基于深度学习的文本表示与分类方法研究

聂　维　刘小豫
咸阳师范学院　计算机学院　陕西咸阳　712000

摘　要：随着信息时代的到来和信息化建设的广泛开展，文本信息也呈现出爆炸式增长的态势。如何从庞杂的信息中心获取到有效的信息，成文文本分析和分类关注的重点。基于此本文详细的阐述了文本表示和文本分类中的存在的问题以及相应的解决方式，以此来不断优化对文本信息的获取方式，提高获取效率。其中文本分类是以文本表示作为基石，在选择文本特征时要避免引入更多的认为因素，使得文本表示的有效性降低。近年来文本表示和文本分类的方式更是多种多样，在带来新的革新的同时，也出现了标签分布不均衡以及泛化能力较差等问题。本文就基于深度学习视域下，对文本表示和文本分类提出新的看法。

关键词：深度学习；文本表示；文本研究；文本分类；现状

1　基于深度学习的文本表示与分类方法现状

1.1 文本表示现状

文本是大量字符的集合，是由非结构化或者半结构化的数字信息组成的，因此并不能直接被分类器识别，这就需要我们在进行文本分类时现将文本表示转换为一种能够被计算机理解的语言。需要注重的是，所转换出来的语言，必须是简洁的，统一的，并且能够被不同的算法和分类器所识别的。目前文本内容既可以用图来表示，也可以用向量来表示。即通过识别所转换的语言来获取文本的内容以及类别信息。

而文本表示中存在的主要问题在于浅层文本表示和深层文本表示。其中在浅层文本表示中，主要存在语义缺失的问题，这直接导致分类器在识别过程中，效率较低并且识别精准度较低。而深层文本大多都是基于现行计算的模型，对这些文本的分类和提取大多依赖于人工特征的选取，通过在分类的过程中加入相应的阈值选择来实现文本的分类，但这样的操作方式破坏了文本的自我学习能力，也忽略的标签数据排布不均匀等问题。

1.2 文本分类现状

文本分类主要是指根据文本内容、文本主题、文本属性等特征将大量的文本归纳到一个或多个类别之中。文本分类在分类方法上可以分为基于规则的分类方法和基于统计的分类方法。其中基于规则的分类方法需要更多的专业知识和规则库来作为支撑，但这种分类方法的应用范围并不广泛，更适用于某一个专业学科或者某一个具体的领域。而基于统计的学习方式更多的是依据某种统计或者采用相关的定律，对样本进行统计和计算，并建立相应的数据模型，实现对文本的分类。同时需要在分类之前，根据样本的参数来对样本进行预测类别。

1.3 深度学习现状

深度学习并不是一种全新的学习方式，而是起源于人工神经网络，是基于深度神经网络一类学习方式的统称，即通过模拟人类大脑认知的肌理来解决相应的问题。

2　文本分类方式分析

信息时代的到来，既是机遇也是挑战，在面对庞杂和海量的资源时，为了能够有效的管理和利用好这些信息资源，就必须对这些信息资源进行分类，这也使得基于内容的信息检索方式成为备受关注的领域。其中文本分类和文本检索技术是进行信息分类和信息挖掘的重要方式。即通过设定预先的类别，再根据文本内容来判断该文本的所属累呗。文本分类和文本检索都需要在自然语言的基础上对文本内容进行处理、理解、信息组织和管理，从而使内容信息得到更加广泛的应用。

文本分类问题是指通过对已知类别的样本尽心学习，来预测未知样本类别的问题。而对于已知的文本分类方法主要是基于分类器的学习和分类结果。针对于分类方法的研讨也是为了能够更好的提高分类效率和分类准确度。尤其是在文本分类中，按照文本标签所属于的类别

基金项目：咸阳师范学院专项科研基金项目（基于深度学习的文本表示与分类方法研究，No.XSYK18010）。

作者简介：
聂维（1977-），女，陕西礼泉人，硕士，讲师。研究领域：网络安全、信息系统开发。
刘小豫（1978-），女，陕西兴平人，硕士，讲师。研究领域：图像处理、信息系统开发。

也可以将文本分为单标签类别和多标签类别。在单标签类别中，可以使用成熟的分类器来完成这项任务，但在多标签分类中，由于部分数据分布的不均匀，导致标签类别较多，使得简单的分类器难以满足多标签的分类需求，因此需要研究出更新的分类方法，使文本资源的检索和分类更加科学，合理。

2.1 经典文本分类方法

多分类器集成学习方法是近年来较为普遍接受和常用的分类方式。但随着信息技术水平的不断发展，对文本的分类方式也呈现出更加多样的化的发展趋势。随着信息资源日益庞杂，传统的经典文本分类方法已经不能满足文本表示和分类的需要。目前比较常见的分类方法还有朴素贝叶斯法、邻近算法、决策树以及集成学习封方法，这些方法都具有其独特的有点和适用范围，极大的提高了对文本的分类速度和分类科学性。

2.2 朴素贝叶斯

朴素贝叶斯是基于一个简单的假设所建立的一种贝叶斯方法。这一假设的主要内容是，假设样的不同特征属性与样本内容的分类影响是互不相干的。朴素贝叶斯的思想基础就是先通过概率估计来得出范围，再通过计算来得出概率。但对于已经给出的待分类项，就需要根据具体的数值来进行分类，即哪个数值较大，就归纳为哪一种类别。

2.3 KNN算法

KNN算法也被成为邻近算法，这一算法的核心是通过从训练集中找到待分类项以及和所需要分类项目相似度最高的若干个文本，再根据这一文本的类别来决定其余若干个分类项的类别，如果这些分类项的相似度较高，就判定为他们同属于同一类别，而该样本也同样属于这一类别。因此若干样本的平均取值是非常重要的，甚至可以说直接影响了分类结果，但同时通过计算若干个样本的取值能够表较好的解决样本内容不平衡的问题，通过求这些样本的临近点来与判断数值进行比较，最后再进行分类。

2.4 决策树

决策树是一种基于规则预测的计算方式，是通过对大量的文本进行计算后得出相应的数据，并对这些数据进行有目的的分类，在分类过程中找到一些具有价值的信息以供决策者做出最正确的决策。决策树的基本思想是：利用的树的结构来将所有的数据记录进行分类，尤其是树中的每一个内部节点都代表着某个条件下的一个记录集，并根据所记录的字段来建立不同取值的分支。这样设计的优点在于能够持续性的在每个分支下重复简历更下层的节点和分支，从而自顶向下构造一颗决策树。在决策树中，叶子的节点就是样本的类别，决策树的分类过程就是从树的根部开始，根据不同节点的特征来进行分类，并测试不同节点的样本属性，并将所得到的数值与分支节点相对性，不断移动找到能够满足所有条件的分支节点。

在文本分类中，决策树的优点是显而易见的，一是计算难度较低，便于日常的分类使用；二是能够比较迅速的处理好不相关的数据。但同时决策树的缺点也较为明显，就是往往容易忽略数据集中属性之间的相关性，从而产生一定的巧合，导致数据和分类的精确程度不高。

2.5 集成学习

集成学习也被称为多重学习或者分类器组合。集成学习主要是用过调用一些简单的分类算法来获得多个不同的分类器，再采用决策优化和覆盖优化的方式将若干个分类器进行组合。以此来达到利用这些分类器来改善整体模型之间的差异和泛化性能，提高分类系统的总体性能。

在集成学习分类中，个体的分类强度和个体之间的相关度越低，那么集成学习器的泛化能力就越强，反之亦然。因此集成学习简单来说就是基础分类器的生成和基础分类器的合并。

3 文本分类技术分析

3.1 标签文本分类技术

文本的分类技术在经过长时间的发展早已广泛应用到人们的日常生活之中。尤其是在传统的分类问题中，每一个文本只能够属于一个标签，根据样本的属性和内容就能够确定某一个最为相近的样本标签。但随着信息内容越来越复杂和丰富，标签的形式也越来越多样化，数据信息也越来越复杂，单一的标签已经无法准确描述文本内容和文本属性，因此常常需要建立相应的标签子集来表达文本的内容，这也就形成了多标签分类的问题。

3.1.1 标签相关性分类

多标签分类可以看成是对单一标签分类的拓展所得到的更加广泛和复杂的分类方式。同时多标签分类对于文本的解释也更加详尽和复杂，其中最主要的原因就是标签的输出空间较大。在多标签分类中，标签的集合越多，组合也就越大。但在分类过程中，数值过大的标签集合或者是标签组合在计算和分类效率以及质量上都具有一定的局限性。因此为了能够更好的解决这一问题，我们根据标签的相关性来区分标签的算法，即一阶方法、二阶方法和高阶方法。

一阶方法是指对每个标签都进行独立处理。也就是目前比较常见的标签分类方法，将分类任务当做多个二分类任务，虽然两个任务在计算过程中是独立的互相不影响的，能够在一定程度上提高分类的准确率和科学性，但由于没有考虑到标签之间的依赖关系，不能很好的解决数据分布不均衡的问题，导致对于部分标签在计算上仍然不能进行很好的分类。

二阶方法更注重两个标签之间的相关性，但也不能包含所有标签的相关情况。就某种意义而言，在本质上仍然是通过关注两个标签之间的相关性来获取相应的数值，对于其他相关性的标签仍然认为是相互独立的，虽然相较于一阶方法有了一定的改善和优化，但由于计算难度较高，很难计算大规模的数据信息，处理相应的学习问题。

3.2 平面分类法

平面分类法直接采用经典的机器学习算法就能够很好的完成文本分类的任务。但同样这些机器学习算法在遇到大规模的分类问题时，仍然会出现数据的倾斜和数据稀疏的问题，从而导致分类的能力降低。其中数据倾斜主要是因为文本类别的差距较大，如果将某一个类别作为正样本，其余的负样本数目远远超过正样本，就会产生数据倾斜的情况，而数据稀疏则是因为样本之间的长度不一，导致大量短文本向量的表示过于稀疏。

3.3 层次分类法

对于多标签，大规模的分类需求就需要通过层次分类法来实现。信息检索作为用户的基本需求，是每一个分类方法在计算之前首先要考虑到的因素。一般来说，信息检索是建立在海量信息文件的规整和分类的基础上，通过文本之间的相互关系来建立多层次多结构的分类体系，从而进一步提高用户的检索速度。相较于标签分类法和平面分类法而言，层次分类法的优点在于能够极大的提高文本分类的准确程度，并通过建立不同层次的系统来提高文本分类的准确程度。在这一过程中，可以将相关性较强的标签类别组成一个大类，再将大类进行区分，从而实现对不同层次的文本内容进行分类的目的，更好的实现分类的需求和分类的科学性，准确性。

4 结论

深度学习下的文本表示和文本分类都需要特定的计算方式和分类方法来实现，尤其是在层级结构中，更需要从不同的分类中提取出相应的特征来解决相应的分类问题。从而为文本的分类提供更加高效精准的文本模型和分类模型。因此在基于深度学习这一背景下，本文详细的分析了文本表示的现状以及文本分类的方法，以期能够更好的提高文本分类的效率和准确度，使文本的分类更加科学。

参考文献：

[1]王甜甜.基于深度强化学习的文本表示与分类研究[D].北京交通大学，2019.

[2]尹凯.基于深度学习的网络新闻文本分类研究[D].山西财经大学，2019.

[3]许奥狄.信息检索中基于深度学习的文本表示与分类方法研究[D].重庆邮电大学，2019.

[4]梁思程.基于深度学习的文本表示与分类研究[D].西安工程大学，2019.

[5]庞丹丹.基于深度学习的文本分类技术的研究[D].北方工业大学，2018.

[6]赖志龙.基于深度学习的多标签文本分类研究[D].电子科技大学，2021.

[7]吴佳君.面向文本分类任务的深度学习方法研究[D].南京信息工程大学，2021.

[8]李心雨.细粒度的新闻文本分类方法[D].哈尔滨工业大学，2020.

[9]秦文帅.基于网络模型融合的新闻长文本表示与分类方法研究[D].河南大学，2020.

[10]王恰.基于Attention Bi-LSTM的文本分类方法研究[D].华南理工大学，2018.

[11]李晓军.基于语义相似度的中文文本分类研究[D].西安电子科技大学，2017.

[12]刘文臻.中文文本多标签分类算法研究[D].电子科技大学，2020.