DOI: 10.3785/j.issn.1008-9497.2011.03.023

矿产预测中的成矿因子选择方法:以滇东南金矿预测为例

俞 乐[1,2] 柏 坚[3,4], 张汉奎[1]

(1.浙江大学 地球科学系, 浙江 杭州 310027; 2.清华大学 地球系统科学研究中心, 北京 100084;
3.中国地质大学 地球科学与资源学院, 北京 100083; 4.云南省地矿局, 云南 昆明 650011)

摘 要:由于矿产地质信息的复杂性和不确定性,难以建立精确的数学模型来确定矿产资源的分布状况.非线性分析建模技术,如人工神经网络(Artificial Neural Network, ANN)、支持向量机(Support Vector Machine, SVM)等,给矿产预测工作提供了新的途径.这类方法在处理数据时可以避免数据分析和建模的困难,即不须理解各种成矿因子与矿床(点)之间的相互关系,只须选择已知的矿床(点)和非矿产(点),进行"黑箱"学习.虽然经过合理的训练,这类方法能够得到较高的预测精度,但由于其分类过程的非线性特性,难以获得容易理解的分类规则,提供成矿因子的知识.本文采用基于 SVM 的迭代特征消去(Recursive Feature Elimination, RFE)技术(SVM-RFE),即在 SVM 模型的训练过程中,采用 RFE 特征选择方法,从所有输入的成矿因子中选择出对矿床(点)能正确预测的重要因子,以提供对输入模型的成矿因子的客观评价.通过对滇东南地区金矿预测的实践表明,采用 SVM-RFE 技术从原始 10 类成矿因子中自动选择 6 类进行预测的精度从 68.42% 提高到 94.74%,并且得到该区域进行矿产预测的成矿因子重要性依次是:Au 异常、As 异常、侵入岩、下三叠统与中三叠统之间的平行不整合面、上二叠统与三叠系的平行不整合面、断裂交汇点密度、石炭系和下二叠统间的平行不整合面、中上泥盆统和石炭系间的平行不整合面、Sb 异常和 Hg 异常,从中选取前 6 类成矿因子进行 SVM 训练得到的预测精度最高.这一结论可为在该区域进行矿产预测的资料选取,以及对成矿因子的理解提供支持.

关 键 词:特征选择; 支持向量机; 迭代特征消去; 金矿; 滇东南

中图分类号:P 628 **文献标志码**:A **文章编号**:1008-9497(2011)03-348-06

YU Le[1,2], BAI Jian[3,4], ZHANG Han-kui[1] (1. *Department of Earth Sciences, Zhejiang University, Hangzhou* 310027, *China*; 2. *Center for Earth System Science, Tsinghua University, Beijing* 100084, *China*; 3. *School of Geoscience and Resources, China University of Geoscience, Beijing* 100083, *China*; 4. *Yunnan Bureau of Geology and Mineral Resources, Kunming* 650011, *China*)

Feature selection approach in mineral prospectivity analysis: a case study of gold deposits in Southeastern Yunnan, China. Journal of Zhejiang University(Science Edition), 2011, 38(3): 348−353

Abstract: Because of the complexity and uncertainty of mineral geological information, it is a difficult task to model the distribution of mineral recourses by using parametric model. Non-linear modeling techniques such as artificial neural network (ANN), support vector machine (SVM) etc. provide a promising mean to handle such kind of information, without acquiring exact relationship between mineral geological information and mineral deposits. In this non-linear approach, all mineral deposits and non-deposits are trained/validated/tested in a "black box"-like classifier. Although a high predict accuracy can be reached if the classifier is trained properly, it is still hard to obtain classify rules, which indicate the preferable metallogenic factors from geological information because of the non-linear structural characteristic the classifier. In this paper, a technique called support vector machine based recursive feature elimination, or SVM-RFE is used to rank all input features in SVM. An experiment of SVM-RFE is conducted on gold prospectivity analysis in southeast Yunnan shows this technique could improve the predict accuracy from 68.42% to 94.74% by shrinking 10 input features to 6. The preferable rank of all 10 features calculated by SVM-RFE is Au abnormity, As abnormity, intrusive rock, parallel unconformity between Lower and Middle Triassic system, parallel unconformity between Upper Permian and Triassic system, density of faults intersection, parallel unconformity between Carboniferous and Permian system, parallel unconformity between Devonian and Carboniferous

收稿日期:2010-02-03.

作者简介:俞 乐(1982−),男,博士,主要从事 GIS 空间分析和遥感数字图像处理研究.

system, Sb abnormity, Hg abnormity; and the former 6 features give the best predict accuracy. This rank benefits to selection and understanding metallogenic factors in this study area.

Key Words: feature selection; support vector machine(SVM); Recursive Feature Elimination(RFE); gold deposit; southeast Yunnan

滇东南地区是我国重要的多金属成矿省,前人对该区成矿作用和成矿地质背景进行了很好的研究.从 90 年代初开始,陆续发现并找到具有工业价值的微细粒型金矿床[1],随后该区成为寻找微细粒型金矿床的热点地区,在该区域内陆续发现金矿床(点)近百处[2].由于区域内易于发现的露头矿、浅部矿日趋减少,找矿难度增大,因此需要开展基于数学地质方法的成矿预测[3],即结合矿区资料、区域地质资料、地球化学资料、地球物理资料等对矿产资源分布进行预测.成矿预测数理模型的建立过程一般是在地质模型简化的基础上抽象的物理模型,进而再抽象为数学模型.其核心是对地质过程的简化,因此,模型中特征的选择需以地质特征分析为基础.

滇黔桂地区金矿化具有多时代、多层位、多类型、多成因的特点[4],成矿因子复杂,传统的线性模型难以满足要求[5].ANN 是一类应用广泛的非线性技术,大量研究表明了基于神经网络进行矿产预测的可行性[6-8],但由于受样本数目(ANN 需要大量样本,但在很多开展矿产预测的实践中,矿点和非矿点样本数目有限)以及"维数灾难(Curse of Dimensionality)"(ANN 只能处理低维数据,如果要处理高维数据,则需要先做降维处理)影响,在矿产预测应用中存在限制.SVM 是近年来出现的一种基于统计学习理论的非线性方法,具有结构风险小、泛化能力好、能处理小样本及高维数据等优点[9],适合矿产预测.SVM 方法能应对其中出现的可供训练的样本(即矿点和非矿点)有限、作为输入的特征(如地层、构造、地球化学异常、地球物理异常等)众多、构成高维数据等情况.

虽然非线性方法(如 ANN、SVM 等)在经过合理的训练后,能够得到较高的预测精度,但由于其分类过程的非线性性,分类规则不易于理解(即 ANN,SVM 的分类规则难以显著地表达不同输入特征对分类结果的贡献).因此,难以从预测过程中认识到适合于特定矿产资源预测的成矿因子的知识.本文采用 SVM-RFE[10]技术,即在 SVM 模型的训练过程中,采用特征选择方法,从各种输入的与成矿相关的特征中,选择对矿床(点)预测的重要特征.

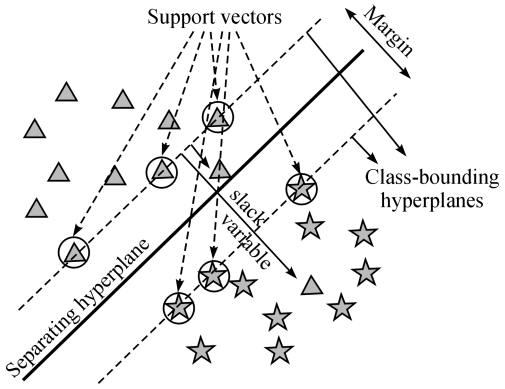
1　SVM

SVM 技术自 20 世纪 90 年代以来大受重视,其原因在于它们在有限样本情况下,对模式识别中的一些根本性问题进行了系统的理论研究,并且在此基础上建立了一种较好的通用学习算法.研究表明,SVM 较传统方法(包括神经网络)具有更好的适应和推广能力[11].另外,SVM 模型不受"维数灾难"影响,适用于高维数据综合信息提取.支持向量机方法的一个重要优点是所获得的分类器的复杂度可以采用支持向量的个数而不是变换空间的维数来刻画,因此 SVM 往往不像其他方法一样容易发生过拟合[12].

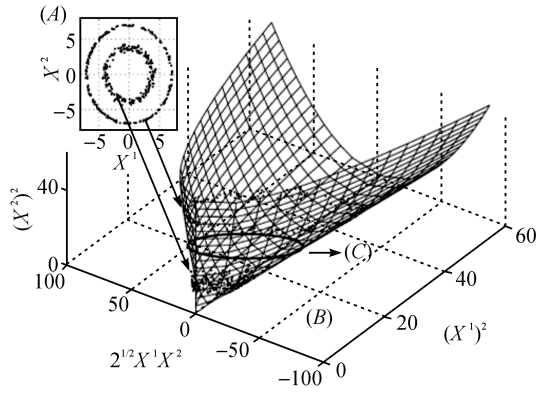
这里结合二分类问题,对 SVM 中的相关概念进行简要介绍(见图 1(a)).关于 SVM 详细的算法可以参考文献[9,12-13]有关模式识别、机器学习的教科书.图 1(a)为三角形点和五角星形点在特征空间的分布.SVM 算法是寻求这 2 个类边缘超平面(class-bounding hyperplane)之间的间隔最大化.特征空间中的沿着类边缘的数据点称为支持向量(support vectors),2 个类的支持向量分别在中心超平面(separating hyperplane)两侧,并与其相互平行的边缘超平面上(图 1(a)).类内的异常值,称为松弛变量(slack variable),它们离边缘超平面比较远,是可能的误分点.为了限制这些松弛变量对正确分类结果的影响,引入一个惩罚系数(C).在训练 SVM 时,惩罚系数值的选择是关键问题之一,过小的惩罚系数,会产生大量不合适的支持向量,但是过大的惩罚系数,可能会导致过拟合.

2 类 SVM 的分类是通过在原始特征空间中用 1 个最大化 2 类之间的边缘(margin)的超平面(hyperplane)对数据集进行分割实现的.但是绝大部分待分类的数据集在原始特征空间中难以用 1 个简单超平面进行分割,SVM 采用高维映射的办法,即将一类在原始低维空间不可用线性超平面划分的数据,映射到高维,然后寻找这样一个线性超平面,以达到分类目的.如图 1(b)所示,原始数据为 2 类同心圆分布的数据,在二维下无法用线性超平面将其

分开,但通过向高维映射后,可以采用 1 个线性超平面将其分开.



(a) SVM示意图



(b) SVM高维映射示例[14]

图 1 SVM 图

Fig.1 Figure of SVM

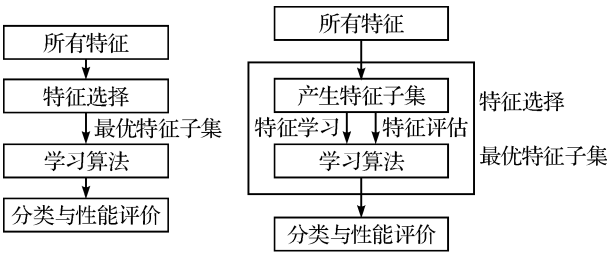
成矿预测中有矿点和无矿点之间的分类,对于成矿因子通常不是线性可分的,因此需要采用 1 个非线性核函数将分类数据集映射到一个更高维的特征空间上,以寻求线性超平面.常用的 SVM 的非线性核函数包括多项式、径向基函数(Radial Basic Function,RBF)和 S 形(双曲线正切)函数等.RBF 在实际应用中使用广泛[15-16],其半径参数 gamma(它决定 RBF 的形状)决定着这一高维映射的形态,直接影响之后的分类精度,过大的 RBF 半径会导致过拟合[17].

由此可知,SVM 的性能很大程度上是由惩罚系数 C 的值和 RBF 核函数的参数 gamma 共同决定的.一般在训练中采用 k 折交叉验证(k-folder cross validation)和网格搜索(grid search)来寻求最佳的 C 和 gamma 组合.

2 特征选择

用于矿产预测的数据往往具有高维特性,特征选择的任务是如何从许多特征中找出那些最有效的特征,即选择出区域中对矿产预测最重要的一些成矿因子(特征).一旦挑选出了这些重要特征,也就是去除了不重要的特征(不相关的、冗余的、噪声等),降低了数据维度,加快了训练、分类速度;更为重要的是使用这些保留的重要特征进行分类时可以提供更容易理解的分类规则[18].

特征选择就是通过对数据的评价,从众多的特征中挑选出用于分类的有限个特征[19],选择最佳特征的最直接也是最优的方法是穷举法,但如果特征空间过于庞大,计算复杂度就会很大,因而导致实用性不强[20].通常采用一些能得到次优解的方法,以性能为代价换取简单、快速的实现,其关键是如何建立一种评价标准来去除冗余或无关的特征组合[23].根据评价函数与分类器本身的关系可以将特征选择方法分为 3 类[21]:过滤法(Filter),评价函数与具体分类方法无关;封装法(Wrapper),采用分类器的错误概率作为评价函数;混合法(Hybrid),是以上两种方法的混合.一般过滤法与封装法相比,计算复杂度低,但精度往往不如封装法.封装法通过运用学习算法对样本数据的特征进行训练和评价,根据训练后的预测精度评价该特征子集,从而选择最优特征子集,其计算复杂度较大(见图 2).



(a) 过滤法 (b) 封装法

图 2 特征选择的两种类型[19]

Fig.2 Two types of feature selection

SVM-RFE[10]技术,是一类封装法,它在 SVM 模型的训练过程中,采用迭代特征消去(Recursive Feature Elimination,RFE)来进行特征选择的方法.RFE 从特征全集出发,进行后向序列约减,每次训练 SVM 后,根据支持向量计算出各个特征对正确预测的贡献程度,然后删除贡献最小(最不重要)的特征.经过多次训练和删除,最终实现对特征的排序,形成若干个嵌套的特征子集,然后利用分类器的预测精度作为特征子集的评估标准;在排序列表中排在前面的那些特征组合,能使分类器获得最优的分类性能[20].SVM-RFE 的实现步骤为(1)令特征集 S 包含所有特征;(2)采用特征集 S 对 SVM 进行训练;(3)对每一个特征 $p\in S$,计算其对分类的贡献度 c_p;(4)找出贡献度最低的特征 $q\in S$,即 $q=$

$\mathrm{argmin}_p\{c_p\}$;(5) 从特征集 S 中删除 q,即 $S \leftarrow S-\{q\}$;(6) 回到第(2)步,直至已经剔除了足够的特征.

3　研究区控矿因素分析

滇东南地区大地构造位置位于环太平洋与特提斯-喜马拉雅两大构造域的构造复合部位,扬子准地台西南缘与华南加里东褶皱系的过渡地带,南岭东西向复杂构造带西段与滇越巨型旋扭构造体系的接合部位.前人对滇东南金矿的研究表明,主要控矿因素有成矿区域构造环境;容矿地层及岩性、岩相与金的初始富集(地层金背景含量;成矿岩相组合型式;构造变形带和组合与金矿的再造富集);断裂构造带对金矿富集的控制(褶皱、断裂组合构造对金矿富集的控制;不整合及其后期转化断裂作用对金矿的再造富集控制);岩浆岩与金矿化;成矿物理化学条件;金成矿作用阶段;成矿方式组合及金成矿作用演化模式等.本文选择地层、构造、化探等 3 类主要特征数据进行量化,作为模型的输入.

该地已知具有金矿化的地层有寒武系、下奥陶统、下泥盆统、下二叠统、上二叠统、下三叠统、中三叠统及上三叠统等多个时代层位.从含金地层岩性看,有火山碎屑岩、硅质岩、石英细-粉砂质黏土岩等多种岩性,其中不纯碳酸盐岩和细碎屑岩接触部位最有利矿化[1].构造对于滇东南金矿化的控制作用甚至强于地层,但构造的影响在不同的矿区有不同的表现[22],如断层、不整合面等.滇东南微细粒型金矿多产于不整合面上,这是该地区的重要找金标志.根据美国卡林型金矿的经验,在有利的构造、岩性条件下,使用化探方法,利用 Au、As、Hg、Sb 等微量元素的异常是发现这类金矿行之有效的方法.在滇东南乃至"金三角"区,在多数金矿的发现过程中,地球化学方法发挥了关键性的作用.另外,本地区微细粒浸染型金矿的成矿作用与岩浆活动有着较为密切的关系.二者在空间上相互依存:基性侵入岩及火山岩是重要的赋矿岩石或矿化主岩;大量产于基性岩浆岩及火山岩区的金矿床(点)及外围金矿床(点),其成矿作用的矿源和热液都与岩浆活动有关[4].

本次研究主要选择侵入岩(从云南省 1∶100 万地质矿产图中提取,应用证据权法确定有利的缓冲区距离为 2 km,形成二值图)、不整合界线(从云南省 1∶100 万地质矿产图中提取,应用证据权法确定合适的缓冲区距离,形成二值图)、断层密度(从云南省 1∶100 万地质矿产图中提取,采用核密度计算法生成交点密度图,并应用证据权法确定合适的缓冲区距离,形成二值图)、地球化学异常(从云南省 1∶20 万区域水系沉积物测量分析成果中提取,采用中位数法进行调平,并使用迭代法剔除特高值对区内数据的影响,确定均值加 3 倍标准偏差为异常下限,形成二值图)等作为成矿预测标志进行矿产潜力的预测,包括中上泥盆统和石炭系间的平行不整合面;石炭系和下二叠统间的平行不整合面;上二叠统与三叠系的平行不整合面;下三叠统与中三叠统之间的平行不整合;Au、As、Sb、Hg 异常;侵入岩及火山岩层;断裂密度,见表 1.

表 1　输入模型的成矿因子(特征)

Table 1　Mineralization facfor of imput model

代号	内容
①	中上泥盆统和石炭系间的平行不整合面
②	石炭系和下二叠统间的平行不整合面
③	上二叠统与三叠系间的平行不整合面
④	下三叠统与中三叠统间的平行不整合
⑤	Au 异常
⑥	As 异常
⑦	Sb 异常
⑧	Hg 异常
⑨	侵入岩
⑩	断层密度

4　训练样本选择及试验结果

训练点包括 2 种类型,为存在目标矿种的点位(矿点)和为不存在目标矿种的点位(非矿点).矿点包括已知的金矿点和化探数据 Au 元素的高值点(通过证据权法对成矿高概率区筛选后得到);非矿点采用选择与金矿差别很大的其他矿点作为无矿点备选,从中剔除落入证据权法高概率区的无矿点、距离与矿点较近的矿点.最终选择 43 个有矿点和 42 个非矿点,为了评价模型的预测精度,从矿点和非矿点中分别随机选择 20% 留作验证点,即从中随机提取出 10 个矿点及 9 个非矿点作为验证点,其余的作为训练样本.

首先对上述 10 个特征进行 SVM 训练,训练过程适用"k-折交叉验证"策略选择最佳的参数(松弛因子 c 和 RBF 核半径参数 gamma),然后在 SVM-RFE 中,每迭代一次,删除对成矿预测结果影响最小的特征,然后 SVM 对剩余的特征重新训练以获得新的特征排序,再次迭代执行这一过程,最后得到特征排序:①Au 异常,②As 异常,③侵入岩,④下三叠统与中三叠统之间的平行不整合面,⑤上二叠统

与三叠系的平行不整合面;⑥断裂密度,⑦石炭系和下二叠统间的平行不整合面,⑧中上泥盆统和石炭系间的平行不整合面;⑨Sb 异常,⑩Hg 异常.最后计算出这一序列不同嵌套子集的预测精度(见表 2).通过对滇东南地区金矿预测的实践表明,采用 SVM-RFE 技术从原始 10 类成矿因子中自动选择前 6 个特征用于预测,其精度最高为 94.74% .通过检查验证数据发现,该次预测中,10 个矿点全部命中,9 个非矿点仅有 1 个被错分.

表 2 成矿因子的分析结果

Table 2 The Analysis results of mineralization factor

排列顺序	特征排序类表	特征子集	预测精度/%
1	⑤	⑤	68.42
2	⑥	⑤,⑥	68.42
3	⑨	⑤,⑥,⑨	73.68
4	④	⑤,⑥,⑨,④	73.68
5	③	⑤,⑥,⑨,④,③	73.68
6	⑩	⑤,⑥,⑨,④,③,⑩	**94.74**
7	②	⑤,⑥,⑨,④,③,⑩,②	78.95
8	①	⑤,⑥,⑨,④,③,⑩,②,①	84.21
9	⑦	⑤,⑥,⑨,④,③,⑩,②,①,⑦	78.95
10	⑧	⑤,⑥,⑨,④,③,⑩,②,①,⑦,⑧	84.21

5 结 论

本文采用 SVM-RFE 技术,从所有输入的成矿因子中选择出对矿床(点)正确预测的重要因子.通过对滇东南地区金矿预测的实践表明,采用 SVM-RFE 技术从原始 10 类成矿因子中自动选择前 6 类进行预测的精度从 68.42% 提高到 94.74% ,并且得到该区域进行矿产预测的成矿因子的重要性排序.这一排序结果与滇东南微粒型金矿成矿模型相吻合,突出了重要的成矿因子,有利于今后指导该区域进行金矿预测的资料选取和提高预测的准确程度和效率.本文只选择了研究区控矿因素的主要特征进行建模,并未采用全部因素和变量进行量化和表达,在后期的研究中可以考虑对全部控矿因素进行特征选择,建立更为全面的矿产预测模型,这样更有利于提供全面的成矿因子分析.

感谢审稿人对本文的有益意见和修改建议.

参考文献(References):

[1] 江鑫培.滇东南地区金矿类型及成矿地质条件分析[J].**云南地质**,1996,15(3):277-281.
JIANG Xing-pei. Analysis of gold deposit type and geological condition in southeast Yunnan[J]. **Yunnan Geology**,1996,15(3):277-281.

[2] 覃文明,何志美,徐自斌.滇东南微细浸染型金矿构造控矿型式[J].**云南地质**,2007,26(3):298-302.
QIN Wen-min, HE Zhi-mei, XU Zi-bing. The structural ore-control pattern of microdisseminated Au deposit in southeast Yunnan[J]. **Yunnan Geology**,2007,26(3):298-302.

[3] 赵鹏大.**矿床统计预测**[M].北京:地质出版社,1994:113-219.
ZHAO Peng-da. **Statistical Prediction of Mineral Deposit**[M]. Beijing: Geological Publishing House, 1994,113-219.

[4] 朱赖民,刘显凡,金景福,等.滇-黔-桂微细浸染型金矿床时空分布与成矿流体来源研究[J].**地质科学**,1998,33(4):463-474.
ZHU Lai-min, LIU Xian-fan, JIN Jin-fu, et al. The study of the time-space distribution and source of ore-forming fluid for the fine-disseminated gold deposits in the Yunnan-Guizhou-Guangxi area[J]. **Scientia Geologica Sinica**,1998,33(4):463-474.

[5] PORWAL A. **Mineral Potential Mapping with Mathematical Geological Models**[D]. Enschode: the University of Twente,2006.

[6] SINGER D A, KOUDA R. Application of a feedforward neural network in the search for Kuroko Deposits in the Hokuroku district, Japan[J]. **Mathematical Geology**,1996,28(8):1017-1023.

[7] HARRIS D P, PAN G C. Mineral favorability mapping: a comparison of artificial neural networks, logistic regression and discriminate analysis[J]. **Natural Resources Research**,1999,8(2):93-109.

[8] BROWN W M, GEDEON T D, GROVES D I, et al. Artificial neural networks: a new method for mineral prospectivity mapping[J]. **Australian Journal of Earth Sciences**,2000,47:757-770.

[9] VAPNIK V N. **The Nature of Statistical Learning Theory**[M]. New York: Springer-Verlag,1995.

[10] GUYON I, WESTON J, BARNHILL S. Gene selection for cancer classification using support vector machines[J]. **Machine Learning**,2002,46:389-422.

[11] BURGES C J C. A tutorial on support vector machines for pattern recognition[J]. **Data Mining and Knowledge Discovery**,1998,2:121-167.

[12] DUDA R O, HART P E, STOCK D G. **Pattern Classification**[M]. 2nd ed, New York: John Wiley & Sons, 2001.

[13] MITCHELL. **Machine Learning**[M]. New York: The McGraw-Hill Companies, Inc,1997.

[14] BELOUSOV A I, VERZAKOV S A, VON FRESE J. A flexible classification approach with optimal generalisation performance: support vector machines[J]. **Chermometrics and Intelligent Laboratory Systems**, 2002,64:15—25.

[15] FOODY G M, MATHUR A. A relative evaluation of multiclass image classi? cation by support vector machines[J]. **IEEE Transactions on Geoscience and Remote Sensing**,2004,42(6):1335—1343.

[16] OOMMEN T, MISRA D, TWARAKAVI N K C, et al. An objective analysis of support vector machine based classification for remote sensing[J]. **Mathematical Geosciences**,2008,40:409—422.

[17] NOBLE W S. What is a support vector machine[J]. **Nature Biotechnology**,2006,24(12):1565—1567.

[18] LIU H, YU L. Toward integrating feature selection algorithms for classification and clustering[J]. **IEEE Transactions on Knowledge and Data Engineering**, 2005,17: 491—502.

[19] 张良培,张立福.**高光谱遥感**[M].武汉:武汉大学出版社,2005:162.
ZHANG Liang-pei, ZHANG Li-fu. **Hyperspectral Remote Senisng**[M]. Wuhan: Wuhan University Press,2005:162.

[20] 卢运梅.**SVM-RFE 算法在数据分析中的应用**[D].长春:吉林大学,2009.
LU Yun-mei. **Application of SVM-RFE Algorithm to Data Analysis**[D]. Changchun: Jilin University, 2009.

[21] MANORANJAN D, LIU H. Feature selection for classification[J]. **Intelligent Data Analysis**, 1997, 1: 131—156.

[22] 仕竹焕,杨云保.初论滇东南地区微细粒型金矿找矿标志[J].**西部探矿工程**,2007,5:65—68.
SHI Zu-huan, YANG Yun-bao. Preliminary study on finding deposit of micro-grain type gold deposits in southeast Yunnan[J]. **Western China Exploration Engineering**,2007,5:65—68.

[23] 柏坚,俞乐.径向基神经网络模型在滇东南金矿潜力预测中的应用[J].**浙江大学学报:理学版**,2011,38(3):354—361.
BAI Jian, YU Le. Application of radial basis function neural networks to prospectivity mapping for gold deposits in Southeastern Yunnan, China[J]. **Journal of Zhejiang University: Science Edition**,2011,38(3):354—361.

(**责任编辑**　寿彩丽)