面向分类的流特征在线特征选择算法

尤殿龙[1,2],郭 松[1,2],赵春慧[1,2],原福永[1,2],申利民[1,2],陈 真[1,2]
(1. 燕山大学信息科学与工程学院,河北秦皇岛 066004;2. 河北省计算机虚拟技术与系统集成重点实验室,河北秦皇岛 066004)

摘 要: 在线流特征选择通过实时过滤无关特征和冗余特征,实现流特征空间降维.针对已有算法,如Alpha-investing分类精度低、SAOLA选择特征数多和OSFS在低冗余高相关数据集下运行时间长的问题,提出了一种面向分类的流特征在线特征选择算法——OSFIC.算法运用四层过滤框架,通过无条件独立过滤不相关新特征、单条件下互信息过滤冗余新特征和候选特征集合中的部分冗余特征,最后通过多条件独立过滤候选特征集中的剩余冗余特征,最终得到分类标签的近似马尔可夫毯.为了分析OSFIC的性能,选择了NIPS 2003和Causality Workbench中的数据集,从预测精度、特征数量、运行时间和AUC方面与已有基准算法进行比较.实验表明,OSFIC平均分类精度比Alpha-investing提升4.41%.在保证精度的前提下,平均特征数量比SAOLA减少41.9%,运行时间比OSFS减少91.59%.最后,在真实的应用场景下验证了OSFIC的有效性.

关键词: 在线特征选择;流特征;互信息;条件独立;近似马尔可夫毯
中图分类号: TP39 **文献标识码:** A **文章编号:** 0372-2112 (2020)02-0321-12
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.02.015

Online Feature Selection with Streaming Features for Classification

YOU Dian-long[1,2],GUO Song[1,2],ZHAO Chun-hui[1,2],YUAN Fu-yong[1,2],SHEN Li-min[1,2],CHEN Zhen[1,2]
(1. *School of Information Science and Engineering, Yanshan University, Qinhuangdao, Hebei* 066004, *China*;
2. *The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province, Qinhuangdao, Hebei* 066004, *China*)

Abstract: Online streaming feature selection achieves stream feature space dimensionality reduction by filtering irrelevant features and redundant features in real time. Existing works, such as Alpha-investing and Online Streaming Feature Selection (OSFS), have been proposed to serve this purpose, but they have drawbacks, including low prediction accuracy and high running time if the streaming features exhibit characteristics such as low redundancy and high relevance. We propose a novel classification-oriented online feature selection algorithm for streaming features, named OSFIC. OSFIC uses a four-layer filtering framework to filter irrelevant new features by null-conditional independence, filter redundant new features and redundant features in a candidate feature set by a single-conditional mutual information, and finally filter the remaining redundancy in the candidate feature set by multi-conditional independence. The approximate Markov blanket of the classify label is finally obtained. To analyze the performance of the algorithm, we selected the datasets in NIPS 2003 and Causality Workbench to compare prediction accuracy, number of selected features, runtime, and AUC with existing state-of-the-art algorithms. Experiments show that the average classification accuracy of OSFIC is 4.41% higher than that of Alpha-investing. Under the premise of high precision, the average number of features is 41.9% lower than SAOLA, and the runtime is 91.59% lower than OSFS. Finally, the efficiency of OSFIC is verified in real scenarios.

Key words: online feature selection; streaming feature; mutual information; conditional independence; approximate Markov blanket

1 引言

特征选择[1~4](Feature selection)的目标是从高维数据[5]中移除冗余特征和不相关特征[6],提取出原始特征空间[3]的"最优特征子集",以便提高分类器的分类精度[7~10],减少时间复杂度.近年来,随着社交网络、

收稿日期:2019-06-10;修回日期:2019-10-17;责任编辑:马兰英
基金项目:国家自然科学基金(No.61772450);中国博士后科学基金(No.2018M631764);河北省自然科学基金(No.F2019203287,No.F2017203307);河北省科技计划项目(No.17210701D)河北省博士后科研项目(No.B2018003009);河北省教育厅科学研究计划项目(No.KCJSX2017028)燕山大学基础研究专项课题(No.16SKY011);燕山大学博士基金(No.BL18003)

传感网络和在线视频网站等的普及,每天产生大量的图片和文字信息,同时,数据产生速度极快[11].例如,微博热门话题的实时更新[12],入侵检测系统中的特征收集[13],CCTV 流图像的动态捕获[14],垃圾邮件的实时过滤[15],直播网站不良内容监管[16]以及环境监测和分析[17]等.在这些应用场景中,随着时间的不断推移,特征以流的形式到达,特征数量不断增加,特征空间持续变化.对于在离线方式下处理已知特征空间的传统特征选择方法,如 Relief[18]、mRMR[19]等,将无法适用于这些流数据场景.因此,有必要依据分类标签实时地进行在线特征选择,以降低空间复杂度.作为特征选择的重要研究方向,流特征选择(Streaming Features Selection,SFS)假设训练实例数量是固定的,而特征数量随时间推移而增加[2].近年来,流特征的在线特征选择受到了极大的关注,其目标是在某一时刻,能够动态地处理当前到来的流特征[20~22].

为提高流特征选择的准确性,Perkins[23]等人提出了一个基于逐步梯度下降的 Grafting 算法去处理在线流特征选择问题.Grafting 是一个嵌入式特征选择方法,用以处理流特征中的冗余特征和不相关特征.但是需要提前获取特征数量,以便于可以得出正则化参数 λ.Zhou[24]等人提出一个基于逐步回归的在线流特征选择 Alpha-investing 算法,但 Alpha-investing 需要在已知候选特征先验知识的前提下,对初始特征进行变换,并且因为只判断特征的相关性,导致候选集合中存在大量冗余特征,导致分类精度降低.Wu[6]等人使用条件独立性检验作为筛选依据,提出一个 OSFS 算法.OSFS 使用 G^2 检验来表示特征间的条件独立或者依赖,然后识别冗余特征和不相关特征.当冗余特征量很多时,能够在选择更少的特征的前提下,提供比 Grafting 和 Alpha-investing 更高的分类精度.然而,当弱相关特征数量增加时,算法的运行时间会呈指数增长.Wang[25]等人提出使用组结构作为一个先验知识进行特征选择的 OGFS 算法,它经过两个阶段生成特征子集:组内特征选择和组外特征选择.但是,OGFS 需要提前选择少量参数.Wu[26]等人使用在线成对比较方法处理具有较高维度的特征集,并提出一个 SAOLA 算法.它采用互信息对特征两两间进行比较并过滤冗余特征,但是只判断在单个条件下是否冗余,无法将全部冗余特征移除,且算法本身无法获得最优特征之间相关性的判断阈值.

针对以上问题,本文提出一个在线过滤流特征(Online Streaming Features)的,采用互信息(Mutual Information)和条件独立性(Conditional Independence)的特征选择算法(简称 OSFIC).OSFIC 通过四层过滤选择相关特征,从特征空间中挖掘分类标签的马尔可夫毯.众所周知,马尔可夫毯[27]是分类标签的最优特征子集,在忠实性条件下,马尔可夫毯是唯一的.但流特征下的数据集往往具有非忠实性,因此,OSFIC 仅挖掘近似马尔可夫毯.研究过程中主要有以下挑战:(1)如何发现具有高分类精度的近似马尔可夫毯;(2)如何在保证高分类精度的前提下,缩短运行时间和减少选择特征量;(3)如何评估算法的性能并解决其不足.

本文创新点和贡献如下:(1)结合互信息和条件独立性检验,采用四层过滤方法处理冗余特征和不相关特征;(2)提出一种适用于流特征选择框架,根据对新到达特征进行相关性检验和对候选集合进行冗余性分析,获得分类标签的近似马尔可夫毯;(3)在理论上分析 OSFIC 算法的有效性和性能;通过与已有算法的实验对比,得出 OSFIC 在保证分类精度提高的前提下,特征选择数量和运行时间均得到明显优化.

2 基于互信息和条件独立性检验的流特征选择框架

流特征之间存在不相关、弱相关且冗余、弱相关且非冗余和强相关四种关系[6].强相关特征对于提高分类的分类精度起到至关重要的作用,因此不能在不影响分类精度的情况下移除强相关特征;弱相关特征分为冗余特征和非冗余特征[6],非冗余特征有助于提高分类精度,因此需要在移除冗余特征的同时保留非冗余特征;不相关特征对于分类精度的影响无关紧要[27],需要从数据集中识别并移除不相关特征.表 1 为本文中所用到的符号及其含义.

表 1 符号及其含义

符号	含义
D	含有流特征的特征空间集合
CFS	候选特征集合
$P(.\mid.)$	特征间的条件概率
$I(.\mid.)$	特征间的互信息量
f_i	在 t_i 时刻到达的第 i 个输入特征
f,x,y,z	在 CFS 中的特征
t_i	第 i 个特征到达的时刻 t
C	分类标签
α	显著性水平
S	当前候选集合的子集
$CFS-\{.\}$	除去某个特征的特征集合或数据集
$O()$	时间复杂度
$\perp$	独立

2.1 问题定义

通常,特征空间 D 定义为 $D=(f_i,c)$,f_i 代表每一个特征向量,c 为分类标签 C 的向量表示. D 中特征可分为与分类标签 C 强相关、冗余、非冗余和不相关四种关系,特征选择问题是从 D 中寻找出一个子集以最大化分类和预测模型的性能,即从 D 中过滤分类标签的冗余特征和不相关的特征,同时保留下强相关特征和非冗余特征.

定义 1[28]　条件独立(Conditional Independence). 如果特征 $x\in D$ 与 $y\in D$ 是条件独立,当且仅当存在一个子集 $S\subseteq D$,使得 $P(x|S,y)=P(x|S)$ 或者 $P(y|S,x)=P(y|S)$ 成立. 条件独立是多随机变量的重要概念. 如果两个变量 x 和 y 是独立的,则它们的联合分布概率 $P(x,y)=P(x)P(y)$,即 $x\perp y$. 如果 x 和 y 在给定条件集合 S 下相互独立,则它们的联合分布概率 $P(x,y|S)=P(x|S)P(y|S)$.

定义 2　互信息(Mutual Information,MI). 给定两个特征 x 和 y,则 x 与 y 的互信息定义为:

$$I(x,y)=H(x)-H(y) \tag{1}$$

其中 $H(x)$ 定义为特征 f 的熵:

$$H(x)=-\sum_{i=1}^{n}P(x_i)\log_2 P(x_i) \tag{2}$$

n 为 x 的元素个数.

定义 3[6]　强相关(Strong relevance). 如果特征 x 与分类标签 C 强相关,当且仅当 $\forall S\subseteq D-\{x\}$,使得 $P(C|S,x)\neq P(C|S)$ 成立. 如果特征 x 在所有给定条件的情况下,对分类标签 C 均具有预测性,则特征 x 和分类标签 C 之间具有强相关性[27].

定义 4[6]　弱相关(Weak relevance). 如果特征 x 与分类标签 C 弱相关,当且仅当特征 x 与分类标签 C 不是强相关,且 $\exists S\subseteq D-\{x\}$,使得 $P(C|S,x)\neq P(C|S)$ 成立.

如果特征 x 在某些给定条件的情况下,对分类标签 C 具有预测性,则特征 x 和分类标签 C 之间具有弱相关性[27].

定义 5[6]　不相关(Irrelevance). 如果特征 x 与类别属性 C 不相关,当且仅当特征 x 与分类标签 C 既不强相关也不弱相关,且 $\forall S\subseteq D-\{x\}$,使得 $P(C|S,x)=P(C|S)$ 成立. 如果特征之间既不存在强相关关系,也不存在弱相关关系,则称它们之间不相关,即如果特征 x 在所有给定条件下,对分类标签 C 均不具有预测性,则特征 x 和分类标签 C 之间不相关[27].

定义 6　马尔可夫毯(Markov Blanket,MB). 马尔可夫毯是特征空间 D 的一个子集,其中特征 x 的马尔可夫毯定义为 $MB\subseteq D-\{x\}$,对于 $\forall S\subseteq D-(MB\cup\{x\})$,使得 $P(C|S,MB)=P(C|MB)$ 成立.

定义 7[6]　冗余特征(Redundant features). 如果特征 x 是冗余特征,当且仅当特征 x 为分类标签 C 的弱相关特征,且特征 x 在 $MB(C)$ 有一个马尔可夫毯 $MB(x)$.

1996 年,Koller 和 Sahami 从信息论角度证明了分类标签 C 的马尔可夫毯包含了所有其他特征对分类标签 C 的预测信息,即马尔可夫毯是分类标签 C 的最优特征子集——排除了不相关特征和冗余特征.

2.2 流特征下在线特征选择框架

本节提出了一个基于互信息和条件独立性检验的流特征选择框架. 如表 2 所示.

表 2　基于互信息和条件独立性检验的流特征选择框架

OSFIC 框架
1 **初始化**:分类标签 C;候选特征集合 $CFS=\{\}$;
2. 在 t_i 时刻得到新到特征 f_i.
3. **阶段 1**:分析新到达特征
3.1 过滤 1:通过条件独立性检验判断 f_i 与 C 的相关性,如果两者不相关则过滤 f_i;否则进行下一步.
3.2 过滤 2:通过互信息判断在 $f\in CFS$ 为条件下,f_i 是否为冗余特征,如果冗余则过滤 f_i;否则进行下一步.
4. **阶段 2**:分析候选特征集合
4.1 过滤 3:对于 $\forall f\in CFS$,通过互信息判断在特征 $x\in CFS-\{f\}$ 的条件下,f 是否为冗余特征,如果冗余则过滤 f;否则进行下一步.
4.2 过滤 4:对于 $\forall f\in CFS$,通过条件独立性检验判断在子集 $S\subseteq CFS-\{f\}$ 的条件下,f 是否为冗余特征,如果冗余则过滤 f;否则继续判断其余特征,直至 CFS 中所有特征判定完成.
5. 重复 2~4 步骤,直至没有新到达特征或者满足停止条件.
6. 输出 CFS

2.2.1 过滤 1　无条件下独立性检验过滤不相关新特征

推论 1　如果特征在过滤 1 中被移除,那么该特征为不相关特征.

证明　假设特征 $x,y\in CFS$,并且结合定义 1 和定义 5,式(3)推论成立.

$$\begin{aligned}x\perp y&\Rightarrow P(x|y)=P(x)\\&\Rightarrow\frac{P(x,y)}{P(y)}=P(x)\\&\Rightarrow P(x,y)=P(x)P(y)\\&\Rightarrow x\perp y\end{aligned} \tag{3}$$

因此,x 和 y 是无条件独立,即 x 和 y 为不相关特征.

2.2.2 过滤 2　单条件下互信息过滤冗余的新特征

引理 1[12]

$$\begin{aligned}I(x,y|z)&=H(x|z)-H(x|yz)\\&=H(x|z)+H(y|z)\\&\quad-H(x,y,z)-H(z)\end{aligned}$$

引理 2　$I(x,y|z)\geq 0$

引理 3　假设 $t-1$ 时刻的候选特征集合 CFS_{t-1},当

新到达的特征 f_i 到达时,如果 $\exists f\in CFS_{t-1}$,使得 $I(f_i,C\mid f)=0$,则 $I(f_i,f)\geq I(f_i,C)$.

证明 通过式(1)和引理1可得:

$$\begin{aligned}I(f_i,C)+I(f_i,f\mid C)&=H(f_i)-H(f_i\mid C)+H(f_i\mid C)\\&\quad-H(f_i\mid fC)\\&=H(f_i)-H(f_i\mid fC)\end{aligned}\tag{4}$$

$$\begin{aligned}I(f_i,f)+I(f_i,C\mid f)&=H(f_i)-H(f_i\mid f)+H(f_i\mid f)\\&\quad-H(f_i\mid fC)\\&=H(f_i)-H(f_i\mid fC)\end{aligned}\tag{5}$$

通过式(4)和式(5)得出:

$$I(f_i,C\mid f)=I(f_i,C)+I(f_i,f\mid C)-I(f_i,f)\tag{6}$$

如果 $I(f_i,C\mid f)=0$,则由式(6)得出:

$$I(f_i,f)=I(f_i,C)+I(f_i,f\mid C)\tag{7}$$

通过引理2和式(7)可得:

$$I(f_i,f)\geq I(f_i,C)$$

引理4 假设 $t-1$ 时刻的候选特征集合 CFS_{t-1},当新特征 f_i 到达时,如果 $\exists f\in CFS_{t-1}$,使得 $I(f_i,C\mid f)=0$,则 $I(f,C)\geq I(f_i,C)$.

证明 通过引理1得出 $I(f,f_i\mid C)=I(f_i,f\mid C)$,并结合式(7)可得:

$$I(f,C\mid f_i)-I(f,C)=I(f,f_i\mid C)-I(f_i,f)\tag{8}$$

由此可得:

$$I(f,C\mid f_i)=I(f,C)-I(f_i,C)\tag{9}$$

如果 $I(f,C\mid f_i)=0$,则 $I(f,C)=I(f_i,C)$;如果 $I(f,C\mid f_i)>0$,则 $I(f,C)>I(f_i,C)$.

推论2 如果特征在过滤2中被移除,那么该特征为冗余特征.

证明 通过引理4可以看出,如果考虑 $I(f_i,C\mid f)=0$ 且 $I(f,C\mid f_i)=0$,则 $I(f_i,C)$ 和 $I(f,C)$ 相等,f_i 和 f 能够相互代替. 如果考虑 $I(f_i,C\mid f)=0$ 且 $I(f,C\mid f_i)>0$,结合引理2可得,如果 $I(f_i,C\mid f)=0$,则下列式(4)成立.

$$I(f,C)>I(f_i,C)\text{且}I(f_i,f)>I(f_i,C)\tag{10}$$

当一个新到达的特征 f_i 在过滤2被移除,证明 f_i 在当前候选特征集合 CFS_{t-1} 已经存在能够满足要求的特征,f_i 为冗余特征.

2.2.3 过滤3 单条件下互信息过滤候选集合中冗余特征

同2.2.2节所述,对于候选特征集合中的特征 f_1,f_2,如果特征 f_1 在过滤3被移除,则证明 f_1 为冗余特征.

$$I(f_1,C)\geq I(f_2,C)\text{且}I(f_1,f_2)\geq I(f_2,C)\tag{11}$$

其中 $f_1\in CFS$, $f_2\in CFS-\{f_1\}$.

2.2.4 过滤4 多条件下独立性检验过滤候选集合中冗余特征

推论3 如果 f 经过过滤4筛选,对于 $\forall f\in CFS$,如果 $\forall S\subseteq CFS-\{f\}$ 满足 $f\perp C\mid S$,那么 $f\notin MB(C)$.

证明 如果 $\forall S\subseteq CFS-\{f\}$ 满足 $f\perp C\mid S$,因为 $MB(C)$ 是 CFS 的一个子集,那么 $\exists S=MB(C)$,满足 $f\perp C\mid MB(C)$,按照定义6,如果 $f\perp C\mid MB(C)$,$f\notin MB(C)$.

经过上述过程筛选,最终保留在 CFS 中的特征为强相关和非冗余特征.

OSFIC 框架有以下优点:(1)对新到来的特征,通过过滤1和过滤2两层过滤,减少进入候选集合的特征数量;(2)在特征间两两比较的基础上,通过对特征进行多条件独立性检验过滤冗余特征;(3)在通过多条件下独立性检验过滤冗余特征时运行时间更少.

3 在线流特征选择算法

3.1 OSFIC 算法及其分析

本文将上述 OSFIC 框架应用于流特征选择,并提出 OSFIC 算法. 使用独立性检验时,对于离散数据,算法使用 G^2 检验[29]进行条件独立性检验,对于连续数据,算法使用 fisher-z 检验[29]进行条件独立性检验.

其中 G^2 定义为:

$$G^2=2\sum_{i,j,k}S_{xyz}^{ijk}\ln\frac{S_{xyz}^{ijk}S_z^k}{S_{xz}^{ik}S_{yz}^{jk}}\tag{12}$$

式(11)中 S_{xyz}^{ijk} 代表特征 x,y,z 分别满足 $x=i,y=j,z=k$ 所满足的个数,S_{xz}^{ik},S_{yz}^{jk} 同理.

其中 fisher-z 检验中 Z 定义为:

$$Z=\frac{1}{2}\sqrt{n-|z|-3}\left(\ln\frac{1+\xi}{1-\xi}\right)\tag{13}$$

式(13)中 n 代表样本大小,z 为条件特征,ξ 代表在给定 z 特征条件下,特征 x 和 y 的总体偏相关系数,具体定义如下:

$$\xi_{(xy\mid z)}=\frac{\xi_{(xy)}-\xi_{(xz)}\xi_{(yz)}}{\sqrt{1-\xi_{xz}^2}\sqrt{1-\xi_{yz}^2}}\tag{14}$$

式(14)中 $\xi(xy)$ 代表特征 x 和特征 y 的相关系数,假设特征 $\boldsymbol{x}=(x_1,x_2,x_3\cdots x_n)^{\mathrm{T}}$,特征 $\boldsymbol{y}=(y_1,y_2,y_3\cdots y_n)^{\mathrm{T}}$. 则 $\xi_{(xy)}$ 定义为:

$$\xi_{(xy)}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}\tag{15}$$

$\xi_{(xz)}$,$\xi_{(yz)}$ 以此类推.

在 fisher-z 检验中,假设在给定 z 的条件下,特征 x 和 y 条件独立为零假设,即 $\xi(x,y\mid c)=0$;ρ 为 fisher-z 检验中返回的 p 值且 α 为显著性水平,如果 $\rho>\alpha$,则零假设成立,特征 x 和 y 之间无关;如果 $\rho\leqslant\alpha$,则零假设被拒绝,特征 x 和 y 之间相关. 根据推论2中式(9)结论可得,在 t 时刻,如果 $\xi_{(yz)}>\xi_{(xz)}$ 且 $\xi_{(xy)}>\xi_{(xz)}$,则在给定条件 z 下,特征 x 与 y 相互独立.

使用互信息检验时,对于离散数据,使用对称不确

定性[30] $SU(x,y)$代替互信息量 $I(x,y)$,对于连续数据,使用上述 fisher-z 检验中的总体偏相关系数 ξ 代替互信息量 $I(x,y)$. 对称不确定性 $SU(x,y)$定义为:

$$SU(x,y)=\frac{2I(x,y)}{H(x)+H(y)} \tag{16}$$

使用 $SU(x,y)$代替 $I(x,y)$的优点是能够将特征间的关系量化在 0 -1 之间,减小 $I(x,y)$对特征间的偏差影响. 具体流程如算法 1 所示.

算法 1　在线流特征选择算法 OSFIC　　$O(\cdot)$

输入:数据集 *Dataset*、阈值 α、分类标签 C,候选特征集合 *CFS*
输出:被选特征集合 *SF*
/ * 将新到达的特征赋值 f_i * /
1. $CFS=\varnothing$,新到达特征→f_i,$flag=0$
2. repeat　　$\times |N|$
/ * 如果在无条件独立性检验下,f_i 与 C 相互独立,表示 f_i 为不相关特征 * /
3.　if $f_i \perp C|[\,]$ then 移除 f_i;　　$O(1)$
4.　end if
/ * 如果在 $\exists x \in CFS$ 的条件下,f_i 与 C 相互独立,证明 f_i 为冗余特征 * /
5.　if $\exists x \in CFS$ 满足 $I(x,C)>I(f_i,C)$ 且 $I(f_i,x)>I(f_i,C)$ then 移除 f_i;　　$O(|CFS|)$
6.　else $CFS=CFS\cup\{f_i\}$,$flag=1$;　　$O(1)$
7.　end if
8.　if $flag=1$　　$O(1)$
9.　　for each $y \in CFS-\{x\}$　　$\times |CFS|$
10.　　if $I(y,C)>I(x,C)$ 且 $I(x,y)>I(x,C)$ then $CFS=CFS-\{x\}$　　$O(1)$
11.　　end if
12.　end for
/ * 遍历 *CFS* 中的每个特征 x,如果 $\exists S \subseteq CFS-\{x\}$,使得 x 与 C 不相关,证明该特征为冗余特征 * /
13.　for each $x \in CFS$　　$\times |CFS|$
14.　　if $\exists S \subseteq CFS-\{x\}$ 满足 $x \perp C|S$ *then* $CFS=CFS-\{x\}$　　$O(2^{|CFS|})$
15.　　end if
16.　end for
17. 直至所有特征完成.
18. $SF=CFS$

3.2　OSFIC 时间复杂度分析

OSFIC 的时间复杂度主要取决于新到达特征分析和候选集合分析两个阶段. 如表 3 所示,其中 $|N|$ 为当前已到达的特征数量,$|N_i|$ 为当前已到达且与分类标签 C 无关的特征数量,$|M|$ 为在进行基于条件独立性检验的多条件过滤前所保留的特征数量,$|CFS|$ 为当前候选集合中特征数量.

如表 3 所示,OSFIC 的时间复杂度为 $O(N+2|N-N_i||CFS|+|M||CFS|2^{|CFS|})$,主要取决于 $|N|$、$|N_i|$、$|M|$ 和 $|CFS|$,各参数之间的大小关系为 $|CFS|<|M|<|N-N_i|<|N|$,且一般情况下,$|CFS|$ 远小于 $|N-N_i|$. 算法中时间主要花费在多条件独立性检验的过程,$|M|$ 和 $|CFS|$ 对算法的时间复杂度影响较大. 随着强相关特征不断加入到 *CFS* 中,OSFIC 的时间复杂度将随之变高,最坏情况为 $O(N+2|N||CFS|+|N||CFS|2^{|CFS|})$,最好情况为 $O(|N|)$. 冗余特征和不相关特征占特征空间的比重越大,OSFIC 的时间复杂度越低.

表 3　算法中两个阶段的时间复杂度

OSFIC 执行阶段	时间复杂度
分析新到达特征	
过滤 1:无条件独立性检验过滤不相关新特征	$O(\|N\|)$
过滤 2:单条件互信息过滤冗余的新特征	$O(\|N-N_i\|\|CFS\|)$
分析候选特征集合	
过滤 3:单条件互信息过滤候选集合中冗余特征	$O(\|N-N_i\|\|CFS\|)$
过滤 4:多条件独立性检验过滤候选集合冗余特征	$O(\|M\|\|CFS\|2^{\|CFS\|})$

3.3　OSFIC 的近似马尔可夫毯

本节研究 OSFIC 仅挖掘流特征的近似马尔可夫毯原因:(1)在现实生活中,很难获得忠实性数据,所以使得分类标签 C 的马尔可夫毯不唯一;(2)最优特征子集应该选择强相关和非冗余的特征,但是在流式传输中无法获得所有特征的先验知识,所以无法找到全部强相关特征和非冗余特征. 根据 2.2 节中所得推论 1 ~3 可知 OSFIC 移除的特征为冗余特征和不相关特征,保留下的为非冗余和强相关特征. 根据定义 6 所得最终特征子集即为分类标签 C 的近似马尔可夫毯.

4　实验结果与讨论

4.1　实验准备

实验测试的 14 个数据集的特征数量及样本数量如表 4 所示. 来自于 NIPS 2003 特征选择挑战提供的数据集 madelon、allaml、ionosphere,以及 Causality Workbench 网站提供的 Sylva、regedi1、lucas0、marti1 和 lucap0,常用的公共微阵列数据集 wdbc 和 lung-cancer 等. 其中,marti1,reged1,lung,prosate_GE,arcene 和 SMK_CAN_187 的特征数量大于样本数. 14 个数据集涵盖了广泛的实际应用领域,包括基因表达,生态学和因果发现.

所有的实验均在 Windows 10 PC,Intel(R)® Core(TM)® i5-7500 CPU @ 3.40GHz,8G RAM 环境中进行,采用 Matlab 2016a 实现 OSFIC 算法,每次实验进行多次 10 折交叉验证,显著性水平 α 为 0.05. 实验设计如下:(1)分析 OSFIC 在每个阶段的特征数量变化;(2)

比较 OSFIC、Alpha-investing、OSFS 和 SAOLA 在上述数据集的分类精度;(3)分析上述四种算法在数据集上的选择特征数量及运行时间.

表 4 实验所用数据集的特征数量和样本数量

数据集	特征数量	样本数量	数据集	特征数量	样本数量
allaml	7129	72	madelon	500	2600
lung-cancer	12533	181	marti1	1024	500
hiva	1617	3845	Prostate_GE	5966	102
ionosphere	34	351	reged1	999	500
lucap0	143	2000	SMK_CAN_187	19993	187
cina0	132	16033	Sylva	216	13086
lung	3312	203	wdbc	30	569

4.2 不同阶段的选择特征数量

本节对 14 个数据集在两个阶段中选择特征的数量分别进行统计,具体结果如表 5 所示.

表 5 在两个阶段中 OSFIC 选择的特征数量

数据集	选择特征数量				
	阶段 1			阶段 2	
	初始	过滤 1	过滤 2	过滤 3	过滤 4
allaml	7129	2100	79	79	4
lung-cancer	12533	5933	137	137	8
hiva	1617	619	27	26	10
ionosphere	34	25	5	5	4
lucap0	143	94	35	35	22
cina0	132	106	20	20	11
lung	3312	2318	94	94	14
madelon	500	41	27	26	13
marti1	1024	1	1	1	1
Prostate_GE	5966	3182	38	37	4
reged1	999	541	30	30	12
SMK_CAN_187	19993	4924	35	34	6
Sylva	216	77	14	14	7
wdbc	30	25	9	8	3

随着特征数量变多,数据集中的冗余特征和不相关特征也随之变多,OSFIC 筛选效率随之变高. 因为 lucap0、lung、madelon、reged1、cina0 五个数据集中的强相关和非冗余特征较多,所以最终候选集合中特征个数较多. 在目标特征分析阶段和候选集合分析阶段中,通过过滤 2 和过滤 3 所得到的特征个数基本保持一致,原因是在过滤 1 与过滤 2 已过滤大部分冗余特征,并且过滤 3 是以单个特征为条件判断,能够找出的冗余特征有限,所以过滤 3 中的特征个数与过滤 2 相近.

4.3 OSFIC 与在线流特征算法的比较

实验使用 LOFS(Library of Online Streaming Feature Selection)中的 Alpha-investing、OSFS 和 SAOLA 作为对比算法,并采用 MATLAB 自带分类工具箱中的 Decision Tree(Complex Tree、Medium Tree、Simple Tree)、KNN(Medium KNN、Coarse KNN、Cosine KNN)、SVM(Linear SVM、Quadratic SVM、Coarse Gaussian SVM)和 Ensemble(Bagged Trees、Subspace Discriminant、RUSBoosted Trees)12 种分类器进行分类. 实验从分类精度,所选特征个数和运行时间三方面比较 OSFIC 与现有流特征选择算法.

4.3.1 分类精度对比

如图 1 ~ 图 4 所示实验中 14 个数据集在上述分类器上的四种算法的分类精度比较. OSFIC、OSFS 和 SAOLA 算法在上述分类器的整体分类性能明显优于 Alpha-investing. 数据集 wdbc 和 Prostate_GE 在 Decision Tree 分类器上,OSFIC 分类性能明显优于 SAOLA;数据集 lung 在 Decision Tree 分类器上以及数据集 allaml 在 KNN 分类器上 OSFIC 分类性能明显优于 OSFS. 通过 14 个数据集在各个分类器上的比较,OSFIC 能够在提高已有算法的分类精度基础上保持较稳定的分类性能.

如表 6 所示 14 个数据集在不同算法下的平均分类精度. 在 Decision Tree、KNN、SVM 和 Ensemble 分类器上,OSFIC 的平均精度(88.07、87.03、89.75、87.88)明显高于 Alpha-investing 算法(83.51、83.34、85.64、82.60)和 OSFS(87.22、86.58、89.34、86.88),分别高出 4.41% 和 0.67%;比较 SAOLA(87.49、86.91、89.73、88.05),在 Ensemble 分类器上分类精度略低,其余三个分类器均高于 SAOLA. 经过对比发现,OSFIC 在分类精度的性能方面明显高于 Alpha-investing 算法. 表 7 为进行多次实验后,13 个数据集在 12 个分类器上的(lung 为多分类数据集,不做比较)OSFIC、OSFS、SAOLA 和 Alpha-investing 四种算法平均 AUC 结果. 从表中可得,OSFIC、OSFS、SAOLA 三种算法在数据集上的 AUC 明显高于 Alpha-investing. 在 allaml、lung-cancer、ionosphere、cina0、madelon、reged1 和 wdbc 数据集上,OSFIC 的 AUC 均优于其他三种算法.

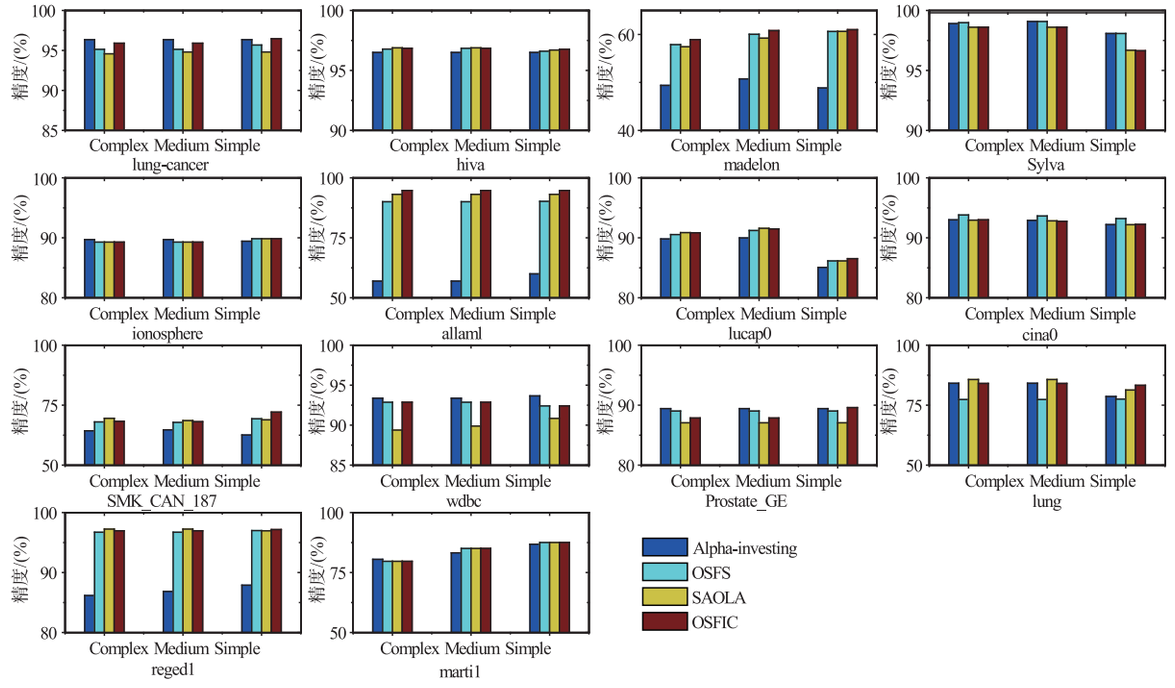

图1　14个数据集下四种算法在Decision Tree分类器精度比较

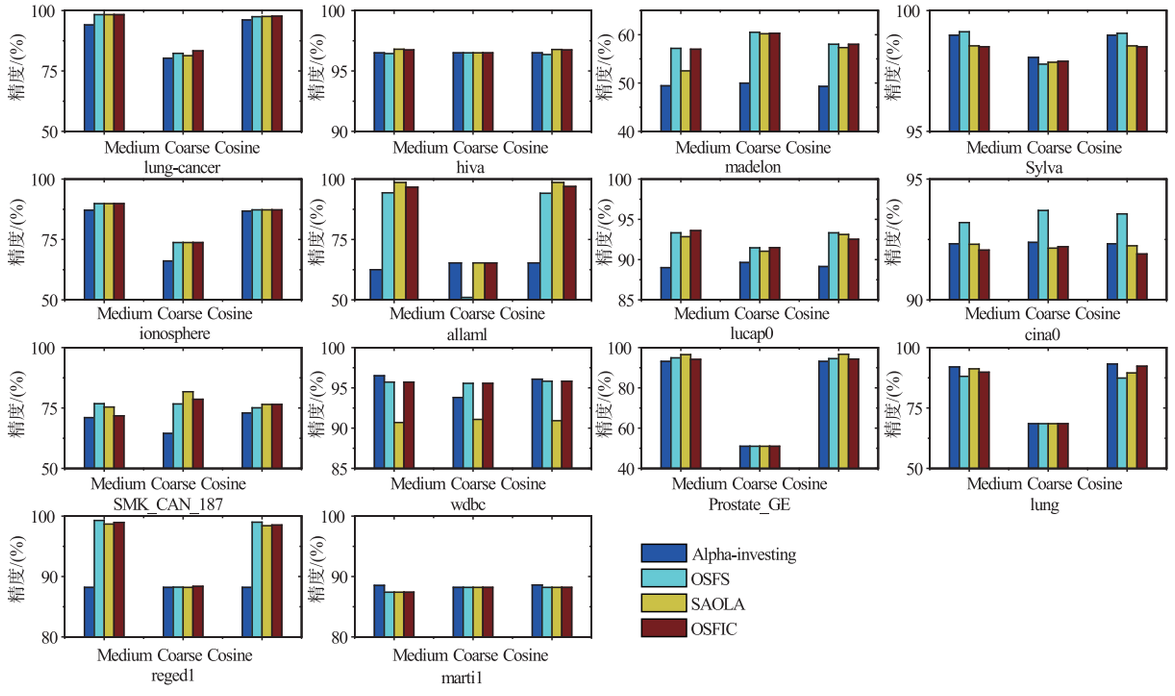

图2　14个数据集下四种算法在KNN分类器精度比较

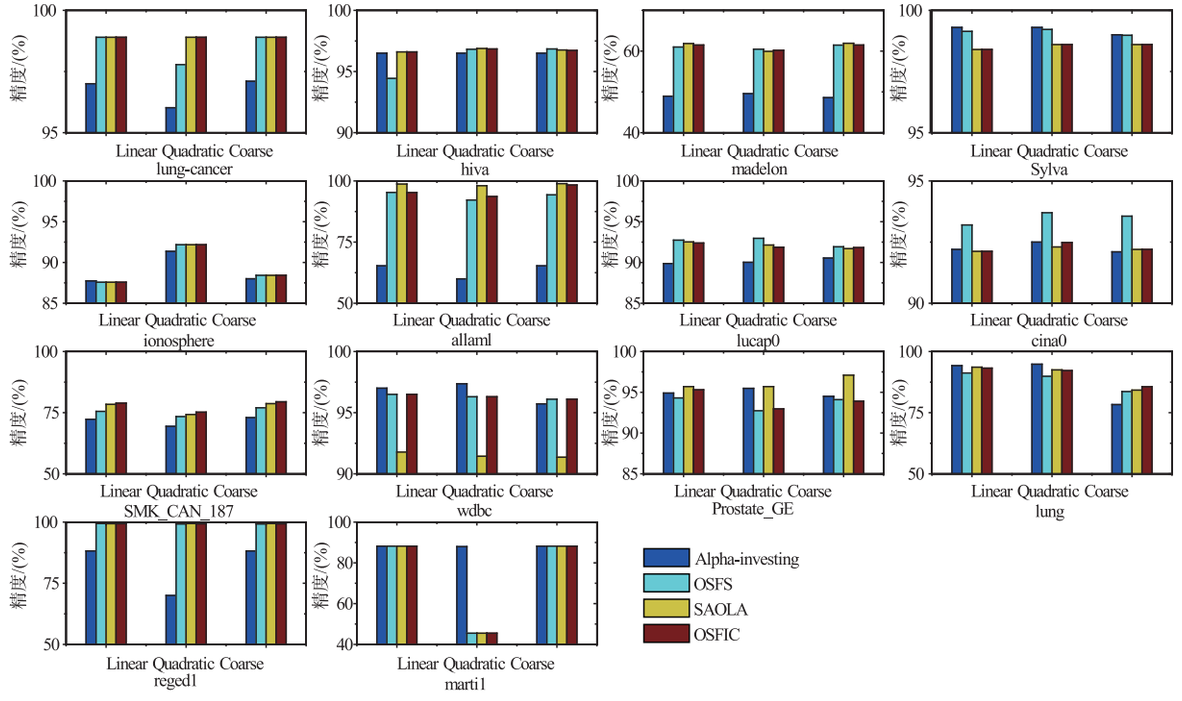

图3 14个数据集下四种算法在SVM分类器精度比较

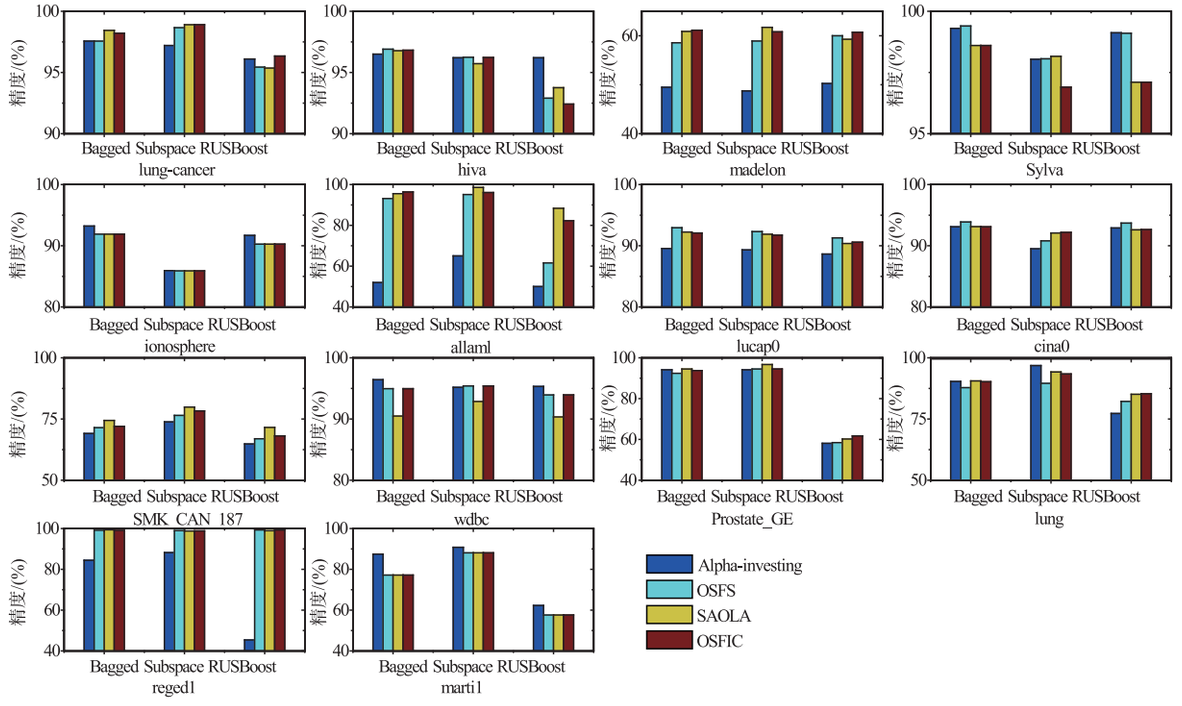

图4 14个数据集下四种算法在Ensemble分类器精度比较

表 6　平均分类精度比较

	14 个数据集的平均分类精度(%)					
OSFIC	Complex Tree	Medium Tree	Simple Tree	Linear SVM	Quadratic SVM	Coarse Gaussian SVM
	87.69	88.23	88.30	91.01	87.58	90.64
	Decision Tree 平均分类精度:88.07			SVM 平均分类精度:**89.75**		
	Medium KNN	Coarse KNN	Cosine KNN	Boosted Trees	Subspace Discriminant	RUSBoosted Trees
	90.03	80.71	90.36	89.67	90.52	83.44
	KNN 平均分类精度:87.03			**ENSEMBLE 平均分类精度:87.88**		
	平均分类精度:88.18					
OSFS	Complex Tree	Medium Tree	Simple Tree	Linear SVM	Quadratic SVM	Coarse Gaussian SVM
	86.86	87.43	87.37	90.53	87.31	90.19
	Decision Tree 平均分类精度:87.22			SVM 平均分类精度:**89.34**		
	Medium KNN	Coarse KNN	Cosine KNN	Boosted Trees	Subspace Discriminant	RUSBoosted Trees
	90.27	79.57	89.91	89.08	89.93	81.62
	KNN 平均分类精度:86.58			**ENSEMBLE 平均分类精度:86.88**		
	平均分类精度:87.50					
SAOLA	Complex Tree	Medium Tree	Simple Tree	Linear SVM	Quadratic SVM	Coarse Gaussian SVM
	87.30	87.84	87.33	91.05	87.70	90.45
	Decision Tree 平均分类精度:87.49			**SVM 平均分类精度:89.73**		
	Medium KNN	Coarse KNN	Cosine KNN	Boosted Trees	Subspace Discriminant	RUSBoosted Trees
	89.97	80.56	90.22	89.57	90.96	83.63
	KNN 平均分类精度:86.91			**ENSEMBLE 平均分类精度:88.05**		
	平均分类精度:88.05					
Alpha-investing	Complex Tree	Medium Tree	Simple Tree	Linear SVM	Quadratic SVM	Coarse Gaussian SVM
	83.46	83.84	83.24	86.55	85.02	85.36
	Decision Tree 平均分类精度:83.51			SVM 平均分类精度:**85.64**		
	Medium KNN	Coarse KNN	Cosine KNN	Boosted Trees	Subspace Discriminant	RUSBoosted Trees
	85.66	78.17	86.18	85.18	86.35	76.28
	KNN 平均分类精度:83.34			**ENSEMBLE 平均分类精度:82.60**		
	平均分类精度:83.77					

4.3.2　选择特征数量对比

本节对比 OSFIC、OSFS 与 SAOLA 选择出的候选特征中的特征数量,如表 8 所示. 定义 $\sum$# 算法在 14 个数据集中选择特征数量总和. 通过平均特征数量比值可得,OSFIC 在三个算法中选择的特征数量最少;与 OSFIC 相比,OSFS 的平均特征数量比为 20.6%,SAOLA 的平均特征数量比为 41.9%,均明显高于 OSFIC. 原因是 OSFIC 采取两种判断标准为依据,在分析目标特征与分类标签的相关性的同时在多条件下移除冗余特征,以此保证输出分类标签的近似马尔可夫毯,所以选取了比 OSFS 和 SAOLA 更少的特征.

4.3.3　运行时间对比

表 9 是 OSFIC、OSFS 与 SAOLA 在 14 个数据集上的运行时间对比. 在所有的数据集中,SAOLA 和 OSFIC 明显快于 OSFS. 在 lung-cancer、hiva、lucap0、cina0、lung、reged1 和 Sylva 数据集上,OSFS 比 OSFIC 运行时间大幅增长,如表 10 所示. 原因是上述数据集的特征间具有低冗余和高相关特性,导致 OSFS 的运行时间呈指数增长.

通过对比得出 OSFIC 和 SAOLA 的运行时间明显少于 OSFS,分析原因如下:

表 7 OSFIC,OSFS,SAOLA 和 Alpha-investing 的 AUC 比较

数据集	OSFIC	OSFS	SAOLA	Alpha-investing
allaml	**0.9165**	0.9152	0.9147	0.5092
lung-cancer	**0.9655**	0.9492	0.9570	0.9600
hiva	0.6673	0.6717	0.6687	0.5233
ionosphere	**0.9268**	**0.9268**	**0.9268**	0.9256
lucap0	0.9433	0.9485	0.9438	0.9250
cina0	**0.9577**	0.9570	0.9565	0.9550
madelon	**0.6353**	0.6310	0.6298	0.4942
marti1	0.5818	0.5818	0.5818	0.6633
Prostate_GE	0.8955	0.9073	0.8963	0.9050
reged1	**0.9808**	0.9798	0.9762	0.4783
SMK_CAN_187	0.7853	0.7985	0.8107	0.7392
Sylva	0.9718	0.9883	0.9718	0.9858
wdbc	**0.9758**	**0.9758**	0.9740	0.9754

表 8 OSFIC,OSFS 和 SAOLA 选择特征数量比较

数据集	OSFIC	OSFS	SAOLA
allaml	4	4	21
lung-cancer	8	4	39
hiva	10	13	9
ionosphere	4	4	4
lucap0	22	36	22
cina0	11	22	9
lung	14	11	30
madelon	13	14	21
marti1	1	1	1
Prostate_GE	4	3	12
reged1	12	13	16
SMK_CAN_187	6	4	11
Sylva	7	18	8
wdbc	3	3	2

OSFS 与 OSFIC 平均特征数量比:

$$\left(\sum \#_{OSFS} - \sum \#_{OSFIC}\right) \Big/ \sum \#_{OSFS} = 20.6\%$$

SAOLA 与 OSFIC 平均特征数量比:

$$\left(\sum \#_{SAOLA} - \sum \#_{OSFIC}\right) \Big/ \sum \#_{SAOLA} = 41.9\%$$

(1) 在 lung-cancer、hiva、lucap0、cina0、lung、reged1 和 Sylva 数据集中,OSFIC 比 OSFS 的运行时间短,原因是 OSFIC 同时对新到达特征进行相关性检验和单条件互信息过滤,移除了大量冗余特征,同时进行多条件独立性检验过滤的时间复杂度大幅降低.

(2) 因为 SAOLA 对新到达特征只进行在单条件下的冗余特征过滤,虽然降低了运行时间,但是会遗漏冗余特征,造成选择特征数量增加.

4.4 实验结果讨论

在分类精度方面,相比较 Alpha-investing,OSFIC 在四种分类器上的平均分类精度提高了 4.41%. 原因是 Alpha-investing 未对冗余特征进行过滤,而 OSFIC 使用单条件互信息过滤掉部分冗余的新特征.

表 9 14 个数据集的算法运行时间比较

数据集	运行时间(s)		
	OSFIC	OSFS	SAOLA
allaml	7.2244	58.2798	1.5102
lung-cancer	**87.5207**	**249.4264**	**5.5198**
hiva	**3.6107**	**5.62×10^2**	**0.4667**
ionosphere	0.0225	0.2315	0.0096
lucap0	**58.6193**	**1.33×10^3**	**0.1055**
cina0	**3.8824**	**709.152**	**0.1222**
lung	**45.6285**	**317.5624**	**1.6644**
madelon	4.6037	11.9491	0.1005
marti1	0.0869	0.0871	0.0807
Prostate_GE	1.7430	7.7718	1.1015
reged1	**3.6922**	**106.1459**	**0.3021**
SMK_CAN_187	7.1018	37.6450	3.4865
Sylva	**0.6850**	**220.4977**	**0.1198**
wdbc	0.0370	0.1298	0.0121

表 10 OSFIC 和 OSFS 的运行时间比较

数据集	运行时间		Av_t	平均 Av_t
	t_{OSFIC}	t_{OSFS}		
lung-cancer	87.5207	249.4264	−0.6491	−0.9159
hiva	3.6107	5.62×10^2	−0.9936	
lucap0	58.6193	1.33×10^3	−0.9560	
cina0	3.8824	709.1520	−0.9945	
lung	45.6285	317.5624	−0.8563	
reged1	3.6922	106.1459	−0.9652	
Sylva	0.6850	220.4977	−0.9969	

OSFIC 与 OSFS 运行时间比率:$Av_t = (t_{OSFIC} - t_{OSFS})/t_{OSFS}$

在选择特征数量方面,相较于 OSFS 和 SAOLA,OSFIC 在 14 个数据集的平均特征选择数量分别降低了 20.6% 和 41.9%,原因是 SAOLA 无法移除多条件下的冗余特征,导致特征数较多;与 OSFS 相比,OSFIC 加入了互信息和条件独立性检验两种判断标准进行过滤,能够筛选出更多的冗余特征.

在运行时间方面,OSFS 直接对筛选出的相关特征进行"k-贪婪"搜索策略找寻特征子集,以此为条件进行冗余特征过滤,当强相关和低冗余特征不断增加,时间呈指数增长. 在此类特征集下,OSFIC 运行时间相比较 OSFS 减少了 91.59%. OSFIC 的运行时间高于 SAOLA,主要原因为 OSFIC 为筛选出更多的冗余特征而花费更长的时间. 对比以上三种算法,OSFIC 能够获得更高的分类精度,在较短的运行时间内减少选择特征的数量.

5 应用场景

选取 UCI 中 PEMS-SF 数据集作为应用场景. PEMS-SF 记录了 963 个传感器在 2008 年 1 月 1 日至 2009 年 9 月 30 日之间,旧金山湾区每一天的高速公路不同车道占用率(包含 440 个实例和 138672 个特征),每一个特征代表一个传感器在一天的车道占用率(介于 0 ~1 之间),分类标签为 1 ~7(代表星期一至星期日). 表 11 显示了 OSFIC、OSFS 和 SAOLA 在 PEMS-SF 数据集上的运行时间. 图中"——"表示 OSFS 在该数据集上存在较高的计算花费(超过 3 天),OSFIC 和 SAOLA 能够在有限时间内完成实例和特征数量较大数据集的筛选. OSFIC 经过 4 层过滤,能够移除大量冗余特征,保留极少部分特征,获得比 SAOLA 更少的特征数量.

表 11　三种算法的运行时间和特征数量比较

数据集	时间(s)	特征数量				
		特征占比	过滤 1	过滤 2	过滤 3	过滤 4
OSFIC	686	25%	14295	130	35	20
		50%	31461	149	40	24
		75%	46636	154	46	26
		100%	61618	158	47	26
OSFS	——	——				
SAOLA	50	25%	25			
		50%	34			
		75%	39			
		100%	37			

如图 5 所示,OSFIC 在四个分类器上的精度均大于 SAOLA,分别提升 4. 27%、2. 04%、5. 62% 和 1. 18%.

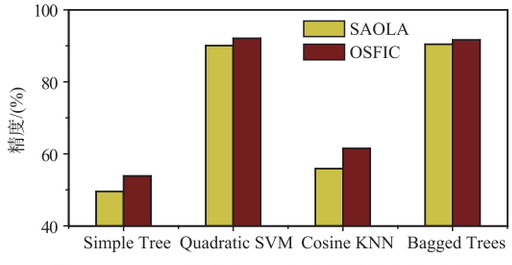

图5　PEMS-SF数据集在四种分类器的分类精度

6 总结

OSFIC 能找出分类标签的近似马尔可夫毯,并且在分类精度提升的同时,缩短运行时间,减少选择特征数量. 在特征间具有低冗余和高相关特性时,同样能够保持高效. 实验得出:(1)相比较已有的在线流特征选择算法,OSFIC 在分类精度上得到提升;(2)特征间具有低冗余高相关特征时,OSFIC 的运行时间得到了明显减少;(3)在保证分类精度的前提下,OSFIC 能够过滤较多的冗余特征和不相关特征;(4)OSFIC 可以找到分类标签的近似马尔可夫毯.

未来对以下内容展开研究:(1)如何更精确地获得分类标签的马尔可夫毯;(2)如何更有效地对多标签分类和多类分类下特征流开展高精度的特征选择;(3)如何发现被选特征与分类标签的因果关系.

参考文献

[1] Charu C Aggarwal. Data Classification:Algorithms and Applications[M]. 1th ed. Boca Raton:CRC press,2014. 37 –64.

[2] 乔立岩,彭喜元,彭宇. 基于微粒群算法和支持向量机的特征子集选择方法[J]. 电子学报,2006,34(3):496 –498.
Qiao Li-yan,Peng Xi-yuan,Peng Yu. BPSO-SVM wrapper for feature subset selection[J]. Acta Electronica Sinica,2006,34(3):496 –498. (in Chinese)

[3] Li J,Cheng K,Wang S,et al. Feature selection:A data perspective[J]. ACM Computing Surveys (CSUR),2018,50(6):94:1 –94:45.

[4] Cai J,Luo J,Wang S,et al. Feature selection in machine learning:A new perspective[J]. Neurocomputing,2018,300(7):70 –79.

[5] Zhang Q,Zhang P,Long G,et al. Online learning from trapezoidal data streams[J]. IEEE Trans,2016,KDE-28(10):2709 –2723.

[6] Wu X,Yu K,Ding W,et al. Online feature selection with streaming features[J]. IEEE Trans,2013,PAMI-35(5):1178 –1192.

[7] Kohavi R,John G H. Wrappers for feature subset selection[J]. Artificial intelligence,1997,97(1 –2):273 –324.

[8] Loscalzo S,Yu L,Ding C. Consensus group stable feature selection[A]. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining[C]. New York:ACM press,2009. 567 –576.

[9] Rodriguez-Lujan I,Huerta R,Elkan C,et al. Quadratic programming feature selection[J]. Journal of Machine Learning Research,2010,11(1):1491 –1516.

[10] Zhao Z,Liu H. Searching for interacting features in subset selection[J]. Intelligent Data Analysis,2009,13(2):207 –228.

[11] 吴信东,何进,陆汝钤. 从大数据到大知识:HACE + BigKE[J]. 自动化学报,2016,42(7):965 –982.
Wu Xin-dong,He Jing,Lu Ru-qian. From big data to big knowledge:HACE + BigKE[J]. Acta Automatica Sinica,2016,42(7):965 –982. (in Chinese)

[12] Yu K,Wu X,Ding W,et al. Scalable and accurate online feature selection for big data[J]. ACM Trans,2016,KDD-11(2):16:1 –16:39.

[13] 张昊,陶然,李志勇. 基于 KNN 算法及禁忌搜索算法的特征选择方法在入侵检测中的应用研究[J]. 电子学报,2009,37(7):1628 – 1632.
Zhang Hao,Tao Ran,Li Zhi-yong. A research and application of feature selection based on KNN and tabu search algorithm in the intrusion detection[J]. Acta Electronica Sinica,2009,37(7):1628 – 1632. (in Chinese)
[14] Jia X,Kuo B C,Crawford M M. Feature mining for hyperspectral image classification[J]. Proceedings of the IEEE,2013,101(3):676 – 697.
[15] Wang D,Irani D,Pu C. Evolutionary study of web spam: Webb spam corpus 2011 versus webb spam corpus 2006 [A]. Proceedings of the 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)[C]. Pittsburgh: IEEE, 2012. 40 – 49.
[16] Singh N ,Lai K H ,Vejvar M ,et al. Big data technology: Challenges,prospects and realities[J]. IEEE Engineering Management Review,2019,47(1):58 – 66.
[17] Wu X,Zhu X,Wu G Q,et al. Data mining with big data [J]. IEEETrans,2014,KDE-26(1):97 – 107.
[18] Kira K,Rendell L A . The Feature Selection Problem:Traditional Methods and a New Algorithm[A]. Proceedings of the Tenth National Conference on Artificial Intelligence [C]. San Jose:AAAI Press,1992. 129 – 134.
[19] Peng H,Long F,Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance,and min-redundancy[J]. IEEE Trans,2005,PAMI-27(8):1226 – 1238.
[20] Yu K,Ding W,Simovici D A,et al. Classification with streaming features:An emerging-pattern mining approach [J]. ACM Trans,2014,KDD-9(4):30:1 – 30:31.
[21] Mairal J,Bach F,Ponce J,et al. Online learning for matrix factorization and sparse coding[J]. Journal of Machine Learning Research,2010,11(1):19 – 60.
[22] Wang J,Zhao P,Hoi S C H,et al. Online feature selection and its applications[J]. IEEE Trans,2014,KDE-26(3): 698 – 710.
[23] Perkins S,Theiler J. Online feature selection using grafting [A]. Proceedings of the 20th International Conference on Machine Learning (ICML)[C]. Washington DC:AAAI Press,2003. 592 – 599.
[24] Zhou J ,Foster D P ,Stine R A ,et al. Streamwise Feature Selection [J]. Journal of Machine Learning Research, 2006,7(9):1861 – 1885.
[25] Wang J,Wang M,Li P,et al. Online feature selection with group structure analysis[J]. IEEE Trans,2015,KDE-27 (11):3029 – 3041.
[26] Yu K,Wu X,Ding W,et al. Scalable and accurate online feature selection for big data[J]. ACM Trans,2016, KDD-11(2):16:1 – 16:39.
[27] Aliferis C F,Statnikov A,Tsamardinos I,et al. Local causal and markov blanket induction for causal discovery and feature selection for classification part i:Algorithms and empirical evaluation[J]. Journal of Machine Learning Research,2010,11(1):171 – 234.
[28] Koller D. Toward optimal feature selection[A]. Proceedings of the Proceedings of the 20th International Conference on Machine Learning (ICML)[C]. Washington DC:AAAI Press,1996. 284 – 292.
[29] Neapolitan R E. Learning Bayesian Networks[M]. 1th ed. Upper Saddle River,NJ:Pearson Prentice Hall,2004. 600 – 604.
[30] Press W H,Teukolsky S A,Vetterling W T,et al. Numerical Recipes in C[M]. Cambridge:Cambridge university press,1992. 632 – 636.

作者简介

尤殿龙 男,1981 年 1 月生,内蒙古赤峰人,博士,副教授. 主要研究方向:流特征选择和因果发现,图信号处理,数据挖掘、人工智能、知识工程.
E-mail:youdianlong@ sina. com

郭 松 男,1994 年 5 月生,河北邯郸人,硕士研究生. 主要研究方向:流特征选择,人工智能.
E-mail:guosongysu@ 163. com

赵春慧 女,1994 年 4 月生,山西阳泉人,硕士研究生. 主要研究方向:数据挖掘,推荐系统.
E-mail:zch635134004@ 163. com

原福永(通信作者) 1958 年 12 月,黑龙江齐齐哈尔人,教授. 主要研究方向:数据挖掘,推荐系统.
E-mail:fyyuan@ ysu. edu. cn