文章编号:1001-2486(2007)03-0065-06

一种支持向量机增量学习淘汰算法*

廖东平,魏玺章,黎 湘,庄钊文
(国防科技大学 电子科学与工程学院,湖南 长沙 410073)

摘 要:针对大规模数据集的分类问题,支持向量机的训练成为一个难题。增量学习是解决这一难题的思路之一。分析了新增样本加入训练集后支持向量集的变化情况,提出了一种基于密度法的支持向量机增量学习淘汰算法,淘汰了对最终分类无用的样本,在保证测试精度的同时减少了训练时间。实验仿真证明这种算法是有效的。

关键词:支持向量机;增量学习;支持向量
中图分类号:TN959 **文献标识码**:A

A Removing Algorithm for Incremental Support Vector Machine Learning

LIAO Dong-ping, WEI Xi-zhang, LI Xiang, ZHUANG Zhao-wen
(College of Electronic Science and Engineering, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: The training of support vector machine is a difficult issue in classifying large-scale data set. Incremental learning is one of the solutions to the difficulty. After new samples were added to training set, the possible changes of support vector set, were analyzed and a removing algorithm based on density for incremental support vector machine learning was presented. It discarded useless samples, kept the testing accuracy and reduced the training time. Experiments show the validity of this algorithm.

Key words: support vector machine (SVM); incremental learning; support vector (SV)

支持向量机(support vector machine,简称SVM)是一种建立在结构风险最小化原则基础之上的全新的机器学习方法[1],具有很强的学习能力和泛化性能,能够较好地解决小样本、高维数、非线性、局部极小等实际问题,因此,这种方法为雷达目标识别提供了一种新的有潜力的途径[2-4]。

但是作为一种新兴的技术,SVM目前还存在着一些局限性,比如当训练样本集的规模很大时,这种情况在雷达目标识别过程中经常出现,此时花费的存储空间和运算时间的代价十分巨大,迫切需要寻找解决的思路;再比如在雷达目标识别过程中,要在训练初期就收集一个非常完整的训练集是非常困难甚至是难以实现的,而更多的情况下,样本是不断加入的,即增量式地加入训练样本,因此希望SVM具有这样的能力,即其学习的精度可以随着应用过程中样本集的不断积累而逐步提高。

增量学习技术(incremental learning technique)是一种得到广泛应用的智能化数据挖掘与知识发现技术。一种机器学习方法是否具有良好的增量学习功能已经成为评价其性能优劣的重要标准之一。但经典的SVM理论与增量式学习并不具备直接的相容性。SVM训练所得的支持向量(support vector,简称SV)能够完全反映分类超平面的信息,而SV通常只占训练样本很小的一部分,这对SVM的增量学习算法的构建具有重要意义[5]。

文献[5]最早开始基于SVM增量学习的研究,在此后的几年之内,不断有关于此方面的研究见诸于各种杂志与会议[5-13]。文献[6-9]利用初始SVM的分类结果对新增样本进行划分,并通过循环迭代方式寻找最优分类超平面。这些算法必须保存全部的历史样本,并且在初始样本不足的情况下,后继的增量学习将出现"振荡"现象,影响了训练的收敛速度。文献[5,10-13]引入淘汰机制,丢弃或完全丢弃非

* 收稿日期:2006-12-23
基金项目:国家部委基金资助项目(41303040203)
作者简介:廖东平(1977—),男,讲师,博士。

SV样本以加速SVM的在线学习速度和减少对历史样本的存储,但这些淘汰机制缺乏新增样本对SV集影响的考虑,在丢弃的同时也导致分类知识的丢失,因为被丢弃的样本同样含有分类知识,甚至在后续学习过程中可能成为SV。

1 支持向量机简介

1.1 支持向量机

给定分类问题,其训练样本集为$\{x_i, y_i\}$, $i = 1,2,3,\cdots,n$, $\{x_i\} \in R^l$, $y_i \in \{\pm 1\}$。SVM的目标就是根据结构风险最小化原则,构造一个最优分类超平面,使得类别间的分类间隔最大[14]。在一般的线性不可分情况下,SVM把分类间隔最大最终归结为如下凸二次优化问题:

$$\begin{cases} \max\ W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j K(x_i \cdot x_j) \\ s.t.\quad \sum_{i=1}^{n} y_i\alpha_i = 0, \quad 0 \leqslant \alpha_i \leqslant C, \quad i = 1,2,3,\cdots,n \end{cases} \tag{1}$$

其中,$\alpha = (\alpha_1,\cdots,\alpha_n)$,是拉格朗日乘子,$K(\cdot)$为核函数,$C$为某个指定的常数,起控制对错分样本惩罚程度的作用。由上述问题得到最优解α,则SVM的分类函数为:

$$f(x) = \mathrm{sign}\{\sum_{i=1}^{n} y_i\alpha_i K(x \cdot x_i) + b\} \tag{2}$$

1.2 KKT条件

最优解$\alpha = (\alpha_1,\cdots,\alpha_n)$使得每个样本满足优化问题的KKT条件[15]:

$$\alpha_i = 0 \Rightarrow y_i f(x_i) \geqslant 1, \quad 0 < \alpha_i < C \Rightarrow y_i f(x_i) = 1, \quad \alpha_i = C \Rightarrow y_i f(x_i) \leqslant 1 \tag{3}$$

其中非零的α_i为SV。考虑函数系$f(x) = h$,可知$f(x) = 0$为分类面,$f(x) = \pm 1$为分类间隔面,则$\alpha = 0$对应的样本分布在分类器分类间隔面之外,$0 < \alpha < C$对应的样本位于分类间隔面之上,$\alpha = C$对应的样本位于分类间隔面关于本类的异侧。

1.3 增量学习后支持向量集变化分析

定理1 $f(x)$为SVM分类决策函数,$\{x_i, y_i\}$为新增样本。满足KKT条件的新增样本将不会改变SV集,违背KKT条件的新增样本将使SV集发生变化。违背KKT条件的样本可以分为三类:(1)位于分类间隔中,与本类在分类边界同侧,可以被原分类器正确分类的样本,满足$0 \leqslant y_i f(x_i) < 1$;(2)位于分类间隔中,与本类在分类边界异侧,被原分类器分类错误的样本,满足$-1 \leqslant y_i f(x_i) \leqslant 0$;(3)位于分类间隔外,与本类在分类间隔异侧,被原分类器分类错误的样本,满足$y_i f(x_i) < -1$。

综合以上三类情况可得,违背KKT条件等价于$y_i f(x_i) < 1$。

定理1的证明可参见文献[15]。定理1表明,对新增样本再学习得到新的SVM分类器时,KKT条件比分类函数的分类判断更合理,分类错误是样本违反KKT条件的特定情况,并且只有违背KKT条件的样本才会影响增量学习后的SV集。依据这个性质,可以大大简化SVM对新增样本的操作。首先,如果新增样本不存在违背原SVM的KKT条件的样本,则说明原SVM已经包含了这部分样本的信息,不需要再对这部分样本学习;其次,如果新增样本中存在违背KKT条件的样本,说明原SVM没有包含这部分样本的信息,需要对这部分样本进行学习。这种学习不仅没有破坏原有的SVM包含的信息,而且大大简化了对于新增样本的学习,所以这种学习是可积累性的。

定理2 新增样本违背KKT条件,则原样本集中非SV可能转化为SV。

定理2的证明可参见文献[7,15]。定理2表明,在增量学习中,只考虑新增样本和原来的SV集的虽然符合定理1,但是这样可能会丢失原来训练集中的信息,而且也不能十分有效地淘汰无用样本。因此对于无用样本的淘汰,必须根据具体情况,选择最有把握的方式来进行。

2 基于密度法的增量学习淘汰算法

对于包括新增样本集中满足 KKT 条件和原样本集中非 SV 的样本,必须采取有效的方法淘汰掉无用样本,保留重要信息。如何有效地淘汰样本,既保证训练的精度又提高训练的速度,成为本文研究的重心。SV 从物理意义上来说,就是在两类的交遇区中,那些靠得最近但是又属于不同类的样本,即两类相对部分的边界向量。因此边界向量成为 SV 的可能性要远大于非边界向量,这从另外一个方面也可以说明,非边界向量对后继的训练影响甚微,可以将其淘汰掉,而只保留边界向量。

对非边界向量的淘汰亦即对边界向量的确定,因此现有的一些边界向量确定方法均可适用[16-17]。但这些方法都需要事先判定样本集合是否线性可分,这会给实际应用带来困难。

如果将一个样本周围(一定范围内)同类样本的数量定义为样本密度,则样本密度从一定程度上反映了此样本在整个样本集合中的相对位置情况。结合雷达目标识别的样本特性,一般而言,同类样本之间具有一定的聚集性,反映在样本密度上,就是处于不同空间位置的样本密度各有不同,类内点的样本密度一般要比类边界点的样本密度要大,因而可以考虑从样本密度的角度出发淘汰类的非边界向量。此外,利用样本密度的不同来实现对非边界向量的淘汰,不需要事先判定样本集合是否线性可分,这在实际应用中具有很大的优势。本文基于以上考虑,提出了一种基于密度法的增量学习淘汰算法,该方法能准确地提取出边界向量,淘汰掉非边界向量,并且计算简单,易于实现。

首先给出密度的定义,使用一个样本的某一周围邻域内的同类样本数量作为此样本的样本密度,其计算公式如下:

$$\rho(x_i) = M(\{x \mid d(x, x_i) \leqslant T\}) \tag{4}$$

式中,x_i 和 x 为同类样本;$\rho(x_i)$为样本 x_i 的密度;$M(X)$为样本集合 X 的样本数量;$d(x, x_i)$为样本 x_i 和 x 相似性度量的一个函数;T 为给定的阈值。

根据不同的相似性度量公式,有以下两种计算密度的方法:

第一种方法为计算不超过距离阈值的样本数量:

$$\rho(x_i) = M(\{x \mid D(x, x_i) \leqslant T\}) \tag{5}$$

有许多种表示距离的函数,其中常用的为 Minkowsky 距离,即

$$D_\lambda(x, x_i) = \left[\sum_{j=1}^{l} \mid x(j) - x_i(j) \mid^\lambda\right]^{1/\lambda} \tag{6}$$

式中,l 为样本的维数,λ 为正整数,$\lambda = 2$ 时的距离为 Euclidean 距离。

第二种方法为计算与样本 x_i 夹角的余弦值大于给定阈值的样本的数量:

$$\rho(x_i) = M(\{x \mid \cos(x, x_i) \geqslant T\}) \tag{7}$$

两种方法的阈值 T 的确定方法有很多种,最常用的方法是计算出任意同类样本之间的距离或其夹角的余弦值,再对这些值求平均获得。

从前面的密度公式可以看出,需要计算任意两同类样本之间的距离或者其夹角的余弦值。即假如某类样本数量为 n,则需要计算 n^2 个距离或者余弦值。假如每个距离需要 4 个字节的话,总共需要 $4n^2$ 字节的存储空间。如果 n 比较小,则计算量和存储量还是可以容忍的。但是当 n 比较大时,则需要花费很长时间计算距离或余弦,且花费很大的存储空间。假如 $n = 5000$,则需要 100MB 的内存空间,这将大大降低计算机的运行速度。

在样本数量较大时,为提高本算法的运行速度,减小内存开销,可采用工兵探雷法计算各样本点周围的样本密度,即计算每个样本一定的周围邻域 $U(x_i, \delta)$内同类样本的数量。这样既免去了大量的距离计算,同时大大减小了内存开销,只需要存储两个 l 维的密度向量,其占用的存储空间仅为距离法的 $1/n$。

首先,将同类样本空间的每一维平均分成 m 份,即将同类样本每一维数据的最大值减去最小值除以 m,这样就分别得到了同类样本的单位空间邻域向量。

$$z = 1/m[\max(x_i(1)) - \min(x_i(1)), \cdots, \max(x_i(l)) - \min(x_i(l))], \quad i \in \{1,2,3,\cdots,n\}$$

当求第 i 个样本的邻域样本密度时，首先确定其邻域 $U(x_i,\delta) = r \cdot z$，其中 r 为系数，然后便可得到其邻域密度为

$$\rho(x_i) = M(\{x \mid |x - x_i| \leqslant r \cdot z\}) \tag{8}$$

向量 z 的大小由 m 来确定，在实际应用中，为了保证单位空间邻域内有一定数量的样本，一般取 m 为同类训练样本数的 $1/k$，这里 $k \leqslant 10$。r 的取值不宜过大，应以保证使类内样本的样本密度大于边界样本的样本密度为原则。因此边界向量应该满足：

$$\rho(x) \leqslant M_0 \tag{9}$$

其中 M_0 为边界向量选取参数，其值一般取所有同类样本密度的平均值。阈值 M_0 的取值，对边界向量的分布区域的大小是至关重要的，如果阈值选定合适的话，边界向量集合就是包含了 SV 集合的最小集合，更有甚者，边界向量集合就是 SV 集合。

3 SVM 增量学习算法

3.1 符号意义

Ω_o^k 表示由第 k 次训练用原始样本集得到的 SVM 分类器；

X_o^k 表示第 k 次训练用原始样本集；

X_I^k 表示新增样本中满足 Ω_o^k 的 KKT 条件的样本集；

$X_o^{sv_k}$ 表示 Ω_o^k 的 SV 集；

X_I^k 表示第 k 次取出的新增样本（$k=0$ 时表示初始样本集）；

$X_o^{nsv_k}$ 表示 Ω_o^k 的非 SV 集；

X_I^{nk} 表示新增样本中违背 Ω_o^k 的 KKT 条件的样本集；

X_u^{nk} 表示 $X_o^{nsv_k}$ 和 X_I^{nk} 的并集；

X_r^{nk} 表示淘汰计算后 X_u^{nk} 的剩余部分。

3.2 学习过程

(1)判断训练样本集是否空集，空则训练结束，否则转(2)；

(2)从训练样本集中取出新增样本 X_I^k。若 $k=0$，则由 X_I^0 训练得到原始分类器 Ω_o^0，$X_o^1 = X_I^0$，$k = k+1$，转(1)；若 $k \neq 0$，则转(3)；

(3)检验 X_I^k 中的样本是否违背 Ω_o^k 的 KKT 条件，如果没有样本违背，则转(1)，否则根据结果 X_I^k 被分为 X_I^{nk} 和 X_I^{vk}；

(4)将 $X_o^{nsv_k}$、X_I^{nk} 合并得 X_u^{nk}，根据标示符，将集合分为正例样本集 A^+ 和负例样本集 A^-，并分别根据淘汰规则进行处理，淘汰对后继训练影响不大的样本，得到剩余正例样本集 A_r^+ 和剩余负例样本集 A_r^-，合并二者得 X_r^{nk}；

(5)将 X_r^{nk}、X_I^{vk}、$X_o^{sv_k}$ 合并得到 X_o^{k+1}，对其训练得到新的分类器 Ω_o^{k+1}，并生成 $X_o^{sv_{k+1}}$ 和 $X_o^{nsv_{k+1}}$，$k = k+1$，转(1)。

图 1 给出了 SVM 增量学习过程示意图。

3.3 淘汰规则

设要处理的样本集为 A，设定一合适的阈值 M_0，根据淘汰法中的定义计算各个样本的样本密度 $\rho(x_i)$，淘汰 $\rho(x_i) > M_0$ 的样本，保留 $\rho(x_i) \leqslant M_0$ 的样本（即保留边界向量），得到剩余样本集 A_r。

4 实验结果

依据上面提出的算法，针对两组数据集进行实验，一组为某机载雷达实测多普勒数据集，共有 710 个样本，样本维数为 32（不包括分类属性），样本类型包括吉普车和坦克两种，其中初始训练样本有 113 个，测试样本有 276 个，并将剩余样本随机分为 3 组，构成 3 个增量学习样本集；另一组为某机载雷达实测一维距离像数据集，共有 1889 个样本，样本维数也为 32（不包括分类属性），样本类型包括卡车和坦克两种，其中初始训练样本有 410 个，测试样本有 616 个，并将剩余样本随机分为 3 组，构成 3 个增量学习

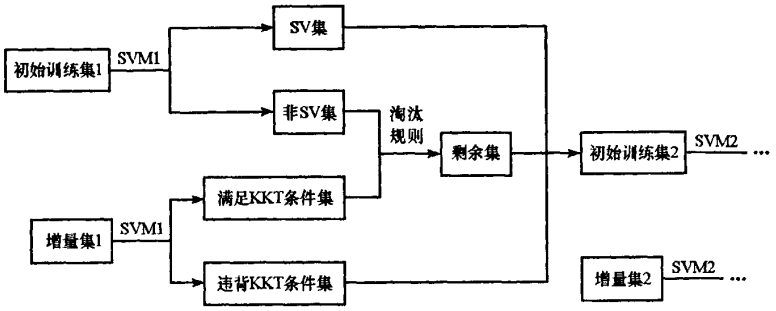

图1 支持向量机增量学习过程示意图

Fig.1 Procedure scheme of incremental support vector machine learning

样本集。

实验中,针对以上数据集,分别采用常规 SVM 训练算法和本文提出的增量 SVM 算法进行训练,训练结果如表 1 和表 2 所示(实验软件环境:Matlab 6.5/Windows XP;硬件环境:PⅣ2.4GHz/256MB)。在 SVM 训练中,C 取 1000,采用径向基内积函数,参数取 0.5。

表1 多普勒数据集增量学习情况

Tab.1 Result of incremental learning based on Doppler data set

训练集	增量集样本数	常规 SVM			增量 SVM		
		实际训练集样本数	时间/s	测试精度	实际训练集样本数	时间/s	测试精度
初始训练	113	113	2.97	87.32%	113	2.97	87.32%
增量 1	126	239	15.19	90.22%	181	9.75	88.77%
增量 2	94	333	38.87	91.67%	176	9.78	89.86%
增量 3	101	434	87.97	91.31%	192	11.28	90.22%

表2 一维距离像数据集增量学习情况

Tab.2 Result of incremental learning based on one-dimensional range profile data set

训练集	增量集样本数	常规 SVM			增量 SVM		
		实际训练集样本数	时间/s	测试精度	实际训练集样本数	时间/s	测试精度
初始训练	410	410	85.17	83.77%	410	85.17	83.77%
增量 1	241	651	342.62	85.23%	432	103.44	84.74%
增量 2	344	995	1085.30	85.88%	554	217.53	85.55%
增量 3	278	1273	2136.60	85.88%	659	359.44	85.23%

从表 1 和表 2 可以看出,除了在初始训练阶段两算法的训练速度相等外,在后继的增量学习阶段,和常规 SVM 学习方法相比,本增量学习算法显著减少了历史数据的存储,在不明显降低训练精度的同时显著提高了训练速度,为 SVM 在线学习提供了一条有效的途径。

5 结束语

本文研究了 KKT 条件和样本之间的关系,分析了样本增加后 SV 集的变化情况,提出了一种基于密度法的 SVM 增量学习淘汰算法,对训练样本进行有效淘汰,实现了 SVM 的增量学习。通过对实测数据集的实验结果表明,使用该方法进行增量学习在保证训练精度的同时,能有效地提高训练速度并降低存储空间的占用。

参考文献:

[1] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York: Springer, 1995.

[2] Zhang L, Zhou W D, Jiao L C. Radar Target Recognition Based on Support Vector Machine[C]//Proceedings of the 2000 International Conference on Signal Processing, 2000, 3: 1453 - 1456.

[3] Li Y, Ren Y, Shan X M. Radar HRRP Classification with Support Vector Machines[C]//Proceedings of the 2001 International Conference on Info-tech and Info-net, 2001, 1: 218 - 222.

[4] Wang X D, Wang J Q. Support Vector Machine for HRRP Classification[C]//Proceedings of the 7th International Symposium on Signal Processing and Its Applications, 2003, 1: 337 - 340.

[5] Syed N A, Liu H, Sung K K. Incremental Learning with Support Vector Machines[C]//Proceedings of Workshop on Support Vector Machines at the International Joint Conference on Artificial Intelligence, Sweden: Stockholm, 1999: 272 - 276.

[6] 萧嵘,王继成,孙正兴,等. 一种SVM增量学习算法[J].南京大学学报(自然科学版),2002,38(2):152 - 157.

[7] 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(自然科学版),2002,41(6):687 - 691.

[8] An J L, Wang Z G, Ma Z P. An Incremental Learning Algorithm for Support Vector Machine[C]//Proceedings of the 2th International Conference on Machine Learning and Cybernetics, Xi'an, 2003(2): 1153 - 1156.

[9] Li Z W, Zhang J P, Yang J. A Heuristic Algorithm to Incremental Support Vector Machine Learning[C]//Proceedings of the 3th International Conference on Machine Learning and Cybernetics, Shanghai, 2004(3): 1764 - 1767.

[10] Mitra P, Murthy C A, Pal S K. Data Condensation in Large Databases by Incremental Learning with Support Vector Machines[C]//Proceedings of the 15th International Conference on Pattern Recognition, Spain, 2000(2): 708 - 711.

[11] Xiao R, Wang J C, Zhang F Y. An Approach to Incremental SVM Learning Algorithm[C]//Proceedings of the 12th International Conference on Tools with Artificial Intelligence, 2000(1): 268 - 273.

[12] 萧嵘,王继成,孙正兴,等. 一种SVM增量学习算法——ISVM[J].软件学报,2001,12(12):1818 - 1824.

[13] Cauwenberghs G, Poggio T. Incremental and Decremental Support Vector Machine Learning[C]//Advances in Neural Information Processing Systems, Cambridge MA: MIT Press, 2001(13): 409 - 415.

[14] Burges C J C. A Tutorial on Support Vector Machines for Pattern Recognition[M]. Boston: Kluwer Academic Publishers, 1998.

[15] 周伟达,张莉,焦李成. 支撑矢量机推广能力分析[J].电子学报,2001,29(5):590 - 594.

[16] Zhang L, Zhou W D, Jiao L C. Pre-extracting Support Vectors for Support Vector Machine[C]//Proceedings of the 5th International Conference on Signal Processing, 2000(3): 1432 - 1435.

[17] Ding A L, Liu F, Li Y. Pre-extracting Support Vector by Adaptive Projective Algorithm[C]//Proceedings of the 6th International Conference on Signal Processing, 2002(1): 21 - 24.

(上接第64页)

4 结论

本文提出了一个基于erasure的数据复制算法Dyre,编码块采用动态分配算法存储到不同的节点中,在每个节点的前驱和后继节点中保存编码块副本,利用数据迁移和恢复算法保证数据在节点失效情况下可用,采用重建算法保证编码数据块的数量大于数据还原所需的块数。模拟实验表明该算法能够获得很高数据的可用性,在相同的环境当中获得的数据可用性高于全复制和块复制算法。

参考文献:

[1] MacWilliams F J, Sloane N J A. The Theory of Error-correcting Codes, Part I[M]. North-Holland Publishing Company, Amsterdam, New York, Oxford, 1977.

[2] Kubiatowicz J, et al. Oceanstore: An Architecture for Global-scale Persistent Storage[C]//Proceedings of ASPLOS 2000, Cambridge, Massachusetts, Nov. 2000.

[3] Brunskill E. Building Peer-to-peer Systems with Chord, a Distributed Lookup Service[C]//HOTOS'01: Proceedings of the Eighth Workshop on Hot Topics in Operating Systems, 2001.

[4] Xu Z C, Mahalingam M, Karlsson M. Turning Heterogeneity into an Advantage in Overlay Routing[R]. Infocom'03, 2003.