Métodos de Reconocimiento de Patrones en la solución de tareas geólogo - geofísicas*

** Ricardo BARANDELA ALONSO

RESUMEN. *Se discuten dos aplicaciones prácticas de los métodos de Reconocimiento de Patrones en el campo de las investigaciones geólogo-geofísicas. Una técnica particular, la así llamada regla NN (Nearest Neighbor rule) y tres de sus variantes más importantes son descritas.*

Se presenta también un sistema de computación: NNINT, desarrollado para trabajos de clasificación con la regla NN en forma interactiva. Este sistema se empleó en las mencionadas aplicaciones.

INTRODUCCIÓN

En los últimos años, las técnicas matemáticas han encontrado creciente empleo en el procesamiento y análisis de los datos de las geociencias. Entre esas técnicas se destacan los métodos de Reconocimiento de Patrones por su utilidad en problemas de clasificación y pronóstico.

El presente trabajo se propone ejemplificar esa utilidad mediante la discusión de dos aplicaciones prácticas en el campo de la exploración de recursos naturales. Se describe la técnica de Reconocimiento de Patrones que se empleó (la regla NN).

También se mencionan brevemente las características principales del sistema interactivo de computación NNINT, empleado en las mencionadas aplicaciones, que son discutidas.

LA REGLA NN. ALGUNAS VARIANTES

La regla NN (así llamada por las siglas de su nombre en inglés: Nearest Neighbor rule) es un método de Reconocimiento de Patrones muy popular entre investigadores y usuarios.

Esta popularidad se debe, entre otras, a las siguientes razones:

1. Es un método no paramétrico, lo que significa que su empleo no depende de funciones de distribución probabilística

* Manuscrito aprobado en julio de 1987.

** Instituto de Geofísica y Astronomía de la Academia de Ciencias de Cuba.

y por tanto no se requiere conocimiento previo al respecto.

2. Se conoce la cota superior de su probabilidad de error (mala clasificación), lo que representa una ventaja sobre la mayoría de los restantes métodos no paramétricos.
3. Su implementación en un programa de computación es muy fácil y expedita.
4. Se dispone de numerosas variantes encaminadas a mejorar su comportamiento o a enfrentar situaciones que ocurren con frecuencia en la práctica. Algunas de estas variantes se discuten en esta misma Sección.
5. Las ideas y conceptos en que se basa resultan muy intuitivas y de rápida comprensión como puede verse en la siguiente descripción.

Como ocurre en todos los métodos supervisados, la regla NN se apoya en la información suministrada por un conjunto de patrones de entrenamiento (prototipos) que representan a todas las clases de interés y que se supone estén perfectamente identificados. Este conjunto de prototipos recibe el nombre de Muestra de Entrenamiento (ME).

Cuando se desea clasificar un patrón desconocido X según la regla NN, se busca su vecino más cercano en la ME. En otras palabras, se determina el prototipo que minimiza una función de distancia predefinida, con respecto a X. Entonces X se asigna a la clase representada por su vecino más cercano.

Entre las ya mencionadas variantes de que se dispone en conexión con la regla NN, las tres que se discuten brevemente a continuación resultan de gran importancia y fueron empleadas en las aplicaciones que se presentan en este trabajo. Las tres consisten en preprocesamientos a los que debe ser sometida la ME antes de iniciar la etapa de clasificación de nuevos patrones.

a) Edición (Wilson, 1972). Aunque se incluye generalmente entre los métodos encaminados a reducir el tamaño de la ME, pues tiene también esa propiedad, su verdadero objetivo estriba en mejorar el comportamiento del clasificador mediante la reducción de su probabilidad de clasificación errónea.

El procedimiento consta de dos etapas:

1. Para todo x_i en la ME se buscan sus k vecinos más cercanos en el resto de la ME, y se determina cuál es la clase más representada entre esos k vecinos.
2. Se elimina x_i de la ME si su identificación original no coincide con la clase determinada en el paso anterior.

El valor del parámetro k debe ser fijado de antemano.

b) Edición Generalizada (Koplowitz y Brown, 1978). Originalmente diseñado también para mejorar el rendimiento del clasificador, se ha podido comprobar (Barandela, en prensa) que este procedimiento brinda magníficos resultados cuando se utiliza en aquellos casos en los que no se cumple el supuesto de que la identificación de todos los prototipos sea correcta.

Estas situaciones ocurren con cierta frecuencia en las aplicaciones prácticas, pues en algunas de ellas la identificación de los prototipos es una tarea costosa y difícil. Se han reportado en diagnóstico médico, clasificación de cultivos por medio de fotos aéreas, elaboración de mapas pronóstico de depósitos minerales y otras aplicaciones.

Con este procedimiento el parámetro k, como en el anterior, debe ser definido pre-

viamente. Entonces se determina otro parámetro k', tal que:

$$(k + 1)/2 \leqslant k' \leqslant k$$

Para cada prototipo x_i se buscan sus k vecinos más cercanos en el resto de la ME. Si una clase tiene al menos k' representantes entre esos k vecinos, entonces x_i se identifica con esa clase con independencia de su identificación original.

En otro caso, x_i se elimina de la ME.

Barandela (1986) propuso el empleo reiterado de este procedimiento y mostró, con experimentos simulados, que sólo con dos repeticiones se obtenían mejoras sustanciales.

c) Subconjunto Selectivo Modificado (SSM) (Barandela, en prensa). La regla NN adolece de una desventaja práctica importante: para clasificar cada patrón desconocido es necesario examinar todos los prototipos. Cuando la ME tiene un tamaño considerable, esto plantea requerimientos excesivos en capacidad de memoria y tiempo de computación.

Por eso se han publicado numerosas variantes encaminadas a buscar una reducción del tamaño de la ME, de manera que se pierda lo menos posible de la información suministrada por la ME original.

La presente variante tiene varias ventajas sobre las publicadas previamente: no depende del orden en que se examinan los prototipos, conserva aquéllos que brindan mayor información discriminatoria, y está definida de manera formal, no intuitiva.

Una descripción completa del algoritmo escapa al alcance de este trabajo. Las ventajas arriba mencionadas han sido confirmadas por resultados de experimentos de Montecarlo, comparativos con otras variantes publicadas.

SISTEMA DE COMPUTACIÓN NNINT PARA EL RECONOCIMIENTO DE PATRONES EN FORMA INTERACTIVA

El valor que debe asignarse a los parámetros k y k' arriba mencionados, la cantidad conveniente de repeticiones de la Edición Generalizada, si la ME se reduce o no, y otros muchos aspectos (por ejemplo, la selección de variables), que forman parte de un proceso de Reconocimiento de Patrones, son decisiones que dependen fuertemente de las características del problema particular que se desea resolver y resulta muy difícil determinarlas de antemano.

Lo anterior permite afirmar que la mejor forma de llevar a cabo la tarea de Reconocimiento de Patrones es mediante un procedimiento interactivo que facilite probar distintas alternativas hasta encontrar la que mejor se ajuste al problema en cuestión.

En el Instituto de Geofísica y Astronomía se ha implementado el sistema de computación NNINT (Barandela y Fuentes, 1986) para la aplicación de la regla NN en forma interactiva con microcomputadoras del tipo NEC. Las principales características del sistema NNINT son:

1. Ha sido diseñado para lograr una comunicación usuario-máquina que resulte fácil y cómoda, incluso para aquellos sin experiencia previa en el empleo de computadoras.

2. Se ha estructurado jerárquicamente, en forma de submenús y módulos. Esto permite:
   a) Segmentación del sistema con el propósito de ahorrar memoria interna.
   b) Brindar al usuario una visión clara y completa de todas las posibilidades disponibles.
   c) Dar al sistema flexibilidad para añadir fácilmente cualquier nuevo algoritmo deseado por el usuario.
3. Se ha previsto la posibilidad de errores por parte del usuario al introducir información. El sistema detecta tales errores, hasta donde esto es posible (por ejemplo, el número de una clase que no existe), alerta al usuario acerca de la información incorrecta y repite la pregunta que dio origen al error. Todo esto sin tener que comenzar de nuevo desde el principio y sin alterar la información ya procesada.
   Por supuesto, algunos errores no pueden ser detectados por el sistema y ocasionarían resultados incorrectos. En estos casos, al presionar la tecla "Stop" se puede interrumpir el proceso y regresar al punto donde se cometió el error.
4. Es posible utilizar también el sistema cuando alguna o todas las variables son del tipo cualitativo, es decir, cuando las mediciones son efectuadas sobre una escala discreta o nominal.

El sistema ofrece los siguientes submenús: Preprocesamiento, Reducción de la ME, Depuración, Estimación del error de clasificación, Selección de variables, Gráficos, Clasificación y Organizativas.

En Clasificación se puede optar por la regla NN (descrita arriba), por una extensión de ésta, la así llamada regla k-NN, o por una variante que permite al clasificador no tomar decisión alguna (rechazo) sobre aquellos patrones cuya identificación es altamente dudosa a la luz de la información disponible.

Dos opciones del submenú Organizativas desempeñan un importante papel en la consecución del objetivo principal del sistema:

— Grabar, para conservar en el disco flexible, cada vez que se desee, el estado vigente de la ME, después de haber sufrido modificaciones por algunos de los restantes algoritmos.
— Reiniciar, para comenzar de nuevo con la ME original, anulando todas las modificaciones previamente efectuadas.

Con estas dos subrutinas es posible probar diferentes combinaciones de las opciones deseadas para buscar una mejor configuración de la ME, conservar la resultante para un posible uso futuro, y recomenzar de nuevo con la ME original para ensayar nuevas combinaciones, hasta que se arribe a una decisión final. Las estimaciones del error de clasificación pueden ser de gran ayuda en esta selección.

DOS APLICACIONES PRÁCTICAS

Ambas aplicaciones fueron desarrolladas en el campo de las investigaciones geólogo-geofísicas, en relación con la prospección de yacimientos gaso-petrolíferos. Los datos fueron amablemente suministrados por el Departamento de Geofísica del ISPJAE.

Primera aplicación

La ME original consistió, en este caso, en 268 patrones no identificados, correspondientes a capas (estratos) de registros de pozos. Seis variables fueron medidas:

1. Cociente potencial espontáneo/Intensidad gamma.
2. Resistividad aparente obtenida con sonda potencial.
3. Resistividad aparente obtenida con sonda gradiente.
4. Intensidad gamma.
5. Porosidad Neotrónica.
6. Variación del diámetro de pozo.

Previamente, estos patrones fueron divididos en 4 clases mediante un algoritmo de agrupación ejecutado en el ISPJAE. Esta configuración fue tomada como punto inicial para la presente aplicación. Puesto que el propósito era explorar la posibilidad de manifestaciones gaso-petrolíferas, aunque al ME comprendía 4 clases, se consideró de la manera siguiente:

— clases 1 y 3, conformadas por estratos sin perspectivas;
— clases 2 y 4, fuertemente asociadas con los estratos con perspectivas.

La Tabla 1 muestra los resultados del trabajo. Para cada algoritmo de los que produjeron mejoría en la ME, se reflejan la cantidad de prototipos resultantes y los estimados del error de clasificación según dos de los métodos incluidos en NNINT: método C y método L (Barandela, 1981).

Como los resultados de un algoritmo de agrupación (cluster) no pueden ser considerados como una identificación enteramente confiable, se comenzó por aplicar la Edición Generalizada con parámetros $k = 5$ que la ME comprendía 4 clases, se consien 50% y el estimado del error en 67%. Una segunda aplicación ($k = 3$, $k' = 2$) de esta técnica, mostró los beneficios que reporta esta reiteración. El tamaño de la ME permaneció inalterable y sin embargo el error estimado disminuyó en más de 85%. Una tercera aplicación brindó resultado no aceptable y fue desechada.

Se evaluó la importancia informativa de cada variable. La eliminación de la variable No. 4 permitió una reducción del tiempo de computación subsiguiente y no produjo deterioro en el clasificador. Se aplicó entonces la Edición de Wilson. En total se logró una configuración de la ME con un tamaño menor que 50% de la original, lo que significa considerable ahorro de tiempo de computación. Además, el estimado del error de clasificación resultante fue prácticamente nulo.

Esto último fue confirmado cuando esta ME se utilizó para clasificar, con la regla NN, otros 628 estratos procedentes de 3 registros de pozos diferentes, cosa que se logró con resultados altamente satisfactorios.

TABLA 1. *Resultados de la primera aplicación.*

Secuencia Empleada	Cantidad de prototipos	Est. del error (%) C	L
ME inicial	268	47,4	34,0
Edición Generalizada	130	15,4	6,2
Repetición (de la E.G.)	130	2,3	1,5
Eliminar variable 4	130	2,3	1,5
Edición	122	0,8	0,0

Con posterioridad se utilizó el método SSM para reducir aún más el tamaño de la ME. Permanecieron sólo 16 prototipos, pero este resultado no se incluyó en la Tabla 1 porque el estimado del error aumentó considerablemente. Sin embargo, cuando esta ME reducida se empleó para clasificar los mencionados 628 patrones independientes, solamente 12 de ellos recibieron una asignación diferente a la obtenida anteriormente (es decir, cuando la ME consistía de 128 prototipos).

Resulta interesante destacar que en todos esos 12 casos, la diferencia consistió en que primeramente (con la ME de tamaño 128) estos se consideraron como secciones de pozos con perspectivas y en la segunda ocasión no. Es decir, ahora se obtuvo una clasificación más conservadora. De todas maneras, 12 prototipos representan menos de 2% de clasificación errónea, mientras que el tiempo de computación requerido disminuyó en más de 85 %. Según las características del problema que se esté analizando, el usuario puede decidir sobre la alternativa más conveniente.

Segunda aplicación

En este caso, el conjunto original de prototipos consistió en 124 estratos pertenecientes a diferentes registros de pozos, pero todos ellos con conocidas manifestaciones gaso-petrolíferas. El objetivo fue clasificar estos patrones de acuerdo con su producción estándar. La información disponible sobre la producción no se utilizó en el trabajo de clasificación (sólo se emplearon mediciones geólogo-geofísicas) y se dejó para emplearse al final como medio de control.

Aquí se tuvieron en cuenta 5 variables:

—Potencial espontáneo.

—Parámetro Duplo Diferencial de la intensidad gamma.

—Parámetro Duplo de la intensidad neutrón-gamma.

—Factor de formación aparente de la zona invadida.

—Factor de formación aparente de la zona virgen.

Primeramente, los prototipos fueron agrupados en dos clases (Grandes y Pequeños colectores) mediante la aplicación de un algoritmo de agrupación (cluster). Entonces se aplicó la Edición Generalizada tres veces (siempre con $k = 5$ y $k' = 4$) y a continuación la Edición de Wilson.

De esta forma el tamaño de la ME disminuyó a 84 prototipos y el estimado del error (según el método C) decreció desde 25% hasta 0%. Desde el punto de vista de la estructura de la muestra se observó también una notable mejoría. La diferencia entre las dos clases se hizo más clara, como se desprende de los valores medios del nivel de producción:

	Antes del proced.	*Después del proced.*
Grupo 1 (grandes colect.)	8,5	10,3
Grupo 2 (pequeños colect.)	6,0	2,8

La cantidad de prototipos colocados en el grupo incorrecto (grandes colectores entre los pequeños o viceversa) disminuyó en 50%. En resumen, se obtuvo una mejoría sustancial de la muestra para utilizarla como referencia en futuros trabajos de pronóstico.

CONCLUSIONES

1. Las dos aplicaciones discutidas constituyen un ejemplo de lo válido que resulta la aplicación de los métodos de Reconocimiento de Patrones, en particular la regla NN, para solucionar importantes problemas geólogo-geofísicos.

2. Es frecuente, en la literatura especializada, recomendar la normalización o tipificación de los datos, cuando se trabaja con clasificadores del tipo de la regla NN. Sin embargo, en los dos casos analizados, cuando se intentó preprocesar la información de esa manera, el comportamiento del clasificador empeoró. Esta cuestión, al igual que la de la selección de variables, requiere estudio adicional.

3. Muy interesante resulta el empleo de la Edición Generalizada como complemento de un algoritmo de agrupación (cluster), metodología no reportada previamente.

REFERENCIAS

Barandela, R. (1981): Estudio comparativo de los métodos empíricos de estimación de la probabilidad de error. *Investigación Operacional*, 1:114-126.

———— (1986): Nuevas variantes para el procesamiento de muestras de entrenamiento imperfectamente identificadas. En *Segundo Congreso Nacional de Matemática* La Habana.

———— [en prensa]: "Métodos para la reducción de la muestra de entrenamiento", pendiente de publicación en la Revista Estructura, ISPJAE.

———— [en prensa]: "Imperfectly labeled training samples in the context of Nearest Neighbor rule", pendiente de publicación en la Revista System Analysis and Modeling Simulation, RDA.

Barandela, R. y N. Fuentes (1986): Un sistema interactivo para el reconocimiento de patrones con la regla NN. En *Quinta Conferencia Científica del ISPAJE.*

Koplowitz, J. y T. A. Brown (1978): On the relation of performance to editing in Nearest Neighbor rules. En *Proc. of the 4th. Int. Joint. Conf. on Pattern Recognition,* Japón.

Wilson, D. L. (1972): Asymptotic properties of Nearest Neighbor rules using edited data. *IEEE Trans. Syst., Man and Cyb.*, SMC, 2:408-421.

Ciencias de la Tierra y del Espacio, 17, 1990

PATTERN RECOGNITION METHODS IN THE SOLUTION OF GEOLOGO-GEOPHYSICAL INVESTIGATIONS

Ricardo BARANDELA ALONSO

ABSTRACT. *Two practical applications of Pattern Recognition methods in the realm of geologo-geophysical investigations are discussed. A particular technique: the so called Nearest Neighbor (NN) rule and three of its more important variants are described.*
A computing system: NNINT, developed for classifying with the NN rule in an interactive setting and employed through these applications is also presented.