2019 年 2 月 第 2 期
Vol. 40 No. 2 2019

结合滑动窗口与模糊互信息的多标记流特征选择

程玉胜[1,2,3],李 雨[1],王一宾[1,3],陈 飞[1]

[1](安庆师范大学 计算机与信息院,安徽 安庆 246011)
[2](数据科学与智能应用福建省高校重点实验室,福建 漳州 363000)
[3](安徽省高校智能感知与计算重点实验室,安徽 安庆 246011)
E-mail: chengyshaq@163.com

摘 要: 特征选择是处理高维度问题的一种有效方法,而传统的大部分算法都基于静态的特征空间.但是有些问题其特征空间和标记空间均呈现增量或动态的特点,传统的特征选择算法不再适用.针对这一问题,结合滑动窗口机制,本文提出了结合滑动窗口与模糊互信息的多标记流特征选择;同时,为了减弱互信息对特征重要程度的判断,对模糊互信息进行正则化处理,并通过正则化重新优化特征重要度目标函数.提出的算法在多标记数据集上进行了大量测试,实验结果和统计假设检验说明本文算法是有效的.

关 键 词: 模糊互信息;多标记学习;数据流;特征选择
中图分类号: TP181 **文献标识码:** A **文 章 编 号:** 1000-1220(2019)02-0320-08

Multi-label Streaming Feature Selection Combining Sliding Window and Fuzzy Mutual Information

CHENG Yu-sheng[1,2,3],LI Yu[1],WANG Yi-bin[1,3],CHEN Fei[1]

[1](School of Computer and Information,Anqing Normal University,Anqing 246011,China)
[2](The University Key Laboratory of Data Science and Intelligence Application of Fujian,Zhangzhou 363000,China)
[3](The University Key Laboratory of Intelligent Perception and Computing of Anhui Province,Anqing 246011,China)

Abstract: Feature selection is an effective method to deal with high dimension problem in multi-label learning. The traditional feature selection algorithm is based on static features and labels space mostly. However,in the real life,some features cannot be known in advance,and their features or labels are incremental or dynamic. The traditional algorithms about the feature selection do not work well any longer. To solve this problem,combining with the sliding window mechanism,a streaming feature selection method is proposed with fuzzy mutual information. At the same time,to weaken the judge of mutual information on the importance of the feature,the fuzzy mutual information is regularized and the degree of importance is redefined to optimize the objective function. The experimental results and the statistical hypothesis test further illustrate the effectiveness of our proposed algorithm.

Key words: fuzzy mutual information;multi-label learning;streaming data;feature selection

1 引 言

多标记学习是机器学习、数据挖掘、模式识别等人工智能领域的研究热点之一,能很好的解决现实生活中多义性问题.在多标记学习框架之下,特征数据往往面临着高维性[1-3]的问题,使得一般的手工标记费时费力,不仅如此,高维数据也会严重影响分类器的分类精度[4-6].但如果利用标记的相关性对特征进行选择,无疑会很好的解决上述问题.

目前,Lee 等[7]提出了基于多变量互信息的多标记特征选择算法(Pairwise Multivariate Mutual Information,PMU),其算法是通过选择与标记空间互信息最大的特征,生成特征子集. Lin 等[8]提出了基于邻域互信息的特征选择算法(Multi-label feature selection based on neighborhood mutual information,MFNMI),其算法基于三种不同的观点,首先计算特征与标记空间的互信息大小,选择与标记空间互信息最大、与已选特征冗余性最小的特征生成特征子集. Zhang 等[9]提出了基于最大相关性的属性约简算法(Multi-Label Dimensionality Reduction via Dependence Maximization,MDDM),其算法利用两种投影策略,根据原始特征与标记空间最大相关性将原始数据投影到较低维的特征空间. Zhang 等[10]提出的多标签朴素贝叶斯分类的特征选择算法(Feature selection for multi-label naive Bayes classification,MLNB)等.

另外,上述常见的特征选择算法都无法处理动态或者增量数据,然而现实世界中有些实例的特征无法一次性全部获取,数据是实时产生并记录的,例如患者诊断就是医生不断的诊断检查再诊断检查,最后才确定患者的病因病情的过程.我们把这种特征逐渐增加的过程称为流特征.传统的特征选择算法,无法解决具有流特征[11-14]问题的特征选择.滑动窗

收稿日期: 2018-04-24 收修改稿日期: 2018-05-24 基金项目: 安徽省高校重点科研项目(KJ2017A352)资助;福建省高校重点实验室开放课题项目(D1801)资助;安徽省高校重点实验室基金项目(ACAIM160102)资助. 作者简介: 程玉胜,男,1969 年生,博士,教授,研究方向为数据挖掘、粗糙集和机器学习等;李 雨,女,1992 年生,硕士研究生,研究方向为多标签学习、机器学习和数据挖掘等;王一宾,男,1970 年生,硕士,教授,研究方向为多标签学习、机器学习和软件安全等;陈 飞,男,1994 年生,硕士研究生,研究方向为特征选择、数据挖掘和机器学习等.

口[15,16]机制能很好的解决该问题. 目前滑动窗口主要集中在数据流聚类和特征选择[17-19]算法研究中,如 Guha 等[20]提出了 STREAM 算法,该算法是针对数据流聚类的典型算法,其根据分治原理,在有限的空间对数据流进行聚类. 常建龙等[16]提出了基于滑动窗口的数据流聚类算法 CluWin,该算法不仅研究了拒伪和纳真误差滑动窗口模型中的聚类问题,还将其推广用于解决 N-n 窗口内的数据流聚类的问题,针对数据流环境下特征选择还主要以单标记特征选择为主,对多标记学习中流特征选择研究还不多.

同时不难发现,在常见的特征选择算法中,目前主要还是基于信息熵方法计算特征与标记之间的相关性较多,但这种方法主要基于香农熵也即传统信息熵[21-23]方法,该方法不具有补的性质,因此计算过程较为复杂. 上述大部分特征选择算法中都利用互信息作为启发式函数来选择重要特征,一般来说互信息越大其特征越重要. 然而,香农信息熵计算"只计字数、不计内容",撇开了人的主观因素和信息本身的含义,如果结合实际情况考虑一个有意义的因素,定量地给出该事件相应的权重,那么修正的加权熵就更有意义[24,25],为了解决这种问题,在模糊信息熵的方法基础上,本文重新构造了重要度目标函数: 首先对互信息进行了加权处理,然后对其进行正则化,并且正则化处理依次在有限大小的窗口中进行.

针对上述问题,本文提出了结合滑动窗口与模糊互信息流特征选择算法,首先将滑动窗口应用到多标记学习的特征选择上,根据判断条件将满足条件的特征保存下来,随着特征的不断流入,窗口也向前滑动. 其次,引入粗糙集理论修正了传统的信息熵,并用新的重要度目标函数度量特征与标记之间的相关性. 最后,本文进行了大量实验,实验结果表明本文算法是有效的,并采用统计假设检验对本文算法进行了有效性检验.

2　多标记学习及其模糊信息熵

2.1　多标记学习框架

定义. [10] $X=R^m$ 表示 m 维样本空间,样本集合 $X=\{x_1,x_2,\cdots,x_q\}$,类别标记集合 $L=\{l_1,l_2,\cdots,l_n\}$,给定多标记训练集 $T=\{(x_i,Y_i)\mid i=1,2,\cdots,q\}$,在特征空间中,样本 x_i 用 m 维属性向量 $x_i=[x_i^1,x_i^2,\cdots,x_i^m]$ 来表示,样本 x_i 对应于标记空间中的标记集合记为 $Y_i=[y_i^1,y_i^2,\cdots,y_i^n]$,当 x_i 含有标记时 $l_a,y_i^a=1$,否则 $y_i^a=-1$.

2.2　模糊信息熵

定义 1[26]. 随机变量 $X=\{x_1,x_2,\cdots,x_q\}$,随机变量 X 的不确定期望为 $H(X)=-\sum_{i=1}^{m}\frac{|x_i|}{|X|}\log_2\frac{|x_i|}{|X|}$,令 $p_i=\frac{|x_i|}{|X|}=\frac{|x_i|}{\sum_{j=1}^{q}|x_j|}$,随机变量的信息熵为

$$H(X)=-\sum_{i=1}^{m}p_i\log_2 p_i \tag{1}$$

信息熵是度量随机变量不确定性的程度,信息熵越大就表示随机变量不确定性程度越大,在多标记的特征选择算法中,常常利用传统信息熵来选择特征空间中与整个标记空间互信息大的特征,但传统信息熵不具有补的性质,基于传统的信息熵,利用粗糙集理论等价划分的思想,引入模糊信息熵的新定义.

定义 2. [23,27] 假设信息系统由样本及其对应的特征所描述,把样本空间的描述记为论域 U,根据某种特征属性可以对论域 U 进行划分,假设按照特征属性 $F=\{f_1,f_2,\cdots,f_n\}$ 对论域 U 进行划分记 $U/F=X=\{X_1,X_2,\cdots,X_m\}$,则模糊信息熵定义如下:

$$E(X)=\sum_{i=1}^{m}\frac{|X_i|}{|U|}\cdot\frac{|X_i^c|}{|U|}=\sum_{i=1}^{m}\frac{|X_i|}{|U|}\left(1-\frac{|X_i|}{|U|}\right) \tag{2}$$

其中 $E(X)$ 即为模糊熵,公式(2)中 X_i^c 是 X_i 的补集,即 $X_i^c=U-X_i$,$|X_i^c|$ 是 X_i^c 的个数大小,$|X_i|$ 是 X_i 的个数大小,$\frac{|X_i|}{|U|}$ 表示等价类 X_i 在论域 U 中的概率,$\frac{|X_i^c|}{|U|}$ 表示在论域 U 中的 X_i 的互补概率. 用 $p=(p_1,p_2,\cdots,p_m)$ 表示论域 U 中的概率,即 $E(p)=E(X)$:

$$\begin{aligned}E(p)&=\sum_{x\in U}p(x)\cdot(1-p(x))=\sum_{i=1}^{q}[p(x_i)-p^2(x_i)]\\&=\sum_{i=1}^{q}p(x_i)-\sum_{i=1}^{q}p^2(x_i)=1-\sum_{x\in U}p^2(x)\end{aligned} \tag{3}$$

性质 1. [23] 知识 p 的信息熵 $E(p)$ 满足: $0\leqslant E(p)\leqslant 1-\frac{1}{q}$ (其中,$|U|=q$).

证明: 易知 $0\leqslant E(p)$,求上式(3)最大值即要求 $\sum_{x\in U}p^2(x)$ 取得最小值,利用拉格朗日乘数法构造函数,构造函数 $H(\partial)=\sum_{x\in U}p^2(x)+\beta\left(\sum_{x\in U}p(x)-1\right)$.

对 $H(\partial)$ 求导得: $H'_{\partial}(\partial)=\sum_{x\in U}p(x)-1=0$ 求得 $p(x)=\frac{1}{q}$,这时 $\sum_{x\in U}p^2(x)$ 取得最小值 $\frac{1}{q}$,$E(p)=\sum_{x\in U}P(x)\cdot(1-p(x))=1-\sum_{x\in U}p^2(x)$ 取得最大值 $1-\frac{1}{q}$. 因此,$0\leqslant E(p)\leqslant 1-\frac{1}{q}$ 成立.

在多标记学习中,按照特征属性对论域 U 划分记为 $X=\{X_1,X_2,\cdots,X_m\}$,按照标记属性对论域 U 的划分记为 $Y=\{Y_1,Y_2,\cdots,Y_n\}$,类似于传统的条件信息熵,模糊条件熵[27]定义如下:

定义 3. [23,27]

$$E(Y\mid X)=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|Y_i\cap X_j|}{|U|}\cdot\frac{|Y_i^c-X_j^c|}{|U|} \tag{4}$$

定义 4. [23,27] 类似的,模糊互信息定义为:

$$E(Y;X)=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|Y_i\cap X_j|}{|U|}\cdot\frac{|Y_i^c\cap X_j^c|}{|U|} \tag{5}$$

模糊自信息熵、条件熵以及模糊互信息之间满足如下关系:

$$E(Y;X)=E(Y)-E(Y\mid X) \tag{6}$$

证明: 因为 $Y_i^c=(Y_i^c\cap X_j^c)\cup(Y_i^c-X_j^c)$

$$\begin{aligned}E(Y)&=\sum_{i=1}^{n}\frac{|Y_i|}{|U|}\frac{|Y_i^c|}{|U|}\\&=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|Y_i\cap X_j|}{|U|}\frac{|Y_i^c|}{|U|}\\&=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|Y_i\cap X_j|}{|U|}\frac{|(Y_i^c\cap X_j^c)\cup(Y_i^c-X_i^c)|}{|U|}\\&=E(Y;X)+E(Y\mid X)\end{aligned}$$

所以,$E(Y;X)=E(Y)-E(Y\mid X)$ 成立.

2.3 一种新的重要度目标函数

目前,很多算法都利用特征与标记空间之间的互信息来判断特征是否为重要特征,然而互信息最大的特征可能并不是最重要的特征,为了避免过拟合,我们先对互信息进行加权,然后对其进行正则化处理,构造的函数如下:

$$para(i) = \frac{\sum_{j=1}^{q}(x_{ji} \neq 0)}{\sum_{j=1}^{q}(x_{ji} = 0)} \cdot FMI(f_{t_i}; L) + \frac{W+i}{q} \quad (7)$$

公式(7)中 q 值表示样本的个数,$x_{ji} \neq 0$ 表示在特征空间中第 i 个特征下所有实例中不为 0 的实例个数,$x_{ji} = 0$ 则表示在特征空间中第 i 个特征下所有实例中为 0 的实例个数; W 为处理窗口的大小,$\frac{W+i}{q}$为正则项.

另外,如果对于某一个特征,存在 $\sum_{j=1}^{q}(x_{ji} = 0)$ 为 0 或 $\sum_{j=1}^{q}(x_{ji} \neq 0)$ 为 0 时,那么这个特征所对应的 $para(i)$ 的取值则为 $FMI(f_i; L) + \frac{W+i}{q}$.

3 基于滑动窗口的流特征选择

3.1 滑动窗口机制

滑动窗口一般有两个端点,分别确定这个窗口的起始位置和终止位置,如图 1 滑动窗口模型[18,19],每一个新的数据流入滑动窗口时,滑动窗口的前端和末端都会同时向前移动,在依次向前滑动的过程中,根据窗口内数据个数是否发生变化可将滑动窗口分为窗口大小固定和窗口大小可变两种形式,一般地,每到达一个新的数据时,滑动窗口的前端和末端同时向前移动一个位置. 本文设定两个窗口大小固定的窗口,即处理窗口和存储窗口,特征进入的窗口即为处理窗口,处理窗口的大小设为 W,处理窗口是动态的,每当一个新的特征到达时,处理窗口整体会向前滑动一个位置,而原处理窗口最末端的数据被删除. 保留重要特征的窗口即为存储窗口,存储窗口的大小设为 H,存储窗口是固定的,不随着新特征的到达而移动. 假设滑动窗口的传输方向是自左向右,如图 1 所示.

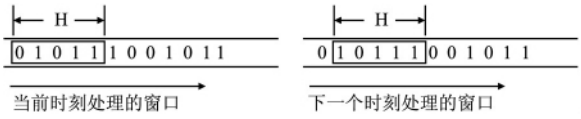

图 1 滑动窗口模型

Fig. 1 Sliding window model

实际情况中每个特征流入滑动窗口中的时间间隔并不是相等的,由于滑动窗口大小是固定的,所以如果仅以相等的时间间隔向后移动窗口是不合理的,窗口大小固定的滑动窗口是在按照时间顺序的基础上以窗口大小 W 为单位依次向后移动,如果某一时刻,特征同时到达的个数几乎不止一个,那么这个时候我们应该尽可能的精确时间,使得每个特征到达的顺序是相差一定时间的.

3.2 结合滑动窗口与模糊互信息的多标记流特征选择

在本文所提出的结合滑动窗口的模糊互信息流特征选择算法中,利用滑动窗口对特征的模糊互信息进行正则化处理,以减小模糊互信息对特征重要程度判断的影响,防止过度拟合,采用正则化方法会自动削弱不重要的特征变量. 首先将特征空间中的特征作为流特征,按照特征排序的顺序依次进入滑动窗口,特征在进入固定大小的滑动窗口后计算特征与整个标记空间的模糊互信息大小,然后根据公式(7)的特征重要程度目标函数依次计算其值的大小,所得的值越大就表示其特征越重要,特征保存到存储窗口后并在存储窗口内根据重要程度由小到大依次进行排序,当存储窗口 H 中的特征个数达到了窗口大小时,窗口仍然向后滑动,如果下一时刻进入的特征的重要程度比存储窗口中的第一个特征大,那么新进入的特征将取代第一个特征,依次循环,直到最后一个特征进入窗口中,最后,将存储窗口 H 的特征序列构成特征子集.

根据上述描述,结合滑动窗口与模糊互信息的多标记流特征选择(*M*ulti-label *S*treaming *F*eature *S*election combining *S*liding *W*indow and *F*uzzy *M*utual *I*nformation: ***SFS-FMI-SW***)描述如下:

输入: 处理窗口的大小 W,存储窗口的大小 H,多标记学习数据集 T.

输出: t_i 时刻所产生的特征子集 f.

1) t_i 时刻产生的特征 f_i;
2) while $|f| \leqslant H$, do
3) for each $f_i \in F$;
4) for each $l_j \in L$;
5) 根据公式(2)计算 $E(f_i; l_j)$;
6) end;
7) 根据 $\sum_{j=1}^{n} E(f_i; L)$,计算流入处理窗口中的特征 f_i 对整个标记空间的模糊互信息;
8) 根据流特征与标记空间的模糊互信息利用公式(7)计算特征的重要程度目标函数 $para(i)$;
9) 特征进入存储窗口中根据公式(7)的值的大小依次从小到大进行排序;
10) end
11) end while;
12) for i = $H+1$: m
13) 如果新流入的特征的重要程度比存储窗口中任一特征重要;
14) 那么将新特征代替存储窗口中的那个特征;
15) 存储窗口内的特征根据重要程度自动进行排序;
16) end;
17) 返回存储窗口 H 中的所有特征作为特征子集 f.

4 实验数据及其结果分析

4.1 实验数据

本文采用了 Yeast、Birds、Cal500、Artificial、Reference、Health、Business、Science 这 8 个数据集来验证本文算法的有效性. 各数据集的相关信息见表 1. 数据均来自于 http://mulan.sourceforge.net/datasets.html.

4.2 评价指标

为了验证该算法的有效性,利用海明损失(Hamming Loss,HL),1-错误(One Error,OE)和覆盖率(Coverage,CV),

排位损失(Ranking Loss,RL),平均准确率(Average Precision,AP)这 5 个评价指标[10]对实验结果进行验证. 样本多标

表 1　多标记数据集

Table 1　Multi-label datasets

数据集	样本数	特征数	类别数	训练样本数	测试样本数	应用领域
Yeast	2417	103	14	1499	918	生物
Birds	3322	260	20	322	3000	音频
Cal500	502	68	174	251	251	音频
Artificial	5000	462	26	2000	3000	文本
Reference	5000	793	33	2000	3000	文本
Health	5000	612	32	2000	3000	文本
Business	5000	438	30	2000	3000	文本
Science	5000	743	40	2000	3000	文本
Computer	5000	681	33	2000	3000	文本

记测试集: $T=\{(x_i,Y_i)\mid i=1,2,\cdots,p\}$,由算法预测得到的标记集合记为 $h(x)$.

Hamming Loss:

$$HL(h)=\frac{1}{p}\sum_{i=1}^{p}\mid h(x_i)\,\Delta Y_i\mid \tag{8}$$

公式(8)中 Δ 用于计算集合 Y_i 和 $h(x_i)$ 的对称差,即对集合进行布尔逻辑中的异或操作,该评价指标用于评估样本在单个标记上的误分类情况,$HL(h)$ 越小则表示分类器 $h(x)$ 性能越优,最优值为 0.

One Error:

$$OE(f)=\frac{1}{p}\sum_{i=1}^{p}[[\operatorname{argmax}_{y\in Y}f(x_i,y)]]\notin Y_i] \tag{9}$$

公式(9)中 $f(x_i,y)$ 为分类器 $h(x)$ 对应的实值函数,该评价指标用于考察样本的标记排序序列中位于第一位的标记不属于样本相关标记集合的次数情况. $OE(f)$ 越小则表示分类器 $h(x)$ 性能越优,最优值为 0.

Coverage:

$$CV(f)=\frac{1}{p}\sum_{i=1}^{p}\max_{y\in Y_i}rank_f(x_i,y)-1 \tag{10}$$

公式(10) $rank_f(x_i,y)$ 是与实值函数 $f(x_i,y)$ 对应的排序函数. 该评价指标用于考察在样本标记排序中,覆盖所有相关标记所需的搜索深度情况. $CV(f)$ 越小则表示分类器 $h(x)$ 性能越优,最优值为 0.

Ranking Loss:

$$RL(f)=\frac{1}{p}\sum_{i=1}^{p}\frac{1}{\mid Y_i\mid\mid Y_i^c\mid}\mid\{(y',y)\mid f(x_i,y')\leqslant f(x_i,y''),(y',y'')\in Y_i\times Y_i^c\}\mid \tag{11}$$

公式(11)中,Y_i^c 是 Y_i 在标记空间 L 中的补集,该评价指标用于考察在所有样本的类别标记不相关排序中,无关标记排序排在相关标记之前次数的情况. 该指标值越小则表示分类器 $h(x)$ 性能越优,最优值为 0.

Average Precision:

$$AP(f)=\frac{1}{p}\sum_{i=1}^{p}\frac{1}{\mid Y_i\mid}\sum_{y\in Y_i}\frac{\mid\{y'\mid rank_f(x,y')\leqslant rank_f(x_i,y),y'\in Y_i\}\mid}{rank_f(x_i,y)} \tag{12}$$

该评价指标用于考察在样本的类别标记排序序列中,排在给定样本标记之前的标记仍属于该样本标记概率的均值,该指标取值越大分类器 $h(x)$ 性能越优,其最优值为 1.

4.3　实验结果及分析

实验代码均在 Matlab2012b 中运行,硬件环境 Inter® CoreTM i5-3230 2.6GHz CPU,4G 内存; 操作系统为 Windows 7. 本文的实验是以 kNN 作为分类器,对比算法有基于多变量互信息的多标记特征选择算法(Pairwise Multivariate Mutual Information,PMU)、基于最大相关性降低多标记维度(Multi-Label Dimensionality Reduction via Dependence Maximization,MDDM)、基于邻域互信息的特征选择算法(Multi-label feature selection based on neighborhood mutual information,MFNMI)、多标签朴素贝叶斯分类的特征选择算法(Feature selection for multi-label naïve Bayes classification,MLNB). kNN[28]参数值设定为默认值,即平滑系数 $s=1$,$k=10$. 表中 ↑ 表示指标数值越大越好,↓ 表示指标数值越小越好,实验结果后的符号 "√" 表示每个数据集在所有算法上取得的最好结果,本文算法只与 MLNB、MDDMspc、MDDMproj、PMU、MFNMIpes 对比时取得最好的结果用黑体字表示. 各实验结果后面 "(·)" 内的值表示每个数据集在各算法上的排序. 为了验证算法的有效性,随机记录一组实验,实验中处理窗口大小 W 设置为 100,本文是基于固定滑动窗口的流特征选择算法,因此依次达到的特征也是随机的,运行十次,最终的结果以十次的平均值来代替. 所有算法的特征子集大小都与 MLNB 所选的特征子集大小相等.

实验结果如下:

表 2　在平均精度上五个特征选择算法的排名

Table 2　Ranks among five feature selection methods in Average Precision(↑)

Methods	ML-KNN	MLNB	MDDMspc	MDDMproj	PMU	MFNMIpes	SFS-FMI-SW
Yeast	0.7511(2)	0.7230(7)	0.7376(4)	0.7392(3)	0.7358(6)	0.7363(5)	**0.7534**(1)√
Birds	0.6949(4)	0.6346(7)	0.6536(6)	0.6595(5)	0.7134(3)	0.7157(2)	**0.7383**(1)√
Cal500	0.4853(1)√	0.4776(6)	**0.4825**(2.5)	**0.4825**(2.5)	0.4715(7)	0.4802(5)	0.4819(4)
Artificial	0.5093(3)	0.4991(5)	0.4968(6)	0.4594(7)	0.5051(4)	0.5101(2)	**0.5185**(1)√
Reference	0.6193(4)	0.6234(3)	0.6316(2)	0.6192(5)	0.6103(7)	0.6141(6)	**0.6327**(1)√
Health	0.6812(6)	0.6880(4)	0.6945(2)	0.6830(5)	0.6757(7)	0.6900(3)	**0.7006**(1)√
Business	0.8798(1)√	0.8713(4)	0.8707(5)	0.8731(3)	0.8628(7)	0.8681(6)	**0.8749**(2)
Science	0.5327(1)√	0.4513(4)	0.4449(6)	0.4400(7)	0.4468(5)	0.4566(3)	**0.4648**(2)
Average	2.75	5	4.1875	4.6875	5.75	4	1.625

1) 由表 2 可发现: 本文算法与 MLNB、MDDMspc、MDDMproj、PMU、MFNMIpes 这 5 个算法进行对比时,在 8 个数

据集上,本文提出的算法 SFS-FMI-SW 在其中 7 个数据集上 Average Precision 值最大,即性能最好; Cal500 数据集上 SFS-FMI-SW 算法的 AP 值的大小仅比 MDDM 算法的 AP 值少 0.0006,这表示 SFS-FMI-SW 算法的性能非常稳定. 本文算法与所有算法进行对比时,在 8 个数据集上有 5 个数据集的 Average Precision 值最大. 根据 8 个数据集上的平均排序结果可以看出,SFS-FMI-SW 位列第一,ML-KNN 排在第二,PMU 排在最后,性能最差.

表 3 在海明损失上五个特征选择算法的排名

Table 3 Ranks among five feature selection methods in Hamming Loss (↓)

Methods	ML-KNN	MLNB	MDDMspc	MDDMproj	PMU	MFNMIpes	SFS-FMI-SW
Yeast	0.2010$_{(2)}$	0.2186$_{(7)}$	0.2132$_{(6)}$	0.2080$_{(3)}$	0.2103$_{(4)}$	0.2116$_{(5)}$	**0.1992**$_{(1)\surd}$
Birds	0.0536$_{(4)}$	0.0610$_{(5)}$	0.0655$_{(7)}$	0.0622$_{(6)}$	0.0529$_{(3)}$	0.0525$_{(2)}$	**0.0503**$_{(1)\surd}$
Cal500	0.1399$_{(3.5)}$	0.1426$_{(6)}$	**0.1395**$_{(1.5)\surd}$	**0.1395**$_{(1.5)\surd}$	0.1431$_{(7)}$	0.1407$_{(5)}$	0.1399$_{(3.5)}$
Artificial	0.0612$_{(5)}$	0.0612$_{(5)}$	0.0615$_{(7)}$	0.0611$_{(3)}$	0.0602$_{(2)}$	**0.0600**$_{(1)\surd}$	0.0612$_{(5)}$
Reference	0.0314$_{(5.5)}$	**0.0296**$_{(1)\surd}$	0.0314$_{(5.5)}$	0.0304$_{(2)}$	0.0311$_{(4)}$	0.0310$_{(3)}$	0.0315$_{(7)}$
Health	0.0458$_{(7)}$	0.0420$_{(2)}$	0.0428$_{(4)}$	0.0430$_{(5)}$	0.0455$_{(6)}$	0.0422$_{(3)}$	**0.0406**$_{(1)\surd}$
Business	0.0269$_{(1)\surd}$	0.0283$_{(5)}$	0.0280$_{(3.5)}$	0.0280$_{(3.5)}$	0.0285$_{(6)}$	0.0286$_{(7)}$	**0.0274**$_{(2)}$
Science	0.0325$_{(1)\surd}$	0.0346$_{(4)}$	0.0347$_{(5.5)}$	0.0348$_{(7)}$	0.0344$_{(3)}$	**0.0343**$_{(2)}$	0.0347$_{(5.5)}$
Average	3.625	4.375	5	3.875	4.375	3.5	3.25

2) 表 3 实验结果表明: 本文算法与 MLNB、MDDMspc、MDDMproj、PMU、MFNMIpes 这 5 个算法进行对比时,8 个数据集中有 4 个数据集 SFS-FMI-SW 性能优于其他算法,在数据集 Cal500 上,MDDMspc 和 MDDMproj 算法的 Hamming Loss 值相对于 SFS-FMI-SW 减少仅 0.0004,在 Artificial 数据集上,MFNMIpes 的 Hamming Loss 值只比 SFS-FMI-SW 算法减少了 0.0012. 本文算法与所有算法进行对比时,有 3 个数据集的 Hamming Loss 值最小,在 Cal500 和 Artificial 这两个数据集上,本文算法与 ML-KNN 的 Hamming Loss 值大小相同,在其他 3 个数据集上 Hamming Loss 值与其他算法相差甚小. 在 Hamming Loss 指标上,SFS-FMI-SW 排名第 1,性能最好.

表 4 在 1-错误上五个特征选择算法的排名

Table 4 Ranks among five feature selection methods in One Error (↓)

Methods	ML-KNN	MLNB	MDDMspc	MDDMproj	PMU	MFNMIpes	SFS-FMI-SW
Yeast	0.2495$_{(3)}$	0.2658$_{(7)}$	0.2451$_{(2)}$	0.2582$_{(5)}$	0.2647$_{(6)}$	0.2527$_{(4)}$	**0.2342**$_{(1)\surd}$
Birds	0.3910$_{(4)}$	0.4861$_{(7)}$	0.4830$_{(6)}$	0.4427$_{(5)}$	0.3375$_{(2.5)}$	0.3375$_{(2.5)}$	**0.2941**$_{(1)\surd}$
Cal500	0.1076$_{(3)}$	0.1434$_{(7)}$	0.1195$_{(5.5)}$	0.1195$_{(5.5)}$	0.1076$_{(3)}$	0.1076$_{(3)}$	**0.1036**$_{(1)\surd}$
Artificial	0.6327$_{(4)}$	0.6433$_{(5)}$	0.6510$_{(7)}$	0.6457$_{(6)}$	0.6303$_{(3)}$	**0.6120**$_{(1)\surd}$	0.6197$_{(2)}$
Reference	0.4730$_{(4)}$	0.4703$_{(3)}$	0.4630$_{(2)}$	0.4770$_{(6)}$	0.4827$_{(7)}$	0.4767$_{(5)}$	**0.4567**$_{(1)\surd}$
Health	0.4207$_{(7)}$	0.3947$_{(4)}$	0.3903$_{(3)}$	0.4020$_{(5)}$	0.4157$_{(6)}$	0.3843$_{(2)}$	**0.3697**$_{(1)\surd}$
Business	0.1213$_{(1)\surd}$	0.1317$_{(6)}$	0.1283$_{(4)}$	0.1263$_{(3)}$	0.1360$_{(7)}$	0.1293$_{(5)}$	0.1247$_{(2)}$
Science	0.5803$_{(1)\surd}$	0.6813$_{(4)}$	0.6903$_{(6)}$	0.6993$_{(7)}$	0.6880$_{(5)}$	0.6713$_{(3)}$	**0.6660**$_{(2)}$
Average	3.375	5.375	4.4375	5.3125	4.9375	3.1875	1.375

3) 表 4 的结果表明: 本文算法与 MLNB、MDDMspc、MDDMproj、PMU、MFNMIpes 这五个算法进行对比时,有 7 个数据集上 SFS-FMI-SW 算法的 One-Error 值都是最小的,在数据集 Artificial 上,MFNMIpes 的 One-Error 值只比 SFS-FMI-SW 减少了 0.0077. 本文算法与所有算法进行对比时,在 8 个数据集上有 5 个数据集的 One-Error 值最大,其他 3 个数据集的 One-Error 值都排在第二位. 这表明算法 SFS-FMI-SW 的性能很好,在 One-Error 指标上,SFS-FMI-SW 排名第 1.

表 5 在排位损失上五个特征选择算法的排名

Table 5 Ranks among five feature selection methods in Ranking Loss (↓)

Methods	ML-KNN	MLNB	MDDMspc	MDDMproj	PMU	MFNMIpes	SFS-FMI-SW
Yeast	0.1760$_{(2)}$	0.2000$_{(7)}$	0.1893$_{(6)}$	0.1835$_{(3)}$	0.1849$_{(4)}$	0.1881$_{(5)}$	**0.1741**$_{(1)\surd}$
Birds	0.1251$_{(2)}$	0.1463$_{(7)}$	0.1369$_{(5)}$	0.1430$_{(6)}$	0.1258$_{(3)}$	0.1268$_{(4)}$	**0.1163**$_{(1)\surd}$
Cal500	0.1892$_{(1)\surd}$	0.1913$_{(6)}$	0.1897$_{(3.5)}$	0.1897$_{(3.5)}$	0.1918$_{(7)}$	0.1900$_{(5)}$	**0.1895**$_{(2)}$
Artificial	0.1520$_{(2)}$	0.1542$_{(6)}$	0.1537$_{(5)}$	0.1548$_{(7)}$	0.1526$_{(3)}$	0.1530$_{(4)}$	**0.1485**$_{(1)\surd}$
Reference	0.0919$_{(7)}$	0.0889$_{(3)}$	**0.0859**$_{(1)\surd}$	0.0890$_{(4)}$	0.0912$_{(6)}$	0.0911$_{(5)}$	0.0862$_{(2)}$
Health	0.0605$_{(2)}$	0.0641$_{(6)}$	**0.0604**$_{(1)\surd}$	0.0638$_{(5)}$	0.0642$_{(7)}$	0.0608$_{(3)}$	0.0618$_{(4)}$
Business	0.0374$_{(1)\surd}$	0.0419$_{(3)}$	0.0433$_{(5.5)}$	**0.0416**$_{(2)}$	0.0459$_{(7)}$	0.0433$_{(5.5)}$	0.0421$_{(4)}$
Science	0.1166$_{(1)\surd}$	0.1364$_{(3)}$	0.1398$_{(4)}$	0.1441$_{(7)}$	0.1429$_{(6)}$	0.1409$_{(5)}$	**0.1349**$_{(2)}$
Average	2.25	5.125	3.875	4.6875	5.375	4.15625	2.125

4) 从表 5 可看出,本文算法与 MLNB、MDDMspc、MDDMproj、PMU、MFNMIpes 这五个算法进行对比时,算法 SFS-

FMI-SW 在 5 个数据集上 Ranking Loss 值最小,也即性能最好,在数据集 Reference 上,MDDMspc 只比 SFS-FMI-SW 的 Ranking Loss 值减少了 0.0003,在 Health 数据集上,MDDM-spc 比 SFS-FMI-SW 的 Ranking Loss 值也只减少了 0.0014. 本文算法与所有算法进行对比时,有 3 个数据集的 Ranking Loss 值最小,Cal500、Artificial 和 Science 这 3 个数据集的 Ranking Loss 值排在第 2 位. 在 8 个数据集上的综合排位中,SFS-FMI-SW 排在第 1.

表 6　在覆盖率上五个特征选择算法的排名

Table 6　Ranks among five feature selection methods in Coverage (↓)

Methods	ML-KNN	MLNB	MDDMspc	MDDMproj	PMU	MFNMIpes	SFS-FMI-SW
Yeast	6.4058$_{(1)}$√	6.7516$_{(7)}$	6.6307$_{(5)}$	6.5490$_{(3)}$	6.5784$_{(4)}$	6.7266$_{(6)}$	**6.4216**$_{(2)}$
Birds	3.3994$_{(2)}$	3.7678$_{(6)}$	3.6099$_{(5)}$	3.8173$_{(7)}$	3.4520$_{(4)}$	3.4458$_{(3)}$	**3.3127**$_{(1)}$√
Cal500	130.6255$_{(6)}$	131.4343$_{(7)}$	130.4861$_{(4.5)}$	130.4861$_{(4.5)}$	129.9801$_{(3)}$	**129.0837**$_{(1)}$√	129.9363$_{(2)}$
Artificial	5.4453$_{(2)}$	5.5040$_{(6)}$	5.4877$_{(4)}$	5.5557$_{(7)}$	5.4850$_{(3)}$	5.4973$_{(5)}$	**5.4023**$_{(1)}$√
Reference	3.5420$_{(7)}$	3.4313$_{(3)}$	**3.3403**$_{(1)}$√	3.4323$_{(4)}$	3.5143$_{(6)}$	3.5040$_{(5)}$	3.3490$_{(2)}$
Health	3.3047$_{(2)}$	3.4163$_{(6)}$	**3.2777**$_{(1)}$√	3.4150$_{(5)}$	3.4333$_{(7)}$	3.3290$_{(3)}$	3.3787$_{(4)}$
Business	2.1847$_{(1)}$√	2.3483$_{(3)}$	2.3913$_{(5)}$	**2.3370**$_{(2)}$	2.4800$_{(7)}$	2.4290$_{(6)}$	2.3597$_{(4)}$
Science	6.0430$_{(1)}$√	6.8467$_{(3)}$	6.9740$_{(4)}$	7.1670$_{(7)}$	7.0957$_{(6)}$	7.0377$_{(5)}$	**6.8193**$_{(2)}$
Average	2.75	5.125	3.6875	4.9375	5	4.25	2.25

5) 根据表 6 所有算法的 Coverage 值可看出: 本文算法与 MLNB、MDDMspc、MDDMproj、PMU、MFNMIpes 这五个算法进行对比时,8 个数据集中有 4 个数据集的 Coverage 值最小,即效果最好,在数据集 Cal500 上的 Coverage 值与 MFNMIpes 算法的值相差不超过 0.8526,在数据集 Reference 上,MDDM-spc 的 Coverage 值相对于 SFS-FMI-SW 仅减小了 0.0087,差距非常小. 本文算法与所有算法进行对比时,算法 SFS-FMI-SW 的 Coverage 值的排序都位于前四. 从平均排序中来看,SFS-FMI-SW 排在第 1 位.

6) 从图 2-图 3 可发现: 图 2 和图 3 分别为数据集 Business 和 Science 在 SFS-FMI-SW 算法上同 MDDMspc、MDDMproj、PMU、MFNMIpes 在 5 个评价指标上的分类性能对

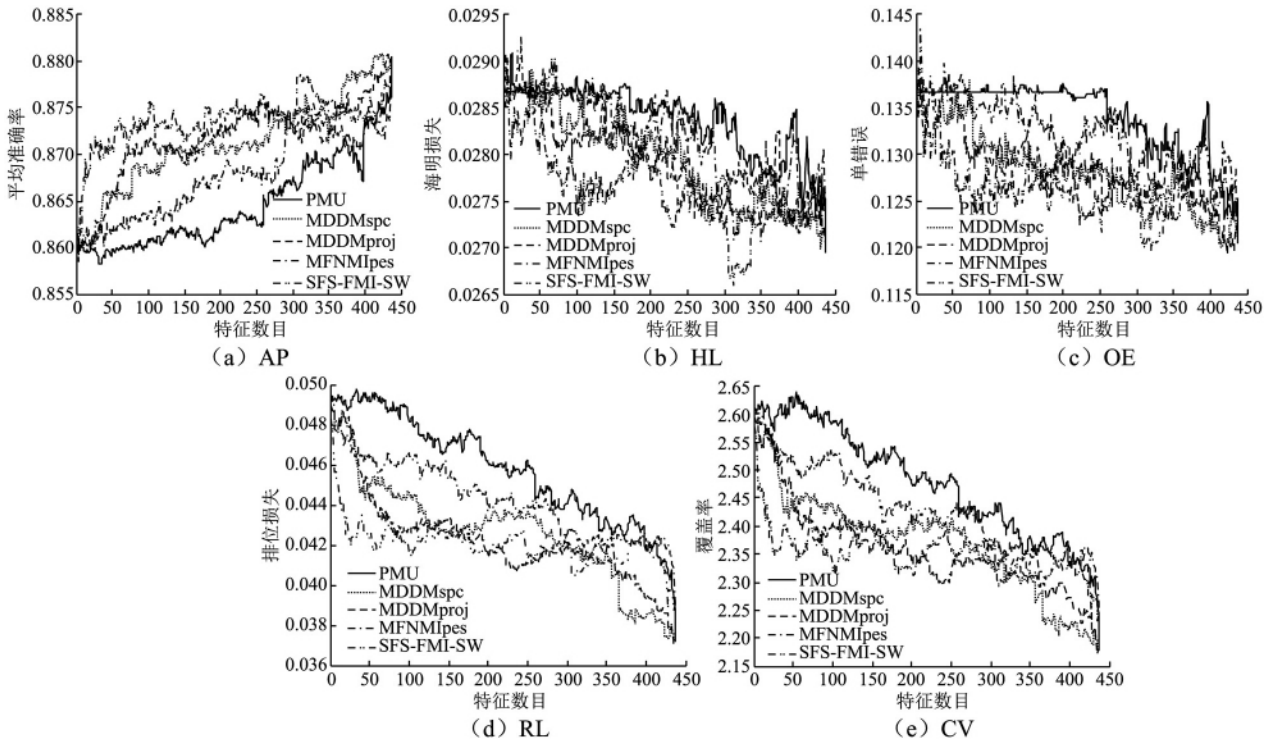

图 2　Business 数据集的各个评价指标的性能变化

Fig. 2　Changes in performance of various evaluation indicators of Business data set

比,由图可能看出在本文 SFS-FMI-SW 算法上分类性能曲线的变化情况明显占优. 类似的,其他数据集也可以绘制分类性能曲线,限于篇幅,略.

7) 根据以上实验结果分析以及算法在 5 个评价指标上的实验结果可知,本文算法在大部分评价指标上优于其他对比算法,特别是在评价指标 Average Precision 和 One Error 上,性能基本优于其他算法,所有算法在各评价指标的综合排位中,本文所提的算法均位列第 1,这也验证了本文提出的 SFS-

FMI-SW 算法是有效的.

4.4 算法有效性统计性假设检验

为了进一步验证本文算法的有效性,运用统计学知识,在8个数据集上,本文采用显著性水平为 0.1 的 Friendman test [29],对于每个评价指标,我们都拒绝零假设,若两个算法在所有数据集上的平均排序的差高于临界差(critical SFS-difference,简称 CD),则认为它们有显著性差异. 图4给出了所有算法在不同评价指标上的比较,根据公式(13)可计算出 CD = 2.9085($K=7$,$N=8$),坐标轴画出了各种算法的平均排序,并且坐标轴上的数字越小则表示平均排序越低,线段连接则表示两种算法性能没有显著差异.

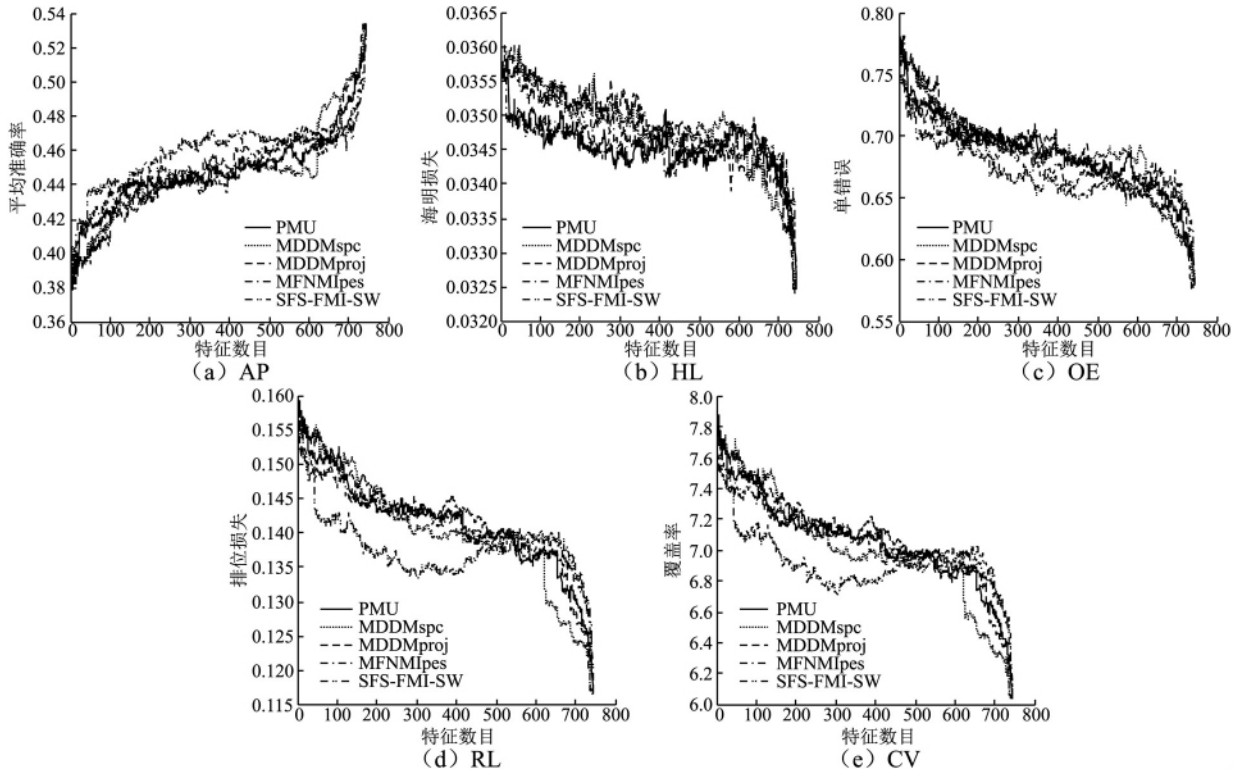

图3 Science 数据集的各个评价指标的性能变化

Fig. 3 Changes in performance of various evaluation indicators of Science data set

$$CD_{\partial} = q_{\partial}\sqrt{\frac{k(k+1)}{6N}} \tag{13}$$

从图4不难发现:在5个评价指标上,本文提出的算法

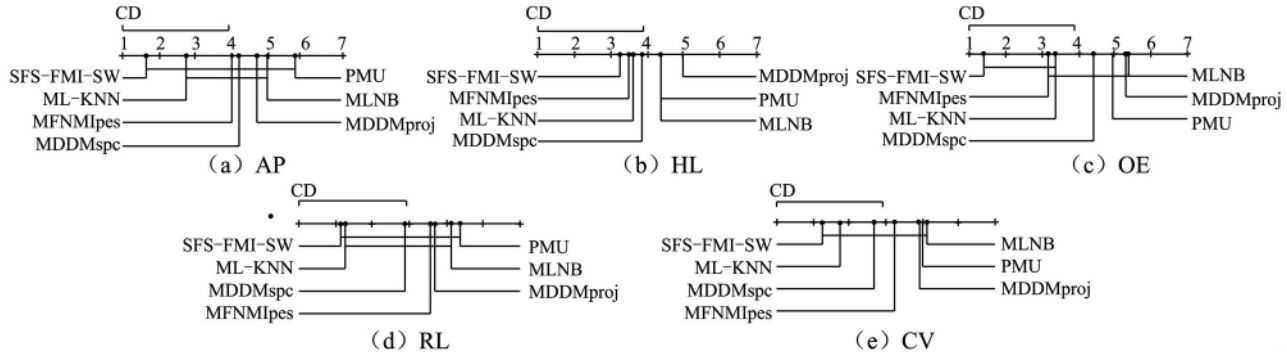

图4 使用 Friendman test 测试比较几种算法的性能

Fig. 4 Comparison of SFS-FMI-SW against other comparing algorithms with the Friendman test

FMI-SW 都基本优于其他算法.

5 结 语

本文引入新的模糊信息熵,提出了结合滑动窗口机制特征选择算法,尝试解决动态或增量问题中流特征选择问题,实验结果和假设检验进一步说明本文算法是有效的. 不足之处是本文算法将窗口设置为固定大小,在一定程度上影响算法的效率和性能. 实际情况中,由于每个时刻产生的特征个数可能不同,窗口过小则不符合实际需求,窗口过大则使内存资源无法得到充分利用. 因此,下一步将考虑根据每个时刻产生特征的个数而调整合适大小的窗口使其更符合对流特征的选择问题.

References:

[1] Kumar V,Minz S. Multi-view ensemble learning: an optimal feature set partitioning for high-dimensional data classification [J]. Knowledge & Information Systems,2015,2015: 1-59.

[2] Kong X,Yu P S. Multi-label feature selection for graph classification [C]. IEEE,International Conference on Data Mining,IEEE,2011: 274-283.

[3] Lee J,Kim D W. Fast multi-label feature selection based on information-theoretic feature ranking [J]. Pattern Recognition,2015,48(9): 2761-2771.

[4] Chen H,Li T,Luo C,et al. A decision-theoretic rough set approach for dynamic data mining [J]. IEEE Transactions on Fuzzy Systems,2015,23(6): 1958-1970.

[5] Javidi M M,Eskandari S. Stream wise feature selection: a rough set method [J]. International Journal of Machine Learning & Cybernetics,2018,9(4): 667-676.

[6] Zhang L,Hu Q,Duan J,et al. Multi-label feature selection with fuzzy rough sets [M]. Rough Sets and Knowledge Technology. Springer International Publishing,2014: 121-128.

[7] Lee J,Kim D W. Feature selection for multi-label classification using multivariate mutual information [J]. Pattern Recognition Letters,2013,34(3): 349-357.

[8] Lin Y,Hu Q,Liu J,et al. Multi-label feature selection based on neighborhood mutual information [J]. Applied Soft Computing,2016,38(C): 244-256.

[9] Zhang Y,Zhou Z H. Multi-label dimensionality reduction via dependence maximization [C]. AAAI Conference on Artificial Intelligence,AAAI 2008,Chicago,Illinois,USA,July. DBLP,2008: 1503-1505.

[10] Zhang M L,Robles V. Feature selection for multi-label naive Bayes classification [J]. Information Sciences,2009,179(19): 3218-3229.

[11] Wu X,Yu K,Ding W,et al. Online feature selection with streaming features [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2013,35(5): 1178-1192.

[12] Lin Y,Hu Q,Zhang J,et al. Multi-label feature selection with streaming labels [J/OL]. Information Sciences,2016,372: 256-275. https://doi.org/10.1016/j.ins.2016.08.039.

[13] Eskandari S,Javidi M M. Online streaming feature selection using rough sets [M]. Elsevier Science Inc.,2016.

[14] Chen H,Li T,Luo C,et al. A decision-theoretic rough set approach for dynamic data mining [J]. IEEE Transactions on Fuzzy Systems,2015,23(6): 1958-1970.

[15] Zhao Lin,Li Jiu-shun,Cheng Jian-hua. Innovation-based adaptive Kalman filter with sliding window for integrated navigation [J]. Systems Engineering and Electronics,2017,39(11): 155-159.

[16] Chang Jian-long,Cao Feng,Zhou Ao-ying. Clustering evolving data streams over sliding windows [J]. Journal of Software,2007,18(4): 905-918.

[17] Yu K,Ding W,Wu X. LOFS: a library of online streaming feature selection [J]. Knowledge-Based Systems,2016,113: 1-3.

[18] Rahmaninia M,Moradi P. OSFSMI: online stream feature selection method based on mutual information [J]. Applied Soft Computing,2018,68: 733-746.

[19] Datar M,GionisA,Indyk P,et al. Maintaining stream statistics over sliding windows [J]. Siam Journal on Computing,2002,31(6): 1794-1813.

[20] Guha S,Meyerson A,Mishra N,et al. Clustering data streams: theory and practice [J]. IEEE Transactions on Knowledge and Data Engineering,2003,15(3): 515-528.

[21] Lin Y,Hu Q,Liu J,et al. Multi-label feature selection based on max-dependency and min-redundancy [J]. Neuro Computing,2015,168(C): 92-103.

[22] Liu Jing-hua,Lin Meng-lei,Wang Chen-xi,et al. Multi-label feature selection algorithm based on local subspace [J]. Pattern Recognition and Artificial Intelligence,2016,29(3): 240-251.

[23] Cheng Yu-sheng,Zhang You-sheng,Hu Xue-gang. Entropy of knowledge and rough set based on boundary region [J]. Journal of System Simulation,2007,19(9): 2008-2011.

[24] Yu S,Huang T Z. Exponential weighted entropy and exponential weighted mutual information [J]. Neuro Computing,2017,249: 86-94.

[25] Kamimura R. Collective mutual information maximization to unify passive and positive approaches for improving interpretation and generalization [J]. Neural Networks,2017,90: 56-71.

[26] Lin Y,Hu X,Wu X. Quality of information-based source assessment and selection [J]. Neurocomputing,2014,133(133): 95-102.

[27] Wu Wei-zhi. An uncertainty measure in partition-based fuzzy rough sets [J]. International Journal of General Systems,2005,34(1): 77-90.

[28] Zhang M L,Zhou Z H. ML-KNN: a lazy learning approach to multi-label learning [J]. Pattern Recognition,2007,40(7): 2038-2048.

[29] Ar J. Statistical comparisons of classifiers over multiple data sets [J]. Journal of Machine Learning Research,2006,7(1): 1-30.

附中文参考文献:

[15] 赵 琳,李久顺,程建华. 基于滑动窗口的新息自适应组合导航算法 [J]. 系统工程与电子技术,2017,39(11): 155-159.

[16] 常建龙,曹 锋,周傲英. 基于滑动窗口的进化数据流聚类 [J]. 软件学报,2007,18(4): 905-918.

[22] 刘景华,林梦雷,王晨曦,等. 基于局部子空间的多标记特征选择算法 [J]. 模式识别与人工智能,2016,29(3): 240-251.

[23] 程玉胜,张佑生,胡学钢. 基于边界域的知识粗糙熵与粗集粗糙熵 [J]. 系统仿真学报,2007,19(9): 2008-2011.