第 37 卷 第 14 期
Vol.37 No.14

计 算 机 工 程
Computer Engineering

2011 年 7 月
July 2011

• 专栏 •

文章编号：1000—3428(2011)14—0018—03 文献标识码：A 中图分类号：O235

核化空间深度包围核的模糊决策异常检测算法

张思懿[a]，王士同[b]

(江南大学 a. 物联网工程学院；b. 数字媒体学院，江苏 无锡 214122)

摘 要：根据核化空间深度异常检测算法中适用性的局限性和最小包围核算法中存在参数影响检测效率的缺点，在引入模糊决策思想下，提出一种将上述 2 种算法相结合的模糊决策异常检测算法。融合后的算法将 2 种算法的优势相结合，并用模糊决策方法提高算法的稳定性和适用性。通过在人工数据集和 UCI 数据集上的实验结果表明，该算法具有较好的异常检测效果。

关键词：核化空间深度；最小包围核；核函数；异常检测；模糊决策

Fuzzy Decision Outlier Detection Algorithm of Kernelized Spatial Depth Sphere

ZHANG Si-yi[a], WANG Shi-tong[b]

(a. School of Internet of Things Engineering; b. School of Digital Media, Jiangnan University, Wuxi 214122, China)

【Abstract】This paper proposes a new idea using the idea of the fuzzy decision. And it is based on the kernelized spatial depth and the idea of the smallest sphere, intending for the problems that kernelized spatial depth function can not have good performance on some datasets and the parameters have the influence on the effectiveness. In this way, the algorithm improves the effectiveness and robustness in outlier detection by using the advantages of the algorithm and weakening the disadvantages. This paper does some experiments on the two artificial datasets and three different UCI datasets. Results show the effectiveness of the proposed idea.

【Key words】kernelized spatial depth; smallest sphere; kernelized function; outlier detection; fuzzy decision

DOI: 10.3969/j.issn.1000-3428.2011.14.005

1 概述

异常检测算法自 20 世纪 80 年代开始发展以来，已经取得了丰硕的成果[1-2]。目前，在异常检测领域研究算法有以下 2 种：基于统计的模型异常检测技术；基于机器学习的异常检测算法，它是用机器学习的方法来建立系统映像。它的最大特点是根据正常来分辨异常，因为其训练数据大多是代表清一色的正常行为。这种方法的优点是检测速度快，且误检率低。但该方法在用户动态行为变化以及单独异常检测方面还有待改善。复杂的相似度量和先验知识加入到检测中可能会提高系统的准确性。机器学习的方法主要包括：(1)基于数据挖掘的异常检测模型；(2)基于免疫学原理的异常检测模型；(3)基于实例的异常检测模型[3]。除此之外，为提高检测精度，同时保持算法复杂度在可接受的范围内，出现了基于特征模式的马尔可夫链异常检测模型[4]。本文在核化空间深度的异常检测算法及最小包围核算法的基础上，提出一种新的模糊决策异常检测算法。

2 核化空间深度的异常检测算法及最小包围核算法

2.1 核化空间深度的异常检测算法

核化空间深度异常检测算法[5]是将计算出的深度值与阈值相比，小于阈值的则可定为异常类。在该算法中，对于给定的核函数[6]来说阈值是仅有一个参数。本文给定基于深度判别的误判率的上界，而且这个上界可以决定此算法的阈值。

2.1.1 核化空间深度函数(KSD 函数)

$$D_k(x,X)=1-\frac{1}{|X\cup\{x\}|-1}\times\left(\sum_{y,z\in X}\frac{k(x,x)+k(y,z)-k(x,y)-k(x,z)}{\delta_k(x,y)\delta_k(x,z)}\right)^{\frac{1}{2}} \tag{1}$$

其中，$\delta_k(x,y)=\sqrt{k(x,x)+k(y,y)-2k(x,y)}$，类似地，当 $x=y$ 或 $x=z$ 时，导致空间函数为 0。

2.1.2 核函数中 σ 的选取

由文献[5]可知，由于在核函数 $k(x,y)$中参数 σ 的值将决定 KSD 函数全局和局部特点的均衡。正确地选择 σ 的值将使得 KSD 函数按照不同模式的几何特点来描绘数据的分布，从而达到不同的分类效果。下面考虑广泛使用的高斯核函数：

$$k(x,y)=\exp(-(x-y)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(x-y)) \tag{2}$$

其中，$\boldsymbol{\Sigma}=Diag[\sigma_1^2,\sigma_2^2,\mathrm{L},\sigma_n^2]$是一个对角矩阵。本文用依照数据的 k 维分布特征计算出 σ 的值组成 $\boldsymbol{\Sigma}$。在此提出 3 种分布的 $\boldsymbol{\Sigma}$ 的估计算法：

(1) $\boldsymbol{\Sigma}1:\sigma_k=mean_{i,j=1,2,\mathrm{L},l}\left|x_{ik}-x_{jk}\right|$，其中，$x_{ik}$ 和 x_{jk} 分别代表样本数据中 x_i 和 x_j 第 k 维的分量。

(2) $\boldsymbol{\Sigma}2:\sigma_k=median_{i,j=1,2,\mathrm{L},l}\left|x_{ik}-x_{jk}\right|$。

(3) $\boldsymbol{\Sigma}3:\sigma_k=mean_{i=1,2,\mathrm{L},l}\left|x_{ik}-mean_{j=1,2,\mathrm{L},l}x_{jk}\right|$。

这 3 种计算方法都有各自的优势，详细可参见文献[5]。

2.1.3 算法总结

对于上述的算法，输入是数据集 $X=\{x_1,x_2,\mathrm{L},x_l\}\in IR^n$，核函数和参数 t。这些数据都是有由 F_{good} 产生的，即参数 t 在决定一个数据是否为异类中起着决定性的作用。而参数 b 只有在训练数据的上界中用到，其目的也是为了选择合适的 t，下面用伪代码总结如下：

基金项目：国家自然科学基金资助项目(60773206)

作者简介：张思懿(1987－)，女，硕士研究生，主研方向：模糊决策异常检测算法，人工智能，模式识别；王士同，教授、博士生导师

收稿日期：2011-02-24 **E-mail：**wear560@126.com

1 FOR(every pair of x_i and x_j in X)

2 $K_{ij}=k(x_i,x_j)$

3 END

4 given input x

5 FOR (every observation x_i in X)

6 $\zeta_i = k(x,x_i)$

7 $\delta_i = \sqrt{k(x,x)+K_{ii}-2\zeta_i}$

8 IF $\delta_i = 0$

9 $z_i = 0$

10 ELSE

11 $z_i = \dfrac{1}{\delta_i}$

12 END

13 END

14 FOR (every pair of x_i and x_j in X)

15 $k_{ij} = k(x,x)+K_{ij}-\zeta_i-\zeta_j$

16 END

17 $D_k(x,X) = 1-\dfrac{1}{|X \cup \{x\}|-1}\sqrt{z^T k z}$

18 OUTPUT(x is an outlier if $D_k(x,X) \leqslant t$)

上述的伪代码是假设核函数 k 是已经给定的。而选择核函数对每一种核方法都是很重要的。对于高斯核函数来说，σ 决定了用来计算 KSD 的数据集的大小范围。

2.1.4 阈值 t 的选取

选择阈值的方法很简单，也就是要选择一个值能够控制 FAP(False Alarm Probability)保持低于特定的值即可。根据文献[5]中理论 1 可计算出阈值 t:

$$IE_{F|X}\left[g_k(x,X)\right] \leqslant \frac{1}{l}\sum_{i=1}^{l} g_k(x,X)+\frac{2}{lb}+(1+\frac{4}{b})\sqrt{\frac{In\dfrac{2}{\delta}}{2l}} \tag{3}$$

其中，异常检测的函数 $g_k(x,X)$ 如下:

$$g_k(x,X)=\begin{cases}1 & D_k(x,X) \leqslant t \\ \dfrac{t+b-D_k(x,X)}{b} & t < D_k(x,X) \leqslant t+b \\ 0 & \text{otherwise}\end{cases} \tag{4}$$

当 $g_k(x,X)=1$ 时，表示 x 被认为是异常数据。$IE_{F|X}$ 为 X 数据集在给定的积分函数下计算出的期望值，然后即可计算 FAP 的上界。

2.2 核化空间异常检测算法存在的问题

核化空间深度异常检测[5]是无监督的检测方法，它利用统计深度和核函数的概念进行检测，并取得了一定的分类效果，但它仍存在一些问题：(1)在 σ 的选取过程中，每一种 σ 都存在一定的误差，从而导致核函数的计算误差，最终对实验的结果有一定的影响。(2)深度函数是只对正类数据进行训练的无监督算法，导致这种算法对于正类异类差别很大的数据集效果很好，否则将影响检测效果。

2.3 最小包围核算法

文献[7]提出一种思想，最基本的观点就是要建立一个可以尽可能地将所有的正类数据都包围起来的一个具有尽可能小的半径的超球体(以下简称 ss 算法)，与此同时，这个超平面的表面要保持与异常数据之间尽可能大的距离。这样就可以形成紧密包围正常数据的边界。为了达到这个目的，只需用支持向量机[8]中解决凸集的优化方法，表达式如下:

$$\min_{R,c,\rho,\xi} R^2-\upsilon\rho^2+\frac{1}{\upsilon_1 m_1}\sum_{i=1}^{m_1}\xi_i+\frac{1}{\upsilon_2 m_2}\sum_{j=m_1+1}^{n}\xi_j \tag{5}$$

其中，

$$\left\|\phi(x_i)-c\right\| \leqslant R^2+\xi_i,\quad 1 \leqslant i \leqslant m_1 \tag{6}$$

$$\left\|\phi(x_i)-c\right\| \leqslant R^2+\rho^2-\xi_j,\quad m_1 < j \leqslant n \tag{7}$$

$$\xi_k \geqslant 0, 1 \leqslant k \leqslant n \tag{8}$$

其中，C 表示超球体的中心；R 表示超球体的半径；ρ 表示超球体与异常数据之间的距离；m_1 表示正类数据的个数；m_2 表示异常数据的个数；$\upsilon_1,\upsilon_2,\upsilon_3$ 是 3 个正的常数；$\boldsymbol{\xi}=[\xi_1,\xi_2,\cdots,\xi_n]^{\mathrm{T}} \in IR^n$ 是一组松弛变量。利用拉格朗日乘子以及二次规划的方法得到各个参数的值。然后根据式(9)对点进行判别:

$$f(x)=\mathrm{sgn}(R^2-\left\|\phi(x)-c\right\|^2) \tag{9}$$

2.4 最小包围核算法存在的问题

最小包围核异常检测算法[7]是将无监督的异常检测算法中引入很少量的异类数据训练的有监督的训练。但这个算法也存在不足之处，如：(1)算法受训练参数的影响较大，参数选择的不同识别结果有很大不同；(2)由于算法中引入了松弛变量使得算法产生了一定的误差。

3 本文方法

本文从以上 2 种算法的缺点出发，在 2 种方法的基础上，加入一种决策能力较好的模糊决策法来提高算法的稳定性。针对核化深度异常检测算法中不适用数据集中异类特征不明显的弱点和最小包围核中识别精度由于参数和松弛变量受到较大影响的缺点，引入模糊决策思想来提高算法的稳定性和识别率。

3.1 模糊决策的方法

文献[9]中提出了特征模糊子集的概念。在形式上形成关于 A 集合的关于 $u_1,u_2,\cdots,u_n$ 的剖分 $A^{(1)},A^{(2)},\cdots,A^{(n)}$。本文考察 $A^{(p)}(p \in J_n)$中每个对象所对应的模糊子集，可知：$u_p=(0,\cdots,0,1,0,\cdots,0)$(其中，1 是第 p 个分量)是诸子集中最体现 $A^{(p)}$ 特征的子集。或者说，通过 $D_{\%p}$ 对它所对应的对象进行决策时最无争议，即该对象无可争议地分在 u_p 类中。因此，把 $D_{\%p}$ 称作是 $A^{(p)}$ 的特征模糊子集。

按照通常的模糊决策方法，必须先得出决策对象 A_i 关于 β 在论域 $U=\{u_1,u_2,\cdots,u_n\}$ 上的模糊子集 $C_i=(c_{i1},c_{i2},\cdots,c_{in})$。因此，对任何 A_i 所对应的 C_i，都能通过 C_i 与 $D_{\%1}D_{\%2}\cdots D_{\%n}$ 的贴近度来决定 A_i 的归属。这种非对称贴近度的计算方法，可以使决策比较准确。而此时选用的贴近度的计算方法为:

$$N(A_{\%},B_{\%})=1-\frac{2}{n(n+1)}\sum_{k=1}^{n}\left|\mu_{A_{\%}}^{p}(u_k)-\mu_{B_{\%}}^{p}(u_k)\right|gk \tag{10}$$

3.2 算法框架及步骤

在模糊决策思想的引入下，本文提出的算法框架如下:

首先，利用空间深度异常检测算法对数据进行训练和测试并记录结果。其次，利用最小包围核的算法思想[7]对相同的数据集进行训练和测试并记录结果。最后，用一种模糊决策法[9]对判断不同的点进行最后的决策并记录最终的结果。

综上所述，本文算法的步骤如下:

(1)选取高斯核函数和 σ 的值，根据训练样本得到 sp 算法的核心参数 t、b 的值，然后对测试数据集按照 2.1.3 节中描述的算法对其进行分类，并记录分类结果 $R1$，与阈值 t 相比较可知是否为异类，将判断结果记录。

(2)选取高斯核函数，根据训练样本得到 ss 算法的参数值，对测试数据集按照步骤(5)进行判断，并记录分类结果 $R2$，

与阈值 0(根据数据集的不同，有时会用 $1e^{-n}$ (n 为正整数)代替)相比较可知是否为异类，将判断结果记录。

(3)筛选出分类结果中 $R1$ 和 $R2$ 判断不相同的数据部分 DIF_ DOT 和判断相同的数据部分 SAME_DOT。

(4)对于 DIF，设 distance1=$|R1-t|$, distance2=$|R2-0|$，将 distance1,distance2 作为该数据归属为哪一类的可能性(概率)。

(5)按照 3.1 节中描述的模糊决策思想，C=(distance1, distance2)，$D_{0\hat{1}}$=(1,0)(表示 C 的第 1 分量为最体现特征的子集)，$D_{0\hat{2}}$=(0,1)(表示 C 的第 2 分量为最体现特征的子集)。计算 $N_1=N(C, D_{0\hat{1}})$, $N_2=N(C, D_{0\hat{2}})$。

(6)比较 N_1、N_2 的值，$\begin{cases} N_1 \geqslant N_2 & R_1\text{正确} \\ N_1 < N_2 & R_2\text{正确} \end{cases}$，把 DIF_DOT 中所有数据都确定其归属类。

(7)在步骤(3)中保存的 SAME_DOT 与步骤(6)确定好分类的 DIF_ DOT 相加综合后就得到最终的判别结果。

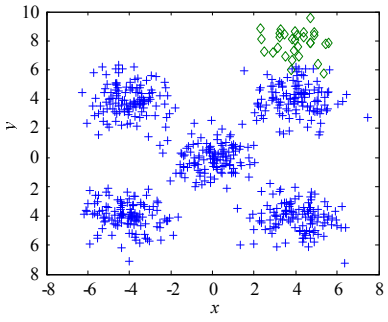
4 实验及分析

在实验中，为了进行综合比较，选择 2 个人工数据集和 3 个国际标准的 UCI 数据集，以验证本文提出的算法所提高的效率。

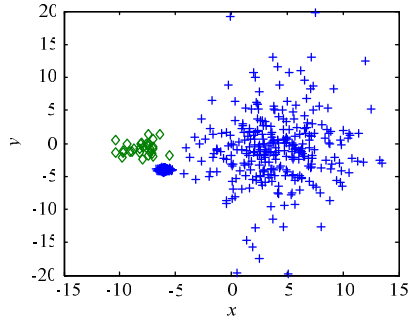
4.1 数据集简介

实验中用到的 2 个人工数据集均选取正类 600 个数据，负类 30 个数据。

第 1 类的数据的正类由以下这五部分满足高斯分布的随机数据组成，其中这五部分数据具有相同的权重 $N1$~$N([0, 0], I)$, $N2$~$([4, 4], I)$, $N3$~$([-4, 4], I)$, $N4$~$([4, -4], I)$, $N5$ $([-4, -4], I)$，负类是由满足高斯分布 $N([4, 8], I)$的随机数据组成，见图 1(a)。第 2 类数据的正类由以下几部分按照相同的权重组成 $N1$~$N([4, -1], [3.938, 2.923; 2.923, 7.313])$, $N2$~$N([-6, -4], [0.293\ 0.117; 0.117\ 0.158])$，负类满足高斯分布 $N([-8, -1], I)$，见图 1(b)。



(a)第 1 类随机数据集



(b)第 2 类随机数据集

图 1 人工数据集的图示分布

实验中用到的 UCI 数据集是 Shuttle、Ann-thyroid 和 Diabetes 数据集，其基本信息如表 1 所示。

表 1 UCI 数据集基本性质

数据集	特征数目	分类数目	样本总数目
Shuttle	9	5	5 012
Ann-thyroid	21	3	3 772
Diabetes	8	2	512

4.2 实验结果及分析

4.2.1 人工数据集

在实验中，第 1 类和第 2 类人工数据集对于 sp 算法选择 600 个正类的训练数据集合 600 个正类的验证数据集；对于 ss 算法选择了与 sp 算法相同的 600 个正类训练数据和 30 个异类训练数据集。测试数据集由 600 个正类数据和 30 个异类数据组成。对于 σ 的 3 种不同取值分别进行了实验，取平均结果进行记录。实验记录如表 2 所示。

表 2 人工数据集的识别率及参数选择

数据集	sp 识别率/(%)	ss 识别率/(%)	本文识别率/(%)	t	b
第 1 类	90.5	97.4	97.5	0.275	0.05
第 2 类	83.7	84.3	84.3	0.210	0.05

从这 2 组实验结果中体现了 sp 算法在针对不同数据分布时会产生较大的差异。而本文正是针对该缺点，增强了算法的适应性，通过实验结果可以明显地看出这一点。

4.2.2 UCI 数据集

在实验中，Shuttle 和 Ann-thyroid 数据集对于 sp 算法选择了数据集中 600 个正类的训练数据集和 600 个正类的验证数据集；对于 ss 算法选择了与 sp 算法相同的 600 个正类训练数据后又外加了 30 个异类的训练数据集。测试数据集由 200 个正类数据和 15 个异类数据组成。对于 σ 的 3 种不同取值分别进行了实验。对于 Shuttle 数据集将第 1 类和第 4 类作为正类，第 5 类作为异类进行实验；对于 Ann-thyroid 数据集分别进行了以数据集中的第 1 类作为异类的实验和以第 2 类作为异类的实验。Diabetes 数据集对于 sp 算法和 ss 算法的训练数据集都选择了 200 个正类数据和 30 个异类数据，测试数据集选择了 150 个正类数据和 20 个异类数据。

参数及实验结果如表 3~表 5 所示。

表 3 Shuttle 数据集算法识别率对比

数据集	σ	ss 识别率/(%)	sp 识别率/(%)	本文识别率/(%)	t	b
Shuttle (第 5 类为异类)	$\Sigma 1$	93.0	99.5	100.0	0.280 0	0.002
	$\Sigma 2$		81.9	87.0	0.291 0	0.001
	$\Sigma 3$		87.0	88.4	0.288 5	0.001

表 4 Ann-Thyroid 数据集算法识别率对比

数据集	σ	ss 识别率/(%)	sp 识别率/(%)	本文识别率/(%)	t	b
Ann-thyroid (第 1 类为异类)	$\Sigma 1$	93.0	92.6	92.6	0.274 0	0.005
	$\Sigma 2$		83.7	92.6	0.286 4	0.003
	$\Sigma 3$		88.4	92.6	0.286 6	0.002
Ann-thyroid (第 2 类为异类)	$\Sigma 1$	93.0	82.8	90.2	0.272 0	0.005
	$\Sigma 2$		86.0	90.2	0.285 9	0.002
	$\Sigma 3$		82.8	90.2	0.285 0	0.002

表 5 Diabetes 数据集算法识别率对比

数据集	σ	ss 识别率/(%)	sp 识别率/(%)	本文识别率/(%)	t	b
Diabetes	$\Sigma 1$	88.2	81.8	94.7	0.287 7	0.002
	$\Sigma 2$		84.1	94.7	0.290 0	0.005
	$\Sigma 3$		83.5	95.3	0.290 3	0.002

表 3 表明在不同 σ 选择值的情况下，对 sp 算法的识别率有较大影响。σ 在 $\Sigma 1$ 时的检测结果最好，相应地本文算法在 $\Sigma 1$ 处也取得最好的检测效率，这就说明本文的算法是基于这 2 种参考算法的，并且会受到这 2 种算法的影响，但在这 2 种算法的基础上又表现出较好的稳定性。

(下转第 26 页)