基于混合编码方式和textRCNN模型的电力数据智能分类方法

1摘要

本发明公开了一种基于混合编码方式和textRCNN模型的电力数据智能分类方法，包括：数据预处理，对原始电力数据进行清洗和标准化处理，确保数据的一致性和完整性；多编码生成，使用多种编码方式（如词袋模型、TF-IDF、词嵌入、one-hot编码等）对预处理后的数据进行编码，生成相应的向量表示；向量拼接，将不同编码方式生成的向量拼接成一个新的综合向量，相对于传统的编码方式，特征更加丰富；textRCNN模型处理，使用拼接后的综合向量作为输入，训练textRCNN模型；得到分类结果，训练完成后，将需要分类的数据通过上述方式生成一个综合向量，经过模型处理得到最终分类结果；通过实验验证，该系统在电力数据分类的任务中表现出更高的准确性和鲁棒性。

2技术领域

本发明属于文本处理技术领域，具体涉及一种基于混合编码方式和textRCNN模型的电力数据只能分类方法。

3背景技术

在电力行业中，随着智能电网的发展和数字化转型的推进，积累了大量的结构化数据。这些数据包括设备运行参数、维护记录、故障报告、用户用电记录等。如何高效地处理和利用这些数据，提高电力系统的运行效率和安全性，成为了一个重要的研究课题。电力数据的特点是数据量大、类型多样且信息丰富。这些数据不仅包括数值型数据（如电流、电压），还包括类别型数据（如设备类型、故障类型）和日期型数据（如维护时间、故障发生时间）。传统的方法主要依赖于手工规则和专家经验，这种方法在面对大规模、复杂的数据时，效率低下且容易出错。此外，现有的许多方法在处理多维度、高复杂度的结构化数据时，往往难以全面捕捉各个特征之间的关联性，导致分类和分级的准确性不高。一些复杂的深度学习模型虽然在理论上表现良好，但在实际应用中往往面临训练困难、计算资源消耗大等问题，尤其是在处理大规模数据时，模型的训练时间和计算资源需求显著增加。

为了解决上述问题，本发明提出了一种基于混合编码方式和TextRCNN模型的电力数据智能分类方法。该方法首先对原始电力数据进行清洗和标准化处理。再使用one-hot编码，bags of words编码和。通过特征选择和融合技术，将多个特征向量拼接或降维，生成最终的输入向量。在特征向量生成后，使用TextRCNN模型进行分类和分级。TextRCNN模型结合了循环神经网络（RNN）和卷积神经网络（CNN）的优点，能够有效捕捉特征之间的依赖关系和局部特征，提高了模型的表达能力和分类准确性。与传统的单一编码方式相比，混合编码方式能够更全面地表示数据特征，避免了单一编码方式的局限性。与复杂的深度学习模型相比，TextRCNN模型在保持高精度的同时，减少了训练时间和计算资源的消耗，更适合大规模数据的处理。此外，本方法不仅适用于电力数据的分类和分级，还可以扩展到其他领域的结构化数据处理，具有较强的适应性和泛化能力。通过这种方式，本发明能够在不增加过多计算资源的情况下，实现对电力数据的高效、准确分类和分级，为电力系统的智能化管理和运维提供有力支持。

4 发明内容

在智能电网和数字化转型的背景下，电力行业面临着大量结构化数据分类的问题，常见的方法都是对模型进行改进，而没有对编码方式进行变动，总是采用单一的编码方式，无法表示数据的丰富特征。为了解决这一问题，本发明提出了一种基于混合编码方式和TextRCNN模型的电力数据智能分类方法，具有准确性和鲁棒性等显著优点。该方法能够综合利用多种编码技术，全面捕捉电力数据中的关键信息，借助混合编码方式，我们的系统能够生成更丰富的特征表示，而TextRCNN模型则有效捕捉特征之间的依赖关系，从而实现更高的分类准确率和更好的模型适应性。本发明通过以下技术方案达到上述目的，包括如下步骤：

1 数据预处理: 我们对原始电力数据进行清洗和标准化处理。数据清洗的具体步骤包括去除冗余的记录、填补缺失值以及处理异常值，以确保数据的一致性和完整性。标准化处理则针对数值型数据，采用归一化方法，使得不同特征在相同的尺度上进行比较，进而提高后续模型训练的效率和稳定性

2 多编码生成: 在本发明中，多编码是最核心的环节，采用多种编码方式生成特征向量，可以结合多种编码方式的优势，充分挖掘数据的潜在信息，首先，我们可以使用词袋模型，该模型通过统计文本中每个词出现的频率，将文本转换为一个固定长度的向量。这种方法能够有效提取出文本数据中的关键词，并为后续分析提供基础。然而，词袋模型忽略了词汇之间的顺序和上下文关系。为了解决这一问题，我们进一步引入了TF-IDF编码。TF-IDF不仅考虑词汇在单个文档中的重要性，还通过计算逆文档频率来降低常见词汇的权重，从而强调在特定上下文中更具代表性的词。这使得模型能够更好地识别出影响分类结果的关键特征，增强分类的准确性。此外，我们采用词嵌入技术，通过将词语映射到稠密的向量空间，使得具有相似语义的词在向量空间中靠得更近。通过这种方式，模型能够捕捉到词语之间的深层次关系，如同义词和上下文关联性，进而提高对复杂文本数据的理解能力。同时我们也采用One-hot编码。这种方法为每个类别分配一个独立的二进制向量，确保每个类别都有唯一的表示。这一方式不仅使得模型能够处理离散数据，还避免了类别之间的顺序关系可能带来的误导。通过结合多种编码方式，我们能够从多个维度对电力数据进行全面的表示。这种综合的特征生成策略，不仅丰富了数据的表达形式，还增强了模型对复杂数据特征的捕捉能力，为后续的分类任务奠定了坚实的基础。

3 向量拼接：在特征向量的拼接过程中，我们将所有编码方式生成的向量合并为一个一维向量，以便为后续的模型训练提供统一的输入格式。在这一过程中，我们首先对每种编码方式生成的多维向量进行处理，使其能够有效地拼接。具体而言，对于词袋模型和TF-IDF生成的向量，这些通常是固定长度的稀疏向量，因此可以直接用于拼接。对于词嵌入技术生成的稠密向量通常具有较高的维度，我们将它展开成一维向量，按照顺序进行拼接。在处理One-hot编码时，每个类别的向量也是一维的，且长度与类别数量相等。这些向量同样可以直接拼接到特征向量中，最终，所有这些处理后的向量将按顺序拼接成一个一维综合向量，形成一个具有丰富信息的特征表示。

4 TextRCNN模型处理：如前文所说，因为电力数据的上下文关联性不是特别强，而且对于我们的编码技术来说，有很强的局部特征，因此使用TextRCNN比较合理，TextRCNN模型结合了CNN和RNN的优点，能够有效提取数据中的局部特征和上下文依赖关系，从而提高分类的准确性和鲁棒性。首先，构建TextRCNN模型的输入层。我们将之前生成的综合一维特征向量输入到模型中。该向量包含了多种编码方式的拼接结果，确保模型能够接触到电力数据的丰富特征。在这一阶段，数据首先经过嵌入层，将输入特征转化为稠密向量，以便于后续处理。接下来，模型的卷积层负责对输入的特征向量进行局部特征提取。我们使用多个卷积核对输入向量进行卷积操作，以捕捉不同大小的特征模式。通过设置不同的卷积核大小，模型能够从不同的视角分析数据，提取出关键的局部特征。这一过程能够有效识别文本中的重要信息。在卷积层之后，应用激活函数引入非线性特征，使得模型能够更好地学习复杂的特征关系。随后，通过池化层，进一步降低特征维度，减少计算复杂性，同时保留重要的特征信息。这一阶段的目的是提高模型的鲁棒性，防止过拟合。接着，输出的特征经过拼接，进入RNN层。RNN层主要用于捕捉特征之间的时间依赖关系。通过循环连接，RNN能够处理序列数据，在分类任务中捕获上下文的依赖性。为了增强模型的学习能力，我们可以在RNN层中使用LSTM）或GRU结构。这些结构能够有效缓解传统RNN在长序列学习中的梯度消失问题，提高模型对长依赖信息的记忆能力。最后，经过卷积层和RNN层的特征提取后，模型输出的特征通过全连接层进行整合，最终生成分类结果。全连接层负责将提取到的特征映射到具体的分类标签上，使用Softmax激活函数来输出每个类别的概率分布。模型的损失函数采用交叉熵损失，以评估预测结果与真实类别之间的差异。训练过程中，我们使用Adam优化器来调整模型参数，以实现更快的收敛速度和更好的性能。

6 具体实施

* 1. 数据预处理：首先对数据进行去重，对各种类型的数据的格式进行统一，然后对文本分词
  2. 多编码生成：
     1. 词袋编码：历文本数据集，统计所有不同单词，建立一个全局词汇表,针对每条记录，创建一个长度等于词汇表大小的向量，每个元素代表对应词汇在该记录中的出现次数。
     2. TF-IDF编码：对于每条记录，计算每个词的词频（TF），对应的值是该词在记录中出现的次数除以记录总词数，同一个词语在长文件里可能会比短文件有更高的词频，而不管该词语重要与否，然后计算逆文档频率（IDF），一些通用的词语对于主题并没有太大的作用， 反倒是一些出现频率较少的词才能够表达文本的分类， 所以单纯使用是TF不合适的。权重的设计必须满足：一个词预测分类的能力越强，权重越大，反之，权重越小。所有统计的文本中，一些词只是在其中很少几个文本中出现，那么这样的词对文本的分类的作用很大，这些词的权重应该设计的较大。IDF就是在完成这样的工作。最后将每个词的TF与IDF相乘得到最终的TF-IDF值。

其中表示单词在文档中出现的次数，那么就是表示单词在文档中出现的频率

其中表示所有文档的数量，表示包含单词的文档的数量，加1是因为 要防止分母为0

* + 1. 词嵌入：使用预训练的通用Word2Vec模型，并在其基础上进行微调，以便更好地适应我们的需求，训练完成后，从模型中提取所有词汇及其对应的高维向量表示。接下来我们可以使用前面创建好的词汇表，它包含所有出现在训练集中的唯一单词及其对应的整数索引。通过这个词汇表，这一映射关系用于后续将文本转换为数值表示。同时，创建一个查找表，其中每个词汇对应着它的高维向量表示。这个查找表将在后续步骤中用于快速获取词嵌入向量。
    2. One-hot编码：和词袋编码类似，每个元素代表对应词汇在该记录中是否出现，1代表出现0代表未出现。
  1. 向量拼接：按照固定顺序先词袋编码,，接着后、TF-IDF再词嵌入最后one-hot将所有编码后的向量先展开成一维，然后连接起来形成一个综合向量。确保所有向量的顺序和维度一致以便于后续模型输入。
  2. TextRCNN模型构建与训练：在本发明的方法中，我们使用了TextRCNN模型来处理混合编码后的电力数据，TextRCNN模型结合了卷积神经网络（CNN）和循环神经网络（RNN）的优点，能够高效地提取文本数据中的局部特征以及长时序特征。
     1. 构建模型
        1. 输入层：一般来说TextRCNN需要先预处理词向量，不过我们以及对词进行了混合编码处理，就不需要预处理这一步了，输入层接收来自多种编码方式拼接后的特征向量，也就是是经过词袋模型（BoW）、TF-IDF编码、词嵌入和One-hot编码后的特征向量，在输入层，数据被传入网络，准备经过下一步处理，要注意的是，TextRCNN为了更好的使用双向RNN，在输入一个词的时候需要把这个词左右两边的词也输入进去，比如文本是“x，x，x，任务，预算，最多，x，x, x”，那么在输入“预算”这个单词时，会输入“任务”+“预算”+“最多”，这三个词对于的特征向量拼接起来
        2. 卷积层：
     2. 训练配置
     3. 模型评估
  3. 最终部署