doi:10.3799/dqkx.2010.032

新型混合重取样算法在岩爆预测中的应用

谷 琼[1,2],蔡之华[2],朱 莉[2],王贤明[3]

1. 襄樊学院数学与计算机科学学院,湖北襄樊 441053
2. 中国地质大学计算机学院,湖北武汉 430074
3. 温州大学瓯江学院信息系,浙江温州 325035

摘要:针对岩爆现象发生的不均衡及发生机理受多因素影响的问题,在分析重取样技术的基础上,设计并实现了自适应选择近邻的混合重取样算法,并将其用于岩爆危险性预测.该方法结合过取样和欠取样方法的优势,改进了SMOTE过取样算法在产生合成样本过程中存在的盲目性及只能复制生成数值属性的问题,新算法能根据实例样本集内部分布的真实特性,自适应调整近邻选择策略,对不同属性的数据采取不同的复制方法生成新的少数类实例,控制和提高合成样本的质量;并通过对合成之后的数据集,用改进的邻域清理方法进行适当程度欠取样,去掉多数类中的冗余实例和边界上的噪音数据,减少其规模,在一定程度上达到相对均衡,从而,可有效地处理非均衡数据分类问题,提高分类器的性能.该算法在VCR采场岩爆实例上进行实验,预测的结果与实际情况完全一致,表明在工程实例岩爆危险性实例数据非均衡情况下实施混合重取样方案是可行的,预测准确率高,具有良好的工程应用前景.采用该方法可找到岩爆发生的主控因素,为深部开采工程的合理设计与安全施工提供科学依据.

关键词:岩爆;灾害;不均衡数据集;预测;合成少数类过取样;欠取样.

中图分类号:TU457 **文章编号**:1000-2383(2010)02-0311-06 **收稿日期**:2009-04-26

A Novel Hybrid Re-Sampling Algorithm and Its Application in Predicting Rockburst

GU Qiong[1,2], CAI Zhi-hua[2], ZHU Li[2], WANG Xian-ming[3]

1. *Faculty of Mathematics & Computer Science, Xiangfan University, Xiangfan* 441053, *China*
2. *School of Computer, China University of Geosciences, Wuhan* 430074, *China*
3. *Department of Information Science & Technology, Oujiang College, Wenzhou University, Wenzhou* 325035, *China*

Abstract: Because of poor understanding about the mechanism of rockbust and about the effect factors, the statistic data of large amounts of rockburst are typical imbalanced data sets (IDS). On the basis of analyzing re-sampling technology, a novel hybrid re-sampling technique based on Automated Adaptive Selection of the Number of Nearest Neighbors (ADSNN-Hybrid RS) is proposed and applied to study the prediction of rockburst. This method takes advantage of both technology of improved Synthetic Minority Over-sampling Technique (SMOTE) method and Neighborhood Cleaning Rule (NCR) data cleaning method. In the procedure of over-sampling with the SMOTE method, blindfold new synthetic minority class examples by randomly interpolating pairs of closest neighbors were added into the minority class; and data sets with nominal features can not be dealt with. These two problems were solved by the automated adaptive selection of nearest neighbors and adjusting the neighbor selective strategy. As a consequence, the quality of the new samples can be well controlled. In the procedure of under-sampling, by using the improved under-sampling technique of neighborhood cleaning rule, borderline majority class examples and the noisy or redundant data were removed. The main motivation behind these methods is not only to balance the training data, but also to remove noisy examples lying on the wrong side of the decision border. The removal of noisy examples might aid in finding better-defined class clusters, therefore, allow the creation of simpler models with better generalization capabilities. In turn, it promises effective processing of IDS and a considerably enhanced classifier performance. The VCR rockburst data sets were employed as a sample IDS for classification and prediction. By adding extra artificial minority class samples as the expanded train-

基金项目:国家高技术研究发展"863"计划(No. 2009AA12Z117);襄樊学院规划项目(No. 2009YA012).
作者简介:谷琼(1973-),女,博士研究生,主要从事地学信息相关方面的智能计算等研究工作. E-mail: gujone@163.com

ing set, experiment was conducted, which yields exactly consistent prediction results with the actual situation. The ADSNN-Hybrid RS and classification scheme we developed is feasible and reasonable for applications of IDS from engineering. Thus this method can be readily implemented to determine the controlling factors of engineering. Such a prediction can provide reasonable and sufficient guidance to design a safe construction scheme in deep mining engineering.

Key words: rockburst; disasters; imbalanced dataset; prediction; SMOTE; under-sampling.

岩爆是深部开挖过程中一种常见地质灾害,是处于高应力或极限平衡状态的岩体或地质结构体在开挖活动的扰动下,其内部储存的应变能瞬间释放,造成开挖空间周围部分岩石从母岩体中急剧、猛烈地突出或弹射出来的一种动力学现象. 岩爆具有突然性和猛烈性,常表现为岩石片状剥落、严重片帮、岩片崩落、岩片弹射等现象,有时还伴有爆裂声响(葛启发和冯夏庭,2008). 随着矿山开采和地下工程规模的不断扩大,岩爆问题日益突出. 岩爆可造成超挖和初期支护失效,对人们的生产活动构成严重危害,轻则影响工程进度,给生产和地下工程以及矿山安全造成一定的经济损失,重则对施工人员的生命构成严重威胁,还可能诱发地震.

岩爆发生机理的研究表明,岩爆的发生既受复杂地质因素的影响,同时又受工程环境因素和人为开挖因素的影响. 在采矿诱发下,岩爆的发生受到地质结构、采矿结构及其布局、采矿的推进方向与地质结构的关系、支护效果等许多因素的控制. 岩爆与其影响因素之间存在着极其复杂的非线性关系,如何有效地预测岩爆以减轻岩爆引起的灾害是非常必要的.

随着人类地下空间开发利用的不断深入,岩爆问题的研究越来越受到人们的重视. 近年来,国内外学者运用各种理论,对岩爆的发生机理、分类预测、数值模拟方法等进行大量研究,在总结岩爆事例特征的基础上,针对岩爆危险性的预测问题,提出了基于专家系统(冯夏庭,2000)、物元模型和可拓学理论预测(杨莹春和诸静,2001)、人工神经网络(陈海军等,2002)、灰色系统最优归类(姜彤等,2003)、支持向量机(赵洪波,2005)等预测方法,均取得了良好的效果.

针对岩爆小样本及岩爆发生的可能性危险类别不均衡的情况,如果采用传统的预测方法,则可能出现总的分类预测精度非常高,但我们真正关心的类别却未达到最佳预测效果,即没有考虑预测训练数据的非均衡情况发生,使分类结果偏好于大类别数据,而忽视了小类别数据. 本文提出一种新型混合重取样算法(hybrid re-sampling algorithm based on automated adaptive selection of the number of Nearest neighbors,简记为 ADSNN-Hybrid RS),并针对南非科学研究院建立的 VCR 采场岩爆实例数据,通过人工生成部分少数类实例作为训练数据进行仿真实验,预测的岩爆危险性状态与实际情况完全一致. 采用该方法还可找到岩爆发生的主控因素,为深部开采工程的合理设计与安全施工提供科学依据.

1 重取样算法

重取样算法通过对训练集进行预处理,再用预处理过的数据训练分类器,以降低在训练过程中对小类别数据的忽视和不公平待遇,减小训练集中各类别不均衡对分类性能造成的影响,从而提高少数类的分类性能. 重取样包括过取样和欠取样. 最简单的过取样方法是随机复制少数类样本,缺点是没有给少数类增加任何新的信息,容易导致过学习. 最简单的欠取样方法是随机地去掉某些多数类样本来减小多数类的规模,缺点是容易丢失多数类的一些重要信息.

针对上述重取样技术所存在的缺点,人们提出了许多改进的方法. 改进的过取样方法有人工合成少数类实例过取样算法(synthetic minority over-sampling technique,简记为 SMOTE)(Chawla *et al.*, 2002),该方法可在一定程度上避免随机过取样出现的过学习问题,受到了众多学者的广泛关注,并出现了很多改进方法,如将 SMOTE 方法同标准 boosting 过程相结合的 SMOTEBoost 算法(Chawla *et al.*, 2003)、Borderline-SMOTE 方法(Han *et al.*, 2005)和改进的 SMOTE 方法(杨智明等,2007),这些方法都在一定程度上提高了少数类的分类精度.

改进的欠取样方法有压缩最近邻法(Hart,1968)、邻居清理方法(Laurikkala,2001)、单类选择(Kubat and Matwin, 1997)、托梅克联系对(Tomek,1976)等,这些方法通过一定的规则和方式,找出边界样本和噪音样本,有选择地去掉对分类

作用不大、远离分类边界或者引起数据重叠的多数类样本,只留下安全样本和小类样本作为分类器的训练集.

重取样方法的关键:如何既能消除大量的噪声信息,显著减小数据不均衡程度,又能保证最小的信息损失,以保留绝大多数对分类学习有用的样本点(Estabrooks.,2000).

2　自适应选择近邻的混合重取样算法

从几何的角度看,SMOTE 算法相当于在少数类样本及其被选中的同类近邻连线上进行取样来获得新样本的循环过程,但该算法建立在相距较近的少数类样本之间仍是少数类这样的假设之上,并没有考虑样本数据的真实分布特性,没有考虑少数类样本附近也可能存在多数类样本的分布情况,因此 SMOTE 算法的近邻选择策略存在一定的盲目性.

为了改进 SMOTE 算法在人工合成样本的过程中所存在的盲目性和只能复制生成数值型属性的问题,我们提出一种新型自适应选择近邻的混合重取样算法 ADSNN-Hybrid RS. 该算法分为两部分:过取样部分解决 SMOTE 算法所存在的问题,可根据不同样本集内部的真实分布特性,自己选择调整少数类的近邻,在近邻候选集合中去除距离较远的少数类样本,并允许部分距离少数类样本较近的多数类样本参与到新样本的生成过程中,对于近邻的不同情况采取不同的生成方法,从而达到控制少数类样本近邻区域、提高合成样本质量的目的;欠取样部分对合成之后的数据集用 NCR 方法进行欠取样,去掉多数类中的冗余实例或边界上的噪音数据,在此基础上再进行分类和预测. 该算法结合了过取样和欠取样两种方法的优势,一方面通过自适应选择近邻的方法增加少数类样本的方式强调了少数类,另一方面对多数类进行适当程度的欠取样,减少其规模,达到多数类和少数类样本在一定程度上的相对均衡,从而可以有效地处理非均衡数据分类问题,提高分类器的性能.

为了便于论述,首先给出如下定义:

假设整个不均衡数据训练集 T,数据集中少数类实例为 P,$P=\{P_1,P_2,\cdots,P_{pnum}\}$,多数类实例为 N,$N=\{N_1,N_2,\cdots,N_{mum}\}$,$pnum$ 和 mum 分别为少数类和多数类实例的个数,每个样本具有 m 个属性,则少数类样本 P_i 的第 j 个属性值为 P_{ij},多数类样本 N_i 的第 j 个属性值为 N_{ij},$j=1,2,\cdots,m$. 样本 P_i 生成新合成实例的 5－近邻候选集合为 $P_i-CAND=\{P_i-cand_k\mid k=1,2,\cdots,5,P_i-cand_k\in$ 训练集 $T\}$;只有候选集合中的样本才能参与新样本的生成. $dpp(i,k)$ 表示少数类样本 P_i 与其少数类近邻之间的距离,$dpn(i,k')$ 表示少数类样本 P_i 与其多数类近邻之间的距离.

ADSNN-Hybrid RS 算法的具体实现过程为:

(1)由于 SMOTE 算法的近邻数设为 5,这里我们同样也设近邻数为 5. 对于少数类 P 中的每一个实例 P_i,在 T 中计算它的 5 个近邻. P_i 的 5 个近邻中属于少数类中的个数记为 $K(0\leqslant K\leqslant 5)$,属于多数类中的个数记为 $K'(0\leqslant K'\leqslant 5)$,$K+K'=5$.

(2)如果 $K'=0$,说明 P_i 的 5 个近邻全部是少数类,属于安全实例,则将其加入 P_i-CAND 集合;如果 $0<K'<K<5$,说明 P_i 的 5 个近邻中少数类实例的个数多于多数类实例的个数,我们认为 P_i 是相对比较安全的实例,我们将 P_i 的 5 个近邻都加入 P_i-CAND 集合,但注明其分别所属的类别如(P_i－少数类、P_i－多数类);如果 $0<K<K'<5$,说明 P_i 的 5 个近邻中多数类实例个数多于少数类实例的个数,我们认为 P_i 是相对比较危险的实例,只有当 $dpp(i,k)$ 全部小于多数类的 $dpn(i,k')$ 时,我们才将 P_i 的五个近邻全部加入 P_i-CAND 集合,否则只将 P_i 的少数类近邻加入 P_i-CAND 集合;如果 $K'=5$,P_i 的 5 个近邻全部是多数类,则可以认为 P_i 是噪声数据,不参加人工生成新的合成样本数据.

(3)在确定全部少数类样本 P_i 的近邻候选集合后,下面将生成新的少数类合成样本. 对于数值型属性分两种情况:若 P_i 的候选近邻为少数类样本,则随机数与 SMOTE 算法相同取 rand[0,1];若 P_i 的候选近邻为多数类样本,随机数取 rand[0,0.5],其他计算过程同 SMOTE,这样做可以控制新产生的合成样本尽可能地靠近少数类实例,更好地避免样本混叠现象的发生. 重复上面的步骤,直到产生足够的少数类样本.

(4)对加入人工合成少数类实例后的数据集进行欠取样处理,这里借鉴 NCR 的思想,为了避免对多数类实例进行欠取样时去掉过多包含有用信息的数据,所以在其基础上稍做改进,适当减少清理的程度. 对于数据集的每一个实例 T_i,找到它的 2 个近邻:如果实例 T_i 属于多数类,当分类后,它的 2 个近

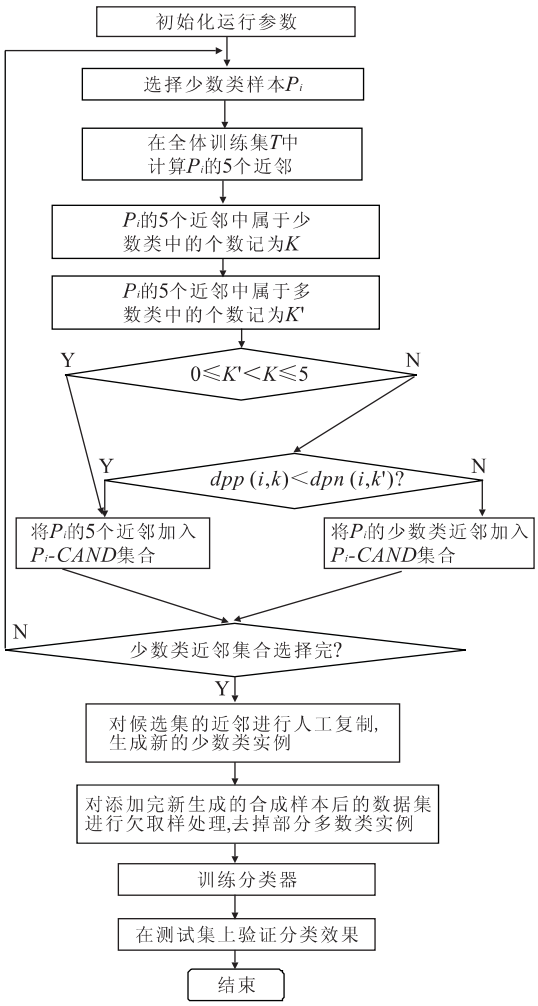

图 1 ADSNN-Hybrid RS 算法处理流程图

Fig. 1 ADSNN-Hybrid RS algorithm flow chart

邻与 T_i 最初类别相反，属于少数类时，则去掉实例 T_i；如果实例 T_i 属于少数类，并且它的 2 个近邻属于多数类，则去掉 T_i 的 2 个多数类近邻.

ADSNN-Hybrid RS 算法流程如图 1 所示.

在计算实例与其近邻的距离时，如果实例属性为数值形式定量类型，使用常用的欧氏距离来衡量；如果实例属性为名词形式定性类型，使用 VDM (value difference metric)(Stanfill and Waltz，1986) 进行衡量. VDM 是 1986 年由 Stanfill 等提出的一个适合于名词型属性的距离计算函数. 对于一个名词型属型 a 的两个属性值分别为 x 和 y 的 VDM 计算方法见式(1)：

$$VDM_a(x,y) = \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^q, \quad (1)$$

式中，$N_{a,x}$ 为训练集 T 中 a 属性值为 x 的样本数目；$N_{a,x,c}$ 为训练集 T 中输出类别为 c 且 a 属性值为 x 的样本数目；q 为常数，一般为 1 或 2.

3 仿真实验

现有的岩爆危险性预测方法是将所有的分类训练数据假设为均衡的前提下进行的分类和预测，但岩爆现象的发生往往属于不均衡情况，为了使分类的结果更有效，我们将 ADSNN-Hybrid RS 算法用于岩爆的危险性预测. 选取南非科学研究院采矿所建立的 VCR 采场岩爆实例数据(冯夏庭，2000)进行实验. VCR 碳化矿床采场工作面的岩爆风险估计模型中，考虑了下列主要影响因素：埋深、地质结构面的倾角、地质结构面的类型、采矿方法(长壁式、破裂式)、临时支护、永久支护和区域支护的类型、效果，采场宽度、走向跨度，岩爆发生的位置、大小和发生后的处理措施等.

VCR 采场岩爆实例数据集通过对岩爆影响因素的分析以及深部开采中岩爆实例的收集建立的岩爆实例数据库，将岩爆的主要影响因素作为输入矢量，岩爆发生与否作为输出标量. 在这里对预分类的数据进行重取样训练，然后再进行预测. 由于我们只考虑岩爆数据重取样前后的分类情况，所以用最简单的决策树算法 J48(Quinlan，1993)来验证混合重取样前后岩爆数据的分类结果. 我们先对少数类实例进行自适应选择近邻过取样，然后再进行欠取样处理.

实验一：从数据集中随机选取 99 个样本(如 VCR 采场岩爆实例数据集的前 99 个实例)用于训练，用所获得的模型对剩下的 5 个进行测试，如果对训练数据不做预处理，用最基本的 J48 做实验仅仅预测正确 2 个，预测错误 3 个.

当对 99 条原始数据进行 ADSNN-Hybrid RS 重取样后，再对测试数据重新进行预测，预测完全正确分类的结果如表 1 所示. VCR 采场岩爆风险预测值的详细分类结果见表 2，第 100、101 这 2 个实例未发生岩爆，第102、103、104这3个实例确实发生

表 1 分类结果

Table 1 Classification results

=== Detailed Accuracy By Class ===					
TP rate	FP rate	Precision	Recall	F-measure	Class
1	0	1	1	1	发生岩爆
1	0	1	1	1	不发生岩爆
=== Confusion matrix ===					
a b <——classified as					
3 0 | a = 发生岩爆					
0 2 | b = 不发生岩爆					

表 2　VCR 采场岩爆预测结果

Table 2　Rockburst prediction results at VCR mining stope

样本编号	特征矢量输入	预测输出	实际情况
100	10010100010100100000000010010001	01	不发生岩爆
101	10010100100100100000000010001010	01	不发生岩爆
102	01001010100100100010000000010001	10	发生岩爆
103	10010100100100001000010000010100	10	发生岩爆
104	01010001100100100000000001010010	10	发生岩爆

了岩爆. 可见分类结果的预测值与实际情况完全一致.

所得到修剪后的分类结果决策树如图 2 所示.

实验二：为了探讨如何有效地控制岩爆的发生，可对岩爆发生的主控因素进行试验研究工作. 我们以表中的实例 100 的特征作线索，实例 100 的原始影响因素值为{1,0,0,1,0,1,0,0,0,1,0,1,0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,1,0,0,0,1}，结果是不发生岩爆，我们将永久支护由现在的“永久支护＝木垛”改为其他支护方式，或将走向跨度由现在的“走向跨度＝100～200 m”改为“走向跨度＞200 m”，则情况发生了变化，岩爆危险状态变为发生岩爆. 由此可以看出，当施工条件与实例 100 的情况相似时，是否有永久支护、走向跨度的多少，对岩爆的发生有很大的影响，不利的支护则使岩爆的风险增加. 因此，可以建议工程建设方通过加强永久支护、或者增大工程的开挖宽度等来减小应集力集中的程度，以减小岩爆发生的风险. 同理，地质结构和采深对岩爆发生的可能性也有很大的影响. 按照上述方法在数据相对均衡的情况下，岩爆训练和测试样本的分类精度能够得到很大的提升，结果与实际比较吻合，该方法具有较好的应用前景.

4　结论

工程实例应用结果表明，岩爆预测结果与实际情况一致，说明本文提出的 ADSNN-Hybrid RS 算法在岩爆的实例数据不均衡的情况下，人工生成部分少数类数据，作为训练数据的取样方法用于岩爆的危险性预测是切实可行的，预测准确率高，具有良好的工程应用前景. 本文方法不必建立复杂的数学方程或计算模型，输入数据是客观存在或易于测量的，具有实现简单的优点. 采用该方法可对其他发生岩爆的实例寻找出相应的发生岩爆的主控因素，为深部开采工程的合理设计与安全施工提供科学依据.

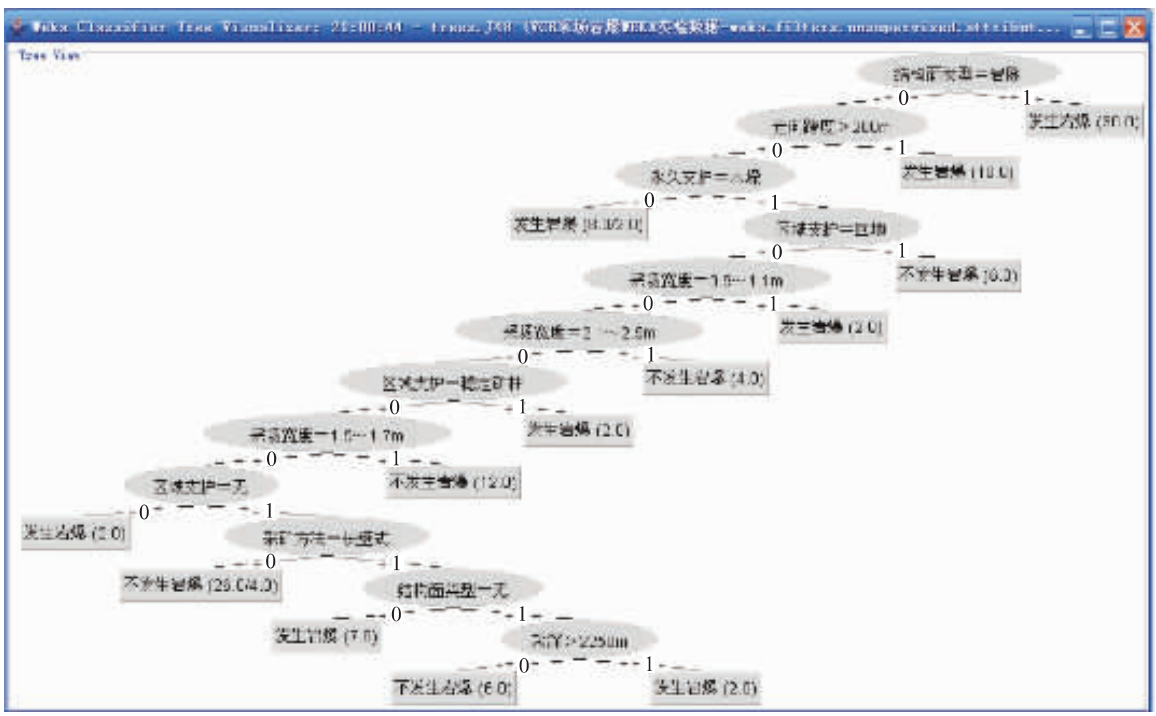

图 2　VCR 采场岩爆实例数据生成的修剪过的决策树

Fig. 2　Pruned decision tree on rockburst instances at VCR mining stope

References

Chawla, N. V., Bowyer, K. W., Hall, L. O., et al., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(3): 321—357.

Chawla, N. V., Lazarevic, A., Hall, L. O., et al., 2003. SMOTEboost: improving prediction of the minority class in boosting. *Lecture Notes in Computer Science*, 2838: 107—119. doi. 10. 1007/b13634

Chen, H. J., Li, N. H., Nie, D. X., et al., 2002. A model for prediction of rockburst by artificial neural network. *Chinese Journal of Geotechnical Engineering*, 24(2): 229—232 (in Chinese with English abstract).

Estabrooks, A., 2000. A combination scheme for inductive learning from imbalanced data sets. Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada.

Feng, X. T., 2000. Introduction to intelligent rock mechanics. Science Press, Beijing (in Chinese).

Ge, Q. F., Feng, X. T., 2008. Classification and prediction of rockburst using AdaBoost combination learning method. *Rock and Soil Mechanics*, 29(4): 943—948 (in Chinese with English abstract).

Han, H., Wang, W. Y., Mao, B. H., 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. *Lecture Notes in Computer Science*, 3644(1): 878—887. doi. 10. 1007/11538059_91

Hart, P., 1968. The condensed nearest neighbor rule(Corresp.). *IEEE Transactions on Information Theory*, 14(3): 515—516.

Jiang, T., Huang, Z. Q., Zhao, Y. Y., et al., 2003. Application of grey system optimal theory model in forecasting rockburst. *Journal of North China Institute of Water Conservancy and Hydroelectric Power*, 24(2): 37—40 (in Chinese with English abstract).

Kubat, M., Matwin, S., 1997. Addressing the curse of imbalanced training sets: one-sided selection. Proceedings of the Fourteenth International Conference on Machine Learning. Morgan Kaufmann Publishers, Inc., 179—186.

Laurikkala, J., 2001. Improving identification of difficult small classes by balancing class distribution. *Lecture Notes in Computer Science*, 2101: 63—66. doi. 10. 1007/3—540—48229—6_9

Quinlan, J. R., 1993. C4. 5: programs for machine learning. Morgan Kaufmann. doi. 10. 1007/BF00993309

Stanfill, C., Waltz, D., 1986. Toward memory-based reasoning. *Communications of the ACM*, 29(12): 1213—1228.

Tomek, I., 1976. Two modifications of CNN. *IEEE Transactions on Systems, Man and Cybernetics*, 6(6): 769—772.

Yang, Y. C., Zhu, J., 2001. An matter-elements model and its application to classified prediction of rockburst. *Systems Engineering—Theory & Practice*, 21(8): 125—129 (in Chinese with English abstract).

Yang, Z. M., Qiao, L. Y., Peng, X. Y., 2007. Research on detamining method for imbalanced dataset based on improved SMOTE. *Acta Electronica Sinica*, 35(12A): 22—26 (in Chinese with English abstract).

Zhao, H. B., 2005. Classification of rockburst using support vector machine. *Rock and Soil Mechanics*, 26(4): 642—644(in Chinese with English abstract).

附中文参考文献

陈海军，郦能惠，聂德新，等，2002. **岩爆预测的人工神经网络模型. 岩土工程学报**，24(2)：229—232.

冯夏庭，2000. **智能岩石力学导论. 北京：科学出版社.**

葛启发，冯夏庭，2008. **基于** AdaBoost **组合学习方法的岩爆分类预测研究. 岩土力学**，29(4)：943—948.

姜彤，黄志全，赵彦彦，等，2003. **灰色系统最优归类模型在岩爆预测中的应用. 华北水利水电学院学报**，24(2)：37—40.

杨莹春，诸静，2001. **物元模型及其在岩爆分级预报中的应用. 系统工程理论与实践**，21(8)：125—129.

杨智明，乔立岩，彭喜元，2007. **基于改进** SMOTE **的不平衡数据挖掘方法研究. 电子学报**，35(12A)：22—26.

赵洪波，2005. **岩爆分类的支持向量机方法. 岩土力学**，26(4)：642—644.