数据挖掘平台中模糊聚类算法的研究与应用[1]

摘　要　数据挖掘是当前很多领域的一个研究热点，引起了大量研究人员的关注。本文具体研究了基于目标函数的模糊聚类算法（FCM）[1]，并对聚类效果的有效性和参数选择进行了详细分析。最后将该算法应用于模型生命表的制作中。

关键字　数据挖掘；模糊聚类；有效性分析

引言

数据挖掘是目前国际上数据库和信息决策领域最前沿、最活跃的研究方向之一。本文的研究主体是数据挖掘方法中的基于目标函数的模糊聚类算法（FCM），重点是对聚类效果的分析。其中，有效性分析的目的是得到理想的聚类数，使聚类结果最佳地反映数据集的结果；加权指数的分析是为了得到最佳的聚类模糊性。关于有效性的实现是目前该算法的一个重点和难点问题，文中针对该问题运用一组实验数据对效果进行了分析。本文还将该算法应用于模型生命表的制作中，取得了很好的效果。

1　基于目标函数的模糊聚类算法

1．1　基于目标函数的模糊聚类算法（FCM）的基本原理

设集合 $X=\{x_1,x_2,\ldots,x_n\}$ 中元素有 m 个特征，即 $x_i=(x_{i1},\ldots,x_{im})$。要把 X 分为 c 类（$2\le c\le n$）。设有 c 个聚类中心 $V=\{v_1,v_2,\ldots,v_c\}$，其中 $v_i\in\{v\mid v=\sum_{i=1}^{n}a_ix_i,a_i\in R,x_i\in X\}$。取 $d_{ik}=\|x_k-v_i\|=[\sum_{j=1}^{m}(x_{kj}-v_{ij})^2]^{1/2}$ 为样本 x_k 与聚类中心 v_i 的欧氏距离，那么理想的分类显然是使目标函数 $J(U,V)=\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}(d_{ik})^2$ 为极小的 U。其中，u_{ik} 表示样本 x_k 对于聚类中心 v_i 的隶属度。

1．2　FCM 算法的实现方法

为了灵活地变动元素的相对隶属程度，把目标函数更一般化为：

$$J(U,V)=\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^r\|x_k-v_i\|^2$$

其中 $r\ge 1$，是待定的参数，$\|\cdot\|$ 是 R^m 空间中的任一种范数。

步骤如下：

（1）取定 c，$2\le c\le n$；取定终止条件 ε；取初始化聚类中心 $V^{(0)}$；逐步迭代（$l=0,1,2,\ldots$）；

（2）对于 $V^{(l)}$，修正 $U^{(l)}$

[1]本文是国务院人口普查办公室重点招标项目（国人字 12 号）的一部分。

$$\frac{1}{\sum_{j=1}^{c}\left(\frac{\|x_k-v_i\|}{\|x_k-v_j\|}\right)^{\frac{1}{r-1}}},\quad \forall i,\forall k,\ \|x_k-v_i\|\neq 0 \text{ 且 } \|x_k-v_j\|\neq 0$$

$$u_{ik}^{(l+1)}=\begin{cases}1 & \text{当}\|x_k-v_i\|=0\\ 0 & \text{当}\|x_k-v_j\|=0\end{cases}$$

（3）计算聚类中心 $v_i^{(l0}=\dfrac{\sum_{k=1}^{n}(u_{ik}^{(l)})^r x_k}{\sum_{k=1}^{n}(u_{ik}^{(l)})^r}$

（4）用一个矩阵范围比较$V^{(l)}$与$V^{(l+1)}$，对取定的$\varepsilon>0$（ε一般取 0.001 和 0.01 之间），若$\|V^{(l+1)}-V^{(l)}\|\le\varepsilon$，则停止迭代，否则$l=l+1$，转向（2）。

1．3　结果的清晰化

本算法迭代所得的 U 是一个模糊划分矩阵，对应着 X 的模糊划分，可用下述两种方法使划分清晰化，得到 X 的普通分类：

方法 1.$\forall x_k\in X$,若$\|x_k-v_{i0}\|=\min_{1\le i\le c}\|x_k-v_i\|$，则将 x_k 归入第 i_0 类。其中 v_{i0} 是第 i_0 类的聚类中心。也就是说，x_k 与哪一个聚类中心最接近，就将它归到哪一类。

方法 2.在 U 的第 k 列中，若$u_{i_0k}=\max_{1\le i\le c}(u_{ik})$，则将 x_k 归入第 i_0 类。也就是说，x_k 对哪一类的隶属度最大，就将它归入到哪一类。这一方法实际上就是最大原则方法。

2　FCM 聚类算法的效果分析

2．1　初始聚类中心的研究

初始聚类中心的选择一般有两种方法：

①随机法：即随机选取前 c 个点作为初始聚类中心。

这种方法的迭代次数多，收敛速度慢，而且可能使结果为局部最优解。

②爬山法：

算法如下：　1．选取第一个点为第一个聚类中心；

2．选出离第一个点最远的那个点为第二个聚类中心；

3．For　i = 3 to c，选出离已有的聚类中心的距离之和最大的那个点为第 i 个聚类中心。

本文采用了爬山算法，在于其能够明显减少迭代次数，并加快聚类速度。而且，能够有效的防止得到局部最优解。

2．2　有效性的研究

由于聚类是无人监督的，因此必须对聚类结果的有效性进行研究，就是应该把数据集分成几类才是最好的，才能最佳反映数据集的结构[6]。有效性问题可以转化为最佳类别数c的确定，基本思想如下：

1.事先给定聚类数的范围$[c_{\min},c_{\max}]$，最佳聚类数在该范围中取得。

2.For c = $c_{\min}$ to $c_{\max}$（或则 For c = $c_{\max}$ to $c_{\min}$）

2.1 初始化聚类中心 V

2.2 应用 FCM 算法更新模糊分类矩阵 U 和聚类中心 V

2.3 判断收敛性，如果没有，转 2.2

2.4 通过有效性指标函数计算指标值$V_d(c)$

3.比较各有效性指标值，最大（或最小）指标值$V_d(c_f)$所对应c_f的就是所求的最佳聚类数。

现有的聚类有效性函数按其定义方式可分为两大类：基于数据集模糊划分和基于数据集集合结构。其中，基于数据集模糊划分理论基础是：好的聚类分析对应于数据集较“分明”的划分。这一类有效性函数包括分离系数V_{PC}[2]和分离熵（平均信息量）V_{PE}[2, 3]。它们的优点是简单、运算量小，适用于本身已经较分明且数据量小的数据集。但是，与数据集的结构特征缺乏直接联系，对于类间有交迭的数据不能很好的处理；基于数据集集合结构的理论基础是：每个子类应当是紧凑的，而且子类间是尽可能分离的。这一类的有效性函数有：Xie-Beni有效性V_{xie}[4]，Fakuyame-Sugeno有效性V_{FS}[1]，Rhee-Ho有效性V_{RH}[1]，Rezaee-Letlieveldt-Reiber有效性V_{RLR}[5]，Sun. H-S. Wang-Q. Jiang有效性V_{WSJ}[6, 7]等。这类方法是在类内紧凑度与类间分离度之间找一个平衡点，以获得最好的聚类。

2．3　加权指数 r 的研究

Bezdek[13]引入了加权指数r，又称为平滑因子，控制着模式在模糊类间的分享程度。由于r的选取影响着目标函数的凹凸性，而且控制聚类的模糊性，因此r的取值必然会对模糊聚类的性能产生重要的影响。然而最佳r的选取目前尚缺乏理论指导。当$r \to 1$，最终分类有较小的模糊性，当r>2，逐渐增大时，最终分类将具有较大的模糊性。从聚类有效性的实验研究中得到r的最佳选取区间为[1.5,2.5]，一般采取折中方案取r=2。

3　FCM 算法模块在数据挖掘平台上的实现

3．1　数据源

选择数据来源：可以是 ACCESS、Foxpro、SQLServer、Oracle 等常用数据库，还可以是 Excel 数据表，Text 文挡的数据等非数据库格式但很常见常用的文件。数据源的提取如图 1，图 2 所示：

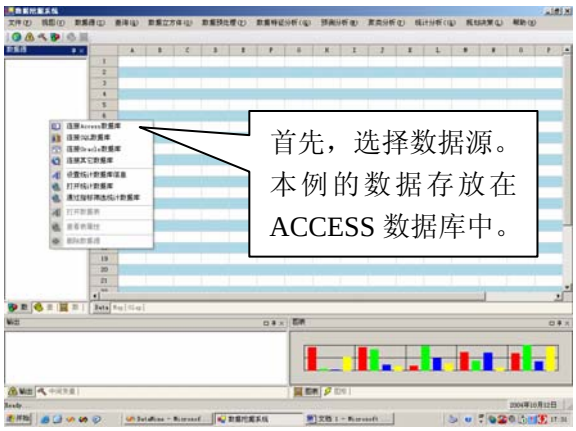

图 1

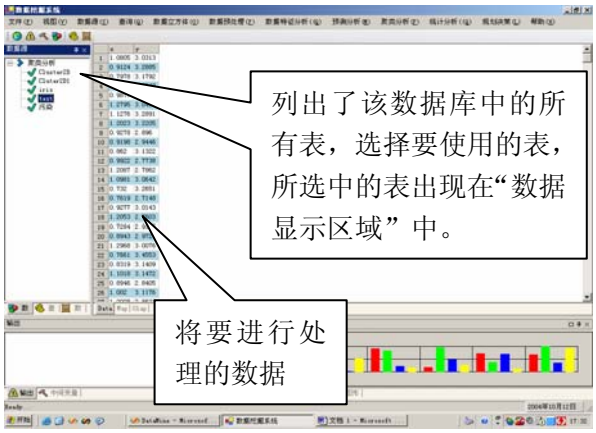

图 2

3．2　参数选择

数据准备就绪后，就可以选择具体的算法。

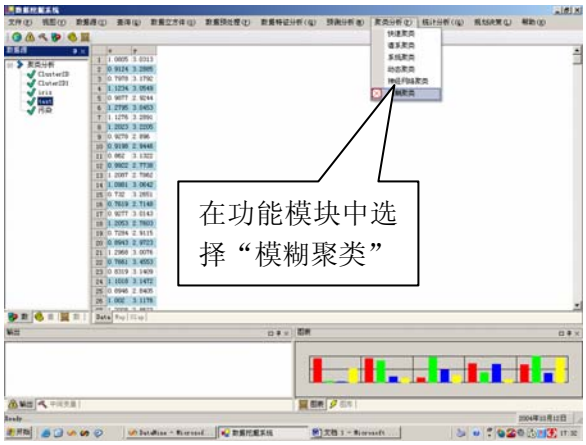

图 3

图 4

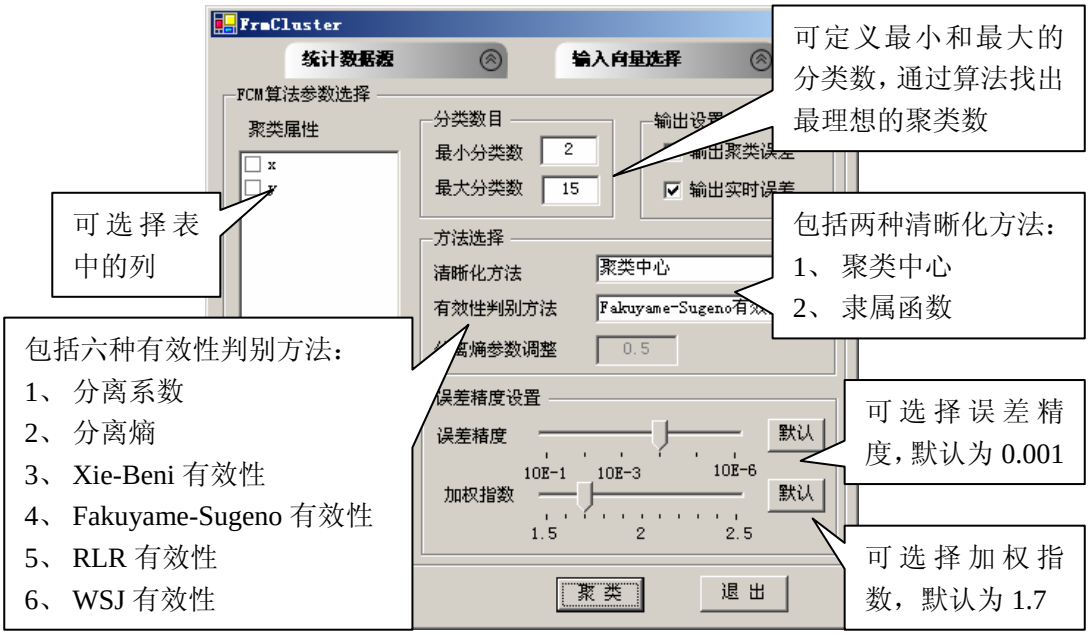

图 5 参数选择对话框

3．3　结果表达

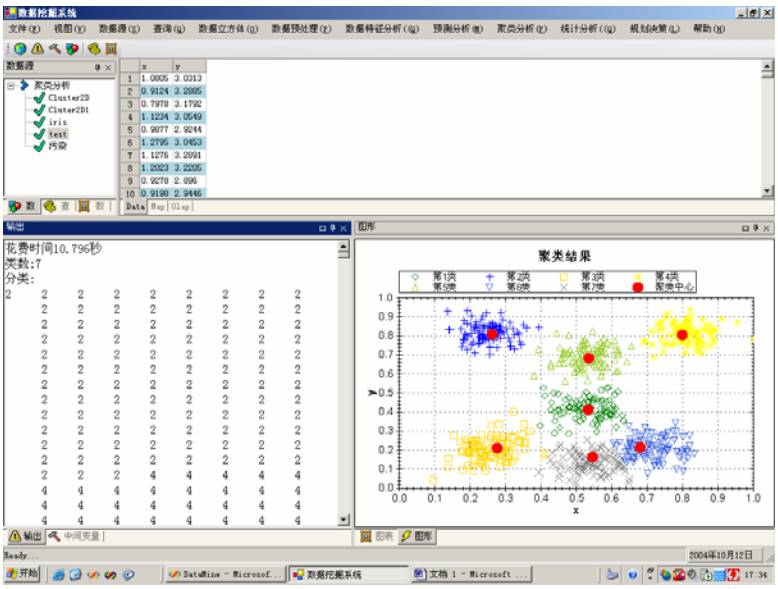

图 8

3． 4 数据分析

在前面的基础上，不改变参数，选择不同有效性方法，得到不同的结果如下：

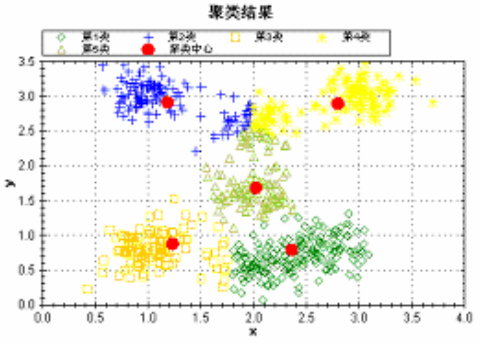

图 9　当参数不变时，RLR 有效性的聚类效果

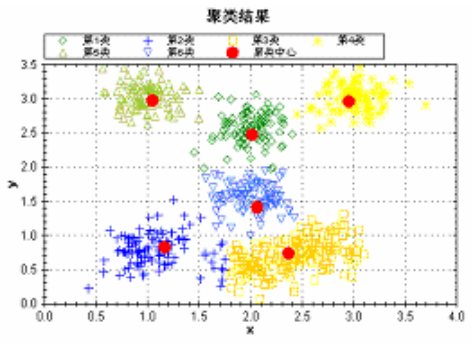

图 10 当参数不变时，WSJ 有效性的聚类效果

由此可见，在相同参数下，不同判别方法的聚类效果相差甚大。为寻找最佳的聚类效果，可通过调整误差精度和加权指数来实现。在聚类过程中，可观察聚类误差和实时误差两项动态指标来确定最佳的判别方法和参数设置。

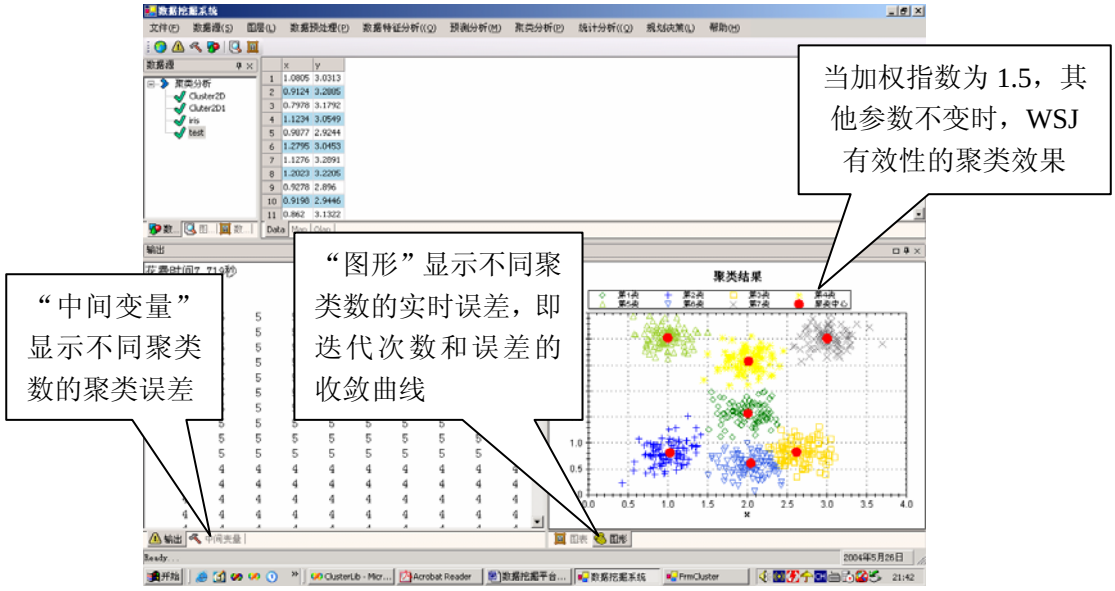

图 11

4 应用案例

以下是该算法在模型生命表制作中的一个应用。

数据来源是第五次人口普查。在对全国 2871 个县区数据建立的 2870 张原始生命表（除西沙群岛）进行一定条件的筛选和合并的基础上，得到 1615 组调整后的县区分年龄段生存死亡数据表。再以这 1615 个合并后的县区数据，生成原始生命表。

在该案例中，利用这 1615 张原始生命表对死亡概率进行聚类，使死亡概率相近的县区归到同一类中，目的是按死亡概率将全国分为若干个区域，建立模型生命表。

由于聚类的属性——分年龄段的死亡概率——是多维的，在多次试验论证的基础上，文中选择了零岁组和 1-5 岁组作为图形结果显示的两个维度。结果如图 12 所示，得到四个类。通过对聚类结果的观察和对具体数据的分析，发现红线区域的点离其对应的聚类中心的距离较大，应将它们归为独立的一类，即最终聚类结果为五类。通过对比，发现经过聚类得到的结果与《中国分类(区域)模型生命表》的分类方式吻合，即分为西南、华中和华东、华北、东北、新疆五个主要地域，其中新疆作为一个特殊的区域独立为一类。通过合并，各类包含的县区数和死亡概率分析如表 1。

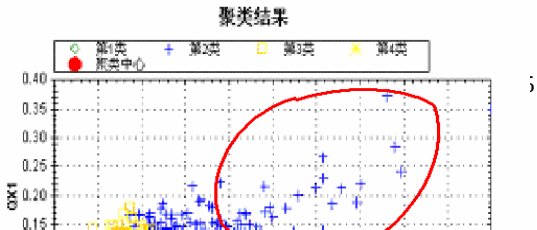

表 1

类别	县区数	死亡概率描述
西南	285	高死亡率
华中和华东	905	低死亡率
华北	172	次低死亡率
东北	187	次高死亡率
新疆	66	极高死亡率

图 12 聚类结果图

结论

本文通过对基于目标函数的模糊聚类算法（FCM）的研究与应用，得到以下主要结论：

一、研究发现，目前所有的有效性判别方法都不具有绝对的优势，都需要适应和针对具体的数据集而进行选择。同时，也发现不存在对任何数据都有效的最佳加权指数，而只能根据数据的具体情况选择较为合适的加权指数。鉴于此，本文在算法的实现上给出了聚类误差和实时误差两项动态指标，通过对这两项指标的观察，选择出了具有较好有效性的判别方法及其对应的最佳加权指数。

二、本文的创新应用之处还在于将算法应用于中国分类模型生命表制作之中。通过将 2000 年第五次人口普查中的 1680 个县区的按年龄段死亡概率进行聚类分析，并划分为 5 组，其结果既与《中国分类(区域)模型生命表》的分类方式吻合（即划分为西南、华中和华东、华北、东北、新疆五个主要地域），也为基于 2000 年人口普查数据的中国分类地区模型生命表的研制和开发提供了重要的聚类算法基础。

参考文献

[1]李相镐,李洪兴,陈世权,等. 模糊聚类分析及其应用. 贵州:贵州科学技术出版社,1994.

[2]Bezdek.J.C. Pattern Recognition with Fuzzy Objective Function Algorithms. New York:Plenum Press,1981

[3]范九伦,裴继红,谢维信. 聚类有效性函数:熵公式. 模糊系统与数学,1998-12(3)

[4]Xie XL,Beni G. A validity method for fuzzy clustering.IEEE PAMI,1991.

[5]Ramze Rezaee M, Lelieveldt B P F and Reiber J H C, A New Cluster Validity Index for the Fuzzy c-mean, Pattern Recognition Letters, (Netherlands) Mar 1998.

[6]Sun.H,S Wang,Q.Jiang Anew Validation Index for Determing the Number of Clustering in a Data Set. Washington DC.July 14-19,2001

[7] Qingshan Jiang. New Validation Index for Determing the Number of Clustering in K-Means Clustering. 2002.

Study and Application of Fuzzy Clustering Algorithm in Data Mining Platform

Abstract Data Mining is a hot topic for many research areas currently, which attracts the interest of many researchers. This article systematically describes the Fuzzy C-Means(FCM) Clustering Algorithm based on objective function. The effectiveness and parameter selecting of FCM are analyzed in detail. Finally the successful application of FCM on Model Life Table is realized.

Keywords Data Mining;Fuzzy Clustering; Effectiveness Analysis