计算机科学
COMPUTER SCIENCE

基于持续同调的过滤式特征选择算法

殷杏子, 彭宁宁, 詹学燕

引用本文

殷杏子, 彭宁宁, 詹学燕.基于持续同调的过滤式特征选择算法[J]. 计算机科学, 2023, 50(6): 159-166.

YIN Xingzi, PENG Ningning, ZHAN Xueyan. Filtered Feature Selection Algorithm Based on Persistent Homology [J]. Computer Science, 2023, 50(6): 159-166.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220500169

基于持续同调的过滤式特征选择算法

殷杏子　彭宁宁　詹学燕
武汉理工大学理学院　武汉 430070
(305849@whut.edu.cn)

摘　要　现有的过滤式特征选择算法忽略了特征之间的关联性。鉴于此,提出了一种新的过滤式特征选择算法——基于持续同调的特征选择算法(Rel-Betti 算法),该算法能够识别特征之间的关联性以及组合效果。通过提出相关贝蒂数概念,筛选出数据集中重要的拓扑特征信息。该算法对数据集进行预处理后,根据类标签将数据集分类,计算不同类中的相关贝蒂数,获得数据信息的特征均值,按特征均值差值大小对特征进行重要性排序。利用 UCI 数据集中的 8 个数据,将该算法与其他常见算法在决策树、随机森林、K 近邻和支持向量机这 4 种学习模型下进行比较实验。结果表明,该算法是一种有效的特征选择算法,其能够提高分类的准确率和 F1 值,并且不依赖于特定的机器学习模型。

关键词:特征选择;持续同调;条形码;贝蒂数;机器学习
中图法分类号　O189.22

Filtered Feature Selection Algorithm Based on Persistent Homology

YIN Xingzi, PENG Ningning and ZHAN Xueyan
School of Science, Wuhan University of Technology, Wuhan 430070, China

Abstract　The existing filtering feature selection algorithm ignores the correlation between features. This paper proposes a new filtering feature selection algorithm —the feature selection algorithm based on persistent homology(Rel-Betti algorithm), which can consider the correlation between features and the combined effect. This paper gives a new definition named by relevant Betti numbers, which can filter out the important topological features in the dataset. The algorithm first preprocesses the data set, classifies the data set according to the class labels, calculates the relevant Betti numbers in different classes, obtains the feature mean of the data information, and uses the feature mean difference to sort the importance of the features. Using eight data in UCI, the algorithm is compared with other common algorithms under four learning models: decision tree, random forest, K-nearest neighbor and support vector machine. Experimental results show that the Rel-Betti algorithm is an effective method that can improve classification accuracy and F1 value, and does not depend on a specific machine learning model.

Keywords　Feature selection, Persistent homology, Barcode, Betti number, Machine learning

1　引言

在大数据时代,特征选择是对数据进行预处理的必要环节。特征选择作为一种数据降维技术,其主要目的是从原始数据中选择出重要的特征,排除冗余与不相关的特征,降低数据的维度和学习任务的难度,进而提升机器学习模型的效率。根据与分类器的关系,特征选择算法被划分为过滤式(Filter)、嵌入式(Embedded)和包裹式(Wrapper) 3 类[1-2]。过滤式方法独立于分类器,先利用特征选择过程对初始特征值进行过滤,然后利用过滤后所得特征值来训练模型。嵌入式方法将机器学习模型的训练过程与特征选择过程融为一体,其中最典型的是决策树算法。包裹式的方法通过不断从初始特征集合中选择出特征子集来训练学习模型,并根据学习模型的性能对子集进行评价,直至筛选出最佳子集,该特征选择方法直接针对给定学习模型进行优化。3 种方法中过滤式特征算法最为常见。虽然过滤式特征选择算法根据传统统计学的度量方式能够快速比较特征的重要性,但其仅仅考虑了每一个特征与目标变量的相关性,忽略了不同特征之间的关联性及组合效果,并且过滤式特征选择算法只能消除不相关的特征,无法消除冗余的特征[3]。

鉴于此,本文将持续同调运用于特征选择。本文的主要贡献如下:

(1)针对传统过滤式方法忽略不同特征之间组合效果的不足,提出了一种考虑不同特征之间关联性和组合效果,且

到稿日期:2022-05-18　返修日期:2022-08-11
基金项目:国家自然科学基金(11701438)
This work was supported by the National Natural Science Foundation of China(11701438).
通信作者:彭宁宁(pengn@whut.edu.cn)

能够消除冗余特征的 Rel-Betti 算法。

(2)在 8 个公共数据集上进行实验验证，相较于卡方检验、F 检验和互信息这 3 种特征选择方法，本文提出的 Rel-Betti 算法能够显著提高分类的准确率和 F1 值，同时也证明了 Rel-Betti 算法可以结合多种机器学习算法，不局限于某一种单一的机器学习模型。

2 相关工作

过滤式方法通过对所有特征进行评分，以分值大小排序来判断特征的重要程度[4]。常用的方法有卡方检验、F 检验、互信息，以及皮尔逊相关性分析等。过滤式方法的时间复杂度低，可以快速缩小特征集规模。它们通常只考虑单个特征的预测能力，忽略了特征与特征之间的相关性，无法消除冗余特征，通常不能得到一个规模较小的优化特征子集。为解决此问题，Ji 等[5]提出了先用 Filter 方法进行特征预选，再用 Wrapper 方法做进一步的特征选择的二阶段组合式特征选择算法。Xu 等[6]提出了一种基于组策略的 MRMR 改进算法。该算法引入的组策略考虑了特征之间的相关性，首先，基于信息熵对所有特征进行排序分块；其次，以划分的特征块作为一个整体特征，采用 CCA 计算与类标签的相关度和特征块间的冗余度，对所划分的数据块进行排序；最后根据用户需求选择排序后的特征块，再次利用 MRMR 算法进行二次降维。Singh 等[7]提出了一种四步混合集成特征选择算法。该算法首先采用交叉验证对数据集分区；然后综合各种基于加权分数的过滤方法生成特征排序；之后利用顺序前向选择算法获得最优特征子集；最后将最优子集用于后续的分类任务。该算法还被应用于医疗数据。Priscilla 等[8]提出了一种在第一阶段采用互信息，在第二阶段使用递归特征消除(RFE)来消除冗余特征的两阶段特征选择方法。Pashaei 等[9]使用最小冗余最大相关度(mRMR)作为第一级过滤器，然后将模拟退火和交叉算子引入二元算术优化算法中以选择最小信息基因集。为规避互信息的局限性，Yang[10]引入了 RReliefF 算法来度量特征与标签的相关性，且引入最大互信息系数来度量特征与标签的相关性以及特征与特征之间的冗余性。

已有文献提出利用混合集成的特征选择方法来考虑特征与特征之间的相关性，消除冗余特征，从而达到提高分类精度和降低计算成本的目的。本文将持续同调应用于特征选择，提出了一种新的非集成的过滤式特征选择方法。

持续同调(Persistent Homology，PH)将几何与拓扑学联系在一起。点云在过滤过程中会产生一系列不同尺度的单纯复形，持续同调就是研究过滤复形中存在的拓扑不变量。这些拓扑不变量被称作贝蒂数[11]。研究发现，一些贝蒂数在单纯复形中“存活”的时间更长，而另一些贝蒂数在过滤值改变时“死亡”的速度更快。这些贝蒂数的“寿命”或“持续时间”与几何性质直接相关，通过条形码来表示其变化过程。2004 年，Carlsson 等[12]利用持续同调提出并研究了拓扑数据分析方法，并从理论上表明了该方法适用于大型数据集。2009 年，Carlsson[13]研究了如何应用几何和拓扑对不同种类数据进行分析处理。2014 年，Lockwood 等[14]提出了一种基于持续同调的癌症基因表达数据挖掘新方法。该方法选择一小部分基因作为地标来构建拓扑结构，从而捕获数据集中持续的、具有拓扑意义的特征子集。2018 年，Cang 等[15]通过构造持续同调过程中的过滤距离矩阵验证了 PH 模型不仅可以保存关键的化学和生物信息，还可以对生物分子的相互作用进行有效的拓扑描述。2018 年，Chi 等[16]对基于持续同调的机器学习(PHML)模型展开综述，并讨论了其在蛋白质结构分类中的应用。

3 相关理论

3.1 Vietoris-rips 复形

n 维单形是欧氏空间中 $n+1$ 个点 $\{a_0, a_1, \cdots, a_n\}(n \geqslant 0)$ 的凸包。如常见的 0 维单形用顶点表示，一条直线表示 1 维单形，三角形表示 2 维单形拓扑结构。单纯复形是若干单形的集簇，且需要满足下面两个条件：1)其中任意一个单形的子单形仍属于这个集簇；2)其中任意两个单形如果有非空交集，则它们的交仍是一个单形。

定义 1 令 (X, d) 为一个度量空间，给定参数 $\varepsilon > 0(\varepsilon \in R)$，$B(x, \varepsilon)$ 表示以 x 为球心、ε 为半径的闭球，则 Vietoris-Rips 复形 $VR(x, \varepsilon)$ 定义如下：

$$VR(x, \varepsilon) = \{\langle \alpha_0, \alpha_1, \cdots, \alpha_k \rangle \mid B(\alpha_i, \varepsilon) \cap B(\alpha_j, \varepsilon) \neq \varnothing, \alpha_i, \alpha_j \in X, 0 \leqslant i, j \leqslant k\} \tag{1}$$

图 1 是由不同的参数半径 ε 构成的 Vietoris-Rips 复形(简称 Rips 复形)。参数 ε 的选择十分重要，ε 太小，点云是一个离散的点集；ε 太大，生成的 Rips 复形是一个单一的高维复形。Rips 复形是应用中使用最为广泛的单纯复形，具有构造简单、计算速度快、能够处理复杂的高维数据等优点[17]。

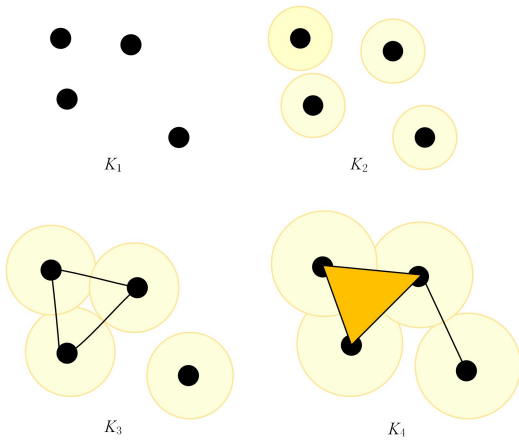


图 1 Vietoris-Rips 复形

Fig. 1 Vietoris-Rips complex

3.2 持续同调

令 K 为一个原始的单纯复形，$H_q(K)$ 为 K 的 q 维同调群，可在某一时刻添加其子单纯复形，构建变化的复形过滤流，建立单纯复形间的嵌套关系。

$$K_0 \subset K_1 \subset K_2 \subset \cdots \subset K_n = K \tag{2}$$

因为 $K_{i-1} \subset K_i$，所以在单纯复形之间引入同态 f：$H_q(K_{i-1}) \subset H_q(K_i)$，使嵌套的单纯复形序列与同态连接的同调群序列一一对应，每个维度对应一个不同的 q，则同调群的序列如下：

$$H_q(K_0) \rightarrow H_q(K_1) \cdots \rightarrow H_q(K_n) = H_q(K) \tag{3}$$

将 q 维同调群 $H_q(K)$ 的秩记为 b_q，那么 $b_q = rank(H_q(K))$，并将 b_q 称为单纯复形 K 的 q 维贝蒂(Betti)数。从几何意义上来讲，b_0 表示复形中连通分支的个数，b_1 表示复形中孔洞的个数。持续同调就是在构建复形过滤流中记录 K_i 到 K_j 持续存在的拓扑特征，即贝蒂数。

3.3 条形码

条形码常被用来可视化贝蒂数的持续性，其优点在于能够定性地过滤掉拓扑噪声并捕获重要特征[18]。贝蒂数以一条直线，即区间的形式出现在条形码图中，直线的起点表示贝蒂数的出生时间，终点表示贝蒂数的死亡时间，通常称区间长度为贝蒂数的寿命。观察 1 维条形码图中不同贝蒂数的持续时间，持续时间短的贝蒂数是噪音，持续时间长的贝蒂数反映了数据中重要的拓扑特征。下面以图 2 为例来介绍条形码图表示贝蒂数的过程。可以看出，随着 ε 增大，贝蒂数也会发生改变。

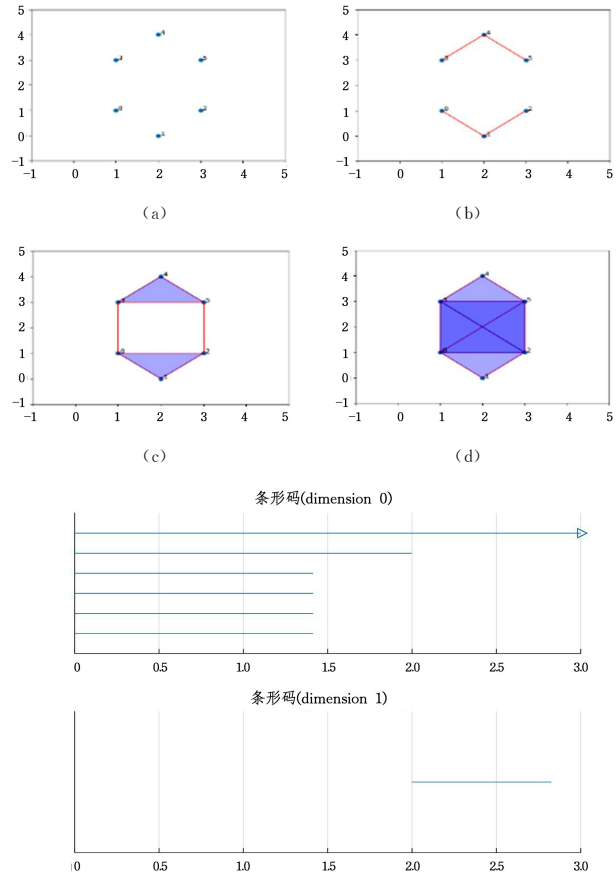


图 2 条形码

Fig. 2 Barcode

当 $0 < \varepsilon < 1.414$ 时，在图 2(a)中生成了点云，$b_0 = 6$，$b_1 = 0$，因此 0 维条形码有 6 条直线，1 维条形码没有直线。

当 $1.414 < \varepsilon < 2$ 时，在图 2(b)中，一些数据点之间形成了边，b_0 减少，有两个连通分支，故 $b_0 = 2$。此时不存在 1 维特征“洞”，故 $b_1 = 0$。因此 0 维条形码有两条直线，1 维条形码没有直线。

当 $2 < \varepsilon < 2.828$ 时，在图 2(c)中，全部点连接在一起，即 $b_0 = 1$，且形成了 1 个一维特征“洞”，故 $b_1 = 1$。因此 0 维条形码有一条直线，1 维条形码中也有一条直线。

当 $\varepsilon > 2.828$ 时，在图 2(d)中，$b_0 = 1$，则 0 维条形码中只有一条直线持续存在，一维“洞”消失，即 $b_1 = 0$，1 维条形码中的直线死亡。

4 Rel-Betti 算法

持续同调在复杂网络、图像分类、基因检测等领域取得了一系列的研究成果。本文运用持续同调的理论对高维数据构建复形，首先根据比率大小给出新概念相关贝蒂数(Relevant Betti Number)，利用其挑选重要的数据信息，从而对特征进行重要性排序，我们称之为 Rel-Betti 算法，算法研究框架如图 3 所示。

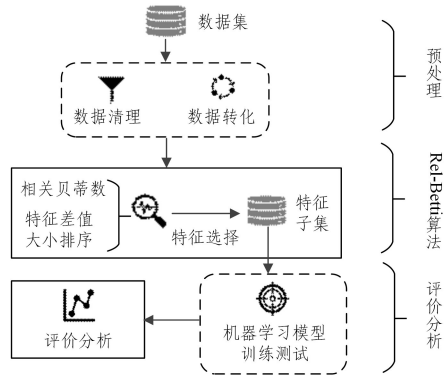


图 3 研究结构框架

Fig. 3 Research structural framework

4.1 相关贝蒂数

定义一维条形码 $B_1 = \{\beta_0, \beta_1, \cdots, \beta_n\}$，其中 $\beta_i = [m_i, n_i)$，m_i 表示贝蒂数 β_i 的出生过滤值，n_i 表示贝蒂数 β_i 的死亡过滤值。

首先，找到 1 维条形码中持续时间最长，即寿命最长的贝蒂数 $\beta_{\max}$，它是一维空间中非常重要的贝蒂数，表示数据集中最显著的拓扑特征。

$$\beta_{\max} = \max_{\beta_i \in B_1}(n_i - m_i) \tag{4}$$

其次，相关贝蒂数 β_{rel} 及其个数 $N_{\beta_{\mathrm{rel}}}$ 定义如下：

$$\beta_{\mathrm{rel}} \geqslant \beta_{\max} \times ratio \tag{5}$$

$$N_{\beta_{\mathrm{rel}}} = \sum_{\beta_i \in B_1} f(n_i - m_i, \beta_{\max}, ratio)$$

$$f(n_i - m_i, \beta_{\max}, ratio) = \begin{cases} 1, & \text{如果 } n_i - m_i \geqslant ratio \cdot \beta_{\max} \\ 0, & \text{其他} \end{cases} \tag{6}$$

适当调整比率大小，得到的相关贝蒂数 β_{rel} 的个数不同。实验结果表明比率不同不影响特征的重要性排序。最后，由于相同生存时间区间的相关贝蒂数中的数据非常相似且可能

重复，因此若一个相关贝蒂数的寿命区间 $\beta_i'=[m_i',n_i')\subseteq[m_i,n_i)=\beta_i$，则舍弃贝蒂数 β_i' 的数据信息，只选择贝蒂数 β_i 中的数据信息。

4.2 特征平均值的差值

设分类型数据集 $X=\{x_1,x_2,\cdots,x_n\}$ 有 n 个样本，其中每个样本点表示为 $x_i=\{x_{i1},\cdots,x_{im}\}$，数据集的特征为 $\{A_1,A_2,\cdots,A_m\}$。

根据类标签将数据集 X 中的 n 个样本分成 X_1 和 X_2 两类，分别在 X_1 和 X_2 上构建 Rips 复形，计算 Rips 复形中显著的寿命最长的贝蒂数 β_{max} 和相关贝蒂数 β_{rel}，输出相关贝蒂数 β_{rel} 中存在的样本数据信息 $x_i=\{x_{i1},x_{i2},\cdots,x_{im}\}$。根据样本数据信息分别计算 X_1 和 X_2 每个特征 A_i 的平均值 A_{i1} 和 A_{i2}，从而计算 X_1 和 X_2 每个特征平均值的差值(Difference of Mean，DM) $DM_i=(A_{i1}-A_{i2})$，按照 DM_i 的大小对数据集特征进行重要性排序。特征均值差值越小，说明该特征在两个分类中越相似，起到的分类作用越小；特征均值差值越大，说明该特征起到的分类作用越大。

4.3 Rel-Betti 算法流程

本算法采用斯坦福大学拓扑计算小组基于 PLEX 库开发的软件包 Java Plex 进行计算。Rel-Betti 算法的流程如算法 1 所示。

算法 1 Rel-Betti 算法

输入：数据集 $X=\{x_1,x_2,\cdots,x_n\}$，过滤半径参数 $\varepsilon_1,\varepsilon_2$ 对应数据类别 X_1,X_2；相关贝蒂数个数 $N_{\beta_{rel}}$ 的比率 ratio

输出：每个特征 A_i 均值差值 DM_i

1. 利用 Min-Max 标准化 $X'=\dfrac{X-X_{min}}{X_{max}-X_{min}}$ 对数据集 X 进行归一化处理；
2. 根据类标签将数据集 $X=\{x_1,x_2,\cdots,x_n\}$ 分成 X_1,X_2；
3. 根据式(1)分别在两类 X_1,X_2 上构建 Rips 复形；
4. 计算 rips 复形的 1 维同调群 $H_1(K)$，得到 X_1,X_2 中 1 维的所有贝蒂数区间 $[m_i,n_i]$；
5. 初始化，存储贝蒂数区间；
6. 比较贝蒂数长度，找到 β_{max}；

```
for 0 到 n
  if max life<n_i-m_i
/* max life 为最长贝蒂数寿命 */
    max life=n_i-m_i
  end
end
```

7. 输出相关贝蒂数区间；

```
for 0 到 n
  if life>max life×ratio
/* life 表示相关贝蒂数寿命 */
    return
  end
end
```

8. 返回 X_1,X_2 的样本数据 $x_i=\{x_{i1},x_{i2},\cdots,x_{im}\}$；
9. 计算 X_1,X_2 的特征均值差值 $DM_i=(A_{i1}-A_{i2})$；
10. 返回每个特征 A_i 均值差值 DM_i。

5 实验与结果分析

本文实验环境为 Inter Core i5-10210U，CPU@1.60GHz，8GB 内存，Windows10 64 位操作系统。

5.1 实验数据集

本文使用 UCI 机器学习库的 8 个公共数据集 iris，breast-cancer，heart-cancer，Parkinson，bone，Autistic Spectrum Disorder Screening Data for Children(ASDSDC)，arrhythmia 以及 ad 进行对比实验。表 1 列出了这些数据集的具体信息。

表 1 样本数据集信息

Table 1 Sample dataset information

数据集	特征个数	类别	样本总数	样本数/类	类型
iris	4	3	150	50,50,50	连续型
breast-cancer	9	2	286	201,85	混合型
heart-cancer	13	2	303	164,139	混合型
Parkinson	22	2	195	48,147	连续型
bone	6	2	310	210,100	连续型
ASDSDC	20	2	290	150,140	混合型
arrhythmia	260	2	451	206,245	混合型
ad	1555	2	3278	458,2820	离散型

5.2 实验设计

为验证本文 Rel-Betti 算法的有效性，在决策树(Decision Tree，DT)、随机森林(Random Forest，RF)、K 近邻(K-Nearest Neighbors，KNN)、以及支持向量机(Support Vector Machines，SVM)这 4 种学习模型下进行了 4 组实验。每组实验中，将 Rel-Betti 算法与卡方检验[19]、F 检验[20]、互信息[21]进行比较，对比特征个数相同时各算法的性能。

为了防止过拟合现象，保证结果的可靠性，本文实验使用 5 次交叉验证的方法。数据集中训练集占样本总量的 80%，测试集占 20%。

5.3 评价指标

评价机器学习模型是利用机器学习解决实际问题过程中极其重要的一步。准确率(accuracy)、精确率(precision)和召回率(recall)是分类算法中最常用的评价指标。

$$accuracy=\frac{TP+TN}{TP+TN+FP+FN}$$

$$precision=\frac{TP}{TP+FP}$$

$$recall=\frac{TP}{TP+FN}$$

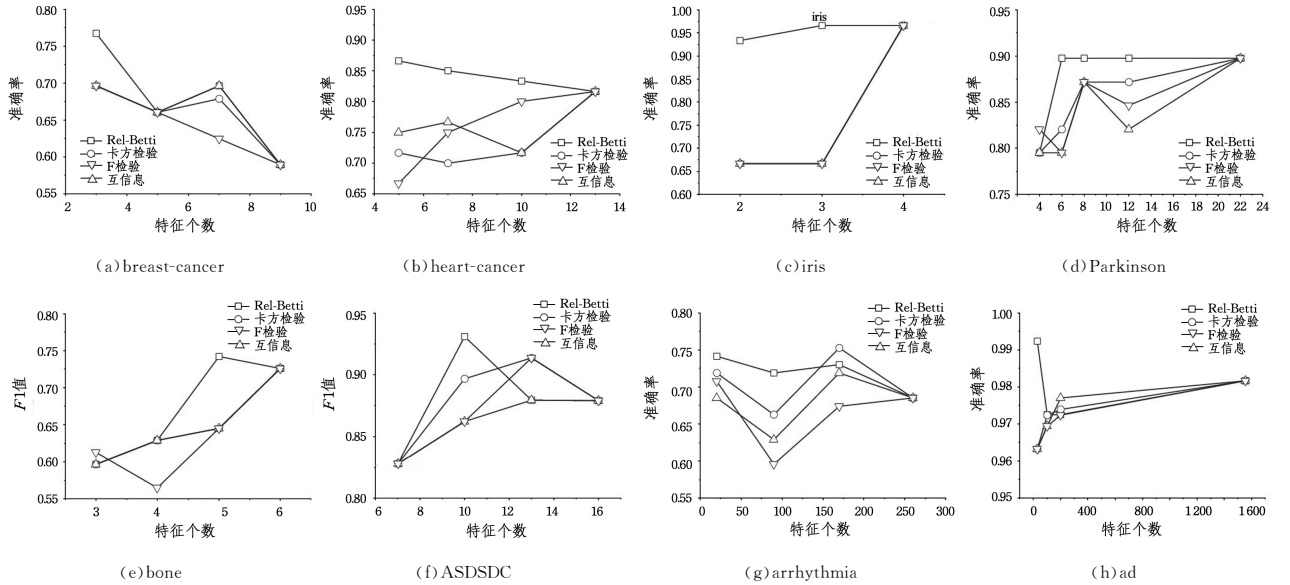
TP 指分类器把正例样本预测为正例的数量，FP 指分类器把负例样本预测为负例的数量，FN 指分类器把负例样本预测为正例的数量。

由于精确率和召回率是相互的，单纯提升其中一个性能可能会导致另一个性能的下降[22]。为平衡精确率和召回率，有效地评价特征选择算法的性能，本实验采用 F1 值度量作为分类性能指标，定义如下：

$$F1=\frac{2\times precision\times recall}{precision+recall}$$

5.4 实验结果

第1组实验是在DT学习模型下，为直观地观察不同特征选择算法之间的性能差异，本文分别在8个数据集上进行实验，并绘制出了不同算法所选特征在DT下的准确率值，如图4所示；以及不同算法所选特征在DT下的F1值，如图5所示。



(a)breast-cancer (b)heart-cancer (c)iris (d)Parkinson

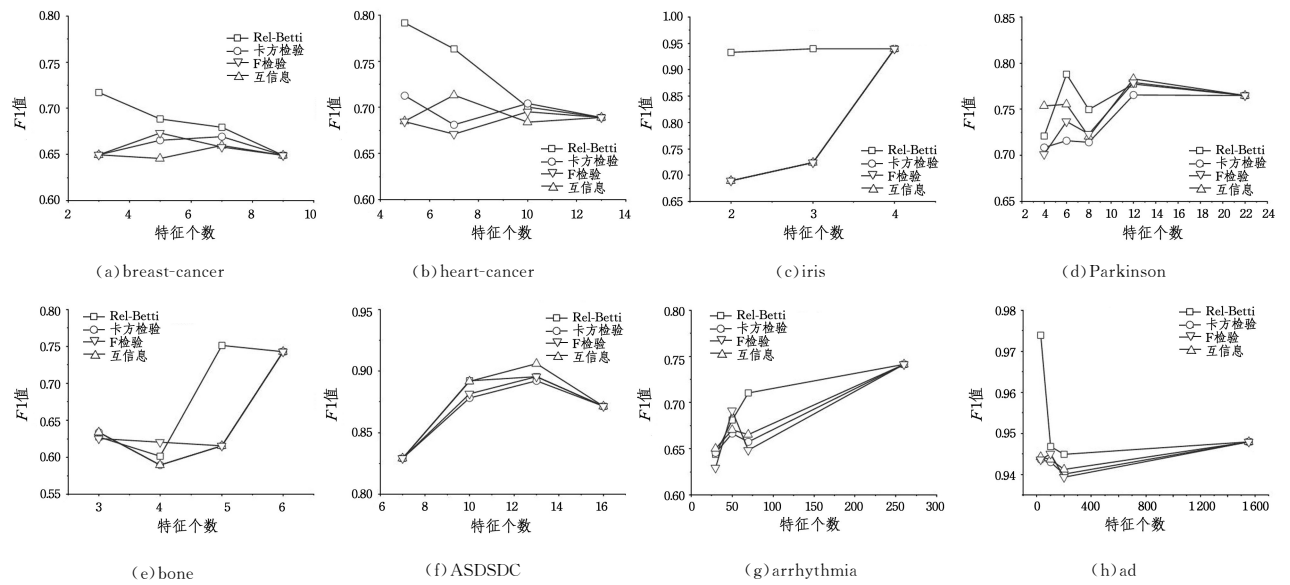
(e)bone (f)ASDSDC (g)arrhythmia (h)ad

图4 决策树下8个数据集的准确率

Fig. 4 Accuracy of eight datasets under decision tree

从图4的实验结果能够看出，Rel-Betti算法在不同的数据集上的准确率普遍优于其他3种方法，具体而言：相较于次优的方法，本文方法在数据集breast-cancer上取3个特征时准确率提高了7.1%，在heart-cancer上取5个特征时提高了3.4%；在iris上取2个特征时提高了27%，在Parkinson上在取6个特征时提高了7.7%，在bone上在取5个特征时提高了9.6%，在ASDSDC上取10个特征时提高了3.5%，在arrhythmia上在取70个特征时提高了2.2%，在ad上在取30个特征时提高了3%。

从图5的实验结果能够看出，本文方法在不同的数据集上的F1值普遍优于其他3种方法，具体而言：相较于次优的方法，本文方法在数据集breast-cancer上取3个特征时F1值提高了6.6%，在heart-cancer上取5个特征时提高了7.9%，在Parkinson上取6个特征时提高了3.2%，在iris上取2个特征时提高了23.7%，在bone上在取5个特征时提高了10.8%，在ASDSDC上在取10个特征时提高了0.1%，在arrhythmia上在取70个特征时提高了5.3%，在ad上在取30个特征时提高了3%。第2－4组实验结果是在8个数据集上，RF，KNN以及SVM模型下的准确率和F1值，分别如表2和表3所列。



(a)breast-cancer (b)heart-cancer (c)iris (d)Parkinson

(e)bone (f)ASDSDC (g)arrhythmia (h)ad

图5 决策树下8个数据集的F1值

Fig. 5 F1 value of eight datasets under decision tree

表 2 8 个数据集中不同特征个数的准确率

Table 2 Accuracy of different numbers of features in eight datasets

数据集	特征个数	RF 下的 Accuracy				KNN 下的 Accuracy				SVM 下的 Accuracy			
		Rel-Betti	卡方检验	F 检验	互信息	Rel-Betti	卡方检验	F 检验	互信息	Rel-Betti	卡方检验	F 检验	互信息
breast-cancer	3	**0.77**	0.70	0.70	0.70	**0.77**	0.70	0.70	0.70	**0.77**	0.70	0.70	0.70
	5	**0.71**	0.66	**0.71**	0.70	**0.77**	0.75	0.68	**0.77**	**0.77**	0.77	0.68	**0.77**
	7	**0.70**	**0.70**	0.68	**0.70**	**0.77**	0.75	**0.77**	**0.77**	**0.77**	0.75	**0.77**	0.71
	9	0.70	0.70	0.70	0.70	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73
heart-cancer	5	**0.88**	0.72	0.78	0.82	**0.80**	0.75	0.77	0.67	0.80	0.75	**0.82**	0.68
	7	**0.90**	0.77	0.80	0.77	**0.85**	0.80	0.85	0.78	**0.83**	0.78	0.82	**0.83**
	10	**0.88**	0.78	0.83	0.78	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**
	13	0.88	0.88	0.88	0.88	0.80	0.80	0.80	0.80	0.88	0.88	0.88	0.88
iris	2	**0.93**	0.53	0.53	0.53	**0.97**	0.67	0.67	0.67	**0.97**	0.60	0.60	0.60
	3	**0.97**	0.56	0.56	0.56	**0.97**	0.63	0.63	0.63	**0.97**	0.67	0.67	0.67
	4	0.97	0.97	0.97	0.96	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
Parkinson	4	**0.90**	0.82	0.82	0.82	**0.87**	**0.87**	**0.87**	**0.87**	0.87	0.85	0.85	0.85
	6	**0.90**	0.85	0.85	0.82	0.87	**0.90**	**0.90**	**0.90**	**0.87**	0.85	0.85	0.85
	8	**0.95**	0.87	0.87	0.87	**0.97**	0.87	0.87	0.87	**0.90**	0.85	0.85	0.85
	12	0.87	**0.92**	0.90	0.90	**0.97**	0.90	0.90	0.90	**0.87**	0.85	0.85	0.85
	22	0.85	0.85	0.85	0.85	0.90	0.90	0.90	0.90	0.92	0.92	0.92	0.92
bone	3	0.65	0.66	**0.66**	0.69	0.69	0.69	0.69	0.69	0.71	0.71	0.71	0.71
	4	**0.69**	0.69	0.69	0.68	**0.68**	0.68	0.68	0.68	**0.68**	0.68	0.68	0.66
	5	**0.74**	0.69	0.69	0.69	**0.82**	0.66	0.66	0.66	**0.76**	0.68	0.68	0.68
	6	0.76	0.76	0.76	0.76	0.77	0.77	0.77	0.77	0.76	0.76	0.76	0.76
ASDSDC	7	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.83	0.83	0.83	0.83
	10	**0.91**	0.90	0.88	0.90	**0.91**	0.86	0.86	0.86	**0.95**	0.90	0.90	0.90
	13	0.95	0.91	**0.98**	0.95	**0.90**	0.88	0.88	0.88	0.97	**0.98**	0.98	0.98
	16	0.95	0.95	0.95	0.95	0.90	0.90	0.90	0.90	0.98	0.98	0.98	0.98
arrhythmia	30	0.74	0.73	**0.75**	**0.75**	**0.67**	0.65	0.65	0.65	**0.71**	0.67	0.67	0.67
	50	**0.80**	0.79	**0.80**	**0.80**	0.63	**0.67**	**0.67**	**0.67**	**0.79**	0.76	0.76	0.76
	70	**0.81**	0.79	0.80	**0.81**	0.64	**0.71**	**0.71**	**0.71**	0.78	0.78	0.78	0.78
	260	0.84	0.84	0.84	0.84	0.56	0.56	0.56	0.56	0.78	0.78	0.78	0.78
ad	30	**0.99**	0.96	0.96	0.96	**0.99**	0.97	0.97	0.97	**0.99**	0.96	0.96	0.96
	100	0.97	0.97	0.97	0.97	0.95	0.96	0.96	0.96	0.97	0.97	0.97	0.97
	200	0.97	0.97	0.97	0.97	0.95	**0.97**	**0.97**	**0.97**	**0.98**	0.97	0.97	0.97
	1555	0.98	0.98	0.98	0.98	0.96	0.96	0.96	0.96	0.98	0.98	0.98	0.98

表 3 8 个数据集中不同特征个数的 F1 值

Table 3 F1 values of different numbers of features in eight datasets

数据集	特征个数	RF 下的 F1 值				KNN 下的 F1 值				SVM 下的 F1 值			
		Rel-Betti	卡方检验	F 检验	互信息	Rel-Betti	卡方检验	F 检验	互信息	Rel-Betti	卡方检验	F 检验	互信息
breast-cancer	3	**0.72**	0.65	0.65	0.65	**0.72**	0.64	0.64	0.64	**0.72**	0.65	0.65	0.65
	5	**0.69**	0.67	0.65	0.67	**0.69**	0.66	0.65	0.69	**0.68**	0.67	0.63	0.68
	7	**0.68**	0.67	0.66	0.66	**0.68**	0.67	0.65	0.66	**0.68**	0.68	0.67	0.64
	9	0.65	0.65	0.65	0.65	0.67	0.67	0.67	0.67	0.66	0.66	0.66	0.66
heart-cancer	5	**0.79**	0.75	0.73	0.75	**0.79**	0.75	0.76	0.74	**0.79**	0.78	0.78	0.73
	7	**0.79**	0.72	0.75	0.78	**0.77**	0.74	0.75	0.76	**0.79**	0.74	0.77	0.79
	10	**0.79**	0.75	0.76	0.76	**0.79**	0.75	0.75	0.75	**0.79**	0.77	0.77	0.77
	13	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
iris	2	**0.93**	0.69	0.69	0.69	**0.95**	0.64	0.64	0.64	**0.95**	0.71	0.71	0.71
	3	**0.94**	0.72	0.72	0.72	**0.94**	0.74	0.74	0.74	**0.95**	0.79	0.79	0.79
	4	0.94	0.94	0.94	0.94	0.95	0.95	0.95	0.95	0.94	0.94	0.94	0.94
Parkinson	4	**0.77**	0.76	0.75	0.75	**0.82**	0.74	0.74	0.74	**0.80**	0.74	0.74	0.74
	6	**0.78**	0.76	0.76	0.75	**0.81**	0.80	0.80	0.80	**0.79**	0.75	0.75	0.75
	8	**0.79**	0.76	0.76	0.76	0.79	**0.82**	0.82	0.82	**0.81**	0.75	0.75	0.75
	12	**0.79**	0.75	0.76	0.74	0.77	**0.84**	0.84	0.84	**0.79**	0.79	0.79	0.79
	22	0.73	0.73	0.73	0.73	0.78	0.78	0.78	0.78	0.79	0.79	0.79	0.79
bone	3	**0.63**	0.63	0.63	0.63	**0.59**	0.59	0.59	0.59	**0.59**	0.59	0.59	0.59
	4	**0.60**	0.59	0.59	0.62	**0.62**	0.62	0.62	0.59	**0.65**	0.64	0.64	0.61
	5	**0.75**	0.62	0.62	0.62	**0.75**	0.60	0.60	0.60	**0.77**	0.65	0.65	0.65
	6	0.74	0.74	0.74	0.74	0.73	0.73	0.73	0.73	0.77	0.77	0.77	0.77
ASDSDC	7	**0.83**	0.83	0.83	0.83	**0.83**	0.83	0.83	0.83	**0.82**	0.82	0.82	0.82
	10	**0.92**	0.90	0.86	0.91	**0.92**	0.86	0.86	0.86	**0.93**	0.90	0.90	0.90
	13	**0.95**	0.95	0.94	0.95	**0.92**	0.91	0.91	0.91	**0.98**	0.98	0.98	0.98
	16	0.93	0.93	0.93	0.93	0.92	0.92	0.92	0.92	0.96	0.96	0.96	0.96

（续表）

数据集	特征个数	RF 下的 F1 值				KNN 下的 F1 值				SVM 下的 F1 值			
		Rel-Betti	卡方检验	F 检验	互信息	Rel-Betti	卡方检验	F 检验	互信息	Rel-Betti	卡方检验	F 检验	互信息
arrhythmia	30	0.70	0.71	**0.73**	0.72	0.64	0.64	0.64	0.64	**0.72**	0.71	0.71	0.71
	50	**0.80**	0.76	0.76	0.76	0.61	**0.64**	**0.64**	**0.64**	0.73	**0.74**	**0.74**	**0.74**
	70	**0.80**	0.75	0.75	0.75	0.63	**0.67**	**0.67**	**0.67**	0.73	**0.75**	**0.75**	**0.75**
	260	0.81	0.81	0.81	0.81	0.53	0.53	0.53	0.53	0.74	0.74	0.74	0.74
ad	30	**0.98**	0.94	0.95	0.95	**0.98**	0.95	0.95	0.95	**0.98**	0.95	0.95	0.95
	100	0.95	0.95	0.95	0.95	0.92	0.92	0.92	0.92	0.95	0.95	0.95	0.95
	200	0.95	0.95	0.95	0.95	0.91	**0.92**	**0.92**	**0.92**	0.96	0.96	0.96	0.96
	1555	0.95	0.95	0.95	0.95	0.91	0.91	0.91	0.91	0.95	0.95	0.95	0.95

通过分析 4 组实验，可以得出以下结论：

（1）本文算法在 4 种分类器下的准确率和 F1 值都普遍优于卡方检验、F 检验和互信息，分析结果说明本文算法不依赖于学习模型，且本文方法选取少数特征进行分类时得到的结果比直接使用数据集的所有特征进行训练并分类的准确率高。

（2）随着特征个数的增加，Rel-Betti 算法选取的特征在 4 种机器学习下的准确率和 F1 值通常呈现先上升后稳定或先下降后稳定的趋势，这是由于数据集中包含许多冗余特征。并且本文算法通常能够取得较小的特征子集，减少训练时间。

（3）当数据集中特征个数较多时，无关特征的比率会变大，此时 Rel-betti 算法与卡方检验、F 检验和互信息选取相同个数的特征后的训练分类结果相比优势不大。

（4）从不同数据集的实验结果能够看出，不同的数据集具有不同的特性，这也说明在特征选择方面仍有探索的空间。

5.5 灵敏度分析

本节研究在 Rel-Betti 算法中参数对实验结果的影响。Rel-Betti 算法中的不确定参数包括过滤半径参数 ε 和比率 *ratio* 两项。

对于参数 ε，其值会稍微影响实验结果。ε 的取值需考虑将绝大部分数据形成复形，从而保证提取的拓扑特征是整个数据集的显著特征。

对于参数 *ratio*，其值对特征排序结果的影响较小。

以数据集 iris 为例，本文给出参数 *ratio* 对特征重要性排序的灵敏度分析。对于数据集 iris 中的 3 个类别，过滤半径参数 ε_1，ε_2，ε_3 分别为 0.3，0.2，0.3。当取 *ratio*＝30%，结果如表 4 所列，Rel-Betti 算法对特征进行重要性降序排序依次为 Petal. Width，Petal. Length，Sepal. Length，Sepal. Width。当取 *ratio*＝40%时，特征排序依旧如上，说明 *ratio* 的取值基本不会影响特征排序结果。

表 4 样本数据集信息

Table 4 Sample dataset information

iris	Sepal. Length	Sepal. Width	Petal. Length	Petal. Width
setosa	0.2114	0.6288	0.0753	0.0527
versicolor	0.4677	0.3171	0.5697	0.5234
virginica	0.5972	0.3875	0.7432	0.8063
差值最小值	0.1294	0.0704	0.1735	0.2828

从上述对参数 *ratio* 的灵敏度分析中能够看出，Rel-Betti 算法保留了持续同调对噪声具有鲁棒性。

5.6 验证 Rel-Betti 算法考虑特征组合效果

采用皮尔逊相关性分析对本文提出的 Rel-Betti 算法与卡方检验、F 检验、互信息特征选择后的数据特征进行分析，在一定程度上说明了 Rel-Betti 算法能够考虑特征与特征之间的组合关系。

皮尔逊系数的范围为（－1，1），其绝对值越大，说明两个变量间的相关程度越强，我们通过表 5 来判断变量之间相关程度的强弱。

表 5 皮尔逊相关程度

Table 5 Pearson correlation

相关系数范围	强弱程度
(0.0,0.2)	极弱相关或无相关
(0.2,0.4)	弱相关
(0.4,0.6)	中等程度相关
(0.6,0.8)	强相关
(0.8,1.0)	极强相关

以数据集 breast-cancer 为例，我们对 Rel-Betti 算法、卡方检验、F 检验以及互信息特征选择出的 3 个特征与类标签之间的相关性进行皮尔逊相关性分析，结果如图 6 所示，其中最后一行和最后一列都表示类标签。

（a）Rel-Betti 算法　（b）卡方检验

（c）F 检验　（d）互信息

图 6 breast-cancer 的皮尔逊相关性分析

Fig. 6 Pearson correlation analysis of breast-cancer

从图 6 能够看出，相对于卡方检验、F 检验、互信息挑选出的特征与类标签之间的相关性程度而言，Rel-Betti 算法挑选出的一个特征与类标签之间的相关性只有 0.041，属于极弱相关或无相关。从前面的实验结果图中能够看出，Rel-Betti 算法特征选择出的 3 个特征在分类器下的准确率和 F1 值

高于卡方检验、F 检验与互信息挑选出的 3 个特征。在一定程度上说明了 Rel-Betti 算法进行特征选择时能够考虑特征之间的组合效果。

结束语 本文基于 PH 提出了 Rel-Betti 算法，既保留了数据的内在信息，又降低了数据维度，提升了机器学习模型的性能。本文通过真实数据集在 4 种不同分类器上的一系列实验对所提算法的效果进行验证，实验结果表明 Rel-Betti 算法可以取得比传统过滤式方法更高的准确率和 F1 值。

本文算法基于持续同调理论，计算高维数据中的拓扑特征，当数据中一维相关贝蒂数不明显时，本文算法的准确率与其他特征算法结果相近。根据皮尔逊相关系数判断特征与分类目标之间的相关性来看，本文算法在处理冗余特征较多的数据集时效果更好，处理无关特征较多的数据集时效果一般。本文未对多分类数据集进行充分的实验，未来我们考虑将本文算法与传统算法相结合来处理无关特征多的数据，从而提高分类准确率，并在多分类数据集上验证本文算法的有效性。

参 考 文 献

[1] SHI Q J, PAN F, LONG F H, et al. A Review of Research on Feature Selection Methods[J]. Microelectronics and Computers, 2022, 39(3): 1-8.

[2] YU L, LIU H. Efficient feature selection via analysis of relevance and redundancy[J]. The Journal of Machine Learning Research, 2004, 5: 1205-1224.

[3] DROTÁR P, GAZDA J, SMÉKAL Z. An experimental comparison of feature selection methods on two-class biomedical datasets[J]. Computers in Biology and Medicine, 2015, 66: 1-10.

[4] BOMMERT A, SUN X, BISCHL B, et al. Benchmark for filter methods for feature selection in high-dimensional classification data[J]. Computational Statistics & Data Analysis, 2020, 143: 106839.

[5] JI Z W, HU M. A Double Filtering Feature Selection Algorithm [J]. Computer Engineering and Applications, 2011, 47(19): 190-193, 206.

[6] XU Y, HU X G, LI P P. A Filtered Feature Selection Algorithm Based on Group Policy[J]. Application Research of Computers, 2016, 33(5): 1322-1326.

[7] SINGH N, SINGH P. A hybrid ensemble-filter wrapper feature selection approach for medical data classification[J]. Chemometrics and Intelligent Laboratory Systems, 2021, 217: 104-131.

[8] PRISCILLA C V, PRABHA D P. A two-phase feature selection technique using mutual information and XGB-RFE for credit card fraud detection [J]. International Journal of Advanced Technology and Engineering Exploration, 2021, 8(85): 1656.

[9] PASHAEI E, PASHAEI E. Hybrid binary arithmetic optimization algorithm with simulated annealing for feature selection in high-dimensional biomedical data[J]. The Journal of Supercomputing, 2022: 78(13): 15598-15637.

[10] YANG Y K. Research on filtering feature selection algorithm based on mutual information[D]. Changchun: Jilin University, 2022.

[11] LUM P Y, SINGH G, LEHMAN A, et al. Extracting insights from the shape of complex data using topology[J]. Scientific Reports, 2013, 3(1): 1-8.

[12] ZOMORODIAN A, CARLSSON G. Computing persistent homology[C]// Proceedings of the twentieth Annual Symposium on Computational Geometry. 2004: 347-356. .

[13] CARLSSON G. Topology And Data[J]. Bulletin of the American Mathematical Society, 2009, 46(2): 255-308.

[14] LOCKWOOD S, KRISHNAMOORTHY B. Topological Features in Cancer Gene Expression Data[J]. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, 2015, 20(7): 108-119.

[15] CANG Z X, MU L, WEI G W, et al. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening [J]. Plos Computational Biology, 2018, 14(1): e1005929.

[16] CHI S P, XIA K, SI X L. Persistent-Homology-based Machine Learning and its Applications－A Survey[J]. Artificial Intelligence Review, 2022, 55(7): 5169-5213.

[17] OTTER N, PORTER M A, TILLMANN U, et al. A roadmap for the computation of persistent homology [J]. Epj Data Science, 2017, 6(1): 1-38.

[18] GHRIST R. Barcodes: The persistent topology of data[J]. Bulletin of the American Mathematical Society, 2008, 45(1): 61-75.

[19] GAO B L, ZHOU Z G, YANG W W, et, al. A Feature Selection Method Combining Category-Based and Improved CHI[J]. Application Research of Computers, 2018, 35(6): 1660-1662.

[20] WANG Y R, TANG M. Evaluation of side channel leakage based on Bartlett and multi-class F-test [J]. Journal of Communications, 2021, 42(12): 35-43.

[21] WU Y, LIU Y H, YANG W W, et, al. Feature Selection Method Based on Nearest Farthest Neighbor and Mutual Information [J]. Research of Computers 2017, 34(12): 3713-3716.

[22] WANG W Y, LIU C, ZHAO Q, et, al. Direct Verified Encapsulated Feature Selection Method [J]. University of Electronic Science and Technology of China Journal, 2016, 45(4): 607-615.

YIN Xingzi, born in 1998, postgraduate. Her main research interest is topology data analysis.

PENG Ningning, born in 1985, associate professor, Ph. D. His main research interests include mathematical logic, computability theory and topology data analysis.

（责任编辑：何杨）