GESTIÓN Y DATA MINING

Raúl Alberto Ercole

Contador Público

Profesor de la Universidad Nacional de Córdoba y Universidad Católica de Córdoba

Correo: ercole3@fibertel.com.ar

Catalina Lucía Alberto

Contadora Pública

Profesor de la Universidad Nacional de Córdoba

Correo: catalina.alberto@gmail.com

Claudia Etna Carignano

Contadora Pública

Profesor de la Universidad Nacional de Córdoba

Correo claudiacarignano@gmail.com

Resumen

La toma de decisiones tiene múltiples facetas y enorme variedad de situaciones.

En algunos casos la información de base es difícil de interpretar adecuadamente. En otros, el modelo de decisión no refleja la realidad en forma concluyente. En algunos más, las posibles acciones no son fácilmente ejecutables o la incertidumbre es demasiado importante.

De allí la existencia de variados métodos de apoyo a las decisiones, justamente porque las situaciones decisorias son disímiles y cambiantes.

Si se intenta visualizar hacia dónde debería en el futuro encaminar sus pasos el profesional de gestión, debe pensarse que la gestión es dinámica y requiere de actualizaciones permanentes, que las necesidades de la gestión de hoy son

marcadamente diferentes, que la comunicación interdisciplinaria es absolutamente necesaria, que los sistemas de información se han automatizado y que en este ámbito compiten profesionales de diversas extracciones y que no es suficiente producir información sino que en cambio debe administrarse la gestión.

Es en este sentido que debe ampliarse la investigación, el conocimiento y la aplicación de nuevas técnicas o métodos de apoyo a decisiones y a la gestión. El presente trabajo orienta su objetivo a este propósito, introduciendo los conceptos elementales y fundamentales del "data mining".

Al final del trabajo se expone en forma sucinta una de las técnicas (denominada "vecinos más cercanos"), entre las varias factibles y conocidas como "data mining" o "minería de datos".

Palabras clave: Gestión, Data, Mining

Abstract

Decision making has multiple facets and wide variety of situations.

In some cases the basic information is difficult to interpret properly. In others, the decision model does not reflect reality conclusively. In some others, the possible actions are not easily enforceable or uncertainty is too important.

Hence the existence of various methods of decision support, precisely because the situations are different and changing makers.

If you try to visualize where in the future should direct his steps the professional management, the management must think that is dynamic and requires constant updates, the management needs of today are markedly different, that interdisciplinary communication is absolutely necessary that information systems have been automated and professionals compete in this area of diverse backgrounds and is not sufficient to produce information but instead be given the management.

It is in this sense that further investigation, knowledge and application of new techniques or methods of decision support and management. This paper focuses its target for this purpose, introducing the basic concepts and fundamental «data mining».

At the end of the paper describes succinctly one technique (called «nearest neighbors»), among several feasible and known as «data mining» or «data mining».

***Keywords**: Management, Data, Mining*

I. Objetivo

Data Mining es un proceso de exploración y análisis de datos con el objetivo de descubrir comportamientos y/o reglas significativos, correlaciones o tendencias.

Si bien es un campo de análisis relativamente nuevo y en proceso de evolución, tiene amplia aplicación en la vida de los negocios.

Clasificar y predecir comportamientos de nuevos clientes, analizar posibles preferencias de consumo, clasificar deudas según el índice de cobrabilidad, estimar posibles ingresos futuros, intentar descubrir cuáles clientes podrían estar más deseosos de abandonar una suscripción de servicio o muchísimas otras cuestiones de negocios, pueden ser abordadas por diferentes métodos y técnicas de data mining.

En general el proceso debe lidiar con grandes bases de datos, lo que dificulta la precisión. El empleo de técnicas estadísticas, conceptos matemáticos y facilidades de la computación ayuda en la creación de diferentes modelos que pueden ser apropiados según el caso considerado, pero se debe ser en todo momento consciente de los errores que se puedan producir y que, desde ya, deben ser mensurados y evaluados en su dimensión relativa.

Existen muchos métodos para predicción y clasificación. Cada uno de ellos es aplicable en determinadas situaciones, según las características del tamaño de la base de datos, el tipo de comportamiento inherente a la misma, las irregularidades o "ruidos" que presenta y , fundamentalmente, el objetivo de la investigación.

Diferentes métodos pueden implicar diferentes resultados y por ello es de utilidad la aplicación de varios de ellos, seleccionando aquél que tenga una mejor correlación con el objetivo propuesto. De hecho, los métodos o técnicas se clasifican según los predictores (variables de ingreso) y las respuestas (variables de salida) sean continuas o categóricas, o si la búsqueda se orienta más a la segmentación.

Data mining toma elementos de la matemática, la estadística y la computación y las combina adecuadamente con aplicaciones concretas en la vida de los negocios o en el mundo científico.

El objetivo del trabajo se centra en el comentario sobre aspectos introductorios del proceso de data mining como una herramienta útil para la gestión y decisiones.

II. Etapas en el proceso de data mining

La aplicación de data mining con el objetivo de construir información sigue, en general, un proceso que puede ser sintetizado en los siguientes pasos:

1) Definir el propósito del estudio.

   El mismo puede encaminarse hacia la búsqueda de respuestas en una sola etapa de aplicación o puede ser un procedimiento continuo de predicción de datos.

2) Obtener la base de datos adecuada.

   Los datos pueden ser internos a la Organización o externos a ella. También pueden combinarse datos de diferentes fuentes para la construcción de una base relevante.

3) Explorar y reprocesar los datos.

   La base de datos debe examinarse, analizarse, graficarse, coordinar unidades de medida y períodos, verificar la consistencia, evaluar la dispersión de datos y valorizar la importancia de "outliers".

4) Determinar la tarea del data mining.

   Implica decidir si la tarea o salida del modelo es de clasificación (segmentación en clases) o de predicción continua (estimación numérica).

5) Particionar la base de datos

   Con el objeto de construir el modelo adecuado para la tarea propuesta y evitar la parcialidad de resultados, la base de datos debe ser particionada aleatoriamente en:

   - base de datos de trabajo: es la utilizada para construir diferentes modelos que luego serán evaluados en su ajuste al propósito. Es normalmente la proporción mayor de la base de datos.

   - base de datos de evaluación: es la utilizada para validar diferentes modelos que se hayan construido y elegir el o los más relevantes.

   El o los modelos escogidos puede, posteriormente, se reevaluados con nuevos datos de prueba.

6) Escoger la técnica de data mining.

   Supone elegir el método de análisis adecuado a la tarea propuesta, entre las posibles técnicas utilizadas frecuentemente en data mining como regresiones, redes neuronales, procesos de jerarquías, árboles de clasificación, análisis discriminante, análisis de cluster, reglas de asociación y muchas otras.

7) Construir y utilizar algoritmos para desarrollar la tarea.

   Comprende la evaluación de distintos algoritmos que puedan ser utilizados para el éxito del propósito.

8) Interpretar resultados

   Comprende la elección del/de los algoritmo/s más adecuado/s y el análisis de resultados en el sector de validación de la base de datos.

9) Desplegar el modelo.

Supone la aplicación definitiva del modelo en sistemas operacionales y en datos reales para control y toma de decisiones.

III. Modelos de clasificación

a) Matriz de clasificación

En una determinada base de datos, las clasificaciones generadas por un método predictor o clasificador pueden ser resumidas en la llamada "matriz de clasificación". Por ejemplo, una matriz de clasificación para 2 clases o categorías que provienen de la base de datos puede adoptar la siguiente estructura:

CLASES REALES	PREDIC CLASE A	PREDIC CLASE B
CLASE A	$n_{a.a}$	$n_{a.b}$
CLASE B	$n_{b.a}$	$n_{b.b}$

En la matriz, la celda superior izquierda muestra el número de observaciones que siendo efectivamente Clase A fueron clasificadas por el predictor de igual forma. Del mismo modo, la celda inferior derecha muestra el número de observaciones clasificadas correctamente en la Clase B.

Por el contrario, las otras celdas en diagonal (inferior izquierda y superior derecha) muestran el número de los errores de clasificación en relación a la situación real de cada observación de la base de datos.

En este caso, con 2 clases tanto reales como predicción, el número total (n) de observaciones es

$$n = n_{a.a} + n_{a.b} + n_{b.a} + n_{b.b}$$

La tasa de error general de clasificación se computa como sigue:

$$t_e = \frac{n_{a.b} + n_{b.a}}{n}$$

del mismo modo que la tasa de precisión general se calcula como:

$$t_p = 1 - t_e = \frac{n_{a.a} + n_{b.b}}{n}$$

La matriz de clasificación brinda estimados de las predicciones acertadas y de los errores de clasificación. Cabe aclarar que son estimaciones, no obstante que si la base de datos es lo suficientemente amplia y significativa, los resultados podrán ser bastante confiables.

En aquellos casos en que los datos se encuentran en algunas bases publicadas (como censos, por ejemplo) que podrían usarse para la estimación de las proporciones de precisión o error, pero la realidad es que en la mayoría de los problemas de negocios o económicos no se conocen estas tasas.

Para obtener un error de estimación confiable, debe utilizarse la matriz de clasificación obtenida con la base de datos de evaluación. Es decir, construido el modelo con la base de datos de trabajo, se aplica el mismo a la base de datos de evaluación.

Es de esperar inferiores resultados en la matriz de clasificación obtenida con la base de datos de evaluación respecto a la matriz que se pueda obtener con la base de datos de trabajo. Sin embargo, como se expresó, un estimador de porcentaje de errores debe buscarse en la primera.

En caso de existir diferencias significativas entre ambas matrices de clasificación, deberá revisarse el modelo y sus resultados.

b) El valor de corte

El primer paso en la mayoría de los algoritmos de clasificación es estimar la probabilidad que un caso concreto pertenezca a cada una de las clases.

En muchos casos, hay una clase que es de especial interés; entonces, el foco estará en esta particular clase y se comparará la probabilidad estimada de pertenecer a dicha clase con un "valor de corte". Si la probabilidad de pertenecer a la clase de interés está arriba del valor de corte, el caso es asignado a dicha clase.

Por default, el valor de corte para un caso de 2 clases es 0,5. Obviamente, cambiando el valor de corte, se altera la tasa de error general de clasificación.

Por ejemplo, supóngase un caso que tiene los siguientes datos base:

CLASE A: Propietario

CLASE B: Inquilino

REGISTRO	PROBABILIDAD CLASE A	CLASE REAL
1	0,99	A
2	0,98	A
3	0,97	A
4	0,96	A
5	0,94	A
6	0,89	A
7	0,85	A
8	0,76	B
9	0,71	A
10	0,68	A
11	0,66	A
12	0,62	B
13	0,51	A
14	0,47	B
15	0,34	B
16	0,22	A
17	0,20	B
18	0,15	B
19	0,07	B
20	0,06	B
21	0,05	B
22	0,04	B
23	0,03	B
24	0,01	B

En base a los datos, se construyen distintas matrices de clasificación según el valor de corte prefijado, lo que se resume en las tablas siguientes:

VALOR DE CORTE 0,5

MATRIZ DE CLASIFICACIÓN

	CLASE PREDICCION		
CLASE REAL	PROPIETARIO	INQUILINO	
PROPIETARIO	11	1	12
INQUILINO	2	10	12
TOTAL	13	11	24

VALOR DE CORTE 0,25

MATRIZ DE CLASIFICACIÓN

	CLASE PREDICCION		
CLASE REAL	PROPIETARIO	INQUILINO	
PROPIETARIO	11	1	12
INQUILINO	4	8	12
TOTAL	15	9	24

VALOR DE CORTE 0,75

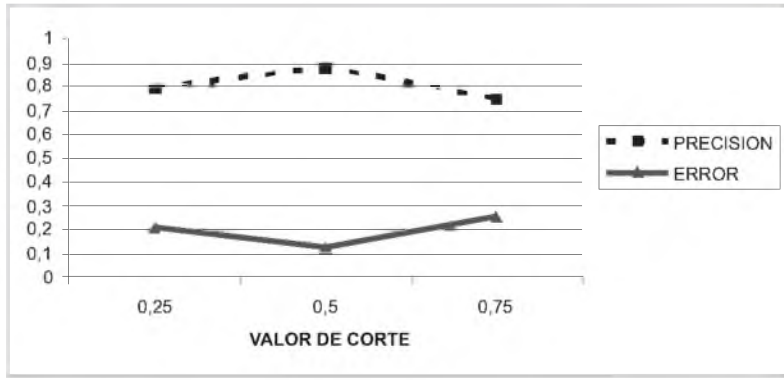
MATRIZ DE CLASIFICACIÓN

	CLASE PREDICCION		
CLASE REAL	PROPIETARIO	INQUILINO	
PROPIETARIO	7	5	12
INQUILINO	1	11	12
TOTAL	8	16	24

Calculando la tasa de precisión y de error se arriba a los siguientes resultados:

CORTE	PRECISION	ERROR
0,25	0,791666667	0,208333333
0,5	0,875	0,125
0,75	0,75	0,25

lo que también puede ser observado en el gráfico siguiente:



Puede ser interesante utilizar valores de corte diferentes al promedio, dado que los costos de errores de clasificación pueden ser asimétricos. En otras palabras, se aceptarán mayores errores de clasificación cuando es menor el costo de error.

Un clásico ejemplo es el de créditos bancarios a otorgar, donde es mucho más importante predecir con precisión eventuales faltas de pago o morosidades que devoluciones en forma, dado que es mucho más costosa la morosidad. En estos casos el error general de clasificación (o la tasa de precisión) no es el mejor indicador de evaluación del modelo clasificador.

Más bien, se usan las siguientes medidas:

La "sensibilidad" del clasificador es la habilidad para detectar los miembros de la clase relevante en forma correcta.

Se calcula como

$$S = \frac{n_{a.a}}{n_{a.b} + n_{a.a}}$$

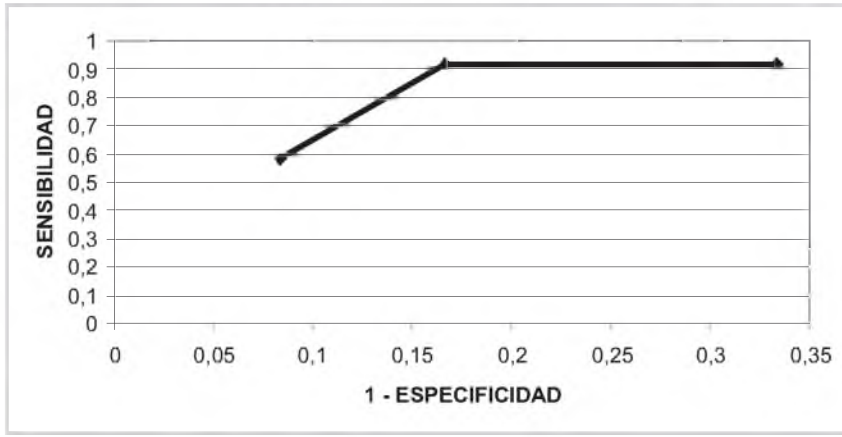
La "especificidad" del clasificador es la habilidad para clasificar miembros no pertenecientes a la clase relevante en forma correcta.

Se calcula como

$$E = \frac{n_{b.b}}{n_{b.b} + n_{b.a}}$$

Para el mismo ejemplo planteado, los cálculos y el gráfico indican lo siguiente:

CORTE	SENSIBILIDAD	ESPECIFICIDAD	1 - ESPECIFIC
0,25	0,916666667	0,666666667	0,333333333
0,5	0,916666667	0,833333333	0,166666667
0,75	0,583333333	0,916666667	0,083333333



Mejores performances del modelo clasificador están reflejados por curvas que están más cercanas al sector izquierdo superior del gráfico (alta sensibilidad y alto índice de especificidad).

c) El gráfico de ganancias

El gráfico de ganancias ayuda en ser efectivos para la selección de un pequeño número de casos y lograr una relativamente gran porción de aciertos.

Supóngase el ejemplo anterior de 24 registros:

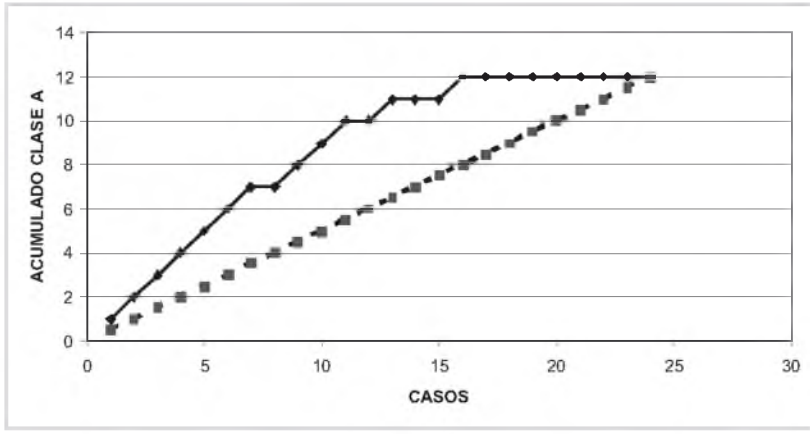
REGISTRO	PROBABILIDAD CLASE A	CLASE REAL	ACUM CL A
1	0,99	A	1
2	0,98	A	2
3	0,97	A	3
4	0,96	A	4
5	0,94	A	5
6	0,89	A	6
7	0,85	A	7
8	0,76	B	7
9	0,71	A	8
10	0,68	A	9
11	0,66	A	10
12	0,62	B	10
13	0,51	A	11

CONTINÚA EN PÁGINA SIGUIENTE

REGISTRO	PROBABILIDAD CLASE A	CLASE REAL	ACUM CL A
14	0,47	B	11
15	0,34	B	11
16	0,22	A	12
17	0,20	B	12
18	0,15	B	12
19	0,07	B	12
20	0,06	B	12
21	0,05	B	12
22	0,04	B	12
23	0,03	B	12
24	0,01	B	12

La tabla muestra 24 casos con la probabilidad de pertenecer a la Clase A (clase de interés) en orden descendiente, clase a la que pertenece y acumulado de pertenencia a la clase de interés (la Clase A).

El gráfico de ganancias se construye con la primera y la última columna del modo siguiente:



Se observa la "ganancia" del gráfico en relación a una línea imaginaria que marque el promedio, o sea la unión de los puntos (0,0) y (24,12).

En el gráfico o en la tabla puede observarse que si se eligiera 10 casos como Clase A se estaría acertando en 9 de ellos, mientras que el promedio simple (azar o aleatorio) sólo "acertaría" en 5 de ellos, lo que da un coeficiente de ganancia de 9/5 = 1,8

De hecho que si se incluyen más casos, el porcentaje de ganancia va decreciendo. Por ejemplo, con 15 casos, el índice es de 11/7,5 = 1,4666

d) Costos asimétricos de errores de clasificación

En general, es factible asumir que el costo (o beneficio) de efectuar correctas clasificaciones es cero. Ello es posible trabajando con costos reales y con costos de oportunidad que se producen en los errores de clasificación. Es decir, en lugar de observar el beneficio por clasificar correctamente, se observa y trabaja con el costo de no clasificar en forma correcta, incluyendo costos de oportunidad.

Supóngase el ejemplo de enviar una oferta de productos a 10.000 clientes, 1% de los cuales responderá, en promedio, con compras.

Usando un método clasificador de data mining, se logra calcular la matriz de clasificación. Supóngase que la misma es la siguiente:

CLASE REAL	PREDIC A	PREDIC B
A no responde	965	25
B compra	3	7

Estas clasificaciones logradas por el método de data mining utilizado tienen un error de clasificación de 28/1.000 = 2,80%.

Supóngase que los costos de envío de ofertas son de $ 1 y los beneficios de una compra de $ 10. En tal caso, y con esa matriz de clasificación, se enviaría la oferta a los 32 clientes clasificados como B y se tendría lo siguiente:

- costo de envío a 25 clientes que no compran: $ 25
- costo de envío a 7 clientes que compran: $ 7
- beneficio de 7 clientes que compran: $ 70
- beneficio neto final: 70 - 32 = $ 38

Este beneficio neto puede ser observado también desde los costos reales y de oportunidad comparando las 2 acciones posibles a efectuar. En efecto:

1) No emplear ningún procedimiento clasificatorio de data mining y simplemente usar la regla del "no responde"; ello implicaría no haber enviado oferta alguna a nadie y tener un costo de oportunidad de $ 100 por los 10 clientes que hubieran comprado.

2) Utilizar el método de data mining, construir la matriz de clasificación y enviar la oferta a los 32 clientes; en este caso los costos reales son los $ 32 de envío y los costos de oportunidad los $ 30 de 3 clientes que hubieran comprado pero que no se les envió oferta dado que fueron clasificados como "no responde", lo que hace un total de $ 62.

La diferencia entre las 2 políticas refleja el beneficio de $ 38.

Para reflejar los costos asimétricos en las clasificaciones, es conveniente utilizar un indicador que tenga en cuenta tal circunstancia.

Una popular medida del "costo promedio de error de clasificación", denotando por (q_a) el costo de error en clasificación de una observación de Clase A (clasificada como Clase B) y por (q_b) el costo de error de clasificación de una observación de Clase B (clasificada como Clase A), es:

$$CPEC = \frac{q_a * n_{a.b} + q_b * n_{b.a}}{n}$$

Esto implica buscar un método clasificador que minimice esta cantidad, la que inclusive puede ser computada para diferentes valores de corte.

Dado que en general es más sencillo estimar el ratio (q_a/q_b) que los valores individuales, muchos softwares trabajan con el ratio considerando que es igual minimizar una expresión que minimizar la misma expresión dividida por una constante.

Si en lugar de 2 clases existen "m" clases, la matriz de clasificación tendrá, por supuesto, "m" filas y "m" columnas. Los costos de errores de clasificación asociados con las celdas diagonales serán, entonces, cero. Sin embargo, evaluar CPEC es mucho más complicado en casos como este, en donde existen *m*(m-1)* tipos de errores de clasificación. Conviene en estos casos, por lo tanto, operar con una clase conceptualizada como "relevante" (definida por el decisor) y dejar las otras como "no importantes".

IV. Modelos de predicción

En el caso de respuestas continuas (predicción con salida numérica) la evaluación del comportamiento del modelo es diferente a la correspondiente a una salida categórica (clasificación o asignación de clase).

La precisión en la predicción de un modelo, que es la base del data mining para la exploración de nuevos casos, no es igual a la bondad del ajuste en relación a los datos de trabajo; esta última se corresponde más al objetivo de las clásicas medidas de coeficientes de regresión o errores estándar de estimación.

Los indicadores usados para medir la precisión en la predicción están calculados en los datos de validación y no en los de trabajo. Esto es así porque el grupo de registros de validación no están utilizados para seleccionar predictores o para estimar los coeficientes de los modelos

La predicción en un nuevo registro surge de la salida promedio de los registros del área de trabajo, y el error de predicción para un registro cualquiera "i" es la diferencia entre el valor real (y) con el valor de su predicción $(\hat{y})$

$$e_i = y_i - \hat{y}_i$$

Las medidas más populares de precisión en las predicciones son:

a) MEDIA DEL ERROR ABSOLUTO

$$MA = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

b) MEDIA DEL ERROR

$$M = \frac{\sum_{i=1}^{n} e_i}{n}$$

c) MEDIA DEL ERROR PORCENTUAL ABSOLUTO

$$MPA = \frac{\sum_{i=1}^{n} \left| \frac{e_i}{y_i} \right|}{n} \times 100\%$$

d) RAIZ DE LA MEDIA DEL ERROR CUADRADO

$$RMC = \sqrt{\frac{\sum_{i=1}^{n} e_i^{\,2}}{n}}$$

e) SUMA DEL ERROR CUADRADO

$$SC = \sum_{i=1}^{n} e_i^{\,2}$$

Estas medidas se utilizan para comparar modelos y evaluar su grado de precisión. Todas las medidas están influenciadas por "outliers" o valores fuera de rango, por lo que se deben evaluar los errores cuidadosamente.

Es importante hacer notar que un modelo con alta precisión predictiva puede no coincidir con un modelo que ajuste la base de datos de trabajo de la mejor forma.

V. Modelo ejemplo: vecinos más cercanos

El modelo de "VECINOS MÁS CERCANOS" es un algoritmo útil para clasificación de una salida categórica o para predicción de una salida numérica.

El método se apoya en encontrar similares registros en los datos de trabajo; estos "vecinos" se usan para la clasificación o predicción de un nuevo registro.

a) Clasificador (respuesta categórica)

La idea en el método es encontrar "k" registros en los datos de trabajo que sean similares al nuevo registro que se desea clasificar, asignando el mismo a la clase predominante en sus vecinos más cercanos.

El método no hace supuestos acerca del tipo de relación existente entre los miembros de una clase (Y) y sus predictores (X).

La cuestión central es determinar la distancia entre registros basándose en los valores de sus predictores, para lo que se utiliza la distancia euclídea.

Dicha distancia entre 2 registros (v_1, v_2,....,v_p) y (w_1, w_2,....., w_p) se calcula con la expresión siguiente:

$$D(v,w)=\sqrt{(v_1-w_1)^2+(v_2-w_2)^2+....+(v_p-w_p)^2}$$

siendo "v" y "w" las variables predictoras.

En el caso de diferentes escalas para los predictores, será preciso normalizar la misma previamente al cálculo de la distancia euclídea. Para la normalización, lo más común es dividir todas las observaciones por su suma, lo que convierte a cada registro en un porcentaje del total y así puede operarse en comparaciones (restas) con otros predictores.

Una vez calculada la distancia, es necesaria una regla de clasificación para asignar una clase al registro deseado, teniendo en cuenta la clase de sus vecinos.

En el caso de k=1, se asigna la clase de su registro más cercano.

Para una extensión de k>1, se encuentran los k vecinos más cercanos, y luego se sigue la regla de la mayoría para la asignación de la clase.

Por ejemplo, puede tratarse de un caso de grado de cumplimiento en el pago de clientes. Los mismos han sido clasificados en 2 grupos (1 = satisfactorio; 2 = no satisfactorio) y se consideran predictores sus índices de patrimonio y de rentabilidad.

La base de datos debe primero segregarse, como se expresó anteriormente, en base de datos de trabajo y base de datos de evaluación.

Supóngase que la siguiente sea la base de datos de trabajo (el análisis sólo es válido para ilustración, dado el pequeño tamaño de la muestra:

Observación	Grupo	Patrimonio	Rentab
1	1	45	36
2	1	43	42
3	1	42	31
4	1	40	43
5	1	42	37

CONTINÚA EN PÁGINA SIGUIENTE

Observación	Grupo	Patrimonio	Rentab
6	1	39	34
7	1	38	40
8	1	38	29
9	1	38	45
10	1	37	43
11	1	36	35
12	2	39	33
13	2	37	36
14	2	36	31
15	2	36	27
16	2	34	39
17	2	33	35
18	2	31	29
19	2	30	38
20	2	30	34

En primer lugar se normalizan los valores de las variables predictoras (índices de patrimonio y rentabilidad), lo que brinda los valores siguientes:

NORMALIZACIÓN	
Patrimonio	Rentabilidad
0,06048387	0,05020921
0,0577957	0,05857741
0,05645161	0,0432357
0,05376344	0,05997211
0,05645161	0,05160391
0,05241935	0,0474198
0,05107527	0,05578801
0,05107527	0,0404463
0,05107527	0,06276151
0,04973118	0,05997211
0,0483871	0,0488145
0,05241935	0,0460251
0,04973118	0,05020921
0,0483871	0,0432357
0,0483871	0,0376569
0,04569892	0,05439331
0,04435484	0,0488145
0,04166667	0,0404463
0,04032258	0,05299861
0,04032258	0,0474198

Ahora se supone el caso de un nuevo cliente (nuevo registro) con índices de 44 en patrimonio y 29 en liquidez. Una vez normalizada la base de datos y los valores correspondientes al nuevo cliente, se calcula la distancia euclídea del nuevo registro (nuevo cliente) con cada una de las observaciones, utilizando para ello los valores normalizados.

El vecino más cercano (menor distancia) para k=1, es el registro 3 que tiene 42 y 31, respectivamente, de índices, por lo que el nuevo cliente sería clasificado como grupo 1 (satisfactorio).

Para k=3, los vecinos más cercanos son 3, 8 (clase 1) y 12 (clase 2), por lo que por mayoría queda asignado al grupo 1 (satisfactorio).

El valor de "k" generalmente es elegido como impar, para evitar empates. Si se usa un valor pequeño de "k", se atiende a las características de la base de datos. Por el contrario, un valor de "k" muy alto simplemente significa que se predecirá la clase más frecuente en todos los casos.

Debe buscarse un valor de "k" que brinde el mejor rendimiento en la clasificación. Para tener una idea de ello, se computa la tasa de error para distintos valores de "k" en la matriz de clasificación de la base de datos de evaluación.

Debe recordarse que la tasa de error de clasificación se mide como:

$$t_e = \frac{n_{a.b} + n_{b.a}}{n}$$

Supóngase, sólo a fines ilustrativos por lo pequeño de la muestra, que la base de datos de evaluación es la siguiente:

Observación	Grupo	Patrimonio	Rentabilidad
21	1	41	33
22	1	41	40
23	2	36	33
24	1	45	30
25	2	30	34
26	2	36	27

Para cada uno de esos registros, se utiliza el método de normalizar / calcular distancias con cada uno de los registros de la base de datos de

trabajo y se clasifica, utilizando varios valores de "k". A posteriori, se confeccionan las matrices de clasificación de los registros de la base de datos de evaluación, para cada valor de "k", y se computa la tasa de error.

En cuanto al valor de corte, la regla de la mayoría está asociada implícitamente al mismo. Por ejemplo, en el caso de k=3, la mayoría es 2/3 (registros 3 y 8) y ése es el valor de corte para asignar clase (a igual resultado se hubiera llegado con un valor de 0,5 que no hace más que respetar la mayoría).

El procedimiento explicado puede perfectamente ser aplicado a un caso en el que existan más de 2 clases.

b) Predicción (respuesta numérica)

La metodología de "vecinos más cercanos" en caso de predicción de valores numéricos en lugar de respuesta categórica (asignación a una clase) permanece inalterable en la etapa inicial de cálculo de distancias euclídeas.

En la segunda etapa, el voto mayoritario para asignación de clase, es reemplazado al tomar el valor de la respuesta promedio de los "k" vecinos más cercanos. El promedio puede ser un promedio ponderado, con ponderaciones más altas para los puntos que están más cercanos al punto del cual es necesaria la predicción.

También cambia la evaluación del mejor "k", pues en lugar de usar la tasa del error de clasificación, se usa algunos de los indicadores típicos de predicción, como los expuestos en el apartado "MODELOS DE PREDICCIÓN" de este trabajo.

c) Evaluación del modelo

El modelo de "vecinos más cercanos" funciona perfectamente bien, tanto para clasificación como para predicción, especialmente si hay buenos y suficientes predictores.

Su mayor problema reside, por un lado, en el prohibitivo tiempo que puede demandar encontrar los vecinos más cercanos en grandes bases de datos. Por otro lado, en la medida que el número de predictores crezca, la base de datos necesaria para ser relevante crece desmesuradamente.

Por ello suelen utilizarse técnicas de reducción de base de datos que permitan trabajar con similar precisión en menor cantidad de registros.

VI. Conclusión

El trabajo no tiene mayor pretensión que un desarrollo de los conceptos fundamentales del data mining.

En el mismo se expusieron.los objetivos, las etapas del data mining y las características de los modelos de clasificación y los modelos de predicción.

En el caso de los modelos de clasificación, se enfatizó la importancia de la matriz de clasificación y el cálculo de la tasa de error general de clasificación.

También se hicieron las consideraciones pertinentes acerca del valor de corte y la relevancia de calcular indicadores de sensibilidad y especificidad para el caso de clasificación o respuesta categórica.

A continuación se explicó cómo el gráfico de ganancias ayuda en ser efectivos para la selección de un pequeño número de casos y lograr una relativamente gran porción de aciertos. También se hizo referencia al caso de costos asimétricos en los errores de clasificación y en función de ello cómo surge la necesidad de calcular el costo promedio del error de clasificación.

Para el caso de los modelos de predicción de un valor numérico, se cambian los indicadores de precisión del modelo y para ello se expusieron varias medidas.

Por último, se expusieron las características generales de un modelo específico de data mining, el de vecinos más cercanos, tanto en su versión de clasificación como en la de predicción.

Los cambios tecnológicos han traído un cambio y un gran avance en los procesos de data mining. Es innumerable la cantidad de softs desarrollados en tal sentido, la necesidad de adaptarse a nuevas situaciones como la importancia que han cobrado los datos no estructurados (texto o páginas de Internet), la necesidad de integrar los algoritmos y resultados obtenidos en sistemas operacionales, la exigencia de que los procesos funcionen prácticamente en línea o los tiempos de respuesta muchas veces necesarios en tiempo real.

Por todo ello el profesional de gestión debe estar abierto y atento a todos los desarrollos vigentes de la disciplina y sumamente predispuesto a la integración interdisciplinaria.

VII. Bibliografía

ERCOLE, Raúl - ALBERTO, Catalina - CARIGNANO, Claudia - «MÉTODOS CUANTITATIVOS PARA LA GESTIÓN» - Segunda Edición -

Asociación Cooperadora de la FCE - UNC - Córdoba, 2007 - ISBN 978-987-1436-01-9

SHMUELI, GALIT - PATEL, NITIN R - BRUCE, PETER C - "Data Mining for Business Intelligence" - Segunda edición - John Wiley & sons - Hoboken - New Jersey (2010) - ISBN 978-0-470-52682-8

ANDERSON, David - SWEENEY, Dennis - WILLIAMS, Thomas - «Quantitative Methods for Business" - 9e - Thomson South Western - USA (2004) - ISBN 0-324-18414-X

BIERMAN, Harold - BONINI, Charles - HAUSMAN, Warren - "ANÁLISIS CUANTITATIVO PARA LOS NEGOCIOS". - Novena Edición - Irwin - McGraw Hill. Bogotá, 2000 - ISBN 0-256-14021-9

RAGSDALE, Cliff T. "SPREADSHEET MODELING AND DECISION ANALYSIS" - 3rd. edition - South Western College Publishing. Cincinnati - Ohio (USA), 2001 - ISBN 0-324-02122

POWELL, Stephen G - BAKER, Kenneth R - "MANAGEMENT SCIENCE - THE ART OF MODELING WITH SPREADSHEETS" - Second Edition - John Wiley & sons - USA, 2007 - ISBN 978-0-470-03840-6