结合无监督学习的数据流分类算法*

徐树良[1] 王俊红[1,2]

[1](山西大学 计算机与信息技术学院 太原 030006)
[2](山西大学 计算智能与中文信息处理教育部重点实验室 太原 030006)

摘 要 为了能有效应对数据流中的概念漂移现象,提出结合无监督学习的数据流分类算法. 该算法以集成式分类技术为基础,在分类过程中引入属性约简,利用聚类算法对数据进行聚类,通过对比分类和聚类结果的准确率,判断是否发生概念漂移. 实验表明,文中算法在综合时间花销和准确率上取得较好效果.

关键词 数据流, 概念漂移, 集成式分类, 属性约简, 无监督学习
中图法分类号 TP 181 **DOI** 10.16451/j.cnki.issn1003-6059.201607011
引用格式 徐树良,王俊红. 结合无监督学习的数据流分类算法. 模式识别与人工智能, 2016, 29(7): 665-672.

Classification Algorithm Combined with Unsupervised Learning for Data Stream

XU Shuliang[1], WANG Junhong[1,2]
[1](*School of Computer and Information Technology*, *Shanxi University*, *Taiyuan* 030006)
[2](*Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education*, *Shanxi University*, *Taiyuan* 030006)

ABSTRACT

An ensemble learning techniques based algorithm combined with unsupervised learning is proposed for concept drift problem of data stream. An attribute reduction mechanism is introduced into classification process and then a clustering algorithm is applied to the data for clustering. Accuracies of classification and clustering are compared to decide whether concept drift appears or not. The experimental results show that the proposed algorithm efficiently decreases time consumption and improves the precision.

Key Words Data Stream, Concept Drift, Ensemble Classification, Attribute Reduction, Unsupervised Learning
Citation XU S L, WANG J H. Classification Algorithm Combined with Unsupervised Learning for Data Stream. Pattern Recognition and Artificial Intelligence, 2016, 29(7): 665-672.

* 国家自然科学基金项目(No. 61305057,61303008,61202018)、山西省青年科技基金项目(No. 2013021018-1)、山西省高等学校科技创新项目(No. 2013102)资助
Supported by National Natural Science Foundation of China (No. 61305057,61303008,61202018), Shanxi Province Science Foundation for Youths (No. 2013021018-1), Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi (No. 2013102)
收稿日期:2015-07-08;修回日期:2015-10-29;录用日期:2015-11-11
Manuscript received July 8, 2015; revised October 29, 2015; accepted November 11, 2015

在生产实践中,随着信息社会的发展,许多领域产生了大量的数据流,这些数据快速到达,动态变化及存在概念漂移,使得传统的数据挖掘方法面临巨大的挑战. 因此,如何从数据流中挖掘人们感兴趣的知识和模式,成为当今数据挖掘领域一个研究热点[1].

自从数据流模型提出后,数据流分类研究得到学者们的广泛关注[2-5]. Street 等[6]提出集成式数据流分类算法(A Streaming Ensemble Algorithm, SEA),通过替换分类性能较差的子分类器实现分类模型的更新. Domingos 等[7]提出基于 Hoeffding 边界的快速决策树算法(Very Fast Decision Tree, VFDT),建立一棵 Hoeffding 树,在训练过程中不断把叶节点变为决策节点,最终新数据通过测试决策节点进入叶节点,完成分类. 由于 VFDT 只能处理离散属性和缺少概念漂移监测机制,Hulten 等[8]提出针对 VFDT 的改进型自适应概念的快速决策树算法(Concept-Adapting Very Fast Decision Tree Learner, CVFDT),在发生概念漂移时,建立一棵备选子树替换当前决策节点的子树,实现概念转移. Wu 等[9]提出基于半监督的数据流分类算法(A Semi-supervised Classification Algorithm for Data Streams with Concept Drifts and Unloabeled Data, SUN),在只有部分数据被标记的情况下,该算法能有效区分概念漂移和噪声. Elwell 等[10]针对周期性概念漂移问题,提出非稳定环境中的增量式加权学习算法(Learners in Nonstationary Environments, Learn++. NSE),在分类过程中保存历史概念,利用当前分类的错误率决定分类器的权值,最终的分类结果采用加权投票决定. Liao 等[11]提出针对概念漂移类型不确定问题的 AGE(Accuracy and Growth Rate Updated Ensemble),对每个分类器设置双权值,考虑历史概念和当前概念,能处理不同类型的概念漂移.

基于上述分析,本文针对数据流问题,设计结合无监督学习的数据流分类算法(Classification Algorithm Combined with Unsupervised Learning for Data Stream, CAU),以集成式分类为基础,在分类过程中引入属性约简和聚类分析,通过对比分类和聚类的结果检测是否发生概念漂移. 最终实验证明本文算法能较好地平衡准确率和时间开销.

1 概念漂移和集成式分类

假设 t 为时间变量,d_t 为 t 时刻进入滑动窗口的数据块,数据流可表示为

$$S=\{\cdots,d_{t-1},d_t,d_{t+1},\cdots\},$$

t 时刻对训练数据进行训练得到的目标概念为 M,经过一段时间 Δt 后,再次对当前数据进行训练,得到的目标概念为 N. 如果概念 M 与概念 N 不同,称数据流在 Δt 时间内发生概念漂移[12].

根据 Δt 的不同可把概念漂移分为如下 2 种类型:当 Δt 很短时,发生的概念漂移称为突变式概念漂移;当 Δt 较长时,发生的概念漂移称为渐进式概念漂移. 由文献[10]可知,当发生概念漂移时,数据的后验概率发生改变,即

$$P_t(\omega\mid x)\neq P_{t+1}(\omega\mid x),$$

由此会引起

$$P(x,\omega)=P(\omega\mid x)\cdot P(x)$$

的改变,最终表现为数据的分布发生改变. 在发生概念漂移后,如果不采取相应措施,会使得系统的分类性能大幅度下降,在实际应用中可利用概念漂移的这种特性进行入侵检测、欺诈检测、网络舆论分析等.

集成式分类是一种有效的数据流分类方法,该方法结合多个分类器,能有效处理概念漂移问题,在分类过程中赋予不同分类器相应的权值,通过动态调整和淘汰基分类器实现分类器的更新,从而适应数据流中的新概念,数据类标签最终通过投票机制做出决策[13]. 在数据流环境下,集成式分类器的分类性能优于单个分类器[14]. 由于集成式分类具有上述优点,本文算法以集成式分类技术作为基础.

2 结合无监督学习的数据流分类算法

在数据流分类算法中,结合无监督学习的分类技术大多应用在数据被部分标记的分类问题中,本文将聚类算法引入到数据被完全标记的分类中. 由于聚类算法的最终结果只依赖于当前数据,并不像分类那样需要根据数据当前的概念调整分类模型,因此可利用数据的先验知识,把聚类结果映射到真实类标签中,从而反映当前数据的真实概念. 通过对比分类与聚类的结果,判断分类模型是否适用于当前数据.

2.1 分类器的权值确定和分类器系统更新

在 CAU 中,当滑动窗口获得一个新数据块 B_i,如果系统中子分类器的数目未达到上限,利用当前到来的数据块训练一个分类器,加入到集成分类器中,新分类器的权值

$$w_j = \frac{1}{MSE_r + \varepsilon}, \tag{1}$$

其中,ε 为一个很小的常量,是为了防止式(1)中除数为0,MSE_r 为随机分类器对数据块 B_i 分类的均方误差,

$$MSE_r = \sum_y p(y)(1 - p(y))^2,$$

y 为数据块 B_i 中不同类标签的集合,$p(y)$ 为随机分类器将实例标记为 y 类的概率.

若系统中的子分类器数目超过上限,此时先利用系统中的集成分类器分类 B_i,分类结果采用投票机制决定,计算集成分类器的错误率 $error_i$:

$$error_i = \frac{1}{winsize}\sum_{i=1}^{winsize} predict_{(x_i,y)\in B_i}(x_i). \tag{2}$$

其中:如果 x_i 分类正确,$predict(x_i) = 0$,如果 x_i 分类错误,$predict(x_i) = 1$;$winsize$ 为数据块 B_i 包含的实例数目.

对于每个分类器,MSE_{ij} 为第 j 个子分类器在数据块 B_i 上分类结果的均方误差,即

$$MSE_{ij} = \frac{1}{|B_i|}\sum_{\{x,y\}\in B_i}(1 - f_y^j(x))^2,$$

其中,$f_y^j(x)$ 为在 B_i 中第 j 个分类器将实例 x 标记为 y 类的概率[14],

$$f_y^j(x) = p(y \mid x) + \beta_y^j + \eta_y^j(x),$$

β_y^j 为在输入 x 时分类器 C_j 的偏倚,根据文献[14]可知,计算 $f_y^j(x)$ 时,β_y^j 可忽略,$\eta_y^j(x)$ 为在输入 x 时,分类器 C_j 根据 0-1 损失计算的输出结果方差,由于 $\eta_y^j(x)$ 一般情况下都很小,为了简化计算,

$$f_y^j(x) = p(y \mid x).$$

对数据块分类完成后,根据每个子分类器对数据块 B_i 的分类结果,更新子分类器的权值,结合上述各公式,权值 w_{ij} 计算如下:

$$w_{ij} = \frac{1}{MSE_r + MSE_{ij} + \varepsilon}. \tag{3}$$

子分类器的权值更新完成后,再利用聚类算法对数据块进行聚类. 为了提高聚类结果的准确率,若聚类算法需要事先指明类簇数 k 值和初始种子点,这两者分别等于数据块 B_i 中不同类标签个数和各类点的均值,最终聚类结果的真实类标签采用多数投票决定,即簇中出现次数最多的类标签为该簇所有点的类标签. 聚类完成后,计算聚类结果错误率 $kerror_i$,$kerror_i$ 的计算方法与式(2)一致. 如果两者的错误率

$$|kerror_i - error_i| > \theta,$$

则认为发生概念漂移,用当前数据块 B_i 训练一个新分类器 C_{new},使用 C_{new} 替换权值最小的分类器.

2.2　本文算法中的属性约简

在 CAU 中,为了消除无关属性的影响,引入属性约简过程,属性约简算法步骤如下所示.

算法 1　属性约简

输入　数据块 B_i,数据块大小 $winsize$,
数据块中数据的维数 col

输出　经过约简的数据集 B_i

```
i = 1;
while i >= 1&&i <= (col - 1)
                        // 第 col 维为实例类标签
{
    对B_i 使用 K-means 算法进行聚类,
    k = length(unique(B_i_label)),
    初始种子点分别为 B_i 中各类点的均值,利用式(2)计算聚类的错误率err_0;
    删除 B_i 中第 i 维数据,得到数据集 D_i,使用上述的聚类算法对 D_i 进行聚类,利用式(2)计算聚类结果的错误率 err_i;
    if (err_0 >= err_i)          // 第 i 维属性可约简
    {
        B_i = D_i;
    }
    else
    {
        i ++;
    }
}
```

2.3　算法实现

基于上述分析,CAU 最终的具体执行过程如下所示.

算法 2　CAU

输入　集成式分类器 $ensemble = NULL$,数据流 S,
子分类器的容量 K,数据块大小 $winsize$,
当前子分类器的数目 $num = 0$

输出　训练好的集成式分类器 $ensemble$

```
While(S! = NULL)
{
    读取 winsize 条数据形成数据块 B_i;
    if(num < k)
    {
        使用B_i 训练一个新分类器 C_j,其中 C_j 的权值使用式(1) 计算;
        ensemble = ensemble ∪ C_j;
        num ++;
```

```
    }
    else
    {
        使用 ensemble 分类 B_i,
        采用投票机制计算错误率 error_i;
        for each ensemble(i) ∈ ensemble
        {
        根据 ensemble(i) 对 B_i 分类结果,使用式(3)更新权值;
        }
        对 B_i 进行属性约简;
        使用聚类算法对 B_i 进行聚类,若聚类算法需指明类簇数 k 和初始中心点,则
        k = length(unique(B_i_label)),
        初始种子点分别为 B_i 中各类点的均值,根据聚类结果计算 kerror_i;
        if | kerror_i - error_i | > θ        // 发生概念漂移
        {
            使用 B_i 创建一个新分类器 C_new, 替换 ensemble(j),
            j = arg min(ensemble(j).weight);
        }
    }
    使用 B_i 训练 ensemble 中每个子分类器;
}
```

在 CAU 中,使用聚类算法把聚类结果作为判断发生概念漂移的基准.考虑到实际数据集中有些维度可能是无关属性,为了消除无关属性对聚类结果的影响,对数据集的属性进行约简. CAU 中的聚类实际是一个充分利用先验知识的聚类,若算法需要指明类簇数 k 和初始中心点,则两者都由训练集决定,并不是依据人的经验,这样做的主要目的是为了提高聚类结果的准确率,使聚类结果尽可能与实际相符.

CAU 采用准确率之差检测概念漂移,这种方法的原理如下:若数据流发生概念漂移,集成式分类器由于缺少对新概念的训练,准确率必定会急剧下降,而聚类算法的准确率只依赖于数据集,与概念变化并不相关,因此聚类结果的准确率保持稳定. 当两者准确率差值增大而超过系统设置的阈值时,就产生分类器的替换,因此可用两者准确率之间差值的绝对值监测数据概念变化的情况.

对于阈值 θ 的取值,可利用 Hoeffding 边界限定取值范围,Hoeffding 边界[8]表述为:通过 n 次独立观察随机变量 r,得出随机变量 r 的取值范围为 R,在 $1-\delta$ 的置信水平下,r 的真实值至少为 $\bar{r}-\nu$,其中

$$\nu=\sqrt{\frac{R^2}{2n}\ln\left(\frac{1}{\delta}\right)},$$

即在[0,1]内,观察值的均值与真实值之差不超过 ν. 由于 $0\leqslant r\leqslant 1$,则 $R=1$,因此可限定 θ 取值不超过 $2\sqrt{\frac{1}{2n}\ln\left(\frac{1}{\delta}\right)}$.

3 实验及结果分析

为了验证 CAU 的有效性,本文选取 AGE[11]、基于 C4.5 决策树和 NaiveBayes 混合模型的数据流分类算法(Classification Algorithm for Mining Data Streams Based on Mixture Models of C4.5 and NB, CDSMM)[15]、基于多分类器的改进型 ID4 算法(An Improved ID4 Algorithm Based on Multi-classifiers, M_ID4)[16] 作为对比算法,分别在 waveform+noise、Hyperplane、sensor_readings_4 这 3 个数据集上进行测试.

实验环境为 Windows 7 操作系统,Intel Core 2.94 GHz 双核 CPU,4 GB 内存,算法程序由 Matlab R2013a 实现.

在本次实验中,AGE 的基分类器使用 C4.5 算法实现,M_ID4 的基分类器采用分类与回归树算法(Classification and Regression Tree, CART)实现,CAU 的基分类器采用 NaiveBayes 算法实现,CDSMM 的基分类器采用文献[15]中给出的算法实现. 在本文实验中,无特殊说明,CAU 中的聚类算法采用 K-means 算法.

3.1 实验数据集

waveform + noise 数据集中共有 5 000 条数据,每条数据 41 维,前 40 个属性的取值为[0,6]内的实数,其中 19 个属性为无关属性,在所有数据中,共包含 3 个不同的类标签.

Hyperplane 数据集为一个人工数据,一个 d 维超平面样本 X 满足如下的数学表达式:

$$\sum_{i=1}^{d}a_ix_i=a_0.$$

在本文中的 Hyperplane 数据集在海量数据在线分析(Massive Online Analysis, MOA)环境[17] 中随机生成,共包含 5 000 条数据,共 21 维,前 20 维的 a_i 取值均在[0,1]内. 当数据满足

$$\sum_{i=1}^{d}a_ix_i\geqslant a_0$$

时,类标记为 1;否则,类标记为 2,其中

$$a_0 = \frac{1}{2}\sum_{i=1}^{d} a_i.$$

sensor_readings_4 数据集共包含 5 456 条数据,每条数据 5 维,数据含有 4 个类标签.

3.2　实验性能分析

为了验证算法的分类性能,选择 waveform + noise 数据集作为实验数据,在实验中 $winsize = 200$,分类器的最大数目 $K = 5$,CDSMM 中 $\partial = 0.95$,M_ID4 中 $\theta = 0.01$,AGE 中 $\varepsilon = 0.000\ 1$,CAU 中 $\theta = 0.2$,$\varepsilon = 1 \times 10^{-9}$. 后续实验中参数设置相同. 算法的测试结果如图 1 和表 1 所示.

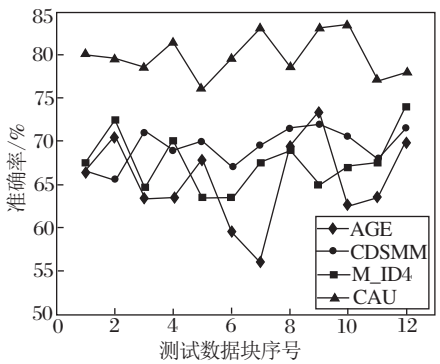

图 1　4 种算法在 waveform+noise 数据集上的测试结果

Fig. 1　Testing results of 4 algorithms on waveform + noise dataset

表 1　4 种算法在 waveform+noise 数据集上的运行时间和平均准确率

Table 1　Running time and average accuracies of 4 algorithms on waveform+noise dataset

	AGE	CDSMM	M_ID4	CAU
平均准确率/%	65.54	69.33	67.63	**79.83**
运行时间/s	1072.190	547.128	612.341	11.6039

由图 1 和表 1 可知,CAU 不但准确率高于其它 3 种算法,而且算法的运行时间也最低,在时间和准确率上取得较好效果. 在本次实验中,AGE、CDSMM 和 M_ID4 的准确率都远低于 CAU,这主要与其分类方法相关. 在这 3 种算法中,都采用决策树构建分类器,在 waveform + noise 数据集中,共有 19 个属性为无关属性,大量无关属性使得算法构建的分类器并不可靠. 而在 CAU 中,使用贝叶斯算法构建分类器,数据分类的结果采用后验概率决定,因此能降低无关属性对决策的影响.

为了进一步验证算法的分类性能,选择 Hyperplane 数据集作为实验数据,向实验数据集中添加 5% 的噪声,结果如图 2 和表 2 所示.

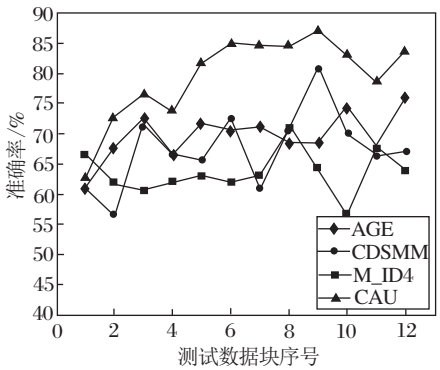

图 2　4 种算法在 Hyperplane 数据集上的测试结果

Fig. 2　Testing results of 4 algorithms on Hyperplane dataset

表 2　4 种算法在 Hyperplane 数据集上的运行时间和平均准确率

Table 2　Running time and average accuracies of 4 algorithms on Hyperplane dataset

	AGE	CDSMM	M_ID4	CAU
平均准确率/%	69.67	67.50	63.54	**79.38**
运行时间/s	282.546	187.106	284.932	2.125

由图 2 和表 2 可知,CAU 的性能仍要明显高于其它 3 种算法. 在 Hyperplane 数据集中,产生渐进式概念漂移. AGE 采用权值衰减机制淘汰分类器,把一个分类器的权值衰减到一个较低的水平往往需要较长的训练时间,因此当发生概念漂移时,不能及时删除分类性能较差的分类器,影响 AGE 的准确率. CDSMM 根据相邻 2 个数据块的准确率差值的统计量判断是否发生概念漂移,但这往往对突变式概念漂移较有效,在发生渐进式概念漂移时,相邻 2 个数据块准确率的差值较小,因此算法有可能会漏检数据中的概念漂移,使得分类器系统未能得到及时更新. M_ID4 以单个事例为对象更新分类器系统,尽管这种分类策略对概念的变化较敏感,能有效检测概念漂移,但在噪声环境下,只要出现一次分类错误,分类器就被新分类器替换,因为新加入的分类器都未得到充分训练,所以 M_ID4 的分类性能较差. 在 CAU 中,采用分类与聚类结果的准确率之差判断是否发生概念漂移,由于聚类结果的准确率只依赖于数据集,与概念的变化无关,当发生概念漂移后,分类结果的准确率必定会有较大幅度下降,因而能有效检测概念变化,更新实现分类器系统,所以最终算法分类结果准确率较高.

为了验证算法对突变式概念漂移的分类效果,

选用 sensor_readings_4 数据集作为实验数据,最终结果如图 3 和表 3 所示.

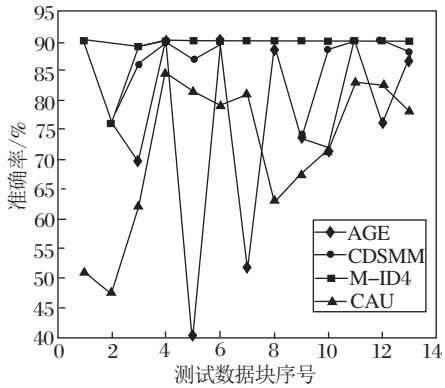

图 3 4 种算法在 sensor_readings_4 数据集上的测试结果

Fig. 3 Testing results of 4 algorithms on sensor_readings_4 dataset

表 3 4 种算法在 sensor_readings_4 数据集上运行时间和平均准确率

Table 3 Running time and average accuracies of 4 algorithms on sensor_readings_4 dataset

	AGE	CDSMM	M_ID4	CAU
平均准确率/%	86.35	93.65	**98.85**	81.69
运行时间/s	129.299	53.435	124.996	1.605

由图 3 和表 3 可知,在 sensor_readings_4 数据集中,M_ID4 的分类准确率最高,CAU 的分类准确率低于 AGE、CDSMM 和 M_ID4. 出现这种现象主要是由于 sensor_readings_4 数据集的数据分布呈现出高度偏斜造成. 在 sensor_readings_4 数据集中,数据的概念呈现重复性突变式变化,在每个数据块中,同类标签的数据往往高度集中,使得数据分布高度偏斜. 在 CAU 中,为了检测概念漂移,使用 *K*-means 算法,*K*-means 算法对数据集的形状较敏感,高度偏斜分布的数据会降低 *K*-means 算法聚类结果的准确率. 在概念漂移的检测中,系统会产生误报,把符合当前概念的分类器淘汰,所以最终影响 CAU 的性能. 而 CDSMM 和 M_ID4 都是以整个数据块的分类结果判断是否需要替换基分类器,同类标签高度集中,会使得训练集和测试集数据分布更相似,训练的分类模型与测试数据的实际情况较一致,所以大幅度提高分类结果的准确率,使得这 2 种算法的准确率最高.

从实验结果来看,相比 AGE、CDSMM 和 M_ID4,CAU 的分类准确率分别相差 4.66%、11.96%、17.16%,但 AGE、CDSMM 和 M_ID4 的时间开销分别为 CAU 的 80.56 倍、33.29 倍和 77.88 倍,在对时间性能要求较高时,选用 CAU 能对时间和准确率取得一个较好平衡.

为了验证 CAU 的抗噪性,选用 waveform + noise 数据集作为实验数据,分别向数据集中添加 5%、10%、15%、20%、25% 的噪声,运行 AGE、CDSMM、M_ID4、CAU,最终结果如图 4 所示.

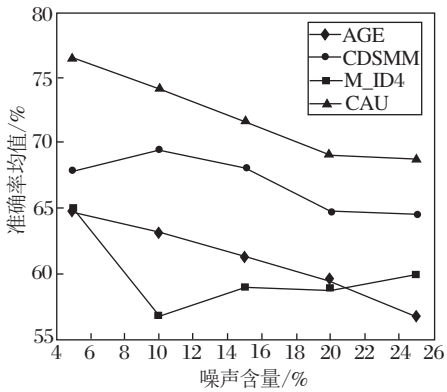

图 4 4 种算法在不同噪声情况下的测试结果

Fig. 4 Testing results of 4 algorithms with different noise

由图 4 可知,在噪声含量较大时,算法的准确率都有不同程度的下降,相比另外 3 种算法,CAU 在不同噪声情况下准确率始终最高,性能表现最佳. 在本次实验中,CAU 使用 *K*-means 算法,具有较好的抗噪能力,主要是因为 *K*-means 算法对噪声较敏感,如果把噪声点选为初始种子点,会大幅降低算法性能,但在实验的 CAU 中,使用结合先验知识的 *K*-means 算法,k 值由训练集中不同类标签数决定,初始种子点为训练集中各类标签数据点的均值,算法的基分类器采用 NaiveBayes 算法实现. NaiveBayes 利用后验概率决定各实例的类标签,本身具有较好的抗噪性,实验中的 CAU 通过求取均值和采用 NaiveBayes 分类器,能有效降低噪声对聚类结果的影响,使得算法在噪声数据集中表现出较好的抗噪能力.

为了进一步研究 CAU 中约简算法对分类性能的影响,选用 waveform + noise 数据集作为实验数据,分别运行 CAU 和未进行属性约简的 CAU(记作 UCAU),基分类器采用 CART. 最终结果如图 5 和图 6 所示.

由图 5 可知,在有冗余属性的数据集中,通过属性约简过程,能提高聚类结果的准确率,从而使得概念漂移的检测更准确,最终提高算法的分类性能. 在 UCAU 参数同 CAU 一致时,UCAU 由于缺少约简机制,会降低聚类结果的准确性,产生大量概念漂移的误报,替换了一些性能较好的分类器. 由于新加入的

分类器未得到足够训练,所以最终分类结果的准确率低于 CAU.

结合图 5、图 6 可得,CAU 中的属性约简过程能取得一定效果. 在 waveform + noise 数据集中共有 19 个无关属性,在 CAU 的约简过程中能基本消除 14 个属性. 由图 5 可知,很少有数据块能被消除 19 个属性,基本都有一定偏差,这主要是与数据集中噪声有关. 在 waveform + noise 数据集中存在一定数量的噪声数据,因此当数据块中噪声达到一定含量时,约简过程会受到影响,约简结果会出现偏差.

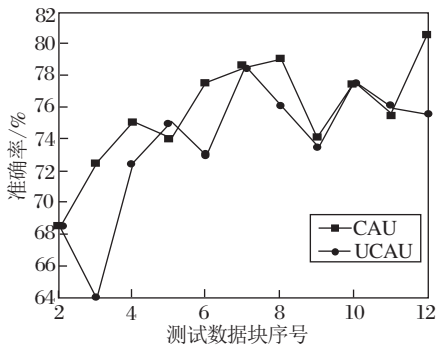

图 5　2 种算法在 waveform+noise 数据集上的准确率

Fig. 5　Accuracies of 2 algorithms on waveform+noise dataset

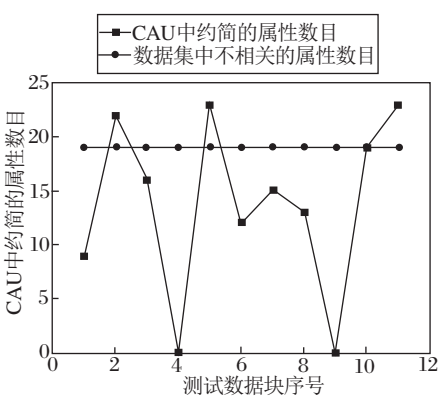

图 6　CAU 约简的属性数目

Fig. 6　Attribute number of CAU reduction

4　结 束 语

针对含有概念漂移的数据流分类问题,本文提出结合无监督学习的数据流分类算法,通过分类与聚类方法的结合,能有效检测概念漂移,在数据流分类上取得较好效果. 实验表明,本文算法对概念较多且频繁变化的数据具有较好的分类能力. 此外,CAU 也有一定的局限性,如聚类算法的选取与数据集相关,这可能需要依赖先验知识,若聚类算法选择不正确,系统会产生概念漂移的误报,最终会降低分类算法的性能. 如何降低聚类算法对 CAU 的影响,将是下一步的研究重点.

参　考　文　献

[1] WOŹNIAK M, KASPRZAK A, CAL P. Weighted Aging Classifier Ensemble for the Incremental Drifted Data Streams // Proc of the 10th International Conference on Flexible Query Answering Systems. Berlin, Germany: Springer-Verlag, 2013: 579-588.

[2] YANG Y, WU X D, ZHU X Q. Combining Proactive and Reactive Predictions for Data Streams // Proc of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. New York, USA: ACM, 2005: 710-715.

[3] SHAO J M, AHMADI Z, KRAMER S. Prototype-Based Learning on Concept-Drifting Data Streams // Proc of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2014: 412-421.

[4] ŽLIOBAITĖ I, BIFET A, READ J, *et al*. Evaluation Methods and Decision Theory for Classification of Streaming Data with Temporal Dependence. Machine Learning, 2015, 98(3): 455-482.

[5] WANG H X, YIN J, PEI J, *et al*. Suppressing Model Overfitting in Mining Concept-Drifting Data Streams // Proc of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2006: 736-741.

[6] STREET W N, KIM Y S. A Streaming Ensemble Algorithm (SEA) for Large-Scale Classification // Proc of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2001: 377-382.

[7] DOMINGOS P, HULTEN G. Mining High-Speed Data Streams // Proc of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2000: 71-80.

[8] HULTEN G, SPENCER L, DOMINGOS P. Mining Time-Changing Data Streams // Proc of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2001: 97-106.

[9] WU X D, LI P P, HU X G. Learning from Concept Drifting Data Streams with Unlabeled Data. Neurocomputing, 2012, 92: 145-155.

[10] ELWELL R, POLIKAR R. Incremental Learning of Concept Drift in Nonstationary Environments. IEEE Trans on Neural Networks, 2011, 22(10): 1517-1531.

[11] LIAO J W, DAI B R. An Ensemble Learning Approach for Concept Drift // Proc of the International Conference on Information Science and Applications. Seoul, Republic of Korea: IEEE, 2014. DOI:10.1109/ICISA2014.6847357.

[12] BRZEZINSKI D, STEFANOWSKI J. Prequential AUC for Classifier Evaluation and Drift Detection in Evolving Data Streams // Proc of the 3rd International Workshop on New Frontiers in Mining Complex Patterns. Zurich, Switzerland: Springer International Publishing, 2015: 87-101.

[13] BRZEZINSKI D, STEFANOWSKI J. Reacting to Different Types of Concept Drift: The Accuracy Updated Ensemble Algorithm. IEEE Trans on Neural Networks and Learning Systems, 2014, 25(1): 81-94.

[14] WANG H X, FAN W, YU P S, *et al*. Mining Concept-Drifting Data Streams Using Ensemble Classifiers // Proc of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2003: 226-235.

[15] 李 燕,张玉红,胡学钢. 基于 C4.5 和 NB 混合模型的数据流分类算法. 计算机科学, 2010, 37(12): 138-142.
(LI Y, ZHANG Y H, HU X G. Classification Algorithm for Data Stream Based on Mixture Models of C4.5 and NB. Computer Science, 2010, 37(12): 138-142.)

[16] 孙 岳,毛国君,刘 旭,等. 基于多分类器的数据流中的概念漂移挖掘. 自动化学报, 2008, 34(1): 93-97.
(SUN Y, MAO G J, LIU X, *et al*. Mining Concept Drifts from Data Streams Based on Multi-classifiers. Acta Automatica Sinica, 2008, 34(1): 93-97.)

[17] BIFET A, HOLMES G, KIRIKBY R, *et al*. MOA: Massive Online Analysis. Journal of Machine Learning Research, 2010, 11: 1601-1604.

作者简介

徐树良,男,1989 年生,硕士研究生,主要研究方向为数据流分类、概念漂移检测、极限学习机. E-mail: xushulianghao@126.com.

(**XU Shuliang**, born in 1989, master student. His research interests include data streams classification, concept drift detection and extreme learning machine.)

王俊红(通讯作者),女,1979 年生,博士,副教授,主要研究方向为数据挖掘、机器学习. E-mail: wjhwjh@sxu.edu.cn.

(**WANG Junhong**(Corresponding author), born in 1979, Ph.D., associate professor. Her research interests include data mining and machine learning.)