第33卷 第24期 计 算 机 工 程 2007年12月
Vol.33 No.24 Computer Engineering December 2007
·软件技术与数据库· 文章编号：1000—3428(2007)24—0060—03 文献标识码：A 中图分类号：TP18

基于分治法的高维大数据集模糊聚类算法

王宝文[1]，阎俊梅[1]，刘文远[1]，石 岩[2]
(1. 燕山大学信息学院，秦皇岛 066004；2. 日本九州东海大学工程学院信息系统工程系)

摘 要：将高维的大数据集随机分成若干个子集，对每个子集聚类采用一种基于遗传算法的高维数据模糊聚类方法。该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，并将高维样本随机初始化到二维平面，利用遗传算法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。将得到的最优的二维样本用模糊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，该算法有较好的聚类效果，且极大地提高了聚类的速度。

关键词：模糊聚类；分治法；遗传算法；模糊非相似矩阵；大数据集；高维

Fuzzy Clustering Algorithm for High-dimensional Large Data Sets Based on Distributed Method

WANG Bao-wen [1], YAN Jun-mei[1], LIU Wen-yuan[1], SHI Yan[2]
(1. Informatin Science and Engineering Institute of Yanshan University, Qihuangdao 066004;
2. Department of Information System Eng., School of Engineering, Kyushu Tokai University, Japan)

【Abstract】 Data sets are randomly divided into several subsets. A high- dimensional datum fuzzy clustering method based on genetic algorithm is used to cluster the subsets, by importing a fuzzy dissimilar matrix to express the dissimilar degree between any two high dimensional datum, 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. At last cluster the two-dimensional datum using FCM algorithm, so avoid dependence of clustering validity on the space distribution of high-dimensional samples. Experimental results show the method has exact clustering result, and improves the clustering speed greatly.

【Key words】 fuzzy clustering; distributed method; genetic algorithm; fuzzy dissimilar matrix; large data sets; high dimension

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，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软聚类算法[1]。随着需要聚类的数据量不断地增加，用传统的方法直接聚类效率很低[2]，鉴于此，本文提出了基于分治方法的聚类算法。如果大数据集是属于二维的，那么对子集的聚类可直接用传统的聚类算法聚类。高维数据集涉及到空间分布的情况，而传统聚类算法的有效性对样本的空间分布有较强的依赖性[3]。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，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]。为了克服聚类有效性对样本空间分布的依赖，对每个高维大数据集的子集聚类采用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。最后对二维样本利用FCM算法聚类即可。

1 基于分治法的高维大数据集模糊聚类算法

算法过程描述：

(1)给定数据集$A=\{a_1, a_2,\cdots, a_n\}$。

(2)将数据集A分成若干个子集$A_1, A_2,\cdots, A_p$。

(3)if(数据集的维数>2) then 对每个高维样本的子集采用一种基于遗传算法的高维数据模糊聚类算法。即将高维样本随机初始化到二维平面，利用遗传算法迭代优化二维样本，将高维样本映射到二维平面，然后对得到的最优的二维样本聚类；else 直接用FCM算法聚类。

(4)对p个子集聚类后分别得到的聚类中心数为$m_1, m_2,\cdots,m_p$。

(5)if$|m_1+m_2+\cdots+m_p|\geq n_0$($n_0$为问题规模的阈值)，then将$m_1+m_2+\cdots+m_p$个聚类中心看成集合$A$，转到第(2)步；else转到 第(6)步。

(6)把$m_1+m_2+\cdots+m_p$个聚类中心进行一次性聚类。

(7)类的合并：第(6)步结束后，if聚类中心x_1和x_2聚为一类，而在第(4)步结束后c_1和c_2分别是以x_1和x_2为聚类中心的类，then将类c_1和c_2合并为一类。

(8)If 数据集是高维样本，then 将第(7)步的聚类结果还原到原始的高维样本中；else 算法到第(7)步就结束。

基金项目：国家科技部高新技术计划资金资助项目(2005EJ000017)；河北省科技研究与发展计划基金资助项目(02547015D)；河北省普通高等学校博士科研资金资助项目(B2002118)

作者简介：王宝文(1957－)，男，副教授，主研方向：软计算，模糊推理，数据挖掘；阎俊梅，硕士研究生；刘文远，教授、博士后；石 岩，副教授

收稿日期：2006-12-21 **E-mail：**junmei689@163.com

2 基于分治法的二维样本聚类分析

实验选取了一部分 IRIS 数据，且使用了 petal length 和 petal width 两个属性。IRIS 数据一共是 3 类，表 1 所选的数据是每一类的子集。

表 1 选取的 IRIS 数据

第 1 子集	第 2 子集	第 3 子集
1.3 0.2; 1.7 0.4; 1.4 0.3; 1.4 0.2; 1.5 0.1; 1.5 0.2;1.6 0.2;1.2 0.2;1.4 0.1; 1.1 0.1	4.7 1.4; 4.5 1.5; 4.9 1.5; 4 1.3; 4.6 1.5; 4.5 1.3; 4.7 1.6; 3.3 1; 4.6 1.3; 3.9 1.4; 3.5 1; 4.2 1.5;4 1; 3.6 1.3	5.9 2.1; 5.6 1.8;5.8 2.2; 6.6 2.1; 6.3 1.8; 5.8 1.8; 6.1 2.5;6.4 2; .3 1.9; 5.5 2.1; 6.1 1.9;5.3 2.3; 5.5 1.8; 6.7 2.2; 6.9 2.3

本文实验将数据集随机分成 2 个或 3 个子集，对每个子集聚类时设置聚类中心数为 3，最后的一次性聚类设置聚类中心数为 3。下面是其中的一次实验。该实验将数据集随机分成 3 个子集 *A*, B, C，结果如表 2~表 4 所示。

表 2 对 *A* 子集聚类的结果

聚类中心	属于该类的数据
4.0133 1.2493	3.3 1;4 1.3;4.6 1.3
1.3551 1.0775	1.5 0.1;1.4 0.1;1.1 0.1
1.4302 2.3097	1.3 0.2;1.7 0.4;1.4 0.3;1.4 0.2;1.5 0.2;1.6 0.2;1.2 0.2

表 3 对 *B* 子集聚类的结果

聚类中心	属于该类的数据
4.5828 1.4506	4.7 1.4;4.5 1.5;4.6 1.5;4.5 1.3;4.7 1.6
4.8857 1.5519	4.9 1.5
5.5838 1.9526	5.9 2.1;5.6 1.8;5.8 1.8;5.3 1.9;5.5 2.1;5.3 2.3;5.5 1.8

表 4 对 *C* 子集聚类的结果

聚类中心	属于该类的数据
4.0614 1.3953	3.9 1.4;4.2 1.5;4 1
3.6125 1.1291	3.5 1;3.6 1.3
6.3785 2.1188	5.8 2.2;6.6 2.1;6.3 1.8;6.1 2.5;6.4 2;6.7 2.2;6.9 2.3;6.1 1.9

表 2~表 4 中的聚类中心构成集合 *D*，*D* 中有 9 个数，可直接对这 9 个聚类中心进行一次性聚类。最后进行类的合并。子集 *D* 聚类后的结中(1.355 1 1.077 5)和(1.430 2 2.309 7)为一类；(5.583 8 1.952 6)和(6.378 5 2.118 8)为一类；其余的 5 个聚类中心为一类，如图 1 所示。

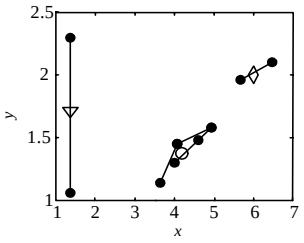

图 1 对集合 D 聚类的结果

那么以(1.355 1 1.077 5)和(1.430 2 2.309 7)为聚类中心的两个类应归为一类，而这两个类正好是集合 *A* 聚类后其中的两个类，合并后正好是表 1 中的第 1 个子集中的数据；同理可以分别得到表 1 中的第 2 个和第 3 个子集中的数据，因此，利用该方法与对原始的数据进行一次性聚类得到了相同的结果。

利用该方法连续做实验 20 次，其中，只有一次产生分歧。产生分歧的结果是将第 2 类中的(4.9 1.5)分到了第 3 类。而(4.9 1.5)正好在第 2 类的边界。可见在边界处的点容易被分出去。

以上是对二维数据利用分治法的过程，实验证明了它的有效性，分析算法的时间复杂度可得：FCM 聚类算法的时间复杂度为 $O(L\times c\times n+L\times c)$，而利用分治法的聚类算法为 $O(n)$ (其中，L 为迭代次数；n 为聚类总数；c 为聚类类别数)，极大地提高了聚类的速度。

所以，对于高维样本利用分治方法时只需证明将高维降为二维来聚类的有效性，即证明基于遗传算法的高维数据模糊聚类算法是有效的。

3 基于遗传算法的高维数据聚类算法

3.1 建立模糊非相似矩阵

设样本空间$X=\{x_1, x_2,\cdots, x_n\}$，$\forall x_i\in X$，其特征为$x_i=(x_{i1}, x_{i2},\cdots, x_{ip})$，$x_{ik}$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

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

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

$$r_{ij}=c\sum_{k=1}^{p}\left|x'_{ik}-x'_{jk}\right| \tag{4}$$

其中，C为[0,1]间的参数，本文选择C=0.01，这样得到一个$n\times n$维的对角线为 0 的模糊非相似$(r_{ij})_{nn}$：

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

3.2 适应度函数的选取

将高维样本先初始化到二维平面，利用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高维样本间的模糊非相似度。因此，遗传算法的误差函数定义为

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

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

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

3.3 算法描述

(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：利用步骤(1)～步骤(5)建立模糊非相似$(r_{ij})_{nn}$。

(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 位二进制表示。若一共有n个样本，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S。

(4)计算适应度：利用步骤(6)～步骤(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。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' 中个体进行交叉构成了群体 S''。

(8)变异操作：对 S'' 中的个体进行变异操作。

(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 =0.005)，则解码。否则转到步骤(5)。

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

3.4 算法可行性分析

根据 3.1 节、3.2 节可知，$r_{ij} \in [0,1]$，$r_{ij}^{'} \in [0,1]$。若 $r_{ij} \approx 0$，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，$r_{ij}^{'} \approx 0$，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。

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 实验仿真

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，每类包括的样本数都为 7。

利用本文提出的方法对表 5 中的数据进行聚类时取种群规模 N=100，迭代次数 G=60，变异概率 Pm=0.5，交叉概率 Pc=0.2，则聚类结果如图 2 所示。

表 5 部分 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
…	…	…	…	…
21	4.9	2.5	4.5	1.7

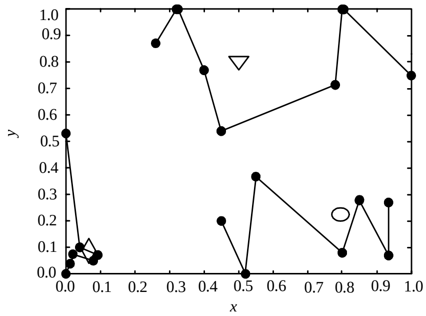

图 2 基于遗传算法的模糊聚类

实验结果表明，该方法有较好的聚类效果，即证明了将高维降为二维聚类的有效性。

5 结束语

本文提出了一种基于分治法的高维大数据集模糊聚类算法。以二维样本为例进行了实验，结果表明利用分治法在大多数情况下与一次性聚类结果一致，并且通过分析可得该方法能极大的提高聚类的速度。对于高维样本通过实验证明了将高维降为二维去聚类的有效性，并分析了可行性，因此，本文提出的方法适合对高维大数据集聚类，能提高聚类的效率，且具有有效性。

参考文献

[1] Zhang Yuanquan, Rueda L. A Geometric Framework to Visualize Fuzzy-clustered Data[C]//Proceedings of IEEE 25th International Conference of the Chilean Computer Science Society. Valdivia, Chile: [s. n.], 2005: 8-13.

[2] Davidson I, Satyanarayana A. Speeding Up *K*-means Clustering by Bootstrap Averaging[C]//Proc. of IEEE Data Mining Workshop on Clustering Large Data Sets. Brighton, UK: [s. n.], 2004: 98-102.

[3] Aggarwal C, Yu P. Finding Generalized Projected Clusters in Dimensional Spaces[C]//Proc. of ACM SIGMOD Conference on Management Data. Dallas, Texas, U.S.A: [s. n.], 2000: 78-52.

[4] Tsai Chengfa, Tsai Chunwei, Chen Chiping. A Novel Multiple Searching Genetic Algorithm for Multimedia Multicast Routing[C]//Proc. of IEEE Congress on Evolutionary Computation. Piscataway, NJ: [s. n.], 2002: 7065-7068.

~~~~~~

(上接第 59 页)

上比传统算法有了明显提高，同时在推荐的多样性上也有了明显的改善。下一步的工作为：将该算法部署到实际的推荐系统中，通过在线测试获得用户对推荐准确性和多样性的满意度的反馈，进一步改进算法。

参考文献

[1] Pazzani M. A Framework for Collaborative, Content-based and Demographic Filtering[J]. Artificial Intelligence Review, 1999, 13(5): 393-408.

[2] 王丽珍. 一种基于语义贴近度的抽象归纳法[J]. 计算机学报, 2000, 23(10): 1114-1121.

[3] Kamahara J, Asakawa T, Shimojo S, et al. A Community-based Recommendation System to Reveal Unexpected Interests[C]//Proc. of the 11th International Multimedia Modeling Conference. Tokyo, Japan: [s. n.], 2005: 433-438.

[4] Sarwar B, Karypis G, Konstan J, et al. Analysis of Recommender Algorithms for E-commerce[C]//Proc. of the 2nd ACM E-commerce Conference. Minneapolis, America: Minnesota Press, 2000: 135-141.

[5] Ziegler C, Mcnee S, Konstan J, et al. Improving Recommendation Lists Through Topic Diversification[C]//Proc. of International World Wide Web Conference. Chiba, Japan: [s. n.], 2005: 22-32.
~~~~~~