基于变异系数的SOM算法在多维分析中的应用研究

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,自组织的,可视化等特性.本文首先引入SOM算法，介绍SOM算法原理,分析了相似性度量的几个方法，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算法具有更好的准确率和凝聚度。

关键词：降维；自组织映射（SOM）；变异系数；权值更新

# 0.引言

多维分析中，常规的统计和基本分析为主的可视化方法有局限性，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）,目标是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必须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（SOM），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，达到了降维并且聚类的效果；

基于SOM算法,国内外学者从不同方向提出了多种的改进方法。基于权值更新公式的改进：陈万振提出了贝叶斯正则化的SOM聚类算法，算法是将SOM权值更新公式中加入经过贝叶斯推理得到的惩罚项，防止过度拟合，提高了聚类的凝聚度[[1]](#endnote-1)。 SOM算法和其他算法结合:张龚[[2]](#endnote-2)等提出基于SOM算法与粒子群优化方法（PSO）相结合的地震分析技术，先利用SOM算法的拓扑结构的特点，将地震数据进行压缩，然后通过粒子群的全局寻优能力来优化Kmeans聚类的效果，算法即把数据进行了降维，又得到了较准确的全局解，兼顾了计算效率和精度；胡晓雪等[[3]](#endnote-3)提出了基于SOM-DMB-PAM混合聚类算法并应用到电力客户细分中，首先利用SOM算法对数据进行降维，然后使用围绕中心点的切分（PAM）对降维后的数据聚类并用Davies-Bouldin指标标定最佳的聚类的个数以保证聚类效果，算法对聚类的个数进行了优化，减少了聚类中个数指定的盲目性。

本文针对SOM算法中相似度计算方式以及聚类结果进行了研究，首先使用利用变异系数调整维度权重并应用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，达到聚类可视化的效果；

# 1.SOM算法介绍

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[[4]](#endnote-4)。

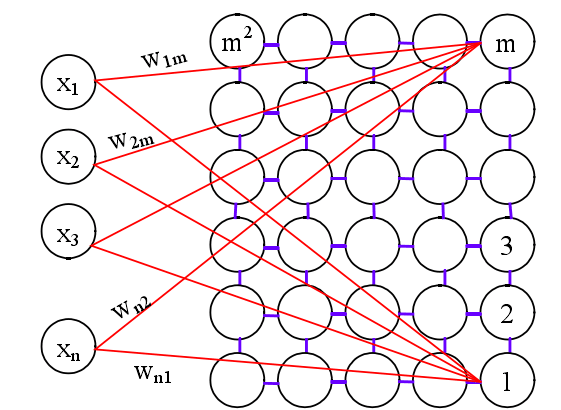


图1 自组织映射网络基本结构

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）权值更新。

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。

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，因此影响权值更新的参数有：迭代次数，神经元距离，学习率，权向量更新权值公式如下：

 (1.1)

公式1.1中为当前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；

公式1.1中是第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；

 （1.2）

公式1.1中优胜领域：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,为此周围神经元与获胜神经元之间的距离是一个影响因素，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：

 （1.3）

公式1.3中是i神经元与j神经元之间的距离，常用的方法是欧式距离，为学习率，高斯函数中随着迭代次数的增加，学习率在减少，获胜神经元的优胜领域也不断的缩小。

SOM算法的具体实现过程:

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：从输入样本总体中随机取得一个输入向量，i=1,2,….p,p为样本的个数；

3.竞争学习：寻找获胜神经元，使用相似性计算公式计算输入向量与所有的神经元的相似性，其中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.调整学习率：根据当前的迭代次数递减学习率；

5.迭代：检查学习率是否衰减到0或者某个特定的正小数，没有满足就回到第二个步骤继续迭代；

# 2.变异系数的作用和引进

两个个体具有n个维度的特征，即与，度量两个个体的相似程度常用方法有欧式距离，标准化欧式距离，余弦相似度等。

欧式距离属于距离度量，它计算欧式空间中两个个体的绝对距离，两个个体越相似其在欧式空间中的距离也就越近，但是计算的是各维度特征值下的绝对距离，并没有考虑量纲，因而在某些情况下欧式距离会失效。公式如下：

 （1.4）

余弦相似度属于相似度度量，比较的是两个向量在方向上的相似性，当两个向量的夹角越小，其夹角余弦也就越大，对于两个向量越相似，但其只能比较个体之间在维之间的差异，没法衡量每个维数值的差异[[5]](#endnote-5)。公式如下：

 （1.5）

标准化欧式距离是为了解决欧式距离的缺点的提出的方法，它将个体的每个向量都标准化到均值和方差都相等，均值为0，方差为1，那标准化过程公式如下：

 （1.6）

其中为原向量的均值，为方差，经推导得标准化欧式距离公式如下：

 （1.7）

根据以上的介绍可以得出欧式距离计算的是绝对距离，个体之间的绝对差异，余弦相似度从方向上区分差异，均未考虑到维度的权值对相似性计算的影响[[6]](#endnote-6)，标准化欧式距离中方差的倒数可以算做一种维度的权值。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[[7]](#endnote-7)，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，而变异系数法是一种客观赋权法，在统计学中，常用与评价数据之间的差异。变异系数越大的指标，权重系数就越大，因此使用变异系数来衡量权重的系数[[8]](#endnote-8)，变异系数公式如下。

V= （1.8）

式1.8中，V代表变异系数，代表标准差，代表算术平均值。

因此本文中利用变异系数来衡量维度的权值。

# 3.改进的SOM算法与实现

为了利用SOM算法的领域函数的特性，降维映射一般结果是在二维的拓扑结构中，其拓扑结构一般定义的范围不限于最终聚类的个数，应该首先使用SOM算法对数据进行一次“粗聚类”，然后使用其他的方法对降维后的数据进行再一次的处理[[9]](#endnote-9)，Kmeans算法简单效率高[[10]](#endnote-10)，本论文采用Kmeans算法进行再次聚类得到聚类结果。

基于变异系数的SOM算法描述：

1. 神经元初始化：对神经元权值赋小随机数，定义初始迭代次数，最大学习率；
2. 样本归一化：由于指标的类型不同，其量级和量纲是由差异的，因此对指标采用指标极值标准化法[[11]](#endnote-11)对样本无量纲化处理；
3. 样本输入：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:

 （1.9）

1. 竞争学习，学习率调整；
2. 迭代:学习率为0或者到某一正整数；
3. Kmeans聚类：使用Kmeans 对降维后的二维数据进行聚类；

# 4.算法比较

论文主要研究基于变异系数的SOM算法，同时搭配Kmeans进行二次聚类，为了验证基于变异系数的SOM算法是否提高了算法的准确度，已经SOM算法搭配Kmeans的可行性，将进行如下实验：

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，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[[12]](#endnote-12)，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；

表1是分别使用欧式距离，标准欧式距离，加入变异系数的的三种SOM算法的测试结果，实验数据是不同算法使用Iris数据集降维聚类10次取得的平均值，类内间距和类间间距是在每一次实验中计算的三个聚类簇之间的类内间距和类间间距的平均值。

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 |

结果分析：对比不同算法下Iris降维聚类结果，基于变异系数的SOM算法，从准确率上，比其他算法准确率提高了3%,从类间间距来看，基于变异系数的SOM算法的聚类簇的类间间距相对较大，说明聚类簇之间的差异性也大，从类内间距来看，基于变异系数的SOM比其他的算法类内间距小，因此聚类簇聚有更高的类内相似性，凝聚度也更高。

# 5.在多维分析中的应用

对大诚中医项目中的客户进行降维聚类，对客户数据清洗，整合，抽取1175个客户的11个维度，分别为性别（sex），年龄(age)，上次下单具今天数(order\_r)，下单总数(order\_f)，下单总额(order\_m)，上次储值具今天数(store\_r)，储值总数(store\_f)，储值总额(store\_m)，上次退款具今天数(tui\_r)，退款总数(tui\_f)，退款总额(tui\_m)。表2是客户不同维度下的变异系数以及相应的权重，mean为均值，SD标准差，CV为变异系数，weight为维度权重。

表2 客户维度变异系数及权重

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Sex | Age | Order\_r | Order\_f | Order\_m | Store\_r | Store\_f | Store\_m | Tui\_r | Tui\_f | Tui\_m |
| Mean | 1.73 | 35.9 | 50.2 | 8.88 | 1719.5 | 177.68 | 0.19 | 1719.14 | 190.4 | 0.15 | 28.8 |
| SD | 0.44 | 18.9 | 31.9 | 11.14 | 2247.8 | 47.97 | 0.43 | 5383.22 | 35.2 | 0.92 | 152.2 |
| CV | 0.26 | 0.53 | 0.64 | 1.25 | 1.31 | 0.27 | 2.21 | 3.19 | 0.18 | 5.99 | 5.28 |
| Weight% | 1.2 | 2.5 | 3.0 | 5.9 | 6.2 | 1.2 | 10.5 | 15.1 | 0.8 | 28.3 | 25.0 |

使用基于变异系数的SOM算法对客户进行降维聚类，聚类结果如图1：

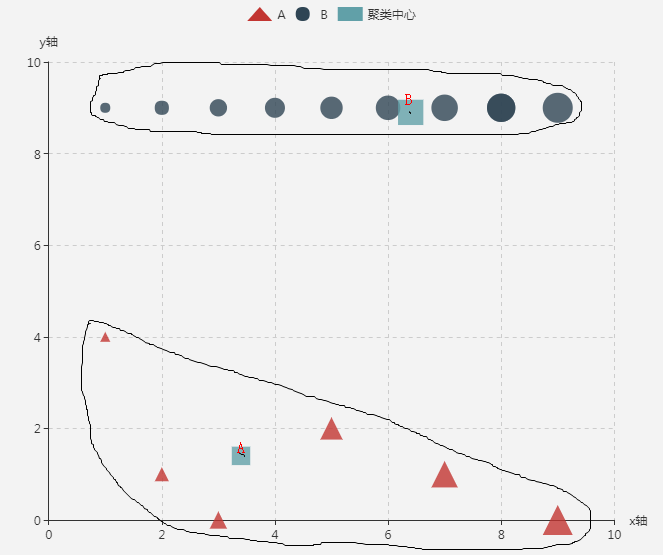


图1 客户聚类拓扑结构图

聚类结果分为两类，正方形为两个聚类簇的中心，多维的客户数据被明显区分为两个类，从而可以为以后的客户分析奠定基础。

# 6.结束语

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，加入变异系数的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:**

1. 陈万振,张予瑶,苏一丹,覃华,蒙祖强. 贝叶斯正则化的SOM聚类算法[J]. 计算机工程与设计,2017,(01):127-131. [↑](#endnote-ref-1)
2. 张,郑晓东,李劲松,路交通,曹成寅,隋京坤. 基于SOM和PSO的非监督地震相分析技术[J]. 地球物理学报,2015,(09):3412-3423. [↑](#endnote-ref-2)
3. 胡晓雪,赵嵩正,吴楠. 基于SOM-DB-PAM混合聚类算法的电力客户细分[J]. 计算机工程,2015,(10):295-301+308. [↑](#endnote-ref-3)
4. 冯如晓. 基于SOM聚类的数据可视化系统研究与实现[D].南华大学,2014. [↑](#endnote-ref-4)
5. 陈大力,沈岩涛,谢槟竹,马颖异. 基于余弦相似度模型的最佳教练遴选算法[J]. 东北大学学报(自然科学版),2014,(12):1697-1700. [↑](#endnote-ref-5)
6. 王星华,陈卓优,彭显刚. 一种基于双层聚类分析的负荷形态组合识别方法[J/OL]. 电网技术,2016,(05):1495-1501. [↑](#endnote-ref-6)
7. 刘帮莉,马玉鑫,侍洪波. 基于加权距离邻域选取策略的多模态过程故障检测[J]. 华东理工大学学报(自然科学版),2015,(02):192-197+215. [↑](#endnote-ref-7)
8. 王芸．基于变异系数权重的灰色关联投影法在水质评价中的应用［j］．地下水

   ，2010，32(2)：61-63. [↑](#endnote-ref-8)
9. Shen, Z., Wei, J., Ma, K. L., & Sundaresan, N. (2012). Visual cluster exploration of web clickstream data. Visual Analytics Science and Technology (pp.3-12). IEEE. [↑](#endnote-ref-9)
10. 张发明. 一种融合SOM与K-means算法的动态信用评价方法及应用[J]. 运筹与管理,2014,(06):186-192. [↑](#endnote-ref-10)
11. 贾凤梅，郭盛昌，柏金凤，等．黑龙江省生态环境质量动态变化及建议［Ｊ］．湖南农业科学，2014（4）：60-62． [↑](#endnote-ref-11)
12. 陈国玉. 基于SOM-K-means的天涯BBS水军帖的聚类分析[D].华中科技大学,2013. [↑](#endnote-ref-12)