Sistema de digitalización y estructuración de información clínica con técnicas de reconocimiento óptico de caracteres y procesamiento del lenguaje natural

Hugo Eduardo Castro Aranzábal
hcastro@aloe.ulima.edu.pe / Universidad de Lima. Lima, Perú

Walter Giancarlo Pinedo Barrientos
walterpinedo.barrientos@gmail.com / Universidad de Lima. Lima, Perú

Recepción: 12-7-2018 / Aceptación: 20-8-2018

Resumen. El presente trabajo busca desarrollar un sistema que permita la digitalización y estructuración de los registros clínicos apuntados por el doctor de forma tradicional mediante técnicas de reconocimiento óptico de caracteres y procesamiento del lenguaje natural, resultando en un proceso no intrusivo al flujo de trabajo. Además, permitirá poner a disposición estos datos para trabajos futuros.

Palabras clave: reconocimiento óptico de caracteres, visión de computadora, procesamiento de lenguaje natural, máquinas de soporte vectorial, clasificación internacional de enfermedades, frecuencia de término-frecuencia inversa de documento

Clinical information digitalization and structuring system using optical character recognition and natural language processing techniques

Abstract. This work aims to develop a system that allows the digitization and structuring of clinical records written in a traditional fashion by a doctor using optical character recognition and natural language processing techniques, proving in a non-invasive workflow. In addition, it will enable the use of the generated data for future work.

Keywords: optical character recognition, electronic computer vision, natural language processing, international statistical classification of diseases, term frequency – inverse document frequency, support vector machine

1. Introducción

La tasa de adopción de sistemas de registros clínicos electrónicos en países desarrollados es bastante alta, de 96,4 % entre hospitales (Charles, Meghan y Searcy, 2015), a diferencia de países en vías de desarrollo donde el uso de estos sistemas es mínimo (Candice y Erasmus, 2016). La realidad peruana no difiere tanto; se muestran pasos hacia la adopción de estos sistemas, los cuales se verán acelerados por leyes; un claro ejemplo es la Ley 30024, dictada en el 2013 como Ley que crea el Registro Nacional de Historias Clínicas Electrónicas (Ministerio de Salud, 2013).

Aun cuando algunos centros de salud han empezado a utilizar registros clínicos digitales, cuyo ingreso de datos es realizado manualmente por el doctor, existen otros centros donde únicamente el lapicero y papel siguen siendo usados para la entrega del diagnóstico al paciente. Por lo tanto, al tener toda esta información relevante tanto para el paciente como para la investigación científica del campo (Hilbert, 2015) en un medio no digital y no estructurado dificulta su análisis y administración.

En este escenario, el presente artículo busca la creación de una herramienta que facilite a las instituciones clínicas la transformación de datos no estructurados de los registros clínicos manuscritos a datos estructurados para luego almacenarlos en una base de datos. la herramienta implementará técnicas de reconocimiento óptico de caracteres (OCR) y procesamiento del lenguaje natural (NLP). El presente artículo se basa en la tesis de pregrado *Sistema de digitalización y estructuración de información clínica con técnicas de reconocimiento óptico de caracteres y procesamiento de lenguaje natural* de Castro y Pinedo (2018).

2. Metodología

2.1 Digitalización de caracteres

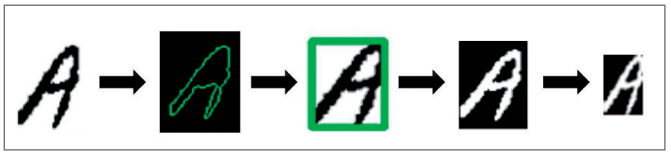
Para el desarrollo del reconocimiento óptico de caracteres se realizaron dos aplicaciones escritas en C++ utilizando las librerías de OpenCV, útiles para el preprocesamiento de las imágenes, y la ejecución del clasificador kNN. Además, las imágenes utilizadas se obtuvieron del NIST Special Database 19 (SD19), hechas por 4169 escritores (Grother y Hanaoka, 2016), con lo que captura una gran cantidad de variaciones por cada letra.

En cuanto a la primera aplicación, esta fue utilizada para la asignación de clases a cada imagen, el preprocesamiento consistió en convertir la imagen a 8 bits, aplicar una binarización adaptativa e inversa de la escala de grises, lo cual simplificará los cálculos dado que las zonas negras tienen un valor de 0 (Ouchtati, Redjimi y Bedda, 2015).

Una vez obtenida la imagen binarizada se procede a identificar su contorno más externo y grande, lo que nos permitirá descartar posibles ruidos y conocer la posición en la que se

encuentra la letra dentro de la imagen, así como realizar el procesamiento solamente en los pixeles donde aparece la letra. Sin embargo, también se eliminará la virgulilla de la ñ, el punto de la i y las tildes.

A partir de la imagen binarizada se extrae solamente en el contorno que conforma la letra y redimensiona a un tamaño de 20 por 30 pixeles (Pradeep, Srinivasan y Himavathi, 2011) para tratar imágenes de distinto tamaño, finalmente esta se vectoriza. Este proceso se puede ver en la figura 1. Finalmente, se le asignará una clase representada por el código ASCII decimal y tanto la imagen como las clasificaciones son guardadas en dos archivos de formato XML.



Nota: de izquierda a derecha, detección de contornos, definir región de interés, extracción de región de interés, imagen redimensionada a 20 por 30 pixeles.

Figura 1. Proceso de tratamiento de imagen
Elaboración propia

Para la segunda solución, a modo de prueba se conformó un texto por imágenes no clasificadas, el cual se convirtió a una imagen de 8 bits; se aplicó un filtro bilateral para eliminar ruido gaussiano para luego realizar la binarización. Posteriormente, para detectar las palabras, se dilató la imagen hasta que las letras que estén contiguas se agrupen en lo que se podría describir como una gran mancha blanca. Como siguiente paso se extraen los contornos más exteriores de la imagen y utilizando las coordenadas de cada contorno, se ordenan de izquierda a derecha y de arriba hacia abajo.

Adicionalmente se redimensionaron las letras a 20 por 30 pixeles y, para mejorar los resultados, se aplicó una técnica de normalización a las letras, la cual permitió enderezarlas a un plano vertical. Con ello, cada una de estas imágenes serán vectorizadas y almacenadas en una matriz. Finalmente, con la data clasificada del proyecto anterior, el kNN podrá encontrar la letra que más se asemeje y se le asignará una clase a cada letra para finalmente conformar las palabras.

2.2 Categorización de texto

Una vez digitalizada la información se plantean dos subprocesos para poder categorizarla y estructurarla: procesamiento del lenguaje natural y clasificación del diagnóstico. En el primero se utilizan librerías especiales para extraer a partir del diagnóstico las posibles categorías de la base ICD-10. En el segundo subproceso se obtienen las posibles categorías y mediante la técnica SVM se logra encontrar la categoría de enfermedad a la que corresponde y se almacena en una base de datos.

El primer subproceso inicia con la clasificación POS (*parts of speech*) que refiere a una clasificación mediante la librería nltk: StanfordPOSTagger (Stanford Natural Language Processing Group, 2015). En esta se utiliza una etiqueta para clasificar cada parte de la oración, como sustantivos ("nc..."), determinantes ("da..."), adjetivos ("aq..."), etcétera (Castro y Pinedo, 2018). Con las etiquetas asociadas a cada palabra se procede con el *chunking*, donde se busca crear patrones que ayuden a seleccionar lo más relevante de la oración, como por ejemplo extraer partes de la oración que contengan sustantivo más adjetivo.

En seguida se aplicará un filtro eliminando las *stop words*, que son palabras que no aportan valor para efectos de análisis de textos. Con este preprocesamiento se obtienen grupos para la evaluación de posibles categorías. Por último, se ejecuta el *stemming*, que es el proceso donde se extrae la raíz de las palabras y con estas se procederá a seleccionar en la base ICD-10 las enfermedades de mayor relación, obteniendo así las posibles categorías de enfermedades para la clasificación.

El segundo subproceso se inicia con la extracción de las posibles categorías. Una vez extraídas se inicia la vectorización. La vectorización es el proceso principal dentro del cual el documento se convierte en un vector, esto se logrará mediante las medidas TF-IDF (Mickevicius, Krilavicius y Morkevicius, 2015), que utilizan el contexto para evaluar la relevancia de una palabra en una colección de documentos. En este caso, la colección de documentos es todo el conjunto de enfermedades de las posibles categorías extraídas, mientras que los términos o *tokens* son las palabras extraídas de las mismas categorías.

Con base en estas medidas se evalúa la relevancia de cada término y se construye el modelo SVM con el vector que indica a qué categoría pertenece cada vector ingresado y el conjunto de vectores, donde existe un vector por posible categoría con los valores TF-IDF en cada columna que corresponde a un *token* cada una; en caso de que el *token* no se encuentre en la posible categoría se colocará el valor de 0, como se observa en la figura 2.

```
Matriz a ingresar:
[ [0, 0, 0, 2, 1, 0, 3, 0, 0, 0],
  [3, 4, 3, 0, 1, 0, 0, 3, 3, 0],
  [0, 0, 3, 0, 1, 0, 3, 0, 3, 0],
  [3, 4, 0, 2, 1, 0, 0, 3, 0, 0],
  [0, 0, 0, 2, 0, 3, 0, 0, 0, 3]]
Vector a predecir (Diagnostico):  [0, 0, 3, 0, 1, 0, 0, 0, 0, 0]
```

Figura 2. Representación del conjunto de vectores a ingresar y el vector a predecir
Elaboración propia

Por último, para predecir se convierte con los valores TF-IDF al diagnóstico en un vector. Dicho vector se podrá evaluar y clasificar. En caso de que no encuentre una categoría por ser probable que usen algunos términos coloquiales que no están registrados en el ICD-10, entonces el diagnóstico tendrá el código "ZZ100" con descripción "no_catalogado", que posteriormente pasará a revisión con la intención de enriquecer el diccionario interno del centro de salud (Castro y Pinedo, 2018).

2.3 Validación

Para el reconocimiento de caracteres se aplicó *11-fold cross-validation* porque permite un mejor aprovechamiento de la data que se tiene disponible puesto que cada dato es validado al menos una vez. Con ella se realizó la optimización del número de vecinos cercanos (k); para ello se calculó la sensibilidad asignando a k valores impares comprendidos entre 1 y 21, y se escogió como óptimo aquel número con el que se obtuvo la mayor sensibilidad.

El siguiente tipo de *cross-validation* a utilizar es el de *holdout*, el cual consiste en separar la data de forma aleatoria para el entrenamiento y pruebas (Dadhania y Dhobi, 2012). El primer grupo consiste en 104 000 imágenes, 2000 por cada letra, con las clases conocidas por el kNN que proporciona una muestra balanceada por cada letra. Para las pruebas se usaron 10 400 imágenes, 200 muestras por cada letra, sin clase conocida por el kNN. Resultado de ello se armó una matriz de confusión multiclase de 52 por 52, lo cual permitió calcular la precisión, sensibilidad y especificidad para cada clase.

Para la validación en la categorización de texto se utiliza la técnica *MicroAveraging*, en esta se crean 10 particiones (*10-fold cross-validation*), donde se repite el experimento 10 veces con distintos conjuntos de datos de entrenamiento y evaluación, calculando así los aciertos y fallos en cada clase acumulativamente y calculando sobre estos valores el indicador final. (Perea, Martín, Mointejo y Diaz, 2009). Además se utilizará la base central del Sistema Metropolitano de la Solidaridad (SISOL) de la municipalidad de Lima que contiene enfermedades dadas en dicho centro del año 2004 al 2012.

3. Resultados

En cuanto al OCR, el número de vecinos cercanos óptimo que se obtuvo fue de cinco con una sensibilidad de 77,5 %; con dicho valor y producto del *holdout cross-validation* se obtuvo una sensibilidad promedio de 78,69 %. Es por ello que se aplicó el enderezamiento de letras, donde se obtuvo una exactitud del modelo de 79,53 % y debido a que posteriormente se convertirán a minúsculas se puede dejar de considerar el evento mencionado anteriormente como un error.

Con lo anterior se obtuvo una precisión de 92 % y una sensibilidad de 92,1 %, bastante comparable con los 95,3 % (Rasmussen, Peissig, McCarthy y Starren, 2012) y 92,4 % (Biondich

et al., 2002) encontrados en la literatura, siendo la mayor diferencia que solamente entrenaron un número limitado de caracteres o dígitos, restricción que no podemos realizar. Por otro lado, la especificidad permite concluir que el clasificador puede obtener resultados confiables y diferenciar entre clases un promedio de 99,84 % de veces, ligeramente mayor a la especificidad de 99,6 % (Rasmussen *et al*., 2012).

Sin embargo, se pueden apreciar clases que resultaron con valores muy diferentes a las demás, como es el caso de la "I", que posee una precisión y sensibilidad baja debido a su gran similitud con la letra "l" y con la misma "i" (figura 3). Algo similar sucede con la letra "h", que es confundida en menor medida con la letra "t"; la "q" con la "g"; la "x" con la "Y", y la "y" con la "v". En otros casos que suceden con mayor frecuencia, la "i" con la "l", la "Q" con la "O", y las letras "a" y "e" con la "c"; la "r" con la "g", la "j" con la "i".

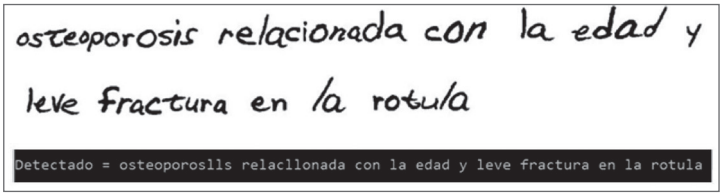


Figura 3. Resultado del reconocimiento óptico de caracteres
Elaboración propia

En cuanto a la clasificación de texto, se evalúa tanto la precisión como sensibilidad individualmente para ver oportunidades de mejora respectivas del modelo y las medidas a tomar. En esta ocasión el mínimo valor que tomó la precisión fue de 83,67 %, mientras que la sensibilidad fue de 79,25 %. Esto se debe a que en ocasiones, al haber algunas categorías iguales de enfermedades, el *software* clasificaba en una categoría "G80" cuando era "G81".

El resultado final bajo el esquema de *MicroAveraging* es una precisión del 91,53 % y una sensibilidad del 86,06 %. Así como en los estudios anteriores explorados, a comparación de otros algoritmos la técnica evidencia ser un eficaz clasificador que logra catalogar multiclases en el contexto de categorización de enfermedades en español (Castro y Pinedo, 2018).

4. Conclusiones

El presente artículo muestra el flujo para realizar la clasificación de términos médicos con letra imprenta partiendo de su digitalización hasta la categorización de enfermedades en base al ICD-10. En cuanto al reconocimiento óptico de caracteres, el preprocesamiento fue apropiado razón por la cual se obtienen buenos resultados con las imágenes obtenidas del NIST SD19.

Sin embargo, el preprocesamiento requerirá optimización de parámetros al digitalizar imágenes no presentes en esta para enfrentar cambios de iluminación, resolución de la imagen y tanto la calidad del papel como la tinta. Además se recomienda la incorporación de un corrector ortográfico, para corregir errores por la "ñ", tildes e inclusive errores simples al digitalizar letras como la "i".

Finalmente se puede concluir que el clasificador kNN utilizado para el reconocimiento óptico de caracteres demuestra un gran potencial a pesar de tratarse de una técnica bastante simple. En suma, el procesamiento de texto mediante técnicas de procesamiento de lenguaje natural y el proceso de categorización mediante la técnica SVM ha sido desafiante en esta oportunidad, debido a que en las investigaciones consultadas no se observa el proceso de vectorización y clasificación de manera tan explícita.

Si bien se pudieron mejorar en gran medida los resultados obtenidos, todavía existe espacio para mejoras, por ejemplo, realizar una extracción de características (Sun, 2015); durante el estudio solo se probó con adelgazamiento de la imagen. Por otro lado, al combinar esta técnica con un enderezado, fue posible mejorar los resultados en un 4 % en algunos casos. Esto se debe a que no solamente se tiene que aplicar esta técnica sola, sino que se tienen que aplicar varias técnicas para extraer múltiples características y obtener mejores resultados (Ouchtati *et al*., 2015).

Respecto a la categorización de textos se recomienda realizar un estudio más profundo en las técnicas de procesamiento de lenguaje natural que no fueron tratadas y afinar mejor la categorización de enfermedades, esto significa un *chunking* mejorado y un mejor POSTagging para encontrar más preciso el rol de la palabra, así como un trabajo más detallado en el contexto de la oración en el diagnóstico.

Referencias

Biondich, P., Overhage, M., Dexter, P., Downs, S., Lemmon, L., y McDonald, C. (2002). A modern optical character recognition system in a real world clinical setting: some accuracy and feasibility observations. Recuperado de http://www.ncbi.nlm.nih. gov/pmc/articles/PMC2244242/

Candice, N. y Erasmus, L. (2016). Electronic Medical Records: A developing and developed country analysis. Recuperado de http://iamot2016.org/proceedings/ papers/IAMOT_2016_paper_32.pdf

Castro, H. y Pinedo, W. (2018). *Sistema de digitalización y estructuración de información clínica con técnicas de reconocimiento óptico de caracteres y procesamiento de lenguaje natural*. Tesis de pregrado. Universidad de Lima.

Charles, D., Meghan, G. y Searcy, T. (abril de 2015). Adoption of Electronic Health Record Systems among U.S. NonFederal. Recuperado de https://www.healthit.gov/ sites/ default/files/data-brief/2014Hospital AdoptionDataBrief.pdf

Dadhania, S., y Dhobi, J. (2012). Improved kNN Algorithm by Optimizing Cross-validation. *International Journal of Engineering Research y Technology (IJERT) 1*(3), pp. 1-6. Recuperado de http://www.ijert.org/download/135/improved-knn-algorithm-by-optimiz ing-cross-validation

Grother, P., y Hanaoka, K. (2016). *NIST special database 19 hand printed forms and characters.* (Segunda edición). Recuperado de National Institute of Standards and Technology: https://s3.amazonaws.com/nist- srd/SD19/sd19_users_guide_edition_2.pdf

Hilbert, M. (2015). *Digital technology and social change* (Open Online Course at the University of California, freely available). Recuperado de https://canvas.instructure. com /courses/949415

Mickevicius, V., Krilavicius, T. y Morkevicius, V. (2015). *Classification of short legal lithuanian texts*. Recuperado de http://bpti.lt/wp-content/uploads/2016/02/bsnlp2015.pdf

Ouchtati, S., Redjimi, M., y Bedda, M. (2015). An offline system for the recognition of the fragmented handwritten numeric vhains. *International Journal of Future Computer and Communication 4*(1), pp. 33-39. Recuperado de http://www.ijfcc.org/vol4/351-C032. pdf

Perea, J., Martín, M., Montejo, A., y Diaz, M. (2008). Categorización de textos biomédicos usando UMLS. *Procesamiento del Lenguaje Natural 40* (pp. 121-127). Recuperado de http://www.sepln.org/revistaSEPLN/revista/40/todo.pdf

Pradeep, J., Srinivasan, E., y Himavathi, S. (marzo de 2011). Neural Network based handwritten character recognition system without feature extraction. En: *International Conference on Computer, Communication and Electrical Technology-ICCCET 2011, 18th y 19th March*, (pp. 40-44). Recuperado de http://ieeexplore.ieee.org/document/5762513/

Rasmussen, L. V., Peissig, P., McCarty, C., y Starren, J. (junio de 2012). Development of an optical character recognition pipeline for handwritten form fields from an electronic health record. Recuperado de https://jamia.oxfordjournals.org/content /19/e1/e90

Stanford Natural Language Processing Group (2015). *Spanish FAQ for Stanford CoreNLP, parser, POS tagger, and NER*. Recuperado de https://nlp.stanford.edu/software/ spanish-faq.shtml

Sun, H. (2015). k- Nearest Neighbour and SVM classifier with feature extraction and feature selection. Recuperado de http://homepages.rpi.edu/~sunh6/15fall6967.pdf

Bibliografía

Ministerio de Salud del Perú. (2013). *Registro Nacional de Historias Clínicas Electrónicas.* Recuperado de http://www.minsa.gob.pe/renhice/?op=1

World Health Organization (1993). *The ICD-10 Classification of Mental and Behavioural Disorders, Diagnostic Criteria for Research*. Recuperado de: http//apps.who.int/iris/bitstream/10665/37108/1/9241544554.pdf