

基于谱聚类欠取样的不平衡漂移 大数据流分类算法

康耀龙¹, 冯丽露², 张景安³

(1. 山西大同大学 计算机与网络工程学院, 山西 大同 037009; 2. 山西大同大学 教育科学与技术学院, 山西 大同 037009; 3. 山西大同大学 网络信息中心, 山西 大同 037009)

摘要:在不平衡数据分类中,多数类样本在数量上占优,其分布会对聚类结果产生较大的“拉扯”作用,而少数类样本由于数量少,其特征在整个数据集中相对不明显,导致数据流出现漂移问题,影响数据流分类效果。为解决此问题,本文研究基于谱聚类欠取样的不平衡漂移大数据流分类算法。通过欠取样技术减少不平衡漂移大数据流中多数类数据的冗余数据量,均衡多数类数据与少数类数据的数据量,缓解数据因聚类“拉扯”导致的漂移问题;选取均衡后大数据流的核心点构成核心点集,通过谱聚类算法对此核心点集实施聚类,基于谱聚类所得到的聚类结构以及选定的核心点,实现不平衡漂移大数据流分类。实验结果显示:该算法可实现不平衡漂移大数据流的均衡处理,处理后平均不平衡度可降至 1.024,几乎接近于平衡状态;可实现不同属性大数据流核心点的选定与有效分组,为此类大数据流的后续有效应用提供保障。

关键词:谱聚类;欠取样;不平衡;漂移大数据流;核心点集;组别划分

中图分类号:TP391 **文献标志码:**A **文章编号:**1671-5497(2026)02-0509-07

DOI:10.13229/j.cnki.jdxbgxb.20241184

Unbalanced drift big data stream classification algorithm based on spectral clustering undersampling

KANG Yao-long¹, FENG Li-lu², ZHANG Jing-an³

(1. School of Computer and Network Engineering, Shanxi Datong University, Datong 037009, China; 2. School of Education Science and Technology, Shanxi Datong University, Datong 037009, China; 3. Network Information Center, Shanxi Datong University, Datong 037009, China)

Abstract: In imbalanced data classification, the majority of class samples have an advantage in terms of quantity, and their distribution will have a significant "pulling" effect on the clustering results. However, the minority class samples, due to their small quantity, have relatively unclear features in the entire dataset, resulting in drift problems in the data stream and affecting the classification performance of the data stream. To address this issue, research is conducted on an imbalanced drift big data stream classification

收稿日期:2024-11-05.

基金项目:国家自然科学基金项目(71601101);山西省艺术科学规划课题项目(24BA152);山西省高等学校教学改革创新项目(J20241141);山西省哲学社会科学规划课题项目(2025YB184);大同市应用基础研究计划项目(2025074);大同市应用基础研究计划项目(2024080);山西大同大学云冈学专项项目(2022YGZX016).

作者简介:康耀龙(1979-),男,副教授,硕士.研究方向:数据挖掘,大数据技术.E-mail:yaolong_king@126.com

algorithm based on spectral clustering undersampling. By using undersampling techniques to reduce the redundant amount of majority class data in imbalanced drift big data streams, balance the amount of majority class data and minority class data, and alleviate the problem of data drift caused by clustering "pulling"; select the core points of the balanced big data stream to form a core point set, and use spectral clustering algorithm to cluster this core point set. Based on the clustering structure obtained from spectral clustering and the selected core points, realize the classification of imbalanced drift big data streams. The experimental results show that the algorithm can achieve balanced processing of imbalanced drift big data streams, and the average imbalance degree after processing can be reduced to 1.024, almost approaching the equilibrium state; it can achieve the selection and effective grouping of core points for different attribute big data streams, providing guarantees for the subsequent effective application of such big data streams.

Key words: spectral clustering; undersampling; out of balance; drift big data stream; core point set; group division

0 引言

在机器学习和模式识别领域,不平衡数据流分类问题一直是一个重要的研究方向。不平衡指的是数据流中某一类(多数类)的样本数量明显多于另一类(少数类)的样本数量^[1]。这种不平衡性往往导致传统的分类算法在分类上产生偏差,无法有效识别出少数类样本。同时,数据流中的漂移现象也使得分类任务更加复杂^[2,3]。漂移是指数据流中数据的统计特性或分布随时间发生变化^[4],这种变化可能是由多种因素引起的,如环境变化^[5]。因此,为应对数据流中的不平衡性和漂移问题,保证机器学习与模式识别领域内不平衡漂移大数据流的精准分类,需研究一种科学合理的不平衡漂移大数据流分类算法。

目前,已有部分国内外学者针对大数据流的分类问题展开了相关研究,例如赵小强等^[6]提出基于密度的空间聚类算法(Density-based spatial clustering of applications with noise, DBSCAN)与过采样的随机森林不平衡数据分类算法,先对多数类用DBSCAN聚类分解,再对少数类过采样,最后通过随机森林算法实现数据分类。该算法可以在不丢失数据的情况下削弱多数类样本的优势,有助于平衡数据集,但如果生成的样本过多或分布不合理,可能引发过拟合风险。黄富幸等^[7]提出使用改进蜉蝣算法(Improved mayfly algorithm, IMA)优化反向传播(Back propagation, BP)神经网络分类模型,通过引入阻尼比系数和非线性惯性权重因子,提高分类稳定性,引入集成学习中的软投票机制,将多个 IMA-BP 弱分类器

结果融合,实现数据分类。该算法能提高 BP 神经网络分类模型的稳定性,但如果阻尼比系数、非线性惯性权重因子等参数设置不当,可能会影响模型的性能,甚至导致分类结果不佳。周尔昊等^[8]提出旋转平衡森林(Rotated balanced forest, ROBF)模型,从安全等级再划分机制和引入约束度不同的控制因子两个角度出发,改进旋转森林模型,利用改进后的旋转平衡森林模型实现数据分类。该方法能满足更广泛的数据分类需求,但复杂的模型结构和较多的参数增加了模型过拟合的风险。Kenger 等^[9]将模糊最小-最大神经网络与混合整数线性规划模型相结合展开数据分类,该方法节省了计算资源和时间成本,可以更快地得到分类结果,但模型结合过程中可能会出现一些难以协调的问题,影响模型的整体性能。

综合以上分析,本文结合欠取样与谱聚类技术,设计一种针对不平衡漂移大数据流的分类算法,旨在减少因数据不平衡和漂移导致的误分类情况,提升数据分类效果。

1 不平衡漂移大数据流的欠取样均衡处理

不平衡漂移大数据流中包含多数类数据与少数类数据,其中,多数类数据可能包含噪声数据或冗余数据,此类噪声冗余数据会造成大数据流谱聚类时的聚类边界向少数类数据方向偏移,如此会增加误聚类概率,从而降低不平衡漂移大数据流的分类精度^[10-13]。为解决此问题,需对不平衡漂移大数据流中的多数类数据进行欠取样处理,减少多数类数据的冗余数据量,均衡多数类数据

与少数类数据的数据量,保证不平衡漂移大数据流的聚类精度^[14],提升数据流分类效果。不平衡漂移大数据流的欠取样处理过程可描述为:

(1) 设 Z 为原始不平衡漂移大数据流; m 为该大数据流的数据点数量; G 为原始不平衡漂移大数据流 Z 中的属性数量; u 为该大数据流中所包含的数据类型数量; U_j 为每种数据类型所包含的数据量。设 m/u 为阈值,将原始不平衡漂移大数据流 Z 中数据量高于阈值 m/u 的数据类型划分成多数类,反之,则划分成少数类。

(2) 运用欠取样方法对划分后的原始不平衡漂移大数据流 Z 中的多数类数据进行欠取样处理,得到数据量为 U_j 的多数类数据集 G_j 。

(3) 如果 G_j 低于所设定阈值 m/u ,则需对此多数类数据集继续进行有放回抽样处理,增加数据量至阈值 m/u 。另外,对原始不平衡漂移大数据流中的少数类数据进行有放回抽样处理,令其数据量同样达到阈值 m/u ,实现漂移大数据流中多数类与少数类数据量的均衡。将均衡后的多数类数据与少数类数据组合,构成平衡漂移大数据流 $\dot{Z}$ 。

2 欠取样均衡漂移大数据流的谱聚类与分类

欠取样均衡后的漂移大数据流可能仍然存在复杂的内在结构,谱聚类有助于将少数类样本更合理地聚类,它不像一些基于距离的聚类方法容易受到多数类样本的影响,能在一定程度上保护少数类样本的特征,提高后续分类算法对少数类样本的识别能力,进而提升整个分类算法在不平衡漂移大数据流场景下的性能。因此,通过谱聚类对上节中所获得的平衡漂移大数据流 $\dot{Z}$ 实施聚类。

由于谱聚类实际运算过程复杂,将其直接应用到大数据流分类会增加一定的困难。为解决此问题,在对第2节中欠取样均衡后的平衡漂移大数据流 $\dot{Z}$ 实施谱聚类之前,先对该大数据流 $\dot{Z}$ 实施核心选取,并构建核心点集,依据所选定核心点集对所构建的大数据流 $\dot{Z}$ 实施聚类,将所得大数据流分组结果与核心点集谱聚类结果相结合,实现不平衡漂移大数据流的聚类处理。

在此结合数据相似性及随机抽样,选定欠取样均衡后的平衡漂移大数据流 $\dot{Z}$ 的核心点,构建其核心点集。大数据流的核心点即同其余数据点具有最高相似性的点。欠取样均衡后的平衡漂移

大数据流 $\dot{Z}$ 的核心点选定过程为:

(1) 通过随机抽样获得大数据流 $\dot{Z}$ 的样本数据集 Z' 。

(2) 对大数据流 $\dot{Z}$ 及其随机抽样的样本数据集 Z' 内全部数据对之间的相似性实施运算,同时选取运算所得相似性最高的数据点作为大数据流 $\dot{Z}$ 的核心点。

(3) 设 $\dot{z}_1^*, \dot{z}_2^*, \dots, \dot{z}_{L-1}^*$ 为所选择的大数据流 $\dot{Z}$ 的核心点,其中, L 为核心点的数量。基于此,可得到第 L 个核心点为:

$$\dot{z}_L^* = \arg \max \left\{ \sum_{z' \in Z'} s(\dot{z}, z') - \gamma \sum_{j=1}^{L-1} s(\dot{z}, \dot{z}_j^*) \right\} \quad (1)$$

式中: $s(\dot{z}, z')$ 为大数据流 $\dot{Z}$ 中数据 $\dot{z}$ 与随机抽样样本数据集 Z' 内数据 z' 之间的相似性; γ 为松弛参数; $s(\dot{z}, \dot{z}_j^*)$ 为大数据流 $\dot{Z}$ 中数据 $\dot{z}$ 与第 j 个核心点 $\dot{z}_j^*$ 之间的相似性。

根据上述过程所选定的大数据流 $\dot{Z}$ 核心点 $\dot{z}_1^*, \dot{z}_2^*, \dots, \dot{z}_L^*$, 构建其核心点集 $\dot{Z}^*$, 在此基础上,展开欠取样均衡漂移大数据流谱聚类,具体过程如下:

(1) 针对欠取样后的平衡漂移大数据流 $\dot{Z}$, 构建所需聚类的核心点集无向图 $A = (W, F)$, 其中, $W = \{w_1, w_2, \dots, w_l\}$ 表示该无向图中的顶点集合,即大数据流 $\dot{Z}$ 的核心点集; l 表示顶点的数量,即该大数据流的核心点数量; F 表示各顶点之间的边;通过各个核心点之间的相似性为边 F 赋权重 ω 。

(2) 如果该无向图 A 属于加权图,那么其中两个核心点 w_j 和 w_i 之间的边的权重为 ω_{ji} ,同时该权重值大于0;当 $\omega_{ji} = W$ 时,表明这两个核心点间无相连的边。

(3) 设 $\omega' = (\omega_{ji})$ 表示用于判定无向图 A 加权相连边矩阵的方程式,同时加权相连边符合 $\omega_{ji} = \omega_{ij}$,此时,判定核心点 $w_j \subset W$ 的方程式为:

$$b_j' = \sum_{i=1}^l \omega_{ji} \times \dot{z}_i^* \quad (2)$$

(4) 对无向图 A 实施判定的度矩阵可表示成:

$$B = \text{diag} b_j' (b_1', b_2', \dots, b_l') \quad (3)$$

(5) 对无向图 A 实施标准化处理后,能够得到其拉普拉斯矩阵为:

$$\text{Laplacian} = J - B^{-1/2} \omega' B^{-1/2} \quad (4)$$

运算得到该拉普拉斯矩阵的前 k 个特征向量 $\eta_1, \eta_2, \dots, \eta_k$, 以此类特征向量为矩阵中的列向量, 生成矩阵 $C \in R^{l \times k}$ 。

(6) 对矩阵 $C \in R^{l \times k}$ 中的各行向量实施标准化处理, 使行向量 ϕ_{ji} 能够符合:

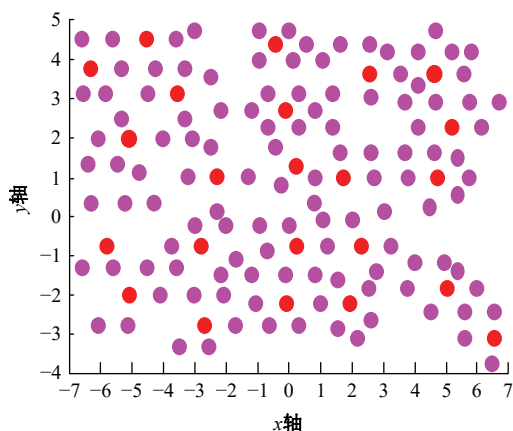
$$\phi_{ji} = \eta_{ji} \times \text{Laplacian} \left(\sum_{i,j=1}^K \eta_{ji}^2 \right)^{1/2} \quad (5)$$

通过式 (4) 可获得标准化处理后的矩阵 $C' \in R^{l \times k}$ 。设 $\phi_j \in R^k$ 表示该矩阵中的第 j 行向量, 且 $j = 1, 2, \dots, l$ 。

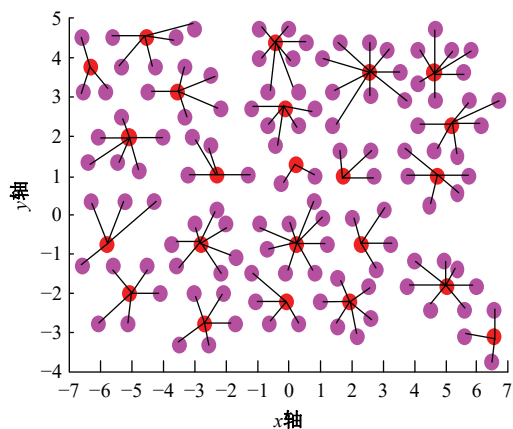
(7) 采用 K -means 聚类算法对行向量中所包含的大数据流核心点实施聚类, 生成一个核心点子集 $D_1, D_2, \dots, D_k$ 。通过该方法完成全部行向量中大数据流核心点的聚类后, 获得该大数据流核心点集的最终谱聚类结果。

运用本文算法对欠取样均衡处理后的大数据流实施核心点选定与分组, 结果如图 1 所示。

由图 1 可知, 通过本文算法可实现大数据流



(a) 核心点选定与分组前



(b) 核心点选定与分组后

图 1 核心点选定与分组效果

Fig. 1 Core point selection and grouping effect

核心点的有效选定, 实现组别准确划分, 为后续大数据流分类奠定坚实的基础。

通过谱聚类对数据进行聚类操作, 将数据划分为不同的簇, 这有助于初步梳理数据的分布情况。基于谱聚类所得到的聚类结构以及选定的核心点, 实现不平衡漂移大数据流分类。

划分大数据流 $\dot{Z}$ 为 L 个组别, 以各核心点 $\dot{z}_j^*$ 表示各个分组 $\dot{Z}_j$ 。另外, 将大数据流 $\dot{Z}$ 内的非核心点数据, 向与其相似性最高的核心点所在组别划分, 此类非核心点数据的组别划分方程式为:

$$\dot{z}_{l_0}^* = \arg \max_{1 \leq j \leq l} s(\dot{z}, \dot{z}_j^*) \quad (6)$$

完成大数据流 $\dot{Z}$ 的遍历之后, 将 $\dot{Z}$ 划分成 L 个组别 $\dot{Z}_1, \dot{Z}_2, \dots, \dot{Z}_L$, 以此实现不平衡漂移大数据流分类。

3 实验分析

实验从 UCI 数据库中选取 12 组数据作为实验大数据流, 该数据库具备机器学习等各种实际应用背景, 运用本文算法对所选取的 12 组实验大数据流进行处理, 检验分析本文算法的实际应用效果。所选取的实验大数据流初始情况见表 1。

实验大数据流属于不平衡漂移数据, 其属性会随时间变化而出现漂移变化, 以 car、seg1、letter 3 个大数据流中的部分漂移数据为例, 呈现数据漂移情况, 如图 2 所示。

运用本文算法先对此类实验大数据流进行欠取样处理, 以实现大数据流的均衡性, 处理后的实验大数据流详情见表 2。

表 1 实验大数据流初始情况

Table 1 Initial situation of experimental big data flow

大数据流名称	属性	多数类数据量	少数类数据量	不平衡度
pima	10	510	277	1.84
car	7	1 200	373	3.22
sonar	25	122	95	1.28
yeast	10	452	418	1.08
ionosphere	36	214	115	1.86
glass	11	75	16	4.69
breast-w	30	504	280	1.80
letter	17	778	755	1.03
seg1	24	2 162	463	4.67
wdbc	31	346	201	1.72
haberman	5	214	115	1.86
artificial	28	5674	813	6.98

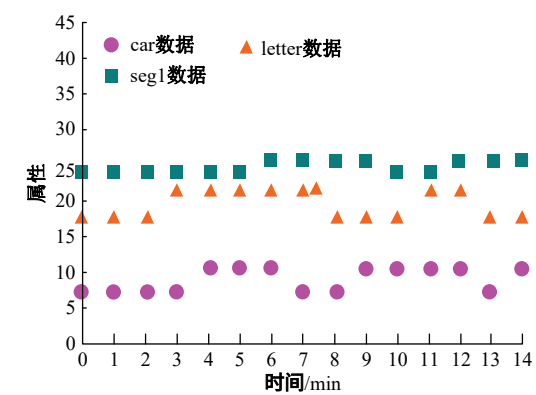


图 2 实验大数据流部分数据漂移情况

Fig. 2 Partial data drift of experimental big data stream

由表 2 可看出,经本文算法的欠取样均衡处理后,各实验大数据流中的多数类数据量均明显下降,且更接近少数类数据量,不平衡度显著降低,平均不平衡度降至 1.024,几乎接近平衡状态。这是因为本文算法将欠取样技术应用于处理不平衡漂移大数据流中的多数类数据冗余问题,通过减少多数类样本数量,实现了数据量的均衡,为后续的分类任务奠定了坚实的基础。

上文对核心点选定与分组效果进行了验证,为了更加全面地验证本文算法的有效性,继续运用本文算法对欠取样均衡处理后的各实验大数据流实施核心点选定与分组,以 seg1、letter 大数据流为例,所得结果如图 3 所示。图中红色区域代表本文算法所选定的核心点。

由图 3 可知,通过本文算法可实现各实验大数据流的核心点选定与组别划分,所选定的 car、

表 2 本文算法欠取样均衡后的大数据流详情

Table 2 Details of the algorithm in this article for undersampling and equalization of big data streams

大数据流名称	属性	多数类数	少数类数	不平衡度
		据量	据量	
pima	10	298	277	1.076
car	7	377	373	1.011
sonar	25	100	95	1.053
yeast	10	419	418	1.002
ionosphere	36	117	115	1.017
glass	11	17	16	1.063
breast-w	30	282	280	1.007
letter	17	757	755	1.003
seg1	24	466	463	1.006
wdbc	31	203	201	1.010
haberman	5	118	115	1.026
artificial	28	815	813	1.002

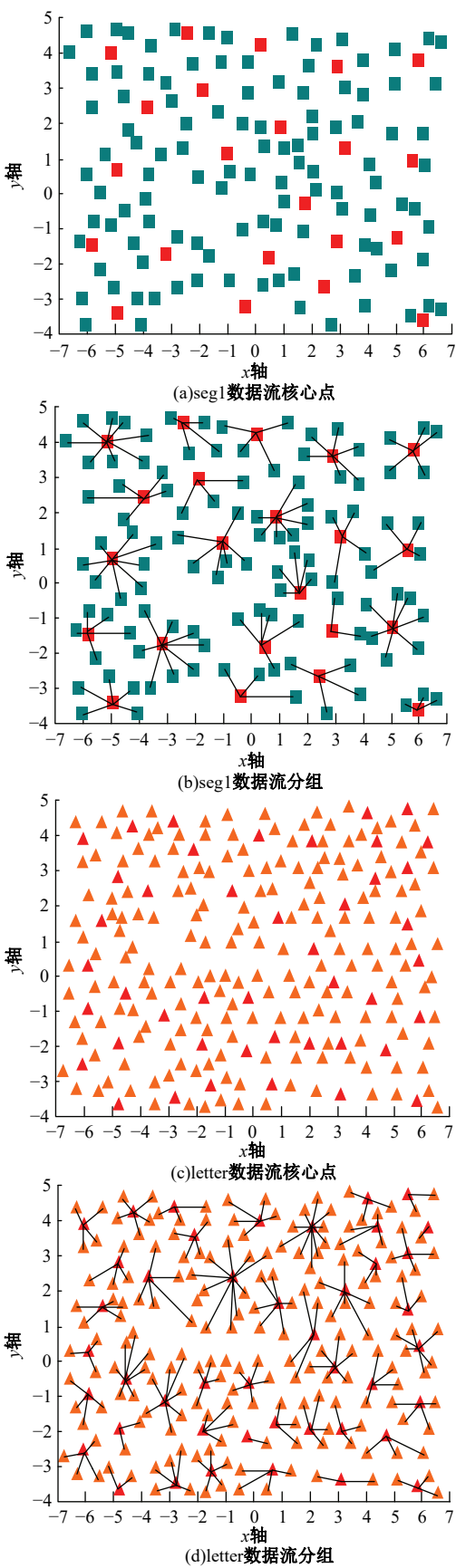


图 3 均衡后各实验大数据流核心点选定与分组结果

Fig. 3 Core point selection and grouping results of each experiment big data flow after equalization

seg1、letter 3个大数据流的核心点数量依次为23、22、44个,依据所选定的各核心点完成各实验大数据流的分组任务。原因在于本文算法在对数据进行均衡处理后,提出了构建核心点集的方法,巧妙地融合了谱聚类算法对核心点集进行聚类分析,有效解决了不平衡数据分类中的漂移问题。

在上述实验基础上,继续通过本文算法对各个大数据流的核心点集实施分类,以car、seg1、letter 3个大数据流为例,呈现本文算法最终所得各个实验大数据流的分类结果,如图4所示。

分析图4可得出,本文算法基于谱聚类所得到的聚类结构以及选定的核心点,实现了对不平衡漂移大数据流的分类,提高了分类的准确性,为不平衡漂移大数据流的有效应用提供了全新的解决方案。因此,采用该算法进行数据分类时能针

对不同大数据流实现有效的分类处理,且分类结果十分准确,不存在误分类问题。

4 结 论

(1)通过欠取样技术,有效减少了多数类数据的冗余数据量,从而实现了多数类数据与少数类数据之间的数据量均衡,有效缓解了数据量不均衡的问题。

(2)在对数据进行均衡处理后,选取大数据流中的核心点构成核心点集,并应用谱聚类算法对其进行聚类,利用谱聚类在处理复杂数据结构方面的优势,解决了不平衡数据分类中的漂移问题,提高了分类效果,为不平衡漂移大数据流的有效应用提供了保障。

(3)实验结果显示:该算法能显著降低不平衡漂移大数据流的不均衡度,几乎达到了平衡状态,并实现了不同属性大数据流核心点的有效选定与分组,验证了该算法的实际效果,为其在实际应用中的推广提供了有力支持。

参考文献:

- [1] 宋婷婷, 吴赛君, 裴颂文. 融合BiLSTM的双图神经网络文本分类模型[J]. 上海理工大学学报, 2023, 45(2): 103-111.
Song Ting-ting, Wu Sai-jun, Pei Song-wen. Dual graph neural networks with BiLSTM for text classification[J]. Journal of University of Shanghai for Science and Technology, 2023, 45(2): 103-111.
- [2] 邓维斌, 王智莹, 高荣壕, 等. 融合注意力与CorNet的多标签文本分类[J]. 西北大学学报: 自然科学版, 2022, 52(5): 824-833.
Deng Wei-bin, Wang Zhi-ying, Gao Rong-hao, et al. Multi-label text classification combining attention with CorNet[J]. Journal of Northwest University (Natural Science Edition), 2022, 52(5): 824-833.
- [3] 崔雨萌, 王靖亚, 刘晓文, 等. 融合注意力和裁剪机制的通用文本分类模型[J]. 计算机应用, 2023, 43(8): 2396-2405.
Cui Yu-meng, Wang Jing-ya, Liu Xiao-wen, et al. General text classification model combining attention and cropping mechanism[J]. Journal of Computer Applications, 2023, 43(8): 2396-2405.
- [4] 张虎, 柏萍. 融入句子中远距离词语依赖的图卷积短文本分类方法[J]. 计算机科学, 2022, 49(2): 279-284.

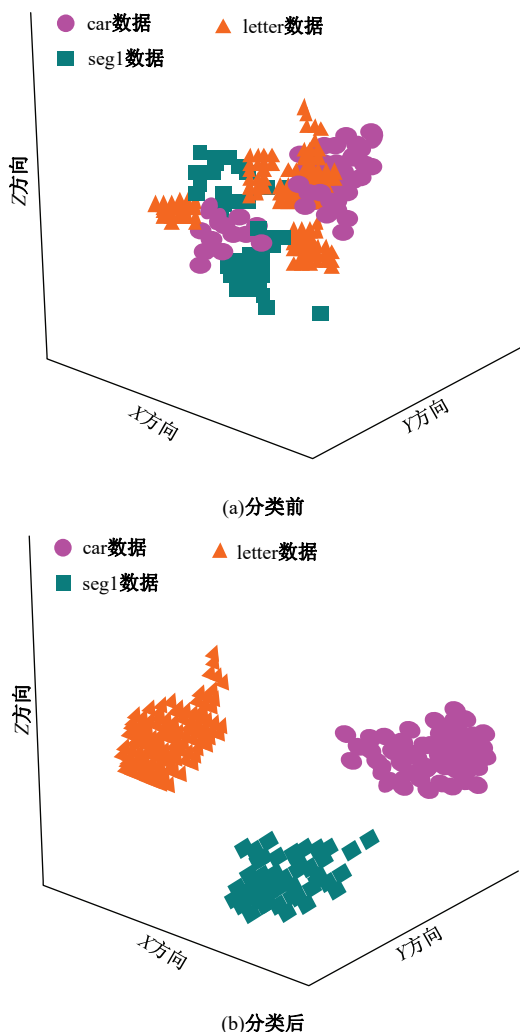


图4 本文算法大数据流分类结果

Fig. 4 This article presents the classification results of big data streams using algorithms

- Zhang Hu, Bai Ping. Graph convolutional networks with long-distance words dependency in sentences for short text classification[J]. Computer Science, 2022, 49(2): 279-284.
- [5] 赵嘉, 姚占峰, 吕莉, 等. 基于相互邻近度的密度峰值聚类算法[J]. 控制与决策, 2021, 36(3): 543-552.
Zhao Jia, Yao Zhan-feng, Lyu Li, et al. Density peak clustering algorithm based on mutual proximity[J]. Control and Decision, 2021, 36(3): 543-552.
- [6] 赵小强, 姚青磊. 基于DBSCAN聚类分解和过采样的随机森林不平衡数据分类算法[J]. 兰州理工大学学报, 2023, 49(6): 80-89.
Zhao Xiao-qiang, Yao Qing-lei. Random forest imbalanced data classification algorithm based on DBSCAN clustering decomposition and oversampling[J]. Journal of Lanzhou University of Technology, 2023, 49(6): 80-89.
- [7] 黄富幸, 韩文花. 基于Voting机制的IMA-BP不平衡数据分类算法[J]. 科学技术与工程, 2023, 23(27): 11698-11705.
Huang Fu-xing, Han Wen-hua. Classification algorithm of IMA-BP for unbalanced data based on voting mechanism[J]. Science Technology and Engineering, 2023, 23(27): 11698-11705.
- [8] 周尔昊, 高尚, 申震. 基于旋转平衡森林的不平衡数据分类算法[J]. 计算机工程与设计, 2022, 43(2): 458-464.
Zhou Er-hao, Gao Shang, Shen Zhen. Classification algorithm of imbalanced data based on rotation balanced forest[J]. Computer Engineering and Design, 2022, 43(2): 458-464.
- [9] Kenger M N, Ozceylan E. A hybrid approach based on mathematical modelling and improved online learning algorithm for data classification[J]. Expert Systems with Applications, 2023, 218(5): 1-16.
- [10] 毕志臻, 杨德刚, 冯骥. 面向超大规模数据的自适应谱聚类算法[J]. 智能系统学报, 2023, 18(2): 251-259.
Bi Zhi-zhen, Yang De-gang, Feng Ji. Adaptive spectral clustering algorithm for very large scale data[J]. Journal of Intelligent Systems, 2023, 18(2): 251-259.
- [11] 古险峰, 汤永利. 基于群体智能算法的混合属性大数据聚类仿真[J]. 计算机仿真, 2023, 40(9): 458-461.
Gu Xian-feng, Tang Yong-li. Clustering simulation of mixed attribute big data based on swarm intelligence algorithm[J]. Computer Simulation, 2023, 40(9): 458-461.
- [12] 张熳, 徐兆瑞, 沈项军. 一种傅里叶域海量数据高速谱聚类方法[J]. 北京航空航天大学学报, 2022, 48(8): 1445-1454.
Zhang Man, Xu Zhao-rui, Shen Xiang-jun. A high-speed spectral clustering method for Fourier domain massive data[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1445-1454.
- [13] 梁浩玮, 王石, 曹存根. 非完美多分类标签体系下的领域短文本分类方法研究[J]. 计算机科学, 2023, 50(1): 185-193.
Liang Hao-wei, Wang Shi, Cao Cun-gen. Study on short text classification with imperfect labels[J]. Computer Science, 2023, 50(1): 185-193.
- [14] 黄伟, 刘贵全. MSML-BERT模型的层级多标签文本分类方法研究[J]. 计算机工程与应用, 2022, 58(15): 191-201.
Huang Wei, Liu Gui-quan. Study on hierarchical multi-label text classification method of MSML-BERT model[J]. Computer Engineering and Applications, 2022, 58(15): 191-201.