

# 基于单簇聚类的数据描述

陈 斌<sup>1),2)</sup> 冯爱民<sup>1)</sup> 陈松灿<sup>1)</sup> 李 斌<sup>2),3)</sup>

<sup>1)</sup>(南京航空航天大学信息科学与技术学院 南京 210016)

<sup>2)</sup>(扬州大学信息工程学院 江苏 扬州 225009)

<sup>3)</sup>(南京大学计算机软件新技术国家重点实验室 南京 210093)

**摘 要** 文中提出了一种基于单簇可能性 C-均值聚类(Possibilistic C-Means, PCM)的数据描述方法并用于单分类. 训练时,其首先进行 P1M(PCM, C 值取 1)聚类,得到所有训练样本对目标类的隶属度;然后设置隶属度阈值,形成相应的数据描述进行单分类. 分类时,计算新样本对目标类的隶属度,若其隶属度小于该阈值则判为异常,否则为正常. 该方法和当前流行的支持向量域数据描述方法以及 Parzen 方法窗具有类似的参数配置和相当的分类性能,由此提供了另一种单分类学习算法. 值得指出的是,尽管是 PCM 的一个特例,但 P1M 拥有 PCM 一般不具备的全局最优特性,而该特性对解决实际问题十分重要.

**关键词** 数据描述;聚类;单簇;可能性 C-均值;单分类

中图法分类号 TP181

## One-Cluster Clustering Based Data Description

CHEN Bin<sup>1),2)</sup> FENG Ai-Min<sup>1)</sup> CHEN Song-Can<sup>1)</sup> LI Bin<sup>2),3)</sup>

<sup>1)</sup>(College of Information Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 210016)

<sup>2)</sup>(Information Engineering College, Yangzhou University, Yangzhou, Jiangsu 225009)

<sup>3)</sup>(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093)

**Abstract** In this paper, a one-cluster clustering based data description method (OCCDD) is proposed for one-class classification. It operates as follows: when training, one-cluster Possibilistic C-Means (P1M) algorithm is firstly performed on the training target samples, then the memberships to the target class of all samples are obtained, a threshold of memberships is set to form the data description. When testing, the memberships of the samples for testing are computed, the samples with less membership than the threshold are thought as the outliers, otherwise as the target objects. The proposed method has the same parameter configuration as the prevalent methods: Support Vector Data Description (SVDD) and Parzen-window method, and leads to an alternative one-class classifier. It is worthy to point out that: although as a special example of traditional PCM algorithm, P1M can obtain a globally optimal solution while traditional PCM generally could not. Moreover, the globally optimal property is of great importance for the practical implementation.

**Keywords** data description; clustering; one-cluster; PCM; one-class classification

收稿日期:2007-03-05;修改稿收到日期:2007-05-23. 本课题得到国家自然科学基金(60603029)、江苏省自然科学基金(BK2004052, BK2007074)和江苏省高校自然科学基金(06KJB520132)资助. 陈 斌,男,1974 年生,博士研究生,讲师,研究方向为模式识别和数据分析. E-mail: b.chen@nuaa.edu.cn. 冯爱民,女,1971 年生,博士研究生,讲师,研究方向为人工智能和模式识别. 陈松灿,男,1962 年生,教授,博士生导师,研究领域为模式识别、生物信息工程、医学图像处理、神经网络. E-mail: s.chen@nuaa.edu.cn. 李 斌,男,1965 年生,教授,研究领域为软件理论与工程以及软件工程中的机器学习.

## 1 引言

单分类器研究作为模式分类的一个重要分支,广泛应用于故障分析诊断、疾病检测、入侵检测以及身份辨识等领域<sup>[1]</sup>.和两类分类问题不同,单分类器在训练时只有一类样本即正常类样本可资利用,而反类(异常类)样本由于很少发生或者其获取代价过高,因而不能获得充分采样,所以单分类器只能从已获取的充分采样的目标类样本中进行学习并形成数据域描述进行单分类.例如电厂的故障诊断,不能因为要获得一个故障数据而故意破坏设备<sup>[2]</sup>,所以只能从正常运行数据中进行学习,形成数据域描述用于故障诊断.

单分类器的设计方法主要可分为基于密度的方法<sup>[3]</sup>、基于支撑域的方法<sup>[2,4]</sup>以及基于聚类的方法等<sup>[2,5]</sup>.基于密度的单分类器设计可以采用高斯模型<sup>[6]</sup>或者混合高斯模型<sup>[6]</sup>等参数化的方法对概率密度进行建模,也可以采用 Parzen 窗进行非参数的概率密度建模<sup>[3,7]</sup>,然后根据经验风险设置相应的概率密度为阈值,分类时将所有概率密度低于该阈值的测试样本判为异常<sup>[3,7]</sup>.基于支撑域的单分类方法<sup>[2]</sup>一般都是最小化目标数据的支撑域,得到的数据描述往往表示为目标类的支撑域的几何形状.其中最典型的就是单类支持向量机(One-Class Support Vector Machine, OCSVM)<sup>[8-9]</sup>和支持向量数据域描述(Support Vector Data Description, SVDD)<sup>[4-5]</sup>.OCSVM 训练时寻找一个由支持向量表示的超平面,并最大化超平面和原点(视为反例)之间的间隔;SVDD 则寻求一个包含所有正常类训练样本的最小超球<sup>[10,4]</sup>,这两者在一定的条件下等价<sup>[8,11]</sup>,但他们的求解最终都归结为二次规划问题,计算复杂性高.基于聚类的 K-均值<sup>[12]</sup>(K 值取大于 1)和 K-中心<sup>[13]</sup>(K 值取大于 1)单分类器开拓了单分类器的设计思路,它们都是通过“两步走”的方法设计单分类器. K-均值单分类器首先进行 K-均值聚类,然后将所有训练样本到最近簇类中心的平均距离作为阈值;K-中心单分类器则首先进行 K-中心聚类,然后选取所有训练样本到其最近簇类中心的最大距离作为阈值.但是由于 K-均值以及 K-中心算法的自身缺陷,对簇类中心的选择非常敏感,一般只能得到局部最优解,并且其中的 K 值的选取还是一个悬而未决的问题<sup>[5-6]</sup>.

本文针对 K-均值<sup>[12]</sup>和 K-中心<sup>[13]</sup>单分类器所

存在的问题,提出了一个基于单簇聚类的数据描述(One-Cluster Clustering based Data Description, OCCDD)方法用于单分类.这种方法同样遵循了“两步走”的方法:首先采用单簇的可能性 C 均值<sup>[14]</sup>(Possibilistic C-Means, PCM)即 P1M(PCM, C 值取 1)算法进行聚类,然后设置样本相对于目标类的隶属度阈值从而形成数据描述.易引起疑问的是当目标类具有多个子簇时, P1M 仍可能有效地保证来自于著名的 Cover 定理<sup>[15]</sup>:多模态类可以通过核化近乎以概率 1 转化为单模态类.尽管采用了传统的 PCM 算法<sup>[14]</sup>,但值得强调的是当簇数为 1 时, P1M 能够获得全局最优解,而传统的 PCM(C 值大于 1)一般则不能保证.因此这种方法避免了 K-中心和 K-均值单分类器中 K 值的选取、初始化依赖以及局部最优解等问题.并且从优化目标上比较, SVDD 最小化样本到圆心的最大距离, OCCDD 则最小化所有样本到圆心的加权平均距离,因此这种方法比 SVDD 对野值更为鲁棒.在部分 UCI Benchmarks 数据集<sup>①</sup>和 USPS 手写体数字上的实验结果表明: OCCDD 具有和 Parzen 窗单分类器<sup>[3,7]</sup>以及当前流行的 SVDD 相当的分类性能;训练时 OCCDD 和 Parzen 窗计算效率比 SVDD 高,而在测试时由于利用了所有训练样本进行计算, OCCDD 和 Parzen 窗效率均不如 SVDD.尽管存在一定的不足,但通过充分挖掘传统 PCM 聚类算法在单分类算法设计中的潜力,不失为是一种新的设计思路.

本文第 2 节简单回顾流行的单分类器 SVDD;第 3 节提出核化后的 P1M 算法及其全局最优性证明,基于 P1M 聚类的单分类器算法 OCCDD 也在第 3 节给出;第 4 节给出在 UCI Benchmarks 数据集和 USPS 手写体数字数据集上的实验以及相关分析;第 5 节总结本文的工作并提出下一步的工作构想.

## 2 支持向量数据域描述(Support Vector Data Description, SVDD)

支持向量数据域描述<sup>[4-5]</sup>寻找一个包含所有的正常类样本的最小超球.如给定训练数据集  $D = \{x_i | x_i \in \mathcal{R}^d\}_{i=1}^n$ , 通过一个非线性映射  $\phi: \mathcal{R}^d \mapsto F$  将数据映射到一个高维(甚至无穷维)的特征空间  $F$  中.为了避免训练集中掺杂异常样本带来的危害,训练

① Blake C, Merz C. UCI repository of machine learning databases, 1998. <http://www.ics.uci.edu/~mllearn/MLRepository.html>

样本到球心的平方距离不应严格小于  $R^2$ , 但应对大于  $R^2$  的平方距离进行惩罚, 因此引入松持变量  $\xi_i$  松弛所有训练样本都应落在球内的约束而允许部分训练样本落在球外, 软间隔的 SVDD 通过优化下面的表达式来发现包含大多数目标类训练样本的超球<sup>[4-5]</sup> (半径为  $R$ ):

$$\min R^2 + C \sum_{i=1}^n \xi_i \tag{1}$$
$$\text{s. t. } \|\phi(x_i) - a\|^2 \leq R^2 + \xi_i, \xi_i \geq 0$$

这里  $a$  表示超球的中心,  $C$  是正则化因子. 引入非负 Lagrange 乘子构建 Lagrange 函数可得

$$L = R^2 + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \mu_i \xi_i - \sum_{i=1}^n \alpha_i (R^2 + \xi_i - \|\phi(x_i) - a\|^2) \tag{2}$$

令  $L$  对  $R, a$  和  $\xi_i$  的偏导数分别为 0, 得到

$$a = \sum_{i=1}^n \alpha_i \phi(x_i), \alpha_i = C - \mu_i, \sum_i \alpha_i = 1 \tag{3}$$

将对偶表示式(3)代入式(2), 并将其中的内积代之以核函数, 可得如下对偶问题

$$\max_{\alpha} \sum_{i=1}^n \alpha_i K(x_i, x_i) - \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j)$$
$$\text{s. t. } \sum_{i=1}^n \alpha_i = 1, \alpha_i \geq 0 \tag{4}$$

由 Karush-Kuhn-Tucker 条件可知, 对应  $\alpha_i$  为 0 的样本在超球内, 对应  $0 < \alpha_i < C$  的样本在超球面上, 故而称作(非边界)支持向量, 其余对应  $\alpha_i = C$  的那部分样本在超球外, 被称作边界支持向量. 通过二次规划求解求得所有样本对应的  $\alpha_i$ , 则球心可以由非零  $\alpha_i$  的对应样本稀疏表示, 而半径可由任意一个支持向量到球心的距离获得.

### 3 基于单簇聚类的数据描述(One-Cluster Clustering based Data Description, OCCDD)

#### 3.1 可能性 1-均值(P1M)聚类算法描述

模糊 C-均值(Fuzzy C-Means, FCM)<sup>[16]</sup> 算法通过放松 K-均值<sup>[6]</sup> 聚类算法中每个样本只能完全属于某一类的约束, 认为每个样本以某种程度模糊隶属于一类, 并且每个样本对所有类的隶属度之和为 1. 当数据集中存在野值时, 该约束使得 FCM 中的隶属度无法真正代表样本隶属于某类的概率或可

能性<sup>[14]</sup>, 导致野值对各类的等隶属度, 因而使 FCM 对野值不鲁棒. 针对 FCM 存在的问题, Krishnapuram 等人通过放宽 FCM 的概率约束, 提出了 PCM 算法<sup>[14]</sup>, 使得隶属度真正代表样本隶属于某一类的可能性, 进而提高了对噪声和野值的鲁棒性. 为鲁棒起见, 本文采用单簇 PCM 算法, 简记为 P1M 算法. 尽管作为传统的 PCM 算法的特例, P1M 算法显示出与 PCM 一般不具有的全局最优特性以及类似 SVDD 性能的数据描述能力. 考虑到核型 P1M 算法的描述能力, 本文在下面仅给出 P1M 算法刻画.

引入核型 P1M 的动机在于: 若给定的数据集  $D = \{x_i | x_i \in \mathcal{R}^d\}_{i=1}^n$  是一个非线性复杂问题, 如多子簇单类问题, 则由 Cover 定理<sup>[15]</sup> 所知, 可通过一个非线性映射  $\phi: \mathcal{R}^d \mapsto F$  将数据映射到一个高维(甚至无穷维)的特征空间  $F$  中可以使非线性问题更有可能转化为易处理的简单问题. 相应的核型 P1M 优化目标定义为

$$\min J(U, V) = \sum_{j=1}^n u_j^m d^2(\phi(x_j), v) + \eta \sum_{j=1}^n (1 - u_j)^m$$
$$\text{s. t. } 0 \leq u_j \leq 1, \tag{5}$$

其中  $\eta$  是一个正的正则化因子,  $u_j$  是第  $j$  个样本对目标类的隶属度,  $v$  是训练集中所有样本的加权平均中心,  $d(\phi(x_j), v)$  是  $\phi(x_j)$  和  $v$  之间的欧氏距离, 参数  $m \in (1, +\infty)$  是一个模糊因子<sup>[6]</sup>, 控制在半径为  $\sqrt{\eta}$  的球内外样本隶属度差异的显著程度. 其中式(5)中优化目标的第二项为正则化项, 其作用是避免使所有的隶属度收敛为 0 而产生平凡解. 比较 P1M 目标函数的第一项和 SVDD 中优化目标的式(1), SVDD 是最小化样本到圆心的最大距离, 而 P1M 则最小化所有样本到圆心的加权平均距离, 并且样本权重与其到圆心的距离成反比, 由此可使 P1M 比 SVDD 鲁棒. 求解该问题得到

$$u_j = \frac{1}{1 + \left( \frac{d^2(\phi(x_j), v)}{\eta} \right)^{1/(m-1)}}, v = \frac{\sum_{j=1}^n u_j^m \phi(x_j)}{\sum_{j=1}^n u_j^m},$$
$$\eta = \frac{\sum_{j=1}^n u_j^m d^2(\phi(x_j), v)}{\sum_{j=1}^n u_j^m} \tag{6}$$

令  $u'_j = u_j^m / \sum_{l=1}^n u_l^m$  和  $\mathbf{u} = (u'_1, u'_2, \dots, u'_n)^\top$  并以核函数

代替内积使  $d^2(\phi(x_j), v) = K(x_j, x_j) - 2 \sum_{i=1}^n K(x_j, x_i) u_i' + \mathbf{u}^T \mathbf{K} \mathbf{u}$ , 这里  $\mathbf{K}$  表示核 Gram 矩阵. 由于 Gaussian 核函数  $K(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{\sigma^2}\right)$  ( $\sigma$  表示核带宽) 诱导的距离度量是鲁棒的<sup>[17]</sup>, 故 SVDD 和 OCCDD 中本文均采用 Gaussian 核函数. 此后 P1M 算法就可按传统 PCM 的交替优化方法进行迭代优化, 获得所有样本对目标类的隶属度和  $\eta$ , 具体算法参看文献[14].

3.2 全局最优性分析

传统 PCM(C 值大于 1) 算法由于对隶属度和簇类中心进行交替优化, 已证明其只能获得一个局部最优解<sup>[18]</sup>, 而作为特例的 P1M 算法, 由于簇数仅为 1, 保证了优化目标对隶属度和簇类中心的 Hessian 矩阵的正定性, 进而保证了优化目标问题的全局最优性.

定理 1. 单簇 P1M 是全局最优算法.

证明. 首先计算 P1M 的目标函数  $J(U, V)$  对所有样本隶属度  $\mathbf{u} = (u_1, u_2, \dots, u_n)^T$  和簇类中心  $\mathbf{v} = (v_1, v_2, \dots, v_{d'})^T$  ( $d'$  是特征空间的维数) 的 Hessian 矩阵

$$\mathbf{H}_{J(U,V)} = \begin{bmatrix} \frac{\partial^2 J}{\partial u_1^2} & \cdots & \frac{\partial^2 J}{\partial u_1 \partial u_n} & \frac{\partial^2 J}{\partial u_1 \partial v_1} & \cdots & \frac{\partial^2 J}{\partial u_1 \partial v_{d'}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 J}{\partial u_n \partial u_1} & \cdots & \frac{\partial^2 J}{\partial^2 u_n} & \frac{\partial^2 J}{\partial u_n \partial v_1} & \cdots & \frac{\partial^2 J}{\partial u_n \partial v_{d'}} \\ \frac{\partial^2 J}{\partial v_1 \partial u_1} & \cdots & \frac{\partial^2 J}{\partial v_1 \partial u_n} & \frac{\partial^2 J}{\partial^2 v_1} & \cdots & \frac{\partial^2 J}{\partial v_1 \partial v_{d'}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 J}{\partial v_{d'} \partial u_1} & \cdots & \frac{\partial^2 J}{\partial v_{d'} \partial u_n} & \frac{\partial^2 J}{\partial v_{d'} \partial v_1} & \cdots & \frac{\partial^2 J}{\partial^2 v_{d'}} \end{bmatrix}$$

(7)

$J(U, V)$  对样本隶属度和簇类中心的二阶偏导数分别是

$$\frac{\partial^2 J}{\partial u_i \partial u_j} = \begin{cases} m(m-1) [u_j^{m-2} d^2(\phi(x_j), v) + \eta(1-u_j)^{m-2}], & i=j \\ 0, & i \neq j \end{cases}$$

(8)

$$\frac{\partial^2 J}{\partial v_k \partial u_j} = \frac{\partial^2 J}{\partial u_j \partial v_k} = -2mu_j^{m-1} (\phi(x_j)^k - v^k)$$

(9)

$$\frac{\partial^2 J}{\partial v_k \partial v_l} = \begin{cases} 2 \sum_{j=1}^n u_j^m, & k=l \\ 0, & k \neq l \end{cases}$$

(10)

其中  $\phi(x_j)^k$  表示  $\phi(x_j)$  的第  $k$  维坐标.

由于  $m > 1$  和  $1 \geq u_j \geq 0$ , 且参数  $\eta$  大于 0, 很明显  $m(m-1) [u_j^{m-2} d^2(\phi(x_j), v) + \eta(1-u_j)^{m-2}] \geq 0, 2 \sum_{j=1}^n u_j^m > 0$ , 因此 Hessian 矩阵是对称非负定的, 当且仅当簇类中心  $v$  等于某一个样本点  $x_i$  在特征

$$\text{空间中映射 } \phi(x_i) \text{ 时, 即 } v = \frac{\sum_{j=1}^n u_j^m \phi(x_j)}{\sum_{j=1}^n u_j^m} = \phi(x_i)$$

Hessian 矩阵的行列式的值为零, 这要求除第  $i$  个样本隶属度为 1 外其他所有样本的隶属度均为零, 这种情况除了整个训练集只有单个样本之外根本不可能存在, 因此完全可以说 Hessian 矩阵是正定的, 也就是说 P1M 算法可以获得全局最优解. 证毕.

3.3 基于单簇聚类的数据描述

3.3.1 数据描述的获得

OCCDD 首先在训练集上执行 P1M 聚类, P1M 收敛后得到所有样本对目标类的隶属度, 然后设置一个隶属度的阈值  $\theta_u$ . 分类时, 将那些隶属度低于阈值  $\theta_u$  的样本判为异常类, 否则判为正常类. 将那些隶属度等于阈值的样本作为边界点, 其到簇类中心的距离作为数据描述域的半径  $R$

$$R = \sqrt{\eta \left( \frac{1 - \theta_u}{\theta_u} \right)^{(m-1)}}$$

(11)

数据描述中心是所有训练样本的加权平均中心  $v$ , 尽管其在特征空间中一般不能显式地表示, 但可以通过核代入计算获得新样本到  $v$  的欧氏距离. 进行异常检测或者新颖性检测时, 可以根据其是否落在支撑域中按下式进行判别

$$I(d(\phi(x_j), v) > R) = \begin{cases} 1, & \text{则 } x_j \text{ 是离群点} \\ 0, & \text{则 } x_j \text{ 是目标项} \end{cases}$$

(12)

其中函数  $I(\cdot)$  是一个指示函数, 当输入为真则返回 1, 否则返回 0. 这和根据隶属度进行判别是等价的, 所以也可以根据隶属度来进行判别, 判别函数为

$$I(u(x_j) < \theta_u) = \begin{cases} 1, & \text{则 } x_j \text{ 是离群点} \\ 0, & \text{则 } x_j \text{ 是目标项} \end{cases}$$

(13)

由于采用隶属度阈值直觉上更容易处理, 所以下文中都采用式(13)进行判别.

3.3.2 隶属度阈值的设置

在 SVDD 中通过正则化参数  $C = 1/(n \times fracj)$  来调控经验风险的大小, 其中  $fracj$  同时也控制着边界支持向量个数的下界以及支持向量个数的上

界<sup>[9]</sup>,而 OCCDD 中同样可通过正常样本的拒绝比  $fracj$  来控制经验风险. 为此 OCCDD 首先对所有的样本隶属度进行排序, 然后根据拒绝比选择第  $(n \times fracj) + 1$  个样本的隶属度作为阈值  $\theta_n$ .

3.3.3 参数优化

由于 OCCDD 中引入了正常样本的拒绝比以及核函数, 必然引入了相应的一些参数, 选用高斯核函数时引入了核带宽. 对于 P1M 优化函数中的模糊因子  $m$ , 其控制着数据描述在  $\sqrt{\eta}$  这个半径内外样本隶属度差异的显著性, 对实际的单分类的意义不大, 所以固定为 2. 核带宽则控制着数据描述边界的光滑

性, 带宽越大, 描述边界越光滑, 带宽越小边界就越粗糙, 但过粗糙会导致过拟合, 而过光滑又会导致欠拟合. 由此核带宽不仅控制着推广能力的优劣, 而且也能实现对多子簇单类数据的描述.

图 1 展现了人工香蕉形数据集<sup>[11]</sup>上参数的作用, 从中可以看出  $\sigma$  越大数据描述边界越光滑;  $fracj$  越大, 会有越多的正常样本落在数据描述边界之外. 在实际应用中对  $fracj$  和核带宽  $\sigma$  两个参数, 采用  $k$  重交叉验证选择参数来实现经验风险和推广能力的折中. 图 1 显示了核带宽对描述边界光滑性的调节作用和对多子簇单类数据的描述能力.

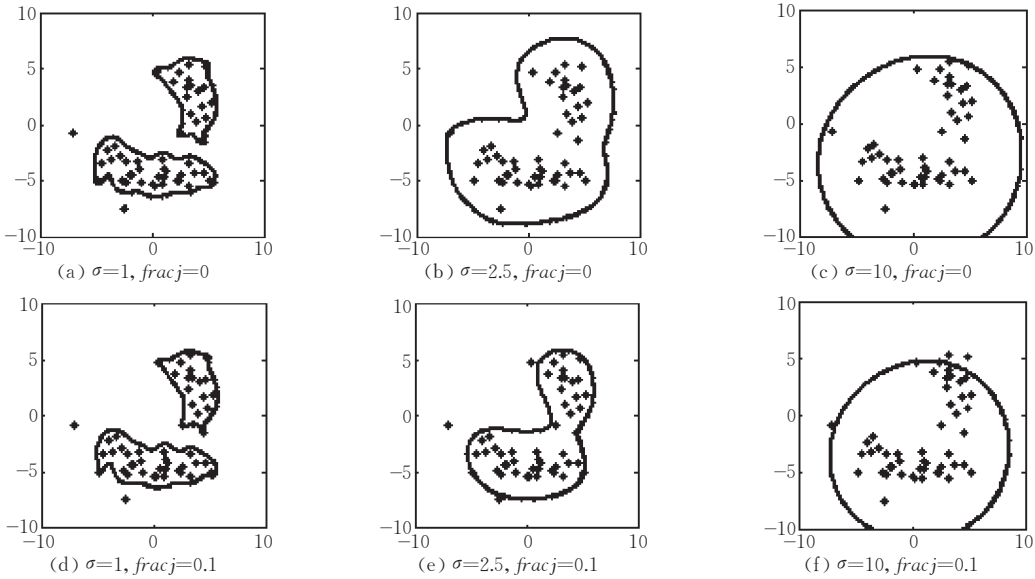


图 1 香蕉型数据集上 OCCDD 的参数  $fracj$  和核带宽  $\sigma$  的作用

3.4 和 SVDD 的比较与分析

下面从计算复杂性、鲁棒性以及稀疏性等方面分析比较 SVDD 和 OCCDD. 首先从稀疏性上比较, SVDD 中超球半径和中心都可以由支持向量来表示, 而支持向量一般仅是训练样本的一部分, 所以 SVDD 在分类时仅需要这些支持向量. 但这种稀疏性并不总是存在, 最坏情况下每个样本都可能是支持向量而导致过拟合. 因 OCCDD 簇类中心需要所有样本计算加权平均中心, 从而丧失了稀疏性, 但直觉上簇类中心的表示并不需要所有训练样本, 因此进一步研究簇类中心表示的稀疏性对提高 OCCDD 的测试效率很有意义.

其次比较两者对野值的鲁棒性, 如果训练样本中存在野值, 则 SVDD 的超球中心必然向该野值倾斜, 超球半径也因此会扩大, 从而导致数据描述包含更多的异常类样本, 降低其推广能力. 而在 OCCDD 中, 由于采用了可能性方法, 根据隶属度的计算公

式, 野值的隶属度必然很小, 其对簇类中心进而对数据描述域半径的决定所起的作用就小. 因而 OCCDD 比 SVDD 更为鲁棒.

最后从计算复杂性而言, SVDD 需要求解二次规划问题, 所以其训练复杂性是  $O(n^3)$ <sup>[19]</sup>; 而 OCCDD 的训练复杂性依赖于 P1M 的复杂性, P1M 算法的复杂性是  $O(nc^2dt)$ <sup>[20]</sup>, 其中  $n$  是样本个数,  $c$  是簇类个数,  $d$  是样本的维数,  $t$  是迭代次数, 则 P1M 的训练复杂度是  $O(ndt)$ , 因而训练效率比 SVDD 高. 但测试时, 因 SVDD 的稀疏性, 效率反而高于 OCCDD.

4 实验与评估

4.1 小规模数据集上的实验

首先在部分 UCI 机器学习数据库中的数据集合 Iris, Heart, Diabetes, Glass 等以及 Stalog 上的

Vehicle 数据集 (<http://ict.ewi.tudelft.nl/davidt/occ/index.html>) 上进行 OCCDD 与 SVDD 以及 Parzen 窗方法单分类性能的比较. 选取其中一类为目标类, 剩余部分为异常类. 所有实验都在同一台计算机 (Xeon 2.8GHz, 1GB RAM) 上运行.

在单分类实验中, 分类的可能结果如表 1 所示.

表 1 单分类的可能结果汇总表

真实类别	分类为正常类	分类为异常类
正常类	True Positive(TP)	False Negative(FN)
异常类	False Positive(FP)	True Negative(TN)

单分类的评价指标<sup>[21]</sup>主要包括: 查准率  $Precision = TP / (TP + FP)$ ; 查全率  $Recall = TP / (TP + FN)$ ; 准确率  $Accuracy = (TP + TN) / (TP + FP +$

$TN + FN)$ ; 以及  $F-Score = 1 / (\alpha / Precision + (1 - \alpha) / Recall)$ ,  $\alpha$  是平衡因子. 由于数据集的不平衡, 最大化准确率将导致算法过分偏向于非目标类样本的正确分类结果, 产生过拟合. 故本文采用  $F-Score$  度量并设置  $\alpha = 1/2$ , 实现查全率和查准率之间的平衡.

表 2 中所有的测试结果都是通过 10 重交叉验证选择参数并使  $F-Score$  最大后的平均值. 粗体字表示性能最优. 从表中可发现 OCCDD 分类性能尽管占优最少, 但和 SVDD 以及 Parzen 窗方法的性能互有上下, 因此可以说它们分类性能相当. 从计算时间上比较, 在训练样本不多的情况下 SVDD 因训练时间比较长, 而导致其耗时比 Parzen 窗方法和 OCCDD 高得多.

表 2 Parzen 窗、SVDD 和 OCCDD 的单分类性能比较

数据集(维)	目标类	训练集	测试集	F-Score / %			耗时 / s		
				Parzen 窗	SVDD	OCCDD	Parzen 窗	SVDD	OCCDD
Iris(4)	Setosa	45	105	94.7	95.8	<b>96.9</b>	0.56	1.56	<b>0.23</b>
	Vesicular	45	105	88.0	89.0	<b>91.3</b>	0.56	0.77	<b>0.17</b>
	Virginica	45	105	85.6	<b>90.0</b>	82.6	0.56	2.05	<b>0.17</b>
Breast(9)	Malignant	413	289	95.1	<b>96.4</b>	95.1	<b>2.58</b>	613.2	4.11
	Benign	217	482	93.5	<b>94.4</b>	93.7	1.547	40.61	<b>1.11</b>
Sonar(60)	Mines	110	108	64.7	<b>65.7</b>	65.5	0.75	1.17	<b>0.44</b>
	Rocks	89	120	65.5	65.3	<b>65.9</b>	0.77	0.98	<b>0.36</b>
Diabetes(8)	Present	450	318	<b>71.0</b>	68.7	67.2	<b>4.25</b>	14.88	4.31
	Absent	242	526	63.3	65.7	63.4	2.34	4.27	<b>1.72</b>
Glass(9)	Building float	63	151	74.4	<b>78.1</b>	74.5	0.58	0.63	<b>0.20</b>
	Building nonfloat	69	145	<b>67.0</b>	66.3	63.8	0.58	0.36	<b>0.25</b>
	Vehicle float	16	198	68.6	<b>74.5</b>	66.7	0.53	0.24	<b>0.17</b>
	Containers	12	202	<b>75.7</b>	58.8	66.7	0.50	0.26	<b>0.16</b>
	Headlamps	27	187	<b>85.0</b>	66.6	67.2	0.58	0.41	<b>0.19</b>
Vehicle(18)	Van	180	666	<b>83.3</b>	81.1	75.3	1.95	20.81	<b>1.19</b>
	Saab	196	640	68.2	<b>68.3</b>	66.2	2.03	2.49	<b>1.23</b>
	Bus	197	649	73.4	<b>77.9</b>	62.7	1.92	33.97	<b>1.45</b>
	Opel	191	655	66.8	66.6	65.4	1.92	2.23	<b>1.25</b>

4.2 USPS 手写数字识别

为了进一步比较三种算法的分类性能, 在著名的美国邮政编码手写数字数据集上进行了测试. USPS 手写体数字数据集 (<http://www.cs.toronto.edu/roweis/data.html>) 包含所有 0~9 数字的 16×16 的灰度图像, 共计 11000 个样本, 每类样本各 1100 个, 每个样本表示为一个 256 维的向量 (16×16 矩阵), 每个像素的灰度值在 0~255 范围之内, 预处理时把所有的像素灰度值除以 255 以使灰度值都落入 0~1 范围之内.

首先选择某类数字为目标类, 其余数字为异常类. 例如对目标类“0”, 三种算法均选取 90% 的数字

“0”进行训练集, 其余部分为测试集. 为了实现在正常命中率  $TP$  和伪命中率  $FP$  的折中, 采用 10 重交叉验证进行参数选择.

图 2 展示了在固定  $frac{j}$  为 0.1 的前提下使用 10 重交叉验证对核带宽进行调控所得的 SVDD、Parzen 窗和 OCCDD 的平均性能 ROC 曲线<sup>[5]</sup>, 其中实线表示 SVDD, 点划线表示 Parzen 窗方法, 虚线表示 OCCDD. 通过 ROC 曲线下面积 (AUC) 的简单比较, 可以发现 SVDD 的性能比 OCCDD 要稍好, 除在数字 1 上相差稍多外, 其余数字上尽管有一定的差距, 但差别都较小; 和 Parzen 窗单分类性能相比, OCCDD 基本相当并稍稍占优.

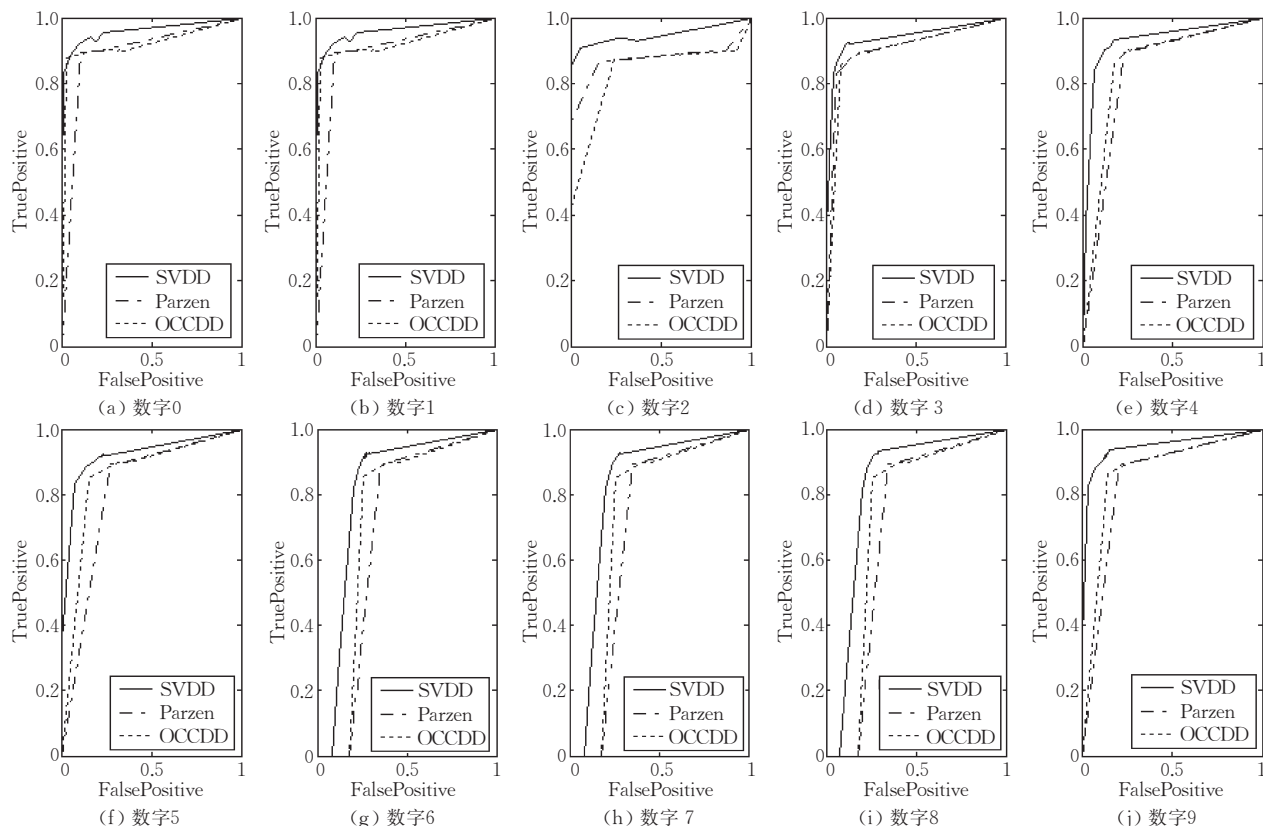


图 2 USPS 手写体数据集上 OCCDD 和 SVDD 以及 Parzen 窗方法的 ROC 性能曲线比较

## 5 结束语

本文提出了一种基于单簇聚类的数据描述方法应用于单分类,该方法具有和 SVDD 以及 Parzen 窗方法相当的性能,但其训练效率更优.然而由于该方法中簇类中心的非稀疏化表示降低了其测试效率,因此进一步研究簇类中心的稀疏化表示对提高 OCCDD 的测试效率非常重要.值得指出的是,尽管采用了传统的 PCM 算法,但是该方法展现了传统 PCM 一般不具有的全局最优特性.该方法通过充分挖掘传统 PCM 算法的潜能,提供了另一种新的基于聚类的单分类器,进而拓展了单分类算法的设计思路供人借鉴并推广应用到其他的聚类算法中.

## 参 考 文 献

- [1] Markos M, Sameer S. Novelty detection: A review-part I: Statistical approaches. *Signal Processing*, 2003, 83(12): 2481-2497
- [2] Juszczak P. Learning to recognize: A study on one-class classification and active learning[Ph. D. dissertation]. Delft University of Technology, Holland, 2006
- [3] Bishop C. Novelty detection and neural network validation.

IEE Proceedings of Vision, Image and Signal Processing, 1994, 141(4): 217-222

- [4] Tax D, Duin R. Support vector domain description. *Pattern Recognition Letters*, 1999, 20(11-13): 1191-1199
- [5] Tax D. One-class classification; Concept-learning in the absence of counter-examples[Ph. D. dissertation]. Delft University of Technology, Holland, 2001
- [6] Duda R, Hart P, Stork D. *Pattern Classification* (2nd Edition). New York, USA: John Wiley & Sons, 2001
- [7] Yeung D-Y, Chow C. Parzen-Window networks intrusion detectors//*Proceedings of the 16th International Conference on Pattern Recognition*. Quebec, Canada, 2002, 4: 385-388
- [8] Schölkopf B, Platt J, Shawe-Taylor J, Smola A, Williamson R. Estimating the support of high-dimensional distribution. *Neural Computation*, 2001, 13: 1443-1471
- [9] Schölkopf B, Williamson R, Smola A, Shawe-Taylor J, Platt J. Support vector method for novelty detection//*Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2000, 12: 582-588
- [10] Schölkopf B, Burges C, Vapnik V. Extracting support data for a given task//*Proceedings of the 1st International Conference on Knowledge Discovery & Data Mining*. Montreal, CA. Menlo Park, CA: AAAI Press, 1995: 252-257
- [11] Tax D, Duin R. Support vector data description. *Machine Learning*, 2004, 54: 45-66
- [12] Jiang M-F, Tseng S-S, Su C-M. Two-phase clustering process for outliers detection. *Pattern Recognition Letters*, 2001, 22(6/7): 691-700

- [13] Hochbaum D, Shmoys D. A best possible heuristic for the  $k$ -center problem. *Mathematics of Operations Research*, 1985, 10(2): 180-184
- [14] Krishnapuram R, Keller J. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1993, 1(2): 98-110
- [15] Cover T. Geometrical and statistical properties of systems of linear inequalities with applications to pattern recognition. *IEEE Transactions on Electronic Computer*, 1965, 14(3): 326-334
- [16] Bezdek J. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 1981
- [17] Chen S C, Zhang D Q. Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure. *IEEE Transactions on Systems, Man and Cybernetics Part B*, 2004, 34(4): 1907-1916
- [18] Pal N, Pal K, Keller J, Bezdek J. A possibilistic fuzzy c-means clustering algorithm. *IEEE Transactions on Fuzzy Systems*, 2005, 13(4): 517-530
- [19] Chin K K. Support vector machines applied to speech pattern classification[M. S. dissertation]. University of Cambridge, Cambridge, UK, 1998
- [20] Kolen J, Hutcheson T. Reducing the time complexity of the fuzzy c-means algorithm. *IEEE Transactions on Fuzzy Systems*, 2002, 10(2): 263-267
- [21] Hand D, Mannila H, Smyth P. *Principle of Data Mining*. Cambridge, MA, USA: MIT Press, 2001



**CHEN Bin**, born in 1974, Ph. D. candidate. His research interests include pattern recognition and data exploration.

**FENG Ai-Min**, born in 1971, Ph. D. candidate. Her research interests include artificial intelligence and pattern recognition.

## Background

The project is partially supported by the National Science Foundations of China (grant No. 60603029), the Science Foundations of Jiangsu Province (grant Nos. BK2004052, BK2007074) and High School Science Foundations of Jiangsu Province (grant No. 06KJB520132). The project is to design a novel classifier for One-Class Classification and apply it to many fields, that is, handwritten digit recognition, intrusion detection and fault detection, etc. To meet such demand, this paper employs the One-Cluster Possibilistic C-Means algorithm for clustering and formulates a data description for One-Class Classification, thus proposes a new One-Class Classifier based on clustering. The proposed method has similar parameter configuration to Support Vector Data Description and Parzen windows method, and has comparable performance in some UCI Benchmarks datasets and USPS handwritten digits, but it is very efficient in training. Moreover,

it provides a new methodology of tailoring the traditional clustering algorithm to new One-Class problems, thus widens the range for designing One-Class Classifier. The project is based on the previous research on intrusion detection with Self-Organizing Map, and Grey theory. Now a prototype system for intrusion detection has been developed by their graduated members. However, the generalization performance of the intrusion detector is still weak, and the robustness is still not strong enough. Another object of this project is to design a one-class classifier with good generalization performance and robustness to outliers. The proposed method in this paper is a good trial to realize these goals. It has shown comparative performance to SVDD and Parzen windows methods, and relatively strong robustness to outliers due to the intrinsic robustness of Possibilistic 1-Means algorithm.

**CHEN Song-Can**, born in 1962, professor, Ph. D. supervisor. His main research interests include pattern recognition, bioinformatics engineering, medical image processing, and neural networks.

**LI Bin**, born in 1965, professor. His main research interests include software theory, software engineering and machine learning in software engineering.