Instrumentos de recolección de datos en Informática o en Ciencias de la Computación

Ciara Mendez Cruz
Universidad Nacional de Trujillo
Trujillo, Perú
t022700920@unitru.edu.pe

***Resumen*—La investigación en informática estudia múltiples dimensiones de la tecnología informática (software, hardware, etc.) y su uso en el mundo real. Investigar en este gran campo consiste en analizar el problema y plantear soluciones innovadoras: un nuevo algoritmo, una nueva arquitectura, etc. Por ello, es necesario la investigación en este campo y sobretodo elegir correctamente el instrumento que se usa para la recolección de datos. Uno de los instrumentos de recolección de datos en investigaciones de ciencias de la computación que se utilizó fue una aplicación que se implementó para la segmentación de imágenes basada en conglomerados, denominado Algoritmos Genéticos K-medias (AGKM) que permite analizar imágenes que con otros instrumentos sería casi imposible de realizar, por otro lado, otro instrumento usado fue el muestreo sistemático de grilla sobre la ubicación geográfica de los cultivos de palta hasta recolectar un total de 630 muestras enfermas y sanas y para el reconocimiento de estos datos se utilizaron 4 algoritmos: Redes Neuronales, Support Vector Machines y Random Forest, Naive Bayes, asimismo se explican estas dos investigaciones, en que consisten y que resultados obtuvieron.**

***Index Terms*—instrumentos, datos, investigación, informática.**

I. Introducción

La informática es un sistema digital que permite realizar tareas de recolección y procesamiento de datos, mientras que en otras investigaciones no relacionadas a este campo utilizan instrumentos de recolección de datos, como: encuesta, observación, entre otros, en investigaciones de ciencias de la computación se suele usar algoritmos y softwares que permiten determinar en gran medida y con mayor rapidez la información.

Por ello, en este informe se presenta información respecto a instrumentos de recolección de datos en investigaciones en el campo de informática o en ciencias de la computación. Se ha considerado para cada una de las investigaciones describir en que consistió su trabajo, el instrumento, método o técnica que usaron y finalmente que se logró con ello.

II. Instrumentos de recolección de datos

Los datos se pueden recopilar de una o más fuentes según sea necesario para proporcionar la información que se busca. Por ejemplo, para analizar las ventas y la eficacia de sus campañas de marketing, un minorista puede recopilar datos de clientes de registros de transacciones, visitas a sitios web, aplicaciones móviles, su programa de fidelización y una encuesta en línea.[3]

Los métodos utilizados para recopilar datos varían según el tipo de aplicación. Algunos implican el uso de tecnología, mientras que otros son procedimientos manuales. Los siguientes son algunos instrumentos comunes de recopilación de datos:

- Funciones automatizadas de recopilación de datos integradas en aplicaciones comerciales, sitios web y aplicaciones móviles.
- Sensores que recopilan datos operativos de equipos industriales, vehículos y otra maquinaria.

- Recopilación de datos de proveedores de servicios de información y otras fuentes de datos externas.
- Seguimiento de redes sociales, foros de discusión, sitios de reseñas, blogs y otros canales en línea.
- Encuestas, cuestionarios y formularios, realizados en línea, en persona o por teléfono, correo electrónico o correo ordinario.
- Grupos focales y entrevistas individuales.
- Observación directa de los participantes en un estudio de investigación.

II-A. Ejemplos

1. En la tesis titulada ***"Implementación de Algoritmos Genéticos para la Segmentación de Imágenes Satelitales por Conglomerados de la Región Puno – 2013"*** [1] se desarrolló una aplicación para la segmentación de imágenes basada en conglomerados, denominado Algoritmos Genéticos K-medias (AGKM). Esta aplicación fue propuesta debido al deficiente método de selección del valor de inicialización del algoritmo K-medias al tomar un número de conglomerados inicial de forma aleatoria o por cálculo de la observación visual, esto puede influir en el desempeño del algoritmo, haciendo que tenga una separación inadecuada o demore más tiempo en la búsqueda del número de conglomerados.

   - **Método de recopilación de datos**

     La fuente de donde se descargaron las imágenes satelitales es proveída de forma gratuita por el INPE - Instituto Nacional de Investigaciones Espaciales (Brasil). Con la unión de recursos financieros y tecnológicos entre Brasil y China, con una inversión superior a U$$ 300 millones, fue creado un sistema de responsabilidades divididas (30 % brasileño y 70 % chino), teniendo como objetivo la implantación de un sistema completo de sensoramiento remoto a nivel internacional.

     Este sistema está basado en una interfaz Web, accesible en www.dgi.inpe.br/CDSR, proyectada para una operación simple y de fácil comprensión por el usuario. El catálogo de imágenes de la DGI/INPE fue íntegramente concebido y desarrollado por la División de Procesamiento de Imágenes (DPI) conjuntamente con la División de Generación de Imágenes (DGI) del INPE (Ver Figura 1.).

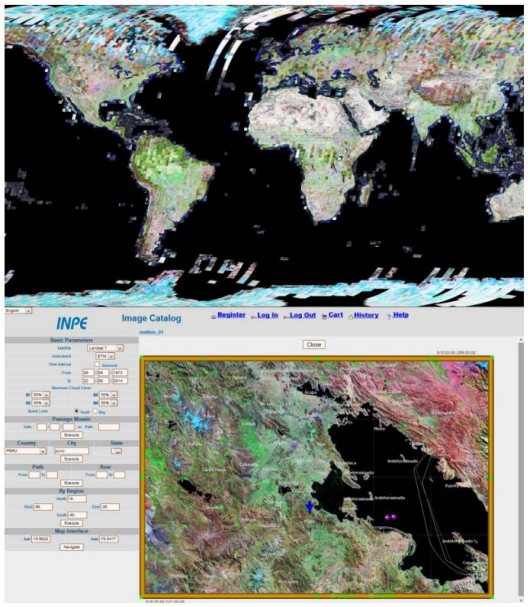

Figura 1. Catalogo de imágenes INPE

   - **Método de tratamiento de datos**

     El tratamiento de las imágenes satelitales comprende dos pasos, el primero trata las imágenes mediante el software AGKM implementado en lenguaje Matlab. Este software tiene implementado los algoritmos K-medias y AGKM, a partir del procesamiento con este software es que se obtiene la entropía de cada imagen las que fueron almacenadas en una tabla y posteriormente evaluadas estadísticamente con asistencia del SPSS para probar la eficacia del algoritmo.

     El algoritmo propuesto es el AGKM compuesto de dos partes, la primera parte se encarga de buscar el número ideal de conglomerados en la imagen satelital(Ver Figura 2.) por medio de los Algoritmos Genéticos (valor de inicialización para Kmedias), la segunda parte comprende el algoritmo K-medias que utiliza como entrada el número de conglome-

rados encontrados por los algoritmos genéticos.

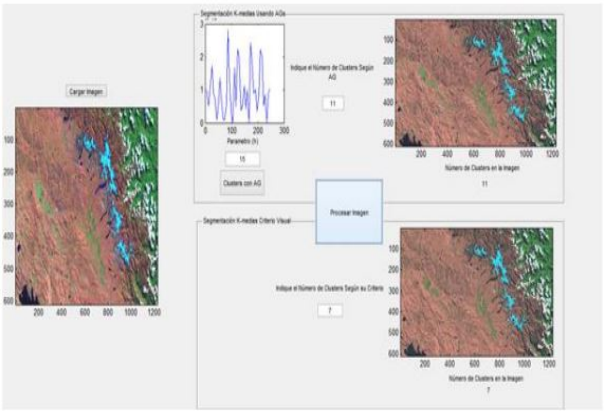

Figura 2. Máscara de la aplicación AGKM

Finalmente, en la tesis anteiormente mencionada se implementó un aplicación usando la metodología de los Algoritmos Genéticos (AGs) y K-medias, el primero tiene la finalidad de encontrar un número de conglomerados existentes en la imagen y el segundo realiza el proceso de separación. La métrica usada para evaluar la eficiencia de este algoritmo es el valor de la entropía en las imágenes, los resultados obtenidos son sometidos a una prueba estadística que nos indica que existe una ligera mejoría. Y se concluyó que el AGKM ofrece una ligera mejoría con respecto al algoritmo K-medias tradicional en la segmentación de imágenes satelitales para la Región Puno.

2. En la tesis titulada ***"Aplicación de Algoritmos Inteligentes para Reconocimiento Automático de Enfermedades Foliares de Cultivo de Palta"*** [2] se aplicaron cuatro algoritmos inteligentes: Naive Bayes, Random forest, Redes Neuronales y Support Vector Machines y tuvo como objetivo general la determinación del algoritmo inteligente más eficaz para el reconocimiento de imágenes de la enfermedad foliar de palta.

   - **Técnica para recopilación de datos**

     Se realizó un muestreo sistemático de grilla sobre la ubicación geográfica de los cultivos de palta hasta recolectar un total de 630 muestras enfermas y sanas.

   - **Instrumento para recopilación de datos**

     Cámara fotográfica profesional - CANÓN EOS REBEL T5I (Full HD + HDMI) – Para recolección de datos.(Ver Figura 3 y 4.).

   - **Adquisición de datos**

     Se hizo el uso de Cámara fotográfica profesional - CANÓN EOS REBEL T5I (Full HD + HDMI) para adquisición de imágenes sanas y enfermas (Oligonychus sp.).

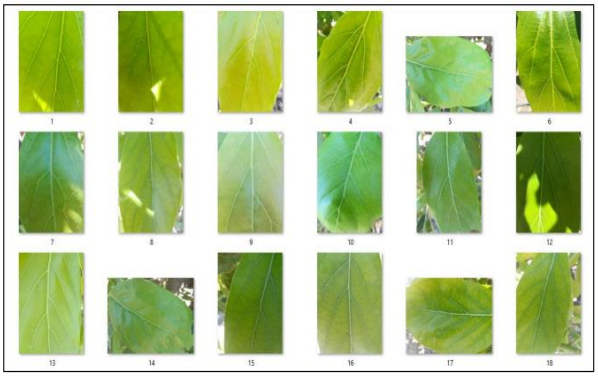

Figura 3. Se recolecto en total 330 imágenes sanas (se adjunta algunas imágenes)

Figura 4. Se recolecto en total 300 imágenes enfermas (Oligonychus sp.), se adjunta algunas imágenes

   - **Reconocimiento de datos**

     En esta fase se realizó el entrenamiento de los modelos y la evaluación de la exactitud. (Ver Figura 5) Se utilizaron 4 algoritmos: Redes Neuronales, Support Vector Machines y Random Forest, Naive Bayes. La matriz con los valores extraídos de las subimágenes se particionó en dos partes de manera

aleatoria: 80 % de las observaciones (filas de la matriz) para entrenamiento de los modelos (dataset de entrenamiento), y 20 % de las observaciones para validación de los modelos (dataset de validación). El dataset de entrenamiento se usó entonces para entrenar los 4 modelos. Una vez que se entrenó el modelo para cada clasificador, se usó el dataset de validación para evaluar la exactitud del modelo respectivo, comparando la clase de referencia de cada observación ('enferma' o 'sana') con la clase predicha por el clasificador.

Evaluación de desempeño de los modelos		
Algoritmos	**Correctos**	**Margen de Error**
Support Vector Machine	95,23810%	4,76190%
Redes Neuronales	94,44444%	5,55556%
Random Forest	91,26984%	8,73016%
Naive Bayes	86,50794%	13,49206%

Figura 5. Evaluación de desempeño de modelos

Luego de la evaluación de su eficacia se obtuvo que el algoritmo Maquina de Soporte Vectorial tiene mayor asertividad de 96 % en el reconocimiento de enfermedades foliares del cultivo de palta, esto después de ser evaluados con Matriz de Confusión.

III. Conclusiones

Este informe presentó información relevante respecto a dos ejemplos de instrumentos de recolección de datos en informática o en ciencias de la computación. Se ha explicado para cada una de las investigaciones en que consistió su trabajo, el instrumento, método o técnica que usaron y finalmente que se logró con ello, además de permitir reafirmar que la informática es un campo futurista avanzado y emergente de innovaciones importantes relacionadas con la forma de vida moderna.

Referencias

[1] M. Apaza Tito. Implementación de algoritmos genéticos para la segmentación de imágenes satelitales por conglomerados de la región puno-2013. 2014. URL http://repositorio.unap.edu.pe/handle/UNAP/4875.

[2] G. T. Castro Alvarez. Aplicación de algoritmos inteligentes para reconocimiento automático de enfermedades foliares de cultivo de palta. 2019. URL http://repositorio.unam.edu.pe/handle/UNAM/98.

[3] E. M. Coravin. Recopilación de datos: métodos, desafíos y pasos clave. URL www.techtarget.com/searchcio/definition/data-collection. (Accessed on 08/30/2022).