一类新型快速模糊支持向量机

施其权,李小明,肖辞源
(西南石油大学 理学院,四川 成都 610500)

摘 要:针对一般模糊支持向量机训练时间过长,训练效率低下的问题,通过定义了一种新的隶属度函数的方法,来改进算法,从而得到了一种快速模糊支持向量机。本算法中的新定义的隶属度函数能够对离分类超平面较远、不可能成为支持向量的数据赋予较小的隶属度,使训练样本集中的数据大大减少。同时,在将二类模糊支持向量机推广到 k 类时,采用了DAGSVMs方法,进一步提高了多类分类问题的分类效率。实验表明,提出的快速模糊支持向量机在保证测试精度的同时,减少了训练时间。

关键词:模糊支持向量机;隶属度函数;边缘数据
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2010)02-0103-03

A Kind of Novel Fast Fuzzy Support Vector Machines

SHI Qi-quan, LI Xiao-ming, XIAO Ci-yuan
(Dept. of Sciences, Southwest Petroleum University, Chengdu 610500, China)

Abstract: Proposes a kind of fast fuzzy support vector machines to solve the problem of long training time and low training efficiency by improved algorithm. The definition of membership function in the new algorithm can give the smaller memberships to the data away from separating hyperplane which can not be the support vector, so it reduces the data in the training sample set. Meanwhile, it expends to multi-classes fast fuzzy support vector machines using DAG. Experimental results indicate that the algorithm reduces the training time on the premise of guaranteeing the testing accuracy.

Key words: FSVMs; membership function; edge data

0 引 言

支持向量机(support vector machine, SVM)是在统计学习理论的VC维理论和结构风险最小原理的基础上发展起来的一种新的机器学习方法[1]。SVMs的基本思想是最大限度地分开两类训练样本,即对于一个两类问题的训练样本,构造一个分类超平面,使得分类间隔达到最大[2,3]。模糊支持向量机(FSVM)是一种改进的SVM[4]。FSVM有两种表现形式:a) 2002年由台湾学者Lin Chun-fu等人提出的,根据训练样本在训练过程中所起的作用不同,对所有数据包括异常数据赋予一个隶属度,加大对容易错分样本的惩罚,以改进SVM性能的模糊支持向量机[5];b)2001和2002年由日本学者Takuga与Shigeo提出的,针对两类问题推广到多分类问题时,决策过程中存在不可分区域,构造隶属函数,以减少不可分区域的模糊支持向量机[6,7]。支持向量机算法的复杂度依赖于数据集的大小,如何提高支持向量机的训练效率,是一个重要的研究方向[8]。文献[9]提出了一种快速模糊支持向量机(FFSVMs)。文中定义了另一种隶属度函数,改进了文献[9]的算法,同时,在将二类模糊支持向量机推广到 k 类时,采用DAGSVMs,进一步提高了多类分类问题的分类效率。

收稿日期:2009-05-30;**修回日期**:2009-08-20
基金项目:四川省教育厅重点基金项目(072A143)
作者简介:施其权(1984-),男,安徽庐江人,硕士研究生,研究方向为模糊数学、支持向量机;李小明,硕士,教授,研究方向为应用概率统计、支持向量机。

1 快速模糊支持向量机

1.1 基本思想与隶属度函数的构造

在FSVM方法中,隶属度函数的设计非常关键。目前,构造隶属度函数的方法主要是基于样本到类中心之间的距离来度量其隶属度的大小。根据SVMs的基本思想,离分类超平面最近的点就是支持向量,离分类超平面次近的点对于支持向量的获取也至关重要。因此,分类超平面附近的数据不仅容易错分,而且成为支持向量的机会多一些,而每类中心附近的样本成为支持向量的可能性要小一些,甚至根本不可能成为支

持向量[4]。基于这一思想，文献[9]对边缘数据赋予了较大的隶属度，而对类中心附近的数据赋予较小的隶属度，体现加大对容易错分样本进行惩罚这一改进策略，其定义的隶属度函数为：

$$\mu_{1i}=\frac{\|x_{1i}-\bar{x}_1\|-\min\limits_{1\le j\le l_1}\|x_{1j}-\bar{x}_1\|}{\max\limits_{1\le j\le l_1}\|x_{1j}-\bar{x}_1\|-\min\limits_{1\le j\le l_1}\|x_{1j}-\bar{x}_1\|}\quad(1)$$

$$\mu_{2i}=\frac{\|x_{2i}-\bar{x}_2\|-\min\limits_{1\le j\le l_2}\|x_{2j}-\bar{x}_2\|}{\max\limits_{1\le j\le l_2}\|x_{2j}-\bar{x}_2\|-\min\limits_{1\le j\le l_2}\|x_{2j}-\bar{x}_2\|}$$

这里，假设训练数据集包含正负两类样本，$\|\cdot\|$ 表示欧式距离，x_{1i} 和 x_{2i} 分别表示正负类的训练样本，l_1 和 l_2 分别为正负样本个数，$\bar{x}_1$ 和 $\bar{x}_2$ 分别为正负类样本的类中心向量，计算公式如下：

$$\bar{x}_1=\frac{1}{l_1}\sum_{i=1}^{l_1}x_{1i}\qquad\bar{x}_2=\frac{1}{l_2}\sum_{i=1}^{l_2}x_{2i}\quad(2)$$

然而以上隶属度函数只考虑样本点离类中心的距离，并未考虑样本点的位置，即样本点是在 H_3 的哪一侧，如图 1 所示。由式(1) 可知，文献[9] 对样本点 A 与样本点 A' 赋予了相同的隶属度。可是，样本点 A 与样本点 A' 对支持向量的构成的影响明显不同，A 点更有可能成为支持向量，或是对支持向量的影响较大，应赋予较大的隶属度；而 A' 对支持向量的构成几乎没有影响，应赋予较小的隶属度。基于这一基本思想，文中对超平面 H_3、H_4 两侧的样本点赋予不同的隶属度，定义如下：

$$\begin{cases}\mu_{1i}=\dfrac{\|x_{1i}-\bar{x}_1\|-\min\limits_{1\le j\le l_1}\|x_{1j}-\bar{x}_1\|}{\max\limits_{1\le j\le l_1}\|x_{1j}-\bar{x}_1\|-\min\limits_{1\le j\le l_1}\|x_{1j}-\bar{x}_1\|},\\ d(x_1,H_2)\le R_1\\ \mu_{1i}=\delta,d(x_{1i},H_2)>R_1\end{cases}$$

$$\begin{cases}\mu_{2i}=\dfrac{\|x_{2i}-\bar{x}_2\|-\min\limits_{1\le j\le l_2}\|x_{2j}-\bar{x}_2\|}{\max\limits_{1\le j\le l_2}\|x_{2j}-\bar{x}_2\|-\min\limits_{1\le j\le l_2}\|x_{2j}-\bar{x}_2\|},\\ d(x_{2i},H_1)\le R_2\\ \mu_{2i}=\delta,d(x_{2i},H_1)>R_2\end{cases}\quad(3)$$

其中，$R_1=\dfrac{g(\bar{x}_1)+1}{\|\omega\|}$，$R_2=\dfrac{g(\bar{x}_2)+1}{\|\omega\|}$，分别为类中心 $\bar{x}_1$ 到 H_2 的距离和类中心 $\bar{x}_2$ 到 H_1 的距离，δ 为较小的正数。

现实中，同类训练数据多数服从正态分布，即同一类的大部分相对集中，少数特殊数据离其类中心较远。文中定义的隶属度函数不仅使得类中心附近的大部分数据不参与训练，还将离分类超平面较远且在错误的一方的边缘数据也排除在外，这使得参与训练的数据约是文献[9] 的一半，这就很大程度上减少了训练时间，能更快速地得到支持向量。

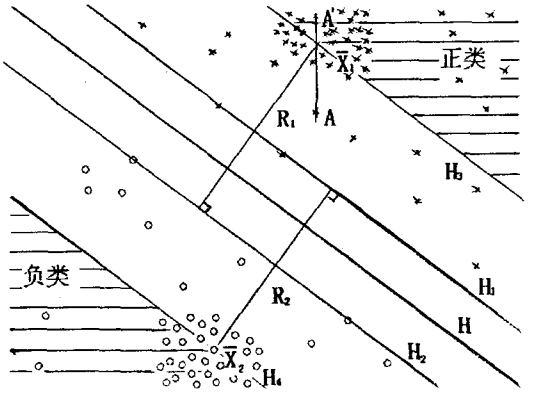

图 1 样本点的分布

1.2 二类快速模糊支持向量机算法

文中按下面两个步骤实现模糊支持向量机的这一思想。首先，根据训练样本在训练过程中的不同作用，构造训练样本的隶属度；其次，引入参数 λ 选取少训练样本集构造学习机。

其实现算法如下[9]：

算法一：

S1：训练传统 SVM 分类器，得到初始支持向量，并构造决策分类面；

S2：用式(3)求出每个样本的隶属度函数；

S3：模糊化训练集

$S_{1f}=\{(x_{1j},y_{1j},\mu_{1j})\mid y_{1j}=+1,\mu_{1j}\in[0,1],j=1,2,\cdots,l_1\}$

$S_{2f}=\{(x_{2j},y_{2j},\mu_{2j})\mid y_{2j}=-1,\mu_{2j}\in[0,1],j=1,2,\cdots,l_2\}$

S4：在模糊集 $S_f=S_{1f}\cup S_{2f}$ 上，利用参数 λ 选取少训练样本集

$S_{1f\lambda}=\{(x_{1j},y_{1j},\mu_{1j})\mid y_{1j}=+1,\mu_{1j}\ge\lambda,j=1,2,\cdots,l_1\}$

$S_{2f\lambda}=\lambda\{(x_{2j},y_{2j},\mu_{2j})\mid y_{2j}=-1,\mu_{2j}\ge\lambda,j=1,2,\cdots,l_2\}$

S5：在模糊子集 $S_{f\lambda}=S_{1f\lambda}\cup S_{2f\lambda}$ 上构造 FSVMs；

S6：检验 FSVMs 的性能，若令人满意，则得到学习机，否则，转 S4。

参数 λ 是用户自定义的参数，按照由大到小的顺序选择，如可依次选取 1,0.9,0.8,…,0.1,0 等。λ 越大，包含的训练样本点越少；λ 越小，包含的训练样本点越多。$\lambda=1$ 时，少训练样本集仅含有两个样本，包括一个正样本和一个负样本；$\lambda=0$ 时，新的学习机退化为传统的模糊支持向量机。只有当 $\lambda\le\delta$ 时，阴影部分的样本点才参与训练。

1.3 基于 DAG 的 k 类 FFSVMs

对于多类问题，目前应用较多的有“一对多(1－a

－r)"和"一对一(1－a－1)"。然而这两种方法都存在一定的缺点,如样本的重复训练率高、需要计算每一个支持向量机分类器的决策函数值等,导致训练速度慢[1]。文中采用 DAG(directed acyclic graph,导向非循环图)多类分类方法构造多类模糊支持向量机。对于 k 类问题,需要构造 $\frac{k(k-1)}{2}$ 个两类分类器,但是,DAG 方法分类时采用了有向无环图的组合策略,分类时不必遍历所有的分类器,这样就大大提高了分类效率。DAG 方法在分类类别数较少时优越性更加明显。

DAG 的基本思想是从任意一子分类器开始,将待分样本划分为两个子类,然后再对两个子类进一步划分,如此循环,直到子类中只包含一个类别为止,如图 2 所示。

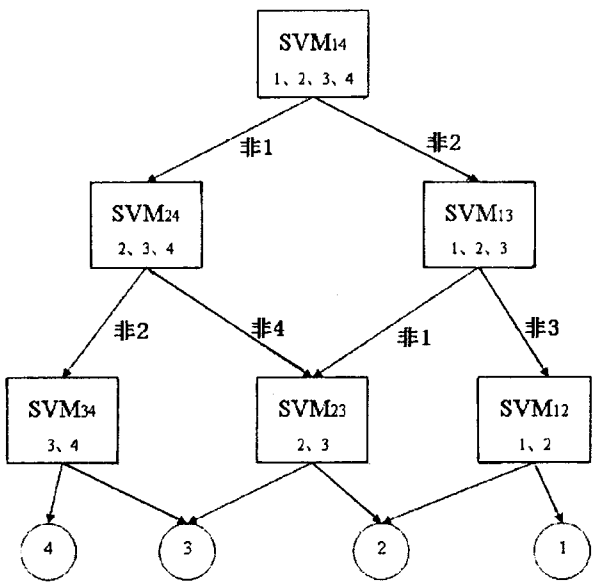

图 2　DAG 流程图

2　实验结果及分析

为了检验文中提出的改进的 FFSVMs 的性能,用文献[9]中的实验一做比较,利用并修改了标准的 LibSVM 软件包,以适应文中的方法。为了具有可比较性,所有学习机的运行环境与文献[9]相同(Windows2000,奔腾 IV 处理器,2.4GHz,256MB 内存)。表 1 列出了性能的比较,其中 FFSVM1 是文献[9]的支持向量机,FFSVM2 是文中提出的支持向量机,Ntr 表示训练样本的个数,Tr(%)表示训练数度,Ts(%)表示测试精度,PS(s)表示核函数的计算时间,QP(s)表示二次优化的时间。

从表 1 可以看出,和文献[9]提出的 FFSVMs 相比,改进的 FFSVMs 具有相同或更好的推广能力,即测试精度,且运行时间减少了很多。

3　结束语

对于 FSVMs,由于要计算每个训练样本的隶属度,增加了计算时间。文献[9]提出了基于边缘数据的少训练样本集上的 FFSVMs,提高了分类速度。文中对文献[9]做进一步改进,对离分类超平面较远、不可能成为支持向量的数据赋予较小的隶属度,使训练样本集中的数据大大减少。

实验结果表明,这类学习机在保证分类精度的前提下,提高了分类时间。

表 1　使用 dot 核时 FFSVM1 与 FFSVM2 的比较 (C=5000000)

Classifier	λ	Ntr	Tr(%)	Ts(%)	PS(s)	QP(s)
FFSVM1	0.5	879	96.966	93.211	0.36	2.389
FFSVM2		445	96.965	93.210	0.29	2.117
FFSVM1	0.4	1477	99.581	96.105	0.935	4.064
FFSVM2		730	99.581	96.105	0.712	3.513
FFSVM1	0.3	2277	99.869	96.327	2.018	8.481
FFSVM2		1053	99.869	96.327	1.156	6.913
FFSVM1	0.2	3127	100	96.439	3.8	13.747
FFSVM2		1524	100	96.439	2.969	10.025
FFSVM1	0.1	3678	100	96.661	5.767	18.703
FFSVM2		1798	100	96.661	4.567	14.563

参考文献:

[1] 张　永.基于模糊支持向量机的多类分类算法研究[D].大连:大连理工大学,2008.

[2] Shawe－Taylor J. Classification Accuracy Based on Observed Margin[J]. Algorithmica,1998,22(1):157－172.

[3] Shawe－Taylor J, Cristianini N. Further Results on the Margin Distribution[C]//Proceedings of the Conference on Computational Learning Theory. New York, USA: ACM Press, 1999:278－285.

[4] 刘太安,梁永全,薛　欣.一种新的模糊支持向量机多分类算法[J].计算机应用研究, 2008,25(7):2041－2042.

[5] LIN Chun－fu, WANG Sheng－de. Fuzzy support vector machines[J]. IEEE Trans on Neural Networks,2002,13(2):464－471.

[6] Inoue T, Abe S. Fuzzy support vector machines for patter classification[C]//Proc of International Joint Conference on Neural Networks. Washington DC:[s. n.],2001:1449－1455.

[7] Tsuj I N Ish I D, Abe S. Fuzzy least squares support vector machines for multiclass problems[J]. Neural Networks,2003, 16:758－792.

[8] 祈　立,刘玉树.基于二阶段聚类的模糊支持向量机[J].计算机工程,2008,34(1):4－6.

[9] 刘宏冰,熊盛武.一类快速模糊支持向量机[J].系统仿真学报,2008,20(24):6664－6667.