基于区间数单簇聚类-单分类器的异常检测

孙 强[1] 魏 伟[2] 侯培鑫[1] 岳继光[1]
(同济大学电子与信息工程学院 上海 201804)[1]
(埃尔兰根-纽伦堡大学电子工程所 埃尔兰根 91058DE)[2]

摘 要 异常检测是系统运行维护的重要工作。在系统运行过程中可获得大量正常的运行数据,但异常数据的获取成本较高,因此可引入单分类器的思想来处理异常检测问题。测量不确定性、环境噪声、存储设备等导致监测数据可能存在不确定性。利用区间数描述不确定的监测数据,提出区间数样本的核可能性1-均值单簇聚类-单分类器异常检测算法。分别考虑聚类中心位于输入空间与特征空间两种情况,并考虑区间数样本具有的区间宽度不均衡性,提出区间细分检测策略。结合人工数据集与UCI数据集给出的算例验证了所提算法的有效性,其与现有SVM-OCC相比具有更高性能。

关键词 区间数样本,单簇聚类,单分类器,区间细分,异常检测
中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.06.032

Anomaly Detection Based on Interval One Cluster and Classification

SUN Qiang[1] WEI Wei[2] HOU Pei-xin[1] YUE Ji-guang[1]
(College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China)[1]
(Institute of Electronics Engineering, Erlangen-Nuremberg University, Erlangen 91058DE, Germany)[2]

Abstract Anomaly detection is crucial in system maintenance. During operating process, normal operation data is easy to obtain, while anomal data usually take high cost to obtain. Therefore, one classifier could be utilized to solve the anmaly detection problem. Due to measurement uncertain, environment noise and storage problem etc., uncertainness could be a characteristic of the montoring data. This paper utilized interval number to describe the uncertainess in the monitoring data, and raised an anomaly detection algorithm based on kernelized possibilistic 1-means clustering and 1-classifier for interval samples. The clustering center was considered both in the input space and the feature space. The interval width of the samples could be unbalanced, therefore, an interval splitting stragy was also proposed. Finally, illustrative numberic examples were given in utilizing artificial dataset and UCI machine learning repository. The effectiveness of the proposed algorithm is verified, and improvement is made by comparing with the existing SVM-OCC algorithms.

Keywords Interval samples, One cluster, One classifier, Interval splitting, Anomaly detection

1 引言

在实际系统中容易通过监测手段获得大量正常运行的数据样本,但获取异常样本的成本一般较高,难以获取大量有效的样本用于分类学习。单分类器是异常检测的重要工具,可通过正常样本的学习建立正常数据描述域,以此为基础检测并分类异常数据样本。监测数据中可能存在不确定性,来源于测量不确定性、背景噪声等,同时还可能由于存储设备的原因造成缺失。区间数[1]是描述数据不确定性的有效工具,既可包含真值,又能有效描述测量值的不准确性。本文考虑区间数样本及其不均衡性,提出一种基于区间数的核可能性C均值单簇聚类-单分类器的异常检测算法。

聚类是将有限未标记类别的样本按照某种准则聚集为若干类簇。模糊C均值(Fuzzy C-Means,FCM)聚类算法是最经典的基于目标函数的模糊聚类算法,可能性C均值(Possibilistic C-Means,PCM)聚类算法放宽FCM中隶属度之和为1的约束条件,使隶属度更加准确地描述样本对类的隶属程度,提高了对噪声的鲁棒性。当聚类数为1时,PCM具有全局最优性[2]。核化聚类算法旨在处理复杂数据形状的样本聚类问题,将样本非线性映射到高维空间以增加线性可分性,同时利用核技巧处理高维空间的内积运算。目前针对区间数样本的核聚类问题的研究较少。任世锦等[3]研究了基于遗传算

到稿日期:2016-11-28 返修日期:2017-02-13 本文受上海市科委科研项目(11JC1413000)资助。
孙 强(1986—),男,博士生,主要研究方向为系统健康监测与管理,E-mail:10qsun@tongji.edu.cn;魏 伟(1988—),男,硕士,主要研究方向为机器视觉;侯培鑫(1991—),男,博士生,主要研究方向为统计学习预测;岳继光(1961—),男,博士,教授,主要研究方向为先进过程控制、计算机控制。

法优化的多项式核化的区间数 FCM 聚类算法，指出区间数空间中的核函数选择的条件与困难；Pimentel 等[4]研究了 Gauss 核化的区间数 FCM 聚类算法，并进一步比较了聚类中心位于输入空间与输出空间时算法的性能[5]。

基于单簇聚类的异常检测实质上是构建正常样本描述的单分类器(One-Class Classifier，OCC)，属于半监督学习。通过单簇聚类处理，可获取样本对单簇的隶属程度输出；通过设置异常决策阈值，构成正常样本描述分界超平面。根据统计学习理论结构风险最小化的思想[6]，在训练中设置松弛变量，这样尽管可能会产生部分训练误差，但可获得更好的泛化能力。据此设置决策阈值的容忍度，即样本检测的拒绝比，以期获得更好的异常检测性能。

SVM-OCC 是目前解决单分类问题的典型方法，最早由 Tax 与 Duin[7]于 1999 年提出，基本思想是寻找包含训练集的最小超球，属于非监督学习。Schoelkopf 等[8]提出构造具有最大间隔的超平面，以分割目标集与测试集。Campbell 与 Bennett[9]提出了 C-BSVM-OCC 模型，其对偶形式可归结为线性规划问题，故 C-BSVM-OCC 模型可通过简单的对偶形式处理原本复杂的单分类问题。Utkin 与 Chekh[10]提出了一种利用 Triangular 核近似 Gauss 核的改进 C-BSVM-OCC，并将其用于区间数样本的异常检测(记为 Int-C-B)，将极小化分类风险转化为约束风险上下界的 minmax 问题。尽管该改进模型能取得较高的检测率，但算法复杂且 Triangular 核适用性有限。

本文结合核技巧，提出聚类中心位于输入空间与特征空间的两种区间数可能性 1-均值单簇聚类算法(IkP1M 系列算法)；提出结合隶属度判断的异常检测方法(IkP1M-OCC)，并考虑训练集与测试集样本的区间不均衡，提出两种检测策略，同时利用基于 GridSearch 的 K-折交叉验证选择算法参数。以人工香蕉数据集、区间化的 UCI 真实数据集为算例，验证了 IkP1M-OCC 的有效性；同时结合算例，说明了 IkP1M-OCC 与典型的 SVM-OCC 算法相比适用性更佳，且具有一定的检测性能优势。

2 理论基础

本节给出区间数样本的定义及其相异度衡量，回顾可能性 C 均值(PCM)聚类算法，为全文奠定理论基础。

2.1 区间数样本及其相异度

1962 年 Moore[1]首次完整阐述了区间分析理论，给出了区间数及其运算的基本概念。本文利用区间数表示具有不确定性的监测数据，并基于这些区间数构建区间数样本空间。

定义 1(区间数) 设$\mathbb{R}$为实数集，$\underline{a},\overline{a}\in\mathbb{R}$，$\underline{a}\leqslant\overline{a}$，记

$$[\![a]\!]=[\underline{a},\overline{a}] \tag{1}$$

称$[\![a]\!]$为(实)区间数，$\underline{a},\overline{a}$ 分别为区间数$[\![a]\!]$的下限和上限。若$\underline{a}=\overline{a}$，则$[\![a]\!]$退化为一个精确的实数。记一维区间数空间为$\mathbb{IR}$，一维实区间数可用实数轴上的区间段表示。

设待分析样本为区间数样本空间$\mathbb{IR}^p$中的样本集$\{[\![\boldsymbol{x}_1]\!],\cdots,[\![\boldsymbol{x}_n]\!]\}$，每个样本向量含 p 个区间特征，即：

$$[\![\boldsymbol{x}_k]\!]=([\![x_{k1}]\!],\cdots,[\![x_{kp}]\!]);[\![x_{kj}]\!]=[\underline{x}_{kj},\overline{x}_{kj}]$$

$$j=1,\cdots,p;k=1,\cdots,n$$

设$[\![\boldsymbol{x}_i]\!],[\![\boldsymbol{x}_j]\!]\subset\mathbb{IR}^p$，$[\![\boldsymbol{x}_i]\!]=([\![x_{i1}]\!],\cdots,[\![x_{ip}]\!])$，$[\![\boldsymbol{x}_j]\!]=([\![x_{j1}]\!],\cdots,[\![x_{jp}]\!])$，$[\![x_{lo}]\!]=[\underline{x}_{lo},\overline{x}_{lo}]$，$m_{lo}=(\underline{x}_{lo}+\overline{x}_{lo})/2$，$r_{lo}=(\overline{x}_{lo}-\underline{x}_{lo})/2$，$l=i,j;o=1,\cdots,p$。

定义 2 $[\![\boldsymbol{x}_i]\!]$，$[\![\boldsymbol{x}_j]\!]$的 Hausdorff 距离[11]定义为：

$$d_{Hij}=\sqrt{\sum_{o=1}^{p}[(m_{io}-m_{jo})^2+(r_{io}-r_{jo})^2+2|m_{io}-m_{jo}||r_{io}-r_{jo}|]} \tag{2}$$

容易证明，d_{Hij} 满足对称性、非负性、自反性及三角不等式等 4 个条件，且计算结果为确定数即退化的区间数。本文采用区间数样本的 Hausdorff 距离衡量相异度。Gauss 核诱导的区间数样本$[\![\boldsymbol{x}_i]\!]$，$[\![\boldsymbol{x}_j]\!]$的 Hausdorff 距离为：

$$d_{\Phi ij}=\sqrt{2-2K_{intG}([\![\boldsymbol{x}_i]\!],[\![\boldsymbol{x}_j]\!];\sigma)} \tag{3}$$

其中，$K_{intG}([\![\boldsymbol{x}_i]\!],[\![\boldsymbol{x}_j]\!];\sigma)=\exp(-d_{ij}^2/2\sigma^2)$为区间数空间中的 Gauss 核。

2.2 PCM 聚类算法

Krishnapuram 等[12-13]于 1993 年首次提出 PCM 聚类算法，放宽了 FCM 算法中隶属度之和为 1 的约束条件，同时为了避免产生隶属度均为 0 的平凡解，在目标函数中添加了正则项$\sum_{i=1}^{c}\eta_i\sum_{k=1}^{n}(1-u_{ik})^m$，其中惩罚因子 $\eta_i>0$。PCM 算法等价为如下二次规划问题：

$$\begin{aligned}&\min J_{PCM}(U,\{\boldsymbol{v}_i\}_c)=\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}^m d^2(\boldsymbol{x}_k,\boldsymbol{v}_i)+\\&\qquad\sum_{i=1}^{c}\eta_i\sum_{k=1}^{n}(1-u_{ik})^m\\&\text{s.t. } 0\leqslant u_{ik}\leqslant 1,0\leqslant\sum_{k=1}^{n}u_{ik}\leqslant n;\\&i=1,\cdots,c;k=1,\cdots,n\end{aligned} \tag{4}$$

聚类中心、隶属度分别按照式(5)、式(6)更新：

$$\boldsymbol{v}_i=\frac{\sum_{k=1}^{n}u_{ik}^m\boldsymbol{x}_k}{\sum_{k=1}^{n}u_{ik}^m} \tag{5}$$

$$u_{ik}=\frac{1}{1+[d^2(\boldsymbol{x}_k,\boldsymbol{v}_i)/\eta_i]^{1/m-1}} \tag{6}$$

Krishnapuram 指出 PCM 算法与 FCM 算法具有本质的不同：FCM 算法是基于划分的聚类算法；而 PCM 算法因为放宽了隶属度约束，样本可自发地被吸引至所属聚簇，因此属于基于模式搜索的聚类算法。但 PCM 算法仍有明显的缺陷，根据式(6)，隶属度更新与类间距离无关，实际迭代过程极易陷入局部极值。

考虑 $c=1$ 的单簇聚类，形成的 P1M 算法则不存在上述问题，又因其对噪声鲁棒，所以适用于离群点检测。陈斌[14]证明了 P1M 算法可保证目标函数对隶属度及聚类中心 Hessian 矩阵的正定性，即 P1M 算法可获得全局最优解。

3 区间数样本的核可能性 1-均值单簇聚类算法

本节将 P1M 算法核化并推广至区间数域，提出区间数样本的核可能性 1-均值单簇聚类算法(Kernelized Possibilistic 1-Means Clustering Algorithm for Interval Samples，IkP1M)。分别考虑聚类中心在输入空间与特征空间两种情况，结合算法目标函数的两种形式，给出具有 4 种形式的 IkP1M 系列算

法，并分析算法的复杂度与关键参数的影响。

3.1 IkP1M 系列算法的理论推导

核化聚类算法是处理非超球复杂形状数据样本的途径之一。本节讨论区间数样本核可能性 C 均值单簇聚类。考虑聚类中心在输入空间与特征空间两种情况，结合算法目标函数的两种表达形式，本节提出含有 4 种形式的 IkP1M 系列算法。

设区间数非线性映射 $\Phi(\cdot)$ 将 $\mathbb{IR}^p$ 上的区间数样本集映射到高维区间数特征空间 $\mathbb{IR}^q(q>p)$ 上。利用核技巧 $K([\![\boldsymbol{x}]\!],[\![\boldsymbol{x}']\!])=\langle\Phi([\![\boldsymbol{x}]\!]),\Phi([\![\boldsymbol{x}']\!])\rangle$ 可处理非线性映射的内积，无需 $\Phi(\cdot)$ 的具体形式。考虑聚类中心位于输入空间（InputSpace）与特征空间（FeatureSpace）两种情况，构成 IkP1M-IS 与 IkP1M-FS 算法。

设输入空间中的聚类中心为：

$[\![\boldsymbol{v}^I]\!]\subset\mathbb{IR}^p;[\![\boldsymbol{v}^I]\!]=([\![v_1^I]\!],\cdots,[\![v_p^I]\!]);[\![v_j^I]\!]=[\![\underline{v}_j^I,\overline{v}_j^I]\!];$

$j=1,\cdots,p$

设特征空间中的聚类中心为：

$[\![\boldsymbol{v}^F]\!]\subset\mathbb{IR}^q;[\![\boldsymbol{v}^F]\!]=([\![v_1^F]\!],\cdots,[\![v_q^F]\!]);[\![v_j^F]\!]=[\![\underline{v}_j^F,\overline{v}_j^F]\!];$

$j=1,\cdots,q$

设第 k 个样本对聚类中心的隶属度为 u_k，$\boldsymbol{u}=(u_k)_n$。

3.1.1 第一种目标函数形式

由 PCM 算法的目标函数进行类比，构成第一种形式的目标函数（相应的算法记作 IkP1M-$*$S，$*=$I,F）：

$$\min J_{*S1}(\boldsymbol{u},[\![\boldsymbol{v}^*]\!])=\sum_{k=1}^{n}u_k^m d_{*\Phi k}^2+\eta\sum_{k=1}^{n}(1-u_k)^m$$
$$\text{s.t. } 0\leqslant u_k\leqslant 1;k=1,\cdots,n \tag{7}$$

其中，$m\in(1,+\infty)$ 为模糊平滑因子，η 为惩罚因子，$d_{*\Phi k}^2$ 为核化相异度。

令 $\partial J_{*S1}/\partial\boldsymbol{u}=\boldsymbol{0}$，可得隶属度更新公式：

$$u_k=\frac{1}{1+(d_{*\Phi k}^2/\eta)^{\frac{1}{m-1}}} \tag{8}$$

令 $\partial J_{IS1}/\partial[\![\boldsymbol{v}^I]\!]=0$，IkP1M-IS$_1$ 算法的聚类中心的更新公式为：

$$[\![\boldsymbol{v}^I]\!]=\frac{\sum_{k=1}^{n}u_k^m\widetilde{K}([\![\boldsymbol{x}_k]\!],[\![\boldsymbol{v}^I]\!])[\![\boldsymbol{x}_k]\!]}{\sum_{k=1}^{n}u_k^m\widetilde{K}([\![\boldsymbol{x}_k]\!],[\![\boldsymbol{v}^I]\!])} \tag{9}$$

其中，$\widetilde{K}([\![\boldsymbol{x}_k]\!],[\![\boldsymbol{v}^I]\!])$ 根据所选择的核函数决定。选择 Gauss 核，则有：

$$\widetilde{K}([\![x_k]\!],[\![v^I]\!])=K([\![\boldsymbol{x}_k]\!],[\![\boldsymbol{v}^I]\!])$$
$$=\exp\left[-\frac{d^2([\![\boldsymbol{x}_k]\!],[\![\boldsymbol{v}^I]\!])}{2\sigma^2}\right] \tag{10}$$

输入空间的相异度 $d_{I\Phi k}^2=\|\Phi([\![\boldsymbol{x}_k]\!])-\Phi([\![\boldsymbol{v}^I]\!])\|^2$ 可展开并进行核代入，得到：

$$d_{I\Phi k}^2=K([\![\boldsymbol{x}_k]\!],[\![\boldsymbol{x}_k]\!])+K([\![\boldsymbol{v}^I]\!],[\![\boldsymbol{v}^I]\!])-2K([\![\boldsymbol{x}_k]\!],[\![\boldsymbol{v}^I]\!]) \tag{11}$$

Gauss 核化的相异度 $d_{I\Phi k}^2$ 为：

$$d_{I\Phi k}^2=2-2K_{imG}([\![\boldsymbol{x}_k]\!],[\![\boldsymbol{v}^I]\!])$$
$$=2-2\exp\left[-\frac{d^2([\![\boldsymbol{x}_k]\!],[\![\boldsymbol{v}^I]\!])}{2\sigma^2}\right] \tag{12}$$

令 $\partial J_{FS1}/\partial[\![v^I]\!]=0$，得 IkP1M-FS$_1$ 算法的聚类中心的计算公式为：

$$[\![\boldsymbol{v}^F]\!]=\frac{\sum_{k=1}^{n}u_k^m\Phi([\![\boldsymbol{x}_k]\!])}{\sum_{k=1}^{n}u_k^m} \tag{13}$$

由于式(13)中含有 $\Phi(\|\boldsymbol{x}_k\|)$，无法显式得到 $[\![\boldsymbol{v}^F]\!]$。将特征空间的相异度 $d_{F\Phi k}^2=\|\Phi([\![\boldsymbol{x}_k]\!])-[\![\boldsymbol{v}^F]\!]\|^2$ 代入式(13)，得：

$$d_{F\Phi k}^2=K([\![\boldsymbol{x}_k]\!],[\![\boldsymbol{x}_k]\!])-\frac{2\sum_{r=1}^{n}u_r^mK([\![\boldsymbol{x}_r]\!],[\![\boldsymbol{x}_k]\!])}{\sum_{r=1}^{n}u_r^m}+$$
$$\frac{\sum_{r=1}^{n}\sum_{s=1}^{n}u_r^mu_s^mK([\![\boldsymbol{x}_r]\!],[\![\boldsymbol{x}_s]\!])}{(\sum_{r=1}^{n}u_r^m)^2} \tag{14}$$

Gauss 核化的相异度 $d_{F\Phi k}^2$ 为：

$$d_{F\Phi k}^2=1-\frac{2\sum_{r=1}^{n}u_r^m\exp\left[-\frac{d^2([\![\boldsymbol{x}_r]\!],[\![\boldsymbol{x}_k]\!])}{2\sigma^2}\right]}{\sum_{r=1}^{n}u_r^m}+$$
$$\frac{\sum_{r=1}^{n}\sum_{s=1}^{n}u_r^mu_s^m\exp\left[-\frac{d^2([\![\boldsymbol{x}_r]\!],[\![\boldsymbol{x}_s]\!])}{2\sigma^2}\right]}{(\sum_{r=1}^{n}u_r^m)^2} \tag{15}$$

将式(15)中的隶属度归一化，令 $\tilde{u}_r=u_r^m/\sum_r u_r^m$，则：

$$d_{F\Phi k}^2=1-2\sum_{r=1}^{n}\tilde{u}_r\exp\left[-\frac{d^2([\![\boldsymbol{x}_r]\!],[\![\boldsymbol{x}_k]\!])}{2\sigma^2}\right]+\tilde{\boldsymbol{u}}^{\mathrm{T}}\boldsymbol{G}\tilde{\boldsymbol{u}}$$
$$\tilde{\boldsymbol{u}}=(\tilde{u}_1,\cdots,\tilde{u}_r)^{\mathrm{T}},\boldsymbol{G}=\left\{\exp\left[-\frac{d^2([\![\boldsymbol{x}_i]\!],[\![\boldsymbol{x}_j]\!])}{2\sigma^2}\right]\right\}_{n\times n} \tag{16}$$

其中，$\boldsymbol{G}$ 为 Gram 矩阵。

惩罚因子 η 用于调节正则项的影响程度，文献[12]给出了一种估计方法：

$$\eta=\gamma\frac{\sum_{k=1}^{n}u_k^md_{*\Phi k}^2}{\sum_{k=1}^{n}u_k^m} \tag{17}$$

其中，γ 为惩罚因子系数，一般设置为 1。与 PCM 算法相同，IkP1M-$*$S 算法根据初始化设定的参数估计惩罚因子 η，在主循环中是否更新 η 则需考查是否造成目标函数值不稳定。由于初始化方式的不同将影响惩罚因子的取值，因此在第一次迭代更新聚类中心和隶属度后，重新代入式(17)可获得更准确的惩罚因子 η，随后进行的第二次循环可获取更准确的隶属度向量。离群点检测需确定隶属度阈值作为异常标志，对于系统早期微弱故障征兆的处理，有必要第二次重估惩罚因子 η 并重新迭代获取更精确的隶属度向量。通过大量实验发现，第二次迭代更新仅需几步即可完成，对算法运行效率的影响很小。

3.1.2 第二种目标函数形式

模糊加权指数 m 控制样本隶属度的差异性，其取值与优化一直是基于划分的聚类算法的难点，目前也尚未形成理论指导。对于单簇聚类算法，尽管 m 的影响小于聚类数 $c>1$ 的情况，但将其进行简化仍是有意义的。Krishnapuram 和 Keller[13] 提出了修改目标函数中的正则项，可简化设置模糊加权指数 $m=1$。

受此启发，形成第二种形式区间数核可能性 C 均值单簇聚类算法（后续简记作 IkP1M-$*$S$_2$，$*=$I,F），目标函数为：

$$\min J_{*S2}(\boldsymbol{u},[\![\boldsymbol{v}^*]\!])=\sum_{k=1}^{n}u_kd_{*\Phi k}^2+\eta\sum_{k=1}^{n}(u_k\ln u_k-u_k)$$

s. t. $0 \leqslant u_k \leqslant 1; k=1,\cdots,n$ (18)

聚类中心与惩罚因子的更新公式均不改变，仅需简化 $m=1$。

隶属度更新公式为：

$$u_k = \exp(-d^2_{*\Phi k}/\eta) \tag{19}$$

3.2 IkP1M 系列算法的统一步骤

设定初始化参数后，IkP1M 系列算法首先根据式(17)估计惩罚因子 η，然后第一次循环迭代更新聚类中心、隶属度；待隶属度稳定后，重新估计惩罚因子并再次循环迭代更新聚类中心和隶属度；在隶属度稳定后，输出优化的聚类中心及隶属度向量。下面给出 IkP1M 系列算法的统一步骤。

(1)初始化：设置模糊加权指数 m(IkP1M-$*S_2$ 算法中 $m=1$)、最大迭代次数 MaxIter、惩罚控制常数 γ，迭代终止精度 ε；随机初始化聚类中心$[\![\boldsymbol{v}^*]\!]$、隶属度向量 $\boldsymbol{u}$；根据式(17)初始化惩罚因子 η；设置循环标志 $t=0$；

(2)循环 1：

```
while(t<MaxIter)
    更新隶属度；
    更新聚类中心(IkP1M-IS算法)；
    更新惩罚因子 η；
    if ‖u(t)−u(t−1)‖<ε 循环终止，转至(3)；
    else t=t+1；
```

(3)根据式(17)初始化惩罚因子 η；设置循环标志 $t=0$；

(4)循环 2：

```
while(t<MaxIter)
    更新隶属度；
    更新聚类中心(IkP1M-IS算法)；
    更新惩罚因子 η；
    if ‖u(t)−u(t−1)‖<ε 循环终止，转至(5)；
    else t=t+1；
```

(5)算法结束，输出隶属度向量 $\boldsymbol{u}$ 及惩罚因子 η。

标准 PCM 算法属于期望最大化(Expectation Maximization，E-M)算法，通过更新公式的交替迭代可稳定地收敛。IkP1M 系列算法的目标函数基于标准 PCM 算法改进所得，因此 IkPCM 系列算法是收敛的。

3.3 IkP1M 系列算法的复杂度分析

IkP1M-IS 与 IkP1M-FS 算法在形式上的差异仅在于相异度。IkP1M-IS 算法的相异度 $d^2_{I\Phi k}$ 是一种核诱导的输入空间中样本与聚类中心的相异度；IkP1M-FS 算法的相异度 $d^2_{F\Phi k}$ 是特征空间中样本的象与聚类中心的相异度。$d^2_{I\Phi k}$ 与 $d^2_{F\Phi k}$ 均可根据核技巧展开，$d^2_{I\Phi k}$ 是样本$[\![\boldsymbol{x}_k]\!]$与聚类中心$[\![\boldsymbol{v}^I]\!]$在特征空间中象的相异度；$d^2_{F\Phi k}$ 则因聚类中心$[\![\boldsymbol{v}^F]\!]$无法显式表达，利用了所有样本的加权核描述。由于两类算法考虑不同空间的聚类中心，这从本质上决定了考虑高维特征空间中的聚类中心的 IkP1M-FS 算法的 VC 维比考虑输入空间中的聚类中心的 IkP1M-IS 算法的 VC 维大，即 IkP1M-FS 算法本质上比 IkP1M-IS 算法学习复杂数据形状样本的能力强。

IkP1M-IS 算法的时间复杂度为 $O(np)$，IkP1M-FS 算法的时间复杂度为 $O(n^3 p)$，其中 n 为样本数目，p 为样本维度。据此，在样本数目与特征维度较大时，IkP1M-FS 算法的运行效率远低于 IkP1M-IS 算法。IkP1M-FS 算法需存储核 Gram 矩阵，因此在样本数目与特征维度较大时，其空间复杂度也远高于 IkP1M-IS 算法的空间复杂度。

模糊加权指数 m 的设置是 IkP1M-$*S_1$ 与 IkP1M-$*S_2$ 算法的区别所在。IkP1M-$*S_2$ 修改正则项，将模糊加权指数简化设置为 $m=1$，隶属度更新公式变为指数函数形式，对较大的 $d^2_{*\Phi k}$ 隶属度衰减更快，因此相比 IkP1M-$*S_1$，IkP1M-$*S_2$ 减少了需要优化的参数数目，可降低优化算法的复杂度。

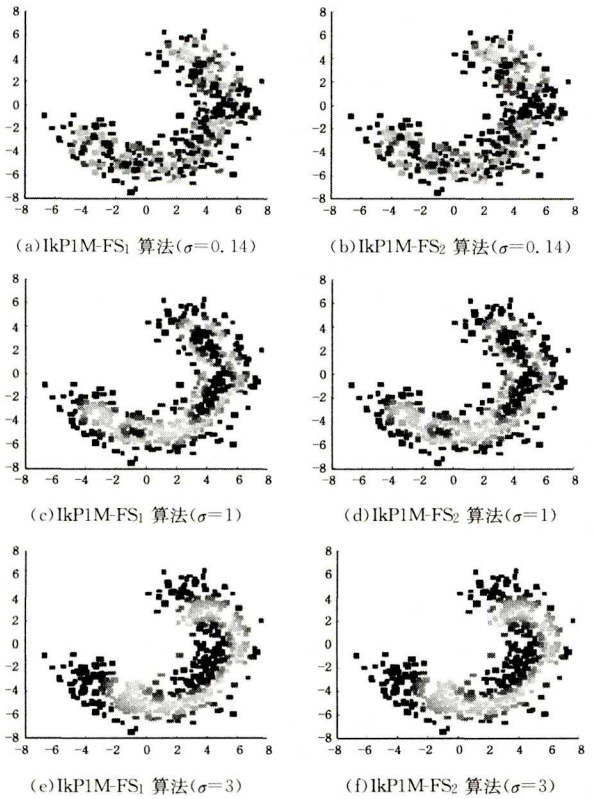
3.4 IkP1M 系列算法的关键参数分析

这里采用区间数向量 Hausdorff 距离作为区间数样本的相异性度量，利用 Gauss 核函数构成核化算法。核宽度 σ、IkP1M-$*S_1$ 算法中的模糊加权指数 m、区间半径 r、惩罚因子系数 γ 是影响相异度 $d^2_{I\Phi k}$ 与 $d^2_{F\Phi k}$ 取值乃至单簇聚类 IkP1M 系列算法的参数。对于单簇聚类，模糊加权指数 m 控制了隶属度的差异显著性。文献[12-13]指出 PCM 聚类算法族适合的取值为 $m=1.5$，本节亦取此值。惩罚因子系数 γ 决定了 IkP1M 算法目标函数中正则项的影响程度，本节按照文献[12]取 $\gamma=1$。核宽度 σ 关系到特征非线性映射的复杂程度，直接影响了算法的 VC 维，其取值直接关系到聚类算法的性能，是最重要的超参数。这里以 IkP1M-FS 算法为例，分析核宽度 σ 对算法学习能力的影响。

基于人工香蕉区间数样本集给出算例。人工香蕉数据集是典型的非超球状数据，较广泛地应用于分类性能检测。首先利用 Prtoolbox[15] 生成 500 个样本点的香蕉形数据簇；然后以人工香蕉数据集样本为中心，随机生成区间半径 $r\in[0.1, 0.25]$的二维区间数样本。固定区间半径加权因子 $\theta=1$，惩罚因子系数 $\gamma=1$，IkP1M-FS_1 算法中的模糊加权指数 $m=1.5$。由于区间半径的差异，即使区间中点位置一致，相异度不同也会造成隶属度不同，因此通常的隶属度等高线图示不适用于描述区间数样本聚类隶属度的分布情况。这里将算法输出的隶属度排序，生成线性 256 级灰阶映像，隶属度由小到大对应由深到浅(灰度 0 至 255)，再由浅到深(灰度值 255 至 0)。灰度图能直观地反映出区间数样本隶属度由小至大的变化情况，且能突出隶属度的极值，即识别区域的边缘及中心。图 1 为 $\sigma=0.14,1,3$ 这 3 种情况下对人工香蕉区间数样本集聚类的结果。由于对于不同的核宽度 σ，聚类算法输出的隶属度范围有所差异，因此图中的灰度仅反映单幅图中样本隶属度的相对关系，而不同图中灰度的样本隶属度没有必然的关联。IkP1M-FS_1 算法与 IkP1M-FS_2 算法聚类输出的隶属度范围具有差异性，但对核宽度 σ 变化的敏感程度基本一致。

经过 IkP1M-FS 算法聚类，人工香蕉区间数据集的隶属度呈现由边缘向中心增大的变化趋势。从图 1 中可以看到，随着核宽度 σ 的增大，相同灰度表示的数据边缘变得光滑，即核宽度 σ 调节了核化相异度分辨率：当 σ 较小时，核化相异度分辨率较高，因此边缘较为粗糙，相同隶属度构成的描述域边界较为紧致；而当 σ 增大时，核化相异度分辨率下降，因此边缘变得平滑，相同隶属度构成的描述域边界开始向外扩张。根据统计学习理论中 VC 维的概念，核宽度 σ 较小时，核函数的复杂度较高，易造成过拟合；核宽度 σ 增大时，核函数的复杂度下降，易造成欠学习。由图 1 所示结果可见，$\sigma=1$ 是 3 种核宽度中较为合适的取值，而 $\sigma=0.14$ 时已有过拟合现象，

$\sigma=3$ 时样本的边缘欠学习。



(a) IkP1M-FS$_1$ 算法($\sigma=0.14$)　(b) IkP1M-FS$_2$ 算法($\sigma=0.14$)

(c) IkP1M-FS$_1$ 算法($\sigma=1$)　(d) IkP1M-FS$_2$ 算法($\sigma=1$)

(e) IkP1M-FS$_1$ 算法($\sigma=3$)　(f) IkP1M-FS$_2$ 算法($\sigma=3$)

图 1　人工香蕉区间数据集的 IkP1M-FS 算法聚类结果

4　异常检测算法

本节提出基于区间数核单簇聚类的单分类器(One Class Classifier based on IkP1M, IkP1M-OCC)异常检测方法。用单分类器处理二分类问题，两类类标分别为正常与离群。IkP1M-OCC 异常检测属于典型的半监督学习，即给定一组已知类标、未知分布规律的训练集：

$$T_{train}=\{([\![\boldsymbol{x}_i]\!],y_i)\}_{i=1}^{n_{tr}};[\![\boldsymbol{x}]\!]=\{[\![x_i]\!]\}_{i=1}^{n_{tr}}$$

$$y=\{y_i\}_{i=1}^{n_{tr}}$$

其中，$[\![\boldsymbol{x}_i]\!]$为区间数样本，y_i 为对应类标(目标类为 1，异常类为 0)，通过学习获得类标分布的规律 $f:[\![\boldsymbol{x}]\!]\rightarrow\boldsymbol{y}$，从而判断未知类标测试集：$T_{test}=\{([\![\boldsymbol{x}_j^*]\!],y_j)\}_{j=1}^{n_{te}}$ 的类标 y_j，$j=1,\cdots,n_{te}$。

本节首先给出了 IkP1M-OCC 的实现方法，然后考虑泛化能力与经验风险的平衡，并详细讨论了异常检测阈值与判别函数的设定。

4.1　IkP1M-OCC 的实现

IkP1M-OCC 的思想是利用 IkP1M 单簇聚类构建目标类描述域，然后设置决策阈值，据此判断测试点的类标。通过设置训练集样本与聚类中心的相异度阈值，可构建一个闭合正常样本描述域，根据相异度可判断检测点是否异常。IkP1M 系列算法输出的隶属度向量给出了样本对聚簇的隶属程度，检测点也可通过聚类算法最终的输出参数结合 IkP1M 系列算法中的计算公式估计隶属度。因此可设置隶属度阈值 u_{th} 作为异常检测的界限，这与相异度的阈值在本质上是等价的。

通过 IkP1M 单簇聚类，获得区间数样本训练集 T_{train} 对目标类的隶属度分布规律：

$$U=\{([\![\boldsymbol{x}_i]\!],u_i)\}_{i=1}^{n_{tr}}$$

其中，u_i 为区间数样本$[\![\boldsymbol{x}_i]\!]$对目标类的隶属度。

聚类算法训练结束后，可用训练输出的参数值计算测试集样本$[\![\boldsymbol{x}^*]\!]$从属于目标类的隶属度 $u([\![\boldsymbol{x}^*]\!])$。设定目标类隶属度阈值 u_{th} 与判别函数 I，判断未知类标测试集的类标。

4.2　异常检测策略

为了保证分类的泛化能力及训练误差(经验风险)的平衡控制，类比于在 SVM 中引入松弛变量，此处引入正常样本拒绝比 υ 来调节隶属度阈值 u_{th}。设训练样本数目为 n，将训练样本输出隶属度向量 $\boldsymbol{u}$ 中的元素由小到大排序得到 $\boldsymbol{u}_r$，然后选取$\boldsymbol{u}_r(n\upsilon+1)$作为隶属度阈值。仅当训练样本数目 n 足够大时，经验风险最小化(即 $\upsilon=0$)才可保证良好的学习效果；当样本数目 n 较小时，拒绝比 υ 减小，虽能保证对训练样本的经验风险下降，但可能带来泛化能力的下降；拒绝比 υ 增大，泛化能力上升，但对训练样本的经验风险却上升。因此需选择合适的拒绝比 υ，以保证分类的经验风险与泛化能力的平衡。

策略 1　确定隶属度阈值 u_{th} 后，根据测试集样本$[\![\boldsymbol{x}^*]\!]$从属于目标类的隶属度 $u([\![\boldsymbol{x}^*]\!])$与 u_{th} 的关系，直接判断测试集样本的类标。判别函数为：

$$I_1[u([\![\boldsymbol{x}^*]\!])]=\begin{cases}0, & u([\![\boldsymbol{x}^*]\!])<u_{th}\Rightarrow[\![\boldsymbol{x}^*]\!]\notin\boldsymbol{S}_{normal}\\ 1, & u([\![\boldsymbol{x}^*]\!])>u_{th}\Rightarrow[\![\boldsymbol{x}^*]\!]\in\boldsymbol{S}_{normal}\end{cases}\tag{20}$$

其中，$\boldsymbol{S}_{normal}$ 为正常样本空间。

策略 2　考虑测试集与训练集样本属性区间半径的不平衡。设测试集样本$[\![\boldsymbol{x}^*]\!]$中各属性区间半径为 $r_1^*,\cdots,r_p^*$，训练样本各属性的区间半径均值为 $\bar{r}_1,\cdots,\bar{r}_p$。根据 $r_1^*,\cdots,r_p^*$ 与 $\bar{r}_1,\cdots,\bar{r}_p$ 之间的关系，分别划分相应属性区间数，即将测试集样本$[\![\boldsymbol{x}^*]\!]$划分为更小的区间数样本。根据子样本与训练集的隶属关系，判断测试集样本$[\![\boldsymbol{x}^*]\!]$的类标。测试集样本$[\![\boldsymbol{x}^*]\!]$第 j 个属性划分数为：

$$k_j^{div}=\lceil r_j^*/\bar{r}_j\rceil\tag{21}$$

其中，$\lceil\cdot\rceil$表示向上取整。根据式(21)，测试集样本被划分为 $\prod_j^p k_j^{div}$ 个子样本。

计算子样本从属于目标类的隶属度，当且仅当 50%以上的子样本的隶属度大于隶属度阈值 u_{th} 时，测试集样本的类标为目标类。据此，策略 2 的类标判别函数为：

$$I_2[u([\![\boldsymbol{x}^*]\!])]=\begin{cases}0, & \dfrac{1}{\prod k_j^{div}}\sum_{k_1=1}^{k_1^{div}}\cdots\sum_{k_p=1}^{k_p^{div}}I_1(u([\![\boldsymbol{x}^*_{\prod k_j}]\!]))<\\ & 0.5\Rightarrow[\![\boldsymbol{x}^*]\!]\notin\boldsymbol{S}_{normal}\\ 1, & \dfrac{1}{\prod k_j^{div}}\sum_{k_1=1}^{k_1^{div}}\cdots\sum_{k_p=1}^{k_p^{div}}I_1(u([\![\boldsymbol{x}^*_{\prod k_j}]\!]))\geqslant\\ & 0.5\Rightarrow[\boldsymbol{x}^*]\in\boldsymbol{S}_{normal}\end{cases}\tag{22}$$

其中，$[\![\boldsymbol{x}^*_{\prod k_j}]\!]$表示划分位置为$(k_1,\cdots,k_p)$的子样本。

策略 1 直接利用测试集样本计算与目标类的隶属度，检测效率较高。策略 2 考虑了测试集样本与训练样本属性区间宽度的关系，在测试集样本属性区间宽度过大时，通过细分测试集样本来平衡训练样本与测试集样本的不确定程度。经过式(21)的划分，子样本的各属性区间半径为训练集属性区间半径均值的 0.5～1 倍，可进一步确定测试集样本中的正常区域。

5 异常检测算法的参数选择

Gauss 核宽度 σ 决定了聚类算法的性能，拒绝比 υ 可保证异常检测中泛化能力与经验风险的平衡，是基于 Gauss 核的半监督区间数核可能性单簇聚类异常检测方法的关键参数。文献[14]指出，单簇聚类中模糊加权因子 m 的取值仅影响了训练样本隶属度变化范围的显著性，对算法性能没有实质影响；又因 IkP1M-$*S_2$ 算法的目标函数已将模糊加权因子 m 取为 1，因此不再考虑模糊加权因子的寻优。利用基于 Grid Search 的 K-折交叉验证(K-fold Cross Validation)选择参数、核宽度 σ 与拒绝比 υ。

基于 Grid Search 的 K-折交叉验证参数优化方法的主要思想是：首先将训练样本集随机等分为 K 个不相交的子集，以 $K-1$ 个子集作为训练集，余下的 1 个子集作为测试集；然后以 1 个子集为测试集，其余 $K-1$ 个子集作为训练集，重复执行 K 次网格遍历寻优，保存每次训练的最优参数值；最后综合 K 次寻优结果，得到全局最优参数值。由于 IkP1M-OCC 算法仅有两个寻优参数，且能根据先验知识压缩参数寻优区域，因此基于 Grid Search 的 K-折交叉验证亦能获得令人满意的效率。对于实时性要求较高的情况，可采用启发式算法，以提高寻优效率。

F-measure 是统计分类中一种常用的性能指标，以加权调和的形式综合了准确率(Precision)和召回率(Recall)。其计算公式为：

$$F_a=\frac{(\alpha^2+1)PR}{\alpha^2P+R} \tag{23}$$

其中，α 为常数，P 为准确率，R 为召回率。取 $\alpha=1$，构成 F_1-measure 指标。这里选取 F_1-measure 指标作为检测性能的评价指标。F_1-measure$\in[0,1]$，越接近 1 说明检测的性能越好。

设训练样本集共有 N_t 个样本，则每一个子集含 N_t/K 个样本，每次聚类与异常检测得到的分类结果与实际类标的数目即混淆矩阵(Confusion matrix)如表 1 所列。

表 1 聚类与异常检测的输出结果

		判断结果	
		正常	异常
实际	正常	TP	TN
	异常	FP	FN

根据表 1 及有关定义可知，目标类判断的准确率、召回率、F_1-measure 指标分别为：

$$\begin{aligned}P_n&=\frac{TP}{TP+FP}\\R_n&=\frac{TP}{TP+TN}\\F_{1n}&=\frac{2P_nR_n}{P_n+R_n}=\frac{2TP}{2TP+TN+FP}\end{aligned} \tag{24}$$

同理，异常类判断的准确率、召回率、F_1-measure 指标分别为：

$$\begin{aligned}P_a&=\frac{FN}{TN+FN}\\R_a&=\frac{FN}{FP+FN}\\F_{1a}&=\frac{2P_aR_a}{P_a+R_a}=\frac{2FN}{TN+FP+2FN}\end{aligned} \tag{25}$$

根据设定的拒绝比 υ，在一次聚类与异常检测中，目标类与异常类的实际数目分别为：

$$\begin{aligned}N_n&=[(1-\upsilon)-(1-\upsilon)/K]N_t\\N_a&=[\upsilon+(1-\upsilon)/K]N_t\end{aligned} \tag{26}$$

检测结果的 F_1-measure 为：

$$F_1=\frac{N_nF_{1n}+N_aF_{1a}}{N_n+N_a} \tag{27}$$

参数选择算法的流程为：

(1)初始化。将训练样本集 T 随机等分为 K 个子集 $T_1,\cdots,T_K$；设置核宽度 σ 与拒绝比 υ 的约束域；设置 IkP1M 聚类算法相关参数的初值。

(2)参数寻优。依次选取 $T_1,\cdots,T_K$ 为测试集、其余子集为训练集进行参数寻优：

1)参数约束域的优化。随着核宽度 σ 的增大，聚类算法的学习能力单调下降。据此，结合训练结果，缩减核宽度的寻优区域。

2)设置参数约束域的网格密度，遍历执行 IkP1M-OCC 算法，保存每次训练参数值与对应的 F_1-measure。

3)输出当前训练集与测试集对应 F_1-measure 最大的核宽度 σ 与拒绝比 υ。

(3)输出。计算全局核宽度 σ 与拒绝比 υ。设当 $T_1,\cdots,T_K$ 为测试集时，输出的参数值分别为 $[\sigma_1,\upsilon_1],\cdots,[\sigma_K,\upsilon_K]$，则全局核宽度 σ 与拒绝比 υ 为：

$$\sigma=\frac{\sum_{i=1}^{K}\sigma_i}{K};\upsilon=\frac{\sum_{i=1}^{K}\upsilon_i}{K} \tag{28}$$

(4)训练结束。

6 算例

本节给出 IkP1M-OCC 的算例分析，以考查 IkP1M-OCC 对人工香蕉样本集与小规模 UCI 真实数据集的检测能力；比较两种异常检测策略，并分析 IkP1M-OCC 与典型 SVM-OCC 进行异常检测的性能。

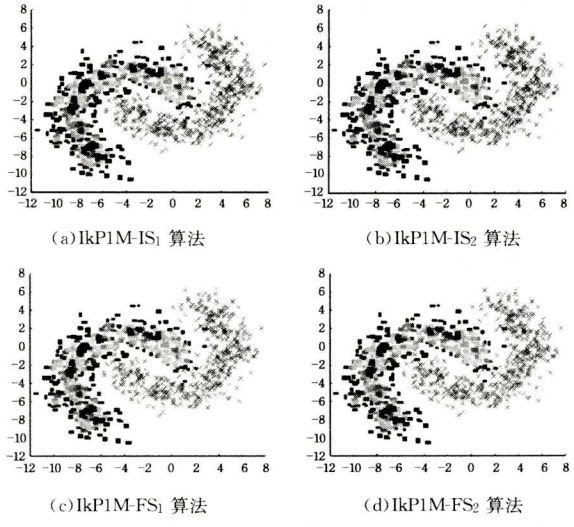
6.1 人工数据集试验

本节利用人工香蕉区间数样本集来考查与比较采用不同 IkP1M 算法 OCC 的检测能力。首先利用 Prtoolbox 生成两簇香蕉数据集，每类含 500 个样本，分布在实数域$(-12\times8)\times(-12\times8)$中；以人工香蕉数据集样本为中心，随机生成区间半径 $r\in[0.1,0.25]$的二维区间数样本。分别设定两簇样本的类标为正常与离群。

为了比较 IkP1M-OCC 的检测能力，统一设置拒绝比 υ 为 0。由于两簇区间数样本集的区间宽度按照同一规则设定，因此可直接按照策略 1 执行检测。设置惩罚因子系数 $\gamma=1$，IkP1M-$*S_1$ 算法的模糊加权指数 $m=1.5$。根据正常样本隶属度范围设置隶属度阈值 u_{th}。IkP1M-FS 算法的核宽度 σ 由 5 折交叉验证训练目标类样本直接得到；由于 IkP1M-IS 算法对香蕉样本集的聚类能力有限，直接以 F_1-measure 指标的交叉验证训练易导致过学习，即所有样本隶属度基本接近，为了防止 IkP1M-IS 算法的过学习，限制了过小的核宽度 σ 取值。IkP1M 算法中两种目标函数形式的聚类效果基本一致，因此 IkP1M-IS_1 与 IkP1M-IS_2 算法、IkP1M-FS_1 与 IkP1M-FS_2 算

法训练得到的核宽度取值一致。

图 2 所示为 IkP1M-OCC 进行一次试验的检测结果。按 3.4 节所述，灰度映像表示正常样本的聚类结果，拒绝的离群样本用灰度为 0 的×表示，检测正确的样本用灰度为 76 的×表示。



(a) IkP1M-IS_1 算法　(b) IkP1M-IS_2 算法

(c) IkP1M-FS_1 算法　(d) IkP1M-FS_2 算法

图 2　IkP1M-OCC 的聚类与离群点检测结果

图 2 所示的聚类结果验证了 IkP1M-FS 算法对复杂非球状数据集的聚类效果优于 IkP1M-IS 算法：图 2(a)、图 2(b)中聚类中心周围的暗色样本呈圆形，样本边界描述效果不佳；图 2(c)、图 2(d)中暗色样本较好地描述了人工香蕉区间数样本集的边缘。IkP1M-OCC 的 50 次试验检测的均值结果如表 2 所列。由于聚类中心初始化的随机性，IkP1M-IS-OCC 检测性能不稳定，最差的检测正确率低于 80%；而 IkP1M-FS-OCC 检测性能稳定，离群检测率基本保持在 99%以上。在 Core i5 6200U、8GB 的计算机上，基于 Matlab 2015b 平台实现的两种形式的 IkP1M-FS 聚类平均耗时均为 180s，多于实现 IkP1M-IS 算法的 47s 左右。实际应用中样本集的形状具有不确定性，由于 Gauss 核 IkP1M-FS 算法具有更强的复杂数据形状样本处理能力，因此尽管牺牲了算法性能，但 Gauss 核 IkP1M-FS 算法比 Gauss 核 IkP1M-IS 算法更适用，且保证了 OCC 检测性能的稳定。

表 2　IkP1M-OCC 的聚类与离群点检测试验结果

聚类算法	核宽度	检测正确率/%	F_1-measure	运行时间/s
IkP1M-IS_1	0.7071	86.2	0.9259	179.8
IkP1M-IS_2	0.7071	87.4	0.9328	180.2
IkP1M-FS_1	0.3450	99.4	0.9950	46.7
IkP1M-FS_2	0.3450	99.6	0.9980	48.3

6.2　区间化的 UCI 真实数据集试验

本节考查 IkP1M-OCC 对区间化的 UCI 真实数据集[16]的检测能力。参照文献[10]的思路，将真实数据集样本作为区间数中心，然后分别考查已知类中各维(特征)的标准差$\boldsymbol{s}=(s_1,\cdots,s_p)$，设置区间扩展因子 IDC。据此扩展原始数据集，第 $j(j=1,\cdots,p)$维属性的最大区间半径为 $s_1\cdot$IDC。

本节采用 4 个小规模 UCI 数据集即 Seeds，Iris，Glass-Identification以及 Ionosphere 进行试验。Seeds，Iris，Glass-Identification数据集包含多个类别，这里采用 one-against-rest 策略，取一类为目标类，余下为异常类。由于需要比较采用不同单簇聚类算法 OCC 的检测效果，因此对每一组数据集中的同一目标类统一设置 4 种单簇聚类算法的拒绝比。聚类算法的核宽度由 5-折交叉验证得到。各数据集的区间扩展因子 IDC 与 IkP1M-OCC 异常检测算法的参数如表 3 所列。

表 3　IkP1M-OCC 对区间化 UCI 数据集的离群检测试验结果

数据集	目标类	扩展因子 IDC	拒绝比 υ	核宽度 σ			
				IS_1	IS_2	FS_1	FS_2
Seeds	Canadian，Kama，Rosa	10	0	2.2361	2.2813	2.2917	2.3269
Iris	Setosa，Versicolour，Virginica	5	0.05	7.0711	7.0711	7.5827	7.5827
Glass Identification	bldg win-f	5	0.05	5	5	5	5
	bldg win-n			5	5	1.5811	1.5811
	veh win-f			7.0711	7.0711	1	1
	containers，tableware，headlamps			2.5435	2.5487	1	1
Ionosphere	good	0.1	0.05	10	10	1	1
	bad			10	10	3.1623	3.1623

在 Core i5 6200U、8GB 的 Matlab 2015b 平台上，对各类中的每种算法均进行 50 次试验。由于所有类生成区间数的规则一致，因此采用检测策略 1。试验检测均值如表 4 所列，其中分别记录了异常类检测的 F_1-measure 值 F_{1a}、全局 F_1-measure 值 F_1 以及运行检测算法的典型时间，最优试验结果加粗标注。

由表 4 可知，一般地，IkP1M-IS-OCC 比 IkP1M-FS-OCC 检测效率更高，但在的小规模低维特征数据集上的优势不明显。IkP1M-FS 算法在运行中首先需运算与存储 Gram 矩阵，其效率和复杂度与样本数目和特征维度直接相关，因此样本规模增大、特征维度升高均会使 IkP1M-FS 算法的效率下降。在试验中，IkP1M-FS 算法的迭代次数远比 IkP1M-IS 算法的少，且聚类结果不受聚类中心随机初始化的影响。

Seeds 数据集包含 3 类麦种数据(Canadian，Kama，Rosa)，每类含 70 个样本，样本特征维度为 7。试验中分别随机取每类中的 60 个样本为目标集，其余 10 个样本与剩余两类作为异常类。由记录的异常类 F_{1a} 与全局 F_1 知，IkP1M-FS-OCC 比 IkP1M-IS-OCC 的检测性能更优。由于数据集中 Canadian 类与 Kama 类、Rosa 类的分离度较好，而 Kama 类、Rosa 类的分离度较差，因此将 Kama 类、Rosa 类作为目标集时，全局 F_1 下降较为明显。

Iris 数据集包含 3 类莺尾花(Setosa，Versicolour 和 Virginica)花瓣与花萼的测量数据，每类含 50 个样本，样本特征维度为 4。试验中分别随机选取每类中的 40 个样本为目标

集，其余 10 个样本与剩余两类作为异常类。Setosa 类数据与 Versicolour 和 Virginica 类数据完全分离，在以 Setosa 类数据为目标集时，IkP1M-IS-OCC 尽管在检测效率上低于 IkP1M-FS-OCC，但取得了更高的异常类 F_{1a} 与全局 F_1。由于 Versicolour 和 Virginica 类数据分离性较差，因此 IkP1M-FS-OCC 的检测性能更佳。

GlassIdentification 数据集包含 6 类玻璃测试数据，共计 214 个样本，样本特征维度为 9。Vehicle_windows_non_flat_processed，tableware，headlamps 类作为目标类时，训练集远小于测试集，属于典型的样本不均衡单分类问题。IkP1M-FS-OCC 在处理此类样本不均衡问题时比 IkP1M-IS-OCC 的性能稳定。

Ionosphere 数据集是美国 GooseBay 的雷达系统的电离层测量数据，共计 351 个样本，样本特征维度为 34，已由样本数据确定了正常(225 个)与异常(126 个)的类标。由于样本特征维度较高，IkP1M-FS-OCC 的检测效率明显低于 IkP1M-IS-OCC，但目标类与异常类分别作为训练集时均能获得稳定的检测效果。

表 4 区间化 UCI 数据集的 IkP1M-OCC 离群检测试验结果

数据集	目标类	训练集	测试集	F_{1a}				F_1				运行时间/s			
				IS_1	IS_2	FS_1	FS_2	IS_1	IS_2	FS_1	FS_2	IS_1	IS_2	FS_1	FS_2
Seeds	Canadian	60	150	0.8571	0.8617	0.9821	**0.9859**	0.6837	0.6884	**0.9192**	0.9185	6.78	**5.94**	9.65	9.67
	Kama	60	150	0.6190	0.9098	0.9643	**0.9655**	0.4739	0.7339	**0.8316**	0.6897	5.83	**3.2**	9.78	9.63
	Rosa	60	150	0.9070	0.8941	**0.9756**	0.9655	0.7703	0.7529	**0.8287**	0.6897	**4.3**	5.53	9.61	9.76
Iris	Setosa	40	110	**0.9950**	0.9901	0.9901	0.9662	**0.9823**	0.9631	0.9631	0.8316	6.7	6.04	4.26	**4.25**
	Versicolour	40	110	0.9412	0.9417	0.9703	**0.9709**	0.7569	0.7287	**0.8893**	0.8644	6.61	**3.77**	4.36	4.21
	Virginica	40	110	0.6111	0.8603	0.9519	**0.9569**	0.5183	0.7350	0.7425	**0.7502**	**0.86**	1.70	4.23	4.24
Glass Identification	bldg win-f	60	144	0.8938	0.8938	0.8706	**0.8986**	0.6912	0.6912	**0.7313**	0.6817	3.30	**2.37**	9.15	9.15
	bldg. win-n	65	149	0.6816	**0.7105**	0.6087	0.6512	0.5346	**0.5582**	0.4920	0.5210	9.04	**6.55**	10.1	10.2
	veh win-f	12	202	0.9585	0.9849	0.9771	**0.9875**	0.9110	0.9247	**0.9325**	0.9321	0.90	**0.83**	1.93	1.93
	containers	8	206	0.2575	0.2193	**0.9877**	**0.9877**	0.2491	0.2123	**0.9508**	**0.9508**	**0.37**	0.38	1.35	1.34
	tableware	6	208	0.9677	0.5423	**0.9927**	**0.9927**	0.9406	0.5275	**0.9649**	**0.9649**	**0.30**	0.32	1.06	1.00
	headlamps	24	190	0.5475	0.6259	**0.9871**	**0.9871**	0.4963	0.5672	**0.8794**	**0.8794**	2.43	**1.28**	3.86	3.90
Ionophere	good	200	151	0.9420	0.9490	0.9323	**0.9575**	0.8205	0.8205	0.7809	**0.8360**	**29.1**	31.5	50.5	50.3
	bad	111	240	0.8223	0.7979	0.9605	**0.9628**	0.6211	0.5994	0.7358	**0.7408**	9.48	**5.40**	28.4	28.2

综合各数据集的试验结果，可得如下结论：

(1)在小规模数据集试验中，IkP1M-IS-OCC 与 IkP1M-FS-OCC 的效率相差不大，且在数据形状不复杂的情况下，它们均能取得较好的检测效果；

(2)在样本不均衡、形状复杂、分离度较差的数据集试验上，IkP1M-FS-OCC 的检测效果优于 IkP1M-IS-OCC。

试验中为了横向比较采用 4 种 IkP1M-OCC 进行检测的效果，规定 OCC 的拒绝比相同，交叉验证参数选择的结果导致异常类检测准确率(Precision)较高，但目标类检测准确率下降，进而导致全局 F_1-measure 值下降。实际应用中，针对具体问题训练核宽度与拒绝比，可更好地平衡准确率与召回率(Recall)。

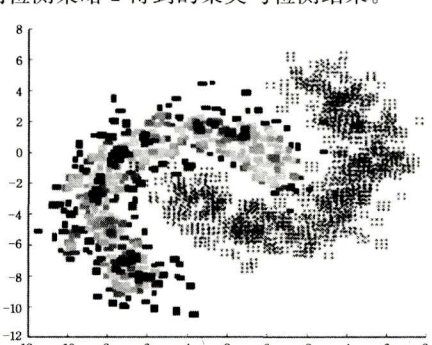
6.3 两种检测策略的比较

本节比较第 4 节提出的两种检测策略。策略 1 利用测试样本的隶属度直接判断类标，策略 2 则考虑了测试样本与训练样本的区间半径不平衡。检测策略 2 比较测试样本与训练样本各属性的区间半径，并将测试样本划分为子样本。本节重新生成人工香蕉区间数样本集与 Iris 区间样本集作为算例，比较两种检测策略的检测结果 TP，TN，FP，FN 以及异常类检测的 F_1-measure 值 F_{1a} 和全局 F_1-measure 值 F_1。本节还对策略 2 的检测结果进行了可视化，以直观地反映策略 2 的检测结果。

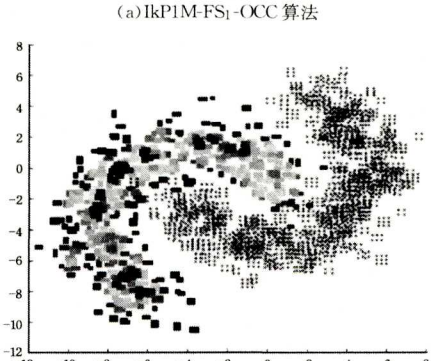
6.3.1 人工香蕉区间数样本集

这里仍利用 6.1 节中的两簇香蕉数据集，以原始数据集样本为中心，一簇随机生成区间半径 $r\in[0.1,0.25]$ 的二维区间数样本作为目标类，另一簇作为异常类测试，其区间半径 $r\in[0.25,0.4]$。考虑到 IkP1M-IS-OCC 性能问题，这里仅测试 IkP1M-FS-OCC 的检测性能。

图 3(a)、图 3(b)分别为 IkP1M-FS_1-OCC 与 IkP1M-FS_2-OCC 采用检测策略 2 得到的聚类与检测结果。



(a)IkP1M-FS_1-OCC 算法



(b)IkP1M-FS_2-OCC 算法

图 3 人工香蕉区间数样本集 IkP1M-FS-OCC 离群检测试验结果

图 3 中，用灰度为 76 的·标记判断类标为异常类的测试

子样本，用灰度为 0 的 · 标记判断类标为目标类的测试子样本，目标类用灰度反映隶属度梯度。

表 5 比较了两种策略下 IkP1M-FS-OCC 的 50 次检测结果的均值，性能较优的结果加粗标注。

表 5　人工香蕉区间数样本集 IkP1M-FS-OCC 离群检测试验结果

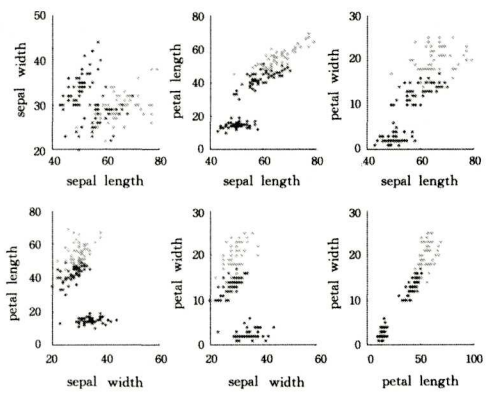
检测策略	聚类算法	核宽度	检测正确率/%	F_1
策略 1	IkP1M-FS$_1$	0.3450	99.4	0.9950
	IkP1M-FS$_2$	0.3450	99.6	0.9980
策略 2	IkP1M-FS$_1$	0.3450	**100**	**1**
	IkP1M-FS$_2$	0.3450	**100**	**1**

由图 3 与表 5 可知，两种检测策略都获得了稳定的检测结果。策略 1 中误判的样本在策略 2 中一半子集的类标为目标类，未达到 50%的阈值，故被判断为异常类。策略 2 将每一测试样本细分，可获得测试样本内部子集的类标，在测试样本具有较大不确定性时有助于获取测试集类标的真实边界，容易获得更优的检测效果。

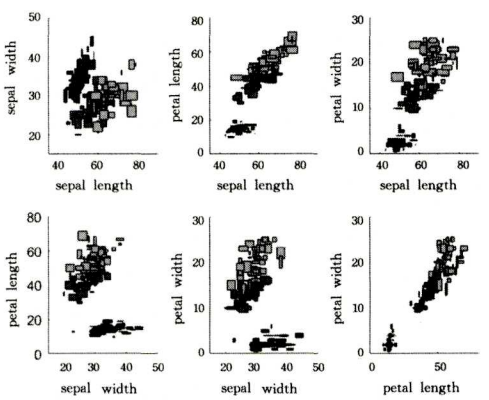
6.3.2　Iris 区间数样本集

本节比较两种检测策略对 Iris 区间数样本集的检测性能，生成区间数样本的规则与 6.2 节一致，设置区间扩展因子 $IDC=0.5$。由于 Iris 数据集含有 4 个属性(sepal width，sepal length，petal length，petal width)，为了将 Iris 数据集可视化，分别将 4 个属性两两组合，构成 6 个二维样本图。

图 4(a)、图(b)分别给出了原始 Iris 数据集与生成的 Iris 区间数样本集的分布情况。



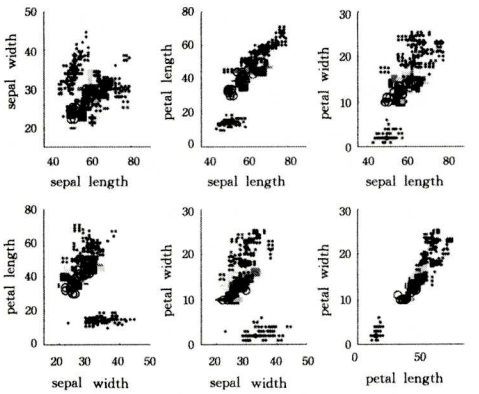
(a) Iris 数据集：3 类 4 维原始数据



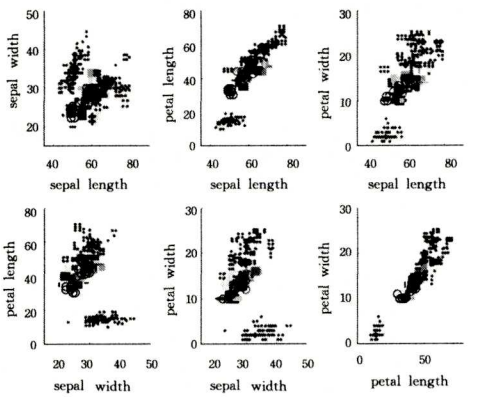
(b) Iris 区间数样本集($IDC=0.5$)

图 4　Iris 数据集及其区间扩展

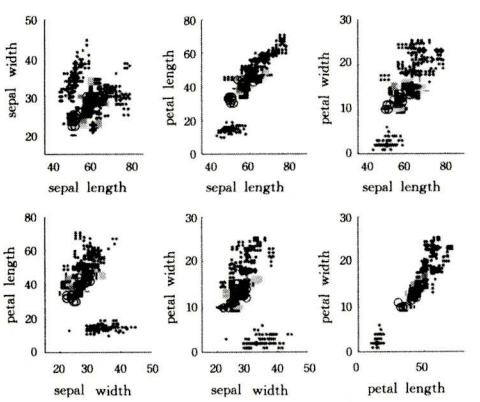
图 4 中 Setosa 类的灰度值为 83，Versicolour 类的灰度值为 109，Virginica 类的灰度值为 168。由图 4(a)可直观看出，Setosa 类与 Versicolour 类、Virginica 类各属性均是分离的，而 Versicolour 类、Virginica 类各属性的分离度较差。由图 4(b)可知，当 $IDC=0.5$ 时，区间化的 Setosa 类、Versicolour 类及 Virginica 类在 4 维空间中仍基本保持了上述关系。



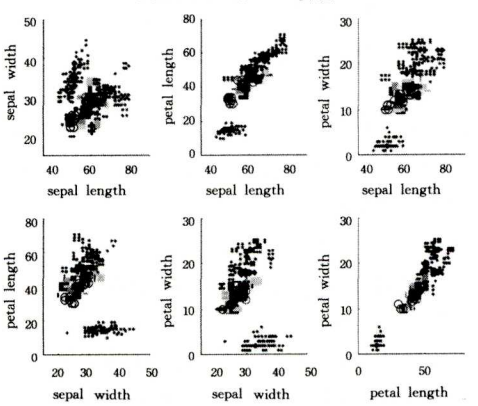
(a) IkP1M-IS$_1$-OCC 算法



(b) IkP1M-IS$_2$-OCC 算法



(c) IkP1M-FS$_1$-OCC 算法



(d) IkP1M-FS$_2$-OCC 算法

图 5　采用检测策略 2 的区间化 Iris 数据集离群检测试验结果

本节随机选择 Versicolour 类中 40 个区间数样本为目标类，其余样本与 Setosa 类、Virginica 类的所有样本构成测试集，以考查采用 4 种 IkP1M-OCC 及两种检测策略的检测效果。图 5(a)、图 5(b)分别为采用策略 2 时 IkP1M-IS_1-OCC 与 IkP1M-IS_2-OCC 的聚类与检测结果，图 5(c)、图 5(d)分别为采用策略 2 时 IkP1M-FS_1-OCC 与 IkP1M-FS_2-OCC 的聚类与检测结果。用灰度为 76 的·标记实际类标为异常类、判断类标为异常类(False Negtive，FN)的测试子样本，用灰度为 0 的·标记实际类标为异常类、判断类标为目标类(False Positive，FP)的测试子样本；用灰度为 76 的。标记实际类标为目标类、判断类标为目标类(Ture Positive，TP)的测试子样本，用灰度为 0 的。标记实际类标为目标类、判断类标为异常类(Ture Negtive，TN)的测试子样本；目标类用灰度反映隶属度梯度。为了比较 4 种算法的检测性能，统一设置拒绝比为 0；聚类算法的核宽度由 5 折交叉验证获得。

表 6 列出了 IkP1M-OCC 的参数及 50 次检测结果的均值，性能较优的结果加粗标注。

表 6 Iris 区间数样本集(IDC= 0.5)IkP1M-OCC 离群检测试验结果

检测策略	聚类算法	核宽度	F_1	F_{1a}	F_{1n}
策略 1	IkP1M-IS_1	7.0711	0.7944	**0.9615**	0.3333
	IkP1M-IS_2	7.0711	0.8702	0.9538	0.6400
	IkP1M-FS_1	4.0825	**0.8928**	0.9645	**0.6957**
	IkP1M-FS_2	4.0825	0.8812	0.9592	0.6667
策略 2	IkP1M-IS_1	7.0711	0.7944	**0.9615**	0.3333
	IkP1M-IS_2	7.0711	0.8596	0.9485	0.6154
	IkP1M-FS_1	4.0825	**0.8812**	0.9592	**0.6667**
	IkP1M-FS_2	4.0825	0.8702	0.9538	0.6400

根据图 5 与表 6 可知，对于 Iris 区间数样本集，IkP1M-FS1-OCC 的总体检测性能较好，可较好地均衡准确率与召回率。IkP1M-IS_1-OCC 在两种检测策略下均获得对异常类 100%的检测率，但对目标类的检测率只有 20%。IkP1M-IS_2-OCC，IkP1M-FS_1-OCC，IkP1M-FS_2-OCC 在两种检测策略下均取得了良好、稳定的检测效果，但对异常类的检测率均下降了 1%；漏判样本中超过 50%的细分子样本的隶属度大于隶属度阈值。这说明 50%的固定阈值的设置还可进一步优化。

由图 4、图 5 与表 6 可知，尽管 Versicolour 类与 Virginica 类在数据结构上分离度较差，但 4 种核化的聚类算法均能获得稳定的目标数据描述，其中 IkP1M-IS_2-OCC，IkP1M-FS_1-OCC，IkP1M-FS_2-OCC 在两种检测策略下均可较好地平衡准确率与召回率。

6.4 与典型区间数样本 SVM-OCC 算法的性能比较

下面比较标准 C-B，Int-C-B 以及 4 种 IkP1M-OCC 对区间化的 GlassIdentification 数据集、Seeds 数据集的异常检测性能。为了方便比较算法性能，区间化 UCI 真实数据集的方法与文献[10]中的一致。数据集的区间扩展系数 IDC 的取值见表 3。在 GlassIdentification 区间样本集中，将前两类作为目标类，余下 4 类作为异常类；在 Seeds 区间数样本集中，将 Kama 类、Rosa 类作为目标类，Canadian 类作为异常类。表 7 列出了标准 C-B，Int-C-B，4 种 IkP1M-OCC 的核参数 KC(Int-C-B 为 Triangular 核参数 γ，IkP1M-OCC 为 Gauss 核宽度 σ)及 50 次重复试验检测准确率(ACC)均值，每一数据集的最高准确率加粗标注。

表 7 C-B，Int C-B，IkP1M-OCC 离群检测试验结果

OCC 类型	策略 1				策略 2			
	Glass		Seeds		Glass		Seeds	
	KC	ACC	KC	ACC	KC	ACC	KC	ACC
C-B	2	0.4068	2	0.6459	2	0.4161	2	0.6460
Int-C-B	2	**0.8126**	2	0.8462	2	**0.8909**	2	0.8400
IkP1M-IS_1-OCC	1.3417	0.8088	2.2361	0.8571	1.3417	0.8215	2.2361	0.8286
IkP1M-IS_2-OCC	1.2561	0.7941	2.2813	0.8714	1.2561	0.8039	2.2813	0.8428
IkP1M-FS_1-OCC	1.2910	0.7647	2.2917	**0.9286**	1.2910	0.7842	2.2917	**0.9121**
IkP1M-FS_2-OCC	1.3156	0.7523	2.3269	0.9142	1.3156	0.7713	2.3269	0.9087

由表 7 可知，在检测的区间数样本具有相对较大的不确定性时，分别采用策略 1 与策略 2，IkP1M-OCC 均取得了远高于标准 C-B 的准确率。对于 Glass Identification 区间数样本集，Int C-B 的准确率在两种策略下均略高于 4 种 IkP1M-OCC；对于 Seeds 区间数样本集，IkP1M-FS-OCC 在两种策略下的准确率最高。当拒绝比 υ 增大时，Int C-B 算法中计算极值点 $\alpha(1),\cdots,\alpha(T)$ 的复杂度上升，但极值点 $\alpha(1),\cdots,\alpha(T)$ 的数目与样本属性数目无关，故处理高维样本的效率较高。为了将分类模型归结为一系列线性规划问题，在 Int C-B 中利用 Trian-gular 核线性近似 Gauss 核，该近似具有一定的限制。IkP1M-OCC 的缺陷是算法复杂度与空间复杂度随着样本属性维数的升高而上升，尤其是 IkP1M-FS-OCC，其复杂度对样本属性维数的升高更加敏感。但 IkP1M-OCC 比 IntC-B 具有更紧凑、更清晰的数学描述，且不存在替换线性化近似核的局限性，适用范围更广。

结束语 针对在实际系统中异常数据难以大量获取的问题，本文将基于单簇聚类-单分类器的思想用于异常检测。考虑监测数据的不确定性，提出了区间数样本的核可能性 1-均值单簇聚类算法，采用 Gauss 核化的 Hausdorff 区间数样本距离，结合两种目标函数形式，建立了在输入空间与特征空间中的两类单簇聚类算法。引入拒绝比来平衡泛化误差与经验风险，构造了具有松弛边界的单分类器。针对区间数样本的属性区间半径的不确定性，讨论了两种检测策略。最后利用人工数据集、区间化的 UCI 真实数据集验证了所提 IkP1M-OCC 的有效性，与典型 SVM-OCC 性能相比，IkP1M-OCC 具有一定的优越性。

区间数具有良好的运算性质，能有效地描述数据的不确定性。如何针对具体问题，从区间数样本的不确定性中寻找确定性的规律以及收缩表示不确定性的区间宽度是值得进一步研究的问题。复杂系统运行中可能产生高维特征数据，本文所提的 IkP1M-OCC 暂未考虑区间数样本具有高维特征的

(下转第 205 页)

Search System Based on Web Information and the Visibility of Scenic Sights [C]// Second International Symposium on Universal Communication (ISUC 2008). 2008:154-161.

[9] CRANDALL D, BACKSTROM L, HUTTENLOCHER D, et al. Mapping the World's Photos [C]// Proc. Int. Conf. on World Wide Web (WWW). 2009:761-770.

[10] ASNBROOK D, STARNER D. Using GPS to Learn Significant Locations and Predict Movement across Multiple Users [J]. Personal and Ubiquitous Computing, 2003, 7(5):275-286.

[11] IWATA T, WATANABE S, YAMADA T. Topic Tracking Model for Analyzing Consumer Purchase Behavior [C]// Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI). 2009:1427-1432.

[12] HOFMANN T. Probabilistic Latent Semantic Analysis [C]// Proc. Conf. on Uncertainty in Artificial Intelligence (UAI). 1999:289-296.

[13] KURASHIMA T, IWATA T, IRIE G. Travel route recommendation using geotags in photo sharing sites [C]// ACM International Conference on Information and Knowledge Management (CIKM). 2010:579-588.

[14] LU E H C, CHEN C Y, TSENG V S. Personalized trip recommendation with multiple constraints by mining user check-in behaviors [C]// Proceedings of the 20th International Conference on Advances in Geographic Information Systems (SIGSPATIAL). 2012:209-218.

[15] WANG H N, LI G L, FENG J H. Group-Based Personalized Location Recommendation on Scenic Sights[C]// Proc. Int. Conf. on APWeb. 2014:68-80

[16] HE W, LI D Y, ZHANG T L, et al. Mining regular routes from gps data for ridesharing recommendations [C]// Proceedings of the ACM SIGKDD International Workshop on Urban Computing(UrbComp'12). New York, NY, USA, 2012:79-86.

[17] LIN Y R, SUN J M, CASTRO P, et al. Metafac: community discovery via relational hypergraph factorization [C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2009:527-536.

[18] MACQUEE Q J. Some Methods for Classification and Analysis of Multivariate Observation [C]// Proceeding 5th Berkley Symposium on Mathematical Statistics and Probability. 1967: 281-297.

[19] SHI J, MALIK J. Normalized Cuts and Image Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8):888-905.

[20] ROUSSEEUW P J. Rousseeuw. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis [J]. Journal of Computational and Applied Mathematics, 1987, 20(20): 53-65.

(上接第 198 页)

问题。引入处理大数据的思想(如 Map Reduce 算法等)来处理高维特征，是处理具有高维特征区间数样本的有效途径之一。

参 考 文 献

[1] MOORE R E. Interval arithmetic and automatic error analysis in digital computing[D]. Palo Alto: Stanford University, 1962.

[2] FILIPPONE M, MASULLI F, ROVETTA S. Applying the possibilistic c-means algorithm in kernel induced spaces[J]. IEEE Transcations on Fuzzy Systems, 2010, 18(3):572-584.

[3] REN S J, LV J H. Genetic algorithm based kernel function FCM clustering algorithm for interval numbers[J]. Journal of System Engineering, 2008, 23(5):611-616. (in Chinese)
任世锦，吕俊怀. 基于遗传算法的区间数核模糊聚类算法[J]. 系统工程学报，2008，23(5)：611-616.

[4] PIMENTEL B, COSTA A, SOUZA R. Kernel-based fuzzy clustering of interval data[C]// Proceedings of 2011 IEEE International Conference on Fuzzy Systems. Taipei, 2011:497-501.

[5] PIMENTEL B, COSTA A, SOUZA R. Input space versus feature space in kernel-based interval fuzzy C-Means clustering[C]// Proceedings of 2015 International Joint Conference on Neural Networks. 2015:1-7.

[6] VAPNIK V N. The Nature of Statistical Learning Theory[M]. London: Springer, 2000.

[7] TAX D M J, DUNI R P W. Support vector domain description [J]. Pattern Recognition Letters, 1999, 20(11):1191-1199.

[8] SCHOELKOPF B, SMOLA A J. Learning with kernels: support vector machines, regularization, optimization, and beyond[M]. Cambridge, Massachusetts: The MIT Press, 2002.

[9] CAMPBELL C, BENNETT K P. A linear programming approach to novelty detection[C] // Proc of the Conference on Neural Information Processing Systems: Natural and Synthetic. Vancouver, Canada, 2001:395-401.

[10] UTKIN L V, CHEKH A I. A new robust model of one-class classification by interval-valued training data using the triangular kernel[J]. Neural Networks, 2015, 69:99-110.

[11] CARVALHO F, SOUZA R, BEZERRA L. A dynamical clustering method for symbolic interval data based on a single adaptive Euclidean distance[C]// Proc of the Ninth Brazilian Symposium on Neural Networks (SBRN'06). 2006.

[12] KRISHNAPURAM R, KELLER J M. A possibilistic approach to clustering[J]. IEEE Transcations on Fuzzy Systems, 1993, 1(2):98-110.

[13] ANDERSON D T, BEZDEK J C, POPESCU M, et al. Comparing fuzzy, probabilistic, and possibilistic partitions[J]. IEEE Transcations on Fuzzy Systems, 2010, 18(5):906-918.

[14] CHEN B. Research on Outlier Detection Method and Its Key Techniques[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2013. (in Chinese)
陈斌. 异常检测方法及其关键技术研究[D]. 南京：南京航空航天大学，2013.

[15] HEIJDEN F, DUIN R, RIDDER D, et al. Classification, parameter estimation and state estimation-an engineering approach using Matlab[M]. Wiley, 2004.

[16] LICHMAN M. UCI Machine Learning Repository[DB/OL]. http://archive.ics.uci.edu/ml.