

# La gestión de clientes en el comercio electrónico.

## Aplicación de Algoritmos Genéticos (AG) al CRM.

---

LAURA NÚÑEZ LETAMENDÍA

*Instituto de Empresa*

**La diferencia entre la gestión de clientes en el entorno del comercio tradicional y el entorno de Internet o del comercio electrónico viene dada precisamente por las posibilidades brindadas en este último**

83

mo caso, como consecuencia de la mayor amplitud de datos almacenados en formato digital por parte de la empresa acerca del comportamiento del consumidor. Además de diversos datos personales (exigidos para darse de alta como comprador en cualquier web) como nombre y apellidos, dirección postal, teléfono de contacto, número de la tarjeta de crédito, etc., las tiendas «online» disponen de otro tipo de información, incluso más valiosa que la anterior: frecuencia y tiempo empleado en los diferentes sitios visitados en la web de la empresa, compras históri-

cas realizadas a través de la web de la empresa, camino por el que se ha llegado a la web de la empresa (buscador, anuncio, enlace de otra página web, e-mail publicitario, etc.), si el cliente ha solicitado o no información a través del e-mail y qué tipo de información, etc.

Precisamente toda esta información que las empresas «online» son capaces de almacenar sin apenas esfuerzo, y el valor que a la misma se le atribuye, ha sido uno de los argumentos sobre los que se ha sustentado la enorme subida bursátil registrada en el

año 1999 y primera mitad del 2000 por las empresas de Internet, despreciándose los mecanismos tradicionales de valoración de empresas (descuentos de cash flows, ratio cotización/beneficio por acción (PER), ratio precio/valor contable, etc.), y justificándose cotizaciones realmente desorbitadas para empresas de nueva creación que mantenían grandes masas de datos de clientes o usuarios.

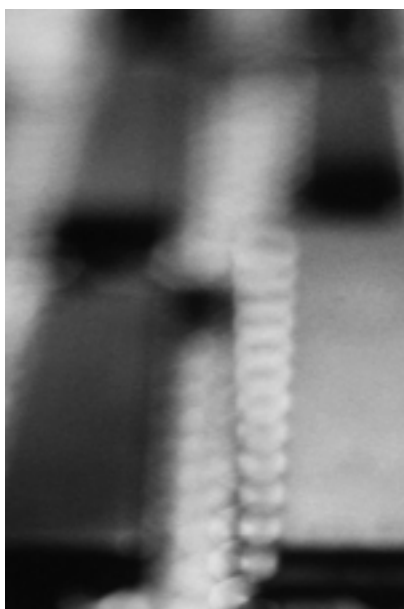
Es pertinente recordar, por ejemplo, que una empresa como Terra llegó a valer en bolsa el doble que Repsol, empresa ésta

última con una trayectoria ya consolidada y que obtenía 44.000 millones de euros en ingresos anuales, con unos fondos propios de 15.000 millones y un resultado neto consolidado de 2.400 millones de euros, frente a los 30 millones de ingresos de Terra, 6.000 millones de euros en fondos propios y unas pérdidas superiores a los 700 millones de euros. ¿Quiere decir entonces, el posterior desplome de las cotizaciones de estas empresas, que la información que poseen las mismas de los clientes o usuarios no es tan valiosa? La respuesta es «DEPENDEN».

Sólo en la medida en la que una compañía sea capaz de gestionar dicha información y transformarla o convertirla en ventas, ingresos, o rentabilidad para la empresa, esta información conforma un activo valioso. Así pues, en este nuevo entorno de Internet (desaparición de las fronteras geográficas, de las restricciones de horarios, acceso casi ilimitado a la información, etc.) se hace, aun más imprescindible que nunca, gestionar de forma óptima la relación con el cliente, ya que, como se suele decir, la competencia está a un «clic», puesto que los clientes tienen un acceso casi ilimitado e inmediato a información sobre empresas, productos, precios, etc., y por otro lado, la empresa dispone también de más información sobre los hábitos y preferencias de los clientes, información que constituye un activo estratégico para retener al cliente.

El término CRM o «Customer Relationship Management», que se ha venido empleando profusamente en los últimos años, literalmente se traduce como gestión de las relaciones con los clientes, sin embargo, hace referencia a un tipo de gestión determinada, basada en la concepción del cliente como elemento central de la compañía.

Son dos, fundamentalmente, las razones que han llevado a las empresas a variar su estrategia desde una percepción asentada en los productos a otra basada en los clientes. Por un lado, el entorno crecientemente competitivo en el que se ha venido desarrollando su actividad en los últimos años, debido a los procesos de globalización, internacionalización, desregulación, etc., ha originado una visión del cliente como uno de los activos fundamentales y



estratégicos de la empresa, y que por tanto debe ser gestionado de forma activa, con el objeto de vincularlo a la empresa permanentemente, aportándole más valor añadido que nuestra competencia.

Por otro lado, la revolución de las tecnologías de la información y las comunicaciones, y en especial la aparición de Internet y el acceso casi infinito a la información que dicho entorno permite (información de empresas, información de usuarios, información de productos, de servicios, etc.) genera posibilidades de interacción entre empresas, entre clientes y entre ambos, que eran impensables hace sólo unos años.

El CRM exige un cambio de orientación en toda la empresa —no sólo en los departamentos de marketing, ventas, o atención al cliente— enfocado a la creación de valor para el cliente a partir del análisis de sus necesidades y preferencias (análisis que adquiere una nueva dimensión en el entorno de Internet), con objeto de maximizar la rentabilidad de éste para la empresa. En definitiva, se trata de ofrecer a los clientes una propuesta de valor superior, basada en el conocimiento que la empresa ha acumulado sobre sus preferencias, comportamientos, etc.

Lógicamente, esta nueva concepción de la estrategia del negocio necesita un soporte computacional que sea capaz de in-

tegrar la información sobre clientes que posee la empresa y que facilite las herramientas de análisis adecuadas para su tratamiento. Este soporte computacional está constituido por las denominadas «aplicaciones de CRM», que conforman, por tanto, un elemento indispensable en la implementación de cualquier estrategia de CRM en la empresa. En este sentido, los paquetes de software que son comercializados por empresas como Oracle, Siebel, E.piphany, etc., trabajan en dos ámbitos para lograr una gestión efectiva de la interacción empresa-cliente:

- ✓ Recopilan, digitalizándola previamente en caso necesario, toda la información que la empresa posee sobre cada cliente (por ejemplo, la información recibida a través de Internet, con la recibida en los «call centres», o la que los diferentes departamentos de la empresa poseen sobre el cliente: pagos, reclamaciones, entregas, marketing, logística, contabilidad, etc.) en una gran base de datos, es lo que se denomina «data warehouse».

- ✓ Emplean las técnicas tradicionales de estadística y econometría para realizar el análisis de esa información «data mining».

Si bien pueden aplicarse Algoritmos Genéticos (en adelante, AG) —técnica de inteligencia artificial subsimbólica inspirada en los mecanismos y teorías de la evolución, la selección natural y la genética— al diseño de bases de datos (Cedeño and Vemuri, 1997), y por tanto del «data warehouse», es en el segundo punto, el de la explotación de la información que la empresa posee de los clientes, en el que las técnicas de inteligencia artificial subsimbólica, y en concreto los AG, pueden aportar más valor añadido, tanto mediante su integración en las propias aplicaciones de CRM como en otras complementarias a las mismas.

Es relativamente frecuente encontrar empresas que han invertido grandes cantidades de dinero en la creación de una «data warehouse» y que posteriormente no han sabido implementar los mecanismos adecuados para sacar beneficio de toda esa información, no siendo capaces de aumentar la rentabilidad de su base de clientes, lo que revela la trascendencia de la fase de análisis de los datos. En este senti-

do, los AG permiten optimizar la explotación de dicha información —mediante, por ejemplo, su aplicación a la caracterización de perfiles de clientes por rentabilidad, o al análisis de la efectividad de las políticas de marketing e interacción de la empresa sobre la base de la información almacenada— y además, de forma dinámica en el tiempo, lo que es importante, ya que las preferencias y hábitos de los consumidores van variando con el tiempo.

Otra aplicación de AG, que puede utilizarse de forma complementaria en la gestión de clientes, es la construcción de indicadores adelantados de consumo para diferentes tipos de bienes, a partir de datos, obtenidos de fuentes secundarias, sobre evolución de la renta, empleo, tipos de interés, datos sectoriales, etc. Y finalmente, un área también de interés para la aplicación de estos algoritmos, y que además adquiere mayor relevancia en el entorno del comercio electrónico, es el de la logística: optimización de la gestión de los stocks, optimización de las rutas de reparto de los pedidos, etc.

Aunque el empleo de AG en la resolución de problemas en el campo de la ciencia y la ingeniería (diseño de barcos, navegación aérea, predicción meteorológica, selección de rutas para circuitos de telecomunicaciones, procesamiento de señales, composición óptima de materiales, etc.) está empezando a ser una realidad y cuenta ya con un cuerpo amplio de bibliografía (véase Davis, 1991), su uso en el entorno de la gestión empresarial es casi inexistente. Sin embargo, si empiezan a generarse en este campo algunas contribuciones a la literatura empírica con resultados, en general, notables —selección de estrategias en mercados oligopolísticos, Marks (1989); gestión de inversiones, Bauer (1994), Allen *et al.* (1999); gestión del riesgo de insolvencia en la banca, Varetto (1998); construcción de indicadores adelantados de coyuntura, Farley and Jones (1994), etc.—, por lo que cabe esperar que en un futuro próximo se vayan produciendo también algunas aplicaciones reales.

No obstante, existen campos de la gestión empresarial prácticamente sin explorar en relación con la aplicación de AG, como es el caso propuesto en este estudio, la gestión de clientes, a pesar de la



especial idoneidad que presenta dicha problemática para la aplicación de esta técnica, especialmente adecuada para el tratamiento masivo de datos de forma dinámica y los procesos de búsquedas combinatorias y numéricas.

Así pues, el objetivo de este artículo es contribuir a la literatura sobre aplicaciones de AG en el ámbito de la empresa, y en concreto en el área de la gestión de clientes en el entorno del comercio electrónico, mediante la modelización de una aplicación de AG desarrollada para caracterizar los perfiles de los clientes más rentables y analizar la efectividad del marketing empleado por una empresa *on-line* concreta, a partir de la información que vía Internet recoge de sus clientes. El próximo apartado describe brevemente las bases y funcionamiento de la metodología de AG, mientras que el siguiente explica el proceso de implementación de la aplicación. Y finalmente, el último epígrafe presenta algunas conclusiones sobre la aplicación de la metodología de AG en el área de la gestión de clientes para el comercio electrónico.

## Concepto y funcionamiento de los AG

Los AG —técnica ideada por John Holland en la década de los setenta— perte-

necen al campo de la inteligencia artificial subsimbólica que trata de simular los sistemas fisiológicos a partir de los cuales se produce la inteligencia en los seres humanos y que básicamente se corresponden con dos modelos. El modelo relativo al funcionamiento del cerebro (neuronas, conexiones entre neuronas, etc.), que ha dado lugar a las técnicas conexionistas, más conocidas como redes neuronales; y el modelo basado en la evolución de la especie a través del intercambio de información genética y la adaptación al entorno, que es el origen conceptual de los AG, a partir de los cuales se han derivado, a su vez, una serie de metodologías similares, y que se engloban en lo que se ha venido denominando técnicas de computación evolutiva, en las cuales el empleo de la terminología biológica es habitual, dado su origen.

Para una introducción a los modelos conexionistas pueden consultarse, entre otros, Barba-Romero y Olmeda (1993) y Kröse and Smagt (1991); una revisión de las diferentes técnicas que conforman el paradigma de los algoritmos evolutivos o computación evolutiva es la presentada en Michalewicz (1996), Dasgupta and Michalewicz (1997) y Fogel (2000).

Los AG son una herramienta de búsqueda que se emplea para resolver problemas complejos de diverso tipo (optimización, clasificación, selección, etc.) en múltiples

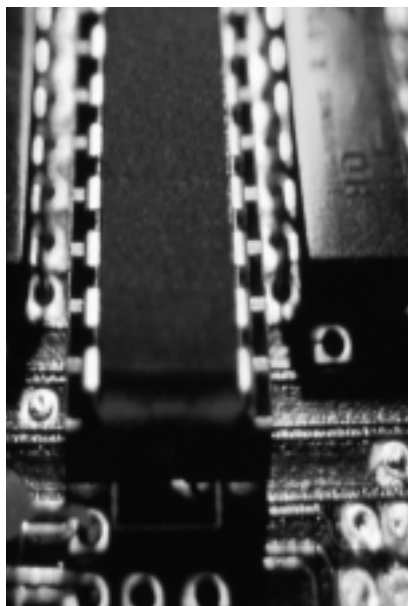
dominios, y que presentan características que los hacen superiores a los métodos tradicionales, basados en cálculo o en procedimientos aleatorios o enumerativos.

Estas características son: que no necesitan de la introducción de conocimientos sobre el problema a resolver —si bien, si se dispone de ellos, pueden ser utilizados—, ya que son técnicas que aprenden por sí mismas a partir de los datos históricos, generando modelos que son capaces de capturar el conocimiento; son técnicas intensivas en computación que trabajan de forma paralela y que permiten la utilización masiva de espacios de datos; tampoco es necesario presuponer formas funcionales determinadas como en los modelos econométricos, por lo que son muy útiles en el estudio de problemas con variables relacionadas de forma no lineal; son capaces de superar el problema de los óptimos locales si han sido diseñados correctamente (no obstante, su diseño adecuado no es una tarea trivial); y finalmente, pueden aplicarse sin la exigencia de ningún tipo de restricciones sobre las distribuciones de las variables objeto de estudio (continuidad, derivabilidad, normalidad, unimodalidad, etc.). A estas ventajas añade Goldberg (1989) su simplicidad y sus potentes efectos.

La aplicación de la técnica de AG requiere, en primer lugar, la representación del problema de forma adecuada para su tratamiento computacional, para lo que se instrumenta dicho problema, de tal forma que las posibles soluciones al mismo puedan representarse a través de vectores, que son denominados cromosomas. Además, es necesario definir una función objetivo, que permita clasificar las soluciones según su bondad, y que será la que guíe el proceso operativo del AG para que éste sea capaz de ir encontrando las mejores soluciones.

El funcionamiento del AG tradicional se basa en una serie de etapas:

**1** Se comienza generando aleatoriamente un grupo de potenciales soluciones para un problema concreto, el que se pretende resolver. Este grupo de soluciones, también denominado individuos, conforma la población de la primera generación del AG. Dichas soluciones o in-



dividuos deben representarse en forma de vectores o cromosomas, para lo que generalmente se emplea la codificación binaria (1). Cada cromosoma identificará un individuo o solución.

**2** En segundo lugar, se calcula el valor que toma la función objetivo para cada uno de los diferentes cromosomas.

**3** El tercer paso consiste en aplicar una serie de mecanismos u operadores que manipulan la información de los cromosomas en la dirección adecuada para incrementar el valor tomado por la función objetivo. Estos mecanismos son habitualmente tres: uno, de reproducción, donde las soluciones o cromosomas que mayor valor generan en la función objetivo son mayormente reproducidos; otro, de cruzamiento, donde la información de las cadenas cromosómicas previamente reproducidas se intercambia, y un tercero, de mutación, donde alguno de los valores de las cadenas cromosómicas es sustituido por otro de forma aleatoria. Los cromosomas resultantes de estas transformaciones conformarán la siguiente generación del AG.

Hay que tener en cuenta que el operador de reproducción no genera nuevas soluciones, ya que se limita a reproducir, en diferente proporción, según la ponderación dada por el valor de la función objetivo, los cromosomas o vectores ya presentes. Por ello, es sumamente importante que exista algún mecanismo que permita

abarcarse nuevas soluciones, ya que de otra forma estaríamos restringiendo nuestro espacio de búsqueda global a la selección aleatoria realizada inicialmente para generar la población de la primera generación del AG, con la consiguiente obtención de óptimos locales.

Este mecanismo lo conforman los operadores de cruzamiento y de mutación. Estos operadores son controlados por los parámetros denominados probabilidad de cruzamiento y probabilidad de mutación, que indican qué porcentaje de individuos del total de la población de cada generación debe someterse al proceso correspondiente de manipulación.

**4** A continuación se calculan los valores de la función objetivo para el nuevo conjunto de cromosomas obtenido tras la aplicación de los operadores mencionados, que forman una nueva generación del AG. Y de nuevo se vuelven a aplicar los operadores de reproducción, cruzamiento y mutación para obtener un nuevo conjunto de cromosomas o soluciones y por tanto una nueva generación.

**5** Se repite de forma iterativa el paso cuarto hasta que no se produzcan mejoras en los valores de la función objetivo o durante un número de iteraciones (generaciones) determinado.

La aplicación de la operativa descrita conduce a la obtención del conjunto de mejores soluciones para el problema que tratamos de resolver. Si bien es cierto que no existe forma de saber si hemos encontrado o no el conjunto de soluciones óptimas del problema, salvo que dichas soluciones puedan ser conocidas también mediante otras técnicas alternativas, no lo es menos que las simulaciones experimentales realizadas con AG han alcanzado, por lo general, resultados óptimos o al menos mejores que los encontrados mediante otras técnicas.

La capacidad de los AG reside en ser una metodología dinámica-evolutiva, estocástica, y que además trabaja en paralelo:

**Dinámica evolutiva** porque utiliza mecanismos inspirados en la reproducción de los seres vivos, donde las soluciones más cercanas a las óptimas (los indivi-

duos más adaptados, que generan mayor valor en la función objetivo) tienen mayores probabilidades de reproducirse, obteniéndose de esta forma sucesivos conjuntos de soluciones cada vez mejores.

**Estocástica**, puesto que emplea procesos aleatorios, partiendo de diversos conjuntos de información previamente seleccionados sobre la base de un objetivo, para permitir abarcar nuevos puntos de búsqueda.

**Paralela**. Los AG utilizan poblaciones de individuos o soluciones (cromosomas) que analizan en paralelo. A este paralelismo explícito hay que sumarle, además, el denominado paralelismo implícito (focalización simultánea de los esfuerzos de búsqueda en diversas regiones del espacio total de soluciones), que genera una exploración del espacio de soluciones potenciales muy superior a la explícitamente analizada y que es proporcional al cubo del número de soluciones explícitas que aparecen a lo largo de la ejecución del AG. La explicación de esta cuestión, que conforma uno de los pilares de los fundamentos matemáticos de los AG, va más allá de los objetivos de este artículo. Una descripción rigurosa y detallada de los fundamentos teóricos de la metodología puede encontrarse en Goldberg (1989).

## Aplicación de AG a la gestión de clientes en el comercio electrónico

A continuación se analiza el caso concreto de una librería «online» que cuenta con una base de datos de X millones de clientes o visitantes y que emplea una aplicación de AG para realizar la caracterización de los perfiles de los clientes que resultan en conjunto más rentables para la empresa y de las estrategias de marketing y de interacción empresa-cliente más efectivas. Esto le permitirá a la librería virtual diseñar políticas orientadas a incrementar el valor de la empresa mediante la retención y adquisición de nuevos clientes de un determinado perfil (los que aportan más rentabilidad) y el fomento de las ventas cruzadas teniendo en cuenta los intereses manifestados explícita o implícitamente por los usuarios.

CUADRO 1  
ESQUEMA DE CODIFICACIÓN DEL SUBVECTOR DE DATOS DE PERFIL PERSONAL

Edad	Sexo	Profesión	Zona de residencia	Nº hijos
010	0	0100	00011	01

Aunque el análisis inicial se hace tomando un período de referencia  $t$  que puede ser el último año, el último semestre, etc., la ventaja de los AG es que dicho período puede ser actualizado continuamente, lo que permite una adaptación de la estrategia de la empresa a los cambios de hábitos de los clientes.

La librería dispone de información almacenada sobre datos personales de los usuarios, páginas web de la empresa (secciones) visitadas por los clientes, enlace a través del cual han entrado en las páginas de la empresa (publicidad en un portal, etc.), información solicitada a través del e-mail por los clientes, clientes a los que la empresa les ha enviado información por e-mail o correo tradicional, compras efectuadas, etc. El análisis de cuestiones como el tipo de productos que han adquirido, la información que han empleado, las estrategias de marketing que han sido más rentables, etc., debe servir para diseñar las políticas de la empresa con el objetivo de maximizar su rentabilidad y el valor aportado a los clientes.

El primer paso para la aplicación de la técnica de AG es representar el problema de forma adecuada para su tratamiento computacional, y posteriormente debe implementarse el proceso operativo del AG definiendo la función objetivo.

### Representación del problema en términos vectoriales para su tratamiento computacional

En el caso aquí planteado, una opción puede ser construir un vector (formado por subvectores) indicativo de posibles perfiles de clientes, según la información obtenida a través de Internet almacenada por la empresa:

✓ Subvectores de datos de perfil personal: edad, sexo, zona de residencia, profesión, número de hijos a cargo, etc.

✓ Subvectores de las secciones visitadas en la web de la librería: economía, dere-

cho, literatura, deportes, salud, informática, historia, poesía, matemáticas, infantil, idiomas, enciclopedias, ocio, etc.

✓ Subvectores de los enlaces empleados para entrar en la web de la empresa: e-mail enviado, publicidad en buscadores, portales, o agregadores, directamente, etcétera.

✓ Subvectores de otras relaciones con el cliente: e-mail enviados, e-mail recibidos, llamadas de teléfono realizadas por el cliente, envíos de correspondencia, etc.

✓ Subvectores de las compras realizadas: cantidad gastada.

Para todos los subvectores emplearemos la codificación binaria, proyectando linealmente los valores dados por dicha codificación en los intervalos que deseamos.

Por ejemplo, el subvector de datos del perfil personal podría estar codificado de la siguiente forma (ver cuadro 1):

**Edad:** codificación binaria con tres dígitos, lo que permite obtener ocho niveles ( $2^3$ ) de segmentación que se proyectan a los siguientes intervalos:

000=0 menos de 16 años  
001=1 entre 16 y 22 años  
010=2 entre 22 y 28 años  
011=3 entre 28 y 35 años  
100=4 entre 35 y 45 años  
101=5 entre 45 y 55 años  
110=6 entre 55 y 65 años  
111=7 más de 65 años.

**Sexo:** codificación binaria con un dígito (0 indica hombre - 1 mujer).

**Profesión:** codificación binaria con cuatro dígitos ( $2^4$ ) para una clasificación de dieciséis sectores profesionales.

**Zona de residencia:** codificación binaria con número de dígitos en función de los niveles de segmentación que se desee

CUADRO 2  
ESQUEMA DE CODIFICACIÓN DEL SUBVECTOR DE DATOS DE SECCIONES VISITADAS

Economía	Derecho	Literatura	Deporte	Salud	Informática	Historia	Poesía	Matem.	Infantil	Idioma	Enciclop.	Ocio	Otros
01	00	01	10	10	00	00	10	11	00	01	00	11	01

(suponemos una segmentación de treinta y dos zonas, lo que implica cinco dígitos <sup>25</sup>).

**Número de hijos a cargo:** dos dígitos (00=0, ninguno; 01=1, uno; 10=2, dos; 11=3, más de dos).

La codificación del subvector de las secciones visitadas (ver cuadro 2) en un periodo  $t$  determinado se podría implementar en una escala de 4 niveles para cada sección, codificada en binario mediante dos dígitos (el 00 sería el nivel más bajo de visitas, por ejemplo, ninguna; el 01 el siguiente, por ejemplo, entre 1 y 10; el 10 más alto, más de 10 y menos de 30 y el 11 el más alto, más de 30).

También se podría trabajar con un número más exacto, en relación con las visitas, ampliando el número de dígitos considerado en la codificación binaria, seis dígitos permitirían identificar, con una exactitud total, de 0 hasta 63 visitas ( $2^6$  visitas): 000000 = 0; 000001 = 1; 000010 = 2; 000011 = 3; 000100 = 4; 000101 = 5; 000110 = 6; 000111 = 7; 001000 = 8; ..... ; 111111 = 63.

El subvector de los enlaces empleados para acceder a la web (ver cuadro 3) podríamos formarlo mediante la identificación previa (con un número) de las diferentes posibilidades existentes para entrar en la web y su posterior codificación binaria con tantos dígitos como necesitemos para abarcar todas las posibilidades. Registraríamos sólo el enlace utilizado la primera vez que el cliente contacta con la empresa. Utilizando 7 dígitos podemos abarcar desde 1 hasta 128 ( $2^7$ ) entradas diferentes.

Por ejemplo, esto resultará útil para saber qué enlaces concretos (qué buscador o qué portal, o la publicidad insertada en qué página, etc.) han conducido al visitante a nuestra web, y especialmente a aquellos visitantes que han sido rentables para la empresa, con objeto de cen-

CUADRO 3  
ESQUEMA EMPLEADO DE CODIFICACIÓN DEL SUBVECTOR DE DATOS DE ENLACE DE ENTRADA

0000000 = 0 = directamente
0000001 = 1 = e-mail enviado
0000010 = 2 = publicidad en página web de la empresa X
0000011 = 3 = publicidad en página web de la empresa Y
0000100 = 4 = publicidad en página F del portal Z
0000101 = 5 = enlace en página H del portal W
.....
1111111 = 127 = agregador S

CUADRO 4  
ESQUEMA DE CODIFICACIÓN DEL SUBVECTOR DE MODOS DE INTERACCIÓN EMPRESA-CLIENTE

e-mails enviados	Llamadas recibidas	Correspondencia	Otros
1	1	0	0

trar nuestros esfuerzos de marketing posteriores en la publicidad que resulta más ventajosa para nuestra empresa.

Codificamos el subvector de otras relaciones de la empresa con el cliente (ver cuadro 4) con un dígito para cada tipo de relación (el gen 0 implica que no se ha realizado ese tipo de relación y el 1 que sí).

Y finalmente, el subvector relativo a las compras realizadas (ver cuadro 5) podría codificarse siguiendo el siguiente patrón de un dígito por cada materia (donde el gen 0 implica que no se han realizado adquisiciones de esa materia y el gen 1 que sí), y 3 dígitos más representando la cantidad gastada (donde, por ejemplo, 000 = 0 sería nada, 001 = 1 sería hasta 150 euros, 010 = 2 sería hasta 300 euros, ....., 110 = 6 sería hasta 900 euros, y 111 = 7 sería más de 900 euros).

Así pues, hemos conseguido codificar toda la información de los clientes en macrocadenas cromosómicas o vectores con 71 genes formados únicamente por ceros

o unos, siendo los 3 últimos genes los que nos indican la cantidad gastada por el cliente en el periodo  $t$  de referencia (ver cuadro 6). Si hubiese más información, tendríamos que emplear este sistema para estructurarla computacionalmente.

### Implementación del proceso operativo del AG y definición de la función objetivo

Nuestro objetivo ahora es caracterizar los perfiles de clientes más rentables para la empresa en términos globales, para enfocar nuestro negocio hacia aquellos segmentos y estrategias que aportan mayor valor. El AG será el encargado de buscar dichos perfiles a través de la generación automática de combinaciones de características potenciales de clientes y su posterior comparación con los de la base de datos real de la empresa, para detectar qué perfil es el que define a los clientes más rentables.

El número máximo de perfiles diferentes que el AG podrá construir será de  $2^{68}$  =

CUADRO 5  
ESQUEMA DE CODIFICACIÓN DEL SUBVECTOR DE COMPRAS REALIZADAS

Economía	Derecho	Literatura	Deporte	Salud	Informática	Historia	Poesía	Matemát.	Infantil	Idiomas	Enciclop.	Ocio	Otros	Cantidad
0	1	0	1	0	0	0	1	1	1	0	1	0	1	001

CUADRO 6  
ESQUEMA DE CODIFICACIÓN DE LOS VECTORES IDENTIFICATIVOS DE LOS PERFILES DE LOS CLIENTES

Personal	Secciones visitadas	Enlaces	Relaciones	Compras
001010101000011	1010001010100111001110011001	1110011	0110	10101000110101110

295.147.905 millones», teniendo en cuenta que no consideramos dentro del perfil los 3 dígitos que codifican la cantidad gastada por cada cliente, ya que dicha cantidad vamos a emplearla dentro de la función objetivo que va a guiar la operativa del AG (2).

A continuación se describen las etapas de implementación del AG siguiendo el esquema descrito en el apartado segundo de este trabajo.

**Generación de n individuos.** El proceso del AG comienza mediante la generación de una población (que conformará la primera generación del AG) de n individuos o cromosomas —perfiles— de forma aleatoria (ver cuadro 7). El número de individuos (n) que conforman la población es uno de los parámetros del AG que es necesario fijar previamente. En general, debe ser una fracción muy pequeña del espacio de búsqueda (3) y debe fijarse teniendo en cuenta y de forma coherente con el parámetro relativo al número de iteraciones o generaciones (g) que el AG va a emplear en el proceso. El problema es, que al igual que para el resto de los parámetros que guían la operativa de los AG —probabilidad de cruzamiento y de mutación—, no existen reglas para fijar los valores que éstos deben tomar, valores que parecen ser dependientes del problema a resolver. Es ésta, de hecho, una de las líneas más actuales de investigación en el campo de los AG. En general, el valor de estos parámetros suele estimarse a través de un procedimiento de aproximación mediante el cual se prueba a introducir valores diferentes para los mismos en varias eje-

CUADRO 7  
GENERACIÓN ALEATORIA DE INDIVIDUOS O PERFILES

Cromosoma 1	0 0 0 1 1 0 1 0 0 1 1 1	.....	0 0 1 0 1 1 1 0 1
Cromosoma 2	0 1 1 1 0 0 1 0 0 0 1 0	.....	0 0 0 0 1 0 1 1 1
Cromosoma 3	1 1 0 1 0 1 1 1 1 0 1 1	.....	1 1 0 1 1 0 1 1 1
Cromosoma 4	1 1 1 0 0 1 1 1 0 1 0 1 0	.....	1 0 1 1 1 1 0 0 1 0
Cromosoma ...	.....	.....	.....
Cromosoma n	0 0 0 1 1 0 1 1 1 0 0 0	.....	0 1 0 0 0 0 0 0 1

cuciones del AG, y se seleccionan aquellos que generan mejores resultados.

**Definición de la función objetivo.** En segundo lugar, es necesario definir una función objetivo que nos clasifique estas soluciones o cromosomas —perfiles— en función de la bondad de los mismos con relación al objetivo planteado, que es encontrar la combinación de características que definen a los clientes que interesa retener y adquirir en la empresa porque maximizan nuestra rentabilidad. Podríamos definir la función objetivo teniendo en cuenta, por ejemplo, la cantidad gastada en compras de libros durante el período t de referencia por cada cliente y las coincidencias entre los perfiles propuestos por el AG y aquéllos más rentables de la base de datos real de clientes. Para ello podemos construir un indicador mixto que pondere el número de coincidencias entre el perfil propuesto por el AG y el del cliente por la cantidad gastada por dicho cliente. De esta forma, cuando el AG seleccione un perfil que coincida en gran medida con el de los

clientes que más dinero se han gastado, dicho perfil o cromosoma generará un valor alto en la función objetivo.

Por ejemplo, si la coincidencia de perfiles es total, es decir, los 68 genes del cromosoma seleccionado por el AG son idénticos a los de un determinado cliente, asignamos el 100% de la cantidad gastada por este cliente al valor de la función objetivo (si el cliente gasta mucho, el valor que toma la función objetivo será mayor que si gasta poco); si la coincidencia es nula, 0 genes de 68, asignamos el 0% de la cantidad gastada. Para el resto de coincidencias definimos una escala, que vendrá dada por el ratio  $100/68$  y que es igual a 1,47. Sumaremos entonces 1,47 por cada gen coincidente: si 47 genes de 68 del cromosoma seleccionado por el AG son iguales a los del perfil del cliente, el valor asignado será el 69 % ( $47 \times 1,47$ ) de la cantidad gastada por el cliente.

Con cada perfil (cromosoma) seleccionado por el AG se calculará este valor de la

CUADRO 8  
CÁLCULO FIGURADO DEL VALOR DE LA FUNCIÓN OBJETIVO DE CADA CROMOSOMA SELECCIONADO POR EL AG

	Personas										Secciones visitadas										Compras										Valor F.objetivo									
Cromosoma 1	0	0	0	1	1	0	1	0	0	1	1	1	.....	0	0	1	0	1	1	1	0	1	.....	0	0	0	0	1	0	1	1	1	87.965							
Cromosoma 2	0	1	1	1	0	0	1	0	0	0	1	0	.....	0	0	0	0	1	0	1	1	1	.....	0	0	0	0	1	0	1	1	1	49.035							
Cromosoma 3	1	1	0	1	0	1	1	1	1	0	1	1	.....	1	1	0	1	1	0	1	1	0	1	.....	1	1	0	1	1	0	1	1	126.500							
Cromosoma 4	1	1	1	0	1	1	1	0	1	0	1	0	.....	1	0	1	1	1	0	0	1	0	.....	1	0	1	1	1	0	0	1	0	13.568							
Cromosoma ...	.....																						.....																	
Cromosoma n	0	0	0	1	1	0	1	1	1	0	0	0	.....	0	1	0	0	0	0	0	0	0	.....	0	1	0	0	0	0	0	1	38.758								
Población	87.965+...+38.758																																							

CUADRO 9  
IMPLEMENTACIÓN DEL PROCEDIMIENTO RULETA EN EL AG (\*)



1	$f_{11} - 87965$
2	$f_{12} - 49035$
3	$f_{13} - 126500$
4	$f_{14} - 13568$
5	$f_{15} - 24325$
6	$f_{16} - 105200$
7	$f_{17} - 6852$
8	$f_{18} - 38758$
9	$f_{19} - 24157$
10	$f_{110} - 38758$
	$F - 515118$
	Siendo $F_1 = \sum f_{1k}$

peso cromos <sub>11</sub> = $f_{11} / F_1 = 17,1\%$
peso cromos <sub>12</sub> = $f_{12} / F_1 = 9,5\%$
peso cromos <sub>13</sub> = $f_{13} / F_1 = 24,6\%$
peso cromos <sub>14</sub> = $f_{14} / F_1 = 2,6\%$
peso cromos <sub>15</sub> = $f_{15} / F_1 = 4,7\%$
peso cromos <sub>16</sub> = $f_{16} / F_1 = 20,4\%$
peso cromos <sub>17</sub> = $f_{17} / F_1 = 1,3\%$
peso cromos <sub>18</sub> = $f_{18} / F_1 = 7,5\%$
peso cromos <sub>19</sub> = $f_{19} / F_1 = 4,7\%$
peso cromos <sub>110</sub> = $f_{110} / F_1 = 7,5\%$

(\*) Donde el primer subíndice indica el número de generación en el que nos encontramos, y el segundo, el número de cromosoma de la generación.

90

función objetivo para cada uno de los clientes<sup>4</sup>, sumándose después la de todos los clientes y obteniéndose de esta forma el valor de la función objetivo para el cromosoma que está analizando el AG en ese momento. Estos valores servirán para realizar la clasificación de todos los cromosomas en cada generación. De esta forma, el AG asignará una puntuación mayor a aquellos perfiles o cromosomas seleccionados por él mismo que coincidan en mayor medida con los de los clientes que mayor rentabilidad generan para la empresa (bajo la hipótesis de que aquellos que más gastan son los más rentables).

Así pues, se estima el valor de la función objetivo para cada uno de los cromosomas (perfiles) que conforman la primera generación del AG (expuesto en el cuadro 8) como el sumatorio para todos los clientes de la base de datos o de la muestra de «(Q × C)», donde Q es la cantidad gastada en libros en la empresa en el período t, y C es

1,47 multiplicado por el número de genes coincidentes entre el perfil propuesto por el AG y el del cliente:

$$f_z = \sum \forall_{\text{clientes}} [Q \times (1,47 \times n.^{\circ} \text{ genes del cliente coincidentes con los del cromosoma } Z)] \quad [1]$$

$$F = \sum \forall_{\text{cromosomas de la generación}} f \quad [2]$$

siendo  $f_z$  la función objetivo del cromosoma z, y F la de toda la población de la generación.

**El siguiente paso implica aplicar el operador de reproducción.** Las cadenas cromosómicas (perfiles posibles de clientes) de la primera generación del AG son reproducidas o copiadas de acuerdo con el valor que toma la función objetivo en las mismas. El AG para realizar este proceso puede ser implementado de múltiples formas, siendo la llamada «ruleta», una de las más comunes y que se estructura tal y como se expone a continuación.

Se construye una ruleta (ver cuadro 9) cuyas opciones son los diferentes perfiles o cromosomas que conforman la población de la primera generación del AG, con un peso en la ruleta igual al del valor de la función objetivo para los mismos (f) en proporción al valor total de la función objetivo para toda la población (F), es decir, los pesos en la ruleta vendrán dados por  $f/F$ . Este cuadro 9 muestra cómo se implementaría la ruleta para una población de 10 cromosomas o perfiles con unos datos figurados.

Una vez creada la ruleta, el proceso de reproducción se genera aleatoriamente sobre la base de la estructura de cromosomas que conforman la misma, de la siguiente forma (ver cuadro 10): se generan tantas tiradas aleatorias a la ruleta como cromosomas hay en la población (en este ejemplo serían 10 tiradas). Cuanto mayor proporción de la ruleta represente un cromosoma más probabilidad tendrá de ser elegido y por tanto de



CUADRO 10  
APLICACIÓN DEL PROCESO DE REPRODUCCIÓN MEDIANTE LA RULETA

Cromosomas	Cs 1	Cs 2	Cs 3	Cs 4	Cs 5	Cs 6	Cs 7	Cs 8	Cs 9	Cs 10
Pesos de los progenitores	17,1	9,5	24,6	2,6	4,7	20,4	1,3	7,5	4,7	7,5
Resultado de las tiradas aleatorias = selección sucesores	2	1	2	0	0	2	0	1	1	1

CUADRO 11  
IMPLEMENTACIÓN DE LOS MECANISMOS DE CRUZAMIENTO Y MUTACIÓN

Antes de la implementación																						
Cs 4	1	1	1	0	1	1	1	0	1	0	1	0	.....	1	0	1	1	1	0	0	1	0
Cs 1	0	0	0	1	1	0	1	0	0	1	1	1	.....	0	0	1	0	1	1	1	0	1
Cs n	0	0	0	1	1	0	1	1	1	0	0	0	.....	0	1	0	0	0	0	0	0	1
Cs 3	1	1	0	1	0	1	1	1	1	0	1	1		1	1	0	1	1	0	1	1	1
Cs....	.....																					
.....																						
Después de la implementación																						
Cs 4	1	1	1	0	1	1	1	0	0	0	1	0	.....	1	0	1	0	1	1	1	0	1
Cs 1	0	0	0	1	1	0	1	0	0	1	1	1	.....	0	0	1	1	1	0	0	1	0
Cs n	0	0	0	1	1	0	1	1	1	0	0	0	.....	0	1	0	1	1	0	1	1	1
Cs 3	1	1	0	1	0	1	1	1	1	0	1	1		1	1	0	0	0	0	0	0	1
Cs....	.....																					

pasar a la siguiente generación. Si el número de individuos de la población es suficientemente grande, los sucesores tenderán a ser proporcionales a los pesos de los progenitores en la ruleta, por lo que los progenitores con mayor valor en la función objetivo tendrán más descendencia que aquellos con menor valor en la función objetivo y por tanto con menor representación en la ruleta. En el cuadro se exponen los que podrían ser los resultados de la ejecución del proceso de reproducción.

Por el funcionamiento del mecanismo de reproducción, las nuevas generaciones tendrán, sucesivamente, una función conjunta (F) superior, ya que estarán formadas por aquellos individuos con mayor función objetivo individual (f).

**Aplicación de los operadores de cruzamiento y mutación.** A algunos de los individuos o cromosomas de la población se les aplicará los operadores de cruzamiento y de mutación (la proporción dependerá del valor que se le asigne a los parámetros de control de las probabilidades de ambos operadores) para generar

individuos con diferente información genética y ampliar de esta forma la búsqueda de perfiles diferentes, que de otra forma quedaría restringida a la presente en los cromosomas seleccionados aleatoriamente en la primera generación del AG.

Para ello, agrupamos aleatoriamente de dos en dos a los cromosomas «sucesores» y elegimos de forma aleatoria un punto de corte en las cadenas cromosómicas. Intercambiamos las cadenas de un porcentaje de las parejas de cromosomas o perfiles (dado por la probabilidad de cruzamiento) a partir de la posición de corte elegida. Posteriormente elegimos aleatoriamente diferentes genes para su mutación (el número vendrá dado por la probabilidad de mutación que hayamos asignado), intercambiando en éstos el valor 0 por el 1 ó viceversa. La descripción de estos procesos se expone en el cuadro 11.

Después de estas transformaciones, se calculan para la nueva población las funciones objetivo  $f_{2i}$  de los individuos o cromosomas de la segunda generación y la de la población en conjunto  $F_2$ , y se repiten ite-

rativamente los procesos de reproducción, cruzamiento y mutación para conseguir nuevas generaciones de cromosomas o perfiles con comportamientos cada vez mejores. El número de iteraciones (generaciones) es uno de los parámetros de control del AG y por tanto debe ser asignado en el proceso de diseño del mismo, antes de ejecutar el AG. En general, suelen utilizarse al menos 50 iteraciones o generaciones, pero el valor de este parámetro depende del número de individuos considerado en cada generación del AG y de la amplitud del problema a resolver (que se reflejará en el número de dígitos de la cadena cromosómica).

## Conclusiones

La aplicación anterior permitiría disponer de información sobre qué características o a que patrones diferentes responden aquellos clientes que en conjunto son más rentables para la empresa; es decir, de qué forma las características analizadas (edad, sexo, hijos, profesión, lugar de residencia, secciones visitadas en las páginas «online»

de la compañía, enlaces empleados para entrar, interacción mantenida con la librería y secciones en las que han realizado compras) se combinan en los clientes que nos aportan valor. Podría ser que descubriéramos que existen 3 tipos de perfiles muy rentables para la empresa, por ejemplo, como los siguientes:

**Perfil 1:** 22-28 años/hombre/economista/residentes en París/sin hijos/interés en Economía-Derecho-Informática-Deporte-Idiomas/entrada por publicidad en Yahoo/interacción por e-mail/compras realizadas en secciones de Informática-Deporte.

**Perfil 2:** 28-35 años/mujer/profesor de Universidad/residentes en Madrid/un hijo/interés en Matemáticas-Idiomas-Ocio/entrada por enlace en agregador/interacción por e-mail-teléfono/compras en Ocio-Matemáticas.

**Perfil 3:** más de 65 años/hombre/ingeniero/residente en Barcelona/ningún hijo a cargo/interés en Ocio-Salud-Otros/entrada por publicidad en página de Terra/interacción e-mail-teléfono/compras en Salud.

A partir de esta información, la empresa puede diseñar estrategias para captar clientes nuevos en esos segmentos —mediante campañas de marketing enfocadas a sectores específicos y sobre productos específicos, mediante publicidad en las páginas web que le generan mayor captación de clientes rentables— o para retener y maximizar la rentabilidad de los actuales —mediante, por ejemplo, el envío de información adaptada a cada perfil de cliente, sobre las novedades editoriales, promociones especiales, productos específicos, etc., o la utilización más amplia de los procedimientos de interacción que resultan más efectivos, etc. En definitiva, se trata de aportar valor al cliente a partir del conocimiento que sobre sus preferencias y hábitos tenemos y que hemos conseguido gracias a las posibilidades de Internet, para que él, a su vez, nos aporte rentabilidad.

Pero, además, una de las enormes ventajas de esta técnica es que permite el seguimiento de las preferencias de los clientes en el tiempo, lo que es muy importante, ya que es obvio que éstas van cambiando en función de las circunstan-

cias personales de cada uno. Los AG, una vez implementada una aplicación concreta de los mismos, permiten esta dinámica mediante su simple ejecución en un período posterior de la muestra. Es una metodología cuyo desarrollo inicial es algo laborioso, pero, una vez implementado el AG para un objetivo concreto, se puede ir ejecutando en el tiempo de forma que nuestro conocimiento, por ejemplo, sobre el comportamiento de los consumidores va evolucionando al mismo tiempo que dicho comportamiento lo hace.

## Notas

(1) Pueden también representarse directamente los parámetros de la solución correspondiente sin necesidad de ser codificados en binario, o utilizando otro tipo de codificación. Sin embargo, un gran número de aplicaciones utilizan dicha codificación por el hecho de que constituyó la base del desarrollo de los fundamentos matemáticos de los AG.

El equivalente binario de un número en sistema decimal se calcula dividiendo el número por dos de forma repetida, siempre que se pueda y tomando los restos de las divisiones, así como el último cociente de éstas. Por ejemplo, el equivalente binario de 8 es el vector {1, 0, 0, 0}, ya que  $8/2=4$ , siendo el resto 0,  $4/2=2$ , siendo el resto 0,  $2/2=1$ , siendo el resto 0.

A su vez, para pasar un número binario al sistema decimal debe multiplicarse cada valor del número binario por la potencia correspondiente del número 2, potencia relativa a la posición de cada valor en la cadena binaria: el valor correspondiente a la posición más a la derecha se multiplica por  $2^0$ , el valor de la posición que le sigue hacia la izquierda se multiplica por  $2^1$ , el siguiente por  $2^2$ , y así hasta finalizar la cadena binaria. Una vez aplicadas las correspondientes potencias se suman todos los valores obtenidos. En el ejemplo anterior {1, 0, 0, 0} la fórmula sería:

$$0 \times 2^0 + 0 \times 2^1 + 0 \times 2^2 + 1 \times 2^3 = 8$$

(2) Sería necesario tener en cuenta ciertas restricciones (obviadas en este cálculo), como, por ejemplo, que es imposible que una persona que ha comprado un libro en una determinada sección de la librería «online» no haya visitado dicha sección o el caso poco probable de que existan compradores menores de 16 años con más de tres hijos, etc.

(3) Número máximo de perfiles diferentes a los que tiene acceso el AG y que es de, aproximadamente, 295.147.905 millones

(4) En vez de usar la base de datos de clientes completa se puede trabajar con una muestra estadísticamente representativa de la misma, lo que implica reducir el tiempo computacio-

nal necesario para ejecutar el AG, esto es especialmente útil en casos como el de AMAZON.COM, cuya base de datos puede contener varios millones de registros.

## Bibliografía

- ALLEN, F. and KARJALAINEN R. E. (1999): «Using genetic algorithms to find technical trading rules». *Journal of Financial Economics*, págs. 245-271.
- BARBA-ROMERO, S. y OLMEDA, I. (1993): *Redes neuronales artificiales: fundamentos y aplicaciones*, Servicio de publicaciones de la Universidad de Alcalá.
- BAUER, R. J. Jr. (1994): *Genetic algorithms and investment strategies*, John Wiley & Sons, Inc.
- CEDEÑO, W. and VEMURI, V. R. (1997): «Database design with genetic algorithms», en *Evolutionary algorithms in engineering applications*, DASGUPTA and MICHALEWICZ, Springer, Alemania.
- DASGUPTA, D. and MICHALEWICZ, Z. (1997): *Evolutionary Algorithms in Engineering Applications*, Springer-Verlag, Germany.
- DAVIS, L. (1991): *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, Nueva York.
- FARLEY, A. M. and JONES, S. (1994): «Using a genetic algorithm to determine an index leading economic indicators», *Computational Economics*, 7, Kluwer Academic Publishers, Netherlands, págs. 163-173.
- FOGEL, D. B. (2000): *Evolutionary Computation. Toward a new philosophy of machine intelligence*, 2.<sup>a</sup> edición, The Institute of Electrical and Electronics Engineers, Inc. Nueva York.
- GOLDBERG, D. E. (1989): *Genetic algorithms in search, optimization and machine learning*, Addison-Wesley, Reading, Massachusetts.
- HOLLAND, J. H. (1975): *Adaptation in Natural and Artificial Systems: an introductory analysis with applications to biology, control and artificial intelligence*, University of Michigan Press, Ann Arbor, MI.
- KRÖSE, B. J. A. and VAN DER SMAGT, P. P. (1991): *An introduction to Neural Networks*, 4.<sup>a</sup> edición, University of Amsterdam.
- MARKS, R. E. (1989): «Breeding hybrid strategies: optimal behaviour for oligopolists», *Proceedings of the Third International Conference on Genetic Algorithms*, Ed. SCHAFER, J. David. Morgan Kaufmann Publishers, Inc. págs. 198-207.
- MICHALEWICZ, Z. (1996): *Genetic Algorithms + Data Structures = Evolution Programs*, 3.<sup>a</sup> edición, Springer-Verlag, Nueva York.
- VARETTO, F. (1998): «Genetic Algorithms applications in the analysis of insolvency risk», *Journal of Banking and Finance*, 22, Elsevier, págs. 1421-39.