改进的 GA-FCM 算法及其在交通事故挖掘中的应用①

杨兴春[1,2] 王 刚[1,2] 张安妮[3] (1.四川警察学院 四川 泸州 646000;
2.四川省公安厅警官培训基地 四川 泸州 646000; 3.山东黄河信息中心 山东 济南 250013)

摘 要: 目前，公安信息化建设和应用正在不断深入推进，如果能从高速增长的公安业务数据中发掘出隐藏的信息并用于指导公安实战，不但能提升信息化的水平，而且能极大提升实战工作的针对性和有效性。论文针对 GA-FCM 算法收敛慢、解质量不高的缺点，利用前期工作的成果，并对 GA-FCM 的种群选择、变异算子进行了改进。将改进后的算法应用于城区交通事故历史数据挖掘，实验表明，挖掘结果对于预防交通事故、改善交通状况具有一定的指导意义。

关键词: 公安信息化; 聚类挖掘; GA-FCM; 交通事故

Improved GA-FCM and Its Application in Traffic Data Mining

YANG Xing-Chun[1,2], WANG Gang[1,2], ZHANG An-Ni[3]

(1. Sichuan Police College, Luzhou 646000, China; 2. Police Officer Training Base of Sichuan Public Security Bureau, Luzhou 646000,China; 3.Information Center of Shandong Yellow River, Jinan 250013, China)

Abstract: With the continual advanvement of Information Construction in Public Security, it is important to boost the direction and the validity of police work and improve the level of Information Construction, if interested information can be mined from the large amount of historical data. Based on an earlier study, this paper improves the GA-FCM from the population selection and the mutation operator. Finally, the improved algorithm is also applied to the traffic accident data mining. The results have shown that the algorihm contributes significantly in preventing traffic accident and improving the traffic situation.

Keywords: information construction in public security ; clustering data mining; gA-FCM; traffic accident

1 引言

当前，公安信息化在公安工作中占有很重要的地位，随着金盾一期工程建设结束，基本形成了公安业务数据的全警采集、全警应用、全警共享的格局。利用数据挖掘技术，可以对这些数据进行更深层次的利用，以提升信息化建设和应用水平，提高公安工作的有效性和针对性，文献[1] 就将聚类算法应用于火警信息的挖掘分析，取得了较好的效果。

聚类挖掘是数据挖掘的一种，属于无指导的挖掘方式，它根据对象间的相似性将对象分成多个类或簇，使得同一类中的对象具有较高的相似性，而不同类中的对象差别较大。针对简单遗传算法(GA)优化的模糊 C 均值(FCM)聚类算法(GA-FCM)收敛慢、解质量不高等缺点，本文在前期工作的基础上，对 GA-FCM 的种群选择和变异算子等进行了改进，并将改进后的算法应用于交通事故数据挖掘，取得了较好的效果，挖掘

① 基金项目:四川省教育厅青年科学基金(07ZB047)
收稿时间:2009-12-24;收到修改稿时间:2010-02-11

结果对于交通事故的预防和交通状况的改善具有较强的指导意义。

2　模糊C-均值(FCM)聚类算法

FCM 算法具有良好的数学理论基础，并且简洁易用，因而应用相当广泛，算法流程如下：

(1) 初始化参数：数据集大小 n，聚类数 c，迭代终止阀值 ε >0，初始聚类中心 V(0)，最大迭代次数 T_{max}，迭代计数器 t=0；

(2) 根据初始聚类中心 V (0)，更新划分矩阵 U(t)：

对于所有的 i,k，如果有 $d_{ik}^{(t)}>0$，则有：

$$m_{ik}^{(t)}=\frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}^{(t)}}{d_{jk}^{(t)}}\right)^{\frac{2}{m-1}}}$$

如果有 i,k 使得 $d_{ik}^{(t)}=0$，则有：$m_{ik}^{(t)}=1$，且 j≠i 时，有 $m_{jk}^{(t)}=0$

(3) 更新聚类中心 V(t+1)：

$$n_i^{t+1}=\frac{\sum_{k=1}^{n}\left(m_{ik}^{(t+1)}\right)^m\cdot\mathbf{x}_k}{\sum_{k=1}^{n}\left(m_{ik}^{(t+1)}\right)^m}\quad(i=1,2,\ldots,c)$$

(4) 若||V(t)-V(t+1)||< ε，或达到最大迭代次数 T_{max}，则终止迭代，输出划分矩阵 U 与聚类中心 V，否则继续步骤(1)。

FCM 算法虽然计算简单，并能从任意给定的初始点开始，沿一个迭代子序列快速地收敛到目标函数的局部极小值点或鞍点[2]。但该算法对初始值尤其是初始聚类中心很敏感，不同的初始值可能得到完全不同的聚类结果，一般不易收敛到全局最优点；在处理大数据集时效率也有待提高。

在实际应用中，通常将遗传算法(GA)与 FCM 算法相结合，通过 GA 的全局寻优能力，克服 FCM 算法对初始聚类中心敏感的缺陷，同时又兼顾了 FCM 的局部快速寻优能力，可以极大地提高聚类的质量，缩短聚类的时间。

3　改进的GA-FCM

GA[3]最早由美国 Michigan 大学的 Holland 教授提出并创建，是一种基于生物自然选择和遗传机制的随机搜索算法，其模拟生物进化过程中的繁殖、交配与基因突变等现象，算法引入优胜劣汰的自然法则，特点是几乎不需要所求问题的任何信息而仅需要目标函数的信息，不受搜索空间是否连续或可导的限制就可找到最优解或满意解，并且非常适合于大规模的并行计算。

但 GA 本身存在早熟收敛、易陷入局部极小值的缺陷，如果只是将简单 GA 与 FCM 相结合来解决聚类问题，聚类的性能与质量一般都不会很理想。因此，若要构造有效的基于 GA 优化的聚类算法，必须改进 GA。文献[4]对 GA 作了有效的改进，本文也是基于此的后续工作。

在文献[4]的基础上，本文将改进的 GA 用于优化 FCM 算法，并且从以下几个方面对此混合聚类算法(IFGA-FCM)也进行了改进：

染色体编码：利用实数编码方式，把聚类中心作为染色体。先随机选取聚类中心，生成 M/2 条染色体，另 M/2 条染色体则由随机生成的模糊聚类矩阵的聚类中心构成。

优质种群从按适应度值降序排列后的正常种群中选择前 μ=20%的个体，劣质种群则选择后 λ =20%的个体。当然，参数 μ 和 λ 值的选择可以根据具体情况而定，此处不一定是最优的取值。

适应度函数：取下式的倒数作为适应度函数

$$F_m(V)=\sum_{k=1}^{n}\left(\sum_{i=1}^{c}d_{ik}^{\frac{1}{1-m}}\right)^{1-m}$$

选择算子：采用简单且常用的适应度比例选择法，并且配合最优保存策略。本文采用将最优与次优个体替换掉最劣个体与次劣个体的最优保存策略，同时记录历史最优个体。

交叉算子：采用线性组合交叉，即如果欲交叉的两条染色体为 x_i、x_j，则交叉后的染色体为：

$$x_i'=r_ix_i+(1-r_i)x_j$$
$$x_j'=r_ix_j+(1-r_i)x_i$$

r_i 为(0,1)上的随机数。

交叉之前，为保证产生有意义的新个体，使算法的收敛性得到改善，所以要进行最短距离基因匹配。

变异算子：正常种群采用多重均匀变异，优质种群采用诱导变异，劣质种群采用大变异。

解码：当 IFGA-FCM 算法运行终止时，得到的最优染色体，即为所对应的数据集的最佳聚类中心 V；将此染色体解码为各聚类中心向量 v_i，然后通过公式：

$$\begin{cases} m_{ik}=\dfrac{1}{\sum_{j=1}^{c}(\dfrac{d_{ik}}{d_{jk}})^{\frac{2}{m-1}}} & I_k=f\text{时} \\ m_{ik}=0 & \forall i\in\overline{I_k},\text{及}\sum_{i\in I_k}m_{ik}=1,\quad I_k\neq f\text{时} \end{cases}$$

计算最优模糊分类矩阵 U，再根据最大隶属度原则确定并输出各数据点的类属。IFGA-FCM 算法总体流程如下：

Step 0：输入聚类样本 X，聚类数 c 及其它相关参数;

Step 1：初始化正常种群及相关参数;

Step 2：初始化优质种群，劣质种群及相关参数;

Step 3：Step 1 与 Step 2 中三类种群按以下各自的遗传策略进化：

Setp 3.1 正常种群执行一般的遗传操作;

Step 3.2 优质种群按适应度比例选择，执行一致交叉与诱导变异;

Step 3.3 劣质种群按适应度比例选择，执行一致交叉与大变异操作;

Step 4：每隔 K 代后，融合三种群，淘汰按适应值降序排列的后 $\mu+\lambda$ 个个体，余下的 n 个个体作为下一次进化的正常种群，并从中选出优质与劣质种群的个体;

Step 5：如果满足终止条件，则结束运行，输出具有最大适应度的个体作为本次计算的最优解，否则转 Step 2 ;

Step 6：聚类中心精确化处理。

IFGA-FCM 算法的主干代码如下：

```
InitializeVar(C,E,k,n,N,NPop,EPop,BPop);
InitializeData(X);
InitializeCenter(Ini_X_center);
CIndividual* nga=new Individual(NPop);  //正常种群
CIndividual* ega=new Individual(EPop);  //优质种群
CIndividual* bga=new Individual(BPop);  //劣质种群
CFcm  fcm;
nga.Initialize();  //初始化正常种群
ega.Initialize();  //初始化优质种群
bga.Initialize();  //初始化劣质种群
fcm.Initialize();
do{ n++;
  if(n%k==0) {
    pmix=Mix(nga,ega,bga);  //种群融合
    ega.Select(EPop,pmix); //选择优质种群
    nga.Select(NPop,pmix); //选择正常种群
    bga.Select(BPop,pmix);  //选择劣质种群
  }
   nga.Evolute();   //正常种群进化
   ega.Evolute();   //优质种群进化
   bga.Evolute();   //劣质种群进化
}while(n<  N&&  abs(ega.CurrentBest-HistroyElitist)>E);
fcm.Computing(ega.CurrentBest);//得到大致全局聚类中
//心后，再进行聚类中心的精确化
fcm.OutPut();
```

算法测试时，随机生成 3 组服从高斯分布的 3 维数据，数据量分别为 5000、10000、15000，并人为加入 2‰的噪声数据，分别测试经简单 GA 优化的 FCM 算法(SGA-FCM)和 IFGA-FCM 算法的性能。算法各自独立运行 20 次，每次运行最大进化代数为数 50 次，交叉率为 0.9，变异率为 0.01；聚类数为 3，聚类有效性采用 Xie-Beni 有效性函数计算；如果在容许的误差范围内能找到聚类中心，则称算法收敛。测试结果见表 1，可以看出：对于大数据集，且由于噪声数据的影响，SGA-FCM 算法的收敛性较差，而 IFGA-FCM 均完全收敛；运行时间上，IFGA-FCM 耗时远比 SGA-FCM 算法少；从聚类有效性上看，IFGA-FCM 的聚类质量远高于 SGA-FCM。

表 1 SGA-FCM 及 IFGA-FCM 算法性能比较

数据量	平均收敛时间比 SGA-FCM/IFGA-FCM	平均聚类有效性比 SGA-FCM/IFGA-FCM	收敛次数 SGA-FCM:IFGA-FCM	
5000	1.651	1.245	20	20
10000	1.920	1.384	17	20
15000	1.847	1.653	16	20

4 交通事故历史数据挖掘实例

将上述的改进算法 IFGA-FCM 用于某市城区

2008 年 2 月至 2008 年 7 月的交通事故数据挖掘，经过数据筛选和清洗后，得到 1054 条有效记录，将其作为挖掘数据集，选取其中最能显著刻画交通事故发生规律的 4 个字段：(月份，时刻，发生地横坐标，发生地纵坐标)进行聚类挖掘。算法参数设置：n=1054，数据维数：d=4，迭代次数：N=60，迭代阈值：Delt=0.001，聚类有效性函数选取常用的 Xie-Beni 函数。对于聚类数 C，可先给定 C 的一个大致范围为 2~10，然后根据最小的聚类有效性值确定最佳聚类数 C 的值。

算法最终挖掘结果为：最佳聚类数 C=5，聚类中心为：C_1(2.358,18.235,232.247,733.531)，C_2(2.816,17.943,657.792,1031.506)，C_3(5.134,12.134,658.790,267.714)，C_4(7.463,12.025,1020.656,1930.978)，C_5(7.562,11.841,431.326,1642.579)。该结果表明，2008 年 2 月至 7 月间，该市交通事故多发生于 2 月，5 月和 7 月的中午和傍晚时段，并集中发生于(232.247,733.531),(657.792,1031.506),(658.790,267.714), (1020.656,1930.978), (431.326,1642.579)五个地理位置附近处。

据此分析，如果在上述时段和地理位置及其附近处有针对性的加强交通警力部署和交通标牌设施的设置，可以有效地减少交通事故的发生，挖掘结果也表明，改进后的聚类挖掘算法是很有效的。

5 结束语

数据挖掘技术应用于公安业务数据，对于指导公安实战工作具有一定的指导意义，并具有较好的应用前景，可以利用其构建具有智能性和主动性的信息平台，以提升信息化办案水平，本文的后续工作将会在这方面作一些相关的研究。另外，挖掘前业务数据的预处理及算法的较优参数选取将会在很大程度上影响挖掘的质量，后续工作也会在这方面作进一步的探究。

参考文献

1 薛京生,孙济洲,孙宇,何宏.基于应急事件响应的模糊聚类分析算法.计算机工程,2006,32(1):201－202.

2 高新波.模糊聚类分析及其应用.西安:西安电子科技大学出版社, 2004.49－55.

3 Zadeh LA. Fuzzy sets. Information and Control. 1965,(8):338－353.

4 杨兴春,李进.一种基于多种群隔代融合的遗传算法.计算机与数字工程, 2008,36(5):30－32.