

基于邻域决策误差率的层次分类在线流特征选择

王晨曦^{1,2}, 刘园奎^{1,2}, 吕彦^{1,2}, 林耀进^{1,2}

(1. 闽南师范大学计算机学院, 福建 漳州 363000)

(2. 闽南师范大学数据科学与智能应用福建省高等学校重点实验室, 福建 漳州 363000)

[摘要] 在实际应用领域中, 存在许多特征空间无法预先给定的场景, 数据以特征流的形式随时间动态流入特征空间, 而样本数量是固定不变的. 同时, 数据的类别中往往存在丰富的层次化结构关系, 传统的特征选择算法在性能上已无法满足需求. 基于此, 本文提出一种面向层次分类学习的在线流特征选择算法. 首先, 利用兄弟节点之间的关系设计了一种基于最大近邻的决策误差率计算公式. 其次, 设计在线重要性选择和在线冗余更新两种在线评估准则, 用于选择决策误差最小的特征子集. 最后, 在 6 个层次数据集上的实验结果表明, 所提算法优于一些现有的在线流特征选择算法.

[关键词] 在线流特征选择, 层次分类, 兄弟关系, 邻域决策误差率

[中图分类号] TP18 **[文献标志码]** A **[文章编号]** 1672-1292(2022)04-0009-10

Online Hierarchical Streaming Feature Selection Based on Neighborhood Decision Error Rate

Wang Chenxi^{1,2}, Liu Yuankui^{1,2}, Lü Yan^{1,2}, Lin Yaojin^{1,2}

(1. School of Computer Science, Minnan Normal University, Zhangzhou 363000, China)

(2. Key Laboratory of Data Science and Intelligence Application, Minnan Normal University, Zhangzhou 363000, China)

Abstract: In many practical application fields, there are numerous scenes in which the entire feature space cannot be available in advance, candidate features flow into the feature space dynamically over time, and the number of samples is fixed. At the same time, there exists a hierarchical structure relationship between classes, and traditional feature selection methods cannot be able to meet the demand. Based on these, an online streaming feature selection algorithm for hierarchical classification learning is presented. Firstly, a decision error rate calculation formula is designed on the basis of the largest nearest neighbor according to sibling relationships. Secondly, two online evaluation criteria of online significance selection and online relevance analysis are proposed to select features with minimum decision error. Finally, experimental results on six hierarchical datasets manifest that the proposed algorithm is better than some existing online streaming feature selection algorithms.

Key words: online streaming feature selection, hierarchical classification, sibling relationships, neighborhood decision error rate

在许多实际领域中, 数据的类别标记间往往存在丰富的层次化结构关系^[1]. 对层次化结构分类学习的研究具有重要的应用价值.

在层次化结构数据的分类建模过程中, 数据通常具有超高维和演化性的特点. 为了解决该类数据的存储和分类问题, 学者们将支持向量机、贝叶斯模型、多核学习、集成学习等机器学习技术应用到层次分类中, 取得了显著成果^[2]. 一些考虑标记间层次化结构的分类算法被先后提出. Freeman 等^[3]利用遗传算法, 提出联合特征选择和层次分类器的方法, 用于提高分类器精度. Song 等^[4]提出一种基于类别区分和特征位置信息的文本特征选择方法. 为了最大程度减少误分类成本, Pan 等^[5]利用贝叶斯决策规则设计了一种

收稿日期: 2022-08-08.

基金项目: 国家自然科学基金项目(62076116)、福建省自然科学基金项目(2020J01811、2020J01792 和 2021J02049).

通讯作者: 林耀进, 博士, 教授, 研究方向: 数据挖掘、粒计算. E-mail: zlzinyaojin@163.com

成本敏感的快速图分类算法. Zhao 等^[6]考虑类结构中父子之间的层次关系,提出一种基于递归正则化的层次分类特征选择框架.

上述关于考虑层次化结构的分类方法假设特征空间是静态的和已知的,并未考虑特征空间的未知性和演化性. 在许多实际应用领域中,具有类别层次化结构的数据特征并不能一次性全部获取,而是逐个流入特征空间. 例如,可以从不同尺度的高分辨率火星图片中获取数以万计的纹理特征,若等到获取整个火星表面的信息后再进行特征选择处理,这必然是不现实的. 对此,学者们提出了在线流特征选择方法. 目前,流特征选择主要可划分为在线单特征流特征选择和在线流组特征选择. 针对单特征流数据,Zhou 等^[7]提出基于流回归的在线流特征选择算法 alpha-investing. Yu 等^[8]提出一种新颖的两两对比的在线流特征选择算法(scalable and accurate online feature selection approach, SAOLA). Lin 等^[9]采用模糊互信息来评估多标签学习中特征的质量,并提出一种多标记流特征选择算法(multi-label streaming feature selection, MSFS)用于处理特征空间完全已知或部分已知时的场景. 针对于具有组结构的特征流数据,Liu 等^[10]提出基于特征交互的多标记在线流组特征选择算法(online multi-label group feature selection, OMCFS),该算法分为在线组选择和在线组间选择两个阶段. Li 等^[11]提出流特征组特征选择算法(group feature selection with streaming features, GFSSF),该算法是一种针对流环境下的组特征选择算法,可同时处理单个特征和流组特征情况.

现有的流特征选择算法假设标记之间是相互独立的,并未考虑标记之间存在的层次化结构关系. 邻域决策误差最小化准则^[12]在处理特征选择时具有较强的鲁棒性,但不能处理具有层次化结构关系的数据. 本文利用层次化数据关系的兄弟策略改进邻域决策误差最小化准则以适应流特征环境下的层次化结构数据,并设计在线重要性分析及在线冗余更新准则,提出一种基于邻域决策误差率的层次分类在线流特征选择算法(online hierarchical streaming feature selection based on neighborhood decision error rate, OHS-NDER). OHS-NDER 通过在线分析使得特征与标记之间的决策误差最小化,从而选择决策误差最小的特征子集. 在 6 个标记具有层次化结构的数据上的实验表明,相比于现有的在线流特征选择算法, OHS-NDER 算法性能更优.

1 背景知识

1.1 类别的层次结构

层次结构可分为树和有向无环图两种结构,均具有“从属关系”的特点. 本文从树结构角度将“从属关系”归纳为反自反性、不可逆性和传递性. 给定一个层次结构的序对 $(D, <)$, 其中 D 为类别标签集合, $<$ 为从属关系,则反自反性、不可逆性和传递性可形式化描述为:

- (1)反自反性: $\forall d_i \in D$, 有 $d_i \not< d_i$;
- (2)不可逆性:若 $d_i < d_j$, $\forall d_i, d_j \in D$, 则 $d_j \not< d_i$;
- (3)传递性:若 $d_i < d_k$ 且 $d_k < d_j$, 对 $\forall d_i, d_j, d_k \in D$,

则 $d_i < d_j$.

一般来说,一个层次结构包含 5 种节点之间的常用关系,如表 1 所示.

1.2 邻域决策误差率

为了估计不同特征子空间中的分类复杂度,使用邻域决策误差率,其既适用于数值特征,也适用于混合特征. 与 KNN 分类器一样,邻域决策误差率能较好地推断出子空间的分类能力.

定义 1^[12] 给定邻域决策系统 $NDT = \langle U, C, D \rangle$, $x_i \in U$, $\delta(x_i)$ 是样本 x_i 在邻域近似空间的邻域, $P(\omega_j | \delta(x_i))$, $j = 1, 2, \dots, c$, 是该邻域内 ω_j 类的概率,则定义 x_i 的邻域决策函数为

$$ND(x_i) = \omega_i, P(\omega_i | \delta(x_i)) = \max_j P(\omega_j | \delta(x_i)), \quad (1)$$

式中, $P(\omega_j | \delta(x_i)) = n_j / N$, N 是邻域内样本的数量, n_j 是第 j 类样本的数量.

引入 0-1 误分类损失函数:

$$\begin{aligned} \lambda(\omega(x_i) | ND(x_i)) &= 0, \text{ s.t. } \omega(x_i) = ND(x_i), \\ \lambda(\omega(x_i) | ND(x_i)) &= 1, \text{ s.t. } \omega(x_i) \neq ND(x_i), \end{aligned}$$

表 1 层次结构中的符号含义

Table 1 Symbol meaning in the hierarchy

符号表示	表达意义
P_i	类别 i 的父节点
C_i	类别 i 的孩子节点
$Anc(d_i)$	类别 d_i 的祖先节点集合
$Des(d_i)$	类别 d_i 的对应的叶子节点集合
$Sib(d_i)$	类别 d_i 的兄弟节点集合

式中, $\omega(x_i)$ 为 x_i 的真实类别; $ND(x_i)$ 为 x_i 的重新分配类别. 若 x_i 是属于决策近似较低的样本, $\omega(x_i) = ND(x_i)$; 否则, 应计算 x_i 邻域类别的概率来给 x_i 分配一个类标签. 若 x_i 附近的多数样本类别都不同, 则 $\omega(x_i) \neq ND(x_i)$.

定义 2^[12] 邻域决策误差率 β 定义为:

$$\beta = \frac{1}{n} \sum_{i=1}^n \lambda(\omega(x_i) | ND(x_i)), \quad (2)$$

式中, n 为样本的总数量.

邻域决策误差率的本质就是根据邻域内样本的分布, 按照多数决策原则重新给每个样本分配类别, 进而统计实际类别与重新分配的类别的差异率.

性质 1^[12] 给定邻域决策系统 $NDT = \langle U, C, D \rangle$, Δ 是 U 上的度量, $\delta \geq 0$, 则以下表达式成立:

(1) $\mu_A \leq 1 - \beta$;

(2) 若邻域决策系统在邻域近似空间中是一致的, 则邻域识别率 $\mu_A = 1 - \beta$, 且 $\mu_A = 1, \beta = 0$.

定义 3^[12] 给定邻域决策系统 $NDT = \langle U, C, D \rangle$, Δ 是 U 上的度量, $\delta \geq 0$. $B \subset C, a \notin B$, 定义给属性 B , a 相对于 D 的重要度为:

$$Hsig(a, B, D) = \mu_{B \cup a}(D) - \mu_B(D). \quad (3)$$

2 基于邻域决策误差率的层次分类在线流特征选择算法

2.1 层次分类的兄弟策略

在传统特征选择算法中, 假设标记之间是相互独立的, 直接使用排斥策略^[13], 即同类表示为 x , 异类则为非 x . 在层次分类学习中, 可利用类别间的相关性来区分样本的同类与异类, 即存在利用父子关系区分样本的包含策略^[13]和利用兄弟关系区分样本的兄弟策略^[14], 如表 2 所示.

表 2 区分同类和异类样本的 3 种策略
Table 2 Three strategies of positive and negative samples' definitions

搜索策略	同类	异类
排斥策略	x	非 x
包含策略	$x + Des(x)$	非 $[x + Des(x)]$
兄弟策略	x	Sib(x)

2.2 面向层次结构化数据的邻域决策误差率模型

从人类认知与决策的角度出发, 包含策略与排斥策略可分别表示乐观与悲观的决策关系, 这两种策略均易产生极端结果. 因此, 本文采用代表中立决策的兄弟策略来构建面向层次结构化数据的邻域决策误差率模型.

定义 4 给定一个层次邻域决策系统 $HDST = \langle U, C, D, I \rangle$, 其中类别标记 D 存在层次关系 $I, \forall x \in U, x$ 基于兄弟策略的最大近邻点集为:

$$\delta^{sib}(x) = \{y | \Delta(x, y) \leq d^{sib}(x), y \in U\}. \quad (4)$$

式中,

$$d^{sib}(x) = \max(d_1^{sib}(x), d_2^{sib}(x)),$$

$$d_1^{sib}(x) = \Delta(x, NM^{sib}(x)),$$

$$d_2^{sib}(x) = \Delta(x, NH^{sib}(x)).$$

式中, $NH^{sib}(x)$ 表示利用兄弟策略得到的样本 x 的最近同类样本; $NM^{sib}(x)$ 表示利用兄弟策略得到的样本 x 的最近异类样本; $d_1^{sib}(x)$ 与 $d_2^{sib}(x)$ 分别表示样本 x 在兄弟策略下到最近异类样本 $NM^{sib}(x)$ 和最近同类样本 $NH^{sib}(x)$ 的距离.

定义 5 给定层次邻域决策系统 $HDST = \langle U, C, D, I \rangle$, 决策属性 D 将论域 U 划分为 N 个等价类 $\{X_1, X_2, \dots, X_N\}$, I 表示标记的层次结构. $P(\omega_j | \delta^{sib}(x_i)), j = 1, 2, \dots, c$, 是邻域 $\delta^{sib}(x_i)$ 内 ω_j 类的概率, 则定义 x_i 的邻域决策函数为

$$ND^{sib}(x_i) = \omega_i, P(\omega_i | \delta^{sib}(x_i)) = \max_j P(\omega_j | \delta^{sib}(x_i)), \quad (5)$$

式中, $P(\omega_j | \delta^{sib}(x_i)) = n_j / N$, N 是邻域内样本的数量, n_j 是第 j 类样本的数量.

引入 0-1 误分类损失函数:

$$\lambda(\omega(x_i) | ND^{\text{sib}}(x_i)) = 0, \text{ s.t. } \omega(x_i) = ND^{\text{sib}}(x_i),$$

$$\lambda(\omega(x_i) | ND^{\text{sib}}(x_i)) = 1, \text{ s.t. } \omega(x_i) \neq ND^{\text{sib}}(x_i),$$

式中, $\omega(x_i)$ 为 x_i 的真实类别.

定义 6 邻域决策误差率 β^{sib} 定义为:

$$\beta^{\text{sib}} = \frac{1}{n} \sum_{i=1}^n \lambda(\omega(x_i) | ND^{\text{sib}}(x_i)). \quad (6)$$

式中, n 为样本的总数量.

性质 2 给定邻域决策系统 $HDST = \langle U, C, D, I \rangle$, Δ 是 U 上的度量, $\delta \geq 0$, 则以下表达式成立:

$$(1) \mu_A^{\text{sib}} \leq 1 - \beta^{\text{sib}};$$

$$(2) \text{ 若层次邻域决策系统在邻域近似空间中是一致的, 则邻域识别率 } \mu_A^{\text{sib}} = 1 - \beta^{\text{sib}}, \text{ 且 } \mu_A^{\text{sib}} = 1, \beta^{\text{sib}} = 0.$$

定义 7 给定层次邻域决策系统 $HDST = \langle U, C, D, I \rangle$, Δ 是 U 上的度量, $\delta \geq 0$. $B \subset C, a \notin B$, 定义给属性 B, a 相对于 D 的兄弟策略重要度为:

$$\text{Hsig}^{\text{sib}}(a, B, D) = \mu_{B \cup a}^{\text{sib}}(D) - \mu_B^{\text{sib}}(D). \quad (7)$$

2.3 基于邻域决策误差率的层次分类在线流特征算法模型

为了从层次结构化数据中选取分类性能较强的特征子集, 并确保选择的特征子集具有最小的决策误差, 本文提出基于面向层次结构化数据的邻域决策误差率模型的两在线特征评估准则: 在线重要性选择和在线冗余更新准则.

2.3.1 在线重要性选择

定义 8 给定在线流特征的层次邻域决策系统 $HDST = \langle U, C, D, I, t \rangle$, 其中非空有限样本集合 $U = \{x_1, x_2, \dots, x_n\}$, $C = \{f_1, f_2, \dots, f_t\}$ 为描述论域 U 的实数型特征集合, f_t 为 t 时刻流进特征空间的特征, D 为决策属性, I 表示类别标记 D 存在的层次关系. f_t 相对于类别标记 D 的兄弟策略重要度为:

$$\text{Hsig}^{\text{sib}}(f_t, S_{t-1}, D) = \gamma_{S_{t-1} \cup f_t}^{\text{sib}}(D) - \gamma_{S_{t-1}}^{\text{sib}}(D), \quad (8)$$

式中, $S_{t-1} \subseteq C$ 表示 $t-1$ 时刻所选的特征集合. 由于决策误差率 $\gamma_{S_{t-1} \cup f_t}^{\text{sib}}(D) \in [0, 1]$, 且 $\gamma_{S_{t-1}}^{\text{sib}}(D) \geq \gamma_{S_{t-1} \cup f_t}^{\text{sib}}(D)$, 因此特征 f_t 的重要度 $\text{Hsig}^{\text{sib}}(f_t, S_{t-1}, D) \in [0, 1]$. 若 $\text{Hsig}^{\text{sib}}(f_t, S_{t-1}, D) > 0$, 则认为特征 f_t 是一个重要的特征. 否则, 当前所选特征集合 S_{t-1} 中存在与 f_t 相互冗余的特征或特征 f_t 是一个无意义的特征.

2.3.2 在线冗余更新

由定义 8 可知, 若 t 时刻特征 f_t 的重要度 $\text{Hsig}^{\text{sib}}(f_t, S_{t-1}, D) = 0$, 则认为已选特征集合中存在与 f_t 互为冗余的特征. 因此, 为了选择最具有最优判别性能的特征子集而建立在线冗余分析.

定义 9 给定在线流特征的层次邻域决策系统 $HDST = \langle U, C, D, I, t \rangle$, $S_{t-1} \subseteq C$ 表示 $t-1$ 时刻所选的特征集合, 已选特征 $f' \in S_{t-1}$, 若满足式(9), 则放弃 f_t ; 若满足式(10), 则认为特征 f_t 比特征 f' 更重要, 并将 f_t 加入已选集合 S_{t-1} 中:

$$\beta_{f_t \cup f'}^{\text{sib}}(D) - \beta_{f'}^{\text{sib}}(D) \leq 0 \text{ and } \beta_{f_t}^{\text{sib}}(D) \leq \beta_{f'}^{\text{sib}}(D), \quad (9)$$

$$\beta_{f_t \cup f'}^{\text{sib}}(D) - \beta_{f'}^{\text{sib}}(D) \leq 0 \text{ and } \beta_{f_t}^{\text{sib}}(D) > \beta_{f'}^{\text{sib}}(D). \quad (10)$$

2.4 基于邻域决策误差率的层次分类在线流特征选择算法

根据定义 8 与定义 9, 本文提出一种基于邻域决策误差率的层次分类在线流特征选择算法, 算法步骤如下所示:

算法 1 基于邻域决策误差率的层次分类在线流特征选择算法 (OHS-NDER)

输入: 在线流特征的层次邻域决策系统 $\langle U, C, D, I, t \rangle$;

t 时刻到达的特征 f_t ;

$t-1$ 时刻已选的特征子集 S_{t-1} ;

输出: t 时刻所选的特征子集 S_t

(1) REPEAT

(2) 在 t 时刻流进新特征 f_t ;

```

(3) 通过公式(6)计算  $\beta_{f_i}^{\text{sig}}(D)$ ;
    /* 在线重要性选择 */
(4) IF  $\beta_{f_i}^{\text{sig}}(D) < \beta_{S_{t-1}}^{\text{sig}}(D)$ 
(5)    $S_t = f_i$ ;
(6)   通过公式(8)计算  $\text{Hsig}^{\text{sig}}(f_i, S_{t-1}, D)$ ;
(7)   IF  $\text{Hsig}^{\text{sig}}(f_i, S_{t-1}, D) > 0$ 
(8)    $S_t = S_{t-1} \cup f_i$ ;
(9) /* 在线冗余更新 */
(10) ELSE
(11)   FOR  $S_t$  中的每一个元素  $f_i$ 
(12)   IF  $\beta_{f_i \cup f_i}^{\text{sig}}(D) - \beta_{f_i}^{\text{sig}}(D) \leq 0$  &  $\beta_{f_i}^{\text{sig}}(D) \leq \beta_{S_{t-1}}^{\text{sig}}(D)$ 
(13)     Discard  $f_i$ , and go to Step 20;
(14)   ELSE IF  $\beta_{f_i \cup f_i}^{\text{sig}}(D) - \beta_{f_i}^{\text{sig}}(D) \leq 0$  &  $\beta_{f_i}^{\text{sig}}(D) > \beta_{S_{t-1}}^{\text{sig}}(D)$ 
(15)      $S_{t-1} = S_{t-1} - f_i$  and  $S_t = S_{t-1} \cup f_i$ ;
(16)   END IF
(17)   END FOR
(18) END IF
(19) END IF
(20) 直到没有新特征到达, 返回  $S_t$ .

```

算法 1 主要包含两个关键步骤:在线重要性选择和在线冗余更新. 在线重要性选择步骤中, 当新特征 f_i 到达时, 执行第(3)步计算特征邻域决策误差率 $\beta_{f_i}^{\text{sig}}(D)$. 若 $\beta_{f_i}^{\text{sig}}(D) < \beta_{S_{t-1}}^{\text{sig}}(D)$, 则使用 f_i 取代 S_{t-1} 中的特征. 若不满足要求, 根据公式(8)计算 $\text{Hsig}^{\text{sig}}(f_i, S_{t-1}, D)$, 若 $\text{Hsig}^{\text{sig}}(f_i, S_{t-1}, D) > 0$, 则将特征 f_i 加入特征子集 S_{t-1} 中, 否则执行第 12 至 18 步. 在线冗余更新步骤中, OHS-NDER 将评估 f_i 是否应该添加进在当前已选定的特征集合 S_{t-1} 中, 若 $\exists f_i \in S_{t-1}$ 使得公式(9)成立, 则放弃 f_i , 不再考虑, 否则将使用公式(10)检查 S_{t-1} 中是否包含冗余特征, 若 $\beta_{f_i \cup f_i}^{\text{sig}}(D) - \beta_{f_i}^{\text{sig}}(D) \leq 0$ 同时 $\beta_{f_i}^{\text{sig}}(D) > \beta_{S_{t-1}}^{\text{sig}}(D)$, 则将 f_i 从已选集合 S_{t-1} 中删除, 并将 f_i 加入 S_{t-1} 中.

假设论域 U 的样本个数为 $|U|$, 条件属性 C 的属性个数为 $|C|$, 当前所选特征子集的数量为 $|S_t|$, 计算邻域决策误差率 β^{sig} 的时间复杂度为 $O(|C| \cdot \lg |U|)$. 在线重要性选择阶段, 遍历所有样本所需要的时间复杂度为 $O(|S_t| \cdot |U|^2 \cdot \lg |U|)$. 执行在线冗余更新阶段时, OHS-NDER 的最差时间复杂度为 $O(|C| \cdot |S_t| \cdot |U|^2 \cdot \lg |U|)$.

3 实验与结果分析

3.1 实验数据及环境设置

本实验选取 6 个层次结构的数据集, 用于验证 OHS-NDER 在层次数据集上进行特征选择的性能. 所选数据集包含 2 个蛋白质数据集和 4 个图像数据集. 数据集详细描述信息如表 3 所示.

表 3 数据集的描述
Table 3 Descriptions of data sets

数据集	样本数	特征数	内部节点	叶子节点	树高
AWA	317	252	17	10	4
Bridges	108	12	8	6	3
CLEF	390	80	25	63	4
DD	291	473	5	27	3
F194	254	473	202	194	3
VOC	354	1 000	30	88	5

本实验采用 KNN ($K=1$) 和线性支持向量机 (LSVM) 两个基分类器对已选择的特征子集进行分类精度的评估, 并采用 10 折交叉验证. 实验平台统一采用 Matlab R2016a, 且所有的实验都在同一台 Inter i5, 2.79 GHz, 4 GB 内存的计算机上运行.

3.2 评价指标

分层分类方法不同于扁平分类方法, 其评价方法也有所不同. 本文采用最近共同祖先 (lowest common ancestor, LCA)、基于集合的评价指标 (Hierarchical-F1, H-F1) 和树诱导误差 (tree induced error, TIE) 3 种基于层次结构的分层分类评价指标.

(1) 最近共同祖先 (LCA): 最近共同祖先的评价方法有如下定义: 假设 d_u 与 d_v 为树结构 D 中的两个类别节点, 则使用 $LCA(d_u, d_v)$ 为距离 d_u 与 d_v 节点最近的共同祖先.

(2) 基于集合的评价指标 (H-F1): 常用的增广方法是利用真实类别、预测类别及其祖先类别作为增广集. 基于集合的度量包括利用层次信息来求真实类别 d 及预测类别 \hat{d} 的增广集和基于增广集合来计算错误的惩罚.

$$d_{\text{aug}} = d \cup \text{Anc}(d),$$

$$\hat{d}_{\text{aug}} = \hat{d} \cup \text{Anc}(\hat{d}).$$

基于增广集合的分层精度和召回率定义如下:

$$P_H = \frac{|\hat{d}_{\text{aug}} \cap d_{\text{aug}}|}{|\hat{d}_{\text{aug}}|},$$

$$R_H = \frac{|\hat{d}_{\text{aug}} \cap d_{\text{aug}}|}{|d_{\text{aug}}|},$$

式中, $|\cdot|$ 为元素的数目. 基于增广集合的分层评价定义如下:

$$F_H = \frac{2 \cdot P_H \cdot R_H}{P_H + R_H}.$$

(3) 树诱导误差 (TIE): 在树结构的类别中, 通过预测类别节点与真实类别节点间需要走的步骤数来定义惩罚. 假设真实类别标签及预测的类别标签分别为 Y 和 \bar{Y} , 则

$$\text{TIE}(Y, \bar{Y}) = |E_H(Y, \bar{Y})|,$$

式中, $|E_H(Y, \bar{Y})|$ 是从树结构中 Y 类别节点与 \bar{Y} 类别节点之间的最小总边数. 若 $\text{TIE}(Y, \bar{Y}) = 0$, 根据对称关系, 有 $\text{TIE}(Y, \bar{Y}) = \text{TIE}(\bar{Y}, Y)$, 即预测类别与真实类别为同一类别.

3 个性能评价指标中, LCA 和 H-F1 的取值越大越好, TIE 的取值越小越好.

3.3 与在线特征选择算法的对比

在流特征环境下, 为了验证 OHS-NDER 算法在层次结构数据集上进行分层流特征选择的有效性, 同时对比考虑数据集类别层次结构关系和忽略类别层次结构关系时的性能差异, 本文采用 6 种不考虑层次结构关系的在线流特征选择算法作为对比算法进行实验, 包括 OSFS^[15]、FOSFS^[15]、OFS-D^[16]、K-OFS^[17]、A3M^[18] 及 SAOLA^[8]. 根据对应文献, 分别将对比算法所涉及到的参数按照所描述的最优参数进行设置, 其中 OSFS、FOSFS 和 SAOLA 中的显著水平参数 α 设置为 0.01, K-OFS 中的近邻个数 K 设置为 7.

3.4 实验结果与分析

为了验证 OHS-NDER 算法的有效性, 首先比较各种算法所选的特征子集在传统分类评价指标和 3 种层次分类评价指标上的效果, 然后进一步使用蜘蛛图直观分析各种算法在以上 4 个分类评价指标上的性能.

表 4-表 11 分别描述了 7 种不同在线流特征选择算法在 LCA、H-F1、TIE 和传统预测精度评价指标上的 KNN 和 LSVM 分类结果. 表 4-表 11 中“ \uparrow ”符号表示取值越大越好, 反之, “ \downarrow ”符号表示取值越小越好. 表中最后一行为算法的平均精度.

表 4 基于 KNN 分类器的分类精度

Table 4 Predictive accuracy using the KNN classifier

数据集	A3M	OSFS	FOSFS	SAOLA	K-OFSD	OFS-D	OHS-NDER
AWA	0.164 1	0.127 6	0.146 7	0.130 8	0.149 7	0.164 8	0.171 5
Bridges	0.548 9	0.630 0	0.417 8	0.630 0	0.477 8	0.507 8	0.595 6
CLEF	0.453 3	0.515 9	0.562 1	0.528 5	0.549 6	0.479 2	0.551 8
DD	0.378 3	0.279 0	0.278 8	0.333 3	0.389 7	0.413 4	0.429 2
F194	0.311 9	0.342 6	0.260 6	0.289 8	0.391 3	0.463 0	0.447 0
VOC	0.221 8	0.185 0	0.210 7	0.212 2	0.195 0	0.167 6	0.234 7
Avg	0.346 4	0.346 7	0.312 8	0.354 1	0.358 9	0.366 0	0.405 0

表 5 基于 KNN 分类器的 LCA 值(↑)

Table 5 LCA score using the KNN classifier(↑)

数据集	A3M	OSFS	FOSFS	SAOLA	K-OFSD	OFS-D	OHS-NDER
AWA	0.428 5	0.412 2	0.425 6	0.411 4	0.411 9	0.441 9	0.435 9
Bridges	0.740 7	0.785 2	0.674 1	0.785 2	0.697 2	0.715 7	0.775 0
CLEF	0.586 1	0.625 7	0.662 9	0.641 0	0.654 7	0.609 3	0.659 9
DD	0.645 5	0.565 9	0.564 7	0.596 8	0.635 2	0.643 2	0.662 7
F194	0.467 2	0.479 0	0.469 8	0.503 9	0.528 9	0.533 5	0.538 7
VOC	0.494 2	0.457 2	0.478 7	0.478 0	0.475 6	0.444 2	0.495 8
Avg	0.560 4	0.554 2	0.546 0	0.569 4	0.567 3	0.564 6	0.594 7

表 6 基于 KNN 分类器的 H-F1 值(↑)

Table 6 H-F1 score using the KNN classifier(↑)

数据集	A3M	OSFS	FOSFS	SAOLA	K-OFSD	OFS-D	OHS-NDER
AWA	0.500 8	0.497 6	0.509 5	0.494 5	0.477 9	0.528 4	0.511 8
Bridges	0.768 5	0.806 8	0.720 4	0.806 8	0.721 9	0.749 7	0.798 1
CLEF	0.602 8	0.640 4	0.676 6	0.651 4	0.677 8	0.629 7	0.681 4
DD	0.719 4	0.622 0	0.619 7	0.651 8	0.689 6	0.694 2	0.721 6
F194	0.506 6	0.519 7	0.522 3	0.574 8	0.587 9	0.584 0	0.615 5
VOC	0.526 8	0.481 7	0.502 8	0.494 7	0.509 2	0.464 7	0.522 2
Avg	0.604 2	0.594 7	0.591 9	0.612 3	0.610 7	0.608 5	0.641 8

表 7 基于 KNN 分类器的 TIE 值(↓)

Table 7 TIE score using the KNN classifier(↓)

数据集	A3M	OSFS	FOSFS	SAOLA	K-OFSD	OFS-D	OHS-NDER
AWA	3.993 7	4.018 9	3.924 3	4.044 2	4.176 7	3.772 9	3.905 4
Bridges	1.203 7	1.009 3	1.481 5	1.009 3	1.416 7	1.314 8	1.027 8
CLEF	2.892 3	2.617 9	2.325 6	2.507 7	2.366 7	2.710 3	2.325 6
DD	1.683 8	2.268 0	2.281 8	2.089 3	1.862 5	1.835 1	1.670 1
F194	2.960 6	2.881 9	2.866 1	2.551 2	2.472 4	2.496 1	2.307 1
VOC	3.033 9	3.409 6	3.214 7	3.257 1	3.104 5	3.488 7	3.098 9
Avg	2.628 0	2.700 9	2.682 3	2.576 5	2.566 6	2.603 0	2.389 2

表 8 基于 LSVM 分类器的分类精度

Table 8 Predictive accuracy using the LSVM classifier

数据集	A3M	OSFS	FOSFS	SAOLA	K-OFSD	OFS-D	OHS-NDER
AWA	0.119 8	0.145 4	0.164 2	0.162 2	0.151 9	0.145 4	0.171 5
Bridges	0.625 6	0.630 0	0.630 0	0.630 0	0.487 8	0.564 4	0.632 2
CLEF	0.468 9	0.517 9	0.547 0	0.529 5	0.528 7	0.509 7	0.550 3
DD	0.391 3	0.250 5	0.264 6	0.303 0	0.303 1	0.328 4	0.422 5
F194	0.178 4	0.118 6	0.261 0	0.336 8	0.390 4	0.450 8	0.414 7
VOC	0.236 7	0.241 4	0.245 9	0.238 1	0.266 3	0.234 7	0.257 6
Avg	0.336 8	0.317 3	0.352 1	0.366 6	0.354 7	0.372 2	0.408 1

表 9 基于 LSVM 分类器的 LCA 值(↑)
Table 9 LCA score using the LSVM classifier(↑)

数据集	A3M	OSFS	FOSFS	SAOLA	K-OFSD	OFS-D	OHS-NDER
AWA	0.418 0	0.435 6	0.449 5	0.447 2	0.445 3	0.430 1	0.456 9
Bridges	0.786 1	0.785 2	0.785 2	0.785 2	0.708 3	0.752 8	0.797 2
CLEF	0.600 8	0.632 7	0.653 4	0.652 3	0.653 7	0.630 0	0.657 5
DD	0.645 5	0.545 2	0.557 3	0.569 3	0.577 9	0.585 3	0.659 2
F194	0.421 3	0.382 5	0.464 6	0.497 4	0.526 2	0.534 8	0.544 6
VOC	0.505 3	0.506 3	0.509 4	0.504 3	0.523 2	0.502 0	0.517 6
Avg	0.562 8	0.547 9	0.569 9	0.576 0	0.572 4	0.572 5	0.605 5

表 10 基于 LSVM 分类器的 H-F1 值(↑)
Table 10 H-F1 score using the LSVM classifier(↑)

数据集	A3M	OSFS	FOSFS	SAOLA	K-OFSD	OFS-D	OHS-NDER
AWA	0.512 6	0.532 3	0.541 8	0.540 2	0.546 5	0.518 1	0.548 9
Bridges	0.803 1	0.806 8	0.806 8	0.806 8	0.745 4	0.785 2	0.817 3
CLEF	0.607 1	0.642 7	0.665 3	0.652 1	0.665 3	0.642 2	0.679 1
DD	0.710 2	0.599 1	0.611 7	0.612 8	0.630 0	0.631 2	0.719 4
F194	0.456 7	0.429 0	0.511 8	0.553 8	0.593 2	0.591 9	0.624 7
VOC	0.511 0	0.506 3	0.509 4	0.504 3	0.528 6	0.502 0	0.524 0
Avg	0.600 1	0.586 0	0.607 8	0.611 7	0.618 2	0.611 8	0.652 2

表 11 基于 LSVM 分类器的 TIE 值(↓)
Table 11 TIE score using the LSVM classifier(↓)

数据集	A3M	OSFS	FOSFS	SAOLA	K-OFSD	OFS-D	OHS-NDER
AWA	3.899 1	3.741 3	3.665 6	3.678 2	3.627 8	3.854 9	3.608 8
Bridges	1.000 0	1.009 3	1.009 3	1.009 3	1.333 3	1.138 9	0.944 4
CLEF	2.733 3	2.548 7	2.400 0	2.469 2	2.451 3	2.538 5	2.359 0
DD	1.738 8	2.405 5	2.329 9	2.469 2	2.219 9	2.213 1	1.683 8
F194	3.259 8	3.582 7	2.929 1	2.677 2	2.440 9	2.448 8	2.252 0
VOC	2.881 4	2.872 9	2.861 6	2.887 0	2.785 3	2.906 8	2.819 2
Avg	2.585 4	2.693 4	2.532 6	2.531 7	2.476 4	2.516 8	2.277 9

根据表 4-表 11 的结果可得如下结论:

(1)从整体角度上,OHS-NDER 在 6 个数据集上的 3 个评价指标的平均性能均排名第一. 在 3 个评价指标中,OHS-NDER 至少在一半以上的数据集上性能最优. 故从总体性能上看,OHS-NDER 在各个评价指标上性能均为最优.

(2)OHS-NDER 在 KNN 和 LSVM 两个分类器中,在 DD 数据集上的分类性能在所有指标上均为最优,且在其他数据集上的分类性能和最优值相差不大,故该算法在 6 个数据集上的分类性能比较稳定.

为了更直观地对比 OHS-NDER 与 6 个对比算法之间分类性能的稳定性,采用蜘蛛图进行实验结果分析. 图 1 所示为在各个数据集下不同算法的稳定性. 从图 1 可以观察到:

(1)在平均预测精度、LCA 和 H-F1 指标下,OHS-NDER 的形状接近于正六方形;在 TIE 指标下,覆盖面积较广;说明 OHS-NDER 得到的解更优.

(2)在各个指标下,OHS-NDER 至少在 3 个数据集上拥有最优性能.

(3)在所有指标下,OHS-NDER 的覆盖面积远大于其他算法,说明 OHS-NDER 能够获得更稳定的解.

根据上述实验结果可知,OHS-NDER 算法比其他算法具有更强的稳定性.

4 结论

本文提出了一种基于邻域决策误差率的层次分类在线流特征选择算法,该算法利用层次结构中的兄弟策略定义了一种新的邻域关系;同时改进邻域决策误差率模型,提出了一种基于层次结构的邻域决策误差率模型;并将所提出的在线流特征选择算法分为两个步骤:在线重要度分析和在线冗余更新. 实验结果显示,在 6 个层次化结构数据集上,OHS-NDER 能够选择出较优的特征子集. 本文仅考虑了单个特征流的在线流特征选择问题,在未来的工作中,将进一步讨论层次化结构数据的在线流组特征选择问题.

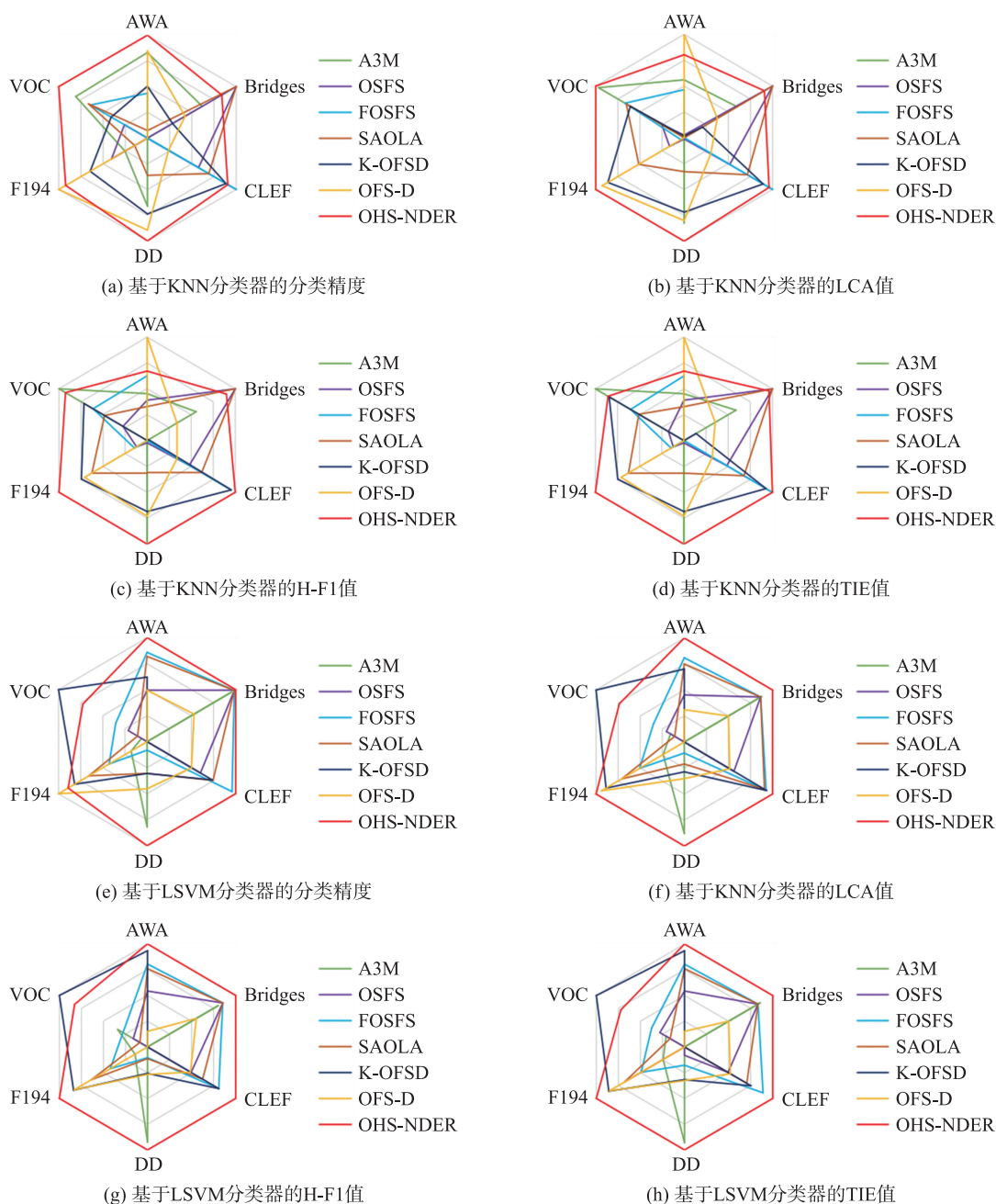


图 1 蜘蛛图显示算法在不同指标和 6 个数据集下的稳定性

Fig. 1 The stability index values obtained on six datasets with different evaluation metrics by spider web diagram

[参考文献] (References)

- [1] 胡清华,王煜,周玉灿,等. 大规模分类任务的分层学习方法综述[J]. 中国科学:信息科学,2018,48(5):7-20.
- [2] 赵红. 面向层次结构数据的特征选择方法[D]. 天津:天津大学,2019.
- [3] FREEMAN C, KULIC D, BASIR O. Joint feature selection and hierarchical classifier design[C]//Proceedings of 2011 IEEE International Conference on Systems, Man and Cybernetics. Anchorage, USA:IEEE,2011.
- [4] SONG J, ZHANG P Z, QIN S J, et al. A method of the feature selection in hierarchical text classification based on the category discrimination and position information [C]//Proceedings of 2015 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration. Wuhan, China:ICIICII,2015.
- [5] PAN S R, WU J, ZHU X Q. Cogboost: boosting for fast cost-sensitive graph classification [J]. IEEE Transactions on Knowledge & Data Engineering,2015,27(11):2933-2946.

- [6] ZHAO H,ZHU P F,WANG P,et al. Hierarchical feature selection with recursive regularization[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia; AAAI Press, 2017:3483–3489.
- [7] ZHOU J,FOSTER D P,STINE R A,et al. Streamwise feature selection[J]. Journal of Machine Learning Research, 2006, 7(1):1861–1885.
- [8] YU K,WU X D,DING W,et al. Scalable and accurate online feature selection for big data[J]. ACM Transactions on Knowledge Discovery from Data, 2016, 11(2):1–39.
- [9] LIN Y J,HU Q H,LIU J H,et al. Streaming feature selection for multi-label learning based on fuzzy mutual information[J]. IEEE Transactions on Fuzzy Systems, 2017, 25(6):1491–1507.
- [10] LIU J H,LIN Y J,WU S X,et al. Online multi-label group feature selection[J]. Knowledge-Based Systems, 2018, 143:42–57.
- [11] LI H G,WU X D,LI Z,et al. Group feature selection with streaming features [C]//Proceedings of 2013 IEEE 13th International Conference on Data Mining. Dallas, USA;IEEE, 2013.
- [12] HU Q H,PEDRYCZ W,YU D R,et al. Selecting discrete and continuous features based on neighborhood decision error minimization[J]. IEEE Transactions on Systems Man & Cybernetics(Part B), 2010, 40(1):137–150.
- [13] EISNER R,POULIN B,SZAFRON D,et al. Improving protein function prediction using the hierarchical structure of the gene ontology[C]//Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. La Jolla, USA;IEEE, 2005.
- [14] CECI M,MALERBA D. Classifying web documents in a hierarchy of categories: a comprehensive study[J]. Journal of Intelligent Information Systems, 2007, 28(1):37–78.
- [15] WU X D,YU K,DING W,et al. Online feature selection with streaming features[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(5):1178–1192.
- [16] ZHOU P,HU X G,LI P P,et al. OFS-Density: A novel online streaming feature selection method[J]. Pattern Recognition, 2019, 86:48–61.
- [17] ZHOU P,HU X G,LI P P,et al. Online feature selection for high-dimensional class-imbalanced data[J]. Knowledge-Based Systems, 2017, 136:187–199.
- [18] ZHOU P,HU X G,LI P P. A new online feature selection method using neighborhood rough set[C]//Proceedings of 2017 IEEE International Conference on Big Knowledge. Hefei, China;IEEE, 2017.

[责任编辑:严海琳]