系统工程导论

开课单位：清华大学自动化系

主讲教师：胡坚明 副教授

模块二：系统分析

聚类分析方法

■本章内容

- 聚类问题的一般性描述
- K-均值聚类方法
- 系统聚类方法
- 动态聚类方法
- 基于自组织映射（SOM）的聚类方法
- 应用实例

■引言

物以类聚，人以群分！

聚类问题出现在我们社会生活的方方面面

- 世界——发达国家/发展中国家 （GDP）
- 中国——东部/中部/西部 （经济状况）
- 大学——重点大学/一般院校 （综合实力）
- 专业——文、理、工、医等 （学习内容）
- 学生——擅长数理、擅长推理等 （学习成绩）
- .....

集合　　子集　　测度

■引言

- 聚类问题实际上是将包含若干元素的集合，按照某种测度，划分成若干子类。
- 测度是指定义在每个类上的函数，我们的目标就是使其达到最大或最小。
- 聚类问题的本质是划分问题——属于NP难问题

■引言

举例——推研面试考核

- 自动化系每年大约有140个毕业生，每名学生推研前完成了45门课程的学习。不同老师如何根据这些课程成绩，考察其所关心的能力？
- 这实际上构成了一个140×45的矩阵，不同老师可以根据不同目的构造一个函数（评价指标、测度）、使得出现不同分类。

■条件

给定一组对象，用以下指标集表示

$$J(M)=\{1,2,\cdots,M\}$$

给定一个对 $J(M)$ 的所有非空子集有定义的实值函数

$$\rho(\varpi)\quad \forall \varpi \subseteq J(M),\ \varpi \neq \varphi$$

其中 φ 表示空集

■任务

确定 $\Omega=\{\varpi_i \subseteq J(M), 1\le i\le k\}$ 满足

$$\bigcup_{1\le i\le k}\varpi_i = J(M),\quad \varpi_i\cap\varpi_j=\varphi\ \ \forall i\neq j$$

使下述目标函数最小(或最大)

$$\sum_{i=1}^{k}\rho(\varpi_i)$$

k 一般未知，要根据最优目标函数的变化情况确定

■对象——向量聚类

$$x(t)=[x_1(t)\ x_2(t)\cdots x_n(t)]^T$$

$$1\le t\le N$$

n—考试科目
N—学生总数
$x_i(t)$—i同学考试成绩

$$\rho(\varpi)=\sum_{t\in\varpi}(x(t)-e_\varpi(x))^T(x(t)-e_\varpi(x))$$

$$\forall\varpi\subseteq J(N)$$

与中心点的距离，即紧密程度

其中 $e_\varpi(x)=\frac{1}{|\varpi|}\sum_{t\in\varpi}x(t)$　子类中心

问题 $\min_{\Omega}\sum_{i=1}^{k}\rho(\varpi_i)$　尽可能紧密！

■对象——变量聚类

$$X_i=[x_i(1)\ x_i(2)\cdots x_i(N)]$$

$$1\le i\le n$$

保守准则：同一类内，最不相关的两个变量。

$$\rho(\varpi)=\min_{i\in\varpi,\ j\in\varpi} r(X_i,X_j)\quad \forall\varpi\subseteq J(n)$$

r是两个变量的协方差，我们希望类内协方差的值越大越好。即：类内两个变量的夹角夹角越小越好。保守准则是先找到类内夹角最大的那两个变量。

其中 $r(X_i,X_j)$

$$=\frac{1}{N-1}\sum_{t=1}^{N}\frac{(x_i(t)-e(x_i))(x_j(t)-e(x_j))}{\sqrt{\delta^2(x_i)\delta^2(x_j)}}$$

类间：我们希望整体分类结果是相关性越强越好！

$$\max_{\Omega}\sum_{i=1}^{k}\rho(\varpi_i)$$

按照“最大最不相关”进行分类

■例：分片线性逼近问题

对象 $(y(t), x(t)),\ y(t) \in R,\ x(t) \in R^n$

$$1 \le t \le N$$

$$\rho(\varpi) = \min_{\alpha \in R,\ \beta \in R^n} \sum_{t \in \varpi} \left(y(t) - \left(\alpha + \beta^T x(t) \right) \right)^2$$

$\forall \varpi \subseteq J(N)$ 使优化问题有唯一解

问题 $\min_{\Omega} \sum_{i=1}^{k} \rho(\varpi_i)$

考虑向量聚类问题

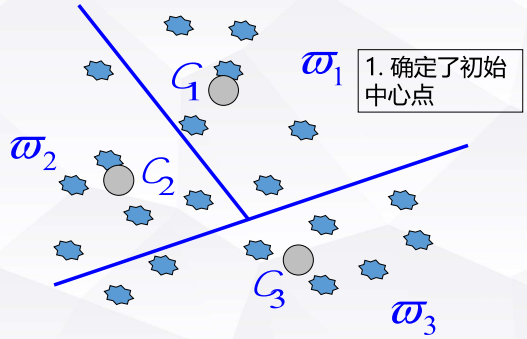
$$\min_{\Omega} \sum_{i=1}^{k} \sum_{t \in \varpi_i} \left(x(t) - e_{\varpi_i}(x) \right)^T \left(x(t) - e_{\varpi_i}(x) \right)$$

其中 $e_{\varpi}(x) = \frac{1}{|\varpi|} \sum_{t \in \varpi} x(t)$

如果已有一个分类 $\Omega = \{\varpi_i, 1 \le i \le k\}$

可以先求出 k 个集合的中心点(样本均值)

$$c_i = e_{\varpi_i}(x) \in R^n \quad 1 \le i \le k$$



1. 确定了初始中心点

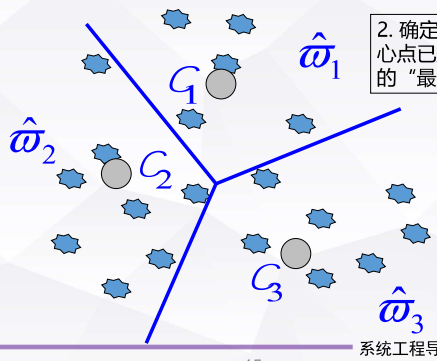
然后确定一个新的分类

$$\hat{\Omega} = \{\hat{\varpi}_i, 1 \le i \le k\}$$

其中

$$\hat{\varpi}_i = \{ t \in J(N) \mid (x(t) - c_i)^T (x(t) - c_i) \le (x(t) - c_j)^T (x(t) - c_j), \forall j \}$$

也就是说新的分类中每个x(t) 都距离ci更近!



2. 确定在初始中心点已知条件下的“最优”分类

由 $\hat{\varpi}_i = \{ t \in J(N) \mid (x(t) - c_i)^T (x(t) - c_i) \le (x(t) - c_j)^T (x(t) - c_j), \forall j \}$

可知 $\sum_{i=1}^{k} \sum_{t \in \hat{\varpi}_i} (x(t) - c_i)^T (x(t) - c_i) \le \sum_{i=1}^{k} \sum_{t \in \varpi_i} (x(t) - c_i)^T (x(t) - c_i)$

回忆向量分类的准则

又因为 $c_i = e_{\varpi_i}(x) \in R^n \quad 1 \le i \le k$

所以 $\sum_{i=1}^{k} \sum_{t \in \hat{\varpi}_i} (x(t) - c_i)^T (x(t) - c_i) \le \sum_{i=1}^{k} \rho(\varpi_i)$

另外, 对于优化问题

$$\min_{\alpha \in R^n} \sum_{t \in \hat{\varpi}_i} (x(t) - \alpha)^T (x(t) - \alpha)$$

由于 $\frac{\partial \left(\sum_{t \in \hat{\varpi}_i} (x(t) - \alpha)^T (x(t) - \alpha) \right)}{\partial \alpha} = -2 \sum_{t \in \hat{\varpi}_i} (x(t) - \alpha)$

所以最优解为 $\hat{\alpha}_i = \frac{1}{|\hat{\varpi}_i|} \sum_{t \in \hat{\varpi}_i} x(t) = e_{\hat{\varpi}_i}(x)$

3. 即最优解应该是该类的中心。

$$\sum_{i=1}^{k} \rho(\hat{\varpi}_i) = \sum_{i=1}^{k} \sum_{t \in \hat{\varpi}_i} \left(x(t) - e_{\hat{\varpi}_i}(x) \right)^T \left(x(t) - e_{\varpi_i}(x) \right) \le \sum_{i=1}^{k} \sum_{t \in \hat{\varpi}_i} (x(t) - c_i)^T (x(t) - c_i) \le \sum_{i=1}^{k} \rho(\varpi_i)$$

4. 判断当前ci是否是最优解的依据。只要有一个分类内部的中心点和前面已知的中心点不吻合，就没有找到最优分类。

并且, 只要有一个 $c_j \ne e_{\hat{\varpi}_j}(x)$

就一定有 $\sum_{i=1}^{k} \rho(\hat{\varpi}_i) < \sum_{i=1}^{k} \rho(\varpi_i)$

■步骤

相比前面做法，此处进行了简化!

1) 在 $x(t), t \in J(N)$ 中随机选取 $c_i, 1 \le i \le k$

2) 确定分类 $\Omega = \{\varpi_i, 1 \le i \le k\}$, 其中

$$\varpi_i = \{ t \in J(N) \mid (x(t) - c_i)^T (x(t) - c_i) \le (x(t) - c_j)^T (x(t) - c_j), \forall j \}$$

3) 如果 $e_{\varpi_i}(x) = c_i, 1 \le i \le k$, 停止, 否则令

$c_i = e_{\varpi_i}(x), 1 \le i \le k$, 回到2) 继续迭代

是否收敛?

■进一步解释

① 选中心点

首先确定分类数目 k，然后在所有样本中挑选 k 个作为初始中心点，

例如，可简单地取

$$c(i) = x(i), \quad i = 1, 2, \cdots, k$$

② 逐个利用每个样本修改中心点

对所有的样本，顺序进行下述计算：

第一、将其归入与其最近的中心点所在的类；

第二、重新计算该类的重心，并用新的重心替换中心点

■进一步解释

③ 停止准则

如果上述步骤开始时和结束后的中心点差别很小，停止分类

可以证明，该迭代算法是收敛的

系统聚类算法是一种贪婪算法！

考虑变量聚类问题

$$\max_{\Omega}\sum_{i=1}^{k}\rho(\varpi_i)$$

其中

$$\rho(\varpi)=\min_{i\in\varpi,\ j\in\varpi} r(x_i,x_j)\qquad \forall\varpi\subseteq J(n)$$

■基本步骤

(1) 首先将每个变量视为一类，得到 n 类变量

(2) 每次选择最相关的两个类合并，顺序得到

$n-1,\ n-2,\ n-3,\cdots$ 直至一类变量

(3) 记录合并过程生成聚类谱系图

(4) 设定阈值，根据聚类谱系图决定最终分类

假定所有变量已经合并成了 m 类

$$\varpi_i\subseteq J(n),\ 1\le i\le m$$

在确定下一步合并哪两类时，选择

$$1\le\alpha<\beta\le m$$

满足

$$\rho(\varpi_\alpha\cup\varpi_\beta)\ge\rho(\varpi_i\cup\varpi_j)\qquad\forall 1\le i<j\le m$$

将 ϖ_α 和 ϖ_β 合并，从而得到 $m-1$ 类

例 四个变量依相关性聚类

$$R=\begin{bmatrix}1 & 0.334 & 0.950 & 0.597\\ & 1 & 0.217 & -0.117\\ & & 1 & 0.569\\ & & & 1\end{bmatrix}\begin{matrix}x_1\\x_2\\x_3\\x_4\end{matrix}$$

森林面积　森林覆盖率　林木蓄积量　草原面积

$$\varpi_1=\{1\}\quad\varpi_2=\{2\}\quad\varpi_3=\{3\}\quad\varpi_4=\{4\}$$

$$\rho=\begin{bmatrix}1 & 0.334 & 0.950 & 0.597\\ & 1 & 0.217 & -0.117\\ & & 1 & 0.569\\ & & & 1\end{bmatrix}\begin{matrix}\varpi_1\\\varpi_2\\\varpi_3\\\varpi_4\end{matrix}$$

$\rho=\min R(x_i,x_j)$

聚类标准　$\rho(\varpi_\alpha\cup\varpi_\beta)\ge\rho(\varpi_i\cup\varpi_j)$

聚类：$\varpi_1\cup\varpi_3$

$$\rho=\begin{bmatrix}1 & 0.334 & 0.950 & 0.597\\ & 1 & 0.217 & -0.117\\ & & 1 & 0.569\\ & & & 1\end{bmatrix}\begin{matrix}\varpi_1\\\varpi_2\\\varpi_3\\\varpi_4\end{matrix}$$

$\rho=\min R(x_i,x_j)$

$\varpi_5=\varpi_1\cup\varpi_3\quad\varpi_6=\varpi_2\quad\varpi_7=\varpi_4$

$\rho(\varpi_5\cup\varpi_6)=\min\{0.334,0.217\}=0.217$

$\rho(\varpi_5\cup\varpi_7)=\min\{0.597,0.569\}=0.569$

$\rho(\varpi_6\cup\varpi_7)=-0.117$

■计算方法

$$\rho=\begin{bmatrix}1 & 0.334 & 0.950 & 0.597\\ & 1 & 0.217 & -0.117\\ & & 1 & 0.569\\ & & & 1\end{bmatrix}\begin{matrix}\varpi_1\\\varpi_2\\\varpi_3\\\varpi_4\end{matrix}$$

$\rho=\min R(x_i,x_j)$

$\varpi_5=\varpi_1\cup\varpi_3\quad\varpi_6=\varpi_2\quad\varpi_7=\varpi_4$

$\rho(\varpi_5\cup\varpi_6)=\rho(\varpi_1\cup\varpi_3\cup\varpi_2)$

$=\min\{R(x_1,x_2),R(x_3\cup x_2)\}$

$=\min\{0.334,0.217\}$

$=0.217$

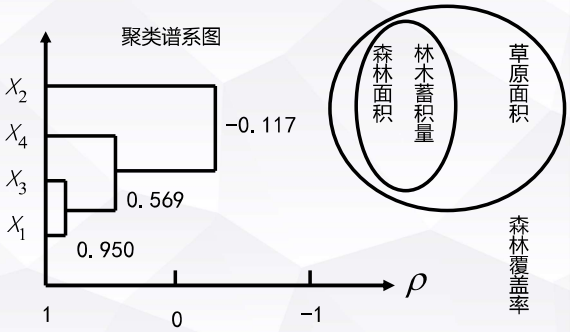
■计算方法

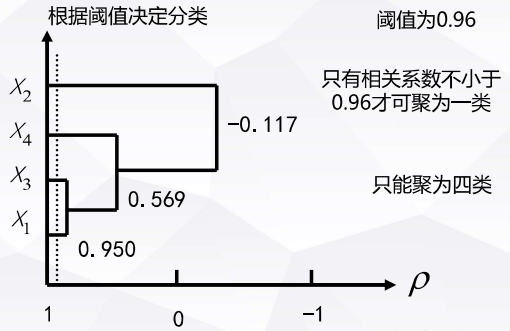
$$\rho=\begin{bmatrix}0.95 & 0.217 & 0.569\\ & 1 & -0.117\\ & & 1\end{bmatrix}\begin{matrix}\varpi_5\\\varpi_6\\\varpi_7\end{matrix}$$

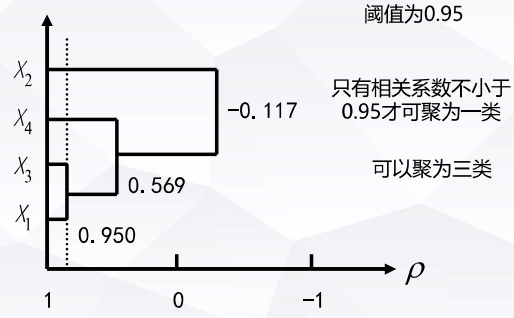
$\varpi_8=\varpi_5\cup\varpi_7\quad\varpi_9=\varpi_6$

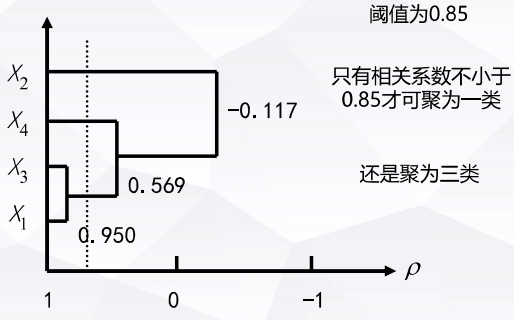
$\rho(\varpi_8\cup\varpi_9)=\min\{0.217,-0.117\}=-0.117$

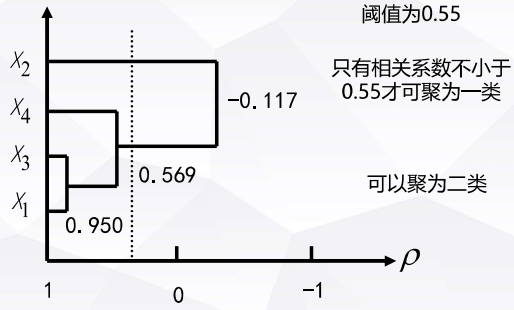
最后 $\varpi_{10}=\varpi_8\cup\varpi_9\quad\rho(\varpi_{10})=-0.117$



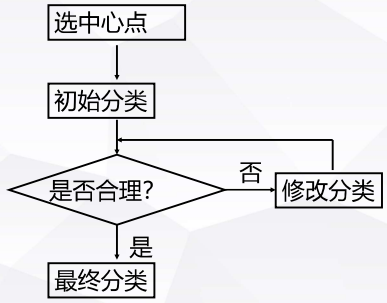








■基本步骤



例：对21个工厂的产品按两个质量指标分类

t	$x_1(t)$	$x_2(t)$	t	$x_1(t)$	$x_2(t)$
1	0	6	12	-2	2
2	0	5	13	-3	2
3	2	5	14	-3	0
4	2	3	15	-5	2
5	4	4	16	1	1
6	4	3	17	0	-1
7	5	1	18	0	-2
8	6	2	19	-1	-1
9	6	1	20	-1	-3
10	7	0	21	-3	-5
11	-4	3			

■按批修改方法

1）选中心点（密度法）

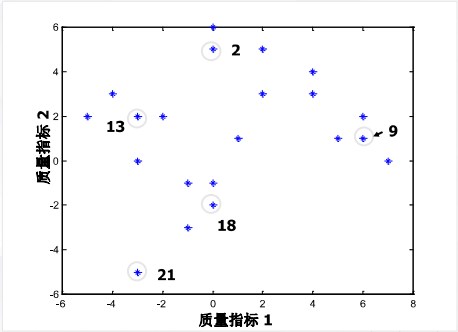
取阈值 $a_1 = 2$ ，对每个样本 $x(t)$ ，用

$\eta(t)$ 表示下述集合元素个数（密度）

$$\left\{ x(k), 1 \le k \le 21, k \ne t \,|\, d_2\left(x(k), x(t)\right) \le a_1 \right\}$$

例如 $\eta(2) = 2, \eta(9) = 3, \eta(13) = 4$

$\eta(18) = 3, \eta(21) = 0$



■按批修改方法

取密度最大的点为第一中心点

$$c(1) = x(13) \quad (\eta(13) = 4)$$

取阈值 $a_2 = 2a_1 = 4$ ，考虑密度第二大的点

$$\eta(9) = 3$$

由于 $d_2\left(x(9), c(1)\right) > a_2$ ，取为第二中心点

$$c(2) = x(9)$$

反之不用该点为中心点

■按批修改方法

考虑密度第三大的点 $\eta(18) = 3$

由于 $d_2\left(x(18), c(1)\right) > a_2$

$$d_2\left(x(18), c(2)\right) > a_2$$

取为第三中心点 $c(3) = x(18)$

最终得到

$$c(1) = x(13) \quad c(2) = x(9) \quad c(3) = x(18)$$
$$c(4) = x(2) \quad c(5) = x(21)$$

■按批修改方法

2）初始分类（最近中心点原则）

距某个中心点最近的点划归该中心点对应的类

如果对任意的 $j \neq i$

$$d_2\left(x(t), c(i)\right) \le d_2\left(x(t), c(j)\right)$$

将 $x(t)$ 划归第 i 类

■按批修改方法

3）修改分类（修改中心点）

用每类的重心替换中心点

重心：每类样本的均值向量

例如，若某类样本为

$$x(7),\ x(8),\ x(9),\ x(10)$$

其重心为

$$\frac{1}{4}\left(x(7)+x(8)+x(9)+x(10)\right)$$

■按批修改方法

4）停止准则

如果当前的中心点和对应的重心点都很接近，停止分类

可以证明，该迭代算法是收敛的。这就是说，只要迭代次数足够多，上述中心点和重心点可以任意接近

■说明

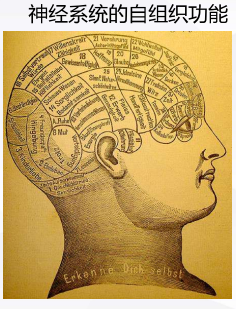
1）动态聚类方法的结果一般和初始中心点有关，选择初始中心点很重要

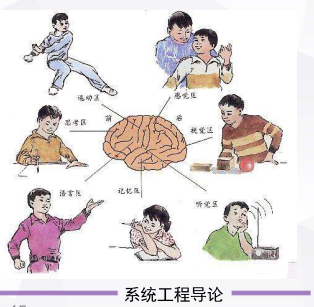
2）聚类分析是实践性很强的问题，所介绍的基本方法有很多变形和发展

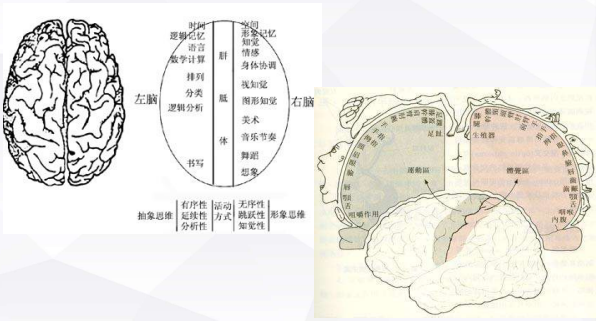
3）同系统聚类法对应的还有一大类逐渐分解的分类方法

■基于自组织映射（SOM）的聚类

神经系统的自组织功能







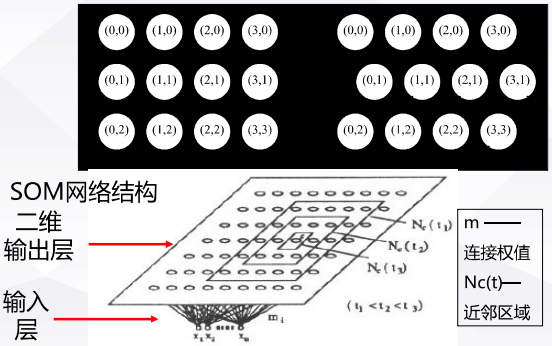
自组织映射：Self-organizing Map, SOM网络是一种有效的聚类和聚类结果可视化工具。

它定义了从n 维输入空间到规则的二维输出节点矩阵的非线性映射。

通常SOM是一个无监督的两层神经网络。

- 第一层是输入层
- 第二层是二维输出层。

输出层的点阵结构可以是矩形，也可以是六边形。



SOM网络结构

- 神经元呈平面分布
- 输入向量的每一维连接到每个结点
- 结点间依分布位置关系而有相互作用

神经元计算特性

- 向量匹配，与输入向量最佳匹配者称为Winner

SOM特性

- 经过适当训练后，结点对输入的响应表现出一定的规律性，这些规律性反映了输入样本集内部的一些特性，即样本分布密度及样本间拓扑关系。

■SOM网络中几个重要概念

- 象：随着不断学习，对于某个输入向量，对应Winner结点逐步趋于固定，该结点称作该样本的象
- 原象：若向量 x 的象为结点 i，则x为i的一个原象
- 象密度：同一结点上原象的数目，即有多少样本映射到该结点。
- 象密度图：按节点位置把象密度的相对值画到一张图上，这张图称为象密度图。

■自组织特性

- 每个样本有且只有一个象（固定映射到某一结点），但一个结点可能有多个原象（也可能没有）
- 在原空间中距离相近的样本趋于映射到相同或相近的象
- 节点象密度与原象附近原空间中的样本密度有某种单调关系

■自组织分析（SOMA）

基本思路

- 用未知样本集训练SOM
- 计算并绘制象密度图
- 根据象密度图划分聚类（把结点代表的小聚类合并）

特点

- 对数据分布形状依赖性小
- 可反映真实存在的聚类数目
- 高维数据的有效二维表示

■SOM学习算法

（1）权值初始化（取小随机数，但不能相等）

（2）在时刻 t, 加入一个输入样本 x（从样本集中选择或抽取，有时须调整顺序）

（3）寻找最佳匹配点c（winner）

如用欧式距离匹配：$c:\|x(t)-m_c(t)\|=\min_{i\in A}\{\|x(t)-m_i(t)\|\}$

（4）权值竞争学习：

$$m_i(t+1)=m_i(t)+\alpha(t)h_{ci}(t)d[x(t),m_i(t)],\forall i\in A$$

其中，$\alpha(t)$ 为步长，$h_{ci}(t)$为结点 i 与 c 之间的近邻函数值，$d[x(t),m_i(t)]$为误差函数，通常采用欧式距离.

■SOM学习算法

如采用矩形领域，则

$$m_i(t+1)=\begin{cases}m_i(t)+\alpha(t)[x(t)-m_i(t)] & i\in N_c(t)\\ m_i(t) & i\notin N_c(t)\end{cases}\quad\forall i\in A$$

其中，$N_c(t)$是结点 c 的矩形领域中的结点集合。

（5）更新步长 $\alpha(t)$ 和领域 $N_c(t)$，如达到终止条件（如迭代次数）则终止，否则置 $t=t+1$，转（2）

（$\alpha(t)$ 和 $N_c(t)$ 一般应随 t 增大而减小或缩小！）

■站点客流集散量聚类研究——研究背景

众所周知，一般的公交线路上乘客出行都具有一定的规律性。反映在公交站点上，就是客流集散量按照一定规律变化。

站点与站点客流集散量之间也同样存在某种相关关系，这是因为在季节因素、天气因素基本相同的条件下，每周同一天同一时间段乘客在何处上车，到何处下车，换乘那条线路也具有规律性，表现的结果就是这些站点间客流集散量具有一定的相关关系。

■站点客流集散量聚类研究——主要步骤

（1）聚类前的数据变换处理——标准化处理

$$x'_{ij}=\frac{x_{ij}-\bar{x}_j}{S_j}\qquad(i=1,2,...,n\quad j=1,2,...,m)$$

其中：$\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$　　$S_j=\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2\right]^{1/2}$

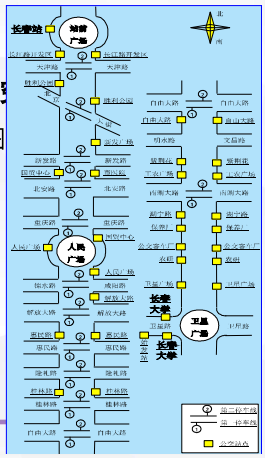
（2）相似系数的确定。即：选择度量样品或变量之间亲疏程度的相似系数，本研究采用平方欧式距离，该值越小，说明亲密程度越大。

（3）计算类间相似系数并进行聚类。

（4）绘制系统聚类谱系图。

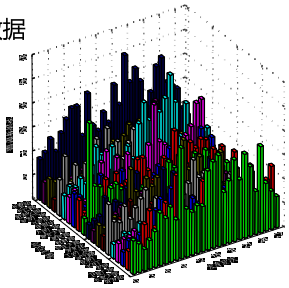
■站点客流集散量聚类研

长春市306路公交车示意图



■站点客流集散量聚类研究——实证研究

客流集散量原始数据



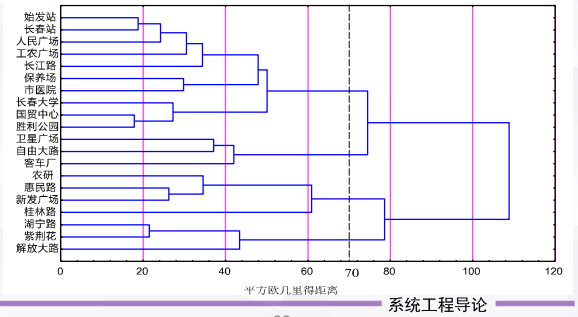
站点间平方欧几里得距离

站号	站点名称	始发站	长春大学	卫星广场	农研	客车厂	保养厂	湖宁路	工农广场	紫荆花	自由大路
1	始发站	0									
2	长春大学	22.0	0								
3	卫星广场	28.3	43.9	0							
4	农研	49.3	50.7	40.7	0						
5	客车厂	49.5	52.9	41.2	50.2	0					
6	保养厂	40.4	30.6	52.4	43.0	55.4	0				
7	湖宁路	42.8	37.3	63.5	58.7	53.9	41.6	0			
8	工农广场	31.1	30.9	36.8	42.3	53.9	32.9	44.3	0		
9	紫荆花	36.0	44.4	46.3	62.7	46.2	54.8	21.5	53.8	0	
10	自由大路	25.4	32.2	37.1	41.5	40.3	41.8	37.2	34.8	36.6	0
11	桂林路	60.9	62.3	63.2	56.1	65.7	52.2	51.2	53.9	55.7	53.3
12	惠民路	39.4	52.9	52.6	34.7	58.7	41.2	38.0	41.9	46.1	40.5
13	解放大路	39.4	53.6	54.6	58.1	66.6	45.1	36.1	60.9	39.7	46.1
14	人民广场	25.5	30.5	45.6	52.6	61.4	36.4	45.0	31.7	45.3	37.0
15	国贸中心	28.7	19.9	35.7	27.5	42.2	30.8	55.1	28.6	58.1	42.1
16	市医院	28.0	38.2	32.9	41.5	64.8	29.8	49.6	30.0	41.4	30.1
17	新发广场	42.8	62.8	44.3	30.3	64.3	43.7	46.9	45.2	48.6	53.4
18	胜利公园	18.9	29.9	29.1	33.1	50.0	33.8	62.9	39.2	55.3	40.6
19	长江路	24.9	33.1	51.9	46.6	66.1	44.4	55.3	36.8	58.6	45.9
20	长春站	18.8	26.9	44.0	36.8	47.6	30.4	36.3	19.7	41.5	26.9

站点间平方欧几里得距离（续）

站号	站点名称	桂林路	惠民路	解放大路	人民广场	国贸中心	市医院	新发广场	胜利公园	长江路	长春站
11	桂林路	0									
12	惠民路	54.4	0								
13	解放大路	55.0	31.6	0							
14	人民广场	51.6	34.9	41.2	0						
15	国贸中心	59.3	47.6	58.0	50.2	0					
16	市医院	38.1	40.9	46.1	34.6	42.8	0				
17	新发广场	41.7	26.3	34.6	40.9	42.9	42.8	0			
18	胜利公园	63.5	47.8	52.0	38.7	17.9	38.0	37.8	0		
19	长江路	53.3	29.1	41.0	30.4	35.6	44.2	41.0	33.8	0	
20	长春站	50.2	30.7	50.1	20.2	26.8	28.8	30.8	28.8	30.6	0

■站点客流集散量聚类研究——聚类谱系图



■站点客流集散量聚类研究——聚类结果

我们选择平方欧几里得距离为70将公交站点进行分类。分类结果为：

- 第一类：始发站、长春站、人民广场、工农广场、长江路开发区、保养场、市医院、长春大学、国贸中心和胜利公园；
- 第二类：卫星广场、自由大路和公交客车厂；
- 第三类：农研、惠民路和新发广场；
- 第四类：湖宁路、紫荆花酒店和解放大路。

■基于SOM网络聚类方法——研究背景

- 网络交通流时变、非线性特征
- 网络交通状态演变规律研究的重要性
- 网络交通流短时预测的重要性
- 交通流数据的海量特性

问题：如何从中挖掘出交通状态的演变规律，对网络层次的交通流进行预测？

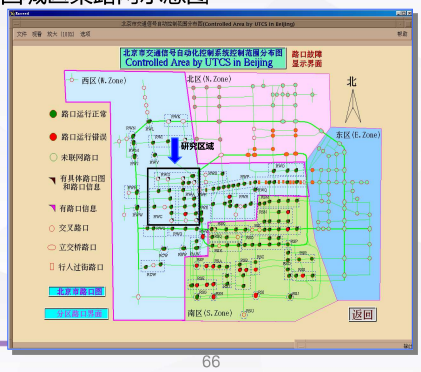
■基于SOM网络聚类方法——研究方法

- 将网络交通流数据转化为N维时间序列

  即：$\mathbf{F}(t)=[f_1(t), f_2(t),\cdots, f_d(t)]^T$

  其中：$f_i(t)$ 表示 t 时段第 i 个路段的交通流量。d 表示所研究交通网络的路段总数。
- 将上述数据作为SOM网络的输入，输出是一个10×10结点的二维平面。
- SOM聚类
- 基于聚类结果的网络交通状态分析与短时交通流预测研究。

北京市西城区某路网示意图



■基于SOM网络聚类方

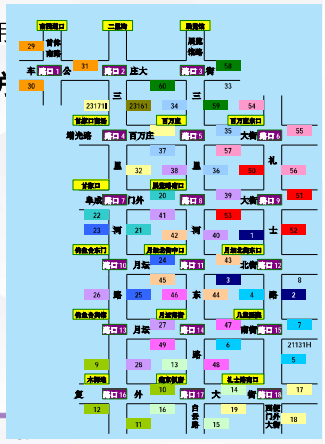
47个路段

2002年3月1日至3月30日一个月的交通流数据（每个路口每15分钟一个数据）

即整个交通数据构成一个47维，长度为2786个有效数据

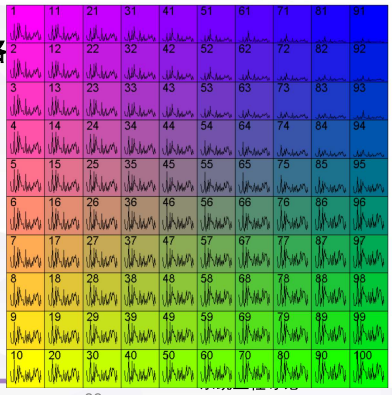
2500个用于训练

286个用于预测



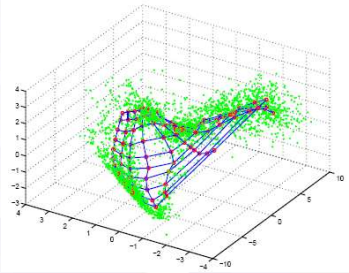
■基于SOM网络

每一类用该类中所有流率向量的平均值来表示。对于每个结点，X轴表示路段名序列，Y轴表示流率 (veh/h). 左上角数字即聚类号。不同类用不同颜色表示，相邻结点具有相似颜色。



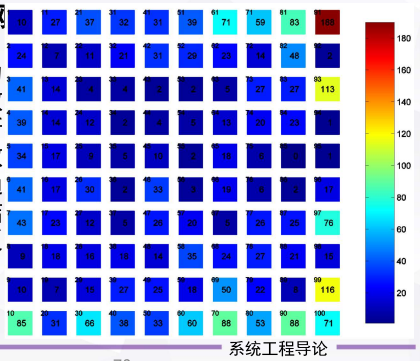
■基于SOM网络聚类方法——聚类结果

绿色点：样本点；红色点：SOM聚类中心结点



■基于SOM网

左上角数字为类号，中间数字为该类中所包含的向量数目。颜色越趋向暖色表示结点数据多，反之少。



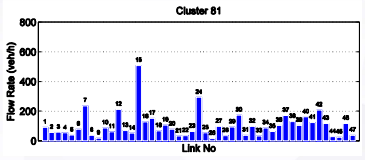
典型交通流分布模式类

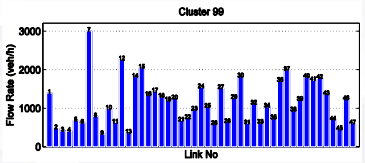
X轴：路段号

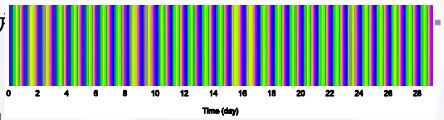
Y轴：流率

类81：包含83个向量

类99：包含116个向量





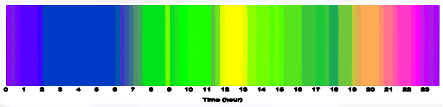


交通流向量时间序列分析

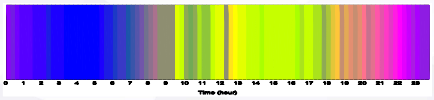
每个时间段的数据都用它所在的类来表示。

X轴：时间点

整个月网络交通流变化



某个周五网络交通流变化



某个星期日网络交通流变化

SOM实质上起到了非线性降维的作用，由47维降到2维。我们还可以将其用于网络交通流“批量”预测。具体方法（kNN法）包含三个步骤：

（1）创建 t 时刻的统计量，预测 t + 1 时刻流率

$$r(t)=[c_1(t),c_2(t),c_1(t-1),c_2(t-1),\cdots,c_1(t-p+1),c_2(t-p+1)]^T$$

（2）计算当前时间的统计量与前面所有统计量的欧拉距离

$$d(s)=\|r(t)-r(s)\|,s=1,2,3,\cdots,l$$

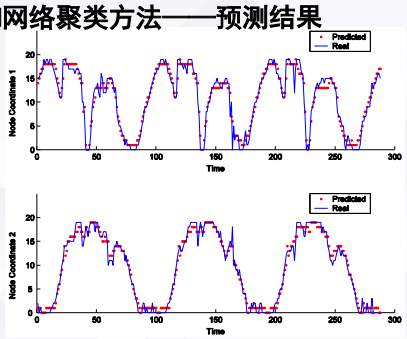
（3）确定k个与当前统计量最为接近的统计量，下一时刻的流量的预测值即为用这k个统计量对应时刻的流率平均值。

$$F(t+1)=\frac{1}{k}[F(s_1)+F(s_2)+\cdots+F(s_k)]$$

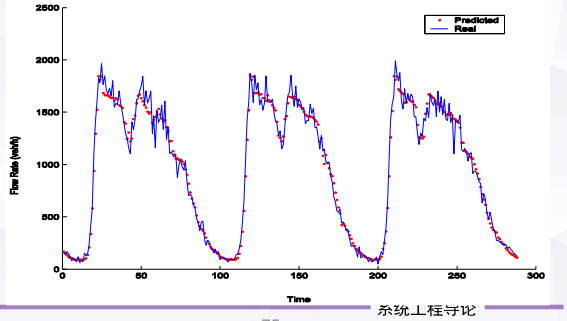
t+1时刻对应的结点坐标值也可以用前k个时段的结点坐标的均值来计算。

■基于SOM网络聚类方法——预测结果

取p = 4



■基于SOM网络聚类方法——预测结果



■基于SOM网络聚类方法——预测结果

预测结果比较（K为最近邻数目）

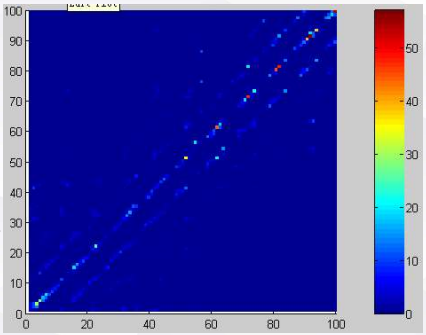
k	kNN with SOM	kNN with PCA
	MAPE	MAPE
5	0.1619	0.1734
10	0.1414	0.1717
20	0.1433	0.1705
30	0.1519	0.1703
40	0.1393	0.1707

$$mape=\frac{\sum_{i=1}^{d}\sum_{t=1}^{l}|f_i(\hat{t})-f_i(t)|/f_i(t)}{d\times l}$$

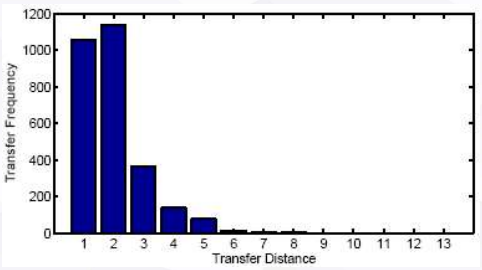
模式转移研究

横坐标是前一时刻的标号，纵坐标是下一时刻的标号

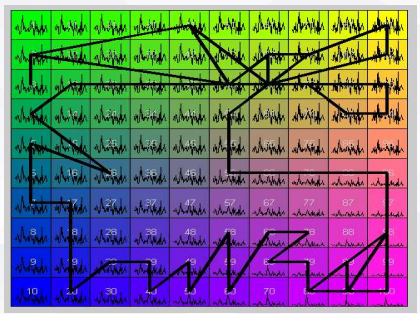
可见集中在对角线上（停留）或行列标号相差10、20的位置（只转移到相邻的pattern）



模式转移分布



第25天内的转移路径



第23天内的转移路径