《改进的SOM算法在多维可视化中的研究》

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.论文首先介绍了SOM算法发展过程中的改进算法,然后针对欧式距离的特点,将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:相比没有指标权重的SOM算法,采用带有指标权重的SOM算法具有更好的降维效果;

1. 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；

1. SOM算法及其改进
   1. SOM算法基本原理

SOM网络结构如图1所示，它由输入层和竞争层(输出层)组成。输入层神经元数为n，竞争层由 m 个神经元组成的一维或者二维平面阵列,网络是全连接的，即每个输入结点都同所有的输出结点相连接。

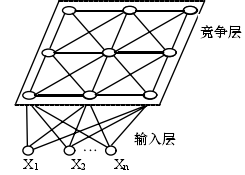


图1 SOM网络结构

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“即以获胜神经元为圆心，对近邻的神经元表现出兴奋性侧反馈，而对远邻的神经元表现出抑制性侧反馈，近邻者相互激励，远邻者相互抑制”。

SOM算法的具体过程:

初始化:神经元初始化,样本归一化;

竞争学习:计算样本和所有神经元的欧式距离,并选择最小的获胜神经元;

权值更新:

学习率调整

迭代

* 1. 改进的SOM算法

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,因此在SOM算法中加入权重系数,以此来改变对于欧式距离中各个维度的贡献率,加权欧式距离；

对于数据的不同维度的权重，权重大的维度对于数据的分类越重要，类之间的间距也就越大，对于数据的离散程度的离散程度，我们可以采用变异系数：标准差/均值来度量，也就是加权欧式距离中的所定义的，但是与均值不能为0的限制，我们可以采用各个维度的“四分位离差系数”作为其权值；

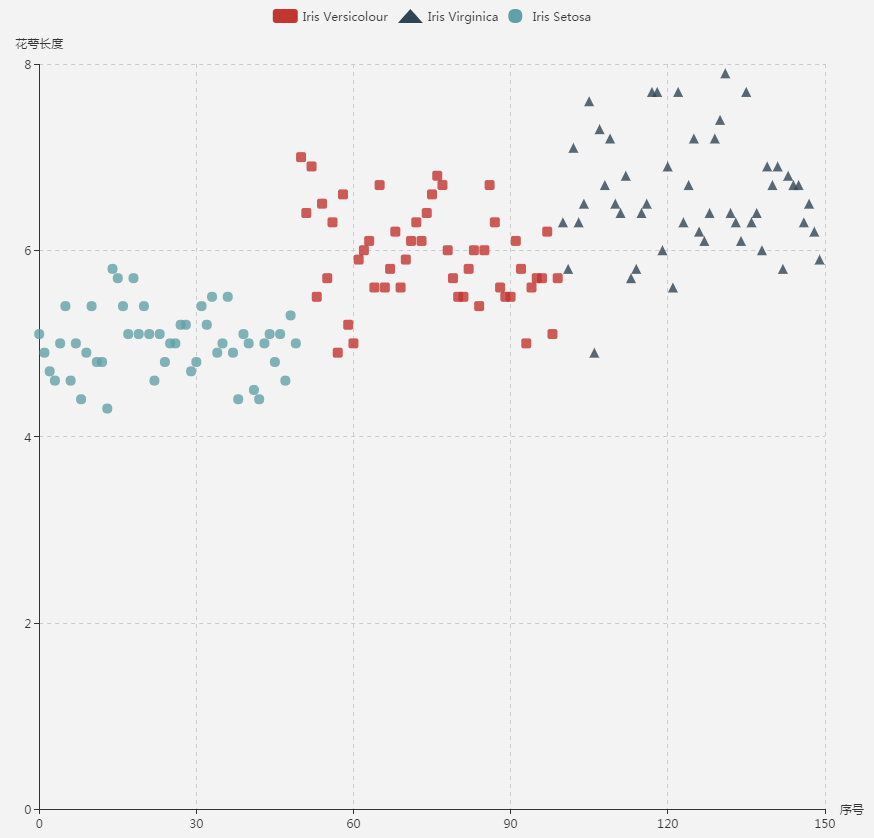
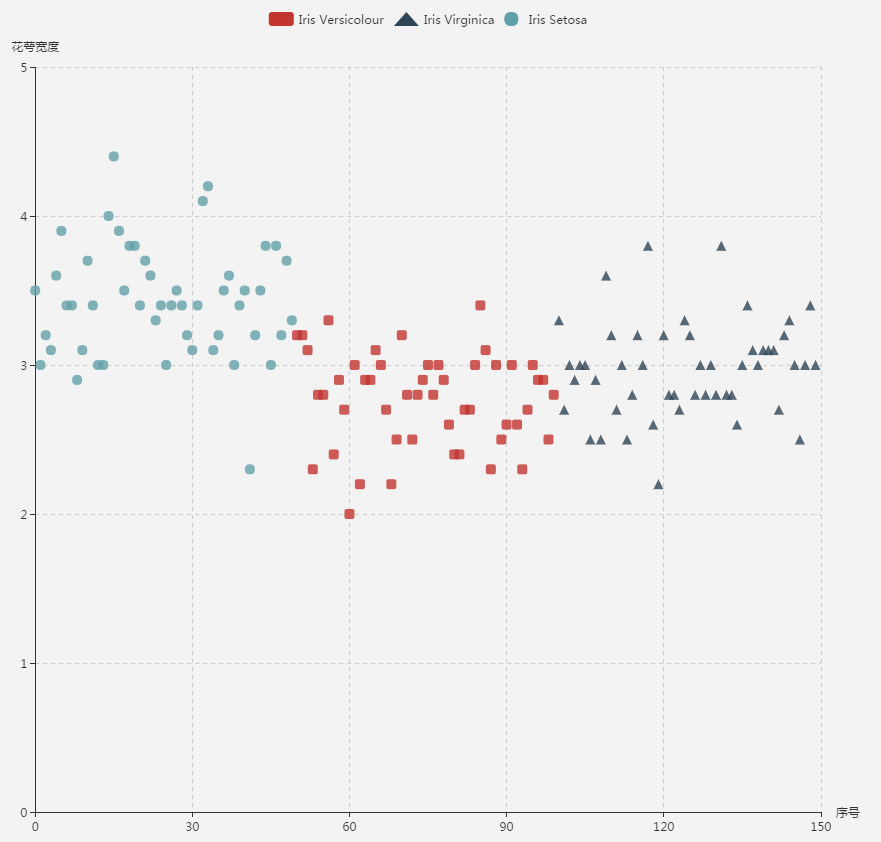
四分位离差系数的定义：

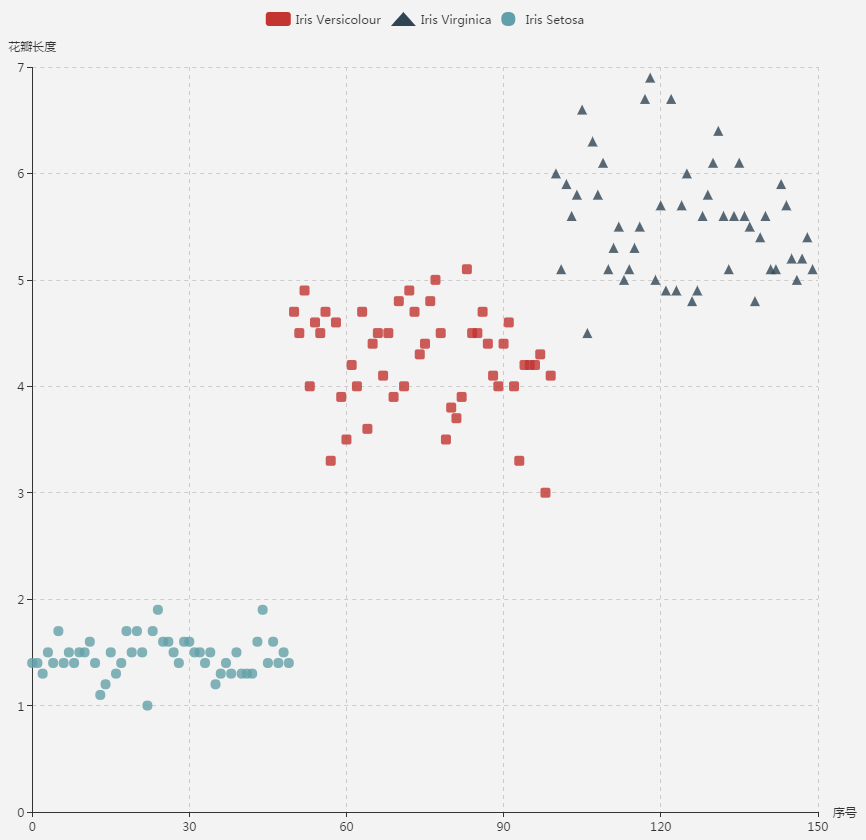
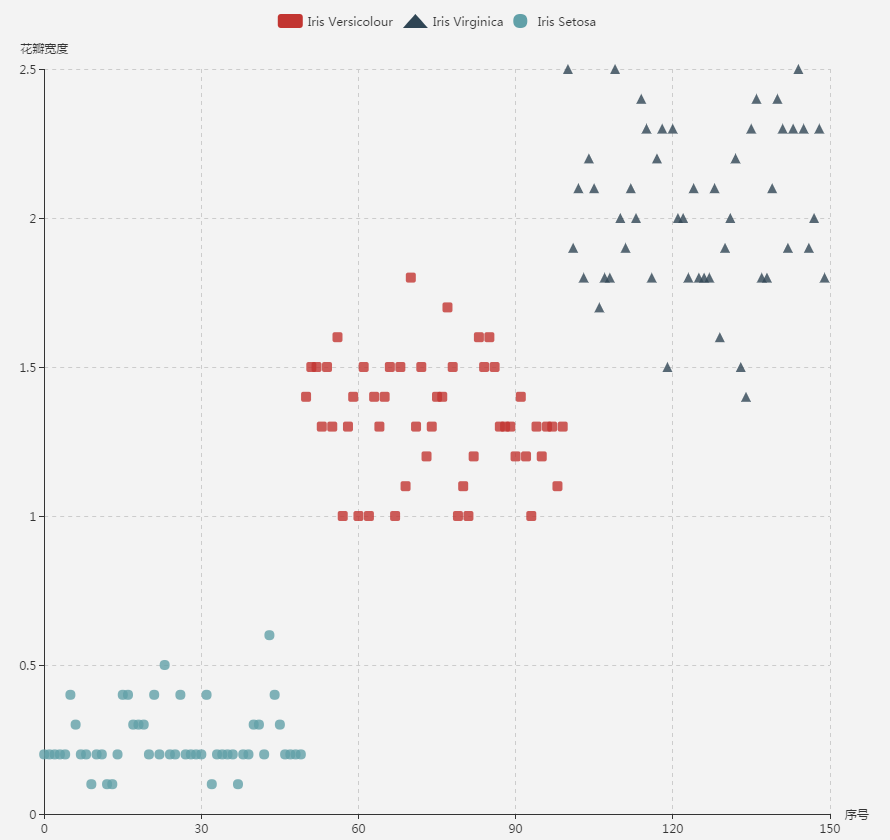
https://wenku.baidu.com/view/9391a1770740be1e640e9a03.html

带有指标权重的SOM算法描述:

….形式语言描述

1. 实验结果分析
   1. 采用鸢尾花数据集，特点是什么？

* 1. 将上述两种算法进行比较，对比结果

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；下面是使用不同距离公式下的som算法聚类结果的比较，采用准确率**（Precision），类间间距（以重心之间的距离），类内间距（类内所有点两两之间的平均距离）来进行比较**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 准确率 | 类间间距 | 类内间距 |
| Som算法（欧式距离） |  |  |  |
| Som算法（标准欧式距离） |  |  |  |
| Som算法（加权欧式距离） | 83% |  |  |

3.3

4 结语

参考文献: