神经网络
Neural Networks
第四章

自组织网络
—自组织映射

史忠植

中国科学院计算技术研究所
http://www.intsci.ac.cn/

内容提要

- 4.1 概述
- 4.2 自组织特征映射神经网络结构
- 4.3 自组织特征映射网络的学习算法
- 4.4 自组织特征映射网络的应用
- 4.5 自组织网络学习算法的MATLAB实现

概　述

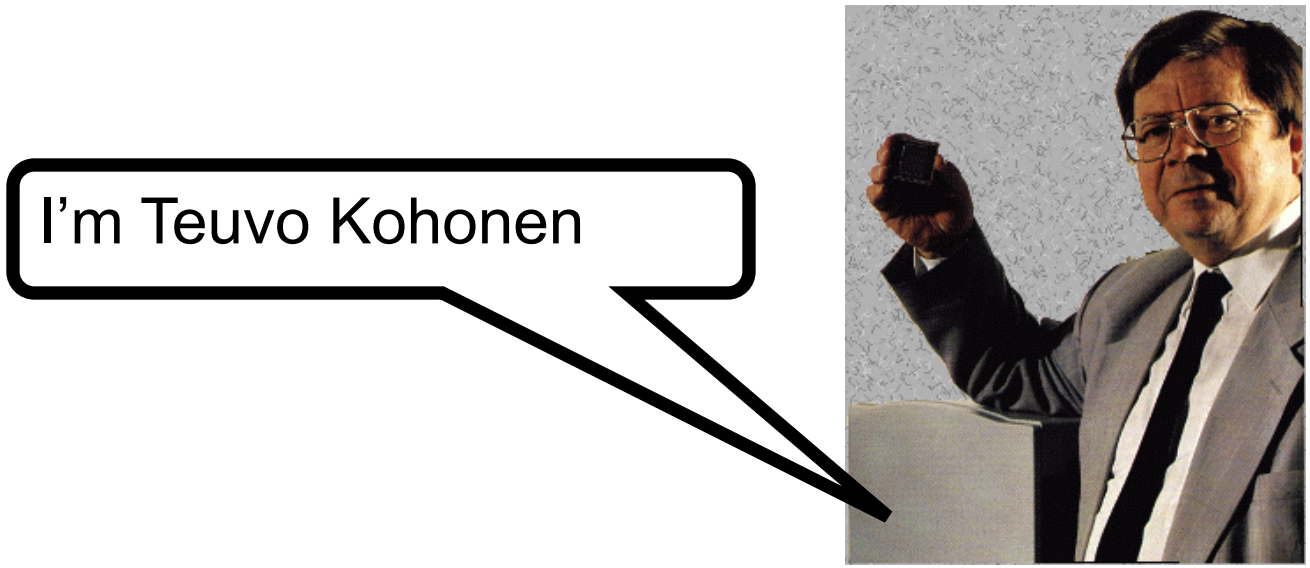
- 自组织神经网络，又称为自组织竞争神经网络，特别适合于解决模式分类和识别方面的应用问题。
- 自组织神经网络属于前向神经网络类型，采用无导师学习算法，
- 自组织特征映射神经网络不仅能够像自组织竞争神经网络一样学习输入的分布情况，而且可以学习神经网络的拓扑结构。

概 述

- 自组织竞争神经网络类型
  - 自组织特征映射(self-Organizing Map, SOM)网络
  - 自适应共振理论(Adaptive Resonance Theory, ART)网络
  - 对传(Counter Propagation, CP)网络
  - 协同神经网络(Synergetic Neural Network . SNN)

自组织特征映射神经网络结构

- 由芬兰学者Teuvo Kohonen于1981年提出



- 基本上为输入层和映射层的双层结构，映射层的神经元互相连接，每个输出神经元连接至所有输入神经元
- Kohonen的思想在本质上是希望解决有关外界信息在人脑中自组织地形成概念的问题。

SOM网的生物学基础

Kohonen认为人的大脑有如下特点：

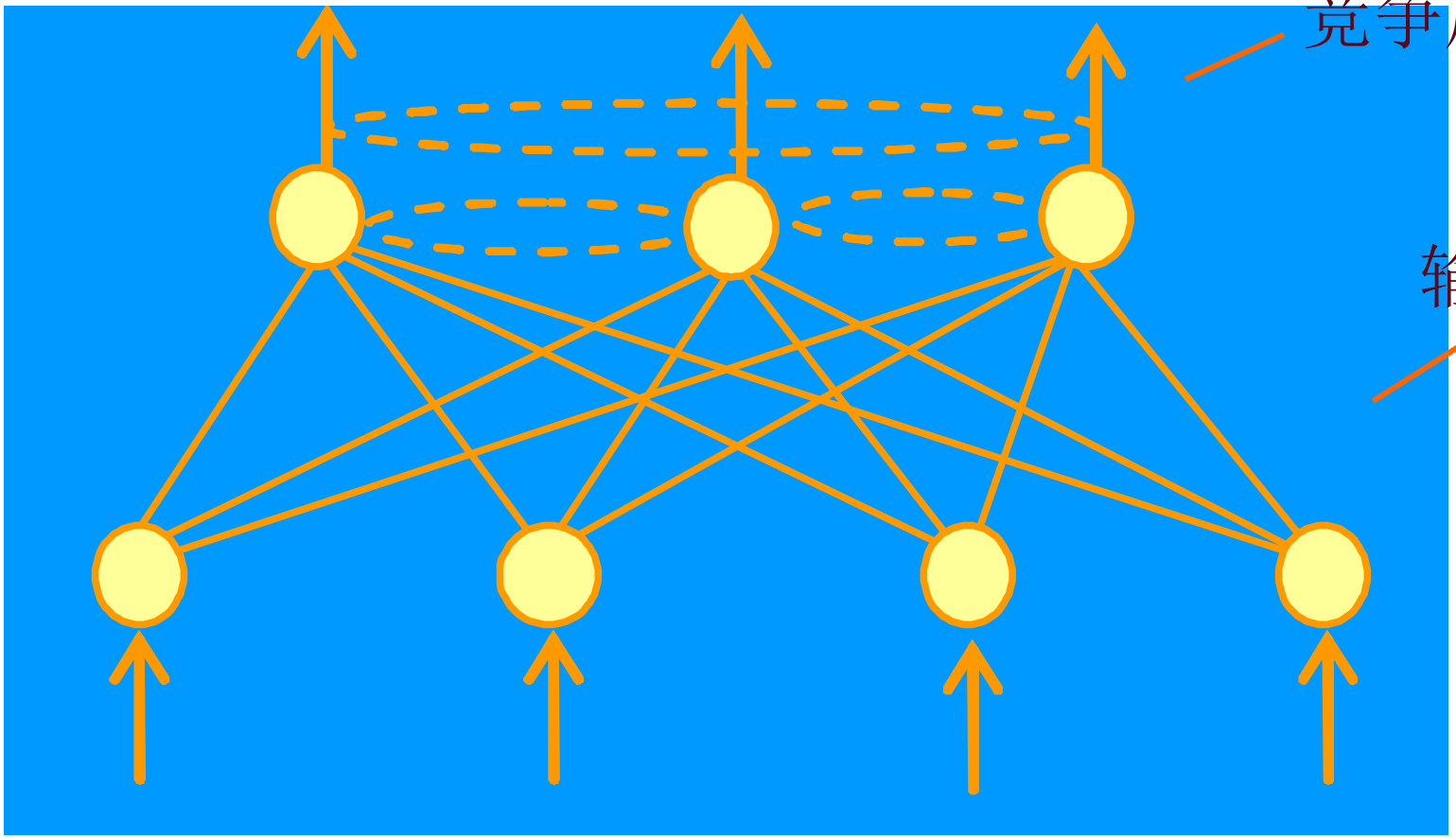
1. 大脑的神经元虽然在结构上相同，但是它们的排序不同。排序不是指神经元位置的移动，而是指神经元的有关参数在神经网络受外部输入刺激而识别事物的过程中产生变动。
2. 大脑中神经元参数在变动之后形成特定的参数组织；具有这种特定参数组织的神经网络对外界的特定事物特别敏感。
3. 根据生物学和神经生理学，大脑皮层分成多种不同的局部区域，各个区域分别管理某种专门的功能，比如听觉、视觉、思维等。
4. 大脑中神经元的排序受遗传决定，但会在外界信息的刺激下，不断接受传感信号，不断执行聚类过程，形成经验信息，对大脑皮层的功能产生自组织作用，形成新功能。

SOM网的生物学基础

生物学研究的事实表明，在人脑的感觉通道上，神经元的组织原理是有序排列。因此当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，而且类似的外界信息在对应区域是连续映象的。

对于某一图形或某一频率的特定兴奋过程，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。

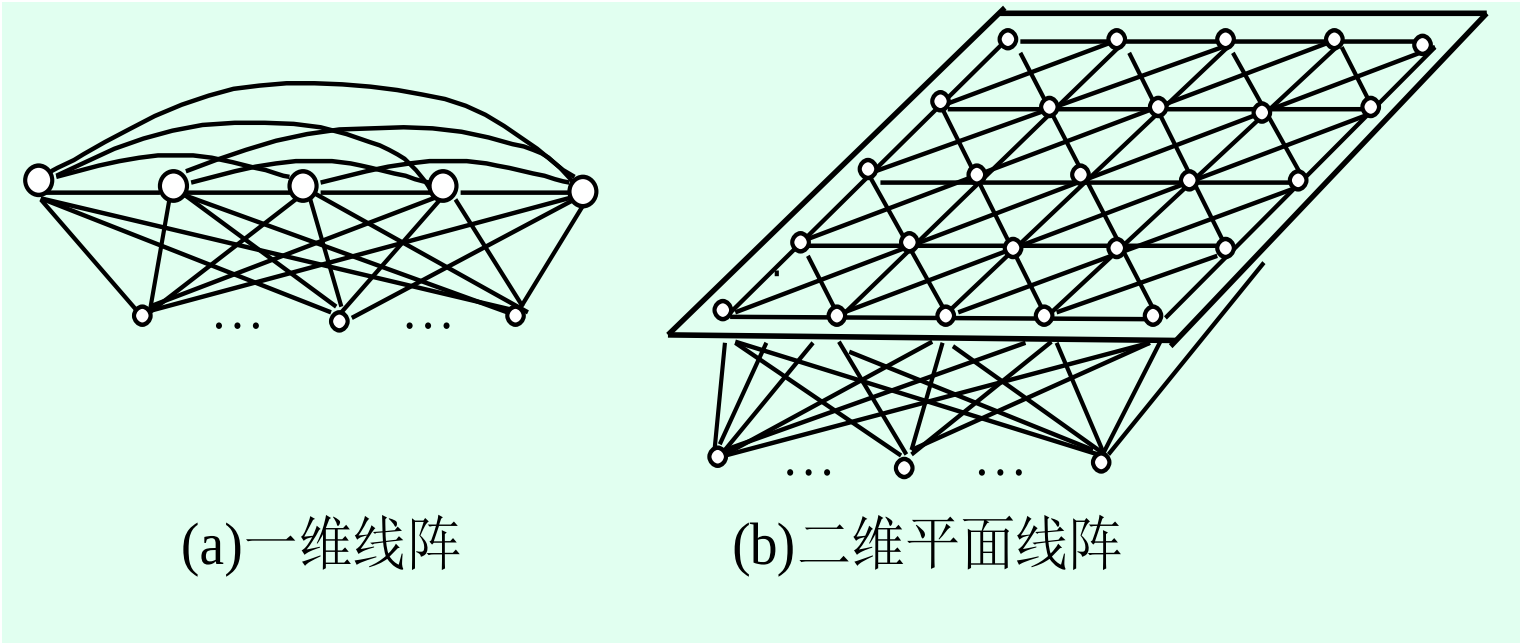
自组织特征映射神经网络结构



SOM神经网络结构

SOM网的拓扑结构

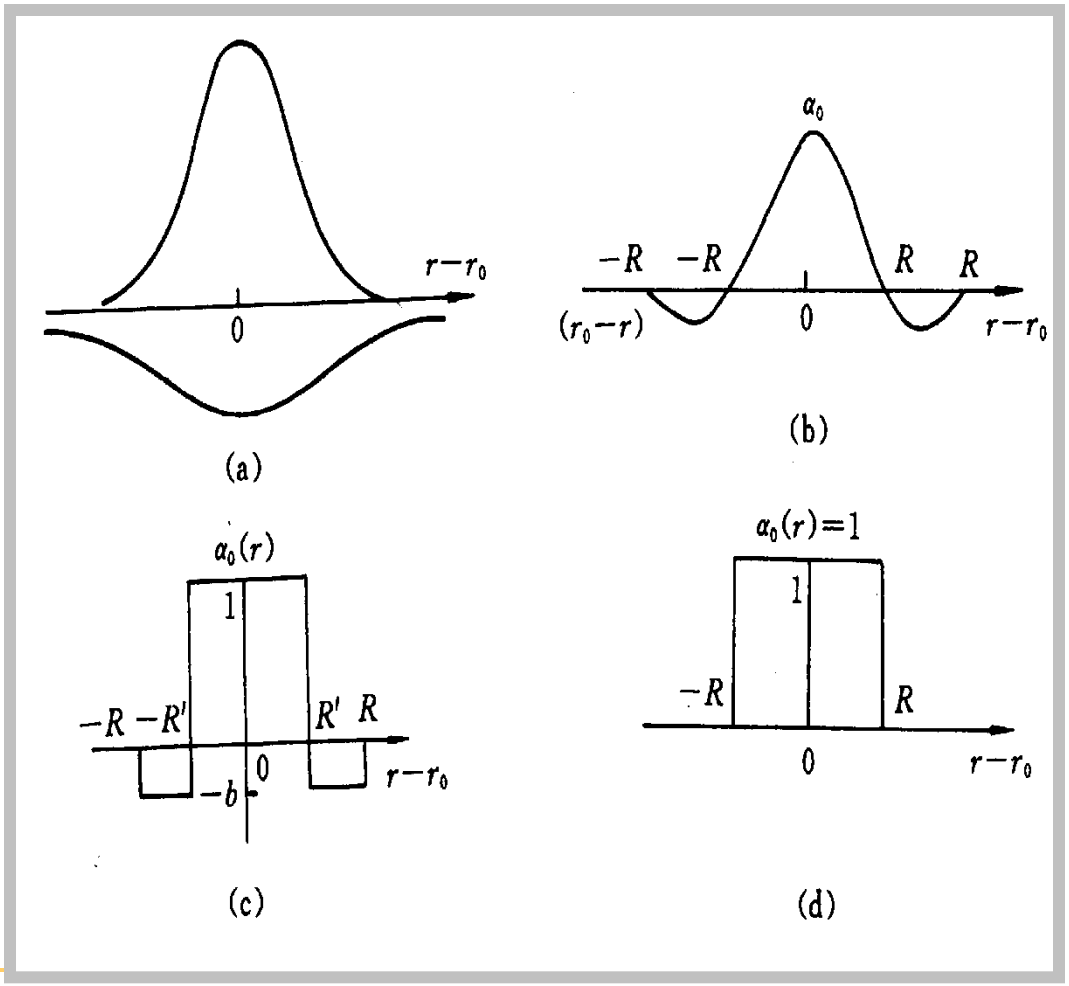
SOM网共有两层，输入层模拟感知外界输入信息的视网膜，输出层模拟做出响应的大脑皮层。



(a)一维线阵　　(b)二维平面线阵

SOM网的权值调整域

SOM网的获胜神经元对其邻近神经元的影响是由近及远，由兴奋逐渐转变为抑制，因此其学习算法中不仅获胜神经元本身要调整权向量，它周围的神经元在其影响下也要程度不同地调整权向量。这种调整可用三种函数表示：



SOM网的权值调整域

以获胜神经元为中心设定一个邻域半径，该半径圈定的范围称为优胜邻域。在SOM网学习算法中，优胜邻域内的所有神经元均按其离开获胜神经元的距离远近不同程度地调整权值。

优胜邻域开始定得很大，但其大小随着训练次数的增加不断收缩，最终收缩到半径为零。

自组织特征映射网络的学习算法

- 自组织特征映射学习算法原理
  - Kohonen自组织特征映射算法，能够自动找出输入数据之间的类似度，将相似的输入在网络上就近配置。因此是一种可以构成对输入数据有选择地给予响应的网络。
- 类似度准则
  - 欧氏距离

$$d_j = \sqrt{\sum_{i=1}^{n} (x_i - w_{ij})^2}$$

自组织特征映射网络的学习算法

- 自组织特征映射学习算法步骤
  - (1)网络初始化
    - 用随机数设定输入层和映射层之间权值的初始值
  - (2)输入向量
    - 把输入向量输入给输入层
  - (3) 计算映射层的权值向量和输入向量的距离
    - 映射层的神经元和输入向量的距离，按下式给出

$$d_j = \sqrt{\sum_{i=1}^{n} (x_i - w_{ij})^2}$$

自组织特征映射网络的学习算法

- 自组织特征映射学习算法步骤
  - **(4)** 选择与权值向量的距离最小的神经元
    - 计算并选择使输入向量和权值向量的距离最小的神经元，把其称为胜出神经元并记为 j^* ，并给出其邻接神经元集合。
  - **(5)**调整权值
    - 胜出神经元和位于其邻接神经元的权值，按下式更新：

$$\Delta w_{ij} = \eta \, \mathrm{h}(j, j^*)(x_i - w_{ij})$$

$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}$$

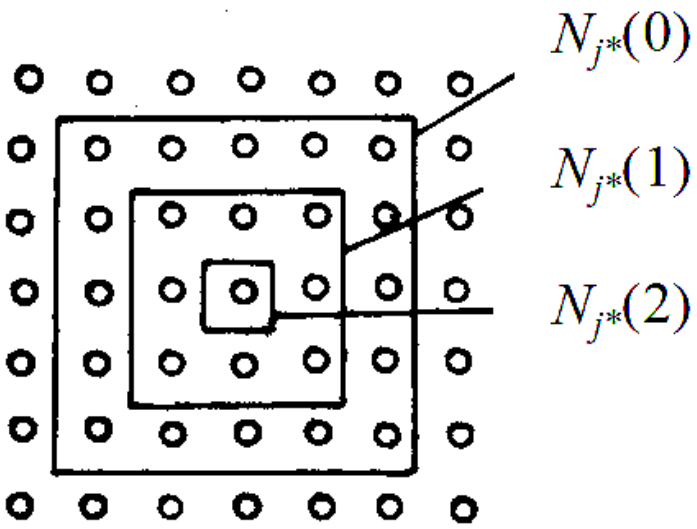
  - **(6)**是否达到预先设定的要求如达到要求则算法结束，否则返回**(2)**，进入下一轮学习

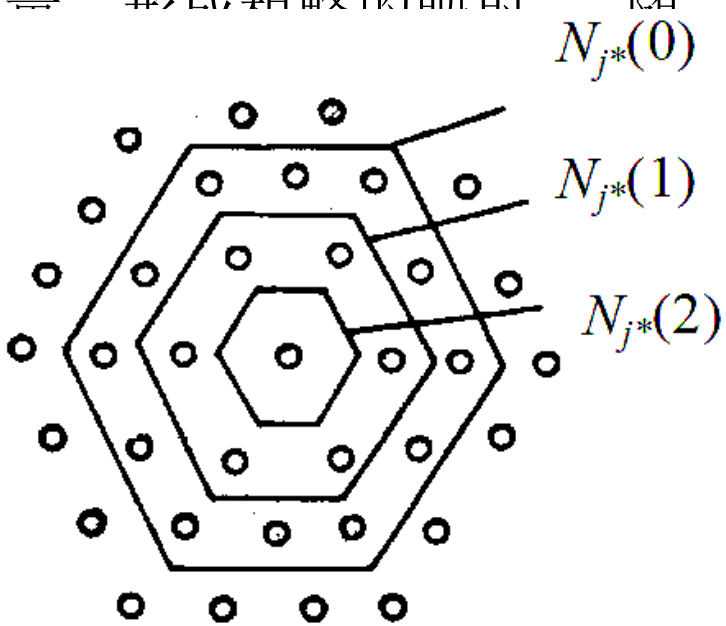
自组织特征映射网络的学习算法

- 邻域函数

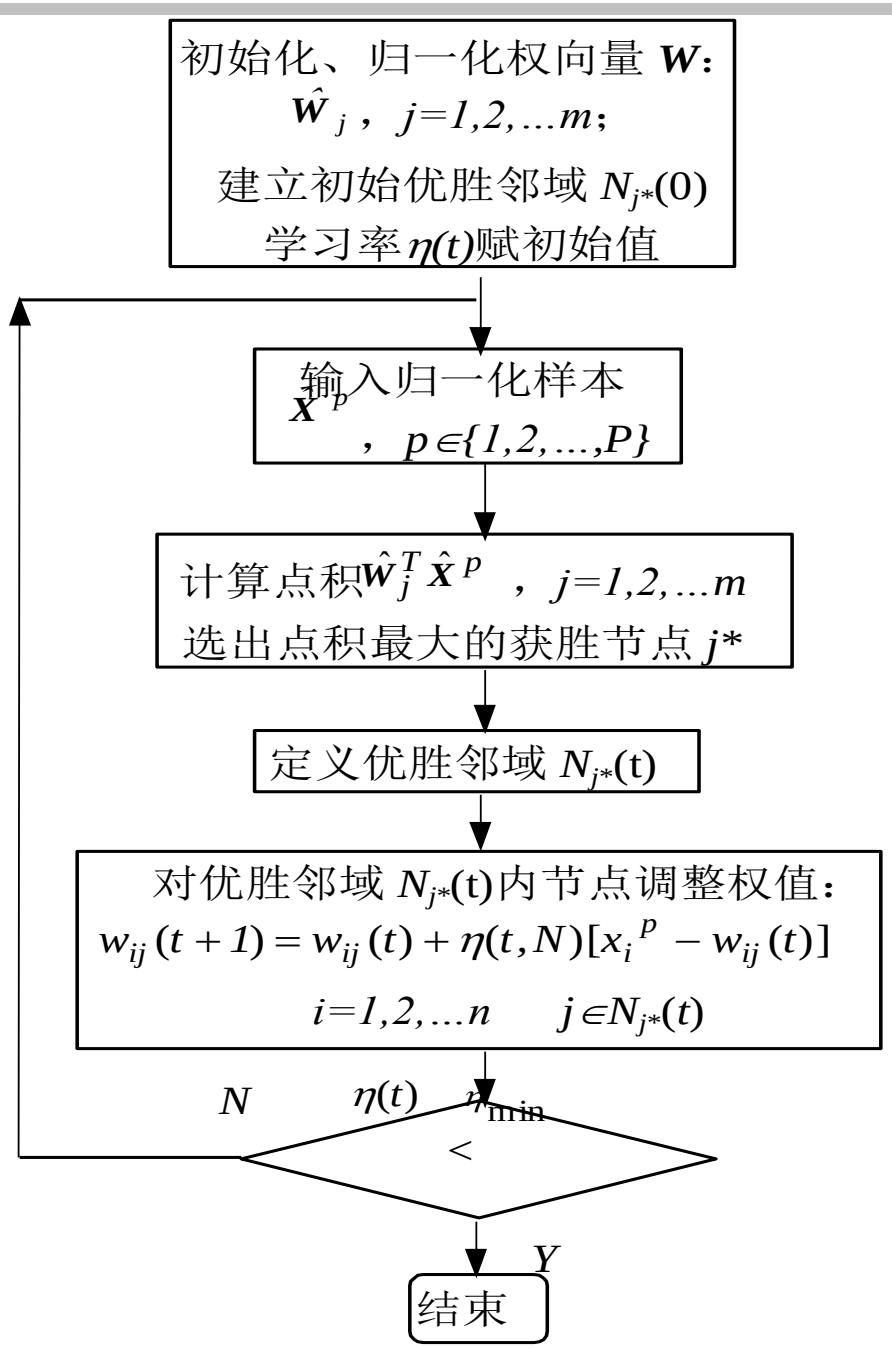
$$h(j, j^*) = \exp\left(-\frac{|j - j^*|^2}{\sigma^2}\right)$$

  - 由邻域函数可以看到，以获胜神经元为中心设定了一个邻域半径，称为**胜出邻域**。学习初期，胜出神经元和其附近的神经元全部接近当时的输入向量，形成粗略的映射，随





Kohonen学习算法程序流程



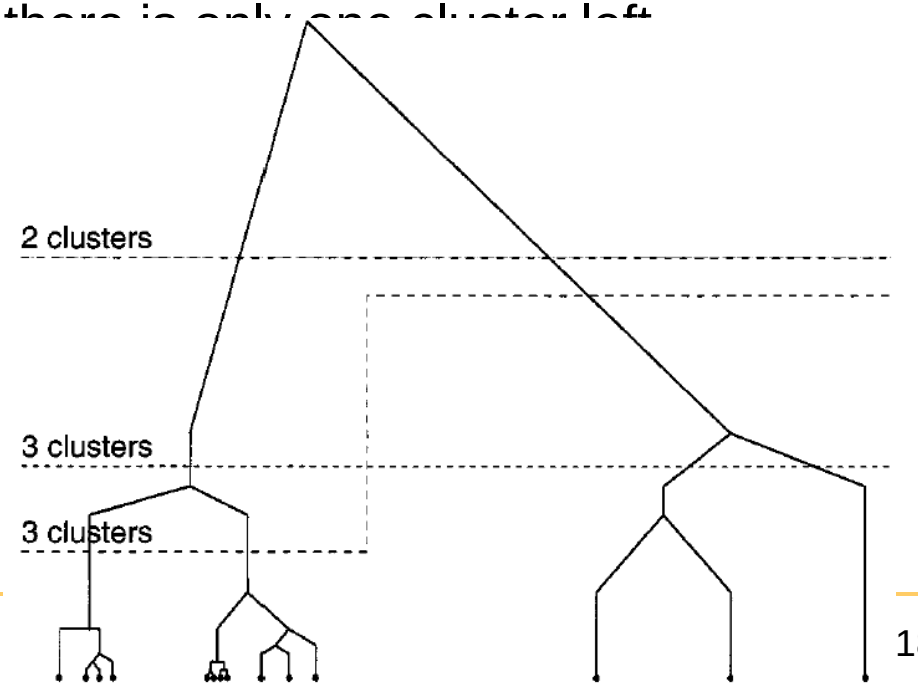
Hierarchical and Partitive Approaches

- Partitive algorithm
  - Determine the number of clusters.
  - Initialize the cluster centers.
  - Compute partitioning for data.
  - Compute (update) cluster centers.
  - If the partitioning is unchanged (or the algorithm has converged), stop; otherwise, return to step 3
- k-means error function
  - To minimize error function

$$E = \sum_{k=1}^{C} \sum_{\boldsymbol{x} \in Q_k} \|\boldsymbol{x} - \boldsymbol{c}_k\|^2$$

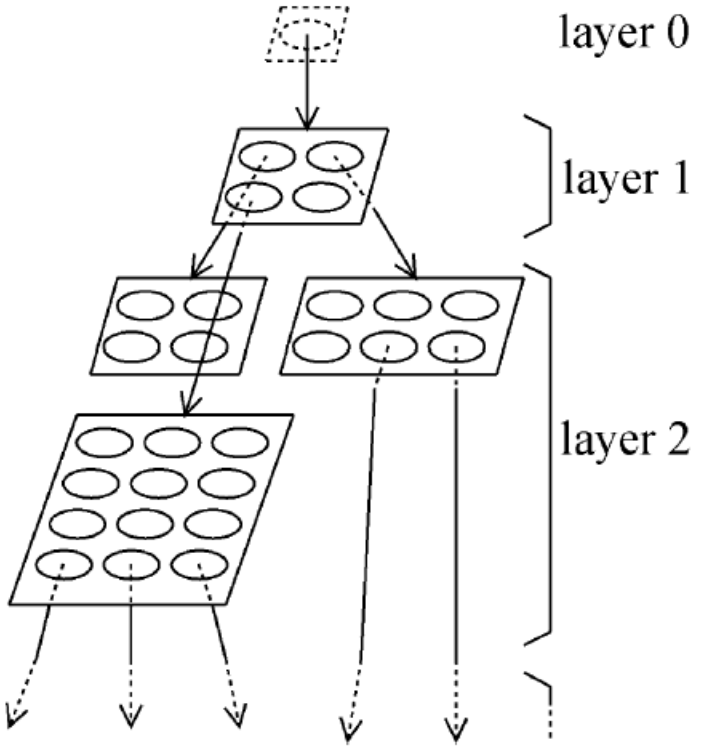
Hierarchical and Partitive Approaches

- Hierarchical clustering algorithm (Dendrogram)
  - Initialize: Assign each vector to its own cluster
  - Compute distances between all clusters.
  - Merge the two clusters that are closest to each other.
  - Return to step 2 until there is only one cluster left.
- Partition strategy
  - Cut at different level



Hierarchical SOM

- GHSOM – Growing Hierarchical Self-Organizing Map
  - grow in size in order to represent a collection of data at a particular level of detail



自组织网络学习算法的MATLAB实现

- MATLAB中自组织神经网络的重要函数和基本功能

函 数 名	功　能
newsom()	创建一个自组织特征映射神经网络
plotsom()	绘制自组织特征映射网络的权值矢量
vec2ind()	将单值矢量组变换成下标矢量
compet()	竞争传输函数
midpoint()	中点权值初始化函数
learnsom()	自组织特征映射权值学习规则函数

自组织网络学习算法的MATLAB实现

- MATLAB中自组织神经网络的重要函数和基本功能
  - newsom()
    - 功能 创建一个自组织特征映射网络函数
    - 格式 net = newsom(PR，[D1，D2，...]，TFCN，DFCN，OLR，OSTEPS，TLR，TND)
    - 说明 net为生成的新BP神经网络；PR为网络输入矢量取值范围的矩阵[Pmin Pmax]；[D1，D2，...]为神经元在多维空间中排列时各维的个数；TFCN为拓扑函数，缺省值为hextop；DFCN为距离函数，缺省值为linkdist；OLR为排列阶段学习速率，缺省值为0.9；OSTEPS为排列阶段学习次数，缺省值为1000；TLR为调整阶段学习速率，缺省值为0.02，TND为调整阶段领域半径，缺省值为1。

自组织网络学习算法的MATLAB实现

- plotsom()
  - 功能　绘制自组织特征映射网络图的权值向量函数
  - 格式
    - (1) plotsom(pos)
    - (2) plotsom(W, D, ND)
  - 说明　式中pos是网络中各神经元在物理空间分布的位置坐标矩阵；函数返回神经元物理分布的拓扑图，图中每两个间距小于1的神经元以直线连接；W为神经元权值矩阵；D为根据神经元位置计算出的间接矩阵；ND为领域半径，缺省值为1；函数返回神经元权值的分布图，图中每两个间距小于ND的神经元以直线连接。

自组织网络学习算法的MATLAB实现

- yec2ind()
  - 功能　将单值向量组变换成下标向量
  - 格式　ind = vec2ind(vec)
  - 说明　式中，vec为m行n列的向量矩阵x，x中的每个列向量i，除包含一个1外，其余元素均为0，　ind为n个元素值为1所在的行下标值构成的一个行向量。

自组织网络学习算法的MATLAB实现

- 例1 人口分类是人口统计中的一个重要指标，现有1999共10个地区的人口出生比例情况如下：
  - 出生男性百分比分别为：0.5512 0.5123 0.5087 0.5001 0.6012 0.5298 0.5000 0.4965 0.5103 0.5003;
  - 出生女性百分比分别为：0.4488 0.4877 0.4913 0.4999 0.3988 0.4702 0.5000 0.5035 0.4897 0.4997

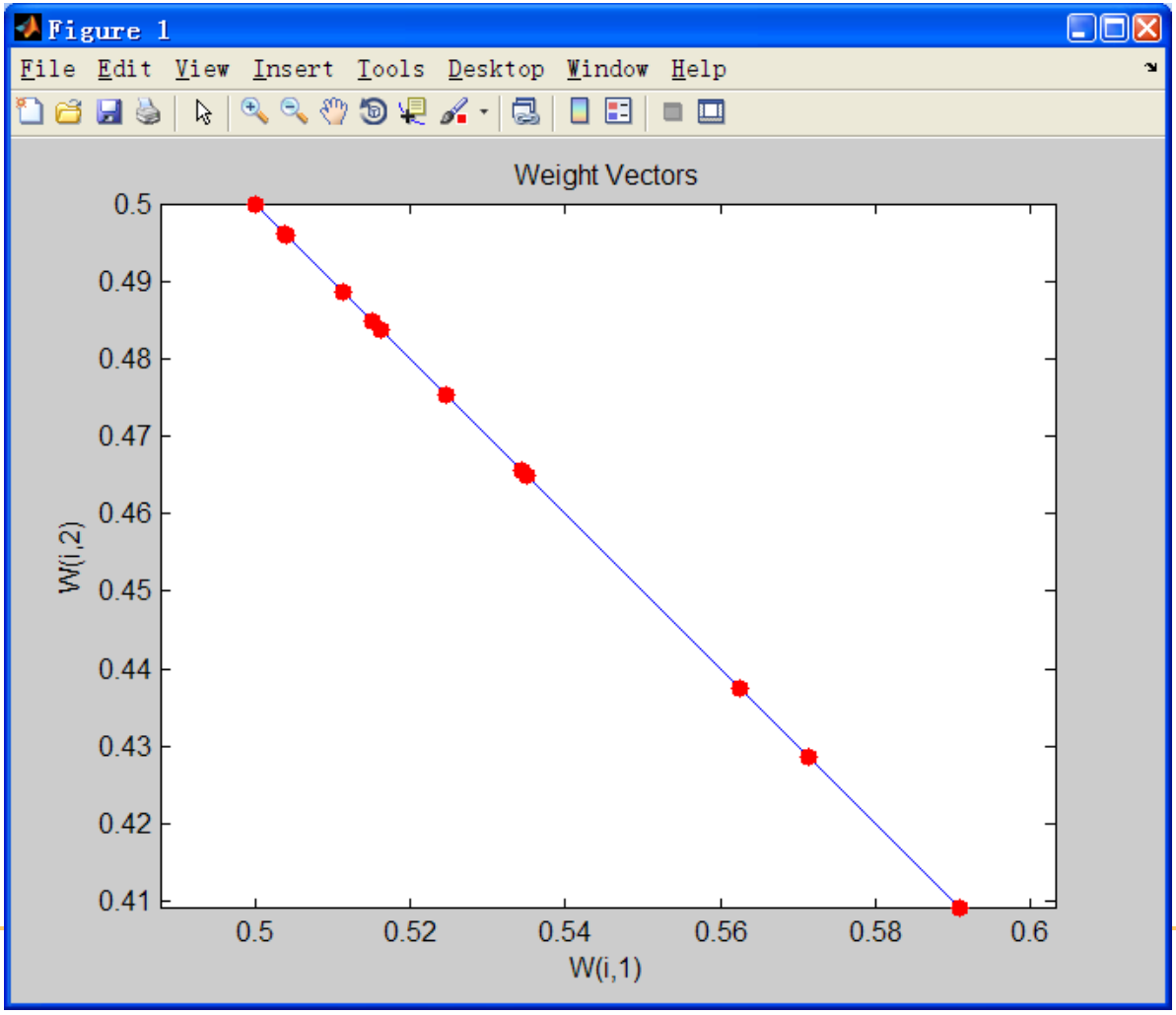
自组织网络学习算法的MATLAB实现

- 例1 源程序

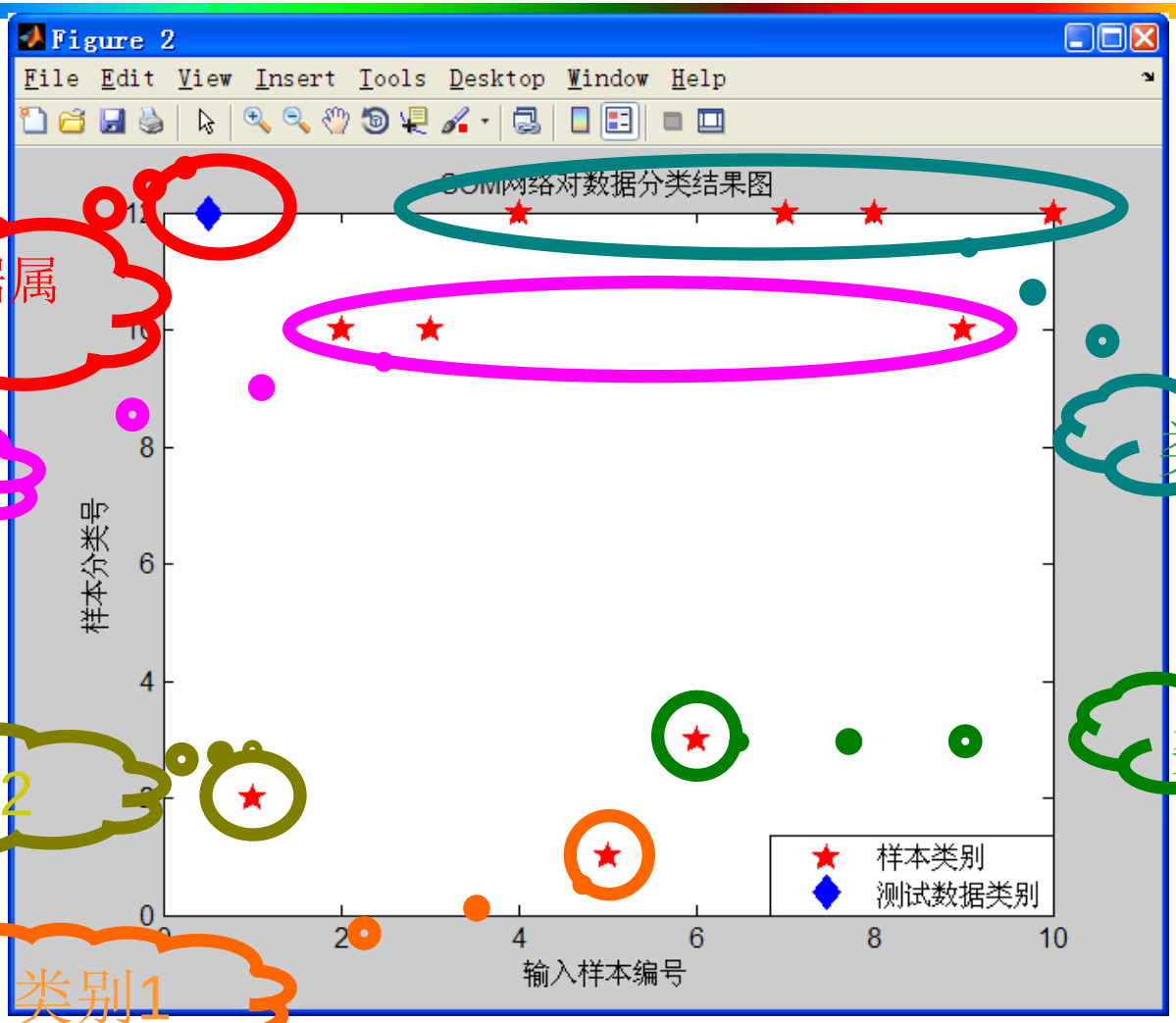
建立一个自组织神经网络对上述数据分类，给定某个地区的男、女出生比例分别为0.5，0.5，测试训练后的自组织神经网络的性能，判断其属于哪个类别。
解答：MATLAB代码如下：

```
P=[0.5512    0.5123   0.5087   0.5001   0.6012   0.5298   0.5000   0.4965
    0.5103   0.5003; 0.4488    0.4877 0.4913   0.4999   0.3988   0.4702
    0.5000   0.5035   0.4897   0.4997];
%创建一个自组织神经网络，[0 1;0 1]表示输入数据的取值范围在[0，1]之
%间，[3，4]表示竞争层组织结构为3 4，其余参数取默认值
net=newsom([0 1;0 1],[3 4]);
net.trainParam.epochs=500;
net=init(net);
net=train(net,P);
y=sim(net,P);
%获取训练后的自组织神经网络的权值
w1=net.IW{1,1};
%绘出训练后自组织神经网络的权值分布图
plotsom(w1,net.layers{1}.distances);
```

例1 SOM网络权值分布图



例1 SOM网络数据分类图



自组织语义图

- 利用词与词在文档中的上下文关系，将词表示成一个向量，然后用表示词的向量作为SOM网络的输入，聚类，输出形成一个词汇类别图（Word category map）
- 在这个图中，意义相近的词聚在一起，组成一个词类，词在词汇类别图中的位置可以通过快速Hash的方法查到。

自组织语义图：一种将词向量化的方法

- 在一个文档集中考虑词与词之间的上下文关系。设$l_i^{(d)}$表示相对于第i个词位移为d的位置上出现的词集合（有可能出现多次），例如，$l_i^{(1)}$表示第i个词的所有前趋邻接词
- 用向量x_i表示第i个词，对位移集$\{d_1,\ldots,d_N\}$：

$$x_i = \left[x_i^{(d_i)},\ldots,x_i^{(d_N)}\right]^T, \qquad x_i^{(d)} = \frac{1}{\left|I_i^{(d)}\right|}\sum_{k\in I_i^{(d)}} e_k$$

其中$\left|I_i^{(d)}\right|$表示$I_i^{(d)}$中词的数量

一般地，为计算简单，只取d=1和d=-1

自组织语义图：另一种对中文词汇向量化的方法

- 对每一个词，在文档集合中出现该词的时候会伴随一些修饰词。因此可以用修饰词来表示该词，提供该词的一些语义信息
- 例如对名词“大学”，会出现一些修饰词如“本科”、“重点”、“合格”等，则定义，大学={本科，重点，合格，…}

自组织语义图：另一种对中文词汇向量化的方法

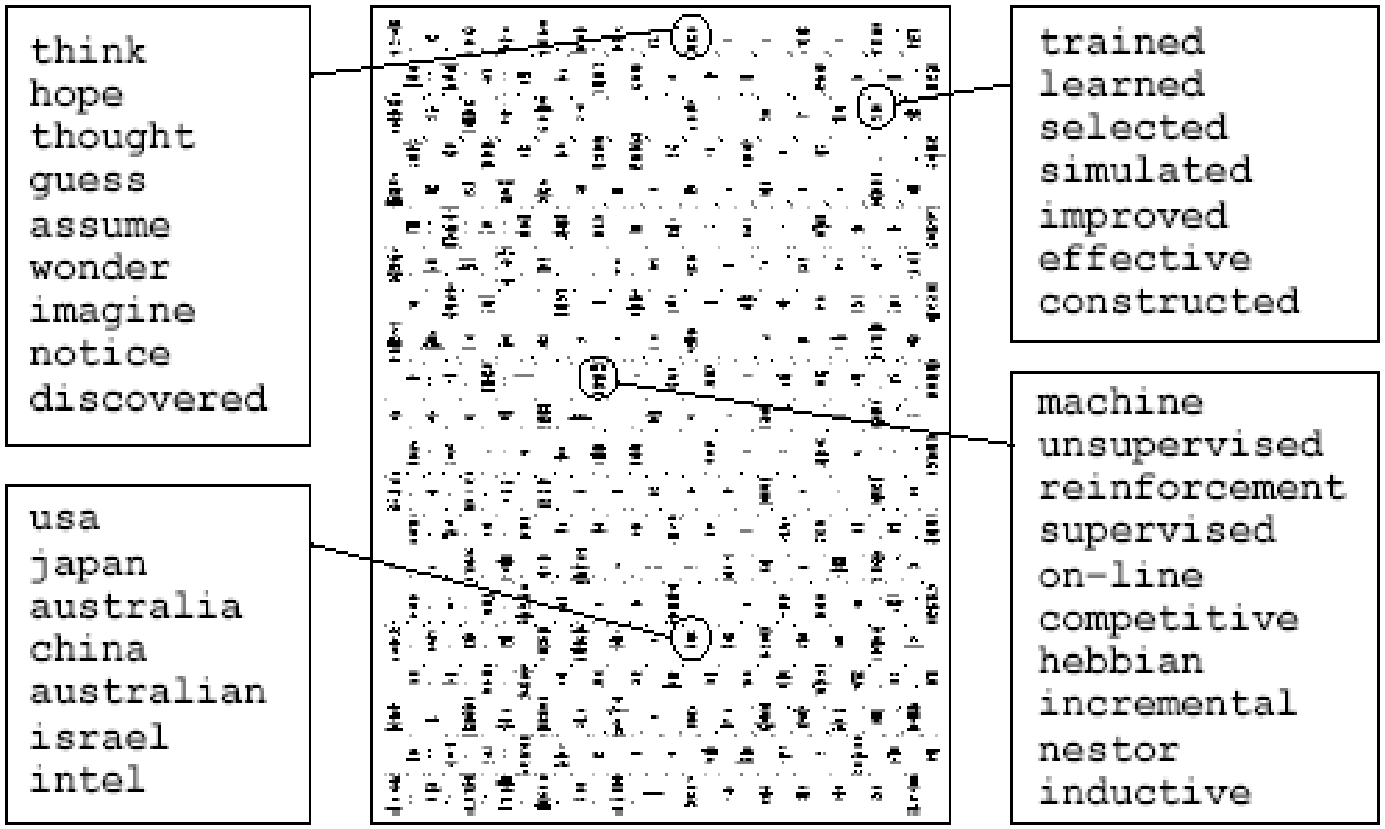
- 一般地，词w_i(i=1,2,…,N)为：$w_i = \{a_1^{(i)}, a_2^{(i)}, ..., a_{n_i}^{(i)}\}$
  其中n_i表示修饰词个数
- 定义词w_i的向量表示为$V(w_i) = [d_{i1}, d_{i2}, ..., d_{iN}]^T$
- 其中，

$$d_{ij} = \frac{(n_i - c_{ij}) + (n_j - c_{ij})}{n_i + n_j - c_{ij}} \quad , \quad i \neq j \quad ; \quad d_{ij} = 0 \quad i = j$$

c_{ij}表示词w_i和词w_j的修饰词集合中都出现的修饰词个数

- 将词用这种向量表示后，作为一个SOM网络的输入，可以聚类形成中文语义图。

自组织语义图：示例


think
hope
thought
guess
assume
wonder
imagine
notice
discovered
usa
japan
australia
china
australian
israel
intel
trained
learned
selected
simulated
improved
effective
constructed
machine
unsupervised
reinforcement
supervised
on-line
competitive
hebbian
incremental
nestor
inductive

利用SOM进行文本聚类：预处理

- 去掉非文本信息
- 去掉在整个文档集合中出现次数小于50次的词
- 去停用词
- 经过上述处理后，词的个数由 1 127 184 减少为 63 773

利用SOM进行文本聚类：Word category map

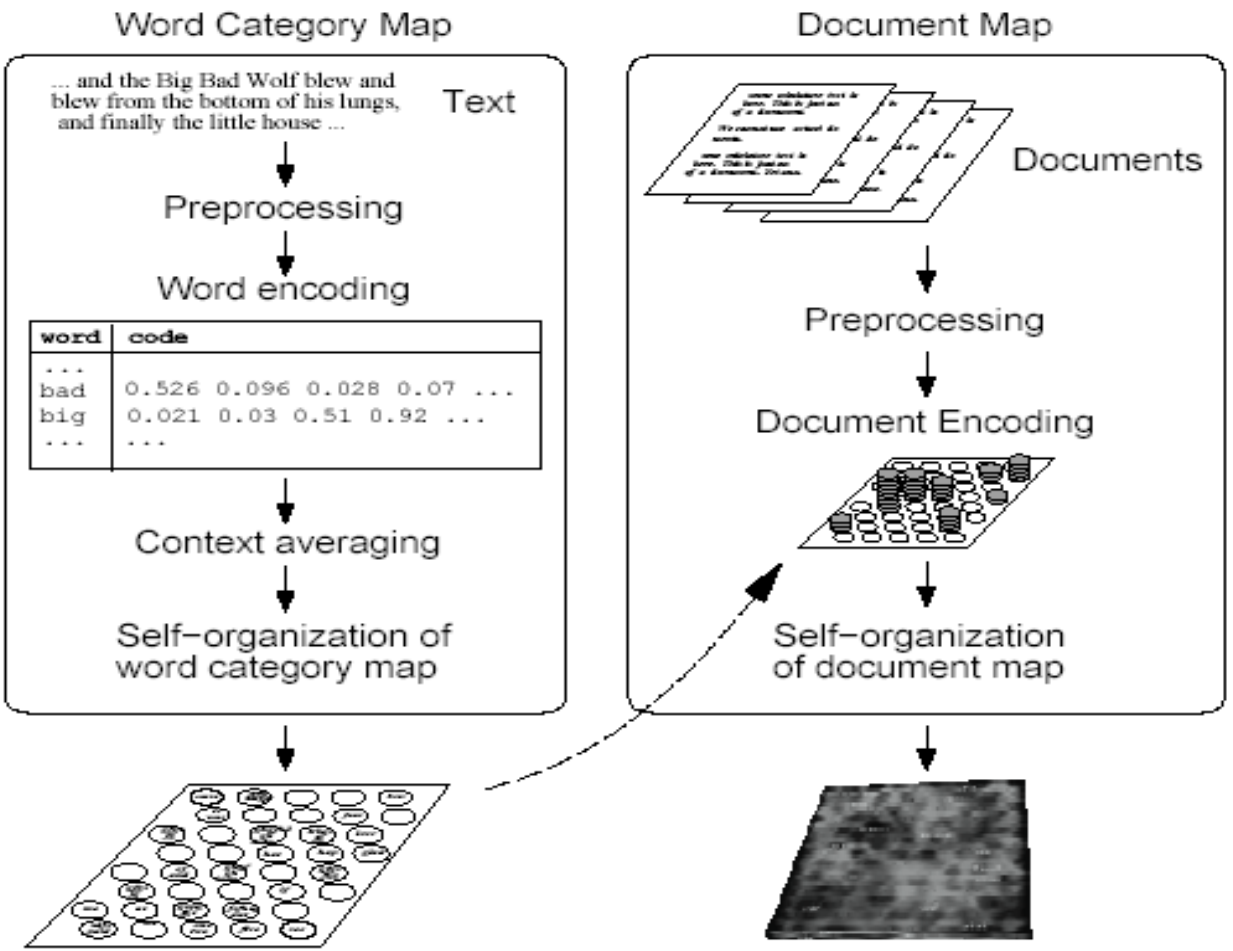
- 将词表示成为一个180维的向量，作为一个SOM网络的输入，进行聚类
- 最终产生 13 432 个词类单元

（ 63 773 → 13 432）

利用SOM进行文本聚类：Document map

- 利用上面产生的词类将文档向量化后，每篇文档表示为一个 13 432 维的向量，再利用随机映射(Random mapping method)的降维方法，向量维数减少到 315 维
- 将这 315 维的向量作为一个 SOM 的输入
- 相关的结果可以参见 http://websom.hut.fi/websom/

利用SOM进行文本聚类



SOM的特点

- 自组织映射（Self-organizing Maps，SOM）算法是一种无导师学习方法
- 具有良好的自组织
- 可视化
- 得到了广泛的应用和研究。

Thank You

Intelligence Science

http://www.intsci.ac.cn/