·先进计算与数据处理· 文章编号:1000-3428(2017)04-0015-06 文献标志码:A 中图分类号:TP391.3

一种动态实时高校建筑能耗异常检测方法

江 航[1],卢 暾[1],顾寒苏[2],丁向华[1],顾 宁[1]

(1.复旦大学 计算机科学技术学院,上海 201203;2.希捷科技有限公司,美国 朗蒙特 80503)

摘 要: 针对静态建筑能耗异常检测方法在动态高校建筑能耗环境中容易出现误判的问题,提出一种改进的高校建筑能耗异常检测方法。采用 SA-DBSCAN 算法根据能耗数据的统计特性自适应识别建筑能耗模式,利用 C4.5 算法构建能耗模式判定树,依据判定树得到实时能耗数据的相应类别后使用 LOF 算法进行离群分析检测异常。将判定正常的能耗增量地更新到建筑能耗模式中,并根据更新结果动态调整异常检测模型。实验结果表明该方法能有效检测异常能耗数据并逐步拟合高校建筑能耗环境的变化来减少误判。

关键词: 动态实时;高校建筑能耗;异常检测;自适应识别;增量更新

中文引用格式: 江 航,卢 暾,顾寒苏,等. 一种动态实时高校建筑能耗异常检测方法[J]. 计算机工程,2017,43(4):15-20,27.

英文引用格式: Jiang Hang, Lu Tun, Gu Hansu, et al. A Dynamic and Real-time Outlier Detection Method for Energy Consumption of Campus Building[J]. Computer Engineering, 2017, 43(4):15-20, 27.

A Dynamic and Real-time Outlier Detection Method for Energy Consumption of Campus Building

JIANG Hang[1], LU Tun[1], GU Hansu[2], DING Xianghua[1], GU Ning[1]

(1. College of Computer Science, Fudan University, Shanghai 201203, China; 2. Seagate Technology Co., Ltd., Longmont 80503, USA)

【Abstract】 The static energy consumption outlier detection method prones to misjudgment of justice in the dynamic campus building energy consumption environment. Therefore, an improved outlier detection method for energy consumption of campus building is proposed. The method uses SA-DBSCAN algorithm based on the statistical characteristics of energy consumption data to identify the building energy consumption mode adpatively. Then it uses C4.5 algorithm to build energy consumption pattern decision tree. After the corresponding category of the real-time energy consumption data is obained, according to the decision tree, it uses LOF algorithm to realize outlier analysis and anomaly detection. The normalized energy consumption is updated incrementally to the building energy consumption mode, and the anomaly detection model is dynamically adjusted according to the update results. Experimental results show that the method can detect the abnormal energy consumption data effectively and fit the change of the campus building energy environment step by step which reduces misjudgments.

【Key words】 dynamic and real-time; campus building energy consumption; outlier detection; adaptive identification; incremental update

DOI: 10.3969/j.issn.1000-3428.2017.04.003

0 概述

高校作为社会的重要组成部分,在能源的使用和消耗方面都占有很大比重,因此,高校建筑的节能有重大的战略意义。在国家“节能减排”政策的号召下,很多高校都建立了能耗监测管理系统并采集了大量的能耗数据。对这些包含建筑运行期间重要信息的能耗数据进行实时的异常研究有利于发现建筑在使用和管理上的不合理之处,及时调整能达到节能的目的。然而大量的能耗数据也带来了“数据灾难”,如复旦大学节约型校园建筑节能监管平台自 2012 年 4 月正式上线以来,共安装智能电表 3 790 个,智能水表 244 个,其他智能控制设备 1 637 个,覆盖了全校 4 个校区的 400 多栋建筑,这 4 年已经稳定地采集了 14 亿

基金项目: 国家自然科学基金重点项目“智能电网信息系统的体系结构和验证环境”(61233016)。

作者简介: 江 航(1990—),男,硕士研究生,主研方向为协同计算、数据挖掘;卢 暾,副教授、博士;顾寒苏,博士;丁向华,副教授、博士;顾 宁,教授、博士、博士生导师。

收稿日期: 2016-04-18 **修回日期:** 2016-05-18 **E-mail:** 13210240016@fudan.edu.cn

4 千万多条能耗数据,每天实时新增的能耗数据多达220 万条,使得管理人员很难通过人工的方式及时地查找到异常的能耗数据。

传统的能耗异常检测方法是设定阈值,但阈值的设定较为困难,过高过低都会对检测结果有较大影响。随着研究的深入,国内外学者均提出了一些智能的建筑能耗异常检测方法[1-5],主要是通过识别正常能耗数据集的模式来达到区别未知能耗的目的。这些建筑能耗异常检测方法分析的是企业或者工厂等采集的能耗数据,这些单位的建筑类型相对单一并且作息比较固定,建筑用能规律比较稳定,所以,它们都是采用一个固定模型进行静态分析,没有根据环境变化来动态调整异常检测模型。

高校建筑的能耗环境存在建筑类型繁多、环境复杂等特点。按《高等学校校园建筑节能监管系统建设技术导则》中规定的高校建筑类型多达 13 类[6],而不同类型的建筑有不同的用能特点,比如教学楼和学生宿舍的用能模式一般不同。特定类型的建筑(如教学楼和图书馆等)还受学期阶段的影响,比如开学和期末的用能模式可能不同。建筑能耗除了受建筑类型的影响外,还受到季节、区域环境以及学校特有的寒暑假等因素的影响。众多的影响因素导致了高校建筑能耗环境比较动态。在这种较为动态的能耗环境下,若仍采用传统的静态模型做实时的异常检测,则容易出现误判。针对现有能耗异常检测方法在高校建筑能耗环境下存在的不足,本文提出一种动态实时的高校建筑能耗异常检测方法。

1 相关工作

在建筑能耗的异常检测中首先要考虑能耗数据的分类问题,因为不同能耗环境下产生的能耗数据不能直接进行比较。传统方法是基于管理者对建筑能耗使用方式的认识进行划分,比如周末和工作日,这样的划分比较粗略,不利于分析。随着研究的深入,研究者提出了图示的方法[7-8],采用时间序列图或者箱线图等描述能耗,通过观察这些图形进行能耗数据的分类,但该方法需要人工参与,不太智能。所以,研究者采用聚类的算法识别建筑的能耗模式[9-10]。

在建筑能耗异常检测研究中,文献[11]提出能耗数据的图示方法,通过观察和建筑能耗数据特性相关的图形来识别异常能耗数据。这种方法比较直观准确,但需要人工参与判定异常,所以,研究者对此进行了改进。文献[1]先通过箱线图对能耗数据进行分类,然后采用 GESD 算法来检测异常的能耗数据。文献[2]采用 DBSCAN 算法对能耗数据进行分类后,利用 GESD 算法检测异常的建筑照明能耗。文献[3]在构建能耗分类的决策树后,采用 GESD 算法进行能耗异常检测。这些建筑能耗异常检测方法的能耗分类是固定的,而现实情况中能耗的分类条件会随着环境的影响而变化。除了以上方法外,研究者还提出了一些其他的智能方法。文献[4]通过神经网络预测能耗值,再根据实际值和预测值的比例是否达到阈值进行能耗判定。文献[5]考虑温度对能耗的影响,通过实际能耗和模拟能耗的偏差来判定异常。然而这 2 种方法都没有考虑能耗的分类问题。因为高校建筑能耗环境较为动态,能耗分类条件的变化也比较频繁,所以这些研究方法在高校建筑能耗环境下进行实时异常检测时准确率不高,容易出现误判。虽然可以通过设定方法的重建周期来拟合能耗环境的变化,但周期也不固定,需要考虑实际能耗环境的变化情况。

本文提出的动态实时高校建筑能耗异常检测方法能自适应地构建高校建筑的异常检测模型,并根据建筑能耗模式类别的变化来动态调整模型,从而能有效检测异常并拟合能耗环境的变化来减少误判。

2 高校建筑能耗异常检测方法

本文的动态实时高校建筑能耗异常检测方法包括如下 4 个步骤:

1)自适应地识别建筑的能耗模式。

2)在对建筑的能耗模式标记分类属性后,构建能耗模式的判定模型。

3)实时能耗数据根据能耗模式判定模型得到相应的能耗模式后,在该模式中采用离群点检测算法进行异常判定。

4)模型的增量调整。

动态实时高校建筑能耗异常检测流程如图 1 所示。

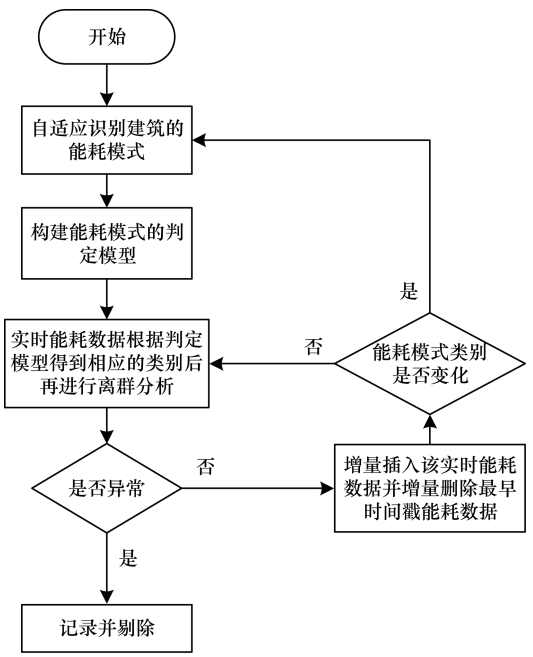

图 1 动态实时高校建筑能耗异常检测流程

2.1　建筑能耗模式的自适应识别

目前主要采用聚类算法识别建筑能耗模式。由于能耗数据簇比较复杂,形状未知,因此不能采用基于划分和基于层次的聚类算法进行处理。而基于密度的聚类算法可以识别任意形状的数据簇并且有较好的抗噪声能力,从而为异常判定提供更为“干净”的数据集,所以,本文方法中采用基于密度的聚类算法。

DBSCAN 算法[12]是经典的基于密度的聚类算法,但它对参数比较敏感并且需要人工参与调节参数,这在自适应的环境下不适用。目前研究者已经对 DBSCAN 算法参数的问题做了大量研究。文献[13]通过改进簇连接信息的方式来降低 DBSCAN 算法对参数的敏感性,但该方法不能自适应地识别参数。文献[14]不使用参数而采用分布密度的概念,但这种方法过于依赖网格的大小,容易失真。SA-DBSCAN 算法[15]通过分析数据集的统计特性来自动确定参数并且聚类准确度也较高,所以,在本文方法中采用该算法来自适应识别建筑的能耗模式。SA-DBSCAN 算法[15]是基于 DBSAN 算法的自适应改进方法,所以,聚类参数的定义和 DBSCAN 算法相同。SA-DBSCAN 算法的具体过程如下:

算法 1　SA-DBSCAN 算法

1)采用 Inverse Gaussian 分布拟合能耗数据集 k 距离的概率分布,Inverse Gaussian 分布的概率密度公式为:

$$P(x)=\sqrt{\frac{\lambda}{2\pi x^3}}\mathrm{e}^{-\frac{\lambda(x-\mu)^2}{2x\mu^2}} \tag{1}$$

2)最大似然估计得到参数 μ_k 和 λ_k:

$$\mu_k=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{2}$$

$$\lambda_k=\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}-\frac{n}{\sum_{i=1}^{n}x_i}} \tag{3}$$

其中,k 为核心对象要求的最小点个数 $MinPts$;x_i 为数据集的第 i 个对象;n 为数据集个数。

3)对 Inverse Gaussian 分布求峰值得到最小点个数 $MinPts=k$ 时,Eps_k 邻域为:

$$Eps_k=\frac{\mu_k\sqrt{9\mu_k^2+4\lambda_k^2}-3\mu_k^2}{2\lambda_k} \tag{4}$$

4)遍历数据集得到所有的核心对象并找出 Eps_k 邻域的直接密度可达点,然后查找最大密度相连集合。在遍历完成后,该最小点个数 $MinPts$ 和邻域 Eps_k 对应的聚类过程结束,那些没有簇标识的数据即为噪声。

5)根据不同的最小点个数 $MinPts$ 和邻域 Eps_k,求得噪声和最小点个数 $MinPts$ 近似相等时对应的最小点个数 $MinPts$ 和邻域 Eps_k,即为聚类的参数,相应的聚类结果为较优的建筑能耗模式。

2.2　能耗模式判定模型的构建

分类算法[16]主要有神经网络、贝叶斯分类、决策树以及支持向量机(Support Vector Machine,SVM)等。因为当前的神经网络存在着训练时间长、不可解释等缺点,所以本文方法中不采用。贝叶斯分类是基于贝叶斯定理预测数据的分类并根据可能性大小来判定数据的类别,但贝叶斯定理的成立需要很强的条件独立性假设,现实中的数据往往不满足这些条件。支持向量机(SVM)一般适合小样本的数据。而决策树分类算法导出的分类规则比较容易理解并且准确率也较高,所以,本文采用经典的 C4.5 决策树算法。C4.5 决策树算法[17]的具体过程如下:

算法 2　C4.5 决策树算法

1)计算各属性集的信息增益。

2)计算各属性集的信息增益率。

3)选取信息增益率最高的属性作为决策树的结点,对该结点进行分裂。计算分裂后各属性集的信息增益率并选取增益率最高的属性进行进一步分裂,循环此过程直到达到决策树的终止条件。

2.3　异常判定

离群点检测算法[18]主要包括基于分布、距离、偏移以及密度的类别等。基于分布的离群点检测算法要求数据满足某个概率分布模型。基于偏移的离群点检测算法的相异函数定义较复杂,对现实复杂数据检测不理想。基于距离的离群点检测算法采用的是全局检测标准,而能耗数据簇比较复杂并且位置关系也不确定,可能存在局部离群点。因此,本文选择可以识别局部离群点的基于密度的离群点检测算法,即局部离群因子(LOF)算法。LOF 算法[19]具体过程如下:

算法 3　LOF 算法

1)计算每个对象的 k 距离邻域。

2)根据对象间的距离以及对象 k 距离的相对大小求得对象间的可达距离。

3)计算每个对象的局部可达密度和局部离群因子,并根据局部离群因子是否达到阈值来判定离群点。

2.4　增量调整

当实时的能耗数据判定为正常时,把它更新到建筑的能耗模式中,这样能不断获得建筑能耗模式的最新数据特征。考虑到效率问题,本文方法采用增量的方式更新建筑的能耗模式。增量 DBSCAN 算法与 DBSCAN 算法聚类具有同样的结果[20],所以,本文采用增量 DBSCAN 算法更新建筑的能耗模式。为了保证建筑的能耗模式数据集不会随着异常检测的推进越来越大,同时较小的能耗模式数

据集也能较好地拟合建筑能耗模式的动态变化,在增量更新实时的正常能耗数据后也增量地剔除最早时间戳的能耗数据。

因为 DBSCAN 算法是基于密度的聚类算法,所以插入或删除对象只影响该对象邻域内点的核心状态。根据其对邻域对象密度影响的不同,插入更新判定为产生噪音、创建新的聚类、归入已有聚类以及合并相邻聚类中的一种[20]。删除更新判定为删除噪音、撤销聚类、减少聚类的个数以及分裂聚类中的一种[20]。

在增量更新后,如果建筑的能耗模式类别发生改变,那么有可能是建筑的能耗模式类别确实发生了改变,也可能是聚类的参数已经不适用于当前的能耗数据集。此时,采用 SA-DBSCAN 算法分析当前能耗数据的统计特征并自适应地识别较优的建筑能耗模式,然后重新构建能耗模式判定决策树用于异常检测。

3 实验结果与分析

高校建筑能耗主要包括水、电以及煤气等。由于用电能耗在总能耗中占有很大比重,因此本文实验主要分析建筑的用电能耗。该方法也可以应用于其他能耗的异常检测。在高校众多的建筑类型中,教学楼建筑除了受区域环境、季节更替以及节假日等因素的影响外,还受学期各个阶段的影响,能耗环境的动态性尤为明显,所以,实验以复旦大学节约型校园建筑节能监管平台 2015 年实时采集的第二教学楼全年的用电量为对象进行动态实时的能耗异常检测分析,以验证本文方法在较为动态的高校建筑能耗环境下异常检测的有效性。选取该教学楼 2015 年 1 月的用电能耗数据建立初始的建筑异常检测模型,逐时分析该楼 2015 年全年的用电异常情况。

在采用 SA-DBSCAN 算法自适应识别建筑的能耗模式时,把数据集的 k 距离划分成 15 份后进行拟合。因为随着邻域的增大噪声一般呈递减的趋势[15],所以实验中设定当噪声小于等于核心对象要求的最小点个数时的参数即为聚类的最优参数。因为用电量在一天的各时段的类别可能不同,工作日和周末相同时刻的类别也可能不一致,所以标注了“小时”和“是否周末”2 个分类属性进行决策树的构建。采用十折交叉验证方法来验证决策树分类的准确率,把分成 10 份的数据集中的任意 9 份用于构建决策树,1 份用于测试。将采用 LOF 算法进行离群分析设定为计算第 35 个距离邻域。正常用电量的局部离群因子应该趋近于 1,但该教学楼的用电量都比较小,较小的变化也能引起局部离群因子较大的变化,所以,设定局部离群因子大于 2 时才判定为异常。该教学楼 2015 年用电异常检测的总体情况如表 1 所示。

表 1 2015 年教学楼用电异常检测的总体情况

重建次数	最短周期/h	最长周期/h	平均周期/h	平均分类准确率/%	异常个数
66	2	617	90	86.695	281

在整个异常分析的过程中共检测出 281 个异常点。其中异常检测模型重建了 66 次,平均重建周期为 90 h,这也说明了高校教学楼能耗环境的动态性。尤其是在该教学楼能耗环境发生变化时异常检测模型的重建会比较频繁。最短的重建周期为 2 h,主要是因为用电能耗模式中数据量较少的能耗模式在增量聚类的过程中发生了类别变化。

由于在整体的用电能耗图中不便观察,因此作出了部分时段异常检测的情况。图 2 为该教学楼 2015 年 3 月第 1 周的用电量折线图,可以观察到 2015-03-03 10:57:00 的用电量 39.526 kWh 以及 2015-03-04 13:58:00 的用电量 35.843 kWh 远高于周边时刻的用电量。

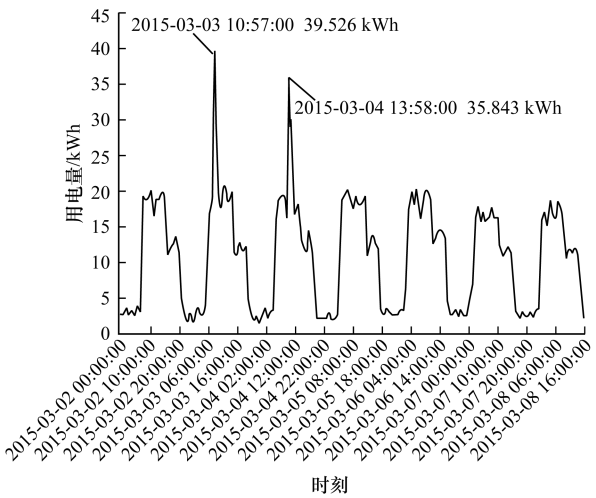

图 2 2015 年 3 月 2 日—2015 年 3 月 8 日教学楼用电量

距离该时段最近的重建时间为 2015-03-03 04:56:00,自适应识别较优的聚类参数为核心对象的最小点个数 35,邻域 1.011 6。构建的能耗模式判定树如图 3 所示,十折交叉验证分类的准确率为 82.323%。因为重建的时段处于该校寒假放假期间,所以周末和工作日的能耗差异不大,从而决策树中没有体现“是否周末”的属性。

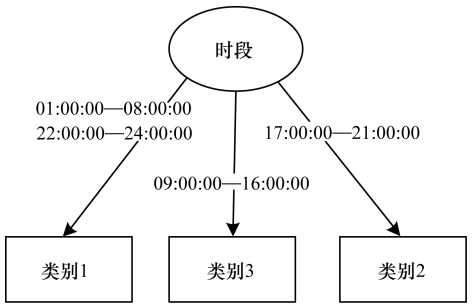

图 3 2015-03-03 教学楼的能耗模式判定树

因为逐时分析 2015-03-03 10:57:00 时教学楼的用电能耗模式类别没有发生变化,所以仍采用 2015-03-03 04:56:00 时刻识别的聚类参数进行增量的聚类以及采用图 3 所示的能耗模式判定树进行实时用电量类别的判定。2015-03-03 09:58:00 时刻增量聚类的用电能耗模式如图 4 所示。为表示方便,横坐标从 1 开始往后标识对应每个时刻。下文实验结果也采用这样的方式表示。

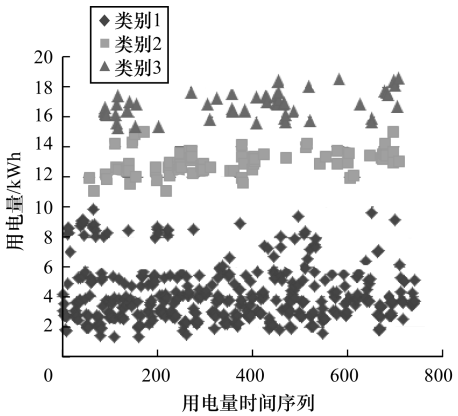

图 4　2015-03-03 09:58:00 教学楼的用电能耗模式增量聚类

由于 2015-03-03 10:57:00 的用电量为 39.526 kWh,因此根据图 3 所示的能耗模式判定树判定为图 4 的类别 3。在该用电量能耗类别中采用 LOF 算法进行离群分析得到各数据的局部离群因子如图 5 所示。用电量 39.526 kWh 的局部离群因子为28.756,远大于设定的阈值,所以,该用电量为异常值,应剔出并及时向管理人员报警。

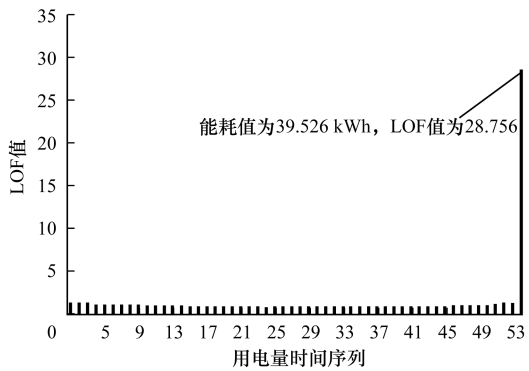

图 5　用电量为 39.526 kWh 时离群分析的 LOF 值 1

因为逐时分析 2015-03-04 13:58:00 时教学楼的用电能耗模式类别没有发生变化,所以仍采用 2015-03-03 04:56:00 时刻的异常检测模型。在相应的增量更新后的用电能耗类别中采用 LOF 算法进行离群分析得到各数据的局部离群因子如图 6 所示。该时刻的用电量 35.843 kWh 的局部离群因子为19.937,大于设定阈值,所以,该用电量为异常值。

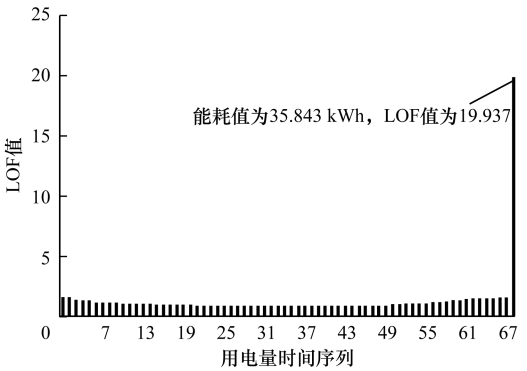

图 6　用电量为 35.843 kWh 时离群分析的 LOF 值 1

如果采用 2015-02-06 05:53:00 时刻构建的异常检测模型对该时段的用电量进行分析,即不采用动态实时重建的异常检测模型而采用 20 多天前的异常检测模型进行离群分析,得到用电量 39.526 kWh 以及 35.843 kWh的局部离群因子分别如图 7、图 8 所示。

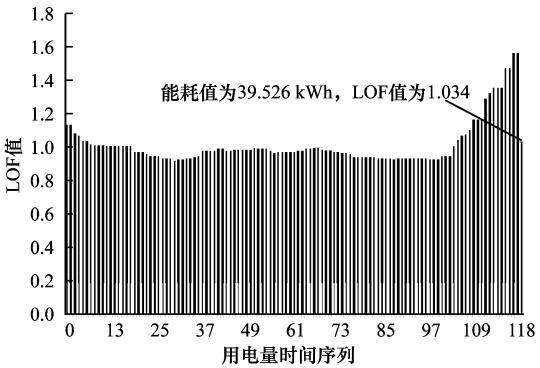

图 7　用电量为 39.526kWh 时离群分析的 LOF 值 2

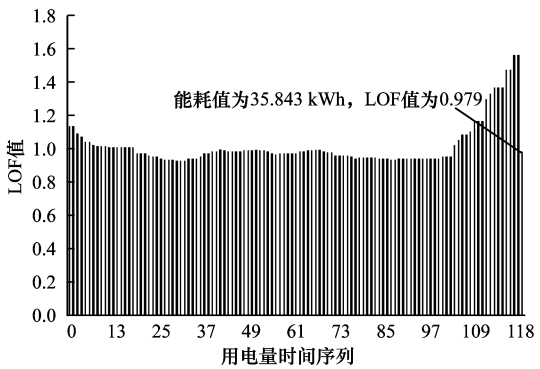

图 8　用电量为 35.843 kWh 时离群分析的 LOF 值 2

用电量 39.526 kWh 的局部离群因子为 1.034 以及用电量 35.843 kWh 的局部离群因子为0.979。因为正常用电量的局部离群因子都趋近于 1 并且也小于设定的阈值,所以这 2 个用电量都判定为正常值。从图 2 可以观察到这 2 个时刻的用电量确实属于异常,采用 2015-02-06 05:53:00 时刻构建的异常检测模型出现误判。

经过调查分析得到之所以出现误判是由于该校在 2015 年 2 月初放寒假,2015 年 3 月 8 日才正式开学,因

此 2015-02-06 05:53:00 构建的异常检测模型中包含了教学时间的用电量,从而用电量较高。寒假期间较高的异常用电量 39.526 kWh 以及 35.843 kWh 在该时段也判定为正常。若不及时地调整异常检测模型,则容易出现误判。该方法在 2015-02-06 05:53:00—2015-03-03 04:56:00 异常检测模型演变的过程中经历了 6 次模型的重建,逐步地拟合了环境变化的因素,如 2015-02-16 05:57:00 时刻重建的异常检测模型中"是否周末"分类属性消失,从而在这个过程中都能有效地减少误判并提供更为准确的异常检测结果。

在该教学楼 2015 年用电异常检测的过程中也有很多其他环境变化的拟合情况,如 2015-12-08 15:57:00 时刻重建的异常检测模型相比 2015-11-18 23:56:00 时刻重建的异常检测模型出现了高能耗模式。经过调查分析主要是由于 2015 年 11 月下旬气温开始下降,教室采用空调制暖较多,从而高用电簇出现。其中经历了 7 次模型的重建,涉及低用电能耗簇合并以及高用电能耗簇生成的过程。若不及时地调整异常检测模型,则容易出现误判。

该方法主要的时间消耗在 SA-DBSCAN 算法拟合能耗数据的统计特征上。SA-DBSCAN 算法是基于 DBSCAN 算法的自适应算法,而 DBSCAN 算法采用树结构实现的时间复杂度可达到 $O(n\text{lb}n)$(n 为数据集的个数),SA-DBSCAN 算法与 DBSCAN 算法的运算时间不存在量级的差异[15],并且高校建筑能耗数据的采集是周期性的,按照《高等学校校园建筑节能监管系统建设技术导则》中规定采集频率为 1 次/15 min ~ 1 次/h[6],所以,在下一个实时能耗到来前会有一定的间隔。本文实验在 Intel(R) Core(TM) i3 CPU@2.93 GHz,内存 4 GB,Windows7 操作系统(64 位)下运行重建异常检测模型的平均时间为 21.52 s,相比于该间隔是比较小的,该方法在高校建筑能耗环境下的效率是可行的。因为高校建筑的能耗模式类别不会一直实时变化,所以没必要有任何一个能耗改变就重建异常检测模型,由此可见,本文方法的动态重建是合理有效的。

以上实验分析说明该方法能有效地检测异常的能耗数据并能逐步地自适应拟合高校建筑节假日、气温等因素的变化动态地调整异常检测模型,从而在异常检测的过程中减少误判并提供比静态异常检测模型更为准确的异常检测结果。

4 结束语

本文提出一种动态实时的高校建筑能耗异常检测方法。该方法使用 SA-DBSCAN 算法自适应地识别较优的建筑能耗模式,然后采用 C4.5 算法构建能耗模式的判定树。实时的能耗数据根据判定树得到相应的类别后在该类别中利用 LOF 算法进行离群分析来检测异常。将判定正常的能耗增量地更新到建筑的能耗模式中并增量地剔除最早时间戳的能耗数据,根据增量更新后建筑能耗模式的类别是否变化来动态地调整异常检测模型。实验结果表明,该方法能有效地检测异常的能耗数据并逐步拟合高校动态环境的变化来减少误判,具有较强的实用性。当能耗数据簇密度差别很大时,虽然 SA-DBSCAN 算法根据噪声选取参数能有效地避免低密度数据簇大量丢失的情况,但还是可能出现高密度数据簇合并的问题,所以,下一步将重新定义聚类中密度的概念以取得更好的自适应聚类效果。

参考文献

[1] Seem J E. Using Intelligent Data Analysis to Detect Abnormal Energy Consumption in Buildings[J]. Energy and Buildings,2007,39(1):52-58.

[2] Khan I,Capozzoli A,Corgnati S P,et al. Fault Detection Analysis of Building Energy Consumption Using Data Mining Techniques[J]. Energy Procedia,2013,42(42):557-566.

[3] Capozzoli A,Lauro F,Khan I. Fault Detection Analysis Using Data Mining Techniques for a Cluster of Smart Office Buildings[J]. Expert Systems with Applications,2015,42(9):4324-4338.

[4] Dodier R H,Kreider J F. Detecting Whole Building Energy Problems[J]. ASHRAE Transactions,1999,105(1):579-589.

[5] Lin G,Claridge D E. A Temperature-based Approach to Detect Abnormal Building Energy Consumption[J]. Energy and Buildings,2015,93(1):110-118.

[6] 国家住房和城乡建设部,国家教育部. 高等学校校园建筑节能监管系统建设技术导则(试行)[EB/OL]. (2009-10-15). http://www.mohurd.gov.cn/.

[7] Li X,Bowers C P,Schnier T. Classification of Energy Consumption in Buildings with Outlier Detection[J]. IEEE Transactions on Industrial Electronics,2010,57(11):3639-3644.

[8] Hoglin D C,Mosteller F,Tukey J W. Understanding Robust and Exploratory Data Analysis[M]. New York,USA:Wiley,1983.

[9] Seem J E. Pattern Recognition Algorithm for Determin-ing Days of the Week with Similar Energy Consumption Profiles[J]. Energy and Buildings,2005,37(2):127-139.

[10] Iglesias F,Kastner W. Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns[J]. Energies,2013,6(2):579-597.

[11] Haberl J S,Abbas M. Development of Graphical Indices for Viewing Building Energy Data:Part II[J]. Journal of Solar Energy Engineering,1998,120(3):162-167.

[12] Ester M,Kriegel H P,Sander J,et al. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]//Proceedings of the 2nd International Conference on Knowledge Discovering in Databases and Data Mining. Washington D. C.,USA:IEEE Press,1996:226-231.

(下转第 27 页)

下,说明合并的数据划分和区域查询步骤对于通信开销并无明显影响,算法可以较好地适应不同节点的并行化环境。

4　结束语

本文基于数据划分的思想,提出 DBSCAN-PSM 算法,利用 KD 树进行数据划分,同时简化数据融合进一步提高算法效率,通过处理海量数据的聚类,使其更加有效地应用于云计算环境中。下一步工作重点为:1)改进现有分区方式,在不同分区中采用不同的 *Eps* 值以进一步提高聚类质量。2)利用数据集的统计特征,自动选取 *Eps* 和 *MinPts* 值,提高算法自动化水平。3)针对不同数据集类型应用不同的分区方式,进一步提高并行化水平。

参考文献

[1] 孙吉贵,刘　杰,赵连宇. 聚类算法研究[J]. 软件学报,2008,19(1):48-61.

[2] Chen Min, Gao Xuedong, Li Huifei. Parallel DBSCAN with Priority R-tree[C]//Proceedings of IEEE International Conference on Information Management and Engineering. Washington D. C., USA: IEEE Press, 2010:508-511.

[3] Wikipedia. Mapreduce[EB/OL]. (2015-04-04). http://en.wikipedia.org/wiki/MapReduce.

[4] 冀素琴,石洪波. 基于 MapReduce 的 K-means 聚类集成[J]. 计算机工程,2013,39(9):84-87.

[5] Dai Biru, Lin I C. Efficient Map/Reduce-based DBSCAN Algorithm with Optimized Data Partition[C]//Proceedings of IEEE International Conference on Cloud Computing. Washington D. C., USA: IEEE Press, 2012:59-66.

[6] He Yaobin, Tan Haoyu, Luo Wuman, et al. MR-DBSCAN: An Efficient Parallel Density-based Clustering Algorithm Using MapReduce[C]//Proceedings of IEEE International Conference on Parallel & Distributed Systems. Washington D. C., USA: IEEE Press, 2011: 473-480.

[7] Zaharia M, Chowdhury M, Das T, et al. Resilient Distributed Datasets: A Fault-tolerant Abstraction for In-memory Cluster Computing[C]//Proceedings of Usenix Conference on Networked Systems Design & Implementation. San Jose, USA: USENIX Association, 2012:2.

[8] 李璐明,蒋新华,廖律超. 基于弹性分布式数据集的海量空间数据密度聚类[J]. 湖南大学学报(自然科学版),2015,42(8):116-124.

[9] Cordova I, Moh T S. DBSCAN on Resilient Distributed Datasets[C]//Proceedings of International Conference on High Performance Computing & Simulation. Washington D. C., USA: IEEE Press, 2015:531-540.

[10] 于亚非,周爱武. 一种改进的 DBSCAN 密度算法[J]. 计算机技术与发展,2011,21(2):30-33.

[11] Berger M J, Bokhari S H. A Partitioning Strategy for Nonuniform Problems on Multiprocessors[J]. Computers, 1987, 100(5):570-580.

[12] Wikipedia. KD-tree[EB/OL]. (2015-04-10). https://en.wikipedia.org/wiki/K-d_tree.

[13] 周水庚,周傲英,曹　晶. 基于数据分区的 DBSCAN 算法[J]. 计算机研究与发展,2000,37(10):1153-1159.

[14] Scikit-learn. Dataset Loading Utilities[EB/OL]. (2015-11-09). http://scikit-learn.org/stable/datasets/.

[15] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python[J]. Journal of Machine Learning Research, 2011, 12(1):2825-2830.

编辑　陆燕菲

(上接第 20 页)

[13] 蔡颖琨,谢昆青,马修军. 屏蔽了输入参数敏感性的 DBSCAN 改进算法[J]. 北京大学学报(自然科学版),2004,40(3):480-486.

[14] Xu X, Ester M, Kriegel H P, et al. A Distribution-based Clustering Algorithm for Mining in Large Spatial Databases[C]//Proceedings of the 14th International Conference on Data Engineering. Washington D. C., USA: IEEE Press, 1998:324-331.

[15] 夏鲁宁,荆继武. SA-DBSCAN:一种自适应基于密度聚类算法[J]. 中国科学院大学学报,2009,26(4):530-538.

[16] 罗　可,林睦纲,郗东妹. 数据挖掘中分类算法综述[J]. 计算机工程,2005,31(1):3-5.

[17] Quinlan T R. C4.5: Programs for Machine Learning[M]. San Mateo, USA: Morgan Kaufamann, 1993.

[18] Han J, Kamber M. Data Mining: Concepts and Techniques[M]. San Mateo, USA: Morgan Kaufmann, 2000.

[19] Breunig M, Kriegel H P, Ng R, et al. LOF: Identifying Density-based Local Outliers[C]//Proceedings of ACM SIGMOD International Conference on Management of Data. New York, USA: ACM Press, 2000:93-104.

[20] Ester M, Kriegel H P, Sander J, et al. Incremental Clustering for Mining in a Data Warehousing Environment[C]//Proceedings of the 24th International Conference on Very Large Data Bases. San Mateo, USA: Morgan Kaufmann, 1998:323-333.

编辑　陆燕菲