# 基于光谱相似尺度的支持向量机蚀变信息提取

傅文杰[1,2],洪金益[1],朱谷昌[3]
(1. 中南大学地学与环境工程学院,长沙 410083;
2. 莆田学院,莆田 351100; 3. 有色金属矿产地质调查中心,北京 100814)

[摘 要]文章提出一种基于光谱相似尺度(spectral similarity scale - SSS)的支持向量机(support vector machines - SVM)遥感数据矿化蚀变信息提取的新方法。该方法选择青海两兰地区作为遥感矿化蚀变信息典型研究区,利用该区域的Landsat7ETM遥感影像结合地面实况调查数据,从图像上选取少量具有代表性的样本点的光谱作为参考光谱,利用SSS方法提取训练样本,然后应用SVM算法进行遥感矿化蚀变信息提取。试验结果经野外检查和验证,效果良好。

[关键词]光谱相似尺度 支持向量机 矿化蚀变信息 遥感数据
[中图分类号]P627 [文献标识码]A [文章编号]0495 - 5331(2006)02 - 0069 - 05

## 0 引 言

遥感蚀变信息提取在区域地质制图和找矿中应用的成效非常显著,尤其是在一些地质工作程度较低的偏远地区,从遥感图像上圈定成矿靶区,能大大的减少了野外工作的盲目性。蚀变信息是一种微弱信息,常受其他地物信息所干扰,因此,增强并提取出蚀变信息存在着一定的难度。比值法与主成分分析法是目前遥感矿化蚀变信息提取的常用方法[2]。但这些方法主要着眼于增强处理,以求目视效果,在背景与噪声的干扰下,蚀变信息的增强效果并不理想。近期新出现的一些方法如:小波分析、神经网络、分数维几何(分形或分几)等,也由于受样本数目和维数的限制,得到的结果一般不能令人满意。支持向量机(SVM)是一种基于统计学理论的机器学算法,在解决有限样本、非线性及高维模式识别问题中表现出许多特有的性能,并且具有强大的泛化能力。但支持向量机算法本身是一种有监督学习方法,需要一定数量标注过类型的训练样本,对遥感蚀变信息而言,训练样本选取的方法主要通过野外实地调查,然后在遥感影像上,利用准确的界线建立感兴趣区,选取训练样本。通过地面调查获取样本信息效率低。

光谱相似尺度是通过同时度量两个光谱之间的大小和形状来判定光谱相似性的,而传统的光谱相似性度量方法(如:欧氏距离、相关系数、光谱角等)只计算两个光谱之间的大小(亮度)或形状。因此本文提出了一种基于光谱相似尺度(SSS)的支持向量机(SVM)遥感矿化蚀变信息提取算法,通过对青海两兰地区的试验证明了该方法在遥感矿化蚀变信息提取应用中的可行性,同时得到了令人满意的效果。

## 1 基于光谱相似尺度的支持向量机算法

### 1.1 SVM基本原理

支持向量机(Suport Vector Machines - SVM)完整的数学描述见文献[3, 4],这里我们只对这个算法进行简单描述。支持向量机的基本思想可用图1说明。图中,实心点和空心点代表两类样本,H为分类线,H1、H2分别为过各类中离分类线最近的样本且平行于分类线的直线,他们之间的距离叫做分类间隔(Margin)。假设训练样本集为$(x_i, y_i)$, $i = 1, \cdots n$, $x \in R^d$,其中$x$为输入向量,$i$为样本数,$y$是输入向量所属的类别,对于两类的分类问题,$y \in \{+1, -1\}$。D维空间中线性判别函数的一般形式为,$g(x) = w \cdot x + b$分类面方程为:

$$w \cdot x + b = 0 \qquad (1)$$

我们将判别函数进行归一化,使两类所有样本都满

[收稿日期]2005 - 08 - 24;[修订日期]2005 - 10 - 11;[责任编辑]曲丽莉。
[基金项目]国家"十五"科技攻关计划项目(编号:2003BA612A - 04)资助。
[第一作者简介]傅文杰(1967年 - ),男,1987年毕业于中南大学,获学士学位,在读博士生,高级工程师,现主要从事遥感技术及GIS应用工作。

足$|g(x)|\geq 1$,即:让离分类面最近的样本的$|g(x)|=1$,这样分类间隔就等于$2/\|w\|$,因此使间隔最大等价于使$\|w\|^2$最小;而要求分类线对所有样本正确分类,就是要求它满足约束条件:

$$y_i[(w\cdot x_i)+b]-1\geq 0, i=1,\cdots,n \quad (2)$$

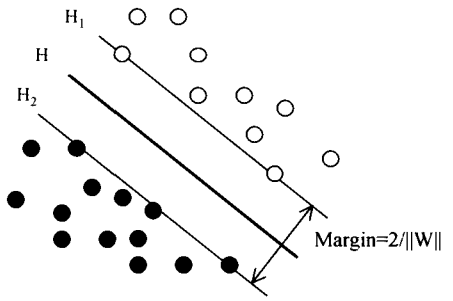


图 1 支持向量机的基本思想

考虑到有些训练样本是线性不可分的,Vapnik 和 Cortes 等人[5]引入了非负的松弛变量,将(2)式放宽为:

$$y_i[(w\cdot x_i)+b]\geq 1-\xi_i, \xi_i\geq 0, i=1,\cdots,n \quad (3)$$

显然,当划分出现错误时$\xi_i$就会大于零。因此在求分类平面的同时,$\sum_i \xi_i$的值也希望愈小愈好。所以原本目标函数是求$\|w\|^2/2$的最小值,会变成求目标函数$\|w\|^2/2+C(\sum_i \xi_i)$的最小值;其中 C>0 是一个常数,是可调的参数,控制对错分样本惩罚的程度,C 越大表示对错误的惩罚越重。这是一个二次规划问题,其对偶形式为:

$$\max \sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j) \quad (4)$$

$$\text{st.}\ 0\leq a_i\leq C\ i=1,\cdots n \quad (5)$$

求解这个对偶问题得到最优分类函数:

$$f(x)=\operatorname{sgn}\{\sum_{i=1}^{n}\alpha_i y_i(x_i\cdot x)+b\} \quad (6)$$

$\alpha_i$为 Lagrange 乘子,$\alpha_i$不为零的样本点就称作支持向量,即位于$H_1$、$H_2$上的样本。这些向量充分描述了整个训练样本集数据的特征,使得对支持向量集的线性划分等价于对整个数据集的分类。

对于非线性可分样本,支持向量机构造分类决策函数的方法是,首先将训练数据从原始模式空间经过特定核函数的非线性变换,映射到高维特征空间[6]。然后,在特征空间中,寻找最优分类超平面,该超平面实际上对应着原始模式空间中的非线性分类面。因此,支持向量机在处理非线性情况时,仅比线性情况多了一个非线性映射环节。其对偶形式变为:

$$\max \sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_y K(x_i,x_j) \quad (7)$$

其中$K(x_i\cdot x)$为满足 Mercer 条件的核函数,目前,应用较多的核函数有 3 种:即多项式核函数、径向基核函数和 Sigmoid 核函数。

对于多类分类问题,可采用多个二分类 SVM 组合的办法解决。

### 1.2 光谱相似尺度

光谱相似尺度(SSS)[7]的数学定义在基于向量同一性的定义(两个相同的矢量有相同的大小和方向)的定义和向量的大小与方向无关的假设。因此,两个相似的反射光谱应具有相似的光谱大小相似的方向。为了定量表示两个光谱间的大小和方向的相似度,我们引入了光谱相似值(spectral similarity value - SSV),SSV 的数学表达式为(8)式:

$$SSV=\sqrt{d_e^2+\hat{r}^2} \quad (8)$$

这里,$d_e$为广义的欧氏距离,主要度量光谱间的大小差异,由式(9)式计算。$\hat{r}$为 1 减去相关系数平方的差,主要度量光谱间的形状差异,计算式为(10)式。

$$d_e=\sqrt{\frac{1}{N}\sum_{i=1}^{N}s(X_i-Y_i)^2} \quad (9)$$

其中:$N$为影像像元光谱的波段数,$X$为影像像元光谱矢量,$Y$为参考光谱矢量。

$$\hat{r}=1-r^2 \quad (10)$$

其中:$r$为相关系数,由(11)式计算

$$r=\left[\frac{\sum_{i=1}^{N}(X-\mu_X)(Y-\mu_Y)}{(N-1)\sigma_X\sigma_Y}\right] \quad (11)$$

其中:$\mu_X$、$\mu_Y$分别为两光谱的均值,$\sigma_X$、$\sigma_Y$分别为两光谱的标准偏差。

SSV 值的大小范围为 0 至$\sqrt{2}$,SSV 越小表示两光谱越相似。从理论上说,SSV 为 0 是同一类地物光谱,但在实际上,同类地物光谱特征也会存在一定的差异,常常设定一个较小的域值,SSV 值小于该域值的像元就划分为该类。

### 1.3 基于 SSS 的 SVM 算法

在该算法中,我们先计算图像中每个像元光谱与每类参考光谱之间的 SSV,将 SSV 值小于设定阈值的像元选为训练样本,然后利用 SVM 构造最优分类器模型,最后利用该分类器模型对遥感图像进行信息提取,其算法描述如下:

1)对每个影像像元光谱计算它与各类实测光

谱间的 SSV 值,并将该值和所设定的初始阈值比较,若小于某类设定阈值,则选为该类的训练样本。这里的技术关键是初始阈值的设置,既要保证采集到足够的样本,又要保证训练样本与提取目标的最大相似性;

2) 如果某类的训练样本数未达到所需的样本数,则重新调整阈值,执行 1)。

3) 利用提取的训练样本集应用 cross - validation 算法确定最优 SVM 模型参数。

4) 根据最优 SVM 模型参数训练 SVM 分类器模型。

5) 采用训练好的 SVM 模型对整个遥感图像进行信息提取

## 2 试验结果及分析

### 2.1 试验数据

试验场选择青海两兰金多金属成矿区。该试验区范围长 14km,宽 25km,面积为 350km$^2$,区内主要褶皱构造为 NWW 走向的阿尔茨托山—吉给申沟南山复式倒转背斜。断裂构造以 NWW 或 NW 向为主,次为 NE 向或 SN 向。区内主要的岩浆侵入岩为加里东期、华力西期和印支期的花岗岩、花岗闪长岩、闪长岩以及元古宙的辉长岩、超基性岩。本区是一个铜、铅、锌、银多金属矿化集中区,成矿条件较好,矿产资源潜力大、找矿前景好。

试验所用的数据为 2001 年 7 月 3 日的 landsat7 ETM 数据,选用其中的 6 个波段(1、2、3、4、5、7 波段),图像大小为 881 行、489 列。对图像进行包括大气校正、几何校正和地理配准等的数据预处理。因为 TM 图像的每个像元值是灰度值,需将它转换为反射率值,具体转换公式为[5]:

$$L = gain * DN + bias \tag{12}$$

$$\rho = \pi L d_s^2 / (E_0 \cos\theta) \tag{13}$$

其中:L 是地物在大气顶部的辐射亮度,DN 是像元灰度值,增益(gain)和偏移值(bias)可从头文件中得到,$\rho$ 是地物反射率,$d_s$ 为日地天文单位距离,$E_0$ 为大气顶部的太阳辐照度,$\theta$ 为成像时的太阳天顶角,从图像的头文件中读取。

### 2.2 训练和测试样本采集

由于支持向量机法是根据训练样本来选择特征参数,建立判别函数进行分类的。因此,训练样本选择的质量数量很大程度上关系到整个分类结果的精度,有研究表明训练样本选择比分类算法的选择对分类精度的影响更大[8]。

本次训练和测试样本采集先是根据已知矿床(点)地质资料(图 2)并结合野外调绘资料,确定本研究区的主要地物有 8 类,其中有两类为本研究要提取的信息,即:铁化蚀变岩及泥土蚀变岩,其他 6 类为:片麻岩、石英砂岩、凝灰质砂岩、砂砾岩、肉红色花岗岩、灰白色花岗岩。利用该区域的 Landsat7ETM 遥感影像结合已知矿床(点)地质资料和野外调绘资料,在典型地区对铁化蚀变岩、泥土蚀变岩、片麻岩、石英砂岩、凝灰质砂岩、砂砾岩、肉红色花岗岩及灰白色花岗岩分别选取了一定数量的具有普遍性、代表性的样本点。进一步,利用 ENVI 软件的光谱工具对样本进行了"n 维散度空间"提纯,得到各类"纯净"样本点,选取经过提纯后的各类样本在对应 TM1 TM2 TM3 TM4 TM5 和 TM7 波段的光谱生成各岩石类别的参考光谱,然后利用 SSS 算法,分别计算每个像元光谱与各类参考光谱之间的 SSV 值,将 SSV 值小于设定阈值的像元选为训练样本。这里每类阈值的设定是以保证各类分别采集到 200 个左右训练样本为准。

### 2.3 SVM 模型选择及实验结果

选择合适的核函数参数和误差惩罚因子 C 对 SVM 的性能至关重要。因此,在使用支持向量机进行分类和预测时,如何选择适当的参数就成了一个非常重要的问题。Vapnik 等人的研究表明,SVM 的性能与所选用的核函数的类型关系不大,而核函数的参数和误差惩罚因子 C 是影响 SVM 性能的主要因素[9]。因此在笔者设计 SVM 分类器时选择 RBF 函数作为核函数,原因主要有以下几点:首先,RBF 函数可以将样本非线性地规划到更高维的空间中,从而解决类标签和属性间非线性的关系问题,这是线性核函数无法解决的。实际上线性核函数是 RBF 核函数的特例。另外,Sigmoid 核函数取某些特定参数时性能和 RBF 相同。因此对本文算法来说,需要选择的参数即为 RBF 核参数 $\gamma$ 和惩罚因子 C。本文采用交叉比对算法(cross - validation)来选取这两个参数。交叉比对法的过程是:将样本集分成 n 个子集,每次将其中 n - 1 个子集代入 SVM 训练,针对 SVM 的惩罚因子 C 和不同的核函数,计算剩余 1 个子集的分类正确率。系统根据误差自动调整惩罚因子的取值。通常,n 的取值一般为 10。通过上述实验选择的最佳 SVM 模型参数为 $\gamma = 0.5$, $C = 32$。最后应用该模型对研究区整个 ETM 影像进行了信息提取,提取结果见图 2。本文实验中采用的 SVM

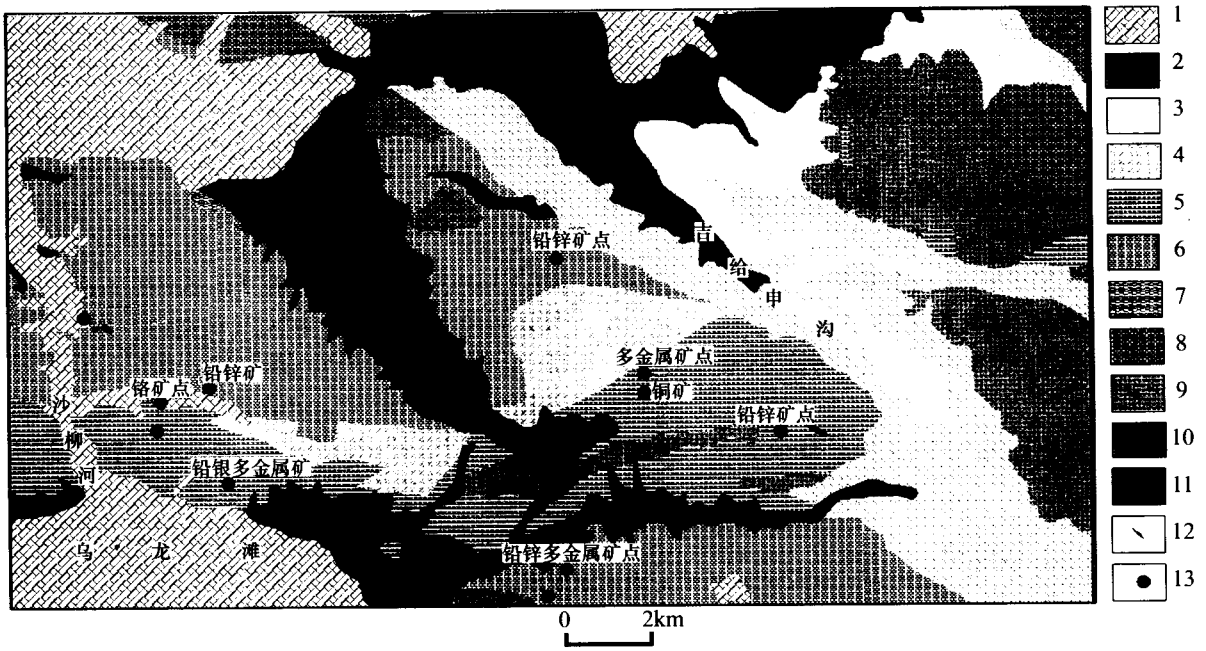


图 2 青海两兰地质矿产图

1—冲积物、风成砂、砂砾石及粉砂质粘土;2—砾石层及粉砂质砾石层;3—黄—棕色砂砾岩;4—凝灰质砾岩夹安山岩及板岩;5—变长石石英砂岩,局部夹结晶灰岩;6—黑云角闪斜长片麻岩;7—肉红色细—中粒花岗岩;8—灰白色细—中粒花岗岩;9—灰白色片麻状花岗岩类;10—灰绿—黑绿色超基性岩类;11—灰绿色变角闪闪长岩;12—花岗斑岩(γ、π);13—矿区(点)


图 3 SVM 提取铁化及泥化信息叠加 ETM 原图(7,4,1)

算法来自台湾大学 Chih - Jen Lin 开发的 LibSVM 软件[10], LibSVM 中采用"一对一"的策略解决多类别分类问题。LibSVM 已广泛用于 SVM、回归和分类估计,并且支持多类分类,它的基本算法结合了 Platt

提出的SMO[11]和Joachims提出的SVMLight[12]算法思想。

### 2.4 精度分析

精度分析是遥感数据分类过程中一项不可缺少的工作。通过精度分析,分类者能确定分类模式的有效性,改进分类模式,提高分类精度。本次实验主要采用两种查证方法:一是对已知矿床(点)的查证,主要用地质矿产图与蚀变遥感异常信息图进行叠合分析,统计吻合程度,对部分典型矿床进行野外查证;二是典型异常信息的野外查证。本试验区地质矿产图上已经登记的矿床(点)12处,发现有11处矿床(点)与提取的蚀变遥感异常信息吻合,吻合率为91.6%。在异常信息提取的基础上,选择部分异常点及矿床(点)进行了野外查证,新发现的一些矿化蚀变异常点在野外均不同程度存在矿化蚀变现象,与物化探异常叠合性好。

## 3 结论与问题

本文提出的一种新的基于光谱相似尺度的支持向量机遥感矿化信息提取方法,既减少野外实地调查工作量,同时保证了较高的分类精度,得到了比较满意的效果。克服了传统的提取遥感矿化蚀变信息方法应用中受样本数目以及维数的限制等局限,同时因为光谱相似尺度是同时度量光谱间的大小和形状,因此可以减少异物同谱的现象。另外,如何选取理想的核函数以及相应的参数,还需要进一步研究。

[参考文献]


[1] Yu X, Reed I S, Stocker A D, Comparative performance analysis of adaptive multispectral detectors[J]. IEEE Transaction on Signal Processing, 1993, 41 (8):2639 ~ 2656.

[2] Ruiz - armenta J R. Prol - ledesma. Techniques for enhancing the spectral response of hydro - thermal alteration minerals in Thematic Mapper images of Central Mexico[J]. Int. J. Remote Sensing, 1998, 19 (10):1981 ~ 2000.

[3] Vapnik V N. The Nature of Statistical Learning theory[R]. New York: Springer Verlag, 1995.

[4] Vapnik V N. Statistical Learning theory[R]. New York: Wiley, 1998.

[5] Cortes C, Vapnik V N, Support - Vector Networks[J]. Machine Learning, 1995, 20(3):273 ~ 297.

[6] Scholkopf B, Mika S, Burges C. Input space vs. feature space in kernel - based methods[J]. Transactions on Neural Networks, 1999, 10(5):1000 ~ 1017.

[7] Granahan J C, Sweet J N. An evaluation of atmospheric correction techniques using the spectral similarity scale[J]. IEEE 2001 International Geoscience and Remote Sensing Symposium, 2001, 5: 2022 ~ 2024.

[8] Hixson M, Scholz D, Fuhs N, et al. Evaluation of several schemes for classification of remotely sensed data[J]. Photogrammetric Engineering and Remote Sensing, 1980, 46:1547 ~ 1553.

[9] Vapnik V N. 统计学习理论的本质[M]. 北京:清华大学出版社. 2000.

[10] Chih - Chung Chang, Chih - Jen Lin. LIBSVM: A library for support vector machines, 2001. software available at http://www.csie.ntu.edu.tw/ ~ cjlin/libsvm

[11] Platt J. Fast training of support vector machines using sequential minimal optimization[A]. Advances in Kernel Metheds - support vector Learning. Cambridge[C]. MA: MITpress, 1999. 185 ~ 208.

[12] Joachims T. Making Large - scale SVM Learning Practiceal [M]. In Advances in Kernel Methods - support Vector Learning, MIT Press, Cambridge, MA, 1998:169 ~ 184


## EXTRACTING ALTERED AND MINERALIZED ROCK INFORMATION FROM REMOTE SENSING IMAGE BASED ON SUPPORT VECTOR MACHINES AND SPECTRAL SIMILARITY SCALE


FU Wen - jie[1,2], HONG Jin - yi[1], ZHU Gu - chang[3]

(1. *School of Geoscience and Environmental Engineering, Central South University, Changsha* 410083;
2. *Putian College, Putian* 351100; 3. *Nonferrous Metals Resource Geological Survey of China, Beijing* 100814)



**Abstract**: A new method for extracting mineralization information from remote sensing image based on support vector machines (SVM) and spectral similarity scale (SSS) is presented. Based on the Landsat 7 ETM data and ground truth data, Lianlan area in Qinghai Province is taken as a typical region to select training size of main rock types with the method of SSS. Then SVM algorithm was applied to extract mineralization information from remote sensing image. Practice has proved that such a method is effective in extracting mineralization information.



**Key words**: spectral similarity scale, support vector machine, altered and mineralized rock information, remote sensing data