**Computer Science and Application** 计算机科学与应用, 2021, 11(12), 3038-3050
Published Online December 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.1112307

Hans汉斯

# 基于云隐私保护的支持向量机训练外包

**邵宇航**

青岛大学计算机科学技术学院，山东 青岛

收稿日期：2021年11月25日；录用日期：2021年12月21日；发布日期：2021年12月28日

## 摘　要

**支持向量机(Support Vector Machine, SVM)是一种监督式学习机器学习分类算法，自出现以来，已经成为应用最广泛的分类算法。本文探讨了云环境下隐私保护的SVM训练问题，并提出了一种新的隐私保护的SVM外包训练(PPOSVM)协议。与已有的方案相比，除了隐藏数据样本的隐私性，新协议首次考虑了样本数据与标签对应关系的隐私性，隐藏了密文数据的访问模式。同时，采用豪斯霍尔德变换以及置换矩阵进行加密操作，使得新协议具有很高的效率。严格论证了协议的输入输出隐私性以及效率，同时，通过广泛的实验分析验证了所提协议的实际性能，进一步证实了理论分析。**

## 关键词

**支持向量机，隐私保护，豪斯霍尔德变换，云计算**

# Support Vector Machine Training Outsourcing Based on Cloud Privacy-Preserving

**Yuhang Shao**

College of Computer Science and Technology, Qingdao University, Qingdao Shandong

Received: Nov. 25th, 2021; accepted: Dec. 21st, 2021; published: Dec. 28th, 2021

## Abstract

**Support Vector Machine (SVM) is a supervised learning machine learning classification algorithm. Since its emergence, it has become the most widely used classification algorithm. This paper discusses the problem of SVM training for privacy-preserving in the cloud environment, and proposes a new privacy-preserving SVM outsourcing training protocol. Compared with the existing schemes, in addition to hiding the privacy of data samples, the new protocol considers the privacy of the**

文章引用: 邵宇航. 基于云隐私保护的支持向量机训练外包[J]. 计算机科学与应用, 2021, 11(12): 3038-3050.
DOI: 10.12677/csa.2021.1112307

**corresponding relationship between sample data and labels for the first time, and hides the access mode of ciphertext data. At the same time, the use of Householder transformation and permutation matrix for encryption operation makes the new protocol highly efficient. The input and output privacy and efficiency of the protocol are strictly demonstrated. At the same time, the actual performance of the proposed protocol is verified through extensive experimental analysis, which further confirms the theoretical analysis.**

## Keywords

**Support Vector Machine, Privacy-Preserving, Householder Transformation, Cloud Computing**



## 1. 引言

目前生活的社会是一个高速发展的社会，科技发达，信息流通，人们之间的交流越来越密切，生活也越来越方便，大数据就是这个高科技时代的产物。国际数据公司(IDC)报告称，2011 年全球创造了约 1.8 ZB (=1021 字节)的数据[1]。如物联网，网络物理系统，智慧城市[2]，智能医疗，智能导航[3]等每天都会产生和收集大量数据[4]。大数据时代给日常工作生活带来了巨大便利，如果能够合理并高效地对海量数据进行处理，就有可能产生巨大的商业价值。

近年来人工智能技术高速发展，已经广泛应用在社会生活的各个方面[5] [6]，成为人类社会智能进化的重要技术手段之一。而人工智能的发展决定了大数据技术的发展，由于人工智能技术具有强大的特征提取和抽象能力，整合多源信息，处理异构数据。大数据技术也为人工智能技术的发展提供了充足的训练样本。两者相得益彰。

人工智能算法——支持向量机(SVM)是在分类与回归分析中分析数据的监督式学习模型与相关的学习算法。给定一组训练实例，每个训练实例被标记为属于两个类别中的一个或另一个，SVM 训练算法创建一个将新的实例分配给两个类别之一的模型，使其成为非概率二元线性分类器。同时支持向量机也是一种广泛应用的机器学习算法，例如在在医疗，金融生物信息等领域。Devikanniga 等[7]使用支持向量机进行有效诊断肝病，Farhadian 等[8]使用支持向量机设计了牙周病决策诊断支持系统。Sivaram 等[9]使用支持向量机进行金融产品收益预测。Byvatov 等[10]将支持向量机应用在生物信息学中。

大数据时代，在本地计算资源和存储资源有限的情况下，对于大规模数据处理是一件很苦难的事情，而云计算服务可以很好地解决这个问题，本地设备可以将数据外包给计算能力和存储能力都很强的云服务器，并向云服务器支付一定的费用，从而换取相应的计算服务，将支持向量机(SVM)外包训练给云计算服务提供商，有利于不熟悉支持向量机技术和本地计算资源有限的数据所有者使用完整的支持向量机分类功能。但支持向量机的训练通常需要庞大的数据样本进行支撑，因此将支持向量机训练外包给云服务器是合理高效的。

但是将支持向量机训练给云服务器会产生数据安全问题，可能会导致意思隐私数据泄露。例如，病人的生理特征数据以及诊疗数据关系到病人的个人隐私，如果可能会造成社会歧视并且对病人造成更大的心理伤害。金融机构的用户数据关系到用户的财产安全，一旦泄露可能会使不法分子有机可乘，给用户带来财产损失。而训练产生的机器学习模型也是模型拥有者的知识产权，其中包含着商业价值，一旦

泄露可能会导致巨大损失。因此在对数据进行外包计算时，需要在不影响计算结果的情况下。设计加密算法保护数据的隐私性。一些现有的工作也考虑外包训练支持向量机的数据的隐私性，Wang 等[11]使用同态加密技术外包了医疗物联网支持向量机模型的训练，Liu 等[12]使用顺序最小优化技术实现了用于药物发现支持向量机模型的安全外包训练，Lin 等[13]使用随机变换技术外包训练支持向量机模型，Serrano 等[14]使用同态加密外包训练了支持向量机模型。

对于支持向量机的隐私保护训练，Laur 等[15]和 Omer 等[16]随后提出了基于安全多方计算和 HE 技术的分布式隐私数据支持向量机的隐私保护协议。Lin 等[13]提出了在外包设置中训练支持向量机的协议。他们通过随机线性变换扰动数据来实现隐私保护。最近，Shen 等[17]和 Wang 等[11]分别研究了智慧城市和互联网中隐私保护的 SVM 训练。尽管如此，他们的协议都涉及耗时的同态加密操作。而耗时的同态加密会增加客户端开销。同时对于训练样本密文数据的访问模式，即：标签相同的密文数据对应的明文数据标签也相同。

本文讨论了外包 SVM 训练中的数据隐私以及分类标签隐私问题，并设计了一种从在加密数据中训练 SVM 的方案。使用 Householder 矩阵对样本数据进行加密，使用 Householder 矩阵加密可以降低计算复杂度，并无需公开数据的实际内容。同时使用置换矩阵盲化样本与其分类标签的对应关系，并且加密后的数据不影响最终的分类模型。同时由于使用的是加密后的样本数据，因此服务提供商无法获得真实的支持向量机模型，除非使用来自数据提供者发送的加密样本测试数据。同时还设计了一个三方框架，来保证训练过程安全可靠并且保持高效。包括数据提供者(Data provider)、云服务器(Cloud server)、分类测试者(Data tester)，基本流程如下所示：首先数据提供者将数据进行加密，然后将 SVM 训练任务外包给云服务器，云服务计算获得相应的模型参数后将其发送给数据提供者，最终机器学习模型部署在数据提供者方以供其他数据测试者进行分类测试。具体来说，本文的贡献可以概括如下：

1) 协议从两个方面实现了隐私保护的目标。(1,1)云服务器无法知道样本数据内容和样本数据的分类标签的真实数据。(1,2)由于使用了 Householder 矩阵对数据进行加密，云服务器无法生成真实的支持向量机训练模型。

2) 协议是高效的。通过使用 Householder 矩阵、置换矩阵实现隐私保护要求，避免了沉重的(完全)同态加密操作。隐私保护技术的简洁性和易用性使本文设计的协议获得了高性能。此外，通过外包，数据提供者(DP)实现了可观的计算节省。去后进行了大量的实验来验证理论主张。

本文的剩余组织部分如下：第 2 节介绍系统架构，威胁模型和设计目标。第 3 节列出本文使用的基本数学工具。第 4 节具体给出了 PPOSVM 协议的设计，并在第 5 节分析了它的正确性、安全性和效率，第 6 节对 PPOSVM 协议性能进行了实验评估。最后总结工作在第 7 节。

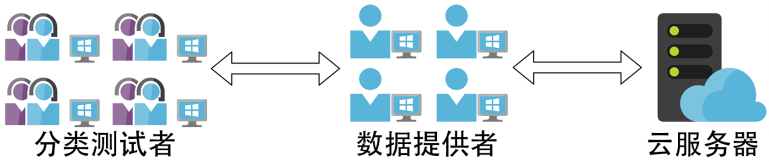


**Figure 1.** System model
**图 1.** 系统模型图

## 2. 系统架构、威胁模型和设计目标

### 2.1. 系统架构

如图 1 所示，PPOSVM 协议由三方组成，数据提供者(Data provider)、云服务器(Cloud server)、分类

测试者(Data tester)。

DP，数据提供者拥有大量的训练样本数据，他希望通过云服务器训练相应的支持向量机分类模型。由于样本数据的内容存在潜在的敏感性，他希望云服务器无法知道真实的样本数据内容以及样本数据对应的分类标签。

CS，它是一个云/边缘服务器，具有强大的计算、数据分析和存储资源。出于经济目的，它可以与资源受限的客户共享计算资源，帮助他们完成计算任务。然而，它可能是好奇的。

DT，当支持向量机模型部署到 DP 上时，他可以将测试样本发送到 DP 以获取分类结果。

### 2.2. 威胁模型

系统中的云服务器采用与之前的隐私保护机器学习分类服务相似的“诚实但好奇”的威胁模型。假设云服务器对数据提供者的数据样本内容，数据样本和分类标签的对应关系以及训练后的支持向量机模型感兴趣，但是它只能访问中间结果，不能访问任何最终结果。假设云服务器会遵守协议内容，同时努力推断自己感兴趣的内容，并且假设各方之间不会串通。

### 2.3. 设计目标

本文设计云辅助的隐私保护支持向量机外包训练(PPOSVM)协议，以确保系统中的所有各方都能高效、安全地运行，并最终实现准确的结果。现在，详细阐述以下目标

1) 正确性，如果所有参与者都忠实地执行分配的计算任务，该协议能够最终获得正确的支持向量机模型。

2) 高效性，由于将支持向量机模型训练任务外包到云服务器，最终 PPOSVM 协议所设计的计算过程成本必须要低于本地计算的成本。

3) 输入隐私，样本数据内容和其分类标签可能包含敏感信息，所以应该对云服务器不可见。

4) 输出隐私，训练后的支持向量机模型含有商业价值，应该向云服务器保密。

## 3. 预备知识

在这一节中将介绍一些基本概念和设计中使用的工具。

### 3.1. 支持向量机

支持向量机(SVM)，首先由 Vapnik [18]提出，是在分类与回归分析中分析数据的监督式学习模型与相关的学习算法。给定一组训练实例，每个训练实例被标记为属于两个类别中的一个或另一个，SVM 训练算法创建一个将新的实例分配给两个类别之一的模型，使其成为非概率二元线性分类器。SVM 模型是将实例表示为空间中的点，这样映射就使得单独类别的实例被尽可能宽的明显的间隔分开。然后，将新的实例映射到同一空间，并基于它们落在间隔的哪一侧来预测所属类别。除了进行线性分类之外，SVM 还可以使用所谓的核技巧有效地进行非线性分类，将其输入隐式映射到高维特征空间中。假设有 $m$ 个训练样本实例，每个实例由 $\left(x_i; y_i\right), x_i \in R^n$ 包含第 $i$ 个实例的属性和它的类别标签 $y_i \in \{+1;-1\}$。支持向量机最终求解以下二次规划问题，找到最优超平面 $w^{\mathrm{T}} x+b=0$。

$$\begin{cases} \min\limits_{w,b} \dfrac{1}{2}|w|^2 \\ \text{s.t. } y_i\left(wx_i+b\right) \geq 1, i=1,\cdots,m \end{cases} \tag{1}$$

在目标函数中最小化 $\dfrac{1}{2}\left|w^2\right|$ 意味着最大化两类数据之间的边际。通过拉格朗日乘子法将有约束问题转

换为无约束问题，最后求解 SVM 可以转化为求解 SVM 的对偶问题。

$$\begin{cases}\min\limits_{\alpha}\sum_{i=1}^{m}\sum_{j=1}^{m}-\dfrac{1}{2}\alpha_i\alpha_j y_i y_j x_i x_j+\sum_{i=1}^{m}\alpha_i \\ \text{s.t. } \alpha_i\geq 0,\sum_{i=1}^{m}\alpha_i y_i=0, i=1,\cdots,m\end{cases} \tag{2}$$

将(2)转换为矩阵形式为

$$\begin{cases}\min\limits_{\alpha}-\dfrac{1}{2}(XDY)^{\mathrm{T}}XDY+Sum(D) \\ \text{s.t. } \alpha_i\geq 0,\sum_{i=1}^{m}\alpha_i y_i=0, i=1,\cdots,m\end{cases} \tag{3}$$

其中训练数据样本矩阵 $X=(x_1\cdots x_m)\in R^{n\times m}$，$n$ 表示数据样本维数，$m$ 表示训练所需的数据样本数量，$D=\mathrm{diag}\{\alpha_1,\cdots,\alpha_m\}$ 为拉格朗日乘子的 $m\times m$ 对角矩阵，对角线为拉格朗日乘子 $\alpha=(\alpha_1\cdots\alpha_m)$，$Y$ 为数据样本的分类标签的 $m$ 维向量，$Y^{\mathrm{T}}=(y_1\cdots y_m)$。$Sum(D)$ 表示对 $D$ 中所有元素求和。

$$D=\begin{pmatrix}a_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & a_m\end{pmatrix}$$

### 3.2. 豪斯霍尔德变换(Householder)变换

在本文的设计中使用了一种特殊的正交变换 Householder 变换，由 $n$ 维向量 $v=(v_1\cdots v_n)^{\mathrm{T}}$ 生成的 Householder 变换 $H$ 可以被定义为：

$$H_v(x)=Qx,\forall x\in R^n$$

这里

$$Q=I-\frac{2vv^{\mathrm{T}}}{|v|^2}$$

$Q$ 是一个 $n\times n$ 的正交矩阵，$I$ 是单位矩阵。对于任何 $x,y\in R^n$，$\langle Qx,Qy\rangle=\langle x,y\rangle$。

### 3.3. 随机置换及其在矩阵上的作用

$n$ 阶(随机)置换 $\pi$ 是集合 $\{1,2,\cdots,n\}$ 上的(随机)双射，可以表示为：

$$\pi=\begin{Bmatrix}1 & 2 & \cdots & n\\ \pi(1) & \pi(2) & \cdots & \pi(n)\end{Bmatrix}$$

其中 $(\pi(1),\pi(2),\cdots,\pi(n))$ 是 $(1,2,\cdots,n)$ 的一个(随机)重排。任一随机置换可由一个众所周知的高效洗牌算法 **Alg.1** 生成[16]。

Algorithm 1　随机置换生成

输入：一个参数 $n$
输出：一个随机置换 $\pi$
集合 $(resp.\ \pi=I_n)$ 是相同的排列
For $i=n:2$
选择一个随机整数 $j=(1\leq j\leq i)$
交换 $\pi(i)$ 和 $\pi(j)$
返回 $\pi$

给定一个 $n$ 阶置换 $\pi$，一个 $n\times m$ 阶矩阵

$$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{im} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix},$$

定义 $A$ 上的左置换 $\pi$ 为

$$\pi\circ A=\begin{pmatrix} a_{\pi(1)1} & a_{\pi(1)2} & \cdots & a_{\pi(1)m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{\pi(i)1} & a_{\pi(i)2} & \cdots & a_{\pi(i)m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{\pi(n)1} & a_{\pi(n)2} & \cdots & a_{\pi(n)m} \end{pmatrix},$$

例如，取 $n=3,m=4$ 的排列

$$\pi=\begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix},$$

矩阵 $A$ 为

$$A=\begin{pmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{pmatrix}.$$

所以

$$\pi\circ A=\begin{pmatrix} 3 & 6 & 9 & 12 \\ 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \end{pmatrix}.$$

接下来给出一个简单但有用的性质

引理 1.对于任何 $n\times m$ 矩阵 $A=(a_1\cdots a_m)$，一个 $n$ 维列向量 $x=(x_1\cdots x_n)^{\mathrm{T}}$，一个 $n$ 阶置换 $\pi$，有

$$\langle\pi\circ A,\pi\circ x\rangle=\langle A,x\rangle.$$

## 4. 隐私保护的支持向量机训练外包(PPOSVM)协议

本节将介绍设计的 PPOSVM 协议，用 $m$ 来表示数据拥有者训练支持向量机模型所需要的样本 $x_i$ 的量 $i\in(1\cdots m)$，可能有时候需要大量数据样本来进行模型训练。根据支持向量机原理，需要使用云服务器来求解一个二次规划问题，并最终由云服务器返回相应的训练结果参数以供数据拥有者计算支持向量机模型 $w^{\mathrm{T}}x_i+b=0$。正如前文所说的，云服务器可能会对数据样本 $x_i$，样本分类标签 $y_i\in(1\cdots m)$ 以及支持向量机模型感到好奇。因此无法将实际的数据样本以及样本分类标签发送给云服务器。本文使用 Householder 矩阵 $Q$ 对数据样本进行保护同时使用置换矩阵 $\pi$ 对分类标签进行加密。

如图 2 所示，PPOSVM=(DPtoCSKEYGEN, DPtoCSENC, CSTraining，DPModleCon，DTTest)的具体过程如下：

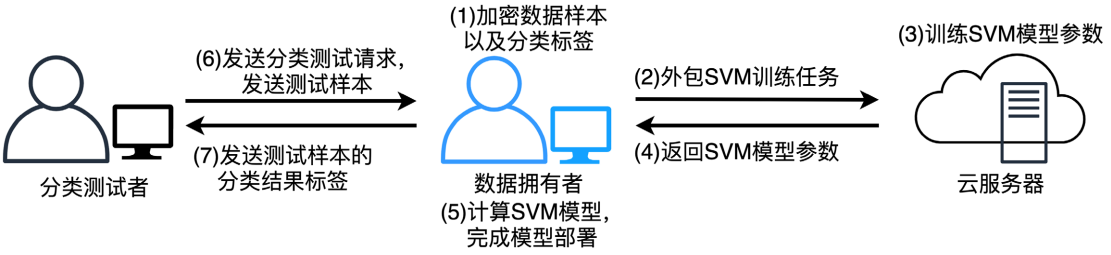


**Figure 2.** Support vector machine training outsourcing model
**图 2.** 支持向量机训练外包模型

1) DPtoCSKEYGEN：数据拥有者(DP)生成 $n$ 维随机向量 $q=(q_1\cdots q_n)$，其中 $q_i$ 为随机均匀选取 $\lambda$ 比特实数。同时，调用 **Alg.1** 输出 $n$ 阶随机排列 $\pi$，所有这些都应该由数据拥有者保密。

2) DPtoCSENC：数据拥有者对数据样本 $X=(x_1\cdots x_m)\in R^{n\times m}$，分类标签 $Y^{\mathrm{T}}=(y_1\cdots y_m)$ 进行加密，

$$X'=QX,\ Y'=\pi Y$$

然后数据拥有者将 $X_i'$ 和 $Y_i'$ 发送给云服务器。(如图 2 步骤(1) (2)所示)

3) CSTraining：云服务器收到支持向量机训练任务后，执行相应的 SVM 训练算法，例如，SMO 算法。计算相应的二次规划问题[19] [20] [21]，

$$\begin{cases}\min\limits_{\alpha}-\dfrac{1}{2}\left(X'D'Y'\right)^{\mathrm{T}}X'D'Y'+Sum\left(D'\right)\\ \text{s.t. }\alpha_i\geq 0,\sum_{i=1}^{m}\alpha_i y_i=0,i=1,\cdots,m\end{cases}\tag{4}$$

其中

$$Q=I-\frac{2qq^{\mathrm{T}}}{|q|^2}$$

表示与向量 $q$ 相关的 Householder 矩阵，$Sum(D')$ 表示对矩阵 $D'$ 所有元素求和。获得相应满足式子(4)条件的的模型参数 $D'$。并将模型参数 $D'$ 返回给数据提供者。(如图 2 步骤(3)、(4)所示)

4) DPModleCon：数据提供者计算 $D=\pi D'$。解密后的 $D$ 按照行序选择 $\alpha=(\alpha_1\cdots\alpha_m)$，然后使用模型参数 $\alpha$ 计算相应的支持向量机模型 $w^{\mathrm{T}}x+b=0$。

$w$ 可以由以下式子求出

$$w=\sum_{i=1}^{m}\alpha_i y_i x_i\tag{5}$$

$b$ 可以由以下式子推出

$$\exists x_k y_k\ \text{ s.t. }1-y_k\left(w^{\mathrm{T}}x_k+b\right)=0$$

由于 $y_k\left(w^{\mathrm{T}}x_k+b\right)=1$，两边同时乘 $y_k$ 得 $y_k^2\left(w^{\mathrm{T}}x_k+b\right)=y_k$，又因为 $y_k\in(+1,-1)$，所以 $w^{\mathrm{T}}x_k+b=y_k$，即

$$b=y_k-w^{\mathrm{T}}x_k.\tag{6}$$

数据拥有者使用(5)、(6)两式可以得到 SVM 最优超平面，数据拥有者将 SVM 模型部署于自身。(如图 2 步骤(5)所示)

5) DTTest：在 SVM 模型部署到数据拥有者方时，数据测试者可以向数据拥有者发送分类请求和测试样本 $x_{ci}$，测试样本的分类结果标签 $y_{ci}$ 由下面的公式判定：

$$\begin{cases} w^{\mathrm{T}} x_{ci} + b \geq 1, y_{ci} = +1 \\ w^{\mathrm{T}} x_{ci} + b \leq 1, y_{ci} = -1 \end{cases}$$

当数据拥有者获取分类结果标签 $y_{ci}$ 时，将其发送给数据测试者。(如图 2 步骤(6)、(7)所示)

## 5. 正确性、安全性和效率分析

这一节将详细分析 PPOSVM 协议的正确性，安全性，高效性。

### 5.1. 正确性分析

根据设计目标中正确性的含义，定义

定义 1 隐私保护的支持向量机外包训练协议是正确的。

使用诚实但好奇的云服务器完成训练任务，那么使用加密数据进行计算不会影响最终的分类结果。支持向量机训练任务需要解决的是以下二次规划问题

$$\begin{cases} \min\limits_{\alpha} -\dfrac{1}{2}(XDY)^{\mathrm{T}} XDY + Sum(D) \\ \text{s.t. } \alpha_i \geq 0, \sum_{i=1}^{m} \alpha_i y_i = 0, i = 1, \cdots, m \end{cases} \tag{7}$$

加密算法使用到 Householder 变换矩阵 $Q$ 以及随机置换矩阵 $\pi$，因此将加密后的二次规划问题表示如下

$$\begin{cases} \min\limits_{\alpha} -\dfrac{1}{2}(QXD'\pi Y)^{\mathrm{T}} QXD'\pi Y + Sum(D') \\ \text{s.t. } \alpha_i \geq 0, \sum_{i=1}^{m} \alpha_i y_i = 0, i = 1, \cdots, m \end{cases} \tag{8}$$

经过展开后可以得到

$$\begin{cases} \min\limits_{\alpha} -\dfrac{1}{2} Y^{\mathrm{T}} \pi^{\mathrm{T}} D'^{\mathrm{T}} X^{\mathrm{T}} Q^{\mathrm{T}} QXD'\pi Y + Sum(D') \\ \text{s.t. } \alpha_i \geq 0, \sum_{i=1}^{m} \alpha_i y_i = 0, i = 1, \cdots, m \end{cases} \tag{9}$$

Householder 变换产生的矩阵 $Q$ 为正交矩阵，而置换矩阵本质上是混淆了训练数据样本与分类标签之间的关系，根据正交矩阵的性质 $Q^{\mathrm{T}}Q = 1$，可以将式子(9)化简为

$$\begin{cases} \min\limits_{\alpha} -\dfrac{1}{2} Y^{\mathrm{T}} \pi^{\mathrm{T}} D'^{\mathrm{T}} X^{\mathrm{T}} XD'\pi Y + Sum(D') \\ \text{s.t. } \alpha_i \geq 0, \sum_{i=1}^{m} \alpha_i y_i = 0, i = 1, \cdots, m \end{cases} \tag{10}$$

对于 $Sum(D)$，仅需要将矩阵 $D$ 中所有元素求和即可。对于式子(10)获得的模型参数矩阵为 $D' = D\pi$，因此数据拥有者收到加密的模型参数矩阵后，只需要对其进行解密可以获得正确的模型参数矩阵 $D = \pi D'$。经过以上数学分析，PPOSVM 协议可以实现与原问题相同的解决方法与结果，因此满足正确性的定义。

### 5.2. 安全性分析

首先讨论输入输出隐私，根据设计目标，可以证明

定理 2 对于任意规模的输入矩阵 $X'$，提出的支持向量机外包训练算法满足输入隐私性。

证明：首先证明数据样本内容的隐私性，即云服务器知道加密后的训练样本矩阵 $X'$，云服务器能够

恢复真实训练样本矩阵 $X$ 的概率可以忽略不计。根据加密算法得到加密后的 $X'$，而 $X'=QX$，$X'$ 取决于正交矩阵 $Q$，那么 $X=Q^{-1}X'$，云服务器恢复 $X$ 的概率为

$$\frac{1}{\left\{X \mid X=Q^{-1}X'\right\}}=\left\{Pr(Q)\right\}\le\frac{1}{\left\{Q \mid Q=I-\frac{2qq^{\mathrm{T}}}{|q|^2}\right\}}=\frac{1}{2^{\lambda n}}.$$

此概率显然是微不足道的。

其次证明样本与标签对应关系的隐私性。即使云服务器知道加密后的训练样本矩阵 $Y'$，云服务器能够恢复真实训练样本分类标签 $Y$ 的概率可以忽略不计。

对于训练样本分类标签 $Y$，使用随机置换对其进行加密 $Y'=\pi Y$，$Y'$ 取决于随机置换 $\pi$，那么 $Y=\pi^{-1}Y'$，云服务器恢复 $Y$ 的概率

$$\frac{1}{\left\{Y \mid Y=\pi^{-1}Y'\right\}}=\left\{Pr(\pi)\right\}\le\frac{1}{m!}$$

这显然是微不足道的。通过上述证明，可以说明设计的协议满足输入隐私。

定理 3 对于任意规模的输入矩阵 $X'$，提出的支持向量机外包训练算法满足输出隐私性。

证明：在云服务器知道输出的参数矩阵 $D'$ 后，证明云服务器获得真实的 $D$ 的概率，由于 $D=\pi D'$，所以云服务器获得真实的 $D$ 的概率为：

$$\frac{1}{\left\{D \mid D=\pi^{-1}D'\right\}}=\left\{Pr(\pi)\right\}\le\frac{1}{m!}$$

这显然是微不足道的。通过上述证明，可以说明设计的协议满足输出隐私。

### 5.3. 效率分析

在本节中，将展示 PPOSVM 协议也满足高效率要求。也就是说，与在没有云服务器帮助的情况下本地实现支持向量机的训练，设计中的数据提供者可以获得可观的计算时间节省。

在 DPtoCSKEYGEN 过程中，数据提供者需要进行次 $n$ 随机数生成操作和 $m$ 次置换操作，这是效率很高的。

在 DPtoCSENC 过程中，由于矩阵乘法的关联性和 Householder 矩阵的简洁构造，数据拥有者可以通过快速矩阵向量乘法高效地计算，避免了耗时的矩阵－矩阵乘法。即，

$$X'=QX=\left(I-\frac{2qq^{\mathrm{T}}}{\|q\|^2}\right)X=X-\frac{2}{\|q\|^2}q\left(q^{\mathrm{T}}X\right)$$

因此在整个支持向量机训练过程中数据提供者的时间开销是 $t=O(nm)$。因此的协议可以实现以下时间节省

$$t_{\text{本地计算svm}}/t_{\text{隐私保护外包计算svm}}=O\left(n^2m/nm\right)=O(n).$$

## 6. 实验分析和实验效果评价

为了全面评估 PPOSVM 协议在实践中的实际性能，在本节中对 PPOSVM 协议进行实验分析。为了响应设计目标中的效率要求，实验着重于模拟数据提供者和云服务器的时间开销，并将其开销与无需外包的算法进行比较。

### 6.1. 实验环境

DP 端实验环境配置：Intel(R)Core(TM)i5-8500T 2.10GHz CPU 和 8GB RAM 的笔记本电脑上，实验中使用 Pycharm community 2021 模拟数据提供者。云端实验环境配置：Intel(R)Core(TM)i7-9750H 2.6GHz 和 16GB RAM 的 MacBookPro，实验中运行 Pycharm community 2021 模拟云服务器。

### 6.2. 实验方法

一个设计良好的外包协议应该确保数据提供者节省计算时间，并且使用协议执行支持向量机训练总时间开销应该大大少于数据提供者自己完成支持向量机训练的时间开销。因此，从以下两个角度通过实验评估 PPOSVM 协议的有效性。1) 比较 PPOSVM 协议中数据提供者的时间开销与数据提供者自己完成支持向量机训练的时间开销。2) 将 PPOSVM 协议中数据提供者的时间开销和云服务器的时间开销的总和与数据提供者自行完成支持向量机训练的时间进行比较。

### 6.3. 实验结果和分析

在实验中有两个重要参数$(m,n)$，由于在实际应用中训练样本维数一般是固定的，因此在确定训练样本维数的情况下改变训练样本数目，进行了表 1~3 三组实验。

**Table 1.** Privacy-preserving support vector machine outsourcing training protocol experimental results $(n=10)$
**表 1.** 隐私保护的支持向量机外包训练协议实验结果$(n=10)$

| | $t_{加密}$ | $t_{解密}$ | $t_{本地计算}$ | $t_{云计算}$ | $t_{本地计算}/t_{加密+解密}$ | $t_{本地计算}/t_{加密+解密+云计算}$ |
|---|---|---|---|---|---|---|
| $m=100$ | 0.0033 s | 0.0002 s | 24.7864 s | 8.1178 s | 7472 | 3.05 |
| $m=150$ | 0.0039 s | 0.0008 s | 47.0321 s | 15.056 s | 11925 | 3.12 |
| $m=200$ | 0.0043 s | 0.0013 s | 87.9824 s | 24.1569 s | 20147 | 3.64 |

**Table 2.** Privacy-preserving support vector machine outsourcing training protocol experimental results $(n=20)$
**表 2.** 隐私保护的支持向量机外包训练协议实验结果$(n=20)$

| | $t_{加密}$ | $t_{解密}$ | $t_{本地计算}$ | $t_{云计算}$ | $t_{本地计算}/t_{加密+解密}$ | $t_{本地计算}/t_{加密+解密+云计算}$ |
|---|---|---|---|---|---|---|
| $m=100$ | 0.0041 s | 0.0002 s | 24.8671 s | 11.9459 s | 6057 | 2.08 |
| $m=150$ | 0.0045 s | 0.0007 s | 55.3462 s | 23.7667 s | 11060 | 2.32 |
| $m=200$ | 0.0050 s | 0.0014 s | 94.8330 s | 37.0008 s | 14675 | 2.56 |

**Table 3.** Privacy-preserving support vector machine outsourcing training protocol experimental results $(n=30)$
**表 3.** 隐私保护的支持向量机外包训练协议实验结果$(n=30)$

| | $t_{加密}$ | $t_{解密}$ | $t_{本地计算}$ | $t_{云计算}$ | $t_{本地计算}/t_{加密+解密}$ | $t_{本地计算}/t_{加密+解密+云计算}$ |
|---|---|---|---|---|---|---|
| $m=100$ | 0.0053 s | 0.0001 s | 25.3667 s | 13.1271 s | 4732 | 1.93 |
| $m=150$ | 0.0071 s | 0.0005 s | 56.1142 s | 25.8637 s | 7812 | 2.16 |
| $m=200$ | 0.0081 s | 0.0013 s | 102.0006 s | 43.5954 s | 11969 | 2.24 |

具体实验描述如下，固定训练样本维数 $n=10$，$n=20$，$n=30$，改变训练样本数量 $m$，$m$ 在 100~200 之间变化，步长为 50。为了避免出现偶然的数据点，因此得到的每个数据都是算法运行 10 次的平均运

行时间，$t_{加密}$表示数据拥有者对数据样本以及分类标签进行 Householder 变换以及随机置换的时间开销，$t_{解密}$表示数据拥有者对模型参数进行解密的时间开销，$t_{本地计算}$表示支持向量机训练任务本地设备独立计算所需要的时间开销，$t_{云计算}$表示将支持向量机训练任务外包到云服务器进行计算所需要的时间开销。同时用两个指标$t_{本地计算}/t_{加密+解密}$、$t_{本地计算}/t_{加密+解密+云计算}$来表示设计的协议的效率。从图 3 和图 4 可以看出，随着训练样本数量不断提高，$t_{本地计算}/t_{加密+解密}$、$t_{本地计算}/t_{加密+解密+云计算}$两个指标也不断提升，这表明，随着训练样本数的增加，所提出的支持向量机训练外包协议可以节省更多的计算量。

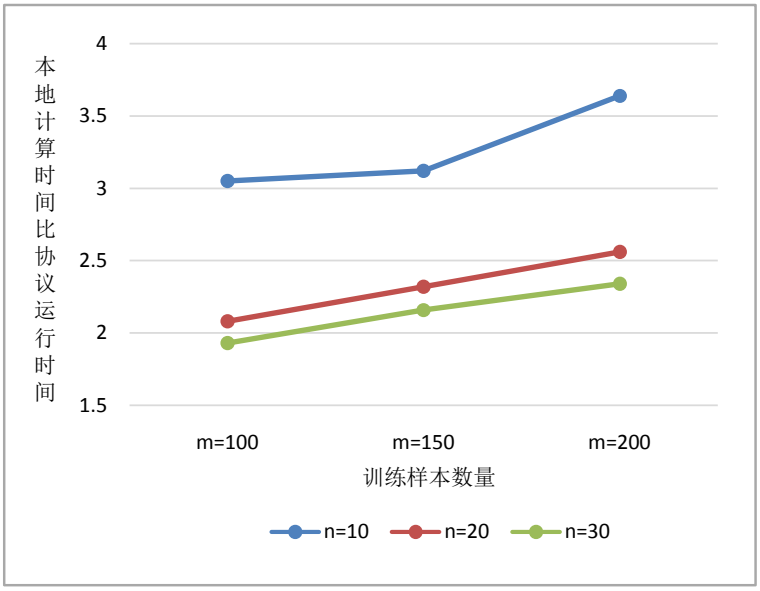


**Figure 3.** Ratio of local computing time to protocol running time
**图 3.** 本地计算时间和协议运行时间之比

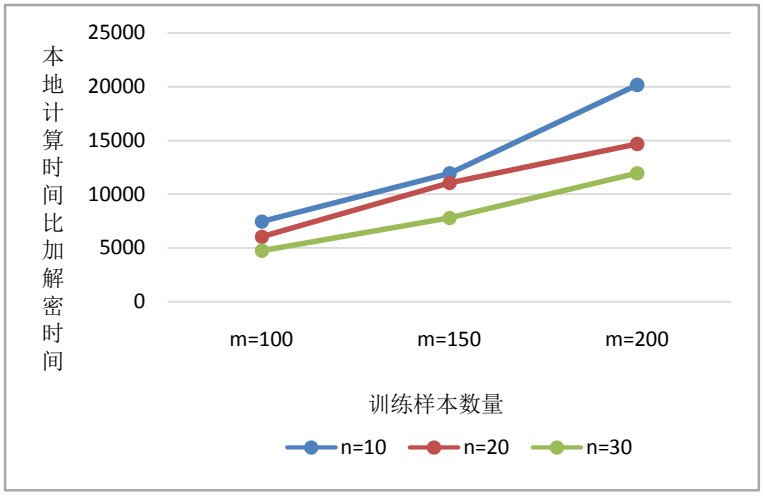


**Figure 4.** Ratio of local computing time to encryption and decryption time
**图 4.** 本地计算时间和加解密时间之比

## 6.4. 对照比较

在这一节中比较了协议和其他的相关保护支持向量机训练外包协议的不同，如表 4 所示。

**Table 4.** Comparison of existing work
**表 4.** 现有工作对比

| 现有工作 | 同态加密 | 保护训练样本隐私 | 保护分类模型 | Householder 变换 | 保护分类标签 |
|---|---|---|---|---|---|
| Lin [13] | × | √ | √ | × | × |
| Maekawa [22] | × | √ | √ | × | × |
| Wang [11] | √ | √ | √ | × | × |
| PPOSVM | × | × | × | √ | √ |

## 7. 总结

本文提出了一种新的高效且隐私保护的外包支持向量机方案，该方案支持安全训练，以保护训练数据样本以及分类标签。同时使用了高效的 Householder 变换以及随机置换对数据样本以及分类标签进行加密，安全性分析表明，该方案在诚实好奇模型下是安全的。性能评估表明方案可以节省数据提供者的计算开销。美中不足的是，协议设计了支持向量机模型的训练外包。使用云服务器对其他机器学习模型的外包训练也是一个非常有趣的问题。

## 参考文献


[1] Gantz, J. and Reinsel, D. (2011) Extracting Value from Chaos. *IDC Iview*, **1142**, 1-12.

[2] Lyu, Z., Li, J., Dong, C., Wang, Y., Li, H. and Xu, Z. (2021) DeepPTP: A Deep Pedestrian Trajectory Prediction Model for Traffic Intersection. *KSII Transactions on Internet and Information Systems* (*TIIS*), **15**, 2321-2338. https://doi.org/10.3837/tiis.2021.07.002

[3] Lyu, Z., Li, J., Li, H., Xu, Z. and Wang, Y. (2021) Blind Travel Prediction Based on Obstacle Avoidance in Indoor Scene. *Wireless Communications and Mobile Computing*, **2021**, Article ID: 5536386. https://doi.org/10.1155/2021/5536386

[4] Luo, C., Zhang, K., Salinas, S. and Li, P. (2017) Secfact: Secure Large-Scale QR and LU Factorizations. *IEEE Transactions on Big Data*, **7**, 796-807, https://doi.org/10.1109/TBDATA.2017.2782809

[5] Lv, Z., Li, J., Dong, C. and Xu, Z. (2021) DeepSTF: A Deep Spatial-Temporal Forecast Model of Taxi Flow. *The Computer Journal*, Article No. bxab178. https://doi.org/10.1093/comjnl/bxab178

[6] Lv, Z., Li, J., Dong, C., Li, H. and Xu, Z. (2021) Deep Learning in the COVID-19 Epidemic: A Deep Model for Urban Traffic Revitalization Index. *Data & Knowledge Engineering*, **135**, Article ID: 101912. https://doi.org/10.1016/j.datak.2021.101912

[7] Devikanniga, D., Ramu, A. and Haldorai, A. (2020) Efficient Diagnosis of Liver Disease Using Support Vector Machine Optimized with Crows Search Algorithm. *EAI Endorsed Transactions on Energy Web*, **7**, Article No. e10.

[8] Farhadian, M., Shokouhi, P. and Torkzaban, P. (2020) A Decision Support System Based on Support Vector Machine for Diagnosis of Periodontal Disease. *BMC Research Notes*, **13**, Article No. 337. https://doi.org/10.1186/s13104-020-05180-5

[9] Sivaram, M., Lydia, E.L., Pustokhina, I.V., Alexandrovich Pustokhin, D., Elhoseny, M., Prasad Joshi, G., *et al*. (2020) An Optimal Least Square Support Vector Machine Based Earnings Prediction of Blockchain Financial Products. *IEEE Access*, **8**, 120321-120330. https://doi.org/10.1109/ACCESS.2020.3005808

[10] Byvatov, E. and Schneider, G. (2003) Support Vector Machine Applications in Bioinformatics. *Applied Bioinformatics*, **2**, 67-77.

[11] Wang, J., Wu, L., Wang, H., Choo, K.-K.R. and He, D. (2020) An Efficient and Privacy-Preserving Outsourced Support Vector Machine Training for Internet of Medical Things. *IEEE Internet of Things Journal*, **8**, 458-473. https://doi.org/10.1109/JIOT.2020.3004231

[12] Liu, X., Deng, R.H., Choo, K.K.R. and Yang, Y. (2018) Privacy-Preserving Outsourced Support Vector Machine Design for Secure Drug Discovery. *IEEE Transactions on Cloud Computing*, **8**, 610-622. https://doi.org/10.1109/TCC.2018.2799219

[13] Lin, K.-P. and Chen, M.-S. (2010) Privacy-Preserving Outsourcing Support Vector Machines with Random Transformation. *Proceedings of the* 16*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*,

Washington DC, 25-28 July 2010, 363-372. https://doi.org/10.1145/1835804.1835852

[14] González-Serrano, F.J., Navia-Vázquez, Á. and Amor-Martín, A. (2017) Training Support Vector Machines with Privacy-Protected Data. *Pattern Recognition*, **72**, 93-107. https://doi.org/10.1016/j.patcog.2017.06.016

[15] Laur, S., Lipmaa, H. and Mielikäinen, T. (2006) Cryptographically Private Support Vector Machines. *Proceedings of the* 12*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, 20-23 August 2006, 618-624. https://doi.org/10.1145/1150402.1150477

[16] Omer, M.Z., Gao, H. and Sayed, F. (2016) Privacy Preserving in Distributed SVM Data Mining on Vertical Partitioned Data. 2016 3*rd International Conference on Soft Computing & Machine Intelligence* (*ISCMI*), Dubai, 23-25 November 2016, 84-89. https://doi.org/10.1109/ISCMI.2016.40

[17] Shen, M., Tang, X., Zhu, L., Du, X. and Guizani, M. (2019) Privacy-Preserving Support Vector Machine Training over Blockchain-Based Encrypted IoT Data in Smart Cities. *IEEE Internet of Things Journal*, **6**, 7702-7712. https://doi.org/10.1109/JIOT.2019.2901840

[18] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, **20**, 273-297. https://doi.org/10.1007/BF00994018

[19] Knuth, D.E. (1997) The Art of Computer Programming. Pearson Education, Loondon.

[20] Platt, J. (1998) Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Microsoft, Redmond.

[21] Chang, C.C. (2001) LIBSVM: A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/cjlin/libsvm

[22] Maekawa, T., Kawamura, A., Nakachi, T. and Kiya, H. (2019) Privacy-Preserving Support Vector Machine Computing Using Random Unitary Transformation. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, **102**, 1849-1855. https://doi.org/10.1587/transfun.E102.A.1849