基于计算智能的聚类算法①

Clustering Algorithm Based on Computational Intelligence

马金亮　成新明　(中南大学 信息科学与工程学院 湖南 长沙 410083)

摘　要： 近年来数据挖掘领域中的聚类和人工智能领域的计算智能都有长足的进步和发展。计算智能自组织、自学习的特性为聚类问题的解决提供了一个有效的途径。当前基于计算智能的聚类算法主要包括：基于神经网络的聚类算法、基于遗传算法的聚类算法和基于蚁群算法的聚类算法。本文针对以上算法进行了阐述，详细说明了算法思想、关键技术和优缺点，并提出了有待进一步研究的问题。

关键词： 聚类 自组织特征映射 遗传算法 蚁群算法

1　前言

聚类分析是数据挖掘(Data Mining)领域的研究热点之一，其基本思想是将数据对象分组成为多个类或簇，使得在同一个簇中的对象具有较高的相似度，而不同簇的对象之间差别较大[1]。经典的聚类算法如K-means、CURE等，都需要预先指定聚类的数目，这种指定多基于经验，具有一定的盲目性，在整体上降低了聚类的效率和结果的可信度，而且增加了数据挖掘流程的复杂性[2]。

近年来计算智能的发展为聚类问题的解决提供了有效的方法，本文介绍了将计算智能中的神经网络、遗传算法和蚁群算法与已有的聚类算法相结合的方法，提高了聚类的性能和准确性，增强了算法的适应性和鲁棒性。

2　基于神经网络的聚类算法

神经网络是模仿人类神经系统的工作原理，通过学习待分析数据中的模式来构造模型，具有高度的非线性、自学习、自组织等特性。

2.1 算法描述

自组织映射(Self-Organizing Map,简称 SOM)是神经网络中的著名算法，SOM是一种竞争学习的神经网络模型，它在学习的过程中，其竞争层的获胜神经元在进行权值调整时，不仅仅调整获胜神经元，而且对其邻近的神经元按与其空间距离的远近也分别作不同程度的调整，使得获胜的神经元与其邻近的神经元保持一定的相似性。SOM能在一维或二维的竞争层神经元阵列中对输入样本在其矢量空间中的拓扑关系进行学习和保持。经过学习，网络将输入样本映射到输出层的相应区域，形成特征映射。不同的连接权值对应不同的模式，改变连接权值的过程就是学习的过程[3-5]。

自组织特征映射网络的拓扑结构如图1所示，由输入层和竞争层(输出层)组成，输入层的神经元个数由输入模式的特征数决定，一般是一个特征对应一个输入神经元，图中表示N个输入神经元，$m\times m=M$ 个输出神经元，且形成一个二维阵列。输入模式的每个元素(既特征值)均连至输出层(即特征平面)上的每个神经元。

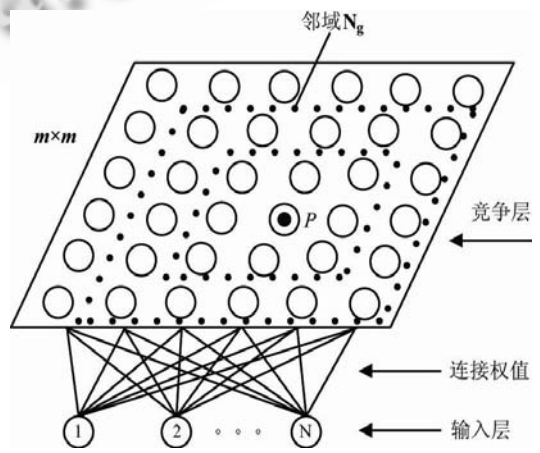

图1　SOM的网络拓扑结构

① 基金项目:湖南省自然科学基金项目(07JJ6126)
收稿时间:2008-09-11

2.2 算法流程

SOM 的目标是试图使输出节点的内部连接权矢量模仿输入信号。其网络训练算法如下：

Step1：权连接初始化。对所有从输入节点到输出节点的连接权值赋以随机的小数。时间计数置 0，即 t=0。

Step2：网络输入模式为 $X^k=(x_1,x_2,\cdots,x_n)$。

Step3：对 X^k 计算 X^k 与全部输出节点所连权向量 W_j^T 的距离。

$$d_j=\sum_{i=1}^{n}(x_i^k-w_{ij})^2,i\in\{1,2,\cdots,n\},j\in\{1,2,\cdots,m\} \quad (1)$$

Step4：具有最小距离的节点 N_{j*} 竞争获胜。

$$d_{j*}=\min_{j\in\{1,2,\cdots,m\}}\{d_j\} \quad (2)$$

Step5：调整输出节点 N_{j*} 所连接的权值以及 N_{j*} 几何邻域 $NE_{j*}(t)$ 内节点所连接的权值：

$$\Delta w_{ij}=\eta(t)(x_i^k-w_{ij}),N_j\in NE_{j*}(t),i\in\{1,2,\cdots,n\}$$

式中 $\eta(t)$ 为标量自增益，$0<\eta(t)<1$，是一个单调减函数，往往选择 $\eta(t)=0.9(1-t/1000)$，公式(1)和(2)组合在一起，就是自组织映射的学习规则。

Step6：若还有输入样本数据，那么 $t\leftarrow t+1$，转到步骤 Step2。

根据学习规则，网络通过对输入模式的反复学习，捕捉各个输入模式所含的模式特征，在输出层将结果表示出来，即网络对输入数据进行自动聚类，输出层反映了聚类结果。

自组织特征映射是一种自组织、无监督的聚类算法，能将任意维输入模式在输出层映射成一维或二维离散图形，并保持其拓扑结构不变，即在无师指导的情况下，通过对输入模式的自组织学习，在竞争层将分类结果表示出来，可以产生较为理想的聚类结果。

其局限性是在学习模式较小时，网络连接权向量的初始状态对网络的收敛性能影响很大，其次，邻域和学习速率如何选择，目前尚无固定的数学方法，一般原则是邻域开始选得较大，然后逐步收缩，学习速率的选择主要是从经验出发,并不具有精确的数学上的意义。另外其网络结构参数难以确定，只适用于小数据集合，对大数据集合容易出现训练过度，时间空间复杂度过高，聚类结果精确度难以保证，输出结果依赖于样本的输入顺序。

因此，针对神经网络算法的局限性可从如下几个方面展开进一步的研究：探索确定邻域和学习速率的新方法；通过确定网络结构参数的新方法，以避免学习过度、降低算法的时间和空间复杂度；消除对样本输入顺序的依赖等。

3 基于遗传算法的聚类算法

遗传算法(Genetic Algorithm)是一类借鉴生物界的进化规律演化而来的随机化搜索方法，具有内在的隐并行性和良好的全局寻优能力[6]。

3.1 算法描述

聚类算法的作用是在 N 维空间适当决定 K 个聚类中心，从而把 n 个无标签数据进行合理的聚类。本文算法中，种群中的每个染色体由一个的浮点数序列表示，第 1 个 N 位表示 N 维空间中的第一个聚类中心，接下来 N 位表示第二个聚类中心，依此类推。

3.2 算法流程

Step1：初始化运行参数。

Step2：初始化种群。

Step3：对种群中的每个个体执行 K-means 操作，然后对个体进行适度值的计算。

Step4：判断结束条件，当条件未满足时进行以下操作：

(1) 按轮盘赌注选择策略从 n 个父代中选出 n 个个体形成交配池；

(2) 将交配池个体随机配对成 n/2 对配对个体，然后进行两两交叉；

(3) 对交叉后的个体进行变异操作；

(4) 产生新一代种群；

Step5：转到 Step3。

3.3 适度函数设计

特征向量与相应聚类中心的平方误差的和作为聚类划分质量的评价函数 J，J 越小表示聚类划分的质量越好。J 的数学表达式为：

$$J=\sum_{i=1}^{k}\sum_{x_j\in c_i}\|x_j-c_i\|^2 \quad (3)$$

遗传算法的目的是搜索 J 值最小的聚类中心，因此适应度函数 f=1/J。

3.4 遗传算子

3.4.1 选择操作

使用比例选择与最优保存策略相结合的混合选择算子。首先采用按适应度分配计算个体被选中的概率，然后用轮盘赌方法进行个体的选择。择优策略就是将种群中的最好的解保存到下一代。在每次新的群体形成的时候用上一代中所记录的最优个体来替换新群体中的最差个体，以防止遗传操作破坏当前的种群中适应度最好的个体。

3.4.2 交叉操作

在交叉操作之前使用随机配对方式将交配池中 n 个染色体配对成 n/2 对配对个体。交叉操作是在这些配对个体组中的两个个体之间进行的。为了尽量保证产生有意义的新个体，以提高遗传算法的收敛速度，采用基于最短距离基因匹配的算术交叉算子。

3.4.3 变异操作

每个染色体以一个固定的概率 p_m 来进行变异。假设在当代中，聚类评价函数的最大值和最小值分别为 $J_{\min}$ 和 $J_{\max}$。对于一个聚类评价函数为 J 的染色体，根据均衡分布原则，产生一个在[-R,+R]范围中的数 δ。其中：

$$R=\begin{cases}\dfrac{J-J_{\min}}{J_{\max}-J_{\min}} & J_{\min}>J\\ 1 & J_{\min}=J\end{cases} \tag{4}$$

如果数据集中的第 i 维(i=1,2,⋯,N)的最大值和最小值分别为 $X^i_{\max}$ 和 $X^i_{\min}$，聚类中心要变异的位置为第 i 维，其值 X^i 在变异操作后变化情况为：$\delta\geq 0$ 时变为 $X^i+\delta\times(X^i_{\max}-X^i)$，否则变为 $X^i+\delta\times(X^i-X^i_{\max})$。

3.5 终止规则

算法的结束条件为：若种群中最优个体的适应值连续 50 代未得到改善，则算法结束。在每次进行选择、交叉和变异操作之后，记录当前子代中适应度最高的个体，算法运行结束后，适应度最高的个体为聚类问题的最优解。

遗传算法模拟生物进化的过程，具有很好的自组织、自适应和自学习能力。基于遗传算法的聚类方法的优点是不需要关于待分类数据的先验分布知识，也不会受初始解选择的影响而得到次优解[7,8]。

遗传算法最主要的缺陷是局部搜索能力差，采用遗传算法求解优化问题一般只能得到准最优解，而不容易得到最优解。其次是容易产生较多的无效解，算法的收敛速度也有待进一步提高。再次，遗传操作中的参数交叉率和变异率的选择和设定目前尚无统一的理论指导，多数都视具体问题而定。其中，变异率的选择最为困难，稍有不适便会产生早熟。

因此，针对遗传算法的不足可从如下几个方面展开进一步的研究：如何做到既防止震荡过大、收敛速度过慢，又防止收敛速度过快、发生早熟现象；如何在进化过程中适当地调节交叉率和变异率；如何使用户通过模拟程序来直接监视种群的多样性等。

4 基于蚁群算法的聚类算法

蚁群算法(Ant Colony Algorithm)是近几年才提出的一种新型模拟进化算法，具有全局优化能力，适用于复杂的组合优化问题的求解[9]。

4.1 算法描述

算法模拟真实蚂蚁的协作过程，由许多蚂蚁共同完成，每只蚂蚁在候选解空间中独立地搜索解，并在所寻得的解上留下一定的信息量，信息量越大的解被选择的可能性也越大。蚁群聚类算法是一种全局优化的启发式算法，能根据聚类中心的信息量把周围数据归并到一起，从而得到数据分类[10,11]。

假定数据样本为一个包含 N 个数据的集合 $\{x_1,x_2,\cdots,x_N\}$，每个样本点表示为 $x_k=\{x_{k1},x_{k2},\cdots,x_{kn}\}$。聚类分析就是将这 N 个对象划分到 K 个类中，使得聚类目标函数 F 最小。定义如下目标函数：

$$F_{\min}(w,m)=\sum_{j=1}^{K}\sum_{i=1}^{N}\sum_{v=1}^{n}w_{ij}\|x_{iv}-m_{jv}\|^2 \tag{5}$$

满足

$$\sum_{j=1}^{K}w_{ij}=1,i=1,2,\cdots N \qquad \sum_{i=1}^{N}w_{ij}\geq 1,j=1,2,\cdots K$$

式中，x_{iv} 是对象 i 的第 v 个属性值；w 是一个的加权矩阵，其元素

$$w_{ij}=\begin{cases}1,\text{如果}\ x_i\in cluster\ j\\ 0,\text{如果}\ x_i\notin cluster\ j\end{cases}$$

m 是一个 $K\times n$ 的聚类中心的矩阵，m_{jv} 表示类 j 中所有样本的属性 v 的均值，计算公式如下：

$$m_{jv}=\frac{\sum_{i=1}^{N}w_{ij}x_{iv}}{\sum_{i=1}^{N}w_{ij}},j=1,2,\cdots,K,v=1,2,\cdots n \tag{6}$$

在蚁群聚类算法中，使用 R 个人工蚂蚁构建解，每个蚂蚁构建的解为一个长度为 N 的字符串 $S=\{c_1,c_2,\cdots,c_N\}$，其中 $\{c_i|i=1,2,\cdots,N\}$ 是对象 i 的类标示，$c_i\in 1,2,\cdots,K$。$c_i=c_j$ 表示对象 x_i,x_j 属于同一个类，$c_i\neq c_j$ 表示对象 x_i,x_j 不属于同一个类。在算法的开始阶段，初始化 $N\times K$ 的信息素矩阵,将其赋予一个初值 τ_0，矩阵元素 τ_{ij} 表示对象 i 相对于类 j 的信息素浓度。在每一次迭代中，每一个人工蚂蚁基于信息素矩阵构造一个解，并且使用交叉算子进一步改善解的质量，然后基于解的质量更新信息素矩阵。于是，在不断更新的信息素矩阵的指引下，蚂蚁不断改善解的质量，直到达到迭代次数。

4.2 算法流程

Step1：将 R 个蚂蚁的解编码为字符串。

Step2：随机生成 K 个聚类中心。

Step3：令 i=1。

Step4：使用状态转移规则构建解 S_i。

Step5：计算样本与聚类中心的加权矩阵。

Step6：计算解 S_i 的目标函数 F_i。

Step7：令 i=i+1。

Step8：如果 $i\leq R$ 转到 Step3，否则转到 Step9。

Step9：对最好的 L 个解运行交叉算子。

Step10：计算新生成解的目标函数值。

Step11：计算全局最优解的 K 个聚类中心。

Step12：进行全局信息素更新。

Step13：判断是否满足结束条件，满足则结束，不满足则转到 Step1。

蚁群算法适用于组合优化问题，把数据对象看作蚂蚁，聚类中心即是食物源，得到了合理、有效的聚类结果[12]。

蚁群算法出现时间较短，其相关数学分析还比较薄弱。蚁群算法要初始化大量的参数，这些参数的选择会对算法的性能产生较大的影响，但其选取的方法和原则目前尚无理论上的依据，只能通过多次实验调优，且计算时间偏长，容易出现停滞现象。另外，目前蚁群算法的收敛性证明还不完善，收敛速度也不高。

因此，针对蚁群算法的缺陷可从如下几个方面做进一步的研究：研究蚁群算法的理论性分析和各种参数设置问题，得出更好的设置规则；给出更强的收敛性证明并得出收敛速度的估计；研究蚁群的并行实现，实现蚁群之间的通讯以及蚁群的调度的并行性；研究蚁群算法和其他先进算法的融合，以改善自身性能。

虽然蚁群算法的研究相对于其他比较成熟的计算智能方法来讲还处于初级阶段，并存在很多有待深入研究和解决的问题，但是可以预言群体智能的研究代表了以后计算机研究发展的一个重要方向。

5 结语

本文综述了三类基于计算智能的聚类算法，包括：基于神经网络的聚类算法、基于遗传算法的聚类算法和基于蚁群算法的聚类算法，详细介绍了算法的实现步骤和实现算法的关键点，阐述了每种算法的优缺点和有待于进一步研究的问题，对基于计算智能的聚类算法研究提供有益的参考。

参考文献

1 Han J W, Kamber M.范明,孟小峰,等译.数据挖掘概念与技术.北京:机械工业出版社, 2001:223－225.

2 杨黎刚,苏宏业,张英,褚健.基于 SOM 聚类的数据挖掘方法及其应用研究.计算机工程与科学, 2007, 29(8):133－136.

3 白耀辉,陈明.利用自组织特征映射神经网络进行可视化聚类.计算机仿真, 2006,23(1):180－183.

4 Tomsich P, Rauber A, Merkl D. Optimizing the parsom neural network implementation for data mining with distributed memory system and cluster computing. Proceedings of 11th International Workshop on Database and Expert Systems Applications. 2000:661－665.

5 张敏灵,陈兆乾,周志华.SOM 算法、LvQ 算法及其变体综述.计算机科学, 2002,29:97－100.

6 周明,孙树栋.遗传算法原理及其应用.北京:国防工业出版社, 1999.

7 傅景广,许刚,王裕国.基于遗传算法的聚类分析.计算机工程, 2004,30(4):122－124.

8 Hall LO, Ozyurt I B, Bezdek J C. Clustering with a geneticallyoptimized approach. IEEE Transactions on Evolutionary Computation, 1999,3(2):103－112.

9 李士勇.蚁群优化算法及其应用研究进展.计算机测量与控制, 2003,11(12):911－913.

10 Shelokar P S , Jayaraman V K, Kulkarni B D. An ant colony approach for clustering. Analytica Chimica Acta, 2004,509:187－195.

11 刘波.一种利用信息熵的群体智能聚类算法.计算机工程与应用, 2004,35:180－182.

12 高坚.基于并行多种群自适应蚁群算法的聚类分析.计算机工程与应用, 2003,29(25):78－82.