基于变异系数的SOM算法在多维分析中的研究

摘要：自组织映射SOM算法是一种无导师学习方法,具有良好的,自组织,可视化等特性.论文首先引入SOM算法，介绍SOM算法原理,然后针对欧式距离的特点,将通过变异系数对维度指标赋予权重来改进SOM算法,最后针对SOM降维数据,采用Kmeans算法进行再次聚类,使结果具有可评价性,结果表明:相比没有维度权重的SOM算法,采用带有指标权重的SOM算法具有更好的降维效果。

关键词：降维；自组织映射（SOM）；变异系数；权值更新

# 引言

多维分析中，常规的统计和基本分析为主的可视化系统分析功能不足，针对多维数据，散点图，表格透镜，平行坐标等更适合多维数据的展示，但是当数据维度非常高（10维），各类可视化方法都无法清晰的表示所有的数据细节，因此采用降维方法把多维数据映射至低维数据。

线性降维方法主要有主成份分析（PCA）多维尺度分析（MDS）,将原本的多个指标重新组合成新的指标，并使新的指标尽可能保留数据原有的指标信息，应用特点是原始数据可以在二维空间或三维空间用线性结构或其组合来描述数据，否则就需要使用非线性降维方法；非线性降维方法主要有等距映射(Isomap),局部线性嵌入（LLE）等，非线性方法大都是在保持高维数据的某些特性的条件下完成数据在低维的可视化，因此不同特性下的降维方法对于数据的展示的效果是不一致的；神经网络中也有一类降维可视化方法，典型的就是自组织神经网络，是一种具有降维，聚类，可视化的无监督学习算法，通过模拟人脑对信号处理的特点而发展起来的一种人工神经网络。SOM算法将高维的数据映射到网络拓扑结构上，通常是正方形或者六边形来映射数据，而网络的远近用来表示数据的相似程度，达到降维并且聚类的效果；

基于SOM算法,国内外从不同方向提出了多种的改进方法。基于动态改变神经元数目的改进：Alahakoon 提出的 GSOM，GSOM 在初始时，竞争层由 4 个神经元构成正方形结构，在训练过程中，对于每一个输入样本 x，计算其获胜结点 c 的累计误差TE，若 TE 大于预先指定的生长阈值 GT，则在 c 的邻域内找一空闲位置生成一个新结点，若 c 的邻域内无空闲位置，则将 c 的累计误差 TE 分配给其邻域内的结点。其不足是不能按需要方便地在合适的位置生成新结点；王莉等提出的树型动态增长模型 TGSOM，它与 GSOM 的不同在于它可以按需要方便地在任意合适位置生成新结点，克服了 GSOM 的缺点；基于匹配神经元策略的改进:SOFM-CV的思想是：把 SOM 网络的权值都初始化为 (n为输入向量的维数)，每个输入向量 x 要经过如下修正后：αx +(1 −α) n (α 随时间从 0 逐渐增大)，再输入网络。SOM算法和其他算法结合: Xiao 等提出了把 SOM 和微粒群优化(Particle swarm optimization，PSO)算法结合用来对基因数据进行聚类，先用 SOM 算法对基因数据进行聚类，得到一组权值，然后用此权值初始化 PSO 算法，用 PSO算法对此聚类结果进行优化。

SOM算法中的相似性的计算通常采用的是欧式距离,而欧式距离在计算过程中是默认按照维度的权重系数是一样的,也就是维度对于相似性的贡献是一样的，这样对于具有维度权重特点的数据计算其相似性是不利的,同时SOM的降维结果是显示在网络拓扑结构上的，对与数据的解释性是不能很好的表达，因此论文中首先使用som算法对数据进行降维，显示在网络拓扑结构中，然后针对拓扑结构中的二维的数据采用快速聚类（Kmeans）进行聚类，从而可以针对聚类结果分析；

# SOM算法及其改进

## SOM算法基本原理

SOM网络结构一般有两层，由输入层和竞争层组成，竞争层也就是输出层，属于单层神经网络。输入层神经元数为n，n为样本的维数，输出层为个神经元组成的网络拓扑结构，神经元与周围的神经元进行相连来表明其网络关系，而输入层的神经元与输出层的每个神经元之间以权值w相连接。

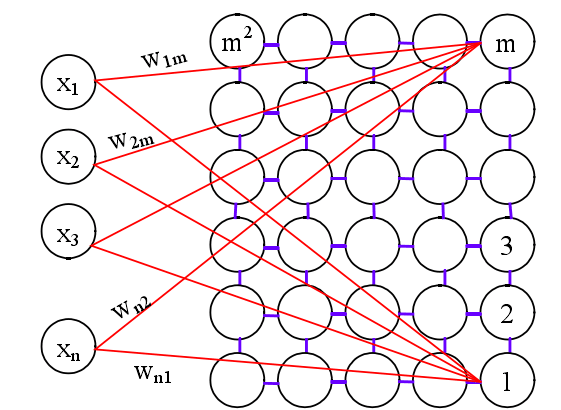


图1 自组织映射网络基本结构

SOM 网络能将任意维输入模式在输出层映射成一维或二维图形，并保持其拓扑结构不变；网络通过对输入模式的反复学习可以使权重向量空间与输入模式的概率分布趋于一致，即概率保持性。网络的竞争层各神经元竞争对输入模式的响应机会，获胜神经元有关的各权重朝着更有利于它竞争的方向调整。自组织映射SOM算法最主要有两个过程：（1）神经元通过竞争学习得到最佳匹配神经元为获胜神经元；（2）更新权值。

### 1.1.1 竞争学习

输出层神经元的获胜是通过计算输入层样本数据与竞争层神经元的相似进行计算的，最大相似度的为获胜神经元，如图1中输入样本的维度为,……,而与1输出神经元与样本相连的权值为，……,因此输入样本转化为，输出神经元转化为，输入样本与输出神经元的相似度计算就转化为与的计算，而获胜神经元就是与，i=1,……m中相似度最大的的神经元。向量相似度的计算常用的方式有欧式距离，标准化欧式距离和余弦相似度等。

### 1.1.2 权值更新

当得到获胜神经元I后，权向量应该得到相应的修改以保证整个学习过程是收敛的，而且并不仅仅是获胜神经元进行调整，每个获胜神经元周围的神经元都应该进行响应的修正，权向量更新权值公式如下：



为当前j神经元第n次迭代的权值，是学习率，为领域函数，是i获胜神经元与z输入样本之间的差距；

是第n次迭代的学习率：为了使收敛更快，学习率开始的时候应该很大，当神经元权值调整到大概位置时，进行小学习率的调整，为初始学习率，一般为1；



优胜领域：根据生物学上所启发的神经网络模型，空间上相邻的神经元的相关性学习可以通过侧反馈与周围神经元的相互作用来实现，为此围绕获胜i神经元设定的一个领域半径，对于优胜领域内的所有的神经元按照距离神经元的距离远近不同程度调整神经元权值，同时优胜领域本身随着迭代次数的增加，半径也不断减少到半径为0，这样会使权值调整的过程是收敛的，常见的领域的形状有正方形，六边形或菱形，而对于其调整的神经元的权值也与神经元之间的间距有关系的，距离获胜神经元近的神经元可以调整更大的权值,为此神经元之间的距离采用欧式距离进行计算，对应不同迭代次数n下的ij神经元的领域公式采用高斯函数：



是i神经元与j神经元之间的距离，为学习率，高斯函数随着迭代次数的增加，优胜领域也不断的减少。

SOM算法的具体过程:

1.神经元初始化:将输出层神经元赋小随机数并进行归一化处理，得到，j=1,2,…m，m为目标输出神经元个数,学习率****赋初始值，迭代次数赋初始值；

2.接受输入：从输入样本总体中取得一个输入模式，i=1,2,….p,p为样本的个数；

3.竞争学习：寻找获胜神经元，将与所有的的相似性，相似性最大的神经元为获胜神经元；

4.权值更新:根据权值更新公式将获胜神经元领域范围内的神经元的权值进行更新；

5.迭代：调整学习率并检查学习率是否衰减到0或者某个特定的正小数

## 1.2 改进的SOM算法

计算向量相似度的常用方法有欧式距离，标准化欧式距离，余弦相似度等。

欧式距离：

标准化欧式距离：

余弦相似度：

在传统SOM算法中,获胜神经元的计算方式是采用欧式距离最小,也就是默认样本维度的权重是一样的,而在常规的多维分析中，每个维度对于数据相似性的贡献是不一样的。因此在SOM算法中加入维度权重,以此来改变SOM算法中神经元相似度的计算方式；

对于数据的不同维度的权重，大权重维度对于数据的相似性越重要，也会使数据的聚类的凝聚性越高，聚类簇间距也就越大，数据的离散程度也就越大，因此可以采用数据指标的离散程度来表示数据指标的权值系数，离散程度越大，单位的差距越大，单位的权值系数也越大；

统计中，我们一般采用平均数来客观表现总体中单一数据指标的水平，但是平均数反映的是样本各指标的平均水平，并不能反映指标的离散程度，常用的离散程度的计算方法有标准差和变异系数,但是由于标准差是以算术平均值为中心，反映的是一个总体中所有指标的离散程度，是绝对指标系数，当用其对不同总体对比时，缺乏可比性，而变异系数更加普遍适用与反映总体各单位的的离散程度，集中趋势等；

V=

上式中，V代表变异系数，代表标准差，代表算术平均值。

基于变异系数的SOM算法是在接受输入步骤中对于取得的输入模式的不同的维度赋予不同的变异系数:



# 实例分析

## 实验环境

操作系统为Windows 7 (64位),处理器为Intel(R) Core(TM) i7-3770 CPU @ 3.40GHz ,内存8 G;本文采用iris数据集进行实验，iris数据集有4个维度，分别为花萼长度，花萼宽度，花瓣长度，花瓣宽度，分别计为,,,, 取自3个种类（setosa,versicolor,virginica）,每个种类取50个样本，共150个样本。分别观察4个变量关于所有样本观测的散点图（图xxx），从图xxx中可以得到4个维度其中花瓣长度和宽度两个维度对于样本的区分度比较大，然后采用“变异系数”分别进行计算，得到样本维度权值系数为：0.141，0.141，0.467，0.634.从计算结果可以看出，花瓣的长度和宽度的权值系数较大，表明用变异系数的合理性；

为了将som算法的聚类结果进行可视化，因此首先使用som算法对数据进行”粗聚类”，然后针对聚类的二维的结果采用kmeans来进行二次聚类，从而将数据可视化，同时也可以对som算法的聚类效果进行有效评价，评价方法采用以下几个参数；

准确率（Precision）:正确结果的数量除以所有返回结果的数量；

类内间距：又叫组内相关，它描述了同一组中单位的强度如何相似，表明类的凝聚性，常见的计算方式是组内所有点两两之间的平均距离；

类间距离：聚类结果中类与类之间的聚类，常见的测度方法有组内平均连接距离，重心距离，离差平方和法等，论文采用重心距离进行计算；

表 1 不同算法下iris降维聚类结果

|  |  |  |  |
| --- | --- | --- | --- |
|  | 准确率 | 类间间距 | 类内间距 |
| Som算法（欧式距离） | 85.4% | 7.602 | 2.325 |
| Som算法（标准欧式距离） | 85.6% | 7.54 | 2.485 |
| Som算法（变异系数） | 88.3% | 8.385 | 2.164 |

## 实验分析

从表1中可以得出对于加权欧式距离的som算法，从准确率上，比其他算法准确率提高了3%,从类间间距和类内间距来看，加权欧式距离下的som算法的聚类簇的类间间距和类内间距相对较高，因此聚类簇聚有更高的类内相似性，凝聚度也更高。

# 结语

本文对SOM算法计算相似度的方法进行优化，采用了变异系数作为维度的权值系数，从而将维度权重加入到了相似度距离计算中，在完全无监督的情况下聚类效果得到了提升。实验结果表明，与传统SOM算法和标准欧式距离SOM算法相比，加入权重系数的SOM算法的降维聚类的准确率和聚类凝聚度都有所提高；

参考文献: