

模式识别中共享近邻法在化探中的初步应用

邓国瑛

(中国有色金属工业总公司矿产地质研究院)

本文主要探讨模式识别中共享近邻法在化探中的应用,并以湘南—粤北地区不同成因铁帽的分类和九毛地区超基性岩体的含矿性评价为例,分别用多类与两类判别进行了计算,分类结果用四层结构的树状图表示。结果表明,共享近邻法不但对已知模型样品分类正确,对未知样品归类清楚,而且能为快速找矿评价提供重要信息。

非参数的模式识别方法已广泛应用于科学研究、生产技术等一些领域。近年来,该方法已开始用于地质成矿统计预测,岩石、岩浆岩的合理分类,地层划分,以及地球化学找矿中对岩体、铁帽、露头 and 地球化学异常的深部含矿性评价等方面,本文主要探讨模式识别中共享近邻法在化探中的应用。

方法简介

模式识别理论的一个中心内容是判别分类,可分为参数的和非参数的两大类。参数的判别分类方法建立在假设模式空间中的样本分布服从某种数理统计模型,并根据统计准则对样品进行分类。非参数判别方法的基本假定是:若样品类型上的不同是由一组特征(即变量)所确定,则同类样品应该比较相似,这种相似性使它们在同一模式空间中相聚,而不同类型的在此空间中应占据不同的区域。

由R. A. Jarvis 和 E. A. Patrick 等提出的共享近邻法,是一种典型的非参数判别方法。主要是根据点与点之间的联系实现分类,可自动调节模式空间中一对点的相似性测量尺度。是一种简便的模式识别方法。其基本设想是:

设 p 维模式空间中的样本矢量集为 $\{\vec{x}_i\}$
 $i = 1, \dots, N$, $\vec{x}_i$ 与 $\vec{x}_j$ 是该集合中两两样品间的欧氏距离,在 $\vec{x}_i$ 与 $\vec{x}_j$ 周围各找 k 个近邻点,若 $\vec{x}_i$ 的 k 个近邻点中有些点能够在 $\vec{x}_j$ 的 k 个近邻点中找到,则称 $\vec{x}_i$ 、 $\vec{x}_j$ 为相似的。共同享有的点越多,

其相似程度就越高,反之亦然。计算步骤如下:

(一) 构造 k 近邻表 将已知模型样的数据按类别依次放在前面,待判样品放在后面。先计算两两样品间的欧氏距离,按欧氏距离的大小找出每一个样品周围比较近的 k 个邻点,并按其距离从小到大(相对从近到远)进行排列,每一个样品自身作为它的第零个近邻。以后的工作都根据该近邻表进行整数运算。

(二) 建立标号表 标号表对应于样品的顺序(从1到 N),主要用于分类时作比较。

(三) 分类 从近邻表的第一、二行开始,若每一行中的第零个近邻点可以在另一行的邻点中发现,且有大于 k_i 个相同的共享近邻数,则从标号表中用对应的较小的标号替换较大的标号。反复进行此过程,直到所有的点都一对一对地比较完。最后根据标号表中相同的标号划归同一类。

(四) 调节相似性测量尺度 重新给出 k_i 、 k_j 值,重复第(二)、(三)步运算,可得到新的分类结果。

k_i 、 k_j 都是调节相似性测量尺度的可控参量。 k_i 是近邻数,即某次分类选取的近邻个数; k_j 是共享近邻数,即要求某次分类必须在多少个共享近邻以上进行的一个阈值。如给定 $k_i = 6$, $k_j = 2$,则在 k 近邻表中只能取前6个近邻进行比较分类,这时,两个样品分为同一类的条件是:在各自的6个近邻中均包含对方,且二者都有大于2的相同共享近邻数。所以, k_i 越大,分类的条件就越高,致使样品很难聚类。

共享近邻法可提供两个重要信息,一是各样品的分类结果,据此可绘制分类树状图;另一个是近邻表,从表中可以具体了解模式空间中每一个点的周围情况及各模式之间的联系。

应用条件试验

为了搞清方法的应用条件,和影响方法使用效果的关键因素,我们进行了以下几方面试验。

(一) k_i 和 k_r 的最佳取值范围 k_i, k_r 值可任意给定,其值不同得到的分类结果也不相同。通过调节 k_i, k_r 值,可根据不同需要了解模式局部或整体的空间结构。对后面的两个应用实例,分别进行了几十种 k_i, k_r 组合计算,表1总结了各种取值情况。表中“不用范围”是使分类不聚群的 k_i, k_r 取值范围,这些范围的计算结果使所有样品(包括已知和待判的)都各自成一类,即有多少个样品就分多少类。近邻数 k_i 的最佳取值范围是 5~7,而共享近邻数 k_r 最佳取值是 1 和 2。该范围是否适用于样品数相当大的情况,还有待进一步试验。

		k_i					
		3	4	5	6	7	8
k_r	不用范围	≥ 1	≥ 2	≥ 3	≥ 4	≥ 5	≥ 6
	可用范围		1	1~2	1~3	1~4	1~5
	最佳范围			1~2	1~2	1~2	

(二) 变量的选择 试验结果表明,经挑选变量后的分类结果,明显地优于将所有变量直接计算的分类结果,所以,计算前优选变量对该方法应用效果至关重要。

通常,变量选择可由计算机用数学方法来完成。但数学方法不能涉及到地质和其他一系列因素,而且,目前对变量选择还没有一种普遍适用的方法和理论,致使机选变量效果往往不够理想。我们认为,在地质与地球化学研究程度较高的地区,元素的指示意义较清楚,根据地化特征来选取类间差异较大的变量是合理的。

(三) 样品数与变量数(维数)的比值范围 共享近邻法对样品数 n 与变量数 d 的比值范围有一

定要求。实验结果得出:多类判别要求 $n/d > 3$; 作两类判别时, $n/d > 2$ 即可。一般 n/d 比值应尽量高一些,以保证其应用效果的可靠性。

应用实例

(一) 铁帽含矿性评价 湘南—粤北地区广泛发育着不同成因的铁帽,在研究建立已知区铁帽数学模型过程中,我们应用了多类逐步判别分析、对应分析、点群分析等方法,其结果都不太好,已知铁帽错分较多。分析其原因,主要是因为这批样品点在模式空间的分布不服从参数的判别分类方法所建立的某种数理统计模型,或者是所用方法的相似性度量不合适,导致分类结果失效。为此,我们利用共享近邻法进行了计算。

湘南—粤北地区的铅锌矿主要有两种成因:岩浆热液型和层控型。在这两类矿床上方及其附近都不同程度地发育有铁帽。另外,在某些含铁质(主要是碎屑岩)和锰质(主要是碳酸盐岩)较高的地层中,经风化、淋滤后也分别形成一些零星分布的铁帽和锰帽。通过对这四类不同成因的铁帽和锰帽地球化学特征的研究,挑选以下类间差异较大的元素和比值: Cu、Mo、Pb、Zn、W、Ni、Hg、Mn/Zn、Pb/Bi 9 个变量。参加计算的样品有 34 个,其中已知模型样品 16 个,待判样品 18 个。按 $k_i = 6, k_r = 1$ 的计算结果列于表 2。

为了使分类结果更直观,我们采用四层结构的树状图来表示(图1)。它主要由以下几部分组成:

第一层:根节点,由各已知大类(铁帽类)构成。

第二层:枝节点,由各已知大类新分出的类别构成。节点上用罗马数字标明类号。

第三层:小枝节点,由归入同一小类的所有样品组成。节点上用等线体阿拉伯数字标明各小类号。

第四层:树叶,由已知样品与待判样品按分归的小类顺序排列而成。为清楚起见,各类已知样品用特殊标志,待判样品用空心圆表示,其上方均标明样品号。如果样品数很多,可以用一个

同类型的大标志代表一些样品。

这种树状图对分类结果一目了然。作图时，

根据表 2 先从树叶开始往下画，不但手绘快而简单，在计算机上也易实现。

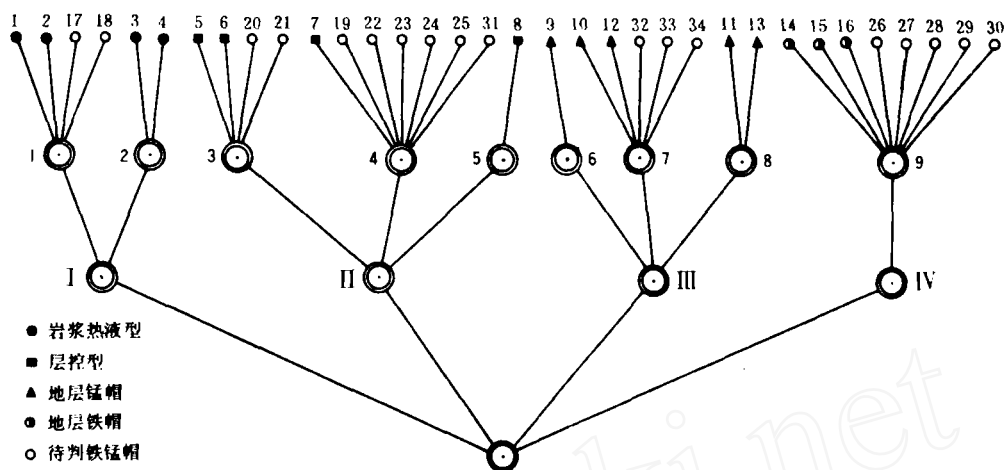


图1 铁帽分类树状图

从图1可见，各类已知铁帽的划分虽然又出现了新的小类，但都能清楚而准确地进入四大类。待判样品归类较好，没有出现孤立和分散的新类。

先看第Ⅰ类，岩浆热液型。该类4个铁帽被分成1、2两个小类。这是因为湖南水口山和黄沙坪都是以Pb、Zn为主的多金属矿床，而湖南宝山和广东大宝山则是以Cu、Pb、Zn（Mo、W）为主的多金属矿床。这种差异也反映到其上部或附近分布的铁帽中，即宝山与大宝山铁帽中含Cu比前者高。共享近邻法能将这二者区别开来，说明该法的确有它的独到之处。判入岩浆热液型的有湖南茶山铁帽和宝山方园锰帽。其中茶山是已知铅锌矿床，由于它的成因有争议，特将茶山铁帽作为待判样品参加计算，结果它被判为第Ⅰ类，茶山铅锌矿床属于岩浆热液成因的说法在此得到进一步证实。宝山方园锰帽分布在宝山铜铅锌多金属矿床的近围，其Pb、Zn、W、Bi含量都很高，推断该锰帽下方可能存在盲的岩浆热液型铅锌多金属矿化体。

对于第Ⅱ类层控型铁帽，共为分3、4、5三个小类。其中广东凡口与杨柳塘都是典型的层控型铅锌矿床，所以这两个铁帽被分进第3小类。而广东西岗寨是含铅锌的黄铁矿矿床，与前者不同，则被分为第4小类。湖南太坪铁帽为什么单独进入第5小类呢？根据样品分析资料，该矿床

外围铁锰黑土中Pb+Zn最高可达20000~30000ppm，而矿床内部产出的铁帽中Pb+Zn同样很高，可达15000ppm。实际上，太坪是氧化堆积形成的富铅锌矿床。共享近邻法能把铁帽中存在的这种差异完全反映出来，这是参数判别方法无能为力的。判入层控型的有广东凡区凡口村、铁石岭、富屋背后山铁帽和广东宝区4、5、6、9、10号铁帽。其中凡区凡口村、铁石岭、富屋背后山铁帽均分布在凡口层控型铅锌矿床的外围，其中Pb、Zn等元素含量很高，推断这三个铁帽可能指示深部有凡口式层控铅锌矿化。广东宝区4、5、6、9、10号铁帽中Cu、Pb、Zn含量较高，且在铁帽体的上方和附近有土壤吸附相态汞异常出现，预测该区可能有层控型铅锌矿化体存在。

第Ⅲ类地层锰帽，分为6、7、8三个小类。进入该类的有湖南和平、洋市、银河三个锰帽。

第Ⅳ类地层铁帽，由于其地球化学特征明显，微量元素普遍很低，则三个已知铁帽都同时进入第9小类。广东凡区羊角山、石塘、筛排岭铁帽和广东宝区7、8号铁帽都归入地层铁帽成因。

地层锰帽和地层铁帽对寻找铅锌矿床，不能提供可靠信息，可以不予考虑。

表3是八近邻表。通过它可以具体了解分类的情况。从近邻表可以看出，岩浆热液型铁帽1、2号与待判样品17、18号分进第1小类，是因为

铁帽分类信息表

表 2

原始分类	样号	铁帽产地 (名称)	新分 类号	对应的 原分类
I 岩浆热液型	1	水口山铁帽	1	I
	2	黄沙坪铁矿	1	
	3	宝山铁帽	2	
	4	大宝山铁帽	2	
II 层控型	5	凡口铁帽	3	II
	6	杨柳塘铁帽	3	
	7	西岗寨铁帽	4	
	8	太坪铁帽	5	
III 地层锰帽	9	六河锰帽	6	III
	10	余田锰帽	7	
	11	银河谢家锰帽	8	
	12	正和锰帽	7	
	13	仁义锰帽	8	
IV 地层铁帽	14	宝区1号铁帽	9	IV
	15	宝区2号铁帽	9	
	16	宝区3号铁帽	9	
待判样品	17	茶山铁帽	1	I
	18	宝山方园铁帽	1	
	19	凡区凡口村铁帽	4	
	20	凡区铁石岭铁帽	3	II
	21	凡区富屋背后山铁帽	3	
	22	宝区4号铁帽	4	
	23	宝区5号铁帽	4	
	24	宝区6号铁帽	4	
	25	宝区10号铁帽	4	
	26	凡区羊角山铁帽	9	IV
	27	凡区石塘铁帽	9	
	28	凡区筛排岭铁锰帽	9	
	29	宝区7号铁帽	9	
	30	宝区8号铁帽	9	
	31	宝区9号铁帽	4	II
	32	和平锰帽	7	III
	33	羊市锰帽	7	
	34	银河锰帽	7	

铁帽八近邻表 表 3

1	2	3	4	5	6	7	8
1	2	18	17	7	19	3	25
2	18	17	1	3	23	9	21
3	4	7	33	23	19	32	11
4	3	33	7	23	19	32	11
5	21	6	20	18	17	1	2
6	21	7	25	20	19	8	5
7	19	25	23	31	24	33	28
8	22	23	19	25	24	7	6
9	11	23	4	13	3	32	7
10	34	32	12	13	11	33	27
11	13	26	28	27	16	30	14
12	13	32	27	28	26	16	29
13	11	27	26	16	28	30	14
14	30	15	16	27	29	26	28
15	14	30	16	27	29	26	28
16	30	15	14	27	29	26	28
17	18	2	1	8	21	5	3
18	17	2	1	21	8	5	7
19	7	25	24	31	23	22	28
20	7	21	26	19	25	31	29
21	5	6	20	1	7	25	19
22	24	19	23	31	25	7	28
23	19	22	24	7	31	25	11
24	31	29	28	25	19	26	27
25	31	19	24	29	28	26	27
26	29	27	16	28	30	15	14
27	16	29	30	28	14	15	26
28	29	27	16	26	30	15	14
29	27	28	16	26	30	15	14
30	14	15	16	27	29	26	28
31	24	25	29	28	26	27	16
32	12	33	10	34	13	11	28
33	32	28	12	13	7	27	19
34	10	32	12	13	11	33	27

这4个样品的近邻中均互相包含,而且它们的共享近邻数都相同。已知地层铁帽14、15、16号的所有近邻完全一样,被同时分为第9小类。再观察进入该类的26、27、28、29、30号的近邻,这几个样不但空间上分布在一起,而且较靠近已知地层铁帽(14、15、16号)。所以,当我们要进一步了解各样点在空间的分布,分类中为什么会出现某种情况,各模式之间的联系到底怎样等问题时,近邻表可以提供这些有用信息。

(二) 超基性岩体含矿性评价 广西九毛锡矿主要产在含凝灰质云母石英片岩和超基性岩中,其中规模与储量都较大的矿体产在超基性岩体的接触带上。该地区已查明的超基性岩体就有60多个,如何快速评价这些岩体的含矿性,对找矿工作是很有用的。

岩体分类的几种计算结果 表4

原始分类	样号	新的分类					
		$k_i = 5$		$k_i = 6$		$k_i = 7$	
		$k_i = 1$	$k_i = 2$	$k_i = 1$	$k_i = 2$	$k_i = 1$	$k_i = 2$
I	1	1	1	1	1	1	1
	2	1	2	1	1	1	1
	3	2	3	2	2	2	2
	4	2	3	2	2	2	2
	5	2	3	2	2	2	2
	6	3	4	1	1	1	1
	7	4	5	3	3	3	3
II	8	5	6	4	4	1	4
	9	6	7	4	5	1	4
	10	5	6	4	4	1	4
	11	5	7	4	5	1	4
	12	6	7	4	5	1	4
	13	5	7	4	5	1	4
待判	14	2	8	2	2	2	2
	15	5	9	4	4	1	4
	16	5	6	4	4	1	4
	17	4	5	3	3	3	3
	18	1	2	1	1	1	1
	19	2	10	2	2	2	2
	20	2	11	2	2	2	2
	21	2	10	2	2	2	2

我们共收集了21个超基性岩体的资料,其中已知含矿岩体样7个,已知无矿岩体样6个,待判岩体8个。根据地化特征挑选出8个有区分意义的元素Sn、Cu、Pb、Zn、As、Ni、In、F进行共享近邻法的计算。 k_i 、 k_j 取不同值的计算结果列于表4。从中可以看出,当 $k_i = 6$, $k_j = 1$ 和 $k_i = 7$, $k_j = 2$ 的分类结果完全一致。而 $k_i = 5$, $k_j = 1$ 与 $k_i = 6$, $k_j = 2$ 虽然在分的小类上有些差异,但都与前两种情况的分类结果相同,即待判样品中只有15、16号判为无矿,而其他样品均判为有矿。当 $k_i = 5$, $k_j = 2$ 时,待判样品不归已知类,并出现新的小类,分类效果不好。根据以上分析,我们只讨论 $k_i = 7$, $k_j = 2$ 的结果(表5)。由表5作分类树状图(图2)。从图2可见,已知含矿超基性岩体分为三类,待判岩体

岩体分类信息表 表5

原始分类	样号	岩体编号	新分类号	对应原分类
已知含矿岩体	1	σ 74-1	1	I
	2	σ 74-2	1	
	3	σ 74-3	2	
	4	σ 74-4	2	
	5	σ 74-5	2	
	6	σ 59-1	1	
	7	σ 60-1	3	
已知无矿岩体	8	σ 东-1	4	II
	9	σ 东-2	4	
	10	σ 东-3	4	
	11	σ 东-4	4	
	12	培地-1	4	
	13	荣塘-1	4	
待判岩体	14	天友寨-1	2	I
	15	天友寨-2	4	II
	16	大平-1	4	
	17	大平-2	3	
	18	大平-3	1	I
	19	σ 38	2	
	20	σ 47	2	
	21	σ 小	2	

中判为含矿岩体的有天友寨—1、大平—2、大平—3、 σ 38、 σ 47和 σ 小岩体。其中大平—2、大平—3岩体的化探次生晕异常很好，而且有重砂异常。而天友寨—1、 σ 38、 σ 47和 σ 小岩体均有较好的化探次生晕异常，可见这些岩体含矿的可能性较大，建议对这几个超基性岩体再进一步工作。

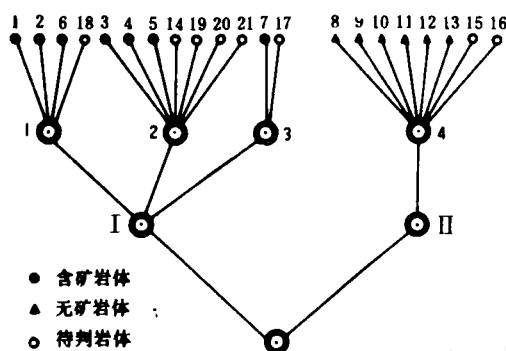


图2 岩体分类树状图

6个已知无矿岩体被分为同一小类，归入无矿的天友寨—2和大平—1两个超基性岩体。

以上两个实例的分类结果表明，共享近邻法不但对已知模型样品分类正确，对未知样品归类清楚，而且能为快速找矿评价提供重要的信息。

共享近邻法与点群分析相比，它有已知模型作为标准，分类结果的好坏可以很直观地从树状图中看出来。所以，当某些问题在使用参数判别方法失效时，可选用共享近邻法，这样能得到较好的分类效果。

本次工作只是对模式识别技术中共享近邻法应用的初步尝试。在实际应用中，还可结合模式识别中其他方法，如势分析法、相似分析法等一起使用。

该方法的程序编制较简单，在微机（如IBM，P C系列等）即可实现，便于在野外队推广使用。我们为PC—1500机编制的程序已通过运行。

An Application of Shared Neighbourhood Method for Pattern Recognition in Geochemical Exploration

Deng Guoying

(Guilin Institute of Mineral Resources and Geology, China National Nonferrous Metal Industry Corporation)

Abstract

The application of shared neighbourhood method for pattern recognition in geochemical exploration is discussed, citing the classification of different genetical gossans in South Hunan North Guangdong district and mineralization appraisal of the ultrabasic intrusive in Jiumao district as examples. Both the simple and multiple discriminant calculation methods were used. Classification results are shown by dendrograms. It has been proved that this method can give a very accurate and distinctive classification for samples of both known and unknown patterns, in order to provide important information for fast evaluation of mineral resources.