河北师范大学软件学院
Software College of Hebei Normal University

第三部分 特征工程

自组织映射神经网络(SOM)

--高维数据的低维可视化、聚类

张朝晖

2018-2019学年 2018年10月

1.分类模块

产生式分类模型 **A.**贝叶斯分类模型

判别式分类模型 **线性分类模型**
- **B. Fisher**判别分类
- **C.** 感知器分类模型
- **D.** 大间隔分类模型(线性***SVM***)

非线性分类模型
- **E.** 核***SVM***(非线性***SVM***)
- **F.** 核**Fisher**判别分类
- **G.** 神经网络(***bp*,*autoencoder*,*som***)

其它分类模型
- ***H*.KNN**分类模型
- ***I*.**决策树分类模型
- ***J*.*Logistic***回归
- ***K*.*Softmax***回归

2.聚类模块
- ***L*.K-**均值聚类
- ***M*.**高斯混合聚类
- ***N*.DBSCAN**聚类
- ***O*.**层次聚类

3.回归模块
- ***P*.KNN**回归
- ***Q*.**回归树
- ***R*.**最小二乘线性回归
- ***S*.**岭回归
- ***T*.**LASSO回归

4.集成学习
- ***U*.*Bagging***
- ***V*.**随机森林
- ***W*.*Boosting***
  - ***AdaBoost*,*GBDT*,*XGBoost***
  - ***lightGBM*,*CatBoost***

5.特征工程
- ***X*.**主成分分析***(PCA)***
- ***Y*.kernel-PCA**
- ***Z*.SOM**

6.评价模块 混淆矩阵(及其相关指标)、ROC曲线、交叉验证

神经网络三种基本模型

(1)**前馈型神经网络**(***feedfroward network***)

多层感知器

BP网络

RBF网络

(2)**反馈网络**(***feedback network***)

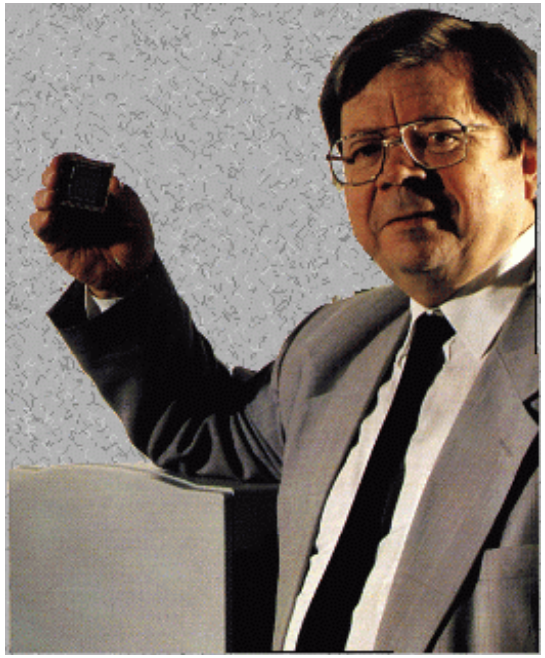
*Hopfield*网络

(3)**竞争学习网络**(***competitive learning network***)

*SOM*网络

1 *SOM*网络的引入

(1) *SOM*的产生



最早由芬兰赫尔辛基大学神经网络专家***Teuvo Kohonen*** 于 **1981**年提出.

SOM模拟大脑神经系统自组织特征映射功能，在训练中无监督地自组织学习。

SOM网络是一种有效聚类和对聚类结果可视化的工具.

book - 2001 - Self - Orginazing Maps

(2) *SOM*网络的生物学基础

生物神经元的“自组织现象”：在人脑的感觉通道上，神经元的组织原理是有序排列。

人脑通过感官接受外界的特定时空信息时，大脑皮层的特定区域兴奋，且类似的外界信息在对应区域是连续映象的。

大脑皮层关于不同类型刺激的特定兴奋过程、神经元的有序排列、以及对外界信息的连续映象是自组织特征映射网中竞争机制的生物学基础。

(3)SOM网络结构--两层网络

输入层

模拟视网膜神经元，**接收输入模式**

样本点 $\boldsymbol{x} = \begin{bmatrix} x_1 & \cdots & x_d \end{bmatrix}^T$

d个输入单元，d是高维输入数据的维数

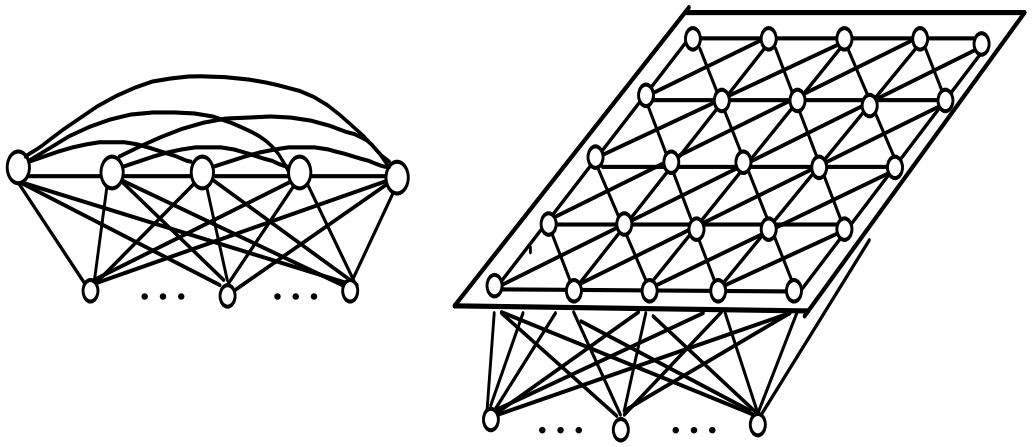
竞争层(输出层,映射层)

模拟大脑皮层神经元，n个可能**映射位置**

竞争层神经元节点集合$\boldsymbol{y} = \{ y_1 \quad \cdots \quad y_n \}$

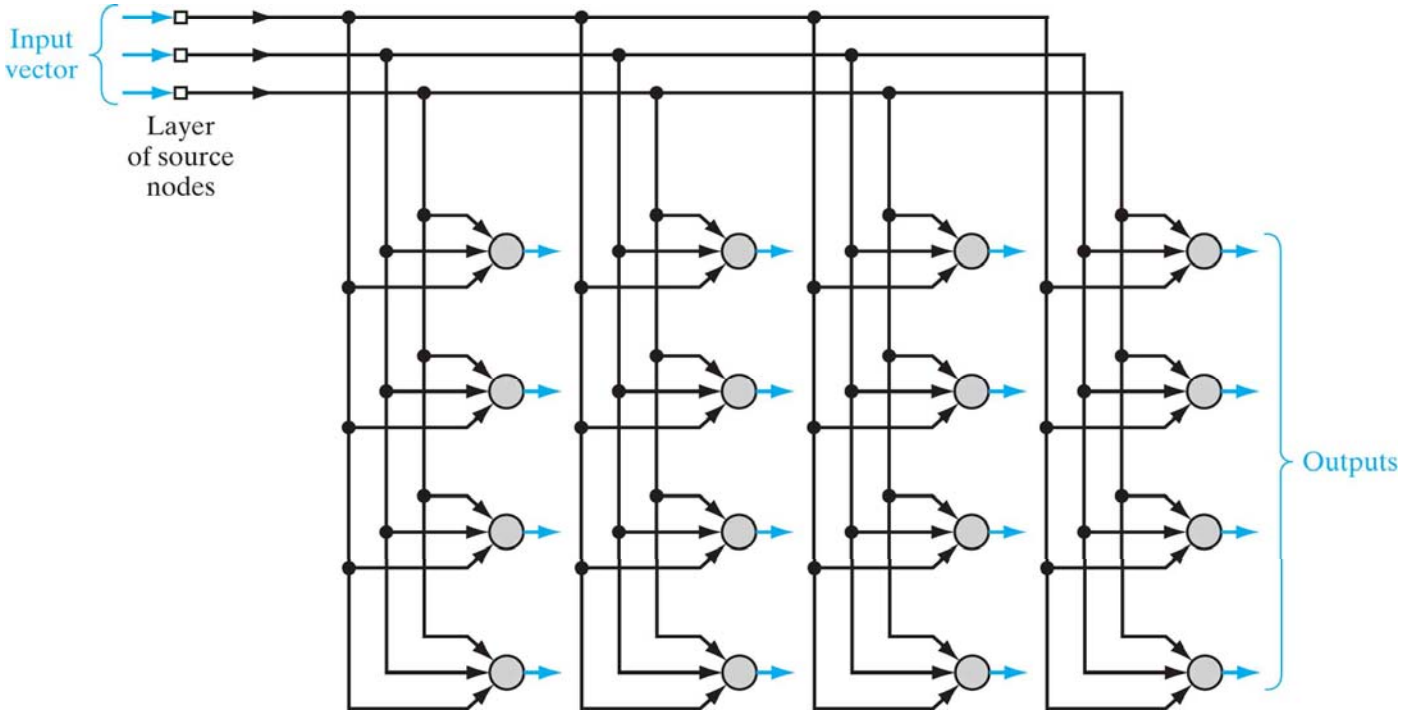
不同层节点之间的连接权值

SOM网络定义了d维输入空间向规则低维输出节点的非线性映射。



(a)一维线阵　　(b)二维平面线阵

The SOM may be described formally as a nonlinear, ordered, smooth mapping of high-dimensional input data manifolds onto the elements of a regular, low-dimensional array.

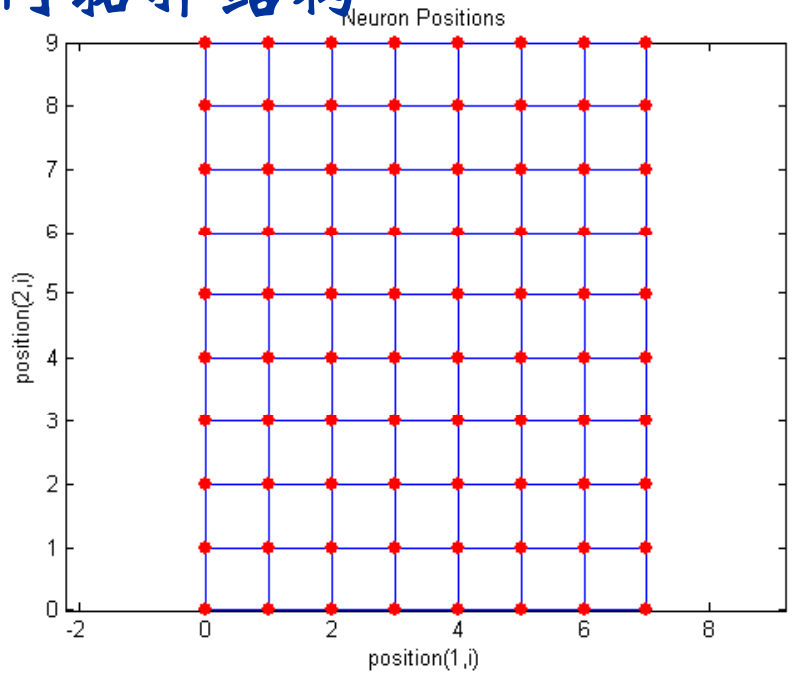


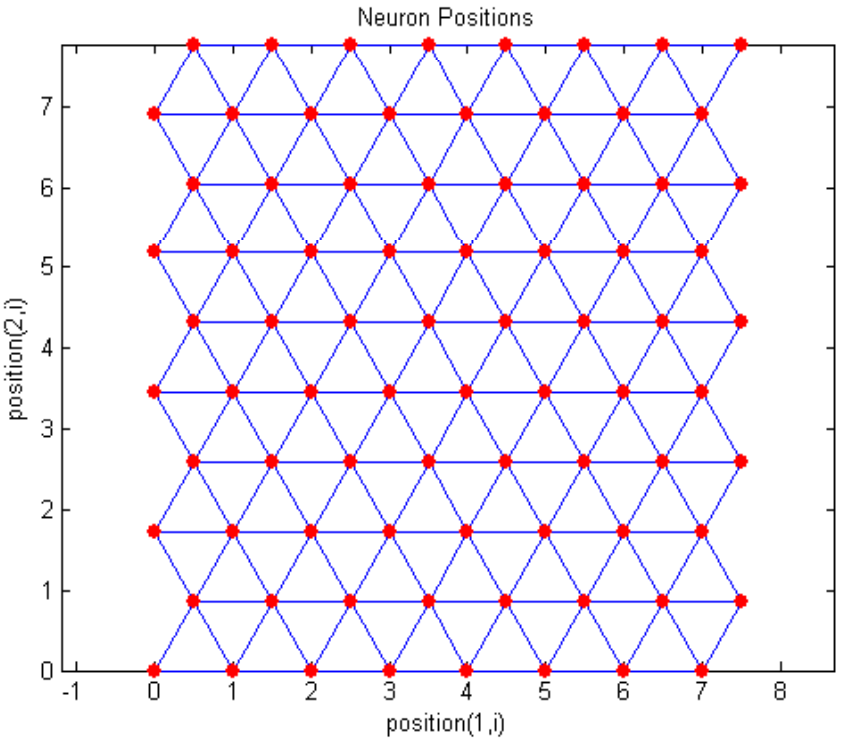
输出层(竞争层)神经元间的不同拓扑结构

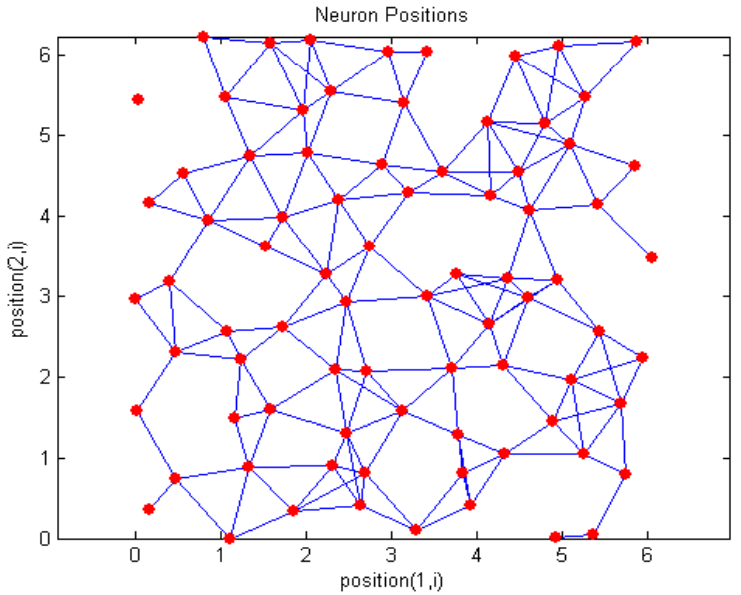
矩形结构（网格结构）

六角形结构

随机拓扑结构







It converts the nonlinear statistical relationships between high-dimensional data into simple geometric relationships of their image points on a low-dimensional display, usually a regular two-dimensional grid of nodes.

As the SOM thereby compresses information while preserving the most important topological and/or metric relationships of the primary data elements on the display, it may also be thought to produce some kind of abstractions.

These two aspects, visualization and abstraction, can be utilized in a number of ways in complex tasks such as process analysis, machine perception, control, and communication.

引自 ***book - 2001 - Self - Orginazing Maps***

2. SOM的学习--竞争式的自学习

(1)输出层节点间的近邻函数(交互作用函数)

若
$$\begin{cases}\sigma(t)\text{--有效窗半径，有效近邻半径}\\ d_{cj}\text{--两输出节点}c,j\text{之间的距离}\\ h_{cj}(t)\text{--节点}c\text{关于节点 }j\text{ 近邻函数}\end{cases}$$

巴拿马草帽形式
$$h_{cj}(t)=\begin{cases}1-\dfrac{d_{cj}}{\sigma(t)} & d_{cj}\le\sigma(t)\\ 0 & \text{其它}\end{cases}$$

矩形窗形式
$$h_{cj}(t)=\begin{cases}1 & d_{cj}\le\sigma(t)\\ 0 & \text{其它}\end{cases}$$

高斯形式
$$h_{cj}(t)=\begin{cases}e^{-\frac{d_{cj}^2}{2\sigma^2(t)}} & d_{cj}\le\sigma(t)\\ 0 & \text{其它}\end{cases}$$

(2)基本学习算法

学习过程包括：竞争, 合作, 更新(突触调节)

设
$$\begin{cases} \boldsymbol{X}=\left\{\boldsymbol{x}\in\boldsymbol{R}^{\boldsymbol{d}}\right\}\text{--}\boldsymbol{d}\text{维输入样本向量集合} \\ \boldsymbol{A}=\left\{1,2,\ldots,\boldsymbol{n}\right\}\text{--竞争层所有神经元节点集合} \\ \boldsymbol{m}_{\boldsymbol{i}}=\left[m_{i1}\cdots m_{id}\right]^{\boldsymbol{T}}\text{--竞争层第}\boldsymbol{i}\text{个神经元的权向量,}\boldsymbol{i}\in\boldsymbol{A} \end{cases}$$

*STEP*0 准备工作.

A.设定竞争层(输出层)神经元节点数目 n

B.设定硬性终止条件(最大迭代次数**T**);

C.归一化各输入样本向量x $\quad x \leftarrow \frac{x}{\|x\|} = \left[\frac{x_1}{\|x\|} \quad \cdots \quad \frac{x_d}{\|x\|}\right]^T$

D.设定时间常数τ_1, τ_2

E.设定拓扑邻域"有效窗半径"，例$\sigma(t)=\sigma_0 e^{-\frac{t}{\tau_1}}$, $t=0,1,...$

σ_0--初始窗半径 $\quad \sigma(t)=\sigma_{initial}\left(\frac{\sigma_{final}}{\sigma_{initial}}\right)^{\frac{t}{t_{max}}}$

F.定义随时间变化的近邻函数，例$h_{cj}(t)=e^{-\frac{d_{cj}^2}{2\sigma^2(t)}}$

d_{cj} --神经元j关于获胜神经元c的欧式距离

G.定义随时间变化的学习率$\alpha(t)$,比如$\alpha(t)=\alpha_0 e^{-\frac{t}{\tau_2}}$ $\quad \alpha(t)=\alpha_{initial}\left(\frac{\alpha_{final}}{\alpha_{initial}}\right)^{\frac{t}{t_{max}}}$

STEP1 权值初始化.

A. 以小随机数初始化输入层与竞争层神经元节点j之间连接权向量$\boldsymbol{m}_j, \forall j \in A$

B. 归一化竞争层神经元节点j的权向量$\boldsymbol{m}_j$

$$\boldsymbol{m}_j = \begin{bmatrix} m_{1j} & \cdots & m_{dj} \end{bmatrix}^T \qquad \forall j \in A$$

$$\boldsymbol{m}_j \leftarrow \frac{\boldsymbol{m}_j}{\|\boldsymbol{m}_j\|} = \begin{bmatrix} \frac{m_{1j}}{\|\boldsymbol{m}_j\|} & \cdots & \frac{m_{dj}}{\|\boldsymbol{m}_j\|} \end{bmatrix}^T$$

STEP2. 取样.

在时刻t按照给定顺序或随机顺序输入一个归一化样本向量，记为$\boldsymbol{x}(t)$.

*STEP*3. 相似匹配.

计算竞争层各神经元响应，并确定获胜节点*c*.

若以**欧式距离**作为匹配规则，则获胜节点c满足

$$\|x(t)-m_c(t)\|=\min_{j\in A=\{1,2,..,n\}}\{\|x(t)-m_j(t)\|\}$$

即 $$c(x(t))=\arg\min_{j\in A=\{1,2,..,n\}}\{\|x(t)-m_j(t)\|\}$$

获胜节点c为**主兴奋神经元**, 是输入向量$x(t)$的最佳匹配.

注：向量间**欧式距离最小等价于内积最大**;

节点竞争基于**胜者为王**(*winner-take-all*)策略.

STEP4. **权值竞争学习.**(**合作+更新**)对于$\forall j \in A$

A.更新竞争层所有神经元节点j的权向量m_j

$$m_j(t+1)=m_j(t)+\alpha(t)h_{cj}(t)\left[x(t)-m_j(t)\right]$$

学习步长 $\alpha(t)=\alpha_0 e^{-\frac{t}{\tau_2}}$

高斯形式的近邻函数$h_{cj}(t)=e^{-\frac{d_{cj}^2}{2\sigma^2(t)}}$

矩形邻域的近邻函数$h_{cj}(t)=\begin{cases}1 & 若d_{cj}\le\sigma(t)\\ 0 & 否则\end{cases}$

B.归一化 $m_j(t+1)\leftarrow\dfrac{m_j(t+1)}{\left\|m_j(t+1)\right\|}$

*STEP*5. 参数更新

A.调整学习步长 $\alpha(t) \leftarrow \alpha(t+1)$

B.更新近邻函数 $h_{cj}(t) \leftarrow h_{cj}(t+1)$

C.更新迭代次数 $t \leftarrow t+1$

D.终止条件判断

若满足终止条件，则学习终止；

否则，重复***STEP***2 ~ ***STEP***5.

(3)**SOM网络"学习"的有关说明：**

A.学习终止条件

- **硬性终止条件**　　迭代次数达到最大$t=T$
- **其它终止条件**$(t<T)$
  - 如：特征映射不再出现明显变化；
  - 学习过程中权向量改变量小于某阈值

B.学习过程　全局 $\rightarrow$ 局部, 粗调 $\rightarrow$ 精调

B.参数调整　可在学习过程根据需要更新有关参数

- **学习步长**可随迭代次数增加缓慢递减
  - 如：$\alpha(t)=\alpha_0 e^{-\frac{t}{\tau_2}}$　　$\alpha(t)=0.995\cdot\alpha(t-1)$　　$\alpha(t)=\alpha_0\left(1-\frac{t}{T}\right)$
- **拓扑邻域"有效窗半径"**
  - 如：$\sigma(t)=\sigma_0 e^{-\frac{t}{\tau_1}}$　　$\sigma(t)=\sigma_0\cdot\left(1-\frac{t}{T}\right)$

3. SOM网络的自组织现象(特征映射)

原像(原始空间样本x)$\xrightarrow{\Phi}$**像**(竞争层最佳响应节点)

样本密度 $\longrightarrow$ 像密度

(1)**输入空间的近似** 以较小原型集实现大型输入空间的近似.

(2)**拓扑排序** 特征映射Φ是拓扑有序的，输出层上神经元空间位置对应于输入模式的特定区域或特征.

(3)**密度匹配** **SOM**反映了输入分布的统计特征变化. 输入空间以高概率产生样本的区域在输出空间映射为较大区域.

(4)**特征选择与特征提取**

从原始特征空间给定数据，可通过**SOM**为逼近固有分布选择一组较好的特征；借助**SOM**非线性变换实现低维映射空间的特征提取.

4. 自组织分析（SOMA）

(1) 基本思路

- 用未知样本集训练SOM，得到样本像和节点原像
- 计算并绘制像密度图
- 根据像密度图划分聚类
  把节点代表的小聚类合并

(2)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

4. SOM用于模式识别

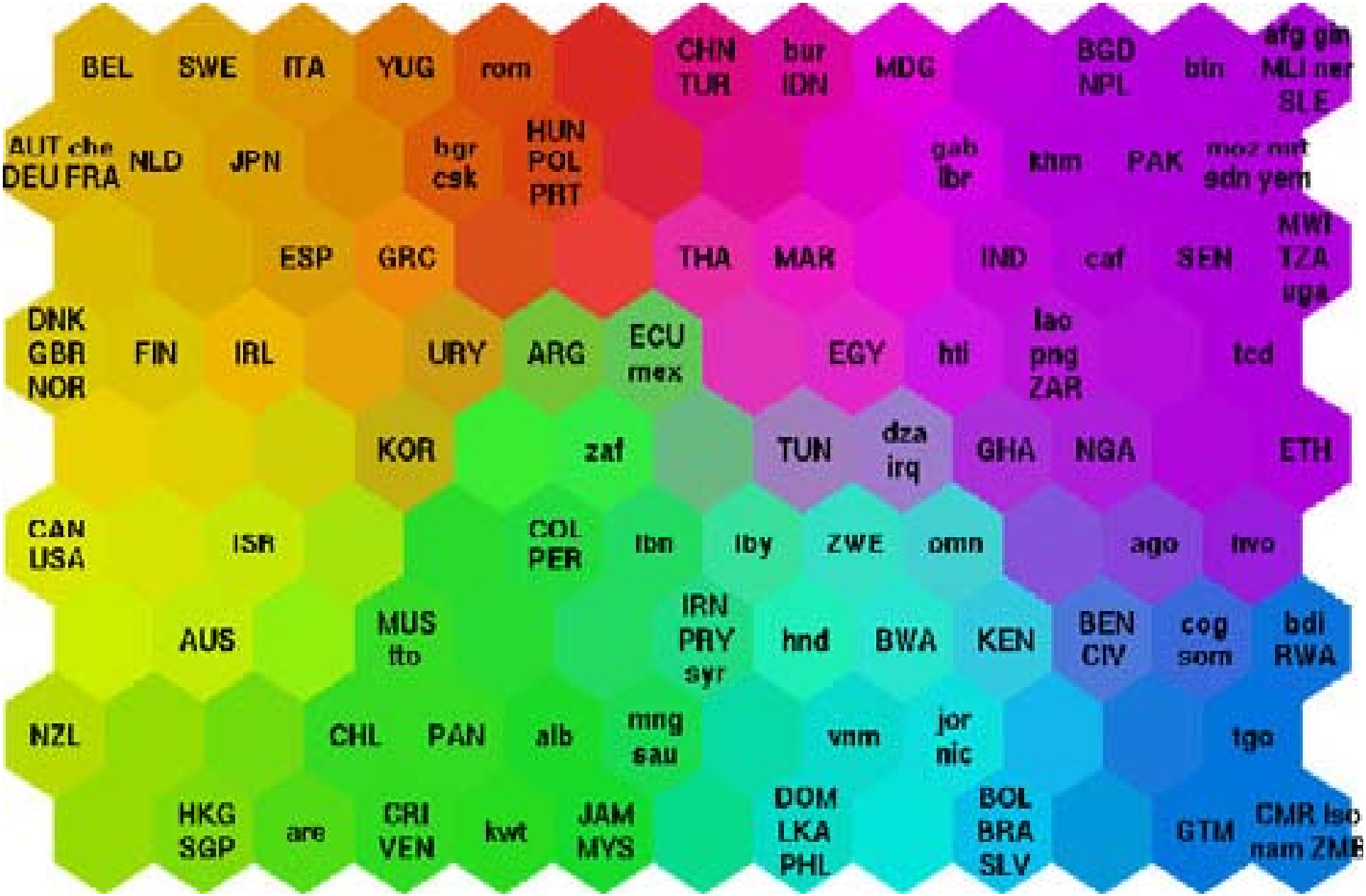
应用：数据降维，数据压缩；数据可视化；聚类,...

例1 不同国家生活质量分析

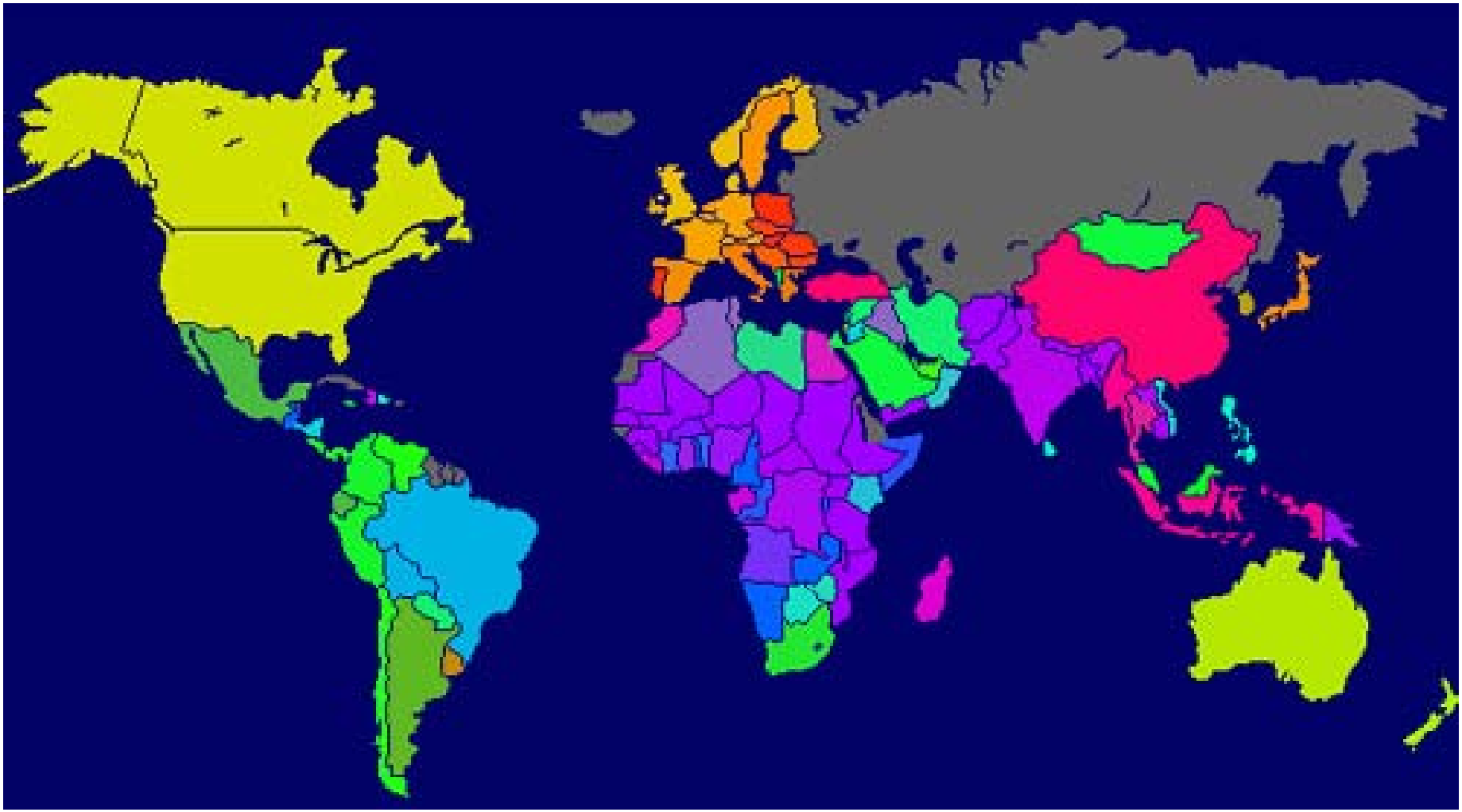
http://www.ai-junkie.com/ann/som/som5.html

- **To classify statistical data describing various quality-of-life factors such as state of health, nutrition, educational services etc.**
- **Countries with similar quality-of-life factors end up clustered together.**
- **The countries with better quality-of-life are situated toward *the upper left* and the most poverty stricken countries are toward *the lower right*.**
- **SOM does not show poverty levels, rather it shows how similar the poverty sets for different countries are to each other. (Similar color = similar data sets).**

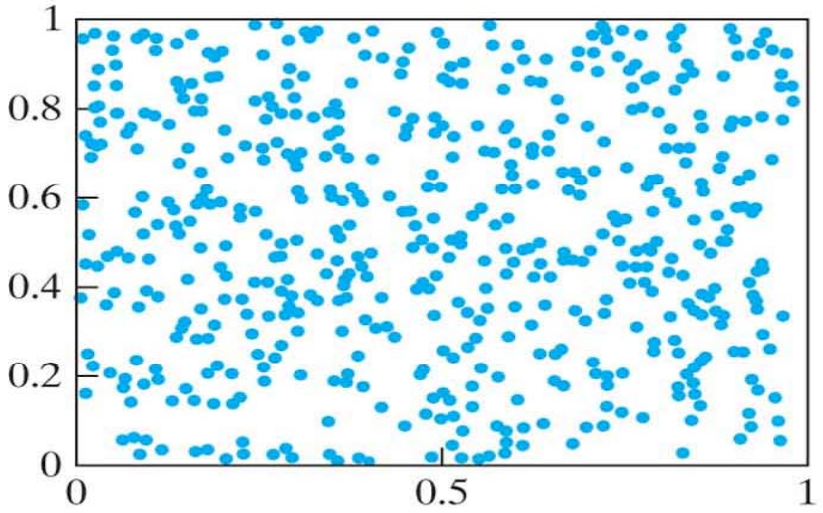
SOM网络


BEL SWE ITA YUG rom
CHN TUR bur IDN MDG
BGD NPL bin
afg gin MLI ner SLE
AUT che DEU FRA NLD JPN
bgr csk HUN POL PRT
gab ibr khm PAK moz nrl sdn yem
ESP GRC
THA MAR IND caf SEN
MWI TZA uga
DNK GBR NOR FIN IRL
URY ARG ECU mex
EGY hti lao png ZAR tcd
KOR zaf
TUN dza irq GHA NGA ETH
CAN USA ISR
COL PER ibn iby ZWE omn ago mvo
AUS
MUS tto
IRN PRY syr hnd BWA KEN
BEN CIV cog som bdi RWA
NZL
CHL PAN alb mng sau vnm jor nic tgo
HKG SGP are
CRI VEN kwt JAM MYS
DOM LKA PHL
BOL BRA SLV
GTM CMR iso nam ZMB

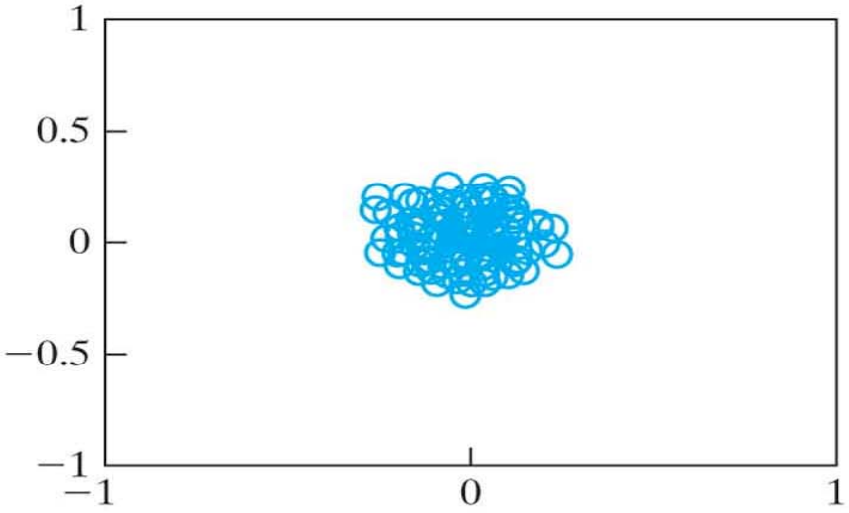
映射结果在地图可视化



例2 二维输入空间在一维离散空间(100个神经元节点)的特征映射

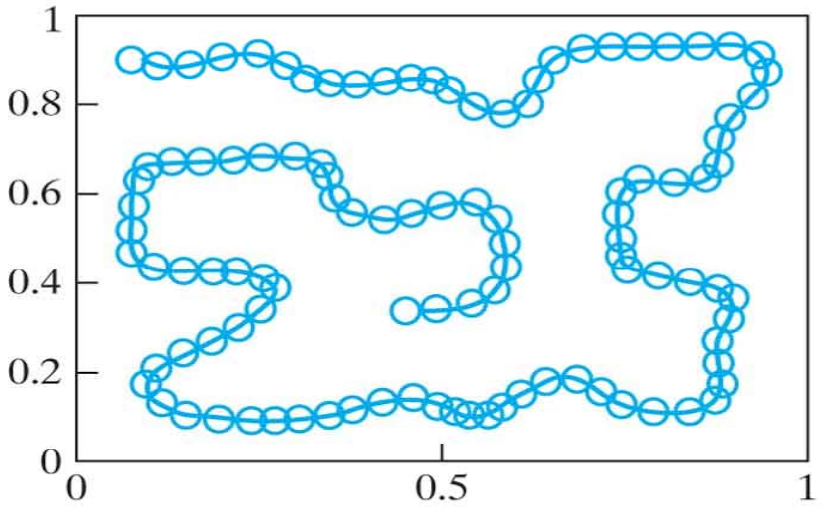


(a) Input distribution



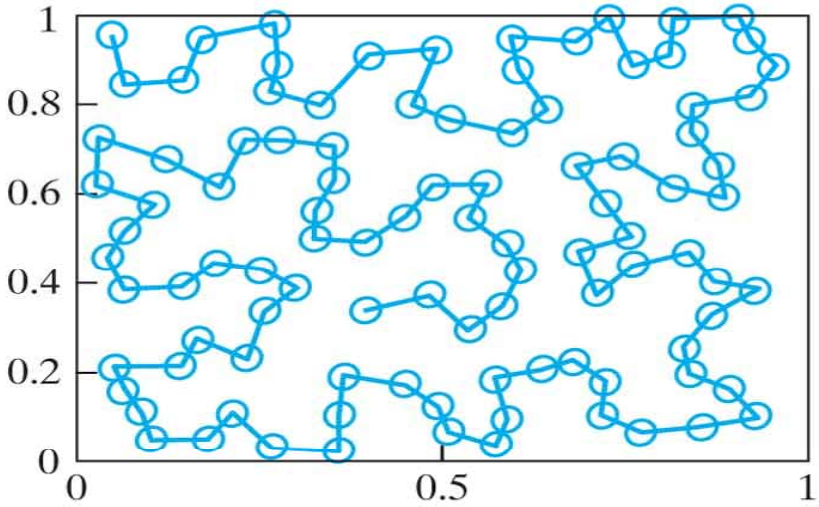
Time = 0

(b) Initial weights



Time = 50 K

(c) Ordering phase

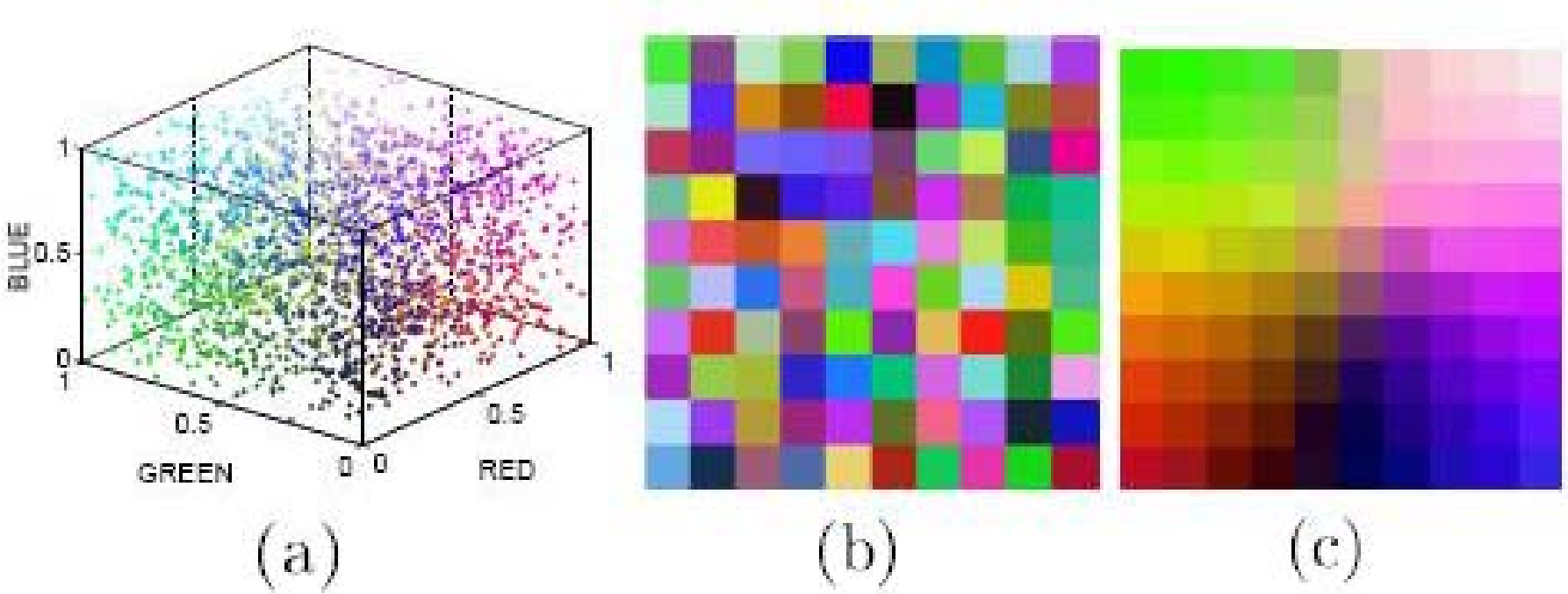


Time = 100 K

(d) Converging phase

例3 RGB color grouping based on SOM

a) Input space
b) Initial weights
c) Final weights



思考题

1. SOM基本模型的结构？各层节点如何设置？代表的意义？
2. SOM网络的学习，学习内容？
3. SOM网络的学习原理？如何体现“竞争”？
4. 如何使用SOM网络进行高维数据的低维可视化？