Un Clasificador de Texto Por Aprendizaje

Peláez J.I. [(a)] La Red D. [(b)] Sánchez P.[(a)]

[(a)] Dpto. Lenguajes y Ciencias de la Computación
E.T.S.I. Informática. Campus de Teatinos. Universidad de Málaga
Málaga 29071. España
E-mail: jignacio@lcc.uma.es

[(b)]Dpto. Informática
Universidad Nacional del Nordeste
Corrientes. Argentina
E-mail: lrmdavid@exa.unne.edu.ar

Resumen

Uno de los problemas más importantes en telemedicina es la derivación de forma automática de pacientes al especialista apropiado de acuerdo a su sintomatología. Esta asignación es normalmente realizada por un profesional en medicina no especializado que partiendo de un prediagnóstico, normalmente expresado en lenguaje natural, determina la especialidad más adecuada. El objetivo de este trabajo es desarrollar un clasificador estomatológico por aprendizaje, que categoriza dicho prediagnóstico en un conjunto de especialidades.

Palabras clave: Clasificación de textos, Selección de Características, Telemedicina, Procesamiento de Lenguaje Natural.

1. Introducción

Es en la década de los años 60 cuando se presentan los primeros clasificadores automáticos de texto [7]. Desde estas fechas hasta la década de los 80 principios de los 90, la clasificación de textos se llevaba a cabo mediante un proceso manual que extraía el conocimiento del experto y lo representaba mediante reglas por medio de técnicas de ingeniería del conocimiento. Estas reglas se construyen como:

if <condición$_i$> **then** <clase$_j$>

donde, si el texto a clasificar satisface la condición *i*-ésima entonces es clasificado en la clase o categoría *j*-ésima. Un ejemplo de este tipo de clasificadores es el *Sistema Construe* [6], construido por Carnegie Group para la agencia de noticias Reuters.

La principal desventaja de este enfoque radica en la dificultad de extraer el conocimiento del experto, lo que provoca, que dichos clasificadores no sean portables, porque las reglas obtenidas son específicas del problema y del dominio; y difícilmente mantenibles, porque pueden surgir nuevas reglas que deben ser definidas por el experto.

Es en la década de los 90, cuando el *paradigma de la máquina que aprende* [9] emerge como un nuevo enfoque de clasificación que atrae el interés de

diferentes investigadores. En dicho enfoque aparece un proceso que se denomina *proceso general inductivo*, que construye de forma automática un clasificador por aprendizaje a partir de un conjunto de textos previamente clasificados. Para ello, este proceso extrae las características que debe tener un texto, desde unos ejemplos de entrenamiento dados por un experto, para pertenecer a una clase. Por lo tanto, con este enfoque el esfuerzo del ingeniero no se dirige hacia la construcción de un clasificador, sino que se dirige, hacia la confección de un proceso automático de construcción de clasificadores. De manera que, si el conjunto original de clases se actualiza o el sistema es portado a un dominio diferente, solamente es necesario realizar un nuevo entrenamiento a partir del nuevo conjunto de textos.

Las principales ventajas que presenta este enfoque son: *Efectividad*, no es necesario que un experto defina las reglas de clasificación; e *Independencia del dominio* de los textos a clasificar.

La gran mayoría de los clasificadores de textos por aprendizaje se basan en métodos de inducción probabilísticos [5] [6], esencialmente cuantitativos (numéricos), lo que conlleva una difícil interpretación de los resultados.

Otra clase de clasificadores que han experimentado un gran auge en los últimos años, son los simbólicos [3]. Estos se basan en la localización y posterior clasificación de los patrones más representativos del texto y determinantes de cada categoría. Los clasificadores construidos bajo este nuevo paradigma están alcanzando resultados que hacen de la clasificación automática por aprendizaje una alternativa cualitativa y comercialmente viable respecto a los clasificadores tradicionales.

El objetivo de este trabajo es presentar un clasificador estomatológico por aprendizaje. Para ello, se utilizarán unos prediagnosticos que han sido elaborados por facultativos en medicina general junto con las especialidades que finalmente atendieron a los pacientes. El trabajo ha sido organizado como sigue: en la sección 2, se presenta el problema de la clasificación de textos; en la sección 3, se presenta el clasificador; y finalmente, se muestran las conclusiones.

2. El Problema de la Clasificación de Textos

El problema de la clasificación de textos se puede definir como la forma de determinar la asignación de un valor $a_{ij} \in \{0, 1\}$ para cada entrada de una matriz de decisión (figura 1). Donde C = $\{c_1,...,c_m\}$ representaría el conjunto de categorías predefinidas, D = $\{d_1,...,d_n\}$ el conjunto de textos a ser clasificados, y a_{ij}, la decisión de clasificar d_j en la categoría c_i, de manera que si a_{ij} = 1 entonces el elemento d_j es clasificado en c_i y 0 en otro caso.

	d_1	...	d_i	...	d_n
C_1	a_{11}	...	a_{1j}	...	a_{1n}
...	...	...	...	...	...
c_i	a_{i1}	...	a_{ij}	...	a_{in}
...	...	...	...	...	...
c_m	a_{m1}	...	a_{mj}	...	A_{mn}

Figura 1. Matriz de decisión para el problema de la clasificación de textos

Dos observaciones para comprender la tarea de clasificación:

- Las categorías son solo etiquetas simbólicas. No se asume ningún conocimiento adicional de su *significado* como ayuda en el proceso de construcción del clasificador.

- La asignación de un texto a una categoría debe ser en general, realizada en base a la semántica del texto y no en base a metainformación. Por ejemplo: fecha de publicación, tipo de texto, fuente de publicación, etc. Es decir, la clasificación debería basarse en conocimiento endógeno (conocimiento que puede ser extraído del texto) y no en conocimiento exógeno (información que puede ser provista, para este propósito, por una fuente externa) [2][8].

Debido a que la semántica de un texto es una noción inherentemente subjetiva, la idea fundamental de la clasificación de textos radica en que la asignación de un texto a una categoría no puede decidirse determinísticamente. Esto se explica con el fenómeno de inconsistencia de interindexación [1]: Cuando dos humanos deben tomar la decisión de clasificar un texto d_j bajo una categoría c_i, ellos pueden no estar de acuerdo, hecho que ocurre con relativa frecuencia [4] [10].

En el diseño de un clasificador de textos por aprendizaje se plantean principalmente dos problemas:

- *Reducción de Dimensionalidad.* Dado un texto formado por un conjunto de r patrones, se extrae aquel subconjunto $r' << r$ formado por los patrones representativos.

  La reducción de dimensionalidad es necesaria por la alta dimensión del espacio de términos existentes en un texto. Por lo tanto, es necesario disponer de técnicas que ayuden a simplificar dicha dimensión reduciendo el espacio de términos de r a $r´ << r$. Una forma de reducir la dimensionalidad de un texto es la utilización de filtros de acuerdo a determinados criterios, que generalmente dependen del dominio.

  Además, la reducción de dimensionalidad permite eliminar parte del *problema de sobre adaptación* (*overfitting*), un fenómeno por el cual un clasificador de textos considera información relevante y no relevante [8].

- *Proceso Inductivo de Construcción* . Este proceso consiste en la construcción automática de un clasificador por aprendizaje a partir de una colección de textos de entrenamiento previamente clasificados.

  La construcción inductiva de un clasificador para una categoría c_i consta de dos fases: La primera es la definición de una función de clasificación: dado un texto d_j, devuelve un valor de clasificación perteneciente al intervalo [0, 1] que representa la certeza de clasificar d_j bajo c_i; y en segundo lugar, la definición de un umbral que determina un nivel de activación para tomar la decisión de clasificar d_j bajo c_i.

  El problema de la construcción inductiva del clasificador ha sido abordada de diferentes formas, destacando el enfoque probabilístico [7]. En este enfoque la clasificación se lleva a cabo en función de la probabilidad (aplicación del teorema de Bayes) que tiene un texto de pertenecer a una categoría. Este enfoque es esencialmente numérico, con la dificultad de interpretación que esto conlleva.

  Un enfoque alternativo es el simbólico. Éste se basa en la localización de patrones representativos en el texto y su posterior clasificación dentro de una estructura particular [3].

3. Un Clasificador Estomatológico

En esta sección se propone un clasificador de textos estomatológico simbólico, que realiza una categorización de los mismos utilizando los propios textos y un diccionario de términos especifico del contexto. Para ello en primer lugar se abordará el problema de reducción de dimensionalidad; en segundo lugar, el proceso de construcción inductivo; en tercer lugar se mostrará el proceso de entrenamiento; y finalmente se analizarán los resultados.

3.1. Reducción de dimensionalidad

Como se mencionó anteriormente, la reducción de dimensionalidad tiene como objetivo la obtención o selección de los patrones más representativos de un texto. En este trabajo dicho proceso se lleva a cabo mediante sucesivos filtrados sobre unos prediagnósticos estomatológicos preestablecidos por personal médico no especializado (figura 2).

Este proceso de filtrado es guiado mediante un diccionario de términos estomatológicos que contiene las palabras relevantes dentro de dicho dominio, y un diccionario de términos no informativos que contienen palabras no relevantes. Este diccionario de términos no relevantes parte de terminos como por ejemplo artículos, pronombres, etc, los cuales no aportan ninguna información (en principio) en la clasificación y se va completando mediante el proceso de aprendizaje del clasificador.

Figura 2. Filtrado de prediagnósticos

Este proceso se compone de cuatro filtros que son aplicados de forma sucesiva sobre el texto a clasificar. Como resultado de este proceso se obtienen dos conjuntos de patrones. El primero de estos conjuntos esta compuesto de patrones con una longitud de una palabra, mientras que el segundo conjunto esta compuesto de patrones con una longitud de dos palabras. El tamaño de estos patrones ha sido establecido de forma empírica, donde se ha comprobado que un aumento en la longitud del

patrón no produce un incremento en el número de elementos que son clasificados correctamente.

A continuación se muestran los filtros utilizados para realizar la reducción de dimensionalidad.

- Filtro de frases: Divide el texto en frases de acuerdo a signos de puntuación, conjunciones, etc.

  → Entrada: Texto.
  ← Salida: Texto dividido en frases.

- Filtro de palabras no relevantes: Se eliminan de las frases las palabras no significativas dentro del dominio.

  → Entrada: Texto dividido en frases.
  ← Salida: Frases sin palabras no relevantes.

- Filtro de diccionario estomatológico: Se eliminan las frases que no contengan palabras existentes en el diccionario de términos estomatológico.

  → Entrada: Frases sin palabras no relevantes.
  ← Salida: Frases con al menos una palabra del diccionario estomatológico.

- Filtro de conjuntos: Se genera el conjunto de palabras y el conjunto de frases. El conjunto de palabras está determinado por todas las palabras del prediagnóstico que pertenecen al diccionario de términos estomatológico. El conjuntos de frases contiene grupos de dos palabras determinados por la concatenación de todas las palabras del prediagnóstico que pertenecen al diccionario de términos estomatológico, con su palabra antecesora y predecesora (figura 3).

  → Entrada: Frases con al menos una palabra del diccionario estomatológico.

  ← Salida: Conjunto de palabras y conjunto de frases.

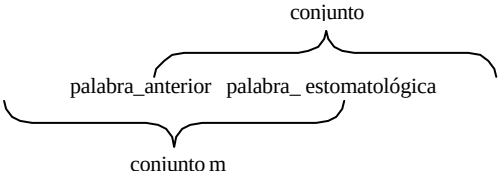

Figura 3. Patrones frase

3.2. Proceso de construcción inductivo

El proceso inductivo para la construcción del clasificador ha sido realizada mediante una red como la que se muestra en la figura 4. Esta red contiene todas las posibles categorías (especialidades en nuestro caso) asociadas a los patrones que las determinan, los cuales se encuentran relacionados mediante puertas lógicas *And* y *Or*. Además, se utilizan en las puertas *And* unos umbrales $m_i \in [0,1]$ que establecen cuando una especialidad debe ser activada dependiendo del número de patrones coincidentes.

La estructura de la red de clasificación es la siguiente:

- Especialidades solución: una por cada categoría propuesta para la clasificación.

- Nodos palabra: patrones representativos de longitud de una palabra.

- Nodos frase: patrones representativos de longitud de dos palabras.

- Puertas *Or*.

- Puertas *And* con umbral de activación que activa dicha especialidad dependiendo del número de patrones de palabras (m_1) y frases (m_2) presentes.

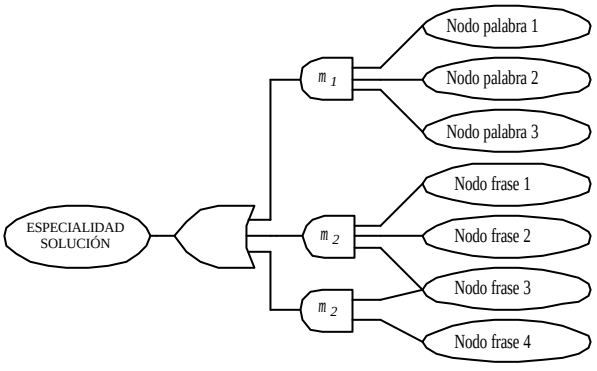

Figura 4. Ejemplo de la red de clasificación

Tanto en la etapa de entrenamiento de la red de clasificación, cuando es necesario decidir si crear un nuevo nodo o reutilizar uno existente, como en la etapa de evaluación, cuando se intenta hacer coincidir un nuevo nodo con alguno de los ya existentes, se utiliza una función de similitud entre

patrones $F(patrón_1, patrón_2)$ que calcula un valor $\in [0,1]$ en función del número de caracteres coincidentes entre dos patrones, junto con un umbral de similitud $m_s \in [0,1]$, de manera que si $F(patrón_1, patrón_2) > m_s$, el $patrón_1$ se considera igual al $patrón_2$. Por lo tanto, cuanto más cercano es el umbral de similitud a 1, más estricta es la comparación de igualdad entre dos nodos. Además, esta función de similitud permite aumentar o disminuir el tamaño de la red de clasificación, de forma que, definiendo un umbral de similitud bajo se obtiene una red de menor tamaño, eso si, a costa de una diferenciación más baja entre patrones, mientras que con un umbral de similitud alto se produce el efecto contrario.

3.3. Proceso de entrenamiento

El entrenamiento de la red de clasificación se realiza en base a unos prediagnósticos estomatológicos, los cuales están formados generalmente por un breve historial estomatológico, dolencias manifestadas y un examen visual previo del paciente, realizado por un profesional de la salud no especializado.

El proceso de entrenamiento seguido se muestra en la figura 5. En éste, para cada prediagnóstico, en primer lugar se realiza un filtrado del texto con los filtros anteriormente expuestos; en segundo lugar, se crean dos puertas *And*, una para los conjuntos de palabras y otra para los conjuntos de frases resultantes del proceso de filtrado; y finalmente, teniendo en cuenta el umbral de similitud, se crean o se reutilizan los nodos necesarios para la inclusión del nuevo prediagnóstico dentro de la red de clasificación.

Una vez que la red ha sido entrenada, se realiza un proceso de poda sobre la misma, con el objetivo de reducir el tamaño de la red. Básicamente, este proceso de poda consiste en eliminar aquellos nodos que por si mismos determinan varias especialidades. En la figura 5 se muestra el algoritmo de entrenamiento.

```
Entradas: Lista de prediagnósticos, Umbral de similitud
procedimiento Entrenamiento
  Para todas los prediagnósticos de la Lista de prediagnósticos
    Aplicar filtrado a prediagnóstico (Prediagnóstico)
    Conjunto de palabras =
      Obtener conjunto de palabras (Prediagnóstico filtrado)
    Conjunto de frases =
      Obtener conjunto de frases (Prediagnóstico filtrado)
    Crear puerta AND de nodos palabra
    Para todos los nodos palabra del Conjunto de palabras
      Si No Existe nodo (Nodo palabra, Umbral de similitud)
        Crear nuevo nodo
      sino Reutilizar nodo existente
    Crear puerta AND de nodos frase
    Para todos los nodos frase del Conjunto de frases
      Si No Existe nodo (Nodo frase, Umbral de similitud)
        Crear nuevo nodo
      sino Reutilizar nodo existente
    fin para
  Podar red de clasificación
fin procedimiento
```

Figura 5. Proceso de entrenamiento

A continuación se muestra mediante un ejemplo los pasos seguidos en el proceso de entrenamiento. El ejemplo se encuentra redactado en lengua inglesa y el diccionario estomatológico utilizado es el correspondiente a la Asociación Médica Americana.

Ejemplo. Proceso de entrenamiento.

Texto de entrada:

"Severe crowding. Very good oral hygiene. Has a severe crowding and right laterals are in cross bite. Several teeth are locked out. Please evaluate"

Especialidad:

Orthodontics

Filtro de frases:

"SEVERE CROWDING"
"VERY GOOD ORAL HYGIENE"
"HAS A SEVERE CROWDING"
"RIGHT LATERALS ARE IN CROSS BITE"
"SEVERAL TEETH ARE LOCKED OUT"
"PLEASE EVALUATE"

Filtro de palabras no relevantes:

"SEVERE CROWDING"
"GOOD ORAL HYGIENE"
"SEVERE CROWDING"
"RIGHT LATERALS CROSS BITE"
"SEVERAL TEETH LOCKED"
"PLEASE EVALUATE"

Filtro de diccionario odontológico:

"SEVERE CROWDING"
"GOOD ORAL HYGIENE"
"SEVERE CROWDING"
"RIGHT LATERALS CROSS BITE"
"SEVERAL TEETH LOCKED"

Filtro de conjuntos:

Nodos palabra:

"CROWDING", "ORAL", "HYGIENE", "CROSS", "BITE", "TEETH"

Nodos frase:

"SEVERE CROWDING", "GOOD ORAL", "ORAL HYGIENE", "LATERALS CROSS", "CROSS BITE", "SEVERAL TEETH", "TEETH LOCKED"

Finalmente, en la figura 6, se muestra como se incorpora el prediagnóstico estomatológico a la red de clasificación.

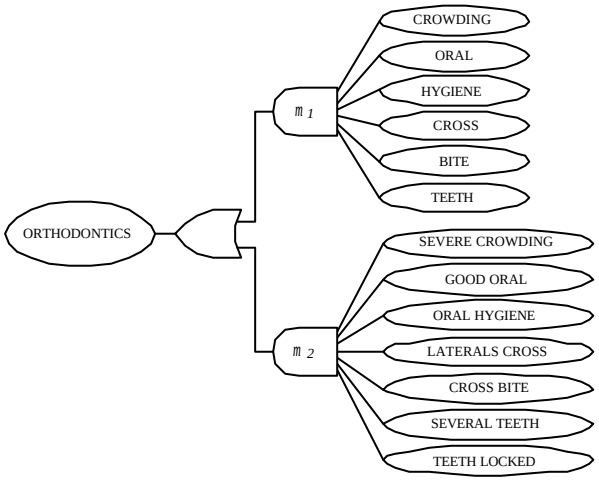

Figura 6. Incorporación del prediagnóstico a la red de clasificación

3.4. Análisis del clasificador

Como previamente se ha mencionado, el enfoque de la máquina inteligente se basa en la existencia de un conjunto de textos previamente clasificados en categorías. Generalmente, para realizar la evaluación de la efectividad de un clasificador se divide el conjunto de textos disponibles en dos subconjuntos, no necesariamente del mismo tamaño:

- Un subconjunto de entrenamiento. Este es el conjunto de textos de ejemplo de los cuales el clasificador extrae las características de cada una de las categorías.

- Un subconjunto de prueba. Este es el conjunto de textos utilizados para evaluar al clasificador.

Para llevar a cabo dicha evaluación, se comparan las decisiones tomadas por clasificador con las decisiones tomadas por el experto.

A continuación se muestran los resultados obtenidos (figura 8) con este clasificador sobre un total 551 prediagnósticos distribuidos en seis categorías (figura 7), de los cuales 270 han sido utilizados como casos de entrenamiento y el resto como casos de prueba.

Como muestra la figura 8, el total de prediagnosticos clasificados de forma correcta es superior al 70%, habiéndose obtenidos resultados del 80% mediante el ajuste de los umbrales de activación y similitud.

Especialidad	Cantidad de casos
Endodontics	89
Oral Pathology	93
Oral Surgery	86
Orthodontics	98
Periodontics	91
Prosthodontics	94
Total	551

Figura 7. Prediagnósticos estomatológicos disponibles

Especialidad	Porcentaje
Endodontics	77,43 %
Oral Pathology	80,91 %
Oral Surgery	65,77 %
Orthodontics	84,41 %
Periodontics	65,13 %
Prosthodontics	66,23 %
m_p de palabras	0,60
m_f de frases	0,50
Similitud	0,70
Media global	**73,31 %**

Figura 8.a. Resultados obtenidos (m_p = 0,60, m_f = 0,50, similitud = 0,70)

Especialidad	Porcentaje
Endodontics	81,67 %
Oral Pathology	84,62 %
Oral Surgery	69,97 %
Orthodontics	87,01 %
Periodontics	66,56 %
Prosthodontics	70,09 %
*m*de palabras	0,40
*m*de frases	0,30
Similitud	0,70
Media global	**76,65 %**

Figura 8.b. Resultados obtenidos (m_p = 0,40, m_f = 0,30, similitud = 0,70)

Especialidad	Porcentaje
Endodontics	87,55 %
Oral Pathology	92,30 %
Oral Surgery	70,27 %
Orthodontics	88,40 %
Periodontics	72,16 %
Prosthodontics	71,39 %
*m*de palabras	0,40
*m*de frases	0,30
Similitud	0.50
Media global	**80,34 %**

Figura 8.c. Resultados obtenidos (m_p = 0,40, m_f = 0,30, similitud = 0,50)

Conclusiones

En este trabajo se ha presentado un clasificador estomatológico por aprendizaje que basándose en los prediagnosticos establecidos por un profesional médico no especializado, y un diccionario de términos estomatológicos, es capaz de clasificar nuevos prediagnosticos en las especialidades correspondientes en un porcentaje superior al 70%.

Además, las pruebas empíricas han mostrado como la utilización conjunta de umbrales de activación y una función de similitud permite aumentar dicho porcentaje por encima del 80%, valores superiores a los obtenidos por los clasificadores tradicionales por reglas, los cuales no llegan a superar dicho porcentaje.

Referencias

[1] Cleverdon, C. 1984. Optimizing convenient online access to bibliographic databases. *Information Services and Use 4*, 1, 37-47.

[2] Hamill , K. Zamora, A. 1980. The use of titles for automatic document classification. *Journal of the American Society for Information Science 33*, 6, 396-402.

[3] Fabrizio, S. 1999. A Tutorial on Automated Text Categorisation. In *Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence.*

[4] Gonzalo Serrano. Dámaso Rodríguez. 1994. Negociación en las Organizaciones. Eudema Psicología. España.

[5] Guzmán A. A. 1998. Finding the main themes in a Spanish document. *Journal Expert Systems and Applications. 14. 139-148.*

[6] Hayes, P. J. Andersen P.M. Nirenburg G.I.B. Schamandt, L.M. 1990. TCS: a shell for a content-based text categorization. In *Proceedings of CAIA-90, 6th IEEE Conference on Artificial Intelligence Applications* (Santa Barbara, US, 1990), pp. 320-326.

[7] Maron, M. 1961. Automatic indexing: an experimental inquiry. *Journal of the Association for Computing Machinery 8*, 3, 404-417.

[8] Martínez-Trinidad J.F. Beltrán-Mártinez F. Ruiz-Shulcloper J. 2000. A tool to discover the main themes in a Spanish or English document. *Expert Systems with Applications 19. 319-327.*

[9] Mitchell, T. M. 1996. *Machine Learning*. McGraw Hill, New York US.

[10] Roubens M. Vincke P. 1985. "Preference Modelling". Lecture Notes in Economics and Mathematical Systems. Springer-Verlag. Berlin. Germany.