模糊蚁群聚类在信息数据判别中的应用①

张 奎

(中石化管道储运分公司聊城输油处，聊城 252000)

摘 要：随着信息技术的高速发展，无论是商业企业、科研机构或是政府部门，在过去若干年的时间里都积累了大量的，并以不同形式存储的信息数据资料。针对信息数据十分复杂，仅靠数据库的查询检索机制和统计学方法远不能满足人们需要的问题，提出一种改进的模糊聚类新方法，能够准确、智能地将信息数据转化为明确的专家知识。最后将之与传统方法作对比，突出这种方法对信息数据判别的优势。

关键词：信息技术；数据资料；模糊聚类；信息判别

Application of Fuzzy Clustering Ant Colony to Information Data Discriminant

ZHANG Kui

(Sinopec Pipeline Storage and Transportation Company, Liaocheng 252000, China)

Abstract: With the rapid development of information technology, the commercial enterprises as well as scientific research institutions or government departments, in the past few years, have accumulated lots of different information datas. In view of the complex information data, only the database query retrieval mechanism and statistics method can not meet people' needs. This paper puts forward an improved fuzzy clustering method, which can put the information data into the accurately, intelligence and clear expert knowledge. Finally, the paper compares it with the traditional method and highlights its advantages.

Key words: information technology; data information; fuzzy clustering; information distinguish

在经济高速发展的今天，各类项目层出不穷，为了能使项目高效率运作，评审成为了一个重要的环节，不仅是考核项目承担方水平的重要手段，而且也是体现公平公开的重要途径。每次评审前编排评审名单是一项非常繁重的工作，往往花费大量的时间去搜集专家资料，要做大量的重复工作。同时，在实际生活中，由于不是时刻与专家保持密切接触，不能及时的更新资料，造成了专家信息管理的极大不便。而且由于传统的专家库系统只是简单的针对数据库的调用。这些都促进了对传统的评审选取制度和专家管理的改革[1]。

而随着计算机在政务等领域的广泛应用，国家为了规范评审和咨询工作，充分发挥科学和技术专家的作用，很多地方都出台了相应的专家管理办法，做到更好的与专家交流，更加公平、合理、公正、公开地进行各类评审，更好地选择突出的人才。同时也为了让工作人员从单调、繁重的重复工作中解脱出来，更多地投身于其他工作和提高专家管理水平，提高工作效率。国内已经推出了许多类型的专家信息系统。而近年来数据库技术与人工智能的交叉学科的迅速发展又为数据库的进一步发展提供了新的途径[2]。

1 信息判别系统的研究现状

目前国内外对自动判别专家知识领域的研究一直是热点，大多的研究成果主要集中在基于数据挖掘相关理论的研究，也取得了较大进展[3]，虽然数据挖掘理论给信息自动判别系统奠定了前所未有的理论基础，但在实际应用中并没有达到我们所需求信息知识的高精度判别，因此，如何能够在判别精度上满足人

① 收稿时间:2011-11-11;收到修改稿时间:2011-12-20

们的要求成为社会各行各业的研究重点。

数据挖掘主要利用了来自如下一些领域的思想：(1)来自统计学的抽样、估计和假设检验，(2)人工智能、模式识别和机器学习的搜索算法、建模技术和学习理论。数据挖掘也迅速地接纳了来自其他领域的思想，这些领域包括最优化、进化计算、信息论、信号处理、可视化和信息检索。一些其他领域也起到重要的支撑作用。

但是专家所写的论文，项目材料等文献资料是很容易收集且保存的，系统自动判断出这些文档材料的领域，则专家的研究领域就能够判断出来[4]。如图 1 所示：

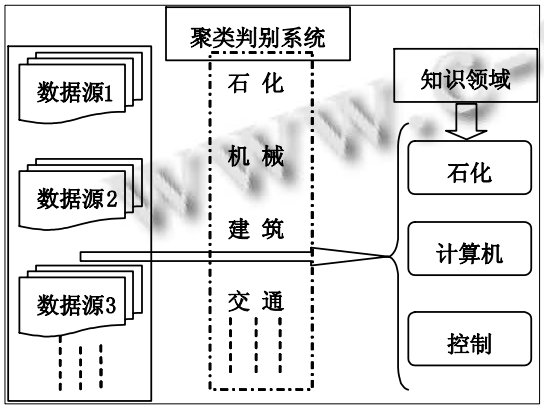

图 1 自动判别信息材料的所属知识领域

2 信息判别系统的聚类方法实现

专家的知识领域就是专家所研究的领域。根据专家知识领域，系统可以整理、分析并从中抽取适合评审的专家，对专家信息的管理和评审工作的开展起到了重要作用，使工作效率明显地提高，在实际管理中有着非常重要的意义。专家知识领域当然可以自己填写或选择，但是如果专家信息系统可以自动判别出专家的知识领域，将大大节省人力资源。根据什么来判断出专家的知识领域呢？如果专家没有自己选择知识领域，对于一个陌生的专家，评审部门从专家的其它

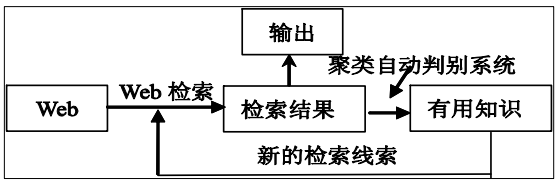

图 2 数据信息判别系统的总流程图

情，更何况还要对专家按领域进行分类。本文针对以上问题，提出一种基于模糊聚类的信息自动判别的方法[5]，将文档按某种搜索要求分成若规则，每一个规则对应一类信息数据。图 2 中明确了本文的研究点。

对数据信息判别研究，前人已经提出了许多聚类算法，其大体思想是：给出聚类指标函数，使下面的目标函数达到最小：

$$J=\sum_{k-1}^{n}\sum_{i-1}^{c}(\mu_{ik})^{w}(d_{ik})^{2} \tag{1}$$

约束条件：$\sum_{i=1}^{c}\mu_{ik}=1,(k=1,2,\cdots n)$，其中 $w\in(1,\infty)$ 是一个加权指数，d_{ik} 为欧几里德形式表示的数据点 x_k 与聚类中心 v_i 是 m 维向量：$v_i=[v_{i1},v_{i2},\cdots v_{im}]$，$v_i$ 的第 j 个特征值为：

$$v_{ij}=\frac{\sum_{k-1}^{n}(\mu_{ik})^{w}x_{kj}}{\sum_{k-1}^{n}(\mu_{ik})^{w}} \tag{2}$$

这里 $j=1,2,\cdots m$。

而在信息数据判别的研究上，各种聚类算法在不同的应用领域中均表现出了不同的性能，也就是说，很少有一种算法能同时适用于若干个不同的应用背景。针对这一问题，提出一种新的聚类方法即模糊蚁群聚类。将其思想归纳如下：

在蚂蚁群寻找食物的过程中，若找到了目标它们就会产生一定量的信息素。这将会影响两个分散的聚集点上的蚂蚁，也会以此点为中心向其它方向运动，在此过程中也制约了其它蚂蚁的运动行为。正是由于信息素所产生的作用，给后面蚂蚁带来很大方便，这样会大大减少寻找食物的时间。用聚类算法的表达为：数据点的分类用转移概率计算，进而更换新的聚类中心位置，从新计算，直到符合要求停止[6]。

设 X 表示聚类集合，为目标所求集合，v_j 表示聚类中心，$d_{ij}=P(X_i-Y_i)$，式中 d_{ij} 为 X_i 到 Y_i 的距离；P 加权系数。

在 t 时刻，对于数据集 A，第 k 只蚂蚁在寻找食物的方向 (i,j) 上释放信息素量时，后面的蚂蚁将分别以 i 和 j 为中心以 r 为半径扩散，这时蚂蚁 k 所释放的信息量。

定义：

$$\tau_{ij}^{k}(t)=\begin{cases}\dfrac{Q}{d_{ij}} & d_{ij}\le R\\ 0 & d_{ij}>R\end{cases} \tag{3}$$

蚂蚁所释放的信息量$W\tau_{ij}^{k}(t)$表示如下：

$$W\tau_{ij}^{k}(t)=\begin{cases} r\dfrac{Q}{d_{ij}}\left(1-\dfrac{d_{ij}\bullet(d_{ij})^{\omega}}{\hat{d}^{\omega+1}}\right) & d_{ij}\le R \\ 0 & d_{ij}>R \end{cases} \tag{4}$$

以 j 为中心扩散的计算过程与上相同。式中 $\omega>1$； $P_{ij}(t)\ge P_0$； R 表示初始设置的聚类半径；设 $\tau_{ij}(0)=0$。

转移概率 P_{ij} 的计算式：

$$P_{ij}(t)=\frac{\tau_{ij}^{\alpha}(t)\mu_{ij}^{\beta}(t)}{\sum_{s\in S}\tau_{ij}^{\alpha}(t)\mu_{ij}^{\beta}(t)} \tag{5}$$

这里，$S=\{X_s\ d_{sj}\le R, s=1,2,\cdots,n\}$ 表示在 v_j 周围的数据集；α 被遗弃信息的百分比，β 表示目标实现系数，μ_{ij} 为 t 时刻蚂蚁由 A 到 B 处的信息素大小。假如 $P_{ij}(t)\ge P_0$（P_0 为初始时所赋的值）。

则有：$CS_j=\{X_i\ d_{ij}\le R, i=1,2,\cdots,J\}$，为所有归结到 v_j 的集合，J 数据点数目。下面两个公式用来计算更新后的 v_j 和 u_{jl}。

$$v_j=\frac{\sum_{l=1}^{J}u_{jl}^{m}X_i}{\sum_{l=1}^{J}u_{jl}^{m}} \tag{6}$$

$$u_{jl}=\frac{1}{\sum_{p=1}^{k}\left(\dfrac{d_{jl}}{d_{pl}}\right)^{2/(m-1)}} \tag{7}$$

数据点聚类时与中心的分离度 ε_j 和本次计算的误差总和 ε 分别由下面两个式子得到：

$$\varepsilon_J=\frac{1}{J}\sum_{i=1}^{J}(X_i-v_j) \tag{8}$$

$$\varepsilon=\frac{1}{J}\sum_{J=1}^{k}\varepsilon_J \tag{9}$$

此算法的优点在于：它采用的是全局优化式，能根据聚类信息量把数据清晰分开，兼顾了以往研究方法的在搜索过程中易产生局部极小情况，而使信息判别正确率极低。模糊蚁群聚类算法的另一优势在于：算法将蚁群寻食理论与模糊聚类理论相结合，较原来的判别系统相比，提高了系统的灵活性，减少了聚类分析时间即信息系统判别时间，从而达到对信息数据快速、准确的判别。

3　几种聚类算法的仿真研究

为了评价改进后的聚类性能，选择 IRIS 数据进行实验。将模糊蚁群聚类算法与传统的 k-means 和 c-means 聚类算法做对比，所有实验均使用 MATLAB7.1 环境下进行仿真研究。文献中提供的著名的 IRIS 实际数据作为测试样本集，IRIS 数据由 150 个样本组成[7]，分别隶属于 3 个不同的类别，在数据集中 IRIS 数据与其他两类间较容易分离，所以此数据常被作为标准的测试数据。仿真结果如下：

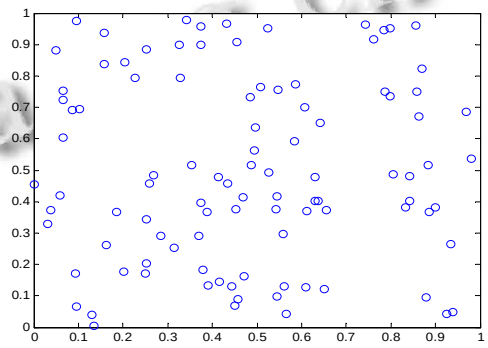

图 3　二维随机数据图

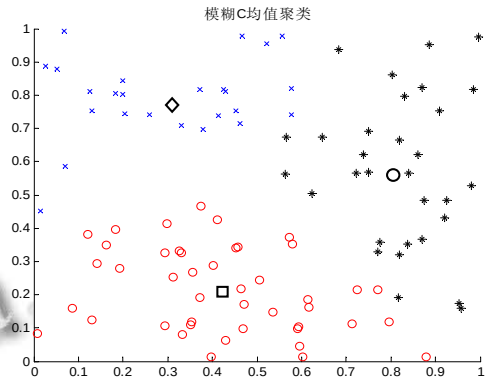

图 4　k-means 聚类结果

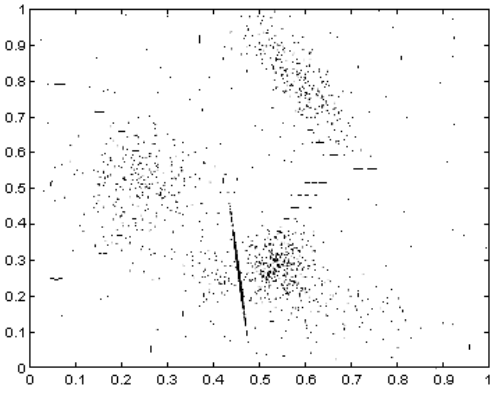

图 5　fuzzy c-means 聚类结果

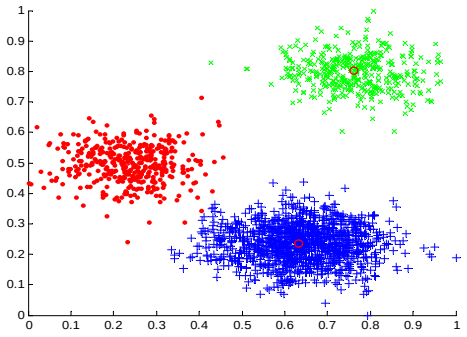

图 6 模糊蚁群聚类结果

4 聚类性能对比分析

从上面的分析研究不难看出，这些现有的聚类算法在不同的应用领域中均表现出了不同的性能，也就是说，很少有一种算法能同时适用于若干个不同的应用背景。总体来说，c-means 算法的应用最为广泛[8]，其收敛速度快，且能够扩展以用于大规模的数据集；缺点在于它倾向于识别凸形分布、大小相近、密度相近的聚类，而不能发现形状比较复杂的聚类，并且初始聚类中心的选择和噪声数据会对聚类结果产生较大的影响。K-means 不仅适用于任意属性和任意形状的数据集，还可以灵活控制不同层次的聚类粒度，因此具有较强的聚类能力，但它大大延长了算法的执行时间；此外，对聚类结构不能进行回溯处理。机器学习中的人工神经网络和模拟退火等方法虽然能利用相应的启发式算法获得较高质量的聚类结果，但其计算复杂度往往较高，同时其聚类结果的好坏也依赖于对某些经验参数的选取。在针对高维数据的子空间聚类和联合聚类等算法中，虽然通过在聚类过程中选维、逐维聚类和降维从一定程度上减少了高维度带来的影响，但它们均不可避免地带来了原始数据信息的损失和相应的聚类准确性的降低，因此，寻求这类算法在聚类质量和算法时间复杂度之间的折中也是一个重要的问题。

5 结语

基于上述分析,得到聚类各个方法的比较结果,如表 1 所示：

表 1

算法	加权指数	聚类正确比率（%）
K-means	1；1.8	83.23；85.12
FCM	1.2；2.0；4.0	88.24；79.25；91.56
模糊蚁群聚类	1.2；2.0；4.0	93.15；95.33；89.12

由上表数据可以看出，本文提出的基于模糊蚁群聚类算法，可以很好的弥补前人研究果的算法的不足。我们从算法仿真图上可看到，采用 FCM 和 K-means 方法划分得到的数据在清晰度上有明显的混合交叉，而模糊蚁群聚类算法可以把数据清晰的划分为三类，并且三类数据划分的界限很明显，无交叉和混杂，验证了此方法对数据信息的自动判别的有效性。

参考文献

1 龙军.国家科技奖励综合业务处理平台研究.长沙:中南大学,2005.

2 Zadeh LA. Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans. on Systems, Man, and Cybernetics, 1973, 28–41.

3 高正源.XX 市科技咨询专家管理信息系统的研发.重庆:重庆大学,2007.

4 Wang L, Langari R. Complex systems modeling via fuzzy logic. IEEE Trans. Syst, Man, Cybem, 1996,6(1):100–106.

5 杨占华.聚类分析研究及其在文本挖掘中的应用.成都:西南交通大学,2006.

6 Bezdek JC, Keller JM, Krishnapuram R, et a1. wIll the RealIRIS Data Please Stand Up. IEEE Trans. on Fuzzy System, 1999,7(3):368–369.

7 Jain AK, Dubes RC. Algorithms for Clustering Data. Prentice Hall, 2010.

8 雷筱珍,赖万钦.一种基于信息素的 FCM 聚类算法.安阳工学院学报,2009:12–16.