文章编号:1671-4598(2020)07-0253-07 DOI:10.16526/j.cnki.11—4762/tp.2020.07.051 中图分类号:TP391 文献标识码:A

基于数据挖掘的建筑能耗异常检测研究

段中兴[1,2], 梅思雨[1]
(1. 西安建筑科技大学 信息与控制工程学院, 西安 710055;
2. 西部绿色建筑国家重点实验室, 西安 710055)

摘要: 建筑能耗异常检测对于建筑管理和运行至关重要, 论文提出了一种基于D—S证据理论的不平衡数据多划分(Multi—partition, MP)聚类算法, 并构建MP算法能耗异常检测模型对建筑能耗中的异常值进行准确检测; 首先通过改进的信任c均值算法将能耗数据集多划分; 利用基于K—NN的均值漂移算法确定数据集的真实类别个数; 然后根据密度合并规则对能耗数据进行合并; 最后对未合并的能耗数据再次划分得到最终的能耗异常检测结果; UCI数据集验证结果表明, MP算法对于不平衡数据聚类效果良好, 能够有效避免样本"均匀效应", 降低错误率; 通过对某大型商场建筑空调和照明用电能耗异常值检测, 验证了MP算法能耗异常检测模型的有效性。

关键词: 能耗异常检测; D—S证据理论; 不平衡数据; 聚类; 异常检测模型

Research on Abnormal Detection of Building Energy Consumption Based on Data Mining

Duan Zhongxing[1,2], Mei Siyu[1]
(1. School of Information & Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China;
2. State Key Laboratory of Green Building in Western China, Xi'an University of Architecture and Technology, Xi'an 710055, China)

Abstract: Abnormal detection of building energy consumption is very important for building management and operation. In this paper, a multi—partition (MP) clustering algorithm for imbalanced data based on D—S evidence theory is proposed, and the energy consumption anomaly detection model of MP algorithm is constructed to accurately detect the abnormal values in building energy consumption. Firstly, the energy consumption data set is divided into multiple parts by the improved credal c—means algorithm. The KNN—based Mean—shift algorithm is used to determine the number of real categories of the data set. Then the energy consumption data is merged according to the density merging rules. Finally, the energy consumption data that is not merged is divided again to get the final abnormal energy consumption detection results. The UCI data set verification results show that the MP algorithm has a good clustering effect for imbalanced data, which can effectively avoid the " uniform effect" of samples and reduce the error rate. Through detecting the abnormal values of the energy consumption of air conditioning and lighting in a large shopping mall, the validity of the energy consumption anomaly detection model of MP algorithm is verified.

Keywords: energy consumption anomaly detection; D—S evidence theory; uncertainty; imbalanced data; clustering; anomaly detection model

0 引言

随着建筑行业的不断发展, 建筑能耗监管系统的运行使海量的能耗数据在数据库中不断积累, 由于能耗监管系统异常、设备存在故障等问题, 建筑能耗数据中往往存在异常值, 利用数据挖掘的方法寻找海量能耗数据中存在的异常能耗并对这些能耗异常值进行分析, 有助于建筑运营管理者及时发现和解决建筑运行过程中可能存在的问题, 针对性地对建筑内部产生的故障进行诊断。目前, 已有许多学者针对能耗异常检测开展了大量研究工作。例如, 文献[1]提出了一种基于数据挖掘技术的建筑能耗实时监测方法, 通过将DBSCAN算法与分类方法相结合, 对建筑能耗值进行类别提取并识别出新产生能耗值所属类别, 从而判断其是否为异常值; 文献[2]在广义离群值检测(GESD)的基础上改进得到了Modified z—score算法, 该算法在检测离群点的同时能够反映出离群数据的离散程度, 适合于建筑能耗数据的检测。这些方法虽然能够实现对建筑异常能耗数据的检测, 但当样本空间密度分布不均或类间距差异很大时, 检测结果会出现偏差, 且不能对能耗数据进行快速处理。从能耗数据本身看, 其中异常值仅在整个能耗数据中占很小的比例, 即正常和异常能耗数据在数量上存在很大差异, 属于不平衡数据类型, 那么对于能耗异常检测问题, 实质上则可以看作不平衡数据聚类, 通过对能耗数据聚类, 得到正常能耗(多数类)和异常能耗(少数类)类别, 从而有效检测出能耗数据中的异常值, 并

收稿日期:2020-04-25; **修回日期**:2020-05-29。
基金项目:国家自然科学基金资助项目(51678470)。
作者简介:段中兴(1969-),男,湖南株洲人,院长,博士生导师,教授,主要从事智能系统与智能信息处理、智能检测与机器视觉、建筑环境控制与节能优化、嵌入式技术等方向的研究。
梅思雨(1995-),女,陕西咸阳人,硕士研究生,主要从事机器学习、模式识别等方向的研究。

给出针对性的诊断。不平衡数据，即数据集中不同类别所含样本在数量上存在很大差异，或不同类别所含样本数量相同但分布不均匀，是数据集中普遍存在的一种数据类型，存在于实际生活中的各个领域（如欺诈检测、网络入侵、医疗检查等）。目前已有的大多数经典聚类方法对于平衡数据聚类能够得到较好的聚类效果，但对于不平衡数据的聚类效果不理想，往往会产生样本“均匀效应”，比如模糊 c 均值（FCM）[3]聚类算法在聚类过程中会均衡化各类别样本数量，使来自多数类中的部分样本被误划分到与其相邻的少数类中，造成很高的误分率。为了避免这个问题，一些学者对此提出了不同的解决思路，例如文献［4］提出了一种多聚类中心算法，通过将样本数量多的类别拆分为若干个类别来减弱不同类别之间的不平衡，避免“均匀效应”。但该算法只适用于不同类别特征之间有明显差异的场景，如果不同类别之间存在数据重叠现象则会产生不理想的聚类效果。Gustafson－Kessel（GK）算法[5]利用马氏距离代替了 FCM 目标函数中的欧式距离，考虑了除球形数据以外的其他簇形对聚类结果产生的影响。

针对以上问题，本文在 D－S 证据理论框架下提出一种不平衡数据多划分（Multi－partition，MP）聚类算法，并将其应用到建筑能耗异常检测中，构建 MP 算法能耗异常检测模型对建筑能耗中的异常值进行检测。实验表明，该算法能够有效避免样本“均匀效应”，极大降低误分率；通过对某商场建筑用电能耗异常值的检测，验证了 MP 算法能耗异常检测模型的有效性。

1 D－S 证据理论概述

Dempster－Shafer（D－S）理论又称证据推理（Evidence Reasoning），1967 年由 Dempster 最先提出[6]，后由 Shafer 于 1976 年对其进行推广形成证据推理理论[7]。在 Shafer 模型中，定义了一个包含了有限个互斥且完备的元素集合 $\Omega=\{\omega_1,\omega_2,\cdots,\omega_n\}$，$\Omega$ 所有子集构成的集合称为 Ω 的幂集，表示为 2^Ω（包含 $2^{|\Omega|}$ 个元素，其中 $|\Omega|$ 表示集合 Ω 中的元素个数）。例如，若辨识框架为 $\Omega=\{\omega_1,\omega_2,\omega_3\}$，则 $2^\Omega=\{\phi,\omega_1,\omega_2,\omega_3,\omega_1\cup\omega_2,\omega_1\cup\omega_3,\omega_2\cup\omega_3,\Omega\}$（其中 $|\Omega|=3$，包含 $2^3=8$ 个元素）。在 Shafer 模型中，从 2^Ω 到［0,1］上的一个映射函数 $m()$ 为一个证据的基本信任指派（basic belief assignment，bba），其满足以下条件：

$$\begin{cases}\sum\limits_{A\in 2^\Omega} m(A)=1\\ m(\phi)=0\end{cases} \tag{1}$$

D－S 理论将传统的辨识框架 Ω 扩展到幂集 2^Ω，使样本类别信息更加丰富（可以属于单类或由若干单类构成的复合类），其优势在于能够满足比概率理论更弱的条件并具有直接表达不确定的能力，因此在模式识别、信息融合领域得到了广泛应用。本文提出的 MP 聚类算法则基于 D－S 证据理论，引入复合类对不确定样本进行了合理表征。

2 不平衡数据多划分（MP）聚类算法

为了对不平衡数据进行有效聚类，避免样本“均匀效应”，本文提出一种基于 D－S 证据理论的不平衡数据多划分（Multi－partition，MP）聚类算法，能够有效处理不平衡数据集，合理表征处在不同类别边缘的不确定样本，极大降低误分率。该算法包含四个子步骤：数据集多划分、真实类别寻找、子数据集合并和剩余数据划分，下面将对 MP 算法的每个子步骤进行详细阐述。

2.1 数据集多划分

MP 聚类算法的第一个步骤即对不平衡数据集中的样本进行子簇划分，受 CCM 算法[8]的启发，本节将提出一种改进的信任 c 均值（Improved credal c－means，ICCM）聚类算法，利用多聚类中心思想（生成多个子簇和若干个复合类，其中子簇个数 N 应大于数据集真实类别个数 c，即 $N>c$），对不平衡数据集中多数类和少数类中的样本数量重新进行平衡，从而有效降低错误率，避免“均匀效应”。由于复合类的引入，那些处在重叠区域的不确定样本能够被合理表征，且 ICCM 的计算复杂度远小于 CCM。对于一个辨识框架为 2^Ω（$\Omega=\{\omega_1,\cdots,\omega_N\}$）的 N 类问题，ICCM 算法分为以下两个部分。

1）子簇的划分：

在这部分中，不平衡数据集中的样本仅允许被划分到子簇和噪声类中，对于一个数据集 $X\in\boldsymbol{R}^{n\times p}$，通过 ICCM 算法对目标函数的最小化将 X 划分为 N 个子簇，能够得到基本信任值 $M=(m_1,\cdots,m_n)\in\boldsymbol{R}^{n\times(N+1)}$ 和矩阵规模为 $N\times p$ 的聚类中心矩阵 $\boldsymbol{V}$，其中 ICCM 算法的目标函数 J_{ICCM} 被定义如下：

$$J_{ICCM}(M,\boldsymbol{V})=\sum_{i=1}^{n}\sum_{i=1}^{n}\delta^2 m_{i\phi}^{\beta}+\sum_{j/A_j\in\Omega} m_{ij}^{\beta}d_{ij}^2 \tag{2}$$

且需满足以下的约束条件：

$$\sum_{j/A_j\in\Omega} m_{ij}+m_{i\phi}=1 \tag{3}$$

其中：$m_{i\phi}$ 表示样本属于噪声类的基本信任值 $m_i(\phi)$，$1\leqslant i\leqslant n$，$1\leqslant j\leqslant N$。参数 β 和 δ 的含义与 CCM 中参数的含义相同，其中 δ 用来控制噪声样本的数量，β 为加权指数（默认值 $\beta=2$）。目标函数 J_{ICCM} 最小化过程类似 FCM 和 CCM，基本信任值 $m(.)$ 通过以下公式更新：

$$m_{ij}=\frac{d_{ij}^{\frac{2}{\beta-1}}}{\sum\limits_{k=1}^{N} d_{ik}^{\frac{2}{\beta-1}}+\delta^{\frac{2}{\beta-1}}} \tag{4}$$

其中：m_{ij} 表示样本 x_i 属于子簇 ω_j 的基本信任值。

2）复合类的产生：

此过程通过设定复合类阈值计算得到复合类的基本信任值，对于样本 x_i，其可能所属的复合类 Λ_i（$\Lambda_i\in 2^\Omega$）被定义如下：

$$\Lambda_i=\{\omega_k\cup,\cdots,\cup\omega_t\,|\,m_i(\omega_k)-m_i(\omega_t)\leqslant\varepsilon\},k\neq t \tag{5}$$

且需满足：

$$m_i(\omega_k)=\max\{m_i(\omega_1),\cdots,m_i(\omega_N)\},?1\leqslant k,t\leqslant N \tag{6}$$

其中：ε 为可调节的复合类阈值，其值大小决定了划分到复合类中的样本数量。对于样本 x_i，其辨识框架拓展为 $\Theta_i=\{\phi,\omega_1,\cdots,\omega_N,\Lambda_i\}$，且不同样本可能得到不同的辨识框

架 Θ，复合类 Λ_i 的基本信任值 $m(\Lambda_i)$ 被定义如下：

$$m(\Lambda_i) = m_i(\omega_k) +, \cdots, + m_i(\omega_t) \tag{7}$$

样本 x_i 通过以下公式对基本信任值 $m(.)$ 归一化并进行更新：

$$m^{\Theta_i}(A) = \frac{m(A)}{\sum_{k=1}^{N} m_i(\omega_k) + m_i(\phi) + m(\Lambda_i)}, A \in \Theta_i \tag{8}$$

其中：$m(A)$ 通过公式（4）计算可得，通过寻找基本信任值中的最大值，将样本 x_i 划分到子簇或者复合类中，这样就可以得到经过 ICCM 划分后的子簇和复合类。

ICCM 算法能够减小由于不同类别样本数量不等或分布不均对结果造成的影响，且能有效避免 CCM“指数爆炸”现象，降低计算复杂度，实现数据的快速处理。在后面的步骤中，将利用子簇和复合类之间的密度关系对划分的子簇进行合并，所提密度合并规则仅允许复合类中所包含单类个数为 2，即样本 x_i 可能所属的复合类 Λ_i 在满足阈值 ε 的条件下仅能包含两个子簇。

复合类阈值 ε 的参数调整规则：在实际应用中，阈值 ε 需要被控制在一个合理范围之内，ε 过大将使原本属于子簇的样本被划分到复合类中，导致不精确率增大；而 ε 过小则会导致复合类中的样本数量极少，极大增加误划分的风险。根据实验，建议阈值 ε 的取值范围为 $\varepsilon \in [0.1, 0.3]$，默认值 $\varepsilon = 0.2$。

2.2　真实类别寻找

利用 ICCM 对不平衡数据集进行划分得到了 N 个子簇和若干个复合类，本节需要对数据集的真实类别个数进行确认，以确保子簇合并的正确性。受均值漂移（Mean－shift）算法[9]的启发，本节将提出一种基于 K－NN 的均值漂移（KNN－based mean shift，KMS）算法，利用 K 近邻（K－NN）思想计算当前样本点的均值漂移向量，使向量沿着密度增大的方向移动直到到达密度峰值处，自适应地确定数据集的真实类别个数 c，克服传统均值漂移算法易受带宽 h 影响的缺点。当数据集分布不平衡时，固定带宽会影响聚类效果，KMS 算法通过 K 近邻思想能够得到灵活的“带宽 h”。具体的，使用一定数量的 K 个最近邻样本点对均值漂移向量进行直接迭代，这样不仅能够保证参与每次迭代的样本数量，而且可以很好适应迭代范围。样本的均值漂移向量 $M_h(x)$ 被定义如下：

$$M_h(x) = \frac{1}{K}\sum_{k=1}^{K}(x_k - x), x_k \in S_h(x) \tag{9}$$

其中：$S_h(x)$ 和 K 分别表示样本 x 的集合和 K 近邻数量。在 KMS 中，仅改进 $M_h(x)$ 以适应样本迭代范围，提高系统的鲁棒性，其他步骤与均值漂移算法相似。为了减小计算负担，这里仅取从 ICCM 算法中获得的 N 个子簇类中心作为均值漂移向量迭代的初始点，由 KMS 的聚类结果可得到数据集的真实类别个数 c。

参数 K 的选取原则：在实际应用中，N 个子簇的类中心被用作迭代均值漂移向量的初始点，因此 KMS 算法对 K 值具有较强的鲁棒性。为了减少迭代次数，推荐 $K = (n/N) \cdot (1 \pm 10\%)$ 作为默认值，其中 n 为不平衡数据集中包含的样本数量。

2.3　子数据集合并

本节将提出一种密度合并规则（Density－ based merging rule，DMR），根据复合类和其所包含的两个子簇之间的密度关系对划分的子簇及部分复合类进行合并，直至得到与原始数据集真实类别个数相同的 c 个单类。复合类被认为是不同子簇之间的不确定类别，样本被划分到复合类意味着样本可能属于复合类所包含的子簇中的任何一个。如果 ICCM 将同属于一个类别的样本划分给了不同的子簇和复合类，表明这些子簇的密度可能非常相似；复合类中的样本通常分布在类别的相对中心，所以复合类的密度应大于或者介于复合类中所包含子簇的密度之间；如果复合类的密度小于其所包含的两个子簇的密度，则意味着这两个子簇属于不同的类别。综上，复合类和其包含的两个子簇之间存在以下三种密度关系：

$$C_1: \rho_{\omega_k}(\rho_{\omega_t}) \leqslant \rho_{\Lambda_i}$$
$$C_2: \rho_{\omega_k}(\rho_{\omega_t}) < \rho_{\Lambda_i} < \rho_{\omega_t}(\rho_{\omega_k})$$
$$C_3: \rho_{\Lambda_i} < \rho_{\omega_k}(\rho_{\omega_t})$$

满足上述 C_1 和 C_2 关系的复合类和子簇能够进行合并，并且满足 C_1 关系的可优先合并。不难发现，子簇合并过程具有传递性，即如果有两个已部分合并的子数据集都与一个未合并的子簇满足密度合并关系，则这两个子数据集也应进行合并。目前已有许多密度计算方法得到了广泛应用，本节提供一种简单的方法对不同类簇 A_i（子簇或复合类）进行密度估计，A_i 的密度被定义如下：

$$\rho_{A_i} = \left[\frac{1}{n_i}\frac{1}{K}\sum_{i=1}^{n_i}\sum_{j=1}^{K} d_{ij}\right]^{-1} \tag{10}$$

其中：ρ_{A_i} 为类簇 A_j 的密度，n_i 表示 A_i 中样本的数量，d_{ij} 表示 A_i 中样本 x_i 与数据集中样本 x_i 的第 j 个近邻之间的欧式距离。这里利用 K 近邻思想来消除噪声带来的影响，默认值 $K = 10$。根据上述合并规则，能够将多划分获得的 N 个子簇以及部分复合类进行合并。

为了表示方便，定义 $\omega_{k,t} \triangleq \omega_k \cup \omega_t$，下面通过一个简单的例子来说明根据密度合并的过程。考虑一个真实类别 $c = 2$，多划分后子簇个数 $N = 4$ 的问题，各个子簇和复合类的密度分别为 $\rho_{\omega_1} = 0.56$，$\rho_{\omega_2} = 0.71$，$\rho_{\omega_3} = 0.47$，$\rho_{\omega_4} = 0.34$，$\rho_{\omega_{1,2}} = 0.24$，$\rho_{\omega_{1,3}} = 0.67$，$\rho_{\omega_{1,4}} = 0.21$ 和 $\rho_{\omega_{3,4}} = 0.42$。此例中各个子簇和部分复合类的具体合并过程如下：1）根据上述 C_1，可得 $\Gamma_1 = \omega_1 \cup \omega_3 \cup \omega_{1,3}$；2）根据上述 C_2，可得 $\Gamma_2 = \omega_3 \cup \omega_4 \cup \omega_{3,4}$；3）由传递性可得，$\omega^1 = \Gamma_1 \cup \Gamma_2 \cup \omega_{1,4}$。因此，通过密度合并最终得到的新的类别结果如下：

$$\omega^1 = \omega_1 \cup \omega_3 \cup \omega_4 \cup \omega_{1,3} \cup \omega_{3,4} \cup \omega_{1,4}; \omega^2 = \omega_2$$

其中：Γ_i 表示已合并的过渡簇（子数据集），ω^i 表示样本最终所属的真实类别。在获得需要的 c 个单类之后，可能仍会存在一些复合类（比如 $\omega_{1,2}$）尚未合并，这些未合并复合类中的样本通常处于不同类别的重叠区域（例如 ω^1 和 ω^2），因此需要采用更加谨慎的策略对这些样本进行划分。

2.4 剩余数据划分

本节提出一种剩余样本再划分规则（Re－partition rule，RPL）对未合并复合类中的样本进行再次划分以得到最终的聚类结果。未合并复合类中存在的少数样本经过再划分后仍很难被划分给某个特定类别，则这些样本将保留成为一个新的复合类，以降低误划分风险。RPL 的关键在于，认为样本处在不同类别重叠区域的条件为该样本到不同类别中与其最近的 K 个近邻的平均距离无明显差异。对于未合并复合类中的样本 x_i，首先将获得 x_i 在与此复合类相关的两个单类中的 K 近邻，定义样本 x_i 到最终类别 ω^k 的距离为 x_i 到 K 个最近邻的平均距离，用公式表示如下：

$$d(x_i,\omega^k)=\frac{1}{K}\sum_{j=1}^{K}d(x_i,x_j^{\omega^k}) \tag{11}$$

其中：$d(x_i,x_j^{\omega^k})$ 表示样本 x_i 与其在类别 ω^k 中第 j 个近邻的距离，采用 K 近邻目的是为了消除噪声的影响，K 太大会增加计算成本，故默认值 $K=5$。为了谨慎划分这些不确定样本，若样本 x_i 分别属于 ω^k 和 ω^t 的概率之差小于参数 χ，则 x_i 将会被划分到新的复合类 $\omega^{k,t}$ 中（$\omega^{k,t}\triangleq\omega^k\cup\omega^t$），复合类 $\omega^{k,t}$ 被定义如下：

$$\omega^{k,t}=\{x_i\mid P_i^{\omega^k}-P_i^{\omega^t}\leqslant\chi\} \tag{12}$$

其中：$P_i^{\omega^k}$ 为样本 x_i 属于类别 ω^k 的概率，表示为：

$$P_i^{\omega^k}=\frac{d^{-1}(x_i,\omega^k)}{d^{-1}(x_i,\omega^k)+d^{-1}(x_i,\omega^t)} \tag{13}$$

再划分参数 χ 的选取原则：$\chi\in[0,1]$ 是一个可调的阈值参数，其值大小会影响最终复合类中的样本数量。χ 越小，最终复合类中的不确定样本越少，这将会增加不确定样本误划分的风险；而随着 χ 增大，更多不确定样本被划入最终的复合类中，这将导致不精确率增高。χ 应根据可接受的不精确程度进行调节。

为了更加清晰表达 MP 算法的基本流程和主要内容，图 1 展示了多划分（MP）聚类算法的流程框图。

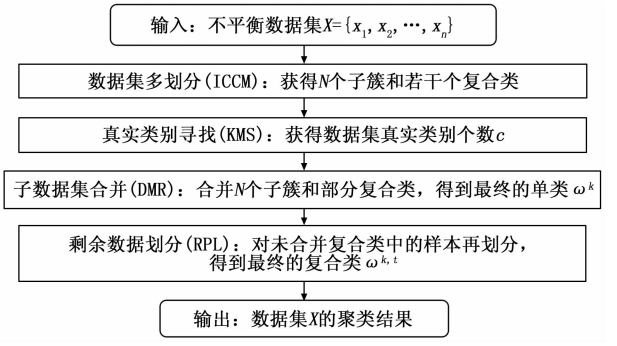

图 1 不平衡数据多划分（MP）聚类算法流程框图

3 实验分析

本文利用 UCI 数据库[10]中五组真实数据集（即 Wine、Bupa、Balancescale、Aggregation 和 WBC）对不平衡数据多划分（MP）聚类算法的性能进行测试和评价，通过与 FCM、GK 和 CCM 三种聚类算法对比验证 MP 算法的性能。Balancescale 数据集共有 3 个类别，其中名为 Left 和 Balanced 的两个类别（分别包含 288 和 49 个样本）满足不平衡数据分布，选择这两类来评估算法性能。同样在 Aggregation 数据集中共有 7 个类别，选择其中分别包含 102、34 和 34 个样本的 3 个类别（即第三、五、七类）来验证算法的有效性。除以上两组数据集外，其余数据集均采用所有类别进行实验，实验所用数据集的详细信息如表 1 所示。

表 1 实验所用五组 UCI 数据集详细信息

数据集名称	类别数	属性数	样本数	各类别样本数
Wine	3	13	178	5 971 48
Bupa	2	6	345	1 452 00
Balancescale	2	4	337	2 884 9
Aggregation	3	2	170	1 023 434
WBC	2	9	683	4 442 39

MP 算法中，参数 N（$N>c$）为 ICCM 生成的子簇个数，可根据用户的具体需求设置，在本实验中，默认 $N=6$，复合类阈值取 $\varepsilon=0.2$，再划分参数 $\chi=0.2$。FCM 和 GK 算法中模糊指数 $m=2$；CCM 中距离权重 γ 分别设置为 $\gamma=0.5,1.0,2.0$，且 $t_c=2$。实验采用错误率 $R_e=\frac{N_e}{N_T}$ 和不精确率 $R_{i_j}=\frac{N_{i_j}}{N_T}$ 两个评价指标对算法性能进行评价，其中 N_e 表示被错误划分的样本数，N_{i_j} 表示被划分到复合类的样本数（j 表示构成复合类的最大单类个数，本实验中 $j=2$），N_T 表示不平衡数据集中的样本总数。除利用以上两个指标来评价算法有效性外，实验还对各算法的运行时间 T（单位：s）进行了比较，参数 T 在一定程度能反映算法的计算复杂度。为了便于标注，记复合类 $\omega_{i,\cdots,j}\triangleq\omega_i\cup\cdots\omega_j$。实验所用平台为惠普笔记本电脑，软硬件参数如下：CPU 为英特尔 i5－8300 H，软件版本为 Matlab 2016 b。四种聚类算法对五组 UCI 数据集进行聚类的结果分别如表 2 所示。

表 2 五组 UCI 不平衡数据集聚类结果 %

		FCM	GK	CCM			MP
				$\gamma=0.5$	$\gamma=1.0$	$\gamma=1.5$	$\varepsilon=0.2$，$\chi=0.2$
Wine	R_e	31.46	39.33	27.53	26.40	24.72	24.49
	R_{i_2}	/	/	5.62	10.67	17.98	0
	T	0.017 1	0.061 5	0.457 7	0.488 4	0.439 7	0.053 2
Bupa	R_e	47.54	48.99	48.12	46.96	46.96	42.03
	R_{i_2}	/	/	2.90	4.93	8.12	0
	T	0.013 4	0.014 3	0.509 0	0.494 3	0.470 2	0.091 2
Balancescale	R_e	47.18	38.58	51.34	47.18	46.88	13.77
	R_{i_2}	/	/	0.30	1.19	1.48	12.82
	T	0.025 2	0.037 5	0.512 3	0.459 3	0.508 5	0.131 8
Aggregation	R_e	48.82	30.00	46.47	45.29	44.71	0
	R_{i_2}	/	/	4.71	7.06	9.41	0
	T	0.005 3	0.008 7	0.668 2	0.677 8	0.712 5	0.048 3
WBC	R_e	4.39	6.73	3.07	2.78	2.34	2.05
	R_{i_2}	/	/	2.64	3.07	4.39	2.64
	T	0.018 7	0.047 5	0.578 5	0.594 6	0.569 8	0.172 3

从表 2 可以看出，MP 算法对 UCI 中五组不平衡数据集的聚类结果均优于其他三种算法，错误率最低。MP 算法中引入的复合类能从一定程度降低样本误划分的风险，合理表征处于重叠区域的不确定样本，降低错误率。从程序运行时间 T 上，FCM 和 GK 算法由于没有复合类，故运行时间最快；CCM 和 MP 算法由于引入了复合类，程序运行时间 T 会比前两种算法时间长，但通过实验数据可看出 MP 运行时间远小于 CCM，说明 MP 计算复杂度远小于 CCM，算法运行效率比较高。

4　MP 算法在建筑能耗异常检测中的应用

4.1　MP 算法能耗异常检测模型

本节将利用提出的 MP 聚类算法原理及内容构建能耗异常检测模型。MP 聚类算法分为四个子步骤：数据集多划分、真实类别寻找、子数据集合并以及剩余样本划分，现将这些步骤运用在建筑能耗异常检测中，构建如图 2 所示的 MP 算法能耗异常检测模型。首先将预处理后的能耗数据集进行多划分，得到 N 个能耗子数据集（$N > c$）和若干个复合类；接着寻找数据集真实类别个数 c，即正常能耗类别和异常能耗类别个数之和；然后对多划分得到的能耗子数据集和部分复合类进行合并；最后，对未合并复合类中的剩余能耗数据进行再划分，得到能耗数据集的类别划分结果，即最终的异常检测结果。从可行性的角度分析，由于能耗数据的分布符合聚类分布的特点，即距离类中心越近的地方样本点分布越密集，这就保证了 MP 算法在第三步密度合并时能够有效利用子数据集和复合类的密度进行子数据集合并，同时保证了 MP 算法能耗异常检测模型的可行性。

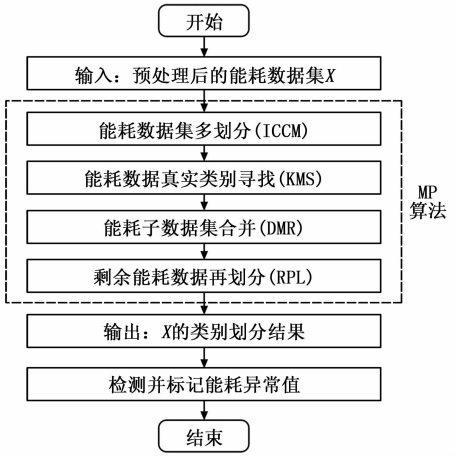

图 2　MP 算法能耗异常检测模型

4.2　能耗异常检测实验结果与分析

本文使用的能耗数据来源于对西安市某大型商场建筑的逐日分项用电监测，通过对该商场能耗监管系统进行调研，采集并记录了 2018 年 3 月 5 日至 2019 年 2 月 28 日的分项日用电能耗情况（共 360 组样本），包括空调、照明、动力、特殊设备用电量以及总用电量。选取该商场 18 年第二季度（6～8 月，共计 92 天）日分项能耗（空调用电和照明用电）数据进行能耗数据异常检测实验，空调和照明用电能耗数据如图 3 所示。由于直接来源于现实生活中的数据经常会存在不完整、不一致等现象，这些对数据挖掘效果都会产生很大影响，因此在进行能耗异常检测实验前，需要对实验所用能耗数据进行预处理，确保数据的完整性和一致性。MP 算法中的两个阈值分别设置为 $\varepsilon = 0.2$，$\chi = 0.2$。为了与 MP 算法得到的结果进行比较，实验还采用 FCM 和 GK 算法对相同用电能耗数据进行处理，得到了相应的聚类结果。

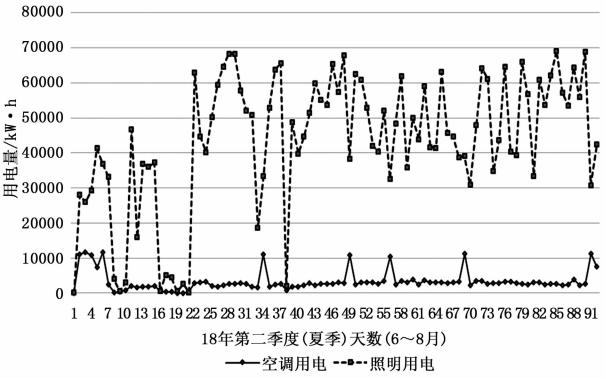

图 3　西安某大型商场建筑 18 年第二季度（6～8 月）空调与照明用电能耗数据

利用 MP 算法对上述用电能耗数据进行聚类，实验取 $N = 8$（即对该能耗数据集多划分得到的子数据集个数为 8），图 4（a）和（b）分别展示了第二季度原始用电能耗数据以及 MP 算法对空调和照明用电能耗数据聚类得到的结果。

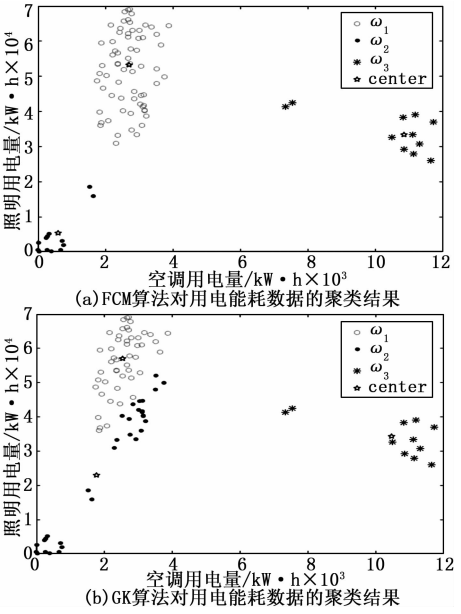

图 4　MP 算法建筑能耗异常检测实验结果

在实验过程中可得到能耗数据的真实类别个数 $c = 3$，从图 4（b）可以看出，MP 算法将该季度用电能耗数据最终被划分为三类：ω^1、ω^2 和 ω^3（等于能耗数据集真实类别个数），其中类别 ω^1 从样本数量上看属于多数类，能够判断其

属于正常能耗类别，其中的能耗数据在范围上相对比较稳定（即空调用电和照明用电量都在一定范围内）；而类别 ω^2 和类别 ω^3 中所含能耗样本的数量很少（属于少数类），且分布上明显偏离类别 ω^1，故将这两个类别所包含的能耗数据认定为异常能耗数据，其中类别 ω^2 中的能耗数据在空调用电量上表现出异常（远超出正常空调用电量水平），类别 ω^3 中的能耗数据在空调用电和照明用电上均表现出异常（均远小于正常用电量水平）。从最终的聚类结果来看，能耗数据集除了被划分为以上三个类别外，还得到了两个复合类（$\omega^{1,2}$ 和 $\omega^{1,3}$），它们所包含的能耗样本虽然不能认定为异常能耗数据，但介于正常与异常能耗之间，需要对这种不确定能耗数据采取更加谨慎的态度，以免导致数据误判的风险。图 5 为根据 MP 算法异常检测结果对异常能耗数据标记之后得到的用电能耗数据折线图（其中三角形和菱形标记分别表示检测出的空调用电异常能耗数据和不确定数据，圆形和正方形标记分别表示检测出的照明用电异常能耗数据和不确定数据），通过 MP 算法能耗异常检测模型，能够有效检测得到建筑能耗数据中的异常值，为建筑能耗监管系统的管理和运行提供必要的帮助，有利于管理人员及时发现并解决建筑中可能存在的问题与故障。

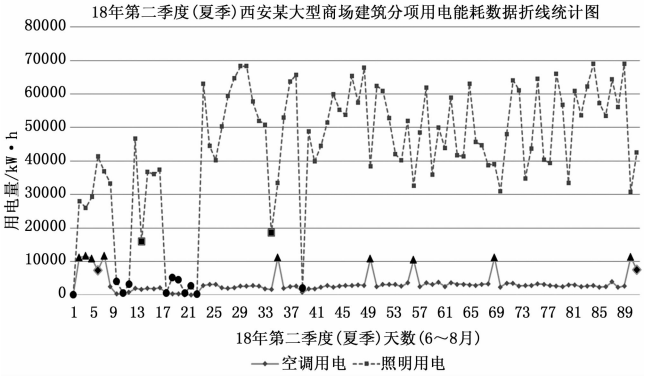

图 5 空调/照明用电能耗异常检测数据图

为了与 MP 算法进行对比，采用 FCM 和 GK 两种算法对相同的能耗数据进行处理，图 6（a）和（b）分别为 FCM 和 GK 算法对用电能耗数据聚类的结果。由图 6（a）可看出，FCM 将用电能耗数据划分为三类，但因其初始聚类时并没有类别先验信息，故在对数据进行聚类时首先需要获得数据的真实类别；从结果来看，FCM 将 MP 算法中划分到复合类中的不确定数据强行划分到异常数据类别中，这样可能会增加数据误判为异常值的风险。从图 6（b）可以看到，GK 将用电能耗数据划分为三类，但同样需要在聚类前对能耗数据的真实类别个数进行判断，最终的聚类结果显示，GK 将原本属于正常能耗类别 ω_1 中的部分数据错误划分到能耗异常类别 ω_2 中，导致了部分正常能耗数据被误判为异常能耗，与 MP 聚类算法对比错误率明显增加。

5 结束语

由于异常能耗值在能耗数据中仅占很小的部分，能耗

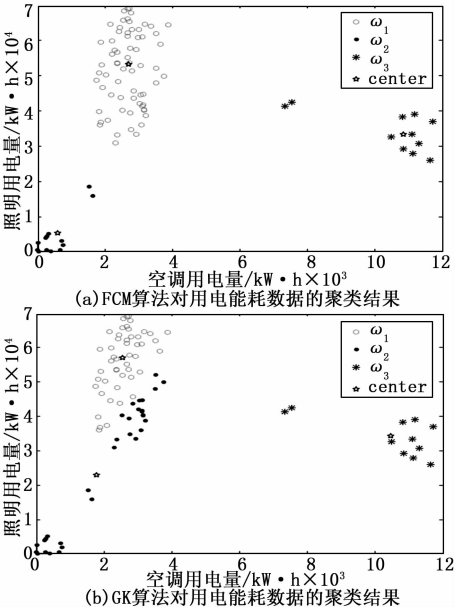

图 6 FCM 和 GK 算法对第二季度用电能耗数据的聚类结果

异常检测可以看作对不平衡数据的聚类，为了对不平衡数据进行有效聚类，避免样本“均匀效应”，本文提出了一种基于 D－S 证据理论的不平衡数据多划分（MP）聚类算法，并将其应用到建筑能耗异常检测中，构建了 MP 算法能耗异常检测模型对建筑能耗中的异常值进行检测。首先对预处理后的能耗数据集进行多划分，得到 N 个能耗子数据集和若干复合类；确定该能耗数据集的真实类别个数；然后对多划分得到的能耗子数据集和部分复合类进行合并；最后对未合并复合类中的剩余能耗数据进行再划分，得到能耗数据集的类别划分结果，即最终的异常检测结果。经 UCI 数据集验证，MP 算法具有良好的聚类效果，通过对某商场建筑用电能耗数据进行能耗异常检测，验证了 MP 算法能耗异常检测模型的有效性。由能耗异常检测实验的结果可以看出，MP 算法对于处在正常和异常能耗数据之间的不确定数据没有强行划分，但同时给算法带来了一定的不精确率，如何谨慎地对这些数据进行划分，从而确定这些能耗数据是否为异常值，是下一步需要深入研究的问题。

参考文献：

[1] 卿晓霞，肖 丹，王 波，等. 能耗实时监测的数据挖掘方法[J]. 重庆大学学报：自然科学版，2012，35（7）：133－137.

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

[3] Bezdek J C，Ehrlich R，Full W. FCM：The fuzzy c－means clustering algorithm [J]. Computers & Geosciences，1984，10（2）：191－203.

[4] Liang J Y，Bai L，Dang C Y，et al. The K－Means－Type Algorithms Versus Imbalanced Data Distributions [J]. IEEE Transactions on Fuzzy Systems，2012，20（4）：728－245.

[5] Chaomurilige，Jian Y，Yang M S. Analysis of Parameter Selec-

tion for Gustafson－Kessel Fuzzy Clustering Using Jacobian Matrix [J]. IEEE Transactions on Fuzzy Systems, 2015, 23 (6).

[6] Dempster A. Upper and Lower Probabilities Induced by a Multi－valued Mapping [J]. The Annals of Mathematical Statistics, 1967, 38 (2): 325－339.

[7] Shafer G. A Mathematical Theory of Evidence [M]. Princeton Univ. Press, 1976.

[8] Liu Z G, Pan Q, Dezert J, et al. Credal c－means clustering method based on belief functions [J]. Knowledge－Based System, 2015, 74: 119－132.

[9] Cheng Y Z. Mean Shift, Mode Seeking, and Clustering [J]. IEEE Transactions on Pattern. Analysis and Machine Intelligence, 1995, 17 (8): 790－799.

[10] Frank A, Asuncion A. UCI Machine Learning Repository [EB/OL]. University of California, School of Information and Computer Science, Irvine, CA, USA, 2010. http://archive. ics. uci. edu/ml.

(上接第 238 页)

图 6　卫星轨道及观测站点分布

值，其中取卫星轨道误差为 10 m、测频噪声为 10 Hz、测高误差为 3 m。在迭代计算过程中取粒子群规模为 N_s 100，迭代次数上限 T 为 100，惯性权重 ω 取 0.9，学习因子 c_1、c_2 都取 1.5，自适应网格步长调整策略中 t 取 10。

通过仿真得到不同时刻各站点计算的定位误差如表 1 所示。

表 1　单星仿真定位误差　　(m)

	时间 1	时间 2	时间 3
站点 1	92.99	91.96	168.43
站点 2	321.01	320.02	477.04
站点 3	268.45	268.43	445.65
站点 4	63.35	64.32	260.31
站点 5	67.40	68.03	115.69

可得到在仿真环境下不同时刻定位精度存在一定差异，并且远离星下点的观测点定位误差较差，总体上单星定位精度在百米量级。

4　结束语

基于伪距和多普勒观测量可实现单星条件下的导航定位，在本文设计的定位方式下，单星定位精度与伪距和多普勒观测误差相关，误差分布呈现出对不同误差项的敏感差异。通过仿真初步验证表明基于伪距和多普勒测量能够达到百米量级的定位精度。未来将进一步探讨单星导航定位的应用前景和工程实现中的相关技术。

参考文献：

[1] Michael G F, Stephen G C. A New Pseudolite Battlefield Navigation System [A]. IEEE Conference on Position Location and Navigation Symposium [C]. 1998: 208－217.

[2] Oktay H, Stepaniak M. Airborne Pseudolites in a Global Positioning System Degraded Environment [A]. 5th International Conference on Recent Advances in Space Technologies [C]. 2011.

[3] Angelo Trunzo, Paul Benshoof. The UHARS Non－GPS Based Positioning System [A]. 24th International Technology Meeting of the Satellite Division of the U. S. Inst. of Navigation [C]. 2011: 3582－3586.

[4] Se Phil Song, Heon Ho Choi, Young－Baek Kim, et al. Verification of GPS Aided Error Compensation Method for eLoran using Raw TOA Measurements [A]. 11th International Conference on Control, Automation and Systems [C]. 2011: 1620－1624.

[5] Gregory W J, Peter F S, Richard J H, et al. An Evaluation of eLoran as a Backup to GPS [A]. Conference on Technologies for Homeland Security [C]. 2007: 95－100.

[6] Hang Yan, Jinkun Cao, Lei Chen. Study on Location Accuracy of Dual－Satellite Geolocation System [A]. 10th International Conference on Signal Processing [C]. 2010: 107－110.

[7] Niemeyer P H. Single Satellite Geolocation [C]. Satellite Communication and Broadcasting, 1987: 69－77.

[8] 郭福成. 基于 WGS－84 地球模型的单星测向定位方法 [J]. 宇航学报, 2011, 32 (5): 1179－1183.

[9] 杨　斌，张　敏，李立萍. 基于 WGS－84 模型的单星 DOA 定位算法 [J]. 航天电子对抗, 2009, 25 (4): 24－26.

[10] 严　航，姚山峰. 低轨单星测频定位技术及其精度分析 [J]. 计算机工程, 2012, 38 (18): 6－10.

[11] 李献斌. 单星测频无源定位技术研究 [D]. 长沙：国防科技大学, 2009.

[12] 曹东波，张　敏，姜文利. 单星多普勒变化率无源定位精度分析 [J]. 航天电子对抗, 2010, 26 (4): 1－4, 64.

[13] 邓　琳，李广侠，田世伟，等. 基于 LEO 增强的 COMPASS 导航系统抗干扰能力研究 [J]. 军事通信技术, 2012, 33 (2): 65－69, 80.

[14] Mark S. Asher, Stephen J. Stafford, Robert J. Bamberger, et al. Radio Navigation Alternatives for US Army Ground Forces in GPS Denied Environments [A]. International Technical Meeting of The Institute of Navigation [C]. 2011: 508－532.

[15] Ashton C, Bruce A S, Colledge G, et al. The Search for MH370 [J]. The Journal of Navigation, 6: 78－81.