基于分形 SMOTE 重采样集成算法圈定区域化探异常*

李 江[a], 金 辉[b], 刘 伟[b]
(中国地质大学 a. 信息工程学院; b. 计算机学院, 武汉 430074)

摘 要: 基于分形自相似性理论改进 SMOTE 算法,实现数据集的均衡化。结合集成学习 Adaboost 技术更新样本权值,改善非均衡数据的分类性能,并对云南个旧锡铜多金属矿床进行了仿真实验,结果表明新算法预测结果能较好地检测成矿异常,为成矿预测与评价提供新的解决途径。

关键词: 化探异常; 不均衡数据; SMOTE; 分形; 集成学习; Adaboost
中图分类号: TP181; P628 **文献标志码**: A **文章编号**: 1001-3695(2012)10-3744-04
doi:10.3969/j.issn.1001-3695.2012.10.036

Novel improved SMOTE resampling integrated algorithm based on fractal for geochemical anomalies evaluation

LI Jiang[a], JIN Hui[b], LIU Wei[b]
(*a. School of Information Engineering, b. School of Computer Science, China University of Geosciences, Wuhan* 430074, *China*)

Abstract: Based on the similarity theory of fractal, this paper put forward a new SMOTE re-sampling algorithm. According to the real distribution of samples, a few sets of data samples should be reconstructed to realize the equalization of data sets. The new algorithm combined Adaboost technology, according to the classification of the error rate updating weights of samples to improve the classification performance of imbalanced data. The new algorithm was based on the simulation experiment on the research objection of polymetallic deposits such as tin and copper from Gejiu, Yunnan province. The experimental results show that predicted results for the new algorithm delineating regional geochemical anomalies are better than traditional methods, which can identify the geochemical anomaly accurately.

Key words: geochemical anomaly; imbalanced data; SMOTE; fractal; integrated learning; Adaboost

0 引言

地质异常是指在成分、结构、构造或成因序次上与周围环境有明显差异的地质体或地质体的组合,其中化探异常识别是成矿预测和资源评价的关键。区域化探异常识别中传统地质统计方法具有无偏、最优等特点,但要求数据呈正态分布,而实际应用往往不符合统计假设。近年来,分形理论被大量应用于地球化学异常确定,但基本思路还是采取单元素值来确定背景值,存在需要平滑处理数据、对样品中特高品位敏感等问题,因此,寻找能体现地球化学数据空间结构和非线性特征的异常识别方法具有重要的研究价值。地理空间的非平稳性使得地质异常现象并不十分均衡,也就是有矿样本数目远远低于无矿样本数目,因此本质上化探异常识别问题可归为不均衡数据分类问题。传统机器学习分类算法往往基于三点假设[1]:a)追求最大分类正确率;b)不同分类错误代价相同;c)数据集中不同类别包含的样本数目大致相当。在区域化探数据集中,有矿、无矿样本数目并不均衡,不符合上述假设,如果采用传统研究方法处理,往往会偏向多数类样本即无矿样本而忽略少数类样本即有矿样本,导致将测试样本全部判别为大类,虽然总体分类正确率很高,但小类有矿异常样本识别率却非常低。而在成矿识别中,人们更关心的是少数类即有矿样本的分类正确率,因此,有效提高少数类的分类性能是成矿异常识别亟待解决的问题。本文拟将非均衡数据分类问题引入到区域化探异常识别中,以有效识别区域化探异常。

目前,非均衡数据挖掘研究主要从数据和算法两个层面来解决分类的有偏性[2]。在数据层面,通过重构训练样本集均衡各类别数据,包括欠采样(under-sampling)和过采样(over-sampling)两类算法。欠取样主要包括 Condensed Nearest Neighbor Rule、Neighborhood Cleaning Rule、One-sided Selection、Tomek Links 等方法。通过寻找边界样本和噪声样本,有选择地去掉部分大类样本,但同时也忽略了部分有用信息[3]。过采样以 Chawla 等人[4]提出的 SMOTE(synthetic minority over-sampling technique)算法为代表,通过保持样本分布特性均衡各类别数据,提高少数类的分类性能,但重构的少数类样本有可能会进入多数类样本空间;Han 等人[5]提出 Borderline-SMOTE 技术,只对边界样本进行过取样,在适当区域内插值增加有价值的样本,但样本的选定存在一定盲目性。针对 SMOTE 算法没有考虑样本真实分布特性的缺点,谷琼[6]提出自适应选择近邻混合重取样算法,并与欠取样算法结合,有效提高合成样本的质量;许丹丹等人[7]采取在少数类实例及其最近邻少数类实例构成的 n 维球体内进行随机插值的方法改进算法;孙晓燕等人[8]采用遗传算法进行过取样。算法层面

收稿日期: 2012-03-06; **修回日期**: 2012-04-17 **基金项目**: 国家自然科学基金资助项目(40972206);中央高校基本科研业务费专项资金资助项目(1323520909)

作者简介: 李江(1968-),男,湖北武汉人,博士研究生,主要研究方向为智能计算、GIS 应用(ls8583@sina.com);金辉(1987-),男,湖北黄石人,硕士研究生,主要研究方向为智能计算;刘伟(1986-),男,山东人,硕士研究生,主要研究方向为智能计算.

主要通过采取各类样本间的代价函数、对不同类样本设置不同权值、改变概率密度、调整分类边界等措施。另外,实际数据的多样性使得单一分类方法往往不够有效,采用集成学习方法构造强分类器也可以提高少数类的分类性能。Chawla 等人[9]将 SMOTE 方法与标准的 Boosting 过程相结合,提出了 SMOTE-Boost 算法,但该算法仅通过 SMOTE 生成新样本改变类分布;李正欣等人[10]将 SMOTE 技术嵌入 Adaboost 算法中,提出 SMOTEboostSVM 集成方法,通过 Adaboost 中的权值更新解决非均衡数据分类的偏斜问题。

过采样的核心思想是希望按数据的真实分布特性生成新样本,而分形理论能较好地描述具有特征维度、极不规则但自相似性的复杂现象,因此,本文利用分形自相似性特点,在 SMOTE 算法基础上提出基于分形的 SMOTE 新算法。

1　算法理论基础

1.1　分形理论

1975 年 Mandelbrot 提出的分形理论,主要研究具有特征维度、极不规则但具有自相似性的复杂现象,如连绵的山脉、矿藏的分布等。自相似性是分形理论的核心,即在任何标度下物体的形状和结构不变[11,12]。定量描述这种自相似性的参数称为分维 D。

定义 1　分形集或分形。假设集合 $A \in E^n$(E^n 是 n 维欧式空间),其中 D_f 为 A 的 Hausdorff 维数,D_t 为 A 的拓扑维。如果公式 $D_f \geqslant D_t$ 成立,那么集合 A 即为分形集。

对地质异常,最有用的分维数计算是相似维或容量维 D_0。

定义 2　相似维或容量维 D_0。

$$D_0 = \lim_{r \to 0} \frac{\ln N(r)}{\ln (1/r)} \tag{1}$$

其中:r 为在测量地质体边界的长度时的测量尺度,$N(r)$ 为覆盖地质体边界的最少次数。可将这一定义推广到 n 维欧式空间 E^n 中。

1.2　SMOTE 算法

传统基于样本复制的过采样算法由于决策区间过小往往会引起过拟合,2002 年 Chawla 等人提出 SMOTE 过采样算法,通过人工合成新样本点来提高少数类的分类性能。

算法 1　SMOTE 算法

输入:少数类样本的数目 N;SMOTE 过采样率 $R\%$;最近邻域的数目 K。

输出:合成$(R/100) \times N$ 个少数类样本。

a)计算少数类样本 n 的 K 个最近少数类邻;

b)随机选取 K 个邻近点中的一个样本 y;

c)计算样本 n 与 y 的全部属性差值 diff = $-y[\text{attr}] - n[\text{attr}]$;

d)随机产生介于 0-1 的随机数 rand;

e)按 newSmaple = $n[\text{attr}]$ + diff × rand 合成少数类样本;

f)循环执行 b) ~ e),直到产生的少数类样本是 $R\%$ 个;

g)循环执行 b) ~ f),直到 n 类少数样本合成完。

1.3　AdaBoost 算法

AdaBoost 算法是一种集成学习算法,在弱分类器基础上组合生成强分类[13]。算法将所有训练样本赋予相同的权值,训练得到弱分类器,按分类的错误率更新样本的权值,增加分类错误样本权重,同时减少正确样本的权重。重复上述操作,最后集成每轮的弱分类器为强分类器。

算法 2　AdaBoost 算法

输入:输入 N 个带有类标签的样本(x_i, y_i),其中 $i = 1,2,\cdots N$,$y_i \in (-1, +1)$;迭代次数 T。

输出:分类器 $H(x)$。

a)初始化样本权值 $w_i = 1/N(i = 1,2,\cdots,N)$。

b)循环迭代;$t < T$ 时循环:

(a)对带有权重的训练样本用 SMO 算法进行训练学习,得到一个弱分类器 h_t;

(b)计算分类器 h_t 的错误率 $\varepsilon_t = \frac{1}{n}\sum_{i=1}^{n} pr(h_t(x_i) \neq y_i)$;

(c)如果 $\varepsilon_t \geqslant 0.5$ 或者 $\varepsilon_t = 0$,则令 $t = 1$,返回 b);

(d)计算加权参数 $\alpha_i = 0.5 \times [\ln (1 - \varepsilon_t)/\varepsilon_t]$;

(e)更新样本的权值 $w_{t+1}(i) = w_t(i)\exp[-\alpha_i y_i h_t(x_i)]/Z_i$,$Z_i$ 为归一化因子;

(f)$t = t + 1$。

c)得到强分类器 $H(x) = \text{sign}(\sum_{t=1}^{T} \alpha_i h_t(x))$。

2　基于分形 SMOTE 的 Adaboost 分类算法

2.1　基于分形的 SMOTE 算法

由于 SMOTE 算法的近邻选择策略存在一定盲目性,重构的少数类样本有的会进入多数类样本空间,因此本文利用分形自相似性提出新的重采样 FXSMOTE 算法,使新样本与原始数据集保持相似的空间分布特性[14]。算法具体描述如下:

算法 3　基于分形的 SMOTE 的算法(FXSMOTE)

输入:少数类样本的数目 N;FXSMOTE 过采样率 $R\%$;最近邻域的数目 K。

输出:合成$(R/100) \times N$ 个少数类样本。

a)计算少数类样本 n 的 K 个最近少数类邻。

b)随机选取 K 个邻近点中的两个不同的样本 y_1、y_2。

c)分别计算样本 n 与样本 y_1、y_2 之间的全部属性差值 diff1、diiff2、diif3:

diff1 = $y_1[\text{attr}] - n[\text{attr}]$;

diiff2 = $y_2[\text{attr}] - n[\text{attr}]$;

diff3 = $y_3[\text{attr}] - y2[\text{attr}]$。

d)合成少数类样本。

由两个近邻样本和原少数类样本组成的三角形进行分形操作,新合成样本如下:

newSample[1] = $n[\text{attr}] + 0.5 \times$ diff1

newSample[2] = $n[\text{attr}] + 0.5 \times$ diff2

newSample[3] = $y2[\text{attr}] + 0.5 \times$ diff3

e)循环执行 b) ~ d),直到产生的少数类样本是 $R\%$ 个。

f)重复执行 b) ~ e),直到 n 个少数类样本都合成完。

2.2　基于分形 SMOTE 的 Adaboost 分类算法(ABFSMOTE)

基于分形 SMOTE 算法虽然能按原始数据的空间分布特征合成新样本,但如果不均衡数据、不平衡率严重歪曲,少数类样本成为困难样本,会出现大数吃小数的情况。集成 AdaBoost 算法在迭代过程中更加关注分类错误的样本,本文将 Adaboost 方法嵌入到分形 SMOTE 算法中,以进一步提高困难样本预测性能。算法基本框架如下:

算法 4　ABFSMOTE 算法

输入:输入 N 个带有类标签的样本(x_i, y_i),其中 $i = 1,2,\cdots,N$,$y_i \in (-1, +1)$;迭代次数 T;FXSMOTE 过采样率 $R\%$;最近邻域的数目 K。

a)获取 N 个带有类标签的样本中的少数类样本,记总数为 m,用 FXSMOTE 方法进行处理。

(a)计算少数类样本 n 的 K 个最近少数类邻;随机选取 K 个邻近点中的两个不同的样本 y_1、y_2;

(b)分别计算样本 n 与样本 y_1、y_2 之间的全部属性差值 diff1、diff2,diif3,由两个近邻样本和原少数类样本组成的三角形进行分形操作,并合成新的样本;

(c)重复执行(a)和(b),直到 n 个少数类样本都合成完,最终生成 $(R/100)\times m$ 个少数类样本。

b)对合成之后的样本集进行权值初始化,此时权值为 $w_i = \frac{1}{[N+m\times(R/100)]}$。

c)循环迭代;$t<T$ 时循环:

(a)对带有权重的训练样本用 SMO 算法进行训练学习,得到一个弱分类器 h_t。

(b)计算分类器 h_t 的错误率 $\varepsilon_t = \frac{1}{n}\sum_{i=1}^{n} pr(h_t(x_i)\neq y_i)$,如果 $\varepsilon_t \geqslant 0.5$ 或者 $\varepsilon_t = 0$,则令 $t=1$,返回(a);

(c)计算加权参数 $\alpha_i = 0.5\times[\ln(1-\varepsilon_t)/\varepsilon_t]$,更新样本的权值 $w_{t+1}(i) = w_t(i)\exp[-\alpha_i y_i h_t(x_i)]/Z_i$,$Z_i$ 为归一化因子;

(d)$t=t+1$。

d)得到强分类器 $H(x) = \text{sign}(\sum_{t=1}^{T}\alpha_i h_t(x))$。

3 仿真实验

本文选取云南个旧锡铜多金属矿床为研究对象,个旧矿区位于云南东南部,其中心地理位置为东经 103°09′26″、北纬 23°22′40″,海拔标高 1 300 m ~ 2 600 m,面积 2 400 km^2,处在欧亚板块被太平洋、印度板块俯冲碰撞相接的部位,属于中国东南微板块西南缘盆地中的南盘江凹断褶束之西南隅。矿区分布在北东、北西和南北方向多个褶皱断裂带的交汇处[15]。个旧地区是锡铜多金属成矿区,笔者深入研究云南个旧锡铜多金属矿区,找出与成矿关系最密切的化探变量,选取 Sn、Cu、Pb、Zn 等 39 种共计 524 条 1:20 万水系沉积物化探数据进行仿真实验,其中已经勘明的有矿点 41 个、无矿点 483 个,无矿与有矿的不平衡率为 11.78(比例为 483:41),为典型非均衡数据集。

3.1 实验参数设定

本文首先采用分形 SMOTE 方法对化探数据进行过采样处理,选择 SMO 算法作为基分类器,将分类结果与标准的 SMO 算法、基于 SMOTE 的 SMO 算法、基于 ABFSMOTE 的 SMO 等算法进行性能对比。实验采用十折交叉验证,将 10 次实验结果的平均值作为最后结果。实验设置分形插值次数为 2,近邻参数 K 取 5,选择 Linear Kernel 作为 SMO 核函数。具体算法参数设置如表 1 所示。

表 1 参数值设置

parameter	SMO	SMOTE + SMO	FXSMOTE + SMO	ABFSMOTE + SMO
nearest neighbores	5	5	5	5
percentage R%	—	400%/1000%	400%/1000%	400%/1000%
random seed	1	1	1	1
kernel	$K(x,y)=\langle x,y\rangle$	$K(x,y)=\langle x,y\rangle$	$K(x,y)=\langle x,y\rangle$	$K(x,y)=\langle x,y\rangle$

3.2 评价指标

传统分类方法对多数类有较高的识别率,而对少数类识别率却很低,以分类精度作为不均衡数据的评价标准并不合适。本文引入 G-mean、ROC Area 等测试指标对算法进行综合评价。

定义 3 G-mean 指标。G-mean 也称几何平均准则,由 Kubat 等人于 1997 年提出,是一种有效衡量不平衡数据分类效果的准则。

$$\text{G-mean} = \sqrt{acc^{+}\times acc^{-}} \tag{2}$$

其中:acc^{+} 为少数样本的精度,acc^{-} 为多数样本的精度[1]。G-mean 指标综合考虑了两类样本的精度,能更好地衡量不平衡数据分类器的性能。

定义 4 ROC Area 指标。ROC 曲线能较全面地描述分类器的性能,由于不能定量分析,采用 ROC Area 值表示[16]。ROC Area 值表示 ROC 曲线下的面积(AUC),其计算公式为

$$\text{ROC Area(AUC)} = \frac{1}{2}\sum_{i=1}^{n^{+}}\sum_{j=1}^{n^{-}} Pr(f(x_i^{+})>f(x_j^{-})) \tag{3}$$

其中:n^{+} 为少数类样本的个数,n^{-} 为多数类样本的个数。对于任一少数类样本,若分类算法 f 将其分类为少数类的概率大于多数类的概率,则记为 1。ROC Area 值越接近 1,模型的预测效果越好。

3.3 实验结果与分析

根据上述实验参数设定,本文对区域化探数据进行仿真实验,其结果如表 2 所示。为方便对比,各算法评测指标表现最好的结果用黑体标出,次好的结果用斜体标出

表 2 数据集 SMOTE、FSMOTE、ABFSMOTE 算法实验结果比较

	TP Rate	FP Rate	Recall	F-measure	ROC	G-mean
原始数据	1	0.976	1	0.96	0.512	0.155
	0.024	0	0.024	0.048		
SMOTE (400%)	0.965	0.717	0.965	0.85	0.624	*0.523*
	0.283	0.035	0.283	0.414		
FXSMOTE (400%)	0.965	0.663	0.965	0.859	*0.651*	**0.570**
	0.337	0.035	*0.337*	*0.474*		
ABFSMOTE (400%)	0.203	0.054	0.203	0.311	**0.686**	0.438
	0.946	0.797	**0.946**	**0.495**		
SMOTE (1000%)	0.814	0.295	0.814	0.779	0.759	*0.758*
	0.705	0.186	0.705	0.74		
FXSMOTE (1000%)	0.81	0.191	0.81	0.815	*0.809*	**0.809**
	0.809	0.19	**0.809**	**0.804**		
ABFSMOTE (1000%)	0.762	0.248	0.762	0.764	**0.837**	0.757
	0.752	0.238	*0.752*	*0.749*		

从表 2 可看出,原始不均衡数据如果采用标准 SMO 分类器,虽然整体分类精度较高,但少数类(有矿类)样本的预测效果很差,几乎为 0。也就是模型在外推时几乎没有识别出有矿样本,而少数类样本正是本文要重点关注的,因此标准 SMO 分类器几乎不能满足实际需求。改进后的 FXSMOTE 分类算法,各项评测指标 F-measure、ROC、G-mean 相对标准的 SMO 均表现较好,虽然多数类样本分类精度降低,但是大大提高了少数类样本的分类精度,这是由于分形的自相似性使得合成的少数类新样本更符合原始数据的空间分布,改变数据的不均衡性,分类器的性能得到大大提高。特别是当过采样率提高到 1000% 时,即数据基本平衡时分类性能最佳,相比过采样率 400% 算法的 ROC Area 和 G-mean 值都有很大的提高。表 2 中嵌入 Adaboost 的分形 SMOTE 算法虽然 G-mean 值反而降低了,但 ROC Area 的值有所提高,这是因为集成 Adaboost 算法更关注少数类样本,通过牺牲多数类的准确率来提高少数类的精度,以达到提高分类器实际性能的目的。

图 1、2 分别为不同过采样率 200% ~ 2000% 下各算法运行结果的比较,其中横坐标代表过采样率,纵坐标代表 G-mean 和 ROC Area 的值。从图 1、2 可以看出,不同的过采样率 R,FXSMOTE 算法的各评测指标表现均优于 SMOTE 算法,而 ABFSMOTE 受到 Adaboost 权值更新的影响,其 G-mean 值在不同的采样率 R 下,相对于 FXSMOTE 算法有所下降,但是 ROC

Area 值在 SMOTE、FXSMOTE 比较中表现最好。另外不平衡率相差越大,其分类性能越不好,只有当过采样率的取值使得数据集中两类样本数趋于平衡时,分类性能最佳。

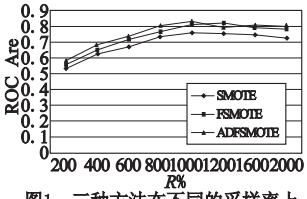

图1　三种方法在不同的采样率上 ROC Area值的比较

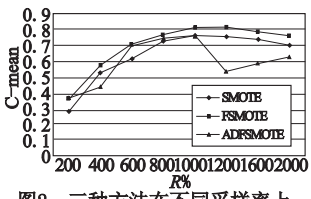

图2　三种方法在不同采样率上 G-mean的比较

为进一步说明新算法的预测效果,从数据集中随机选取 101 个样本(包括 6 个有矿样本,其他为无矿样本)。标准的 SMO 和分形 SMOTE 算法得到的混淆矩阵结果分别如图 3、4 所示。

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-measure   ROC Area  Class
                 1         0.333     0.979     1         0.99      0.833     无矿
                 0.667     0         1         0.667     0.8       0.833     有矿
Weighted Avg.    0.98      0.314     0.981     0.98      0.978     0.833

=== Confusion Matrix ===

  a  b   <-- classified as
 95  0 |  a = 无矿
  2  4 |  b = 有矿
```

图3　标准SMO算法分类结果

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-measure   ROC Area  Class
                 1         0         1         1         1         1         无矿
                 1         0         1         1         1         1         有矿
Weighted Avg.    1         0         1         1         1         1

=== Confusion Matrix ===

  a  b   <-- classified as
 95  0 |  a = 无矿
  0 12 |  b = 有矿
```

图4　分形SMOTE算法分类结果

图 5 为原始数据(左)与分形 SMOTE 处理(右)预测结果的对比。从图 5 可看出,标准 SMO 算法预测时有两个样本(表 2 的第 33 条和第 55 条)出错,而图 5 用分形 SMOTE 算法对这组数据集重采样后预测结果完全正确。

inst#,	actual,	predicted,	error,	probability distribution		inst#,	actual,	predicted,	error,	probability distribution	
33	2:有矿	1:无矿	+	*1	0	33	2:有矿	2:有矿		0	*1
34	1:无矿	1:无矿		*1	0	34	1:无矿	1:无矿		*1	0
35	1:无矿	1:无矿		*1	0	35	1:无矿	1:无矿		*1	0
36	1:无矿	1:无矿		*1	0	36	1:无矿	1:无矿		*1	0
37	1:无矿	1:无矿		*1	0	37	1:无矿	1:无矿		*1	0
38	1:无矿	1:无矿		*1	0	38	1:无矿	1:无矿		*1	0
39	2:有矿	2:有矿		0	*1	39	2:有矿	2:有矿		0	*1
40	1:无矿	1:无矿		*1	0	40	1:无矿	1:无矿		*1	0
41	1:无矿	1:无矿		*1	0	41	1:无矿	1:无矿		*1	0
42	1:无矿	1:无矿		*1	0	42	1:无矿	1:无矿		*1	0
43	1:无矿	1:无矿		*1	0	43	1:无矿	1:无矿		*1	0
44	1:无矿	1:无矿		*1	0	44	1:无矿	1:无矿		*1	0
45	1:无矿	1:无矿		*1	0	45	1:无矿	1:无矿		*1	0
46	1:无矿	1:无矿		*1	0	46	1:无矿	1:无矿		*1	0
47	1:无矿	1:无矿		*1	0	47	1:无矿	1:无矿		*1	0
48	1:无矿	1:无矿		*1	0	48	1:无矿	1:无矿		*1	0
49	1:无矿	1:无矿		*1	0	49	1:无矿	1:无矿		*1	0
50	1:无矿	1:无矿		*1	0	50	1:无矿	1:无矿		*1	0
51	1:无矿	1:无矿		*1	0	51	1:无矿	1:无矿		*1	0
52	1:无矿	1:无矿		*1	0	52	1:无矿	1:无矿		*1	0
53	1:无矿	1:无矿		*1	0	53	1:无矿	1:无矿		*1	0
54	1:无矿	1:无矿		*1	0	54	1:无矿	1:无矿		*1	0
55	2:有矿	1:无矿	+	*1	0	55	2:有矿	2:有矿		0	*1
56	1:无矿	1:无矿		*1	0	56	1:无矿	1:无矿		*1	0

图5　原始数据(左)与分形SMOTE处理(右)预测结果对比

通过上述实验分析可以看出,基于分形的 SMOTE 过采样新算法较好地利用了分形的自相似性特点,生成的新样本更符合数据的真实分布。在过采样新算法均衡化数据集基础上,通过嵌入集成学习 Adaboost 算法,能更关注分类错误的样本,进一步提高小类有矿异常样本识别率。对云南个旧锡铜多金属矿床进行的仿真实验结果与研究区域的实际情况比较吻合,圈定的区域化探异常好于传统方法,能更准确地找出成矿异常,为实际工程应用提供依据。

4　结束语

本文将非均衡数据分类问题引入到异常识别中以有效识别区域化探异常。针对成矿异常识别中人们更关心的是少数类即有矿样本的分类正确率,本文在传统重取样算法的基础上,提出一种基于分形的 SMOTE 新型重取样算法。该算法较好地利用了分形的自相似性特点,生成的新样本更符合数据的真实分布;同时,嵌入的集成学习 Adaboost 算法进一步提高了小类有矿异常样本识别率。在云南个旧锡铜多金属矿床进行的仿真实验结果与研究区域的实际情况比较吻合,圈定的区域化探异常好于传统方法,能更准确地找出成矿异常,为实际工程应用和进行矿产资源定量预测与评价提供了新的解决途径。

参考文献:

[1] PROBOST F. Machine learning from imbalanced data sets 101[C]// Proc of AAAI Workshop on Imbalanced Data Sets. 2000.

[2] CHAWLA N V, JAPKOWICZ N, KOTCA A. Editorial: special issue on learning from imbalanced data sets[J]. SIGKDD Explorations, 2004,6(1):1-6.

[3] CHEN Lei-chen, CAI Zhi-hua, CHEN Lu, *et al*. A novel differential evolution-clustering hybrid resampling algorithm on imbalanced datasets[C]//Proc of the 3rd International Conference on Knowledge Discovery and Data Mining. 2010:81-85.

[4] CHAWLA N V, BOWYER K W, HALL L O, *et al*. SMOTE: synthetic minority over-sampling technique[J]. Journal of Articial Intelligence Research,2002,16(1):321-357.

[5] HAN Hui,WANG Wen-yuan, MAO Bing-huan. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]// Lecture Notes in Computer Science, vol 3644. Berlin: Springer-Verlag,2005:878-887.

[6] 谷琼. 面向非均衡数据集的机器学习及在地学数据处理中的应用[D]. 武汉:中国地质大学,2009.

[7] 许丹丹,王勇,蔡立军. 面向不均衡数据集的 ISMOTE 算法[J]. 计算机应用,2011,31(9):2399-2401.

[8] 孙晓燕,张化祥. 用于不均衡数据集分类的 KNN 算法[J]. 计算机工程与应用,2011,47(28):143-146.

[9] CHAWLA N V, LAZAREVIC A, HALL L O, *et al*. SMOTEBoost: improving prediction of the minority class in boosting[C]//Proc of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases. 2003:107-119.

[10] 李正欣,赵林度. 基于 SMOTEBoost 的非均衡数据集 SVM 分类器[J]. 系统工程, 2008,26(5):116-119.

[11] 谢淑云,鲍征宇. 多重分形与地球化学元素的分布规律[J]. 地质地球化学,2003,31(3):97-102.

[12] 成秋明. 地质异常的奇异性度量与隐伏源致矿异常识别[J]. 地球科学(中国地质大学学报),2011,36(2): 307-316.

[13] 吴广潮,陈奇刚. 不平衡数据集中的组合分类算法[J]. 计算机工程与设计,2007,28(23):5687-5690.

[14] ZHANG Dong-mei, LIU Wei. A novel improved SMOTE resampling algorithm based on fractal[J]. Journal of Computational Information Systems,2011,7(6):1027-1034.

[15] 刘才泽,胡光道. 个旧地区化探数据的各向异性及东西矿区的对比研究[J]. 地质与勘探,2007,43(6):81-85.

[16] FAWCETT T. ROC graphs: notes and practical considerations for researchers, HPL-2003-4[R]. [S.l.]: HPLaboratories,2004.

[17] 周智勇. 基于分形与地质统计学理论的矿床建模技术研究及实践[D]. 长沙:中南大学,2005.