一种基于时态约束的关联规则隐私保护方法

李军怀[1] 刘海玲[2] 彭 军[2] 张 璟[1] 陈晓明[1]

(西安理工大学计算机科学与工程学院 西安 710048)[1] (重庆科技学院电子信息工程学院 重庆 400050)[2]

摘 要 时间性是现实世界数据库本身固有的因素,更是构成隐私数据的基本属性,把它作为约束条件,就可以研究更为真实的现实情况。基于隐私保护和数据安全的考虑,现将数据的时态特性与不同粒度数据的安全级相结合,提出了数据安全级的时效性概念,然后对不同的项目或事务对应的安全级采用不同等级进行属性分层及数据的隐私保护处理,并提出了时态约束下的关联规则隐私保护算法。最后通过实验对算法的信息损失度和效能进行了分析和验证。

关键词 隐私保护,时态约束,数据挖掘,概化,数据安全级

中图法分类号 TP311 **文献标识码** A

Method of Association Rule Privacy Protection Based on Temporal Constraint

LI Jun-huai[1] LIU Hai-ling[2] PENG Jun[2] ZHANG Jing[1] CHEN Xiao-ming[1]

(School of Computer Science & Engineering, Xi'an University of Technology, Xi'an 710048, China)[1]

(College of Electronic Information Engineering, Chongqing University of Science and Technology, Chongqing 400050, China)[2]

Abstract Time is an inherent property of data in the real world and also is an important basic characteristic of privacy data. We can delve into the problem about data privacy protection if we take time property as a constraint condition. On the consideration about privacy protection and data security, we combined the data temporal property with security level of the different data granularity and then proposed the conception of time effectiveness about data security. Furthermore, we applied the different level to generalize the different items or transactions in order to protect data privacy, and presented the algorithm of association rule privacy protection based on temporal constraint. Finally, we analyzed and evaluated the method performance, which include information loss and algorithm effectiveness.

Keywords Privacy preserving, Temporal constraints, Generalization, Data confidential level

数据挖掘研究的是如何从大量数据中发现有用的知识。但是随着挖掘工具的普及,它们可能对隐私和信息安全构成威胁,所以需要研究新的方法来保证数据挖掘中的数据隐私不被泄露。目前,在数据隐私保护方面,许多的文献对此提出了各种不同的保护方法,主要包括对挖掘数据集应用随机化方法、对隐私建立度量度评估、取代样本真实数据、对记录进行交换等方法,同时有在分布式环境下的数据挖掘隐私保护以及通过对原始数据的混乱或扭曲进行隐私保护、敏感数据隐藏算法、规则混乱、取样法等[1-3]。但是,这些方法大都基于如下两个前提:①所有人的保护需求是一样的,或者从数据的角度说,所有数据的安全等级是一样的;②所保护的对象数据库是永远有效的。在第一个前提下,容易造成对部分隐私的保护不充分,而对部分隐私数据做出了过量的保护,浪费了资源。在第二个前提下,没有任何约束表明数据何时变得有效,何时又被认为无效,不考虑这些与时间相关的属性显然是不合理的。现实世界中存在着大量与时间密切相关的数据,例如股票市场的行情变化、超市的交易记录、病员的病历数据记录、天气数据日志文件等等[4]。时间性是现实世界数据库本身固有的因素,更是构成隐私数据的基本属性。同样,隐藏在这些数据库中的知识也必然有相应的时态约束,以表明所发现的知识何时是有效的。目前,规则事实上都是假定永远有效的。在这种情况下,没有任何东西表明规则何时变得有效,何时又被认为无效。在现实中,附加上时间特性的这种时态规则将可以更好地描述客观现实情况,因而也会更有价值。因此,在隐私保护数据挖掘研究中,考虑知识的时态特性具有重要的意义。

不同于以往的隐私保护方法,本文所提出的基于时间约束的隐私保护中,所研究的方法主要有以下两个特点:

第一,隐私数据保护方法会针对用户的个性化需求作出相应的处理;

第二,隐私数据保护方法会针对不同的时间层次、不同的时间区间作出相应的处理。

对于隐私保护,所采取措施的根据是用户隐私数据的数据安全级。正如前面说到的,由于时效性是数据的基本特性,那么数据安全级同样是与时间紧密相关的,具有时效性。通过时间约束,对不同的数据赋以不同的数据安全级;通过设置

到稿日期:2008-11-03 返修日期:2009-05-15 本文受国家863重点项目(编号:2007AA010305和2007AA010402),陕西省教育厅科技计划项目(编号:07JK333),西安市科技计划项目(编号:GX07026)资助。

李军怀(1969-),男,博士,副教授,主要研究方向为分布式计算、Web数据挖掘等,E-mail:lijunhuai@xaut.edu.cn;刘海玲 女,讲师;彭 军 教授,主要研究方向为密码学、数据安全;张 璟 教授,博士生导师,主要研究方向为服务计算、网络化制造;陈晓明 男,硕士研究生。

数据安全级，对隐私进行不同级别的保护。

1 时态数据挖掘及隐私保护相关技术

1.1 时态数据挖掘技术

时态数据是在传统的数据库基础上加上时间维，时态数据的特点决定了时态数据库中的挖掘技术及所发现的时态知识都具有其自身的特点，目前的研究主要集中在时态关联挖掘、周期性挖掘、趋势性挖掘和序列模式挖掘等方面。在时态关联挖掘中，主要包括单维、多维、单层次、多层次、量化、基于距离的关联等等。当前的时态关联研究大多将已有的关联分析运用到时态数据中，主要考虑该关联成立的时间范围，提出了一些时间区间合并、延展技术[5]。此外，一些时态关联挖掘的算法大都是基于 Apriori 算法的变形[6]。

周期模式挖掘可视为一组分片序列为持续时间的序列模式挖掘，分为全周期模式的挖掘、部分周期模式的挖掘及循环或周期关联规则的挖掘。有关周期性的分析大都应用了 Apriori 启发式特性和变通的 Apriori 挖掘方法。趋势性挖掘主要针对连续型数值，通过对数字曲线模式利用统计时序中的方法进行分析，以获得属性随时间变化的趋势，从而制定出长期或短期的预测[7]。序列模式挖掘的目的是为了寻找一段特定时间以外的可预测行为模式[3]。

1.2 隐私保护数据挖掘技术

目前在隐私保护数据挖掘领域已经采用许多技术方法，在实际应用中，国内外的学者已经提出了诸多算法，它们主要集中在每一种特定情形下的算法讨论上。现有的一些隐私保护技术大体上可基于下面的因素对它们分类：数据分布、隐私保护技术、数据或规则更改方法、数据挖掘算法。

Stanley R. M. Oleveira 等人在文献[3]中提出了一种基于启发式的隐私保护方法，该算法通过一种单次扫描算法来实现对敏感规则的保护。

Elena Dasseni 等在文献[8]中提出了一种基于混乱的方法，通过隐藏与敏感规则相关的频繁项集，以及通过设定阈值来减少置信度，防止敏感规则的产生。

在基于隐私保护的分类技术方面则是在数据挖掘过程中，建立一个没有隐私泄露的、准确的分类模型[9]。

文献[10]中提出的一种隐私保护方法结合了分类规则和吝啬降级法(parsimonious downgrading)。

Evfimievski 等在文献[11]中提出了一种基于重建式的隐私保护算法，使用了一种称之为统一随机化的方法(Uniform Randomization)。R. Agrawal 在文献[1]中提出了一种基于重建式的技术，算法针对数值型数据概率分布的重建，通过添加随机偏移量对原始数据进行随机化混乱，然后使用贝叶斯公式，根据原始数据的分布来重建决策树。通过重建数据分布可以建立一种准确程度接近真实数据分布的分类标记。

对于分布式环境下的隐私保护问题，安全多方计算(SMC)是最为常用的一个协议。安全多方计算是在一个互不信任的多用户网络中，各用户能够通过网络来协同完成可靠的计算任务，同时又能保持各自数据的安全性[11]。

Murat Kantarcioglu 在文献[12]中提出了一种数据水平分布下针对关联规则的隐私保护数据挖掘算法。各个站点不必知道其他站点的具体记录信息，就可以计算出全局的关联规则。算法提出的目标就是各个站点除了知道全局的结果之外，对其他各站点的信息一无所知。Jaideep Vaidya 在文献[13]中提出了一种数据垂直分布条件下的基于关联规则隐私保护算法。数据按照属性分布在各个站点，在这种条件下，可以通过发现项集的支持计数来进行数据挖掘，如果某个项集的支持计数可以被安全地计算，那么通过检查计数和预先设定的阈值比较，就可以知道该项集是否是频繁项集。

2 相关定义

2.1 数据安全级

数据安全级来源于密级的概念，密级属于安全范畴。在隐私保护领域，同样可以认为不同的信息、不同的数据具有不同的密级，本文称之为数据密级。

定义 1(数据密级) 密级实际是一个按密级高低的升序排列的线性有序的名称序列。令 q_i 表示第 i 个密级，且 $q_1 < q_2 < q_3 < \cdots < q_i$，则 $Q=\{q_1, q_2, q_3, \cdots, q_i\}$ 表示一类密级，其中 $q_1, q_2, q_3, \cdots, q_i$ 称为密级项。

定义 2(隐私项) 设集合 $P=\{p_1, p_2, p_3, \cdots, p_m\}$ 中每一元素都表示一个数据项，每一个数据项实际上都是一个需要保护的隐私信息，称之为隐私项。

定义 3(含有时间约束的数据安全级) 设安全级 $S=\{s_1, s_2, s_3, \cdots, s_n\}$ 是由 $X_i(1 \leqslant i \leqslant m)$ 组成的集合，S 是由密级项 Q 与隐私项 P 以及隐私项有效时间 T 组成的三元组 $S=(Q, P, T)$，其中 T 是一个对象在现实世界中发生并保持生存状态的那段时间，则称 S 是一个带有时间约束的安全级集合，简称为时效安全级，它表示了一类具有相同数据安全级在同一时间粒度的隐私项集合。

2.2 概化

所谓概化就是将大的任务相关的数据集从较低的概念层抽象到较高的概念层[3]。本文所关心的是在一张数据表 T 的任一属性列 $\{A_0, A_1, A_2, \cdots, A_n\}$ 上进行层次概化。

定义 4(概化) 假定一张数据表 T，针对某一个属性 A，在 A 上的概化是关于 A 的一个关系，即 $f: A \to B$。

进一步说，称以下形式为一个概化序列，即：

$$A_0 \xrightarrow{f_0} A_1 \xrightarrow{f_1} A_2 \cdots \xrightarrow{f_{n-1}} A_n$$

其中，$A_n=\{A_{n-1,0}, A_{n-1,1}, \cdots, A_{n-1,i}\}$，$A_n$ 是 T 的第 n 个属性，$A_{n-1,i}$ 是在 A_n 上的一个划分，f_i 表示属性 A 上的第 i 层概化。这个序列还称作是属性 A 上的一个域概化层次(Domain Generalization Hierarchy，DGH[4])，它是一个从 A_0 到 A_n 的线性结构。域是指事物某一属性的取值范围。在层次概化中，往往处于最下层的域包含最多的属性元素，此时对事物的描述最具体，越高一级也越抽象概括。而到达最高层的时候，域中只包含一个元素。

2.3 时态关联规则

对于基于关联规则的隐私保护数据挖掘，通常是指使得关联规则的支持度、置信度分别不大于用户指定的最小支持度阈值(MST)和最小置信度阈值(MCT)。那么，在基于时态约束的数据挖掘中，需要对关联规则加入时间约束。

定义 5(时态约束关联规则) 设 D 是原数据库，$X_{(g1)}$ 和 $Y_{(g2)}$ 分别是满足时态条件 $g_1(T_1, P_1)$，$g_2(T_2, P_2)$ 的项集，且 $X \cap Y=\varphi$。如果存在规则 $r: X_{(g1)} \Rightarrow Y_{(g2)}$，则称 r 是含有时态约束的关联规则，记作 $r(T, P)$。

如果 r 满足以下条件，那么称 r 为含有时态约束的强规则或敏感规则：

(1)supp $(X_{(g1)} \cup Y_{(g2)}) > MST$；

(2)conf$(X_{(g1)} \Rightarrow Y_{(g2)}) > MCT$。

隐私保护主要目标是隐藏一些含有高敏感知识的频繁项集-敏感规则。

3 基于时间约束的隐私保护模型和算法

3.1 隐私保护模型

如图 1 所示，隐私保护模型包含以下几个部分：原始数据库与发布数据库，分别含有待处理的隐私数据和经过算法处理的数据，预处理，生成概化层次，然后根据隐私保护算法，对敏感项进行隐藏。

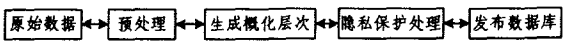

图 1 隐私保护模型

3.2 预处理与概化

主要是针对数据安全级以及概化层次的设置，实际上安全级和概化层次可以看作是用户时间轴划分的函数。也就是说，根据用户对时间区间的划分，可以将相应的区间映射到相应的数据安全级(S)，再根据安全级设置概化层次(DGH)。那么知识库定义为如下形式，即

$S = f_1(partition\ (UT))$；

$DGH = f_2(S)$；

其中，$partition(UT) = [T_i \cdots T_j)$，$T_i$，$T_j$ 是时间轴(UT)上的两个时刻，并且 $i \leqslant j$。

为了生成数据安全级到域概化层次上的映射关系，必须首先根据原始数据库生成所有属性的概念分层，各个属性的概念分层是一种树形结构，因而也常被称作是概念树。通常采用的方法是 FP 算法。FP 算法是一种频繁模式增长算法，它采用如下分治策略：将提供频繁项集的数据库压缩到一棵频繁模式树(FP-tree)，但仍保留项集关联信息，然后将这种压缩后的数据库分成一组条件数据库(一种特殊类型的投影数据库)，每个关联一个频繁项，并分别挖掘每个条件数据库。

当进行安全级到概化层次的映射时，由于不同数据间的概念分层是不同的，例如某地区区号的概化等级可以分成 7 级{710048，71004 * ，7100 * * ，710 * * * ，71 * * * * ，7 * * * * * ，* }，年龄的概化等级可以划分为 3 级{32，[30，40]，* }，这样在实际应用中不可能对所有的概念划分同样多的等级，所以就需要根据具体情况进行设置。

为了方便讨论，根据有效结束时间与当前时间的差值进行知识库设置。设当前时间表示为 Now，事务有效开始时间表示为 Begin_Time，有效结束时间表示为 End_Time，令 Δt = Now－End_Time，那么有：

$S = f_1(\Delta t)$，$DGH = f_2(S)$，并且假定数据安全级 S 为一整数集合：

将 f_1 定义为时间跨度与安全级的二元关系，即：

$f_1(\Delta t) = \{s_i \mid s_i \in S, \Delta t \in partition(UT)\}$；

将 f_2 定义为安全级到概化等级的二元关系。由于存在安全等级与概化等级不同的情况，因此需要进行分类设置。以 maxlevel(DGH)表示层次概化的最高等级数，maxlevel(S)表示数据安全级的最高等级数：

maxlevel(DGH)$\leqslant$maxlevel(S)时：

$$DGH_i = \begin{cases} s_i, & i \in (0, \text{maxlevel})(DGH) \\ *, & i \in (\text{maxlevel}(DGH), \text{maxlevel}(S)) \end{cases}$$

maxlevel ($DGHA$)$\geqslant$maxlevel(S)时：

$$DGH_i = \begin{cases} s_i, & i \in (0, \text{maxlevel})(S) \\ *, & i \in (\text{maxlevel}(S), \text{maxlevel}(DGH)) \end{cases}$$

在进行隐私保护处理的过程中，细节信息的损失是不可避免的，而主要导致信息损失的来源在于概化层次的选择。由前面的定义可以看出，概化层次是直接取决于数据安全级设置的，数据安全级划分的粒度不同，将会导致不同的处理结果，而安全级的设置与时间区间的划分有着密切的关系。但事实上，由于事先并不知道时间区间的划分为多大的跨度才合适，这样时间跨度设定得不合适也就在所难免。因此，一般采取试探性的办法。为了达到处理结果有效，必然要不断地调整。

3.3 敏感项处理

定义 6(敏感项) 设 D 是一个交易事项数据库，FR 是可以从 D 中挖掘的所有频繁项集，R_h 是根据某些安全策略必须隐藏的规则。存在一些频繁项集 I_{FR}，如 $I_{FR} \subset FR$，而且当且仅当规则 R_h 只能从 I_{FR} 中挖掘出来，那么 I_{FR} 称作敏感项。存在另一些项集 $\sim I_{FR}$，并且 $I_{FR} \cup \sim I_{FR} = FR$，则 $\sim I_{FR}$ 称作非敏感项。

定义 7(含有时态约束的敏感规则隐藏) 设 D 是原数据库，D' 是经过隐私保护算法处理后的数据库，D 中有规则 r：$X_{(g1)} \Rightarrow Y_{(g2)}$ 在 D' 中隐藏需至少满足以下条件之一：

(1)supp $(X_{(g1)} \cup Y_{(g2)}) < MST$；

(2)conf $(X_{(g1)} \Rightarrow Y_{(g2)}) < MCT$。

通常来说，在进行关联规则隐私保护时，一个简单而有效的方法是降低敏感项的支持度。其方法是可以删除或修改一些敏感交易项中的项或项集，来达到隐藏敏感项的目的，进而对关联规则进行隐私保护。因此，在进行算法处理前必须对原始数据库进行关联规则的分析，设置 MST 或者 MCT，作为隐私保护算法的根据。

目前有 IGA[3] 算法和 SWA[14] 算法，通过从一组包含敏感规则集的事务集中移出部分项集，从而使敏感规则集的支持度或置信度低于安全阈值的要求；同时算法中还有一个可由数据库用户本身控制的开放阈值参数 Ψ，用来表示用户对敏感度的具体要求。当 $\Psi = 0\%$ 时，所有敏感规则都必须隐藏；当 $\Psi = 100\%$ 时，所有的规则都可认为不是敏感的。

3.4 隐私保护算法描述

基于时间约束的隐私保护数据挖掘问题可以描述为：从原始数据 D 构建知识库，挖掘所有频繁项集 P，R_h 是根据某些安全策略必须隐藏的规则($R_h \subset P$)，通过运用保护算法将原有数据库 D 转变成对外公开的数据库 D^*，使得从 D^* 挖掘不到包含敏感信息的规则，即隐藏 R_h，从而达到隐私保护的目的。一方面要注意在实施过程中尽量减少对原有数据库的改动以及造成的影响；另一方面还要在隐私信息保护和数据挖掘之间寻找到一个平衡点。

算法 1 基于时空约束隐私保护算法

输入：发布数据表 T；知识库定义 S，DGH，需要隐藏的一组关联规则集 R_h，MST，MCT；

输出：经算法处理后的数据表 T*，表中不产生规则集 R_h；

方法：

(1)Load(T)；

(2)Create FP-tree;
(3)For all rules r in T do
(4)Δt =Now−EndTime;// 计算当前时间与其有效结束时间的距离;
(5)S=$f_1(\Delta t)=\{s_i|_{i\in}S,\Delta t\in$ partition (UT) $\}$;// 根据时效数据安全级空间定义,计算元组的安全级
(6)DGH = $f_2(S)$;//对元组进行概化处理
(7)foreach rule r{
(8) if r in R_h then{
(9) foreach item i in r{
(10) TimePartition(i);//计算 r 中每个元组的有效时间所在的时间划分;
(11) Confidential_Level(i);//由知识库定义,根据第(5)、(6)步的计算结果,计算每个元组的数据安全级 S;
(12) Compute_DGH(i);//根据所在元组的数据安全级 S,计算元组的隐私保护等级 DGH,对元组进行概化处理;}
(13) min_conf(r),min_supp(r);//计算 r 的最小置信度,最小支持度
(14) repeat until min_conf(r)<MCT or min_supp<MST do
(15) delete one row with the lowest S;
(16) delete r;//从 R_h 删除 r;}
(17) }
(18) output(T*);//保存结果,输出数据表 T*。

在进行算法之前,还可以对数据按照字典序列进行排序,以加快算法处理速度。分析以上过程,在隐藏关联规则的过程中,使用的是删除关联规则中的项集的方法。删除的规则是数据安全级低优先,因为数据安全级低的元组,通常概化等级也低,也就是说元组中包含的项比较具体真实,那么通过删除这一部分数据,很明显就可以达到隐藏规则的目的。

4 实验讨论与分析

4.1 实验

本节对上述算法进行实验测试。从以下几个方面进行分析,比较:(1)元组长度固定,进行约束的属性组变化;(2)元组长度变化,进行约束的属性组固定。实验数据由随机数据发生器产生,采用元组长度为 10 的表作为数据集。数据集中的各特征值是均匀分布的,数据集的大小由 1000 变化到 10000,记作 1k～10k,对属性集采用包含 2～8 个属性的方法进行。

4.2 性能分析

4.2.1 执行时间分析

通过图 2 可以看出数据集大小对算法执行时间的影响,就某一曲线来讲,算法的执行时间与数据集的大小并不是一个正比关系,尤其是在属性集数目较少的情况下,算法的执行时间变化不是很快。随着属性集的增加,算法在数据变化方面出现了差异,基本是按照属性集的增加而增加的。这是因为数据量相同时,算法需要更多的时间去过滤符合条件的数据元组;而在同一属性集数量情况下,因为数据量的增加,元组需要搜索更大的数据空间,从而增加了执行时间。

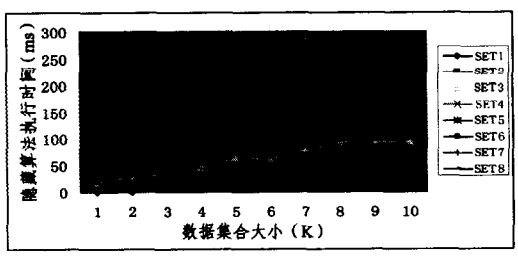

图 2 执行时间随数据量和属性集数目的变化

4.2.2 算法效能

图 3 显示了不同时间区间、不同数据安全级下规则数的变化。其中横轴为时间区间,T1:1980 年到 1982 年,T2:1982 到 1984 年,依此类推,一直到 T10:1998 年到 2000 年。图中的 3 条曲线 S1,S2,S3 分别代表了 3 种时间约束,即用安全等级阈值分别为 0,2,4 的元组进行统计。从图中可以得出不同时间区间规则数目的变化情况,同时可以得出随着数据安全级的降低可挖掘出的关联规则数目呈减少的趋势。这是因为随着安全级的降低,进行隐藏的数据也就越多,导致频繁集数目的降低,从而减少了规则数目。

图 4 所示的是随着 MCT 的变化,算法产生的规则数的变化情况。算法中所取的时间约束条件为 1990 年到 1998 年,可以得出:在不断增加 MCT 的情况下,经过算法处理后,所挖掘出的规则数逐步减少。这是因为随着置信度阈值的提高,符合约束条件的数据也就随之增加,这样算法所要处理的隐私数据增多,导致频繁集的比例降低,减少了所能挖掘出的规则的数目。

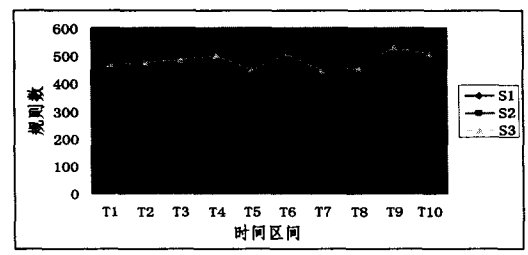

图 3 不同时间区间的关联规则数

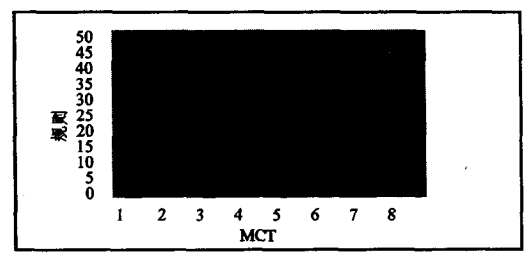

图 4 规则数随 MCT 的变化

结束语 时间是现实世界数据库本身固有的因素,所以在数据中常常会发现时态语义问题。时态数据的出现使我们有必要在知识发现过程中考虑时间因素。本文将数据的时间特性与数据安全级相结合,通过层次概化方法进行数据隐私保护处理,可以根据实际保护的最小需要,将保护对象的需求划分等级,分类进行保护,使其具有很好的灵活性。进一步的工作是将数据的时效性与空间特性相结合,研究数据隐私保护方法。

参 考 文 献

[1] Agrawal R, Srikant R. Privacy-preserving data mining[C]//Proceedings of the 2000 ACM SIGMOD Conference on Management of Data. 2000

[2] Verykios V S, Bertino E, Fovino I N. State-of-the-art in Privacy preserving Data Mining[J]. SIGMOD Record, 2004, 33(1)

[3] Oleveira S R M, Zaiane O R. Protecting Sensetive Knowledge by Data Sanitization[C]//Proceedings of the Third IEEE International Conference on Data Mining. 2003

(下转第 217 页)

件 ENVI 4.1。试验用计算机硬件配置：Intel P4 CPU 1.8GHz，512M 内存。

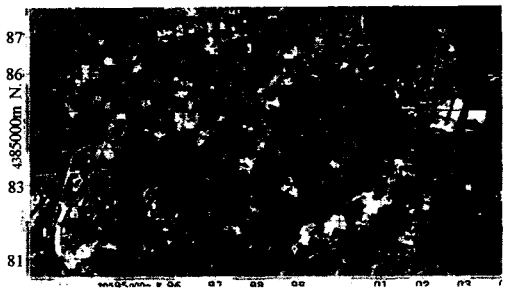

图 3　唐山地区 1∶50000TM 局部图(黑色部分为水面)

本文试验使用多类问题的分类求解。设训练样本集为 $(x_1,y_1),\cdots,(x_l,y_l),x_i\in R^n$，$x_i$ 为训练样本输入，$i=1,\cdots n$，y_i 为期望输出，表示类别标号。对于本文所讨论的问题，采用 C-SVM 分类方法，核函数为径向基核函数，$c=128$，$\gamma=0.125$。共有 4 类地物，分别表示房屋、水面、绿地和公共用地，即 $y_i\in\{4,3,2,1\}$。训练样本通过遥感软件 ENVI 的 ROIs 工具来获取。

PSO 中基本参数设置为 $c_1=2,c_2=2,\omega=1$，粒子群数量为 15，最大迭代次数为 20。PSO 算法运行的结束条件是达到最大迭代次数即刻停止。表 1 所列为训练样本经本文所述策略缩减前后分类相关参数的比较。

表 1　SVM 训练样本缩减前后分类指数比较

	训练样本总数	缩减率	最优分类精度	总耗时(秒)	支持向量(sv)个数				
缩减前	3756	0	83%	41	房屋 33	水面 608	绿地 601	公共 78	合计 1320
缩减后	1516	59%	84%	20	17	237	237	46	537

图 4 所示为样本缩减前的 SVM 分类结果-水面(蓝色部分)，图 5 所示为缩减样本后的试验结果。结合表 1 可知，基于 PSO 的缩减策略，不但丢弃了非支持向量数部分，也去除了训练样本中大量冗余的支持向量，使得训练样本得到了很大的缩减，训练和分类总消耗时间缩短了大约一半，其分类精度提高了 1%。

图 4　没缩减样本的分类结果-水面(黑色部分)

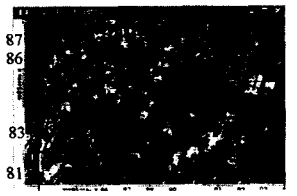

图 5　缩减样本后的分类结果-水面(黑色部分)

结束语　基于 SVM 的图像分类已经得到了广泛的应用，但也面临着挑战，特别是对大规模数据集样本的训练依然非常困难，本文采用粒子群算法对训练样本进行优化缩减，尽可能地将训练样本中非支持向量和冗余的支持向量去除，以提高训练效率。通过对多类遥感图像的分类试验证明，此方法在保证分类精度的前提下，达到了缩减的目的，这对于大规模数据集支持向量样本的训练效率的提高有非常重要的现实意义。如何进一步提高分类精度，是下一步需要研究的内容。

参 考 文 献

[1] Zhang Xue gong. Introduction to statistical learning theory and support vector machines[J]. Acta Automatica Sinica, 2000, 26(1): 32-42

[2] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines [M]. Cambridge, UK: Cambridge University Press, 2000

[3] 罗瑜，易文德，王丹琛，等. 大规模数据集下支持向量机训练样本的缩减策略[J]. 计算机科学，2007，34(10)：211-213

[4] Kennedy J, Eberhart R C, Shi Y. Swarm Intelligence[M]. San Francisco: Morgan Kaufman Publishers, 2001

[5] 王凌，刘波. 微粒群优化与调度算法[M]. 北京：清华大学出版社，2008

[6] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods[M]. House of Electronics Industry, 2005

[7] Kennedy J, Eberhart R C. A discrete binary version of the particle swarm algorithm[C]// Proceedings of the Word Multiconference on Systemics, Cybernetics and Informatics. Piscataway, NJ, 1997

(上接第 204 页)

[4] 欧阳为民，蔡庆生. 在数据库中发现具有时态约束的关联规则[J]，软件学报，1999，10(5)：527-532

[5] 欧阳为民，蔡庆生. 数据库中的时态数据发掘研究[J]. 计算机科学，1998，25(4)：60-63

[6] 徐敏，金远平. 一种新的周期性关联规则模型[J]. 计算机工程与科学，2000，22(4)：78-81

[7] Garofalakis M. Mining sequential patterns with regular expression constraints[J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(3): 120-136

[8] Dasseni E, Verykios V S, Elmagarmid A K, et al. Hiding Association Rules by using Confidence and Support[C]// Proceedings of the 4th Information Hiding Workshop. 2001: 369-383

[9] Pinkas B. Cryptographic techniques for privacy - preserving data mining[J]. SIGKDD Explorations, 2002, 4(2)

[10] Chang Liwu, Moskowitz I S. Parsimonious downgrading and decisions trees applied to the inference problem[C]// Proceedings of the 1998 New Security Paradigms Workshop. 1998: 82-89

[11] Evfimievski A, Srikant R, Agrawal R, et al. Privacy preserving mining of association rules[C]// Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, 2002: 217-228

[12] Kantarcioglu M, Clifton C. Privacy- preserving Distributed Mining of Association Rules on Horizontally Partitioned Data[C]// ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. 2002

[13] Vaidya J, Clifton C. Privacy preserving association rule mining in vertically partitioned data[C]// the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002: 639-644

[14] Oliveria S R M, Zaiane O R. Privacy Preserving Frequent Itemset Mining[C]// Workshop on Privacy, Security and Data Mining at The 2002 IEEE International Conference on Data Mining (ICDM'02). Maebashi City, Japan, December 2002