DOI:10.20033/j.1003-7241.(2025)12-0147-04

基于孤立森林的多源加权序列数据挖掘方法

李毓丽
(广州软件学院,广东 广州 510990)

摘　要:研究旨在解决多源加权序列数据集中存在的噪声和缺失值问题,以提高数据挖掘性能,提出一种基于孤立森林的多源加权序列数据挖掘方法。首先,利用相关系数度量多源数据间的相似性,并进行多源融合处理。接着,采用 MAP 方法去除数据噪声,并根据序列匹配模式进行数据平滑。然后,基于孤立森林构建数据挖掘模型,并优化其内置算法,以处理多源加权序列数据。最后,通过车流量信息数据库序列的降维拟合实验,验证所提方法的有效性。实验结果表明,该方法能显著提高拟合优度,从而验证了其在多源加权序列数据挖掘中的有效性。

关键词:孤立森林;加权序列;多源融合;多源序列;序列匹配;数据挖掘
中图分类号:TP311.13　　**文献标志码**:A　　**文章编号**:1003-7241(2025)12-0147-04

Multi-source weighted sequence data mining method based on isolated forests

LI Yuli
(School of Software Engineering Institute of Guangzhou, Guangzhou 510990, China)

Abstract: This study aims to address the issues of noise and missing values in multi-source weighted sequence datasets, in order to improve data mining performance. A multi-source weighted sequence data mining method based on isolated forests is proposed. Firstly, it uses correlation coefficients to measure the similarity between multi-source data and perform multi-source fusion processing. Next, the MAP method is used to remove data noise, and data smoothing is performed based on sequence matching patterns. Then, a data mining model is constructed based on isolated forests, and its built-in algorithms are optimized to process multi-source weighted sequence data. Finally, the effectiveness of the proposed method is verified through dimensionality reduction fitting experiments on the sequence of the traffic flow information database. The experimental results show that this method can significantly improve the goodness of fit, thus verifying its effectiveness in multi-source weighted sequence data mining.

Keywords: isolated forests; weighted sequence; multi source fusion; multiple source sequences; sequence matching; data mining

0　引言

多源融合与数据挖掘是大数据分析与挖掘的核心内容之一,在复杂背景下多源数据的处理和挖掘技术面临诸多挑战,如多源数据的融合[1]、多源数据的特征提取、多源数据的融合等[2]。目前很多学者针对多源加权序列数据挖掘方面展开了研究,如文献[3]中探讨教育数据挖掘中的不平衡分类问题,利用随机过采样、随机欠采样及混合重采样技术提升预测准确性。实验表明,随机过采样对中度不平衡数据、混合重采样对极度不平衡数据效果最佳。文献[4]中提出基于 FP-Growth 关联规则算法的库岸边坡监测数据挖掘方法,解决数据归一化难题,通过因果关联规则和空间关联规则挖掘数据间的潜在因果关系和多测点效应量关联性,提取有价值信息。文献[5]中计及供给侧出力,对数据挖掘负荷预测展开研究,先聚类气象数据,再利用灰色关联分析挖掘气象因素与供电侧的关系,并通过支持向量机算法完成负荷预测分析。

以上方法在教育数据挖掘中取得了一定效果,但在实际应用中,对含有噪声或缺失值等问题的多源加权序列数据集时,这些方法的数据挖掘性能会有所降低。孤立森林在数据挖掘领域中是一种非常流行的序列挖掘方法,它在数据库分析、文本挖掘、社交网络分析等领域都有着广泛的应用,因此本文设计一种基于孤立森林的多源加权序列数据挖掘方法。

1　多源加权序列数据挖掘方法研究

1.1　多源加权序列数据相似性度量

在多源加权序列数据挖掘中,不同的数据源通常具有不同的属性,但它们在数据量和计算成本上存在相似之处,因此多源加权序列数据的相似性度量通常由以下两种形式定义:1)多源序列的样本点之间的距离定义为样本点到相似度矩阵的距离;2)多源序列的样本点之间的相似度矩阵定义为样本点到相似度矩阵的距离[6-7]。上述两种形式认为,前一种是对多源加权序列数据进行聚类之前进行的预处理步骤,即对多个数据源进行比较和分类,将多源数据分离。后一种形式则是在聚类过程中使用。在进行相似性度量的过程中,使用相关系数对两个随机变

*基金项目:广东省教育厅特色专业项目(JXTD201901)
收稿日期:2024-05-27

量之间线性关系强弱进行描述[8]。计算公式为

$$\rho_{XY}=\frac{\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})}{\sqrt{D(\boldsymbol{X})}\ \sqrt{D(\boldsymbol{Y})}} \tag{1}$$

式中，$\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})$ 表示多源加权序列 $\boldsymbol{X}$、$\boldsymbol{Y}$ 的协方差，$D(\boldsymbol{X})$、$D(\boldsymbol{Y})$ 多源加权序列 $\boldsymbol{X}$、$\boldsymbol{Y}$ 的方差。多源加权序列相似性度量本文使用的是基于权重的朴实距离度量法，假设每个多源加权序列数据集合由多个不同类型、不同数量和不同质量的样本点构成[9]，所以可以使用以下公式来计算多源加权序列数据之间的相似性为

$$L(\boldsymbol{X},\boldsymbol{Y})=\sqrt{\sum_{t=1}^{n}\left(w_X\boldsymbol{X}_t-w_Y\boldsymbol{Y}_t\right)^2} \tag{2}$$

式中，w_X、w_Y 表示多源加权序列 $\boldsymbol{X}$、$\boldsymbol{Y}$ 数据的权重，$\boldsymbol{X}_t$ 表示多源加权序列 $\boldsymbol{X}$ 数据集合；表示 $\boldsymbol{Y}_t$ 表示多源加权序列 $\boldsymbol{Y}$ 数据集合，n 为序列长度。多源加权序列相似性度量中，样本点之间的相似度矩阵在多源加权序列数据集合中选取 k 个样本点作为计算相似度矩阵的基础[10]；通过对所有样本点计算其到相似度矩阵中对应元素的权重并进行加权求和。当 n 趋于无穷大时，由于多源加权序列数据集合中样本点之间距离会随着 n 增加而趋于无穷大，因此针对多个样本点之间距离较大时，具有较强的稳定性。

1.2 数据平滑处理

在多源序列数据挖掘中，首先要对数据进行预处理，即去除噪声。在孤立森林算法中，使用最大后验概率 MAP 方法来实现噪声数据的去除[11-12]。MAP 是一种基于贝叶斯定理的最大后验概率估计方法，MAP 可以用于任意大的数据集上，因为它可以处理非高斯和连续分布数据。在此，将对序列进行平滑处理，即

$$D_{\mathrm{dtw}}(\boldsymbol{X},\boldsymbol{Y})=D_{\mathrm{base}}[(\mathrm{head}),\mathrm{head}]+\begin{cases}D_{\mathrm{dtw}}[\boldsymbol{X},\ \mathrm{rest}(\boldsymbol{Y})]\\ D_{\mathrm{dtw}}[\mathrm{rest}(\boldsymbol{X}),\boldsymbol{Y}]\\ D_{\mathrm{dtw}}[\mathrm{rest}(\boldsymbol{X}),\mathrm{rest}(\boldsymbol{Y})]\end{cases} \tag{3}$$

式中，$D_{\mathrm{dtw}}(\boldsymbol{X},\boldsymbol{Y})$ 是原始序列中所有元素的最大后验概率，$D_{\mathrm{base}}[(\mathrm{head}),\mathrm{head}]$ 是原始序列中元素的最小后验概率，rest 是原始序列中元素的方差[13]。根据经验和实验可知，对于具有不同类型和大小的序列数据，MAP 估计会产生不同的结果。在多源序列数据挖掘中，可以根据数据源类型、大小和分布等因素来选择合适的 MAP 方法。在进行平滑处理的过程中，两个序列之间会产生一定的匹配，如图 1 所示。

从上图中可以看出，当数据量较小时，MAP 估计会产生较大的估计结果；当数据量较大时，MAP 估计会产生较小的估计结果；而当数据量很大时，MAP 则不会产生任何结果。

1.3 建立基于孤立森林的数据挖掘模型

在建立模型之前，需要定义一些基本的属性，其中最重要的是模型的输入参数。例如，一个序列中的第一个元素代表了序列中所有可能出现的位置，因此需要定义第一个参数为第一个元素。当执行此操作时，将首先根据一个初始的参数（例如，第二个元素）创建一个孤立森林模型，该模型包含了所有可能出现在序列中的元素[14]。然后，需要定义一个参数（例如，第三个元素）来表示第一个参数与第二个参数之间的关系。在此之后，需要创建一个函数来表示该参数与第三个和第四个元素之间的关系。在所有这些属性之后，可以使用 MATLAB 等工具来实现简单的孤立森林算法。模型如图 2 所示。

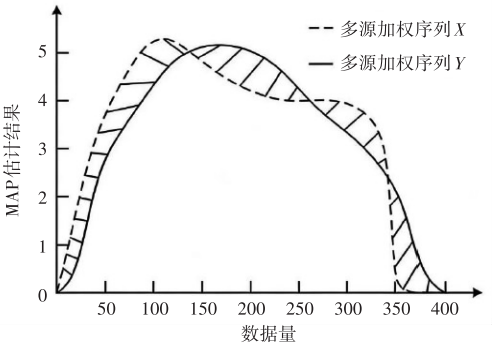

图 1 序列匹配示意图

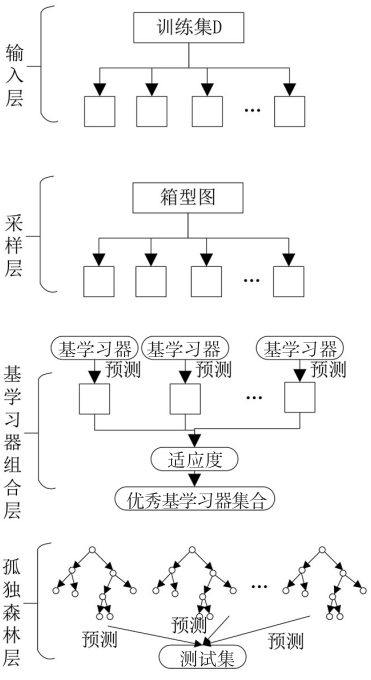

图 2 基于孤立森林的数据挖掘模型

在建立孤立森林模型后，还需要对模型进行预处理。预处理步骤包括以下两个步骤：1）优化模型内置算法；2）将多源数据转化为单一源数据。使用 MAP 方法来估计数据集的均值和方差[15]。在算法优化的过程中，可以通过定义一个加权矩阵来计算该序列中不同元素之间的相关系数 C_{ij}。用户序列为 C，i 为元素编号，j 表示元素位置，对于每个元素来说，计算该元素与第 i 个元素之间的相关系数为

$$C_i=\frac{1}{\sum i}=p_j(ij) \tag{4}$$

将每个元素与第 i 个元素进行加权得到多源加权序列。对于不同序列的有序事务位置向集,可以分别计算这两个特征之间的相关系数 C_{ij}。然后可以使用 MSWFST 算法对多源加权序列进行聚类以获得序列模式。在序列的扩展过程中,假设其中一个序列 X_1 为频繁序列,即

$$X_1 = \langle w_1, w_2, \cdots, w_n \rangle \tag{5}$$

在该序列上进行扩展,得到新的序列,有

$$X_1 = \langle w_1, w_2, \cdots, w_n, \{i_x\} \rangle \tag{6}$$

经过扩展之后,将其中的数据按照用户进行分组处理,并将权值引进到序列模式挖掘。经过以上算法,搭配相关的模型,实现多源加权序列的数据挖掘。

2 仿真实验

2.1 实验环境与数据集

为了验证本文设计的基于孤立森林的多源加权序列数据挖掘方法的有效性,实验数据的选择来源于典型的多源加权序列:某一路段车流量信息形成的数据库,随机选择三组序列进行降维拟合仿真实验。本次仿真实验的运行环境选择如下:操作系统为 Win10 专业版,内存 64 G,硬盘容量为 1 TB,利用 C#开发语言进行编程。监测数据点的选择按照时间信息,间隔 15 min,共 96 个数据点,时间设定为 6:00 至 21:00。另外需要考虑到不同时段车流量的波动情况,因此要考虑不同阈值状态的拟合效果。本文序列数据挖掘方法还设置了转折角度和均线距离等条件进行细节补充确保拟合效果。

2.2 实验结果对比与分析

在对拟合效果进行判定时,拟合效果的评估通常采用拟合优度进行评估,计算公式为

$$R = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\sum_{i=1}^{n} (y_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y_i})^2} \tag{7}$$

式中,y_i 代表序列点的真实值;y_i 代表序列点的拟合值;$\overline{y_i}$ 代表序列点的平均;拟合效果的数值区间为[0,1],在此区间数值越大拟合效果越好。

针对几种不同的序列数据挖掘方法在同一阈值下不同的时间数据点的仿真实验结果如下。

表 1 $K_0=2$ 时,不同方法拟合情况

方法名称	数据序列		
	5 月 10 日	7 月 20 日	9 月 30 日
传统交叉方法	0.727	0.749	0.715
本文加权方法	0.976	0.983	0.972

由上表中的实验结果可以看出,虽然两种方法都可以完成序列数据的拟合,但是拟合优度存在较大差距;本文设计孤立森林加权方法整体的拟合优度保持较高水平,较传统的交叉验证方法高出 35.7%,且拟合效果较为稳定表现更优异。

为了进一步评估不同序列数据挖掘方法的拟合效果,分别对处在不同阈值及时段的数据进行对比分析,实验结果如下所示。

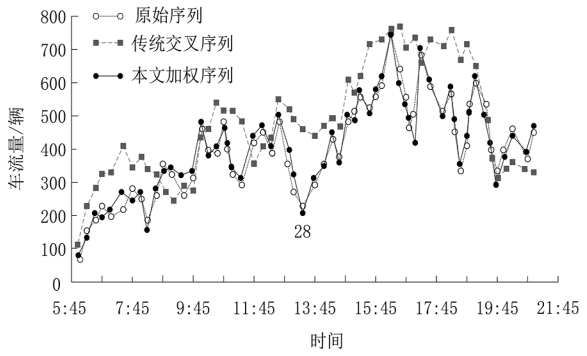

图 3 $K_0=2$ 时,5 月 10 日车流量拟合效果对比图

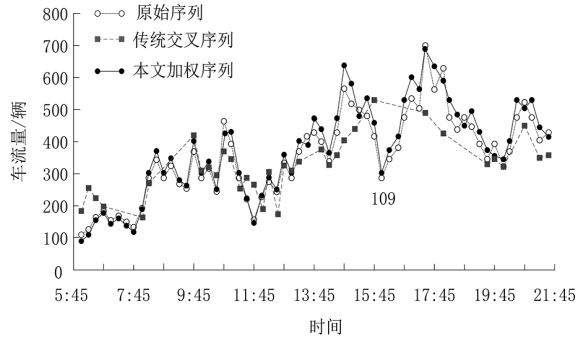

图 4 $K_0=2$ 时,7 月 20 日车流量拟合效果对比图

观察以上两图可以看出,传统的交叉序列方法在处于变换较为频繁时间段进行数据点信息提取时,存在较大部分数据点不满足极值点保持时间要求,因此被舍弃掉,较为明显地可见上图中不同取样时间标记的 28 和 109 点。原始的时序列在日常实际状态下波动情况与试验情况相似,极值点数量也较多,传统的交叉数据挖掘方法由于不满于阈值条件,大量数据信息丢失,无法真实地识别还原序列情况,无法满足需求。

本文设计的基于孤立森林的多源加权序列数据挖掘方法对于关键特征点的识别,不但准确性较高,数据信息相对完整,可以很好地保留关键点和转折点,更有利于序列的后续分析。

考虑到阈值可能对不同序列数据挖掘方的拟合优度值产生一定的影响,对阈值 $K_0=3$ 时拟合度情况进行对比分析,实验结果如下所示。

表 2 $K_0=3$ 时,不同方法拟合情况

方法名称	数据序列		
	5 月 10 日	7 月 20 日	9 月 30 日
传统交叉方法	0.542	0.497	0.503
本文加权方法	0.958	0.971	0.964

与表 1 中数据进行对比可以看出,当阈值改变拟合优度也相应改变,当 K_0 增大时,拟合优度呈现下降趋势,传统交叉方法整体下降较为明显在 28.6%左右;本文设计的孤立森林多源加权序列数据挖掘方法虽然也受到一定

程度的影响,但是波动较小,且仍然保持较高的拟合优度值,稳定性较高。

对两种挖掘方法的适用性和稳定性进一步验证得到如下的实验结果,即

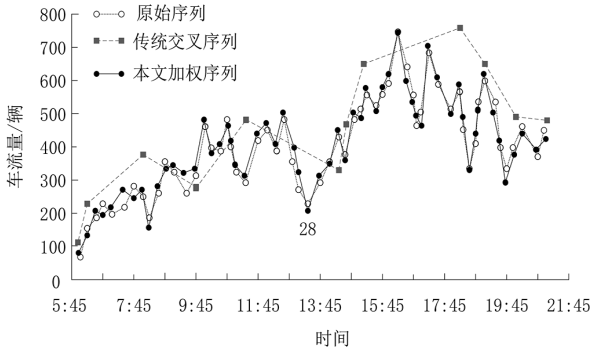

图 5 $K_0=3$ 时,5 月 10 日车流量拟合效果对比图

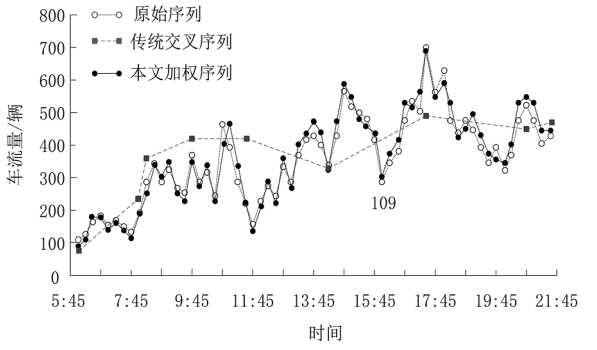

图 6 $K_0=3$ 时,7 月 20 日车流量拟合效果对比图

当阈值设定为 3 时,与前文图 1 和图 2 的实验结果进行对比,可以看出阈值的改变对于传统的交叉序列挖掘方法影响较为明显,在波动频繁的序列区间失去识别功能,丧失了较大部分的极值点,整体缺陷较为突出,表现极差,进而失去了预设的降维拟合功能,不具有参考性。

本文设计的基于孤立森林的多源加权序列数据挖掘方法由于具有特定的补充筛选条件,在波动频繁时间段进行细化分析补充,可以很好地对关键点和转折点均进行保留,不同的阈值条件下拟合曲线几乎无改变,可以较好的保留序列数据的特征,可以避免条件改变对实验结果带来的影响,适用性更强更稳定。

3 结束语

在多源序列数据的挖掘中,孤立森林是一种广泛使用的方法,本文提出了一种基于加权孤立森林的多源序列挖掘方法,它能够有效地处理多源数据,实验结果表明,该算法在处理大规模数据时表现出良好的性能。但该方法也有一些缺点,如对噪声敏感,难以处理大规模数据等。在今后的研究中,还有一定的优化空间。

参考文献

[1] 白帆, 李雪贞, 马国学, 等. 基于时间序列分析的环境 γ 辐射剂量率数据预处理方法研究及评估[J]. 辐射防护, 2023, 43(2):128-136.

[2] 任其亮, 徐韬, 刘媛, 等. 考虑载客状态的改进孤立森林浮动车异常数据检测算法[J]. 交通运输系统工程与信息, 2024, 24(1):124-131.

[3] WONGVORACHAN, TARID, SURINA HE, AND OKAN BULUT. A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining[J]. Information, 2023, 14(1):54.

[4] 陈波, 詹明强, 黄梓莘. 基于关联规则的库岸边坡监测数据挖掘方法[J]. 长江科学院院报, 2022, 39(8): 58-64.

[5] 刘庆彪, 张桂红, 许德操, 等. 计及供给侧出力的数据挖掘负荷预测方法[J]. 沈阳工业大学学报, 2022, 44 (3):259-264.

[6] YUAN, LIYUN, AND JIAXING CAO. Application of data mining in female sports behavior prediction based on FCM algorithm[J]. Soft Computing, 2023, 27(14):10045-10055.

[7] 汪力纯, 刘水生. 基于混合采样和特征选择的改进随机森林算法研究[J]. 南京邮电大学学报(自然科学版), 2022, 42(1):81-89.

[8] 祝和明, 蔡榕, 周长江, 等. 基于融合指标的电力专利可信数据挖掘方法研究[J]. 自动化技术与应用, 2024, 43(3):139-142, 164.

[9] 李瑞峰, 杨海峰, 蔡江辉, 等. 一种基于加权深度森林的离群数据挖掘算法[J]. 小型微型计算机系统, 2022, 43(7):1426-1431.

[10] 张淏凌, 冯杰成, 陈炳海, 等. 基于时空数据挖掘的无人机航线自动寻优规划[J]. 电子设计工程, 2024, 32 (8):153-156, 161.

[11] 李跃辉, 方愉冬, 徐峰, 等. 基于关联数据挖掘的继电保护定值风险评估方法研究[J]. 科学技术与工程, 2023, 23(24):10355-10361.

[12] 赵林锁, 陈泽, 丁琳琳, 等. 基于 RELM 的时间序列数据加权集成分类方法[J]. 计算机工程与科学, 2022, 44 (3):545-553.

[13] 曹兴武, 姚頔, 孙樊荣, 等. 基于雷达数据挖掘的空域扇区规划方法[J]. 北京航空航天大学学报, 2023, 49 (12):3237-3244.

[14] 黄光球, 赵羲轩, 陆秋琴. 基于 KPCA-IF-WRF 模型的多源 VOCs 数据清洗方法研究[J]. 安全与环境学报, 2022, 22(6):412-3423.

[15] 蔺万科, 宋华, 南新元, 等. 一种基于最优聚类中心与权重欧式距离的多源异质传感器数据融合方法[J]. 传感技术学报, 2022, 35 (1):49-56.

作者简介:李毓丽(1980—),女,硕士,副教授,研究方向:计算机应用,Web 应用开发,数据分析与数据挖掘。

(上接第 114 页)

[13] LIU W, ANGUELOVD, ERHAN D, et al. SSD: sing leshot multibox detector [C]//European conference on computer vision (ECCV). Amsterdam, The Netherlands: Springer, 2016:21-37.

[14] TAN M, LE Q V. EfficientNet: rethinking model scaling for convolutional neural networks[C]// Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019:6105-6114.

[15] 钱志杰、胡恩德,等. 基于图像识别的电力安全工器具智能管理系统设计[J]. 中国信息化,2021(12):53-54.

[16] 韩路,高社民,等. 智能工器具柜三重身份验证功能的实现[J]. 数字通信世界, 2023(5):54-55.

作者简介:韩文芝(1982—),男,本科,副高级工程师,研究方向:电网智能运检、设备智能运维。

通信作者:李 锐(1983—),男,本科,中级工程师,研究方向:电网智能运检、设备智能运维。