一种面向高维数据挖掘的隐私保护方法

杨 静,赵家石,张健沛
(哈尔滨工程大学计算机科学与技术学院,黑龙江哈尔滨 150001)

摘 要: 为解决隐私保护数据挖掘中的维数灾难问题,提出了一种基于随机投影技术的隐私保护方法.该方法考虑了攻击者可以通过推测随机投影矩阵重建原始数据的情况,首先提出了安全子空间和安全子空间映射的概念,然后利用通用哈希函数生成的随机投影矩阵构造了一个安全子空间映射,实现低失真嵌入的同时保证了数据的安全,最后证明了安全子空间能够保护原始数据间的欧式距离和内积.实验结果表明,在保护数据隐私的前提下,该方法能够有效的保证数据挖掘应用中的数据质量.

关键词: 隐私保护; 高维数据挖掘; 哈希技术; 随机投影; 安全子空间
中图分类号: TP309.2 **文献标识码:** A **文章编号:** 0372-2112 (2013) 11-2187-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.11.012

A Privacy Preservation Method for High Dimensional Data Mining

YANG Jing, ZHAO Jia-shi, ZHANG Jian-pei
(*College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang* 150001, *China*)

Abstract: This paper proposes a privacy preservation method based on random projection to overcome the curse of dimensionality in privacy preserving data mining. To prevent leaks of random matrix which can lead to the reconstruction attack, it first proposes the concepts of secure subspace and secure subspace mapping. Then, it constructs a secure subspace mapping using hash technique, which is implemented by a random projection matrix, and it achieves a low distortion embedding while preserving the data privacy. Finally, it proves that the secure subspace can preserve the Euclidean distance and inner product between any two original points. The experimental results show that the proposed technique can ensure the data quality in different data mining applications effectively under the precondition of preserving data privacy.

Key words: privacy preservation; high dimensional data mining; hash technique; random projection; secure subspace

1 引言

信息技术的发展使得相关机构可以收集大量组织和个人的信息进行数据挖掘与分析,从而带来商业价值和科研价值.但是,交易数据、医疗数据以及人工普查数据等涉及大量个人隐私信息的数据发布和分析都面临着隐私泄露问题.因此,隐私保护问题已成为数据挖掘领域的重要研究课题[1].目前隐私保护研究主要包括:k-匿名技术及其衍生技术[2~4]、扰动技术[5~7]以及源于密码学领域的多方安全计算[8].然而,这些方法都面临一个重要威胁:维数灾难问题[9].在高维情况下,隐私保护面临两个困难,计算复杂度和算法的有效性.

处理高维数据的一个有效方法是随机投影技术.一方面该技术能够降低数据维数,计算复杂度较低;另一方面随机投影保护原始数据间的欧式距离.文献[10]首先提出了基于随机投影的乘法扰动方法,提升隐私保护水平的同时仍然保护数据的某些统计特征.文献[11]提出了一种基于梗概技术的稀疏文本数据的隐私保护方法,采用 AMS 梗概技术来实现数据转换.AMS 梗概技术本质上也是一种随机投影方法.以上两种方法的数据转换过程简单,执行效率高,适合高维数据的隐私保护.但是文献[12]指出,在获得随机矩阵的情况下,可以重建原始数据.本文针对已有隐私保护方法在高维数据中的不足,在随机投影技术的基础上,提出了安全子空间方法,既继承了随机投影方法的计算复杂度低的特点,又解决了现有随机投影方法中由于的投影矩阵泄露而导

收稿日期:2012-08-16;修回日期:2013-07-09;责任编辑:蓝红杰
基金项目:国家自然科学基金(No. 61370083, No. 61073043, No. 61073041);高等学校博士学科点专项科研基金(No. 201123041100011, No.20122304110012);哈尔滨市科技创新人才研究专项资金(优秀学科带头人)(No.2011RFXXG015)

致的原始数据泄漏问题.本文的主要贡献为:①提出了安全子空间和安全子空间投影的概念;②构造了一个安全子空间映射;③证明了安全子空间方法的有效性.

2 相关概念及定义

2.1 随机投影

随机投影是一种低失真嵌入,理论依据是 Johnson-Lindenstrauss 引理[13].

引理 1(Johnson-Lindenstrauss 引理) 对于任意整数 $d>0$,以及任意的 $0<\varepsilon<1,\delta<1/2$,对于 $k=O(\varepsilon^{-2}\log(1/\delta))$,存在一个映射 $f:R^d\to R^k$,使得对于任意的 $x\in R^d$,

$$\Pr[\,|\,\|f(x)\|_2^2-\|x\|_2^2\,|>\varepsilon\|x\|_2^2]<\delta \tag{1}$$

引理 1 指出在 d 维空间中的 n 个点可以嵌入到 k 维空间中,同时保证原始空间中任意两点之间的距离近似不变,其中 $k<d$.引理 1 的主要证明思路是:对于一个给定的 d 和适当的 k,定义一个映射 $f:R^d\to R^k$,证明其满足式(1).

2.2 通用哈希函数

通用哈希函数的主要思想是从一个均匀分布的 k-universal 哈希函数族中随机选取一个哈希函数[14],对于一个给定的输入,随机选择的哈希函数将会在已知的概率范围内生成同样的哈希值,也就是说,给定一个从哈希函数族 H 中随机选择的哈希函数 h 和一个哈希值 y,满足 $h(x)=y$ 的 x 的可能值是均匀分布的.本文工作主要使用 2-universal 哈希函数族.

3 安全子空间方法

安子空间方法是在随机投影的基础上提出安全子空间的概念,针对随机投影隐私保护方法中泄露随机投影矩阵会导致攻击者推测出原始数据的问题,构造难以被推导的随机矩阵,建立安全的投影变换,利用随机投影的低失真嵌入实现近似保护原始数据间距离和内积的数据转换.

3.1 安全子空间

定义 1(安全子空间) 安全子空间是一个度量空间(X',d'),经过原始度量空间(X,d)到该度量空间的一个低失真嵌入后,未授权的用户无法重建投影到该度量空间内的数据.

由定义可知一个目标空间为安全子空间需满足两个条件:一是数据从原始空间到目标空间是一个低失真嵌入;二是嵌入到目标空间的数据无法被未授权的用户恢复到原始空间.

原始数据投影到安全子空间后转变为匿名数据,这一变换通过安全子空间映射实现,安全子空间映射的定义如下:

定义 2(安全子空间映射) 映射 $f:X\to X'$ 是从原始度量空间到安全子空间的一个低失真嵌入,并且未授权的用户无法推测该映射过程.

3.2 安全子空间映射

本节详细描述安全子空间映射构造方法并证明其符合安全子空间映射定义条件.

假设原始度量空间的维数为 d,安全子空间的维数为 k,h 是从通用哈希函数族中随机选取的哈希函数 $h:[d]\to[k]$,其中$[d]$表示$\{1,2,\cdots,d\}$,$[k]$表示$\{1,2,\cdots,k\}$.定义一个随机矩阵 $\boldsymbol{P}=(P_{ij})$,$P_{ij}=g_{ih(j)}\,r_j$,当 $i=h(j)$时,$g_{ih(j)}=1$;当 $i=h(j)$时,$g_{ih(j)}=0$.r 为独立同分布的随机变量:$r_j=\{+1,-1\}$,$\Pr[r_j=1]=\Pr[r_j=-1]=1/2$,$i\in[k]$,$j\in[d]$.为证明使用随机矩阵 $\boldsymbol{P}$ 实现的映射 $f:R^d\to R^k$ 为安全子空间映射,首先证明其满足引理 1.

假定原始数据已经规范化.令 $S_i=\sum_j P_{ij}x_j$,$\sigma_i^2=E_r[S_i^2]$,则

$$\sigma_i^2=E[S_i^2]=E_r[(\sum_{j\in[d]}P_{ij}x_j)^2]=E_r[(\sum_{j\in[d]}g_{ih(j)}r_jx_j)^2] \tag{2}$$

E_r 代表关于随机变量 $r=\{r_j\}$ 的期望.由 $\Pr[r_j=1]=\Pr[r_j=-1]=1/2$,可得 $E_r[S_i^2]=\sum_{j\in[d]}g_{ih(j)}^2x_j^2$,由于 $g_{ih(j)}=\{0,1\}$,所以 $g_{ih(j)}^2=g_{ih(j)}$,带入式(2)可得 $\sigma_i^2=\sum_{j\in[d]}g_{ih(j)}x_j^2$.

由于哈希函数 h 中对于每一个 $i\in[k]$,都存在一个 $h(j)=i$,所以 $\sum_{i=1}^{k}\sum_{j\in[d]}g_{ih(j)}x_j^2=\sum_{j\in[d]}x_j^2=\|x\|_2^2$,即 $\sum_{i=1}^{k}\sigma_i^2=x_2^2$,令 $T_i=S_i^2-E_r[S_i^2]$,则 $\sum_{i=1}^{k}T_i=\sum_{i=1}^{k}S_i^2-\sum_{i=1}^{k}\sigma_i^2=\sum_{i=1}^{k}S_i^2-1$.若映射 f 满足引理 1,需要证明 $\Pr[\,|\,\|\boldsymbol{P}x\|_2^2-\|x\|_2^2\,|>\varepsilon\|x\|_2^2]<\delta$,

即 $\Pr[\,|\sum_{i=1}^{k}S_i^2-1|\geqslant\varepsilon]\leqslant\delta$,即 $\Pr[\,|\sum_{i=1}^{k}T_i|\geqslant\varepsilon]\leqslant\delta$.

证明 由于 $\Pr[T_i\geqslant\varepsilon]=\Pr[\exp(uT_i)\geqslant\exp(u\varepsilon)]$,根据马尔可夫不等式可知:

$$\Pr[\exp(uT_i)\geqslant\exp(u\varepsilon)]\leqslant\frac{E[\exp(uT_i)]}{\exp(u\varepsilon)} \tag{3}$$

下面首先界定 $E[\exp(uT_i)]$,$E[\exp(uT_i)]=\sum_{t\in(-\infty,+\infty)}\exp(ut)p(t)$,其中 $p(t)$ 表示 $T_i=t$ 时的概率,令 $E[\exp(uT_i)]=\sum_{t\in(-\infty,\lambda]}\exp(ut)p(t)+\sum_{t\in(\lambda,+\infty)}\exp(ut)p(t)$,其中 $\lambda\leqslant\frac{1}{u}$.首先界定 $t\in(-\infty,\lambda]$:根据指数函数不等式可知 $\sum_{t\in(-\infty,\lambda]}\exp(ut)p(t)\leqslant\sum_{t\in(-\infty,+\infty)}(1+$

$ut+u^2t^2)p(t)=1+uE[T_i]+u^2E[T_i^2]$. 由于 $T_i=S_i^2-E[S_i^2]$, $S_i=\sum_j g_{ih(j)}r_jx_j$, 可得出 $E[T_i]=0$, 则

$$\sum_{t\in(-\infty,\lambda]}\exp(ut)p(t)\leqslant 1+u^2E[T_i^2] \tag{4}$$

下面界定 $t\in(\lambda,+\infty]$:

$$\sum_{t\in(\lambda,+\infty)}\exp(ut)p(t)\leqslant\sum_{v=u\lambda}^{\infty}\exp(v+1)Pr[T_i\geqslant\frac{v}{u}+1]\leqslant\sum_{v=u\lambda}^{\infty}\exp(v+1)Pr[S_i^2\geqslant\sigma_i^2+\frac{v}{u}+1] \tag{5}$$

由于 $S_i=\sum_{j:h(j)=i}r_jx_j$, $\Pr[r_j=1]=\Pr[r_j=-1]=\frac{1}{2}$, 所以 $E_r[\exp(uS_i)]=\prod_{j:h(j)=i}E_r[\exp(ur_jx_j)]=\prod_{j:h(j)=i}(\frac{1}{2}\exp(ux_j)+\frac{1}{2}\exp(-ux_j))$.

根据指数函数不等式可知:

$$E_r[\exp(uS_i)]\leqslant\prod_{j:h(j)=i}\exp(\frac{u^2x_j^2}{2}) \tag{6}$$

由于 $\sum_{j:h(j)=i}x_j^2=\sigma_i^2$, 可得:

$$\prod_{j:h(j)=i}\exp(\frac{u^2x_j^2}{2})=\exp(\frac{u^2\sigma_i^2}{2}) \tag{7}$$

根据马尔可夫不等式和式(6)、(7)可得:

$$\Pr[S_i\geqslant t]\leqslant\frac{E_r[\exp(uS_i)]}{\exp(ut)}\leqslant\exp(\frac{u^2\sigma_i^2}{2}-ut) \tag{8}$$

选择 $u=\frac{t}{\sigma_i^2}$ 可得 $\frac{u^2\sigma_i^2}{2}-ut=-\frac{t^2}{2\sigma_i^2}$, 代入到式(8)得:

$$\Pr[S_i\geqslant t]\leqslant\exp(-\frac{t^2}{2\sigma_i^2}) \tag{9}$$

根据式(5)、(9)可知:

$$\sum_{t\in(\lambda,+\infty)}\exp(ut)p(t)\leqslant\sum_{v=u\lambda}^{\infty}\exp(v+1)\exp(-\frac{\frac{v}{u}+\sigma_i^2}{2\sigma_i^2})=\sqrt{e}\sum_{v=u\lambda}^{\infty}\exp(v-\frac{v}{2u\sigma_i^2}) \tag{10}$$

对于 $u\leqslant\frac{1}{4\sigma_i^2}$, 可得 $v-\frac{v}{2u\sigma_i^2}\leqslant-\frac{v}{4u\sigma_i^2}$, 代入(10)得:

$$\sum_{t\in(\lambda,+\infty)}\exp(ut)p(t)\leqslant\sqrt{e}\sum_{v=u\lambda}^{\infty}\exp(-\frac{v}{4u\sigma_i^2}).$$

通过几何级数的极限界定可知 $\sqrt{e}\sum_{v=u\lambda}^{\infty}\exp(-\frac{v}{4u\sigma_i^2})\leqslant 2\sqrt{e}\exp(-\frac{u\lambda}{4u\sigma_i^2})$, 因此:

$$\sum_{t\in(\lambda,+\infty)}\exp(ut)p(t)\leqslant 2\sqrt{e}\exp(-\frac{\lambda}{4\sigma_i^2}) \tag{11}$$

所以结合两段(4)、(11)可得:

$$E_r[\exp(uT_i)]\leqslant 1+E_r[T_i^2]u^2+2\sqrt{e}\exp(-\frac{\lambda}{4\sigma_i^2}) \tag{12}$$

选择 $\lambda=4\sigma_i^2\ln(\frac{k}{\delta})$, 代入(12)得:

$$E_r[\exp(uT_i)]\leqslant 1+E_r[T_i^2]u^2+\frac{4\delta}{k} \tag{13}$$

由指数函数不等式和式(13)可得 $E_r[\exp(uT_i)]\leqslant\exp(E_r[T_i^2]u^2+\frac{4\delta}{k})$. 由于 $E_r[Z_i^2]=\sum_{j\neq j',j,j'\in[d]}g_{ih(j)}g_{ih(j')}x_j^2x_{j'}^2$, h 为 2-universal 哈希函数, $\Pr[h(j)=h(j')=i]\leqslant 1/k^2$, 可得 $E_{h,r}[Z_i^2]\leqslant\frac{1}{k^2}$, 那么 $E[\exp(uT_i)]\leqslant\exp(\frac{u^2}{k^2+\frac{4\delta}{k}})$, 则:

$$E[\exp(u\sum_i T_i)]=\prod E[exp(uT_i)]\leqslant exp(\frac{u^2}{k}+4\delta) \tag{14}$$

根据式(3)、(14)可得:

$$\Pr[\sum T_i\geqslant\varepsilon]\leqslant\frac{E[\exp(u\sum T_i)]}{\exp(u\varepsilon)}\leqslant\exp(\frac{u^2}{k}+4\delta-u\varepsilon) \tag{15}$$

$U=\frac{u^2}{k}+4\delta-u\varepsilon$, 当 $u=\frac{k\varepsilon}{2}$ 时 U 取最优值. 将 $k=\frac{C}{\varepsilon^2}\log(\frac{1}{\delta})$ 代入式(15)可得: $\Pr[\sum T_i\geqslant\varepsilon]\leqslant\exp(-\frac{C\log(\frac{1}{\delta})}{4}+4\delta)$. 当 $\delta\leqslant\frac{1}{5}$, 常数 $C\geqslant 6$, $\Pr[\sum T_i\geqslant\varepsilon]<\delta$. 与上述证明思想相同可得 $\Pr[\sum T_i\leqslant-\varepsilon]=\Pr[\exp(-u\sum T_i)\geqslant\exp(u\varepsilon)]=\frac{E[\exp(-u\sum T_i)]}{\exp(u\varepsilon)}$

由于 $E[\exp(-u\sum T_i)]=E[\prod\exp(-uT_i)]=\prod E[\exp(u(-T_i))]$, 将上述证明过程中的 T_i 替换为 $-T_i$ 即可得出 $\Pr[\sum T_i\leqslant-\varepsilon]<\delta$. 综上所述, $\Pr[|\sum T_i|\geqslant\varepsilon]<\delta$, 因此该随机矩阵满足引理 1. 证毕.

因此映射 $f:R^d\to R^k$ 是一个低失真嵌入. 若映射 f 为安全子空间映射还需满足该映射过程无法被攻击者推测, 也就是投影矩阵 $\boldsymbol{P}$ 是安全的, 这将在下一小节安全性分析中说明.

3.3　安全性分析

在获得随机投影矩阵的情况下, 攻击者能够根据样本数据, 通过最大后验概率估计和欠定盲源分离法来重建原始数据[12]. 在仅知转换后数据的情况下, 攻击者只能根据投影矩阵可能的概率分布随机猜测, 无法获得任何有关原始数据的信息[10]. 因此, 安全性取决于随机投影矩阵被攻击者获知的可能性. 直接存储投影矩阵以及使用确定的随机数生成方法生成投影矩阵存在投影矩阵泄漏的风险. 确定的随机数生成方法找到

种子就可生成同样的随机矩阵.

安全子空间法无需存储投影矩阵,而是通过随机种子生成通用哈希函数族,然后从中随机选择哈希函数来隐式的表达投影矩阵 $\boldsymbol{P}=(P_{ij})$. P_{ij} 的值实质上由两个随机函数确定:$g_{ih(j)}$ 和 r_j,其中 $g_{ih(j)}$ 确定矩阵中非零元素的位置,r_j 确定非零项的值.因此,安全子空间法的投影过程可以表示为 $y_i=\Sigma_j g_{ih(j)} r_j x_j$,其中 x_j 为原始 d 维数据记录,y_i 为投影后的 k 维数据记录.由随机种子生成两组通用哈希函数族 H 和 R,$h\in H$,$h:[d]\rightarrow[k]$,$r\in R$,$r:[d]\rightarrow[\pm 1]$.设 H 中包含 n_h 个哈希函数,R 中包含 n_r 个哈希函数,非零项的个数为 n_{nz},那么 $\boldsymbol{P}$ 的可能组合有 $n_h\times n_r\times n_{nz}$ 种.攻击者即便知道所用的哈希函数族也无法确定投影矩阵,因此由该方法构造的映射 $f:R^d\rightarrow R^k$ 是安全子空间映射,目标空间 R^k 是一个安全子空间.

综上所述,攻击者仅知转换后数据而无法重建原始数据,因此,安全子空间可以防止原始数据泄漏风险.

3.4 数据可用性分析

经过隐私保护算法转换后的数据在数据挖掘中的挖掘结果越接近原始数据的挖掘结果,数据的可用性就越高.欧式距离与内积是许多数据挖掘算法中相似度的度量标准,本节证明安全子空间近似保护原始数据间的欧式距离与内积.

推论 1(欧式距离) 对于任意一个包含 n 个点的集合 $A\subset R^d$,点 $u,v\in A$,至少在 $1-\delta$ 的概率下,存在安全子空间映射 $f:R^d\rightarrow R^k$ 满足:$(1-\varepsilon)\|u-v\|_2^2\leqslant\|f(u)-f(v)\|_2^2\leqslant(1+\varepsilon)\|u-v\|_2^2$.

证明 根据 3.2 节的内容可知安全子空间映射满足 $\Pr[\,|\,\|\boldsymbol{P}x\|_2^2-\|x\|_2^2\,|\geqslant\varepsilon\|x\|_2^2]<\delta$,即在至少 $1-\delta$ 的概率下,$|\,\|\boldsymbol{P}x\|_2^2-\|x\|_2^2\,|\geqslant\varepsilon\|x\|_2^2$,也就是 $(1-\varepsilon)\|x\|_2^2\leqslant\|\boldsymbol{P}x\|_2^2\leqslant(1+\varepsilon)\|x\|_2^2$.令 $u-v=x$,则可得出 $(1-\varepsilon)\|u-v\|_2^2\leqslant\|P(u-v)\|_2^2\leqslant(1+\varepsilon)\|u-v\|_2^2$,即 $(1-\varepsilon)\|u-v\|_2^2\leqslant\|\boldsymbol{P}u-\boldsymbol{P}v\|_2^2\leqslant(1+\varepsilon)\|u-v\|_2^2$,因此可得 $(1-\varepsilon)\|u-v\|_2^2\leqslant\|f(u)-f(v)\|_2^2\leqslant(1+\varepsilon)\|u-v\|_2^2$. 证毕.

推论 2(内积) 对于任意一个包含 n 个点的集合 $A\subset R^d$,点 $u,v\in A$,$\|u\|_2^2\leqslant 1$,$\|v\|_2^2\leqslant 1$,至少在 $1-\delta$ 的概率下,存在安全子空间映射 $f:R^d\rightarrow R^k$ 满足:$|f(u)(f(v)-u\cdot v|\leqslant\varepsilon$.

证明 根据 3.2 节内容可知 $(1-\varepsilon)\|x\|_2^2\leqslant\|f(x)\|_2^2\leqslant(1+\varepsilon)\|x\|_2^2$,将向量 $\boldsymbol{x}$ 分别替换为 $u+v$ 和 $u-v$ 可得:

$$(1-\varepsilon)\|u+v\|_2^2\leqslant\|f(u+v)\|_2^2\leqslant(1+\varepsilon)\|u+v\|_2^2 \tag{16}$$

$$(1-\varepsilon)\|u-v\|_2^2\leqslant\|f(u-v)\|_2^2\leqslant(1+\varepsilon)\|u-v\|_2^2 \tag{17}$$

由式(16)和(17)可知:

$$\|f(u+v)\|_2^2-\|f(u-v)\|_2^2\geqslant 4u(v-2\varepsilon(\|u\|_2^2+\|v\|_2^2) \tag{18}$$

$$\|f(u+v)\|_2^2-\|f(u-v)\|_2^2\leqslant 4u(v+2\varepsilon(\|u\|_2^2+\|v\|_2^2) \tag{19}$$

由于 $\|f(u+v)\|_2^2-\|f(u-v)\|_2^2=4f(u)f(v)$,代入式(18)、(19)得 $4u\cdot v-2\varepsilon(\|u\|_2^2+\|v\|_2^2)\leqslant 4f(u)\cdot f(v)\leqslant 4u(v+2\varepsilon(\|u\|_2^2+\|v\|_2^2)$.由于 $\|u\|_2^2\leqslant 1$,$\|v\|_2^2\leqslant 1$,可以得出 $4u\cdot v-4\varepsilon\leqslant 4f(u)\cdot f(v)\leqslant 4u\cdot v+4\varepsilon$,即 $|f(u)\cdot f(v)-u\cdot v|\leqslant\varepsilon$. 证毕.

4 实验分析

4.1 实验环境及数据

实验选取了三个数据集,其中两个来自 UCI 机器学习数据库,分别是 arcene 数据集和 arrhythmia 数据集.arcene 数据集包含 900 个样本和 10000 个属性,该数据集是二分类问题;arrhythmia 是对患者进行心律不齐的分类,包含 452 个样本和 279 个属性.另外一个数据集是 RCV1(Reuters Corpus Volume 1)数据集的子集 Reuters_5topic,本实验选取其中 320 个实例,通过去除非关键词汇最终整理出 4185 个属性.实验环境为:intel core i5 处理器,4GB 内存,1TB 硬盘,Microsoft Windows 7 操作系统.使用 32 位的 matlab(2010a)测试.

4.2 实验结果

实验首先比较安全子空间法与传统高斯随机投影对原始数据间的欧式距离与内积的保护程度,对其在数据可用性方面的性能进行评估;然后分别选取支持向量机分类算法和 k-均值聚类算法进行测试,评估其在数据挖掘应用中的有效性.通过隐私保护后的挖掘精度与原始数据的挖掘精度的比值来度量其有效性,假设在原始数据上的挖掘结果精度是 C_o,在隐私保护数据上的挖掘结果精度是 C_p,那么数据有效度为 $Q_c=C_p/C_o$.

实验中哈希函数采用乘法通用哈希[15],令 $A=\{a\,|\,a([2^l],a$ 为奇数$\}$,则通用哈希函数族 $H=\{h_a\,|\,a(A\}$,其中 $h_a(x)=(ax \bmod 2^l)\ \mathrm{div}\ 2^{l-m}$,当 d 为偶数时 $l=\log_2 d$,否则 $l=\log_2(d+1)$,当 k 为偶数时 $m=\log_2 k$,否则 $m=\log_2(k+1)$,d 为原始数据维数,k 为子空间维数,mod 表示模运算,div 表示取商的整数部分.

实验 1:欧式距离与内积

实验选取 arcene 数据集,每次实验运行 10 次(由于矩阵随机生成,所以进行多次运行),取相对误差绝对值的平均值.图 1 显示了不同投影维数下的欧式距离和

内积的相对误差.可以看出,安全子空间对欧式距离和内积的保护几乎与高斯投影相当,且投影维数越大,越接近于高斯投影.相对误差随着投影维数的增大而降低,当投影维数达到 3000(原始维数的 30%),相对误差开始低于 0.2%.这说明了在合理的投影维数内,安全子空间可以保证数据可用性.

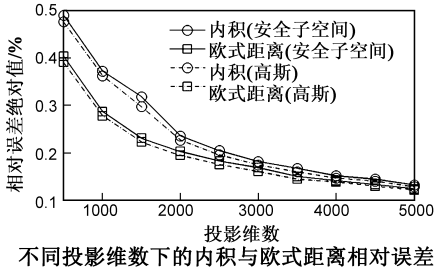


图1　不同投影维数下的内积与欧式距离相对误差比较

实验 2:分类

分类测试采用支持向量机算法实现,分两部分进行,一部分对原始数据进行分类,测试分类精度.另一部分对通过安全子空间法变换后的数据进行分类测试.实验选取 arrhythmia 数据集,采用十倍交叉验证来估计分类精度.原始数据集上的分类精度是 98.42%,不同投影维数下分类精度如表 1 所示.可见不同投影维数下的数据有效度 Q_c 分别可以达到 82.10%,87.48%,89.70%,90.50%和 91.25%.这说明了利用安全子空间转换后的数据进行分类依然能够获得较好的分类结果,转换后的数据仍然有效.

表 1　不同投影维数下的分类精度

子空间维数		10	50	100	150	200
分类精度(%)	1	73.34	76.51	77.53	76.67	78.33
	2	82.33	85.97	99.01	85.70	82.21
	3	80.67	80.20	88.40	85.70	99.30
	4	81.22	90.51	92.59	99.12	82.21
	5	78.24	92.03	92.59	92.81	95.41
	6	80.67	90.51	80.85	85.70	82.21
	7	91.08	80.95	88.40	92.81	94.02
	8	77.16	79.15	83.85	86.57	95.41
	9	82.33	85.97	87.00	92.81	94.50
	10	78.92	80.20	92.59	92.81	94.50
	平均值	80.80	86.10	88.28	89.07	89.81

实验 3:聚类

实验采用 K 均值算法及 Reuters_5topic 数据集来测试安全子空间法在聚类中的有效性,分别测试原始数据的聚类精度和转换后数据的聚类精度.K 均值算法中的 K 值设为 5,相似度度量距离选择欧式距离.不同投影维数下的聚类精度如表 2 所示.可以看出,转换到不同维数子空间内数据的有效度分别为 91.28%,94.08%,96.51%,98.25%和 99.30%,在投影维数达到原始数据维数的 50%左右时(2000、2500),投影数据的聚类结果已经接近实际聚类.并且本实验结果要好于实验 2,除了数据挖掘算法本身的作用,主要是数据集的影响.通过对比可知,本实验数据集维数更高,由此可见,原始维数越高,安全子空间法效果越好.

表 2　不同投影维数下的聚类准确率

子空间维数	聚类实例					聚类精度(%)
	1	2	3	4	5	
500	71	48	19	62	110	81.87
1000	69	45	20	64	122	84.38
1500	63	47	33	64	113	86.56
2000	64	49	21	65	120	88.12
2500	67	47	26	77	123	89.06
原始数据	61	46	28	70	113	89.69
实际聚类	66	43	24	68	119	100

5　结论

本文为解决隐私保护数据挖掘中的维数灾难问题,在随机投影技术的基础上提出了一种可以防止重建原始数据的隐私保护方法,称为安全子空间法.提出了安全子空间和安全子空间映射的概念并构造了一个安全子空间映射,利用投影转换实现原始数据的保护,使用哈希技术加密生成投影矩阵,使得恶意用户无法通过重建技术恢复扰动数据.对安全子空间法的有效性给出了数学证明,提供了理论依据.该方法能够有效的处理高维数据挖掘中的隐私保护问题.下一步将研究如何根据数据的稀疏度和分布情况来构造映射,以及根据子空间维数与数据质量、执行时间和隐私保护程度间的关系找到最优的子空间维数.

参考文献

[1] C C Aggarwal, P S Yu. A General Survey of Privacy-preserving Data Mining Models and Algorithms [M]. NewYork: Springer US, 2008. 11 – 52.

[2] L Sweeney. K-anonymity: A model for protecting privacy[J]. International Journal of Uncertainty Fuzziness and Knowledge Based Systems, 2002, 10(5): 557 – 570.

[3] 韩建民,岑婷婷,虞慧群,等.数据表 k – 匿名化的微聚集算法研究[J].电子学报,2008,36(10):2021 – 2029.
Han Jian-min, Cen Ting-ting, Yu Hui-qun, et al. Research in microaggregation algorithms for k-anonymization [J]. Acta Electronica Sinica, 2008, 36(10): 2021 – 2029. (in Chinese)

[4] 王波,杨静.一种基于逆聚类的个性化隐私匿名方法[J].

电子学报,2012,40(5):883 – 890.
Wang Bo, Yang Jing. A personalized privacy anonymous method based on inverse clustering[J]. Acta Electronica Sinica,2012,40(5):883 – 890.(in Chinese)

[5] 李光,王亚东.一种改进的基于奇异值分解的隐私保持分类挖掘方法[J].电子学报,2012,40(4):739 – 744.
Li Guang, Wang Ya-Dong. An improved privacy-preserving classification mining method based on singular value decomposition[J]. Acta Electronica Sinica,2012,40(4):739 – 744.(in Chinese)

[6] S Lee, M G Genton, R B Arellano-Valle. Perturbation of numerical confidential data via skew-t distributions[J]. Management Science,2010,56(2):318 – 333.

[7] K K Chen, L Liu. Geometric data perturbation for privacy preserving outsourced data mining[J]. Knowledge and Information Systems,2011,29(3):657 – 695.

[8] 张锋,孙雪冬,常会友等.两方参与的隐私保护协同过滤推荐研究[J].电子学报,2009,37(1):84 – 89.
Zhang Feng, Sun Xue-Dong, Chang Hui-You. Research on privacy-preserving two-party collaborative filtering recommendation[J]. Acta Electronica Sinica,2009,37(1):84 – 89.(in Chinese)

[9] C C Aggarwal. On randomization, public information and the curse of dimensionality[A]. Proceedings of 23rd International Conference on Data Engineering[C]. Istanbul: IEEE,2007,136 – 145.

[10] K Liu, H Kargupta, J Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining[J]. IEEE Transactions on Knowledge and Data Engineering,2006,18(1):92 – 106.

[11] C C Aggarwal, P S Yu. On privacy-preservation of text and sparse binary data with sketches[A]. Proceedings of the Seventh SIAM International Conference on Data Mining[C]. Minnesota: SIAM,2007.57 – 67.

[12] Y Sang, H Shen, H Tian. Reconstructing data perturbed by random projections when the mixing matrix is known[A]. Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases[C]. Berlin: Springer,2009.334 – 349.

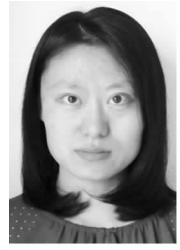
[13] W B Johnson, J Lindenstrauss. Extensions of lipschitz mappings into a hilbert space[J]. Contemporary Mathematics,1984,26:189 – 206.

[14] P Anna ,P Rasmus. Uniform hashing in constant time and optimal space[J]. SIAM Journal on Computing,2008,38(1):85 – 96.

[15] M Dietzfelbinger, T Hagerup, J Katajainen, M Penttonen. A reliable randomized algorithm for the closest-pair problem [J]. Journal of Algorithms,1997,25(1):19 – 51.

作者简介

杨 静 女,1962 年生于黑龙江哈尔滨.哈尔滨工程大学计算机科学与技术学院教授、博士生导师.主要研究方向为数据库与知识工程、数据挖掘、隐私保护、软件理论等.
E-mail: yangjing@hrbeu.edu.cn



赵家石 女,1985 年生于吉林松原.哈尔滨工程大学计算机科学与技术学院博士研究生.主要研究方向为数据挖掘、隐私保护.
E-mail: zhaojiashib09@hrbeu.edu.cn