基于遗传算法的高维数据模糊聚类

王宝文[1],阎俊梅[1],刘文远[1],石 岩[2]
WANG Bao-wen[1],YAN Jun-mei[1],LIU Wen-yuan[1],SHI Yan[2]

1.燕山大学 信息学院,河北 秦皇岛 066004
2.日本九州东海大学 工程学院 信息系统工程系
1.Informatin Science and Engineering Institute of Yanshan University,Qinhuangdao,Hebei 066004,China
2.Department of Information System Engineering,School of Engineering,Kyushu Tokai University,Japan
E-mail:junmei689@163.com

WANG Bao-wen,YAN Jun-mei,LIU Wen-yuan,et al.High dimensional datas fuzzy clustering based on genetic algorithm.Computer Engineering and Applications,2007,43(16):191-192.

Abstract: A high dimensional datas fuzzy clustering method is presented based on genetic algorithm,by importing a fuzzy dissimilar matrix to express the dissimilar degree between any two datas,and initializing the high dimensional samples to two dimensional plane.And then iteratively optimize the coordinate value of two dimensional plane using genetic algorithm,which makes the euclidean distance between the two dimensional plane approximate to the fuzzy dissimilar degree between samples gradually,and the high dimensional samples are mapped into two dimensional plane.At last,using FCM algorithm to the two dimensional datas,avoids the dependence of the validity of clustering on the space distribution of high dimensional samples. Experimental results show that the method this paper proposed has more exact clustering result and faster convergence speed than FCM algorithm.

Key words: fuzzy clustering;fuzzy dissimilar matrix;genetic algorithm;high dimensional datas

摘 要:提出了一种基于遗传算法的高维数据模糊聚类方法。引入了一个模糊非相似矩阵来表示高维样本之间的非相似程度,并将高维样本初始化到二维平面。利用遗传算法进行迭代优化二维样本的坐标值,实现二维样本之间的欧氏距离向样本间的模糊非相似度的趋近,使高维样本映射到二维平面。最后将得到的最优的二维样本利用模糊 C-均值聚类(FCM)算法聚类,克服了聚类有效性对高维样本空间分布的依赖。实验仿真表明利用该方法有较好的聚类效果,且比用 FCM 算法直接聚类收敛速度快。

关键词:模糊聚类;模糊非相似矩阵;遗传算法;高维数据
文章编号:1002-8331(2007)16-0191-02 **文献标识码**:A **中图分类号**:TP18

1 引言

聚类是依据事物的某些属性将其聚集成类,使类内相似性尽量大,类间的相似性尽量小,在这一过程没有教师的指导,是一种无监督的模式识别问题。传统的聚类方法有 C-均值聚类和 FCM 软聚类,这些算法都是基于目标函数的[1],而基于目标函数的聚类对样本的空间分布有较强的依赖性[2]。例如,C-均值聚类对于特征空间呈超球体的情况聚类效果较好,而对于呈任意形状簇分布的情况则聚类效果较差[3],FCM 软聚类对于特征空间呈椭球体结构的情况聚类效果较好[4],而且 FCM 算法对高维数据聚类时速度较慢[5]。为了克服聚类有效性对样本分布的依赖以及提高聚类的效率,本文提出了一种基于遗传算法的高维数据模糊聚类算法,目的是将高维样本间的模糊非相似程度转化为二维样本间的欧氏距离,即将高维样本的差异性转化为二维样本的差异性, 实现高维样本向二维样本的映射,最后再对二维样本利用 FCM 算法聚类即可。

2 基于遗传算法的高维数据模糊聚类

2.1 建立模糊非相似矩阵

模糊非相似矩阵用于存储样本之间的非相似程度,用[0,1]之间的数来表示。设样本空间 $X=\{x_1,x_2,\cdots,x_n\}$,$\forall x_i\in X$,其特征矢量为 $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots x_{ip})$,$x_{ik}$ 为第 i 个样本的第 k 个特征属性。

记 n 个样本第 k 个特征属性的平均值和标准方差分别为

$$\mu_k=\frac{1}{n}\sum_{i=1}^{n}x_{ik} \tag{1}$$

$$s_k=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ik}-\mu_k)^2} \tag{2}$$

则原始样本可标准化为

基金项目:国家科技部高新技术计划项目(No.2005EJ000017);河北省科技研究与发展计划(02547015D);河北省普通高等学校博士科研资助基金(No.2002B2002118)。

作者简介:王宝文(1957-),男,副教授,研究方向:软计算,模糊推理,数据挖掘;阎俊梅(1981-),女,硕士生,研究方向:模糊聚类;刘文远(1968-),男,教授,博士后,研究方向:虚拟企业、电子商务和数据挖掘;石岩,副教授,研究方向:软计算。

$$x'_{ik}=\frac{x_{ik}-\mu_k}{s_k} \tag{3}$$

利用汉明距离可得到第 i 个样本与第 j 个样本之间的模糊非相似性为：

$$r_{ij}=c\sum_{i=1}^{n}|x'_{ik}-x'_{jk}| \tag{4}$$

C 为[0,1]间的参数，在这里选择 C=0.01 这样就得到了一个 $n\times n$ 维的对角线为 0 的模糊非相似矩阵 $(r_{ij})_{nn}$：

$$\begin{bmatrix} 0 & & & & \\ r_{21} & 0 & & & \\ r_{31} & r_{32} & 0 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & \cdots & 0 \end{bmatrix} \tag{5}$$

2.2 适应度函数的选取

将高维样本随机初始化到二维平面，通过遗传算法对各个二维样本的坐标值进行迭代优化，使各样本间的欧氏距离逐渐趋近于模糊非相似度。因此，遗传算法的误差函数定义为：

$$E=\sum_{i=1}^{n}\sum_{j=i}^{n}|r'_{ij}-r_{ij}| \tag{6}$$

其中，r'_{ij} 表示二维平面的样本 i 和 j 之间的欧氏距离。设 i 和 j 的坐标值分别为 (a_i,b_i) 和 (a_j,b_j)，i=1,2,…,n，j=1,2,…,n，则 r'_{ij} 为：

$$r'_{ij}=\sqrt{(a_i-a_j)^2+(b_i-b_j)^2} \tag{7}$$

误差函数值越小，个体的适应值越高，因此个体的适应度函数定义为：

$$f=1/(1+E) \tag{8}$$

2.3 基于遗传算法的高维数据模糊聚类算法

算法描述如下：

(1)初始化：将待聚类样本随机分布在二维平面的一定区域内，即随机赋给每个样本一对坐标值 (a_i,b_i)，其中 $a_i,b_i\in[0,1]$，i=1,2,…,n。

(2)建立模糊非相似矩阵：利用式(1)~(5)建立模糊非相似矩阵 $(r_{ij})_{nn}$。

(3)建立初始群体：将每个样本的一对坐标值 (a_i,b_i) 用 8 位二进制编码，即 a_i 和 b_i 分别用 4 位二进制来表示。若一共有 n 个样本，将所有的基因链接起来构成一个染色体(即个体)，则染色体长度 L=8n 位。设置种群大小 $Size$=N，即由上述的 N 条染色体构成了初始群体。

(4)计算适应度：利用式(6)~(8)计算出每个个体的适应度值。

(5)选择父本：利用保留最优策略和轮盘选择法。首先选择群体中适应度值最大的个体作为一个父本，然后，计算其余每个个体的选择概率 $p_k=f_k/\sum_{i=1}^{n}f_m$ 以及累计概率 $q_i=\sum_{j=1}^{i}p_j$，产生一个[0,1)区间的均匀随机数 r，若 $r<q_1$，则选择第一个个体；否则若 k 满足 $q_{k-1}\leqslant r<q_k$，则选择个体 k，旋转 M-1 次。即共可选择出 M 个个体，构成子群体 S'，$S'\subset S$。

(6)随机地将 S'中的个体两两配对。

(7)交叉操作：S'中的每对个体产生[0,1]之间的随机数 r，若 $r<Pc$(Pc 为选定的交叉概率)，则进行交叉操作。随后产生[1,8n]之间的随机数以确定交叉的位置，交叉后的新染色体构成了群体 S''。

(8)变异操作：对 S''中的每一个体的每一位产生[0,1]之间的随机数 r，若 $r<Pm$(Pm 为选定的变异概率)，则该位变异。

(9)计算 S+S''中所有个体的适应度，并淘汰掉适应度小的 M 个个体，形成新一代群体 S。

(10)终止操作：如果新一代个体的最大的适应度与上一代个体的最大适应度的差值小于 ε(ε 取值为 0.005)，则解码。否则转到(5)。

(11)对解码后的二维坐标值应用 FCM 算法，并把得到的聚类结果对应回原始的高维样本中。

3 算法可行性分析

该算法是利用遗传算法将二维样本间的欧氏距离逐渐趋近于高维样本间的模糊非相似性，使得误差函数 $E=\sum_{i=1}^{n}\sum_{j=i}^{n}|r'_{ij}-r_{ij}|$ 达到最小值，实现由高维样本向二维样本的映射。

根据 2.1 和 2.2 可知，$r_{ij}\in[0,1]$，$r'_{ij}\in[0,1]$。若 $r_{ij}\approx0$，即高维样本 i 与样本 j 的非相似性几乎为 0，说明样本 i 与样本 j 为一类。又 r'_{ij} 趋近于 r_{ij}，所以 $r'_{ij}\approx0$，即这两个高维样本映射到二维平面上的二维样本间的欧氏距离几乎为 0，根据类内距离小，类间距离大可得该二维样本经 FCM 聚类后应为一类。

若 $r_{ij}\approx1$，即高维样本 i 与样本 j 的非相似性几乎为 1，说明样本 i 与样本 j 属于不同的类。又 r'_{ij} 趋近于 r_{ij}，所以 $r'_{ij}\approx1$，同样对应的二维样本经 FCM 聚类后应属于不同的类。

同理，任何两个高维样本，若它们的模糊非相似性越大，那么它们对应的二维样本间的欧氏距离越大。而欧氏距离越大，则相似性越小，即二维样本之间的非相似性越大，这样就将高维样本间的差异程度转化为二维样本间的差异程度，因此对映射后的二维样本聚类就相当于对原始的高维样本聚类，具有可行性。

4 仿真实验

实验选取了一部分 IRIS 数据作为样本，样本总数为 21，样本属性为 4，聚类类别为 3，其中每类所包括的样本数都为 7。

表 1 所选取的 IRIS 数据

数据编号	属性 1	属性 2	属性 3	属性 4
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	7.0	3.2	4.7	1.4
9	6.4	3.2	4.5	1.5
10	6.9	3.1	4.9	1.5
11	5.5	2.3	4.0	1.3
12	6.5	2.8	4.6	1.5
13	5.7	2.8	4.5	1.3
14	6.3	3.3	4.7	1.6
15	6.3	3.3	6.0	2.5
16	5.8	2.7	5.1	1.9
17	7.1	3.0	5.9	2.1
18	6.3	2.9	5.6	1.8
19	6.5	3.0	6.6	2.1
20	7.6	3.0	6.6	2.1
21	4.9	2.5	4.5	1.7

(下转 221 页)