

一种基于关系熵和 J 量值的网络事件关联模式漂移检测方法

杨英杰 刘 帅 常德显

(解放军信息工程大学 郑州 450001)

摘 要 针对传统的模式漂移检测方法无法直接适用于关联规则分析的问题,提出了一种基于关系熵和 J 量值的模式漂移检测方法。抽取并定义了4种特征属性:关系量 RC 、关系熵 RE 、窗口数 WC 和 J 量值,并通过对时间相邻的两个滑动窗口内的关系熵和 J 量值的取值分布进行假设检验,判断和定位模式漂移的发生,得到模式漂移影响较大的事件集和事件对集,它们可为关联规则的调整和更新提供支持。实验数据表明该方法准确可行。

关键词 关联模式,模式漂移,关系熵, J 量值,假设检验,模式漂移检测

中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.1.056

Pattern Drift Detection in Association Analysis Based on Relation Entropy and J Measure

YANG Ying-jie LIU Shuai CHANG De-xian

(PLA Information Engineering University, Zhengzhou 450001, China)

Abstract Aiming at the problem that traditional concept drift detection technology cannot be applied in association analysis directly, this paper proposed a method of concept drift detection based on relation entropy and J measure. Four features such as relation count, relation entropy, windows count and J measure are extracted. Concept drift is detected and located through hypothesis test of features distribution in two sliding windows. Event set and event pair set mostly effected by concept drift which can provide support to adjustment and refreshment of concept are obtained. Experimental results show that this method is accurate and feasible.

Keywords Association pattern, Pattern drift, Relation entropy, J measure, Hypothesis test, Pattern drift detection

网络事件关联分析是数据挖掘领域的一个研究热点。网络事件的关联模式会随着时间的推移、环境的转变以及用户行为方式的变化而变化,即模式漂移,这就要求学习系统能尽早检测到模式漂移,并对自身进行适应调整,以对不断到来的数据尽可能地做出正确判断。

目前的模式漂移检测大致分为3种方法:性能法、距离法和性质法^[1,2]。性能法是根据当前分类器或匹配器对最新数据进行分类或匹配后的性能变化判断模式是否发生漂移。比较经典的方法有Widmer和Kubat提出的FLORA系列算法^[3]、Last等提出的OLIN算法^[4]、罗秀等提出的基于误差率的检测方法^[5]等。这些基于性能的检测方法受其它因素的影响较大,误报率较高。距离法是将一个数据块映射成一个概念向量,然后对多个概念向量进行聚类,并通过计算新到来的数据集所对应的概念向量与各个聚类中心的距离来检测模式漂移的发生,Katakis^[6]用该方法实现了对重复模式漂移的检测。李南等人提出了一种基于少量类标签的模式漂移检测算法^[7],该算法根据当前数据块未被模型簇覆盖的实例数目是否发生显著增大,来判断是否发生模式漂移。但距离法的局限性在于:一个数据块中的数据应当属于同一概念。性质法^[8-11]是通过分析最近获得的数据的一些统计特性来实现对

模式漂移的检测。Alippi等^[8]利用中心极限定理,设计了不依赖数据分布模型、不需要任何先验信息的模式漂移检测算法。这种方法不需要任何先验信息,并且识别率较高,但对于数据特征值的选取要求非常苛刻。

当前学术界针对关联分析的模式漂移研究并不多,因为网络关联分析的侧重点在于网络事件的关系,所以以往针对分类技术的模式漂移检测方法并不完全适用。本文通过对事件关联模式的研究,选择定义了4种特征属性:关系量 RC 、关系熵 RE 、窗口数 WC 和 J 量值,提出了基于以上特征的模式漂移检测方法,通过对这些特征取值分布的假设检验,最终检测并定位模式漂移的发生,并同时得到模式漂移影响较大的事件集和事件对集,它们可为网络关联模式匹配的适应性调整提供支持。

1 基于事件关系的关联模式特征抽取

网络事件关联分析主要针对网络事件分析事件间的顺序关系及其所反映的关联模式。这里给出网络事件的相关符号表示: A 表示事件集,包含出现在该网络环境内部的所有事件; A^+ 表示 A 中所有的非空事件序列的集合,其中的一个事件序列表示为 $t=t(1)t(2)t(3)\cdots t(n)$, $|t|=n$ 表示事件序列

到稿日期:2014-12-24 返修日期:2015-05-08 本文受国家“863”计划资助项目(2012AA012704),郑州市科技领军人才项目(131PLJRC644)资助。

杨英杰(1971—),男,博士,副教授,硕士生导师,主要研究方向为网络安全、数据分析;刘 帅(1991—),硕士生,主要研究方向为网络安全、数据分析, E-mail: lius0228@163.com;常德显(1976—),男,博士,讲师,主要研究方向为网络安全。

的长度, $t(k)$ 表示事件序列中的第 k 个事件, $t(i, j)$ 表示事件序列 t 从第 i 个事件到第 j 个事件的子序列; $t^i = t(i, |t|)$ 表示从第 i 个事件开始之后的序列; 一种关联模式 M 由多个非空事件序列组成, 比如两个事件序列 $abdc$ 和 $acdb$ 同属于一种关联模式。模式漂移检测的目的就是发现关联模式 M 随时发生的变化。

事件之间一般存在两种基本关系: 跟随关系和前驱关系。对于任意两个事件 $a, b \in A$, 事件序列 $t = t(1)t(2)t(3) \cdots t(n) \in A^+$, 给出下列定义^[11]。

定义 1(跟随关系, follows)

$b \text{ follows } a \Leftrightarrow t(i) = a, \exists j(i < j < n), t(j) = b$

定义 2(前驱关系, precedes)

$a \text{ precedes } b \Leftrightarrow \forall j(1 < j < n), t(j) = b, \exists i(1 \leq i < j), t(i) = a$

上述两种关系定义的是两个事件在同一个事件序列中的关系, 而在一个事件流中, 根据统计量的累积, 每种关系均可延伸出 3 种关系。定义如下:

定义 3(总是跟随, always follows)

$b \text{ always follows } a \Leftrightarrow \forall t \in L, b \text{ follows } a \text{ in } t$

即当且仅当在事件流 L 所有的事件序列 t 中, b 均和 a 存在跟随关系。

定义 4(有时跟随, sometimes follows)

$b \text{ sometimes follows } a \Leftrightarrow \exists t \in L, b \text{ follows } a \text{ in } t$

即当且仅当在事件流 L 中, 存在一个事件序列 t , b 和 a 存在跟随关系。

定义 5(从不跟随, never follows)

$b \text{ never follows } a \Leftrightarrow \forall t \in L, b \text{ not follows } a \text{ in } t$

即当且仅当在所有的的事件序列 t 中, b 和 a 都不存在跟随关系。

例如在一个事件流 $L = [acaebfh, ahijebd, aeghijk]$ 中, 存在关系: e 总是跟随 a , b 有时跟随 a , e 从不跟随 b 。图 1 即示出了该事件序列的关系矩阵, 矩阵中 S 代表有时跟随, A 代表总是跟随, N 代表从不跟随。例如横坐标为 d 、纵坐标为 e 的取值 S 表示 d 有时跟随 e 。

	a	b	c	d	e	f	g	h	i	j	k	f_{RC}^L	f_{RE}^L
a	S	S	S	S	A	S	S	A	S	S	S	<2,9,0>	0.684
b	N	N	N	S	N	S	N	S	N	N	N	<0,3,8>	0.845
c	A	A	N	N	A	A	N	A	N	N	N	<5,0,6>	0.994
d	N	N	N	N	N	N	N	N	N	N	N	<0,0,11>	0.000
e	N	S	N	S	N	S	S	S	S	S	S	<0,8,3>	0.845
f	N	N	N	N	N	N	N	A	N	N	N	<1,0,10>	0.440
g	N	N	N	N	N	N	N	A	A	A	A	<4,0,7>	0.946
h	N	S	N	S	S	N	N	N	S	S	S	<0,6,5>	0.994
i	N	S	N	S	S	N	N	N	N	A	S	<1,4,6>	1.322
j	N	S	N	S	S	N	N	N	N	N	S	<0,4,7>	0.946
k	N	N	N	N	N	N	N	N	N	N	N	<0,0,11>	0.000

(a)

(b)

(c)

图 1 事件流 L 中事件的关系矩阵、关系量和关系熵

前驱关系在事件流中的延伸定义与跟随关系类似。关联模式的模式漂移必然会带来行为内部事件关系的变化, 继而引发整个事件流中事件关系的变化。反言之, 通过抽取事件

关系的特征, 并检测该特征值的变化, 即可反映出关联模式的模式漂移。本文基于上述事件关系抽取了 4 种特征属性: 对于事件流中的单一事件, 定义了其在某段事件流中的关系量 RC 和关系熵 RE , 以反映该事件在这段事件流中与其它事件的关联程度; 对事件序列中的某一事件对, 定义了窗口数 WC 和 J 量值, 以反映两个事件间的关系信息量。具体定义如下:

定义 6(关系量 RC , Relation Count) 在一段事件流 L 中, 事件 a 的关系量 RC 值表示为一个三元组 $\langle c_A, c_S, c_N \rangle$, 其中 c_A, c_S, c_N 分别表示总是跟随、有时跟随以及从不跟随在事件 a 之后的事件个数。即 $f_{RC}(a) = \langle c_A, c_S, c_N \rangle$ 。对于一个包含 $|A|$ 个事件的事件流而言, 其关系量 RC 值是一个矩阵, 如图 1 所示, 矩阵第一行表示其他事件与事件 a 的跟随关系, 事件 a 的关系量表示为 $\langle 2, 9, 0 \rangle$, 表示在事件序列 L 中, 有 2 个事件总是跟随 a , 9 个事件有时跟随 a , 0 个事件从不跟随 a 。若仅考虑跟随关系或者前驱关系中的一个, 关系量 RC 值的维度为 $3 \times |A|$, 若同时考虑两种关系, 则关系量 RC 值维度为 $3 \times 2 \times |A|$ 。关系量可以很好地反映一个事件与其它事件的关系紧密程度, 根据关系量, 可得如下关系熵 RE 的定义。

定义 7(关系熵 RE , Relation Entropy) 事件流 L 中, 事件 a 的关系熵 RE 值是其关系量 RC 值的熵, 即

$$f_{RE}(a) = -p_A \log_2(p_A) - p_S \log_2(p_S) - p_N \log_2(p_N) \quad (1)$$

其中 $p_A = \frac{c_A}{|A|}$, $p_S = \frac{c_S}{|A|}$, $p_N = \frac{c_N}{|A|}$, 并且 $\langle c_A, c_S, c_N \rangle = f_{RC}(a)$ 。例如图 1 中, 事件 a 的关系熵为 0.684。对于一段事件流而言, 若仅考虑跟随关系或前驱关系中的一个, 则其维度为 $|A|$ 。若同时考虑两种关系, 则其维度为 $2 \times |A|$, 通过检测事件关系熵的变化可以发现该事件与整体的关系情况的变化, 从而检测模式漂移的发生并得到事件受漂移影响的程度。

针对事件序列中的一对事件, 提出以下窗口数 WC 和 J 量值的定义。

定义 8(窗口数 WC , Windows Count) 对于事件流 L 中的一段事件序列, 本文设计一个长度为 l 的滑动窗口进行分析, $S^{i:l}(a)$ 表示以事件 a 开始的滑动窗口集合。 $\Gamma^{i:l}(a, b)$ 表示以事件 a 开始并且包含事件 b 的窗口集合, 即

$$\Gamma^{i:l}(a, b) = [s \in S^{i:l}(a) \mid \exists 1 < k \leq |s|, s(k) = b] \quad (2)$$

窗口数 WC 值即

$$f_{WC}^{i:l}(a, b) = |\Gamma^{i:l}(a, b)| \quad (3)$$

定义 9(J 量值) J 量值由 Smyth 和 Goodman 在文献 [14] 中提出, 是一种表示规则信息量的矩阵。由于可以将两个事件间的跟随关系看成一条规则, 因此对其进行改进用以表示事件 a 和事件 b 的信息量。 $p^i(a)$ 和 $p^i(b)$ 分别表示两事件在事件序列 t 中出现的概率。 $p^{i:l}(a, b)$ 表示事件 b 出现在从 a 开始的窗口中的概率, 根据定义 8, $p^{i:l}(a, b)$ 为同时包含事件 a 和事件 b 的窗口数与只包含事件 a 的窗口数之间的比值, 表示为

$$p^{i:l}(a, b) = \frac{|\Gamma^{i:l}(a, b)|}{|S^{i:l}(a)|} \quad (4)$$

那么, 一个大小为 l 的窗体的 J 量值表示为

$$f_J^{i:l}(a, b) = p^i(a) CE^{i:l}(a, b) \quad (5)$$

其中, $CE^{i:l}(a, b)$ 是 a 和 b 的相对熵, 定义为

$$CE^{t,t}(a,b) = p^{t,t}(a,b) \log_2 \left(\frac{p^{t,t}(a,b)}{p^t(b)} \right) + (1 - p^{t,t}(a,b)) \log_2 \left(\frac{1 - p^{t,t}(a,b)}{1 - p^t(b)} \right) \quad (6)$$

事件流上两段事件流之间 J 量值的不同可以很好地反映事件 a 和 b 间关系信息量的变化,本文通过捕获这些变化来检测模式漂移的发生,并可根据两个 J 量值间的差异程度找到受漂移影响较大的事件对,为关联模式的更新提供支持。

2 基于关系熵和 J 量值分布假设检验的模式漂移检测

模式漂移检测时,事件是以事件序列流的形式到来,本文主要根据关系熵和 J 量值在事件流上分布的变化来判断模式漂移的发生。本文设计了时间紧邻的双滑动窗口,一般情况下,窗口大小设为事件序列的平均长度。然后,假设在两个滑动窗口中特征值(关系熵或 J 量值)所遵从的分布是相同的,再依据窗口中实际数据进行假设检验,若假设错误,则证明发生模式漂移,否则,未发生模式漂移。具体检测流程如图2所示。

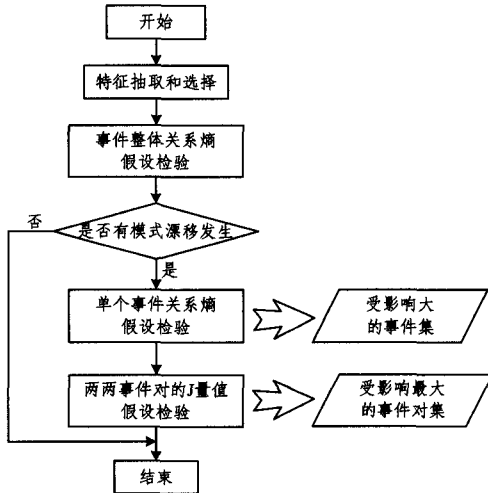


图2 模式漂移检测流程

2.1 双滑动窗口设计

本文首先将整个事件流分割成若干个较小的子流,然后分别根据式(1)一式(6)计算各子流中的关系量、关系熵、窗口数、 J 量值4种特征值,将4种特征的取值按时间顺序排列即可得到特征流。如前所述,两事件间的关系量和关系熵是以子流为单位进行计算的,而窗口数和 J 量值以事件序列为单位进行计算,故两种数据流的长度是不同的,如图3所示。

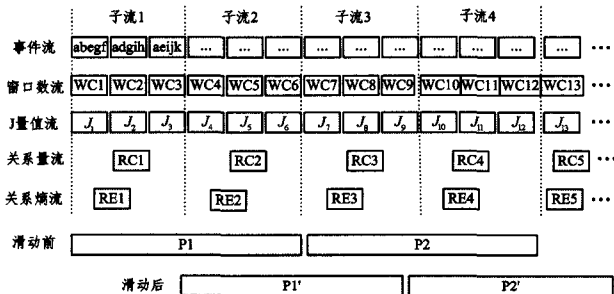


图3 抽取流特征与滑动窗口示意图

当模式漂移发生时,变化点前后的特征值分布将会发生

变化,为了检测到这种变化,我们同时设计在数据流上紧邻的两个大小为 w 的滑动窗口 P_1 和 P_2 ,假设在上述4种特征数据流中选取一种特征流 D ,其上的特征为 d ,在时间为 t 时, $P_1 = \{d_{t+1}, d_{t+2}, \dots, d_{t+w}\}$, $P_2 = \{d_{t+w+1}, d_{t+w+2}, \dots, d_{t+2w}\}$; 经过一次滑动后,时间为 $t+1$ 时, $P_1' = \{d_{t+2}, d_{t+3}, \dots, d_{t+w+1}\}$, 同时 $P_2' = \{d_{t+w+2}, d_{t+w+3}, \dots, d_{t+2w+1}\}$ 。图3则为选取关系熵流进行分析时,设定窗口大小为2时的实例。

在时间 t 时,关系熵或 J 量值在 P_1 和 P_2 内的分布相同,本文认为两窗口内未发生模式漂移,若在时间 $t+1$ 时,关系熵或 J 量值在 P_1' 和 P_2' 内的分布不再相同,则认为在从 t 到 $t+1$ 的这段时间内发生了模式漂移。本文采用下述假设检验的方法对两窗口中的特征分布进行比较。

2.2 关系熵和 J 量值分布假设检验

假设检验又称统计假设检验,是用来判断样本与样本、样本与总体的差异是由抽样误差引起还是由本质差别造成的统计推断的方法。目前假设检验方法有很多^[13],根据参数的有无可分为有参数假设检验和无参数假设检验两种,根据样本数目的不同可以分为单样本假设检验、双样本假设检验、多样本假设检验3种,根据变量的数目可分为单变量假设检验和多变量假设检验。本文采用了单变量双样本 Kolmogorov-Smirnov 检验(KS 检验)和 Mann-Whitney U 检验(MW 检验)以及多变量双样本 Hotelling T^2 检验方法。

由于要对两个窗口中的样本进行分析并且不知道特征值在窗口中的分布,因此采用双样本无参数假设检验。同时,只考虑一对事件的关系特征值时,需进行单变量假设检验,采用双样本 Kolmogorov-Smirnov 检验(KS 检验)和 Mann-Whitney U 检验(MW U 检验);同时考虑若干对事件时,需进行多变量假设检验,此时采用 Hotelling T^2 检验方法。

1) 事件整体关系熵假设检验

当时间为 t 时,对事件集 A 中任一事件 a ,根据式(1)可计算其在子流 i 中的关系熵为 RE_i^a ,然后分别计算其事件 a 在滑动窗口的所有子流中的关系熵取值,并按照时间顺序进行排列,可得到双滑动窗口下关系熵集合 P_1 和 P_2 :

$$P_1 = \{(RE_1^1, RE_1^2, \dots, RE_1^n), (RE_2^1, RE_2^2, \dots, RE_2^n), \dots, (RE_w^1, RE_w^2, \dots, RE_w^n)\}$$

$$P_2 = \{(RE_{w+1}^1, RE_{w+1}^2, \dots, RE_{w+1}^n), (RE_{w+2}^1, RE_{w+2}^2, \dots, RE_{w+2}^n), \dots, (RE_{2w}^1, RE_{2w}^2, \dots, RE_{2w}^n)\}$$

其中, $a_1, a_2, \dots, a_n \in A$, w 为滑动窗口的大小。假设 P_1 和 P_2 内的关系熵分布相同,进行相邻样本多变量 Hotelling T^2 假设检验,随着窗口滑动,可得到假设检验的 P 值变化曲线(P 值表示拒绝原假设的最小显著性水平), P 值变化的谷底点是最有可能拒绝原假设的时间点,若该点 P 值小于设定阈值,则拒绝原假设,即认为 P_1 和 P_2 内的关系熵分布不同,模式漂移发生。

2) 单个事件关系熵假设检验

当确定模式漂移发生后,在双滑动窗口内对每个事件的关系熵进行计算可得单个事件的关系熵集合 P_1 和 P_2 ,此时:

$$P_1 = \{RE_1^a, RE_2^a, \dots, RE_w^a\}$$

$$P_2 = \{RE_{w+1}^a, RE_{w+2}^a, \dots, RE_{2w}^a\}$$

假设 P_1 和 P_2 内的事件关系熵分布相同,进行相邻样本单变量 KS 假设检验,随着窗口滑动,可得假设检验的 P 值变

化曲线,然后根据 P 值在漂移位置的取值判断其受模式漂移影响的程度,最终得到受模式漂移影响最大的事件集合 A_m 。

3) 两两事件对的 J 量值假设检验

得到受模式漂移影响最大的事件集合 A_m 后,对任意 $a_i, a_j \in A_m$,根据式(5)可计算其在第 n 段事件序列中的 J 量值为 $J_n^l(a_i, a_j)$,将双滑动窗口中所计算的 J 量值按时间顺序排列可得事件 a_i 与事件 a_j 的 J 量值集合 P_1^l 和 P_2^l ,此时:

$$P_1^l = \{J_1^l(a_i, a_j), J_2^l(a_i, a_j), \dots, J_n^l(a_i, a_j)\}$$

$$P_2^l = \{J_{w+1}^l(a_i, a_j), J_{w+2}^l(a_i, a_j), \dots, J_{w+n}^l(a_i, a_j)\}$$

假设 P_1^l 和 P_2^l 内事件 a_i 与 a_j 的 J 量值所遵从的分布相同,对事件集进行相邻样本 MW U 假设检验,同样根据假设检验的 P 值变化曲线,可以判断模式漂移对每个事件对的影响程度,从而得到受模式漂移影响最大的事件对集合。

3 实验分析与讨论

为验证本文检测方法的可行性与准确性,以网上购物为例进行实验分析。如图 4 所示,可知购物行为由一系列相互关联的事件组成。图 4 中,由左边开始节点到右边末节点的任意一条路径均代表一个事件序列,所有的事件序列均属于网络购物这一行为模式,即关联模式。比如,在购物过程中用户会进行多次搜索和浏览来选择要购买的东西,之后用户登录并提交订单,最后完成付款,这表示一次成功的购物过程。若在完成付款前退出或中止,虽未成功购物但同样属于一种购物行为。由于购物网站的设计或者用户行为习惯发生了变化,关联模式从图 4(a)转变成图 4(b),即发生了一次模式漂移。

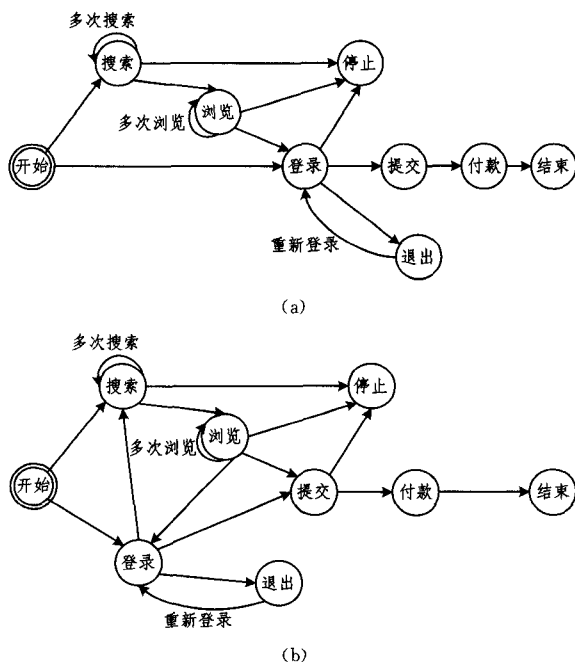


图 4 网上购物行为中的模式漂移

3.1 数据准备

本文从网上购物行为中提取了 9 个事件,如表 1 所列。购物关联模式在漂移发生之前有 8 种事件序列,发生之后有 11 种事件序列,如表 2 所列。

表 1 事件标识对应表

标识	a	b	c	d	e	f	g	h	i
事件	开始	搜索	浏览	登录	退出	停止	提交	付款	结束

分别在模式漂移发生之前和之后模拟 4400 次购物行为,前者的 8 种事件序列每种设定 550 个,后者的 11 种事件序列每种设定 400 个,并分别将两阶段的事件序列顺序打乱,然后将这 8800 次购物时间序列划分成 200 个子流,即每个子流中有 44 次购物行为发生。

表 2 模式漂移前后事件序列的变化

模式漂移发生前	模式漂移发生后
abf	abf
abcf	abcf
abcdf	abcdgf
abcdghi	abcdghi
adghi	adbcghi
ade	adbf
adedghi	adbcf
abcdedghi	adbcgf
	ade
	adedbcghi
	abcdedghi

3.2 假设检验及结果分析

首先对 9 个事件的关系熵进行双样本多变量 Hotelling T^2 假设检验,判断模式漂移的发生。假设设定的两窗口中的特征值服从相同分布, P 值在整个事件流上的变化曲线如图 5 所示。

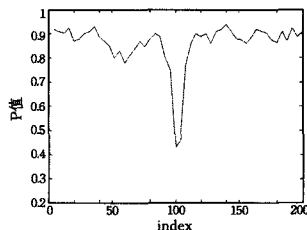


图 5 事件整体关系熵多变量 Hotelling T^2 假设检验结果

由图 5 可判断事件流上存在模式漂移,且能够大致定位到在第 100 个左右的子流上发生了模式漂移。然后在事件流中的 7 个事件(开始和结束除外)的关系熵进行双样本单变量 KS 假设检验,分别假设两窗口中的特征值服从相同分布,则得到的 P 值变化曲线如图 6 所示(由于篇幅限制,只列出了 b, d, g, h 这 4 个事件的 P 值变化)。

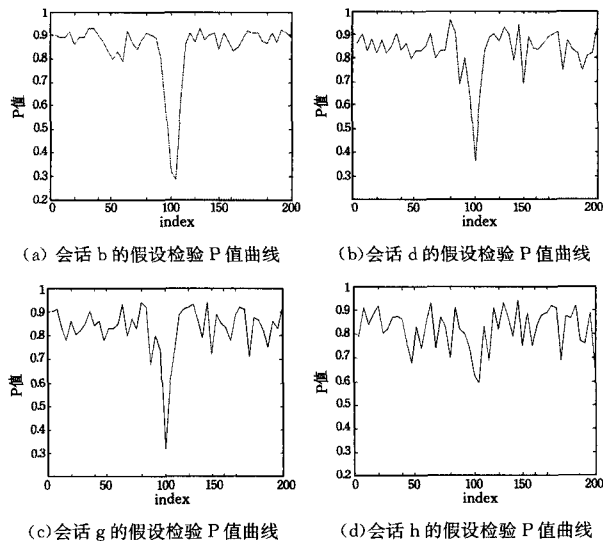


图 6 4 个事件关系熵的 KS 假设检验结果

图6中(a),(b),(c),(d)分别表示 b, d, g, h 4个事件关系熵假设检验的 P 值变化曲线。前3种事件的 P 值在模式漂移处的变化坡度较大,且降到一个较低的位置,可知这3种事件受模式漂移的影响较大,而图中第4种事件则受模式漂移影响较小。将这 b, d, g 3种事件两两组合,形成 $(b, d), (b, g), (d, g)$ 3个事件对,然后对每个事件对的 J 量值的分布进行双样本单变量MW假设检验,假设分布相同,以对 (d, g) 的分析为例,结果如图7所示。

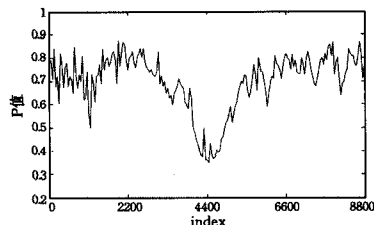


图7 (d, g) 事件对 J 量值MW假设检验的结果

从图7中可知, (d, g) 的假设检验 P 值大概在4400的位置降低到一个较低的水平,并出现低谷,表明事件 d 和事件 g 之间的关系受模式漂移影响较大。考虑到进行模拟数据准备时的设定,模式漂移之后,用户购物的行为习惯发生变化,逐渐从先搜索浏览商品,再登录网站进行购买转变为先登录网站,然后搜索浏览商品并进行购买,即在事件 d 与事件 g 之间往往需要事件 b 和事件 c 的发生,导致两者关系发生较大变化。这表明,本文的实验结果与实际是相符的。

3.3 结论分析

传统的面向分类的概念漂移检测由于对分类标签依赖度大、不能有效提取特征,对事件关联模式漂移是没有检测能力的,本文方法能成功检测模式漂移的发生,并能定位到模式漂移发生的大致位置,同时在最后可以得出受模式漂移影响最大的几个事件对。这些事件间的关系在模式漂移之后发生了较大的变化,可用于学习系统进行购物关联模式图的更新和改进,以适应模式漂移的变化。实验表明,本文的检测方法是准确可行的。

结束语 针对网络关联模式挖掘中面临的模式漂移问题,提出了一种基于关系熵和 J 量值的关联模式漂移检测方法。通过对网络事件及其相互关系的研究,定义了4种特征属性:关系量、关系熵、窗口数和 J 量值,将事件流转化成特征数据流,并对基于双滑动窗口分析获取的特征分布进行Hotelling T^2 假设检验和模式漂移的判定与定位,单个事件的KS假设检验和事件对的MW假设检验分别能够给出受模式漂移影响最大的事件集和事件对集,从而为关联模式的调整和更新提供支持。但如何具体指导关联模式库的更新从而适应模式漂移,本文还未作深入研究,这是下一步需要研究和完善的方向。

参考文献

[1] Wen Yi-min, Qiang Bao-hua, Fan Zhi-gang. A survey of the classification of data streams with concept drift [J]. CAAI Transac-

tions on Intelligent Systems, 2013, 8(2): 95-104 (in Chinese)

文益民, 强保华, 范志刚. 模式漂移数据流分类研究综述[J]. 智能系统学报, 2013, 8(2): 95-104

- [2] Dongre P B, Malik L G. A review on real time data stream classification and adapting to various concept drift scenarios[C]// 2014 IEEE International Advance Computing Conference, 2014: 533-537
- [3] Widmer G, Kubat M. Effective learning in dynamic environments by explicit context tracking [C]// Proceedings of the Sixth European Conference on Machine Learning, Vienna, Austria, 1993: 69-101
- [4] Last M. Online classification of nonstationary data stream [J]. Intelligent Data Analysis, 2002, 6(2): 1-16
- [5] Luo Xiu, Wang Da-ling, Feng Shi, et al. An algorithm for classifying data stream with recurrent concept drift [J]. Journal of Computer Research and Development, 2009, 46 (Suppl.): 400-405 (in Chinese)
- 罗秀, 王大玲, 冯时, 等. 一种面向周期性模式漂移的数据流分类算法[J]. 计算机研究与发展, 2009, 46 (Suppl.): 400-405
- [6] Katakis I, Tsoumakas G, Vlahavas I. Tracking recurring contexts using ensemble classifiers: an application to email filtering [J]. Knowledge and Information Systems, 2010, 22 (3): 371-391
- [7] Li Nan, Guo Gong-de, Chen Li-fei. Concept drift detection method with limited amount of labeled data [J]. Journal of Computer Applications, 2012, 8(1): 2176-2185 (in Chinese)
- 李南, 郭躬德, 陈黎飞. 基于少量类标签的模式漂移检测算法[J]. 计算机应用, 2012, 8(1): 2176-2185
- [8] Alippi C, Boracchi G, Rocheri M. An effective just-in-time adaptive classifier for gradual concept drifts [C]// Proceedings of the 2011 International Joint Conference on Neural Networks, San Jose, USA, 2011: 1675-1681
- [9] Carmona J, Gavalda R. Online techniques for dealing with concept drift in process mining [C]// Proc. Int. Conf. IDA, 2012: 90-102
- [10] Du Fei, Zhang Li-qun, Ni Guang-yun. A algorithm for detecting concept drift based on context in process mining [C]// 2013 Fourth International Conference on Digital Manufacturing & Automation, 2013: 5-8
- [11] J R P, Wil M P, et al. Dealing with concept drifts in process mining [J]. IEEE Transaction on Neural Networks and Learning Systems, 2014, 25(1): 154-171
- [12] Wei Hao, Chen Xing-yuan, Wang Chao, et al. Network behavior detection based on data stream analysis [J]. Application Research of Computers, 2013, 12(30): 3800-3803 (in Chinese)
- 魏浩, 陈性元, 王超, 等. 一种基于数据流分析的网络行为检测[J]. 计算机应用研究, 2013, 12(30): 3800-3803
- [13] Sheskin D. Handbook of parametric and nonparametric statistical procedures [M]. London, U. K: Chapman & Hall/CRC, 2004
- [14] Smyth P, Goodman R M. Rule induction using information Theory [M]. Washington, D C, USA: AAAS Press, 1991: 159-176