文章编号:1672-3961(2025)02-0088-09 DOI:10.6040/j.issn.1672-3961.0.2024.184

基于图结构的概念漂移检测

周彦冰,马士伦,文益民*
(广西图像图形与智能处理重点实验室(桂林电子科技大学),广西 桂林 541004)

摘要:为了解决传统的概念漂移检测方法,仅依赖错误率进行漂移检测不可靠的问题,提出一种基于图结构的概念漂移检测方法。该方法使用 k 关联最优图表示当前数据分布,定义样本的漂移率表示分类器与当前数据分布的不一致性,利用漂移率形成比特流,使用概念漂移检测器在比特流上检测概念漂移。通过与传统的使用错误率的概念漂移检测方法的对比和分析,结果表明在人工数据集上基分类器的准确率提高 1%~5%,在真实数据集上提高 1%~2%。所提出的方法有效提高概念漂移检测的准确性,帮助基分类器更好适应概念漂移。

关键词:数据挖掘;数据流;概念漂移;图结构;k 关联最优图
中图分类号:TP181 **文献标志码**:A
引用格式:周彦冰,马士伦,文益民. 基于图结构的概念漂移检测[J].山东大学学报(工学版),2025,55(2):88-96.
ZHOU Yanbing, MA Shilun, WEN Yimin. Concept drift detection based on graph structure[J]. Journal of Shandong University (Engineering Science), 2025, 55(2):88-96.

Concept drift detection based on graph structure

ZHOU Yanbing, MA Shilun, WEN Yimin*
(Guangxi Key Laboratory of Image and Graphic Intelligent Processing, Guilin University of Electronic Technology, Guilin 541004, Guangxi, China)

Abstract: In order to solve the problem that the traditional concept drift detection method only relied on the error rate for drift detection was not reliable enough, a concept drift detection method based on graph structure was proposed. In this method, the k-associated optimal graph was used to represent the current data distribution, and the drift rate of the sample was defined to represent the inconsistency between the classifier and the current data distribution. The drift rate was used to form a bit stream, and the concept drift detector was used to detect the concept drift on the bit stream. Compared with the traditional concept drift detection method using error rate, the results showed that the accuracy of the base classifier was improved by 1%-5% on artificial datasets and 1%-2% on real-world datasets. The proposed method could effectively improve the accuracy of concept drift detection and help base classifiers better adapt to concept drift.

Keywords: data mining; data stream; concept drift; graph structure; k-associated optimal graph

0 引言

在动态环境下进行分类是一项十分具有挑战性的任务。这项任务的困难在于数据分布可能会随时间推移而产生变化。这种数据分布随着时间而改变的情况被称为概念漂移。一旦在学习过程中发生了概念漂移,将直接导致模型泛化性能下降。为解决这个问题,许多数据流分类算法都采用概念漂移检测方法来探测这种变化,通过模型调整以适应概念漂移。在监督学习中,通常通过监控分类模型的泛化性能来检测概念漂移。当模型的泛化性能在一段时间内发生显著变化时,可以认为发生了概念漂移。

收稿日期:2024-07-24
基金项目:国家自然科学基金资助项目(62366011);广西重点研发计划资助项目(桂科 AB21220023);广西图像图形与智能处理重点实验室资助项目(GIIP2306)
第一作者简介:周彦冰(2001—),男,湖南益阳人,硕士研究生,主要研究方向为机器学习。E-mail:18074392274@163.com
*通信作者简介:文益民(1969—),男,湖南桃江人,教授,博士生导师,博士,主要研究方向为机器学习、数据流分类、媒体分析与数据挖掘。E-mail: ymwen@guet.edu.cn

现有的概念漂移检测算法,如漂移检测方法[1](the drift detection method, DDM),Hoeffding 漂移检测方法[2](hoeffding drift detection method, HDDM)等往往仅依赖于模型的错误率来进行漂移检测,这在某些情况下不够可靠。例如,在分类面不变的情况下,从左到右数据分布发生变化,分类精度却未发生改变。在这种情况下仅仅依赖错误率无法准确检测到概念漂移的发生。

本研究提出了基于图结构的概念漂移检测算法。该算法利用分类器的预测结果与 k 关联最优图计算漂移率[3],再通过漂移率来检测概念漂移。利用 k 关联最优图表示数据分布,将分类器的分类结果与图结构进行比较,判断分类结果是否符合数据分布。漂移率的本质是计算分类器与 k 关联最优图非一致性。漂移率大于阈值,认为该样本可能发生了漂移,将其标记为 0,反之标记为 1;利用检测器在比特流上进行概念漂移检测。本研究不依赖模型错误率,根据分类器分类结果与数据分布的非一致性,检测概念漂移。

1　相关工作

在监督环境下的概念漂移检测算法已有许多的研究,产生不少高质量的成果。基于模型性能的概念漂移检测方法大致可以分为 3 类[4]:统计过程控制方法、滑动窗口方法、集成方法。

统计过程控制方法,即通过监控模型的在线错误率变化来检测概念漂移。如果模型错误率变化超过显著性检验水平,便认为发生了概念漂移。这类方法中最著名的算法是由文献[1]提出的 DDM,该算法将误差认为是具有二项分布的伯努利随机变量,监控 t 时刻下的错误率 p_t 与标准差 s_t。在数据流中,当 $p_t+s_t<p_{\min}+s_{\min}$ 时,便使用 p_t 和 s_t 替换 $p_{\min}$ 和 $s_{\min}$,使 $p_{\min}$ 和 $s_{\min}$ 两者的和始终保持最小。当 $p_t+s_t\geqslant p_{\min}+2s_{\min}$ 时,DDM 发出漂移警告,当 $p_t+s_t\geqslant p_{\min}+3s_{\min}$ 时,DDM 检测到漂移。在 DDM 的基础上,文献[2]提出了 HDDM,该算法使用 Hoeffding 不等式来对 DDM 算法进行改进,得到了 HDDM_A 与 HDDM_W 两种变体算法。文献[5]在 HDDM 的基础上提出了快速 Hoeffding 漂移检测方法(fast hoeffding drift detection method, FHDDM),该算法采用滑动窗口来比较最大准确率与当前窗口内的准确率,如果差值大于 Hoeffding 界则认为发生了概念漂移。文献[6]对 FHDDM 进行扩展提出了堆叠快速 Hoeffding 漂移检测方法(stacking fast hoeffding drift detection method, FHDDMS),该算法使用不同大小滑动窗口来检测不同类型漂移。文献[7]提出精确概念漂移检测方法(accurate concept drift detection method, ACDDM)利用 Hoeffding 不等式来分析错误率的不一致性,检测概念漂移。由文献[8]提出的早期漂移检测方法(early drift detection method, EDDM)通过跟踪两个连续错误分类样本之间距离来检测概念漂移。这种方法可以更有效检测渐进概念漂移。

滑动窗口方法根据时间顺序将数据流划分为窗口,通过监测模型性能在窗口内的变化检测概念漂移。文献[9]提出的自适应窗口(adaptive windowing, ADWIN)是最著名的滑动窗口方法。该方法利用 Hoeffding 界来比较两个不同大小子窗口内的均值,如果大于 Hoeffding 界则认为发生了概念漂移。另一个著名方法是由文献[10]提出的等比例统计检验检测方法(statistical test of equal-proportion eetection, STEPD),该算法设置了最近窗口和整体窗口,使用等比例统计检验来比较两个窗口间的精度。文献[11]对 STEPD 进行改进并提出了 3 种新的概念漂移检测方法,Fisher 比例漂移检测器(fisher proportions drift detector, FPDD),Fisher 平方漂移检测器(fisher square drift detector, FSDD)和 Fisher 试验漂移检测器(fisher test drift detector, FTDD)。这些方法都是利用 Fisher 检验来计算显著性水平 p[12]。文献[13]提出的 Wilcoxon 漂移检测器(wilcoxon rank sum test drift detector, WSTD),受到 STEPD 的启发,采用 Wilcoxon 秩和检验来检测漂移[14]。文献[15]提出的余弦相似度漂移检测器(cosine similarity drift detector, CSDD)的工作原理类似于 WSTD,它基于每个窗口的 P_{PV} 和 F_{DR} 来计算混淆矩阵,其中 $P_{\mathrm{PV}}=T_{\mathrm{P}}/(T_{\mathrm{P}}+F_{\mathrm{P}})$, $F_{\mathrm{DR}}=F_{\mathrm{P}}/(T_{\mathrm{P}}+F_{\mathrm{P}})$, T_{P} 为真正例数,F_{P} 为假正例数。通过计算两个窗口的混淆矩阵创建出的向量之间余弦相似度来检测漂移。

集成方法通过考虑所有基学习器的准确率或每个单独基学习器的准确率来监控整体的性能。学习器性能显著下降则认为发生了概念漂移。文献[16]提出利用多样性处理漂移(diversity for dealing with drifts, DDD)通过融合低多样性和高多样性集成机制来控制集成中学习器的多样性水平。该方法利用低多样性集成来检测概念漂移,一旦检测到漂移,随即切换至高多样性集成,以应对变化。文献[17]提出的多样化在线集成检测(diversified online ensembles ddtection, DOED),设计了两个具有不同多样性程度的集成,分别标记为 E_0 和 E_1。在该方法中,通过单一的显著性水平 p 来判断 E_0 和

E_1 是否发生了概念漂移。一旦检测到任一集成出现漂移,就会对所有集成进行重置。两个集成都发现了漂移现象,特别对精度较低的集成执行重置操作。文献[18]提出的 k 类问题的在线漂移检测器(the online drift detector for the k-class problem, ODDK)通过构建一个列联表来记录每一对分类器之间的多样性变化情况,利用 PH 检验方法来检测概念漂移。

2 k 关联最优图

k 关联最优图(k-associated optimal graph, KAOG)由文献[3]在 2011 年提出,在 2013 年提出了其增量学习的版本[19]。KAOG 在各种分类任务中都有应用,文献[20]提出的 KAOGSS 为 KAOG 在半监督学习中的应用,文献[21]利用 KAOG 来进行反事实解释。

KAOG 无需设置参数便能很好表示空间中数据样本间拓扑结构,本研究主要利用 KAOG 来表示当前概念下的数据分布。KAOG 的基础是 k 关联图(k-asssociated graph, KAG)。KAG 将样本数据作为顶点,根据样本数据之间的相似性构建边,边的方向为从当前样本到近邻样本。构建好的 KAG 由一系列连通子图组成,这些连通子图被称为 KAG 组件。对于每一个组件会定义一个纯度度量来表示组件的紧密性。纯度的数学定义为

$$\Phi_\alpha=\frac{D_\alpha}{2k}, \tag{1}$$

式中, D_α 为组件内顶点的平均度数,k 为人为指定的近邻数量。

在构建 KAG 中其 k 是固定的。在同一个样本集上对于不同的 k 可以构建出不同的 KAG,这些 KAG 组件各不相同,其纯度也不相同,有的组件纯度高,有的纯度低。只用一个固定的 k 来构建图,其生成的组件纯度不能保证最优。通过逐步增加 k 来构建 KAG,将高纯度组件替换低纯度组件,组件纯度一样会保留具有较小 k 的组件,直到图中所有组件平均纯度不再有提高。图中所有组件拥有不同的 k,到达最高的纯度,这样的图就为 KAOG。

3 基于图结构的概念漂移检测

本研究通过引入图结构摆脱了概念漂移检测对分类器错误率的依赖,具体算法如下所述。

3.1 漂移率的定义

本研究利用 KAOG 表示当前数据分布,通过比较分类器预测结果与 KAOG 的非一致性形成比特流,使用概念漂移检测器在比特流上检测概念漂移。为了表示非一致性,定义了样本的漂移率,漂移率数值大,意味着对于当前样本分类器与 KAOG 的非一致性程度高。图 1 所示为一个 KAOG 示例(省略边的方向),其中红色方块表示正例样本,蓝色圆形表示负例样本,虚线表示 KAOG 分类面,每一个连通子图表示为 KAOG 中的一个组件(图 1 中表示有 4 个组件),由正例样本构成的组件为正类组件,由负例样本构成的组件为负类组件。

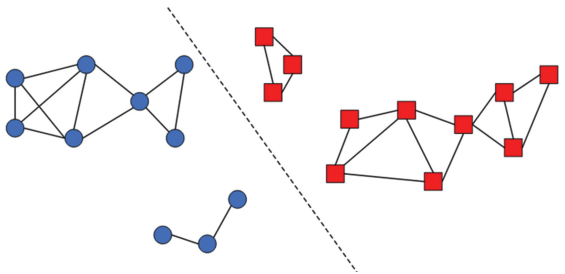

图 1 KAOG 示意图

Fig.1 Illustration of KAOG

数据流一直保持稳定,由分类器预测出的负例样本应落在 KAOG 分类面左边,预测出的正例样本则应该落在 KAOG 分类面右边。负例样本落到右边而正例样本落到左边,表明当前分类器预测结果与 KAOG 所表示的数据分布出现了差异,认为发生了概念漂移。将分类器预测结果中落在正确分类范围内的样本标记为 1,落在错误分类范围内的样本标记为 0,使用概念漂移检测器来监控比特流,检测概念漂移。

为了判断样本是否落在了正确分类范围,本研究通过定义漂移率来描述样本错误偏离 KAOG 分类面的程度,人为设置一个漂移阈值 β,当前样本漂移率小于 β,认为数据落在正确的分类范围内,即标记为 1,反之则标记为 0。

3.2 漂移率的计算

由于无法准确获取 KAOG 分类面的确切位置,本研究使用样本的邻域信息来计算样本漂移率。如图 2 所示,一个新的样本 v 到来,分类器预测 v 的类别为正例。

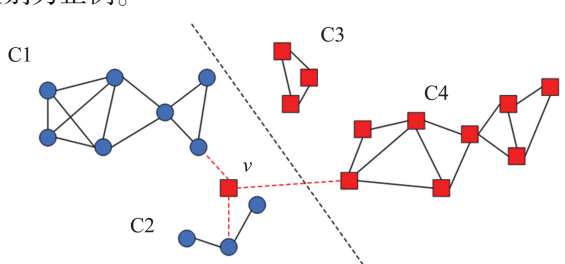

图 2 一个新的样本到来

Fig.2 A new sample arrives

当 v 到来后,开始寻找 v 的 k 近邻组件。由于组件间的不对等性,根据样本到组件的距离来寻找 k 近邻不够可靠。受万有引力公式启发,本研究利用一种新的引力模式来寻找样本 k 近邻。万有引力公式为

$$F=G\frac{M_1M_2}{R^2}, \tag{2}$$

式中,G 为重力常数,M_1 和 M_2 为两个质点质量,R 为质点间的距离。

受万有引力公式启发,本研究定义组件的引力公式为

$$F=\frac{\Phi N}{D^2}, \tag{3}$$

式中,Φ 为组件纯度,N 为组件顶点数量,D 为组件样本之间的距离。

根据引力公式,将寻找那些纯度高、包含顶点数量多、距离样本近的组件。这样的组件通常更为稠密,更具有代表性。在计算组件与样本之间距离中,将组件抽象为一个质点,再进行距离计算。这种抽象化的方法可以简化计算过程,同时忽略组件内部细节,使得距离计算更加直接。将组件抽象为点的公式为

$$V^{\alpha}=\sum_{i=1}^{n}w_i\boldsymbol{V}_i,\ V_i\in C_{\alpha}, \tag{4}$$

$$w_i=\frac{d_i^{\text{in}}+d_i^{\text{out}}}{d_{\alpha}}, \tag{5}$$

式中,w_i 为组件 α 中顶点 $\boldsymbol{V}_i$ 的度数占组件 α 总度数的比例,d_i^{in} 为顶点 V_i 的入度,d_i^{out} 为顶点 $\boldsymbol{V}_i$ 的出度,d_{α} 为组件 α 的总度数。

计算样本 v 与组件 α 间距离 D

$$D=\text{dist}(V^{\alpha},v), \tag{6}$$

式中 dist( ) 为距离函数。

漂移率计算公式为

$$D_R=\frac{1}{k}\sum_{i=1}^{k}\Phi_i\,\text{Ⅱ}(c\neq c_i), \tag{7}$$

式中:Φ_i 为 k 近邻组件中第 i 个组件的纯度;c_i 为第 i 个组件的类别;c 为样本预测类别;Ⅱ( ) 为指示函数,Ⅱ($c\neq c_i$) 表示当 c 与 c_i 不同输出 1,反之输出 0。

根据式(7)可知,漂移率的本质是样本 k 近邻组件中与样本不同类组件的平均纯度,其中与样本同类的组件可以看作是一个纯度为 0 的不同类组件。当样本落在错误分类范围中,其 k 近邻组件中多数为不同类组件,此时漂移率大;当样本落在正确分类范围中,其 k 近邻组件中多数为同类组件,此时漂移率小。这样设置一个漂移阈值 β,当漂移率大于 β 就认为样本落在错误分类范围内;当漂移率小于 β 就认为落在正确分类范围内。

3.3　基于图结构的概念漂移检测算法

基于图结构的概念漂移检测算法的伪代码如算法 1 所示。首先是初始化阶段,使用数据流中一小批数据对分类器 H 进行预训练,并构建初始 KAOG。随后当新样本到来计算样本漂移率,根据漂移率来向检测器中添加比特位,进行概念漂移检测。然后根据检测器的检测结果对 KAOG 进行更新,当检测器发出漂移预警会向 window 中添加当前样本,一旦检测到概念漂移,就使用 window 中保存的样本构建一个新的 KAOG 替换旧的 KAOG 来表示新概念下的数据分布。通过这样的方法可以保证 KAOG 始终代表的是最新的数据分布。完整的检测流程如图 3 所示。

算法 1　基于图结构的概念漂移检测算法

输入　数据流 D, 组件数 k, 漂移阈值 β;

输出　预警状态 S_w, 漂移状态 S_d;

创建一个空的缓存 W;

for $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ in D

向缓存 W 中添加样本 $\boldsymbol{x}_i$;

end for

根据缓存 W 中的样本初始化分类器 H 与 k 关联最优图 G_K;

清空缓存 W;

for $\boldsymbol{x}_{n+1},\boldsymbol{x}_{n+2},\boldsymbol{x}_{n+3},\cdots$ in D

使用分类器 H 对当前样本进行分类 $c=H(\boldsymbol{x}_j)$;

根据式(7)计算漂移率 $D_R=C_{\text{dr}}(\boldsymbol{x}_j,G_K,k,c)$;

if $D_R<\beta$ then

向漂移检测器 d 中添加比特位 1;

else

向漂移检测器 d 中添加比特位 0;

end if

if $d.S_w$ is True then

发出预警警报 $S_w=T$;

向缓存 W 中添加样本 $\boldsymbol{x}_j$;

end if

if $d.S_d$ is True then

发出漂移警报 $S_d=T$;

根据缓存 W 中的样本创建新的 k 关联最优图 G_K,并清空缓存 W;

end if

end for

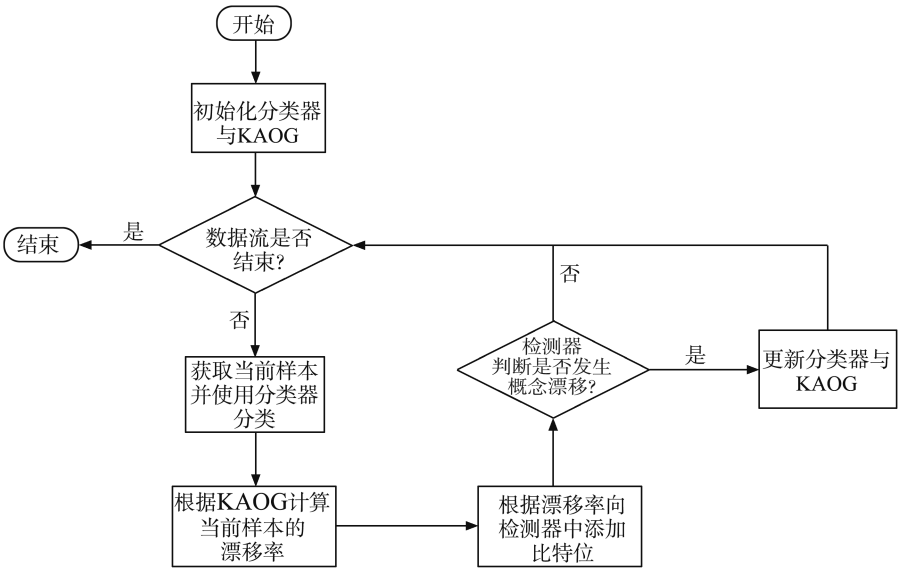

图 3 基于图结构的概念漂移检测流程图

Fig.3 Concept drift detection flow chart based on graph structure

4 试验验证

4.1 试验数据集

为了证明算法的有效性,在 5 个人工数据集和 3 个真实数据集上进行了广泛的试验。表 1 为数据集具体信息。这些数据集在概念漂移研究领域广泛使用,具有不同的特点并来自于不同的数据源。其中前 5 个为人工数据集,后 3 个为真实数据集。

表 1 多维人工和真实数据集

Table 1 Multidimensional artificial and real data sets

数据集	属性	标签	实例数量/个	漂移类型
Mixed	4	2	30 000	突变,重现
Agrawal	6	3	30 000	突变,重现
Sine	2	2	30 000	突变,重现
STAGGER	3	2	30 000	突变,重现
SEA	3	2	30 000	突变,重现
Electricity	7	2	45 312	未知
Covertype	54	7	581 012	未知
Weather	8	2	18 159	未知

注:对于人工数据集生成了 30 000 个实例,设置每 1 000 个实例发生一次概念漂移,一共发生了 29 次概念漂移。

4.2 对比算法和配置

在 DDM、HDDM_A 和 HDDM_W3 个概念漂移检测器上对错误率和漂移率进行比较,主要的评价指标为基分类器的准确率,准确率的数值越高表示概念漂移检测的越准确。HDDM_A 和 HDDM_W 的参数设置都与原论文中相同[2],即设置 $\alpha_W = 0.005$, $\alpha_D = 0.001$, HDDM_W 中 $\lambda = 0.05$。使用霍夫丁树(hoeffding tree,HT)作为基分类器,所有试验中设置其参数为 $\delta = 10^{-7}$, $\tau = 0.05$, $n_{\min} = 200$。对于漂移率的计算,所有试验都设置 $k = 5$,漂移阈值 $\beta = 0.4$。对于 k 的选择,在 Agrawal 数据集上使用不同检测器对不同的 k 进行比较试验,试验结果如表 2 所示。分析试验结果可知,算法对 k 的敏感性不高,但在 $k = 5$ 时效果最优。

表 2 不同 k 下分类器准确率的比较

Table 2 Comparison of classifier accuracy at different k

单位:%

检测器	$k=3$	$k=5$	$k=7$	$k=10$
HDDM_A	68.43	69.16	68.45	68.83
HDDM_W	72.86	73.17	70.07	70.56
DDM	65.40	66.01	64.50	64.59

注:在试验中,当检测到概念漂移,将直接训练新的分类器替换旧的分类器适应概念漂移。

4.3 概念漂移检性能试验结果及分析

在生成的二维正态分布数据集上,比较不同概念漂移检测算法下,利用错误率(E_R)或漂移率(D_R)检测概念漂移的准确性。试验结果如表 3 所示,每一个结果进行 10 次重复试验。试验使用人工生成的二维正态分布数据集来实现,数据集设置两组正态分布参数来表示两个类别,其中表示正例的参数为均值 $\mu_1 = 0.2$,标准差 $\sigma_1 = 0.1$,协方差 $\rho_1 = 0$;表示负例的参数为均值 $\mu_2 = 0.5$,标准差 $\sigma_2 = 0.1$,协方差 $\rho_2 = 0$。一共生成了 50 000 个实例,每 1 000 个实例表示一个概念,一共发生了 49 次概念漂移。在一个概念中包含 500 个正例和 500 个负例。通过对正例和负例的正态分布均值增加 0.2 改变一次数据分布,即制造概念漂移。

在试验中,如果检测器在发生概念漂移后 100 个实例内检测到概念漂移则为 N_{Detected};若在 100 个实例

后检测到概念漂移则为 N_{Delay};如果检测器在1 000个实例内检测到多次概念漂移则认为除了第一次的检测外其余检测发生了错误,则为 N_{False};若检测器在1 000个实例内没有检测到概念漂移则为 N_{Missed}。

分析表 3 的试验结果,可以发现在 HDDM_A 和 HDDM_W 两个检测器上漂移率明显要比错误率检测到的概念漂移更加准确,在 DMM 检测器上 D_R 则稍逊于 E_R。在 HDDM_A 和 HDDM_W 上,使用 D_R 进行检测会出现少量检测错误情况,这说明利用 D_R 来进行检测提高了准确度,降低了稳定性。

表 3　在 3 种检测器上使用漂移率与错误率的平均检测性能的比较

Table 3　Comparison of the average detection performance using drift rate versus error rate on the three detectors

检测器	指标	N_{Detected}	N_{Delay}	N_{False}	N_{Missed}	检测时间/s
HDDM_A	E_R	0.4	34.2	0	14.4	58.55
	D_R	44.0	2.8	2.8	2.2	82.43
HDDM_W	E_R	0.2	32.6	0	16.2	82.05
	D_R	45.6	0.2	4.4	3.2	93.78
DDM	E_R	0	20.0	0	29.0	31.74
	D_R	0	16.0	0	33.0	52.64

表 4　在 3 种检测器上使用漂移率与错误率的分类器平均准确率的比较

Table 4　Comparison of the average accuracy using drift rate versus error rate on the three detectors

数据集	分类器准确率/%						
	无检测器	HDDM_A-ER	HDDM_A-DR	HDDM_W-ER	HDDM_W-DR	DDM-ER	DDM-DR
Mixed	52.08	53.33	**54.97**	54.55	**54.87**	**56.96**	53.77
Agrawal	63.13	68.77	**69.16**	70.83	**73.17**	65.36	**66.01**
Sine	55.29	**68.96**	68.15	68.68	**73.97**	**65.57**	65.06
STAGGER	71.33	75.02	**75.13**	78.08	**82.68**	**73.02**	71.98
SEA	79.75	80.36	**80.68**	80.15	**80.53**	**80.09**	80.05
Electricity	79.38	83.55	**84.16**	83.00	**84.56**	80.33	**82.27**
COVERTYPE	82.40	82.88	**83.56**	82.70	**83.21**	81.55	**81.61**
Weather	**73.45**	69.81	70.69	69.72	70.03	69.45	70.01

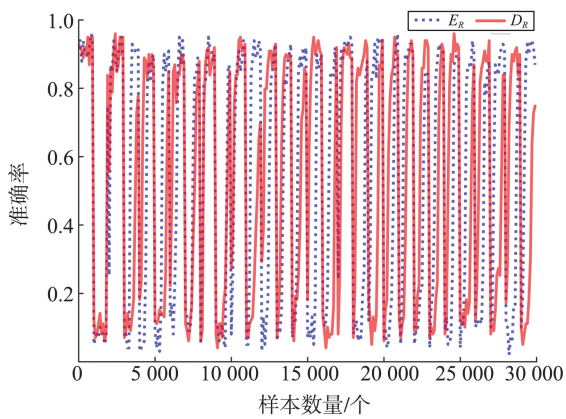

图 4　Mixed 数据集 HDDM_A 实时准确率比较

Fig.4　Comparison of real-time accuracy on HDDM_A on the Mixed dataset

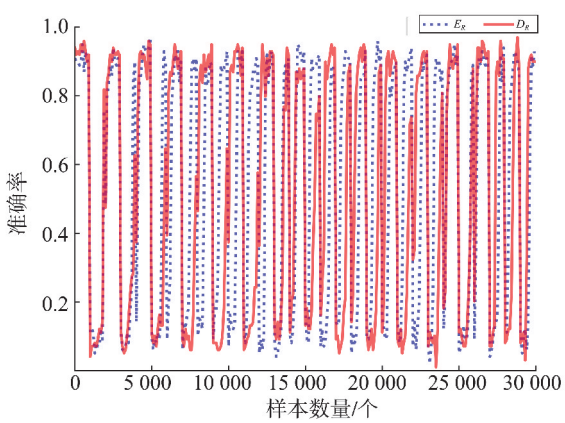

图 5　Mixed 数据集 HDDM_W 实时准确率比较

Fig.5　Comparison of real-time accuracy on HDDM_W on the Mixed dataset

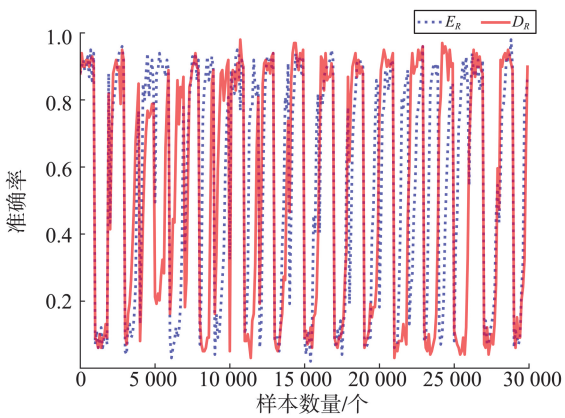

图 6　Mixed 数据集 DDM 实时准确率比较

Fig.6　Comparison of real-time accuracy on DDM on the Mixed dataset

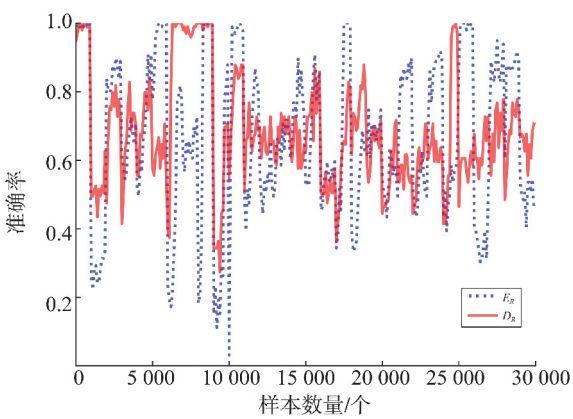

图 7　Agrawal 数据集 HDDM_A 实时准确率比较

Fig.7　Comparison of real-time accuracy on HDDM_A on the Agrawal dataset

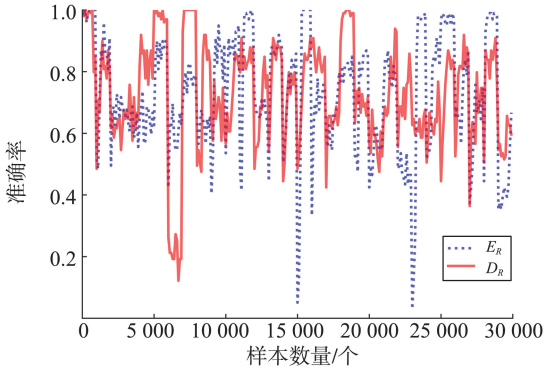

图 8 Agrawal 数据集 HDDM_W 实时准确率比较

Fig.8 Comparison of real-time accuracy on HDDM_W on the Agrawal dataset

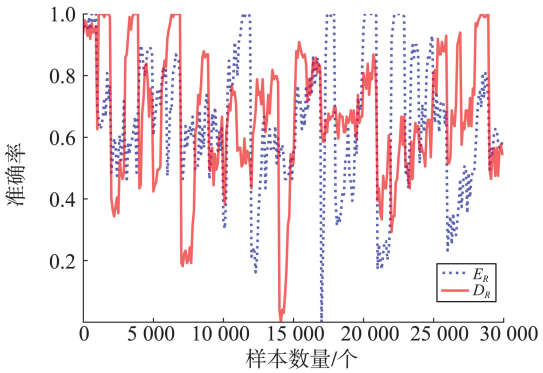

图 9 Agrawal 数据集 DDM 实时准确率比较

Fig.9 Comparison of real-time accuracy on DDM on the Agrawal dataset

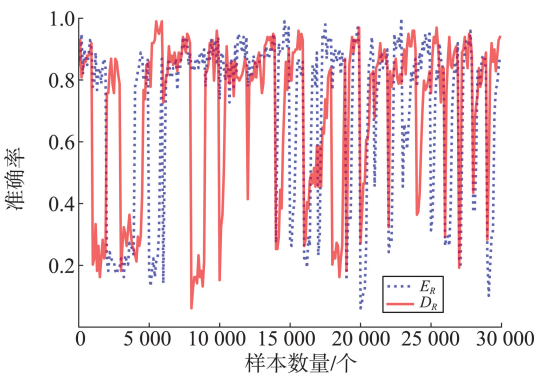

图 10 Sine 数据集 HDDM_A 实时准确率比较

Fig.10 Comparison of real-time accuracy on HDDM_A on the Sine dataset

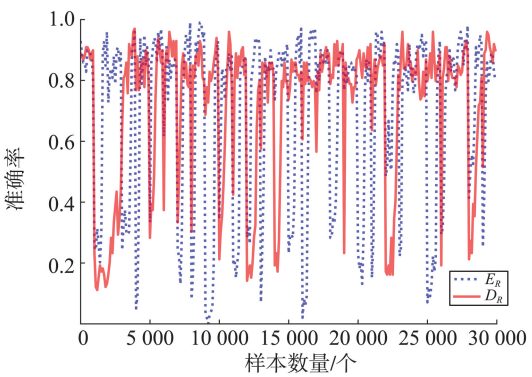

图 11 Sine 数据集 HDDM_W 实时准确率比较

Fig.11 Comparison of real-time accuracy on HDDM_W on the Sine dataset

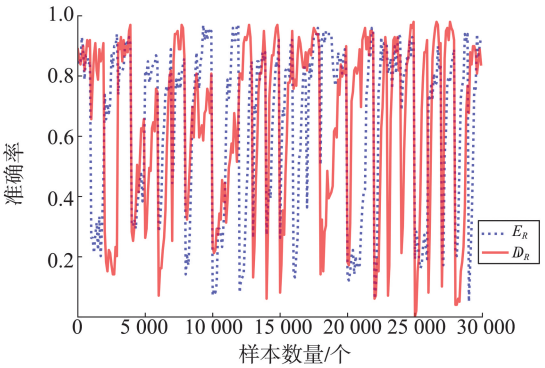

图 12 Sine 数据集 DDM 检测器比较

Fig.12 Comparison of real-time accuracy on DDM on the Sine dataset

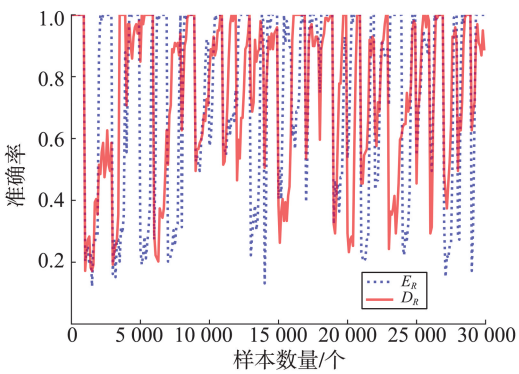

图 13 STAGGER 数据集 HDDM_A 准确率比较

Fig.13 Comparison of real-time accuracy on HDDM_A on the STAGGER dataset

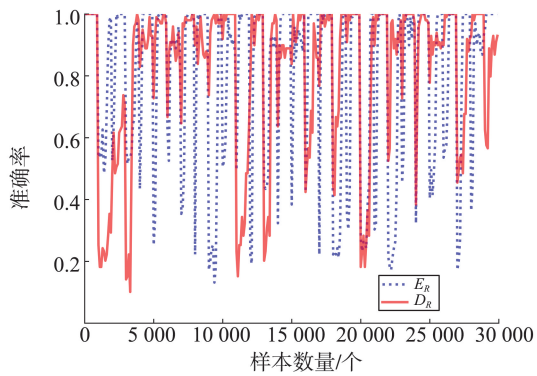

图 14 STAGGER 数据集 HDDM_W 准确率比较

Fig.14 Comparison of real-time accuracy on HDDM_w on the STAGGER dataset

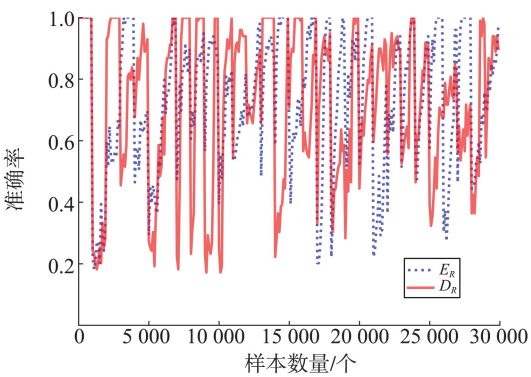

图 15 STAGGER 数据集 DDM 实时准确率比较

Fig.15 Comparison of real-time accuracy on DDM on the STAGGER dataset

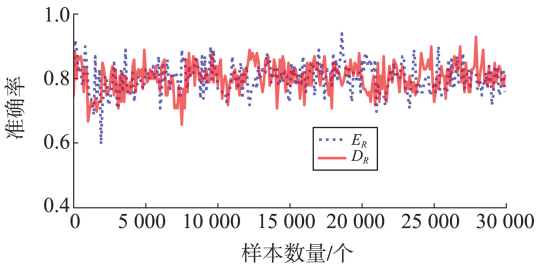

图 16　SEA 数据集 HDDM_A 实时准确率比较

Fig.16　Comparison of real-time accuracy on HDDM_A on the SEA dataset

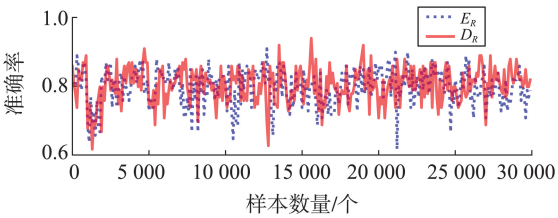

图 17　SEA 数据集 HDDM_W 实时准确率比较

Fig.17　Comparison of real-time accuracy on HDDM_W on the SEA dataset

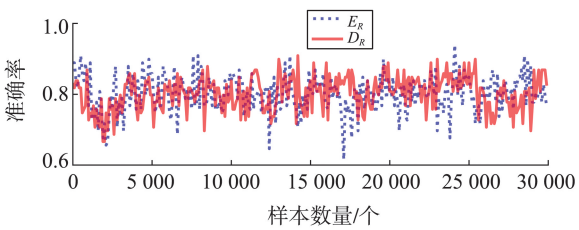

图 18　SEA 数据集 DDM 实时准确率

Fig.18　Comparison of real-time accuracy on DDM on the SEA dataset

4.4　分类器准确率试验结果及分析

使用分类器分别在 5 个人工数据集和 3 个真实数据集上进行试验,比较在使用不同检测器和指标下分类器的准确率。累积准确率试验结果如表 4 所示,其中前 5 个数据集为人工数据集,后 3 个数据集为真实数据集,对于人工数据集每个试验结果都进行了 20 次的重复试验,对于真实数据集每个试验结果都进行了 10 次的重复试验,表 4 中用黑体标注了同一检测器下分类器的最佳性能;实时准确率的结果如图(4-18)所示,在试验中使用每 100 个样本来计算一次准确率。根据表 4 中的试验结果,对于人工数据集,在 HDDM_A 与 HDDM_W 这两个检测器上,使用漂移率来检测概念漂移要比使用错误率更加有效,除了在数据集 Sine 上使用错误率的 HDDM_A 要比使用漂移率更加有效外,其余结果均是使用漂移率更加有效;对于 DDM,使用错误率要比使用漂移率更加有效,原因是 DDM 检测器本身的误检率比较高,导致使用漂移率效果不佳。对于真实数据集,发现在 HDDM_A、HDDM_W 和 DDM 这 3 个检测器上均是使用漂移率更加有效。在 Weather 数据集上未使用漂移检测情况下准确率要更高,原因是 Weather 数据集的样本数量较少,当检测到漂移后只能使用很少量的样本来训练新分类器,导致新训练的分类器的质量不如旧分类器。在实时精度图中,红线为基于漂移率检测,蓝线为基于错误率检测,试验结果表明当准确率下降时,使用漂移率检测可以更快发现这种变化并开始适应概念漂移,提高分类器的准确率。本研究提出的算法在大部分情况下要优于传统仅依赖错误率的漂移检测算法。

5　结论

本研究提出了一种基于图结构的概念漂移检测方法,该方法利用 KAOG 表示当前数据分布,根据当前数据分布与分类器的不一致性定义漂移率,使用漂移率形成比特流,在比特流上进行概念漂移检测。与传统的概念漂移检测方法不同,本研究的方法不依赖于分类器的错误率,通过分析分类器与图结构之间的不一致性来进行概念漂移检测。这种方法有效避免了传统方法仅依赖错误率所带来的不可靠性,特别是在数据分布改变而分类器错误率没有发生变化的情况下,传统检测方法无法有效检测,本研究所提出的方法可以有效进行检测。利用多个人工数据集和真实数据集评估方法的有效性,试验结果表明本研究提出的方法在大部分情况下要比传统依赖错误率的检测算法更加准确。

参考文献:

[1] GAMA J, MEDAS P, CASTILLO G, et al. Learning with drift detection [C]// Advances in Artificial Intelligence-SBIA 2004: 17th Brazilian Symposium on Artificial Intelligence. Sao Luis, Brazil: Springer, 2004: 286-295.

[2] FRIAS-BLANCO I, DEL CAMPO-ÁVILA J, RAMOS-JIMENEZ G, et al. Online and non-parametric drift detection methods based on Hoeffding's bounds[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 27(3): 810-823.

[3] BERTINI J R, ZHAO L, MOTTA R, et al. A nonparametric classification method based on k-associated graphs[J]. Information Sciences, 2011, 181(24): 5435-5456.

[4] BAYRAM F, AHMED B S, KASSLER A. From concept

drift to model degradation: an overview on performance-aware drift detectors [J]. Knowledge-Based Systems, 2022, 245: 108632-108651.

[5] PESARANGHADER A, VIKTOR H L. Fast hoeffding drift detection method for evolving data streams [C]// Machine Learning and Knowledge Discovery in Databases: European Conference. Riva del Garda, Italy: Springer, 2016: 96-111.

[6] PESARANGHADER A, VIKTOR H, PAQUET E. Reservoir of diverse adaptive learners and stacking fast hoeffding drift detection methods for evolving data streams [J]. Machine Learning, 2018, 107 (11): 1711-1743.

[7] YAN M M W. Accurate detecting concept drift in evolving data streams [J]. ICT Express, 2020, 6(4): 332-338.

[8] BAENA-GARCIA M, DEL CAMPO-ÁVILA J, FIDALGO R, et al. Early drift detection method [C]//Fourth international workshop on knowledge discovery from data streams. Berlin, Germany: ACM, 2006, 6: 77-86.

[9] BIFET A, GAVALDA R. Learning from time-changing data with adaptive windowing [C]//Proceedings of the 2007 SIAM international conference on data mining. Minneapolis, USA: SIMA, 2007: 443-448.

[10] NISHIDA K, YAMAUCHI K. Detecting concept drift using statistical testing[C]//International conference on discovery science. Berlin, Germany: Springer, 2007: 264-269.

[11] DE LIMA CABRAL D R, DE BARROS R S M. Concept drift detection based on Fisher's exact test[J]. Information Sciences, 2018, 442: 220-234.

[12] FISHER R A. On the interpretation of χ^2 from contingency tables, and the calculation of P[J]. Journal of the Royal Statistical Society, 1922, 85(1): 87-94.

[13] DE BARROS R S M, HIDALGO J I G, DE LIMA CABRAL D R. Wilcoxon rank sum test drift detector [J]. Neurocomputing, 2018, 275: 1954-1963.

[14] WILCOXON F. Individual comparisons by ranking methods[M]. New York: Springer, 1992: 196-202.

[15] HIDALGO J I G, MARIÑO L M P, DE BARROS R S M. Cosine similarity drift detector [C]//International Conference on Artificial Neural Networks. Munich, Germany: Springer, 2019: 669-685.

[16] MINKU L L, YAO X. DDD: A new ensemble approach for dealing with concept drift[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 24 (4): 619-633.

[17] SIDHU P, BHATIA M P S. An online ensembles approach for handling concept drift in data streams: diversified online ensembles detection[J]. International Journal of Machine Learning and Cybernetics, 2015, 6(6): 883-909.

[18] MAHDI O A, PARDEDE E, ALI N. A hybrid block-based ensemble framework for the multi-class problem to react to different types of drifts[J]. Cluster Computing, 2021, 24(3): 2327-2340.

[19] BERTINI J R, ZHAO L, LOPES A A. An incremental learning algorithm based on the K-associated graph for non-stationary data classification [J]. Information Sciences, 2013, 246: 52-68.

[20] BERTINI J R, LOPES A A, ZHAO L. Partially labeled data stream classification with the semi-supervised K-associated graph[J]. Journal of the Brazilian Computer Society, 2012, 18: 299-310.

[21] DA SILVA A T, BERTINI J R. Using the *k*-associated optimal graph to provide counterfactual explanations [C]//IEEE International Conference on Fuzzy Systems. Padua, Italy: IEEE, 2022: 1-8.

(编辑:陈燕)