第47卷 第11期 计 算 机 工 程 2021年11月
Vol.47 No.11 Computer Engineering November 2021

·人工智能与模式识别· 文章编号：1000-3428(2021)11-0100-08 文献标志码：A 中图分类号：TP18

# 基于不平衡数据的特征选择算法研究

王俊红[1,2]，赵彬佳[1]

（1.山西大学 计算机与信息技术学院，太原 030006；2.山西大学 计算智能与中文信息处理教育部重点实验室，太原 030006）

**摘 要：**不平衡分类问题广泛存在于医疗、经济等领域，对于不平衡数据集分类，特别是高维数据分类时，有效的特征选择算法至关重要。然而多数特征选择算法未考虑特征协同的影响，导致分类性能下降。对FAST特征选择算法进行改进，并考虑特征的协同作用，提出一种新的特征选择算法FSBS。运用AUC对特征进行评估，以相互增益衡量协同作用大小，选出有效特征，进而对不平衡数据进行分类。实验结果表明，该算法能有效地选择特征，尤其在特征数量较少的情况下可保持较高的分类准确率。

**关键词：**特征选择；不平衡数据；FSBS算法；特征协同；分类准确率

开放科学（资源服务）标志码（OSID）：

**中文引用格式：**王俊红，赵彬佳.基于不平衡数据的特征选择算法研究[J].计算机工程，2021，47(11)：100-107.

**英文引用格式：**WANG J H，ZHAO B J. Research on feature selection algorithms based on unbalanced data[J]. Computer Engineering，2021，47(11)：100-107.

# Research on Feature Selection Algorithms Based on Unbalanced Data

WANG Junhong[1,2]，ZHAO Binjia[1]

（1.School of Computer and Information Technology，Shanxi University，Taiyuan 030006，China；2.Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education，Shanxi University，Taiyuan 030006，China）

**【Abstract】**The problem of unbalanced classification widely exists in medical，economic and other fields. Research shows that for the classification of unbalanced data sets，especially when the data is high-dimensional，an effective feature selection algorithm is crucial. However，most feature selection algorithms do not consider the impact of feature synergy，resulting in a decrease in classification performance. Considering the synergy of features，this paper proposes a new feature selection algorithm，FSBS，on the basis of the improved FAST feature selection algorithm. This algorithm employs AUC to evaluate the features，and the mutual gain is used to measure the magnitude of synergy. Then the effective features are selected and the unbalanced data are classified. Experimental results show that the proposed algorithm can effectively select features and improve the classification performance，especially when the number of features is small.

**【Key words】**feature selection；unbalanced data；FSBS algorithm；feature synergy；classification accuracy

**DOI：**10. 19678/j. issn. 1000-3428. 0059373

## 0 概述

特征选择是数据降维的一种重要方法[1]，最早出现在上世纪60年代，它的本质是从最开始的$M$个特征中选取某种标准下表现最好的$N$个特征组成特征子集[2]，以便训练出更加精确的模型，得出更满意的分类效果，取得更清楚的分析结论[3]。特征选择具有减少储存需求、减少训练时间、提高学习的准确率等优点，在机器学习等诸多领域具有不可替代的作用[4]。

不平衡数据是指数据集中的某个或某些类的样本量远远高于其他类，而某些类样本量较少，通常把样本量较多的类称为多数类，样本量较少的类称为少数类。不平衡分类问题广泛存在于多个领域，例如文本分类、医学诊断、信用卡欺诈检测、垃圾邮件判断等[5]。而在这些应用中，人们对研究少数类更感兴趣，少数类中的样本往往更有价值。

随着社会的发展和进步，越来越多的数据是高维且不平衡的，这给数据挖掘工作带来前所未有的挑战。在分类任务中，这些高维数据中存在大量无

**基金项目：**国家自然科学基金（61976128）；山西省自然科学基金（201701D121051）。

**作者简介：**王俊红（1979—），女，副教授、博士，主研方向为数据挖掘、机器学习；赵彬佳，硕士研究生。

**收稿日期：**2020-08-27 **修回日期：**2020-10-21 **E-mail：**wjhwjh@sxu.edu.cn

关特征和冗余特征，不仅需要大量的运行时间，还会影响分类效果。当这些高维数据中存在类别不平衡现象时，传统的分类算法在少数类上的分类效果总是不尽人意[6]。因此，对于不平衡数据集分类，特别当数据同时也是高维时，特征选择有时比分类算法更重要[7]。

FAST（Feature Assessment by Sliding Thresholds）算法[8]根据AUC值（ROC曲线下面积）评估特征，而AUC值是一个较好的预测性能的指标，尤其是对于不平衡数据分类问题，但该方法没有考虑特征之间的协同作用。具有协同作用的特征是指各自对目标集不相关，但两者相结合却对目标集高度相关的特征，若不考虑则可能删除这些有协同作用的特征，导致分类性能下降。本文对FAST算法进行改进，提出一种基于特征协同的FSBS特征选择算法。运用UCI数据集进行实验，将原始数据进行直接分类，并与特征选择后进行分类的FAST算法作为对比，以验证FSBS方法的分类性能。

## 1　相关研究

特征选择算法一般需要确定搜索起点和方向、搜索策略、特征评估函数和停止准则4个要素。在特征选择中，搜索策略和特征评估函数是比较重要的2个步骤，因此将特征选择算法分为两大类。

按照搜索策略可以将特征选择方法分为全局最优搜索、序列搜索和随机搜索3种[9]。全局最优搜索有穷举法和分支定界法[10]。序列搜索根据搜索方向的不同可分为前向搜索、后向搜索和双向搜索3类。随机搜索策略具有较强的不确定性，遗传算法（GA）[11]、模拟退火算法（SA）[12]和蚁群算法（ACO）[13]是常用的随机搜索方法。

特征评估函数对特征选择至关重要，特征子集的好坏取决于评估函数，现有的特征评估函数大致分为信息度量、距离度量、依赖性度量、一致性度量和分类性能5类[14]。

Relief[15]算法是一种特征权重算法，特征和类别的相关性决定特征权重大小，特征权重越大，该特征的分类能力越强，特征权重小的特征将会被删除，此方法只能用于二分类问题。为了使Relief算法支持多分类问题，KONONENKO等[16]对其进行改进，提出一种ReliefF算法。信息增益（Information Gain，IG）[17]常用来进行特征选择，通过计算每个特征对分类提供的信息量判断特征的重要性程度。卡方统计[18]是在文本分类中经常出现的一种特征选择方法，能有效提高文本分类的性能。互信息也经常用作特征选择，文献[19]提出一种新的基于迭代的定性互信息的特征选择算法，利用随机森林算法求出每个特征的效用得分，并将其与每个特征的互信息及其分类变量相结合。随机森林能够处理高维数据，并且有着较好的鲁棒性，常被用来做特征选择。

针对不平衡数据的分类问题，目前主要从数据层面和算法层面来解决。数据层面可以使用过采样方法增加少数类样本，也可以使用欠采样方法减少多数类样本，还可以使用混合采样方法平衡数据集。SMOTE[20]是一种经典的过采样方法，通过合成样本增加正类样本数量。文献[21]提出一种利用朴素贝叶斯分类器的欠采样方法，以随机初始选择为基础，从可用的训练集中选择信息最丰富的实例。过采样方法和欠采样方法都存在缺陷，过采样方法容易造成数据过拟合，而欠采样方法容易丢失有用信息。

在算法层面，传统的分类器假定类别之间的误分类代价是相同的，但是在不平衡数据中，少数类的误分类代价往往高于多数类。因此，很多针对不平衡数据的分类过程都伴随代价敏感学习，当少数类被错分时，赋予更高的惩罚代价，使分类器对少数类的关注增加。文献[22]利用HCSL算法建立代价敏感的决策树分类器。文献[23]将模式发现与代价敏感方法相结合来解决类别不平衡问题。

将特征选择应用到不平衡数据，在近年来已经引起了研究者的关注。CHEN等[8]提出一种基于AUC评价标准的特征选择方法。文献[24]将递归特征消除和主成分分析运用到随机森林中，并结合SMOTE方法处理不平衡数据。SUN等[25]提出一种基于分支与边界的混合特征选择（BBHFS）与不平衡定向多分类器集成（IOMCE）相结合的不平衡信用风险评估方法，在减少特征数量的同时保留更多的有用信息。BIAN等[26]将代价敏感学习应用到特征选择中，提出一种代价敏感特征选择方法，通过给正类赋予更高的惩罚代价值使得算法对正类关注增加。SYMON等[27]使用对称不确定性衡量特征与标签之间的依赖性，同时还使用harmony搜索以选择最佳的特征组合，是一种适用于高维不平衡数据集的特征选择方法。

FAST算法[8]根据AUC值来评估特征，该方法通过在多个阈值上对样本进行分类并计算每个阈值处的真正例率和假正例率，从而构建ROC曲线并计算该曲线下的面积。AUC大的特征往往具有更好的预测能力，因此将AUC用作特征等级。

## 2　基于特征协同的FSBS特征选择算法

多数特征选择算法通常只有单一的决策边界[28]，当改变这个决策边界时，可能产生更多的真正例和更少的真反例，也可能产生更少的真正例和更多的真反例。在不平衡数据中，这种情况的发生会严重影响分类效果。因此，需要不断滑动阈值来确定哪个特征集更好，这样才能够更好地对不平衡数

据集进行分类。ROC曲线对样例进行排序，按此顺序逐个把样本作为正例进行预测（即滑动阈值），每次预测后计算假正例率和真正例率，分别以它们作为横纵坐标作图得到ROC曲线。研究发现，那些在不平衡数据集上表现较好的分类器，是因为在特征选择时使用了主要关注少数类的度量标准[8]。因此，基于不平衡数据的特征选择算法需要使用适合不平衡数据的度量标准。利用ROC曲线进行评估的一个巨大优势是：即便正负样本数量发生变化，ROC曲线形状基本保持不变。因此，ROC曲线能够更加客观也衡量学习器本身的好坏，适合于不平衡数据集。如果需要对学习器进行量化分析，比较合适的标准是ROC曲线下的面积，即AUC。

FAST方法通过计算每一个特征的AUC值，选取AUC值大于预设阈值的特征组成特征子集，非常适合不平衡数据集上的特征选择。此方法只是对单个特征进行评估，因此有一个不足之处，即没有考虑特征与特征之间的相互影响。如果数据集中存在协同的特征，则该方法容易将这些特征忽略，导致分类性能下降。为解决该问题，提高分类准确率，本文在原算法的基础上考虑特征之间的协同作用，提出了FSBS方法。

协同的特征是指各自对目标集不相关或弱相关，但两者相结合却对目标集高度相关的特征。本文以相互增益评价协同作用大小，在介绍相互增益前，先引入信息熵和信息增益的概念。

信息熵是SHANNON在1948年提出的，用来评估样本集合纯度的一个参数。给出一个样本集合，该集合中的样本可能属于多个不同的类别，也可能只属于一个类别，那么如果属于多个不同的类别，则该样本是不纯的，如果只属于一个类别，则该样本是纯洁的。信息熵是计算一个样本集合中的数据是否纯洁，信息熵越小，表明这个数据集越纯洁。信息熵的最小值为0，此时数据集$D$中只含有一个类别。

信息增益是指信息熵的有效减少量。一个特征的信息增益越大，说明该特征的分类性能越强。假设特征为$X=\{x_1,x_2,\cdots,x_n\}$，标签为$Y=\{y_1,y_2,\cdots,y_m\}$，它们的联合概率分布为$p(x_i,y_j)$，$x_i$代表特征中的某个取值，$y_j$代表标签中的某个取值，其中，$i=1,2,\cdots,n$，$j=1,2,\cdots,m$。则信息增益$I(X;Y)$的计算公式如下：

$$I(X;Y)=\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_i,y_j)\mathrm{lb}\frac{p(x_i|y_j)}{p(x_i)} \tag{1}$$

JAKULIN等[29-30]提出了相互增益的概念。相互增益可以衡量协同作用的大小，可以为正、零，还可以为负。假设$X$和$Y$是特征，$Z$是标签，则相互增益的计算公式如下：

$$I(X;Y;Z)=I(X,Y;Z)-I(X;Z)-I(Y;Z) \tag{2}$$

式(2)可改变如下：

$$I(X;Y;Z)=\sum_{k}\sum_{j}\sum_{l}p(x_k,y_j,z_l)\mathrm{lb}\frac{p(x_k,y_j)p(y_j,z_l)p(x_k,z_l)}{p(x_k)p(y_j)p(z_l)p(x_k,y_j,z_l)} \tag{3}$$

从上式可以看出，当2个特征在一起时的信息增益大于它们单独存在时的信息增益，相互增益为正值。相互增益为正值，说明特征之间存在协同作用，相互增益越大，协同作用越强。

FSBS方法首先计算所有特征两两之间的相互增益，然后求每个特征与其他特征相互增益的算术平均值Ave（即为某个特征的平均相互增益），再用每个特征的AUC值和平均相互增益Ave按式(4)计算得到C_AUC。式(4)将AUC值和平均相互增益融合形成一个新的标准评价每个特征。

$$\mathrm{C\_AUC}=\mathrm{AUC}\times(1+(\mathrm{Ave}\times 100)) \tag{4}$$

其中：AUC的取值范围为0.5~1.0；Ave的值可能为正数，也可能为负数。在算法的最后设置阈值$q$，选取C_AUC超过阈值的特征组成特征子集。

FSBS算法流程如图1所示。

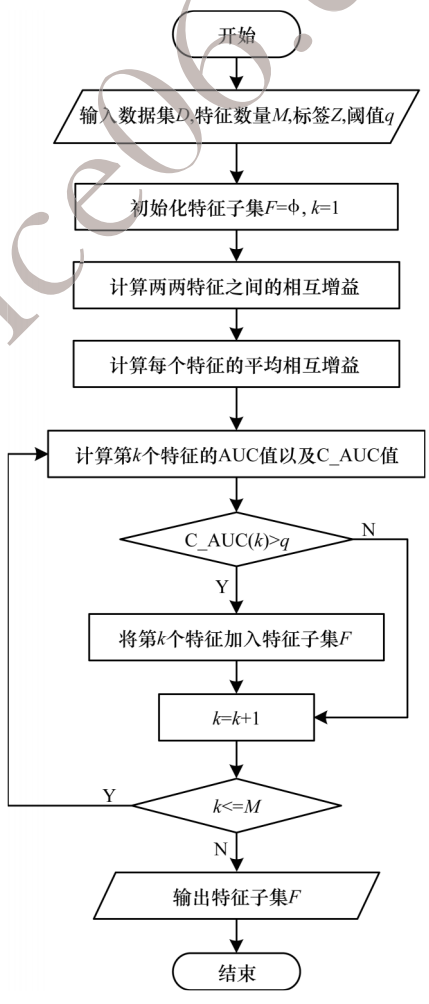


**图1 FSBS算法流程**

**Fig.1 FSBS algorithm procedure**

FSBS算法的伪代码如下：

**算法1**　FSBS方法

**输入**　数据集$D$，特征数量$M$，标签$Z$，阈值$q$

**输出**　特征子集$F$

1.初始化：特征子集F = ϕ
2.FOR r=1 to M
3.FOR s=1 to Mand r! =s
4.计算相互增益$I(X_r, Y_s, Z)$
5.END
6.计算第r个特征的平均相互增益Ave
7.END
8.FOR k=1 to M
9.计算第k个特征的AUC
10.计算第k个特征的C_AUC
11.IF C_AUC(k)>q
12.将第k个特征加入特征子集F
13.END
14.END

## 3　实验结果与分析

### 3.1　实验数据

为评估FSBS算法的性能，以证实FSBS方法能有效地用于实践，本文从UCI数据库中选择了14个数据集来进行实验对比分析。这14个数据集的特征个数最少为4，最大为1 470，数据的不平衡比大小不同。数据集如表1所示。

**表1　实验数据集**

**Table 1　Experimental Datasets**

| 数据集 | 样本数量 | 特征数量 | 正负样本比例 |
|---|---|---|---|
| auto | 205 | 25 | 27:178 |
| cancer | 683 | 9 | 239:444 |
| cars | 392 | 8 | 147:245 |
| crx | 690 | 15 | 307:383 |
| dermatology | 366 | 34 | 130:236 |
| ecoli | 336 | 343 | 151:185 |
| german | 1 000 | 20 | 300:700 |
| glass | 214 | 9 | 96:118 |
| heart | 270 | 13 | 120:150 |
| ionosphere | 351 | 34 | 126:225 |
| iris | 150 | 4 | 50:100 |
| vote | 435 | 16 | 168:267 |
| wine | 178 | 13 | 71:107 |
| yeast | 1 484 | 1 470 | 671:813 |

本文实验以少数类的分类准确度、多数类的分类准确度、总的分类准确度以及分类时间作为评价指标。以原始数据进行分类和FAST算法特征选择后进行分类作为对比。

### 3.2　实验参数

本文在实验中运用式(4)计算C_AUC。为确定式(4)中Ave扩大多少倍，实验中使用上文的14个数据集，以决策树为分类器进行实验，分别将Ave扩大1倍、10倍、100倍以及1 000倍，以G-mean值作为评价指标。不同情况下各数据集的G-mean值对比如图2所示。

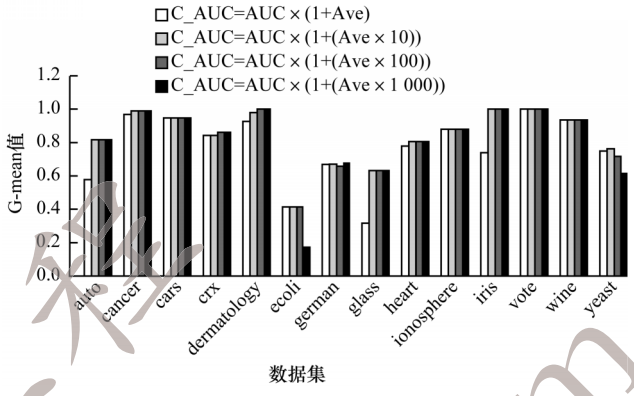


**图2　不同情况下C_AUC值对比**

**Fig.2　Comparison of C_ AUC values in different cases**

实验结果表明，在Ave扩大100倍时效果最佳，有12个数据集G-mean值最高，因此式(4)中确定为扩大100倍。

阈值$q$的选取会影响特征数量，实验中将阈值$q$分别设置为所有特征的C_AUC中最大值与最小值之和的0.3倍、0.4倍、0.5倍、0.6倍、0.7倍。当最大值与最小值之和为正数时，随着倍数的提高，阈值$q$增大，特征数量呈下降趋势。反之，当最大值与最小值之和为负数时，随着倍数的提高，阈值$q$减小，特征数量呈增加趋势。不同阈值下特征数量如表2所示。

**表2　不同阈值下的特征数量比较**

**Table 2　Comparison of features number different thresholds**

| 数据集 | $q$=(max+min)·0.3 | $q$=(max+min)·0.4 | $q$=(max+min)·0.5 | $q$=(max+min)·0.6 | $q$=(max+min)·0.7 |
|---|---|---|---|---|---|
| auto | 15 | 11 | 11 | 7 | 6 |
| cancer | 8 | 8 | 6 | 5 | 2 |
| cars | 6 | 6 | 4 | 3 | 3 |
| crx | 5 | 7 | 7 | 7 | 7 |
| dermatology | 28 | 28 | 31 | 32 | 33 |
| ecoli | 342 | 341 | 340 | 338 | 338 |
| german | 18 | 18 | 18 | 19 | 19 |
| glass | 6 | 6 | 3 | 2 | 1 |
| heart | 8 | 9 | 9 | 9 | 10 |
| ionosphere | 32 | 32 | 32 | 32 | 32 |
| iris | 2 | 2 | 2 | 2 | 2 |
| vote | 11 | 10 | 7 | 5 | 3 |
| wine | 8 | 6 | 6 | 4 | 3 |
| yeast | 1 470 | 1 470 | 1 467 | 0 | 0 |

分析表2可知：在阈值$q$为0.3倍和0.7倍时，会导致在部分数据集上特征相对较多，而另一部分数据集上特征相对较少；在阈值为0.4倍和0.6倍时，此种情况减轻；在阈值为0.5倍时，情况最好。因此，本在文实验中，FSBS方法在生成特征子集时的阈值$q$最终设置为$M$个特征的C_AUC值中最大值和最小值的算术平均数（即0.5倍）。

### 3.3 实验结果

为更好地比较2个算法选择的特征子集的优劣，验证算法的性能，分别使用SVM、决策树、随机森林3种分类器进行分类。在本文实验中训练集为总样本数的90%，测试集为总样本数的10%。

表3所示为原始数据的特征个数、FAST方法选择后的特征个数以及FSBS方法选择后的特征个数。分析表3可以发现，FSBS方法对原始数据进行特征选择后特征的数量都比原始数据少，显然这和阈值$q$的设置有关。与FAST相比，两者对原始数据进行特征选择后特征数量没有必然联系，即使这2种特征选择方法处理后特征数量一样，特征也不一定相同。

在14个数据集中，FSBS算法有3种不同的效果：第1种提升了分类准确率（见表4）；第2种使用更少的特征，却保持了最高的分类准确率（见表5）；第3种提升了正类的准确率，负类准确率只有轻微下降（见表6）。在表4~表6中：①代表原数据；②代表FAST算法选择后的数据；③代表FSBS方法选择后的数据。

**表3 不同方法的特征数量比较**

**Table 3 Comparison of feature numbers of different methods**

| 数据集 | 原始数据特征数量 | FAST选择后特征数量 | FSBS选择后特征数量 |
|---|---|---|---|
| auto | 25 | 8 | 11 |
| cancer | 9 | 8 | 6 |
| cars | 8 | 6 | 4 |
| crx | 15 | 3 | 7 |
| dermatology | 34 | 6 | 31 |
| ecoli | 343 | 340 | 340 |
| german | 20 | 3 | 18 |
| glass | 9 | 2 | 3 |
| heart | 13 | 8 | 9 |
| ionosphere | 34 | 8 | 32 |
| iris | 4 | 1 | 2 |
| vote | 16 | 10 | 7 |
| wine | 13 | 4 | 6 |
| yeast | 1 470 | 5 | 1 467 |

**表4 分类准确率1**

**Table 4 Classification accuracy 1** %

| 数据 | SVM | | | 决策树 | | | 随机森林 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 正类准确率 | 负类准确率 | 总准确率 | 正类准确率 | 负类准确率 | 总准确率 | 正类准确率 | 负类准确率 | 总准确率 |
| cancer① | 100.00 | 97.78 | 98.55 | 91.67 | 97.78 | 95.65 | 100.00 | 97.78 | 98.55 |
| cancer② | 100.00 | 97.78 | 98.55 | 95.83 | 97.78 | 97.10 | 100.00 | 97.78 | 98.55 |
| cancer③ | 100.00 | 97.78 | 98.55 | 100.00 | 97.78 | 98.55 | 100.00 | 97.78 | 98.55 |
| crx① | 77.42 | 97.44 | 88.57 | 74.19 | 97.44 | 87.14 | 64.52 | 97.44 | 82.86 |
| crx② | 77.42 | 97.44 | 88.57 | 58.06 | 100.00 | 81.43 | 77.42 | 97.44 | 88.57 |
| crx③ | 77.42 | 97.44 | 88.57 | 74.19 | 100.00 | 88.57 | 74.19 | 97.44 | 87.14 |
| glass① | 20.00 | 100.00 | 63.64 | 0.00 | 100.00 | 54.55 | 10.00 | 100.00 | 59.09 |
| glass② | 10.00 | 100.00 | 59.09 | 10.00 | 100.00 | 59.09 | 10.00 | 100.00 | 59.09 |
| glass③ | 3.00 | 100.00 | 68.18 | 40.00 | 100.00 | 72.73 | 30.00 | 100.00 | 68.18 |
| heart① | 61.54 | 93.75 | 79.31 | 69.23 | 87.50 | 79.31 | 69.23 | 93.75 | 82.76 |
| herat② | 61.54 | 100.00 | 82.76 | 69.23 | 87.50 | 79.31 | 69.23 | 100.00 | 86.21 |
| heart③ | 61.54 | 100.00 | 82.76 | 69.23 | 93.75 | 82.76 | 69.23 | 93.75 | 82.76 |
| ionosphere① | 69.23 | 100.00 | 88.89 | 76.92 | 91.30 | 86.11 | 84.62 | 95.65 | 91.67 |
| ionosphere② | 61.54 | 86.96 | 77.78 | 84.62 | 82.61 | 83.33 | 84.62 | 91.30 | 88.89 |
| ionosphere③ | 69.23 | 95.65 | 86.11 | 84.62 | 91.30 | 88.89 | 84.62 | 95.65 | 91.67 |

表5　分类准确率2

Table 5　Classification accuracy 2

| 数据 | 数据大小 | SVM/% | | | 决策树/% | | | 随机森林/% | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 正类准确率 | 负类准确率 | 总准确率 | 正类准确率 | 负类准确率 | 总准确率 | 正类准确率 | 负类准确率 | 总准确率 |
| auto① | 205×26 | 0.00 | 100.00 | 85.71 | 100.00 | 100.00 | 100.00 | 33.33 | 100.00 | 90.48 |
| auto② | 205×9 | 0.00 | 100.00 | 85.71 | 33.33 | 100.00 | 90.48 | 0.00 | 100.00 | 85.71 |
| auto③ | 205×12 | 0.00 | 100.00 | 85.71 | 66.67 | 100.00 | 95.24 | 33.33 | 100.00 | 90.48 |
| cars① | 392×9 | 80.00 | 88.00 | 85.00 | 93.33 | 96.00 | 95.00 | 100.00 | 96.00 | 97.50 |
| cars② | 392×7 | 86.67 | 64.00 | 72.50 | 93.33 | 96.00 | 95.00 | 100.00 | 96.00 | 97.50 |
| cars③ | 392×5 | 80.00 | 68.00 | 72.50 | 93.33 | 96.00 | 95.00 | 100.00 | 96.00 | 97.50 |
| dermatology① | 366×35 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| dermatology② | 366×7 | 92.86 | 100.00 | 97.37 | 85.71 | 100.00 | 94.74 | 92.86 | 100.00 | 97.37 |
| dermatology③ | 366×32 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| iris① | 150×5 | 0.00 | 100.00 | 64.71 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| iris② | 150×2 | 0.00 | 100.00 | 64.71 | 66.67 | 81.82 | 76.47 | 50.00 | 90.91 | 76.47 |
| iris③ | 150×3 | 0.00 | 100.00 | 64.71 | 100.00 | 100.00 | 10.00 | 100.00 | 100.00 | 100.00 |
| vote① | 435×17 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| vote② | 435×11 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| vote③ | 435×8 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| wine① | 178×14 | 87.50 | 100.00 | 94.74 | 87.50 | 100 | 94.74 | 100.00 | 100.00 | 100.00 |
| wine② | 178×5 | 100.00 | 90.91 | 94.74 | 100.00 | 90.91 | 94.74 | 100.00 | 90.91 | 94.74 |
| wine③ | 178×7 | 87.50 | 100.00 | 94.74 | 87.50 | 100.00 | 94.74 | 100.00 | 100.00 | 100.00 |

表6　分类准确率3

Table 6　Classification accuracy 3　　%

| 数据 | SVM | | | 决策树 | | | 随机森林 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 正类准确率 | 负类准确率 | 总准确率 | 正类准确率 | 负类准确率 | 总准确率 | 正类准确率 | 负类准确率 | 总准确率 |
| ecoli① | 75.00 | 31.58 | 51.43 | 68.75 | 36.84 | 51.43 | 87.50 | 10.53 | 45.71 |
| ecoli② | 81.25 | 42.11 | 60.00 | 81.25 | 10.53 | 42.86 | 81.25 | 10.53 | 42.86 |
| ecoli③ | 81.25 | 42.11 | 60.00 | 81.25 | 21.05 | 48.57 | 81.25 | 5.26 | 40.00 |
| german① | 48.39 | 87.32 | 75.49 | 29.03 | 76.06 | 61.76 | 48.39 | 92.96 | 79.41 |
| german② | 35.48 | 98.59 | 79.41 | 45.16 | 77.46 | 67.65 | 51.61 | 88.73 | 77.45 |
| german③ | 38.71 | 85.92 | 71.57 | 58.06 | 74.65 | 69.61 | 54.84 | 88.73 | 78.43 |
| yeast① | 76.47 | 79.27 | 78.00 | 72.06 | 69.51 | 70.67 | 79.41 | 85.37 | 82.67 |
| yeast② | 76.47 | 81.71 | 79.33 | 77.94 | 71.95 | 74.67 | 73.53 | 81.71 | 78.00 |
| yeast③ | 82.35 | 74.39 | 78.00 | 69.12 | 74.39 | 72.00 | 83.82 | 80.49 | 82.00 |

从表4可以看出,5个数据集使用FSBS算法后,在不同程度上提高了分类准确率。其中1个数据集在3个分类器上的分类准确率均有提高。5个数据集在决策树上的分类准确率都得到了提高。

从表5可以看出,6个数据集使用FSBS算法在特征数量尽可能少的情况下,可以保持最高的分类准确率。当数据维度较大时,能在特征数量较少的情况下保持较高的准确率是至关重要的,可以缩短运行时间,减少存储成本。

从表6可以看出,3个数据集上负类分类准确率有轻微减少,但在正类上提升效果明显。在不平衡数据分类中,本文对正类更感兴趣,正类错分比负类错分代价更大,因此可以以牺牲部分负类分类准确率为代价,提高正类分类准确率。但在数据ecoli中,虽然负类分类准确率下降较多,但是测试集少,只分错3个样本。

如表7所示,在进行分类时重复10次,求取平均值得到分类时间(SVM、决策树、随机森林3个分类器的分类总时间)。对比发现在多数数据集上,FAST和FSBS都可以使分类时间变短,原因是经过特征选择后,特征数量下降,所以分类时间变短。而在个别数据集上(ecoli和yeast)分类时间增加,是由

于特征选择后,特征数量几乎保持不变(详见表3),而且经过特征选择后,数据集中特征位置发生变化,使得分类器耗时更长。在表7中,分类时间不含特征选择时间,但考虑到在高维数据中特征选择花费的时间不可忽略,在表8中,将特征选择时间包含进去进行比较。表8显示在部分数据集上,特征选择后的分类时间依然比原始数据进行分类的时间要短,但当数据集中特征数量较多或样本数量较多时,FSBS方法处理后的分类时间较高。

**表7 不含特征选择时间的分类时间对比**

**Table 7 Classification time comparison of excluding feature selectiontime** s

| 数据 | 原数据分类时间 | FAST处理后分类时间 | FSBS处理后分类时间 |
|---|---|---|---|
| auto | 0.369 5 | 0.100 8 | 0.121 4 |
| cancer | 0.476 4 | 0.267 5 | 0.243 3 |
| cars | 0.334 8 | 0.136 6 | 0.111 4 |
| crx | 0.877 1 | 0.163 1 | 0.289 1 |
| dermatology | 0.578 3 | 0.152 4 | 0.392 0 |
| ecoli | 6.920 5 | 10.192 6 | 10.599 2 |
| german | 1.765 6 | 0.278 4 | 1.521 4 |
| glass | 0.324 9 | 0.091 8 | 0.097 6 |
| heart | 0.417 3 | 0.172 4 | 0.196 4 |
| ionosphere | 0.720 6 | 0.192 3 | 0.528 0 |
| iris | 0.237 5 | 0.074 6 | 0.061 8 |
| vote | 0.495 9 | 0.231 1 | 0.182 2 |
| wine | 0.288 4 | 0.078 4 | 0.081 3 |
| yeast | 691.684 2 | 1.019 1 | 1 127.990 6 |

**表8 含特征选择时间的分类时间对比**

**Table 8 Classificationtime comparison of including feature selectiontime** s

| 数据 | 原数据分类时间 | FAST处理后分类时间 | FSBS处理后分类时间 |
|---|---|---|---|
| auto | 0.369 5 | 0.108 5 | 0.762 4 |
| cancer | 0.476 4 | 0.274 1 | 0.332 7 |
| cars | 0.334 8 | 0.143 0 | 0.426 4 |
| crx | 0.877 1 | 0.171 4 | 1.792 4 |
| dermatology | 0.578 3 | 0.163 4 | 0.838 9 |
| ecoli | 6.920 5 | 10.267 5 | 40.794 6 |
| german | 1.765 6 | 0.289 9 | 2.554 2 |
| glass | 0.324 9 | 0.097 3 | 0.474 3 |
| heart | 0.417 3 | 0.180 9 | 0.360 6 |
| ionosphere | 0.720 6 | 0.203 1 | 29.649 3 |
| iris | 0.237 5 | 0.078 5 | 0.090 8 |
| vote | 0.495 9 | 0.238 8 | 0.313 2 |
| wine | 0.288 4 | 0.084 0 | 0.647 0 |
| yeast | 691.684 2 | 1.668 0 | 2 113.335 7 |

## 4 结束语

本文通过对FAST算法进行改进,在FAST算法的基础上考虑特征的协同作用,提出一种新的特征选择算法。通过计算特征之间的相互增益,分析特征与特征之间的协同度,从而使协同作用大的特征更容易被选择到。实验结果表明,该算法能在一定程度上提高分类性能,尤其是少数类的准确率。但该方法在计算特征之间的相互增益时会增加运行时间,因此快捷有效地选择有协同作用的特征并进行高效应用是下一步的研究重点。

## 参考文献


[1] LI J, LIU H. Challenges of feature selection for big data analytics[J]. IEEE Intelligent Systems, 2017, 32(2): 9-15.

[2] BOLON-CANEDO V, SANCHEZ-MARONO N, ALONSO-BETANZOS A. Feature selection for high dimensional data [J]. Progress in ArtificialIntelligence, 2016, 5(2): 65-75.

[3] DESSÌ N, PES B. Similarity of feature selection methods: an empirical study across data intensive classification tasks [J]. Expert Systems with Applications, 2015, 42(10): 4632-4642.

[4] 初蓓,李占山,张梦林,等. 基于森林优化特征选择算法的改进研究[J]. 软件学报,2018,29(9):2547-2558.
CHU B, LI Z S, ZHANG M L, et al. Research on improvements of feature selection using forest optimizationalgorithm[J]. Journal of Software, 2018, 29(9): 2547-2558. (in Chinese)

[5] 于巧,姜淑娟,张艳梅,等. 分类不平衡对软件缺陷预测模型性能的影响研究[J]. 计算机学报,2018,41(4):809-824.
YU Q, JIANG S J, ZHANG Y M, et al. The impact study of class imbalance on the performance of software defect prediction models[J]. Chinese Journal of Computers, 2018, 41(4): 809-824. (in Chinese)

[6] HUANG X, ZOU Y, WANG Y, et al. Cost-sensitive sparse linear regression for crowd counting with imbalanced training data[C]//Proceedings of International Conference on Multimedia and Expo. Washington D. C., USA: IEEE Press, 2016: 1-6.

[7] FORMAN G. An extensive empirical study of feature selection metrics for text classification [J]. Journal of Machine Learning Research, 2003, 3(2): 1289-1305.

[8] CHEN X, WASIKOWSKI M. Fast: a roc-based feature selection metric for small samples and imbalanced data classificationproblems[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2008: 124-132.

[9] SUN Z, BEBIS G, MILLER R H, et al. Object detection using feature subset selection [J]. Pattern Recognition, 2004, 37(11): 2165-2176.

[10] SOMOL P, PUDIL P, KITTLER J, et al. Fast branch &bound algorithms for optimal feature selection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(7): 900-912.

[11] WUTZL B, LEIBNITZ K, RATTAY F, et al. Genetic algorithms for feature selection when classifying severe chronic disorders of consciousness [J]. PloSone, 2019,

14(7)：1265-1278.
[12] 李元香，项正龙，张伟艳. 模拟退火算法的弛豫模型与时间复杂性分析[J]. 计算机学报，2020，43(5)：796-811.
LI Y X，XIANG Z L，ZHANG W Y. A relaxation model and time complexity analysis for simulated annealing algorithm[J]. Chinese Journal of Computers，2020，43(5)：796-811. (in Chinese)
[13] GHIMATGAR H，KAZEMI K，HELFROUSH M S，et al. An improved feature selection algorithm based on graph clustering and ant colony optimization[J]. Knowledge Based Systems，2018，159：270-285.
[14] IMOLA F. A survey of dimension reduction techniques[J]. Neoplasia，2002，7(5)：475-485.
[15] KIRA K，RENDELL L A. Feature selection problem：traditional methods and a new algorithm[C]//Proceedings of the 20th National Conference on Artificial Intelligence. San Jose，USA：AAAI Press，1992：129-134.
[16] KONONENKO I. Estimating attributes：analysis and extensions of RELIEF[C]//Proceedings of European Conference on Machine Learning. Berlin，Germany：Springer，1994：171-182.
[17] 张振海，李士宁，李志刚，等. 一类基于信息熵的多标签特征选择算法[J]. 计算机研究与发展，2013，50(6)：1177-1184.
ZHANG Z H，LI S N，LI Z G，et al. Multi-label feature selection algorithm based on information entropy[J]. Journal of Computer Research and Development，2013，50(6)：1177-1184. (in Chinese)
[18] PÁRRAGA-VALLEJ，GARCÍA-BERMÚDEZ R，ROJAS F，et al. Evaluating mutual information and chi-square metrics in text features selection process：a study case applied to the text classification in PubMed[C]//Proceedings of International Work-Conference on Bioinformatics and Biomedical Engineering. Berlin，Germany：Springer，2020：636-646.
[19] NAGPAL A，SINGH V. Feature selection from high dimensional data based on iterative qualitative mutual information[J]. Journal of Intelligent and Fuzzy Systems，2019，36(6)：5845-5856.
[20] CHAWLA N V，BOWYER K W，HALL L O，et al. SMOTE：synthetic minority over—sampling technique[J]. Journal of Artificial Intelligence Research，2002，16(1)：321-357.
[21] ARIDAS C K，KARLOS S，KANAS V G，et al. Uncertainty based under-sampling for learning naive Bayes classifiers under imbalanced data sets[J]. IEEE Access，2020，7：2122-2133.
[22] ZHANG S. Decision tree classifiers sensitive to heterogeneous costs[J]. Journal of Systems and Software，2012，85(4)：771-779.
[23] LOYOLAGONZALEZ O，MARTINEZTRINIDAD J F，CARRASCOOCHOA J A，et al. Cost-sensitive pattern-based classification for class imbalance problems[J]. IEEE Access，2019，7：60411-60427.
[24] ABDOH S F，RIZKA M A，MAGHRABY F A，et al. Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques[J]. IEEE Access，2018，5：59475-59485.
[25] SUN J，LEE Y，LI H，et al. Combining B&B-based hybrid feature selection and the imbalance-oriented multiple-classifier ensemble for imbalanced credit risk assessment[J]. Technological & Economic Development of Economy，2015，21(3)：351-378.
[26] BIAN J，PENG X，WANG Y，et al. An efficient cost-sensitive feature selection using chaos genetic algorithm for class imbalance problem[J]. Mathematical Problems in Engineering，2016(10)：1-9.
[27] MOAYEDIKIA A，ONG K，BOO Y L，et al. Feature selection for high dimensional imbalanced class data using harmony search[J]. Engineering Applications of Artificial Intelligence，2017，57：38-49.
[28] GUYON I，ELISSEEFF A. An introduction to variable and feature selection[J]. Journal of Machine Learning Research，2003，3(6)：1157-1182.
[29] JAKULIN A，BRATKO I. Testing the significance of attribute interactions[C]//Proceedings of the 21st IEEE International Conference on Machine Learning. Washington D. C.，USA：IEEE Press，2004：409-416.
[30] JAKULIN A. Attribute interactions in machine learning[D]. [S. l.]：University of Ljubljana，2003.

编辑　索书志