Brazilian Journal of Development

Análisis de algoritmos aplicados al Churn Analysis

Analysis of algorithms applied to Churn Analysis

Recebimento dos originais: 05/02/2019
Aceitação para publicação: 08/03/2019

Verónica Fabbro
Ingeniera en Sistemas de Información
Instituição: Universidad Tecnológica Nacional (UTN - FRBA)
Endereço: Av. Medrano 951 - Ciudad Autónoma de Buenos Aires, Argentina
E-mail: vnfabbro@gmail.com

Ariel Deroche
Ingeniero en Sistemas de Información
Instituição: Universidad Tecnológica Nacional (UTN - FRBA)
Endereço: Av. Medrano 951 - Ciudad Autónoma de Buenos Aires, Argentina
E-mail: arielderoche@gmail.com

Diego Basso
Magíster en Ingeniería de Sistemas de Información, Ingeniería informática, Ingeniero en Informática.
Instituição: Universidad Tecnológica Nacional (UTN - FRBA)
Endereço: Av. Medrano 951 - Ciudad Autónoma de Buenos Aires, Argentina
E-mail: diebasso@yahoo.com.ar

Florencia Pollo-Cattaneo
Directora del Grupo. Doctora en Ciencias Informáticas por la Facultad de Informática de la Universidad Nacional de La Plata. Magíster en Ingeniería en Software – Doble Titulación: Universidad Politécnica de Madrid / Instituto Tecnológico de Buenos Aires. Especialista en Construcción de Sistemas Expertos (ITBA). Ingeniera en Sistemas de Información por la UTN FRBA.
Instituição: Universidad Tecnológica Nacional (UTN - FRBA)
Endereço: Av. Medrano 951 - Ciudad Autónoma de Buenos Aires, Argentina
E-mail: flo.pollo@gmail.com

ABSTRACT

This meta-paper describes the style to be used in articles and short papers for SBC conferences. For papers in English, you should add just an abstract while for the papers in Portuguese, we also ask for an abstract in Portuguese (“resumo”). In both cases, abstracts should not have more than 10 lines and must be in the first page of the paper.

Keyword: Algoritmos, churn analysis, fuga de clientes, minería de datos, retención de clientes.

RESUMEN

Las empresas diseñan constantemente estrategias para ser más eficientes y posicionarse mejor dentro de su categoría. Para ello necesitan analizar diferentes indicadores que los

Brazilian Journal of Development

ayuden a tomar decisiones a mediano y largo plazo. Algunos de los objetivos empresariales a los cuales hay que dar respuesta son: cómo aumentar las ganancias, disminuir los costos, cómo aumentar o mantener a los clientes. Las empresas necesitan conseguir y mantener clientes. La captación de nuevos clientes a veces desvía el foco de otros aspectos sumamente importantes como puede ser la satisfacción de los clientes actuales. Ante un escenario en donde ganar un nuevo cliente cuesta cinco veces más que retenerlo, se hace imperiosa la necesidad de predecir con antelación la posible fuga de clientes, la segmentación en grupos, la identificación de factores de baja de los mismos, la detección de perfiles de clientes, entre otras, de tal manera de establecer un adecuado plan de contingencia para la empresa. Para ello resulta necesaria la aplicación de procesos, identificando y seleccionando los algoritmos de minería de datos más apropiados. La importancia de poder relevar los distintos algoritmos a ser tenidos en cuenta en diversas implementaciones constituye un paso hacia su entendimiento y posibilidad de ponerlos en práctica.

Palabra clave: Algoritmos, análisis de abandono, fuga de clientes, extracción de datos, retención de clientes.

1 INTRODUCCIÓN

La minería de datos (data mining) resulta ser esencial para proporcionar el conocimiento de poblaciones enteras de datos basándose en el análisis de los mismos e históricos. Relacionada a la tecnología, se encuentra en el ámbito de la programación asociándose con los algoritmos a implementar [1].

Teniendo en cuenta las grandes cantidades de datos que hoy en día acumulan las empresas, se hace imperiosa la necesidad de analizarlos para obtener un provecho de cara a cumplir con objetivos redituables para las mismas [2].

Entre las distintas estrategias de negocio de las compañías, la retención de clientes es una de ellas y las causas por las cuales éstos dejan de consumir los servicios son múltiples. Los clientes prestan atención a la experiencia, servicio personalizado, la diversidad y agilidad que les brindan las compañías, intensificando la competitividad entre ellas [3]. Se calcula que el costo de obtener un nuevo cliente es 5 veces más alto que mantener a un cliente existente [3, 4, 5, 6]. De aquí surge el término “churn analysis” o “attrition analysis”, son técnicas de minería de datos aplicadas a predecir la fuga de clientes o, explicar las causas de por qué los mismos se van de la compañía. El objetivo principal de este tipo de análisis es retener al cliente, para que éste no desista de los servicios o elija la competencia [5, 6, 7]. En este contexto, surge la tasa de abandono de los clientes (churn rate), calculándose de la siguiente manera:

(Número de clientes perdidos en un período determinado) / (Número de clientes al comienzo de ese período) x100. Los períodos más utilizados para el cálculo de la tasa de abandono son: mensual, trimestral y anual [8].

Así, el objetivo principal del análisis es identificar un grupo de clientes con alta probabilidad de fuga, para que la empresa tome las decisiones y acciones necesarias para retener a aquellos que valgan la pena. Empresas de las industrias de seguro, telecomunicaciones y bancarias enfrentan una intensa competencia, deseando mantener a tantos clientes como sea posible debido al alto costo de inversión en marketing para captar un nuevo cliente. Es por ello que estos rubros son los que más se abocan al estudio de retención de clientes, descubriendo así patrones de fuga para ayudar a los gerentes de marketing a comprender los motivos de la rotación de clientes, y mejorando a su vez la comunicación y relación con los mismos [9,10].

Por otro lado, el proceso de data mining aplicado al análisis de fuga puede describirse mediante los pasos que se describen a continuación [11] (los mismos se encuentran enmarcados dentro de la metodología CRISP-DM [14]):

1. Definición del problema: formular el problema de la compañía en términos de análisis de fuga.

2. Revisión de datos y selección inicial.

3. Formulación de problemas en términos de datos existentes.

4. Recopilación de datos, catalogación y formato.

5. Procesamiento de datos (limpieza de datos, generación de nuevas variables, transformaciones de los datos, análisis estadísticos, análisis de sensibilidad, detección de pérdidas, etc.).

6. Modelización de datos a través de modelos de clasificación.

7. Revisión y análisis de los resultados: utilizar el modelo de minería de datos para predecir los clientes que están por fugarse por sobre los clientes actuales.

8. Despliegue de resultados: el resultado final del modelo es un grupo de clientes clasificados como “churn”.

El término “customer churn” [7], refiere a identificar aquellos clientes que van a dejar la compañía. Este concepto puede variar dependiendo la naturaleza de la compañía o los distintos servicios que brinde, y por dicha razón, el análisis de desgaste está ligado con la definición que la compañía le dé al “churn”. La empresa debe definir qué significa que un cliente deje sus servicios. Por ejemplo, una empresa que brinde servicios a través de un

contrato puede definir como “churn” a la baja del mismo y, por ende, el fin de su prestación de servicios con el cliente. Por otro lado, otra empresa que no opere con modalidad contractual (como una empresa de “retail”), mide la frecuencia de compra de sus clientes y el monto de la misma y define un umbral para ambos; cuando un cliente esté debajo de este umbral el mismo se definirá como que ha dejado de consumir en la compañía. [12].

La lealtad del cliente es el término opuesto al customer churn, y se mide como: Lealtad del cliente = 1 - Tasa de abandono de los clientes (Customer loyalty = 1-Churn rate Modeling). Un cliente leal es aquel que compra con frecuencia, y exhibe un patrón de comportamiento regular.

En este contexto, se detecta la necesidad de considerar oportuna una comparación de los diferentes algoritmos existentes para aplicar al concepto mencionado de Churn Analysis, sirviendo así de guía recopiladora para poder encarar futuras investigaciones a fin de facilitar su posterior aplicación dependiendo el caso.

El presente trabajo introduce la definición del problema (sección 2) y la propuesta de solución describiendo cada algoritmo (sección 3). Se realiza un resumen comparativo de los algoritmos analizados (sección 4) y finalmente, se exponen las conclusiones y futuras líneas de trabajo (sección 5).

2 DEFINICIÓN DEL PROBLEMA

Considerando la información que una empresa tiene de cada cliente, es posible realizar un análisis que permita relacionar distintas variables y detectar anticipadamente a aquellos con alto potencial de abandono de la empresa, estableciendo así estrategias de fidelización y de retención cuando un cliente resulta interesante. Este enfoque puede ocasionar costos importantes para la empresa si la predicción es errónea, ya que se estaría perdiendo dinero en incentivos o promociones especiales a clientes que se irán de todos modos.

En la actualidad, existen varias técnicas y algoritmos que permiten abordar el problema de la fuga de clientes, o “Churn Analysis”. Sin embargo, la elección de un algoritmo sobre otro resulta ser un trabajo arduo, generando cada uno de ellos resultados diferentes. Esto dificulta además su proceso de aplicación por la no existencia de herramientas que unifiquen esta actividad [7, 12, 14].

Partiendo de lo enunciado y acompañando las actuales investigaciones sobre cada uno de los algoritmos de manera individual, surge el interrogante sobre la posibilidad de realizar un análisis comparativo de los diferentes algoritmos a fin de resultar una buena

práctica del estilo de guía al momento de decidir qué algoritmo aplicar. Asimismo, si es posible aportar información comparativa de los diferentes algoritmos facilitando así su aplicación.

3 SOLUCIÓN DEL PROBLEMA

A continuación, se presentan los distintos algoritmos que se han tenido en cuenta en base a diferentes investigaciones, y que a menudo se suelen utilizar para construir modelos de predicción de churn [13]. Por cada uno se tiene en cuenta una breve descripción junto a sus ventajas, desventajas y recomendación de uso. Se ha relevado la base de datos científica Science Direct [15] enfocándose en los trabajos de los últimos diez años. Asimismo, se han recolectado investigaciones utilizando el buscador académico Google Scholar y el buscador académico de la Universidad Tecnológica Nacional (UTN) [16, 17]. El estudio comparativo tiene como objetivo principal responder las preguntas planteadas en la sección anterior.

3.1. REGRESIÓN LOGÍSTICA

Es un tipo de algoritmo estadístico donde se desea conocer la relación entre una variable dependiente en función de más variables explicativas independientes o predictoras. El objetivo que resuelve esta técnica es modelar la probabilidad de aparición del suceso dicotómico ante el valor de los demás factores, a través de una función numérica matemática. Esta función está compuesta por coeficientes que son multiplicados por los valores que tomen las variables de entrada como se puede observar a continuación: [7, 18, 19, 20, 21].

$$P(Y = 1/x_1, x_2, ..., x_m) = \frac{1}{1 + e^{-(\beta_0 + \beta_1.x_1 + \beta_2.x_2 + ... + \beta_m.x_m)}}$$

Figura 1. La fórmula1 (fórmula de regresión logística - multinomial), tiene en cuenta lo siguiente: P(Y=1|xm) es la probabilidad de que la variable Y tome el valor de 1 (presencia de la característica estudiada), en presencia de las variables predictoras; xm es un conjunto de m variables predictoras; β0 es la constante del modelo o término independiente; βm son los coeficientes de las variables predictoras.

Un ejemplo que se puede mencionar es el de un banco que desea predecir aquellos clientes que dejan su entidad. Para el caso, se trabaja con un historial de datos de clientes en un tiempo determinado. Con los mismos se entrenará al modelo y se obtendrán los coeficientes correspondientes que se multiplicarán con variables como pueden ser edad del

cliente, zona geográfica donde vive, salario mensual, gastos promedio con la tarjeta, límite de la tarjeta, entre otras. Para predecir la fuga de los clientes actuales, se utilizará el modelo entrenado con los datos anteriores y se calculará la probabilidad que éstos clientes dejen la entidad.

3.1.1. Ventajas

Los coeficientes de la fórmula que utiliza el algoritmo, reflejan la relación entre las variables independientes y la dependiente, pudiendo realizar un análisis de la importancia de la presencia de las mismas para explicar la variable dependiente [22]. En el ejemplo nombrado anteriormente, podría pasar que el modelo indique que la variable "gasto promedio de la tarjeta" revele mayor importancia que edad del cliente, teniendo en cuenta los valores de los coeficientes.

3.1.2. Desventajas

Esta técnica requiere actualización constante en ambientes dinámicos donde las variables independientes se van modificando y actualizando [23]. En línea al ejemplo presentado se puede mencionar que si la entidad bancaria realiza constantes modificaciones en los límites de las tarjetas, o la economía del país es cambiante y se modifican los sueldos, se sugiere que el modelo se actualice por lo menos anualmente.

Asimismo, requiere de una gran inversión de tiempo para la codificación de las variables. Las variables de entrada son numéricas ya que es un algoritmo matemático. Aquellas variables que no sean numéricas pueden ser transformadas mediante una cuantificación en categorías o rangos numéricos. Por ejemplo la variable Zona Geográfica cuyas instancias son Capital Federal, Zona Norte, Zona Sur, entre otras pueden ser transformadas mediante una tabla de equivalencias a 1,2,3 y así con las demás categorías que presente la variable.

3.1.3. Recomendación de uso

En primer lugar, es útil cuando el resultado que se busca es una probabilidad de ocurrencia y no una clasificación rígida. La regresión logística binaria se utiliza cuando la variable de interés es dicotómica (es decir, toma valores binarios). La regresión logística multinomial se utiliza cuando la variable interés es una variable categórica (que puede adoptar un número limitado de categorías) en función de las variables independientes o

predictoras (categóricas y/o cuantitativas) [22, 18]. El equipo que desarrolle este modelo debe tener un fuerte conocimiento estadístico por la naturaleza del algoritmo, para poder interpretar correctamente los resultados y una buena elaboración del modelo.

3.2 ÁRBOLES DE DECISIÓN

El árbol de decisión permite describir gráficamente los posibles sucesos que pueden ocurrir a partir de una decisión asumida, permitiendo además ayudar a la toma de decisiones. Realiza predicciones basándose en la tendencia hacia un resultado concreto. Proporciona un conjunto de reglas que se van aplicando sobre los ejemplos nuevos para decidir qué clasificación es la más adecuada a sus atributos. La construcción del árbol de decisión depende del orden en que se hacen las preguntas sobre los atributos a encontrar [24, 18, 20].

Los algoritmos utilizados para la construcción de árboles son diferentes variaciones del genérico "Greedy algorithm" que va desde la raíz hacia las hojas (top-down) buscando recursivamente los atributos que generan el mejor árbol hasta encontrar el óptimo global con una estructura de árbol lo más simple posible [25]. De esta manera, dependiendo del corte en cada nodo, el mecanismo de división o segmentación del espacio de ejemplos y criterio de parada, se da lugar a distintos algoritmos de clasificación como CART, ID3, C4.5 y C5.0. Estos algoritmos, de la familia TDIDT (Top Down Induction of Decision Trees), pertenecen a los métodos inductivos del aprendizaje automático que aprenden a partir de ejemplos preclasificados y generan árboles y reglas de decisión a partir de estos ejemplos [18, 26].

Entre los mencionados, CART brinda la posibilidad de obtener valores reales (o continuos) como resultado contra los valores discretos o categóricos de ID3, C4.5 y C5.0 [25].

Debido a su capacidad de representar reglas de modo si-entonces, que permiten una mayor comprensión de los resultados, el algoritmo C4.5 es uno de los más populares en la minería de datos. Es una extensión mejorada del algoritmo ID3, ya que cuenta con la capacidad de tratar atributos continuos y no sólo discretos como éste último [27]. El algoritmo C5.0 es una evolución de la versión C4.5, ya que construye árboles considerablemente más pequeños, en menos tiempo y con la misma capacidad predictiva [25].

En [4] se plantea que los algoritmos como árboles y reglas de asociación, que son descriptivos de la situación, describen patrones secuenciales que logran explicar por qué los

clientes se van de la empresa. Si estos patrones son aplicados de forma inversa, se puede mejorar la situación.

3.2.1. Ventajas

Permiten procesar un gran volumen de información de manera eficiente [25]. Tiene buen manejo del ruido u outliers (errores en los valores o en la clasificación de estos) [25]. Pueden ser utilizados para representar visualmente y de forma explícita la toma de decisiones, siendo fáciles de construir, interpretar y comprender luego de una breve explicación. [8].

3.2.2. Desventajas

Pocos flexibles en su estructura para generalizar sobre datos de prueba excesivamente ramificados y complejos (situación conocida como overfitting o sobreajuste) [28]. Existen conceptos que difícilmente se puedan representar por medio de los árboles de decisión como XOR o problemas de multiplexor [11, 28].

3.2.3. Recomendación de uso

Cuando el grupo de variables o atributos predictores contiene una mezcla de variables numéricas y factores. Es un tipo de algoritmo descriptivo, que refleja gráficamente las relaciones entre las variables. Esto facilita el entendimiento de las relaciones. Los árboles de decisión se pueden utilizar para modelizar problemas de clasificación binaria (SI/NO) o clasificación multiclase. Esto resulta útil para medir el grado de satisfacción de los clientes, y de regresión (para el caso del análisis de gastos o pagos de clientes). Todo esto confluye en el plan de marketing de cualquier empresa de retail orientado a retener compradores frecuentes [11, 29].

3.3 SVN

Es un tipo de algoritmo dentro del área del aprendizaje automático o aprendizaje de máquinas (Machine Learning), disciplina dentro del ámbito de la Inteligencia Artificial, encargada de crear sistemas que aprenden automáticamente [3, 19].

El SVM o Máquinas de vectores de soporte (Support Vector Machines) define límites entre las clases de datos mediante la optimización del error cuadrático medio. El algoritmo es entrenado de forma tal que logra representar los puntos en un espacio multidimensional y

generar hiperplanos que logren separar y clasificar los elementos [3, 19, 30, 10]. Amazon utiliza este tipo de algoritmos para identificar los clientes con alta probabilidad de abandono, permitiéndole interactuar con ellos proactivamente a través de promociones o contactos de atención al cliente [13].

3.3.1.Ventajas

Dentro de los algoritmos de clasificación binaria, es considerado uno de los modelos más precisos y robustos [31]. Minimiza el error estructural al clasificar nuevos objetos, logrando la habilidad de generalizar de forma correcta.

3.3.2. Desventajas

La clasificación que devuelve el modelo resulta puramente dicotómica y no provee una probabilidad de pertenencia a la clase [19]. No está diseñado para identificar los atributos importantes para construir la regla discriminante [19, 31].

3.3.3. Recomendación de uso

Los campos de aplicación suelen ser variables. En [3] se brinda un ejemplo de cómo los clientes de los bancos se caracterizan por permanecer en la misma compañía y mantener su cuenta y sus depósitos en la misma, sin elegir otra o cambiar de banco. Así es que para los bancos esta relación con el cliente provee un ambiente estable. Por este motivo, la deserción de clientes provoca grandes pérdidas en este tipo de compañías y se realizan numerosos estudios de retención de clientes. Se destacan, además, la aplicación de este algoritmo en detección de comportamiento de compras para el ámbito del e-commerce, teniendo como objetivo la fidelidad de los clientes [3].

3.4 REDES NEURONALES ARTIFICIALES

Las Redes Neuronales Artificiales (RNA) son un tipo de algoritmo del área de Inteligencia Artificial donde se modela el comportamiento del cerebro. Las neuronas son unidades interconectadas que poseen un peso numérico, y cada una recibe un número de valores de entrada, se procesa y produce distintos valores de salida. Las entradas y las respuestas pueden provenir o servir de entrada a otra unidad [11, 18, 32, 19, 20, 21]. El resultado del modelo es una clasificación categórica.

Tomando el mismo ejemplo de la entidad bancaria, este modelo puede ser utilizado previamente entrenado con los datos históricos. Los clientes de los cuales se desea conocer si dejarán, o no, la entidad son la entrada al modelo previamente entrenado y la salida es una clasificación de su posible comportamiento.

3.4.1. Ventajas

Un modelo de RNA correctamente entrenado, logra tener buenos resultados de clasificación [33].

3.4.2. Desventajas

Es necesario establecer su arquitectura de la red y la tasa de aprendizaje, aunque aún definida, no existen garantías de que converja a una solución aceptable. Estos parámetros afectan directamente el tiempo de entrenamiento, el rendimiento y la tasa de convergencia de la red neuronal. [28,33]. Se dificulta la interpretación de los resultados debido a que no contiene más que un conjunto de pesos para la red, dificultando ver las relaciones en el modelo y por qué son válidas [34].

3.4.3. Recomendación de uso

El Perceptrón (red de 3 capas) es el algoritmo más conocido de RNA con aprendizaje supervisado. El mismo podría ayudar a predecir si un cliente puede irse, o no, basándonos en sus características de comportamiento (variables de entrada), y contando con los datos históricos del comportamiento de los clientes, se puede entrenar la red para predicción/clasificación de nuevos casos [33].

3.5. NAÏVE BAYES

Este algoritmo calcula la probabilidad condicional entre atributos de entrada y de predicción y supone que las relaciones de dependencias entre los atributos del conjunto de datos son condicionalmente independientes entre sí dado un atributo clase o target [41]. De esta manera, se suele utilizar cuando se desea estimar la probabilidad de que ocurra un suceso determinado [11, 42, 9, 10, 21].

3.5.1. Ventajas

Es menos complejo que otros algoritmos, resultando útil para generar rápidamente modelos de minería de datos, para descubrir relaciones entre columnas de entrada y columnas de predicción [9]. Sólo se requiere de una pequeña cantidad de datos de entrenamiento para estimar los parámetros necesarios para la clasificación [8]. Es muy eficiente, tiene buen manejo del ruido (outliers) en los datos de entrada y es robusto a atributos irrelevantes [8].

3.5.2. Desventajas

Cuenta con la limitación de suponer la independencia condicional de todas sus variables predictivas, ya que este hecho es relativamente infrecuente. En algunas ocasiones o problemas complejos esto puede hacer disminuir el rendimiento del algoritmo. Asimismo, la redundancia en los atributos puede causar problemas [43].

3.5.3. Recomendación de uso

Este algoritmo de clasificación es útil cuando se quiere identificar las características o factores de mayor incidencia sobre un determinado resultado de un problema [44]. En nuestro análisis, el algoritmo podría aplicarse para ponderar las características más significativas que presentan los clientes con tendencia a darse de baja, focalizando las acciones en minimizar esta situación de riesgo. Asimismo, podría utilizarse combinado con los árboles de decisión, para ponderar los resultados de las reglas descubiertas [44]. Otra de las aplicaciones de interés de este algoritmo es en la clasificación de textos, concretamente para el filtrado de spam [45].

3.6 REGLAS DE ASOCIACIÓN

Las reglas de asociación encuentran relaciones de interés entre los valores de los atributos en una base de datos [35, 36]. Su objetivo se centra en identificar patrones de asociación de ítems que aparecen juntos en un determinado conjunto de datos, representando esas asociaciones de dependencia mediante reglas. A través de éstas es posible conocer la probabilidad de que la ocurrencia de un conjunto de ítems implique la ocurrencia de otro conjunto de ítems [35, 36].

Las reglas de asociación reflejan, mediante un umbral indicador de soporte y confianza [35], la validez, certeza y utilidad de las mismas para el descubrimiento de relaciones

interesantes en grandes conjuntos de datos [4, 37, 18, 20]. A pesar de que constituyen una forma de aprendizaje no supervisado, en [48] se menciona la existencia de investigaciones proponiendo el uso de reglas de asociación para resolver problemas de clasificación.

Cabe destacar, que existe una clasificación llamada Patrones Secuenciales. En ella se enmarcan las reglas de asociación secuenciales, que se usan para determinar patrones secuenciales en los datos. Se basan en secuencias temporales de acciones y difieren de las reglas de asociación en que las relaciones entre los datos son temporales [4, 39].

3.6.1. Ventajas

El resultado del algoritmo en forma de reglas es de fácil entendimiento y ayuda a la toma de decisiones del negocio. Siempre tratan de buscar muchas reglas, cada una de las cuales puede tener una conclusión diferente, a diferencia de otros algoritmos que generan reglas con una única conclusión. La lógica del algoritmo es sencilla, no requiere de un equipo con mucha experiencia [37, 18, 20, 38].

3.6.2. Desventajas

Tratan de encontrar patrones en un espacio de búsqueda potencialmente muy amplio y, por tanto, pueden necesitar mucho más tiempo de ejecución que otros algoritmos como árboles de decisión [4, 39, 38].

3.6.3. Recomendación de uso

Se utiliza regularmente para obtener información acerca de la compra de productos que realizan los clientes en los supermercados, tarea conocida como “análisis de canasta de compra” o Market Basket Analysis [40]. Mediante el uso de reglas de asociación, podrían detectarse diferentes relaciones de dependencia entre los clientes y aplicar acciones de seguimiento sobre aquellos que poseen mayor tendencia a la fuga [38].

3.7. SNA

El Análisis de Redes Sociales (Social Network Analysis, SNA), emerge como una técnica clave del nuevo paradigma de la sociología moderna, ciencias de la información, comunicación y ciencias políticas, entre otras. Es utilizado para observar las relaciones sociales en términos de la teoría de redes. Permite reconocer las relaciones entre la gente para plasmarlas en un mapa que facilite la identificación del flujo de conocimiento: de quién

toma la gente información y conocimiento, con quiénes lo comparten o quién conoce a quién; a diferencia de los organigramas que sólo muestren relaciones formales [46, 47].

Teniendo en cuenta la clasificación de una determinada población con técnicas de SNA, es posible diseñar estrategias activas para aquellos que no sólo tienen una mayor predisposición a darse de baja sino, además, los que podrían ser arrastrados por los primeros, por el contagio que cualquier tipología de red puede tener [47].

Técnicas de aprendizaje automático como árboles de decisión, Naive Bayes, redes neuronales y algoritmos genéticos, a menudo se utilizan para construir modelos de predicción de churn en SNA [49].

3.7.1. Ventajas

Permite a los investigadores recoger datos cualitativos y preguntas de aclaración mientras que la red de datos se traza [47].

3.7.2. Desventajas

Por su incipiente aparición, muchos investigadores cuentan con ciertos reparos a la hora de realizar análisis mediante esta técnica. De igual manera, no existe por el momento demasiados artículos destinados a fin de poder conocer más de esta técnica [47].

3.7.3. Recomendación de uso

Tal y como se mencionaba en párrafos anteriores, esta técnica es útil para medir el Churn desde el punto de vista de la propagación, es decir, la señal que podría transmitir cierto cliente que se da de baja en una compañía influenciando sobre otros. Es importante también destacar que su aplicación se recomienda para mejorar la efectividad de los canales de comunicación formal e informal dentro del marco de una organización, así como también para visualizar las relaciones y personas más influyentes [46,47]

4 TABLA COMPARATIVA DE ALGORITMOS

Se procede a comparar los siguientes algoritmos mediante diferentes valores descriptos a continuación.

En primer lugar se toman cada uno de los algoritmos descriptos comparándolos mediante su tipo de Algoritmo y Resultado del Algoritmo (Tabla 1). En un primer agrupamiento se identifican los algoritmos del tipo “Estadísticos” como los de Regresión

Logística y Naive Bayes. Un segundo agrupamiento considera a los algoritmos SVM, RNA y SNA como del tipo IA (Inteligencia Artificial). En tercer lugar se establece el agrupamiento denominado de clasificación, tomando a los algoritmos Naive Bayes, Árboles de Decisión y SNA. Un cuarto y último agrupamiento, denominado de Asociación, incluye al algoritmo Reglas de Asociación.

Como segunda tabla (Tabla 2) podemos encontrar un agrupamiento por 3 variables. La primera es la recomendación de experiencia del equipo, debido principalmente a la dificultad del algoritmo para ser aplicado. La segunda tiene en cuenta el grado de información extra que provee cada algoritmo para el entendimiento del Negocio. Como tercera y última variable, el tipo de datos de entrada que necesita cada algoritmo. De esta manera, se puede observar que la Regresión Logística, SVM y RNA tienen como requerimiento el tipo de dato de entrada numérico, distinto a los algoritmos de Reglas de Asociación, Árboles de Decisión y SNA que no tienen restricciones.

Tabla 1. Agrupamiento por Tipo y Resultado.

	Tipo de Algoritmo	**Resultado del Algoritmo**
Regresión Logística	Estadístico	Fórmula para calcular una probabilidad
Naive Bayes	Estadístico / Clasificación	Árbol de ponderación de dependencias significativas
SVM	IA	Fórmula para calcular una probabilidad
RNA	IA	Función con el valor de salida de la neurona, en base al estado de activación de la misma.
SNA	IA / Clasificación	Árbol/Grafo de la estructura de la Red
Arboles de Decisión	Clasificación	Reglas con una única conclusión (atributo clase o target), expresadas mediante un árbol descriptivo. (Gráfico)
Reglas de	Asociació	Reglas con conclusiones

	Tipo de Algoritmo	Resultado del Algoritmo
Asociación	n	diferentes, por asociación de atributos

Tabla 2. Agrupamiento por Tipo de entrada, información extra que brinde para el negocio y experiencia recomendable del equipo.

	Experiencia del equipo	Brinda datos extra para entender el Negocio	Tipos de Datos de entrada
Regresión Logística	Alta	SI	numéricos
SVM	Alta	NO	numéricos
RNA	Media	NO	numéricos
Naive Bayes	Media	SI	discretos
Reglas de Asociación	Baja	SI	todos
Arboles de Decisión	Baja	SI	todos
SNA	Baja	SI	todos

5 CONCLUSIONES

Las empresas necesitan analizar diferentes indicadores que los ayuden a tomar decisiones a mediano y largo plazo con el fin de implementar estrategias que mejoren su eficiencia y su posicionamiento en el mercado.

La minería de datos es fundamental para proporcionar el conocimiento de poblaciones enteras de datos basándose en el análisis de los mismos e históricos. Relacionada a la tecnología, se encuentra en el ámbito de la programación asociándose con los algoritmos a implementar.

Entre las distintas estrategias de negocio de las compañías, la retención de clientes es una de ellas y las causas por las cuales éstos dejan de consumir los servicios son múltiples. No obstante, muchas empresas asumen la tasa de abandono como una condición normal del negocio y sus esfuerzos se enfocan en la captación continua de nuevos clientes, sin

considerar que esto tarde o temprano es más costoso que tratar de conservar los clientes existentes. En ese sentido, los esfuerzos deben enfocase en seleccionar y retener a los clientes adecuados e incrementar el valor de éstos.

La segmentación de los clientes que abandonan la empresa permite generar alertas de los clientes propensos al Churn. Asimismo con procesos adecuados y algoritmos minería de datos es posible detectar qué clientes entran en la fase previa del abandono y analizar sus causas y factores significativos.

Considerando lo planteado, se ha detectado la necesidad de considerar oportuna una comparación de los diferentes algoritmos existentes para aplicar al concepto mencionado de Churn Analysis, sirviendo así de guía recopiladora para poder encarar futuras investigaciones a fin de facilitar su posterior aplicación dependiendo el caso.

Con esta finalidad, se presenta un análisis comparativo de distintos algoritmos de minería de datos que pueden ser de aplicación al análisis del Churn, a efectos de identificar cuál se comporta mejor para predecir la fuga o abandono de clientes, basándonos en un conjunto de factores de relevancia. Esta comparación resulta de gran utilidad para el entendimiento conceptual de cada uno, pudiendo así contar con una base sólida a la hora de su elección e implementación.

Acorde a lo expuesto, como futuras líneas de desarrollo se propone seleccionar campos específicos de aplicación e implementar cada uno de los algoritmos propuestos, a fin de analizar y medir la capacidad predictiva de los mismos. La información entregada por cada algoritmo servirá de aporte para que las empresas puedan aplicar anticipadamente acciones de retención, permitiéndoles reducir los costos de retención, disminuir la tasa de abandono, focalizar los esfuerzos de marketing y mejorar la comunicación con sus clientes.

Asimismo, se propone establecer un conjunto de características y factores que permitan clasificar distintos proyectos de Churn, asociándole el/los algoritmo/s más apropiado/s según los resultados obtenidos.

REFERENCES

Boulic, R. and Renault, O. (1991) “3D Hierarchies for Animation”, In: New Trends in Animation and Visualization, Edited by Nadia Magnenat-Thalmann and Daniel Thalmann, John Wiley & Sons ltd., England.

Dyer, S., Martin, J. and Zulauf, J. (1995) "Motion Capture White Paper", http://reality.sgi.com/employees/jam_sb/mocap/MoCapWP_v2.0.html, December.

Holton, M. and Alexander, S. (1995) "Soft Cellular Modeling: A Technique for the Simulation of Non-rigid Materials", Computer Graphics: Developments in Virtual Environments, R. A. Earnshaw and J. A. Vince, England, Academic Press Ltd., p. 449-460.

Knuth, D. E. (1984), The TeXbook, Addison Wesley, 15th edition.

Pollo-Cattaneo. M., Pytel, P., García-Martínez, R., Vegega, C., Amatriain, H., Ramón, H., Mansilla, D., Deroche, A., Cigliuti, P., Saavedra-Martínez, P., Garbarini, R., Rodriguez, D., Britos, P., Tomasello,M. (2013). Prácticas y Aplicaciones de Ingeniería de Requisitos en Proyectos de Explotación de Información. Proceedings del XV Workshop de Investigadores en Ciencias de la Computación, Pág. 171-175. ISBN 978-9-872-81796-1.

Barrientos, F., Ríos, S.A. (2013). Aplicación de Minería de Datos para Predecir Fuga de Clientes en la Industria de las Telecomunicacione. Universidad de Chile, Santiago, Chile.

He, B., Shi, Y., Wan, Q., & Zhao, X. (2014). Prediction of customer attrition of commercial banks based on SVM model. Procedia Computer Science, 31, 423-430.

Chiang, D. A., Wang, Y. F., Lee, S. L., & Lin, C. J. (2003). Goal-oriented sequential pattern for network banking churn analysis. Expert Systems with Applications, 25(3), 293-302.

Ahn, J. H., Han, S. P., & Lee, Y. S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications policy, 30(10), 552-568.

Keramati, A., & Ardabili, S. M. (2011). Churn analysis for an Iranian mobile operator. Telecommunications Policy, 35(4), 344-356.

Mutanen, T. (2006). Customer churn analysis–a case study. Journal of Product and Brand Management, 14(1), 4-13.

Lozano Núñez, D. (2015). Modelos Predictivos del Churn - Abandono de clientes - para operadores de telecomunicaciones. Universidad de Vigo.

Alvarado Bustos Valdivia, J.M (2011). Diseño e Implementación de un modelo predictivo para detectar patrones de fuga en los servicios de telefónica del sur, Universidad Austral de Chile.

Kisioglu, P., & Topcu, Y. I. (2011). Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey. Expert Systems with Applications, 38(6), 7151-7157.

Hu, X. (2005). A data mining approach for retailing bank customer attrition analysis. Applied Intelligence, 22(1), 47-60.

Ahn, J. H., Han, S. P., & Lee, Y. S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications policy, 30(10), 552-568.

AManso, F. (2015) Análisis de Modelos de Negocios basados en Big Data para Operadores móviles. Universidad de San Andrés. Tesis de Maestría en Gestión de Servicios Tecnológicos y Telecomunicaciones (http://repositorio.udesa.edu.ar/jspui/bitstream/10908/10920/1/%5bP%5d%5bW%5d%20T.%20M.%20Ges.%20Manso%2c%20Fernando.pdf).

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. CRISP-DM 1.0 Step-by-step Data Mining Guide. http://tinyurl.com/crispdm, 2000 (accessed 02.05.17).

Science Direct, http://www.sciencedirect.com/, Último acceso Julio de 2017.

Google Académico, https://scholar.google.com.ar/, Último acceso Julio de 2017.

Bibliotecas Electrónicas UTN, http://portal.bibliotecas.utn.edu.ar/proxy/, Último acceso Julio de 2017.

Britos, P. V., & Britos, P. V. (2005). Minería de datos basada en sistemas inteligentes. Nueva Librería.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics, 35(5), 352-359.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert systems with applications, 36(2), 2592-2602.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, 1-9.

Peláez, I.M. (2016). Modelos de regresión: lineal simple y regresión logística. Revista SEDEN.

Van den Poel, D., & Lariviere, B. (2004). Customer attrition analysis for financial services using proportional hazard models. European journal of operational research, 157(1), 196-217.

Arnejo Calviño, H.A. (2016). Métodos para la mejora de predicciones en clases desbalanceadas en el estudio de bajas de clientes (CHURN)

Dupouy Berrios, C. (2014). Aplicación de Árboles de decisión para la estimación del escenario económico y la estimación de movimiento la tasa de interés en Chile. Santiago, Chile.

Lopez Nocera, M. (2012). Descubrimiento de Conocimiento Mediante la Integración de Algoritmos de Explotación de Información. Tesis de Magister en Ingeniería de Sistemas de Información. Facultad Regional Buenos Aires. Universidad Tecnológica Nacional.

Ramírez Cabrera, J.L. (2013) Análisis comparativo DBDT vs otros Algoritmos para el manejo de datos no escalares. Departamento de Sistemas Informáticos y Computación. Universidad Politécnica de Valencia.

Goddard J.C.; Cornejo J.M.; Martínez F.M.; Martínez A.E., Rufiner H.L.; Acevedo R.C. (1995) Redes Neuronales y Árboles de Decisión: Un Enfoque Híbrido. SINC.

Martín Arevalillo, J. (2013) Data mining con árboles de decisión en Bioinformática. In Ciclo de conferencias de la Facultad de Informática 2012/2013, 18 de junio de 2013, Sala de Grados de la Facultad de Informática de la Universidad Complutense de Madrid.

Orrù, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., & Mechelli, A. (2012). Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neuroscience & Biobehavioral Reviews, 36(4), 1140-1152.

Álvaro Flores,Sebastián Maldonado, Richard Weber (2015). Selección de atributos y support vectormachines adaptado al problema de fugade clientes. Revista Ingeniería de Sistemas.

Pyle, D. (1999). Data preparation for data mining (Vol. 1). morgan kaufmann.

Mihaich, F. (2014). Aplicación de redes neuronales en la clasificación de imágenes. Trabajo Especial de Licenciatura en Ciencias de la Computación. Universidad Nacional de Córdoba.

Tang Z & MacLennan J (2005). Data Mining with SQL Server 2005. Wiley Publishing, Inc.

Kotsiantis S., Kanellopoulos D. (2006). Association Rules Mining: A Recent Overview. GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), pp. 71-82

Lavrac, N., Kavsek, B., Flach, P., Todorovski, L. (2004). Subgroup discovery with CN2-SD. Journal of Machine Learning Research, 5. Pp. 153-188.

T. Menzies, Y. Hu. (2013). Data Mining For Busy People. IEEE Computer, pgs. 18-25.

Jélvez Caamaño, A., Moreno Echeverría M., Ovalle Retamal, V., Torres Navarro, C., Troncoso Espinosa F. (2014). Modelo predictivo de fuga de clientes utilizando minería de datos para una empresa de telecomunicaciones en Chile. Académico Departamento de Ingeniería Industrial, Universidad del Bío-Bío.

Álvaro Pita Martín, D. (2012). Una metaheurística para la extracción de reglas de asociación. Aplicación a terremotos. Escuela Técnica Superior de Ingeniería Informática Máster Oficial en Ingeniería y Tecnología del Software.

Chen Y., Tang K., Shen R., Hu, Y. (2005). Market basket analysis in a multiple store environment. Original Research Article Decision Support Systems, Volume 40, Issue 2, Pages 339-354

Langley P., Sage S. (2013). Induction of Selective Bayesian Classifiers. Institute for the Study of Learning and Expertise. Palo Alto, CA-USA. Pp. 399-406.

Nath, S. V., & Behara, R. S. (2003, November). Customer churn analysis in the wireless industry: A data mining approach. In Proceedings-annual meeting of the decision sciences institute (pp. 505-510).

Ruiz Sánchez, D.R., (2006) Heurística de selección de atributos para datos de gran dimensionalidad. Departamento de Lenguajes y Sistemas Informáticos. Universidad de Sevilla.

Britos, P. (2008). Procesos de Explotación de Información basados en Sistemas Inteligentes. Tesis Doctoral. Universidad Nacional de La Plata. Facultad de Informática. Argentina.

García Serrano, A. (2013). Inteligencia artificial. Fundamentos, práctica y aplicaciones. Alfaomega.

Análisis de Redes Sociales (ARS), http://www.kstoolkit.org/An%C3%A1lisis+de+Redes+Sociales+%28ARS%29/, Último acceso Julio de 2017.

Klepac, G., (2015) Developing Churn Models Using Data Mining Techniques and Social Network Analysis. IGI Global. ISBN 1466662891, 9781466662896.

Lucas J. (2010). Métodos de Clasificación Basados en Asociación Aplicados a Sistemas de Recomendación. Tesis Doctoral. Universidad de Salamanca, Departamento de Informática y Automática, España.

Manso, F. (2015). Análisis de Modelos de Negocios Basados en Big Data para Operadores Móviles. Tesis de Magíster en Gestión de Servicios Tecnológicos y de Telecomunicaciones. Universidad de San Andrés.