

基于多示例学习的时序离群点检测算法研究

钱景辉, 卢屹韦

(南京工业大学 计算机科学与技术学院, 江苏 南京 211816)

摘 要: 在 FIndCBLOf 算法的基础上, 提出了一种基于多示例学习的 FindCBLOF 离群点检测算法 (MIL-FindCBLOf) 用于时序离群点检测。算法首先对数据进行分段聚集, 再利用多示例框架, 封装每个对象, 以此保留每个对象的属性, 然后采用全局策略计算对象的因子数值, 最后通过计算平均因子来确定离群序列。在实际企业的实时采集监控系统中, 将 MIL-FindCBLOf 算法与其它经典离群点检测算法进行实验对比, 结果表明本文提出的 MIL-FindCBLOf 算法相对其它算法提高了检测的全面性和准确性。

关键词: 机器学习; 时序离群点; 多示例学习; 聚类; 平均因子

中图分类号: TP181

文献标识码: A

文章编号: 1000-7180(2018)01-0046-04

Multi-instance-learning Based Time Series Outlier Detection Algorithm

QIAN Jing-hui, LU Yi-wei

(College of Computer Science and Technology, Nanjing Tech University, Nanjing 211816, China)

Abstract: On the basis of the FIndCBLOf algorithm, this paper proposes a FindCBLOF outlier detection algorithm (MIL-FindCBLOf) for Multi instance learning based the detection of time series outliers. Firstly, the data were gathered using piecewise multi instance framework to package each object in order to preserve the properties for each object, to calculate the object factor with global strategy, to determine the outlier sequence by calculating the average factor. In the real-time monitoring system of practical enterprise MIL-FindCBLOf algorithm will be compared with other classical outlier detection algorithm by experiments. The results show that the proposed MIL-FindCBLOf algorithm compared with other algorithms in this paper can improve the accuracy and comprehensiveness of detection.

Key words: machine learning; sequential outlier; multi instance learning; clustering; average factor

1 引言

随着信息数据的发展, 信息量的增大和丰富, 离群点分析在数据挖掘模式中的作用也越来越引起一些专家和学者们的关注。离群点是一些行为目标和正常数据非常不同的对象, 一些数据挖掘方法在挖掘数据时, 往往会把这些对象当作是噪声或者是异常而舍弃掉, 但在金融市场、系统监测^[1]、生物学数据、用户行为序列等研究中, 这些罕见的离群点往往又是行业人士所感兴趣的, 它们的存在也暗示了正常的环境发生了失衡。因此, 离群点检测也在这些行

业中发挥着至关重要的作用。

然而在一些比较复杂的数据中, 离群点往往彼此之间有相似的特性, 并且会自行聚集成一个簇, 因此, 文献[2]一种 FindCBLOF 算法, 该算法能够通过识别小簇或者稀疏簇来判定离群点。这样, 可以识别一些不同于正常的特殊簇集, 提高检测的全面性。但随着数据内容呈现多义且复杂, FindCBLOF 算法在识别这些数据时, 往往把数据所有的特征堆杂在一个对象中来进行识别, 这忽略了对对象本身内部的属性和对象与对象之间的关联, 使得其聚类的效果有所下降, 导致了其准确率不高。

多示例学习(MIL)是可以对复杂的、多语义的对象进行计算^[3],这可以对一些传统的算法进行优化改进^[4].因为多示例学习可以充分考虑对象本身的各个属性,在保留属性不缺失的情况下,对对象进行学习^[5],这样对计算结果的准确率有着重要作用. J. Wang 和 J. -D. Zucker^[6]采用 Hausdorff 距离,扩展了 k-邻近算法,使之能够处理多示例学习问题;G. Ruffo^[7]扩展了决策树,构造了多示例决策树;Z. -H. Zhou 和 M. -L. Zhang^[8]通过使用自定义的误差函数反向传播神经网络,对 BP 神经网络算法进行了扩展,构造了多示例学习算法 BP-MIP.

本文提出了一种基于 MIL(多示例学习)的 FindCBLOF 算法(MIL-FindCBLOF),充分保留对象的各个属性,独立考虑每个对象每个属性来提高此算法的准确率.

2 MIL-FindCBLOF

2.1 相关定义

定义 1 示例包: 设集合 $S = \{S_1, S_2, S_3, \dots, S_i, \dots, S_n\}$, $n \in N$, n 为数据的总个数, S_i 即为一个多示例包, S 为多示例包集合.

定义 2 示例: 设集合 $S_i = \{S_{i1}, S_{i2}, S_{i3}, \dots, S_{ij}, \dots, S_{il}\}$, $S_i \in S$, $l \in N$, l 为示例包中示例的总个数, S_{ij} 即为示例包中的一个示例.

定义 3 特征属性: 设集合 $S_{ij} = \{S_{ij1}, S_{ij2}, S_{ij3}, \dots, S_{ijm}, \dots, S_{ilk}\}$, $S_{ij} \in S_i$, $k \in N$, k 为特征的总个数, $\{S_{ij1}, S_{ij2}, S_{ij3}, \dots, S_{ijm}, \dots, S_{ilk}\}^T$ 即为 S_{ij} 的特征属性.

定义 4 若 $C = S^T$, 则定义 C 为 S 对应的示例层集合, $B = \{C_{ij} | C_{ij} \in C_i, C_i \in C\}$, B 为示例层的集合, 称之为示例集.

以上定义的多示例数据框架如图 1 所示.

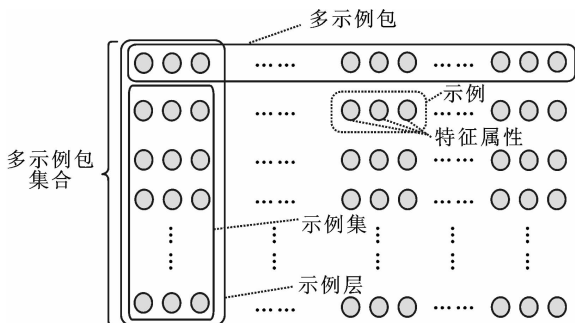


图 1 多示例数据框架图

定义 5 相似度函数: 相似度, 函数采用 Z. He 等人提出的 Sim 函数, 设 L 为包 B 中的一个示例, 则

示例 L 的相似度函数定义为 $\text{Sim}(C, L) = \frac{\sum_{a_i \in \text{VAL}_i(C)} \text{Sup}(a_i)}{\sum_{a_j \in \text{VAL}_i(C)} \text{Sup}(a_j)}$, $L, A_i = a_i, C$ 为一个已经聚类好的簇, m 为示例 L 的特征个数, $\text{Sup}(a_i)$ 表示在 C 中所有示例的第 i 种特征值为示例 L 的第 i 种特征值 a_i 的个数, $\sum_{a_j \in \text{VAL}_i(C)} \text{Sup}(a_j)$ 表示 C 的示例集合中所有示例的第 i 种特征值的总个数, $L, A_i = a_i$ 表示示例 L 的第 i 种特征值为 a_i .

定义 6 LC 和 SC: LC 和 SC 采用 Z. He 等人提出的定义, 对示例通过其 Squeezer 算法聚类后, 设示例簇 $C = \{C_1, C_2, C_3, \dots, C_k\}$, 且 $|C_1| \geq |C_2| \geq |C_3| \geq \dots \geq |C_k|$, 定义 $\begin{cases} |C_1| + |C_2| + |C_3| + \dots + |C_a| \geq |D| * \alpha \\ |C_a| / |C_{a-1}| \geq \beta \end{cases}$, a 为大簇和小簇的分割边界, α 为数量级参数, β 为大小参数, 则

$$\text{LC} = \{C_k | k \leq a\} \quad (1)$$

$$\text{SC} = \{C_k | k > a\} \quad (2)$$

定义 7 CBLOF(t): CBLOF(t) 采用 Z. He 等人提出的定义, 对于单个示例 L ,

$$\text{CBLOF}(L) =$$

$$\begin{cases} |C_i| * \min(\text{dis}(L, C_j)), L \in C_i, C_i \in \text{SC} \text{ 且 } C_j \in \text{LC} \\ |C_i| * \text{dis}(L, C_i), L \in C_i \text{ 且 } C_i \in \text{LC} \end{cases} \quad (3)$$

定义 8 平均因子: 对于示例包 S_i 中的所有实例, 其平均因子 $\text{Avg}(S_i)$ 给出定义为:

$$\text{Avg}(S_i) = \sum_{S_{ij} \in S_i, j=1}^l \omega_j * \text{CBLOF}(S_{ij}) \quad (4)$$

ω_j 为 S_{ij} 对应的权重系数 W , W 定义为:

$$W = \left\{ \frac{|C_1|}{\sum_{j=1}^l |C_j|}, \frac{|C_2|}{\sum_{j=1}^l |C_j|}, \frac{|C_3|}{\sum_{j=1}^l |C_j|}, \dots, \frac{|C_l|}{\sum_{j=1}^l |C_j|} \right\} \quad (5)$$

采用权重系数可以保证在即使不同的示例中也能获得较好的计算效果.

定义 9 联合离群点: 对于每一个示例包 S_i , 若其在整个示例包集中的特征明显不同于其它示例包, 则把示例包 S_i 定义为一个正例, 反之, 为反例. 对于行业专家给定的范围, 若 $\text{Avg}(S_i)$ 在此范围之外, 则示例包 S_i 为一个正例, 即联合离群点.

2.2 算法设计

MIL-FindCBLOF 算法利用了 MIL 框架, 来学

习时间序列数据集. 数据在 MIL 结构下封装, 形成示例和示例包. MIL 框架充分考虑每个数据的各个属性, 并无损地保留下来. MIL-CBLOF 对数据的属性进行全局计算, 聚类出示例集, 各个示例包中的每个示例分别在各个示例集中进行计算, 最后确定多示例包的正负性, 来判断离群点.

为了减少数据中的噪声或者异常给数据计算带来的误差, 本文, 首先把时间序列数据点的序列分隔成一系列连续的且不重叠的子序列数据段, 分段聚集, 每段子序列作为一个对象, 从而把整个复杂的数据序列转化为一个个对象. 在每个对象中的每个数据可以作为每个对象的属性. 这样既简化了数据, 又保留了信息. 对象代表一个时间段的数据, 转化成对象计算又可对时序数据进行时间段的分析.

为了使 FindCBLOF 算法适应复杂数据, MIL-FindCBLOF 算法采用了退化策略, MIL 数据通过矩阵分割和属性提取, 把多示例学习的问题转化为单实例学习的问题, 然后采用基于全局的思想把所有示例包中的所有示例代入计算.

MIL-FindCBLOF 算法把每一个对象数据段看成是一个包, 包内的数据点看作是多个示例, 所有示例包中的所有示例在基于全局思想下进行计算, 充分考虑了每个实例在整个示例集的影响因子, 所以在每个示例包中计算所有示例的平均比重因子来表示示例包的影响因子, 最后判断出示例包的正负性, 计算出联合离群点, 来提高检测的准确率.

MIL-FindCBLOF 算法处理时序数据时分为 5 个步骤:

- 第一步: 预处理, 进行数据采集与清洗;
- 第二步: 切割数据段, 分段聚集成对象;
- 第三步: 初始化, 根据 MIL 框架, 对数据进行转化处理;
- 第四步: 对退化后的单示例根据定义 5、6、7, 利用相似函数进行聚类, 计算其 CBLOF 因子;
- 第五步: 根据定义 8, 计算子序列的平均因子, 判定示例包的正负性, 确定联合离群点.

2.3 算法描述

输入:

- X: 数据序列
- len: 分段数据长度
- α, β : 簇的数量参数和大小参数
- min: 正常域下限;
- 过程:

(1) 把时间序列 $X = \{x_1, x_2, x_3, \cdots, x_m\} (m \in N)$ 按不重叠等分划分为长度为 len 的子序列 $X_1, X_2, X_3, \cdots, X_g$, 多余长度不足 len 的序列单独成子序列;

- (2) 初始化 S 结构:
- for($m = 1, 2, 3, \cdots, k$), 依次初始化 $\{S_{ij1}, S_{ij2}, S_{ij3}, \cdots, S_{ijm}, \cdots, S_{dk}\}$ 的值;
- for($j = 1, 2, 3, \cdots, l$), 依次初始化 $\{S_{i1}, S_{i2}, S_{i3}, \cdots, S_{ij}, \cdots, S_{il}\}$ 的值;
- for($i = 1, 2, 3, \cdots, n$), 依次初始化 $\{S_1, S_2, S_3, \cdots, S_i, \cdots, S_n\}$ 的值.

(3) for($i = 1, 2, 3, \cdots, m$), 依次用 Squeezer^[2] 算法结合相似函数, 聚类所有示例;

(4) 用公式(1) 和(2) 计算出示例集 LC 和示例集 SC;

(5) for($i = 1, 2, 3, \cdots, m$), 用公式(3) 计算出每个示例的 CBLOF(L);

(6) for($\forall C_i \in C$), 用公式(4) 计算每个包中的 $Avg(S_i)$, 当 $Avg(S_i) < \min$ 时, 则 S_i 标记为联合离群点;

输出: 联合离群点集合

3 实验验证与结果分析

3.1 实验验证

本文实验数据均来自于某油储企业的现场实时监控数据, 数据通过现场储罐中传感器传送到 OPC 服务端, 客户端读取采集信息, 计算后存入数据库. 测试数据选用该厂 2013 年 11 月 12 日当天作业中的储罐进行测试.

测试环境: 处理器为 Intel(R) Core(TM) i3-3220 CPU 3.30 GHz, 内存为 4 GB, 操作系统为 32 位 windows7, 编程工具为 Matlab 2014a.

本文选取当天 13 个作业罐中的随机 10 罐进行测试, 储罐物料和原始数据集如表 1、表 2 所示.

表 1 储罐物料表

储罐号	物料名	密度/(g/cm³)	储罐号	物料名	密度/(g/cm³)
T ₁	双环戊二烯	0.971	T ₆	甲醇	0.7912
T ₂	甲醇	0.7912	T ₇	石脑油	0.70001
T ₃	醋酸乙烯	0.93235	T ₈	甲醇	0.7912
T ₄	石脑油	0.70001	T ₉	二甲基酰胺	0.9487
T ₅	甲醇	0.7912	T ₁₀	二甲基酰胺	0.9487

表 2 原始数据集

特征项	动态实时特征	
T_{101}	16.6	254.657
T_{102}	19.8	2365.61
.....		
.....		
T_{401}	16.7	5 241.13
T_{402}	16.8	31 861.9
T_{403}	17.1	1 827.61

离群点检测的三个重要参数分别为检测率(Diagnosis Rate, DR)、误检率(False Alarm Rate, FAR)和漏检率(Failed Report Rate, FRR),三者公示如下:

$$DR = \frac{\text{正确检测数目}}{\text{所有检测数目}} \tag{6}$$

$$FAR = \frac{\text{漏检测数目}}{\text{所有离群点数目}} \tag{7}$$

$$FRR = \frac{\text{正常点检测为异常点数目}}{\text{正常数据数目}} \tag{8}$$

3.2 结果分析

本文运用 Mil-FindCBLof 算法、FindCBLof 算法、LOF 算法和 DBSCAN 算法分别对 10 个储罐进行检测,检测率、误检率和漏检率结果对比分别如图 2、图 3、图 4 所示。

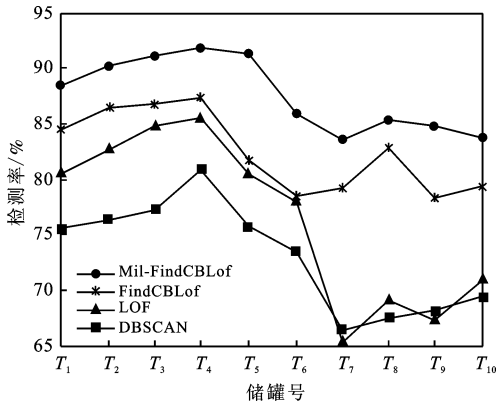


图 2 检测率结果对比图

由图 2、图 3 和图 4 所示, MIL-FindCBLOF 算法的精准性相对于其它算法都有明显的优势,检测率高,误检率和漏检率都低.检测率都保持在 82% 以上,误检率保持在 15% 以下,漏检率保持在 6% 以下.然而在 T_7 、 T_9 和 T_{10} 的检测效果上,结果偏不理想,但在对 T_2 、 T_3 、 T_4 和 T_5 的检测效果上,有明显的优势.

MIL-FindCBLOF 算法因为使用了多示例的框架,在聚类的效果上,尽可能得把对象的属性特征进行聚类.而其它算法在学习时把属性特征都堆杂在一

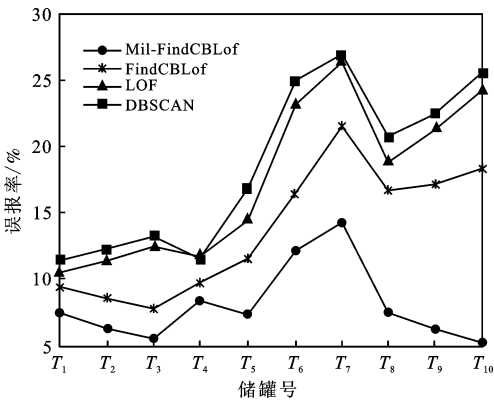


图 3 误检率结果对比图

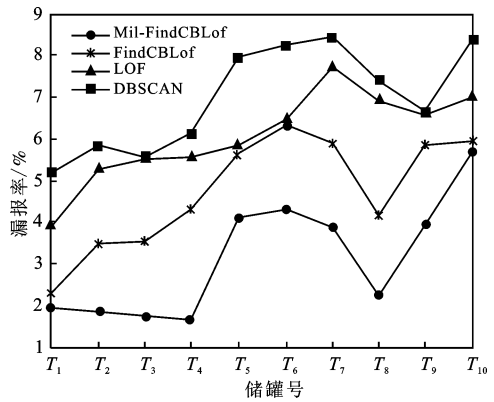


图 4 漏检率结果对比图

起,在聚类的时候往往会聚类出一些多余的无用簇,或者是计算时属性揉杂,从而使度精度和效率下降.综合仿真的结果,本文提出的 MIL-FindCBLOF 算法提升了 FindCBLOF 算法的精度和全面性.

4 结束语

本文提出了一种基于多示例学习的聚类离群点检测算法(MIL-FindCBLof)运用于时序离群点检测,对工业实时监测进行了扩充.将属性复杂的时序数据转化示例的形式进行计算,在保留属性的同时运用全局策略和权重分配保证了数据的完整性,降低了 FindCBLof 算法在检测多义数据时易聚类出无用簇的概率,提高了整体检测的准确率.今后在时序数据的分割上开展进一步研究,如:根据数据量进行动态的分割,以提高检测的稳定性.

参考文献:

[1] 彭勇,江常青,谢丰,等. 工业控制系统信息安全研究进展[J]. 清华大学学报:自然科学版,2012,97(10): 1396-1408.

的细节信息.

6 结束语

提出了一种基于图像纹理分解的图像放大新模型. 并对图像分解成的卡通与纹理两个部分分别采用不同的放大模型, 对卡通部分采用非凸二阶总广义变分模型, 可以有效的在图像的平坦区域抑制“阶梯效应”. 对纹理部分采用分数阶变分, 结合分数阶的特性, 从而能有效的增强图像的细节信息. 实验结果表明, 新的模型无论在峰值信噪比还是图像相似程度上都有明显的提高.

参考文献:

- [1] Rudin LI, Osher S, Fatemi E. Nonlinear total variation based noise removal algorithms[J]. Physica D: Nonlinear Phenomena, 1992, 60(1): 259-268.
- [2] Malgouyres F. Total variation based oversampling of noisy images [C] // Scale-Space and Morphology in Computer Vision, Lecture Notes in Computer Science, Vancouver, Canada, 2006(2106): 111-122.
- [3] Chambolle A. An algorithm for total variation minimization and applications[J]. Journal of Mathematical Imaging and Vision, 2004, 20(1/2): 89-97.
- [4] 吴玉莲, 冯象初, 姜东焕. 二阶总广义变分图像放大[J]. 光子学报, 2013, 41(6): 732-736.
- [5] Bredies K, Kunisch K, Pock T. Total generalized variation[J]. SIAM Journal on Imaging Sciences, 2010, 3(3): 492-526.
- [6] 刘亚男, 杨晓梅, 陈超楠. 基于全变分正则化的超分辨率图像重建[J]. 计算机科学, 2016, 43(5): 274-278.
- [7] Peter Ochs, Alexey Dosovitskiy, Thomas Brox, et al.

An iterated l1 algorithm for non-smooth non-convex optimization in computer vision[J]. Computer Vision & Pattern Recognition, CVPR 2013. 2013, 86(11): 1759-1766.

- [8] Chambolle A, Pock T. A first-order primal-dual algorithm for convex problems with applications to imaging [J]. Journal of Mathematical Imaging and Vision, 2011, 40(1): 120-145.
- [9] Bredies K, Valkonen T. Inverse problems with second-order total generalized variation constraints[C] // Proceedings of Samp TA 2011-9th international conference on sampling theory and applications. Singapore, 2011.
- [10] Bai J, Feng X C. Fractional—order anisotropic diffusion for image denoising[J]. IEEE Transactions on Image Processing, 2007, 16(10): 2492-2502.
- [11] Odlubny I. Fractional differential equations[M]. New York: Academic Press, 1999.
- [12] Aujol J, Aubert G, Blanc-Feraud L, et al. Image decomposition into a bounded variation component and an oscillating component [J]. Math Imaging and Vision, 2005, 22(1): 71-88.
- [13] Wang Z, Bovik A, Sheikh H. et al. Image quality visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 1-14.

作者简介:

谢斌 男, (1977-), 博士研究生, 副教授. 研究方向为图像处理与现代通信技术.

刘壮(通讯作者) 男, (1992-), 硕士研究生. 研究方向为图像处理. E-mail: 18115163704@163.com.

丁成军 男, (1992-), 硕士研究生. 研究方向为图像处理.

(上接第 49 页)

- [2] He Z Y, Xu X F, Deng S C. Discovering cluster-based local outliers[J]. Pattern Recognition Lett, 2003, 24(9/10): 1641-1650.
- [3] Wei X S, Wu J, Zhou Z H. Scalable multi-instance learning[C] // 14th IEEE International Conference on Data Mining (ICDM14). Piscataway, NJ, USA: IEEE, 2014: 1037-1042.
- [4] Wang Z G, Zhao Z S, Weng S F, et al. Incremental multiple instance outlier detection[J]. Neural Computing & Applications, 2015, 26(4): 957-968.
- [5] Zhang W J, Zhou Z H. Multi-instance learning with distribution change [C] // 28th AAAI Conference on Artificial Intelligence (AAAI14). San Francisco, CA, USA: AAAI, 2014: 2184-2190.

- [6] Wang J, Zucker J D. Solving the multiple-instance problem: a lazy learning approach[C] // 17th International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann, 2000: 1119-1125.
- [7] Ruffo G. Learning single and multiple instance decision trees for computer security applications[D]. Torino: University of Turin, 2000.
- [8] Zhou Z H, Zhang M L. Neural networks for multi-instance learning[R]. Nanjing: Nanjing University, 2002.

作者简介:

钱景辉 男, (1978-), 硕士, 讲师. 研究方向为计算机控制自动化.

卢屹伟(通讯作者) 男, (1992-), 硕士研究生. 研究方向为数据挖掘、人工智能. E-mail: 362781537@qq.com.