计算机科学
COMPUTER SCIENCE

STK:基于对比学习嵌入的聚类方法

刘晋霞，张曦

引用本文

刘晋霞，张曦. STK:基于对比学习嵌入的聚类方法[J]. 计算机科学，2024，51(11A)：240400011-6.

LIU Jinxia, ZHANG Xi. STK:Clustering Method Based on Contrastive Learning Embedding[J]. Computer Science, 2024, 51(11A): 240400011-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240400011

STK:基于对比学习嵌入的聚类方法

刘晋霞　张　曦
太原科技大学经济与管理学院　太原 030024
(liujinxia@tyust.edu.cn)

摘　要　SimCSE作为一种对比学习方法,在文本嵌入和聚类中表现出了良好的性能。文中旨在优化SimCSE训练模型生成的句子嵌入使其适用于聚类任务,通过多个算法组合和训练参数调整,解决聚类算法选择、噪声及异常值的影响等问题。文中提出一种联合KL散度和KMeans算法的无监督聚类模型STK(SimCSE t-SNE KMeans),使用SimCSE对文本进行编码;随后采用t-SNE算法对高维嵌入进行降维,通过最小化KL散度保留低维空间中高维数据点之间的相似性关系,降维的同时改善文本嵌入表示;最后使用KMeans算法对降维后的嵌入进行聚类,得到聚类结果。通过将本研究的聚类结果与Bert,UMAP,HDBSCAN等算法得到的结果进行比较,发现文中提出的模型在制氢领域专利和论文数据集上表现出更好的聚类效果,尤其在轮廓系数这一评价指标上。

关键词:SimCSE;句嵌入;KL散度;聚类;轮廓系数
中图分类号　TP391

STK:Clustering Method Based on Contrastive Learning Embedding

LIU Jinxia and ZHANG Xi
School of Economics and Management,Taiyuan University of Science and Technology,Taiyuan 030024,China

Abstract　SimCSE,as a contrastive learning method,has shown good performance in text embedding and clustering. The aim of this paper is to optimize the sentence embedding generated by SimCSE training models to make them suitable for clustering tasks. By combining multiple algorithms and adjusting training parameters,the problems of clustering algorithm selection,noise,and outliers can be solved. This paper proposes an unsupervised clustering model SimCSE t-SNE KMeans(STK) that combines KL divergence and K-Means algorithm. SimCSE is used to encode the text,and then the t-SNE algorithm is used to reduce the dimensionality of high-dimensional embeddings. By minimizing KL divergence and preserving the similarity relationship between high-dimensional data points in low dimensional space,the dimensionality is reduced while improving the text embedding representation. Finally,the KMeans algorithm is used to cluster the reduced embeddings and obtain clustering results. By comparing the clustering results of this study with those obtained by algorithms such as Bert,UMAP,HDBSCAN,etc.,it is found that the model proposed in the paper showed better clustering performance in the field of hydrogen productionpatent and paper datasets,especially in the evaluation index of Silhouette coefficient.

Keywords　SimCSE,Sentence embedding,KL divergence,Clustering,Silhouette coefficient

1　引言

随着信息时代的到来,海量文本数据的持续涌现为我们提供了丰富的信息资源,然而,有效地处理和理解这些数据仍然是一个重大挑战。为了解决这一问题,自然语言处理领域的研究者们在句嵌入(Sentence Embeddings)、降维和聚类等技术上进行了深入的研究,旨在从多维、高度抽象的文本数据中提取有意义的模式和结构[1]。

句嵌入技术的崛起为文本表示学习提供了一种强大的方式,它能够将句子映射到连续的向量空间,使得语义信息得以更紧凑的表示。相较于传统文本表示方法,句嵌入技术具有以下几方面优势。首先,句嵌入技术通过深度学习模型自动学习句子的语义表示,避免了传统手动设计特征的繁琐过程。其次,端到端学习的方式简化了建模过程,提高了模型的可扩展性和灵活性[2]。通过在大规模语料库上进行训练,句嵌入方法能够学到更通用、更具泛化性的语义表示,使得模型在不同任务和领域中都能表现出色。最后,句嵌入方法的可迁移性使得模型在新任务或领域上能够更好地适应新的数据。这些优势,使得句嵌入技术成为自然语言处理任务中强有力的工具。

随着数据规模的增大,高维度的句嵌入向量可能导致计算复杂度增加,而这正是降维技术所致力于解决的问题。通过将高维向量映射到低维空间,能够保留主要的语义信息,同时减少计算负担[3]。常用的降维技术有t-SNE和UMAP等,这些降维技术在聚类任务应用中各有千秋。其中,t-SNE是一种非线性降维技术,它可以保留数据中的局部结构,并在低维空间中保持相似度关系,通常用于数据可视化,尤其是在二维或三维空间中展示高维数据的分布,使得数据的结构更容易被理解[4];UMAP是一种流形学习技术,用于降维和数

基金项目:教改项目(JG2023092)
This work was supported by the Education Reform Projects(JG2023092).
通信作者:张曦(efdcad@163.com)

据可视化，相比于其他降维方法，其更注重于保留数据的全局结构，而不仅仅是局部结构。针对本文数据的特性和可视化需求，更倾向于选择 t-SNE 作为降维方法，实际实验结果也证明 t-SNE 比 UMAP 更加合适[5]。

在嵌入和降维之后，聚类技术作为一种无监督学习方法，能够发现文本数据中的内在结构，将相似的文本归为一类，有助于在大规模文本数据中发现潜在的主题结构，减少冗余信息，确定异常文本，为信息检索、主题分析等任务提供基础。将句嵌入、降维和聚类相结合，成为处理大规模文本数据的一种前沿方法，有望提高文本数据的组织、检索和理解效率。

对比学习方法 SimCSE(Simple Contrastive Learning of Sentence Embeddings)能将相似的句嵌入向量靠近，不相似的句嵌入向量远离，从而学习到具有语义信息的句子表示[6]。虽然 SimCSE 在语义相似度和迁移学习等任务上取得了与先前方法更好的性能，但它没有在大规模数据集上对可拓展性进行分析，处理大规模数据集时需要降维和聚类来减少计算负担并发现文本数据中的内在结构，而降维时有破坏高维数据相似性关系导致最终结果受到影响的风险，同时噪声和异常值对聚类的影响也是不可忽视的[7]。针对以上问题，要将 SimCSE 应用于聚类任务，就需要找到一种既能减少计算负担又不破坏嵌入文本之间相似性的方法，使 SimCSE 不仅在句子表示上游刃有余，而且能在聚类任务中大显身手。

2 相关研究

SimCSE 是由 Gao 等[8-9]提出的对句子嵌入学习的对比学习模型，该方法通过使用无监督和有监督的训练信号来学习句子嵌入，适用于未标记和标记数据。无监督 SimCSE 只需输入一个句子并在对比学习框架中进行自我预测，并应用不同的隐藏丢弃掩码来优化模型参数，改进表达能力。有监督 SimCSE 利用自然语言处理数据集，将蕴含(前提-假设)对作为正例，将矛盾对以及其他批次内的实例作为负例，将 NLI 数据集中带注释的对合并到对比学习中，进一步提高了句子嵌入的对齐性和表达能力[10-11]。

SimCSE 自诞生之后，便被研究者们广泛应用于各种自然语言处理任务。大致可以分为以下两个方面。

1)对 SimCSE 本身的改进。例如 Zhang 和 Lan 提出了 S-SimCSE，他们把应用丢弃掩码的网络视为其自身的一个子网络，其期望规模由 dropout 率决定，通过推送具有不同期望的子网络来学习相同句子的相似嵌入，算法性能优于 SimCSE 算法 1% 以上。Mohanty 等提出了一种修改的无监督 SimCSE 方法学习句子嵌入，称之为 E-SimCSE，并确定 E-SimCSE 在 BERT-base，BERT-large，RobERTa-base 和 RobERTa-large 上的表现优于 unsup-SimCSE。Guo 和 Yuan 等提出一种改进的无监督句嵌入方法 SimCSE-PSER，采用 dropout 和位置嵌入扰动联合进行数据增强，引入 R-Drop 正则化方法，降低无监督 SimCSE 使用 dropout 作为数据增强方法带来的训练与预测阶段的不一致性，结果证明该方法优于其他无监督句嵌入方法[12-13]。

2)聚类与主题建模。He 和 Wu 等在短文本聚类研究中融合 SimCSE 提出 SSKU 模型，该模型采用 SBERT 对短文本进行嵌入表示，利用无监督 SimCSE 方法联合深度聚类 KMeans 算法对文本嵌入模型进行微调，改善短文本的嵌入表示，使其适于聚类[14]。Wang 等[15]以农业机器人为例提出 SimCSE-LDA，为大数据背景下的领域技术主题识别和颠覆性技术的判定提供了更科学的方法。

综上所述，目前国内外关于 SimCSE 的研究已有一定成果，大多专注于对 SimCSE 算法的改进，这些改进方法如 S-SimCSE 和 E-SimCSE 等虽然在句子表示学习方面更加如鱼得水，但对于将 SimCSE 应用于聚类任务还存在降维时数据结构被破坏、噪声和异常值影响等问题。因此，本文提出一种联合 KL 散度和 KMeans 算法的无监督聚类模型 SimCSE t-SNE KMeans，并以我国制氢领域专利和论文为数据来源，使用此模型进行聚类任务，以在降维减少计算负担的同时尽量维持高维数据点之间的相似性关系，优化嵌入结果，使其更适用于聚类任务。

3 STK 模型

3.1 模型组成

SimCSE t-SNE KMeans 模型由文本嵌入、降维、聚类 3 部分组成。1)文本嵌入，使用 SimCSE 训练模型把文本转化为向量表示。2)降维，使用 t-SNE 对嵌入文本进行降维，利用 KL 散度最小化可以帮助算法调整嵌入空间中点的位置，保持原始高维空间中文本嵌入的局部结构，使得在低维空间中相似的文本嵌入仍然接近，以在降维后使数据易于处理的同时尽量保持原始文本嵌入的相似性关系。3)聚类，在 SimCSE 的嵌入空间中，相似语义概念之间的嵌入差异降低，这使得 KMeans 对数据中的噪声和异常值更具鲁棒性，避免它们对簇中心的影响。此外，由于嵌入空间中的距离关系与语义相似性相关，因此得到的簇会更具有实际意义。图 1 给出了本模型的工作过程。

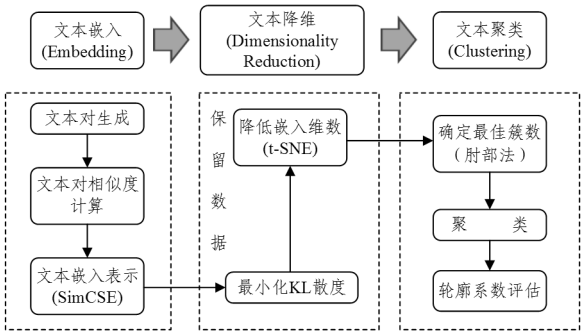


图 1 SimCSE t-SNE KMeans 模型

Fig. 1 SimCSE t-SNE KMeans model

3.2 文本嵌入方法

本文采用对比学习方法 SimCSE 训练模型作为文本嵌入方法。SimCSE 分为无监督和有监督两种训练方式[16]，此处仅介绍无监督训练方式。取一个句子集合$(x_i)_{i=1}^m$，并使$x_i^+ = x_i$，使其与相同的正对一起工作的关键因素是对x_i和x_i^+使用独立采样的 dropout 掩码。在全连接层上设置 dropout 掩码以及注意概率 p(默认 $p=0.1$)。使用$h_i^z = f_\theta(x_i, z)$表示，其中 z 是随机的丢弃掩码，将相同的输入送入编码器两次，得到两个具有 dropout 掩码 z 和z'的嵌入，即生成了一个文本对。然后计算各个文本对的余弦相似度，计算公式如下：

$$Cosine\ Similarity(h_1, h_2) = \frac{h_1 \cdot h_2}{\| h_1 \| \cdot \| h_2 \|} \tag{1}$$

得到所有文本对的余弦相似度之后，模型通过对比损失函数来最大化正样本对的余弦相似度，最小化负样本对的余弦相似度。对比损失函数如下：

$$l_i=-\log\frac{\mathrm{e}^{\frac{sim(h_i^{z_i},h_j^{z_j'})}{\tau}}}{\sum_{j=1}^{N}\mathrm{e}^{\frac{sim(h_i^{z_i},h_j^{z_j'})}{\tau}}} \tag{2}$$

其中，z 和 z' 是同一样本两次通过模型时的不同 dropout 掩码；$sim(h_1,h_2)$ 是 h_1 和 h_2 的余弦相似度；N 为训练批次内的句子数。其中参数设置与文献[17]相同。在每个训练样本的前向传播过程中，算法以 dropout 率随机丢弃一些神经元的输出，本文设置 dropout 率为 0.3[17]。这样可以降低每个神经元之间的相互依赖性，提高模型的泛化能力，减少模型过拟合的风险。因此，使用 Sim-CSE 进行文本嵌入比单纯使用 BERT 能更准确地捕获文本的语义，从而在后续的聚类工作中得到更好的结果[18]。

3.3 t-SNE 降维方法

t-SNE(t-Distributed Stochastic Neighbor Embedding)是一种用于高维数据降维和可视化的非线性降维算法。通过最小化 KL 散度来在低维空间中保留高维数据点之间的相似性关系。它在数据可视化和聚类领域非常流行，因为它可以帮助我们更好地理解数据的结构，特别是在高维数据集中[19]。

本文选择 t-SNE 作为降维方法，主要有以下几点原因：1)t-SNE 在保持高维数据的局部结构方面表现得尤为出色，对于本文数据来说，需要更加注重局部结构；2)在进行数据可视化时，t-SNE 产生了更直观和解释性更强的结果；3)通过具体的实验结果，发现 t-SNE 结合 KMeans 聚类后的聚类效果优于 UMAP。因此本文选择 t-SNE 作为降维方法[20]。

t-SNE 有 3 个参数，分别是维度(n_components)、困惑度(perplexity)、迭代次数(n_iter)。维度指定了降维后的数据应该具有多少维度；困惑度决定了每个数据点周围近邻数据点的数量[21]；而 t-SNE 的每次迭代会优化数据点在降维空间中的位置，以最大程度地保持原始数据点之间的相似性关系，过多的迭代次数会导致过拟合。本文设置 $n\_components=2$，即将数据降到 2 维空间，以便于后续的数据可视化。通过对不同参数组合的可视化，最终设置 $perplexity=30$，$n\_iter=300$[22]。

降维工作过程如下：

1)构建概率分布：t-SNE 算法使用高斯分布来建立条件概率分布，以度量数据点之间的相似性。对于每个数据点，计算其与其他数据点之间的条件概率分布。

2)构建目标分布：构建一个低维空间中基于 t 分布(t-distribution)的对称的概率分布，以便在降维后的空间中保持数据点之间的相对距离[23]。

3)最小化 KL 散度：t-SNE 的主要目标是最小化高维空间和低维空间之间的分布差异，以便保留数据点之间的相似性关系。这通过使用梯度下降算法调整低维嵌入空间中数据点的位置以最小化 KL(Kullback-Leibler)散度来实现[24]，使高维空间和低维空间中的条件概率分布尽可能相似。KL 散度是一个衡量两个概率分布之间差异的指标，其计算式如下：

$$KL(P\parallel Q)=\sum_i P(i)\log\left(\frac{P(i)}{Q(i)}\right) \tag{3}$$

其中，$P(i)$ 表示高维空间中的条件概率分布；$Q(i)$ 表示低维空间中的目标概率分布。

4)降维结果：一旦 t-SNE 算法完成了优化过程，就可以得到低维嵌入空间中每个数据点的位置，这就是最终的降维结果。

3.4 KMeans 聚类方法

KMeans 聚类是一种常用的无监督学习算法，用于将数据点分成具有相似特征的簇或群，每个簇由一个代表性的中心点(质心)表示，而每个数据点被分配到离其最近的簇[25]。以下给出算法的具体步骤。

1)初始化质心：随机选择 K 个数据点作为初始质心，或者根据某种启发式方法选择初始质心。这些初始质心将成为每个簇的中心点。本文根据实验得到的经验设置 $n\_init=10$，以确保算法有足够的机会找到最佳的聚类效果。

2)确定最佳簇数，分配数据点：对于每个数据点，计算它与每个质心的距离，通常使用欧氏距离或其他距离度量。将数据点分配到距离最近的质心所属的簇。本文通过肘部法得到最佳簇数 $n\_clustters=5$。在肘部法中，随着簇数 K 的增加，每个簇内的平均距离将逐渐减小，但当 K 变得很大时，每个簇内的平均距离将趋近于 0，因为每个数据点都可能成为自己的簇。肘部法可以找到一个 K 值，该值对应于簇内平均距离急剧下降的点，在此点之后增加 K 值将导致每个簇更小，不一定具备信息价值[26]。

利用肘部法确定最佳簇数的实验结果如图 2 所示。

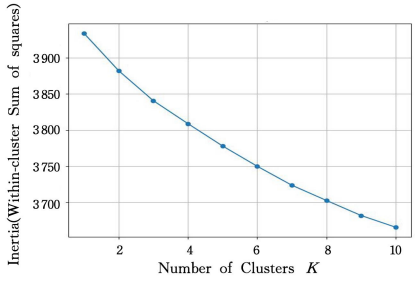


图 2 肘部法确定最佳簇数

Fig. 2 Elbow method to determine the optimal number of clusters

3)更新质心：对于每个簇，计算其所有成员的平均值，以得到新的质心位置。

4)重复步骤 2)和 3)，直到质心的位置不再发生显著变化，或者达到了预定的迭代次数。当质心不再变化或达到迭代次数后，算法终止。此时，每个数据点都被分配到一个簇中，形成 K 个簇。

综上所述，通过仔细选择模型每部分的参数和有效的训练方式，确保模型可以胜任文本聚类任务。这些训练方法的细节以及参数设置的合理性将在后续实验结果中得到验证[27]。

4 实验结果与分析

4.1 数据来源

本文数据来源于我国制氢领域的专利和论文文献。

1)专利。在专利之星检索平台使用表格检索，检索式为‘F XX(制氢/TI) * (制氢/AB) * (制氢/TX)’(TI 表示标题，AB 表示摘要，TX 表示关键词)‘F XX(20000101>20231231/AD)’(AD 表示申请日，20000101>20231231 表示时间跨度为 2000－2023 年)，共检索到 9401 篇专利文献。其中，发明专利 5878 篇，实用新型专利 3367 篇，外观专利 156 篇。为保证数据多样化，选择全部专利类型。专利的法律状态包括有效、审中、失效，在检索结果中，有效专利 4481 篇，审中专利 2616 篇，失效专利 2304 篇。审中专利不能确定是否全部有效；失效专利虽说明此项技术曾经被认可，但现在已经被淘

汰。因此，选择有效专利作为本文研究对象[28]。

2)论文。在中国知网文献检索平台使用高级检索，主题定为“制氢”，因为质量高的期刊比较有研究价值，所以文献来源限制为 SCI、EI、北大核心、CSSCI，作者和期刊均不限，发表时间限定为 2000－2023 年。

根据以上检索条件，共检索得到 4 481 篇专利文献和 4 829 篇论文文献，因为专利之星检索系统每次只能导出 15 条数据，将其批量导出后使用 Excel 中的 powerquery 插件把导出的所有数据表格合并。知网检索系统可以直接导出所有数据，每条数据中包含申请号、标题、摘要、关键词、分类号、地址等 21 项内容，提取标题、摘要、关键词 3 项内容作为本文的研究对象。

4.2 数据预处理

从专利和论文数据中提取出标题、摘要、关键词 3 项信息并合并为一列便于后续处理，删除数据中的缺失项和重复项。制氢专利领域数据在摘要中会出现一些化学方程式，这些化学方程式对于数据整体聚类无特殊作用，甚至会增加数据的噪音，对聚类结果产生负面影响，因此定义一个正则表达式来匹配常见的化学方程式，对其进行删除，最后把处理好的数据保存。

4.3 实验步骤

1)文本嵌入。处理过的数据集通过无监督 SimCSE 训练的语言模型 unsup-simcse-bert-base-uncased 获取一个 767 维向量作为文本的嵌入表示。由于数据庞大，为了提高模型处理效率，可以选择在 GPU 上运行程序或者把数据分批次处理之后再合并。本文选择分批次处理，每个批次大小设置为 32。最终将每个批次中的嵌入向量连接在一起，形成表示整个数据集的嵌入。

2)对嵌入向量进行降维。利用 t-SNE 算法将嵌入文本降维到二维空间，在此过程中设置初始概率分布为随机，最小化 KL 散度帮助模型调整嵌入空间中点的位置，避免模型陷入局部最优解。其他参数设置分别为 $n\_components=2$，$perplexity=30$，$n\_iter=300$，如此便可以保持原始高维空间中文本嵌入的局部结构，使得在低维空间中相似的文本嵌入更加靠近，以在降维后使数据易于处理的同时增强原始文本嵌入的相似性关系[29]。

3)聚类。将降维后的文本输入到 KMeans 聚类算法，最佳簇数设置为 5，初始化中心点的次数 $n\_init=10$，利用聚类结果反向优化训练模型与聚类参数，从而提高聚类准确度。

4.4 评估指标

本文研究无监督聚类模型，不提供真实标签，而常用的聚类精度 ACC 和互信息的评估方法需要提供真实标签，所以本文采用轮廓系数(Silhouette_score)作为评估指标。轮廓系数是无监督聚类算法评估的常用指标。取降维后的两维数据作为特征列，利用特征列与聚类标签列计算轮廓系数，如此便可将聚类结果在降维空间中的分布和分离程度量化为分数，从而评估聚类质量。计算公式如下：

$$Silhouette_{score}=\frac{b-a}{\max(a,b)} \tag{4}$$

其中，a 为该数据点到同一簇中其他数据点的平均距离；b 为从数据点到不同簇中的数据点的最小平均距离。

轮廓系数得分越接近 1，表示聚类效果越好；接近 0 表示该对象位于或非常接近两个相邻簇之间的决策边界；接近－1 表示该对象可能放置在错误的簇中。由于 KMeans 和 t-SNE 算法中包含随机性元素，这就导致即使对于相同的嵌入文本和运行参数，轮廓系数也会出现不同的结果，因此采取 10 次实验的平均值作为最终实验结果。

4.5 对比方法和实验结果分析

4.5.1 对比方法

为证明本文方法在文本聚类上表现出良好的性能，在降维和聚类步骤中使用其他模型进行替换。在降维中使用 UMAP 替换 t-SNE，在聚类中使用 HDBSCAN 替换 KMeans，使用替换后的方法对同一数据集进行聚类操作，通过轮廓系数和聚类图谱评估聚类质量并与本文方法进行比较。同时，在对比实验完成之后，采用消融实验的方法对是否经过 SimCSE 和 t-SNE 改善的句子嵌入进行了聚类实验，以验证这两个部分的必要性[30]。

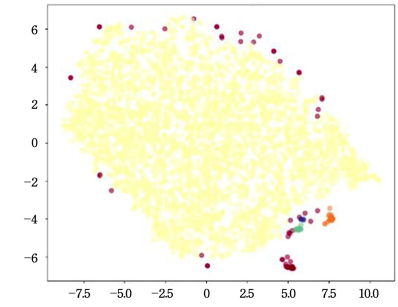
4.5.2 对比实验结果分析

本实验以我国制氢领域数据为例，使用轮廓系数作为评价指标，结果保留小数点后 3 位。在相同的数据集上做了 4 组对比实验，可视化聚类图谱展示了各方法在制氢领域专利数据集上的聚类结果。表 1 列出了轮廓系数分别在制氢领域专利数据集和论文数据集上的评估结果。实验结果如图 3 和图 4 所示。

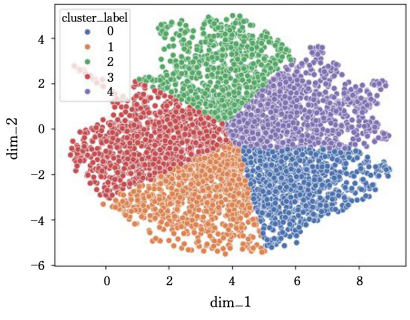
表 1 对比实验轮廓系数

Table 1 Silhouette coefficient of contrast experiments

Methods	Silhouette_score	
	Patent Dataset	Thesis dataset
SimCSE t-SNE HDBSCAN	0.007	0.012
SimCSE UMAP KMeans	0.301	0.291
SimCSE UMAP HDBSCAN	0.132	0.106
Bert t-SNE KMeans	0.343	0.332
SimCSE t-SNE KMeans	0.388	0.381



(a)SimCSE t-SNE HDBSCAN



(b)SimCSE UMAP KMeans

图 3 SimCSE t-SNE HDBSCAN 和 SimCSE UMAP KMeanss 聚类可视化结果

Fig. 3 Visualization results of SimCSE t-SNE HDBSCAN cluster and SimCSE UMAP KMeans cluster

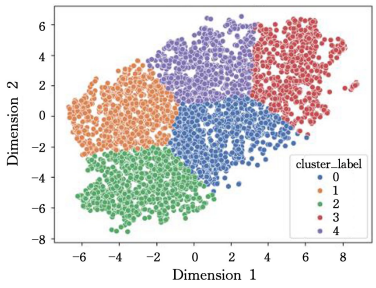


图 4　SimCSE t-SNE KMeans 聚类可视化结果

Fig. 4　Visualization results of SimCSE t-SNE KMeans cluster

1)从图 3 和图 4 可以看出，图 4 中的数据点分布更为均匀，更全面地展示了整个数据集的结构，没有过度强调某些数据区域，使数据分析更为客观。t-SNE 更注重数据的局部结构，而在制氢领域数据中，数据的局部结构往往包含了重要的反应特征和催化活性信息，这些信息至关重要，所以 t-SNE 比 UMAP 更适合本文数据。而且在表 1 的轮廓系数对比中，图 4 方法的轮廓系数要比图 3(b)中方法的轮廓系数高出百分之 22.4%，说明 t-SNE 比 UMAP 更有利于发现数据中的局部结构和簇，通过最小化 KL 散度在低维空间中保留高维数据点之间的相似性关系，在改善可视化和解决高维数据处理困难性问题的同时，使得嵌入文本更加适用于聚类任务[31]。

2)图 3(a)显示黄色数据点占据大部分面积，而其他数据点只有很小的一部分，没有表现出明显的聚类特征；图 3(b)的数据簇呈扇形分布，聚类特征明显。结合图 3(a)和图 4 的对比，在相同的嵌入和降维算法之下，Hdbscan 的表现仍然远不及 KMeans，这是因为使用 SimCSE 方法得到嵌入文本之后，KMeans 对数据中的噪声和异常值更具有鲁棒性，有效降低了它们对簇中心的影响，因此其效果比 HDBSCAN 更好。

4.5.3　消融实验结果分析

为了更好地验证本文提出的聚类框架的有效性，我们对是否经过 SimCSE 与 t-SNE 改善的句子嵌入进行了聚类实验。从消融实验中，我们可以得到以下信息。

1)从图 4 和图 5 中可以看出，缺少 SimCSE 训练的句子嵌入在聚类图谱上表现出数据点稀疏的特征，而经过 SimCSE 训练的嵌入方法得到的聚类图谱中簇与簇间的划分和边界更加明显，聚类效果更好。这说明 SimCSE 提供了更好的特征表示，使得聚类算法能够更好地识别数据中的模式和结构。从表 2 可知，STK 在制氢领域专利和论文数据集的轮廓系数上与单纯的 Bert 嵌入方法相比分别高出了 13.2% 和 14.8%，并且与其他方法相比轮廓系数更接近于 1，这说明了 STK 在制氢领域专利和论文数据集的轮廓系数上与单纯的 Bert 嵌入方法相比分别高出了 13.2% 和 14.8%，并且与其他方法相比轮廓系数更接近于 1，这说明了 STK 在制氢领域专利和论文数据集中表现出良好的性能。

2)对失去降维的数据聚类后无法将聚类结果直观地可视化在二维平面上，但仍然可以通过轮廓系数对聚类结果进行分析和评估。由表 2 可知，失去降维后的聚类结果在轮廓系数上的评估并不理想，即便是经过 SimCSE 嵌入改善了文本表示，轮廓系数仍然不及 STK。同时结合对比实验中 SimCSE UMAP KMeans 方法的轮廓系数来看，t-SNE 与 UMAP 相比，仍然是经过 t-SNE 降维后的嵌入文本在聚类任务中更胜一筹。

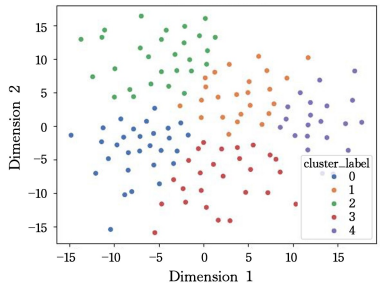


图 5　Bert t-SNE KMeans 专利数据集可视化结果

Fig. 5　Visualization results of Bert t-SNE KMeans patent dataset

表 2　消融实验轮廓系数

Table 2　Silhouette coefficient of ablation experiments

Methods	Silhouette_score	
	Patent Dataset	Thesis dataset
Bert t-SNE KMeans	0.343	0.332
SimCSE KMeans	0.097	0.101
SimCSE t-SNE KMeans	0.388	0.381

结束语　本文基于对比学习模型 SimCSE 和现有的嵌入、降维、聚类方法在制氢领域数据的聚类效果上做了不同组合的对比实验和消融实验，并提出一种在此类文本上聚类效果较好的模型。首先利用 SimCSE 训练模型进行句子嵌入，获取文本的向量表示；然后联合 KL 散度，使用非线性降维方法 t-SNE 对嵌入文本进行降维，使得到的文本嵌入更加适用于聚类；最后使用深度聚类模型 KMeans 进行聚类，得到聚类标签。得到聚类结果之后，使用轮廓系数对无监督聚类结果进行评估，分析对比实验结果，得出本文模型在轮廓系数这一评估指标上表现良好。

本文使用的是专利之星检索系统中的中文专利和论文数据，未来可以在外文数据和其他文本数据上进行深入研究，进一步评估此模型在其他语言和领域的数据集的表现。此外，KMeans 需要提前确定最佳聚类簇数，这对于聚类结果有一定的影响，本文中使用肘部法确定最佳簇数，具体是否还有其他更好的方法还有待探索。t-SNE 的结果受到不同初始化的影响，导致在不同运行中会产生不同的嵌入结果，本文采用随机初始化帮助模型避免陷入局部最优解，并且在对聚类结果进行评价时采用了取平均数的方法以减小此影响，提高评价的可靠性。未来可以根据参数分布的先验知识选择一个合适的初始概率分布，进一步优化此模型。

参 考 文 献

[1] MOHANTY I, GOYAL A, DOTTERWEICH A. Emotions are subtle: Learning sentiment based text representations using contrastive learning[J]. arXiv: 2112.01054, 2021.

[2] ZHANG J. S-SimCSE: sampled sub-networks for contrastive learning of sentence embedding[J]. arXiv: 2111.11750, 2021.

[3] RETSINAS G, STAMATOPOULOS N, LOULOUDIS G, et al. Nonlinear manifold embedding on keyword spotting using t-SNE [C]// 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017: 487-492.

[4] CAI T T, MA R. Theoretical foundations of t-sne for visualizing high-dimensional clustered data [J]. The Journal of Machine Learning Research, 2022, 23(1): 13581-13634.

[5] WANG Y, HUANG H, RUDIN C, et al. Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization[J]. The Journal of Machine Learning Research, 2021, 22(1): 9129-9201.

[6] HAO H B. Disease Knowledge Graph Q&A System Based on SimCSE [J]. Computer and Information Technology, 2023, 31(2): 97-100.

[7] WU X, GAO C, ZANG L, et al. Esimcse: Enhanced sample building method for contrastive learning of unsupervised sentence embedding[J]. arXiv: 2109.04380, 2021.

[8] GAO T, YAO X, CHEN D. Simcse: Simple contrastive learning of sentence embeddings[J]. arXiv: 2104.08821, 2021.

[9] CAO R, WANG Y, LIANG Y, et al. Exploring the impact of negative samples of contrastive learning: A case study of sentence embedding[J]. arXiv: 2202.13093, 2022.

[10] DENG J, WAN F, YANG T, et al. Clustering-Aware Negative Sampling for Unsupervised Sentence Representation[J]. arXiv: 2305.09892, 2023.

[11] GUO J H, YUAN Y C, WANG K J, et al. Unsupervised sentence embedding method based on improved SimCSE [J]. Computer Engineering and Design, 2023, 44(8): 2382-2388.

[12] ZHANG J, LAN Z, HE J. Contrastive Learning of Sentence Embeddings from Scratch[J]. arXiv: 2305.15077, 2023.

[13] HE W H, WU C J, ZHOU S J, et al. Short text clustering research using unsupervised SimCSE fusion [J]. Computer Science, 2023, 50(11): 71-76.

[14] WANG X H, WANG X, WANG S F, et al. A Disruptive Technology Identification Method Based on SimCSE-LDA and Anomaly Detection: Taking Agricultural Robots as an Example [J]. Intelligence Theory and Practice, 2023, 46(5): 135-143.

[15] MELIT DEVASSY B, GEORGE S, NUSSBAUM P. Unsupervised clustering of hyperspectral paper data using t-SNE[J]. Journal of Imaging, 2020, 6(5): 29.

[16] HE H, ZHANG J, LAN Z, et al. Instance smoothed contrastive learning for unsupervised sentence embedding[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2023: 12863-12871.

[17] POLIČAR P G, STRAŽAR M, ZUPAN B. openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding[J]. BioRxiv, 2019: 731877.

[18] GONZÁLEZ-MÁRQUEZ R, BERENS P, KOBAK D. Two-dimensional visualization of large document libraries using t-SNE [C]// ICLR 2022 Workshop on Geometrical and Topological Representation Learning. 2022.

[19] ZHOU Y, SHARPEE T O. Using global t-SNE to preserve inter-cluster data structure[J]. bioRxiv, 2018: 331611.

[20] DAMRICH S, BÖHM N, HAMPRECHT F A, et al. From t-SNE to UMAP with contrastive learning[C]// The Eleventh International Conference on Learning Representations. 2022.

[21] CAO Y, WANG L. Automatic selection of t-SNE perplexity[J]. arXiv: 1708.03229, 2017.

[22] GARE S, CHEL S, KURUBA M, et al. Dimension reduction and clustering of single cell calcium spiking: comparison of t-SNE and UMAP[C]// 2021 National Conference on Communications (NCC). IEEE, 2021: 1-6.

[23] ROBINSON I, PIERCE-HOFFMAN E. Tree-sne: Hierarchical clustering and visualization using t-sne[J]. arXiv: 2002.05687, 2020.

[24] GISBRECHT A, MOKBEL B, HAMMER B. Linear basis-function t-SNE for fast nonlinear dimensionality reduction[C]// The 2012 International Joint Conference on Neural Networks (IJCNN). IEEE, 2012: 1-8.

[25] BAJAL E, KATARA V, BHATIA M, et al. A review of clustering algorithms: comparison of DBSCAN and K-mean with oversampling and t-SNE[J]. Recent Patents on Engineering, 2022, 16(2): 17-31.

[26] LIU W, SHAO W, XIN Y. Method based on t-sne reduction and K-means clustering to identify the household-transformer relationship in low-voltage distribution network[C]// Second International Conference on Electronic Information Technology (EIT 2023). SPIE, 2023: 74-79.

[27] ZHANG D, LI S W, XIAO W, et al. Pairwise supervised contrastive learning of sentence representations[J]. arXiv: 2109.05424, 2021.

[28] LI Z Y. K-SimCSE: Research on Text Retrieval Integrating Domain Knowledge [D]. Wuhan: Central China Normal University, 2022.

[29] AI A W S. Pairwise Supervised Contrastive Learning of Sentence Representations[J]. arXiv: 2109.05424, 2021.

[30] FEI Y, NIE P, MENG Z, et al. Beyond prompting: Making pretrained language models better zero-shot learners by clustering representations[J]. arXiv: 2210.16637, 2022.

[31] POTRATZ G L, CANCHUMUNI S W A, CASTRO J D B, et al. Automatic lithofacies classification with t-SNE and K-nearest neighbors algorithm [J]. Anuário Do Instituto De Geociências, 2021, 44.

LIU Jinxia, born in 1973, Ph. D, associate professor, master's supervisor. Her main research interests include intelligent decision-making, data analysis, and innovative management.

ZHANG Xi, born in 2000, postgraduate. His main research interests include big data-driven management and decision-making.