硕士学位论文

基于安全多方计算的隐私保护支持向量机算法研究

RESEARCH ON PRIVACY PRESERVING SUPPORT VECTOR MACHINE ALGORITHM BASED ON SECURE MULTI-PARTY COMPUTATION

孙文礼

哈尔滨工业大学

2018 年 12

国内图书分类号：TP391.4　　学校代码：10213
国际图书分类号：621.3　　密级：公开

工学硕士学位论文

基于安全多方计算的隐私保护支持向量机算法研究

硕士研究生　孙文礼
导　　师　蒋琳助理教授
申请学位　工学硕士
学　　科　计算机科学与技术
所在单位　哈尔滨工业大学（深圳）
答辩日期　2018 年 12 月
授予学位单位　哈尔滨工业大学

Classified Index: TP391.4
U.D.C: 621.3

A dissertation submitted in partial fulfillment of
the requirements for the academic degree of
Master of Engineering

RESEARCH ON PRIVACY PRESERVING SUPPORT VECTOR MACHINE ALGORITHM BASED ON SECURE MULTI-PARTY COMPUTATION

Candidate: Wenli Sun
Supervisor: Asst. Prof. Lin Jiang
Academic Degree Applied for: Master's Degree of Engineering
Specialty: Computer Science and Technology
Affiliation: Harbin Institute of Technology, Shenzhen
Date of Defence: December, 2018
Degree-Conferring-Institution: Harbin Institute of Technology

摘 要

数据挖掘是指从海量的数据中借助算法查找数据中潜在信息的过程。为了提高数据挖掘的准确性一方面需要对算法进行改进，另一方面需要在大量的数据上做数据挖掘，而这些数据一般来源于不同的单位或用户。由于本地存储和计算资源的限制，随着云计算的发展，越来越多的用户选择把数据上传到云上实现存储外包和计算外包。云是不完全可信的第三方，会使得用户对自己数据的拥有权和控制权分离，进而导致数据隐私信息泄露的风险产生。另外数据挖掘是一把“双刃剑”，直接在具有隐私信息的数据上做数据挖掘也会导致数据隐私信息泄露的风险产生。本文基于安全多方计算，提出了一种利用支持向量机算法在多用户的加密的数据上做具有隐私保护功能的数据挖掘方案。

针对整数域上多用户加密数据的计算问题，本文提出了具有存储外包和计算外包功能的双云框架模型。设计了支持在多密钥加密的整数域上做加法和乘法计算的同态加和同态乘协议。该协议首先基于“盲化”技术对密文数据进行盲化处理，然后通过两个云之间的交互计算将多密钥加密的数据转为同一个单密钥加密的数据，最后利用单密钥同态协议的性质完成在密文上做加和乘的计算。

针对有理数域上多用户加密数据的计算问题，本文沿用了在整数域上设计的双云框架。有理数包含整数和小数，而小数的加解密计算和存储不同于整数，本文首先通过将小数转为分数，然后分别对分子和分母利用整数域上设计的同态加和同态乘的协议进行计算，最后在多密钥加密的有理数域上完成加法和乘法的计算。

基于本文设计的支持在密文上做加法和乘法的同态协议，可以利用支持向量机算法在多密钥加密的有理数域和整数域上做数据挖掘。在半诚实安全模型中可以证明在保证数据挖掘准确性的前提下，本文设计的算法可以保护用户的数据隐私、中间计算结果的隐私、分类模型的隐私和最后分类预测结果的隐私。本文基于设计的算法，搭建了利用支持向量机算法做具有隐私保护功能的数据挖掘系统，并在医疗环境中进行了应用示范。

关键词：安全多方计算；支持向量机；多密钥；隐私保护

Abstract

Data mining refers to the process of searching for hidden information from massive data. In order to improve the accuracy of data mining, on the one hand, the algorithm needs to be improved. On the other hand, data mining needs to be done on a large amount of data, which is generally derived from different units or users. Due to the limitations of local storage and computing, with the development of cloud computing, more and more users choose to upload data to the cloud to achieve storage outsourcing and computing outsourcing. The cloud is a third party that is not completely trusted. It will cause the user to separate the ownership and control of their own data, which in turn leads to the risk of data privacy information leakage. In addition, data mining is a "double-edged sword". Data mining directly on data with private information can also lead to the risk of data privacy information leakage. Based on secure multi-party computation, this paper proposes a data mining scheme with privacy preserving function on multi-user encrypted data by using support vector machine algorithm.

Aiming at the computational problem of multi-user encrypted data in integer domain, this paper proposes a dual cloud framework model with storage outsourcing and computing outsourcing. A homomorphic addition and homomorphic multiplication protocol supporting addition and multiplication on the integer domain of multi-key encryption is designed. The protocol first blinds the ciphertext data based on the "blind" technology, and then converts the multi-key encrypted data into the same single-key encrypted data through the interaction calculation between the two clouds, and finally，the calculation of adding and multiplying on the ciphertext is completed by using the properties of the single-key homomorphic protocol.

Aiming at the computational problem of multi-user encrypted data on the rational number domain, this paper follows the dual-cloud framework designed on the integer domain. Rational numbers contain integers and decimals, and the encryption and decryption calculation and storage of decimals are different from integers. This paper first calculates the numerator and denominator by using the homomorphic addition and homomorphic multiplication protocol designed on the integer domain. Finally, the addition and multiplication calculations are performed on the rational number field of multi-key encryption.

Based on the homomorphic protocol designed to support addition and multiplication on ciphertext, the support vector machine algorithm can be used for data mining on the rational number field and integer domain of multi-key encryption. In the semi-honest security model, it can be proved that under the premise of ensuring the

accuracy of data mining, the algorithm designed in this paper can protect the user's data privacy, the privacy of intermediate calculation results, the privacy of the classification model and the privacy of the final classification prediction results. Based on the algorithm designed in this paper, we built a data mining system with privacy preserving function using the support vector machine algorithm, and demonstrated the application in the medical environment.

Keywords: secure multi-party computation, support vector machine, multi-key, privacy-preserving

目　录

第1章 绪 论

1.1 课题研究背景与意义

随着科技的发展，越来越多的结构化和非结构化的数据被产生出来。为了更好地利用数据，数据挖掘技术也被越来越多的学者研究。数据挖掘一般是通过机器学习的方法对数据进行处理、建模进而根据训练的模型做分类和回归的预测，近年来数据挖掘在金融领域、医疗领域、零售和电商领域、电信领域、交通领域等发挥着越来越重要的作用。随着机器学习、人工智能等技术的研究和应用使得数据挖掘的能力越来越强大，与此同时也给个人敏感信息的安全带来了严重的威胁。在含有个人或企业的敏感信息上进行数据上挖掘，会使得这些敏感信息被分析和泄露出来，从而给个人和企业带来严重的安全威胁。所以如何在不泄露数据隐私的前提下提高挖掘数据的效率是当前数据挖掘领域的关键问题，隐私保护数据挖掘就是为了解决这个问题被提出来的。隐私保护数据挖掘需要在数据产生、存储、发布和挖掘的整个生命周期中保证数据隐私信息不被泄露。

机器学习是隐私保护数据挖掘技术的关键，优秀的机器学习算法能够显著地提高数据挖掘的性能。随着数据挖掘技术的应用范围越来越广，应用层次越来越高，对机器学习算法的要求也越来越严格。目前针对隐私保护数据挖掘的机器学习算法有支持向量机（Support Vector Machine，SVM）、K-近邻、K-means、ID-3、C4.5 和深度学习等，与其他机器学习算法相比，支持向量机算法基于 VC 维理论和结构最小化原则能有效地避免维数灾难和过拟合的问题，并且具有求解速度快、预测精度高等优点，另外还具有较好的推广能力和解的全局最优性。更加重要的是支持向量机算法不仅可以用来解决分类问题，也可以用于求解回归问题。因此本课题主要针对具有隐私保护功能的支持向量机算法进行研究。

在数据的整个生命周期中为了防止隐私信息泄露，需要对数据进行处理。一般的方法是对原始数据进行匿名化或扰动处理，通过在数据中加入噪声，达到对数据脱敏的目的。不过由于数据中有噪声，会影响接下来利用机器学习算法进行训练和预测的过程，使得最终计算出的模型在对预测数据进行处理时会出现很大的偏差。而当利用加密的方法对数据进行处理时可以保证分类或回归预测结果的高准确性，而且加密后得到的密文看起来就是一些无意义的数或字符串，另外加密算法本身可以保证任何用户不能从密文中推出原始明文的信息。因此本课题在研究如何对隐私数据进行保护时采用了加密的方法对数据中的隐私信息进行处理。而为了使得机器学习算法能够在加密后的密文上进行数据挖掘，需要进一步针对机器学习算法的计算过程设计相应的密码协议。本课题主要研究的是支持向量机算

法，通过研究可以发现在利用支持向量机算法进行数据挖掘时主要的计算是点积又称内积。而点积可以分为乘法和加法的计算。因此可以通过设计能够在密文上做加法和乘法计算的密码协议来完成支持向量机算法在密文上进行数据挖掘的目的。经过调研发现通过安全多方计算可以在保护数据隐私的前提下在密文上做计算。因此本课题主要研究的是基于安全多方计算的具有隐私保护功能的支持向量机算法。

综上所述，基于本课题设计的方案可以集合多个用户或多个单位的数据，在不泄露各自数据隐私的前提下，借助云端的存储和计算能力进行数据挖掘。并且在数据挖掘的过程中可以在保证最后分类预测结果的准确性的前提下保护用户原始数据的隐私、中间计算结果的隐私、最后挖掘出的分类模型的隐私、以及分类预测结果的隐私。

1.2 国内外研究现状

本文在安全多方计算的基础上，提出了具有隐私保护功能的支持向量机分类算法，下面主要对安全多方计算和隐私保护数据挖掘的国内外研究现状进行分析。

1.2.1 安全多方计算研究现状

安全多方计算主要用来解决互不信任的用户或参与方之间如何合作共同计算一个函数的问题，并且需要保证在计算完成后，每个用户只能得到最后的计算结果以及自己的输入，而不能知道其他用户的输入。一般可以通过混淆电路的方法[1]、基于秘密分享的方法[2]以及基于同态加密的方法[3]解决多个参与方之间安全计算的问题。

同态加密的概念最早是由 Rivest 等人于 1978 年提出[4]。同一年他们提出了具有乘法同态性质的 RSA 加密算法[5]。利用混淆电路求解安全多方计算的方法最早是由 Yao 提出来的，1986 年，Yao 基于混淆电路和茫然传输协议构造了通用的安全两方计算协议[1]。Yao 提出的混淆电路主要是为了解决两方的计算问题，其中一方是电路构造方，负责生成混淆电路，另一方是电路计算方，负责对电路生成方构造的电路进行计算。1987 年，Goldreich 等人[6]基于零知识证明和承诺协议提出了一种将一个在半诚实模型下安全的协议通过 GMW 编译器转为一个在恶意模型下安全协议的方案。不过该方案的效率比较低，仅具有理论意义。1999 年，Paillier[7]针对同态协议进行研究，提出了具有加法同态性质的 Paillier 加密算法。2005 年，Boneh 等人[8]针对同态加密协议首次提出了可以执行多次加法同态运算和一次乘法同态运算的方案。2007 年，Lindell 等人[9]基于 Yao 的混淆电路和 Cut-and-Choose

技术，研究了如何在恶意模型下构建安全的两方参与的通用计算协议。安全多方计算协议除了可以用来解决多方参与者之间的计算，也可以推广应用到其他领域，例如 Zhang 等人[10]基于同态加密协议设计了具有隐私保护功能的数据挖掘方案。2009 年，Gentry[11]首次提出了基于理想格的全同态加密协议，这是全同态加密协议研究的一个重大突破。2011 年，Kamara 等人[12]研究了在云环境下如何进行安全计算，并首先提出了将基于 Yao 混淆电路构造的安全多方计算协议中的计算外包给第三方服务器实现计算外包。2012 年，Brakerski 等人[13]基于环的 LWE 假设[14]提出了一个全同态加密方案，该方案中使用了密钥交换技术和模交换技术使得方案的效率要高于 Gentry 设计的全同态加密方案。同年 López-Alt 等人[15]设计了支持多密钥的全同态加密方案，该方案可以在多密钥加密的数据上做同态加和同态乘的计算，不过该方案的效率比较低满足不了实际的需求。随着移动互联网的发展，2013 年，Carter 等人[16]基于 Yao 混淆电路的安全多方计算协议，研究了在移动端如何安全计算外包的相关问题。同年，Gentry 等人[17]基于近似特征向量设计了一个无需密钥交换和模交换的全同态加密方案。2014 年，Jakobsen 等人[18]针对如何将基于秘密共享的安全多方计算协议外包到多个云服务器上进行计算的问题进行了研究。2016 年，Brakerski 等人[19]针对多密钥全同态加密算法进行了研究，并提出了一个基于格的支持无限次操作的全同态加密协议，不过该协议的效率依然不高。2017 年，Furukawa 等人[20]针对只有一个恶意敌手，诚实者占大多数的环境中提出了用于三方之间的通用安全多方计算协议。同年，Alagic 等人[21]针对量子计算和同态加密协议进行了研究，提出了具有验证功能的可用于量子计算的全同态加密协议。2018 年，Cheon 等人[22]针对全同态加密协议进行了研究，并基于 Gentry 提出的全同态加密协议中的 bootstrapping 技术[11]，设计了一种新的用于密文刷新的方案。

1.2.2 隐私保护数据挖掘研究现状

隐私保护数据挖掘目前主要是通过扰动、差分、加密的方法来保护数据的隐私信息[23]，不过扰动和差分的方法会在数据中加入噪声，使得最终进行分类或回归预测时的准确性不高[24]。而通过加密的方法对数据进行处理，后续的数据挖掘都是在密文上进行，具体的计算也是针对密文。而这可以通过安全多方计算实现，使得在密文上计算的结果就是对明文数据进行同样的计算后加密的密文结果[25]。不过采用加密的方法通常是在密文上进行数据挖掘，会使得数据挖掘的时间比较长，但是这种方法不影响做分类或回归预测时的准确性，使用这种方法和在明文数据上用同样的机器学习算法进行预测的准确性是一样的，另外随着云计算的发

展，可以借助云端的存储和计算能力进行数据挖掘，从而可以在很大程度上提高算法的效率和性能。

除了上述保护数据隐私的方法不同，各种方法所针对的数据分布方式也不尽相同。另外有的隐私保护数据挖掘方法借助了云端的计算能力实现了计算外包，有的借助了云端的存储和计算能力，同时实现了计算外包和存储外包。接下来将根据数据的分布方式、是否是计算外包、以及是否是存储和计算外包这三个方面来介绍近年来隐私保护数据挖掘及其相关理论的发展概况。

2000 年，Lindell 等人[25]在水平分布的数据集上，设计了安全对数计算协议并应用到了 ID3 这种机器学习算法中，使得 ID3 可以在加密后的水平数据集上做数据挖掘。2003 年，Vaidya 等人[26]针对 K-means 算法设计了相应的同态加密协议，使得 K-means 可以在垂直分布的数据集上做具有隐私保护功能的数据挖掘。2005 年，Lin 等人[27]设计了在任意分布的数据上利用 EM 算法做具有隐私保护功能的数据挖掘方案。2006 年，Yu 等人[28]在水平分布的数据集上，针对非线性核函数设计了可以利用支持向量机做具有隐私保护功能的数据挖掘方案。在同一年他们同样在不泄露敏感数据的前提下，基于垂直分布的数据集提出了具有隐私保护功能的 SVM 分类算法[29]。2008 年，Vaiday 等人[30]也提出了具有隐私保护功能的 SVM 分类算法，他们的方案不仅可以应用于水平分布和垂直分布的数据集，更加重要的是可以应用到任意分布的数据集上。同年，Samet 等人[31]也设计了可以在水平分布的数据集上利用 ID3 做具有隐私保护功能的数据挖掘方案。另外，Jaideep 等人[32]针对贝叶斯算法设计了相应的具有隐私保护功能的数据挖掘方案。2010 年，Hu 等人[33]针对任意分布的数据集也设计了具有隐私保护数据挖掘功能的 SVM 算法，而且他们设计的分类方案即使公开也不会泄露数据的隐私信息。2011 年，Skarkala 等人[34]在水平分布的多方数据集上设计了具有隐私保护功能的贝叶斯算法。2012 年，Lory 等人[35]基于切比雪夫多项式设计了可以安全求解对数的协议。

随着信息技术的发展，人类已经慢慢地进入了大数据的时代，各种结构类型和非结构类型的数据也在急剧地快速增长，此时为了解决用户在本地对大规模的数据进行数据挖掘效率比较慢的问题，一些云服务商开始对外提供计算来帮助用户做一些在本地不能做或做起来比较慢的计算，计算外包就是在这种背景下被提出来了。在分布式数据挖掘中，各方用户都把自己需要计算的部分外包到云端，借助云端的强大计算能力进行计算。2011 年，Kamara 等人[12]基于云服务器设计了一个安全多方计算外包的框架。2012 年，Ma 等人[36]设计了一个可以用于求解多模指数计算的安全外包协议。2013 年，Lei 等人[37]提出了可用于矩阵计算的安全外包协议。2015 年，任艳丽等人[38]设计了一个具有隐私保护功能的可以验证的多元多项式计算外包的方案。同年，刘晓燕[39]设计了利用 K-means 做聚类的具有隐

私保护功能的计算外包方案。2016 年，Zhang 等人[40]基于线性矩阵方程组的研究提出了一个相应的计算外包方案。同年，孙茂华等人[41]设计了一个可以安全求并集的计算外包方案。2017 年，张兴兰等人[42]针对如何对线性问题进行安全外包计算进行了研究。同年，蔡建兴等人[43]设计了一个用于求解线性问题的计算外包方案。

在计算外包的隐私保护数据挖掘方法中，因为用户的数据都是分布式地存储在本地，只是借助了云端的计算能力来完成数据挖掘的计算过程，此时需要云端和用户之间传输大量的数据来进行交互，这样会带来通信量过大的问题。为了解决这个问题，在计算外包的基础上有学者提出了存储外包的方案。2013 年，Peter 等人[44]设计了一个可以在多密钥加密的数据上做计算的安全多方计算协议，在这个协议中用户除了需要做基本的加密和解密操作外，不需要做其他额外的计算，而且用户的数据都是加密后存储在云端的，这样也在计算外包的同时实现了存储外包。2014 年，Liu 等人[45]基于门陷加密的方法，设计了支持单个用户存储和计算外包的具有隐私保护功能的 K-means 数据挖掘方案。2015 年，Liu 等人[46]基于全同态加密协议，针对垂直分布的数据集设计了可以利用 SVM 做隐私保护数据挖掘的方案，在他们的方案中也实现了存储外包和计算外包。不过用户除了做基本的加密和解密操作外还是需要参与做一些其他的计算。而且他们的方案中由于采用了全同态的加密方案，所以效率相对比较低。同年，靳亚宾[47]设计了支持存储和计算外包的隐私保护 K-means 聚类方案。2016 年，Li 等人[48]基于对称密钥设计了同态加密协议，并应用到了关联规则算法中做具有隐私保护功能的数据挖掘。2017 年，Zhang 等人[49]基于整数向量加密协议，提出了一个支持存储外包和计算外包的利用 SVM 做隐私保护数据挖掘的方案。在她们设计的方案中由于数据是用用户各自的公钥加密的，因此需要用户协商出一个密钥转换矩阵来实现密文转换的目的。同年 Zhang 等人[10]基于 BCP 加密协议[50]，针对垂直分布的数据集设计了具有隐私保护功能的 SVM 算法。而 Li 等人[51]基于 BCP 加密协议[50]和全同态加密协议[11]设计了一个可以在多密钥加密的数据集上利用深度学习做隐私保护数据挖掘的方案。

1.3 主要研究内容与组织结构

接下来主要介绍本文的研究内容和具体的组织结构。

1.3.1 主要研究内容

通过对利用支持向量机算法和其他机器学习算法做隐私保护数据挖掘的研究现状进行调研分析发现目前的隐私保护方案一般都是通过对数据进行扰动或加密

来实现的。通过扰动的方法来保护数据的隐私会在数据中添加噪声，使得最后挖掘出的结果不准确。而通过加密的方法来保护数据的隐私则不会影响最后的挖掘结果。因此为了不影响挖掘结果的准确性，本文主要研究如何通过加密的方法来利用支持向量机算法做具有隐私保护功能的数据挖掘。为了在加密后的隐私数据上做数据挖掘可以基于安全多方计算设计相应的数据挖掘算法。

安全多方计算主要包括 Yao 的混淆电路、秘密共享和同态加密这三种方法。不过 Yao 的混淆电路方法和秘密共享的方法主要针对的是单密钥的场景。当通过机器学习算法做数据挖掘时，为了提高挖掘的准确性，需要在海量的数据上训练一个机器学习模型。当海量的数据分别来自多个用户或多个单位时，各个用户的公钥是不同的，因此为了在多密钥的场景中做具有隐私保护功能的数据挖掘，本文主要研究了如何基于同态加密协议设计具有隐私保护功能的支持向量机算法。在保护数据隐私的前提下当利用安全多方计算的方法对密文数据进行计算时会因为数据的分布方式不同、数据类型的不同和参与方个数的不同而出现无法进行数据挖掘的问题。为了解决上述问题，本课题主要针对以下几个方面进行具体研究：

（1）研究了如何在分布式存储的数据上做具有隐私保护功能的数据挖掘。当用户的数据分布方式是水平分布、垂直分布或任意分布时，为了在这些分布式存储的数据上做数据挖掘，本文设计了具有存储外包和计算外包功能的方案，通过将分布式存储的多方数据加密后上传到云上，利用云的存储和计算能力做具有隐私保护功能的数据挖掘来解决数据分布式存储的问题。

（2）研究了如何基于多个用户的数据设计支持向量机算法进行数据挖掘。为了在保护数据隐私的前提下保证数据挖掘结果的准确性，本文研究了如何基于多方的数据训练支持向量机并进行分类预测。因为单个用户的数据比较少，因此需要在多方数据的基础上做数据挖掘，而因为每个用户的公钥都是不同的，无法直接利用安全多方计算的方法进行计算，所以本文设计了一个支持在多个不同公钥加密的数据上做计算的安全多方计算协议，从而达到通过支持向量机算法做数据挖掘的目的。

（3）研究了如何在整数域和有理数域上利用支持向量机进行数据挖掘。因为有理数中小数的加解密计算和存储是不同于整数的。因此除了需要设计支持在多密钥加密的整数域上做数据挖掘的方案，还需要研究如何在有理数域上进行数据挖掘。本文通过将有理数域上的计算和存储转为整数域上的计算和存储，进而通过设计的整数域上的安全多方计算协议来解决有理数域上的计算和存储问题，从而完成在整数域和有理数域上利用支持向量机算法做具有隐私保护功能的数据挖

掘。

（4）研究了如何在半诚实的安全模型下保护数据的隐私信息。假设存在一个外部敌手，该敌手除了可以窃听参与方之间的通信，还可以和参与方合谋来猜测用户的数据隐私。本文在敌手存在的前提下设计了算法并且可以通过现实和理想模型证明该算法可以保证用户的数据隐私、中间计算结果的隐私、分类模型的隐私和分类预测结果的隐私。

1.3.2 组织结构

本论文的组织结构主要分为四章，每章的具体内容如下：

第 1 章绪论。首先说明了课题的来源、研究背景和意义。然后从安全多方计算和隐私保护数据挖掘这两方面介绍了本课题的国内外研究现状。安全多方计算的国内外研究现状主要是从 Yao 的混淆电路、秘密共享和同态加密协议这三个方向进行了介绍。隐私保护数据挖掘的国内外研究现状主要是从分布式、计算外包、存储和计算外包这三个方面作了介绍。最后介绍了本文的主要研究内容和组织结构。

第 2 章安全多方计算与支持向量机算法。首先介绍了本课题在研究具有隐私保护功能的数据挖掘方案时对隐私数据进行保护和计算的安全多方计算协议，然后介绍了方案中在密文数据上做数据挖掘用到的支持向量机算法。最后以银行存储的数据为例介绍了本课题所研究的数据分布形式，分别是水平分布、垂直分布和任意分布。

第 3 章基于安全多方计算的支持向量机算法。首先介绍了本课题的安全需求、系统框架和安全模型，然后针对本课题设计的在水平分布、垂直分布和任意分布的整数域和有理数域上利用支持向量机做具有隐私保护功能的分类算法进行了具体的介绍，最后对算法的复杂度进行了分析，并在半诚实的安全模型下对算法进行了安全证明。

第 4 章隐私保护支持向量机算法的实现与分析。首先介绍了对第 3 章设计的算法进行实现时所用到的开发环境、训练和测试的数据集。然后基于实验的结果对系统的性能进行了分析。最后对本课题设计的算法从功能和性能上与其他利用支持向量机做隐私保护数据挖掘的方案进行了比较分析。

第2章 安全多方计算与支持向量机算法

为了能够在保护数据隐私的前提下做数据挖掘，本课题采用了密码学中的加密方法来实现对隐私数据的保护。为了进行数据挖掘，本课题采用了支持向量机的机器学习算法，该算法不仅可以做分类，还能够进行回归。另外不仅可以在线性可分的数据集上进行处理，也能够利用核函数在线性不可分的数据集上进行分类或回归的处理。而为了能够在密文上利用 SVM 做隐私保护的数据挖掘，本课题采用了密码学中的安全多方计算的方法，来实现 SVM 在密文上的计算，在保证计算结果正确性的同时，可以保护数据的隐私信息。

2.1 安全多方计算

接下来首先介绍安全多方计算的定义，然后介绍安全模型，最后对本文设计的算法中用到的安全多方计算中的同态加密协议进行介绍。

2.1.1 安全多方计算定义

安全多方计算[1]是来自于姚期智的百万富翁问题，具体是指两个富翁在不泄露各自财富的情况下，比较谁更富有的问题。通用的安全多方计算是指两个或两个以上的参与方，在不泄露各自的输入的情况下，共同计算并得到一个输出结果。在计算的整个过程期间，每个参与方只能知道自己的输入和最后的输出，而不能知道其他参与方的输入信息。安全多方计算的具体定义如下：

定义 2.1（安全多方计算）：网络中存在n个互不信任的参与方，他们分别具有输入$x_1,\cdots,x_n$，然后准备通过一个安全多方计算协议π来协作计算一个函数$f(x_1,\cdots,x_n)=(y_1,\cdots,y_n)$，当完成计算后每个参与方除了自己的输入$x_i$和输出$y_i$外不能获得其他任何的额外信息。

一个安全的多方计算协议应该满足以下几个特性。隐私性指的是每个参与方从协议的计算开始到计算结束仅仅只能知道自己的输入和最后的输出结果外。正确性指的是经过安全多方计算协议计算出的结果应该是正确无误的。输入独立性指的是各个参与方之间的输入是互不影响的相互独立的。公平性指的是最后计算出的结果要么所有的参与方都可以知道，要么都不能知道。

2.1.2 安全多方计算隐私安全模型

在分析安全多方计算协议的安全性时，通常会假设在外部存在一个敌手A，敌手可以可以控制一部分参与者，和这些参与者合谋来猜测其他参与者的输入信

息。根据敌手 A 的攻击能力不同，可以把安全模型分为半诚实安全模型和恶意安全模型。与敌手合谋的参与方在半诚实安全模型中会继续按照协议的规则执行协议，不会破坏协议的规则和流程。但是他们会记录计算的中间结果，并尝试从中推测出其他参与者的输入信息。在恶意安全模型中，和敌手合谋的参与者可以更改协议的流程和规则，并能够篡改或伪造数据，敌手的目标包括阻止其他参与者获得正确的输出，并尝试破解其他参与方的输入。

证明协议的安全性能够通过理想（Ideal）和现实（Real）模型来证明。在 Ideal 模型中有一个可信任的第三方（Trusted Third Party，TTP）。所有的参与者都可以通过安全的信道把自己的秘密输入发送给这个可信的第三方，然后这个可信的第三方基于收到的数据完成一个函数的计算，并把最后的计算结果通过安全的信道传递给所有的参与方。在 Real 模型中是没有这个可信的第三方，所有的参与者通过执行一个交互协议来完成对目标函数的计算。相比 Ideal 模型，这种 Real 模型更加符合实际的情况。当 Real 模型中存在一个敌手时，可以在 Ideal 模型中模拟这个敌手在 Real 模型中的能力和行为，最后如果敌手在 Ideal 模型中的输入输出和在 Real 模型中的输入输出不可区分，那么就可以说 Ideal 模型完全模拟了 Real 模型，进而证明了在 Real 模型中所执行协议的安全性。令 $IDEAL_{f,S,Z}(\lambda,x)$ 表示在 Ideal 模型 Z 中，输入为 x，存在一个外部敌手 S，安全参数为 λ，在 Ideal 模型中计算一个目标函数 f 所得到的输出结果；令 $REAL_{f,A,Z}(\lambda,x)$ 表示在 Real 模型 Z 中，输入为 x，存在一个敌手 A，安全参数为 λ，计算目标函数 f 所得到的输出。令 $X \equiv_c Y$ 表示 X 和 Y 的计算不可区分。那么一个安全多方计算协议的安全性定义如下所示：

定义 2.2（安全多方计算安全性）：假设 π 为安全多方计算协议，需要计算的函数是 $f:\{0,1\}^* \times \cdots \times \{0,1\}^* \to \{0,1\}^* \times \cdots \times \{0,1\}^*$。如果在 Real 模型中存在一个非均匀的概率多项式时间敌手 S，并且在 Ideal 模型中也有一个非均匀的概率多项式时间敌手 A 与之对应，满足公式 2-1

$$\left\{IDEAL_{f,S,Z}(\lambda,x)\right\} \equiv_c \left\{REAL_{f,A,Z}(\lambda,x)\right\} \tag{2-1}$$

那么就认为安全多方计算协议 π 能够安全地计算目标函数 f。

2.1.3 同态加密算法

同态加密算法是指利用同态加密协议对数据进行加密后，可以在密文上进行代数运算，而得到的结果就是对应明文进行加或乘计算后的密文结果。

如果一个加密算法满足下面的条件，那么就称该加密算法是具有同态加法性

质的加密算法：首先利用该算法的密钥生成算法 $Gen(1^n)$ 生成一对公私钥 (pk, sk)，然后利用公钥 pk 对明文空间中的两个明文 $m_1, m_2 \in M$ 进行加密（其中 M 是明文消息空间），并得到两个密文 $c_1 = Enc_{pk}(m_1)$ 和 $c_2 = Enc_{pk}(m_2)$，且 $c_1, c_2 \in C$，C 是密文空间。对密文进行如下操作

$$c_1 \Theta c_2 = Enc_{pk}(m_1 \otimes m_2) \tag{2-2}$$

其中 Θ 为代数计算，如果 $\otimes$ 操作是加法操作，那么就称该算法是具有加法同态性质的加密算法，如果 $\otimes$ 操作是乘法操作，那么就称该算法是具有乘法同态性质的加密算法。如果一个加密算法既具有乘法同态性质，又具有加法同态那么就称该加密算法为全同态加密算法。如果一个算法只具有乘法同态性质或加法同态性质中的一种，那么就称该加密算法是半同态加密算法。接下来主要介绍两个具有加法同态性质的算法：一个是 Paillier 的加法同态算法[7]，以及在 Paillier 算法上改进的加法同态算法[52]。

Paillier 加密算法具有加法同态的性质，当利用同一个公钥 pk 分别对两个明文 $m_1, m_2 \in M$ 进行加密，可以得到对应的密文 $c_1 = Enc_{pk}(m_1)$ 和 $c_2 = Enc_{pk}(m_2)$，并且这两个密文满足：

$$c_1 \bullet c_2 = Enc_{pk}(m_1 + m_2) \tag{2-3}$$

该算法包括密钥生成算法、加密算法、解密算法，具体如下所示：

- 密钥生成算法 $Gen(1^k) \to (sk, pk)$：

其中 k 是安全参数，随机选择两个大素数 p, q 满足 $\gcd(pq, (p-1)(q-1)) = 1$，并计算 $n = pq$ 和 $\lambda = lcm(p-1, q-1)$。另外随机选择 $g \in Z^*_{N^2}$，满足 $\gcd(L(g^\lambda \bmod n^2), n) = 1$，其中函数 $L(x) = (x-1)/n$。此时公钥 $pk = (n, g)$，私钥 $sk = \lambda$。

- 加密算法 $Enc_{pk}(m) \to c$：

当利用公钥 $pk = (n, g)$ 对明文 m 进行加密时，需要随机选择一个随机数 r 且 $r < n$，则密文为：

$$c = g^m \bullet r^n \bmod n^2 \tag{2-4}$$

- 解密算法 $Dec_{sk}(c) \to m$：

可以利用私钥 $sk = \lambda$ 对密文做如下计算得到明文：

$$m = \left(L(c^\lambda \bmod n^2) / L(g^\lambda \bmod n^2) \right) \bmod n \tag{2-5}$$

Liu 等人对 Paillier 算法做了一些改进，使得一个密文除了可以利用用户的私钥解密外，还可以利用系统中的主密钥进行解密。他们的方案也具有加法同态性质。具体方案如下所示：

- 公共参数和主密钥生成算法 $Setup(k) \to (PP, MK)$

其中 k 是一个安全参数， PP 是公共参数， MK 是一个主密钥可以用于解密系统中用户的密文。随机选择两个大素数 p', q'，并计算 $p = 2p'+1$， $q = 2q'+1$，和 $N = pq$，其中 p 和 q 的长度是 k。另外再随机选择一个数 $g \in Z_{N^2}^*$。那么公共参数 $PP = (N, k, g)$，主密钥 $MK = lcm(p-1, q-1)$，其中 MK 需要分成两部分 k_1, k_2，并满足 $k_1 + k_2 \equiv 0 \bmod MK$ 和 $k_1 + k_2 \equiv 1 \bmod N^2$。

- 密钥生成算法 $KeyGen(PP) \to (pk, sk)$

首先随机选择一个数 $\alpha \in [1, N/4]$，然后计算 $h = g^{\alpha} \bmod N^2$，那么此时公钥 $pk = (N, g, h)$，私钥 $sk = \alpha$。

- 加密算法 $Enc_{pk}(m) \to c$

当利用公钥 $pk = (N, g, h)$ 对一个明文数据 m 加密时，需要先选一个随机数 $r \in [1, N/4]$，那么 m 对应的密文就可以表示成 $Enc_{pk}(m) \to (A, B)$，其中 $A = g^r \bmod N^2$， $B = h^r (1+mN) \bmod N^2$。

- 解密算法 $Dec_{sk}(c) \to m$

当利用用户的私钥 $sk = \alpha$ 对一个密文进行解密时，可以通过如下的计算得到明文数据

$$m = \left(\left(B / A^{sk} \right) - 1 \bmod N^2 \right) / N \tag{2-6}$$

当利用主密钥 k_1, k_2 进行解密时可以做如下的计算：

- 解密算法 $SDec1_{k_1}(c) \to m_1$， $SDec2_{k_2}(c, m_1) \to m$

首先利用 k_1 进行解密如下所示：

$$SDec1_{k_1}(c) = g^{rk_1 sk}(1 + mNk_1) \bmod N^2 = m_1 \tag{2-7}$$

然后利用 k_2 进行解密如下所示：

$$m_2 = g^{rk_2 sk}(1 + mNk_2) \bmod N^2 \tag{2-8}$$

$$SDec2_{k_2}(c, m_1) = \left(\left(m_1 \cdot m_2 - 1 \bmod N^2 \right) / N \right) \bmod N = m \tag{2-9}$$

另外当一个密文，是由多个用户公钥的乘积作为新的公钥加密得到时，例如用 n 个公钥的乘积作为一个新的密钥 pk_{Π}，也即 $pk_{\Pi} = (N, g, h_{\Pi} = g^{\alpha_1 + \alpha_2 + \cdots + \alpha_n})$，当密文 c 是由公钥 pk_{Π} 加密得到时，那么为了解密 $c = (A, B)$ 可以做如下的计算：

$$A^{\alpha_i} = g^{r\alpha_i} \bmod N^2 \tag{2-10}$$

接下来就可以在上面求解的基础上，通过做如下的计算解密得到 m

$$m = \left(\left(B / (\prod_1^n A^{\alpha_i}) \bmod N^2 - 1 \right) / N \right) \bmod N \tag{2-11}$$

这个改进的 Paillier 算法满足加法同态性，当利用同一个公钥 pk 对两个明文数

据 m_1, m_2 进行加密可以得到 c_1, c_2，也即 $c_1 = Enc_{pk}(m_1)$ 和 $c_2 = Enc_{pk}(m_2)$。当对这两个密文 c_1, c_2 做乘法计算时，可以得到下式（2-12）的结果：

$$c_1 \cdot c_2 = \left(g^{r_1+r_2} \bmod N^2, h^{r_1+r_2}\left(1+(m_1+m_2)N \bmod N^2\right)\right) = Enc_{pk}(m_1+m_2) \quad (2\text{-}12)$$

另外当对密文做 $N-1$ 乘法计算时，会得到下式（2-13）的结果：

$$\left(Enc_{pk}(m)\right)^{N-1} = Enc_{pk}(-m) \quad (2\text{-}13)$$

2.2 支持向量机算法

支持向量机算法是基于统计学理论所提出的一种机器学习算法，它能够较好地解决小样本、非线性、高维数据和局部极小点等问题。如图 2-1 所示为利用支持向量机算法对两类叶子数据进行二分类的示例，

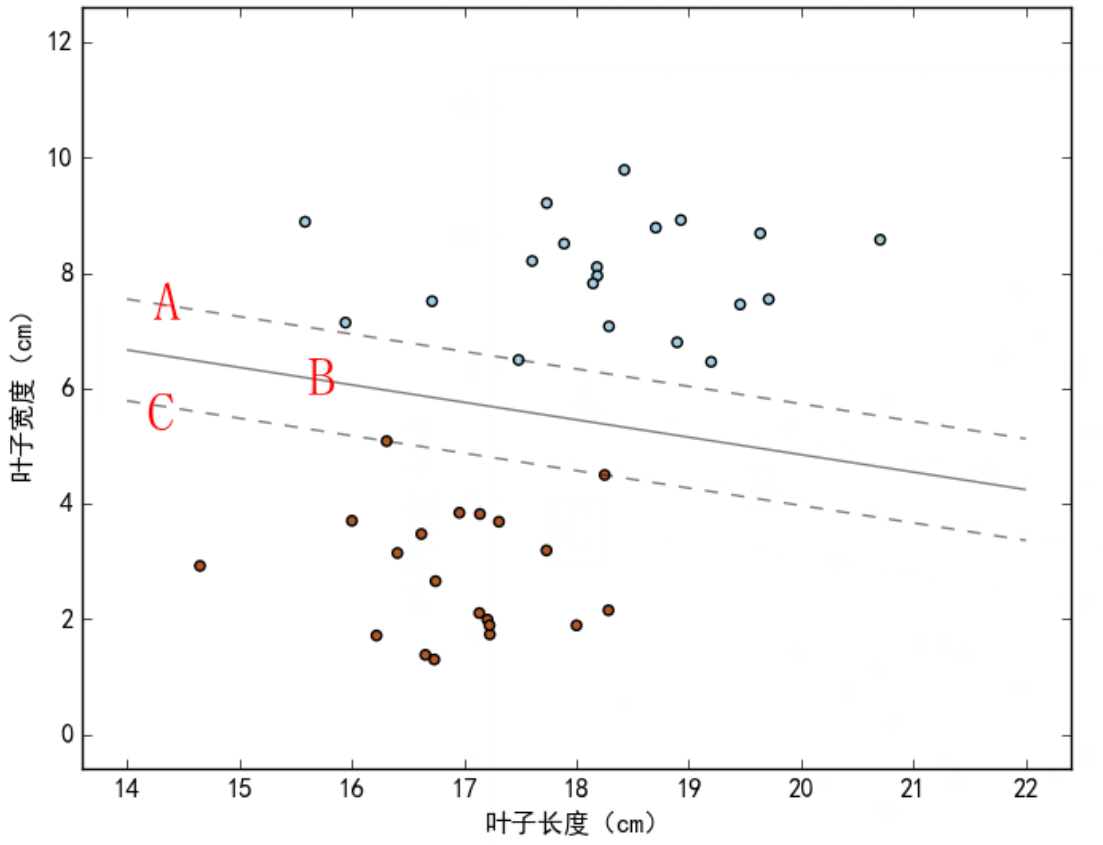

图 2-1 支持向量机分类示例

支持向量机算法的基本思想是通过在样本空间中寻找一个超平面来对样本进行分割，而寻找超平面的最大原则是使间隔最大化，这可以转为一个凸二次规划问题进行求解。当数据线性不可分时，可以使用一种映射将低纬空间中的训练数据映射到高维的空间，进而计算出一个最优划分超平面，使得不同类别的样本点线性可分。最优划分超平面是指使得不同类别的样本点之间距离间隔最大的那个

超平面。如图 2-1 所示为了划分图中的两个类别的样本需要找到一个最优的分割超平面 *B* 使得所有的样本点都尽可能的远离 *B* 也就是使图 2-1 中 *A* 和 *C* 的距离最大，而 *A* 和 *C* 上的样本点称为支持向量的样本点。

当数据线性不可分时，可以引入松弛变量来控制训练过程中的经验风险，从而使得训练得到的模型可以在满足分类预测准确的前提下，又具有很好的泛化能力。当样本的维数不断变大时，此时为了在这些样本上训练得到一个模型往往需要付出很高的计算代价，为了解决维数灾难的问题降低计算的代价目前有很多降低样本维数的方法，主要包括主分量分析、独立分量分析以及各种特征提取的方法。

通过支持向量机进行分类预测时，通过对图 2-1 进行观察可以发现在训练过程中求解最好的分割超平面 *B* 的主要计算是和样本上边界上的支持向量机的样本点有关，也即并不是所有的样本点都参与计算求解分割超平面，只需要找到边界上的支持向量机的样本点即可，然后最大化他们之间的距离。图 2-1 所示的是样本点线性可分的情况，当样本点线性不可分时，此时利用支持向量机进行分类预测可以借助核函数这个工具，把线性不可分的样本点映射到另一个高维度的空间中进行求解计算。而在高纬度空间中的计算也可以通过在原来维度中的样本点的点积计算来求解。

支持向量机的主要特点如下：

（1） 支持向量机以数学理论为基础，通过一步步推导计算克服了传统神经网络学习中靠经验和启发式的先验成分的特点，一切都有公式可寻。

（2） 支持向量机综合了经验风险最小化和结构风险最小化的原则来求目标函数，这样也避免了训练过程中的过拟合问题，提高了模型在新的测试数据集上的泛化能力。

（3） 支持向量机的分割超平面只与样本边界上的支持向量机的点相关，而这些点和全部的样本点相比是比较少的，因此利用支持向量机进行数据挖掘当数据比较多时，和其他机器学习算法相比，支持向量机可以在很大程度上减少计算的复杂性，降低维数灾难的影响。

（4） 针对不同的问题，模型选择的好坏影响支持向量机的分类性能。模型选择主要是指选择模型的判断因素和支持向量机中的参数的搜寻方法。针对不同的具体问题，需要选择符合问题实际情况的相应模型。

在利用支持向量机算法进行数据挖掘时主要包括两个步骤：训练和预测。训练主要是在已知类别的样本点上进行有监督的训练，预测是根据训练得到的模型在新的未知类别的数据上通过计算来预测数据的类别，而结果预测准确的高低和支持向量机的训练过程密切相关，支持向量机的训练方法主要有增量算法、分解

算法和块算法。

其中增量算法[53]是在利用 SVM 算法进行训练时，如果此时又有新的样本点数据添加进来，那么不需要把新的样本点数据和原来的数据混合起来从头开始训练，而是通过修改或删除新的样本点与原来的模型之间有关联的部分完成训练，其他没有关联的部分则不进行修改或删除。另外利用该算法进行训练时不是一次性就能完成的，而是通过不断增加数据集，不断进行迭代优化来完成的。

块算法[54]一般是借助一种迭代的策略将非支持的向量逐步减去，达到改变训练过程中数据集大小的目的。具体做法是将一个大的二次规划问题通过不断分解为一些小的二次规划问题进行求解，并把矩阵中的所有 Lagrange 乘数为零的行和列进行删除。块算法可以提高训练的速度，特别是当支持向量的样本数目远远小于总的训练样本数目时，利用这种方法在数据上进行训练可以很快的得到一个用于分类或回归预测的模型。

分解算法[55]是目前为止解决大规模二次规划问题效率最高的方法，也是用的最频繁的方法。分解算法的工作原理也是通过将一个二次规划问题不断分解为一系列小的二次规划子问题进行后续的迭代计算。利用该算法进行分类训练时需要将用于训练的数据集分成两部分，一部分是工作集，另一部分是非工作集，主要是在工作集上进行训练，因此需要使得在工作集尽可能少的情况下又能保证训练的效果，而这就需要一个好的划分方法来将训练数据集有效的划分两部分，使得工作集尽可能的少尽可能的包含所有的支持向量机的样本点。目前在解决实际问题中应用最多的划分方法是序列最小化（Sequential Minimal Optimization，SMO）划分算法[56]。

SMO 算法是一种启发式的算法，该算法的基本原则是假如所有变量的解都能达到最优化问题的 KKT（Karush Kuhn Tucker）要求，那么最优化问题的解就可以通过计算求得。因为 KKT 条件是求解最优化问题的充要前提。如果满足不了条件，那么就可以选择两个变量，单独基于这两个变量设计一个二次规划问题，而这个问题对应这两个变量的解，是近似原来问题的解。通过这样不断迭代地把原二次规划问题划分为一些小的二次规划问题进行求解，从而达到计算原二次规划问题的目的。

2.3 数据分布形式

在分布式数据挖掘中，参与方的数据都是分布式地存储在各个参与方本地，数据的分布方式也不是单一固定的，主要包括三种分布方式，分别是：水平分布、垂直分布和任意分布，接下来将针对这三种分布方式进行具体的介绍。

2.3.1 水平分布数据集

当数据是水平分布时，表示分布式存储的各方都拥有数据的全部属性，但是各方存储的具体记录可能是不一样的，如表 2-1 和表 2-2 所示：

表 2-1. 银行 A 的客户数据

客户	年龄	存款	薪水	贷款	透支
Cust.1	31	1000	2001	2012	Yes
Cust.2	23	2000	1230	2033	No
Cust.3	51	50000	5032	3032	Yes
Cust.4	49	1000	18909	4046	Yes
Cust.5	34	50000	2300	10824	No
Cust.6	35	10000	26000	20385	Yes

表 2-1 中存储的是银行 A 的客户数据，其中每条记录有六个属性分别是：客户、年龄、存款、薪水、贷款和透支，银行 B 存储的数据如表 2-2 所示。

表 2-2. 银行 B 的客户数据

客户	年龄	存款	薪水	贷款	透支
Cust.7	45	11320	2423	4332	No
Cust.8	28	276331	19972	1222	No
Cust.9	58	5852	52999	50982	Yes
Cust.10	43	15889	27220	70211	Yes

通过对表 2-1 和表 2-2 的观察可以发现，这两个银行存储的数据属性都一样，而记录条数不一样。表 2-1 和表 2-2 的数据就是水平分布的数据集。

2.3.2 垂直分布数据集

和水平分布的数据集不同的是，垂直分布的数据集中，各方的数据属性个数可能不一样，不过可以通过某一个属性，把两方的数据连接起来，如下表 2-3 所示：

表 2-3. 银行 A 的客户数据

客户	年龄	存款
Cust.1	31	1000
Cust.2	23	2000
Cust.3	51	50000
Cust.4	49	1000
Cust.5	34	50000
Cust.6	35	10000

和表 2-1 不同的是，表 2-3 中存储的数据属性只有三个，分别是客户、年龄和存款。

表 2-4. 银行 B 的客户数据

客户	薪水	贷款	透支
Cust.1	2001	2012	Yes
Cust.2	1230	2033	No
Cust.3	5032	3032	Yes
Cust.4	18909	4046	Yes
Cust.5	2300	10824	No
Cust.6	26000	20385	Yes

通过对表 2-4 进行观察可以发现，表 2-4 中每条记录有四个属性，分别是客户、薪水、贷款和透支。表 2-3 和表 2-4 是分别存储在银行 A 和银行 B，两个表中的记录个数是一样的，不过数据属性是不同，但是可以通过表中的“客户”这个属性，把两张表连接起来。

2.3.3 任意分布数据集

和水平分布、垂直分布不同的是，在任意分布的数据集中，分布式存储在各方的数据集中每条记录的具体属性值可能会有缺省。如表 2-5 和表 2-6 所示：

表 2-5. 银行 A 的客户数据

客户	年龄	存款	薪水	贷款	透支
Cust.1	31	——	2001	——	Yes
Cust.2	——	2000	——	2033	No
Cust.3	51	50000	——	——	Yes
Cust.4	49	——	——	4046	Yes
Cust.5	——	50000	2300	10824	No

通过对表 2-5 进行分析，可以发现在表 2-5 中，每条记录所对应的属性值可能都存在缺省。不过每条记录的都有六个属性值分别是客户、年龄、存款、薪水、贷款和透支。

表 2-6. 银行 B 的客户数据

客户	年龄	存款	薪水	贷款	透支
Cust.1	——	1000	——	2012	Yes
Cust.2	23	——	1230	2033	No
Cust.3	——	——	5032	3032	Yes
Cust.4	——	1000	18909	——	Yes
Cust.5	34	50000	——	——	No
Cust.6	35	10000	26000	20385	Yes

对表 2-6 进行分析，同样可以发现在表 2-6 中每条记录对应的属性值也存在缺省，但是每条记录也有六个和表 2-5 一样的属性值。而表 2-5 和表 2-6 是分别存储在银行 A 和银行 B 的以任意方式划分的数据集，其中各个数据集中每条记录的属性值可能都不是很全，但是也可以根据数据表中的“客户”这个属性把两张表连

接起来。

本课题主要是针对以上三种方式分布的数据集进行研究，因此需要设计可以在多密钥加密的水平分布、垂直分布和任意分布的数据集上，能够利用 SVM 算法做具有隐私保护功能的数据挖掘的方案。

2.4 本章小结

本课题主要研究如何在多密钥加密的整数域和有理数域上设计可以利用支持向量机做具有隐私保护功能的数据挖掘算法。在本章主要对本课题所设计的算法中用到的安全多方计算协议和支持向量机算法进行了介绍，另外对本课题所研究数据的三种分布形式通过两个银行存储的客户数据集为例进行了说明。

第3章 基于安全多方计算的支持向量机算法

为了解决在具有隐私信息的分布式存储的各方数据上利用支持向量机算法做数据挖掘，并且在数据挖掘的过程中不泄露数据隐私信息的问题，本课题设计了基于安全多方计算的隐私保护支持向量机算法。基于本课题设计的算法在具有隐私的数据上做数据挖掘，可以保护原始数据的隐私，中间计算结果的隐私，最终分类模型的隐私，和预测结果的隐私。

3.1 隐私需求

在具有隐私信息的数据上利用机器学习算法做隐私保护数据的过程中会面临以下的隐私需求：

存储隐私：由于本地的存储和计算资源相对比较紧张，因此目前越来越多的数据拥有者会将自己的数据放到云服务商那里。但是云服务商是不可信的第三方，因此为了保护数据的隐私信息，需要在数据上传到云之前，对数据做一个保护处理，并且要保证处理后的结果不影响后续的数据挖掘。

计算隐私：由于数据是存储在云端，因此为了提高数据挖掘的性能，减少非必要的数据传输过程，一般数据挖掘的计算过程也是在云端进行的。而在计算的过程中也要确保不能从中间的计算结果中得到原始数据的隐私，因此需要对数据挖掘的方法进行一个全新的设计。

为了解决第一个意思需求，传统的方案中是对原始数据做扰动的处理，不过在扰动处理的过程中，会在数据中加入噪声，而这不仅会破坏原始数据的信息，也会对最终的分类模型和分类结果有负面的影响。因此在本课题设计的方案中，为了避免这种影响，采用了密码学的方法来对隐私数据进行处理。数据拥有者需要在上传数据之前利用自己的公钥对数据进行一个加密的处理，并把加密后得到的相应密文上传到云端。此时云端存储的都是密文，因为云端没有密文对应的解密密钥，所以云端不能够从密文中得到原始的数据信息，这样也就保护了数据的隐私信息。而为了不影响后续的数据挖掘过程，本课题所设计的加密方案是具有同态性质的，因此可以满足在密文上做代数运算的需要。

因为在整个数据挖掘的过程中，所有的计算都是在密文上进行的，而中间的计算结果、最后的分类模型和预测的结果无论是在云端还是传输的过程中都是密文形式，而敌手没有相应的私钥是得不得相应的明文结果，这样也就解决第二个安全需求。

除了上面的隐私需求以外，当利用加密的方法做隐私保护数据挖掘时，还会遇到下面几种困难。

由于在本课题设计的方案中，数据拥有者是利用自己的公钥加密数据并上传到云端，而云端此时存储的是多个公钥加密的数据，因此不能简单的利用同态加密算法来对数据做加密和计算的操作，因为传统的同态加密算法只支持在同一个公钥加密的数据上做计算，而针对在不同公钥加密的数据上做计算的问题，需要设计一种新的同态加密算法来解决，使得利用这种算法可以在多密钥加密的数据上做代数计算。

除此之外一般的数据加密算法只是支持对整型的数据，做加解密和其他计算的操作。而这也限制了利用加密的方法做隐私保护数据挖掘的范围和性能，特别是当数据拥有者的数据是浮点数时，这种方法就不适用了，而为了解决这个问题也需要对目前的加密算法做改进。

另外数据拥有者之间的数据分布方式也不仅是水平分布，还会出现垂直分布和任意分布，特别是当各个数据拥有者之间的数据是用各自不同的公钥加密的，因此需要设计支持在多密钥加密的分布式存储的各个参与方的数据集上做具有隐私保护功能的数据挖掘。

为了解决上面的三个困难，本课题分别设计了支持在多密钥加密的整数域和有理数域上，利用支持向量机做具有隐私保护功能的数据挖掘，而且各个参与方的数据可以以分布式的方式存储。另外在数据挖掘的整个过程中可以保护原始数据的隐私信息、中间计算结果的隐私信息、分类模型的隐私信息和最终分类预测结果的隐私信息。

3.2 系统架构

本课题主要基于 Liu 的算法[52]设计了利用 SVM 做隐私保护数据的方案，方案中主要包括两个云端服务器，多个数据拥有者，一个数据挖掘者，还有一个密钥生成中心。具体架构如图 3-1 所示：

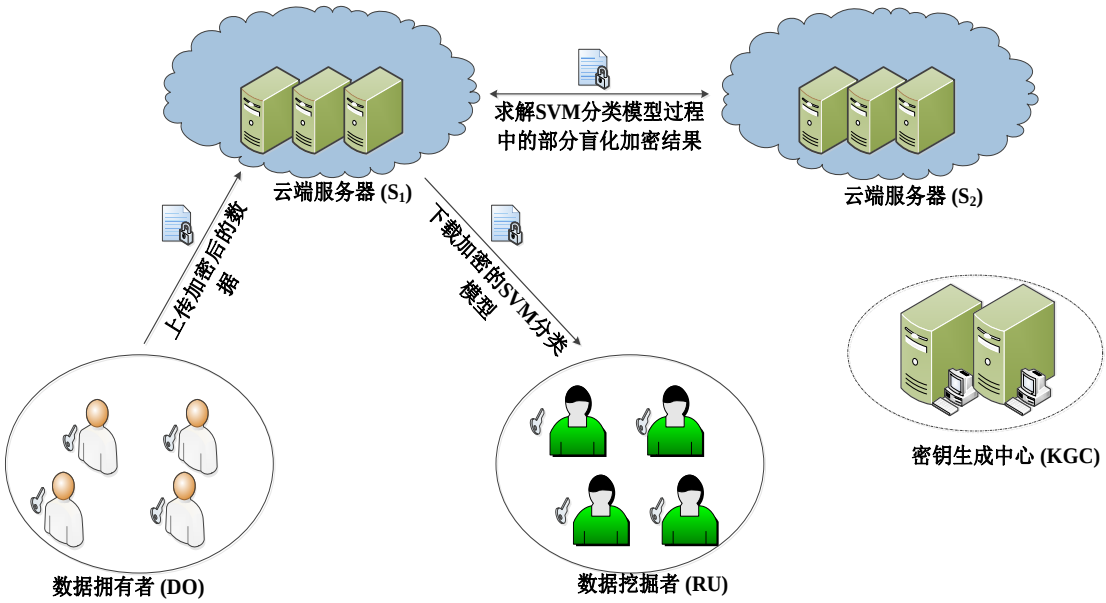

图 3-1. 系统架构

其中的密钥生成中心（*KGC*）：主要为数据拥有者和数据挖掘者提供生成密钥的参数，以及为云端服务器生成主密钥，并把主密钥分成两部分，分别发给两个云端服务器。

云端服务器（S_1）：主要存储数据拥有者加密后的数据，并通过和云端服务器 S_2 的交互计算来实现在多密钥加密的数据上，利用数据挖掘者 *RU* 所提供的数据挖掘方案做具有隐私保护功能的数据挖掘目的。

云端服务器（S_2）：主要通过和云端服务器 S_1 的交互计算来实现在多密钥加密的数据上，利用数据挖掘者 *RU* 所提供的机器学习算法做具有隐私保护功能的数据挖掘目的。

数据拥有者（*DO*）：数据的拥有者可以是银行和医院等单位。每个数据拥有者的数据中都具有一些隐私信息，因此这些数据不能公开也不能让其他数据拥有者和第三方知道。由于本地的存储和计算资源的限制，各个数据拥有者利用密钥生成中心发来的参数生成自己的公私钥对，并利用各自的公钥把自己的数据加密后上传到云端服务器 S_1 中。

数据挖掘者（*RU*）：数据挖掘者可以是研究所和一些企业等，其主要的工作是设计一种具有隐私保护功能的数据挖掘方案。利用这种方案可以在多个数据拥有者的数据上做数据挖掘，并且也数据挖掘的整个过程中不会泄露数据的隐私信息。数据挖掘的训练过程完成后，数据挖掘者会得到一个对应的分类模型，当数据拥有者有新的数据需要预测数据对应的类别时，可以利用之前训练得到的模

型进行分类预测处理，并且预测结果只能让数据拥有者知道。

在本课题设计的方案中，多个数据拥有者和数据挖掘者分别利用密钥生成中心发送的参数生成自己的公私钥对，然后数据拥有者利用自己的公钥加密自己的数据并上传到云端 S_1 中，然后云端 S_1 和 S_2 利用数据挖掘者所提供的 SVM 训练方法，通过交互来完成 SVM 的训练过程，并得到在多密钥加密数据上的分类模型。当数据拥有者有新的数据时，可以再次利用自己的公钥加密数据并上传到云端 S_1，然后 S_1 和 S_2 通过交互把新的密文数据代入到之前训练得到的分类模型中完成计算，并得到预测的分类结果，最后把分类结果发送给数据拥有者，数据拥有者就可以通过解密得到最终的明文分类结果。

3.3 隐私安全模型

在本课题设计方案中，KGC 我们认为是可信的，而数据拥有者、数据挖掘者、云端服务器 S_1 和 S_2 是半诚实的。他们会严格执行方案中设计的协议，不过在执行的过程中，也会记录一些中间的计算结果，并尝试从这些结果中去猜测原始数据的信息。为了更好地分析本文设计的方案的安全性，我们引入了一个敌手 A，这个敌手的能力主要包括以下两点：

（1） 敌手可以窃听两个云端之间的交互过程，以及数据拥有者和数据挖掘者与云端的交互过程。

（2） 敌手不能同时和云端 S_1 和 S_2 合谋，但是可以和云端 S_1 和 S_2 中的一个合谋。

敌手的主要目标是获得原始数据的信息、分类的模型和最后的分类预测结果。

3.4 水平分布整数域上的支持向量机分类算法

本课题设计的方案是支持多个数据拥有者的，为了方便解释和说明下面主要以两个数据拥有者（ DO_1， DO_2 ），一个数据挖掘者（ RU ）举例介绍本文设计的支持在多密钥加密的整数和有理数上做数据挖掘的支持向量机算法。

本课题在利用 SVM 算法进行分类训练时，首先有一堆训练数据的正负样本，标记为 $\{x_i, y_i\}, i=1,\cdots,l, y_i \in \{-1,1\}, x_i \in R^d$，假设有一个超平面 $H: w\cdot x+b=0$ 可以把这些样本准确无误地分割开来，并且同时存在两个平行于 H 的超平面 H_1 和 H_2：

$$\begin{aligned} H_1: &\quad w\cdot x+b=1 \\ H_2: &\quad w\cdot x+b=-1 \end{aligned} \tag{3-1}$$

结合图 2-1 进行分析，可以把 H 看成图中的 B，而 H_1 和 H_2 分别是图中的 A 和

C，并有如下约束：

当 $y_i=1$ 时

$$w\cdot x+b\geq 1 \tag{3-2}$$

当 $y_i=-1$ 时

$$w\cdot x+b\leq -1 \tag{3-3}$$

结合以上的式子可以得到：

$$y_i\left(w\cdot x_i+b\right)-1\geq 0 \tag{3-3}$$

超平面 H_1 和 H_2 的距离为：

$$M=2/\|w\| \tag{3-4}$$

SVM 算法的目标就是计算出一个 H 把不同种类的数据分隔开，并要求类间的距离最大。为了计算出最好的 H，需要最大化不同类别样本点的间隔 $M=2/\|w\|$，相应即最小化 $\|w\|^2$。那么问题可以转为下面式（3-5）的形式：

$$\min\|w\|^2/2 \qquad s.t.\quad y_i\left(w\cdot x_i+b\right)-1\geq 0 \tag{3-5}$$

为了计算上式（3-5），可以利用拉格朗日方法求解。于是可以把要计算的问题转为下式（3-6）的形式：

$$L(w,b,\alpha_i)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{t}\alpha_i\left(y_i\left(w\cdot x_i+b\right)-1\right)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{t}\alpha_i y_i\left(w\cdot x_i+b\right)+\sum_{i=1}^{t}\alpha_i \tag{3-6}$$

其中：$\alpha_i\geq 0$

则上述的规划问题变为：

$$\min_{w,b}\max_{\alpha_i\geq 0}L(w,b,\alpha_i) \tag{3-7}$$

对此可以对式（3-7）做对偶变化可以得到：

$$\min_{w,b}\max_{\alpha_i\geq 0}L(w,b,\alpha_i)=\max_{\alpha_i\geq 0}\min_{w,b}L(w,b,\alpha_i) \tag{3-8}$$

进一步可以转换为：

$$\max_{\alpha_i\geq 0}\min_{w,b}L(w,b,\alpha_i)=\max_{\alpha_i\geq 0}\left\{\sum_{i=1}^{t}\alpha_i-\frac{1}{2}\sum_{i=1}^{t}\sum_{j=1}^{t}\alpha_i\alpha_j y_i y_j\left(x_i\cdot x_j\right)\right\} \tag{3-9}$$

由此上述式（3-7）的规划问题就变为：

$$\max_{\alpha_i\geq 0}\left\{\sum_{i=1}^{t}\alpha_i-\frac{1}{2}\sum_{i=1}^{t}\sum_{j=1}^{t}\alpha_i\alpha_j y_i y_j\left(x_i\cdot x_j\right)\right\} \tag{3-10}$$

并且需要满足：

$$\begin{cases} \sum_{i=1}^{t} \alpha_i y_i = 0 \\ \alpha_i \geq 0 \end{cases} \tag{3-11}$$

另外如果一个样本是 H_1 和 H_2 上的点，那么其对应的拉格朗日系数不是零，假如不是 H_1 和 H_2 上的点，那么其对应的拉格朗日系数为 0，结合：

$$w = \sum_{i=1}^{t} \alpha_i y_i x_i \tag{3-12}$$

就可以计算求解 H 的法向量 w，基于约束支持向量可以计算分割阈值 b，那么进而能够计算出 H_1 和 H_2，也即 SVM 算法模型。

由于在上述求解最优分割面的过程中需要不断对数据向量进行内积的运算，而为了在多密钥加密的数据上完成上述的计算过程，需要基于安全多方计算设计安全的内积运算。

首先 KGC 生成参数 P 和主密钥 MK，并把主密钥 MK 分成两部分 k_1, k_2 分别发送给 S_1 和 S_2，另外 DO_1， DO_2，和 RU 利用 P 分别生成自己的公私钥对 (pk_i, sk_i)，其中 $i \in \{1,2,3\}$。

接下来 DO_1 和 DO_2 分别利用自己的公钥 pk_1 和 pk_2 加密自己的数据，并把加密后的数据上传到云端 S_1。当 S_1 收到 DO_1 和 DO_2 上传的数据后， RU 把自己利用 SVM 做数据挖掘的方案也上传到 S_1，然后 S_1 就开始通过和 S_2 的交互计算来求解 SVM 的分类模型 $wx + b = 0$，具体的交互计算过程如下所示：

在前面的介绍中，可以知道直接求解凸二次规划问题比较麻烦，可以通过引入拉格朗日因子 α，把凸二次规划问题转为对偶问题进行求解。而在求解的过程中可以通过 SMO 算法来计算 α，在上面的计算过程中主要的计算是内积的计算，而内积可以分为先进行乘法计算，然后再做加法计算。下面以 m_1 的密文 $Enc_{pk_1}(m_1)$ 和 m_2 的密文 $Enc_{pk_2}(m_2)$ 为例来介绍本文设计的支持在多密钥加密的整型数据上做乘法和加法计算的方案，首先乘法的计算过程如下所示：

首先 S_1 选择四个随机数 $r_1, r_2, R_1, R_2 \in Z_n$ 并基于加法同态的性质做如下的计算：

$$\begin{aligned} C_1 &= Enc_{pk_1}(m_1) \cdot Enc_{pk_1}(r_1) = Enc_{pk_1}(m_1 + r_1) \\ C_2 &= Enc_{pk_2}(m_2) \cdot Enc_{pk_2}(r_2) = Enc_{pk_2}(m_2 + r_2) \\ C_3 &= Enc_{pk_1}(R_1) \cdot \left[Enc_{pk_1}(m_1) \right]^{N - r_2} = Enc_{pk_1}(R_1 - m_1 \cdot r_2) \\ C_4 &= Enc_{pk_2}(R_2) \cdot \left[Enc_{pk_2}(m_2) \right]^{N - r_1} = Enc_{pk_2}(R_2 - m_2 \cdot r_1) \end{aligned} \tag{3-13}$$

接下来S_1根据上面的计算结果并结和自己的部分主密钥k_1做下面的计算：

$$\begin{aligned} C_1^{'} &= SDec1_{k_1}(C_1) \\ C_2^{'} &= SDec1_{k_2}(C_2) \\ C_3^{'} &= SDec1_{k_1}(C_3) \\ C_4^{'} &= SDec1_{k_1}(C_4) \end{aligned} \tag{3-14}$$

当S_1完成上面的计算后，把计算结果C_1、C_2、C_3、C_4、$C_1^{'}$、$C_2^{'}$、$C_3^{'}$和$C_4^{'}$发给S_2，然后S_2利用自己的部分主密钥k_2完成下面的计算。

$$\begin{aligned} C_5 &= SDec2_{k_2}(C_1, C_1^{'}) \cdot SDec2_{k_2}(C_2, C_2^{'}) \\ C_6 &= SDec2_{k_2}(C_3, C_3^{'}) \\ C_7 &= SDec2_{k_2}(C_4, C_4^{'}) \end{aligned} \tag{3-15}$$

随后，S_2利用一个新的公钥$pk_{\Pi} = (N, g, g^{\alpha_1+\alpha_2+\alpha_3})$对上面计算的结果进行加密：

$$\begin{aligned} C_5^{'} &= Enc_{pk_{\Pi}}(C_5) \\ C_6^{'} &= Enc_{pk_{\Pi}}(C_6) \\ C_7^{'} &= Enc_{pk_{\Pi}}(C_7) \end{aligned} \tag{3-16}$$

当S_2把上面计算后得到的结果$C_5^{'}$，$C_6^{'}$，$C_7^{'}$发给S_1后，S_1需要利用新的公钥$pk_{\Pi} = (N, g, g^{\alpha_1+\alpha_2+\alpha_3})$加密$r_1, r_2, R_1, R_2$：

$$\begin{aligned} CR_1 &= [Enc_{pk_{\Pi}}(r_1 \cdot r_2)]^{N-1} \\ CR_2 &= [Enc_{pk_{\Pi}}(R_1)]^{N-1} \\ CR_3 &= [Enc_{pk_{\Pi}}(R_2)]^{N-1} \end{aligned} \tag{3-17}$$

随后S_1通过计算$C_5^{'}$，$C_6^{'}$，$C_7^{'}$，CR_1，CR_2，和CR_3的乘积来完成多密钥的同态乘计算，也即：

$$Enc_{pk_{\Pi}}(m_1 \cdot m_2) = C_5^{'} \cdot C_6^{'} \cdot C_7^{'} \cdot CR_1 \cdot CR_2 \cdot CR_3 \tag{3-18}$$

在已知pk_1加密m_1获得密文$Enc_{pk_1}(m_1)$，和pk_2加密m_2获得密文$Enc_{pk_2}(m_2)$的情况下，可以经由两个云端S_1和S_2的交互计算获得通过公钥pk_{Π}加密的密文$Enc_{pk_{\Pi}}(m_1 \cdot m_2)$。并且在整个计算期间中$S_1$和$S_2$得不得$m_1$，$m_2$，和$m_1 \cdot m_2$的明文信息，从而保护了数据隐私。接下来介绍已知$Enc_{pk_1}(m_1)$和$Enc_{pk_2}(m_2)$如何计算获得同态加的结果$Enc_{pk_{\Pi}}(m_1 + m_2)$。

首先S_1选择两个随机数$r_1, r_2 \in Z_n$，并计算下式：

$$C_1 = Enc_{pk_1}(m_1)\cdot Enc_{pk_1}(r_1) = Enc_{pk_1}(m_1 + r_1)$$
$$C_2 = Enc_{pk_2}(m_2)\cdot Enc_{pk_2}(r_2) = Enc_{pk_2}(m_2 + r_2) \quad (3\text{-}19)$$

然后 S_1 利用自己的部分主密钥 k_1 计算下式：

$$C_1^{'} = SDec1_{k_1}(C_1)$$
$$C_2^{'} = SDec1_{k_2}(C_2) \quad (3\text{-}20)$$

并把计算结果 C_1， C_2， $C_1^{'}$，和 $C_2^{'}$ 发给 S_2，然后 S_2 利用自己的部分主密钥 k_2 计算下式：

$$C_1^{''} = SDec2_{k_2}(C_1, C_1^{'}) = m_1 + r_1$$
$$C_2^{''} = SDec2_{k_2}(C_2, C_2^{'}) = m_2 + r_2 \quad (3\text{-}21)$$

随后， S_2 利用公钥 $pk_{\Pi} = (N, g, g^{\alpha_1+\alpha_2+\alpha_3})$ 加密 $C_1^{''} + C_2^{''}$，也就是

$$C = Enc_{pk_{\Pi}}(C_1^{''} + C_2^{''}) \quad (3\text{-}22)$$

并把加密后的结果 C 发给 S_1，最后 S_1 通过下面的计算完成多密钥的同态加计算：

$$Enc_{pk_{\Pi}}(m_1 + m_2) = C\cdot[Enc_{pk_{\Pi}}(r_1 + r_2)]^{(N-1)} \quad (3\text{-}23)$$

在已知 pk_1 加密 m_1 获得密文 $Enc_{pk_1}(m_1)$，和 pk_2 加密 m_2 获得密文 $Enc_{pk_2}(m_2)$ 的情况下，可以利用两个云端 S_1 和 S_2 的交互计算获得利用公钥 pk_{Π} 加密的密文 $Enc_{pk_{\Pi}}(m_1 + m_2)$。并且在计算的过程中 S_1 和 S_2 得不得 m_1， m_2，和 $m_1 + m_2$ 的明文结果，从而保证了数据隐私。

这样 SVM 算法在密文上进行训练时，就能够在多密钥加密的密文数据上通过做同态加和同态乘的计算完成点积的计算，从而获得最后利用公钥 pk_{Π} 加密的分类模型 $sign(wx+b)$。之后 DO_1 和 DO_2 可以利用自己的私钥计算公式 2-10 的值，并把计算的结果发给 RU，那么 RU 之后就根据自己的私钥计算公式 2-10 的值，并结合 DO_1 和 DO_2 计算的结果计算公式 2-11 的值来分别解密 w 和 b 的值，由此得到明文形式的分类模型。

当 DO_1 或 DO_2 有新的整数类型的数据想要分类时，可以把新的数据利用自己的公钥加密并上传到云端，利用之前通过 RU 的分类算法训练得到的 SVM 分类模型做分类。通过观察分类模型，可以看出模型中主要包括乘和加的计算，因此可以通过方案中设计的支持在多密钥加密的整数上做同态加和乘的算法完成计算，不过在计算的过程中需要把之前利用公钥 pk_{Π} 做加密的过程换成利用 DO_1 或 DO_2

的公钥进行加密，并把计算的结果发给 DO_1 或 DO_2 。这时的结果是用 DO_1 或 DO_2 的公钥进行加密的，因此 DO_1 或 DO_2 可以通过自己的私钥解密得到明文结果，并通过和 0 进行比较得到最终的分类结果是正类还是负类。

3.5 水平分布有理数域上的支持向量机分类算法

上一小节中介绍的是在多密钥加密的整数上利用 SVM 算法做分类，不过当用户的数据是有理数类型时，上述方案就不适用了。因为有理数的数据可以是小数和整数，当是小数时，由于小数的存储和计算都不同于整数，因此需要对小数的加密、解密，和其他的计算设计不同于整数的，新的算法。

本课题在解决这个问题时，是把小数先表示成分数的形式，然后通过分别对分子和分母做运算来解决小数的加密、解密和其他的计算问题。例如 0.3，通过表示成 $(3,10)$，然后分别对分子 3 和分母 10 做加密、解密和其他的计算，来解决当数据拥有者的数据是小数时无法进行数据挖掘的问题。在本课题设计的方案中分子的范围是 $[-N_1, N_1]$，而分母的范围是 $[0, N_1]$，假如用 $length()$ 表示比特长度，那么 $length(N_1) < (length(N)/8) - 1$。

当云端 S_1 存储有 DO_1 通过自己的公钥 pk_1 加密明文 $(m_1, m_1^{'})$ 获得的密文 $(Enc_{pk_1}(m_1), Enc_{pk_1}(m_1^{'}))$，$DO_2$ 通过自己的公钥 pk_2 加密明文 $(m_2, m_2^{'})$ 获得的密文 $(Enc_{pk_2}(m_2), Enc_{pk_2}(m_2^{'}))$ 时。根据前面的介绍可以知道，当利用 SVM 算法在数据上进行训练时，为了求解分类模型，所做的计算是内积。而内积可以分为先求解乘法，然后做加法计算。那么需要设计在多密钥加密的小数上做乘法和加法的安全协议。

当需要做乘法计算时，由于密文 $(Enc_{pk_1}(m_1), Enc_{pk_1}(m_1^{'}))$ 和 $(Enc_{pk_2}(m_2), Enc_{pk_2}(m_2^{'}))$ 中的 m_1，$m_1^{'}$，m_2，和 $m_2^{'}$ 都是整数，因此对于 $Enc_{pk_1}(m_1)$ 和 $Enc_{pk_2}(m_2)$，可以首先根据前面设计的在多密钥加密的整数上，利用云端 S_1 和 S_2 的交互计算来完成同态乘的方案，进而通过计算得到 $Enc_{pk_{\Pi}}(m_1 \cdot m_2)$。同理对于 $Enc_{pk_1}(m_1^{'})$ 和 $Enc_{pk_2}(m_2^{'})$ 可以计算得到 $Enc_{pk_{\Pi}}(m_1^{'} \cdot m_2^{'})$，这时就可以得到 $(m_1, m_1^{'})$ 和 $(m_2, m_2^{'})$ 乘积的密文形式也就是 $(Enc_{pk_{\Pi}}(m_1 \cdot m_2), Enc_{pk_{\Pi}}(m_1^{'} \cdot m_2^{'}))$，由此完成了同态乘的计算。

当需要在云端基于密文 $(Enc_{pk_1}(m_1), Enc_{pk_1}(m_1^{'}))$ 和密文 $(Enc_{pk_2}(m_2), Enc_{pk_2}(m_2^{'}))$ 计算同态加时，可以首先根据 $Enc_{pk_1}(m_1)$ 和 $Enc_{pk_2}(m_2^{'})$ 计算出 $Enc_{pk_{\Pi}}(m_1 \cdot m_2^{'})$，同理可以基于 $Enc_{pk_1}(m_1^{'})$ 和 $Enc_{pk_2}(m_2)$ 计算出 $Enc_{pk_{\Pi}}(m_1^{'} \cdot m_2)$，然后通过 $Enc_{pk_1}(m_1^{'})$ 和

$Enc_{pk_2}(m_2^{'})$ 计算得到 $Enc_{pk_\Pi}(m_1^{'}\cdot m_2^{'})$。另外对于利用同一个公钥 pk_Π 加密的密文 $Enc_{pk_\Pi}(m_1\cdot m_2^{'})$ 和 $Enc_{pk_\Pi}(m_1^{'}\cdot m_2)$，根据算法同态加的性质可以得到 $(Enc_{pk_\Pi}(m_1\cdot m_2^{'})+Enc_{pk_\Pi}(m_1^{'}\cdot m_2))$，那么此时就可以计算得到 $(m_1,m_1^{'})$ 和 $(m_2,m_2^{'})$ 加法和的密文形式 $\left(\left(Enc_{pk_\Pi}(m_1\cdot m_2^{'})+Enc_{pk_\Pi}(m_1^{'}\cdot m_2)\right),Enc_{pk_\Pi}(m_1^{'}\cdot m_2^{'})\right)$。

依据上面的方案可以完成在多密钥加密的小数上做同态加和同态乘的计算，并且在计算的过程中不泄露明文数据的隐私信息。因此就可以利用 SVM 算法在多密钥加密的小数上进行训练，并得到最终的分类模型 $sign(wx+b)$。而且当 DO_1 和 DO_2 利用自己的私钥计算完公式 2-10 的值，并把计算的结果发给 RU 后，RU 就可以根据自己的私钥计算公式 2-10 的值，并结合 DO_1 和 DO_2 计算的结果计算公式 2-11 的值来分别解密 w 和 b 的值，由此得到明文形式的分类模型。

当 DO_1 和 DO_2 有新的有理数类型的数据需要做分类预测时，依然可以利用自己的公钥加密新的数据并上传到云端，利用之前训练得到的模型做分类。通过观察分类模型，可以看出所需的计算主要包括乘法和加法的运算，因此可以通过上述方案中设计的支持在多密钥加密的有理数上做同态加和乘的算法进行计算，不过在计算的过程中也需要把之前利用公钥 pk_Π 做加密的过程换成利用 DO_1 或 DO_2 的公钥进行加密，并把计算的结果发给 DO_1 或 DO_2。这时 DO_1 或 DO_2 就可以通过自己的私钥解密得到明文结果，并通过和 0 进行比较得到最终的分类结果是正类还是负类。

3.6 垂直和任意分布数据上的支持向量机分类算法

当数据分布式的存储在多方时，因为本地的存储和计算资源受限，因此可以借助云的存储和计算能力进行数据挖掘。当数据具有隐私信息时，为了保护数据的隐私，需要各方利用自己的公钥加密数据，然后把密文上传到云端，此时云上存储的是多密钥加密的数据。而数据的分布形式可以是水平分布、垂直分布和任意分布，下面以两个数据拥有者 DO_1 和 DO_2 为例介绍本文所设计的方案如何在分布式存储的数据上做具有隐私保护功能的数据挖掘。

当利用本文设计的方案处理水平分布的数据时，结合前面的分析可以发现当数据是水平分布时，每个用户的数据记录都有全部的数据属性，当利用支持向量机算法进行分类时，主要做的是点积的计算，而点积的具体计算又可以分为三种情况，即对 DO_1 的数据做点积的计算，对 DO_2 的数据做点积的计算，结合 DO_1 和 DO_2 的数据做点积的计算。由于 DO_1 和 DO_2 的数据都是利用自己的公钥加密后上

传到云上的，因此可以首先根据算法的单密钥同态加和同态乘的性质分别对 DO_1 和 DO_2 的数据做点积的计算。当需要结合 DO_1 和 DO_2 的数据做点积的计算时，由于此时要处理的数据是分别利用 DO_1 和 DO_2 的公钥加密的，因此需要利用本方案中设计的支持在多密钥加密的数据上做同态加和同态乘的计算完成支持向量机的训练过程。而当数据是以垂直分布的方式存储时，由于不同用户的数据属性可能不一样，但是每个属性都是固定的。因此也可以把支持向量机的计算分为三种情况：即对 DO_1 的数据做点积的计算，对 DO_2 的数据做点积的计算，结合 DO_1 和 DO_2 的数据进行点积的计算。所以可以首先基于算法的单密钥同态加和同态乘的性质对 DO_1 和 DO_2 的密文数据做加和乘的计算，然后再依据多密钥的同态加和同态乘的性质，把通过不同公钥加密的计算结果转为同一个公钥加密的结果进行后续的计算。当数据是以任意形式分布时，这种情况下的计算比水平分布和垂直分布下的计算要复杂，不能单独对 DO_1 和 DO_2 的密文数据做加和乘的计算，需要直接利用方案中的多密钥同态加和同态乘的性质，把所有利用不同公钥加密的数据转为利用同一个公钥加密的数据进行计算。而此时所需要的计算量要比数据是水平分布和垂直分布时要高。

当数据线性不可分时，可以通过核函数把数据转到高维空间进行分类求解，支持向量机中可以利用的核函数有高斯核函数和拉普拉斯核函数等。当核函数是非线性函数时，由于所有的计算都是针对密文进行计算，因此可以利用泰勒展开式或麦克劳林函数把非线性函数转为线性函数进行计算求解，然后针对线性函数设计具体的安全协议完成密文上的计算。从而在线性不可分的数据集上可以利用本课题所设计的支持向量机算法做具有隐私保护功能的数据挖掘。

3.7 算法性能分析

假设幂指数长度为 $|N|$ 的一次常规幂运算需要 $1.5|N|$ 次的乘法计算（$|N|$ 表示 N 的比特长度）[57]，结合前面对同态加密算法的介绍可以看出在本文的方案中加密和求逆的过程包含两部分，每部分都需要一次指数运算，那么对于本文算法中的加密和求逆的计算就需要 $3|N|$ 次的乘法计算。对解密算法进行分析可以发现当用户通过私钥解密时需要 $1.5|N|$ 次的乘法计算。当密文是通过方案中用户公钥的乘积加密时，如果想要解密这个密文，结合公式 2-10 和公式 2-11 可以发现此时需要两次指数计算，因此解密此时的密文需要 $3|N|$ 的乘法计算。结合两个云的交互过程和具体计算进行分析可以发现两个云通过部分主密钥解密密文数据需要

4.5$|N|$次乘法计算。算法中对多密钥加密的数据进行加法计算时需要首先利用盲化技术并结合算法的同态性质对用户上传的密文数据进行处理，然后基于处理后的密文数据上进行后续的交互计算，通过对具体的计算进行分析可以发现云S_1需要21$|N|$次乘法计算，云S_2需要12$|N|$次乘法计算。当对多密钥加密的密文数据进行乘法计算时两个云的交互过程比做加法时要更加复杂一些，此时为了完成在多密钥加密的数据上进行乘法计算需要对多个数据进行盲化处理和加解密的操作，结合具体的计算可以发现云S_1需要45$|N|$次乘法计算，云S_2需要27$|N|$的乘法计算。

接下来对本文设计方案的通信复杂度进行分析，在用户加密自己的数据并上传到云S_1的过程中主要传输的是原始数据对应的密文，因此此时需要传输的密文长度是4$|N|$比特，当两个云在多密钥加密的数据上做加法计算时，两个云之间传输的主要是盲化后的密文数据以及两个云利用自己的部分主密钥解密计算的密文结果，此时需要传输16$|N|$比特。做乘法计算的过程类似于加法计算，此时需要传输36$|N|$比特。

3.8 算法比较

通过对近些年隐私保护数据挖掘的发展进行调研和分析，可以发现有一些利用 SVM 算法在密文数据上做隐私保护数据挖掘的工作，通过对比分析可以发现，很多论文的方案是基于全同态加密算法设计了具有隐私保护功能的 SVM 算法，论文[46]的方案中基于全同态加密算法设计了 SVM 算法，该算法用到了一个云端服务器，并且只能针对两个用户的垂直分布数据进行数据挖掘。因为全同态加密算法可以支持多种运算，所以这种算法的效率不高，而论文[46]由于采用了全同态的加密算法，所以相应方案的效率也不高，另外在计算的过程中需要用户在线下载、传输数据并参与计算。而在本文设计的方案中利用了半同态的加密算法，半同态加密算法由于只能支持一种运算所以相应的效率要远远高于全同态加密算法，另外本文设计的方案不仅能支持垂直分布的数据也能支持水平分布和任意分布的数据，并且用户也不需要在线参与计算。

在论文[49]中也是利用 SVM 对多密钥加密的数据做隐私保护的数据挖掘，为了能够在多密钥加密的数据上做计算，在论文[49]设计的方案中是通过参与方之间的交互来协商出一个新的密钥，通过把密文转成该密钥对应的密文进行后续的计算，所以为了协商出新的密钥需要参与方在线参与部分计算。而且随着用户的增加，为了协商出新的密钥相应的交互过程也会更加复杂，而在本文所提出的方案

中，当多个用户把加密后的数据上传到云上后是通过两个云之间的交互把用户的密文转成用户公钥的乘积对应的密文，来完成后面的计算，在整个计算的过程中不需要用户的参与，并且密文转换过程的计算性能并不随着用户的增加而出现明显的变化。

在论文[10]设计的方案中为了在多密钥加密的密文上做乘和加的计算，也是引入了两个云来做计算，不过论文[10]中的方案仅仅支持在垂直分布的数据上做具有隐私保护功能的数据挖掘，而本文设计的方案除了可以在垂直分布的数据上做具有隐私保护功能的数据挖掘，也能够拓展应用到水平分布和任意分布的数据上。并且在论文[10]设计的方案需要用户除了做基本的加解密外，还需要在线参与其他的计算。

不过上述论文的方案都只是针对在整数域的密文数据上做具有隐私保护功能的数据挖掘，并不支持在有理数域的密文数据上做具有隐私保护功能的数据挖掘。而本文设计的方案除了在多密钥加密的整数域上，还可以在多密钥加密的有理数域上利用 SVM 算法做具有隐私保护功能的数据挖掘。

3.9 算法隐私安全性分析

前面分别介绍了在多密钥加密的整数和有理数上利用 SVM 算法做分类的方案，该方案可以保护原始数据的隐私，中间计算结果的隐私、分类模型的隐私，和最后分类预测结果的隐私。下面就对这两个方案的安全性进行详细的分析和证明。

因为在本文设计的方案中用户的密文除了可以通过用户的私钥解密外，还可以通过系统中的主密钥进行解密，因此这样方案是一个双陷门的方案。那么首先可以得到：

定理 3-1：基于 $Z_{N^2}^*$ 上 DDH 困难性的假设，可以证明前面介绍的方案中所用的加密算法是语义安全的。

证明：由于方案中所用的加密算法是一个双陷门的方案，因此方案的安全性就基于双陷门算法的安全性。因为双陷门方案的安全性，在标准模型中基于 $Z_{N^2}^*$ 上 DDH 困难性的假设，已经被证明是语义安全的[50]。另外方案中主密钥分成两部分这个过程的安全性，也可以通过 Shamir 秘密共享算法的安全性去保证[58]。那么因为用户的私钥都在用户自己手里，如果一个敌手没有同时拿到分开后的两部分主密钥，那么敌手是无法破解密文得到明文数据的隐私信息的。

因为方案中主要针对多密钥加密的整数和有理数设计了同态乘和同态加的方

案，那么接下来就主要分析一下这两个方案的安全性。在分析时主要根据“理想/现实”框架来进行分析：在理想模型中有一个可信的第三方，其他参与方都是把自己的输入发给这个可靠的第三方，然后通过第三方进行计算，并把计算的结果返回给所有的参与方。而在现实模型中是没有这个可信的第三方，所有的计算都是通过各个参与方之间的交互进行的。基于前面介绍的安全多方计算的安全模型，需要通过在理想模型中模拟敌手在现实模型中的行为证明协议的安全性。

定理 3-2：方案中所设计的同态加的算法可以安全地计算多密钥加密数据的加法和。

证明：假设有四个敌手 A_{DO_1}，A_{DO_2}，A_{S_1} 和 A_{S_2} 分别与数据拥有者 DO_1 和 DO_2 合谋，以及云端 S_1 和 S_2 合谋，接下来需要构造四个模拟器：Sim_{DO_1}，Sim_{DO_2}，Sim_{S_1}，和 Sim_{S_2}。

其中 Sim_{DO_1} 选择 m_1 作为模拟器的输入，并和 A_{DO_1} 做如下交互：Sim_{DO_1} 利用公钥 pk_1 加密 m_1，并把得到的密文 $Enc_{pk_1}(m_1)$ 发送给敌手 A_{DO_1}。然后输出 A_{DO_1} 的视图，主要是密文数据。根据双门陷加密算法的语义安全性，A_{DO_1} 的视图无论是在理想环境还是现实环境中都是不可区分的。而 Sim_{DO_2} 和 A_{DO_2} 也做如上的操作。

Sim_{S_1} 随机选择两个随机数 $r_1, r_2 \in Z_N$，并基于公钥 pk_1 和 pk_2 分别加密两个随机选择的消息 m_1 和 m_2 获得 $Enc_{pk_1}(m_1+r_1)$ 和 $Enc_{pk_2}(m_2+r_2)$。然后 Sim_{S_1} 利用公钥 pk_1 和 pk_2 对应的私钥 sk_1 和 sk_2 计算公式 2-10 的值，并把计算的结果发给 A_{S_1}，如果 A_{S_1} 输出无意义的值，那么 Sim_{S_1} 也输出无意义的值。无论是在现实模型还是理想模型中，敌手 A_{S_1} 得到的只是密文 $Enc_{pk_1}(m_1+r_1)$，$Enc_{pk_2}(m_2+r_2)$，以及 Sim_{S_1} 根据私钥 sk_1 和 sk_2 计算 2-10 的值。在现实模型中，这些由双陷门加解密方案的语义安全性，以及数据拥有者是诚实的来保证。因此敌手 A_{S_1} 在现实模型和理想模型中的视图是不可区分的。

Sim_{S_2} 随机选择一个消息 M，并基于公钥 pk_{Π} 加密，把加密的结果 $Enc_{pk_{\Pi}}(M)$ 发给敌手 A_{S_2}，假如 A_{S_2} 回复无意义的值，相应的 Sim_{S_2} 也输出无意义的值。A_{S_2} 的视图主要包括密文 $Enc_{pk_{\Pi}}(M)$。在现实模型中，这些同样可以根据双陷门加解密方案的语义安全性来保证。因此 A_{S_2} 的视图无论是在现实模型，还是理想模型中都是不可区分的。

定理 3-3：方案中所设计的同态乘的算法可以安全地计算多密钥加密数据的乘

积。

证明：这个同态乘证明过程和同态加的证明过程比较相似，在这里就不详细介绍。

定理 3-4：方案中所设计的同态加和同态乘的算法，可以在多密钥加密的有理数上安全地计算加法和以及乘积。

证明：由于方案中设计的支持在多密钥加密的有理数上做同态加和同态乘的计算，是把有理数转换成分数的形式，通过对分子和分母分别进行计算来达到对有理数计算的目的。而分数中的分子和分母都是整数，因此这个证明过程和在整数上的证明过程比较类似，在这里也不再详细解释。

接下来再对方案中交互过程的安全性进行分析。

因为如果有一个敌手 A，那么敌手 A 可以监听数据拥有者，数据挖掘者，和云端之间的交互过程，并得到他们相互之间传输的数据。不过敌手 A 不能同时和两个云端 S_1 和 S_2 合谋，只能和其中一个合谋。当敌手 A 监听两个云端之间的交互过程时，如果敌手和云端 S_1 合谋，那么敌手可以得到数据拥有者上传到 S_1 上的密文数据，不过因为这些密文数据都是用数据拥有着各自的公钥加密的，因此敌手是无法解密获得明文数据。

另外因为本文设计的方案采用的是一个双陷门的加解密方案，那么用户的密文数据也可以利用主密钥进行解密，不过在设计方案时，本文把主密钥分成了两部分，分别放到了两个云端 S_1 和 S_2 上，因此解密时需要 S_1 和 S_2 协助进行解密才能得到明文数据。所以当敌手 A 和云端 S_1 合谋时，敌手可以拿到 S_1 的那部分密钥，不过却拿不到 S_2 的密钥，所以敌手是无法解密获得数据拥有者的明文数据。当敌手和 S_2 合谋时，因为在设计方案时为了保护数据的隐私信息，在 S_1 向 S_2 发送数据之前，需要对 S_1 上存储的密文数据做了盲化处理，那么当 S_2 利用自己的那部分主密钥解密 S_1 发来的数据时，得到的是盲化后的明文数据，因此敌手 A 此时即使与 S_2 合谋也拿不到用户的原始明文数据。另外基于 Li 的分析[59]，能够发现如果在两个云端 S_1 和 S_2 之间加一个认证协议，那么即使敌手与数据挖掘者合谋，并发起“Bypass”攻击，敌手 A 依然得不得数据拥有者的原始数据。

基于前面的分析，可以知道最后的分类模型是通过公钥 pk_{Π} 加密的。数据挖掘者如果想要解密这个分类模型，需要在数据拥有者的帮助下进行计算，从而解密得到明文形式的分类模型。而这可以避免数据挖掘者滥用数据拥有者的数据，从而达到数据控制的目的。

3.10 本章小结

本章首先分析了隐私保护数据挖掘的隐私需求、本文设计的系统架构和所研究的安全模型。然后重点解释了本课题设计的可以在水平分布、垂直分布和任意分布的整数域和有理数域上利用支持向量机做具有隐私保护功能的数据挖掘算法。并从计算复杂度和通信复杂度这两个方面对算法的性能进行了理论分析。最后对算法的安全性在半诚实模型下进行了证明，可以发现利用本课题所设计的算法可以在保证分类预测结果高准确度的同时保护用户的原始数据隐私信息、中间计算结果的隐私信息、挖掘出的分类模型隐私信息和最后分类预测结果的隐私信息等。

第4章 隐私保护支持向量机算法的实现与分析

在上一章中主要介绍了本课题提出的基于安全多方计算的隐私保护支持向量机算法，并作了安全证明，下面主要介绍算法的系统实现并分析其性能。

4.1 系统实现

下面主要从开发环境、训练与测试数据集和实验结果等三个方面具体介绍本课题的系统实现。

4.1.1 开发环境

本课题的开发环境主要包括以下几个部分：

（1）PC 机型号为 DELL，处理器为 64 位的 Ubuntu，版本为 7.5.1804，处理器为 Intel Core TM i5-4570, 3.20GHz，内存是 1GB。

（2）开发语言是 Python2.7。

（3）所使用的库是 PBC，Charm，GMP。

4.1.2 训练与测试数据集

分别针对不同大小的数据集进行训练和测试，在训练和测试的过程中，把数据集分成三部分：训练集、验证集和测试集，每一个参与方的数据如表 4-1 所示。

表 4-1. 数据样例

客户	年龄	存款	薪水	贷款	透支
Cust.1	31	1000	2001	2012	Yes
Cust.2	23	2000	1230	2033	No
Cust.3	51	50000	5032	3032	Yes
Cust.4	49	1000	18909	4046	Yes
Cust.5	34	50000	2300	10824	No
Cust.6	35	10000	26000	20385	Yes

首先在训练集上做有监督的训练，并计算出一个 SVM 的二分类模型。然后通过验证集，对训练出的模型进行优化，主要是因为在训练 SVM 之前需要对 SVM 中的松弛参数和核函数进行人为的设定，这时就可以通过训练得到的 SVM 模型在验证集上的表现来调整 SVM 中的参数。最后再利用测试集进行模型预测和模型性能的评估，由此来检验本文方案中设计的隐私保护 SVM 算法的性能。

4.1.3 系统性能分析

首先由于本方案中提出的算法主要是在多密钥加密的密文数据上，利用 SVM 做隐私保护的数据挖掘。在挖掘的过程中主要做的计算是加密、解密，以及同态加和同态乘的计算。因此首先在不同大小的数据集上测试了加密和解密的消耗时间。如表 4-2 所示，可以发现当数据量不是很大时，加密时间和解密时间没有什么区别，当数据集的大小是 600KB 时，可以发现加密的时间是 301.61 秒，解密的时间是 297.05 秒。不过随机数据量的增加，解密时间会逐渐比加密时间长。当数据集的大小是 1000KB 时，可以发现此时的加密时间是 458.83 秒，而解密时间已经达到了 556.35 秒。

表 4-2. 加解密耗时

加解密消耗时间						
数据集大小（kb）	100.00	200.00	400.00	600.00	800.00	1000.00
加密耗时（s）	54.12	104.92	199.95	301.61	396.01	458.83
解密耗时（s）	57.93	114.72	231.54	297.05	412.83	556.35

接下来也在不同大小的数据集上对本方案中的同态加和同态乘的计算进行了实验，所消耗的时间如表 4-3 所示：

表 4-3. 同态加和乘计算的耗时

同态加和乘消耗时间						
数据集大小（kb）	100.00	200.00	400.00	600.00	800.00	1000.00
同态乘耗时（s）	425.65	787.84	1473.92	2204.07	2836.48	3372.23
同态加耗时（s）	247.35	454.41	873.51	1299.10	1607.67	1903.08

通过对表 4-3 进行分析发现，在多密钥加密的密文上做同态乘的计算时间要比做同态加的计算时间长一些，当在 600KB 的数据集上进行同态加和同态乘的计算可以发现同态乘的计算需要消耗 2204.07 秒，而同态加的计算耗时为 1299.10 秒。这也可以从方案中进行证明，因为对比同态乘和同态加的方案会发现，当需要对两个不同公钥加密的密文做乘法时，两个云端之间交互和传输的过程要比做加法时更加复杂，而且需要传输的数据量也会更多一些，从而致使计算同态乘所需要的时间比计算同态加所需要的时间更长。

最后对设计的具有隐私保护功能的 SVM 算法的分类准确度进行了测试，通过在不同大小的数据上利用本文设计的方案进行分类实验，可以得到具体的分类准确度和消耗的时间，如表 4-4 所示，随着数据的增加，分类所需要的时间也在逐渐增加，另外分类的准确度也是可以达到 93%。而这个分类准确度主要是取决于所使用的 SVM 算法，另外本课题所提出的方案也可以应用到其他形式的 SVM 算法中，或者是其他的主要计算是代数运算的机器学习算法中。本方案主要是提出了

一个可以在多密钥加密的密文上做数据挖掘的工具，并可以保护用户的数据隐私、中间计算结果的隐私、分类模型的隐私和最后分类预测结果的隐私。

表 4-4. 分类准确度和对应消耗的时间

分类准确度和耗时						
数据集大小（kb）	100.00	200.00	400.00	600.00	800.00	1000.00
分类准确度（%）	92.51	93.02	92.49	92.68	93.41	92.89
分类耗时（s）	787.12	1467.17	2787.43	4163.12	5356.21	6521.32

4.2 实验比较

为了更好地分析本文设计的方案性能，我们通过实验和论文[10]设计的方案进行了对比分析，论文[10]设计的方案是目前最新的利用支持向量机算法做具有隐私保护功能的数据挖掘方案。并且在论文[10]设计的方案中也是通过两个云的框架实现了存储外包和计算外包。

首先是通过在不同大小的数据集上比较了加解密的效率如图 4-1 和图 4-2 所示：

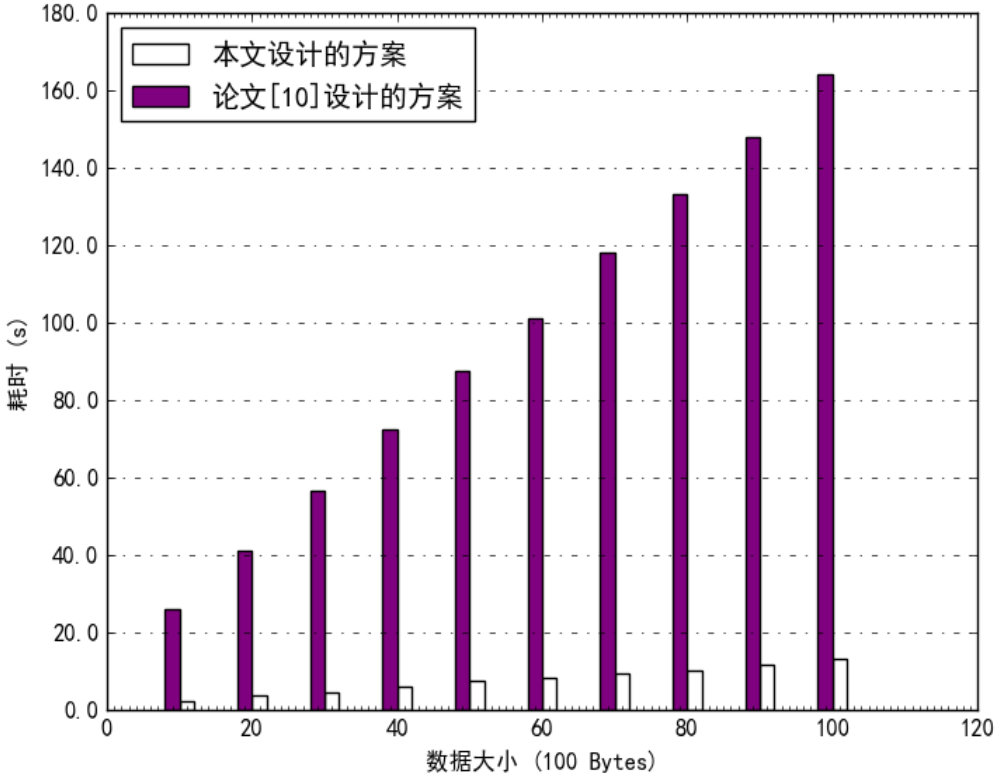

图 4-1. 加密时间对比

通过对图 4-1 进行分析可以发现，当数据量增加时，两种方案加密操作所花费的时间也在逐渐增加，不过在同样大小的数据上，本文设计的方案加密操作所消

耗的时间是小于论文[10]中的方案。

另外可以发现两种方案中加密消耗的时间和处理的数据大小之间是一个线性关系，通过对论文[10]的方案进行进一步地分析发现该方案在加密时消耗的时间主要是花费在产生系统的参数和公私钥的过程中，这是一个迭代的过程，为了找到一个合适的满足条件的安全参数需要进行多轮迭代，因此消耗的时间也随着迭代的轮数增加而变大。而本课题所设计的方案中在加密阶段的主要耗时是在对数据的处理上，产生公私钥的时间花费是比较少的。

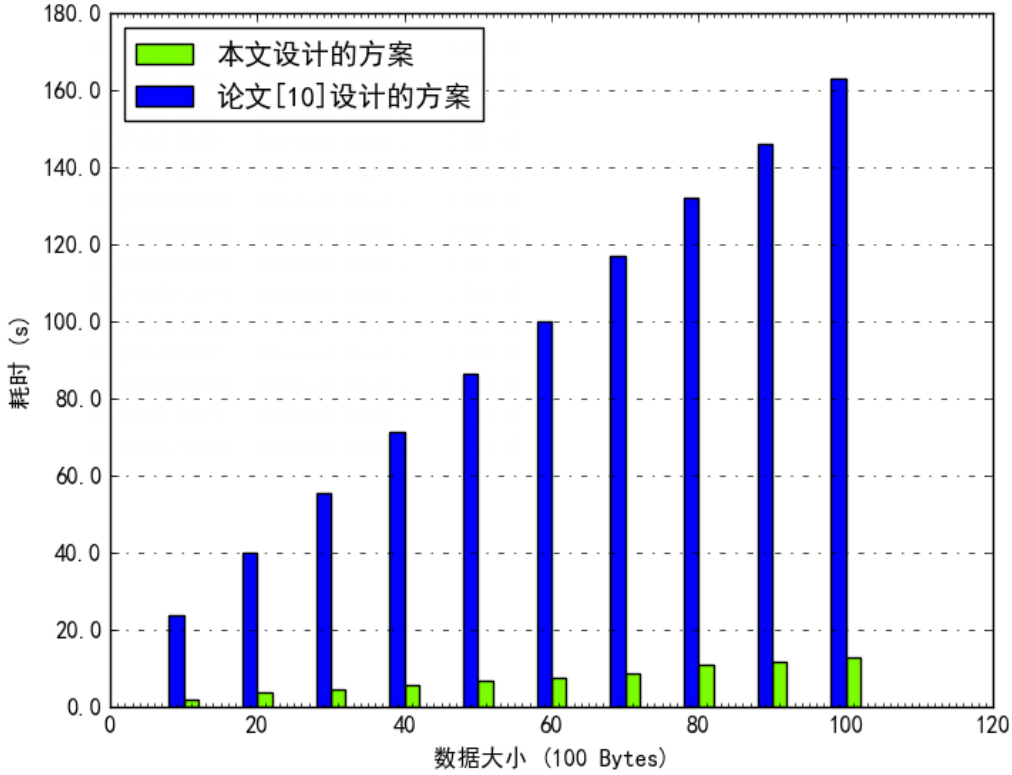

图 4-2. 解密时间对比

通过对图 4-2 进行分析也可以发现，随着数据量的增加，两种方案解密所消耗的时间也在逐渐加大，不过在同样的数据集上，本文设计的方案解密所消耗的时间是要小于论文[10]的方案，而且随着数据量的增加，这种差异表现的也越来越明显。同样可以发现这两种方案中的解密耗时和解密的数据大小之间是一个线性关系。

两种方案解密阶段的耗时差距主要是在算法的具体构造上，在论文[10]的方案中是以密文数据的变大为代价来完成在多密钥加密的数据上的计算，因此同样大小的数据利用本文设计的方案进行处理所产生的密文是要小于论文[10]的方案，而且论文[10]中的这种代价随着处理数据的增加而急剧变大，从而使得在解密阶段需

要消耗大量的时间。

另外在不同大小的数据集上测试了两种方案做同态加法和同态乘法计算所消耗的时间，如图 4-3 和图 4-4 所示:

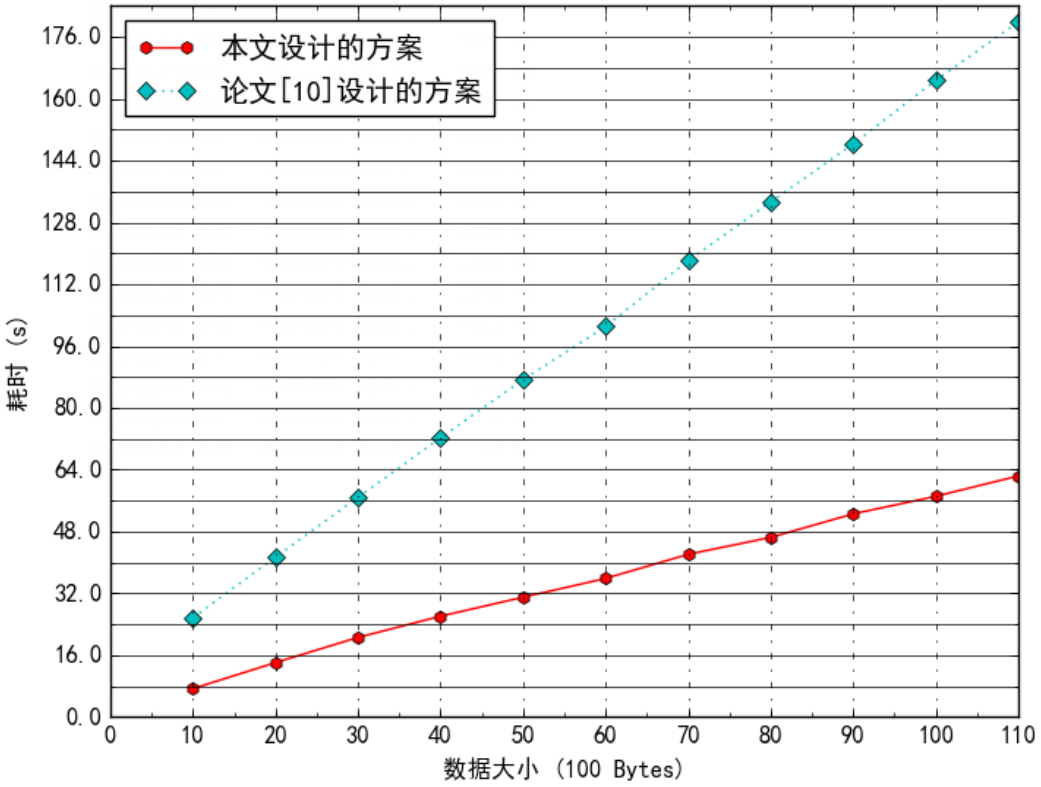

图 4-3. 同态乘耗时对比

通过对图 4-3 进行分析，可以发现随着数据量的增加，两种方案计算同态乘所消耗的时间都逐渐增大，而在同样的数据集上论文[10]的方案消耗的却比本文设计的方案要大，而且随着数据量的增加，这种差异也越来越明显。

两种方案做做同态乘法操作时，消耗的时间和处理数据集的大小之间是一个线性关系，在论文[10]中为了完成在多密钥加密的整数域上的乘法计算引入了一个转换的技术，通过转换来完成对多密钥数据的乘法计算，不过在转换的过程中也会造成密文数据的增加，从而导致计算的耗时也随着数据的增加而变得越来越大。特别是随着用户的增多导致密文数据增加的问题会越来越严重。

而对图 4-4 进行分析可以发现，随着数据量的增加，两种方案计算同态加的时间也在逐渐增大，不过在同样的数据集上，论文[10]的方案所需要消耗的时间依然比本文设计的方案要大，在数据量小的时候，这种差异不是很突出，不过随着数据量的变大，这种差异将更加突出。

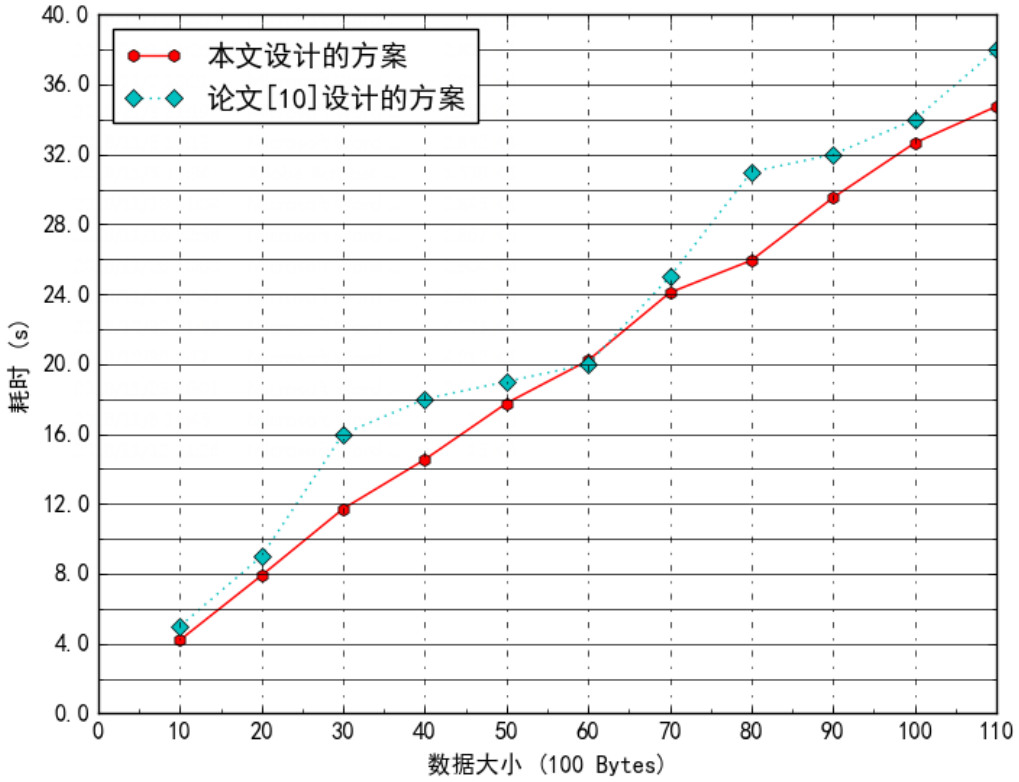

图 4-4. 同态加耗时对比

通过以上的分析可以看出在功能上本文所设计的方案可以在水平分布、垂直分布和任意分布的多密钥加密的整数域和有理数域上做数据挖掘，并且能够保护用户的数据隐私、中间计算结果的隐私、分类模型的隐私和分类预测结果的隐私。而在性能上，本文设计的算法在加密、解密、同态加和同态乘上的表现都优于同类型的算法。另外在本文所设计的算法中实现了存储和计算外包，因此用户不需要实时在线参与计算。

4.3 本章小结

本章主要对第三章提出的基于安全多方计算的隐私保护支持向量机算法进行了实现和分析。主要从两个方面对算法的实现进行了介绍，一方面从开发环境、训练与测试数据集和系统性能这三个方面对系统实现的过程和实验的结果进行了介绍和分析。另一方面又从功能和安全性上和其他利用支持向量机算法做具有隐私保护功能的数据挖掘方案进行了比较和总结。

结　论

随着信息技术的发展，人们为了更好地利用数据理解数据背后的知识，越来越多地把机器学习算法被应用到大数据挖掘中。而当数据涉及隐私或敏感信息时，假如直接在这些数据上进行数据挖掘会泄露数据的隐私信息。所以需要设计的方案能够保护数据的隐私。当数据量很大时，由于本地的存储和计算资源受限，很难在本地进行数据挖掘。而随着云计算的发展，可以借助云的存储和计算能力进行数据挖掘。但是云是由第三方提供的，是不完全可信的，如果直接在云上进行数据挖掘，会泄露用户数据的隐私信息，所以如何在不泄露数据隐私的前提下在云上进行数据挖掘是一个难题。本文针对这个问题提出了基于安全多方计算的具有隐私保护功能的支持向量机算法，该算法结合云实现了计算外包和存储外包的功能，可以显著地提高数据挖掘的效率。利用本文提出的算法进行数据挖掘，可以保护用户的数据隐私、中间计算结果的隐私、分类模型的隐私和最后分类预测结果的隐私。本文的研究成果主要包括以下几个部分：

（1）设计了支持在多密钥加密的数据上做加法和乘法计算的安全多方计算协议。由于当数据分布式地存储在多方时，为了保护数据的隐私，需要各方利用自己的公钥加密数据并上传到云上进行数据挖掘。通过对支持向量机算法在数据挖掘中的主要计算进行分析后，设计了支持在多密钥加密的数据上做隐私保护数据挖掘的同态加和同态乘的协议，而且这两个协议也可以推广应用到其他需要代数计算的机器学习算法中，进行隐私保护的数据挖掘。

（2）提出了可以在水平分布、垂直分布和任意分布的整数域和有理数域上利用支持向量机算法做具有隐私保护功能数据挖掘方案。在方案中针对多密钥加密的数据不能直接通过安全多方计算的方法进行处理的难题，以及当前的加密算法不支持对有理数域中的小数进行处理的问题，分别设计了方案进行解决。当数据分布式地存储在各方时，基于该方案可以在多密钥加密的整数域和有理数域上做隐私保护的数据挖掘，并且在挖掘的整个期间可以保护用户的数据隐私、中间计算结果的隐私、分类模型的隐私和最后分类预测结果的隐私。

（3）实现和比较分析了本文提出的基于支持向量机算法做具有隐私保护功能的数据挖掘方案。在系统实现的过程中通过把数据集分成训练集、验证集和测试集对方案的性能进行了测试。并和利用支持向量机做隐私保护数据挖掘的其他方案进行了比较分析，通过比较可以发现本文所提出的方案在整体上的性能表现要

优于其他方案。另外也对本文所提出的方案进行了安全分析和证明，当敌手和方案中的参与方存在合谋行为时，本文所提出的方案在半诚实模型中，可以在保护数据隐私的前提下进行数据挖掘。

此外，本文还有进一步的改进空间，首先如何进一步地提高方案的效率是一个难点。另外在本文设计的方案中，是通过两个云之间的交互来完成在多密钥加密的数据上做加法和乘法的计算，而且允许敌手可以和两个云之中的任何一个云进行合谋攻击。而当敌手可以同时和两个云进行合谋攻击时，如何在保护数据隐私的前提下进行数据挖掘是一个难点。以上的这两个难点都需要进一步的学习，通过优化或重新设计方案进行解决。

参考文献

[1] Yao A C C. How to Generate and Exchange Secrets[C]//Foundations of Computer Science, 27th Annual Symposium on. IEEE, 1986: 162-167.

[2] Cramer R, Damgård I, Maurer U. General Secure Multi-party Computation from any Linear Secret-Sharing Scheme[C]// International Conference on the Theory and Applications of Cryptographic Techniques. Springer, Berlin, 2001:280-300.

[3] Cramer R, Damgård I, Nielsen J B. Multiparty Computation From Threshold Homomorphic Encryption[C]//International Conference on the Theory and Applications of Cryptographic Techniques. Springer, Berlin, 2001:280-300.

[4] Rivest R, Shamir A, Adleman L. On Data Banks and Privacy Homomorphisms[J]. Foundations of Secure Computation, Georgia Institute of Technology, 1978, 21(2):169-180.

[5] Rivest R, Shamir A, Adleman L. A Method for Obtaining Digital Signatures and Public Key Cryptosystems[J]. Communications of the ACM, 1978, 21(2):120-126.

[6] Goldreich O, Micali S, Wigderson A. How to Play any Mental Game or A Completeness Theorem for Protocols with Honest Majority[J]. Proc Stoc, 1987:218-229.

[7] Paillier P. Public-Key Cryptosystems Based on Composite Degree Residuosity Classes[C]//International Conference on Theory and Application of Cryptographic Techniques. 1999: 223-238.

[8] Boneh D, Goh E J, Nissim K. Evaluating 2-DNF Formulas on Ciphertexts[C]//Theory of Cryptography Conference. Springer, Berlin, Heidelberg, 2005: 325-341.

[9] Lindell Y, Pinkas B. An Efficient Protocol for Secure Two-Party Computation in the Presence of Malicious Adversaries[J]. Journal of Cryptology, 2015, 28(2): 312-350.

[10] Zhang J, He M, Yiu S M. Privacy-Preserving Elastic Net for Data Encrypted by Different Keys-With an Application on Biomarker Discovery[C]//IFIP Annual Conference on Data and Applications Security and Privacy. Springer, 2017:185-204.

[11] Gentry C. Fully Homomorphic Encryption Using Ideal Lattices[J]. Stoc, 2009, 9(4):169-178.

[12] Kamara S, Mohassel P, Raykova M. Outsourcing Multi-Party Computation[J]. IACR Cryptology ePrint Archive, 2011, 2011(3):435-451.

[13] Brakerski Z, Gentry C, Vaikuntanathan V. (Leveled) Fully Homomorphic Encryption without Bootstrapping[J]. ACM Transactions on Computation Theory

(TOCT), 2014, 6(3):1-36.

[14] Regev O. On Lattices, Learning with Errors, Random Linear Codes, and Cryptography[J]. Journal of the ACM (JACM), 2009, 56(6): 34.

[15] López-Alt A, Tromer E, Vaikuntanathan V. On-the-fly Multiparty Computation on the Cloud via Multikey Fully Homomorphic Encryption[C]//Proceedings of the Forty-Fourth annual ACM Symposium on Theory of Computing. ACM, 2012:1219-1234.

[16] Carter H, Mood B, Traynor P, Butler K. Secure Outsourced Garbled Circuit Evaluation for Mobile Devices[C]//Usenix Conference on Security. USENIX Association, 2013:289-304.

[17] Gentry C, Sahai A, Waters B. Homomorphic Encryption from Learning with Errors: Conceptually-Simpler, Asymptotically-Faster, Attribute-Based[C]//Advances in Cryptology-CRYPTO. Springer, Berlin, 2013:75-92.

[18] Jakobsen T P, Nielsen J B, Orlandi C. A Framework for Outsourcing of Secure Computation[C]//Proceedings of the 6th edition of the ACM Workshop on Cloud Computing Security. ACM, 2014:81-92.

[19] Brakerski Z, Perlman R. Lattice-based Fully Dynamic Multi-key FHE with Short Ciphertexts[C]//Annual Cryptology Conference. Springer, Berlin, Heidelberg, 2016: 190-213.

[20] Furukawa J, Lindell Y, Nof A, et al. High-throughput Secure Three-party Computation for Malicious Adversaries and an Honest Majority[C]//Annual International Conference on the Theory and Applications of Cryptographic Techniques. Springer, Cham, 2017: 225-255.

[21] Alagic G, Dulek Y, Schaffner C, et al. Quantum Fully Homomorphic Encryption with Verification[C]//International Conference on the Theory and Application of Cryptology and Information Security. Springer, Cham, 2017: 438-467.

[22] Cheon J H, Han K, Kim A, et al. Bootstrapping for Approximate Homomorphic Encryption[C]//Annual International Conference on the Theory and Applications of Cryptographic Techniques. Springer, Cham, 2018: 360-384.

[23] Xu L, Jiang C, Wang J, et al. Information Security in Big Data: Privacy and Data Mining[J]. IEEE Access, 2014, 2: 1149-1176.

[24] 康海燕, 马跃雷. 差分隐私保护在数据挖掘中应用综述[J]. 山东大学学报(理学版), 2017, 52(3) : 16-23.

[25] Lindell Y, Pinkas B. Privacy Preserving Data Mining[C]//International Cryptology Conference on Advances in Cryptology. 2000: 36-54.

[26] Vaidya J, Clifton C. Privacy-Preserving K-means Clustering over Vertically Partitioned Data[C]//Proceedings of the Ninth ACM SIGKDD International

Conference on Knowledge Discovery and Data Mining. ACM, 2003: 206-215.

[27] Lin X, Clifton C, Zhu M. Privacy Preserving Clustering with Distributed EM Mixture Modeling[J]. Knowledge and Information Systems, 2005, 8(1): 68-81.

[28] Yu H, Jiang X, Vaidya J. Privacy-preserving SVM Using Nonlinear Kernels on Horizontally Partitioned Data[C]//Proceedings of the 2006 ACM symposium on Applied computing. ACM, 2006: 603-610.

[29] Yu H, Vaidya J, Jiang X. Privacy-preserving SVM Classification on Vertically Partitioned Data[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Berlin, 2006:647-656.

[30] Vaidya J, Yu H, Jiang X. Privacy-preserving SVM Classification[J]. Knowledge and Information Systems, 2008,14(2): 161-178.

[31] Samet S, Miri A. Privacy Preserving ID3 Using Gini Index over Horizontally Partitioned Data[C]//IEEE ACS International Conference on Computer Systems and Applications. 2008: 645-651.

[32] Vaidya J, Kantarcıoğlu M, Clifton C. Privacy-Preserving Naive Bayes Classification[J]. The International Journal on Very Large Data Bases, 2008, 17(4): 879-898.

[33] Hu Y, He G, Fang L, et al.: Privacy-Preserving SVM Classication on Arbitrarily Partitioned Data[C]//IEEE International Conference on Progress in Informatics and Computing. 2010:543-546.

[34] Skarkala M E, Maragoudakis M, Gritzalis S, et al. Privacy Preserving Tree Augmented Naive Bayesian Multi-Party Implementation on Horizontally Partitioned Databases[C]//International Conference on Trust, Privacy and Security in Digital Business. 2011: 62-73.

[35] Lory P. Enhancing the Efficiency in Privacy Preserving Learning of Decision Trees in Partitioned Databases[C]//International Conference on Privacy in Statistical Databases. 2012: 322-335.

[36] Ma X, Li J, Zhang F. Efficient and Secure Batch Exponentiations Outsourcing in Cloud Computing[C]//International Conference on Intelligent Networking and Collaborative Systems (INCoS). IEEE, 2012: 600-605.

[37] Lei X, Liao X, Huang T, et al. Outsourcing Large Matrix Inversion Computation to A Public Cloud[J]. IEEE Transactions on Cloud Computing, 2013, 1(1): 1-1.

[38] 任艳丽, 蔡建兴, 黄春水等. 基于身份加密中可验证的私钥生成外包算法[J]. 通信学报, 2015, 36(11): 61-66.

[39] 刘晓燕. 基于安全多方计算的隐私保护 K-means 聚类算法的外包计算[D]. 哈尔滨：哈尔滨工业大学，2015.

[40] Zhang J, Yang Y, Wang Z. Outsourcing Large-Scale Systems of Linear Matrix

Equations in Cloud Computing[C]//2016 IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS). IEEE, 2016: 438-447.

[41]孙茂华, 宫哲. 一种保护隐私集合并集外包计算协议[J]. 密码学报, 2016, 3(2): 114-125.

[42]张兴兰, 刘祥. 安全高效的可验证大型线性方程组求解外包计算方案[J]. 网络与信息安全学报, 2017, 3(6): 1-7.

[43]蔡建兴, 任艳丽. 大型线性方程组求解的可验证外包算法[J]. 计算机应用研究, 2017, 34(2)536-538.

[44]Peter A, Tews E, Katzenbeisser S. Efficiently Outsourcing Multiparty Computation Under Multiple Keys[J]. IEEE Transactions on Information Forensics & Security, 2013, 8(12): 2046-2058.

[45]Liu D, Bertino E, Yi X. Privacy of Outsourced K-means Clustering[C]//Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security. ACM, 2014: 123-134.

[46]Liu F, Ng W K, Zhang W. Encrypted SVM for Outsourced Data Mining[C]//Cloud Computing (CLOUD), IEEE 8th International Conference on. IEEE, 2015:1085-1092.

[47]靳亚宾. 云环境下具有隐私保护的 K-means 聚类算法研究与设计[D]. 哈尔滨：哈尔滨工业大学, 2016.

[48]Li L, Lu R, Choo K K R, et al. Privacy-Preserving-Outsourced Association Rule Mining on Vertically Partitioned Databases[J]. IEEE Transactions on Information Forensics & Security, 2016, 11(8): 1847-1861.

[49]Zhang J, Wang X, Yiu S M, et al. Secure Dot Product of Outsourced Encrypted Vectors and its Application to SVM[C]//Proceedings of the Fifth ACM International Workshop on Security in Cloud Computing. ACM, 2017:75-82 .

[50]Bresson E, Catalano D, Pointcheval D. A Simple Public-key Cryptosystem with a Double Trapdoor Decryption Mechanism and its Applications[C]//International Conference on the Theory and Application of Cryptology and Information Security. Springer, Berlin, 2003:37-54.

[51]Li P, Li J, Huang Z, et al. Multi-key Privacy-preserving Deep Learning in Cloud Computing[J]. Future Generation Computer Systems, 2017, 74: 76.

[52]Liu X, Deng, R H, Choo, K K R, et al. An Efficient Privacy-preserving Outsourced Calculation Toolkit with Multiple Keys[J]. IEEE Transactions on Information Forensics and Security, 2016, 11(11): 2401-2414.

[53]Syed N, Liu H, Sung K. Incremental Learning with Support Vector Machines[C] //International Joint Conference on Artificial Intelligence. Sweden: Morgan Kaufmann Publishers, 1999: 352-356.

[54] Boser B E, Guyon I M, Vapnik V N. A Training Algorithm for Optimal Margin Classifiers[C]//Proceedings of the Fifth Annual Workshop on Computational Learning Theory. New York: ACM Press, 1992: 144-152.

[55] Osuna E, Frenud R, Girosi F. An Improved Training Algorithm for Support Vector Machines[C]//Proceedings of IEEE Workshop on Neural Networks for Signal Processing. USA, 1997: 276-285.

[56] PLATT J C. 12 Fast Training of Support Vector Machines Using Sequential Minimal Optimization[J]. Advances in Kernel Methods, 1999:185-208.

[57] Knuth G. The Art of Computer Programming, Seminumerical Algorithms, Vol. 2, Addition Wesley[J]. Reading, Massachusetts, 1998.

[58] Shamir, A. How to Share a Secret[J]. Communications of the ACM, 1979, 22(11): 612-613.

[59] Li C, Ma W. Comments on An Efficient Privacy-Preserving Outsourced Calculation Toolkit With Multiple Keys[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(10): 2668-2669.

攻读硕士学位期间发表的学术论文及其他成果

学术论文

[1] **Wenli Sun**, Zoe L. Jiang, Jun Zhang, S. M. Yiu, et al. Outsourced Privacy Preserving SVM with Multiple Keys[C]// The 18th International Conference on Algorithms and Architectures for Parallel Processing(ICA3PP 2018). Guangzhou, 2018:415-430.（EI 检索，CCF C 类）

[2] Zoe L. Jiang, **Wenli Sun**, Jun Zhang, S. M. Yiu, et al. Outsourced Privacy Preserving Data Mining on Integer and Rational Data Using SVM [J]. Sensor, 2018.（已投，SCI 检索，IF 2.475）

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《基于安全多方计算的隐私保护支持向量机算法研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：孙文礼　　日期：2019 年 1 月 2 日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：孙文礼　　日期：2019年 1月2 日

导师签名：薛琳林　　日期：2019年 1月2日

致　谢

流光如箭，硕士研究生的旅程马上就要走到终点了，回想起这两年半的旅程心中思绪万千，往事一幕幕的在脑海中闪现有欢声、有笑语、有失意也有收获，不管怎么样，时光都无法倒流，无法让我回到旅行的起点重新开始。此时此刻感触最深的就是一分耕耘一分收获，另外在这里也很想对陪我一路走来的老师、同学和家人说一声感谢。

首先，我要感谢我的导师蒋琳老师。蒋老师在我硕士生涯给了我很多帮助，尤其在本课题的研究中，她牺牲了很多自己的休息时间来指导我的课题，从开题、中期、再到答辩一次次不厌其烦地指出我课题中存在的问题，帮助我修改论文。而且在每次的组会中，蒋老师都会指导我如何做汇报。在我写小论文遇到挫折的时候，蒋老师也会细心地帮我梳理思路调整论文结构。在生活中，蒋老师给予了我很多的关心和帮助，不仅指导了我高效率做事的方法，更教会了我许多为人处世的道理。蒋老师潜心钻研的科研精神，严格认真的工作态度都是值得我终身学习的。

其次，我要感谢王轩老师，感谢他在学习和生活中给我的指导和帮助。在研究中心这两年多的时间里我收获很多，在王老师的指导下，我从一名不是很懂科研的本科生，一步一步成长起来懂得了如何写项目申请和项目的结题报告，熟悉了项目的完整开发流程并明白和掌握了写一篇好论文的方法，另外还养成了每周写周报，总结自己每周工作的好习惯。特别是在做“云计算环境用户数据隐私保护关键技术”这个项目时，王老师会经常给我们开会，耐心地指导我们如何做项目，写文档。在找工作期间，王老师会耐心地帮助我们做规划，修改简历，并传授我们找工作的一些经验，这些对我的帮助都很大。即将离开校园步入社会，我会时刻要求自己传承和发扬哈工大“规格严格，功夫到家”的校训。

另外我还要感谢信息安全组的同学们，每一次的组会都使我受益匪浅。当然，我还要感谢研究中心十六级的同学们，在学校的两年半里，我们共同学习共同进步。即将毕业，希望在以后的日子里，我们能携手拼搏出属于自己的一片天。

最后，我要感谢我的家人。虽然我独自在外求学，但是我的成长和进步离不开他们的鼓励和支持。每当我遇到挫折时，是他们作为我的精神支柱不断地鼓励着我克服困难。他们是让我不断奋斗，不断前行的动力和源泉。