第 35 卷　第 4 期　世界科技研究与发展　Vol. 35　No. 4
2013 年 8 月　459 - 461 页　WORLD SCI-TECH R&D　Aug. 2013　pp. 459 - 461

半监督增量式 SVM 在故障诊断中的应用研究*

罗显科**　柴　毅　李华锋　梁奕欢
（重庆大学自动化学院，重庆 400044）

摘　要：基于半监督学习能够有效降低人工标注成本，以及增量学习可以加快训练速度，避免数据量大时训练时间过长等特性，本文提出了一种半监督增量式 SVM 算法。在算法中，首先对已标记样本进行训练得到初始分类器，然后利用此分类器对新增样本进行标记，最后结合 KKT 条件选择合适的样本对分类器进行更新。每当有新样本加入便执行以上过程，以保证分类器得到及时更新。将该算法运用于 6135D 型柴油机的故障诊断中，并与传统 SVM 算法和增量式 SVM 算法进行了对比，证实了本文所提算法的可行性与有效性。

关键词：支持向量机（SVM）；半监督学习；增量学习；故障诊断
中图分类号：TP183　　**文献标识码**：A　　doi：10.3969/j.issn.1006 - 6055.2013.04.005

Application Research of Semi-supervised Incremental SVM on Fault Diagnosis*

LUO Xianke**　CHAI Yi　LI Huafeng　LIANG Yihuan
（College of Automation，Chongqing University，Chongqing 400044）

Abstract：For that the employment of the semi-supervised learning based approach decreases the cost of manually-marking，and the employment of the incremental learning method gain the learning speed，which avoid the long-time learning when the number of the data is too large. a semi-supervised incremental learning SVM（Support Vector Machine）algorithm is proposed. In this algorithm，the initial classifier is obtained by training the pre-labeled samples. Then the new samples are labeled by this classifier. At last，KKT constraint is used for the selection of the proper samples which are employed for the updating of classifier. The classifier is renewed by new samples in this process. In the last，this algorithm is applied into the fault diagnosis of the 6135D diesel engine. A performance comparison between the conventional SVM algorithm and semi-supervised incremental learning SVM algorithm are also made. The experiment results demonstrate the effectiveness of the algorithm.

Key words：Support Vector Machine（SVM）；semi-supervised learning；incremental learning；fault diagnosis

1　引言

随着科学和技术的不断进步，设备的功能和自动化水平越来越高，设备结构越来越复杂。设备不同模块之间交叉关联耦合，导致设备故障呈现出非线性、不确定性，故障一旦发生，极可能造成重大损失。因此，快速准确的对其进行诊断，以便及时排除故障就显得至关重要。故障诊断由于相关知识获取困难，以至于存在已知知识规模小、推广能力差、实时性不强等问题[1]。针对这些问题，国内外学者和科研人员对智能故障诊断系统进行了深入细致的研究，提出了基于神经网络、基于模糊逻辑等智能故障诊断算法[2]。然而神经网络存在过拟合、易陷入局部极值、推广能力差等问题[3]，模糊逻辑存在过分依赖专家知识、模糊规则库难建立等瓶颈，而支持向量机（Support Vector Machine，SVM）[4]能有效避免这些问题。

然而，传统的 SVM 算法需要对训练数据进行预先标注，当每次有新数据加入时，都需要对所有样本进行重新训练，这导致大量的资源消耗。此外，随着样本数量的逐渐增大，训练时间也不断增加，这不仅不能满足实时性要求，且训练出来的模型推广能力也受到极大限制。而机械设备的故障诊断是一个连续的过程。随着设备运行，时刻有新的未经标注的数据产生并加入到训练集中，以保证分类模型能够得到及时更新。鉴于半监督学习[5]能有效减少人工标注成本，以及增量式学习[6-8]能提升训练速度，避免数据量大时训练时间过长等特性，本文提出了一种半监督增量式 SVM 算法。该算法通过使用半监督学习方法解决样本标注问题，引入增量学习算法来解决运算速度慢、训练时间长的问题。

最后将该算法应用到柴油机故障诊断中，实验结果表明，半监督增量式 SVM 算法是可行的，并且用较少的时间获得了较高的诊断精度。

2　SVM 原理与增量学习算法

2.1　SVM 基本理论

支持向量机（SVM）[9,10]方法是建立在统计学习理论的 VC 维（Vapnik-Chervonenkis Dimension）理论基础上的，它追求结构风险最小化（Structural Risk Minimization，SRM），可根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷，以期获得最好的泛化能力。它是从线性可分情况下的最优分类面发展而来的。所谓最优分类面就是要求不仅能够将不同的类别正确分开，还要使分类间隔最大。对于高维空间，最优分类面就表现为一个超平面。SVM 算法的一般表达为：

$$\begin{aligned} &\max: \quad Q(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \\ &s.t. \qquad \sum_{i=1}^{n} y_i\alpha_i = 0 \\ &\qquad 0 \leqslant \alpha_i \leqslant C \qquad i = 1,\cdots,n \end{aligned} \tag{1}$$

式中，α 为每个约束条件对应的 Lagrange 乘子；$K(x_i,x_j)$ 为满足 Mercer 条件的核函数；C 为设定的惩罚因子。最优分类

* 国家自然科学基金（60974090），重庆市攻关项目（2010ac3055）资助
** E-mail：lockelk@126.com；Tel：13637948841

面的求解问题就转化为了一个凸二次规划寻优问题。相应的,决策函数可表示为:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i^* y_i K(x_i, x) + b^*\right) \quad (2)$$

对于式(1)这种最优化模型,最优解必然满足 KKT 条件,此处可描述为:

$$\begin{aligned} &\alpha_i = C \Rightarrow y_i f(x_i) \leqslant 1 \\ &0 < \alpha_i < c \Rightarrow y_i f(x_i) = 1 \\ &\alpha_i = 0 \Rightarrow y_i f(x_i) > 1 \end{aligned} \quad (3)$$

式中,当 $\alpha_i = C$ 时,x_i 在分类间隔以内,此时可能为错分;当 $0 < \alpha_i < c$ 时,x_i 在分类间隔上,就是我们所关注的支持向量;$\alpha_i = 0$ 时,x_i 在分类间隔以外。可见,满足 KKT 条件[11]可表示为 $y_i f(x_i) \geqslant 1$。也就是说,所有样本中满足 KKT 条件的必为那些位于分类器的分类间隔之外且被正确分类的样本以及位于分类间隔之上的支持向量。相反的,我们可以得出违反 KKT 条件的表达为 $y_i f(x_i) < 1$。

2.2　基于 KKT 条件的 SVM 增量学习

由以上分析可知,通过 KKT 条件对训练样本集进行筛选可以近似最大限度的保留影响最终分类器的特征信息,舍弃无用样本。样本选择时只需要判断该样本是否违反了 KKT 条件,这样不仅减少了计算量、缩短了训练时间,而且更利于实现。因而,采用 KKT 条件作为分类标准比使用判别函数更加合理。

设原始样本集为 x_k,由原始样本集训练得到的分类器为 Γ_k、支持向量集为 SV_k、非支持向量集为 $\overline{SV_k}$,新增样本集为 x_{k+1}。那么,基于 KKT 条件的 SVM 增量学习算法主要步骤可以描述如下[6]:

Step1:用原始样本集 x_k 所得到的分类器 Γ_k 检测新增样本集 x_{k+1} 中是否包含违背 KKT 条件的样本。若有,进入 Step2;否则,分类器 Γ_k 即为最终分类器 Γ,算法终止。

Step2:对新增样本集 x_{k+1} 进行训练,得到新的分类器 Γ_{k+1},以及支持向量集 SV_{k+1} 和非支持向量集 $\overline{SV_{k+1}}$。用分类器 Γ_{k+1} 检测原始样本集 x_k 中是否包含违背 KKT 条件的样本。若有,进入 Step3;否则,分类器 Γ_{k+1} 即为最终分类器 Γ,算法终止。

Step3:将两个样本集中所有违反 KKT 条件的样本与 $\overline{SV_k}$ 和 $\overline{SV_{k+1}}$ 组合成为一个新的训练样本集,进行训练后得到新的分类器即作为最终分类器 Γ,算法结束。

3　半监督增量式 SVM 算法

3.1　半监督学习与增量式 SVM

传统机器学习技术按照其利用的训练集是否已标记可分为两类:只利用已标记样本集的叫做有监督学习(Supervised Learning),如神经网络等;只利用未标记样本集的叫做无监督学习(Unsupervised Learning),如聚类、PCA 降维等。但在现实中大多样本是已标记与未标记并存。为了能够更好地利用这些数据,半监督学习应运而生。半监督学习是介于监督学习和无监督学习之间的学习技术,它所利用的数据集同时包括了已标记和未标记样本[12]。

传统的 SVM 算法是有监督学习,即需要训练集中的样本都为已标记样本,而现实中往往未标记样本的数量远远大于已标记样本,而且对数据进行标记代价很高,通常不能获取大量有着正确类别标签的样本,特别是在实时性要求较高的故障诊断中。如果能把大量未标记样本所包含的数据特征加入到学习算法的设计中去,就可以弥补传统 SVM 的这个固有缺陷,以获得更好的分类效果。这也是学者将半监督学习的思想引入到支持向量机学习算法中并将其应用于故障诊断的主要原因[5]。

然而,现实中的故障诊断面临两个问题。一是随着系统运行,数据量逐渐增大,而且获取的数据为未标记数据;二是要求算法有较高的实时性。通常,增量式 SVM 算法无需保存大量的历史数据,能有效减少对内存空间的占用,且在新一轮的训练中能充分利用历史有效数据,大大减少后序训练的时间。因而增量式 SVM 能够很好的解决数据量大和实时性要求高的问题。而半监督学习算法可以解决未标记数据的使用问题。基于该思想,本文提出了半监督增量式 SVM 算法(Semi-Supervised Incremental Support Vector Machines, SSISVM),并将其应用到柴油机的故障诊断中。

3.2　半监督增量式 SVM 算法

本文提出的半监督增量式 SVM 算法 SSISVM 不仅继承了增量 SVM 算法训练速度快,能有效解决数据量大时的存储问题等优点,而且和半监督学习算法一样可以有效利用未标记数据。该算法描述如下:

首先,用获取的已标记样本集 S_a 作为初始训练集 x_0 进行训练,得到分类器 Γ_0,支持向量集 SV_0 和非支持向量集 $\overline{SV_0}$。当有新增未标记样本加入时,执行以下步骤:

Step1:用分类器 Γ_0 对新增的未标记样本集 s_{u1} 进行分类,标记 s_{u1} 得到新增样本集 x_1;

Step2:用分类器 Γ_0 判断 x_1 中是否存在违反 KKT 条件的样本。若有,转入 Step3;否则,分类器 Γ_0 作为本轮新增未标记样本后的最终分类器 Γ,完成本轮训练。

Step3:对样本集 x_1 进行训练,得到新的分类器 Γ_1,得到支持向量集 SV_1 和非支持向量集 $\overline{SV_1}$。用分类器 Γ_1 检测初始训练集 x_0 中是否存在违反 KKT 条件的样本。若有,进入 Step4;否则,分类器 Γ_1 作为本轮新增未标记样本后的最终分类器 Γ,完成本轮训练。

Step4:将违反 KKT 条件的样本与 $\overline{SV_0}$ 和 $\overline{SV_1}$ 组合成为一个新的训练样本集,进行训练后得到新的分类器即作为本轮新增未标记样本后的最终分类器 Γ,完成本轮训练。

当有新的未标记样本集加入时,重复以上步骤,进行新的一轮训练。

4　实验与结果分析

为验证本文提出算法的有效性和优越性,本文以 6135D 型柴油机的故障诊断为例来说明。在采样频率为 25 KHz 下,通过采用加速度传感器在缸盖处分别测量出柴油机在出现故障和正常状态下的 8 个振动信号,以此构成一组数据。通过实验,共获取数据10 000余组。本文随机选取1 400组作为训练和测试数据集,包括故障情况 683 组,标记为 -1;正常情况 717 组,标记为 1。其中前 500 组作为初始训练集,501 到1 000组作为增量训练集,1 001到1 400组作为测试集。

4.1　仿真实验

为验证本文提出的半监督增量式 SVM 算法 SSISVM 的有效性，本文实验主要从训练时间和诊断精度两方面来进行分析。首先，对初始训练集进行训练，得到初始分类器；然后，进行增量训练，每次增量样本数为 50，增量次数为 10 次，得到新的分类器。每次在获得分类器的同时，获取训练时间，并对测试数据集进行分类测试，获取测试精度。以线性支持向量机为基础，选取 RBF 核函数，设置惩罚因子 C = 4，松弛变量 g = 0.025。实验平台为：CPU：AMD Athlon 4 000 + 2.11 GHz；内存：2 G；操作系统：大地 Windows7 旗舰版。仿真平台：Matlab 7.11.0(R2010b)。

4.2　结果分析

用标准 SVM，基于 KKT 条件的增量 SVM(ISVM)，以及本文提出的半监督增量式 SVM(SSISVM)三种算法分别对训练样本集进行训练，得到结果如表一所示。训练消耗时 对比如图 1 所示；分别利用三种算法训练后得到的分类模型对测试样本集进行测试，得到测试精度对比如图 2 所示。

表 1　结果对比

Table 1　Results comparison

训练样本集	SVM		ISVM		SSISVM	
	训练时间(ms)	测试精度(%)	训练时间(ms)	测试精度(%)	训练时间(ms)	测试精度(%)
初始样本集	83	89.50	83	89.50	80	89.50
增量样本集 1	73	89.25	39	89.50	41	88.50
增量样本集 2	93	90.25	27	91.25	22	89.75
增量样本集 3	95	88.75	24	89.00	21	90.00
增量样本集 4	109	89.75	24	90.25	23	90.00
增量样本集 5	115	90.50	22	91.00	22	89.75
增量样本集 6	134	89.50	24	89.75	20	89.50
增量样本集 7	167	89.75	22	89.50	19	89.75
增量样本集 8	171	90.00	21	89.50	21	89.50
增量样本集 9	183	90.75	24	88.75	20	89.75
增量样本集 10	199	90.75	24	89.75	23	89.50

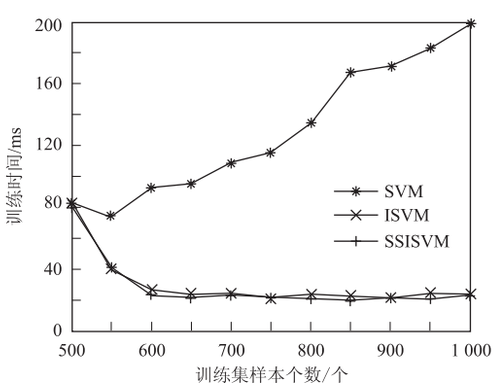

图 1　训练时间对比图

Figure 1　Trainging time comparison

由图 1 可知，传统 SVM 算法随着训练样本数量的增加，训练时间逐渐增大；而 ISVM 算法和 SSISVM 算法的训练时间不是随着训练样本数量的增加而增大，而是趋于稳定。而且 SSISVM 算法的训练时间小于 ISVM 算法，这证明了本文提出的算法在训练时间上相对于传统算法的优越性。

由图 2 可知，对于同样的测试样本集，三种算法的分类精度差别很小，都在可以接受的范围内。然而现实中绝大多数采集到的数据都是未经标记的，SSISVM 算法对于 501 到

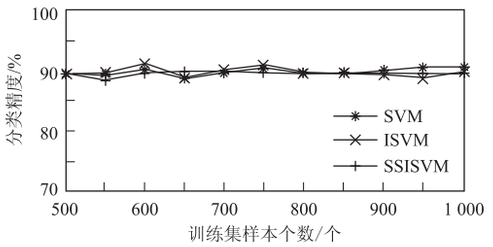

图 2　测试精度对比图

Figure 2　Test accurary comparison

1 000这 500 组样本进行训练时是假设他们为未标记样本的，并没有利用他们的类别标签，由此可见 SSISVM 可以有效利用未标记样本。

5　结束语

本文提出了半监督增量式 SVM 算法，该算法不仅继承了半监督学习以及增量式学习算法的优越性，而且克服了它们各自的不足。最后将该算法应用到柴油机故障诊断中，并与传统 SVM 和增量式 SVM 算法进行对比实验。实验结果表明了该算法的可行性，并且验证了该算法能有效克服传统 SVM 算法需要所有训练样本预先标记以及数据量较大时训练时间长等缺点。

参考文献

[1] ISERMANN R. Supervision, fault-detection and fault – diagnosis methods-An introduction[J]. Control Engineering Practice, 1997, 5(5): 639-652.

[2] 李晗，萧德云. 基于数据驱动的故障诊断方法综述[J]. 控制与决策，2011，26(1)：1-9.

[3] 曹龙汉，武明亮，何俊强，等. 基于 DE-SVM 的柴油机气门故障诊断方法及应用[J]. 仪器仪表学报，2011，32(2)：323-328.

[4] HUANG Jian, HU Xiaoguang, YANG Fan. Support vector machine with genetic algorithm for machinery fault diagnosis of high voltage circuit breaker[J]. Measurement, 2011, 44(6): 1 018-1 027.

[5] 赵莹. 半监督支持向量机学习算法研究[D]. 哈尔滨：哈尔滨工程大学，2010.

[6] 王小燕. 基于加权增量的支持向量机分类算法研究[D]. 杭州：浙江大学，2008.

[7] 萧嵘，王继成，孙正兴. 一种 svm 增量学习算法 α – ISVM[J]. 软件学报，2001，12(12)：1 818-1 823.

[8] 刘叶青，刘三阳，谷明涛. 一种改进的支持向量机增量学习算法[J]. 计算机工程与应用，2008，44 (10)：142-143，187.

[9] CRISTIANINI N, SHAWE T J. An introduction to support vector machines and other kernel-based learing methods[M]. Cambridge: Cambridge University Press, 2000.

[10] VAPNIK V N. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995.

[11] PRONOBIS A, LUO J, Caputo B. The more you learn, the less you store: Memory-controlled incremental SVM for visual place recognition[J]. Image and Vision Computing, 2010, 28(7): 1 080-1 097.

[12] SHI Lei, MA Xinming, XI Lei, et al. Rough set and ensemble learning based semi-supervised algorithm for text classification[J]. Expert Systems with Applications, 2011, 38(5): 6 300-6 306.