Clasificación automática de Historias Clínicas basada en Prototipos utilizando técnicas de Procesamiento de Lenguaje Natural

Roque E. López Condori[1] Dennis Barreda Morales[2] Javier Tejada Cárcamo[2]
Luis Alfaro Casas[1]

[1]Universidad Nacional de San Agustín

[2]Universidad Católica San Pablo

rlopezc27@gmail.com, dennis.barreda@ucsp.edu.pe, jtejadac@ucsp.edu.pe,

casas@unsa.edu.pe

Resumen

Este trabajo presenta un método supervisado para resolver el problema de clasificación de historias clínicas. El clasificador propuesto consta de dos etapas: la primera utiliza técnicas de procesamiento de lenguaje natural para la generación de prototipos. En la segunda etapa se calculan las similitudes entre la historia clínica a clasificar y cada uno de los prototipos, eligiendo el más similar. Los resultados experimentales son alentadores pues indican que el método propuesto es adecuado para la clasificación automática de historias clínicas superando los resultados de enfoques tradicionales, tales como Naive Bayes y k-Nearest Neighbour.

1. Introducción

La clasificación automática de textos consiste en asignar un documento dentro de un grupo de clases previamente definidas (Coyotl, 2007). Si el documento pertenece sólo a una de las categorías, se trata de una *clasificación de una sola etiqueta*, caso contrario, es una *clasificación multi-etiqueta* (Cardoso, 2007)(Sebastiani, 2002). Por la naturaleza de las historias clínicas, el tipo de clasificación en este trabajo es de *una sola etiqueta*. Una historia clínica es el conjunto de documentos que contiene los datos sobre la situación y evolución clínica de un paciente a lo largo del proceso asistencial (Gisbert and Villanueva, 2004).

En la actualidad existe una gran cantidad de historias clínicas disponible. Toda esta información es improductiva si no se cuenta con mecanismos adecuados para su acceso, clasificación y análisis. La necesidad de poder utilizar esta información ha llevado a la creación de diversos medios de manipulación de información, entre las que se encuentra la *clasificación*. Sin embargo, el incremento constante de las historias clínicas, hace que la tarea de clasificación manual sea costosa y además consuma mucho tiempo, por lo que ha surgido un interés en realizar la clasificación de forma automática.

El método que se propone consta de dos etapas: entrenamiento y clasificación. En la primera se utilizan algunas técnicas de procesamiento de lenguaje natural para la extracción de características de las historias clínicas y con éstas se generan los prototipos. En particular se emplean estrategias de lematizacion de palabras, eliminación de palabras vacías y parametrización de historias clínicas. En la etapa de clasificación se calculan las similitudes entre la historia clínica a clasificar y cada uno de los prototipos, eligiendo el más similar.

El resto del documento se organiza de la siguiente manera. En la sección 2 se presentan algunos trabajos previos. En la sección 3 se explican los pasos que se realizan antes de que una historia clínica sea procesada por un clasificador. La sección 4 describe la etapa de entrenamiento. En la sección 5 se explica la etapa de clasificación. Los experimentos y resultados se encuentran en la sección 6. La discusión de los experimentos realizados se muestra en la sección 7. Finalmente en la sección 8 se exponen nuestras conclusiones.

2. Trabajos Previos

Un ejemplo de clasificación de textos en el ámbito médico se da en (Chapman et al., 2005), donde los textos describen las razones por la cual un paciente es internado. En este trabajo se utiliza una red bayesiana construida manualmente y las probabilidades se actualizan en la fase de entrenamiento.

En (Olszewski, 2003), se aplica el método Naive Bayes para la tarea de clasificación de triajes médicos. En (Larkey and Croft, 1996), tres clasificadores (K-Nearest Neighbor, relevance feedback y el clasificador bayesiano independiente) se aplican para asignar automáticamente códigos ICD-9 (International Classification of Diseases, ninth revision). Ellos muestran que la combinación de estos clasificadores obtienen el mejor rendimiento en la clasificación. (Argraw et al., 2007) presenta un método basado en aprendizaje supervisado, donde el clasificador se entrena en un corpus formado por historias clínicas etiquetadas.

3. Pre-procesamiento de las Historias Clínicas

El pre-procesamiento consiste en transformar las historias clínicas, de su formato original, a un modelo matemático adecuado para la tarea de clasificación. El pre-procesamiento consta de dos fases: la extracción de las características principales y la parametrización de las historias clínicas.

3.1. Extracción de características

El objetivo de esta fase es eliminar las partes de las historias clínicas que no sean importantes, es decir, que no aporten significado. Esta fase consta de los siguientes pasos:

- Eliminación de palabras vacías (*stop words*). Elimina palabras que no transmiten información (pronombres, preposiciones, conjunciones, etc.).
- Eliminación de símbolos de puntuación.
- Lematización de palabras. Elimina sufijos y afijos de una palabra de tal modo que aparezca sólo su raíz léxica. Por ejemplo, los vocablos *medicina*, *médico* y *medicinal* tienen la raíz léxica *medic*. Para este paso se utiliza el Algoritmo de Porter (Porter, 1980).

3.2. Parametrización de Historias Clínicas

Existen varias maneras de representar un documento; pero la más usada es el *modelo vectorial* (Salton et al., 1975). En este modelo, las historias clínicas se representan por vectores de palabras en un espacio de n dimensiones, siendo n el número de palabras en el texto (ver figura

1). De esta manera, las historias clínicas quedan representados como un vector $d = (w_1, ..., w_n)$, donde cada término indexado corresponde a una palabra en el texto y tiene un peso (w_i) que refleja la importancia del término. En nuestros experimentos el peso se representa por la frecuencia de aparición de un término en el texto.

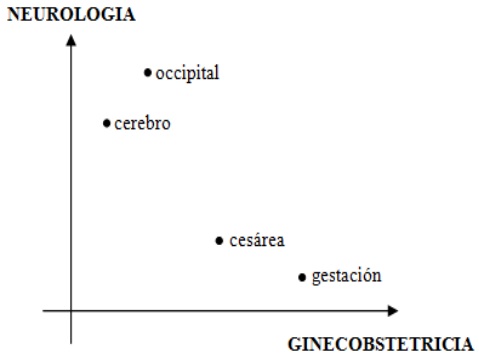

Figura 1: Espacio vectorial

4. Etapa de Entrenamiento

La construcción de un clasificador de textos mediante métodos de aprendizaje supervisado requiere de una colección de documentos previamente clasificados, conocido como *conjunto de entrenamiento* (Figuerola et al., 2004). En este trabajo se usaron 181 historias clínicas distribuidas en 6 clases para formar el conjunto de entrenamiento. En el primer paso se realiza la extracción de características y transformación de las historias clínicas al modelo vectorial (ver figura 2), tal como se explica en la sección 3.

Figura 2: Etapa de entrenamiento

Una vez que se tengan todas las historias clínicas de entrenamiento en el modelo vectorial, el siguiente paso consiste en buscar una historia clínica representante de cada categoría. El problema es cuál de ellas elegir como representante o, si es necesario, crear una historia clínica virtual que represente de mejor manera a cada categoría, a este representante se le conoce como *prototipo* (Cardoso and Oliveira, 2006)(Ramirez et al., 2010).

Palabras tales como *paciente*, *dolor*, *enfermedad*, *medicamento*, etc., aparecen de manera frecuente en todas las historias clínicas, y éstas no aportan información importante acerca de la clase a la cual pertenecen. Por este motivo, se propone un modelo de prototipo en el cual, el peso de cada palabra indique la importancia que tiene ésta para una categoría, y al mismo tiempo

sea discriminante para las demás categorías. Con este modelo, una palabra tendrá mayor valor para una clase, cuando más veces aparezca en ella y menos en las demás.

El peso de la palabra i-ésima $w_{i_{clase}}$, en relación a la clase se calcula como:

$$w_{i_{clase}} = tf_i \cdot log(\frac{N_{clases}}{n_{i_{clases}}}) \quad (1)$$

donde tf_i es el número de historias clínicas en la clase en los que la palabra i-ésima aparece, este valor es normalizado entre el total de documentos en la clase; N_{clases} es el total de clases; y $n_{i_{clases}}$ es el número de clases que tienen historias clínicas con la i-ésima palabra.

De esta manera se calculan los pesos de las palabras presentes en las historias clínicas para cada una de las categorías, formando así los prototipos representantes. El proceso de formar esos patrones se conoce como *entrenamiento* o *aprendizaje* (Debole and Sebastiani, 2003).

5. Etapa de Clasificación

Una vez que se tienen todos prototipos representantes, la etapa de entrenamiento o aprendizaje está concluida. Para clasificar nuevas historias clínicas, se estima la similitud entre el nuevo documento y cada uno de los prototipos (ver figura 3). El que obtenga un índice mayor nos indica la categoría a la cual se debe asignar la historia clínica.

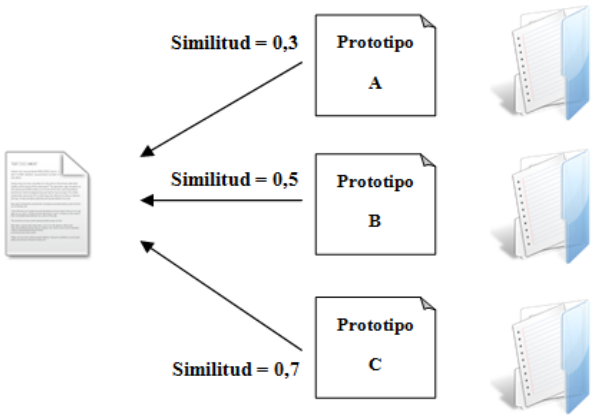

Figura 3: Etapa de clasificación

Es necesario establecer una manera de comparar la similitud entre la historia clínica a clasificar y cada prototipo. Se puede considerar como un problema en el cual se desea saber qué tantas características comparten la nueva historia clínica con los prototipos, y no sólo eso, sino también saber si las características que comparten son importantes o no. Esta importancia se puede saber con el peso que se asignó a los prototipos en la ecuación 1. En (Alvarez, 2009) se denota intersección pesada a esta forma de comparar documentos con las clases y se define como:

$$similitud(d, p) = \sum_{i \in d} w_{i_{doc}} \cdot w_{i_{clase}} \quad (2)$$

donde *d* es el documento que se quiere clasificar, *p* es el prototipo de la categoría, $w_{i_{clase}}$ es el peso de la palabra i-ésima en el prototipo *p*, $w_{i_{doc}}$ representa el peso de la palabra i-ésima en el documento, que en este caso es la frecuencia de aparición de la palabra.

6. Experimentos y Resultados

Se realizaron experimentos para evaluar el rendimiento del clasificador propuesto en base a *precisión* y *recall*, medidas de evaluación ampliamente utilizadas en este campo. También se realizó una comparación de la tasa de aciertos con los métodos de Naive Bayes y k-Nearest Neighbour(k=3). Para los experimentos se utilizó una colección de historias clínicas formada por 261 documentos los cuales están distribuidos de manera casi uniforme en 6 clases. Esta colección se elaboró con el fin de realizar pruebas sobre historias clínicas reales. La colección se creó de forma manual y cada historia fue asignada a una sola categoría por un médico profesional. La tabla 1 muestra la composición de esta colección.

Clases	**Documentos**
Gastroenterología	41
Ginecología	48
Neurología	42
Oncología	43
Traumatología	42
Urología	45
Total	**261**

Cuadro 1: Composición de la colección Historias Clínicas

La tabla 2 muestra una comparación de las precisiones obtenidas por los 3 clasificadores. La tabla 3 muestra el recall de cada método. El porcentaje de aciertos de los clasificadores se muestran en la figura 4.

Clases	**Naive Bayes**	**K-Nearest Neighbour**	**Método Propuesto**
Gastroenterología	0.92	0.79	**0.92**
Ginecología	0.9	1	0.93
Neurología	0.89	0.81	0.83
Oncología	0.71	0.71	**0.85**
Traumatología	0.9	0.9	0.85
Urología	0.71	0.79	**0.94**

Cuadro 2: Precisión para la colección Historias Clínicas

7. Discusión de los Experimentos

El objetivo de realizar los experimentos fue comprobar si el método propuesto es adecuado para la clasificación de historias clínicas. Los resultados mostrados en la figura 4, indican que el método propuesto obtiene la mayor tasa de aciertos en la clasificación automática de historias clínicas.

Clases	Naive Bayes	K-Nearest Neighbour	Método Propuesto
Gastroenterología	1	1	**1**
Ginecología	0.72	0.72	**0.77**
Neurología	0.71	0.75	**0.83**
Oncología	0.65	0.77	**0.85**
Traumatología	0.75	0.75	**0.92**
Urología	1	1	**1**

Cuadro 3: Recall para la colección Historias Clínicas

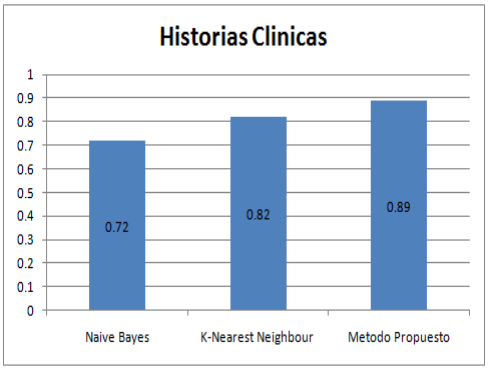

Figura 4: Tasa de aciertos en la clasificación de historias clínicas

En los textos de las historias clínicas aparecen términos médicos que son importantes para cada categoría, por ejemplo: *menstruación*, *embarazo*, *gestación*, son significativos para la clase ginecología, estas palabras aparecen con poca frecuencia en el texto. Sin embargo el clasificador propuesto toma en cuenta la importancia que tiene un término en una clase y no en el documento, a diferencia de los enfoques Naive Bayes y k-Nearest Neighbour. Es por eso que en el método propuesto los términos *menstruación*, *embarazo*, *gestación* tienen más importancia para la clase ginecología que para las demás, y es en base a estos términos importantes que se mejora la tasa de aciertos.

El mejor rendimiento, según la tabla 3, se obtiene en las clases gastroenterología y urología ya que son las que poseen términos médicos más diferenciados, tales como *epigástrico*, *digestivo*, *micción*, *próstata*, etc.

8. Conclusiones

En este trabajo se presentó un método para clasificación automática de historias clínicas el cual mejora los resultados del método Naive Bayes y k-Nearest Neighbour. Este método consta de dos etapas, en la etapa de entrenamiento se utilizan algunas técnicas de procesamiento de lenguaje natural para la extracción de características de las historias clínicas y con éstas se generan los prototipos. En la etapa de clasificación se calculan las similitudes entre la historia clínica a clasificar y cada uno de los prototipos, eligiendo el más similar.

Los puntos más importantes a resaltar del presente trabajo son, en primer lugar, que el método propuesto alcanzó niveles de exactitud que pueden competir con el desempeño de personas

capacitadas en la clasificación de historias clínicas. En segundo lugar, que la mayoría de los errores surgieron generalmente en las clases que contaban con una menor cantidad de texto en las historias clínicas.

En el presente trabajo también se proporciona una comparación entre los clasificadores anteriormente mencionados y el método propuesto. Estas comparaciones nos dan la conclusión de que el método propuesto es adecuado para la clasificación automática de historias clínicas. Otra conclusión muy importante es que la mejora de los resultados depende también del nivel de dificultad de la colección de datos que se utiliza.

Entre los principales trabajos futuros se encuentran: (1) Experimentar con otros métodos de parametrización de historias clínicas, como por ejemplo a través de *n-grams*. (2) Utilizar mecanismos de extracción de keywords (palabras claves), y en base a estos realizar la clasificación de los textos médicos.

Referencias

Alvarez, J. D. (2009). Clasificación automática de textos usando reducción de clases basada en prototipos. *Tesis de Maestría, Instituto Nacional de Astrofísica, Óptica y Electrónica*.

Argraw, A., Hulth, A., and Megyesi, B. (2007). General-purpose text categorization applied to the medical domain. *DSV Research report - 2007-016*.

Cardoso, A. (2007). Improving methods for single-label text categorization. *Ph.D. thesis, Universidade Técnica de Lisboa*.

Cardoso, A. and Oliveira, A. (2006). Empirical evaluation of centroid-based models for single-label text categorization. *INSEC-ID Technical Report 7/2006, pp. 3-7*.

Chapman, W., Christensen, L., Wagner, M., Haug, P., Ivanov, O., Dowling, J., and Olszewski, R. (2005). Classifying free-text triage chief complaints into syndromic categories with natural language processing. *Artificial Intelligence in Medicine 33, pp. 31-40*.

Coyotl, R. M. (2007). Clasificación automática de textos considerando el estilo de redacción. *Tesis de Maestría, Instituto Nacional de Astrofísica Óptica y Electrónica, México*.

Debole, F. and Sebastiani, F. (2003). Supervised term weighting for automated text categorization. *ACM (2003), pp. 784-788*.

Figuerola, C., Berrocal, J., Zazo, A., and Rodríguez, E. (2004). Algunas técnicas de clasificación automática de documentos. *pp. 1-3*.

Gisbert, J. A. and Villanueva, E. (2004). Médicina legal y toxicología. *pp. 102-103*.

Larkey, L. and Croft, W. (1996). Combining classifiers in text categorization. *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 289-297*.

Olszewski, R. (2003). Bayesian classification of triage diagnoses for the early detection of epidemics. *Proceedings of the FLAIRS Conference, pp. 412-416*.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, vol.14, no. 3, pp. 130-137*.

Ramirez, G., Montes, M., and Villaseñor, L. (2010). Enhancing text classification by information embedded in the test set. *Springer (2010), pp. 2-5*.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communication of the ACM*.

Sebastiani, F. (2002). Machine learning in automated text categorizatio. *ACM 34 (2002), pp. 1-47*.