**Revista Anales de la Academia de Ciencias de Cuba Vol.8 No.1**

# DESCUBRIMIENTO DE SECUENCIAS FRECUENTES Y SU APLICACIÓN A LA CLASIFICACIÓN DE DOCUMENTOS.

**UNIDAD EJECUTORA PRINCIPAL:** Centro de Aplicaciones de Tecnologías de Avanzada (División de Investigaciones CENATAV, DATYS Soluciones Tecnológicas).

**AUTOR PRINCIPAL:** Dr.C. José Kadir Febrer Hernández (CENATAV, Datys Soluciones Tecnológicas)

**OTROS AUTORES:** Dr.C. José Hernández Palancar (CENATAV, Datys Soluciones Tecnológicas) Dr.C. Raudel Hernández León (CENATAV, Datys Soluciones Tecnológicas)

**COLABORADORES CIENTÍFICOS:** Dr.C. Claudia Feregrino Uribe (INAOE, México), Dr.C. Andrés Gago Alonso (CENATAV, Datys Soluciones Tecnológicas), Lic. Yadira Font Rezk (CENATAV, Datys Soluciones Tecnológicas)

**AUTOR PARA LA CORRESPONDENCIA:**

José Kadir Febrer Hernández

Dirección: 7maA #21406 e/ 214 y 216, Reparto Siboney, Playa, C.P. 12200

Fax: (+) 537.273.0045

correo electrónico: jfebrer@cenatav.co.cu

**RESUMEN:**

Esta investigación teórica aborda, en una primera etapa, el problema de la minería de secuencias frecuentes sobre conjuntos de datos estáticos. En este proyecto se propone un nuevo algoritmo para la obtención de todas las secuencias frecuentes, siguiendo como estrategia principal la generación de las secuencias candidatas a partir de las secuencias frecuentes de tamaño dos. Además, este algoritmo introduce una nueva estrategia de poda que permite reducir la cantidad de secuencias candidatas, lo cual coadyuva a la eficiencia del algoritmo. En una segunda etapa se propone un nuevo clasificador basado en secuencias frecuentes, el cual, sin pérdida de generalidad, se evalúa en colecciones de documentos. Para ello, se propone un nuevo algoritmo para el cálculo de reglas de clasificación basadas en secuencias, que utiliza la medida de calidad Netconf. Todas las propuestas se validaron a través de experimentos sobre conjuntos de datos sintéticos y conjuntos de datos internacionales utilizados en los trabajos reportados. En los experimentos se utilizaron algoritmos del estado del arte, algunos de ellos proporcionados por sus autores. La novedad científica de este trabajo está avalada principalmente por dos artículos publicados en revistas de impacto internacional, por cuatro artículos publicados en memorias de eventos de impacto

internacional y especializados en el tema, así como por la Tesis de Doctor en Ciencias Matemáticas del autor principal. El aporte de esta investigación viene dado, fundamentalmente, por el desarrollo de nuevos métodos, estrategias y algoritmos que mejoran la eficiencia en el cálculo de las secuencias frecuentes y la eficacia de los clasificadores basados en secuencias que utilizan reglas de clasificación.

## COMUNICACIÓN CORTA DE LOS RESULTADOS

1. Descripción de la problemática existente

Los grandes avances tecnológicos de los últimos años han provocado un aumento considerable en los volúmenes de información generados por los sistemas informáticos. Actualmente, casi cualquier dispositivo electrónico o electrodoméstico es capaz de producir información y enviarla a través de las redes. Actividades tan simples como enviar un correo electrónico o realizar una compra por Internet se generan diariamente en volúmenes tan grandes que son imposibles de analizar por un humano. Como solución para enfrentar este problema surge la Minería de Datos, la cual consiste en extraer información no trivial, previamente desconocida y útil de los grandes conjuntos de datos.

Específicamente, la Minería de Secuencias es uno de los campos de investigación de la Minería de Datos que ha alcanzado un gran auge en los últimos años debido al amplio campo de aplicaciones que presenta. Dentro del mismo existen diferentes técnicas y formas de analizar y extraer conocimiento de los datos como son la minería de secuencias frecuentes y la clasificación basada en secuencias.

La minería de secuencias frecuentes consiste en el descubrimiento de secuencias interesantes que se encuentran en un conjunto de datos secuencial una cantidad de veces mayor que un valor especificado. La complejidad que presenta este proceso es exponencial, causado por la gran cantidad de posibles combinaciones de secuencias que se pueden formar para un conjunto finito de elementos (ítems). Por lo que las investigaciones en este campo van dirigidas principalmente a lograr una mayor eficiencia en el proceso de obtención o descubrimiento del conjunto de las secuencias frecuentes.

Por otro lado, la clasificación basada en reglas consiste en encontrar el conjunto de reglas representativas de cada una de las clases involucradas en un determinado problema y clasificar, a partir de estas, a cualquier nuevo elemento que se desee. Para esto, primero, y a través de un proceso conocido como proceso de entrenamiento, se obtiene el conjunto de reglas que representarán a cada una de las clases. Luego, mediante una estrategia de ordenamiento y selección de las reglas, se escoge a que clase corresponderá un elemento nuevo que se quiera clasificar. En este caso, las investigaciones van dirigidas hacia una mejora en la eficacia que logran los clasificadores en el momento de clasificar a un nuevo elemento, ya sea por mayor cantidad de reglas que representan a las clases, reglas de una mayor calidad o de más representatividad dentro de las clases, o el uso de mejores medidas de calidad con las que se determinan y escogen las reglas representativas. En el caso de la clasificación

basada en reglas dentro de la minería de secuencias, se utiliza como antecedente de la regla (parte izquierda de la regla) una secuencia y en el consecuente la clase a la que pertenece dicha secuencia.

En este trabajo investigativo se propusieron nuevos métodos, estrategias y algoritmos con el objetivo de mejorar la eficiencia en la obtención de secuencias frecuentes en conjunto de datos estáticos. Para lograr estos resultados el trabajo se centró, fundamentalmente, en una nueva forma candidatas que se generan para como consecuencia de esto mejorar la eficiencia de este tipo de minería.

También, en una segunda etapa, se propone un nuevo clasificador basado en reglas donde el antecedente de estas son secuencias frecuentes, logrando así mejorar la eficacia de los clasificadores existentes basados en secuencias, sobre todo para colecciones de documentos.

## 2. Novedad científica

Sobre la base de lo explicado anteriormente, el estudio del estado del arte de la minería de secuencias y los problemas detectados en la literatura relacionada, se obtuvieron varios resultados importantes los cuales se describen a continuación:

1. Se propuso y desarrollo una nueva estrategia de generación de secuencias candidatas y una nueva estrategia de poda que, de conjunto, reducen el número de secuencias candidatas y hacen más eficiente el proceso de la minería de secuencias frecuentes en conjuntos de datos estáticos. La ventaja de la propuesta está en generar un menor número de secuencias candidatas a través del desarrollo de una nueva forma de generar las secuencias que permite de esta forma reducir las secuencias que se deben analizar, así como la nueva estrategia de poda que mejora la eficiencia de este proceso al poder conocer si las secuencias son frecuentes o no, sin tener que realizar el proceso de conteo de las mismas.

2. Se propuso y desarrollo un nuevo algoritmo para el cálculo de secuencias frecuentes en conjuntos de datos estáticos más eficiente que los existentes en la literatura para este tipo de minería de datos. El nuevo algoritmo desarrollado incluye los aportes y las propuestas que aparecen en el punto anterior.

3. Se propuso utilizar una nueva medida de calidad para el cálculo del conjunto de reglas de asociación de clases basada en patrones secuenciales, la cual no tenga las limitaciones que presentan las medidas del soporte y la confianza. Además, se introduce una nueva estrategia de ordenamiento basado en la medida de calidad seleccionada. Las medidas de calidad más utilizadas en el cálculo para la obtención de reglas son el soporte y la confianza las cuales, a pesar de su amplio uso, son medidas que presentan una serie de problemas las cuales se tratan de suplir con la utilización de la medida seleccionada.

4. Se propuso y desarrollo un nuevo algoritmo que utiliza la medida de calidad seleccionada de la propuesta anterior, para obtener el conjunto de las reglas. Además, se introduce una nueva estrategia de poda que permite generar reglas de mejor calidad. La calidad de las reglas está determinada por el valor de la medida que se seleccionó y la forma en que se podan (eliminan) las reglas.

5. Se propuso y desarrollo un nuevo clasificador basado en las reglas obtenidas por el algoritmo del punto anterior que alcanza mayor eficacia que los clasificadores existentes en la literatura, particularmente en colecciones de documentos. El nuevo clasificador con las estrategias de ordenamiento y selección de las reglas es capaz de determinar con mayor eficacia a que clase pertenece un nuevo elemento.

## 3. Aportes, impactos y novedad del resultado

Los aportes de este trabajo contribuyen a avanzar en el conocimiento del área de la minería de datos, específicamente en el campo de la minería de secuencias, tanto a nivel nacional como internacional. El aporte de esta investigación viene dado, fundamentalmente, por el desarrollo de nuevos métodos, estrategias y algoritmos que mejoran la eficiencia en el cálculo de las secuencias frecuentes y la eficacia de los clasificadores basados en secuencias que utilizan reglas de clasificación.

Los nuevos aportes en el área de la minería de secuencias frecuentes permiten mejorar, con relación al estado del arte, la eficiencia del cálculo del conjunto de las secuencias frecuentes y reducir la cantidad de secuencias candidatas que se generan, aspectos de gran importancia en este tipo de minería.

Los métodos y estrategias para la generación y poda de las secuencias candidatas desarrollados permiten reducir la cantidad de secuencias candidatas que se generan contribuyendo así en la mejora de la eficiencia del cálculo del conjunto de las secuencias frecuentes. En este sentido, el impacto del aporte se manifiesta en el hecho de que estos métodos permiten reducir el tiempo (la eficiencia) de la búsqueda de secuencias frecuentes en grandes conjuntos de datos sin afectar la eficacia del resultado.

Por otro parte, los nuevos aportes realizados en el área de clasificación basada en secuencias permiten obtener reglas de mayor calidad, mejorando de esta manera la eficacia de los clasificadores basados en reglas. Las nuevas estrategias y algoritmos desarrollados permiten obtener el conjunto de las reglas más reducido y de mayor calidad que los del estado del arte. La selección de una medida de calidad que no presenta los problemas del soporte y la confianza (medidas más utilizadas para este tipo de clasificadores) y una nueva estrategia de poda que permite que las reglas obtenidas sean de mayor calidad ayudan a crear un nuevo clasificador que tiene mayor eficacia que los del estado del arte.

El impacto del uso de la minería de secuencias en problemas de minería de datos va más allá del hecho de proporcionar soluciones superiores en eficiencia (algoritmo de minería de secuencias frecuentes) y eficacia (clasificador basado en secuencias) a los existentes en la actualidad. “Las actividades de vigilancia e inteligencia, en el contexto de la sociedad de la Información, se considera como la transformación de la información en conocimiento, y del conocimiento en acción. Es por esto la importancia del análisis de la información en el seno de todas las actividades que se trata de transformar los datos brutos con el fin de extraer los conocimientos que puedan ser explotados y útiles en un determinado campo de acción”. Una manera de extraer este conocimiento de los datos es a través de la Minería de Secuencias.

El contar con métodos con *know how* propio, constituye un aspecto básico y un gran aporte para garantizar la sostenibilidad de los resultados alcanzados y le da un valor agregado a las aplicaciones y productos que se desarrollen a partir de estos resultados.

La novedad de estas propuestas está avalada por la aceptación en eventos y revistas, en algunos casos de un alto impacto a nivel internacional, de los artículos donde se describen los mismos. Estos artículos deben transitar por un proceso de revisión a ciegas por revisores internacionales asignados por un comité de manera que los revisores conozcan de los temas, y una de las principales características que revisan estos revisores es la novedad de lo que se propone. Debido a esto, la publicación de estos artículos demuestra el impacto a nivel mundial que representan los resultados obtenidos.

Los resultados de esta investigación forman la base teórica del conocimiento en el campo de la Minería de Datos, particularmente en la Minería de Secuencias. Dirigidas para aplicaciones donde sea importante obtener patrones secuenciales o clasificar. Específicamente, los métodos, estrategias y algoritmos propuestos en este trabajo proveen una solución a un grupo de problemas de la minería de datos a partir de patrones secuenciales.

Estos resultados investigativos permiten incorporar conocimientos propios a nuestros sistemas y evitar la compra e inversión en sistemas extranjeros de altos costos de compra y soporte, que muchas veces no pueden ser adquiridos por nuestro país debido al bloqueo impuesto por el gobierno de los Estados Unidos.

Los métodos propuestos proveen una solución a los problemas ministeriales relacionados con el análisis de datos. Con las propuestas desarrolladas se pueden, entre otras cosas, detectar fraudes bancarios, problemas en el flujo de los procesos o anomalías, comportamiento de las personas, clasificación de documentos o de hechos, predicción de hechos, el flujo de los usuarios en Internet, entre muchos otros.

Nivel de introducción en la práctica y/o potencialidades concretas de introducción a corto plazo

Los métodos y algoritmos desarrollados, como resultado de esta investigación, serán introducidos en la práctica a través de proyectos relacionados a la inteligencia financiera desarrollados de conjunto con organismos nacionales. Los algoritmos desarrollados que formarán parte de estos proyectos permitirán entre otras cosas:

- Obtener las secuencias de operaciones más frecuentes que realizan las personas.
- Predecir posibles operaciones o tipos de operaciones a ejecutarse.
- Ayudar en el análisis del flujo de operaciones que se llevan a cabo diariamente.
- Clasificar los tipos de secuencias de operaciones que se realizan.
- Determinar el comportamiento de las personas mediante las secuencias de eventos realizados.

Lo expresado anteriormente permitirá ayudar a un mejor control y análisis de los diferentes procesos que son ejecutados diariamente en el sector financiero, ayudando de esta manera a un mayor control y desarrollo de nuestro país.

También, se está trabajando en estos momentos en varias propuestas de aplicaciones de los resultados obtenidos como son el algoritmo de secuencias frecuentes y el clasificador basado en reglas, en proyectos que están siendo desarrollados por la división de Datys-Santiago. Además, se está evaluando la posibilidad de introducir estos resultados en productos de la Empresa relacionados con la minería de textos y el análisis en redes sociales.

Por otro lado, se está analizando la posibilidad de utilizar los algoritmos desarrollados para predecir y detectar el movimiento de las personas a través de llamadas desde celulares. Igualmente, se está estudiando aplicar estos métodos en determinar el comportamiento de grupos de turistas.