基于相关熵和距离方差的支持向量数据描述选择性集成

邢红杰[1] 魏勇乐[2]
(河北大学数学与信息科学学院河北省机器学习与计算智能重点实验室 保定 071002)[1]
(河北大学计算机科学与技术学院 保定 071002)[2]

摘 要 提出基于信息理论学习中相关熵和距离方差的支持向量数据描述选择性集成。利用相关熵代替均方误差来度量集成的紧致性,构造出更为紧致的分类边界;利用距离方差集成度量集成中基分类器间的差异性,以提高集成模型的差异性;在目标函数中增加基于 ℓ_1 范数的正则化项,实现选择性集成。此外,利用半二次优化技术对所提选择性集成模型进行求解。与单个支持向量数据描述、基于 Bagging 的支持向量数据描述集成以及基于 AdaBoost 的支持向量数据描述集成相比,所提方法取得了更优的分类性能。

关键词 单类分类,支持向量数据描述,相关熵,选择性集成
中图法分类号 TP391.4 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.5.047

Selective Ensemble of SVDDs Based on Correntropy and Distance Variance

XING Hong-jie[1] WEI Yong-le[2]
(Key Laboratory of Machine Learning and Computation Intelligence, College of Mathematics and Information Science, Hebei University, Baoding 071002, China)[1]
(College of Computer Science and Technology, Hebei University, Baoding 071002, China)[2]

Abstract Selective ensemble of support vector data description (SVDD) based on correntropy of information theoretic learning and distance variance was proposed. Correntropy is utilized to substitute mean square error to measure the compactness of ensemble and construct more compact classification boundary. Distance variance is used to measure the diversity of base classifiers to enhance the diversity of the ensemble model. An ℓ_1 norm based regularization term is introduced into the objective function to implement the selective ensemble. Moreover, the half-quadratic optimization technique is utilized to solve the proposed selective ensemble model. In comparison with single SVDD, Bagging based ensemble of SVDDs, and AdaBoost based ensemble of SVDDs, the proposed method achieves better classification performance.

Keywords One-class classification, Support vector data description, Correntropy, Selective ensemble

1 引言

单类分类[1]是介于监督学习和无监督学习之间的机器学习任务,它能够有效地解决类别极端不平衡问题。在训练阶段,单类分类器仅需对正常数据加以训练,在测试阶段,它能够将待测样本分类为正常或异常数据。迄今为止,已涌现了大量的单类分类方法,其中最为常用的是一类支持向量机[2]和支持向量数据描述[3]。一类支持向量机(One-class Support Vector Machine, OCSVM)利用核函数将正常样本映射到高维特征空间,然后在特征空间中将正常样本的像与原点以最大间隔分开。支持向量数据描述(Support Vector Data Description, SVDD)亦利用核函数将正常样本映射到高维特征空间,然后在特征空间中寻找包含所有正常样本的像的最小包围球。当采用径向基核函数时,Tax 和 Duin[3]证明了一类支持向量机等价于支持向量数据描述。

为了使单类分类器取得更为紧致的分类边界,Tax 和 Duin[4]提出了单类分类器集成,其明显提高了单类分类器的分类性能。Seguí 等[5]提出了基于加权 Bagging 的单类分类器集成方法,所使用的基分类器为最小生成树类描述器,该方法在基准数据集上产生了优于单类分类器的性能。Zhang 等[6]首先使用局部保留映射特征提取方法对原始数据集进行维数约减,然后训练多个单类分类器,最终对各个单类分类器的输出结果加以集成。Wilk 和 Wozniak[7]提出了基于模糊组合器的单类分类器集成,该集成方法以模糊误差修正输出编码和模糊决策模板为集成策略,以基于模糊规则的分类方法为基分类器,在基准数据集上验证了所提方法的有效性。Casale 等[8]提出了近似多面体单类分类器集成方法,该方法首先利用凸壳定义目标类的分类边界,然后利用随机投影和集成决策判断样本是否属于高维空间中的凸壳模型,最后提出用于对非凸结构进行建模的平铺策略。在 200 个数据集上

到稿日期:2015-04-03 返修日期:2015-06-05 本文受国家自然科学基金项目(61473111),河北省自然科学基金项目(F2013201060),河北大学基金项目(3504020)资助。

邢红杰(1976—),男,博士,教授,主要研究方向为模式识别、机器学习,E-mail:hjxing@hbu.edu.cn。

的实验结果验证了所提集成方法的有效性。Krawczyk 等[9]提出了基于聚类划分的单类分类器集成，该集成方法首先利用聚类算法将目标类所在的数据空间划分成不相交的子区域，然后在每个子区域上训练一个单类分类器，最后融合所有单类分类器的输出。

虽然分类器集成能够提高单个分类器的泛化能力，但是随着基分类器个数的增加，集成模型所需的存储空间会变大且测试速度会变慢。为了克服分类器集成的上述缺陷，Zhou 等[10]提出了选择性分类器集成，并从理论和实验两方面证明了由部分基分类器训练得到的选择性集成模型能够保持甚至提高分类性能。此外，Wang 和 Li[11]提出了两种基于约束映射的支持向量机的选择性集成方法，第一种方法利用遗传算法寻找最优的组合权重，并设定一个阈值，剔除小于该阈值的基分类器，保留大于该阈值的基分类器；第二种方法是直接通过最小化集成分类器的期望均方误差得到集成模型。Li 和 Zhou[12]提出了基于正则化框架的选择性集成方法，该方法通过求解组合权重的二次规划问题得到稀疏解，从而得到基分类器的组合权重，实现选择性集成。Zhang 和 Zhou[13]提出了基于线性规划的稀疏性集成方法。对于单类分类选择性集成，Krawczyk[14]提出了一种单类分类器集成修剪方法，该方法利用群体智能算法中的萤火虫算法对集成规模进行约减；实验结果表明，该方法优于其他单类分类器集成修剪方法。

然而，已有的单类分类器集成并未综合考虑差异性和选择性集成对集成性能的影响；此外，传统的单类分类器集成方法的分类边界不够紧致。基于此，为了提高单类分类器集成分类边界的紧致性，并同时考虑集成的差异性和选择性集成，提出了一种基于相关熵和距离方差的 SVDD 选择性集成方法（Correntropy and Distance Variance Based Selective Ensemble of SVDDs，CDVSE-SVDDs）。主要有 3 方面工作：第一，利用信息理论学习中的相关熵度量来度量 SVDD 集成的紧致性，构造出紧致的分类边界；第二，计算训练样本到各最小包围球中心的距离，并利用距离方差作为集成的差异性度量，以提高基分类器的差异性；第三，在目标函数中引入了基于 ℓ_1 范数的正则化项，用于实现选择性集成。实验结果表明，所提方法能够有效地减少 SVDD 中基分类器的个数，并能使分类准确率不低于甚至高于使用所有基分类器的集成策略。

本文第 2 节简要回顾了 SVDD、相关熵以及基于方差的差异性度量；第 3 节详述了所提的基于相关熵和距离方差的 SVDD 选择性集成；第 4 节通过实验验证了所提方法在人工数据集和标准数据集上的有效性；最后总结全文。

2 相关知识

本节简要回顾了 SVDD 单类分类方法、信息理论学习中的相关熵、神经网络集成中基于方差的差异性度量。

2.1 SVDD

支持向量数据描述（SVDD）是由 Tax 和 Duin[3]在 SVM 的基础上提出的对一类数据分布区域进行描述的算法，其基本思想是利用核函数将样本从低维空间映射到高维特征空间，并寻找尽可能包围所有正常类样本的最小包围球。

给定包含 N 个多维数据的数据集 $\{\boldsymbol{x}_i\}_{i=1}^N$，首先定义一个包含该数据集的包围球，其中心和半径分别为 a 和 R，SVDD 通过最小化 R^2 寻找体积最小的包围球，用于对所给数据集进行描述，以便使整个数据集都包围在该球中。球外的数据看作异常数据，为了减小其影响，使用松弛变量 ξ_i 将异常数据排除在球外。因此，SVDD 的优化问题为：

$$\begin{aligned}&\min_{R,a,\xi} R^2+C\sum_{i=1}^N\xi_i\\&\text{s. t. } \|\boldsymbol{x}_i-\boldsymbol{a}\|^2\leqslant R^2+\xi_i,i=1,2,\cdots,N\\&\quad\xi_i\geqslant 0,i=1,2,\cdots,N\end{aligned}\tag{1}$$

其中，R 为包围球的半径，C 是控制球体体积和误差之间的折衷参数，ξ_i 是松弛变量，$\boldsymbol{a}$ 是包围球的中心。优化问题(1)可以由拉格朗日乘子法求解，此外，用核函数代替内积，可得下面的对偶优化问题：

$$\begin{aligned}&\min_{\alpha}\sum_{i=1}^N\sum_{j=1}^N\alpha_i\alpha_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)-\sum_{i=1}^N\alpha_iK(\boldsymbol{x}_i,\boldsymbol{x}_j)\\&\text{s. t. }\sum_{i=1}^N\alpha_i=1\\&\quad 0\leqslant\alpha_i\leqslant C,i=1,2,\cdots,N\end{aligned}\tag{2}$$

其中，α 为拉格朗日乘子，$K(\cdot,\cdot)$是核函数，对于核函数的选取，可以参看文献[15]。

待测样本的类别标号可由下式判别：

$$\begin{aligned}\|\phi(\boldsymbol{x})-\boldsymbol{a}\|^2=&K(\boldsymbol{x},\boldsymbol{x})-2\sum_{i=1}^N\alpha_iK(\boldsymbol{x},\boldsymbol{x}_i)+\\&\sum_{i=1}^N\sum_{j=1}^N\alpha_i\alpha_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)\leqslant R^2\end{aligned}\tag{3}$$

其中，$\phi(\boldsymbol{x})$是待测样本 $\boldsymbol{x}$ 在高维特征空间中的像。若满足上式，则待测样本是正常数据，否则就是异常数据。

2.2 相关熵

相关熵源于信息理论学习[16]，可以作为一种新的相似性度量，它比均方误差（Mean Squared Error，MSE）更具有鲁棒性，在文献[16]中有具体描述。下式给出了任意两个随机变量 X 和 Y 之间的相关熵：

$$V_\sigma(X,Y)=E[K_\sigma(X-Y)]\tag{4}$$

其中，$K(\cdot)$是核函数。然而，在实际中 X 和 Y 的联合概率密度函数是未知的，并且可用的数据 $\{\boldsymbol{x}_i,\boldsymbol{y}_i\}_{i=1}^N$ 是有限的。因此，相交相关熵的样本估计量为：

$$\hat{V}_\sigma(X,Y)=\frac{1}{N}\sum_{i=1}^N K_\sigma(\boldsymbol{x}_i-\boldsymbol{y}_i)\tag{5}$$

本文 $K_\sigma(\cdot)$取为高斯核函数，即：

$$k_\sigma(\boldsymbol{x}_i-\boldsymbol{y}_i)=G(\boldsymbol{x}_i-\boldsymbol{y}_i)=\exp\left(-\frac{(\boldsymbol{x}_i-\boldsymbol{y}_i)^2}{2\sigma^2}\right)\tag{6}$$

则式(5)可以改写为：

$$\hat{V}_{N,\sigma}(X,Y)=\frac{1}{N}\sum_{i=1}^N G(\boldsymbol{x}_i-\boldsymbol{y}_i)\tag{7}$$

2.3 基于方差的差异性度量

除了基分类器的精度之外，分类器集成的性能还依赖于基分类器之间的差异性[17]。提高基分类器之间的差异性能够有效地减少它们之间的相关性。最近，Yin 等[18]为神经网络集成提出了基于方差的差异性度量方法，该方法源于处理回归问题时的误差-差异度分解（Error-ambiguity Decomposition）[19]。

集成模型的平均误差公式如下：

$$\bar{E}=\sum_{m=1}^M w_mE_m\tag{8}$$

其中，M 为基分类器的个数，w_m 和 E_m 分别是第 m 个基分类器的组合权重和集成误差。E_m的表达公式如下：

$$\begin{aligned}E_m&=\int(y_m(\boldsymbol{x})-\bar{y}(\boldsymbol{x}))^2p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\&=\int\left(y_m(\boldsymbol{x})-\sum_{k=1}^M w_ky_k(\boldsymbol{x})\right)^2p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\end{aligned}\tag{9}$$

其中，x 是样本向量，且 $p(x)$ 为其概率密度，$y_m(x)$ 是第 m 个基分类器对于 x 的输出，$\bar{y}(x)$ 是所有基分类器的加权平均输出。假设所有样本都满足独立同分布，则第 m 个基分类器的集成误差为：

$$E_m \approx \sum_{n=1}^{N} \frac{1}{N} (y_m(x_n) - \sum_{k=1}^{M} w_k y_k(x_n))^2 \tag{10}$$

其中，N 为样本个数。利用式(8)和式(10)，集成模型的差异性可表示为[18]：

$$\begin{aligned} f_{diversity}(w) &= N \times \sum_{m=1}^{M} w_m \left(\sum_{n=1}^{N} \frac{1}{N} (y_m(x_n) - \sum_{k=1}^{M} w_k y_k(x_n))^2 \right) \\ &= \sum_{n=1}^{N} \left(\sum_{m=1}^{M} w_m (y_n^m)^2 - (w^{\mathrm{T}} x_n)^2 \right) \end{aligned} \tag{11}$$

其中，$y_n^m = y_m(x_n)$，$y_n = (y_n^1, y_n^2, \cdots, y_n^M)^{\mathrm{T}}$，$w = (w_1, w_2, \cdots, w_M)^{\mathrm{T}}$ 为 M 个基分类器的组合权重向量。

3 基于相关熵和距离方差的 SVDD 选择性集成

本节将详细描述基于相关熵和距离方差的 SVDD 选择性集成方法，并利用半二次优化技术求解相应的优化问题。

3.1 选择性集成模型

设有 M 个 SVDD，它们的集成半径定义为：

$$\bar{r} = \sum_{k=1}^{M} w_k r_k = (w, r) \tag{12}$$

其中，r_i 是第 i 个 SVDD 的半径，$r = (r_1, r_2, \cdots, r_M)^{\mathrm{T}}$ 和 $w = (w_1, w_2, \cdots, w_M)^{\mathrm{T}}$ 分别为 M 个 SVDD 的半径构成的向量及其对应的组合权重向量。为了衡量集成模型的紧致性，采用相关熵度量各个基分类器的半径与集成半径之间的相似性程度，如下：

$$\frac{1}{M} \sum_{k=1}^{M} k_\sigma(r_k - \bar{r}) = \frac{1}{M} \sum_{k=1}^{M} \exp\left(-\frac{(r_k - \bar{r})^2}{2\sigma^2}\right) \tag{13}$$

其中，σ 为宽度参数。

为了增加基分类器之间的差异性，提出了一种基于距离方差的差异性度量方法，其表达公式如下：

$$\tilde{f}_{diversity}(w) = \sum_{i=1}^{M} \sum_{n=1}^{N} [d_i(x_n) - \bar{d}(x_n)]^2 \tag{14}$$

其中，$d_i(x_n)$ 为样本 x_n 在高维特征空间中的像 $\phi(x_n)$ 到第 i 个 SVDD 中心的距离，如图 1 所示。

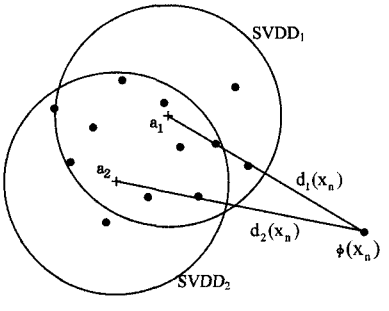


图 1 样本 x_n 的像 $\phi(x_n)$ 到各个 SVDD 中心的距离

设第 i 个 SVDD 的中心为 a_i，则 $d_i(x_n)$ 可表示为：

$$\begin{aligned} d_i(x_n) &= \| \phi(x_n) - a_i \| \\ &= \sqrt{K(x_n, x_n) - 2\sum_{j=1}^{N} \alpha_{ij} K(x_n, x_j) + \sum_{j=1}^{N} \sum_{l=1}^{N} \alpha_{ij} \alpha_{il} K(x_j, x_l)} \end{aligned} \tag{15}$$

$\bar{d}(x_n)$ 为 $\phi(x_n)$ 到集成中心的距离：

$$\bar{d}(x_n) = \sum_{k=1}^{M} w_k d_k(x_n) \tag{16}$$

为了实现选择性集成，在集成模型中增加基于 ℓ_1 范数的正则化项，即 $\| w \|_1$，此处要求组合权重的元素均大于零，即 $w_k \geqslant 0 (k = 1, 2, \cdots, M)$。

综上，所提基于相关熵和距离方差的 SVDD 选择性集成模型为：

$$\begin{aligned} \max\ & \frac{1}{M} \sum_{k=1}^{M} \exp\left(-\frac{(r_k - \bar{r})^2}{2\sigma^2}\right) - \lambda \sum_{i=1}^{M} \sum_{n=1}^{N} [d_i(x_n) - \sum_{k=1}^{M} w_k d_k(x_n)]^2 - \gamma \| w \|_1 \\ \text{s.t. } & w_k \geqslant 0, k = 1, 2, \cdots, M \end{aligned} \tag{17}$$

上述优化问题可以利用半二次优化技术[20]进行求解。

根据凸共轭函数理论[21]，则有：

定理 1 对于 $G(z) = \exp(-\frac{z^2}{2\sigma^2})$，存在凸共轭函数 φ，使得

$$G(z) = \sup_{\alpha \in R^-} \left(\alpha \frac{z^2}{2\sigma^2} - \varphi(\alpha)\right) \tag{18}$$

其中，α 为辅助共轭变量。对于一个固定的 z，上式在 $\alpha = -G(z)$ 处取得最大值。

根据定理 1，式(10)可以改写为：

$$F(w, P) = \frac{1}{M} \sum_{k=1}^{M} [p_k \frac{(r_k - \bar{r})^2}{2\sigma^2} - \varphi(p_k)] - \lambda \sum_{i=1}^{M} \sum_{n=1}^{N} [d_i(x_n) - \sum_{k=1}^{M} w_k d_k(x_n)]^2 - \gamma \sum_{k=1}^{M} w_k \tag{19}$$

其中，P 为对角矩阵，其主对角线元素为 $P_{kk} = p_k$，且 $p_k (k = 1, 2, \cdots, M)$ 为辅助共轭变量。

根据半二次优化技术，可对目标函数(19)分两个步骤求解。

1. 计算辅助共轭变量 $p_k (k = 1, 2, \cdots, M)$：

$$p_k^{\tau+1} = -\exp\left(-\frac{(r_k - \bar{r}^{\tau})^2}{2\sigma^2}\right) \tag{20}$$

2. 将辅助变量代入式(19)，得到：

$$\begin{aligned} w^{\tau+1} &= \arg\max_{w} \frac{1}{M} \sum_{k=1}^{M} p_k^{\tau} \frac{(r_k - \bar{r})^2}{2\sigma^2} - \lambda \sum_{i=1}^{M} \sum_{n=1}^{N} [d_i(x_n) - \sum_{k=1}^{M} w_k d_k(x_n)]^2 - \gamma \sum_{k=1}^{M} w_k \\ &= \arg\max_{w} \frac{1}{2} w^{\mathrm{T}} \left(\frac{\sum_{k=1}^{M} p_k^{\tau} r r^{\mathrm{T}}}{\sigma^2 M} - 2\lambda M \sum_{n=1}^{N} d_n d_n^{\mathrm{T}} \right) w + \\ &\quad \left(-\frac{\sum_{k=1}^{M} p_k^{\tau} r_k r}{\sigma^2 M} + 2\lambda \sum_{n=1}^{N} d_n d_n^{\mathrm{T}} 1 - \gamma 1 \right)^{\mathrm{T}} w + \text{const} \end{aligned} \tag{21}$$

其中，τ 和 $\tau+1$ 分别表示第 τ 和第 $\tau+1$ 次迭代，$d_n = (d_1(x_n), d_2(x_n), \cdots, d_M(x_n))^{\mathrm{T}}$，1 为元素均为 1 的列向量，const 是与 w 无关的常数。

对式(21)的目标函数关于组合向量 w 求导数，并令导数为零，可得：

$$\begin{aligned} w^{\tau+1} = & \left(\frac{\sum_k p_k^{\tau} r r^{\mathrm{T}}}{\sigma^2 M} - 2\lambda M \sum_n d_n d_n^{\mathrm{T}} \right)^{-1} \times \left(\frac{\sum_k p_k^{\tau} r_k r}{\sigma^2 M} - \right. \\ & \left. 2\lambda \sum_n d_n d_n^{\mathrm{T}} 1 + \gamma 1 \right) \end{aligned} \tag{22}$$

本文中 λ 和 γ 均取值为 1。

3.2 学习算法

基于相关熵和距离方差的 SVDD 选择性集成的整个训练过程概括在算法 1 中。集成模型训练完成之后，将最优组合权重向量 w^* 中元素值小于 $1/M$ 的权重所对应的 SVDD 从

集成模型中剔除，从而实现选择性集成。

算法 1 基于相关熵和距离方差的 SVDD 选择性集成

输入：训练样本$\{\mathbf{x}_n\}_{n=1}^{N}$，宽度参数 σ，最大迭代次数 T_{max}，距离方差项的系数 λ，ℓ_1 范数正则化项系数 γ，基分类器 SVDDR 的个数 M，基分类器 SVDD 的核函数参数 ζ 和折衷参数 C

输出：最优组合权重向量 $\mathbf{w}^*$

初始化：组合权重向量 $\mathbf{w}=1/M$。

步骤 1 训练 M 个基分类器 SVDD。

步骤 2 由式(12)计算集成半径 $\bar{r}$，由式(15)计算所有样本到 M 个基分类器 SVDD 的距离 $d_i(\mathbf{x}_n)(i=1,2,\cdots,M;n=1,2,\cdots,N)$。

步骤 3 更新辅助共轭变量 $p_k(k=1,2,\cdots,M)$和组合权重向量 $\mathbf{w}$。

for $\tau=1,2,\cdots,T_{max}$

a)根据式(20)更新辅助共轭变量；

b)根据式(22)更新组合权重向量。

end for

4 实验结果

为了验证所提基于相关熵和距离方差的 SVDD 选择性集成方法的有效性，本节首先利用两个人工数据集考察了其可行性，然后在 13 个标准数据集上将 CDVSE-SVDDs 与单个 SVDD 以及其他单类分类器集成方法进行了比较。在 SVDD 和 CDVSE-SVDDs 中，核函数使用径向基核函数 $K(\boldsymbol{x},\boldsymbol{y})=\exp(-\zeta\|\boldsymbol{x}-\boldsymbol{y}\|^2)$。从所给数据集的正常样本中随机选取 70%构成训练集。此外，以下实验均采用几何平均值(g-means)[22]来衡量不同单类分类方法的性能。g-means 定义为：

$$g=\sqrt{a^+ \cdot a^-} \tag{23}$$

其中，a^+ 和 a^- 分别是正常样本和异常样本的分类正确率。

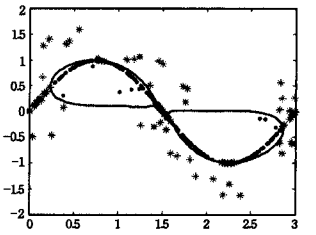
4.1 人工数据集

本节所使用的两个人工数据集分别为 Sine_noise 和 Square_noise，各自的描述如下。

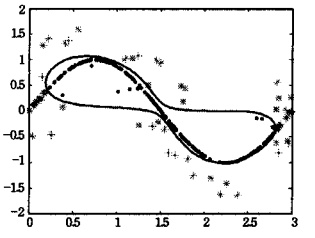
Sine_noise：该数据集包含 250 个二维样本点。从集合$\{(x,y):x\in[0,3],y\in[0,3]\}$中随机选取 200 个样本点作为正常数据，然后从集合$\{(x,y)\mid y=\sin(\frac{2\pi}{3}x)\pm\varepsilon,\varepsilon\sim N(0,1)\}$中随机选取 50 个样本点作为异常数据。

Square_noise：该数据集也包含 250 个二维样本点。从集合$\{(x,y)\mid x\in[0.3,2.7],y\in[0.3,0.6]\cup[2.4,2.7]\}\cup\{(x,y)\mid x\in[0.3,0.6]\cup[2.4,2.7],y\in[0.3,2.7]\}$中随机选取 200 个样本点作为正常数据，然后从集合$\{(x,y)\mid x\in[0,3],y\in[0,3]\}$中随机选取 50 个样本点作为异常数据。

对于两个数据集，CDVSE-SVDDs 的基分类器个数和半二次优化的迭代次数均设为 20，相关熵的宽度参数 $\sigma=0.3$。SVDD 和 CDVSE-SVDDs 基分类器的核函数参数 $\zeta=0.3$，折衷参数 $C=1$。SVDD 和 CDVSE-SVDDs 在 Sine_noise 上的分类效果分别如图 1(a)和图 1(b)所示，两者的 g-means 值分别为 0.7710 和 0.8438，最终 CDVSE-SVDDs 的基分类器个数为 2。SVDD 和 CDVSE-SVDDs 在 Square_noise 上的分类效果分别如图 2(a)和图 2(b)所示，两者的 g-means 值分别为 0.7909 和 0.8224，最终 CDVSE-SVDDs 的基分类器数为 5。

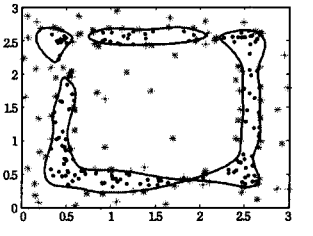


(a)SVDD

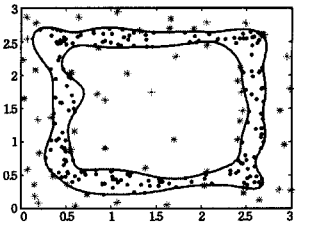


(b)CDVSE-SVDDs

图 2 SVDD 与 CDVSE-SVDDs 在 Sine_noise 人工数据集上的分类效果



(a)SVDD



(b)CDVSE-SVDDs

图 3 SVDD 与 CDVSE-SVDDs 在 Square_noise 人工数据集上的分类效果

综上，CDVSE-SVDDs 在两个人工数据集上均取得了优于单个 SVDD 的分类性能。

4.2 标准数据集

本节选用 UCI 机器学习数据库[23]中的 13 个标准数据集来验证所提方法的有效性。从每个标准数据集中选用某一类样本作为正常数据，剩余的另一类样本作为异常数据。所有标准数据集的信息如表 1 所列。

表 1 实验所使用 13 个标准数据集的相关信息

数据集	特征数	正常类个数	异常类个数
Cancer	9	444	239
Breast_cancer	9	186	77
Cleverland_heart	13	83	214
Diabetis	8	500	268
Flare_solar	9	94	50
German	20	700	300
Heart	13	150	120
Hepatitis	19	123	32
Housing	13	261	245
Liver	6	145	200
Pima	8	500	268
Sonar	60	97	111
Thyroid	5	150	65

在以下实验中，从正常数据和异常数据中分别随机选取70%和0.5%作为训练集，剩余的正常数据和异常数据用作测试集。CDVSE-SVDDs的基分类器的个数设置为50。首先在训练集上采用五折交叉验证选取单个SVDD的最优参数，即径向基核函数参数ζ和折衷参数C，最优参数确定之后，将其直接用作CDVSE-SVDDs中基分类器的参数。然后，再使用五折交叉验证法选取CDVSE-SVDDs的其他参数，包括相关熵的宽度参数σ、系数λ及γ。SVDD和CDVSE-SVDDs的最优参数值概括在表2中，此外，表2还给出了CDVSE-SVDDs最终选择的基分类器个数。

表2 SVDD及CDVSE-SVDDs的参数设置

数据集	C	ζ	σ	最终基分类器个数
Cancer	0.07	0.1	0.01	2
Breast_cancer	0.4	0.01	0.7	4
Cleverland_heart	0.2	0.01	0.05	17
Diabetis	0.01	2	7	2
Flare_solar	0.6	0.1	0.9	12
German	0.4	0.1	0.8	2
Heart	0.3	0.1	0.1	5
Hepatitis	0.4	0.1	7	4
Housing	0.02	0.5	0.01	3
Liver	0.5	2	0.7	6
Pima	0.4	1	0.7	2
Sonar	0.4	0.01	0.7	4
Thyroid	0.06	0.06	1	31

由表2可知，CDVSE-SVDDs可以有效减少集成中所使用的基分类器个数，从而有效提高测试速度。为了减小训练集的随机选取对结果的影响，将上述实验在每个数据集上重复20次，将平均训练准确率和平均测试准确率分别作为最终的训练准确率和测试准确率。实验结果如表3和表4所列，20次实验结果的标准差也概括在表中。此外，两种传统分类器集成方法即Bagging和AdaBoost的结果也分别在表3和表4中给出。

表3 4种不同方法在标准数据集上的训练结果(平均值±标准差(%))

数据集	SVDD	Bagging	AdaBoost	CDVSE-SVDDs
Cancer	96.48±0.56	95.83±2.24	96.32±0.35	96.70±0.74
Breast_cancer	74.95±8.11	78.12±6.44	75.89±1.24	78.16±4.72
Cleverland_heart	89.80±3.73	90.99±2.36	86.32±2.25	90.36±2.50
Diabetis	92.45±5.27	90.78±2.75	88.74±1.59	92.48±4.44
Flare_solar	63.14±3.74	62.99±4.17	59.47±2.85	63.91±5.46
German	77.47±1.21	76.89±0.48	76.01±0.57	77.27±1.14
Heart	82.86±3.86	82.85±2.70	78.68±1.89	83.44±4.32
Hepatitis	76.12±4.50	75.88±3.94	71.33±2.27	76.53±5.27
Housing	94.99±3.52	92.28±4.43	90.91±1.82	93.57±3.81
Liver	70.35±2.67	69.46±2.07	67.93±1.75	70.97±3.39
Pima	77.30±1.29	76.91±0.40	75.99±0.82	77.48±1.40
Sonar	75.55±2.08	77.57±2.01	73.75±1.51	77.30±4.23
Thyroid	96.89±0.78	96.70±1.28	96.33±0.68	96.47±1.48

表4 4种不同方法在标准数据集上的测试结果(平均值±标准差(%))

数据集	SVDD	Bagging	AdaBoost	CDVSE-SVDDs
Cancer	94.95±1.81	94.30±3.69	95.02±1.11	95.16±1.41
Breast_cancer	58.70±5.73	59.89±2.91	59.10±3.66	60.06±3.62
Cleverland_heart	75.71±2.75	75.56±3.38	75.61±3.54	75.99±3.76
Diabetis	64.05±2.14	64.20±1.74	64.07±1.59	64.27±1.79
Flare_solar	42.30±2.99	42.43±2.91	42.51±3.54	42.82±3.81
German	55.39±1.57	55.41±1.60	55.46±1.52	55.62±1.57
Heart	71.75±3.56	71.40±3.05	70.47±3.39	72.13±2.93
Hepatitis	61.27±4.64	61.18±5.06	59.70±5.48	62.30±4.38
Housing	75.45±2.41	75.02±2.29	75.20±2.82	76.02±2.49
Liver	54.25±3.33	54.81±3.44	54.20±3.69	54.83±3.33
Pima	64.70±1.79	64.83±1.75	64.41±1.80	64.95±1.56
Sonar	60.15±4.88	59.86±4.66	58.73±5.03	60.64±4.69
Thyroid	94.06±2.14	93.98±1.88	93.80±2.54	94.23±1.81

由表4可知，综合考虑测试准确率及其标准差，CDVSE-SVDDs在标准数据集上取得了优于单个SVDD、基于Bagging的SVDD集成以及基于AdaBoost的集成的测试结果。

综合以上实验结果，本文所提CDVSE-SVDDs在标准数据集上均取得了优于相关方法的分类性能，并且有效地减少了基分类器的数量。CDVSE-SVDDs的优势概括如下：

(1)利用基分类器半径和集成半径之间的相关熵度量来度量SVDD集成的紧致性，可以构造出更为紧致的分类边界。

(2)使用距离方差作为单类分类器集成中的差异性度量，可以提高基分类器之间的差异性。

(3)在集成模型中引入基于ℓ_1范数的正则化项，可以有效地剔除对集成贡献较小的基分类器，提高集成性能。

结束语 基于信息理论中的相关熵和神经网络集成中基于方差的差异性度量，提出了一种支持向量数据描述集成策略，同时，在所提模型的目标函数中引入基于ℓ_1范数的正则化项，从而实现了选择性集成。利用半二次优化技术对所提模型进行求解。实验结果表明，与相关模型相比，所提基于相关熵和距离方差的支持向量数据描述选择性集成取得了更优的分类性能。

为了使所提方法更具吸引力，在以后的工作中，将从两个方面对所提选择性集成加以研究和探讨。首先，相关熵的宽度参数σ、支持向量数据描述的核函数参数ζ和折衷参数C对集成模型的性能影响很大，它们均由“试差法”进行选取，非常耗时，以后会考虑使用启发式方法对上述参数进行选取。其次，基于相关熵的准则已被证明比基于均方误差的准则具有更强的抗噪声能力，以后会进一步检查所提集成模型的抗噪声能力。

参 考 文 献

[1] Tax D M J. One-class classification: concept learning in the absence of counter examples[D]. Delft University of Technology, 2001

[2] Schölkopf B, Williamson R C, Smola A J, et al. Support vector method for novelty detection[C]// Conference: Advances in Neural Information Processing Systems. 2000: 582-588

[3] Tax D M J, Duin R P W. Support vector data description[J]. Machine Learning, 2004, 54(1): 45-66

[4] Tax D M J, Duin R P W. Combining one-class classifiers[C]// Proceedings of the 2nd International Workshop on Multiple Classifier Systems. 2001: 299-308

[5] Seguí S, Igual L, Vitrià J. Weighted bagging for graph based one-class classifiers[C]// Proceedings of the 9th International Workshop on Multiple Classifier Systems. 2010: 1-10

[6] Zhang J, Lu J, Zhang G. Combining one class classification models for avian influenza outbreaks[C]// Proceedings of the 2011 IEEE Symposium on Computational Intelligence in Multicriteria Decision-Making. 2011: 190-196

[7] Wilk T, Wozniak M. Soft computing methods applied to combination of one-class classifiers[J]. Neurocomputing, 2012, 75: 185-193

[8] Casale P, Pujol O, Radeva P. Approximate polytope ensemble for one-class classification[J]. Pattern Recognition, 2014, 47: 854-864

(下转第264页)

表 3 人工模拟数据测试结果

(l,d)	Algorithms		
	PROJECTION	CEM	SCEM
(11,2)	0.88(10s)	0.96(10s)	0.97(10s)
(15,4)	0.94(4m)	0.98(45s)	1(26s)
(18,6)	0.89(33m)	0.98(9.8m)	0.98(5.8m)
(19,7)	0.86(1.5h)	0.95(18.2m)	0.96(8.9m)
(27,8)	0.84(2h)	0.94(31.2m)	0.96(20.1m)
(36,11)	0.78(2.5h)	0.91(40.5m)	0.96(34.6m)

通过表 3 可以看到，相比随机投影和 CEM，SCEM 不但能保证较高的算法精度，并且具有较低的时间开销。在寻找长植入模式时，其算法精度都保持在 95%以上，时间开销远小于随机投影。

结束语 本文提出的算法可以在合理的时间内有效解决 (l,d)植入模体发现问题，并且在解决长植入模体时也展现出了良好的性能，与目前流行的算法相比极具竞争力。通过对期望最大化算法中隐变量 Z 的约束，加速各种子集收敛求精，并且由于各种子集的运算相互独立，所提算法也易于并行化。而结合 ChIP-seq 等新技术，处理更大规模的数据，则是我们后续所关注的研究内容。

参 考 文 献

[1] Park P J. ChIP-seq: advantages and challenges of a maturing technology[J]. Nat Rev Genet, 2009, 10(10): 669-80

[2] Zhang Yi-pu, Wang Ping. A Fast Cluster Motif Finding Algorithm for ChIP-Seq Data Sets[C]// BioMed Research International, 2015. 2015

[3] Bailey T L, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in bipolymers[C]// Second International Conference on Intelligent Systems for Molecurlar Biology. 1994: 28-36

[4] Zambelli F, Pesole G, Pavesi G. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era[J]. Briefings in Bioinformatics, 2013, 14(2): 225-237

[5] Zhang Yi-pu, Huo H, Yu Qiang. A Heuristic Cluster-based EM Algorithm for the Planted(l,d) Problem[J]. Journal of Bioinformatics and Computational Biology, 2013, 11(4): 1350009

[6] Pevzner P A, Sze S H. Combinatorial approaches to finding subtle signals in DNA sequences[C]// ISMB. 2000: 269-278

[7] Reid J, Wernisch L. STEME: efficient EM to find motifs in large data sets[J]. Nucleic Acids Research, 2011, 39(18): 126

[8] Buhler J, Tompa M. Finding motifs using random projections [J]. Journal of Computational Biology, 2002, 9(2): 225-242

[9] Chen Kun, Zhang Xiao-jun. MCI Clustering Algorithm Solving Planted (l,d) Motif Identification[J]. Journal of Henan University(Natural Science), 2015, 45(1): 102-107(in Chinese)
陈昆，张小骏. MCL 聚类算法求解植入(l,d)模体识别问题[J]. 河南大学学报(自然科学版)，2015, 45(1): 102-107

[10] Sun De-cai, Wang Xiao-xia. Research on Filtering Algorithm of Approximate String Matching[J]. Computer Technology and Development, 2015(4): 171-176(in Chinese)
孙德才，王晓霞. 近似串匹配过滤算法研究[J]. 计算机技术与发展，2015(4): 171-176

[11] Xu Yong-kang, Yang Guang-lu, Lu Song-feng, et al. Approximate string matching algorithm based on compressed suffixarray [J]. Computer Engineering and Applications, 2015, 51(23): 139-142(in Chinese)
胥永康，杨光露，路松峰，等. 基于压缩后缀数组的近似字符串匹配算法[J]. 计算机工程与应用，2015, 51(23): 139-142

[12] Chan H L, Lam T W, Sung W K, et al. Compressed indexes for approximate string matching[J]. Algorithmica, 2010, 58(2): 263-281

[13] Atallah M J, Grigorescu E, Wu Y. A lower-variance randomized algorithm for approximate string matching[J]. Information Processing Letters, 2013, 113(18): 690-692

(上接第 256 页)

[9] Krawczyk B, Woźniak M, Cyganek B. Clustering-based ensembles for one-class classification[J]. Information Sciences, 2014, 264: 182-195

[10] Zhou Z H, Wu J X, Tang W. Ensembling neural networks: many could be better than all[J]. Artificial Intelligence, 2002, 137(1/2): 239-263

[11] Wang L, Li Q. Effective selective ensemble algorithms for support vector machines[J]. Neurocomputing, 2010: 287-295

[12] Li N, Zhou Z. Selective ensemble under regularization framework[J]. Lecture Notes in Computer Science, 2009, 5519: 293-303

[13] Zhang L, Zhou W. Sparse ensembles using weighted combination methods based on linear programming[J]. Pattern Recognition, 2011, 44(1): 97-106

[14] Krawczyk B. One-class classifier ensemble pruning and weighting with firefly algorithm[J]. Neurocomputing, 2015, 150(B): 490-500

[15] Vapnik V N. Statistical Learning Theory[M]. New York: Wiley, 1998

[16] Principe J C. Information Theory Learning: Renyi's Entropy and Kernel Perspectives[M]. Springer, 2012

[17] Tang E K, Suganthan P N, Yao X. An analysis of diversity measures[J]. Machine Learning, 2006, 65(1): 247-271

[18] Yin X C, Huang K, Yang C, et al. Convex ensemble learning with sparsity and diversity[J]. Information Fusion, 2014, 20: 49-59

[19] Vedelsby J, Krogh A. Neural network ensembles, cross-validation and active learning[J]. Advances in Neural Information Processing Systems, 1995, 7: 231-238

[20] He R, Hu B G, Zheng W S, et al. Robust principal component analysis based on maximum correntropy criterion[J]. IEEE Transactions on Image Processing, 2011, 20(6): 1485-1494

[21] Eockfellar R. Convex analysis[M]. Princeton University Press, Princeton, 1970

[22] Wu M, Ye J. A small sphere and large margin approach for novelty detection using training data with outliers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(11): 2088-2092

[23] Frank A, Asuncion A. UCI machine learning repository[OL]. University of California, Irvine, School of Information and Computer Science, Irvine, CA. http://archive.ics.uci.edu/ml