基于Tabu搜索的聚类算法研究*)

钟 将 吴中福 吴开贵 杨 强
(重庆大学计算机学院 重庆400030)

摘 要 聚类分析的两个基本任务是分析数据集中簇的数量以及这些簇的位置。大多数的聚类方法通常只关注后一个问题。为了在聚类数不确定的情况下实现聚类分析,本文提出了一种新的结合人工免疫网络和Tabu搜索的动态聚类算法—DCBIT。新算法主要包含两个阶段:先使用人工免疫网络算法获得一个候选聚类中心集,然后使用Tabu搜索在候选聚类中心集上实现动态聚类。仿真实验结果表明与现有方法相比,新方法具有更好的收敛概率和收敛速度。

关键词 动态聚类,人工免疫网络,Tabu搜索

A Novel Dynamic Clustering Algorithm Based on Tabu Search

ZHONG Jiang WU Zhong-Fu WU Kai-Gui YANG Qinag
(Computer College of Chongqing University, Chongqing 400030)

Abstract Cluster analysis aims at answering two main questions: how many clusters there are in the data set and where they are located. Usually, the traditional clustering algorithms only focus on the last problem. In order to solve the two problems at the same time, this paper proposes a novel dynamic clustering algorithm called DCBIT, which is based on the immune network and Tabu search. The algorithm includes two phases, it begins by running immune network algorithm to find a candidate clustering center set, and then it employs Tabu search to search the optimum number of clusters and the location of each cluster according to the candidate centers. Experimental results show that the new algorithm has satisfied convergent probability and convergent speed.

Keywords Dynamic clustering, Artifical immune network, Tabu search

1 引言

聚类分析作为一种可以从研究对象的特征数据中发现有用规则的无监督学习方法,已经广泛应用于数据挖掘、图像分割、模式识别、网络入侵检测等诸多领域,并取得了令人满意的效果。聚类分析主要解决两个问题:数据集中存在多少聚类簇以及这些簇的位置。如果聚类分析前已知簇的数量就被称为静态聚类,反之,要在聚类过程中计算簇的数量就被称为动态聚类[1]。使用 p 维的特征向量集 $X=\{x_1,x_2,\cdots,x_n\}\subset R^p$ 来表示要分析的数据集合,聚类分析就是寻找对象集 $V=\{v_1,v_2,\cdots,v_k\}$,它将 X 划分成 k 个子集,即 k 个簇。

目前最常用的聚类算法是 K-means 聚类算法,该方法算法简单,收敛速度快,但对初始值的选取敏感,容易陷于局部最优。研究者使用各种方法来避免算法陷入局部极值,如Krovi,Hall 等提出了使用遗传算法[2,3]来改进 K-means 聚类的方法,2002年 Hong-Bing Xu 等提出使用 fuzzy tabu search 的方法来改进 K-means 聚类[4],2003年行小帅等使用免疫规划的方法,并对算法的收敛性进行了证明[5]。

以上方法能够较好解决静态聚类问题,对于动态聚类问题,本文巧妙地将人工免疫网络和Tabu搜索结合起来,提出了一种新的算法。该方法的基本思想是:首先通过人工免疫网络分析数据集的轮廓性分布特征,然后使用Tabu搜索求解最优的簇数量及中心位置。仿真试验表明新算法不仅可以发现聚类簇的数量,有效避免局部极值,而且收敛速度比传统的聚类算法快。

文中的第2节介绍获取候选聚类中心集的人工免疫算法;第3节给出了基于候选聚类中心集的聚类算法;第4节给出了新算法的完整描述;第5节是在两个人工数据集和 Iris 数据集上的仿真实验结果;最后总结全文。

2 获得候选聚类中心集

2.1 算法背景

近年来,研究人员从不同的角度模拟生物免疫网络的工作原理来进行数据分析[7]。2000年 Leandro Nunes de Castro 等人提出使用演化的免疫网络来实现聚类分析[8],算法模拟生物免疫系统中抗体克隆选择机制,将数据集作为抗原集合,而生成的抗体集合作为聚类结果。通过仿真实验发现,该算法难以确定聚类簇的数量,聚类中心也常常偏离正确的位置。由于该算法减少了数据的冗余,且生成的抗体集合能够反映数据集的总体分布特征,因此本文在该算法的基础上作适当的修改,用来获得候选聚类中心集。

2.2 算法描述

算法1 候选聚类中心集求解算法

输入:数据集 X,以及免疫抑制阈值 t_s。

输出:抗体集合 M,也即候选聚类中心集 R。

*)基金项目:国家自然科学基金资助(No. 60073047)。钟 将 讲师,博士生,主要研究方向为网络安全、免疫计算。吴中福 教授,博士生导师,主要研究方向为计算机网络与通、宽带综合业务数字网。吴开贵 副教授,博士后,主要研究方向为密码学,网络安全。杨 强 博士生,主要研究方向为SVM,图像处理。

基本步骤：

step1 随机产生一个抗体集合 Abs，令 $M=\{\}$。

step2 按照随机顺序选择 X 中每一个抗原 Ag，进行以下运算。

2.1 计算 Ag 与 Abs 中抗体的距离；

2.2 选择距离 Ag 最近的 k 个抗体，构成临时抗体集合 $tmpM$；

2.3 $tmpM$ 上执行克隆运算，每个抗体最大的克隆数为 c；

2.4 $tmpM$ 上进行退火变异；

2.5 删除 $tmpM$ 中与 Ag 的距离最远的部分抗体；

2.6 在 $tmpM$ 上进行免疫抑制操作；

2.7 将 Ag 插入到 $tmpM$ 最前面；

2.8 将 $tmpM$ 添加到 M 中；

2.9 在 M 上进行免疫抑制操作；

step3 循环结束条件判断。若不满足终止条件，令 $Abs=M$，$M=\{\}$ 转到步骤2执行，否则返回，结束算法。

2.3 计算复杂度分析

假定抗体集合中最大的抗体数量为 m，步骤2.1到2.9总的计算量为 $O(m)$，抗原数据集大小为 n，那么步骤2的计算复杂度为 $O(m*n)$。步骤2最大循环次数为 t，算法总的计算复杂度为 $O(t*m*n)$，即 $O(m*n)$。经过归范化后的数据集 $X\subset[0,1]^r$，如果免疫抑制参数为 t_s，那么 m 的最大取值为 $[(2/t_s)]^r$。

3 基于 Tabu 搜索的动态聚类

如果已经获得一个候选聚类中心集 R，动态聚类过程就是在 R 上选择一个子集 C^*，使 C^* 满足 $f(c^*)=\min_{\forall C\subset R}(f(C))$。其中函数 $f(C)$ 表示 C 作为 K-means 聚类算法的初始聚类中心的聚类评估函数。基于 R 的聚类问题可看成一个组合优化问题，因此可使用遗传算法(GA)，模拟退火(SA)，Tabu 搜索等算法来解决此问题。TS(Tabu Search)算法作为一个最优工具来求解非线性覆盖问题，到目前为止作为一种通用的启发式最优技术，Tabu 搜索在许多领域已经取得了令人瞩目的成功。尽管缺乏理论证明，但多数实际应用表明：Tabu 搜索的收敛速度比 GA 和 SA 快，而且可获得更好的解。因此本文使用 TS 方法来实现动态聚类分析。TS 方法通过利用灵活记忆的特殊形式来避免搜索陷入局部最优。该算法在结合具体应用时需要定义：解的编码方式，邻近解的产生方式和 Tabu 表的结构。

3.1 算法中的编码方式及评估函数

本文采用二进制的编码方式来表示聚类的初始聚类中心集 C。假设 R 是数据集的一个候选聚类中心集，R 中包含的候选初始聚类中心数 $r=|R|$，可以使用长度为 r 的二进制串表示如何在 R 中选择初始聚类中心集 C，其中每一位表示 R 中相对应候选中心是否包含在 C 中。例如，如果 $Cb=001100$，则表示 R 含了6个候选聚类中心，C_b 使用了 R 中的第3和第4个对象作为聚类中心。

目前常用的动态聚类评估函数有：Modified Hubert Static(MHT)，Davies-Bouddin (DB)，Dunn's CS(DCS)[9]。本文采用 Davies-Bouddin (DB)方法来评价聚类的质量，该方法的评估值越小，其聚类的质量也越好(本文中的 $v_{DB,q,t}$ 函数中的参数 $q=2,t=2$)。

3.2 邻近解集的产生方式

若当前最优解为 I，其邻近解集 $N(I)$ 可通过'移动'操作来生成。文中定义了两种基本'移动 $m_-(j)$ 和 $m_+(j)$，分别表示移去或添加 R 中的第 j 个候选簇中心。例如 $I=001111$，$I\oplus m_-(6)=001110$，$I\oplus m_+(1)=101111$。

本文还定义了交换'移动' $m_e(i,j)$，即将 i 和 j 位上的取值互换，显然只有当第 i 和第 j 位上的取值不同，交换操作才有意义。交换'移动'可以通过 $m_-(j)$ 和 $m_+(j)$ 来实现。例如 $I=001111$，其邻近解 $I'=I\oplus m_e(1,6)=101110$，同样 $I'=I\oplus m_-(6)\oplus m_+(1)$。

3.3 Tabu 表的结构

为了避免循环搜索问题，Tabu 搜索将最近若干次移动的反向移动存放在 Tabu 表中。由于邻近解集都可以通过最基本的两种移动来产生，因此在 Tabu 表中只需要存放基本的移动即可。

由于 $m_-(j)$ 与 $m_+(j)$ 之间是互为反向的'移动'操作，当 I 通过 $m_-(j)$ 得到 I'，就在 Tabu 表中就添加一条记录 $m_+(j)$，反之亦然。如果 I' 是通过交换'移动'获得，那么在 Tabu 表中就需要添加两条记录。例如当前的解 $I=001111$，如果通过交换移动 $m_e(1,6)$ 获得邻近解 $I'=101110$，由于该交换'移动'等价于两次基本移动 $m_+(1)$ 和 $m_-(6)$，因此在 Tabu 表中需存放 $m_-(1)$ 和 $m_+(6)$ 两条记录。

3.4 Tabu 搜索中的重要参数

Tabu 表的长度将影响算法的行为，如果采用较大的 Tabu 表，有利于扩大搜索空间，如果采用较小的 Tabu 表，将有利于在当前最优解附近搜索。因此很多 Tabu 搜索算法采用自适应的方式，即在算法不同的运行期更改 Tabu 表的大小[4]。为简化起见，本文采用固定长度的 Tabu 表，长度为 $\lfloor\sqrt{r}\rfloor+1$，r 为 R 中对象的数量。

为防止有价值的'移动'在 Tabu 表中而被限制执行，TS 中设计了释放水平函数，用来释放那些有价值的'移动'。为了使算法在初始阶段倾向于扩大搜索空间，而在算法后期则倾向于在已知最优解附近搜索，本文定义满足条件：

$$(f(I)-f(I\oplus m))/f(I)\geqslant\nabla/K \tag{1}$$

为有价值移动，其中 ∇ 为一个常数(本文设置为0.02)，K 为算法迭代的次数，因此定义释放水平函数为：

$$A=f(I_{best})*(I-\nabla/K) \tag{2}$$

3.5 基于 Tabu 搜索的动态聚类算法

算法2 基于 Tabu 搜索的动态聚类算法

输入：候选聚类中心集 R。

输出：聚类中心集 C。

主要步骤如下：

step1：随机生成一个长度为 $|R|$ 的二进制串作为初始解 I_b，计算 I_b 的评价函数。令 $I_{best}=I_b$，初始化 Tabu 表以和释放水平 A，设置迭代计数器 $K=0$；

step2：如果 K 等于最大的迭代次数，将 I_{best} 所代表的聚类中心集作为 K-means 算法的初始聚类中心集，并输出最终的聚类结果 C；否则令 $K=K+1$，并转到 step3；

step3：通过定义的移动操作来产生邻近解集的一个子集 $N(I_{best})$，并从中选出评价函数最优的一个邻近解 $I'=I_{best}\oplus m$，如果 $f(I')<f(I_{best})$ 且 m 不在 Tabu 表中，或者 m 在 Tabu 表中，但是 $f(I')<A$，令 $I_{best}=I'$，更新 Tabu 表以及释放水平

A 并转到 step2；否则转 step4；

step4：在 $N(I)$ 中选择产生的移动不在 Tabu 表中，且评估值最小的解 I^L，如果 $f(I^L)<f(I_{best})$，且更新 Tabu 表和释放水平 A。转到 step2。

4 DCBIT 算法的完整描述

下面给出基于 Tabu 搜索的动态聚类算法(DCBIT，Dynamic Clustering Algorithm Based on Artifical Immune Network And Tabu Search)的完整描述。

算法3 DCBIT 算法

输入：数据集 X 和免疫阈值 t_s

输出：最佳的聚类中心集 C

step1：使用算法1求解 X 的聚类候选中心集 R。

step2：删除数据集上和候选聚类中心集的异常对象。(是可选步骤)

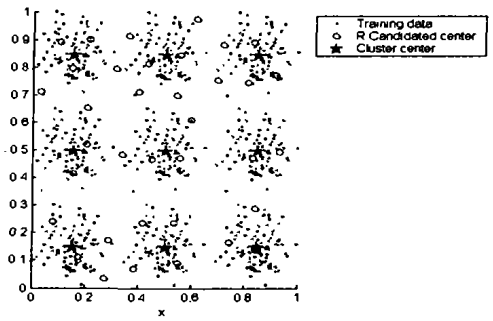
如果一个抗体其半径为 t_s 的邻域中识别的抗原数目小于指定值就将抗体从免疫网络中清除。

step3：使用算法2在候选聚类中心集 R 寻找最优的初始聚类中心集合 C^*。

step4：使用 K-means 方法获得聚类中心集 C，并将 X 划分成 $k=|C|$ 个簇。

5 仿真实验

5.1 人工数据集上的实验结果



(数据集●，候选聚类中心○和聚类中心☆)

图1-a 数据集1的一次试验结果

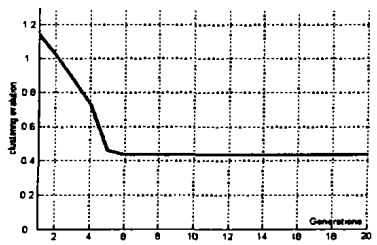
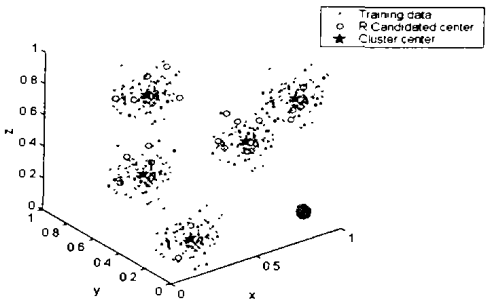


图2-a 最佳的聚类评估值的变化情况

5.3 算法比较

动态聚类的目标是确定数据集中簇的数量以及这些簇的位置，因此可以从四个方面来考察算法的性能：V_{best}，能够获得的最佳的聚类评估值；V_{avg}，所有聚类评估值的平均值；T_k，收敛到正确聚类数目的次数；T_b，收敛到最佳评估值的次数。本文分别使用 GA[3]，aiNet[7]以及 DCBIT 在 Iris 数据集上试验50次。试验中遗传算法和 DCBIT 最多迭代40次。BHCM 算法的结果采用文[8]中提供的值。

为了直观显示实验结果，首先使用两个在簇中心附近产生的高斯分布的人工数据集试验。数据集1是包含9个簇的二维数据集，数据集2是包含5个簇的三维数据集。每个簇包含100个样本。两个数据集使用的免疫抑制参数分别为 $ts=0.1$ 和 $ts=0.15$。



(数据集 ，候选聚类中心○和聚类中心☆)

图1-b 数据集2的一次试验结果

在两个数据集上分别进行了50次试验，发现算法100%得到正确的聚类数量 k，并收敛到最佳的聚类中心的位置。其中一次的试验结果见图1所示。对于数据集1，候选聚类中心集的规模平均为37。从图1可见，候选聚类中心集 R 几乎覆盖了整个数据空间，甚至在稀疏的区域上存在候选聚类中心，这有利于提高算法的收敛性。

5.2 IRIS 数据集上的实验结果

为了进一步验证算法的有效性，使用 UCI 机器学习库中的 Iris 数据集，该数据集有4个特征维，故不能直观显示其聚类结果。数据集的物理标识分为3类，但是根据其数据特征，第2类和第3类之间区域相互覆盖。因此其最佳的聚类的数量一直存在争论。根据文[8]，使用 $\nu_{DB,22}$ 评价函数，IRIS 数据集的最佳的聚类数量应为2。

使用 $ts=0.25$ 作为免疫网络的抑制参数，候选聚类中心的平均大小为16。动态聚类阶段一般在小于10代就可获得与文[8]提供的结果，即簇的数量为2，评估值 $\nu_{DB,22}=0.46$。实际在多数情况下算法收敛到更优解0.44，图2表示其中一次动态聚类过程中最佳的聚类数量以及最佳的聚类评估值的变化情况。

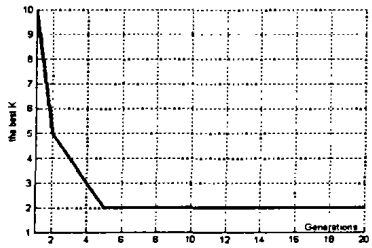


图2-b 最佳的聚类数量的变化情况

表1 算法性能比较(N/A：文献未提供)

Algorithms	T_A	V_{best}	V_{avg}	T_k	T_b
GA[2]	50	0.44	0.77	22	17
aiNet[7]	50	0.44	0.46	50	20
DCBIT	50	0.44	0.44	50	48
BHCM[9]	N/A	0.46	N/A	N/A	N/A

表1中的实验结果说明，直接使用 GA 尽管可以获得最优

(下转第189页)

表1

	数据点——用(x,y)坐标表示
A^B	(51.7608,76.7328)
	(54.6401,76.1603)
	(59.2785,75.0582)
	……
	(79.3904,65.3305)
	(82.2814,63.6042)
	(84.2326,62.4981)

从表1可以看出，主曲线 A^B 是一个数据点集，而数据点集中的第一个点和最后一个点就是主曲线的两个端点。而本文正是通过对主曲线端点进行分析从而提取出指纹的细节特征点(算法思想在本节最后阐述)。接下来我们分析指纹线图5(b)的主曲线(见图7)。

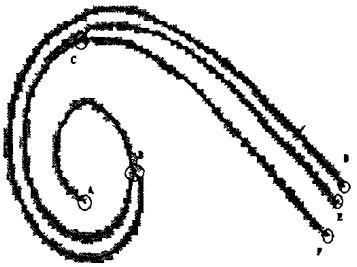


图7

通过图7可知，图5(b)中的指纹线的骨架一共由5条主曲线(A^B,B^C,B^D,C^E,C^F)组成，这五条主曲线分别是五个数据点集，每个数据点集的第一个和最后一个点就使曲线的两个端点。

基于以上的实验结构，我们在这一节的最后一部分提出提取细节特征点算法：首先遍历主曲线的数据点集的两个端点；如果端点只在一个数据集中出现，那么这个点是纹线端点；如果端点在三个数据集中出现，那么这个点是纹线分叉点。

通过上述算法，可以得到图6中的点 A 和 B 以及图7中的点 A 是纹线端点；图7中的 B、C 两点是纹线分叉点；而图7中的点 D、E、F 由于是指纹的边界，所以过滤它们，不作为细节特征点提取。

结论 在本文，我们提出了利用主曲线来进行指纹细节特征点提取的方法，这是一种新的提取指纹特征点的手段。从实验的结果看，该方法对大部分的指纹线都取得了比较好的效果，当然，对一些比较复杂的指纹线，例如曲率比较大的指纹线以及指纹中心部分，还有一些不尽如人意的地方，这些地方需要我们在下一步的工作中进行改进。通过对所做实验进行分析，我们认为利用主曲线对指纹的细节特征点的提取在理论和实践上都是可行的。我们计划在今后的研究中完善算法以便取得更好的效果，同时我们还要将这种方法与传统的特征提取方法进行比较，分析它与传统方法在效率和性能上的优劣。除此之外，我们还计划把主曲线扩展到对指纹的宏观特征分析上，以及实现指纹的识别。

参 考 文 献

1 Hastie T. Principal Curves and surfaces. Laboratory for Computational Statistics. Stanford University, Department of Statistics: [Technical Report 11]. 1984

2 Kegl B. Principal Curves: Learning, Design, and Applications: [Dissertation for Ph. D]. 1999

3 Lin Hong. Automatic personal identification using fingerprints [D]: [Dissertation for Ph. D.]. Michigan State University, 1998. 5～46

4 Hong L, Jain A. Integrating faces and fingerprints for personal identification. IEEE-PAMI, 1998, 20(12) :1295～1307

5 Hrechak A K, McHugh J A. Automated fingerprint recognition using structural matching. Pattern Recognition, 1990,23(7):893～904

6 Verma M R, Majumder A K, Chatterjee B. Edge Detection in fingerprint. Pattern Recongnition, 1987,20(5):513～523

7 Malleswara Rao T Ch. Feature extraction for Fingerprint classification. Pattern Recognition, 1976,8:181～191

8 张军平，王珏. 主曲线综述[J]. 计算机学报，2003,26(2):129～146

9 尹义龙，宁新宝，张晓梅. 改进的指纹细节特征提取算法. 中国图象图形学报，2002,7(12):1302～1306

10 张雄，贺贵明. 基于宏观曲率的指纹特征提取和分类. 计算机研究与发展，2003,40(3)

(上接第174页)

评估值，由于其搜索空间大，因此收敛到最优解的概率较低。BHCM 算法通过尝试所有可能的取值来发现最优的聚类数目，由于直接使用 *K*-means 方法，因此没能发现最优评估值 0.44。aiNet 方法存在收敛到最优评估值的概率较低的问题(即偏离理想聚类中心集)，尽管在 Iris 数据集上收敛到正确聚类数目的概率较高，但对于本文中的两个人工数据集，收敛到正确聚类数目的概率约为0.5。这表明新算法不仅能够获得正确的聚类数目，且在同样的计算量(*K*-means 的计算次数)下收敛到最优值的概率最高，即收敛速度更快。

结论 文中给出了一个全新的动态聚类算法，它通过免疫网络算法快速地求出一个候选聚类中心集，然后使用 Tabu 搜索在候选聚类中心集上实现动态聚类。通过仿真试验证明了新算法具有较高的收敛概率，并极大地提高了动态聚类过程的收敛速度。

在大量实际的聚类应用中，如网络入侵检测，数据挖掘，由于事先很难获得理想的聚类簇数量，因此本算法将具有广阔的应用前景。

参 考 文 献

1 Karkkainen I, et al. Dynamic local search for clustering with unknown number of clusters[A]. IEEE 16th Intl. Conf. on Pattern Recognition[C], Quebec CANADA,2002,2:240～243

2 Hall L O, et al. Clustering with a genetically optimized approach [J]. IEEE Trans. on Evolutionary Computation,1999,3(2):103～112

3 Krovi R. Genetic algorithms for clustering: a preliminary investigation. IEEE In: Proc. of the Twenty-Fifth Hawaii Intl. Conf. on System Sciences, 1992,4:540～544

4 Xu Hong-Bing, et al. Fuzzy tabu search method for the clustering problem. In: IEEE In: Proc. of the first Intl. Conf. on Machine Learning and Cyberneteics, 2002. 876～880

5 行小帅，等. 基于免疫规划的 k-means 聚类算法. 计算机学报，2003, 26(5):605～610

6 Timmis. Artificial immune system: an novel data analysis technique inspired by immune network theory. Wales university, 2001

7 de Castro L N. An Evolutionary Immune Network for Data Clustering[A]. In:Proc. of the IEEE SBRN, 2000. 84～89

8 Bezdek J C,Pal N R. Some new indexes of cluster validity. IEEE Transactions on Systems, Man and Cybernetics, Part B,1998,28 (3):301～315