第32卷 第21期 计 算 机 工 程 2006年11月
Vol.32 No.21 Computer Engineering November 2006
·人工智能及识别技术· 文章编号：1000—3428(2006)21—0177—03 文献标识码：A 中图分类号：TP18

基于支持向量的分层并行筛选训练样本方法

文益民
(湖南工业职业技术学院信息工程系，长沙 410007)

摘 要: 基于支持向量能够代表训练集分类特征的特点，该文提出了一种基于支持向量的分层并行筛选训练样本的机器学习方法。该方法按照分而治之的思想将原分类问题分解成若干子问题，将训练样本的筛选过程分解成级联的2个层次。每层采用并行方法提取各训练集中的支持向量，这些被提取的支持向量将作为下一层的训练样本，各层训练集中的非支持向量通过学习被逐步筛选掉。为了保证问题的一致性，引入了交叉合并规则，仿真实验结果表明该方法在保证分类器推广能力的情况下，缩短了支持向量机的训练时间，减少了支持向量的数目。

关键词: 分而治之；训练样本筛选；支持向量机；交叉合并规则

Method for Flatted and Parallel Training Samples Selection Based on Support Vectors

WEN Yimin
(Department of Information Engineering, Hunan Industry Polytechnic, Changsha 410007)

【Abstract】 In order to handle large-scale classification problems, this paper presents a machine learning method for hierarchically and parallel training samples selection, based on the characteristics of support vectors that represent the classification information of training data. In this method, according to the principle of divide and conquer, the original classification problem is divided into several small sub-problems. After that, the training procedure is separated into two cascade phases. In each phase, all of the sub-problems are processed, their support vectors are extracted and so the non-support-vectors are filtered out step by step. In order to keep the consistency, a cross-merger principle is introduced. The simulation results indicate that the method speeds up training while maintaining the generalization accuracy of support vector machines(SVMs), and reduces the number of support vectors (SVs).

【Key words】 Divide and conquer; Training sample selection; Support vector machines; Cross-merger rules

1 概述

支持向量机[1]是一种重要的模式分类方法，已被广泛地应用于文本分类、人脸检测与识别、语音识别、生物信息学等领域。基于结构风险最小化原则，支持向量机方法通常将训练样本映射到高维空间，然后在其中构造最优分类超平面，从而使分类器具有较强的分类能力和良好的推广能力。支持向量机训练过程的本质是求解一个二次凸优化问题[1]，其时间复杂度为$O(N^3)$，N为训练样本数。经过Joachims等人的工作[2]，目前常用的支持向量机训练算法的时间复杂度为$O(N^2)$，当问题规模很大时，支持向量机的训练时间将会很长。

使用支持向量机解决大规模模式分类问题，通常有2种方法：串行学习方法[2,3]和并行学习方法[4,5]。串行学习方法按照分而治之的原则将一个问题分成若干子问题，然后将各个子问题串行处理。工作集方法就是典型的串行学习方法。工作集方法通常包括Chunking算法、分解算法[2]和SMO算法。工作集方法每次只针对一部分变量进行优化而将其他变量视为常量。然后根据一些启发式规则，选择下一个工作集进行优化，如此迭代直至求得最优解。虽然这种方法已广为使用，但是在处理大规模模式分类问题时，由于某些训练样本反复进入工作集，导致了迭代次数过多和训练时间过长等问题。并行学习方法按照分而治之的原则将原问题分解成若干子问题，将各个子问题并行处理以后再进行集成[4,5]。并行学习方法的优点是能缩短训练时间，具有良好的可修改性和可扩充性，其缺点是支持向量有所增加。

根据Syed[3]和Scholkopf的工作，支持向量集包含了训练样本集的全部分类信息。为了解决大规模模式分类问题，根据支持向量的这个特点，本文在前期工作[6]的基础上提出了一种新的分层并行筛选训练样本方法，以缩短训练时间。首先将一个包含两类样本的训练集按照样本类别分别分解成K个大致相等的子集，然后将这些子集组合成K^2个子问题。通过并行训练提取各自的支持向量，然后按照交叉合并规则将其中每K个支持向量集合并成一个训练集。同样通过并行训练提取各自的支持向量，将来自K个训练集的支持向量再进行合并以作为最终分类器的训练集，该方法易于实现。

2 基于支持向量的分层并行筛选训练样本方法

假设两类分类问题中正类的样本集为

$$X^+=\{(X_i,+1)\}_{i=1}^{N^+}$$

反类的样本集为

$$X^-=\{(X_i,-1)\}_{i=1}^{N^-}$$

基金项目：湖南省青年骨干教师基金资助项目(湘教通[2001]204号)
作者简介：文益民(1969－)，男，博士生，主研方向：统计学习理论，生物信息学，图像处理
收稿日期：2006-01-19 **E-mail：**ymwen2004@yahoo.com.cn

其中，x_i 表示第 i 个训练样本，N^+ 和 N^- 分别表示正类和反类样本的数目，则训练样本集可表示为 $S = X^+ \cup X^-$，样本总数为 $N = N^+ + N^-$。

经过 2 层筛选，非支持向量被逐步过滤掉。S_{final} 含有比原训练集 S 少得多的样本。由于在分层筛选过程中采用了交叉合并方法，S_{final} 中包含了原训练集 S 中全部的分类信息。以上分层筛选训练样本的过程构成一个 3 层的满 K 叉树，如图 1 所示。

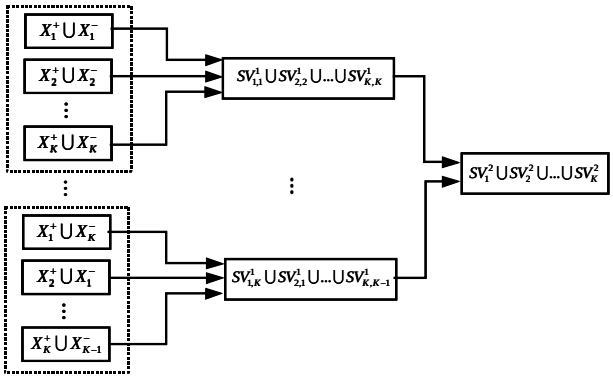

图 1 分层并行筛选训练样本方法

分层筛选训练样本的过程由以下两步构成：

(1)根据总训练样本数 N 和事先确定的分解常数 K，然后将原训练集 X^+ 和 X^- 分别分解为 K 个大致相等且互不相交的子集：

$$\begin{aligned} X^+ &= \bigcup_{i=1}^{K} X_i^+, X_i^+ = \{(X_j, +1)\}_{j=1}^{N_i^+}, i = 1,2,\ldots,K \\ X^- &= \bigcup_{i=1}^{K} X_i^-, X_i^- = \{(X_j, -1)\}_{j=1}^{N_i^-}, i = 1,2,\ldots,K \end{aligned} \tag{1}$$

其中，

$$N_i^+ = \left\lfloor N^+ / K \right\rfloor, i = 1,2,\ldots,K-1;\ N_K^+ = N^+ - \sum_{i=1}^{K-1} N_i^+;$$

$$N_i^- = \left\lfloor N^- / K \right\rfloor, i = 1,2,\ldots,K-1;\ N_K^- = N^- - \sum_{i=1}^{K-1} N_i^-$$

于是，原两类分类问题 S 被分解成下列 K^2 个规模较小的两类分类子问题：

$$S_{i,j}^1 = X_i^+ \cup X_j^-, 1 \le i, j \le K \tag{2}$$

由于这 K^2 个子问题在处理时不需要数据交换，可将得到的这 K^2 个子问题 $S_{i,j}^1\ (1 \le i, j \le K)$按照通常 SVMs 的训练方法并行训练，得到 K^2 个支持向量集合，表示为 $SV_{i,j}^1$，$1 \le i, j \le K$。

(2)将以上 K^2 个支持向量集按照下面的方法合并成 K 个集合：

$$S_i^2 = \bigcup_{j=1}^{K} SV_{j,permu(j+i-1)}^1, 1 \le i \le K \tag{3}$$

其中，$permu(m)$ 的定义为

$$permu(m) = \begin{cases} m - K & if \quad m > K \\ m & otherwise \end{cases} \tag{4}$$

这种将 $SV_{i,j}^1\ (1 \le i, j \le K)$合并的方法被定义为“交叉合并”。交叉合并的目的是为了保证原问题分解后包含的分类信息不被损失，因为在每个集合 $S_i^2\ (1 \le i \le K)$中尽可能多地包含有原整个训练集中的分类信息。同样按照通常 SVMs 的训练方法并行地对 $S_i^2\ (1 \le i \le K)$进行处理而得到 k 个支持向量集合。分别表示为 SV_i^2，$1 \le i \le K$。不同的 $SV_i^2\ (1 \le i \le K)$是对原全部训练样本中包含分类信息的一种近似。最后将它们按式(5)合并：

$$S_{final} = \bigcup_{i=1}^{K} SV_i^2 \tag{5}$$

其中，S_{final} 将作为最终分类器的训练集。

3 仿真实验

为了系统地研究本文所提方法的有效性，在仿真实验中采取了 3 种策略：(1)选择分类难度不等的数据集。本文从 UCI[7]中挑选了 5 个数据集。前 2 个数据集的分类难度较大，后 3 个数据集的分类难度较小。(2)选择训练集规模不等的数据集。本文中的训练集之间的规模变化较大(576~290 508)。(3)K值变化范围较大。前 4 个实验中K值取 1,2,…,30。K=1 时表示不对训练集进行分层筛选，也就是直接用整个训练集训练支持向量机。K＝2,3,…,30 表示将训练集中的每类数据分解成K(>1)个子集，以进行分层筛选。由于第 5 个实验的训练时间较长，只选择了K=1、K=2、K=10、K=30 4 种情形进行实验。通过以上措施，可以研究当K变化时，在各种分布、各种规模的训练集上分类器的准确率、训练时间和支持向量数目的变化规律。

在实验中，第 1、2、4 个数据集未作规范化处理，一律采用数据集中的原始数据，第 3 个和第 5 个数据集进行了规范化。实验中多类问题的解决采用一对一分解策略，也就是一个 M 类分类问题通过将其分解成 $M(M-1)/2$ 个两类分类问题来解决。测试时的最终结果由 $M(M-1)/2$ 个分类器作相对多数投票得到。本文采用SVM^{light}[2]作为训练支持向量机的工具，实验平台为：Pentium 4/3.0GHz/1000MB RAM/ PC机。各实验中采用的核函数均为径向基函数为

$$\exp(-\frac{1}{2\sigma^2}\|X - X_i\|^2)$$

实验中两类问题的时间统计为分层并行的时间和。多类问题的时间统计为各两类分类问题的时间和。多类分类问题中的支持向量数目包含各两类分类问题中重复的支持向量。为了方便比较，在图 2 中将 K=1 时的支持向量数目和分层并行训练时间都用 1 表示。K=2,3,…,30 时的支持向量数目和训练时间都用百分比表示。各个实验的数据分布、K=1 时的支持向量数目和支持向量机参数选择如表 1 所示。

表 1 实验数据分布、支持向量数目和支持向量机参数选择情况

数据集	分类问题	特征维数	类别数	训练集	测试集	K=1 时的支持向量数目	C	σ
Indians-diabetes	A_1	8	2	576	192	371	10	20
German credit	A_2	24	2	750	250	505	10	7
Letter recognition	A_3	16	2	15 000	5 000	2 562	16	$\sqrt{2}$
Space shuttle	A_4	9	5	43 483	14 494	544	1 000	50
Forest coverType	A_5	54	7	290 508	290 504	183 000	128	0.25

3.1 实验数据说明

在India diabetes 和German credit两个实验中，采用了 4-cross-validation策略。表 2 和图 2 中的相应数据是 4 次实验的平均值。Letter recognition 数据集中的前 13 类数据被当成正类数据，后 13 类数据作为反类数据。分别将其中 3/4 作为训练集，其中 1/4 作为测试集。针对上述 3 个数据集的实验

分别被标记为A_1、A_2和A_3。

Challenger Space Shuttle数据集有训练数据 43 500 条，测试数据 14 500 条。由于其中的第 6 类、第 7 类包含的数据很少，训练集分别为 6 条和 11 条，测试集分别为 4 条和 2 条，抽取第 1 类~第 5 类共 5 类数据构成一个分类问题A_4。Forest CoverType数据用来预测原始地区森林的植被覆盖情况。数据集包括 7 类数据共计 581 012 条。本文利用第 1 类~第 7 类所有数据构成一个 7 类分类问题A_5。在A_5中随机抽取各类的一半数据作为训练数据，另一半作为测试数据。

3.2 实验结果与讨论

从表 2、图 2 和表 3 可知，本文所提方法在 K 取不同值时，在各种不同分布、规模不等的数据集上，都能保持分类器的准确率。

表 2 支持向量机载测试集上的准确率的变化

数据集	分类问题	K=1	K=2,3,…,30	
			均值	方差
Indians-diabetes	A_1	0.720 05	0.720 05	0.000 90
German credit	A_2	0.718 00	0.717 83	0.000 53
Letter recognition	A_3	0.989 98	0.989 74	0.000 37
Space shutlle	A_4	0.998 97	0.998 94	0.000 36

表 3 Forest coverType 数据集上的实验结果

划分数 K	测试准确率	支持向量数目	训练时间(s)
K=1	0.948 91	183 000	88 584
K=2	0.948 92	175 314	9 243
K=10	0.948 92	171 661	5 462
K=30	0.948 97	171 883	4 839

由图 2 可知，本文所提方法能减少支持向量。这是因为对训练集的分解导致各个子问题的规模变小，各层得到的支持向量集总是与更简单的分类面相联系。

本文提出的方法产生了类似样本修剪[8]的效果，从而使支持向量数目减少。可以预测，对于同一训练集，随着K的进一步增大，支持向量数目会逐步增加。极端时，第 1 层可能失去数据筛选功能，导致 $S_i^2 = S\ (1 \le i \le K)$，使最终支持向量数目等于不作训练集分解时的支持向量数目，图 2(a)、图 2(b)中支持向量数目的变化体现了这种趋势，图 2(c)、图 2(d)中支持向量数目随着 K 值增加也呈现了上升趋势。

在表 3 中，K=30 时支持向量的数目也开始增加。在 India diabetes 和 German credit 2 个实验中，由于支持向量在整个训练集中所占的比例较大(60%以上)，分层并行筛选掉的非支持向量不多，因此训练时间会增加。在 Letter recognition、Challenger Space Shuttle 和 Forest CoverTyp 的实验中，由于支持向量在整个训练集中所占的比例较小，大量的非支持向量被筛选掉，不再重复进入训练过程，因此训练时间大幅度减少。在很多大规模模式分类问题中，支持向量在训练样本集中所占的比例通常比较小，本文提出的方法特别适应于这种大规模模式分类问题。

分层并行筛选训练样本方法在保证分类器推广能力的同时，能将大规模问题分解成若干规模较小的问题，在一定条件下不但能减少分类器的训练时间，而且能减少支持向量。在第(1)步和第(2)步的非支持向量筛选过程中，如果能够采取一些措施在不求出各子优化问题解的前提下找出支持向量的一个大致集合，则本文提出的方法会更快。

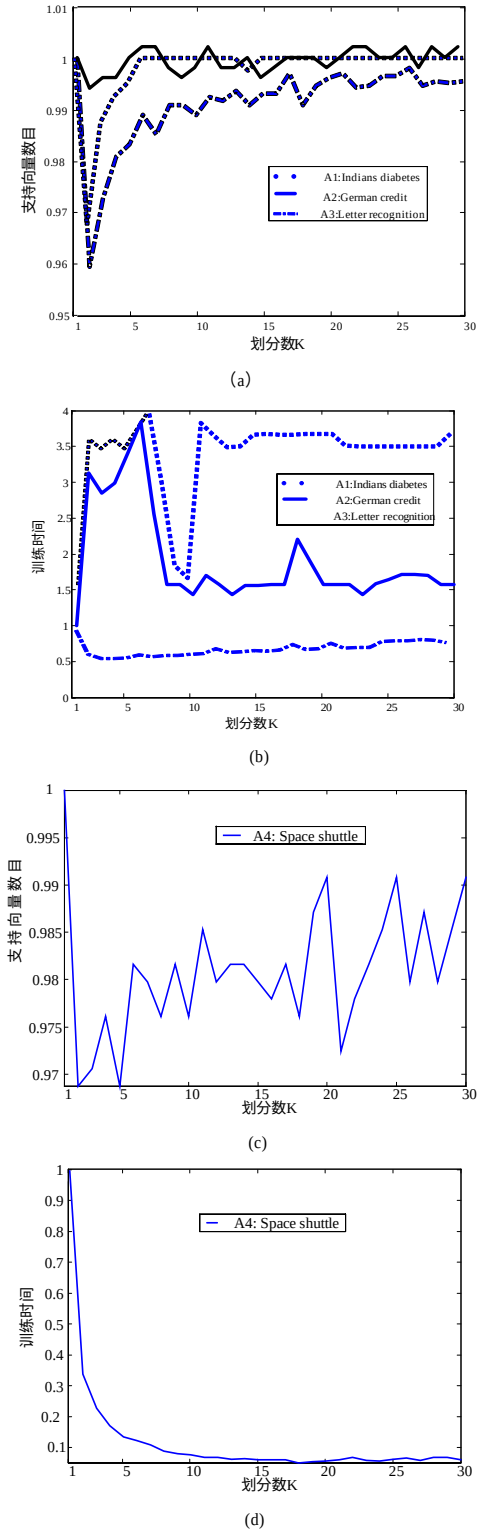

图 2 分层并行训练时间和支持向量数目的变化

4 结束语

该方法降低了问题的规模，为支持向量机的并行实现提供了一个可行的框架。多个不同的数值仿真试验表明，当支持向量所占比重较小时，本文所提方法与通常训练支持向量机的方法相比有 2 个优点：(1)在保证分类器推广能力的前提下，能提高支持向量机的训练速度；(2)减少了支持向量的数目，这一优点有利于提高支持向量机的响应速度，降低支持向量机在软件和硬件实现时的成本，本方法是解决大规模模式分类问题的一种简单有效的方法。 （下转第 182 页）