基于人工免疫原理的数据流聚类算法研究

胡 伟[1,2] 徐福缘[1] 马庆国[3]
(上海理工大学管理学院 上海 200093)[1] (绵阳师范学院数学与计算机科学学院 绵阳 621000)[2]
(浙江大学管理学院 杭州 310058)[3]

摘 要 针对传统数据流聚类算法自适应性不强、对问题的依赖性过高以及聚类质量不够理想、聚类效率低下等缺陷,提出一种基于人工免疫原理的数据流聚类 IMStream 算法。该算法通过引入衰减函数和时刻权重来反映过去的数据与当前流入的数据在整个数据流中的地位,通过计算抗体期望克隆率 $E(x_i)$ 来限制抗体克隆的数目以及保持抗体的多样性,通过采取网络中的淘汰策略使最终的网络结构更符合原始数据流的内在特性。在真实数据集和人工数据集上的实验表明,IMStream 算法比传统的数据流聚类算法具有更好的性能。

关键词 数据流,聚类,免疫算法,数据挖掘
中图法分类号 TP311 **文献标识码** A

Research of Data Stream Clustering Algorithms Based on Artificial Immune Principle

HU Wei[1,2] XU Fu-yuan[1] MA Qing-guo[3]
(School of Business, University of Shanghai for Science&Technology, Shanghai 200093, China)[1]
(School of Mathematics and Computer Science, Mianyang Normal University, Mianyang 621000, China)[2]
(School of Business, Zhejiang University, Hangzhou 310058, China)[3]

Abstract In traditional data stream clustering algorithm, adaptive problem is not strong, the dependence is too high and clustering quality is not ideal, clustering efficiency is low, this paper presented a data stream clustering algorithms based on artificial immune principle—IMStream. This algorithm introduces attenuation function and moment weight to reflect the past data weights with the current data in the data stream of position, through calculating antibody expect cloning rate $E(x_i)$ to limit the antibody cloning number and keep the diversity of antibodies, take out eliminated strategy to made the final network structure more to accord with the original data stream inherent characteristics. The experiments based on the real data set and artificial data set show that the algorithm IMStream has better clustering performance.

Keywords Data stream, Clustering, Immune algorithm, Data mining

近年来许多应用中待处理的数据以一种动态、流式的形式出现,如实时监控系统、互联网数据传输系统、金融市场证券信息发布等方面,均会在每时每刻产生大量的数据,对此类数据的分析和挖掘已日益成为一个热点问题。聚类分析作为一种基础的数据挖掘手段,在数据流环境下得到了学术界和工业界的广泛关注[1]。人工免疫系统是模仿自然免疫系统功能的一种智能方法,它受生物免疫系统启发,通过学习外界物质的自然防御机理,提供噪声忍耐、无教师学习、自组织、记忆等进化学习机理,结合了分类器、神经网络和机器推理等系统的一些优点,因此它提供了新颖的解决问题的方法和途径[2]。在所构造的人工免疫系统形式中,免疫算法是被应用得最为广泛的一个部分,其中大多数被用于实现全局优化问题。由于人工免疫原理的数据流聚类算法借鉴了生物免疫系统的功能和原理,因此可以利用免疫系统所具有的强大的学习、记忆、适应动态环境变化及自组织特性,以设计适用于数据流聚类问题求解的方法和构造出适合于数据流聚类的高效算法。

本文第1节介绍数据流的相关研究工作;第2节对相关的概念和定义进行描述;第3节详细描述了 IMStream 算法的各个步骤;第4节基于真实数据集合和人工数据集进行实验分析;最后进行了总结,并指出了下一步的研究方向。

1 相关研究

在传统的聚类算法中,最有代表性的是由 Aggarwal 等人[3]提出的 CluStream 算法。该算法首次把数据流看成是一个随时间变化的过程,提出了解决数据流聚类问题的框架,把聚类过程分为联机和脱机2个部分。之后出现了许多改进 CluStream 算法的方法,如常建龙等人[4]提出的基于滑动窗口的进化数据流聚类算法;邹凌君等人[5]提出了基于相关系数的多数据流聚类算法;朱蔚恒等人[6]提出了基于网格的聚类算法;也有人提出了基于密度的数据流聚类、基于网络和密度的数据流聚类以及基于模型的数据流聚类方法等等。

以上这些研究成果对数据流聚类的性能都有不同程度的

到稿日期:2011-03-30 返修日期:2011-07-28 本文受国家自然科学基金项目(70672110),上海市(第三期)重点学科项目(S30504)资助。
胡 伟(1979—),男,博士生,讲师,主要研究方向为系统工程、企业供需网及其管理、超网络等,E-mail:hw18721348510@126.com;徐福缘(1948—),男,教授,博士生导师,主要研究方向为系统工程、企业供需网及其管理、超网络等;马庆国(1945—),男,教授,博士生导师,主要研究方向为决策科学。

提高，但始终未能解决数据流聚类问题中的对初始值敏感、易陷入局部最优、效率低、聚类能力不强等缺陷。针对以上问题，本文提出一种基于人工免疫原理的数据流聚类 IMStream 算法，该算法通过引入衰减函数 $f(t)=e^{-\alpha t}$ 和时刻权重来反映过去的数据与当前流入的数据在整个数据流中的地位，通过计算抗体期望克隆率 $E(x_i)$ 来限制抗体克隆的数目及保持抗体的多样性，在网络中采取淘汰策略以减少参数设置数目。抗体经过克隆和变异后，从种群 B 中选出一定数量的亲和度高的抗体更新记忆细胞库中的抗体，使得记忆细胞库一直保持固定数目为 M 的最高亲和度抗体，从而有效地克服了传统算法自适应性不强，对问题的依赖性过高等缺陷。实验结果表明，在相同条件下与 IMStream 算法相比较，该算法内存占用量较少，同时具有良好的聚类质量及较快的数据处理能力。

2 基本概念

现代免疫学理论认为，生物的免疫系统对于外来侵犯的抗原可由 B 细胞分泌相应的抗体来抵御。抗体和抗原结合后，通过一系列的反应来破坏抗原。根据上述机理可将抗原对应于数据流中的数据点，抗体对应于各模式的特征值，它在免疫网络中表现为一个数据项。抗体通过相互作用构成一个免疫网络，应用目标就是使免疫网络经过学习后，能够反映原始数据集所包含的模式类[7]。

设原始数据流是一系列不断增长的 $d(d>1)$ 维数据记录 $X_1,X_2,\cdots,X_i,\cdots$，数据流中每一个记录 X_i 是 d 维，各数据点到达的时间分别为 $T_1,T_2,\cdots,T_i,\cdots$；对任意的 $1\leqslant i\leqslant d$（i 为离散型变量）和 $T_i<T_d$，有 $X_i=(X_{i1},X_{i2},\cdots,X_{id})$，其中 X_i 对应于免疫网络中的抗原，Y_j 对应于抗体。

定义 1（衰减函数 $f(t)$） 数据流中的数据是一种连续到达、持续增长、动态演化的数据，过去的数据与当前流入的数据在整个数据流中的地位是不一样的。对于历史数据，距离现在的时间越长，其重要性越低。现定义一个衰减函数来控制数据流中不同时刻观测到的数据的重要性。衰减函数如下所示：

$$f(t)=e^{-\alpha t}\ (0<\alpha<1) \tag{1}$$

可以看出，α 越大，历史数据的重要性随时间降低得越快。

定义 2（时刻权重 $WD(i,j)$ 和 $BD(i,j)$） 免疫系统识别抗原的过程就是寻求一个能够与抗原亲和力最大的抗体的过程。时刻权重 $WD(i,j)$ 和 $BD(i,j)$ 用来衡量抗原与抗体、抗体与抗体之间的匹配程度，在 T_i 时刻任意 X_i 与 Y_j 的时刻权重 $WD(i,j)$ 为：

$$\begin{aligned}WD(i,j)&=\sqrt{\sum_{t=1}^{d}(X_i-Y_j)^2}\times\sum_{t=1}^{d}f(t)\\&=\sqrt{\sum_{t=1}^{d}(X_i-Y_j)^2}*\sum_{t=1}^{d}e^{-\alpha t}\end{aligned} \tag{2}$$

式中，$1\leqslant i\leqslant d,1\leqslant j\leqslant d$。

在 T_i 时刻，任意 Y_i 与 Y_j 之间的时刻权重 $BD(i,j)$ 为：

$$BD(i,j)=\sqrt{\sum_{t=1}^{d}(Y_i-Y_j)^2}*\sum_{t=1}^{d}e^{-\alpha t} \tag{3}$$

式中，$1\leqslant i\leqslant d,1\leqslant j\leqslant d$。

定义 3（激励度 aff_{ij}） 激励度 aff_{ij} 用来表示抗原与抗体之间的刺激度，设 ρ_{ij} 表示抗原对抗体激励的阈值。当 X_i 与 Y_i 的时刻权重 $WD(i,j)$ 小于阈值 ρ_{ij} 时，其激励度 aff_{ij} 为 $\frac{1}{0.1+WD(i,j)^{\frac{1}{2}}}$，若 $WD(i,j)$ 大于或等于 ρ_{ij}，则激励度 aff_{ij} 为 0。具体表示如下：

$$aff_{ij}=\begin{cases}\dfrac{1}{0.1+WD(i,j)^{\frac{1}{2}}}, & WD(i,j)<\rho_{ij}\\ 0, & WD(i,j)\geqslant\rho_{ij}\end{cases} \tag{4}$$

$$\rho_{ij}=2\lambda\cdot\frac{\sum_{i=1}^{d}\sum_{i=1}^{d}WD(i,j)}{d^2} \tag{5}$$

式中，λ 为(0,1)之间的随机数。

3 基于人工免疫原理的数据流聚类算法 IMStream

3.1 期望克隆率的计算

抗体克隆的数目与其所受到的激励水平成正比，受激励越大的抗体，其克隆数目越多，受激励越小的抗体克隆的数目越少。为了保持抗体的多样性，抗体的期望克隆率与时刻权重 $BD(i,j)$ 成反比。抗体期望克隆率 $E(x_i)$ 可表示为：

$$E(x_i)=\frac{aff_{ij}}{BD(i,j)} \tag{6}$$

3.2 抗体克隆

通过抗体克隆和超变异机制来产生适当的抗体，从而使网络中的抗体结构可以反映遇到的抗原特征。当抗体所受到来自抗原的刺激达到一定程度时，免疫系统被激活，系统开始对抗体进行克隆增殖。设 $C(i,j)$ 为受抗原 X_i 激励所生成的抗体 Y_i 的克隆数目，且：

$$C(i,j)=E(x_i)*(1+aff_{ij}^2) \tag{7}$$

由式(2)—式(4)、式(6)、式(7)可推出：

$$C(i,j)=\frac{1}{\left(0.1+\left(\sqrt{\sum_{t=1}^{d}(X_i^i-Y_j^j)^2}*\sum_{t=1}^{d}e^{-\alpha t}\right)^{\frac{1}{2}}\right)*\sqrt{\sum_{t=1}^{d}(Y_i^i-Y_j^j)^2}*\sum_{t=1}^{d}e^{-\alpha t}}*\left[1+\left(\frac{1}{0.1+\left(\sqrt{\sum_{t=1}^{d}(X_i^i-Y_j^j)^2}*\sum_{t=1}^{d}e^{-\alpha t}\right)^{\frac{1}{2}}}\right)^2\right] \tag{8}$$

则新生成的抗体集合 C 为：

$$C=\lfloor C_1,C_2,\cdots,C_j,\cdots,C_d\rfloor$$

式中，$C(i,j)$ 为抗体克隆数，C 为抗体集合。

3.3 抗体变异

设 $\delta(i,j)$ 为抗体受到抗原刺激后新产生的抗体的变异率。变异率 $\delta(i,j)$ 与激励度 aff_{ij} 成反比，即抗体在变异的过程中，若受抗原的激励度越大，变异越小；若受抗原的激励度越小，变异越大[8,9]。由抗原 X_i 入侵引起的 Y_j 的克隆所生成的 C_j 变异为：

$$C_j'=C_j+\delta(i,j)*(C_j-X_i) \tag{9}$$

式中，

$$\begin{aligned}\delta(i,j)&=\frac{1}{aff_{ij}}=0.1+WD(i,j)^{\frac{1}{2}}\\&=0.1+\left(\sqrt{\sum_{t=1}^{d}(X_i-Y_j)^2}*\sum_{t=1}^{d}e^{-\alpha t}\right)^{\frac{1}{2}}\end{aligned} \tag{10}$$

3.4 记忆细胞生成和亲和度成熟

在抗体经过克隆和变异后，克隆的抗体种群 C 和变异产生的抗体种群 C' 就构成了抗体种群 B。所谓记忆细胞的生

成，就是从种群 B 中选出一定数量的亲和度高的抗体更新记忆细胞库中的抗体，使得记忆细胞库一直保持固定数目为 M 的最高亲和度抗体[10]。

亲和度成熟是根据设定的条件判断抗体是否达到最优，如果达到则结束寻优过程，否则从种群 B 中选择出抗体进行下一次操作。

3.5 算法描述

基于人工免疫原理的数据流聚类算法 IMStream 描述如下。

Step1 初始化。设要识别的原始数据流 X_i 为免疫网络中的抗原，此时抗体 Y_i 通过随机产生。

Step2 判断是否满足终止条件。若满足则退出，输出集合 M；否则，执行 Step3。

Step3 计算抗原与抗体时刻权重 $WD(i,j)$、抗体与抗体之间的时刻权重 $BD(i,j)$；计算任意抗体来自抗原的激励度 aff_{ij}，同时确定激励阈值 ρ_{ij}。

Step4 判定抗体是否要进行克隆。对任意抗体，它与抗原间的时刻权重大于激励度 aff_{ij} 时，则抗体自身要进行克隆，其克隆的数目为 $C(i,j)$，最后新生成的抗体集合为 C。

Step5 抗体变异。根据抗体的变异率，由抗原 X_i 入侵引起的 Y_j 的克隆所生成的 C_j 变异为 C_j'，变异产生的抗体种群为 C'。

Step6 抗体经过克隆和变异后，就构成了抗体种群 B。记忆细胞生成就是从种群 B 中选出一定数量的亲和度高的抗体更新记忆细胞库中的抗体，使得记忆细胞库一直保持固定数目为 M 的最高亲和度抗体。

Step7 判断抗体是否达到最优，如果达到则结束寻优过程，否则从种群 B 中选择出抗体进行下一次操作，跳转到 Step2。

4 实验结果与分析

为测试本算法的有效性，对算法进行仿真，并与传统 CluStream 算法进行性能上的对比。实验环境为：Intel PentiumD 主频2.8GHz，2GB 内存，操作系统为 Microsoft Windows XP Professional Edition。

本实验数据集包括人工数据集和真实数据集。人工数据集中数据点的个数、维数及各种分布特性可以预先设定，用人工数据集来测试 IMStream 算法和 CluStream 算法的处理速度与内存消耗。设定人工数据集为 50 万个数据，数据点的分布随时间变化。图 1 是 IMStream 算法与 CluStream 算法内存占用量的比较，其中横坐标表示的是时间点，纵坐标表示的是内存占用的数量(单位为 kb)。

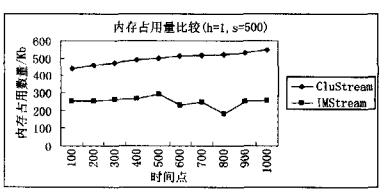

图 1 内存占用量的比较

从图 1 可以看出，本文所提出的 IMStream 算法较之传统的 CluSteam 算法在相同的条件下所占用的内存较少。图 2 为 IMStream 算法与 CluStream 算法处理速度的比较，其中横坐标为时间点，纵坐标为处理时间(单位为 ms)，从图 3 中可以看出，IMStream 算法的平均速度要比 CluStream 算法的平均速度快。

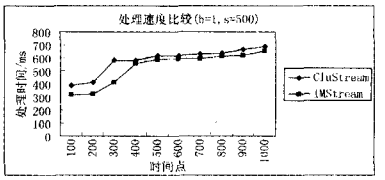

图 2 处理速度的比较

使用真实数据集 KDD-CUP-99 对 IMStream 算法和 CluStream 算法的聚类精度进行比较。图 3 是 IMStream 算法与 CluStream 算法聚类精度的比较结果，其中横坐标表示的是数据流时间(单位为 ms)，纵坐标表示的是聚类精度。从图 3 中可以看出，IMStream 算法的聚类精度要高于 CluStream 算法的聚类精度。

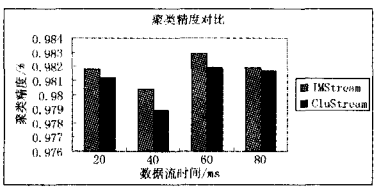

图 3 聚类精度对比

结束语 本文研究了基于人工免疫原理的数据流聚类问题，提出了 IMStream 算法，并通过实验验证了该算法的有效性，即相比于传统 CluStream 算法，本算法在聚类精度、处理速度及内存消耗方面都有明显的优势。因此，本文算法会有其自身的理论与应用价值。

参 考 文 献

[1] 黄德才，吴天虹. 基于密度的混合属性数据流聚类算法[J]. 控制与决策，2010，25(3)：416-421

[2] 丁永生，任立红. 人工免疫系统：理论与应用[J]. 模式识别与人工智能，2000，13(1)：52-59

[3] Aggarwal C C, Han Jia-wei, Wang Jian-yong, et al. A framework for clustering evolving data streams[C]// Proc the 29th VLDBConference. Berlin：Morgan Kaufmann，2003：81-92

[4] 常建龙，曹锋，周傲英. 基于滑动窗口的进化数据流聚类[J]. 软件学报，2007，18(4)：905-918

[5] 邹凌君. 流数据的聚类分类算法研究[D]. 扬州：扬州大学，2008：50-80

[6] 朱蔚恒，印鉴，谢益煌. 基于数据流的任意形状聚类算法[J]. 软件学报，2006，3：379-387

[7] 王述云. 数据流频繁项挖掘与聚类分析的研究[D]. 上海：复旦大学，2009：106-125

[8] Forrest S, Perelson A S, Cherukupi R, et al. Self-nonself discrimination in a computer[C]// The Symposium on Research in Security and Privacy. Oakland，CA，USA，1994：202-212

[9] Farmer J D，Packard N H，Perelson A S，et al. The immune system，adaptation，and machine learning [J]. Physical D，1986，22：87-204

[10] Liu T，Wang Y C，Wang Z J，et al. The distance concentration artificial immune algorithm[J]. Journal of China University of Mining & Technology，2005，2：81-85