文本分类研究综述

汪 岿 刘柏嵩(宁波大学 信息科学与工程学院 浙江宁波 315211)

摘 要:在大数据时代,网络上的文本数据日益增长。采用文本分类技术对海量数据进行科学地组织和管理显得尤为重要。文本分类算法的研究起源于上个世纪50年代,一直受到科研人员的广泛关注。本文围绕文本分类的关键技术和基本流程进行重点阐述,主要包括文本预处理、词和文本的分布式表示、特征降维、分类算法等多个模块。其中详细分析了几种分类模型与分类方法,如深度学习、迁移学习、强化学习等等。此外,本文简单介绍了文本分类的评价指标与应用场景,并对当前面临的挑战及未来的发展趋势进行总结、预测。

关键词:文本分类;特征降维;机器学习
中图分类号:TP391.1 文献标识码:A

1 引言

随着科技的进步和互联网技术的发展,数字化资源已经渗透到当今社会的各个行业。然而这种爆发式的增长也给信息检索带来了困扰。文本作为分布最广、数据量最大的信息载体,如何对这些数据进行有效地组织和管理是亟待解决的难题。

文本分类是自然语言处理任务中的一项基础性工作,其目的是对文本资源进行整理和归类,同时其也是解决文本信息过载问题的关键环节。

早在上个世纪中叶,有关文本信息处理的研究开始走进人们的视野。词匹配法是最早被提出的分类算法,这种方法仅根据文档中是否出现了与类名相同或相近的词来判断文档是否属于某个类别。很显然,这种过于简单机械的方法无法带来良好的分类效果。

20世纪70年代,Salton等人[1]提出了向量空间模型。在后来一段时间内,知识工程成为文本分类的主要技术。然而这种技术高度依赖专业人员的帮助,需要为每个类别定义大量的推理规则和模板,造成了人力、物力的大量浪费。直至上个世纪90年代,基于统计和机器学习的文本分类方法逐渐兴起。通过机器从文档中挖掘出一些能够有效分类的规则,训练得到分类器,成为目前的主流方法。迄今为止,经过数十年的演变,文本分类已经初步形成了相对完整的理论体系。

文本分类按照任务类型的不同可划分为问题分类[2]、主题分类[3]以及情感分类[4]。常用于数字化图书馆、舆情分析、新闻推荐、邮件过滤等领域,为文本资源的查询、检索提供了有力支撑,是当前的主要研究热点之一。

本文以文本分类的相关工作为研究对象,全文组织结构如下:第2节简单描述文本分类的基本概念;第3节、第4节围绕文本分类的关键技术进行重点阐述;第5节、第6节简要介绍文本分类的评价指标,概括分析了文本分类的应用场景与挑战;第7节总结全文,并对未来的发展趋势做出预测。

2 文本分类概述

文本分类是指按照一定的分类体系或标准使用机器对文本集进行自动分类标记的过程。从宏观上看,整个分类流程可以近似地看作数学上做映射的过程。因此,我们可用映射关系诠释文本分类的概念。

文本分类的数学定义如下:

假设给定文档集合 $D=\{d_1,d_2,\ldots,d_m\}$,类别集

基金项目:国家社会科学基金项目/后期资助项目"学术型大数据知识组织与服务标准研究"(项目编号:15FTQ002)、省部级实验室/开放基金"数字图书馆知识组织与标引标准规范研究"(项目编号:B2014)。

合 $C=\{c_1,c_2,\ldots,c_n\}$。其中，d_i，c_j 分别表示集合中第 i 篇文档和第 j 个类别；m，n 为集合 D 的文档总数和集合 C 的类别数。我们可以发现文档集合和类别集合之间存在一定的映射关系 $f: D\times C\rightarrow R$，$R\in\{0,1\}$。当 $f(d_i,c_j)=1$ 时，表示文档 d_i 属于 c_j 类；反之，当 $f(d_i,c_j)=0$ 时，文档 d_i 不属于 c_j 类，f 为分类器。

文本分类从流程上可分为文本预处理、文本表示、特征提取、分类器训练等过程，其中最关键的步骤是特征提取和分类器训练。接下来，我们将对文本分类的关键技术进行详细分析。

3 文本分类的关键技术

3.1 文本预处理

在处理文本数据时，首先要对原始信息进行预处理。由于中文数据词语之间没有明确的分隔符且存在一定的噪音信息，所以在预处理阶段通常要经过分词、去除停用词、低频词过滤等过程。现有的分词算法可分为三大类：分别为基于字符串匹配（词典）的分词方法、基于理解的分词方法和基于统计的分词方法。近年来也有研究人员探索将深度学习技术应用于中文分词任务[5]。

受限于英文的语言学特征，在对英文文本进行预处理时，通常还包括词形还原、词干提取等步骤，数据预处理的质量直接影响后续的相关工作。

3.2 词向量与文本表示

文本是由词和短语构成的符号序列。要将自然语言处理问题转化成机器可学习的数学模型，首先要对词和文本进行向量化建模。

One－hot 表达方式是常用的词向量表示方法。假设 N 为整个词表空间，则每个词的词向量可表示为：$x=\{0,0,\cdots,1,0\cdots\}\in R^{(1\times|N|)}$，该词在词表中对应位置编号的维度为 1，其它维度全为 0。

词向量是语言模型（Language Model）的产物，为了弥补浅层表示学习的不足，Hinton、bengio 等人提出了分布式表达和词嵌入的概念。经典的神经网络语言模型包括 HLBL[6]、RNNLM[7] 等，其中最具有代表性的是：Mikolov 等人基于 CBOW 和 Skip－gram 模型提出的，结合哈夫曼编码的词向量训练方法 Word2vec[8]。通过该方法可训练得到低维、连续、实值、定长的词向量，进而可以高效、准确地计算词语之间的相似度。

传统的文本表示模型有布尔模型、向量空间模型 VSM（Vector Space Model）、概率模型以及图空间模型。然而，这些传统的文本表示方法缺乏语义表征能力。伴随着 Word2vec、Glove[9] 等分布式单词表示技术的兴起，深度文本表示模型得到了广泛的研究和应用。如 Joulin 等人基于浅层神经网络设计并开发出一款词向量训练和文本分类的工具 FastText[10]。此外，为了弥补模型在词向量处理阶段忽略单词之间排列顺序的缺陷，Doc2vec[11] 在 Word2vec 算法的基础上引入段落信息，增强模型表示文本语义的准确性和完整性。

3.3 特征驱动与文本分类

特征降维是文本信息处理的关键环节。传统的文本分类算法基于词袋模型和向量空间模型，特征空间具有高维性。然而这种高维、离散的特征给相关计算带来了不便——时空复杂度较高。同时，特征的冗余以及缺乏有效关联也会影响分类性能。

特征选择

特征降维的方法包括特征选择和特征抽取。特征选择是指从原始的特征空间中筛选部分重要特征组成新的特征集合，从而提高文本分类的准确率和效率，不改变原始空间的性质。常用的特征选择算法有文档频率、期望交叉熵、互信息等等。本文对比分析了几种常用方法，如下表所示。

表 1 特征选择算法对比分析

特征选择算法	主要原理	优点	缺点
词频-逆文本 频率	某词条在一篇文章中出现的频率越高，且文档集中包含该词条的文档数较少，则该词条的特征权重越大	原理简单，直观高效，具有普适性。适合在单篇文档中提取特征	没有考虑特征在类内、类间的分布情况
期望交叉熵	用来衡量某个特征对训练集整体的重要性。其值表示：出现某特定词的条件下类别的概率分布与类别本身概率分布的距离	不考虑特征项缺失的情况，降低稀有特征的干扰，提高分类效率	缺少对类间集中度、类内分散度的度量
互信息	一种信息度量方法，表示一个随机变量中包含的关于另一个随机变量的信息量	适用于局部信息（单一类别）和全局信息的特征选择	低频词的互信息较大，容易引起过学习；忽略了文本量对词条在每个类别中出现概率的影响
信息增益	用以度量两种变量的概率分布差异，具有非对称性。通过计算不同情况下的条件概率，选择信息增益较大的词条构成特征空间	综合考虑了特征项出现与缺失的情况	只适用于全局信息的特征选择，计算量大
卡方检验	通过观察实际值与理论值的偏差来确定理论的正确与否，是一个归一化的统计量	适用于局部和全局信息的特征选择，忽略词频的影响	计算开销大,过于注重一篇文档中某个特征的出现与否,对低频词的统计结果有所偏袒

特征抽取

特征抽取是特征降维的重要手段，其对原始的特征空间进行压缩、变换生成新的语义空间，能够较好地解决自然语言中的一义多词、一词多义问题，并降低原始空间的维度[12]。经过特征抽取后，新的特征子集能够更加简洁、准确地刻画文本的语义信息。常用的特征抽取方法有潜在语义索引(Latent Semantic Indexing)、主成分分析(Principle Component Analysis)和非负矩阵分解(Non-negative Matrix Factorization)。

除上述方法之外，也有其它特征降维技术用于文本分类任务。如优势比(Odds Ratio)、文本证据权(Weight of Evidence for Text)、基尼指数[13]、特征聚类等等。总的来说，以上算法各有优劣。在文本分类时，通常还要结合文本数据和分类器的特性，融合、改进各种特征降维方法。

分类算法

从宏观上划分，文本分类的方法包括三类：分别为无监督、半监督、有监督的文本分类。

无监督的文本分类无需带类别标记的训练数据。在实践中，通过文本聚类、种子词匹配[14]，潜在主题挖掘[15]等方法减少分类任务对标记数据的依赖。

半监督的文本分类算法只需少量带有标记的数据。通过学习少量标记数据和大量无标签数据的潜在特征，建立分类模型，并对新数据做出预测。

有监督的机器学习方法需要大量带有标签的训练数据。通常情况下，其分类准确率高于无监督和半监督方法。然而，标注数据的价值不断提高，有监督的文本分类方法高度依赖人为标记的结果，耗时费力。当前，广泛使用的传统分类模型有：朴素贝叶斯、K 最近邻、支持向量机、决策树等等。本文对各种模型的特点进行比较分析，如表 2 所示。

3.4 深度学习与文本分类

随着计算机软硬件技术的不断进步，成本不断下降。近年来，深度学习技术在文本、语音、图像等多媒体信息的处理任务中取得突破性进展。

深度学习作为机器学习的重要分支，相对于传统的文本分类模型，深度学习方法能够通过多层语义操作，获得更高层更抽象的语义表征，并将特征提取工作融合于模型的构建过程中，减少人为设计特征的不完备性与冗余。在文本分类中，常用的深度学习模型有卷积神经网络、循环神经网络等等。

表 2 传统分类算法比较分析

分类算法	主要原理	优点	缺点
朴素贝叶斯	基于特征条件独立假设与贝叶斯定理，通过先验和数据决定后验概率	原理简单，分类性能稳定。参数估计少，对缺失数据不敏感，适合增量式训练	由于假设的先验模型导致预测效果不佳。属性较多或属性之间关联性较强时，分类效果差
K 最近邻	依据特征空间中最邻近的一个或多个样本的类别来判断待分样本的类别	训练代价低，易处理类域交叉或重叠较多的样本集。适用于样本容量较大的文本集合	时空复杂度高，样本容量较小或数据集偏斜时容易误分，K 值的选择影响分类性能
支持向量机	通过学习，寻找间隔最大化的超平面对样本进行分割	高维稀疏、小样本数据集处理效果好，可解决非线性问题	训练速度慢，超平面的确定依赖少量实例，对数据缺失敏感。核函数的选择缺乏统一标准
决策树	在已知各种情况发生概率的基础上，将样本所有特征的判断级联起来，通过一系列规则对数据进行分类	易于理解和解释，适用于数值型和标称型数据，计算复杂度不高。能够根据分类规则推出相应的逻辑表达式并通过静态测试对模型进行测评	处理连续型、时序型、缺失数据较为困难，存在过拟合以及忽略属性之间关联性的问题
Rocchio 算法	使用训练语料为每个类构造一个原型向量(质心向量)，通过计算待分类文档(向量化表示)与原型向量的相似度，划分类别	算法简单极易实现，训练和分类效率高，易被理解	受样本分布影响，原型向量可能落于所属的类域外

卷积神经网络

卷积神经网络(CNN)由输入层、卷积层、池化层、全连接层以及 softmax 层构成，如图 1 所示。其中，输入层为向量化的文本矩阵。假设某文本包含 n 个单词，词向量维数为 k，则输入可表示为 $n \times k$ 维的文本矩阵，记作 X。

$$X = \{x_1, x_2, \ldots, x_n\}^T \quad (1)$$

其中，x_i 为第 i 个词的词向量。

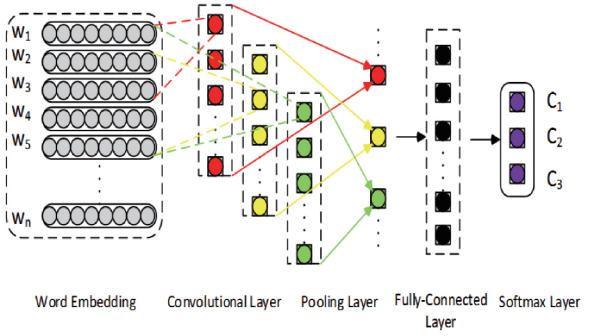

图 1 卷积神经网络文本分类模型

卷积层和池化层是 CNN 的核心组件，经过卷积和池化操作，将原始特征映射到更高层次维度的语义空

间。通过设置不同长度的卷积核，CNN 可以进行丰富的局部特征提取。假设卷积核的尺寸为 h，权值矩阵和偏置分别为 $W \in R^{h\times k}$ 和 $b \in R$。将输入 X 分为 $\{x_{1:h}, x_{2:h+1}, \ldots, x_{i:i+h-1}, \ldots, x_{n-h+1:n}\}$，每次卷积操作对第 i 步时滑动窗口内的信息 $x_{i:(i+h-1)} \in R^{h\times k}$ 进行特征提取，得到的属性值 $c_i \in R$，$i \in [1, n-h+1]$，计算如下：

$$c_i = f(W \otimes x_{i:(i+h-1)} + b) \tag{2}$$

其中，f 是非线性的激活函数，$\otimes$ 为卷积操作符。则卷积特征图可描述为：

$$C = \{c_1, c_2, \ldots, c_{n-h+1}\} \tag{3}$$

池化层负责对卷积层提取到的信息进行采样，并保留其最重要的部分。同时为下一层的计算减少参数，加快模型的训练速度。通过对卷积网络提取到的特征映射向量取最大值(max pooling)或平均值(mean pooling)的方法，使得 CNN 可以接受变长的文本输入。该过程描述如下：

$$\hat{c} = \text{mean}(C) \tag{4}$$

$$\hat{c} = \max(C) \tag{5}$$

其中，$\max(\cdot)$，$\text{mean}(\cdot)$ 分别代表最大池化和平均池化操作。对于 m 个卷积核，生成的池化特征图可表示为：

$$\hat{C} = (\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_m) \tag{6}$$

此外，经研究发现，池化层在减少过拟合方面也发挥着不小的作用。

全连接层的功能是将样本从特征空间映射到标记空间。主要参数为权值矩阵 W_f 和偏置 b_f。将池化层得到的特征信息输入到全连接层，然后通过 softmax 层输出归一化的分类概率。

$$y = softmax(W_f \cdot \hat{C} + b_f) \tag{7}$$

假设 d 维向量 V，V_j 表示 V 中的第 j 个元素，则其 softmax 值计算如下：

$$S_j = \frac{e^{V_j}}{\sum_1^d e^{V_j}} \tag{8}$$

通过参数共享机制，CNN 能够较好地处理高维数据，并且在建模过程中，无需人为选择特征。除了上述简单的五层结构，对于具有复杂特征的数据集，通过卷积层－池化层反复堆叠，构建深层的特征提取网络[16]，CNN 能够挖掘更为丰富的文本语义。

输入层(文本向量化层)的词向量可通过随机初始化或预训练得到。在模型的训练过程中其值始终保持不变(Static)或作为参数动态优化(Not Static)。此外，在该层使用字符级[17]，字符和词级别的双输入[18]，词、位置、词性等多通道组合特征[19]能够在一定程度上减少模型对词向量和语法、句法结构信息的依赖，提高泛化能力。对于新闻、博客类长文本，其内容中包含大量与主题无关的信息，部分研究人员提出将注意力机制[20]、句法依存关系[21]、主题模型[5]与卷积神经网络结合，从而提取更深层次的语义特征，增强模型的可解释性。

循环神经网络

循环神经网络(RNN)是一种常用的文本信息处理深度学习模型。RNN 擅长于处理序列数据，并且具备变长输入和发掘长期依赖的能力。RNN 通过循环机制将上一时刻的隐状态传递至下一时刻，并计算得到新的状态信息，如图 2 所示。其隐藏层的计算过程描述如下：

$$h^{(t)} = \sigma(Ux^{(t)} + Wh^{(t-1)} + b) \tag{9}$$

其中，$h^{(t)}$ 为 RNN 在隐藏层 t 时刻的状态，$U \in R^{|h^{(t)}|\times|x^{(t)}|}$，$W \in R^{|h^{(t)}|\times|h^{(t)}|}$，$b \in R^{|h^{(t)}|}$ 为模型的参数。σ 是非线性的激活函数。

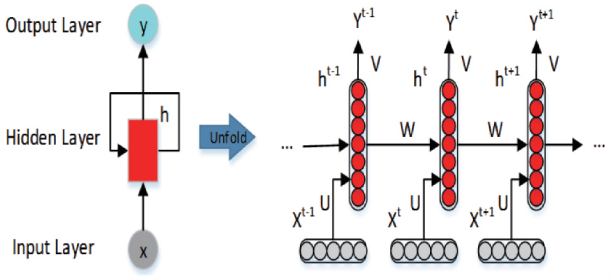

图 2 循环神经网络

当前，循环神经网络已经广泛应用于文本分类[22]、语音识别[23]、机器翻译[24]等诸多领域。对于文本分类任务，每一层的输入对应文本中的词向量表达 $x^{(i)}$。t 时刻的隐状态由当前层的输入 $x^{(t)}$ 和前一时刻的隐状态 $h^{(t-1)}$ 共同决定，从而实现信息的持久化。在实际运用中，考虑到上下文信息对文本语义的影响，也有学者采用双向 RNN(BiRNN)模型对文本序列进行建模：

$$\overrightarrow{h^{(t)}} = \sigma(\overrightarrow{U}x^{(t)} + \overrightarrow{W}\overrightarrow{h^{(t-1)}} + \overrightarrow{b}) \tag{10}$$

$$\overleftarrow{h^{(t)}} = \sigma(\overleftarrow{U}x^{(t)} + \overleftarrow{W}\overleftarrow{h^{(t+1)}} + \overleftarrow{b}) \tag{11}$$

其中 $\rightarrow$，$\leftarrow$ 分别代表前向和后向 RNN。此时，当前词的语义表示取决于其上下文向量。双向 RNN 可以捕捉序列中的历史信息与未来信息，学习更加丰富的句子表达。然后，将最后一层的输出或对各层输出取平均值的结果输入到全连接层和 softmax 层。此外，为了防止过拟合，通常会在全连接层引入 Dropout 策略。

对于深度神经网络，最常采用的训练方法为梯度下降法。然而，经研究发现，随着网络层数的不断加深，通过反向传播训练参数时往往面临着梯度不稳定的问题（梯度消失或梯度爆炸）。这些现象的发生不仅导致模型的收敛速度变慢，也使得网路难以发挥深度结构的优势[25]，从而影响文本分类的性能。因此，研究人员提出了一系列改进方法：

1）改变训练方法，如采用无监督策略预训练参数，再进行微调。

2）优化节点计算，使用激活函数（ReLU），批归一化（batch normalization）、梯度截断等技巧。

3）调整网络结构，如跳跃连接（跨多个时间步的长距离连接），门控单元（LSTM，GRU）等。

LSTM（Long Short－Term Memory）和 GRU（Gated Recurrent Unit）是常用的循环神经网络门控算法。LSTM 的核心部件为细胞（cell）单元，它采用门控存储结构替换 RNN 的隐状态来解决梯度不稳定和长距离依赖问题。在每一层，其通过遗忘门（forget gate）、输入门（input gate）以及输出门（output gate）控制数据的迭代与更新，实现信息传递。

GRU 是 LSTM 网络的常用变体。相较于长短期记忆网络，GRU 的网络结构更加简单，并且保持了 LSTM 的效果。其控制信息传输的结构由更新门（update gate）和重置门（reset gate）组成，同时合并了系统状态 $h^{(t)}$ 和单元状态 $C^{(t)}$。

当前，循环神经网络及其变体在文本分类中的应用较多。例如，Shi 等人[26]提出使用深度长短期记忆网络学习句子的特征表示，大幅提升了查询分类的准确率。Yogatama 等[27]针对文本分类问题构建了基于 LSTM 网络的生成模型和判别模型。Wang[28]提出了中断循环神经网络 DRNN（Disconnected Recurrent Neural Networks），通过限制 RNN 的信息流动，并将位置不变性引入 GRU 模型。这使得 DGRU（Disconnected Gated Recurrent Unit）模型既能捕捉长距离依赖关系，又可以很好地抽取关键短语信息，在多个文本分类数据集上取得了当前最好的结果。Dai[29]将 GRU 网络和迁移学习技术相结合，处理产品评论中的情感分类问题。在时间性能方面，SRU[30]（Simple Recurrent Unit）在分类和问答数据集上的训练和收敛速度相比传统的 LSTM 网络提高了 5－9 倍。SRNN[31]（Sliced Recurrent Neural Networks）将初始输入序列切分成多个子序列来实现并行化，从而获得多个层级的高级信息，且不需要额外增加参数。另外，也有研究人员探索将 RNN 和 CNN 的体系结构结合，如 RCNN[32]等，应用循环结构在学习单词表示时尽可能地捕获上下文信息，并采用最大池化方法提取在分类过程中起重要作用的文本特征。

注意力机制

受人类注意力的启发，注意力机制（Attention Mechanism）最早被用于计算机视觉[33]和机器翻译领域[24]。视觉注意力是人类视觉特有的信号处理机制。人类视觉通过快速扫描全局图像，聚焦关键区域，然后对该区域投入更多注意力资源，以获取更多的细节信息。在机器翻译领域，注意力机制主要着力于解决从源语言到目标语言的翻译对齐问题。在文本摘要[34]任务中，注意力机制的作用是筛选文本中的概括性语句，抑制次要和无关信息。

因此，从本质上说，注意力机制是一种资源分配机制。其主要目的是利用有限的注意力资源从大量信息中快速筛选出高价值的信息，对文本分类工作同样适用。如 Wan 等人[35]使用 LSTM 网络和注意力机制解决跨语言情感分类任务。Zhou 等人[36]使用基于注意力的 BLSTM 来构建对话表示，提升中文会话主题分类任务的整体性能。Yang 等人[37]提出层级注意力模型 HAN（Hierarchical Attention Networks）。该模型通过单词级别和句子级别的注意力感知机制，分别计算单词对句子的重要性以及每个句子对分类结果的贡献度，提高文本分类的准确率，增强模型的可解释性。

卷积神经网络和注意力机制的结合也被广泛使用。例如，文献[20]提出结合注意力机制的句子过滤方法及分类模型。文献[38]提出基于多注意力的卷积神经网络情感分析模型，结合词向量、词性、位置三个维度的注意力输入矩阵，有效弥补了传统模型仅依赖内容层面注意力的不足。

自注意力（Self－Attention）机制是一种特殊的注意力方法，序列中的每一个单元都需要和该序列中的

所有单元进行注意力计算。自注意力的特点在于其直接计算词语之间的关联,而无视它们的距离,能够学习句子的内部结构,捕获长距离依赖关系。此外,其实现过程较为简单,且可以并行计算。多头自注意力能够提升句子在复杂语义上的表示结果,从不同的表示子空间里学习句子语义,弥补单注意力的不足。同时,合理使用惩罚项,减少多注意力在句子学习多样性不足时产生的信息冗余[39]。

4 文本分类的其它方法

4.1 其它深度学习模型

除 RNN 和 CNN 之外,也有其它深度学习模型应用于文本分类任务,如深度信念网络(Deep Belief Network)和堆叠自动编码器(Stacked Auto - Encoder)。DBN 由多层受限玻尔兹曼机(Restricted Boltzmann Machine)堆叠而成,是一种概率生成模型。经典的 DBN 结构包括若干层 RBM 网络和一层 BP 网络。RBM 基于能量模型,由可见层(输入层)和隐层构成,可见层和隐层之间全连接,层内无连接。通过最大化模型在训练数据上的对数似然函数来拟合观测数据,调整参数,实现输入数据的重构。RBM 是一种无监督模型,在深度信念网络的最顶层通常连接文本分类层,将样本从特征空间映射到标记空间[40]。整个过程可以描述为:1)自底向上逐层训练多个 RBM,得到预训练的权值参数;2)使用标记数据在反向传播时对整个网络进行微调。这种参数预训练方法相比其它神经网络中的参数随机初始化更具优势,能够快速生成一个深层网络。

有关研究还包括堆叠自动编码器,和 DBN 不同的是它采用自动编码器替换 DBN 中的 RBM 结构[41]。最后和 DBN 类似,也可以使用有监督方法微调文本分类层和整个网络。堆叠自动编码器每一层学习到的语义编码可以看作输入数据的不同表达或可以充分代表原始输入的句子特征。自动编码器是一种判别模型,这使得网络难以捕获句子的内部结构。常用的改进模型有降噪自动编码器和稀疏自动编码器。在文本分类任务中,降噪自动编码器能够减少噪声的干扰,学习更加鲁棒的文本表达,提高泛化能力[42]。

4.2 集成学习与文本分类

集成学习(Ensemble Learning)的概念非常广泛,其主旨思想是利用多个同质或异质的学习器处理同一问题,并将它们的学习结果融合,从而获得比单一学习器更好的实践效果。

针对文本分类问题,融合多个分类器的分类结果能在一定程度上提升文本分类的准确率[43]。多个分类器可以是同质的,例如都是决策树,也可以是异质的,比如包含朴素贝叶斯、支持向量机等多个不同类型的分类模型。通过多个弱分类器(基分类器)组合生成一个强分类器,提升分类效果。因此,集成学习的重点在于两个方面:1)如何生成多个基分类器;2)基分类器的融合策略。常用的获得基分类器的方法包括 Boosting 和 Bagging。Boosting 方法生成基分类器是串行的,分类器之间存在依赖关系。而 Bagging 方法可以通过独立并行的方式生成多个基分类器。保持基分类器之间的差异性和多样性能够提升集成学习的效果,增强模型的鲁棒性[44]。

此外,基分类器的融合策略也十分重要。当前,融合方法主要包括三大类,分别是平均法、投票法以及学习法。平均法是对多个弱分类器的学习结果进行算术平均或加权平均,从而得到最终的输出结果,常用于数值型数据的回归预测任务。投票法,顾名思义,对多个弱分类器的分类结果进行投票,少数服从多数。也可以类似平均法,对弱分类器的投票结果进行加权统计,即加权投票法。不同于以上方法,学习法的规则更加复杂。其主要做法是在初始的多个基分类器上,再加上一层学习器,以基分类器的学习结果为输入,训练集的输出为输出,重新训练一个次级分类器,拟合样本数据。

4.3 迁移学习与文本分类

迁移学习(Transfer Learning)的主要思想:将源领域学习到的知识迁移到目标领域,用以辅助完成新任务的过程。对于人类来说,就是掌握举一反三的学习能力。按照情景不同,迁移学习可以分为三类,分别是

归纳式迁移学习、直推式迁移学习和无监督迁移学习[45]。

在文本分类任务中应用最广泛的是直推式迁移学习。对于归纳迁移学习,近年来,随着深度学习的发展,使用预训练的词向量再进行调优是大多数模型的应用方式,如 Word2vec、Glove、FastText 等。然而这些简单的迁移技术仅用在模型的输入层。ELMo[46]、CoVe[47] 等基于不同层输出和上下文特征的拼接嵌入方法,可以更好地捕捉语法和语义层面的信息,但仍然需要从头训练任务模型。语言模型、Transformer 架构

的提出及相关方法(如 GPT[48]、BERT[49]、ULMFiT[50])的使用,使得迁移学习在包括文本分类在内的多个自然语言处理任务上取得了重要进展。

在大规模通用语料库上训练语言模型,然后针对具体的数据集和分类任务进行微调(Fine - tuning)是一种有效的解决办法。通过一次训练,多次使用可以降低模型的训练难度,节省计算资源。此外,将从其它语料中学习到的通用知识进行迁移,可以快速提高模型在小数据集或者只有少量标记数据上的分类效果。

4.4 强化学习与文本分类

强化学习(Reinforcement Learning)是智能体(Agent)以试错的方式进行学习,通过与环境之间的交互,获得奖赏、指导行为[51]。其目标是学习从环境状态到动作的映射,使得智能体选择的动作能够获得最大的累积回报。"试错学习"和"延迟回报"是强化学习的两个主要特征[52]。

Agent 根据当前状态选择一个动作作用于环境,环境接受该动作后更新状态,同时产生一个奖惩信号反馈给 Agent。Agent 再依据反馈的信号和新的环境状态选择下一步动作[53]。如果 Agent 的某个行为策略导致环境正的奖赏,那么 Agent 以后产生这个行为的趋势便会加强。如此循环往复,智能体与环境之间不断交互,学习从状态到动作的映射策略,从而达到优化系统性能的目的[54]。

强化学习为文本分类等任务提供了新的解决思路和训练策略。将文本分类问题建模成顺序、离散的决策过程,通过强化学习方法优化模型,训练参数。符号化表征模型的决策过程具有很好的可解释性,同时分类效果也得到了提升[55-56]。

5 评价指标

文本分类的性能评价指标主要有精确率、召回率和 F - Measure。

针对分类体系中的任意类别 C 构建混淆矩阵,如表 3 所示。

表3 混淆矩阵

预测值 / 真实值	正例	负例
正例	TP(True Positive)	FN(False Negative)
负例	FP(False Positive)	TN(True Negative)

在上表中,TP、FN、FP、TN 分别表示:属于类 C 的样本被正确分类到类 C、属于类 C 的样本被错误分类到其它类、其它类的样本被错误分类到类 C 以及其它类的样本被正确分类到其它类。

因此,各评价指标可计算如下:

(1) 精确率(Precision),也叫查准率,表示正确分类的正例个数占分类为正例的实例个数的比例。

$$P = \frac{TP}{TP + FP} \tag{12}$$

(2) 召回率(Recall),也叫查全率,表示正确分类的正例个数占实际正例个数的比例。

$$R = \frac{TP}{TP + FN} \tag{13}$$

(3) 查准率和查全率的综合评价指标 F - Measure。

$$F_\beta = \frac{(1 + \beta^2) \times P \times R}{\beta^2 \times P + R} \tag{14}$$

当 $\beta = 1$ 时:

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{15}$$

6 应用场景及其面临的挑战

6.1 文本分类的应用

文本分类的应用非常广泛。如在医疗领域,智能分诊技术的使用能够节约大量医疗资源,提升服务质量和效率。在一些企业,依靠智能客服代替人工提供全天候的客户服务,可以有效降低运营成本,改善用户体验。问题分类在问答系统(Question Answering System)中起着重要作用,提高问题分类的准确率有助于构建更加鲁棒的 QA 系统[2]。

在图书情报领域,专利[57]、图书[58]、期刊论文[59]、学术新闻[60]等跨类型学术资源的自动组织与分类是数字化图书馆的关键技术,有利于工业企业、科研院所的研究人员更快地掌握各类前沿动态。

随着移动互联网的发展,人们获取信息的方式发生了变化,由单纯的信息检索转变为"搜索 + 推荐"的双引擎模式。但无论是搜索还是推荐,其背后都离不开机器对内容的理解能力。文本作为网络上分布最广、数据量最大的信息载体,准确的分类标签为资源检索和新闻资讯的个性化推荐提供了有力支撑,使得推荐的信息能够尽可能地满足千人千面的用户

需求[21,59]。

情感分类(情感极性分析)是文本分类的重要分支。如在社交媒体中,对用户评论的情感倾向进行分析[8](积极、消极等)。情感极性分析能帮助企业理解用户消费习惯、分析热点话题和危机舆情监控,为企业提供有力的决策支持。此外,情感分析技术还可以用在商品和服务领域。例如对产品[29]、电影[60]、图书[61]评论的情感分类。

智能手机的普及促进了在线即时消息和短信使用的增长。将文本分类技术应用于邮件检测和短信过滤任务[62-63],可以帮助人们快速筛选有用信息。

6.2 当前面临的挑战

(1)数据标注瓶颈。数据和算法是推动人工智能向前发展的主要动力。高质量的标记数据有助于提升文本分类的准确率。然而,网络上存在大量杂乱无章的无标签数据,依赖人工标注的成本高,效率低。无监督数据的特征学习和半监督学习自动标注过程中的噪音剔除是当前的研究热点和难点。

(2)深度学习的可解释性。深度学习模型在特征提取,语义挖掘方面有着独特的优势,在文本分类任务中取得了不俗的成绩。然而,深度学习是一个黑盒模型,其训练过程难以复现,隐语义和输出结果的可解释性较差。例如,结合迁移学习理论的文本分类方法,初始预训练的语言模型学习到哪些知识,在参数迁移、特征迁移、针对目标域的训练数据和分类任务进行微调时,保留了哪些特征,我们很难了解。这使得模型的改进与优化失去了明确的指引,也大大加深了研究人员调参的难度。

(3)跨语种或多语种的文本分类。在经济全球化的大背景下,跨语言的文本分类在跨国组织和企业中的应用越来越多。将在源语言中训练的分类模型应用于另一种语言(目标语言)的分类任务,其挑战性在于源语言数据的特征空间与目标语言数据之间缺乏重叠[31]。各国的语言、文字包含不同的语言学特征,这无疑加大了跨语言文本分类的难度。当前,基于机器翻译技术的跨语言文本分类方法过于依赖双语词典和平行语料,在一些小语种上的表现较差[64]。通过跨语言文本表示技术和迁移学习方法训练得到独立于语言的分类模型是未来的重点研究方向[35,65]。

7 总结与展望

近年来,移动互联网和大数据快速发展,网络上的文本数据日益增长。使用计算机对海量数据进行自动化分类是当前的研究热点。本文围绕文本分类的处理流程进行重点阐述,对比分析了词和文本表示、特征工程、分类算法等多个模块的关键技术。我们相信随着计算机技术的不断发展,文本分类研究仍有广阔的前景。主要趋势预测如下:

(1)对传统方法进行优化。如常用机器学习模型的改进;传统的机器学习算法、特征提取方法与深度学习模型的融合。

(2)新理论、新方法的提出。如将图卷积神经网络(Graph Convolutional Networks)应用于文本分类任务[66]。

(3)引入知识库、知识图谱等结构化的外部知识,优化文本表示和预训练的语言模型,进而提升文本分类的性能。

(4)在自然语言处理领域,很多任务具有较强的内部关联性,采用多任务联合学习或对抗学习的效果更好。例如,将关键词抽取、文本分类、文本摘要等多个任务联合训练,寻找最优的参数组合。同时,网络上存在大量的多媒体信息,文本分类、语音识别、图像处理与计算机视觉等跨领域的多任务联合学习也是未来的发展趋势。

(5)今日头条等资讯平台兴起,面对大规模文本数据,采用在线增量学习和离线学习相结合的办法,在分布式平台上处理不断增长的信息洪流。

参考文献

[1] Salton G. A vector space model for automatic indexing [J]. Communications of the Acm, 1974, 18(11):613-620

[2] Mohasseb A, Baderelden M, Cocea M, *et al*. Question categorization and classification using grammar based approach [J]. Information Processing and Management, 2018, 54(6): 1228-1243

[3] 杜雨萌,张伟男,刘挺. 基于主题增强卷积神经网络的用户兴趣识别[J]. 计算机研究与发展, 2018(1):188-197

[4] Wijayanti R, Arisal A. Ensemble approach for sentiment polarity analysis in user-generated Indonesian text [C]//

2017 International Conference on Computer, Control, Informatics and its Applications (IC3INA). IEEE, 2018

[5] Mengge Wang, Xiaoge Li, Zheng Wei, Shuting Zhi, and Haoyue Wang. 2018. Chinese Word Segmentation Based on Deep Learning. [C] // In Proceedings of the 2018 10th International Conference on Machine Learning and Computing (ICMLC 2018). ACM, New York, NY, USA, 16 - 20

[6] Mnih A, Hinton G. A Scalable Hierarchical Distributed Language Model. [C] // International Conference on Neural Information Processing Systems. Curran Associates Inc. 2008

[7] Mikolov T, Karafiat M, Burget L, *et al.* Recurrent eural network based language model [C] // Conference of the International Speech Communication Association, 2010: 1045 - 1048

[8] Mikolov T, Chen K, Corrado G, *et al.* Efficient Estimation of Word Representations in Vector Space [J]. Computer Science, 2013

[9] Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation [C] // Conference on Empirical Methods in Natural Language Processing. 2014: 1532 - 1543

[10] Joulin A, Grave E, Bojanowski P, *et al.* Bag of Tricks for Efficient Text Classification [J]. 2016: 427 - 431

[11] Le Q V, Mikolov T. Distributed Representations of Sentences and Documents [J]. 2014, 4: II - 1188

[12] 陈涛，谢阳群. 文本分类中的特征降维方法综述 [J]. 情报学报，2005，24(6)：11 - 11

[13] 尚文倩，黄厚宽，刘玉玲，*et al.* 文本分类中基于基尼指数的特征选择算法研究 [J]. 计算机研究与发展，2006，43(10)：1688 - 1694

[14] Zagibalov T, Carroll J. Automatic seed word selection for unsupervised sentiment classification of Chinese text [C] // International Conference. 2008

[15] Li C, Xing J, Sun A, *et al.* Effective Document Labeling with Very Few Seed Words: A Topic Model Approach [C] // the 25th ACM International. ACM, 2016

[16] Conneau A, Schwenk H, Barrault, Loïc, *et al.* Very Deep Convolutional Networks for Text Classification [J]. 2016

[17] Adams B, Mckenzie G. Crowdsourcing the character of a place: Character - level convolutional networks for multilingual geographic text classification [J]. Transactions in Gis, 2018(1)

[18] Yu B, Zhang L, Management S O. Chinese short text classification based on CP - CNN [J]. Application Research of Computers, 2018

[19] 陈珂，梁斌，柯文德，许波，曾国超. 基于多通道卷积神经网络的中文微博情感分析 [J]. 计算机研究与发展，2018，55(05)：945 - 957

[20] 卢玲，杨武，王远伦，等. 结合注意力机制的长文本分类方法 [J]. 计算机应用，2018，38(5)：1272 - 1277

[21] 夏从零，钱涛，姬东鸿. 基于事件卷积特征的新闻文本分类 [J]. 计算机应用研究，2017(4)：991 - 994

[22] Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi - task learning [C] // Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2016: 2873 - 2879

[23] Sak, Haşim, Senior A, Rao K, *et al.* Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition [J]. Computer Science, 2015

[24] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate [J]. Computer Science, 2014

[25] 陈建廷，向阳. 深度神经网络训练中梯度不稳定现象研究综述 [J]. 软件学报，2018，v.29(07)：249 - 269

[26] Shi Y, Yao K, Tian L, *et al.* Deep LSTM based Feature Mapping for Query Classification [C] // Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 1501 - 1511

[27] Yogatama D, Dyer C, Wang L, *et al.* Generative and Discriminative Text Classification with Recurrent Neural Networks [J]. 2017

[28] Wang B. Disconnected Recurrent Neural Networks for Text Categorization [C] // Meeting of the Association for Computational Linguistics. 2018: 2311 - 2320

[29] Dai M, Huang S, Zhong J, *et al.* Influence of Noise on Transfer Learning in Chinese Sentiment Classification using GRU [C]. International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery. 2017

[30] Lei T, Zhang Y, Wang S I, *et al.* Simple Recurrent Units for Highly Parallelizable Recurrence. [J]. empirical methods in natural language processing, 2018: 4470 - 4481

[31] Yu Z, Liu G. Sliced Recurrent Neural Networks [J]. international conference on computational linguistics, 2018: 2953 - 2964

[32] Lin R, Fu C, Mao C, *et al.* Academic News Text Classification Model Based on Attention Mechanism and RCNN [C] // CCF Conference on Computer Supported Cooperative Work, and Social Computing. Springer, Singapore, 2018

[33] Mnih V, Heess N, Graves A, *et al.* Recurrent Models of Visual Attention [J]. Advances in neural information processing systems, 2014

[34] Rush A M, Chopra S, Weston J, *et al.* A Neural Attention Model for Abstractive Sentence Summarization [J]. empirical

methods in natural language processing, 2015: 379 – 389

[35] Zhou X, Wan X, Xiao J, *et al.* Attention – based LSTM Network for Cross – Lingual Sentiment Classification [C]. empirical methods in natural language processing, 2016: 247 – 256

[36] Zhou Y, Li C, Xu B, *et al.* Hierarchical Hybrid Attention Networks for Chinese Conversation Topic Classification [C]. international conference on neural information processing, 2017: 540 – 550

[37] Yang Z, Yang D, Dyer C, *et al.* Hierarchical Attention Networks for Document Classification [C]. north american chapter of the association for computational linguistics, 2016: 1480 – 1489

[38] 基于多注意力卷积神经网络的特定目标情感分析 [J]. 计算机研究与发展, 2017(8)

[39] Vaswani A, Shazeer N, Parmar N, *et al.* Attention is All you Need [J]. neural information processing systems, 2017: 5998 – 6008

[40] 张庆庆,刘西林. 基于深度信念网络的文本情感分类研究 [J]. 西北工业大学学报(社会科学版),2016,36(01):62 – 66

[41] Cross – lingual sentiment classification with stacked autoencoders [J]. Knowledge and Information Systems, 2016, 47(1):27 – 44

[42] 刘红光,马双刚,刘桂锋. 基于降噪自动编码器的中文新闻文本分类方法研究 [J]. 现代图书情报技术,2016(06):12 – 19

[43] 高元,刘柏嵩. 基于集成学习的标题分类算法研究 [J]. 计算机应用研究,2017,34(04):1004 – 1007

[44] 徐禹洪,黄沛杰. 基于优化样本分布抽样集成学习的半监督文本分类方法研究 [J]. 中文信息学报,2017,31(06):180 – 189

[45] 夏禹. 迁移学习在文本分类中的应用研究 [D]. 哈尔滨工程大学,2014

[46] Peters M E, Neumann M, Iyyer M, *et al.* Deep Contextualized Word Representations [J]. north american chapter of the association for computational linguistics, 2018: 2227 – 2237

[47] Mccann B, Bradbury J, Xiong C, *et al.* Learned in Translation: Contextualized Word Vectors. [J]. neural information processing systems, 2017: 6294 – 6305

[48] Radford A, Narasimhan K, Salimans T, *et al.* Improving language understanding by generative pre – training [J]. 2018

[49] Devlin J, Chang M, Lee K, *et al.* BERT: Pre – training of Deep Bidirectional Transformers for Language Understanding [J]. arXiv: Computation and Language, 2018

[50] Howard J, Ruder S. Universal Language Model Fine – tuning for Text Classification [J]. meeting of the association for computational linguistics, 2018: 328 – 339

[51] 万里鹏,兰旭光,张翰博,郑南宁. 深度强化学习理论及其应用综述 [J]. 模式识别与人工智能,2019,32(01):67 – 81

[52] 张素芳,翟俊海,王聪,沈矗,赵春玲. 大数据与大数据机器学习 [J]. 河北大学学报(自然科学版),2018,38(03):299 – 308 + 336

[53] 刘建伟,高峰,罗雄麟. 基于值函数和策略梯度的深度强化学习综述 [J/OL]. 计算机学报,2018: 1 – 38 [2019 – 04 – 03]. http://kns.cnki.net/kcms/detail/11.1826.TP.20181022.1231.002.html

[54] Yang P, Ma S, Zhang Y, *et al.* A Deep Reinforced Sequence – to – Set Model for Multi – Label Text Classification [J]. arXiv: Computation and Language, 2018

[55] Zhang T, Huang M, Zhao L, *et al.* Learning Structured Representation for Text Classification via Reinforcement Learning [C]. national conference on artificial intelligence, 2018: 6053 – 6060

[56] Liu X, Mou L, Cui H, *et al.* JUMPER: Learning When to Make Classification Decisions in Reading [C]. international joint conference on artificial intelligence, 2018: 4237 – 4243

[57] Shamsi F A, Aung Z. Automatic patent classification by a three – phase model with document frequency matrix and boosted tree [C]// 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA). IEEE, 2016

[58] 马芳,黄翠玉. 中文科技期刊论文多标签分类研究 [J]. 图书情报导刊,2019,4(02):26 – 32

[59] 李静,杨小帆,孙启干. 面向 Web 信息检索的虚核文本分类算法 [J]. 计算机工程,2012,38(10):182 – 184 + 187

[60] Ye Q, Shi W, Li Y J. Sentiment Classification for Movie Reviews in Chinese by Improved Semantic Oriented Approach [C]// Hawaii International Conference on System Sciences. 2006

[61] Ye Q, Li Y, Zhang Y. Semantic – Oriented sentiment classification for Chinese product reviews: An experimental study of book and cell phone reviews [J]. 清华大学学报:自然科学英文版, 2005, 10(s1):797 – 802

[62] Almeida T A, Silva T P, Santos I, *et al.* Text Normalization and Semantic Indexing to EnhanceInstant Messaging and SMS Spam Filtering [J]. Knowledge – Based Systems, 2016: S0950705116300909

[63] Ezpeleta E, Garitano I, Zurutuza U, *et al.* Short Messages Spam Filtering Combining Personality Recognition and Sentiment Analysis [J]. International Journal of Uncertainty, Fuzziness and Knowledge – Based Systems, 2017, 25(Suppl. 2):175 – 189

[64] Mhamdi M, West R G, Hossmann A, *et al*. Expanding the Text Classification Toolbox with Cross - Lingual Embeddings [J]. arXiv: Computation and Language, 2019
[65] 高国骥. 基于跨语言分布式表示的跨语言文本分类 [D]. 哈尔滨工业大学,2018
[66] Yao L, Mao C, Luo Y, *et al*. Graph Convolutional Networks for Text Classification [J]. national conference on artificial intelligence, 2019

作者简介: 汪岿(1991 -),男,在读硕士,研究领域为文本分类、数据挖掘; 刘柏嵩(1971 -),男,博士,研究馆员,博士生导师,研究领域为知识工程及计算机网络。■

简讯

中国信科 5G 新突破|中国信科联合重庆电信、中国汽研发布 5G 远程驾驶一阶段示范应用

5 月 15 日,由中国信科集团旗下大唐移动、重庆电信、中国汽研三方合作的 5G 智能网联汽车试点项目在中国汽车工程研究院园区正式发布,月内重庆市民就有望体验到基于 5G 网络的人车分离式远程驾驶。这是重庆首次发布的 5G 远程驾驶一期成果,也是全国首个城市交通场景下的 5G 远程驾驶应用示范。

5G 远程驾驶正式亮相山城 市民可实现远程观看

本月,重庆电信、中国汽研、大唐移动三方将在重庆仙桃国际大数据谷开展 5G 智能网联汽车一阶段试点体验。该阶段展示的三个主要场景为 5G + VR 应用、5G 视频直播应用、基于 5G 的车辆远程控制应用。

据悉,5G 远程驾驶是通过 5G 网络高带宽低时延,来实现人车分离的远程驾驶。用户可通过汽研院园区的驾驶模拟舱,实时远程操控远在仙桃数据谷的车辆: 5G 远程控制系统通过 5G 网络 + 天翼云实现用户远程对车辆的操作及控制; 车辆实时接收远程用户的控制命令,对自身进行控制; 同时实时反馈车辆自身运行状态及周边环境; 当车辆运行存在安全隐患或者数据传输中断时,能够主动靠边停车。

驾驶员在汽研院园区的驾驶模拟舱远程操控远在仙桃数据谷的车辆,而市民则可以在位于几十公里之外的解放碑利用 5G 网络亲眼见证车辆现场的实时情况。通过佩戴 VR 眼镜设备,体验者将获得 360°全景浸入式体验,观看车辆完成启动、加减速、转向等各种动作,如同身临其境。除此之外,车内搭载了多路摄像头,拍摄的画面也将利用 5G 网络传回并在大屏幕上直播,让体验者能够一睹车内的奥秘。

5G + VR 应用通过 5G 超大带宽,超低时延传输特性,使全景直播技术和 VR 虚拟现实技术相融合,通过 5G 网络实时传送至观看端,观看到位于仙桃数据谷的 5G 智能网联汽车行驶的过程,使用户有身临其境之感。

5G 视频直播实现基于 5G 网络的高带宽低时延车载视频与车辆轨迹的直播,车辆内部将安装 5 路摄像头,通过 5G 网络进行车内行驶直播。

强强联合 助力重庆智慧网联汽车产业

2019 年 1 月 16 日,由重庆电信、中国汽研、大唐移动三家企业签约,依托中国电信 5G 网络技术优势,发挥各自优势,强强联合,通过基础设施建设改造,技术融合、测试研发等多个方面在仙桃数据谷打造 5G 自动驾驶应用示范公共服务平台,探索天翼云、车联网、自动驾驶、智慧交通等系统解决方案。这是我国首个启动建设的 5G 自动驾驶应用示范公共服务平台,其中包括重庆首次发布的 5G 远程驾驶项目。

该项目落户仙桃数据谷,得益于其定位。据介绍,仙桃国际大数据谷位于渝北区,构建以数据为驱动的"1 +3 +5 +10 +N"创新生态圈,大力发展大数据、人工智能、云计算、物联网、智能汽车等智能产业,打造中国大数据产业生态谷。作为重庆市重要的数字经济创新高地,此次以车联网为切入,在谷内探索打造国内首个 5G 车联网及远程驾驶示范基地,加速推动重庆智慧网联汽车产业新生态,助力传统汽车企业转型升级。

目前,三方共同推进中的 5G 智能网联汽车项目,已完成重庆市仙桃数据谷和中国汽车研究院园区内的 5G 智能网联汽车行驶线路的综合设计,在仙桃数据谷和汽研院园区均已实现了 5G 基站信号连续覆盖。

据悉,实现一阶段示范应用后,后续将实现二阶段目标: 5G 高精度地图下载、危险场景预警、连续绿灯通行等能力。

来源于: 中国信科