计算机科学
COMPUTER SCIENCE

基于动态机器学习的信用评估模型

陈奕君，高浩然，丁志军

引用本文

陈奕君，高浩然，丁志军.基于动态机器学习的信用评估模型[J]．计算机科学, 2023, 50(1): 59-68.

CHEN Yijun, GAO Haoran, DING Zhijun. Credit Evaluation Model Based on Dynamic Machine Learning [J]. Computer Science, 2023, 50(1): 59-68.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220800191

基于动态机器学习的信用评估模型

陈奕君　高浩然　丁志军
嵌入式系统与服务计算教育部重点实验室(同济大学)　上海 201804
上海市网络金融安全协同创新中心(同济大学)　上海 201804
(2232918@tongji.edu.cn)

摘　要　随着计算机技术的发展,利用机器学习算法构建自动化评估模型已经成为金融机构进行信用评估的重要手段。然而,目前信用评估模型仍存在一些问题:信用数据本身存在类别不平衡和高维特征的问题,并且不同的时间下外界环境的改变会影响信用主体的行为,即数据会产生概念漂移现象。为此,文中提出了一个动态的信用评估模型,通过集成学习在新的增量数据上训练基分类器,并对各个基分类器的权重进行动态调整来适应概念漂移,以实现模型的动态更新。当发生概念漂移时,会针对概念漂移的检测结果对高维不平衡的信用数据进行不同形式的均衡化和特征选择。特别地,针对特征选择,文中提出了结合历史代表性样本的增量特征选择算法,该算法能够进行高效准确的特征选择,从而使模型可以同时解决增量信用数据存在的高维不平衡和概念漂移问题。最后,文中选取了真实的增量高维信用数据集,验证了所提算法相比其他主流算法在准确率和效率上的优越性。

关键词:信用评估;特征选择;概念漂移;滑动窗口;动态模型
中图法分类号　TP3-05

Credit Evaluation Model Based on Dynamic Machine Learning

CHEN Yijun,GAO Haoran and DING Zhijun
Key Laboratory of Embedded System and Service Computing of Ministry of Education,Tongji University,Shanghai 201804,China
Shanghai Network Finance Security Collaborative Innovation Center,Tongji University,Shanghai 201804,China

Abstract　With the development of computer technology,using machine learning algorithms to build automated evaluation models has become an important tool to for the financial institutions to conduct credit evaluation. However,currently,the credit evaluation model is still facing challenges:credit data is class-imbalanced and high-dimensional,meanwhile,the behavior of customers can be influenced by the changeable external environment,namely,the concept drift will occur. As a result,this paper proposes a dynamic credit evaluation model,which can achieve the flexible model update by using ensemble learning algorithm to continuously add base classifiers trained on new incremental data,and dynamically adjusting the weight of each base classifier to adapt to concept drift. When concept drift occurs,according to the detection results of concept drift,the model is able to use different forms of equalization and feature selection on credit data. In particular,for feature selection,this paper proposes an incremental feature selection algorithm combining the choice of representative samples that makes the feature selection efficient and accurate,enabling the model to simultaneously process the high-dimensional imbalanced data and adapt the concept drift of incremental credit data. Finally,this paper manages to demonstrate that the proposed dynamic model is more efficient and accurate than other prevailing algorithms on real incremental high-dimensional credit datasets.

Keywords　Credit evaluation,Feature selection,Concept drift,Sliding window,Dynamic model

1　引言

信用评估问题是银行或其他评估机构利用数据信息通过专家经验或数学规律分析等方法,对个人、企业等主体的各方面信用水平进行评价的一种工作[1]。在本文的借贷场景下,可以将信用评估问题看作通过模型分析信用数据来预测贷款人是否会违约的过程,其本质是一个二分类问题,即通过客户的个人信息和属性将其预测分类为信用良好和信用不良两个类别。随着计算机技术的不断发展,基于机器学习算法的信用风险评估方法凭借其高精度、高效率的优势,逐渐成为了

到稿日期:2022-08-19　返修日期:2022-09-21
基金项目:上海市科技创新行动计划(19511101300)
This work was supported by the Shanghai Science and Technology Innovation Action Plan(19511101300).
通信作者:丁志军(dingzj@tongji.edu.cn)

主流的信用评估方法。但是，对于目前主流的机器学习算法来说，解决实际的信用评估仍旧存在许多问题。

首先，一个人的信用数据在不同的时间和经济社会环境下反映的个人信用状态是不同的。个人信用状态会受到外界其他因素的影响，个人的还款意愿也会随之改变，这种随着时间的推移导致样本特征发生变化的现象被称作概念漂移[2]。数据发生概念漂移时，说明历史数据和最新数据的分布已经发生变化。如果仍通过历史数据的规律来对新数据进行预测，不对模型进行及时的调整，预测结果也会出现偏差。同时，加入新到达的增量数据进行训练能使数据更加充分，使模型更加准确。因此，信用评估需要构建一个动态的机器学习模型，使其能够根据时间的推移和新增的数据进行模型的动态更新和调整。

另外，现实中获取的贷款信用数据通常显示出高维不平衡的特点。当数据中不同类别的样本数量差异非常大时，称该数据是不平衡的[3]。具体来说，即信用评估中大部分数据都属于信用良好的类别，只有少部分数据是信用不良的。当两个类别的数据严重不平衡时，模型通常会偏向含有更多数据的多数类，出现忽略少数类的情况，这在很大程度上会影响信用不良数据的检出，在实际中造成很大的风险隐患。另外，信用数据包含客户多维度的特征信息，如信用历史、收入状况、履行能力等，其中会包含一些冗余的特征信息，冗余的特征信息不仅会对模型的训练产生额外的计算负担，还可能会对模型结果产生负面影响。同时，当每次获取到新增的信用数据时，都需要重新结合所有的历史数据进行全量数据的特征处理，因此历史的数据信息会被多次计算，即造成重复性计算，从而影响计算效率。因此，在面向信用评估的动态机器学习模型中，同样需要解决信用数据本身存在的高维不平衡问题。

目前的信用评估模型大多都不能同时解决信用数据中概念漂移以及高维不平衡的问题。这些模型在对新增的信用数据进行特征处理时，简单结合历史信息进行特征选择的方式会造成重复性计算和空间冗余；在发生概念漂移时，大多数静态模型也无法很好地适应概念漂移现象。因此，本文围绕信用评估中高维不平衡的增量数据以及随之出现的概念漂移问题展开研究，目标是通过设计一个动态模型，来解决信用数据的特点导致的模型计算复杂度增加、预测偏向于多数类样本以及概念漂移导致的准确率下降等问题。具体思路是设计一个信用评估的增量学习框架，利用概念漂移检测算法来判断新增数据是否出现了概念漂移现象，根据不同的检测结果进行不同的数据处理：如果没有发生概念漂移，则结合历史数据通过增量式特征选择高效地去除冗余特征，并进行数据均衡化；反之，则只利用新增数据进行特征选择和均衡化。在对新增数据进行处理后，对其构建一个新的基分类器，通过集成学习不断加入新的基分类器，并动态调整基分类器集成的权重来适应可能出现的概念漂移现象，最终完成一个动态机器学习的信用评估模型，具体流程如图 1 所示。

本文的主要贡献如下：

(1)完成了一个由概念漂移检测驱动的信用评估流程框架，可以先检测数据是否存在概念漂移现象，并根据不同的检测结果对新增数据进行不同方式的特征选择，利用增量集成学习进行模型动态调整，从而在信用评估的同时解决概念漂移和数据高维不平衡的问题。同时，将模型运用在真实的信用评估场景中，证明了模型的优越性。

(2)提出了一种结合代表性样本集的增量特征选择算法，将该算法引入最终的信用评估框架中，使其能够在新增数据没有发生概念漂移时，只利用增量计算的信息熵和有限的代表性样本集进行高效的特征选择工作。该算法与传统特征选择算法相比，在时间和空间上减少了冗余；与纯增量算法相比，提高了特征选择的准确率，在性能和效率两个方面都表现良好。

(3)提出了一种基于时间和性能因素的动态增量集成学习算法，能够根据基分类器的创建时间和性能表现对各数据的权重进行动态调整，从而使其能够很好地适应概念漂移现象。

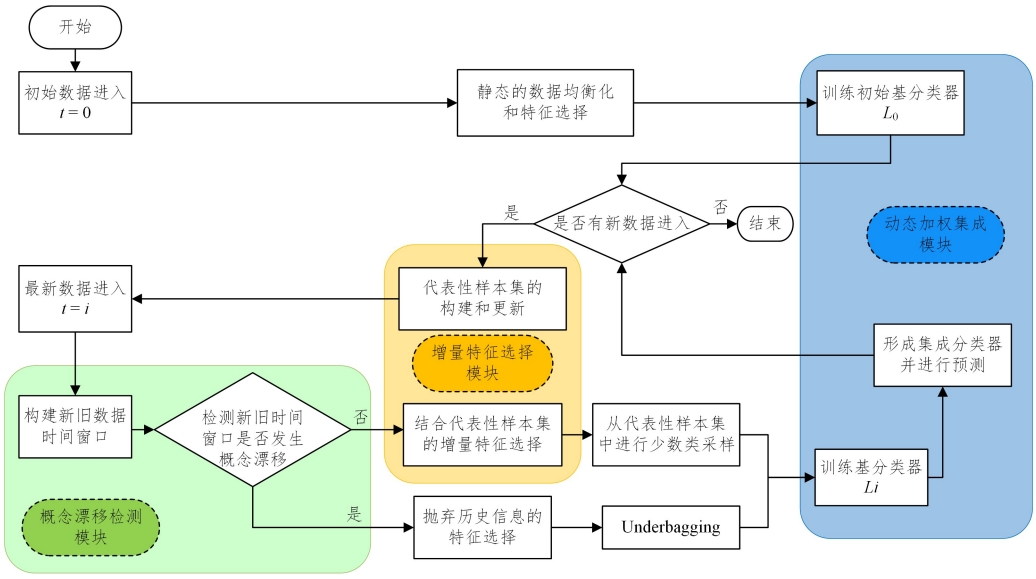


图 1 动态信用评估模型的总体框架流程

Fig. 1 Flow chart of dynamic credit evaluation model

2 相关工作

2.1 信用评估模型

目前常用的信用评估模型一般分为静态和动态两种。静态信用评估模型一般是根据静态的数据进行固定模型的训练，模型训练完成后不会发生变化。最经典的评估模型是美国 Fair Isaac 公司提出的 FICO 评分模型，该模型主要聚焦于客户 5 个方面的基本信用数据信息来对客户进行评分[4]。随着计算机技术的发展，一些传统的单一机器学习算法逐渐被运用到信用评估领域，如逻辑回归[5]、K 近邻[6]、支持向量机[7-8]、遗传算法[9]以及神经网络[10]等。针对信用评估数据中的类别不平衡数据及高维特征的问题，研究人员提出了许多结合数据均衡化方法和特征降维方法的静态信用评估模型。例如，Sun 等[11]基于差分采样率的 Bagging 集成学习(DSR)算法和合成少数过采样技术，提出了一种新的决策树集成模型，能够有效解决类别不平衡问题。Zhang 等[12]提出了一种新的混合遗传算法，用于获得最优的特征子集和基分类器子集，使用多数投票策略扩展了硬阈值方法来处理数据不平衡的现象，能够解决静态信用数据下的高维不平衡问题。

基于静态机器学习的模型在信用评估领域占据了主流地位，但由于一个人的信用数据在不同的时间和经济社会环境下反映的个人信用状态是不同的，会随着时间的推移而不断变化，并且获取到的信用数据量也会不断增加，因此静态信用评估模型通过历史数据来预测即将到来的新数据结果将会产生较大的偏差。为了解决这些问题，需要构建一个动态的信用评估模型来适应最新的概念，及时进行模型的更新和调整。

一些研究也将增量学习和流数据挖掘的相关知识运用到信用评估领域。Barddal 等[13]利用一些常见的动态数据流挖掘方法来构建信用评估模型，并将其与传统的静态信用评分模型进行对比，结果表明数据流挖掘方法具有更好的模型性能。Cai 等[14]考察了信用评估中的数据增量问题，提出了基于 Karush-Kuhn-Tucker(KKT)[15]条件和 SVDD[16]的增量学习算法，并利用该算法对信用数据进行预测。Tian 等[17]提出了一种基于 SMOTE 和加权多数投票法的增量学习集成方法，在解决数据增量问题的同时也解决了信用评估中数据不平衡的问题。但是，目前还没有模型能够同时解决类别不平衡、高维特征和概念漂移的增量数据的问题，大部分算法只是着重研究了其中的一两个方面，或者没有考虑到信用数据本身的高维特点，单纯从增量数据的角度来思考问题。因此，本文构建了一个由概念漂移检测驱动的动态信用评估模型，根据概念漂移检测结果对数据进行不同形式的特征选择和数据均衡化，并利用增量集成学习对模型进行动态调整，从而同时解决信用评估中增量数据的概念漂移和高维不平衡问题。

2.2 特征选择算法

特征选择的本质在于在不损失信息量的前提下，保留对模型有用的特征，尽量去除冗余的或可能对实验结果产生不良影响的特征，使得最终所选的特征子集是最小、最优的。常见的特征选择方法可以分为过滤法、封装法和嵌入法 3 类[18]。过滤法根据方差、相关系数、信息价值(IV)等特征的性能指标来进行特征选择；封装法利用算法本身的目标函数对特征进行选择，如递归特征消除法、序列前后向选择算法[19]等；嵌入法直接使用模型训练算法的评估结果来得到各个特征的权重，并根据权重大小进行特征选择，将模型的训练与特征的选择相结合，如基于正则项的方法、基于树模型的方法等。

传统的特征选择算法都是通过全量数据对特征进行筛选，当新增数据到来时，可能会因为新增数据的不充分，每次新增都需要结合历史数据重新进行特征选择。在增量数据的场景下，这样的方式会造成重复性计算与空间冗余等现象，因此出于处理动态数据和大数据的需求，增量式的特征选择非常重要。增量式特征选择的目的是利用新数据与历史数据之间增量的信息对新数据进行特征选择，从而避免重复多次使用历史数据或完全丢弃历史数据。

在增量特征选择中，主流算法是基于模糊粗糙集理论的，将通过该理论中定义的模糊相似性关系来描述数据样本之间的相似性、特征与标签之间的不一致性等，从而对特征集合的信息量进行衡量。这类算法可以通过增量计算信息粒度来为不断到来的新增数据选择最佳的特征子集。Shu 等[20]针对动态变化的混合型数据，提出了一种基于邻域粗糙集的增量特征选择算法，通过增量计算邻域熵进行特征过滤，能够有效进行新增数据的特征选择，并且降低了时间复杂度。Sang 等[21]定义了一种新的粗糙集模型，即模糊优势邻域粗糙集(Fuzzy Dominance Neighborhood Rough Sets，FDNRS)，利用 FDNRS 和条件熵设计了一种新的增量特征选择算法，并在公共数据集上进行了实验，验证了该算法在动态有序数据上进行特征选择的有效性和高效性。但在实际应用时，这些算法纯增量的计算方式通常会因为增量熵的近似计算而导致特征选择的偏差，并且对模型准确率的影响较大。因此，本文提出了结合代表性样本集的增量特征选择算法，该算法筛选保留小部分具有代表性的历史样本，利用具有代表性的历史样本来补充新增数据的信息，并且通过增量计算数据的信息熵来减少特征选择时的重复计算。

2.3 概念漂移适应算法

概念漂移的问题是在数据流处理中提出的，与时间和环境变化相关的数据通常有可能出现概念漂移现象。概念漂移指目标域的统计属性以任意方式随时间推移而变化的现象[22]，概念漂移适应就是通过调整或更新模型来解决模型中数据存在的概念漂移问题的方法。目前的适应算法通常建立在数据流分类算法和增量集成模型的研究上。其中，常用的算法是 Elwell 等[23]提出的 Learn++.NIE 和 Learn++.CDS 算法，这两种算法可以为历史数据和增量数据划分时间窗口，并在每个时间窗口内训练基分类器，利用时间衰减函数对基分类器进行权重调整，最后集成总分类器以完成预测，两种算法都能利用新增的数据块实现对模型的调整更新。Zhang 等[24]针对存在概念漂移的数据流，提出了一种自适应的在线增量学习算法(Adaptive Online Incremental Learning，AOIL)，通过构建带有内存模块的自动编码器来检测概念漂移并及时调整模型参数。Li 等[25]对非平稳不平衡的数据流进行了研究，提出了一种基于块的增量集成算法(Dynamic Updated Ensemble，DUE)，该算法能够通过选择数据流中的

样本集合来快速处理概念漂移和类别不平衡的联合问题，并在真实数据集上验证了该算法的有效性。

由于流数据是大量实时到达的有序实例序列，具有数据数量持续增加、按照时序依次到达和时效性的特点，因此信用数据具有与流数据相似的性质，两者也同样存在概念漂移现象，但是两者的区别在于，信用评估中用到的信用数据不要求模型的强实时性。因此，在本文提出的信用评估框架中，将信用数据作为分批到达的增量数据，利用增量集成学习对新增数据构建基分类器，并对基分类器的权重进行动态调整，从而使模型适应概念漂移现象。另外，目前的适应算法大多是通过丢弃历史基分类器或者降低历史分类器权重的形式，只根据基分类器的创建时间对模型进行调整，这样很可能会丢弃一些有效的历史基分类器而保留了无效的新分类器，因此本文对此进行了改进，从性能和时间两个角度进行衡量，使基分类器权重调整更为有效。

3 面向信用评估的动态增量集成学习模型

3.1 概念漂移检测驱动的动态信用评估模型

由于数据量不断增加，模型必须适应概念漂移现象，并对其做出及时的反馈，在新数据的基础上不断更新模型。现有的更新方法大多是丢弃历史数据，只关注新数据，利用新数据来产生最新的模型。但历史数据不是完全与新数据割裂的，历史数据中也可能蕴含着有用的信息，且与历史数据相比，新数据的数据量不足通常会导致模型容易出现过拟合的现象。因此，模型需要在保留有用历史信息的基础上，根据新数据得到新知识、构建新模型。另外，数据的类别不平衡和高维特征问题在增量式数据中需要进行额外的处理。当数据以增量的形式到达时，需要利用历史数据信息来对最新数据进行相应的数据均衡化和特征选择，并利用处理后的最新数据进行模型的更新。

根据信用评估数据中出现的类别不平衡、高维特征以及概念漂移的问题，本文设计了一个基于时间窗口和增量学习的动态加权集成模型(Credit Evaluation on Dynamic Machine Learning，CEDML)，其设计思想是将到达的信用数据按照时间顺序划分为不同的窗口，通过检测最新窗口和历史窗口中数据分布的偏差来判断数据是否出现了概念漂移现象，并根据概念漂移的检测结果对类别不平衡和高维特征数据选择不同的均衡化和特征选择方式。集成学习通过保留历史基分类器模型的方式来保留有效的历史信息，无须对全体历史数据进行存储，减少了多余的存储，将基分类器加权结合得到集成分类器，分类器将随着新数据的到来而不断更新，以适应概念漂移。具体的模型构建框架如图 1 所示，模型中其他具体算法将在 3.2 节与 3.3 节中详细阐述。

首先，当 $t=0$ 时，初始数据进入，由于初始数据是第一批数据，因此需要进行数据均衡化与特征选择，并生成初始基分类器 L_0。然后，新的数据不断到达，将数据划分为新旧两个滑动窗口，窗口大小为 q 个数据块，则新窗口中包含最新到达的 q 个数据块，旧窗口中包含新窗口中到达的 q 个数据块，两个窗口将随着数据的到来不断向后滑动。如图 2 所示，窗口大小 $q=5$，当 $t=k$ 时新窗口 W_k^{new} 包含时间区间$[k-4,k]$到达的数据块 $S_{k-4\sim k}$，旧窗口 W_k^{old} 包含时间$[k-9,k-5]$到达的数据块 $S_{k-9\sim k-5}$；下一时刻 $t=k+1$ 时，W^{new} 与 W^{old} 都向后滑动一个数据块，W_{k+1}^{new} 包含数据块 $S_{k-3\sim k+1}$，W_{k+1}^{old} 包含数据块 $S_{k-8\sim k-4}$，以此类推。将新窗口中的数据分布与历史窗口中数据的分布进行比较，检测当前新时刻的数据是否发生了概念漂移。如果没有发生概念漂移，则说明新数据的分布与历史数据类似，可以利用历史数据的信息结合新信息进行数据的均衡化和特征选择；如果发生了概念漂移，则说明历史数据的分布与最新数据的分布不一致，那么此时将只对新数据进行均衡化和特征选择。t 时刻新到来的数据 S_t 经过预处理后，将其作为训练集得到基分类器 L_t。因此，在时间区间$[0,t]$上，会利用数据$\{S_0,\cdots,S_t\}$训练得到 t 个基分类器，如果某些基分类器的错误率太高，性能较差，超过了设定的阈值，那么就将其删去，最终得到 m 个有效的基分类器。每次新到来的数据所训练的新基分类器都会被加入总模型中，每个基分类器 L_i 有其对应的分类器权重 w_i，通过加权集成的方式形成总的分类器 $\mathcal{L}$，最后利用分类器 $\mathcal{L}$ 对下一时刻的数据 S_{t+1} 进行预测，通过其预测准确率来对总模型进行评估。

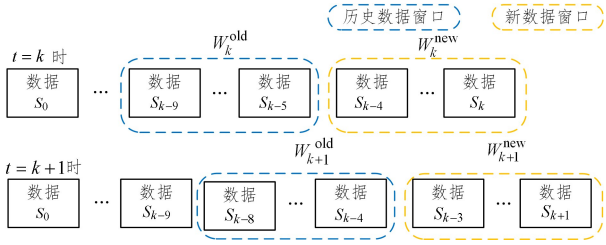


图 2 滑动窗口示意图

Fig. 2 Sliding windows

算法 1 构建动态信用评估模型

输入：增量数据 S_i

输出：每个时刻的最新信用评估模型 L

1. $S=\{S_0,S_1,\cdots\}$，S_i 为时刻 i 到来的最新数据
2. 对初始数据 S_0 进行初始数据均衡化和特征选择，训练初始基分类器 L_0
3. while 新数据 S_i 到达 then
4. 检测 S_i 是否发生了概念漂移现象
5. if S_i 没有发生概念漂移 then
6. 在数据 S_{i-1} 中筛选出历史代表性样本集 M_i
7. 结合 M_i 和最新数据 S_i 进行增量式特征选择
8. 结合 M_i 和最新数据 S_0 进行数据均衡化
9. else
10. 只利用最新数据 S_i 进行均衡化和特征选择
11. $S_i'\leftarrow S_i$ 进行特征选择和数据均衡化后的结果
12. 用 S_i' 训练基分类器 L_i
13. 将$\{L_0,L_1,\cdots,L_i\}$进行加权集成，得到最新模型 L，并用 L 对下一时刻的数据进行预测
14. end for

3.2 结合代表性样本集的增量特征选择算法

3.2.1 模糊粗糙集理论与增量信息熵

定义一个模糊信息系统为(U,A)，其中 $U=\{x_1,x_2,\cdots,x_n\}$是非空有限对象集，称为论域，$A=\{a_1,a_2,\cdots,a_m\}$是属性集。对于 A 的每个属性 $a_t\in A$，都有一个映射关系 $U\rightarrow V_{a_t}$

成立，其中 V_{a_t} 是 a_t 的域，并且可以定义模糊关系 $R_{\{a_t\}}$。其中，模糊关系 $R_{\{a_t\}}$ 是定义在模糊幂集 $F(U\times U)$ 上的一个模糊集，$R_{\{a_t\}}(x_i,x_j)$ 用来反映对象 x_i 和 x_j 在属性 a_t 上的相似程度。

同时，可以定义基于模糊关系 R 的模糊集 X 的下近似算子 $\underline{R}_X(x_i)$ 和上近似算子 $\overline{R}_X(x_i)$[26]，来衡量 x_i 肯定属于 X 和 x_i 可能属于 X 的程度，因此可以用 $(\underline{R}_X,\overline{R}_X)$ 来定义 X 的模糊粗糙集。每个 $x_i\in U$ 可以表示为：

$$\underline{R}_X(x_i)=\inf_{x_j\in U}\max\{1-R(x_i,x_j),X(x_j)\} \tag{1}$$

$$\overline{R}_X(x_i)=\sup_{x_j\in U}\min\{1-R(x_i,x_j),X(x_j)\} \tag{2}$$

在模糊信息系统 (U,A) 中加入决策属性集 $D=\{d_i\}$，$A\cap D=\varnothing$，可以得到一个模糊决策系统 $(U,A\cup D)$，其中 A 为条件属性集，D 为决策属性集。则可以为 d 定义一个等价关系 R_D，由 R_D 将全域 U 划分为一组不相交子集 $U/R_D=\{[x_i]_D,x_i\in U\}$，其中 $[x_i]_D=\{x_j\in U,d(x_i)=d(x_j)\}$ 称为对象 x_j 所属的决策类。此外，x_j 的隶属函数为：

$$[x_i]_D(x_j)=\begin{cases}1, & x_j\in[x_i]_D\\ 0, & \text{其他}\end{cases} \tag{3}$$

当数据集的特征没有通过原始属性的组合而改变时，特征选择又可以叫做属性约简。一般情况下，模糊决策系统 $(U,A\cup D)$ 包含冗余属性，并且属性约简的目标是找到一个最小的条件属性子集来保存信息。

对于一个 $U=\{x_1,x_2,\cdots,x_n\}$ 的模糊决策系统 $(U,A\cup D)$，决策属性集 D 对于条件属性子集 B 的 $\lambda-$ 条件熵，即 $H_\lambda(D|B)$，有以下定义[[27]]：

$$H_\lambda(D|B)=-\frac{1}{|U|}\sum_{i=1}^{n}(\mathbb{I}|[x_i]_D\cap[x_i]_B^{\lambda_i}|\log\frac{|[x_i]_D\cap[x_i]_B^{\lambda_i}|}{|[x_i]_B^{\lambda_i}|}) \tag{4}$$

其中，$\lambda_i=\underline{R}_A[x_i]_D(x_i)$ 是 x_i 的模糊下近似值，$|\cdot|$ 是一个模糊集的基数，$[x_i]_B^{\lambda_i}(x_j)$ 是 x_i 相对于 B 的模糊粒子，其计算式为：

$$[x_i]_B^{\lambda_i}(x_j)=\begin{cases}\lambda_i, & 1-R_B(x_i,x_j)<\lambda_i\\ 0, & \text{其他}\end{cases} \tag{5}$$

随着新数据不断到来，将会在 U 的基础上添加一个新的样本集 $U_0=\{x_{n+1},\cdots,x_{n+n_0}\}$，而对于新的模糊决策系统 $(U\cup U_0,A\cup D)$，传统方法会根据式(4)计算加入增量数据后的决策属性集 D 相对于条件属性子集 B 的 $\lambda-$ 条件熵 $H_\lambda^{U\cup U_0}(D|B)$，这种方式在每次数据新增时都需要重新计算所有的历史数据。因此，为了减少重复计算量，将对信息熵进行增量计算，根据 Zhang 等[28]的研究，可以推导出新条件熵的转化形式：

$$H_\lambda^{U\cup U_0}(D|B)=\frac{1}{|U|+|U_0|}(|U|H_\lambda^U(D|B)+|U_0|H_\lambda^{U_0}(D|B))-\Delta \tag{6}$$

其中，$H_\lambda^U(D|B)$ 和 $H_\lambda^{U_0}(D|B)$ 用式(4)计算，Δ 的计算式为：

$$\Delta\approx\frac{1}{|U\cup U_0|^2}\sum_{j=1}^{m}\left(\sum_{i=1}^{n}|[x_j]_D\cap[x_i]_B^{\lambda_i'}|_U\log\frac{|[x_j]_D\cap[x_i]_B^{\lambda_i'}|_U}{|[x_i]_B^{\lambda_i'}|_U}+\sum_{i=n+1}^{n+n_0}\left(|[x_j]_D\cap[x_i]_B^{\lambda_i'}|_{U_0}\log\frac{|[x_j]_D\cap[x_i]_B^{\lambda_i'}|_{U_0}}{|[x_i]_B^{\lambda_i'}|_{U_0}}\right)\right) \tag{7}$$

其中，λ_i' 为加入最新样本后的模糊下近似，也可以进行增量计算，计算式为：

$$\lambda_i'=\begin{cases}\lambda_i\wedge\inf\limits_{\substack{x_k\in U\\ d(x_k)\neq d(x_i)}}\{1-R_A(x_i,x_k)\}, & x_i\in U\\ \lambda_i^0\wedge\inf\limits_{\substack{x_k\in U_0\\ d(x_j)\neq d(x_i)}}\{1-R_A(x_i,x_j)\}, & x_i\in U_0\end{cases} \tag{8}$$

根据式(6)，数据新增后的信息熵将由 3 部分组成：原始数据 U 已经计算过的 H_λ^U、由新数据 U_0 单独计算的条件熵 $H_{\lambda'}^{U_0}$ 和增量信息 Δ。这种增量信息熵的计算形式可以减少对历史数据的重复计算，只需要直接获取历史计算结果并对最近数据进行单独计算即可。另外，特征子集的搜索将采用后向消除的启发式搜索方式。通过对比各个特征子集的信息熵，可以选择出特征子集最小且信息熵变化值在阈值内的最佳特征子集。

3.2.2 代表性样本集与增量特征选择

由于增量式信息熵的近似计算会导致特征选择结果出现偏差，为了弥补增量信息量的不足，本文对历史样本中的代表性样本进行了选择，目的是选择历史数据中最具有代表性的小部分样本予以保留，用保留的历史数据信息来丰富增量信息。这样既能够增加数据的信息量，也无须担心保留所有原始数据带来的存储负担。

代表性样本的选择基于最近邻的思想，通过样本点分布的距离远近来度量样本之间相似性。如果几个样本之间距离足够小，则说明这些样本之间是相似的，可以选择其中的一个样本予以保留，将其余冗余的样本删去。具体的实现方式是最近邻聚类，通过距离度量，把数据分为多个不同的类别，如果样本与某一类的中心的距离最近，则分为该类。最后得到的聚类样本中心点将作为代表性样本集。构建好代表性样本集后，每次的增量特征选择将会把代表性样本集与新增数据合并以计算增量信息熵。最终得到的结合代表性样本集的增量特征选择算法（Sample Set and Incremental Wrapper Attribute Reduction，SSIWAR）的具体流程如算法 2 所示。

算法 2 结合代表性样本集的增量特征选择算法

输入：增量数据 S_i，决策属性集 D，代表性样本集 M，样本集容量 σ

输出：每个时刻的特征集合 B^i

1. $S=\{S_0,S_1,\cdots,S_t\}$，S_i 为时刻 i 到来的最新数据
2. M←通过聚类算法在 S_0 上选择具有代表性的样本集合
3. for i←1 to t do
4. $S'\leftarrow M\cup S_i$
5. 在数据 S' 中增量计算各个特征子集 B 的信息熵 $H_\lambda^{U\cup U_0}(D|B)$
6. 利用信息熵进行特征选择，得到最佳特征子集 B^i
7. 利用最近邻算法得到 S_i 中的代表性样本集 $M^{(i)}$
8. $M\leftarrow M\cup M^{(i)}$
9. if $|M|>\sigma$ then
10. 在 M 中删除最早加入的样本
11. end for

3.3 基于动态权重的集成学习分类模型

对于每个划分的时间窗口，对其内的数据进行训练，得到一个新的基分类器，基分类器可以用于解决数据中的类别不平衡问题。首先，基分类器会进行数据均衡化，具体的措施基于 UnderBagging[29]，对数据进行 bagging 迭代，每次迭代都对多数类进行欠采样操作，使训练数据的类别趋于平衡状态。处理完不平衡数据后，会在新数据上训练基分类器，通过集成方式将新基分类器加入到总分类器中，并通过调整基分类器的权重来适应数据中可能发生的概念漂移现象。

基分类器的权重调整将基于两个因素，即时间因素与性能因素。在时间方面，距离当前时间越远的基分类器越不符合当前的概念，因此其权重将降低。在性能方面，如果历史数据产生的基分类器的性能在最新数据上的表现欠佳，那么说明该基分类器在目前的数据概念上不适用，数据可能发生了概念漂移，因此性能较差的基分类器的权重也会降低。以下是基分类器权重动态更新的计算流程。

给定一个在时间区间$[0,t]$内获取到的贷款数据样本集合 $S_{0\sim t}=\{S_0,\cdots,S_t\}$。其中，$S_t$ 是时间 t 的一组数量为 n_t 的样本集合 $S_t=\{s_0,\cdots,s_{n_t}\}$，$s_i=(X_i,y_i)$是一个样本实例。模型对每次到达的数据都会训练一个基分类器并赋予其相应的权重，并且更新所有其他历史基分类器的权重。权重低于阈值 θ 的历史基分类器将被删去，其余基分类器则会被保留。

时刻 t 时，保留的权重大于 θ 的 m_t 个基分类器的集成可以表示为 $L^{(t)}=\{L_1^t,\cdots,L_{m_t}^t\}$。对于每个 L_i^t 对应的分类器权重，记为 $W^{(t)}=\{w_1^t,\cdots,w_{m_t}^t\}$，其中 w_i^t 表示在时刻 t 第 i 个分类器的权重，其计算和更新的公式为：

$$w_i^t=[1-b*\alpha_i^{(t)}-(1-b)*\beta_i^{(t)}] \tag{9}$$

其中，$\alpha_i^{(t)}$ 表示时刻 t 第 i 个分类器的性能评价参数，可以是错误率等评估函数，$\beta_i^{(t)}$ 是时间衰减参数，数据到达时间越早，其值越高。而 b 与 $1-b$ 为性能指标和时间指标的比重大小。w_i^t 是随着时间推移而不断改变的。

最后，当 $t+1$ 时刻的数据到来时，会在之前的分类器中加入新数据训练得到的基分类器，新形成一个分类器 L'，随之 $W^{(t)}$ 权重将会被重新计算，更新为 $W^{(t+1)}$。则最终 $t+1$ 时刻将会结合权重形成集成分类器 $L^{(t+1)}$，再用分类器 $L^{(t+1)}$ 就可以预测将会到达的数据 S_{t+1}：

$$a_x=\mathrm{sign}(\sum_{i=1}^{m}w_i^{(t)}L_i^{(t)}(x)) \tag{10}$$

其中，a 为样本 x 的预测值，且 $x\in S_{t+1}$，sign 为符号函数，具体计算式为：

$$\mathrm{sign}(z)=\begin{cases}1, & z>0\\ -1, & z\leqslant 0\end{cases} \tag{11}$$

利用集成分类器训练的整个动态机器学习模型(Dynamic Machine Learning, DML)的具体训练如算法 3 所示。

算法 3 DML 模型训练

输入：增量数据 S_t，权重阈值 θ，分类器性能权重 b

输出：基分类器集合 $L^{(t)}$，基分类器的权重 $w_j^{(t)}$

1. $S_t=\{s_0,\cdots,s_{n_t}\}$为时刻 t 到达的一组数量为 n_t 的样本集，$s_i=(X_i,y_i)$是一个样本实例
2. $L^{(0)}\leftarrow S_0$ 训练的初始基分类器上一时刻 t−1 的基分类器集合 $L^{(t-1)}$及其权重 $W^{(t-1)}$

   $L^{(t-1)}=\{L_1^{t-1},\cdots,L_m^{t-1}\}$

   $W^{(t-1)}=\{w_1^{t-1},\cdots,w_m^{t-1}\}$
3. for i←1 to n_t do
4. 通过集成模型 $L^{(t-1)}$ 预测 x_i 的类别：

   $a_x\leftarrow\mathrm{sign}(\sum_{i=1}^{m}w_i^{(t)}L_i^{(t)}(x))$
5. end for
6. for j←1 to m do
7. 计算每个基分类器 L_j^t 在 S_t 上的测试误差 $\beta_j^{(t)}$ 和时间衰减 $\alpha_j^{(t)}$
8. 时间因素占比为 1−b，更新基分类器的权重

   $w_j^t\leftarrow[1-b*\alpha_j^{(t)}-(1-b)*\beta_j^{(t)}]$
9. end for
10. 去除权重小于 θ 的基分类器：

    $L^{(t)}\leftarrow L^{(t-1)}\setminus\{L_j^{(t-1)}\mid w_j^t<\theta\}$
11. 对数据训练新基分类器，并初始化权重：

    L 训练的基分类器(S_t，M)

    $L^{(t)}\leftarrow L^{(t)}\cup L$，$w_j^{(t)}\leftarrow 1$，$m_t\leftarrow|L^{(t)}|$

4 实验研究与结果分析

4.1 主要数据集描述

首先，为了测试各种算法模型在信用数据上的表现，本文选取了两个真实的信贷数据集：Lending club 数据集及巴西银行贷款数据集。其次，为了验证增量特征选择模块的有效性，除了真实信用数据集，还从机器学习数据库 UCI[30] 的存储库中选取了 3 个高维数据集，进行特征选择算法的对比实验。另外，为了验证集成学习和动态模型对概念漂移的适应能力，还选取了 3 个公开的构造漂移数据集，进行概念漂移对比实验。数据集的基本信息如表 1 所列，具体描述如下。

(1)Lending Club Loan Data(LC)：该数据是是 P2P 借贷平台 Lending Club 的贷款数据，包括贷款者的信息与还贷情况。需要说明的是，由于原始数据量过大，本实验将从 2016−2018 年的每个月的数据中随机选取 2000 条数据，共 36 个数据块。

(2)CSDS：该数据集为巴西一家大型银行分析、批准和让出的 9.7 万笔贷款的相关信息，用于确定客户在 3 个月内是否违约。此数据集包含从 2013 年 10 月到 2015 年 1 月的贷款请求，每次选取 2000 条数据，共 48 个数据块。

(3)Hill-Valley(HV)：山谷特征数据集，数据集中包含“突起”“凹陷”两类，每个样本具有 100 个山势走向特征，通过分析山势特征来预测某一位置是“山”还是“谷”。

(4)Multiple Features(MFeat)：手写数字特征数据集，数据集包含“0”～“9”10 个类别，每个样本具有 649 个特征，通过分析数字特征来预测某一样本属于哪个数字。需要说明的是，为了符合本实验的二分类问题，实验中只选取了其中的“1”和“0”两个类别的样本。

(5)Hyper Plane(HP)：该数据集有 10 个属性，属性与类别存在一定的数学关系。

(6)Moving Gaussian(MG)：该数据集中包含两个特征，

由两个具有协同方差的高斯分布组成。

(7)Checkerboard(CB)：该数据集中包含 3 个特征，其中第三个特征属于噪声，与类别无关。

表 1 数据集的基本信息

Table 1 Statistics of data sets

Data Set	Number of Samples	Number of Chunks	Chunk Size
LC	72000	36	2000
CSDS	96000	48	2000
HV	606	6	101
MFeat	2000	6	333
HP	50000	25	2000
MG	50000	25	2000
CB	200000	100	2000

4.2 评价指标

针对二分类问题，可以将实例分成正类或负类。在使用模型对实例进行预测分类时，根据预测结果与真实结果出现的 4 种情况可以通过混淆矩阵来描述，如表 2 所列。

表 2 混淆矩阵

Table 2 Confusion matrix

Actual	Predict	
	True	False
True	TP	FN
False	FP	TN

ROC 曲线是一种常用于二元分类器的评价方式，横纵轴分别对应真正类率(True Positive Rate，TPR)和假正类率(False Positive Rate，FPR)，两者的计算式为：

$$TPR=\frac{TP}{TP+FN} \tag{12}$$

$$FPR=\frac{FP}{FP+TN} \tag{13}$$

AUC 是 ROC 曲线下的面积。由于 AUC 的计算中引入了真正类率和假正类率的概念，因此 AUC 值在不平衡的数据集中能很好地反映模型的性能，其具体计算式为：

$$AUC=\frac{\sum_{\text{positives}} r-\frac{n_{\text{pos}}(n_{\text{pos}}+1)}{2}}{n_{\text{pos}}n_{\text{neg}}} \tag{14}$$

其中，r 为数据集中预测为正样本的概率，N_{pos} 为数据集中正样本的数量，N_{neg} 为数据集中负样本的数量。

4.3 对比实验设置

实验 1 为了检验本文提出的增量特征选择算法(SSIWAR)的性能，本文利用公开高维数据集 HV，Mfeat 和真实信用数据集 LC，CSDS，与几个典型的增量、非增量式特征选择算法进行对比，对比算法包含增量式算法 FWAR，IARM[31] 和 FDNRS，以及非增量式算法 RFFS[32] 和 BGSO[33]。需要说明的是，本实验在数据集 LC 和 CSDS 中只采用了最新的 6 个数据块。

实验 2 将 3.3 节中提出的 DML 算法与主流的动态机器学习模型 REA[34]，AUE[35]，Learn++.NIE 进行对比，利用公开的构造漂移数据集 MG，CB，HP 和真实的信用数据集 LC，CSDS，验证单独的动态集成学习模型对概念漂移问题的解决能力。本实验不包含特征选择模块。

实验 3 将本文提出的整个信用评估模型 CEDML 与没有加入特征选择部分的 DML 算法以及主流的动态机器学习模型进行对比，利用真实的高维不平衡且可能存在概念漂移的信用评估数据 LC 和 CSDS 来验证所提出的整个信用评估模型框架的性能。

算法的描述如表 3 所列。

表 3 对比算法描述

Table 3 Descriptions of comparative algorithms

算法	主要思想
FWAR	SSIWAR 删除代表性样本集选择后的算法
IARM	基于相对辨别关系的增量特征选择算法
FDNRS	基于模糊优势邻域粗糙集的增量特征选择算法
RFFS	基于随机森林特征重要性排序的选择算法
BGSO	遗传算法与粒子群算法相结合的特征选择算法
AUE	基于集成决策树的数据流分类算法
REA	处理不平衡数据流的递归集成增量学习算法
Learn++	基于时间衰减权重的集成学习

同时，各个实验的具体设置如表 4、表 5 所列，其余对比算法中的设置将选择原论文中的最佳参数。

表 4 实验 1 的参数设置

Table53 Parameter settings of experiment 1

参数	含义	设置
L	基分类器	决策树
N_{chunk}	数据块数量	6
$\max\|S'\|$	代表性样本集的最大值	总数据量的 1/10

表 5 实验 2、实验 3 的参数设置

Table 5 Parameter settings of experiment 2 and 3

参数	含义	设置
$\mathscr{L}$	基分类器	决策树
n	每次固定增量的数据块大小	2000
q	滑动窗口中数据块的数量	5
θ	基分类器被移除的权重阈值	0.001

4.4 实验结果及分析

4.4.1 增量特征选择实验

本文提出的 SSIWAR 算法将与 FWAR，IARM，FDNRS，RFFS 和 BGSO 等算法在时间效率和 AUC 两个指标上进行对比。其中，在时间效率上 5 种算法之间的对比结果如图 3 所示，在 AUC 值上的对比结果如图 4 所示。表 6 列出了新数据到来的 5 个时刻，各种特征选择算法在每个数据集上的平均处理时间。表 7 列出了在每个数据集上的平均 AUC 值。

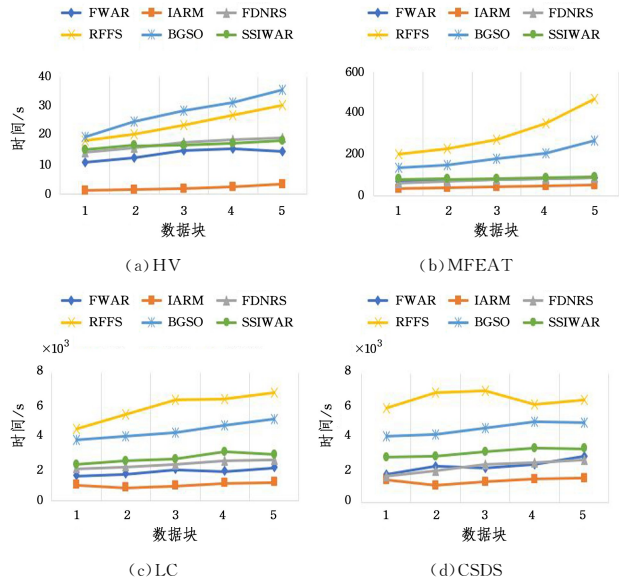
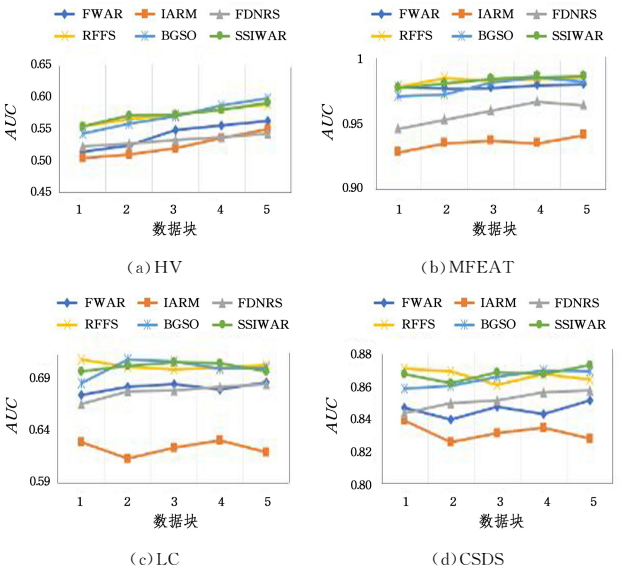


图 3 实验 1：时间效率对比

Fig. 3 Experiment 1：comparison of time efficiency



(a) HV (b) MFEAT

(c) LC (d) CSDS

图 4 实验 1:AUC 变化趋势

Fig. 4 Experiment 1: trend of AUC

表 6 实验 1:平均处理时间

Table 6 Experiment 1: average processing time

(单位:s)

算法	数据集			
	HV	Mfeat	LC	CSDS
FWAR	13.7	77.2	1803.8	2271.8
IARM	**2.2**	**41.9**	**981.6**	**1327.0**
FDNRS	27.9	75.0	2287.4	2214.8
RFFS	23.9	301.8	5850.4	6395.2
BGSO	17.1	186.1	4372.2	4574.2
SSIWAR	16.9	83.4	2681.4	3110.2

表 7 实验 1:平均 AUC 值

Table 7 Experiment 1: average value of AUC

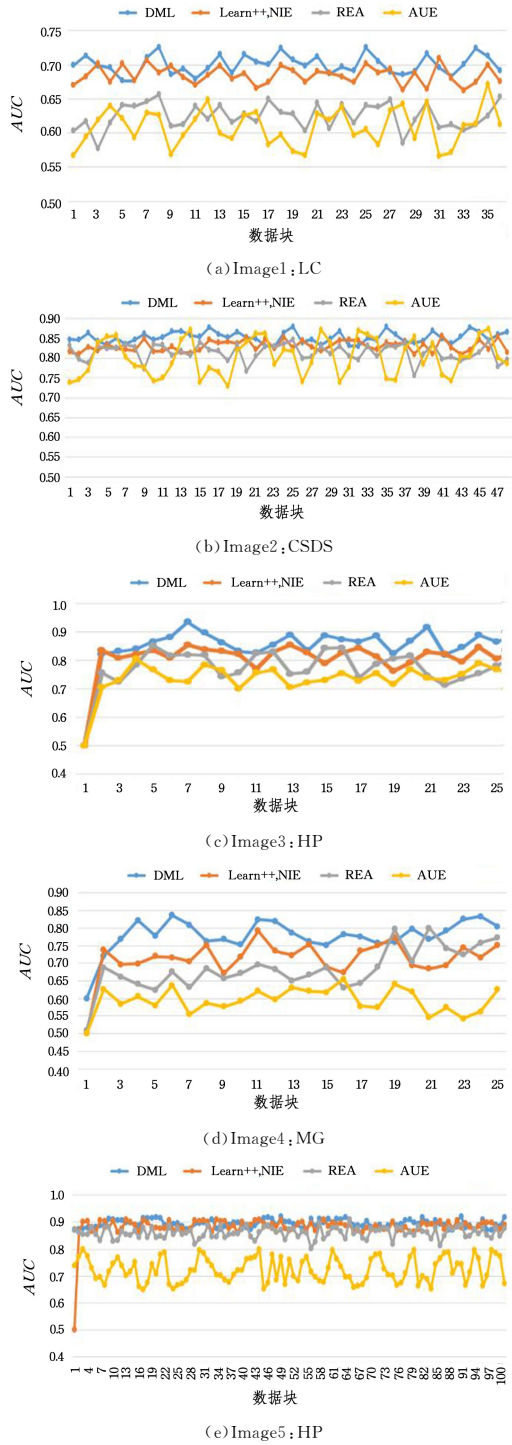
(单位:s)

算法	数据集			
	HV	Mfeat	LC	CSDS
FWAR	0.54040	0.97974	0.67912	0.84572
IARM	0.52336	0.93582	0.62282	0.83154
FDNRS	0.53212	0.95896	0.67556	0.85182
RFFS	0.57178	**0.98458**	**0.69962**	0.86672
BGSO	0.57042	0.97990	0.69740	0.86500
SSIWAR	**0.57280**	0.98454	0.69848	**0.86843**

根据实验结果可以看出:SSIWAR 与非增量的特征选择算法 RFFS 和 BGSO 相比,时间效率有所提高,并且在 AUC 上与 RFFS 算法的性能相似,比 BGSO 算法的效果更好;虽然结合代表性样本集选择的算法 FWAR 以及增量式特征选择算法 IARM 和 FDNRS 在时间效率上要略高于 SSIWAR 算法,但是在 AUC 上表现较差,远不及非增量的特征选择算法。因此,综合考虑时间和性能因素,本文提出的结合代表性样本集的增量特征选择算法的表现较好。

4.4.2 概念漂移适应实验

为了验证整个概念漂移适应算法的稳定性和 AUC,算法模型会对下一时刻的数据进行预测,并计算 AUC 值。图 5 给出了新数据到来的各个时刻,DML 算法与其他对比算法分别在各个数据集上的 AUC 值的变化。表 8 列出了所有时刻 AUC 的平均值。



(a) Image1: LC

(b) Image2: CSDS

(c) Image3: HP

(d) Image4: MG

(e) Image5: HP

图 5 实验 2:AUC 变化趋势

Fig. 5 Experiment 2: trend of AUC

表 8 实验 2:平均 AUC 值

Table 8 Experiment 2: average value of AUC

算法	数据集			
	DML	Learn++	REA	AUE
MG	**0.78402**	0.74551	0.71795	0.60285
CB	**0.89488**	0.88381	0.85946	0.72241
HP	**0.85969**	0.81016	0.78528	0.74126
LC	**0.69243**	0.68453	0.61029	0.60652
CSDS	**0.85457**	0.84251	0.83572	0.82845

首先,由曲线的波动可以清晰地看出,本文提出的 DML 算法在具有概念漂移的数据中有较好的稳定性,当数据具有

明显的概念漂移现象时，模型的效果也不会受到较大的影响，在概念漂移中的 AUC 都能保持较高的水平。而其他算法都不同程度地受到了较大的影响，其中 AUE 算法因为没有考虑数据的不平衡问题，所以总体的 AUC 值表现是最差的，而 Learn++.NIE 算法的性能是最接近 DML 算法的，原因是 Learn++.NIE 算法中对基分类器的权重调整考虑了时间因素，但没有考虑基分类器的性能因素，并且不是所有的历史基分类器都比新的基分类器的效果差，因此 DML 的集成权重加入了对基分类器错误率的衡量，这将会提升集成分类器的性能。

4.4.3 动态信用评估模型实验

为了考察总体信用评估算法框架在实际信用数据上的表现，将 CEDML 算法与不进行特征选择的 DML 算法以及其他主流算法进行对比，表 9 列出了各算法分别在数据集 LC 和 CSDS 上的平均 AUC 值。为了考察整个模型的稳定性，将会对模型进行重复实验，计算每个模型平均 AUC 值的标准差，并将 CEDML 算法与其他对比算法（包括不进行均衡化和特征选择操作的 DML 算法）的 AUC 标准差进行对比。表 10 列出了各算法分别在数据集 LC 和 CSDS 上的 AUC 标准差。

表 9 实验 3：平均 AUC 值

Table 9 Experiment 3: average value of AUC

算法	数据集				
	CEDML	DML	Learn++	REA	AUE
LC	**0.70782**	0.69243	0.68453	0.61029	0.60652
CSDS	**0.86223**	0.85457	0.84251	0.83572	0.82845

从实验结果可以看出，在平均 AUC 值上，本文提出的 CEDML 算法在真实信用评估数据上的表现比主流算法 Learn++.NIE，REA，AUE 等都更好，并且通过计算每次重复实验的 AUC 标准差，可以看到，CEDML 和 DML 算法的 AUC 稳定性也比其他算法更好。同时，将 CEDML 与没有加入数据特征选择模块的 DML 算法进行对比，可以看到，加入了特征选择模块后，模型的整体平均 AUC 值也有了一定的提升，并且在两个数据集上的稳定性高于 DML 算法，这也验证了特征选择模块的必要性。

表 10 实验 3：AUC 的标准差（$\times 10^{-4}$）

Table 10 Experiment 3: standard deviation of AUC($\times 10^{-4}$)

算法	数据集				
	CEDML	DML	Learn++	REA	AUE
LC	**102.85**	123.96	282.15	246.86	465.23
CSDS	**145.45**	156.47	234.56	189.23	396.46

结束语 本文构建了一个动态的信用评估模型，能够根据最新到来的数据不断更新，在处理好类别不平衡问题和进行高效准确特征选择的基础上，使其适应了概念漂移现象。实验结果表明，在公开数据集和真实信用数据上，本文提出的结合代表性样本选择的增量特征选择算法和面向信用评估的动态增量集成模型在时间和性能上都有不错的表现，能够较好地处理高维不平衡数据和概念漂移问题。未来将进一步深入自动化特征工程和深度学习的动态信用评估等其他动态信用评估模型的研究。

参 考 文 献

[1] YUAN Y, GONG X, GUO M, et al. Research on Personal Credit Evaluation of Commercial Banks Under Ensemble Learning Framework[C]// 2020 2nd International Conference on Applied Machine Learning(ICAML). IEEE, 2020: 29-38.

[2] LU J, LIU A, DONG F, et al. Learning Under Concept Drift: A Review[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(12): 2346-2363.

[3] KRAWCZYK B. Learning from Imbalanced Data: Open Challenges and Future Directions[J]. Progress in Artificial Intelligence, 2016, 5(4): 221-232.

[4] ARYA S, ECKEL C, WICHMAN C. Anatomy of the Credit Score[J]. Journal of Economic Behavior & Organization, 2013, 95: 175-185.

[5] DONG G, LAI K K, YEN J. Credit scorecard based on logistic regression with random coefficients [J]. Procedia Computer Science, 2010, 1(1): 2463-2468.

[6] HAND D J, HENLEY W E. Statistical Classification Methods in Consumer Credit Scoring: A Review [J]. Journal of the Royal Statistical Society, 1997, 160(3): 523-541.

[7] DANENAS P, GARSVA G. Selection of Support Vector Machines Based Classifiers for Credit Risk Domain [J]. Expert Systems with Applications, 2015, 42(6): 3194-3203.

[8] HARRIS T. Credit Scoring Using the Clustered Support Vector Machine [J]. Expert Systems with Applications, 2015, 42(2): 741-750.

[9] ONG C S, HUANG J J, TZENG G H. Building Credit Scoring Models Using Genetic Programming [J]. Expert Systems with Applications, 2005, 29(1): 41-47.

[10] WEST D. Neural Network Credit Scoring Models [J]. Computers & Operations Research, 2000, 27(11): 1131-1152.

[11] SUN J, LANG J, FUJITA H, et al. Imbalanced Enterprise Credit Evaluation with DTE-SBD: Decision Tree Ensemble Based on SMOTE and Bagging with Differentiated Sampling Rates [J/OL]. Information Sciences, 2018, 425: 76-91. https://www.sciencedirect.com/science/article/pii/S0020025517310083.

[12] ZHANG W, HE H, ZHANG S. A Novel Multi-stage Hybrid Model with Enhanced Multi-Population Niche Genetic Algorithm: An Application in Credit Scoring[J/OL]. Expert Systems with Applications, 2018, 121: 221-232. https://www.sciencedirect.com/science/article/pii/S0957417418307887.

[13] BARDDAL J P, LOEZER L, ENEMBRECK F, et al. Lessons Learned From Data Stream Classification Applied to Credit Scoring [J/OL]. Expert Systems with Applications, 2020, 162: 113899. https://www.sciencedirect.com/science/article/pii/S0167268111001259.

[14] CAI Y, JIANG Y. Credit Scoring Using Incremental Learning Algorithm for SVDD[C]// 2016 International Conference on Computer, Information and Telecommunication Systems (CITS). IEEE, 2016: 1-4.

[15] PONTIL M, VERRI A. Properties of Support Vector Machines [J]. Neural Computation, 1998, 10(4): 955-974.

[16] TAX D M J, DUIN R P W. Support Vector Data Description

[J]. Machine learning, 2004, 54(1): 45-66.
[17] TIAN J, LIU X, LI M. An Incremental Learning Ensemble Method for Imbalanced Credit Scoring[C]// 2019 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2019: 754-759.
[18] VENKATESH B, ANURADHA J. A Review of Feature Selection and Its Methods[J]. Cybernetics and Information Technologies, 2019, 19(1): 3-26.
[19] GUYON I, ELISSEEFF A. An Introduction to Variable and Feature Selection[J]. Journal of Machine Learning Research, 2003, 3(5): 1157-1182.
[20] SHU W, QIAN W, XIE Y. Incremental Feature Selection for Dynamic Hybrid Data Using Neighborhood Rough Set[J/OL]. Knowledge-Based Systems, 2020, 194: 105516. https://www.sciencedirect.com/science/article/pii/S0950705120300289.
[21] SANG B, CHEN H, YANG L, et al. Incremental Feature Selection Using a Conditional Entropy Based on Fuzzy Dominance Neighborhood Rough Sets[J]. IEEE Transactions on Fuzzy Systems, 2021, 30(6): 1683-1697.
[22] ŽLIOBAITE I, PECHENIZKIY M, GAMA J. Big Data Analysis: New Algorithms for a New Society[M]. Cham, Switzerland: Springer International Publishing, 2016: 91-114.
[23] ELWELL R, POLIKAR R. Incremental Learning of Concept Drift in Nonstationary Environments[J]. IEEE Transactions on Neural Networks, 2011, 22(10): 1517-1531.
[24] ZHANG S, LIU J, ZUO X. Adaptive Online Incremental Learning for Evolving Data Streams[J/OL]. Applied Soft Computing, 2021, 105: 107255. https://www.sciencedirect.com/science/article/pii/S1568494621001782.
[25] LI Z, HUANG W, XIONG Y, et al. Incremental Learning Imbalanced Data Streams with Concept Drift: The Dynamic Updated Ensemble Algorithm[J/OL]. Knowledge-Based Systems, 2020, 195: 105694. https://www.sciencedirect.com/science/article/pii/S095070512030126X.
[26] DUBOIS D, PRADE H. Rough Fuzzy Sets and Fuzzy Rough Sets[J]. International Journal of General System, 1990, 17(2/3): 191-209.
[27] ZHANG X, MEI C, CHEN D, et al. Feature Selection in Mixed Data: A Method Using a Novel Fuzzy Rough Set Based Information Entropy [J/OL]. Pattern Recognition, 2016, 56: 1-15. https://www.sciencedirect.com/science/article/pii/S0031320316000844.
[28] ZHANG X, MEI C, CHEN D, et al. Active Incremental Feature Selection Using a Fuzzy-Rough-Set-Based Information Entropy [J]. IEEE Transactions on Fuzzy Systems, 2019, 28(5): 901-915.
[29] BARANDELA R, VALDOVINOS R M, SÁNCHEZ J S. New Applications of Ensembles of Classifiers[J]. Pattern Analysis & Applications, 2003, 6(3): 245-256.
[30] CHANG S, SHIHONG Y, QI L. Clustering Characteristics of UCI Dataset [C] // 2020 39th Chinese Control Conference (CCC). IEEE, 2020: 6301-6306.
[31] YANG Y, CHEN D, WANG H, et al. Fuzzy Rough Set Based Incremental Attribute Reduction from Dynamic Data with Sample Arriving[J/OL]. Fuzzy Sets and Systems, 2017, 312: 66-86. https://www.sciencedirect.com/science/article/pii/S0167404820301231.
[32] LI X K, CHEN W, ZHANG Q, et al. Building Auto-Encoder Intrusion Detection System Based on Random Forest Feature Selection [J/OL]. Computers & Security, 2020, 95: 101851. https://www.sciencedirect.com/science/article/pii/S0167404820301231.
[33] GHOSH M, GUHA R, ALAM I, et al. Binary Genetic Swarm Optimization: A Combination of GA and PSO for Feature Selection[J]. Journal of Intelligent Systems, 2020, 29(1): 1598-1610.
[34] CHEN S, HE H. Towards Incremental Learning of Nonstationary Imbalanced Data Stream: A Multiple Selectively Recursive Approach[J]. Evolving Systems, 2011, 2(1): 35-50.
[35] SUN Y, TANG K, MINKU L L, et al. Online Ensemble Learning of Data Streams with Gradually Evolved Classes[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(6): 1532-1545.

CHEN Yijun, born in 2000, postgraduate. Her main research interests include data mining and machine learning.

DING Zhijun, born in 1974, Ph.D, professor, Ph. D supervisor, is a senior member of China Computer Federation. His main research interests include intelligent software engineering, cloud computing and services, big data credit reporting and financial risk control.

(责任编辑:喻藜)