说明文档

一．Batch SOM 简介

自组织特征映射模型被称为 Self-organizing map，由芬兰学者 Teuvo Kohonen 于 1981 年提出。该网络是一个由全互连的神经元阵列形成的无教师自组织自学习网络。Kohonen 认为，处于空间中不同区域的神经元有不同的分工，当一个神经网络接受外界输入模式时，将会分为不同的反应区域，各区域对输入模式具有不同的响应特征。自组织神经网络模型能将高维数据投影到低维数组中。这种非线性投影生成的二维特征网络可以帮助分析和检测输入空间的特征。其中，Batch SOM 所有的网络权值在每个训练周期的末尾更新，所以可以应用在并行算法中。
在 Batch SOM 神经网络中，权重公式为：

$$W_k(t_f) = \frac{\sum_{t'=t_0}^{t'=t_f} h_{ck}(t')x(t')}{\sum_{t'=t_0}^{t'=t_f} h_{ck}(t')}$$

网络中共有 K 个神经元，t_0和 t_f分别代表训练周期的开始和结束。

其中h_{ck}为标准高斯邻域方程：

$$h_{ck}(t) = \exp\left(-\| r_k - r_c \|^2/\sigma(t)^2\right)$$

r_k和r_c是坐标点，$\sigma(t)$是网络宽度，$\sigma(t)$将在每个周期数据输入时收敛直至最后到聚焦到一个神经元。这表征了 SOM 神经网络的自组织特性：每个输入向量的出现将调整获胜神经元，并使得邻域神经元更靠近输入向量，最终这些神经元将反馈输入向量的似然分布。

在每次网络获得输入向量$x(t)$时，获胜神经元由以下公式决定：

$$d_k(t) = \| x(t) - W_k(t_0) \|^2$$

$$d_c(t) = \min_k d_k(t)$$

其中,获胜点为 c，$W_k(t_0)$为前一个周期网络经过调整的神经元。

二．算法流程

参数	
/opt/hadoop-1.2.1/input	输入向量路径
/opt/hadoop-1.2.1/output	输出路径
/opt/hadoop-1.2.1/Neuron	初始神经元输入路径
numberOfNeuron	神经元个数 K
numberOfAttribute	网络维度 2
neighbourhoodSize	网络大小 100
decay	邻域递减速率 1.5
iter	迭代次数 3

表 1.代码参数

流程图

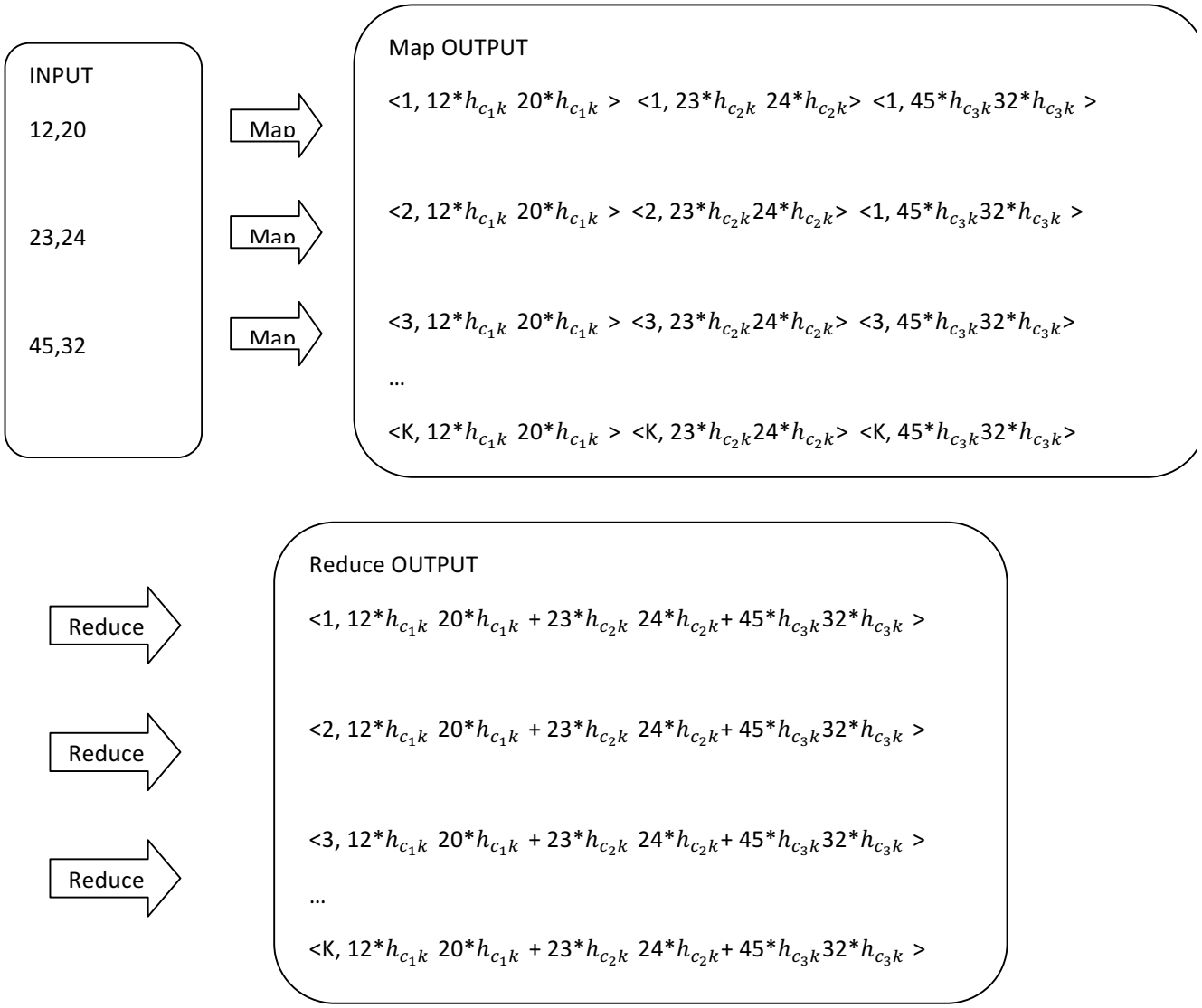

图 1.MapReduce 过程

三．算法解析

先将初始神经元读入 centerlist 容器中，进入 map 过程。

每行 INPUT 对应一个 map 过程。

1.计算与输入向量欧氏距离最近的神经元，获得被激活神经元r_c；

$$d_k(t) = \| x(t) - W_k(t_0) \|^2$$

$$d_c(t) = \min_k d_k(t)$$

2.根据高斯邻域方程

$$h_{ck}(t) = \exp(-\| r_k - r_c \|^2 / \sigma(t)^2)$$

获得激活神经元周围神经元的增益情况。

先算出分子 tempup $=-\| r_k - r_c \|^2$,分母 tempwidth= $\sigma(t)^2$，其中

$\sigma(t) =$ neighbourhoodSize*exp(-t/decay)，在每个 map 过程中时间 t 加一，达到网络收敛效果；

3.得到h_{ck}= exp(tempUp/tempWidth)；

4.将权重公式$W_k(t_f)$的分子$h_{ck}(t')x(t')$和分母$h_{ck}(t')$送入 reduce；

每个神经元对应一个 reduce 过程。

1. 将每个从 map 中获得的键值对按照 key 值分解，获得每个神经元对应的分子$h_{ck}(t')x(t')$和分母$h_{ck}(t')$；
2. 将权重公式中的分子和分母累加，即：$\sum_{t'=t_0}^{t'=t_f} h_{ck}(t')x(t')$， $\sum_{t'=t_0}^{t'=t_f} h_{ck}(t')$；
3. 得到一个神经元经过多个向量刺激的结果，即：
   $W_k(t_f) = \frac{\sum_{t'=t_0}^{t'=t_f} h_{ck}(t')x(t')}{\sum_{t'=t_0}^{t'=t_f} h_{ck}(t')}$；
4. 逐个输出 K 个神经元至 centerlist 中。

进行下一次迭代，重新开始 mapreduce。