**PROGRAMAS Y LÍNEAS DE INVESTIGACION**

**PROGRAMS AND LINES OF RESEARCH**

Área C Ingenierías al que pertenece el Instituto de Investigación - Facultad de Ingeniería de Sistemas e Informática, Aprobado según Resolución Rectoral N° 00017 del 07 de enero del 2014.

**Base Legal:**

* Estatuto de la UNMSM
* Ley Universitaria 23733
* Reglamento de Gestión de Actividades de Investigación dé la Universidad Nacional Mayor de San Marcos aprobado por RR N° 03063-R-13 del 05 de julio 2013.

**AREA C: INGENIERIAS**

**PROGRAMA C.0.3. TECNOLOGIAS DE LA INFORMACION Y COMUNICACION**

**LINEAS DE INVESTIGACION**

C.0.3.1. Automatización en adquisición de datos

C0.3.3. Desarrollo de modelos y aplicación de las tecnologías de información y comunicaciones

C.0.3.4. Inteligencia de negocios

C..0.3.5 Comunicaciones inalámbricas

C.0.3.6 Comunicaciones ópticas

C.0.3.7. Tecnologías IP

C.0.3.10. Redes de datos

C.0.3.13. Herramientas colaborativas para el aprendizaje e-learning

C.0.3.14. Procesamiento de voz

C.0.3.15. Computación Grafica

C.0.3.16. Procesamiento de imágenes y visión por computadora

C.0.3.17. Reconocimiento de patrones

C.0.3.18. Informática y Educación

C.0.3.19. Informática y sociedad

C.0.3.20. Gestión de sistemas informáticos y de información

C.0.3.21. Técnicas de programación

C.0.3. 22. Ingeniería de Software

C.0.3.23. Administración de base de datos

C-0.3.24. Almacenamiento y recuperación de datos

C..0.3.25. Tecnología de información y aplicaciones de sistemas

C.0.3.26. Procesamiento de datos administrativos

C:0.3.27. Aplicaciones de internet

C.0.3.28. Comunicaciones/redes y sistemas distribuidos

**PROGRAMA C.0.4. OPTIMIZACION DE PROCESOS DE TRANSFORMACION Y MANUFACTURA**

**LINEAS DE INVESTIGACION**

C.0.4.1. Desarrollo de modelos de simulación y optimización de procesos

C0.4.10. Sistemas de Control

C0.4.11. Robótica

C.0.4.12. Instrumentación virtual

C.0.4.13. Instrumentación digital

C.0.4.15. Simulación, modelado y visualización

**PROGRAMA C.0.5. DISEÑO Y APLICACION DE NUEVAS TECNOLOGÏAS**

**LINEAS DE INVESTIGACION**

C.0.5.5. Sistemas embebidos

C.0.5.6. Arquitectura de computadoras

C.0.5.7. Arquitectura de VLSI

C.0.5.8. Inteligencia Artificial

**PROGRAMA C.0.6. SEGURIDAD, SALUD Y SOCIEDAD**

**LINEAS DE INVESTIGACION**

C.0.6.8. Seguridad de la información y de la comunicación

**INSTRUCTIVOS PARA LA PRESENTACION DE ARTICULOS**

La Revista Peruana de Ingeniería de Sistemas e Informática RISI es una publicación científica editada por el Instituto de Investigaciones de Ingeniería de sistemas e Informática, Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos, Lima Perú. Tiene una periodicidad semestral, y se publica tanto en su versión impresa como digital.

La Revista RISI recibe artículos completos, originales e inéditos en los temas de relacionados con el campo de la Ingeniería de Sistemas, Ingeniería de Software, Ciencias de la Computación e Informática, elaborados según las normas indicadas en las presentes instructivos.

El articulo deberá ser presentado acompañado de una carta dirigida al Editor Jefe, firmada por el responsable del trabajo con quien se tendrá comunicación, indicando además el carácter inédito, original y completo del articulo presentado y su disposición para que sea revisado y editado.

Las publicaciones pueden estar dentro de las categorías:

1. **TRABAJOS ORIGINALES. A**rtículos inéditos que presentan resultados de trabajos de investigación y constituye aportes al conocimiento. Todo el artículo debe tener como máximo 12 páginas, las ilustraciones deben ser solo las necesarias para una mejor exposición de los resultados.
2. **NOTAS CIENTÍFICAS**. Reportes de resultados cuya información es de interés para la comunidad científica. La extensión del texto no será mayor de 6 páginas.
3. **ARTÍCULOS DE REVISIÓN**. Trabajos que constituyen una exhaustiva revisión del tema de investigación del autor, se incluyen aquí tesis, revisiones taxonómicas y recapitulaciones.

**REDACCION DEL ARTICULO**

**GENERALIDADES**

1. El documento deberá redactarse cumpliendo lo siguiente:  
   1. Idioma : Español , utilizando Microsoft Word
   2. Formato : A4 , tipo de letra : Time New Román 12
   3. Márgenes: 25 mm, a espacio y medio (con excepción de los cuadros). Izquierda 30 mm y derecha 25 mm.
   4. Numeración : Inicie con el numero 1, consecutivamente en la parte central inferior de cada página.
2. Los títulos principales (encabezados de cada sección: Resumen, Abstracto, Introducción, Material y Métodos, Resultados, Discusión y Literatura Citada) van centrados, en mayúsculas, y en negrita.

1. Los títulos de primer orden se colocan en negrita, en mayúsculas tamaño 12, al margen izquierdo, en renglón separado, y sin puntuación final. El texto que se le sigue se ubica en párrafo aparte. Los títulos de segundo y tercer orden se colocan en cursiva, al margen izquierdo, en minúscula. Pueden ir en renglón separado sin puntuación final, o al inicio de la primera línea del párrafo, seguidos de un punto.
2. Los párrafos se justifican a ambos lados y se separan entre sí con un renglón en blanco.
3. Utilice frases breves y precisas en la redacción del documento, con los verbos en forma activa y evitando el uso de la primera persona.
4. Los cuadros y figuras se ordenan con numeración arábiga y con los títulos auto-explicativos, de manera consecutiva en cada artículo y en la parte inferior.
5. Los nombres Científicos, productos y herramientas tecnológicas se colocan en cursiva.
6. En caso que el articulo exponga sobre experimentos con humanos y animales, los procedimientos deberán de ceñirse a la Declaración de Helsinki de 1975 y a las leyes peruana vigentes (Ley 27265), debiendo presentarse las declaraciones pertinentes y mencionadas en el texto.

**Página Inicial**

1. La primera página del documento deberá contener:

a) Título del artículo en español e inglés,

b) Titulo corto,

c) Nombre y apellidos de cada autor y, al pie, su afiliación institucional, correo electrónico,

2 El Titulo deberá:

* + 1. Contener a lo más 20 palabras
    2. Identificar el contenido real del estudio.
    3. Ser descriptivo, breve y claro
    4. No contener siglas ni abreviaciones.
    5. El titulo corto es un titulo abreviado, no mayor de 80 caracteres incluyendo espacios, que aparecerá en la parte superior de las páginas impares del trabajo publicado.

1. El Autor

Deberá haber participado lo suficiente como para asumir la responsabilidad pública del contenido del trabajo. Además, debe haber contribuido en forma sustancial en:

1. La concepción y el diseño del estudio, o en el análisis y la interpretación de los datos;
2. La redacción del artículo o la revisión crítica de una parte sustancial de su contenido intelectual; y
3. La aprobación final de la versión que será publicada.

**Resumen**

1. El Resumen deberá contener el resumen en español seguido del abstract en inglés conteniendo la misma información que el resumen.
2. El Resumen deberá estar contenido en una sola pagina
3. El resumen deberá contener los objetivos, procedimientos, resultados principales y las conclusiones de la investigación en un solo párrafo, no mayor de 300 palabras y en forma clave y concisa.
4. No puede contener cuadros, figuras, citas bibliográficas, ni abreviaturas o acrónimos, salvo que sean de uso común o se definan en el mismo Resumen.

**Palabras claves**

Después del Resumen (y del Abstract), dejando un renglón en blanco, escriba 4 a 6 palabras claves (key words) para efectos de indización. Estas deberán ser colocadas en minúscula, a excepción de nombres propios, separadas por comas y sin puntuación final.

**Introducción**

Deberá cumplir lo siguiente

* 1. Describir la naturaleza del problema en estudio,
  2. Presentar trabajos relacionados o realizados por otros autores en el campo bajo estudio.
  3. Indicar como se pretende solucionar el problema, de que manera se aumenta el acervo de conocimientos existentes sobre el tema.
  4. Utilizar no más de tres referencias para apoyar un concepto o idea.

**Materiales y Métodos**

En caso se utilice, deberá indicarse:

1. **Materiales :** Describir el objeto de estudio, sus características, la localización geográfica y periodo del estudio, así como los mecanismos utilizados para el desarrollo del trabajo, incluyendo la forma de recolección de datos, el diseño experimental (si lo tuviera) y el tipo de análisis estadístico utilizado.
2. **Métodos :** Indicar los métodos, proporcionando las referencias cuando sean necesarias y lo equipos de importancia utilizados.
3. **Técnicas :** Usar referencias para las técnicas utilizadas, o describa en detalle si son nuevas o han sido modificadas.

**Resultados**

Deberá:

* 1. Presentar los resultados y una discusión de los mismos
  2. Presentar la información en forma de texto, cuadros y figuras, y siguiendo una secuencia lógica.
  3. Describir los hallazgos encontrados y el resultado de los análisis, apoyado con cuadros y figuras.
  4. No repita en el texto los datos de los cuadros o figuras, sino resalte o resuma las observaciones más importantes. No repita conceptos indicados en la metodología.
  5. Puede combinarse la sección de resultados con la Discusión

**Discusión**

Deberá:

* + 1. Explicar el significado de los resultados con el apoyo de referencias bibliográficas, así como sus implicancias en futuras investigaciones. Compare las observaciones realizadas con aquellas de otros estudios pertinentes. Enfatice en aspectos nuevos e importantes del estudio y en las conclusiones que se deriven de ellos.
    2. Discutir las implicancias teóricas o prácticas del estudio.
    3. Evitar afirmaciones o conclusiones insuficientemente avaladas por los datos o con resultados sin soporte estadístico.
    4. No repetir los datos de información presentada en las secciones previas.
    5. No hacer afirmaciones sobre costos o beneficios económicos, salvo que en el trabajo se incluyan datos y análisis económicos.

**Conclusiones**

* + - * 1. Presentar en forma muy breve los principales hallazgos y planteamientos que fueran producto del estudio. Use viñetas si se tiene varias conclusiones.
        2. Evitar conclusiones y opiniones personales que no estén respaldadas por los datos presentados.
        3. No sugiera continuar investigando sobre el tema. Esto no constituye una conclusión.

**Agradecimientos**

De haberlas:

Reconocer la ayuda de personas e instituciones que aportaron significativamente al desarrollo de la investigación, así como la entidad que financió el estudio.

**Literatura Citada**

Las referencias son importantes en todo trabajo académico pues indican al lector las distintas fuentes de sus citas o ideas trabajadas en su investigación. El mundo académico es sumamente estricto en este sentido y una falta al indicar las fuentes equivale a plagio (o hurto literario). En este sentido, el propósito de un sistema de referencias es describir las fuentes de manera exacta y rigurosa, para indicar -en el mismo trabajo- dónde y cómo fueron empleadas.

**Deberá:**

1. Ordenarse alfabéticamente por el apellido del autor. El formato requiere que los títulos de libros, diarios, etc. sean destacados utilizando tipografía itálica (conocida también como cursiva). Si tiene dos o más referencias de los mismos autores se ordenan por año de publicación; si estos están publicados en el mismo año se ordenan por orden alfabético de los títulos de los artículos, agregando un sufijo al año de publicación (por ejemplo, 2005ª, 2005b); y si tiene dos o más referencias del mismo año de publicación y con tres o más autores, donde el primer autor es el mismo, se les agrega el sufijo al año de publicación como en el caso anterior.
2. Limitar el número de referencias a las más pertinentes o de mayor actualidad.
3. Solo la primera palabra y los nombres propios en los títulos de las publicaciones van en mayúscula.
4. Evitar citar resúmenes y literatura de publicaciones que no sean científicas.

**Libros**

La información para citar un libro usualmente está consignada al inicio de la obra, en el reverso de la página del título. Asegúrese de identificar correctamente el nombre la casa editorial y no confundirla con el de la imprenta. Esta es una información muy importante en su lista de referencias. El Catálogo de su Biblioteca le proporcionará el nombre de la editorial si tuviese alguna duda. No considere las fechas de reimpresión; lo que debe considerar es la fecha en que se publicó la primera, segunda o la tercera edición de la obra, la misma que debe corresponder con la edición que esté empleando en su trabajo. En algunos casos, es importante señalar la fecha de la primera edición o versión original del libro.

Moran Cortez, Enma Teolinda,. *Compactación de memoria*, 6ª ed. McGrawHill 2008. Madrid España150p.

Pro Concepcion, Luzmila,. *Algoritmia, Técnicas Algorítmicas*, CEPREDIM 2012 Lima Perú. 390p.

**Artículos de Revistas Científicas**

La información necesaria para citar un artículo de revista puede encontrarse, generalmente, en el índice, portada o en el artículo mismo.

Cortez Lama, Parcemon,. Sistemas de razonamiento basado en reglas para determinar recomendación de cirugía refractiva Rev. Algorithmic Lima Perú, vol. 1 N°1 2010, pag. 23-28.

**Fuentes electrónicas**

El patrón básico para una referencia electrónica es:

Autor, inicial(es) de su nombre (año). Título. Mes, día, año, de la dirección en Internet.

Bancos, I. (n.d.). Las redes neuronales y la solución de problemas del sector salud. Obtenida el 29 de agosto de 2008, de <http://www.coreguide.direct.nhs.uk/>

Si no consigue identificar la fecha en que el documento fue publicado, utilice la abreviatura n.d. (no date [sin fecha]).

Si no consigue identificar al autor, empiece su referencia con el título del documento.

Si el documento se ubica dentro de una página institucional, como la de alguna universidad o departamento gubernamental, primero cite el nombre de la organización o del departamento en cuestión, antes de dar la dirección electrónica:

Cortez Vasquez , A., & Villen, J. A. (2003). Evaluando las Fuentes Electrónicas. Consultado el 21 de agosto de 2001, Widener University, página web conmemorativa de la biblioteca Wolfgram: [http://www.ridener.edu/ Library/webevaluation /](http://www.ridener.edu/%20Library/webevaluation%20/) webeval.htm

Construyendo una idea. (2005). Consultado el 5 de septiembre de 2001, Portsmouth University, página web de Servicios Profesionales:

http://www.port.ac.uk/departments/careers/plancareer/deciding-your-future.htm

Artículos electrónicos de revistas científicas que a su vez son reproducción de la

versión impresa

Emplee el mismo formato de referencia que utiliza para un artículo de revista científica impresa y agregue "versión electrónica" entre corchetes, después del título del artículo:

Pariona Quispe, Hugo. (2001). Un modelo de sistemas expertos para dietas caseras [versión electrónica ]. Journal of Common Market Studies, 39(3), 228- 239.

Si tiene que citar un artículo electrónico cuya versión se diferencia de la versión impresa, o incluye datos o comentarios adicionales, debe agregar la fecha en que usted consultó el documento en la web y su respectiva dirección (URL).

**Artículos de revistas científicas que sólo se publican en la web*.***

Utilice la fecha completa de publicación que figura en el artículo.

Cerciórese de que no tenga paginación.

Siempre que sea posible, procure que la dirección electrónica que cite (URL) remita directamente al artículo.

Evite citar una dirección electrónica en dos líneas y cuide que el enlace (URL) no se corte después de un guión o antes de un punto. No inserte guiones en el enlace cuando esto ocurra.

Korda, L. (2001, Julio). La fabricación de un traductor. Translation Journal, 5(3). Consultada el 21 de agosto de 2001, http://accurapid.com/journal/17prof.htm

**Artículos obtenidos de una base de datos**

Utilice el formato apropiado al tipo de trabajo obtenido y agregue la fecha de recuperación del material más el nombre de la base de datos:

Moran Cortez, T. (2000, Julio 9). Distribución de archivos en un sistema inteligente. The Observer, p.7. consultado el 10 de septiembre de 2001, en The Porter y The Observer, en su base de datos en CD-ROM.

1. ***Capítulos de libro***

Lama J. 1999. Sistemas expertos. En: Hafez ESE, Hafez B, eds. Reproducción e inseminación artificial en animales. 7ª ed. México: McGrawHill. p224-242.

1. ***resúmenes de congresos o reuniones***

Argenti P, Ly J, Espinoza F. 2007. Evaluación del efecto de secuestrantes en el control de micotoxinas en alimentos para cerdos en crecimiento. En: XX Reunión ALPA. Cusco: Asociación Latinoamericana de Producción Animal.

1. ***Tesis***

PARIONA.J 2008 Detección de errores en el análisis sintáctico de un compilador. Tesis de Lic. en Computación. Lima: Universidad Nacional Mayor de San Marcos. 48 p.

1. ***Informe científico o técnico***

[UNMSM] Instituto de Investigación. 2001. Factores de riesgos en la transmisión electrónica de datos. Lima Peru.541 p.

1. ***Material electrónico en CD ROM***

Amadeo Vásquez Tavara. 2003. Redes Neuronales Artificiales [CD-ROM]. 13ª ed. Madrid: Editorial Medica Panamericana. 363 p.

1. ***Articulo en revista electrónica***

Aparicio-Acosta FM. 2000. Universidad , Ciencia y Tecnología [Internet], [28 noviembre 2007]. Disponible en: <http://www.uk.es/bdARTE/v6n.htm>

1. ***Sitio de Web***

Portal Agrario. 2007. Lima: Ministerio de Agricultura. [Internet], [12 agosto 2007]. Disponible en: <http://www.minag.gob.pe/>

**Cuadros**

1. Las tablas deben de haber sido citados en el texto.
2. El titulo en la parte inferior comienza con la palabra “Tabla”, seguida del número que le corresponda. Va en minúscula con excepción de los nombres propios y acrónimos. No lleva puntuación al final del título.
3. Cada columna debe ser identificada y solo la primera letra del encabezado va en mayúscula. Las explicaciones se colocan como notas al pie.

**Figuras**

1. Las figuras incluyen gráficos, fotografías, imágenes, mapas y diagramas. Deben ser citadas en el texto.
2. El titulo y la leyenda se colocan en la parte inferior de cada figura. El titulo comienza con la palabra “Figura”, seguida del número que le corresponda. Va en minúsculas, a excepción de los nombres propios y acrónimos.
3. Si la figura ha sido previamente publicada (a excepción de documentos de dominio público), cite la fuente original al final de la leyenda.

**Unidades de medida**

* + - 1. En los valores numéricos, el punto señala la separación entre los números enteros y las fracciones (ejemplo: 6.35). Se usa la coma para separar números con más de tres dígitos (ejemplo: 9,232.5). Se coloca el 0 a la izquierda del punto en número y la unidad que le corresponde.
      2. No comience una frase con un número. Utilice otra expresión o deletree el número y la unidad que le corresponde.
      3. Al hacer referencia a porcentajes, el símbolo se coloca junto al número sin espacio en blanco entre ellos (85%).
      4. Cada vez que se mencione un número inferior a 10 y que no vaya seguido de una unidad de medida, éste debe deletrearse (ejemplo: cuatro archivos, 14 procesos, dos gigabytes, etc.). Si se menciona una serie de elementos semejantes que incluyen numero mayores y menores de 10, todos se colocan en caracteres numéricos.

**PRESENTACIÓN DEL ARTICULO**

* + 1. La presentación del articulo deberá hacerla personalmente el responsable del artículo, debiendo presentar dos copias impresas del manuscrito en papel Bond A4 (Las figuras que deban ir a color deberán ser impresas a color) y en un CD debidamente rotulado, conteniendo el trabajo en un archivo de Word, las ilustraciones en formato jpg. La Oficina de la RISI: Av. Amezaga s/n CU San Marcos – Pabellón Facultad Ingeniería de Sistemas, 3er piso Oficina. Lima. Tel 6197000-3604
    2. Si hubiera más de un autor, el responsable debe ser el primero de la lista consignada en la presentación.
    3. La RISI no cobra costos de impresión a los autores.

**PROCESOS DE REVISIÓN**

1. El Comité Editorial (CE) de la RISI verificara que el articulo presentado se circunscriba a la temática de la revista y que el formato del documento se encuadre dentro de las especificaciones solicitadas a los autores.
2. Los trabajos que cumplan con estos requisitos serán derivados a los revisores científicos para su revisión.
3. Los revisores mantienen la confidencialidad del contenido del manuscrito, realizan una revisión profesional al documento y preparan su informe con comentarios y sugerencias en forma clara y precisa.
4. Los autores cuyos trabajos sean sujeto de correcciones deberán resolverlas y devolver una carta aceptando las sugerencias, presentando las modificaciones o justificando las razones para no modificar.
5. El CE, en base a la respuesta de los autores, aprobará o rechazará el articulo y le comunicara la decisión final a los autores. Así mismo, enviara una carta electrónica a los autores de trabajos indicando las razones correspondientes.

Diseño de un Sistema de Recuperación de Imágenes de individuos malhechores para seguridad ciudadana

Design a system for recovering images of individuals criminals for citizen security

**Nora La Serna Palomino, Luzmila Pró Concepción**

**Universidad Nacional Mayor de San Marcos**

**Facultad de Ingenieria de Sistemas e Informatica**

[**nlasernap@unmsm.edu.pe**](mailto:nlasernap@unmsm.edu.pe)**, lproc2003@hotmail.com**

# RESUMEN

El trabajo que se presenta en este artículo se desarrolla en el marco de Los Sistemas de Recuperación de Imágenes faciales, en particular se desenvuelve en el contexto de la seguridad ciudadana. Se ha realizado el análisis y diseño de una herramienta de software, que es un Sistema de recuperación de imágenes faciales de individuos. Las imágenes que se registran son de individuos que han cometido actos delictivos y tienen antecedentes penales. El sistema consta de dos módulos: 1) Almacenamiento o registro de descripciones textuales de imágenes faciales; y 2) Recuperación de una imagen en particular ante una consulta de un usuario. Su desarrollo se basa en el Modelo de Espacio Vectorial. De esta manera, el Sistema de Recuperación de Información de Imágenes de individuos malhechores son una herramienta tecnológica de apoyo al proceso de combatir la delincuencia, en la que a partir de un repositorio de imágenes de individuos malhechores, el sistema permite identificar individuos sospechosos que se encuentran en un ambiente público, como en buses, metro, calle, etc.

**Palabras claves**: Sistema de Recuperación de Imágenes, Seguridad ciudadana, Modelo de espacio vectorial, algoritmos de recuperación.

# ABSTRACT

The work that is presented in this article has being developed in the area of The Image Retrieval Systems. Especially, it is unrolled in the context of the civil safety. There has been realized the analysis and design of a tool of software, which is a Retrieval System of facial images of individuals. The images that are registered belong to individuals who have committed criminal acts and have penal precedents. The system consists of two modules: 1) Storage of textual descriptions of facial images; and 2) Recovery of an image before a consultation of a user. Its development is based on the Vector Space Model. Hereby, the Image Retrieval System of Recovery of individual Images of malefactors they are a technological tool of support to the process of attacking the delinquency. In that, from a repository of images of individual malefactors, the system allows to identify suspicious individuals who are in a public environment, since in buses, meter, street, etc.

**Key words**: Image Retrieval System, Civil safety, Vector space Model, retrieval algorithms.

**1. INTRODUCCIÓN**

El trabajo que se presenta en este artículo se desarrolla en el marco de Los Sistemas de Recuperación de Imágenes, estos sistemas tienen muchas aplicaciones en diferentes sectores del quehacer humano, pero en particular ayudan significativamente en temas de seguridad ciudadana, los cuales se podrían aplicar en la ciudad de Lima, en donde a pesar de los esfuerzos de las autoridades presentan un alto índice delincuencial.

Estos Sistemas de Recuperación de información visual tales como imágenes y video, permiten recuperar una imagen específica desde un conjunto de imágenes almacenadas. Para buscar la imagen solicitada por un usuario, se utilizan los siguientes tipos de información: a) Textual, b) Basado en el contenido de la imagen, y c) Podrían ser ambos información textual y basados en contenido de la imagen.

De otra manera, el Perú tiene una alta incidencia en delitos y faltas como hurto, robo, robo con arma blanca y de fuego. Incluyen la violencia familiar y delito sexual. En la sección 2.2 se presenta un gráfico en el que se observa que la delincuencia va en aumento, y es uno de los problemas graves en el país en los últimos años [6,7]. Gráficamente, en la Figura 1, se muestra el panorama que día a día ocurre en nuestro país y que se muestran en los medios de comunicación.

El presente trabajo desarrolla una herramienta que es un Sistema de Recuperación de imágenes faciales de individuos que han cometido actos delictivos y tienen antecedentes penales. El sistema consta de dos módulos: 1) Almacenamiento o registro de descripciones textuales de imágenes faciales; y 2) Recuperación de una imagen en particular ante una consulta de un usuario. Su desarrollo se basa en el Modelo de Espacio Vectorial [4,14].



Figura 1. El problema de la delincuencia en el país [Fuente: medios de comunicación].

La estructura del resto del presente artículo es la siguiente: En la sección 2 se desarrolla el Marco Teórico Referencial de los Sistemas de Recuperación de imágenes, y la Seguridad Ciudadana en nuestro país. En la sección 3 se describen los métodos de búsqueda para imágenes; Mientras que en la sección 4 se presenta el diseño del sistema propuesto; la sección 5 corresponde a las Conclusiones del trabajo realizado y se proponen tareas futuras para su implementación; y finalmente en 6 se presentan las referencias bibliográficas utilizadas.

**2. MARCO TEÓRICO REFERENCIAL**

En esta sección se desarrollan los Sistemas de recuperación, y la Seguridad ciudadana en el Perú.

**2.1 Sistemas de recuperación**

Un sistema de recuperación de información almacena grandes volúmenes de documentos, los cuales pueden venir de procesadores de textos, de páginas Web, de bases de datos en línea, documentos electrónicos e inclusive otros archivos. A la vez el sistema debe disponer de interfaces que permitan hacer consultas en lenguaje natural acerca de los documentos que se encuentran almacenados. Para ello utiliza estructuras de datos y operaciones que permiten el almacenamiento y recuperación eficiente de la información. Uno de los modelos mayormente utilizados en estos sistemas, es el *espacio vectorial*, el que se introduce en la sección 3.2 [4, 14].

Sin embargo, también existen Sistemas de Recuperación de información visual tales como imágenes y video, inclusive para audios. En este trabajo nos enfocamos a la recuperación de imágenes, el cual tiene muchas aplicaciones, en investigaciones, en seguridad, etc. en estos sistema para buscar una imagen utilizan las siguientes: información textual, información basado en el contenido de la imagen- conocidos también como CBIR-, y también podrían ser ambos información textual y basados en contenido de la imagen.

Algunos ejemplos de sistemas comerciales basados en información textual son: Google image search (<http://images.google.com/>), Lycos (http://multimedia.lycos.com/), and AltaVista photo finder (http://image.altavista.com/), es el tipo de sistema que emplearemos en nuestro trabajo.

Algunos ejemplos de software basados en el contenido son: Color and Scale representative Image Regions (CSIR) [13], SIMPLICITY [15], CIRES [16], QATRIS IMANAGER [17], ASSERT [18].

**2.2 Seguridad ciudadana**

Seguridad ciudadana es la acción integrada que desarrolla el estado con la colaboración de la ciudadanía, destinada a asegurar la convivencia pacífica, la erradicación de la violencia y la utilización pacífica de las vías y espacios públicos. Asimismo, busca contribuir a la prevención de la comisión de delitos y faltas. Uno de los objetivos estratégicos de la Policía Nacional es combatir la delincuencia.

Por otro lado, está la inseguridad ciudadana que es un fenómeno social producido por actos delictivos, actos que generan violencia y situaciones que generan miedo. Este fenómeno genera las siguientes consecuencias: el normal desenvolvimiento de la vida de los ciudadanos, amenaza el fortalecimiento de la democracia, deteriora el nivel de vida de los ciudadanos, impide la vigencia plena de los derechos humanos y las garantías constitucionales.

La inseguridad ciudadana no solamente se da en espacios y ambientes públicos, sino también podría darse en el hogar, en el barrio, en instituciones educativas y en los centros de trabajo [6,7].

El Perú tiene una alta incidencia en delitos y faltas como hurto, robo, robo con arma blanca y de fuego. Incluyen la violencia familiar y delito sexual. En la Figura 2, se observa el porcentaje de los principales problemas del Perú ocurridos en los años 2008, 2010, 2012, en donde se observa que la delincuencia es uno de los problemas graves y que va en aumento en el país en los últimos años [8].

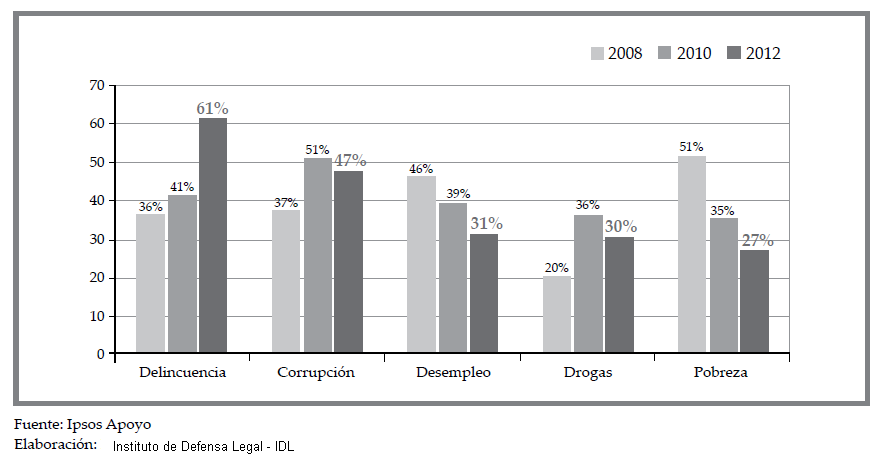


Figura 2. La delincuencia como uno de los principales problemas del País [6].

**3. MÉTODOS DE BÚSQUEDA PARA IMÁGENES [1, 6]**

En esta sección se explican los métodos de búsqueda: a) Basados en el contenido visual de la imagen, y b) basados en la descripción textual de la imagen.

**3.1 Basados en contenido**

Los Sistemas de recuperación de imágenes basados en contenido (CBIR) presentan dos módulos prioritarios: a) La representación y almacenamiento de las imágenes, y b) La recuperación de imágenes ante una consulta del usuario. Se puede observar que, a pesar del considerable número de trabajos de investigación realizados en estos sistemas, todavía no hay una solución aceptable universalmente que pueda interpretar la caracterización de la visión humana. Generalmente, los estudios son extensiones de otros trabajos realizados o surgen nuevas investigaciones.

En el primer caso, las características de la imagen también conocidas como características primitivas o de bajo nivel, es la información visual por ejemplo del color, textura o forma. Estas se presentan como de bajo nivel en contraste a las características de alto nivel que definen conceptos como montaña, cielo, células, José, etc. Se distinguen dos métodos de extracción de características: global si se realiza sobre la imagen entera y local si se aplica a un grupo pequeño de pixeles de la imagen. En la Figura 3, podemos observar cómo se divide la imagen para obtener características locales.



Figura 3. Extracción de características locales [12].

En el segundo caso, un descriptor de una imagen es una forma de representar a una imagen por sus características, con fines de almacenamiento y recuperación, en donde usualmente el descriptor es una formulación matemática. En [5] se clasifica de manera general a los descriptores en dos tipos: a) vectores y b) distribuciones. Aunque, uno de los descriptores más usados últimamente son los basados en regiones, ellos son representados como un conjunto de vectores con peso, obtenidos a partir de distribuciones discretas. Los descriptores de imágenes, también se pueden clasificar en: a) Histogramas, b) Basados en particiones, y c) Basados en regiones.

En la figura 4, se observan los histogramas para cada imagen de textura que se presenta. Las características extraídas desde el histograma de color para cada imagen son: media, varianza, coeficiente de asimetría y coeficiente de kurtosis, cuyos valores se muestran para cada imagen respectivamente.

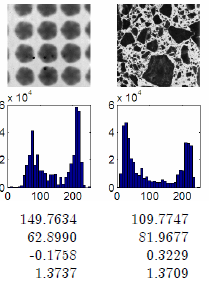


Figura 4. Ejemplos de imágenes de textura, histogramas de color y sus características [3].

En el tercer aspecto, la segmentación subdivide una imagen en sus partes constituyentes u objetos, con el fin de separar las partes de interés del resto de la imagen, por lo tanto el nivel al que se lleva a cabo esta subdivisión depende del problema a resolver [13].

Los varios métodos para extracción de descriptores visuales vienen con ventajas y limitaciones. Mientras atributos globales dan la “imagen grande”, atributos locales representan los detalles. Por tanto, dependiendo de los patrones, una representación apropiada debería ser seleccionada. En este sentido, representaciones híbridas podrían algunas veces ser más atractivas, aunque podría venir con un costo de complejidad adicional. Mientras que segmentación intenta reconocer objetos en una escena, segmentación precisa todavía es un problema abierto. Por consiguiente, investigaciones alternativas para caracterizar estructuras podrían ser más convenientes.

**3.2 Basados en información textual**

Los sistemas basados en información textual, utilizan los métodos y técnicas utilizados en los Sistemas de Recuperación de Información. Uno de los modelos mayormente utilizados en estos sistemas, es el *espacio vectorial*. Según este modelo, cada documento es representado mediante un vector de *n* términos, en donde un término es la unidad mínima de información, por ejemplo una palabra o la raíz sintáctica de una palabra. En el modelo, a cada término se le asigna un peso para medir la importancia de un término y de esta manera un término permite distinguir un documento de otro en la colección de documentos.

Siguiendo el modelo de espacio vectorial, las consultas de los usuarios también son representadas mediante un vector de términos, en donde dicho vector debe coincidir con los términos de la matriz que se forma a partir de la colección de documentos. Asimismo, se calcula el peso de los términos de la consulta.

Posteriormente, se seleccionan aquellos documentos que se aproximan más a la pregunta del usuario, mediante un cálculo denominado *similaridad*, que podría ser por ejemplo el producto del vector de la consulta del usuario con cada vector de la matriz de términos. Finalmente, se ordenan los documentos seleccionados de mayor a menor valor de similaridad. Este modelo será utilizado en la construcción del sistema, en donde los términos representan rasgos faciales de individuos, como color de ojos, color de cabello, etc. [14].

**4. DISEÑO DEL SISTEMA DE RECUPERACIÓN DE IMÁGENES [7, 11]**

En esta sección se explican los siguientes puntos: a) Los módulos del sistema en desarrollo, b) se proponen los datos que son descripciones faciales de individuos, y una propuesta de formulario de cómo se podría solicitar una imagen de un individuo, c) los algoritmos de almacenamiento y recuperación de imágenes, y d) una representación gráfica de cómo se procesan los datos.

**4.1 Módulos del sistema en desarrollo**

Se han considerado dos componentes principales del sistema:

1) El subsistema de almacenamiento de la colección de datos de cada individuo, y

2) El subsistema de recuperación de imágenes a partir de la consulta del usuario.

En la figura 5, se observan los dos subsistemas, los flujos de datos e información y los procesos principales que abarcan el sistema.



Figura 5. Módulos de almacenamiento y recuperación del Sistema [Fuente propia].

**4.2 Datos de individuos propuestos**

En esta sección presentamos los datos que pueden ser registrados en la base de datos y que corresponden a descripciones faciales de individuos.

**4.2.1 Características faciales**

**Cabello:**

**Color:** negro, castaño, rojizo, rubio

**Forma:** lacio, ondulado, crespo

**Tinte:** si, no

**Cara:**

**Forma:** larga, redonda, ovalada

**Tamaño:** grande, pequeño, mediano

**Color de piel:** blanco, negro, amarillo, mestizo

**Grupo racial:** Blanco, Del Mediterráneo o Hispánico, Africano/Caribeño, Indio/Pakistaní/Bangalí, u otras de Asia, Chino/Japonés, o personas del sudeste asiático, Árabe/Egipcio/Magreb, Origen desconocido

**Nariz:**

**Forma:** Pequeñas alteraciones: nariz estética.

Corta, larga, ancha: dorso y punta.

Nariz senil. Nariz traumática, nariz de boxeador. Nariz fisurada.

**Rinomegalia:** nariz grande en todas sus dimensiones.

**Nariz racial:** negroide, mestiza, mestiza andina, asiática.

**Ojos:**

**Color:** negro, gris, marron, castaño, verde, azul

**Tamaño:** grandes, pequeños, medianos

**Tipo:** saltones, hundidos

**Labios:** Delgados, gruesos

**Cejas:** Gruesas, finas, arqueadas, rectas

**Orejas:** Grandes, pequeñas, redondas, punteagudas, pegadas a la cara

**Frente:** ancha, estrecha, ovalada, cuadrada, trapezoidal

**Cabeza:** grande, pequeña, mediana

**Pómulos:** altos, bajos, centrados, pequeños

**Otros:** Pecas, manchas.

**4.2.2 Formulario de características faciales**

En la figura 6, se muestra un formulario de características faciales para un individuo sospechoso, que se ingresan al sistema como pregunta del usuario.



Figura 6. Ejemplo de un formulario de características faciales.

**4.3 Algoritmos de almacenamiento y recuperación de imágenes**

En esta sección se describen los algoritmos de los módulos del sistema propuesto, los que se basan en el modelo de espacio vectorial.

**4.3.1 Algoritmo de almacenamiento**

* Se registra en una base de datos para cada individuo, su identificación y una colección de características físicas por ejemplo: cabello color negro, cara forma larga, etc. En donde cada característica en el sistema es un término o unidad mínima de información.
* Para medir la importancia de un término de un individuo o en una colección de individuos, se asignan pesos a cada uno de los términos. Por ejemplo, si es un malhechor reincidente se le puede asignar mayor peso.
* Para cada individuo se registra también su imagen.

**4.3.2 Algoritmo de recuperación de imágenes**

* Ante una consulta de un individuo, es decir un conjunto de rasgos físicos, el sistema realiza dos procesos: a) Búsqueda del individuo en la Base de datos, b) El sistema devuelve imágenes parecidas de individuos, presentadas por orden de mayor probabilidad.
* En el método de espacio vectorial para la búsqueda del individuo se realiza mediante el método producto de espacio vectorial. El resultado es un conjunto de rasgos de individuos.
* El producto de espacio vectorial se da calculando la distancia que existen entre los vectores de los individuos y la solicitud del usuario. La fórmula corresponde al coseno del ángulo entre los dos vectores:

di xdj

cos(di \*dj) = -------------

|di |x|dj |

* 1. **Representación de los datos**

En este punto se presenta gráficamente como se desarrollan los datos dentro del sistema.

En la figura 7, se observa la estructura de datos del módulo de almacenamiento, una matriz de M x N, en donde las filas de la matriz representan a los individuos, y en las columnas se encuentran todos los rasgos faciales que podrían tener los individuos. En las intersecciones de la matriz se encuentran los pesos de las características de cada individuo.

Para obtener los individuos sospechosos en función de la pregunta mediante características de un individuo, se utiliza como se mencionó anteriormente el producto de espacio vectorial de términos, obteniendo como resultado un conjunto de individuos con rasgos similares a la pregunta inicial, ordenados de mayor a menor probabilidad, como se observa en las Figuras 7 y 8.



Figura 7. Representación de los datos en los módulos de almacenamiento y recuperación del Sistema [Fuente propia]



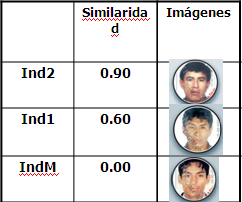


Figura 8. Individuos malhechores en la base de datos, individuos sospechosos ordenados de mayor a menor probabilidad.

**5. CONCLUSIONES Y TRABAJOS FUTUROS**

* Se ha realizado el análisis y diseño de una herramienta de software, que es un Sistema de recuperación de imágenes de individuos para Seguridad Ciudadana, concretamente en las ciudades del Perú en donde a pesar de los esfuerzos de las autoridades, presenta un alto índice delincuencial. El sistema consta de dos módulos: 1) Almacenamiento o registro de las imágenes faciales y sus descripciones; y 2) Búsqueda de una imagen en particular. Se utiliza el algoritmo Modelo de Espacio Vectorial.
* Se está elaborando la codificación y pruebas del Sistema.
* Como trabajos futuros, se pueden considerar los sistemas basados en su contenido, con el mismo objetivo para ayudar en la seguridad ciudadana. O una combinación de los dos sistemas basados en el texto y en el contenido de la imagen para ganar eficiencia del sistema.
* Asimismo, el sistema puede ser ampliado para considerar la implementación en dispositivos móviles.

**Agradecimientos**

El presente trabajo se desarrolla en el marco del proyecto de investigación 2012 “Elaboración de un sistema de recuperación de imágenes de individuos sospechosos para seguridad ciudadana”, financiado parcialmente por el Vicerrectorado de Investigación de la Universidad Nacional Mayor de San Marcos.

**6. REFERENCIAS BIBLIOGRÁFICAS**

[1] Castleman, K.R. "Digital Image Processing", Prentice-Hall, Englewood Cliffs, New Jersey 07632, 1996.

[2] Gonzalez, R.C., Woods, R.E. Tratamiento Digital de Imágenes, Addison-Wesley Publishing Co, Reading, Washington, 1996.

[3] Mazo, Q. M. &Boquete, V.L. &Barea N.R. Visión Artificial. Publicaciones UAH. ISBN 84 8138 101 2, DL. M 1321 1996.

[4] La Serna N., Román U. Implementación de un sistema de recuperación de información. Revista de Investigación de Sistemas e Informática (RISI) UNMSM. Vol 6 No 1 2009.

[5] Sonka M., Hlavac V., y Boyle R. Image Processing, Analysis, and Machine Vision. Thomson 2008.

[6] Ley del Sistema Nacional de Seguridad Ciudadana, 27933 del 11 de Febrero del 2003.

[7] Guillermo Bonilla Arévalo. La seguridad ciudadana. Desafío actual. Instituto de defensa legal. 2008.

[8] Intituto de DefensaLegal IDL, Informe Annual 2012 sobre Seguridad Ciudadana.

[9] Carpineto C., Romano G.. A Survey of Automatic Query Expansion in Information Retrieval. ACM Computing Surveys, Vol. 44, No. 1, Article 1, Publication date: January 2012.

[10] KHERFI M. and BERNARDI A. Image Retrieval From the World Wide Web:

Issues, Techniques, and Systems. ACM Computing Surveys, Vol. 36, No. 1, March 2004, pp. 35–67.

[11] Ritendra Datta, et.al. Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Computing Surveys, Vol. 40, No. 2, 2008.

[12] Chen, et al. Machine Learning and Statistical Modeling Approaches to Image Retrieval. Kluwer Academic Publishers, 2004.

[13] Jurandy Almeida et.al. Making colors worth more than a thousand words. SAC’08, Fortaleza Brazil, 2008.

[14] Figuerola C., Alonso J., y Zazo A. Diseño de un motor de recuperación de la información para uso experimental y educativo. BID Num.4 junio 2000.

[15] James Z. Wang, Member, IEEE, Jia Li, Member, IEEE. Simplicity. 2001.

[16] CIRES. Department of Electrical and Computer Engineering The University of Texas at Austin. 2007. [amazon.ece.utexas.edu/~qasim/research.htm](http://amazon.ece.utexas.edu/~qasim/research.htm).

[17] SICUBO S.L. en estrecha colaboración con el grupo de investigación de ingeniería de medios de la Universidad de Extremadura. 2006. [www.sicubo.com/index/es/Soluciones/Qatris\_Imanager](http://www.sicubo.com/index/es/Soluciones/Qatris_Imanager)

[18] ASSERT. The School of Electrical and Computer Engineering at Purdue University,West Lafayette. rvl2.ecn.purdue.edu/˜cbirdev/WEB\_ASSERT/assert.html.

**UN ALGORITMO GENÉTICO PARA LA DETECCIÓN DE FRAUDE ELECTRÓNICO EN TARJETAS DE DÉBITO EN EL PERÚ**

A Genetic Algorithm For Fraud Detection Electronic Debit Cards In Peru

Luis Enrique Lavado Napaico

Universidad Nacional Mayor de San Marcos

Facultad de Ingeniería de Sistemas e Informática

[llavado@msn.com](mailto:llavado@msn.com)

|  |
| --- |
| RESUMEN  La gran problemática de la banca peruana es detectar las transacciones fraudulentas que se encuentran dispersas con las transacciones genuinas. Las soluciones propuestas no son suficientes para detectar estas operaciones ilícitas con precisión porque están orientadas a mercados diferentes al peruano. Consecuentemente, se han revisado las principales aportaciones dentro de este ámbito, tales como teoría de Dempster-Shafer y aprendizaje bayesiano, sistema evolutivo Fuzzy Darwinian evolutionary system, etc. se presenta el modelado y diseño de un algoritmo genético para obtener las reglas más representativas de compra de los tarjetahabientes dentro del universo de datos transaccionales recopilados de un banco peruano. De las pruebas experimentales se obtuvo una precisión del 95.5% en el canal internet y 95.8% para el canal punto de venta. Finalmente, Se concluyó que el empleo de la estrategia de algoritmo evolutivo obtuvo una aceptable exactitud en la predicción.  **Palabras clave:** Comercio electrónico, algoritmos genéticos, fraude electrónico.  ABSTRACT  The great problem of the bank is to detect fraudulent transactions which are scattered with genuine transactions. The proposed solutions are not sufficient to detect these illegal operations precisely because they are aimed at different Peruvian markets. Consequently, we reviewed the main contributions in this area, such as Dempster–Shafer theory and Bayesian learning, Fuzzy Darwinian, etc. presents the modeling and design of a genetic algorithm to obtain more representative rules cardholder purchase within the universe of transactional data collected from a Peruvian bank. From experimental evidence obtained an accuracy of 95.5% in the internet channel 95.8% for the point of sale. Finally, it was concluded that the use of evolutionary algorithms strategy got an acceptable accuracy in prediction.  **Palabras clave:** Electronic commerce, genetic algorithms, fraud electronic. |

1. INTRODUCCION

El impacto que las tecnologías de información tienen en el mundo actual es impresionante. Cada vez se crean nuevos paradigmas y la forma de realizar operaciones comerciales está cambiando. Las entidades comerciales y financieras en el mundo están apostando fuerte por la sustitución del papel moneda por el dinero electrónico, sin embargo, para expandir el comercio electrónico se requiere de un sistema de pago que se ajuste a las necesidades de compra de los usuarios, transmitiendo la seguridad y confianza en sus operaciones. El estudio presentado por Richardson Robert [1] director de Computer Security Institute (CSI) indica que entre los principales riesgos de seguridad, los ataques por fraude financiero y las vulnerabilidades relacionadas con aplicaciones Web de comercio electrónico están aumentando y que postulan como nuevas tendencias de riesgo. El informe realizado por CyberSource Corporation [2] revela que en el año 2011 los comercios informaron que perdieron un promedio de 1% de los ingresos totales por fraudes cometidos en transacciones de comercio electrónico. Los comercios informaron una reducción del 33% de los porcentajes de órdenes perdidas por fraude. Las operaciones fraudulentas o ilegales suelen seguir patrones característicos que permiten, con cierto grado de probabilidad, distinguirlas de las legítimas y desarrollar así mecanismos para tomar medidas rápidas frente a ellas.

1.1. Problemática

La seguridad se ha convertido en el principal problema e inquietud del comercio electrónico. Existe el temor de los consumidores y negocios debido a las grandes olas de estafas que el mundo está experimentando. El informe realizado por Frost & Sullivan [3] revela que el 22% de los usuarios en Latinoamérica han dejado de usar la banca en línea, y el 10% han cambiado de banco debido a incidentes de fraude. Un impresionante 95% de los usuarios de transacciones en línea creen que su banco debe implementar mayores y mejores soluciones de seguridad, a fin de minimizar los riesgos por fraude. La realidad del Banco de la Nación del Perú indica que los servicios de banca por internet y operaciones VISA ha sufrido una baja en los últimos años debido a la desconfianza y el temor del tarjetahabiente por el uso de estos servicios no tradicionales.

1.2. Propuesta

Se propone la utilización de un modelo heurístico basado en el comportamiento transaccional de los clientes y la determinación de los patrones de desviación que sean catalogadas como sospechosas, para ello se empleará técnicas basadas en algortimos genéticos. En la primera fase se recopilará datos históricos transaccionales de los canales: punto de venta (POS) y operaciones en línea internet de un periodo representativo (julio y agosto del 2011); posteriormente, se diseñará un algoritmo genético cuyo objetivo será maximizar la precisión de la predicción frente a los datos reales. Los objetivos alternos serán: la minimización de los falsos positivos que son las transacciones que el sistema deja pasar aún siendo fraudulentas y la minimización los falsos negativos que son aquellas que el sistema devuelve una transacción como sospechosa siendo una operación válida.

2. ESTADO DEL ARTE

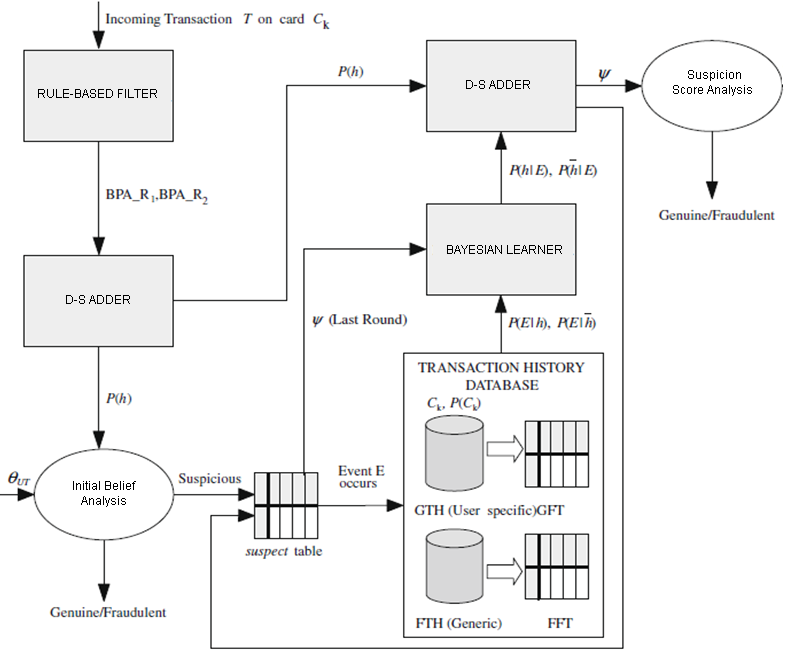
La detección del fraude consiste en analizar el comportamiento de los usuarios cuando realiza una operación financiera para estimar, detectar o evitar las transacciones indeseables. Para una lucha eficaz contra el fraude de transacciones de tarjeta de débito es necesario comprender las estrategias empleadas en investigaciones científicas que derivaron en reglas con alto valor predictivo.

2.1. Paradigmas de detección de fraude

*2.1.1. Fusión teoría Dempster–Shafer y Bayesian learning*

Panigrahi S. et al. [4], proponen un modelo para la detección de fraudes de tarjetas crédito, que combina las evidencias actuales así como el comportamiento pasado almacenado en el historial de transacciones. La propuesta se basa en la integración de los enfoques: rule-based filtering, teoría Dempster–Shafer y Bayesian learning. En el modelo propuesto, un número de reglas se utilizan para analizar la desviación de cada transacción entrante de un perfil normal de tarjetahabiente. Una transacción entrante es primero manejada por el componente Rule-based. La creencia de los valores son combinados para tener una creencia inicial para la transacción mediante la aplicación de la teoría Dempster–Shafer. La creencia general es, además, reforzada o debilitada conforme a similares transacciones históricas de fraude u operaciones genuinas usando Bayesian learning. El modelo se representa en la figura 1.

La simulación del modelo obtuvo 98% verdadero positivo (TP) y menos que el 10% de Falso Positivo (FP). La utilización de la teoría de Dempster-Shafer ofrece una buena precisión, sobre todo en términos de verdaderos positivos, el aprendizaje bayesiano ayuda a mejorar la exactitud del sistema.

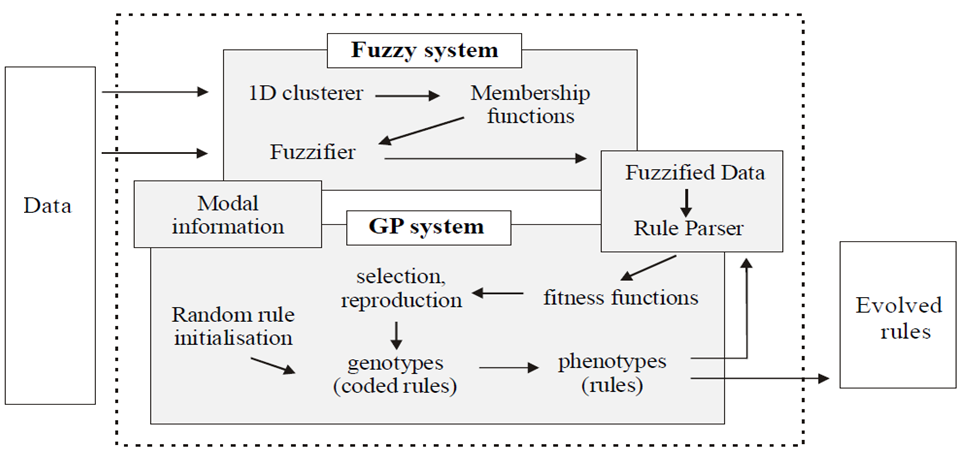


**Figura 1.** Propuesta de detección de fraude Dempster–Shafer theory and Bayesian learning (Panigrahi et al., 2009)

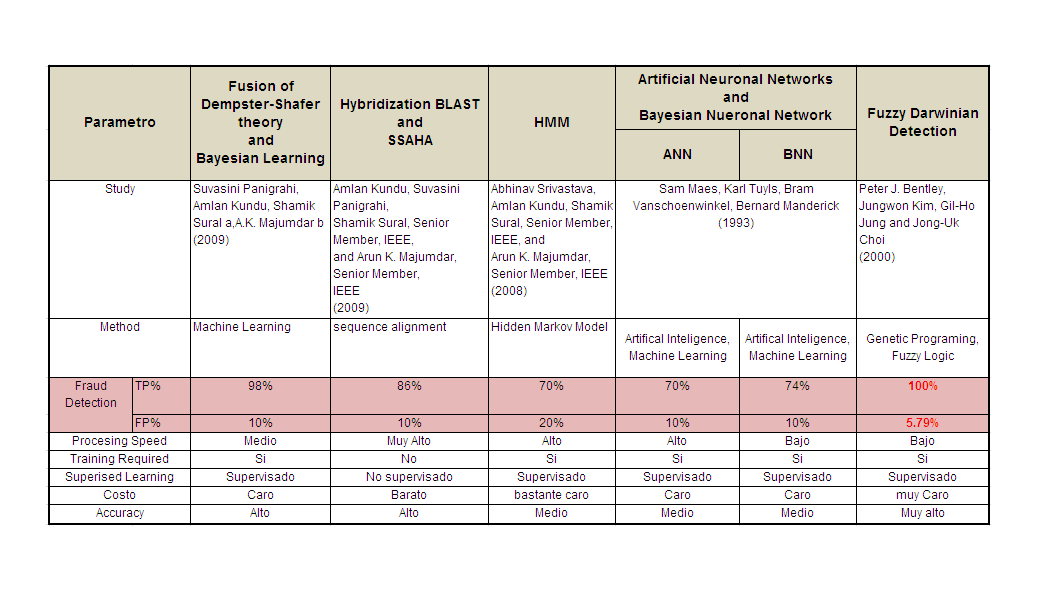
*2.1.2. Fuzzy Darwinian*

Bentley Peter J. et al. [5], propone un modelo de detección Fuzzy Darwinian capaz de clasificar las transacciones de tarjetas de crédito en clases sospechosos y no sospechoso. El sistema se compone de un algoritmo de búsqueda en programación genética (GP) y un sistema experto difuso. El sistema híbrido genético difuso es un sistema adaptativo donde la estructura se puede construir y los parámetros se pueden ajustar por un algoritmo de aprendizaje. La Figura 2 provee una visión de este modelo.

Los resultados experimentales muestran que este método es capaz de alcanzar una buena precisión y los niveles de inteligibilidad de los datos reales. Se tiene una precisión muy alta y produce una baja falsa alarma, detectando el 100% de las sospechas verdadero positivo (TP) y un 5.79% de Falso Positivo (FP).

****

**Figura 2.** Arquitectura Evolutionary-Fuzzy System (Peter J. Bentley, 2000).



2.2. Comparación de métodos de detección de fraude

En las publicaciones Benson y Portia [6], se presenta un análisis comparativo de métodos detección de fraude, ver Tabla 1. Para la comparación de diferentes métodos se emplea la matriz de confusión. Los autores comparan las siguientes técnicas: Dempster-Shafer theory and Bayesian learning [4], BLAST-SSAHA Hybridization [7], Hidden Markov Model [8], Fuzzy Darwinian [5] y Bayesian and Neural Networks [9]. Los parámetros usados para la comparación de las técnicas de detección de fraude son la precisión, ratios en término de verdadero positivo (TP-denota el número de casos positivos que han sido predichos como tal) y falso positivo (FP-denota el número de casos negativos que han sido predichos como positivos). Los resultados muestran que la técnica de detección de fraude Fuzzy tienen alta ocurrencia, el ratio TP llegó al 100% en comparación de los otros y para el caso del ratio FP fue de 5.79% el menor de todos.

**Tabla 1.** Comparación de métodos de detección de fraude (Benson y Portia, 2011).

3. METODOLOGIA DE LA INVESTIGACION

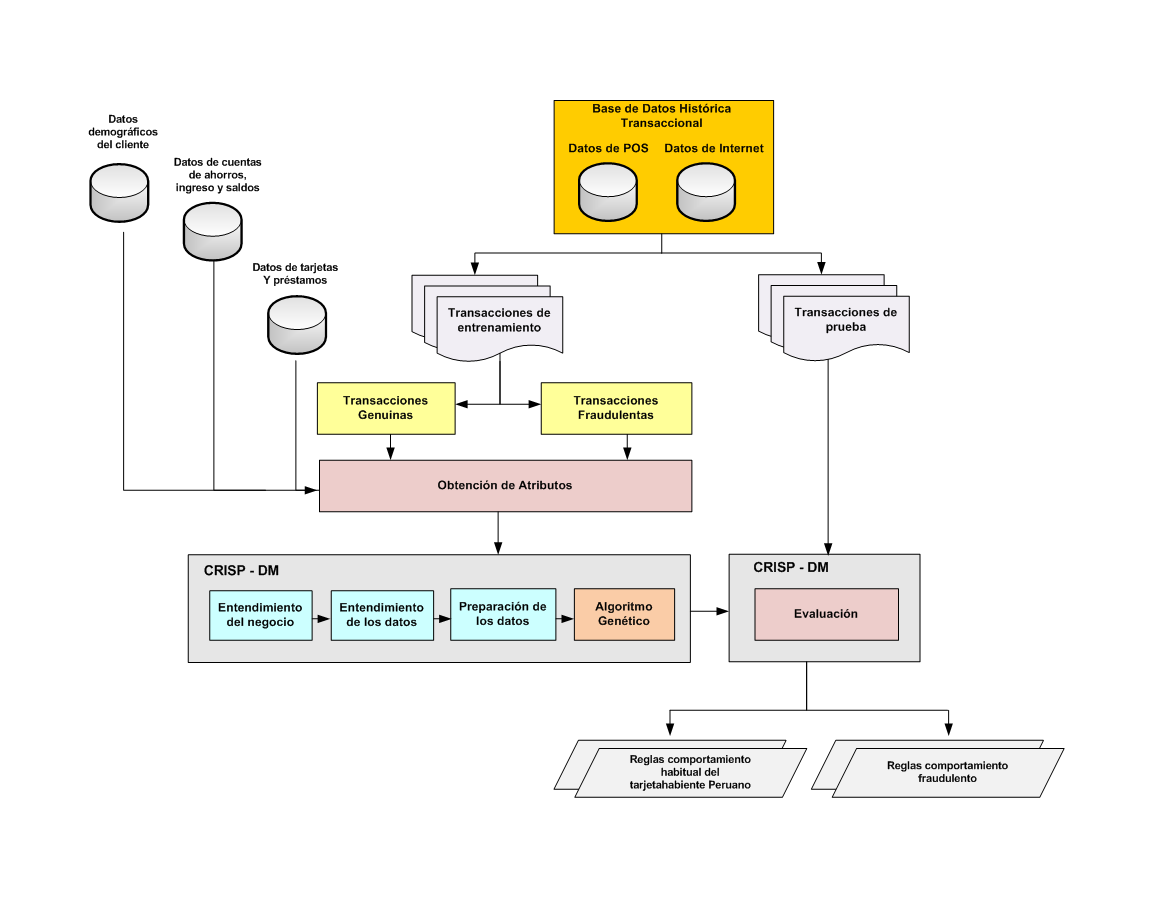
3.1. Modelo Conceptual de la propuesta

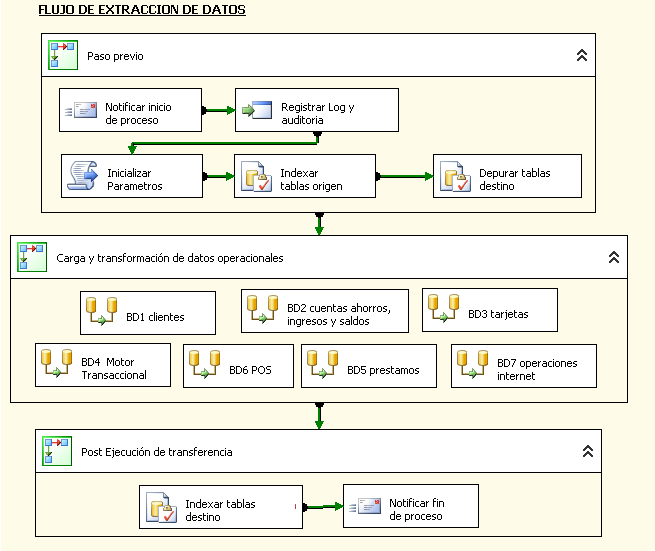
La propuesta de extracción de reglas de comportamiento esbozado en la figura 3, consiste en recopilar los datos históricos transaccionales de los canales punto de venta (POS) y operaciones en internet. De los datos de muestreo el 70% servirá para el proceso de entrenamiento del algoritmo genético y la diferencia para el proceso de evaluación. Los datos de transacciones genuinas y fraudulentas se conjugan con las bases de datos demográficos del cliente, datos de las cuentas de ahorros, ingresos y saldos del cliente y la base de datos de tarjetas y préstamos con el fin de obtener atributos relevantes para el análisis. Se aplica la metodología CRISP-DM como herramienta que guía el proceso de descubrimiento de conocimiento y permite la identificación de variables. En la fase de modelado del CRISP-DM se emplea una técnica basada en algoritmos genéticos, la resultante del proceso de entrenamiento del algoritmo genera reglas interesantes que identifiquen a las transacciones genuinas y fraudulentas. En la fase de evaluación se verifica la calidad de las reglas obtenidas del proceso de entrenamiento. La salida del esquema propuesto establece dos grupos de reglas de decisión, (1) las que son consideradas como comportamiento habitual y (2) las reglas que fueron catalogadas como fraudulentas.

**Figura 3.** Esquema del Modelo Conceptual de la propuesta.

3.2. Fases para la obtención de variables

*3.1.1. Entendimiento de los datos*

En la Figura 4, se representa el flujo del proceso de extracción y transformación de los datos. En la etapa inicial del paquete el primer grupo de tareas se encargan del registro de eventos, inicialización de variables, re-indexación de tablas, y a la depuración de los datos en las tablas de destino. En la segunda etapa del paquete se procede a leer los datos de cada sistema origen y transferirlos a los nuevos esquemas; en este proceso se adicionó tareas de conversión y limpieza que permita verificar si los datos cumplen con la estructura solicitada y la calidad esperada. En la etapa final del paquete se procede a indexar las tablas de destino para un mejor rendimiento.



**Figura 4.** Flujo de extracción de los datos iniciales.

*3.1.2. Preparación de los datos*

El empleo y representación de las variables dependen de los objetivos definidos en la fase “entendimiento del negocio” y de estudios previos de investigadores cuyo modelos han demostrado una propuesta de solución al fraude electrónico. En la tabla 2 se lista las variables utilizadas en el algortimo genético.

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Alias** | **Descripción** | **Canal** |
| **Variables de entrada** | | | |
| 1 | HCHP | Horario de compra habitual del cliente dentro de las 24 horas | POS |
| 2 | HCHI | Horario de compra habitual del cliente dentro de las 24 horas | WWW |
| 3 | CSCP | Situación civil del cliente asociada a la compra | POS |
| 4 | CSCI | Situación civil del cliente asociada a la compra | WWW |
| 5 | CCEP | Categoría de establecimiento más concurrida por el cliente | POS |
| 6 | CCEI | Categoría de establecimiento más concurrida por el cliente | WWW |
| 7 | CREP | Rango de edad del cliente asociado al hábito de compra | POS |
| 8 | CREI | Rango de edad del cliente asociado al hábito de compra | WWW |
| 9 | TETP | Tiempo transcurrido (días) desde su enrolamiento de la tarjeta y su primera compra en el periodo de muestra | POS |
| 10 | TETI | Tiempo transcurrido (días) desde su enrolamiento en internet y su primera compra en el periodo de muestra (enrolamiento en la compra) | WWW |
| 11 | TUCP | Tiempo transcurrido (horas) desde la última compra a la penúltima compra con la misma tarjeta | POS |
| 12 | TUCI | Tiempo transcurrido (horas) desde la última compra a la penúltima compra con la misma tarjeta | WWW |
| 13 | MMXP | Monto máximo de compra | POS |
| 14 | MMXI | Monto máximo de compra | WWW |
| 15 | NMXP | Número máximo de transacciones efectivas realizadas con una misma tarjeta en el día | POS |
| 16 | NMXI | Número máximo de transacciones efectivas realizadas con una misma tarjeta en el día | WWW |
| 17 | NEVP | Número de establecimientos visitados en un día | POS |
| 18 | NEVI | Número de establecimientos visitados en un día | WWW |
| 19 | NMRP | Número máximo de rechazos detectados antes de ser efectiva la transacción de compra en el mismo establecimiento | POS |
| 20 | NMEI | Número máximo de intentos de enrolamiento antes de la primera compra | WWW |
| **Variable de salida** | | | |
| 1 | OUPT | Situación de la transacción | POS / WWW |

**Tabla 2.** Variables utilizadas en el algortimo genético.

4. INGENIERIA DEL ARTEFACTO

4.1. Planteamiento del algoritmo genético propuesto

El algoritmo genético planteado sigue el enfoque Iterative Rule Learning (IRL). Según esta estrategia cada individuo de la población representa una única regla, sólo el mejor individuo es considerado como parte de la solución descartándose al resto de individuos de la población. Por lo tanto, cada ejecución de los algoritmos genéticos proporciona una solución parcial al problema de aprendizaje, ya que en cada iteración aportará una nueva regla a la solución problema [10].

Para la aplicación del paradigma de un algoritmo genético presenta dos factores críticos: la codificación de los individuos y la evaluación de éstos. Ambos factores, entre otros, influyen en la efectividad del algoritmo y su convergencia, siendo por tanto los aspectos donde se centra gran parte del esfuerzo.

*4.1. Codificación del Cromosoma*

La forma de codificación propuesta sigue el enfoque IRL, en el que cada cromosoma representa una regla “Cromosoma = Regla”, como SLAVE [11], MOGUL [12] y la propuesta de Carvalho y Freitas [13]. Cada individuo de la población representa una única regla de decisión que describe la relación entre los valores de los atributos y las etiquetas de clase.

Donde:

son atributos y valores asignados para cada atributo, y etiqueta discreta de clasificación.

Las reglas se representan normalmente mediante cadenas binarias de bit ***(0’s y 1’s)*** de una longitud determinada ***L*** que vendrá impuesta por el número de variables existentes en la solución y por el número de bits necesarios para codificarla, el espacio de búsqueda tendrá un tamaño de . El antecedente o descripción de la regla (parte izquierda) es una conjunción entre atributos y varias disyunciones sobre los valores de los atributos. El esquema de codificación binario, por cada atributo se almacena un bit para cada uno de los valores categóricos que puede tomar, de forma que si el bit correspondiente tiene el valor de 0 indica que no pertenece a la condición y si tiene el valor 1 que sí pertenece. Si en un individuo todos los bits correspondientes a un atributo tienen valor 1, indica que dicha variable no es relevante para la información aportada en la regla (cualquier valor de la variable verifica la condición de la regla), por lo que esta variable se ignora. El código 0 es omitido en el conjunto de valores codificados al carecer éste de sentido. En la tabla 3 se ejemplifica el diseño del antecedente del canal POS conformado por 10 variables. Para el caso del atributo ***HCHP*** los valores categóricos que participan se encuentran en la posición 2 y 4 que representa el horario de compra realizada entre 12:00 y 15:00 horas y el rango entre las 20:00 y 23:00 horas, existen 16 posibles combinaciones de ésta variable.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **VAR** | **Etiquetas Lingüísticas** | | | | | | | | | **Comentario** | **Comb** |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **HCHP** | **0** | **1** | **0** | **1** |  |  |  |  |  | Compras realizadas entre las horas: \* 12:00 <= t <= 15:00  \* 20:00 <= t <= 23:00 | **24 = 16** |
| **CSCP** | **1** | **1** | **0** | **0** | **0** |  |  |  |  | Situación de los clientes que realizaron una compra: \* Soltero \* Casado | **25 = 32** |
| **CCEP** | **0** | **0** | **0** | **0** | **1** | **0** | **0** | **0** | **0** | Categoría del establecimiento donde se realizó la compra: \* Restaurantes | **29 = 512** |
| **CREP** | **0** | **1** | **1** | **1** | **0** |  |  |  |  | Rango de edad de las personas con hábitos de compra: \* 33 <= e <= 40 \* 41 <= e <= 49 \* 50 <= e <= 60 | **25 = 32** |
| **TETP** | **1** | **0** | **0** | **1** | **0** |  |  |  |  | Tiempo transcurrido en días desde su enrolamiento y su primera compra: \* t <= 690 \* 1108 <= t <= 1444 | **25 = 32** |
| **TUCP** | **1** | **1** | **1** | **0** | **0** |  |  |  |  | Tiempo transcurrido en horas desde la última compra a la penúltima: \* t <= 1 \* 2 <= t <= 10 \* 11 <= t <= 37 | **25 = 32** |
| **MMXP** | **0** | **0** | **0** | **1** | **1** |  |  |  |  | Monto máximo de compra: \* 201 <= m <= 400 \* 401 <= m | **25 = 32** |
| **NMXP** | **1** | **1** | **0** |  |  |  |  |  |  | Número máximo de transacciones efectivas: \* n = 1 \* n = 2 | **23 = 8** |
| **NEVP** | **1** | **0** | **0** |  |  |  |  |  |  | Número de establecimientos visitados en un día: \* n = 1 | **23 = 8** |
| **NMRP** | **1** | **1** | **0** |  |  |  |  |  |  | Número máximo de rechazos detectados: \* n = 1 \* n = 2 | **23 = 8** |

**Tabla 3.** Diseño del antecedente para el canal POS.

Respecto a la clase, es importante señalar que la operación se realiza sobre conjuntos de datos donde cada ejemplo es etiquetado con una única etiqueta de clase discreta ***Clase=E***, donde los valores posibles son: 0 de tratarse de una transacción genuina y 1 si es una transacción fraudulenta. Tanto para el canal POS e internet la clase es representada por el valor discreto de la variable de salida ***OUPT***.

Se ha seleccionado la Codificación Natural [14] como la forma de representación del individuo que reducirá el cardinal del conjunto de posibles soluciones, sin que ello produzca pérdida en la precisión en las mismas. Partiendo de la codificación binaria que asigna un bit a cada posible valor, denotando con 1 y 0 la presencia o ausencia del valor en la condición, respectivamente, la codificación natural transforma esa cadena binaria en su representación decimal. Así, un gen discreto es un número natural que identifica un conjunto de valores discretos y pertenece al intervalo , donde ***|A|*** es el cardinal del conjunto de valores posibles del atributo. En la tabla 4 muestra un ejemplo de codificación natural de la variable ***HCHP***, conformada por 4 etiquetas lingüísticas. La codificación natural asociada a las compras realizadas entre 12:00 y 15:00 horas y el rango entre las 20:00 y 23:00 horas tiene el valor de 5, que equivale a su representación binaria {0 1 0 1}.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Etiquetas Lingüísticas de la Variable HCHP** | | | | **Codificación Natural** |
| ***HCHP1*** | ***HCHP2*** | ***HCHP3*** | ***HCHP4*** |
| 0 | 0 | 0 | 1 | 1 |
| 0 | 0 | 1 | 0 | 2 |
| … | … | … | … | … |
|  |  |  |  |  |
| **0** | **1** | **0** | **1** | **5** |
| 0 | 1 | 1 | 0 | 6 |
| … | … | … | … | … |
| 1 | 1 | 1 | 1 | 15 |

**Tabla 4.** Diseño del antecedente para el canal POS.

*44.2. Operador genético*

*4.2.1. Mutación del antecedente*

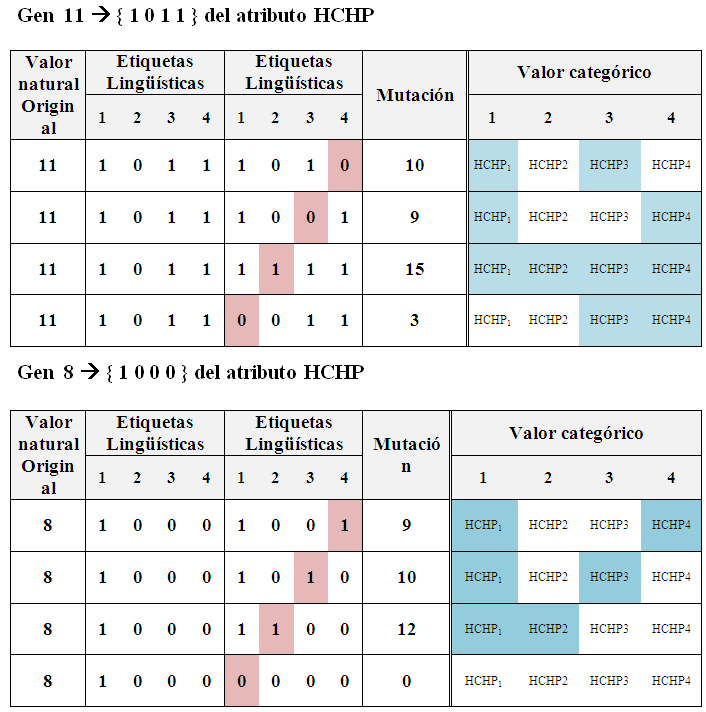
Partimos de un número natural cuyos bits en representación binaria denotan la presencia o ausencia de un valor en una condición. La mutación consiste en cambiar el valor del gen por otro símbolo del alfabeto que represente un conjunto de valores distinto al inicial donde simplemente se ha agregado o suprimido un valor. Esta mutación aplicada en la codificación binaria tiene una implementación consistente en seleccionar un bit al azar y cambiar su valor de 0 a 1 o al revés, según el caso. La tabla 5 ilustra la mutación del primer gen del atributo ***HCHP*** el valor original del gen ***HCHP1*** es reemplazado su valor de O a 1, generando el nuevo valor natural de 13.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Atributo HCHP original.** | | | | | | | | | |
| **VAR** | **Valor categórico** | | | | **Etiquetas Lingüística** | | | | **Cod.**  **Nat.** |
| **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** |
| **HCHP** | HCHP1 | HCHP2 | HCHP3 | HCHP4 | **0** | **1** | **0** | **1** | **5** |
| **Atributo HCHP mutado en el primer gen.** | | | | | | | | | |
| **Atributo** | **Valor categórico** | | | | **Etiquetas Lingüística** | | | | **Cod.**  **Nat.** |
| **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** |
| **HCHP** | **HCHP1** | HCHP2 | HCHP3 | HCHP4 | **1** | **1** | **0** | **1** | **13** |

**Tabla 5.** Esquema de mutación del primer gen del atributo HCHP.

*4.2.2. Cruce del antecedente*

El cruce se basa en la mutación natural definida anteriormente. Cada gen que interviene en el cruce proporciona un conjunto de candidatos. Estos candidatos son el resultado de unir el conjunto de posibles mutaciones del gen con el propio gen. Así, la descendencia de dos genes se calcula como la intersección de los conjuntos de candidatos que cada gen aporta. Cuando los padres no ofrecen candidatos comunes, se calculan nuevos candidatos para cada padre mutando los conjuntos iniciales hasta que la intersección no esté vacía [14]. Se detalla un ejemplo de los operadores naturales para atributos discretos basados en ***HCHP***. En la tabla 6 el gen codificado con el número natural ***11*** tiene como código binario el ***1011***. El bloque a la derecha da las posibles mutaciones que este gen puede sufrir. Para el gen codificado con el valor natural ***8 (1000)*** es similar. Así, los conjuntos de posibles mutaciones de ***11*** y ***8*** son ***{10; 9; 15; 3}*** y ***{9; 10; 12; 0}***, respectivamente.



**Tabla 6.** Tabla de comprobación de mutaciones del gen discreto HCHP.

El cruce entre ambos genera el conjunto de genes ***{9; 10}***, pues es la intersección de los dos conjuntos anteriores.

Mut(11) = {10,9,15,3}

Mut(8) = {9,10,12,0}

Cruce(11,8) = { Mut(11) Mut(8) }

Cruce(11,8) = { {10,9,15,3} {9,10,12,0}}

Cruce(11,8) = {9,10}

*4.3. Descripción del algoritmo genético*

El algoritmo propuesto se divide en un procedimiento principal ***MainDetectFraud***, el cual construye el conjunto de reglas; y la función complementaria ***AlgoritmoEvolutivo***, que implementa el algoritmo genético propiamente dicho. Inicialmente, el conjunto de reglas ***R*** está vacío y en cada iteración se añade la regla que devuelve ***AlgoritmoEvolutivo***. El parámetro ***DE*** es el conjunto de datos de entrenamiento, el cual es representada con base a la propuesta de Codificación Natural [14], con el fin de reducir el espacio de búsqueda y acelerar la convergencia del algoritmo. La transformación da como resultado el conjunto de ejemplos codificados ***xDE*** que será usado durante toda la ejecución. La variable ***n*** almacena el número inicial de ejemplos de ***xDE***, ya que éste será reducido en cada iteración. Dicha reducción se produce eliminando aquellos ejemplos de ***xDE*** que son cubiertos por la regla ***r*** obtenida de ***AlgoritmoEvolutivo***. La regla obtenida r se adiciona al conjunto de reglas solución ***R***. El parámetro ***Poda*** (factor de poda de datos de entrenamiento) controla el número de ejemplos ***xDE*** que aún no han sido cubiertos durante el proceso. El proceso iterativo finaliza cuando el número de ejemplos que restan el conjunto de entrenamiento ***xDE*** no supera el factor establecido por ***Poda*** sobre el número inicial de ejemplos ***n***. El módulo ***AlgoritmoEvolutivo*** inicializa la población ***P*** y ejecuta la función evolucionar que devuelve la regla con la mejor fitness obtenida en el proceso evolutivo. Finalmente, se ejecuta la función ***reducirEscenario*** que elimina los datos de entrada que clasificaron con la regla seleccionada. El módulo ***Evolucionar*** lleva a cabo la evolución siempre y cuando el número de generaciones es menor igual al parámetro ***num\_generaciones*** ingresada por el usuario. En cada iteración, el procedimiento ***Evaluacion*** asigna un valor de bondad a cada individuo de la población actual ***P***, la función de bondad está conformada por dos factores: la complejidad y completitud de la regla [15].

El procedimiento ***Reemplazo*** genera la nueva población mediante los operadores genéticos: selección, cruce y mutación de individuos. Cuando el número de generaciones preestablecido ***num\_generaciones*** es alcanzado, la población final es evaluada para seleccionar el mejor individuo de ésta. Para relajar el modelo, en la fase de selección de los individuos para el cruce se empleará el método de la ruleta para aceptar soluciones malas y para salir del óptimo local. El individuo seleccionado en el procedimiento ***Evaluacion*** es devuelto al módulo principal para eliminar los ejemplos cubiertos ***xDE***, se incluirá en el conjunto de reglas ***R*** y se continuará el proceso.

5. EXPERIMIENTOS Y RESULTADOS

En las siguientes secciones se detalla los resultados de cada experimento realizados en forma independiente para el canal POS e internet.

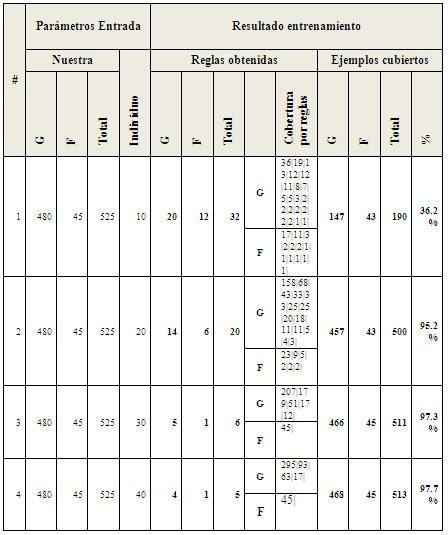
*5.1. Canal POS*

*5.1.1. Parámetros iniciales de entrenamiento*

El proceso de entrenamiento consiste en ejecutar el algoritmo genético para 4 escenarios, en los cuales se modificó el parámetro número de individuos de la población con el fin de obtener el conjunto de reglas con la mejor bondad. Para una muestra de 520 registros (480 transacciones genuinas y 45 transacciones fraudulentas), se inicializa con 10 individuos hasta llegar a 50 individuos. El factor de mutación se mantiene uniforme en 2% de la cantidad de individuos. El parámetro de completitud es mayor a la complejidad porque la búsqueda está orientada a la obtención de reglas con una alta capacidad predictiva (número de ejemplos cubiertos por la regla).

*5.1.2. Resultados del entrenamiento*

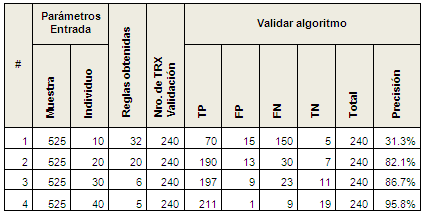
En la tabla 7 se detalla los escenarios probados y la cantidad de reglas genuinas y fraudulentas obtenidas. En el escenario 1 se ejecutó el algoritmo genético con una población de 10 individuos obteniéndose 32 reglas, sucesivamente se aumentó el número de individuos. En el escenario 4 se obtuvo 5 reglas (4 reglas para transacciones genuinas 1 una regla para transacciones fraudulentas) que han cubierto 513 ejemplos de las 525 muestras, haciendo un 97.7% de cobertura del total de la muestra. De las simulaciones realizadas se deduce que el aumento de individuos en el proceso genético se obtiene como resultado la reducción del número de reglas y un aumento en el número de ejemplos cubiertos.



**Tabla 7.** Resultados del entrenamiento canal POS.

*5.1.3. Validación del entrenamiento*

En la tabla 8 se detalla los resultados obtenidos del proceso de verificación de las reglas generadas en el proceso de entrenamiento. El conjunto de datos de prueba está conformado de 240 transacciones (220 transacciones genuinas y 20 fraudulentas). En el escenario 1 el resultado de la precisión fue muy desalentador obteniéndose sólo un 31.3%. Al mejorarse la bondad de las reglas en el proceso de entrenamiento se obtuvo un mejor rendimiento en cada escenario. En el escenario 4 la precisión alcanzó un 95.8%, alcanzando un valor de TP de 211 de las 220 transacciones genuinas y un valor de TN de 19 de las 20 transacciones fraudulentas.



**Tabla 8.** Validación de reglas obtenidas canal POS.

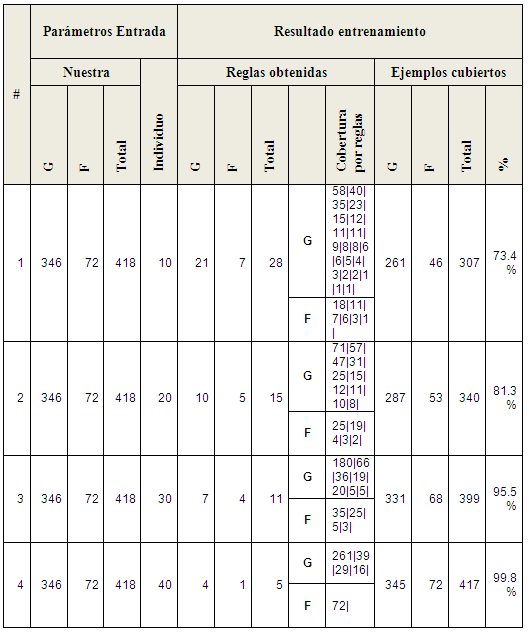
*5.2. Canal Internet*

*5.2.1. Parámetros iniciales de entrenamiento*

El proceso de entrenamiento ejecuta el algoritmo genético para 4 escenarios. Se tomó como muestra 418 transacciones por internet distribuidas en 346 transacciones genuinas y 72 transacciones fraudulentas. La población es inicializada con 10 individuos los cuales son incrementados aritméticamente en 10 unidades llegando a un tope de 40 individuos. El factor de mutación se mantiene uniforme en 2% de la cantidad de individuos. El porcentaje del parámetro completitud es del 98% y la diferencia es para el parámetro complejidad. El factor de poda que controla el número de ejemplos que aún no han sido cubiertos durante el proceso es del 5% del total de transacciones de muestra.

*5.2.2. Resultados del entrenamiento*

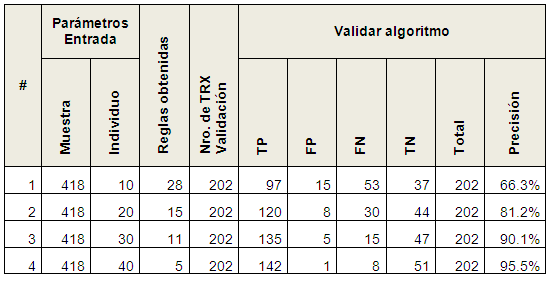
Las reglas obtenidas del proceso de entrenamiento se describen en la tabla 9. En la columna reglas obtenidas se cuantifica los individuos derivados de la ejecución del algoritmo genético, diferenciando las transacciones genuinas de las fraudulentas, además se detalla el nivel de cobertura de cada individuo. En la columna ejemplos cubiertos se totaliza las coincidencias encontradas, se elige el escenario que posea el menor número de reglas que cubran la mayor cantidad de transacciones de muestra. En el escenario 1 se ejecutó el algoritmo genético con una población de 10 individuos obteniéndose 28 reglas y un 73.4% de cobertura, sucesivamente se aumentó el número de individuos con el fin de mejorar estos resultados. En el escenario 4 se obtuvo 5 reglas (4 reglas para transacciones genuinas 1 una regla para transacciones fraudulentas) que han cubierto 417 ejemplos de las 418 muestras, haciendo un 99.8% de cobertura del total de la muestra.



**Tabla 9.** Resultados del entrenamiento canal Internet.

*5.2.3. Validación del entrenamiento*

Con el objetivo de determinar la precisión de las reglas obtenidas en el proceso de entrenamiento, se verificó los resultados con un conjunto de datos de prueba conformada por 202 transacciones (150 transacciones genuinas y 52 transacciones fraudulentas). En la tabla 10 se detalla los resultados del proceso de verificación, en función de los valores TP, FP, FN y TN. En el escenario 4 la precisión alcanzó un 95.5%, alcanzando un valor de TP de 142 de las 150 transacciones genuinas y un valor de TN de 51 de las 52 transacciones fraudulentas.



**Tabla 10.** Validación de reglas obtenidas canal Internet.

6. CONCLUSIONES

En este estudio se analizó la problemática del descubrimiento de amenazas de fraude electrónico en transacciones financieras que involucra los canales POS e Internet. Los resultados experimentales de la investigación son alentadores mostrando que el método empleado es capaz de alcanzar una buena precisión.

En la revisión de literaturas para la detección de fraude; los resultados de los autores muestran que las técnicas de detección de fraude basado en algoritmos evolutivos han alcanzado un mejor nivel de precisión, el ratio TP llegó a 100% a comparación de los otros modelos.

El algoritmo genético planteado sigue el enfoque Iterative Rule Learning (IRL) donde, la solución global está formada por las mejores reglas de una serie de ejecuciones sucesivas. En el diseño del algoritmo, se ha seleccionado la Codificación Natural como la forma de representación del individuo, permitiéndonos disminuir el cardinal del conjunto de posibles soluciones.

Se ha desarrollado un algortimo genético, el cual permite conocer el hábito de compra del tarjetahabiente peruano con base a la discretización de 20 variables continuas, de las cuales 10 son analizadas para el canal POS y 10 para el canal Internet. En el proceso de entrenamiento del algoritmo se observó que el aumento en el número de individuos de la población se obtuvo como resultado la reducción del número de reglas de decisión y un mejor nivel de cobertura. En el análisis del cuarto escenario del canal POS se obtuvo 5 reglas (4 reglas de transacciones genuinas y 1 regla de transacciones fraudulentas) con un porcentaje del 97.7% de cobertura de los 525 registros de muestra. Para el caso del canal Internet, el cuarto escenario obtuvo 5 reglas (4 reglas de transacciones genuinas y 1 regla de transacciones fraudulentas) con un porcentaje del 99.8% de cobertura de 418 registros.

En la verificación del modelo la precisión de la predicción aumentó progresivamente con base al incremento de los individuos en cada escenario. En la revisión del cuarto escenario del canal POS, la precisión de las variables TP y TN bordeó los 230 aciertos que equivale al 95.8% de un total de 240 registros; las transacciones que el sistema dejó pasar siendo fraudulentas (FP) y aquellas transacciones que no permitió realizar la operación (FN) de pago sumó 10 registros. Para el caso del canal Internet el cuarto escenario obtuvo la precisión de 95.5% que equivale a 193 aciertos de 202 operaciones, y fueron 9 transacciones no predichas correctamente por las reglas.

7. REFERENCIAS BIBLIOGRAFICAS

[1] Richardson Robert, 2010/2011 Computer Crime and Security Survey [Encuesta]. 15ta edición anual de CSI - Computer Security Institute, 2011.

[2] CyberSource Corporation, Informe sobre fraude en comercio electrónico 2012 – Tendencias de fraude en pagos en línea, prácticas de los comercios y comparativos [Encuesta]. 13ra edición anual de CyberSource, 2012.

[3] Frost A. y White Sullivan, Retos Clave Contra el Fraude Electrónico en las Instituciones Bancarias y Financieras de Latinoamérica [Publicación]. Octubre del 2010 Frost & Sullivan.

[4] Panigrahi Suvasini, Kundu Amlan, Sural Shamik, Majumdar A.K., Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning [Publicación]. Journal Information Fusion, octubre de 2009. Vol. 10 issue 4, págs 354 – 363.

[5] Bentley Peter J., Kim Jungwon, Jung Gil H. y Choi Jong U., Fuzzy Darwinian Detection of Credit Card Fraud [Publicación]. In 14th Annual Symposium of the Korean Information Processing Society, 2000.

[6] Benson Edwin Raj, Portia A. Annie, Analysis on Credit Card Fraud Detection Methods [Publicación]. International Conference on, Computer, Communication and Electrical Technology (ICCCET), marzo del 2011. Págs 152 – 156.

[7] Kundu Amlan, Panigrahi Suvasini, Sural Shamik y Majumdar Arun K., BLAST-SSAHA Hybridization for Credit Card Fraud Detection [Publicación], Journal IEEE Transactions on: dependable and secure computing, diciembre 2009. Vol. 6, issue 4, págs 309 – 315.

[8] Srivastava Abhinav, Kundu Amlan, Sural Shamik, Majumdar Arun K., Credit Card Fraud Detection Using Hidden Markov Model [Publicación]. IEEE transactions on, dependable and secure computing, enero de 2008. Vol. 5.

[9] Adeyiga J. A., Ezike J. O., Omotosho A. y Amakulor W., A Neural Network Based Model for Detecting Irregularities in e-Banking Transactions [Publicación]. African Journal of Computing y ICT, diciembre del 2011. Vol. 4, issue 2.

[10] Herrera Francisco y Magdalena Luis, Genetic Fuzzy Systems: A Tutorial [Publicación]. Tatra Mountains Mathematical Publications, junio 1997. Vol. 13, págs 93-121.

[11] González Antonio. and Pérez Raúl, SLAVE: a genetic learning system based on an iterative approach [Publicación]. Journal IEEE Transactions on: Fuzzy Systems, 1999. – Vol. 7, págs. 176 – 191.

[12] Cordón O., Herrera F., Lozano M. y M. J. del Jesus, MOGUL: A Methodology to Obtain Genetic fuzzy rule-based systems Under the iterative rule Learning approach [Publicación]. International Journal of Intelligent Systems, 1998. – Vol. 14.

[13] Carvalho Deborah R. y Freitas Alex A., A genetic algorithm for discovering small disjunct rules in data mining [Publicación]. Applied Soft Computing 2002.

[14] Giráldez R. Raúl, Mejoras en eficiencia y eficacia de algoritmos evolutivos para aprendizaje supervisado [Tesis]. Sevilla: Universidad de Sevilla, departamento de lenguajes y sistemas informáticos, setiembre 2003.

[15] Noda E., Freitas Alex A. y Lopes H. S., Discovering Interesting Prediction Rules with a Genetic Algorithm [Publicación]. Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on. – Vol. 2, ISBN: 0-7803-5536-9.