Caso de estudio: pozos de petróleo

Luis Carlos Molina Félix
Ramon Sangüesa i Solé

PID_00165735

Universitat Oberta de Catalunya
www.uoc.edu

Índice

Introducción

Este módulo presenta un caso de estudio que pretende que el estudiante se acerque a la realidad de *data mining* a partir de una situación real.

El caso presentado en este módulo se basa en una situación real. Los datos que se dan han sido extraídos de un caso concreto y adaptados para que puedan presentarse como ejemplo válido de las técnicas que utiliza *data mining*.

Objetivos

Los materiales didácticos asociados a este módulo permitirán al estudiante alcanzar los objetivos siguientes:

1. Poner al estudiante en contacto con una situación real en la que la *data mining* es una metodología potente que permite realizar un análisis detallado de un problema y extraer conclusiones al respecto.

2. Permitir al estudiante conocer las dificultades que se plantean con las técnicas de *data mining* y las ventajas que presentan dichas técnicas.

1. El problema de los pozos de petróleo

La permeabilidad de un pozo es un factor importante a la hora de decidir la conveniencia o inconveniencia y la dificultad de perforar un pozo petrolífero, y para tener una idea de sus posibles costes de mantenimiento. El hecho de conocer la permeabilidad de un pozo es, pues, de gran relevancia. Se trata de un problema de predicción. Uno de los atributos importantes que hay que considerar es la porosidad del terreno, que ofrece una idea de la permeabilidad de los pozos vecinos y la profundidad.

Presentaremos cómo se afrontó el problema. En concreto, nos fijaremos en los aspectos siguientes:

a) Cómo se eligieron los datos que se debían considerar como factores predictivos.

b) Las técnicas de preparación de datos utilizados.

c) Los diferentes métodos aplicados y los modelos resultantes.

d) La evaluación y comparación de los diferentes resultados.

e) La comparación con el conocimiento de los expertos sobre el mismo tema.

El caso tiene como características más interesantes la influencia de las técnicas de preparación de datos en el resultado final y la importancia relativa del conocimiento aportado por los expertos en la cuestión.

1.1. ¿Cuál es el problema exactamente?

El trabajo realizado se basa en métodos desarrollados y datos recogidos por S.J. Rogers, H.C. Chen, D.C. Kopaska-Merkeli y J.H. Fang publicados en *Predicting Permeability from Porosity Using Artificial Networks*. En este trabajo, se utilizaron redes neuronales para predecir la permeabilidad de un pozo petrolífero en función de su porosidad y su profundidad. El conjunto de datos procede de medidas tomadas en seis pozos petrolíferos en Big Escambia Creek, Alabama, en Estados Unidos. A partir de estos datos y de consideraciones geológicas aportadas por expertos en la materia se prepararon tres escenarios de prueba diferentes:

Lectura complementaria

Encontraréis los datos del problema considerado en este apartado en la obra siguiente:

S.J. Rogers; H.C. Chen; D.C. Kopaska-Merkel; J.H. Fang (1995). "Predicting Permeability from Porosity Using Artificial Neural Networks". *American Association of Petroleum Geologists Bulletin* (vol. 79, diciembre, pág. 1.786-1.797).

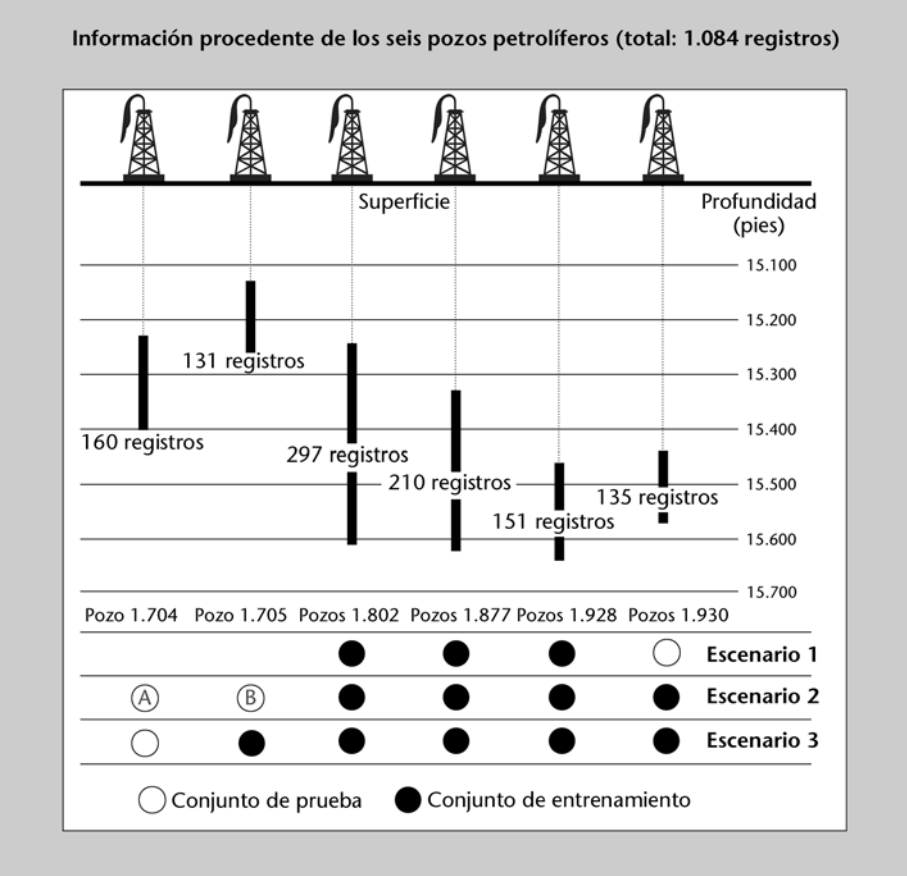
1) En el primer escenario se utilizaron los datos procedentes de los pozos 1.877 y 1.928 como conjunto de entrenamiento; los del pozo 1.802, como conjunto de validación, y los del pozo 1.930, como pozo para predecir.

2) En el segundo escenario, el conjunto de entrenamiento estaba formado por los pozos 1.802, 1.877 y 1.930; el conjunto de validación estaba formado por los datos procedentes del pozo 1.928 y los pozos para predecir eran el 1.704 y el 1.705.

3) Finalmente, en el tercer escenario los pozos 1.928, 1.877 y 1.930 aportaban los datos del conjunto de entrenamiento, el pozo 1.705 contribuía con sus datos al conjunto de validación y el pozo 1.704, como conjunto de prueba, era el pozo sobre el que se deseaba realizar la predicción.

Una de las características que dificultan este problema en gran medida es la ausencia de datos, tanto en lo que concierne a la porosidad y la permeabilidad como en cuanto a la falta de secuencia entre las profundidades.

En la figura siguiente podemos ver las profundidades relativas de los seis pozos (en pies), el número de casos que corresponden a cada una de las profundidades, y también los tres escenarios considerados:



Compararemos dos tipos de modelos:

- Los obtenidos mediante la traducción a reglas del método C4.5-*rules*, que inicialmente construye árboles de decisión.

- Los obtenidos con el algoritmo de inducción de reglas de clasificación CN2.

2. Metodología

A continuación, precisaremos el tipo de herramientas utilizadas en este caso y los distintos pasos que se han llevado a cabo.

2.1. Herramientas y métodos

Para las tareas de análisis previo se recurrió a las herramientas STATISTICA™, versión 5.0, de StatSoft Inc. y Statistics Visualizer™, de Silicon Graphics. El Scatter Visualizer™ de Silicon Graphics se utilizó para efectuar el análisis multidimensional de la conducta de los datos procedentes de los seis pozos.

La discretización se realizó con el *software* MineSet™, versión 2.01, una herramienta de la empresa Silicon Graphics (Silicon Graphics, 1998) que conjuga la *data mining* con herramientas de visualización multidimensional y ofrece al mismo tiempo algunas herramientas de discretización y de análisis de datos para encontrar, por ejemplo, clasificaciones y asociaciones entre elementos de la base de datos.

Como hemos dicho, se utilizaron dos sistemas de inducción de reglas: CN2 (Clark & Niblett, 1989) y C4.5-*rules*. Con su ayuda, se pudieron determinar y comparar las tasas de error que correspondían a varios tipos de discretización para el atributo clase *Permeabilidad*.

Lecturas complementarias

Con relación al algoritmo C4.5 podéis consultar las obras siguientes:

J.R. Quinlan (1987). "Generating Production Rules from Decision Trees". En: *Proceedings of 4th International Machine Learning Workshop* (págs. 304-307). San Mateo: Morgan Kaufmann.

J.R. Quinlan (1993). *C4.5 Programs for Machine Learning*. San Mateo: Morgan Kaufmann Publishers.

1) El **método CN2** es un algoritmo no incremental que toma un conjunto de ejemplos y genera reglas del tipo "si... entonces..." para clasificar los ejemplos.

2) El **método C4.5-*rules*** es un generador de árboles de decisión y la C4.5-*rules* crea reglas del tipo "si... entonces..." a partir del árbol de decisión resultante.

Ambos algoritmos fueron usados utilizando la librería MLC++, *Machine Learning Library in C++* (Kohavi y otros, 1994), un paquete de *software* desarrollado en la Universidad de Stanford que contiene varios algoritmos de aprendizaje automático y que permite mantener el formato de los datos de entrada cuando se cambia de algoritmo. Asimismo, ofrece métodos normalizados para llevar a cabo experimentación con los diferentes modelos obtenidos.

2.2. Fases del proyecto

En este subapartado presentaremos con detalle las fases del proyecto tal como las hemos ido explicando a lo largo de toda la asignatura.

2.2.1. Análisis preliminar y preparación de los datos

El dominio se describe por medio de los atributos siguientes:

- **Permeabilidad**. Se define como la facilidad con la que un gas o un líquido atraviesa un material a través de sus poros cuando está sometido a presión (medida en darcios).

- **Porosidad**. Es la propiedad que tiene un material de contener poro o intersticios (grietas que separan las moléculas de un sólido). Se define como la relación entre el volumen de intersticios y el volumen de la masa del material, y depende del número, la forma y la distribución de los espacios vacíos. Se expresa en porcentaje.

- **Profundidad**. Expresada en pies, se refiere al nivel de perforación alcanzado en un pozo.

Los datos originales sobre profundidad, porosidad y permeabilidad aparecen descritos en la tabla –que vemos en el subapartado siguiente– en forma de atributos continuos.

2.2.2. Selección y limpieza de datos

La tabla siguiente incluye los datos originales:

Datos utilizables del conjunto inicial			
	Profundidad	**Porosidad**	**Permeabilidad**
Total de registros	1.084	1.084	1.084
Valores ausentes	0	96	96
Valores perdidos	0	6	6
Valores conocidos	1.084	982	982

De este conjunto de datos, primero se eliminaron noventa y seis casos en los que faltaban valores tanto para la permeabilidad como para la porosidad. También se eliminaron seis casos considerados como "perdidos", que correspondían a situaciones en las que el instrumento de medida había perdido los valores de porosidad y permeabilidad para una profundidad determinada. Finalmente quedó un total de 982 casos para efectuar el primer análisis.

2.2.3. Análisis previo de datos

Para ver lo que indican los datos y tener una mejor comprensión del dominio, se realizó una primera fase de estadística descriptiva; se encontraron los pará-

metros siguientes: la media, la desviación estándar, la moda, y los valores mínimo y máximo.

En la tabla siguiente presentamos los resultados obtenidos con STATISTICA™ sobre el total de los 1.084 registros de datos:

Distribución de los valores nulos e inferiores a 0,01 para la permeabilidad				
Pozo	**Casos**	**Valor 0**	**Valores < 0,01**	**%**
1.704	159	66	0	41,5
1.705	129	0	78	60,4
1.802	249	121	0	48,6
1.877	182	0	58	31,8
1.928	130	0	38	29,2
1.930	133	0	15	11,3
Total	982	187	189	38,3

Se observó que los datos originales mostraban bastantes valores de permeabilidad menores de 0,01 y que muchos otros valores eran igual a cero.

Los valores nulos e inferiores a 0,01 se sustituyeron por el valor 0,009, de modo que hay que interpretar los resultados que se obtienen tanto con CN2 como con C4.5-*rules* en los que aparece el valor *Permeabilidad* = 0,009 como *Permeabilidad* < 0,01.

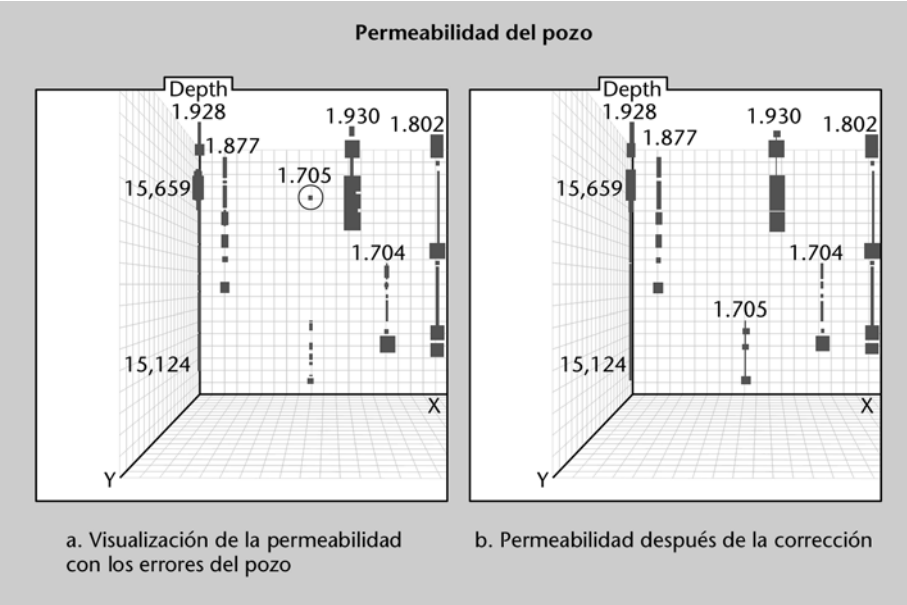
2.2.4. Visualización de datos

La posibilidad de visualizar los datos previamente facilita una mejor comprensión del dominio. Quizá es tanto o más importante disponer de un conocimiento previo que nos permita saber qué tipo de herramienta de visualización hay que aplicar para obtener el tipo de gráfico que retorne mejor información.

Utilizamos Scatter Visualizer® a fin de obtener una primera visualización de los datos. Los cuadrados de los gráficos siguientes representan los valores de los atributos. Un cuadrado pequeño indica que el atributo correspondiente tiene un valor pequeño y un cuadrado grande, lo contrario. En los gráficos **a** y **b** de la figura 2 podemos ver los gráficos que corresponden al análisis de la permeabilidad y la porosidad en relación con la profundidad. En los gráficos **a** y **b** de la figura 2 podemos apreciar la clara ausencia de datos que muestra el pozo 1.705.

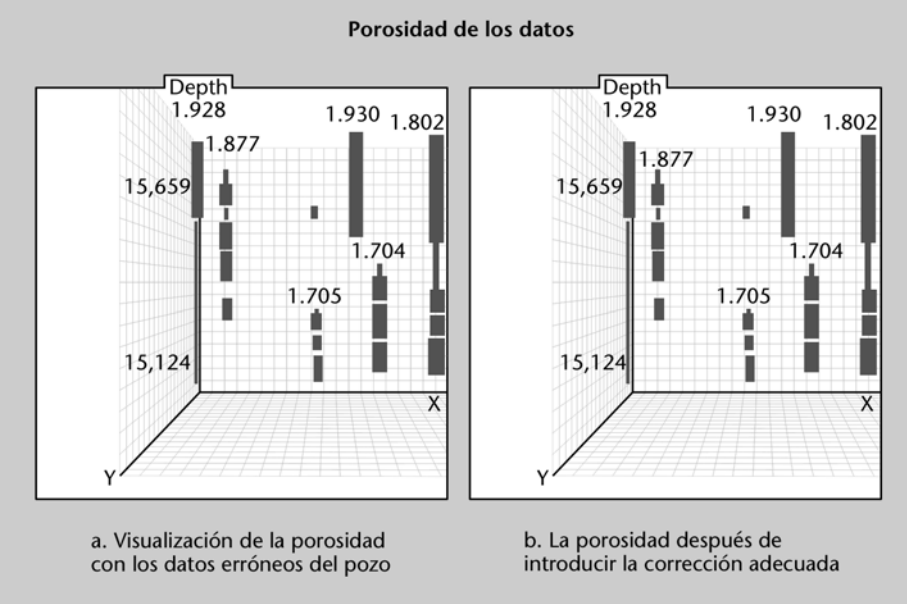
Volvimos a analizar los datos originales y pudimos concluir que seguramente se había introducido un error a causa de la secuencia de valores que mostraba y de los valores que tenían otros atributos. Observad que después de una pro-

fundidad de 1.520 pies, el valor siguiente que aparece es de 1.521 pies y a continuación se produce un salto de 1.541 y 1.542:



a. Visualización de la permeabilidad con los errores del pozo

b. Permeabilidad después de la corrección



a. Visualización de la porosidad con los datos erróneos del pozo

b. La porosidad después de introducir la corrección adecuada

Si analizamos los gráficos que aparecen en las dos figuras anteriores, podremos observar una gran entropía entre los valores de la permeabilidad y los de la porosidad, que presentan un comportamiento más homogéneo.

Asimismo, podemos observar que el pozo 1.930 tiene una porosidad y unas permeabilidades más altas que los otros pozos. Es importante destacar que los pozos 1.705 y 1.802 poseen valores de permeabilidad más bajos, pero mucha más porosidad con respecto a los otros pozos.

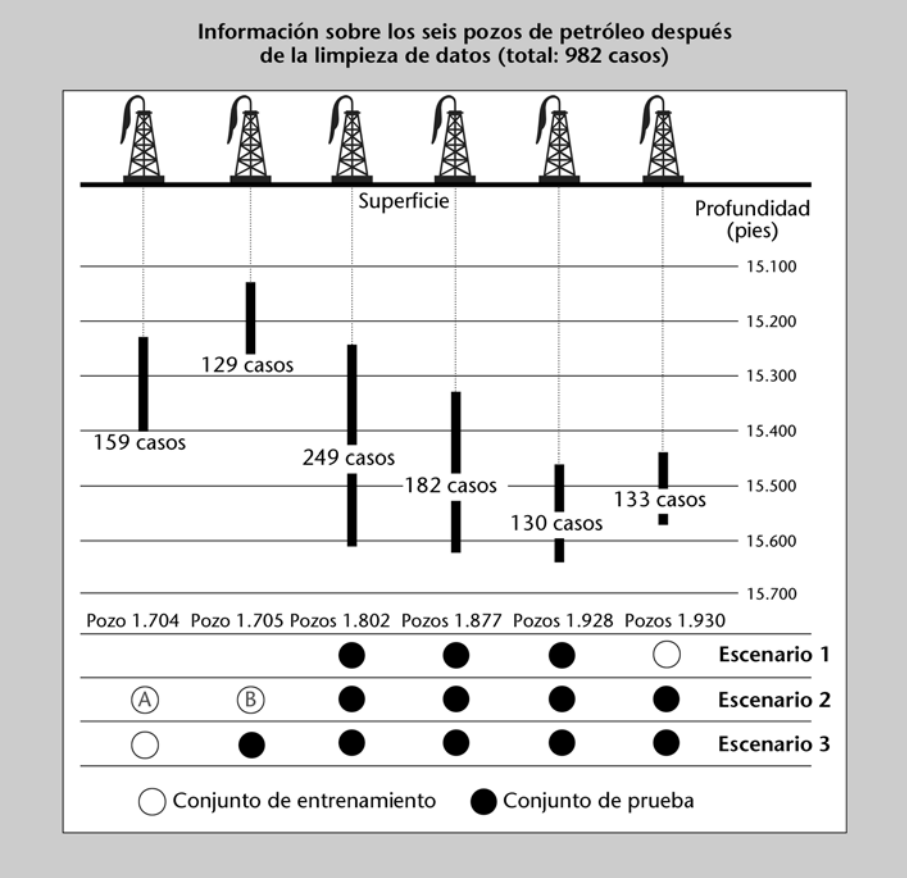
Una vez efectuada la limpieza y la corrección de datos, se procedió a realizar un análisis estadístico de los datos (media, mediana, desviación estándar, histogramas y cuartiles) que se muestra en forma gráfica en la tabla que presentamos a continuación:

Detección y corrección de errores en los datos del pozo 1.705					
Datos originales			**Datos corregidos**		
Profundidad	**Porosidad**	**Permeabilidad**	**Profundidad**	**Porosidad**	**Permeabilidad**
15.217	2	< 0,01	15.217	2	0,009
15.218	0,7	< 0,01	15.218	0,7	0,009
15.219	0,7	< 0,01	15.219	0,7	0,009
15.220	0,6	< 0,01	15.220	0,6	0,009
15.521	0,7	< 0,01	15.221	0,7	0,009
15.522	0,7	< 0,01	15.222	0,7	0,009
15.523	1,3	< 0,01	15.223	1,3	0,009
15.524	1,1	< 0,01	15.224	1,1	0,009
...	...	...	...	...	...
...	...	...	...	...	...
15.538	9,4	0,9	15.238	9,4	0,9
15.539	9,8	0,8	15.239	9,8	0,8
15.540	7,8	0,05	15.240	7,8	0,05
15.541	0,8	< 0,01	15.241	0,8	0,009
15.242	7,1	0,57	15.242	7,1	0,57
15.243	8,3	0,05	15.243	8,3	0,05
15.244	15,1	0,34	15.244	15,1	0,34
15.245	11,7	0,29	15.245	11,7	0,29

En la figura siguiente mostramos los resultados de aplicar estadística descriptiva con Statistics Visualizer™ a 982 casos:

Descripción estadística de los seis pozos (total: 982 casos)			
	Profundidad	**Porosidad**	**Permeabilidad**
Valores diferentes	500	213	142
Valores [Máx., Mín.]	[15.568, 15.124]	[15,1, 0,6]	[0,57, 0,009]
Media	15.415 ± 127,9	8,2 ± 6,5	2,5 ± 6,7
Mediana	15.439	8,2	0,05

Sólo se presentan los resultados correspondientes a los 982 casos resultantes de la selección y limpieza de datos.



2.2.5. Preparación de datos: proceso de discretización

Una vez realizado el análisis inicial de datos y las correcciones de datos mencionado, se procedió a discretizar el atributo de clase *Permeabilidad*. De esta manera, ya estábamos en condiciones de aplicar los algoritmos CN2 y C4.5- *rules*.

Utilizamos tres de los métodos de discretización que ofrece MineSet™:

1) Discretización *stand-alone* del atributo de clase *Permeabilidad* basada en la frecuencia de distribución.

2) Un método de discretización empírica sugerido por un petrolero experto que discretiza el atributo de clase *Permeabilidad* en función de los valores de la porosidad.

3) Un método híbrido que **intenta** mejorar la discretización sugerida por el experto con respecto a la precisión de los algoritmos CN2 y C4.5-*rules*.

Primer método de discretización

El número de intervalos generados para el atributo de clase *Permeabilidad* se especificó de entrada, sin considerar los otros atributos (porosidad y profundidad), y se trató de obtener una buena distribución de los datos entre todos los intervalos. Para obtener la discretización, se utilizaron datos de los seis pozos de petróleo (982 casos) con la intención de poder realizar las comparaciones correctas para los tres escenarios de prueba. Se obtuvieron tres discretizaciones con dos, tres y cinco grupos.

Discretización sugerida por el experto

Según el experto, cada terreno donde se efectúa una perforación es diferente, con lo que resulta muy difícil determinar cuáles son los mejores intervalos de discretización para cada uno de los tres escenarios. El experto que aporta la mejor discretización es aquel que conoce muy bien una clase particular de terreno, puesto que los valores de permeabilidad son diferentes en cada caso.

Según el experto, también es importante saber que en la mayoría de los casos una porosidad alta en las rocas es indicativa de la existencia de petróleo. Por tanto, según el experto, para establecer los criterios de discretización es conveniente considerar la porosidad del pozo. Aunque el experto no determinó los valores de discretización, sugirió discretizar la permeabilidad utilizando el valor correspondiente de la porosidad como factor de ponderación. Mineset™ ofrece la posibilidad de discretizar un atributo en función del peso de otro atributo. Los resultados de las diferentes discretizaciones obtenidas con la utilización de un método *stand-alone* con pesos uniformes y el que sugería el experto se muestran en la tabla siguiente:

Intervalos de discretización obtenidos por dos métodos diferentes

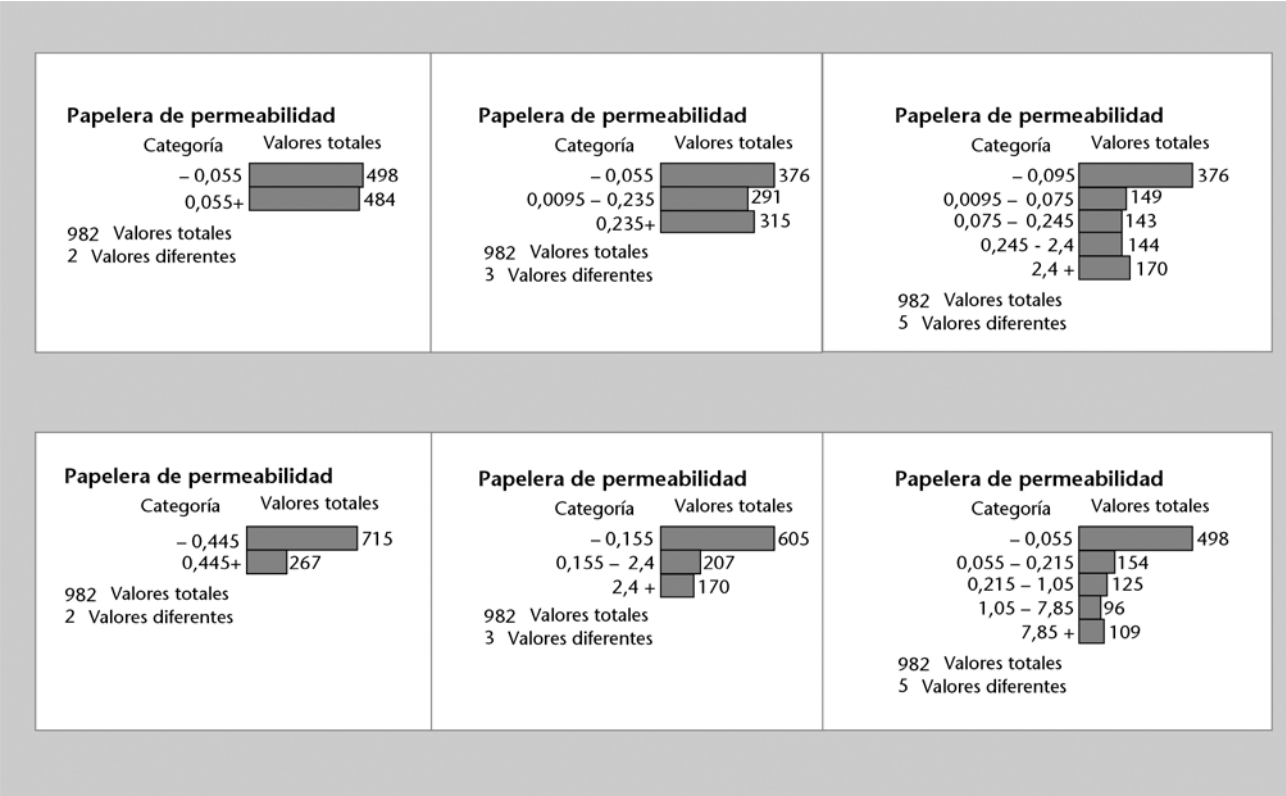
	Discretización *Stand-alone*	Discretización sugerida por el experto
2 intervalos	< 0,055 > 0,055	< 0,445 > 0,445
3 intervalos	< 0,0095 [0,0095, 0,235] > 0,235	< 0,155 [0,155, 2,4] > 2,4
5 intervalos	< 0,0095 [0,0095, 0,075] [0,075, 0,245] [0,245, 2,4] > 0,235	< 0,055 [0,055, 0,215] [0,215, 1,05] [1,05, 7,85] > 7,85

Lectura complementaria

Podéis consultar otros experimentos parecidos en la siguiente obra:

J. Dougherty; R. Kohavi; M. Sahami (1995). "Supervised and Unsupervised Discretizations of Continuous Features". En: *Proceedings of the 12th International Conference on Machine Learning* (pág. 194-202). Morgan Kaufmann Publishers.

En la figura siguiente presentamos las diferentes distribuciones de los valores de discretización según los dos métodos comentados. El algoritmo de discretización utilizado fue el que se basa en la entropía porque había presentado mejores resultados en otros experimentos parecidos:



Realizaremos una discusión más profunda de los resultados obtenidos con estas discretizaciones más adelante.

También se aplicaron otros criterios de discretización. En concreto, se aplicó un algoritmo automático que intenta encontrar el número de intervalos. El resultado fue la propuesta de doce intervalos discretos que generaban una gran cantidad de errores al aplicar los algoritmos CN2 y C4.5-*rules*.

2.2.6. *Data mining*: modelos de clasificación

Se utilizaron los métodos CN2 y C4.5-*rules* para determinar qué tasas de errores se obtenían con las diferentes discretizaciones propuestas para el atributo de clase *Permeabilidad*. Ambos algoritmos se implementaron con MLC++, *Machine Learning Library in C++* (Kohavi y otros, 1994). En ambos casos, los algoritmos se ejecutaron con los parámetros por defecto de la librería MLC++. Las tasas de error para cada algoritmo y escenario, así como el error de clasificación del conjunto de entrenamiento se muestran en las tres tablas que veremos más abajo.

Lecturas complementarias

Podéis consultar los métodos CN2 y C4.5-*rules*, respectivamente, en las obras siguientes:

P. Clark; T. Niblett (1989). "The CN2 Induction Algorithm". *Machine Learning* (vol. 3, pág. 261-283).

J. Dougherty; R. Kohavi; M. Sahami (1995). "Supervised and Unsupervised Discretizations of Continuous Features". En: *Proceedings of the 12th International Conference on Machine Learning* (pág. 194-202). Morgan Kaufmann Publishers.

De los resultados podemos inferir que sólo en un pequeño conjunto de casos la tasa de error es relativamente baja para ambos métodos (CN2 y C4.5-*rules*). Asimismo, podemos ver que hay una gran variabilidad en las tasas de error. Tanto en el escenario 1, con 3 y 5 discretizaciones, como en el escenario 2, con las discretizaciones indicadas por el experto, la tasa de error es más alta que la tasa de error de clasificación mayoritaria, lo cual es completamente inaceptable.

Tasas de error para el escenario 1						
	Discretización *Stand-alone*			Discretización sugerida por el experto		
Número de discretizaciones	Tasa de error con CN2	Tasa de error C4.5-*rules*	Tasa de error de la clase mayoritaria	Tasa de error con CN2	Tasa de error con C4.5-*rules*	Tasa de error de la clase mayoritaria
2	12,8%	22,6%	47,2%	23,3%	13,5%	26,7%
3	36,1%	27,1%	61,3%	42,1%	28,6%	36,5%
5	46,6%	41,4%	61,3%	67,7%	68,4%	47,2%

Tasas de error para el escenario 2						
	Discretización *Stand-alone*			Discretización sugerida por el experto		
Número de discretizaciones	Tasa de error con CN2	Tasa de error C4.5-*rules*	Tasa de error de la clase mayoritaria	Tasa de error con CN2	Tasa de error con C4.5-*rules*	Tasa de error de la clase mayoritaria
2 c. prueba *A*	13,8%	6,9%	47,3%	8,8%	10,1%	30,8%
2 c. prueba *B*	13,2%	9,3%	47,3%	6,2%	5,4%	30,8%
3 c. prueba *A*	17,0%	16,4%	64,7%	30,2%	17,0%	42,2%
3 c. prueba *B*	0,8%	7,8%	64,7%	5,4%	7,8%	42,2%
5 c. prueba *A*	28,9%	27,7%	66,6%	34,0%	32,7%	52,7%
5 c. prueba *B*	22,5 %	22,5 %	66,6%	25,6%	19,4%	52,7%

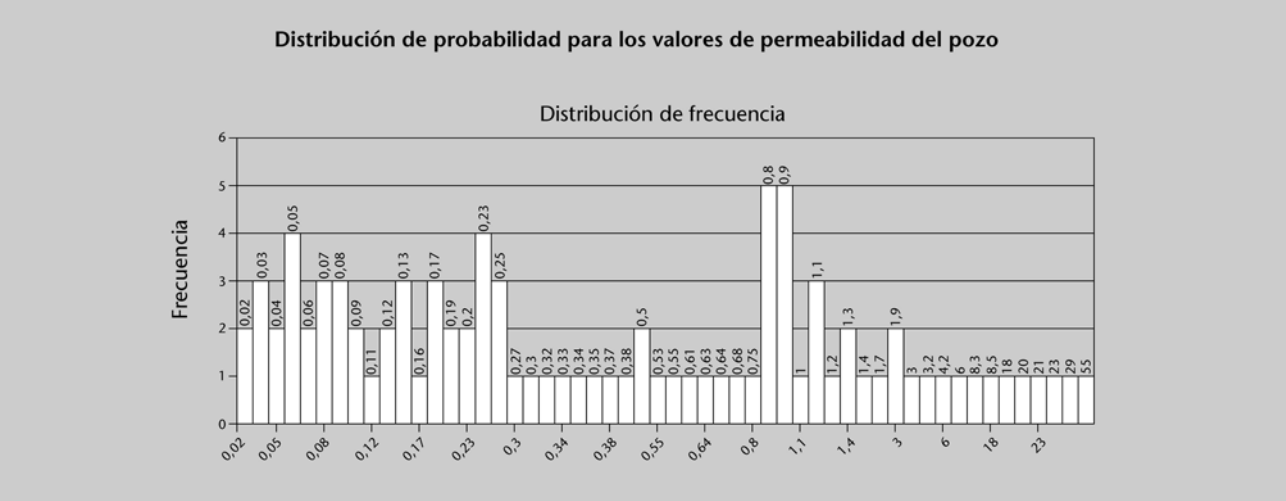
Tasas de error para escenario 3						
	Discretización *Stand-alone*			Discretización sugerida por el experto		
Número de discretizaciones	Tasa de error con CN2	Tasa de error C4.5-*rules*	Tasa de error de la clase mayoritaria	Tasa de error con CN2	Tasa de error con C4.5-*rules*	Tasa de error de la clase mayoritaria
2	14,5%	6,9%	48,8%	8,8%	10,1%	27,2%
3	21,4%	15,7%	62,3%	29,6%	18,2%	37,8%
5	27,7%	25,2%	62,3%	28,9%	32,1%	48,8%

Nos concentramos ahora en analizar un escenario concreto con una discretización fija con el fin de mejorar la clasificación y obtener reglas de clasificación de una calidad más alta.

Basándonos en las consideraciones previas, elegimos el escenario 2 (con tres discretizaciones dadas por el experto con el conjunto de prueba *A*). La matriz de confusión para estos supuestos es la que aparece representada en la tabla siguiente.

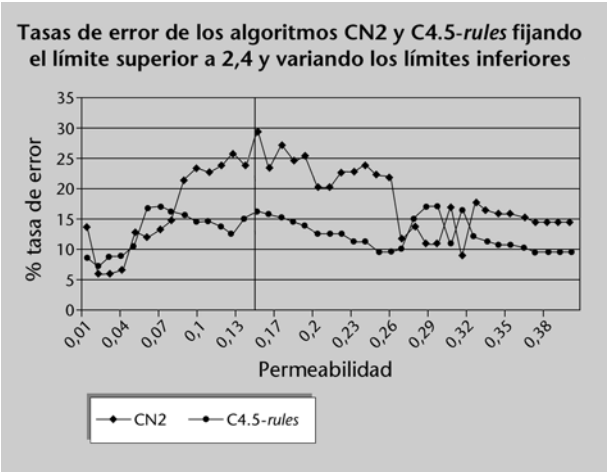
Matriz de confusión para el escenario 2 con tres discretizaciones con ayuda del experto utilizando el conjunto de prueba *A*										
Discretización	CN2					C4.5-rules				
		(a)	(b)	(c)	Error		(a)	(b)	(c)	Error
< 0,155 [0,155, 2,4] > 2,4	(a)	90	3	0	3,2%	(a)	92	1	0	1,0%
	(b)	41	9	14	83,3%	(b)	25	28	1	48,1%
	(c)	0	0	12	0%	(c)	0	0	12	0%

La distribución de los valores de permeabilidad de los valores del pozo 1.704, que era lo que se utilizó como conjunto de prueba de este escenario, se muestran en la figura siguiente. Al valor 0,009 le corresponden sesenta y seis elementos que no se muestran en el gráfico.



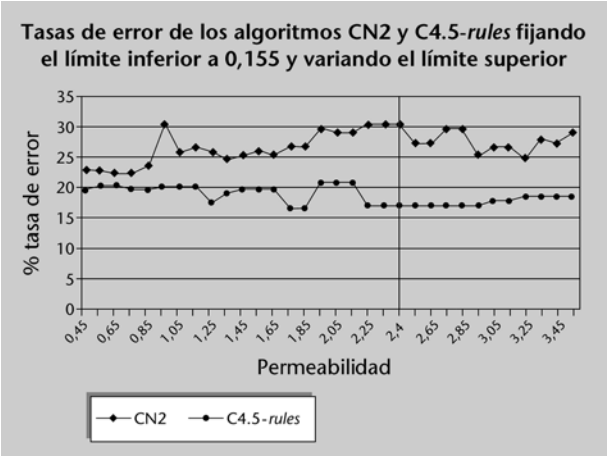
Con el objetivo de mejorar la precisión de los resultados tanto del CN2 como de C4.5-*rules,* establecimos uno de los límites del intervalo en un valor prefijado e hicimos variar el otro límite con valores más altos y más bajos. Para cada variación se determinaron las tasas de error de cada algoritmo.

Empezamos por fijar un límite superior de 2,4 y por hacer variar el límite inferior (0,15). En la figura siguiente podemos ver los resultados obtenidos. Las tasas de error más bajas se obtuvieron con el algoritmo CN2, fijando el límite inferior en 0,015, 0,025 y 0,035. Para el caso del algoritmo C4.5-*rules*, las tasas de error más bajas se obtuvieron con límites inferiores de 0,015, 0,025 y 0,035.



Después fijamos el límite inferior en 0,155 e hicimos variar el límite superior utilizando valores más altos y más bajos. En la figura siguiente mostramos los resultados obtenidos. Con CN2 los valores 0,65 y 0,75 dieron las tasas de error más bajas. Para el caso del C4.5-*rules,* los valores que correspondían a la tasa

de error más baja fueron 1,75 y 1,85. Es importante subrayar aquí que sólo se utilizó el valor 1,75 para las otras pruebas, y no el valor 1,85, dado que ambos valores ofrecían resultados parecidos.



Utilizando como referencia las tasas de error mínimas, se eligieron los mejores límites inferiores y superiores próximos (pero no iguales) a 0,015 y 2,4. A continuación, se generaron las matrices de confusión para cada caso a fin de intentar determinar cuál era el intervalo con la tasa de error mínima y considerando la distribución de los conjuntos de datos. En este escenario se utilizaron 694 casos de entrenamiento y 159 de prueba.

Las matrices de confusión para el método CN2 que utilizan las tasas de error mínimas de este algoritmo aparecen en el anexo 1; las que corresponden al método C4.5-*rules* podéis verlas en el anexo 2.

Podéis ver las matrices de confusión para el método CN2 en el anexo I, y las que corresponden al método C4.5-*rules,* al final de este módulo.

Basándonos en estos resultados, pudimos establecer que las discretizaciones con tasa de error mínima y distribución relativa a sus datos eran 0,015 y 0,75 y 0,015 y 1,75. Puesto que consideramos que estos puntos de corte daban un buen resultado en este escenario, se utilizaron en otros con el fin de determinar su conducta. Las matrices de confusión para estos dos nuevos intervalos de discretización se muestran en el anexo 3.

Podéis ver las matrices de confusión para los nuevos intervalos de discretización en el anexo 3, que se halla al final de este módulo.

3. Resultados

La tabla siguiente muestra la comparación entre los resultados de la discretización sugerida por el experto y la obtenida con los puntos de 0,015 y 1,75, que es con la que se obtuvieron los mejores resultados para todos los escenarios.

Comparación de la discretización sugerida por el experto con la discretización con puntos de corte de 0,015 y 1,75						
	Discretización sugerida por el experto con [0,155, 2,4]			Discretización usando los puntos de corte [0,015, 1,75]		
Escenario	Tasa de error con CN2	Tasa de error C4.5-*rules*	Tasa de error de la clase mayoritaria	Tasa de error c on CN2	Tasa de error con C4.5-*rules*	Tasa de error de la clase mayoritaria
1	42,1%	28,6%	36,5%	38,3%	39,8%	64,4%
2 c. prueba *A*	30,2%	17,0%	42,2%	13,8%	6,9 %	59,9%
2 c. prueba *B*	5,4%6	7,8 %	42,2%	3,9%	3,1%	59,9%
3	29,6%	18,2 %	37,8%	8,2%	5,7 %	56,7%

Podemos observar que, a excepción del escenario 1, la nueva discretización mejora la precisión en la clasificación.

El algoritmo C4.5-*rules* generó diez reglas a partir del conjunto de entrenamiento del escenario 2 (694 casos). Estas reglas no cubren los noventa y nueve casos de los 694 casos totales (14,3%).

Podéis ver las reglas generadas a partir del algoritmo C4.5-*rules* en el anexo 4 que encontraréis al final del módulo.

Presentamos un poco más adelante el resultado obtenido al aplicar tres reglas al conjunto de prueba.

a) Regla 20:

Porosidad > 15.7 ⇒
⇒ clase > 1,75 [11,1%] [112 14] [14,3%] [12 2]

b) Regla 1:

Porosidad ≤ 2,9 ⇒
⇒ clase < 0,015 [0,0%] [203 0] [1,6%] [60 1]

c) Regla 19:

Clase > 3,9 ⇒
⇒ clase ≤ 15,7
⇒ clase [0,015:1,75] [17,5%] [188 40] [7,3%][76 6]

La conclusión de cada regla muestra la clase, además de información en el formato siguiente:

- [AA %]. Error de clasificación al aplicar la regla al conjunto de entrenamiento.
- [BB y CC]. Número de ejemplos clasificados correctamente [BB] y clasificados incorrectamente [CC] dentro del conjunto de entrenamiento.
- [DD %]. Error de clasificación al aplicar la regla al conjunto de prueba.
- [EE y FF]. Número de ejemplos clasificados correctamente [EE] y clasificados incorrectamente [FF] dentro del conjunto de prueba.

Con el mismo escenario y conjunto de entrenamiento, CN2 generó ciento veinte reglas que dejan sin cubrir dieciocho casos (2,6%). Ahora bien, más del 50% de estas reglas son especializaciones para poder cubrir uno, dos o tres ejemplos.

3.1. Valoración

Los resultados indican que la elección de un método de discretización u otro afecta directamente a los métodos CN2 y C4.5-*rules*.

Para el caso de la discretización *Stand-alone*, el algoritmo C4.5 ha dado mejores resultados, en el sentido de que la precisión es mejor que la del CN2 en un 83,3% de los casos, y la misma en un 8,3 %.

Si se utiliza una discretización que recoge las indicaciones del experto, con C4.5-*rules* se obtiene un resultado mejor que con CN2 en un 41,6% de los casos, y en un 25% presenta los mismos resultados.

Además, debemos tener en cuenta la matriz de confusión para poder analizar la distribución de los elementos. Por ejemplo, se muestra una precisión del 100% con CN2 para doce elementos del conjunto *C*, pero esto carece de toda relevancia si lo comparamos con el conjunto *A* con CN2, que presenta una precisión del 96,8% para noventa y tres elementos.

La discretización híbrida aumenta la precisión considerablemente en todos los escenarios menos en el escenario 1.

A pesar de la gran mejora alcanzada, el papel del experto es fundamental en la definición de la discretización. La discretización con métodos no supervisados ofrece al experto una buena ayuda para determinar los mejores intervalos. Finalmente, el número de intervalos también puede influir en la precisión de los métodos de *data mining* aplicados. El método híbrido tiende a incrementar la complejidad al incluir más intervalos.

Bibliografía

Breiman, L.; Friedman, J.; Olshen R.; Stone C. (1993). *Classification and Regression Trees*. Nueva York: Chapman and Hall.

Catlett, J. (1991). *Megainduction: Machine Learning on Very Large Databases*. Tesis doctoral. Universidad de Sydney.

Chmielewski, M.; Grzymala-Busse, J. (1995). "Global Discretization of Continuous Attributes as Preprocessing for Machine Learning". En: T.Y. Lin; A.M. Wildberger, (eds.). *Soft Computing, Society for Computer Simulation* (págs. 294-301). San Diego.

Clark, P.; Niblett, T. (1989). "The CN2 Induction Algorithm". *Machine Learning* (vol. 3, págs. 261-283).

Dougherty, J.; Kohavi, R.; Sahami, M. (1995). "Supervised and Unsupervised Discretizations of Continuous Features". En: *Proceedings of the 12th International Conference on Machine Learning* (págs. 194-202). Morgan Kaufmann Publishers.

Fayyad, U.M.; Irani K.B. (1993). *Multi-interval discretization of continuous-valued attributes for classification learning*. En: *Proceedings of the 13th International Conference on Machine Learning* (págs. 1.022-1.027). Morgan Kaufmann Publishers.

Fayyad, U.M.; Piatesky-Shapiro, G.; Smyth, P.; Uthurusamy, R. (1996). *Advanced in Knowledge and Data Mining*. Menlo Park. AAAI/MIT Press.

Grzymala-Busse, W. (1992). "Lers - A System for Learning from Examples Based on Rough Sets". En: R. Slowinski (ed.). *Intelligence Decision Support. Handbook of Applications and Advances of the Rough Set Theory* (págs. 3-18). Dordrecht: Kluwer Academic Publishers.

Kerber, R. (1992). *ChiMerge: Discretization of Numeric Attributes* (págs. 123-127) En: *Proceedings of the 10th National Conference on Artificial Intelligence*.

Kohavi, R.; John, G.; Long, R.; Manley, D.; Pfleger, K. (1994). "MLC++: A Machine Learning Library in C++". En: *Tool with Artificial Intelligence* (pág. 740-743). IEEE Computer Society Press.

Lenarcik, A.; Piasta, Z. (1992). *Discretization of Condition Attribute Space* (págs. 373-389) En: R. Slowinski (ed.). *Intelligence Decision Support. Handbook of Applications and Advances of the Rough Set Theory*. Dordrecht: Kluwer Academic Publishers.

Lenarcik, A.; Piasta, Z. (1993). "Probabilistic Approach to Decision Algorithm Generation in the Case of Continuous Condition Attributes". *Foundations of Computing and Decision Sciences* (vol. 18, núm. 3-4, págs. 213-224). Poznan.

Lenarcik, A.; Piasta, Z. (1995). "Minimizing the Number of Rules in Deterministic Rough Classifiers". En: T.Y. Lin; A.M. Wildberger (eds.). *Soft Computing (*págs. 32-35). San Diego: Society for Computer Simulation.

Molina, F.L.C.; Oliveira, R.S.; Doi, C.Y.; De Paula, M.F.; Romanato, M.J. (1998). "MLC++: Biblioteca de Aprendizado de Máquina em C++". *Technical Report 72*. São Paulo: ICMSC/USP.

Nguyen, S.H.; Skowron, A. (1995). "Quantization of Real Value Attributes: Rough Set and Boolean Reasoning Approach". *Proceedings of the Second Joint Annual Conference on Information Sciences* (págs. 34-37). Wrightsville Beach.

Pfahringer, B. (1995). "Compression-based discretization of continuous attributes". En: A. Prieditis; S. Russel (eds.). *Proceedings of the 12th International Conference on Machine Learning*. Morgan Kaufmann Publishers.

Quinlan, J.R. (1987). "Generating Production Rules from Decision Trees". En: *Proceedings of 4th International Machine Learning Workshop* (págs. 304-307). San Mateo: Morgan Kaufmann.

Quinlan, J.R. (1990). "Induction of Decision Trees". En: Shavlik, J.W.; Dietterich, T.G. (eds.). *Readings in Machine Learning* (págs. 57-69). San Mateo: Morgan Kaufmann Publishers.

Quinlan, J.R. (1993). *C4.5 Programs for Machine Learning*. San Mateo: Morgan Kaufmann Publishers.

Rogers, S.J.; Chen, H.C.; Kopaska-Merkel, D.C.; Fang, J.H. (1995). "Predicting Permeability from Porosity Using Artificial Neural Networks". *American Association of Petroleum Geologists Bulletin* (vol. 79, diciembre, págs. 1.786-1.797).

Silicon Graphics. *MineSet* (1997). *http://www.sgi.com/Products/software/MineSet/*.

Ventura D.; Martinez T.R. (1994). "BRACE: A Paradigm For the Discretization of Continuously Valued Data". En: *Proceedings of the 7th Florida Artificial Intelligence Research Symposium* (págs. 117-121).

Ventura D.; Martinez, T.R. (1995). "An Empirical Comparison of Discretization Methods". En: *Proceedings of the 10th International Symposium on Computer and Information Sciences* (págs. 443-450).

Water Resource Research Center (1998). *Glossary of Organizations and Acronyms. College of Agriculture*. The University of Arizona [*on-line*]. *http://Ag.Arizona.Edu/Azwater/ Gloss.Html*.

Anexos

Anexo 1

Tasas de error mínimas para el algoritmo CN2

Intervalo	CN2					C4.5-*rules*					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Tasa de error de la clase mayoritaria (%)
< 0,015	(a)	61	5	0	7,5	(a)	60	6	0	9,1	
[0,015,0,65]	(b)	11	41	5	28	(b)	1	50	6	12,2	
> 0,65	(c)	4	4	28	22,2	(c)	0	7	29	19,4	
Total					18,2					12,6	59,9

Intervalo	CN2					C4.5-*rules*					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Tasa de error de la clase mayoritaria (%)
< 0,015	(a)	63	3	0	4,5	(a)	60	6	0	9,1	
[0,015,0,75]	(b)	2	58	0	3,3	(b)	1	54	5	10	
> 0,75	(c)	0	18	15	54,5	(c)	0	7	26	21,2	
Total					14,5					11,9	59,9

Intervalo	CN2					C4.5-*rules*					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Tasa de error de la clase mayoritaria (%)
< 0,025	(a)	64	4	0	5,8	(a)	61	7	0	10,2	
[0,025,0,65]	(b)	15	35	5	36,3	(b)	0	49	6	10,9	
> 0,65	(c)	2	6	28	22,2	(c)	0	7	29	19,4	
Total					20,1					12,6	57,8

Intervalo	CN2					C4.5-*rules*					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Tasa de error de la clase mayoritaria (%)
< 0,025	(a)	62	6	0	8,8	(a)	61	7	0	10,2	
[0,025,0,75]	(b)	8	44	6	24,1	(b)	0	53	5	8,6	
> 0,75	(c)	0	7	26	21,2	(c)	0	7	26	21,2	
Total					17					11,9	57,8

Intervalo	CN2					C4.5-*rules*					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Tasa de error de la clase mayoritaria (%)
< 0,035	(a)	65	6	0	8,4	(a)	61	10	0	14	
[0,035,0,65]	(b)	7	40	5	23	(b)	0	46	6	11,5	
> 0,65	(c)	1	7	28	22,2	(c)	0	7	29	19,4	
Total					16,4					14,5	56,5

Intervalo	CN2					C4.5-*rules*					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Tasa de error de la clase mayoritaria (%)
< 0,035	(a)	65	6	0	8,4	(a)	61	10	0	14	
[0,035,0,75]	(b)	5	45	5	18,1	(b)	0	54	1	1,8	
> 0,75	(c)	0	7	26	21,2	(c)	0	12	21	36,3	
Total					14,5					14,5	56,5

Anexo 2

Tasas de error mínimas para el algoritmo C4.5-*rules*

Intervalo	CN2					C4.5-*rules*					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Tasa de error de la clase mayoritaria (%)
< 0,015	(a)	62	4	0	6	(a)	60	6	0	9	
[0,015,1,75]	(b)	11	63	5	20,2	(b)	1	76	2	3,7	
> 1,75	(c)	1	1	12	14,2	(c)	0	2	12	14,2	
Total					13,8					6,9	59,9

Intervalo	CN2					C4.5-*rules*					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Tasa de error de la clase mayoritaria (%)
< 0,025	(a)	62	6	0	8,8	(a)	61	7	0	10,2	
[0,025,1,75]	(b)	17	54	6	29,8	(b)	0	75	2	2,5	
> 1,75	(c)	1	1	12	14,2	(c)	0	2	12	14,2	
Total					19,5					6,9	57,8

Intervalo	CN2					C4.5-*rules*					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Tasa de error de la clase mayoritaria (%)
< 0,035	**(a)**	63	8	0	11,2	**(a)**	61	10	0	14	
[0,035,1,75]	**(b)**	16	52	6	29,7	**(b)**	0	72	2	2,7	
> 1,75	**(c)**	0	2	12	14,2	**(c)**	0	2	12	14,2	
Total					20,1					8,8	56,5

Anexo 3

Matrices de confusión correspondientes a los valores 0,015 y 0,75, y 0,015 y 1,75

Escenario	CN2					C4.5-*rules*					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Tasa de error de la clase mayoritaria (%)
1	**(a)**	15	5	2	31,8	**(a)**	14	7	1	36,3	
	(b)	10	23	18	54,9	**(b)**	2	32	17	37,2	
	(c)	4	10	46	23,3	**(c)**	0	14	46	23,3	
Total					36,8					30,8	54,4

Matrices de confusión con los puntos de corte 0,015 y 0,75 en todos los escenarios

Escenario	CN2					C4.5-*rules*					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Tasa de error de la clase mayoritaria (%)
2 B	**(a)**	78	0	0	0	**(a)**	77	1	0	1,2	
	(b)	5	35	4	20,4	**(b)**	1	39	4	11,3	
	(c)	2	2	3	57,1	**(c)**	0	5	2	71,4	
Total					10,1					8,5	59,9

Escenario	CN2					C4.5-*rules*					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Tasa de error de la clase mayoritaria (%)
2 B	**(a)**	62	4	0	6	**(a)**	62	4	0	6	
	(b)	7	46	7	23,3	**(b)**	1	58	1	3,3	
	(c)	0	7	26	21,2	**(c)**	0	12	21	36,3	
Total					15,7					11,3	56,7

Matrices de confusión con los puntos de corte 0,015 y 1,75 en todos los escenarios

Escenario	CN2					C4.5-*rules*					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Tasa de error de la clase mayoritaria (%)
1	**(a)**	17	4	1	22,7	**(a)**	16	2	4	27,2	
	(b)	14	30	16	50	**(b)**	8	21	31	65	
	(c)	1	15	35	31,3	**(c)**	0	8	43	15,6	
Total					38,3					39,8	54,4

Escenario	CN2					C4.5-*rules*					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Tasa de error de la clase mayoritaria (%)
2 B	**(a)**	77	1	0	1,2	**(a)**	77	1	0	1,2	
	(b)	1	47	0	2	**(b)**	1	47	0	2	
	(c)	1	2	0	0	**(c)**	0	2	1	66,6	
Total					3,9					3,1	59,9

Escenario	CN2					C4.5-*rules*					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Tasa de error de la clase mayoritaria (%)
3	**(a)**	62	4	0	6	**(a)**	62	4	0	6	
	(b)	6	72	1	8,8	**(b)**	1	76	2	3,7	
	(c)	0	2	12	14,2	**(c)**	0	2	12	14,2	
Total					8,2					5,7	56,7

Anexo 4

Reglas generadas por el algoritmo C4.5-*rules*

```
C4.5 [release 8] rule generator     Thu Jul 9 03:04:38 1998

-------------------------------

 Options:
      Rulesets evaluated on unseen cases
      File stem </var/tmp/AAAa000ox>

Read 694 cases (2 attributes) from /var/tmp/AAAa000ox

-------------------------------

Processing tree 0
Final rules from tree 0:

Rule 20:
      porosidad > 15.7
      -> class 1.75+ [86.4%]

Rule 18:
      profundidad > 15599
      porosidad > 3.9
      -> class 1.75+ [79.4%]
```

```
Rule 12:
      profundidad > 15426
      profundidad <= 15496
      porosidad > 10.8
      -> class 1.75+ [78.7%]

Rule 17:
      profundidad > 15582
      porosidad > 11.1
      -> class 1.75+ [64.5%]

Rule 1:
      porosidad <= 2.9
      -> class - 0.015 [99.3%]

Rule 4:
      profundidad > 15412
      porosidad <= 3.9
      -> class - 0.015 [96.3%]

Rule 16:
      profundidad > 15582
      profundidad <= 15599
      porosidad <= 11.1
      -> class - 0.015 [86.7%]

Rule 14:
      profundidad > 15515
      porosidad <= 10.3
      -> class - 0.015 [77.2%]

Rule 19:
      porosidad > 3.9
      porosidad <= 15.7
      -> class 0.015-1.75 [65.3%]

Rule 3:
      porosidad > 2.9
      porosidad <= 3.2
      -> class 0.015-1.75 [54.6%]

Default class: 0.015-1.75

Evaluation on training data (694 items):
```

Rule	Size	Error	Used	Wrong	Advantage	Class
20	1	13,6%	126	14 (11,1%)	53 (64|11)	> 1,75+
18	2	20,6%	4	0 (0,0%)	4 (4|0)	> 1,75+
12	3	21,3%	33	11 (33,3%)	11 (22|11)	> 1,75+
17	2	35,5%	6	2 (33,3%)	3 (4|1)	> 1,75+
1	1	0,7%	203	0 (0,0%)	56 (56|0)	< 0,015
4	2	3,7%	22	4 (18,2%)	1 (4|3)	< 0,015
16	3	13,3%	8	1 (12,5%)	3 (3|0)	< 0,015
14	2	22,8%	60	27 (45,0%)	6 (33|27)	< 0,015
19	2	34,7%	228	40 (17,5%)	0 (0|0)	[0,015, 1,75]
3	2	45,4%	2	0 (0,0%)	0 (0|0)	[0,015, 1,75]

```
Tested 694, errors 99 (14.3%) <<


   (a)     (b)     (c)  <-classified as
  ----    ----    ----
   261      16       1  (a): class - 0.015
    32     192      26  (b): class 0.015-1.75
     0      24     142  (c): class 1.75+
```

```
Evaluation on test data (159 items):
```

Rule	Size	Error	Used	Wrong	Advantage	Class
1	1	0,7%	61	1 (1,6%)	59 (60|1)	< 0,015
19	2	34,7%	82	6 (7,3%)	0 (0|0)	[0,015, 1,75]
20	1	13,6%	14	2 (14,3%)	10 (12|2)	> 0,75

```
Tested 159, errors 11 (6.9%) <<


   (a)     (b)     (c)  <-classified as
  ----    ----    ----
    60       6       0  (a): class < 0.015
     1      76       2  (b): class [0.015,1.75]
     0       2      12  (c): class > 1.75
```