·人工智能及识别技术· 文章编号：1000-3428(2018)04-0268-06 文献标志码：A 中图分类号：TP311

面向不平衡数据集分类模型的优化研究

温雪岩[1]，陈家男[1]，景维鹏[1]，徐克生[2]

（1. 东北林业大学 信息与计算机工程学院，哈尔滨 150040；2. 国家林业局 哈尔滨林业机械研究所，哈尔滨 150086）

摘 要： 为提高不平衡数据集的分类效率，建立一种分类模型，从样本采样和分类算法两方面进行优化。对决策边界的少类样本进行循环过采样生成新样本集，并与决策边界外合成的少类样本集合并，提高样本的重要度。针对传统 ε-支持向量机（ε-SVM）在对不平衡数据集分类时超平面偏移的问题，引入正负惩罚系数和混合核函数，并利用客观的熵值法选取惩罚系数，提高分类算法的性能。实验结果表明，与标准的 SVM 算法相比，该分类模型在不平衡数据集分类上 F-measure 值平均提高 18.1%，具有较好的分类效果。

关键词： 文本分类；不均衡数据集；数据挖掘；样本重采样；熵值法

中文引用格式： 温雪岩，陈家男，景维鹏，等. 面向不平衡数据集分类模型的优化研究[J]. 计算机工程，2018，44(4)：268-273，293.

英文引用格式： WEN Xueyan，CHEN Jianan，JING Weipeng，et al. Research on Optimization of Classification Model for Imbalanced Data Set[J]. Computer Engineering，2018，44(4)：268-273，293.

Research on Optimization of Classification Model for Imbalanced Data Set

WEN Xueyan[1]，CHEN Jianan[1]，JING Weipeng[1]，XU Kesheng[2]

（1. College of Information and Computer Engineering，Northeast Forestry University，Harbin 150040，China；
2. Harbin Forestry Machinery Research Institute，State Forestry Administration，Harbin 150086，China）

【Abstract】 In order to improve the classification efficiency of unbalanced data sets，this paper proposes a classification model. The sample sampling and classification algorithm are optimized. A new sample set is generated by cyclic sampling of the few samples of the decision boundary，combined with the small sample sets synthesized outside the boundary of the decision-making，then the importance of the sample is improved. Aiming at the problem of hyperplane offset in classification of imbalanced data sets by traditional ε-Support Vector Machine（ε-SVM），the positive and negative penalty coefficients and the mixed kernel function are introduced. The objective entropy value method is used to select the penalty coefficients and the performance of the classification algorithm is improved. Experimental results show that compared with the standard SVM algorithm，the classification is better in the classification of imbalanced data sets，the average F-measure value is increased by 18.1%，and the better classification results are achieved.

【Key words】 text categorization；imbalanced data set；data mining；sample resampling；entropy method

DOI： 10.3969/j.issn.1000-3428.2018.04.043

0 概述

在这个信息大爆炸的时代，为了从海量数据中挖掘出有效信息[1]，许多实际应用的数据集需要进行分类处理，如防火墙过滤、入侵检测[2]和缺陷预测[3]等。但多数情况下，这些数据集是不平衡的，表现出来的现象是，数据集中各个样本之间的数量差距悬殊。在机器学习过程中，一般将数据集中关于类别分布的不均衡问题称为数据集的不均衡问题（Class Imbalance Problem of Data Set，CIPD），体现在样本的数量差异较大。采用传统分类方法解决 CIPD 时，分类结果往往倾向于多数类。对 CIPD 学习效果进行改善，提高 CIPD 的分类准确率是当前机器学习算法领域的热点之一[4-6]。

支持向量机以其效果稳定、精确度高的优点得到了广泛应用。但是在利用支持向量机（Support Vector Machine，SVM）对不均衡数据集分类时效果都不够理想，原因是 SVM 算法学习得到的超平面倾向于少数类样本，导致分类器性能较差。

过采样通过生成少类样本来减少数据的不均衡性。文献[7]提出 SMOTE 算法，该算法通过随机合成而不是复制少类样本的方式有效解决了过拟合的现象，但是由于没有对少类样本进行区域划分，致使合成的样本分布区域存在局限性。

基金项目： 国家重点研发计划项目（2016YFD0702105）。

作者简介： 温雪岩（1971—），男，副教授、硕士，主研方向为机器学习、数据挖掘；陈家男，硕士研究生；景维鹏，副教授、博士；徐克生，研究员。

收稿日期： 2017-11-06 **修回日期：** 2017-12-12 **E-mail：** wenxy2005@nefu.edu.cn

针对 SMOTE 算法的不足,文献[8]提出了 B-SMOTE 算法,用 SMOTE 算法对决策边界的少数类样本进行人工合成。文献[9]提出了对错分样本进行循环采样人工合成新样本的方法(L-SMOTE)。虽然这些方法有效地提升了 SMOTE 算法的性能,但是仍然存在一些不足。如 B-SMOTE 算法在执行过程中,忽略了决策边界外的少类样本中的重要信息;L-SMOTE 算法在执行过程中,忽视了错分样本中的噪声点,不断合成新的噪声样本,影响了分类精确度。

文献[10]通过精确选择参数 ε 值提高了 ε-SVM 在均衡与不均衡数据集上的分类精度。文献[11]引入双隶属度的非对称加权算法对混合核 SVM 的核函数进行优化,并将其应用到不平衡数据集分类中。以上 2 种方法有效改善了分类算法对不平衡数据集的分类效果,但是到目前为止,对于混合核 ε-SVM 的优化方法只涉及到预测方面,而关于混合核 ε-SVM 对不平衡数据集分类方面的优化方法还没有提出过。

针对以上不足,本文提出一种从样本采样和分类算法两方面同时优化的分类模型。在样本采集方面,给出一种面向决策边界少类样本循环过采样的 LD-SMOTE 算法,并将新生成的样本集与决策边界外新生成的少类样本进行合并。在分类算法方面,将正负惩罚系数引入到混合核 ε-SVM 中,并将更具有客观性的熵值法运用到惩罚系数的选择上。

1　基于决策边界的 L-SMOTE 优化方法

1.1　L-SMOTE 算法

和传统的 SMOTE 算法不同,L-SMOTE 算法关注的是影响分类平面的错分样本,根据错分样本循环合成新样本,提升这些关键样本的质量,提高分类的精确度。

但是该算法在执行时存在一定的缺陷,如图 1 所示,P_3、P_4 和 P_5 是少数类样本,P_1 和 P_2 是新生成的样本,P_2 是较为合理的合成样本,但是 P_1 的有效性却是值得商榷的,因为 P_1 生成的位置正好位于多数类的散列点中间,属于噪声点,根据 L-SMOTE 算法,P_1 点是错分样本点,采取错分样本的重采样,那么生成的新样本也必然是噪声点,循环执行将会严重影响分类效果。

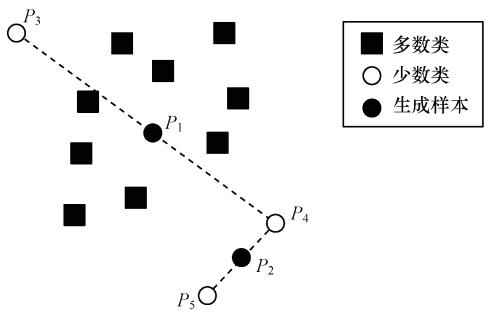

图 1　合成样本的有效性

1.2　D-SMOTE 算法

因为错分类样本主要集中在决策边界,只对决策边界的少类样本进行循环重采样就会有效避免噪声点的不断生成。针对决策边界少类样本的人工合成,本文提出一种基于样本间距的决策边界过采样算法(D-SMOTE)。

该算法的具体步骤如下:

步骤 1　计算出各个少数类样本点 a_i($i=1,2,\cdots,m$)到多数类样本点 b_j($j=1,2,\cdots,n$)的距离之和为 $\sum_{j=1}^{n} d(a_i,b_j)$。

步骤 2　求平均距离:$\bar{d}=\dfrac{\sum_{i=1}^{m}\sum_{j=1}^{n} d(a_i,b_j)}{m}$。

步骤 3　当少数类样本 a_i 满足 $\sum_{j=1}^{n} d(a_i,b_j) < \bar{d}$ 时,将 a_i 划分到决策边界样本集中,a_i 称为决策样本。

步骤 4　对各个决策样本计算在少数类样本集中的 k 近邻,从中任取一个 a_j,利用 a_j 和 a_i 两个样本,结合 SMOTE 算法合成新的样本。公式如下:

$$a_{\text{new}} = a_i + random(0,1) \times |a_i - a_j| \tag{1}$$

在对决策边界的少类样本进行人工合成时,本文用 D-SMOTE 算法取代传统的 B-SMOTE 算法,因为 B-SMOTE 算法在处理少数类样本极少的样本集时,往往会造成合成的新样本分布不均、过于集中的现象,而 D-SMOTE 通过比对数类和多数类样本的间距来确定决策边界样本,有效地控制了决策样本的分布范围,样本分布更均匀,提升了决策边界样本集的质量。

1.3　LD-SMOTE 算法

将本文提出的 D-SMOTE 算法与 L-SMOTE 算法相结合,得到 LD-SMOTE 算法。该算法的具体操作步骤如下:

输入　设原始样本数据 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_t,y_t)\}\in(\mathbb{R}^n\times Y)^t$ 中少数类为正类 P,多数类为负类 N,样本数量分别为 nP 和 nN。

步骤 1　用 D-SMOTE 算法选出少数类样本的决策样本集合,记为 P_d。

步骤 2　用标准 SMOTE 算法对少数类样本进行人工合成,合成后的新样本集合记作 P_l。

步骤 3　用标准 SMOTE 算法对 P_d 中的样本进行人工合成,生成新的样本集合,记为 P_e。

步骤 4　令 $P_d=P_e$,$P_{ld}=P_d+P_l$,重复步骤 3,直到 $P_{ld}=nN$。

P_{ld} 就是 LD-SMOTE 算法执行后最终得到的少数类样本集,与 B-SMOTE 算法合成的样本集不同,该样本集包含了非决策边界的少数类样本的重要信息,而且通过循环合成让决策边界的少数类样本能够反复学习,从而提高了最终合成的少数类样本集的质量。

该算法的伪码如下:

输入　样本集 T,少数类样本集 P,多数类样本集 N,少数

类样本数量 nP,多数类样本数量 nN

输出 最终生成的少数类样本集:P_{ld}

```
1.  P_d = D-SMOTE(P)
2.  P_l = SMOTE(P)
3.  P_e = SMOTE(P_d)
4.  P_d = P_e, P_ld = P_d + P_l
5.  While P_ld ! = nN
6.      P_e = SMOTE(P_d)
7.      P_d = P_e, P_ld = P_d + P_l
8.  Endwhile
```

2 基于熵值法的混合核 ε-SVM 优化方法

2.1 ε-SVM

SVM 分为线性可分、非线性可分以及需要核函数映射 3 种情况。设训练样本 $T=\{(x_i,y_i)\}$ $(i=1,2,\cdots,l)$,x_i为 SVM 的输入特征,y_i为类别标签,l为训练样本个数。基于二分类目标核函数 SVM 实现非线性划分的分类算法,其模型的原始问题可表示为:

$$\min_{w,b,\xi}\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i-1}^{l}\xi_i$$

$$\text{s.t. } y_i((\boldsymbol{w}\cdot\phi(x_i))+b)\geqslant 1-\xi_i,i=1,2,\cdots,$$

$$\xi_i\geqslant 0,i=1,2,\cdots,l \tag{2}$$

其中,$\boldsymbol{w}$ 是一个被确定的权重向量,C 和 ζ_i分别为惩罚系数和松弛变量。

传统的 ε-SVM 是在 SVM 算法模型中引入一个不敏感损失函数[12-13],如式(3)所示。ε 为参数,给定数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\}\in(\mathbb{R}^n\times\mathbb{R})^l$,其中,$x_i\in\mathbb{R}^n$,$y_i\in\mathbb{R}$,$i=1,2,\cdots,l$,运用式(4)的线性函数集合来估计回归函数。

$$L(y,f(x,a))=L(|y-f(x,a)|\varepsilon) \tag{3}$$

$$f(x_i)=\omega\cdot\phi(x_i)+b \tag{4}$$

其中,ω 为回归系数,$\varphi(x_i)$为输入空间到特征空间的映射函数,b 为阈值。

2.2 混合核函数

混合核函数是指通过组合的方式将单个核函数合并成新的核函数,同时考虑局部核函数和全局核函数的特性,将两者的优势充分发挥,弥补两者在应用上的不足。由于 Polynomial 核函数有着良好的全局性质,而 RBF 核函数则是局部性强,本文将这 2 种核函数组合起来,得到学习能力和推广性都很强的混合核函数,其构造形式如下:

$$k_{\text{Poly}}=[(x\times x_i)+1]^q \tag{5}$$

$$k_{\text{RBF}}=\exp(-\|x-x_i\|^2/\sigma^2) \tag{6}$$

$$k(x,x')=\lambda k_{\text{Poly}}(x,x')+(1-\lambda)k_{\text{RBF}}(x,x') \tag{7}$$

$$\int_{-\infty}^{\infty}\phi^2(x)\mathrm{d}x<\infty$$

$$\int_{-\infty}^{\infty}k(x,x')\phi(x)\phi(x')\mathrm{d}x\mathrm{d}x'>0 \tag{8}$$

式(5)和式(6)分别表示 Polynomial 核函数和 RBF 核函数。式(7)表示构造的混合核函数,其中的 λ 表示的是单个核函数在混合核函数中占有的比重,$0<\lambda<1$。式(8)表示的是 Mercer 核函数约束条件。将 $k(x,x')$带入到式(8)中,符合 Mercer 核函数约束条件[14-15]。文献[14]已对 $k(x,x')$的线性组合进行验证,满足 Mercer 条件,这里不作具体论证。

将混合核函数植入到传统的 ε-SVM,构造成混合核 ε-SVM,分类算法具有了更强大的学习能力和泛化能力。

2.3 混合核 ε-SVM 的优化

通过 LD-SMOTE 算法生成新样本能够使样本数据集变得均衡,但是扩充样本集合时,并不能改变原有样本分布的外围轮廓特征,这就意味着对分类问题中分类边界的影响比较小,所以利用混合核 ε-SVM 训练样本时超平面依然会偏向少数类,分类效果依然会受到影响。受文献[16]的启发,在样本训练过程中,将正负惩罚系数 $C+$ 和 $C-$ 引入到混合核 ε-SVM 中,并在正负惩罚系数的选择上运用了熵值法进行优化。

1)正负惩罚系数

二分类平面图如图 2 所示。圆和星分别表示多数类样本和少数类本,虚线表示的是使用一个惩罚系数时的分割效果。在这种情况下,如果对 2 类样本赋予不同的惩罚系数 $C+$ 和 $C-$,灵活地调节误差代价,最终就会得到理想的分类效果,图中的实线表示调整正负惩罚系数后的分割效果。

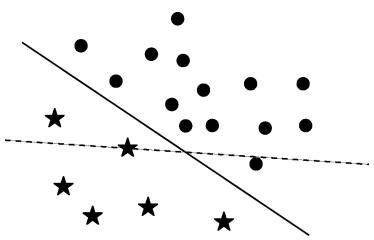

图 2 二分类平面图

通过以上分析,结合式(2)~式(4)、式(7),最终推导出改进的混合核 ε-SVM 的约束化问题:

$$\min\frac{1}{2}\|\omega\|^2+C^+\frac{1}{l}\sum_{y_i=+1}^{l}(\xi_i+\xi_i^*)+$$

$$C^-\frac{1}{l}\sum_{y_i=-1}^{l}(\xi_i+\xi_i^*)$$

$$\text{s.t.}\begin{cases}y_i-(\omega\phi(x_i))-b\leqslant\varepsilon+\xi_i\\(\omega(\phi(x_i))+b-y_i\leqslant\varepsilon+\xi_i^*\\\xi_i,\xi_i^*\geqslant 0\end{cases}$$

$$i=1,2,\cdots,l \tag{9}$$

其中,ζ_i和 ζ_{i*} 为松弛因子,$C+$ 和 $C-$ 表示少类样本(正类)和多类样本(负类)的惩罚系数。

在惩罚系数 $C+$ 和 $C-$ 的选择上,传统方法都没有考虑到样本内各个属性的相对变化程度,使得惩罚系数在选择上过分依赖个人经验,具有很强的主观性。

2)熵值法确定正负惩罚系数

本文将信息熵的思想用于到惩罚系数的选择上,提出熵值法[17]确定惩罚系数的方法。根据多数类和少数类样本的离散程度确定不同的惩罚系数,避开主观人为因素的干扰,即一种客观的赋值方法,选出的惩罚系数更具有价值,其具体实现方法如下:

假设原始样本 T 中正类样本 S^+ 包含 n 个子类:$S^+ = \{s_1^+, s_2^+, \cdots, s_n^+\}$,各个事件的概率分布 $P^+ = \{p_1^+, p_2^+, \cdots, p_n^+\}$,$\sum_{i=1}^{n} \ln p_i^+ = 1$。其中,$s_{i+}$ 的信息量可用 $-\ln p_i^+$ 来衡量,正类样本的 S^+ 熵值为:

$$H(I+) = H(p_1^+, p_2^+, \cdots, p_n^+) = -\sum_{i=1}^{n}(p_i^+ \ln p_i^+) \tag{10}$$

同理,负类样本 S^- 包含 m 个子类,负类样本 S^- 的熵值为:

$$H(I-) = H(p_1^-, p_2^-, \cdots, p_m^-) = -\sum_{i=1}^{m}(p_i^- \ln p_i^-) \tag{11}$$

计算正类样本 S^+ 和负类样本 S^- 的差异性系数,将式(10)、式(11)代入得:

$$d^+ = 1 - H(I+) = 1 + \sum_{i=1}^{n}(p_i^+ \ln p_i^+) \tag{12}$$

$$d^- = 1 - H(I-) = 1 + \sum_{i=1}^{m}(p_i^- \ln p_i^-) \tag{13}$$

其中,d^+、d^- 分别表示正类和负类的差异性系数。令 $C^+ = C$,得:

$$C^- = C^+ \times \frac{d^-}{d^+ + d^-} \tag{14}$$

通过以上优化方法,使得分类算法在对不平衡数据集分类时的性能进一步提高。在参数的选择上,本文利用文献[18]提出的 AMPSO 算法进行参数寻优。将优化后的混合核 ε-SVM 算法和 LD-SMOTE 算法相结合,最终得出本文的分类模型,如图 3 所示。

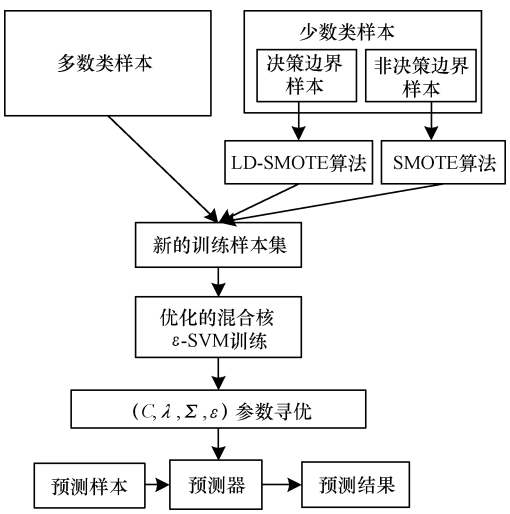

图 3　本文的分类模型

本文的分类模型伪码如下:

输入　训练样本集中的多数类样本 D1,训练样本集中的少数类样本 D2,测试数据集 D3

输出　D3 数据集的分类结果

```
1. 计算 LD-SMOTE 决策边界样本语料库
2. 计算 SMOTE 非决策边界样本语料库
3. DNEW = LD-SMOTE + SMOTE
4. 使用式(11)~式(16)训练模型 ε-SVM 参数
5. result = [ ]
6. for i in range(0,len(D3))
7.       result_D3 = ε-SVM(D3[i])
8.       result.append(result_D3)
9. end for
10. return result
```

3　实验设置与结果分析

3.1　数据来源

为了验证本文提出的分类模型的分类效果,采用 UCI 数据集[19]中的 6 个不平衡数据集作为测试性能的数据,各个数据集的信息如表 1 所示,其中的比例表示的是少数类与多数类的比值。

表 1　不平衡数据集

数据集	少数类数量	总数	特征数量	比例
abalone (20 vs 8,9,10,11)	26	2 404	8	2:183
car (good vs others)	69	1 728	6	1:24
yeast (vac vs cyt)	30	493	10	2:31
computer hardware (0-20 vs others)	31	209	6	16:89
ecoli (pp vs others)	52	336	7	2:11
breasttissue (con vs others)	14	106	9	5:33

3.2　分析指标

在对不平衡数据集进行分类时,常用的分析指标有 3 种,分别是查准率(Precision)p、敏感度(Sensitivity)s 和综合考虑 F-measure 指标 f,具体公式如下:

$$p = \frac{TP}{TP + FP} \tag{15}$$

$$s = \frac{TP}{TP + FN} \tag{16}$$

$$f = \frac{2 \cdot s \cdot p}{s + p} \times 100\% \tag{17}$$

其中,FP 表示将负类样本错分成正类的数目,FN 是指将正类样本错分成负类的数目,TP 表示正类样本

被正确分类的个数。

3.3 实验结果分析

将数据的 70% 作为样本的训练集,30% 作为样本的测试集。利用 word2vec 对样本进行词向量的训练,生成向量空间。实验中所有的数据集都采用了 5 折交叉验证,以便于验证分类模型的性能。

1)近邻值参数 k 值的确定

k 值的选择对于本文提出的 LD-SMOTE 算法至关重要,将 k 值范围设置在 2 ~ 10 之间进行讨论。实验数据采用 UCI 不平衡数据集中 30% 的测试数据,对 6 个数据集分别进行测试,将不同 k 值下的 F-measure值作为评价指标,F-measure 取 6 个数据集的平均值。用本文提出的改进混合核 ε-SVM 作为分类算法,图 4 表示的是在本文分类算法下,不同 k 值取得 F-measure 值的折线图。当 k 值到 6 时,F-measure达到最高值,因此在接下来的实验中,将 LD-SMOTE 算法的 k 值设定为 6。

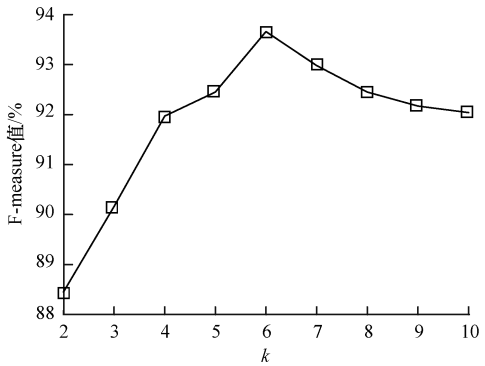

图 4 不同 k 值下的实验结果

2)3 种分类算法的实验结果对比

实验采用 abalone 作为测试数据集,该数据集是一个极度不均衡的数据集。该实验用本文提出的 LD-SMOTE 算法进行样本过采样处理,然后用改进的混合核 ε-SVM 算法、改进的单核 ε-SVM 算法(采用 RBF 核函数,运用熵值法确定正负惩罚系数)和传统的 ε-SVM 算法(采用 RBF 核函数)进行学习和最终的预测,利用文献[18]提出的 AMPSO 算法对 3 种分类算法进行参数优化,下面的实验均用该方法进行参数优化。实验采用查准率、敏感度和 F-measure 值作为评估标准。利用 AMPSO 算法寻找出最优参数组合如表 2 所示,实验结果如图 5 所示。

表 2 参数寻优结果

分类算法	$(C,\varepsilon,\lambda,\sigma)$和$(C,\varepsilon,\sigma)$
单核 ε-SVM	(20.412,0.192 1,1.025)
混合核 ε-SVM	(12.524,0.211,0.863,1.423)

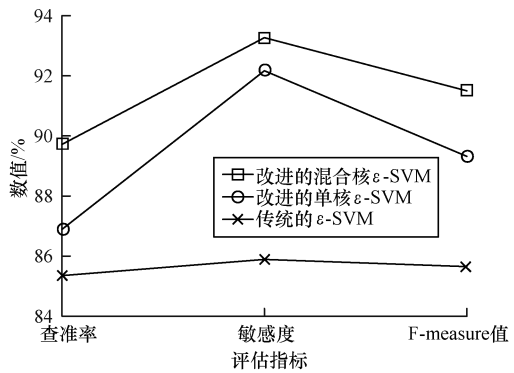

图 5 3 种分类算法的实验结果

如图 5 所示,本文提出的改进混合核 ε-SVM 的 3 个评估指标比其他 2 种分类算法明显提高。因为采用了熵值法确定正负惩罚系数,所以在处理极度不均衡数据集时,2 种改进算法的分类精度要比传统 ε-SVM 算法有所提高。而混合核比单核分类精确度高是因为混合核函数具有更强的泛化能力和鲁棒性。

3)传统 SMOTE 算法和 LD-SMOTE 算法的分类结果对比

实验采用 6 个不平衡数据集作为测试数据集,分类算法均采用标准 SVM,F-measure 值作为评估标准,实验结果如图 6 所示。

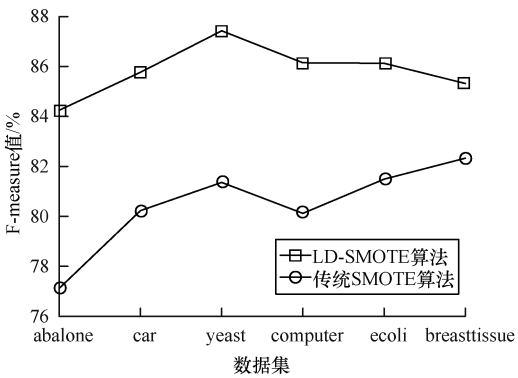

图 6 2 种采样算法的实验结果

实验结果表明,LD-SMOTE + SVM 的分类精确度比 SMOTE + SVM 算法有明显提升。但是当样本集极度不均衡时(abalone 数据集),只对训练样本进行重采样处理,不对分类算法进行改进,分类精确度明显偏低。

4)3 种分类方法实验结果对比

为了更好地验证本文提出的分类模型的性能,在相同实验条件下,与标准的 SVM[20] 和 SD-ISMOTE + SVM[21] 进行实验比较。实验选用 F-measure指作为评价标准。实验结果如表 3 和图 7 所示。

表 3　3 种分类方法的 F-measure 值对比　%

数据集	标准 SVM	SD-ISMOTE + SVM	本文分类模型
abalone	62.31	82.12	91.72
car	74.31	86.21	92.58
yeast	82.54	91.36	94.33
computer	80.36	93.09	94.74
ecoli	81.82	94.23	95.15
breasttissue	79.33	92.16	95.12

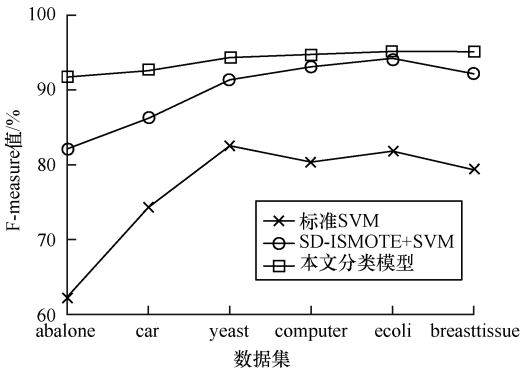

图 7　3 种分类方法的实验结果

表 3 显示了 3 个算法对 6 个数据集进行分类预测的实验结果。实验结果表明,本文提出的分类模型比 SD-ISMOTE + SVM 和标准 SVM 在 F-measure 值上取得了明显提升。F-measure 值比标准 SVM 平均高出 18.1%,比 SD-ISMOTE + SVM 平均高出 4.35%,说明本文提出的分类模型在对不平衡数据集进行分类时具有明显优势。

在图 7 中,标准 SVM 算法的折线始终在图像的最下方,尤其是在 car 和 abalone 两个数据点上,其 F-measure 值达到了最低。产生的原因是,标准 SVM 算法没有对训练样本做任何处理,尤其是当数据集的正负类样本数量差距悬殊,分类平面严重向另一侧倾斜时,如果直接采用 SVM 算法对测试样本进行分类,分类的精确度会大大降低。而本文的分类模型和 SD-ISMOTE + SVM 都针对不平衡的训练样本集进行过采样处理,都获得了较好的分类效果。但是本文的分类模型的分类精确度更高一些,原因在于在分类算法的改进上,将正负惩罚系数、熵值法和多核学习引入到支持向量机中,进一步提高了分类模型的分类性能。

4　结束语

本文构建一种面向不平衡数据集的分类模型。在样本集过采样优化方面,针对 L-SMOTE 算法对错分样本进行循环采样时不断生成噪声点的问题,通过对决策边界样本进行循环过采样的方法生成新的样本集,并将第一次过采样时生成的决策边界范围外的少类样本添加到新生成的样本集中,提升了样本的重要度。在算法优化方面,针对传统的 ε-SVM 算法在对不平衡数据集分类时超平面偏移的问题,把正负惩罚系数引入到支持向量机模型中,并且采用了更具有客观性的熵值法选取惩罚系数。同时构造了混合核 ε-SVM,加强了支持向量机的泛化能力和学习能力,分类精确度明显提高。下一步将改进粒子群算法,选出最优参数,并减少算法运行消耗的时间。

参考文献

[1] GARCA S,LUENGO J,HERRERA F. Data preprocessing in data mining[M]. Berlin,Germany:Springer,2016.

[2] 沈夏炯,王　龙,韩道军. 人工蜂群优化的 BP 神经网络在入侵检测中的应用[J]. 计算机工程,2016,42(2):190-194.

[3] YU Qiao, JIANG Shujuan, ZHANG Yanmei. The performance stability of defect prediction models with class imbalance: an empirical study [J]. IEICE Transactions on Information & Systems,2017,100(2):265-272.

[4] ZHANG Chunkai,WANG Guoquan,ZHOU Ying,et al. A new approach for imbalanced data classification based on minimize loss learning[C]//Proceedings of the 2nd International Conference on Data Science in Cyberspace. Washington D. C. ,USA:IEEE Press,2017:82-87.

[5] NAPIERALA K,STEFANOWSKI J. Types of minority class examples and their influence on learning classifiers from imbalanced Data [J]. Journal of Intelligent Information Systems,2016,46(3):563-597.

[6] HERRERA F. Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced Big Data[J]. Fuzzy Sets & Systems,2015,258(3):5-38.

[7] 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.

[8] HAN Hui, WANG Wenyuan, MAO Binghuan. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning [C]//Proceedings of International Conference on intelligent Computing. Berlin,Germany:Springer,2005:878-887.

[9] 衣柏衡,朱建军,李　杰. 基于改进 SMOTE 的小额贷款公司客户信用风险非均衡 SVM 分类[J]. 中国管理科学,2016,24(3):24-30.

[10] 杨俊燕,张优云,朱永生. ε 不敏感损失函数支持向量机分类性能研究[J]. 西安交通大学学报,2007,41(11):1315-1320.

[11] 赵淑娟. 基于非对称加权和核方法的不平衡数据集[D]. 南京:南京邮电大学,2013.

[12] ALZATE C, SUYKENS J. Kernel component analysis using an epsilon-insensitive robust loss function [J]. IEEE Transactions on Neural Networks,2008,19(9):1583-1598.

[13] WATANABE K. Vector quantization based on ε-insensitive mixture models[J]. Neurocomputing,2015,165(3):32-37.

(下转第 293 页)

5　结束语

本文提出的算法通过对裂缝图像全局和局部视觉显著性的有效融合,增强裂缝信息的同时削弱了噪声干扰,从而准确地分割出裂缝区域。实验结果表明,该算法可以应用于复杂环境下的高速路面裂缝检测。后续工作将研究裂缝特性,基于视觉显著性模型进一步增强裂缝显著信息。

参考文献

[1] CHENG H D, MIYOJIM M. Automatic pavement distress detection system [J]. Journal of Information Sciences,1998,108(1):219-240.

[2] ZALAMA E,GÓMEZ-GARCÍA-BERMEJO J,MEDINA R, et al. Road crack detection using visual features extracted by gabor filters [J]. Computer-Aided Civil and Infrastructure Engineering,2013,29(5):342-358.

[3] SHI Y, CUI L, QI Z, et al. Automatic road crack detection using random structured forests [J]. IEEE Transactions on Intelligent Transportation Systems, 2016,17(12):3434-3445.

[4] 高建贞,任明武,唐振民,等. 路面裂缝的自动检测与识别[J]. 计算机工程,2003,29(2):149-150.

[5] LI Q, LIU X. Novel approach to pavement image segmentation based on neighboring difference histogram method[C]//Proceedings of 2008 Congress on Image and Signal Processing. Washington D. C., USA: IEEE Press,2008:792-796.

[6] 闫茂德,伯绍波,贺昱曜. 一种基于形态学的路面裂缝图像检测与分析方法[J]. 工程图学学报,2008,29(2):142-147.

[7] NEJAD F M, ZAKERI H. An optimum feature extraction method based on wavelet-radon transform and dynamic neural network for pavement distress classification [J]. Expert Systems with Applications, 2011,38(8):9442-9460.

[8] 马常霞,赵春霞,胡　勇,等. 结合 NSCT 和图像形态学的路面裂缝检测[J]. 计算机辅助设计与图形学学报,2009,21(12):1761-1767.

[9] 徐　威,唐振民,吕建勇. 基于图像显著性的路面裂缝检测[J]. 中国图象图形学报,2013,18(1):69-77.

[10] OLIVEIRA H, CORREIA P L. Automatic road crack detection and characterization [J]. IEEE Transactions on Intelligent Transportation Systems,2013,14(1):155-168.

[11] HU Y,ZHAO C,WANG H. Automatic pavement crack detection using texture and shape descriptors [J]. IETE Technical Review,2010,27(5):398-405.

[12] 钱　彬,唐振民,徐　威,等. 子块鉴别分析的路面裂缝检测[J]. 中国图象图形学报,2015,20(12):1652-1663.

[13] ZHANG L,YANG F,ZHANG Y D,et al. Road crack detection using deep convolutional neural network[C]//Proceedings of IEEE International Conference on Image Processing. Washington D.C.,USA:IEEE Press,2016:3708-3712.

[14] 张巧荣,景　丽,肖会敏,等. 利用视觉显著性的图像分割方法[J]. 中国图象图形学报,2011,16(5):767-772.

[15] LI S,LU H,LIN Z,et al. Adaptive metric learning for saliency detection [J]. IEEE Transactions on Image Processing,2015,24(11):3321-3331.

[16] ITTI L,KOCH C,NIEBUR E. A model of saliency-based visual attention for rapid scene analysis [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998,20(11):1254-1259.

[17] LAND E H. The retinex theory of color vision [J]. Scientific American,1977,237(6):108.

[18] PERONA P,MALIK J. Scale-space and edge detection using anisotropic diffusion [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1990,12(7): 629-639.

[19] ACHANTA R, HEMAMI S, ESTRADA F, et al. Frequency-tuned salient region detection [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2009: 1080-1085.

[20] 徐志刚,赵祥模,宋焕生. 基于直方图估计和形状分析的沥青路面裂缝识别算法[J]. 仪器仪表学报,2010,31(10):2260-2266.

[21] 李立寒,黄天元,王国培,等. 沥青路面破损程度分级与破损换算系数取值的探讨[J]. 公路交通科技,2005,22(9):40-43.

[22] ACHANTA R, ESTRADA F, WILS P, et al. Salient region detection and segmentation[C]//Proceedings of International Conference on Computer Vision Systems. Berlin,Germany:Springer-Verlag,2008:66-75.

[23] ZHAI Y,SHAH M. Visual attention detection in video sequences using spatiotemporal cues[C]//Proceedings of ACM Multimedia. New York, USA: ACM Press, 2006:815-824.

编辑　金胡考

(上接第 273 页)

[14] 唐　奇,王红瑞,许新宜,等. 基于混合核函数 SVM 水文时序模型及其应用[J]. 系统工程理论与实践,2014,34(2):521-529.

[15] 颜根廷,马广富,肖余之. 一种混合核函数支持向量机算法[J]. 哈尔滨工业大学学报,2007,39(11):1704-1706.

[16] 刘东启,陈志坚,徐　银,等. 面向不平衡数据分类的复合 SVM 算法研究[EB/OL]. [2017-11-06]. http://kns.cnki.net/kcms/detail/51.1196.TP.20170401.1738.050.html.

[17] 朱喜安,魏国栋. 熵值法中无量纲化方法优良标准的探讨[J]. 统计与决策,2015(2):12-15.

[18] FRANK A, ASUNCION A. UCI machine learning repository [EB/OL]. [2017-11-06]. http://archive.ics.uci.edu/ml.

[19] 刘文贞,陈红岩,李孝禄,等. 基于自适应变异粒子群算法的混合核 ε-SVM 在混合气体定量分析中的应用[J]. 传感技术学报,2016,29(9):1464-1470.

[20] 常甜甜. 支持向量机学习算法若干问题的研究[D]. 西安:西安电子科技大学,2010.

[21] 古　平,杨　炀. 面向不均衡数据集中少数类细分的过采样算法[J]. 计算机工程,2017,43(2):241-247.

编辑　顾逸斐