文章编号:1001 - 9081(2008)07 - 1642 - 03

基于增量学习的直推式支持向量机算法

肖建鹏,张来顺,任 星
(信息工程大学 电子技术学院, 郑州 450004)
(betret@sohu.com)

摘 要:针对直推式支持向量机在进行大数据量分类时出现精度低、学习速度慢和回溯式学习多的问题,提出了一种基于增量学习的直推式支持向量机分类算法,将增量学习引入直推式支持向量机,使其在训练过程中仅保留有用样本而抛弃无用样本,从而减少学习时间,提高分类速度。实验结果表明,该算法具有较快的分类速度和较高的分类精度。

关键词:支持向量机;直推式学习;增量学习
中图分类号: TP181 **文献标志码**:A

Transductive support vector machines based on incremental learning

XIAO Jian-peng, ZHANG Lai-shun, REN Xing
(*Institute of Electronic Technology, Information Engineering University, Zhengzhou Henan* 450004, *China*)

Abstract: Aiming at the problem of lower precision, slower training speed and more back learning steps when transducitve support vector machine learning algorithm carry on a great deal of data classification, a new transductive support vector machine based on incremental learning was proposed. The incremental learning was introduced into transductive support vector machine; thereby, the algorithm only employed the useful samples and discarded unwanted samples in the training process, reduced the study time and improved the classification speed of the algorithm. The experimental results show that the algorithm is of faster classification speed and higher classification accuracy.

Key words: support vector machine; transductive inference; incremental learning

0 引言

基于结构风险最小化原则的支持向量机(Support Vector Machine,SVM),由于能够较好地解决小样本、高维数、非线性等问题,因此得到广泛的应用并成为机器学习领域的研究热点[1-2]。

虽然 SVM 有比较坚实的理论基础和严格的理论分析,但是其从理论到应用还有很多尚未得到充分研究的问题[3]。比如,在对未知样本进行预测时,可直接采用从已知样本出发对特定未知样本进行识别,而放弃对所有可能的样本进行识别的方法,与传统的方法相比,这种学习方法被称为直推式学习(Transductive Inference)[4]。基于直推式学习的支持向量机叫作直推式支持向量机(Transductive Support Vector Machine,TSVM),它是一种通过少数已标记(或标注)样本和较多未标记样本共同训练的学习方法,虽然一些模型被先后提出[5-7],但这些方法存在训练速度慢、回溯式学习步骤多等缺点。

本文在针对现有 TSVM 算法不足的基础上,提出一种基于增量学习的直推式支持向量机分类算法,在克服以往算法中对未标记样本的正负比例容易错估的同时,将增量学习的思想引入其中,每次抛弃非支持向量,只将支持向量同新增未标注样本进行混合构成新的训练集进行训练,进而提高算法的速度。

1 支持向量机

SVM 理论是在统计学领域中结构风险最小化原则基础上发展起来的机器学习方法。SVM 的思想可以描述为:

考虑使用某个特征空间的超平面对给定训练数据集作二值分类的问题,对于给定样本点:

$$(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_i,y_i);\boldsymbol{x}_i\in\mathbf{R}^n,y_i\in\{-1,+1\} \tag{1}$$

其中向量 $\boldsymbol{x}_i$ 可能是从目标样本集中抽取的某些特征而直接构造的向量,也可能是原始向量通过某个核函数映射到核空间中的映射向量。

在特征空间中求一个分类超平面 $(w\cdot\boldsymbol{x})+b=0$,关键是求其系数 w 和 b。由于 SVM 理论要求分离超平面具有良好的分类特性,即必须满足最优分类超平面的条件:

$$\begin{cases} y_i[(w\cdot\boldsymbol{x}_i)+b]\geqslant 1;\ i=1,2,\cdots,m \\ \min\limits_{w}\varphi(w)=\|w\|^2 \end{cases} \tag{2}$$

为了找到最优分类超平面,根据最优理论和借助 Lagrange 函数将原问题转化成为求解标注型二次规划问题:

$$\max W(a)=\sum_{i=1}^{m}a_i-\frac{1}{2}\sum_{i,j=1}^{m}a_ia_jy_iy_jK(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{3}$$

$$\text{s.t.}\quad \sum_{i=1}^{m}a_iy_i=0$$

$$a_i\geqslant 0$$

$$i=1,2,\cdots,m$$

求最优超平面的关键在于求出 $a_i\geqslant 0$ 的 a_i 以及

收稿日期:2008 - 01 - 02;**修回日期**:2008 - 03 - 21。

作者简介:肖建鹏(1979 -),男,辽宁锦州人,硕士研究生,主要研究方向:机器学习、信息抽取; 张来顺(1963 -),男,河北安国人,教授,主要研究方向:网络安全、软件工程; 任星(1982 -),女,重庆人,硕士研究生,主要研究方向:数据库安全。

$$b_0 = \frac{1}{2}[K(w_0, x(1)) + K(w_0, x(-1))] \tag{4}$$

最优超平面为

$$f(x) = \text{sign}\{\sum_{\alpha>0} a_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) - b_0\} \tag{5}$$

通常 $a_i \geqslant 0$ 对应的样本点为支持向量。

2　改进的直推式支持向量机算法

2.1　传统直推式支持向量机

直推式支持向量机是传统 SVM 在未知样本上的一种扩展。传统 SVM 试图通过已标记样本对未标记样本进行预测,训练用于识别未标记样本的分类器;而 TSVM 则通过使用少量已标记样本和较多未标记样本共同进行分类学习,找到的最优分类超平面能够满足对原始未知样本的分类具有最小的泛化误差,比单纯使用有标签训练得到的分类器在性能上有明显提高。TSVM 算法实现原理如下[8]:

给定一组独立同分布的已标记训练样本 $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_i, y_i)$, $\boldsymbol{x}_i \in \mathbf{R}^n$, $y_i \in \{-1, +1\}$ 和另一组有相同分布的未标记样本 $\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \cdots, \boldsymbol{x}_k^*$。

根据 SVM 理论,分类超平面满足式(2)。为了保证输入向量在线性不可分的情况下,允许错分样本的存在,引入了松弛变量 ξ_1。然后利用 Lagrange 优化方法,引入 Lagrange 乘子 $\alpha_i, i = 1,2,\cdots,l$ 后,问题式(1)就转变为

$$\text{Minimiz over } (y_1^*, \cdots, y_k^*, w, b, \xi_1, \cdots, \xi_n, \xi_1^*, \cdots, \xi_k^*)$$

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + C^*\sum_{j=1}^{k}\xi_j^* \tag{6}$$

$$\text{subject to: } \forall_{i=1}^{n}: y_i[w \cdot \boldsymbol{x}_i + b] \geqslant 1 - \xi_i$$

$$\forall_{j=1}^{k}: y_j[w \cdot \boldsymbol{x}_j^* + b] \geqslant 1 - \xi_j^*$$

$$\forall_{i=1}^{n}: \xi_i \geqslant 0$$

$$\forall_{j=1}^{k}: \xi_j^* \geqslant 0$$

可以通过控制惩罚因子 C 和 C^* 的大小,来调节误分类样本对分类超平面的影响。测试集样本就是通过包含惩罚因子 C^* 的项来影响分类超平面的。

2.2　渐进式直推式支持向量机

虽然 TSVM 成功地将未标记样本中所隐含的分布信息引入支持向量机的学习过程中,比传统的 SVM 在性能上有显著的提高,但是 TSVM 仍然存在不足,主要表现在每次算法执行之前必须人为指定待训练的无标签样本中的正标签样本数 N,而在一般情况下 N 值是很难做出比较准确的估计的,尤其在已标记样本较少的情况下很容易导致较大的误差。

文献[3]$^{451-460}$和[9]对此进行改进,提出一种渐进直推式支持向量机学习算法(Progressive Transductive Support Vector Machine, PTSVM)。在该算法中,没有事先设定无标签样本中的正标签样本,而是在训练过程中根据一定的原则渐进地对无标签样本赋予标签并动态地给予调整,采用成对标注法则,每次标注两个样本,从而更好地描述样本的分布特征。具体算法如下:

1)指定参数 C 和 C^*,使用归纳式学习对有标签样本进行一次初始学习,得到一个初始分类器。

2)用初始分类器对无标签样本进行学习,计算每一个无标签样本的判别函数输出,用成对标注的法则在当前边界区域内的无标签样本标注一个新的正标签和一个新的负标签。

3)对所有样本重新训练,计算每一个无标签样本的判别函数输出。如果发现某一个早期标注的无标签样本的标签值和当前判别函数的输出值不一致,则按照标签重置的法则取消对该样本的标注。

4)用成对标注法寻找当前边界区域内符合新加标注条件的未标注的无标签样本。如果存在这样的无标签样本,则对其加以标注并返回第 3 步;如果不存在这样的无标签样本,则用当前的分割平面对剩下的全部无标签样本做分类并加标签。算法结束,并输出结果。

PTSVM 在使用成对标注法的过程中,可能在某一次训练后发现一个或多个已标注的无标签样本和用当前分类器对标签分类所得到的值不一致。在这种情况下,就把不一致的样本重新置为无标签样本并继续进行迭代,这种方法叫作标签重置法。

2.3　改进直推式支持向量机

尽管 PTSVM 采用成对标注法则和标签重置法动态地对无标签样本给予调整,解决了在算法结束的时候应该将哪类样本加入到未标记样本库的问题。但 PTSVM 算法并没有考虑到在训练过程中,当未标注样本较多时,频繁地成对标注和重复训练会加重学习机负担,最终导致训练速度明显下降的问题。

通过对 SVM 原理分析可以发现,SVM 最终可转化为解二次规划问题,即对所有样本进行优化,但并不是所有的样本在学习中都起同样重要的作用,只有满足 SV 的样本才对最优超平面和决策函数有贡献。一般来说,将与拉格朗日乘子 a_i 的取值为 $0 < a_i < C$ 对应的 $\boldsymbol{x}_i$ 称为普通支持向量,而将与 $a_i = C$ 对应的 $\boldsymbol{x}_i$ 称为边界支持向量。前者代表了所有不能被正确分类的样本;而后者则代表了大部分样本的分类特征,它们共同决定了最终分类器的形式。也就是说,支持向量集能够充分描述整个训练样本集的特征,对它的划分等价于对整个样本集的划分。

然而,支持向量仅仅占整个训练样本集的很少一部分,如果在学习中去花费大量时间优化非支持向量,就会影响学习的效率。如果在每次训练过程中,仅保留样本集中的支持向量而抛弃非支持向量,进而取代训练样本集并同新增样本混合进行训练,在不影响分类精度的同时能够降低训练时间,提高分类器的性能,使分类学习具有增量学习能力,这就是增量式 SVM 的分类思想。支持向量机增量式学习问题可以描述为:

前提条件:设历史数据集 T^1,增量数据集 T_{inc}^1,并且假定对于两个数据集 $T^1 \cap T_{inc}^1 = \varnothing$, Γ^1 和 $\boldsymbol{A}_{sv}^1$ 分别为数据集 A 上的初始 SVM 分类器以及对应的 SV 集。

问题的目标:在初始分类器 Γ^1 和对应的 SV 集 $\boldsymbol{A}_{sv}^1$ 的基础上,寻找基于样本集合 $A \cup B$ 上的 SVM 分类器 Γ 和对应的 SV 集 A_{sv}。

具体算法描述如下[10]:

第 1 步　标记初始训练样本集 T^1。

第 2 步　根据训练样本集 T^1 训练 SVM,得到分类器 Γ^1,取得支持向量 $\boldsymbol{A}_{sv}^1$。

第 3 步　如果满足训练精度或增量样本集 T_{inc}^1 为空,则跳至第 4 步,否则从 T_{inc} 中取出 k 个样本,组成的集合记为 X, $T_{inc} = T_{inc}^1 - X$, $A_{sv} = \boldsymbol{A}_{sv}^1 \cup X$, 跳至第 2 步。

第 4 步　输出最终分类器 Γ。

从以上可以看出,增量式 SVM 算法与传统的 SVM 之间

最大的区别是:增量式 SVM 的学习样本仅由已学习样本中的支持向量和新增学习样本组成;而传统的 SVM 则由所有的已学习样本和新增的学习样本组成。增量式 SVM 是在不损失分类精度的前提下丢弃一些样本点,以较小的时间空间代价实现新样本的学习。由此,本文将增量式学习思想引入 TSVM 中,并结合 PTSVM 算法中成对标注法则与标签重置法则的优化方法,提出一种基于增量学习的直推式支持向量机算法(Incremental Progressive Transductive Support Vector Machine,IPTSVM)。与 TSVM 相比,IPTSVM 在每次训练过程中,使用成对标注法则从当前边界区域内获得一对样本,只与前一次使用分类器所获得的支持向量相结合,而不是同全部样本相结合去获得新的训练集。这样在每次训练过程中抛弃大量对分类超平面影响不大的样本,只保留一小部分对分类超平面起决定作用的支持向量,可以极大地缩短训练时间、减小迭代次数,进而提高训练速度。具体的算法描述如下:

前提:设历史数据集 T^0,未标记样本集 T_{ul}^0,Γ^0 和 A_{sv}^0 分别为数据集 A 上的初始 SVM 分类器以及对应的 SV 集。

第 1 步 获得初始样本集 T^0,使用归纳式学习进行训练,得到初始的分类器 Γ^0,取得支持向量 A_{sv}^0。

第 2 步 判断是否有未标记样本,如果有则执行第 3 步;否则算法结束,输出当前分类器。

第 3 步 使用分类器 Γ^0 对当前无标签样本进行学习,用成对标注的法则在当前边界区域内的无标签样本中标注一个新的正样本 A^+ 和一个新的负样本 A^-。

第 4 步 执行 $T_{ul}^0 = T_{ul}^0 - A^+ - A^-$,将两个使用成对标注法则标注的标签从未标记样本集 T_{ul}^0 中删除,获得新的未标记样本集 T_{ul}^0。

第 5 步 将新标记的两个样本加入到原来训练样本的支持向量集 A_{sv}^0 中,获得新的训练样本 $A_{sv}^0 = A_{sv}^0 + A^+ + A^-$。

第 6 步 重新训练,等到新的分类器 Γ^1,获取支持向量 A_{sv}^1。

第 7 步 使用 Γ^1 重新计算所有无标记样本的类别。如果发现早期标记的样本的类值和当前判别函数输出值不一致,则用标记重置法取消对该样本的标注,记这些样本的集合为 N。

第 8 步 重新将 N 放回到未标记样,再次获得新的未标记样本 $T_{ul}^0 = T_{ul}^0 + N$。

第 9 步 用成对标记法在当前边界区域内寻找符合条件的未标记样本,如果不存在这样的未标记样本,则用当前分类器对无标记样本进行分类,输出分类器 Γ,清空未标记样本集,跳到第 2 步;否则跳到第 4 步。

算法中,在第 5 步和第 6 步进行重新训练 SVM 时,训练是在支持向量和新增样本上进行的,体现出增量学习的优点。第 3、7、9 步的过程中采用的成对标记法则和标记重置法则,其中标记重置法则保证在每次训练过程过后是 IPTSVM 算法具有一定的修复能力,进而使 SVM 分类器的性能可以不断地得到提高。

3 实验结果与分析

3.1 评估标准和实验条件

为了验证本文提出的 IPTSVM 算法的有效性,从以下两个方面对其性能进行评价:

1)学习精度。算法能对输入的数据进行正确、精确的学习和分类。

2)学习速度。算法能在有限的时间内完成学习任务。

使用 P4 2.0 GHz、512 MB 内存、系统实验环境为 Windows XP 和 Matlab 7.0 的 PC 机进行实验,采用径向内积函数 $K(\boldsymbol{x}, \boldsymbol{x}_i) = \exp(-|\boldsymbol{x} - \boldsymbol{x}_i|^2/2s^2)$,其中参数 s 取 0.5。实验数据使用来自于 http://www.ics.uci.edu/~mlearn/MLRepository.html 提供的二类别标准样本数据集 Promoter-Gene,其中所有数据都处理为均值为 0,标准差 1。所有实验的训练集均采用一组有标签样本,即事先手工分类 10 个正标签样本和 10 个负标签样本。训练集中的无标签样本则按照不同的数量和不同的比例来选取,以测试算法在各种可能的无标签样本数据分布下的性能。所有实验共用含有 200 个正标签样本和 200 个负标签样本的测试集。

3.2 实验结果和分析

本文使用 TSVM 和 IPTSVM 分别对数据集进行学习,每次采用不同的初始训练样本集得出结果,如表 1 所示。

表 1 IPTSVM 和 TSVM 算法性能比较

学习算法	无标签样本数	训练时间/s	支持向量数	P/%	R/%	测试精度/%
TSVM	Pos=20	0.51	39	71.19	98.01	79.19
IPTSVM	Neg=20	0.59	39	77.44	98.24	84.79
TSVM	Pos=40	0.81	121	90.65	78.33	84.78
IPTSVM	Neg=60	1.92	113	89.46	84.00	86.54
TSVM	Pos=80	2.12	154	73.32	99.67	81.83
IPTSVM	Neg=120	3.71	147	77.67	99.33	84.50
TSVM	Pos=160	5.51	238	95.58	98.33	96.83
IPTSVM	Neg=160	8.41	249	90.49	98.00	94.00

从表 1 可以看出 IPTSVM 算法在整体性能上优于 TSVM 算法,这主要是因为 IPTSVM 算法引入增量学习的思想,在每次训练过程中只使用对分类函数有用的支持向量,而抛弃无用的非支持向量,缩短训练时间进而提高分类效率。同时,算法结合成对标注法则与标签重置法则,对于训练集中无标签样本所赋标签的正负比例总是比 TSVM 更接近实际正负标签的比例,使得 IPTSVM 算法比 TSVM 算法有更好的性能。

4 结语

本文在简单介绍支持向量机理论和直推式学习的基础上,提出一种基于增量学习的直推式支持向量机算法 IPTSVM。该算法在将增量学习的思想引入 TSVM 的同时结合成对标注法则与标签重置法则,不但继承渐进赋值和动态

(下转第 1648 页)

为“-s 0 -c 1000 -t 2 -wi 1”,相关参数的意义和取值参见 LIBSVM[6] 对应的网站,其他参数取默认值。

表1 语种语音样本的分布情况

样本	时间/s	样本数量			
		中文	英文	日文	总计
片长1	1.008	2323	4254	1007	7584
片长2	1.504	1325	2449	623	4397
片长3	2.000	827	1611	427	2865
片长4	2.496	567	1119	315	2001

表2 语种语音样本的识别比较

维数	PONN			
	中文	英文	日文	平均
992	91.031	84.177	95.327	87.562
1488	91.928	84.177	88.785	86.940
1984	89.238	76.160	80.374	80.348
2480	77.130	86.709	83.178	83.831
维数	SVM			
	中文	英文	日文	平均
992	89.256	84.758	93.443	87.140
1488	92.562	82.156	95.082	86.696
1984	90.083	75.836	86.885	81.153
2480	81.818	90.706	80.328	87.140

通过比较 PONN 和 LIBSVM 的语种语音识别结果发现,CASA 算法在一定情况下(如表2中带下划线数据所示)要比 SVM 好,具有一定竞争力,但不是绝对。但由于 SVM 方法转化为最优化问题的求解,在高维数据情况下难以计算,而 PONN 方法利用同类、异类样本点集关系,更具可计算性,具体时间复杂度的分析有待进一步研究。

4 结语

本文研究了 PONN 网络的学习算法框架,提出一般学习过程描述,对特定算法 RPA、CASA 进行了描述、分析,通过实验可以看出:

1)结合事物高维空间分布特点,有利于提高网络构造性能和泛化能力。

2)简单的 CASA 测试结果在一定情况下比 SVM 好,具有很强的竞争力。

PONN 网络通过动态构造方法进行学习,把网络结构与神经元分开,从而克服传统网络神经元数量难以确定、收敛速度慢等问题。如何更好地分析高维空间点之间的关系,构造合适的网络结构,是进一步要解决的问题。

参考文献:

[1] HAYKIN S. 神经网络原理[M]. 北京: 机械工业出版社, 2004.

[2] BAKRY H M, MASTORAKIS N. New fast normalized neural networks for pattern detection[J]. Image and Vision Computing, 2007, 25(11): 1767-1784.

[3] ISLAM M M, YAO X, MURASE K. A constructive algorithm for training cooperative neural network ensembles[J]. IEEE Transactions on Neural Network, 2003, 14(4): 820-834.

[4] LIU T, Moore A W, Gray A. New algorithms for efficient high-dimensional nonparametric classification [J]. Journal of Machine Learning Research, 2006, 7(6): 1135-1158.

[5] PAREKH R, YANG J H, HONAVAR V. Constructive neural network learning algorithm for pattern classification[J]. IEEE Transactions on Neural Network, 2000, 11(2): 436-451.

[6] CHANG C C, LIN C J. LIBSVM-A library for support vector machines [EB/OL]. [2007-10-10]. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[7] IKEDA K, YAMASAKI T. Incremental support vector machines and their geometrical analyses[J]. Neurocomputing, 2007, 70(13/15): 2528-2533.

[8] WANG S J. Blomimetic pattern recognition(in Chinese)[J]. Acta Electronica Sinica, 2002, 30(10): 1417-1420.

[9] WANG SHOU-JUE, LAI JIANG-LIANG. Geometrical learning, descriptive geometry, and biomimetic pattern recognition[J]. Neurocomputing, 2005, 67: 9-28.

[10] KIM S, ERIKSSON T, KANG H G, *et al*. A pitch synchronous feature extraction method for speaker recognition[C]// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Montreal, Canada, 2004, 1: 405-408.

[11] INGER I M, THORPE J A. 拓扑学与几何学基础讲义[M]. 干丹岩, 译. 上海: 上海科学技术出版社, 1985: 1-3.

(上接第1644页)

调整的规则,而且使用增量学习方法,保留支持向量、抛弃非支持向量,有效地克服了传统 TSVM 算法中训练速度慢,训练准确率低等问题,具有现实的推广意义。实验结果表明,IPTSVM 算法在各种样本分布情况下都取得了较好的分类效果。

参考文献:

[1] BURGES C J C. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121-167.

[2] VAPNIK V N. The nature of statistical learning theory[M]. New York: Springer-Verlag, 2000.

[3] 陈毅松, 汪国平, 董士海. 基于支持向量机的渐进直推式分类学习算法[J]. 软件学报, 2003, 14(3): 451-460.

[4] VAPNIK V N. Statistical learning theory[M]. New York: John Wiley and Sons, 1998.

[5] WANG Y E, HUANG SHANG-TENG. Training TSVM with the proper number of positive samples[J]. Pattern Recognition Letters, 2005, 26(14): 2187-2194.

[6] JOACHIMS T. Transductive learning via spectral graph partitioning [C]// Proceeding of The Twentieth International Conference on Machine Learning. Washington, DC: [s.n]. 2003: 290-297.

[7] ZELIKOVITZ S. Transductive LSI for short text classification problems[C]// Proceedings of 17th International Florida Artificial Intelligence Research Symposium Conference. Miami: AAAI Press, 2004.

[8] JOACHIMS T. Transductive inference for text classification using support vector machines[C]// Proceedings of the 16th International Conference on Machine Learning (ICML). San Francisco: Morgan Kaufmann Publishers, 1999: 200-209.

[9] CHEN YI-SONG, WANG GUO-PING, DONG SHI-HAI. Learning with progressive transductive support vector machine[C]// Second IEEE International Conference on Data Mining(ICDM'). Maebashi Japan: IEEE Press, 2002: 67-74.

[10] RATSABY J. Incremental learning with sample queries[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 883-888.