

doi: 10.7690/bgzd.2025.12.007

基于高频数据共现聚类算法的异构网络信息增量式更新方法

李昌伟, 王 凯, 张立博, 李 明, 李 帅

(中国绿发投资集团有限公司, 北京 100010)

摘要: 为解决异构网络信息聚类过程中存在数据漂移, 且目前的增量更新方法不能识别伪相关数据的问题, 提出一种基于高频数据共现聚类算法的异构网络信息增量式更新方法。利用漂移点偏差挖掘算法搜索数据集中的漂移点位置, 将漂移点较多的数据集识别为异常数据并进行剔除。统计数据集中的数据属性同时标记相同数据集的频繁次数, 采用关联特征选择算法提取代表性语素属性, 提取与用户检索匹配度一致的真实信息进行增量式更新。实验结果表明, 该方法能成功收敛漂移数据点, 伪相关数据查准率为 98.2%, 具有较强的可行性。

关键词: 高频数据; 共现聚类; 漂移点偏差; 异构网络; 增量式更新; 数据离散

中图分类号: TP393 **文献标志码:** A

Incremental Update Method for Heterogeneous Network Information Based on High-frequency Data Co-occurrence Clustering Algorithm

Li Changwei, Wang Kai, Zhang Libo, Li Ming, Li Shuai

(China Green Development Investment Group Co., Ltd., Beijing 100010, China)

Abstract: In order to solve the problem of data drift in the information clustering process of heterogeneous networks and the inability of current incremental update methods to identify pseudo-relevant data, an incremental update method for heterogeneous network information based on a high-frequency data co-occurrence clustering algorithm is proposed. The drift point deviation mining algorithm is utilized to search for the locations of drift points in the dataset, identifying datasets with a high number of drift points as abnormal and subsequently removed. Data attributes in the dataset are statistically analyzed, and frequent occurrences within the same dataset are marked. An association feature selection algorithm is employed to extract representative morpheme attributes, and genuine information that matches user queries is extracted for incremental updates. Experimental results demonstrate that this method can successfully converge drift data points, achieving a pseudo-relevant data precision rate of 98.2%, indicating strong feasibility.

Keywords: high-frequency data; co-occurrence clustering; drift point bias; heterogeneous network; incremental updating; data discretization

0 引言

随着互联网领域的高速发展, 单一的网络结构已经无法满足用户的业务需求。伴随网络信息兼容能力的不断提高, 基于异构网络的融合越来越被广泛应用。由于异构网络接入技术的泛化特性, 导致高频数据信息真伪共存, 用户在检索过程中容易受到与原始数据并无关联的虚假信息误导; 因此, 如何在多源异构数据更新过程中聚类相关数据已然成为当下研究的重点。

文献[1]采用对抗学习方法设立映射函数并构建约束条件, 利用循环生成模型策略将文字与图片属性进行转换获得其相关性, 通过模型训练实现异构网络数据的一致化检索与更新; 然而, 该方法难以准确标定漂移数据点, 导致聚类效果不佳。文献[2]提出一种基于聚类分析的数据更新方法, 利用

并行计算来获得关联数据和数据集的不确定性。通过标签聚类方法对数据集标记编码, 引入差分匿名泛化模型将分类树中的父值降维替换为子值, 从而提高异构数据的聚类效果。这种方法存在的缺点是难以精准判定数据集中的伪相关数据, 导致更新数据的真实性不高。

针对以上问题, 笔者提出一种基于高频数据共现聚类算法的异构网络信息增量式更新方法, 在提高更新数据与源数据相关性的基础上, 有效解决高频数据漂移问题, 具有较强的可靠性。

1 高频数据共现聚类算法

异构网络由于包含多种接入技术, 在信息传输过程中存在局部离散漂移问题, 需要采用共现聚类算法实现高频数据的聚类标记。在高频共现数据更新过程中对真伪相关度进行判定, 防止用户在更新

收稿日期: 2024-10-26; 修回日期: 2024-11-26

基金项目: 中国绿发投资集团有限公司科技项目 (CGDG529000220008)

第一作者: 李昌伟(1987—), 男, 山东人, 硕士。

数据过程中存在蹭热度导致的数据伪相关度高，而降低用户在网络信息检索过程中的体验，故采用了基于漂移点偏差挖掘算法发现数据集中的异常数据，由此实现高频数据共现聚类研究。

1.1 算法参数设置及规则

在高频数据集中，个别数据会与大多数数据产生一定偏差，其距离大于某个阈值的节点被称为漂移点，包含漂移点较多的数据集合即判定为异常数据^[3]。在进行漂移点搜索时，可以通过划分网格判定漂移数据空间位置，从而挖掘偏离于其他数据的漂移数据^[4]。

按等距进行网格划分，假设各维度分段节点数量为 f ，如果单个维度中包含 M 个高频数据集，那么 f 的计算公式为：

$$f = \left\lceil \sqrt[d+1]{M} \right\rceil + 2 \cdot d。 \quad (1)$$

式中 d 为全部数据之间距离的平均值，大于 d 值的数据点即为漂移点。参照单个维度， f 值的大小由漂移度阈值 T_i 确定，当阈值固定时，无法精确获得区间内高频数据分布状态。由此，结合数据的实际分布特征对阈值动态取值：

$$T_i = (1/(f-1)) \sum_{u=1}^{f-1} \delta_u。 \quad (2)$$

式中 δ_u 为邻接区间的共现程度，若 $T_i < \delta_u$ ，则表示邻接区间的共现程度较高，若 $T_i > \delta_u$ ，则表示邻接区间伪共现度较高，判定为异常数据。将共现程度较高的区间重新组合，通过任意维度的所有邻接区间共现程度平均值确定 T_i 数值，提高重组区间数据的规范性。

基于漂移点偏差挖掘算法参数设置规则如下：

规则 1：按照等距划分网格，搜索网格区间内每个高频数据集并进行比对，将漂移点较多的高频数据集判定为异常数据，并进行剔除^[5]。

规则 2：在剩余的高频数据上再次进行漂移点偏差挖掘计算，计算重组网格区间内存在的高频数据点数量，将每个数据集合中漂移点数量递减排序，获取漂移点较多的高频数据并进行标记。

1.2 高频数据共现聚类

当高频数据集中出现共现度较高的高频数据信息时，有关特性频繁出现且相似度较高，这些信息属性被称作高频数据的共现聚类^[6]。对高频数据集进行漂移点偏差挖掘计算，寻找高频数据集中的信

息特性，对数据进行提取并对共现度进行计算，由此实现共现聚类^[7]。

高频数据集在完成漂移点标定后，往往会发生一个数据集合被 2 个不同的数据属性所标记的现象，这种现象被称为数据 s_1 与 s_2 共现。当 2 个数据属性频繁标记一个数据集，通过对 2 个数据属性同时标记单一数据集的次数进行统计，就可以判定这 2 个数据标属性的共现程度。2 个数据属性标记记录同一个数据频繁程度与这 2 个数据的共现程度成正比^[8]。

数据属性 s_x 与数据属性 s_y 的共现度计算公式为：

$$R(s_x, s_y) = \frac{|\{ve_{v,x} > 0 \& e_{v,y} > 0\}|}{\text{AVERAGE}(|\{ve_{v,y} > 0\}|, |\{ve_{w,y} > 0\}|)}。 \quad (3)$$

式中： $R(s_x, s_y)$ 为数据属性 s_x 与 s_y 的共现度； $ve_{v,x}$ 为数据属性 s_x 标记数据集 v_v 的频繁程度。通过计算 n 个数据属性中的两两数据属性的共现程度，由此得到数据属性相似度矩阵 $\mathbf{S}_{t \times t}$ ， $\mathbf{S}_{t \times t}$ 组成因子等于 $R(s_x, s_y)$ 。

数据属性 s_x 与数据类别之间的共现关系可表示为：

$$R(s_x, o_n) = \frac{1}{n} \sum_{y=1}^n R(s_x, s_y)。 \quad (4)$$

式中： n 为数据属性个数； O 为由数据属性分类。由此获得数据属性与和数据类别之间的共现程度，然后得到数据属性和数据类别相似度矩阵 $\mathbf{L}_1$ ，矩阵的组成因子与 $R(s_x, o_n)$ 一致。

由此推断各个类别之间的共现程度，可表示为：

$$R(o_w, o_n) = \frac{1}{w * n} \sum_{x=1}^w R(s_x, s_y)。 \quad (5)$$

式中： o_w 为具有 w 属性的数据类别； o_n 为具有 n 属性的数据类别； $R(o_w, o_n)$ 为 o_w 类与 o_n 类的共现程度。在此基础上计算类别之间的共现度矩阵 $\mathbf{L}_2$ ，矩阵内的组成因子与 $R(o_w, o_n)$ 相拟合。

经过以上计算，解决了高频数据共现聚类过程中数据漂移问题，提高异构网络数据的分类能力^[9]。

2 异构网络信息增量式更新方法

通过上述漂移点偏差挖掘算法实现了异构网络高频数据的聚类操作，增强了信息分类能力。由于在异构网络体系中，经常采用信息增量来表示网络中的动态变化，随着使用人数和访问量的增加，网络的规模也在逐渐扩大，因此，一次发布全部的数

据已经不能满足对数据的实时获取和持续获取的要求, 必须在数据持续变化的情况下, 对其进行及时的增量更新和发布^[10]。

2.1 增量式更新相关性设置

笔者将高频信息的相关性特征, 利用关联特征选择法判定增量数据的相关性和伪相关性, 筛选与用户检索匹配度一致的真实信息进行更新发布^[11-12]。关联特征选择法是在符合关联规则 A 的基础上满足 B 值要求的算法, 通过将关键特征信息真伪判定实现数据的实时增量更新^[13]。首先确定被插入的数据集合, 划分为等效数据集合, 需要符合以下规则:

1) 伪相关数据判定原则 A : 设数据集为 X , 相关性特征集为 C , 在相关性特征集 C 内, 若相关性特征取值的最低门限为 W , 则 A 最大取值应小于等于 W 。

2) 真相关数据判定原则 B : 设数据集为 X , 相关性特征集为 $C(C_1, C_2, C_3, \dots, C_w)$, o 为 X 中包含相关性特征的个数, 则每个相关性特征伪相关门限最大值判定公式为:

$$sd_i = \max(sd_i^j) (1 \leq i \leq \min(C_1))。 \quad (6)$$

式中: C_j 是不同相关性特征 C_j 的个数 ($1 \leq j \leq o$); d_i^j 为升序排列相关性特征 C_j 的频次值 ($1 \leq j \leq o$ 且 $1 \leq i \leq C_j$); sd_i^j 为累加升序排列每个相关特征 C_j 频次值 ($1 \leq j \leq o$ 且 $1 \leq i \leq C_j$)。

由此推算可判定伪相关数据最低门限为:

$$\min C_2 = \lfloor (h - sd_{p-1}) / i \rfloor。 \quad (7)$$

式中 h 为待插入元素序列的个数。

设预备更新的数据集为 d_1^1 、 d_2^1 、 d_3^1 和 d_4^1 4 组, 每组 30 条信息内容, 将相关性属性对应取值为 C_1 、 C_2 、 C_3 、 C_4 , 其中相关性属性递增排序关系为 $C_1 < C_2 < C_3 < C_4$, 根据式(7)内容可判定为 C_2 为伪相关数据最低门限, 得到属性为 C_1 的元素序列伪相关数据为 8 组, 相关性属性为 C_2 的元素序列伪相关数据为 11 组, 相关性属性为 C_3 的元素序列伪相关数据为 4 组, 相关性特征为 C_4 的元素序列伪相关数据为 7 组, 将数据判定结果整理, 得到表 1 和 2。

表 1 伪相关特征频次统计

伪相关性特征	$d_1^1 < C_1$	$d_2^1 < C_2$	$d_3^1 < C_3$	$d_4^1 < C_4$
频次	8	11	4	7

表 2 真相关特征累计频次统计

真相关性特征	$d_1^1 > C_2$	$d_2^1 > C_2$	$d_3^1 > C_3$	$d_4^1 > C_4$
频次	22	19	26	23

通过上述研究, 采用关联特征选择法判定信息真伪相关性, 能够改善信息在增量更新发布过程中存在的虚假信息。

2.2 异构网络信息增量更新流程

将待插入元素序列的相关性特征进行关联, 与临时列表中等效特征集合元素序列对比, 将匹配成功的待计算元素序列的相关性特征子集识别并增量式更新, 具体流程如图 1 所示。

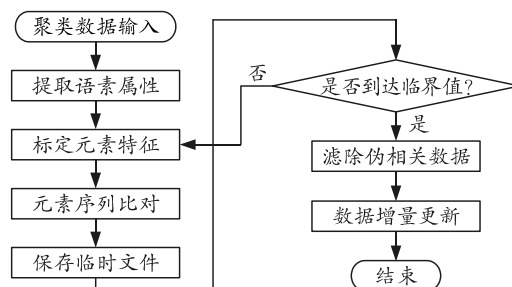


图 1 算法流程

根据图 1 可知: 提取待更新信息的语素属性, 将相关性特征数据判定, 保证数据在临时列表中的隐秘性。判断待插入元素序列是否与临时列表类目匹配, 匹配成功到达临界值时对相关性特征数据识别并进行更新, 匹配失败的元素序列重新判定并保存至临时列表中, 等待下次数据更新。如元素序列长期匹配失败, 则将其视为伪相关数据进行删除^[14]。

上述操作可确保更新数据的有效性, 在保证数据质量的基础上实现了异构网络数据的增量更新^[15]。

3 实验

3.1 高频漂移数据聚类验证

为了验证笔者提出的基于高频数据共现聚类算法的异构网络信息增量式更新方法, 进行实验。设定更新时间为 10 s, 采用 Intel core i7-930CPU 2.8 GHz 处理器, 操作系统为 Windows10 企业版。实验随机抽取某网站 4 类已发布高频数据集为实验样本, 每类数据为 50 条数据内容, 原始数据集分布状态, 如图 2 所示。

分别使用生成对抗网络方法、基于聚类分析方法与本文中提出基于高频数据共现聚类算法的异构网络信息增量式更新方法进行聚类实验对比, 对比

结果如图 3 所示。

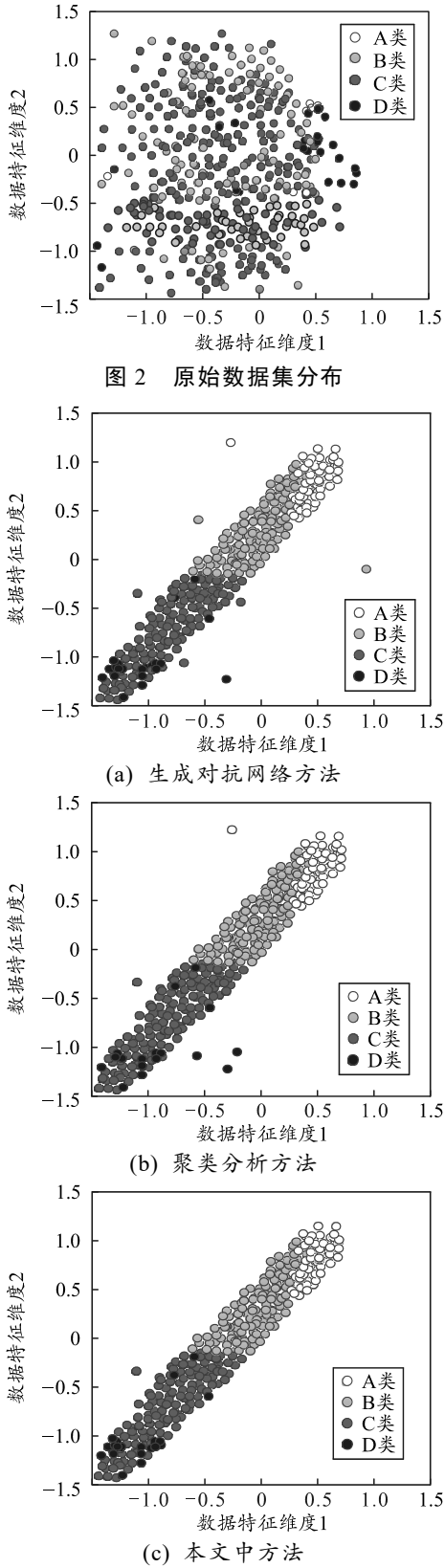


图 3 3 种方法聚类对比结果

由图 3 可知：基于生成对抗网络方法在进行数据聚类后存在 7 个漂移数据点，聚类占比为 96.5%，

聚类效果较差。基于聚类分析方法在进行数据聚类后仍存在 5 个漂移数据点，聚类占比为 97.5%，聚类效果仍需进一步提升。本文中方法在数据聚类后仅存在一个漂移数据点，聚类占比为 99.5%，说明该方法可以使漂移数据有效收敛，具有良好的聚类效果。

3.2 异构网络数据更新的相关性验证

为了进一步证明本文中研究的可行性，设定 4 组待更新增量数据集，每组真相关数据为 2 486 条，伪相关数据为 14 条，导入生成对抗网络方法(文献[1])、基于聚类分析方法(文献[2])与本文中方法进行异构网络条数更新的相关性验证，结果如表 3 所示。

表 3 异构网络数据更新的相关性对比结果

相关性判定	额定条数	文献[1]方法	文献[2]方法	本文中方法
真相关性	9 944	9 950	9 947	9 945
伪相关性	56	50	53	55

由于待更新数据中的伪相关数据判定难度大于真相关数据判定难度，故笔者对比实际检测出的伪相关数据与额定伪相关数据的查准率占比。根据表 3 可知，基于生成对抗网络方法检测出的伪相关数据为 50 条，查准率占比为 89.2%，说明该方法在数据增量更新过程中容易受到虚假信息误导，数据的真实性不高。基于聚类分析方法检测出的伪相关数据为 53 条，查准率占比为 94.6%，该方法对于伪相关数据有一定的判定能力，但数据的可信性难以达到实际需求。而本文中方法检测出的伪相关数据为 55 条，查准率占比为 98.2%，能够从大量数据中准确判定伪相关数据，使更新数据与原始数据相拟合，具有较高的可靠性。

4 结论

基于高频数据共现聚类算法的异构网络信息增量式更新方法，通过实验分析，得出如下结论：

- 1) 设定高频数据共现聚类算法参数及规则，为实现数据聚类提供数据基础。
- 2) 通过网格划分搜索高频数据集中漂移点，从而挖掘数据集中的异常数据。
- 3) 采用关联属性选择算法，判定数据的相关性和伪相关性。
- 4) 根据代表性语素判定结果，提取与用户检索匹配度一致的真实信息进行更新发布。

本文中研究在异构网络增量式更新过程中能够

提高更新数据与原始数据的相关性,在此基础上有效实现高频数据共现聚类,具有较高的可行性。后续研究将进一步扩大实验样本规模,以达到更加准确的聚类效果。

参考文献:

- [1] 聂为之,王岩,杨嵩,等. 基于循环生成对抗网络的跨媒体信息检索算法[J]. 计算机学报, 2022, 45(7): 1529-1538.
- [2] 聂静,常涛,刘维,等. 基于聚类分析的个性化异构数据发布[J]. 科学技术与工程, 2021, 21(14): 5813-5821.
- [3] 田夏利,熊莹. 融入新的属性选择机制的文本数据聚类算法[J]. 计算机工程与设计, 2021, 42(3): 734-741.
- [4] 廖彬,黄静莱,王鑫,等. SCEA: 一种适应高维海量数据的并行聚类集成算法[J]. 电子学报, 2021, 49(6): 1077-1087.
- [5] 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.
- [6] 王雪,刘京,孙佳妮,等. 基于谱聚类的异构蜂窝超密集网络高效资源分配算法[J]. 通信学报, 2021, 42(7): 162-175.
- [7] 刘建航,武万萍,王树庆,等. 一种异构网络多协议并行数据分发方法[J]. 电子与信息学报, 2021, 43(6):

1715-1723.

- [8] 王志刚,王辰阳,李恒武. 多维异构网络数据智能预加密推送方法研究[J]. 计算机仿真, 2022, 39(6): 417-420, 463.
- [9] 屠莉,陈峻. 衰减窗口中的不确定数据流聚类算法[J]. 计算机应用研究, 2021, 38(9): 2673-2677, 2682.
- [10] 张春英,高瑞艳,范雨祥,等. 一种面向不完备数据的集对粒层次聚类算法[J]. 小型微型计算机系统, 2021, 42(3): 522-530.
- [11] 洪金鑫,吴英杰,蔡剑平,等. 基于属性分割的高维二值数据差分隐私发布[J]. 计算机研究与发展, 2022, 59(1): 182-196.
- [12] 吴万青,赵永新,王巧,等. 一种满足差分隐私的轨迹数据安全存储和发布方法[J]. 计算机研究与发展, 2021, 58(11): 2430-2443.
- [13] 罗洪斌,张珊,王志远. 共生网络——异构网络安全高效互联的体系结构与机理[J]. 通信学报, 2022, 43(4): 36-49.
- [14] 马天祥,贾伯岩,范伟,等. 一种基于博弈论的 5G 异构网络多目标优化干扰算法[J]. 电讯技术, 2022, 62(6): 702-710.
- [15] 龙恩,李伟,鲁江丽,等. 基于匹配理论的非正交多址异构网络资源分配算法[J]. 计算机工程, 2021, 47(1): 165-171.

(上接第 18 页)

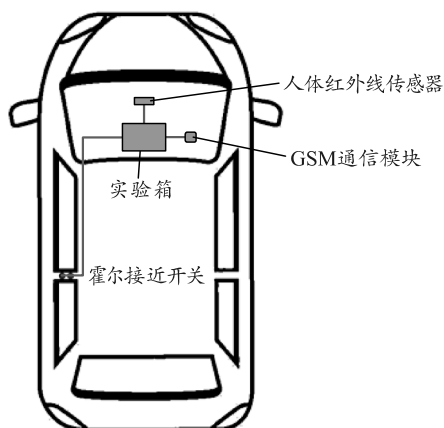


图 5 系统安装

5 结束语

该系统借助热释电人体红外感应技术,依靠人体发出的特定波长来检测是否有人遗留车内,并且可通过调节探测距离和安装角度,实现只探测车内部而不受车外部活动的人和实物的影响,灵敏度高,误报、漏报率极低。采用 GSM 通信技术,信号覆盖范围极广,基本无信号盲区,并可实现远距离报警,儿童发生险情时,即使驾驶员不在现场,也能

被系统发送的短信或电话及时提醒,为营救被困车内儿童赢取更多时间。利用霍尔传感器检测汽车车门是否关闭,可以更科学精确地探测车内状况,避免误报。采用温度传感器来获取车内实时温度,测温范围广,精度高,设定阈值后,一旦超标即进入警戒状态,有效避免了儿童夏天被遗忘车内由于高温窒息或冬季温度过低威胁生命安全的情况。系统采用单片机控制,各传感器均为低功耗,整个装置耗电极低,且汽车启动时自动给装置充电,续航时间持久,基本可实现不间断运行,成本低廉,性能可靠,而且易于安装,无需拆车,对提升乘车儿童的生命安全保障有巨大的现实意义。

参考文献:

- [1] 夏若宸. 基于 STM32 的儿童误锁车内远程报警系统[D]. 杭州: 浙江海洋大学, 2021.
- [2] 舒强,张亚南. 加强儿童乘车安全管理对策建议研究[J]. 汽车与安全, 2017(2): 104-107.
- [3] 吴淑霞. 儿童乘员乘车安全分析[J]. 汽车工程师, 2018(12): 12-13.
- [4] 毕廷良. 基于 GSM 的儿童反锁车内远程报警系统设计[D]. 上海: 上海应用技术大学, 2017.