ESTRUCTURA DE DATOS PARA LA COMPRESION DE ARCHIVOS

Andrés Julian Caro Restrepo
Universidad Eafit
Colombia
ajcaror@eafit.edu.co

David Alvarez Grisales
Universidad Eafit
Colombia
dalvarezg@eafit.edu.co

Simón Marín
Universidad Eafit
Colombia
smaring1@eafit.edu.co

Mauricio Toro
Universidad Eafit
Colombia
mtorobe@eafit.edu.co

RESUMEN

A medida que avanza el mundo también se hace necesario que los avances tecnológicos se expandan para apoyar a todas las áreas y procesos del ser humano. Para el presente informe encontramozs una problemática en la clasificación de la salud animal en el contexto de la Ganadería de precisión, ya que esta trabaja con imágenes que deben de ser de alta precisión, que por lo general son pesadas y se dificulta el uso de estos sistemas en infraestructuras de red limitadas, como los que se tienen en zonas rurales. Para lo anterior proponemos usar el mismo sistema de análisis, apoyado por un algoritmo de compresión sin perdida (que todavía no sabemos cuál), para lograr los mismos resultados con imágenes que consuman menos capacidad de red y procesamiento.

Palabras clave

Algoritmos de compresión, aprendizaje de máquina, aprendizaje profundo, ganadería de precisión, salud animal.

1. INTRODUCCIÓN

Se hace necesaria la compresion de imágenes debido a que en los lugares en las que se requieren de estos tipos de sistemas para reconocer la salud del animal de ganado no se cuenta con una buena conexión a internet que permite el flujo de datos a la velocidad necesaria para descargar todo el peso de una imagen original.

Problema

En los últimos años se ha encontrado un aumento en el índice de enfermedad de las vacas en el área ganadera, lo que ha afectado la producción de la carne. Se es necesario un sistema que reconozca y clasifique los animales según su estado de salud en el contexto de Ganadería de Precisión, a través de imágenes, pero que este actúe de una manera rápida y eficaz, capaz de trabajar en la zona rural en la que no se cuenta con la mejor conexión de red posible, por lo que las imágenes deberán ser comprimidas sin perder la exactitud de los análisis necesarios.

1.2 Solución

En este trabajo, utilizamos una red neuronal convolucional para clasificar la salud animal, en el ganado vacuno, en el contexto de la ganadería de precisión (GdP). Un problema común en la GdP es que la infraestructura de la red es muy limitada, por lo que se requiere la compresión de los datos.

1.3 Estructura del artículo

En lo que sigue, en la Sección 2, presentamos trabajos relacional con el problema. Más adelante, en la Ssección 3, presentamos los conjuntos de datos y los métodos utilizados en esta investigación. En la Sección 4, presentamos el diseño del algoritmo. Después, en la Sección 5, presentamos los resultados. Finalmente, en la Sección 6, discutimos los resultados y proponemos algunas direcciones de trabajo futuras.

2. TRABAJOS RELACIONADOS

En lo que sigue, explicamos cuatro trabajos relacionados. en el dominio de la clasificación de la salud animal y la compresión de datos. en el contexto del PLF.

2.1 Sistema de gestión sistemática basado en RFID

La falta de garantía de la salud del ganado y poca productividad en los sistemas convencionales de gestión de ganado, provocan dificultades a los ganaderos por lo que es requerido hacer un cambio de los sistemas obsoletos y viejos, y empezar a desarrollar/usar nuevas tecnologías que permitan una mayor productividad gracias a que son mas eficaces que los viejos sistemas.

Algoritmo usado: Se hizo uso de un algoritmo basado en la identificación por radiofrecuencia (RFID), implementando plataformas como MySQL, Apache y JSP, además de programación en java como lenguaje base, MySQL como base de datos y NetBeans IDE 7.2 para el desarrollo del sistema.

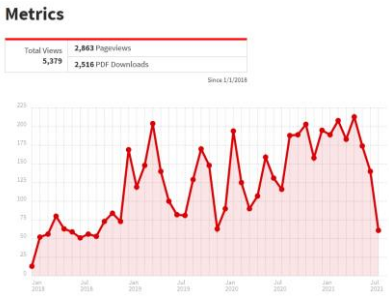
M. H. Ariff, I. Ismarani and N. Shamsuddin, "RFID based systematic livestock health management system," *in 2014 IEEE Conference on Systems*, (Kuala Lumpur, Malaysia,2014*),*111-116, doi: 10.1109/SPC.2014.7086240.

2.2 Ganadería de precisión para los porcinos.

Enfoque de la ganadería de precisión en los cerdos, ya que aunque el uso del GdP no es nuevo, no esta totalmente extendido y es necesario hacer enfoques para poder que funcione de manera óptima, por lo que enfocarse

específicamente en los cerdos ayuda a tener un sistema mas optimo y especializado.

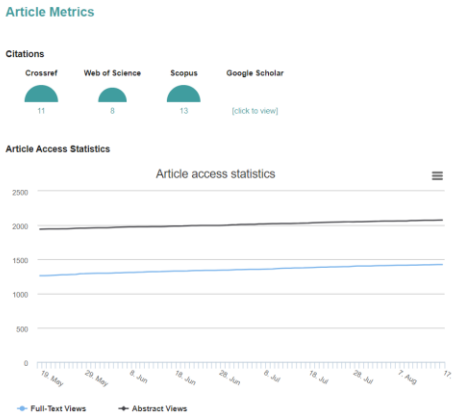
Algoritmo usado: En este problema el animal (los cerdos) son los sensores, se monitorea la temperatura, humedad, ingesta de alimento y agua, cámaras, micrófonos, etc. Posteriormente los algoritmos se encargan de traducir la información recolectada, para así conocer el estado del animal y verificar que se encuentre en un bienestar óptimo.



Erik Vranken, Dries Berckmans, Precision livestock farming for pigs, *Animal Frontiers*, 7 (1) 32–37, https://doi.org/10.2527/af.2017.0106

2.3 Algoritmo combinado para la ganadera de precisión.
Crear un algoritmo offline y online combinado para la clasificación del comportamiento de las ovejas en el contexto de la GdP., a pesar del alto potencial de un sistema de este tipo, es difícil realizarlo y que funcione bien porque existen problemas derivados al combinarlos, pero implementarlos de buena manera y en el contexto del GdP es una ayuda optima y largo plazo para las personas en el mundo de la ganadería.

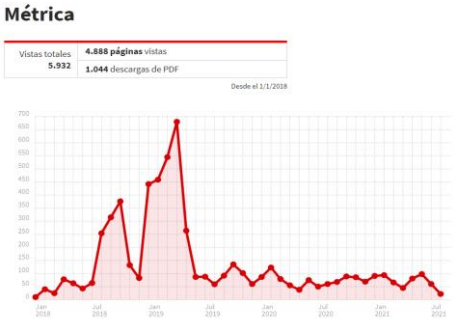
Algoritmo: Se creo un algoritmo online-offline combinado para monitorear el comportamiento de las ovejas a largo plazo, a pesar de los problemas que pueda presentar un algoritmo así, lograron hacerlo funcionar de manera óptima, este algoritmo clasifica tres comportamientos relevantes de las ovejas en tiempo real y también bajo condiciones de cambio.



Vázquez-Diosdado JA, Paul V, Ellis KA, Coates D, Loomba R, Kaler J. A Combined Offline and Online Algorithm for Real-Time and Long-Term Classification of Sheep Behaviour: Novel Approach for Precision Livestock Farming. *Sensors*. 2019; 19(14), https://doi.org/10.3390/s19143201.

2.4 Herramientas de ganadería de precisión en la industria láctea.
La falta de herramientas de ganadería de precisión, para la industria láctea hace que el ganadero tenga más difícil evitar las perdidas y sacar provecho total al ganado, por lo que es necesario también especializarse en partes especificas del ganado, en este caso las vacas lecheras y la industria láctea.

Algoritmo: Se creo un algoritmo que pudiera monitorear a tiempo real el estado de las vacas lecheras y en caso de que hubiera algún problema, dar la señal para intervenir de manera rápida y evitar que continúe el problema, el algoritmo recolecta la información y hace una valoración del estado de la vaca en tiempo real.



T. Norton, D. Berckmans, Developing precision livestock farming tools for precision dairy farming, *Animal Frontiers*, 7(1), 18–23, https://doi.org/10.2527/af.2017.0104

3. MATERIALES Y MÉTODOS

En esta sección, explicamos cómo se recogieron y procesaron los datos y, después, diferentes alternativas de algoritmos de compresión de imágenes para mejorar la clasificación de la salud animal.

3.1 Recopilación y procesamiento de datos

Recogimos datos de *Google Images* y *Bing Images* divididos en dos grupos: ganado sano y ganado enfermo. Para el ganado sano, la cadena de búsqueda era "cow". Para el ganado enfermo, la cadena de búsqueda era "cow + sick".

En el siguiente paso, ambos grupos de imágenes fueron transformadas a escala de grises usando Python OpenCV y fueron transformadas en archivos de valores separados por comas (en inglés, CSV). Los conjuntos de datos estaban equilibrados.

El conjunto de datos se dividió en un 70% para entrenamiento y un 30% para pruebas. Los conjuntos de datos están disponibles en https://github.com/mauriciotoro/ST0245-Eafit/tree/master/proyecto/datasets .

Por último, utilizando el conjunto de datos de entrenamiento, entrenamos una red neuronal convolucional para la clasificación binaria de imágenes utilizando *Teachable Machine* de Google disponible en https://teachablemachine.withgoogle.com/train/image.

3.2 Alternativas de compresión de imágenes con pérdida

En lo que sigue, presentamos diferentes algoritmos usados para comprimir imágenes con pérdida.

3.2.1 Tallado de costuras

Este algoritmo consta de tres pasos básicos.

- Asignarle a cada pixel un valor de energía.

- Encontrar ocho rutas conectadas del pixel con el menor valor de energía.

- Eliminar todos los pixeles de dicha ruta

Estos pasos se repiten las veces que sean necesarios de manera vertical u horizontal hasta llegar a la escala de imagen a la que se desea llegar.

Lo que hace es determinar en la asignación de valores de energía aquellos pixeles que en la ruta en sí no “cobran relevancia”, realizando así, una especie de recorte de lo no importante.

Original:

Con algoritmo:

Complejidad: O(n) constante.

3.2.2 Escalado de imágenes

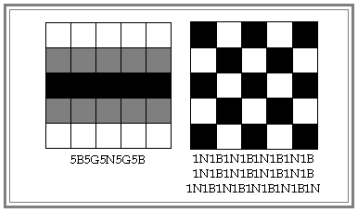
Consiste en el cambio de tamaño de una imagen digital. Se puede aumentar el número de pixeles o como es el caso, disminuirlos y es aquí, en donde se genera una pérdida de calidad visible.

Complejidad: O(n log(n))

By User:Kieff, User:Mysid - Derived from Image:Pixelart-tv-iso.png, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=1805834

3.2.3 Lzw

Realiza un análisis inicial del texto para identificar cadenas repetidas para armar un diccionario de equivalencias, asignando códigos En una segunda etapa, se convierte el texto utilizando los códigos equivalentes para las cadenas repetidas. Esto requiere dos etapas, una de análisis y una segunda de conversión.



3.2.4 LZS

Es un algoritmo que usa una combinación de LZ77 y el código de huffman. n compresor LZS busca coincidencias

entre los datos que se van a comprimir y los últimos 2 KB de datos. Si encuentra una coincidencia, codifica una referencia de desplazamiento / longitud al diccionario. Si no se encuentra ninguna coincidencia, el siguiente byte de datos se codifica como un byte "literal". El flujo de datos comprimidos termina con un marcador de fin.

3.3 Alternativas de compresión de imágenes sin pérdida

En lo que sigue, presentamos diferentes algoritmos usados para comprimir imágenes sin pérdida. *(En este semestre, ejemplos de tales algoritmos son la transformada de Borrows y Wheeler, LZ77, LZ78, la codificación Huffman y LZS).*

3.3.1 Formato PNG - Deflación

Es un formato de Compresión que puede almacenar hasta 16 millones de colores, permitiendo comprimir imágenes sin perder calidad.

Usa el método conocido como Deflación, que combina la codificación LZSS y la de Huffman, comprimiendo los datos por bloques, precedido por una cabecera de 3 bits.

Cabecera de 3 bits:

Primer bit: Marca si el bloque es el último del archivo

1: Es el último bloque del archivo

0: Hay más bloques por procesar después.

Segundo y Tercer bit: Determinan la codificación del bloque

00: Sección almacenada en bruto y literal.

01: Bloque de Huffman estático comprimido, usando un árbol de Huffman ya definido.

10: EBloque comprimido completado con la tabla de Huffman dada.

11: Reservado, no está en uso.

Los datos que sean compresibles se codifican en mayor parte a través del método de codificación dinámica de Huffman, que produce un árbol de Huffman optimizado, personalizado para cada bloque de datos individualmente. Las instrucciones para generar este árbol aparecen inmediatamente después del bloque de la cabecera.

La compresión se lleva a cabo en dos pasos:

- Se buscan cadenas de bits dubplicado y se reemplazan con punteros.
- Se reemplazan símbolo con otros nuevos basados en la frecuencia de uso.

Sacado de: https://es.wikipedia.org/wiki/Deflaci%C3%B3n_(algoritmo)

Imágenes sacadas de: https://es.wikipedia.org/wiki/Portable_Network_Graphics

3.3.2 Delta encoding

Es una técnica de compresión que se basa en la compresión compacta del archivo. Una función Delta, en el caso de los archivos, se refiere a la diferencia que existe dentro de un mismo archivo en dos versiones distintas.

Esta codificación tiene como objetivo obtener solamente los bytes que han sido modificados desde la última versión del archivo. Ya una vez comprimido el archivo, se puede obtener el original con la versión de referencia del archivo y el generado por el algoritmo Delta encoding.

https://es.wikipedia.org/wiki/Compresi%C3%B3n_de_imagen#La_compresi%C3%B3n_con_y_sin_p%C3%A9rdida_(LOSSLESS)

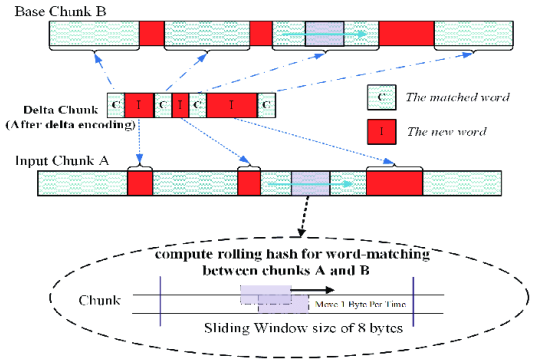


Imagen sacada de: https://www.researchgate.net/figure/A-run-length-coding-principle-example_fig3_331404982

3.3.3 Run-length encoding

Es una forma de compresión de datos en la que secuencias de datos con el mismo valor son almacenados como un único valor más su recuento. Este método se usa principalmente en datos que contengan muchas de estas secuencias, como iconos y logos, que son gráficos sencillos con áreas de color plano.

Usando esta compresión pasamos de:

```
BBBBBBBBBBBBNBBBBBBBBBBBBNNNBBBBBBBBBB
BBBBBBBBBBBBBBNBBBBBBBBBBBBBB
```

A:

```
12B1N12B3N24B1N14B
```

https://es.wikipedia.org/wiki/Run-length_encoding

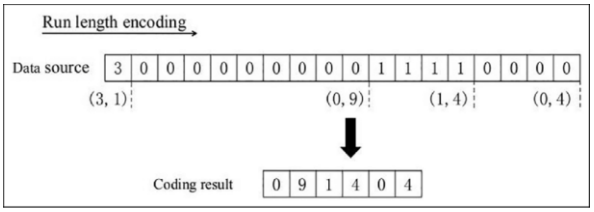


Imagen sacada de: https://www.researchgate.net/figure/A-run-length-coding-principle-example_fig3_331404982

3.3.4 DPCM

Codifican la diferencia entre un valor y el anterior, se utiliza para codificar los coeficientes DC de la Transformada Coseno Discreta. Estas presentan el nivel de gris medio del bloque, y muchas de las veces el nivel medio de un bloque será similar al del bloque anterior. Esto quiere decir que la diferencia entre dos bloques consecutivos es mínima.

Por ejemplo, al codificar una imagen donde hay gran parte de cielo, entonces se enviaran números más pequeño, ya que la diferencia entre los bloques será casi 0.

https://es.wikipedia.org/wiki/Codificaci%C3%B3n_entr%C3%B3pica

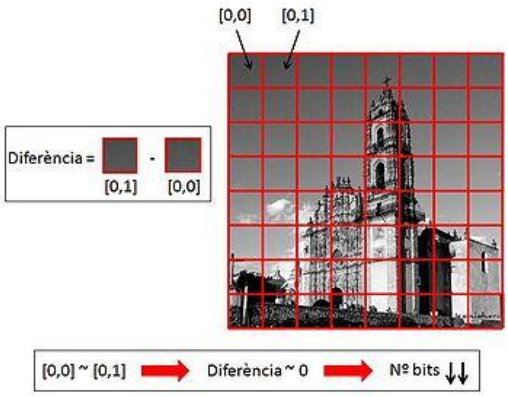


Imagen sacada de: https://es.wikipedia.org/wiki/Codificaci%C3%B3n_entr%C3%B3pica

REFERENCIAS

1. algoritmo de talla de costura - programador clic. (2020-2021). Programmer Click. https://programmerclick.com/article/4510360417/
2. M. H. Ariff, I. Ismarani and N. Shamsuddin, "RFID based systematic livestock health management system," *in 2014 IEEE Conference on Systems*, (Kuala Lumpur, Malaysia,2014*),*111-116.
3. Erik Vranken, Dries Berckmans, Precision livestock farming for pigs, *Animal Frontiers*, 7 (1) 32–37, https://doi.org/10.2527/af.2017.0106
4. Vázquez-Diosdado JA, Paul V, Ellis KA, Coates D, Loomba R, Kaler J. A Combined Offline and Online Algorithm for Real-Time and Long-Term Classification of Sheep Behaviour: Novel Approach for Precision Livestock Farming. *Sensors*. 2019; 19(14), https://doi.org/10.3390/s19143201.
5. T. Norton, D. Berckmans, Developing precision livestock farming tools for precision dairy farming, *Animal Frontiers*, 7(1), 18–23, https://doi.org/10.2527/af.2017.0104