第 42 卷 第 5 期 仪 器 仪 表 学 报 Vol. 42 No. 5
2021 年 5 月 Chinese Journal of Scientific Instrument May 2021

DOI: 10.19650/j.cnki.cjsi.J2107586

基于高维空间聚类的集中供热末端数据异常检测*

孙文慧[1],张海伦[2],王 雷[3]

(1. 山东建筑大学信息与电气工程学院 济南 250101; 2. 山东大学信息科学与工程学院 青岛 266237;
3. 山东大学控制科学与工程学院 济南 250061)

摘 要:因集中供热建筑结构、住户行为习惯等差异,末端住户供暖数据具有特征差异大、非线性强、数据量大、响应时间长等特征,在原数据空间中利用聚类分析进行异常检测造成类间数据交叉,精确性无法保证。本文提出高维高斯混合聚类算法,将数据集映射到高维空间进行聚类,利用核函数映射、内积运算与高维特征空间分解等计算方法,提高精确度,规避维数灾难。搭建工业大数据分析平台,对比 K-Means、高斯混合、恒虚警率、高维高斯混合算法聚类结果与异常检测精确度,本文所提算法将准确性提高到 90.72%,误报率降低到 5.92%,结合该算法完成 4 类异常用热数据集的解释与辨识。高维高斯混合聚类可以有效分析用户用热特征、检测异常数据,辅助降低采暖能耗,实现建筑节能。

关键词: 集中供热;异常检测;高维高斯混合聚类;空间映射

中图分类号: TP391.5 TH81 **文献标识码**: A **国家标准学科分类代码**: 520

Anomaly detection of residential data in the district heating system based on high dimensional Gaussian mixture clustering

Sun Wenhui[1], Zhang Hailun[2], Wang Lei[3]

(*1. School of Information and Electrical Engineering, Shandong Jianzhu University, Ji'nan 250101, China*;
2. School of Information Science and Engineering, Shandong University, Qingdao 266237, China;
3. School of Control Science and Engineering, Shandong University, Ji'nan 250061, China)

Abstract: The building structure and household behavior of end-users are different. The heating datasets of end-users have features of large amount, strong nonlinearity, long response time, etc. In the original data space, it is hard to implement anomaly detection by the clustering analysis. The problem is the serious data crossing that greatly reduces the accuracy. In this paper, the high dimensional Gaussian mixture clustering (HGMM) is proposed to map datasets in original space to high-dimensional space for clustering. Kernel function mapping, inner product, decomposition of high-dimensional feature space are used to improve clustering accuracy and avoid dimensional disaster. Industrial big data ingestion and analysis platform (IBDP) is established. The clustering and anomaly detection accuracy of K-Means, Gaussian mixture model (GMM), constant false alarm rate, and HGMM are compared. The proposed method could improve the clustering accuracy to 90.72% and reduce the detection error rate to 5.92%. Four types of abnormal heating patterns are identified and analyzed. The proposed HGMM could be used to effectively analyze the residential heating characteristics, detect the abnormal datasets, help reduce the heating energy consumption, and realize the building energy saving.

Keywords: district heating; anomaly detection; high dimensional Gaussian mixture clustering; space mapping

0 引 言

随着人民生活水平的不断提高,能源的需求和消耗量也在日益激增。根据国际能源署统计,全球范围内,建筑能耗占社会能源消费总量的 32%,其中供暖、通风和空调消耗量占 34.8%[1]。根据中国供热行业发展回顾与发展趋势分析报告(2020~2026 年),近年来

收稿日期:2021-03-07 Received Date: 2021-03-07
*基金项目:国家自然科学基金(61903226)、山东省自然科学基金(ZR2020QF061,ZR2020QF068)项目资助

我国集中供热面积达到 61.1 亿平方米,城市供热行业仍以燃煤为主要燃料(占比超过 80%),加剧了能源的日益匮乏,也带来了严重的环境问题。其中,居民采暖热力消费增速高于工业领域,占全国热力消费总量约 30%且比重不断提高。

居民采暖集中供热系统回路中一次管网由热电厂向换热站提供高温高压蒸汽或者高温热水;二次管网由换热站向用户提供 65 ℃左右热水;一次管网与二次管网在换热站处进行热量交换。相较于一次管网,二次管网情况复杂,不同小区因其地理位置、建筑结构、建筑材料、住户行为习惯等差异明显,导致供暖数据量大、特征差异较大。在现在集中供热分户计量逐渐普及的背景下,大量的分户计量数据无法进行有效地分析,用户侧异常用热、供热失调等问题无法实时发现并解决。本文提出一种居民采暖异常用热检测算法,并利用该算法对多小区集中供热数据集进行聚类与异常检测分析,识别开窗、加装换热器等异常用热行为,为管理者提供合理化建议,辅助降低采暖能耗,提高能源使用效率,改善室内热环境质量,促进居民行为节能,实现建筑节能目标。

1 相关工作

异常检测是近二十年来的一个研究热点,早期基于物理和模型的诊断方法一直是主流方法;近年来,随着智能传感器、数据分析和深度学习技术的快速发展,数据驱动算法得到广泛认可(例如,支持向量机、卷积神经网络、深度学习[2-4]等)。Ravanbakhsh 等[5]利用深度网络支持向量机进行特征提取,训练异常检测模型。然而,这些深层次的特性是次优的,因为它们没有针对整个问题进行设计或优化。王小飞等[6]针对大样本飞参数据,提出基于支持向量合并的在线支持向量数据描述算法,提升发动机监控性能效率。Chen 等[7]结合自编码和 Adaboost 提出一种新的图像离群点检测方法,将 Adagrad 和最近梯度下降相结合优化学习目标,成功应用于高维图像数据集。张栋等[8]提出融合目标检测的级联卷积神经网络算法,检测和提取包括质控峰和检测峰的区域,并对峰值进行快速定位。然而,上述文章往往假设多数据集间彼此独立,实际应用中的数据集往往包含一些相关性,因此上述方法难以对数据进行完善的建模。

聚类算法因其耗时较少、人工监督成本较低,被认为是最流行的描述性数据挖掘算法,成功的应用于建筑能源系统大数据分析中[9]。聚类算法主要被用于识别负荷模式[10]以及负荷预测等数据挖掘技术的数据预处理部分[11-12]。K-means 算法被认为是最著名且最常用的聚类算法[13],近年来,许多学者将 K-means 算法及其改进算法用于建筑能耗模式识别中,Do Carmo 等[14]用 K-means 聚类分析了丹麦住宅日热负荷分布模式,确定了两种集群分配模式,工作日和周末集群,负荷分布相对恒定的主集群和变化明显的小集群。文献[15]将 K-shape 算法应用于不同层次的建筑用能模式检测,证明了利用 K-shape 聚类的结果显著提高了支持向量回归模型的预测精度。高斯混合聚类(Gaussian mixture model, GMM)将服从相似正态分布的向量分为一类,因其普适性,文献[16]选用 GMM 聚类方法来识别温度相关子模式和人的行为相关子模式,并进一步利用聚类结果来提高集中供热负荷预测模型的准确性。文献[17]首先介绍了集中供热变换热站的参数化调度策略,然后利用 GMM 聚类方法对这些策略进行了反向识别,进而提出等价供需匹配系数指标,评估管制策略的运作效果和诊断管制策略的无效性。

综上所述,现阶段,针对集中供热分户计量数据的异常用热检测鲜有研究;聚类分析已被成功运用在建筑能源系统的大数据分析中;由于各小区供暖数据特征差异明显,在不知道特征数据的前提下利用聚类算法对数据进行分析,可以简化计算过程,提高精确度。因此,本文拟采用聚类算法 K-means、GMM 对用户侧集中供热数据集进行分析,并针对集中供热系统大数据量、非线性、响应时间长的特征对 GMM 算法进行改进,提出高维高斯混合聚类(high dimensional Gaussian mixture clustering, HGMM)。在实验部分,参与搭建工业大数据分析平台(industrial big data ingestion and analysis platform, IBDP),完成 HGMM 算法准确性的分析,并得到多种异常用热数据集。

2 算法基础

2.1 K-means 算法思想阐述

K-means 算法是一种典型的无监督学习算法[13],主要用于将相似的数据点自动归到一个类别中。该算法计算过程与框架如算法 1 所示。首先确定常数 k,常数 k 意味着最终的聚类类数,可以通过标准测度函数即最小平方和来进行 k 的计算与选取。然后,随机选定 k 个初始点为质心,并通过计算每一个数据点与质心之间的相似度(此处为欧式距离),将数据点归到最相似的类中(即离该点距离最小的质心)。最后,重新计算每个类的质心,重复这样的过程,直到质心不再改变,最终就确定了每个数据点所属的类别以及每个类的质心。

K-means 拟合出来的类为以质心为圆心的圆形,实际数据分布可能是椭圆,因此实际聚类时导致多圆之间的重叠,即多类之间数据交叉造成误差的出现。

算法 1 基于 K-means 算法的异常检测算法

输入:传感器数据集 $\{x_1,x_2,\cdots,x_N\}$

输出:异常点及其所属类

1:根据 K-means 聚类标准测度函数的计算,计算聚类类数 k

2:输入 k 的值,即我们指定希望通过聚类得到 k 个分组

3:在$\{x_1,x_2,\cdots,x_N\}$中随机选取 k 个数据点$\{y_1,y_2,\cdots,y_k\}$作为质心

4:for i=1 to N

5: {

6: for j=1 to k

7: {计算距离 d(x_i,y_j) }

8: 取最小 $d_{\min}(x_i,y_j)$ 时 $y=n$,则 x_i 属于第 n 类

9: }

10: 计算每个分组数据的平均值作为新的质心 $\{z_1,z_2,\cdots,z_k\}$

11: 若 $|y_i-z_i|<T$,则聚类过程结束;否则根据$\{z_1,z_2,\cdots,z_k\}$ 迭代 4 ~ 10 步

12: 数据对象数<阈值的类且结合数据集横向纵向比来进行异常数据集的标记

2.2 GMM 算法思想阐述

GMM 可以理解为是 K-means 算法的一个优化,GMM 基于多变量高斯分布[16],试图找到多维高斯模型概率分布的混合表示,从而拟合出任意形状的数据分布。该算法中心思想即对于某个数据点,把该数据点代入到 K 个高斯分布中,求出该点属于每个类别的概率;然后选择概率值最高的类别作为该数据点的所属类;最后得到每个数据点被分配到每一个类的概率,该聚类方式被称为软聚类。高斯混合聚类计算过程如下所示:

变量 x 的高斯混合模型概率密度函数定义由式(2)所示。

$$p(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\times\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\sum{}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{1}$$

$$\mathrm{P_r}(\boldsymbol{x})=\sum_{k=1}^{K}\pi_k p(\boldsymbol{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \tag{2}$$

其中,$p(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ 表示每个分类的概率密度函数,$\boldsymbol{x}$ 表示数据点向量,$\boldsymbol{\mu}$ 为该类中数据集的平均向量,$\boldsymbol{\Sigma}$ 为该类中数据集的方差,$(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 标识了唯一的类,且表征了该类数据集的组合分布。D 表示变量 $\boldsymbol{x}$ 的维度。K 为分类的个数且该值需要在概率密度函数之前进行计算与设置,π_k 表示第 k 个类的先验概率,当 π_k 大于阈值时,数据点 x 属于第 k 个类。随着 K 值的增大,混合模型更加复杂,K 值足够大时,高斯混合模型可以用来近似任意连续的概率密度分布。

上述参数的估计可以通过最大化生成给定数据的方式来进行计算,即最大化似然函数 $\sum_{i=1}^{N}\log p(\boldsymbol{x}_i)$。上述计算函数中 N 表示数据点个数,$p(\boldsymbol{x}_i)$ 可以通过 EM 算法来进行计算。

期望值的计算过程如式(3)所示。

$$P^{(l)}(\boldsymbol{x}_i\in\boldsymbol{C}_k)=P^{(l)}(\boldsymbol{C}_k\mid\boldsymbol{x}_i)=\frac{\pi_k^{(l)}p(\boldsymbol{x}_i\mid\boldsymbol{\mu}_k^{(l)},\boldsymbol{\Sigma}_k^{(l)})}{\sum_{j=1}^{K}\pi_j^{(l)}p(\boldsymbol{x}_i\mid\boldsymbol{\mu}_j^{(l)},\boldsymbol{\Sigma}_j^{(l)})} \tag{3}$$

式中:l 表示第 l 次迭代;$\boldsymbol{x}_i$ 表示第 i 个数据点向量;$\boldsymbol{C}_k$ 表示第 k 个高斯分类。

最大化的计算过程中,第 k 个分类的概率密度参数如式(4)~(7)所示。

$$\boldsymbol{\mu}_k^{(l+1)}=\frac{\sum_{i=1}^{N}P^{(l)}(\boldsymbol{x}_i\in\boldsymbol{C}_k)\boldsymbol{x}_i}{\sum_{i=1}^{N}P^{(l)}(\boldsymbol{x}_i\in\boldsymbol{C}_k)} \tag{4}$$

$$\boldsymbol{\Sigma}_k^{(l+1)}=\frac{\sum_{i=1}^{N}P^{(l)}(\boldsymbol{x}_i\in\boldsymbol{C}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_{i=1}^{N}P^{(l)}(\boldsymbol{x}_i\in\boldsymbol{C}_k)} \tag{5}$$

$$\pi_k^{(l+1)}=\frac{\max\left\{0,\sum_{i=1}^{N}P^{(l)}(\boldsymbol{x}_i\in\boldsymbol{C}_k)-V/2\right\}}{\sum_{j=1}^{K}\max\left\{0,\sum_{i=1}^{N}P^{(l)}(\boldsymbol{x}_i\in\boldsymbol{C}_j)-V/2\right\}} \tag{6}$$

$$V=\frac{1}{2}D^2+\frac{3}{2}D \tag{7}$$

重复迭代期望值和最大化计算步骤,极大似然函数 $\sum_{i=1}^{N}\log p(\boldsymbol{x}_i)$ 收敛到稳定值,迭代过程停止。

GMM 多针对线性数据集,而本文针对的集中供暖数据集有非线性、响应时间长的特点,因此需要对 GMM 进行改进,使其适应于集中供热数据集的聚类与检测。

3 针对集中供热末端数据的异常检测算法—高维高斯混合聚类(HGMM)

将非线性数据集从低维空间映射到高维空间[18],可以把低维空间中线性不可分的两类点变成高维空间线性可分的集合。本文提出高维高斯混合聚类(HGMM)算法,它的中心思想为:将原数据集由低维空间映射到高维空间,进而进行高斯混合聚类;为了简化高维空间的计算同时避免维数灾难,将高维数据集的内积转化为低维数据集的核计算。HGMM 算法具体计算过程如下所示。

数据集由低维空间映射为高维空间,即 $\boldsymbol{x}\rightarrow\boldsymbol{\varphi}(\boldsymbol{x})$;高维特征空间 $\boldsymbol{\varphi}(\boldsymbol{x})$ 概率密度函数定义为式(9)。

$$p(\boldsymbol{\varphi}(\boldsymbol{x});\tilde{\boldsymbol{\mu}},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{\varphi}(\boldsymbol{x})-\tilde{\boldsymbol{\mu}})^{\mathrm{T}}\sum{}^{-1}(\boldsymbol{\varphi}(\boldsymbol{x})-\tilde{\boldsymbol{\mu}})\right] \tag{8}$$

$$P_{\mathrm{r}}(\boldsymbol{\varphi}(\boldsymbol{x})) = \sum_{k=1}^{K}\tilde{\pi}_k p(\boldsymbol{\varphi}(\boldsymbol{x});\tilde{\boldsymbol{\mu}}_k,\tilde{\boldsymbol{\Sigma}}_k) \tag{9}$$

式中：$\tilde{\boldsymbol{\mu}}$ 为 $\boldsymbol{\varphi}(\boldsymbol{x})$ 的平均向量；$\boldsymbol{\Sigma}$ 为 $\boldsymbol{\varphi}(\boldsymbol{x})$ 的协方差矩阵；$\tilde{\boldsymbol{\mu}}_k$ 为高维空间中各个类的平均向量；$\boldsymbol{\Sigma}_k$ 为高维空间中各个类的协方差矩阵。

高维特征空间中 EM 算法如式(11)~(14)所示。

期望值的计算过程如式(10)所示。

$$\tilde{P}^{(l)}(\boldsymbol{\varphi}(\boldsymbol{x}_i)\in\tilde{\boldsymbol{C}}_k) = \tilde{P}^{(l)}(\tilde{\boldsymbol{C}}_k\mid\boldsymbol{\varphi}(\boldsymbol{x}_i)) = \frac{\tilde{\pi}_k^{(l)}p(\boldsymbol{\varphi}(\boldsymbol{x}_i)\mid\tilde{\boldsymbol{\mu}}_k^{(l)},\boldsymbol{\Sigma}_k^{(l)})}{\sum_{j=1}^{K}\tilde{\pi}_j^{(l)}p(\boldsymbol{\varphi}(\boldsymbol{x}_i)\mid\tilde{\boldsymbol{\mu}}_j^{(l)},\boldsymbol{\Sigma}_j^{(l)})} \tag{10}$$

公式(10)中，l 表示迭代次数，$\boldsymbol{\varphi}(\boldsymbol{x}_i)$ 表示高维空间中第 i 个数据点向量，$\tilde{\boldsymbol{C}}_k$ 表示高维空间中第 k 个高斯分类。

最大化的计算过程如式(11)~(14)所示。

$$\tilde{\boldsymbol{\mu}}_k^{(l+1)} = \frac{\sum_{i=1}^{N}\tilde{P}^{(l)}(\boldsymbol{\varphi}(\boldsymbol{x}_i)\in\tilde{\boldsymbol{C}}_k)\boldsymbol{\varphi}(\boldsymbol{x}_i)}{\sum_{i=1}^{N}\tilde{P}^{(l)}(\boldsymbol{\varphi}(\boldsymbol{x}_i)\in\tilde{\boldsymbol{C}}_k)} \tag{11}$$

$$\boldsymbol{\Sigma}_k^{(l+1)} = \frac{\sum_{i=1}^{N}\tilde{P}^{(l)}(\boldsymbol{\varphi}(\boldsymbol{x}_i)\in\tilde{\boldsymbol{C}}_k)(\boldsymbol{\varphi}(\boldsymbol{x}_i)-\tilde{\boldsymbol{\mu}}_k)(\boldsymbol{\varphi}(\boldsymbol{x}_i)-\tilde{\boldsymbol{\mu}}_k)^{\mathrm{T}}}{\sum_{i=1}^{N}\tilde{P}^{(l)}(\boldsymbol{\varphi}(\boldsymbol{x}_i)\in\tilde{\boldsymbol{C}}_k)} \tag{12}$$

$$\tilde{\pi}_k^{(l+1)} = \frac{\max\left\{0,\sum_{i=1}^{N}\tilde{P}^{(l)}(\boldsymbol{\varphi}(\boldsymbol{x}_i)\in\tilde{\boldsymbol{C}}_k)-\tilde{V}/2\right\}}{\sum_{j=1}^{K}\max\left\{0,\sum_{i=1}^{N}\tilde{P}^{(l)}(\boldsymbol{\varphi}(\boldsymbol{x}_i)\in\tilde{\boldsymbol{C}}_j)-\tilde{V}/2\right\}} \tag{13}$$

$$\tilde{V} = \frac{1}{2}\tilde{D}^2+\frac{3}{2}\tilde{D} \tag{14}$$

式(14)中，参数 $\tilde{D}$ 表示 $\boldsymbol{\varphi}(\boldsymbol{x})$ 的维度。

由于本算法针对集中供热用户侧数据，数据点多达数十万条，直接通过最大化似然函数估计的方法对高维空间计算会产生维数灾难。为了减少计算量同时避免维数灾难，运用核映射，将高维数据集的内积转化为低维数据集的核计算，如式(15)所示。

$$\boldsymbol{G}_{ij} = \boldsymbol{G}(\boldsymbol{x}_i,\boldsymbol{x}_j) = \langle\boldsymbol{\varphi}(\boldsymbol{x}_i),\boldsymbol{\varphi}(\boldsymbol{x}_j)\rangle \tag{15}$$

式(15)中，$\boldsymbol{G}(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 表示核矩阵，$\langle\boldsymbol{\varphi}(\boldsymbol{x}_i),\boldsymbol{\varphi}(\boldsymbol{x}_j)\rangle$ 表示 $\boldsymbol{\varphi}(\boldsymbol{x}_i)$ 和 $\boldsymbol{\varphi}(\boldsymbol{x}_j)$ 的内积。通过上述映射，可以利用原低维空间内积之间的计算取代式(10)~(14)中高维空间中的计算。为了简化计算过程，本文选取参数复杂度低、具有普适性的核函数-径向基函数(radial basis function，RBF)来进行相应的非线性变换。

径向基函数的变化过程如式(16)所示。

$$\boldsymbol{G}_{ij} = \boldsymbol{G}(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\lambda^2}\right) \tag{16}$$

式(16)中，定义参数 $\lambda=1$，用来调整径向基函数的宽度。

$$\overline{\boldsymbol{G}} = \boldsymbol{G}-\boldsymbol{E}_N\boldsymbol{G}-\boldsymbol{G}\boldsymbol{E}_N+\boldsymbol{E}_N\boldsymbol{G}\boldsymbol{E}_N \tag{17}$$

$$(\tilde{\boldsymbol{G}}_k^{(l)})_{ij} = \pi_{ki}^{(l)}\pi_{kj}^{(l)}\overline{\boldsymbol{G}}_{ij} \tag{18}$$

$$(\hat{\boldsymbol{G}}_k^{(l)})_{ij} = \pi_{kj}^{(l)}\overline{\boldsymbol{G}}_{ij} \tag{19}$$

$$\pi_{ki}^{(l)} = \left\{\frac{P^{(l)}(\tilde{\boldsymbol{C}}_k\mid\boldsymbol{\varphi}(\boldsymbol{x}_i))}{\sum_{m=1}^{K}P^{(l)}(\tilde{\boldsymbol{C}}_m\mid\boldsymbol{\varphi}(\boldsymbol{x}_i))}\right\}^{1/2} \tag{20}$$

式(17)中，参数 $\overline{\boldsymbol{G}}$ 定义为中心核矩阵的平均值，$\boldsymbol{E}_N$ 定义为所有元素均为 $1/N$ 的 $N\times N$ 矩阵。式(18)中，参数 $\tilde{\boldsymbol{G}}_k^{(l)}$ 定义为第 l 次迭代过程中第 k 个分类的加权核矩阵。式(19)中，$\hat{\boldsymbol{G}}_k^{(l)}$ 定义为第 l 次迭代过程中第 k 个分类的加权映射核矩阵。

将高维特征空间分解为两部分：主空间、残差子空间。残差部分影响较小可忽略。针对第 k 个聚类，$\boldsymbol{\varphi}(\boldsymbol{x})$ 的概率密度函数可以简化为式(21)所示。

$$p(\boldsymbol{\varphi}(\boldsymbol{x}_i)\mid\tilde{\boldsymbol{\mu}}_k^{(l)},\boldsymbol{\Sigma}_k^{(l)}) = \frac{1}{(2\pi)^{\tilde{D}/2}\prod_{s=1}^{\tilde{D}}(\alpha_{ks}^{(l)})^{1/2}}\exp\left[-\frac{1}{2}\sum_{s=1}\frac{x_s^2}{\alpha_{ks}^{(l)}}\right] \tag{21}$$

$$x_s = (\boldsymbol{\xi}_{ks}^{(l)})^{\mathrm{T}}\boldsymbol{\eta}_i^{(l)} \tag{22}$$

式(21)中，参数 $\alpha_{ks}^{(l)}$ 定义为 $\tilde{\boldsymbol{G}}_{\mathrm{k}}^{(l)}$ 的第 s 个特征值；式(22)中，参数 $\boldsymbol{\xi}_{ks}$ 定义为 $\tilde{\boldsymbol{G}}_{\mathrm{k}}^{(l)}$ 的特征向量，$\boldsymbol{\eta}_i^{(l)}$ 定义为 $\hat{\boldsymbol{G}}_{\mathrm{k}}^{(l)}$ 的第 i 个列向量。

$P^{(l)}(\boldsymbol{\varphi}(\boldsymbol{x}_i)\in\tilde{\boldsymbol{C}}_k)$ 收敛，则迭代过程停止。

4 实验与结果分析

4.1 实验环境与平台

本实验所用平台为工业大数据采集分析平台(IBDP)，该平台系统架构如图 1 所示。本文集中供热末端数据为关系型数据，数据源来自数据库中数据表格，数据获取层采用 Sqoop 将数据批量导入 Hadoop 的 HDFS 中，数据分析层采用 Spark 计算框架，使用 MLlib 库中 GMM 算法并对其进行改进得到 HGMM，实现模型的计算。

图 2 为供热数据采集、计量、监控和分析平台，采用浏览器和服务器架构，可实时读取住户信息、热量表、温度采集器等的实时数据；本文 IBDP 平台数据源为该监

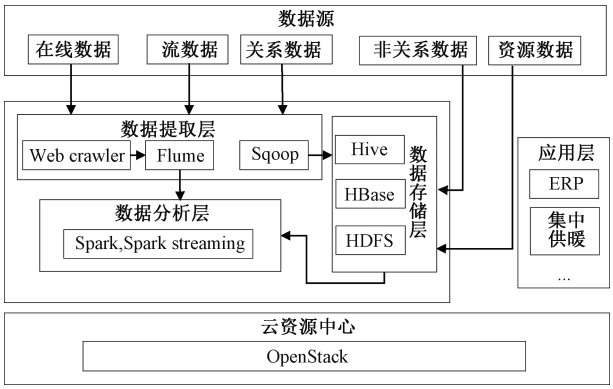

图 1 工业大数据采集分析平台系统架构图

Fig. 1 System architecture of industrial big data ingestion and analysis platform

分别使用 K-means、GMM、恒虚警率、HGMM 3 类聚类算法对数据进行分析与对比。

图 2 济南市某小区监控平台示意图

Fig. 2 Schematic diagram of the monitoring platform for a community in Ji′nan

控分析平台数据库。为了使数据更有普适性,本次实验选取济南市两类小区的数据对比分析,A 类为建于 2000 年左右暖气片供暖的多层住宅,选取了 50 户用户进行分析,B 类为建于 2010 年后地面辐射供暖的高层住宅,选取了 320 户用户数据进行分析。针对上述住户,选取 2018 年的集中供热数据(2018 年 11 月 15 日~2019 年 3 月 15 日,进水温度,回水温度,瞬时流量,热功率),以小时为时间步长,约 100 万个数据点。针对上述数据,

4.2 异常检测算法验证

经过 HGMM 聚类,将数据集原空间维度 $D=4$,经过 2 次多项式核计算,得到维度 $\tilde{D}=16$ 的数据集。通过对两类小区供热数据分析得到表 2 检测结果,表中给出了 K-means、GMM 和 HGMM 聚类下,异常检测结果的准确性(DR)和误报率(FPR),同时与文献[19]中提出的恒虚警率异常检测算法进行对比,该算法将 GMM 聚类与相似性度量结合进行异常模式的检测。

表 1 K-means, GMM, 恒虚警率和 HGMM 异常检测算法的 DR 和 FPR 对比

Table 1 Comparison of DR and FPR of K-means, GMM, constant false alarm rate and HGMM detection algorithms %

小区 ID (用户数)	K-Means		GMM		恒虚警率		HGMM	
	DR	FPR	DR	FPR	DR	FPR	DR	FPR
A1 (20)	75.22	12.1	82.33	9.92	85.53	6.74	87.13	6.84
A2 (30)	75.14	12.51	82.43	9.78	86.15	6.45	88.55	6.65
B1 (120)	80.13	10.1	85.54	9.08	87.56	6.02	90.72	5.92
B2 (200)	80.03	10.14	84.66	9.38	86.45	6.03	90.45	6.05

结合表 1 可以看出,针对评估指标 DR,K-means<GMM<恒虚警率<HGMM;针对 FPR,K-means>GMM>HGMM≈恒虚警率。HGMM 算法检测结果明显优于 GMM 聚类,HGMM 聚类算法异常检出率高达 90.72%,误报率仅为 5.92%。同时,HGMM 和恒虚警率检测相比误报率大小无差别,但由于恒虚警率检测在原空间进行相似度度量,数据集间有一定交叉,而 HGMM 算法对数据空间的映射,使得多类别之间离散度增加,使得聚类准确性有一定程度提升。通过 A、B 类小区的数据对比,B 类小区检测结果优于 A 类小区,归纳原因在于现代住宅建筑结构稳定、材料系数稳定、传感器精度高。综上所述,HGMM 算法将供暖非线性数据映射到高维核特征空间,使其线性分离,优于 K-means、GMM、恒虚警率方法;同时该算法更适用于新建住宅,应用前景广泛。

4.3 异常用热数据与用户行为分析

针对 A 小区,跟据 K-means 标准测度函数的计算,K 值为 6 时,标准测度函数斜率最大,因此 GMM 聚类和 HGMM 聚类的 K 值为 6,即分为 6 类,聚类结果如表 2 所示。表中权重值表示该类中数据量占所有数据点总数的百分比。其中,类 0~3 权重值明显偏低,分别为 5.3%,9.7%,2.9%,2.3%,类 4、5 权重值分别为 34.2%,45.6%。因类 0~3 权重值低于类 4 权重值的 30%,将类 0~3 初步判定为异常用热数据,后面内容将对各个类数据进行详细解释与分析。

表 2 A 类住宅小区 GMM 和 HGMM 算法聚类结果对比

Table 2 Comparison of results of GMM and HGMM clustering for residential community A

集群	GMM					HGMM				
	权重值/%	平均向量				权重值/%	平均向量			
		进水温/℃	回水温/℃	流量/($m^3 \cdot h^{-1}$)	热功率/($kW \cdot m^{-2}$)		进水温/℃	回水温/℃	流量/($m^3 \cdot h^{-1}$)	热功率/($kW \cdot m^{-2}$)
0	4.7	62.3	48.6	0.83	0.063	5.3	61.3	46.7	0.86	0.061
1	9.4	62.2	46.7	0.46	0.055	9.7	61.5	46.5	0.48	0.054
2	3.1	37.9	66.9	0.46	0	2.9	37.3	67.9	0.46	0
3	2.7	51.1	60.5	0.46	0.032	2.3	49.9	59.8	0.47	0.037
4	36.1	58.9	50.3	0.45	0.029	34.2	58.3	50.2	0.43	0.027
5	44.0	63.5	52.3	0.59	0.047	45.6	63.5	52.3	0.59	0.045

集群 0 中,供水流量偏高约 0.86 m^3/h,温差偏大约 14.6℃(进水温度正常约 61.3℃,回水温度偏低约 46.7℃),热功率偏高约 0.06 $kW \cdot m^{-2}$。选取存在此集群数据的某一住户,该住户供暖数据和其同楼层住户、楼上住户数据对比图如图 3~5 所示,图中横坐标为时间,每个点对应纵坐标为该住户该集群多天该时间段平均值。该类数据可解释为私自放出管道热水,造成流量偏高,且回水较长时间静止造成管道散热量大,回水温度低。

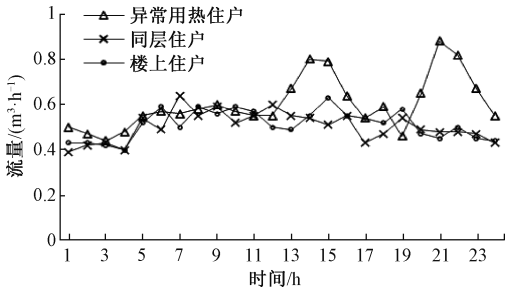

图 3 集群 0 异常用热住户及其临近住户供水流量对比图

Fig. 3 Comparison of supply water's volume flowrate between abnormal heat users and nearby households in cluster 0

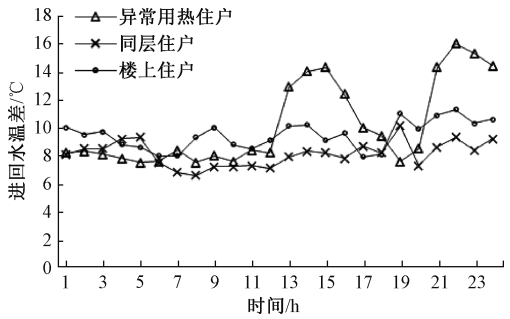

图 4 集群 0 异常用热住户及其临近住户供回水温差对比图

Fig. 4 Comparison of supply and return water temperature difference between abnormal heat users and nearby households in cluster 0

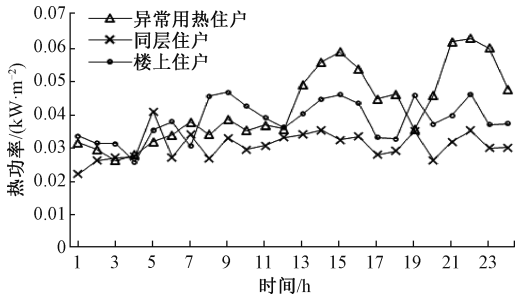

图 5 集群 0 异常用热住户及其临近住户热功率对比图

Fig. 5 Comparison of thermal power between abnormal heat users and adjacent households in cluster 0

集群 1 中,供水流量属于正常范围,温差偏大约 15℃(进水温度正常约 61.5℃,回水温度偏低约 46.5℃),热功率偏高约 0.054 $kW \cdot m^{-2}$。选取存在此集群数据的某一住户,该住户供暖数据和其同楼层住户、楼上住户数据对比图如图 6~7 所示,结合其异常数据出现时间,该住户应为工作时间开窗通风换气住户。该类数据可解释为开窗散热或者安装换热器等换热设备,在管道流量不变的情况下,增加管道散热,造成回水温度偏低。

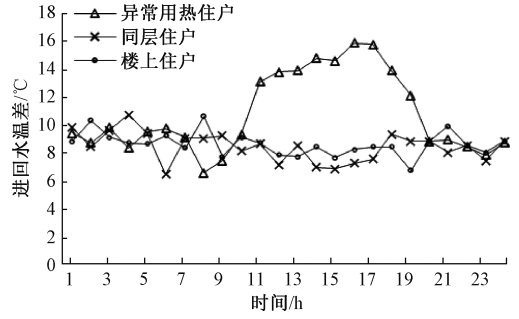

图 6 集群 1 异常用热住户及其临近住户供回水温差对比图

Fig. 6 Comparison of supply and return water temperature difference between different common heat households in cluster 1 and their neighboring households

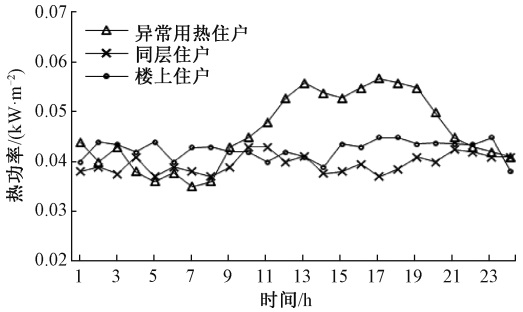

图 7 集群 1 异常用热住户及其临近住户热功率对比图

Fig. 7 Comparison of thermal power between abnormal heat users and adjacent households in cluster 1

集群 2 中,供水流量属于正常范围,温差为负值(回水温度过高平均值为 67.9℃或进水温度过低平均值为 37.3℃),热功率为 0 kW·m^{-2}。该类数据可解释为测温探头损坏,这种情况下,因回水温度或进水温度过低,数据库中计算时判断为异常,将热功率输出为 0 值。

集群 3 中,供水流量属于正常范围,温差为负值(进回水温度均在正常值范围内),热功率为正常值约 0.037 kW·m^{-2}。该类数据可解释为进回水管道测温探头装反,数据库计算时直接以温差绝对值来计算,因此热功率值在正常范围内。

集群 4 和集群 5 中,进回水温度、流量和热功率数值均在正常范围内。集群 4 中进回水温差约为 6℃,低于集群 5 中进回水温差 11℃,查询数据发现,集群 4 中多为供暖季开始和结束多半个月时间内的用户侧供暖数据,因室外温度偏高,室内需热量较少。集群 5 权重最大,供水温度高于集群 4,其功率值为供暖季室外温度偏低时的室内耗热量。

5 结 论

集中供热末端住户侧数据因其数据量大、强非线性、响应时间长等特征,在原空间中对数据进行聚类分析易造成类间数据交叉严重。本文提出了一种高维空间聚类算法(HGMM),将数据集映射到高维空间进行聚类,提高聚类精确度;利用核函数映射、内积运算与高维特征空间分解等,规避维数灾难,简化计算过程。

实验过程中,搭建工业大数据分析平台,在 Spark 计算框架下,利用 MLlib 算法库,并对库中 GMM 算法进行改进,实现了 K-means、GMM、恒虚警率、HGMM 4 种聚类算法对集中供热末端住户侧数据的聚类分析。

实验结果表明,HGMM 算法能有效解决原低维空间中多类数据交叉降低聚类精确度的问题,将聚类准确性提高到 90.72%,聚类误报率降低到 5.92%,且该算法更适合于新建住宅。结合 HGMM 聚类结果,对集中供热末端住户数据中异常用热和行为特征进行分析,HGMM 聚类将末端数据分为 6 类:私放热水,开窗散热或装散热器,测温探头损坏,测温探头装反,供暖季开始和结束时正常用热,较冷月份正常用热。上述聚类结果中,前四类为异常用热数据集。

综上所述,本文提出高维空间聚类算法(HGMM)有效分析用户侧用热特征、检测异常用热数据,且更适用于现代住宅,辅助提高系统运行效率,降低能耗,应用前景广泛。

参考文献

[1] MIRNAGHI M S, HAGHIGHAT F. Fault detection and diagnosis of large-scale HVAC systems in buildings using data-driven methods: A comprehensive review [J]. Energy and Buildings, 2020, 229(1): 110492.

[2] HABEEB R, NASARUDDIN F, GANI A, et al. Real-time big data processing for anomaly detection: A Survey[J]. International Journal of Information Management, 2019, 45(4): 289-307.

[3] 吕侃徽, 张大兴. 基于改进 Hough 变换耦合密度空间聚类的车道线检测算法[J]. 电子测量与仪器学报, 2020, 34(12): 172-180.
LU K H, ZHANG D X. Lane detection algorithm based on improved hough transform coupled density space clustering[J]. Journal of Electronic Measurement and Instrumentation, 2020, 34(12): 172-180.

[4] 苏鹏, 王常顺, 卢萌萌. 基于变分自编码器的视频异常事件检测方法[J]. 电子测量与仪器学报, 2020, 34(10): 179-185.
SU P, WANG CH SH, LU M M. Video anomaly detection and localization via variational autoencoder[J]. Journal of Electronic Measurement and Instrumentation, 2020, 34(10): 179-185.

[5] RAVANBAKHSH M, NABI M, MOUSAVI H, et al. Plug-and-Play CNN for crowd motion analysis: An application in abnormal event detection [C]. IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018.

[6] 王小飞, 王元鑫, 曲建岭, 等. 面向大样本飞参数据的航空发动机性能监控方法[J]. 仪器仪表学报, 2020, 41(7): 175-184.
WANG X F, WANG Y X, QU J L, et al. An aero-engine performance monitoring method based on large scale flight data [J]. Chinese Journal of Scientific Instrument, 2020, 41(7): 175-184.

[7] CHEN Z, YEO C K, LEE B S, et al. Evolutionary multi-objective optimization based ensemble autoencoders for image outlier detection[J]. Neurocomputing, 2018,

309: 192-200.

[8] 张栋, 杜康, 韩文念, 等. 基于级联卷积神经网络的荧光免疫层析图像峰值点定位方法研究[J]. 仪器仪表学报, 2021, 42(1): 217-227.
ZHANG D, DU K, HAN W N, et al. Peak point location of fluorescence immunochromatography image based on the cascaded convolutional neural network[J]. Chinese Journal of Scientific Instrument, 2021, 42(1): 217-227.

[9] DEB C, LEE S E. Determining key variables influencing energy consumption in office buildings through cluster analysis of pre- and post-retrofit building data[J]. Energy and Buildings, 2018, 159: 228-245.

[10] GUNAY H B, SHI Z. Cluster analysis-based anomaly detection in building automation systems[J]. Energy and Buildings, 2020, 228: 110445.

[11] LI K H, MA Z J, ROBINSON D, et al. A data-driven strategy to forecast next-day electricity usage and peak electricity demand of a building portfolio using cluster analysis, cubist regression models and particle swarm optimization[J]. Journal of Cleaner Production, 2020, 273: 123115.

[12] PAN D, XIE K G, GUO T T, et al. Short-term load forecasting for electric power systems using the PSO-SVR and FCM clustering techniques[J]. Energies, 2010, 4: 173-184.

[13] RAMOS S, DUARTE J, DUARTE F J, et al. A data-mining based methodology to support MV electricity custormers' characterization[J]. Energy and Buildings, 2015, 91: 16-25.

[14] DO CARMO C M R, CHRISTENSEN T H. Cluster analysis of residential heat load profiles and the role of technical and household characteristics[J]. Energy and Buildings, 2016, 125: 171-180.

[15] YANG J J, NING C, DEB C, et al. K-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement[J]. Energy and Buildings, 2017, 146: 27-37.

[16] LU Y, ZHE T, PENG P, et al. GMM clustering for heating load patterns in-depth identification and prediction model accuracy improvement of district heating system[J]. Energy and Buildings, 2019, 190: 49-60.

[17] LU Y, ZHE T, PENG P, et al. Identification and evaluation of operation regulation strategies in district heating substations based on an unsupervised data mining method[J]. Energy and Buildings, 2019, 202: 109324.

[18] MA X, AMINIAN M, KIRBY M. Error-adaptive modeling of streaming time-series data using radial basis functions[J]. Journal of Computational & Applied Mathematics, 2019, 362: 295-308.

[19] 王培志, 张一迪, 杨沁, 等. 非侵入式交通信号灯异常检测[J]. 传感技术学报, 2019, 32(12): 1911-1916.
WANG P ZH, ZHANG Y D, YANG Q, et al. Non-invasive anomaly detection for traffic light[J]. Chinese Journal of Sensors and Actuators, 2019, 32(12): 1911-1916.

作者简介

孙文慧,2011 年于齐鲁工业大学获得学士学位,2014 年于东华大学获得硕士学位,2018 年于山东大学获得博士学位,现为山东建筑大学讲师,主要研究方向为集中供热异常检测与热负荷预测。
E-mail: sunwenhui1203@163.com

Sun Wenhui received her B. Sc. degree from Qilu University of Technology in 2011, received her M. Sc. degree from Donghua University in 2014, and received her Ph. D. degree from Shandong University in 2018. She is currently a lecturer at Shandong Jianzhu University. Her main research interests include anomaly detection of and heat load forecast in district heating system.

张海伦(通信作者),2011 年于齐鲁工业大学获得学士学位,2014 年于齐鲁工业大学获得硕士学位,2018 年于山东大学获得博士学位,现为山东大学博士后,主要研究方向为喷射理论与燃料电池系统热管理与控制。
E-mail: zhl_sdu@163.com

Zhang Hailun (Corresponding author) received his B. Sc. degree and M. Sc. degree both from Qilu University of Technology in 2011 and 2014, and received his Ph. D. degree from Shandong University in 2018. He is currently a postdoctoral scholar at Shandong University. His main research interests include jet theory, PEMFC thermal management and system control.

王雷,1993 年于山东工业大学获得学士学位,2000 年于山东大学获得硕士学位,2004 年于浙江大学获得博士学位,现为山东大学教授,主要研究方向为现代检测技术、多相流检测、低品位能源利用中的控制问题。
E-mail: leiwang@sdu.edu.cn

Wang Lei received his B. Sc. degree from Shandong Polytechnic University in 1993, received his M. Sc. degree from Shandong University in 2000, and received his Ph. D. degree from Zhejiang University in 2004. He is currently a professor at Shandong University. His main research interests include modern detection technology, multiphase flow detection and the control problems in low-grade energy resource utilization.