第 42 卷 第 7 期 天 津 大 学 学 报 Vol.42 No.7
2009 年 7 月 Journal of Tianjin University Jul. 2009

基于核函数优化的相符预测器在故障检测中的应用

杨 帆[1]，罗 键[1]，王华珍[1]，彭彦卿[1,2]，米 红[1]
（1. 厦门大学自动化系，厦门 361005；2. 厦门理工学院电子与电气工程系，厦门 361005）

摘 要：为了提高相符预测器的计算效率，在算法中引入基于核的度量学习. 将其学习过程分解成 2 部分：先通过提高 75%的训练样本的类可分性获得 1 个优化核；然后在优化的核空间中采用 k 近邻方法设计奇异度函数，并使用剩下的 25%的样本实现标准的相符预测器算法. 将新算法应用于田纳西-伊斯曼过程的多类故障诊断问题，实验结果表明，在保证高的预测效率的同时，新算法可以显著降低计算时间.

关键词：故障检测；相符预测器；度量学习；田纳西-伊斯曼过程
中图分类号：TP19 **文献标志码**：A **文章编号**：0493-2137(2009)07-0614-08

Optimized Kernel-Based Conformal Predictor for Online Fault Detection

YANG Fan[1], LUO Jian[1], WANG Hua-zhen[1], PENG Yan-qing[1,2], MI Hong[1]
(1. Department of Automation, Xiamen University, Xiamen 361005, China;
2. Department of Electronic and Electrical Engineering, Xiamen University of Technology, Xiamen 361005, China.)

Abstract: In order to improve the computational efficiency of conformal predictor, a procedure of adaptive kernel-based distance metric learning was incorporated in the algorithm. The learning process was divided into two stages. Firstly, an optimized kernel was obtained by increasing the class separability of 75% of the training samples. Secondly, the k nearest neighbor classifier was used to design a nonconformity measure function in the optimized kernel space. And then the standard conformal predictor algorithm was conducted on the remaining 25% of the training samples. The new method was applied to the multiple fault diagnosis of Tennessee Eastman process. The results show that the new algorithm provides substantial reductions in computational time, and ensures high predictive efficiency as well.

Keywords: fault detection; conformal predictor; distance metric learning; Tennessee Eastman process

随着大工业系统的集成度和复杂度不断提高，对产品生产的管理和控制提出了各种新的挑战[1-2]. 故障检测是大工业系统异常事件管理的核心问题. 由于故障检测具有模式识别的特点，目前已有许多学者研究出基于历史数据的故障检测算法[3-5]. 大多数机器学习算法采用离线学习方式，其首要的假设是已经获得了充分的历史数据用于构建分类模型，对当前数据进行预测. 现实条件中搜集大量故障数据往往代价高昂，故障检测算法应尽可能利用工业过程中的新信息，进行在线学习和推理.

与此同时，考虑到在线学习的时效性，必须提高算法的计算效率和速度，应尽可能少地利用新数据来重新训练分类器模型，尽量避免存储或者频繁访问大量历史数据，而代之以存储“历史知识”.

各种相关研究关注的重点往往放在提高算法的准确率和计算效率上，而轻视甚至忽视预测结果的可靠程度. 因此，各种判决需要高的可靠性. 然而当分类器判断一个样本的类别时，这个判断的准确性如何？一般的机器学习方法往往依据训练数据或先验分布假设，采用交叉验证（cross validation，CV）方法估计一个有偏的、方差较大的置信度. 相符预测器（conformal predictor，CP）[6-8]是最近发展起来的一种置信学习模型，它为每一次预测结果提供精确可控的置信度，具有理论保证的、严格的校准性（well-

收稿日期：2008-08-12；**修回日期**：2009-04-08.
基金项目：厦门大学 985 二期工程信息创新平台资助项目（0000–x07204）；厦门市科技计划资助项目（3502Z20083028）.
作者简介：杨 帆（1982— ），男，博士研究生，yang@xmu.edu.cn.
通讯作者：罗 键，jianluo@xmu.edu.cn.

calibrated)，预测的准确率恰好等于使用者预先设定的置信度，因而算法的错误风险具有预先的、完全的可控性. 与贝叶斯方法相比，CP 算法对数据的先验分布要求较弱，只要求满足独立同分布条件，因而适用范围广泛[9]. CP 算法在本质上是在线的，即测试样本依序出现，每一个测试样本在下一个测试样本到来前获得真实类别，并被加入到训练样本集中，因此特别适合于大工业系统的在线故障检测.

CP 算法将流行的机器学习方法引入到样本奇异度函数设计中，对包括测试样本和已知样本的样本集进行直推式学习，得到每一个样本的奇异值：测试样本对于每一类数据总体分布的相符或不一致程度. 现有的研究包括使用支持向量机(support vector machine，SVM)、核感知器(kernel perceptron，KP)、k 近邻(k nearest neighbors，kNN)和线性判别函数(linear discriminant classifier，LDC)等来设计奇异度函数，从而在 CP 框架下产生了 CP-SVM、CP-KP、CP-kNN、CP-LDC 等算法. 从另一个角度来说，也可以认为 CP 为这些机器学习算法的预测结果提供了一个可靠性评估(hedging predictions)[7].

上述算法有其各自不同的本质缺陷. 其共同缺陷在于：大多在原空间中设计奇异度函数，如 CP-kNN、CP-NC 等，而原空间的距离度量往往无法适应大多数的实际问题需要；少数算法如 CP-SVM 和 CP-KP 通过选定一个核函数，在核空间中设计奇异度函数，但核参数的选取缺乏指导. 算法把主要的计算代价放在了在线推理上，除了 CP-kNN 外，大部分算法面对在线问题时，每次预测都要利用待测样本和所有历史样本进行直推，需要不断重新训练分类器来计算所有样本的奇异值，严重影响计算效率.

CP-kNN 也存在两个问题，一是距离度量难以确定，原空间的欧式距离或其他距离不适应大多数分类问题的需要，尤其对于高维、小样本的情形分类效果较差；二是需要存储大量的历史数据，在推理过程中频繁访问和寻找 k 近邻，大大降低了计算效率.

由此可见，在利用相符预测器进行在线的直推式学习时，需要存储和频繁访问大量数据，并不断重新训练分类器，这对于大数据集问题，如大工业过程，是难以承受的. 因此，目前对于 CP 算法的研究，主要集中在计算效率和预测效率之间的折中：Papadopoulos 等[10]提出归纳式相符预测器(inductive conformal predictor，ICP)，先利用部分样本离线学习一个归纳模型，并采用这个固定的模型来计算其他所有样本的奇异度，从而把部分的在线计算代价转移到离线学习上，但由于损失大量在线信息，预测效率比 CP 差；Ho 等[11]提出主动学习相符预测器(active learning conformal predictor，AL-CP)，根据样本属于各类的置信度选择满足条件的新样本加入学习序列，控制数据规模，但 Vovk 等认为[7]这种主动学习破坏了序列的随机性，不再满足算法随机性理论，不仅预测效率降低，而且校准性也不再严格满足，因此并不可取.

本文的研究目的在于，在提高 CP 算法在线运算效率的同时，保持算法的预测效率和性能. 为此，引入基于核空间的度量学习，得到基于核函数优化的相符预测器(optimized kernel-based conformal predictor)，由于采用 kNN 来设计样本奇异度函数，简称其为 CP-KerNN.

为比较算法的有效性、稳定性和可靠性，采用工业过程领域使用的基准平台——田纳西-伊斯曼过程(Tennessee Eastman process，TEP)[3,5]进行测试.

1　相关研究

以工业系统的在线故障检测为例，研究对象产生训练样本序列 $(x_1,y_1),\cdots,(x_{n-1},y_{n-1})$ 和一个未知类别的待测数据 x_n，$x_i\in X, y_i\in Y$，$i=1,2,\cdots,n$，$n\in N$；其中 X 和 Y 分别代表属性空间和类别空间，并记样本空间为 $Z=X\times Y$. 不断有新的待测样本加入，其真实标签在下一个待测样本到来之前给出，本文将上述在线学习问题标记为 E.

1.1　算法随机性理论与相符预测器

CP 是一种基于 Kolmogorov 算法随机性理论的直推式(transductive inference)学习机器，即对包括待测样本和已知样本的样本序列进行随机性检验，给出序列符合假设分布(独立同分布)的量化估计，这个估计值就是待测数据被正确分类的置信度. 预先指定一个置信度水平，算法把所有大于该置信度的分类作为预测结果，实现域预测(region prediction)，而不是传统机器学习算法输出的点预测(point prediction). CP 算法的一个核心问题是样本奇异值度量(nonconformity measure)，即每一个样本对于每一类数据总体分布的一致或不一致程度，从样本空间到样本奇异值(实数空间)的映射被称为样本奇异度函数，通过它计算包括待测样本和已知样本在内的每一个样本的奇异值.

定义 1[6]　函数 $A_n: Z^{(n-1)}\times Z_n\to R$ 是一个样本奇异度函数，产生的样本奇异值(nonconformity score)为

$$\forall n\in N,\alpha_i := A_n(z_i,\{z_1,\cdots,z_{i-1},z_{i+1},\cdots,z_n\})$$

$$i=1,2,\cdots,n-1$$
$$\alpha_n := A_n(z_n, \{z_1,\cdots,z_{n-1}\}) \tag{1}$$

式中：$\{\}$称为数据包，其包含的元素可以任意交换位置；α_i称为样本z_i对于某类数据分布的奇异值，且α_i与该样本在序列中的位置无关. 如果待测样本的奇异值相对于某一类训练样本的奇异值较高，则待测样本隶属该类的置信度就较低.

定义 2[6] 给定学习问题 E，当使用者给定置信度$1-\varepsilon$和样本奇异度量函数A_n，一个平滑的相符预测器输出该风险水平ε下的域预测Γ^{ε}为

$$\Gamma^{\varepsilon,\tau_n}(z_1,\cdots,z_{n-1},x_n,\tau_n)=\{y\in Y: p_y=\frac{|\{i=1,\cdots,n:\alpha_i>\alpha_n\}|+\tau_n|\{i=1,\cdots,n:\alpha_i=\alpha_n\}|}{n}>\varepsilon\} \tag{2}$$

式中：y是x_n的所有可能类别；p_y是随机性检验值，称为 p 值，衡量y是待测数据真实类别的可靠程度；τ_n服从[0,1]上的均匀分布.

由式(2)可知算法的输出是某个风险水平下含有多个可能类别的域预测. 如果待测样本x_n的真实类别包含在预测集中，则算法预测正确，否则预测错误.

定理 1[6] 相符预测器满足可校准性，即

$$\limsup_{n\to\infty}\frac{E_n^{\varepsilon}}{n}\leqslant\varepsilon \tag{3}$$

式中E_n^{ε}表示风险水平ε下n次预测中错误的次数，证明略. 由于算法的风险是可控的，预测的置信度是可靠的，因此人们关心的是在给定ε后，预测集的规模. 在同一个置信水平下，预测集规模应尽量小，这时候选类别的个数少，提供给使用者的信息量大，能够更明确地选出真实的类别. 这被称为 CP 算法的预测效率问题，通常采用以下指标来评价.

(1)确定率(certain prediction rate)，指含单个元素的域预测所占比率.

(2)不确定率(uncertain prediction rate)，指含有多于一个元素的域预测的比率.

(3)空集率(empty prediction)，空集的比率. 本文进一步提出了“偏好率”，即仅含单个元素且预测准确的域预测的比率.

1.2 *k* 近邻分类器

k 近邻分类器是模式识别领域一种广泛应用的直推式方法[12]. 在 CP 框架中，利用 *k*NN 设计样本奇异度函数简单直观，不需要构建和保存模型，适用于训练样本集不断扩大的在线学习；但它又是一种懒惰式学习策略，需要保存所有的历史数据，计算代价被放在了每次的预测过程，并频繁访问和寻找近邻，加重了在线学习的负担；同时依赖于所采用的距离度量，由于缺少数据分布信息，实际应用中一般采用欧氏距离，预测效果受到限制.

采用 *k*NN 设计的样本奇异度函数为

$$\alpha_i=\frac{\sum_{j=1}^{k}d_{ij}^{-y}}{\sum_{j=1}^{k}d_{ij}^{y}}\qquad i=1,2,\cdots,n,j\neq i \tag{4}$$

式中：$\sum_{j=1}^{k}d_{ij}^{-y}$表示类别不为$y$的所有样本中，在欧式空间中与样本$i$最近的 k 个样本与其邻近度之和；$\sum_{j=1}^{k}d_{ij}^{y}$表示类别为y的所有样本中，与样本i最近的k个样本与其邻近度之和.

1.3 基于核空间的度量学习

能否准确反映样本之间的相似或相异程度，将极大影响分类器的效果，尤其是对于 *k*NN 这样的基于相似度或相异度评价的学习机器. 解决手段之一是提出各种度量学习(distance metric learning)[13]，从输入数据的信息中学习一个好的度量或映射，由于这个映射是从输入数据的分析中得来的，因此被称为经验特征空间(empirical feature space). 核映射的选择改变数据在特征空间中的几何机构，因而被广泛使用. 一般的核方法如 SVM 等往往采取交叉验证来选取核函数，操作性差，缺乏理论指导. 为此，人们提出基于核的度量学习，即通过优化训练样本的某个目标函数值来寻找一个最优的核函数[14].

2 CP-KerNN 算法

2.1 算法原理

本文将训练样本集划分为压缩样本集和校验样本集 2 部分，并将学习过程划分 2 个阶段.

1)离线学习阶段

引入基核映射 K_0，将压缩样本集投入核空间，以样本在核空间中的类可分性为目标函数，优化的目标是数据的类分开性更好. 通过对压缩样本集的度量学习得到一个优化的核空间 K_1. 该阶段的目的是，将部分数据压缩成知识，降低存储规模，转移在线计算代价，该部分的数据将不再保存.

2)在线学习阶段

将校验样本集和待测样本投入优化的核空间K_1，通过核映射自适应地调整和重新计算样本间的相似度. 由第 1 阶段压缩样本集得到的优化核具有一定的泛化性，不仅增大了压缩集的类分开性，也提高了校验样本集和待测样本在该空间中的可分性，据此

设计的奇异函数更有效，预测效率将得到提高. 同时，在线计算的代价显著降低.

与同样采取 2 阶段学习策略的 ICP 相比，本文方法的特点在于，压缩样本集的信息实际上是以间接方式提供给奇异度函数，样本奇异值的计算仍然是在线的；而 ICP 的样本奇异度函数完全通过第 1 阶段的数据离线学习而决定，是固定的，忽略了不断加入的在线信息. 从这个角度来说，CP-KerNN 由于比 ICP 更多地考虑了在线信息，因此计算效率会略差.

在优化后的核空间中，同样可以采取各种机器学习方法来设计奇异度函数. 由于 *k*NN 简单直观，无需重新训练分类器，适于在线学习，仍采用它来设计奇异度函数，称为 CP-KerNN（Kernel *k*NN）. 可以看出，CP-KerNN 与 CP-*k*NN 不同之处在于前者在优化后的核空间中进行在线运算，而后者是在原空间中. 由于在线运算过程时所采用的奇异度函数仍然基于 *k*NN 原理设计，满足定义 1 的要求，因而也能满足定理 1，从而保证了算法的可校准性[6,8].

2.2　基于输入数据的核函数优化

适用于度量学习的核函数一般定义为

$$k(x_i,x_j)=q(x_i)q(x_j)k_0(x_i,x_j) \tag{5}$$

其中 $x_i,x_j\in X$，$i,j=1,2,\cdots,n$；$k_0(x_i,x_j)$ 为基核，本文取 $k_0(x_i,x_j)=\mathrm{e}^{-\gamma_0\|x_i-x_j\|^2}$，即高斯径向基函数. $q(\cdot)$ 为优化因子函数，形式为

$$q(x_i)=\omega_0+\sum_{j=1}^{n}\omega_j k_1(x_i,x_j) \tag{6}$$

式中：$k_1(x_i,x_j)=\mathrm{e}^{-\gamma_1\|x_i-x_j\|^2}$；$\omega_0$ 和 ω_j 为优化组合系数. 这种定义使得用于度量学习的核函数 $k(x_i,x_j)$ 仍满足 Mercer 定理.

上述核函数可用矩阵形式表示为

$$\boldsymbol{K}=\boldsymbol{Q}\boldsymbol{K}_0\boldsymbol{Q} \tag{7}$$

式中：$\boldsymbol{K}=[k(x_i,x_j)]_{n\times n}$；$\boldsymbol{K}_0=[k_0(x_i,x_j)]_{n\times n}$；$\boldsymbol{Q}=[q(x_1),q(x_2),\cdots,q(x_n)]^{\mathrm{T}}$. $\boldsymbol{W}=(\omega_0,\omega_1,\cdots,\omega_n)^{\mathrm{T}}$，于是可得 $\boldsymbol{Q}=\boldsymbol{K}_1\boldsymbol{W}$，且有

$$\boldsymbol{K}_1=\begin{pmatrix}1 & k_1(x_1,x_1) & \cdots & k_1(x_1,x_n)\\ \vdots & & & \vdots\\ 1 & k_1(x_n,x_1) & \cdots & k_1(x_n,x_n)\end{pmatrix}_{n\times(n+1)} \tag{8}$$

核函数优化过程本质上是优化参数 $\boldsymbol{W}$，优化的目标函数是样本集在特征空间的类分开性函数，即

$$\boldsymbol{J}=\frac{\mathrm{tr}(\boldsymbol{S}_{\mathrm{b}})}{\mathrm{tr}(\boldsymbol{S}_{\mathrm{w}})} \tag{9}$$

式中 $\boldsymbol{S}_{\mathrm{b}}$ 和 $\boldsymbol{S}_{\mathrm{w}}$ 分别为类间和类内离散矩阵. 当衡量两类数据（数据量 $n=n_1+n_2$）之间的可分性时，将基核矩阵 $\boldsymbol{K}_0$ 按照样本的类别调整并划分成 4 个子矩阵，即

$$\boldsymbol{K}_0=\begin{pmatrix}\boldsymbol{K}_{11}^0 & \boldsymbol{K}_{12}^0\\ \boldsymbol{K}_{21}^0 & \boldsymbol{K}_{22}^0\end{pmatrix} \tag{10}$$

其中各子矩阵 $(\boldsymbol{K}_{11}^0)$、$(\boldsymbol{K}_{12}^0)$、$(\boldsymbol{K}_{21}^0)$和$(\boldsymbol{K}_{22}^0)$ 的规模分别为 $n_1\times n_1$、$n_1\times n_2$ 、$n_2\times n_1$和$n_2\times n_2$. 类分开性表达式变换为

$$\boldsymbol{J}(\boldsymbol{W})=\frac{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}_0\boldsymbol{W}}{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{N}_0\boldsymbol{W}}=\frac{\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_0\boldsymbol{Q}}{\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}_0\boldsymbol{Q}}=\frac{\boldsymbol{J}_1}{\boldsymbol{J}_2} \tag{11}$$

其中 $\boldsymbol{M}_0=\boldsymbol{K}_1^{\mathrm{T}}\boldsymbol{B}_0\boldsymbol{K}_1$，$\boldsymbol{N}_0=\boldsymbol{K}_1^{\mathrm{T}}\boldsymbol{D}_0\boldsymbol{K}_1$，且有

$$\boldsymbol{B}_0=\begin{pmatrix}\frac{1}{n_1}\boldsymbol{K}_{11}^0 & \boldsymbol{0}\\ \boldsymbol{0} & \frac{1}{n_2}\boldsymbol{K}_{22}^0\end{pmatrix}-\frac{1}{n}\boldsymbol{K}_0 \tag{12}$$

$$\boldsymbol{D}_0=\mathrm{diag}(\boldsymbol{K}_0)-\begin{pmatrix}\frac{1}{n_1}\boldsymbol{K}_{11}^0 & \boldsymbol{0}\\ \boldsymbol{0} & \frac{1}{n_2}\boldsymbol{K}_{22}^0\end{pmatrix} \tag{13}$$

本文采用梯度下降法进行迭代优化，需要设置的参数有迭代次数 m、学习率 η、基核参数 γ_0 和优化因子参数 γ_1. 对于多分类问题，可采用一对一的方法分别构建每两类间的类分开性度量数据，在每一次迭代过程选择类分开性度量值最小的两类样本数据进行优化，核函数的优化过程如下.

输入：训练样本 $(x_1,y_1),\cdots,(x_{n-1},y_{n-1})$，参数 m、η、γ_0、γ_1.

输出：优化核参数 $\boldsymbol{W}=(\omega_0,\omega_1,\cdots,\omega_n)^{\mathrm{T}}$.

步骤 1　构建样本的基核函数矩阵，根据类别调整并划分为各种子矩阵 $\boldsymbol{K}_0^{ij}$，$i,j=1,2,\cdots,c$；计算相应的 $\boldsymbol{B}_0^{ij}$和$\boldsymbol{D}_0^{ij}$.

步骤 2　初始化组合参数 $\boldsymbol{W}^{(0)}=(1,0,\cdots,0)^{\mathrm{T}}$ 和学习率 $\eta(0)$.

步骤 3　计算一对一方法构成的每两类之间的分开性度量值 $\boldsymbol{J}^{ij}(\boldsymbol{W}^t)$，$i,j=1,\cdots,c$.

步骤 4　选择 $(u,v)=\arg\min(\boldsymbol{J}^{ij})$，计算 $\boldsymbol{M}_0^{u,v}$、$\boldsymbol{N}_0^{u,v}$、$\boldsymbol{J}_1^{u,v}$、$\boldsymbol{J}_2^{u,v}$.

步骤 5 按式

$$\boldsymbol{W}^{t+1}=\boldsymbol{W}^t+\eta(t)\frac{2}{(\boldsymbol{J}_2^{u,v})^2}(\boldsymbol{J}_2^{u,v}\boldsymbol{M}_0^{u,v}-\boldsymbol{J}_1^{u,v}\boldsymbol{N}_0^{u,v})$$

更新 $\boldsymbol{W}^t$，并对 $\boldsymbol{W}^{t+1}$ 进行归一化，使得 $\|\boldsymbol{W}^{t+1}\|=1$，其中 $\eta(t)=\eta(0)(1-\frac{t}{n})$.

步骤 6 判断迭代次数是否到 m，否则转步骤 3.

2.3 CP-KerNN 算法

当学习问题为 E 时，将样本集划分成两个部分，选取 $m_k(m_k<n)$ 个数据作为第 1 阶段的压缩数据，即压缩样本集；剩余的 $n-1-m_k$ 个样本用作在线学习的初始训练样本，即校验样本集. 在算法过程中，首先采用算法 1 和压缩样本集得到一个经验特征空间，将 m_k 个数据以经验知识的形式压缩在核映射模型 M 中；在第 2 阶段，利用核映射模型 M、校验样本集以及待测样本 z_n，定义奇异度函数如下.

定义 3

$$\begin{aligned}\alpha_i &:= A_n(z_i, \{z_{m_k+1},\cdots,z_{i-1},z_{i+1},\cdots,z_n\}, M)\\ &\quad i=m_k+1,\cdots,n-1\\ \alpha_n &:= A_n(z_n, \{z_{m_k+1},\cdots,z_{n-1}\}, M)\end{aligned} \tag{14}$$

预测输出为

$$\begin{aligned}\Gamma_M^{\varepsilon}(z_1,\cdots,z_{n-1},x_n,\tau_n,m_k) = \{y\in Y:\ p_y = \\ \frac{|\{i=m_k+1,\cdots,n:\alpha_i>\alpha_n\}|}{n-m_k} + \\ \frac{\tau_n|\{i=m_k+1,\cdots,n:\alpha_i=\alpha_n\}|}{n-m_k} > \varepsilon\}\end{aligned} \tag{15}$$

在下一个待测样本到来之前，得到 x_n 的真实标签，并加入到校验样本集中.

将 kNN 引入优化核空间中计算样本奇异值时，变形为

$$\alpha_i = \frac{\sum_{j=1}^{k}\boldsymbol{\theta}_{ij}^{-y}}{\sum_{j=1}^{k}\boldsymbol{\theta}_{ij}^{y}} \qquad i=m_k+1,\cdots,n, j\neq i \tag{16}$$

式中：$\sum_{j=1}^{k}\boldsymbol{\theta}_{ij}^{-y}$ 表示序号为 $m_k+1,\cdots,n$、类别不为 y 的所有样本中，与样本 i 最近的 k 个样本与其邻近度之和；$\sum_{j=1}^{k}\boldsymbol{\theta}_{ij}^{y}$ 表示序号为 $m_k+1,\cdots,n$、类别为 y 的所有样本中，与样本 i 最近的 k 个样本与其邻近度之和. 当样本在经验特征空间可分性良好时，分子的值比分母的值小得多，奇异值就很小，即样本偏离数据分布的奇异程度很小. 算法过程描述如下.

输入：$(x_1,y_1),\cdots,(x_{n-1},y_{n-1})$，待测数据 x_n，参数 m、η、γ_0、γ_1、m_k、k.

输出：x_n 的域预测 Γ^{ε}.

步骤 1 调用第 2.2 节中的算法，利用 $(x_1,y_1),\cdots,(x_{m_k},y_{m_k})$ 进行核函数优化，得到核映射 M.

步骤 2 通过核映射 M 计算数据 $\{x_{m_k+1},\cdots,x_n\}$ 的核矩阵 $[\boldsymbol{\theta}(i,j)]_{(n-m_k)\times(n-m_k)}$.

步骤 3 初始化预测集 $\Gamma^{\varepsilon}=\varnothing$ 和校验样本集

$$V=\{(x_{m_k+1},y_{m_k+1}),\cdots,(x_{n-1},y_{n-1})\}$$

步骤 4 for j=1 to c（c 为样本的类别数）

指定 j 为 x_n 的假设类别，根据式(16)计算 $\{(x_{m_k+1},y_{m_k+1}),\cdots,(x_{n-1},y_{n-1}),(x_n,j)\}$ 的奇异值；根据式(15)计算 (x_n,j) 的 p 值，得到 p_n^j. 若 $p_n^j>\varepsilon$，则 $\Gamma^{\varepsilon}=\Gamma^{\varepsilon}\cup\{j\}$.

end for j

步骤 5 获得 x_n 的真实类别 y_n，扩充校验样本集 $V=V\cup\{(x_n,y_n)\}$.

随着在线学习的不断进行，校验样本集 V 增大到一定程度后，算法的运算效率又会降低；同时，故障数据的分布也可能发生“漂移”，由压缩样本建立的经验核的推广性下降，更重要的是，数据的原始分布发生改变，将会破坏算法随机性. 为此，可以移除离当前时刻点较远的训练数据，仅保留较近时刻的历史数据，并重新划分压缩样本集和校验样本集，生成新的 CP-KerNN 模型. 需要指出的是，在新模型的在线运算过程中，只要新数据样本的产生满足算法随机性，新的模型依然严格满足可校准性.

2.4 时间复杂度分析和各种 CP 算法的比较

为了便于比较，不考虑用来压缩信息和设计奇异度函数的模型的具体形式.

(1) CP 算法必须首先对规模为 n 的样本序列构建奇异度算法模型(如 SVM)，所需的时间复杂度记为 $B_{\text{train}}(n)$，而由奇异度算法模型作用在 n 个样本上得到样本奇异性度量，所需的时间复杂度记为 $B_{\text{apply}}(n)$；该过程要重复 c 次(类别数)，因此经典 CP 算法的时间复杂度可表示为 $c[B_{\text{train}}(n)+B_{\text{apply}}(n)]$.

(2) ICP 算法首先对规模为 m_k 的训练样本构建奇异度算法模型，所需的时间复杂度记为 $B_{\text{train}}(m_k)$，由奇异度算法模型作用在 $n-m_k$ 个剩下样本上得到样本奇异性度量，时间复杂度记为 $B_{\text{apply}}(n-m_k)$，由于 ICP 针对校验样本不再重复训练奇异度算法，因此总的时间复杂度为 $B_{\text{train}}(m_k)+cB_{\text{apply}}(n-m_k)$.

(3) 本文所提算法首先将规模为 m_k 的训练样本的信息以模型的形式压缩和提取，所需的时间复杂度记为 $O_1=I_{\text{train}}(m_k)$；由该算法模型将信息传递给后续 $n-m_k$ 个检验样本，时间复杂度记为 $O_2=I_{\text{apply}}(n-m_k)$；在后续个 $n-m_k$ 检验样本上实现标准 CP 算法，该过程所需的时间复杂度记为 $O_3=c\cdot[B_{\text{train}}(n-m_k)+B_{\text{apply}}(n-m_k)]$，因此总的时间复杂度为 $O_1+O_2+O_3$.

由以上分析可知，ICP 算法复杂度最小；本文所提算法的时间复杂度与 CP 相比，多了 O_1 和 O_2 两项，但由于一般情况下 $n-m_k$ 比 n 小得多，因此 O_3 耗费时间要比 CP 小得多. 另外，由于只需 O_1 和 O_2 部分运行

1 次，其时间开销非常小，与在线学习相比可忽略不计. 另外，在实际应用中，O_1 部分可以看作是预计算，对在线学习而言其时间复杂度便可以忽略不计；由 m_k 个训练样本引起的空间复杂性也可忽略，大大降低了算法的空间复杂度.

各种 CP 算法性能对比见表 1.

表 1　CP算法性能对比

Tab.1　Performance comparison of different CP

算法性能	CP-			AL-CP	ICP
	KerNN	*k*NN	SVM		
需重新训练分类器	否	否	是	是	否
特征空间	优化核	原始	基核	原始	原始/基核
运算效率	高	低	低	高	高
预测效率	高	高	高	低	低
满足算法随机性	是	是	是	否	是

理论分析说明，CP-KerNN 的计算效率虽然比 ICP 稍差，但一定优于标准 CP 算法；AL-CP 和 ICP 的预测效率又比标准 CP 算法差. 因此实验部分将只需说明 CP-KerNN 的预测效率不差于标准 CP 算法，即与 CP-*k*NN 相比较.

3　实验与讨论

3.1　样本集创建和实验设置

田纳西-伊斯曼过程设有 52 个监测器，它检测生产过程的压强、温度等. 当某一种故障发生时，52 个监测器同时采样一段时间，累积的样本可以充分表现故障的分布特性. 相关研究指出[3-5]，当采样间隔设置为 3 min 时，每类故障累积的样本量不能少于 480 个. 本文针对 4 种生产过程中的典型故障进行研究，采样间隔为 3 min，每类的采样样本量为 800（480 用于训练，320 用于测试）；实验数据共有 5 个类别，包括 4 个故障类和 1 个正常类，因此 TEP 仿真器将产生 4 000 个 52 维的样本数据. 受篇幅限制，TEP 仿真系统模型和数据生成过程在此不再赘述. 表 2 简要描述了本文研究的 4 种典型故障，详细的机理可参见文献[15].

表 2　TEP 4种典型故障描述

Tab.2　Description of four kinds of TEP fault diagnosis

故障类别	故障描述	类 型
1	进料 1/进料 3 的比率，进料 2 成分不变	阶跃
2	反应器冷却水的入口温度	阶跃
3	进料 4 的进料温度	随机变量
4	反应器冷却水的入口温度	随机变量

实验的第 1 部分用于说明优化核函数的效果，即数据集被映射到不同空间中的类分开性；第 2 部分检验 CP-KerNN 预测置信度的有效性；第 3 部分比较欧式空间、基核空间和优化核空间中采用 *k*NN 设计奇异度时，算法的预测效率.

实验构建了 3 个样本集：压缩集 T、校验集 V 和测试集 S. 将 4 000 个原始数据随机分配到 3 个样本集中，总的训练样本数目为 480×5=2 400，其中压缩集 T 的规模为 1 800，校验集 V 的规模为 600，即压缩比例为 3/4. 测试样本 S 的数目为 320×5=1 600. 重复实验 20 次，所有实验结果均为平均值.

在学习的第 1 阶段，为了避免核优化时间耗费过多，迭代次数设为 m =1 000，初始学习率 $\eta(0)$=0.01. 采用网格法选取核参数，基核参数候选值 γ_0 分别取{1.0，0.8，0.4，0.1，0.08，0.04，0.01}，优化因子参数 γ_1 分别取{0.9，0.6，0.1，0.01，0.001}，评判标准为校验样本集 V 的类分开性度量 J，根据实验数据，最终选取 γ_0=0.8，γ_1=0.9（过程略）.

当采用 *k*NN 设计奇异度函数时，近邻数 k 的候选值为{1，3，9，21，51，99}.

3.2　数据集在不同空间里的类分开性

由表 3 可以看出，压缩集映射到基核函数 K_0 后，类可分性有了很大提高；通过核函数优化，可进一步提高类分开性，如果精细调整优化算法参数，则效果可能更加明显.校验集 V 和测试集 S 的类分开性的变化情况和压缩集一致，说明核函数优化算法具有很好的泛化性. V 和 S 在核空间中的类分开性好于原始欧氏空间，将提高在线学习的预测效率.

表 3　数据集在不同空间中的类分开性

Tab.3　Class separabilities of the data set in different spaces

数据集	欧氏空间	基核空间	优化核空间
T	0.42	0.51	0.55
$V+S$	0.28	0.44	0.48

3.3　CP-KerNN 的置信预测及其校准性

表 4 为针对某一个真实类别是 1 的样本进行的随机性检验，其中近邻数 k=1. 由表 4 可知，除了类别 1 所对应的 p 值（0.930）较大外，其他类别对应的 p 值都较小，说明样本在类别为 1 的校验数据序列中的随机性水平较高，即隶属于第1 类的置信度较高. 最大 p 值与其他p值之间的差距越大，则对数据分布越敏感，奇异度函数越能反映出样本在分布中的异常程度.

从表 4 可以分析在不同置信度1−ε下算法的输出结果. 当预设的置信度在 86.3%以上时，预测集包含 2 个可能的真实类别：1 或 2，此时为不确定预测；当置信度在区间[7.0%，86.3%]时，给出唯一的输出

表4　待测数据的随机性检验(真实类别为1)

Tab.4　Randomness testing of a test example in class 1

候选类别	p值
故障1	0.930
故障2	0.137
故障3	0.012
故障4	0.086
正常	0.007

结果,此时为确定性预测;当置信度在 7.0%以下时,预测集的规模为零,此时为空预测,样本被提交给专家识别. 预设的风险水平越小,域预测的规模越大,反之亦然.根据需要设置风险水平,算法会做出相应的输出,预测准确率恰好对等于预设的置信度,实现了预测风险的可控性. 如图1所示,斜率等于对应的ε值,如 99%置信度对应的斜率约为 0.01,95%置信度对应的斜率约为 0.05,80%置信度对应的斜率约为0.2.

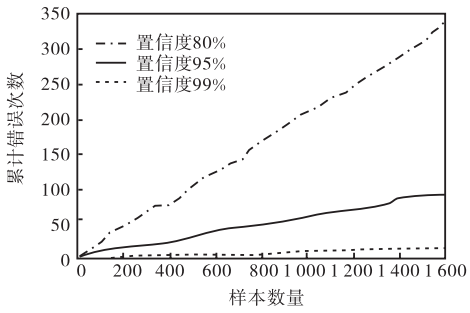

图1　CP-KerNN的校准性

Fig.1　Calibration property of CP-KerNN

3.4　CP-KerNN 的预测效率和计算效率

由第3.3节可知,为保证预测结果的高置信度,算法会给出较大的预测集,以减少犯错误的可能性. 域预测提供了备选结果,降低了接受唯一结果的风险,这时可利用专家知识结合实际做出判断;同时也造成不确定和模糊性,降低了信息量. 在控制风险的同时,更希望得到确定的预测结果. 表5分析了3种不同置信度水平下 CP-KerNN 的预测效率(近邻数 k=1).

表5　CP-KerNN算法的预测效率

Tab.5　Predictive efficiencies of CP-KerNN　%

置信度	准确率	确定率	偏好率	空集率
80	80.97	79.44	71.26	13.21
95	94.65	91.75	84.57	2.65
99	98.23	70.64	62.45	1.33

从表 5 可知,在各种置信度下,实际准确率几乎等于置信度,说明 CP-KerNN 提供的置信度非常有效.确定率和偏好率虽有波动,但总是接近相应的准确率,说明算法输出的大部分预测都是确信的和使用者偏好的;空集率非常接近算法风险水平,说明错误预测几乎都是由于空预测引起的. 由此可知,CP-KerNN 的预测效率也是比较高的.

为了进一步展示 CP-KerNN 的预测效率,在欧氏空间、基核空间和优化核空间中同时采用 kNN 设计奇异度函数,进行对比实验. 3 种算法的测试集均为 S=1 600;CP-KerNN 只有检验集 V 参与在线训练,在线训练集 $T'=V=600$ 基于欧氏空间(即 CP-kNN)和基核空间的 CP 算法没有核优化过程,因此不需要划分压缩集和校验集,在线训练集 T''=2 400.

为了方便对比,k 值均取为 1. 从表 6 可以看出,在 3 种置信度水平下,CP 算法在优化核空间中的预测效率最高,基核空间次之,欧氏空间中预测效率最低. 在线学习阶段中,CP-KerNN 依赖较小的样本集 T' 获得比基于较大样本集 T'' 的 CP 算法还高的效率,说明由第 1 阶段学习到的优化核空间 K_1 具有良好的泛化性,能够有效地提取压缩集的数据分布信息,并传递到第 2 阶段,使得校验集在核空间中也具有良好的类分开性,据此设计的样本奇异函数能够有效地反映样本在分布中的奇异性,压缩集的信息间接参与预测过程,预测效率不降反升;同时,压缩比例高达 3/4,计算代价被部分转移到了离线学习阶段,压缩集的样本不再保存和访问,降低了数据的存储规模,提高了计算效率. 在大数据集中,例如大工业过程,主要的计算负担来自于在线学习,这时 CP-KerNN 的这种优势显得非常重要.

表6　不同空间中的预测效率比较(k=1)

Tab.6　Comparison of certain prediction rates in different spaces(k=1)　%

置信度	欧氏空间	基核空间	优化核空间
80	68.76	69.87	79.44
95	76.26	85.78	91.75
99	63.89	66.34	70.64

k 值的选取是采用 kNN 来设计奇异度函数的关键问题之一,k 值越小则算法的运算效率越高,但预测效率可能下降. 由表7可知,在欧式空间中,当 $k=9$ 时预测确定率最高,在基核空间和优化核空间中,$k=1$ 时取得最高的预测确定率,并且优化核空间中预测确定率显著高于基核空间. 由于与欧式空间相比,数据在核空间具有更好的类分开性,同类的样本相互靠近,异类远离,k 取较小值时样本奇异度函数就能很好反映数据分布特征;由于基核参数 γ_0 较大,数据分布在核空间里比较紧密,k值较大时效率反而下降. 在应用中,k 取较小值能简化计算,避免过多搜索近邻,

能显著提高计算效率.

表 7　不同空间中的确定率与 k 值的关系(置信度取 95%)

Tab.7　Relationship between certain prediction rates and different values of k in different spaces (with confidence level 95%)

参数 k	欧氏空间/%	基核空间/%	优化核空间/%
1	76.26	85.78	91.75
3	77.78	81.35	87.75
5	78.56	81.89	86.64
9	78.96	80.46	86.02
19	76.00	80.08	81.45
51	72.14	75.37	78.89
99	70.00	74.89	76.59

4　结　语

本文提出了基于核函数优化的 CP 算法,即 CP-KerNN,理论分析与 CP-*k*NN 的对比实验都说明了该方法能同时提高预测性能和计算效率. 与 ICP 相比,CP-KerNN 在提高计算效率的同时不会降低预测效率;与 AL-CP 相比,CP-KerNN 仍然严格满足算法随机性理论和可校准性. 因此,本文提出的方法是对 CP 算法的一种较好的改进形式.

参考文献:

[1] 万百五. 工业大系统优化与产品质量控制[M]. 北京:科学出版社,2003.
Wan Baiwu. *Optimization and Product Quality Control of Large-Scale Industrial Processes* [M]. Beijing:Science Press,2003(in Chinese).

[2] 张　钊,吴爱国,裴燕玲. 模糊控制的模糊推理分析[J]. 控制与决策,2005,20(8):905-908.
Zhang Zhao,Wu Aiguo,Pei Yanling. The fuzzy reasoning analysis of fuzzy control [J]. *Control and Decision*, 2005,20(8):905-908(in Chinese).

[3] Venkat V,Rengaswamy R,Yin K,et al. A review of process fault detection and diagnosis(partⅢ):Process history based methods [J]. *Computers and Chemical Engineering*,2003,27(1):327-346.

[4] Lyman P R,Georgakis C. Plant-wide control of the Tennessee Eastman problem [J]. *Computers and Chemical Engineering*,1995,19(2):321-331.

[5] Abhijit K,Jayaraman V K,Kulkarni B D. Knowledge incorporated support vector machines to detect faults in Tennessee Eastman process[J]. *Computers and Chemical Engineering*,2005,29(10):2128-2133.

[6] Vovk V,Gammerman A,Shafer G. *Algorithmic Learning in a Random World* [M]. Heidelberg: Springer,2005.

[7] Gammerman A,Vovk V. Hedging predictions in machine learning [J]. *Computer Journal*,2007,50(2):151-163.

[8] Vovk V. A universal well-calibrated algorithm for on-line classification[J]. *Journal of Machine Learning Research*,2004,5:575-604.

[9] Melluish T,Saunders C,Nouretdinov I,et al. Comparing the Bayes and typicalness frameworks[C]//*Proceedings of the* 12*th European Conference on Machine Learning*. Freiburg,Germany,2001:143-152.

[10] Papadopoulos H. Qualified Predictions for Large Data Set [D]. London:Computer Learning Research Centre, Royal Holloway,University of London,2004.

[11] Ho S S,Wechsler H. Transductive confidence machine for active learning[C]//*Proceedings of the International Joint Conference on Neural Network*. Oregon,Portland, 2003:1435-1440.

[12] Cover T,Hart P. Nearest neighbor pattern classification [J]. *IEEE Transaction on Information Theory*,1967, 13(1):21-27.

[13] Weinberger K Q,Blitzer J,Saul L K. Distance metric learning for large margin nearest neighbor classification[J]. *Advances in Neural Information Processing Systems*,2006,18:1473-1480.

[14] Xiong H L,Swamy M N S,Ahmad M O. Optimizing the kernel in the empirical feature space [J]. *IEEE Transactions on Neural Networks*,2005,16(2):460-474.

[15] 蒋浩天. 工业系统的故障检测与诊断[M]. 段建民译. 北京:机械工业出版社,2003.
Chiang L H. *Fault Detection and Diagnosis in Industrial Systems* [M]. Duan Jianmin Trans. Beijing:China Machine Press,2003(in Chinese).