

基于构造型神经网络的异常模式发现方法

张贤骥 王伦文

(电子工程学院 404 室 合肥 230037)

摘要 数据流的异常模式检测中,有时受噪声等因素影响发生概念漂移,影响了检测效率。针对此问题,提出一种基于构造型神经网络增量学习的异常模式动态检测方法,以提取滑动窗口内数据的数据概要,修正全局数据概要,更新已有的学习模型。另外,数据流速、流量等因素也影响检测效率,采用粒度分析思想改进检测方法,设置合适的时间滑动窗口,根据数据量自适应选择分析粒度,进而更准确地发现异常模式。无线电通信信号监测数据异常模式检测实验验证了本方法的有效性。

关键词 数据流,异常检测,动态检测,构造型神经网络

中图分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.07.061

Outlier Detection Method Based on Constructive Neural Networks

ZHANG Xian-ji WANG Lun-wen

(Research Room 404, Electronic Engineering Institute, Hefei 230037, China)

Abstract Outlier detection efficiency in data stream can be influenced by concept drift when there is noise. An outlier dynamic detection method based on constructive neural networks incremental learning was presented to solve this problem. The outline of the data in the sliding window is acquired, and the learning model is modified. On the other hand, data moving speed and flux can influence the efficiency as well. In this paper, the method of granular analysis was used to improve our method. The best analysis granularity in suitable sliding window was set to find outlier accurately. The simulate experiment and the experiment of outlier detection in radio communication demonstrated the efficiency of this method.

Keywords Data stream, Outlier detection, Dynamic detection, Constructive neural networks

1 引言

Hawkins 指出异常是在数据源中与众不同的数据,使人怀疑这些数据并非随机偏差,而是产生于完全不同的机制^[1]。这种定义偏于抽象,文献[2]认为异常是一种模式,这种模式的数据并不满足我们事先定义的正常数据的特征或者行为,而在整个数据集中找出符合这种模式的数据就称为异常检测。异常检测经常用于网络流量管理、入侵和欺诈检测、金融数据分析、环境监测、故障诊断、工业控制等领域,因此深受人们的重视。

国内外许多学者进行了广泛和深入的研究,文献[3]概括现有的异常检测的技术主要有基于统计的、基于模式预测的、基于数据挖掘的、基于神经网络的和基于免疫系统的异常检测等。近年来,由于异常检测实时性需求和数据流技术的产生,人们提出了许多基于数据流技术的异常模式检测方法。文献[4]提出了一种基于距离的分布式 RFID 数据流孤立点检测算法,但其对全局孤立点检测的准确率还有待提高。文献[5]提出了一种基于 Filter-ary-Sketch 数据结构的异常检测方法,该算法在一定精度条件下时间复杂度和空间复杂

度相对较低,但在基于熵的异常检测维度选择以及多维度整合方面还需要进一步的理论和实验证明。文献[6]使用基于采样的方法定义特异性因子,可在不影响精度的情况下,将采样特异性因子和最优采样频率结合起来进行实时异常检测。文献[7]通过引进 Mahalanobis 距离,并改进了传统的 k 近邻孤立点检测方法,提出了适用于关联属性的样本自适应参数孤立点检测法,提高了检测精度,降低了误检率,但算法复杂度较高。文献[8]深入分析异常产生的原因,广泛阐述了异常检测的应用场景,按异常检测原理进行方法分类,但未能在监督和半监督学习方面深入分析现状。

现实生活中许多传感器经常受入侵、噪声等因素影响,被截获的数据流会产生概念漂移,代表数据对象的特征值的分布产生了变化,因此降低了异常模式检测效率。文献[2]针对数据流中概念漂移的难题,引入了概念漂移检测和容忍策略来适应数据流分布的变化,但必须选择和更新训练数据,影响了算法效率。本文利用构造型神经网络运算复杂度低、适合增量学习等特点^[9],提出基于构造型神经网络的数据流异常模式的动态检测方法。

本文第2节介绍基于构造型神经网络与数据概要间的关

到稿日期:2013-09-09 返修日期:2013-12-15 本文受国家自然科学基金(61273302),安徽省自然科学基金(1208085MF98,1208085MF94)资助。

张贤骥(1990—),男,硕士生,主要研究方向为智能信息处理理论及应用,E-mail:hfzhangxj@126.com;王伦文(1965—),男,博士生导师,主要研究方向为智能信息处理。

系,并在此基础上构建和更新异常检测模型,检测异常模式;第3节采用粒度分析思想改进异常模式的检测方法;第4节将该方法应用到无线电信号异常检测中,用实验验证该方法的效率。

2 基于构造型神经网络的模型更新与异常检测

由于概念漂移在大多数情况下是渐变的(若某一时刻突发的概念漂移,本文称这种情况为异常模式),因此当数据流中发生概念漂移时,其变化之后的数据对象在开始阶段并不会表现得太过明显,但随着时间的推移,新的数据对象的特征将呈现和漂移发生之前完全不同的分布。这里采用构造型神经网络对每个新的滑动窗口数据再学习,微调网络参数,提高异常模式动态检测效率。

2.1 构造型神经网络与数据概要

构造型神经网络的核心内容是覆盖算法^[9,10],其主要思想是原空间样本(n 维)经式(1)变换,投到 $n+1$ 维球面上。

$$T: D \rightarrow S^n, T(x) = (x, \sqrt{r^2 - \|x\|^2}) \quad (1)$$

其中, $r \geq \max\{\|x_i\|\}$ 。在超球面上,通过覆盖算法实现样本空间的划分,这样就将模式识别问题转化成某种求最优覆盖的问题。第 k 类样本 X_k 的“球形领域”的构造方法为任取尚未被覆盖的点 a_i ,按式(2)计算覆盖半径 $\theta = r(k)$ 。

$$\begin{cases} r^1(k) = \max_{x \in X_k} \{\langle a_i, x \rangle\} \\ r^2(k) = \min_{x \in X_k} \{\langle a_i, x \rangle \mid \langle a_i, x \rangle > r^1(k)\} \\ r(k) = (r^1(k) + r^2(k)) / 2 \end{cases} \quad (2)$$

其中, $\langle x, y \rangle$ 表示样本 x 与 y 的内积, $r^1(k)$ 是样本点 a_i 与异类样本 x 的最小距离, $r^2(k)$ 是 a_i 与同类样本 x 的最大距离。

为便于说明问题,以在二维空间构建覆盖为例,如图1所示。数据经过覆盖算法后,呈现图1的分布,有3个覆盖,分别代表3个类。A类中心为 ω_1 ,半径为 r_1 ,对应神经元的权重和阈值,这将空间划分直接与神经网络构建联系起来。故该网络在处理给定的数据的过程中,同时给出网络的结构和参数,不像BP算法需反复迭代。

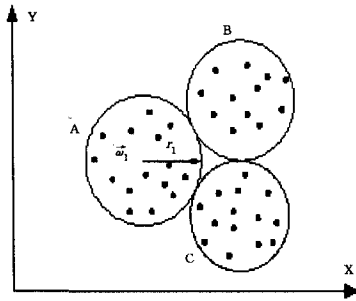


图1 球面覆盖簇投影示意图

一个数据簇对应一个覆盖,如图1所示,数据簇的数据概要就是覆盖的中心和半径,也为一个神经元的权重和阈值。因此可用构造型神经网络描述数据概要。设数据流为 n 维欧氏空间的点集 $\vec{N}_i = \{\vec{N}_1, \vec{N}_2 \dots \vec{N}_n\}$,经覆盖算法得到数据概要 $\vec{C}_i = \{\omega_i, r_i, t_\alpha, t_\beta\}$, $i=1, 2, \dots, n$,其中 ω_i 是第 i 个覆盖簇的圆心(第 i 个数数据模式均值), r_i 表示覆盖半径(第 i 个数数据模式波动范围), t_α 表示第 i 个覆盖簇的出现时间, t_β 表示第 i 个覆盖簇的结束时间。

数据概要的提取沟通了数据与神经网络的联系,新的滑

动窗口数据相当于增量学习样本,按式(1)与式(2)计算新的数据概要,即得到网络参数。因此,构造型神经网络适合增量学习。

2.2 增量学习与模型动态更新

与处理静态数据时只需要一次性学习不同,在数据流环境下,初期可以学习的信息是有限的、不完备的。其次代表数据对象的特征值的分布会随时间的变化而产生缓慢变化,最初建立的学习模型就不再适应新的数据。这时就需要进行增量学习,增量学习就是根据新的数据,结合已有的知识更新新知识。

构造型神经网络进行增量学习时,不需要对整个样本集进行重新学习,只需要根据已学知识和新样本进行训练即可。数据流新的窗口数据为增量学习样本,若数据流中存在概念漂移,增量学习将修正模型结构,为动态检测异常模型提供有利依据。

基于构造型神经网络增量学习就是针对于初始覆盖簇,结合新到数据的特征,不断更新覆盖簇数据概要,使其能更好地覆盖新数据的过程。增量学习能够有效消除概念漂移带来的影响。下面给出基于构造型神经网络增量学习的具体算法步骤。

算法1 Updata()

输入:初始数据 $\vec{N}_1$,流数据 $\vec{X}_i = \vec{X}_1 \dots \vec{X}_j \dots \vec{X}_n$;

输出:经过增量学习更新的覆盖簇 $\vec{C}_i$ 。

Step 1 begin

Step 2 $\vec{C}_1 = \text{cluster}(\vec{N}_1) / *$ 利用覆盖聚类算法对数据非监督学习 $*/[11]$

Step 3 $\vec{C}_1 = \{\omega_1, r_1, t_\alpha, t_\beta\} / *$ 获取初始覆盖簇的数据概要,包括覆盖簇的圆心 ω_1 、覆盖半径 r_1 、覆盖簇的出现时间 t_α 和结束时间 t_β $*/$

Step 4 for $j=1:n$, for $l=1:s / * n$ 为待识别窗口的个数, s 为窗口中所含数据的个数 $*/$

Step 5 if $d(\vec{X}_{j_l}, \vec{C}_1) < r_1, \vec{X}_{j_l} \in \vec{C}_1 / *$ 如果新数据 $\vec{X}_{j_l}$ 能被现有覆盖簇 $\vec{C}_1$ 吸收 $*/$

Step 6 $\omega'_1 = [(\omega_1 * s) + \vec{X}_{j_l}] / (s+1), r'_1 = \max(d_i) / *$ 更新覆盖簇参数,其中 s 为覆盖簇中所含数据的个数, d_i 为覆盖簇重心到各样本的距离 $*/$

Step 7 else $\vec{C}_{n+1} = \text{cluster}(\vec{X}_{j_l}) / *$ 以 $\vec{X}_{j_l}$ 为中心新建覆盖 $\vec{C}_{n+1}$ $*/[11]$

Step 8 end for

Step 9 end

Step 2将初始数据聚类;Step 3获得数据概要,同时也建立了神经网络;Step 4—Step 7为增量学习,修正数据概要和神经网络。通过对新的滑动窗口数据的学习可实时调整检测模型,为准确检测异常模式提供依据。

2.3 异常模式动态检测

异常模式中的数据并不满足我们事先定义的正常数据的特征或者行为。异常模式的检测本质上就是对这些异常数据的检测。为了检测出这些异常数据,首先需要提取定义的正常数据的特征,通过覆盖聚类算法处理过的数据,能将特征类似的数据聚集在一起形成一个覆盖簇。这样不仅能够消除噪音和干扰带来的影响,而且使得数据特征明显,易于下一步的判断。

定义正常数据为 n 维欧氏空间的点集 $\vec{N}_i = \{\vec{N}_1, \vec{N}_2 \dots$

$\vec{N}_n$ }, 利用覆盖聚类算法^[11]对 $\vec{N}_i$ 进行处理, 提取其数据概要 $\vec{C}_i = \{\omega_i, r_i, t_{ai}, t_{\beta i}\}$, $i=1, 2, \dots, s$, 其中 s 为覆盖个数, ω_i 为第 i 个覆盖重心, r_i 为第 i 个覆盖半径; t_{ai} 表示覆盖簇出现时间, $t_{\beta i}$ 表示覆盖簇结束时间。通过对数据概要的观察发现, 覆盖簇 $\vec{C}_i$ 的特征主要可视为两方面, 1) 以 ω_i 为圆心、 r_i 为半径的覆盖范围; 2) 覆盖簇的持续时间 $\Delta t_i = t_{ai} - t_{\beta i}$ 。从这两方面对被检测数据进行分析, 若数据不能被覆盖簇吸收, 即数据不在以 ω_i 为圆心、 r_i 为半径的覆盖范围内, 则认为数据异常; 若数据持续时间不符合模本, 亦为异常模式。

下面给出异常模式的动态检测的算法:

算法 2 Burst()

输入: 初始覆盖簇 $\vec{C}_i$, 流数据 $\vec{X}_i = \vec{X}_{i_1} \dots \vec{X}_{i_j} \dots \vec{X}_{i_n}$

输出: 异常数据 $\vec{Y}_i$

Step 1 begin

Step 2 for $j=1:n/n$ 为第 i 个窗口中的数据个数 *

Step 3 if $|\vec{X}_{i_j} - \omega_i| > r_i$ or $\Delta t_{\vec{X}_{i_j}} > (1+\alpha)\Delta t_i$ or $\Delta t_{\vec{X}_{i_j}} < (1-\alpha)\Delta t_i$

then $\vec{Y}_{i_j} = \vec{X}_{i_j}$ /* 记录下异常数据, 其中 α 为阈值 */

Step 4 else update($\vec{C}_i, \vec{X}_{i_j}$) /* 未见异常, 构造型神经网络增量学习 */

Step 5 end for

Step 6 end

以上从数据概要出发, 给出了异常模式检测方法。由于采用了增量学习的方法, 有效地抑制了概念漂移给异常检测带来的影响。然而, 流数据的流速和流量时有变化, 给异常检测带来了不确定的因素, 这里采用粒度分析方法加以改进。

3 基于粒度分析的模型改进

数据流处理模型按所选取的时序范围, 分为界标模型、滑动窗口模型和快照模型。在这 3 种模型中, 界标模型和滑动窗口模型应用较为广泛。由于构造型神经网络进行增量学习时, 不需要对旧样本重新学习, 只需要根据已学知识和新样本进行训练即可。而滑动窗口模型不断接收新数据, 所以滑动窗口模型适用于基于构造型神经网络的增量学习。

由于数据流量、数据流速等因素都影响滑动窗口中所包含的数据个数, 如何定义滑动窗口呢? 滑动窗口大小的定义有很多种, 最常见的是由一个时间区间定义以及由窗口所包含的数据个数定义。由于在异常模式检测中, 时间的周期性是一个非常重要的属性, 而采用窗口所包含的数据个数定义的窗口大小无法反映这一属性, 且与流速等因素有关, 因此本文采用由一个时间宽度定义的滑动窗口大小。

一个固定时间的滑动窗口中所包含的样本个数受到数据流量、数据流速等因素影响。这些因素严重影响挖掘效果, 为此前人开展了许多研究。文献[12]通过对应用于数据挖掘领域的代表性取样方法的比较研究和分析总结, 提出了一个取样算法分类框架, 同时指出基于滑动窗口模型的取样技术存在着内存上界不确定、附加成本较高和滑动窗口较小等诸多限制, 并且大多局限于处理只插入或少量删除的情形, 需要进一步研究能处理带任意顺序的插入和删除操作或存在频繁增删情况的数据取样问题。文献[13]构造了基于环形缓冲区循环的逻辑滑动窗口, 提出了基于环形循环滑动窗口的数据流抽取算法, 将环形缓冲区和环形循环滑动窗口紧耦合, 保证了数据流的挖掘持续正常实施。文献[14]引入了类型变化界限的概念, 提出频繁项集的挖掘算法 FIMoTS, 其能在内存开销

基本不变的前提下显著提高运算效率。

本文采用基于商空间的变化粒度挖掘方法解决上述问题。如图 2 所示, 根据数据流的数据量大小和流速情况, 自适应地调整分析粒度。设置固定的时间窗口作为滑动窗口, 若流速快或流量大, 窗口内数据多, 采用抽样技术减少数据量 (即粗粒度分析) 以提高运算速率, 降低运算消耗。初始粒度根据历史知识 (规则) 设置, 粒度调整由数据数量大小、数据流速、数据密度与参考粒度等决定, 采用关系合成与分解形成新的分析粒度。异常检测是指在新的粒度下发现异常模式, 粒度判定由知识 (规则) 的价值决定, 根据挖掘的知识决定粒度是否调整以及调整的份额^[14]。具体算法如下:

算法 3 GranularSet()

输入: 初始覆盖簇 $\vec{C}_i$, 窗口数据 $\vec{N}_i = \vec{N}_{i_1} \dots \vec{N}_{i_j} \dots \vec{N}_{i_n}$

输出: 规则输出 $\vec{X}_i = \vec{X}_{i_1} \dots \vec{X}_{i_j} \dots \vec{X}_{i_m}$

Step 1 begin

Step 2 load($\vec{N}_i$) /* 读取第 i 窗口的数据 */

Step 3 $\Gamma = \{\Delta T, v\}$ /* 获取窗口数据信息, 其中 ΔT 为数据时间长度, v 为数据流速 */

Step 4 if $\Delta T \cdot v \leq \lambda$ then $\vec{X}_i = \vec{N}_i$, go Step 7 /* λ 为阈值, 如果数据流量和流速超过这个阈值, 则对窗口数据进行采样处理 */

Step 5 else $\alpha = \text{int}(\Delta T \cdot v / \lambda + 0.5) / \alpha$ 为抽样率, 根据流速设定, 流速越快, α 越大, 则分析粒度越粗 */

Step 6 $\vec{X}_i = \vec{N}_i / \alpha$ /* 以抽样率 α 对窗口数据 $\vec{N}_i$ 进行采样, 并存入 $\vec{X}_i$ 中 */

Step 7 end

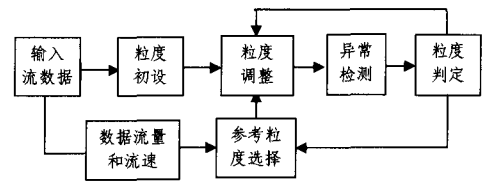


图 2 粒度调整原理框图

通过增量学习能抑制概念漂移给异常检测带来的影响, 利用粒度分析改进数据流处理模型, 克服数据流流速波动带来的影响, 在保证异常检测精度的前提下, 有效降低运算内存消耗, 提高运算速度和适应能力。

4 实验与分析

本文所有实验均在 CPU 为 Pentium E5300@2.60GHz 2.59GHz, 内存为 2GB, 操作系统为 WINDOWS XP 的电脑上运行。

为了说明本文所提算法 (简称 GBurst) 的性能, 这里从 3 个方面将其与基于构造型神经网络增量学习的异常模式动态检测算法 (Burst) 和文献[4]中的基于距离的 LSOD 算法 (LSOD) 进行比较。1) 在一定的流速下, 3 种算法的运算时间; 2) 对一定数量的滑动窗口 (60 个窗口), 在不同流速下的处理时间; 3) 3 种算法的检测精度。

实验数据来源于无线电监测数据, 将无线电接收机调谐到某频率 (本实验为 15.06MHz) 上接收信号, 在接收机中频输出端持续采集时域信号, 经快速傅里叶变换得到图 3 所示的频域数据的时间和幅度二维图 (时幅图)。图中横轴代表时间, 单位为分钟; 纵轴代表信号功率谱幅度大小, 归一化幅值最大为 90, 相当于中频输出 50 毫伏。由于多径效应、噪声等

因素影响,无线电信号幅度不稳定。

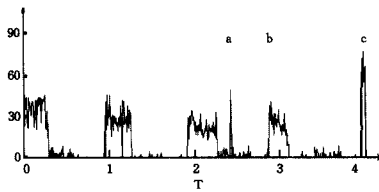


图3 15060kHz 信号时幅图

图3中的数据实际上分为两类,即幅度大于零代表有信号一类,幅度接近于零或等于零代表没有信号一类。

常规的定频无线电信号通常工作在固定频率上,在时间上呈现一定规律性,如无线电广播信号在时间上有一定的周期性,在能量上也相对稳定,基于此规律我们可以检测异常信号。如图3中,前两分钟的信号为规律信号;信号a为突发信号,信号b的持续时间比历史同期信号短,信号c的能量超出平均范围;信号a,b,c均为异常信号。

实验采用固定时间滑动窗口,窗口长度为1s。首先设定数据流速为10kB/s,分别比较3种算法的运算时间,实验结果如图4所示。由于构造型神经网络需要对初始流数据进行学习,构造样本覆盖簇,因此Burst算法和GBurst算法初始处理时间较长。而这两种算法对之后的数据只需要用样本覆盖簇去识别,处理时间大大减少。LSOD算法在初始数据的学习耗时较少,但对之后的数据的学习需经过删除过期对象、插入新对象、孤立点查询3步,处理时间较长。

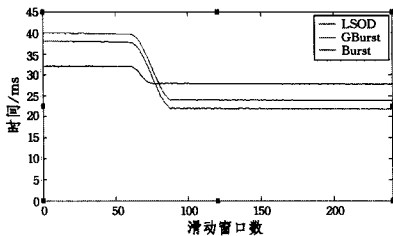


图4 3种算法运行时间比较

为了体现采用粒度分析的效果,实验将在不同流速的条件下,比较3种算法对第150个到第210个滑动窗口的处理时间,但窗口长度为1s。实验结果如图5所示。

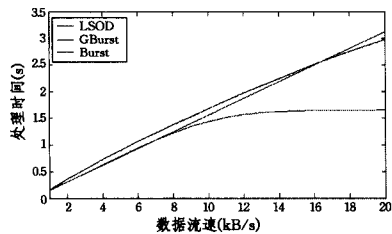


图5 不同流速条件下运行时间比较

由图5可以看出,在流速较低的时候,无需对窗口数据进行抽样,因此GBurst算法和Burst算法所消耗时间并没多大区别,而LSOD算法耗时较多。而当流速提升后,GBurst算法由于采用了粒度分析思想,对窗口数据进行采样,减少需要处理的数据量,提高了运算效率,明显减少了消耗时间;LSOD算法由于采用了随机采样技术以及删除过期数据来减少运算时间,因此在较高的流速下,LSOD算法的运算时间也优于Burst算法。表1列出了3种方法是否能检测出异常信号。

表1 不同方法异常信号检测结果表

算法	信号a	信号b	信号c
LSOD	是	否	是
GBurst	是	是	是
Burst	是	是	是

由表1可以看出,GBurst算法和Burst算法均能检测出3个异常信号;LSOD算法由于主要是针对于孤立点检测的算法,无法对异常信号b进行有效的检测。

实验结果证明,GBurst算法可以在减少处理时间的同时,对无线电信号中的异常信号进行有效、准确的检测。

结束语 针对流数据受噪声等因素影响发生概念漂移,影响数据流中异常模式检测效率的问题,本文将增量学习的概念引入到异常模式检测的问题中,提出一种基于构造型神经网络增量学习的异常模式动态检测方法。由于构造型神经网络将数据学习问题转化为空间覆盖问题,适合增量学习,另外神经网络参数直接从覆盖中获得,减少了运算量。

针对数据流速的变化影响检测效率问题,采用了粒度分析思想改进挖掘方法,合理设置了滑动窗口,提取了数据概要,更新学习模型,准确发现后续测试数据的异常模式。将增量学习的概念引入到异常模式检测中,可以提高异常模式检测的精度。实际数据实验结果表明本文提出的方法能够有效地发现异常信号。该方法适合高维数据流异常检测,有一定推广价值,但对混合属性数据流异常检测存在不足,这也是下一步研究的方向。

参考文献

- [1] Hawkins D. Identification of Outliers [M]. London: Chapman and Hall,1980;2-26
- [2] 陈霏. 数据流异常检测技术研究与实现[D]. 成都: 电子科技大学,2010
- [3] 杨宇舟. 基于特征提取的聚类异常检测技术的研究与实现[D]. 成都: 电子科技大学,2012
- [4] 廖国琼,李晶. 基于距离的分布式RFID数据流孤立点检测[J]. 计算机研究与发展,2010,47(5):930-939
- [5] 郑黎明,邹鹏,韩伟红,等. 基于Filter_ary_Sketch数据结构的骨干网异常检测研究[J]. 通信学报,2011,32(12):152-160
- [6] 牛之贤,孙静宇,石淑萍. 基于采样特异性因子的实时异常检测[J]. 计算机科学,2013,40(3):283-286
- [7] 刘胜宗,樊晓平,廖志芳. 适用于关联属性的样本自适应参数孤立点检测法[J]. 计算机应用研究,2012,29(9):3259-3262
- [8] Chandola V,Banerjee A,Kumar V. Anomaly Detection:a survey [J]. ACM Computing surveys,2009,41(3):1-58
- [9] 王伦文,张铃. 构造型神经网络综述[J]. 模式识别与人工智能,2008,21(1):49-55
- [10] Zhang Ling,Zhang Bo. A Geometrical Representation of McCulloch-Pitts Neural Model and Its Applications [J]. IEEE Transactions on Neural Networks,1999,(10):925-929
- [11] 章季阳,王伦文. 一种领域覆盖的数据流聚类算法[J]. 小型微型计算机系统,2012,33(9):1913-1916
- [12] 胡文瑜,孙志辉,吴英杰. 数据挖掘取样方法研究[J]. 计算机研究与发展,2011,48(1):45-54
- [13] 詹英,吴春明,王宝军. 一种与缓冲区紧密耦合的环形循环滑动窗口的数据流抽取算法[J]. 电子学报,2011,39(4):894-898
- [14] 李海峰,章宁,朱建明,等. 时间敏感数据流上的频繁项集挖掘算法[J]. 计算机学报,2012,35(11):2283-2292