Reglas de Asociación en Flujos de Datos para Monitorizar Actividad de Teléfonos Móviles

Elena Ruiz, Jorge Casillas
DaSCI (Centro de Investigación en Ciencia
de Datos e Inteligencia Computacional)
Universidad de Granada, Granada, España
Email: {eruiz, casillas}@decsai.ugr.es

***Abstract*—Los algoritmos de minería de flujo de datos trabajan sobre datos con altas tasas de llegada, que evolucionan a lo largo del tiempo y requieren respuesta en tiempo real. Este tipo de técnicas, que procesan los datos al vuelo, han captado la atención tanto del ámbito científico como del industrial. El aprendizaje descriptivo en flujos de datos nos permite tener un modelo que se adapta a la evolución de los datos para explicar qué está pasando en tiempo real. En este trabajo, mostramos el potencial de este campo usando información real registrada a través de teléfonos móviles durante meses (por el MIT Human Dynamics Lab). El objetivo es evolucionar de forma dinámica reglas de asociación que expliquen la actividad del usuario en cualquier momento de forma muy eficiente para que pueda incorporarse en un dispositivo móvil. Para conseguir este objetivo empleamos un algoritmo evolutivo que aprende y mantiene de forma incremental una población de reglas de asociación.**

***Keywords*—reglas de asociación; flujo de datos; algoritmo genético; aprendizaje automático; aprendizaje online**

I. Introducción

Vivimos en la era de los datos, donde todos nuestros movimientos y actividades son registrados (o podrían serlo) y, a veces, almacenados y procesados. Obviamente, una gran parte de la información generada carece de interés. Elegir qué información es relevante, sintetizarla y extraer conocimiento de ella es, cada vez, un aspecto más crítico en la sociedad actual. En ocasiones es posible utilizar esta información para obtener modelos (data mining) que simplifican la compleja realidad que contiene dicha información.

La necesidad de extraer información relevante de fuentes de datos ordenados cronológicamente en forma de un flujo continuo, veloz y que cambia con el tiempo, excediendo las capacidades habituales de almacenamiento y procesamiento son cada vez más comunes tanto en el entorno industrial como en el científico [1]. Para solucionar este tipo de problemas es posible gestionar flujos de datos, secuencias infinitas de registros estructurados que se reciben de forma continua [1]. La característica clave de estos sistemas es que los datos producidos por estos flujos no se almacenan de forma permanente, sino que se procesan sobre la marcha. Cada dato es analizado, procesado y olvidado, haciendo posible la gestión de grandes cantidades de datos en tiempo real, incluso con capacidades de almacenamiento y procesamiento reducidas.

Los principales problemas de aprendizaje estudiados en la minería de flujo de datos [1], [2] son: (1) clasificación [3], (2) clustering [4] y (3) patrones frecuentes. En los últimos años, la mayor parte de la literatura especializada en este área se ha centrado en clasificación (y *concept-drift*); a pesar de su falta de aplicabilidad en casos reales, lo que ha derivado en experimentaciones basadas únicamente en benchmarks y datos sintéticos. Sería más realista generar modelos descriptivos e interpretables que permitan monitorear sistemas.

En general, el aprendizaje no supervisado es más directamente aplicable a problemas reales de flujo de datos, por lo que el clustering incremental ha experimentado un desarrollo significativo. Sin embargo, el conocimiento que se descubre (segmentación) resulta a menudo insuficiente para ayudar en la toma de decisiones. Por lo tanto, el descubrimiento de patrones frecuentes y reglas de asociación se considera una muy buena manera de abordar muchos problemas de flujo de datos cuyo propósito consiste en supervisar o monitorear (no predecir) en tiempo real usando modelos independientes, significativos, legibles y simples. Más concretamente, el descubrimiento de asociaciones en flujos de datos mediante la producción de reglas de asociación en un proceso completamente on-line, es particularmente interesante debido a: (1) la demanda de interpretabilidad de los patrones descubiertos en los datos, (2) la necesidad de descubrir patrones a medida que suceden, y (3) los altos y continuos volúmenes de datos a procesar, que exigen algoritmos escalables. Un caso real que supone un buen ejemplo de la utilidad de este campo es la detección de amenazas potenciales para los sitios web y las infraestructuras de red [5]. Existen otras estrategias de detección de anomalías (estadísticas o basadas en densidad), pero típicamente se basan en datos etiquetados y, por lo tanto, no se adaptan a nuevos conceptos. Otro posible caso de uso es el analizado en este documento, en el que se trabaja sobre distintos tipos de información relacionada con el uso del teléfono móvil.

El objetivo de este trabajo es mostrar el potencial de la minería de reglas de asociación en flujo de datos al tratar con varios meses de datos reales de uso de teléfonos móviles (tasa de muestreo de un minuto) proporcionados por el MIT Media Lab. El objetivo final es mantener dinámicamente un conjunto de reglas de asociación que expliquen la actividad del usuario

Este trabajo ha sido financiado por los fondos MINECO/FEDER (TIN2017-89517-P), y por el Proyecto BigDaP-TOOLS - Ayudas Fundación BBVA a Equipos de Investigación Científica 2016. E. Ruiz disfruta de un contrato vinculado al proyecto TIN2014-57251-P del MINECO.

en cualquier momento de forma muy eficiente.

El resto de este documento está organizado de la siguiente manera: La sección II presenta las principales características de Fuzzy-CSar (el algoritmo utilizado en este estudio). La sección III proporciona una descripción del estudio realizado por los investigadores del MIT y explica el proceso de preparación que hemos aplicado a los datos. La sección IV presenta los resultados obtenidos. Finalmente, la Sección V resume y concluye el trabajo.

II. Fuzzy-CSar

Fuzzy-CSar [6] está diseñado para extraer reglas de asociación difusas de flujos de datos mediante la combinación de un algoritmo genético (GA) y mecanismos de aportación de crédito de forma on-line. Es uno de los pocos algoritmos capaces de generar directamente reglas de asociación (no solo *itemsets* frecuentes) con atributos tanto cuantitativos como cualitativos de forma puramente on-line, sin emplear ventana deslizante ni ninguna otra técnica para almacenar datos. Gracias a estas propiedades, el algoritmo encaja perfectamente con el propósito de este trabajo.

Fuzzy-CSar mantiene una población de individuos, donde cada uno está representado por una regla de asociación difusa y un grupo de parámetros que evalúan la calidad de la regla. La regla de asociación difusa consiste en un antecedente y un consecuente. Se permite que el antecedente tenga un número arbitrario de atributos mientras que el consecuente consiste en un solo atributo que no debe estar presente en el antecedente de la misma regla. Cada variable puede estar representada en la regla por una disyunción de términos lingüísticos (etiquetas) para facilitar una mayor generalización. Cada individuo tiene un total de ocho parámetros de calidad.

En cada iteración del proceso de aprendizaje de Fuzzy-CSar se recibe un nuevo ejemplo y el algoritmo lleva a cabo una serie de pasos para actualizar los parámetros de los individuos de la población y descubrir nuevas reglas relevantes. Para descubrir estas nuevas reglas prometedoras, se aplica un algoritmo genético estacionario basado en nichos [7]. Además, se aplican operadores de cruce, distintos tipos de mutación y *covering* con ciertas probabilidades. Podemos ver un esquema de la fase de aprendizaje de Fuzzy-CSar en el algoritmo 1. Complementariamente, explicamos brevemente algunos componentes de la fase de aprendizaje, una explicación más detallada se puede encontrar en [6].

Algoritmo 1: Esquema de la fase de aprendizaje de Fuzzy-CSar [6]

```
proceso TrainFuzzy-CSar( ejemploEntrenamiento e_t,
 Población [P] en el instante t )
Data: e_t tiene la forma {x_i}_{i=1}^{ℓ}
Result: Población [P] en el instante t + 1
begin
    e'_t ← granulación(e_t);
    genera [M] a partir de [P] usando e'_t;
    if |[M]| < θ_mna then
        genera θ_mna − |[M]| individuos que concuerdan usando e'_t y
          actualizando [P];
    end
    agrupa individuos en [M] por su antecedente formando distintos [A]_i;
    selecciona [A] según probabilidad;
    subsume individuos en [A];
    actualiza individuos en [M] ; // Por lo tanto, todos los
      [A]_i se actualizan.
    if el tiempo medio en [A] desde la última vez de GA > θ_GA then
        se lleva a cabo un evento genético en [A] considerando e'_t y
          actualizando [P];
    end
end
```

A. Parámetros de Calidad: Soporte y Confianza

En Fuzzy-CSar tenemos ocho parámetros de calidad para cada individuo de la población. Vamos a explicar cómo se calculan dos de ellos, los dos más utilizados y los más significativos para este trabajo: soporte y confianza (información sobre el cálculo de otros parámetros en [6]).

Si definimos formalmente el campo de minería de reglas de asociación como sigue [8]: Siendo $I = i_1, i_2, ..., i_l$ un conjunto de características binarias (ítems) de l elementos. Siendo $Tr = tr_1, tr_2, ..., tr_N$ un conjunto de N transacciones donde cada transacción tr_j contiene un vector binario que indica en cada posición si un ítem en particular está presente o no. Entonces un ítem X ($X \subset I$) va a tener asociado un soporte que es una medida de su importancia en T y se calcula como $supp(X) = |X(T)|/|T|$, donde $X(T)$ es el conjunto de variables en el antecedente de la regla. Si el soporte de un determinado *itemset* (conjunto de ítems) supera un umbral definido por el usuario (*minsupp*) este *itemset* se considera como un *conjunto frecuente de ítems*. Si X e Y son ambos conjuntos frecuentes de ítems y $X \cap Y = \emptyset$, podemos definir una regla de asociación como una implicación del tipo $X \to Y$. El soporte y la confianza de una regla de asociación son las medidas cualitativas tradicionalmente más usadas:

$$supp(X \to Y) = \frac{supp(X \cup Y)}{|T|}, conf(X \to Y) = \frac{supp(X \cup Y)}{supp(X)}. \quad (1)$$

Donde el soporte indica la frecuencia con la que se cumplen los patrones y la confianza evalúa la fuerza de la implicación indicada en la regla de asociación.

Sea $I = \{i_1, i_2, \ldots i_\ell\}$ un conjunto de ℓ características, $A \subset I$, $C \subset I$ y $A \cap C = \emptyset$. Una regla de asociación difusa es una implicación de la forma $X \to Y$ en la cual:

$$X = \bigwedge_{i_i \in A} \mu_{\widetilde{A}}(i_i) \quad \text{e} \quad Y = \bigwedge_{i_j \in C} \mu_{\widetilde{C}}(i_j), \quad (2)$$

donde $\mu_{\widetilde{C}}(i_j)$ es el grado de pertenencia de la variable en el consecuente y $\mu_{\widetilde{A}}(i_i)$ es el grado de pertenencia de las variables del antecedente. En esta situación, el soporte se extiende usando el producto T-norm y la confianza se extiende usando la *implicación de Dienes* [9]:

$$supp(X \to Y) = \frac{1}{|T|} \sum \mu_{\widetilde{A}}(X) \cdot \mu_{\widetilde{C}}(Y) \quad (3)$$

$$conf(X \to Y) = \frac{\sum(\mu_{\widetilde{A}}(X) \cdot \max\{1 - \mu_{\widetilde{A}}(X), \mu_{\widetilde{C}}(Y)\})}{\sum \mu_{\widetilde{A}}(X)}. \quad (4)$$

donde $\mu_{\widetilde{A}}(X)$ es el grado de pertenencia de la parte del antecedente de la regla y $\mu_{\widetilde{C}}(Y)$ es el grado de pertenencia de la parte del consecuente de la regla.

B. Operador de Covering

Este operador genera nuevas reglas de asociación difusas cuando hay menos de θ_{mna} (siendo θ_{mna} un parámetro de configuración) individuos en el *match set [M]*, individuos de la población actual que concuerdan con el ejemplo recibido e.

El operador de covering construye un nuevo individuo que concuerda con e en el máximo grado posible: para cada variable de entrada de e, e_i, el operador decide aleatoriamente si e_i va a formar parte del antecedente de la regla. Después, elige la variable que estará en el consecuente de entre aquellas que no han sido seleccionadas para formar parte del antecedente.

C. Subsunción de Reglas

Para cada regla de $[A]$ se comprueba su posible subsunción con cada una de las otras reglas del conjunto. Una regla r_i es una candidata para subsumir r_j si: (1) r_i es más general que r_j, y (2) ambas reglas tienen confianzas similares y r_i tiene la suficiente experiencia. Una regla r_i es considerada más general que r_j si todas las variables de r_i están también definidas en r_j y, para cada una de estas variables, r_i tiene, al menos, los mismos términos lingüísticos que r_j.

D. Descubrimiento de Nuevas Reglas

Fuzzy-CSar usa un algoritmo genético incremental estacionario basado en nichos para descubrir nuevas reglas. Este algoritmo genético que se aplica al *association set* seleccionado, cuenta con tres tipos diferentes de mutación: (1) mutación del antecedente; (2) mutación del consecuente, y (3) mutación de términos lingüísticos.

En Fuzzy-CSar, como es común entre los miembros de la familia Michigan-style LCS [10], el coste del algoritmo se incrementa linealmente con el tamaño máximo de la población de reglas, el máximo número de variables por regla, y semilogarítmicamente con el coste de ordenar el *match set*. Es importante resaltar que Fuzzy-CSar no depende directamente del número de transacciones, lo que lo hace muy adecuado para trabajar con grandes bases de datos. En el estudio aquí presentado, el algoritmo se aplica en un problema real (el cual se explica a continuación). La eficiencia es un requisito muy importante en minería de flujo de datos. En el problema aquí abordado, hemos estimado que el tiempo medio que Fuzzy-CSar tarda en procesar cada dato es de unos 15 milisegundos, siendo así factible procesar más de 60 muestras por segundo.

III. Estudio 'Friends and Family'

A. Datos Originales

El *estudio Friends and Family*[1] es una investigación llevada a cabo por el MIT Media Lab, durante los años 2010 y 2011 [11]. Este estudio transforma una comunidad residencial cercana a una conocida universidad norteamericana en un *laboratorio viviente* durante 15 meses. Durante estos meses, los investigadores del Media Lab presentaron su sistema de registro de interacciones sociales y comportamiento basado en teléfonos móviles. Durante cerca de un año, toda actividad, comunicación y detalle social de las vidas de un gran número de miembros de la mencionada comunidad fue registrado mientras ellos realizaban sus tareas cotidianas con normalidad. Un total de 130 sujetos formaron parte del estudio. Durante el período del estudio, se recogió una gran cantidad de datos que resultó en un conjunto de datos muy completo y longitudinal, bautizado como *Friends and Family dataset*. Dicho conjunto de datos incluye una gran colección de señales basadas en la actividad del teléfono móvil incluyendo comunicación (llamadas y mensajes), aplicaciones instaladas, ejecución de aplicaciones, acelerómetro, dispositivos bluetooth próximos...

El estudio se dividió en dos fases. Una fase piloto de 6 meses de duración que comenzó en marzo de 2010, y una segunda fase iniciada en septiembre de 2010. Hasta 130 sujetos participaron en esta segunda fase.

Se proporcionaron *smartphones* a los participantes del estudio con la condición de que estos debían ser sus teléfonos principales durante su participación en el estudio. Estos dispositivos harían el papel de sensores sociales *in-situ* para registrar las características de la actividad de los sujetos.

Parte de la colección de datos obtenida del "Estudio Friends and Family" fue publicada y ha servido como punto de partida para nuestro estudio. El tamaño de esta parte publicada de la colección supera los 7GB. Además, esta gran cantidad de datos está distribuida en distintos archivos cuyo origen, formato y estructura varían.

Los datos recogidos de los teléfonos móviles son el núcleo principal de la colección de datos construida a partir del estudio. La tabla I enumera algunos tipos de información incluidos en esta colección y utilizados en nuestro estudio, junto con sus frecuencias de muestreo originales.

Tabla I: Principales tipos de información incluidos en los datos originales y usados en nuestro estudio

Información	Frecuencia muestreo
Dispositivos bluetooth próximos	cada 5 minutos
Registro de llamadas	cuando una llamada es enviada/recibida
Registro de SMS	cuando un SMS es enviado/recibido
Aplicaciones en el dispositivo	cada 10 minutos
Aplicaciones en ejecución	cada 30 segundos

Dado el origen y las peculiaridades de la información recogida, el estudio se llevó a cabo bajo estrictos protocolos que aseguran la privacidad de todos los participantes.

B. Preparación del Flujo de Datos

La estructura original de los datos no era la más adecuada para el proceso de extracción de conocimiento ni para obtener resultados de calidad. Por lo tanto, los datos tuvieron que ser tratados antes de aplicar el algoritmo. Los datos se encontraban distribuidos en diferentes archivos con diferente formato, estructura y frecuencia de muestreo dependiendo del tipo de información y del método de recolección utilizado. Cada tipo de información había sido registrado con sus propias peculiaridades, los datos de todos los participantes estaban mezclados y no todos los sujetos estuvieron implicados al mismo nivel en el estudio. Fue necesario aplicar un proceso de preparación sobre los datos en bruto para obtener un conjunto de datos completo, unificado y específico para cada sujeto.

[1] http://realitycommons.media.mit.edu/friendsdataset.html

En primer lugar, se realizó un estudio exploratorio de los datos para conocer y comprender mejor el problema. No toda la información original resulta útil para nuestros objetivos. Es necesario decidir qué información es relevante y cuál no.

Un algoritmo evolutivo completamente en línea presenta algunas peculiaridades dado que cada dato va a ser procesado una sola vez y el algoritmo no va a tratar con el conjunto de datos completo en ningún momento. Esto puede hacer que cierta información aparezca solo por un momento para acabar desapareciendo para el algoritmo a pesar de que puedan proporcionar información relevante. Tratando de minimizar este riesgo, se generaron variables del tipo "cuántas llamadas se han registrado en los últimos X minutos". Finalmente, se eligió una frecuencia de muestreo unificada de un minuto para los conjuntos de datos finales tratando de no perder demasiado detalle pero, al mismo tiempo, intentando evitar que el nivel de granularidad de los datos se vuelva innecesariamente bajo.

Para clarificar y resumir, describimos cada paso de la preparación de datos.

1) **Filtrado de datos por fecha**: se seleccionan los datos a partir del 01/10/2010, excluyendo los de la fase piloto.
2) **Agrupamiento de datos por participante**: se agrupa todos los datos de cada participante.
3) **Duplicados y orden cronológico**: eliminamos registros duplicados y ordenamos cronológicamente.
4) **Integración de datos**: los datos libres de duplicados y ordenados cronológicamente recogidos durante la segunda fase del estudio para cada participante se integran en conjuntos de datos individuales.

La figura 1 representa la evolución en el tiempo para los principales atributos después del proceso de preparación de los datos. Los datos presentados en dicha figura corresponden a un sujeto elegido como ejemplo para el análisis de resultados (Sección IV). El gráfico de la figura 1 tiene una resolución diaria, por lo que debemos tener en cuenta que los valores mostrados para las variables acumulativas (llamadas y contadores de mensajes) en los gráficos para un día dado son diez veces el número real de llamadas/mensajes. Esta figura nos ayuda a entender lo difícil que sería extraer información útil y descubrir asociaciones interesantes sin la asistencia proporcionada por el algoritmo de minería de flujo de datos.

IV. Experimentación y Resultados

A. Diseño de Experimentos

Los atributos de entrada utilizados en la experimentación realizada para obtener los resultados presentados en este trabajo son los siguientes: día de la semana; minuto del día [0, 1439]; porcentaje de batería; tres contadores de SMS en los últimos 10 minutos (entrantes, salientes y global); cuatro contadores de llamadas durante los últimos 10 minutos (entrantes, salientes, perdidas y global); tres variables referentes a acelerómetro; alguna aplicación desinstalada y alguna aplicación en ejecución.

La tabla II muestra los valores asignados a los principales parámetros de configuración de Fuzzy-CSar para los experimentos (cuyos resultados se discuten a continuación).

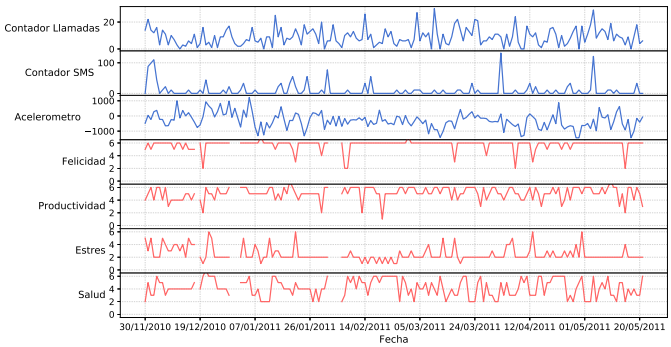


Figura 1: Representación gráfica de la evolución de algunos atributos que forman el conjunto de datos de uno de los participantes (Identificador del participante: sp10-01-24)

Tabla II: Principales parámetros de configuración de Fuzzy-CSar junto con los valores asignados en los experimentos

Parámetro	**Valor**
Tamaño máximo población	10.000
Comprobar subsunción	Sí
Tipo de *association sets*	Antecedente
Número máximo de conjuntos difusos	Dependiente del rango de la variable
Forzar etiquetas adyacentes	Sí
Número máximo de término lingüísticos	60% de los conjuntos difusos
Mutación	Sí
Comprobar subsunción en el GA	Sí

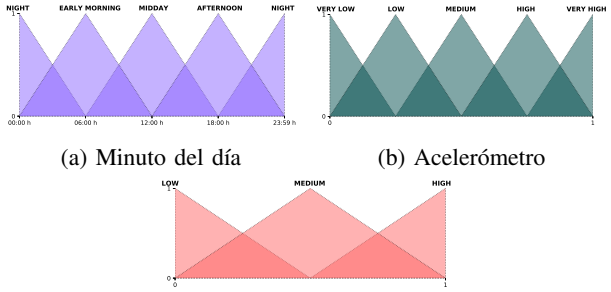
B. Análisis de Resultados

Como se ha explicado, Fuzzy-CSar mantiene continuamente una población de reglas. Es complicado encontrar una manera de trazar la evolución completa de toda esta población. Para poder representar y analizar esta evolución, nos centramos en ciertas reglas completas, consecuentes y antecedentes.

La figura 2 puede ayudar a entender mejor la descripción de algunas reglas de asociación que se muestran en esta sección. En ella es posible observar un esquema de la nomenclatura empleada para referirnos a los conjuntos difusos usados por Fuzzy-CSar para atributos numéricos.

Según estas nomenclaturas, la tabla III muestra tres reglas obtenidas por Fuzzy-CSar. Estos ejemplos ayudan a entender y visualizar mejor la estructura de las reglas de asociación obtenidas por Fuzzy-CSar. En la figura 3 podemos ver la evolución del número de copias almacenadas en la población (numerosidad), lo que representa la importancia relativa de la regla en cada momento, para cada una de las reglas representadas en la tabla III. Podemos observar cómo esta evolución es completamente diferente para cada regla. R_1 (rojo) aparece en un cierto momento, luego la numerosidad de la regla aumenta, para posteriormente disminuir hasta que la regla desaparece. Pero después de eso, la regla aparece de nuevo y repite el mismo proceso. Sin embargo, la regla R_2 (azul) aparece cerca del mes de abril para continuar aumentando su numerosidad hasta el final del experimento. Finalmente, la regla R_3 (verde) existe en la población desde el principio y su numerosidad es siempre creciente. Si analizamos la asociación representada por R_3 esta evolución parece muy lógica ya que normalmente

la gente duerme por la noche y no hace ninguna llamada por lo que esta condición es casi siempre cierta.



(a) Minuto del día (b) Acelerómetro

(c) Llamadas y SMS

Figura 2: Nomenclatura empleada para los conjuntos difusos en las variables referentes a: (a) minutos del día, (b) acelerómetro, y (c) contadores de llamadas y mensajes

Tabla III: Ejemplos de reglas generadas por Fuzzy-CSar durante los experimentos (figura 3)

ID	Rule	Soporte*	Confianza*
R_1	Si hora es MIDDAY o AFTERNOON entonces acelerómetro es MEDIUM, HIGH, VERY HIGH	0.124	0.802
R_2	Si número de SMS salientes es LOW entonces acelerómetro es HIGH, VERY HIGH	0.606	0.709
R_3	Si hora es NIGHT o EARLY MORNING y número de SMS entrantes es LOW entonces número de llamadas salientes es LOW	0.496	0.997

* El soporte y confianza se refieren al momento con máximo número de copias

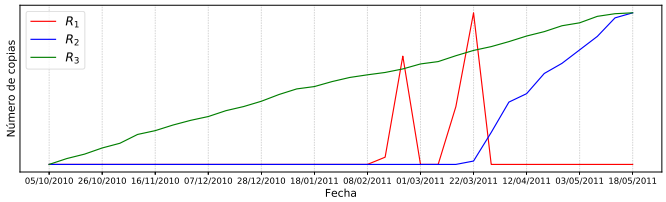


Figura 3: Evolución del número de copias en la población (numerosidad) de: R_1 (máximo 21 copias), R_2 (máximo 90 copias) y R_3 (máximo 537 copias) (tabla III)

Generalizando un poco nuestro análisis podemos centrarnos en un consecuente determinado en lugar de en una regla determinada. Siguiendo esta idea presentamos una serie de gráficos en los que analizamos la evolución en numerosidad de un conjunto específico de reglas que contienen la misma variable en su consecuente o incluso la misma etiqueta. Como prueba de concepto, a continuación se analiza la evolución de la cantidad de actividad física practicada por el participante.

La variable *AccelAccum*, incluida en el conjunto de datos usado, presenta valores relacionados con la información de acelerometría recogida de los *smartphones* de los participantes. Los investigadores del MIT relacionan dicha información de acelerometría con la actividad física, como también lo hacen otros investigadores [12]. Siguiendo esta interpretación,

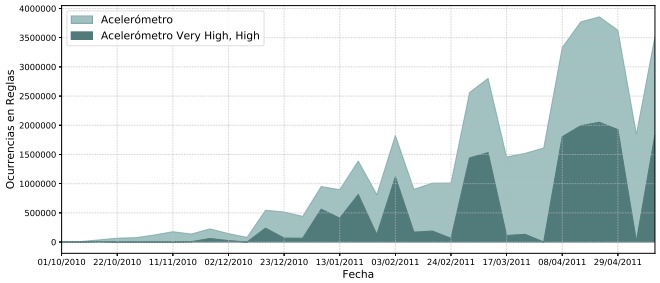


Figura 4: Comparativa entre la evolución en el nº total de reglas con *AccelAccum* en el consecuente y el número de ellas con la etiqueta VERY HIGH, HIGH ($supp \geq 0.1$ y $conf \geq 0.7$). Datos del participante *sp10-01-24*

entendemos que un mayor número de reglas cuyos consecuentes apuntan a valores *HIGH, VERY-HIGH* de la variable *AccelAccum* puede ser interpretado como un indicador de un aumento de la actividad física practicada por el sujeto. La figura 4 representa la evolución en el número de reglas, que superan los umbrales de soporte y confianza, cuyos consecuentes contienen la variable *AccelAccum* con cualquier etiqueta (área color claro) y las reglas cuyos consecuentes contienen valores altos de la variable *AccelAccum* (área color oscuro). Teniendo esto en cuenta, el gráfico representado en la figura 4 muestra cómo el participante *sp10-01-24* aumenta la cantidad de actividad física durante ciertos períodos de tiempo.

La figura 5 nos ayuda a entender la distribución de los valores de *AccelAccum* durante un cierto periodo de tiempo. Como se observa, no hay concentraciones especiales de valores altos o muy altos. El algoritmo comienza a encontrar relaciones que explican los altos valores de *AccelAccum* y, como consecuencia, aumenta el número de reglas con este consecuente específico y suficiente confianza.

Dado que se ha utilizado un algoritmo de reglas de asociación, y no uno de patrones frecuentes, podemos utilizar las relaciones establecidas por las reglas de asociación entre antecedentes y consecuentes como parte de nuestro análisis.

Así, profundizamos en los resultados analizando la composición del antecedente de las reglas con altos valores de *AccelAccum* en el consecuente. La figura 6 representa la evolución de estas reglas de calidad distinguiendo entre variables en el antecedente. En esta figura se muestra cómo la mayoría de los antecedentes se relacionan con la no utilización de varias funciones del teléfono móvil, por ejemplo, llamadas, mensajes, aplicaciones.... Los antecedentes agrupados bajo la etiqueta "Otro" se relacionan principalmente con la hora y el día de la semana. Este hecho refuerza la idea de que los valores altos del acelerómetro están relacionados con la actividad física y no con otros tipos de uso del *smartphone*.

V. Conclusiones

En este trabajo hemos mostrado una aplicación real de un algoritmo evolutivo (Fuzzy-CSar) para minería de reglas de asociación en flujo de datos, cuyo objetivo es descubrir de forma on-line y en tiempo real las relaciones entre los atributos

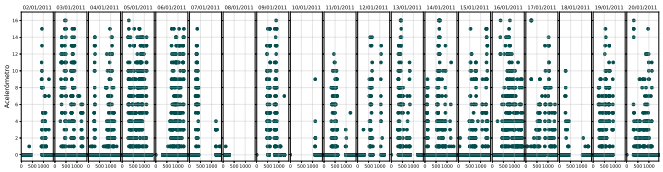


Figura 5: Distribución de los valores de la variable *AccelAccum* a lo largo de los minutos de cada día del 2 al 21 de enero de 2011. Datos del participante *sp10-01-24*

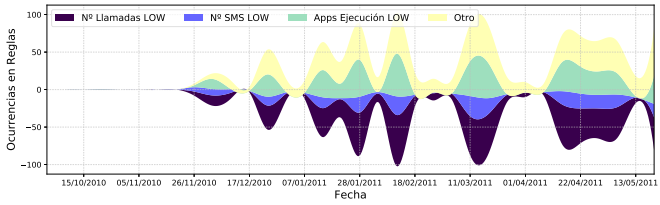


Figura 6: Comparación entre la evolución del nº total de reglas ($supp \geq 0.1$ y $conf \geq 0.7$) cuyo consecuente contiene *AccelAccum* con etiqueta VERY HIGH, HIGH y el nº de ellas con antecedentes referidos a la inactividad del teléfono. Datos del participante *sp10-01-24*

de un flujo de datos. Esta aplicación simula un sistema de monitorización en el que los datos de entrada consisten en muestras reales basadas en información sobre llamadas, SMS, acelerómetros, aplicaciones, etc. recogida desde *smartphones*. Estos datos constituyen un claro ejemplo de problema real en el que la información se genera como un flujo continuo e infinito y puede ser explotada de forma muy productiva sin necesidad de ser almacenada en grandes conjuntos de datos que requieran altas capacidades de procesamiento.

En este caso, el algoritmo evolutivo descubre reglas en tiempo real para explicar lo que está sucediendo en todo momento. Los resultados muestran la evolución que está experimentando la población de reglas de asociación a medida que aumenta la cantidad de datos procesada. Estos son solo algunos ejemplos que muestran cómo el algoritmo Fuzzy-CSar y las reglas de asociación descubiertas por él, hacen posible descubrir nuevas relaciones que no habrían sido descubiertas directamente a partir de datos en bruto. Además, esto se consigue de una manera muy eficiente, ya que Fuzzy-CSar tarda solo 15 ms en procesar cada dato, es decir, en el caso específico de datos con esta estructura es capaz de lidiar con una frecuencia de entrada de unos 67 Hz. En un entorno donde los datos llegan en tiempo real en forma de flujo infinito, como en este trabajo, es posible generar y actualizar un modelo en tiempo real, permitiendo un proceso de monitoreo que puede ayudar en un sistema de toma de decisiones. Otro posible caso de uso podría ser integrar este algoritmo en una aplicación móvil que utilice directamente el conocimiento descubierto por el algoritmo para tomar decisiones (por ejemplo, recomendaciones musicales o sugerencia de aplicaciones) basadas en reglas específicas. Dado que el algoritmo no asume ninguna estructura de problema a priori, es capaz de adaptarse a las características de los datos de cada sujeto. Cabe destacar la importancia de utilizar un algoritmo con capacidad para adaptarse a los cambios conceptuales, ya que a menudo estos cambios proporcionan la información más relevante.

En conclusión, en este trabajo se muestra el potencial de la minería de reglas de asociación en flujo de datos para la monitorización de problemas reales. El trabajo realizado y los resultados obtenidos revelaron que el desarrollo de un sistema capaz de monitorizar el uso que una persona está haciendo de su teléfono a través de un algoritmo de minería de asociaciones en flujos de datos (utilizando técnicas on-line e incrementales), descubriendo información útil, es una opción factible. En este momento, se están considerando varias líneas para continuar con este trabajo, entre las que se incluyen las siguientes: (1) continuar con las mejoras de Fuzzy-CSar y pulir las adaptaciones del algoritmo a los datos de actividad de teléfonos móviles, (2) estudiar más profundamente los resultados obtenidos en este conjunto de datos, (3) aprovechar la buena eficiencia de Fuzzy-CSar para la integración en dispositivos móviles, y (4) desarrollar aplicaciones para obtener más datos de este tipo utilizando luego la información relevante descubierta a través de las asociaciones aprendidas.

References

[1] J. Gama, *Knowledge discovery from data streams*. Chapman & Hall/CRC, 2010.

[2] M. Sayed-Mouchaweh and E. Lughofer, *Learning in non-stationary environments : methods and applications*. Springer, 2012.

[3] A. Orriols-Puig and J. Casillas, "Fuzzy knowledge representation study for incremental learning in data streams and classification problems," *Soft Computing*, vol. 15, no. 12, pp. 2389–2414, dec 2011.

[4] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan, "Clustering data streams: theory and practice," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 3, pp. 515–528, may 2003.

[5] G. Corral, A. Garcia-Piquer, A. Orriols-Puig, A. Fornells, and E. Golobardes, "Analysis of vulnerability assessment results based on CAOS," *Applied Soft Computing*, vol. 11, no. 7, pp. 4321–4331, oct 2011.

[6] A. Sancho-Asensio, A. Orriols-Puig, and J. Casillas, "Evolving association streams," *Information Sciences*, vol. 334-335, pp. 250–272, mar 2016.

[7] S. W. Wilson, "Classifier Fitness Based on Accuracy," *Evolutionary Computation*, vol. 3, no. 2, pp. 149–175, jun 1995.

[8] R. Agrawal, T. Imieliński, A. Swami, R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93*, vol. 22, no. 2. New York, New York, USA: ACM Press, 1993, pp. 207–216.

[9] D. Dubois, E. Hüllermeier, and H. Prade, "A systematic approach to the assessment of fuzzy association rules," *Data Mining and Knowledge Discovery*, vol. 13, no. 2, pp. 167–192, sep 2006.

[10] A. Orriols-Puig, J. Casillas, and F. J. Martínez-López, "Unsupervised Learning of Fuzzy Association Rules for Consumer Behavior Modeling," *Mathware & Soft Computing*, vol. 16, pp. 29–43, 2009.

[11] N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland, "Social fMRI: Investigating and shaping social mechanisms in the real world," *Pervasive and Mobile Computing*, vol. 7, no. 6, pp. 643–659, dec 2011.

[12] J. Bort-Roig, N. D. Gilson, A. Puig-Ribera, R. S. Contreras, and S. G. Trost, "Measuring and Influencing Physical Activity with Smartphone Technology: A Systematic Review," *Sports Medicine*, vol. 44, no. 5, pp. 671–686, may 2014.