词和停用词，常见的中文分词器有中科院分词、哈工大分词、jieba 分词等。

(5) **构建词向量空间**：统计文本词频，生成文本的词向量空间。

(6) **特征提取**：根据生成的词向量空间，提取出区分文本类别能力较强的特征词语。常见的提取特征词语的方法有：信息增益、TF-IDF、卡方校验。

(7) **训练分类器**：模型训练阶段可以使用基于支持向量机、决策树、临近算法或逻辑回归等机器学习模型算法，也可以使用基于卷积神经网络或循环神经网络等基于深度学习的模型算法。选取某一种分类器模型对训练集数据进行训练，在训练的过程中调整参数，以取得更好的分类效果。

(8) **分类结果评估**：使用测试集对训练模型进行验证评估模型算法的优劣。

2.3 词向量技术

词向量技术是自然语言处理中语言建模和特征学习技术的统称，主要目的是将人类的自然语言转化为便于计算机运行的数据，建立词语与实数向量之间的映射关系。目前常见的词向量技术为独热编码和词嵌入技术，下面将分别简要介绍一下这两种方法。

1. 独热编码技术

独热编码技术是一种通过离散特征取值将自然语言进行向量化的技术。在对离散型特征或者分类值数据进行向量化时，具有较好的效果。独热编码使用 N 位寄存器与 N 个状态一一对应，且寄存器中只有一位有效，又称之为一位有效编码。使用独热编码进行自然语言处理时，典型的用法是通过独热编码表述属性的某一个特征。例如，在记录学生成绩时，可将“优”、“良”、“达标”、“不达标”，这四种等第编码成“1000”、“0100”、“0010”和“0001”这四种。编码中，只有一位为 1，代表着独热编码中只有一位有效编码。在针对文本中的词语进行独热编码时，需要先统计该文本中出现的词语的数量，该数量将确定了独热编码后向量的维数。然后。将词语的编码成只有一位为 1，且相互之间乘积为 0 的词向量。完成了词向量的构建，就可以通过该向量集来表示一句话或者一个文本。

独热编码解决了分类器中数据属性不好处理的问题，在一定程度上扩充了数据的特征。当数据的属性或列表数量很多时，特征空间就会变得很大，硬件

运算的时空复杂度也将增加，因此其适用于文本数据集体量较小的应用场景。当文本数据集中的词语量过多，使用独热编码进行词语向量化时，将带来一些问题。例如，将词语进行独热编码后，由于所有编码后所得的词向量都只有一位为 1，所以词与词之间的矢量距离是相同的，这就无法体现词语之间的关联性。同时，在使用独热编码对词语进行向量化时，就默认了词与词之间的关系是相互独立的，但在大多数情况下，词与词之间是相互影响的。为解决上述这些问题，基于稠密空间的词嵌入技术应运而生。

2. 词嵌入技术

该技术的特点是将所有的词语表示成低维度的稠密向量，通过固定维数的连续向量空间，可以评估出词与词之间的相似性。其中，相似程度越大的词语，其词向量在向量空间的夹角就会越小。固定的向量维度，相比于独热编码将大大提高计算机处理数据的效率。目前词嵌入技术常用的模型是跳字模型和连续词袋模型。

（1）跳字模型

跳字模型（SGM）的主要作用是基于文本中的某个词语来生成该词语周围的若干个词语。在进行 SGM 模型训练时，所有词向量都将被表示成两个固定维数的词向量来计算条件概率。进行词向量训练前，需要事先设置背景词窗口大小。在词向量训练的过程中，每个词语将作为中心词和背景词两种角色，分别将作为中心词和背景词时的向量大小存放在对应的两个词向量中。假设文本序列由$\{w_1, \ldots, w_n\}$，n 个词语组成。以w_c作为中心词，设置背景窗口的大小为 2。跳字模型的训练过程是，根据中心词w_c，生成与其距离不超过两个词语的背景词w_{c+1}、w_{c+2}、w_{c-1}和w_{c-2}的联合概率，计算公式为：

$$P(w_{c+1}, w_{c+2}, w_{c-1,} w_{c-2} | w_c)$$

(2) 连续词袋模型

连续词袋模型（CBOW）通过文本中通过给定数量的背景词来生成中心词。假设以w_c作为中心词，设置背景窗口的大小为 2，CBOW 模型研究的重点在于，通过给定背景词$w_{c-1}, w_{c-2}, w_{c+1}$和$w_{c+2}$，计算生成中心词$w_c$的条件概率，计算公式为：

$$P(w_c|w_{c-1}, w_{c-2}, w_{c+1}, w_{c+2})$$

2.4 特征提取算法

- **卡方检验**是通过计算每个特征和每个类别的关联程度，然后选择那些关联程度高的特征来实现降维。其基本思想就是衡量实际值与理论值的偏差来确定理论的正确与否。
- **信息增益**是通过计算每个特征对分类贡献的信息量，贡献越大信息增益越大，然后可以选择那些信息增益较高的特征实现降维。
- **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。其主要思想是：如果某个词或短语在一篇文章中出现的频率（TF）高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。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。

2.5 分类器训练算法

1. 临近算法

临近算法（KNN）是机器学习分类技术中一个较为简单的分类算法。其中参数 K 表示 K 个最接近的邻居，即每个样本所属的类别都可以通过最接近它的 K 个邻居来进行判断。在判定某个样本所属的类别时，仅根据最邻近的 K 个样

本的类别来决定待分类样本所属的类别，而不需要其他判别条件作为支撑。由于 K 是一个有限的整数，那么当数据集的数量较为庞大时，在使用 KNN 算法进行类别决策时，只有极少数的样本参与决策。因为该算法主要依靠待分类样本周围有限个相邻样本进行判别，而不依靠判别类域的方法来确定，所以算法简单高效，运行速度快。

KNN 算法通过维护一个数量为 K 的优先级队列来进行分类。通过计算待预测样本与数据集中已有样本的矢量距离，来维护优先级队列。优先级队列中将保存 K 个距离由大到小的临近数据。通过遍历训练集，计算待预测样本和已有样本之间的距离，若所得距离大于等于优先级队列中距离最大的样本，将舍弃该数据；反之，若距离小于优先级队列中最大的样本，则删除优先级队列中最大距离的元素，并将该元素存入优先级队列。在将数据集遍历完毕后，通过统计优先级队列中元素最多的类别，即为待预测元素的类别。

2. 贝叶斯分类器

贝叶斯分类器是各种分类器中分类错误概率最小或者在预先给定代价的情况下平均风险最小的分类器。它的设计方法是一种最基本的统计分类方法。其分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。

2.6 评价指标

- TP：True Positive
- FP：False Positive
- TN：True Negtive
- FN：False Negtive

以上四个定义是基础，Positive 表示对样本作出的是正的判断，T 表示判断正确，F 表示判断错误（Negtive 类似）。比如 TP 表示样本为正，我们模型也判断为正，FP 则表示模型判断为正，但是判断错误，样本为负。下面是常用的四个评价指标：

- **Accuracy** = （TP+TN）/（TP+FP+TN+FN）

  准确率，表示在所有样本中分对（即正样本被分为正，负样本被分为负）的样本数占总样本数的比例。

- **Precision** = TP /（TP+ FP）

  精确率，表示模型预测为正样本的样本中真正为正的比例。

- **Recall** = TP /（TP + FN）

  召回率，表示模型准确预测为正样本的数量占所有正样本数量的比例。

- **F1** = 2*P*R /（P+ R）

  F1，是一个综合指标，是 Precision 和 Recall 的调和平均数，因为在一般情况下，Precision 和 Recall 是两个互补关系的指标，因此通过 F 测度来综合进行评估。F1 越大，分类器效果越好。

三、参考文献

[1] 何铠. 基于自然语言处理的文本分类研究与应用[D].南京邮电大学,2020.

[2] Google. 机器学习术语表. https://developers.google.com/machine-learning/crash-course/glossary?hl=zh-cn#r

[3] 于小勇. 二分类相关评估指标（召回率、准确率，精确率，f1，auc 和 roc）. https://blog.csdn.net/weixin_36670529/article/details/84302609

[4] ECHO. NLP ATC （automation text classification） 文本分类. https://houbb.github.io/2020/01/20/nlp-atc#%E5%9B%9B%E6%96%87%E6%9C%AC%E5%88%86%E7%B1%BB%E4%BC%A0%E7%BB%9F%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0%E6%96%B9%E6%B3%95

[5] Walter_Jia. 【文本分类】文本分类流程及算法原理. https://blog.csdn.net/jiayanhui2877/article/details/19764317

[6] Bao Xizao. Machine Learning 学习笔记之中文文本分类. https://baoxizhao.com/2017/05/20/MachineLearning%E5%AD%A6%E4%B9%A0%E7%A

C%94%E8%AE%B0%E4%B9%8B%E4%B8%AD%E6%96%87%E6%96%87%E6%9C%AC%E5%88%86%E7%B1%BB/

[7] marsjhao. 机器学习算法笔记之 4：贝叶斯分类器. https://blog.csdn.net/marsjhao/article/details/69055634