一种基于特征选择的不平衡数据分类算法

肖 鹰[1] 吴哲夫[1] 张 彤[1] 王中友[2]

[1](浙江工业大学信息学院 杭州 310023)
[2](浙江省通信产业服务有限公司 杭州 310050)

摘 要 不平衡数据分类是当前机器学习的研究热点，传统分类算法通常基于数据集平衡状态的前提，不能直接应用于不平衡数据的分类学习。针对不平衡数据分类问题，文章提出一种基于特征选择的改进不平衡分类提升算法，从数据集的不同类型属性来权衡对少数类样本的重要性，筛选出对有效预测分类出少数类样本更意义的属性，同时也起到了约减数据维度的目的。然后结合不平衡分类算法使数据达到平衡状态，最后针对原始算法错分样本权值增长过快问题提出新的改进方案，有效抑制权值的增长速度。实验结果表明，该算法能有效提高不平衡数据的分类性能，尤其是少数类的分类性能。

关键词 机器学习；特征选择；不平衡提升算法；分类预测
中图分类号 TP 18 **文献标志码** A

Feature Selection Based Classification Algorithm with Imbalanced Data

XIAO Ying[1] WU Zhefu[1] ZHANG Tong[1] WANG Zhongyou[2]

[1](*College of Information Engineering*, *Zhejiang University of Technology*, *Hangzhou* 310023, *China*)
[2](*Zhejiang Branch of China Communications Services Company*, *Hangzhou* 310050, *China*)

Abstract At present, imbalanced data classification is the research hotspot of machine learning. Traditional machine learning classification algorithms are usually used on balanced datasets, which cannot be directly applied to the imbalanced data. A new kind of imbalanced boosting algorithm based on feature selection was proposed to balance the importance for the minority class samples from different types of attributes of datasets, which not only could select the more meaningful attributes for the prediction of the minority class samples, but also reduce data dimension. Then, the imbalanced boosting algorithm was combined to make the datasets balanced. Finally, since the wrong sample weight of the original algorithm grew fast, a new algorithm which could restrain the growths of sample weight effectively was put forward. Experimental results show that the proposed algorithm can effectively improve the classification performance of imbalanced datasets, especially that of the minority class.

Keywords machine learning; feature selection; imbalanced boost algorithm; classification and predication

收稿日期：2015-03-30 **修回日期**：2015-05-05
基金项目：浙江省自然科学基金资助项目(LY13F010011)；浙江省重大科技专项(2014NM002)
作者简介：肖鹰，硕士研究生，研究方向为机器学习和大数据处理；吴哲夫(通讯作者)，副教授，研究方向为数据网络和数据挖掘，E-mail：wzf@zjut.edu.cn；张彤，硕士研究生，研究方向为云计算和数据挖掘；王中友，高级工程师，研究方向为物联网和信息处理。

1 引　言

分类算法是目前机器学习的一个热点内容。随着机器学习技术的愈趋成熟，分类算法得到了广泛的研究和应用，例如贝叶斯、决策树、支持向量机算法等[1,2]。这些算法大多都基于数据平衡状态，当数据不平衡时，分类器的分类往往偏向于多数类的类别，造成少数类无法被正确识别。而在许多实际应用中，少数类的判别往往比多数类的判别更具有价值，例如银行欺诈用户识别、医学疾病识别和网络黑客入侵等[3,4]。

数据不平衡是指数据集类间呈现不均等的数据分布[5]，在一些具体的案例中数据集有时会出现类别严重不平衡的现象，不平衡度甚至达到50∶1 以上。数据不平衡会导致分类器在预测新样本数据类别时出现泛化能力不足的问题，从而影响分类精度。

为了解决上述问题，近年来诸多学者提出的研究方案可分为两个方面：数据层面处理和算法层面处理[6-8]。

(1)数据层面

数据层面处理是指通过改变数据的原始分布使数据达到平衡状态，常用方法有欠采样和过采样。欠采样的基本思想是删除部分多数类样本使数据达到平衡状态，但这种方式会导致原始数据信息的大量丢失；过采样的基本思想是不断复制少数类样本，扩大其规模使数据达到平衡状态，这种方式的优点是能使原始数据信息得到较好的保留。一些改进的过采样算法主要包括：Chawla等[9]提出的少数类过采样技术(Synthetic Minority Over-sampling Technique，SMOTE)，这是一种通过在少数类样本之间进行线性插值的技术来实现增加少数类样本的方法，在一定程度上克服了分类器过度拟合的问题。此外，为了避免产生噪声和边际样本，Batista 等[10]提出了 SMOTE＋Tomek 算法。

(2)算法层面

算法层面处理是指修改算法在数据集上的偏置，从而使最终的决策平面偏向于少数类，提高少数类样本的识别率。此外，主动学习、集成学习也是处理不平衡数据的常见策略。例如，集成学习里的提升算法是通过关注错分样本、组合多个弱分类器的分类结果的方式来建立强分类器，从而提高分类性能。国内外的一些改进算法，诸如 AdaBoost(Adaptive Boosting)[11]；代价敏感学习技术与提升技术结合的算法，如 AdaCost(分类的时候采用 C4.5 算法作为基分类器)等；数据合成方法与提升技术结合的算法，如不平衡分类提升(SMOTEBoost)算法[11]等。其中，SMOTEBoost 算法是将 AdaBoost 与 SMOTE 算法结合，在每次迭代中插入新的合成样本，使每个子分类器更加关注少数类。

2 算　法

本文提出一种新的不平衡数据分类算法。首先采用特征选择算法权衡数据集中每个属性对预测分类属性的重要程度，筛选出对预测分类少数类样本更具有代表性的属性，再结合改进的 SMOTEBoost 算法，最后将数据集应用于分类建模，从而达到提高分类性能的目的。设计的算法流程如图 1 所示。

2.1 特征选择算法

当数据集维度较大时，数据集中往往有很多属性是冗余的，这就需要通过数据降维技术来处理。传统的数据降维技术有主成分分析[13]、因子分析等，这些技术的原理都是将原始的高维空间特征经过线性组合映射成新的特征，但合成的新特征往往不具备很好的解释性。特征选择是根据一定规则从数据集 D 的特征中选择数量为 $d(d<D)$ 的一组最优特征。一些较好的特征选择算法，如信息增益、互信息[14]等算法在非结构化数据集(如文本)分

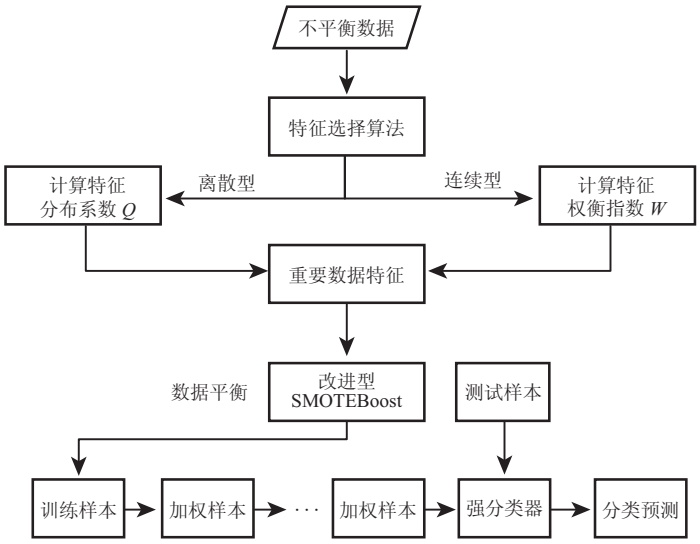

图 1 基于特征选择的不平衡数据分类算法设计流程图

Fig. 1 Design flowchart of imbalanced data sorting algorithm based on feature selection

类中应用广泛且具有良好的效果。而对于结构化数据集特征选择算法研究主要是从特征选择的搜索策略进行开展，如随机搜索策略引入遗传算法[15]、启发式搜索策略的前向选择和后向选择算法，这些算法都没有考虑特征之间的相关性，并且要综合考虑计算代价。另外，2009 年 Mark[16]提出的属性选择算法，消除了冗余的特征。本文将采取权衡属性对于预测分类少数类样本重要性的方式来进行特征选择，即通过两组不同的评估参数分析少数类样本在数据集中的特征分布来进行特征筛选，这种方式不仅具备解释性，也能有效减少分类器训练时间和计算代价。本文特征选择算法的方式有离散型属性特征选择和连续型属性特征选择两种。

2.1.1 离散型属性特征选择

通过分析少数类样本在每一个离散型属性里的分布特性来判别该离散型属性是否筛选，引进分布系数 Q 来权衡离散型属性的重要程度。分布系数 Q 是概率分布离散程度的归一化量度，计算方法如公式(1)所示。

$$Q=\frac{\overline{T}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(T_i-\overline{T}\right)^2}} \tag{1}$$

其中，T_i 是离散型属性在少数类样本中所占的比例；$\overline{T}$ 是离散型属性在少数类样本中的比例均值。当少数类样本在该离散型属性中的分布较为均匀时，此时 $\overline{T}\approx T_i$，$Q\approx 1$。为了能够更好地识别少数类样本，要保证少数类样本在该离散型属性中的分布尽可能不均匀。经过多次实验测试，当 $Q<<1$ 时，可以作为筛选离散型属性的参考标准。表 1 为 Monk2 数据集的离散型特征选择示例。

表 1 离散型属性特征选择

Table 1 Feature selection for discrete attributes

数据集	预测属性	离散属性	分布系数	是否筛选
Monk2	class	a2（1,2,3）	0.91	否
Monk2	class	a6（1,2,3,4）	0.43	是

2.1.2 连续型属性特征选择

通过引进权衡参数 W 来度量连续型属性对于少数类样本的重要程度，计算方法如公式(2)所示。

$$W=\frac{\left|Avg_1-Avg_2\right|}{\sqrt{\frac{Var_1}{Num_1}+\frac{Var_2}{Num_2}}} \tag{2}$$

其中，Avg_1 和 Avg_2 分别表示多数类和少数类的样本均值；Var_1 和 Var_2 分别为多数类和少数类的

样本方差；Num_1 和 Num_2 分别为多数类和少数类的样本个数。W 越大，则该属性对于预测少数类样本影响越显著。经过多次实验测试，当 $W>>3$ 时，可以作为筛选连续型属性的参考标准。表 2 为 Vowel 数据集的连续型特征选择示例。

本文基于属性重要性的特征选择算法的具体流程如下：

(1) 输入原始数据集 D；

(2) 分离出 D 中的多数类和少数类样本，多数类样本存放集合 L，少数类样本存放集合 S，计算集合 L 所包含的样本个数 Num_1 以及集合 S 所包含的样本个数 Num_2；

(3) 分离 D 中离散属性集合 X 和连续属性集合 Y；

(4) 对离散属性集合 X 进行遍历，计算评估指标 Q，如果 $Q<<1$，则输出筛选离散属性集合 X_{FS}；

(5) 对离散属性集合 X 和连续属性集合 Y 进行遍历，计算均值 Avg_1 和 Avg_2；

(6) 对离散属性集合 X 和连续属性集合 Y 进行遍历，计算方差 Var_1 和 Var_2；

(7) 对连续属性集合 Y 进行遍历，计算评估指标 W，如果 $W>>3$，输出筛选连续属性集合 Y_{FS}；

(8) X_{FS} 和 Y_{FS} 合并后，输出经过特征选择后的数据集 D_{FS}。

2.2 不平衡分类提升算法及改进

SMOTEBoost 算法实质上是一种数据合成与集成技术的结合算法，通过 SMOTE 技术对少数类样本进行人工合成，提高少数类样本的比例并降低数据过度倾斜。SMOTE 技术与 AdaBoost 的结合可以有效避免由于赋予少数类样本更大权重而造成的过度拟合，提高分类器的泛化能力。其中，AdaBoost 算法通过多轮训练得到若干弱分类器，并在每一轮训练后调整每个样本的权值，增加错分样本的权值，更加关注少数类样本，最后通过加权弱分类器得到强分类器。原始 SMOTEBoost 算法的具体流程如下：

(1) 输入训练样本集，$(x_1, y_1),(x_2, y_2),\cdots,(x_m, y_m)$，其中，$x_i \in X, y_i \in Y=\{-1,1\}$；

(2) 对样本权值进行初始化 $w_1(i)=1/m$；

(3) 用 SMOTE 算法合成 N 个少数类样本；

(4) 循环 $t=1, 2,\cdots, T$ 次，在加权训练样本集上用弱分类学习算法训练弱分类器 C_t 得到 h_t；

(5) 计算弱分类器 C_t 的训练误差 ε_t，其中，$\varepsilon_t=\sum_{h_t(x_i)\neq y_i} w_t(i)$，当 $\varepsilon_t=0$ 或 $\varepsilon_t>0.5$ 时结束训练；

(6) 计算弱分类器 C_t 的权值 $\alpha_t=\frac{1}{2}\ln(\frac{1-\varepsilon_t}{\varepsilon_t})$；

(7) 更新样本权值，Z_t 为归一化因子，

$$w_{t+1}(i)=\frac{w_{t+1}(i)}{Z_t}\times\begin{cases}e^{-\alpha_t}, & h_t(x_i)=y_i\\ e^{\alpha_t}, & h_t(x_i)\neq y_i\end{cases};$$

(8) 输出得到的强分类器

$$H(x)=sign(\sum_t \alpha_t h_t(x))$$

虽然 SMOTE 算法实现了数据平衡，但 AdaBoost 算法有个缺点就是错分样本的权值增长速度过快，这样会导致分类器过于夸大错分样本的作用，影响到总体的分类精度。针对这个问

表 2　连续型属性特征选择

Table 2　Feature selection for continuous attributes

数据集	类别	预测属性	连续属性	均值	方差	权衡指数 W	是否筛选 W
Vowel	多数类	type	F6	−0.002	0.199	0.5	否
Vowel	少数类	type	F6	−0.020	0.364	0.5	否
Vowel	多数类	type	F8	−0.269	0.319	6.4	是
Vowel	少数类	type	F8	−0.636	0.262	6.4	是

题，本文对原始算法步骤(5)更新训练样本权值的规则做了如下改进：

$$w_{t+1}(i)=\frac{w_t(i)}{Z_t}\times\begin{cases}e^{-\alpha_t/M}, & h_t(x_i)=y_i\\ e^{\alpha_t/M}, & h_t(x_i)\neq y_i\end{cases}$$

其中，M 为数据集的不平衡度(多数类样本/少数类样本)。由于 $\alpha_t>0$ 且 $M>1$，当样本被错误分类时，有 $0<\frac{\alpha_t}{M}<\alpha_t$。由于 $f(x)=e^x$ 是增函数，故 $e^{\frac{\alpha_t}{M}}<e^{\alpha_t}$，减缓了权值的增长速度；同理，当样本正确分类时也可以减缓权值的减慢速度。改进后的权值可以有效解决分类器夸大错分样本的问题，对总体的分类效果起到积极作用。

3 实验验证

3.1 数据集

为检验和评估本文算法的性能，实验选取了 10 个来自 UCI 的公开数据集[17]。表 3 为数据集的基本信息，包括数据集大小、不平衡度、属性类别以及本文特征算法筛选后的属性个数等。一般情况下，将不平衡度介于[1.5, 3.5)、[3.5, 9.5)、[9.5, +∞)分别定义为低度不平衡、中度不平衡、高度不平衡。由于改进的算法涉及不平衡度 M，故选择的 10 个数据集采自不同的不平衡度。

3.2 评估指标

机器学习中的分类器性能评估往往通过分类精度来度量，但对不平衡数据集采用这种方式就显得不大适合。一些针对少数类样本的分类评价指标可以有效评估不平衡数据集的分类性能[18]，例如查全率(recall)和查准率(precision)。在二分类情形下将更为重要的少数类视为正类，多数类视为负类。最终的预测可以通过混淆矩阵来表示，如表 4 所示。其中，TP(True Positive)是指被分类器正确分类的正类；TN(True Negative)是指被分类器正确分类的负类；FP(False Positive)是指被分类器错误分类的正类；FN(False Negative)是指被分类器错误分类的负类。

表 4 机器学习分类混淆矩阵

Table 4 Machine learning confusion matrix

分类	预测正类	预测负类
实际正类	TP	FN
实际负类	FP	TN

通过混淆矩阵产生的几个评价指标：

$$recall=\frac{TP}{TP+FN} \tag{3}$$

$$precision=\frac{TP}{TP+FP} \tag{4}$$

表 3 UCI 数据集的信息

Table 3 UCI datasets information

数据集	样本数	少数类	多数类	不平衡度	属性类别	特征选择后
Monk2	169	64	105	1.7	0 连续 6 离散	0 连续 3 离散
pima	768	268	500	1.8	9 连续 0 离散	4 连续 0 离散
IPLD	584	168	416	2.5	10 连续 0 离散	6 连续 0 离散
Glass	214	29	184	6.3	9 连续 0 离散	6 连续 0 离散
ecoli	336	35	301	8.6	8 连续 0 离散	5 连续 0 离散
Satimage	6 435	626	5 809	9.3	33 连续 0 离散	9 连续 0 离散
Vowel	988	90	898	9.9	10 连续 3 离散	6 连续 0 离散
Zoo	101	5	96	19.2	0 连续 17 离散	0 连续 3 离散
Yeast4	1 484	51	1 433	28.1	8 连续 0 离散	7 连续 0 离散
Page-block	5 473	115	5 358	46.6	10 连续 0 离散	9 连续 0 离散

$$G-mean=\sqrt{\frac{TP}{TP+FN}\times\frac{TN}{TP+FP}} \tag{5}$$

其中，式(3)和式(4)可以有效衡量少数类在分类预测中的准确性；而式(5)是综合考虑两个类的分类性能，兼顾正类和负类分类精度的平均。实验通过这三个评价指标来检验该算法的性能优劣。

3.3　实验结果和评估

在 R-3.0.3 实验环境下对本文算法进行了验证，其中，SMOTEBoost 采用的子分类器为 10 个，分类算法采用决策树(Decision Tree)算法，验证方式采用十折交叉验证(10-fold-cross-validation)。图 2～图 4 分别为收集的 10 个 UCI 数据集在不同算法下的查全率、查准率和 *G-mean*。通过比较原始 SMOTE 算法、原始 AdaBoost 算法、原始 SMOTEBoost 算法和本文算法的评价指标可以得出以下结论：

(1)分类性能上，原始 AdaBoost 算法和原始 SMOTEBoost 算法性能相当，并且均比原始 SMOTE 算法有较大幅度的提升。可见 AdaBoost 算法无论在处理平衡数据集还是不平衡数据集时，性能都是可观的。

(2)本文算法在特征选择前提下进一步改进 SMOTEBoost 算法的不足，使得绝大多数的数据集分类性能都得到了提升，特别是查全率提升幅度最大，而查准率的提升空间相对较小。

(3)实验验证发现，经过有效的特征选择后，少数类样本的查全率和查准率会有改善。由于样本被正确分类时样本权值会减小，导致个别数据集会出现多数类样本的准确率略微下降的现象。由于少数类样本更受关注，通过经过 *G-mean* 的有效调节，最终能够保证数据集总体及少数类样本的分类评价性能。

(4)对于不同的不平衡度数据集，*G-mean* 指标都有一定幅度的提高，特别是针对高度不平衡的数据集。由于改进了 Adaboost 算法的权值调整策略，对于高度不平衡的数据集，权值的改变速度都有所减弱，保证了总体分类性能的有效提升。

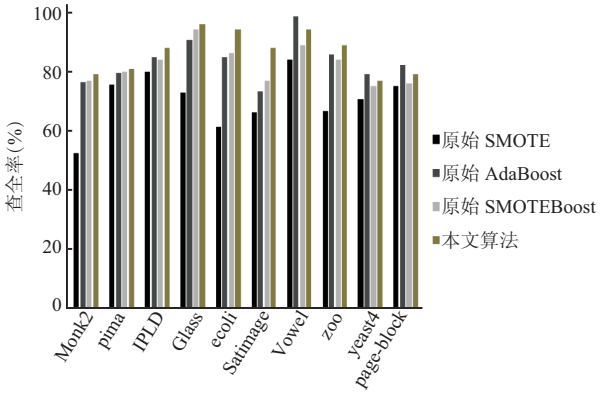

图 2　不同算法下的查全率比较

Fig. 2　Recall ratio of different algorithms

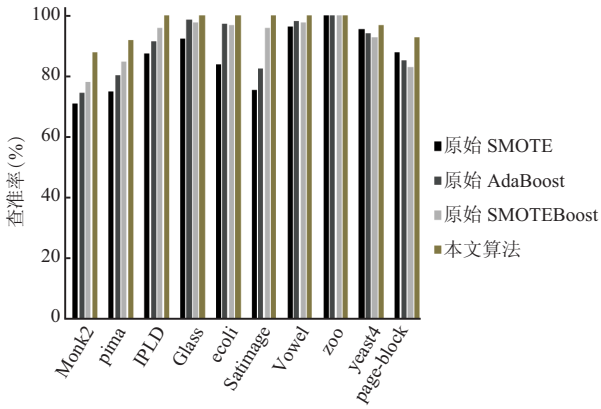

图 3　不同算法下的查准率比较

Fig. 3　Precision of different algorithms

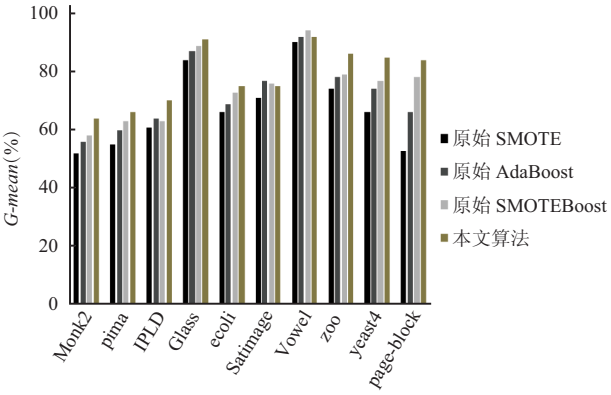

图 4　不同算法下的 *G-mean* 比较

Fig. 4　*G-mean* of different algorithms

4 结束语

本文提出了一种基于特征选择的不平衡数据分类算法，该算法通过有效的评估参数，筛选出有利于预测分类少数类样本的重要属性；再借鉴集成学习思想改进 SMOTEBoost 算法的不足，解决了错分样本权值改变速度过快的问题，并且在保证多数类分类性能损失不大的前提下，提高了少数类的分类性能，更具有实际意义。实验结果表明，该算法对提高不平衡数据中少数类的分类性能有一定程度的改善，对分类预测少类别样本具有很好的指导意义。

参 考 文 献

[1] 李玲俐. 数据挖掘中分类算法综述 [J]. 重庆师范大学学报: 自然科学版, 2011, 28(4): 44-47.

[2] 范明, 孟小峰. 数据挖掘概念与技术 [M]. 北京: 机械工业出版社, 2001.

[3] Chan PK, Stolfo SJ. Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection [C] // Proceedings of the Fourth International Conference on Knowledge Discovery & Data Mining, 1998.

[4] 朱卫, 沈玉琨. 乳腺癌普查资料的分析 [J]. 疾病控制杂志, 2002, 6(3): 253-254.

[5] He H, Garcia EA. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[6] 林智勇, 郝志峰, 杨晓伟. 不平衡数据分类的研究现状 [J].计算机应用研究, 2008, 25(2): 332-336.

[7] Garcia S, Derrac J, Isaac T, et al. Evolutionary-based selection of generalized instances for imbalanced classification [J]. Knowledge-Based Systems, 2012, 25(1): 3-12.

[8] 翟云, 杨炳儒, 曲武. 不平衡类数据挖掘研究综述 [J]. 计算机科学, 2010, 37(10): 27-32.

[9] Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[10] Tomek I. Two modifications of CNN [J]. IEEE Transactions on Systems, Man and Communications, 1976, 6: 769-772.

[11] 李勇, 刘战东, 张海军. 不平衡数据的集成分类算法综述 [J]. 计算机应用研究, 2014, 31(5): 1287-1291.

[12] Chawla NV, Lazarevic A, Hall LO, et al. SMOTEBoost: improving prediction of the minority class in boosting [C] // The 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, 2003: 107-119.

[13] Camacho J, Pico J, Ferrer A. Data understanding with PCA: structural and variance information plots [J]. Chemometrics and Intelligent Laboratory Systems, 2010, 100(1): 48-56.

[14] 徐燕, 李锦涛, 王斌, 等. 不均衡数据集上文本分类的特征选择研究 [J]. 计算机研究与发展, 2006, 44(Z2): 58-62.

[15] Goldberg DE. Genetic Algorithms in Search, Optimization, and Machine Learning [M]. New York: Addison-Wesley Professional, 1989.

[16] 邵进智. 基于属性间相关分析的属性选择算法研究 [D]. 北京: 北京交通大学, 2009.

[17] UCI. Machine Learning Repository [DB/OL]. [2015-03-09]. http://archive.ics.uci.edu/ml/datasets.html.

[18] 林智勇, 郝志峰, 杨晓伟. 若干评价准则对不平衡数据学习的影响 [J]. 华南理工学报: 自然科学版, 2010, 38(4): 147-155.