ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 1,Jan. 2022,pp. 183－193
DOI:10. 16337/j. 1004-9037. 2022. 01. 016
© 2022 by Journal of Data Acquisition and Processing
http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于核极限学习机的多标签数据流集成分类方法

张海翔[1,2]，李培培[1,2]，胡学钢[1,2]

(1. 大数据知识工程教育部重点实验室(合肥工业大学),合肥 230601;2. 合肥工业大学计算机与信息学院,合肥 230601)

摘　要：极限学习机因具有高效处理、性能优越以及更少人工参数设定等优点,已成功应用于批处理多标签分类问题。然而,实际应用领域涌现的数据流呈现海量快速、多标签和概念漂移等特点,使得这些传统的多标签分类算法面临精度与时空的挑战。本文提出一种基于核极限学习机的多标签数据流集成分类方法。首先,为适应数据流环境,利用滑动窗口机制将数据流划分为数据块,在前k个数据块上构建k个核极限学习机的集成分类模型;同时,考虑类标签相关性,利用Apriori算法得到每个数据块的标签间的关联规则,并将关联规则中的同现标签的置信度引入到基于集成模型的预测过程中,以提高整体的分类精度;其次,引入MUENLForeset模型检测新到来的数据块是否发生概念漂移,对分类器设置损失函数更新集成模型以适应概念漂移问题。最后,在实际多标签数据上的大量实验表明:与经典多标签批处理和流数据分类方法相比,所提方法不仅能适应多标签数据流中的概念漂移问题,同时在分类精度上具有显著优势。

关键词：多标签分类;数据流;核极限学习机;标签相关性;概念漂移

中图分类号：TP183　　**文献标志码**：A

Multi-label Data Stream Ensemble Classification Approach Based on Kernel Extreme Learning Machine

ZHANG Haixiang[1,2], LI Peipei[1,2], HU Xuegang[1,2]

(1. Key Laboratory of Big Data Knowledge Engineering Ministry of Education (Hefei University of Technology), Hefei 230601, China;2. School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601,China)

Abstract: Extreme learning machine has a series of achievements on batch processing due to high-activity processing, superior performance, less manual parameter settings and so on, which has been successfully applied in multi-label classification. However, data streams emerging in the real-world applications present the characteristics of high-volume, high-speed, multi-label and concept drift, which poses the challenges in accuracy, time and space consumptions for traditional multi-label classification algorithms. Therefore, this paper proposes a multi-label classification data stream ensemble approach based on kernel extreme learning machine (KELM). Firstly, to adapt to the environment of data streams, the sliding window mechanism is used to partition data chunks, and an ensemble model consisted of k KELM models is built on k data chunks. Meanwhile, considering the label correlation, the Apriori algorithm is used to achieve the association rules of labels, and the confidence of label occurrence is introduced in the prediction using the

基金项目:国家自然科学基金(61976077, 62076085)。
收稿日期:2020-07-12;**修订日期**:2020-11-11

generated model. Secondly, the MUENLForest model is introduced to detect whether a concept drift occurs in the new arriving data chunk, correspondingly the loss function is specified to update the ensemble model for adapting to concept drifts. Finally, massive experiments on the real multi label data sets demonstrate that the proposed approach outperforms the traditional multi label classification methods in accuracy and can adapt data drifts in multi label data streams quickly.

Key words: multi-label classification; data stream; kernel extreme learning machine; label correlation; concept drift

引　言

为了克服传统单标签分类的缺陷,多标签分类(Multi-label classification, MLC)[1],即一个事物对应多个类标签概念的研究变得尤为重要。在实际应用领域中多标签数据流呈现出海量快速、概念漂移等特点,使得传统多标签分类算法无法直接解决此类问题。因而,如何在有限的时间和内存下快速处理这些新到来的数据,并适应数据流环境下的概念漂移等,设计鲁棒的多标签数据流分类方法成为重要而具有挑战的任务之一。

目前,已有的多标签分类方法主要包括:批处理方法和在线学习方法[1]。其中批处理方法默认每次训练与测试的数据集一次性到来,根据已有全部信息采用问题转化、算法自适应[2]解决多标签分类问题。Huang[3-4]提出的极限学习机(Extreme learning machine, ELM)及其改进算法[5]具有高速和高效等特点,避免了繁琐的迭代学习过程以及传统前馈神经网络的迭代学习引起的学习参数随机设置、容易陷入局部最小值等问题,同时改进算法能进一步提高分类精度。因此,基于(核)极限学习机的相关研究被广泛应用于多标签分类问题,并取得了一系列的成果[6-10]。然而,实际应用领域涌现的数据流由于具有海量快速等特点,难以一次性全部获取。同时,当新数据到来时这些批处理算法不断对新数据重新训练而抛弃旧模型,导致有效历史数据的大量丢失,因此能够处理数据流环境下的学习模型也越来越受到重视[11]。目前已有一些成果[12]采用滑动窗口技术将极限学习机应用解决数据流多标签分类,但该方法未考虑多标签间的类标签相关问题以及数据流环境下的概念漂移等问题。另一方面,文献[13]指出在处理数据流时需要考虑模型在有限的时间和内存下做出精准预测并包含应对概念漂移问题的解决方案。这些需求为多标签数据流分类带来更多的挑战。数据流环境下的多标签分类算法[14]大多采用问题转化,将分类转化为一系列稳定的学习任务,虽然在一定程度上该方法能够适用,却忽略了标签之间的相关性[15]。同时未考虑到新到来的数据中高速、多变特性,而且其中隐含的概念漂移问题也是问题转化方法难以解决的[16]。

1　相关工作

本节将简要概述基于ELM的多标签分类方法与多标签数据流分类方法。

1.1　基于ELM的多标签分类方法

Huang等[3-4]提出的ELM在一次学习中就可得到一个恰当的解,避免了如误差反向传播算法(Back propagation,BP)[17]等基于梯度下降的复杂耗时方法。因而利用ELM学习速度快、实验效果更好的性能,在多标签分类问题上将ELM作为训练模型成为一种新的研究方向。文献[18-20]利用给定数据集将ELM应用多标签分类问题。然而,一方面由于ELM的隐藏层节点设置的随机性会引起隐藏层输出矩阵的振荡,从而降低网络结构的稳定性。另一方面,考虑到ELM隐藏层将输入样本映射至线性可分的空间,该映射过程与内核函数的内积运算将特征向量从高维映射到低维空间原理一致[5]。因此,Luo等[10]提出基于核极限学习机(Kernel extreme learning machine,KELM)的多标签分类方法ML-KELM,

相较于ELM,ML-KELM只需确定内核函数和相关参数,就可得到稳定结果。针对高维数据环境,Lin等[21]提出多标签核判别分析方法,利用核函数技术整体处理多标签并实现非线性降维。为了解决神经网络中的过拟合问题,Zhang等[22]提出一种基于径向基函数(Radial basis function, RBF)的多层ELM网络模型用于多标签分类问题ML-ELM-RBF。Kongsorot等[23]提出基于模糊集理论的增量核ELM方法,将实例及其对应类之间的关系定义为模糊成员。Law等[24]提出一种用于多标签数据分类的级联神经网络,将堆叠式自动编码器(Stacked auto-encoder, SAE)和ELM合并协作。Wang等[25]利用标签相关性和非平衡参数得到非平衡标签补全矩阵,将其与核极限学习机进行联合学习。上述方法探索了在多标签分类问题上ELM模型的应用,并取得了一定的成果。然而,这些方法都是批处理方法,难以直接应用于海量快速的数据流分类问题。

1.2 多标签数据流分类方法

为解决多标签数据流分类问题,已有方法多采用问题转化与算法适应的策略[26]。其中,基于问题转化策略相关工作包括:Qu等[27]提出基于二元相关(Binary relevance, BR)的多标签数据流分类方法,采用增量批处理技术,其模型在顺序到来的同等大小数据块上学习。Xioufis等[28]采用BR通过将多标签任务转换为若干二进制分类任务来解决MLC,通过为每个标签维护2个可变大小窗口来处理概念漂移。文献[29]提出增量多标签决策树方法,将Hoeffding树[30]转化为适应数据流多标签分类;多标签Hoeffding树与Pruned Set分类器[31]合并,当此节点中获取样本的缓冲区已满时,修剪每个叶节点处的标签组合;此外Hoeffding树还与ADWIN Bagging[32](EaHTps)结合以解决概念漂移问题。Shi等[33]使用Apriori和集成方法(Ensemble methods,EM)算法将类标签集基于依赖性划分为不同的子集,这些子集被视为新类标签,用于注释每个到达的样本;同时,提出一种基于标签组合与阈值的概念漂移检测方法。然而上述方法忽略标签间的相关性,造成分类精度较差。

基于算法适应策略相关工作包括:Shi等[34]通过动态识别新频繁标签组合并更新标签组合集方法解决类标签相关分析问题;Osojnik等[14]将多目标回归应用于数据流的多标签学习,但该方法仅侧重于学习静止概念问题;Nguyen等[35]提出一种基于贝叶斯的多标签数据流学习方法,可以从每个真实标签的样本中学习标签间相关性,并根据霍夫丁不等式和标签基数来调整预测标签的数目,通过“未确定的值”方法扩展了标签特征值表来解决缺失值问题。此外,作者进一步提出基于加权聚类模型的增量在线多标签分类方法[36],利用衰减机制来适应概念漂移。虽然文献[35-36]利用模型本身优势学习标签间的相关性,并引入损失函数降低历史数据影响以适应概念漂移,但在计算特征与标签间联合分布过程中消耗过多时间。

2 数据流多标签分类集成方法

首先给出数据流多标签分类问题的定义:给定一个多标签数据流D,根据滑动窗口机制,将所述多标签数据流D等分成n个数据块集合$D=\{D_1, D_2, \cdots, D_k, \cdots, D_N\}, k=1, 2, \cdots, N$, 其中$D_k$为所述多标签数据流$D$中的第$k$个数据块,$D_k=\{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \cdots, (\boldsymbol{x}_i, \boldsymbol{y}_i), \cdots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$表示所述多标签数据流$D$中的第$k$个数据块$D_k$中的第$i$个多标签示例,$x_i \in \mathbf{R}^m$表示样本$m$维特征空间,$\boldsymbol{y}_i$表示所述第$k$个数据块$D_k$中的第$i$个多标签示例的类标签,满足$\boldsymbol{y}_i \in Y$,$Y$表示标签空间中包含$L$个不同标签,记为$Y=\{l_1, l_2, \cdots, l_L\}$。在线多标签分类器任务是学习从多标签数据流块中找到其实例的类标签,即$f_{\sum_k D_k}: \boldsymbol{x} \to 2^M, \boldsymbol{x}_i \in \boldsymbol{x}, \boldsymbol{y}_i \in \boldsymbol{y}$。对于新到来未知标签数据块$D_{k+1}$中的样本$\boldsymbol{x}_j \in D_{k+1}$,分类器$f(\cdot)$预测$f(\boldsymbol{x}_j) \subseteq Y$作为它的可能标签集合。

本文所提方法采用增量批处理技术,算法分为4个步骤:(1)初始假设选取前k个数据块构成基分类

器集合 $D=\{D_1,D_2,\cdots,D_k\}$;(2)根据已有的 k 个训练数据分别分析其内部类标签关系,得到关联规则,并构建MUENLForeset概念漂移检测机制;(3)利用得到的关联规则构建基于核的极限学习机KELM的多标签数据流集成分类模型OS-KELM,然后对于新到来的数据块 D_{k+1},先利用训练好的模型和关联规则输出预测结果,并将新数据块替换基分类器中效果最差的数据块数据;(4)在预测过程中判断 D_{k+1} 是否发生漂移,若发生漂移则对基分类器数据块引入权重损失函数降低旧数据的贡献程度。算法框架流程图如图1所示。

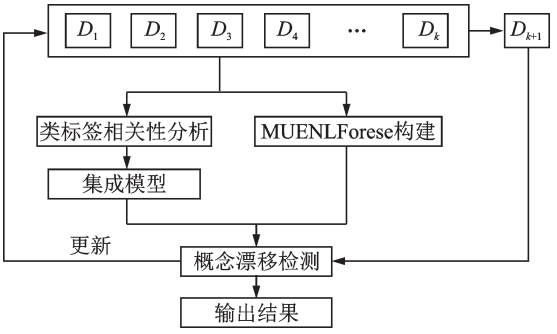

图1 本文方法整体框架图

Fig.1 Framework of the proposed method

2.1 基于Apriori算法的类标签相关性分析

在多标签分类过程中,样本实例与多个标签相对应并且标签集合中可能存在标签关联[34-37]性质,即标签数据集合中存在一种关联关系使得一个标签属于该样本隐含着另一个标签也属于该样本。通过找到这些成对的标签间关系并在预测过程中引入以提高整体的分类精度。基于上述分析,提出基于Apriori关联规则算法的类标签相关性分析策略。

针对到来的每个数据块 D_i,在利用 D_i 中所包含信息训练KELM模型之前,对其标签集 $Y_i\subseteq Y$ 采用Apriori算法[37]计算此数据块标签集中所蕴含的标签间关联规则集合rules,根据关联规则找到所有满足置信度的成对标签,将同现标签的置信度引入到基于集成模型的预测过程中以提高整体的分类精度。本文Apriori算法支持度设置为0.3,置信度为0.6。

2.2 基于核极限学习机的集成模型构建与预测

随着前 k 个数据块 $D_i(1\leqslant i\leqslant k)$ 的到来,分别构建核极限学习机,针对第 i 个数据块 $D_i=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}$,特征向量可表示为 $m\times n$ 的矩阵,m 表示将每一数据块中特征维度,n 表示数据块中实例个数,所有实例的类标签分布表示为 $Y_i=\{\boldsymbol{y}_i\}$。由特征映射形成隐层矩阵 $\boldsymbol{h}_i(\boldsymbol{x})=[\boldsymbol{h}_{i1}(\boldsymbol{x}),\boldsymbol{h}_{i2}(\boldsymbol{x}),\cdots,\boldsymbol{h}_{in}(\boldsymbol{x})]$,由文献[10]得到关于ELM的数学模型和KTT条件得到该数据块的ELM模型输出,即

$$\boldsymbol{\beta}_i=\boldsymbol{H}_i^{\mathrm{T}}\left(\frac{\boldsymbol{I}}{\boldsymbol{C}}+\boldsymbol{H}_i\boldsymbol{H}_i^{\mathrm{T}}\right)^{-1}Y_i \tag{1}$$

式中:$\boldsymbol{H}$ 表示训练数据隐层输出矩阵;Y 为训练数据的标签集合;$\boldsymbol{C}$ 和 $\boldsymbol{I}$ 分别表示岭回归参数和单位矩阵。最后采用核函数 $\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}(i,j)=K(\boldsymbol{x}_i,\boldsymbol{x}_j)$,$(\boldsymbol{\Omega}_{\mathrm{ELM}})_{ij}=\boldsymbol{h}(\boldsymbol{x}_i)\cdot\boldsymbol{h}(\boldsymbol{x}_j)=K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 代替ELM的隐层映射使神经网络结构趋于稳定思想得到OS-KELM输出模型为

$$f_i(\boldsymbol{x})=\boldsymbol{H}_i\boldsymbol{\beta}_i=\boldsymbol{H}_i\boldsymbol{H}_i^{\mathrm{T}}\left(\frac{\boldsymbol{I}}{\boldsymbol{C}}+\boldsymbol{H}_i\boldsymbol{H}_i^{\mathrm{T}}\right)^{-1}Y_i=\begin{bmatrix}K(\boldsymbol{x},x_{i1})\\ \vdots\\ K(\boldsymbol{x},x_{im})\end{bmatrix}^{\mathrm{T}}\left(\frac{\boldsymbol{I}}{\boldsymbol{C}}+\boldsymbol{\Omega}_{i\mathrm{ELM}}\right)^{-1}Y_i \tag{2}$$

式中:$\boldsymbol{\Omega}$ 为核函数矩阵。本文采用径向基核函数

$$K(\boldsymbol{x},\boldsymbol{x}_i)=\exp\left(-\frac{\left\|\boldsymbol{x}-\boldsymbol{x}_i\right\|^2}{2\sigma^2}\right) \tag{3}$$

式中 σ 为径向基函数。初始设定基分类器个数为 k,对其中每一个数据块均做以上处理后得到基分类器集成模型为 $f=\{f_1(\boldsymbol{x}),f_2(\boldsymbol{x}),\cdots,f_k(\boldsymbol{x})\}$。

当第$k+1$个数据块到来时,集成模型f对该数据块中的每个实例x进行预测

$$P(l_j|\boldsymbol{x})=f(l_j|\boldsymbol{x})+\sum_{i\neq j}f(l_i|\boldsymbol{x})\cdot\mathrm{conf}(l_i\Rightarrow l_j)\quad l_i,l_j\in Y$$

$$Y^*=\{l_j|P(l_j|\boldsymbol{x})>\tau\}\tag{4}$$

式中Conf为标签间置信度。根据式(4)计算结果$P(l_j|\boldsymbol{x})$,若大于τ,表示该标签属于此样本,反之不属于(本文设定阈值$\tau=0$),最后所有满足阈值的类标签集作为当前实例$\boldsymbol{x}$的类标签$\boldsymbol{Y}^*$。

2.3 基于MuENLForest模型的概念漂移检测与模型更新

为处理多标签数据流中由于数据分布变化引起的概念漂移问题,本文引入MuENLForest模型[16],通过检测新数据特征对应的标签数据分布发生变化来判断概念漂移是否发生。随着前k个数据块的到来,所提方法会相应构建k个MuENLTree决策树模型组成MuENLForest决策森林模型。其中每个MuENLTree决策树模型是由内部节点和叶节点组成的二叉树,令$a=[X_k, Y_k]$表示第k个具有预测值的训练样本。每个内部节点以下划分策略:$\left\|a^q-p_1\right\|\leqslant\left\|a^q-p_2\right\|$分为2个子节点,其中$p_1$和$p_2$是2个具有$q$属性的聚类中心,$a^q$为$a$的$q$投影,$q=[q_1,q_2]$,$q_1$和$q_2$分别是从输入样本属性集和特征样本属性集中随机选择的K个属性,每个叶子节点定义半径为$r=\max_{x\in S}\|a-m\|$的球覆盖S(即属于该叶子节点的所有训练实例集合),其中$m=\mathrm{mean}(S)$。为了在训练过程中生成MuENLTree,在构建内部节点之前递归地划分训练集,直到满足以下任何条件:(1) 树达到限制高度h;(2)$|S|=1$;(3) S中的所有实例具有相同的x^q值。对于MuENLForest的构建,采用其预测值扩充每个实例,以便可以同时考虑特征和标签信息。 投影在一组随机选择的属性上,每个内部节点基于任一分支上的群集中心进行拆分,结果同一叶节点内的实例在要素或预测值或两者的某些属性上必须相似。

构造好MuENLForest后,当第$k+1$个数据块到来时,需要对该数据块中的每个未知实例进行类分布变化的检测。当新实例到来后检测结果落在球外即大于半径r,则认为当前的实例数据分布发生变化。统计新到来数据块中实例发生概念漂移的个数是否满足阈值,若满足,则认为发生概念漂移,并对基分类器所有数据块设置损失函数2^{-et}以降低其权重,削弱旧数据的影响程度,同时新的数据块替代效果最差预测对应数据块,构建新的基分类器C',设置新的数据块所在位置权重为1;若不满足阈值,则表示无概念漂移发生,不引入损失函数,仅进行更新操作构建新的基分类器C',所有分类器权重设置为1。

2.4 时间复杂度分析

本文算法在样本(X, Y)上的训练过程复杂度为$O(6\times\max\begin{Bmatrix}O\left(\sum|Q|\times n\right),O(n^3),\\O\left(|q|ghn\right)\end{Bmatrix}$,其中$|q|$表示构建MuENLTree时随机选择分割属性的个数;Q表示候选项目组成的集合;n表示数据块的实例个数;h表示构建MuENLTree的最大树高;g表示构建的MuENLTree的个数。Apriori算法时间复杂度为$O\left(\sum_{c\in Q}(g(c)+s(c))\right)$,$g(c)$表示生成每个候选项目的复杂度,$s(c)$表示计算候选项目的复杂度,由于计算每一个项目的支持度都需扫描数据库使$s(c)\gg g(c)$,因而上式复杂度表示为$O\left(\sum_{c\in C}(s(c))\right)$,每计算一个$c$的支持度都需要扫描当前数据块中单个实例数据,假定每个数据块中实例数为n,则上式又可以表示为$O\left(\sum|Q|\times n\right)$。OS-KELM的计算过程主要体现在使用核函数$K(u,v)$代替ELM的隐层映射,使网络结构趋于稳定,其训练时间为样本数N的3次幂,因而时间复杂度为$O(n^3)$。概念漂移检测时间消耗在检测器MuENLForest的构建过程,每个节点都涉及k个均值聚类,其中包含2个具有时间复杂度$O(|q|n)$的聚类。此外还涉及n个实例引入树高度h限制,因此对于带有g个MuENLTrees的MuENLForest时间复杂度为$O(|q|ghn)$。

3 实验及其结果分析

3.1 实验数据集与评价指标

实验选择10个广泛用于多标签分类的真实数据集。表1总结了10组数据集的数据规模、属性维、标签个数和标签基数。

表1 数据集

Table 1 Datasets

数据集	数据规模	数据维	标签个数	标签基数
20NG	19 300	1 001	20	1.100
IMDB	120 919	1 001	28	2.000
OHSUMED	13 929	1 002	23	1.700
SLASHDOT	3 782	1 079	22	1.200
TMC2007	28 596	500	22	2.200
Enron	1 702	1 001	53	3.378
Yeast	2 417	103	14	4.237
Scene	2 407	294	6	1.070
Corel5k	5 000	499	374	3.520
Medical	978	1 449	45	1.250

本文的算法评价指标分类两类[1]：基于实例的评价指标(Example-based metrics)与基于类标签排名的指标(Label ranking-based)。其中前者包括Hamming_loss、Accuracy与F_1-measure；后者包含Coverage、Ranking loss与Average precision。

3.2 基准算法与参数设置

本文在数据集批处理方式上采用数据块方式，其中每个数据块的大小经过反复实验调整得到最佳实验效果。对比实验选择了3个基准算法，在线序列极限学习机OS-ELM[12]、基于贝叶斯网络权重损失的半监督多标签学习方法(DS-BW-MLC)[35]与基于加权聚类模型的增量式在线多标签分类方法(OMLC-WC)[36]。此外，文中也分别在OS-KELM基础上增添关联规则和概念漂移后的实验对比。所提算法在实验运算前选取k个数据块构建模型来预测新数据分类精度。3个基准算法的分类器分别采用ELM、贝叶斯网络和聚类模型。为更好地模拟出流式数据环境，在参数设置上选取6个基分类器，基分类器大小根据不同数据集设置不同值，新到来数据块大小根据不同数据集设置不同大小，正则化系数设置为{50, 100, 200, 500}。测试环境基于Intel Core i5处理器、频率2.90 GHz和内存8 GB的一体机。

3.3 性能分析

本节主要考察所提方法的2大实验性能：一是与3个基本算法对比，考察所提方法在多标签数据流上的分类性能；另外，由于一些数据集内部可能存在概念漂移问题，导致原始基分类器难以适应当前概念，所以在所提方法中又引入基分类器更新策略，并通过实验验证新增方法在数据集上的分类性能。主要讨论OS-KELM、OS-KELM-Ap、Proposed method与OS-ELM算法在Yeast、Scene、Corel5k、Enron、Medical数据集上结果对比以及与DS-BW-MLC和OMLC-WC算法在20NG、Enron、IMDBF、OHSUMED、SLASHDOT、TMC2007数据集上结果对比，如表2～5所示。各类算法描述如下：

(1) OS-KELM：采用滑动窗口方法将数据流环境以数据块形式不断到来。选取一定规模的数据块作为基分类器训练集，使用KELM神经网络训练得到集成模型对新数据块预测结果。

(2) OS-KELM-Ap：在OS-KELM基础上引入关联规则，利用关联规则分析标签空间中标签的关联性，在预测过程中对KELM基分类器的预测结果额外使用关联性调优，得到数据块的最终结果。

(3) OS-ELM：基于OS-ELM模型对多标签数据流直接进行预测，模型本身具有迭代优化措施，可以很好应对各种概念漂移情况，但数据块规模过大对其计算过程的时间消耗存在负担。

(4) DS-BW-MLC：基于贝叶斯网络权重损失的半监督多标签学习方法，从每个样本标签空间中自主学习标签间关联关系，使用霍夫丁不等式与标签基数动态调整预测的标签个数，此外还通过未确定值方法扩展标签特征值解决缺失问题。

表2 2种算法在5个数据集上的实验结果

Table 2 Experimental results of two multi-label algorithms on five datasets

度量标准	算法	Yeast	Scene	Corel5k	Medical	Enron
Accuracy	Proposed method	**0.550**	**0.658**	**0.158**	0.656	**0.452**
	OS-ELM	0.493	0.610	0.060	0.713	0.404
F_1-measure	Proposed method	**0.690**	**0.811**	**0.244**	0.707	0.534
	OS-ELM	0.632	0.637	0.093	0.750	0.536
Hamming loss	Proposed method	**0.200**	0.131	0.010	0.012	0.055
	OS-ELM	0.206	0.098	0.009	0.011	0.049

表3 3种算法在5个数据集上2种指标的实验结果

Table 3 Experimental results of three multi-label algorithms on five datasets regarding two evaluation metrics

度量标准	算法	TMC2007	OHSUMEDF	20NG	IMDBF	SLASHDOT
Average precision	Proposed method	0.575	0.499	0.535	0.390	0.513
	DS-BW-MLC	0.552	0.557	0.687	0.268	0.557
	OMLC-WC	0.662	0.381	0.389	0.473	0.359
Ranking loss	Proposed method	**0.040**	**0.160**	0.070	0.200	**0.100**
	DS-BW-MLC	0.105	0.172	0.097	0.314	0.151
	OMLC-WC	0.132	0.275	0.190	0.177	0.258

表4 3种算法在5个数据集上所有指标的实验结果

Table 4 Experimental results of three multi-label algorithms on five datasets regarding all evaluation metrics

度量标准	算法	Yeast	Scene	Corel5k	Medical	Enron
Accuracy	OS-KELM	0.450	0.368	0.077	0.546	0.277
	OS-KELM-Ap	0.530	0.590	0.078	0.670	0.430
	Proposed method	**0.550**	**0.658**	**0.158**	0.656	0.452
F_1-measure	OS-KELM	0.597	0.550	0.120	0.701	0.474
	OS-KELM-Ap	0.680	0.610	0.124	0.730	0.510
	Proposed method	0.690	**0.811**	**0.244**	**0.707**	**0.534**
Hamming loss	OS-KELM	0.248	0.249	0.014	0.016	0.060
	OS-KELM-Ap	0.246	0.247	0.014	0.015	0.060
	Proposed method	**0.200**	**0.131**	**0.010**	**0.012**	**0.055**
Average precision	OS-KELM	0.693	0.541	0.141	0.775	0.545
	OS-KELM-Ap	0.696	0.542	0.220	0.775	0.546
	Proposed method	**0.770**	**0.897**	**0.261**	**0.86**	**0.588**
Ranking loss	OS-KELM	0.236	0.420	0.622	0.098	0.196
	OS-KELM-Ap	0.236	0.413	0.615	0.098	0.194
	Proposed method	**0.160**	**0.107**	**0.399**	**0.030**	**0.158**
Coverage	OS-KELM	0.532	0.280	0.518	0.102	0.405
	OS-KELM-Ap	0.533	0.274	0.511	0.102	0.403
	Proposed method	**0.450**	**0.080**	**0.475**	**0.083**	**0.345**

表5 3种算法在另外5个数据集上所有指标的实验结果

Table 5 Experimental results of three multi-label algorithms on remaining five datasets regarding all evaluation metrics

度量标准	算法	TMC2007	OHSUMEDF	20NG	IMDBF	SLASHDOT
Accuracy	OS-KELM	0.166	0.097	0.090	0.086	0.116
	OS-KELM-Ap	0.166	0.097	0.074	0.086	0.116
	Proposed method	**0.360**	**0.110**	**0.110**	**0.120**	**0.170**
F_1-measure	OS-KELM	0.284	0.177	0.160	0.180	0.205
	OS-KELM-Ap	0.285	0.177	0.160	0.180	0.205
	Proposed method	**0.520**	**0.200**	0.160	**0.200**	**0.290**
Hamming loss	OS-KELM	0.457	0.619	0.569	0.528	0.391
	OS-KELM-Ap	0.457	0.619	0.569	0.528	0.391
	Proposed method	**0.380**	**0.570**	**0.390**	**0.470**	**0.310**
Average precision	OS-KELM	0.687	0.493	0.433	0.322	0.487
	OS-KELM-Ap	0.689	0.493	0.433	0.322	0.487
	Proposed method	0.575	0.499	**0.535**	**0.390**	**0.513**
Ranking loss	OS-KELM	0.119	0.216	0.262	0.250	0.221
	OS-KELM-Ap	0.117	0.216	0.262	0.250	0.221
	Proposed method	**0.040**	**0.160**	**0.070**	0.200	**0.100**
Coverage	OS-KELM	0.235	0.297	0.253	0.493	0.228
	OS-KELM-Ap	0.233	0.297	0.253	0.493	0.228
	Proposed method	**0.030**	**0.230**	**0.070**	**0.340**	**0.100**

(5) OMLC-WC:基于加权聚类模型的流多标签分类方法,利用损失函数的衰减机制保证分类器能够适应潜在的概念漂移。

根据表2~5实验结果可知:

(1)由表2所示,本文所提方法在Yeast、Scene、Corel5k数据集上的Accuracy、F_1-measure上显著优于OS-ELM。因为所提算法不仅利用ELM的高效快速高精度优势,还引入核ELM克服ELM网络结构不稳定等缺陷,并发掘到来数据块标签空间中蕴含的标签间相关性。此外,为应对数据流环境下多标签分类过程出现由于样本数据分布变化导致精度下降问题,构建概念漂移检测器并使用损失函数降低历史数据的影响程度,因此实验结果显著优于OS-ELM。

(2)通过比较表2和表4~5中OS-ELM和OS-KELM、OS-KELM-Ap以及所提方法的4组实验结果可以得出,OS-KELM在流式环境中并没有OS-ELM分类能力高,原因在于OS-ELM方法对模型更新过程中可以充分利用已到来的所有历史信息,而OS-KELM只能对单独到来数据块分别进行集成分类器训练,虽然在批处理环境中引入核函数技术相较于极限学习机在分类上有一定优势,但是在流式环境中忽略了数据之间的关系和历史数据的保留。在引入标签相关性的OS-KELM-Ap和添加标签相关性、概念漂移的所提方法后实验效果才有明显提升,通过4组方法在实验数据集Yeast、Scene、Enron上的Accuracy、F_1-measure、Hamming loss评价指标结果也验证了方法元素的有效性。

(3)此外与DS-BW-MLC和OMLC-WC在6个高维特征数据集上对比,表3中实验结果可知:在TMC2007和IMDBF、Enron数据集上的Average precision实验指标,所提算法优于DS-BW-MLC;在SLASHDOT和OHSUMEDF上的Average precision优于OMLC-WC,其他5个数据集上所提算法在

Ranking loss优于OMLC-WC。原因在于所提算法在高维大样本数据集的处理能力有待改进,仅依赖标签相关性和概念漂移检测技术远远不够,还应考虑高维数据环境下的特征降维。

(4)在TMC2007数据集上,由于数据集中旧样本数据占比较新样本更重,所提方法更新策略虽然在一定程度上能够适应新的数据,但是忽略了对于历史数据的保留,针对这类数据集本文方法在与未引入更新策略的OS-KELM和关联规则OS-KELM-Ap实验效果对比明显降低。

(5)所提方法使用集成模型的核极限学习机与基于Apriori算法的类标签相关性分析以及基于MuENLForest模型的概念漂移检测与模型更新机制。为验证所用技术有效性,本文分别在10个数据集上做增量式自对比实验,结果如表4~5所示。从10个数据集上的实验结果可以看出,所提方法明显优于OS-KELM和OS-KELM-Ap,进一步反应出在多标签数据流分类过程中合理使用标签间相关性、适应概念漂移方法的重要性。无论是在低维的Yeast、Scene还是高维20NG、IMDBF数据集上,所提方法都比OS-KELM在所有评价指标上实验结果有很大提升。为单独验证标签相关性的作用,对比OS-KELM与OS-KELM-Ap的实验结果可知,在数据集Yeast、Scene、Enron上Accuracy、F_1-measure提升效果明显,但在高维数据集上由于部分数据集的标签平均相关度不高(样本数据标签间的相关性值,取该数据集整体标签空间中标签对的相关性平均值,例如数据集20NG的标签平均相关度为0.364 1、OHSUMED:0.380 2、Enron:0.881 1、Scene:0.674 4)导致标签相关性技术不能很好地发挥作用,实验结果也显示出标签相关性技术带来的提升效果不明显。但所提方法的概念漂移检测与模型更新机制有效性与OS-KELM-Ap实验相比,无论是高维数据TMC2007、SLASHDOT和OHSUMEDF还是低维数据Yeast、Scene、Corel5k,实验精度Accuracy、Average precision、Coverage、Ranking loss显著提高,验证了概念漂移机制的有效性,可以及时发现新样本标签数据分布变化,集成模型迭代更新的正确时机。

(6)使用Nemenyi-Test检验对比算法的实验性能是否存在显著差异。将本文工作的统计检验视为控制算法,记录每个算法在运行数据集上的平均等级,其中各算法之间的差异用临界差异(Critical difference, CD)校准。如果它们的平均等级相差至少1个CD(本文中CD分别为2.728和2.097 6;比较算法分别为4个和5个,数据集N=5),则认为性能差异是显著的。

为了在直观上显示算法间的性能差异,图2~6给出了相应指标下的CD图,其在Accuracy、Average precision、Ranking loss等方面。在每个CD图中,算法间的平均等级沿着轴标记右下方的等级。此外,任何具有一个CD内的平均等级与本文所提算法的平均等级的比较算法用粗线相互连接,否则

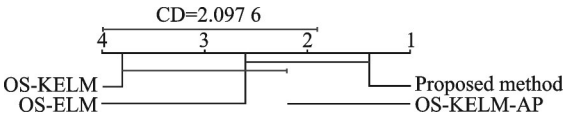

图2 在Accuracy度量标准上的统计结果

Fig.2 Statistic test on Accuracy

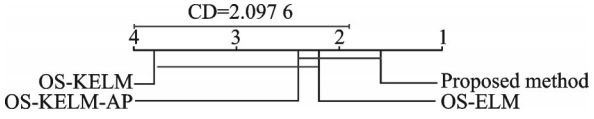

图3 在F_1-measure度量标准上的统计结果

Fig. 3 Statistic test on F_1-measure

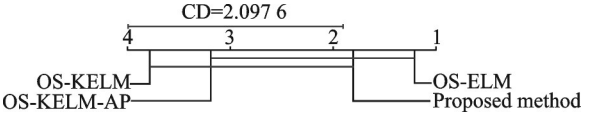

图4 在Hamming loss度量标准上的统计结果

Fig. 4 Statistic test on Hamming loss

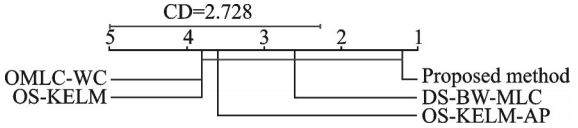

图5 在Ranking loss度量标准上的统计结果

Fig.5 Statistic test on Ranking loss

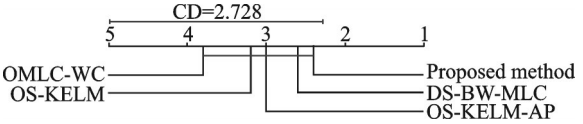

图6 在Average precision度量标准上的统计结果

Fig. 6 Statistic test on Average precision

其性能被认为与本文算法存在显著差异。根据实验结果,所提算法在统计显著性方面效果良好。

4 结束语

本文提出一种基于核极限学习机的多标签数据流集成分类方法,利用核极限学习机通过对已有的多标签信息进行在线分类预测处理,为充分利用潜在标签集合间的相关性,在分类过程中利用Apriori算法得到标签间关联规则用以提高分类精度。同时为了适应不断到来的新数据可能引起概念漂移问题,本文采用利用旧数据构建概念漂移检测森林,并对分类器更新策略采用最差原理,每次使用由最新数据块得到的分类器替换当前效果最差的分类器来完成对基分类器的更新操作。大量对比实验表明,所提方法具有良好的分类效果,同时能够适应小样本数据流中概念漂移问题。下一步工作将针对高维空间中的大样本稀疏问题展开研究探讨,并将进一步合理利用已抛弃的有效历史数据对未来重现概念漂移的问题研究。

参 考 文 献:

[1] ZHANG M, ZHOU Z. A review on multi-label learning algorithms[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1819-1837.

[2] TSOUMAKAS G, KATAKIS I. Multi-label classification: An overview[J]. International Journal of Data Warehousing and Mining, 2007, 3(3): 1-13.

[3] HUANG G, ZHU Q, SIEW C K, et al. Extreme learning machine: Theory and applications[J]. Neurocomputing, 2006, 70(1): 489-501.

[4] HUANG G, ZHU Q, SIEW C, et al. Extreme learning machine: A new learning scheme of feedforward neural networks[C]//Proceedings of International Joint Conference on Neural Network. [S.l.]: IEEE, 2004: 985-990.

[5] HUANG G, ZHOU H, DING X, et al. Extreme learning machine for regression and multiclass classification[J].Proceedings of Systems Man and Cybernetics, 2012, 42(2): 513-529.

[6] VENKATESAN R, ER M J. Multi-label classification method based on extreme learning machines[C]//Proceedings of International Conference on Control, Automation, Robotics and Vision. [S.l.]: IEEE, 2014: 619-624.

[7] ZHENG W, QIAN Y, LU H, et al. Text categorization based on regularization extreme learning machine[J]. Neural Computing and Applications, 2013,22(3): 447-456.

[8] KONGSOROT Y, HORATA P. Multi-label classification with extreme learning machine[C]//Proceedings of International Conference on Knowledge and Smart Technology. [S.l.]: IEEE, 2014: 81-86.

[9] SUN X, XU J, JIANG C, et al. Extreme learning machine for multi-label classification[J]. Entropy, 2016, 18(6): 225.

[10] LUO F, GUO W, YU Y, et al. A multi-label classification algorithm based on kernel extreme learning machine[J]. Neurocomputing, 2017, 260: 313-320.

[11] NGUYEN T T, NGUYEN T T, PHAM X C, et al. An ensemble-based online learning algorithm for streaming data[EB/OL]. (2017-02-13)[2020-07-02]. https://arxiv.org/abs/1704.07938.

[12] VENKATESAN R, ER M J, WU S, et al. A novel online real-time classifier for multi-label data streams[C]//Proceedings of 2016 International Joint Conference on Neural Networks (IJCNN). [S.l.]:IEEE, 2016: 1833-1840.

[13] BIFET A, GAVALDA R. Adaptive learning from evolving data streams[C]//Proceedings of Intelligent Data Analysis. Berlin Heidelberg: Springer, 2009: 249-260.

[14] OSOJNIK A, PANOV P, DŹEROSKI S, et al. Multi-label classification via multi-target regression on data streams[J]. Machine Learning, 2017, 106(6): 745-770.

[15] ZHANG M, LI Y, LIU X, et al. Binary relevance for multi-label learning: An overview[J]. Frontiers of Computer Science in China, 2018, 12(2): 191-202.

[16] ZHU Y, TING K M, ZHOU Z, et al. Multi-label learning with emerging new labels[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(10): 1901-1914.

[17] RUMELHART D E, HINTON G E, WILLIAMS R J, et al. Learning representations by back-propagating errors[J]. Nature, 1988, 323(6088): 696-699.

[18] SUN X, WANG J, JIANG C, et al. ELM-ML: Study on multi-label classification using extreme learning machine[C]// Proceedings of ELM-2015 Volume 2. [S.l.]: Springer International Publishing, 2016.
[19] ER M J, VENKATESAN R, WANG N, et al. A high speed multi-label classifier based on extreme learning machines[C]// Proceedings of ELM-2015 Volume 2. Cham:Springer, 2016: 437-454.
[20] VENKATESAN R, ER M J. Multi-label classification method based on extreme learning machines[C]//Proceedings of International Conference on Control, Automation, Robotics and Vision. [S.l.]: IEEE, 2014: 619-624.
[21] FENG L, WANG J, LIU S, et al. Multi-label dimensionality reduction and classification with extreme learning machines[J]. Journal of Systems Engineering and Electronics, 2014, 25(3): 502-513.
[22] ZHANG N, DING S, ZHANG J, et al. Multi layer ELM-RBF for multi-label learning[J]. Applied Soft Computing, 2016, 43: 535-545.
[23] KONGSOROT Y, HORATA P, MUSIKAWAN P, et al. Kernel extreme learning machine based on fuzzy set theory for multi-label classification[J]. International Journal of Machine Learning and Cybernetics, 2019, 10(5): 979-989.
[24] LAW A, GHOSH A. Multi-label classification using a cascade of stacked autoencoder and extreme learning machines[J]. Neurocomputing, 2019, 358: 222-234.
[25] WANG L, SHEN H, TIAN H, et al. Weighted ensemble classification of multi-label data streams[C]//Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham: Springer, 2017: 551-562.
[26] KARPONI K, TSOUMAKAS G, An Empirical comparison of methods for multi-label data stream classification[C]// Proceedings of INNS Conference on Big Data. Cham: Springer, 2017.
[27] QU W, ZHANG Y, ZHU J, et al. Mining multi-label concept-drifting data streams using dynamic classifier ensemble[C]// Proceedings of Asian Conference on Machine Learning. Berlin, Herdelberg: Springer, 2009: 308-321.
[28] XIOUFIS E S, SPILIOPOULOU M, TSOUMAKAS G, et al. Dealing with concept drift and class imbalance in multi-label stream classification[C]//Proceedings of International Joint Conference on Artificial Intelligence.[S.l.]:[s.n.], 2011: 1583-1588.
[29] READ J, BIFET A, HOLMES G, et al. Scalable and efficient multi-label classification for evolving data streams[J]. Machine Learning, 2012, 88(1/2): 243-272.
[30] DOMINGOS P, HULTEN G. Mining high-speed data streams[C]//Proceedings of Knowledge Discovery and Data Mining. [S.l.]: ACM, 2000: 71-80.
[31] READ J, PFAHRINGER B, HOLMES G, et al. Multi-label classification using ensembles of pruned sets[C]//Proceedings of International Conference on Data Mining. [S.l.]: IEEE, 2008: 995-1000.
[32] BIFET A, HOLMES G, PFAHRINGER B, et al. New ensemble methods for evolving data streams[C]//Proceedings of Knowledge Discovery and Data Mining. [S.l.]: ACM, 2009: 139-148.
[33] SHI Z, WEN Y, FENG C, et al. Drift detection for multi-label data streams based on label grouping and entropy[C]// Proceedings of International Conference on Data Mining. [S.l.]: IEEE, 2014: 724-731.
[34] SHI Z, XUE Y, WEN Y, et al. Efficient class incremental learning for multi-label classification of evolving data streams[C]// Proceedings of International Joint Conference on Neural Network. [S.l.]: IEEE, 2014: 2093-2099.
[35] NGUYEN T T, NGUYEN T T, LUONG A V, et al. Multi-label classification via label correlation and first order feature dependance in a data stream[J]. Pattern Recognition, 2019, 90: 35-51.
[36] NGUYEN T T, DANG M T, LUONG A V, et al. Multi-label classification via incremental clustering on an evolving data stream[J]. Pattern Recognition, 2019, 95: 96-113.
[37] LIU B, HSU W, MA Y, et al. Integrating classification and association rule mining[C]//Proceedings of Knowledge Discovery and Data Mining. [S.l.]: ACM, 1998, 98: 80-86.

作者简介:

张海翔(1996-),男,硕士研究生,研究方向:数据挖掘与人工智能,E-mail: 529644348@qq.com。

李培培(1982-),**通信作者**,女,博士,副研究员,研究方向:数据挖掘和知识工程,E-mail: peipeili@hfut.edu.cn。

胡学钢(1961-),男,博士,教授,研究方向:数据挖掘和知识工程,E-mail:jsjxhuxg@hfut.edu.cn。

(编辑:刘彦东)