# 基于混沌和免疫应答的增量聚类新算法

李向华 [1,2] 王钲旋 [1] 吕天阳 [3] 车翔玖 [1]

**摘 要** 受免疫应答原理的启发, 提出了一种适用于增量数据聚类的人工免疫系统框架, 以及在此框架上的结合混沌的自组织增量聚类新算法, 称为免疫应答算法 (Immune response algorithm, IRA). 新算法利用 Logistic 混沌序列生成初始抗体种群, 利用其多样性识别新增的不属于任何已知簇的数据, 该过程模拟了初次免疫应答. 同时, 初次免疫应答形成的记忆抗体可用于二次免疫应答, 即识别新增的属于已知簇的数据. 为了减少数据冗余, 算法用中心点和代表点表示已知簇并动态更新其识别区域, 这样算法不但能动态、自组织地形成聚类, 而且实现了数据特征的提取. 模拟实验充分显示出该算法无论在聚类质量上还是数据特征的提取上, 都具有一定优势, 且具有参数数量少、速度快、对数据输入次序不敏感的优点, 在实际问题中有一定应用价值.

**关键词** 人工免疫系统, 增量聚类, 免疫应答, 混沌, 特征提取
**DOI** 10.3724/SP.J.1004.2010.00208

## A Novel Incremental Clustering Algorithm Based on Chaos and Immune Response

LI Xiang-Hua[1,2] WANG Zheng-Xuan[1] LV Tian-Yang[3] CHE Xiang-Jiu[1]

**Abstract** Inspired by the immune response principle, an artificial immune system framework for incremental data clustering is proposed. Meanwhile, a novel self organizing incremental clustering algorithm called IRA (Immune response algorithm) is also proposed based on the framework. IRA uses Logistic chaotic sequence to produce the initial antibody population. The diversity of the chaotic sequence is used for recognizing the incremental data which do not belong to any existing clusters. This process simulates the primary immune response. At the same time, the memory antibodies produced by the primary immune response are used for the secondary immune response, that is, they can recognize the incremental data which belong to the existing cluster. In order to reduce the data redundancy, the algorithm uses the center and representative points to represent the existing clusters. The recognizing scopes of them are updated dynamically. Therefore, the algorithm not only can form clusters dynamically and self organization, but also can achieve data feature selection. The experimental results show that the algorithm has advantages on both the clustering quality and the data feature selection. Furthermore, it has some other merits, such as few parameters, fast speed, insensitivity to input and so on. So the algorithm has some value for practical problems.

**Key words** Artificial immune system, incremental clustering, immune response, chaos, feature selection

人工免疫系统 (Artificial immune system, AIS) 作为软计算的又一范例[1], 模拟了生物免疫系统的识别、学习和记忆能力, 以及分布式、自组织和多样性等特性, 因而其应用广泛, 研究成果涉及控制[2]、数据处理[3−10]、优化学习[11−12] 和故障诊断[13] 等许多领域. 本文的研究重点是基于 AIS 的增量数据聚类.

Timmis 等提出的资源受限人工免疫系统 (Resource limited AIS, RLAIS)[4] 和 de Castro 等构造的进化人工免疫网络 (aiNet)[5] 是两个比较著名的基于 AIS 的聚类算法, 二者都是抽取了免疫网络隐喻实现了对静态数值数据的聚类, 显示出 AIS 在数据分析方面的潜能. 在基于 AIS 的增量聚类方面, 也取得了一定的成果. Neal 提出的自稳定人工免疫系统 (Self-stabilizing AIS, SSAIS)[6] 能够处理数据流且保持免疫网络稳定, 文献 [7] 中提出的亚稳定记忆人工免疫系统 (Meta-stable memory in an artificial immune network, MSMAIS) 是对 SSAIS 的改进和简化, 提高了速度. Nasraoui 等[8] 为了提高 AIS 模型的可扩展性, 融合了 K-means 算法. Hart 等[9] 提出了用于处理二进制串的自组织系统 (Self-organizing sparse distributed memory, SOSDM), 该方法可扩展且以增量的方式工作. 岳训[10] 在 aiNet 的基础上, 提出的基于免疫的增量特征提取算法 (An immune-inspired incremental feature selection algorithm, ISFaiNet) 实现了对数据流的增量特征提取, 并应用在反垃圾邮件中. 在以上

收稿日期 2009-02-05 录用日期 2009-06-11
Manuscript received February 5, 2009; accepted June 11, 2009
国家自然科学基金 (60773096, 60773098), 高等学校博士学科点专项科研基金 (20060183041) 资助
Supported by National Natural Science Foundation of China (60773096, 60773098) and Foundation for the Doctoral Program of the Chinese Ministry of Education (20060183041)

1. 吉林大学计算机科学与技术学院 长春 130012 2. 吉林大学珠海学院计算机科学与技术系 珠海 519041 3. 哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001
1. College of Computer Science and Technology, Jilin University, Changchun 130012 2. Department of Computer Science and Technology, Zhuhai College, Jilin University, Zhuhai 519041 3. College of Computer Science and Technology, Harbin Engineering University, Harbin 150001

各种增量聚类算法中, 有的对新模式识别速度慢[6]; 有的存储了大量的冗余数据, 计算复杂[6−9]; 有的算法参数较多, 且对参数设置敏感[10]. 而且这些算法都集中在对生物功能的模拟处理上, 没有关注免疫系统中存在的混沌现象[14].

鉴于此, 本文提出了用于增量聚类的 AIS 框架和在此框架上的一个自组织的增量聚类算法. 该框架主要模拟免疫系统中的免疫应答过程而实现对增量数据的识别, 通过抗原数据和抗体数据的交互过程, 动态、自组织地形成聚类. 为了更好地描述抗体多样性, 将混沌引入算法中, 即在传统 Logistic 混沌序列特性的基础上, 模拟了整个的免疫应答过程. 该算法不但能实现高质量的增量数据聚类, 同时为了提高算法的可扩展性、减少数据冗余, 用中心点和代表点表示已知簇, 从而实现了数据特征的提取. 此外, 算法中参数少, 操作简单. 实验结果表明, 该算法对数据输入次序不敏感, 适应新模式速度快, 聚类质量较高, 特征提取相对较好, 且对高维、复杂的数据聚类问题也十分有效.

## 1　适用于增量聚类的 AIS 框架及其表示

AIS 的框架结构至少包含三个基本要素[1]: 个体表示、亲和度度量和免疫算法, 如图 1 左部所示. 根据免疫应答原理, 我们提出了适用于增量聚类的 AIS 框架, 如图 1 右部所示. 在文中, 与抗原、抗体和记忆抗体相对应的数据构成了 AIS 的基本元素, 它们均为实数值, 这里采用欧氏距离度量数据之间的亲和度, 控制整个 AIS 动态过程的核心部分就是能实现增量数据聚类的动态自组织的免疫应答算法 (Immune response algorithm, IRA).

为了进一步说明增量数据聚类过程与免疫应答过程的对应关系, 以及更好地描述算法, 现给出如下形式化定义.

**定义 1.** 抗原数据 ($\boldsymbol{Ag}$): 在 IRA 中, 将待识别的增量数据看作是免疫系统中的外来物质—抗原, 称之为抗原数据, 表示为 $\boldsymbol{Ag}: \boldsymbol{Ag} = \{\boldsymbol{Ag}_1, \boldsymbol{Ag}_2, \cdots, \boldsymbol{Ag}_n\}$, 其中, $\boldsymbol{Ag}_i = \{Ag_{i1}, Ag_{i2}, \cdots, Ag_{iL}\} \in \mathbf{R}^L$, $1 \le i \le n$.

**定义 2.** 抗体数据 ($\boldsymbol{Ab}$): 在 IRA 中, 将分布在数据空间中的数据看作是免疫系统中的自由抗体, 称之为抗体数据, 表示为 $\boldsymbol{Ab}: \boldsymbol{Ab} = \{\boldsymbol{Ab}_1, \boldsymbol{Ab}_2, \cdots, \boldsymbol{Ab}_{Number_{Ab}}\}$, 其中, $\boldsymbol{Ab}_j = \{Ab_{j1}, Ab_{j2}, \cdots, Ab_{jL}\} \in \mathbf{R}^L$, $1 \le j \le Number_{Ab}$ ($Number_{Ab}$ 为 $\boldsymbol{Ab}$ 种群的数量).

**定义 3.** 记忆抗体数据 ($\boldsymbol{MAb}$): 在 IRA 中, 将用于表示已知簇的中心点和代表点看作是免疫系统中用于识别相似抗原的记忆抗体, 称之为记忆抗体数据, 表示为 $\boldsymbol{MAb}: \boldsymbol{MAb} = \{\boldsymbol{MAb}_1, \boldsymbol{MAb}_2, \cdots, \boldsymbol{MAb}_{Number_{ec}}\}$, 第 $k$ 个已知簇表示为 $\boldsymbol{MAb}_k = \{\boldsymbol{ct}_k, \boldsymbol{rp}_{k1}, \cdots, \boldsymbol{rp}_{kNum}\}$, 其中, $\boldsymbol{ct}_k$ 表示第 $k$ 个簇的中心点, $\boldsymbol{rp}_{k1}, \cdots, \boldsymbol{rp}_{kNum}$ 表示第 $k$ 个簇的 $Num$ 个代表点, 且 $\boldsymbol{ct}_k = \{ct_{k1}, ct_{k2}, \cdots, ct_{kL}\} \in \mathbf{R}^L$, $\boldsymbol{rp}_{ki} = \{rp_{ki1}, rp_{ki2}, \cdots, rp_{kiL}\} \in \mathbf{R}^L$, $1 \le i \le Num$, $1 \le k \le Number_{ec}$ ($Number_{ec}$ 表示 $\boldsymbol{MAb}$ 种群的数量).

**定义 4.** 中心点的识别半径 ($R\_ct$): 将已知簇的中心点到其所有代表点的平均欧氏距离 (用 $Dist(\cdot,\cdot)$ 表示) 称之为该中心点的识别半径, 表示为 $R\_ct$.

假设第 $k$ 个已知簇有 $Num$ 个代表点, 则它的识别半径 $R\_ct_k$ 为

$$R\_ct_k = \frac{\sum_{i=1}^{Num} Dist(\boldsymbol{ct}_k, \boldsymbol{rp}_{ki})}{Num} \tag{1}$$

**定义 5.** 代表点的识别半径 ($R\_rp$): 将已知簇中的代表点按照与其中心点距离递增的顺序排列, 代表点的识别半径与其中心点的距离有关, 即距离越小, 半径越大, 反之亦然. 代表点的识别半径表示为 $R\_rp$.

假设第 $k$ 个已知簇有 $Num$ 个代表点, 其按照与中心点距离由近及远的顺序排列, 那么第 $i$ ($i = 1, 2, \cdots, Num$) 个代表点的识别半径 $R\_rp_{ki}$ 为

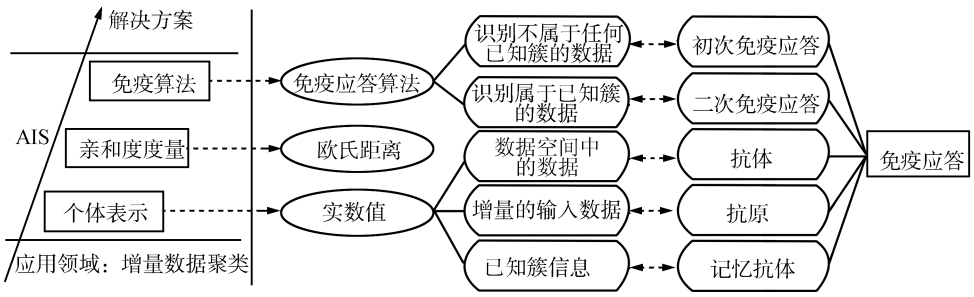



图 1　用于增量数据聚类的 AIS 框架

Fig. 1　AIS framework for incremental data clustering

$$R\_rp_{ki} = \frac{Dist(\boldsymbol{ct}_k, \boldsymbol{rp}_{ki})}{i} \tag{2}$$

**定义 6.** 相似抗体数据 ($\boldsymbol{sAb}$): 在 IRA 中, 将位于记忆抗体数据中心点识别范围之内的抗体数据称之为相似抗体数据, 表示为 $\boldsymbol{sAb}$ : $\boldsymbol{sAb} = \{\boldsymbol{sAb}_1, \boldsymbol{sAb}_2, \cdots, \boldsymbol{sAb}_p\}$, 其中 $\boldsymbol{sAb} \subset \boldsymbol{Ab}$ 且 $Dist(\boldsymbol{sAb}_j, \boldsymbol{ct}) < R\_ct$, $1 \le j \le p$.

**定义 7.** 亲和度 ($Affinity_{Ag-Ab}$): 在 IRA 中, 将 $\boldsymbol{Ag}$ 与 $\boldsymbol{Ab}$ 之间的匹配程度看作是免疫系统中抗原和抗体的结合程度, 称之为亲和度, 表示为 $Affinity_{Ag-Ab}$. 它与 $\boldsymbol{Ag}$ 和 $\boldsymbol{Ab}$ 之间的欧氏距离有关, 它们之间的距离越小, 则亲和度越大, 反之亦然. 计算公式为

$$Affinity_{Ag-Ab} = \frac{1}{1 + Dist(\boldsymbol{Ag}, \boldsymbol{Ab})} \tag{3}$$

## 2 基于混沌和免疫应答的增量聚类新算法

### 2.1 Logistic 混沌序列

混沌是自然界普遍存在的非线性现象, 它是包含于无序中的有序模式, 具有随机性、遍历性和规律性等性质. 混沌理论已成为一种新颖有潜力的优化工具, 被引入到软计算的许多范例中.

在免疫系统中, 正是由于抗体多样性使得免疫系统能够识别和清除不同的抗原. 在 IRA 中, $\boldsymbol{Ab}$ 种群的多样性对整个算法的性能起着至关重要的作用. 为了模拟抗体的多样性, 我们借鉴了混沌的遍历性和随机性. 考虑到 Logistic 映射较其他混沌迭代映射使用方便、计算量小, 故本文采用式 (4) 来产生混沌序列

$$x_{n+1} = \mu x_n (1 - x_n) \tag{4}$$

式中 $\mu$ 为控制参数, $\mu$ 值确定后, 由任意初值 $x_0 \in [0, 1]$, 可迭代出一个确定的时间序列, 如图 2 所示. 随着 $\mu$ 值的增加, 式 (4) 呈现出不同的性质, 当 $3 \le \mu \le 4$ 时, 系统由倍周期通向混沌. 可以证明, 当 $\mu = 4$ 时, 系统完全处于混沌状态, 系统遍历了整个 [0,1] 区间.

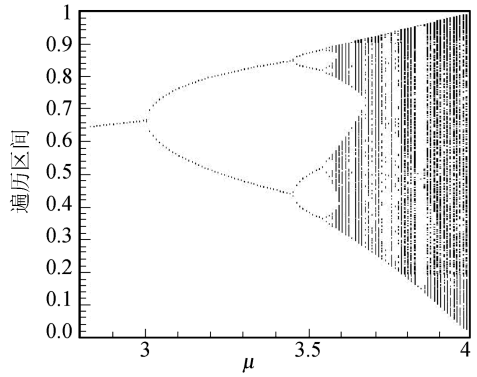



图 2 Logistic 分岔图

Fig. 2 Bifurcation diagram of Logistic

### 2.2 免疫应答算法 (IRA)

根据第 1 节提出的 AIS 框架可知, IRA 直接利用了初次免疫应答和二次免疫应答的机制, 即模拟初次免疫应答的过程实现对不属于任何已知簇的增量数据的识别, 同时获得记忆抗体数据, 用于模拟二次免疫应答, 即识别那些属于已有簇的增量数据.

文中用克隆选择原理[15] 来描述免疫应答过程, 具体如下: 当前抗原数据 $\boldsymbol{Ag}_i$ 如果不能被任何记忆抗体数据识别, 则进行初次免疫应答, 即进行亲和度计算、克隆选择、混沌变异和免疫记忆.

亲和度计算: 按照式 (3) 计算 $\boldsymbol{Ag}_i$ 与 $\boldsymbol{Ab}$ 种群的亲和度.

克隆选择: 具有最高亲和度的抗体数据按式 (5) 进行克隆, $N_{\text{clones}}$ 为克隆数量, $\alpha$ 是克隆率, round$(\cdot)$ 表示取整:

$$N_{\text{clones}} = \text{round}(\alpha \times Affinity_{Ag-Ab} \times Number_{Ab}) \tag{5}$$

混沌变异: 克隆的抗体数据按照变异率 $\beta$, 依据式 (4) 进行混沌变异.

免疫记忆: 由于 $\boldsymbol{Ag}_i$ 是一个全新数据, 它形成一个新簇 $\boldsymbol{MAb}_i = \{\boldsymbol{ct}_i, \boldsymbol{rp}_{i1}, \cdots, \boldsymbol{rp}_{iNum}\}$, 其中 $\boldsymbol{ct}_i = \boldsymbol{Ag}_i$, $\boldsymbol{rp}_i = \{\boldsymbol{Ab}_m | \min_{m=1,2,\cdots,Num}(Dist(\boldsymbol{ct}_i, \boldsymbol{Ab}_m)), \boldsymbol{Ab}_m \in \boldsymbol{Ab} \wedge m = 1, 2, \cdots, Number_{Ab}\}$, 即新增数据为新簇的中心点, 与之距离最近的 $Num$ 个抗体数据成为代表点, 根据式 (1) 和 (2) 计算新簇的识别半径.

于是, IRA 可描述如下:

**输入.** $\boldsymbol{Ag}_i$ ($i = 1, 2, \cdots, n$, $\boldsymbol{Ag}_i \in \boldsymbol{Ag}$), 它们以一次一个的增量方式进行提交.

**输出.** 增量数据的聚类结果及它们的数据特征.

**步骤 1.** 混沌初始化抗体数据种群, $\boldsymbol{Ab} = \{\boldsymbol{Ab}_1, \boldsymbol{Ab}_2, \cdots, \boldsymbol{Ab}_{Initial_{Ab}}\}$.

**步骤 2.** 对每个增量的 $\boldsymbol{Ag}_i$:

**步骤 2.1.** $r_{ik} = Dist(\boldsymbol{Ag}_i, \boldsymbol{MAb}_k)$ ($k$ = 1, 2, $\cdots$, $Number_{ec}$);

**步骤 2.2.** 如果 $r_{ik} > R\_ct_k$ 且 $r_{ik} > R\_rp_k$, 进行亲和度计算、克隆选择、混沌变异和免疫记忆;

**步骤 2.3.** 如果 $r_{ik} \le R\_ct_k$ 或 $r_{ik} \le R\_rp_k$, 记录识别 $\boldsymbol{Ag}_i$ 的簇标号、更新中心点、代表点和识别半径;

**步骤 2.4.** 抑制: 删除 $\boldsymbol{sAb}$;

**步骤 2.5.** 多样性: 如果 $Number_{Ab} < Initial_{Ab}$, 产生 ($Initial_{Ab} - Number_{Ab}$) 个新抗体数据并加入到 $\boldsymbol{Ab}$ 中;

**步骤 2.6.** 如果未达到终止条件, 转步骤 2, 否

则结束.

### 2.3 IRA 的计算复杂度

对于计算复杂度, 在最坏的情况下, 即每个新增的抗原数据都自成一簇, 即 $n = Number_{ec}$. 计算量最大的步骤是步骤 2.1 的亲和度计算和步骤 2.2. 中的免疫记忆部分. 因此, IRA 的计算复杂度由两部分组成: $\mathrm{O}((Num+1) \times Number_{ec})$ (步骤 2.1) 和 $(Num + Number_{Ab})$ (步骤 2.2). 于是, 总的计算复杂度为 $\mathrm{O}(n \times ((Num+1) \times Number_{ec} + (Number_{Ab} + Num)))$. 由于 $Number_{Ab} >> Num$ 和 $n >> Num + 1$, 所以 IRA 的计算复杂度是 $\mathrm{O}(n \times (n + Number_{Ab})) = \mathrm{O}(n^2 + n \times Number_{Ab})$.

## 3 实验结果与分析

IRA 既能实现数据聚类, 又能实现数据特征的提取. 为了评估其性能, 将其与最新的基于 aiNet 的特征提取方法 ISFaiNet[10] 比较, 测试其对数据特征提取的能力; 与 MSMAIS[7] 比较, 评估 IRA 的聚类质量. IRA、ISFaiNet 和 MSMAIS 三种方法的总体比较见表 1, 其中 $Number_{\mathrm{ARB}}$ 为 MSMAIS 中 ARB 的总数量. 为直观显示不同方法的性能和效果, 首先采用人工二维数据集测试三种算法, 然后选用来自 UCI 数据库[16] 经典的 Iris, Wine 和 Wisconsin breast cancer 数据集测试不同方法.

### 3.1 评估方法和参数选择

我们采用与文献 [10] 同样的方法来评价 IRA 的数据特征提取效果: 即算法的运行时间、同质度 (Homogeneity)$H$ 与分离度 (Separation)$S$. 同质度 $H$ 计算属于同一特征类中的各点与中心点的平均距离, 反映同一类数据之间的紧密程度; 分离度 $S$ 计算各个特征类的中心点之间的平均加权距离, 反映不同特征类之间的分离程度. $H$ 越小, $S$ 越大, 数据特征提取的质量就越好, 计算公式如下:

$$H = \frac{1}{n}\sum_{i=1}^{n} Dist(d_i, center_k) \tag{6}$$

$$S = \frac{\sum\limits_{i\neq j} |C_i| \times |C_j| \times Dist(center_i, center_j)}{\sum\limits_{i\neq j} |C_i| \times |C_j|} \tag{7}$$

对于 IRA 的聚类质量, 也从两方面进行评价: 算法的运行时间和平均分类正确率.

在以下的实验中, IRA 和 ISFaiNet 均取 50 次实验的平均值, 而 MSMAIS 由于没有变异机制, 故只取一次实验结果即可, IRA 的参数取值见表 2.

### 3.2 算法性能评价与比较

图 3 为三种不同的算法对二维数据集的一次学习结果, 其中图 3 (a) 为二维数据集, 图 3 (b) ∼ 3(d) 分别为 IRA, ISFaiNet 和 MSMAIS 对该数据集的学习结果. IRA 和 ISFaiNet 采用不同机制对数据特征进行了提取, IRA 用中心点和其代表点来表示每一个簇, ISFaiNet 利用 5 个时间窗口 (Time window) 提取了数据特征, 而 MSMAIS 利用人工免疫网络生成了数据集的拓扑结构. 此外, IRA 和 MSMAIS 还能获得每个增量数据的所属类别.

将 IRA 和 ISFaiNet 分别应用在二维数据集、Iris、Wine 和 Wisconsin breast cancer 获得的结果见表 3. 从实验结果可以看出, IRA 利用了较少的运行时间取得了更好的特征提取效果, 这主要是因为 ISFaiNet 需要反复迭代, 而 IRA 只需一次遍历数据集即可得到学习结果. 另外, 对于无序数据集 (如 Wisconsin breast cancer), 无论 ISFaiNet 的时间窗口如何选取, 都很难得到较好的效果.

表 4 显示的是 IRA 和 MSMAIS 分别对以上四个数据集的学习结果. 尽管 MSMAIS 对某些数据集有运行时间上的优势, 但其分类正确率却低于 IRA 的结果. 而且随着数据集的增大, MSMAIS 中 ARB

表 1　IRA 与 ISFaiNet 和 MSMAIS 的比较

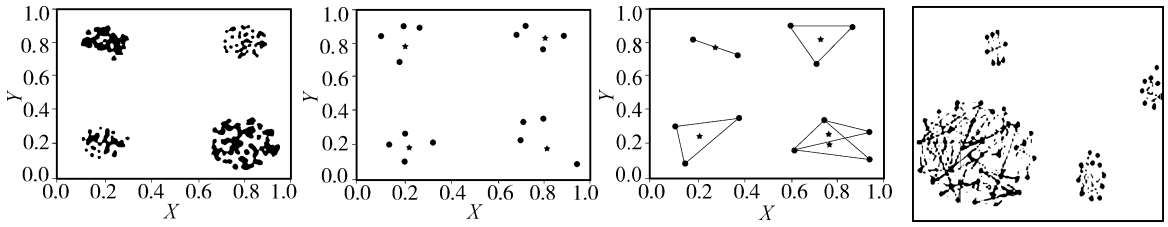
Table 1　Comparison IRA with ISFaiNet and MSMAIS

| 算法 | 计算复杂性 | 参数个数 | 理论基础 | 学习结果 |
|---|---|---|---|---|
| IRA | $\mathrm{O}(n^2 + n \times Number_{Ab})$ | 4 | 克隆选择原理描述免疫应答 | 实现了数据的聚类和特征提取 |
| ISFaiNet | $\mathrm{O}(n^3)$ | 10 | 免疫网络理论 (基于 aiNet) | 实现了数据的特征提取 |
| MSMAIS | $\mathrm{O}(n \times Number^2_{\mathrm{ARB}} - n \times Number_{\mathrm{ARB}})$ | 6 | 免疫网络理论 (基于 RLAIS) | 实现了数据的聚类 |

表 2　IRA 的参数取值

Table 2　Parameter values for IRA

| 参数 | 二维数据 | Iris | Wine | Wisconsin breast cancer | 意义 |
|---|---|---|---|---|---|
| $Initial_{Ab}$ | 50 | 500 | 500 | 50 | 初始抗体种群数量 |
| $Num$ | 4 | 2 | 2 | 3 | 代表点数目 |
| $\alpha$ | 0.1 | 0.1 | 0.1 | 0.1 | 克隆率 |
| $\beta$ | 0.1 | 0.1 | 0.1 | 0.1 | 变异率 |




(a) 二维数据集 (b) IRA (c) ISFaiNet (d) MSMAIS

(a) 2D data set (b) IRA (c) ISFaiNet (d) MSMAIS

图 3 三种算法对二维数据的学习结果

Fig. 3 Learning results for two-dimensional data set from three different algorithms

表 3 IRA 与 ISFaiNet 对二维数据、Iris、Wine 和 Wisconsin breast cancer 的测试结果

Table 3 Testing results from both IRA and ISFaiNet on 2D, Iris, Wine, and Wisconsin breast cancer

| 数据集 | 2D | | Iris | | Wine | | Wisconsin breast cancer | |
|---|---|---|---|---|---|---|---|---|
| 算法 | IRA | ISFaiNet (window = 5) | IRA | ISFaiNet (window = 5) | IRA | ISFaiNet (window = 5) | IRA | ISFaiNet (window = 10) |
| 平均运行时间 (s) | 0.027 | 0.410 | 0.237 | 0.464 | 0.202 | 3.959 | 0.228 | 41.852 |
| 同质度 | 0.088 | 0.150 | 0.116 | 0.304 | 0.549 | 0.473 | 0.639 | 0.845 |
| 分离度 | 0.608 | 0.560 | 0.779 | 0.605 | 1.182 | 0.622 | 1.165 | 0.223 |

表 4 IRA 与 MSMAIS 对二维数据、Iris、Wine 和 Wisconsin breast cancer 的测试结果

Table 4 Testing results from both IRA and MSMAIS on 2D, Iris, Wine, and Wisconsin breast cancer

| 数据集 | 2D | | Iris | | Wine | | Wisconsin breast cancer | |
|---|---|---|---|---|---|---|---|---|
| 算法 | IRA | MSMAIS (NAT = 0.1) | IRA | MSMAIS (NAT = 0.7) | IRA | MSMAIS (NAT = 0.7) | IRA | MSMAIS (NAT = 0.7) |
| 平均运行时间 (s) | 0.027 | 0.031 | 0.237 | 0.031 | 0.202 | 0.062 | 0.228 | 0.406 |
| 平均正确率 (%) | 100 | 91.3 | 95.9 | 95.3 | 93.6 | 90.6 | 97.4 | 96.6 |

表 5 IRA 对二维数据、Iris、Wine 和 Wisconsin breast cancer 不同输入次序的测试结果

Table 5 Testing results from IRA on 2D, Iris, Wine, and Wisconsin breast cancer when inputting data in different orders

| 数据集 | 2D | | Iris | | Wine | | Wisconsin breast cancer | |
|---|---|---|---|---|---|---|---|---|
| 数据输入次序 | 顺序 | 随机 | 顺序 | 随机 | 顺序 | 随机 | 顺序 | 随机 |
| 平均运行时间 (s) | 0.027 | 0.031 | 0.237 | 0.367 | 0.202 | 0.291 | 0.228 | 0.246 |
| 平均正确率 (%) | 100 | 100 | 95.9 | 94.9 | 93.6 | 93.6 | 97.4 | 97.1 |
| 同质度 | 0.088 | 0.094 | 0.116 | 0.114 | 0.549 | 0.831 | 0.639 | 0.648 |
| 分离度 | 0.608 | 0.613 | 0.779 | 0.796 | 1.182 | 1.687 | 1.165 | 1.162 |

的数量也不断增多, 最终使得 MSMAIS 不可扩展.

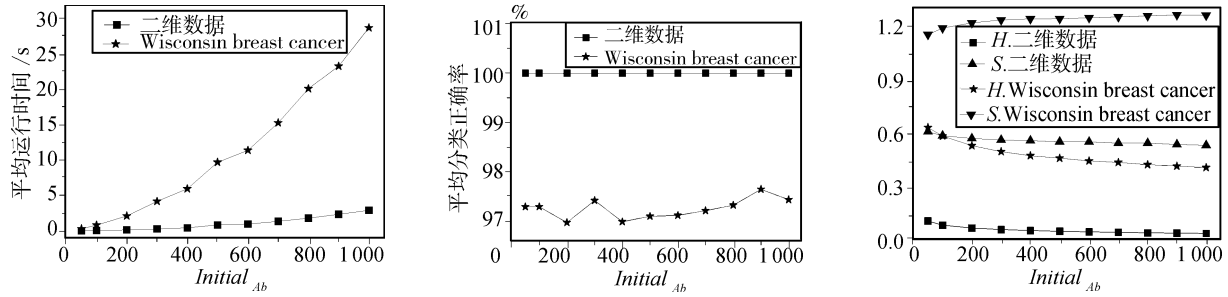
### 3.3 数据输入次序对 IRA 的影响

一些聚类算法对数据输入的次序很敏感, 同一个数据集, 当以不同次序提交给同一个算法时, 可能生成差别很大的聚类结果. 因此, 考察 IRA 对数据输入次序的敏感性十分必要. 在第 3.2 节中, 数据的输入是按照已知的所属簇的次序依次输入的, 称之为顺序输入. 本节中我们对以上四组数据集中的数据按照任意次序输入, 称为随机输入, 取 50 次 (50 次数据输入次序均随机产生) 实验的平均值列于表 5, 并与顺序输入的结果进行了对比. 实验结果表明, IRA 对数据输入次序不敏感.

从表 5 可以看出, 无论数据以何种次序进行增量输入, IRA 在聚类质量和数据特征提取上均没有明显差异. 这是因为在数据空间中存在着由混沌序列产生的多样的抗体数据, 无论抗原数据以怎样的次序输入, 都能被抗体数据或者记忆抗体数据所识别, 而这个识别过程与抗原数据的输入次序无关. 当顺序输入时, 执行了更多次的二次免疫应答, 当随机次序输入时, 执行初次免疫应答的次数增加.

### 3.4 参数对 IRA 的影响

IRA 中共有四个参数: $Initial_{Ab}, Num, \alpha$ 和 $\beta$,

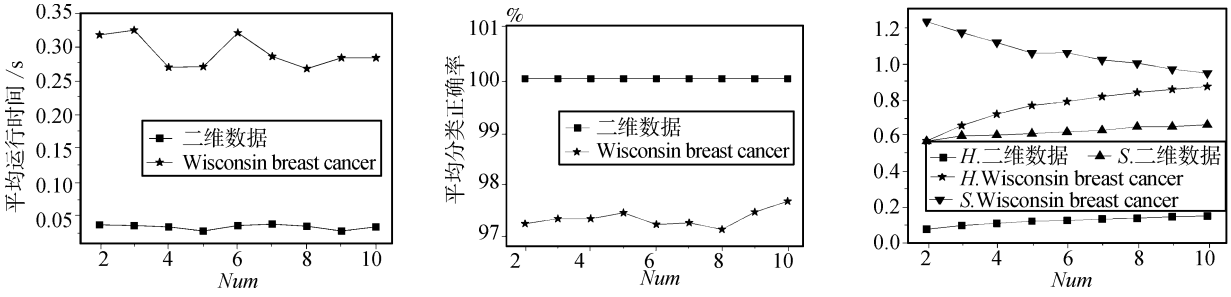



(a) 平均运行时间　(b) 平均分类正确率　(c) $H$ 和 $S$

(a) Average running time　(b) Average correct rate　(c) $H$ and $S$

图 4　参数 $Initial_{Ab}$ 对 IRA 的影响

Fig. 4　Effects of parameter $Initial_{Ab}$ on IRA




(a) 平均运行时间　(b) 平均分类正确率　(c) $H$ 和 $S$

(a) Average running time　(b) Average correct rate　(c) $H$ and $S$

图 5　参数 $Num$ 对 IRA 的影响

Fig. 5　Effects of parameter $Num$ on IRA

这里主要讨论它们对同质度 $H$、分离度 $S$ 以及对算法的平均运行时间和平均分类正确率的影响. 我们以最简单的二维数据集和最复杂的 Wisconsin breast cancer 数据集为例来讨论参数对 IRA 的影响.

通过分析 IRA 可知, 每处理完一个抗原数据, ***Ab*** 种群数量总是围绕在 $Initial_{Ab}$ 左右 (步骤 2.5), 参数 $Initial_{Ab}$ 对 ***Ab*** 种群的多样性起到了关键作用, $Initial_{Ab}$ 取值越大, 多样性越丰富, 分类就越细致, 相应的 $H$ 就会越小, 但是计算时间也会相应增加, 反之亦然. 而 $S$ 的大小取决于最终的簇数目以及各中心点之间的距离, 故其趋势不定. 参数 $\alpha$ 和 $\beta$ 的作用与 $Initial_{Ab}$ 相同, 故不重复讨论它们对 IRA 的影响. 图 4 分别是不同取值的 $Initial_{Ab}$ 对算法的平均运行时间和平均分类正确率以及对 $H$ 和 $S$ 的影响.

$Num$ 是 IRA 的另一个重要参数, 如果 $Num$ 取值变大, 每个簇的中心点的识别范围就会相应扩大, 代表点与中心点的距离也会变大, 即 $H$ 有增加的趋势, 反之亦然, 而 $S$ 的变化趋势依赖不同的数据集. 实验证明, $Num$ 的变化对 IRA 运行时间和正确率都影响不大, 如图 5 所示.

## 4　结论

本文提出了一种适用于增量数据聚类的 AIS 框架, 并在该框架基础上, 提出了动态的、自组织的增量聚类新算法 IRA. 该算法在总体上模拟了免疫应答过程, 实现对增量数据的识别. 同时考虑到了生物系统的非线性混沌本质, 并借助混沌的特性来提高算法性能. IRA 的参数少且限制少, 并利用中心点和代表点表示已知簇, 使得 IRA 具有聚类功能的同时, 也获得了待分析数据的特征, 在一定程度上减少了数据冗余. 实验结果表明, IRA 无论在聚类质量上还是数据特征提取上都取得了较好的效果. 但是对线性不可分的数据集, IRA 的效果还有待进一步提高.

## References

1 de Castro L N, Timmis J I. Artificial immune systems as a novel soft computing paradigm. *Soft Computing*, 2003, **7**(8): 526−544

2 Li Tao. An immune based model for network monitoring. *Chinese Journal of Computers*, 2006, **29**(9): 1513−1520 (李涛. 基于免疫的网络监控模型. 计算机学报, 2006, **29**(9): 1513−1520)

3 Gong Mao-Guo, Jiao Li-Cheng, Ma Wen-Ping, Zhang Xiang-Rong. Unsupervised classification and recognition using an artificial immune system based on manifold distance. *Acta Automatica Sinica*, 2008, **34**(3): 367−375
(公茂果, 焦李成, 马文萍, 张向荣. 基于流行距离的人工免疫无监督分类与识别算法. 自动化学报, 2008, **34**(3): 367−375)

4 Timmis J, Neal M. A resource limited artificial immune system for data analysis. *Knowledge-Based Systems*, 2001, **14**(3-4): 121−130

5 de Castro L N, Von Zuben F J. An evolutionary immune network for data clustering. In: Proceedings of the 6th Brazilian Symposium on Neural Networks. Rio de Janeiro, Brazil: IEEE, 2000. 84−89

6 Neal M. An artificial immune system for continuous analysis of time-varying data. In: Proceedings of the 1st International Conference on Artificial Immune Systems. Berlin, Germany: Springer-Verlag, 2002. 76−85

7 Neal M. Meta-stable memory in an artificial immune network. In: Proceedings of the 2nd International Conference on Artificial Immune Systems. Berlin, Germany: Springer-Verlag, 2003. 168−180

8 Nasraoui O, Gonzalez F, Cardona C, Rojas C, Dasgupta D. A scalable artificial immune system model for dynamic unsupervised learning. In: Proceedings of Genetic and Evolutionary Computation Conference. Chicago, USA: Springer-Verlag, 2003. 219−230

9 Hart E, Ross P. Exploiting the analogy between the immune system and sparse distributed memories. *Genetic Programming and Evolvable Machines*, 2003, **4**(4): 333−358

10 Yue X, Mo H W, Chi Z X. Immune-inspired incremental feature selection technology to data streams. *Applied Soft Computing*, 2008, **8**(2): 1041−1049

11 Timmis J, Edmonds C, Kelsey J. Assessing the performance of two immune inspired algorithms and a hybird genetic algorithm for optimization. In: Proceedings of Genetic and Evolutionary Computation Conference. Berlin, Germany: Springer-Verlag, 2004. 308−317

12 Du Hai-Feng, Gong Mao-Guo, Liu Ruo-Chen, Jiao Li-Cheng. Adaptive chaos clonal evolutionary programming algorithm. *Science in China, Series E*, 2005, **35**(8): 817−829
(杜海峰, 公茂果, 刘若辰, 焦李成. 自适应混沌克隆进化规划算法. 中国科学 E 辑, 2005, **35**(8): 817−829)

13 Zhou Peng, Qin Shu-Ren. Rotating machinery fault diagnosis based on slice spectrum-AIS. *Chinese Journal of Scientific Instrument*, 2008, **29**(6): 1198−1202
(周鹏, 秦树人. 基于切片谱免疫系统的旋转机械故障诊断. 仪器仪表学报, 2008, **29**(6): 1198−1202)

14 Qi An-Shen, Du Chan-Ying. *Nonlinear Models in Immunity*. Shanghai: Shanghai Scientific and Technological Education Publishing House, 1999
(漆安慎, 杜蝉英. 免疫的非线性模型. 上海: 上海科技教育出版社, 1999)

15 de Castro L N, Von Zuben F J. The clonal selection algorithm with engineering applications. In: Proceedings of Workshop on Artificial Immune Systems and Their Applications. Las Vegas, USA: Springer-Verlag, 2000. 36−37

16 UCI. Machine learning repository [Online], available: http://archive.ics.uci.edu/ml/, June 10, 2009

**李向华** 吉林大学博士研究生, 吉林大学珠海学院讲师. 主要研究方向为基于智能算法的聚类和分形.
E-mail: li_xianghua@163.com
(**LI Xiang-Hua** Ph. D. candidate at Jilin University, lecturer at Zhuhai College, Jilin University. Her research interest covers clustering based on intelligence algorithms and fractal.)

**王钲旋** 吉林大学教授. 主要研究方向为计算机图形学、分形与混沌和数据挖掘. E-mail: wangzx@jlu.edu.cn
(**WANG Zheng-Xuan** Professor at Jilin University. His research interest covers computer graphics, fractal and chaos, and data mining.)

**吕天阳** 哈尔滨工程大学讲师. 主要研究方向为数据挖掘和信息检索.
E-mail: raynor1979@163.com
(**LV Tian-Yang** Lecturer at Harbin Engineering University. His research interest covers data mining and information retrieval.)

**车翔玖** 吉林大学教授. 主要研究方向为计算机图形学和小波理论与应用. 本文通信作者. E-mail: chexj@jlu.edu.cn
(**CHE Xiang-Jiu** Professor at Jilin University. His research interest covers computer graphics, wavelet theory and application. Corresponding author of this paper.)