第 25 卷 第 4 期
2005 年 7 月

长安大学学报(自然科学版)
Journal of Chang'an University(Natural Science Edition)

Vol. 25 No. 4
July 2005

文章编号:1671-8879(2005)04-0066-04

公路隧道交通流的数据挖掘

许宏科,揣锦华,张炤华,樊海玮
(长安大学 信息工程学院,陕西 西安 710064)

摘 要:在阐述数据挖掘技术和方法的基础上,研究了基于统计理论的聚类方法。利用微软 SQL Server 2000 提供的聚类数据挖掘方法对某个公路隧道交通流的数据进行了聚类分析,并对数据结果进行了详细的解析,得到该隧道交通流的一些特性信息,如:何时该隧道交通流最大等。根据这些信息,可以针对不同的交通量特点安排隧道监控设备的控制方案及系统的维护方案,以保障公路隧道交通的畅通和安全。

关键词:交通工程;公路隧道;隧道监控;交通流;数据挖掘;聚类分析
中图分类号:U491.112 **文献标识码**:A

Data mining of traffic flow in road tunnel

XU Hong-ke, CHUAI Jin-hua, ZHANG Zhao-hua, FAN Hai-wei
(School of Information Engineering, Chang'an University, Xi'an 710064, China)

Abstract: On the basis of expounding the technology and method of data mining, this paper mainly studies the cluster method based on the theory of statistics, analyzes the traffic flow data of a road tunnel by the cluster data mining method of Microsoft SQL server 2000, makes detailed analysis of the result and gets some characteristic information of the tunnel traffic flow, such as the time interval of maximum traffic flow. Using this information, the control project of tunnel monitoring devices and the system maintenance project can be arranged reasonablly to quarantee highway tunnel traffic smooth and security. 1 tab, 3 figs, 6 refs.

Key words: traffic engineering; highway tunnel; tunnel monitoring and control; traffic flow; data mining; cluster analysis

0 引 言

数据挖掘是 20 世纪 90 年代新崛起的一门学科。它作为目前国际上数据库和信息决策领域最前沿的研究方向之一,引起了学术界和工业界的广泛关注。其研究目标主要是发展有关的方法论、理论和工具,从大量数据中提取有用的知识和模式。数据挖掘最初应用在商业领域,但随着研究的不断深入,它已经广泛地应用于各个领域。然而,资料表明[1],在公路隧道交通领域,数据挖掘技术的研究和应用却极少。

本文将数据挖掘技术应用到公路隧道交通控制中。通过研究数据挖掘中聚类分析的理论和方法,在 SQL Server 2000 平台上,对某个公路隧道交通流的数据进行聚类分析。对数据挖掘技术在公路隧道交通控制和交通信息处理方面的应用进行探讨。

收稿日期:2004-03-19
基金项目:国家西部交通建设科技项目(2004 318 812 22)
作者简介:许宏科(1963-),男,陕西风翔人,长安大学副教授,博士研究生.

1　数据挖掘技术

数据挖掘(Data Mining)就是从大量不完全的、有噪声的、模糊的、随机的数据中提取隐含在其中潜在而有用的信息和知识的过程。它不仅是面向特定数据库的简单检索查询调用，而且要对这些数据进行微观、中观乃至宏观的统计、分析、综合和推理，并试图发现事件间的相互关联，甚至利用已有的数据对未来的活动进行预测，以指导实际问题的求解。

1.1　知识发现的过程

首先是数据准备，此阶段又可分为：数据集成、数据选择、数据预处理和数据转换。其次是数据挖掘，即利用机器学习、统计分析等方法从数据库中发现有用的模式或知识。最后是结果表达和解释，即根据用户的决策目的，对提取的信息进行分析，将有用的信息区分出来，并且通过决策支持工具提交给决策者。

1.2　数据挖掘的方法和技术

归纳方法，包括信息方法(决策树方法)、集合论方法；仿生物技术方法，包括神经网络方法、遗传算法；公式发现，包括物理定律发现系统 BACON、经验公式发现系统 FDD；统计分析方法；模糊数学方法；可视化技术。在实际应用中，通常将上述几种方法和技术综合使用，以解决问题。

2　聚类分析算法

聚类分析是在机器学习、数据挖掘、统计数据分析、数据压缩、向量量化、图象处理等领域的 1 个重要应用。聚类就是“物以类聚”，即把具有相似特性的数据归结为一类。因此，经过聚类计算后，数据集就变为类集。在同一类集中，数据具有相似特性的变量值，不同类集的变量值之间不具有相似性。根据这一特性，在控制领域可以针对不同类集事例采用不同的控制策略，提高效率，得到好的结果。

作为统计学的 1 个分支，聚类分析已经被广泛地研究，研究主要集中在基于距离的聚类分析。最常用的距离计算公式是 Euclidean distance(欧氏距离)，公式定义为

$$d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\cdots+|x_{ip}-x_{jp}|^2)} \tag{1}$$

其中，$i=(x_{i1},x_{i2},\cdots,x_{ip})$，$j=(x_{j1},x_{j2},\cdots x_{jp})$，分别表示 1 个 p-维数据对象。

另 1 个常用的距离计算公式是 Manhattan distance，它的具体定义为

$$d(i,j)=|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\cdots+|x_{ip}-x_{jp}|^2 \tag{2}$$

式(1)、式(2)都满足距离函数的有关数学性质，即

$d(i,j)\geqslant 0$，则表示对象之间距离为非负数的 1 个数值；$d(i,j)=0$，则表示对象之间距离为 0；$d(i,j)\leqslant d(j,i)$，则表示对象之间距离为对称函数；$d(i,j)\leqslant d(i,h)+d(h,j)$，则表示对象之间距离满足“两边之和不小于第三边”的性质。

如果每个变量都可赋予 1 个权值，以表示所代表属性的重要性，那么带权的 Euclidean distance 计算公式为

$$d(i,j)=\sqrt{(w_1|x_{i1}-x_{j1}|^2+w_2|x_{i2}-x_{j2}|^2+\cdots+w_p|x_{ip}-x_{jp}|^2)} \tag{3}$$

基于 k-means(k-平均值)、k-medoids(k-中心点)和其他一些方法的聚类分析工具已经被加入到许多统计分析软件包或系统中。这里主要介绍在微软 SQL Server 2000 环境下的 k-means 聚类算法。

k-means 以 k 为参数，把 n 个对象分为 k 个簇，使簇内具有较高的相似度，而且簇间的相似度较低。相似度的计算根据 1 个簇中对象的平均值(被看作簇的重心)来进行。

k-means 的处理流程如下：首先随机地选择 k 个对象，且每个对象初始地代表 1 个簇的平均值或中心。对剩余的对象，根据其与各个簇中心的距离将它赋给最近的簇。然后重新计算每个簇的平均值。这个过程不断重复，直到准则函数收敛。通常采用平方误差最小准则，其定义为

$$E=\sum_{i=1}^{k}\sum_{p\in c_i}|p-m_i|^2 \tag{4}$$

其中，E 为数据库中所有对象平方误差的总和；p 为空间中的点，表示给定的数据对象；m_i 为簇 c_i 的平均值(p 和 m_i 都是多维的)。这个准则试图使生成的结果簇尽可能地紧凑和独立。

图 1 给出了 k-means 算法的简单过程。

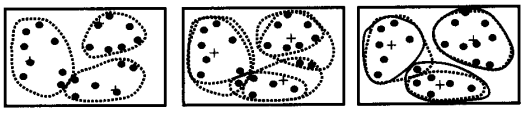

图 1　k-means 算法的过程

k-means 算法如下：

输入：簇的数目 k 和包含 n 个对象的数据库。

输出:满足平方误差最小准则的 k 个簇。

k-means 的处理流程:

(1)任意选择 k 个对象作为初始的簇中心;

(2)循环下述(3)到(4),直到每个簇不再发生变化为止;

(3)根据簇中对象的平均值,计算每个对象与这些中心对象的距离,并根据最小距离将每个对象重新赋予最类似的簇;

(4)更新簇的平均值,即计算每个簇中对象的平均值;

这个算法尝试找到使平方误差函数值最小的 k 个划分。当结果是密集的,而簇与簇之间区别明显时,它的效果较好。如果处理大数据集,该算法是相对可伸缩的和高效率的。因为它的复杂度是 $O(nkt)$,其中,n 是所有对象的数目,k 是簇的数目,t 是迭代的次数。通常,$k \ll n$,且 $t \ll n$。这个算法经常以局部最优结束。

3 公路隧道交通流数据的聚类分析

在微软 SQL Server 2000 环境下,利用系统所提供的微软聚类数据挖掘方法,对杭州绕城高速公路黄鹤山隧道交通流数据进行数据挖掘分析。

3.1 数据描述

黄鹤山隧道是杭州绕城高速公路北段的重要组成部分。隧道为双向 6 车道,呈东西走向。通过对所有数据的分析,决定对黄鹤山隧道右洞 vd2_tatolf=>400 的车流数据进行分析,这些数据共有 368 条。另外,为了使得到的结果更容易分析,将时间信息分为两部分,一部分为年、月、日;另一部分为小时。

3.2 聚类分析

微软聚类采用的是 k-means 算法,它要求用户预先提供聚类的个数,即 k 值,系统默认 $k=10$。当选择 $k=5$,vd2_totalf>=400 时,得到的聚类结果如图 2 所示。其中,根节点是用来训练模型中所有记录的聚类的,每 1 个聚类都有 1 组描述该聚类中记录的规则。聚类编号的顺序是说明聚类中记录数量的另外 1 个重要的特征。聚类结果按照降序进行编号和排列,即 Cluster1 中包含的记录最多,Cluster2 次之,最后底层聚类中包含的记录最少。聚类中包含的记录越多,该聚类中数据的共通性越显著,其可信度也越高;相反,底层聚类可以非常显著地分辨出数据中的异常值,这对孤立点的分析很有用。

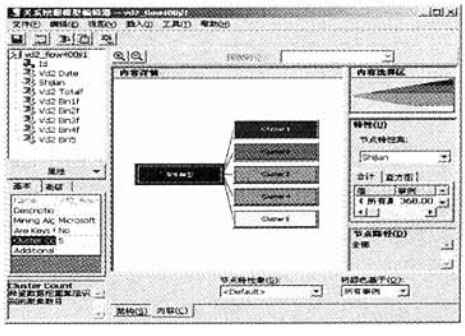

图 2 聚类图

将 5 个簇的数据结果汇总后得到表 1。

表 1 聚类数据表($k=5$)

小时信息	全部		Cluster1		Cluster2		Cluster3		Cluster4		Cluster5	
	事例	可能性/%	事例	可能性/%	事例	可能性/%	事例	可能性/%	事例	可能性/%	事例	可能性/%
10:00	25	6.79	3.94	3.65	7.60	7.35	0.20	0.26	3.85	9.28	9.41	23.34
11:00	45	12.23	22.97	21.27	7.59	7.34	0.02	0.03	11.98	28.89	2.44	6.04
12:00	26	7.07	6.30	5.84	6.01	5.81	1.22	1.63	6.43	15.51	6.04	14.98
13:00	11	2.99	1.34	1.24	2.20	2.13	0.64	0.85	0.96	2.31	5.86	14.53
14:00	34	9.24	14.57	13.50	7.70	7.44	0.25	0.34	10.84	26.13	0.64	1.59
15:00	53	14.40	22.01	20.38	17.77	17.18	8.00	10.70	1.60	3.85	3.62	8.98
16:00	64	17.39	18.21	16.87	26.19	25.32	11.47	15.33	3.20	7.73	4.93	12.21
17:00	64	17.39	16.87	15.62	19.34	18.70	27.09	36.22	0.12	0.30	0.58	1.44
18:00	37	10.05	1.73	1.60	9.04	8.74	23.85	31.89	0.44	1.07	1.94	4.80
19:00	2	0.54	0.03	0.03	0.00	0.00	1.05	1.40	0.04	0.11	0.88	2.17
20:00	2	0.54	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	4.95
21:00	2	0.54	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	4.96
8:00	1	0.27	0.00	0.00	0.00	0.00	0.00	0.00	1.00	2.41	0.00	0.00
9:00	1	0.27	0.00	0.00	0.00	0.00	0.00	0.00	1.00	2.41	0.00	0.00
缺少	1	0.27	0.00	0.00	0.00	0.00	1.00	1.34	0.00	0.00	0.00	0.00

由节点路径和数据表，可以对 5 个簇逐一地进行分析。Cluster1 中包含的信息为：在 9：00 时，轿车、中小型车、加长车有比较大的流量；在 15：00～17：00，总车流量有 1 个高峰。Cluster2 中包含的信息为：轿车、中小型车的流量较高；在 15：00～17：00，总车流量有 1 个高峰。Cluster3 中包含的信息为：在 15：00～18：00，特别是在 18：00，总车流量有 1 个高峰。Cluster4 中包含的信息为：在 18：00～19：00，特长型车流量较高。

另 1 个模型为 $k=5$，聚类特性选择为小时信息，得到的聚类结果如图 3 所示。

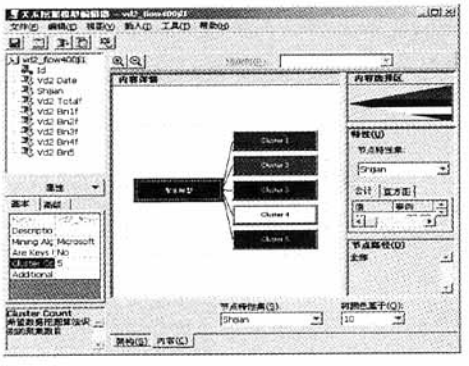

图 3　时间聚类图

利用与上个模型相似的分析方法可得到：数据量最大的是 Cluster1，其对应的时间范围是 10<time<17。这表示 10：00～17：00，车流量最大；数据密度最大的为 Cluster3 和 Cluster5，分别对应的是：Cluster3，time=17；Cluster5，time=10。也就是说 17：00 和 10：00 时，数据密度最大。需要说明的是颜色深表示数据密度量大，而不一定是数据量大，这和 k-平均算法的特点有关，即算法的聚类半径大小决定了 1 个 Cluster 中数据的多少。

通过上述聚类分析，可以得到以下结论：

(1)一般在 10：00 和 17：00 时隧道内交通密度最大；

(2) 从 9：00～17：00 是全天中车流量最大的时间段。

4　结　语

(1)通过对隧道交通流数据的聚类分析，得到了针对某公路隧道交通流的一些特性信息。

(2)根据隧道交通流的特性信息，可以针对不同的交通量特点，安排对隧道监控设备的控制方案及系统的维护方案，以提高工作效率，减少对正常交通的影响。例如，根据交通量在不同时间的分布，安排人员值班及对设备的检修和更换时间；根据交通量确定隧道内排风机的开启时间等。

(3)根据得到的交通流分布结果，可以制定相应的公路隧道交通控制方案，以保障隧道交通的畅通和安全。

参考文献：

References：

[1] 宋颖华. 高速公路隧道监控系统的方案设计[J]. 东北公路，2000，23(3)：80－83.
SONG Ying-hua. A design plan of monitor and control system of expressway tunnel[J]. Northeast Highway，2000，23(3)：80－83.

[2] 朱　明. 数据挖掘[M]. 合肥：中国科学技术大学出版社，2002.
ZHU Ming. Data Mining[M]. Hefei：China Science Technology Press，2002.

[3] 范　明，孟小峰. 数据挖掘概念与技术[M]. 北京：机械工业出版社，2001.
FAN Ming，MENG Xiao-feng. The concept and technology of data mining[M]. Beijing：Machine Industy Press，2001.

[4] 王　颖，谢海红. 基于路网规划的道路立体交叉交通量预测方法[J]. 交通运输工程学报，2003，3(3)：106－109.
WANG Ying，XIE Hai-hong. Traffic volume forecasting method at road interchange based on road net work planning[J]. Journal of Traffic and Transportation Engineering，2003，3(3)：106－109.

[5] 徐丽群，季建华. 模糊逻辑推理在消除交通流诱导负效应中的作用[J]. 中国公路学报，2004，17(3)：78－81.
XU Li-qun，JI Jian-hua. Appliation of fuzzy logic reasoning to elimination of negative effects for traffic flow guidance[J]. China Journal of Highway and Transport，2004，17(3)：78－81.

[6] 刘伟铭. 高速公路系统控制方法[M]. 北京：人民交通出版社，1998.
LIU Wei-ming. The control method of highway system[M]. Beijing：People's Communications Press，1998.