

基于稀疏子空间的分布外文本检测

王祉苑¹, 彭涛¹, 杨捷²

(1. 武汉纺织大学 计算机与人工智能学院, 湖北 武汉 430200;
2. 伍伦贡大学 计算机与信息技术学院, 新南威尔士州 伍伦贡 2522, 澳大利亚)

摘要: 分布外(OOD)检测的目标是识别潜在的不符合训练过程中数据分布的样本, 以规避模型对于异常情况的预测。基于预训练语言模型的OOD检测方法面向文本分类任务时通常过度依赖数据标签, 因此在实际应用中性能受到限制, 而无监督OOD检测的相关研究目前尚不充分。为突破这一局限性, 提出新的OOD文本检测框架: 基于稀疏子空间的分布外文本检测(SSOD)。该框架无需标注数据, 利用稀疏子空间联合建模已知数据的特征分布, 构建观测样本在最相邻子空间的概率密度函数并将其作为OOD检测的评分指标。实验结果表明, SSOD面对不同分布偏移的平均AUROC和平均FAR95分别优于基准线2.2%和4.1%, 在综合性能上超过现有的监督方法。

关键词: 分布外检测; 预训练语言模型; 深度学习; 稀疏子空间聚类; 文本分类

DOI: 10.11907/rjdk.241251

开放科学(资源服务)标识码(OSID):

中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(2025)005-0070-09



Sparse Subspace Based Out-of-Distribution Text Detection

WANG Zhiyuan¹, PENG Tao¹, YANG Jie²

(1. School of Computer and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, China;
2. School of Computer and Information Technology, University of Wollongong, Wollongong 2522, Australia)

Abstract: The goal of Out-of-Distribution (OOD) detection is to identify potential samples that do not conform to the data distribution during the training process, in order to avoid the model making predictions on unusual cases. OOD detection methods based on pre-trained language models often overly rely on data labels for text classification tasks, which limits their performance in practical applications. Currently, there is insufficient research on unsupervised OOD detection. To overcome this limitation, a new OOD text detection framework called Sparse Subspace-based Out-of-Distribution Text Detection (SSOD) is proposed. This framework does not require labeled data and utilizes sparse subspaces to jointly model the feature distribution of known data. It constructs the probability density function of observed samples in the nearest subspace as the scoring metric for OOD detection. Experimental results show that SSOD achieves an average AUROC and average FAR95 that are respectively 2.2% and 4.1% higher than the baseline across different distribution shifts, surpassing existing supervised methods in overall performance.

Key Words: out-of-distribution detection; pre-trained language model; deep learning; sparse subspace clustering; text classification

0 引言

在自然语言处理(Natural Language Processing, NLP)领域,许多先进模型的性能评估都基于一个假设:训练数据和测试数据来源于同一分布。然而,这一假设在实际应用

中往往不成立,导致NLP模型在面对分布外(Out-of-Distribution, OOD)数据时性能急剧下降^[1-2]。将输入数据中的OOD样本与训练阶段的分布内(In-Distribution, ID)样本进行区分的过程,被称为OOD检测。有效的OOD检测能够防止模型在面对未见过的数据时产生误导性预测,这对文本分类和信息抽取等NLP任务至关重要,因为它们需

收稿日期: 2024-04-08

扫描二维码阅读全文:



基金项目: 中国高校产学研创新基金项目(2021ITA05012)

作者简介: 王祉苑(2000-),女,武汉纺织大学计算机与人工智能学院硕士研究生,研究方向为自然语言处理;彭涛(1981-),男,博士,武汉纺织大学计算机与人工智能学院教授、硕士生导师,研究方向为机器视觉、信息安全等;杨捷(1984-),男,博士,伍伦贡大学计算机与信息技术学院讲师、博士生导师,研究方向为自然语言处理、机器视觉等。本文通讯作者: 彭涛。

要模型具备对语义的理解和处理能力^[3-5]。

在 NLP 领域的研究中,预训练语言模型(如 BERT、RoBERTa)已被广泛认为是最先进的模型之一。基于 BERT 系列的 OOD 检测方法在多种 NLP 任务中都显示了良好效果^[6-8]。主流的 OOD 检测方法依赖于模型的输出置信度分数和某些特定的中间层表示,它们的核心思想是根据输入样本的特征表示以设计分数函数,通过该分数函数为 OOD 和 ID 样本赋予不同的分数,用于区分 ID 样本和 OOD 样本^[9-10]。例如,对于 ID 样本,分数普遍较高;而对于 OOD 样本,分数普遍较低。当文本分类作为下游任务时,这些方法通常需要大量的标注数据训练分布外检测器,而在开放环境中,模型获取足够的标签数据通常要花费大量的人工和时间成本。然而,目前针对无监督的 OOD 检测方法的相关研究相对有限^[11]。聚类作为一种常见的无监督学习方法,通过将数据分组为若干簇,使得同一簇内的数据点相似度高,而不同簇间的数据相似度低。在 OOD 检测中,聚类算法提供了一种基于数据自身特性进行探索的方法。

因此,本文考虑稀疏子空间聚类算法(Sparse Subspace Clustering, SSC),并结合多元高斯分布,提出了一种无监督 OOD 检测框架:基于稀疏子空间的分布外文本检测(SSOD)。通过预训练语言模型编码后的文本数据通常具有较高维度,传统的聚类方法可能难以解决高维数据空间中的聚类问题。稀疏子空间的优势在于其能够根据提取的数据特征将高维空间划分为多个低维子空间,并将特征相似的训练样本聚类到同一子空间中,更有效地处理高维数据。聚类后对每个子空间上的数据拟合高斯分布,使数据不再被视为一个抽象的数据矩阵,从而更直观地反映数据的分布情况。最后,在检测过程中通过寻找距离测试样本最近的子空间并利用其在该子空间上出现的似然性以实现对外来样本的有效区分。本文实验部分在多个文本分类任务上对 SSOD 的检测性能进行了评估,并将其应用于不同的预训练语言模型。实验结果证实,SSOD 在无需数据标签的情况下仍然能够有效地检测 OOD 样本,在不同的配置环境下也展现出了其稳定性和通用性。

综上,本文研究主要贡献如下:①将稀疏子空间聚类与多元高斯分布引入 OOD 检测框架,能够直观地描述数据在高维空间的内在表示,并对其特征分布进行建模,为深度学习中的 OOD 检测提供了新思路;②提出的 SSOD 方法不受限于特定模型,可以适用于多种基准模型,具有广泛的普适性。

1 相关工作

1.1 分布外数据

在许多 NLP 任务中,模型在预测过程中时常会遇到来自于与训练数据不同分布的数据点,这些数据点被称为分

布外(OOD)数据。OOD 数据会造成模型在执行任务时表现不佳,如在情感分析中,模型可能遇到与训练数据集中情境或话语完全不同的句子,这类句子往往会误导模型并给出错误预测。

遵循 Arora 等^[6]的分类体系,根据判断标签空间是否变化,将异常分布的 OOD 样本分为语义偏移(Semantic Shift, SS)和背景偏移(Background Shift, BS)两类。语义偏移发生在模型遭遇具有未知标签的输入时,例如一个用科技类和经济类新闻训练得到的新闻主题分类器接收到一个体育类的新闻文本;而背景偏移则是指模型处理已知标签但来自不同领域或风格的文本,例如用于电影评论的情感分析模型接收到一条餐厅评论。具体示例如表 1 所示。

Table 1 Specific examples of ID/OOD data

表 1 ID/OOD 数据具体实例

示例	标签	数据类型
这是我看过的最老套的电影。	负面	ID 数据
这部电影真是一部杰出的作品!	正面	ID 数据
我需要设计一张图表。	未知	语义偏移类型 OOD 数据
服务员的态度非常友好。	正面	背景偏移类型 OOD 数据
食物太咸,难以下咽。	负面	背景偏移类型 OOD 数据

OOD 数据帮助研究者意识到模型的局限性,并采取适当的措施以应对这种局限性,确保模型在实际应用中的鲁棒性和可靠性。

1.2 分布外检测

随着深度学习和 NLP 的迅速发展,处理训练分布以外的样本已成为一项紧迫而富有挑战性的任务。在这一背景下,分布外(OOD)检测成为近年来备受关注的研究领域,它旨在识别推理过程中偏离训练分布的示例,使得模型可以拒绝对它们进行预测^[12-13]。

Hendrycks 等^[14]在 OOD 检测领域的开创性研究为后续工作提供了基准,他们提出的最大 Softmax 概率(Maximum Softmax probability, MSP)方法已成为目前研究的参考标准^[15]。观察不同工作在训练方式上的差异性,将现有的 OOD 检测方法分为 3 类:

(1)有监督的 OOD 检测。这类方法直接引入 OOD 样本作为负类参与模型的训练过程。Zeng 等^[16]提出利用对抗攻击机制增强 OOD 数据,并利用高斯判别分析估计类条件分布,设计出一种基于生成距离的检测器。Lin 等^[17]提出的 FlatS 方法使用公共语料库作为辅助 OOD 数据,并计算输入数据基于特征的似然比分数以实现 OOD 检测。

(2)半监督的 OOD 检测。OOD 数据在训练中不可用,但 ID 标签为已知,利用 ID 样本的标签信息以辅助模型学习区分 ID 样本和潜在的 OOD 样本。文献[18]考虑到监督对比学习,通过最大化类间方差和最小化类内方差以优化上游表示。文献[19]也发现了类似工作。BARLE 利用掩码语言模型(Mask Language Model, MLM)对 ID 样本中的关键信息逐步进行[MASK]操作生成相应的伪 OOD 数据

集,并设计两种损失函数以提高检测效果^[20]。

(3)无监督的OOD检测。这类方法不引入ID样本的标签信息和OOD数据,模型需要依靠数据的内在结构和分布以尝试发现OOD样本。例如,Zong等^[21]提出一种无监督的OOD检测模型(DAGMM),该模型利用深度自编码器为每个数据点生成低维表示和重构误差,并将这些信息反馈至高斯混合模型中加以处理。

实际上,当ID标签信息未知时,问题设置回到了经典的异常检测问题^[22-23]。然而,这类问题设置在NLP领域中很少被研究。无监督OOD检测方法的主要侧重点在于提取更鲁棒的特征,对数据分布进行更准确的估计,这也是本文引入稀疏子空间聚类算法的动机。SSOD无需依赖ID数据的标签信息,省去了采集时间和成本,并使其更适用于各种应用场景。此外,本文在检测阶段集成了密度建模和距离计算这两种机制的优势,在提高检测精度的同时准确地捕捉了数据的内在结构和特征分布。

2 方法描述

在文本分类任务中,考虑一组训练句子 X_{train} ,这些数据被设定为从ID数据集中抽样得到。引入此假设的目的是引导模型根据训练数据的分布加以训练和优化。OOD检测则是判断测试样本 X_{test} 来自于ID分布还是OOD分布。本文提出的SSOD利用稀疏子空间聚类算法,根据训练集中的输入样本在RoBERTa-base模型中的特征表示,将数据的高维空间划分为若干个低维子空间,并对训练集中的所有数据点进行分簇。接下来,利用主成分分析重新描述每个子空间,使子空间中的数据近似地服从多元高斯分布。最后,通过计算距离找出最接近测试样本的子空间,将测试样本出现在该子空间的似然性作为OOD检测的度量指标。SSOD的整体框架如图1所示,其中输入样本表示文本数据,经过RoBERTa-base模型将输入的文本数据转化为嵌入空间中对应的特征向量(图1右侧实线矩阵框中的黑色方块)。

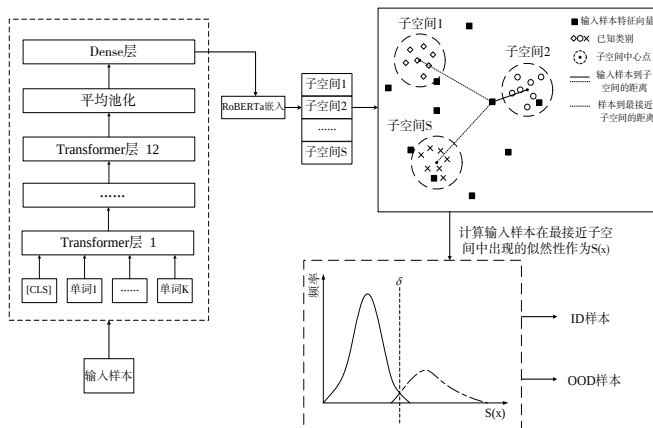


Fig. 1 SSOD overall framework

图1 SSOD整体框架

2.1 稀疏子空间聚类

稀疏子空间聚类(SSC)是一种通过数据的自表示性质对数据进行聚类的方法,它假设数据集中的每个点都可被看作是位于某个低维线性子空间中的一个点,而整个数据集是这些低维线性子空间的并集。换言之,虽然数据点可能位于高维空间中,但每个数据点实际上是由更低维度的线性子空间中的基向量线性组合而成。在这样的假设下,SSC的目标是划分出这些子空间,并将数据点正确地聚类到它们相应的子空间中。在SSC中,同一子空间中的数据点在某种程度上具有相似的特征表示,通过将这些相似的数据点聚集在一起,从而发现数据中的潜在结构。

设输入句子 b 为训练集中的某个数据点,经过预训练模型获得它相应的[CLS]特征向量 $\mathbf{x}$,通常模型输出的[CLS]标记的嵌入向量被视为整个输入序列的信息表示。再给定一个数据矩阵 $\mathbf{X}_S$,其中包含训练集中所有样本的特征向量,即 $\mathbf{X}_S = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{N_S}] \in \mathbb{R}^{N_S \times D}$,其中 N_S 是样本数量, D 是环境空间维度(例如,RoBERTa-base模型中的维度为768), $\mathbf{x}_i$ 表示某一个数据点的向量表示。SSC旨在将这些数据点分配到 S 个子空间中以实现数据点的聚类。

2.1.1 稀疏系数矩阵 $\mathbf{C}$ 构建

构建稀疏系数矩阵 $\mathbf{C}$ 的目的是挖掘数据点之间的关系,使得每个数据点可以被表示为其他数据点的线性组合。这种表示强调了子空间的成员关系,因为只有处在同一子空间中的点才会被用于构造这样的线性组合。与传统的子空间聚类方法不同,SSC通过 l_1 正则化以促进稀疏性,这意味着每个数据点的表示只依赖于少量的其他点。这种稀疏性是SSC的特点,它确保了只有同一子空间中的数据点会有非零系数,从而推动聚类。首先,根据数据矩阵 $\mathbf{X}_S$ 构建出一个稀疏系数矩阵 $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_N] \in \mathbb{R}^{N \times N}$,其中 $\mathbf{c}_i$ 表示数据点 $\mathbf{x}_i$ 的稀疏表示, $\mathbf{c}_i$ 可以通过以下优化问题求解获得:

$$\min_{\mathbf{c}_i} \|\mathbf{c}_i\|_1 \quad \text{s.t.} \quad \mathbf{x}_i = \mathbf{X}_S \mathbf{c}_i, \quad \mathbf{c}_{ii} = 0 \quad (1)$$

其中, $\|\cdot\|_1$ 是 l_1 范数,即向量元素的绝对值之和,并提出约束条件以避免数据点用其自身进行表示。

考虑训练集中的所有样本,则稀疏系数矩阵 $\mathbf{C}$ 可以表示为:

$$\min \|\mathbf{C}\|_1 \quad \text{s.t.} \quad \mathbf{X} = \mathbf{X}\mathbf{C}, \quad \text{diag}(\mathbf{C}) = 0 \quad (2)$$

2.1.2 亲和力矩阵 $\mathbf{W}$ 构建

得到稀疏系数矩阵 $\mathbf{C}$ 后,构建亲和力矩阵 $\mathbf{W}$ 以描述数据点之间的相似性。通常被定义为:

$$\mathbf{W} = |\mathbf{C}| + |\mathbf{C}|^T \quad (3)$$

其中, $|\cdot|$ 表示矩阵元素的绝对值。此时得到的亲和力矩阵 $\mathbf{W}$ 已经包含哪些样本属于同一子空间的初步信息,但为了追求对数据点的更合理划分,后续进行了谱聚类。

2.1.3 谱聚类

利用构建好的亲和力矩阵 $\mathbf{W}$,将数据点聚类到 S 个子空间中。谱聚类旨在找到一种合理的分割方式,将具有高

亲和力的数据点划分到同一类中, 同时将亲和力低的数据点划分到不同类别中。具体地, 先构建图拉普拉斯矩阵 L :

$$L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \quad (4)$$

其中, I 是单位矩阵, D 是一个对角矩阵, 其对角线上的元素 d_{ii} 等于亲和力矩阵 W 第 i 行元素之和。即:

$$d_{ii} = \sum_{j=1}^N w_{ij} \quad (5)$$

然后, 取图拉普拉斯矩阵 L 中最小的前 S 个特征值对应的特征向量 (其中 S 是预期的子空间数), 这些特征向量提供了数据在低维子空间中的新表示形式, 可使用这些特征向量获取数据在子空间上的投影。最后, 利用 K-means 聚类方法对数据点进行分簇, 从而分配数据点到相应的子空间中。需注意的是, SSC 主要是为了解释数据的子空间结构, 实际上是确定每个数据点在空间上的归属, 而不是改变数据本身的维度。

2.1.4 SSC 对噪声与异常值的处理

在 SSC 中, 每个数据点都被表示为其他数据点的稀疏线性组合, 这意味着在表示一个数据点时, 只有少数其他数据点 (这些数据点同属一个子空间) 会被用到, 而大多数其他点的系数将被强制为零, 而求解过程旨在最小化这种表示的稀疏性。因此, SSC 对噪声和异常值的处理主要依赖于其独特的稀疏表示。具体体现在如下两个方面:

(1) 稀疏性约束。SSC 利用 l_1 范数作为稀疏性正则化项, 鼓励每个数据点的表示尽可能少地使用其他数据点。这种稀疏性约束自然地降低了噪声和异常值对聚类结果的影响, 因为噪声和异常值很小可能被用于稀疏表示中的基元素。

(2) 重构误差。引入重构误差以衡量数据点通过其他数据点线性表示的准确性, 可以通过添加一个额外的约束加以实现, 该约束最小化重构误差的大小, 从而提高 SSC 对噪声和异常值的鲁棒性, 优化问题表示为:

$$E_i = \min_{c_i} \|c_i\|_1 + \lambda \|x_i - Xc_i\|_2^2 \quad (6)$$

其中, $\lambda \|x_i - Xc_i\|_2^2$ 表示 x_i 和其重构之间的欧几里得距离 (重构误差), 即原始数据和基于其他数据点的重构数据之间的差异; λ 是控制重构误差影响的权重参数。

通过最小化重构误差和稀疏性约束, SSC 能够在一定程度上抑制噪声和异常值的影响。

2.2 主成分分析与拟合高斯分布

聚类完成后, 训练集中的数据点在子空间中的位置将被确定。而在高维数据集中, 很多维度可能是冗余的, 或者不携带有效信息。因此, 本文对每个子空间进行主成分分析, 通过对数据降维, 原始数据的大部分信息会被压缩到较少的主成分中。通过观察提取出的主成分, 可以更容易地识别出 ID 样本中的 OOD 样本。同时, PCA 可以减少数据维度, 提高 OOD 检测时的工作效率。

设 $X = [x_1, x_2, \dots, x_N] \in \mathbb{R}^{V \times D}$ 为某一个子空间的数

据矩阵, 其中包含该子空间上的所有数据点。首先, 计算该数据矩阵 X 的均值向量 m 和协方差矩阵 Σ , 计算公式如下:

$$m = \frac{1}{N} \sum_i x_i \quad (7)$$

$$\Sigma = \frac{1}{N-1} \sum_i x_i^T x_i \quad (8)$$

然后, 对协方差矩阵 Σ 进行特征值分解, 从而获取 X 中关于所有数据点的所有成分:

$$\Sigma = U \Lambda U^T \quad (9)$$

其中, $U = [u_1, u_2, \dots, u_D]$ 包含了所有的正交向量。同时, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_D)$ 是一个对角矩阵, 它包含了数据的所有特征值, 本文方法中, 假设特征值大小均按降序排列。接下来, 选取数据中的前 d 个成分作为数据的主成分, 通过计算前 k 个主成分的累计方差与总方差的比值以确定预保留的主成分数量 d :

$$d = \arg \min_k \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^D \lambda_i} \geq \tau \quad (10)$$

其中, $\tau \in (0, 1]$ 是一个预定阈值, 通常 $\tau = 0.95$ 。

接着, 利用主成分分析后的结果描述子空间。具体而言, 选取矩阵 U 中提取出的前 d 个特征向量, 将其作为该子空间的坐标系, 并使这些坐标与中心点均值 m 对齐。这些坐标是基于数据矩阵 X 中的所有观测数据推导而来。将这些前 d 个特征向量组合成新的矩阵, 即提取出的主成分重新表示数据点, 记作 $P = [u_1, u_2, \dots, u_d]$ 。对于任意一个子空间 s , 可通过三元组描述它:

$$(P_s, m_s, \Sigma_s) \quad (11)$$

其中, P_s 代表数据点在子空间 s 上的坐标, 这些坐标由子空间包含的主成分决定, 简言之, 它们定义了子空间在整个数据空间中的方向和结构; m_s 代表子空间 s 的中心或均值点; Σ_s 代表子空间 s 的协方差矩阵, 它描述了子空间内数据点的分散程度。

依照主成分分析理论, 可以认为子空间 s 中的特征遵循一个多元高斯分布 $\mathcal{N}(m_s, \Sigma_s)$, 其中 m_s 是分布的中心, 而 Σ_s 描述了该分布的形状和扩散情况。换言之, 在 PCA 背景下, 人们认为数据特征在其主成分子空间中是服从高斯分布的。

2.3 OOD 检测

检测 OOD 样本的任务本质上是一个二元假设检验, 它的目标通常是设计一个分数函数 $S(x)$, 该函数将每个输出的数据点映射到一个标量上, 并利用这个标量区分 ID 样本和 OOD 样本。从数学形式化的角度讲, OOD 检测器 G 可以描述为:

$$G(S(x), \delta) = \begin{cases} \text{ID} & S(x) \geq \delta \\ \text{OOD} & S(x) < \delta \end{cases} \quad (12)$$

其中, δ 代表一个预定义阈值, 它可以通过研究需求进

行调整。例如,本文在实验中设置阈值以确保 ID 样本的召回率为 95%。

在本文所提出的检测背景下,当模型在测试阶段面对一个新的数据时,它无法判断这个数据是否为一个 OOD 样本。对于一个测试样本 x ,首先通过计算 x 到每个子空间中心点 m 的距离找出最接近 x 的子空间 h ,即确定新数据点 x 最有可能属于哪个子空间。公式如下:

$$h = \arg, \min \| (I - P_s P_s^\top) (x - m_s) \|_F^2 \quad (13)$$

其中, I 是单位矩阵, $P_s P_s^\top$ 是子空间 s 的投影矩阵, m_s 是子空间 s 的中心点。 $\| \cdot \|_F^2$ 是 Frobenius 范数的平方, 计算矩阵所有元素的平方和。

在确定了测试样本 x 最可能所属的子空间 h 之后, 计算它在该子空间中出现的似然性, 该似然性是基于一元高斯分布计算得到, 而该高斯分布的均值和标准差分别与该子空间的中心和协方差对应。高的似然性意味着 x 来自于子空间 h 的可能性大, 反之则 x 不一定来自于该子空间。这可以通过以下公式得到:

$$\mathcal{L}(x) = p(x | m_h, \Sigma_h) \quad (14)$$

其中, $p(x | m_h, \Sigma_h)$ 表示多元高斯分布的概率密度函数, 用于计算测试样本 x 在均值为 m_h , 协方差矩阵为 Σ_h 的高斯分布中出现的概率。将基于 x 计算出的似然性作为本文 OOD 检测的分数函数 $S(x)$, 即:

$$\mathcal{L}(x) = S(x) \quad (15)$$

最后, 根据分数对测试样本进行二分类任务, 从而实现 OOD 检测。

3 实验结果与分析

3.1 数据集与基准方法

3.1.1 数据集

实验中, 将 SSOD 方法分别用于检测语义偏移和背景偏移类型的 OOD 样本, 以对 SSOD 的 OOD 检测性能进行评估。因此, 设置两种不同类型的 OOD 场景, 每种场景包含不同的数据集, 使用文本数据并忽略相应的类别标签。

(1) 语义偏移的 ID/OOD 场景。考虑新闻主题分类这一常见的 NLP 任务, 因为它通常要面对出现新类别的 OOD 样本。实验选择 News Category (NC) 数据集^[24]构建该 ID/OOD 场景, 其中共包含 42 个类别。具体而言, 取 NC 数据集中排名前 12 个标签中的文本数据, 将其中 8 个类别的样本作为 ID 数据用于模型训练, 剩余的 4 个类别作为 OOD 数据模拟模型检测时遭遇的未知类别。

(2) 背景偏移的 ID/OOD 场景。在这一场景下考虑情感分析任务, 因为在情感分析中, 对于不同数据集上的样本, 它们共享相同的类别标签 (正面情感和负面情感)。选择 SST2^[25] 和 IMDB^[26] 这两个常用数据集, 将 IMDB 数据集作为 ID 数据, SST2 数据集作为 OOD 数据。其中, SST2 数据集包含观众简短的电影评论, 而 IMDB 数据集包含更长

且更专业的电影评论。

此外, 引入 AGNews 数据集用于评估跨任务设置下的 OOD 检测性能。语义偏移和背景偏移的 ID/OOD 场景下所包含的数据集框架如表 2 所示。

Table 2 The dataset framework used in the experiment

表 2 实验所用的数据集框架

场景	任务	ID	OOD
语义偏移	新闻主题分类	NC 前 12 类中的 8 类	NC 前 12 类中剩余 4 类
背景偏移	情感分析	IMDB	SST2

3.1.2 基准方法

实验中引入上文提及的 4 个 OOD 检测方法与 SSOD 进行比较: ①对于有监督的 OOD 检测方法 (OOD 数据可用 +ID 标签信息可用), 测试了 FLaTS 方法^[17]; ②对于半监督的 OOD 检测方法 (OOD 数据不可用 +ID 标签信息可用), 选取 BARLE 方法^[20]和 doSCL-cMaha 方法^[19]; ③对于无监督的 OOD 检测方法 (OOD 数据 +ID 标签信息不可用), 评估了 DAGMM 模型^[21]。

3.2 评估指标

遵循以往研究, 采用两个广泛使用的评估指标: AUROC 和 FAR95^[7]。AUROC 是评估二元分类器性能的常用指标, 它衡量了模型在不同阈值下的真阳性率 (True Positive Rate, TPR) 与假阳性率 (False Positive Rate, FPR) 之间的权衡关系。TPR 表示在实际为正例的样本中, 模型正确地将其判定为正例的比例; 而 FPR 表示在实际为负例的样本中, 模型错误地将其判定为正例的比例。AUROC 本质上是计算 ROC 曲线下的面积, 该面积的取值范围为 0~1 区间, 数值越接近 1 表示模型性能越好。AUROC 的数学表达式如下:

$$AUROC = \int_0^1 (TPR \times \frac{1}{FPR}) d(\frac{1}{FPR}) \quad (16)$$

在本文实验中, AUROC 用于衡量模型在区分 ID 和 OOD 样本时的准确性, 它可以被解释为当随机地从 ID 分布和 OOD 分布中各选 1 个样本, OOD 检测器输出的 ID 样本置信分数高于 OOD 样本分数的概率。

FAR95 用于计算在 TPR 为 95% 时, 二元分类器将负例错误地判定为正例的比例。FAR95 的公式可以表示为:

$$FAR95 = FPR \times \frac{0.95}{TPR} \quad (17)$$

在本文实验中, FAR95 指当有 95% 的 ID 样本被召回时, 其中被误认为 ID 样本的 OOD 样本比例。较高的 AUROC 和较低 FAR95 表明 OOD 检测性能更优。

3.3 实验设置

本文提出的 SSOD 框架中, 使用官方发布的 Transformer 库中的 RoBERTa-base 模型作为上下文编码器, 并将所有层的 Dropout 率设置为 0.1。通过 RoBERTa-base 模型从输入文本中提取特征向量, 在后续检测模型的嵌入空间中对这些特征进行处理和分析。采用 Adam 优化器对 SSOD 检测模型进行训练, 学习率为 $2e^{-5}$ ^[27]。每个输入序列的最

大长度为 512 个 tokens,批量大小设置为 16,epoch 为 5,以上超参数用于在训练集上进行训练。每完成一个 epoch 后,都会在验证集上对模型进行评估,并选择最佳检查点作为最终检测模型。为确定最佳子空间数,本文遵循了 Cai 等^[28]提出的聚类稳定性方法。将参数范围设置在 2~10 并依次进行尝试,对于不同任务上的训练集,分别进行了 30 次随机采样,在每次聚类中,使用相同的参数设置。通过计算聚类结果的 Jaccard 系数,选择具有最高平均 Jaccard 系数的子空间数作为最佳子空间数。最后,为确保实验中的检测方法能够进行公平比较,从 ID 测试集中随机选择 q 个样本,并从 OOD 测试集中随机选择 $\frac{q}{2}$ 个样本。具体而言,对于主实验中(3.4.1 节)引用的 NC_ID、IMDB 数据集,设置了不同的 q 值,分别为 2 000 和 4 000;对于跨任务场景下引用的 AGNews 和 SST2 数据集, q 值分别设置为 2 000 和 1 000。本文所有数据集的训练集、验证集和测试集等统计信息如表 3 所示。

Table 3 Statistical information of the dataset used in the experiment							
表 3 实验所用数据集统计信息							(个)
场景	任务	数据集	类别数	训练集	验证集	测试集	平均长度
语义偏移	新闻主题	NC_ID	8	68 000	8 160	8 160	30
	分类	NC_OOD	4	-	-	10 000	29
背景偏移	情感分析	IMDB	2	23 000	2 000	25 000	230
		SST2	2	6 920	872	1 821	19
跨任务设置	文本分类	AGNews	4	115 000	3 450	3 450	23

3.4 实验结果

3.4.1 语义偏移和背景偏移 ID/OOD 场景下的实验结果

为了验证本文 SSOD 方法在 OOD 检测中的有效性,将 SSOD 与上述基准方法进行对比实验,并对它们在语义偏移和背景偏移两种场景下的 OOD 检测性能分别进行了评估。所有方法在不同数据集上的 OOD 检测结果如表 4 所示,其中 AUROC 的值越高说明检测性能越好(↑),FAR95 的值越低说明检测性能越好(↓),平均 AUROC 和平均 FAR95 代表了每个方法在两种场景下的综合性能表现。

从实验结果可以看出,本文 SSOD 方法在 OOD 检测中的综合性能上优于其他方法;doSCL-cMaha 方法在检测语

义偏移类型的 OOD 样本时表现出色,但在检测背景偏移类型的 OOD 样本时性能并不理想。相较于表现最佳的基准方法,SSOD 的平均 AUROC 提高了 2.2%,平均 FAR95 降低了 4.1%。这表明 SSOD 不仅能够有效地检测出 OOD 样本,还具有良好的通用性。对于 SSOD 方法综合检测性能优于 doSCL-cMaha 方法的原因,可能在于 doSCL-cMaha 是基于距离测算得到评分函数,而 SSOD 方法结合了密度和距离两种评分机制的优势,能够更好地理解和表示数据之间的关联程度,从而增强了簇之间的区分度。

此外,实验涉及的所有方法在语义偏移场景下的检测性能普遍优于背景偏移场景下的检测性能,这可能意味着大部分检测方法更容易识别出任务中包含未知类别的 OOD 样本。关于 OOD 检测方法在检测背景偏移类型 OOD 数据时性能的欠缺,推测这可能与模型在训练过程中获得的特征表示有关,模型更专注于学习与任务相关的特征,但同时可能忽略或削弱了在预训练阶段学习到的与任务无关的通用特征。

为了更直观地体现出本文方法的优势所在,绘制了 DAGMM 方法和 SSOD 方法的 OOD 检测评分直方图,如图 2 所示。其中,ID 数据来自于 IMDB 数据集,OOD 数据来自于 SST2 数据集。

对于 DAGMM 方法,ID 和 OOD 曲线之间的重叠面积较大,这表示 ID 和 OOD 分数比较接近,导致 OOD 检测效果不够好(AUROC 为 65.8%);而对于本文提出的 SSOD 方法,ID 和 OOD 曲线之间重叠面积较小,这表示 ID 和 OOD 分数差异较大,比 DAGMM 方法更容易区分 ID 数据和 OOD 数据,从而更有效地进行 OOD 检测(AUROC 为 90.1%)。

3.4.2 跨任务设置下的实验结果

值得注意的是,语义偏移和背景偏移这两种 OOD 数据类型并不是相互独立存在,在现实世界中的 OOD 数据时常出现兼具语义和背景两种偏移的情况。因此,除验证 SSOD 方法在检测不同类型 OOD 数据时的有效性外,本文模拟了两种偏移同时发生的情景,以评估 SSOD 和基准方法在跨任务设置下的 OOD 检测性能。

遵循文献[29]的研究工作,考虑情感分析和新闻分类这两类 NLP 任务进行跨任务实验。具体而言,选择 SST2 数据集和 AGNews 数据集以构建一个 ID/OOD 对,当其中一个数据集作为 ID 数据时,另一个数据集被视为 OOD 数据。AGNews 数据集是一个广泛应用于文本分类任务的新闻语料库,涵盖 4 个主题领域,具体信息如表 3 所示。对于每个 ID 数据集,显示了所有方法在对应 OOD 数据集上 AUROC 和 FAR95 的宏平均值,该值是基于 5 个随机种子计算平均得到的结果。跨任务设置下的实验结果如表 5 所示。

根据实验结果显示,SSOD 方法在跨任务设置中的检测性能优于所有基准方法,相较于表现最佳的基准方法,本文方法在 AUROC 值上提高了 2.9%,并在 FAR95 值上降低了 4.4%。这可能是由于当语义偏移和背景偏移情况同

Table 4 Experimental results of semantic shift and background shift in ID/OOD scenarios						
表 4 语义偏移和背景偏移 ID/OOD 场景下的实验结果 (%)						
方法	平均 AUROC	AUROC		平均 FAR95	FAR95	
		NC_ID-NC_OOD	IMDB-SST2		NC_ID-NC_OOD	IMDB-SST2
DAGMM	70.7	75.6	65.8	69.2	58.3	80.1
BARLE	86.2	88.4	83.9	41.9	37.6	46.1
FLatS	90.1	91.6	88.5	29.5	24.8	34.2
doSCL-cMaha	88.5	95.2	81.7	15.9	9.5	22.3
Ours	92.3	94.4	90.1	11.8	11.2	12.4

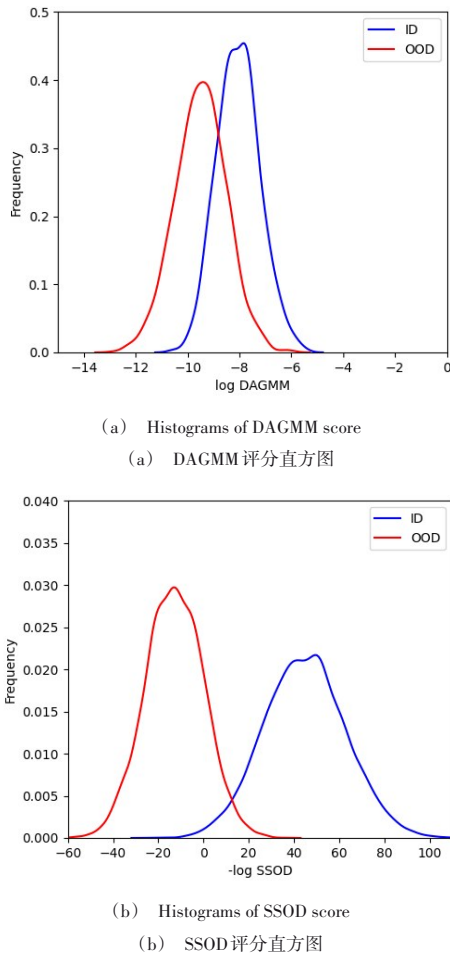


Fig. 2 Histograms of OOD detection scores for DAGMM and SSOD methods

图2 DAGMM方法和SSOD方法的OOD检测评分直方图

Table 5 Experimental results under cross-task settings

表5 跨任务设置下的实验结果 (%)

ID	OOD	方法	AUROC	FAR95
SST2	AGNews	DAGMM	67.9	64.4
		BARLE	78.1	57.9
		FLatS	71.3	66.8
		doSCL-cMaha	80.7	52.6
		Ours	83.5	48.2
AGNews	SST2	DAGMM	69.7	47.1
		BARLE	82.6	28.4
		FLatS	80.0	27.5
		doSCL-cMaha	83.5	21.4
		Ours	86.4	20.7

时发生时,SSOD方法能够更直观地揭露数据在子空间中独特的分布结构,保持了SSOD在跨任务场景设置下检测的稳定性。但由于任务的复杂性增加,可能导致特征表示发生冲突,因而与单一的语义偏移或背景偏移的ID/OOD场景相比,跨任务设置下检测方法的性能表现略有下降。

此外,在将AGNews数据集作为ID数据时,所有方法的性能表现都优于将SST2数据集作为ID数据时的情况。对于这一现象,本文从如下3个方面分析其中原因:①数

据集特征分布:AGNews数据集包含4个类别,而SST2数据集包含2个类别,因而AGNews数据集存在更丰富的特征分布,这使得该数据集作为ID数据时,模型可以学习到更多样化的特征表示,从而提高检测性能;②类别样本均衡性:AGNews数据集中每个类别的样本数量相对均衡,这可以引导模型学习到各主题的区别性信息,从而更容易识别出其中的OOD样本;③数据集样本数量:相比SST2数据集,AGNews数据集的样本数量更多,这为模型提供了更多样本,使得模型在训练过程中能够更准确地捕捉数据的特征分布,从而取得更好的检测性能。

3.4.3 SSC与其他聚类算法在OOD检测中的效果比较

为验证在SSOD框架中,SSC相较于其他传统聚类方法在OOD检测上的优势,本文设计了一组消融实验。根据主实验(3.4.1节)中的实验设置,模拟了语义偏移和背景偏移两种场景,并将SSC替换为其他聚类算法。具体而言,遵循3.3节中的参数设置,在不同的训练集上均使用RoBERTa-base模型作为文本编码器;随后,使用其他聚类算法对提取出的数据特征进行分簇和聚类,与原始SSOD模型的训练参数保持一致。模型训练完成后,根据第3.3节中 q 值的设置,分别选用NC_ID和IMDB数据集中的测试集进行OOD检测,观察不同聚类方法在检测语义偏移和背景偏移OOD数据时的性能表现。

在本次实验中,选择两种传统聚类算法与SSC进行比较,分别是基于划分的K-means聚类算法和基于密度的DBSCAN聚类算法。以上两种算法均来自于Sklearn库。引入轮廓系数(Silhouette Score)作为评价指标,调整其中的参数以保证这两种算法的聚类质量。轮廓系数的范围为 $[-1, 1]$,且接近1的轮廓系数表示较好的聚类效果。

对于K-means聚类算法,在NC_ID和IMDB数据集上分别尝试不同的 k 值(范围设置为2~10),通过计算每个 k 值下的平均轮廓系数(数据集中所有样本轮廓系数的平均值)以评估聚类质量。根据此评估,最终确定在NC_ID数据集上的最优 k 值为8,在IMDB数据集上的最优 k 值设置为2。此外,K-means算法实验还设置最大迭代次数为100。DBSCAN算法不需要预定义簇的个数,而是通过调整两个主要参数确定聚类结果:领域半径Epsilon(Eps)和每个簇中必须包含的最小点数MinPoint。准备一组Eps值(0.5, 0.35, 0.2, 0.1, 0.05)和一组MinPoint值(2, 5, 10, 15, 20),通过评估不同的参数组合在不同数据集上的平均轮廓系数以确定最佳参数配置。在NC_ID数据集上,选择Eps值为0.5和MinPoint值为10;在IMDB数据集上,选择Eps值为0.2和MinPoint值为5作为最佳参数配置。

当不同聚类算法应用于SSOD框架时,语义偏移和背景偏移两种场景下的OOD检测表现(FAR95↓)如表6所示。

实验结果表明,在两种偏移类型下,SSC在OOD检测任务中的表现优于其他聚类算法。相比于K-means算法

Table 6 OOD detection performance of different clustering algorithms in two scenarios

表 6 不同聚类算法在两种场景下的 OOD 检测表现 (%)		
聚类算法	FAR95	
	NC_ID-NC_OOD	IMDB-SST2
k-Means	31.1	46.8
DBSCAN	27.8	31.9
SSC (Ours)	11.2	12.4

和 DBSCAN 算法,SSC 算法在检测中的平均 FAR95 值分别降低 27.2% 和 18.1%。此外还发现,无论更换哪种聚类方法,语义偏移场景下 SSOD 框架及其变体的 OOD 检测性能依然优于背景偏移场景下的性能,这再次验证了之前的猜想:OOD 检测性能与训练过程中模型学习到的特征表示有关。

为了直观地理解不同聚类算法对 OOD 检测的影响,比较了 K-means 和 SSC 这两种聚类方法的检测结果:选取 IMDB 作为 ID 数据集,SST2 作为 OOD 数据集的场景设置,并使用 t-SNE^[30]进行数据特征的可视化。如图 3 所示,在 K-means 算法作为聚类手段时(图 3(a)),ID 和 OOD 数据点出现明显重叠;而 SSC 算法(图 3(b))能够将 OOD 类从 ID 样本中分离,使其得以明显区分。可以推测,SSC 能够改善 OOD 检测效果,可能是因为它更擅长处理高维数据,从而弥补了传统聚类算法在维度上的限制。因此,在高维空间中,模型对数据的处理能力可能会影响 OOD 检测方法的效果。

3.4.4 其他预训练语言模型推广

本文使用两个额外的预训练语言模型对数据进行预训练,并从中提取文本特征。选择两个不同的模型进行测试,分别是 BERT-base 和 RoBERTa-large。通过对不同词嵌入结果的测试,以验证 SSOD 方法的通用性。遵循 3.4.1 节中相同的实验与参数设置,对于语义偏移场景,在 NC 数据集上进行 OOD 检测;对于背景偏移场景,选择 IMDB 数据集(ID)和 SST2 数据集(OOD)进行 OOD 检测。SSOD 在不同模型上的 OOD 检测性能(FAR95↓)如表 7 所示。

从实验结果可以看出:①在不同的模型上普遍存在语义—背景偏移不平衡问题,由于 OOD 数据的偏移类型不同,导致同一种检测方法在不同场景下的性能出现差异;②相比于 BERT-base 模型,RoBERTa-base 模型在预训练阶段使用了更多样化的数据集,这表明在多样化的数据上进行预训练可以提升文本的 OOD 检测能力;③RoBERTa-large 模型的表现不如 RoBERTa-base 模型,这意味着更大的模型并不一定能为 OOD 检测带来更好的性能。

4 结语

分布外检测的目标是检测来自于训练数据分布以外的样本,为达到这一目标,本文提出了 OOD 检测框架:基于稀疏子空间的分布外文本检测(SSOD)。通过分解和提

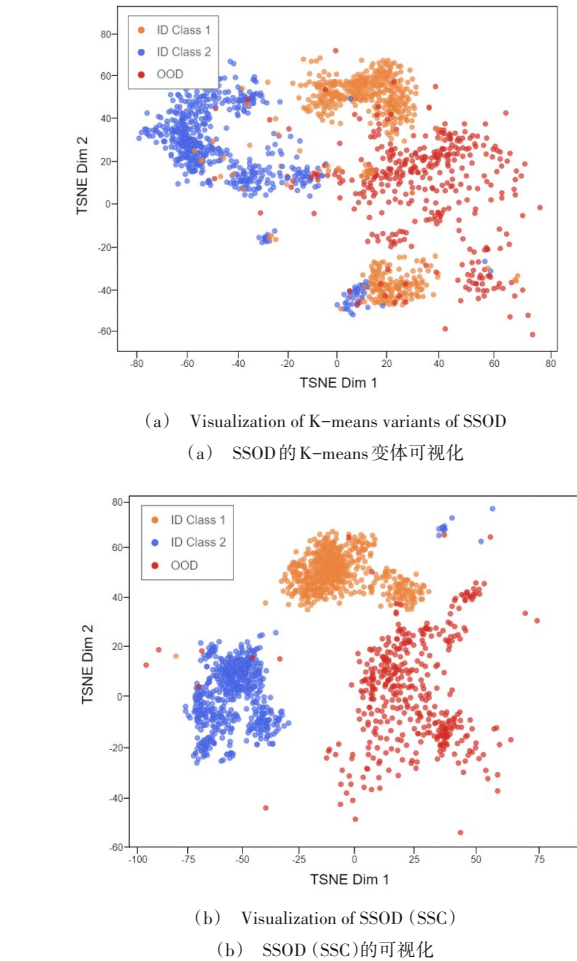


Fig. 3 t-SNE visualization of SSOD and its K-means variant in OOD detection

图 3 SSOD 及其 K-means 变体在 OOD 检测中的 t-SNE 可视化

Table 7 OOD detection performance on different pre-trained language models

表 7 不同预训练语言模型上的 OOD 检测性能 (%)			
模型	方法	FAR95	
		NC_ID-NC_OOD	IMDB-SST2
BERT-base	doSCL-cMaha	14.7	28.5
	Ours	12.6	17.7
RoBERTa-base	doSCL-cMaha	9.5	22.3
	Ours	11.2	12.4
RoBERTa-large	doSCL-cMaha	13.8	30.9
	Ours	15.3	19.2

取数据中的特征,找出数据在空间中独特的分布规律,以检测外来分布的数据。实验结果证明,SSOD 能够有效识别 ID 数据和 OOD 数据,并在不同文本分类任务上达到最佳的综合检测性能。

未来,将探索不同的特征提取方法和聚类策略,并将 SSOD 方法应用于其他 NLP 任务。此外,还可以通过在真实场景里评估 SSOD 方法的性能和可行性,更深入地了解其应用潜力。这些拓展研究将进一步推动 OOD 检测领域的发展,为未来研究提供更多探索空间。

参考文献:

- [1] HEIN M, ANDRIUSHCHENKO M, BITTERWOLF J. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 41–50.
- [2] WU Q, JIANG H, YIN H, et al. Multi-level knowledge distillation for out-of-distribution detection in text [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 7317–7332.
- [3] KAMATH A, JIA R, LIANG P. Selective question answering under domain shift [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 5684–5696.
- [4] YILMAZ E, TORAMAN C. D2U: distance-to-uniform learning for out-of-scope detection [C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022: 2093–2108.
- [5] GOU P F, SONG C Y. Outlier detection method using adaptive neighbor graph[J]. Application Research of Computers, 2023, 40(11): 3309–3314. 邬鹏飞, 宋承云. 基于自适应邻居图的离群点检测方法[J]. 计算机应用研究, 2023, 40(11): 3309–3314.
- [6] ARORA U, HUANG W, HE H. Types of out-of-distribution texts and how to detect them [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021: 10687–10701.
- [7] PODOLSKIY A, LIPIN D, BOUT A, et al. Revisiting mahalanobis distance for transformer-based out-of-domain detection [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 13675–13682.
- [8] YANG J, WANG P, ZOU D, et al. Openood: benchmarking generalized out-of-distribution detection [C]// Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022: 32598–32611.
- [9] LEE K, LEE K, LEE H, et al. A simple unified framework for detecting out-of-distribution samples and adversarial attacks [C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018: 7167–7177.
- [10] LIU W, WANG X, OWENS J D, et al. Energy-based out-of-distribution detection [C]// Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020: 21464–21475.
- [11] LANG H, ZHENG Y, LI Y X, et al. A survey on out-of-distribution detection in NLP [DB/OL]. <https://arxiv.org/abs/2305.03236>.
- [12] BARAN M, BARAN J, WOJCIK M, et al. Classical out-of-distribution detection methods benchmark in text classification tasks [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 119–129.
- [13] CHO H, PARK C, KIM J, et al. Probing out-of-distribution robustness of language models with parameter-efficient transfer learning methods [C]// Proceedings of the 12th Joint Conference on Lexical and Computational Semantics, 2023: 225–235.
- [14] HENDRYCKS D, GIMPEL K. A baseline for detecting misclassified and out-of-distribution examples in neural networks [C]//Toulon: International Conference on Learning Representations, 2017.
- [15] ZHANG Y, DENG W, ZHENG L. Unsupervised evaluation of out-of-distribution detection: a data-centric perspective [DB/OL]. <https://arxiv.org/abs/2302.08287>.
- [16] ZENG Z, XU H, HE K, et al. Adversarial generative distance-based classifier for robust out-of-domain detection [C]// IEEE International Conference on Acoustics, Speech and Signal Processing, 2021: 7658–7662.
- [17] LIN H, GU Y. FLatS: principled out-of-distribution detection with feature-based likelihood ratio score [C]//Proceedings of the 2023 Conference on Empirical Method in Natural Language Processing, 2023: 8956–8963.
- [18] ZENG Z, HE K, YAN Y, et al. Modeling discriminative representations for out-of-domain detection with supervised contrastive learning [C]//Proceeding of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 870–878.
- [19] WANG B, MINE T. Optimizing upstream representations for out-of-domain detection with supervised contrastive learning [C]//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023: 2585–2595.
- [20] DUAN H, YANG Y, ABBASI A, et al. BARLE: background-aware representation learning for background shift out-of-distribution detection [C]// Findings of the Association for Computational Linguistics: EMNLP 2022, 2022: 750–764.
- [21] ZONG B, SONG Q, MIN M R, et al. Deep autoencoding gaussian mixture model for unsupervised anomaly detection [C]//Vancouver: International Conference on Learning Representations, 2018.
- [22] CHALAPATHY R, CHAWLA S. Deep learning for anomaly detection: a survey [DB/OL]. <https://arxiv.org/abs/1901.03407>.
- [23] PANG G, SHEN C, CAO L, et al. Deep learning for anomaly detection: a review [J]. ACM Computing Surveys (CSUR), 2021, 54(2): 1–38.
- [24] MISRA R. News category dataset [DB/OL]. <https://arxiv.org/abs/2209.11429>.
- [25] SOCHER R, PERELYGIN A, WU J, et al. Recursive deep models for semantic compositionality over a sentiment treebank [C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013: 1631–1642.
- [26] MAAS A, DALY R E, PHAM P T, et al. Learning word vectors for sentiment analysis [C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011: 142–150.
- [27] KINGMA D P, BA J. Adam: a method for stochastic optimization [DB/OL]. <https://arxiv.org/abs/1412.6980>.
- [28] CAI J, FAN J, GUO W, et al. Efficient deep embedded subspace clustering [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 1–10.
- [29] OUYANG Y W, GAO Y, ZONG S, et al. Pobe: a generative model based out of distribution text detection method [J]. Journal of Software, 2024, 35(9): 4365–4376.
欧阳亚文, 高源, 宗石, 等. Pobe: 一种基于生成式模型的分布外文本检测方法 [J]. 软件学报, 2024, 35(9): 4365–4376.
- [30] MAATEN L, HINTON G. Visualizing data using t-SNE [J]. Journal of Machine Learning Research, 2008, 9(11): 2579–2605.

(责任编辑: 孙娟)