基于数据挖掘的GIS在车辆自动导航系统中的应用

周 戈[1] 王蔚韬[1] 何光辉[2]

(重庆大学计算机学院 重庆400044)[1] (重庆大学数理学院 重庆400044)[2]

摘 要 车辆自动导航系统是智能运输系统的一个重要组成部分。本文主要介绍了一种数据挖掘的复合聚类分析算法及其在自动导航系统线路设计方面的应用。

关键词 数据挖掘,地理信息系统,车辆导航系统

A New Application of GIS Based on Data Mining in Vehicle Navigation System

ZHOU Ge[1] WANG Wei-Tao[1] HE Guang-Hui[2]

(College of Computer,Chongqing University,Chongqing 400044)[1]

(College of Mathematicai and Physicai Science,Chongqing University,Chongqing 400044)[2]

Abstract The vehicle navigation system is one of the important parts of the intelligent transportation system. The paper gives a multiplex clustering algorithm of data digging,and describes a new application in the design of the route with the geographical information system.

Keywords Data digging,Geographical information system,Vehicle navigation system

数据挖掘(Data Mining 简称 DM)是20世纪末刚刚兴起的数据智能分析技术,它可以从数据库或数据仓库,以及其他各种大量数据类型中,自动抽取或发现有用的模式知识,DM作为一个新兴的多学科交叉应用领域,正在许多行业的决策支持活动中扮演重要的角色。目前有许多种DM方法,聚类分析把每个分类对象称为样品,并根据对象的性质和分类的目的选定若干指标(变量),对每一个样品测出所有的指标值,将得到的结果列一个数据矩阵,这个资料阵是聚类分析的出发点[1,2]。

数据挖掘的定义很多,包含的范围也很广,其中也包含着很多差别细微的地方,但总的来说其核心定义是:数据挖掘是一个从大量的数据中,抽取出潜在的、有价值的知识(模型或规则)的过程[3]。

从商业的角度讲,数据挖掘是一种新的商业信息处理技术[4]。其主要特点是对商业数据中的大量业务数据进行抽取、转换、分析和其他模型化处理,从中提取扶助商业决策的关键性数据,为企业经营决策、市场策划提供依据[5]。

K均值聚类算法是最常用和最知名的划分方法之一,首先从n个数据对象任意选择k个对象作为初始聚类中心,而对于所剩下的其他对象,则根据它们与这些聚类中心的相似度(距离),分别分配给与其最相似的(聚类中心所代表的)聚类;然后再计算每个所获新聚类的聚类中心(该聚类中心中所有对象的均值),不断重复这一过程直到标准测度函数开始收敛为止。这种算法使得各聚类本身尽可能紧凑,而各聚类之间尽可能地分开。但它不适合用于发现非凸性状的聚类,或具有各种不同大小的聚类,对异常数据也很敏感。然而,基于密度的聚类方法却能够帮助发现具有任意形状的聚类,但它仍然需要用户负责设置可帮助发现有效聚类的参数[6]。

在本文里,我们提出了一种复合聚类算法,将K均值算法的思想与基于密度的方法相融合,它把定义在欧氏空间的K均值聚类分析算法,推广到非欧氏空间,扩大了应用范围,同时能够获得更精确的聚类效果。进一步研究将这种算法用于地理信息系统(GIS)利用GIS提供的车辆数据,进行车辆全球定位线路的自动设计。分析结果表明,该方法在GIS数据挖掘中具有重要意义。

1 GIS的概念

1.1 GIS的定义[7]

GIS即地理信息系统,是在计算机硬、软件系统支持下,对现实世界(资源与环境)各类空间数据及描述这些空间数据特性的属性进行采集、储存、管理、运算、分析、显示和描述的技术系统,它作为集计算机科学、地理学、测绘遥感学、环境科学、城市科学、空间科学、信息科学和管理科学为一体的新兴边缘学科而迅速地兴起和发展起来。地理信息系统中"地理"的概念并非指地理学,而是广义地指地理坐标参照系统中的坐标数据、属性数据以及以此为基础而演绎出来的知识。

1.2 GIS的特点

为了满足GIS对地球表面、空中和地下若干要素空间分布和相互关系的研究,GIS必须具备以下基本特点。

①公共的地理定位基础

所有的地理要素,要按经纬度或者特有的坐标系统进行严格的空间定位,才能使具有时序性、多维性、区域性特征的空间要素进行复合和分解,将隐含其中的信息变为显示表达,形成空间和时间上连续分布的综合信息基础,支持空间问题的处理与决策。

②标准化和数字化

将多信息源的空间数据和统计数据进行分级、分类、规格化和标准化,使其适应于计算机输入和输出的要求,便于进行社会经济和自然资源、环境要素之间的对比和相关分析。

③多维结构

在二维空间编码基础上,实现多专题的第三维信息结构的组合,并按时间序列延续,从而使它具有信息存贮、更新和

转换能力，为决策部门提供实时显示和多层次分析的方便。这显然是常规二维或二维半的地形图所不具备的。

④具有丰富的信息

GIS 数据库中不仅包含丰富的地理信息，还包含与地理信息有关的其它信息，如人口分布、环境污染、区域经济情况、交通情况等。纽约市曾经对其数据库进行了调查，发现有80%以上的信息为地理信息或与地理信息有关。

1.3 GIS 与其它系统的区别

GIS 有别于 DBMS(数据库管理系统)。GIS 具有以某种选定的方式对空间数据进行解释和判断的能力，而不是简单的数据管理，这种能力使用户能得到关于数据的知识，GIS 是能对空间数据进行分析的 DBMS，GIS 必须包含 DBMS。

GIS 有别于 MIS(管理信息系统)。GIS 要对图形数据和属性数据库共同管理、分析和应用。MIS 则只有属性数据库的管理，即使存贮了图形，也是以文件形式管理，图形要素不能分解、查询，没有拓扑关系。管理地图和地理信息的 MIS 不一定就是 GIS，MIS 在概念上更接近 DBMS。

GIS 有别于地图数据库。地图数据库仅仅是将数字地图有组织地存放起来，不注重分析和查询，不可能去综合图形数据和属性数据进行深层次的空间分析和提供辅助决策的信息，它只是 GIS 的一个数据源。

GIS 有别于 CAD 系统。二者虽然都有参考系统，都能描述图形，但 CAD 系统只处理规则的几何图形，属性库功能弱，更缺乏分析和判断能力。

1.4 GIS 的系统组成

如图1所示，GIS 由5个主要的元素构成：硬件、软件、数据、人员和模型。

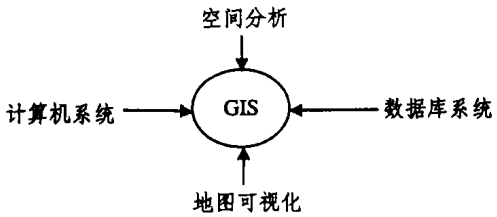

图1 GIS 的组成

1)硬件是 GIS 所操作的计算机。今天，GIS 软件可以在很多类型的硬件上运行。从中央计算机服务器到普通的 PC 机，从单机到网络环境。

2)GIS 软件提供所需的存储、分析和显示地理信息的功能。其要素有：

·地理信息输入和处理的工具；

·数据库管理系统(DBMS)；

·空间查询、分析、可视化表达工具；

·图形用户工具(GUI)。

3)数据。一个 GIS 系统中最重要的部件就是数据，GIS 系统必须建立在准确使用地理数据基础上，数据来源包括从商业组织购买，以及从其他数据的转换。数据类型分为空间数据、属性数据，并与关系数据库相互连接。

4)GIS 人员。GIS 应用的关键是人员的素质，即能掌握 GIS 技术解决现实问题。GIS 的用户范围包括从设计和维护系统的技术专家，到那些使用该系统并完成他们每天工作的人员。

5)模型。GIS 专业模型和经验，是 GIS 应用系统成败的至关重要的因素。

1.5 GIS 的功能

一个 GIS 系统的主要功能包括：①数据输入、存储、编辑；②操作运算；③数据查询、检索；④应用分析；⑤数据显示、结果输出；⑥数据更新。

利用 GIS 应能回答和解决以下5类问题：

·定位(Location)：对象在何处？

·条件(Condition)：即满足一定条件的实体在哪里？

·趋势(Trends)：从何时起发生了哪些变化？

·模式(Patterns)：即在某个地方的空间实体的分布模式。模式分析揭示了地理实体之间的空间关系。

·模型(Modeling)：即某个地方如果具备某种条件会发生什么。通过基于模型的分析实现。

2 复合聚类分析算法

设 $Z=\{z_1,z_2,\cdots,z_n\}\subset R^r$ 为一有限数据集合，n 是数据集中元素的个数，将该数据集中的数据分为 k 类($1<k<n$)，则 Z 将被划分为 k 个 $n*k$ 的分类区域 $U=\{u_j\}\in R^{n\times k}$，$k$ 的确定是根据不同对象的需要。对于每一个小区域 u_j，按照密度计算法选取凝聚点：

$\forall C_1,C_2(C_1<C_2)$为球半径，通常$(C_2=2C_1)$。

$$S(i,j)=|Z(i)-Z(j)| \tag{1}$$

式中：$S(i,j)$为两样本点 $Z(i)$，$Z(j)$之间的距离，

1)If $S(i,j)\leqslant C_1$(样本点落入 C_1球域)，
Then 计算落入 C_1球域的样本点，
And then 选择密度最大的样本点作为第一凝聚点 P_1；
2)对于密度次大的样本点
If $S(P_1,j)\leqslant C_2$(样本点落入 C_2球域)，
Then 忽略此样本点，
If $S(P_1,j)\geqslant C_2$(样本点不落入 C_2球域)，
Then 选择此样本点作为第二凝聚点 P_2；

这样，按照样品密度由大到小一直选下去，每次和已选的任一凝聚点的距离不小于 C_2的样品作为新的凝聚点。

对于以上所求得的凝聚点再求平均密度中心利用重心连接算法公式：

$$D_{ca}(X_{ca},Y_{ca})=\left(\frac{1}{n}\sum_{i=1}^{n}x_i,\frac{1}{n}\sum_{i=1}^{n}y_i\right) \tag{2}$$

其中 x_i,y_i 是样本点 Z_i 的横坐标和纵坐标。

车辆自动导航系统是智能运输系统的重要组成部分。自动导航系统的应用将大幅度提高道路的通行能力，缓解交通拥挤和阻塞。GIS 在车辆自动导航系统中发挥着十分重要的作用。GIS 条件下的电子地图数据库为车辆自动导航系统提供了存放和管理自动导航信息的一个可视化载体。GIS 在车辆自动导航系统中的应用研究将会成为智能运输系统的一个重要的发展方向。

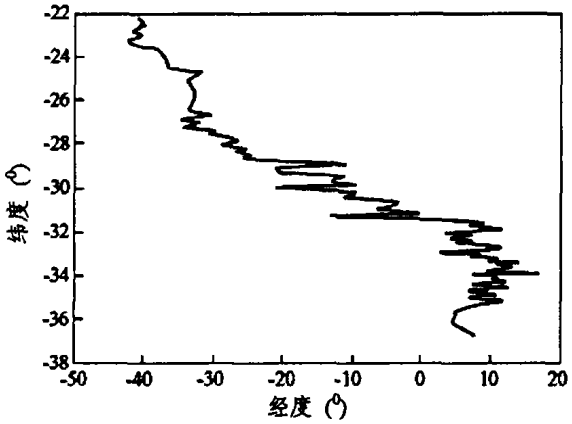

图2 复合聚类方法的聚类结果

(下转第175页)

参 考 文 献

1 Alur R, Holzmann G J, Peled D. An Analyzer for Message Sequence Charts. Software-Concepts and Tools, 1996, 17: 70～77
2 Ben-Abdallah H, Leue S. Expressing and analyzing timing constraints in message sequence chart specifications: [Technical Report 97-04]. Department of Electrical & Computer Engineering, University of Waterloo
3 Ben-Abdallah H, Leue S. Timing Constraints in Message Sequence Chart Specifications. In: Formal Description Techniques X, Proc. of the Tenth Intl. Conf. on Formal Description Techniques FORTE/PSTV'97, Osaka, Japan, Chapman & Hall, 1997
4 Booch G, Rumbaugh J, Jacobson I. The Unified Modeling Language User Guide. Addison-Wesley, 1998
5 France R, Evans A, Lano K, et al. The UML as a formal modeling notation. Computer Standards & Interfaces, 1998, 19: 325～334
6 Seemann J, WvG J. Extension of UML Sequence Diagrams for Real-Time Systems. In: Proc. Intl. UML Workshop, Lecture Notes in Computer Science, Springer, 1998
7 Li Xuandong, Lilius J. Timing Aanlysis of Message Sequence Charts: [TUCS Technical Report 255]. Turku Centre for Computer Science, Finland, March 1999
8 Li Xuandong, Lilius J. Timing Analysis of UML Sequence Diagrams. In: UML 99 - The Unified Modeling Language. Lecture Notes in Computer Science 1723, Springer, 1999. 661～674
9 Rumbaugh J, Jacobson I, Booch G. The Unified Modeling Language Reference Guide. Addison-Wesley, 1998
10 ITU-T Recommendation Z. 120. Message Sequence Chart (MSC). March 1993
11 Muscholl A, Peled D, Su Zhendong. Deciding Properties for Message Sequence Charts
12 Levin V, Peled D. Verification of Message Sequence Charts via Template Matching
13 A Toolset of Requirement Engineering using Message Sequence Charts. Jan. 1998

(上接第146页)

车辆自动导航系统中的一个重要部分就是车辆的行驶路线的设计。图2和图3两幅图片是在两维平面上 GIS 中车辆行驶路线设计的结果。其中图2是采用复合聚类分析方法在 MATLAB 中实现的结果；图3是采用传统的基于密度的方法所得的聚类结果。由这两幅图，我们可以得出结论：复合聚类分析方法比传统的基于密度的方法设计的行车路线要清晰明朗得多。

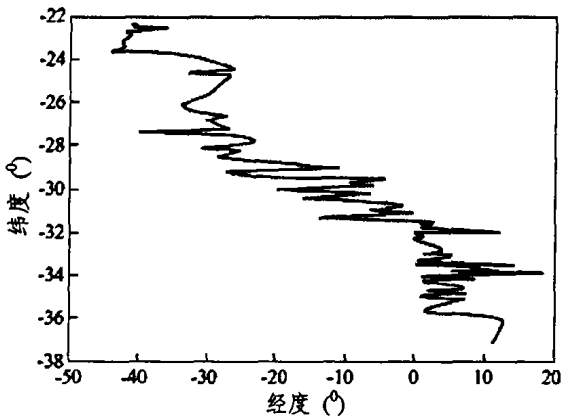

图3 密度方法的聚类结果

3 DM 和 GIS 在车辆自动导航系统中的应用

本文的实例数据来自于 GPS，通过 SQL SERVER 建立专门的数据库，见图4。对数据进行存储、组织和管理。利用 SQL SERVER 建立数据库时，设定相应的规则，保证数据的有效性、安全性和完整性。

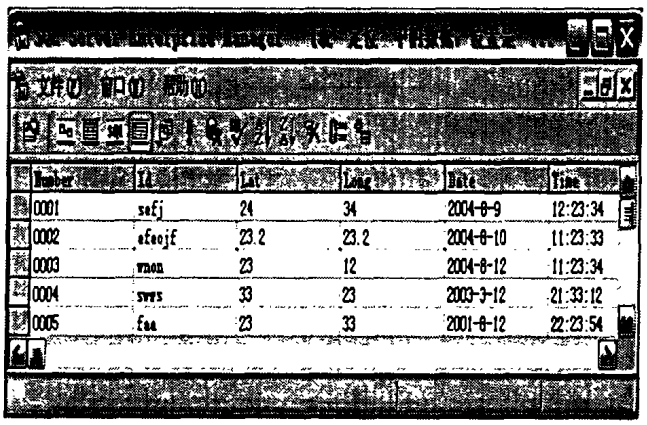

图4 SQL SERVER 建立专门的数据库

这个数据库就包含1张表，表中的每条记录包括5个字段，分别为 Number(编号)、Id(车辆名称)、Lat(纬度坐标)、Long(经度坐标)、Date(日期)、Time(时间)，这样每条记录都完整地记录着某辆车在某一个时刻的确切的位置。所用的地理信息系统的软件为 MapInfo，在 MapInfo 中调用 SQL SERVER 中的数据，并且利用 MapInfo 中提供的工具去除不必要的数据，对空间数据进行处理后，可以将数据库中包含的所有的车辆的位置，描到电子地图的相应地点。由于数据量非常的庞大，因此需要采用 DM 的方法从数据库中提取有价值的信息。我们就以自动导航系统中车辆行驶路线的设计为例来说明一下 DM 是如何应用在 GIS 中的。主要的步骤如下：

1) 从卫星上得到所有加入到自动导航系统中的车辆的分布数据。

2) 选取样本数据，例如，选择任意的两个城市，并导出该区域所包含的数据。

3) 用我们前面所介绍的复合聚类算法进行聚类分析，提取能确定最优路线的数据，最终将所得数据所确定的路线，在地图中以描点的方式显示出来。

因为从 GPS 得来的自动导航系统的车辆分布数据是杂乱无章的、随机的，所以通过测试数据量、样本排序将数据进行预处理，再进行区域分割，经多次试验证明 k 的选取与 n 有关，至少保证每个 μ_j 内样本点数大于10，因为，如果样本点太少，聚类效果不明显；反之，如果样本点数太多的话，设计的路线将会失真。我们取的两个城市之间的样本点为1908，取 $k=120$，试验证明，此时的聚类效果最好。

结论 我们在本文中提出了复合聚类的方法，这种方法在初始的时候设定多个聚类中心，这样的初始中心在数据的空间分布上是很广泛的，具有多样性的特点。这种特点使得最初的聚类基本上能保证每个小区域 μ_i 有一个密度中心，然后再根据适当的准则在小区域 μ_i 找出几个子中心，删除冗余数据，再计算这个区域 μ_i 的平均密度中心，来修正原来的密度中心。事实证明，这种方法在基于 GIS 的自动导航系统中非常有效。

参 考 文 献

1 朱明. 数据挖掘[M]. 合肥：中国科学技术大学出版社，2002
2 Wang X Z. Data Mining and Knowledge Discovery for Process Monitoring and Control[M]. Spring-Verlag London limited, 1999
3 Liu Bing. Knowledge Discovery and Data Mining. 21世纪青年科学论坛，2001, 19(6): 70～74
4 Han Jiawei. Micheline Kamber. Data Mining: Concepts and Techniques. Morgan Kufnamm Publishers, Aug. 2000
5 Business Objects Ltd. Growth in Decision Supports System. Database&Network Jounal, 1998, 28(1): 3～4
6 朱勇华，郜淑彩，孙福玉. 应用数理统计[M]. 武汉：武汉水利电力大学出版社，1999. 362～376
7 龚健雅. 当代 GIS 的若干理论与技术. 武汉测绘科技大学出版社，1999