文章编号:1001 - 9081(2008)09 - 2423 - 04

融合先验知识的模糊最小二乘支持向量机模型及其应用

许 亮
(广东工业大学 自动化学院,广州 510090)
(xul_gsut@sina.com)

摘 要:为了解决最小二乘支持向量机(LSSVM)对噪声或孤立点敏感的问题,融合数据样本中的先验知识,提出一种基于噪声分布模型和样本紧密度的模糊最小二乘支持向量机模型。在训练的过程中,考虑样本的噪声分布信息。为了区分有效样本和噪声,研究了基于样本紧密度的策略。运用该策略和噪声分布模型,可自动生成相应样本的模糊隶属度。该方法提高了最小二乘支持向量机的抗噪声能力以及处理含有噪声或孤立点样本的灵活性。将提出的方法运用于润滑油精制生产过程的故障诊断。实验结果表明,该方法具有很好的分类精度和鲁棒性。

关键词:最小二乘支持向量机;模糊隶属度;先验知识;噪声;故障诊断
中图分类号: TP182 **文献标志码**:A

Incorporating prior knowledge in fuzzy least squares SVM model and its application

XU Liang
(*School of Automation, Guangdong University of Technology, Guangzhou Guangdong* 510090, *China*)

Abstract: To address the drawback that the Least Squares Support Vector Machines (LSSVM) is sensitive to noises or outliers, a LSSVM model incorporating with prior knowledge on data was proposed based on noise distribution and sample affinity. Information of noise distribution for samples was introduced in the training process. A strategy based on the sample affinity was presented to discriminate data with noises. A fuzzy membership was automatically generated and assigned to each corresponding data point in the sample set by using the strategy and the noise model. The ability of FLSSVM was improved to resist noises. The flexibility was increased to treat data points with noises or outliers. The proposed method was applied to fault diagnosis for the lubricating oil refining process. The experimental result shows that the proposed method has better robustness.

Key words: Least Squares Support Vector Machines (LSSVM); fuzzy membership; prior knowledge; noise; fault diagnosis

0 引言

支持向量机(Support Vector Machines, SVM)是由 Vapnik[1] 提出的一种研究小样本情况下机器学习规律的统计学习方法,它通过结构风险最小化原理来提高泛化能力,较好的解决了小样本、非线性、高维数、局部极小等实际问题。近年来,文献[2]提出的最小二乘支持向量机(Least Squares SVM, LSSVM)是对标准 SVM 的扩展,采用具有等式约束并且满足 KKT(Karush-Kuhn-Tucker)条件的规则化最小二乘函数作为损失函数,代替了 SVM 计算复杂的 QP(Quadratic-Programming)问题,求解速度相对加快。然而最小二乘支持向量机也存在不足。由于平方损失函数没有正则化,导致对孤立点的鲁棒性较差。而且需要假定误差服从正态分布,在实际应用中,这种假设条件是很难满足[3]。

为了解决对于噪声或孤立点过于敏感并由此带来的"欠学习"或"过学习"问题,文献[3]将模糊隶属度引入 C-均值聚类方法,构建模糊 C-均值聚类,然后应用到 LSSVM 模型中,降低了 LSSVM 对孤立点的敏感度。文献[4]提出基于支持向量数据域描述的模糊隶属度函数模型。根据样本偏离数据域的程度赋予不同的隶属度,该方法提高了 LSSVM 的抗噪声能力。然而在模糊 LSSVM(Fuzzy LSSVM, FLSSVM)模型训练过程中,如果缺乏样本噪声分布的先验知识,则很难给样本分配合适的模糊隶属度。

本文在模型的训练过程中,通过融合样本数据的先验知识,引入噪声分布模型。为了区分数据和噪声,提出基于样本紧密度的策略。运用该策略以及噪声分布模型,自动生成相应样本数据点的模糊隶属度。噪声分布模型可估计样本是噪声或孤立点的概率,据此调整模糊 LSSVM 中的模糊隶属度。当训练含有噪声或孤立点的样本时,简化模型学习的复杂度。

1 模糊最小二乘支持向量机

1.1 LSSVM 结构

给定带有类别标号的训练集$(x_1, y_1),(x_2, y_2),\cdots,(x_l, y_l)$,其中训练样本输入点$x_i \in R^N$,$y_i \in \{-1, 1\}$,类别标号,$i = 1, 2, 3, \cdots, l$。

在优化目标中,LSSVM 的损失函数为误差ξ_i的二次项,用等式约束方式代替标准 SVM 中的不等式约束方程。因此,优化问题通过求解一组线性方程组解决,则优化问题为[1-2]:

$$\min J = \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}w + \frac{C}{2}\sum_{i=1}^{l}\xi_i^2 \tag{1}$$

收稿日期:2008 - 03 - 21;**修回日期**:2008 - 06 - 20。
作者简介:许亮(1971 -),男,甘肃靖远人,讲师,博士,主要研究方向:统计机器学习、智能系统、故障诊断。

其约束方程为：

$$y_i[\boldsymbol{w}^{\mathrm{T}} \cdot \phi(x_i) + b] = 1 - \xi_i;\ i = 1, 2, 3, \cdots, l, \xi_i > 0 \tag{2}$$

其中：$\phi(\cdot): R^n \to R^{n_h}$ 是非线性映射，权矢量 $\boldsymbol{w}^{\mathrm{T}} \in R^{n_h}$，误差变量 $\xi_i \in R^n$，b 是偏差变量，C 是最大分类间隔与最小分类误差之间的折中。

式(1) 和式(2) 的对偶形式的推导，可根据目标函数和约束条件建立 Lagrange 函数如下：

$$L(w,b,\alpha,\xi) = \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}w + \frac{C}{2}\sum_{i=1}^{l}\xi_i^2 - \sum_{i=1}^{l}\alpha_i\{y_i(\boldsymbol{w}^T\phi(x_i) + b) - 1 + \xi_i\} \tag{3}$$

其中 $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_l)^{\mathrm{T}}$ 是拉格朗日乘子。对 w, b, ξ 和 α_i 分别求偏导，令等式为零，则整理得到：

$$\begin{bmatrix} \Lambda & Y \\ Y^{\mathrm{T}} & 0 \end{bmatrix}\begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} I \\ 0 \end{bmatrix} \tag{4}$$

其中：

$$\Lambda = y_i y_j \phi(x_i)\varphi(x_j) + \frac{\delta_{i,j}}{C},\ \delta_{i,j} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases},$$

$$Y = (y_1, y_2, \cdots, y_l)^{\mathrm{T}} \tag{5}$$

根据式(5) 解得最优解 α^* 和 b^*，并定义 $K(x,x_i) = \phi(x)\phi(x_i)$ 代替非线性映射，最后得到最小二乘支持向量机分类器的决策函数为：

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l}\alpha_i^* y_i K(x,x_i) + b^*\right) \tag{6}$$

核函数 $K(x,x_i)$ 是满足 Mercer 条件的任意对称函数。

1.2 模糊 LSSVM 结构

在标准 SVM 理论中，由于在构造最优分类面时样本数据点具有相同的作用，因此，当训练样本中含有噪声或孤立点时，会导致获得的分类面不是真正的最优分类面。针对这种情况，文献[5-6]提出模糊支持向量机方法。将这一思想引入 LSSVM，构造模糊 LSSVM 模型(FLSSVM)。首先给不同的数据点赋予不同的模糊隶属度，重新构造模糊样本集 $(x_1, y_1, \mu_1), (x_2, y_2, \mu_2), \cdots, (x_l, y_l, \mu_l)$，$\mu_i$ 是模糊成员函数，$\delta \leqslant \mu_i \leqslant 1, i = 1, 2, \cdots, l$，$\delta$ 是一个充分小的正数，则式(1) 中的优化目标函数改写为：

$$\min J = \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}w + \frac{C}{2}\sum_{i=1}^{l}\mu_i\xi_i^2 \tag{7}$$

其中 C 是常数(分类间隔和分类误差之间的折中)。较小的 μ_i 值，降低了参数 ξ_i 对优化问题的作用，相应样本数据点 x_i 对 LSSVM 模型训练的重要性也减少。

优化目标求解过程与 LSSVM 类似，需要构造 Lagrange 函数，最后得到矩阵方程：

$$\begin{bmatrix} \Lambda & Y \\ Y^{\mathrm{T}} & 0 \end{bmatrix}\begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} I \\ 0 \end{bmatrix}$$

其中：

$$\Lambda = y_i y_j \phi(x_i)\phi(x_j) + \frac{\delta_{i,j}}{\mu_i C},\ \delta_{i,j} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases},$$

$$Y = (y_1, y_2, \cdots, y_l)^{\mathrm{T}} \tag{8}$$

求解矩阵方程(7) 即可得到最优解，代入式(6)，就可得到模糊 LSSVM 最优超平面的决策函数。在 FLSSVM 中，不同的 μ_i 值，对于不同的样本点，在分类间隔和分类误差之间的折中是不同的。较小的 μ_i 值其相应样本点的重要性也较小。对含有噪声或孤立点的样本赋予较小的 μ_i 值，降低这些样本对模型的影响作用，从而改善模型的抗噪声能力。

2 融合先验知识的 LSSVM 模型

在实际的样本数据中，通常都隐含着目标问题预先未知的信息。如果在学习的过程中，考虑数据样本蕴涵的先验知识，可提高识别的质量，改善模型的性能[7]。

为了提高计算效率，LSSVM 选择最小平方作为误差函数 $\sum_{i}^{l}\xi_i^2$，并且使用正则化的参数 C 调整最优超平面的最大分类间隔和误差函数最小值。实际上，这种方法已经预估了一个误差概率分布。当误差概率分布的先验知识不同时，误差函数也是不同的，从而就有不同的优化问题。因此，在优化过程中，应该选择正确的误差模型。但是，误差概率分布是很难估计的，甚至对于各种应用而言，概率密度函数都是未知的。与之相比，样本数据点的噪声分布模型是可估计的。设 $p_x(x)$ 是样本数据点(非噪声或非孤立点) 的概率密度函数。对于 $p_x(x)$ 值较高的样本数据点 x_i，意味着这些数据点是非噪声数据的概率较高，而在训练过程中，应该赋予 ξ_i 较低的值[8]。为此，需要修改误差函数：

$$\sum_{i=1}^{l}p_i(x)\xi_i^2 \tag{9}$$

从而优化目标函数也要修改，则有：

$$\min J = \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}w + \frac{C}{2}\sum_{i=1}^{l}p_x(x)\xi_i^2 \tag{10}$$

其中 $p_x(x)$ 可看作一种模糊隶属度函数，使 $\mu_i = p_x(x_i)$ 代入式(7)，则上述的优化问题就转换为 FLSSVM 的优化问题。因此，可用 FLSSVM 的算法求解优化问题。

2.1 噪声分布模型

FLSSVM 模型的性能与样本中所蕴含的知识有关。若存在一个样本的噪声分布模型，则可假设一个函数是样本数据点 x_i 为非噪声的概率。设 p_i 是数据点 x_i 为非噪声的概率。此时，如果训练数据中包含噪声分布的潜在知识，则样本中每个数据点的模糊成员函数为 $\mu_i = f(p_i)$。在训练 FLSSVM 模型过程中，利用这些信息可得到最优超平面。但是，在许多应用中，这种信息是无法直接获得。于是，假设存在一个启发式函数 $h(x)$，它与概率密度函数 $p_x(x)$ 密切相关，两者之间存在某种映射关系。为此定义：

$$p_x(x) = \begin{cases} 1, & h(x_i) < h_C \\ \delta, & h(x_i) > h_N \\ 1 - (1-\delta)\left(\dfrac{h(x) - h_C}{h_N - h_C}\right)^d, & \text{其他} \end{cases} \tag{11}$$

其中：h_C 为可信因子，h_N 为无价值因子，$\delta > 0$ 是一个较小的正数。如图 1 所示，这两个因子可调整概率密度函数 $p_x(x)$ 和启发式函数 $h(x)$ 之间的映射关系，参数 d 是控制映射函数的自由度。

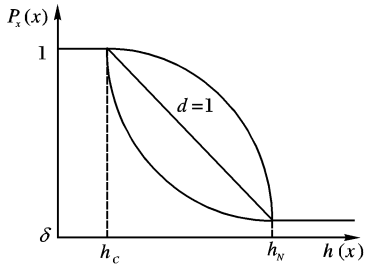


图 1 $p_x(x)$ 和 $h(x)$ 的映射关系

根据样本中数据点与可信因子 h_C 和无价值因子 h_N 之间

的位置关系,可以把样本划分为三个部分。当启发式函数 $h(x)$ 的值小于 h_C,数据点将会位于 $h(x) < h_C$ 的区域,其为非噪声的概率较高,隶属度可设为1;当启发式函数 $h(x)$ 的值大于 h_N,数据点将会位于 $h(x) > h_N$ 的区域,可被视为噪声数据或者孤立点,给其隶属度分配一个较小的 δ 值;位于其余区域的数据点被视作具有不同概率分布的噪声数据。为使问题简化,我们选择多项式函数估计概率密度函数 $p_x(x)$ 和启发式函数 $h(x)$ 的映射关系,自由度 d 控制两者之间映射程度。

2.2　基于样本紧密度的启发式函数

一般样本点到聚类中心的距离,可区分孤立点和有效样本。文献[6]提出基于样本到类中心之间的距离来度量该样本对分类的重要程度。通常远离类中心的数据点大部分是噪声数据和孤立点,对分类的贡献度较低,但是也有少部分噪声数据会位于类边缘。由于决定最优超平面的支持向量也位于类边缘,因此这种情况时,有效样本和噪声数据是很难区分。

本文为区分数据点和噪声,提出基于样本紧密度的策略。借助文献[11]的支持向量数据域描述方法,在特征空间中,用一个紧凑的球或超球将样本集包围起来。此时,样本的紧密度可用包围样本的最小球半径来度量。

设样本集中有 l 个样本表示为 $\{x_i \mid i = 1,2,3,\cdots, l\}$,其中 $x_i \in R^n$ 为输入空间。为了建立数据域描述,需要寻找一个包含所有对象的最小超球。当对样本集的情况未知时,引入一个松弛变量 $\rho_i (i = 1, 2, 3, \cdots, l)$ 来允许一部分数据点位于球的外面。通过对下面的目标函数最小化,寻找最小超球:

$$S(R,a,\xi) = R^2 + E(\sum_{i=1}^{n}\rho_i) \tag{12}$$

其中:参数 $E > 0$ 是惩罚因子,是超球体积和位于球外的样本个数之间的折中;R 是超球半径;a 是球心。

式(12)的约束条件为:

$$(x_i - a)^T(x_i - a) \leqslant R^2 + \rho_i;\ \rho_i \geqslant 0,\ i = 1,\cdots,l \tag{13}$$

为求解带约束的优化问题,构造 Lagrange 函数:

$$L(R,a,\beta_i,\rho_i) = R^2 + E(\sum_{i=1}^{n}\rho_i) - \sum_{i=1}^{n}\beta_i\{R^2 + \rho_i - (x_i^2 - 2ax_i + a^2)\} - \sum_{i=1}^{n}\gamma_i\rho_i \tag{14}$$

其中:Lagrange 系数 $\beta_i \geqslant 0, \gamma_i \geqslant 0$。对上式求偏导,并令等于0,得到:

$$\sum_i \beta_i = 1,\ a = \sum_i \beta_i x_i \tag{15}$$

$$D - \beta_i - \rho_i = 0;\ i = 1,\cdots,l \tag{16}$$

将约束条件式(15)和式(16)代入式(14)并合并整理,有:

$$L(R,a,\beta,\gamma,\rho) = \sum_{i=1}^{n}\beta_i(x_i \cdot x_i) - \sum_{i,j=1}^{n}\beta_i\beta_j(x_i \cdot x_j) \tag{17}$$

其约束条件为:

$$\sum_{i=1}^{n}\beta_i = 1;\ 0 \leqslant \beta_i \leqslant E, i = 1,\cdots,l \tag{18}$$

大多数情况下,样本并不是一个球形分布,此时一个紧密的超球是很难获得的。然而,利用一个非线性映射把输入空间的样本集映射到特征空间,然后选择一个合适的特征空间,于是一个更紧密的超球就可获得。因此,用核函数代替内积运算,式(17)及其约束条件在核空间的表达形式为:

$$L = \sum_{i=1}^{n}\beta_i K(x_i,x_i) - \sum_{i,j=1}^{n}\beta_i\beta_j K(x_i,x_j)$$

$$\sum_{i=1}^{n}\beta_i = 1;\ 0 \leqslant \beta_i \leqslant E,\ i = 1,\cdots,l \tag{19}$$

由式(15)的第二个方程可知,超球球心是带权系数 β_i 的样本线性组合。当 $\beta_i \neq 0$ 时,对应的样本为支持向量;当 $\beta_i = E$ 时,对应的样本位于超球的外边,称为孤立点或含噪声的样本。超球的最小半径由 $0 < \beta_i < E$ 中对应的样本与球心之间的距离来确定,即:

$$R = \| x_i - a \| \tag{20}$$

定义 $d(x_i)$ 为样本集中的样本 x_i 到其超球球心 a 的距离,计算公式为:

$$d(x_i) = \| x_i - a \| \tag{21}$$

样本 x_i 到超球球心的距离越大,则该样本属于该样本集的概率越小,而成为噪声或孤立点的概率越大,因此对分类的贡献度越小。所以,启发式函数 $h(x)$ 可定义为:

$$h(x) = \begin{cases} \| x_i - a \|, & d(x_i) \geqslant R \\ \min(d(x_i)), & d(x_i) < R \end{cases} \tag{22}$$

在噪声分布模型,设置参数可信因子 $h_C = \{d(x_i)\}$,无价值因子 $h_N = \{d(x_i)\}$,其中 $i = 1, 2, 3,\cdots, l$。

2.3　FLSSVM 多类分类方法

故障诊断本质上是一个多类分类问题。因此,利用 FLSSVM 进行故障诊断,需要研究多类分类问题的解决方法。其与 SVM 多类分类方法相似,主要是将标准两类分类方法进行扩展,构建多类分类器。Vapnik[1] 使用一对多分类方法,该方法对于 k- 类分类问题,构造 k 个 2- 类分类器,每一类对应其中的一个,将它与其他的类分开。然而,采用这种方法,对于 LSSVM 而言,需要求解一组变量个数等于训练数据个数的线性方程组。当训练样本很大时,求解一个问题的时间开销也就很大。考虑到训练过程时的计算效率,本文采用一对一多类分类方法构造 LSSVM 多类分类器。在一对一的分类器中,需要构造所有可能的2-类分类器,每一个分类器的训练数据集都只取自相应的两类。这样,对于 k 类问题共需要构造 $k(k-1)/2$[10]。

2.4　参数选择

对于模糊 LSSVM 模型核参数的选择,最常用、也较易使用的方法的是 k 重交叉验证方法。本文提出的融合先验知识的模糊 LSSVM 模型参数的选择可以分为两个主要部分,具体过程如下。

1)使用 k 重交叉验证方法确定 LSSVM 模型的核参数和误差惩罚参数 C。

2)设置已获取的核参数和误差惩罚参数 C 不变,寻找模糊 LSSVM 中的其他参数。

(1)确定启发式函数 $h(x)$,计算超球的球心、球半径以及数据点到球心的最大值、最小值。

(2)预先给定可信因子 h_C、无价值因子 h_N、模糊隶属度下界 δ 和映射函数的自由度 d 等参数的范围。搜索这些参数的过程分为两步:首先给定可信因子 h_C 和无价值因子 h_N 的值,寻找 δ 和 d;其次,给定已获取的 δ 和 d,寻找 h_C 和 h_N。

3　FLSSVM 模型在故障诊断中的应用

3.1　应用问题描述

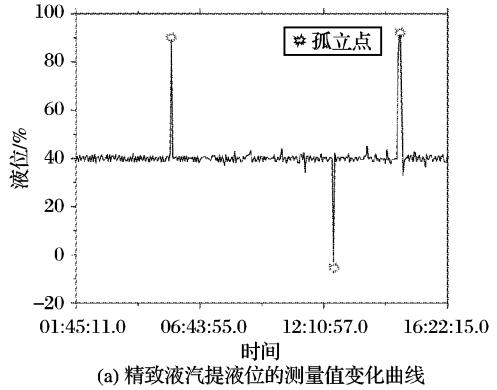
润滑油生产是石油炼制企业主要基础油品的一个重要生产环节。润滑油生产过程具有工艺流程长、工序多、原料种类多及多种产品结构,其生产过程较为复杂。润滑油生产过程包括糠醛精制和酮苯脱蜡两个过程。润滑油精制过程是炼油

生产中工艺较为复杂的过程装置,由原料脱气、糠醛抽提、精制液汽提、抽出液汽提、水溶液回收等子系统组成,它的安全和稳定生产,对维持产品的质量和企业的经济效益意义重大。

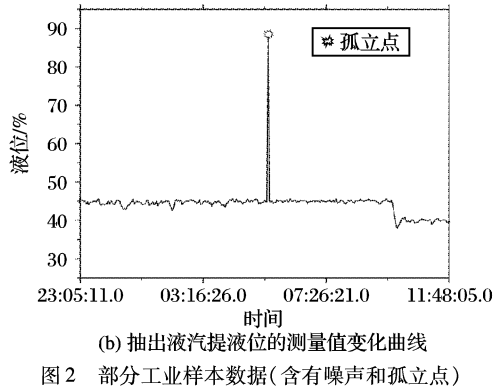
故障诊断技术是保证化工生产过程安全生产运行的核心技术。目前,随着各种智能计算技术的发展,以多元统计分析为代表的故障诊断方法得到了更多研究人员的关注[11]。这类故障诊断方法建模时,可避开机理模型的约束。但是,由于故障样例过少以及化工过程的数据具有高噪声的特点,抑制了这种方法的应用范围和效果。本文将 FLSSVM 方法用于润滑油精制生产过程的故障诊断,并与标准 LSSVM 作了比对。

3.2 训练样本的获取和预处理

训练样本和测试数据来自某石化公司润滑油生产过程。影响润滑油精制生产过程,造成系统运行不稳定的因素主要有:温度、压力、流量和液位。因此,与系统运行相关的有四类,369 个过程变量,每个变量有五类运行状态:高高限、高限、正常、低限、低低限。原始数据保存在控制系统的实时数据库的历史记录表中。表结构中主要有三个属性:位号、采样时间、采样值。样本数据的采样时间段从 2006 年 9 月 10 日 19:43:10 到 2006 年 9 月 15 日 17:00:58,有 232764 条实测数据。为了保证实验的效果,随机选择 2000 个样本点,60 个变量,类别数为 5,经过预处理和归一化等步骤后,分为两个集合:S_1 有 1000 个经过滤波、去噪等处理的样本;S_2 有含有噪声和孤立点 1000 个样本,其中含有 10 个孤立点。部分样本的曲线图如图 2 所示。图 2(a)采样周期为 2006-09-14,含有 3 个孤立点;图 2(b)采样周期为 2006-09-12,含有 1 个孤立点。



(a) 精致液汽提液位的测量值变化曲线



(b) 抽出液汽提液位的测量值变化曲线

图 2 部分工业样本数据(含有噪声和孤立点)

3.3 结果和讨论

将所有数据按照 T_1(训练集):T_2(测试集) = 3:1 进行划分。训练集中的数据要有类别标号。定义分类集 C 为{1, 2, 3, 4, 5},表示为不同的运行状态,其中"1" 表示"低低限","2" 表示"低限","3" 表示"正常","4" 表示"高限","5" 表示"高高限"。因此,对于 LSSVM 训练集表示为$\{x_i, y_i\}$,其中 x_i 是标准化的样本数据,$y_i \in C$ 为类别标号。

对于本文所有实验,核函数均选择高斯径向基:

$$K(x_i, x_j) = \exp(-\lambda \| x_i - x_j \|^2) \tag{23}$$

用 LSSVM 进行训练,给出参数 C 和 γ 范围,$C = \{2^0, 2^1, \cdots, 2^7\}$,$\gamma = \{2^{-7}, 2^{-2}, \cdots, 2^7\}$,通过 10 重交叉验证方法搜寻参数,最后得到 C 和 γ 的值(8,4)。用得到的模型对测试样本进行测试,分类结果如表 1 所示。

用 FLSSVM 训练时,由于比标准 LSSVM 参数多,因此搜寻参数的过程分为两个步骤:寻找核参数 γ 和参数 C;固定寻找到的核参数,搜索模糊隶属度映射函数的参数。

在模型训练过程中,核参数 γ 和参数 C 选用 LSSVM 模型的参数。然后搜寻模糊隶属度映射函数的参数。首先确定启发式函数中的参数,计算输入样本 $\{x_1, x_2, \cdots, x_l\}$ 的最小包围球球心以及任意样本到球心的最大距离 $\mathrm{Max}(d(x_i))$ 和最小距离 $\mathrm{Min}(d(x_i))$,其中核参数 $\lambda = 0.5$,$C = 8$。其次,确定模糊映射函数的参数。先设置参数可信因子 $h_C = \mathrm{Min}(d(x_i))$,无价值因子 $h_N = \mathrm{Max}(d(x_i))$,进行二维搜寻参数 δ 和 d。参数 δ 的值范围 0.1 ~ 0.9,搜寻步长 0.01,d 的值范围 2^{-8} ~ 2^8;再固定参数 $\delta = 0.05$ 和 $d = 32$,进行二维搜寻参数可信因子 h_C 和无价值因子 h_N。最后,用训练后得到的模型对测试样本测试,并与 LSSVM 分类相对比,分类结果见表 1。

表 1 LSSVM 和 FLSSVM 故障分类结果

测试样本集	LSSVM/%	FLSSVM/%
S_1	97.23 ± 0.25	97.86 ± 0.30
S_2	89.65 ± 2.32	95.53 ± 1.24

从表 1 可看出,对不含有噪声和孤立点的样本集 S_1,LSSVM 和 FLSSVM 模型的分类正确率和标准偏差在数值上都很接近。而对含有噪声和孤立点的样本集 S_2,两种模型的分类正确率都出现了下降,标准偏差也较大。但是,FLSSVM 模型的分类正确率较明显好于 LSSVM 模型。从实验结果说明,FLSSVM 模型的抗噪能力优于 LSSVM 模型。

4 结语

针对标准 LSSVM 对噪声或孤立点敏感的问题,本文在 LSSVM 模型中融合样本数据噪声分布的先验知识,构造模糊 LSSVM 模型。在模型的训练过程中,考虑了样本数据的噪声分布信息。为了区分有效样本和噪声样本,研究了基于样本紧密度的启发式函数。然后结合噪声分布模型和启发式函数,确定模糊隶属度函数模型。该方法改善了 FLSSVM 在处理含有噪声或孤立点样本的灵活性,提高了模型的抗噪能力。将提出的方法运用于润滑油精制过程的故障诊断。实验结果表明,该方法可有效增强 LSSVM 的鲁棒性。

与标准 LSSVM 相比,由于需要选择多个参数,FLSSVM 的训练速度相对较慢。大多数情况下,训练时间主要用于核参数和惩罚因子 C 的选择上,较少的时间用于搜索建立模糊隶属度模型的参数。

参考文献:

[1] VAPNIK V N. Statistical learning theory [M]. New York: John Wiley and Sons Inc, 1998.

[2] SUYKENS J A K, VANDEVALLE J. Least squares support vector machine classifiers [J]. Neural Processing Letters, 1999, 9(3): 293-300.

(下转第 2429 页)

应的方式,寻找到最好解,提高了收敛速度。

6) 替换更新。在变异后的 N 个新抗体中,选择 d 个亲和度最高的抗体,替换种群中的 d 个亲和度最低的抗体,形成新一代的抗体群。

7) 终止条件。当种群中绝大多数个体所代表的解的代价和执行时间都基本相同时,则终止演化,否则转到步骤 3)。具体是:进行完一代演化之后,随机选择 k 个个体对 (S_1,S_{1+k}),(S_2,S_{2+k}),…,(S_k,S_{k+k}),每个个体所代表的解的系统处理时间为 $time(S_i)$,系统代价为 $cost(S_i)$,演化的终止条件为:

$$\sum_{i=1}^{k} | time(S_i) - time(S_{i+k}) | /k \leqslant x_t \quad (3)$$

且:

$$\sum_{i=1}^{k} | cost(S_i) - cost(S_{i+k}) | /k \leqslant x_c \quad (4)$$

其中,k,x_t,x_c 是设计时确定的常量参数,当 $N \leqslant 150$ 时,可以取 $k = 15$,当 $N > 150$ 时,可取 $k = \lceil N/10 \rceil$,x_t,x_c 需要根据系统的具体时间、代价参数确定。本文取 $x_t = 0.5$,$x_c = 0.5$。这种终止策略具有很强的自适应性,既可以避免迭代次数过多而进行无用搜索,又可以防止迭代次数太少而达不到最优解。x_t,x_c 可以控制目标种群中个体之间的差异大小和算法的搜索时间。

3　算法结果与分析

为评价本文提出的算法性能,随机生成了 20、50、100 及 200 个节点的 CDFG,并随机生成了各个节点的性能参数,同时根据对节点的性能参数的分析,确定了系统的约束条件。对每个 CDFG 都进行 100 次的测试,并将系统总代价、处理时间及算法执行时间同模拟退火算法和普通遗传算法求出的值进行了比较,实验数据如表 2 所示。

表 2 算法执行代价和处理时间的结果比较

节点数	时间约束	模拟退火算法			遗传算法			克隆选择算法		
		系统总代价	总处理时间	算法执行时间/ms	系统总代价	总处理时间	算法执行时间/ms	系统总代价	总处理时间	算法执行时间/ms
20	200	300	198	1 250	305	196	660	299	197	320
50	500	630	498	2 896	637	495	1 289	630	497	670
100	900	892	895	5 330	898	892	2 300	890	896	1 028
200	1 600	1 360	1 597	7 890	1 370	1 595	3 120	1 355	1 595	1 780

根据实验结果可出如下结论:

1)我们的模型可以有效的指导软硬件划分;

2)三种算法都可以得到较优解;

3)克隆选择算法较模拟退火算法和遗传算法收敛速度快。

4　结语

本文提出了一种嵌入式系统的软硬件划分模型,此模型基于单处理器的体系结构。构造了划分的目标函数和系统约束,在介绍克隆选择算法的基础上,通过实验比较了克隆选择算法、模拟退火算法和遗传算法在软硬件划分问题中的应用。为了把注意力集中到系统的划分问题,本文中没有考虑不同任务之间的通信开销及调度,在实际的应用中,可以综合考虑这些因素而得到更为完善的划分。

参考文献:

[1] 高健,李涛. 三种软硬件划分算法的比较分析[J]. 计算机工程与设计,2007,28(14):3426 - 3428.

[2] 李晖,姚放吾,邓新颖,等. 基于免疫算法的嵌入式系统软硬件划分方法[J]. 计算机工程与设计,2006,27(22):4239 - 4242.

[3] 吴泽俊,钱立进,梁意文. 入侵检测系统中基于免疫的克隆选择算法[J]. 计算机工程,2004,30(6):50 - 52.

[4] 周泉,章兢. 基于克隆选择原理的免疫算法[J]. 计算机工程与应用,2005,41(21):61 - 63.

[5] De CASTRO L N, Von ZUBEN F J. The clonal selection algorithm with engineering applications[C]// Workshop Proceedings of GECCO, Workshop on Artificial Immune Systems and Their Applications. Oralndo: IEEE Press, 2000: 36 - 37.

[6] KALAVADE A, SUBRAHMANYAM P A. Hardware/software partitioning for multi-function systems[C]// Proceedings of the 1997 IEEE/ACM International Conference on Computer-aided Design. Washington, DC: IEEE Computer Society, 1997: 516 - 521.

(上接第 2426 页)

[3] JOOYONG S, CHANGHA H, SUNKYUN N. Robust LS-SVM regression using fuzzy C-Means clustering [J]. Advances in Natural Computation, 2006, 4221: 157 - 166.

[4] 张英,苏宏业,褚健. 基于模糊最小二乘支持向量机的软测量建模[J]. 控制与决策, 2005, 25(6):621 - 624.

[5] LIN C F, WANG S D. Fuzzy support vector machines [J]. IEEE Transactions on Neural Networks, 2002, 13(2):464 - 471.

[6] HUANG H P, LIU Y H. Fuzzy support vector machines for pattern recognition and data mining [J]. International Journal of Fuzzy Systems, 2002, 4(3):826 - 835.

[7] LAUER F, BLOCH G. Incorporating prior knowledge in support vector machines for classification: A review[J]. Neurocomputing, 2008, 71(7/9):1578 - 1594.

[8] LIU F C, WANG D S. Training algorithms for fuzzy support vector machines with noisy data [J]. Pattern recognition letters, 2004 (25):1647 - 1656.

[9] DAVID M J T, ROBERT P W D. Support vector machines [J]. Pattern recognition letters, 1999(20):1191 - 1199.

[10] KERBEL U H-G. Pairwise classification and support vector machines[M]. Advances in kernel methods: support vector learning. Cambridge, MA: MIT Press, 1999: 255 - 268.

[11] HYUN W C. Nonlinear feature extraction and classification of multivariate data in kernel feature space[J]. Expert System With Application, 2007, 32(2):534 - 542.

[12] 许亮. 基于核函数和知识的化工过程安全运行智能支持系统研究[D]. 广州:华南理工大学, 2007.