同态加密隐私保护数据挖掘方法综述*

钱 萍[1a,2]，吴 蒙[1b]

(1. 南京邮电大学 a. 计算机学院; b. 通信与信息工程学院，南京 210003; 2. 江苏科技大学 计算机科学与工程学院，江苏 镇江 212003)

摘 要: 近年来国内外学者对隐私保护数据挖掘(privacy-preserving data mining，PPDM)进行了大量研究，适时地对研究成果进行总结，能够明确研究方向。从分类挖掘、关联规则挖掘、聚类挖掘和安全多方计算等几个方面，总结了现有的基于同态加密技术的算法，分析了其基本原理和特点，并在此基础上指出了 PPDM 技术今后发展的方向。

关键词: 隐私保护; 数据挖掘; 同态加密; 安全多方计算

中图分类号: TP311 **文献标志码:** A **文章编号:** 1001-3695(2011)05-1614-04

doi:10.3969/j.issn.1001-3695.2011.05.004

Survey of privacy preserving data mining methods based on homomorphic encryption

QIAN Ping[1a,2], WU Meng[1b]

(1. *a. College of Computer Science, b. College of Telecommunications & Information Engineering, Nanjing University of Posts & Telecommunications, Nanjing* 210003, *China*; 2. *College of Computer Science & Technology, Jiangsu University of Science & Technology, Zhenjiang Jiangsu* 212003, *China*)

Abstract: In recent years, researchers at home and abroad have done much work on privacy-preserving data mining(PPDM), so summarizing the research results timely can make research directions clear. This paper surveyed the general PPDM methods based on homomorphic encryption and analyzed their theories and characteristics. Finally, discussed future research directions on PPDM.

Key words: privacy preservation; data mining; homomorphic encryption(HE); secure multi-party computation

数据挖掘是一类深层次的数据分析方法，可被用于科学研究或者商业领域，对大量数据进行较为复杂的分析和建模，发现各种规律和有用的信息。不论科学研究还是商业应用，数据挖掘都取得了可喜的成果。但与此同时，数据挖掘也面临着很多问题的挑战。其中，数据挖掘的个人隐私与信息安全问题尤其受到关注。误用、滥用数据挖掘可能导致用户数据特别是敏感信息的泄露，越来越多的人们对此表示担忧，甚至拒绝提供真实的数据。如何在不暴露用户隐私的前提下进行数据挖掘，也就成为人们非常感兴趣的课题。

在 1995 年召开的第一届知识发现(knowledge discovery in database，KDD)与数据挖掘国际学术会议上，隐私保护的数据挖掘就成为了一个专门的研究主题[1]。1999 年 Rakesh Agrawal 在 KDD99 上进行了一场精彩的主题演讲[2]，将隐私保护的数据挖掘作为未来的研究重点之一。自此以后，隐私保护的数据挖掘越来越受到人们的重视，迅速成为近年来数据挖掘领域研究的热点之一。

1 数据挖掘算法

目前主要的数据挖掘隐私保护方法划分为数据修改和数据加密两类。基于数据修改的保护方法中，主要采用数据扰乱技术，该类方法通过对原始数据的局部或全局修改隐藏原始隐私数据或规则，会在一定程度上造成原始数据的损失，影响挖掘的准确性，并且所有经过干扰的数据均与真实的原始数据直接相关，降低了对隐私数据的保护程度。基于数据加密的保护方法中，通过密码机制实现了他方对原始数据的不可见性以及数据的无损失性，因此可以实现与原始挖掘同样准确度的挖掘结果，并实现隐私保护的目的。本文研究基于数据加密的隐私保护数据挖掘算法，特别是针对基于同态加密技术的算法。

2 同态加密技术

同态加密(homomorphic encryption，HE)最初由 Rivest 等人[3]于 1978 年提出，是一种允许直接对密文进行操作的加密变换技术。但是由于其对已知明文攻击是不安全的，后来由 Domingo-Ferrer 等人[4]作了进一步的改进。HE 技术最早用于对统计数据进行加密，由算法的同态性保证了用户可以对敏感数据进行操作但又不泄露数据信息。

秘密同态技术建立在代数理论之上，其基本思想如下:

假设 E_{k_1} 和 D_{k_2} 分别代表加密和解密函数，明文数据是有限集合 $M=\{m_1,m_2,\cdots,m_n\}$，α 和 β 代表运算，若

$$\alpha(E_{k1}(m_1),E_{k1}(m_2),\cdots,E_{k1}(m_n))=E_{k1}(\beta(m_1,m_2,\cdots,m_n)) \quad (1)$$

收稿日期: 2010-12-06; 修回日期: 2011-01-15 基金项目: 国家“973”计划资助项目(2011CB302903)

作者简介: 钱萍(1978-)，女，江苏镇江人，讲师，博士研究生，主要研究方向为网络安全、隐私保护(qptracy@163.com)；吴蒙(1963-)，男，教授，博导，主要研究方向为无线通信、信息安全.

成立,则称函数族(E_{k1} , D_{k2} , α , β) 为一个秘密同态。秘密同态技术近年来得到学术领域广泛的重视,国外很多学者对此作了深入研究[5],而国内学者也对秘密同态技术作了系统研究。杨勇等人[6]提出了同态加密的具体算法; 尹春勇等人[7~9]对秘密同态进行了改进和优化; 武汉大学的研究者把同态加密扩展到实数,操作扩展到加法、减法、乘法和除法[10,11]。

本文将根据数据挖掘的分类方法,以分类挖掘、关联规则挖掘和聚类挖掘三个方面为主,并结合其他分类方法,对基于同态加密技术的隐私保护数据挖掘算法进行介绍和总结。

3　基于同态加密技术的隐私保护

现有的面向数据挖掘的隐私保护方法主要是结合分类挖掘、聚类挖掘、关联规则挖掘这三种挖掘方法来实现的,另外对于基于加密技术的隐私保护方法、多方安全计算方法可以看成是其中的一种特例,还有一些隐私保护算法是综合了几种加密技术和数据挖掘技术。本章将从这样几个方面展开讨论,研究采用了同态加密技术的各种隐私保护算法。

另外需要说明的是,本章中讨论的都是在分布式环境下的隐私保护方法。分布式环境下的数据隐私表示为分布单元在整个数据挖掘过程中对自身持有数据实际数值的唯一可知性。分布式应用采用两种模式存储数据: 垂直划分(vertically partitioned) 和水平划分(horizontally partitioned) 。垂直划分数据是指分布式环境中每个站点只存储部分属性的数据,所有站点存储的数据不重复; 水平划分数据是将数据记录存储到分布式环境中的多个站点,所有站点存储的数据不重复。

3.1　基于同态加密技术的隐私保护分类挖掘算法

分类挖掘算法是数据挖掘中常用的一类方法。分类是这样一个过程,它找出描述并区分数据类或概念的模型或函数,以便能够使用模型预测类标记未知的对象类。导出模型时基于对训练数据集(即其类标记已知的数据对象) 的分析。分类的目标就是要构造一个分类模型,从而预测未来的数据趋势。从目前来看,分类采用的方法主要有决策树、神经网络、贝叶斯算法和 KNN 算法等。而基于隐私保护的分类技术则是要在数据挖掘过程中建立一个没有隐私泄露的准确的分类模型。

3.1.1　决策树分类

决策树分类是数据挖掘中的一种常用方法,以其挖掘结果易理解、精度高和鲁棒性好而著称。一般来说,决策树的建立基于一个从顶而下的模式,基于贪婪策略选择子树的根。一个决策树就是一个类别分类器,它递归地隔开训练集,直到每一个隔开的部分都由来自同一个类的样例组成。即从“选择最佳分裂属性作为根节点被测试”开始,基于训练样例集 S,使用统计分裂规则(如信息增益、增益率和 Gini 指标) 来确定分类能力强的属性作为根节点的测试; 然后为根节点属性的每个可能值产生一个分支,把训练样例集 S 排列到适当的分支(即样例的该属性值对应的分支下) ; 再重复该过程,用每个分支节点关联的训练样例来选取在该点被测试的最佳分裂属性。ID3 就是一个很著名的建立决策树分类的算法[12]。隐私保护的决策树挖掘就是在不精确访问原始数据集的条件下,尽可能准确地构造出分类决策树。

Zhan 等人在文献[13,14]中提出了一种同态加密和数字信封的方法以实现合作决策树分类,参与的合作方不需要分享私有数据,并分别给出了水平多方合作[13]和垂直多方合作[14]情况下决策树分类的例子和解决方案。

Fang 等人在文献[15,16]中也提出了一种基于同态加密的决策树挖掘方法,主要研究在多个水平划分数据库(即数据集按记录分布在不同站点) 的联合样本集上实现隐私保护的决策树挖掘。通过计算器端直接在加密数据上计算加密的全局统计信息,半可信第三方挖掘者在解密后的全局统计信息上进行决策树构建,从而实现了原始信息的隐私保护。

下面以水平分布情况为例,介绍决策树隐私保护挖掘的方法。图 1 是总体架构图。

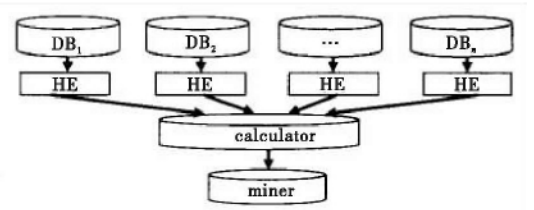

图1　分布式决策树隐私保护挖掘的总体架构图

图 1 中 HE 代表同态加密 HE。假设该架构模型满足以下条件:

a) 所有数据记录以水平划分形式存储在各站点,且各站点之间不进行任何通信;

b) 计算器只负责按照固定公式执行计算,对于各站点数据库信息、参加计算的数据及结果的含义均一无所知;

c) 挖掘服务器负责数据挖掘,并将最终结果传输给各站点,而对于各站点的数据库信息一无所知。

在每次产生最佳分裂属性时,分为三步完成:

a) 各水平划分数据库在本地计算局部统计信息,并利用同态加密技术进行加密,然后传输给计算器;

b) 计算器获得加密信息后,按照固定公式求和,并传输给半可信第三方挖掘者;

c) 半可信第三方挖掘者依据同态解密算法进行解密,选择全局 Gini 值最小的属性作为最佳分裂属性,然后将其结果发送给各节点。

产生当前最佳分裂属性后,各节点将本地训练集 RS 排列到适当的分支(即样例的该属性值对应的分支下) ,再重复该过程。

3.1.2　Naive Bayesian 分类

Naive Bayesian 分类方法是分类领域最成功的分类算法之一,虽然大家还不能很好地解释 Naive Bayesian 分类器之所以具有如此好的性能原因,但结果却是很明显的,Naive Bayesian 分类器可以成为未来分类器发展的一个绝佳起点。

贝叶斯分类器是基于贝叶斯公式,即

$$P(C|X) = \frac{P(C|X) \times P(C)}{P(X)} \tag{2}$$

其中: $P(C|X)$ 为条件 X 下 C 的后验概率; $P(C)$ 为 C 的先验概率; $P(X|C)$ 为条件 C 下 X 的后验概率; $P(X)$ 为 X 的先验概率。

贝叶斯分类器的预测依据就是取后验概率最大的类别。

Raju 等人在文献[17]中提出了基于贝叶斯分类的隐私保护挖掘算法,可用于多方分布式计算。该算法中没有能够访问所有数据的中心信任方,从而保证了数据的私有性。为了解决多方计算中隐私泄露问题,文中首先研究了用于隐私保护的加法和乘法运算,接着提出了基于同态加密方法和数字信封技术

的加法和乘法协议，通过该协议可以解决 Naive Bayesian 分类挖掘过程中的隐私保护问题。

3.1.3 KNN 分类

KNN 算法是数据挖掘技术中比较常用的分类算法，由于其实现的简单性，在很多领域得到了广泛的应用。KNN 分类算法的主要思想是：先计算待分类样本与已知类别的训练样本之间的距离或相似度，找到距离或相似度与待分类样本数据最近的 K 个邻居；再根据这些邻居所属的类别来判断待分类样本数据的类别。如果待分类样本数据的 K 个邻居都属于一个类别，那么待分类样本也属于这个类别；否则，对每一个候选类别进行评分，按照某种规则来确定待分类样本数据的类别。

Kumar 等人[18]提出了基于最近邻居查找的隐私保护方法，并利用同态加密技术在数据用户终端对私有数据进行加密。尽管训练样本中的数据是加密后的数据，根据同态加密的性质，这些数据仍然可以用于计算最近邻居。在计算过程中，需要第三方参与，该第三方被认为是一个准诚信方。

Zhu Jian-ming 在文献[19]中提出了基于同态加密和 EIGamal 加密系统的隐私保护 KNN 分类挖掘方法，在计算 KNN 时利用同态加密技术以保证数据的隐私性，该方法也是假定在准诚信的环境下。

3.2 基于同态加密技术的隐私保护关联规则挖掘算法

关联规则挖掘由 Agrawal 等人在 1993 年提出，其意义是寻找在同一事件中出现的不同项的相关性，即找出事件中频繁发生的项或属性的所有子集，以及它们之间应用的相互关联性。关联规则挖掘的对象一般是大型事务数据库，目的是发现大量数据中项集之间有趣的关联或相关联系，最近几年已被业界所广泛研究。

规则支持度和置信度是关注规则中的两个重要概念，它们分别代表了所发现规则的有用性和确定性。规则 $A \Rightarrow B$ 在事务数据库 D 中成立，具有支持度 support。其中 support 是 D 中事务包含 $A \cup B$（即 A 和 B 两者）的百分比，它是概率 $P(A \cup B)$。规则 $A \Rightarrow B$ 在事务集 D 中具有置信度 confidence，D 中包含 A 的事务同时也包含 B 的百分比是 confidence，这是条件概率 $P(B|A)$。关联规则挖掘就是在事务数据库 D 中找出具有用户给定的最小支持度阈值（min_sup）和最小置信度阈值（min_conf）的规则。

关联规则的挖掘由两步过程组成：

a）找出所有频繁项集。找出存在于事务数据库中所有项集 x，其支持度 support（x）和置信度 confidence（x）分别不小于用户给定的 min_sup 和 min_conf。

b）由频繁项集产生强关联规则。根据定义，这些规则必须满足最小支持度和最小置信度。

在步骤 a）的基础上，b）相对容易实现，因为它与事务数据库 D 无关。因此关联规则数据挖掘的关键在于找到正确的频繁项集。

文献[20～22]中分别给出了利用同态加密技术实现隐私保护关联规则挖掘的算法。文献[20]中给出了用于水平分布处理的隐私保护关联规则挖掘协议 P3ARM（privacy-preserving protocol for association rule mining），该协议的关键思想是利用同态加密技术，在加密的条件下获得项集支持度。文献[21]中结合 RSA 公钥密码系统和同态加密，提出在关联规则挖掘算法中加入密码管理模型，在挖掘的过程中引入加密管理服务器和数据挖掘服务器，利用事务的相似矩阵产生全局 K 频繁项集。Zhan 等人[22]提出了在安全两方或多方计算的情况下，利用同态加密技术实现关联规则挖掘，即在需要用到多方数据进行挖掘计算时，用同态加密算法对各方数据进行加密。

3.3 基于同态加密技术的隐私保护聚类挖掘算法

聚类是一个将物理或者抽象对象的集合分组组成由类似的对象组成的多个类的过程。由聚类所生成的簇是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其他簇中的对象相异。在许多应用中，一个簇中的数据对象可以被作为一个整体来对待。相似或不相似的度量是基于数据对象描述的取值来确定的，通常就是利用对象与对象之间的距离来进行描述的。聚类分析就是从给定的数据集中搜索数据对象之间所存在的有价值联系。在许多应用中，一个聚类中所有对象常常被当做一个对象来进行处理或分析。聚类的方法有很多，如 K-均值和 K-中心点是比较常用的聚类方法。

文献[23]中设计了一种隐私保护 K-均值聚类算法。在该算法的各个步骤中，参与方只知道聚类均值，而不知道其他方的私有信息。该算法的关键步骤是聚类均值的隐私保护计算，在该计算中用到了两种协议：基于健忘多项评估的协议和基于同态加密技术的信息。

3.4 同态加密技术与多方安全计算结合的隐私保护数据挖掘算法

安全多方计算 SMC（secure multi-party computation）起源于姚期智的百万富翁比较问题，主要研究无可信第三方情况下，如何无信息泄露地比较双方的数值。一个安全多方计算协议，如果对于拥有无限计算能力攻击者而言是安全的，则称做是信息论安全的或无条件安全的；如果对于拥有多项式计算能力的攻击者是安全的，则称为是密码学安全的或条件安全的。在隐私保护数据挖掘算法中，安全多方计算可以看成是基于加密技术的一个特例，用于支持分布式环境中隐私保持的挖掘工作，主要利用非对称加密机制形成交互计算的协议，实现无信息泄露的分布式安全计算。

对分布式环境下的站点（参与者），根据其行为可分为准诚信攻击者（semi-honest adversary）和恶意攻击者（malicious adversary）。准诚信攻击者是遵守相关计算协议但仍试图进行攻击的站点，恶意攻击者是不遵守协议且试图披露隐私的站点。一般地，假设所有站点为准诚信攻击者，准诚信环境是分布式计算中讨论较多的一个假设环境。准诚信环境中可能存在多个共谋节点，共享过程数据信息，以寻求发现其他节点的原始数据，该类攻击也被称为 collusion attack，即共谋攻击。

众多分布环境下基于隐私保护的数据挖掘应用都可以抽象为无信任第三方（trusted third party）参与的 SMC 问题，即怎样使两个或多个站点通过某种协议完成计算后，每一方都只知道自己的输入数据和所有数据计算后的最终结果。以在分布式下计算集合的并为例，假设有 N 个独立站点 $S_1, S_2, \cdots, S_N$，站点 S_i 拥有数据 D_i，这 N 个站点在不暴露每个站点具体数据情况下，计算出 $\bigcup_{i=1}^{N} D_i$。具体过程[24]如下所示：

输入：站点 $S_1, S_2, \cdots, S_N$；数据 $D_1, D_2, \cdots, D_N$。

输出：$\bigcup_{i=1}^{N} D_i$。

a) $\forall i, i \in [1, N]$,站点 i 产生私有密钥 K_i,且任意两个站点 i、j 的密钥满足可交换; $E_{K_i}(E_{K_j}(D)) = E_{K_j}(E_{K_i}(D))$。其中 D 为任意数据。

b) 对任意站点 i:

(a) 加密数据 D_i 并将 $E_{K_i}(D_i)$ 传递给其他站点。

(b) 将其他站点传递给自己的加密数据用 K_i 加密后传递给其他站点。

c) 任意数据 D_i 都被所有站点的密钥加密后,为 $E_{K_x}(\cdots E_{K_i}(D_i)\cdots)$,直到全部数据都被所有站点的私有密钥加过密后,去掉重复的数据项,将剩余的数据项再依次传递给所有站点解密。

d) 最后对数据进行解密的站点将得到 $\bigcup_{i=1}^{N} D_i$。

Ouyang Wei-min 等人在文献[25,26]中分别提出了在两方安全计算和多方安全计算的情况下实现连续模式挖掘的方法,即在两方或多方的情况下,每个参与方在不泄露自己私有数据的前提下,通过合作计算发现事物的连续模式。连续模式挖掘与关联规则挖掘类似,不同之处在于,关联规则是找出事务的内在联系,如 75% 买了商品 A 的顾客也可能买商品 B,而连续模式挖掘是尽力找出事务之间的联系。为了保证数据的私有性,两篇文章在合作计算的过程中都使用了同态加密技术进行加密。

3.5　其他综合的隐私保护数据挖掘算法

除了上面讨论的各类隐私保护数据挖掘方法,还有一些结合了多种隐私保护和数据挖掘技术的综合算法,这些算法一般都是结合了多种加密技术和数据挖掘算法,以便更好地实现隐私保护与数据挖掘之间的平衡。

Li 等人[27]提出了在数据挖掘的过程中结合扰乱技术和同态加密技术,利用扰乱技术可以保证隐私保护算法的效率,而在扰乱技术中使用同态加密技术可以更进一步保证数据的隐私性。他[28]还结合同态加密技术,提出了欧式空间下的三方及多方安全距离比较协议,该协议能够直接支持基于欧氏距离及其扩展的数据挖掘方法,如 K-近邻、K-均值等,在分类挖掘、聚类挖掘、Web 挖掘等领域具有通用性和普遍意义。

文献[29]将同态加密与保序加密相结合,提出了一种新的支持隐私保护的分类算法。该算法在保证挖掘结果的前提下,解决了密文的数学运算和数值比较问题,满足挖掘的隐私保护需求,并且降低了算法的通信和计算复杂度。文中还利用该加密技术,通过改变数据垂直分布条件下的聚类步骤,提出了一种新的聚类方法,并将该方法应用到了 K-中心聚类算法当中,在实现保护数据隐私安全的同时,能进行有效的聚类挖掘。

Ge 等人在文献[30]中将同态加密技术、阈值加密技术与安全多方计算相结合,提出了用于分布式关联规则挖掘的抵抗共谋攻击(collusion-resistance)协议。由于该协议使用了阈值同态加密技术,与其他文献相比,该协议不仅可用于准诚信环境,还可用于恶意环境。

4　结束语

面向数据挖掘的隐私保护是隐私保护研究领域的一个重要分支,对该领域的研究已取得了不少成果。如何既保证数据挖掘的正确性,又实现隐私数据的保密性,成为研究人员要解决的重要问题。利用同态加密技术可以较好地解决这一问题,该方法已经被用到许多不同类型的数据挖掘算法中,如分类挖掘、关联规则挖掘和聚类挖掘等,这些方法还可以与其他加密技术结合,如多方安全计算等。本文正是从这几个方面对基于同态加密技术的、面向数据挖掘的隐私保护方法进行了综述。

从文中的介绍和分析可以看出,利用同态加密技术的同态性质,可以保证数据挖掘的计算在隐私数据加密的情况下进行,而不会影响数据挖掘的结果。这类方法特别适用于分布式环境中需要进行多方计算的情况,即需要多方合作进行计算,但任何一方都只知道自己的私有数据,每一方的私有数据不会被泄露给其他参与方,且不存在可以访问任何参与方数据的中心信任方。这一类方法的研究目前仍处在起步阶段,今后可以从这样几个方面进一步展开研究:

a) 由于同态加密技术一般都基于公钥密码体制,其算法复杂度通常要高于其他基于共享密钥的加密技术,也高于一般的扰乱技术。因此,如何有效地降低算法的复杂度和提高效率,是今后研究的一个重要方向。

b) 目前常用的同态加密技术一般只能实现加法和乘法两种运算,显然这两种运算对于更多的数据挖掘方法是不够的,如何进一步扩展同态技术的运算种类是值得研究的课题。

c) 现有的方法一般都只局限于某一种数据挖掘方法的隐私保护,不具有通用性。设计适用于多种数据挖掘方法的算法将是今后研究的热点。

参考文献:

[1] KDD 1995 proceedings [EB/OL]. (1995). http://www.aaai.org/Press/Proceedings/kdd95.php.

[2] KDD 1999 proceedings [EB/OL]. (1999). http://www.sigkdd.org/kdd1999/1999.

[3] RIVEST R L, ADLEMAN L, DETROUZOS M L. On data banks and privacy homomorphism [C]//Proc of Foundations of Secure Computation. New York: Academic Press, 1978: 169-179.

[4] DOMINGO-FERRER J, HERRERA-JOANCOMARTI J. A new privacy homomorphism and applications [J]. Information Processing Letters, 1996, 60(5): 277-282.

[5] YU Yu, LEIWO J, PREMKUMAR B. A study on the security of privacy homomorphism [C]//Proc of the 3rd International Conference on Information Technology. Washington DC: IEEE Computer Society, 2006: 470-475.

[6] 杨勇,方勇,周安民. 秘密同态技术研究及其算法实现 [J]. 计算机工程,2005,31(2): 157-159.

[7] 尹春勇,李秦伟,孙汝霞. 一种秘密同态的改进模式 [J]. 贵州工业大学学报,2007,36(1): 68-70.

[8] 尹春勇,李坚石,李少波. 一种结合秘密同态的在线安全计算研究 [J]. 计算机应用研究,2007,24(12): 189-190,222.

[9] 尹春勇,李秦伟,孙汝霞. 安全科学计算系统 SSCS 实现 [J]. 计算机工程与设计,2007,28(13): 3070-3081.

[10] XIANG Guang-li, CHEN Xin-meng, ZHU Ping. A method of homomorphic encryption [J]. Wuhan University Journal of Natural Sciences, 2006, 11(1): 181-184.

[11] ZHU Ping. Homomorphic encryption based on fraction [C]//Proc of International Symposium on Distributed Computing and Applications to Business, Engineering and Science. 2007: 487-490.

[12] QUINLAN J R. Induction of decision trees [J]. Machine Learning, 1986, 1(1): 81-106.

[13] ZHAN J. Privacy-preserving collaborative data mining [D]. Ottawa: University of Ottawa, 2006.

[14] ZHAN J, CHANG Zi-wu, MATWIN S. Using homomorphic encryption for privacy-preserving collaborative decision tree classification [C]//Proc of Computational Intelligence and Data Mining. Honolulu, HI: IEEE Press, 2007: 637-645.

(下转第 1622 页)

yphonic MIDI files based on melody similarity [C] //Proc of International Symposium on Information Science and Engineering. Washington DC: IEEE Computer Society, 2008: 232-235.

[8] ISIKHAN C, OZCAN G. A survey of melody extraction techniques for music information retrieval [C] //Proc of the 4th Conference on Interdisciplinary Musicology. Thessaloniki: [s. n.], 2008: 82-83.

[9] 张俊杰. 基于和谐泛音检测的主旋律提取技术 [D]. 上海: 上海交通大学, 2007.

[10] 韩圣龙. 一种新的音乐旋律提取方法及其应用 [J]. 现代图书情报技术, 2007(3): 77-79.

[11] RAO V, RAO P. Improving vocal melody extraction in the presence of pitched accompaniment in polyphonic music [J]. IEEE Trans on Audio, Speech, and Language Processing, 2010, 18(8): 2145-2154.

[12] 杨军. MIDI 消息和标准 MIDI 文件格式剖析及应用 [J]. 中南民族大学学报, 2003(S1): 6-8.

[13] ZIV J, LEMPEL A. A universal algorithm for sequential data compression [J]. IEEE Trans on Information Theory, 1977, 23(3): 337-343.

[14] ZIV J, LEMPEL A. Compression of individual sequences via variable-rate coding [J]. IEEE Trans on Information Theory, 1978, 24(5): 530-536.

[15] SHIH H H, NARAYANAN S S, KUO C C. Automatic main melody extraction from MIDI files with a modified Lempel-Ziv algorithm [C] //Proc of International Symposium on Intelligent Multimedia, Video and Speech Processing. Hong Kong: IEEE Press, 2001: 9-12.

[16] LI Jiang-tao, YANG Xiao-hong, CHEN Qing-cai. MIDI melody extraction based on improved neural network [C] //Proc of the 8th International Conference on Machine Learning and Cybernetics. Baoding: IEEE Press, 2009: 1133-1138.

[17] LERDAHL F, JACKENDOFF R. A generative theory of tonal music [M]. Cambridge: MIT Press, 1983: 124.

[18] PFEIFFER S, FISCHER S, EFFELSBERG W. Automatic audio content analysis [C] //Proc of the 4th ACM International Conference on Multimedia. New York: ACM Press, 1996.

[19] SOFIANOS S, ARIYAEEINIA A, POLFREMAN P. Towards effective singing voice extraction from stereophonic recordings [C] //Proc of IEEE International Conference on Acoustics Speech and Signal Processing. 2010: 233-236.

[20] DURRIEU L J, RICHARO G, DAVID B. Singer melody extraction in polyphonic signals using source separation methods [C] //Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2008: 169-172.

[21] GAO Ping, FTENG G, YANG Duan-li, *et al*. Melody information extraction of humming based on wavelet analysis [C] //Proc of the 2nd International Conference on Innovative Computing, Information and Control. 2007.

[22] THORNBURG H, LEISTIKOW R J, BERGER J. Melody extraction and musical onset detection via probabilistic models of framewise STFT peak data [J]. IEEE Trans on Audio, Speech, and Language Processing, 2007, 15(4): 1257-1272.

[23] GAO Liang, WU Ya-dong. A system for melody extraction from various humming inputs [C] //Proc of IEEE International Symposium on Signal Processing and Information Technology. 2006: 680-684.

[24] CAO Chuan, LI Ming, LIU Jian, *et al*. Singing melody extraction in polyphonic music by harmonic tracking [C] //Proc of International Conference on Music Information Retrieval. [S. l.]: Austrian Computer Society, 2007: 3-4.

[25] ZHU Yong-wei, GAO Sheng. Extracting vocal melody from karaoke music audio [C] //Proc of IEEE International Conference on Multimedia and Expo. [S. l.]. IEEE Press, 2005.

(上接第1617页)

[15] FANG Wei-wei, YANG Bing-ru, SONG Ding-li, *et al*. A new scheme on privacy-preserving distributed decision-tree mining [C] //Proc of the 1st International Workshop on Education Technology and Computer Science. 2009: 517-520.

[16] 方炜炜, 胡健, 杨炳儒, 等. 分布式决策树挖掘的隐私保护研究 [J]. 计算机科学, 2009, 36(4): 239-242.

[17] RAJU R, KOMALAVALLI R, KESAVKUMAR V. Privacy maintainance collaborative data mining: a practical approach [C] //Proc of the 2nd International Conference on Emerging Trends in Engineering and Technology. 2009: 307-311.

[18] KUMAR P, SINGH M, SAXENA A. HEMIN: a cryptographic approach for private KNN classification [C] //Proc of International Conference on Data Mining. 2008: 500-505.

[19] ZHU Jian-ming. A new scheme to privacy-preserving collaborative data mining [C] //Proc of the 5th International Conference on Information Assurance and Security. 2009: 468-471.

[20] IMAN S, ALAA M, AMIN S, *et al*. P3ARM: privacy-preserving protocol for association rule mining [C] //Proc of IEEE Workshop on Information Assurance. 2006: 76-83.

[21] GUI Qiong, CHENG Xiao-hui. A privacy-preserving distributed method for mining association rules [C] //Proc of International Conference on Artificial Intelligence and Computational Intelligence. 2009: 294-297.

[22] ZHAN J, MATWIN S, CHANG Li-wu. Privacy-preserving collaborative association rule mining [J]. Journal of Network and Computer Applications, 2007, 30(3): 1216-1227.

[23] SOMESH J, LUIS K, PATRICK M. Privacy preserving clustering [C] //Proc of the 10th European Symposium on Research in Computer Security. 2005: 397-417.

[24] 周水庚, 李丰, 陶宇飞, 等. 面向数据库应用的隐私保护研究综述 [J]. 计算机学报, 2009, 32(5): 847-861.

[25] OUYANG Wei-min, HUANG Qin-hua. Privacy preserving sequential pattern mining based on secure two-party computation [C] //Proc of International Conference on Machine Learning and Cybernetics. 2006: 1227-1232.

[26] OUYANG Wei-min, HUANG Qin-hua. Privacy preserving sequential pattern mining based on secure multi-party computation [C] //Proc of IEEE Internatio-nal Conference on Information Acquisition. 2006: 149-154.

[27] LI Feng, MA Jin, LI Jian-Hua. An adaptive privacy preserving data mining model under distributed environment [C] //Proc of International Conference on Signal Image Technologies and Internet Based Systems. 2007: 60-68.

[28] 李锋. 面向数据挖掘的隐私保护方法研究 [D]. 上海: 上海交通大学, 2008.

[29] 俞迪. 支持隐私保护的数据挖掘研究 [D]. 长沙: 湖南大学, 2009.

[30] GE Xin-jing, ZHU Jian-ming. Collusion-resistant protocol for privacy-preserving distributed association rules mining [C] //Proc of the 11th International Conference on Information and Communications Security. 2009: 359-369.