

基于特征交互的层次分类在线流特征选择

孔令蔚^{1,2}, 蔡林晟^{1,2}, 林少杰^{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] 在开放动态环境下的分类学习任务中, 数据特征空间具有动态性, 标记空间存在层次化结构. 现有的层次分类在线流特征选择算法可以选择较优的特征子集, 但这些算法忽略了特征之间存在的交互作用. 基于此, 提出了一种基于特征交互的层次分类在线流特征选择算法. 首先, 设计了一种基于层次邻域依赖度去判断特征交互的计算方法; 其次, 针对层次化结构数据, 根据层次结构中不同节点间的兄弟关系定义邻域粗糙集模型; 最后, 设计了具有在线重要性分析、在线冗余性分析以及在线交互性分析的层次分类在线流框架, 用于选择强相关和存在交互作用特征子集. 在 6 个层次数据集上的实验验证了所提算法具有较优的综合性能.

[关键词] 在线流特征选择, 层次分类, 特征交互, 兄弟策略, 邻域粗糙集

[中图分类号] TP181 **[文献标志码]** A **[文章编号]** 1672-1292(2024)02-0034-09

Online Hierarchical Streaming Feature Selection Based on Feature Interaction

Kong Lingwei^{1,2}, Cai Linsheng^{1,2}, Lin Shaojie^{1,2}, Lin Yaojin^{1,2}

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

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

Abstract: In classification learning tasks in open dynamic environments, the data feature space is dynamic and there is a hierarchical structure in the labelling space. Existing hierarchical classification online streaming feature selection algorithms can select a superior subset of features, but these algorithms ignore the interactions that exist between the features. Therefore, this paper proposes a feature selection algorithm for hierarchical classification online streaming based on feature interaction. Firstly, a computational method based on hierarchical neighborhood dependency is designed to judge the feature interaction. Secondly, for hierarchical structure data, a neighborhood rough set model is defined on the basis of sibling relationships between different nodes in the hierarchical structure. Finally, the online streaming framework is designed for hierarchical classification with online importance analysis, online redundancy analysis and online interaction analysis for selecting the subset of features that are strongly correlated and have interaction. The proposed algorithm is experimentally verified on six hierarchical datasets to have superior comprehensive performance.

Key words: online streaming feature selection, hierarchical classification, feature interaction, sibling strategy, neighborhood rough set

在现实分类学习任务中, 分类建模面临着海量数据样本和高维且动态特征空间的挑战, 数据的类别空间往往具有层次化结构关系, 如图像数据^[1]、蛋白质数据^[2]. 利用类别标记的层次结构关系设计分治策略, 可以有效降低分类难度^[3]. 基于此, 研究人员利用特征选择的方法筛选关键特征, 提高模型分类性能并改善预测精度, 提出了考虑类别标记层次化结构关系的特征选择算法. Shi 等^[4]提出了一种利用层次化结构和特征关系最大化类间独立性及最小化类内冗余的分层分类方法. Liu 等^[5]提出了一种带有标签增强的层次分类特征选择算法, 以解决语义差距为分类任务带来的错误传播问题. Guo 等^[6]提出了一种具有多粒度聚类结构的层次特征选择方法, 能够有效解决层次结构中的语义差距问题. 林耀进等^[7]通过研究类别标记层次结构来分析标签相关性, 提出了基于标签关联性的分层分类共有与固有特征选择算法. 然

收稿日期: 2023-11-14.

基金项目: 国家自然科学基金面上项目(62076116).

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

而,在实际应用中数据特征往往随时间推移不断产生,逐一或成组流入特征空间,具有动态性和未知性^[8].

上述算法均假设特征空间固定且已提前获取所有特征,忽略了特征空间的动态性.例如,新闻热点每天都在变化,新闻关键词作为新闻分类任务的重要特征也随之改变.为在动态特征空间中进行特征选择,在线流特征选择算法被广泛提出.林耀进等^[9]针对不平衡数据,考虑边界样本,提出了基于最大决策边界的高维类不平衡数据在线流特征选择算法.Wu等^[10]为了在动态的特征空间中选择强相关特征并剔除冗余特征,提出了一种单标记在线流算法.You等^[11]提出了一种新的考虑标签相关性的在线多标签流特征选择方法.Lin等^[12]使用模糊互信息作为特征评价方式,提出了一种针对多标记数据的在线流特征选择算法.上述算法是根据特征与类别标记的相关性强弱选择特征^[13],去除不相关、弱相关特征,选择强相关特征.

在实际应用中,弱相关特征也尤为重要,因为存在两个弱相关特征结合使用时与类别高度相关的情况,称之为特征交互^[14].例如,在疾病诊断任务中,“咳嗽”与“肺癌”相关性弱,“胸痛”与“肺癌”相关性弱,但当“咳嗽”和“胸痛”特征结合时却与“肺癌”标签密切相关,说明特征交互可以达到“1+1>2”的效果^[15].若在层次在线流算法中加入特征交互判断,不仅能够选择强相关特征,还能重视弱相关特征,提高特征选择的精确度.基于此,本文提出一种基于特征交互的层次分类在线流特征选择算法,区别于传统层次分类特征选择算法,考虑了特征之间的交互作用.首先,设计了一种基于层次邻域依赖度去判断特征交互的计算方法;其次,针对层次化结构数据,根据层次结构中不同节点间的兄弟关系定义邻域粗糙集模型;然后,根据依赖度计算特征之间的相关性,设计出层次在线流框架,包括在线重要性分析、在线冗余性分析和在线交互性分析这3种策略,以达到选择最优特征子集的目的;最后,设计算法并进行实验,将本算法与5个在线流特征选择算法进行对比,验证所提算法的有效性.

1 背景知识

1.1 类别的层次结构

类别的层次结构分为树状结构和有向无环图结构^[16],本文关注前者.层次树结构中具有从属关系,表现为3个特性:反自反性、不可逆性、传递性^[17].层次结构中的不同类别节点之间具有多种关系,如表1所示.

1.2 面向层次结构化数据的邻域粗糙集

在面向标记层次结构数据的特征选择过程中,可以使用层次关系中的兄弟关系来区分同异类样本,也称为兄弟策略^[18],即同类样本为 x ,异类样本为 $\text{Sib}(x)$.

定义1^[19] 给定层次邻域交互决策系统 $\text{HNIDS} = \langle U, C, D, H \rangle$,其中, U 为样本集合, C 为条件属性集合,决策属性集合 D 存在层次关系 H .样本 $x_i \in U$,基于兄弟策略样本 x 的最大近邻点集为

$$\delta^{\text{Sib}}(x) = \{y \mid \Delta(x, y) \leq d^{\text{Sib}}(x), y \in U\}. \quad (1)$$

式中,

$$\begin{aligned} d^{\text{Sib}}(x) &= \max(d_1^{\text{Sib}}(x), d_2^{\text{Sib}}(x)), \\ d_1^{\text{Sib}}(x) &= \Delta(x, \text{NH}^{\text{Sib}}(x)), d_2^{\text{Sib}}(x) = \Delta(x, \text{NM}^{\text{Sib}}(x)). \end{aligned}$$

式中, Δ 是度量函数,例如欧式距离函数; $\text{NH}^{\text{Sib}}(x)$ 表示样本 x 的最近同类样本; $\text{NM}^{\text{Sib}}(x)$ 表示样本 x 的最近异类样本; $d_1^{\text{Sib}}(x)$ 表示样本 x 到 $\text{NH}^{\text{Sib}}(x)$ 的距离; $d_2^{\text{Sib}}(x)$ 表示样本 x 到 $\text{NM}^{\text{Sib}}(x)$ 的距离.

定义2^[19] 给定 $\text{HNIDS} = \langle U, C, D, H \rangle$,样本 x 的分类间隔定义为

$$m^{\text{Sib}}(x) = d_2^{\text{Sib}}(x) - d_1^{\text{Sib}}(x). \quad (2)$$

定理1^[19] 给定 $\text{HNIDS} = \langle U, C, D, H \rangle$,条件属性子集 $B \subseteq C$,决策属性 D 关于 B 的兄弟策略邻域下近似为

$$\underline{R}_B^{\text{Sib}} D = \{x_j \mid m_B^{\text{Sib}}(x_j) > 0, x_j \in U\}. \quad (3)$$

证明 若 $m_B^{\text{Sib}} > 0$,则 $d_2^{\text{Sib}}(x) > d_1^{\text{Sib}}(x)$,于是 $\Delta(x_j, \text{NH}^{\text{Sib}}(x_j)) < d_2^{\text{Sib}}(x_j)$,存在 $\delta_B^{\text{Sib}}(x_j) \subseteq D$,所以 $x_j \in \underline{R}_B^{\text{Sib}} D$ 成立.

表1 层次关系符号说明

Table 1 Description of hierarchical relationship symbols

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
$\text{Anc}(d_i)$	类别 d_i 的祖先节点集合
$\text{Des}(d_i)$	类别 d_i 的子孙节点集合
$\text{Sib}(d_i)$	类别 d_i 的兄弟节点集合

定义 3^[19] 给定 HNIDS= $\langle U, C, D, H \rangle$, 决策属性 D 对条件属性子集 $B \subseteq C$ 的兄弟策略依赖度为

$$\gamma_B^{\text{Sib}}(D) = \frac{\text{Card}(R_B^{\text{Sib}}(D))}{\text{Card}(U)}. \quad (4)$$

定义 4^[19] 给定 HNIDS= $\langle U, C, D, H \rangle$, 特征子集 $B \subseteq C$, $\forall f \notin B$, 特征 f 的兄弟策略重要度为

$$\text{HSig}_B^{\text{Sib}}(f, B, D) = \gamma_{B \cup f}^{\text{Sib}}(D) - \gamma_B^{\text{Sib}}(D). \quad (5)$$

定义 5^[20] 给定 HNIDS= $\langle U, C, D, H \rangle$, 特征 $A \subseteq C$, 特征 $A' \subseteq C - A$, 若特征 A 与特征 A' 存在交互作用, 则满足:

$$\gamma_{A, A'}^{\text{Sib}}(D) > \gamma_A^{\text{Sib}}(D) + \gamma_{A'}^{\text{Sib}}(D). \quad (6)$$

2 基于特征交互的层次分类在线流特征选择算法

2.1 在线流特征选择框架

基于特征交互的层次分类在线流特征选择算法能够选择关键特征, 通过动态更新最终选择最优特征子集. 本文提出在线流框架的 3 种特征分析评估策略.

2.1.1 在线重要性分析

定义 6 结合在线流特征随时间流入特征空间的特点, 给定层次交互决策系统 HIDST= $\langle U, C, D, H, A, t \rangle$, 其中, U 为数据集中样本集合, C 为特征集合, D 为类别集合, H 为 D 上存在的层次关系, $A \subseteq C$ 为待交互判断特征子集. 设 f_t 表示新特征在 t 时刻流入特征空间, 根据式(4), 可定义 f_t 基于兄弟策略的重要度为

$$\text{HSig}_{S_{t-1}}^{\text{Sib}}(f_t, S_{t-1}, D) = \gamma_{S_{t-1} \cup f_t}^{\text{Sib}}(D) - \gamma_{S_{t-1}}^{\text{Sib}}(D). \quad (7)$$

式中, $S_{t-1} \subseteq C$ 为 $t-1$ 时刻选择的特征子集. $\text{HSig}_{S_{t-1}}^{\text{Sib}}(f_t, S_{t-1}, D)$ 越大, 说明当前选择的特征 f_t 越重要; 相反则说明特征子集 S_{t-1} 中存在与特征 f_t 冗余的特征, 需要进行下一步在线冗余更新.

2.1.2 在线冗余性分析

根据定义 6, 基于兄弟策略的重要度 $\text{HSig}_{S_{t-1}}^{\text{Sib}}(f_t, S_{t-1}, D) \leq 0$ 是判断特征重要性的方法. 通过设计类似重要度公式的冗余度判断方法, 可以准确地分析冗余特征, 并更新特征子集, 提高特征选择的精确度.

定义 7 给定 HIDST= $\langle U, C, D, H, A, t \rangle$, f_t 对于 D 的兄弟策略冗余度为

$$\text{HRSig}_{S_t}^{\text{Sib}}(f, S_t, D) = \gamma_{S_t}^{\text{Sib}}(D) - \gamma_{S_t - f}^{\text{Sib}}(D). \quad (8)$$

通过在线冗余更新, 将 $\text{HRSig}_{S_t}^{\text{Sib}}(f, S_t, D) \leq 0$ 作为冗余特征的判断依据, 找到冗余特征, 对冗余特征进行交互分析.

2.1.3 在线交互性分析

大多数传统特征选择方法选择直接丢弃冗余特征, 存在忽略了冗余特征中部分特征可交互的情况. 根据式(6), 将冗余特征进行特征交互分析后, 可在线识别可交互的特征.

定义 8 给定 HIDST= $\langle U, C, D, H, A, t \rangle$, A 是待交互判断的特征子集, f_i 和 f_j 是 A 中两个特征, 若满足

$$\gamma_{f_i, f_j}^{\text{Sib}}(D) > \gamma_{f_i}^{\text{Sib}}(D) + \gamma_{f_j}^{\text{Sib}}(D), \quad (9)$$

则称特征 f_i 和特征 f_j 存在交互作用, 将 f_i 和 f_j 加入特征子集 S_t , 否则将 f_i 和 f_j 丢弃.

以上 3 种策略组成流特征选择模型框架, 首先进行在线重要性分析, 然后进行在线冗余性分析, 最后进行在线交互性分析, 从而得到最终特征子集.

2.2 OHFS-FI 算法设计

根据以上多个定义, 得到在线流特征选择框架, 基于此, 算法 OHFS-FI 详细步骤如下所示:

算法 OHFS-FI

输入: 层次交互决策系统 $\langle U, C, D, H, A, t \rangle$;

属性依赖度阈值 $\delta (0 \leq \delta \leq 1)$;

输出: t 时刻的特征子集 S_t ;

(1) REPEAT

(2) t 时刻流入新特征 f_t ;

(3) 通过公式(4)计算 $\gamma_B^{\text{Sib}}(D)$;

```

(4) IF  $\gamma_B^{\text{Sib}}(D) \geq \delta$  /* 在线重要性分析 */
(5) 通过公式(7)计算  $\text{HSig}_t^{\text{Sib}}(f_i, S_{t-1}, D)$ ;
(6) IF  $\text{HSig}_t^{\text{Sib}}(f_i, S_{t-1}, D) > 0$ 
(7)  $S_t = S_{t-1} \cup f_i$ ;
(8) ELSE /* 在线冗余性分析 */
(9)  $S_t = S_{t-1} \cup f_i$ ;
(10) FOR  $S_t$  中的每个特征
(11) 随机选择  $S_t$  中的一个特征  $f$ ;
(12) 通过公式(8)计算  $\text{HRSig}_t^{\text{Sib}}(f, S_t, D)$ ;
(13) IF  $\text{HRSig}_t^{\text{Sib}}(f, S_t, D) \leq 0$ 
(14)  $S_t = S_t - f$ ;
(15)  $A_t = A_{t-1} \cup f$ ;
(16) END IF
(17) END FOR
(18) 选择  $A_t$  中  $\gamma_f^{\text{Sib}}(D)$  最大的特征  $f_i$ ; /* 在线交互性分析 */
(19)  $A_t = A_t - f_i$ ;
(20) FOR  $A_t$  中的每个特征
(21) IF  $\gamma_{f_i, f_j}^{\text{Sib}}(D) > \gamma_{f_i}^{\text{Sib}}(D) + \gamma_{f_j}^{\text{Sib}}(D)$ 
(22)  $S_t = S_t \cup f_i \cup f_j$ ;
(23)  $A_t = A_t - f_j$ ;
(24) END IF
(25) END FOR
(26) END IF
(27) END IF
(28) 直到没有新的特征到达, 返回  $S_t$ .

```

本算法中, 当 t 时刻特征 f_t 到达时, 先执行第 3 步计算特征依赖度 $\gamma_B^{\text{Sib}}(D)$. 然后进行重要性分析, 当 $\gamma_B^{\text{Sib}}(D) < \delta$ 时, 将特征 f_t 删除; 否则执行第 5 步计算 f_t 对于已选特征子集的重要度 $\text{HSig}_t^{\text{Sib}}(f_i, S_{t-1}, D)$. 若重要度大于 0, 执行第 7 步, 将特征 f_t 加入选择特征子集 S_t , 否则执行第 9 至 17 步冗余性分析. 先将特征 f_t 加入 S_t , 然后在 S_t 中随机选择一个特征 f , 在第 12 步计算冗余度 $\text{HRSig}_t^{\text{Sib}}(f, S_t, D)$, 若冗余度小于等于 0, 从 S_t 中删除特征 f , 再将 f 加入待交互判断特征子集 A_t . 最后执行第 18 至 25 步交互性分析, 找到 A_t 中依赖度最大的特征 f_i , 并从 A_t 中删除该特征. 接下来执行第 21 至 23 步, 从 A_t 选择一个特征 f_j , 若满足 $\gamma_{f_i, f_j}^{\text{Sib}}(D) > \gamma_{f_i}^{\text{Sib}}(D) + \gamma_{f_j}^{\text{Sib}}(D)$, 执行第 22 步, 将选择的特征加入特征子集 S_t , 再将特征 f_j 从 A_t 中删除, 否则直接将特征 f_j 从 A_t 中删除.

3 实验

3.1 实验数据及环境设置

本实验选取 6 个数据集用于验证 OHFS-FI 算法进行特征选择的性能, 数据集均具有层次化结构. 数据集描述信息如表 2 所示.

所有实验均采用两个分类器 (KNN、LSVM) 进行性能评估, 并使用 10 折交叉验证. 实验环境为 Matlab 2021b, 所有实验都在同一台 Intel XeonE5, 3.00GHz, 256GB 内存的计算机上运行.

表 2 层次数据集

Table 2 Hierarchical datasets

数据集	样本数	特征数	内部节点	叶子节点	层数	数据集	样本数	特征数	内部节点	叶子节点	层数
AWA	6 405	252	17	10	4	VOC	7 178	1 000	30	88	5
DD	3 020	473	5	27	3	Bridges	108	12	8	6	3
F194	7 015	473	202	194	3	CLEF	8 368	80	25	63	4

3.2 评价指标

评价指标选择传统预测精度 (AP) 和两种层次分类指标 (分别为树诱导误差 (TIE)^[21]、最近共同祖先

F1(LCA-F1)^[22]). 两种层次评价指标的定义和计算公式如下:

(1) 树诱导误差(TIE): 根据样本的预测和真实类别在层次节点间距离定义惩罚. 设 D 表示真实类别, \hat{D} 表示预测类别, $|E_H(D, \hat{D})|$ 表示在 D 与 \hat{D} 节点之间的最小总边数, 则树诱导误差可表示为:

$$\text{TIE}(D, \hat{D}) = |E_H(D, \hat{D})|.$$

(2) 最近共同祖先 F1(LCA-F1): $\text{LCA}(D, \hat{D})$ 表示真实和预测类别节点的最近共同祖先, 可以避免两个节点有多个共同祖先而影响评价结果的情况:

$$D_{\text{aug}} = D \cup \text{Anc}(D), \hat{D}_{\text{aug}} = \hat{D} \cup \text{Anc}(\hat{D}),$$

$$\text{LCA-F1} = \frac{2P_{\text{LCAH}}R_{\text{LCAH}}}{P_{\text{LCAH}} + R_{\text{LCAH}}}.$$

式中,

$$P_{\text{LCAH}} = \frac{|D_{\text{aug}}^{\text{LCA}} \cap \hat{D}_{\text{aug}}^{\text{LCA}}|}{|\hat{D}_{\text{aug}}^{\text{LCA}}|}, R_{\text{LCAH}} = \frac{|D_{\text{aug}}^{\text{LCA}} \cap \hat{D}_{\text{aug}}^{\text{LCA}}|}{|D_{\text{aug}}^{\text{LCA}}|}.$$

对于上述评价指标, LCA-F1 和预测精度的值越大越好, TIE 指标的值越小越好.

3.3 实验结果分析

为验证 OHFS-FI 算法的性能, 本文设计了 3 种实验: 参数分析、算法比较以及统计验证.

3.3.1 参数分析

为使算法达到最佳的性能, 选择不同的参数 δ 取值进行实验, δ 取值分别为 0.01, 0.02, 0.03, 0.04, 0.05, 选择 3 个数据集在 KNN 和 LSVM 分类器上进行实验, 结果如图 1、图 2 所示.

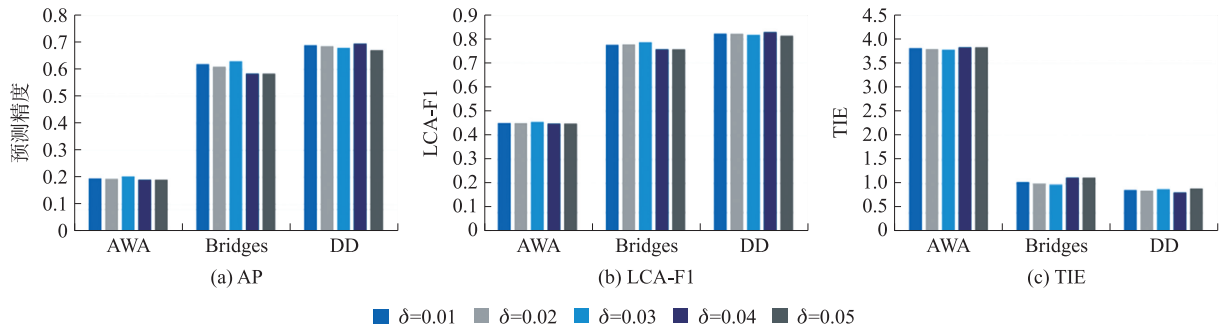


图 1 δ 不同时 OHFS-FI 在 KNN 分类器上的性能对比

Fig. 1 Performance comparison of OHFS-FI using KNN classifier with different δ values

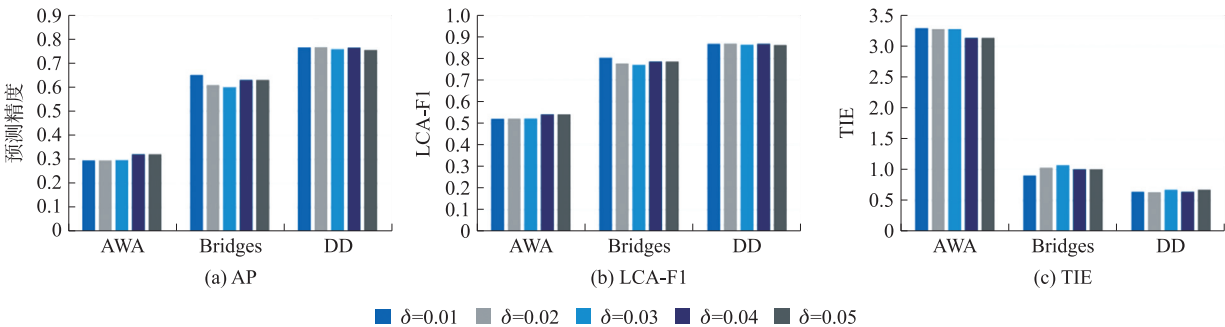


图 2 δ 不同时 OHFS-FI 在 LSVM 分类器上的性能对比

Fig. 2 Performance comparison of OHFS-FI using LSVM classifier with different δ values

根据上述实验结果可以看出, 当使用 KNN 分类器时, 参数 $\delta=0.03$ 时可以在 Bridges 和 AWA 数据集上取得最好效果, 而 $\delta=0.01$ 时可以取得次优效果; 对于 DD 数据集, $\delta=0.04$ 时可以在 3 种评价指标上得到最好效果, 其次是 $\delta=0.01$. 当使用 LSVM 分类器时, 参数 $\delta=0.01$ 时在 Bridges 和 DD 数据集上的 3 种评价指标均有最佳表现; 对于 AWA 数据集, 参数 $\delta=0.01$ 时可以在 TIE 评价指标上获最佳结果. 通过分析结果可以发现, 算法 OHFS-FI 在参数 $\delta=0.01$ 时表现稳定, 能够均衡各种因素影响, 可以代表算法 OHFS-FI 的真实性能. 因此, 以下实验参数均设置 $\delta=0.01$.

3.3.2 算法性能比较

为验证 OHFS-FI 算法在层次结构数据集上的特征选择性能,选取了 5 种在线流算法进行算法性能比较:在线流特征选择(OSFS)^[10]、快速在线特征选择(FOSFS)^[10]、一种新的在线特征选择方法(OFSD)^[23]、准确且可扩展的在线流特征选择算法(SAOLA)^[24]、基于邻域粗糙集的大规模层次分类在线流特征选择(OHFS)^[19]。实验根据算法对应文献中最优参数进行设置。

分类结果如表 3-表 8 所示,表中“↑”表示该指标越大越好,“↓”表示该指标越小越好,所有数据集在各算法中表现最优的值用黑体字表示,最后一行为平均值。

表 3 各算法在 KNN 分类器的分类精度(↑)

Table 3 Predictive accuracy of algorithms with KNN classifier(↑)

数据集	OSFS	FOSFS	OFSD	SAOLA	OHFS	OHFS-FI
AWA	0.199 1	0.193 9	0.180 2	0.171 7	0.192 5	0.195 3
DD	0.405 8	0.405 8	0.558 4	0.317 2	0.624 2	0.689 2
F194	0.242 9	0.289 3	0.221 7	0.259 6	0.469 3	0.534 6
VOC	0.192 3	0.202 6	0.223 9	0.176 8	0.284 1	0.282 0
Bridges	0.630 0	0.417 8	0.507 8	0.630 0	0.630 0	0.618 9
CLEF	0.671 9	0.672 3	0.489 6	0.459 7	0.810 2	0.824 3
Avg	0.390 3	0.363 6	0.363 6	0.335 8	0.501 7	0.524 1

表 4 各算法在 KNN 分类器的 LCA-F1 值(↑)

Table 4 LCA-F1 score of algorithms with KNN classifier(↑)

数据集	OSFS	FOSFS	OFSD	SAOLA	OHFS	OHFS-FI
AWA	0.452 5	0.452 0	0.443 9	0.434 8	0.449 1	0.451 2
DD	0.659 8	0.659 8	0.749 9	0.596 9	0.792 4	0.823 6
F194	0.566 4	0.600 5	0.563 0	0.583 0	0.698 1	0.733 9
VOC	0.471 4	0.479 3	0.497 4	0.460 7	0.537 6	0.535 2
Bridges	0.785 2	0.674 1	0.715 7	0.785 2	0.788 0	0.776 9
CLEF	0.774 6	0.775 6	0.644 4	0.623 4	0.874 5	0.884 4
Avg	0.618 3	0.606 9	0.602 4	0.580 6	0.690 0	0.700 8

表 5 各算法在 KNN 分类器的 TIE 值(↓)

Table 5 TIE score of algorithms with KNN classifier(↓)

数据集	OSFS	FOSFS	OFSD	SAOLA	OHFS	OHFS-FI
AWA	3.825 8	3.795 5	3.832 9	3.932 6	3.837 3	3.815 8
DD	1.705 4	1.705 4	1.234 2	2.106 5	0.988 1	0.874 2
F194	2.175 2	1.951 0	2.130 9	2.042 0	1.497 7	1.330 3
VOC	3.217 6	3.154 9	2.995 8	3.292 3	2.735 4	2.764 6
Bridges	1.009 3	1.481 5	1.314 8	1.009 3	0.981 5	1.037 0
CLEF	1.527 6	1.511 5	2.476 2	2.599 8	0.812 4	0.746 5
Avg	2.243 5	2.266 6	2.330 8	2.497 1	1.808 7	1.761 4

表 6 各算法在 LSVM 分类器的分类精度(↑)

Table 6 Predictive accuracy of algorithms with LSVM classifier(↑)

数据集	OSFS	FOSFS	OFSD	SAOLA	OHFS	OHFS-FI
AWA	0.208 0	0.219 5	0.186 7	0.178 1	0.318 8	0.294 9
DD	0.370 4	0.370 7	0.307 9	0.292 9	0.711 8	0.766 4
F194	0.219 7	0.253 7	0.101 0	0.225 2	0.496 0	0.586 6
VOC	0.259 4	0.267 1	0.289 8	0.256 8	0.360 8	0.341 0
Bridges	0.630 0	0.630 0	0.564 4	0.630 0	0.600 0	0.651 1
CLEF	0.657 5	0.655 7	0.504 4	0.466 9	0.799 5	0.824 2
Avg	0.390 8	0.399 5	0.325 7	0.341 7	0.547 8	0.577 3

表 7 各算法在 LSVM 分类器的 LCA-F1 值(↑)

Table 7 LCA-F1 score of algorithms with LSVM classifier(↑)

数据集	OSFS	FOSFS	OFSD	SAOLA	OHFS	OHFS-FI
AWA	0.464 3	0.471 2	0.448 2	0.448 4	0.539 3	0.521 2
DD	0.633 8	0.633 9	0.574 3	0.578 6	0.839 6	0.868 4
F194	0.555 3	0.580 9	0.452 6	0.563 5	0.717 1	0.767 3
VOC	0.518 4	0.523 6	0.541 1	0.516 5	0.589 8	0.576 7
Bridges	0.785 2	0.785 2	0.752 8	0.785 2	0.770 4	0.803 7
CLEF	0.763 8	0.762 9	0.656 2	0.628 9	0.866 6	0.884 2
Avg	0.620 1	0.626 3	0.570 9	0.586 9	0.720 5	0.736 9

表 8 各算法在 LSVM 分类器的 TIE 值(↓)
Table 8 TIE score of algorithms with LSVM classifier(↓)

数据集	OSFS	FOSFS	OFSD	SAOLA	OHFS	OHFS-FI
AWA	3.686 8	3.643 4	3.815 1	3.762 7	3.164 1	3.300 5
DD	1.875 9	1.875 9	2.339 9	2.228 4	0.771 9	0.643 7
F194	2.213 3	2.041 8	2.971 3	2.137 0	1.375 2	1.137 8
VOC	2.817 5	2.787 7	2.680 0	2.826 8	2.390 6	2.470 7
Bridges	1.009 3	1.009 3	1.138 9	1.009 3	1.074 1	0.907 4
CLEF	1.612 0	1.613 6	2.372 7	2.538 4	0.875 1	0.748 1
Avg	2.202 5	2.162 0	2.553 0	2.417 1	1.608 5	1.534 7

通过对比表 3-表 8 的结果可得出如下结论:

(1) 基于 3 种评价指标,OHFS-FI 算法的平均性能在所有层次结构数据集上排名第一,且在一半以上的数据集上性能表现最优. 从整体的角度看,算法 OHFS-FI 优于其他对比算法,且在 3 个评价指标上性能最优.

(2) OHFS-FI 算法在 DD、F194、CLEF 和 Bridges 4 个数据集上的分类性能可以达到所有指标最优,在 AWA、VOC 数据集上算法的性能仅次于最优值,故 OHFS-FI 算法在 6 个数据集上的分类性能较优且稳定.

3.3.3 统计验证

使用 Friedman 检验^[25]和 Nemenyi 检验^[26]进一步讨论各算法分类性能上是否有显著差异以及实验结果在统计上的意义.

首先进行 Friedman 检验. 设 k 表示比较算法数量, N 表示数据集数量, 算法的平均排名为 $R_j = \frac{1}{N} \sum_{i=1}^N r_i^j$, 其中, 第 j 个算法在第 i 个数据集上的序列排名表示为 r_i^j , 则 Friedman 检验中统计量定义如下:

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}.$$

式中,

$$\chi_F^2 = \frac{12N}{k(k+1)} \left(\sum_{i=1}^k R_i^2 - \frac{k(k+1)^2}{4} \right).$$

表 9 和表 10 描述了在 3 种评价指标、2 种分类器下的 Friedman 检验 F_F 值和相应的临界值. 从两表可知, 当显著性水平 $\alpha=0.05$ 时, “所有算法在不同评价指标上的性能相等” 假设被拒绝.

表 9 基于 KNN 分类器不同评价指标的 F_F 值和临界值 Table 9 F_F -value and critical value for different evaluation indicators based on KNN classifier			表 10 基于 LSVM 分类器不同评价指标的 F_F 值和临界值 Table 10 F_F -value and critical value for different evaluation indicators based on LSVM classifier		
评价指标	F_F	临界值	评价指标	F_F	临界值
AP	4.197 1		AP	10.291 3	
LCA-F1	4.305 8	2.603	LCA-F1	10.594 1	2.603
TIE	4.170 3		TIE	10.328 5	

使用 Nemenyi 检验进一步比较算法之间的性能差异, 采用 CD 临界值进行比较, CD 的定义如下:

$$CD_\alpha = q_\alpha \sqrt{\frac{k(k+1)}{6N}}.$$

式中, 当 $\alpha=0.05$ 时, $q_\alpha=2.850$. 实验中比较算法数量 $k=6$, 数据集数量 $N=6$, 因此得到 $CD_{0.05}=3.078 4$. 根据各算法的平均排序绘制 Nemenyi 检验 CD 图如图 3、图 4 所示, 根据图中坐标轴上的平均排序差值判断是否存在显著性差异. 从图 3、图 4 可以得出以下结论: 基于 3 个评价指标, 在 KNN 分类器上, OHFS-FI 算法与 SAOLA 算法有显著性差异; 在 LSVM 分类器上, OHFS-FI 算法与 OFSD 算法、SAOLA 算法有显著性差异.

上述的实验结果表明算法 OHFS-FI 与其他比较算法存在显著性差异, 且算法 OHFS-FI 的性能明显优于其他算法, 具有更好的分类性能.

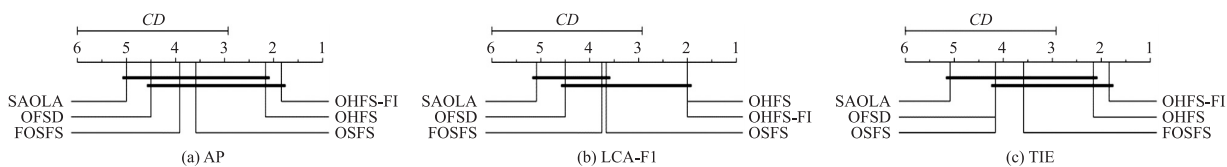


图3 基于 KNN 分类器的 Nemenyi 测试比较算法性能差异

Fig. 3 Differences in performance of comparative algorithms based on the Nemenyi test of KNN classifiers

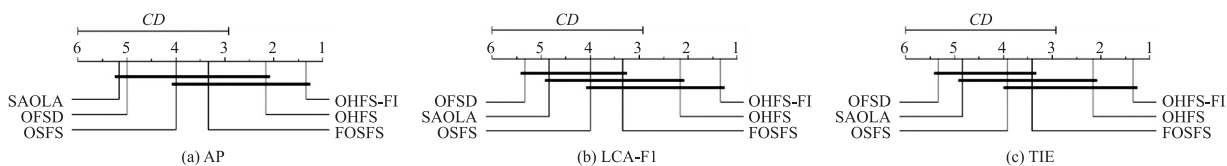


图4 基于 LSVM 分类器的 Nemenyi 测试比较算法性能差异

Fig. 4 Differences in performance of comparative algorithms based on the Nemenyi test of LSVM classifiers

4 结论

本文提出了一种基于特征交互的层次分类在线流特征选择算法,在特征选择过程中考虑了特征交互。首先,设计基于层次邻域依赖度去判断特征交互的方法;其次,根据层次结构中不同节点间的兄弟关系定义邻域粗糙集模型;最后,构建在线流特征选择框架:在线重要性分析、在线冗余性分析和在线交互性分析。通过在6个层次数据集上进行实验,发现所提出的 OHFS-FI 算法在不同评价指标上均取得较优结果,算法具有稳定性和有效性。在未来的工作中,将考虑存在特征不仅仅是逐一流入特征空间,而是成组流入的情况,进一步研究层次化结构数据的流组特征选择问题;同时考虑将类别标记的层次结构进行多粒度划分,提高特征选择的精度和效率。

[参考文献] (References)

- [1] KRAUSE J, STARK M, DENG J, et al. 3d object representations for fine-grained categorization[C]//Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops. Sydney, Australia: IEEE, 2013.
- [2] WEI L Y, LIAO M H, GAO X, et al. An improved protein structural classes prediction method by incorporating both sequence and structure information[J]. IEEE Transactions on NanoBioscience, 2014, 14(4): 339-349.
- [3] 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): 487-500.
- [4] SHI J, LI Z Y, ZHAO H. Feature selection via maximizing inter-class independence and minimizing intra-class redundancy for hierarchical classification[J]. Information Sciences, 2023, 626: 1-18.
- [5] LIU H Y, LIN Y J, WANG C X, et al. Semantic-gap-oriented feature selection in hierarchical classification learning[J]. Information Sciences, 2023, 642: 119241.
- [6] GUO S X, ZHAO H, YANG W Y. Hierarchical feature selection with multi-granularity clustering structure[J]. Information Sciences, 2021, 568: 448-462.
- [7] 林耀进, 白盛兴, 赵红, 等. 基于标签关联性的分层分类共有与固有特征选择[J]. 软件学报, 2022, 33(7): 2667-2682.
- [8] LI H G, WU X D, LI Z, et al. Group feature selection with streaming features[C]//Proceedings of the IEEE 13th International Conference on Data Mining. Dallas, USA: IEEE, 2013.
- [9] 林耀进, 陈祥焰, 白盛兴, 等. 基于最大决策边界的高维类不平衡数据在线流特征选择[J]. 模式识别与人工智能, 2020, 33(9): 820-829.
- [10] WU X D, YU K, DING W, et al. Online feature selection with streaming features[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(5): 1178-1192.
- [11] YOU D L, WANG Y, XIAO J W, et al. Online multi-label streaming feature selection with label correlation[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(3): 2901-2915.
- [12] LIN Y J, HU Q H, LIU J H, et al. Streaming feature selection for multilabel learning based on fuzzy mutual information[J]. IEEE Transactions on Fuzzy Systems, 2017, 25(6): 1491-1507.

- [13] KOHAVI R,JOHN G H. Wrappers for feature subset selection[J]. Artificial Intelligence,1997,97(1/2):273-324.
- [14] JAKULIN A,BRATKO I. Analyzing attribute dependencies[C]//Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases(PKDD2003). Cavtat-Dubrovnik, Croatia:PKDD,2003.
- [15] ZHOU P,WANG N,ZHAO S. Online group streaming feature selection considering feature interaction[J]. Knowledge-Based Systems,2021,226:107157.
- [16] WU F H,ZHANG J,HONAVAR V. Learning classifiers using hierarchically structured class taxonomies[C]//Proceedings of the 6th International Symposium on Abstraction, Reformulation and Approximation(SARA 2005). Airth Castle, Scotland, UK: SARA,2005.
- [17] SILLA JR C N,FREITAS A A. A survey of hierarchical classification across different application domains[J]. Data Mining and Knowledge Discovery,2011,22(1/2):31-72.
- [18] 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.
- [19] 白盛兴,林耀进,王晨曦,等. 基于邻域粗糙集的大规模层次分类在线流特征选择[J]. 模式识别与人工智能,2019,32(9):811-820.
- [20] ZENG Z L,ZHANG H J,ZHANG R, et al. A novel feature selection method considering feature interaction[J]. Pattern Recognition;the Journal of the Pattern Recognition Society,2015,48(8):2656-2666.
- [21] DEKEL O,KESHET J,SINGER Y. Large margin hierarchical classification[C]//Proceedings of the 21th International Conference on Machine Learning. Banff, Canada:ACM,2004.
- [22] KOSMOPOULOS A,PARTALAS I,GAUSSIER E, et al. Evaluation measures for hierarchical classification: A unified view and novel approaches[J]. Data Mining and Knowledge Discovery,2015,29(3):820-865.
- [23] ZHOU P,HU X G,LI P P, et al. OFS-Density: A novel online streaming feature selection method[J]. Pattern Recognition;the Journal of the Pattern Recognition Society,2019,86:48-61.
- [24] 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):16.
- [25] FRIEDMAN M. A comparison of alternative tests of significance for the problem of m rankings[J]. The Annals of Mathematical Statistics,1940,11(1):86-92.
- [26] NEMENYI P B. Distribution-Free Multiple Comparisons[M]. Princeton, USA:Princeton University ProQuest Dissertations Publishing,1963.

[责任编辑:严海琳]