基于混合聚类的数字电影流动放映轨迹点分析模型

国家新闻出版广电总局电影数字节目管理中心 黄昭婷
北京工业大学信息学部信息与通信工程学院 张春梅 王明玉

【摘要】针对我国数字电影流动放映轨迹点特征，本文提出一种基于DBSCAN（Density Based Spatial Clustering of Applications with Noise）的密度聚类算法和基于STING（Statistical Ingormation Crid）的网格聚类算法相结合的数字电影流动放映轨迹点分析模型。本文首先借鉴文献［1］中的改进DBSCAN算法对流动放映轨迹点进行一次微聚类，挖掘出放映密集区域。然后，利用STING算法对放映密度区域进行二次聚类，获得放映密集区域的代表中心点，使聚类结果更加细化。最后以某省份的回传GPS放映轨迹数据进行实际聚类分析。实验结果表明，该算法能够快速反映并确定放映轨迹的空间分布，挖掘出放映密集区域及其中心点，可以较为直观的为管理部门合理分配和调度放映员提供科学理论依据。

【关键词】数字电影流动放映 放映轨迹点 混合聚类 DBSCAN算法 STING算法

1 引言

数字电影流动放映是数字电影放映的一个重要组成方面，旨在解决广大农村、社区和偏远地区看电影难的问题。为了使各个行政村都能受益于“一村一月一部电影”的公益电影放映工程，县市区工作站点在对放映员的放映轨迹进行布局编排时要求其空间分布尽可能均匀，然而实际执行过程中，有可能存在放映密集的区域。一方面是由于偏远地区的行政村间隔甚远，而某些放映员为了减少奔波，经常去邻近的区域放映。另一方面也存在放映工作安排分配不合理的情况，使得多个放映员在一个月内重复去同一区域放映。这使得各行政村放映场次分布不均衡，局部放映密度高。因此对放映高密集区域进行挖掘并以数据分析的结果为依据对放映员的放映轨迹进行局部优化，具有重要的研究意义。

空间聚类是当前地理位置数据挖掘与知识发现的一项关键技术，在无先验知识的情况下，可以从大量地理位置数据中挖掘隐含规律，可以识别空间中对象的密集区域，揭示空间实体的分布规律，适合本文数字电影流动放映轨迹的研究。目前国内外学者已对空间聚类进行深入研究，并广泛应用于城市管理、生态环境、市场分析等不同的领域。如文献［3］利用基于K－means的空间聚类算法在城市应急机构选址方面进行了应用；文献［4］利用空间聚类分析发现入侵物种分布规律；文献［5］利用基于密度和基于网格的空间聚类算法在企业管理的客户划分方面取得了较好的效果；文献［6］利用空间聚类分析方法，结合农机空间运行轨迹的特点，实现农机作业状态自动识别。到目前为止，国内外关于对数字电影流动放映轨迹数据分析的文献报道还比较少。

空间聚类按照一定的距离或相似性为准则，根据空间实体的特征，对空间数据集进行聚类。与传统的聚类相比，空间聚类主要考虑空间模式间的远近、拓扑、方位、稀疏关系等空间上的结构特征。现有的空间聚类算法按照隶属度的取值范围可以分为两类：一类是软聚类。在这类算法中，一个样本可以同时属于几个聚类集合，但是属于各个类别隶

属度在区间［0，1］之间，代表算法有模糊 C 均值聚类（FCM）、期望最大化算法（EM）等。软聚类通常能够直观的将聚类结果进行展示，但是由于该聚类方法需要反复迭代计算，计算工作量会比较大。另一类是硬聚类。硬聚类可以看做模糊聚类的一个特例，隶属度为 0 或 1，即每个样本只能属于一个聚类集合。硬聚类方法根据空间对象之间的相似度大致又可分为两大类[7-9]：（1）距离。通常以欧氏距离计算点间距离，常用于划分、层次、模型。基于划分的方法如 K－means 和 K－medoids 算法。基于层次的方法，如 CURE 和 BIRCH。基于模型的方法如 COBWEB。然而，距离相似度仅表征点对点的相似信息，缺乏多点或局部区域目标的聚集特征的判断。（2）密度。通过计算指定范围内点目标的分布密度判断聚集程度，常用于密度和网格聚类。基于密度的方法，如 DBSCAN 和 OPTICS。基于网格的方法，如 STING 和 CLIQUE。但是基于密度和基于网格的计算依赖于用户对区域半径、密度阈值、格网尺度等核心参数的设置，设置不当可能造成聚类效果下降。

数字电影流动放映轨迹是放映员在地理空间中移动放映过程的记录，其数据表现为携带放映位置与时间信息的离散空间点 P（x，y，t），包含了放映员在整个移动过程中的时间信息、位置信息以及特定的放映行为或移动模式。考虑数字电影流动放映具有流动性、集中性、随机性的特点，其放映轨迹点形成的簇不一定是球状，可能是任意形状。基于密度的方法和基于网格的算法无需事先确定簇的个数，可以发现任意形状的簇，适合于对未知内容的数据集进行聚类。因此，本文采用这两类算法中的代表算法 DBSCAN 和 STING 对数字电影流动放映轨迹点进行研究。

本文在研究空间聚类算法的基础上，结合放映员放映轨迹点特征，借鉴文献［1］中改进的 DBSCAN 算法对流动放映轨迹点进行一次微聚类，挖掘出放映密集区域，然后利用 STING 算法对放映密度区域进行二次聚类，获得放映密集区域的代表中心点，使聚类结果更加细化，最后以某省份的回传 GPS 放映轨迹数据进行实际聚类分析，并以核密度热点分布图展示放映分布情况。实验表明该算法可以快速反映并确定各个县级地区数字电影流动放映的总体分布情况，实现了放映高密度区域的识别，并确定了放映高密度区域的中心点，能够直观的为管理部门合理分配放映员提供决策支持。

2 基于 DBSCAN 和 STING 的混合聚类算法

2.1 DBSCAN 算法

基于 DBSCAN 聚类方法的放映轨迹点聚类分析的主要思想是：以数据集在空间分布上的稠密程度为依据，搜索 Eps 邻域半径内放映次数 Minpts 以上的区域，过滤放映次数较低的区域，以实现对局部放映密集区域的识别。涉及到的相关定义如下[8]：

①直接密度可达（directly density reachable）：给定一个点集合 D，如果 p 在 q 的 Eps 邻域内，且 q 是一个核心点，则我们说点 p 从点 q 出发是直接密度可达的。

②密度可达（density－reachable）：对于样本集 D，如果存在一个对象链 $p_1=q$，$p_n=p$，$p_i\in D$（$1\leq i\leq n$）对于 P_{i+1} 是 P_i 从关于 Eps 和 Minpts 直接密度可达的，则点 p 从点 q 是密度可达的。

③密度相连（density－connected）：如果存在点 $o\in D$，使点 S 和 R 都是从 o 关于 Eps 和 Minpts 密度可达的，那么对象 S 到 R 是关于 Eps 和 Minpts 密度相连的。

④核心点（core points）和边界点（border points）：如果一个点 p 的 Eps 邻域内包含的点数大于等于 Minpt，则称 p 为核心点，否则称其为边界点。

DBSCAN 算法描述如下：

步骤 1、扫描预处理过的放映轨迹点数据集 D，在属性数据库中增加一个新字段 ID（数值型），用于存储聚类结果，初始化所有记录的 ID 值为零。定义搜索数据集 S，用于临时存储检索结果，初始化参数 Minpts 和 Eps。

步骤 2、遍历数据集 D，依次搜索每个点的 Eps 邻域，并为该点建立邻接表来存储邻域内所有大于 Minpts 的数据点；

步骤 3、遍历数据集 D，将每个点作为种子点

进行考察：

①对于点 P_i，如果 ID 为零，则搜索其邻接表；如果邻接表链表中的坐标点数≥Minpts，那么标记点 P_i 为核心点，将 ID 设为 cluster，同时将 P_i 的邻接表链表包含的所有点存入 S 中。

②遍历 S，将每个点作为种子点进行考察，对于点 q_i，如果 ID 为零，搜索其邻接表，如果邻接表链表中的坐标点数≥Minpts，则 q_i 也是个核心点，同时它是点 P_i 的直接密度可达点，与 P_i 属于同一类，将 q_i 的 ID 设置为 cluster；否则 q_i 为边界点，同样将 q_i 的 ID 设置为 cluster。如果 q_i 是核心点，点 o 存在于 q_i 邻接表链表中并且 o 不属于 S，则将点 o 存入 S 中。最后将点 q_i 从 S 中删除。

③搜索 S 中的下一个坐标点，若 S 非空，则执行②。

步骤 4、搜索 D 中的下一个坐标点，若 S 非空，并将 cluster 加 1，执行步骤 2，直至遍历完数据集。

步骤 5、删除搜索数据集 S。

至此 DBSCAN 聚类结束，对数据集 D 仅搜索一次即可得到最终结果。属性数据库中记录了聚类结果，其中字段 ID 值为零的点为噪声点。

2.2 STING 算法

STING 算法的主要思想是采用网格划分技术将数据空间网格化，利用存储在网格单元中的统计信息，寻找密集单元格的连通区域完成聚类。涉及到的相关定义如下：

1. 网格划分参数 K：数据空间 S 的每一维等分为 K 个部分，默认情况下采用均匀分布估计单元格边长进行划分。

2. 密度阈值 Den（g）：每个单元格内落入的点个数的阈值，点个数≥阈值 Den（g），则该单元格显著。

3. 连通区域阈值（Min）：一个连通区域包含的最小单元格的个数。这里，我们设为固定值 5。

STING 算法的描述如下：

步骤 1、根据自动网格划分参数 K，将空间划分为大小相等的网格，因为空间数据库是二维的，所以网格为矩形区域；

步骤 2、将待处理的数据集映射至已划分的网格中，统计落入每个网格中的数据点个数 count，若 count ＞网格单元密度 den（g），则为高密度网格单元；

步骤 3、扫描所有高密度网格单元，当发现第一个高密度网格时，便以该网格开始扩展，搜索与该网格邻接并且也是高密度的网格，加入密集区域，如此迭代，直至不再有这样的网格出现；

步骤 4、遍历所有密集区域，统计密集区域内网格的个数，将网格个数小于连通区域阈值的区域删掉，即为连通区域。提取连通区域，求出每个连通分支中总密度最大的网格的质心，将其作为类簇代表中心；

步骤 5、输出聚类结果与代表中心点坐标。

2.3 基于 DBSCAN 和 STING 的混合算法

图 1 混合聚类流程图

由于空间数据具有二重性，即在最高层次上，空间对象呈现为若干任意形状的高密度密集分布“带”，而在次一层次上，即密集带内部的子区域，空间对象又往往会以几个特别密集的区域为中心，呈现若干辐射状的密集分布“圈”[10]。文献［1］直接用 DBSCAN 混合聚类算法得到的是次一层次上的密集区域，即一组密集点的集合，无法直接提供给管理部门查询。因此本文在文献［1］的基础上，提

出基于 DBSCAN 和 STING 的混合算法聚类模型，进一步提取密集子区域的代表中心点，剔除密集区域中的离群点，使结果更加细化，方便日后推荐系统使用。混合算法的流程如图 1 所示。

3 实验分析

3.1 实验环境及数据预处理

本实验的数据来源于某个省份 2016 年 9 月 1 号到 2016 年 10 月 15 号的实际放映位置数据，共计 5029 条，每条记录中包含的主要属性如下：ID、回传时间、放映时间、回传 GPS 经纬度坐标、GPS 换算后的百度坐标等。实验平台为 Windows 7 操作系统，Pentium（R）Dualcore 4.0G CPU，Matlab 2013b 编程软件。

数据清洗在保证数据完整性和有效性的前期处理过程中有着重要意义，在本实验中需要对原始数据进行以下情况的过滤清洗：①缺失值清洗。若位置信息和时间中任一字段值缺失，则该条记录就失去了实际应用价值，这种情况应剔除该项记录，以减少应用的复杂程度。②重复记录清洗。针对放映员在同一地点回传多次的情况，应根据时间轴对同

图 2 POI 示意图

一放映点的放映记录进行顺序排序，只取该放映点第一次回传的放映信息。③经纬度数据越界清洗。原始数据中有一些明显超出某省份范围的记录，应将其删除。

得到清洗后的数据集后，将转换后的百度坐标映射到百度地图上，如图 2 所示。从图 2 中可以看出，由于村与村之间距离较近，POI（point of interest）数据之间互相重叠，很难用肉眼直接判断到底哪些区域点密集度较高。因此需要对流动放映轨迹点进行挖掘分析，将挖掘的结果较为直观的提供管理部门做决策。

3.2 参数分析

①Eps 参数的选取。由于本文数据集簇密度差异不是特别明显，因此这里根据文献［10］提出的方法，计算出输入数据集的距离分布矩阵 DISTn×n。用 DISTi×n 表示 DISTn×n 中第 i 列的值，对 DISTi×n 中每一列进行升序排列得到 KNN 分布，如图 3 所示。从图 3 中可以看出，从 K＝6 处的红色曲线开始，曲线从下往上逐渐变密集，可以反映出其它 K－dist 曲线的形状。通过观察 K＝6 的 K－dist 曲线，将急剧发生变化的位置所对应的值确定为半径 Eps 的值，为 0.03。

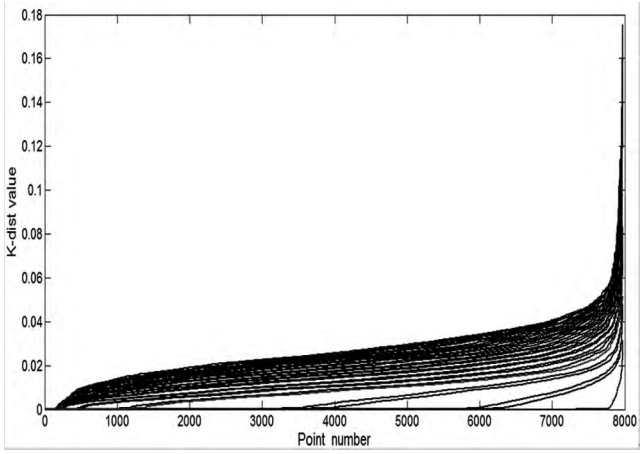

图 3 KNN 分布图

②Minpts[11] 参数的选取。在上述 Eps 确定的情况下，统计数据集中每个点的 Eps 域内点的数目，然后对整个数据集中每个点的 Eps 域内点的数目求数学期望得到 Minpts，此时的 Minpts 为每个聚类中核心对象 Eps 内的数据点个数的最优值。Minpts 的计算公式为：

$$\mathrm{Minpts}=\frac{1}{n}\sum_{i=1}^{n}p_i \tag{1}$$

其中，p_i 为 i 点的 Eps 领域内点的个数。

③网格参数的选取。这里我们采用人工多次实验的选取方法，由于篇幅有限，在此就不一一赘述。

3.3 结果分析与性能比较

为了验证混合算法的有效性，我们将本文提出的 DBSCAN 和 STING 混合聚类算法与模糊聚类算法 FCM 进行分析对比实验。DBSCAN 算法中，参数 Minpts 和 Eps 的值分别取为 0.03 和 35。STING 聚类中，网格大小设置 0.017×0.016、密度阈值设为 5。在 FCM 算法聚类中，聚类个数 K 取为 6。

①基于 DBSCA 的放映密集区域挖掘

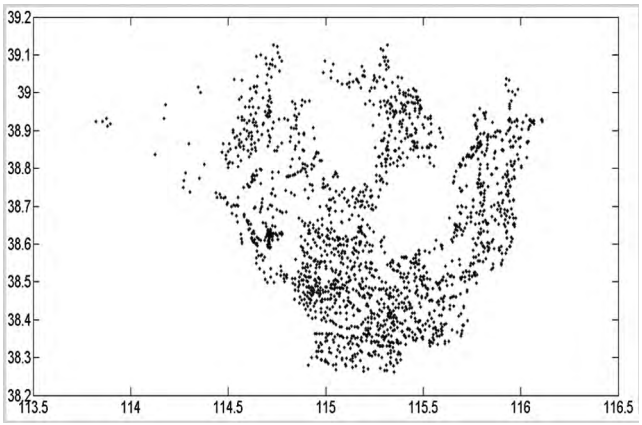

图 4 二维散点图

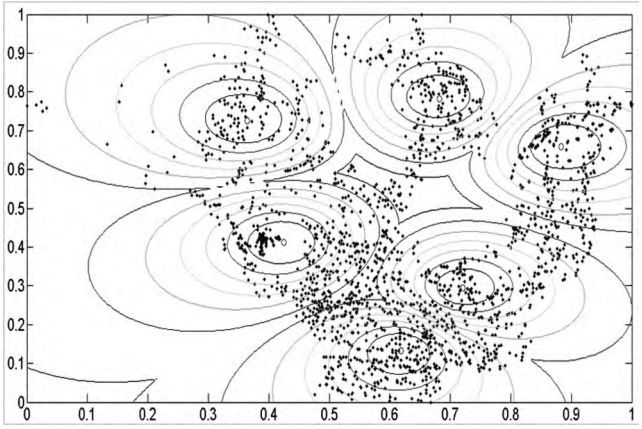

图 5 FCM 聚类结果图

本文首先画出清洗后的放映轨迹原始数据的二维散点图，如图 4 所示，然后利用放映轨迹原始数据分别进行 FCM 聚类和 DBSCAN 聚类，如图 5 和图 6 所示。

从图 5 中可以看出，FCM 聚类算法只是按照基于目标函数最优解方法将实验样本数据进行分类，并没有有效的将放映密集区域与稀疏区域分开，尤其邻近类间存在较多的重叠结构时，分类效果不理想。同时，受噪点的影响，每个簇的中心点不一定是高密度区域的中心。而 DBSCAN 算法不仅能对实验数据进行分类，还能对放映密集区域进行识别，如图 6 所示，其中红色（色块）部分为密集区域。

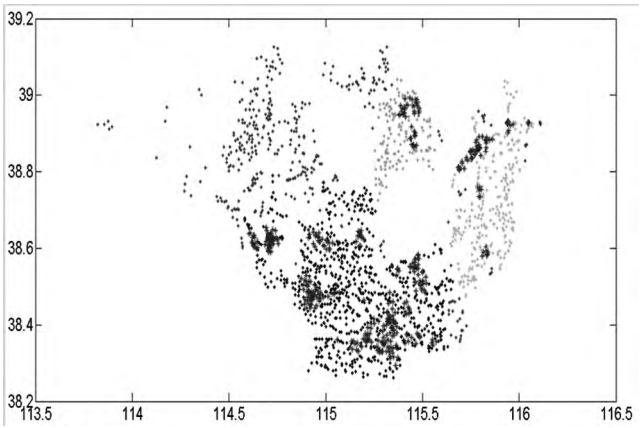

图 6 DBSCAN 聚类结果

此外，由于 FCM 算法对初值的选取十分敏感，容易陷入迭代局部最优，使得聚类效果较差，且需要事先确定类的数量 K，然而本例中的聚类数量是未知的。因此，FCM 模糊聚类算法不适合本文研究。而 DBSCAN 算法对实验样本数据进行聚类时，可以发现任意形状的簇，而且不需要事先确定簇的个数，也可以较为容易的对参数的取值进行选取，因此本文选取 DBSCAN 聚类算法进行下一步的实验。

②基于 STING 的类簇代表中心点确定

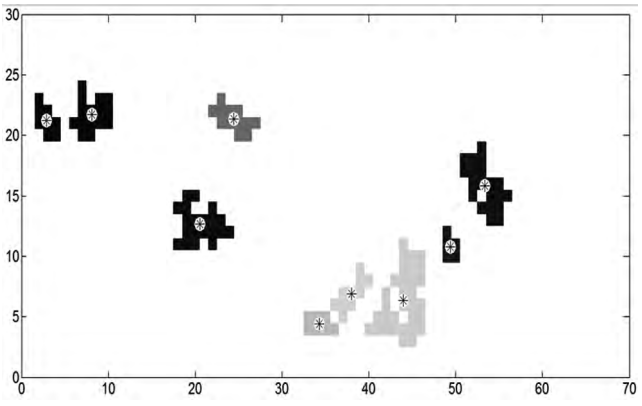

图 7 STING 聚类结果

类簇代表中心点是指局部密度较大，距离其他局部密度较大的点较远的点[12]。本文根据上述放映密集区域的挖掘结果，取1260个样本点，利用STING聚类算法再次进行细分聚类，寻找连通区域的质心来作为类簇代表中心点，聚类结果如图7所示。

从图7中可以看出，STING对密集区域中的噪声数据进行了清理，而这些噪声数据实际上是一些零散的放映较为密集的点，但不是最密集的区域，在这里我们予以删除，不做研究。因此，从总体来看，STING基本获取了连通区域以及连通区域的质心，使代表中心点的值更加准确。但STING算法的聚类质量通常取决于网格的大小和密度阈值，还有待研究。该算法获得的代表中心点坐标如表1所示。

表1 类簇代表中心点坐标值

个数	中心点经纬度坐标
1	(114.6357，38.6267)
2	(114.7251，38.6346)
4	(115.0019，38.6281)
5	(115.1697，38.3508)
6	(115.2328，38.3914)
7	(115.3343，38.3828)
8	(115.4266，38.4553)
9	(115.4937，38.5373)

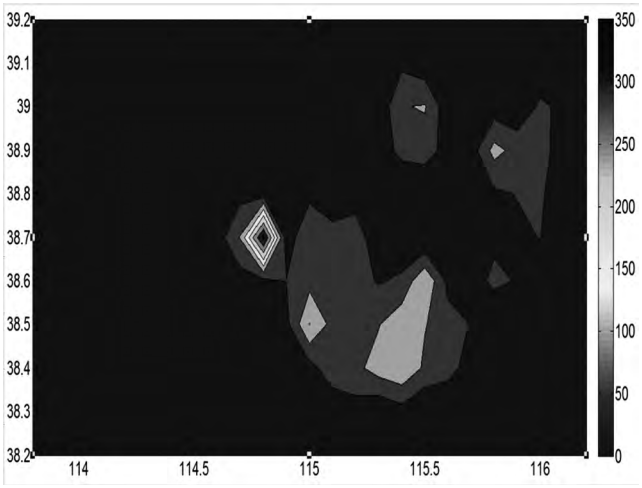

图8 放映热点图

为了更形象、直观的反映放映分布情况，方便管理者进行决策分析，本文还将清洗后的原始数据进行放映热点可视化分析，如图8所示。在图8中，不同的数值范围采用不同的颜色，数值越大，代表放映越密集，从侧面验证本文的实验结果。

4 结论

本文通过与模糊聚类FCM算法进行对比实验，验证了混合聚类方法对数字电影流动放映轨迹点分析的有效性。通过应用本模型，不仅解决了大数据量时兴趣点数据重叠遮盖的问题，又从宏观角度识别出了高密度区域，实现了高密度区域代表中心点的准确定位，为管理部门合理安排各区域放映员轨迹提供科学决策支持。同时，本文也有很多不足之处，比如本文算法获得的中心点还可以与基于欧式距离划分方法的中心点进行比较，以及网格参数选取方面，还需要进一步的讨论和分析。✣

参考文献

[1] Chunmei Zhang，Xinfeng Zhang，etal. An Efficient Clustering Algorithm for Trajectory Points in Digital Movies Mobile Playing Systems. 9th International Congress on Image and Signal Processing，BioMedical Engineering and Informatics (CISP－BMEI)，Datong，2016，pp. 1911－1915.

[2] 吴宏涛，黄昭婷．数据融合和数据挖掘技术在数字电影流动放映信息回传体系中的应用研究[J]．现代电影技术，2011(6)：27－33.

[3] 樊博．基于空间聚类挖掘的城市应急救援机构选址研究[J]．管理科学学报，2008，11(3)：16－28.

[4] Allstadt A，Caraco T，Korniss G. Ecological invasion：spatial clustering and the critical radius [J]. Evolutionary Ecology Research，2007，9(3)：375－394.

[5] Wan L H，Li Y J，Liu W Y，etal. Application and study of spatial cluster and customer partitioning [C] // International Conference on Machine Learning and Cybernetics. 2005：1701－1706 Vol. 3.

[6] 王培，孟志军，尹彦鑫等. 基于农机空间运行轨迹的作业状态自动识别试验[J]. 农业工程学报，2015，31(3)：56—61.

[7] 余莉，甘淑，袁希平，等．基于空间邻近的点目标聚类方法[J]．计算机应用，2016，36(5)：1267－1272.

[8] Ester M，Kriegel H P，Sander J，et al. A Density－Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise [C] // 2008：226－－231.

[9] B Borah and D. K Bhattacharyya，"DDSC：A density differentiated spatial clustering technique，" Journal of Computers，vol. 3，no. 2，pp. 72－79，2008.

[10] 王博．GIS系统中基于网格密度的空间聚类算法的研究与应用[D]．大连理工大学，2005.

[11] 周红芳，王鹏．DBSCAN算法中参数自适应确定方法的研究[J]．西安理工大学学报，2012，28(3)：289－292.

[12] MARTIN E. A density－based algorithm for discovering clusters in large spatial databases with noise [C]. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland，USA：AAAI Press，1996：226－231.