第35卷 第22期 计 算 机 工 程 2009年11月
Vol.35 No.22 Computer Engineering November 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)22—0212—04 文献标识码：A 中图分类号：TP181

改进的在线支持向量机训练算法

潘以桢，胡越明
(上海交通大学计算机科学与工程系，上海 200240)

摘 要：传统支持向量机基于批量训练方法，无法适应环境污染预测中的海量数据与实时性要求。在分析研究一种典型的在线支持向量机回归算法[4]的基础上，指出原算法在训练过程中存在样本重复移动问题，导致模型训练速度下降。提出一种改进算法，消除重复移动问题。实验结果表明，该改进在线支持向量机算法建模精度高，训练速度较原算法有显著提高。

关键词：污染预测；支持向量机；在线学习；增量式学习

Improved Online Training Algorithm of Support Vector Machine

PAN Yi-zhen, HU Yue-ming
(Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai 200240)

【Abstract】 Traditional Support Vector Machine(SVM), which based on batch training, can't satisfy the real-time requirement of environmental pollution prediction with large scale data. With the analysis of a typical kind of online support vector regression algorithm, this paper indicates that repeated sample move exists in the training process would lead to decrease the training speed, and proposes an improved algorithm. Simulation and analysis results show that the proposed algorithm performs high modeling precision, and training speed is increased remarkably compared with the aforementioned algorithm.

【Key words】 pollution prediction; Support Vector Machine(SVM); online learning; incremental learning

1 概述

污染预测是环境保护中的一项重要工作，它能为环境管理部门提供及时、准确、全面的环境质量信息，反映环境污染的变化趋势，有针对性地控制和预防污染事件的发生，从而推动环境监测工作的发展[1]。

目前，污染预测方法主要以污染物排放相关因素为基础进行建模预测，常用的方法有灰色理论模型GM(1, 1)、模糊识别方法和人工神经网络预测法方法[2]等。灰色理论模型形式简单，能反映事物的发展趋势，但对变化幅度较大的数据则无能为力，而且对数据的分布有一定的限制。模糊识别方法根据环境污染具有模糊性以及污染物浓度与多种污染指标存在相关性的特点，应用模糊数学的模糊聚类与识别模型进行预测，要实现较为精确的预测，需要大批量的样本数据。人工神经网络(ANN)方法能较好地模拟污染因素中的非线性关系，但其具有推广能力差、易于陷入局部最优、学习过拟合等缺点。

支持向量机(Support Vector Machine, SVM)是基于统计学习理论的机器学习技术[3]，在解决小样本、非线性问题中表现出独特优势，其遵循结构风险最小化原则，具有很强的泛化能力，目前已经成为污染预测应用领域的研究热点。然而，由于其基于二次规划的优化计算方法，计算复杂度与样本量密切相关，当样本量增加时，其训练时间将呈非线性增长，不适用于大数据量训练，已成为阻碍其进一步应用的瓶颈。

与批量式学习方法不同，在线学习方法能对每次迭代过程中增加的样本进行学习，利用前一次迭代的运算结果，减少计算复杂度，实现在较小的时间代价下的新样本学习，增量式学习方法是其中重要的一种。近年来，有多种SVM增量式在线学习算法被提出[4-5]。其中比较有代表性的是针对回归问题的增量式算法[4]，它以KKT条件为基础，当有新样本加入时，通过修改拉格朗日乘子和偏置值来更新SVM模型。SVM增量式学习算法已经在文本分类、系统辨识等很多领域得到应用，并取得了令人满意的结果[6]。然而，在环境监测领域，还没有发现其应用的实例。

本文针对环境预测这一特定领域，研究增量式支持向量机学习算法，在文献[4]的基础上，针对其在处理临界样本时存在的问题进行分析与研究，对原有算法进行改进，在保证预测精度的前提下，使训练速度提高了30%。

2 基于支持向量机的环境预测

环境预测问题可以归结为回归问题，用于解决此类问题的SVM数学描述为：假设给定 l 个样本数据 $\{x_i, y_i\}_{i=1}^{l}$，其中，$x_i \in R^n$ 为 n 维样本输入；y_i 为样本输出，则回归问题就是找到一个函数 f 使之经过训练后，对于样本以外的 x，通过 f 能够找出对应的 y。所求函数具有如下形式：

$$f(x)=\sum_{i}^{l}\theta_i \boldsymbol{K}(x_i, x)+b \tag{1}$$

其中，$\theta_i=\alpha_i-\alpha_i^*$，$\alpha_i, \alpha_i^*$ 为拉格朗日乘子；$\boldsymbol{K}$ 为核函数矩阵；b 为偏置值。

目前，回归支持向量机(SVR)的训练方式有2种：批量训练与在线训练。其问题表述和最优化求解的原理相同，不同点在于：在线支持向量机不断将新数据加入到训练集，通

作者简介：潘以桢(1983－)，男，硕士研究生，主研方向：机器学习，计算机系统结构；胡越明，副教授

收稿日期：2009-04-23 **E-mail：**panyizhen@gmail.com

过在线调整模型参数，实现预测模型的不断优化，而不是批量式的一次训练所有样本。这样不但能提高其训练时间，而且能提高模型的适应能力，使环境情况发生变化时，模型能相应发生变化。增量式 SVM 的基本思想如图 1 所示。

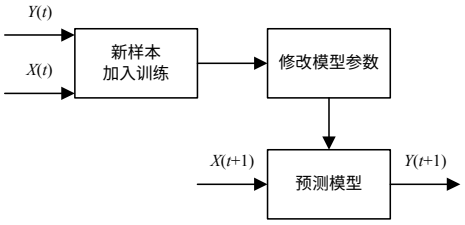

图 1　增量式 SVM 的基本思想

3　增量式在线学习算法

3.1　基本思想

文献[4]提出一种用于解决回归问题的支持向量机在线学习算法，通过严格的数学推导，给出了增加或删除一个样本时 SVM 模型的修正方法。

算法首先定义误差函数

$$h(x_i)=f(x_i)-y_i \tag{2}$$

根据每个样本的 $h(x_i)$ 与参数 θ_i 的值，将训练样本集分成 3 个子集：错误支持向量集合 E、边缘支持向量集合 S 以及保留样本集合 R，定义如下：

$$\begin{aligned}&E=\{i\,|\,(\theta_i=-C\wedge h(x_i)\geq+\varepsilon)\vee(\theta_i=+C\wedge h(x_i)\leq-\varepsilon)\}\\&S=\{i\,|\,(\theta_i\in[-C,0]\wedge h(x_i)=+\varepsilon)\vee(\theta_i\in[0,+C]\wedge h(x_i)=-\varepsilon)\}\\&R=\{i\,|\,\theta_i=0\wedge|h(x_i)|\leq\varepsilon\}\end{aligned} \tag{3}$$

其中，参数 C,ε 分别为回归支持向量机优化问题中的容量因子和允许误差，具体推导过程参见文献[4]。

模型的训练目标是将新样本 $\{x_c,y_c\}$ 加入上述 3 个集合之一，同时使所有样本仍满足 KKT 条件。其方法为：当 $|h(x_c)|<\varepsilon$ 时，新样本加入 R 集，原有训练样本集与 θ_i 值无须改变；当 $|h(x_c)|\geq\varepsilon$ 时，新样本加入 S 集或 E 集，加入前须保证新样本引起的 θ_i 和 $h(x_i)$ 值的变化不会影响原有 3 个集合的组成，否则必须先对原有集合组成进行调整，即样本移动。

样本移动可分为 3 种情况：从 S 集移动到 R 集或 E 集；从 E 集移动到 S 集；从 R 集移动到 S 集。移动的对象和方式取决于移动造成的新样本 θ_c 的变化量($\Delta\theta_c$ 值)。每次迭代选择使 $\Delta\theta_c$ 最小的方向进行样本移动，直至新样本加入 S 集或 E 集。$\Delta\theta_c$ 的计算公式及样本移动后参数调整方式详见文献[4]。

3.2　存在的问题

通过算法分析与实验发现，与常用的 SVM 软件包 LibSVM[7]相比，文献[4]提出的算法速度提高并不明显，在样本数小于 200 时用时相差无几，当样本数为 400 时提高也不到 20%。

在对该算法进行仔细分析后发现，在新样本的训练过程中，算法进行了多次重复样本移动，即相同样本在 2 个集合之间进行了多次移动，消耗了很多时间。

从式(3)中可以看出，3 个集合并非互不相交。集合 E 中存在 $|\theta_i|=C$ 且 $|h(x_c)|=\varepsilon$ 的样本，集合 R 中存在 $|\theta_i|=0$ 且 $|h(x_c)|=\varepsilon$ 的样本，而此样本同时也符合集合 S 的要求；集合 S 中存在 $|\theta_i|=C$ 或 0 且 $|h(x_c)|=\varepsilon$，此样本同样也符合集合 E 或 R 的要求。

文献[5]的算法没有考虑临界位置样本的处理：根据算法中 $\Delta\theta_c$ 的计算方法，移动临界位置的样本将使 $\Delta\theta_c=0$，从而成为首选移动的样本；另外，当集合 E 或 R 中的临界样本移入集合 S 之后，其 θ_i 并没有改变，仍然为 0 或 $\pm C$，根据算法定义，进行下一轮样本移动选择时，此样本对应的 $\Delta\theta_c$ 仍然为 0，因此将再次被选中，移回集合 R 或 E 中。如此循环往复，造成样本的重复移动，理论上将无法完成训练。原算法最终能完成训练是由于浮点数计算的误差，随着移动次数的增多，误差的不断累积导致 $\Delta\theta_c$ 增大从而完成训练。

3.3　改进的算法

为提高训练速度，本文算法改进了集合中临界样本的处理方法，避免了样本的重复移动。当集合 S 中存在对应 $\Delta\theta_c$ 值为 0 的样本时，$\Delta\theta_c$ 值不作更改，即令其移动到集合 R 或 E 中；而当集合 R 或 E 中存在对应 $\Delta\theta_c$ 值为 0 的样本时，则将 $\Delta\theta_c$ 值设为无限大(INF)，即确保其不会被选为此次移动的样本。

本文算法使位于临界位置的样本只能从 S 中移出，不能从 R 集或 E 集中移入 S 集，从而避免样本的重复移动。由于此类样本的移动不会对其余训练样本的参数产生影响，因此禁止其移动后仍能保证训练算法的正确性，而允许 S 集中临界样本移出的原因是算法中参数矩阵计算的复杂度决定于边缘支持向量的数目，将 S 中的样本移出更有利于加快算法速度。

3.4　算法复杂度分析

算法的时间复杂度主要取决于样本移动的次数和样本移动后参数的更新。参数更新的时间复杂度为 $O(n\times l_s)\times O(Kernel)$，其中，$n$ 为训练样本总数；l_s 为边缘支持向量数目；$Kernel$ 表示 SVM 核函数。本文算法在最坏情况下的时间复杂度为 $O(n^3)\times O(Kernel)$。

计算核函数的计算复杂度 $O(Kernel)$ 可以通过缓存核矩阵去除，代价是增加 $O(n^2)$ 的空间复杂度。因此，尽管本文算法在最坏情况下 $O(n^3)$ 的时间复杂度较高，但考虑到样本移动次数的可变性，在平均情况下，算法的时间复杂度为 $O(cn^2)$。

在本文算法中，新样本加入时每个原样本的每种移动至多进行一次，因此，样本的移动总次数为 $O(n\times3)$。文献[4]算法中存在样本重复移动，样本的每种移动方式可能进行多次，其移动次数必然超过 $O(n\times3)$。

4　算法实现与应用

4.1　算法实现步骤

在线 SVM 算法的主要实现部分包括递增算法和递减算法 2 个部分。

(1)递增部分：将新样本 $\{x_c,y_c\}$ 样本加入训练集中，并逐步调整 θ_i 及偏置值 b，直到新样本进入 3 个子集中的一个。具体步骤如下：

Step1　置 $\theta_c=0$。

Step2　计算 $h(x_c)$；如果 $|h(x_c)|<\varepsilon$，转 Step3；否则转 Step4。

Step3　将新样本加入集合 R，算法结束。

Step4　计算 $h(x_i),i=1,2,\cdots,l$，l 为样本总数。

Step5　更新 $\Delta\theta_c$ 计算参数。

Step6　求出新样本加入 S、新样本加入 E 造成的 $\Delta\theta_c$ 值，以及 3 种原样本移动方式中各自造成 $\Delta\theta_c$ 最小的样本移动对象和相应的 $\Delta\theta_c$ 值。对于 $E\rightarrow S$ 和 $R\rightarrow S$ 这 2 种移动方式，如果最小的 $\Delta\theta_c$ 值为 0 的话则取 $\Delta\theta_c$ 次小的样本进行移动。

Step7 求出 5 种样本移动方式中最小的 $\Delta\theta_c$ 值，确定移动方式及移动的样本。

Step8 更新 b，θ_c，$\theta_i, i=1,2,\cdots,l$ 及 $h(x_i), i=\in E\cup R$。

Step9 样本移动；如果新样本进入 S 或 E，算法结束；否则转 Step5。

(2)递减部分：目的是将一个样本从训练集合中去除，而使剩余样本仍满足 KKT 条件。算法逐步调整 θ_i 及偏置值 b，使要去除的样本 θ_c 变为 0，也使样本 $\{x_c, y_c\}$ 移入保留样本集合 R。具体步骤如下：

Step1 如果 $\{x_c, y_c\}\in R$，将 $\{x_c, y_c\}$ 移出训练集；否则，转 Step2。

Step2 将 $\{x_c, y_c\}$ 从 S 或 E 中移出。

Step3 计算 $h(x_i), i=1,2,\cdots,l$，l 为样本总数。

Step4 更新 $\Delta\theta_c$ 计算参数。

Step5 求 $\{x_c, y_c\}$ 加入 R 造成的 $\Delta\theta_c$ 值，以及 3 种样本移动方式中各自造成 $\Delta\theta_c$ 最小的样本移动对象和相应的 $\Delta\theta_c$ 值。同递增算法中一样，对于 $E\to S$ 和 $R\to S$ 这 2 种移动方式，如果最小的 $\Delta\theta_c$ 值为 0，取 $\Delta\theta_c$ 次小的样本进行移动。

Step6 求出 4 种样本移动方式中最小的 $\Delta\theta_c$ 值，确定移动方式及移动的样本。

Step7 更新 b，θ_c，$\theta_i, i=1,2,\cdots,l$ 及 $h(x_i),\ i=\in E\cup R$。

Step8 样本移动；如果 $\{x_c, y_c\}$ 进入 R，将其从训练样本集中删除，算法结束；否则转 Step4。

4.2 增量式 SVR 在污染预测中的应用

污染预测的目标值为某种有害气体的浓度，根据对环境参数进行分析，首先确定影响目标值的关键变量，如污染源状态、温度、风速、数据取得的时间日期等，作为预测模型中的输入特征值。参加训练的样本数据为 $\{x_i, y_i\}$，其中，y_i 为目标值；$x_i=(v_{i_1}, v_{i2},\cdots, v_{i_n})$ 为输入值；$v_{i_1}, v_{i2},\cdots, v_{i_n}$ 代表各个关键变量的值。则基于此模型的污染预测系统的实现过程如下：

Step1 初始化预测模型。

Step2 等待新数据。

Step3 构造新样本 $\{x_c, y_c\}$，如训练数据量未超出阈值，转 Step5。

Step4 使用递减算法，删除最早进入训练集的样本。

Step5 使用递增算法，将 $\{x_c, y_c\}$ 加入训练集，返回 Step2。

经过一段时间训练能获得污染预测模型，根据测试输入值即可通过模型计算实现预测，而后，利用预测目标的实际值，与相应输入一起作为新的训练样本用于模型调整，并可据此分析模型预测精度。

5 仿真实验

实验采用 StatLib 数据库(http://lib.stat.cmu.edu/)中的 NO_2 数据集，其样本目标值为 NO_2 浓度，输入 7 维特征向量，分别是汽车流量(单位：辆/h)、距地面 2 m 高处温度、风速、距地面 25 m 高处与距地面 2 m 高处的温差、风向、所处时刻距 2001 年 1 月以来的天数。数据集中包括 500 个样本。

实验环境是 Dell Inspiron 6000 笔记本电脑，CPU 为 Pentium M, 1.70 GHz, 内存为 1 GB。SVM 模型采用 RBF 核函数，核宽度 σ 取 0.6, 容量因子 C 取为 1，允许误差 ε 取为 0.01。首先对数据进行归一化处理，将数据值范围限制在区间[0, 1]中，然后利用归一化数据进行实验。

5.1 基于增量式在线 SVM 的训练时间

依次将数据集中的数据加入训练集进行训练，数据量阈值设为 300。实验结果如图 2 所示。

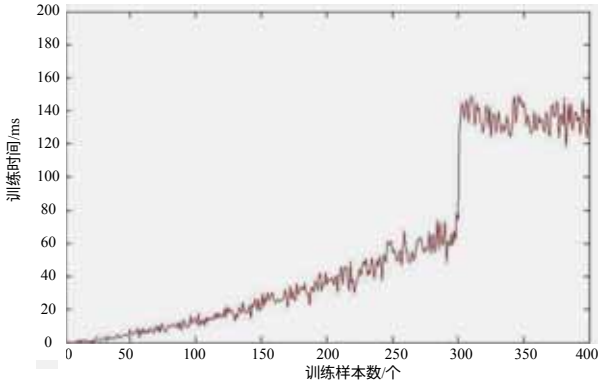

图 2　在线 SVM 模型训练时间

当训练样本数目小于 300 时，训练时间逐渐增加，样本数为 300 时训练时间约为 70 ms；训练样本数目超过 300 后，模型训练时间有一个突变，这是由于先进行样本删除操作，再进样本新增，共 2 次训练，此时训练时间稳定在 130 ms 左右。实验结果显示的训练用时情况完全可以满足污染预测中的实时性要求。

5.2 在线 SVM 与批量式 SVM 训练时间的比较

为减少试验偶然性，本文采用多次实验求均值的方法，即从数据集全部样本中随机选取 400 个作为训练样本，分别采用本文的在线 SVM 算法、文献[4]中的在线 SVM 算法以及 LibSVM 进行训练。对 20 次实验的均值作图并进行比较，结果如图 3 所示。

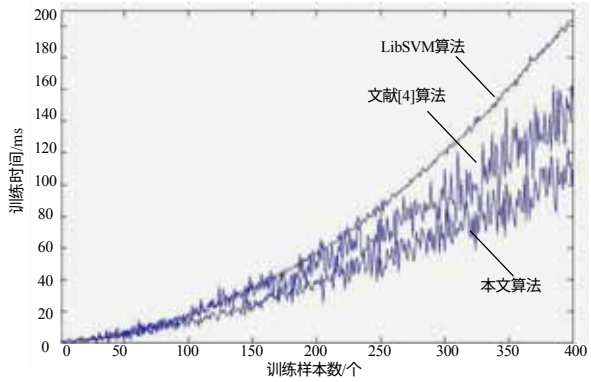

图 3　单个样本训练时间比较

实验结果表明，本文算法的样本训练时间比文献[4]算法减少约 30%，比 LibSVM 减少约 50%，验证了本文算法的有效性。

5.3 预测精度比较

本文将在线 SVM 与 LibSVM 进行预测精度比较。将数据集前 200 样本作为 LibSVM 的训练样本，第 201 个~第 500 个样本为测试样本；在线 SVM 数据量阈值设为 200，只训练及预测第 201 个~第 500 个样本，除首个样本只加入训练外，其余样本先进行预测，后将其将加训练集训练。实验结果分 3 个部分，如图 4~图 6 所示，分析结果如表 1 所示。可以看出，随着训练样本数目的增多，在线 SVM 的预测精度逐步提高，当训练样本数达到 200 个时，其实际精度与 LibSVM 相差很小。

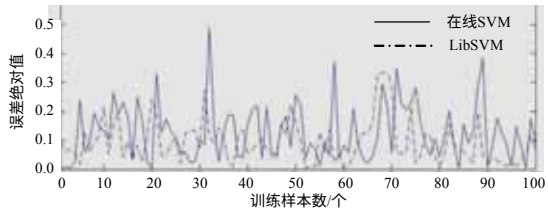

图 4　第 1 个~第 100 个样本的预测精度比较

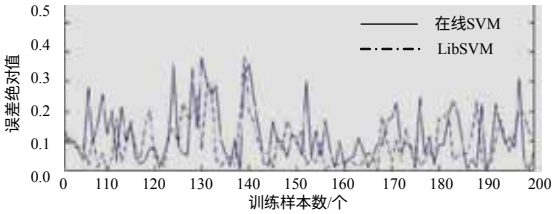

图 5　第 101 个~第 200 个样本的预测精度比较

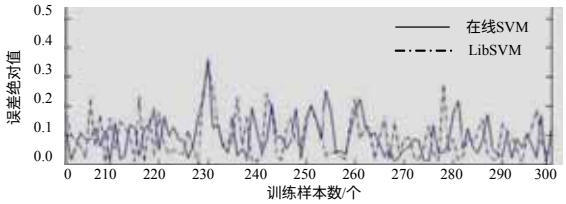

图 6　第 201 个~第 300 个样本的预测精度比较

表 1　在线 SVM 与 LibSVM 的平均误差比较

样本序号	在线 SVM	LibSVM
1~100	0.130 51	0.094 66
101~200	0.115 79	0.093 07
201~300	0.095 60	0.094 92

6　结束语

本文分析在线 SVM 回归算法[4]，针对该算法在处理位于临界位置样本时存在的问题，提出一种有效的解决方法，使算法的训练速度显著提高。改进后的在线 SVM 回归算法能够更快速有效地应用于环境污染预测。后续研究将进一步改进算法，减少对计算机内存的需求，提高预测精度，并加强其在环境监测领域内的应用。

参考文献

[1] 佟彦超. 中国重点城市空气污染预报及其进展[J]. 中国环境监测, 2006, 22(2): 69-70.

[2] 刘　永，郭怀成. 城市大气污染物浓度预测方法研究[J]. 安全与环境学报, 2004, 4(4): 60-61.

[3] Vapnik V. The Nature of Statistical Learning Theory[M]. Berlin, Germany: Springer-Verlag, 1999.

[4] Parrella F. Online Support Vector Regression[D]. Genoa, Italy: University of Genoa, 2007.

[5] Ralaivola L, d'AlcheBuc F. Incremental Support Vector Machine Learning: A Local Approach[C]//Proceedings of ICANN'01. Vienna, Austria: [s. n.], 2001.

[6] Shalev-Shwartz S, Singer Y. Tutorial on Theory and Applications of Online Learning[C]//Proc. of ICML'08. Helsinki, Finland: [s. n.], 2008.

[7] Chang Chih-Chung, Lin Chih-Jen. LIBSVM[Z]. (2008-01-01). http://www.csie.ntu.edu.tw/~cjlin/libsvm.

编辑　顾姣健

(上接第 211 页)

表 2　IISVDD 与常规增量学习算法及非增量学习算法的比较

类	增量样本数	IISVDD 训练数	IISVDD 用时/s	IISVDD 正确率/(%)	常规增量学习算法 训练数	常规增量学习算法 用时/s	常规增量学习算法 正确率/(%)	非增量学习算法 训练数	非增量学习算法 用时/s	非增量学习算法 正确率/(%)
0	*C*1=50	36	0.80	83.15	64	0.95	83.15	126	2.68	76.97
	*C*2=100	44	0.21	89.89	116	1.42	89.89	226	11.05	64.61
	*C*3=150	37	0.14	94.94	164	6.66	94.94	376	75.70	94.94
1	*C*1=50	36	0.93	80.22	67	1.28	80.22	139	3.45	70.88
	*C*2=100	42	0.18	84.07	118	1.71	84.07	239	16.17	56.04
	*C*3=150	49	0.21	91.76	183	7.41	70.88	389	85.28	91.76
2	*C*1=50	39	0.73	84.18	65	1.07	84.18	130	2.62	72.32
	*C*2=100	44	0.17	88.14	119	1.80	88.14	230	14.33	88.14
	*C*3=150	43	0.17	90.96	171	5.74	67.23	380	78.89	90.96
3	*C*1=50	28	0.97	76.50	65	1.14	76.50	139	2.93	49.73
	*C*2=100	41	0.16	83.61	119	1.64	83.61	239	15.61	83.61
	*C*3=150	44	0.17	86.89	170	5.03	57.92	389	83.98	86.89
4	*C*1=50	56	1.05	78.45	75	1.24	79.07	137	3.02	78.45
	*C*2=100	70	0.43	87.29	120	1.62	87.29	237	15.27	87.29
	*C*3=150	74	0.49	91.16	174	5.35	90.61	387	89.34	69.61

6　结束语

本文提出了一种改进的 SVDD 增量学习算法 IISVDD，算法找出了原始样本集中非 SV 集中可能成为 SV 的样本，同时挖掘出了新增样本中对增量学习有影响的样本群。对标准数据集的实验结果表明，本算法在保障分类正确率的同时，有效压缩了训练集的大小，提高了训练速度。

参考文献

[1] Tax D M J, Duin R P W. Support Vector Data Description[J]. Machine Learning, 2004, 54(1): 45-66.

[2] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York, USA: Springer Verlag, 1999.

[3] Xin Dong, Wu Zhaohui, Zhang Wanfeng. Support Vector Domain Description for Speaker Recognition[C]//Proc. of 2001 IEEE Signal Proocessing Society Workshop. Falmouth, UK: [s. n.], 2001.

[4] 杨　敏, 张焕国, 傅建明, 等. 基于支持向量数据描述的异常检测方法[J]. 计算机工程, 2005, 31(3): 39-42.

[5] 李凌均, 张周锁, 何正嘉. 基于支持向量数据描述的机械故障诊断研究[J]. 西安交通大学学报, 2003, 37(9): 910-913.

[6] 燕东渭, 孙田文, 杨　艳, 等. 支持向量数据描述在西北暴雨预报中的应用实验[J]. 应用气象学报, 2007, 18(5): 676-681.

[7] 徐　图, 罗　瑜, 何大可. 超球体单类支持向量机的 SMO 训练算法[J]. 计算机科学, 2008, 35(6): 178-180.

[8] 李　瑜, 郑敏娟, 程国建. 基于支持向量数据描述的分类方法研究[J]. 计算机工程, 2009, 35(1): 235-236, 239.

[9] Syed N, Liu H, Sung K. Incremental Learning with Support Vector Machines[C]//Proc. of the Workshop on Support Vector Machines at the International Joint Conference on Artificial Intelligence. Stockholm, Sweden: Morgan Kaufmann, 1999: 876-892.

[10] 曾文华, 马　健. 一种新的支持向量机增量学习算法[J]. 厦门大学学报, 2002, 41(6): 687-690.

[11] Alpaydin E, Kaynak C. UCI Repository of Machine Learning Databases[EB/OL]. (1998-07-01). http://www.ics.uci.edu/-mlearn/MLRepository.html.

编辑　任吉慧