计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20140275
52(5):1071-1079,2015

基于并行约简的概念漂移探测

邓大勇[1] 徐小玉[1] 黄厚宽[2]
[1](浙江师范大学数理与信息工程学院 浙江金华 321004)
[2](北京交通大学计算机与信息技术学院 北京 100044)
(dayongd@163.com)

Concept Drifting Detection for Categorical Evolving Data Based on Parallel Reducts

Deng Dayong[1], Xu Xiaoyu[1], and Huang Houkuan[2]
[1](*College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinhua, Zhejiang* 321004)
[2](*School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044)

Abstract Data stream mining is one of the hot topics of data mining and concept drifting detection is one of its research directions. There have been many methods to detect concept drifting, but there are some drawbacks in current methods to detect concept drifting, such as no reducing redundant attributes integrally in sliding windows, and detecting concept drifting according to outer properties, etc. Based on the basic principles of rough sets and F-rough sets, the sliding windows in a data stream are regarded as decision subsystems, and the attribute significance of conditional attributes is used to detect concept drifting. This new method is divided into two steps: the redundant attributes in a streaming data are reduced through parallel reducts at first, then the concept drifting is detected according to the change of attribute significance. Different from other existing methods, the inner properties of data stream are used to detect concept drifting. Experiments show that this method is valid to reduce redundant attributes integrally and detect concept drifting, and that the attribute significance based on the mutual information is more effective than the attribute significance based on the positive region when they are used to detect concept drifting. For data stream mining, this paper provides a new method to detect concept drifting. For rough set theory, this paper offers a new application area.

Key words data streams; concept drift; rough sets; F-rough sets; parallel reducts

摘 要 数据流挖掘是当前数据挖掘研究的一个热点,概念漂移检测是数据流挖掘的一个重要研究方向.虽然有不少概念漂移的探测方法,但是它们都有一些共同的缺陷:没有整体上删除冗余属性以及利用外部属性去探测概念漂移(比如利用对外部数据的分类准确率)等.利用粗糙集和F-粗糙集的基本原理和基本方法,把数据流中的滑动窗口当成决策子表簇,提出了一种对数据流进行并行约简、整体删除冗余属性的方法,并运用并行约简后数据流决策子表簇中属性重要性的变化探测概念漂移现象.与传统的方法不同,新方法利用数据的内部特性对概念漂移进行探测.实验结果显示,该方法能够有效地整体删除冗余属性、探测概念漂移现象,并且基于互信息的属性重要性在概念漂移探测效果方面比基于正区域的属性重要性要好些.

关键词 数据流;概念漂移;粗糙集;F-粗糙集;并行约简

中图法分类号 TP18

收稿日期:2014-03-30;**修回日期**:2014-10-13

基金项目:国家自然科学基金项目(61473030);浙江省自然科学基金项目(Y15F020044);浙江省自然科学青年基金项目(Q13F020006);浙江师范大学计算机软件与理论省级重中之重学科开放基金项目(ZSDZZZZXK27)

现实中的数据往往随着时间的变化而变化,例如证券交易数据、微博、视频、传感器数据等这种类型的数据称为数据流[1].数据流具有按照时间顺序排列、快速变化、海量甚至无限并且可能出现概念漂移现象等特征[2-3].数据流挖掘是当前数据挖掘研究的一个热点,概念漂移的探测以及数据流的分类是当前数据流挖掘的主要研究方向.

滑动窗口技术是探测概念漂移常用的技术[1],窗口或者是固定大小或者大小可变化.数据流的分类策略主要有2类:1)单一分类器;2)集成分类器.单一分类器通过对初始模型进行增量式更新,适应概念漂移[4-6],单一分类器更新速度低,准确率较差;集成分类器将数据实例划分成不同的数据块,每个数据块训练一个基础分类器,多个基础分类器构成集成分类器,随着时间的变化,淘汰更新一部分分类准确率低的基础分类器,从而使集成分类器的分类准确率不断更新和提高[3,7-11],因此集成分类器更能适应概念漂移,分类准确率也较高.

无论是单一分类器还是集成分类器都较少考虑不同的属性对分类的不同作用,有些属性是冗余的,对分类不起作用,可以删除,也较少考虑删除掉冗余属性后再检测概念漂移.文献[12]对决策树进行剪枝;文献[13]提出了一种解决概念漂移的增量式算法并进行剪枝;文献[14]对商业数据流进行特征选择.但是它们都是对单个的数据块或单棵决策树进行冗余属性删除,没有从整体上考虑删除冗余属性问题.

粗糙集理论[15-17]是一种处理不精确、不完全、含糊数据的有效数学工具,是数据挖掘和分类的重要方法.属性约简是粗糙集理论的一个重要应用.传统的粗糙集理论不太适合研究海量的、动态变化的数据,也不太适合研究数据流;F-粗糙集方法[18-19]将粗糙理论从单个信息表或决策表推广到多个,并行约简是与F-粗糙集相对应的属性约简理论和方法.并行约简和F-粗糙集比较适合研究动态变化的数据,能够研究数据流和概念漂移.

利用粗糙集理论研究数据流和概念漂移比较少见.文献[20]利用粗糙集的上、下近似检测概念漂移,并利用粗糙率度量概念漂移;文献[21]运用F-粗糙集方法提出了概念漂移的8个度量指标.

本文首先利用F-粗糙集的并行约简理论,将数据流的各个滑动窗口(子决策表)中对分类不起作用的冗余属性整体删除,然后运用各个子表(滑动窗口)中属性重要性的变化探测概念漂移.传统方法主要依靠分类准确率的变化,利用外部特性进行比较,探测概念漂移现象.本文方法与传统的概念漂移探测方法不同,利用数据的内部特性——并行约简后,属性发生重要性的变化——探测概念漂移现象.

1 基础知识

本节简单介绍粗糙集[15-17]、F-粗糙集和并行约简[18-19]的基本知识.

1.1 粗糙集

$IS=(U,A)$是一个信息系统,其中U是论域,A是论域U上的条件属性集.对于每个属性$a\in A$都对应着一个函数$a:U\to V_a$,V_a称为属性a的值域,U中每个元素称为个体、对象或行.

对于每一个属性子集$B\subseteq A$和任何个体$x\in U$都对应着一个如下的信息函数:

$$Inf_B(x)=\{(a,a(x)):a\in B\}.$$

B-不分明关系(或称为不可区分关系)定义为

$$IND(B)=\{(x,y):Inf_B(x)=Inf_B(y)\}.$$

任何满足关系$IND(B)$的2个元素x,y都不能由属性子集B区分,$[x]_B$表示由x引导的$IND(B)$等价类.

对于信息系统$IS=(U,A)$、属性子集$B\subseteq A$和论域子集$X\subseteq U$:

$$\underline{B}(X)=\underline{B}(IS,X)=\{x\in U:[x]_B\subseteq X\},$$

$$\overline{B}(X)=\overline{B}(IS,X)=\{x\in U:[x]_B\cap X\neq\varnothing\}$$

分别称为B-下近似和B-上近似.B-下近似也称为正区域,记为$POS_B(X)$.序偶$(\underline{B}(X),\overline{B}(X))$称为粗糙集.

在决策系统$DS=(U,A,d)$中,$\{d\}\cap A=\varnothing$,决策属性d将论域U划分为块,$U/\{d\}=\{Y_1,Y_2,\cdots,Y_p\}$,其中$Y_i(i=1,2,\cdots,p)$是等价类.决策系统$DS=(U,A,d)$的正区域定义为

$$POS_A(d)=\bigcup_{Y_i\in U/\{d\}}POS_A(Y_i).$$

有时决策系统$DS=(U,A,d)$的正区域$POS_A(d)$也记为$POS_A(DS,d)$或$POS(DS,A,d)$.

1.2 F-粗糙集

在粗糙集的所有模型中,三支决策粗糙集模型[22-24]在一个决策表中通过调整参数表示不同的人对同一件事物的不同看法;F-粗糙集[18-19]和其他任何粗糙集模型不同,它是关于信息系统簇或者决策系统簇的粗糙集模型,这个粗糙集模型比较适合并行计算,也适合研究事物的动态变化.

众所周知,一个概念在不同情形下的意义是不一样的,比如我们说一个人是好人,这个"好人"的概念在不同的情形下,由不同的人说出来的意义是不一样的.下面用 $FIS=\{IS_i\}(i=1,2,\cdots,n)$ 表示与决策系统簇 F 相对应的信息系统簇,其中 $IS_i=(U_i,A)$,而 $DT_i=(U_i,A,d)$.

定义 1. 假设 X 是一个概念,N 是一种情形,$X|N$ 表示在情形 N 下的概念 X.在一个信息系统簇 $FIS=\{IS_1,IS_2,\cdots,IS_n\}$ 中,$X|IS_i=X\cap IS_i$,$X|FIS=\{X|IS_1,X|IS_2,\cdots,X|IS_n\}$.如果不引起混淆,$X|N$ 可以缩写为 X.

假设 X 是信息系统簇 FIS 中的一个概念,那么 X 关于属性子集 $B\subseteq A$ 的上近似、下近似、边界线区域、负区域的定义如下:

$$\overline{B}(FIS,X)=\{\overline{B}(IS_i,X):IS_i\in FIS\}=\{\{x\in U_i:[x]_B\cap X\neq\varnothing,X\subseteq IS_i\}\};$$

$$\underline{B}(FIS,X)=\{\underline{B}(IS_i,X):IS_i\in FIS\}=\{\{x\in U_i:[x]_B\subseteq X,X\subseteq IS_i\}\};$$

$$BND(FIS,X)=\{BND(IS_i,X):IS_i\in FIS\}=\{\overline{B}(IS_i,X)-\underline{B}(IS_i,X):X\subseteq IS_i\};$$

$$NEG(FIS,X)=\{NEG(IS_i,X):IS_i\in FIS\}=\{U_i-\overline{B}(IS_i,X):X\subseteq IS_i\}.$$

概念 X 关于信息系统簇 FIS 的上下近似、边界线区域、负区域分别是 FIS 中的信息子系统关于概念 X 的上下近似、边界线区域、负区域组成的集合.序偶$(\underline{B}(FIS,X),\overline{B}(FIS,X))$称为 F-粗糙集.如果$\underline{B}(FIS,X)=\overline{B}(FIS,X)$则称序偶$(\underline{B}(FIS,X),\overline{B}(FIS,X))$是精确的.

概念 X 关于信息系统簇 FIS 的下近似通常也称为概念 X 关于信息系统簇 FIS 的正区域.对于元素 $x\in U$,我们不能简单地称它是属于正区域、负区域或边界线区域,因为 $x\in U$ 对于不同的 U_i(或 IS_i $(i=1,2,\cdots,n)$)所在的区域不同,它可能既在正区域,也在负区域或边界线区域.我们只能说,对于元素 $x\in U$,它在某个信息子系统 IS_i 中属于正区域、负区域或边界线区域.

F-粗糙集模型的基本概念被定义后,Pawlak 粗糙集和其他粗糙集几乎所有的概念、所有的知识都可以迁移到 F-粗糙集模型中,这就是说,F-粗糙集模型具有非常好的适应性和可扩展性.

定义 2. 设 $DS=(U,A,d)$ 是一个决策系统,$P(DS)$是 DS 的幂集,$F\subseteq P(DS)$,则 F-正区域定义如下:

$$POS(F,A,d)=\{POS(DT,A,d):DT\in F\}.$$

定义 3. 设 $DS=(U,A,d)$ 是一个决策系统,$P(DS)$是 DS 的幂集,$F\subseteq P(DS)$,则 $B\subseteq A$ 称为 F-并行约简,当且仅当 $B\subseteq A$ 满足下面条件:

1) $POS(F,A,d)=POS(F,B,d)$;

2) 对任意 $S\subset B$,都有:

$$POS(F,S,d)\neq POS(F,A,d).$$

2 并行约简方法对概念漂移的探测

在 F-粗糙集[18-19]中决策子系统簇 F 中的元素可以是大数据中的一部分,也可以是数据流中的一部分或一个滑动窗口.本文假设决策子系统簇 F 中的元素是数据流中的一部分,每一个子表可以看作一个滑动窗口.在探测概念漂移之前,我们先用并行约简算法整体删除对分类不起作用的冗余属性,以减少计算量,并探测真正使概念发生漂移的属性之变化.

2.1 并行约简算法

属性重要性是粗糙集中一个重要的概念,多用于求取属性约简的启发式信息.在求取约简时,首先根据属性重要性求取核属性,然后根据各个属性的属性重要性的不同选取那些最重要的属性添加到约简中,并删除那些对约简不起作用的冗余属性.粗糙集中有多种度量属性重要性的指标,其中最常用的是基于正区域的属性重要性,它的定义如下:

定义 4[15-17]. 在决策系统 $DS=(U,A,d)$中,称决策属性 d 以程度 $h(0\leqslant h\leqslant 1)$依赖条件属性集 A,其中:

$$h=\gamma(A,d)=\frac{POS_A(d)}{|U|},$$

符号$|\cdot|$表示集合的势.

定义 5[15-17]. 在决策系统 $DS=(U,A,d)$中,条件属性 $a\in A$ 的属性重要度定义为

$$\sigma(a)=\frac{\gamma(A,d)-\gamma(A-\{a\},d)}{\gamma(A,d)}=1-\frac{\gamma(A-\{a\},d)}{\gamma(A,d)}.$$

在 F-粗糙集中依赖度与属性重要度的定义与传统粗糙集中的定义类似,它们的定义如下:

定义 6[18-19]. 给定一个决策子系统簇 F,$DT_i=(U_i,A,d)\in F$,$i=1,2,\cdots,n$,定义 F 中 d 依赖于 A 的程度为

$$h=\gamma(F,A,d)=\frac{\sum_{DT\in F}|POS(DT,A,d)|}{\sum_{i=1}^{n}|U_i|}.$$

定义 7[18-19]. 给定一个决策子系统簇 $F, DT_i=(U_i,A,d)\in F, i=1,2,\cdots,n$, 定义 F 中属性 $a\in B$ 或 $a\in A-B$ 相对于 B 的重要度为

$$\sigma(B,a)=\frac{\gamma(F,B,d)-\gamma(F,B-\{a\},d)}{\gamma(F,B,d)}=1-\frac{\gamma(F,B-\{a\},d)}{\gamma(F,B,d)}$$

或

$$\sigma'(B,a)=\frac{\gamma(F,B\cup\{a\},d)-\gamma(F,B,d)}{\gamma(F,B,d)}=\frac{\gamma(F,B\cup\{a\},d)}{\gamma(F,B,d)}-1.$$

其中 $\gamma(F,B,d)$ 可能为 0, 这是因为决策子系统簇 F 中的正域可能为空, 因此在实际情况中我们可以重新定义决策子系统簇中的属性重要度.

定义 8[18-19]. 给定一个决策子系统簇 $F, DT_i=(U_i,A,d)\in F, i=1,2,\cdots,n$, 定义 F 中属性 $a\in B$ 或 $a\in A-B$ 相对于 B 的重要度为

$$\sigma(B,a)=\gamma(F,B,d)-\gamma(F,B-\{a\},d)$$

或

$$\sigma'(B,a)=\gamma(F,B\cup\{a\},d)-\gamma(F,B,d).$$

运用属性重要度可以比较容易地求出并行约简, 见算法 1:

算法 1. 基于属性重要性的并行约简算法[18-19].

输入: $F\subseteq P(DS)$;

输出: F 的一个并行约简 B.

步骤 1. $B=\varnothing$;

步骤 2. 对于任意 $a\in A$, 如果 $\sigma(A,a)>0$, 那么 $B=B\cup\{a\}$; /* B 是并行约简的属性核 */

步骤 3. $E=A-B$;

步骤 4. 重复进行如下步骤, 直到 $E=\varnothing$:

步骤 4.1. 对任意 $a\in E$, 计算 $\sigma'(B,a)$;

/* $\sigma'(B,a)=\gamma(F,B\cup\{a\},d)-\gamma(F,B,d)$ */

步骤 4.2. 对任意 $a\in E$, 如果 $\sigma'(B,a)=0$, 那么 $E=E-\{a\}$;

步骤 4.3. 选择 F 中属性重要度非零且最大的元素 $a\in E, B=B\cup\{a\}, E=E-\{a\}$; /* 添加属性集 E 中属性重要度非零且最大的属性到 B 中 */

步骤 5. 输出并行约简 B.

假设采用文献[25]中的算法求属性重要度, 算法的复杂度为 $O(|B||U|\mathrm{lb}|U|)$, 其中 $|U|$ 代表决策表中数据的个数, $|B|$ 代表条件属性的个数. 并行约简的时间复杂度[18-19]为 $O(m^3\sum_{U_i\in F}|U_i|\sum_{U_i\in F}\mathrm{lb}|U_i|)$, 其中 m 表示决策子表簇中条件属性的个数.

并行约简将决策子表簇作为一个整体考虑, 删除了决策子表簇中对分类不起作用的冗余属性, 使得在概念漂移的探测和分类的时候减少了计算量, 并将注意力真正集中到对分类起关键作用的属性集合上. 如果对决策子表各自进行约简, 而不是并行约简, 则每个子表中保留的条件属性不完全相同, 在探测概念漂移的时候, 由于缺乏同样的属性和同样的标准, 得到结论的可理解性与可靠性就会大打折扣.

2.2 概念漂移探测

通过并行约简删除了数据流中对分类不起作用的冗余属性. 受文献[18-19]中属性重要性矩阵的启发, 我们建立数据流约简后的属性重要性矩阵, 用于描述在不同的决策子表(滑动窗口)中并行约简中的每个属性对分类的贡献, 它的定义如下:

定义 9. $DS=(U,A,d)$ 是一个数据流决策系统, $P(DS)$ 是 DS 的幂集, $F\subseteq P(DS)$ 是数据流 $DS=(U,A,d)$ 的若干个滑动窗口的集合, $B\subseteq A$ 是 F 的并行约简, 并行约简 $B\subseteq A$ 关于 F 的属性重要度矩阵定义为

$$\boldsymbol{M}(B,F)=\begin{bmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m}\\ \vdots & \vdots & & \vdots\\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nm}\end{bmatrix},$$

其中 $\sigma_{ij}=\gamma_i(B,d)-\gamma(B-\{a_j\},d)$, $a_j\in B, DT_i=(U_i,A,d)\in F, \gamma_i(B,d)=\dfrac{|POS(DT_i,B,d)|}{|U_i|}$, n 代表 F 中子决策子表的个数, m 代表 B 中条件属性的个数.

并行约简 B 的属性重要性矩阵与文献[18-19]中属性重要性矩阵最大的区别在于: 前者是计算并行约简中各个属性的重要性从而形成属性重要性矩阵, 它选择了那些对分类起作用的属性, 并删除了那些对分类不起作用的属性; 而后者是约简前的属性集合中的属性重要性形成的属性重要性矩阵. 当然, 在具体计算的过程中, 我们不一定需要形成矩阵的形式, 但这种矩阵形式可读性强、可理解性强.

根据属性重要性矩阵的定义, 很容易证明属性重要性矩阵的相关属性.

定理 1. 数据流决策子表簇 $F\subseteq P(DS)$ 中, $B\subseteq A$ 为并行约简, 属性重要性矩阵 $\boldsymbol{M}(B,F)$ 中的元素与属性重要性矩阵 $\boldsymbol{M}(A,F)$ 中相应的元素有 2 种关系:

1) 若属性 $b\in B\subseteq A$ 为并行约简的核属性, 则 b 在 $\boldsymbol{M}(B,F)$ 中对应的元素值小于等于 b 在 $\boldsymbol{M}(A,F)$ 对应的元素值.

2) 若属性 $b\in B\subseteq A$ 为并行约简的非核属性,则 b 在 $\boldsymbol{M}(B,F)$ 中对应的元素值大于等于 b 在 $\boldsymbol{M}(A,F)$ 对应的元素值.

推论 1. 数据流决策子表簇 $F\subseteq P(DS)$ 中,$B\subseteq A$ 为并行约简,属性重要性矩阵 $\boldsymbol{M}(B,F)$ 中非零元素的个数大于等于 $\boldsymbol{M}(A,F)$ 中非零元素的个数.

度量概念漂移的指标有多种,大部分关于数据流分类的文献都将分类器的分类准确率作为概念漂移的度量,这种方法存在一定的缺陷.在构成分类器的过程中往往进行了特征选择或剪枝,各部分数据都是独立训练分类器的,特征选择或剪枝也是独立进行的,这就使得分类器的分类准确率缺乏可比性,这是因为即使是同一个数据集,特征选择或剪枝不同,训练出来的分类器对同一个新数据集的分类准确率也不同.

运用粗糙集理论对概念漂移进行度量的指标[20-21]往往依赖于上、下近似,这种度量方法对固定窗口影响不大,但如果是滑动窗口的大小不是固定的而是大小可变,这样的指标就显得不够灵活甚至完全不能度量.下面我们运用属性重要性的变化对概念漂移进行度量,它独立于上、下近似的变化,也独立于滑动窗口的大小.它的定义如下:

定义 10. 数据流决策子表簇 $F\subseteq P(DS)$ 中,$B\subseteq A$ 为并行约简,2 个滑动窗口 $DT_i,DT_k\in F$ 相对于属性 $b\in B,B\subseteq A$ 的概念漂移定义为

$$PRCD_b(DT_k,DT_i)=|\sigma_{kj}-\sigma_{ij}|,$$

其中,j 为属性 $b\in B,B\subseteq A$ 在 $\boldsymbol{M}(B,F)$ 中所对应的列.

定义 11. 数据流决策子表簇 $F\subseteq P(DS)$ 中,$B\subseteq A$ 为并行约简,$DT_i,DT_k\in F$,基于并行约简 $B\subseteq A$ 的概念漂移量定义为

$$PRCD_B(DT_k,DT_i)=\frac{1}{|B|}\sum_{j=1}^{|B|}|\sigma_{kj}-\sigma_{ij}|.$$

定理 2. 基于并行约简的属性重要性的概念漂移量 $PRCD_b(DT_k,DT_i)$,$PRCD_B(DT_k,DT_i)$ 对称、非负、满足三角不等式.

证明. 根据定义容易证明上述定理.

定理 3. 数据流决策子表簇 $F\subseteq P(DS)$ 中,$DT_i,DT_k\in F,B\subseteq A$ 为并行约简,则 $\boldsymbol{M}(B,F)$ 中相邻 2 行对应属性重要性变化的元素个数大于等于 $\boldsymbol{M}(A,F)$ 中相邻 2 行对应属性重要性变化的元素个数.

证明. 在 $\boldsymbol{M}(A,F)$ 中除了核属性的属性重要性非零外,其余元素都为 0,所以 $\boldsymbol{M}(A,F)$ 属性重要性的变化仅仅在核属性中,而 $\boldsymbol{M}(B,F)$ 中除了核属性外,还可能存在属性重要性非零的非核属性,所以 $\boldsymbol{M}(B,F)$ 中相邻 2 行对应属性重要性变化的元素个数大于等于 $\boldsymbol{M}(A,F)$ 中相邻 2 行对应属性重要性变化的元素个数. 证毕.

定理 1、定理 3 说明冗余属性的存在干扰了概念漂移的探测,删除了一些冗余属性后概念漂移更明显.下面我们利用基于并行约简的概念漂移量来探测概念漂移.

算法 2. 概念漂移检测算法.

输入:数据流 $F\subseteq P(DS)$,阈值 δ;

输出:数据流 $F\subseteq P(DS)$ 中是否发生概念漂移.

步骤 1. 调用算法 1,求出 F 的并行约简 $B\subseteq A$;

步骤 2. 计算约简后 F 中各个属性的重要性,形成属性重要性矩阵 $\boldsymbol{M}(B,F)$;

步骤 3. 计算相邻 2 行之间任意属性重要性的差异 $PRCD_b(DT_k,DT_i)$,并算出 $PRCD_B(DT_k,DT_i)$;

步骤 4. 概念漂移值 $PRCD_b(DT_k,DT_i)$,$PRCD_B(DT_k,DT_i)$ 与给定的阈值 δ 比较,判定是否发生了概念漂移.

阈值 $\delta\in[0,1)$ 的选取我们将在实验中进行说明.

算法 2 的主要时间消耗在步骤 1 求取并行约简和步骤 2 形成属性重要性矩阵上.步骤 1 的时间复杂度为 $O(m^3\sum_{U_i\in F}|U_i|\sum_{U_i\in F}\mathrm{lb}|U_i|)$,步骤 2 的时间复杂度为 $O(\sum_{U_i\in F}m|U_i|\mathrm{lb}|U_i|)$.

所以算法 2 的时间复杂度为

$$O(m^3\sum_{U_i\in F}|U_i|\sum_{U_i\in F}\mathrm{lb}|U_i|)+O(\sum_{U_i\in F}m|U_i|\mathrm{lb}|U_i|)=O(\max\{m^3\sum_{U_i\in F}|U_i|\sum_{U_i\in F}\mathrm{lb}|U_i|,\sum_{U_i\in F}m|U_i|\mathrm{lb}|U_i|\})=O(m^3\sum_{U_i\in F}|U_i|\sum_{U_i\in F}\mathrm{lb}|U_i|),$$

其中 m 表示决策子表簇中条件属性的个数.

例如设 $F=\{DT_1,DT_2\}$,如表 1、表 2 所示.a,b,c 是条件属性,d 是决策属性.

Table 1 Decision Subsystem DT_1

表 1 决策子系统 DT_1

U_1	a	b	c	d
x_1	0	0	1	1
x_2	1	1	0	1
x_3	0	1	0	0
x_4	1	1	0	1

Table 2 Decision Subsystem DT_2

表 2 决策子系统 DT_2

U_2	a	b	c	d
y_1	0	1	0	0
y_2	1	1	0	1
y_3	1	1	0	1
y_4	0	1	0	0
y_5	1	2	0	0
y_6	1	2	0	1

调用算法 1,很容易得到 F 的并行约简为 $B=\{a,b\}$,属性重要性矩阵 $\boldsymbol{M}(A,F)$ 与 $\boldsymbol{M}(B,F)$ 为

$$\boldsymbol{M}(A,F)=\begin{bmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13}\\ \sigma_{21} & \sigma_{22} & \sigma_{23}\end{bmatrix}=\begin{bmatrix}0.75 & 0.00 & 0.00\\ 0.67 & 0.33 & 0.00\end{bmatrix},$$

$$\boldsymbol{M}(B,F)=\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}=\begin{bmatrix}0.50 & 0.50\\ 0.67 & 0.33\end{bmatrix}.$$

DT_1 与 DT_2 之间的概念漂移为

$$PRCD_a(DT_2,DT_1)=|\,0.67-0.50\,|=0.17,$$

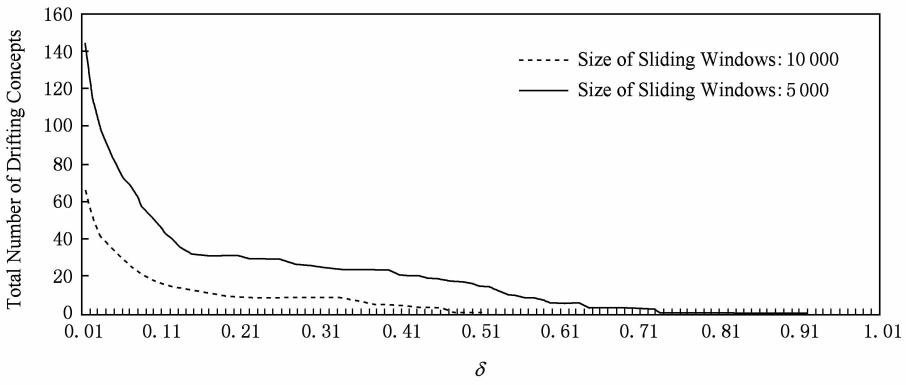
$$PRCD_b(DT_2,DT_1)=|\,0.33-0.50\,|=0.17,$$

$$PRCD_B(DT_2,DT_1)=\frac{1}{m}\sum_{j=1}^{m}|\,\sigma_{2,j}-\sigma_{1,j}\,|=\frac{1}{2}(|\,0.67-0.50\,|+|\,0.33-0.50\,|)=0.17.$$

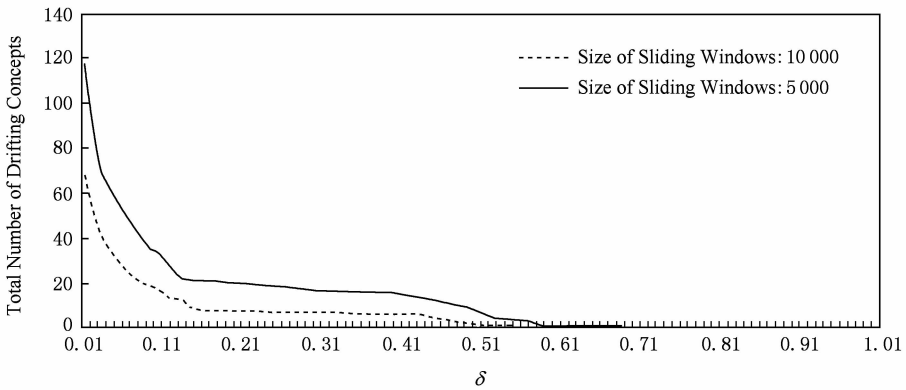
因为条件属性 c 对分类不起作用,是冗余的属性,删除之后对分类起作用的属性 a,b 的概念漂移就彰显出来了.如果取 $\delta=0.1$,那么相对于单个属性 a,b 具有概念漂移,相对于整个并行约简 B 也具有概念漂移;如果取 $\delta=0.2$,那么相对于单个属性 a,b 及相对于并行约简 B 都不具有概念漂移.实际的数据流中,滑动窗口一般情况下是多个,我们可以类似地求出 2 个相邻窗口之间的基于并行约简的概念漂移量.

3 实验结果

第 2 节我们主要讨论了在数据流下基于正区域的属性重要性矩阵的性质,并提出了基于属性重要性矩阵的概念漂移探测算法,粗糙集中属性重要性的度量方式有多种,基于并行约简的概念漂移探测算法并不依赖于某种属性重要性.本节的实验我们



(a) Mutual information



(b) Positive region

Fig. 1 The relation between the total number of drifting concepts based on the parallel reducts and the threshold δ.

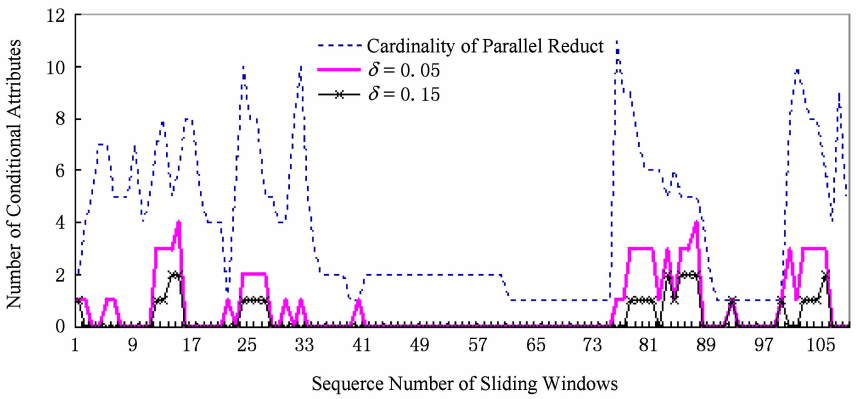
图 1 相对于并行约简的概念漂移总数与阈值 δ 之间的关系

比较了基于互信息的属性重要性矩阵[19]和基于正区域的属性重要性矩阵在探测概念漂移方面的异同,并探索概念漂移和阈值 δ 之间的联系.

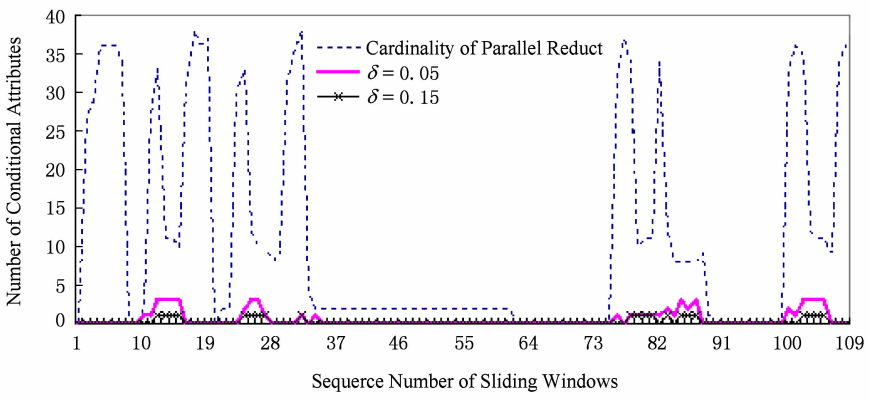
实验数据为 KDD-CUP99 网络入侵检测数据 10%的子集.该数据包含 494 021 条记录、42 个属性.实验中滑动窗口的大小为 1 000～30 000,步长为 1 000;阈值 δ 取值为 0.01～1,步长为 0.01,相邻滑动窗口之间有 10%的数据重复率.所有的实验结果类似,我们以滑动窗口为 5 000 和 10 000 为例进行说明.

图 1 显示滑动窗口分别为 5 000 和 10 000 时,相邻 2 个滑动窗口之间发生概念漂移与阈值 δ 之间的关系.当 $PRCD_B(DT_i,DT_{i+1})\geqslant\delta$ 时计 1 次概念漂移,其中 B 为相邻 2 窗口之间的并行约简.实验结果表明,当 $\delta\geqslant 0.15$ 时,无论是基于互信息的属性重要性(图 1(a))还是基于正区域的属性重要性(图 1(b))的概念漂移现象很少,甚至完全没有.

图 2 和图 3 分别是滑动窗口为 5 000 和 10 000 时,相邻 2 个滑动窗口之间基于互信息的属性重要性并行约简(图 2(a)与图 3(a))和基于正区域的属性重要性并行约简(图 2(b)与图 3(b))后单个属性重要性的变化,即当属性 b 为并行约简 B 中的属性,且 $PRCD_b(DT_i,DT_{i+1})\geqslant\delta$ 时计数 1 次.图 2 和图 3 都用并行约简包含的属性个数与发生概念漂移的属性个数作为参照.实验结果显示,当 $\delta\geqslant 0.15$ 时,单个属性的概念漂移现象很少,甚至完全没有.用基于互信息的属性重要性作为指标探测概念漂移现象比基于正区域的属性重要性探测概念漂移现象效果更明显.



(a) Mutual information



(b) Positive region

The size of sliding windows is 5 000 after parallel reduct.

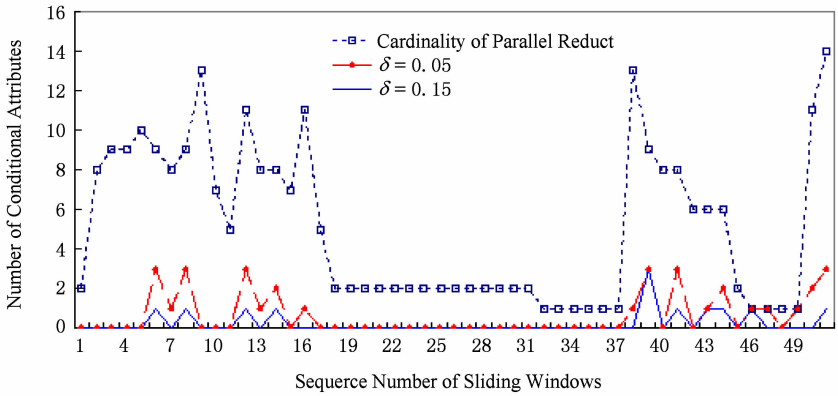
Fig. 2 The number of drifting concepts based on the single attribute.

图 2 相对于并行约简中单个属性的概念漂移情形

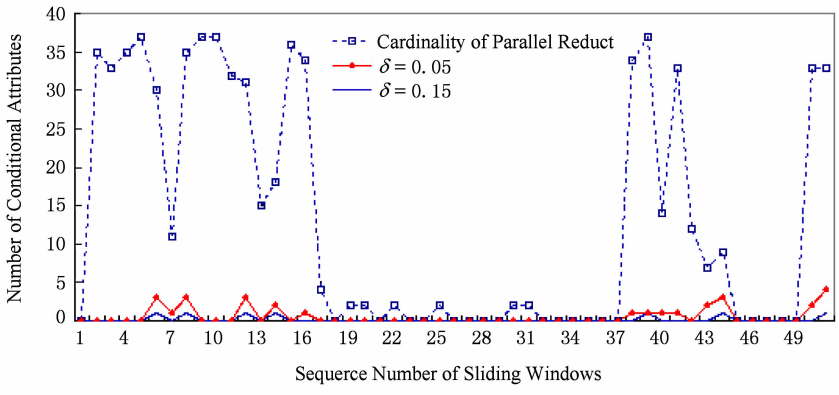
根据 2 种并行约简属性重要性矩阵的特性和实验结果,一般情况下阈值 $\delta\leqslant 0.15$,最好是在 0.01～0.1 之间,而且用基于互信息的属性重要性探测概念漂移现象比用基于正区域的属性重要性效果要好些.

4 结论及展望

传统的概念漂移探测方法主要利用分类准确率的变化对概念漂移现象进行探测.本文提出了一种



(a) Mutual information



(b) Positive region

The size of sliding windows is 10 000 after parallel reduct.

Fig. 3 The number of drifting concepts based on the single attribute.

图 3 相对于并行约简中单个属性的概念漂移情形

基于并行约简的概念漂移探测方法. 该方法利用数据的内部特性——并行约简后的属性重要性变化——探测概念漂移现象. 对滑动窗口进行并行约简可以整体删除冗余属性,彰显对分类起作用的属性重要性的变化;利用属性重要性的变化探测概念漂移,可以不必利用分类器对外部数据的分类能力探测概念漂移;此外,对相邻滑动窗口进行并行约简,使得探测概念漂移的时候比较标准得到了统一.

下一步的研究工作是运用并行约简构建分类器,与传统的概念漂移方法进行深入地分析比较.

参 考 文 献

[1] Babcock B, Babu S, Dater M, et al. Models and issues in data stream systems [C] //Proc of the 21st ACM SIGACT-SIGMOD-SIGART Symp on Principles Database Systems. New York: ACM, 2002: 1-30

[2] Wang Tao, Li Zhoujun, Yan Yuejin, et al. A survey of classification of data streams [J]. Journal of Computer Research and Development, 2007, 44(11): 1809-1815 (in Chinese)
(王涛, 李舟军, 颜跃进, 等. 数据流挖掘分类技术综述[J]. 计算机研究与发展, 2007, 44(11): 1809-1815)

[3] Xu Wenhua, Qin Zheng, Chang Yang. Semi-supervised learning based ensemble classifier for stream data [J]. Pattern Recognition and Artificial Intelligence, 2012, 25(2): 292-299 (in Chinese)
(徐文华, 覃征, 常扬. 基于半监督学习的数据流集成分类算法[J]. 模式识别与人工智能, 2012, 25(2): 292-299)

[4] Gehrke J, Ganti V, Ramakrishnan R, et al. Boat-optimistic decision tree construction [C] //Proc of the 18th ACM SIGMOD Int Conf on Management of Data. New York: ACM, 1999: 169-180

[5] Domingos P, Hulten G. Mining high-speed data stream [C] //Proc of the 6th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2000: 71-80

[6] Hulten G, Spencer L, Domingos P. Mining time-changing data sreams [C] //Proc of the 7th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2001: 97-106

[7] Sun Yue, Mao Guojun, Liu Xu, et al. Mining Concept drifts from data streams based on multi-classifiers [J]. Acta Automatica Sinica, 2008, 34(1): 93-96 (in Chinese)

(孙岳,毛国君,刘旭,等. 基于多分类器的数据流中的概念漂移挖掘[J]. 自动化学报,2008,34(1):93-96)

[8] Wan Haixun, Fan Wei, Yu P S, et al. Mining concept-drifting data streams using ensemble classifiers [C] //Proc of the 9th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2003: 226-235

[9] Scholz M, Klinkenberg R. An ensemble classifier for drifting concepts [C] //Proc of the 2nd Int Workshop on Knowledge Discovery in Data Streams. Berlin: Springer, 2005: 53-64

[10] Aggarwal C C, Han J, Wang Jianyong, et al. A framework for on-demand classification of evolving data streams [J]. IEEE Trans on Knowledge and Data Engineering, 2006, 18(5): 577-589

[11] Bifet A, Holmes G, Pfahringer B, et al. New ensemble methods for evolving data streams [C] //Proc of the 15th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2009: 139-148

[12] Jin R, Agrawal G. Efficient decision tree construction on streaming data [C] //Proc of the 9th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2003: 571-576

[13] Xin Yi, Guo Gongde, Chen Lifei, et al. IKnnM-DHecoc: A method for handing the problem of concept drift [J]. Journal of Computer Research and Development, 2011, 48(4): 592-601 (in Chinese)
(辛轶,郭躬德,陈黎飞,等. IKnnM-DHecoc:一种解决概念漂移问题的方法[J]. 计算机研究与发展,2011,48(4):592-601)

[14] Ju Chunhua, Shuai Zhaoqian, Feng Yi. Granular computing based concept drift features selection for business data streams [J]. Journal of Nanjing University: Natural Sciences, 2011, 47(4): 391-397 (in Chinese)
(琚春华,帅朝谦,封毅. 基于粒计算的商业数据流概念漂移特征选择[J]. 南京大学学报:自然科学版,2011,47(4):391-397)

[15] Pawlak Z. Rough sets [J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341-356

[16] Pawlak Z. Rough Sets—Theoretical Aspect of Reasoning about Data [M]. Amsterdam, Netrerlands: Kluwer Academic Publishers, 1991

[17] Wang Guoyin. Rough Set Theory and Knowledge Acquisition [M]. Xi'an: Xi'an Jiaotong University Press, 2001 (in Chinese)
(王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社,2001)

[18] Deng Dayong, Chen Lin. Parallel Reducts and F-rough Sets [G] //Cloud Model and Granular Computing. Beijing: Science Press, 2012: 210-228 (in Chinese)
(邓大勇,陈林. 并行约简与 F-粗糙集[G] //云模型与粒计算. 北京:科学出版社,2012:210-228)

[19] Chen Lin. Parallel reducts and decision in various levels of granularity [D]. Jinhua, Zhejiang: Zhejiang Normal University, 2013 (in Chinese)
(陈林. 粗糙集中不同粒度层次下的并行约简及决策[D]. 浙江金华:浙江师范大学,2013)

[20] Cao Fuyuan, Huang Zhexue. A concept-drfting detection algorithm for categorical evolving data [G] //LNAI 7819: Proc of the 17th Pacific-Asia Conf on Knowledge Discovery and Data Mining. Berlin: Springer, 2013: 485-496

[21] Deng Dayong, Pei Minghua, Huang Houkuan. The F-rough sets approaches to the measures of concept drift [J]. Journal of Zhejiang Normal University: Natural Sciences, 2013, 36(3): 303-308 (in Chinese)
(邓大勇,裴明华,黄厚宽. F-粗糙集方法对概念漂移的度量[J]. 浙江师范大学学报:自然科学版,2013,36(3):303-308)

[22] Yao Yiyu. Three-way decision with probabilistic rough sets [J]. Information Sciences, 2010, 180(3): 341-353

[23] Yao Yiyu. The superiority of three-way decisions in probabilistic rough set models [J]. Information Sciences, 2011, 181(6): 1080-1096

[24] Liu Dun, Yao Yiyu, Li Tianrui. Three-way decision-theoretic rough sets [J]. Computer Science, 2011, 38(1): 245-250 (in Chinese)
(刘盾,姚一豫,李天瑞. 三支决策粗糙集[J]. 计算机科学,2011,38(1):245-250)

[25] Liu Shaohui, Sheng Qiujian, Shi Zhongzhi. A new method for fast computing positive region [J]. Journal of Computer Research and Development, 2003, 40(5): 637-642 (in Chinese)
(刘少辉,盛秋戬,史忠植. 一种新的快速计算正区域的方法[J]. 计算机研究与发展,2003,40(5):637-642)

Deng Dayong, born in 1968. PhD and associate professor. His main research interests include rough sets, granular computing and data mining.

Xu Xiaoyu, born in 1990. MSc candidate. Her main research interests include rough sets and data mining.

Huang Houkuan, born in 1940. Professor and PhD supervisor. His main research interests include computational intelligence, data mining and multi-agent system.