**Advances in Applied Mathematics** 应用数学进展, 2020, 9(2), 195-203
Published Online February 2020 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2020.92023

Hans汉斯

# Pre-Selection of Support Vectors Base on Distance Pairing Sorting

**Chengzhi Han*, Entao Zheng, Guochun Ma#**

College of Science, Hangzhou Normal University, Hangzhou Zhejiang
Email: 1224764141@qq.com, 851347680@qq.com, #maguochun@163.com

Received: Jan. 30th, 2020; accepted: Feb. 12th, 2020; published: Feb. 19th, 2020

## Abstract

**Support Vector Machine is a binary classification method based on statistical learning theory. The Sequential minimal optimization algorithm is an efficient algorithm developed for the dual problem of SVM. In the data training process, support vectors play a decisive role in the determination of separation hyperplane. However, support vectors are only a small part of the original sample set and distributed in the boundary of two types of data. If a boundary vector set containing most support vectors is used to replace the original sample set for training, the training time can be shortened and the classification speed can be improved on the premise of guaranteeing the classification accuracy. Pre-selection of support vector is difficult. In order to solve this problem, this paper proposes a support vector pre-selection algorithm based on distance pairing sort. The experimental results show that the proposed algorithm can effectively pre-select the set of boundary vectors containing support vectors.**

## Keywords

**Support Vector Machine, Support Vector, Pre-Selection, The Boundary Vector Set, Distance Pairing Sorting**

# 基于距离配对排序的支持向量预选取算法

**韩成志*，郑恩涛，马国春#**

杭州师范大学理学院，浙江 杭州
Email: 1224764141@qq.com, 851347680@qq.com, #maguochun@163.com

收稿日期：2020年1月30日；录用日期：2020年2月12日；发布日期：2020年2月19日

## 摘　要

**支持向量机是一种基于统计学习理论的二分类方法。支持向量机最小序列化算法(SMO)是针对支持向**

*第一作者，#通讯作者。

文章引用: 韩成志, 郑恩涛, 马国春. 基于距离配对排序的支持向量预选取算法[J]. 应用数学进展, 2020, 9(2): 195-203.
DOI: 10.12677/aam.2020.92023

**量机的对偶问题开发的高效算法。在数据训练过程中，支持向量对于分离超平面的确定起着决定性作用，但是支持向量仅占原始样本集的一小部分，并且分布在两类数据的边界上。如果用一个包含大多数支持向量的边界向量集来替换原始样本集进行训练，这样便能在保证分类精度的前提下，缩短训练时间，提高分类速度。然而支持向量的预选取比较困难，因此为了解决该问题，本文提出了一种基于距离配对排序的支持向量预选取算法。数值实验结果表明本文的算法能够有效地预选取包含支持向量的边界向量集。**

## 关键词

**支持向量机，支持向量，预选取，边界向量集，距离配对排序**



## 1. 引言

统计学习理论最早是在 1960 [1]年提出的。这是一种主要研究小样本统计学习规律的理论。在 20 世纪 90 年代中期，Vapnik [2]和他的团队基于这一理论提出了一种新的学习算法，称为支持向量机。作为研究小样本的一种学习方法，支持向量机具有很强的适用性，在模式识别和回归估计等领域得到了广泛的应用[3]。

支持向量机是统计学习方法中的一种二分类模型，在特征空间中，支持向量机的学习策略是使可分的二分类数据间隔最大化，这样二分类问题会转化成一个求解凸二次规划问题。而 SMO 算法是支持向量机的一种快速算法，它通过不断地将原二次规划问题分解为只有两个变量的二次规划子问题，通过启发式的方法得到原二次规划问题的最优解。我们将凸二次规划问题最优解的非零变量对应的样本称为支持向量。根据实验研究发现，支持向量样本通常只占了全部样本的一小部分，且分布在离分类超平面较近的两类样本集的边界上，所以若能提前从原样本集中选取一个包含大多数支持向量的边界向量集作为训练集，通过 SMO 训练得到分类超平面，将能够减少训练样本，缩减训练时间，在保证原分类精度的基础上，提高训练速度。

在支持向量预处理的研究领域中，很多学者提出了许多方法。2005 年李青[4]提出了向量投影的边界向量预选取方法，该方法是找到两类中心点，将所有样本点做中心点连线的投影，并将所有样本按照其投影与同类中心距离远近进行预选取构造边界向量集。2006 年琚旭[5]提出了一种利用同心超球面来选取支持向量，找出两类中心。以每类中心为球心，以同类最远和最近样本点分别作为半径，然后将超球面间隔划分，最后对每一类样本点从外到内作为工作集进行训练。2013 年胡志军[6]提出基于距离排序的快速支持向量机分类算法，找出两类类中心，分别对两类样本按与异类中心点的距离远近进行预选取构造边界向量集，操作简单，有效实用。2013 年李庆[7]提出基于 KNN 思想的 K 边界近邻法支持向量预处理方法，该方法的优点在于预选取过程简单，且无论参数 K 取何值时都能够有效预选取支持向量。2016 年邱静[8]通过 FMC 聚类将两类中心点的位置进行优化，然后再使用其他以类中心距离为基础来进行支持向量预选取方法。2018 年王石[9]提出在 K 边界邻近法的基础上通过 CNN 方法再次压缩边界向量集，减少训练样本。以上支持向量预处理方法中构造出的边界向量集都期望最大程度上保留支持向量样本，减少原始训练样本。这样便能够在保持分类效果基本不变的情况下，缩短训练

时间，提高训练速度。

本文为满足预选取样本的要求提出了一种更简单实用的支持向量的预选取算法。从两类样本中分别任意选取一个样本进行配对，然后求其距离，并将所得的所有距离按照从小到大排序，排序过程中距离对应的样本点不能重复。这样便能构造出按距离从小到大排列的样本对序列，然后按一定比例抽取排在前面的样本对所关联的样本作为边界向量集，最后将构造出来的边界向量集通过 SMO 训练得到分类器。该算法基于距离配对排序，我们将其称为基于距离配对排序的支持向量预选取算法 DPS-SMO (Pre-selection of support vectors based on distance pairing sorting)。实验结果表明，该方法简单实用，在一定程度上预选取出大多数的支持向量，避免将所有样本参与训练，使得计算量变小，缩短了训练时间，提高了训练速度。

## 2. 支持向量机简介

假设训练样本集是 $\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中 $x_i=R^n$， $y_i\in\{-1,1\}$， $i=1,2,\cdots,N$，$N$ 是一个正整数，$R^n$ 表示输入模式的特征空间。根据结构风险最小化的思想，训练的目的就是寻找最优判决函数 $f(x)$，使得两类数据可以被分开，且泛化误差达到最小(或有上界)。Vapnik [2]指出：具有最大间隔的分类超平面可以满足上述条件。假定该分离超平面为 $w\cdot x+b=0$，它是由法向量 $w$ 和截距 $b$ 决定。在不同的情况下为找到该分离超平面，我们需要求解以下凸二次规划问题：

$$\begin{cases}\min \quad w(\alpha)=-\sum_{i=1}^{N}\alpha_i+\frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i,x_j),\\ s.t. \quad 0\le\alpha_i\le C,\ i=1,2,\cdots,N,\\ \qquad \sum_{i=1}^{N}y_i\alpha_i=0,\end{cases}\tag{1}$$

其中 $C\ge 0$ 是惩罚因子，$\alpha_i$ 是拉格朗日乘子，$K(x_i,x_j)=\phi(x_i)^{\mathrm{T}}\phi(x_j)$ 是核函数。通过求解(1)得 $\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_N^*)^{\mathrm{T}}$。计算法向量 $w^*=\sum_{i=1}^{N}\alpha_i^* y_i\phi(x_i)$，选择 $\alpha^*$ 的一个正分量 $0<\alpha_j^*<C$，计算截距 $b^*=y_j-\sum_{i=1}^{N}\alpha_i^* y_i K(x_i,x_j)$。引入核函数以后，最优判决函数可表示为 $f(x)=sign(w^*\cdot\phi(x)+b^*)$，其中 $sign(\cdot)$ 是一个符号函数。由凸二次规划问题得到最终二分类模型分类器为：

$$f(x)=sign\left(\sum_{i=1}^{N}y_i\alpha_i^*K(x,x_i)+b^*\right).\tag{2}$$

那些 $\alpha_i^*\ne 0$ 对应的样本被称为支持向量。显然只有支持向量会影响分类器的确定，但支持向量只占了所有样本的一小部分。如果支持向量可以被预选取出来进行 SMO 训练，这样将减少训练样本数量，进而缩短训练时间，提高训练速度。从几何的角度来看，支持向量主要分布在两类样本的边界上[9]。因此，本文提出的算法思想是通过距离配对排序选取两类样本边界上的样本构造边界向量集。

本文接下来第 3 节将具体描述基于距离配对排序的支持向量预选取算法，第 4 节通过数值实验结果及比较来表明基于距离配对排序的支持向量预选取算法的有效性。

## 3. 基于距离配对排序支持向量预选取算法

### 3.1. 样本距离

定义：由两个样本之间的特征差异定义样本距离，可先通过映射函数 $\phi(\cdot)$ 将原来的输入空间变换

到高维的特征空间中，使原始样本集在高维特征空间中线性可分[9]。任意两个样本 $x_1$，$x_2$，则其距离为：

$$d\left(x_1,x_2\right)=\sqrt{K\left(x_1,x_1\right)-2K\left(x_1,x_2\right)+K\left(x_2,x_2\right)}, \tag{3}$$

其中的 $K\left(\cdot,\cdot\right)$ 是一个核函数。若 $K\left(\cdot,\cdot\right)$ 是一个线性核函数，则(4)式将退化成欧氏距离：

$$d\left(x_1,x_2\right)=\left\|x_1-x_2\right\|_2=\sqrt{\sum_{i=1}^{n}\left(x_1^i-x_2^i\right)^2}, \tag{4}$$

其中上标 $i$ 是对应样本的第 $i$ 个分量。

### 3.2. 边界向量集

两类样本中离得较近的样本组成的集合为边界向量集。假设 $\left\{\left(x_1,y_1\right),\left(x_2,y_2\right),\cdots,\left(x_N,y_N\right)\right\}$ 为原始样本集，其中 $x_i\in R^n$，$y_i\in\left\{-1,1\right\}$，$i=1,2,\cdots,N$，$R^n$ 表示输入模式的特征空间。不妨设

$$X=\left\{\left(x_i,y_i\right)\middle|y_i=1,i=1,2,\cdots,n_1\right\}, \tag{5}$$

$$X'=\left\{\left(x'_j,y_j\right)\middle|y_j=-1,j=1,2,\cdots,n_2\right\}, \tag{6}$$

其中 $n_1+n_2=N$，易知 $x\cap x'=\phi$。两类样本之间任意两个样本进行配对然后按照 3.1 定义的样本距离求距离矩阵 $D$，即：

$$D=\begin{bmatrix} d\left(x_1,x'_1\right) & d\left(x_1,x'_2\right) & \cdots & d\left(x_1,x'_{n_2}\right) \\ d\left(x_2,x'_1\right) & d\left(x_2,x'_2\right) & \cdots & d\left(x_2,x'_{n_2}\right) \\ \vdots & \vdots & \ddots & \vdots \\ d\left(x_{n_1},x'_1\right) & d\left(x_{n_1},x'_2\right) & \cdots & d\left(x_{n_1},x'_{n_2}\right) \end{bmatrix}, \tag{7}$$

其中 $D_{ij}=d\left(x_i,x'_j\right)$，$x_i\in X$，$x'_j=X'$。将距离矩阵(7)中的元素按从小到大的顺序排序，排序过程中元素对应的样本不能重复使用。不失一般性，设 $n_1\le n_2$，那么排完序后得到 $n_1$ 个元素有序对，即 $D'=\left(d_1,d_2,\cdots,d_{n_1}\right)$，其中 $d_k<d_{k+1}$，$k=1,2,\cdots,n_1-1$。按照一定比例 $a$ 来截取前 $l$ 个元素，即：

$$l=\lceil n_1a\rceil, \tag{8}$$

$D'$ 经抽取之后的集合用 $D''$ 来表示，即：

$$D''=\left(d_1,d_2,\cdots,d_l\right), \tag{9}$$

其中每一个分量对应原始样本集中一个互异类样本对。为了方便描述，设 $S$ 为(9)式对应的原始样本组成的样本对集合，即：

$$S=\left\{\left(x_{k_1},x'_{k_1}\right),\left(x_{k_2},x'_{k_2}\right),\cdots,\left(x_{k_l},x'_{k_l}\right)\right\}. \tag{10}$$

把样本对集合式(10)分别表示成两类样本集合，第一类样本集合是 $\left\{x_{k_1},x_{k_2},\cdots,x_{k_l}\right\}$，第二类样本集合是 $\left\{x'_{k_1},x'_{k_2},\cdots,x'_{k_l}\right\}$。因此最终得到的边界向量集为：

$$A=\left\{x_{k_1},x_{k_2},\cdots,x_{k_l}\right\}\cup\left\{x'_{k_1},x'_{k_2},\cdots,x'_{k_l}\right\}. \tag{11}$$

下图 1 是两类样本距离矩阵中前三个最小距离对应的三个样本对抽取示意图。

因为选择的都是两类样本中互相最近的样本，选中的样本一定是边界向量，由结构风险最小化知，由边界向量构成的边界向量集一定能够很好的包含大多数的支持向量。最后将得到的边界向量集 $A$ 作为 SMO 算法的训练集进行训练，能够较速得到分类器。

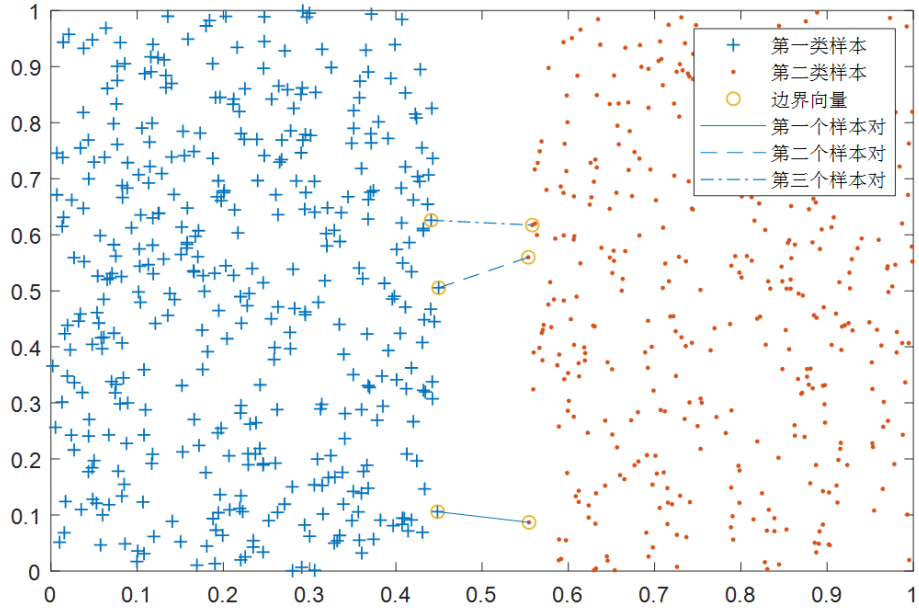


**Figure 1.** Diagram based on distance pairing sorting (first three sample pairs)
**图 1.** 基于距离配对排序示意图(前三个样本对)

### 3.3. 基于距离配对排序的支持向量预选取算法

基于距离配对排序的支持向量预选取算法 DPS-SMO (Pre-selection of support vectors based on distance pairing sorting)的伪码如下：

输入：原始样本集有两类样本集分别是

$X=\{(x_i,y_i)|y_i=1,i=1,2,\cdots,n_1\}$,

$X'=\{(x'_j,y_j)|y_j=-1,j=1,2,\cdots,n_2\}$,

其中 $n_1+n_2=N$，且 $n_1\le n_2$。抽取边界向量的比例值为 $a$。

输出：边界向量集 $A$

方法：

//计算两类样本之间的距离矩阵

For $i$ = 1 to $n_1$

For $j$ = 1 to $n_2$

$$d(x_i,x_j)=\sqrt{K(x_i,x_i)-2K(x_i,x_j)+K(x_j,x_j)}$$

end for

end for

$$D=\begin{bmatrix} d(x_1,x'_1) & d(x_1,x'_2) & \cdots & d(x_1,x'_{n_2}) \\ d(x_2,x'_1) & d(x_2,x'_2) & \cdots & d(x_2,x'_{n_2}) \\ \vdots & \vdots & \ddots & \vdots \\ d(x_{n_1},x'_1) & d(x_{n_1},x'_2) & \cdots & d(x_{n_1},x'_{n_2}) \end{bmatrix}$$

//对距离矩阵中的距离从小到大排序，过程中距离对应的样本不能重复，抽取前 $l$ 个较小距离

$l=\lceil n_1a\rceil$

$D'' = \{d_1, d_2, \cdots, d_l\}$

//排序抽取之后的距离集合对应的样本对集合

$S = \{(x_{k_1}, x'_{k_1}), (x_{k_2}, x'_{k_2}), \cdots, (x_{k_l}, x'_{k_l})\}$

//抽取的每类样本集合

$S_1 = \{x_{k_1}, x_{k_2}, \cdots, x_{k_l}\}$

$S_2 = \{x'_{k_1}, x'_{k_2}, \cdots, x'_{k_l}\}$

//数量为 $2l$ 个样本的边界向量集

$A = \{x_{k_1}, x_{k_2}, \cdots, x_{k_l}\} \cup \{x'_{k_1}, x'_{k_2}, \cdots, x'_{k_l}\}$

### 3.4. 算法复杂度分析

DPS-SMO 算法的复杂度主要取决于边界向量集的确定和 SMO 算法对边界向量集的训练。边界向量集确定的复杂度主要由计算两类样本之间的距离矩阵的复杂度和两类样本之间的距离排序的复杂度决定。

1) 设任意两个样本距离的计算复杂度为 $O(1)$。两类样本之间的距离矩阵需要 $n_1 n_2$ 次距离计算，其中 $n_1$ 为第一类样本的个数，$n_2$ 为第二类样本的个数，则两类样本之间的距离矩阵 $D$ 的计算复杂度为 $O(n_1 n_2)$。

2) 两类样本之间的距离矩阵有 $n_1 n_2$ 个不同的距离。把这些距离进行从小到大排序，根据排序理论[5]可得到其计算复杂度为 $n_1 n_2 \log_2 n_1 n_2$。然后要排除重复使用的样本，从第一个最小距离开始，第一个距离对应的两个样本，后面只要有距离对应与之相同的样本就删去此距离，后面的距离再补上来，以此类推下去，最终会得到一个距离排序的集合 $D'$，排除重复样本的距离总共有 $n_1 - 1$ 步，所以其计算复杂度为 $(n_1 - 1)O(1)$。两类样本之间的距离排序的计算复杂度为：

$$O(n_1 n_2 \log_2 n_1 n_2 + n_1 - 1). \tag{12}$$

3) 由于用标准的 SMO 算法训练原始样本集的复杂度为 $O(N^3)$，其中 $N$ 为原始样本集的样本个数，且 $n_1 + n_2 = N$。DPS-SMO 算法预选取的边界向量集里面的样本个数为 $2l$，其中 $l = n_1 a$。则 DPS-SMO 算法对边界向量集训练的计算复杂度为 $O(8l^3)$。

综上所述，当 $a$ 足够小时，DPS-SMO 的算法复杂度为：

$$O(n_1 n_2 \log_2 n_1 n_2 + n_1 - 1 + n_1 n_2 + 8l^3) \ll O(N^3). \tag{13}$$

因此基于距离配对排序的支持向量预选取算法可以以小于 $O(N^3)$ 的算法复杂度来训练，计算复杂度减小，训练时间也会随之缩短。

## 4. 基于距离配对排序支持向量预选取算法数值实验结果及比较

下面对 DPS-SMO 算法的可行性和有效性来进行实验和讨论。使其与无预处理 SMO 算法进行比较。本文的实验在 python3.6.3 环境中，2.3 GHz，Pentium，Dual CPU，4 GB 内存的硬件平台上进行。实验数据采用随机产生的线性可分数据和非线性可分数据以及 UCI 数据库中的 Banknote authentication 数据。DPS-SMO 算法和 SMO 算法对数据训练采用的核函数都是径向基核函数：

$$K(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{r^2}\right), \tag{14}$$

其中 $r = 1.3$。训练时间的计算是预选取支持向量构造边界向量集和 SMO 算法训练时间的总和。

### 4.1. 实验一

在线性可分情况下，DPS-SMO 算法与 SMO 算法的有效性及性能比较。我们随机产生两类线性可分的均匀分布点集数据。第一类数据 $S_1 \subset [0,0.45]\times[0,1]$，第二类数据 $S_2 \subset [0.55,1]\times[0,1]$。两类数据总共 1800 个，选取其中的 800 个数据作为训练数据，另外 1000 个数据作为测试数据[6]。首先将两类数据采用 3.1 定义的样本距离求其距离矩阵，然后采用 DPS-SMO 算法的算法步骤确定边界向量集，其中抽取比例值 $a=10\%$，最后将边界向量集作为训练集进行 SMO 训练。表 1 为线性可分情况实验下 DPS-SMO 算法与 SMO 算法分类性能的结果比较，其中数值结果是 100 次独立数值实验结果的平均值。

**Table 1.** Comparison of classification performance in linear separable case
**表 1.** 线性可分情况下分类性能比较

| 训练算法 | 边界向量数 | 支持向量数 | 训练时间(s) | 预测准确度(%) |
|---|---|---|---|---|
| SMO | 800 | 12 | 13.05 | 100 |
| DPS-SMO | 80 | 8 | 0.62 | 100 |

通过实验结果比较，我们发现原始训练数据经过 DPS-SMO 算法提取边界向量集再进行 SMO 训练与单纯的用 SMO 算法训练原始数据相比，训练数据大大减少，特别是训练时间降低的效果非常的明显，而且 DPS-SMO 算法预测准确度与 SMO 算法一样。实验结果说明了 DPS-SMO 算法在线性可分情况下是有效的，训练速度大大提升。图 2 是本数值实验下 DPS-SMO 算法预选取支持向量的效果图。

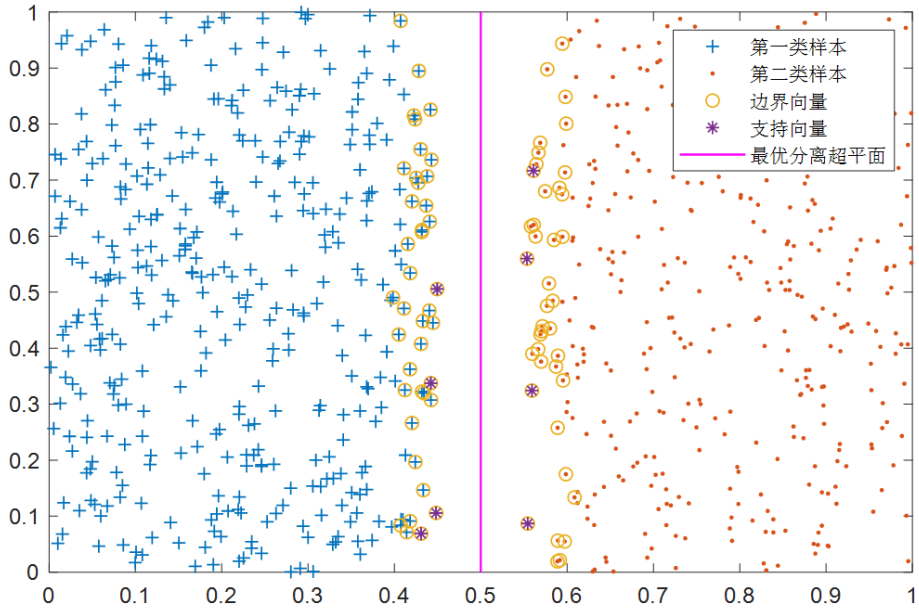


**Figure 2.** Pre-selection of support vector by DPS-SMO in the case of linear separability
**图 2.** 线性可分情况下 DPS-SMO 对支持向量的预选取

从图 2 中可以观察出 DPS-SMO 算法构造的边界向量集中的样本点都分布在两类样本点的相邻边界区域中，且边界向量集包含所有的支持向量。从图 2 中还可以观察出 DPS-SMO 算法排除了大量的非支持向量，在很大程度上减少了原始训练样本中无效样本数量，缩短了训练时间，提高了训练速度。

### 4.2. 实验二

在非线性可分情况下，DPS-SMO 算法与 SMO 算法的有效性及性能比较。我们随机产生两类非线性

可分的均匀分布的点集数据[6]。如随机产生两类均匀分布的同心圆样本：

$$\begin{cases} x_1 = \rho\cos\theta \\ x_2 = \rho\sin\theta \end{cases}, \quad \theta \in [0, 2\pi] \tag{15}$$

第一类数据采用半径为 $\rho_1 \in [0, 4.5]$，第二类数据采用半径为 $\rho_2 \in [5.5, 10]$，其中 $\rho_1$，$\rho_2$，$\theta$ 都是在对应的区域内均匀取值的。两类数据共有 1800 个。其中 800 个为训练数据，1000 个为测试数据。首先将两类数据采用 3.1 定义的样本距离求其距离矩阵，然后采用 DPS-SMO 算法的算法步骤确定边界向量集，其中抽取比例值 $a = 42.5\%$，最后将边界向量集作为训练集进行 SMO 训练。表 2 为非线性可分情况下 DPS-SMO 算法与 SMO 算法实验分类性能结果比较，其中数值结果是 100 次独立数值实验结果的平均值。

**Table 2.** Comparison of classification performance in nonlinear separable case
**表 2.** 非线性可分情况下分类性能比较

| 训练算法 | 边界向量数 | 支持向量数 | 训练时间(s) | 预测准确度(%) |
|---|---|---|---|---|
| SMO | 800 | 118 | 27.17 | 100 |
| DPS-SMO | 340 | 70 | 7.31 | 97.04 |

通过表 2 中的实验结果分析得出：由 DPS-SMO 算法步骤确定的边界向量集作为 SMO 算法的训练样本集与单纯地用 SMO 算法训练原始样本集相比较，训练样本的数量减少，时间上大大缩短，提高训练速度的效果很明显，而且预测准确度达到了 97.04%。这表明了在非线性可分情况下，DPS-SMO 算法可在保持很高的分类精度下，有效缩短训练时间，提高训练速度，达到了预期的效果。图 3 是本数值实验下 DPS-SMO 算法预选取支持向量效果图。

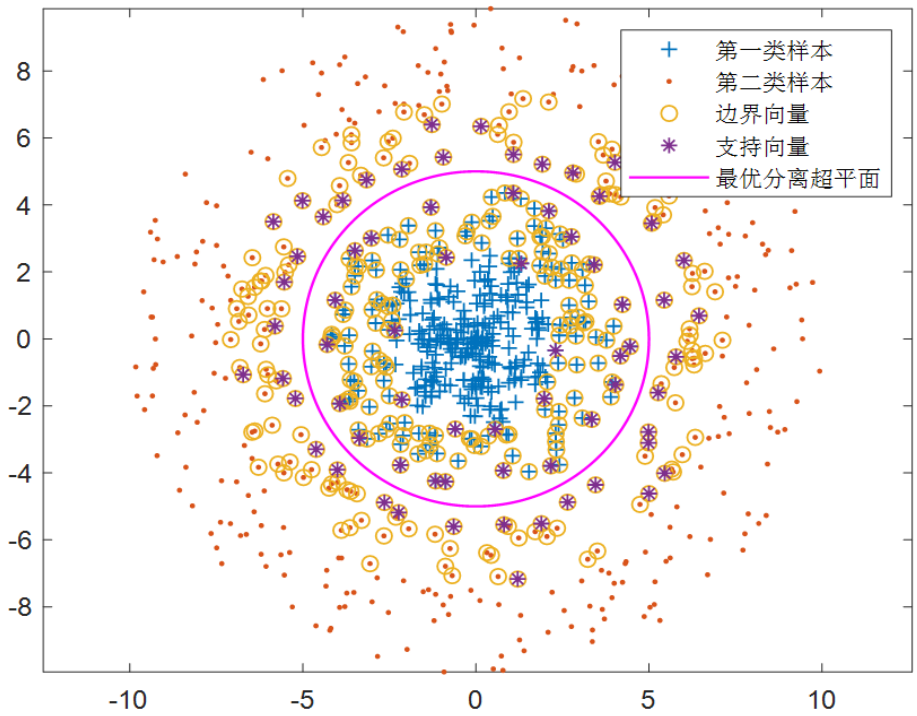


**Figure 3.** Pre-selection of support vector by DPS-SMO in the case of nonlinear separability
**图 3.** 非线性可分情况下 DPS-SMO 对支持向量的预选取

从图 3 明显可看出，通过 DPS-SMO 算法构造的边界向量集位于两类样本相邻的边界区域上。从中还可看出 DPS-SMO 算法去除了很多非支持向量，避免了 SMO 算法在非支持向量上训练时间的浪费，这样便能在保持良好的分类精度下有效提高训练速度。

### 4.3. 实验三

从 UCI 数据库中选取 Banknote authentication 数据用来测试 DPS-SMO 算法的性能。Banknote authentication 数据的每个样本是由 4 个特征属性和 1 个类别属性构成的，总共有 1372 个样本。我们分别用 DPS-SMO 算法与 SMO 算法对 Banknote authentication 数据进行测试。首先将两类数据采用 3.1 定义的样本距离求其距离矩阵，然后采用 DPS-SMO 算法的算法步骤确定边界向量集，其中抽取比例值：

$$a = 36.4\%.$$

数据测试的结果如表 3 所示，其中数值结果是 100 次独立数值实验结果的平均值。

**Table 3.** Comparison of classification performance under UCI data test
**表 3.** UCI 数据测试下分类性能比较

| 训练算法 | 边界向量数 | 支持向量数 | 训练时间(s) | 预测准确度(%) |
|---|---|---|---|---|
| SMO | 1372 | 230 | 123.96 | 100 |
| DPS-SMO | 500 | 140 | 20.98 | 93.15 |

通过观察表 3 可以得出采用 DPS-SMO 算法与单纯地使用 SMO 算法来测试数据相比，DPS-SMO 算法在大大缩短训练时间的同时，还能够保持很高的分类准确度。

## 5. 结语

本文根据支持向量的几何分布特征，提出了一种预选取支持向量的方法，简称 DPS-SMO 算法。该方法能够有效预选取一个包含大多数支持向量的边界向量集作为实验的训练集。通过实验得到分类器的过程中保证了很好的分类精度，又极大地减少了训练样本的个数，提高了支持向量机的训练速度，且训练样本冗余度越大，即支持向量所占比例越小时，方法效果越明显。DPS-SMO 算法与单纯地使用 SMO 算法训练原始数据作对比，实验验证了 DPS-SMO 方法有着缩短训练时间，提高训练速度的很好效果；如果时间允许，可以通过调整抽取比例的取值来提高 DPS-SMO 算法的性能，新方法预选取支持向量的过程简单，效果明显，有很强的使用广泛性，为支持向量机的应用提供了一种新的有效方法。

## 参考文献


[1] Vapnik, V.N. (2000) The Nature of Statistical Learning Theory. Spring-Verlag, New York.

[2] Vapnik, V.N. (1999) An Overview of Statistical Learning Theory. *IEEE Transactions on Neural Network*, **10**, 988-999. https://doi.org/10.1109/72.788640

[3] Yang, J., Yu, X. and Xie, Z.Q. (2011) A Novel Virtual Sample Generation Method Based on Gaussion Distribution. *Knowledge-Based Systems*, **24**, 740-748. https://doi.org/10.1016/j.knosys.2010.12.010

[4] 李青, 焦李成, 周伟达. 基于向量投影的支持向量预选取[J]. 计算机学报, 2005, 28(2): 145-151.

[5] 琚旭, 王浩, 姚宏亮. 基于同心超球面分割的支持向量预抽取方法[J]. 计算机工程与应用, 2006(31): 55-56 + 83.

[6] 胡志军, 王鸿斌, 张惠斌. 基于距离排序的快速支持向量机分类算法[J]. 计算机应用与软件, 2013, 30(4): 85-87 + 100.

[7] 李庆, 胡捍英. 支持向量预选取的 K 边界近邻法[J]. 电路与系统学报, 2013, 18(2): 91-96.

[8] 邱静, 徐晓钟, 邓松, 王婷. 基于优化模糊 C 均值聚类选取相似日的燃气负荷预测[J]. 上海师范大学学报(自然科学版), 2017, 46(4): 560-566.

[9] 王石, 蒋宁宁, 杨舒卉. 基于压缩 K 近邻边界向量的支持向量预抽取算法[J]. 海军工程大学学报, 2018, 30(6): 74-79.