张忠贵，芦娅. 一种通用的生命线工程网络事件空间聚类分析算法[J]. 灾害学，2015，30(1)：29－33. [Zhang Zhonggui andLu Ya. A General Spatial Clustering Analysis Algorithm of Lifeline Network[J]. Journal of Catastrophology，2015，30(1)：29－33.]

一种通用的生命线工程网络事件空间聚类分析算法*

张忠贵[1,2]，芦 娅[3]

（1. 中国地质大学(武汉）信息工程学院，湖北 武汉 430074；2. 武汉中地数码科技有限公司，湖北 武汉 430074；3. 湖北省地震局，湖北 武汉 430071）

摘 要：网络事件空间聚类分析可发现供水、排水、燃气、电力等生命线工程爆管、漏损事件的高发区域。生命线工程事件由网络边约束，可抽象为网络事件。若不考虑网络拓扑关系，将产生网络事件空间聚类结果与实际分类不符的问题。基于事件网络距离，提出了一种通用的网络事件空间聚类方法，给出了核心概念的形式化定义以及算法描述，可广泛应用于生命线工程事件高发区域的发现，具有较强的实用性。并结合供水管网生命线工程爆管事件高发区域分析实例，给出算法参数的确定原则和范围，验证了所提出的算法的有效性。

关键词：生命线工程；网络；事件；空间聚类；网络距离

中图分类号：P208；X4　　**文献标志码**：A　　**文章编号**：1000－811X(2015)01－0029－05

doi：10.3969/j.issn.1000－811X.2015.01.007

生命线工程设施是维系现代城市与区域经济功能的基础性工程设施[1]，包括供水、排水、燃气、电力等基础设施。随着我国城镇化进程显著加快，城市生命线工程设施规模迅速扩大，运营条件愈发复杂，管道爆管、燃气泄露、漏损等生命线工程事件也进入高发期，根据对规划范围内184 个城市的不完全统计，2000－2003 年因供水管网爆管停水的事故达 13.7 万次，影响的用水范围涉及到总计 3 819 万人次[2]，严重影响了社会公众的生产生活和生命财产安全。生命线工程设施的维护以及改造是降低事件发生频率的根本。但是整个城市建成区范围很大，而每年网络基础设施改造的经费和人力有限，先改造哪些区域才能显著降低事件发生的频率，是管理部门面对的难题。

生命线工程设施具有典型的网络特征，具有紧密的网络拓扑关系。爆管、泄露、漏损等生命线工程事件必然发生于网络上，具有空间聚集性。对于生命线工程事件，仅在城市建成区欧式空间上进行聚类分析，而不考虑其网络拓扑关系，难以准确地契合生命线工程事件的空间分布规律。

本文将生命线工程事件抽象为网络事件，针对目前网络事件空间聚类存在的问题，提出一种普适性的网络事件空间聚类算法，可广泛应用于供水、排水、燃气、电力、电信等具有网络特征的生命线工程事件高发区域的发现，从而为制定合理的改造计划提供支撑，降低突发事件发生的频率。

1 聚类方法分析

1.1 聚类分析原理

聚类分析一个非监督分类的过程，可以形式化描述为：令 $P=\{p_1,p_2,\cdots,p_N\}$ 表示一个包含 N 个实体的数据集，根据相似性度量函数将 P 划分为 $k+1(k\geqslant 1)$ 个簇，即 $P=\{C_0,C_1,C_2,\cdots,C_k\}$，同一簇内实体的相似度要尽可能大于不同簇的实体之间的相似度；其中，C_0 为噪声，$C_i(i\geqslant 1)$ 为簇，且需要满足以下条件[3-4]：

$$\bigcup_{i=0}^{k} C_i = P。 \tag{1}$$

对于 $\forall C_m, C_n \subseteq P, m \neq n$，需同时满足：

$$C_m \cap C_m = \Phi, \tag{2}$$

$$\mathrm{MIN}_{\forall P_i,P_j\in C_m}(\mathrm{Similar}(p_i,p_j)) > \mathrm{MAX}_{\forall P_x\in C_m,\forall P_y\in C_n}(\mathrm{Similar}(p_x,p_y))。 \tag{3}$$

式中：Similar( ) 表示相似性度量函数。

1.2 聚类分析方法选择

聚类分析算法从方法上可分为：划分方法、层次方法、基于密度的方法、基于网格的方法[5-6]。聚类分析算法分类如图 1 所示。

* 收稿日期：2014－04－25　　修回日期：2014－08－11

基金项目：国家“十二五”科技支撑计划(2012BAB11B05，2011BAH06B04)；武汉市应用基础研究计划(2013010501010123)；武汉市关键技术攻关计划(2013010602010191)

作者简介：张忠贵(1981－)，男，湖北宜昌人，博士，工程师，主要从事数字城市研究. E-Mail：zzg_ yc@163.com

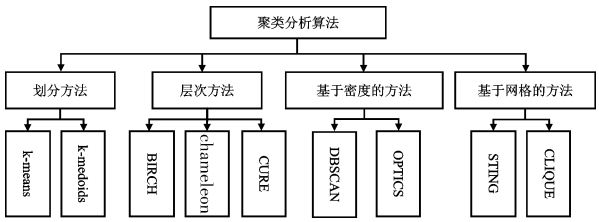

图1 聚类分析算法分类

上述聚类算法中，划分方法和基于密度的方法[6]的研究和应用最为深入和广泛，本文重点对划分方法和基于密度的方法进行分析。其中，k-medoids 难以处理大数据集[7]，K-Means 算法对于离群点敏感。生命线工程事件频发[1]，要求聚类分析算法能够处理较大的数据集。网络事件的本体为网络设施，不能脱离网络设施而存在，均分布在线形网络上，而不是在欧式空间上泊松分布，在空间上可能存在较多噪声点。因此，要求聚类分析算法对离群点不敏感。

根据基于网络突发事件的空间聚类应用场景，结合聚类分析算法特征，可知 k-medoids、K-Means 算法难以适应生命线工程事件空间聚类应用需求。

因此，基于聚类分析算法的应用广泛性、算法效率、离群点敏感程度等指标，选择 DBSCAN[8]作为实现网络事件空间聚类分析算法的基础。

2 面临的问题

网络事件点数据集由网络边约束，决定邻域距离参数需顺着网络边搜索。目前 DBSCAN 算法常用的距离度量方法包括欧式距离、曼哈顿距离等[9]。生命线工程事件总是分布在线形网络上，采用欧氏距离判定核心对象、密度可达、密度相连，会产生与实际分类不符的情况。

本文以图 2 为例，详细描述上述问题。某街区生命线工程管道 A 上发生 3 起漏损事件，生命线工程管道 B 发生 1 起漏损事件、管道 C 发生 1 起漏损事件、管道 D 发生 1 起漏损事件。漏损事件的空间及属性信息如表 1 所示。

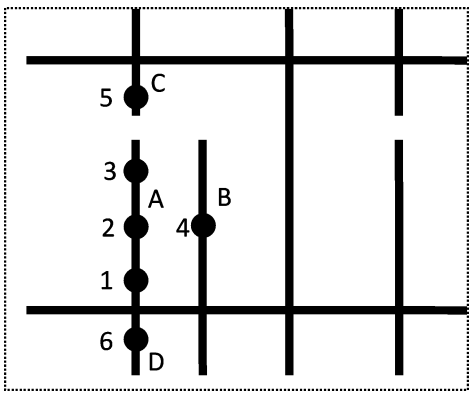

图2 漏损事件分布

表1 事件表

事件编码	事件坐标	设施类型	设施编码
1	(5.0，3.0)	支管	A
2	(5.0，4.0)	支管	A
3	(5.0，5.0)	支管	A
4	(6.0，4.0)	支管	B
5	(5.0，6.0)	支管	C
6	(5.0，2.0)	支管	D

对于图 2 所示的网络事件集合，设空间半径 $\varepsilon=2.0$、邻域密度阈值(Min$Pts=3$)，基于欧氏距离 DBSCAN 算法可能得到的聚类结果如图 3 所示。

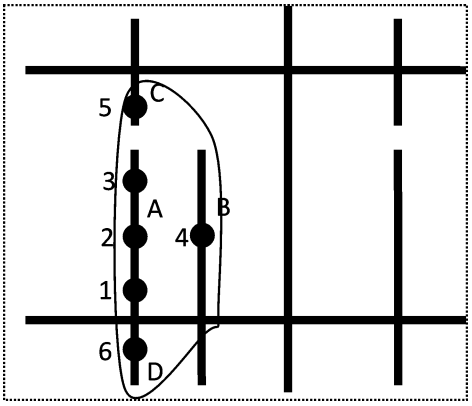

图3 漏损事件聚类结果(基于欧氏距离)

基于欧氏距离的 DBSCAN 算法将所有的事件划为一个簇：$C=\{1，2，3，4，5，6\}$，管道 A、B、C、D 改造的优先级相同，难以提高改造的效率。

3 核心概念形式化定义

网络事件空间聚类算法的主要原则是：网络事件具有空间聚集性，基于网络距离，以核心网络事件为基础，通过密度可达，发现网络事件簇。以下对算法中涉及核心概念的进行形式化定义[8,10]。

定义 1 网络事件 E 的 ε－邻域 $N_\varepsilon(E)$：以事件 E 为中心，网络事件 E_i与事件 E 的网络距离(网络最短距离)小于等于半径参数 ε 的事件集合。形式化描述为：

$$N_\varepsilon(E)=\{E，E_i\in D \mid \mathrm{NetWorkDist}(E，E_i)\leqslant\varepsilon\}。\quad(4)$$

定义 2 核心网络事件 E：如果网络事件 E 的 ε－邻域至少包含最小数目 *MinPts* 的网络事件，则称该网络事件是核心网络事件。形式化描述为：

$$\text{核心网络事件}\ E：\mid N_\varepsilon(E)\mid\geqslant\mathrm{min}Pts。\quad(5)$$

定义 3 直接密度可达：如果网络事件 E_j是在 E_i的 ε－领域内，而 E_i是一个核心网络事件，则称网络事件 E_j从网络事件 E_i出发是直接密度可达的。形式化描述为：

$$E_j\text{从}\ E_i\text{直接密度可达：}E_j\in N_\varepsilon(E_i)，\mid N_\varepsilon(E_i)\mid\geqslant\mathrm{min}Pts。\quad(6)$$

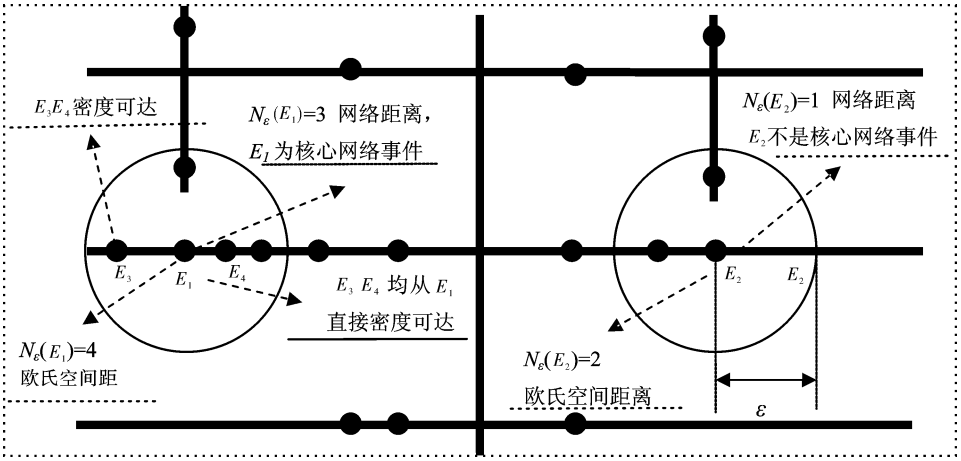

图 4　网络事件空间聚类算法核心概念

定义 4　密度可达：存在网络事件序列 E_1，E_2，…，E_n，若 E_{i+1}从 E_i直接密度可达，则 E_N从 E_1出发密度可达。形式化描述为：

$$\text{若 } E_i,\ E_j \in N_\varepsilon(E_n),\ \text{则 } E_i,\ E_j \text{密度可达。} \tag{7}$$

定义 5　网络事件簇与噪声：基于密度可达的最大密度相连的网络事件集合称为网络事件簇。不属于任何网络事件簇的网络对象为噪声。

上述核心概念的形式化定义可通过图 4 直观说明。基于上述核心概念的定义，提出网络事件空间聚类的核心数据结构：事件网络距离矩阵、*st* - 邻域、邻接表，如表 2 所示。

表 2　核心数据结构

概念	定义
事件网络距离矩阵	*double* * * *MatrixEventNetDis*; *MatrixEventNetDis* * = *new double* *[*max*(*EventID*)]; *MatrixEventNetDis*[*i*] = *new double*[*max*(*EventID*)];
事件 ε - 邻域	*typedefCArray*<*long*, *long*&>*Link_ Data*;
邻接表	*CMap*<*long*, *long*&, *Link_ Data*, *Link_ Data*&>*EventLinkhashList*;

4　网络事件空间聚类算法描述

给定：网络 *NetWork*、事件集合 *EventSet* = {I_0，I_1，I_2，…，I_n}、空间半径 ε、邻域密度阀值（*MinPts*），网络事件空间聚类算法可描述为：

STEP 1 构建事件网络距离矩阵 *MatrixEventNetDis* = *BulidMatrix*（*NetWork*，*EventSet*）；

STEP 2 构建事件邻接表 *EventLinkhashList* = *BulidLinkList*（*MatrixEventNetDis*，ε）；

STEP 3 构建事件访问标识数组 *NewEventVisitFlag*[*Max*(*EventID*)]，均置 0；

STEP 4 构建噪声簇 *New Noise*；初始化事件均为噪声 *Noise* = *EventSet*；

STEP 5 构建簇集合 *NewClusterSet*；*ClusterSet* += *Noise*；

STEP 6 对于任意网络事件 E_i，$E_i \in$ *EventSet*，设置访问标识为 1。根据事件邻接表 *EventLinkhashList* 查找事件 E_i的 ε - 邻域 $N_\varepsilon(E_i)$，并判断 E_i是否为核心网络事件。若 E_i为核心网络事件，则建立一个新簇 *C* 包含 E_i[5]，并从噪声集合中移除 E_i。

STEP 7 对于 $E_j \in N_\varepsilon(E_i)$，设置访问标识为 1。根据事件邻接表 *EventLinkhashList* 查找事件 E_j的 ε - 邻域 $N_\varepsilon(E_j)$，并判断 E_j是否为核心网络事件。若 E_j为核心网络事件，则根据密度连接原则，扩展 $N_\varepsilon(E_i)$[4]：$N_\varepsilon(E_i)$ += $N_\varepsilon(E_j)$；若事件 E_j不属于任意簇，添加事件 E_j到簇 *C*，并从噪声集合中移除 E_j。

STEP 8 循环执行 STEP7，直至 $N_\varepsilon(E_i)$遍历完成，则簇 *C* 构建完成，添加簇 *C* 到簇集合 *ClusterSet*。

STEP 9 循环执行 STEP6、STEP7、STEP8，直到所有网络事件的访问标识为 1 时，算法结束，输出簇集合 *ClusterSet*。

为了便于算法描述，本文以图 2 为例，设空间半径 ε = 2.0、邻域密度阀值（*MinPts* = 3），对于网络事件空间聚类算法的核心步骤进行说明。

4.1　事件网络距离矩阵与事件邻接表构建

基于网络的最短路径分析算法，实现事件网络距离矩阵的构建，算法伪码如下所示。

STEP 1 求解网络事件 $Event_i$、$Event_j$所属的网段 Lin_i、Lin_j，以及与之最近的网段 Lin_i、Lin_j链接的结点 $Node_i$、$Node_j$。

STEP 2 求解 $Node_i$、$Node_j$网络最短路径 *ShortPath* = {Lin_1，Lin_2，…，Lin_n}。网络最短路径算法主要包括最短优先搜索算法（LS 算法，如 Dijkstra 算法[11]），列表搜索算法（DS 算法，如 queue、deque 算法）。长期以来，专家学者们对 LS 和 LC 算法的效率比较进行了深入的研究[12]，在 GIS 中已经提供了成熟的接口，本文不再赘述。

STEP 3 求解最短路径距离 *PathDis*，最短路径距离即为最短路径网段长度的和。

$$PathDis = \sum_{k=0}^{n-1} Len(lin_k)。 \tag{8}$$

STEP 4 求解事件点 $Event_i$与 $Node_i$的网段距离 ΔDis_i；若 $Lin_i \in ShortPath$，则 $\Delta Dis_i = -\Delta Dis_i$。

STEP 5 求解事件点 $Event_j$ 与 $Node_j$ 的网段距离 ΔDis_j；若 $Lin_j \in ShortPath$，则 $\Delta Dis_j = -\Delta Dis_j$。

STEP 6 求解线性网络空间距离 $NetDis = pathDis + \Delta Dis_i + \Delta Dis_j$，并赋值：

$MatrixEventNetDis[Event_i][Event_j] = NetDis$；$MatrixEventNetDis[Event_j][Event_i] = NetDis$。

基于上述算法，构建的网络距离矩阵如下：

$$\begin{bmatrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 0 & 1.0 & 2.0 & 3.0 & 10.0 & 1.0 \\ 2 & 1.0 & 0 & 1.0 & 4.0 & 11.0 & 2.0 \\ 3 & 2.0 & 1.0 & 0 & 5.0 & 12.0 & 3.0 \\ 4 & 3.0 & 4.0 & 5.0 & 0 & 9.0 & 3.0 \\ 5 & 10.0 & 11.0 & 12.0 & 9.0 & 0 & 10.0 \\ 6 & 1.0 & 2.0 & 3.0 & 3.0 & 10.0 & 0 \end{bmatrix}。 \quad (9)$$

基于线性距离矩阵，以距离半径为判断标准，构建网络事件邻接表，如图 5 所示。

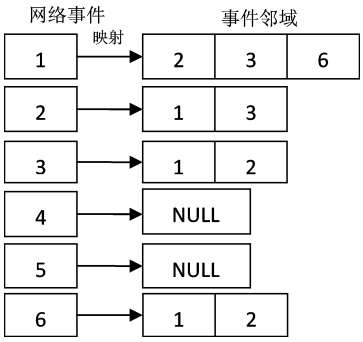

图 5 网络事件邻接表

4.2 事件簇构建算法

以核心网络事件为基础，构建事件簇。执行如下构建簇的算法，得到如图 6 所示的漏损事件聚类结果：

遍历事件集合，当所有事件均访问时终止

For ($E_i \in EventSet \&\& EventVisitFlag[E_i] == 0 \&\& \forall\ (EventVisitFlag) == 1$)

{

$EventVisitFlag[E_i] = 1$;

求解事件 E_i 的 ε – 邻域 $N_\varepsilon(E_i)$

if ($EventLinkhashList.Lookup(E_i, N_\varepsilon(E_i)) != TRUE$) continue;

若事件 E_i 为网络核心事件

if($Num(N_\varepsilon(E_i)) \geqslant MinPts$)

{

创建新簇，添加事件 E_i，并从噪声移出

$New\ Cluster$; $Cluster += E_i$; $Noise -= E_i$;

For($E_j \in N_\varepsilon(E_i)$)

{

$EventVisitFlag[E_j] = 1$;

求解事件 E_j 的 ε – 邻域 $N_\varepsilon(E_j)$;

if ($EventLinkhashList.Lookup(E_j, N_\varepsilon(E_j)) != TRUE$) continue;

若事件 E_j 为网络核心事件，E_i 与 E_j 直接密度可达，扩展事件 E_i 的 ε – 邻域

if($Num(N_\varepsilon(E_j)) \geqslant MinPts$) $N_\varepsilon(E_i) += N_\varepsilon(E_j)$

若事件 E_j 不属于任意簇，添加事件 E_j，并从噪声移出

if ($E_j \notin \forall\ (ClusterSet)$) $Cluster += E_j$; $Noise -= E_j$

}

添加簇 $Cluster$ 到簇集合 $ClusterSet += Cluster$

}

}

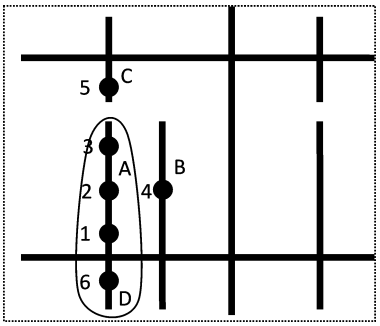

图 6 漏损事件聚类结果

网络事件空间聚类将事件划分为 2 簇：$C = \{\{4,5\},\{1,2,3,6\}\}$，其中 $\{4,5\}$ 为噪声。在进行管网改造计划制定时，可首先改造管道 A、D，从而可在人力和资金制约的情况下，降低事件发生的频率。

5 网络事件空间聚类算法的应用

网络事件空间聚类算法需要确定半径参数、邻域密度阈值。参数的选择对于聚类分析结果的合理性具有很大的影响。参数的确认很大程度依赖于领域知识、网络空间分布情况、以及网络事件发生的频率。本文以供水管网事件空间聚类为例，提出了以下原则供参考：

(1)半径参数的取值范围：网络事件空间聚类的目的之一为制定更合理的管网养护或改扩建计划提供参考依据。因此，半径参数的取值范围为历史维修或改扩建管网长度的最大值和最小值[10]。半径参数的取值可选择接近历史维修管网长度的平均值。

(2)邻域密度阈值的取值范围：MinPts 选择为 2，会导致聚类分析产生的簇太小，从而产生过多离散的簇，这对于发现生命线工程事件高发区域没有太大意义。因此，MinPts 必须在 3 及以上[10]，从而避免产生过多太小的簇。

结合爆管事件的网络空间散点图，本文选择半径参数为 300 m、邻域密度阈值为 3，基于网络事件空间聚类算法，进行供水管网爆管事件高发区域分析，获取一定时间内投诉事件最多、较多、一般或较少的区域，如图 7 阴影区域所示。同时还可通过空间统计分析，了解该区域内爆管的主要原因，辅助管理人员采取有针对性的措施：如加强管网巡检和养护，制定管网改造计划等，从而

降低事件发生的频率。

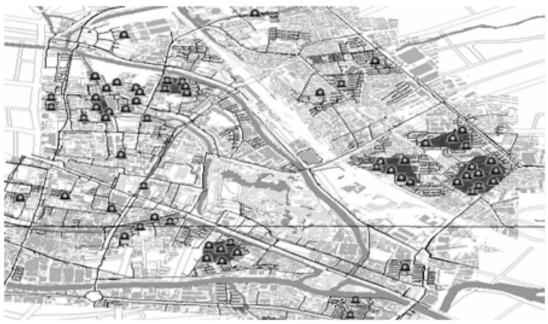

图 7　爆管事件高发区域分析

6　结束语

当前我国供水、排水、燃气、电力等生命线工程设施的爆管、漏损事件进入高发期。通过网络事件的空间聚类分析，在整个城市范围内发现突发事件高发区域，从而为制定合理的改造计划提供支撑，降低突发事件发生的频率，具有较强的现实和理论意义。

本文从网络事件空间聚类面临的问题出发，给出了网络事件空间聚类核心概念的形式化定义，提出了一种通用的网络事件空间聚类算法，并成功运用于城市供水管网的爆管事件高发区域分析，证明了算法的有效性和实用性。

网络事件的发生与时间密切相关，本文后续需综合考虑网络事件的时空变化特征，进一步研究从网络事件时空聚类方法，发现网络事件的时空分布规律。

参考文献：

[1] 李杰，刘威，卫书麟. 生命线工程网络抗震拓扑优化研究[J]. 灾害学，2010，25(Supp. 1)：4－9.

[2] 宋兰合. 城市供水管网问题突出亟待改造[J]. 建设科技，2008(5)：72－73.

[3] 唐建波，邓敏，刘启亮. 时空事件聚类分析方法研究[J]. 地理信息世界，2013(1)：38－45.

[4] 邓敏，刘启亮，王佳璨等. 时空聚类分析的普适性方法[J]. 中国科学：信息科学，2012，42(1)：111－124.

[5] Jianwei Han，Micheline Kamber，Jian Pei. 数据挖掘概念与技术[M]. 范明，孟小峰，译. 北京：机械工业出版社，2012.

[6] Jain A k. Data clustering：50 years beyond K-means[J]. Pattern Recognition Letters，2010，31(8)：651－666.

[7] 宋爱琪，刘晓红，吴国洋. GML 时空聚类算法性能综述[J]. 测绘标准化，2011，27(4)：9－11.

[8] Ester M，Kriegel H，Sander. A density-based algorithm for discovering clusters in large spatial database with noise[C]//Proceedings of the 2nd International Conference on Knowledge and Discovery and Data Mining. Portland，OR：[s. n.]，1996：45－50.

[9] 程显洲，肖兰喜，董翔，等. 基于烈度衰减椭圆闭值空间散点聚类研究——以 12322 灾情速报平台为例[J]. 灾害学，2013，28(4)：205－208.

[10] De Oliveira D P，Garrett Jr J H，Soibelman L. A density-based spatial clustering approach for defining local indicators of drinking water distribution pipe breakage[J]. Advanced Engineering Informatics，2011，25(2)：380－389.

[11] 刘春年，邓青菁. 应急决策信息系统最优路径研究——基于路阻函数理论及 Dijkstra 算法[J]. 灾害学，2014，29(3)：18－23.

[12] 陆锋. 最短路径算法：分类体系与研究进展[J]. 测绘学报，2001，30(3)：269－275.

A General Spatial Clustering Analysis Algorithm of Lifeline Network Event

Zhang Zhonggui[1, 2] and Lu Ya[3]

(1. *Faculty of Information Engineering*, *China University of Geosciences*, *Wuhan* 430074, *China*; 2. *Wuhan Zondy Cyber-tech Co.*, *ltd.*, *Wuhan* 430074, *China*; 3. *Earthquake Administration of Hubei Province*, *Wuhan* 430071, *China*)

Abstract: Spatial clustering analysis of network events can be used to find high incident area of lifeline infrastructure (such as water supply, drainage, gas and electricity). Lifeline events are constrained by the network side and can be abstracted as network events. Spatial clustering will result inconsistent with the actual classification problem without considering the network topology. Based on events network distance, a general spatial clustering analysis algorithm of network event is proposed, which could be widely used to find high incidence areas of lifeline, and with strong practicality. Combined with the example analysis of events in high-risk areas of water supply pipe network of lifeline engineering blasting, the effectiveness of the proposed algorithm are verified with the example of the analysis burst high incidence area.

Key words: lifeline; network; event; spatial clustering; network distance