Hans Journal of Data Mining 数据挖掘, 2024, 14(1), 10-19
Published Online January 2024 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2024.141002

Hans汉斯

支持城市功能街区划分的有序语义聚类算法

张冉冉[1,2]，刘俊岭[1,2*]，孙焕良[1,2]，许景科[1,2]

[1]沈阳建筑大学计算机科学与工程学院，辽宁 沈阳
[2]辽宁省城市建设大数据管理与分析重点实验室，辽宁 沈阳

收稿日期：2023年12月2日；录用日期：2024年1月5日；发布日期：2024年1月12日

摘 要

城市中的功能大都分布在沿街道的两侧建筑，表现为线性街区，识别城市街区功能划分的特征可为城市空间结构及资源的全面规划、合理配置、统筹安排等提供帮助。传统线性语义聚类算法可用于划分单功能城市街道区，但城市街区不仅包括单一功能分区，还包括混合功能区。本文提出一种支持城市功能街区划分的有序语义聚类算法，在发现单一功能区的同时，也发现混合区并定义了一种新的度量混合功能区的方法。提出的算法基于层次聚类思想，具体算法分为两阶段，第一阶段为层次树生成，采用凝聚的方法将相邻的相似分段合并，得到层次树；第二阶段为功能区提取，进行单一功能区与混合功能区识别，获取给定街区的线性功能区。在真实数据集上的实验结果表明，所提出的算法可以有效发现混合功能区。

关键词

POI，有序聚类，功能区划分，混合功能区

An Ordered Semantic Clustering Algorithm Supporting Urban Block Knowledge Graph

Ranran Zhang[1,2], Junling Liu[1,2*], Huanliang Sun[1,2], Jingke Xu[1,2]

[1]School of Computer Science and Engineering, Shenyang Jianzhu University, Shenyang Liaoning
[2]Liaoning Province Big Data Management and Analysis Laboratory of Urban Construction, Shenyang Liaoning

Received: Dec. 2nd, 2023; accepted: Jan. 5th, 2024; published: Jan. 12th, 2024

Abstract

The functions in a city are mostly distributed along the buildings on both sides of the street, manifested as linear blocks. Identifying the characteristics of the functional division of urban blocks can provide assistance for the comprehensive planning, rational allocation, and overall arrangement

*通讯作者。

文章引用: 张冉冉, 刘俊岭, 孙焕良, 许景科. 支持城市功能街区划分的有序语义聚类算法[J]. 数据挖掘, 2024, 14(1): 10-19. DOI: 10.12677/hjdm.2024.141002

of urban spatial structure and resources. The traditional linear semantic clustering algorithm can be used to divide the single function urban street area, but the city block includes not only the single function area, but also the mixed function areas. This article proposes an ordered semantic clustering algorithm that supports the division of urban functional blocks. While discovering a single functional area, it also discovers mixed areas and defines a new method for measuring mixed functional areas. The proposed algorithm is based on the idea of hierarchical clustering, which is divided into two stages. The first stage is the generation of a hierarchical tree, which uses the aggregation method to merge adjacent similar segments to obtain a hierarchical tree. The second stage involves extracting functional areas, identifying single and mixed functional areas, and obtaining linear functional areas for a given block. The experimental results on real datasets show that the proposed algorithm can effectively discover mixed functional areas.

Keywords

POI, Ordered Clustering, Functional Area Division, Mixed Functional Area

1. 引言

城市中的各种功能建筑通常沿街道两侧线性分布，呈现一种城市资源集约高效的现象，可以代表性地展示城市街区功能结构分布。西班牙工程师索里亚·玛塔于 1882 年首次提出了基于线型城市的城市区域规划概念。在识别城市功能区时选择针对线性街道进行研究，对街区功能分布的划分，有利于改善兴趣点推荐、设施选址以及路线推荐等应用。

随着数据采集能力的增强，特别是在信息化背景下，诸多学者在如何充分挖掘大数据所隐含的城市功能区信息展开了一系列探索与尝试。例如，邬群勇等借助出租车轨迹数据分析每个地块交通起止点数量的时空分布特征将厦门岛划分为工作区、居住区和休闲娱乐区[1]；赵莹等基于手机漫游用户数将张家界划分为日常居住区、重点旅游区和其他波动区[2]；王俊珏等基于核密度与兴趣点大数据开展了城市功能分区研究[3]。

兴趣点(point of interest, POI)数据包含名称、类型、经度和纬度四方面信息，具有体量大、精确度高和时效性强等特点。POI 数据记录社会经济各行业部门的空间位置信息，能够为精细识别城市功能区类型提供可能，从而为更好地认识城市空间结构、指导区域空间优化调控提供科学支撑。目前，一些学者基于 POI 数据进行了城市功能区识别研究，例如：Jiang 等根据计算每个单元内 POI 出现频率，配对算法识别城市功能区[4]；Wang 等通过核密度聚类实现功能区识别[5]；康雨豪等使用密度分数进行武汉市功能区识别[6]。上述研究中功能单元划分粒度较大，通常以区域为研究对象进行功能划分。而街道单元本身是认知城市的基本单元，因此本文选择包含街道、建筑和兴趣点数据的街区单元作为基本空间单元进行城市功能区识别研究，可以更准确地识别人们的活动，把握城市的功能结构分布。

观察发现，城市街区功能分布常常存在多种功能的混合，即混合功能区。现有的功能识别算法无法直接发现这些混合功能区，聚类算法结果簇标签通常为单一类型。识别混合功能区存在的挑战为当街区内出现两种及以上功能时，如何判断该区域功能混合，如何提取该街区功能区。基于以上观察，本文定义了城市街区混合功能区，提出了一种基于语义的街道功能有序聚类划分算法。以开放街道地图的路网

数据和 POI 数据为主要数据源，基于层次聚类思想提出了针对线性街区的语义聚类算法。在判断该区域功能是否混合时，提出了利用度量街区的功能分布的纯度，并结合语义强度共同定义混合功能区，在提取该街区功能区时，对语义化层次树通过层次遍历进行提取。

本文的主要贡献如下：

(1) 提出了针对城市街区的单一功能区和混合功能区定义，对城市功能区的划分更加精确合理。

(2) 提出了支持城市功能街区划分的有序语义聚类算法，可以实现对单一功能区和混合功能的识别。

(3) 利用真实数据进行实验，验证了算法的有效性。

2. 相关工作

本文相关工作涉及城市功能区划分以及聚类算法两个方面。

2.1. 城市功能区划分

根据采用的数据不同，城市功能区划分方法可分为基于静态数据的功能区划分、基于动态数据的功能区划分。

早期的基于静态数据的功能区划分主要是基于遥感影像、土地利用、面板数据等，根据土地的自然属性对城市功能进行分类。另一类静态数据为 POI 数据，通过 POI 数据反映各种设施的多样性和混合程度。Zhai 等通过构建 Place2vec 模型来获取 POI 地理信息，改进了功能区识别框架[7]；Ran 等基于 POI 数据对长沙市生活服务业的空间格局进行了深入分析[8]。随着数据获取途径的增加以及 GPS 的普及，利用用户轨迹数据进行功能区识别与划分突破了传统的功能分区思路，Song 等根据公园地理位置的照片数量和视觉内容(来自 Instagram 和 Flickr 平台)探索了公园的使用情况[9]。然而，这些方法受到用户数量和用户足迹定位精度的限制，所以分类结果的准确性较差。

将 POI 数据与动态轨迹数据结合的现有研究中，都选择采取矢量栅格化划分城市地块，将城市不同等级的路网分割为互不相同的空间单元。冯慧芳等利用出租车 GPS 轨迹和 POI 数据，利用城市栅格方法构建栅格关联规则矩阵，识别兰州市城市功能区[10]。陈泽东等利用出租车 GPS 数据提取地块的居民出行时序特征，采用期望最大化算法，进行北京城市区识别与空间交互研究[11]；由于路网数据本身原因，对城市内部功能区范围界定尺度较大，且对于功能区划分研究以定性为主，缺乏对功能区混合现象的研究。

现有城市功能区划分以区域为单位，而本文研究对象为线性分布的街区，采用的划分方法应考虑对象的顺序关系。另外，现有研究中大多仅将城市划分为单一功能区，如商业功能区、居住功能区、工业功能区等，缺乏对城市功能区混合现象(如商业与居住功能区混合)的研究。对此问题，本文研究单一功能区与混合功能区发现方法。

2.2. 聚类算法

聚类是按照特定标准把数据集分割成不同的类或簇，使得同一个簇内的数据对象的相似性尽可能大，同时不同簇中的数据对象高度相异。

本文数据为有序的数据，针对有序数据的聚类算法，Fisher 最优分割法采用的最优二分割法只能求得局部最优，不适合由于样本长度较大时的情况。高苏等通过 K-modes 聚类方法构建的有序聚类方法得到海员职业幸福感指数的等级划分及其相应的语义描述[12]。姚尧等通过时序出租车出行数据和 POI 数据描述居民出行模式，结合动态时间规整和 K-MEDOIDS 聚类算法识别城市的功能属性和空间结构[13]。苏月同等通过对站点有序客流数据聚类，提出了一种基于有序样本聚类的站点级差异化高峰时段识别方

法，识别出城市轨道交通站点高峰时段[14]。本文数据为自然顺序下的街道单元 POI 数据，现有的有序数据聚类，无法直接用于本文。

基于本文数据特点，提出构建一种支持城市功能街区划分的有序语义聚类算法，在不改变数据自然顺序前提下，对其进行聚类分析，完成城市街道单元功能区的划分，功能区包含单一功能区、混合功能区。

3. 问题定义

3.1. 定义 1 (城市街区)

城市街区 S 为城市中一段道路及其两侧所属的空间对象集，空间对象集 P 由有序的对象组成表示为 $(p_1, p_2, p_3, \ldots, p_n)$，其中 p_i 的序号为 i。每个对象 p_i 表示为$(ID, loc, c, info)$，其中 ID 为对象的序号，loc 为街区 S 上的相对位置，c 为对象类别，$info$ 为对象描述信息。

本文将对象类别定义为 6 个，分别为居住、商业、行政、公共服务、医疗和教育。为了对象简化表示，将道路两侧的对象 p_i 投影到街道上，利用映射后街道内包含的对象 p_i 作为研究对象。为了方便表示，将居住、商业、行政、公共服务、医疗和教育分别用字母 *abcdef* 代替。例如某街区 POI 经过简化后可表示为 *cecea*，表示 POI 点分别是行政－医疗－行政－医疗－居住等。

3.2. 定义 2 (城市街区单位区段)

给定城市街区 S 以及街区内有序对象集 P，按单元长度对街区划分生成多个单位区段，表示为$(s_1, s_2, s_3, \ldots, s_m)$。对于区段 s_i 根据区段中的对象分布可以得到区段特征向量 f_i，f_i 的每一个维 v_j 对应一个区域功能，其值可由相应空间对象数目获得。

例如某条街区的 POI 表示为 *ceaceabbcc ccccccccbb aeacaaabac*，生成 3 个分段 s_1、s_2、s_3 分别为 |*ceaceabbcc*|、|*ccccccccbb*|、|*aeacaaabac*|，区段特征向量 s 则会产生 3 个向量 f_1 (0.2, 0.2, 0.4, 0, 0.2, 0), f_2 (0, 0.2, 0.8, 0, 0, 0), f_3 (0.6, 0.1, 0.2, 0, 0.1, 0)，向量的各维数值为相应功能对象的所占比例。

给定城市划分区段的特征向量，通过特征向量的各维数值可以确定街区功能，选择维数值大的功能作为街区功能。本文定义了两种功能区，一种为单一功能区，另一种为混合功能区。

3.3. 定义 3 (单一功能区)

给定城市划分区段 S 的特征向量 f，当 f 的最大取值维 $f \cdot v_j \geq \theta$ 时，则称该区域为单一功能区，功能区标签为该维的功能类型，其中 θ 为功能确定阈值。

本文提出了混合功能区的概念，混合功能指街区的功能表现为两个以上功能特征，并且这些功能分布均匀。为了评价一个线性城市街区 S 的功能分布均匀性，本文提出了线性层次基尼指数度量 Gini(S)，如定义 4 所示。

3.4. 定义 4 (线性层次基尼指数)

给定城市划分区段 S 及其单位区段，设单位区段组为最细粒度层表示为 S_0，将 S_0 相邻区段两个合并生成 S_1 层，依次合并直到合成为一个区段，为顶层表示为 S_q，则区段 S 的线性层次基尼指数为各层基尼指数之和，表示为 $\mathrm{Gini}(s) = \sum_{i=1}^{q} G_i(v)$，其中 $G_i(v) = 1 - \sum_{v_j \in v} v_j^2$ 。

线性层次基尼指数是为了衡量区段 S 内功能类型分布的不纯度，基尼指数越大，不纯度越高，功能的混合程度越高。设计线性层次基尼指数为了度量街区的总体的功能分布，采用某一划分粒度的基尼指数无法度量。例如区段 S 共包含 6 个单位区段分别为|*aa*|*aa*|*bb*|*bb*|*aa*|*bb*|，线性层次基尼指数 $\mathrm{Gini}(S) = 1.5$，区段 S' 共包含 6 个单位区段分别为|*ab*|*ab*|*ab*|*ab*|*ab*|*ab*|，线性层次基尼指数 $\mathrm{Gini}(S') = 5.5$ 。

3.5. 定义 5 (混合功能区)

给定城市划分区段 S 的特征向量 f，存在一个 f 的维的子集 V，满足 $\sum_{v_i \in V} f \cdot v_i \geq \theta$ 且任意两个维度 $f \cdot v_i$ 与 $f \cdot v_j$ 满足 $\left| f \cdot v_i - f \cdot v_j \right| \leq \alpha$，则 S 为候选混合功能区。当 S 的单位区段功能分布均匀性满足 $\text{Gini}(S) \geq \beta$ 则称 S 为混合功能区，标签为 V 相应功能。其中 Gini(S)为 S 的线性层次基尼指数，β 为功能分布均匀性阈值。

混合功能区由语义强度以及混合分布性确定。语义强度需满足特征向量内任意两个及以上的各维数值相加大于等于阈值 θ 时，任意两维数值差小于等于 α。例如特征向量 f(0.1, 0.2, 0.4, 0, 0.3, 0)、$\theta = 0.6$、$\alpha = 0.1$ 时，$0.4 + 0.3 \geq 0.6$，$\left| 0.4 - 0.3 \leq 0.1 \right|$，则此街区为候选混合功能区，其功能特征为子集 V(0.4, 0.3)。

混合分布性由 V 内线性层次基尼指数 Gini(V)确定，当 $\text{Gini}(V) \geq \beta$ 时说明子集 V 内的混合分布性越高，子集 $V(0.4, 0.3)$ 由两段(3, 2) (2, 2)组成，$\beta = 0.4$ 时，$\text{Gini}(V) \geq 0.4$，则称区段为由公共服务和教育混合的混合功能区。

将相邻区段进行合并时，簇间相似度采用余弦相似度计算，当两个簇 R_i、R_{i+1} 合并时，需要生成一个新簇，新簇的特征向量采用加权平均来计算。具体的计算方法为式 1：

$$f_{i,i+1} = \frac{l_i}{l_i + l_{i+1}} f_i + \frac{l_i}{l_i + l_{i+1}} f_{i+1} \tag{1}$$

例如，给定两个簇长度 l_1、l_2 分别为 2、4，向量分为别(0, 0.6, 0.4, 0, 0, 0)、(0, 0.2, 0.6, 0, 0.2, 0)，则新簇的特征向量为(0, 0.333, 0.534, 0, 0.133, 0)。

4. 基于语义的有序聚类算法

本节介绍基于语义的有序聚类算法，用于将城市街区按功能特征划分，并生成功能区。算法包括数据序处理、语义化层次树生成和功能区提取 3 个阶段。首先，在数据序处理阶段将城市街区进行单位划分，根据单位区段内的对象分布得到特征向量线性序列。在语义化层次树生成阶段采用凝聚层次聚类的方法将相邻的相似分段合并，得到层次树，并将层次树结点进行语义化，得到语义化层次树。在功能区提取阶段，进行单一功能区与混合功能区识别，获取给定街区的线性功能区。

4.1. 数据预处理

给定待处理的街区 S，需要将街区周围的一定距离 POI 对象投影到 S，生成线性功能类别有序序列。然后，将线性有序序列划分为街区单元区段，提取出街区单元区段的特征向量，生成特征向量线性序列。

本文采用 2 种方法划分街区单位区段，第 1 种是将 POI 点按距离均等划分，给定一组 POI 功能数据，将其按照等长进行划分；第 2 种是将功能相同的类别合并后，再采用近似等距离划分。

例如：$str\{aaabbccccaaaccc\}$，当采用第 1 种将 POI 点按距离均等划分时按每 5 个 POI 划分，划分结果为$\{aaabb\}$，$\{cccca\}$，$\{aaccc\}$。采用第二种划分，相同字符合并处理后可划分位置数组 $merge = \{3,5,9,12,15\}$。第一次位置划分位置数值取 5，$5 \in merge$，则 $X = \{5\}$，第二次位置划分位置数值取 10，$10 \notin merge$，向后平移取 12，则 $X = \{5,12\}$，$15 \in merge$，则 $X = \{5,12,15\}$，输出划分结果为$\{aaabb\}$，$\{ccccaaa\}$，$\{ccc\}$。

在提取出线性有序序列后，根据本文研究的六个功能类型进行类别重新分类，形成包含功能标注的线性类别序列，对线性类别序列按照划分位置数组 X 进行划分，得到单位区段 s_1、s_2、s_3，...，s_m，根据定义 2 生成每段街区分段的特征向量($f_1, f_2, \ldots, f_m$)。

4.2. 语义化层次树生成

在聚类生成语义化层次树阶段，将提取的街区分段的特征向量作为结点，通过余弦相似度计算相邻

结点的相似度，选择相似度最高的两个结点进行合并，并根据公式 1 生成新的特征向量，以此反复生成层次树。将层次树结点进行语义化，结点的语义化分析指根据结点区段的特征向量将该结点定义为单一功能区或混合功能区，最终得到语义化层次树。

算法 1 给出了层次树生成过程，初始化层次树 T 为空(第 1 行)。当区段集中多于 1 个区段时，进行结点合并生成层次树(第 3~12 行)。首先将最大相似度 *simMax* 初始化为 0，最大相似度的相邻区段标号初始化为 0 (第 3~4 行)，计算相邻区段特征向量的余弦相似度，查找一对相似度最大的相邻区段 f_i，f_{i+1}，将相似度赋值给 *simMax* (第 5~7 行)。i 赋值给 m，将找出的相邻结点及合并后的结点插入 T 中，并将区段集 S 中删除 s_m，s_{m+1}，插入合并区段 $s_{m,\,m+1}$ (第 8~12 行)。最后返回层次树 T (第 13 行)。

算法 1. CreateTree(S, F)

Input：区段集 $S(s_1,s_2,\cdots,s_n)$，各区段的特征向 $F(f_1,f_2,\cdots,f_n)$
Output：层次树 T
(1) T=Φ
(2) **while**(|S|>1)
(3) simMax=0;
(4) m=0;
(5) **for**(i=0; i<|S|; i++)
(6) **if** (sim(f_i, f_{i+1})>simMax) **then**
(7) simMax=sim(f_i, f_{i+1});
(8) m=i;
(9) 将 s_m, s_{m+1} 将插入 T;
(10) 合并 s_m, s_{m+1} 得到 $s_{m,m+1}$ 并插入 T;
(11) S=S+{$s_{m,m+1}$};
(12) S=S-{$s_{m,\,sm+1}$};
(13) **return** T;

算法 1 由两层循环构成，区段集中区段个数为 n，特征向量个数为 n，则算法的时间复杂度为 O(n^2)。

算法 2 给出了结点语义化过程。层次树 T 中每一个结点进行语义化分析(第 1~5 行)，当某一结点的特征向量 f_m 中某一维大于定义 3 定义的阈值 θ 时，则确定该区段为单一功能区，功能区标签为该维的功能类型(第 2~3 行)。当某一结点的特征向量 f_m 中任意两个及以上的各维数值相加大于等于定义 3 定义的阈值 θ，且任意两个维度差满足定义 5 定义的阈值 α，该结点的线性层次基尼指数大于等于定义 5 定义的阈值 β，则确定区段 S_m 为混合功能区，功能区标签为 i，j 代表的功能类型(第 4~5 行)，返回节点语义化的层次树 T' (第 6 行)。

算法 2. Semanticfunction(T,F,θ,α,β)

Input：层次树 T，阈值 θ,α,β 各区段的特征向量 $F(f_1, f_2, \cdots, f_n)$
Output：语义化的层次树 T'
(1) **for** (m=0; m≤n; m++) **do**
(2) **if** ($f_m.v_j \geq \theta$) **then**
(3) 结点 S_m 以 j 为代表功能的单一功能区;
(4) **else** $|f_m.v_i - f_m.v_j| \leq \alpha$, Gini($S_m$)$\geq\beta$ **then**
(5) 结点 S_m 为 i 和 j 为代表的混合功能区;
(6) **return** T';

算法 2 中时将整个层次树 T 结点访问并语义化，所以其时间复杂度为 O(n)。

4.3. 功能区提取

功能区的提取需要对语义化层次树提取结点。提取层次树结点是指根据功能区个数 k 从上往下提取层次树的结点，提取过程中若该结点为单一功能区，则可以分解该结点的左右孩子结点，若该结点为混合功能区，则将该结点及其包括的所有孩子结点视为整体，提取时不再进行分解，提取至结点个数等于 k。

算法 3 给出了层次树结点提取过程。算法采用层次遍历的方法进行结点提取，初始化一个队列 Q，先将根节点加入队列(第 1~2 行)。当队列不为空时进入循环，如果队列中的结点语义化为单一功能区，将其左右孩子结点加入队列并删除该结点(第 3~6 行)。如果队列中的结点语义化为混合功能区，则访问该结点，将其加入最终遍历序列中(第 7~8 行)。当队列中元素个数等于所提取聚类个数时，停止循环，输出队列中结点(第 9~10 行)。

算法 3. C=TreeExtractPartition(T',Q,k)

Input：语义化层次树 T'，聚类数 k
Output：结点队列 Q
(1) initqueue(Q);
(2) enqueue(Q,T');
(3) **while**(!=IsEmpty(Q))
(4) **if**(S_m 为单一功能区)**then**
(5) dequeue(Q,S_m);
(6) enqueue(Q,S_m->rchild,S_m->lchild);
(7) **else** (S_m 为混合功能区)**then**
(8) visit(S_m);
(9) len|Q|=k;
(10) **return** Q;

5. 实验分析

对于本节采用真实数据集对所提出的算法模型进行实验验证，对比分析各方法的性能。

5.1. 实验数据集与实验参数

本文选取的线性街道范围为复兴路至建国路，全长 20 千米。POI 数据来自百度地图，北京市的路网数据来自开放地理数据平台(http://www.locaspace.cn/)。本文将 POI 数据与路网数据进行匹配后，只选取街道两侧各 300 米内的 POI 数据投影到街道上，形成的沿街道路两侧的线性有序序列，生成共 1102 个 POI 点，并根据《城市用地分类与规划建设用地标准》结合北京市实际情况，将 POI 数据分成居住、商业、行政、公共服务、医疗和教育 6 大类，其中各类 POI 数为居住 54 个，商业 694 个，行政 298 个，公共服务 32 个，医疗 10 个，教育 14 个。

本文采用轮廓系数(Improved Silhouette Coefficient, ISC)来度量所提出来的聚类方法的效果。由于所聚类数据对象为有序 POI，将轮廓系数进行了修改。具体计算公式如式 2 所示。

$$S(i)=\frac{1}{n}\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{2}$$

其中，$a(i)$为 i 向量到所有它属于的簇中其它点的平均距离。由于所聚类数据对象为有序 POI，所以簇间距离只包括相邻簇之间的距离，则 $b(i)$为 i 向量到相邻簇内的所有点的平均距离。n 为簇内向量个数。

轮廓系数取值范围为[−1, 1]，越趋近于 1 代表内聚度和分离度都相对较优，聚类效果越好。

在判断功能区类型时，通过设置参数的不同的取值，对比评价不同聚类方法的性能，参数取值如表 1 所示。

Table 1. Parameter value range
表 1. 参数取值范围

参数	取值范围
划分距离 g (单位：米)	200, 300, 400, 500
功能确定阈值 θ	0.4, 0.5, 0.6, 0.7
特征向量维度阈值 α	0.1
功能分布均匀性阈值 β	0.4, 0.5, 0.6, 0.7
簇数 k	4, 5, 6, 7, 8, 9

5.2. 实验评价

利用现有的聚类方法与本文提出的聚类方法进行比较，具体方法表示如下，其中方法(1)~(4)为比较方法，(5)，(6)为本文提出来的聚类方法。

(1) 等分划分层次聚类法(Equidistant Method, EM)：等分城市街区，层次聚类生成层次树后，取 k 簇。

(2) 等分去孤立点法(Equidistant remove isolated points method, ER)：等分城市街区，层次聚类生成层次树，去除孤立点后，取 k 簇。

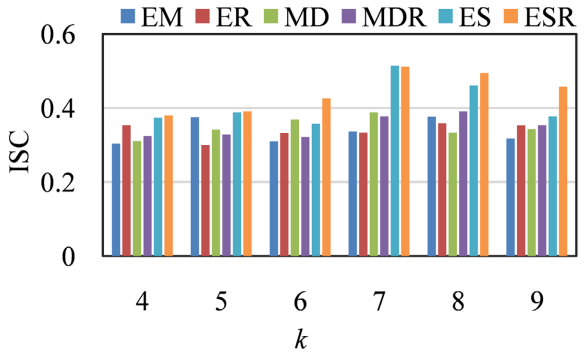
(3) 合并等分法(Merge divide method, MD)：将功能相同类别合并后，再近似等分街区，层次聚类生成层次树，取 k 簇。

(4) 合并等分去孤立点法(Merge divide remove isolated points method, MDR)：将功能相同类别合并后，再近似等分街区，层次聚类生成层次树，去除孤立点，取 k 簇。

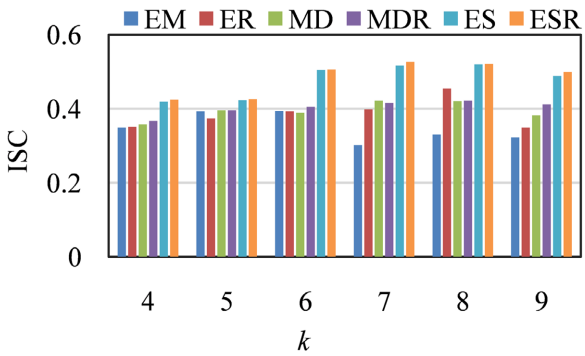
(5) 等分语义法(Equal division semantic method, ES)：等分城市街区，对层次聚类生成的层次树，进行结点语义分析生成混合功能区后取 k 簇。

(6) 等分语义去孤立点法(Equidistant semantic remove isolated points method, ESR)：等分城市街区，对层次聚类生成的层次树，进行结点语义分析生成混合功能区，并且去除孤立点后取 k 簇。

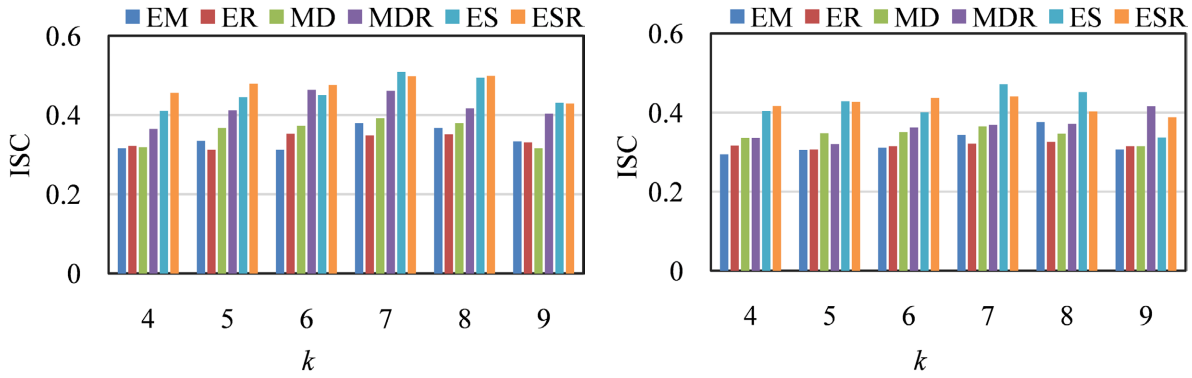
本文通过设置不同的参数，对比分析六种方法聚类效果。第 1 个实验是当 k 变化时，各算法的效果。实验结果如图 1(a)~(d)所示。其中 $\theta = 0.5$，$\alpha = 0.1$，$\beta = 0.5$ 时，g 分别取 200 m，300 m，400 m，500 m。



(a) g = 200 m



(b) g = 300 m



(c) $g = 400$ m (d) $g = 500$ m

Figure 1. Improved silhouette coefficient map for dividing distances
图 1. 划分距离的轮廓系数图

由实验 1 结果可知，ESR 优于其他方法，且 $g = 300$ m，$k = 7$ 时，6 种方法的聚类效果均为最优。

第 2 个实验是改变β的取值，评估 6 种聚类方法的效果，实验结果如图 2 所示，其中 $g = 300$ m，$k = 7$，$\theta = 0.5$，$\alpha = 0.1$。由实验 2 结果可知，ESR 优于其他方法，且当 $\beta = 0.5$ 时，6 种方法的聚类效果均为最优。

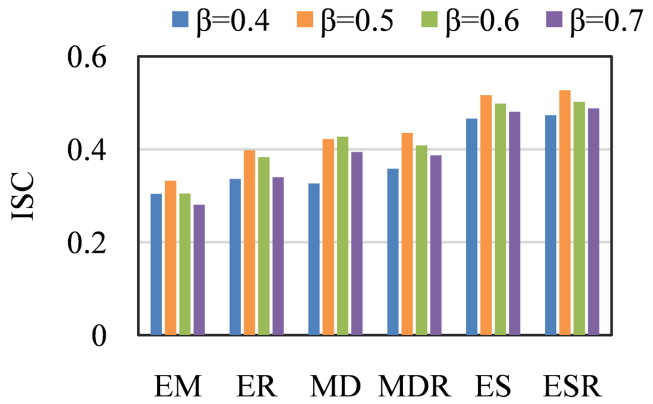


Figure 2. Improved silhouette coefficient plots with different β value
图 2. 不同 β 取值的轮廓系数图

功能区的提取需要对语义化层次树提取结点。提取层次树结点是指根据功能区个数 k 从上往下提取层次树的结点，第 3 个实验是改变 θ 的取值，评估 6 种聚类方法的效果，实验结果如图 3 所示。其中，$g = 300$ m，$k = 7$，$\beta = 0.5$，$\alpha = 0.1$。由实验 3 结果可知，ESR 优于其他方法，且当 $\theta = 0.5$ 时，6 种方法的聚类效果均为最优。

功能区的提取需要对语义化层次树提取结点。提取层次树结点是指根据功能区个数 k 从上往下提取层次树的结点，由以上实验结果观察可知，6 种方法中同样操作方法时去除孤立点方法优于未去除孤立点方法。方法 ES 和 ESR 优于其他 4 种方法，原因是对于混合功能区的度量，使得更精确的计算区段内各类 POI 分布纯度，以及整个区段的功能语义强度。对于街道的划分距离 g，当 $g = 200$ m 时六种方法均为最优，当划分距离过近或过远，单位区段内 POI 个数较少或较多，都无法精确表示该区段的功能特征。对于功能确定阈值 θ，即为单一功能区确定阈值，当该值越小，单一功能区的纯度越低。功能分布均匀性阈值 β 的取值参考基尼指数，当该值越大，不纯度越高，功能的混合程度越高。

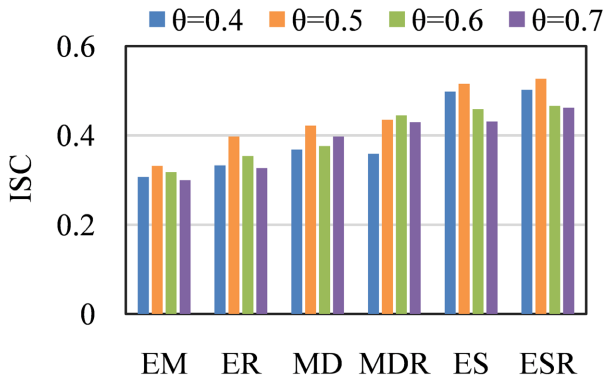


Figure 3. Improved silhouette coefficient plots with different θ value
图 3. 不同 θ 取值的轮廓系数图

6. 结论

本文研究了城市功能街区划分问题，提出了一种支持城市功能街区划分的有序语义聚类算法。该算法重新定义了混合功能区的度量，根据其语义强度以及混合分布性确定混合程度，并根据语义分析结点，进行簇的提取。为了评估所提出算法的性能，将其与基本方法进行比较，并利用真实数据集进行了实验验证。

参考文献

[1] 邬群勇, 吴祖飞, 张良盼. 出租车 GPS 轨迹集聚和精细化路网提取[J]. 测绘学报, 2019, 48(4): 10. https://doi.org/10.11947/j.AGCS.2019.20180256

[2] 赵莹, 张朝枝, 金钰涵. 基于手机数据可靠性分析的旅游城市功能空间识别研究[J]. 人文地理, 2018, 33(3): 8.

[3] 王俊珏, 叶亚琴, 方芳. 基于核密度与融合数据的城市功能分区研究[J]. 地理与地理信息科学, 2019, 35(3): 7.

[4] Jiang, S., Alves, A., Rodrigues, F., *et al.* (2015) Mining Point-of-Interest Data from Social Networks for Urban Land Use Classification and Disaggregation. *Computers Environment & Urban Systems*, **53**, 36-46. https://doi.org/10.1016/j.compenvurbsys.2014.12.001

[5] Wang, Z., Ma, D., Sun, D., *et al.* (2021) Identification and Analysis of Urban Functional Area in Hangzhou Based on OSM and POI Data. *PLOS ONE*, **16**, e0251988. https://doi.org/10.1371/journal.pone.0251988

[6] 康雨豪, 王玥瑶, 夏竹君, 等. 利用 POI 数据的武汉城市功能区划分与识别[J]. 测绘地理信息, 2018, 43(1): 5.

[7] Zhai, W., Bai, X., Shi, Y., *et al.* (2019) Beyond Word2vec: An Approach for Urban Functional Region Extraction and Identification by Combining Place2vec and POIs. *Computers Environment and Urban Systems*, **74**, 1-12. https://doi.org/10.1016/j.compenvurbsys.2018.11.008

[8] Ran, Z., Zhou, G., Jiamin, W.U., *et al.* (2019) Study on Spatial Pattern of Consumer Service Industry in Changsha Based on POI Data. *World Regional Studies*.

[9] Song, X.P., Richards, D.R., He, P., *et al.* (2020) Does Geo-Located Social Media Reflect the Visit Frequency of Urban Parks? A City-Wide Analysis Using the Count and Content of Photographs. *Landscape and Urban Planning*, **203**, 103908. https://doi.org/10.1016/j.landurbplan.2020.103908

[10] 冯慧芳, 杨文亮. 融合 GPS 轨迹和 POI 数据关联规则的城市功能区识别[J]. 测绘科学技术学报, 2020, 37(4): 7.

[11] 陈泽东, 谯博文, 张晶. 基于居民出行特征的北京城市功能区识别与空间交互研究[J]. 地球信息科学学报, 2018, 20(3): 11.

[12] 高苏, 鲍君忠, 王昕, 等. 可解释性有序聚类方法及其应用分析[J]. 计算机应用, 2022, 42(2): 6.

[13] 姚尧, 张亚涛, 关庆锋, 等. 使用时序出租车轨迹识别多层次城市功能结构[J]. 武汉大学学报: 信息科学版, 2019, 44(6): 10.

[14] 苏月同, 徐天捷, 蒲一超, 等. 基于有序样本聚类的城市轨道交通站点差异化高峰时段识别方法[J]. 交通运输工程与信息学报, 2023, 21(2): 123-140.