文章编号:1000-582X(2012)07-133-05

能耗实时监测的数据挖掘方法

卿晓霞[a],肖 丹[b],王 波[b]
(重庆大学 a.城市建设与环境工程学院;b.计算机学院,重庆 400044)

摘 要:针对能耗监测常用的设定能耗阈值方法和基于历史数据的数据分析方法在实时性和智能性方面的不足,提出了一种基于数据挖掘技术的能耗实时监测方法。该方法通过对历史能耗数据进行聚类分析识别耗能体特有的能耗模式集合,对数据分类后获得能耗模式判定树,在能耗实时监测过程中对动态采集的能耗数据进行模式匹配,与相同模式历史数据进行离群点分析,可判别当前能耗是否异常。结合某综合大楼能耗数据进行了实验,验证了该方法及时发现能耗数据异常的有效性。

关键词:能耗;能源管理;数据挖掘;实时;监测
中图分类号:TK01+2; TP391 **文献标志码**:A

A real-time monitoring method of energy consumption based on data mining

QING Xiao-xia[a], *XIAO Dan*[b], *WANG Bo*[b]
(a. Faculty of Urban Construction and Environmental Engineering; b. College of Computer Science, Chongqing University, Chongqing 400044, China)

Abstract: A real-time monitoring method of energy consumption based on data mining techniques is proposed to compensate the deficiency of common energy consumption methods in real time and intelligence. The new method can identify energy consumption patterns by clustering analysis of historical energy consumption data, get the decision tree of energy consumption pattern by classifying the energy consumption data, match the real-time energy consumption data with the energy consumption patterns, make outlier analysis with historical data of the same pattern, and then determine whether the current energy consumption is abnormal. The experiment with energy consumption data from the comprehensive building proves that the new method is effective in detecting the abnormal data of energy consumption real-timely.

Key words: energy consumption; energy management; data mining; real-time; monitoring

中国正处于能源消耗高同时能源相对短缺的时期,大规模的城市建设使得建筑能耗在社会总能耗中的比重达到了 46%[1],环保的更高要求需要在城市及城镇建设运行大量的、能耗高的污水处理厂;另外,许多工业企业也是耗能大户。通过自动监测和控制手段降低运行时能耗是实现节能的一个重要途径。目前,许多高耗能企业已实现能耗数据的采集,在运行过程中积累了大量的能耗数据,这为从数据上分析能耗情况提供了基础。但是,大量的数据也带来了"数据灾难",使得管理人员难以快速有效地发现存在的能耗异常问题。传统的监测方法[2-3]是由能源管理人员设定一个阈值,由此决定能耗是否

收稿日期:2012-02-08
基金项目:国家科技重大专项资助项目(2009ZX07315-005);重庆大学研究生创新基金个人项目资助项目(CDJXS11180016)
作者简介:卿晓霞(1963-),女,博士,重庆大学副教授,主要研究方向为自动控制技术及应用,(E-mail) qxx118@126.com。

出现异常，但这存在两个问题，一是阈值难以确定，过高或过低都可能会影响到实际检测结果；二是没有考虑季节、区域环境特征等因素，不能动态适应变化。因此，一些智能检测方法被提出：在建筑能耗分析方面，Seem 使用基于统计的离群点检测方法 GESR 检测历史能耗数据中存在的异常[4]；Li 等使用离群点检测方法 GESR 剔除能耗异常的数据点，然后利用典型变量分析（CVA）对能耗数据进行了分类和预测[5]；在污水处理厂能耗分析方面，杨凌波等采用统计分析、聚类分析等方法对能耗状况及其影响因素进行了定量规律识别[6]。但是，这些方法都只是基于历史数据对能耗进行静态分析，不能实时准确地检测能耗异常。

笔者提出了一种基于数据挖掘技术的能耗实时监测方法，该方法通过识别耗能体（如建筑、污水处理厂等，以下简称耗能体）存在的能耗模式，建立能耗模式判定树，对实时采集的能耗数据进行模式匹配，与相同能耗模式的历史数据进行比较分析，进而判定能耗是否异常。该方法具有实时性、通用性和鲁棒性等特点。

1 新方法概述

图 1 是新方法的总体流程图。新方法为了提高能耗异常检测的准确性，需要对蕴含的能耗模式进行挖掘，由于能耗模式可能随建筑物设备配置、污水处理量、季节等因素的变化而变化，因此需要定期对耗能体的能耗模式进行识别并重建能耗模式判别树。

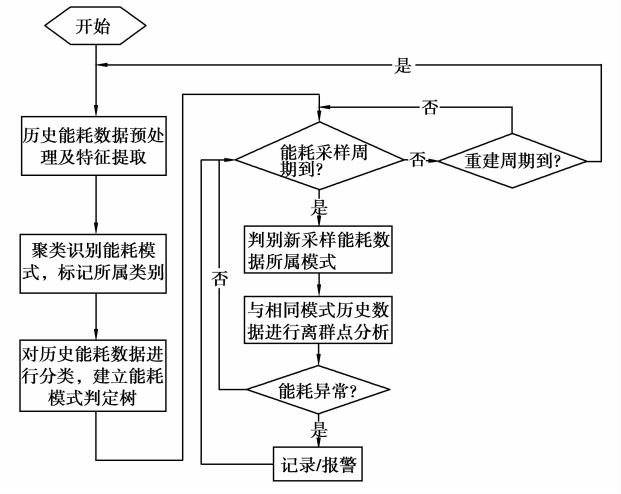

图 1 一种基于数据挖掘的能耗实时监测方法流程图

具体过程为：首先对历史能耗数据进行预处理和特征提取，决定主要能耗特征，如每小时平均能耗和每小时最高能耗量；然后对历史能耗数据中所蕴含的能耗模式进行识别[7]，并在数据中进行标记；接着根据标记结果对历史能耗数据进行分类训练，建立能耗模式判定树。当采集到新数据时，对当前能耗数据判定所属能耗模式，与相同能耗模式的历史数据进行离群点分析，判断当前能耗数据是否为离群点，若是则表明能耗异常，应报警并记录。当未采集到能耗数据时，则判断是否到达重建周期，若到达则执行前述识别能耗模式和重建能耗模式判定树的过程。

2 算法优选

2.1 聚类算法

聚类算法[8-9]可分为划分聚类、层次聚类、密度聚类、约束聚类等类别。划分聚类（K-Means 等）和层次聚类（CURE、ROCK、CHAMELEON 等）不能发现形状比较复杂的簇；密度聚类（DBSCAN、OPTICS 等）可以发现具有任意形状的聚类，并能有效处理异常数据；约束聚类通常只用于处理某些特定应用领域中的特定需求。由于能耗数据集一般比较大，可能存在噪声数据，且对其形状无任何先验知识，所以选择基于密度的聚类算法 DBSCAN[10]。

DBSCAN 算法步骤[11]：

1）确定参数 ε 和 MinPts；

2）从数据集中任意选取一数据点 p，判断其邻域内点的数目 n，如果 n 大于给定的参数 MinPts，则 p 为核心点，否则 p 为边界点；

3）如果 p 为核心点，则建立以 p 为核心点的簇，并依次检查 p 的邻域内的各点 q，赋予簇标识，并将可合并的簇赋予同样的簇标识；

4）若数据集中没有未遍历的点时，聚类过程结束。

2.2 分类算法

分类算法[12]主要有决策树、贝叶斯分类、支持向量机等，贝叶斯分类（Bayesian）需要假定各属性之间相互独立，而实际中往往不成立；支持向量机（SVM）在处理大规模数据集时速度较慢；决策树算法具有较高的分类准确率和较好的分类效率，且生成规则容易理解，选其用于建立能耗模式判定树。C4.5[13]是一种被广泛应用的经典决策树分类算法。

C4.5 算法步骤[14]：

1）计算每个属性的信息增益率

$$\mathrm{GainRatio}(A)=\frac{\mathrm{Gain}(A)}{\mathrm{Splitl}(A)},\qquad(1)$$

式中：Gain 是信息增益，$\mathrm{Splitl}(A)=-\sum_{i=1}^{n}p_i\log_2(\mathrm{p_i})$；

2）选取具有最高增益率的属性作为给定集合的分裂属性。对被选取的分裂属性创建一个结点，并以该属性标记，对该属性的每个值创建一个分支，并据此划分样本。

2.3　离群点检测算法

离群点检测算法[15]可分为基于分布的、基于距离的、基于密度的等类别，基于分布的算法（ESD、GESR 等）需要假设数据集符合某种概率分布模型；基于距离的算法（k-NN 等）的检测标准是全局的、绝对的；基于密度的算法（LOF 等）可以同时检测全局离群点和局部离群点。考虑到能耗数据集中可能存在多个密度不一致的簇，且簇之间的位置关系复杂，为提高异常检测准确度，选用基于密度的离群点检测算法。LOF 算法[16]作为一种经典的基于密度算法，摈弃了以前所有的异常定义中非此即彼的绝对异常观念，更加适合现实生活中的应用[15]。

LOF 算法步骤[17]：

1）对数据集 D 中每个对象进行邻域查找，计算其 MinPts 邻域，并存储每个对象与其邻域中对象的距离；

2）计算每个对象的局部可达密度和局部离群点因子 LOF；

3）根据设定的 LOF 阈值，将 LOF 值大于该阈值的对象判定为离群点。

3　方法验证

为验证新方法的有效性，以分析某综合楼的能耗数据为例进行说明，但该方法不仅限于建筑能耗监测，也可应用于污水处理厂等高耗能行业。

某综合大楼是一座集科研、办公、教学为一体的综合楼，实行能耗分项计量，每隔 10 min 对能耗数据采样一次。不失一般性，选取夏季（7—9 月）能耗数据（如图 2 所示），以耗电量作为主要研究对象。水、燃气等用量数据也可使用该方法进行分析。

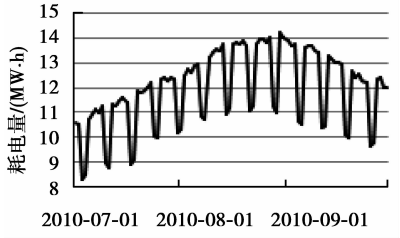

图 2　某综合楼 7—9 月电耗数据

3.1　数据预处理及特征提取

为了使用聚类分析发现建筑潜在的能耗模式，需要从能耗时间序列数据中提取出反映能耗情况的重要的特征向量：每小时平均能耗 C_{avg} 和每小时最高能耗 C_{max}，即特征向量 $\boldsymbol{C}=(C_{avg}, C_{max})$。

由于能耗数据数值较大，且不同时间段能耗值可能变化较大，因此在聚类之前使用 z-score 方法[18]进行规范化处理。在 z-score 规范化中，属性 A 的值基于 A 的均值和标准差规范化。A 的值 v 规范化为 v'，由下式计算：

$$v' = \frac{v-\overline{A}}{\sigma_A}, \tag{2}$$

式中，$\overline{A}$ 和 σ_A 分别为属性 A 的均值和标准差。

对于能耗时序数据中可能存在的缺失值，为简便起见，采取丢弃处理。

3.2　识别能耗模式

不同的公共建筑因为建筑类型、使用方式等原因在能源消耗上表现出一定的差异性。为了准确有效地对当前能耗情况进行分析，首先需要识别该建筑物的能耗模式。

对某综合大楼 2010 年夏季的历史数据应用 DBSCAN 算法进行聚类的结果如图 3。从图中可得知某综合楼夏季能耗情况存在 3 种模式，通过调查分析，这 3 种能耗模式分别对应“工作日上班时段”、“节假日上班时段、工作日下班时段”和“节假日下班时段”时的用能情况。由于某综合楼晚上 10 点关门，且节假日也在使用，因此每晚和节假日也有较高的能耗值。

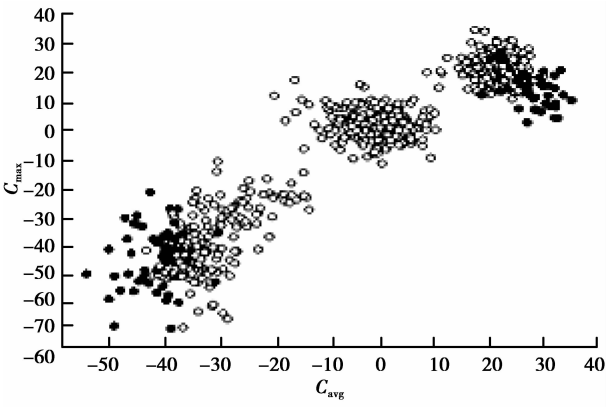

图 3　历史能耗数据聚类图

在识别建筑存在的能耗模式之后，为便于分析能耗模式的判别条件，需要构造 2 个新的属性：能耗标签和是否周末。将 3.2 节聚类得到的数据簇结果填入能耗标签属性中，根据日期判断是否周末将结果填入周末属性中。构造新属性之后的部分能耗数据如表 1 所示。

对构造了新属性之后的能耗数据应用 C4.5 算法进行分类，得到能耗模式判定树如图 4 所示。使用 3/4 的数据作为训练集，剩下的 1/4 数据作为检验集，分类误差为 2.71%。

表1 构造了新属性之后的部分能耗数据

日期	周末	时刻	能耗/(kW·h)	能耗标签
2010-07-16	否	20:10:00	149.68	Cluster3
2010-07-16	否	20:20:00	156.34	Cluster3
2010-07-16	否	20:30:00	152.41	Cluster3
		……		
2010-07-16	否	22:40:00	28.03	Cluster1
2010-07-16	否	22:50:00	31.22	Cluster1
2010-07-16	否	23:00:00	27.10	Cluster1
		……		
2010-07-16	是	9:50:00	82.58	Cluster2
2010-07-16	是	10:00:00	88.65	Cluster2
2010-07-16	是	10:10:00	83.79	Cluster2
2010-07-16	是	10:20:00	85.52	Cluster2

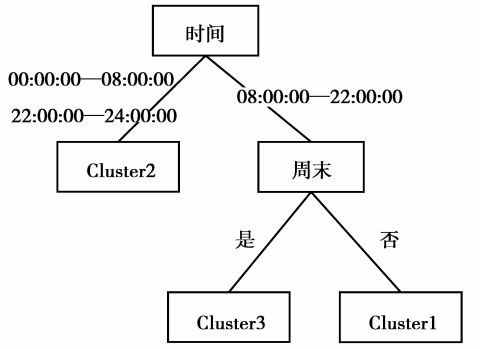

图4 能耗模式判别决策树

3.3 分析实时能耗数据

为说明该方法的实际监测效果，选取2010年8月25日的能耗数据为例，如图5所示。

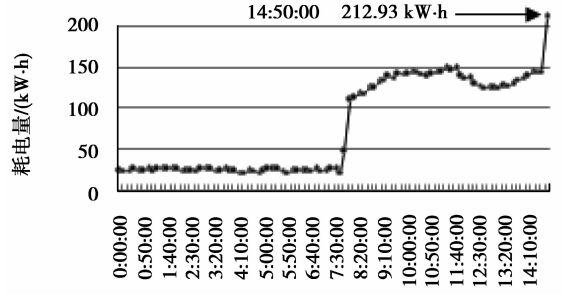

图5 2010年8月25日能耗数据时序图

14:50:00采样的能耗数据值为212.93 kW·h，根据之前生成的能耗模式判定树，该点的能耗模式为Cluster1，为判断此时的能耗情况，与相同能耗模式的历史数据一起使用LOF算法进行离群点分析。

设定MinPts=40，计算各数据点的局部离群点因子LOF结果如图6所示，其中离群点LOF的最大值为5.621，最小值为1.929；聚类簇中点LOF的最大值为1.274，最小值为0.906。14:50:00对应的数据点的LOF值为4.373，因此应为离群点，表明能耗出现异常，应立即向建筑管理人员报警。经管理人员检查发现，造成能耗异常的原因是因为1号冷水机组出现故障，2号冷水机组自动启动承担负载，但1号冷水机组重启后2号冷水机组并未关闭，出现2台冷水机组同时工作的情况，导致能耗异常。在关闭2号冷水机组后，能耗又恢复正常。

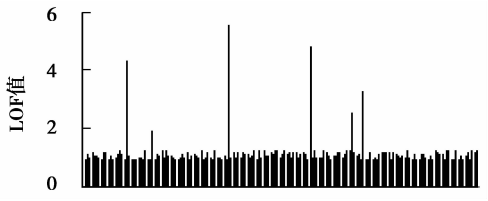

图6 各能耗数据点的LOF数值

图7显示的是2010年9月5日的能耗时序图。11:40:00采样的能耗数据值为186.308 3 kW·h，根据能耗模式判定树判断该点的能耗模式为Cluster3，为判断此时能耗情况，与相同能耗模式的历史数据一起使用LOF算法进行离群点分析。

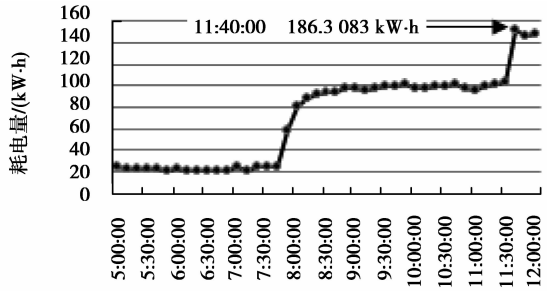

图7 2010年9月5日能耗数据时序图

设定MinPts=40，计算各数据点的局部离群点因子LOF结果如图8所示，其中离群点LOF的最大值为3.574，最小值为1.952；聚类簇中点LOF的最大值为1.302，最小值为0.933。14:50:00对应的数据点的LOF值为2.652，因此应为离群点，表

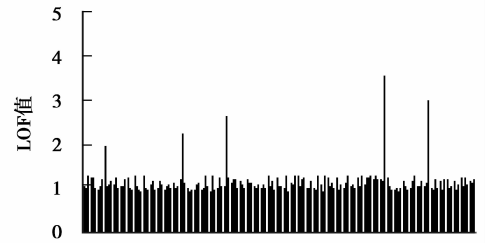

图8 各能耗数据点的LOF数值

明能耗出现异常，应立即向建筑管理人员报警。经检查得知，当天综合楼需要加装一些电子设备以增强外观显示效果，调试设备增加了综合楼的整体能耗，从而导致能耗出现异常。

4 结　语

通过对某综合大楼能耗数据进行的分析表明，新方法能够在耗能设备系统实际运行过程中对每个采样点的能耗数据进行分析，及时发现和报告能耗异常。该方法还具有通用性，能够广泛应用于公共建筑、污水处理厂、工业生产企业等高耗能场所。此外，由于采用的聚类算法具有抗噪声干扰能力，该方法也具有较好的鲁棒性。在实际运用中，该方法可嵌入能源管理系统中，帮助管理人员及时有效地了解能耗情况，为采取有效的节能措施提供决策支持。

参考文献：

[1] 李百战，张宇，丁勇. 重庆市公共建筑能源管理现状分析[J]. 暖通空调，2010，40(9)：112-117.
LI Baizhan，ZHANG Yu，DING Yong. Status analysis of public building energy management in Chongqing[J]. Heating Ventilating & Air Conditioning，2010，40(9)：112-117.

[2] 李峥嵘，李浩翥，郁盛，等. 建筑能效当量能耗方法研究[J]. 同济大学学报：自然科学版，2010，38(3)：353-357.
LI Zhengrong，LI Haozhu，YU Sheng，et al. Equivalent energy consumption of building energy efficiency[J]. Journal of Tongji University：Natural Science，2010，38(3)：353-357.

[3] Lee W S，Lee K P. Benchmarking the performance of building energy management using data envelopment analysis[J]. Applied Thermal Engineering，2009，29(16)：3269-3273.

[4] Seem J E. Using intelligent data analysis to detect abnormal energy consumption in buildings[J]. Energy and Buildings，2007，39(1)：52-58

[5] Li X L，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

[6] 杨凌波，曾思育，鞠宇平，等. 我国城市污水处理厂能耗规律的统计分析与定量识别[J]. 给水排水，2008，34(10)：42-45
YANG Lingbo，ZENG Siyu，JU Yuping，et al. Statistical analysis and quantitative recognition of energy consumption of municipal wastewater treatment plants in China[J]. Water & Wastewater Engineering，2008，34(10)：42-45

[7] Seem J E. Pattern recognition algorithm for determining days of the week with similar energy consumption profiles[J]. Energy and Buildings，2005，37(2)：127-139.

[8] Han J W，Kamber M. 数据挖掘概念与技术[M]. 范明，孟小峰，译. 2 版. 北京：机械工业出版社，2007.

[9] 贺玲，吴玲达，蔡益朝. 数据挖掘中的聚类算法综述[J]. 计算机应用研究，2007，24(1)：10-13.
HE Ling，WU Lingda，CAI Yichao. Survey of clustering algorithms in data mining[J]. Application Research of Computers，2007，24(1)：10-13.

[10] Wu Y，Yang K，Zhang J Z. Using DBSCAN clustering algorithm in spam identifying[C]// ICETC 2010，2nd International Conference on Education Technology and Computer，June 22 - 24，2010，Shanghai，China. United States：IEEE Computer Society，2010：1398-1402.

[11] Nasibov，Efendi N. Robustness of density-based clustering methods with various neighborhood relations[J]. Fuzzy Sets and Systems，2009，160(24)：3601-3615.

[12] Kotsiantis S B. Decision trees：a recent overview[J]. Artificial Intelligence Review，2011：1-23

[13] Kemal P，Salih G. A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems[J]. Expert Systems with Applications，2009，32(6)：1587-1592.

[14] Taherkhani A. Recognizing sorting algorithms with the C4.5 decision tree classifier[C]// ICPC 2010，18th IEEE International Conference on Program Comprehension，June 30 - July 2，2010，Braga，Minho，Portugal. United States：IEEE Computer Society，2010：72-75.

[15] 薛安荣，姚林，鞠时光，等. 离群点挖掘方法综述[J]. 计算机科学，2008，35(11)：13-18.
XUE Anrong，YAO Lin，JU Sshiguang，et al. Survey of outlier mining [J]. Computer Science，2008，35(11)：13-18.

[16] Breunig M，Kriegel H P，Ng R，et al. LOF：identifying density-based local outliers [C] // Proceedings of the ACM SIGMOD International Conference on Management of Data，May 16 - May 18，2000. New York：ACM Press，2000：93-104.

[17] 李健，阎保平，李俊. 基于记忆效应的局部异常检测算法[J]. 计算机工程，2008，34(12)：4-6.
LI Jian，YAN Baoping，LI Jun. Memory-effect-based local outlier detection algorithm [J]. Computer Engineering，2008，34(12)：4-6.

[18] 蔡维玲，陈东霞. 数据规范化方法对 K 近邻分类器的影响[J]. 计算机工程，2010，36(22)：175-177.
CAI Weiling，CHEN Dongxia. Influence of data normalization methods on K-nearest neighbor classifier[J]. Computer Engineering，2010，36(22)：175-177.

（编辑　郑　洁）