DOI:10.20079/j.issn.1001-893x.240618002

基于可扩展子空间学习的数据流聚类方法*

尹宏伟[1,2,3],倪钰洲[1,2],胡文军[1,2,3]

(1.湖州师范学院 信息工程学院,浙江 湖州 313000;
2.浙江省现代农业资源智慧管理与应用研究重点实验室,浙江 湖州 313000;
3.湖州市水域机器人技术重点实验室,浙江 湖州 313000)

摘　要:传统数据流聚类方法缺乏对高维数据的在线降维能力,导致其聚类性能受限。为解决此问题,提出了一种基于可扩展子空间学习的数据流聚类方法(Scalable Subspace Learning for Clustering Data Streams,S^2LCStream)。首先,通过可扩展子空间学习建立历史数据与新增数据之间的投影关系,将新增数据投影至历史数据张成的子空间中,以实时获取其聚类划分。其次,为保持不同时刻聚类划分的准确性,对持续到达的数据流进行数据分布的一致性检测,捕获其中存在的概念漂移,并结合回溯机制对聚类划分进行调整以适应动态变化的数据分布。最后,通过在多个真实数据集上进行测试,验证了所提方法在处理高维数据流的效能。所提方法在保持较高聚类性能的同时,能够高效处理数据流中的概念漂移。

关键词:数据流聚类;子空间学习;可扩展子空间学习;概念漂移检测

开放科学(资源服务)标识码(OSID):

微信扫描二维码
听独家语音释文
与作者在线交流
享本刊专属服务

中图分类号:TP311.13　　**文献标志码**:A　　**文章编号**:1001-893X(2025)11-1836-08

Scalable Subspace Learning for Clustering Data Streams

YIN Hongwei[1,2,3], NI Yuzhou[1,2], HU Wenjun[1,2,3]

(1. School of Information Engineering, Huzhou University, Huzhou 31300, China;
2. Zhejiang Province Key Laboratory of Smart Management and Application of Modern Agricultural Resources, Huzhou 313000, China; 3. Huzhou Key Laboratory of Aquatic Robot Technology, Huzhou 313000, China)

Abstract: Traditional data stream clustering methods lack online dimensionality reduction capabilities for high-dimensional data, leading to limited clustering performance. To address this issue, a Scalable Subspace Learning for Clustering Data Streams (S^2LCStream) method is proposed. Firstly, this method establishes a projection relationship between historical data and new data through scalable subspace learning, projecting the new data into the subspace spanned by historical data to obtain its clustering assignment in real-time. Secondly, to maintain the accuracy of clustering assignments over time, the method performs consistency detection of data distribution on the continuously arriving data stream, capturing concept drifts and adjusting clustering assignments through a backtracking mechanism to adapt to dynamically changing data distributions. Finally, the proposed method is validated on multiple real-world datasets, demonstrating its efficiency in handling high-dimensional data streams. Specifically, S^2LCStream maintains high clustering accuracy while efficiently handling concept drift.

Key words: data stream clustering; subspace learning; scalable subspace learning; concept drift detection

* **收稿日期**:2024-06-18;**修回日期**:2024-08-05
基金项目:国家自然科学基金资助项目(62206094);湖州市公益性应用研究项目(2021GZ05);江苏省网络空间安全工程实验室开放课题(SDGC2237);湖州师范学院研究生科研创新项目(2024KYCX62)
通信作者:胡文军　　Email:hoowenjun@foxmail.com

0　引　言

随着信息技术的迅速发展,数据流正逐渐成为数据处理与分析的核心对象。作为一种持续到达且动态变化的数据对象,数据流广泛存在于交通监管、电子商务、社交媒体以及医疗诊断等多领域[1]。例如,通过分析交通数据流,可以预测交通拥堵模式并优化交通信号控制[2];在金融市场中,通过分析实时交易数据流,可以捕捉异常交易以降低投资风险[3];此外,通过分析用户行为数据流,可以构建用户画像,实现个性化内容推荐[4]。为了在资源受限且缺少监督信息的条件下,实时发现海量数据流中的潜在规律与关联,并从中提取出关键知识,无监督的数据流聚类已经成为当前机器学习与数据挖掘的重要任务之一[5-7]。

与面向封闭静态数据的传统聚类方法不同,数据流聚类面向开放动态的流式数据。开放是指其样本规模会随时间持续增加,动态是指其潜在数据分布同样会随时间产生动态变化。针对持续增加的样本规模,数据流聚类需要持续到达的新增数据进行高效的在线聚类分析。此外,针对动态变化的数据分布,又被称为概念漂移[8],数据流聚类需要具备检测并适应概念漂移的能力,从而实现对动态变化数据分布的精准描述[7-8]。

为实现在线获取数据流聚类划分,现有数据流聚类通常采用两阶段方法,又称为在线-离线双层框架[9],包括 CluStream[10] 和 DenStream[11] 等。在这些方法中,数据流的概要信息被在线生成并存储在特定的数据结构中。当新增样本到达时,数据流的概要信息会被实时更新,以保持对数据分布的准确描述。在离线聚类阶段,定期对生成的数据概要执行聚类算法。为进一步提高实时聚类性能,近年来提出了完全在线聚类,旨在对每个不断到达的数据实例进行重新聚类,以此来保持最新的聚类结果,包括 DPClust[12]、FEAC-Stream[13] 和 Adaptive Stream k-means[14] 等。

通过建立数据流中的概念漂移检测机制,能够有效提高聚类划分的准确性。DenStream[11] 和 CluStream[10] 提出通过淘汰部分历史数据,以保持在线组件中准确更新数据概要,从而适应动态变化的数据分布。但是,此类方法无法显式捕捉概念漂移。SVStream[15] 通过支持向量描述建立数据流聚类方法框架,通过迭代维护数据的最小球体,以动态维持数据流中各类簇的边界,但仍无法显式捕捉概念漂移。为实现概念漂移的显式捕捉,一种基于等密度分区的概念漂移检测方法[16] 被提出。该方法通过对数据进行等密度分区,利用卡方检验对每个分区进行统计和计算,从而检测数据分布变化,以达到概念漂移检测的目的。在文献[17]中,通过统一流形逼近与投影对演化数据流进行在线嵌入和聚类,能自适应捕捉概念漂移,从而提高聚类性能。

尽管以上方法在数据流在线聚类与概念漂移检测上取得了较为理想的结果,但由于缺乏高效的在线降维机制,其处理高维数据流的能力受到较大限制,难以捕捉高维数据流中存在的概念漂移。为解决此问题,本文提出一种基于可扩展子空间学习的新型数据流聚类方法(Scalable Subspace Learning for Clustering Data Streams,S^2LCStream)。首先,该算法通过可扩展子空间学习[18] 建立历史数据与新增数据之间的投影关系,将新增样本投影至历史数据张成的子空间中,以实时获取其聚类划分。其次,为保持不同时刻聚类划分的准确性,对持续到达的数据流进行数据分布的一致性检测,捕获其中存在概念漂移的窗口,并结合回溯机制对聚类划分进行调整以适应动态变化的数据分布。通过在多个仿真数据集以及真实数据流上的实验,验证本文所提算法在聚类性能上优于当前数据流聚类算法,并且能够高效捕捉高维数据流中存在的概念漂移。

1　提出的方法

1.1　可扩展子空间学习

在传统自表示子空间学习方法[19-21] 中,对于数据集 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}\in\mathbb{R}^{d\times n}$,通过设定数据集本体为字典,将各数据样本表示为其他样本的仿射组合,可捕获数据在低维子空间的几何结构,其目标函数如下所示:

$$\min_{\boldsymbol{Z},\boldsymbol{E}}\|\boldsymbol{Z}\|_*+\lambda\|\boldsymbol{E}\|_\ell \quad \text{s.t. } \boldsymbol{X}=\boldsymbol{X}\boldsymbol{Z}+\boldsymbol{E} \tag{1}$$

式中:$\boldsymbol{Z}\in\mathbb{R}^{n\times n}$ 为原始数据的自表示矩阵;$\boldsymbol{E}\in\mathbb{R}^{d\times n}$ 为噪声的扰动矩阵;λ 为权衡参数。根据不同的结构保持规则,$\|\cdot\|_*$ 可选择多种矩阵范数对自表示矩阵进行约束。例如,引入 ℓ_0 范数来捕获数据的局部结构,又因为 ℓ_0 范数诱发的 NP-Hard 问题,通常采用其最优凸近似的 ℓ_1 范数进行替代。引入核范数对自表示矩阵的进行低秩约束,可增强保持原始数据的全局结构,并降低异常值的影响。

在公式(1)的模型中,所获取自表示矩阵 $\boldsymbol{Z}$ 被

视为表示原始数据 $\boldsymbol{X}$ 中所有样本间相似度关系的邻接矩阵。进一步利用谱聚类等方法,可获得原始数据的聚类划分。但是,传统的自表示子空间学习只能实现对历史数据学习,无法扩展至新增数据。令 $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_l\}\in\mathbb{R}^{d\times l}$ 表示新增数据,通过建立历史数据与新增数据之间的投影关系,将样本 $\boldsymbol{y}_i$ 投影至由 $\boldsymbol{X}$ 张成的 n 维子空间中,以获得其聚类标签。可扩展子空间学习的目标函数如下所示:

$$\min_{\boldsymbol{c}_i}\|\boldsymbol{y}_i-\boldsymbol{X}\boldsymbol{c}_i\|_2^2+\gamma\|\boldsymbol{c}_i\|_2^2 \tag{2}$$

其中通过最小化 ℓ_2 范数建立历史数据和新增数据之间的投影。此外,通过对投影向量的二次约束避免过拟合,γ 为权衡参数。该优化问题的解为

$$\boldsymbol{c}_i=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+\gamma\boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}_i \tag{3}$$

为获取新增数据的聚类标签,在获得最优投影向量 $\boldsymbol{c}_i$ 后,取该向量中第 j 个非零元素为 $\delta_j(\boldsymbol{c}_i)$,则样本 $\boldsymbol{y}_i$ 与 $\boldsymbol{X}$ 张成的子空间在第 j 维度的残差定义如下:

$$r_j(\boldsymbol{y}_i)=\|\boldsymbol{y}_i-\boldsymbol{X}\delta_j(\boldsymbol{c}_i)\|_2 \tag{4}$$

通过比较样本 $\boldsymbol{y}_i$ 在各维度上的残差值,确定具有最小残差的维度为该样本的聚类标签 $f(\boldsymbol{y}_i)$,公式如下:

$$f(\boldsymbol{y}_i)=\underset{j}{\operatorname{argmin}}\{r_j(\boldsymbol{y}_i)\} \tag{5}$$

1.2 基于可扩展子空间学习的数据流聚类

对于数据流 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$,其特征维度为 d,样本数据持续到达,且样本总量趋于无穷。为实时获取新增数据的聚类划分,通过可扩展子空间学习建立历史数据和新增数据之间的投影关系。首先,采用滑动窗口 h 对数据流中不同时刻到达的数据进行表示,令 t 时刻滑动窗口内样本集合为 h_t,则第 1 时刻和第 2 时刻获取的样本集合分别为 h_1 与 h_2。根据可扩展子空间学习的原理,联合 h_1 与 h_2 构成历史数据,并通过公式(1)获得其聚类标签 $\boldsymbol{H}^*$。

当新增数据 h_t 到达时,通过公式(2)将新增数据投影至历史数据张成的子空间内,进而通过公式(5)可实时获取其聚类标签 $\boldsymbol{H}_t$。

由于数据分布会随着时间持续变化,在数据流聚类任务中,根据历史数据建立的学习模型无法适应新增数据的数据分布。为保持聚类划分的准确性,需要对持续到达的数据流进行数据分布的一致性检测,捕获其中存在的概念漂移,即概念漂移检测。为实现此目标,在数据流聚类过程中设置概念漂移检测周期 p,对周期内获取的新增数据进行两次聚类,包含通过可扩展子空间学习方法获取各窗口的聚类划分,另一次则通过传统子空间表示学习独立获取周期内所有新增数据的聚类划分。通过计算两次聚类划分结果之间的调整兰德指数(Adjusted Rand Index, ARI),判断周期内是否存在概念漂移。

如图 1 所示,以首次概念漂移检测周期为例,联合 h_1 与 h_2 窗口构成历史数据,并通过公式(1)获得其聚类标签 $\boldsymbol{H}_{1:2}$。当新增数据窗口 h_3、h_4 和 h_5 到达时,通过公式(2)将新增数据投影至历史数据张成的子空间内,进而通过公式(5)可实时获取其聚类标签 $\boldsymbol{H}_{3:5}$,最后联合获得聚类标签 $\boldsymbol{H}_{1:5}$,再联合 h_1、h_2、h_3、h_4 和 h_5 窗口并通过公式(1)获得其聚类标签 $\boldsymbol{Q}_{1:5}$。因而对于相同的数据有两个不同的聚类标签 $\boldsymbol{H}_{1:5}$ 和 $\boldsymbol{Q}_{1:5}$,其中 $\boldsymbol{Q}_{1:5}$ 是不使用任何先验知识得到的聚类标签。在获得两次的聚类标签后,计算标签 $\boldsymbol{H}_{1:5}$ 和 $\boldsymbol{Q}_{1:5}$ 之间的一致性,令 $\zeta=\mathrm{ARI}(\boldsymbol{H},\boldsymbol{Q})$ 表示标签 $\boldsymbol{H}_{1:5}$ 和 $\boldsymbol{Q}_{1:5}$ 之间的调整兰德系数,再令 θ 表示概念漂移的阈值,来进行概念漂移的决策。当标签一致性小于概念漂移的阈值,即 $\zeta<\theta$,这意味着数据特征发生了变化,存在概念漂移,此时自适应减少概念漂移检测周期 p 的大小和自适应调整概念漂移的阈值 θ,并返回初始化阶段,重新计算两次聚类标签 $\boldsymbol{H}_{1:4}$ 和 $\boldsymbol{Q}_{1:4}$ 之间的调整兰德系数,直至满足 $\zeta\geqslant\theta$。当标签一致性不小于概念漂移的阈值,即 $\zeta\geqslant\theta$,这说明输入数据的特征保持一致,未发生概念漂移,此时算法输出当前周期内的聚类结果并继续处理下一周期内的新增数据。

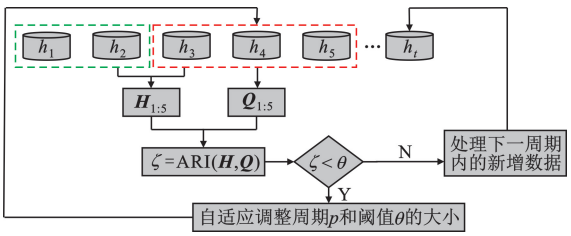


图 1 本文方法的主要框架

结合以上描述,本文提出了一种基于可扩展子空间学习的数据流聚类方法(S^{2}LCStream)。S^{2}LCStream 算法从初始化阶段开始,随后持续按滑动窗口投影数据并定期进行概念漂移检查。到达检查周期时,算法聚类最近一段数据并检测概念漂移。如无漂移,输出聚类结果并继续数据投影直至下一周期;若检测到漂移,则回溯至上一检查点并重新初始化。

基于可扩展子空间学习的数据流聚类(S^2LCStream)算法具体描述如下:

输入:$\boldsymbol{X}$:数据流,k:类簇数,h:窗口大小,nDim:特征维度。

输出:$\boldsymbol{X}$ 的聚类标签。

1 初始化算法,基于可扩展子空间学习建立历史数据和新增数据之间的投影关系。

2 通过公式(1)获得历史数据的聚类标签 $\boldsymbol{H}^*$,再通过公式(5)实时获取新增数据的聚类标签 $\boldsymbol{H}_t$。

3 通过公式(1)获取新增数据的聚类标签 $\boldsymbol{Q}_t$。

4 到达概念漂移检测周期,计算 $\zeta=\mathrm{ARI}(\boldsymbol{H}_t,\boldsymbol{Q}_t)$ 来进行概念漂移检测。

5 检测到概念漂移发生,返回第 1 步。

6 检测到未发生概念漂移,输出聚类结果并继续处理新增数据。

1.3　时间复杂度分析

S^2LCStream 按照窗口大小处理输入数据。h 是窗口大小,即滑动窗口内样本集合。S^2LCStream 检查一次概念漂移,并输出结果。这个过程称为漂移检查周期(p),并设置为($3\times h$),当检测到概念漂移时,概念漂移的阈值 θ 和漂移检查周期 p 会自适应调整。在漂移检查期间,S^2LCStream 在大小为 h 的窗口数据上进行 3 次投影聚类。为了漂移检查的目的,在大小为 p 的窗口数据上进行 1 次投影聚类。

可扩展子空间学习算法针对大小为 h 的窗口数据进行 3 次投影聚类,针对大小为 p 的窗口数据进行 1 次投影聚类。可扩展子空间学习算法的时间复杂度是 $O(t_1(hm^2+h^3)+nh^2+t_2h^3)$,其中,$h$ 是窗口实例数量,m 是特征维数,n 是数据点总数,t_1 是增广拉格朗日乘子法(Augmented Lagrange Method, ALM)迭代的次数,t_2 是 k-means 迭代次数。

对于每个周期,可扩展子空间学习算法在大小为 h 的窗口数据上运行 3 次,针对大小为 p 的窗口数据运行 1 次,因此整体的时间复杂度为 $O(4(t_1\cdot(hm^2+h^3)+nh^2+t_2h^3))$。这是单个周期内的计算复杂度。对于整个数据流处理,如果有 N 个周期,整体复杂度为 $O(N\times4\times(t_1(hm^2+h^3)+nh^2+t_2h^3))$。该复杂度依赖于数据维度 m、窗口大小 h、数据点总数 n、k-means 的迭代次数 t_2 和 ALM 迭代的次数 t_1。

2　实验与结果分析

2.1　实验设置

本实验旨在验证 S^2LCStream 算法的性能和效率,选择在 8 个不同的数据集上进行实验,并将其与 2 种静态聚类算法以及 6 种数据流聚类算法进行比较。实验使用 Python 及相关数据处理和机器学习库。本实验采用了独立测试重复实验,多次运行整个算法,每次都使用不同的随机种子以确保数据顺序不同。计算多次实验的平均值和标准差,以评估算法的稳定性和准确性。实验包括对比实验,以比较各算法在各数据集上的表现;参数实验,以评估关键参数变化对聚类效果的影响;效率评估,记录算法的运行时间对比。采用两种常见聚类评估指标来衡量算法的性能:归一化互信息(Normalized Mutual Information, NMI)和准确率(Accuracy, ACC)。ACC 数值范围为[0-1],趋近 1 时代表标签和聚类结果接近。NMI 是一种从信息论的角度衡量聚类效果的方法。NMI 对互信息进行了归一化处理,使其取值范围固定在 0~1 之间,0 表示两个聚类结果完全不相关,1 表示两个聚类结果完全一致。归一化互信息使用聚类结果的熵将互信息归一化至同一取值范围,使之能够对比不同聚类结果的优劣。NMI 越大,聚类效果与真实分类越接近。通过这些实验,深入分析了 S^2LCStream 在高维数据流聚类任务中对概念漂移的适应能力和处理效率。

2.2　真实数据集

为验证本文所提方法有效性,在 8 个真实数据集上进行试验,如表 1 所示。PEMS-SF[22] 数据集包含 440 个交通流序列数据,描述了旧金山湾区高速公路不同车道的占用率,其中每一天为一个单独的时间序列,其特征维度为 138 672。此数据集将每一天分类到正确的一周中的某一天,共有 7 个类别。AR[23] 数据集包含 1 400 张人脸图像,涵盖 100 名受试者,图像特征维度为 2 200。ExYaleB[24] 数据集包含 38 名受试者的 2 414 张人脸图像,特征维度为 2 016。MPIE[25] 数据集包含 286 个个体在不同环境下的 8 916 张面部图像,并通过主成分分析处理以保留 98% 的信息。NusWide[26] 数据集包含 30 000 张网络图像,属于 31 个类别,其特征维度为 639。Electricity[27] 电厂数据集提供了 45 312 个电力能源的市场价格波动情况的数据,其中有 8 个影响价格的因素,分 2 个类别。此外,在本实验中,采用两个真实电梯数据集验证所提方法对于真实数据流的聚类性能。2023_1_WX 收集了湖州市吴兴区 2023 年 1 月 1 218 台电梯发生的 125 种故障信息,被划分为 2 种类型的风险等级。2022_WX 收集了湖州市吴兴区 2022 年 11 741 台电梯发生的 773 种故障信息,被划分为 4 种类型的风险等级。

表 1 数据集信息

数据集	样本数	类簇数	特征维数
PEMS-SF[22]	440	7	138 672
AR[23]	1 400	100	2 200
ExYaleB[24]	2 414	38	2 016
MPIE[25]	8 916	286	115
NusWide[26]	30 000	31	639
Electricity[27]	45 312	2	8
2023_1_WX	1 518	2	125
2022_WX	11 741	4	773

2.3 对比算法

在本实验中,采用 2 种静态聚类算法 k-means 和 SLRR[18],以及 6 种数据流聚类算法作为基准对比。CluStream[10]通过在线微簇计算与离线宏聚类分析相结合实现数据流聚类任务。DenStream[11]对密度聚类进行拓展,强化数据流聚类过程的孤立点检测,并将聚类过程分为微簇在线更新和微簇离线处理。EmCStream[17]利用 UMAP 技术在线获取数据的二维嵌入,并通过滑动窗口模型对概念漂移进行检测。SVStream[15]基于支持向量描述,通过迭代维护数据的最小球体,以获取各类簇的边界。TSSRC[28]通过精确的簇数量评估标志窗口中数据对象之间的有效关系,这有效地将以前学到的知识随时间传递到当前的标志窗口。OSRC[29]通过引入低维投影到稀疏表示中以适应性地减少高维数据的噪声和冗余,并利用 $l_{2,1}$ 范数优化技术选取代表性数据对象,形成特定字典,从而有效地评估演变数据流中高维数据对象之间的关系并适应性地利用其演变子空间结构。

2.4 对比实验结果分析

在本实验中,通过比较在多个数据集上的聚类指标及单窗口的平均指标,来对实验结果进行分析。图 2 和图 3 所示为各方法在多个数据集上的聚类准确率及聚类互信息,表 2 和表 3 所示为各方法在多个数据集上的单窗口平均聚类准确率及平均聚类互信息。如图表所示,尽管 S^{2}LCStream 在整体性能上可能不如传统聚类算法如 k-means,但在处理分块数据方面表现出显著优势。特别是在高维图像数据集 AR 和交通流量数据集 PEMS-SF 上,S^{2}LCStream 的平均 ACC 值相较于其他 6 个数据流对比算法提升显著。这表明 S^{2}LCStream 能更好地处理高维数据流,这主要得益于两个关键策略:使用可扩展的子空间学习来进行投影学习和自适应概念漂移检测机制。

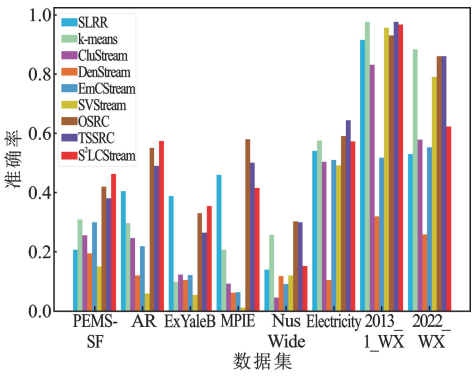


图 2 各数据集聚类准确率

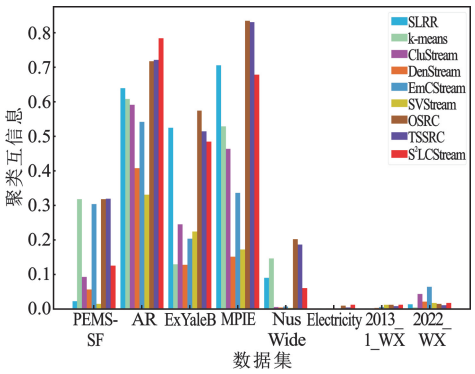


图 3 各数据集聚类互信息

表 2 单窗口聚类平均准确率及标准差

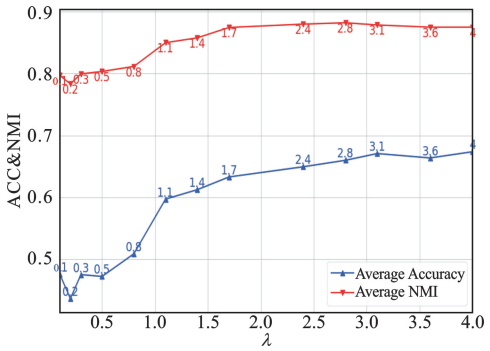
数据集	平均准确率±标准差						
	CluStream	DenStream	EmCStream	SVStream	OSRC	TSSRC	S^{2}LCStream
PEMS-SF[22](h=50)	0. 433±0. 041	0. 278±0. 064	0. 338±0. 033	0. 215±0. 028	0. 37±0. 028	0. 41±0. 041	**0. 478±0. 053**
AR[23](h=200)	0. 289±0. 022	0. 379±0. 023	0. 275±0. 100	0. 151±0. 040	0. 552±0. 022	0. 504±0. 035	**0. 671±0. 080**
ExYaleB[24](h=200)	0. 156±0. 016	0. 311±0. 046	0. 175±0. 049	0. 093±0. 005	0. 285±0. 028	0. 244±0. 019	**0. 443±0. 120**
MPIE[25](h=500)	0. 212±0. 011	0. 233±0. 014	0. 082±0. 045	0. 050±0. 010	**0. 578±0. 003**	0. 482±0. 057	0. 513±0. 135
NusWide[26](h=500)	0. 135±0. 016	0. 162±0. 021	0. 105±0. 009	0. 202±0. 059	**0. 269±0. 043**	0. 258±0. 042	0. 184±0. 015
Electricity[27](h=500)	0. 556±0. 04	0. 199±0. 033	0. 531±0. 038	0. 482±0. 033	0. 589±0. 063	**0. 612±0. 056**	0. 574±0. 023
2023_1_WX(h=100)	0. 957±0. 026	0. 401±0. 167	0. 721±0. 184	0. 721±0. 184	0. 93±0. 02	**0. 97±0. 019**	0. 966±0. 038
2022_WX(h=500)	0. 664±0. 056	0. 418±0. 204	0. 670±0. 119	**0. 886±0. 048**	0. 884±0. 048	0. 884±0. 048	0. 769±0. 138

表 3　单窗口聚类平均互信息及标准差

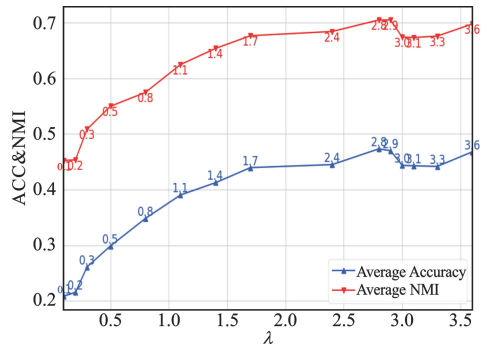
数据集	平均互信息±标准差						
	CluStream	DenStream	EmCStream	SVStream	OSRC	TSSRC	S²LCStream
PEMS-SF[22](h=50)	**0.441±0.040**	0.147±0.136	0.334±0.042	0.002±0.023	0.239±0.062	0.335±0.042	0.114±0.073
AR[23](h=200)	0.734±0.011	0.469±0.047	0.583±0.145	0.179±0.092	0.816±0.004	0.825±0.01	**0.882±0.037**
ExYaleB[24](h=200)	0.377±0.046	0.189±0.069	0.177±0.081	0.509±0.147	0.547±0.018	0.525±0.014	**0.673±0.098**
MPIE[25](h=500)	0.756±0.010	0.585±0.026	0.346±0.201	0.136±0.052	0.83±0.004	0.824±0.014	**0.833±0.061**
NusWide[26](h=500)	0.026±0.018	0.080±0.018	0.023±0.053	0.006±0.006	0.194±0.018	**0.198±0.016**	0.176±0.046
Electricity[27](h=500)	0.013±0.015	**0.060±0.029**	0.005±0.017	0.0001±0.007	0.009±0.017	0.02±0.025	0.015±0.003
2023_1_WX(h=100)	0.004±0.004	0.040±0.028	0.055±0.146	0.0007±0.0006	0.042±0.071	0.013±0.355	**0.075±0.257**
2022_WX(h=500)	0.020±0.015	0.098±0.030	**0.146±0.138**	0.006±0.004	0.059±0.046	0.024±0.021	0.039±0.053

2.5　参数分析

在本文中,算法中涉及一个平衡参数 λ。λ 用来平衡目标函数中的不同部分,参数的选择依赖于数据的分布。对于 S²LCStream,当 λ 的取值范围在 2.0~3.9 之间时,其准确率和 NMI 趋于平缓。图 4 展示了在两个公共数据集 AR 和 ExYaleB 上,不同 λ 的 ACC 和 NMI 结果。



(a) AR



(b) ExYaleB

图 4　λ 在不同数据集的参数实验

2.6　概念漂移检测分析

图 5 展示了 S²LCStream 和 EmCStream 在 8 个数据集上的概念漂移次数和聚类性能对比。在数据集 PEMS-SF、AR、MPIE 和 NusWide 上,S²LCStream 相比 EmCStream 检测到一样的漂移次数,但聚类精度显著增强,证明了本文算法在处理高维数据上的性能较为优异。对于两个真实电梯数据集,尽管 S²LCStream 检测到的漂移次数少于 EmCStream,但其聚类性能更优越。结合以上数据集的分析,S²LCStream 在多个高维数据集上均表现出色,不仅能检测到高维数据流发生的概念漂移,而且在聚类准确率上也远超 EmCStream,突显了所提算法处理高维数据流的高效性和实用性。

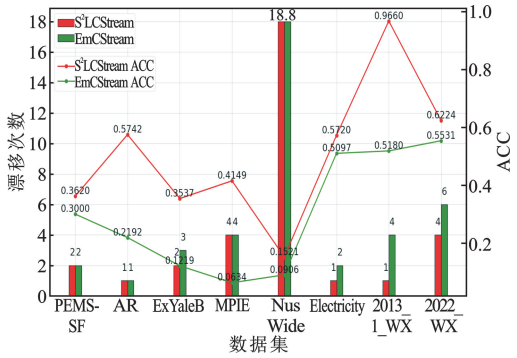


图 5　各数据集的概念漂移次数及其聚类精度

2.7　执行时间分析

代表性的数据流聚类方法在 8 个数据集上的执行时间结果如图 6 所示,这些数据流聚类方法的时间复杂度通常处于不同的水平。所提的算法复杂度为 $O(N\times4\times(t_1(hm^2+h^3)+nh^2+t_2h^3))$。从时间复杂度分析来看,所提算法的时间复杂度相对较高。同时,从执行时间比较来看,所提算法耗时与 EmCStream 相当,当处理大规模的数据集时,其运行时间是优于 EmCStream 算法的。由于所提算法将投影和回溯过程结合到一个模型中,确实需要相当多的时间。在对比图 2 和图 3 中的聚类结果后可以得出结论,虽然 S²LCStream 比其他数据流聚类方法

耗时,但它可以实现更好的聚类性能,同时可以自适应检测和适应概念漂移。

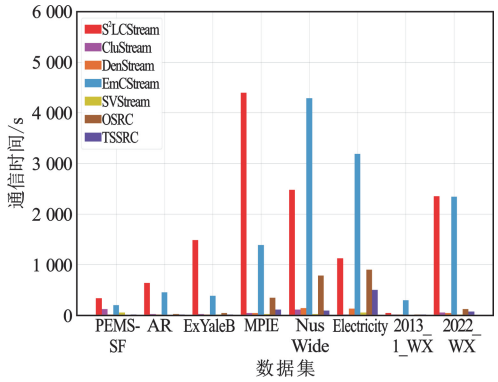


图 6 不同数据集的运行时间对比

3 结束语

本文基于可扩展子空间学习处理高维数据流聚类,并与概念漂移检测机制相结合。在此基础上,提出了一种新型基于可扩展子空间学习的数据流聚类方法(S²LCStream)。在 6 个公开的高维和大规模的数据集和 2 个真实电梯状态数据集上进行实验评估,结果显示 S²LCStream 在聚类质量方面显著优于现有的 DenStream 和 CluStream 等算法。

综上,S²LCStream 算法在处理具有高动态性和多样性的数据流中展示出了优越的聚类性能和适应能力。通过智能地应对概念漂移并利用低秩表示来提高数据处理的精度和稳定性,S²LCStream 成功地克服了传统聚类算法在复杂数据环境中面临的挑战。但是当数据规模过大时,本文方法的处理能力仍然存在可提升空间。未来将进一步针对此类问题进行深入研究,以适应更广泛的应用场景。

参考文献:

[1] SUÁREZ-CETRULO A L, QUINTANA D, CERVANTES A. A survey on machine learning for recurring concept drifting data streams [J]. Expert Systems with Applications, 2023, 213: 1-17.

[2] GAO Y J, FANG Z Q, XU J C, et al. An efficient and distributed framework for real-time trajectory stream clustering [J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(5): 1857-1873.

[3] LIN C C, CHEN C S, CHEN A P. Using intelligent computing and data stream mining for behavioral finance associated with market profile and financial physics[J]. Applied Soft Computing, 2018, 68: 756-764.

[4] 庞兴龙,朱国胜. 基于半监督学习的网络流量分析研究[J]. 计算机科学, 2022, 49(增刊 1): 544-554.

[5] KASHANI E S, BAGHERI SHOURAKI S, NOROUZI Y. Evolving data stream clustering based on constant false clustering probability [J]. Information Sciences, 2022, 614: 1-18.

[6] 张国毅,王晓峰,张旭洲. 基于数据流聚类的动态信号分选框架[J]. 电讯技术, 2011, 51(9): 65-68.

[7] BEZDEK J C, KELLER J M. Streaming data analysis: clustering or classification? [J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 51(1): 91-102.

[8] LI J P, YU H, ZHANG Z Y, et al. Concept drift adaptation by exploiting drift type[J]. ACM Transactions on Knowledge Discovery from Data, 2024, 18(4): 1-22.

[9] SILVA J A, FARIA E R, BARROS R C, et al. Data stream clustering: a survey [J]. ACM Computing Surveys, 2013, 46(1): 1-31.

[10] AGGARWAL C C, HAN J W, WANG J Y, et al. A framework for clustering evolving data streams[C]//The 29th International Conference on very Large Data Bases. Berlin: Morgan Kaufmann, 2003: 81-92.

[11] CAO F, ESTERT M, QIAN W N, et al. Density-based clustering over an evolving data stream with noise[C]// The 2006 SIAM International Conference on Data Mining. Bethesda: Society for Industrial and Applied Mathematics, 2006: 328-339.

[12] XU J, WANG G Y, LI T R, et al. Fat node leading tree for data stream clustering with density peaks [J]. Knowledge-Based Systems, 2017, 120: 99-117.

[13] DE ANDRADE J, HRUSCHKA E R, GAMA J. An evolutionary algorithm for clustering data streams with a variable number of clusters [J]. Expert Systems with Applications, 2017, 67: 228-238.

[14] PUSCHMANN D, BARNAGHI P, TAFAZOLLI R. Adaptive clustering for dynamic IoT data streams [J]. IEEE Internet of Things Journal, 2017, 4(1): 64-74.

[15] WANG C D, LAI J H, HUANG D, et al. SVStream: a support vector-based algorithm for clustering data streams[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(6): 1410-1424.

[16] 陈圆圆,王志海. 基于聚类分区的多维数据流概念漂移检测方法[J]. 计算机科学, 2022, 49(7): 25-30.

[17] ZUBAROGLU A, ATALAY V. Online embedding and clustering of evolving data streams [J]. Statistical Analysis and Data Mining: the ASA Data Science Journal, 2023, 16(1): 29-44.

[18] PENG X, TANG H J, ZHANG L, et al. A unified framework for representation-based subspace clustering of out-of-sample and large-scale data [J]. IEEE Transactions on Neural Networks and Learning Systems,

2016,27(12):2499-2512.
[19] 刘博,谢博鋆,朱杰,等. 快速可扩展的子空间聚类算法[J]. 模式识别与人工智能,2016,29(1):11-21.
[20] 朱林,雷景生,毕忠勤,等. 一种基于数据流的软子空间聚类算法[J]. 软件学报,2013,24(11):2610-2627.
[21] 陈金立,付善腾,朱熙铖,等. 阵元失效下基于核范数和 SCAD 惩罚的 MIMO 雷达 DOA 估计[J]. 电讯技术,2023,63(1):39-46.
[22] CUTURI M,UCI machine learning repository[EB/OL]. [2024-05-25]. https://doi. org/10. 24432/C52G70.
[23] MARTINEZ A,BENAVENTE R. The AR face database [R]. Columbus:Ohio State University,1998:318-323.
[24] GEORGHIADES A S,BELHUMEUR P N,KRIEGMAN D J. From few to many:illumination cone models for face recognition under variable lighting and pose[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2001,23(6):643-660.
[25] GROSS R,MATTHEWS I,COHN J,et al. Multi-PIE[J]. Image and Vision Computing,2010,28(5):807-813.
[26] CHUA T S,TANG J H,HONG R C,et al. NUS-WIDE:a real-world web image database from National University of Singapore [C]//The 8th ACM International Conference on Image and Video Retrieval. Santorini:ACM,2009:1-9.
[27] 陈圆圆. 数据流概念漂移检测及自适应聚类算法研究[D]. 北京:北京交通大学,2022.
[28] CHEN J,WANG Z,YANG S X,et al. Two-stage sparse representation clustering for dynamic data streams[J]. IEEE Transactions on Cybernetics, 2023, 53 (10): 6408-6420.
[29] CHEN J, YANG S X, FAHY C, et al. Online sparse representation clustering for evolving data streams[J]. IEEE Transactions on Neural Networks and Learning Systems,2025,36(1):525-539.

作者简介:

尹宏伟　男,1990 年生于安徽宿松,2019 年获博士学位,现为副教授,主要研究方向为机器学习、数据挖掘和聚类分析等。

倪钰洲　男,1999 年生于江苏苏州,2018 年获学士学位,现为硕士研究生,主要研究方向为聚类分析。

胡文军　男,1977 年生于安徽绩溪,2012 年获博士学位,现为教授,主要研究方向为机器学习、模式识别、数据挖掘、智能系统等。