第 39 卷 第 5 期
Vol.39 No.5
计 算 机 工 程
Computer Engineering
2013 年 5 月
May 2013

·人工智能及识别技术· 文章编号：1000—3428(2013)05—0165—04 文献标识码：A 中图分类号：TP18

基于多核支持向量数据描述的单类分类方法

吴定海[a]，张培林[a]，王怀光[a]，傅建平[b]

(军械工程学院 a. 车辆与电气工程系；b. 火炮工程系，石家庄 050003)

摘 要：针对异常检测模型中，单核支持向量数据描述存在映射形式单一以及核函数、核参数选择困难的问题，提出一种多核优化组合的支持向量域描述的单类分类方法。在分析多核映射的核空间基础上，建立多核支持向量数据描述模型，以更灵活地描述训练样本在高维特征空间的边界分布情况。采用目标函数的梯度下降法对该模型的多核组合权重进行分步寻优，并引入异常类测试样本来控制和评价分类器的描述精度和推广能力。仿真实验结果表明，该方法具有更好的学习能力和计算效率。

关键词：模式识别；单类分类；多核学习；支持向量数据描述；异常检测

One-class Classification Method Based on Multi-kernel Support Vector Data Description

WU Ding-hai[a], ZHANG Pei-lin[a], WANG Huai-guang[a], FU Jian-ping[b]

(a. Department of Vehicle and Electrical Engineering; b. Department of Ordnance Engineering, Artillery Engineering College, Shijiazhuang 050003, China)

【Abstract】Considering the support vector data description model having disadvantages of simple form with only one kernel information and hard to choose the best kernel and its parameters, the multi-kernel one-class classification with a linear combination of multi-kernel is proposed. The multi-kernel support vector data description model which can descript the data distribution boundary in eigenspace more flexibly is built after analysing the space of multi-kernel mapping. The optimal combination kernels' weight is solved by reduced gradient algorithm. Test dataset which includes abnormal samples is introduced to control and evaluate the description accuracy and expansibility of hyper spherical interface. Experimental results show the method has better learning ability and computing efficiency.

【Key words】pattern recognition; one-class classification; multi-kernel learning; support vector data description; anomaly detection

DOI: 10.3969/j.issn.1000-3428.2013.05.036

1 概述

支持向量机(Support Vector Machine, SVM)建立在统计学习理论的基础上，遵循结构风险最小化原理，能够获得更好的泛化性能[1]。近年来，针对模式识别中数据样本严重不平衡的问题，支持向量数据描述(Support Vector Data Description, SVDD)方法只需要一类数据样本即可建立超球分界面，且具有计算速度快、鲁棒性强、算法复杂度低等优点，成为模式识别领域研究和应用的热点[2-3]。

SVDD 在很大程度上取决于核映射方式的选择，采用不同的核函数具有不同的映射效果。目前 SVDD 均只采用单核映射，映射的形式较为单一，具有一定的局限性，如果优化组合多种核函数或同一核函数不同核参数的映射效果，将能够提高核特征空间映射的灵活性，多核学习的引入将能很好地解决支持向量数据描述中核函数种类和核函数参数的选择问题[4]。另外，SVDD 采用无监督学习方式，核参数的选择存在较大的随意性。

基于以上分析，本文提出多核支持向量数据描述(Multi-kernel Support Vector Data Description, MSVDD)的方法。通过对目标函数的梯度下降方法寻找多核学习的权重，针对 SVDD 无监督的问题，给出一种目标数据驱动的优化控制方法，最终获得最优的权重向量。

2 MSVDD 的核空间假设

多核学习问题[4-6]描述如下：假设核函数由 M 个正定核的线性组成，其中任意一个核 K_m 均对应某一再生核 Hilbert

作者简介：吴定海(1981－)，男，讲师、博士，主研方向：机械状态监测，信号处理，人工智能；张培林，教授、博士生导师；王怀光，讲师、博士研究生；傅建平，副教授、博士

收稿日期：2012-02-24 **修回日期：**2012-07-16 **E-mail：**wudh81@163.com

空间 $\boldsymbol{H}_m$ 及相应的内积运算 $\langle\cdot,\cdot\rangle_m$，对于 $\boldsymbol{d}_m\in[0,1]$，任一 $\boldsymbol{K}(x,x')=\boldsymbol{d}_m\boldsymbol{K}_m(x,x')$，则对应某一 Hilbert 空间 $\boldsymbol{H}_m'$，内积运算为：

$$\langle f,g\rangle_{\boldsymbol{H}_m'}=\frac{1}{\boldsymbol{d}_m}\langle f,g\rangle_m \tag{1}$$

$\forall f\in\boldsymbol{H}_m'\in\boldsymbol{H}_m$，有：

$$f(x)=\langle f,\boldsymbol{K}_m(\cdot,x)\rangle_m=\frac{1}{\boldsymbol{d}_m}\langle f,\boldsymbol{d}_m\boldsymbol{K}_m(\cdot,x)\rangle_m=\langle f,\boldsymbol{d}_m\boldsymbol{K}_m(\cdot,x)\rangle_{\boldsymbol{H}_m'} \tag{2}$$

因此，根据再生核的性质[7]，可知 $\boldsymbol{H}_m'$ 也是再生核 Hilbert 空间。定义多核学习的核空间 $\boldsymbol{H}$ 为 $\boldsymbol{H}_m'$ 空间的集合，$\boldsymbol{H}=\bigoplus_{m=1}^{M}\boldsymbol{H}_m'$，则核空间 $\boldsymbol{H}$ 也是定义在核函数式(3)下的再生核 Hilbert 空间：

$$\boldsymbol{K}(x,x)=\sum_{m=1}^{M}\boldsymbol{d}_m\boldsymbol{K}_m(x,x) \tag{3}$$

因此，多核学习的问题转化为求解组合核的权重系数 $\{\boldsymbol{d}_m\}$，通过调整组合核函数的权重从而获得自适应核映射特征空间和自适应的内积运算。

3 MSVDD 模型

文献[8-9]提出了 SVDD 的基本思想是：在通过核函数映射的高维特征空间中寻求一个包围目标样本点的超球体，并通过最小化该超球体所包围的体积让目标样本点尽可能地被包围在超球体中，而非目标样本点排除在超球体中，达到非目标数据检测的目的。

根据多核空间的构建理论，多核支持向量数据描述的核空间映射模式[10]变为：

$$\boldsymbol{K}(x,x)=\sum_{m=1}^{M}\boldsymbol{d}_m\boldsymbol{K}_m(x,x)=\sum_{m=1}^{M}\boldsymbol{d}_m\langle\boldsymbol{\phi}(x),\boldsymbol{\phi}(x)\rangle=\sum_{m=1}^{M}\langle\boldsymbol{d}_m^{1/2}\boldsymbol{\phi}(x),\boldsymbol{d}_m^{1/2}\boldsymbol{\phi}(x)\rangle \tag{4}$$

对于正常数据样本 $\{x\mid x_i\in\boldsymbol{R}^d,i=1,2,\cdots,N\}$，在多核映射的高维空间内，确定包含该数据样本的最小超球体的中心 a 和半径 R，转化为求解以下优化问题：

$$\begin{cases}\min\limits_{R,\xi}R^2+C\sum\limits_{i=1}^{N}\xi_i\\ \text{s.t.}\ \left\|\sum\limits_{m=1}^{M}\boldsymbol{d}_m^{1/2}\boldsymbol{\phi}_m(x_i)-a\right\|^2\leqslant R^2+\xi_i\\ \xi_i\geqslant 0\quad\forall i\\ \sum\limits_m\boldsymbol{d}_m=1,\quad\boldsymbol{d}_m\geqslant 0\quad\forall m\end{cases} \tag{5}$$

可见，对于组合多核支持向量数据描述来说，需要在原问题上增加一个变量以及由这个变量衍生出来的 2 个约束，同样可以利用 Lagrange 函数将上述问题转化为求解对偶问题[11]，简化问题的求解。而实际这种求解方式的复杂度高，往往难以实现。在这里通过 2 步优化求解该凸二次优化问题的变量，优化问题按照以下步骤求解：

(1)初始化多核映射权重向量 $\boldsymbol{d}$，利用 $\boldsymbol{d}$ 计算输入数据组合映射的核矩阵 $\boldsymbol{K}$。

(2)由所得的核矩阵 $\boldsymbol{K}$，利用 SVDD 求解超球分界面和优化目标函数值。

(3)更新和迭代权重向量 $\boldsymbol{d}$，返回第(2)步求解，直到达到需要优化的目标。

4 多核支持向量数据描述问题求解

4.1 优化目标设计的描述

根据以上分步优化的思想，结合支持向量数据描述的特点，将优化的目标定义为：

$$\min_{d}\boldsymbol{J}(\boldsymbol{d})\quad\text{s.t.}\ \sum_{m=1}^{M}\boldsymbol{d}_m=1,\ \boldsymbol{d}_m\geqslant 0 \tag{6}$$

其中，$\boldsymbol{d}=\{d_1,d_2,\cdots,d_M\}$ 为组合核对权重向量，且：

$$\boldsymbol{J}(\boldsymbol{d})=\begin{cases}\min\limits_{R,\xi}R^2+C\sum\limits_{i=1}^{N}\xi_i\\ \text{s.t.}\ \left\|\sum\limits_{m=1}^{M}\boldsymbol{d}_m^{1/2}\boldsymbol{\phi}_m(x_i)-a\right\|^2\leqslant R^2+\xi_i\\ \xi_i\geqslant 0\quad\forall i\end{cases} \tag{7}$$

在多核机器学习算法中，采用多个核函数的线性组合式(4)作为核函数。当权重向量 $\boldsymbol{d}$ 确定后，就可以直接用 SVDD 进行求解，多核支持向量数据描述的对偶问题为：

$$\begin{cases}\max\limits_{\boldsymbol{\alpha}}\sum\limits_{i=1}^{N}\boldsymbol{\alpha}_i\sum\limits_{m=1}^{M}\boldsymbol{d}_m\boldsymbol{k}_m(x_i,x_j)-\sum\limits_{i=1}^{N}\boldsymbol{\alpha}_i\sum\limits_{j=1}^{N}\boldsymbol{\alpha}_j\sum\limits_{m=1}^{M}\boldsymbol{d}_m\boldsymbol{k}_m(x_i,x_j)\\ \text{s.t.}\ \sum\limits_{i=1}^{N}\boldsymbol{\alpha}_i=1\quad 0\leqslant\boldsymbol{\alpha}_i\leqslant C\end{cases} \tag{8}$$

根据最优化理论中的强对偶问题，上述优化目标转变为对偶问题的最大值，即：

$$\boldsymbol{J}(\boldsymbol{d})=\sum_{i=1}^{N}\boldsymbol{\alpha}_i\sum_{j=1}^{N}\boldsymbol{\alpha}_j\sum_{m=1}^{M}\boldsymbol{d}_m\boldsymbol{k}_m(x_i,x_j)-\sum_{i=1}^{N}\boldsymbol{\alpha}_i\sum_{m=1}^{M}\boldsymbol{d}_m\boldsymbol{k}_m(x_i,x_i) \tag{9}$$

在更新权重系数时最小化 $\boldsymbol{J}(\boldsymbol{d})$，便可以获得较好的描述效果。

4.2 最优权重的求解

优化目标函数的梯度计算为权重变量的变化指明了方向。因此，对于多核权重的学习问题可以用较为简单的梯度下降方法[12]进行迭代更新，直到满足优化的目标。由于优化目标函数是凸的且是可微的，因此对其求微分有：

$$\frac{\partial\boldsymbol{J}(\boldsymbol{d})}{\partial\boldsymbol{d}_m}=\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j\boldsymbol{k}_m(x_i,x_j)-\sum_{i=1}^{N}\boldsymbol{\alpha}_i\boldsymbol{k}_m(x_i,x_i) \tag{10}$$

可知目标函数的梯度与权重系数无关，为了最小化目标函数 $\boldsymbol{J}(\boldsymbol{d}_m)$，$-\nabla\boldsymbol{J}(\boldsymbol{d}_m)$ 就是变量 $\{\boldsymbol{d}_m\}$ 优化的方向。为了获得更好的数值稳定性，取 $\boldsymbol{d}_\mu=\max\{\boldsymbol{d}_m\}$，则目标函数的梯度为：

$$\begin{cases}\nabla \boldsymbol{J}_m = \dfrac{\partial \boldsymbol{J}}{\partial \boldsymbol{d}_m} - \dfrac{\partial \boldsymbol{J}}{\partial \boldsymbol{d}_\mu} & \forall m \neq \mu \\ \nabla \boldsymbol{J}_\mu = \sum\limits_{m \neq \mu} \left(\dfrac{\partial \boldsymbol{J}}{\partial \boldsymbol{d}_\mu} - \dfrac{\partial \boldsymbol{J}}{\partial \boldsymbol{d}_m}\right) \end{cases} \tag{11}$$

则权重变量的更新法则为 $\boldsymbol{d} = \boldsymbol{d} + \gamma \boldsymbol{D}$。其中，$\gamma$ 为迭代步长，在梯度下降算法中，迭代步长的控制非常重要，关系到收敛的速度和精度。在这方面的研究也有很多，运用也较为成熟，一旦更新的方向确定，最优步长的确定采用的是一维线性搜索的方法：

$$\boldsymbol{D}_m = \begin{cases} 0 & \boldsymbol{d}_m = 0 \,\&\, \dfrac{\partial \boldsymbol{J}}{\partial \boldsymbol{d}_m} - \dfrac{\partial \boldsymbol{J}}{\partial \boldsymbol{d}_\mu} > 0 \\ \dfrac{\partial \boldsymbol{J}}{\partial \boldsymbol{d}_\mu} - \dfrac{\partial \boldsymbol{J}}{\partial \boldsymbol{d}_m} & \boldsymbol{d}_m > 0 \,\&\, m \neq \mu \\ \sum\limits_{m \neq \mu, d_v > 0} \left(\dfrac{\partial \boldsymbol{J}}{\partial \boldsymbol{d}_\mu} - \dfrac{\partial \boldsymbol{J}}{\partial \boldsymbol{d}_m}\right) & m = \mu \end{cases} \tag{12}$$

多核学习算法在每次更新权重系数后都需要利用支持向量数据描述重新计算凸二次优化，会耗费一定的计算量。但在每次解凸二次优化时可以利用上次计算的结果 α 进行初始化，从而大幅提高了计算的效率。

4.3　目标优化算法的终止迭代收敛条件和评估

根据以上的算法描述，迭代的终止条件可以设计为迭代前后优化目标值的变化量，并通过设定一个合适的阈值 ε 作为迭代终止的条件。阈值的选择应小于目标函数值的数量级，并根据精度的需要来设定，在本文中取 $\varepsilon = 10^{-5}$。

$$DG_{\mathrm{SVDD}} = \left\| \boldsymbol{J}(\boldsymbol{d}) + \sum_{i=1}^{N} \boldsymbol{\alpha}_i \sum_{m=1}^{M} \boldsymbol{d}_m \boldsymbol{k}_m \left(x_i, x_i\right) - \sum_{i=1}^{N} \sum_{j=1}^{N} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j \sum_{m=1}^{M} \boldsymbol{d}_m \boldsymbol{k}_m \left(x_i, x_j\right) \right\| \leqslant \varepsilon \tag{13}$$

然而，支持向量数据描述不同于 2 类分类支持向量机，可以用分类正确率来表征，目标函数 $\boldsymbol{J}(\boldsymbol{d})$ 刻画了超球空间包围目标数据的情况，如果不加以控制，容易出现极端的现象。作为分类器，在训练时完全不考虑异常样本是不科学的，常常会使得模型在训练过程中由于缺少异常类样本的参与，致使训练模型出现欠拟合或过拟合的现象。因此，必须对优化的过程进行控制和评估，从而使得优化后的多核支持向量数据描述能够具有很好的描述精度，同时也具有良好的泛化性能。

本文提出了一种基于目标数据驱动的多核支持向量权重优化评估方法，即通过引入含有异常类的测试样本来综合考察多核超球分类器的分类性能。对于支持向量数据描述来说，评价其模型的好坏最为有效的是 ROC 曲线，从以下 2 个方面来刻画模型的性能：(1)正常的样本被识别成异常，形成虚假报警，称为虚警率；(2)异常的样本被识别为正常，造成漏检，称为漏检率。在追求最小漏检率的条件下，对正常样本具有较高的正确检测率。在二维平面上体现为 ROC 曲线下方与坐标轴围成面积的大小(Area Under Curve, AUC)，所围成的面积越大则分类器的性能越好，因此，本文以此作为评价控制指标。以某次实验为例，迭代过程中目标函数值的下降曲线如图 1 所示，若仅以 $\boldsymbol{J}(\boldsymbol{d})$ 作为控制优化的目标，优化的变量 $\{\boldsymbol{d}_m\}$ 会迅速收敛到初始化的 $\{\boldsymbol{d}_m\}$ 中使优化目标 J 最小的核，同时观察相应的 ROC 曲线评价指标的变化情况。如图 2 所示，MSVDD 的性能并非逐步递增的过程，$\boldsymbol{J}$ 最小时 MSVDD 并非性能最佳，而通过跟踪 AUC 变化可以寻找出最优的权重组合。

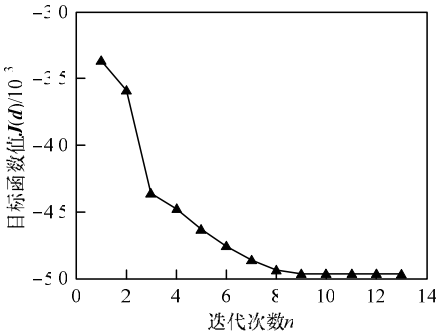


图 1　目标函数的变化过程

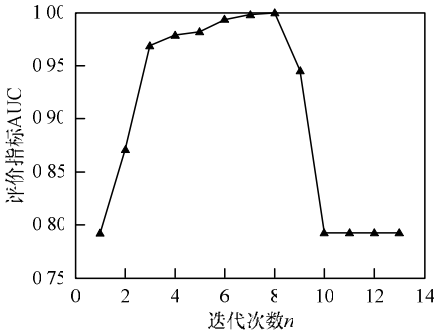


图 2　AUC 评价指标变化过程

5　应用实例

首先运用一个简单的例子描述组合多核支持向量数据描述的有效性。测试数据采用 prtool[8]工具箱产生的两维 2 类 Banana 数据集，用其测试多核支持向量数据描述的分类性能，目标类数据样本 30 个。多核支持向量数据描述由高斯和多项式核组成，高斯核宽度为 2，多项式核参数为 1，控制错分比例参数为 0.01，则高斯核支持向量数据描述效果如图 3 所示，而多项式核支持向量数据描述如图 4 所示。

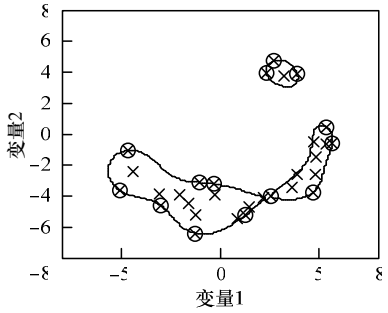


图 3　高斯核描述效果

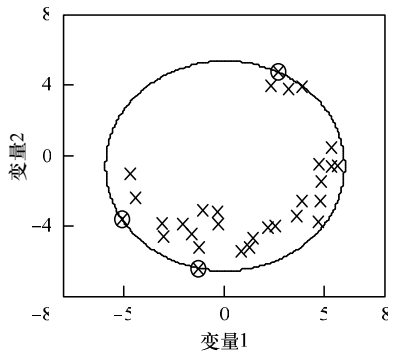


图 4 多项式核描述效果

对于 2 个图可以发现，高斯核函数具有较好的描述性能，但是由于参数的设置并非最优化，该描述的效果过于紧凑，描述精度较高但是损失了一定的推广能力。而多项式核描述则过于宽松，成为一个整圆，描述精度低，存在较大的漏检率。

图 5 显示了组合这 2 个核特征空间后的描述效果，取简单的组合权重为{0.85 0.15}，则该多核支持向量数据描述权衡了 2 种核函数对目标数据的拟合精度，在平面成较好的香蕉型数据分布形状，因而能够得到更高的推广能力和更小的误判率。

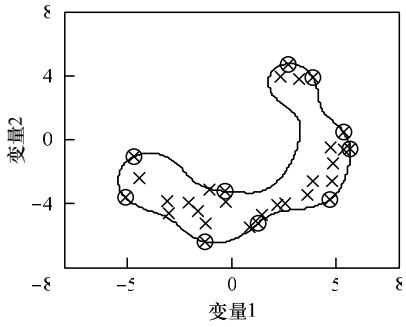


图 5 多核支持向量数据描述效果

实验时，产生 30 个目标类数据样本作为训练样本集和 200 个评价样本集。模型核函数选择高斯核和多项式核函数，核函数的初始参数集则设置以下常用数值的集合，分别为{3, 4, 5}和{1, 2}，因此，多核支持向量数据描述采用以上 5 个核特征空间的优化组合来表示。根据以上梯度下降方法寻求最优的权重，得{0.624 40, 0.353 50, 0.022 10}，由该组合权重即可以建立多核支持向量数据描述的模型。测试集中目标类数据 102 个，异常类数据 98 个，分类结果由 ROC 曲线来度量。

由图 6 可见，不同的核函数及不同的核参数下单类分类器的分类性能差异较大，随着参数的增大，单核的分类器性能变差，多项式核的分类性能变化更为明显。而建立在以上单核组合的基础上，多核分类器的数据描述能力明显优于任何一个单核分类器，能够获得比较理想的分类精度。表 1 则列出各单核及组合多核支持向量数据描述对测试数据集的实验结果，采用 ROC 曲线下方的面积作为评价指标，量化评价了各单类分类器的分类性能。

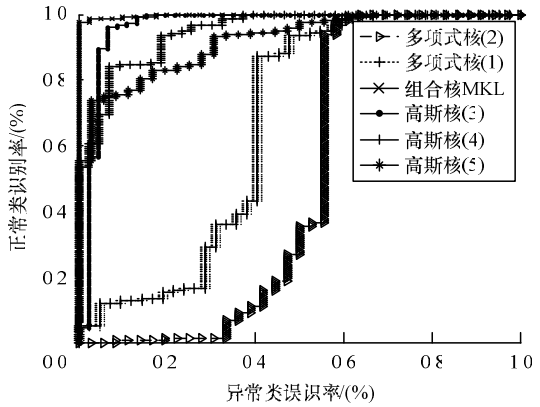


图 6 不同核函数和核参数下的 ROC 曲线

表 1 支持向量数据描述分类性能比较

核函数	核参数	评价指标
guass 核	3	0.953 6
	4	0.914 3
	5	0.847 7
ploy 核	1	0.689 5
	2	0.578 2
MKL	—	0.997 6

6 结束语

机械故障领域训练样本的获取存在严重不平衡问题，支持向量域描述单类分类器在故障诊断中得到广泛的应用。本文分析单核支持向量域描述存在的局限性，提出运用多个核函数映射的特征空间组合这个更加灵活的描述方式来建立单类分类器，并介绍了基于目标函数的梯度下降的方法来搜索最优的组合权重。在优化过程中引入了 ROC 曲线围成的面积作为控制搜索最优权重的评价指标，解决了支持向量域描述的核函数选择和参数优化问题，不仅能够充分利用原有的 SVDD 进行计算，而且具有更好的学习能力和计算效率。

参考文献

[1] 胡 桥, 何正嘉, 张周锁, 等. 基于提升小波包变换和集成支持矢量机的早期故障智能诊断[J]. 机械工程学报, 2006, 42(8): 16-23.

[2] 王自营, 邱绵浩, 安 钢, 等. 基于一类超球面支持向量机的机械故障诊断研究[J]. 振动工程学报, 2008, 21(6): 553-558.

[3] 李凌均, 韩 捷, 王 昆, 等. 基于单分类的机械故障诊断研究及其应用[J]. 机械强度, 2008, 30(5): 697-701.

[4] 张军峰, 胡寿松. 基于多重核学习支持向量回归的混沌时间序列预测[J]. 物理学报, 2008, 57(5): 2708-2712.

[5] Soren S, Gunnar R, Christin S, et al. Large Scale Multiple Kernel Learning[J]. Journal of Machine Learning Research, 2006, 7(1): 1531-1565.

(下转第 173 页)