

基于多聚类结果融合的轨迹聚类方法

李 静, 张 磊, 韩陈寿

(中国矿业大学 计算机学院, 江苏 徐州 221116)

摘 要: 针对轨迹聚类结果的不可靠性, 提出一种基于多聚类结果融合的轨迹聚类方法 MRMTC. 对于多聚类器产生的多个聚类代表轨迹, 提出了轨迹合并算法, 实现了多个聚类代表轨迹的合并. 代表轨迹合并算法以平均扫描线距离函数作为共识函数, 通过共识函数对代表轨迹间的相似度进行比较, 最后合并相似的代表轨迹. 实验表明基于融合的轨迹聚类方法, 可以获得比单一聚类更有效更稳定的聚类结果.

关键词: 轨迹; 聚类; 聚类融合

中图分类号: TP312

文献标识码: A

文章编号: 1000-7180(2011)08-0063-04

Trajectory Clustering Method Based on Multi-clustering Results Merging

LI Jing, ZHANG Lei, HAN Chen-shou

(School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: In view of the unreliable of trajectory clustering results, a trajectory clustering method based on multi-clustering results merging (MRMTC) is proposed in this paper. For the representative trajectories of multi-clustering generated by clustering devices, a trajectory merging algorithm is proposed to merge them. The merging algorithm uses average scan line distance function as consensus function to compare similarities of representative trajectories, and then merges the similar representative trajectories. Finally, experiments results show that the proposed method MRMTC can produce more stable and effective clustering results.

Key words: trajectory; clustering; clustering merging

1 引言

目前对轨迹聚类方法的研究有很多. Yutaka Yanagisawa^[1]等人提出一种基于移动对象的形状和速度的多维度模型进行轨迹聚类的方法. 文献[2]提出 TCMM 轨迹聚类算法用于轨迹的增量聚类. 文献[3]也提出类似的增量轨迹聚类方法, 并且在真实的视频轨迹内做在线增量轨迹聚类有较好的表现. 但这些方法大多是对采样点的位置进行聚类, 没有把轨迹所在环境考虑进去. 而对于相同的轨迹数据, 采用不同的轨迹聚类算法或对同一个聚类算法使用不同的聚类参数设置时, 可能会得出完全不同

的聚类结果. 融合方法就是将得到的这些聚类结果进行合并, 以获得比单一算法更优越的结果. 在回归模型和分类算法中, 融合方法的使用已经比较成熟^[4]. 但在聚类分析领域, 在近几年才开始出现聚类融合方法的研究. A. Strehl 和 J. Ghosh 于 2002 年在文献[5]中最早提出聚类融合的概念.

本文提出基于多聚类结果融合的轨迹聚类方法 MRMTC (Multi-Results Merging based Trajectory Clustering), 该方法采用两种典型的聚类器: 基于多约束的轨迹聚类、基于环境相关的轨迹聚类. 并采用这两种聚类器的多种参数组合进行聚类, 然后采用本文提出的轨迹合并算法对这些聚类结果的代

收稿日期: 2011-05-18; 修回日期: 2011-06-15

基金项目: 国家自然科学基金项目(50674086); 高等学校博士学科点专项科研基金项目(20060290508); 江苏省博士后基金(0802023C)

表轨迹进行合并,实现多聚类结果的融合.

2 两个典型的轨迹聚类器

定义 1 Environment $A = \{A_1, A_2, \dots, A_n\}$ 为轨迹点的环境相关属性集合, A_i 为一个轨迹点的第 i 个环境相关属性, Environment A_i 为第 i 个轨迹点的环境相关属性集合. 环境相关属性是直接跟轨迹点所在环境和轨迹点都相关的量.

定义 2 Inherent $B = \{B_1, B_2, \dots, B_n\}$ 为轨迹点的内在属性集合, B_i 为一个轨迹点的第 i 个内在属性, Inherent B_i 为第 i 个轨迹点的内在属性集合. 内在属性是不依靠具体环境就可以直接测量到的属性.

轨迹 (Trajectory) 是多维度的包含环境信息的有序序列: $TR_i = \{P_1, P_2, \dots, P_m\} (1 \leq i \leq n)$, 其中轨迹点 $P_j (1 \leq j \leq m)$ 表示为 $\langle \text{EnvironmentProperties}, \text{InherentProperties} \rangle$ 包含环境相关信息和内在信息.

定义 3 子轨迹段的环境相关属性值 (EnvironmentA) 定义为

$$\text{Environment } A_{\tau}^j = \frac{(A_i^j + A_{i+1}^j + \dots + A_{i+k}^j)}{k} \quad (1)$$

式中, k 为子轨迹段内的轨迹点数目, A_i^j 为第 i 个轨迹点的第 j 个环境相关属性量值. 对某个具体的环境相关属性量, 子轨迹段的环境相关属性值为子轨迹段内所有点的平均值. 如图 1 所示, 两条子轨迹段离最近的公路 (Road A) 的最近距离分别是 $\text{Environment } A_{\tau 1}$ 和 $\text{Environment } A_{\tau 2}$.

基于多约束的轨迹聚类器: 用以考虑轨迹内在属性, 对轨迹进行聚类. 这里把改进后用于轨迹聚类的 OPTICS 算法命名为 T-OPTICS 算法, 以对子轨迹段进行聚类.

基于环境相关的轨迹聚类器: 用以考虑轨迹点与环境相关的属性量, 对轨迹进行聚类. 为了简化计算, 只选取环境相关属性中一个主要属性进行聚类. 该方法第一步根据环境相关属性量采用 K-Means 算法进行宏观层次的轨迹聚类; 第二步在第一步的每个宏观聚类结果内采用基于密度的 T-OPTICS 根据轨迹位置坐标进行微观聚类.

3 基于多聚类结果融合的轨迹聚类方法

3.1 算法基本思想

基于融合的轨迹聚类方法, 先将轨迹进行划分得到轨迹段^[6], 然后考虑轨迹的内在属性特征和环

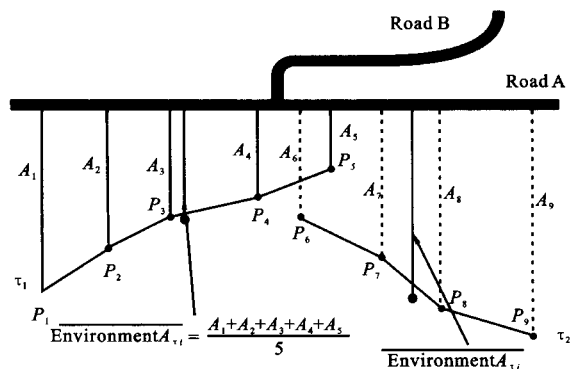


图 1 环境相关属性量示意图

境信息, 对轨迹段多次聚类得到每次聚类结果产生的代表轨迹; 最后采用共识函数对聚类结果代表轨迹进行融合, 得到最终的聚类结果代表轨迹.

3.2 多聚类代表轨迹合并方法

本文采用平均扫描线距离函数作为聚类融合方法中的共识函数. 通过共识函数对代表轨迹间的相似度进行比较; 两条代表轨迹间的垂直线段距离平均值 $\overline{wd}$ 与子聚类器的子轨迹段距离比较阈值的加权平均值 ρ 的比较.

$$\overline{wd} = \frac{\overline{P_1 Q_1} + \overline{P_p Q_q} + \sum_{i=1}^{p-1} \overline{P_i P'_i} + \sum_{j=1}^{q-1} \overline{Q_j Q'_j}}{p + q} \quad (2)$$

$$\rho = \sum_{i=0}^k \epsilon_i \times n_i / \sum_{i=0}^k n_i \quad (3)$$

式(2)中, $\overline{wd}$ 中 p 为代表轨迹 P 的代表点个数, q 为代表轨迹 Q 的代表点个数. $\overline{P_i P'_i}$ 是两条代表轨迹过 p_i 点的垂直连线距离. 式(3)中, ρ 中 k 是所有聚类器聚类次数, n_i 是每次子聚类的聚类结果数, ϵ 是子轨迹段的基于空间的距离阈值.

如图 2 所示, 对于要合并的两条轨迹 P, Q , 计算它们的平均方向向量 v , 然后通过转置矩阵使它们主轴方向平行于 X' 方向. 把两条轨迹按照 X' 轴排序, 得到排序后的 P, Q 首尾两条连接线 (图 2 的 L_s 和 L_e). 计算它们的中点 $\overline{P_s}$ 和 $\overline{P_e}$. 从左到右用扫描线扫描所有的点, 除起始点和结束点外, 对每个点做垂

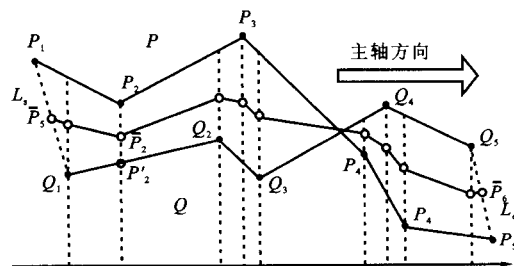


图 2 代表轨迹合并

直主轴(即 X' 轴)的垂线交于另外一条轨迹,如在轨迹 P 上过 P_2 的垂线交于轨迹 Q 的 P'_2 点,连接该两点得到 $P_2P'_2$ 线段,并取它的中点 $\overline{P_2}$. 得到各个垂直连接线的中点序列 $\overline{P_1} \overline{P_2} \cdots \overline{P_k}$, 加入前面的 $\overline{P_s}$ 和 $\overline{P_e}$, 得到序列 $\overline{P_s} \overline{P_1} \overline{P_2} \cdots \overline{P_k} \overline{P_e}$, $k = Plen + Qlen - 2$. 连接这些点,得到合并的轨迹.

3.3 MRMTC 的融合算法描述

假定 C_i 为采用某个聚类器某种参数组合的聚类结果, RTR_k 为 C_i 内的第 k 条代表轨迹, D 为 C_i 集合, 由此 $C_i = \{RTR_k \mid 1 \leq k \leq len_i\}$ (len_i 为第 i 次聚类结果产生的代表轨迹数量), $D = \{C_i \mid 2 \leq i \leq len\}$ (len 为聚类总次数).

输入: A set of Clusters $D = \{C_i \mid 2 \leq i \leq len\}$

输出: 代表轨迹集合 $C = \{RTR_1, RTR_2, \dots, RTR_{num}\}$

算法:

Let $C = C_1, D = \{D - C_1\}$;

While ($D \neq \emptyset$) do{

Let X be the first cluster in D ;

$D = \{D - X\}$;

MergeCluster(C, X);

}

MergeCluster(C, X) {

While ($X \neq \emptyset$) do{

Let R be the first representative trajectory in X ;

$X = \{X - R\}$;

CompareTrajectory(C, R, ρ);

}

CompareTrajectory(C, R, ρ) {

BeInsert = true;

For each ($Y \in C$) do{

If $\delta_{df}(Y, R) < \rho$ then{

MergeTwoTrajectory(Y, R); //3.2 节中的轨迹合并

算法

BeInsert = false;

Break;

}

If BeInsert = true then

Append P_e to C ;

}

4 实验及分析

4.1 实验数据及运行环境

本文用文献[6]中的两个真实数据集(大西洋飓风数据: <http://weather.unisys.com/hurricane/atlantic/>; 动物运动数据: <http://www.fs.fed.us/>

pnw/starkey/data/tables/)进行实验分析. 实验的软硬件环境: Windows 7, Visual Studio 2008, CPU (CORE 2 DUO 2.0 GHz), 内存 2 GB.

4.2 实验分析

(1) 飓风运动数据集实验

采用的分段参数和聚类密度阈值限制如表1所示, 每次聚类随机的选择这个范围内的值参与实验.

表1 飓风数据多结果融合参数范围

	$d_{\min}$	$d_{\max}$	$\alpha_{\max}$	V_e	θ_e	δ_e	MinLns
nnum	2.5	5	145	25	50	4	24
ximum	4	7.5	155	32	70	6	30

上面这七个参数中 $d_{\min}$ 、 $d_{\max}$ 、 $\alpha_{\max}$ 、 V_e 、 θ_e 、 V_e 是用于多约束的轨迹分段, V_e 、 θ_e 即用于分段, 又用于子轨迹段聚类, δ_e 、MinLns 则完全用于轨迹的子轨迹段聚类. 参数 ρ 的赋值通过式(4)获得. 下面讨论聚类结果融合的效果, 算法的性能.

表2是飓风不同个数子聚类结果融合情况, 当融合次数大于22次后, 得到的聚类结果趋于稳定的结果5个. 图3是飓风数据多聚类融合的结果, 可以看到聚类融合次数越大, 其结果越趋近于稳定. 聚类总时间是最大子聚类器线程运行时间加上多聚类结果融合时间. 多聚类结果融合, 以空间代价换取多次串行聚类的时间的减少, 所以内存使用情况随着聚类融合次数的增加, 内存使用量也逐渐成线性的增加.

表2 飓风不同个数子聚类结果融合情况

聚类数	16	18	20	22	24	26
ρ	4.83	4.97	4.90	5.56	4.82	5.27
time/s	1 887.4	1 787.3	1 809.7	2 361.9	1 976.9	2 183.0
Clutsters	6	7	4	5	5	5

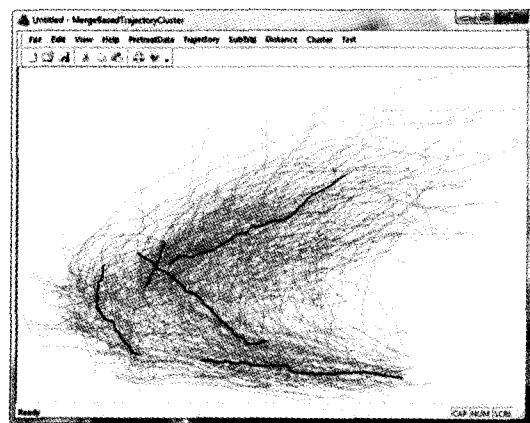


图3 飓风数据多聚类融合结果

(2) 动物运动数据集实验

下面分析两种聚类器的多种参数组合下,多聚类结果融合.以 Deer1995 为实验数据集.

表 3 中 $d_{\min}$ 、 $d_{\max}$ 、 $\alpha_{\max}$ 是用于多约束的轨迹分段, k (整数) 用于环境相关聚类的第一步聚类, ϵ 和 MinLns 用于两种聚类器里的子轨迹段密度聚类的阈值和密度参数. 其取值范围通过对子聚类器的实验发现. 这里分析多聚类结果融合的效果.

表 3 Deer1995 数据多结果融合参数范围

	$d_{\min}$	$d_{\max}$	$\alpha_{\max}$	K	ϵ	MinLns
Minmum	0.4	0.7	145	2	0.5	6
Maximum	0.6	0.8	155	4	0.7	10

由表 4 可知,当融合次数大于 40 次后,得到的聚类结果趋于稳定的结果 5 个. 图 4 是本次实验众多聚类结果中的一种,聚类融合次数越大,其结果越趋近于稳定. 图 5 是 Elk93 数据多次融合后较稳定的一个多聚类融合结果图.

表 4 Deer95 不同个数子聚类结果融合情况

聚类数	20	24	28	32	36	40	44
ρ	0.672	0.575	0.521	0.627	0.695	0.641	0.638
Clusters	13	15	10	8	7	6	5

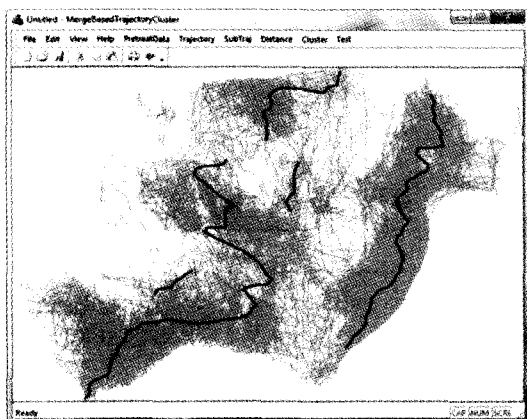


图 4 Deer95 数据多聚类器多参数组合融合结果

5 结束语

本文提出了一种基于多聚类结果融合的轨迹聚类方法 MRMTC,使得聚类结果更为稳定有效. 该方法采用基于多约束的轨迹聚类和基于环境相关的轨迹聚类器同时使用多种参数组合,对聚类结果进行了融合,得到最终的轨迹聚类结果. 实验表明基于融合的轨迹聚类方法,可以获得比单一聚类更有效

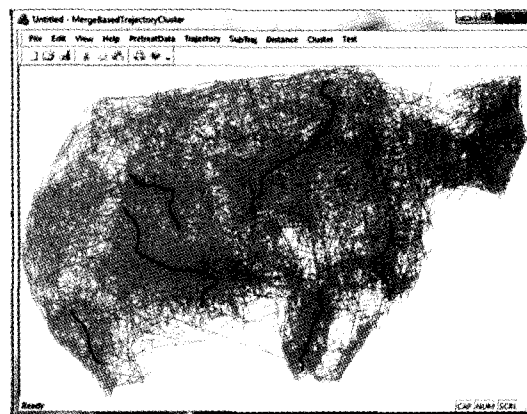


图 5 Elk93 数据多聚类器多参数组合融合结果

更稳定的聚类结果.

参考文献:

- [1] Yanagisawa Y, Saoh T. Clustering multidimensional trajectories based on shape and velocity[C]// IEEE, 2006. Maryland: Baltimore, 2006.
- [2] Zhenhui Li, Jae-Gil Lee, Xiaolei Li, et al. Incremental clustering for trajectories[C]// DASFAA 2010. Japan, Tsukuba, 2010.
- [3] Luis Patino, Francois Bremond, Monique Thonnat. Incremental learning on trajectory clustering[J]. Intelligent Paradigms in Safety and Security, Springer-Verlag, 2010(1): 47-70.
- [4] 阳琳赞, 王文渊. 聚类融合方法综述[J]. 计算机应用研究, 2005(12): 8-10.
- [5] Strehl A, Ghosh J. Cluster ensembles: a knowledge reuse framework for combining multiple partitions [J]. Journal of Machine Learning Research, 2003, 3(3): 583-617.
- [6] Jae-Gil Lee, Jiawei Han, Kyu-Young Whang. Trajectory clustering: a partition- and -group framework [C]// 2007 ACM SIGMOD international conference on Management of data, China: Beijing, 2007: 593-604.

作者简介:

- 李 静 女, (1987-), 硕士研究生. 研究方向为轨迹数据挖掘.
- 张 磊 男, (1977-), 副教授. 研究方向为语义与本体、轨迹数据挖掘.
- 韩陈寿 男, (1984-), 硕士研究生. 研究方向为轨迹数据挖掘.