doi: 10.12068/j.issn.1005-3026.2019.09.021

基于机器学习的钻孔数据隐式三维地质建模方法

郭甲腾, 刘寅贺, 韩英夫, 王徐磊
(东北大学 资源与土木工程学院, 辽宁 沈阳 110819)

摘 要: 针对基于钻孔数据的传统显式三维地质建模方法存在过程繁琐、模型质量难以保证等缺点,本文提出了一种基于机器学习的隐式三维地质建模方法,将地层三维建模问题转换为地下空间栅格单元的属性分类问题. 分别基于支持向量机、BP神经网络等分类算法,实现了钻孔数据的自动三维地质建模. 实际建模结果表明,对于有限、稀疏的钻孔数据,支持向量机方法建模准确率较高,建模效率、效果优于显式建模方法. 最后通过敏感性分析研究了超参数对建模结果准确率、模型形态的影响,为可控的自动三维地质建模提供了一种新的解决思路.

关 键 词: 机器学习;支持向量机;三维地质建模;隐式建模;钻孔数据
中图分类号: TD 67 **文献标志码**: A **文章编号**: 1005-3026(2019)09-1337-06

Implicit 3D Geological Modeling Method for Borehole Data Based on Machine Learning

GUO Jia-teng, LIU Yin-he, HAN Ying-fu, WANG Xu-lei
(School of Resources & Civil Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: GUO Jia-teng, E-mail: guojiateng@mail.neu.edu.cn)

Abstract: Considering the complex modeling process and difficulty in guaranteeing the model quality of traditional explicit 3D modeling methods, an implicit 3D geological modeling method for borehole data based on machine learning was proposed, which transformed the strata 3D modeling problem into a process of geological attribute classification of the underground spatial grid units. Based on the classification algorithms of support vector machine and BP neural network, automatic 3D geological modeling from borehole data was realized. The results demonstrate that for sparse and limited borehole data, support vector machine can generally perform better than explicit methods. Finally, the influence of hyper-parameter on modeling accuracy and model shape is studied through sensitivity analysis, which provides a new solution for controllable 3D geological modeling.

Key words: machine learning; support vector machine; 3D geological modeling; implicit modeling; borehole data

由于地下空间的不可见性及地质构造发育情况的复杂性,地下地质构造的探知难度较大、不确定性高. 目前有效的地下勘探数据来源有钻探、声波测试、重力勘探、电磁波探测等方法,钻探法作为获得地质空间信息最为直观、可靠的手段,其获得的钻孔数据所记录的信息准确率非常高,是用来构建三维地质模型的重要依据. 针对基于钻孔数据构建三维地质模型的问题,朱良峰等[1]将钻孔资料与专家经验相结合,实现了快速三维地质建模;Wu[2]提出了基于广义三棱柱(GTP)进行三维地质模型的构建;在此基础上车德福等[3]改进了基于GTP的断层三维交互建模方法;郭甲腾等[4]实现了地上下三维集成建模.

上述方法属于传统的显式建模方法,通常需要较多的人工交互,建模过程繁琐、效率较低,且容易出现主观错误;或者建立的模型棱角尖锐、光

收稿日期: 2018-09-19
基金项目: 国家自然科学基金资助项目(41671404); 国家级大学生创新创业训练计划资助项目(201810145060); 中央高校基本科研业务费专项资金资助项目(N170104019); 中国地质调查局智能地质调查支撑平台建设项目(DD20160355).
作者简介: 郭甲腾(1980-),男,安徽桐城人,东北大学副教授.

滑度不高. 与之相对应的建模方式称为隐式建模[5-7],是指采用空间插值方法,基于采样数据拟合空间曲面函数,进而生成三维可视化模型,因此隐式三维地质建模的核心问题在于选择适合于地质构造特征的空间插值函数. 常用的空间插值方法包括距离反比插值、离散光滑插值以及克里金插值等[8],而这些传统的插值手段往往要求大量的采样数据[9]. 由于钻孔数据成本较高,在研究区域内往往只能获得有限、稀疏的钻孔资料,为解决这一问题,必须提出一种基于稀疏地质数据的有效隐式三维地质建模方法.

近些年迅速发展的机器学习方法在计算机视觉、自然语言处理等领域已取得了突破性进展,在某些方面其效果甚至已经超过了人类的表现[10]. 如何将机器学习的方法引入到地质问题中来已成为近期地学建模领域的研究热点[11-14].

本文提出了一种基于机器学习的钻孔数据隐式三维地质建模方法:如图 1 所示,与传统显示建模方法不同,将地质建模问题转换为地下空间栅格单元的地层属性分类问题. 以建模区域中每一点的三维坐标作为分类特征,将部分钻孔数据作为训练集,训练出分类器作为中间产物. 再由训练好的分类器对建模区域进行分类,从而获得该区域的三维栅格地质模型,实现基于机器学习的隐式三维地质建模.

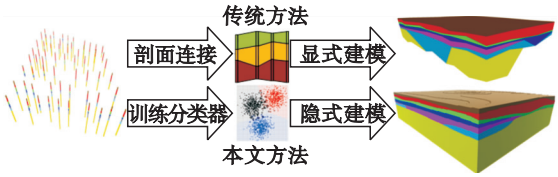

图 1 本文建模方法示意图

Fig. 1 Schematic diagram of proposed modeling method

1 基于机器学习的隐式三维地质建模

基于机器学习的钻孔数据三维地质隐式建模方法的关键问题在于如何利用钻孔数据训练得到分类器. 由于钻孔数据的原始记录格式难以直接应用到机器学习方法中,在训练之前须对钻孔数据进行预处理. 同时还要针对不同的地学问题选择适合训练的分类器,最后讨论相应的超参数选取方法.

1.1 钻孔数据的预处理

原始钻孔数据包含勘探点的平面坐标、高程、各地层分界点的深度和地层类别. 首先,由于计算机不能认识到地层上下界限点之间为同一属性地层这一地学意义,这使得特征空间十分稀疏,难以获得理想的训练效果. 因此需要进行重采样,使得钻孔数据成为一系列具有空间位置和地层(编号为 A,B,C)属性的点. 其次,为使分类器分辨出“地表”这一概念,需要在地表之上虚拟出“空气”训练集(T),如图 2 所示. 最后,为了消除不同坐标量级之间的影响,需要对数据进行标准化处理,其转化函数为

$$X^* = (X - \mu)/\sigma \ . \tag{1}$$

式中:X^* 为标准化后的值;X 为待标准化的值;μ 为样本数据的均值;σ 为样本数据的标准差.

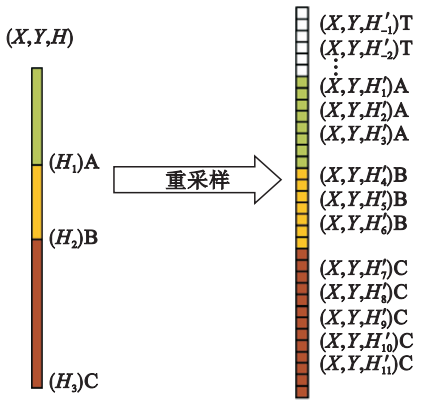

图 2 钻孔数据重采样示意图

Fig. 2 Schematic diagram of borehole data re-sampling

1.2 分类器的选择

本文建模方法实质上是将三维地质建模问题转换为地下空间栅格单元的地层属性分类问题,因此不同的分类器会在根本上决定建模的结果. 贝叶斯分类、决策树方法、支持向量机(support vector machine, SVM)和神经网络等是常见的机器学习算法. 近年来,神经网络特别是深度学习算法在相关领域取得了显著的应用成果,但对于稀疏、非均衡的地下空间数据,支持向量机往往有更好的表现[15]. 本文主要利用这两种算法,研究探讨机器学习方法在钻孔数据隐式三维地质建模中的应用.

1.3 超参数的确定

所谓超参数是指在开始学习过程之前设置的参数. 根据所选的分类器不同,需要确定的超参数也不同. 一些超参数只会影响到模型训练的效率,而某些参数则会直接影响模型分类效果的准确性,并最终影响三维模型的构建效果.

由于地质构造发育复杂程度高,钻孔数据的空间分布情况差异性较大,导致一些超参数很难有效地人工给定. 因此,需要针对建模区域的钻孔

数据特征，设计合适的算法，自动地寻找最优的超参数.

本文采用的 Scikit – learn[16] 机器学习包提供了网格搜索和随机搜索两种参数调优的模式，考虑到采用网格搜索即穷举法寻找最优参数需要耗费大量时间且需要人工干预，本文采用对参数空间随机采样的方式优化参数. 具体方法为：从参数的随机分布中抽样进行交叉验证，每次迭代进行 200 次充分抽样，每轮迭代选出最优的 5 个候选参数作为初值进行下一轮迭代，直到交叉验证结果不再显著提高. 而随机分布函数的选择，则需要在网格法分析基础上根据经验构建.

由于地质模型复杂、训练数据量大，每次测试超参数都需要耗费大量时间. 为了提高超参数的选取效率，需要先对钻孔数据进行稀疏采样，在该数据集上寻找最优参数，之后再精细采样，使用搜索得到的超参数完成三维地质建模.

1.4　基于机器学习的钻孔数据三维地质建模

综合上述问题，本文提出的基于机器学习的钻孔数据隐式三维地质建模流程如图 3 所示. 关键步骤包括：1）对钻孔数据进行预处理，根据不同的采样间隔获得稀疏数据集和精细数据集；2）利用稀疏数据集交叉验证寻找最适合当前问题的最优超参数组合；3）根据最优超参数，在精细采样的数据集上训练分类器；4）由训练好的分类器对建模区域进行预测分类，最终获得三维地质模型.

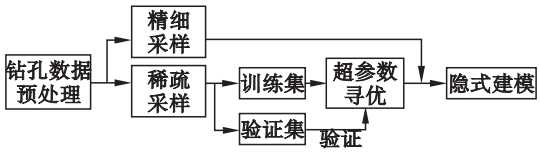

图 3　基于机器学习的钻孔数据隐式三维地质建模流程

Fig. 3　Process of implicit 3D geological modeling for borehole data based on machine learning

2　方法验证与应用

在上述算法基础上，基于 Scikit – learn 和 Keras 开源机器学习库以及 Unity 3D 引擎，采用沈阳某区域的岩土工程勘探钻孔数据开展了隐式三维地质建模实验.

为了验证建模的准确率，随机选择 6 个钻孔的数据（占全部钻孔的 10%）作为验证集进行预处理，不参与建模. 用剩余的钻孔数据进行建模后，检查验证集的建模总体精度（overall accuracy，OA）和 Kappa 值，用来评判建模效果.

2.1　常见分类器对比

本文选择支持向量机作为分类器，同时与几种经典分类算法进行了对比. 采用 Scikit – learn 机器学习库的默认参数训练后，不同分类器在该数据集上的表现如表 1 所示. 可以看出：支持向量机在该问题上的分类效果最好. 由于神经网络的参数难以确定，因此不参与该项对比.

为了避免高维空间中的“维数灾难”问题而引入的核函数[17]，是支持向量机算法中对分类效果影响最为显著的参数. 常用的核函数有线性核函数、多项式核函数、径向基核函数（radial basis function，RBF）以及 Sigmoid 核函数. 其中径向基核函数能够将原始特征映射到无穷维的特征空间，可以有效处理非线性问题，符合钻孔数据三维地质建模的高度非线性特征. 因此，本文选用 RBF 核函数作为支持向量机的优选核函数. 不同核函数使用默认参数在该训练集上的分类结果（表 2）也验证了该结论，最终建模结果如图 4 所示.

表 1　不同分类器的分类结果

Table 1　Classification results of different classifiers

分类器	准确率	Kappa 值	分类器参数
SVM	0.864 2	0.811 2	C = 1.0，kernel = ′rbf′，γ = 0.333 3
K 临近	0.860 8	0.807 0	leaf _ size = 30，n _ neighbors = 5
随机森林	0.845 9	0.785 3	n_estimators = 10，criterion = ′gini′，min_samples_leaf = 1，min_samples_split = 2，决策树深度、叶节点数无上限
决策树	0.828 2	0.761 1	criterion = ′gini′，min _ samples _ leaf = 1，min _ samples_split = 2，决策树深度、叶节点数无上限
逻辑回归	0.730 5	0.592 0	C = 1.0，solver = ′liblinear′，penalty = ′L2′，intercept_scaling = 1

表 2　不同核函数分类结果

Table 2　Classification results of different kernel functions

核函数	准确率	Kappa 值	分类器参数
RBF	0.864 2	0.811 2	C = 1.0，γ = 0.333 3
线性	0.837 0	0.772 6	C = 1.0
多项式	0.826 9	0.757 0	C = 1.0，degree = 3，γ = 0.333 3，coef0 = 0.0
Sigmoid	0.602 9	0.464 6	C = 1.0，γ = 0.333 3，coef0 = 0.0

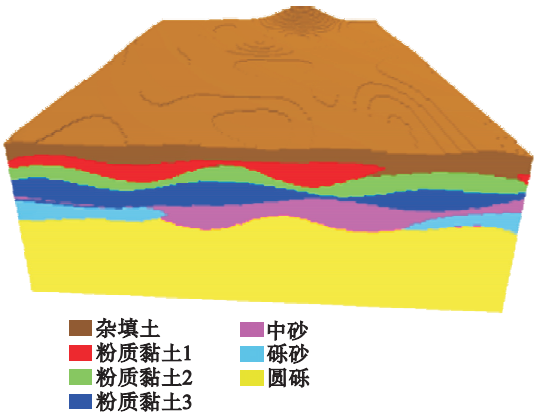

图 4 RBF 核函数的支持向量机建模结果
Fig. 4 Modeling results based on SVM with RBF kernel function

2.2 支持向量机与 BP 神经网络建模结果对比

本文将经过交叉验证获得最优参数的支持向量机(RBF 核函数,$C=16$,$\gamma=1.1$)与一个经过充分训练的 BP 神经网络(输入 *XYZ* 坐标,输出地层类别,中间包括 5 层隐含层,每层 20 个节点,共 2 348 个参数)进行了建模准确率对比,结果如表 3 所示. 可以得出:利用神经网络与支持向量机进行分类获得的地质建模准确率相差不大,都可以获得比较理想的分类结果.

表 3 支持向量机及神经网络建模结果准确率对比
Table 3 Modeling results accuracy comparison of SVM and neural network

钻孔编号	支持向量机准确率	神经网络准确率
2	0.888 5	0.892 7
20	0.902 3	0.910 0
26	0.863 4	0.914 3
41	0.867 1	0.891 3
53	0.881 6	0.864 8
60	0.923 8	0.917 3
总体	0.899 5	0.897 6

通过对比两种方法生成的三维地质模型细节(图 5)可以看出:支持向量机生成的模型更加光滑自然,而神经网络生成的模型一些细节则比较尖锐、突兀,在整体效果上劣于支持向量机. 因此,对于小样本、三维坐标作为分类特征的分类模型,支持向量机更加适合. 同时,由于神经网络的训练耗时远远大于支持向量机,且参数繁多难以确定,综合考虑应选择支持向量机作为钻孔数据三维地质建模的优选分类器.

3 超参数敏感性分析

在方法验证过程中发现,对于不同的建模数据,超参数难以确定唯一值. 而不同的超参数组合会对建模结果产生极大影响. 本节从建模准确率和最终模型形态两个方面,对惩罚因子 C 和 RBF 核函数参数 γ 两个超参数的组合进行了敏感性分析.

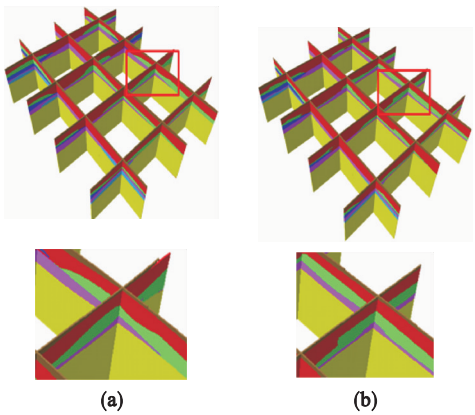

图 5 模型栅状图
Fig. 5 Fence diagram of models
(a)—支持向量机;(b)—BP 神经网络.

3.1 超参数对建模准确率的影响

分类的准确率是评估当前参数组合优劣的标准,可以认为更高的准确率代表当前参数组合更适合当前区域的建模问题. 如图 6 所示,C 值在准确率达到峰值后趋于平稳,没有显著下降,因此可在搜索 C 值时适当调大并采用更大的搜索步长. 随着 γ 值的增大,建模的准确率迅速增加到峰值,下降后趋于长时平稳后又急速下降,表明 γ 值对建模结果的准确率影响较大. 因而需要谨慎确定 γ 值,在参数空间中应以较小的初值和步长进行 γ 值搜索.

3.2 超参数对三维模型形态的影响

不同超参数对建模方法最直观的影响就是对模型形态的影响,如表 4 所示. 当 γ 值过小时,核函数的幅宽过大,导致平滑效应过大,使本不应该相连的尖灭被强行平滑到一个地层;当 γ 值过大时,核函数的幅宽又过小,使钻孔数据难以控制周围的区域,最终占多数的底部地层“渗透”到没有钻孔数据控制的区域,导致在顶部出现了底部地层的问题,模型出现了严重的建模错误. 同时,当 C 过大时,也会因为过拟合使地层出现褶皱.

可以看出,不同的超参数组合,尤其是 γ 的取值,会对模型的形态产生极大影响. 由于 γ 影响核函数的幅宽,而钻孔数据在水平方向上分布比较稀疏,因此对模型的水平方向形态影响较大;而 C 值影响分类器对分类错误的容忍程度,主要控制地层的厚度与起伏,即对模型的竖直方向形态影

响较大.

本文方法中,超参数通过在参数空间中自动搜索得到,仅能够保证在验证集上的准确率达到最优. 因此可通过人为调节 C 值和 γ 值进一步控制模型的形态. 但实验结果表明,采用本文方法搜索得到的参数结果已能够获得良好的建模结果.

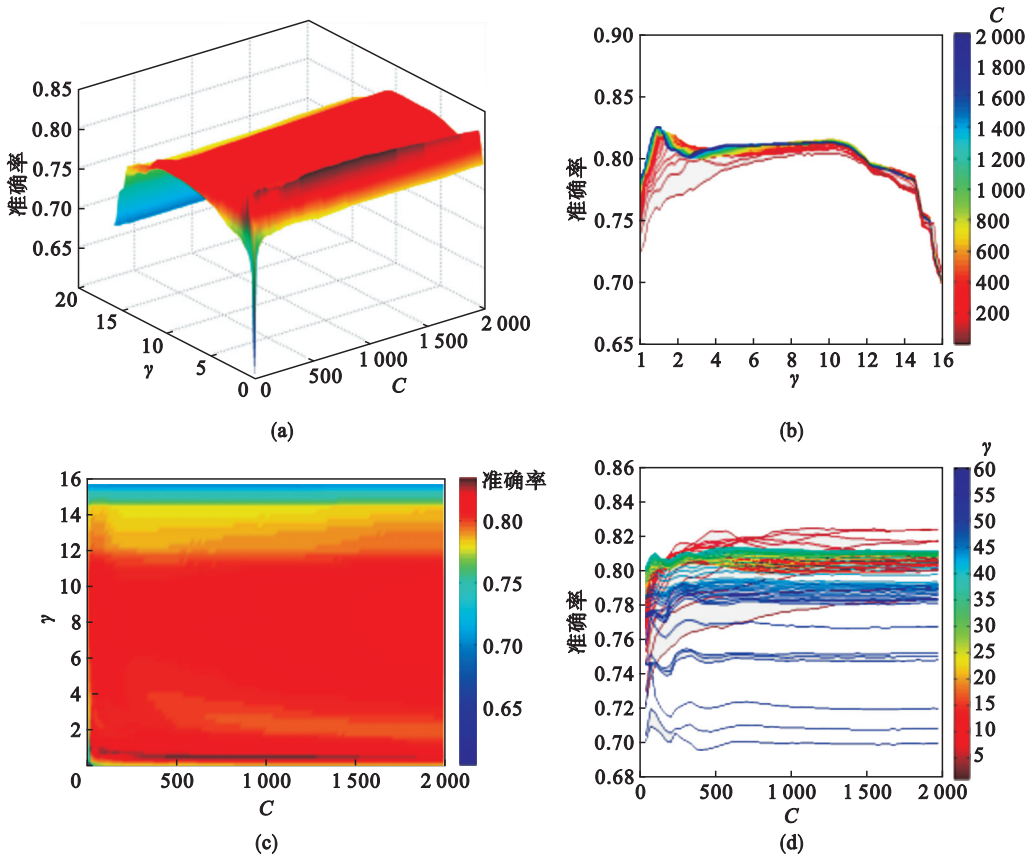

图 6　超参数敏感性分析

Fig. 6　Hyper-parametric sensitivity analysis

(a)—C,γ,准确率视图;(b)—γ,准确率视图;(c)—C,γ 视图;(d)—C,准确率视图.

表 4　不同超参数下三维地质模型对比

Table 4　Comparison of 3D geological model under different hyper-parameters

参数	$\gamma=0.062\,5$	$\gamma=1$	$\gamma=32$
$C=256$			
$C=16$			
$C=0.125$			

4　结语与展望

本文针对钻孔数据的三维地质建模问题,提出了一种基于机器学习的自动隐式建模方法,并成功应用于城市区域岩土工程勘探钻孔的三维地质建模. 通过实例分析与验证认为,针对有限、稀疏的钻孔数据,采用基于 RBF 核函数支持向量机作为分类器较为合适,能够保证较高的分类精度,同时获得的三维模型也较为光滑、合理,具有实际的工程应用价值.

该建模方法在以下方面仍需研究和改进:

1) 需进一步研究超参数在建模中的地学意义,探索最优超参数的经验值确定方法;

2) 支持向量机算法在基于钻孔数据的沉积地层建模方面表现了其优越性,但对于断层、褶皱等复杂地质构造的建模适用性方面需要进一步研究.

参考文献:

[1] 朱良峰,吴信才,刘修国,等. 基于钻孔数据的三维地层模型的构建[J]. 地理与地理信息科学,2004,20(3):26－30. (Zhu Liang-feng, Wu Xin-cai, Liu Xiu-guo, et al. Reconstruction of 3D strata model based on borehole data [J]. *Geography and Geo-information Science*,2004,20(3):

26 – 30.)

[2] Wu L X. Topological relations embodied in a generalized triprism(GTP) model for a 3D geoscience modeling system [J]. *Computers & Geosciences*, 2004, 30(4): 405 – 418.

[3] 车德福,吴立新,殷作如,等. 基于 GTP 的断层三维交互建模方法[J]. 东北大学学报(自然科学版),2008,29(3): 395 – 398.
(Che De-fu, Wu Li-xin, Yin Zuo-ru, et al. On the GTP-based 3D interactive modeling method for geological faults [J]. *Journal of Northeastern University(Natural Science)*, 2008, 29(3): 395 – 398.)

[4] 郭甲腾,吴立新,杨宜舟,等. 岩土工程勘察场地立体空间建模与可视化信息管理[J]. 东北大学学报(自然科学版),2014,35(1): 122 – 125.
(Guo Jia-teng, Wu Li-xin, Yang Yi-zhou, et al. Three-dimensional modeling and visualization information management of geotechnical engineering investigation sites [J]. *Journal of Northeastern University(Natural Science)*, 2014, 35(1): 122 – 125.)

[5] Calcagno P, Chilès J P, Courrioux G, et al. Geological modelling from field data and geological knowledge: part I. modelling method coupling 3D potential-field interpolation and geological rules [J]. *Physics of the Earth & Planetary Interiors*, 2011, 171(1): 147 – 157.

[6] Caumon G, Gray G, Antoine C, et al. Three-dimensional implicit stratigraphic model building from remote sensing data on tetrahedral meshes: theory and application to a regional model of La Popa Basin, NE Mexico[J]. *IEEE Transactions on Geoscience & Remote Sensing*, 2013, 51(3): 1613 – 1621.

[7] Hillier M J, Schetselaar E M, Kemp E A D, et al. Three-dimensional modelling of geological surfaces using generalized interpolation with radial basis functions [J]. *Mathematical Geosciences*, 2014, 46(8): 931 – 953.

[8] 邢延涛,李利军. 三维地质体重构中空间数据插值方法的研究[J]. 计算机与数字工程,2006,34(12): 45 – 47.
(Xing Yan-tao, Li Li-jun. Research of spatial interpolation methods about the 3D-reconstruction of geological models [J]. *Computer & Digital Engineering*, 2006, 34(12): 45 – 47.)

[9] Smirnoff A, Boisvert E, Paradis S J. Support vector machine for 3D modelling from sparse geological information of various origins[J]. *Computers & Geosciences*, 2008, 34(2): 127 – 143.

[10] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: surpassing human-level performance on imagenet classification [C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, 2015: 1026 – 1034.

[11] ítalo G G, Kumaira S, Guadagnin F. A machine learning approach to the potential-field method for implicit modeling of geological structures[J]. *Computers & Geosciences*, 2017, 103: 173 – 182.

[12] Abedi M, Norouzi G H, Bahroudi A. Support vector machine for multi-classification of mineral prospectivity areas [J]. *Computers & Geosciences*, 2012, 46: 272 – 283.

[13] Edwards J, Lallier F, Caumon G, et al. Uncertainty management in stratigraphic well correlation and stratigraphic architectures: a training-based method [J]. *Computers & Geosciences*, 2017, 111(2): 1 – 17.

[14] Cracknell M J, Reading A M. Geological mapping using remote sensing data: a comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information [J]. *Computers & Geosciences*, 2014, 63(1): 22 – 33.

[15] 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.
(Gu Qiong. Research of machine learning on imbalanced data sets and its application in geoscienes data processing [D]. Wuhan: China University of Geosciences, 2009.)

[16] Pedregosa F, Gramfort A, Michel V, et al. Scikit-learn: machine learning in Python[J]. *Journal of Machine Learning Research*, 2011, 12(10): 2825 – 2830.

[17] Chaudhuri A, De K, Chatterjee D. A comparative study of kernels for the multi-class support vector machine [C]// International Conference on Natural Computation. Jinan, 2008: 3 – 7.