(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106295685 A
(43)申请公布日 2017.01.04

(21)申请号 201610624342.X

(22)申请日 2016.08.01

(71)申请人 浙江大学
地址 310058 浙江省杭州市西湖区余杭塘路866号

(72)发明人 杨春节 安汝峤 潘怡君

(74)专利代理机构 杭州求是专利事务所有限公司 33200
代理人 林松海

(51)Int.Cl.
G06K 9/62(2006.01)

权利要求书3页 说明书9页

(54)发明名称

改进的直推式支持向量机的大型高炉故障分类算法及应用

(57)摘要

本发明公开了一种改进的直推式支持向量机的大型高炉故障分类算法及应用，属于工业过程监控与诊断技术领域。首先，针对工业采集数据，利用训练数据中包含的正负标签的数据进行归纳式学习，得到一个原始的样本分类器。其次，利用原始的样本分类器对无标签样本进行分类。最后，通过迭代计算的方法获得最优的样本分类器。本发明提出了一种改进的基于直推式支持向量机的故障分类算法，从平衡数据样本类别的数量入手，对无标签的样本进行了初步的预测，并对该过程进行了优化，因此与其它现有的方法相比，本发明方法在流程工业模拟试验中取得了较好的分类效果，并具有更高的准确率。

CN 106295685 A

1.一种改进的直推式支持向量机的大型高炉故障分类算法，其特征在于，主要采用对N个点的数据采取随机选择的策略，通过L次的选择，分别计算出L次的准确率，选取准确率最高的一次作为该模型的分类器，步骤如下：

步骤一：初始化惩罚因子C，利用训练数据中包含的正负标签的数据进行归纳式学习，得到一个原始的样本分类器；

步骤二：初始化惩罚因子C*，用原始的样本分类器对无标签样本进行分类；

步骤三：迭代计算。

2.根据权利要求1所述的方法，其特征在于，所述的步骤一建模过程如下：

对于支持向量机，给定数据样本集：

$(x_1,y_1),(x_2,y_2),..,(x_1,y_1)$ (1)

$y\in\{-1,1\}$代表不同类，分类的任务是构建最优超平面$f(x)=<w,\Phi(x)>+b$，把属于不同类的向量x_i分开，其中w为参数向量，$\Phi(\cdot)$为输入空间到特征空间的映射函数，定义损失函数如下：

$$\min R(w,b)=\frac{1}{2}\langle w,w\rangle+c\cdot R_{emp} \tag{2}$$

第一部分定义了模型的结构复杂度；第二部分R_{emp}为经验风险；c为调节常数，用于控制模型复杂度与逼近误差的折中，当经验风险取不同的函数时，得到不同的SVM分类器，当经验风险$R_{emp}=0$，即仅仅考虑分类器的模型复杂度时，损失函数变为：

$$R(w\cdot b)=\frac{1}{2}\langle w,w\rangle \tag{3}$$

优化问题描述为：

$$\min R(w,b)=\frac{1}{2}\langle w,w\rangle \tag{4}$$

$$\text{subject to } y_i\left(\langle w,x_i\rangle-b\right)\geq 1,\quad i=1,2,\cdots,l$$

为了得到对偶的优化问题，引入拉格朗日乘子，得到拉格朗日方程：

$$L=\frac{1}{2}\langle w,w\rangle-\sum_{i=1}^{l}a_i\left(y_i\left(\langle w,x_i\rangle-b\right)-1\right) \tag{5}$$

求该函数关于原始变量的微分：

$$\frac{\partial L}{\partial w}=w-\sum_{i=1}^{l}a_iy_ix_i=0 \tag{6}$$

$$\frac{\partial L}{\partial b}=\sum_{i=1}^{l}a_iy_i=0 \tag{7}$$

将公式(6)(7)带入拉格朗日方程：

$$L=\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_{i=1}^{l}a_ia_jy_iy_j\langle x_i,x_j\rangle \tag{8}$$

得到对偶的优化问题：

$$\max\ W(a)=\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_{i=1}^{l}a_ia_jy_iy_j\langle x_i,x_j\rangle \qquad (9)$$

$$\text{subject to}\ \sum_{i=1}^{l}a_iy_i=0,\quad a_i\geq 0,\quad i=1,2,\cdots,l$$

为了容忍训练集中噪声和异常数据，定义间隔松弛向量，以其1范数作为经验风险，即得到1范数软间隔分类器。优化问题描述为：

$$\min R(w,b)=\frac{1}{2}\langle w,w\rangle+c\sum_{i=1}^{l}\xi_i \qquad (10)$$

$$\text{subject to}\ y_i\left(\langle w,x_i\rangle-b\right)\geq 1-\xi_i,\quad i=1,2,\cdots,l$$

其中，ξ_i为松弛变量，它使得可以容忍训练数据的错误分类，当取$\xi_i=0$，$i=1,2,\cdots,l$时，软间隔分类器退化成为硬间隔分类器，该优化问题的对偶问题为：

$$\max\ W(a)=\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}a_ia_jy_iy_j\langle x_i,x_j\rangle \qquad (11)$$

$$\text{subject to}\ \sum_{i=1}^{l}a_iy_i=0,\quad 0\leq a_i\leq C$$

3.根据权利要求1所述的方法，其特征在于，所述的步骤二建模过程如下：基于迭代算法的直推式支持向量机给定一组独立同分布的有标签训练样本点$(x_1,y_1),(x_2,y_2),..,(x_l,y_l)$，$x\in R^m$，$y\in\{-1,1\}$和另一组来自同一分布的无标签样本点$x_1^*,x_2^*,x_3^*,..,x_k^*$，在一般的线性不可分条件下，TSVM的训练过程描述为以下的优化问题：

$$\left(y_1^*,\ldots,y_k^*,w,b,\xi_1,\ldots,\xi_n,\xi_1^*,\ldots,\xi_k^*\right)$$

$$\min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i+C^*\sum_{j=1}^{k}\xi_j^* \qquad (12)$$

$$\text{subject to}\ \forall_{i=1}^{l}: y_i[w\cdot x_i+b]\geq 1-\xi_i$$
$$\forall_{j=1}^{k}: y_j[w\cdot x_j^*+b]\geq 1-\xi_j^*$$
$$\forall_{i=1}^{l}: \xi_i\geq 0$$
$$\forall_{j=1}^{k}: \xi_j^*\geq 0$$

其中参数C和C^*为用户指定和调节的参数，参数C^*是未标识样本在训练过程中的影响因子，C^*与ξ_j^*称为未标识样本x_j在目标函数中的影响项。

4.根据权利要求1所述的方法，其特征在于，步骤三所述的迭代计算过程如下：

1)计算每一个样本到超平面的距离$|f(x)|$，选取N个距离$|f(x)|\leq d$的样本点，在N个样本点中随机取出M个样本点；

2)假定训练集中M个无标记样本中正负样本的比例为1:1，并指定一个训练集中无标签样本的临时惩罚因子C^{*temp}；

3)用得到的样本分类器对训练集中的无标签样本进行重新分类，根据分类器对无标签数据记录的判别结果，对无标签数据做出正负分类判决，并将判决值较大的一半样本标记为正标签，另外一半标记为负样本；

4)用得到的经过重新标记的训练集数据对TSVM学习机进行重新训练,得到新的分类器,然后,按一定的规则交换一对标签值不同的训练样本的标签符号,即把起初标记为正样本的未标记样本中标记为负样本,起初标记为负样本的标记为正样本,计算式(3)的值,使得问题(3)的值获得最大下降;反复执行训练样本标签的变换,直到找不出满足交换条件的样本为止;

5)均匀地增加未标记样本的惩罚因子C^{*temp}的值,并重新执行步骤(4),直到$C^{*temp} \geq C$时,TSVM的学习结束;

6)测试分类器的效果,并重复(1)至(5)的操作L次,选择具有最优正确率的分类器。

5.一种根据权利要求1-4任一项所述的方法用于高炉冶炼过程故障分类。

改进的直推式支持向量机的大型高炉故障分类算法及应用

技术领域

[0001] 本发明属于工业过程监控与故障诊断领域，特别涉及一种改进的基于直推式支持向量机的大型高炉系统故障分类算法。

背景技术

[0002] 工业生产是国家重要的经济发展内容，针对工业过程的故障分类研究，对保证安全高效的生产具有十分重要的意义。目前常见的故障分类方法包括定性与定量的分析方法。其中定性分析方法包括图论方法、专家系统、定性仿真。定量的方法又包括基于解析模型的方法与数据驱动的方法。而目前研究的热门领域包括机器学习、多元统计分析、信号处理等都属于数据驱动的方法。对于复杂的工业过程而言，很难构建精确的机理模型，也很难收集全面的专家系统知识，因此基于数据的方法具有很好的应用前景。工业生产过程中，各种传感器可以获取大量的数据，通过计算机的运算存储功能，数据以海量的规模进行增长，为数据分析提供了充足的资源。目前应用较多的数据驱动方法，如主元分析(PCA)、偏最小二乘(PLS)、支持向量机(SVM)、人工神经网络(ANN)等。很多学者对这些方法进行了改进，也对一些方法进行融合，从而大大提高了故障诊断的效果。

[0003] 对于半监督支持向量机算法，最早是由创始者Vapnik等人提出的直推式学习方法，后来又引入了局部组合搜索、梯度下降、连续优化技术、凸凹过程、半正定编程、不可微方法、决定退火、分支定界等方法。其中直推式学习假定未标记示例就是测试例，即学习的目的就是在这些未标记示例上取得最佳泛化能力。直推式支持向量机(transductive SVM，TSVM)很好地利用了这部分数据，在有标签数据的运算基础上加入无标签数据，通过一些列算法将无标签的数据进行分类，从而有效的解决学习过程中产生的模型的准确问题。

发明内容

[0004] 为了克服现有技术的不足，本发明的目的在于针对直推式支持向量机算法的特点，提供一种基于改进的直推式支持向量机的大型高炉故障分类方法，并将这种方法应用在大型高炉系统的故障分类应用中。

[0005] 一种改进的基于直推式支持向量机的大型高炉系统故障分类算法，主要采用对N个点的数据采取随机选择的策略，通过L次的选择，分别计算出L次的准确率，选取准确率最高的一次即作为该模型的分类器，步骤如下：

[0006] 步骤一：初始化惩罚因子C，利用训练数据中包含的正负标签的数据进行归纳式学习，得到一个原始的样本分类器。

[0007] 对于支持向量机，给定数据样本集：

[0008] $(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)$ (1)

[0009] $y\in\{-1,1\}$代表不同类。分类的任务是构建最优超平面$f(x)=\langle w,\Phi(x)\rangle+b$，把属于不同类的向量$x_i$分开。其中w为参数向量，$\Phi(\cdot)$为输入空间到特征空间的映射函数。定义损失函数如下：

[0010] $$\min R(w,b)=\frac{1}{2}\langle w,w\rangle+c\cdot R_{emp} \tag{2}$$

[0011] 第一部分定义了模型的结构复杂度;第二部分R_{emp}为经验风险;c为调节常数,用于控制模型复杂度与逼近误差的折中。当经验风险取不同的函数时,得到不同的SVM分类器。当经验风险$R_{emp}=0$,即仅仅考虑分类器的模型复杂度时,损失函数变为:

[0012] $$R(w\cdot b)=\frac{1}{2}\langle w,w\rangle \tag{3}$$

[0013] 优化问题描述为:

[0014] $$\min R(w,b)=\frac{1}{2}\langle w,w\rangle \tag{4}$$
$$\text{subject to } y_i\left(\langle w,x_i\rangle-b\right)\geq 1,\quad i=1,2,\cdots,l$$

[0015] 为了得到对偶的优化问题,引入拉格朗日乘子,得到拉格朗日方程:

[0016] $$L=\frac{1}{2}\langle w,w\rangle-\sum_{i=1}^{l}a_i\left(y_i\left(\langle w,x_i\rangle-b\right)-1\right) \tag{5}$$

[0017] 求该函数关于原始变量的微分:

[0018] $$\frac{\partial L}{\partial w}=w-\sum_{i=1}^{l}a_i y_i x_i=0 \tag{6}$$

[0019] $$\frac{\partial L}{\partial b}=\sum_{i=1}^{l}a_i y_i=0 \tag{7}$$

[0020] 将公式(6)(7)带入拉格朗日方程:

[0021] $$L=\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_{i=1}^{l}a_i a_j y_i y_j\langle x_i,x_j\rangle \tag{8}$$

[0022] 得到对偶的优化问题:

[0023] $$\max\ \mathrm{W}(a)=\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_{i=1}^{l}a_i a_j y_i y_j\langle x_i,x_j\rangle \tag{9}$$
$$\text{subject to}\sum_{i=1}^{l}a_i y_i=0,\quad a_i\geq 0,\quad i=1,2,\cdots,l$$

[0024] 为了容忍训练集中噪声和异常数据,定义间隔松弛向量,以其1范数作为经验风险,即得到1范数软间隔分类器。优化问题描述为:

[0025] $$\min R(w,b)=\frac{1}{2}\langle w,w\rangle+c\sum_{i=1}^{l}\xi_i \tag{10}$$
$$\text{subject to } y_i\left(\langle w,x_i\rangle-b\right)\geq 1-\xi_i,\quad i=1,2,\cdots,l$$

[0026] 其中,ξ_i为松弛变量,它使得可以容忍训练数据的错误分类。当取$\xi_i=0$,i=1,2,…,l时,软间隔分类器退化成为硬间隔分类器。该优化问题的对偶问题为:

[0027]
$$\max\ W(a)=\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_{i=1,j=1}^{l}a_ia_jy_iy_j\langle x_i,x_j\rangle \qquad (11)$$
$$\text{subject to}\sum_{i=1}^{l}a_iy_i=0,\quad 0\le a_i\le C$$

[0028] 步骤二：初始化惩罚因子C*，用原始的样本分类器对无标签样本进行分类。基于迭代算法的直推式支持向量机给定一组独立同分布的有标签训练样本点$(x_1,y_1),(x_2,y_2),\ldots,(x_l,y_l)$，$x\in R^m$，$y\in\{-1,1\}$和另一组来自同一分布的无标签样本点$x_1^*,x_2^*,x_3^*,\ldots,x_k^*$。在一般的线性不可分条件下，TSVM的训练过程可以描述为以下的优化问题：

$$\left(y_1^*,\ldots,y_k^*,w,b,\xi_1,\ldots,\xi_n,\xi_1^*,\ldots,\xi_k^*\right)$$

[0029]
$$\min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i+C^*\sum_{j=1}^{k}\xi_j^* \qquad (12)$$
$$\text{subject to}\ \forall_{i=1}^{l}:y_i[w\cdot x_i+b]\ge 1-\xi_i$$
$$\forall_{j=1}^{k}:y_j[w\cdot x_j^*+b]\ge 1-\xi_j^*$$
$$\forall_{i=1}^{l}:\xi_i\ge 0$$
$$\forall_{j=1}^{k}:\xi_j^*\ge 0$$

[0030] 其中参数C和C*为用户指定和调节的参数，参数C*是未标识样本在训练过程中的影响因子，C*与ξ_j^*称为未标识样本x_j在目标函数中的影响项。

[0031] 步骤三：迭代计算。

[0032] 1)计算每一个样本到超平面的距离|f(x)|，选取N个距离|f(x)|≤d的样本点，在N个样本点中随机取出M个样本点。

[0033] 2)假定训练集中M个无标记样本中正负样本的比例为1:1，并指定一个训练集中无标签样本的临时惩罚因子C^{*temp}。

[0034] 3)用得到的样本分类器对训练集中的无标签样本进行重新分类，根据分类器对无标签数据记录的判别结果，对无标签数据做出正负分类判决，并将判决值较大的一半样本标记为正标签，另外一半标记为负样本。

[0035] 4)用得到的经过重新标记的训练集数据对TSVM学习机进行重新训练，得到新的分类器。然后，按一定的规则交换一对标签值不同的训练样本的标签符号，即把起初标记为正样本的未标记样本中标记为负样本，起初标记为负样本的标记为正样本，计算式(3)的值，使得问题(3)的值获得最大下降。反复执行训练样本标签的变换，直到找不出满足交换条件的样本为止。

[0036] 5)均匀地增加未标记样本的惩罚因子C^{*temp}的值，并重新执行步骤(4)，直到$C^{*temp}\ge C$时，TSVM的学习结束。

[0037] 6)测试分类器的效果，并重复(1)至(5)的操作L次，选择具有最优正确率的分类器。

[0038] 所述的工业故障为高炉冶炼过程故障。

[0039] 一种所述的方法用于高炉冶炼过程故障分类。

[0040] 本发明具有以下有益效果：

[0041] 1.本发明首次提出一种应用于高炉冶炼过程故障的改进直推式支持向量机算法，并且基于这个改进方法利用了大量的无标签数据，利用样本的多次迭代筛选的方法，实现了对复杂过程的故障分类；

[0042] 2.本发明能够针对改进的直推式支持向量机算法，通过平衡数据样本类别的数量入手，对无标签的样本进行了初步的预测，并对该过程进行了优化。本算法采用的筛选机制能够比较有效的利用无标签样本对原始模型进行正确修正，使得分类准确率得到提高，有效提高算法的学习精度。

具体实施方式

[0043] 本发明首先，针对工业采集数据，利用训练数据中包含的正负标签的数据进行归纳式学习，得到一个原始的样本分类器。其次，利用原始的样本分类器对无标签样本进行分类。最后，通过迭代计算的方法获得最优的样本分类器。

[0044] 本发明提出了一种改进的基于直推式支持向量机的故障分类算法，从平衡数据样本类别的数量入手，对无标签的样本进行了初步的预测，并对该过程进行了优化。

[0045] 一种改进的基于直推式支持向量机的大型高炉系统故障分类算法，主要采用对N个点的数据采取随机选择的策略，通过L次的选择，分别计算出L次的准确率，选取准确率最高的一次即作为该模型的分类器，步骤如下：

[0046] 步骤一：初始化惩罚因子C，利用训练数据中包含的正负标签的数据进行归纳式学习，得到一个原始的样本分类器。

[0047] 对于支持向量机，利用工业过程采集的离线数据集：

[0048] $(x_1,y_1),(x_2,y_2),..,(x_1,y_1)$ (1)

[0049] $y\in\{-1,1\}$代表不同类。分类的任务是构建最优超平面$f(x)=\langle w,\Phi(x)\rangle+b$，把属于不同类的向量$x_i$分开。其中w为参数向量，$\Phi(\cdot)$为输入空间到特征空间的映射函数。定义损失函数如下：

[0050] $$\min R(w,b)=\frac{1}{2}\langle w,w\rangle+c\cdot R_{emp} \quad (2)$$

[0051] 第一部分定义了模型的结构复杂度；第二部分R_{emp}为经验风险；c为调节常数，用于控制模型复杂度与逼近误差的折中。当经验风险取不同的函数时，得到不同的SVM分类器。当经验风险$R_{emp}=0$，即仅仅考虑分类器的模型复杂度时，损失函数变为：

[0052] $$R(w\cdot b)=\frac{1}{2}\langle w,w\rangle \quad (3)$$

[0053] 优化问题描述为：

[0054] $$\min R(w,b)=\frac{1}{2}\langle w,w\rangle \quad (4)$$

$$\text{subject to } y_i\left(\langle w,x_i\rangle-b\right)\geq 1,\quad i=1,2,\cdots,l$$

[0055] 为了得到对偶的优化问题，引入拉格朗日乘子，得到拉格朗日方程：

[0056] $$L=\frac{1}{2}\langle w,w\rangle-\sum_{i=1}^{l}a_i\left(y_i\left(\langle w,x_i\rangle-b\right)-1\right) \quad (5)$$

[0057] 求该函数关于原始变量的微分：

[0058] $$\frac{\partial L}{\partial w}=w-\sum_{i=1}^{l}a_i y_i x_i=0 \tag{6}$$

[0059] $$\frac{\partial L}{\partial b}=\sum_{i=1}^{l}a_i y_i=0 \tag{7}$$

[0060] 将公式(6)(7)带入拉格朗日方程：

[0061] $$L=\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_{i=1}^{l}a_i a_j y_i y_j\left\langle x_i,x_j\right\rangle \tag{8}$$

[0062] 得到对偶的优化问题：

[0063] $$\max\ \mathrm{W}(a)=\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_{i=1}^{l}a_i a_j y_i y_j\left\langle x_i,x_j\right\rangle \tag{9}$$
$$\text{subject to}\sum_{i=1}^{l}a_i y_i=0,\quad a_i\geq 0,\quad i=1,2,\cdots,l$$

[0064] 为了容忍训练集中噪声和异常数据，定义间隔松弛向量，以其1范数作为经验风险，即得到1范数软间隔分类器。优化问题描述为：

[0065] $$\min R(w,b)=\frac{1}{2}\langle w,w\rangle+c\sum_{i=1}^{l}\xi_i \tag{10}$$
$$\text{subject to}\ y_i\left(\langle w,x_i\rangle-b\right)\geq 1-\xi_i,\quad i=1,2,\cdots,l$$

[0066] 其中，ξ_i为松弛变量，它使得可以容忍训练数据的错误分类。当取$\xi_i=0$，i＝1，2，…，l时，软间隔分类器退化成为硬间隔分类器。该优化问题的对偶问题为：

[0067] $$\max\ \mathrm{W}(a)=\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}a_i a_j y_i y_j\left\langle x_i,x_j\right\rangle \tag{11}$$
$$\text{subject to}\sum_{i=1}^{l}a_i y_i=0,\quad 0\leq a_i\leq C$$

[0068] 步骤二：初始化惩罚因子C*，用原始的样本分类器对无标签样本进行分类。基于迭代算法的直推式支持向量机给定一组独立同分布的有标签训练样本点$(x_1,y_1),(x_2,y_2),\ldots,(x_l,y_l)$，$x\in R^m$，$y\in\{-1,1\}$和另一组来自同一分布的无标签样本点$x_1^*,x_2^*,x_3^*,\ldots,x_k^*$。在一般的线性不可分条件下，TSVM的训练过程可以描述为以下的优化问题：

$$\left(y_1^*,\ldots,y_k^*,w,b,\xi_1,\ldots,\xi_n,\xi_1^*,\ldots,\xi_k^*\right)$$

[0069] $$\min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i+C^*\sum_{j=1}^{k}\xi_j^* \tag{12}$$

[0070] $$\text{subject to}\ \forall_{i=1}^{l}: y_i[w\cdot x_i+b]\geq 1-\xi_i$$
$$\forall_{j=1}^{k}: y_j[w\cdot x_j^*+b]\geq 1-\xi_j^*$$
$$\forall_{i=1}^{l}: \xi_i\geq 0$$
$$\forall_{j=1}^{k}: \xi_j^*\geq 0$$

[0071] 其中参数C和C^*为用户指定和调节的参数，参数C^*是未标识样本在训练过程中的影响因子，C^*与ξ_j^*称为未标识样本x_j在目标函数中的影响项。

[0072] 步骤三：迭代计算。

[0073] 1)计算每一个样本到超平面的距离|f(x)|，选取N个距离|f(x)|≤d的样本点，在N个样本点中随机取出M个样本点。

[0074] 2)假定训练集中M个无标记样本中正负样本的比例为1:1，并指定一个训练集中无标签样本的临时惩罚因子C^{*temp}。

[0075] 3)用得到的样本分类器对训练集中的无标签样本进行重新分类，根据分类器对无标签数据记录的判别结果，对无标签数据做出正负分类判决，并将判决值较大的一半样本标记为正标签，另外一半标记为负样本。

[0076] 4)用得到的经过重新标记的训练集数据对TSVM学习机进行重新训练，得到新的分类器。然后，按一定的规则交换一对标签值不同的训练样本的标签符号，即把起初标记为正样本的未标记样本中标记为负样本，起初标记为负样本的标记为正样本，计算式(3)的值，使得问题(3)的值获得最大下降。反复执行训练样本标签的变换，直到找不出满足交换条件的样本为止。

[0077] 5)均匀地增加未标记样本的惩罚因子C^{*temp}的值，并重新执行步骤(4)，直到C^{*temp}≥C时，TSVM的学习结束。

[0078] 6)测试分类器的效果，并重复(1)至(5)的操作L次，选择具有最优正确率的分类器。

[0079] 上述实施例用来解释说明本发明，而不是对本发明进行限制，在本发明的精神和权利要求的保护范围内，对本发明做出的任何修改和改变，都落入本发明的保护范围。

[0080] 实施例

[0081] 高炉炼铁是钢铁生产中的重要环节，是衡量一个国家的经济水平和综合国力的重要指标。保证大型高炉系统安全稳定的运行在经济和安全上都是十分必要的，所以对大型高炉非正常工况诊断与安全运行方法进行研究具有重要意义。

[0082] 高炉冶炼是一个连续的生产过程，全过程在炉料自上而下，煤气自下而上的相互接触过程中完成。炉料按一定批料从炉顶装入炉内，从风口鼓入由热风炉加热到1000－1300℃热风，炉料中焦炭在风口前燃烧，产生高温和还原性气体，在炉内上升过程中加热缓慢下降的炉料，并还原铁矿石中的氧化物为金属铁。矿石升至一定温度后软化，熔融滴落，矿山中未被还原的物质形成熔渣，实现渣铁分离。渣铁聚集于炉缸内，发生诸多反应，最后调整成分和温度达到终点，定期从炉内排放炉渣和铁水。上升的煤气流将能量传给炉料而使温度降低，最终形成高炉煤气从炉顶导出管排出，进入除尘系统。

[0083] 成立于1958年的某钢炼铁厂，是一个有着56年辉煌历史的设备先进、装备水平较高的大型冶炼企业，主要产品为生铁，副产品有炉尘、炉渣、高炉煤气等。它拥有7座现代化高炉，高炉整体有效容积为11750立方米，其中2号高炉有效容积为2000立方米，是目前该省最大的高炉。新高炉投产后，炼铁厂将具备年产生铁1000万吨以上的综合能力。

[0084] 接下来结合该具体过程对本发明的实施步骤进行详细地阐述：

[0085] 步骤一：初始化惩罚因子C，利用训练数据中包含的正负标签的数据进行归纳式学习，得到一个原始的样本分类器。

[0086] 对于支持向量机，利用工业过程采集的离线数据集：

[0087] $(x_1,y_1),(x_2,y_2),\ldots,(x_1,y_1)$ (1)

[0088] $y\in\{-1,1\}$代表不同类。分类的任务是构建最优超平面$f(x)=\langle w,\Phi(x)\rangle+b$，把属于不同类的向量$x_i$分开。其中w为参数向量，$\Phi(\cdot)$为输入空间到特征空间的映射函数。定义损失函数如下：

[0089] $$\min R(w,b)=\frac{1}{2}\langle w,w\rangle+c\cdot R_{emp} \quad (2)$$

[0090] 第一部分定义了模型的结构复杂度；第二部分R_{emp}为经验风险；c为调节常数，用于控制模型复杂度与逼近误差的折中。当经验风险取不同的函数时，得到不同的SVM分类器。当经验风险$R_{emp}=0$，即仅仅考虑分类器的模型复杂度时，损失函数变为：

[0091] $$R(w\cdot b)=\frac{1}{2}\langle w,w\rangle \quad (3)$$

[0092] 优化问题描述为：

[0093] $$\min R(w,b)=\frac{1}{2}\langle w,w\rangle \quad (4)$$
$$\text{subject to } y_i\left(\langle w,x_i\rangle-b\right)\geq 1,\quad i=1,2,\cdots,l$$

[0094] 为了得到对偶的优化问题，引入拉格朗日乘子，得到拉格朗日方程：

[0095] $$L=\frac{1}{2}\langle w,w\rangle-\sum_{i=1}^{l}a_i\left(y_i\left(\langle w,x_i\rangle-b\right)-1\right) \quad (5)$$

[0096] 求该函数关于原始变量的微分：

[0097] $$\frac{\partial L}{\partial w}=w-\sum_{i=1}^{l}a_iy_ix_i=0 \quad (6)$$

[0098] $$\frac{\partial L}{\partial b}=\sum_{i=1}^{l}a_iy_i=0 \quad (7)$$

[0099] 将公式(6)(7)带入拉格朗日方程：

[0100] $$L=\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_{i=1}^{l}a_ia_jy_iy_j\langle x_i,x_j\rangle \quad (8)$$

[0101] 得到对偶的优化问题：

[0102] $$\max\ W(a)=\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_{i=1}^{l}a_ia_jy_iy_j\langle x_i,x_j\rangle \quad (9)$$
$$\text{subject to}\sum_{i=1}^{l}a_iy_i=0,\quad a_i\geq 0,\quad i=1,2,\cdots,l$$

[0103] 为了容忍训练集中噪声和异常数据，定义间隔松弛向量，以其1范数作为经验风险，即得到1范数软间隔分类器。优化问题描述为：

[0104]
$$\min R(w,b)=\frac{1}{2}\langle w,w\rangle+c\sum_{i=1}^{l}\xi_i \qquad (10)$$
$$\text{subject to } y_i\left(\langle w,x_i\rangle-b\right)\geq 1-\xi_i,\quad i=1,2,\cdots,l$$

[0105] 其中，ξ_i为松弛变量，它使得可以容忍训练数据的错误分类。当取$\xi_i=0$，i＝1，2，…，l时，软间隔分类器退化成为硬间隔分类器。该优化问题的对偶问题为：

[0106]
$$\max\ W(a)=\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_{i=1,j=1}^{l}a_ia_jy_iy_j\langle x_i,x_j\rangle \qquad (11)$$
$$\text{subject to } \sum_{i=1}^{l}a_iy_i=0,\quad 0\leq a_i\leq C$$

[0107] 步骤二：初始化惩罚因子C*，用原始的样本分类器对无标签样本进行分类。基于迭代算法的直推式支持向量机给定一组独立同分布的有标签训练样本点$(x_1,y_1),(x_2,y_2),\ldots,(x_l,y_l),x\in R^m,y\in\{-1,1\}$和另一组来自同一分布的无标签样本点$x_1^*,x_2^*,x_3^*,\ldots,x_k^*$。在一般的线性不可分条件下，TSVM的训练过程可以描述为以下的优化问题：

$$\left(y_1^*,\ldots,y_k^*,w,b,\xi_1,\ldots,\xi_n,\xi_1^*,\ldots,\xi_k^*\right)$$

$$\min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i+C^*\sum_{j=1}^{k}\xi_j^* \qquad (12)$$
$$\text{subject to } \forall_{i=1}^{l}: y_i[w\cdot x_i+b]\geq 1-\xi_i$$
$$\forall_{j=1}^{k}: y_j[w\cdot x_j^*+b]\geq 1-\xi_j^*$$
$$\forall_{i=1}^{l}: \xi_i\geq 0$$
$$\forall_{j=1}^{k}: \xi_j^*\geq 0$$

其中参数C和C*为用户指定和调节的参数，参数C*是未标识样本在训练过程中的影响因子，C*与ξ_j^*称为未标识样本x_j在目标函数中的影响项。

[0108] 步骤三：迭代计算。

[0109] 1)计算每一个样本到超平面的距离|f(x)|，选取N个距离|f(x)|≤d的样本点，在N个样本点中随机取出M个样本点。

[0110] 2)假定训练集中M个无标记样本中正负样本的比例为1:1，并指定一个训练集中无标签样本的临时惩罚因子C^{*temp}。

[0111] 3)用得到的样本分类器对训练集中的无标签样本进行重新分类，根据分类器对无标签数据记录的判别结果，对无标签数据做出正负分类判决，并将判决值较大的一半样本标记为正标签，另外一半标记为负样本。

[0112] 4)用得到的经过重新标记的训练集数据对TSVM学习机进行重新训练，得到新的分类器。然后，按一定的规则交换一对标签值不同的训练样本的标签符号，即把起初标记为正样本的未标记样本中标记为负样本，起初标记为负样本的标记为正样本，计算式(3)的值，使得问题(3)的值获得最大下降。反复执行训练样本标签的变换，直到找不出满足交换条件的样本为止。

[0113] 5)均匀地增加未标记样本的惩罚因子C^{*temp}的值，并重新执行步骤(4)，直到C^{*temp}

≥C时，TSVM的学习结束。

[0114] 6)测试分类器的效果，并重复(1)至(5)的操作L次，选择具有最优正确率的分类器。

[0115] 上述实施例用来解释说明本发明，而不是对本发明进行限制，在本发明的精神和权利要求的保护范围内，对本发明做出的任何修改和改变，都落入本发明的保护范围。