·安全技术·　　文章编号：1000-3428(2017)03-0163-09　　文献标志码：A　　中图分类号：TP391

一种保护私有信息的空间离群点检测方法

俞庆英[a,b]，罗永龙[a,b]，陈付龙[a]，郑孝遥[a,b]

（安徽师范大学 a. 数学计算机科学学院；b. 国土资源与旅游学院，安徽 芜湖 241003）

摘　要： 针对现有空间离群点检测方法难以同时保证数据安全性和检测结果有效性的问题，提出一种隐私保护的空间离群点检测方法。该方法基于空间邻域行为属性值的统计结果及马哈拉诺比斯距离进行空间离群点的检测，通过对基于半诚实模型的安全多方距离、合并向量的中位数及标准化等计算协议的定义和应用，实现私有信息的保护。实验结果表明，该方法在保护隐私信息的同时保证了检测结果的准确性。

关键词： 空间离群点检测；空间邻域；空间属性；非空间属性；安全多方计算

中文引用格式： 俞庆英，罗永龙，陈付龙，等. 一种保护私有信息的空间离群点检测方法[J]. 计算机工程，2017，43(3)：163-171.

英文引用格式： Yu Qingying，Luo Yonglong，Chen Fulong，et al. A Privacy-preserving Spatial Outlier Detection Method[J]. Computer Engineering，2017，43(3)：163-171.

A Privacy-preserving Spatial Outlier Detection Method

YU Qingying[a,b]，LUO Yonglong[a,b]，CHEN Fulong[a]，ZHENG Xiaoyao[a,b]

（a. School of Mathematics and Computer Science；b. School of Territorial Resources and Tourism，Anhui Normal University，Wuhu，Anhui 241003，China）

【Abstract】 Foucusing on the issue that the existing spatial outlier detection methods fail to effectively solve the problem of guaranteeing both the data security and the validity of detection results at the same time，a privacy-preserving spatial outlier detection method is proposed，which uses statistical results of behavior attributes in spatial neighborhood and Mahalanobis distance to detect spatial outliers，and protects the privacy information by using the secure multi-party computation protocol based on semi-honest model，including secure distance computation，secure median computation of the combined vector and secure standardization protocols. Experimental results show that the method guarantees both the ability of privacy preserving and the effect of spatial outlier detection.

【Key words】 spatial outlier detection；spatial neighborhood；spatial attribute；non-spatial attribute；Secure Multi-party Computation(SMC)

DOI： 10.3969/j.issn.1000-3428.2017.03.028

0　概述

随着卫星、红外、CT 成像等各种传感器的广泛应用，空间数据的数量和复杂性都在飞快地增长，空间数据挖掘已成为一个新的研究领域。作为空间数据挖掘技术的一个重要分支，空间离群点检测可用于地理信息系统的多个应用领域，其目的是发现非空间属性与其空间邻居大不相同的空间参考对象[1-2]，它们在交通控制、卫星图像分析、天气预报、城镇化空间布局、医疗诊断、公共安全等很多应用中都可以揭示重要的现象，例如，帮助找到像龙卷风和飓风等极端气象事件，识别疾病的爆发和肿瘤细胞，发现反常的高速公路交通模式，决定潜在的气/油井的位置，检测水污染事件等[3]。

空间数据不仅仅包括空间属性，也包括非空间(行为)属性。其中，空间属性记录了空间信息，如位置、边界、方向，这些决定了邻居间的空间关系；非空间(行为)属性刻画了对象的本质特征，基于空间邻

基金项目： 国家自然科学基金(61370050，61572036)；安徽省自然科学基金(1508085QF134)；安徽师范大学创新基金(2016XJJ074，2015cxjj11)。

作者简介： 俞庆英(1980—)，女，讲师、博士研究生，主研方向为信息安全、空间数据处理；罗永龙，教授、博士、博士生导师；陈付龙，教授、博士；郑孝遥，副教授、博士研究生。

收稿日期： 2016-06-06　　**修回日期：** 2016-07-22　　**E-mail：** ahnuyuq@mail.ahnu.edu.cn

居关系,非空间(行为)属性用来识别反常的观察数据。对于空间离群点检测,空间维和非空间维需独立处理,空间维用来定义邻居关系,非空间维用来定义差异程度[2-3]。空间离群点检测方法既要识别单个行为属性的离群点,也要识别包含多个行为属性的离群点[1,3-6]。然而,为各种目的进行的空间离群点检测已经导致很多个人隐私泄露的问题,因此,保护隐私的空间离群点检测方法倍受关注[7]。

目前,国内外很多研究人员在隐私保护的传统离群点检测方法和不保护隐私的空间离群点检测方法方面都进行了深入研究并取得了一些研究成果。

针对隐私保护的传统离群点检测,文献[7]基于数据的同构和异构分布,提出了保证信息披露量的离群点检测技术。文献[8]基于层次聚类的数据扰动方法,提出了一种具有隐私保护功能的离群点检测技术。文献[9]立足于基于密度的离群点检测方法:局部离群因子(Local Outlier Factor,LOF),结合安全多方计算和 k-距离邻居搜索技术,提出了一种隐私保护的离群点检测方法。文献[10]针对垂直划分数据集分布在两方的情况,提出了一种隐私保护的 LOF 离群点检测协议。文献[11]以离群点检测为实例,提出了一种定制差分隐私方法。以上方法研究的都不是空间参考对象,即均未考虑数据对象的空间属性。由于空间数据的特殊性,传统的离群点检测方法虽然可以应用于空间数据,然而,方法的性能无法得到保证[3]。

另一方面,空间离群点的检测目前也已取得很多研究成果。文献[4]提出一套空间离群点检测方法,能准确有效地检测空间离群点并应用于人口普查数据的统计结果上。文献[12]提出了空间加权离群点检测方法,应用于病毒检测问题;而后,文献[13]又提出了一种基于图的空间离群点检测方法,应用于公寓出租问题;这 2 种方法均是基于单个非空间属性的离群点检测。文献[3]设计了 2 个分别用于单属性和多属性检测的空间离群点检测方法。文献[14]提出了一种改进的基于加权空间离群点(Weighted Spatial Outlier,WSO)的空间离群点检测方法。针对空间分类属性的离群点检测问题,文献[6]提出了基于成对关联函数(Pair Correlation Function,PCF)的方法,通过参考点和它的空间邻居之间的平均 PCR 计算离群点的分数等。

然而,具有隐私保护能力的空间离群点检测方法研究并不多见。为此,本文基于多个安全多方协议[15]的设计,提出一种保护私有信息的空间离群点检测方法。基于安全多方计算(Secure Multi-party Computation,SMC)理论,提出一种保护私有信息的空间离群点检测方法 PPSOD。采用基于空间邻域行为属性值的统计结果及马哈拉诺比斯距离进行空间离群点的检测,并证明 PPSOD 方法的安全性。

1 基本概念和安全协议

1.1 基本概念

在以下定义中,相关标识的含义如表 1 所示。

表 1 相关标识

标识名称	含义
DS	n 个空间数据对象集合
SAD	空间属性集合
$NSAD$	非空间属性集合
k	邻居数目
m	离群点数目
(lx,ly)	空间位置坐标

空间属性包含位置、边界和面积;非空间属性为属性约简或者数据预处理的结果。一般来说,m 不大于数据对象数 n 的 5% 。

定义 1(空间距离) 设 $o_1,o_2\in DS$,则对象 o_1 对象 o_1 和 o_2 之间的空间距离定义为其空间属性的差异程度,记为 $sdist(o_1,o_2)$,以空间位置坐标 (lx,ly) 为评价标准,$lx,ly\in SAD$,可定义如下:

$$sdist(o_1,o_2)=\sqrt{(o_1.lx-o_2.lx)^2+(o_1.ly-o_2.ly)^2} \quad (1)$$

定义 2(空间 k 最近距离) 设 $o\in DS$,则对象 o 的空间 k 最近距离定义为 o 与其第 k 个最近邻居之间的空间距离,记为 $sdist_k(o)$。

定义 3(空间 k 邻域) 设 $o\in DS$,则对象 o 的空间 k 邻域定义为基于空间属性与空间关系的与对象 o 距离最小的 k 个邻居的集合(不包括 o 本身),记为 $kNNA\_SP(o)$,即对于 $\forall p\in kNNA\_SP(o)$,$\forall q\in DS-kNNA\_SP(o)$,有 $sdist(q,o)\geqslant sdist(p,o)$,可定义如下:

$$kNNA\_SP(o)=\{p\mid sdist(p,o)\leqslant sdist_k(o),\ p\in DS\backslash\{o\}\} \quad (2)$$

定义 4(k 邻域非空间距离) 设 $\mathbf{R}$ 是一个正实数,有函数 $f:DS\rightarrow\mathbf{R}^d$,对 $\forall o\in DS$,$\boldsymbol{f}(o)$ 表示对象 o 的 d 维非空间属性向量,设计 k 邻域非空间属性统计函数 $g:DS\rightarrow\mathbf{R}^d$,则对象 o 的 k 邻域非空间距离定义为基于 $kNNA\_SP(o)$ 中所有对象非空间属性向量的统计向量,记为 $\boldsymbol{g}(o)$,本文定义 $g_q(o)$ 为第 q 维 $(q=1,2,\cdots,d)$ 非空间属性值集合 $FS_q=\{f_q(p):p\in kNNA\_SP(o)\}$ 的中位数,可计算如下:

$$g_q(o)=\begin{cases}f_q(p_{\lceil\frac{k}{2}\rceil}),k\text{ 是奇数}\\(f_q(p_{\frac{k}{2}})+f_q(p_{\frac{k}{2}+1}))/2,k\text{ 是偶数}\end{cases} \quad (3)$$

当 $d=1$ 时,对象的非空间属性是一维的,$\boldsymbol{g}(o)$ 即 $g_1(o)$,是一维数值。

定义 5(空间离群因子) 设 $\mathbf{R}$ 是一个正实数,有函数 $f:DS\rightarrow\mathbf{R}^d$,对 $\forall o\in DS$,对象 o 的空间离群因子定义为 $\boldsymbol{f}(o)$ 和 $\boldsymbol{g}(o)$ 之间的相异程度,记为 $spOF(o)$。分 2 种情况:

1)若 $d=1$,代表空间数据对象具有一维非空间属性,则 $\boldsymbol{f}(o)$ 是 o 的一维非空间属性值,$spOF(o)$ 是一维的,其值可计算如下:

$$spOF(o)=\boldsymbol{f}(o)-\boldsymbol{g}(o) \tag{4}$$

2)若 $d>1$,代表空间数据对象具有多维非空间属性,则对 $\forall o\in DS$,$\boldsymbol{f}(o)$ 是 o 的多维非空间属性向量,$\boldsymbol{f}(o)=(f_1(o),f_2(o),\cdots,f_d(o))$,对象 o 的 k 邻域非空间距离为向量 $\boldsymbol{g}(o)=(g_1(o),g_2(o),\cdots,g_d(o))$,则定义一个相异度向量 $\boldsymbol{h}(o)=(h_1(o),h_2(o),\cdots,h_d(o))$,其中,$h_q(o)(q=1,2,\cdots,d)$ 为 $f_q(o)$ 和 $g_q(o)$ 之间的相异程度,$h_q(o)=f_q(o)-g_q(o)$,用马哈拉诺比斯(Mahalanobis)距离[16]计算 $spOF(o)$ 的值,计算如下:

$$\boldsymbol{\mu}^*=\frac{1}{|DS|}\sum_{o\in DS}h(o) \tag{5}$$

$$\boldsymbol{S}=\frac{1}{|DS|}(h(o)-\boldsymbol{\mu}^*)^{\mathrm{T}}(h(o)-\boldsymbol{\mu}^*) \tag{6}$$

$$spOF(o)=(h(o)-\boldsymbol{\mu}^*)\boldsymbol{S}^{-1}(h(o)-\boldsymbol{\mu}^*)^{\mathrm{T}} \tag{7}$$

其中,$\boldsymbol{\mu}^*$ 和 $\boldsymbol{S}$ 分别是基于数据集 DS 中相异度向量的均值向量和协方差矩阵,计算公式如式(5)和式(6)所示。

定义 6(m-空间离群点) 空间离群点是一个空间参考对象,其非空间属性值显著不同于其附近的值[4],m-空间离群点可定义为 $spOFspOF$ 值最大的 m 个对象之一。

定义 7(半诚实安全计算模型) 参与安全多方计算的各方都是半诚实的,均遵从协议进行操作,不会恶意输入虚假信息或中途退出。在协议执行过程中半诚实参与方可能会保留所有它能收集到的其他参与方的信息,并在协议执行结束后对这些信息进行分析以期望得到其他参与方的输入信息。半信任行为也被称为诚实而好奇的行为,参与方也被假设为不串通的。在半诚实模型下安全的协议都可以转化为恶意模型下安全的协议,本文基于半诚实模型下的研究可以转化为恶意模型[17-18]。

1.2　安全多方计算协议

自安全多方计算[15]于 1982 年被提出以来,已获得了广泛的理论研究,并且可用于数据挖掘、统计分析等领域[17-18]。为实现隐私保护的空间离群点检测方法,本文设计了一系列安全多方计算协议,将空间数据集划分为 2 个子集,分别由 2 个数据参与方拥有,在求和、距离计算、合并向量的中位数计算和标准化过程中,可以确保所有数据方在获取所需数据的同时都不会泄露自己的私有数据。

协议 1(安全求和协议)[19]　假设有 r 个数据方 $P_1,\cdots,P_r$,P_i 方拥有数据 D_i,所有参与方不共享数据。每一方都想获取所有数据的总和 $D=\sum_{i=1}^{r}D_i$,而又不想泄露自己的数据值。由服务方 SP 生成一个随机数 R,发送给 P_1 方,P_1 方计算 D_1+R 后发送结果到 P_2 方,P_2 方计算 $\sum_{i=1}^{2}D_i+R$ 后发送结果到 P_3 方,直到 P_r 方得到 $\sum_{i=1}^{r}D_i+R$ 值并发送到 SP,SP 计算 $\sum_{i=1}^{r}D_i+R-R=\sum_{i=1}^{r}D_i$,并将求和结果发送到所有的参与方。

在求和过程中,每一方均保护了自己的数据,而 SP 方也无法获取各方数据,所以协议 1 是安全的,总共通信 $2r$ 次。该协议是其他安全协议的基础。

协议 2(安全点积协议)[20]　P_1 方拥有一个私有向量 $\boldsymbol{X}(x_1,x_2,\cdots,x_n)$,$P_2$ 方拥有一个私有向量 $\boldsymbol{Y}(y_1,y_2,\cdots,y_n)$,双方进行一个协同计算,计算过程中不会向对方泄露自己的数据,P_1 需要得到值 $u=\boldsymbol{X}\cdot\boldsymbol{Y}+v=\sum_{i=1}^{n}x_iy_i+v$,其中,$v$ 是 P_2 选取的一个随机数。满足:P_1 不能从结果中推断出向量 $\boldsymbol{Y}$ 的信息,同样 P_2 也不能推断出向量 $\boldsymbol{X}$ 的信息。点积协议是安全计算的一个基本协议,应用非常广泛,基于该协议,本文提出了协议 3。

协议 3(安全距离计算协议)　有服务方 SP、参与方 P_1 和 P_2,假设 P_1 方有一维私有向量 $\boldsymbol{X}(x_1,y_1)$,P_2 方有一维私有向量 $\boldsymbol{Y}(x_2,y_2)$,采用安全点积协议,由三方共同参与计算向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的距离。P_1 和 P_2 共享一个随机整数 R,具体处理过程如下:

1)计算出 P_1 和 P_2 共享的随机向量 $\boldsymbol{RV}=((-1)^R,(-1)^R,(-1)^R)$。

2)P_1 方计算出向量 $\boldsymbol{X}'(x_1^2,2x_1,1,y_1^2,2y_1,1)$,将向量 $4\boldsymbol{X}''=\boldsymbol{X}'\cdot RV$ 发送给 SP 方。

3)P_2 方计算出向量 $\boldsymbol{Y}'(1,-x_2,x_2^2,1,-y_2,y_2^2)$,将向量 $\boldsymbol{Y}''=\boldsymbol{Y}'\cdot RV$ 发送给 SP 方。

4)SP 方计算 $dist(\boldsymbol{X},\boldsymbol{Y})=\sqrt{\boldsymbol{X}''\cdot\boldsymbol{Y}''}$,然后将结果发送给 P_1 和 P_2 方。

协议 3 主要使用了安全点积协议计算 2 个向量的距离,从而保证了其安全性,通信为 4 次。应用该协议,可安全计算出空间对象的局部空间 k 邻域。

协议 4(合并向量的安全中位数计算协议)　有服务方 SP、参与方 P_1 和 P_2,假设 P_1 方有一维私有向量 $\boldsymbol{X}(x_1,x_2,\cdots,x_m)$,$\boldsymbol{P}_2$ 方有一维私有向量 $\boldsymbol{Y}(y_1,y_2,\cdots,y_n)$,假设 $m\leqslant n$,由三方共同参与计算 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 合并后向量的中位数。P_1 和 P_2 方共享一个随机数 R,具体处理过程如下:

1)P_1 和 P_2 方分别对各自向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的属性值进行升序排列,得出排序后的向量 $\boldsymbol{X}'$ 和 $\boldsymbol{Y}'$。

2)如果 $|\boldsymbol{X}'|=|\boldsymbol{Y}'|=1$(即 $m=n=1$),则 P_1 方

将 $x'_1 = x_1 + R$ 发送给 SP 方，P_2 方将 $y'_1 = y_1 - R$ 发送给 SP 方，由 SP 方计算 $M = (x'_1 + y'_1)/2$，将 M 分别发送给 P_1 和 P_2 方，结束协议；否则，转到步骤 3）。

3）P_1 方计算出 $\boldsymbol{X}'$的中位数 M_1，将 $M'_1 = M_1 + R$ 发送给 SP 方。

4）P_2 方计算出 $\boldsymbol{Y}'$的中位数 M_2，将 $M'_2 = M_2 + R$ 发送给 SP 方。

5）SP 方计算 $M'_1 - M'_2$，得出 M'_1 和 M'_2 的大小关系，然后将比较的结果分别发送给 P_1 和 P_2 方，如果 $M'_1 - M'_2 = 0$，P_1 和 P_2 方均收到“E”，否则，结果大的一方收到“B”，小的一方收到“S”。

6）如果两方均接收到“E”，则合并向量的中位数 $M = M_1 = M_2$，结束协议；否则，转到步骤 7）。

7）如果 P_1 方接收到“B”即 P_2 方接收到“S”，则 P_1 方删除 $\boldsymbol{X}'$中大于等于 M_1 的元素（假设有 k 个），P_2 方删除 $\boldsymbol{Y}'$中前 k 个元素，转到步骤 2）；否则，转到步骤 8）。

8）如果 P_1 方接收到“S”即 P_2 方接收到“B”，则 P_1 方删除 $\boldsymbol{X}'$中小于等于 M_1 的元素（假设有 k 个），P_2 方删除 $\boldsymbol{Y}'$中后 k 个元素，转到步骤 2）。

SP 方在步骤 2）中计算 $M = \dfrac{x'_1 + y'_1}{2} = \dfrac{x_1 + R + y_1 - R}{2} = \dfrac{x_1 + y_1}{2}$，在步骤 5）中计算 $M'_1 - M'_2 = (M_1 + R) - (M_2 + R) = M_1 - M_2$，运算结果不受影响，并且在处理过程中 P_1 和 P_2 方均未泄露自己的数据给对方，而 SP 方也无法获取两参与方的数据，所以，协议 4 是安全的。SP 方进行每一次比较通信 4 次，总的通信次数为 4lb(m)，计算复杂度为O(lb(m))。

应用该协议，可安全计算出空间对象的 k 邻域非空间距离。

协议 5（安全标准化协议） 假设有二维私有向量 $\boldsymbol{X}\begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}$，$\mu_j$ 和 σ_j 分别指的是 $\begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{mj} \end{bmatrix}$的平均值和标准差，标准化的向量为 $\boldsymbol{SX}\begin{bmatrix} sx_{11} & \cdots & sx_{1n} \\ \vdots & & \vdots \\ sx_{m1} & \cdots & sx_{mn} \end{bmatrix}$，则有：

$$sx_{ij} = \left| \frac{x_{ij} - \mu_j}{\sigma_j} \right| \tag{8}$$

具体步骤如下：

1）使用协议 1 计算第 j 维属性值的和 $sum_j = \sum_{i=1}^{m} x_{ij}$以及平均值 $\mu_j = sum_j/m(j = 1,2,\cdots,n)$。

2）使用协议 1 计算第 j 维属性值平方差之和 $s_j = \sum_{i=1}^{m} (x_{ij} - \mu_j)^2$ 以及标准差 $\sigma_j = \sqrt{\sum_{i=1}^{m} s_j/m}$ $(j = 1,2,\cdots,n)$。

3）SP 将 μ_j 和 σ_j 发送到各参与方，由各参与方使用式（5）对数据进行标准化。

协议 5 使用了安全求和协议计算平均值和标准差，从而保证了其安全性。通信次数为 $4r$ 次，计算复杂度为 $O(mn)$。

应用该协议，可安全计算出分布在 2 个参与方的空间离群因子的平均值和标准差，进而根据马哈拉诺比斯距离公式计算出对象的空间离群因子，保证了各方私有数据的安全性。

2 隐私保护的空间离群点检测方法

2.1 方法描述

在分析空间数据时，空间邻域起到了非常重要的作用。为了更好地保护非空间属性，假设空间数据对象分布在 2 个参与方，通过服务方计算空间邻域。

首先，每个参与方基于空间属性和空间关系独立构建每个对象的局部空间邻域，然后发送空间邻域集合到服务方，服务方构建出整个空间邻域关系。最后，服务方基于隐私保护协议计算每个对象的空间离群因子。

本文提出一种隐私保护空间离群点检测方法，简称 PPSOD 方法，本文设计的各个安全协议，该方法在检测空间离群点的同时可以避免对象的非空间属性对外泄露。其计算方法如下：如果对象的所有邻居都在同一方，则在他自己一方直接计算 $spOF$ 值；否则，使用安全协议计算它的 $spOF$ 值。PPSOD 方法主要步骤如下：

方法 1 PPSOD 方法

输入 空间数据集 $O = \{o_1, o_2, \cdots, o_n\}$，2 个数据参与方 P_1, P_2，第 i 方 P_i 拥有的空间数据集 $O_i = \{o_{i1}, o_{i2}, \cdots, o_{in}\}$ $(i = 1,2)$，$O = \bigcup_{i=1}^{2} O_i$，$n = \sum_{i=1}^{2} n_i$，最近邻的数目 k，空间离群点的检测个数 m，非空间属性函数 $f: O \rightarrow \boldsymbol{R}^d$，维数 d

输出 离群程度最高的 m 个（离群）点

步骤：

1）各参与方 $P_i(i = 1,2)$在其内部计算每一个空间对象 $o_{ij}(j = 1,2,\cdots,n_i)$与其他对象之间的空间距离，如果空间对象都在 P_i 方，则使用式（1）计算，否则，使用协议 3 进行计算，得到其局部空间 k 邻域 $kNNA\_SP(o_{ij})$。

2）各参与方计算每一个空间对象 o_{ij}的 k 邻域非空间距离 $g(o_{ij})$，如果 $kNNA\_SP(o_{ij})$ 中的空间对象都在 P_i 方，那么使用式（3）进行计算，否则，使用协议 4 进行计算。

3)如果 d 等于 1,则执行:

(1)各参与方分别用式(4)计算空间离群因子 $spOF_{ij}=spOF(o_{ij})=f(o_{ij})-g(o_{ij})$。

(2)使用协议 5 计算数据集 $\{spOF_{11},\{spOF_{12},\cdots,spOF_{1n_1},spOF_{21},spOF_{22},\cdots,spOF_{2n_2}\}$ 的平均值 μ 和标准差 σ,由各方计算出标准化后的值 $std\_spOF_{ij}=\left|\frac{spOF_{ij}-\mu}{\sigma}\right|$,更新各离群因子 $spOF_{ij}$ 为 std_spOF_{ij} 值。

如果 $d>1$,那么执行:

(1)在各参与方内部,分别用式(5)和式(6)计算向量 $\boldsymbol{\mu}_i^*$ 和协方差矩阵 $\boldsymbol{S}$。

(2)用式(7)计算空间离群因子 $spOF_{ij}=spOF(o_{ij})=(h(o_{ij})-\boldsymbol{\mu}_i^*)^{\mathrm{T}}S_i^{-1}(h(o_{ij})-\boldsymbol{\mu}_i^*)$。

4)各参与方将 $spOF_{ij}(i=1,2;j=1,2,\cdots,n_i)$ 的值发送到 SP 方构建全部对象的离群因子数据集。

5)设 $i_1,i_2,\cdots,i_m$ 是 $\{spOF_1,spOF_2,\cdots,spOF_n\}$ 中值最大的 m 个序号,由 SP 方计算得到 m 个空间离群点 $\{o_{i_1},o_{i_2},\cdots,o_{i_m}\}$。

2.2　安全性分析

方法 1 中的第 1)步对不同参与方的数据计算空间距离时使用了安全距离计算协议,没有泄露各自的空间属性,是安全的。

第 2)步使用了合并向量的安全中位数计算协议对来自不同参与方的非空间属性值集合进行了中位数的计算,上面已证明是安全的。

第 3)步分 2 种情况:

1)当 $d=1$ 时,首先在参与方内部计算空间离群因子 $spOF_{ij}$,没有涉及到其他方的数据,是安全的;然后使用安全标准化协议(协议 5)对数据集进行标准化,已证明其安全性。

2)当 $d>1$ 时,在参与方内部计算向量 $\boldsymbol{\mu}_i^*$、协方差矩阵 $\boldsymbol{S}$ 和空间离群因子 $spOF_{ij}$,没有涉及到其他方的数据,所以,是安全的。

第 4)步是各参与方发送各自的离群因子集到 SP 方,是标准化的统计值集合或一元 Mahalanobis 距离数据集,不会泄露各方私有对象的空间和非空间属性,是安全的。

第 5)步由 SP 方对离群因子集进行查找,没有涉及到两方私有对象的空间和非空间属性值,所以,也是安全的。

综上,PPSOD 方法是安全的。

3　实验结果与分析

本文在 Intel(R) Core(TM) 2 Duo 3.3 GHz CPU,4 GB 内存,Windows 7 操作系统上,用Matlab 8.3 实现了 PPSOD 方法,验证该方法在一维和多维非空间属性集上的有效性。采用 2014 年、2015 年安徽省统计年鉴的相关数据,具体包括各市、县、区户数、人口数(2013 年)和全省分县(市)主要经济指标及位次(2014 年)[21]。根据安徽省各县(市辖区)的中心地理坐标,基于不同的非空间属性(集),该方法检测最可能拥有异常信息的 5 个县市(市辖区)。

安徽省各县域空间位置分布数据是 ESRI 公司的 Shapefile 文件格式,整个数据集包含多个 *.shp 文件,代表各县市(市辖区)的区域和一大套包含普查结果的表("概要文件")。

在以下实验中,设置参数 $k=10$,$m=5$,将空间离群因子最高的 m 个县域作为空间离群点,实验前将所有非空间属性值进行了标准化处理。

3.1　一维非空间属性数据集

城市化是非常复杂的社会现象,对城市化水平的度量难度也很大。虽然有多个指标都可以在一定程度上反映城市化水平,但能被普遍接受的是人口统计学指标,其中,最常用的指标是非农业人口占总人口的比重。应该说,这种方法在单指标方法中是最科学的[22],它反映了人口在城乡之间的空间分布,具有很高的实用性。

本文验证并比较了 PPSOD 方法与没有隐私保护能力的 Iterative r,Iterative z 和 z 方法在一维数据集上的实验结果,采用 2014 年安徽省人口统计数据集,选择非农业人口所占比例作为非空间属性。为了更好地反映城镇化水平,进行 2 组实验,一组是采用含 76 个县市(包含市辖区)的数据集 DS_S1,另一组采用含 61 个县市(不包含市辖区)的数据集 DS_S2。

3.1.1　基于 DS_S1 数据集的实验

安徽省 $spOF$ 值最高的 5 个地区及其相关属性值(DS_S1),如表 2 所示。从表 2 可以看出,方法检测出的 5 个 $spOF$ 值最高的地区均为市辖区,非农业人口所占比例较高,表明市辖区的非农业人口数量与其空间相邻各县相比有着明显不同,这与市辖区的城镇化水平高是一致的。

表 2　安徽省 *spOF* 值最高的 5 个地区及其相关属性值(DS_S1)

Rank	ID	地名	经度/(°)	纬度/(°)	非农业人口所占比例
1	46	芜湖市辖区	118.38	31.33	1.000 0
2	57	铜陵市辖区	17.82	30.93	0.843 6
3	1	合肥市辖区	117.27	31.86	0.814 9
4	7	淮北市辖区	116.77	33.97	0.733 8
5	71	黄山市辖区	118.31	29.72	0.779 0

所得空间位置分布如图 1 所示。PPSOD 方法与其他 3 种方法可视化的运行结果如图 2 所示。其中,X 代表经度;Y 代表纬度。从图 2 可以看出,Iterative z 和 PPSOD 方法检测出了明显的离群点,优

于 z 和 Iterative r 方法。方法运行结果具体信息的比较如表 3 所示。

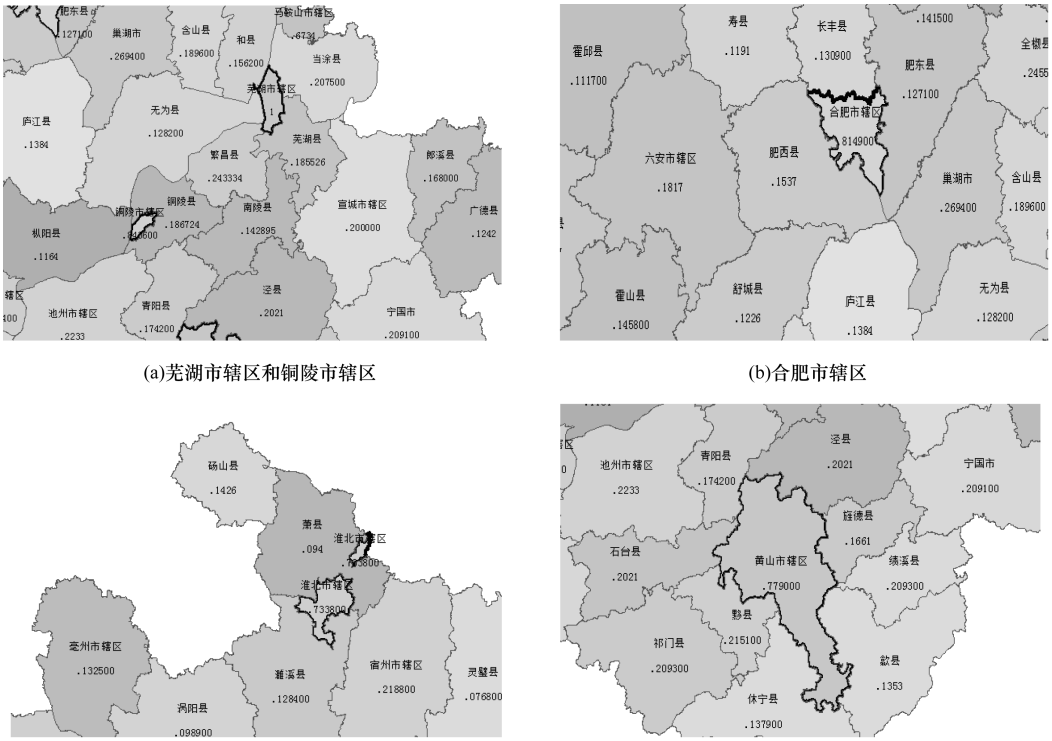


图 1 安徽省 *spOF* 值最高的 5 个县域空间位置分布(基于数据集 DS_S1)

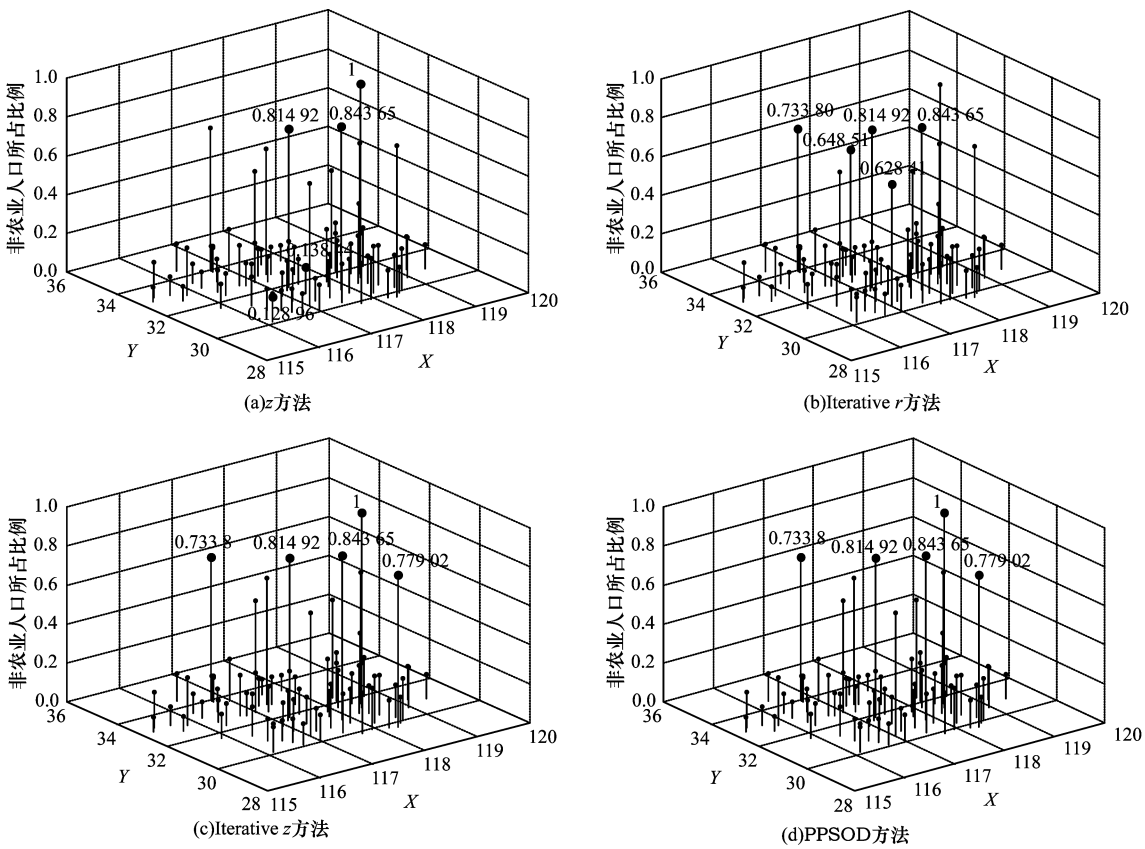


图 2 PPSOD 方法与其他 3 种方法基于数据集 DS_S1 的检测结果对比

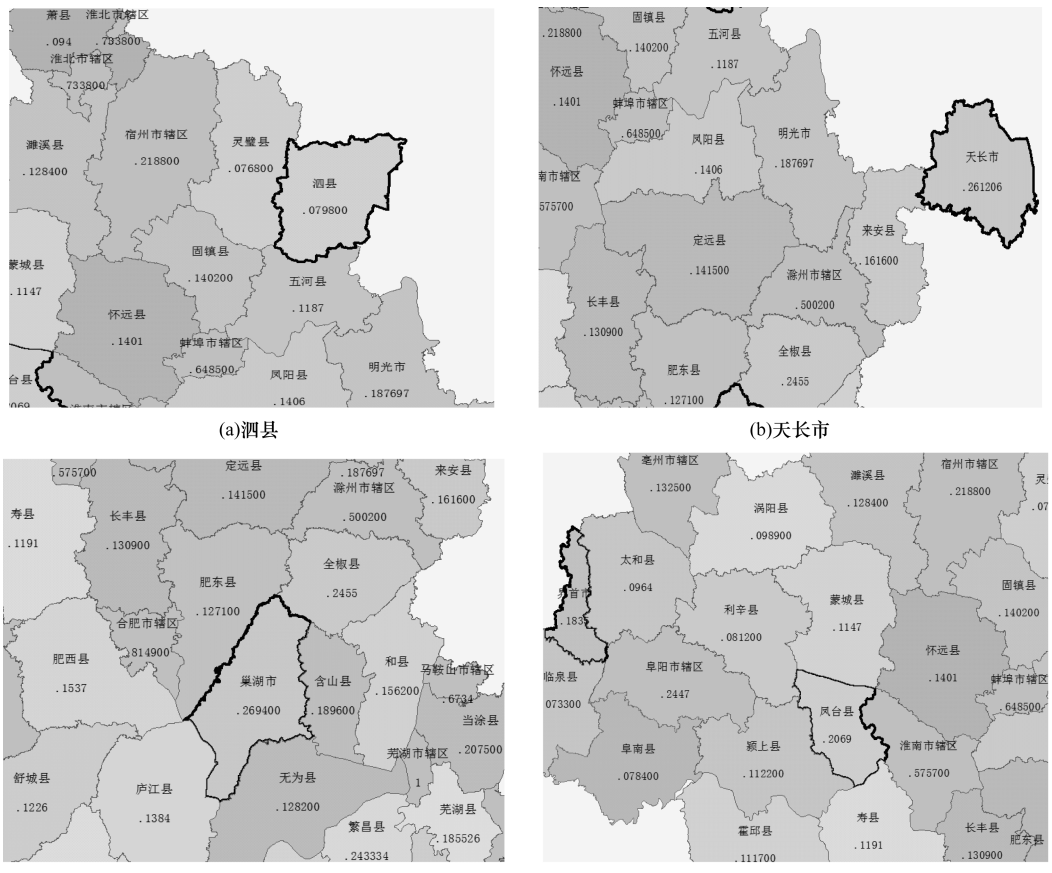
表 3　基于数据集 DS_S1 检测出离群程度最高的 5 个地区

Rank	z 方法	Iterative r	Iterative z	PPSOD
1	57,铜陵市辖区	57,铜陵市辖区	57,铜陵市辖区	46,芜湖市辖区
2	46,芜湖市辖区	7,淮北市辖区	46,芜湖市辖区	57,铜陵市辖区
3	1,合肥市辖区	1,合肥市辖区	1,合肥市辖区	1,合肥市辖区
4	6,庐江县	18,蚌埠市辖区	7,淮北市辖区	7,淮北市辖区
5	67,宿松县	62,安庆市辖区	71,黄山市屯溪区	71,黄山市屯溪区

表 3 显示了 PPSOD 方法与 Iterative r,Iterative z 和 z 方法[3-4]在数据集 DS_S1 上的运行结果,通过对前几个离群点的比较,PPSOD 也优于 Iterative z 方法,说明本文方法不仅安全而且有效,PPSOD 方法在保护隐私信息的同时保证了高准确率。

3.1.2　基于 DS_S2 数据集的实验

为了更好地反映县城的城镇化水平,基于去除所有市辖区的 61 个县市数据集,对离群信息进行检测。图 3 显示了 PPSOD 方法在数据集 DS_S2 上检测出的 5 个县域空间位置分布,详细信息如表 4 所示。从图 3 可以看出,对于方法求解出的 *spOF* 值最高的几个县市,其非农业人口比例相比于周围邻居较为异常。



(a)泗县　　(b)天长市

(c)巢湖市　　(d)界首市和凤台县

图 3　安徽省 *spOF* 值最高的 5 个县域空间位置分布(基于数据集 DS_S2)

表 4　安徽省 *spOF* 值最高的 5 个地区及其相关属性值(DS_S2)

Rank	ID	地名	经度/(°)	纬度/(°)	非农业人口所占比例
1	13	泗县	117.89	33.49	0.079 8
2	27	天长市	119.00	32.68	0.261 2
3	1	巢湖市	117.87	31.62	0.269 4
4	22	凤台县	116.71	32.68	0.206 9
5	21	界首市	115.34	33.24	0.183 5

PPSOD 方法与其他 3 种方法可视化的运行结果如图 4 所示。从图 4 可以看出,PPSOD 方法检测出了明显的离群点,优于 z,Iterative r 和 Iterative z 方法。运行结果具体信息的比较如表 5 所示。表 5 显示了 PPSOD 方法与 Iterative r,Iterative z 和 z 方法在在数据集 DS_S2 上的运行结果,通过对前几个离群点的比较,PPSOD 优于其他几种方法,几个县级市的非农业人口比例在周围邻居中相对是异常的,同样说明本文方法安全且有效。

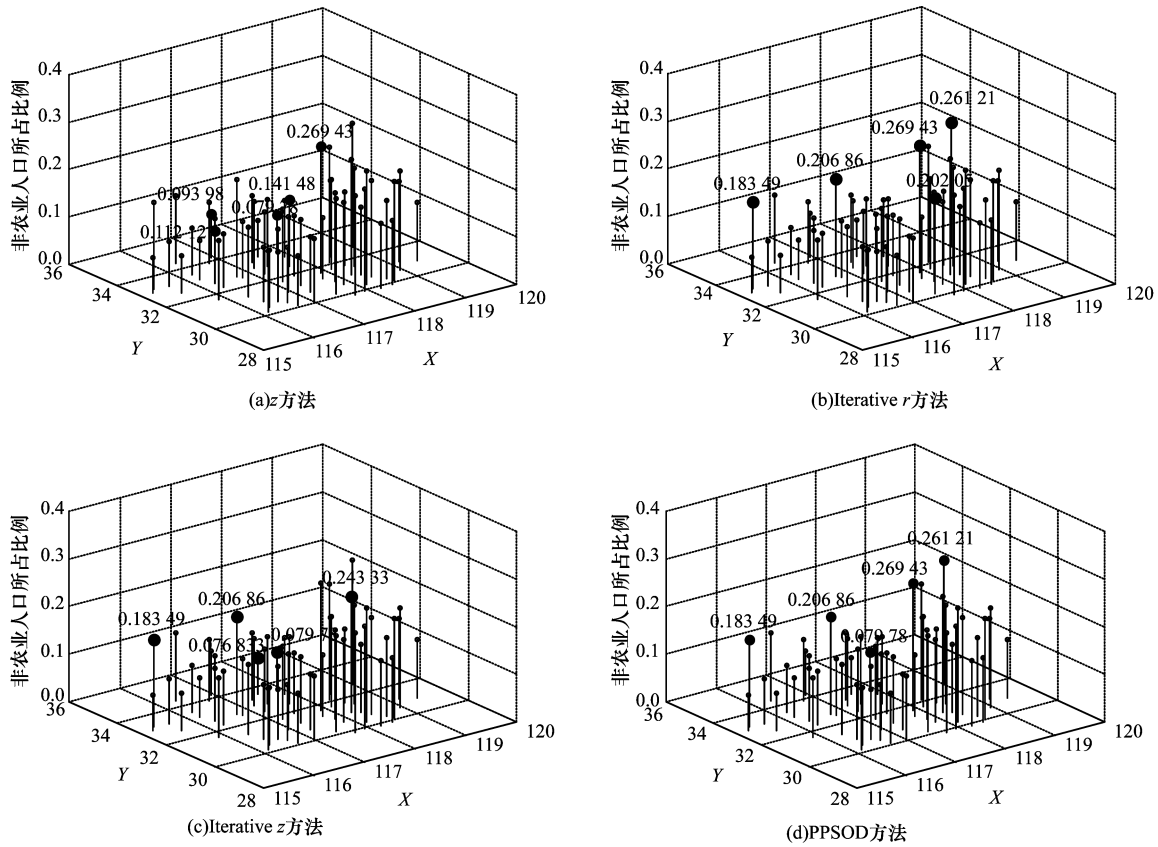


图 4 PPSOD 方法与其他 3 种方法基于数据集 DS_S2 的检测结果对比

表 5 基于数据集 DS_S2 检测出的离群程度最高的 5 个地区

Rank	z algorithm	Iterative r	Iterative z	PPSOD
1	13,泗县	21,界首市	13,泗县	13,泗县
2	1,巢湖市	22,凤台县	22,凤台县	27,天长市
3	25,定远县	1,巢湖市	21,界首市	1,巢湖市
4	20,颍上县	27,天长市	12,灵璧县	22,凤台县
5	11,萧县	46,石台县	36,繁昌县	21,界首市

3.2 多维非空间属性数据集

本文实验是基于单指标方法的,而城市化是一个包括人口、经济、社会等子系统在内的复杂系统的复杂变化过程。因此,城市化是一个以人口非农化、经济非农化和空间利用非农化为基础的区域变化过程[22]。

本节中验证 PPSOD 方法在多维数据集上的实验结果,采用全省分县(市)主要经济指标(2014 年)的数据集 *DS_M*(只有县(市),不包含市辖区),非空间属性集包括生产总值、人均生产总值、地方财政收入等 13 个非空间属性。使用 Mahalanobis 距离计算 *spOF*,检测离群点。

PPSOD 方法在多维数据集 DS_M 上检测出的 5 个县域空间位置分布的详细信息如表 6 所示。

表 6 安徽省 *spOF* 值最高的 5 个地区及其相关属性值(基于数据集 DS_M)

Rank	ID	地名	经度/(°)	纬度/(°)	生产总值/亿元	人均生产总值/元	地方财政收入/万元	人均地方财政收入/元
1	19	阜南县	115.60	32.63	124.72	7 202.00	61 437.00	354.79
2	41	泾县	118.41	30.68	77.46	21 781.00	103 097.00	2 898.96
3	58	休宁县	118.19	29.81	67.85	24 602.00	75 506.00	2 737.93
4	7	涡阳县	116.21	33.49	205.19	12 534.00	112 527.00	687.35
5	6	桐城市	116.94	31.04	217.47	28 742.00	151 269.00	1 999.25

本文对城市化的内涵进行了简化,仅选取了最基本的指标作为实验数据集,可能并不能全面揭示该省城市化发展的水平与格局。结合城镇化空间格局的地理学知识,后续研究还可以基于更能反映城镇化进程的指标作为非空间属性集进行实验。

4　结束语

保护私有信息的空间离群点检测方法目前倍受关注,在气象、交通、城镇化空间布局等地理信息系统领域有着重要的应用前景。本文设计一系列安全协议,基于数据对象与其空间 k 邻域的相关性研究,提出一种隐私保护的空间离群点检测方法。针对单个非空间属性和多维非空间属性 2 种情况,在安徽省人口及经济等真实数据集上进行了实验,结果表明,该方法是高效实用的。

参考文献

[1] Shekhar S, Lu Chang-Tien, Zhang Pusheng. A Unified Approach to Detecting Spatial Outliers [J]. GeoInformatica, 2003, 7(2): 139-166.

[2] Cai Qiao, He Haibo, Man Hong. Spatial Outlier Detection Based on Iterative Self-organizing Learning Model[J]. Neurocomputing, 2013, 117(14): 161-172.

[3] Chen Dechang, Lu Chang-Tien, Kou Yufeng, et al. On Detecting Spatial Outliers [J]. GeoInformatica, 2008, 12(4): 455-475.

[4] Lu Chang-Tien, Chen Dechang, Kou Yufeng. Algorithms for Spatial Outlier Detection [C]//Proceedings of the 3rd IEEE International Conference on Data Mining. Washington D. C., USA: IEEE Press, 2003: 597-600.

[5] Chawla S, Sun P. SLOM: A New Measure for Local Spatial Outliers [J]. Knowledge and Information Systems, 2005, 9(4): 412-429.

[6] Liu Xutong, Chen Feng, Lu Chang-Tien. On Detecting Spatial Categorical Outliers [J]. GeoInformatica, 2014, 18(3): 501-536.

[7] Vaidya J, Clifton C. Privacy-preserving Outlier Detection[C]//Proceedings of the 4th IEEE International Conference on Data Mining. Washington D. C., USA: IEEE Press, 2004: 233-240.

[8] Challagalla A, Dhiraj S S S, Somayajulu D V L N, et al. Privacy Preserving Outlier Detection Using Hierarchical Clustering Methods[C]//Proceedings of the 34th IEEE Annual Computer Software and Applications Conference Workshops. Washington D. C., USA: IEEE Press, 2010: 152-157.

[9] Dai Zaisheng, Huang Liusheng, Zhu Youwen, et al. Privacy Preserving Density-based Outlier Detection[C]// Proceedings of International Conference on Communications and Mobile Computing. Washington D. C., USA: IEEE Press, 2010: 80-85.

[10] Li Lu, Huang Liusheng, Yang Wei, et al. Privacy-Preserving LOF Outlier Detection [J]. Knowledge and Information Systems, 2015, 42(3): 579-597.

[11] Lu E, Pass R. Outlier Privacy [C]//Proceedings of the 12th Theory of Cryptography Conference. Warsaw, Poland: University of Warsaw, 2015: 277-305.

[12] Kou Yufeng, Lu Chang-Tien, Chen Dechang. Spatial Weighted Outlier Detection[C]//Proceedings of the 6th SIAM International Conference on Data Mining. Bethesda, Maryland: Society for Industrial and Applied Mathematics, 2006: 613-617.

[13] Kou Yufeng, Lu Chang-Tien, Santos R F D. Spatial Outlier Detection: A Graph-based Approach [C]// Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence. Washington D. C., USA: IEEE Press, 2007: 281-288.

[14] Cao Lijun, Liu Xiyin, Wang Yubin, et al. WSO-based Spatial Outlier Detection Algorithms [J]. Journal of Networks, 2013, 8(7): 1582-1588.

[15] Yao A C. Protocols for Secure Computations [C]// Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science. Washington D. C., USA: IEEE Press, 1982: 160-164.

[16] Han Jiawei, Kamber M, Pei J. Data Mining: Concepts and Techniques [M]. 3rd Edition. Burlington, USA: Morgan Kaufmann Publishers, 2012.

[17] 罗永龙,黄刘生,徐维江,等. 一个保护私有信息的多边形相交判定协议[J]. 电子学报, 2007, 35(4): 685-691.

[18] Gordon D S, Carmit H, Katz J, et al. Complete Fairness in Secure Two-party Computation[C]//Proceedings of the 40th Annual ACM Symposium on Theory of Computing. New York, USA: ACM Press, 2010: 157-176.

[19] Xue Anrong, Duan Xiqiang, Ma Handa, et al. Privacy Preserving Spatial Outlier Detection [C]//Proceedings of the 9th International Conference for Young Computer Scientists. Washington D. C., USA: IEEE Press, 2008: 714-719.

[20] Murugesan M, Jiang Wei, Clifton C, et al. Efficient Privacy-preserving Similar Document Detection [J]. VLDB Journal, 2010, 19(4): 457-475.

[21] 安徽省统计局. 2014 和 2015 年安徽省统计年鉴[EB/OL]. http://www.ahtjj.gov.cn/tjj/web/tjnj_view.jsp?strColId=13787135717978521&_index=1#.

[22] 薛俊菲,陈　雯,张　蕾. 中国市域综合城市化水平测度与空间格局研究[J]. 经济地理, 2010, 30(12): 2005-2011.

编辑　刘　冰