计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200400140

基于最优输运和 k-近邻的离群文档检测

水泽农　张星宇　沙朝锋
复旦大学计算机科学技术学院　上海 200433
(shuizenong@fudan.edu.cn)

摘　要　离群点或异常检测是数据挖掘和机器学习等领域的研究热点之一,研究人员已提出了多种离群点检测方法,并将其应用于入侵检测和异常交易检测等问题。但多数离群点检测方法主要针对表数据或时间序列数据等,无法直接应用于离群文档检测。现有基于相近性的离群文档检测方法一般用文档与整个文档集的距离来衡量离群性,无法发现基于局部考量的离群文档,而且采用欧几里德距离可能无法刻画出文档间的语义相近性。基于概率模型的离群文档检测方法过于复杂,并且同样只从全局来定义文档的离群值。针对这些问题,文中提出了一种新的基于相近性的离群文档检测方法。该方法引入最优输运距离,基于利用文档词嵌入向量的语义信息,在文档之间使用最优输运算法以度量距离,并利用 LDA 主题模型对文本进行层级抽象,通过最优输运算法算出主题之间的距离后,再计算文档距离,文中基于这两种最优运输距离计算文档与它的 k 近邻文档之间的距离来衡量该文档的离群程度。该方法从局部视角来定义文档的离群性,所采用的文档距离能体现文档之间的语义相近性。在两个开源数据集上进行了较细致的对比实验,实验结果显示,所提方法在多个指标上优于基准离群文档检测方法;还检验了基于 k 近邻离群文档定义的有效性以及 k 值的选取对结果的影响。

关键词:离群文档检测;最优输运;词搬动距离;层次型最优主题输运
中图法分类号　TP311

Outlier Document Detection via Optimal Transport and k-nearest Neighbor

SHUI Ze-nong, ZHANG Xing-yu and SHA Chao-feng
School of Computer Science, Fudan University, Shanghai 200433, China

Abstract　Outlier or anomaly detection is one of the research hotspots in areas such as data mining and machine learning, and researchers have proposed a variety of outlier detection methods that can be applied to problems such as intrusion detection and anomalous transaction detection. However, most outlier detection methods mainly target tabular data or time series data, etc. and cannot be directly applied to outlier document detection. Existing outlier detection methods based on proximity generally measure proximity by the distance of a document to the entire document set, failing to find outliers based on local considerations, and may not be able to characterize semantic proximity between documents using Euclidean distance. Probabilistic model-based outlier document detection methods are too complex and define document outliers only globally. In response to these questions, this paper proposes a new proximity-based outlier document detection method where we measure the outlier of a document by the distance between the document and its k-nearest neighbor document. We introduce the optimal transport algorithm to calculate the distance between documents, based on the semantic information of the document obtained from word embedding vector and the topic model. The method defines document outliers from a local perspective, using document distances that reflect the semantic proximity between documents. This paper conducts extensive experiments on two open source document datasets, and the results show that the proposed methods outperform the benchmark outlier document detection methods in terms of four evaluation metrics. Experiments also demonstrate the effectiveness of proposal of k-nearest neighbor based outliers and the impact of value k.

Keywords　Outlier document detection, Optimal transport, Word mover's distance, Hierarchical optimal topic transport

1　引言

离群点是数据集中非典型的数据点,与数据集中的大部分数据点相异。离群点检测指寻找严重偏离数据整体分布范围的离群数据,如何有效地进行离群点检测是数据挖掘和机器学习等领域的研究热点之一[1]。研究人员已经提出了大量基于概率模型或相近性等的离群点检测方法,并将其广泛应用于计算机安全(如入侵检测)、数据质量(如去除噪声数据)

到稿日期:2020-04-29　返修日期:2020-09-05
基金项目:国家重点研发计划(2018YFB0904503)
This work was supported by the National Key Research and Development Program of China(2018YFB0904503).
通信作者:沙朝锋 (cfsha@fudan.edu.cn)

和金融(如异常交易检测)等诸多领域。文献[2]以局部密度为基础,提出了一种基于谱嵌入和局部密度的离群点检测算法。文献[3]针对现有离群点检测算法运用于大规模数据集上时间效率较低的问题,提出了一种并行离群点检测算法。文献[4]提出了一种基于混合式聚类算法,用于在异常检测中寻找离群点。

目前离群点检测主要针对表数据、图数据以及时间序列数据等,用于在文档集中发现离群文档的研究工作还相对较少,文献[5-6]是其中较有代表性的工作。文献[5]提出了一个概率模型 vMF 来对文档集建模,由此定义每个文档的离群值,但该模型较为复杂,且只从全局视角考虑了文档的离群性。而文献[6]则采用基于距离的离群点检测方法,以对话系统中的句子集为文本数据,通过计算句子向量和句子集合向量平均值之间的欧氏距离(Euclidean Distance,EUC)来衡量该句子的离群值。本文主要关注基于相近性的离群文档检测方法。

大多数离群点检测算法针对的都是向量型数据。若是高维数据,则可先采用 PCA 或 SVD 对数据进行降维,然后再进行离群点检测。而对于文本这类非结构化数据,首先要得到文档的表示,如向量化。可以采用向量空间模型(如将文档表示为 TF-IDF 向量),或者采用主题模型 LDA(Latent Dirichlet Allocation)[7]来得到每个文档的主题分布,或者用最新的深度神经网络学习词(如 word2vec[8])、句子(如 sentence2vec[9])以及文档(如采用基于注意力机制的 LSTM 模型)的表示。然后,基于文档的向量表示,计算文档之间的距离,由此定义文档的离群值。本文借鉴文献[10]的离群点定义,提出了基于文档距离的离群文档检测方法,用每个文档到它的第 k 个邻居之间的距离作为离群值,并且与由文献[6]中方法演化而来的通过计算文档到文档集中心的距离来衡量文档离群值的方法进行对比。本文采用了不同的词嵌入模型和主题模型来表示文档,使用两种基于最优输运的文档距离度量方法:词搬动距离(Word Mover's Distance,WMD)[11]和层次型最优主题输运文档(Hierarchical Optimal Topic Transport,HOTT)[12]距离。由于本文方法是通过文档到 k 近邻文档的距离来衡量文档的离群程度,因此可以发现从局部视角考量认为是离群点的文档;而采用最优输运方法计算的文档距离则能体现文档的语义差异,适于检测离群文档。最后,通过在两个开源文档数据集上进行细致对比实验,检验了本文新提出方法的有效性和 k 值选取对检测结果的影响。

2 离群文档检测基准方法

本文重点关注离群文档的检测方法,因此不再赘述针对其他数据类型的离群检测方法,这些内容可参考文献[1]或关于深度学习的异常检测方法综述。下文主要介绍文献[5]中提出的离群文档检测方法,我们也将在实验部分将这两种离群文档检测方法作为基准方法。

2.1 基于余弦相似度的方法

首先将每个文档表示为一个向量,然后将每个文档和文档集的平均余弦相似度(COS)的负数作为该文档的离群值(度)。文献[5]采用了两种将文档向量化的方法,即 TF-IDF 和 paragraph2vec[13],我们将这两种方法分别命名为 TFIDF-COS 和 P2V-COS,但 TFIDF-COS 的性能比 P2V-COS 差,因此我们在实验部分只报告了 P2V-COS 方法的检测结果。

2.2 KL 散度相似度

将每个文档表示为一个概率分布,同样将整个文档集表示为另一个概率分布,然后将每个文档和整个文档集的 KL 散度作为该文档的离群值。文献[5]同样采用了两种方法来计算概率分布:1)分别对每个文档和整个文档集估计 unigram 的分布(UNI-KL);2)首先在整个文档集上训练 LDA 模型,由此得到每个文档和整个文档集的主题分布,本文实验部分将该方法命名为 LDA-KL-C。

这两类基于文本表示和距离相近性的离群文档检测方法都使用文档与整个文档集的距离来衡量离群性,无法发现基于局部考量的离群文档;而采用欧几里德距离(EUC)或主题直接的 KL 散度可能无法刻画出文档间的语义相近性。下文我们将引入基于最优输运的离群文档检测方法,实验部分将证实本文方法相对于上述基准方法更具优势。

3 基于最优输运和 k-近邻的离群文档检测方法

本文采用基于距离的离群文档检测方法[1,10,14],先给出文档离群值的定义和计算流程,然后给出两种基于最优输运的文档距离定义。

如相关工作所述,现有基于距离的离群文档检测基准方法一般采用单个文档与整个文档集的距离来衡量文档的离群程度,这类全局标准会将从局部来看正常的文档误认为是离群文档从而导致假阳性,也可能无法检测出与周围邻居文档差异较大的离群文档。为避免这些缺陷,本文提出了一种新的基于 k 近邻的离群文档检测方法。

本文采用局部视角来衡量文档的离群值,如果一个文档 d 离它的邻居较远,那么属于离群点的可能性较高。一种可选的定义为,首先计算文档 d 在某个半径 r 范围内的邻居个数,如果其小于某个阈值 k,则认为该文档 d 为离群文档。但该定义中的半径 r 较难确定,而且无法按离群值对文档进行排名[10]。本文采用文献[10]提出的基于 k 近邻的离群值定义,即如果某个数据点离它的 k 近邻越远,则离群值越大。

具体地,给定文档集 D 和文档距离函数 $dist(\cdot,\cdot)$,用户指定的整数 k,对每个文档 d_i 做 k 近邻查询得到 k 近邻集合 $N_k(d_i)$,以文档 d_i 到它的第 k 个邻居文档的距离 $dist(d_i, N_k(d_i))$ 作为文档 d_i 的离群值——该距离越远,则该文档为离群文档的可能性就越高。在返回离群文档结果时,我们将该离群值排序,返回前 n(或用户定义的百分比 p)个离群文档,整个计算流程如算法 1 所示。

算法 1 基于最优输运的异常文档检测算法

输入:文档集 D,文档距离函数 dist(·,·),整数 k,n

输出:离群文档集 O

```
1. for i=1 to |D|
2.    for j=i to |D|
  //采用第 3.1 节或第 3.2 节中的方法计算文档之间的最优输运距离
3.    计算 dist(d_i,d_j)
  //计算每个文档 d 的离群打分 o(d)
```

//以下$N_k(d)$表示文档 d 的 k 近邻集合

4. for i=1 to |D|

5. 采用最小优先级队列计算$N_k(d_i)$

6. $o(d_i)\leftarrow dist(d_i,N_k(d_i))$

7. 以 o(d)作为文档 d 的优先级，采用最大优先级队列返回前 n 个文档构成的集合 O.

在应用以上离群文档检测方法时，我们需要确定合适的文档表示方法以及相应的文档间距离定义。采用标准的词袋(BoW)模型以及欧几里德距离或余弦距离可能无法体现文档的语义相近性，因此本文采用两种较新的基于最优输运的文档距离来定义，一种是基于词嵌入向量这一语义信息，在文档之间使用词搬动距离来进行距离度量；另一种通过 LDA 主题模型来对文本进行层级抽象，先通过最优输运算法算出主题之间的距离，而文本表示为主题的分布，再使用最优输运算法算出文档之间的距离。具体方法将在第 3.1 节和第 3.2 节中介绍。

3.1 词搬动距离

最优输运(Optimal Transport，OT)已成为机器学习领域的一个重要工具，相关的介绍可参考文献[15]，本文主要关注基于最优输运计算文档之间的距离。给定两个文档 d_i 和 d_j，它们之间的词搬动距离 WMD 定义为将一个文档转换为另一个文档所需的最小费用[11]，如图 1 所示，将文档 1 转换为文档 2。

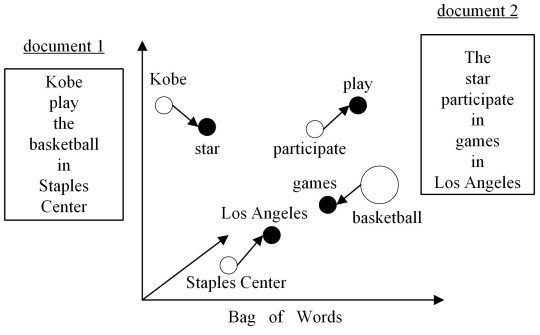


图 1 词搬动距离

Fig. 1 Word mover's distance

具体的实现过程是，首先获取预训练或基于文档集学习的词向量表示，将每个文档表示为 BoW，定义词之间的搬动费用 $c(w_a,w_b)$，然后将词搬动距离 WMD 定义为 1-Wasserstein 距离，即求解以下优化问题：

$$WMD(d_i,d_j)=\min_{\Gamma\geqslant 0}\sum_{w_a\in d_i,w_b\in d_j}c(w_a,w_b)\Gamma_{ab} \tag{1}$$

$$\text{s.t.}\ \sum_{w_b}\Gamma_{ab}=p(w_a|d_i),\forall w_a\in d_i \tag{2}$$

and

$$\sum_{w_a}\Gamma_{ab}=p(w_b|d_j),\forall w_b\in d_j \tag{3}$$

其中，$p(w_a|d_i)$为词w_a在文档d_i中的经验概率，即词在文档中的词频与文档长度的比值 $p(w_a|d_i)=\frac{n_{w_a}}{n_{d_i}}$。按照文献[11]中的 WMD 计算方法，我们也将词与词之间的搬动费用$c(w_a,w_b)$定义为这两个词向量之间的欧氏距离$\|w_a-w_b\|_2$。以上优化问题可采用 Sinkhorn 算法来进行快速求解[16]，文献[11]探索了 WMD 在文本分类中的应用，文献[17]在 WMD 算法的基础上充分挖掘了文本语义中有价值的特征项，并结合知识词典中的语言学知识构架和句法依存关系，提出了用于文本相似性研究的改进算法。

3.2 层次型最优主题输运文档距离

本文还采用了文献[12]提出的层次型最优主题输运文档(HOTT)距离，该距离的定义结合了词向量和主题模型，通过计算词向量之间的距离和主题之间的距离来计算文档间的距离，在计算速度和可解释性方面都优于词搬动距离。给定一个文档集 D，我们首先训练 LDA 模型[7]得到每个主题 z 的词分布 $p(w|z)$，以及每个文档 d 的主题分布 $p(z|d)$。对于两个主题 z_i 和 z_j，它们之间的主题搬动距离(Topic Mover's Distance，TMD)定义为 1-Wasserstein 距离：

$$TMD(z_i,z_j)=W_1(p(\cdot|z_i),p(\cdot|z_j))=\min_{\Gamma\geqslant 0}\sum_{(w_a,w_b)}c(w_a,w_b)\Gamma_{ab} \tag{4}$$

$$\text{s.t.}\ \sum_{w_b}\Gamma_{ab}=p(w_a|z_i),\forall w_a \tag{5}$$

and

$$\sum_{w_a}\Gamma_{ab}=p(w_b|z_j),\forall w_b \tag{6}$$

这里词之间的搬动费用函数还是采用L_2距离$\|w_a-w_b\|_2$。然后，文献[12]将两个文档之间的 HOTT 距离定义为它们的主题分布之间的 1-Wasserstein 距离：

$$HOTT(d_i,d_j)=W_1(\sum_{z=1}^{T}p(z|d_i)\delta_z,\sum_{z=1}^{T}p(z|d_j)\delta_z) \tag{7}$$

这里在计算文档之间的距离时应用了预先计算好的主题之间的距离，由此 HOTT 距离利用了文本的层次信息。

4 实验

本节首先介绍了两个开源文档数据集以及离群文档的生成方法，随后介绍了我们对比的几个离群文档检测基准方法以及各种方法的具体实现和参数设置，最后基于合适的评估指标进行对比实验，并对实验结果进行分析。

4.1 数据集

为了便于复现本文方法，在对比实验中采用了两个开源文本数据集 CLASSIC[1)] 和 REUTERS8[2)]，这两个文档集的统计信息如表 1 所列，Avg.|D| 为每个标签包含的文档数量，Avg.|d| 为每篇文档的平均长度，|V| 为整个文档集的词汇量。

表 1 数据集统计信息

Table 1 Dataset statistics

Dataset	Avg.|D|	Avg.|d|	|V|
CLASSIC	1769.5	42.0	15640
REUTERS8	959.25	35.3	15950

4.1.1 CLASSIC 数据集

实验将直接使用文献[11]中预处理后的版本(共包含 7075 篇文档和 4 类标签)，并将同一标签下的所有文档单独组成一个文档集，总共得到 4 个文档集$\mathbb{D}=\{D_0,D_1,D_2,D_3\}$。

1) http://www.dataminingresearch.com/index.php/2010/09/classic3-classic4-datasets

2 http://www.daviddlewis.com/resources/testcollections/reuters21578/

4.1.2 REUTERS8 数据集

实验将使用文献[18]中预处理好的版本(共包括7674篇文档和8类标签)。采用与上述CLASSIC数据集相同的处理方法后,得到8个文档集。需要注意的是,两个数据集的标签是数据集中固有的统计信息,本文将其用于构造"离群文档"以及作为实验各项指标的测试依据。文档的主题则是通过主题模型在"正常文档"和"离群文档"上进行无监督训练学习得到的。

因为这两个语料集缺少离群文档标注信息,本文按照文献[5]中采用的方法采样生成用于检测离群文档的语料库。对于每一个数据集的整个文档集$\mathbb{D}$,先采样一个标签对应的整个文档集$D_s \in \mathbb{D}$,并将D_s中所有文档标记为"正常文档",然后随机采样另一个文档集$D_s' \in \mathbb{D}$,$D_s' \neq D_s$,并随机选取D_s'中的m篇文档($m \ll |D_s|$),并将其标记为"离群文档",本文设置离群文档数量是正常文档的0.01倍。在构造离群文档时,选用一类标签下的所有文档作为正常文档,并随机选用另一类标签下极少量的文档作为离群文档,两类文档混合后,后者所占比例远远低于前者,又因为主题标签不同,所以产生"离群性"。其中,将上述"正常文档"与"离群文档"结合,得到一个包含小部分离群文档的数据集。

对于CLASSIC数据集,重复上述过程6次,得到了6个包含离群文档的语料库$\mathbb{C}_{\text{CLASSIC}} = \{C^{(1)}_{\text{CLASSIC}}, \cdots, C^{(6)}_{\text{CLASSIC}}\}$;对于REUTERS8数据集,重复上述过程10次,得到了10个包含离群文档的语料库$\mathbb{C}_{\text{REUTERS8}} = \{C^{(1)}_{\text{REUTERS8}}, \cdots, C^{(10)}_{\text{REUTERS8}}\}$。本文的离群文档检测任务将在按照上文构造方法产生的数据集上分别进行测试,并取测试后的结果的均值作为对应数据集下的实验评估结果。

4.2 实验和参数设置

本文主要考察了基于文本距离的离群文档检测方法,需要确定文本表示方法和文本距离度量函数。

4.2.1 文本表示

本文采用归一化的词袋模型(NBoW)、paragraph2vec与LDA主题模型这3种文本表示方法进行实验。本文直接使用文献[19]在English Wikipedia Articles数据集上预训练好的P2V模型,对本文所用语料库中的文档进行推断,得到所处理文本的向量表示。对于LDA表示方法,实验对每一个$C^{(i)} \in \boldsymbol{C}$进行LDA建模,采用吉布斯采样进行参数估计与推断,得到$C^{(i)}$中每篇文档的主题分布,并将该分布作为该文档在$C^{(i)}$下的LDA表示。

4.2.2 文本距离表示

对于文本距离度量,我们选用了欧氏距离、KL散度、余弦距离,并与最优输运距离WMD和HOTT进行对比实验。由于NBoW下的文本表示过于稀疏,因此不选取KL散度作为距离度量。基于NBoW文本表示计算基于最优输运距离的WMD[11]时,将词与词之间的费用函数定义为对应词向量的欧氏距离,词向量分别使用Google News数据集上预训练好的Word2vec[1)]和文献[20]提供的预训练好的300维版本GloVe[2)]。在计算文档之间的WMD距离[11]和HOTT距离[12]时,本文采用开源库POT(Python Optimal Transport)[3)]来计算最优输运距离。

4.2.3 离群文档定义

为了检验本文提出的离群文档检测方法(实验结果中记为NBoW-WMD和LDA-HOTT),我们根据离群文档的不同定义设计了两组实验与基准方法进行对比。

(1)基于最近邻的方法(局部)。基于文档的各种表示(NBoW,P2V,LDA),计算文档之间的距离(欧氏距离EUC、KL散度、余弦距离COS以及本文采用的基于最优输运OT的距离),将每个文档与它的k近邻之间的距离作为该文档的离群值。

(2)基于中心表示的方法(全局)。参考文献[6]的离群句子检测方法,基于文档的各种表示,计算整个文档集的表示(中心点即平均值),将每个文档与中心文档之间的距离作为该文档的离群值。

4.2.4 参数设定

参照文献[5]中参数的设定,将生成语料库的离群文档个数设定为$m = 0.01|D_s|$,即对于每一个用于检测离群文档的语料库,设置正常文档的数量是离群文档的100倍。

对于超参数k近邻中k的取值和主题数T,我们将在第4.4.3节中进行敏感性分析。

4.3 估指标

本文的离群文档检测尽管是无监督学习任务,但也存在正常文档和离群文档类别占比不平衡的问题,传统的度量指标(如$F1$ score)受类别不平衡性的影响,不适于作为这里的评估指标。本文使用文献[21]建议的3种比较适合评估离群文档检测方法性能的指标:MAP(Mean Average Precision),AUC(Area Under Curve)和Precision@N。为了进一步反映各种检测方法的查全率,本文还展示了不同方法在Recall@N这一度量上的表现。

在报告具体的实验结果时,将Precision@N中的N设定为当前语料库文档数量的0.005倍,记为Precision@.5%,将Recall@N中的N设定为当前语料库文档数量的0.02倍,记为Recall@2%。

4.4 实验结果

本小节将分别报告各种离群文档检测方法的性能评估结果、离群值定义的对比结果和参数敏感性分析。

4.4.1 离群文档检测方法性能评估

我们首先考察基于最近邻的局部离群文档检测方法的性能差异,评估不同文本表示下各种距离度量函数对离群文档检测效果的影响,在CLASSIC与REUTERS8数据集上得到的结果如表2所列。可以发现:

(1)在REUTERS8数据集上,LDA-HOTT方法的实验

1) http://code.google.com/archive/p/word2vec/

2) http://nlp.stanford.edu/data/glove.6B.zip

3) http://github.com/rflamary/POT

结果优于其他方法，包括基于词搬动距离的 NBoW-WMD 方法，可能因为在这些文档集上学习得到的主题分布包含更多的文本语义信息，因此 HOTT 能更好地体现文档之间的距离。对于 LDA-HOTT 而言，采用 GloVe 词向量嵌入方法时，检测性能优于基于 Word2vec 的词向量嵌入方法。

(2)在 CLASSIC 数据集上，采用词袋表示的 NBoW-WMD 方法优于 NBoW-EUC 和 NBoW-COS，这说明词搬动距离比欧氏距离和余弦距离能更好地表示文本间的距离。而采用 LDA 表示的 LDA-HOTT 优于 LDA-EUC，LDA-KL 和 LDA-COS，说明了层次型最优主题输运距离比欧氏距离、KL 散度和余弦距离能更好地表示文本间的距离。而 LDA-HOTT 只是在 MAP 上优于 NBoW-WMD，可能是因为对于这个文档集中的文档而言，词频本身已足够体现文本的整体结构信息，不再需要文本的语义信息。总体而言，两种基于最优输运距离(WMD 和 HOTT)在不同数据集和文本表示下都表现出了最好的检测效果，显示了最优输运距离能体现文档之间的语义差异，适用于基于距离的离群文档检测任务。

表 2 基于不同距离的离群文档检测性能对比

Table 2 Outlier documents detection performance comparison based on different distances

数据集	方法	MAP	AUC	Pre	Rec
CLASSIC	NBoW-EUC	0.231	0.647	0.333	0.214
	NBoW-COS	0.525	0.962	0.595	0.669
	NBoW-WMD	**0.633**	**0.979**	**0.804**	**0.737**
	LDA-EUC	0.582	0.876	0.667	0.611
	LDA-KL	0.431	0.860	0.501	0.530
	LDA-COS	0.632	0.910	0.643	0.673
	LDA-HOTT (GloVe)	**0.638**	**0.935**	**0.667**	**0.697**
	LDA-HOTT (Word2vec)	0.630	0.880	0.667	0.655
	P2V-EUC	0.118	0.564	0.202	0.148
	P2V-COS	0.455	0.852	0.513	0.574
REUTERS8	NBoW-EUC	0.015	0.476	0.026	0.056
	NBoW-COS	0.348	0.933	0.486	0.567
	NBoW-WMD	**0.391**	**0.962**	**0.515**	**0.576**
	LDA-EUC	0.429	0.946	0.405	0.439
	LDA-KL	0.183	0.908	0.242	0.396
	LDA-COS	0.430	0.938	0.568	0.525
	LDA-HOTT (GloVe)	**0.584**	**0.964**	**0.692**	**0.653**
	LDA-HOTT (Word2vec)	0.527	0.949	0.634	0.618
	P2V-EUC	0.053	0.806	0.099	0.104
	P2V-COS	0.327	0.951	0.429	0.469

4.4.2 离群值定义对比

本节考察了第 4.2 节提到的基于距离的局部和全局两种视角下两种不同离群文档定义的检测方法的性能差异，即基于到 k 近邻的距离和到中心点的距离来衡量文档的离群程度。如前文所述，参照文献[6]，本文将中心点定义为文档集中所有文档的向量表示的平均值，一篇文档与文档集中心点的距离越大，离群值就越高。表 3 列出了在 REUTERS8 数据集上，基于中心点距离检测方法的性能，表 3 中各种方法带后缀 C，表示基于到中心点的距离。与表 2 所列的在 REUTERS8 数据集上基于 k 近邻距离检测结果进行对比，可以发现：

(1)本文采用的基于 k 近邻距离的检测方法在性能上优于基于到中心点距离的全局检测方法。

(2)本文使用的最优输运距离在两种离群文档定义下都表现出了最优的检测效果，表明本文提出的最优输运距离受离群值定义的影响较小，能够很好地表示文本间的距离。

表 3 基于不同离群文档定义检测方法的性能对比

Table 3 Outlier document detection performance comparison based on different definition

定义	方法	MAP	AUC	Pre	Rec
基于中心点距离	NBoW-EUC-C	0.011	0.385	0.005	0.022
	NBoW-COS-C	0.204	0.916	0.327	0.401
	NBoW-WMD-C	**0.307**	**0.925**	**0.433**	**0.430**
	LDA-EUC-C	0.061	0.673	0.113	0.091
	LDA-KL-C	0.271	0.894	0.413	0.403
	LDA-COS-C	0.329	0.926	0.446	0.479
	LDA-HOTT-C	**0.446**	**0.932**	**0.543**	**0.596**
	P2V-EUC-C	0.037	0.734	0.042	0.086
	P2V-COS-C	0.298	0.946	0.389	0.427

4.4.3 参数敏感性分析

在前文的 k 近邻查询实验中，我们将邻居个数 k 值设定为 $k=0.015|C|$。很多采用 k 近邻的方法在衡量 k 的变化对性能的影响时，一般选择与数据集大小无关的设定，如 $k=10,20,\cdots$，但本文选择了 k 与数据集大小相关的设定。因为本文选择语料库中 0.01 倍的文档作为离群文档，同理类推，离群文档到 $k\approx0.01$ 左右的近邻还是离群文档。表 4 列出了 k 取不同的 5 个值时，对比的离群文档检测方法的性能。可以发现，随着 k 值的增大，离群文档检测方法的性能先提升后下降。这与我们的实验设想是吻合的，当 k 的取值过小时，很多离群文档会被误认为正常(假阴性)。但当 k 值过大时，即使是正常的文档也会被误认为离群文档(假阳性)。

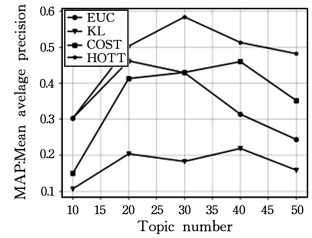
表 4 不同方法的 k 值敏感性对比

Table 4 Sensitivity performance comparison based on different methods and k value

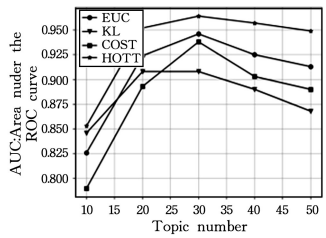
k 值	方法	MAP	AUC	Pre	Rec
$k=0.005$	LDA-EUC	0.096	0.720	0.155	0.152
	LDA-KL	0.047	0.778	0.595	0.669
	LDA-COS	0.028	0.584	0.047	0.062
	LDA-OT	0.094	0.777	0.121	0.158
$k=0.01$	LDA-EUC	0.432	0.941	0.561	0.528
	LDA-KL	0.129	0.886	0.171	0.225
	LDA-COS	0.261	0.882	0.342	0.363
	LDA-OT	0.490	0.958	0.569	0.605
$k=0.015$	LDA-EUC	0.429	0.946	0.568	0.557
	LDA-KL	0.182	0.908	0.225	0.330
	LDA-COS	0.430	0.938	0.555	0.504
	LDA-OT	**0.584**	**0.964**	**0.692**	**0.653**
$k=0.02$	LDA-EUC	0.351	0.926	0.486	0.513
	LDA-KL	0.248	0.909	0.332	0.399
	LDA-COS	0.458	0.931	0.532	0.556
	LDA-OT	0.563	0.953	0.689	0.648
$k=0.025$	LDA-EUC	0.290	0.917	0.433	0.450
	LDA-KL	0.253	0.917	0.337	0.500
	LDA-COS	0.421	0.932	0.489	0.542
	LDA-OT	0.567	0.953	0.684	0.642

本文还对基于不同主题个数的敏感性进行了分析。基于文本的 LDA 表示的方法在 REUTERS8 数据集上的检测性能优于其他方法，其中设置的主题个数 $T=30$，接下来将对评

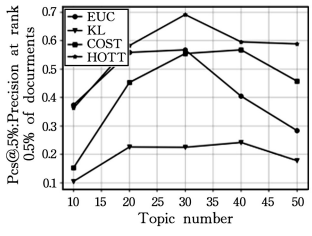
估指标受主题个数(超参数)的影响进行敏感性分析。图 2 给出了在 REUTERS8 数据集上运行各种基于距离度量的检测方法时,各个评估指标随不同主题个数的变化。



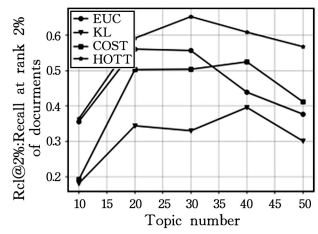
(a)MAP



(b)AUC



(c)Pre@.5%



(d)Rec@2%

图 2 不同主题数量的敏感性分析

Fig. 2 Sensitivity analysis based on different topic numbers

结束语 本文提出了基于最优输运和 k 近邻的离群文档检测方法,采用了定义在文本的词袋表示上的词搬动距离(WMD),并结合词向量和主题分布的层次型最优主题输运(HOTT)距离来表示文本的距离,将文档与它的 k 近邻之间的距离作为该文档的离群值,并验证了 k 值对离群文档检测结果的影响。本文在两个开源文本数据集上进行了对比实验,显示了本文方法优于其他基于最近邻或全局的离群文档检测基准方法。后续我们将如文献[22]、文献[23]一样,考虑把词嵌入和主题模型相结合,以更好地表示文本,结合最优输运距离进行离群文档检测工作。

参 考 文 献

[1] AGGARWALC C. Outlier Analysis[M]. Springer Publishing Company, Incorporated, 2015: 237-263.

[2] LI C J, ZHAO S N, CHI Y X. Outlier Detection Algorithm Based on Spectral Embedding and Local Density[J]. Computer Science, 2019, 46(3): 260-266.

[3] GOU J, MA Z T, ZHANG Z C. PODKNN: A Parallel Outlier Detection Algorithm for Large Dataset[J]. Computer Science, 2016, 43(7): 251-274.

[4] YIN N, ZHANG L. Research on Application of Outlier Mining Based on Hybrid Clustering Algorithm in Anomaly Detection [J]. Computer Science, 2017, 44 (5): 116-119.

[5] ZHUANG H, WANG C, TAO F, et al. Identifying semantically deviating outlier documents[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 2748-2757.

[6] LARSON S, MAHENDRAN A, LEE A, et al. Outlier Detection for Improved Data Quality and Diversity in Dialog Systems [C]// NAACL. 2019: 517-527.

[7] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(Jan.): 993-1022.

[8] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Advances in Neural Information Processing Systems. 2013: 3111-3119.

[9] ARORA S, LIANG Y, MA T. A simple but tough-to-beat baseline for sentence embeddings[C]// 5th International Conference on Learning Representations, ICLR. 2017.

[10] RAMASWAMY S, RASTOGI R, SHIM K. Efficient algorithms for mining outliers from large data sets[C]// Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. 2000: 427-438.

[11] KUSNER M, SUN Y, KOLKIN N, et al. From word embeddings to document distances[C]// International Conference on Machine Learning. 2015: 957-966.

[12] YUROCHKIN M, CLAICI S, CHIEN E, et al. Hierarchical optimal transport for document representation[C]// Advances in Neural Information Processing Systems, 2019: 1599-1609.

[13] LE Q, MIKOLOV T. Distributed representations of sentences and documents [C] // International Conference on Machine Learning. 2014: 1188-1196.

[14] ROUSSEEUW P J, DRIESSEN K V. A fast algorithm for the minimum covariance determinant estimator[J]. Technometrics, 1999, 41(3): 212-223.

[15] PEYRÉ G, CUTURI M. Computational optimal transport[J]. Foundations and Trends in Machine Learning, 2019, 11(5/6): 355-607.

[16] CUTURI M. Sinkhorn distances: Lightspeed computation of optimal transport[C]// Advances in Neural Information Processing Systems. 2013: 2292-2300.

[17] XU X X. Research on Text Similarity Algorithm Based on WMD Distance[D]. Taiyuan: Taiyuan University of Technology, 2019.

[18] JESUS A M D, CACHOPO C. Improving methods for single-label text categorization[D]. Instituto Superior Técnico, Portugal, 2007: 1-141.

[19] LAU J H, BALDWIN T. An Empirical Evaluation of doc2vec

with Practical Insights into Document Embedding Generation [C]// Proceedings of the 1st Workshop on Representation Learning for NLP. 2016:78-86.

[20] PENNINGTON J,SOCHER R,MANNING C D. Glove:Global vectors for word representation[C]// Conference on Empirical Methods in Natural Language Processing. 2014:1532-1543.

[21] CAMPOS G O,ZIMEK A,SANDER J,et al. On the evaluation of unsupervised outlier detection:measures,datasets,and an empirical study[J]. Data Mining and Knowledge Discovery,2016,30(4):891-927.

[22] XU H,WANG W,LIU W,et al. Distilled wasserstein learning for word embedding and topic modeling[C]// Advances in Neural Information Processing Systems. 2018:1716-1725.

[23] DIENG A B,RUIZ F J R,BLEI D M. Topic modeling in embedding spaces[J]. arXiv:1907.04907,2019.

SHUI Ze-nong, born in 1994, postgraduate. His main research interests include natural language processing, data mining and software engineering.

SHA Chao-feng, born in 1976, Ph.D, associate professor. His main research interests include machine learning and data mining, natural language processing.