面向城市热点区域的不平衡数据聚类挖掘研究

蔡 莉[1,2] 李英姿[2] 江 芳[2] 梁 宇[2]
(复旦大学计算机科学技术学院 上海 200433)[1] (云南大学软件学院 昆明 650091)[2]

摘 要 在大数据时代,数据来源众多,因此多源数据的融合成为数据挖掘领域的一个研究热点。现有的多源数据融合研究主要聚焦于相同领域内平衡数据集的融合模型和算法,对来自不同领域的非平衡数据集的聚类挖掘关注较少。DBSCAN(Density-Based Spatial Clustering of Applications with Noise)算法是挖掘热点区域的主要算法,但其无法处理不平衡的融合数据,少数类数据形成的聚类结果很难被发现。针对不平衡数据的融合,文中提出了一种基于时空特征的位置数据融合模型;同时,从数据层面和算法层面提出新颖的方法来解决不平衡数据的挖掘问题。鉴于目前的聚类算法的评价指标并不适用于不平衡数据的聚类结果评估,提出了一种新的综合评价指标来反映聚类质量。将来自交通领域的GPS轨迹数据(多数类数据)和社交领域的微博签到数据(少数类数据)进行融合,然后采用所提方法来挖掘热点区域。实验结果表明:基于多源数据融合的热点区域挖掘结果优于单源挖掘结果,所发现的热点区域位置、分布和数量与实际情况一致。文中所提出的融合模型、改进算法和评估指标法是有效且可行的,还可用于其他来源的位置数据的融合与分析。

关键词 不平衡数据,数据融合,城市热点区域,聚类评价标准,位置数据
中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.08.003

Study on Clustering Mining of Imbalanced Data Fusion Towards Urban Hotspots

CAI Li[1,2] LI Ying-zi[2] JIANG Fang[2] LIANG Yu[2]
(School of Computer Science, Fudan University, Shanghai 200433, China)[1]
(School of Software, Yunnan University, Kunming 650091, China)[2]

Abstract In the era of big data, multi-source data fusion is a trending topic in the field of data mining. Previous studies have mostly focused on fusion models and algorithms of balanced data sets, but seldom on issues of clustering mining for imbalanced data sets. DBSCAN algorithm is a classical algorithm for mining urban hotspots. However, it cannot deal with imbalanced location data, and the clustering results generated by the minority class are difficult to discovery. Aiming at the imbalanced data fusion, this paper proposed a novel fusion model based on spatio-temporal features, at the same time, proposed a novel approach to solve the mining problem of imbalance data from data aspect and algorithm aspect. Since the evaluation index of current clustering algorithm is not suitable for the evaluation of unbalanced data clustering results, a new comprehensive evaluation index was proposed to reflect the clustering quality. GPS trajectory data (the majority class data) from the traffic field and microblog check-in data (the minority class data) from the social field are fused, and then the proposed method is used to mine hot spots. The mining results of hot spots based on multi-source data fusion are better than those of single source data fusion. The location, distribution and number of hot spots are consistent with the actual situation. The proposed fusion model algorithm and evaluation index method are effective and feasible, and can also be used for the fusion and analysis of location data from other sources.

Keywords Imbalanced data, Data fusion, Urban hotspots, Clustering criteria, Location data

1 引言

城市热点区域是指人流量大、交通需求旺盛、公共配套设施较完善或商业较发达的区域,是居民出行聚集程度的体现。城市热点区域的研究对于城市公共基础设施的规划、交通疏导和管理、土地价值评估、公共安全、个人信息保护等具有重要的现实意义和应用价值[1-2]。传统城市热点区域的研究一般采用单源数据,例如,出租车轨迹数据是一种包含GPS经纬度坐标的数据,它不仅记录着乘客的日常出行信息,而且能够反映城市的交通状况。但是,出租车一般不能驶入某些特

到稿日期:2018-11-27 返修日期:2019-01-17 本文受国家自然科学基金(61663047)资助。
蔡 莉(1975-),女,博士生,副教授,主要研究方向为数据挖掘、智能交通;李英姿(1994-),女,硕士,主要研究方向为数据挖掘、数据质量;江 芳(1993-),女,硕士生,主要研究方向为数据挖掘、数据质量;梁 宇(1964-),男,硕士,教授,主要研究方向为智能交通、云计算,E-mail:yuliang@ynu.edu.cn(通信作者)。

定区域,如景区、高校、步行街等,乘客只能在这些区域的周边上下车。因此,单独使用某一领域的数据进行热点区域挖掘,会造成研究结果的片面性。

大数据时代的来临使得数据来源日益多样化。反映居民出行信息的位置数据源除了 GPS 轨迹数据之外,还有微博签到数据、公交卡数据、地铁数据以及手机定位数据等[3]。微博签到数据是一种特殊的微博数据,用户利用带有 GPS 功能的智能终端记录某一时刻所处的位置并写下当时的感言,从而产生带有时空信息和文本内容的数据。签到数据记录了用户的兴趣爱好,反映了人们的生活轨迹,具有很高的研究价值,近年来也成为研究城市热点区域的一种重要来源[4]。出租车 GPS 轨迹数据的采样频率较高,大约每 15～60 s 会记录一个采样点,而且一个城市的出租车数量从几千辆到上万辆不等,每天累计的数据量约有 1000 多万条记录。相比之下,微博签到数据每天的数量只有几千条记录。尽管微博用户的数量庞大,但使用签到功能的用户较少,将这两种数据集融合后就会出现数据不平衡的现象。少数类的数据特征容易被多数类的数据特征所覆盖,而前者往往又蕴藏着极具价值的特征信息。

2　相关工作

近年来,国内外在城市热点挖掘研究方面取得了一系列成果。Lee 等[5]利用 K 均值聚类算法挖掘出租车的轨迹数据并从中发现热点区域,以为司机提供位置推荐。Kisilevich 等[6]提出改进的 DBSCAN 算法,并将其应用于华盛顿热点区域的分析中。Verma 等[7]通过空间聚类技术分析出租车的轨迹数据,从而预测乘客的需求热点区域。宁鹏飞等[8]获取了深圳市的新浪微博签到数据,在对 POI 点数据进行空间多密度聚类的基础上,结合用户签到频率,实现了城市空间的功能区识别。

在不平衡数据研究方面,Orriols-Puig 等[9]研究类的不平衡对不同 LCS(Michigan-style Learning Classifier Systems)分量的影响,然后采用一种与 LCS 最相关的学习分类系统来解决数据的高度不平衡分类问题。Krawczyk 等[10]提出了一种局部集成学习的方案来应对高维和多类不平衡数据。Sebastián 等[11]利用缩放因子技术对特征集的基数进行策略优化,并结合成本敏感型 SVM 来处理高维和不平衡数据的分类学习问题。翟云等[12]综述了近年来国内外对不平衡类数据挖掘的主要研究进展,并从数据层面和算法层面分别对目前存在的各种处理不平衡类数据挖掘的技术方法进行了深入剖析和全面比较。Zhu 等[13]在现有的不平衡分类学习模型的基础上提出了一个采用伪逆线性判别法的边界消除模型,然后根据测试样本和训练集间的启发式测量做出最终决策。

综上所述,现有城市热点挖掘研究所使用的数据集大多为单源数据,较少关注多源数据融合下的热点区域挖掘问题;而且,不平衡问题的研究主要集中在分类过程,很少有人关注不平衡数据的聚类问题。虽然 Li 等[14]指出数据不平衡比例越大,聚类算法的效果就越差,但是他们没有解决聚类过程中的不平衡问题。因此,本文提出了在聚类挖掘中解决不平衡问题的方法,并给出了一种新颖的聚类算法和相关评估指标来验证所提方法的有效性。

3　不平衡数据融合的处理技术

3.1　数据融合概述

数据融合最早产生于 20 世纪 70 年代,相关应用研究从最初的军事领域逐步拓展到资源管理、城市规划、气象预报等多个领域[15]。数据融合的最初模型可以分为数据级融合、特征级融合和决策级融合,后来又被扩展成目标提炼、态势分析、威胁估计和过程精炼这 4 级[16]。根据最近的研究成果,多源数据融合的方法可以分成三大类[3]:1)阶段性的方法,即先用一种数据再用另一种数据;2)基于特征拼接的方法,深度学习方法,还有传统的特征串联加上一些正则化方法,都是这种方法的分支;3)基于语义信息融合的方法,包括基于概率模型、基于相似度以及迁移学习等方法。本文采用特征融合的方法,从微博签到数据和出租车 GPS 轨迹数据中提取时间特征数据和能够反映人们出行活动区域的空间特征数据进行融合,从而得到一个融合数据集。微博签到数据和出租车 GPS 数据的融合模型如图 1 所示。

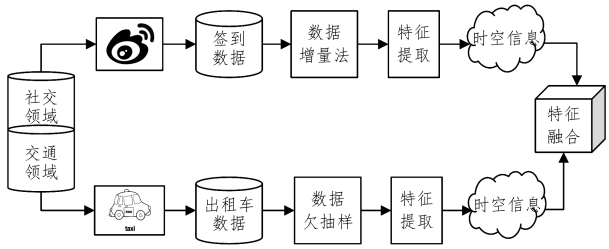

图 1　基于特征层的多源位置数据融合模型

Fig. 1　Multi-source location data fusion model based on features

3.2　基于不平衡数据的聚类挖掘方法

借鉴分类问题中不平衡数据的处理方法[17-18],本文提出不平衡数据融合下的聚类挖掘方法,其包括数据层面和算法层面的改进。数据层面的改进是指对少数类数据(微博签到数据)执行数据增量处理,同时对多数类数据(GPS 数据)进行基于聚类的欠抽样处理。算法层面的改进则是提出了一种基于密度的 IDF-DBSCAN(Imbalanced Data Fusion for DBSCAN)聚类算法。

3.2.1　数据增量法

SMOTE(Synthetic Minority Oversampling Technique)是一种经典的少数类过采样技术[19],其算法基本流程为:1)对于少数类中的每一个样本 x_i,以欧氏距离计算它到少数类样本集中其他样本的距离,得到其 k 近邻;2)根据样本不平衡比例设置采样倍率 N,对于每一个少数类样本 x_i,从其 k 近邻中随机选择若干个样本,假设选择的近邻为 x_n;3)对于每一个随机选出的近邻 x_n,分别按照式(1)与原样本构建新的样本 x_{new}:

$$x_{new}=x_i+rand(0,1)*|x_i-x_n| \tag{1}$$

但是,SMOTE 算法得到的样本点是虚拟的,可能会产生噪声点,而本文采用真实数据来扩充少数类样本,这种方法称为数据增量方法。

图 2 给出了昆明市新浪微博签到数据集连续两个月的数据量。

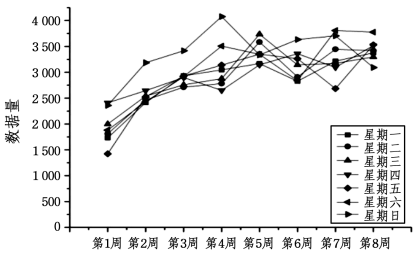

图 2 2015 年 9 月和 10 月的签到数据量

Fig. 2 Check-in data volume in Setp. & Oct. 2015

从图 2 中可以看出:连续 8 周的工作日和休息日的签到数量的变化趋势基本一致,对应数据累加后不会影响数据的变化规律。因此,本文的数据增量方法是指将多个星期的真实微博签到数据累加到一个星期,得到待融合的签到数据集 C,以减小两个数据集中数据量不平衡的比例。

3.2.2 数据欠抽样

数据欠抽样一般是为了减少多数类数据的数据量,按一定的比例系数抽取其中的部分数据进行研究[20]。传统的数据抽样,特别是对多数类数据直接进行欠抽样处理,会造成数据特征的丢失。为避免丢失某些热点区域,先对 GPS 轨迹数据进行单源聚类,得到聚类簇集合 $Q=\{Q_1,Q_2,\cdots,Q_n\}$,然后根据聚类后的不平衡比例系数确定采样倍率 N,并对每个聚类簇进行数据抽样,最后得到欠抽样后的 GPS 数据集 G。采样倍率 N 的计算方法如下:

$$N=round\left(\frac{\sum_{\Gamma=1}^{m}VC_{\Gamma}}{\sum_{\Gamma=1}^{m}\left(\sum_{i=1}^{n}VQ_{\Gamma i}\right)}\right) \tag{2}$$

其中,Γ 表示时间段,根据居民的作息规律,本文将一天的数据分为 13 个时间段,即 $m=13$;n 表示某个时间段内 GPS 数据的聚类簇个数;$VQ_{\Gamma i}$ 表示第 Γ 个时间段内聚类簇 Q_i 的数据量;VC_{Γ} 表示第 Γ 个时间段内签到数据 C 的数据量。

3.2.3 DBSCAN 聚类算法的基本原理

由于热点区域的数量不可预知且常常随时空发生变化,因此采用基于密度的聚类算法来挖掘热点区域是该类研究中的常用方法。DBSCAN 算法是一种基于密度的经典聚类算法,但其无法准确识别不同密度的数据集,即很难发现少数类数据形成的热点区域。为此,本文提出了适用于不平衡数据融合的 IDF-DBSCAN 聚类算法。DBSCAN 算法涉及距离半径 Eps 和最小点 $MinPts$ 两个参数;而 IDF-DBSCAN 算法则根据核心对象的数据来源选取对应的聚类参数,并计算与核心对象密度可达的点和与核心对象密度相连的点。假设数据集 $D=\{x_1,x_2,\cdots,x_m\}$,下面给出算法中的相关定义。

定义 1(ε-邻域) 对于 $x_j\in D$,其 ε-邻域包含样本集 D 中与 x_j 的距离不大于 ε 的样本,即 $N\varepsilon(x_j)=\{x_i\in D|dist(x_i,x_j)\leqslant\varepsilon\}$。

定义 2(核心对象,core object) 若 x_j ε -的邻域至少包含 $MinPts$ 个样本,即 $N\varepsilon(x_j)\geqslant$MinPts,则称 x_j 是一个核心对象。

定义 3(密度直达,directly density-reachable) 如果 x_j 位于 x_i 的 Eps 邻域内,且 x_i 是一个核心对象,则称 x_j 从 x_i 密度直达。

定义 4(密度可达,density-reachable) 对于 x_i 和 x_j,若存在样本序列 $p_1,\cdots,p_i,\cdots,p_n$,其中 $p_1=x_i,p_n=x_j$,且 p_i 是从 p_{i+1}由 p_i 密度直达,则称 x_j 从 x_i 密度可达。

定义 5(密度相连,density-connected) 对于 x_i 和 x_j,若存在 x_k 使得 x_i 与 x_j 均由 x_k 密度可达,则称 x_j 从 x_i 密度相连。

有关密度直达、密度可达和密度相连的实例如图 3 所示。设 $Eps=r$,$MinPts=3$,点 M,P,O 和 R 的 Eps 近邻均包含 3 个以上的点,因此它们都是核心对象。点 M 是从点 P 直接密度可达,点 Q 是从点 M 直接密度可达,因此点 Q 从点 P 密度可达,但点 P 从点 Q 无法密度可达。类似地,S 和 R 是从 O 密度可达的点,且 O,R 和 S 密度相连。

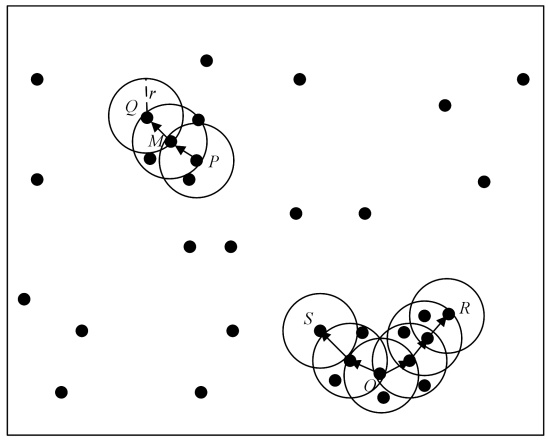

图 3 密度相连的图例

Fig. 3 Example of density-connection

以图 4 为例来说明不平衡数据融合后的聚类效果,其中,圆点代表 GPS 轨迹数据,三角形代表签到数据。由于两者是不平衡数据集,因此会出现密度分布不均匀的情况。IDF-DBSCAN 算法根据 GPS 数据集 G 和微博签到数据集 C 分别设置两组参数,一组参数为{$G.eps=40$ m,$G.MinPts=15$},另一组参数为{$C.eps=40$ m,$C.MinPts=7$}。在计算时,先根据样本 x_i 的数据来源选择合适的邻域参数,找出其对应的邻域 $Neps(x_i)$并确定核心对象集合 Ω。参数的选择过程如下:

$$(\varepsilon,MinPts)=\begin{cases}G.eps,G.M\text{ int }Pts, & \text{if } x_i.Source=G\\ C.eps,C.M\text{ int }Pts, & \text{if } x_i.Source=C\end{cases} \tag{3}$$

得到的核心对象集合 $\Omega=\{G_1,G_2,G_3,C_1,C_2\}$。从 Ω 中随机选取一个核心对象(如 G_1),找出由它密度可达的所有样本,这就构成了聚类簇 $Cluster_1$;接着,IDF-DBSCAN 将 G_1 从 Ω 中去除。不断重复上述操作,直至 Ω 为空。

图 4 中,$Cluster_1$ 由 GPS 数据和签到数据混合构成,$Cluster_2$ 由 GPS 数据构成,$Cluster_3$ 和 $Cluster_4$ 则由签到数据构成。$Cluster_2$ 的密度明显高于 $Cluster_3$ 和 $Cluster_4$ 的密度。在大部分情况下,指定 Eps 后,从核心对象出发构成的直接密度可达点或密度可达点的数据来源会相同,例如从 GPS 数据点 G_1,G_2 和 G_3 出发,直接密度可达的点大部分是 GPS 数据点,核心对象的邻域点会较多。类似地,从签到数据点 C_1 和 C_2 出发,直接密度可达的多数是签到数据点,其密度低而且邻域点较少。如果采用传统 DBSCAN 算法,只根据 GPS

数据的密度来设置聚类密度参数,那么 $Cluster_3$ 和 $Cluster_4$ 将被划分为噪声点,导致签到数据的特性丢失;而根据签到数据的密度设置聚类密度参数,会将一些噪声点划分为聚类簇,例如 G_4 和其周围的 GPS 数据点。在多源不平衡数据集聚类时,IDF-DBSCAN 算法能够根据不同来源数据集的数据量来确定多个聚类参数。

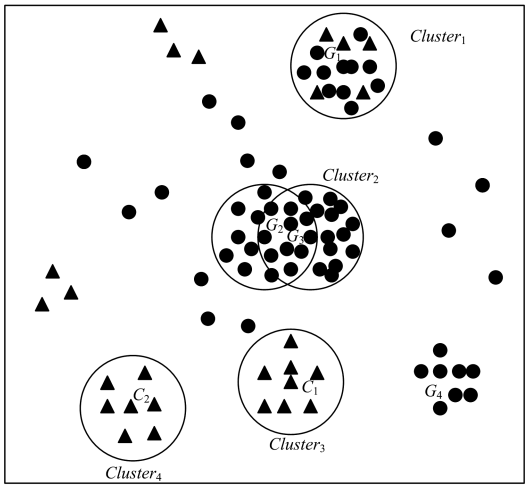

图 4　非平衡数据集的空间分布

Fig. 4　Spatial distributions of imbalanced datasets

3.2.4　IDF-DBSCAN 聚类算法

IDF-DBSCAN 聚类算法中使用的变量及其含义如表 1 所列。

表 1　IDF-DBSCAN 聚类算法中使用的变量

Table 1　Notations for IDF-DBSCAN algorithm

变量	含义
D	数据集合,$D=\{D_1,D_2,\cdots,D_n\}$
C	聚类结果集合,$C=\{C_1,C_2,\cdots,C_n\}$
$G.eps$	核心点为 GPS 数据点时的聚类半径
$G.Minpts$	核心点为 GPS 数据点时的聚类密度
$C.eps$	核心点为签到数据点时的聚类半径
$C.Minpts$	核心点为签到数据点时的聚类密度
$dist$	数据集中任意两点间的球面距离

IDF-DBSCAN 聚类算法的伪代码如算法 1 所示。

算法 1　IDF-DBSCAN

```
Input:D,G. eps,G. Minpts,C. eps,C. Minpts
Output:C={C1,C2,…,Cn}
1. While (D!=null)
2.   for each unvisited point P in D do
3.     dist=dist(P,Di)
4.     Sr=P. Source
5.     if Sr. equals("G")then / * 对象 P 是 GPS 数据 * /
6.     if dist<G. eps then
7.       Neps(P). add(Di) / * P 的邻域 Neps(P) * /
8.       if Neps(P). size < G. Minpts then
9.         mark P as NOISE
10.      else C=new cluster / * 创建一个新类簇 * /
11.      expandCluster(P,Neps(P),C,G. eps,G. MinPts)
12.  else / * 点 P 是签到数据 * /
13.  if dist < C. eps then
14.    Neps(P). add(Di) / * 得到 P 点的邻域 Neps(P) * /
15.    if Neps(P). size < C. Minpts then
16.    mark P as NOISE
17.    else C=new cluster
18.  expandCluster(P,Neps(P),C,C. eps,C. MinPts)
19. expandCluster(P,Neps(P),C,C. eps,C. MinPts)
20.  add P to clusters set C / * 核心点先加入 * /
21.  for each point P′ in Neps(P) do
22.  Neps(P′)=GetNeibors(P′,D) / * 计算 P′点的邻域 * /
23.  if Neps(P′). size≥Source. MinPts then
24.    add Neps(P′) to clusters set C
25.  else add P′ to clusters set C
26. return C
```

IDF-DBSCAN 聚类算法考虑了多源数据的不平衡性,保留了少数类数据的价值,其具体分为两个阶段。第一阶段,先遍历数据集 D 中所有未被处理过的对象 P,计算对象点 P 与剩下所有未被处理过的点的距离;然后根据对象 P 的数据来源选择相应的聚类参数,从而得到 P 的邻域 $Neps(P)$;接着判断 P 是否是噪声点,如果 P 不是噪声点,则以 P 为核心点新建聚类簇 C_i。第二阶段是扩充聚类簇 C_i,首先遍历 $Neps(P)$中的所有点 P',计算得到 P'的邻域 $Neps(P')$,如果 P'是核心点且 $Neps(P')$包含至少 $Source.MinPts$ 个对象,则将 $Neps(P')$加入聚类簇 C_i,否则只将 P' 加入聚类簇 C_i。

对于每个点,该算法只需要维持少量数据,即簇标号和每个点的标识(核心点、边界点或噪声点),因此其空间复杂度为 $O(N)$。第一阶段的时间复杂度为 $O(N)$,第二阶段在最坏情况下的时间复杂度为 $O(N)$,因此总的时间复杂度为 $O(N^2)$。

4　实验分析

4.1　数据来源

本文的数据包括 2015 年 9 月到 11 月的 272 612 条新浪微博签到数据和 2015 年 9 月 7 日到 9 月 13 日一周的出租车上下车点数据,共计 1 245 308 条记录。签到数据的格式为时间、用户 ID、签到经度、签到纬度和签到内容。GPS 数据的格式为车牌 ID、时间、经度、纬度、载客状态、行车速度、道路名称等。融合后的数据格式简化为:时间、经度、纬度和数据来源。数据来源中,0 代表 GPS 数据,1 代表签到数据。表 2 列出了一周内两种数据源的数据量,可以发现出租车的 GPS 数据量是微博签到数据量的 60 倍左右;当对签到数据作一个月增量处理和两个月增量处理后,比例分别降到 12 倍和 6 倍左右,如图 5 所示。

表 2　签到数据与 GPS 数据的数量

Table 2　Data volume of GPS data and check-in data

时间	GPS 数据量	签到数据量	签到数据一个月增量	签到数据两个月增量
2015-09-07	165 952	2 414	11 376	23 948
2015-09-08	167 621	2 463	9 768	23 115
2015-09-09	172 093	2 540	10 174	23 511
2015-09-10	174 725	2 645	10 604	23 319
2015-09-11	186 676	2 525	10 016	22 652
2015-09-12	198 616	2 427	10 732	23 978
2015-09-13	179 625	3 186	13 037	26 791

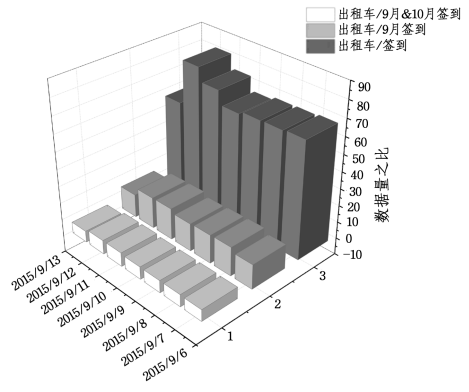

图 5 GPS 数据量与签到数据量的比例

Fig. 5 Ratio of data volume of GPS dataset and check-in dataset

4.2 实验方法

针对融合聚类数据集不平衡的问题，本文从数据层面、算法层面、数据与算法相结合的层面设计了 3 种实验方法，分 6 种方式进行实验对比，具体如表 3 所列。其中，数据增量包括 9 月签到数据增量处理、9 月和 10 月签到数据增量处理。为避免聚类参数对实验效果的影响，在每种方法下都采用 5 组参数进行测试，且 5 组参数都遵循低幅变动的原则。

表 3 实验方法的概述

Table 3 Overview of experimental methods

方法	特点	方法描述
1	不做处理	数据融合后直接使用 DBSCAN 算法进行聚类
2	IDF-DBSCAN	数据融合后使用 IDF-DBSCAN 算法进行聚类
3	数据增量	对签到数据进行增量处理后再融合，使用 DBSCAN 算法进行聚类
4	聚类抽样	对 GPS 数据进行聚类抽样处理后与签到数据融合，使用 DBSCAN 算法进行聚类
5	数据增量＋IDF-DBSCAN	对签到数据进行增量处理与 GPS 数据融合，使用 IDF-DBSCAN 算法进行聚类
6	增量＋聚类抽样＋IDF-DBSCAN	分别对签到数据和 GPS 数据进行增量处理和聚类抽样，之后进行数据融合，再使用 IDF-DBSCAN 算法进行聚类

5 实验结果及分析

5.1 聚类有效性的评价标准

聚类结果的有效性评价指标包括内部评价指标和外部评价指标。谢娟英等[21]利用方差来度量类间分离度和类内紧密度，将类间分离度与类内紧密度之比作为内部度量指标，即 STDI 指标，并验证了 STDI 的性能优于现有的内部指标。虽然相关文献研究了不平衡数据和不同密度类簇的聚类有效性，但它们还没有考虑类偏斜问题。在 STDI 的基础上，本文提出了一种内部评价指标和外部评价指标相结合的综合评价模型 SIID(Synthetic Clustering Index for Imbalanced Datasets)来衡量不平衡数据聚类结果的质量。SIID 的计算如式(4)所示：

$$SIID=\left\| \frac{1}{K}\left(\sum_{k=1}^{K}\left\| c_k-\bar{x}\right\|^2\right)/\sum_{k=1}^{K}\frac{1}{n_k}\left(\sum_{i=1}^{n_k}\left\| x_i-c_k\right\|^2\right)\right\|_{\text{norm}}\times W_{in}+\frac{N_f}{N_g+N_c-N_p}\times W_{ex} \tag{4}$$

其中，K 表示聚类簇的个数，c_k 表示类簇 k 的质心，$\bar{x}$ 表示所有聚类簇的质心，n_k 表示类簇 k 的数据量，x_i 表示类簇 k 的第 i 个数据点，N_f 是多源数据融合后聚类热点区域的个数，N_g 是多数类数据单源聚类挖掘到的热点区域数，N_c 是所有少数类数据单源聚类挖掘到的热点区域数，N_p 是多数类数据和少数类数据所共有的热点区域数。W_{in} 和 W_{ex} 分别是内部指标和外部指标的权重，各设为 0.5。SIID 中，内部指标的分子是各类簇间的距离方差，分母是各类簇内的距离方差之和。分母越小，类内方差越小，说明聚类簇内部的紧密度越高；分子越大，即类间方差越大，说明各类簇之间越分离。整体来看，SIID 的值越大，聚类效果就越好；反之，聚类效果越差。

5.2 实验结果分析

6 种方法在 5 组参数下的聚类结果有效性评估如图 6 所示，其中签到数据为一个月增量数据。

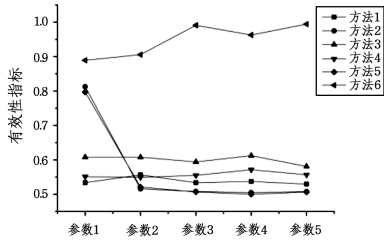

图 6 签到数据一个月增量下 6 种方法的 *SIID* 指标值

Fig. 6 *SIID* values of six methods under 5 sets of parameters using by one-month data accumulation

方法 1、方法 3 和方法 4 都采用 DBSCAN 算法进行聚类，它们在 5 组参数下的 *SIID* 值很平稳，但指标值偏低，基本都在 0.530～0.630 之间。方法 2、方法 5 和方法 6 均采用 IDF-DBSCAN 聚类算法进行聚类，在 5 组参数下的 *SIID* 值不稳定，最低值为 0.500，最高值达到 0.994。方法 2 和方法 5 在第 2 组参数下的指标值从 0.800 下降到 0.500，这是因为聚类簇数量明显增多，聚类簇分布较密集，导致聚类结果的内部指标明显降低。这说明在数据量相同时，聚类参数低幅变动对 DBSCAN 算法聚类结果的影响不大。但是，在多源不平衡数据聚类时，DBSCAN 算法难以挖掘出少数类的数据特征，造成整体指标值偏低，而 IDF-DBSCAN 算法易受参数变化的影响。

对签到数据做两个月增量处理时的评估结果如图 7 所示。方法 2 和方法 6 依然不稳定但偏高；但对比图 6 的结果，有效性指标整体偏低，因为签到数据量急剧增加，加大了签到数据的密度，从而出现了一些与实际不符的聚类簇，导致聚类簇密度增大，聚类结果的内部指标明显降低。

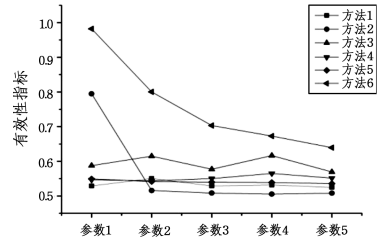

图 7 签到数据两个月增量下 6 种方法的 *SIID* 指标值

Fig. 7 *SIID* value of six methods under 5 sets of parameters using by two-month data accumulation

为了避免聚类参数对实验对比结果的影响，将每种方法下的 5 组参数结果取均值进行对比，结果如图 8 和图 9 所示。

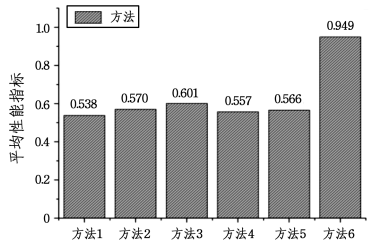

图 8　一个月增量下 6 种方法的 *SIID* 均值

Fig. 8　Average *SIID* of six methods using by one-month data accumulation of check-in data

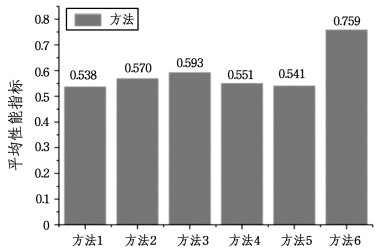

图 9　两个月增量下 6 种方法的 *SIID* 均值

Fig. 9　Average *SIID* of six methods using by two-month data accumulation of check-in data

无论对签到数据进行一个月增量处理还是两个月增量处理，单独使用数据增量、基于聚类抽样、IDF-DBSCAN 聚类方法都优于不做处理的结果，但效果不明显。第 6 种方法的评估结果都明显优于其他 5 组方法的结果。对比结果表明：在对不平衡数据进行聚类时，对多数类数据、少数类数据都进行抽样处理后再与算法改进相结合的方法，优于单纯进行数据层面的处理或单纯进行算法改进的处理方法。此外，对签到数据进行一个月增量处理时，第 6 种方法在第 5 组参数下的 *SIID* 值最高。

下面继续对比 SIID 指标的有效性，如图 10 所示。

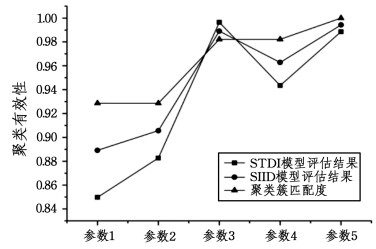

图 10　STDI 与 SIID 评价指标的对比

Fig. 10　Comparisons of STDI and SIID

图 10 中，第 6 种方法在参数 3 下的 STDI 指标最高。但是，采用参数 3 的聚类结果作为最终挖掘结果并不是最优的，因为其挖掘出来的聚类簇与实际热点区域的匹配度为 98%，还有 2%的真实热点区域没有被挖掘。而在 SIID 评价指标下，第 5 组参数的聚类结果最好，其挖掘出来的聚类簇与实际热点区域的匹配程度达到 100%。对比结果证明：SIID 指标更适合评估多源不平衡数据融合下的聚类质量，且真实有效。

最后，利用 ArcGIS 平台展示第 6 种方法的聚类结果，如图 11 所示。同时，将出租车 GPS 数据聚类得到的热点区域图、签到数据聚类得到的热点区域图以及它们所形成的热点区域叠加图与图 11 进行对比，结果如图 12－图 14 所示。

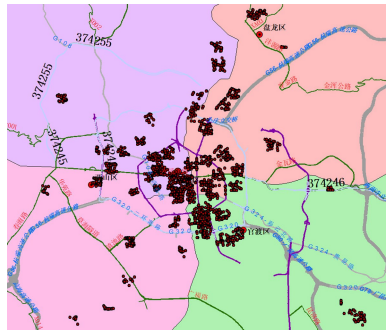

图 11　数据融合后形成的热点区域图

Fig. 11　Urban hotspots generated by the fused data

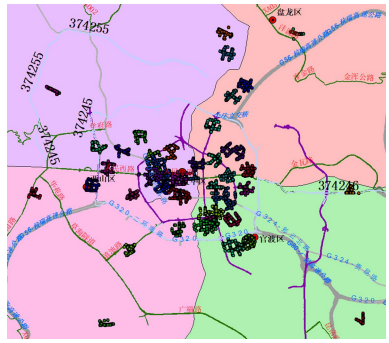

图 12　出租车 GPS 数据形成热点区域图

Fig. 12　Urban Hotspots generated by the GPS data

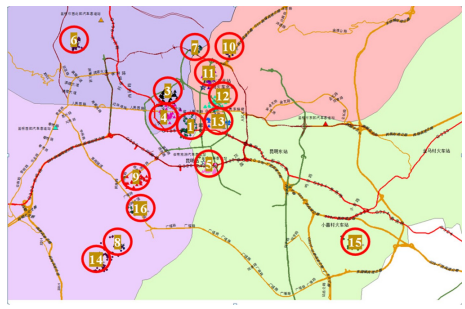

图 13　签到数据形成的热点区域图

Fig. 13　Urban hotspots generated by the check-in data

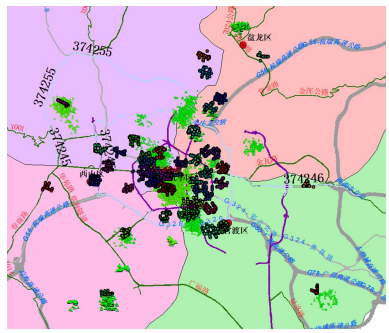

图 14　两种数据集热点区域叠加图

Fig. 14　Overlay effect based on the GPS and check-in data

从图 11－图 14 可以发现：多源数据融合后挖掘到的热点区域个数明显多于单源数据挖掘到的热点区域个数。这表明利用多源融合后的数据来挖掘城市热点区域可以解决使用单源数据带来的聚类结果不全面的问题。而且，当数据不平衡比例在 12 倍左右时，将数据增量、基于聚类的欠抽样和 IDF-DBSCAN 算法三者相结合是处理不平衡数据聚类问题

的一种高效、可行的方法。

结束语 针对多源位置数据在融合后出现的数据量不平衡问题,本文从模型、数据、算法和评估指标等多个方面提出了一套完整的解决方案。利用所提方案对两种来源的位置数据执行聚类挖掘后,发现了 11 个新增的热点区域,这些区域主要分布在昆明市的远郊旅游景点、步行街和高校周边。由于出租车的工作特性,单纯使用 GPS 数据无法挖掘到这些新增区域。使用传统的 DBSCAN 算法只能发现 83%的热点区域,由少数类数据形成的热点区域几乎全部丢失,而使用 IDF-DBSCAN 聚类算法能够发现 100%的热点区域。本文所提方案不仅适用于 GPS 轨迹数据和签到数据,还可以对其他融合后的位置数据进行分析和评价,如公交卡数据、地铁数据和手机定位数据等。由于 IDF-DBSCAN 聚类算法对 Eps 和 $Minpt$ 两个参数的取值比较敏感,不同取值会造成聚类结果的波动,而且两个参数的确定也比较困难,下一步可对 IDF-DBSCAN 聚类算法进行参数自适应方面的研究。此外,还可以考虑将一些索引结构(如 KD 树或者 Octree)应用在 IDF-DBSCAN 算法中,以提高其运行效率,并降低时间复杂度。

参 考 文 献

[1] YUAN J, ZHENG Y, XIE X. Discovering Regions of Different Functions in a City Using Human Mobility and POIs[C]// Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2012: 186-194.

[2] CHEN Y, YUAN P, QIU M, et al. An Indoor Trajectory Frequent Pattern Mining Algorithm Based on Vague Grid Sequence [J]. Expert Systems With Applications, 2019, 118: 614-624.

[3] ZHENG Y. Methodologies for Cross-Domain Data Fusion: An Overview[J]. IEEE Transactions on Big Data, 2015, 1(1): 16-34.

[4] DING Z Y, JIA Y, ZHOU B. Research Summary of Wei bo Data Mining[J]. Journal of Computer Research and Development, 2014, 51(4): 691-706. (in Chinese)
丁兆云,贾焰,周斌. 微博数据挖掘研究综述[J]. 计算机研究与发展, 2014, 51(4): 691-706.

[5] LEE J, SHIN I, PARK G, et al. Analysis of the Passenger Pick-up Pattern for Taxi Location Recommendation [C] // 2008 Fourth International Conference on Networked Computing and Advanced Information Management. New York: IEEE, 2008, 1: 199-204.

[6] KISILEVICH S, MANSMANN F, KEIM D. P-DBSCAN: A Density Based Clustering Algorithm for Exploration and Analysis of Attractive Areas Using Collections of Geo-tagged photos [C]// Proceedings of the First International Conference and Exhibition on Computing for Geospatial Research & Application. New York: ACM, 2010: 38-41.

[7] VERMA N, BALIYAN N. PAM Clustering Based Taxi Hotspot Detection for Informed Driving [C] // 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). New York: IEEE, 2017: 1-7.

[8] NING P F, WANG Y, SHEN Y R, et al. Identification of Urban Interest Function Region by Using Social Medida Check-in Data [J]. Journal of Geomatics, 2018, 43(2): 110-114. (in Chinese)
宁鹏飞,万幼,沈怡然,等. 基于签到数据的城市热点功能区识别研究[J]. 测绘地理信息, 2018, 43(2): 110-114.

[9] ORRIOLS-PUIG A, BERNADO-MANSILLA E, GOLDBERG D E, et al. Facetwise Analysis of XCS for Problems With Class Imbalances[J]. IEEE Transactions on Evolutionary Computation, 2009, 13(5): 1093-1119.

[10] KRAWCZYK B, MCINNES B T. Local ensemble learning from imbalanced and noisy data for word sense disambiguation[J]. Pattern Recognition, 2017, 78: 103-119.

[11] SEBASTIÁN M, JULIO L. Dealing with High-dimensional Class-imbalanced Data sets: Embedded Feature Selection for SVM Classification[J]. Applied Soft Computing, 2018, 67: 94-105.

[12] ZHAI Y, YANG B R, QU W. Survey of Mining Imbalanced Datasets[J]. Computer Science, 2010, 37(10): 27-32. (in Chinese)
翟云,杨炳儒,曲武. 不平衡类数据挖掘研究综述[J]. 计算机科学, 2010, 37(10): 27-32.

[13] ZHU Y J, WANG Z, ZHA H Y, et al. Boundary-Eliminated Pseudo Inverse Linear Discriminant for Imbalanced Problems [J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(6): 2581-2594.

[14] LI X, CHENG Z G, Fan Y, et al. Exploring of Clustering Algorithm on Class-imbalanced Data [C] // 2013 8th International Conference on Computer Science & Education. New York: IEEE, 2013: 89-93.

[15] PAN Q, WANG Z F, LIANG Y, et al. Basic Methods and Progress of Information Fusion [J]. Control Theory & Applications, 2012, 29(10): 1234-1244. (in Chinese)
潘泉,王增福,梁彦,等. 信息融合理论的基本方法与进展[J]. 控制理论与应用, 2012, 29(10): 1234-1244.

[16] HALL D L, LLINAS J. Handbook of Multi-sensor Data fusion [M]. New York: CRC Press, 2001.

[17] BRODINOVÁ Š, ZAHARIEVA M, FILZMOSER P, et al. Clustering of Imbalanced High-dimensional Media data [J]. Advances in Data Analysis and Classification, 2018, 12(2): 261-284.

[18] GUO H X, LI Y J, JENNIFER S, et al. Learning from Class-imbalanced Data: Review of Methods and Applications[J]. Expert Systems with Applications, 2017, 73: 720-739.

[19] LI K, ZHANG W, LU Q, et al. An Improved SMOTE Imbalanced Data Classification Method Based on Support Degree [C]// 2014 International Conference on Identification, Information and Knowledge in the Internet of Things. New York: IEEE, 2014: 34-38.

[20] DENG X, ZHONG W, REN J, et al. An Imbalanced Data Classification Method Based on Automatic Clustering Under-sampling [C]// Proceedings of IEEE Conference on Performance Computing and Communications. New York: IEEE Press, 2016: 1-8.

[21] XIE J Y, ZHOU Y, WANG M Z, et al. New Criteria for Evaluating the Validity of Clustering[J]. CAAI Transactions on Intelligent Systems, 2017, 12(6): 873-882. (in Chinese)
谢娟英,周颖,王明钊,等. 聚类有效性评价新指标[J]. 智能系统学报, 2017, 12(6): 873-882.