Predicción del CTR de los anuncios de Internet usando redes orgánicas artificiales

Luis Miralles Pechuán, Hiram Ponce

Facultad de Ingeniería, Universidad Panamericana,
México

{lmiralles,hponce}@up.edu.mx

Resumen. Para que las redes publicitarias aumenten sus ingresos es necesario darle prioridad a los anuncios más rentables. El factor más importante en la rentabilidad de un anuncio es el CTR (Click-through-rate) que es la probabilidad de que un usuario realice un clic en el anuncio de una página web. Para predecir el CTR, hemos entrenado varios modelos de clasificación supervisados y hemos comparado su rendimiento con las redes ROA (Redes Orgánicas Artificiales). La conclusión es que estas redes son una buena solución para predecir el CTR de un anuncio.

Palabras clave: predicción de CTR, redes orgánicas artificiales en publicidad, modelos supervisados de clasificación, redes de publicidad CPC.

1. Introducción

Desde que se mostrara el primer banner en octubre de 1994, la publicidad en Internet ha tenido un crecimiento constante [1]. La infraestructura y la tecnología de la red permiten lanzar campañas con importantes ventajas para los anunciantes que la publicidad tradicional no puede ofrecer.

Una de las principales ventajas es que permite a los anunciantes segmentar los clientes en función de ciertos parámetros como: localización, horario, aficiones o tipo de dispositivo desde el que se accede. Esto se conoce como *microtargeting* y permite lanzar campañas a un pequeño grupo con intereses comunes. Estas campañas resultan muy eficaces [2].

La publicidad en Internet tiene una amplia variedad de tipos de anuncios como son los *pop-up*, los *pop-under* o los *interstitial* [1]. Pero los anuncios más utilizados debido a su gran impacto y a que no son intrusivos son los banners, los anuncios de texto y los videos.

[1] Los pop-up son ventajas emergentes, los pop-under son nuevas ventanas del navegador que se abren de forma automática y los anuncios interstitial son animaciones que suelen ocupar toda la página.

El modelo de publicidad en Internet está formado por cuatro roles claramente diferenciables. Los anunciantes, la red publicitaria, los editores y los usuarios. Los anunciantes son las empresas que pagan a la red de publicidad para que se muestre su anuncio en las páginas de los editores. Los editores son las personas que tienen al menos una página web y que perciben una retribución económica por prestar un espacio de sus páginas en el que se muestran los anuncios. Los usuarios son personas que navegan por Internet y que compran un producto cuando les interesa.

La red de publicidad es la encargada de gestionar todo el proceso de publicidad. Entre las principales tareas se encuentran: la detección de fraude [3], la preservación de la privacidad de los roles y la gestión de pagos. Pero sin duda, la tarea más importante consiste en optimizar el rendimiento del modelo publicitario para que los anunciantes tengan campañas exitosas y para que tanto la red de publicidad como los editores obtengan el máximo rendimiento por sus anuncios.

En la publicidad online existe una multitud de modelos de cobro a los anunciantes como son el pago por mostrar un anuncio un cierto número de días, el pago por enlaces que apunten a la web del anunciante o el pago por canjeo de cupones[2]. Pero los modelos más utilizados con gran diferencia son el CPM, el CPC y el CPA.

En el modelo CPM *(Cost-per-mile)* los anunciantes pagan por cada mil impresiones[3] de un anuncio. En el modelo CPC *(Cost-per-click)* se paga cada vez que un usuario hace clic sobre un anuncio y por último, en el modelo CPA *(Cost-per-acquisition)* se paga cuando un usuario que hace clic en el banner de un editor realiza una compra o contrata un servicio.

El modelo más utilizado es el CPC, principalmente porque es el más sencillo de entender para los anunciantes y porque lo han adoptado las principales empresas del sector[4]. Para optimizar el rendimiento de la redes CPC hemos de calcular la rentabilidad de los anuncios en las campañas que lanzan los anunciantes.

Dentro del modelo de cobro CPC existen dos conceptos relacionados con el pago de los anunciantes que son el CPC Máximo y el CTR. El CPC Máximo representa el importe más alto que un anunciante está dispuesto a pagar por un clic. Sin embargo, en la mayoría de los casos se le cobra un precio menor que se denomina CPC Real. El CTR *(Click-through rate)* de un anuncio se define como el ratio entre el número de clics en el anuncio y el número impresiones. Conocer con total seguridad el CTR [4] que tendrá un determinado anuncio en el futuro no es posible, pero si logramos predecir el CTR de los anuncios con un margen de error muy pequeño podremos elegir el anuncio que proporcione mayores beneficios.

Existen pocas publicaciones respecto a los algoritmos utilizados por las redes publicitarias[5] [5]. Esto es algo muy razonable puesto que si se publicasen estos

[2] El canjeo de cupones consiste en repartir un código asociado con un editor que al asignarle le otorga una comisión.

[3] Una impresión significa que un anuncio se muestra en alguna página.

[4] Las principales empresas del sector son Google, Yahoo y Microsoft.

[5] Si bien es cierto que existe un famoso artículo de Tuzhilin sobre el sistema de detección de fraude de Google, no hemos encontrado documentación sobre los algoritmos que utilizan las redes de publicidad.

algoritmos las empresas perderían la ventaja competitiva. Ya que cualquiera podría copiar las investigaciones que conllevan tantos años y tantos recursos económicos.

Por otra parte, a las personas con intención de cometer fraude se les facilita vulnerar las redes publicitarias.

Para calcular el CTR utilizaremos modelos supervisados de tipo clasificación. La creación de modelos supervisados es una parte importante del *Data Science*. El *Data Science* consiste en obtener información de utilidad para las organizaciones a partir de datos recolectados. Las técnicas que se aplican provienen de ciencias como lógica, estadística, probabilidad, diseño de modelos o reconocimiento de patrones.

Se han desarrollado numerosas herramientas para simplificar y automatizar estos procesos. Entre las herramientas de software libre más completas y con mejores resultados se encuentra R Studio. Utilizaremos esta herramienta para predecir resultados con diversos métodos como: Splines de regresión multivariante de adaptación, Centroide disminuido más cercano o Análisis discriminante lineal, y posteriormente utilizaremos las ROA (Redes Orgánicas Artificiales) para contrastar los resultados.

Primeramente, entrenaremos las redes ROA y el reto de los modelos con un dataset. Una vez entrenados, probaremos la eficacia de estos con un conjunto de entradas y posteriormente evaluaremos los modelos.

2. Redes de hidrocarburos artificiales y sus aplicaciones en la predicción CTR

En esta sección explicamos brevemente la técnica de aprendizaje automático llamada redes de hidrocarburos artificiales *(artificial hydrocarbon networks, AHN)* y posteriormente exponemos el modelo de predicción CTR que proponemos utilizando este método.

2.1. Redes de hidrocarburos artificiales

La técnica de redes orgánicas artificiales *(artificial organic networks, AON)* es una clase de algoritmos inspirados en los compuestos de química orgánica, la cual permite el empaquetamiento de información (patrones de datos) en módulos llamados moléculas. Además, esta técnica define mecanismos similares a los compuestos químicos orgánicos (heurísticas) que generan estructuras organizadas y óptimas en términos de la energía química. Finalmente, las ROA preservan algunas características químicas como: modularidad, herencia, organización y estabilidad estructural [6].

Con ayuda del método explicado anteriormente, la técnica de redes de hidrocarburos artificiales es una clase de algoritmos de aprendizaje supervisado basados en las ROA. De hecho, el algoritmo está inspirado en los hidrocarburos. De manera similar a los hidrocarburos químicos, las redes de hidrocarburos artificiales únicamente utilizan dos unidades: átomos de hidrógeno y de carbono que pueden ser relacionadas como máximo con uno o cuatro átomos, respectivamente. La unión de estos átomos en unidades pequeñas se conoce como moléculas y la unión de estas

últimas se conocen como compuestos, los cuales incluyen relaciones no lineales entre moléculas. Este tipo de redes son adecuadas para problemas de modelado, así como de sistemas de predicción y búsqueda de patrones en datos desconocidos e inciertos. Algunos de los ejemplos de aplicación de esta técnica pueden encontrarse en la bibliografía [7], [8].

2.2. Modelo de predicción CTR basado en las redes de hidrocarburos artificiales

Nosotros proponemos el uso de redes de hidrocarburos artificiales como un método de aprendizaje supervisado para la construcción de un modelo de predicción CTR. Para esta aplicación, decidimos el uso de redes de hidrocarburos artificiales formadas de un compuesto lineal y saturado que reciba múltiples entradas (categorías de un conjunto de datos especificado) y genere una salida (el valor CTR). La Figura 1 muestra nuestra propuesta del modelo de predicción CTR usando AHN.

Como se muestra en la Figura 1, el modelo de predicción CTR es entrenado usando una base de datos con información del tipo atributo-valor. El proceso de entrenamiento se compone del uso del algoritmo de redes de hidrocarburos artificiales multi-categorías. Después, se puede llevar a cabo una petición basada en el mismo tipo de atributos descritos en la base de datos. Finalmente, el modelo de redes de hidrocarburos artificiales genera el valor respectivo de CTR.

En la siguiente sección, probamos nuestro modelo de predicción CTR basado en las redes de hidrocarburos artificiales utilizando un conjunto de datos de publicidad con su valor CTR esperado.

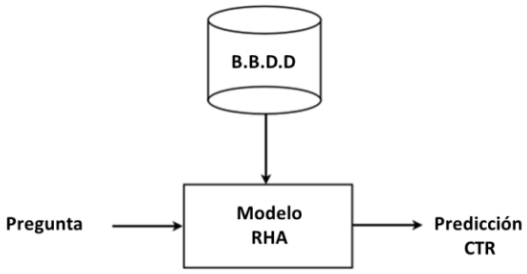

Fig. 1. Nuestra propuesta del modelo de predicción CTR basado en las redes de hidrocarburos artificiales.

2.3. Predicción del CTR del resto de algoritmos mediante Caret de R Studio

R Studio es un entorno de programación orientado al cálculo científico. Esta plataforma permite a los usuarios crear paquetes que implementan un conjunto funciones o métodos matemáticos. Caret es un paquete diseñado para simplificar el proceso de construcción y evaluación de modelos. Caret dispone de un total de 180

métodos para la construcción de modelos predictivos de los cuales 138 son de clasificación y 42 de regresión [9].

Esta herramienta dispone de métodos que automatizan la selección de variables. Esto permite construir modelos más precisos, menos complejos y en menor tiempo. Los métodos para la selección de variables descartan aquellas características que no aportan información valiosa o que son redundantes. Para este problema utilizaremos un método llamado *Recursive Feature Elimination* [10]. Este método lo aplicamos a todos los modelos menos para el ROA, puesto que este modelo integra un método alternativo llamado PCA (Análisis por componentes principales).

El método RFE es de tipo *Wrapper*, es decir, tiene como entradas una combinación de predictores que utiliza para construir modelos. Y como salida una métrica que evalúa la precisión del modelo mediante el método *random forest*. Esto sirve para ir probando combinaciones de variables hasta lograr la mejor combinación.

La principal ventaja de Caret es la simplicidad del código para construir modelos con distintos métodos y con distintas configuraciones de parámetros. Así como la implementación de una serie de métricas para seleccionar automáticamente la mejor configuración de parámetros de varios modelos que se generan con un método. Para medir la precisión del modelo se suelen utilizar métricas como el ROC, el AUC, la *Accuracy* o el error RMSE.

Cada método tiene un número de parámetros determinado que suele estar entre cero y tres. Por ejemplo, las redes neuronales tienen dos parámetros que son el número de nodos y el nivel de aprendizaje. Caret hace un barrido de posibles valores de los parámetros pero también permite configurar esos parámetros por el programador [11].

Para evaluar los métodos se suelen hacer varios data sets y para ello se utiliza el CV *(Cross-validation)*. El método CV sirve para crear distintos training set de forma que podemos construir modelos con el mismo método pero con muestras distintas. En lugar de tomar la precisión de un solo modelo calcula el promedio de todos y de forma que los resultados son más fiables.

Por último, R Studio cuenta con un conjunto de librerías que permite crear gráficos muy ilustrativos y con gran calidad. Esto permite visualizar los resultados obtenidos de una forma mucho más comprensible.

3. Metodología para la predicción de CTR mediante métodos supervisados

Para resolver el problema que se plantea hemos utilizado un dataset proporcionado por la página web: Kaggle.com[6]. Este dataset recoge algunos parámetros de las visitas de los usuarios a la web www.criteo.com durante un periodo de siete días. Sobre estas visitas, se ha reducido en mayor medida el número de muestras en la que los usuarios no hicieron clic en el anuncio que los que sí lo hicieron.

[6] Esta web se llama Kaggle.com y es famosa por que organiza muchos concursos a nivel mundial.

Las visitas vienen representadas en una tabla donde cada fila representa la visita de un usuario y cada columna representa una característica de un usuario o de la página web. Por ejemplo, en la primera columna de la tabla se representa con "0" o con "1" si el usuario hizo clic en el anuncio. La tabla tiene trece columnas con valores de tipo entero y 26 valores de tipo string que representan ciertas categorías. Los valores de las categorías han sido codificados mediante un hash a un valor de 32 bits para garantizar la privacidad. Las filas están ordenadas cronológicamente y cuando el valor de cierto parámetro se desconoce, simplemente se deja un espacio en blanco.

Para realizar el test de prueba utilizaremos un fichero con el mismo formato que la tabla de entrenamiento pero sin la columna que indica si el usuario hizo clic. Para evaluar el rendimiento de las ROA en la predicción del CTR utilizaremos dos métricas. La primera métrica está basada en el logaritmo de la función de probabilidad para una distribución aleatoria de Bernoulli y la segunda métrica se basa en algo tan sencillo como el porcentaje de aciertos sobre el total de pruebas.

Para entrenar las redes se eligieron un total de 1,000,000 muestras aleatorias, ya que las ROA no soportan excesiva información. De este conjunto de datos, utilizamos el 80% de observaciones aleatorias para el proceso de entrenamiento. El conjunto de datos de entrenamiento se depuró, es decir, se quitaron los datos incompletos. Estas características se eliminaron con ayuda del PCA. Este método selecciona y ordena las columnas más importantes. Posteriormente los datos se estandarizaron usando (1), donde x es una columna del dataset, μ es la media del valor de dicha columna, σ es la desviación estandar de dicha columna, y x_{std} es la columna estandarizada.

Para desarrollar el modelo de predicción CTR la red orgánica se creó un compuesto de 100 moléculas artificiales empleando la técnica de estimación por mínimos cuadrados. Esta red se entrenó con un coeficiente de aprendizaje de 0.5. El coeficiente de aprendizaje es un parámetro de las redes de hidrocarburos artificiales que permite regular la tasa de asimilación o regresión de los datos reales.

$$x_{std} = \frac{x - \mu}{\sigma} \quad (1)$$

4. Resultados obtenidos

En la primera métrica, la probabilidad de que un usuario haga clic se expresa en el rango [0,1]. Cuanto menor sea la estimación de que el usuario no haga clic más se aproximará a cero y cuanta mayor probabilidad estimemos que el usuario hará clic más se aproximará a uno.

Una vez que hayamos establecido los resultados calcularemos el Log Loss del modelo mediante la siguiente fórmula:

$$\text{Log Loss} = -1n\sum i=1n\ [\ y_i \log(y_i) + (1-y_i)\log(1-y_i)\ ]$$

El uso del logaritmo permite que una equivocación no tenga un castigo excesivo en el resultado general. Con esta fórmula si decimos que la probabilidad de que haga clic es de un 99% y acertamos nos dará una gran recompensa y si fallamos tendremos una gran penalización.

Los resultados obtenidos con el logaritmo de la función de probabilidad para una distribución aleatoria de Bernoulli han sido muy positivos puesto que hemos obtenido un resultado de Log Loss = 0.6499.

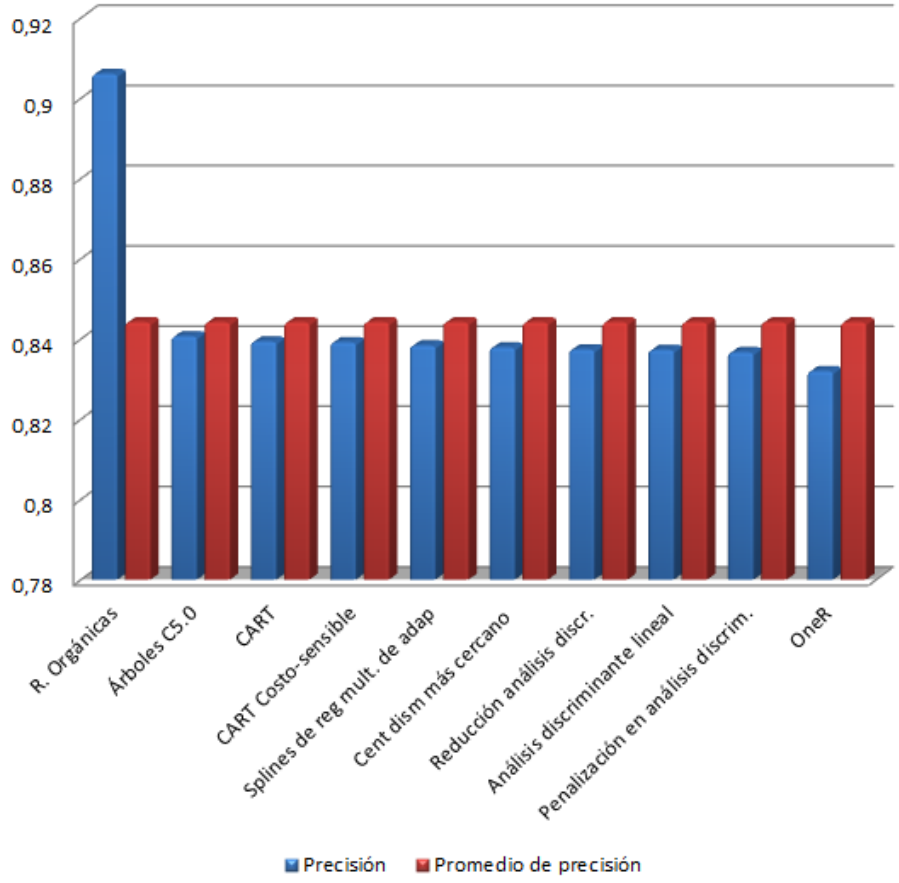

Fig. 2 Gráfica con los resultados de la precisión y el tiempo en horas de las redes.

Vamos a utilizar una segunda métrica más fácil de interpretar. Para medir la eficacia del algoritmo lo expresaremos como el ratio del número de aciertos entre el número de casos totales. Diremos que el usuario hará clic en caso de que la probabilidad estimada sea superior a 0,5 y diremos que no en caso de que sea inferior. El sistema de predicción puede arrojar cuatro posibles resultados:

- Verdadero positivo (VP): Predecimos que el usuario hará clic y el usuario lo hace.
- Verdadero negativo (VN): Predecimos que el usuario no hará clic y si lo hace clic.
- Falso Positivo (FP): Predecimos que el usuario hará clic y no hace clic.
- Falso Negativo (FN): Predecimos que el usuario no hará clic y si hace clic.

Si tenemos en cuenta que VP + VN + FP + FN = N. Entonces la tasa de precisión de nuestro sistema es igual a (VP + VN) / N y la tasa de error de (FP + FN) / N.

Los resultados obtenidos en esta métrica han sido VP = 19,597, VN = 161,614, FP = 14,935 y FN = 3,854. Por lo que la precisión ha sido de 90.61% y la Tasa de Error = 9.39%.

En la tabla 1 mostramos los resultados obtenidos con los distintos métodos supervisados. Estos modelos los hemos creado con el paquete Caret de R Studio y con el método "repeatedcv", con 10 particiones y 3 repeticiones. Estos modelos se han creado con 800,000 muestras y la precisión del modelo se ha medido con 200,000 muestras del mismo modo que las redes ROA.

Tabla 1. Nuestra propuesta del modelo de predicción CTR basado en las redes de hidrocarburos artificiales.

	R. Orgá-nicas	Árboles C5.0	CART	CART Costo-sensible	Splines de reg mult. de adap	Cent dism más cercano	Reduc. análisis discr.	Análisis discrim lineal	Penal. en análisis discrim.	OneR
Precisión	0,9061	0,8407	0,8394	0,8392	0,8384	0,8379	0,8374	0,8373	0,8367	0,832
Tiempo construir modelo	0:00:12	0:00:26	0:00:10	0:00:26	0:00:28	0:00:03	0:00:03	0:00:02	0:00:05	0:00:07

En la Figura 2 mostramos la precisión de cada uno los modelos en comparación con el promedio de todos los métodos.

5. Conclusión

Consideramos que las ROA son un modelo de datos muy efectivo y que se pueden adaptar a muchos tipos de problema. Además tienen la ventaja de que permiten crear modelos en un tiempo reducido.

Las ROA permiten una predicción de CTR de aproximadamente el 90% de precisión lo cual es un buen resultado. Una mejora podría ser aplicar una heurística que mejorara los resultados de la estimación por mínimos cuadrados. También se podrían utilizar técnicas híbridas de aprendizaje (p.e., redes bayesianas para datos incompletos, árboles AVL para mejorar la búsqueda de información, etc.), ya que una sola técnica limita el potencial de aprendizaje en problemas reales.

Una desventaja de las ROA es que no soportan ser entrenadas por un gran número de muestras. Una solución para esto sería entrenar un conjunto redes orgánicas. Por ejemplo crear 1,000 redes orgánicas artificiales y predecir el resultado como el promedio de todas las redes.

Por otra parte, consideramos que este tipo de redes pueden ser un buen aliado en la lucha contra el fraude que se produce en la publicidad en Internet. Existen muchas

amenazas para las redes como son los *botnets* [7] [12], las granjas de clics[8] o los propios editores[9]. Todas las redes de publicidad están expuestas a este tipo de amenazas y muchas de ellas son muy difíciles de detectar. Las ROA son muy útiles para la detección basada en anomalías. Es decir, en primer lugar se podrían entrenan las redes con parámetros que representen el comportamiento habitual de los usuarios. Una vez entrenadas, cuando las redes perciban un comportamiento distinto al acostumbrado podrán mandar un mensaje de alarma.

La principal ventaja que tienen las ROA en este campo es que son capaces de señalar el parámetro por el cual han considerado la visita no es válida. Es decir, que además de descartar el clic nos proporciona cierta información sobre esta decisión. Junto con las ROA, sería conveniente utilizar algunos filtros que comprueben si existen demasiadas IPs de la misma zona, usuarios reincidentes o un excesivo número de clics en un determinado anuncio.

6. Trabajos futuros

Las ROA son modelos predictivos muy eficaces y que tienen utilidad para un gran número de problemas. Un proyecto interesante con este tipo de redes sería crear un paquete para R Studio o para Matlab. Esto facilitaría que muchas personas las usaran y pudiesen comprobar su precisión frente a otros modelos de una forma sencilla.

Otra mejora consistiría en crear un conjunto de predictores para aumentar el porcentaje de aciertos o de predicciones tanto para clasificación como para regresión. En el caso de clasificación la clase vendrá determinada por la clase que predigan la mayoría de los modelos. En el caso de los modelos de regresión, el valor que se prediga será el promedio de los valores que predigan todos los modelos.

Las redes orgánicas artificiales pueden ser muy útiles para la detección de fraude en la publicidad en Internet. Estas redes podrían entrenarse con visitas normales realizadas por los usuarios y en el caso de que detectaran un comportamiento anómalo avisaran a la red publicitaria. También se puede hacer a la inversa, es decir, entrenar las redes con las visitas de los usuarios fraudulentos y en caso de que sea parecido notificar a la red de publicidad.

Otra línea de investigación futura podría ser hacer una estimación del tiempo necesario para construir el modelo con ROA. De este modo, si el tiempo es muy alto el usuario podrá abortar la operación o planificarse teniendo en cuenta el tiempo de cálculo.

Para ello, deberíamos crear un modelo con varias muestras que tuviera como entradas: el número de muestras, el número de procesadores, la memoria RAM, la velocidad de los procesadores, el sistema operativo y el número de entradas. Este algoritmo se podría ir actualizando según se vaya ejecutando el programa de la misma

[7] Los botnets son pequeños programas que se insertan en las computadoras de los usuarios y que pueden navegar y hacer clicks sin que el usuario lo sepa.

[8] Las granjas de enlaces son equipos de personas que pueden estar en países como la india y que se dedican a hacer click para arruinar campañas de publicidad.

[9] Los editores pueden hacer trampas para aumentar sus propios ingresos y muchas veces piden ayuda a otras personas para hacerlo.

forma que lo hacen las descargas. Pues no tenemos garantizado que el 100% del procesador vaya a ser para este programa.

Referencias

1. Coopers, P. W. H., IAB internet advertising revenue report. URL: http://www.iab.net/insights_research/industry_data_and_landscape/adrevenuereport (2014)
2. Moe, W. W.: Targeting display advertising. Advanced database marketing: Innovative methodologies & applications for managing customer relationships. Londres: Gower Publishing (2013)
3. Stone-Gross, B., Stevens, R., Zarras, A., Kemmerer, R., Kruegel, C., Vigna, G.: Understanding fraudulent activities in online ad exchanges. In: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, pp. 279–294, ACM (2011)
4. McMahan, H. B., Holt, G., Sculley, D., Young, M., Ebner, D., Grady, J. Kubica, J.: Ad click prediction: a view from the trenches. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1222–1230, ACM (2013)
5. Tuzhilin, A.: The lane's gifts v. google report. Official Google Blog: Findings on invalid clicks. pp. 1–47 (2006)
6. Ponce, H., Ponce, P., Molina, A.: Artificial Organic Networks: Artificial Intelligence Based on Carbon Networks. Studies in Computational Intelligence, Vol. 521, Springer (2014)
7. Ponce, H., Ponce, P., Molina, A.: A New Training Algorithm for Artificial Hydrocarbon Networks Using an Energy Model of Covalent Bonds. In: 7th IFAC Conference on Manufacturing Modelling, Management, and Control, Vol. 7(1), pp. 602–608 (2013)
8. Ponce, H., Ponce, P.: Artificial Organic Networks. In: IEEE Conference on Electronics, Robotics, and Automotive Mechanics CERMA, pp. 29–34 (2011)
9. Kuhn, M.: Building predictive models in R using the caret package. Journal of Statistical Software, 28(5), 1–26 (2008)
10. Granitto, P. M., Furlanello, C., Biasioli, F., Gasperi, F.: Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometrics and Intelligent Laboratory Systems, 83(2), 83–90 (2006)
11. Kuhn, W., Wing, J., Weston, S., Williams, A., Keefer, C., et al.: Caret: classification and regression training. R package, v515 (2012)
12. Miller, B., Pearce, P., Grier, C., Kreibich, C., Paxson, V.: What's clicking what? Techniques and innovations of today's clickbots. In: Detection of Intrusions and Malware, and Vulnerability Assessment, pp. 164–183, Springer (2011)