基于信息熵的城市隧道实时交通事件检测聚类

李晓峰[1],杨春山[2],丁树春[3]
(1. 东北农业大学成栋学院 计算机科学与技术系,黑龙江 哈尔滨 150025;
2. 北京理工大学 计算机科学与技术学院,北京 100081;
3. 黑龙江大学 电子工程学院,黑龙江 哈尔滨 150080)

摘 要:空间聚类是空间数据挖掘中的一种重要手段,采用空间聚类技术进行有用信息的获取具有重要的现实意义。针对城市隧道实时路况的特点,在常见的城市隧道实时交通事件检测数据挖掘-关联规则算法的情况下,把与城市隧道相关的对象看作实时路况的属性,计算城市隧道间的信息熵,根据信息熵的变化实现城市隧道实时路况的聚类,提出了基于信息熵的城市隧道实时交通事件检测聚类算法(Entropy-based City Tunnel Real-time,ECRT)。实验表明,通过其在实际数据集上进行的测试,算法 ECRT 高效地解决了拓扑关系的复杂空间数据集中对象的聚类问题。

关键词:信息熵;城市隧道实时;空间数据挖掘;聚类
中图分类号:TP399 **文献标识码**:A **文章编号**:1673-629X(2013)10-0212-04
doi:10.3969/j.issn.1673-629X.2013.10.053

Entropy-based City Tunnel Real-time Traffic Incident Detection Clustering

LI Xiao-feng[1], YANG Chun-shan[2], DING Shu-chun[3]
(1. Department of Computer Science and Technology, Cheng-Dong College of Northeast Agricultural University, Harbin 150025, China;
2. College of Computer Science and Technology, BIT, Beijing 100081, China;
3. College of Electronic Engineering, Heilongjiang University, Harbin 150080, China)

Abstract: Spatial clustering is an important tool in spatial data mining, spatial clustering technology accessing useful information has important practical significance. In view of city tunnel real-time traffic characteristics in common urban tunnel real-time traffic incident detection data mining - the case of the association rules algorithm, the city tunnel object as properties of the real-time traffic, calculate the entropy of information between the city tunnel, achieve real-time city tunnel traffic clustering based on entropy changes, the ECRT (entropy-based city tunnel real-time) algorithm is proposed. The experiments show algorithm ECRT used in the actual data set test can effectively resolve the problem of object clustering in complex spatial data set of topology relation.

Key words: information entropy; city tunnel real-time; spatial data mining; clustering

0 引 言

伴随着信息化技术水平的逐步提高,空间数据的复杂性与数量不断地增长,仅通过数据库技术进行查询已经不能提供给用户有用的信息,导致空间数据库中的信息不能充分利用和挖掘。在此背景下,空间数据库应用通过空间数据挖掘(Spatial Data Mining, SDM)应运而生。

实现空间数据挖掘的主要手段是空间聚类,通常是把空间数据的对象通过类组成簇,在同一个簇里,对象一般具有相似性,在不同的簇间,对象的差异较大。不同簇之间的组合生成了不同密集区域,通常空间对象在一定空间内分布较大密度值,若该密度值大于一定范围的密度值时,这个区域是密集,称密集区域。在知识发现领域聚类是主要应用。根据国内外一些学者的研究,提出了一些相关高效的聚类算法。国外学者 Deneubourg[1],依据蚁群尸体堆积方式提出了一种有

收稿日期:2012-12-13 **修回日期**:2013-03-16 **网络出版时间**:2013-05-09
基金项目:国家青年科学基金项目(61102071)
作者简介:李晓峰(1978-),男,哈尔滨人,副教授,博士,CCF 高级会员,ACM 会员,研究方向为数据挖掘、文本挖掘、智能算法。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20130509.1101.060.html

效的基本模型(Basic Model,BM)。在 BM 基础上 Lumer[2]等人对 BM 实现扩展,提出了针对数据对象的一种度量表达式,在聚类分析中实现了这种表达式。翁怀荣[3]提出了一种改进的蚁群聚类的算法,把感觉特征与随机扰动移入了信息素更新中[4-5]。在其他的应用领域蚁群聚类算法也在研究中。文中主要在论述常见的城市隧道实时交通事件检测数据挖掘——关联规则算法的情况下,提出了一种改进算法——基于信息熵的城市隧道实时交通事件检测聚类算法(ECRT),通过 ECRT 进行实验表明,该算法能提高空间数据的应用与分析的水平。该算法主要考虑了空间数据间的联系与复杂性。主要是精练,隐含知识间关系。

1　相关工作

1.1　空间数据挖掘

空间数据挖掘中的对象是空间数据仓库与数据库,多边形,点,线和面等是粒度。通过尺度空间数据说明了由细至粗的多分辨率与多比例尺的变换过程。尺度越小空间目标的表达越微观与精细。通常关系数据库与商业数据库等是空间挖掘对象,通过关系与事务型数据实现存储。空间数据挖掘以其数据的复杂性、数据源十分丰富,数据量非常庞大,数据类型多,存取方法复杂,从而不同于一般的事物数据挖掘,这些特性也使得大多数算法比较复杂,难度大[6]。

1.2　空间数据挖掘的过程

空间数据挖掘过程一般由 4 部分组成:

1)通过掌握知识领域用户目标与预先知识,构建目标数据集,形成一个或多个子数据集[7-8]。

2)进行数据的转换与清理,去掉无关数据与噪声,并且考虑数据变化与时间顺序,通过转换或者维变化的方法查找数据不变式或者通过这种方法减少变量数目。

3)对数据挖掘的任务进行选择,然后对选取的算法实现挖掘[9-10]。

4)对挖掘的结果进行评估,对有用的模式进行转换,去掉多余模式。把挖掘出的报告给用户,并把挖掘的结果应用到系统中。

1.3　城市隧道实时交通事件检测数据挖掘的主要技术方法

通过查阅大量文献发现,目前大多数空间聚类算法是利用空间对象非空间属性方式实现。这种空间对象除了具有非空间属性,还有拓扑结构的空间特征。城市隧道的实时路况数据挖掘技术主要还是集中在传统的方法。文献[6]通过对道路交通事故特征,导致交通事故的人、车、路等因素进行关联规则数据挖掘,并用虚拟现实技术模拟交通事故的发生过程,实现对交通事故数据的分析。文献[7]则主要阐述了 Floating Car Data(FCD) 技术在实时路况仿真模拟中的应用模式,采用多元线性回归模型模拟隧道通行速度的方法,可以合理、有效地模拟隧道内的实时路况,从而能够真实反映汽车在隧道中通行的真实速度,为交通事故的预防提供了策略[11]。

但是不可否认,传统非空间属性聚类算法具有一定的局限性,如必须通过大量数据的关联对比才能得出一定条件下适用的结论,且该结论在突发事件的情况下不具有适用性[12]。

为了能够更好地对城市隧道实时交通事件进行检测,必须提出一种算法即便在突发事件下也能较真实地反映当前情况,笔者提出了基于信息熵的城市隧道实时交通事件检测聚类算法。

2　城市隧道实时交通事件检测聚类的算法

2.1　算法设计

空间数据集一般通过空间对象间的属性,数量,对象间的距离,拓扑与方向等关系体现。空间对象的以上特性,针对空间数据集的分析不能应用现有的聚类算法,通过信息熵方式解决了应用现有聚类算法的问题。针对空间,空间对象不相似时的信息熵大于空间对象相似时的信息熵,通过这个理论,把信息熵引入聚类算法中应用。通过这种算法高效地解决了拓扑关系的复杂空间数据集中对象的聚类问题。

空间数据集 D 建立空间对象群集 $S=\{S_1,S_2,\cdots,S_m\}$,$\mathrm{CL}=\{C_1,C_2,\cdots,C_k\}$,CL 是 S 一个聚类,C_l 是第 l 个簇,$C_l\subseteq S, l\in[1,k]$,$k$ 是聚类数,通过 CL 使得 F 值最小,依据公式 1 计算目标函数 F,依据公式 2 计算空间对象群信息熵值。

$$F=\sum_{l=1}^{k}E_l \tag{1}$$

$$E_l=E(\{S_i \mid S_i\in C_l\})=-\sum_{h=1}^{n}\sum_{t=1}^{T}p_{ht}*\lg p_{ht} \tag{2}$$

2.2　算法步骤

基于拓扑关系,依据空间数据对象在一定范围内信息熵的变化,通过对蚂蚁捡起与放下空间数据的指导,对空间数据集实现聚类。信息熵的城市隧道实时交通事件检测聚类算法过程如下:

1 For iter=1 to $N_{\text{iteration}}$

2 For q = 1 to N_{ant}

3 If(Space Object$_q$isload = flase)then 空间对象 Space Object$_q$ 随机拾起子空间对象 $S_i, i\in[1,m]$

4 else 计算 S_i 邻域范围内 Space Object$_q$ 未放下 S_i 前的信息熵的值 E_1;计算 S_i 邻域范围内 Space Object$_q$ 放下 S_i 后的信息熵的值 E_2

5 If($E_1 >= E_2$ || fail > N_{fail}) then Space Object$_q$ 在当前位置放下 S_i

6 else Space Object 负载 S_i 随机选择一个方向移动步长 d 到下一个位置, fail ++

通过以上过程的迭代,能动态地对空间对象 Space Object 的步长 d 与搜索半径 r 值进行调整,这种方式避免了算法陷入局部最优,提高了算法的效率。

例如图 1 为空间对象分布在二维空间中,通过迭代运行 ECRT 算法,即通过蚂蚁不断地捡起或放下空间对象后,结果如图 2 所示,相似的空间对象被搬到了一起,而不相似的则被搬离开来。

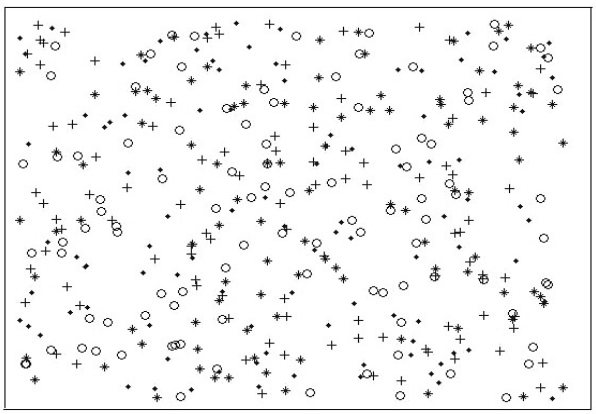

图 1 原始数据分析输出

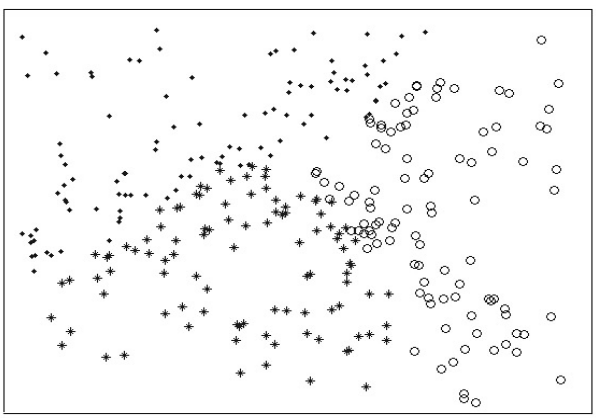

图 2 蚂蚁多次捡起或放下后空间对象的分布

2.3 ECRT 算法优化

从图 2 可以看出,虽然相似的空间对象被搬到了一起,不相似的被分离开来,但是,簇之间的分界线还是模糊的,图 2 说明通过信息熵的蚁群实现的聚类结果是粗糙的。

图 2 通过使用 ECRT 算法进行优化,结果如图 3 所示,显然,簇与簇之间的分界线很明显,提高了聚类的准确性。

聚类过程中空间对象 Space Object 仅实现了相异空间对象的分离与相似空间对象的堆积,由于没有实现对不同簇的拉开,因此簇之间的划分很模糊。为使该算法更贴近实际,提出以下优化过程:

1) For i = 1 to m // m 表示空间数据的总数

2) S_i 当前位置记为 p ,蚂蚁拾起 S_i ,空间对象 Space Object 失败次数 fail = 0

3)分别计算空间对象 Space Object 在位置 p 以及在六个方向上,对每个方向移动一步后 S_i 一定范围内的信息熵值。在范围内取最小值,记录对应的空间位置 p'

4) If(fail < N_{fail} && $p \neq p'$) then 空间对象 Space Object 负载 S_i 步行到 p' 位置, fail ++ , $p = p'$,转 3)

5) else 空间对象 Space Object 在 p' 位置将 S_i 放下

3 实验与结果分析

3.1 算法验证

为了验证算法的有效性,以不同时刻上海市的大连路隧道数据为例,通常采用 GPS 测出车辆的速度,由于在隧道内无法接收 GPS 数据,各自选取隧道进出口连通的重要道路路段,测得其实时速度如表 1 所示。

表 1 大连路隧道重要进出口道路路段

时刻	路段编号	相关路段名	相关起点名	相关终点名	实时速度
04:20:02 ~ 04:24:56	3488	大连路	唐山路	长阳路	28.06
08:20:02 ~ 08:24:56	3498	大连路	长阳路	霍山路	18.84
12:20:02 ~ 12:24:56	3495	东方路	乳山路	商城路	20.86
14:20:02 ~ 14:24:56	695	商城路	东方路	福山路	26.85
16:20:02 ~ 16:24:56	1996	东方路	栖霞路	乳山路	22.69
18:20:02 ~ 18:24:56	2017	长阳路	许昌路	大连路	21.58
20:20:02 ~ 20:24:56	1492	霍山路	东大名路	临潼路	26.54
22:20:02 ~ 22:24:56	1469	长阳路	临潼路	大连路	27.63
00:20:02 ~ 00:24:56	3847	霍山路	临潼路	大连路	29.01
02:20:02 ~ 02:24:56	1471	乳山路	东方路	崂山路	29.67

上表中实时速度的检测,是通过连续 30 天的观察,取平均值得到的。通过以上的实验分析,并运用 ECRT 算法,得出了如下结果。原始输入数据的分布情况如图 1 所示,使用 ECRT 优化算法后所得到的聚类结果如图 3 所示。

通过图 3 可以看出,若相似的数据则会聚类到同一簇,若出现异常状况则会是孤立的点。这能很清楚地显示大量数据聚类后的特征。通过图 3 也可以看出

本次实验的隧道并不存在异常事件。

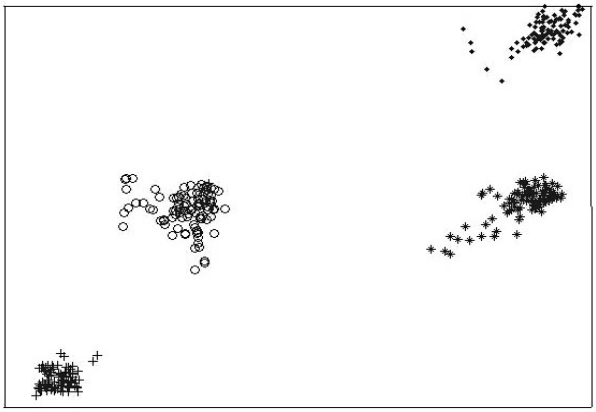

图 3　ECRT 优化算法输出

3.2　算法性能分析

为了验证该算法的性能,笔者在相同环境下,分别使用关联规则算法和文中算法同时对上文实验数据进行了比对,其算法误差精度如图 4 所示,可见文中算法具有更低的误差,尤其是在大量数据的情况下,其精确度较高。

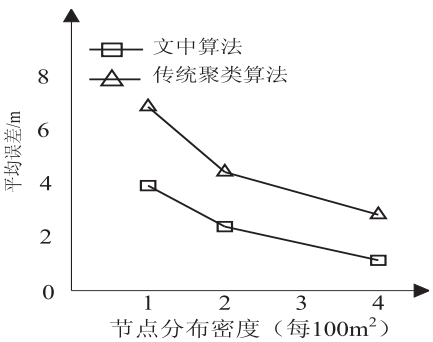

图 4　两种算法精确度情况

4　结束语

文中根据隧道实时监测事件的需要,在原有的关联规则算法分析的基础上,提出了一种新的算法,用实验验证了改算法该有效性,并通过与关联规则算法的比对,得出该算法的性能更高,为该算法的广泛使用奠定了基础。

参考文献:

[1] 吴　斌,郑　毅,傅伟鹏,等. 一种基于群体智能的客户行为分析算法[J]. 计算机学报,2003,26(8):913-918.

[2] 颜宏文,马　瑞,晏弼成. 基于信息熵构造判定树的数据挖掘算法的设计与实现[J]. 计算机工程与应用,2003,40(23):180-182.

[3] 翁怀荣,张洪伟,钟　响,等. 基于改进的蚁群算法的聚类分析及其在 HRM 中的应用[J]. 计算机应用,2008,25(8):1908-1912.

[4] 王雪松,石　琦,高　珍. 基于视频数据的城市隧道交通运行特征与安全研究[J]. 中国安全科学学报,2011(8):129-137.

[5] 周悦来,谭建豪. 基于网格和信息熵的多密度聚类算法[J]. 计算机系统应用,2011,20(10):189-192.

[6] Lin Chengru, Liu Kenhao, Chen Ming-syan. Dual Clustering: Integrating Data Clustering over Optimization and Constraint Domains[J]. IEEE Transactions on Knowledge and Data Engineering,2005,17(5):628-637.

[7] Zhang Xueping, Wang Jiayao, Wu Fang. A Novel Spatial Clustering with Obstacles Constraints Based on Genetic Algorithms and K-Medoids[J]. International Journal of Computer Science and Network Security,2006,6(10):605-610.

[8] Huang Ming, Bian Fuling. A Grid and Density Based Fast Spatial Clustering Algorithm[C]//Proc. of International Conference on Artificial Intelligence and Computational Intelligence. San Sebastian, Spain:[s. n.],2009:260-263.

[9] Hu Caiping, Qin Xiaolin. A novel spatial cluster algorithm with sampling[M]. Heidelberg: Springer-Verlag,2007:216-225.

[10] 毛德梅,丁瑞国. 对数据挖掘中关联规则算法的比较研究[J]. 皖西学院学报,2006,22(5):27-30.

[11] 李　芸,李青山. 数据挖掘中关联规则挖掘方法的研究及应用[D]. 西安:西安电子科技大学,2007.

[12] 刘义安,羊　斌. 关联规则挖掘中对 Apriori 算法的一种改进研究[J]. 计算机应用,2007,27(2):418-420.

(上接第 211 页)

[7] 宗成庆. 统计自然语言处理[M]. 北京:清华大学出版社,2011:341-342.

[8] 罗　可,蔡碧野,吴一帆,等. 数据挖掘中聚类的研究[J]. 计算机工程与应用,2003,39(20):182-184.

[9] Choy S O, Lui A K. Web information retrieval in collaborative tagging systems[C]//Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. Washington, DC: IEEE Computer Society,2006:352-355.

[10] Baldassarri A, Cattuto C, Loreto V, et al. Ranking and community detection in undirected networks[EB/OL]. [2008-10-05]. http://www.tagora-project.eu/wp-content/2007/04/talk-servedio-folkrank.pdf.

[11] 边肇祺,张学工. 模式识别[M]. 第 2 版. 北京:清华大学出版社,2000.

[12] 田莹颖. 基于社会化标签系统的个性化信息推荐探讨[J]. 图书情报工作,2010,54(1):50-53.

基于信息熵的城市隧道实时交通事件检测聚类

作者： 李晓峰， 杨春山， 丁树春， LI Xiao-feng， YANG Chun-shan， DING Shu-chun

作者单位： 李晓峰,LI Xiao-feng(东北农业大学成栋学院 计算机科学与技术系,黑龙江 哈尔滨,150025)， 杨春山,YANG Chun-shan(北京理工大学 计算机科学与技术学院,北京,100081)， 丁树春,DING Shu-chun(黑龙江大学 电子工程学院,黑龙江 哈尔滨,150080)

刊名： 计算机技术与发展

ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2013(10)