计算机科学2002Vol. 29№. 6

一种基于遗传算法的聚类新方法*)

A New Algorithm for Clustering Analysis Based on Genetic Algorithm

张 伟[1,2] 廖晓峰[1] 吴中福[1]

(重庆大学计算机科学与工程学院 重庆400044)[1] (重庆教育学院计算机与现代教育技术系 重庆400067)[2]

Abstract Data Mining aims at big data in large database. In this paper, we present a new algorithm for clustering analysis based on genetic algorithm. There are two characteristics in our methods. Firstly, the algorithm is general-purposed and our cluster analyzer can cluster large data set with mixed numeric and categorical attributes. Secondly, it improves the efficiency of data mining and the quality of the knowledge.

Keywords Data Mining, Genetic Algorithm, Clustering

1 引言

数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。数据挖掘更广义的说法是[1]：数据挖掘意味着在一些事实或观察数据的集合中寻找模式的决策支持过程。人们把原始数据看作是形成知识的源泉，就像从矿石中采矿一样。原始数据可以是结构化的，如关系数据库中的数据，也可以是半结构化的，如文本、图形、图像数据，甚至是分布在网络上的异构型数据。发现知识的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。发现了的知识可以被用于信息管理、查询优化、决策支持、过程控制等，还可以用于数据自身的维护。因此，数据挖掘是一门广义的交叉学科，它汇聚了不同领域的研究者，尤其是数据库、人工智能、数理统计、可视化、并行计算等方面的学者和工程技术人员都投入很大的精力和人力进行研究，并取得了可喜的成果。例如：目前已有不少 KDD 的原型系统、实用化系统和开发工具出现，已广泛地应用于 Internet 网、市场销售预测、金融投资、社会保险、医学、地质等领域。例如 Rogian 大学的 KDD-R 已被用于医学数据分析和电信工业的市场分析；Kansas 大学开发的基于 Rough 集理论的学习系统 LERS，被美国 NASA 的 Johnson 空间中心作为专家系统开发工具用于医学及全球气候变化分析；Lock Head Martin 公司的 AI 中心开发的 Recon 系统用来辅助预测某种股票的趋势或推断是否可能出现异常变化等等，特别要指出的是，数据挖掘技术从一开始就是面向应用的，它不仅是面向特定数据库的简单检索查询调用，而且要对这些数据进行微观、中观乃至宏观的统计、分析、综合和推理，以指导实际问题的求解，企图发现事件间的相互关联，甚至利用已有的数据对未来的活动进行预测。

聚类是数据挖掘的一个重要内容和基本形式之一，聚类通过比较数据的相似性和差异性，能发现数据的内在特征及分布规律，从而获得对数据更深刻的理解与认识，所以聚类的挖掘技术受到了科技界的广泛关注。目前的聚类算法大都是适用于数值属性或符号属性中的一种，而我们研制的一种基于遗传算法的聚类新方法，可以对包含数值属性和符号属性的大数据集进行聚类，适用面很广。本文重点分析了该算法以及应用实例。

2 遗传算法

聚类问题是一个全局最优问题，鉴于遗传算法具有计算简单、优化效果好的特点，以及它在处理组合优化问题方面所具有的优势，本文利用遗传算法设计了一种新的聚类方法。为了描述的方便，下面简述一下遗传算法。

遗传算法[2](Genetic Algorithm)是美国 Michigan 大学 John Holland 教授根据生物进化论和遗传学的思想提出的一种全局启发式优化算法，它利用遗传算子(选择、交叉和变异)，促进解集合类似生物种群在自然界中自然选择、优胜劣汰、不断进化，最终收敛于最优状态。

遗传算法的实现步骤为：1)对求解空间进行编码，初始化；2)初始化种群 $X(i)=(x_1,x_2,\cdots,x_N)$，$i=1,2,\cdots,m$；3)对当前种群 X(i)中每个染色体 x_i 计算其适应度 $f(x_i)$，适应度表示了该个体适应性能；4)应用遗传算子产生新一代群体 X(i+1)；5)判断终止条件，不满足返回。

遗传算法中遗传算子包括：(1)选择算子(Selection)。从群体按概率选择个体，种群 X(i)中个体 x_i 的选择概率与其适应度成正比，如轮盘赌模型；(2)交叉算子(Crossover)。将选中的一对个体基因型按概率进行交叉，如单点或多点交叉；(3)变异算子(Mutation)。将个体基因的等位基因按概率进行变异。

遗传算法比较适合于传统搜索方法所不能解决的复杂问题和非线性问题。然而，实际应用中遗传算法受编码、迭代次数、种群规模的限制，造成种群多样性和选择性压力的调和冲突，即强选择性压力导致遗传搜索过早收敛，强种群多样性导致遗传搜索效率低下。遗传算法的改进应考虑到：为了保证算法能全局收敛，必须保护种群的多样性；另一方面，为了加快算法的收敛，必须使种群中个体尽快向最优解聚集。

3 一种聚类新算法

3.1 问题描述

从空间 X 中给定一个有限的取样点集合(或从数据库中取得有限例子的集合)，聚类的目标是将数据聚集成类，使得

*)重庆市科技计划项目资助。张 伟 博士研究生，主要研究方向为远程教育，数据挖掘。廖晓峰 教授，博士生导师，主要研究方向为网络安全，数据挖掘。吴中福 教授，博士生导师，主要研究方向为远程教育，网络通信，数据挖掘。

类间的相似性尽量小，而类内的相似性尽量大。

假定算法把 n 个向量 $X_j(j=1,2,\cdots,n)$ 分为 c 个组 $C_i(i=1,2,\cdots,c)$，并求每组的聚类中心，使得非相似性（或距离）指标的目标函数达到最小。当选择欧几里德距离为组 i 中向量 X_j 与相应聚类中心 C_i 的非相似性指标时，目标函数可定义为：

$$\min \quad J(u,c)=\sum_{i=1}^{c}J_i=\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})\|X_j-C_i\|^2$$

$$\text{s.t.} \quad \sum_{i=1}^{c}u_{ij}=1,\forall\ j=1,2,\cdots,n$$

这里，u_{ij} 是向量 X_j 属于组 C_i 的程度，它的值介于0和1之间。

$$u_{ij}=\begin{cases}1 & \text{对每个 } k\neq i\text{，如果}\|X_j-C_i\|^2\leqslant\|X_j-C_k\|^2\\ 0 & \text{其它}\end{cases}$$

3.2 算法设计

由于知识表示等方面的原因，一般的遗传算法通常不适合于大型数据库的分析处理，但是我们可以使用改进的遗传算法 MGA（a Modified Genetic Algorithm）来解决问题。在 MGA 中，我们使用聚类中心来表示组成问题的结果，亦即一个聚类中心向量表示为一个基因，而所有聚类中心则组成染色体。这种表示更适合问题的特征，同时由于聚类中心在数量上远远小于数据向量，因而使用 MGA 可以大大减少分析量，从而提高了算法效率。

在我们的数据挖掘研究中，使用 MGA 主要是提高非线性优化的效率以及研究推理机制的应用。图1中给出了 MGA 的算法流程图。

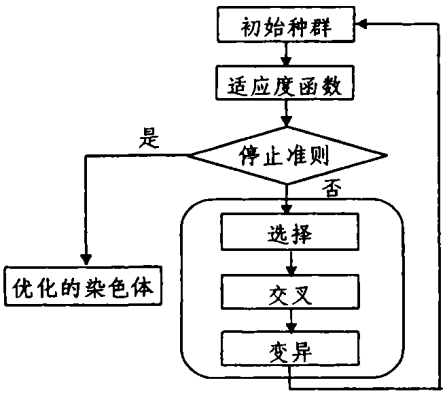

图1 MGA 的算法流程图

初始种群。MGA 的初始化包括设置聚类大小 m、种群和聚类中心表。在 MGA 的迭代过程中，由于聚类中心的组合，聚类数目会变小，因而 m 值不能很小，通常设置为数据向量的一半。种群大小与聚类大小 m 相等，基因设置则可以从数据向量中随机选择。聚类中心表保留当前结果集中所有的聚类中心，同时也保留数据向量集合，这些数据向量是按照一定规则得到，属于相应的聚类中心。使用聚类中心表的好处是可以使用最后一次计算的结果，从而减少了计算量。

适应度函数。MGA 使用向量距离的离差平方和 DSE（Distance Square Error）作为适应度函数。由于 MGA 在每一次迭代过程都保留了聚类中心表，因此 DSE 的计算非常简单：

$$J(u,c)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}\cdot Dis(X_j,C_i)$$

当 $J(u,c)$ 小于某个预先指定的参数 ε（DSE 门限）时，算法就结束。

遗传算子。MGA 使用两种遗传算子：交叉和变异，两种遗传算子在 MGA 中的使用不是任意的。在交叉操作中，应对相同或距离接近的聚类中心进行合并，从而减少种群大小。变异操作不仅会改变值域，而且能够进入向量内部，从而产生一个新的聚类中心。相对于领域知识来说，这两种算子是 MGA 的关键。

（1）**交叉**：交叉操作如图2所示。交叉频率和起始位置由参数 β_c 控制。如果介于0和1之间的随机数小于 β_c，将选择两个染色体，然后进行交换。起始位置由 β_c 和更小的结果大小的乘积决定。交换长度由参数 α_c 控制。在图2中，父染色体是 P1 和 P2，子染色体是 C1和 C2。假定交叉操作从第 i 个基因开始，在第 j 个基因结束。

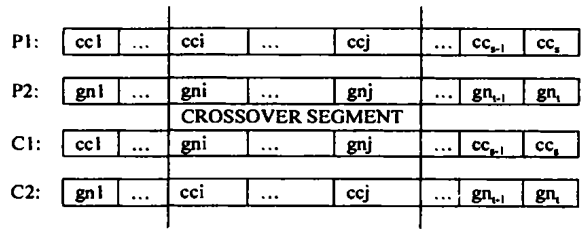

图2 MGA 中交叉操作

在 MGA 中，交叉操作的结果与双亲一起保留。至于它们是否被最终保留则由适应度函数决定，这样一来，种群就更加丰富了。如果在交叉操作结果中，一些基因有重复，它们将合并以使种群数量减少。

（2）**变异**：变异操作如图3所示。变异频率和变异位置由参数 β_m 控制。如果介于0和1之间的随机数小于 β_m，将选择一个染色体进行变异操作。变异起始位置由 β_m 和更小的结果大小的乘积决定。

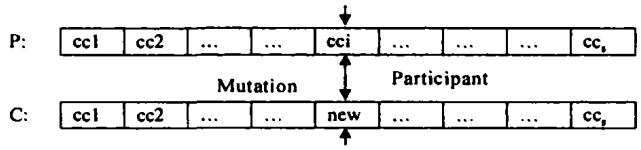

图3 MGA 中变异操作

在图3中，P 是变异前的染色体，C 是变异后的染色体。指针指向变异前的基因 cci 和变异后的新基因 new。通常情况下，变异后的新基因 new 不属于以前基因的值域，它是变异前的基因值发生改变后的结果。

在 MGA 的每一迭代过程，不只是聚类中心进行调整，由聚类确定的数据向量也要进行调整以找到其合适的类别。同时，在迭代过程的每一步还要再次计算聚类中心表来适合变异后的数据向量。

得到优化的染色体。在 MGA 的每一迭代过程，算法将计算每一候选染色体的适应度函数值，并由此产生新的后代染色体。重复执行迭代过程，就能得到聚类结果，直到到达停止规则。这样一来，每个聚类有一个聚类中心 Z_i 和一个最大半径 R_i。

$$R_i=\max(\|X_j-Z_i\|),\quad i=1\cdots c,j=1\cdots n_i$$

最后，得到聚类规则。即是

$$if\ (\|X_j-Z_i\|\leqslant R_i)\ \text{then}\ X_j\in C_i$$

3.3 算法分析

遗传算法在理论上还远未成熟，遗传算法的设计常靠经

验，算法的收敛性研究也才开始。但是，大量实验和理论证实，遗传算法用于优化领域效果显著。MGA 使用了遗传算法的基本思想，并根据领域知识设计了遗传算子。虽然遗传算法的效率不是很高，但是 MGA 比 ISODATA 更为简单。MGA 的算子代替了聚类组合、聚类分裂和聚类再计算等复杂的数学计算。

4 实验结果

实验中，我们使用 MGA 来处理一个汽车贸易公司的汽车数据集。数据包含有五个属性：1个数值属性和4个符号属性。数值属性有参考价(万元)，符号属性有编码、品牌、车型和类别。表1中，我们只用了12条记录来说明数据集。

表1 汽车描述

编码	品牌	车型	参考价(万元)	类别
0101001	奔驰	S级	60	进口车
0101002	奔驰	S级	60	进口车
0101003	奔驰	S级	60	进口车
0401001	捷达	1.6升	15	中档车
0401002	捷达	1.6升	15	中档车
0401003	捷达	1.6升	15	中档车
0401004	捷达	1.6升	15	中档车
0201001	富康	1.6升	10	中档车
0202002	富康	1.4升	8	中档车
0203001	富康	1.8升	12	中档车
0301001	派力奥	1.5升	10	经济型车
0302001	派力奥	1.3升	8	经济型车

实验中的初始参数为：$m=6$，$\alpha_c=4$，$\beta_c=0.7$，$\beta_m=0.2$，$\varepsilon=6$。当系统收敛到挖掘聚类规则的命令后，它把通过 SQL 检索到的数据映射到用户感兴趣的子空间。子空间只有两个属性：品牌和参考价。因为品牌不是一个数值属性，它需要正规化。通过一个预先定义的正规化函数 f(x)来进行正规化：f(奔驰)=1，f(富康)=2，f(派力奥)=3，f(捷达)=4，从而得到正规化后的向量(1,60)，(1,60)，(1,60)，(4,15)，(4,15)，(4,15)，(4,15)，(2,10)，(2,8)，(2,12)，(3,10)，(3,8)。

根据这些向量，我们用 MGA 来初始化种群和聚类中心表。然后对种群使用交叉和变异操作，从而产生新的聚类中心表。当一次迭代过程结束时，使用适应度函数来计算种群，并为下一次迭代保留一些结果，直到种群满足适应度函数。最后的聚类中心如表2所示。

表2 结果表

聚类中心	向量	DSE
(2.5,9)	(2,8),(2,10),(3,8),(3,10)	4.47
(1,55)	(1,50)	5.00
(4,14.5)	(4,14),(4,15)	1.00

最后，我们得到如下聚类规则：

$if(\|X-(2.5,9)\|\leqslant 1.12)$ then $X\in C_1$

$if(\|X-(1,55)\|\leqslant 5)$ then $X\in C_2$

$if(\|X-(4,14.5)\|\leqslant 0.5)$ then $X\in C_3$

结束语 数据挖掘技术包含很多知识领域，但是，由于它的发展历史不长，还没有完整的理论，因此，目前的数据挖掘技术还存在很多问题。我们在以前的一些研究成果基础上，提出了一种基于遗传算法的聚类新方法。该方法有两个优点：一是通用性强，它可以对包含数值属性和符号属性的大数据集进行聚类；二是它提高了数据挖掘的效率和质量。

参考文献

1 Weldon J-L. Data Mining and Visualization. Database Programming and Design, 1996, 9 (5): 21～24

2 Michalewicz Z. Genetic Algorithms + Data Structures = Evolution Programs. Springer-Verlag, 1996

3 Robertson G G. A Table of Two Classification Systems. Machine Learning, 1988, 3(23)

4 Jain A K, Dubes R C. Algorithms for Clustering Data. Prentice Hall, 1988

5 Murty M N, Jain A K. Knowledge-based Clustering Schema for Collection Management and Retrieval of Library Books. Pattern Recognition, 1995, 28(7)

6 Shortland R, Scarfe R. Digging for Gold (Data Management). IEE Review, 1995, 41 (5): 213～217

7 Fayyad U. Knowledge Discovery and Data Mining towards a Unifying Framework, KDD'. In: 96 Proc. 2nd Intl. Conf. on Knowledge Discovery & Data Mining, AAAI Press, 1996

8 Han J. Conference tutorial notes: Data Mining Techniques. In: Proc. of ACM SIGMOD Intl. Conf. 96 on Management of Data (SIGMOD' 96). Montreal, Canada, June, 1996

9 Data Mining: Discovering Hidden Value in Your Data Warehouse. Pilot Software. WWW files, 1996 (http:// www. pilot. com/ dmpaper/ dmindex. html)

10 陈栋，徐洁盘. Knight：一个通用知识挖掘工具，计算机研究与发展，1998，35(4)：338～343

(上接第136页)

CLIPS，分析了它的知识表示方式，规则结构和软件架构中四个部件的实现。我们下一步的工作是以此为基础，考虑如何将规则语言和程序设计语言集成，使规则能够在程序设计语言的上下文中引用定义的数据结构。

参考文献

1 Buchanan B G, Shortliffe E H. Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Addison-Wesley, 1984

2 Brownston L, et al. Programming Expert Systems in OPS5: An Introduction to Rule-Based Programming. Addison-Wesley, 1985

3 ILOG Rules User's Manual, ILOG Corp, 1996

4 Shaw M, Garlan D. Software Architecture: Perspectives on an emerging discipline. Prentice-Hall, 1996

5 Hayes-Roth F. Rule-based systems. Communications of the ACM, 1985, 28(9): 921～932

6 Hayes-Roth B, Pfleger K, et al. A domain-specific software architecture for adaptive intelligent systems. IEEE Transactions on Software Engineering, special Issue on software Architecture, 1995, 21(4): 288～301

7 Hayes-Roth B. Architectural foundations for real-time performance in intelligent agents. The Journal of Real-Time Systems, Kluwer Academic Publishers, 1990, 2: 99～125

8 Miranker D P. Performance estimates for the DADO machine: A comparison of TREAT and Rete. In Fifth Generation Computer Systems, ICOT, Tokyo, 1984

9 Miranker D P, Lofaso B J. The Orgnization and performance of a TREAT-Based Production System Compiler. IEEE Transactions on Knowledge and Data Engineering, 1991, 3(1): 3～9

10 Scales D. Efficient matching algorithms for the SOAR/OPS5 production system, Knowledge Systems Laboratory, Department of Computer Sciences, Stanford University, Stanford, CA, 1986

11 Forgy C L. Rete: a fast algorithm for the many pattern/many object pattern match problem. Artificial Intelligence, 1982, 19: 17～37

12 Nii H P. Blackboard systems. AI Magazine, 1986, 7(3): 38～53, 7(4): 82～107