DOI:10.3979/j.issn.1673-825X.2016.03.023

传感器网络时间序列数据的事件分类研究

叶　李
(中国西南电子技术研究所,四川 成都 610036)

摘　要:目前智能环境中传感器网络所采集的海量数据面临着进行有效事件的模式分类及异常检测的难题。为了有效对智能环境中传感器网络采集的时间序列数据所表征的事件进行分类,提出了基于协方差特征空间映射数据的聚类分析方法。通过对采集得到的时间序列数据按时隙进行划分,映射到协方差特征空间,然后对映射后的数据进行了动态密度聚类,从而实现对事件的分类;并根据聚类结果建立分类模板,作为对日常事件进行分类划分的检测方法,同时利用所得的分类模板,实现对异常事件的检测。实验结果表明,基于协方差特征空间映射数据的聚类分析方法能有效对传感器网络采集的时间序列数据所表征的事件进行分类,并能有效提升异常事件的检测及筛选效果。

关键词:事件分类;时间序列分析;密度聚类;智能环境
中图分类号:TP3　　**文献标志码**:A　　**文章编号**:1673-825X(2016)03-0421-05

Research of time series data events classification in sensor networks

YE Li
(Southwest China Institute of Electronic Technology, Chengdu 610036, P. R. China)

Abstract: The big data collected by sensor networks in intelligent environments are faced with the problem of pattern classification and anomaly detection of time series data events. For the efficient event classification of time series data collected by sensor networks, the clustering analysis in covariance feature space was proposed. By partitioning the time series data which were collected by sensor networks, then mapping to the covariance feature space and using density-based clustering algorithms, the classification template was built to classify the usual event. The classification template was used in anomaly detection. The experiments results show that the proposed methods can effectively classify the time series data events, and enhance the performance of anomaly detection.

Keywords: event classification; times series analysis; density-based clustering; intelligent environments

0　引　言

随着自组织无线传感器网络(wireless ad-hoc sensor network, WASN)技术的发展,智能环境技术得到了高速发展[1-4]。在智能环境的技术应用中,各个传感器节点将定期采集周围环境中的事件和数据,包括温度、湿度、噪声等,并发送到中心节点进行分析处理。目前,对 WASN 所采集的海量数据进行分析处理已经成为重要的研究课题。

传感器网络所采集的数据是一系列按照时间顺

收稿日期:2016-03-18　**修订日期**:2016-05-05　**通讯作者**:叶　李　forefell@sohu.com
基金项目:国家自然科学基金(61379159);重庆市科委自然科学基金(cstc2014jcyjA1350)
Foundation Items: The National Natural Science Foundation of China (61379159); The Natural Science Foundation Project of CQ CSTC (cstc2014jcyjA1350)

序排列的信号,在实际的传感器部署环境中,发生的各种事件影响了多种类型的传感器信号的数值。不同事件促发的信号类别、时间间隔、发生频率均有不同。目前,已有的相关研究工作集中在如何从传感器网络采集的时间序列信号中发现部署环境下发生的事件是事件序列时间探测所关注的问题,如何从事件探测中筛选出异常事件仍然是一个极具挑战性的工作,困难的原因之一是在新布置的智能环境中,没有已知的异常模式进行识别的匹配。

在过去的工作中,通过对时序数据在协方差特征空间的映射,利用基于马氏距离的 T2 检验进行异常事件的检测,实验结果表明,此方法能有效检测出 WASN 采集数据中存在的异常事件,得出日常事件模板[5]。但此方法仍然存在一些问题,如筛选参数的选择对异常事件的筛选影响较大,容易筛选掉可能是正常的事件。为了解决上述问题,同时实现对事件的分类,本文提出了基于协方差特征空间映射数据的聚类分析方法,通过对时间序列数据按时隙进行划分,映射到协方差特征空间,在此基础上进行基于聚类的分析处理,进而对事件进行分类划分。

实验结果表明,本文提出的方法能够有效对传感器网络的时间序列数据进行事件分类的分析处理,相对以前的工作,能更准确地对异常事件进行筛选。

1 相关理论

1.1 协方差特征空间

数据的协方差特征空间的应用已较为广泛[6-7],但针对智能环境中传感器网络所采集的数据处理有所不同。实际上,传感器网络的采集数据是一种时间序列数据。假设有 q 个传感器,则某一时刻采集得到的观测值 $\boldsymbol{x}$ 包含了 q 个数值,即 $\boldsymbol{x}=\{f_1,f_2,\cdots,f_q\}$。在 $l(1\leqslant l\leqslant\infty)$ 时间段 T_l 内,包含了 n 个时刻的观测值的数据形成一个矩阵,该矩阵的协方差矩阵记为

$$\boldsymbol{M}^l=\begin{bmatrix}\sigma_{f_1^lf_1^l} & \sigma_{f_1^lf_2^l} & \cdots & \sigma_{f_1^lf_q^l}\\ \sigma_{f_2^lf_1^l} & \sigma_{f_2^lf_2^l} & \cdots & \sigma_{f_2^lf_q^l}\\ \vdots & \vdots & & \vdots\\ \sigma_{f_q^lf_1^l} & \sigma_{f_q^lf_2^l} & \cdots & \sigma_{f_q^lf_q^l}\end{bmatrix} \quad (1)$$

(1)式中,$\sigma_{f_i^lf_i^l}(1\leqslant i\leqslant q)$ 为第 i 个传感器观测值的方差,而数值 $\sigma_{f_i^lf_j^l}(1\leqslant i\leqslant q,1\leqslant j\leqslant q)$ 代表第 i 个和第 j 个传感器观测值之间的协方差。$\sigma_{f_u^lf_v^l}=\mathrm{cov}(f_u^l,f_v^l)=\frac{1}{n}\sum_{k=1}^{n}(f_u^{l,k}-\mu_{f_u^l})(f_v^{l,k}-\mu_{f_v^l})$ 其中,f_u^l 和 f_v^l 分别为时间段 T_l 中对应的第 u 和 v 个观测值;$f_u^{l,k}$ 和 $f_v^{l,k}$ 分别为时间段 T_l 中时刻 k 对应的第 u 和 v 个观测值。$\mu_{f_u^l}=E(f_u^l)=\frac{1}{n}\sum_{k=1}^{n}f_u^{l,k}$。

由于 $\mathrm{cov}(X,Y)=\mathrm{cov}(Y,X)$,可知协方差矩阵 $\boldsymbol{M}^l$ 为一对称阵,可只针对矩阵的上三角数值进行分析。通过对 $\boldsymbol{M}^l$ 的计算,将时间段 T_l 内的传感器观测值映射到了协方差特征空间,每对传感器之间的关系构成了协方差特征空间的坐标。

在协方差特征空间中,按时间序列采集的各个时刻的原始数据是按时间段 T_l 为单位进行的汇总,与原始数据包含的 $q\times l$ 个维度相比,协方差特征空间包含的数据维度为 $[q\times(q+1)]/2$。当时间段 T_l 包含了 $l>(q+1)/2$ 个时刻观测值的情况下,到协方差特征空间的映射方法在一定程度上压缩了数据处理量。

将传感器网络在每个时刻采集的数据进行时间段分片后映射至协方差特征空间进行分析处理有 2 个优势:①就实际情况来说,各个自然事件的跨度通常由数分钟至一个小时不等,而按秒或分钟所采集的单时刻数据并不能从整体上反映出事件的自然属性;②根据文献[6]的研究分析表明,将单一数据分组汇总后映射至协方差特征空间的异常检测技术相对单一数据检测方式而言,能获得更高的检测准确率和运行效率。

1.2 聚类分析

聚类分析就是在无先验知识的情况下,将数据集划分成簇,保证相同簇中的数据之间具有较高的相似度,不同簇的数据之间相似度较低,簇内数据的相似度越高而簇间差别度越大,表明聚类的质量越高。聚类算法大体上包括了层次法(hierarchical)、划分法(partitioning)、基于密度的方法(density-based)、基于网格的方法(grid-based)以及基于模型的方法(model-based)。

层次法聚类包括凝聚和分解 2 种方式,凝聚是从单个数据开始,不断合并 2 个或多个最合适的簇,分解法是将所有的数据点看作一个大簇,然后按照准则不断进行合适的分裂直到满足预设条件,代表算法有 Birch, Cure 和 Chameleon 等;划分法聚类只是创建数据集的一个单层的划分,其代表算法包括 K-means 和 K-medoids 算法等;基于密度的方法是基

于数据密度来度量类间相似度,通过判断区域中的点密度是否大于设定的阈值讲数据点聚类;基于网格的方法是将数据空间划分成有限数目的数据单元,通过数据网格结构进行聚类处理,其代表算法有Sting,Clique,WaveCluster 等;基于模型的聚类算法运行在根据目标数据集的分布特征假设的函数模型之上,典型的聚类模型包括统计的模型和神经网络的模型。

由于同类事件通过传感器所采集的数据具有类似属性值的特点,选用基于密度的方法对采集数据进行聚类分析处理。基于密度的聚类方法识别类的一个主要依据是每个类中的点的密度明显高于类外点的密度。

基于密度的聚类算法通过检查数据集中每个对象的 ε-领域,即邻域半径来寻找聚类。其输入参数为 ε-领域和 *MinPts*(指在任何簇中的点的最小数目)。它通过反复寻找核心对象的直接密度可达对象,合并密度可达簇,当没有新的对象可以添加到任何簇时,完成聚类。基于密度的聚类算法涉及的基本概念包括以下5方面。

1)密度。定义为在某给定的距离内,包含的对象的最小数目;

2)核心对象。根据对象的 ε-领域包含对象的数目确定,如果包含大于或等于 *MinPts* 个对象,则称其为核心对象;

3)直接密度可达。o 是从 p 直接密度可达是指给定对象集合 D 和核心对象 p,对象 o 在对象 p 的 ε-领域内;

4)密度可达。对象 o_n 是从对象 o 密度可达是指存在一个对象链 $o_1,o_2,\cdots,o_n,o_{i+1}$ 是从 o_i 关于 ε-领域直接密度可达的;

5)密度相连。对象 p 和 q 是密度相连是指对象集合 D 中存在一个对象 o,使得对象 p 和 q 是从 o 密度可达的。

经典的基于密度的聚类算法,如具有噪声的基于密度的聚类(density-based spatial clustering of applications with noise, DBSCAN)方法等,通过检查数据集中每个对象的 ε-领域来寻找类簇。算法中的 ε-领域及密度阈值 *MinPts* 为全局参数,因此,对密度不均匀的数据分类效果较差。在处理密度不均匀的数据集上,共享最近邻密度聚类(sharing nearest neighbor,SNN)[8]的聚类效果相对具有优势,相对DBSCAN 而言,SNN 改进了密度的定义方法,引入新的全局参数 k,将对象所有 k 近邻中与其共享的邻居个数总和定义为密度。通过密度定义的改变动态反映簇的密度变化,从而识别密度不均匀的簇[9-11]。

2　事件分类分析

传感器网络时间序列数据的事件分类分析处理主要包含3个过程:首先,对传感器网络原始采集数据在协方差特征空间进行映射处理;然后,针对映射后的数据进行动态密度聚类,对按照时间片划分的数据聚类处理形成事件分类簇;最后,针对日常数据,根据事件分类簇进行事件分类判别。

2.1　数据映射处理

由于传感器网络所采集的数据是某一时刻的瞬时数值,并不能准确反映环境中实际的事件变化。因此,将传感器网络采集的一系列原始数据集 D 按时间片划分成包含 n 个时刻采集数据的 T_l 的形式。

按照时间顺序将包含 n 个时刻采集数据的时间片数据 T_l 映射到协方差特征空间,取 T_l 数据生成的协方差矩阵的上三角数值,并拓展成列向量。

将所有拓展成列向量的时间片数据按时间汇聚形成数据矩阵,从而形成特征空间数据集 M。数据集 M 中每个时间片的数据即视为需要聚类处理的一个对象。

2.2　动态密度聚类

使用动态密度聚类算法,为了自动区分密度不同的簇,需对数据集 M 中的所有对象计算其密度值。选择密度值最大的对象开始建立一个新簇,循环聚集从该对象直接密度可达的对象,直到该簇无新增对象;对剩余对象迭代此过程,直到完成全部对象的处理,形成聚类结果簇列表 S。S 中的每项都为一个类簇 C,包含内容为数据集 M 中的一个或多个数据项。

动态密度聚类分析的实现步骤描述如下。

输入:数据集 M、近邻个数参数 k、ε-领域及密度阈值 *MinPts*。

输出:聚类簇列表 S。

1)计算数据集 M 中所有数据项对象的距离矩阵,根据距离矩阵及近邻个数 k、ε-领域参数计算每个对象 o 的密度值;

2)根据各个对象的密度值对所有对象排序;

3)选取未归入类簇的密度最大的对象 o_l,建立新簇 C_l;

4)遍历距离矩阵,寻找所有从对象 o_i 出发的关

于 ε-领域及密度阈值 *MinPts* 密度可达的对象 o_j，如果o_j 未加入其他簇中，则加入对象 o_j 所对应的簇 C_i；

5)重复步骤4)，直到没有新的对象加入对象 o_i 所对应的簇 C_i；

6)重复步骤3)，直到所有对象都已处理完毕；

7)输出聚类完成的簇列表 $S=\{C_1,C_2,\cdots,C_m\}$ 。

通过动态密度聚类方法，将协方差特征空间的数据进行了分类，分类的数据保存在聚类簇列表 S 中。后续数据可通过比对列表 S 中的类簇 C 中的模板数据进行分类处理。

2.3 日常数据处理

获得聚类分析结果后，剔除包含对象数目比例较少的聚类簇，将其包含的数据对象划分为噪声点或异常事件。而剩余的聚类簇列表 S'，体现了传感器网络所监测到的事件分类。聚类簇列表 S' 包含各簇 C_i 所对应的核心对象集 $\{o_1,o_2,\cdots,o_m\}$，可作为日常数据分类处理的判据。

分类判别的实现步骤描述如下。

1)系统获得传感器网络采集了对应时间片的数据后，将其映射到协方差特征空间，形成待分类数据 p；

2)将数据 p 与聚类簇列表 S' 包含各簇 C_i 所对应的核心对象集 $\{o_1,o_2,\cdots,o_m\}$ 进行距离计算，并获得对应距离最小的类簇标识；

3)如果数据 p 与各簇 C_i 所对应的核心对象集 $\{o_1,o_2,\cdots,o_m\}$ 的距离均大于 ε-领域，则将该时间片事件标识为异常事件报警；否则，将该事件分类为对应距离最小的类簇所标识的事件。

3 实验分析

实验环境中，传感器网络采集的数据主要包括6种类型，分别为声强、光强、动作、温度、湿度以及二氧化碳浓度。6种不同类型的传感器安装在实验室的一个测试房间内，系统每分钟采集并记录一次传感器数据。测试房间在日常的使用中包括了闲置、会议、清洁等模式。

实验将传感器网络初始2个月的采集数据作为训练集，按15 min的间隔划分时间片。通过对记录半年多的实验数据进行事件分类分析，聚类分析选择近邻个数参数 $k=8$、ε-领域 $=0.15$、密度阈值 $MinPts=5$。

对实验数据进行聚类分析处理后可以获取该房间7×24小时的事件网格分类模板，如图1所示。从图1可以看出，工作日早上6—7点，为空调系统运行事件；晚上22—23点，为清扫事件；在工作日中存在4种不同强度的工作或会议事件；其余时刻均为空闲事件。除去筛选后的异常数据外，实验通过数据的聚类分析处理将事件分成了7类，为日常行为模式的模型建立奠定了数据基础，这是前期工作所不能得到的。

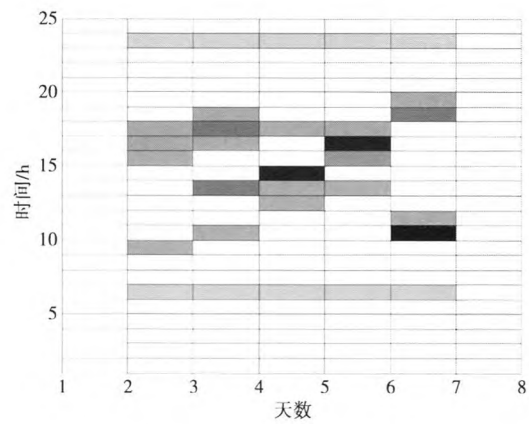

图1 事件分类网格模板

Fig. 1 Event classification grid template

未进行事件分类的实验结果如图2所示，可以看出，通过聚类分析处理后的事件分类模板提供了更加明确的关于监控房间的事件分类模式。

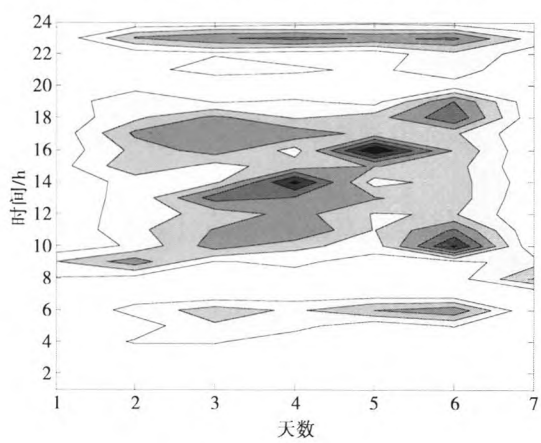

图2 事件分类模板

Fig. 2 Event classification template

基于T2异常检测后获得的分类模板如图3所示，与图2中的基于聚类分析方法进行的异常事件检测结果相比，可以看出，基于聚类的异常事件筛选形成的时间模板更加精细准确。与实际日常事件记录比对可以确定，基于聚类的检测方法提高了对异常事件的检测准确率。

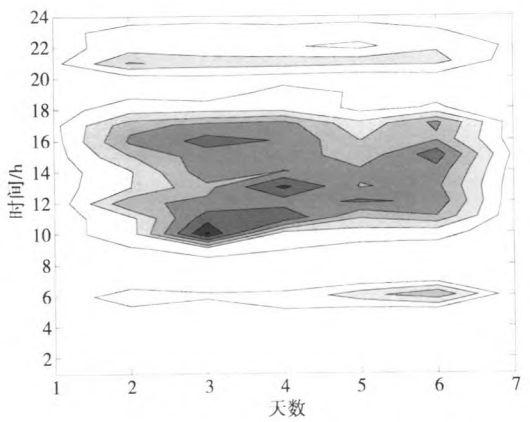

图 3　基于 T2 异常检测处理的分类模板

Fig. 3　Event classification template after original anomaly detection

4　结　论

本文在前期基于异常事件检测的传感器网络数据分析工作的基础上,提出了基于协方差特征空间映射数据的聚类分析方法。对传感器网络在每个时刻采集的数据按设定时间段进行分片,将分片数据映射至协方差特征空间,对映射后数据进行基于密度的聚类分析处理,获得了日常数据分类模板。

实验结果表明,利用聚类分析的方法获得了关于传感器网络时间序列数据的事件分类,能更准确地对异常事件进行筛选,有效完善了 7 × 24 h 的事件分类模板,对于智能环境的事件分析处理具有实际参考价值。

参考文献:

[1] GHAYVAT H, MUKHOPADHYAY S C, GUI X. Intelligent Environmental Sensing[M]. Berlin:Springer International Publishing, 2015:1-31.

[2] HAGRAS H, ALGHAZZAWI D, ALDABBAGH G. Employing Type-2 Fuzzy Logic Systems in the Efforts to Realize Ambient Intelligent Environments(Application Notes)[J]. IEEE Computational Intelligence Magazine, 2015, 10(1):44-51.

[3] 田会峰, 袁明新. 基于无线传感网的智能环境监控系统设计[J]. 测控技术, 2014, 33(9):36-39.
TIAN Huifeng, YUAN Mingxin. Design of Intelligent Environment Monitoring System Based on Wireless Sensor Network[J]. Measurement & Control Technology, 2014, 33(9):36-39.

[4] 刘练, 周凤星. 基于 APP 的智能家居环境监测系统的设计与实现[J]. 计算机测量与控制, 2014, 22(7):2018-2023.
LIU Lian, ZHOU Fengxing. Design and Implementation of Environment Monitoring System at Smart Home Based on APP[J]. Computer Measurement & Control, 2014, 22(7):2018-2023.

[5] YE Li, QIN Zhiguang, WANG Juan, et al. Anomaly Event Detection in Temporal Sensor Network Data of Intelligent Environments[C]//IEEE. 2nd International Conference on Computer Engineering and Technology (ICCET). New York:IEEE, 2010, 7:414-420.

[6] JIN S Y, YEUNG D S, WANG X Z. Network intrusion detection in covariance feature space[J]. Pattern Recognition, 2007, 40(8):2185-2197.

[7] 王邦军, 李凡长, 张莉, 等. 基于改进协方差特征的李-KNN 分类算法[J]. 模式识别与人工智能, 2014(2):173-178.
WANG Bangjun, LI Fanzhang, ZHANG Li, et al. Improved Covariance Feature Based Lie-KNN Classification Algorithm[J]. Pattern Recognition and Artificial Intelligence, 2014(2):173-178.

[8] ERTOZ L, MICHAEL S, KUMAR V. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data[C]//Proceedings of the third SIAM International Conference on Data Mining (SIAM 2003). San Francisco, CA: SIAM, 2003: 47-58

[9] 李霞, 蒋盛益, 张倩生, 等. 适用于大规模文本处理的动态密度聚类算法[J]. 北京大学学报:自然科学版, 2013, 49(1):133-139.
LI Xia, JIANG Shengyi, ZHANG Qiansheng, et al. A Dynamic Density-Based Clustering Algorithm Appropriate to Large-Scale Text Processing[J]. Acta Scientiarum Naturalium Universitatis Pekinensis:Natural Science Edition, 2013, 49(1):133-139.

[10] 万静, 张义, 何云斌, 等. 基于 KD-树和 K-means 动态聚类方法研究[J]. 计算机应用研究, 2015, 32(12):3590-3595.
WAN Jing, ZHANG Yi, HE Yunbin, et al. Dynamic clustering algorithm based on KD-tree and K-means method[J]. Application Research of Computers, 2015, 32(12):3590-3595.

[11] MENARDI G. A Review on Modal Clustering[J]. International Statistical Review, 2015(6):1-12.

作者简介:

叶　李(1977－),男,四川达州人,工程师,博士,主要研究方向为侦察对抗总体设计。E-mail: forefell@sohu.com。

(编辑:王敏琦)