改进的模糊 C 均值聚类算法

刘坤朋,罗 可
LIU Kun-peng,LUO Ke

长沙理工大学 计算机通信与工程学院,长沙 410076
College of Computer and Communication Engineering,Changsha University of Science & Technology,Changsha 410076,China
E-mail:kunpeng55@yahoo.com.cn

LIU Kun-peng,LUO Ke.Improved fuzzy C-means clustering algorithm.Computer Engineering and Applications,2009,45(21):97-98.

Abstract: Self-adaptive strategy with the traditional fuzzy C-means clustering algorithm forms a new fuzzy clustering algorithm. Without prejudice to the speed of convergence,it can resolve the problems of local optimal and sensitivity to initial values.With the two data sets in the database of UCI machine learning for the study,the experimental results indicate that it does not lose the precision to the adaptive immune clustering algorithm.The number of clusters is accurate and its faster convergence is more important in the nowadays of high-speed network data changing.

Key words: fuzzy C-means clustering;self-adaptive;cluster adjustment

摘 要:把自适应的策略与传统的模糊 C 均值聚类算法结合起来,形成新的模糊聚类算法。在不影响收敛速度的情况下,它能够很好解决局部最优以及对初始值敏感的问题。以 UCI 机器学习数据库中的两组数据集为研究对象,实验结果表明,它的精确度与自适应免疫聚类算法相当,能够得到准确的簇的数目,并且它的收敛速度更快,这对于如今网络数据的高速变化来说,该方法显得更为重要。

关键词:模糊 C 均值聚类;自适应;簇的调整
DOI:10.3778/j.issn.1002-8331.2009.21.028 **文章编号:**1002-8331(2009)21-0097-02 **文献标识码:**A **中图分类号:**TP18

聚类分析已被广泛应用于数据分析、模式识别、图像处理等方面。传统的聚类分析要求把数据集中的每一点都精确地划分到某个类中,即所谓的硬划分。但实际上大多数事物在属性方面存在着模糊性,即事物间没有明确的界限,不具有非此即彼的性质,所以模糊聚类的概念更适合事物的本质,能更客观地反映现实。目前,模糊 C 均值(fuzzy C-means)聚类算法是应用最广泛的一种模糊聚类算法。但是,由于 FCM 算法本质是用梯度下降的方法寻找最优解,因此不可避免地会陷于局部最优值,同时算法的收敛速度受初始值的影响较大,特别是在簇类数较大的情况下,这一缺点更为突出。模糊 C 均值算法还有个致命的缺点就是要事先给定数据簇的数目,而在动态的网络中数据簇的数目往往是事先不知道的。在文献[1]中提出了一种自适应的免疫聚类算法, 将该文献中的自适应方法应用到文中,来分析自适应的模糊 C 均值聚类算法在数据挖掘中的特性。

1 模糊 C 均值(FCM)聚类算法

设样本空间 $X=\{x_1,x_2,\cdots,x_n\}$,将 X 分为 m 类,m 为大于 1 的正整数,可以用模糊矩阵 $\boldsymbol{u}=(u_{ij})$ 来表示,u_{ij} 表示第 i 个样本属于第 j 类的排斥性度量。为此定义:

$$u_{ij}=\frac{\| x_i-c_j \|}{\sum_{k=1}^{m}\| x_i-c_k \|} \tag{1}$$

其中 C_j 为第 j 个聚类中心点, $\| x_i-c_j \|$ 表示 x_i 到 c_j 的欧氏距离。很显然,如果对于样本点 x_i 到第 j 个聚类中心 c_j 最近,则排斥性越小,即 u_{ij} 越小。其中

$$u_{ij}<1,\ \sum_{i=1}^{n}u_{ij}=1,i=1,2,\cdots,n,j=1,2,\cdots,m$$

FCM 的目标函数 $F(u,C)$ 定义为:

$$F(u,C)=\sum_{i=1}^{n}\sum_{j=1}^{n}(u_{ij}^{b}*\| x_i-c_j \|) \tag{2}$$

其中 b 为模糊指数。显然, $\| x_i-c_j \|$ 越小,u_{ij} 越小,F 就越小。不断修改聚类中心的值,然后再依次计算 $F(u,C)$,直到得到理想的结果。因此,FCM 算法是将目标函数 $F(u,C)$ 最小化的迭代收敛过程。由于要不断修改聚类中心,因此将聚类中心点设置为:

$$C=\frac{\sum_{i=1}^{n}u_{ij}^{b}*x_i}{\sum_{i=1}^{n}u_{ij}^{b}} \tag{3}$$

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60474070,No.10471036);湖南省科技计划项目(No.05FJ3074);湖南省教育厅重点项目(No.07A001)。
作者简介:罗可(1961-),男,教授,博士,研究方向为数据挖掘,计算机应用等;刘坤朋(1985-),男,硕士,研究方向为数据挖掘,计算机软件与理论。
收稿日期:2008-04-28 **修回日期:**2008-07-28

由此来确定 FCM 聚类算法的迭代过程：

步骤 1 设定簇的数目和模糊指数 b。

步骤 2 随机得到 m 个聚类中心 c_m^0,0 表示当前的迭代次数为 0。

步骤 3 用当前得到的聚类中心根据式(1)去计算目标函数,然后由当前得到的排斥度 u_{ij} 去更新聚类中心 c_j,重复第三步运算,直到目标函数达到最小值或低于预先给定的阈值 ε。

2 自适应模糊 C 均值聚类算法

为了达到自适应的效果,先要从上面得到的模糊矩阵 $\boldsymbol{u}$ 中分离出属于各个聚类的样本矩阵 $\boldsymbol{M}$,$\boldsymbol{M}=[m_{ij},0]$, 其中 m_{ij} 表示属于第 j 类的 i 个样本点,其余的点用 0 来填充,一共 m 个簇,则共有 m 行。$m_{ij}=\min\sum_{j=1}^{m}u_{ji}$ 表示 m_{ij} 取模糊矩阵的每列的最小值,即与 i 类(簇)排斥性最小的样本点 j 最有可能属于该类,将该样本点放置到 $\boldsymbol{M}$ 矩阵的第 i 行第 j 列。另外还得到 $\boldsymbol{M}$ 的映射函数 A,$A=(A_i,a_i)$,A_i 为 i 类的聚类中心,$a_i=\sum_{i=1}^{m}\|m_i-A_i\|$,为每类的样本点到该样本中心的欧式距离。

调整簇的数目:对于某两个簇的样本点,它们有可能属于同一簇,因此就有必要将它们合并为一个簇。如果两个簇的中心点的欧式距离小于一个事先给定的值,且它们各自的样本元素到另一个簇的中心点的距离很接近, 则可以视它们为一个簇,这个时候可以考虑合并。

$$d_{ij}=\|A_i-A_k\| \tag{4}$$

$$l_{ij}=\left|\frac{\sum_{i=1}^{n}\|M_{jk}-A_j\|}{\sum_{i=1}^{n}\|M_{ik}-A_i\|}\right| \tag{5}$$

循环计算 d_{ij} 和 l_{ij} 的值,如果得到 $d_{ij}<\delta_1$,$l_{ij}<\delta_2$ 时,将 $\boldsymbol{M}$ 矩阵的 i,j 行合并为一行,将另外一行删除。

合并了相似的簇后,再随机生成几个新的聚类中心,重新对它们进行迭代聚类算法。该步骤的目的是考虑到数据的动态性,每时每刻都会发生数据的变化,有新的数据补充进来,进一步充实样本空间,因此原来得到的簇的中心点就有可能发生变化,重新对样本空间进行迭代,便可能得到更加精确的聚类中心。

具体情形如下图：

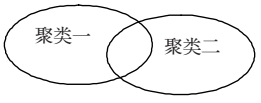

图 1 原聚类结构

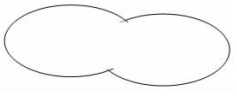

图 2 聚类合并

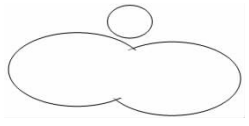

图 3 增加新的聚类

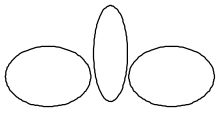

图 4 引起原聚类的变化

3 仿真模拟实验

为了验证提出的算法的有效性,利用 UCI 机器学习数据库中的 Wine 和 Iris 两数据集进行仿真实验,然后将它与文献[1]中利用的自适应免疫聚类算法做个比较。实验源数据如表 1。

实验中有关的参数做如下初始化:第一代簇的数目为 10,模糊指数 b 为 2,迭代次数为 40 代,$\delta_1=0.5$,$\delta_2=0.05$。将它与自适应免疫算法做个对比,比较结果如表 2：

表 1 数据源信息

数据集	样本数目	属性数目	类别数	各类样本所包含的数目
Wine	178	13	3	59 71 48
Iris	150	4	3	50 50 50

表 2 自适应免疫算法与自适应 FCM 算法比较

算法	指标	Wine	Iris
自适应免疫算法	准确	142	135
	错误	36	15
	准确率	79.7%	90%
	得到簇的数目	3	3
自适应 FCM 算法	准确	139	141
	错误	39	9
	准确率	78%	94%
	得到簇的数目	3	3

从表中可以看出,两算法在精确度上基本上差不多,都得到了准确的簇的数目。自适应 FCM 算法搜索的速度更快,只要迭代 30 次就能得到比较满意的结果。利用 Matlab 得到的图 5 如下(小黑点表示 Iris 图,星号表示 Wine 图)：

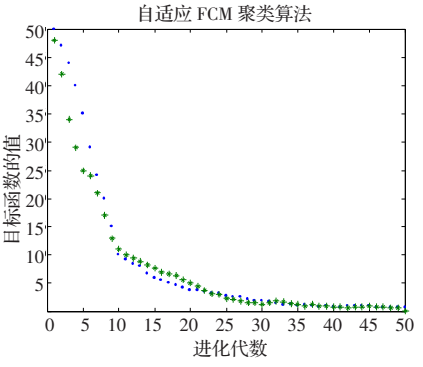

图 5 两种算法的比较结果

4 结论

基于目前应用最广泛的 FCM 聚类算法及结合参考文献[1]而提出的自适应 FCM 聚类算法,很好地解决了 FCM 算法对初始值的敏感以及容易陷入局部最优的问题,若选择一个合适的模糊指数将会对精度的提高有很大的促进作用。

参考文献：

[1] Castro L N D,Zuben F J V.An evolutionary immune network for data clustering[C]//Proceedings of the 6th Brazilian Symposium on Neural Networks,2000:84-89.

[2] YANG Minshen,WANG Peiyuan,CHEN Dehua.Fuzzy clustering algorithms for mixed feature variables[J].Fuzzy Sets and Systems,2004,141(2):301-317.

[3] Chen J J,Gao J,Liao B S,et al.Dynamic semantic clustering approach for web user interest[C]//GCC Workshops,Wuhan,2004:59-66.

[4] Nasraoui,Dasgupta D,Gonzalez F.An artificial immune system approach to robust data mining[C]//Genetic and Evolutionary Computation Conf(GECCO).New York,2002:356-363.

(下转 188 页)