文章编号:1001 - 9081(2005)04 - 0937 - 03

交通流时间序列分离方法

任江涛,谢琼琼,印 鉴
(中山大学 计算机科学系,广东 广州 510275)
(renjt@ yahoo. com)

摘 要:采用聚类分析方法对交通流时间序列进行分析可以发现典型的交通流变化模式。通常可采用欧式距离及 K 均值算法进行时间序列聚类,但经分析发现单凭此方法还难以实现不同变化趋势的交通流时间序列的有效分离。针对此问题,提出了将动态时间弯曲及灰色关联度引入交通流时间序列相似性度量,且结合层次化聚类方法对交通流时间序列进一步分离的方法。通过实验研究,发现基于灰色关联度的层次化聚类方法能较好地实现交通流时间序列的进一步有效分离。

关键词:交通流;时间序列;分离
中图分类号: TP274⁺.2 **文献标识码**:A

Traffic flow time series separation methods

REN Jiang-tao, XIE Qiong-qiong, YIN Jian
(*Department of Computer Science, Zhongshan University, Guanghou Guangdong 510275, China*)

Abstract: By clustering of traffic flow time series, the typical traffic fluctuation patterns can be found. Generally, the euclidean distance and K-means algorithm can be used to clustering the time series, but it is hard to separate the time series with great different variability well. To solve this problem, fluctuation similarity measure, such as dynamic time warping and gray relation grade, and the hierarchical clustering algorithm were used to further separate the traffic flow time series. The experiments show that the proposed method can work and the gray relation grade measure is better suited for the problem than the dynamic time warping measure.

Key words: traffic flow; time series; separation

0 引言

将具有相似的交通流变化趋势的检测站进行聚类,一方面可以发现典型的交通流变化趋势规律,另一方面可以实现对具有不同流量特性的交通检测点所在路段进行合理分组,使得组内的路段具有相对类似的交通流特性,不同组之间的路段其交通流量具有区别明显的变化规律及特性,进而结合空间信息可以发现一些有意义的交通流时空分布规律。另外每组内的路段形成各个相对独立的特征区域,作为进一步进行交通规划及控制优化的依据之一。目前对这方面的研究通常采用基于欧氏距离 K 均值算法,但若仅采用 K 均值算法及欧氏距离度量还不能很好地满足实现交通流时间序列有效分离的要求,其结果只反映了不同检测点的交通流在流量水平上的相似性,而在变化趋势上的相似性还体现不足。因此有必要研究对这些流量水平相似的交通流时间序列基于变化趋势相似度进行进一步分离的方法。

实际上,相似性度量问题是时间序列聚类中的最基本的问题之一。典型的时间序列聚类的相似性度量包括欧氏距离、相关系数、动态时间弯曲(Dynamic Time Warping, DTW)度、灰色关联度等。在本研究中,采用了这两种相似性度量来研究交通流时间序列的进一步分离问题,并比较了两种度量方法在此应用中的优劣。

1 动态时间弯曲度

动态时间弯曲度(Dynamic Time Warping,DTW)最初是应用于文本及字符串匹配及视觉模式识别等领域的相似性度量方法,研究表明这种基于非线性弯曲技术的算法可以获得较高的识别及匹配精度。其具体定义如下:

设现有二时间序列 Q 和 C,其数据长度分别为 n 和 m,有:

$Q = q_1, q_2, \cdots, q_n$

$C = c_1, c_2, \cdots, c_m$

为了利用 DTW 将两个时间序列对准,事先给出两个定义:距离相异矩阵和弯曲路径。

定义 1 n 行 m 列矩阵,矩阵中的元素为不同时间序列数据对象之间的点的欧几里的距离,即 $d(q_i, c_j) = (q_i - c_j)^2$,为距离矩阵,记为 ***D_matrix***。

矩阵中的 $d(q_i, c_j)$ 是二个时间序列数据点之间的距离值,可以看作是对象 q 和对象 c 直接的相异性的量化表示。当对象 q 和 c 越相似或越接近,其值越接近 0;两个对象越不相同,其值越大。将两个时间序列分别置于二维坐标的两轴,如图 1 所示。

定义 2 在两个不同时间序列间的距离矩阵中,定义时

收稿日期:2004 - 09 - 27;**修订日期**:2004 - 12 - 08
基金项目:国家自然科学基金资助项目(60374059);广东省自然科学基金资助项目(04300462)
作者简介:任江涛(1975 -),男,广西柳州人,讲师,博士,主要研究方向:数据挖掘与知识发现、智能交通; 谢琼琼(1981 -),女,广东河源人,硕士研究生,主要研究方向:数据挖掘、信息处理和电子商务; 印鉴(1968 -),男,湖北武汉人,教授,博士,主要研究方向:人工智能、数据挖掘与数据仓库.

间序列间相异性关系的一组连续的矩阵元素的集合，称为弯曲路径。

$$W = w_1, w_2, \cdots, w_k, \cdots, w_K, \max(m,n) \leqslant K \leqslant m+n-1$$

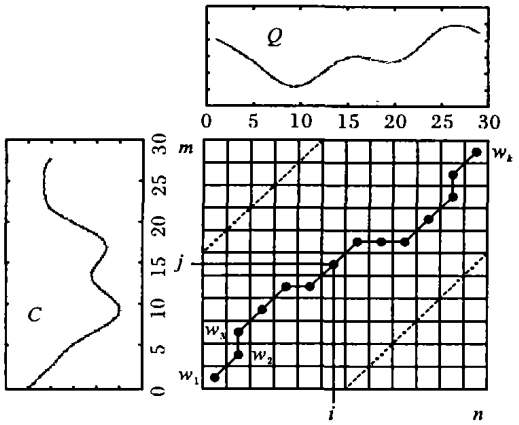


图 1 动态时间弯曲示例

对距离矩阵分析可知，弯曲路径存在多解，但是人们关心的实际上仅仅是弯曲总长度最小的路径，在逻辑意义上，两数据相似性程度最大（距离值最小）的作为相似搜索的判据，如下式：

$$DTW(Q,C) = \min[(\sum_{k=1}^{K} w_k)/K]$$

式中分母 K 是为了在比较不同长度的路径时有统一的标准。

理论上，可以利用穷举搜索法寻找满足条件的弯曲路径，但是完全穷举在大型数据库分析中往往是不切实际的，因为路径的解很多，且与距离矩阵中的元素数成指数关系。由动态规划理论可知，设点 (i,j) 在最佳路径上，那么从点 $(1,1)$ 到 (i,j) 的子路径也是局部最优解，也就是说从点 $(1,1)$ 到点 (m,n) 的最佳路径可以由时间起始点 $(1,1)$ 到终点 (m,n) 之间的局部最优解通过递归搜索获得。即：

$$S_{1,1} = d(q_1, c_1)$$

$$S = d(q_i, c_j) + \min\{S(i-1,j); S(i-1,j-1); S(i,j-1)\}$$

最终时间序列弯曲路径最小累加值为 $S_{m,n}$，从 $S_{m,n}$ 起沿弯曲路径按最小累加值倒退直到起始点 $S_{1,1}$ 即可找到整个弯曲路径。

2 灰色关联度

灰色关联度的概念来源于灰色系统理论，力图寻求系统中各子系统（或因素）之间的数值关系。简言之，若两个因素变化的态势是一致的，即同步变化程度较高，则可以认为两者关联性较强，即灰色关联度较大；反之，则两者灰色关联度较小。因此，灰色关联分析依据各因素时间序列曲线形状的相似程度做发展态势的分析，为一个系统内各因素发展变化态势的同步性提供了量化的度量。

设有 $m+1$ 个时间序列 x_0、x_i、x_m，其中，x_0 称为母序列，又称作参考序列，$x_1, x_2, \cdots, x_m$ 称为子序列，即比较序列。如图 2 所示的 4 个时间序列，其中 x_0 为参考序列，x_1, x_2, x_3 为比较序列。由图中各序列所呈现出的几何相似程度，可判断各序列间的关联度。可明显看出，在三个比较序列当中，序列 x_1 相较于 x_2, x_3，与 x_0 的折线形状较为类似，因此，x_1 与 x_0 的灰色关联度较高。

由文献[5]的定义，若 X 为灰色关联因子空间，即：

$$X = \{x_i \mid i \in I = \{0,1,2,\cdots,m\}, m \geqslant 2, x_i = (x_i(1), x_i(2), \cdots, x_i(n)), n \geqslant 3\}$$

则子序列 x_i 之于母序列 x_0 在 k 点的灰关联系数为：

$$r(x_0(k), x_i(k)) = \frac{\min_j \min_l |x_0(l) - x_j(l)| + \sigma}{|x_0(k) - x_i(k)| + \sigma}$$

其中，σ 为一指定之正实数，若其定义如下：

$$\sigma = \zeta * \max_j \max_k |x_0(k) - x_j(k)|$$

其中 ζ 为分辨系数，$\zeta \in (0,1)$，一般取值 0.5。

根据以上定义，灰色关联系数之量化模型如下所示：

$$r(x_0(k), x_i(k)) = \frac{\min_j \min_l |x_0(l) - x_j(l)| + \zeta * \max_j \max_l |x_0(l) - x_j(l)|}{|x_0(k) - x_i(k)| + \zeta * \max_j \max_l |x_0(l) - x_j(l)|}$$

当计算出灰色关联系数后，各因子序列之间的灰色关联度便可通过计算灰关联系数的平均值得到，如下式所示：

$$r(x_0, x_i) = \frac{1}{n} \sum_{k=1}^{n} r(x_0(k), x_i(k))$$

设有 $m+1$ 个时间序列 $x_0, x_1, x_2, \cdots, x_m$，分别以各个时间序列为参考序列，则可以求得其两两不同时间序列间的灰色关联度为 $r(x_i, x_j), i \neq j$；当 $i = j$ 时 $r(x_i, x_j) = 1$。

因为 $r(x_i, x_j) \neq r(x_j, x_i)$，所以取它们的均值作为其相似度，即 $r_{ij} = r_{ji} = \frac{r(x_i, x_j) + r(x_j, x_i)}{2}$。采用上式计算两两样本间的灰色关联度，即得到的灰色关联度相似度矩阵。

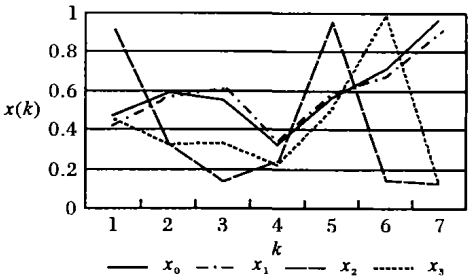


图 2 灰色关联度例图

3 时间序列分离方法

为解决引言中所提出的对经过 K 均值聚类所得到的聚类结果进一步分离的问题，可基于一定上述相似性度量，即动态时间弯曲度或灰色关联度，尝试应用层次化聚类算法实现基于变化趋势相似的聚类，从而实现交通流时间序列的分离。

层次化聚类作为一种被广泛采用的聚类策略应用于聚类领域并不断得到发展，层次化聚类算法将样本集中的样本进行层次分解，形成基于树结构的聚类结果。根据层次分解是自底向上还是自顶向下，层次化聚类算法分为两类：1）凝聚式层次聚类：采用自底向上的策略首先将每个样本视为一类，然后逐步选择最为相似的类进一步合并为更大的类，依次向上直到所有的对象都在一个类中，或者某个终止条件被满足；2）分裂式层次聚类：采用自顶向下的策略，首先将所有样本视为一类，然后根据一种划分原则将该类分裂为若干个较小的类，依次继续分裂直至每个样本自成一类，或达到某一终止条件后停止分裂。在本研究中，采用凝聚式层次聚类方法实现不同变化趋势的时间序列的分离。具体方法如下：

(1)利用相似性度量,如动态时间弯曲或灰色相似度,计算类的相似度矩阵;

(2)基于相似度矩阵,应用层次化聚类算法对时间序列样本进行聚类,分离出变化趋势相似的时间序列;

(3)若分离结果的全部或部分类效果不理想,重复步骤(2),直至得到理想的分离结果。

4　实验研究

为比较分析前述两种反映时间序列变化趋势相似性的相似性度量在进行交通流时间序列分离中的效果,并检验所提出的分离方法是否可行,采用高速公路道路占用率时间序列数据进行了实验研究。数据来源于美国华盛顿州的高速公路监控系统,选择了路网中的 158 个检测点在 2004 年 2 月 23 日的交通流时间序列数据。首先采用 K 均值算法对这些样本进行聚类,得到其中一类如图 3 所示,从图中可以看出,此类明显包含了若干类变化趋势不同的交通流时间序列,有必要采取措施进行进一步分离。

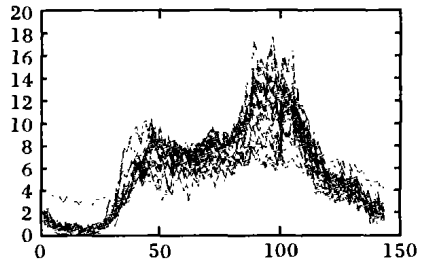
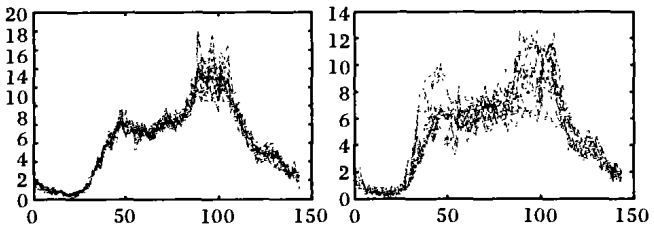


图3　经过 k-means 聚类得到的一类

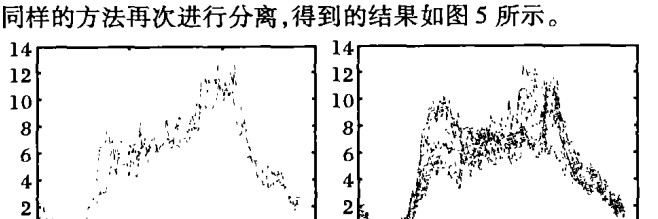
用动态时间弯曲法计算相似度,采用层次化聚类算法将图 1 中的时间序列样本进行分离,得到的结果如图 4 所示。



(a) 第1类　　(b) 第2类

图4　采用动态时间弯曲相似度分离得到的结果

从图 4 可看出基于动态时间弯曲相似度,能将原来的样本进行分离,得到了类内相似度较为明显的第 1 类,同时也发现第 2 类实际上包含着两类趋势的曲线,尝试针对第 2 类用同样的方法再次进行分离,得到的结果如图 5 所示。

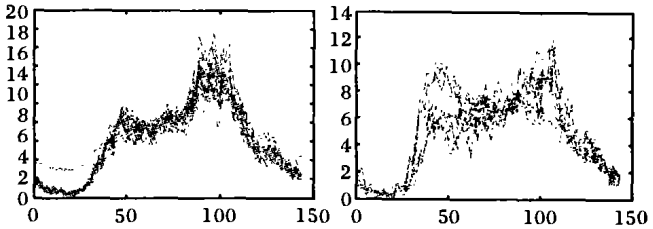


(a) 第1类　　(b) 第2类

图5　采用动态时间弯曲相似度进行二次分离得到的结果

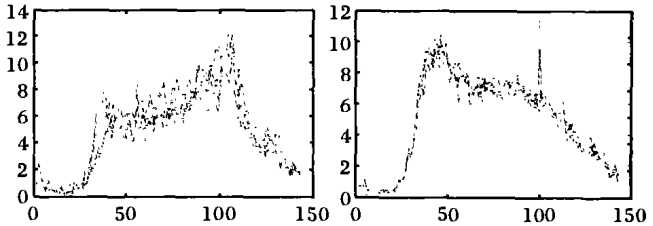
由图 5 可以看出,采用动态时间弯曲相似度进行二次分离的效果不是太理想,接着尝试进行基于灰色关联度进行上述实验。首先,第一次分离的结果如图 6 所示。

观察发现图 6(a)所示的一类样本变化趋势较为一致,因此不需再次进行分离;而图 6(b)中样本则呈现出两类明显不同的变化趋势,因而再次建立灰色关联度矩阵并进行层次化聚类,得到的聚类结果如图 7 所示。



(a) 第1类　　(b) 第2类

图6　采用灰色关联相似度分离得到的结果



(a) 第1类　　(b) 第2类

图7　采用灰色关联相似度进行二次分离得到的结果

从图 7 可以看出,经过再次分离后,所分离出的两类在类内都各自具有较好的类内相似度,即每类都反映了一种交通流时间序列变化趋势。从实验的结果来看,针对本文所采用的交通流时间序列数据,在进行基于变化趋势的分离中,灰色关联度作为相似性度量的结果要优于动态时间弯曲作为相似性度量时的结果,同时也说明了反映时间序列变化趋势相似性的相似性度量与层次化聚类算法相结合实现交通流时间序列分离的可行性。

5　结语

本文研究了交通流时间序列的基于变化趋势的进一步分离问题,对动态时间弯曲及灰色关联度这两类时间序列变化趋势相似性度量方法进行了比较研究,并提出了基于这些变化趋势相似性及层次化聚类算法的交通流时间序列分离方法,通过比较实验分析说明了此方法的可行性。

参考文献:

[1] SHEKHAR S, LU CT, CHAWLA ST, *et al*. Data Mining and Visualization of Twin-Cities Traffic Data[R]. Department of Computer Science Technical Report TR 01 -015, University of Minnesota, 2001.

[2] 翁颖钧,朱仲英. 基于动态时间弯曲的时序数据聚类算法的研究[J]. 计算机仿真,2004,21(3):37 -40.

[3] (加)韩家炜. 数据挖掘:概念与技术[M]. 范明等,译. 北京:机械工业出版社,2001.

[4] HAND D. 数据挖掘原理[M]. 张银奎,译. 北京:机械工业出版社,2003.

[5] 邓聚龙. 灰色系统理论教程[M]. 武汉:华中理工大学出版社,1992.

[6] 徐中祥,吴国平,周新良. 金矿灰色关联分析预测法[J]. 中国地质大学学报,1994,19(1):87 -93.