K-means 聚类分析

主要的聚类算法可以分为以下几种：划分聚类、层次聚类、密度型聚类、网格型聚类和基于模型的聚类。

划分聚类算法把数据点集分为 k 个划分，每个划分作为一个聚类。它一般从一个初始划分开始，然后通过重复的控制策略，使某个准则函数最优化，而每个聚类由其质心来代表(k-means 算法)，或者由该聚类中最靠近中心的一个对象来代
表(k-medoids 算法)。划分聚类算法收敛速度快，缺点在于它倾向于识别凸形分布大小相近、密度相近的聚类，不能发现分布形状比较复杂的聚类，它要求类别数目 k 可以合理地估计，并且初始中心的选择和噪声会对聚类结果产生很大影响。主要的划分聚类算法有 k-means，EM，k-medoids，CLARA，CLARANS 等。

下面主要介绍 K-means 聚类方法。

k-means 算法首先随机选择 k 个对象，每个对象代表一个聚类的质心。对于其余的每一个对象，根据该对象与各聚类质心之间的距离，把它分配到与之最相似的聚类中。然后，计算每个聚类的新质心。重复上述过程，直到准则函数会聚。通常采用的准则函数是平方误差准则函数(squared-error criterion)，即

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m|^2$$

E 是一个数据集中所有对象的误差平方和，P 是一个对象，mi 是聚类 Ci 的质心，即

$$m_i = \frac{\sum_{q \in C_i} q}{\|C_i\|}$$

k-means 聚类算法的具体步骤如下：

①从数据集中选择 k 个质心 C1，C2，…，Ck 作为初始的聚类中心；

②把每个对象分配到与之最相似的聚合。每个聚合用其中所有对象的均值来代表，“最相似”就是指距离最小。对于每个点 Vi，找出一个质心 Cj，使它与其间
的距离 d(Vi,Cj)最小，并把 Vi 分配到第 j 组；

③把所有的点都分配到相应的组之后重新计算每个组的质心 Cj；

④循环执行第②步和第③步，直到数据的划分不再发生变化。

该算法具有很好的可伸缩性，其计算复杂度为 O(nkt)，其中，t 是循环的次数。K-means 聚类算法的不足之处在于它要多次扫描数据库，此外，它只能找出球形的
类，而不能发现任意形状的类。还有，初始质心的选择对聚类结果有较大的影响，该算法对噪声很敏感。

问题探究

近年来，数据挖掘成为越来越热的一个研究方向，而聚类(clustering)作为数据挖掘的主要方法之一，也越来越引起人们的关注。所谓聚类，就是把大量的 d 维数据对象(n 个)聚集成 k 个聚类(k<n)，使同一聚类内对象的相似性尽可能最大，

而不同聚类内对象的相似性尽量达到最小。也就是说，形成聚类之后，同一个聚类内对象具有很高的相似性，而与不属于该聚类的对象有迥然的差异(即不相似)。聚类与分类相比，分类算法分析的是类别已知的数据集，而聚类算法分析的是类别未知的数据。

聚类的输入是一组未分类的记录，而且事先也不知道要分成几类，它通过分析数据，根据一定的分类准则，合理划分记录集合，从而确定每个记录所属的类别。不同的聚类算法中，用于描述相似性的函数也有所不同，有的采用欧氏距离或马氏距离，有的采用向量夹角的余弦，也有的采用其他的度量方法。

当预先不知道类型数目，或者用参数估计和非参数估计难以分辨不同类型的类概率密度函数时，就需要采用聚类分析。有些聚类分析算法可以自动地确定类型的数目 k，而不必以预知 k 为前提条件，也可以给定 k 作为算法的终止条件若没有给定 k，那么如何在聚类过程中自动地确定 k，这是聚类分析中的一个关键问题。

确定 k 值的方法很多，具体哪种较好还有待探究。基于自组织特征神经网络（SOM）的聚类分析是其中的一种解决方案。

自组织特征神经网络简介

在人类的神经系统及脑的研究中，人们发现:人脑的某些区域对某种信息或感觉敏感，如人脑的某一部分进行机械记忆特别有效;而某一部分进行抽象思维特别有效。这种情况使人们对大脑的作用的整体性与局部性特征有所认识。对大脑的研究说明，大脑是由大量协同作用的神经元群体组成的。大脑的神经网络是一个十分复杂的反馈系统;在这个系统中含有各种反馈作用，有整体反馈，局部反馈:另外，还有化学交互作用。在大脑处理信息的过程中，聚类是极其重要的功能。大脑通过聚类过程从而识别外界的信号，并产生自组织过程。根据大脑对信号的处理特点，在 1981 年，T.Kohonen 提出了一种神经网络模型，也就是自组织特征映射 SOM(Self 一　orzanizingFeatureMap)。

SOM 网络算法是一种聚类算法，它能根据其学习规则对输入的模式进行自动分类，即再在无监督的情况下，对输入模式进行自组织学习，通过反复地调整连接权重系数，最终使得这些系数反映出输入样本之间地相互关系，并在竞争层将分类结果表示出来。因此，SOM 神经网络在结构上模拟了大脑皮层中神经元旱二维空间点阵的结构，并在功能上通过网络中神经元间的相互作用和相互竞争，模拟了大脑信息处理的聚类功能、自组织和学习功能。该算法被广泛应用于各种模式识别和分类问题中。

SOM 网络的结构如图 1 所示，它由输入层和输出层组成。其中输入层的神经元个数的选取视输入网络的向量个数而定。输入神经元接收网络的输入信号,输出层则是由神经元按一定的方式排列成一个平面。输入层的神经元与输出层的神经元通过权值相互连接在一起。在每个输入样本学习过程中，SOM 网络找出与之距离最短的输出层单元，即获胜节点（BMU），然后更新 BMU 及其相邻区域的权值，使得输出节点保持输入向量的拓扑特征。SOM 网络的具体训练方法如下：

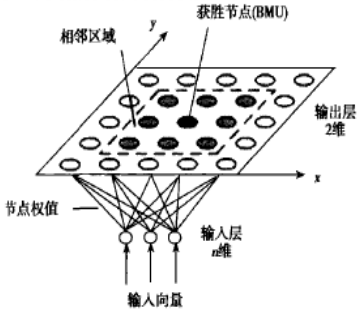

图 1　SOM 网络结构图

1)初始化。对输出层的每个神经元节点的权值 W_j 随机赋予较小的初值，定义训练结束条件。

2)从输入样本中随机选取一个输入模式 X_i。

3)寻找获胜节点（BMU）。定义 g 为获胜节点，其应该满足下式：

$$d_g = \min(d_j) = \min_j \| X_i - W_j \| \qquad (1)$$

式中：$\| \cdot \|$ 表示距离函数，对于连续数值属性的数据集，通常采用欧氏距离。

4)确定获胜单元的相邻区域 $N_g(t)$。区域 $N_g(t)$随着时间的增长而不断缩小。对区域 $N_g(t)$内的节点的权值按下式(2)进行调整使其向 X_i 靠拢：

$$W_j(t+1) = W_j(t) + \eta(t) h_{gj}(t)[X_i(t) - W_j(t)] \qquad (2)$$

式中：t 表示时间，$\eta(t)$为学习率，$h_{gj}(t)$为 g 的邻域函数。

5)选取新的输入模式，重复步骤 3) 4)，直到所有输入样本都已经完成训练。

6)按一定规则收缩相邻区域 $N_g(t)$，减小学习率$\eta(t)$，重复步骤 2)到 5)，当达到训练结束条件后停止训练，输出聚类结果。

由于输出层各节点互相激励学习，训练后的邻近节点具有相似的权值，因此 SOM 网络输出节点的空间位置体现了输入样本的内在联系，即具有相似属性的输入会映射在邻近的 SOM 输出节点上。利用人眼对低维数据的快速把握能力，实现聚类的可视化。

通过计算结果作图分析可以确定 k 值。再结合 k-means 算法快速分类。

参考文献：

【1】　基于 SOM 神经网络和 K-均值聚类的分类器设计　曹金平

【2】　基于自组织映射神经网络的低压故障电弧聚类分析　邹云峰，吴吴为麟李智勇

【3】　基于自组织特征映射网络的聚类算法研究　吴红艳

附录：

k-mean 和 SOM 神经网络 matlab 程序（未调试）。