

基于特征变权的动态模糊特征选择算法

王 玲 孟建瑶

(北京科技大学自动化学院 北京 100083)
(工业过程知识自动化教育部重点实验室(北京科技大学) 北京 100083)
(lingwang@ustb.edu.cn)

Dynamic Fuzzy Features Selection Based on Variable Weight

Wang Ling and Meng Jianyao
(College of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083)
(Key Laboratory of Knowledge Automation for Industrial Processes (University of Science and Technology Beijing), Ministry of Education, Beijing 100083)

Abstract In this paper, a new scheme for dynamic fuzzy feature selection based on variable weight is proposed to optimize the fuzzy feature subset with the important features dynamically. Firstly, the sliding window is adopted to divide the fuzzy dataset. In the first sliding window, the off-line fuzzy features selection algorithm is proposed to access the candidate fuzzy feature subset by calculating the weight of each fuzzy input feature according to the mutual information between the fuzzy input features and the output feature. Based on this, the optimal fuzzy feature subset are obtained by combining the backward feature selection method with the fuzzy feature selection index. With the new sliding window, the on-line fuzzy features selection algorithm is proposed, by integrating the optimal fuzzy feature selection result in the previous sliding window with the candidate fuzzy feature set in the current sliding window, the importance of the fuzzy input feature is calculated to obtain the optimal feature subset in the current window. Finally, the evolving relationship of the fuzzy input features is found with the fuzzy feature weights between the sliding windows. The simulation results show that the proposed algorithm has a significant improvement in the adaptability and prediction accuracy.

Key words dynamic; feature selection; fuzzy; mutual information; weight; silding window

摘 要 针对大多数动态特征选择算法不能实时地根据特征重要性的变化动态优化模糊特征的问题,提出了基于特征变权的动态模糊特征选择算法.该算法利用滑动窗口分割模糊化后的数据,在第1个窗口中进行离线模糊特征选择,根据输入模糊特征与输出特征的互信息量,计算各个模糊输入特征的权重,获取候选模糊特征子集,并采取后向特征选择的方式和模糊特征筛选指标得到优化模糊特征子集;在随后的窗口中进行在线模糊特征选择,结合当前窗口的候选模糊特征子集和已有模糊特征选择结果,计算模糊输入特征的重要度,获得当前窗口中的优化模糊特征子集.通过计算窗口之间模糊特征权重的变化,发现输入模糊特征的演化关系.不同数据集的仿真结果表明,所提出算法在自适应性和预测准确性方面均有显著提高.

关键词 动态;特征选择;模糊;互信息量;权重;滑动窗口

中图法分类号 TP18; TP273

特征选择是数据挖掘和机器学习中的一个重要课题,它不仅有助于理解数据、节约计算成本、减少特征之间的相互影响,而且可以提高预测的准确率^[1-2].特征选择也称特征子集选择,是指从原始特征中选择出一些最有效特征以降低数据集维度的过程^[3].目前,特征选择已经应用于多个领域,例如工业传感器数据分析^[4]、空气质量分析^[5]、健康数据分析^[6]等等.

在一些现实的系统中,人们往往利用一些模糊语义将原始特征集转化为模糊特征集,提高系统的可解释性.然而,模糊化后的特征空间比原始的特征空间维度更高,为了减少系统模糊化后的复杂性,从模糊特征中选择有价值的特征显得更为重要.此外,实际应用往往是具有多特征的动态复杂系统,随着时间的变化,特征的重要程度也在逐渐发生变化,例如一些重要的特征随着时间的推移变得冗余,或者起初不重要的特征随着时间的推移变得越来越重要.虽然国内外已经涌现了一系列特征选择算法,例如过滤式特征选择算法^[7]、封装式特征选择算法^[8]、嵌入式特征选择算法^[9]等用于提高学习算法的性能.但上述方法不能体现特征选择的实时性,导致算法延迟和不连续,影响了可解释性.因此,动态特征选择算法应运而生.文献[10]提出基于在线分类器的动态特征选择算法,该算法选择所有可以代表整体数据集信息的特征,但利用分类器评估所选特征的好坏,增加了计算负担;文献[11]提出基于贝叶斯分类器的封装式动态特征选择算法,该算法利用贪婪算法寻找最优的特征子集,增加了算法的时间复杂度和空间复杂度;文献[12]介绍了4种新的特征质量检测指标,然后用检测指标和聚类算法动态地选择有用的特征,但是聚类算法需要人为设置聚类个数,降低了算法的自适应性;文献[13]提出基于 K 均值算法和遗传算法的动态特征选择算法,该算法分别利用遗传算法和 K 均值算法确定目标函数和寻优范围,但该算法得到的特征子集不是全局最优的.尽管如此,针对动态系统,如何自适应地进行模糊特征选择仍然是一个挑战性的课题.

本文主要有4个方面的贡献:

1) 提出一种基于特征变权的动态模糊特征选择算法,包括离线模糊特征选择和在线模糊特征选择2部分;

2) 通过计算每个模糊特征的权重,动态优化所选模糊特征子集,保证了特征选择过程的平滑性和时效性;

3) 在特征变权的基础上,离线模糊特征选择利用后向选择算法和模糊特征筛选指标获得优化模糊特征子集,在线模糊特征选择根据重要度不断更新优化模糊特征子集;

4) 应用不同数据集的实验结果表明:我们提出的动态模糊特征选择算法在无需设置任何参数的情况下,提高了算法的自适应性和预测准确性.

1 相关工作

动态特征选择算法的目的是在动态数据集中选择出能够实时刻画主要特性的特征.关于动态特征选择的研究有很多^[14-18],其中,一些动态特征选择算法只能处理类别数据,例如文献[14]提出基于粗糙集的动态特征选择算法,以粗糙集为基础,动态地计算特征的独立性,从而进行特征选择;文献[15]提出基于邻域粗糙集的动态特征选择算法,邻域粗糙集摆脱了粗糙集只能处理离散型数据的约束;文献[16]提出基于信息粒度的动态特征选择算法,利用非矩阵的信息粒度结构减少了算法时间和空间的消耗,提高了算法效率.然而,上述算法都是针对类别数据集,适用范围较小.除此之外,研究者提出一些针对其他类型数据集的动态特征选择算法;文献[17]提出动态无监督特征选择算法,该算法将特征选择直接嵌入聚类算法中,利用聚类算法保存局部信息结构,帮助选择具有辨识能力的特征,然而,该算法需要设置参数控制聚类算法目标函数的构建,导致算法自适应能力不佳;文献[18]提出新的无监督动态特征选择算法,该算法近似地保存数据,根据在近似数据上的回归结果确定特征的重要程度,从而进行特征选择.然而,该算法依赖回归模型进行特征选择,没有独立的特征评价指标,限制了算法的使用范围.

针对模糊数据的特征选择算法较少,文献[19]提出基于最小-最大学习规则和扩展矩阵的模糊特征选择算法,利用模糊化后的数据构建扩展矩阵,根据最小-最大学习规则衡量每个模糊特征包含重要信息的多少,从而获得模糊特征选择结果,但是此算法

只适用于静态数据集,不适用于动态系统中的模糊特征选择;文献[20]提出基于在线特征权重的动态模糊特征选择算法,该算法计算每个模糊特征的权重,根据可分离判据,选择有助于分类的模糊特征,然而该算法中需要设置分离阈值,降低了算法的自适应能力。

鉴于此,本文提出了基于特征变权的动态模糊特征选择算法(dynamic fuzzy features selection based on variable weight, DFFS-VW). 主要包括 2 个阶段:

1) 离线模糊特征选择. 首先,利用滑动窗口分割模糊数据,在第 1 个滑动窗口中进行离线模糊特征选择,计算每一个模糊输入特征与输出特征的互信息量;然后根据互信息量计算每一个模糊输入特征的权重,并从大到小排序,计算相邻模糊输入特征的权重梯度和截断点阈值进而获得候选模糊特征子集,缩小了寻找最优模糊输入特征子集的空间,提高了算法效率;在此基础上,根据模糊特征筛选指标既考虑模糊特征子集的输入特征与输出特征之间的互信息量又兼顾模糊输入特征之间的互信息量来筛选特征,并通过后向特征选择算法得到离线的优化模糊特征子集,保证了所选模糊特征子集为全局最优。

2) 在线模糊特征选择. 在依次到来的滑动窗口中进行在线模糊特征选择,综合利用当前时刻滑动窗口中的候选模糊特征子集和前一个滑动窗口中的优化模糊特征子集,自适应地跟踪输入模糊特征权重的变化,得到当前滑动窗口中的优化模糊特征子集,提高了模型的解释性并保证学习的平滑性和实效性;根据模糊输入特征与初始化优化模糊特征子集之间的重要度获取最终的优化模糊特征子集,同时考虑历史特征选择结果和当前数据分布,确保最终的优化模糊特征子集的优越性. 在 DFFS-VW 算法中,输出特征可以是类别标签,也可以是连续数据,无论是离线模糊特征选择,还是在线模糊特征选择,都无需依赖分类器或者回归模型,无需人为设置任何参数,提高了算法的自适应性。

2 相关定义

本文以滑动窗口分割模糊化数据为基础,从模糊特征互信息量入手,建立起模糊输入特征与输出特征之间的联系,进而实现动态模糊特征选择. 为了确定滑动窗口的大小,这里采用 Hoeffding 边界检测数据的分布。

定义 1. Hoeffding 边界^[21]. 对于取值范围为 $\mathbf{R}$ 的特征,假设每个滑动窗口包含 n 条数据样本,在置信程度是 $1-\delta$ (δ 一般取 0.05) 的条件下, Hoeffding 边界 ϵ 为

$$\epsilon = \sqrt{\frac{\mathbf{R} \cdot \mathbf{R}^T \ln(1/\delta)}{2n}}, \quad (1)$$

其中, $\mathbf{R}$ 为特征的取值范围:

$$\mathbf{R} = (x_{1, \max} - x_{1, \min}, x_{2, \max} - x_{2, \min}, \dots, x_{j, \max} - x_{j, \min}, \dots, x_{L, \max} - x_{L, \min}),$$

其中, $x_{j, \max}$ 是第 j 个变量的最大值, $x_{j, \min}$ 是第 j 个变量的最小值, L 是数据集变量的个数。

随着 n 的增加, Hoeffding 边界 ϵ 在变小, 这表明当 n 足够大, Hoeffding 边界 ϵ 将趋近于 0。

根据 Hoeffding 边界, 样本集的最小容量 N_H 可以确定:

$$N_H = \frac{\mathbf{R} \cdot \mathbf{R}^T \ln(1/\delta)}{2\epsilon^2}. \quad (2)$$

根据式(2)可知, Hoeffding 边界 ϵ 是获得样本集的最小容量 N_H 的关键. 假设相邻滑动窗口 W_{p-1} 和 W_p 所包含的样本集均值分别是 μ_{p-1} 和 μ_p , 在置信程度是 $1-\delta$ 的条件下, $|\mu_{p-1} - \mu_p| \leq 2\epsilon$, 则样本集的最小容量 N_H 的计算公式可以调整为

$$N_H = \frac{2\mathbf{R} \cdot \mathbf{R}^T \ln(1/\delta)}{(\mu_{p-1} - \mu_p)^2}. \quad (3)$$

由于对数值数据进行模糊化处理不仅有助于提取高层次的含义, 而且不需要对数据进行简单地“硬”划分, 它被广泛应用于智能系统中^[19]. 通常的模糊化方法是根据经验或者知识建立隶属度函数. 但是, 在多数情况下, 很难获取所需要的信息或者专家知识. 本文利用以前的研究基础^[22]对数据特征进行聚类实现特征模糊化. 设第 p ($p \geq 0$) 个滑动窗口包含 n 条数据, L 维输入特征和一维输出特征, 可以表示为

$$\mathbf{D}_p = \begin{bmatrix} x_p^{11} & \cdots & x_p^{1L} & y_p^1 \\ x_p^{12} & \cdots & x_p^{2L} & y_p^2 \\ \vdots & & \vdots & \vdots \\ x_p^{1n} & \cdots & x_p^{nL} & y_p^n \end{bmatrix}. \quad (4)$$

令输入特征和输出特征共同组成的特征向量为 $\mathbf{X}_p = (X_p^1, X_p^2, \dots, X_p^j, \dots, X_p^L, Y_p)$.

定义 2. 输入特征模糊化. 假设特征 $X_p^1, X_p^2, \dots, X_p^j, \dots, X_p^L$ 的聚类结果中分别包含 $|C_p^1|, |C_p^2|, \dots, |C_p^j|, \dots, |C_p^L|$ 个类别, 根据每个输入特征的聚类结果将输入特征向量 $\mathbf{X}_p^{\text{in}} = (X_p^1, X_p^2, \dots, X_p^j, \dots, X_p^L)$ 转化为输入模糊特征向量 $\mathbf{X}_p^{\text{fin}}$:

$$\mathbf{X}_p^{\text{fin}} = (\tilde{X}_p^{11}, \tilde{X}_p^{12}, \dots, \tilde{X}_p^{1|C_p^1|}, \tilde{X}_p^{21}, \tilde{X}_p^{22}, \dots, \tilde{X}_p^{2|C_p^2|}, \dots, \tilde{X}_p^{jq}, \dots, \tilde{X}_p^{L1}, \tilde{X}_p^{L2}, \dots, \tilde{X}_p^{L|C_p^L|}), \quad (5)$$

其中, $\tilde{X}_p^{jq}$ 是第 p 个滑动窗口中第 j ($1 \leq j \leq L$) 个特征的第 q ($1 \leq q \leq |C_p^j|$) 个模糊特征, 可以表示为

$$\tilde{X}_p^{jq} = \frac{\mu_{\tilde{X}_p^{jq}}(x_p^{1j})}{x_p^{1j}} + \frac{\mu_{\tilde{X}_p^{jq}}(x_p^{2j})}{x_p^{2j}} + \dots + \frac{\mu_{\tilde{X}_p^{jq}}(x_p^{ij})}{x_p^{ij}} + \dots + \frac{\mu_{\tilde{X}_p^{jq}}(x_p^{nj})}{x_p^{nj}}, \quad (6)$$

$$\mu_{\tilde{X}_p^{jq}}(x_p^{ij}) = \exp\left(-\frac{(x_p^{ij} - c_{X_p^j})^2}{2\sigma_{X_p^j}^2}\right), \quad (7)$$

$$1 \leq i \leq n, 1 \leq j \leq L, 1 \leq q \leq |C_p^j|,$$

其中, n 是数据集的样本个数, L 是数据集的输入维度, x_p^{ij} 是第 p 个滑动窗口中第 i 个样本的第 j 个输入特征值, $c_{X_p^j}$ 是第 p 个滑动窗口中第 j 个输入特征的均值, $\sigma_{X_p^j}^2$ 是第 p 个滑动窗口中第 j 个输入特征的方差。

为了度量模糊特征之间的相关性, 有必要引入新的度量。

定义 3. 模糊信息熵. 设 f_p^s 是第 p 个滑动窗口的模糊特征子集, f_p^s 的模糊信息熵定义如下:

$$H(f_p^s) = -\frac{1}{n} \sum_{i=1}^n \lg \frac{B_p^{is}}{n}, \quad (8)$$

其中, B_p^{is} 是特征子集 f_p^s 相对于第 i 条样本的基:

$$B_p^{is} = \sum_{k=1}^n \frac{1}{|f_p^s|} \sum_{j=1}^L \sum_{q=1}^{|C_p^j|} (1 - |\mu_{\tilde{X}_p^{jq}}(x_p^{ij}) - \mu_{\tilde{X}_p^{jq}}(x_p^{kj})|), \quad \tilde{X}_p^{jq} \in f_p^s, k \neq i, \quad (9)$$

其中, $|f_p^s|$ 是 f_p^s 中包含的模糊特征个数, $\tilde{X}_p^{jq}$ 是第 p 个滑动窗口中第 j 个特征的第 q 个模糊特征, $\mu_{\tilde{X}_p^{jq}}(x_p^{ij})$ 是第 p 个滑动窗口中第 i 条数据的第 j 个特征值对于模糊特征 $\tilde{X}_p^{jq}$ 的隶属度值, $\mu_{\tilde{X}_p^{jq}}(x_p^{kj})$ 是第 p 个滑动窗口中第 k 条数据的第 j 个特征值对于模糊特征 $\tilde{X}_p^{jq}$ 的隶属度值。

定义 4. 联合模糊信息熵. 设 f_p^s 是第 p 个滑动窗口的模糊特征子集, $\tilde{X}_p^{jq}$ 是第 p 个滑动窗口中第 j 个特征的第 q 个模糊特征, f_p^s 和 $\tilde{X}_p^{jq}$ 的联合模糊信息熵:

$$H(\tilde{X}_p^{jq}, f_p^s) = -\frac{1}{n} \sum_{i=1}^n \lg \frac{|B_{\tilde{X}_p^{jq}}^{is} \cap B_{f_p^s}^{is}|}{n}, \quad \tilde{X}_p^{jq} \notin f_p^s, \quad (10)$$

其中, $B_{\tilde{X}_p^{jq}}^{is}$ 是模糊特征 $\tilde{X}_p^{jq}$ 相对于第 i 条样本的基:

$$B_{\tilde{X}_p^{jq}}^{is} = \sum_{k=1}^n (1 - |\mu_{\tilde{X}_p^{jq}}(x_p^{ij}) - \mu_{\tilde{X}_p^{jq}}(x_p^{kj})|). \quad (11)$$

3 基于特征变权的动态模糊特征选择算法 (DFFS-VW)

3.1 DFFS-VW 算法框架

为了提高模糊动态特征选择的实效性和快速性, 降低算法复杂性, 发现各个模糊输入特征权重的演化规律. 本文提出了 DFFS-VW 算法, 该算法的基本框架如图 1 所示, 首先利用滑动窗口 $W_0, W_1, \dots, W_p$ 分割模糊数据, 以第 1 个滑动窗口中的数据作为初始数据进行离线模糊特征选择, 在随后依次到来的滑动窗口中进行在线模糊特征选择. 在离线模糊特征选择中, 计算每个模糊输入特征的权重, 根据权重得到候选模糊特征子集; 结合后向特征选择算法和模糊输入特征筛选指标迭代地筛选候选模糊特征子集, 每一步都忽略其中权重值最小的模糊特征, 得到优化模糊特征子集. 在在线模糊特性选择中, 根据当前滑动窗口中的数据获取候选模糊特征子集, 利用上一个滑动窗口的模糊优化特征子集与当前窗口的候选模糊特征子集的交集作为当前快照的初始优化模糊特征子集, 计算其余的模糊输入特征与初始优化模糊特征子集的重要度, 若重要度大于 0, 将模糊输入特征加入优化模糊特征子集. 最后, 根据各个模糊输入特征的权重对模糊输入特征进行演化分析。

3.2 滑动窗口的确定

在动态环境中, 滑动窗口是一种被广泛应用的技术. 每个滑动窗口中包含着最新产生的样本, 并丢弃过去的样本. 目前主要有 2 种确定滑动窗口的方法: 1) 使用固定的窗口大小; 2) 根据衰减因子确定滑动窗口中丢弃的数据和保留下来的数据, 利用新数据代替已丢弃的数据. 这 2 种方法都需要人为设置参数, 然而, 这些参数都很敏感, 不易确定. 根据数据分布直接确定窗口大小, 可以解决上述问题. Hoeffding 边界考虑了特征的变化范围, 不断地调整滑动窗口包含的样本个数 n , 使得通过滑动窗口中的数据求得的最小容量 N_H 不大于实际滑动窗口样本个数 n , 将 n 作为固定的滑动窗口大小. 在动态特征选择过程中, 随着信息不断的采集输入, 滑动窗口的任务是以有效的方式管理输入数据, 存储信息. 相邻滑动窗口记录将被用来进行演化过程中特征之间的对等比较, 获取各个输入模糊特征对输出特征的权重演化规律。

表示模糊输入特征对输出特征的影响急剧减小. 我们可以将相邻模糊权重与相邻模糊权重均值和相邻模糊权重标准差的和作比较, 自适应地发现模糊输入特征对输出特征的影响急剧减小的截断点, 将相邻模糊权重均值和相邻模糊权重标准差的和称作是截断点阈值. 根据第 p 个滑动窗口中前 $d-1$ 个权重梯度, 计算第 p 个滑动窗口中第 d 个权重梯度的截断点阈值 ρ_p^d :

$$\rho_p^d = \frac{\sum_{d1=1}^{d-1} \Delta_p^{d1}}{d-1} + \sqrt{\sum_{d1=1}^{d-1} \left(\Delta_p^{d1} - \frac{\sum_{d2=1}^{d-1} \Delta_p^{d2}}{d-1} \right)^2}, d \geq 3. \quad (18)$$

当 $\Delta_p^d > \rho_p^d$ 时, 排序后的第 $d+1$ 个特征是截断点.

无论是模糊特征权重, 还是相邻模糊特征权重梯度, 都只能衡量单个模糊特征的好坏, 特征选择是要保证所选的输入特征子集整体最优. 第 p 个滑动窗口中第 j 个特征的第 q 个模糊特征 $\tilde{X}_p^{jq}$ 和第 u 个特征的第 v 个模糊特征 $\tilde{X}_p^{uv}$ 之间的互信息量:

$$I(\tilde{X}_p^{jq}, \tilde{X}_p^{uv}) = H(\tilde{X}_p^{jq}) + H(\tilde{X}_p^{uv}) - H(\tilde{X}_p^{jq}, \tilde{X}_p^{uv}), \quad (19)$$

其中, $H(\tilde{X}_p^{jq})$ 和 $H(\tilde{X}_p^{uv})$ 分别是第 p 个滑动窗口的模糊输入特征 $\tilde{X}_p^{jq}$ 和 $\tilde{X}_p^{uv}$ 的模糊信息熵, $H(\tilde{X}_p^{jq}, \tilde{X}_p^{uv})$ 是 $\tilde{X}_p^{jq}$ 和 $\tilde{X}_p^{uv}$ 的模糊联合信息熵. 单个的模糊特征是模糊特征子集的特殊形式, 所以可根据定义 3 和定义 4 计算模糊特征的模糊信息熵和模糊联合信息熵.

为此, 我们将模糊输入特征筛选指标 FCI 作为衡量动态模糊特征选择的衡量标准. 它不仅考虑模糊特征子集的输入特征与输出特征之间的互信息量, 还考虑模糊输入特征与模糊输入特征之间的互信息量:

$$FCI(f_p^s) = \frac{|f_p^s| |CF_p|}{\sqrt{|f_p^s| + |f_p^s| (|f_p^s| - 1) FF_p}}, \quad (20)$$

其中, $|f_p^s|$ 是第 p 个滑动窗口中模糊特征集合 f_p^s 包含的模糊特征个数, CF_p 是第 p 个滑动窗口中 f_p^s 包含的各个模糊特征与输出特征之间互信息量的均值, FF_p 是第 p 个滑动窗口中 f_p^s 包含的两两模糊特征的互信息量的均值:

$$CF_p = \frac{\sum_{j=1}^L \sum_{q=1}^{|C_p^j|} I(\tilde{X}_p^{jq}, Y_p)}{|f_p^s|}, \tilde{X}_p^{jq} \in f_p^s, \quad (21)$$

$$FF_p = \frac{\sum_{j,u=1, j \neq u}^L \sum_{q,v=1, q \neq v}^{|C_p^{j,u}|} I(\tilde{X}_p^{jq}, \tilde{X}_p^{uv})}{|f_p^s| (|f_p^s| - 1)}, \quad (22)$$

$$\tilde{X}_p^{jq} \in f_p^s, \tilde{X}_p^{uv} \in f_p^s.$$

3.4 离线模糊特征选择

第 1 个滑动窗口中的数据作为初始数据进行离线模糊特征选择, 主要分为 2 部分: 1) 根据输入模糊特征和输出特征的互信息量计算各个输入模糊特征的权重, 将模糊输入特征根据权重由大到小排序, 获取候选模糊特征集; 2) 在候选模糊特征集中, 根据模糊特征筛选指标同时考虑输入模糊特征和输出特征的互信息量与输入模糊特征之间的互信息量, 利用后向特征选择和模糊输入特征筛选指标得到与输出特征相关性最大、冗余度最小的优化模糊特征子集.

算法 1. 离线模糊特征选择算法.

输入: 所有模糊特征;

输出: 最优模糊特征子集.

Step1. 初始化. 初始化筛选后的模糊特征子集 $f_p^s = \emptyset (p=0)$.

Step2. 获取候选模糊特征子集 f_p^{cs} .

Step2. 1. 分别计算各个模糊特征与输出类别的互信息量.

Step2. 2. 根据互信息量计算各个模糊特征的权重, 并根据权重从大到小排列模糊输入特征 $X_p^{ofin} = \{Fa_p^1, Fa_p^2, \dots, Fa_p^d, \dots, Fa_p^m\}$, 依次计算排序后相邻权重的梯度.

Step2. 3. 初始化候选模糊特征集 $f_p^{cs} = \{Fa_p^1, Fa_p^2, Fa_p^3\}$.

Step2. 4. 确定候选模糊特征集 f_p^{cs} . 如果 $\Delta_p^d > \rho_p^d (d \geq 3)$, 则选取排序后的模糊输入特征中的前 d 个特征作为候选模糊特征集 f_p^{cs} .

Step3. 获取最优模糊特征子集 f_p^s .

Step3. 1. 计算候选模糊特征集 f_p^{cs} 的综合指标 $FCI(f_p^{cs})$.

Step3. 2. 按照 f_p^{cs} 中模糊特征权重由小到大的顺序.

Step3. 3. 依次不考虑排序后模糊特征权重对应的模糊特征 $Fa_p^d (1 \leq d \leq |f_p^{cs}|)$ 时, 计算模糊特征子集 $f_p^{cs}(\overline{Fa_p^d}) = \{Fa_p^1, Fa_p^2, \dots, Fa_p^{d-1}, Fa_p^{d+1}, \dots, Fa_p^{|f_p^{cs}|}\}$ 的综合指标 $FCI(f_p^{cs}(\overline{Fa_p^d}))$.

Step3. 4. 如果 $FCI(f_p^{cs}(\overline{Fa_p^d})) > FCI(f_p^{cs})$, 则 $f_p^{cs} = f_p^{cs}(\overline{Fa_p^d})$, 更新综合评价指标 $FCI(f_p^{cs})$, 跳到 Step3. 5.

Step3. 5. 循环执行 Step3. 2 ~ Step3. 4, 直到 Step3. 4 中所产生的模糊特征子集的综合指标都不大于 $FCI(f_p^{cs})$, 则不再循环.

Step4. 最优模糊特征子集 $f_p^s = f_p^{cs}$.

3.5 在线模糊特征选择

为了提高算法的效率,在线模糊特征选择首先利用第 $p-1$ 个滑动窗口中的 $|C_{p-1}|$ 个聚类中心 μ_{p-1}^q 和第 p 个滑动窗口中的所有数据组成 D_p , 其中 $p \geq 1, 1 \leq q \leq |C_{p-1}|$; 然后在此数据集中, 根据权重的相邻梯度获取候选模糊特征集; 最后以第 $p-1$ 个滑动窗口中的优化模糊特征子集 f_{p-1}^s 与候选模糊特征集的交集作为当前滑动窗口的初始优化模糊特征子集, 根据重要度获得第 p 个滑动窗口中的优化模糊特征子集。

算法 2. 在线模糊特征选择算法。

输入: 前一个滑动窗口的聚类结果和当前滑动窗口的所有数据;

输出: 优化模糊特征子集。

Step1. 构建 D_p . 当 $1 \leq q \leq |C_{p-1}|$ 时, 利用第 $p-1$ 个滑动窗口中的 $|C_{p-1}|$ 个聚类中心 μ_{p-1}^q 和第 p 个滑动窗口中的所有数据组成 D_p .

Step2. 获取候选模糊特征集 f_p^{cs} .

Step2.1. 输入输出互信息量的计算. 根据第 p 个滑动窗口中的所有数据计算模糊输入特征 $\tilde{X}_p^{jq}$ 与输出特征 Y_p 的输入输出互信息量 $I(\tilde{X}_p^{jq}, Y_p)$.

Step2.2. 根据输入输出互信息量计算各个模糊特征的权重, 并根据权重从大到小排列模糊输入特征 $X_p^{\text{fin}} = \{Fa_p^1, Fa_p^2, \dots, Fa_p^d, \dots, Fa_p^m\}$, 依次计算排序后相邻权重的梯度。

Step2.3. 初始化候选模糊特征集 $f_p^s = \{Fa_p^1, Fa_p^2, Fa_p^3\}$.

Step2.4. 确定候选模糊特征集 f_p^{cs} . 当 $d \geq 3$ 时, 如果 $\Delta_p^d > \rho_p^d$, 则选取排序后的模糊输入特征中的前 d 个特征作为候选模糊特征集 f_p^{cs} .

Step3. 初始第 p 个快照中优化模糊特征子集 $f_p^s = f_{p-1}^s \cap f_p^{cs}$.

Step4. 计算模糊特征 $Fa_p^d (Fa_p^d \notin f_p^s, Fa_p^d \in f_p^{cs})$ 在 f_p^{cs} 中的重要度:

$$SIG(Fa_p^d, f_p^s) = H(Fa_p^d, f_p^s) - H(f_p^s). \quad (23)$$

Step5. 选择具有最大重要度的模糊特征. 选择满足下面条件的属性:

$$H(Fa_p | f_p^s) = \max_d (SIG(Fa_p^d, f_p^s)). \quad (24)$$

Step6. 获取优化模糊特征子集 f_p^s . 如果 $H(Fa_p | f_p^s) > 0$, 则 $f_p^s = f_p^s \cup Fa_p$, 并且返回 Step4; 否则, 直接输出 f_p^s .

3.6 DFFS-VW 算法计算复杂度分析

DFFS-VW 算法的时间复杂度取决于模糊特征个数. 假设 m 是所有模糊特征个数, DFFS-VW 算法包括离线模糊特征选择和在线模糊特征选择. 离

线模糊特征选择初始化首先计算各个模糊特征的互信息量, 所需要的计算量是 $O(m)$, 根据互信息量计算各个模糊特征的权重, 所需要的计算量是 $O(m)$; 然后根据权重从大到小对模糊特征进行排序, 所需要的计算量是 $O(m)$, 依次计算排序后相邻模糊特征权重的梯度 Δ_p^d , 其中 $d \geq 3$, 如果 Δ_p^d 大于截断点阈值, 则选取排序后的模糊特征中的前 d 个特征作为候选模糊特征集 f_p^{cs} . 最坏情况下, 候选模糊特征集 f_p^{cs} 包含的模糊特征个数是 m , 获取候选模糊特征集 f_p^{cs} 所需要的计算量是 $O(m)$. 最好情况下, 候选模糊特征集 f_p^{cs} 包含的模糊特征个数是 3, 因此, 获取候选模糊特征集 f_p^{cs} 所需要的计算量是 $O(3)$. 在此基础上, 根据综合评价指标不断更新 f_p^{cs} , 直到综合评价指标 $FCI(f_p^{cs})$ 不再增加. 最坏情况下, f_p^{cs} 所需要的更新次数是 $m-3$, 所以得到最终的优化模糊特征子集所需要的计算量是 $O(m-3)$. 综上所述, 无论是最坏情况还是最好情况, 该算法的复杂度是 $O(4m)$, 经化简算法的最终复杂度是 $O(m)$.

在线模糊特征选择获取当前窗口的候选模糊特征集的过程与离线模糊特征选择的相似, 最坏情况下所需要的时间复杂度是 $O(m+m+m+m)$; 最好情况下所需要的时间复杂度是 $O(m+m+m+3)$. 在此基础上, 依次计算其余的模糊特征相对于候选模糊特征集的重要度, 直到其重要度不大于 0, 获得最终的优化模糊特征子集, 最坏情况下, 所需要的计算复杂度数 $O(m-3)$. 综上所述, 该算法的复杂度是 $O(4m)$, 经化简算法的最终复杂度是 $O(m)$.

4 实验结果

4.1 实验环境

本文将 DFFS-VW 算法与基于邻域粗糙集的动态特征选择算法 (incremental feature selection based on fuzzy neighborhood rough set, IFSFNRS)^[13]、基于贝叶斯分类器的封装式动态特征选择算法 (wrapper feature selection algorithm with incremental Bayesian classifier, WFSIBC)^[11]、动态无监督特征选择算法 (online unsupervised multi-view feature selection, OMVFS)^[20]、改进的动态无监督特征选择 (feature selection on data streams, FSDS)^[20] 对比, 选取 1 个人工合成数据集、UCI 数据库中的 10 个数据集^[23]、2 个分类数据流^[24] 和 2 个回归数据流作为实验数据进行分析, 以测试 DFFS-VW 算法在无需设置任何参数的情况下, 算法执行效率、自适应和预测

准确性方面均得到提高. 对于所有数据, 各个维度均进行输入特征模糊化处理, 使得数据集中所有数据值均在 $[0, 1]$ 区间. 所有实验都在 Inter[®] CPU 2.30 GHz 工作台上运行, 利用 matlab 2014. a 软件进行仿真. 数据集的详细信息如表 1 所示:

Table 1 Description of Datasets
表 1 数据集描述

Dataset	# Pattern	Dimensionality
SD	2 000	2
BCH	1 372	4
Glass	214	9
Liver	341	6
Wine	178	13
Yeast	1484	8
SEA	50 000	3
Weather	18 159	8
Airfoil Self-Noise	1 503	5
Concrete_Data	1 030	8
Energy Efficiency	768	8
Red Wine Quality	1 599	11
White Wine Quality	4 898	11
Beijing PM2.5 Data	43 800	6
PPPTS	45 730	9

4.2 评价指标

本文采用 2 种评价指标对算法的动态特征选择效果进行评价: 分类准确率(Acc)和平均绝对百分比误差($MAPE$).

$$Acc = \frac{1}{n} \sum_{k=1}^{|C|} a_k,$$

(25)

其中, a_k 代表第 k 个类别中算法的分类结果和实际数据集分类情况相一致的数据个数, $|C|$ 是类别个数, n 是数据集中包含的数据个数.

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|\hat{y}_i - y_i|}{y_i},$$

(26)

其中, y_i 是第 i 个样本的实际目标值, $\hat{y}_i$ 是第 i 个样本的模型预测值.

4.3 滑动窗口对特征选择结果的影响测试

为了分析滑动窗口大小对 DFFS-VW 算法模糊特征选择结果的影响, 本文利用不同数据集形成大小各异的滑动窗口, 以数据集 70% 作为训练集, 30% 作为测试集, 分析比较模糊特征选择的性能. 表 2 分别对比了 CMAR(classification based on multiple association rules) 分类器^[25]、C4.5 分类器^[26]和 DFFS-VW 算法对不同大小滑动窗口的模糊特征选择结果的性能, 其中 CMAR 中的最小支持度设置为 10%, 最小置信度为 70%, C4.5 选取信息增益最大的作为划分节点. 从表 2 可以看出, 本文所提方法自动确定窗口大小的分类性能优于人为定义窗口大小的分类性能. 表 3 分别对比了 BP 神经网络^[27]、支持向量回归(support vector regression, SVR)^[28]和 DFFS-VW 算法对不同大小滑动窗口的模糊特征选择结果的预测性能, 其中 BP 神经网络有 1 层隐含层, 隐含层中包含 10 个节点. SVR 的核函数为径向基函数, 不敏感系数为 0.05, 惩罚因子为 10. 从表 3 可以看出, 本文所提出方法的预测结果优于人为定义窗口大小的预测结果. 这是由于本文所提出的窗口大小确定方法根据 Hoeffding 边界考虑了数据的分布情况, 提高了分割后的模糊数据的分类性能和预测性能.

Tabel 2 The Comparison of the Classification Accuracy(Acc) for CMAR/C4.5 on Different Window Sizes
表 2 不同窗口大小的 CMAR/C4.5 分类性能(Acc)对比

Dataset	The Size of Window				The Proposed Method	
	5%		10%			
	CMAR	C4.5	CMAR	C4.5	CMAR	C4.5
SD	0.875	0.881	0.864	0.875	0.906	0.912
BCH	0.883	0.896	0.874	0.896	0.947	0.958
Glass	0.904	0.912	0.891	0.901	0.962	0.967
Liver	0.892	0.881	0.874	0.869	0.952	0.948
Wine	0.903	0.921	0.886	0.897	0.957	0.969
Yeast	0.911	0.916	0.895	0.903	0.963	0.974
SEA	0.921	0.934	0.904	0.912	0.983	0.989
Weather	0.878	0.896	0.858	0.869	0.948	0.957
Mean	0.896	0.905	0.881	0.890	0.952	0.959

Tabel 3 The Comparison of the Prediction Accuracy(MEAP) for BP/SVR on Different Window Size
表 3 不同窗口大小的 BP/SVR 预测性能(MEAP)对比

Dataset	The Size of Window				The Proposed Method	
	5%		10%			
	CMAR	C4.5	CMAR	C4.5	CMAR	C4.5
Airfoil Self-Noise	0.194	0.185	0.185	0.176	0.141	0.042
Concrete_Data	0.149	0.133	0.137	0.131	0.116	0.053
Energy Efficiency	0.224	0.216	0.216	0.244	0.181	0.057
Red Wine Quality	0.174	0.161	0.168	0.159	0.124	0.053
White Wine Quality	0.153	0.142	0.144	0.136	0.104	0.031
Beijing PM2.5 Data	0.254	0.243	0.241	0.246	0.173	0.026
PPPTS	0.254	0.249	0.236	0.219	0.194	0.057
Mean	0.197	0.190	0.190	0.184	0.148	0.043

4.4 特征选择效果测试

为了充分验证 DFFS-VW 算法的有效性,对比 WFSIBC 算法、IFSFNRS 算法、OMVFS 算法和 FSDS 算法在多个数据集下的特征选择效果.图 2 给出 5 种特征选择算法的模糊特征选择结果与原始模糊特征个数的对比结果.从图 2 可以看出,除数据集 SD,经过 5 种动态特征选择算法后,所选模糊特

征个数都比原始模糊特征个数少,算法 DFFS-VW 最为明显.DFFS-VW 算法所选的模糊特征个数均比其余 4 种算法的少,这是因为 DFFS-VW 算法根据模糊权重获取候选模糊特征子集,同时考虑输入模糊特征与输出特征的互信息量与输入模糊特征之间的互信息量得到最优模糊特征子集,使得 DFFS-VW 算法的特征选择效果更加明显.

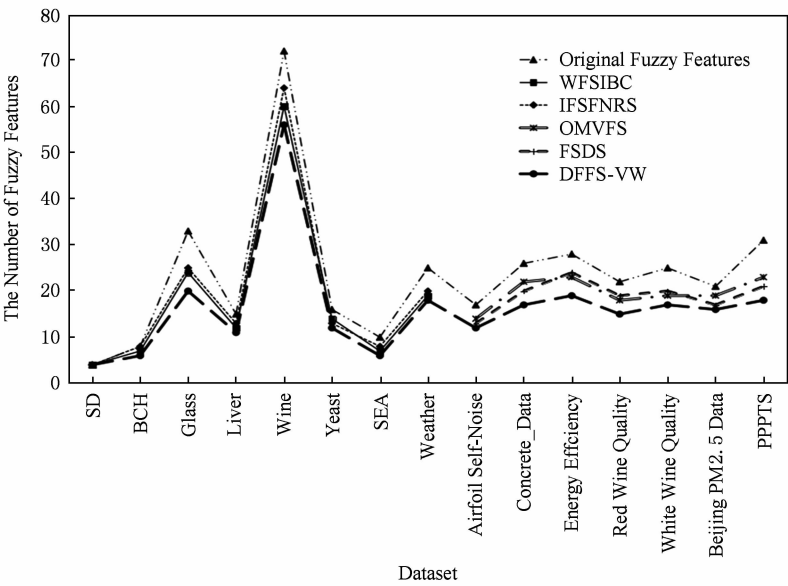


Fig. 2 Fuzzy features comparison of different algorithms on different datasets
图 2 模糊特征个数对比

表 4 给出了 CMAR 分类器和 C4.5 分类器对原始模糊特征集和 3 种动态特征选择算法的所选模糊特征集的分类准确性的对比.从表 4 可以看出,除了 SD 数据集,DFFS-VW 算法所选的模糊特征子集对于 CMAR 分类器的分类效果最好;经特征选择后,C4.5 分类器对于各个数据集的分类准确率均得

到了提高,其中 DFFS-VW 算法所选的模糊特征子集提高得最明显.

表 5 给出了 BP 神经网络和支持向量回归对原始模糊特征集和 3 种动态特征选择算法的所选模糊特征集的预测精确性的对比.从表 5 可以看出,除 SD 数据集外,BP 神经网络对 DFFS-VW 算法所选

的模糊特征子集的预测精确性最好;经模糊特征选择后,支持向量回归对于各个数据集的预测精确性

均得到了提高,其中 DFFS-VW 算法所选的模糊特征子集提高得最明显.

Tabel 4 The Comparison of the Classification Accuracy(Acc)
表 4 CMAR/C4.5 分类性能(Acc)对比

Dataset	Original		WFSIBC		IFSFNRS		DFFS-VW	
	CMAR	C4.5	CMAR	C4.5	CMAR	C4.5	CMAR	C4.5
SD	0.906	0.912	0.906	0.912	0.906	0.912	0.906	0.912
BCH	0.921	0.925	0.940	0.945	0.936	0.938	0.947	0.958
Glass	0.932	0.906	0.951	0.958	0.945	0.949	0.962	0.967
Liver	0.916	0.896	0.932	0.933	0.928	0.931	0.952	0.948
Wine	0.925	0.913	0.939	0.936	0.938	0.932	0.957	0.969
Yeast	0.915	0.924	0.943	0.946	0.928	0.935	0.963	0.974
SEA	0.923	0.909	0.956	0.961	0.944	0.925	0.983	0.989
Weather	0.919	0.918	0.945	0.943	0.936	0.928	0.948	0.957
Mean	0.920	0.913	0.939	0.942	0.933	0.931	0.952	0.959

Tabel 5 The Comparison of the Prediction Accuracy (MEAP)
表 5 BP/SVR 预测性能(MEAP)对比

Dataset	Original		OMVFS		FSDS		DFFS-VW	
	CMAR	C4.5	CMAR	C4.5	CMAR	C4.5	CMAR	C4.5
Airfoil Self-Noise	0.248	0.075	0.183	0.055	0.209	0.062	0.141	0.042
Concrete_Data	0.224	0.094	0.194	0.042	0.217	0.051	0.116	0.033
Energy Efficiency	0.235	0.096	0.204	0.065	0.218	0.072	0.181	0.057
Red Wine Quality	0.232	0.104	0.181	0.067	0.205	0.069	0.124	0.052
White Wine Quality	0.206	0.087	0.192	0.064	0.186	0.068	0.104	0.031
Beijing PM2.5 Data	0.311	0.076	0.254	0.054	0.264	0.065	0.173	0.026
PPPTS	0.341	0.106	0.272	0.074	0.295	0.086	0.194	0.057
Mean	0.257	0.092	0.211	0.061	0.228	0.068	0.148	0.043

如图 3 所示,本文采用不同的数据集对 5 种不同算法的运行时间进行了对比,可以看出,针对同一个数据集,DFFS-VW 算法的运行时间最短.由于 FSDS 算法利用回归模型提高所选模糊子集的预测正确率,在回归模型寻参期间需要不断迭代,时间复杂度大约为 $O(n^3)$;WFSIBC 算法使用贪婪算法发现最优特征子集,而贪婪算法的时间复杂度是 $O(n \log n)$,当数据量较大时,算法复杂度也会大幅度地增加;IFSFNRS 算法以粗糙邻域为基础,在最坏情况下,需要对整个数据集进行重新特征选择,时间复杂度不低于 $O(\log n^2)$;OMVFS 算法将特征选择嵌入到聚类算法中,使得时间复杂度近似为 $O(kn_l nL|C|)$,其中 k 是迭代至收敛的平均迭代次

数, n_l 是属于第 l 个聚类的样本个数, L 是输入特征个数, $|C|$ 是聚类个数,由此可见 OMVFS 算法的时间复杂度也很大. DFFS-VW 算法利用权重梯度选取候选模糊特征子集,减少了最优模糊特征子集的寻找范围,从而减少了算法的运行时间,时间复杂度为 $O(m)$,其中 m 是模糊特征个数.其中,DFFS-VW 算法对 SD 数据集的运行时间最长,对 Wine 数据集的运行时间最短.从图 3 可以得出,原始模糊特征个数越多,算法的运行时间越长.

4.5 模糊特征变权分析

图 4 是数据集 SEA 部分模糊特征在不同滑动窗口中权重的变化,可以看出,滑动窗口 $W_0 \sim W_4$ 分别包括 10 000 条样本,随着滑动窗口中数据的不

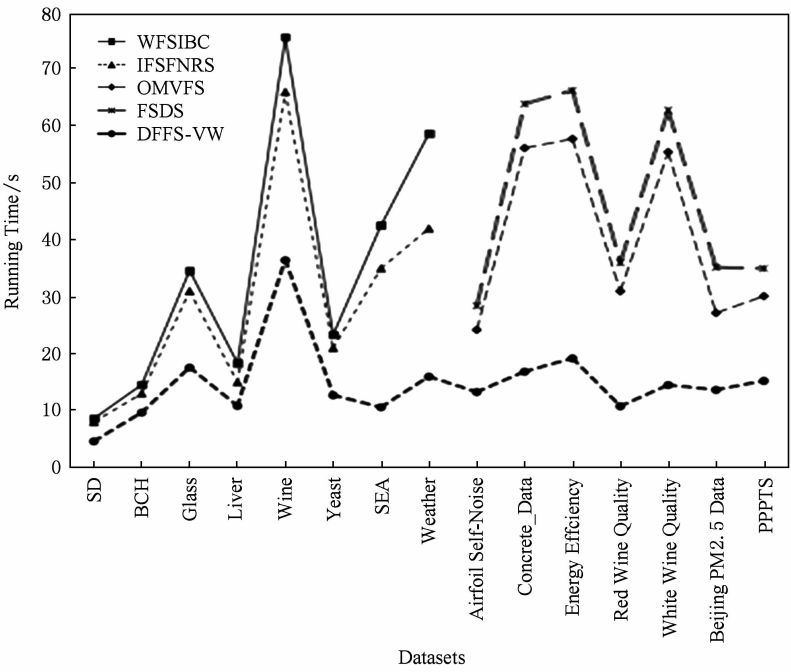


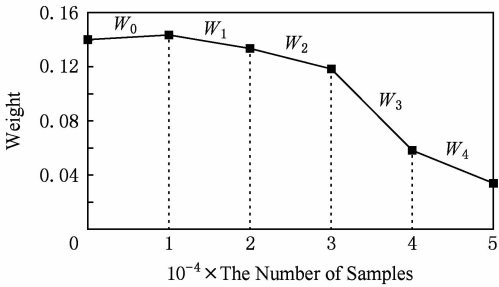
Fig. 3 Comparison of the running time for feature selection with different algorithms on different datasets

图 3 不同数据集上不同算法特征选择运行时间的比较

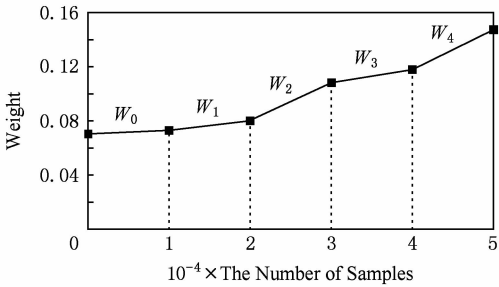
断变化,所选的模糊特征在不断地发生变化. 其中,模糊特征 $\tilde{X}_{11}$ 的权重在不断减少,在滑动窗口 W_0 中,模糊特征的权重达到了最小,对输出特征几乎没有影响;模糊特征 $\tilde{X}_{13}$ 的权重在不断增加,在滑动窗口 W_4 中,模糊特征的权重达到了最大;模糊特征 $\tilde{X}_{23}$, $\tilde{X}_{33}$ 的权重先减小后增大,滑动窗口 W_0 和滑动

窗口 W_4 中模糊特征 $\tilde{X}_{34}$ 的权重几乎相等.

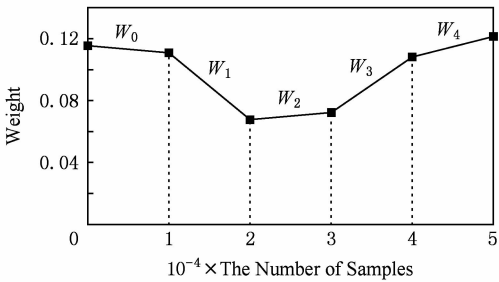
为了分析权重变化对分类准确率的影响,以数据集 SEA 为例,滑动窗口 $W_0 \sim W_4$ 分别包括 10 000 条样本,按照 $\tilde{X}_{11}$, $\tilde{X}_{13}$, $\tilde{X}_{23}$, $\tilde{X}_{33}$ 的顺序,依次将模糊输入特征从原始模糊输入特征集中忽略,对比忽略前后的分类准确率,结果如图 5 所示. 其中,由于模



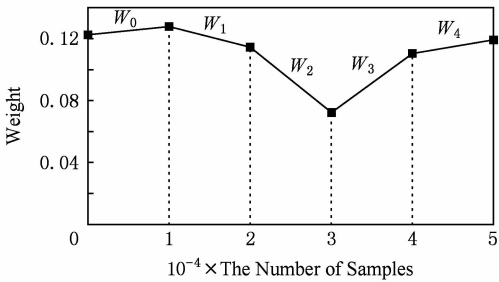
(a) Fuzzy feature $\tilde{X}_{11}$



(b) Fuzzy feature $\tilde{X}_{13}$



(c) Fuzzy feature $\tilde{X}_{23}$



(d) Fuzzy feature $\tilde{X}_{33}$

Fig. 4 The variable weight of some fuzzy features on the dataset SEA

图 4 数据集 SEA 部分模糊特征变权

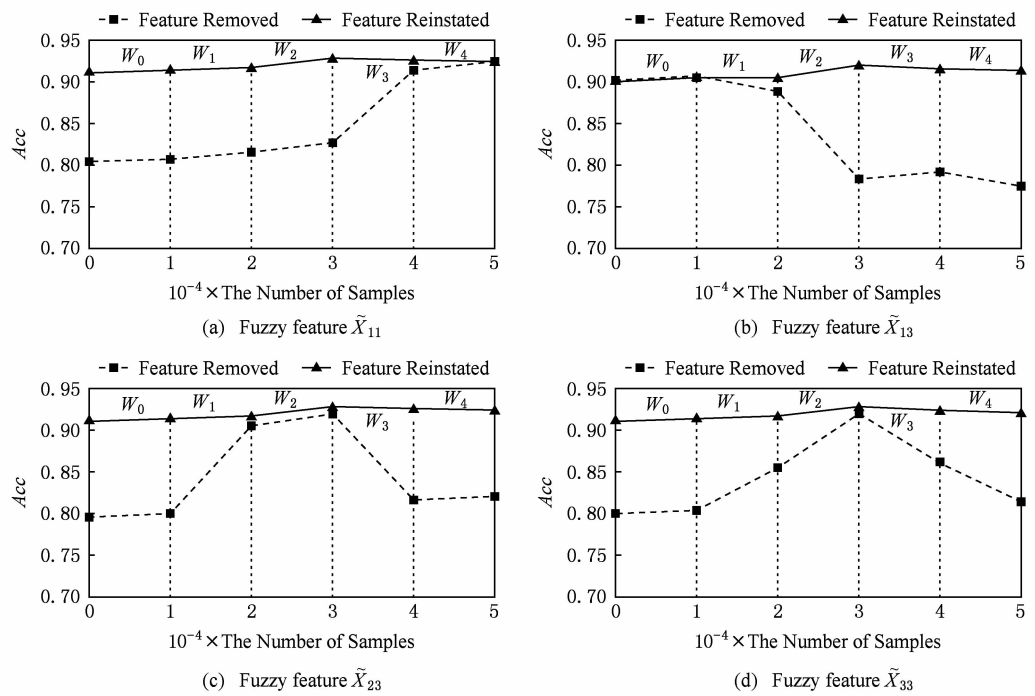


Fig. 5 The effect on classification accuracy of ignoring some fuzzy features on dataset SEA

图 5 数据集 SEA 忽略部分模糊特征前后分类准确率

糊输入特征 $\tilde{X}_{11}$ 的权重在不断减少,所以输入特征 $\tilde{X}_{11}$ 对分类准确性影响越来越小,在滑动窗口 W_0 中,忽略模糊特征前和忽略模糊特征后的分类准确率几乎相等;模糊特征 $\tilde{X}_{13}$ 的权重在不断增加,模糊特征 $\tilde{X}_{13}$ 对分类准确性影响越来越大,在滑动窗口 W_0 中,忽略模糊特征前和忽略模糊特征后的分类

准确率相差最大;模糊特征 $\tilde{X}_{23}, \tilde{X}_{33}$ 的权重先减小后增大,模糊特征 $\tilde{X}_{23}, \tilde{X}_{33}$ 对分类准确率的影响先增大后减小。

为了分析权重变化对预测精确性的影响,以数据集 PPPTS 为例,滑动窗口 $W_0 \sim W_7$ 分别包括 6 000 条样本,取其中的 4 个模糊输入特征 $\tilde{X}_{12},$

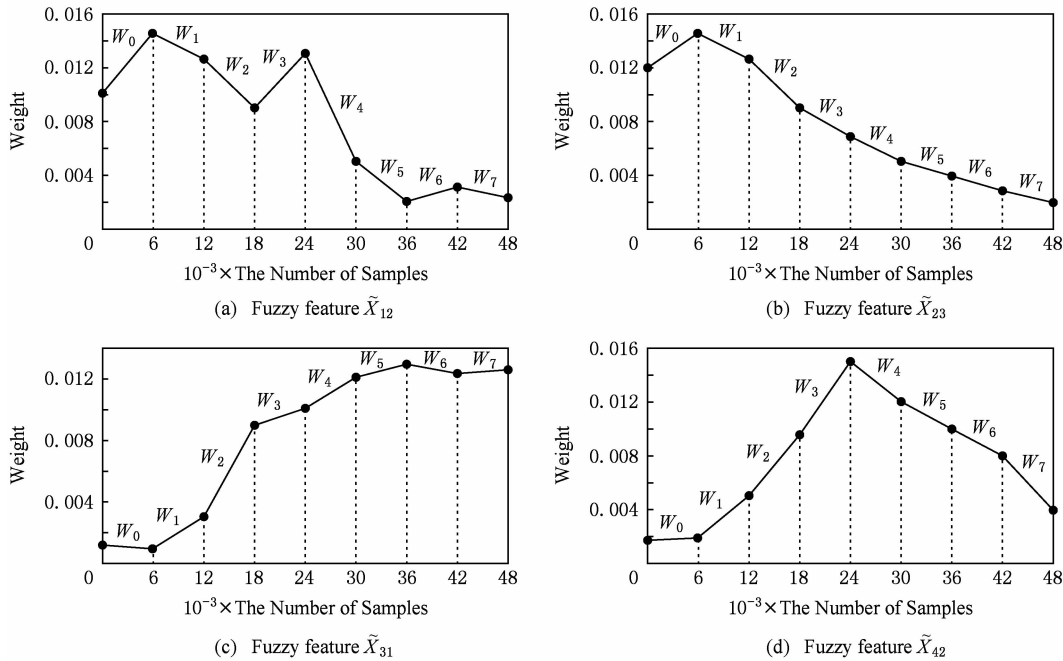


Fig. 6 The effect on prediction accuracy of ignoring some fuzzy features on dataset PPPTS

图 6 数据集 PPPTS 部分模糊特征权重的演化

$\tilde{X}_{23}, \tilde{X}_{31}, \tilde{X}_{42}$, 各个模糊输入特征的权重在不同滑动窗口的变化如图 6 所示. 从图 6 中可以看出, 随着数据的不断更新, 模糊输入特征 $\tilde{X}_{12}, \tilde{X}_{23}$ 的权重呈下

模糊特征子集, 根据模糊输入特征相对于初始优化模糊特征子集的重要度, 获得最终的优化模糊特征子集. 通过不同滑动窗口模糊特征权重的演化趋势,

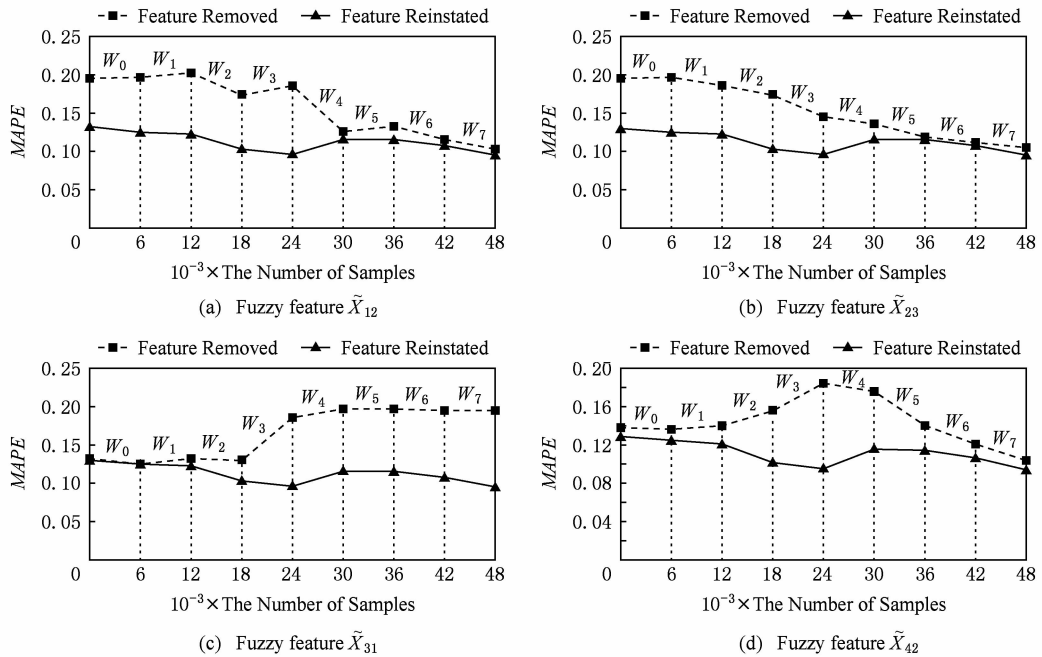


Fig. 7 The effect on prediction accuracy of ignoring some fuzzy features on dataset PPPTS

图 7 数据集 PPPTS 忽略部分模糊特征前后预测精确率对比

降趋势, 模糊输入特征 $\tilde{X}_{31}$ 的权重呈上升趋势, 模糊输入特征 $\tilde{X}_{42}$ 的权重先上升后下降. 依次将以上模糊输入特征从原始模糊输入忽略, 对比忽略前后的预测准确率, 结果如图 7 所示. 其中, 由于模糊输入特征 $\tilde{X}_{12}, \tilde{X}_{23}$ 的权重在不断减少, 所以输入模糊特征 $\tilde{X}_{12}, \tilde{X}_{23}$ 对预测精确率影响越来越小, 在滑动窗口 W_7 中, 忽略模糊特征前后的预测精确率几乎相等; 模糊输入特征 $\tilde{X}_{31}$ 的权重在不断增加, 模糊特征 $\tilde{X}_{31}$ 对预测精确率影响越来越大, 忽略模糊特征前和忽略模糊特征后的预测精确率的差值在不断变大; 模糊特征 $\tilde{X}_{42}$ 的权重先增大后减小, 模糊特征 $\tilde{X}_{42}$ 对预测精确率的影响先减小后增大.

综上所述, DFFS-VW 算法无需人为设置参数, 能够自动得到最优模糊特征子集, 具有良好的自适应性; 利用互信息量计算特征之间的相关性, 不仅考虑了特征之间的线性关系, 还考虑了非线性关系, 提高了特征选择结果的分类准确度和预测精确率; 考虑各个模糊特征的权重, 得到候选模糊特征子集, 并采取后向搜索的方式得到离线模糊特征选择的优化模糊特征子集, 减少了算法运行时间, 提高了算法的效率; 将当前窗口的候选模糊特征子集和历史模糊特征选择结果的交集作为当前窗口的初始优化

发现模糊输入特征对分类准确性或者预测精确性的影响.

5 总 结

本文研究动态模糊特征选择问题, 无论是连续数据集, 还是类别数据集, 以输入特征相对与输出特征的权重作为衡量输入特征相对于输出特征之间的重要性, 提出基于特征变权的动态模糊特征选择算法. 在离线模糊特征选择中, 首先根据特征权重, 获取候选模糊特征子集. 然后, 结合后向特征选择方式和模糊特征筛选指标, 筛选候选模糊特征子集得到优化模糊特征子集; 在线模糊特征选择中, 以上一个滑动窗口的优化模糊特征子集与当前滑动窗口中的候选模糊特征集的交集为基础, 根据模糊输入特征在模糊特征子集中的重要度, 获得当前窗口中的优化模糊特征子集. 更进一步, 我们分析了窗口之间模糊特征权重的变化, 发现输入模糊特征的演化关系和模糊输入特征对分类准确性或者预测误差的影响. 实验结果表明了本文所提动态模糊特征选择算法的有效性. 但目前的滑动窗口实质上仍然是在离线采集数据的基础上反复学习获得的, 本文未来将

研究随着在线数据变化,自适应地确定滑动窗口大小,进一步提升动态模糊特征选择算法的性能。

参 考 文 献

- [1] Xiao Jin, Xiao Yi, Huang Anqiang, et al. Feature-selection-based dynamic transfer ensemble model for customer churn prediction [J]. Knowledge and Information Systems, 2015, 43(1): 29-51
- [2] Zhang Xiangrong, He Yudi, Jiao Licheng, et al. Scaling cut criterion-based discriminant analysis for supervised dimension reduction [J]. Knowledge and Information Systems, 2015, 43(3): 633-655
- [3] Guyon I, Elisseeff A. An introduction to variable and feature selection [J]. Journal of Machine Learning Research, 2003, 3(3): 1157-1182
- [4] Li Dan, Zhou Yuxun, Hu Guoqiang, et al. Optimal sensor configuration and feature selection for AHU fault detection and diagnosis [J]. IEEE Trans on Industrial Informatics, 2016, 13(3): 1369-1380
- [5] Siwek K, Osowski S. Data mining methods for prediction of air pollution [J]. International Journal of Applied Mathematics and Computer Science, 2016, 26(2): 467-478
- [6] Soguero-Ruiz C, Hindberg K, Rojo-Alvarez J, et al. Support vector feature selection for early detection of anastomosis leakage from bag-of-words in electronic health records [J]. IEEE Journal Biomedical Health Informatics, 2014, 20(5): 1404-1415
- [7] Zhao Peilin, Steven C, Jin Rong. Double updating online learning [J]. Journal of Machine Learning Research, 2011, 12(may): 1587-1615
- [8] Kohavi R, John G H. Wrappers for feature subset selection [J]. Artificial Intelligence, 1997, 97(1/2): 273-324
- [9] Zhou Yang, Jin Rong, Steven C. Exclusive lasso for multi-task feature selection [J]. Journal of Machine Learning Research, 2010, 9: 988-995
- [10] Wang Jialei, Zhao Peilin, Steven C, et al. Online Feature selection and its applications [J]. IEEE Trans on Knowledge & Data Engineering, 2006, 26(3): 698-710
- [11] Li Yang, Gu Xueping. Feature selection for transient stability assessment based on improved maximal relevance and minimal redundancy criterion [J]. Proceedings of the Chinese Society of Electrical Engineering, 2013, 33(34): 179-186 (in Chinese)
(李扬, 顾雪平. 基于改进最大相关最小冗余判据的暂态稳定评估特征选择[J]. 中国电机工程学报, 2013, 33(34): 179-186)
- [12] Naqvi S S, Browne W N, Hollitt C. Feature quality-based dynamic feature selection for improving salient object detection [J]. IEEE Trans on Image Processing, 2016, 25(9): 4298-4313
- [13] Zhao Wei, Wang Yafei, Li Dan. A dynamic feature selection method based on combination of GA with K-means [C] // Proc of the Int Conf on Industrial Mechatronics and Automation. Piscataway, NJ: IEEE, 2010: 271-274
- [14] Raza M S, Qamar U. An incremental dependency calculation technique for feature selection using rough sets [J]. Information Sciences, 2016, 343(12): 41-65
- [15] Bilal I S, Keshav P D, Alamgir M H, et al. Diversification of fuzzy association rules to improve prediction accuracy [C] // Proc of the Int Conf on Fuzzy System. Piscataway, NJ: IEEE, 2010: 1-8
- [16] Jing Yuege, Li Tianrui, Luo Chuan, et al. An incremental approach for attribute reduction based on knowledge granularity [J]. Knowledge-Based Systems, 2016, 104(7): 24-38
- [17] Shao Weixiang, He Lifang, Lu Chunta, et al. Online unsupervised multi-view feature selection [C] //Proc of the 16th Int Conf on Data Mining. Piscataway, NJ: IEEE, 2016: 1203-1208
- [18] Huang H, Yoo S, Kasiviswanathan S P. Unsupervised feature selection on data streams [C] //Proc of the 24th ACM Int Conf on Information and Knowledge Management. newyork. New York, ACM, 2015: 1031-104
- [19] Li Yun, Wu Zhongfu. Fuzzy feature selection based on min-max learning rule and extension matrix [J]. Pattern Recognition, 2008, 41(1): 217-226
- [20] Lughofer E. On-line incremental feature weighting in evolving fuzzy classifiers [J]. Fuzzy Sets & Systems, 2011, 163(1): 1-23
- [21] Du Lei, Song Qinbao, Jia Xiaolin. Detecting concept drift: An information entropy based method using an adaptive sliding window [J]. Intelligent Data Analysis, 2014, 18(3): 337-364
- [22] Wang Ling, Meng Jianyao. The dynamic clustering algorithm of Bayesian adaptive resonance theory based on local distribution [J]. Control and Decision, 2018, 33(3): 471-478 (in Chinese)
(王玲, 孟建瑶. 基于局部分布的贝叶斯自适应共振理论增量聚类算法[J]. 控制与决策, 2018, 33(3): 471-478)
- [23] Comas D S, Meschino G J, Nowe A, et al. Discovering knowledge from data clustering using automatically-defined interval type-2 fuzzy predicates [J]. Expert Systems with Applications, 2016, 68(2): 136-150
- [24] Ghazikhani A, Monsefi R, Yazdi H S. Online neural network model for non-stationary and imbalanced data stream classification [J]. International Journal of Machine Learning and Cybernetics, 2014, 5(1): 51-62
- [25] Li Wermin, Han Jiawei, Pei Jian. CMAR: Accurate and efficient classification based on multiple class-association Rules [C] //Proc of 2001 IEEE Int Conf on Data Mining (ICDM). Piscataway, NJ: IEEE, 2001: 369-376

[26] Mu Yashuang, Liu Xiaodong, Yang Zhihao, et al. A parallel C4.5 decision tree algorithm based on MapReduce [J]. Concurrency and Computation, 2017, 29(6): 1-12

[27] Cui Yongfeng, Ma Xiangqian Li, Liu Zhijie. Application of improved BP neural network with correlation rules in network intrusion detection [J]. International Journal of Security and Its Applications, 2016, 10(4): 423-430

[28] Wang Lihuan, Guo Yonglong, Xin G. An asymmetric weighted SVR for construction final accounting prediction [J]. Journal of Information and Computational Science, 2014, 11(5): 1387-1394



Wang Ling, born in 1974. PhD, associate professor. Her main research interests include data mining and machine learning.



Meng Jianyao, born in 1992. Master. Her main research interests include data mining (jianyaomeng@qq.com).

2018 年《计算机研究与发展》专题(正刊)征文通知

——分布式安全与区块链技术研究

计算机网络已经渗透于万物之中,我们正处在一个万物互联的时代,无人驾驶汽车、家庭机器人以及家庭安保系统等各种网络、智能设备,都需要分布式网络系统的支撑。但是,随着分布式网络给人们带来越来越便利的同时,安全性威胁也越来越严重。区块链正在以让人们始料未及的速度和规模迅速发展,其去中心化、可验证、防篡改的特性以及进行价值传递的功能,也许会引起人们社会生活的深刻变革,改变社会某些领域的管理模式。区块链是以密码技术构造的新型信息系统体系,从设计到运行都涉及大量密码学与信息安全问题。

为推动我国学者在分布式安全及区块链技术领域研究,促进信息安全在基础理论、重要应用领域、新兴安全技术的发展,及时报道我国学者的最新研究成果,《计算机研究与发展》将于 2018 年 10 月出版网络与信息安全专题——分布式安全及区块链技术研究,主要聚焦分布式安全及智能安全、区块链技术等,欢迎相关领域的专家学者和科研人员踊跃投稿。

征文内容

本专辑包括(但不限于)下列主题:

- 1) 分布式密码算法与协议;
- 2) 分布式安全理论与智能安全;
- 3) 分布式网络攻防;
- 4) 物联网安全;
- 5) 电子货币与区块链技术;
- 6) 安全多方计算与隐私保护。

投稿要求

- 1) 论文应属于作者的科研成果,数据真实可靠,具有重要的学术价值或推广应用价值,未在国内外公开发行的刊物或会议上发表过,不存在一稿多投问题。作者在投稿时,需向编辑部提交版权转让协议。
- 2) 论文应包括题目、作者信息、摘要、关键词、正文和参考文献,论文一律用 Word 排版,论文格式体例格式请参考《计算机研究与发展》近期文章。
- 3) 论文需附通信作者的联系地址、电话或手机及 E-mail 地址。
- 4) 论文请通过期刊网站(<http://crad.ict.ac.cn>)进行投稿,并在作者留言中注明“网络与信息安全 2018 专题”(否则按自由来稿处理)。

重要日期

征文截止日期:2018 年 6 月 11 日

作者修改稿提交日期:2018 年 8 月 6 日

录用通知日期:2018 年 7 月 20 日

出版日期:2018 年 10 月

专题特邀编委

徐秋亮 教授 山东大学 xql@sdu.edu.cn

张玉清 教授 中国科学院大学、国家计算机网络入侵防范中心 zhangyq@ucas.ac.cn

董晓蕾 教授 华东师范大学 dongxiaolei@sei.ecnu.edu.cn

联系方式

联系人:徐秋亮 xql@sdu.edu.cn

编辑部:crad@ict.ac.cn, 010-62620696, 010-62600350

通信地址:北京 2704 信箱《计算机研究与发展》编辑部

邮政编码:100190