REVISTA COLOMBIANA DE REUMATOLOGÍA
VOL. 14 No. 4, Diciembre 2007, pp. 287-296

INVESTIGACIÓN ORIGINAL

El sistema inmunológico como herramienta diagnóstica de enfermedades reumatológicas

McCarthy Newball[1], Gerardo Quintana L.[2], Luis Fernando Niño I.[3]

Resumen

Introducción: los sistemas biológicos han sido objeto de muchas observaciones y recientemente se han convertido en modelos para ser emulados en diversos ambientes y ofrecer soluciones a problemas de la vida real. El sistema inmune es uno de los más representativos y en la actualidad constituye motivo de inspiración para la implementación de sistemas computacionales que respondan a diversas tareas, constituyendo los Sistemas Inmunes Artificiales.

Objetivo: este estudio busca desarrollar mecanismos computacionales inspirados en la inmunología para el diagnóstico de enfermedades reumatológicas que contribuyan en la educación y la toma de decisiones diagnósticas en reumatología. Se pretende obtener una herramienta computacional que, partiendo de un conjunto de historias clínicas como datos de entrenamiento, obtenga una efectividad en el diagnóstico comparable a los sistemas de clasificación de documentos actuales. El sistema está inspirado en la interacción entre los tejidos y los linfocitos B, y se apoya en conceptos de la teoría de la información para extraer relaciones entre términos. Los linfocitos B tendrán la función de discriminar la enfermedad reumatológica de un paciente con base en su historia clínica.

Materiales y métodos: se utilizó un conjunto de datos compuesto por 54 historias clínicas de 54 pacientes en reumatología, entre los cuales 21 padecían artritis reumatoide, y el resto padecían otras enfermedades reumatológicas. El conjunto de datos se dividió en dos grupos: pacientes con artritis reumatoide y pacientes sin artritis reumatoide. Se hizo un procesamiento manual de las historias clínicas para eliminar toda la información que no fuera relevante para el sistema en la tarea de diagnóstico. La efectividad del sistema fue comparada frente a otros tres algoritmos de clasificación de texto ampliamente utilizados en tareas de clasificación de documentos (ID3, BayesNet y PsoSVM).

Resultados: el sistema obtuvo resultados de efectividad prometedores en comparación con los demás algoritmos, con un promedio de 87,65% de efectividad en el diagnóstico. Sin embargo, debido a la limitación de datos, cabe la posibilidad de sesgo en los resultados. Se observó, como se había previsto, que los anticuerpos que representan la información en varios casos son redundantes. Adicionalmente, la información que representan no corresponde necesariamente a conocimiento médico, sino a reglas de clasificación de texto.

1 Facultad de Ingeniería de Sistemas y Computación, Universidad de los Andes.
2 Departamento de Medicina Interna, Facultad de Medicina. Universidad Nacional de Colombia y Servicio de Reumatología, Hospital Militar Central. Bogotá, Colombia.
3 Departamento de Ingeniería de Sistemas e Industrial. Universidad Nacional de Colombia.

Recibido para publicación: agosto 27/2007
Aceptado en forma revisada: noviembre 26/2007

Conclusiones: la teoría de la información, ayudada por la teoría del sistema inmunológico adaptativo y un mecanismo de señalización, muestra tener un potencial grande para la clasificación de historias clínicas. Debido a la posibilidad de sesgo observada en los resultados, será necesario realizar experimentos adicionales sobre un conjunto de historias clínicas más numeroso y más heterogéneo. Aunque entre los experimentos no se obtuvo anticuerpos que representaran claramente los conceptos, de tal manera que puedan ayudar a un profesional médico en el aprendizaje para la toma de decisiones, el trabajo a seguir consiste en adaptar técnicas de procesamiento de lenguaje natural (i.e., sintaxis y semántica), para así llegar a un sistema de obtención de conocimiento en lugar de un sistema de obtención de reglas de clasificación de texto.

Palabras clave: sistemas inmunes artificiales, enfermedades reumatológicas, diagnóstico.

Summary

Introduction: the biological systems have been observed and analyzed carefully and they have transformed into models to be emulated in many types of scenery and these offer solutions to problems of the real life, more recently. The immune system is one of the most representatives and at the moment is used for implementation of computational systems to respond to many tasks, constituting the Artificial Immune Systems.

Objective: in this work a computational method inspired by immunology for diagnosis of rheumatologic diseases is developed. The goal is to obtain a computational tool that, given a group of clinical histories as training data, performs rheumatologic diagnosis comparable to the current systems used in document classification. The proposed tool is expected to contribute in education and decision making in rheumatologic diagnosis. The proposed system is inspired by the interaction between tissues and B lymphocytes, and it relies on concepts of information theory to extract relationships among terms. The B lymphocytes will have the function of discriminating a patient's rheumatic diseases based on its clinical history.

Materials and methods: a dataset consisting of 54 medical records from 54 patients with rheumatologic diseases was used; 21 patients suffered rheumatoid arthritis, and the rest suffered other rheumatologic diseases. The dataset was divided into two groups: patients with and without rheumatoid arthritis. A manual process on the clinical histories was performed to eliminate the irrelevant information in the diagnosis task. The effectiveness of the system was compared to other three text classification algorithms widely used in document classification tasks, namely, ID3, BayesNet and PsoSVM.

Results: the proposed system obtained promising results in comparison with other algorithms, with an average of 87,65% effectiveness in the diagnosis. However, due to the limitation of the data, there is a possibility that the results are biased. It was observed, as expected that the antibodies that represent the information in several cases are redundant. Additionally, the information that it represents not necessarily corresponds to medical knowledge, but to rules of text classification.

Conclusions: information theory in conjunction with an adaptive immune system and a signaling mechanism showed great potential for the classification of medical records. Due to the possibility of a bias in the results, it will be necessary to carry out additional experiments on a larger and more heterogeneous group of medical records. From the experiments, antibodies that clearly represented concepts explaining rheumatoid arthritis were not obtained, which could help medical trainees in the learning process and medical doctors in decision making. Therefore, in future work, the task to continue consists on adapting natural language processing methods (i.e., syntax and semantics) to obtain a knowledge extraction system instead of a set of rules for text classification.

Key words: artificial immune systems, rheumatologic diseases, diagnosis.

Introducción

El diagnóstico de enfermedades reumatológicas (y de cualquier índole) no constituye una tarea simple, incluso para un médico experimentado. La complejidad de esta tarea se debe principalmente a factores como:

1. La complejidad intrínseca en la fisiopatología de las enfermedades auto-inmunes.
2. Experiencia insuficiente en el abordaje de pacientes con este tipo de enfermedades.
3. La existencia, relativamente alta, de cuadros clínicos difíciles de interpretar por parte del especialista.
4. La falta de estandarización de pruebas de laboratorio en inmunología.
5. La cantidad de historias clínicas sin un nivel adecuado de detalle.

El aprendizaje de máquina inspirado en la biología brinda una alternativa a los enfoques convencionales (e.g. Sistemas Expertos) para la solución de problemas de diagnóstico, al facilitar la tarea de búsqueda y reconocimiento de patrones. Diversos mecanismos computacionales han sido aplicados al problema de diagnóstico de enfermedades partiendo de diferentes fuentes de información como pruebas de laboratorio[1], radiografías e imágenes médicas[2], y mediciones en tiempo real[3,4]. Existen también trabajos realizados sobre historias clínicas[5-7]. Estas últimas son el elemento que contribuye en mayor medida en la interpretación de un diagnóstico[8]. Sin embargo, muchos de estos trabajos hacen uso de datos clínicos estructurados, adecuados para el procesamiento y la aplicación directa de técnicas de minería de datos, en contraste con trabajos que se basan en el procesamiento de textos en lenguaje natural[6,7].

La práctica médica a nivel mundial se caracteriza por el uso de narrativa (lenguaje natural) en reportes médicos e historias clínicas, la forma más natural y flexible de transmitir (y de recuperar posteriormente) la información adquirida en un proceso médico. La narrativa consta de una secuencia de letras, números, signos de puntuación, que a pesar de la enseñanza de la cátedra de Semiología en escuelas de Medicina, falta rigurosidad en la escritura de las historias clínicas. Aunque existen mecanismos para representar la narrativa por medio de datos estructurados[9,10], estos no son los más adecuados para la representación de conceptos y conocimiento, pasando por alto nociones como la localización espacial y temporal de ciertos objetos, sus cantidades, y las relaciones entre los mismos. Existen métodos que tienen en cuenta el contexto de una palabra con el fin de encontrar correlaciones entre palabras[11]. Estos no son suficientes para solucionar problemas semánticos (e.g. no es lo mismo la sentencia "presenta sinovitis" que la sentencia "no tiene sinovitis"), la sinonimia (varias palabras pueden hacer referencia al mismo concepto: gastrointestinales también hace referencia a intestinales) y la homonimia (una palabra puede hacer referencia a distintos conceptos: cara puede referirse a la cara anterior de una pierna o a la cara de la cabeza).

Además de no solucionar adecuadamente estos problemas lingüísticos expuestos, los métodos referenciados tienen tan solo una capacidad limitada para la representación de conocimiento, a partir de narrativa. Los sistemas de reglas y los árboles de decisión son ideales para representar el conocimiento inferido desde datos estructurados. Sin embargo, en la presencia de datos sin estructura, el conocimiento que representan no es lo suficientemente intuitivo. Los métodos probabilísticos (e.g. Naive Bayes, BayesNet), así como los métodos estadísticos (e.g. SVM), aunque más efectivos, son aún menos intuitivos.

Sistemas inmunológicos artificiales

En años relativamente recientes y a través de la interacción entre diferentes disciplinas, biólogos, matemáticos e ingenieros, entre otros, han buscado destacar y aprovechar los comportamientos biológicos puros como modelos para dar solución o agilizar tareas y/o problemas de la vida real en cualquier escenario que son tediosos, altamente complejos por el número de variables que se involucran en un toma de decisiones y en general, procesos que requieren gran entrenamiento, demanda de tiempo y a veces conocimientos avanzados. Numerosos estudios en biología teórica[12,13] han revelado la capacidad cognitiva del sistema inmunológico natural. Varios trabajos[14-16] han tomado como inspiración diversas características del sistema inmunológico como lo son sus facilidades de adaptación, de memoria asociativa, de generalización y de reconocimiento de patrones, para construir sistemas computacionales. A estos últimos se les conoce como sistemas inmunológicos artificiales.

Diferentes aspectos, principalmente del sistema inmunológico adaptativo, han sido aplicados a una

variedad de problemas, desde la detección de virus, pasando por el monitoreo de redes, hasta el entrenamiento de robots. Más recientemente, con base en el trabajo realizado por Matzinger[17-19] sobre la teoría del peligro, se han aplicado técnicas de detección de peligro por parte del sistema inmunológico innato a la clasificación de correo electrónico, y la detección de tráfico irregular en redes[15,20].

Para este trabajo, es importante destacar los trabajos realizados sobre clasificación de documentos. Entre estos encontramos el sistema de clasificación AIRS[21], que hace uso del concepto de Bola de Reconocimiento Artificial (ARB, por sus siglas originales en inglés), que se refiere al volumen dentro del espacio de forma (shape-space) que un anticuerpo es capaz de reconocer. Este sistema hace uso de técnicas inspiradas en el sistema inmunológico humano como la selección clonal y la hiper-mutación somática.

También se encuentra el sistema para la detección de spam AISEC[15], que hace uso de anticuerpos y selección clonal sobre estos para obtener un clasificador de correos electrónicos.

Tanto AIRS como AISEC convierten los documentos de entrada en un vector de palabras, sin tener en cuenta su significado, únicamente la ocurrencia o no de una palabra dentro del documento. Esta representación es poco adecuada en el momento de obtener conocimiento, tal y como se discutirá en la última parte de este artículo. Este estudio aplica ideas inspiradas tanto en la teoría del peligro, basándose en el trabajo realizado[20], así como del sistema inmunológico adaptativo, más específicamente de las células B.

En este artículo se introduce un nuevo sistema para la clasificación de historias clínicas de pacientes con casos reumatológicos, inspirado en el funcionamiento de la respuesta inmunológica, apoyado en conceptos de la teoría de la información.

Materiales y métodos

Teoría de la información

Para este trabajo se hace uso de la noción de entropía de la teoría de la información. En este contexto la entropía puede verse como la cantidad de información que aporta una palabra en la clasificación de un documento determinado.

Si se define $P_z(w)$ como la probabilidad de ocurrencia de la palabra *w* dentro de la categoría o clase *z* (que corresponde a una enfermedad) entonces se puede definir la entropía de una palabra *w* dentro de la categoría *z* como

$$E_z(w) = -P_z(w) \cdot \log(P_z(w)) \quad (1)$$

La entropía total de la categoría *z* está dada por

$$E_z = \sum_{w \in z} E_z(w) \quad (2)$$

La entropía de la palabra *w* dentro de todo el conjunto de documentos está dada por

$$E_{total}(w) = -P_{total}(w) \cdot \log(P_{total}(w)) \quad (3)$$

La entropía de todo el conjunto de documentos está dada por

$$E_{total} = \sum_{w} E_{total}(w) \quad (4)$$

La entropía condicional de la palabra *w* dentro de la categoría *z* está dada por

$$E_{z|w} = -\sum_{v \in z} P_{total}(v \mid w) \cdot \log(P_{total}(v \mid w)) \quad (5)$$

La entropía total de la palabra *w* dentro del conjunto de documentos está dada por

$$E_{total|w} = -\sum_{v} P_{total}(v \mid w) \cdot \log(P_{total}(v \mid w)) \quad (6)$$

La ganancia de información de la palabra *w* dentro de la categoría *z* está dada por

$$GI_x(w) = E_z - E_{z|w} \quad (7)$$

La ganancia de información de la palabra *w* dentro de la totalidad de los documentos está dada por

$$GI_{total}(w) = E_{total} - E_{total|w} \quad (8)$$

Un sistema inmunológico artificial para la clasificación de historias clínicas en reumatología

Inicialmente se propone un sistema capaz de aprender reglas de clasificación a partir de un conjunto de historias clínicas, que constan de texto narrado en español. Las reglas de clasificación por lo general no son aptas para representar el conocimiento, ya que se dejan de lado asociaciones de mucho más alto nivel como las correlaciones de incidencias de dolor en ciertas partes del cuerpo, e incluso las diferencias en exámenes de laboratorio. Para trabajo futuro se deja la posibilidad de aprender conocimiento médico a partir

de los documentos. Este primer estudio se enfoca en la efectividad de un sistema inspirado en la inmunología en el reconocimiento de patrones dentro de historias clínicas reumatológicas. Cada historia clínica corresponde a una enfermedad (categoría) reumatológica determinada. Por simplicidad solo se tendrán en cuenta dos categorías: artritis reumatoide y no artritis reumatoide.

El sistema propuesto consta de dos capas: una capa de tejido cuya función es la de determinar las palabras más relevantes para una enfermedad determinada, y una capa de linfocitos B donde se determina qué combinaciones de palabras identifican a la enfermedad. Adicionalmente, se cuenta con un centro germinal encargado de la producción de nuevos linfocitos B. La capa de tejido recoge aquellos términos que son más relevantes para una enfermedad, y los envía a la capa de anticuerpos para así iniciar una respuesta inmunológica donde se seleccionarán aquellos anticuerpos con la mayor afinidad contra los documentos que se refieran a la misma enfermedad.

Representación de los documentos

El primer paso consiste en convertir un documento de historia clínica a una representación adecuada para el sistema. La manera convencional de representar un documento en los problemas de clasificación es asociar una dimensión con una palabra dentro del conjunto de datos. La ocurrencia de una palabra en el texto se representa con el número 1, y su ausencia con el número 0. Esto se conoce como un vector de características. Para este sistema se hará uso de una representación similar a la utilizada por Secker y colaboradores[15], debido a que no se necesita conocer las palabras existentes en un conjunto de datos sino únicamente las que ocurren dentro del documento en cuestión. El siguiente ejemplo muestra como se representaría un documento dentro del sistema:

Motivo de Consulta:

El paciente asiste por dolores intensos en manos y tobillos.

Enfermedad Actual:

El paciente presenta dolores en mano derecha, rodillas y tobillo derecho.

La representación dentro del sistema es la siguiente:

{<MC>{el, pacient, asist, por, dolor, intens, en, man, y,tobill},<EA>{el, pacient, present, dolor, en, man, derech, rodill, y, tobill}}

Nótese que el documento, al estar representado como un conjunto de palabras, no tiene en cuenta las palabras que se repiten. De igual manera, las secciones de motivo de consulta y de enfermedad actual se toman como independientes; así que el paciente, aunque ocurre dos veces dentro del documento, la agrupación de términos se realiza a nivel de secciones y no de la totalidad del documento. Nótese además cómo se suprimen ciertas letras de algunas palabras. Esto se debe a la aplicación de un algoritmo de lematización, para así tener en cuenta únicamente la raíz de la palabra, y que términos como derecho y derecha se tomen como palabras iguales, aún cuando se escriben de manera distinta.

El algoritmo

Esta representación del documento puede ser vista como un patógeno que entra al organismo. Este patógeno puede contener uno o más antígenos, cada uno compuesto por una serie de moléculas, donde cada molécula representa un área de conocimiento presente en la historia clínica (e.g. enfermedad actual, examen físico, antecedentes familiares, etc.). Cada molécula a su vez está compuesta por átomos, donde cada átomo constituye una palabra lematizada dentro de una sección de la historia clínica. Una vez el patógeno ingresa al sistema, primero deberá pasar por la capa de tejido. De acuerdo a la teoría del peligro, el tejido es el responsable de iniciar la respuesta inmune, al arrojar señales de peligro al sistema inmunológico. En el sistema propuesto, la capa de tejido sirve como una capa de pre-procesamiento, que luego alerta a la capa de anticuerpos sobre la presencia de los términos más relevantes dentro de los documentos. Cada vez que se presenta un patógeno al sistema, en la capa de tejido se determina qué términos del patógeno son los más relevantes, y aquellos que no aportan significativamente para distinguir entre la existencia o no de la enfermedad son descartados. Para obtener los términos más relevantes de una enfermedad, se hace uso de una técnica de la teoría de información denominada Ganancia de Información, cuyas fórmulas están dadas por (7) y (8).

Se buscan todas aquellas palabras que tengan una alta ganancia de información para la totalidad de los documentos, pero una baja ganancia de información para la categoría, es decir, las palabras que sean comunes para una categoría, pero no en la totalidad de los documentos. Además de los términos relevantes, es importante tener en cuenta también todos los demás términos que co-curren frecuentemente con los primeros. Esto para garantizar que en una frase como "no presenta dolor", en la cual dolor es un término relevante, la palabra "no" también sea tenida en cuenta, ya que sería completamente distinto a únicamente "dolor". Para determinar qué términos co-ocurren frecuentemente con los términos relevantes, es necesario obtener todas aquellas palabras que tengan la mayor información mutua con los términos relevantes. La información mutua se calcula por medio de la siguiente ecuación:

$$I(x;y) = E_z(w) - E_z(x|y) \quad (9)$$

Una vez el tejido ha determinado todos los términos relevantes para clasificar un documento, el paso a seguir consiste en enviar una señal de peligro por cada uno de los términos que hayan sido considerados relevantes. Esta señal tiene asociada una fuerza, la cual está dada por la ganancia de Información del término. El tejido consta de células, y cada célula está asociada a un término relevante. Cada célula arroja una señal que contiene información sobre la fortaleza de la misma, y el término al que hace referencia la señal. Estas señales llegan al centro germinal y a los anticuerpos. Cuando el centro germinal recibe las señales agrega los términos que sean nuevos a su librería de genes, para la generación de futuros anticuerpos. Las señales que llegan a los anticuerpos son tenidas en cuenta posteriormente para calcular la estimulación de un anticuerpo frente a un anticuerpo. La fortaleza de las señales está dada por la ganancia de información del término que la señal representa.

La fortaleza de las señales se encuentra normalizada, así que su valor se encuentra en el rango [0, 1]. Estas señales únicamente son enviadas cada vez que se haga necesario reconstruir el tejido. Esto sucede después de analizar un cierto número de documentos, ya que la reconstrucción del tejido es una tarea costosa. La reconstrucción del tejido consiste en la creación de nuevas células de tejido y la eliminación de células inservibles. El resto de células permanecen inalteradas.

Cada patógeno presentado al sistema es pasado luego la capa de linfocitos B, donde es presentado a cada uno de los linfocitos existentes y esta presentación resulta en la estimulación de los mismos. Luego de recibir las señales provenientes del tejido, se induce una respuesta inmunológica adaptativa donde los linfocitos con el mayor promedio de estimulación, teniendo en cuenta además la estimulación producida por las señales del tejido, serán seleccionados para su clonación y posterior mutación. El número de clones producido por un anticuerpo está dado por la siguiente fórmula:

$$NC_a = l \cdot \frac{S_n}{2}, \quad (10)$$

donde l es una constante que determina el número máximo de clones que puede tener un anticuerpo, y Sn es el valor de la última estimulación del anticuerpo.

La capa de anticuerpos se divide en dos conjuntos: un conjunto de células inexpertas y un conjunto de células de memoria. Estos nuevos anticuerpos harán parte del conjunto de células inexpertas, ya que aún no han sido estimuladas. También se eliminará un porcentaje de los anticuerpos con menor estimulación del conjunto de células de memoria. Posteriormente, el centro germinal generará un número aleatorio de anticuerpos nuevos que será introducido al conjunto de células inexpertas. Esto último contribuirá a que el sistema cubra la mayor cantidad de términos posible, ya que en cada respuesta inmunológica pueden haberse introducido nuevos términos al centro germinal que pueden ser importantes en la clasificación. Por último, aquellos anticuerpos del conjunto de células inexpertas, que tengan una estimulación promedio superior a la estimulación de alguna de las células pertenecientes al conjunto de células de memoria, pasará a ser parte del conjunto de células de memoria.

Algunas definiciones necesarias para el algoritmo se muestran en la Tabla 1. El algoritmo de entrenamiento del sistema se resume en la Figura 1 y la Figura 2. Una vez entrenado el sistema, para determinar la enfermedad a la que corresponde una historia clínica, esta es convertida a un patógeno y es presentada directamente a cada uno de los linfocitos de la capa de linfocitos. La estimulación de los

Algoritmo 1

```
for all hc ∈ HC do
  patogeno ← ConvertirAPatogeno(hc)
  actualizar estadísticas términos patógeno
  redirigir patogeno a capa de anticuerpos
  if pa + 1 > pp then
    GIpatogeno ← ∅
    for all termino ∈ Tpatogeno do
      GIpatogeno ← GIpatogeno ∪ {GIx(termino)}
    end for
    GIs ← maxs(GI)
    MIpatogeno ← ∅
    for all terminox, terminoy ∈ GIs do
      MIpatogeno ← MIpatogeno ∪ {I(terminox, terminoy)}
    end for
    MIs ← maxs(MI)
    TS ← GIs ∪MIs
    Ct ← CrearCelulasTejido(TS)
    S ← ObtenerSeñales (Ct)
    InducirRespuestaInmune(S)
  end if
end for
```

Figura 1. *Algoritmo de entrenamiento*

linfocitos de cada categoría es promediada, y aquella categoría con el promedio de estimulación más alto será a la que corresponda la historia clínica.

Este algoritmo cuenta con una característica poco deseable. Debido a que no se está obligando a que los linfocitos difieran el uno del otro, es muy posible que se formen linfocitos muy similares que reconocen términos muy similares, lo que puede representar un desperdicio importante de recursos computacionales. Esto puede llevar también a que en el momento de la clasificación de las historias clínicas, debido a la similitud de los linfocitos, si uno de estos es altamente estimulado, los demás también lo serán, lo que podría favorecer a la discriminación entre una enfermedad y otra. Estas limitaciones esperan ser solucionadas en la siguiente fase de este estudio.

Algoritmo 2 *InducirRespuestaInmune*(*S*)

```
if CM = ∅ then
  CM = CI
  CI = ∅
end if
for all ab ∈ CI do
  if ∃abm ∈ CM: abm.estimulacion < ab.estimulacion then
    CM ← CM ∪ {ab}
    CI ← CI - ab
  end if
end for
CT ← ∅
for all ab ∈ CM do
  ab.estimulacion ← (ab.estimulacion+ab.stim(S))/2
  if ab.estimulacion > u then
    CT ← CT ∪ {ab}
  end if
end for
for all ab ∈ CT do
  NCab ← CalcularNumeroClones(ab)
  CC ← GenerarClones(ab, NCab, mut)
  CI ← CI ∪ CC
end for
mor ← ActualizarTasaDeMortalidad()
NM ← |CM| · Td
CM ← CM - {los peores NM anticuerpos de CM }
CI ← CI [ {número aleatorio n de nuevos anticuerpos: 0 ≤ n ≤
MAXn }
```

Figura 2. *Algoritmo 2: Inducir Respuesta Inmunes*

Resultados

Durante la primera parte de este proyecto se realizaron experimentos con un número diverso de clasificadores, los cuales han sido ampliamente estudiados y cuya efectividad ha sido medida en distintos dominios de aplicación. Los métodos elegidos para la experimentación fueron: ID3 (algoritmo de árboles de decisión), PsoSVM (implementación de algoritmo

Tabla 1. *Parámetros del algoritmo de entrenamiento.*

Parámetro	Descripción
HC	Conjunto de historias clínicas
pa	Número de patógenos analizados por el tejido después del último envío de señales
pp	Número de patógenos máximo que el tejido debe analizar antes de enviar señales
Tx	Conjunto de términos de un patógeno x
S	Señales enviadas por el tejido
maxs(GI)	Función que devuelve los términos con mayor ganancia de información dentro de GI
maxs(MI)	Función que devuelve los términos con mayor información mutua dentro de MI
stim(S)	Función que calcula la estimulación de un anticuerpo frente a las señales S
u	Umbral de estimulación para que un anticuerpo sea clonado
mor	Tasa de mortalidad
mut	Tasa de mutación
CM	Conjunto de anticuerpos de memoria
CI	Conjunto de anticuerpos inexpertos
MAXn	Máximo número de anticuerpos nuevos a ingresar en CI

SVM utilizando Particle Swarm Optimization) y BayesNet (implementación del algoritmo de redes bayesianas). Un conjunto de 54 historias clínicas (21 artritis reumatoide, 33 no artritis reumatoide) fue utilizado. Inicialmente se realizó un conjunto de pruebas donde se aplicó lematización y selección de los términos más relevantes por medio de una técnica denominada *Information Gain Ratio*.

Las primeras pruebas se realizaron sobre las historias clínicas sin curar, es decir, se tuvo en cuenta todo el contenido original de las mismas. Los resultados del experimento se muestran en la Tabla 2. Este documento fue previamente lematizado (stemming), y se le aplicó un algoritmo de reducción de dimensionalidad *Information Gain Ratio* para escoger los atributos más relevantes. En la Tabla 2, los números entre paréntesis corresponden al número de ejemplos de positivos y negativos, respectivamente. Aunque los resultados a primera vista son de una efectividad elevada, la mayoría de documentos contaba con palabras que podrían discriminar fácilmente la artritis reumatoidea de las demás enfermedades, como, por ejemplo, las palabras artritis reumatoide y la abreviación ar discriminan de manera trivial una historia clínica correspondiente a artritis reumatoide. Esto se pudo comprobar en el árbol de decisión resultante del algoritmo ID3.

Tabla 2. *Desempeño de clasificadores sobre el conjunto de historias clínicas adaptadas a vectores de atributos, donde cada atributo es una palabra del documento.*

Algoritmo	Aciertos	Fallos	Efectividad
ID3	51(20-31)	3(1-2)	94,44%
BayesNet	50(18-32)	4(3-1)	92,59%
PsoSVM	44(15-29)	10(6-4)	81,48%

Para evitar este tipo de sesgos, se organizó un segundo conjunto de pruebas en el cual se removió de las historias clínicas toda la información que pudiera determinar trivialmente la enfermedad de un paciente. Una vez realizada esta tarea, se procedió a realizar el mismo experimento. Los resultados se muestran en la Tabla 3. Puede observarse una reducción notable en el desempeño de cada uno de los algoritmos. El que menos redujo su desempeño fue PsoSVM, debido a que su funcionamiento está dado por conjuntos de atributos, en lugar de ciertos atributos específicos. En cambio BayesNet e ID3 dependen de la efectividad de los atributos para discriminar entre categorías. Estos resultados son un reflejo de lo disperso que es el lenguaje utilizado dentro de las historias clínicas. Cada uno de los términos ocurre un número limitado de veces dentro

Tabla 3. *Desempeño de clasificadores sobre el conjunto de historias clínicas adaptadas a vectores de atributos, donde cada atributo es una palabra del documento.*

Algoritmo	Aciertos	Fallos	Efectividad
ID3	40(15-25)	14(6-8)	74,07%
BayesNet	32(0-32)	22(21-1)	59,26%
PsoSVM	34(8-26)	20(13-7)	62,96%

de las historias clínicas. Términos, que pueden ser muy relevantes, en 54 historias clínicas ocurren tan solo entre dos y tres veces.

Los resultados corresponden a los experimentos realizados sobre las historias clínicas revisadas. Este documento fue previamente lematizado (stemming), y se le aplicó un algoritmo de reducción de dimensionalidad *Information Gain Ratio* para escoger los atributos más relevantes. Los números entre paréntesis corresponden al número de ejemplos de positivos y negativos, respectivamente.

Estos experimentos sirvieron como referencia para determinar la efectividad del algoritmo presentado en este artículo. En la Tabla 4 se muestran los resultados de los experimentos realizados sobre el algoritmo. Estos resultados corresponden a los experimentos realizados sobre las historias clínicas revisadas. Cada documento fue previamente lematizado (stemming). Puede observarse una efectividad superior a cada uno de los algoritmos de clasificación utilizados para los experimentos anteriores.

En la Figura 3 se muestran extractos de dos de los anticuerpos mayormente estimulados durante una de las ejecuciones del algoritmo.

Tabla 4. *Desempeño del clasificador presentado en este modelo.*

Experimento	Aciertos	Fallos	Efectividad
1	51(20-31)	3(1-2)	94,44%
2	45(19-26)	9(2-7)	83,33%
3	46(19-27)	8(2-6)	85,18%
Promedio	47,33	6,66	87,65%

Anticuerpo 1 (artritis reumatoide)	
cabeza-cuello	hipocrom oj mucos la men bi humed rotacion ros disminución
Osteomuscular	en art inclinacion glut trocant schub shoeb lumbosacr flexoextension articular mayor superior cm palpacion torac intermaleol lateral entesopati preserv descrit
Ea	flexoextensi articular maner progresivament an talon men patologi sacr postraci fasci prim entesopati ultim gl artralgi han much tip espinal igual aument persistenci
rxs	ac pic fiebr talon poc matinal piolaquiuri sacroil per inflamtori piens inferior nivel plant normal hu posterior hi esquelet bilateral parcial mucos dificult raynaud miembr inguinal dpn intestinal uveitis lumbr

Anticuerpo 2 (artritis reumatoide)	
cabeza-cuello	alteracion hipocrom faci disminución mucos oj
Osteomuscular	en articular cm schoob glut trocant schub lumbosacr flexoextension art mayor superior palpacion torac
Ea	articular nieg igual progresivament stic talon sistom patologi sacr prim gl remision y entesopati par fasci artralgi han much tip espinal moviliz agudiz persistenci diciembr entesitis articulaci anquilosis inflamatori
Rxs	sintomatologi bilateral mucos lumbalgi fenomen mejotri sacroil disminuci fiebr adormec hu talon hi cronico- articulaci dificult piens lumbr intestinal uveitis matinal

Figura 3. *Ejemplos de anticuerpos resultantes de un proceso de entrenamiento.*

Aunque la efectividad del sistema es aparentemente elevada, en los anticuerpos obtenidos se pudo observar que los más estimulados compartían casi los mismos términos, pero en diferente orden, y además abarcaban un gran número de términos. Esto implica que los anticuerpos intentan abarcar el mayor número de términos posible, desperdiciando espacio valioso, y haciendo muy difusos los conceptos, ya que no se sabe qué términos corresponden juntos. Por otra parte, debido a lo reducido que resultó ser el conjunto de datos inicial, no se puede llegar a resultados concluyentes sobre la efectividad del algoritmo. Sin embargo, los resultados son alentadores. También se pudo observar que aunque algunos términos contribuyen en la discriminación entre una enfermedad u otra, estos no son suficientes para que un médico pueda determinar las reglas tenidas en cuenta para el diagnóstico de las enfermedades.

Para un trabajo futuro se propone la introducción de otros elementos de procesamiento de lenguaje

natural para la obtención de conocimiento a partir de las historias clínicas.

Discusión

El sistema inmunológico constituye una fuente de inspiración inigualable para problemas de reconocimiento de patrones, ya sea que provenga desde el sistema inmunológico innato, el sistema inmunológico adaptativo o incluso la interacción de otras células del organismo con el sistema inmunológico. Para este trabajo se tomó inspiración de la interacción entre los tejidos del cuerpo y el sistema inmunológico adaptativo.

Se mostró que con la aplicación de teoría de la información es posible determinar los grupos de palabras más relevantes dentro de una enfermedad. Debido a la posibilidad de sesgo observada en los resultados, será necesaria la repetición de los experimentos sobre un conjunto de historias clínicas más numeroso y más heterogéneo. Aunque entre los experimentos no se obtuvo anticuerpos que representasen claramente los conceptos, de tal manera que puedan ayudar a un profesional médico en el aprendizaje para la toma de decisiones, el trabajo a seguir consiste en adaptar técnicas de procesamiento de lenguaje natural (i.e., sintaxis y semántica), para así llegar a un sistema de obtención de conocimiento en lugar de un sistema de obtención de reglas de clasificación de texto. La teoría de la información con la ayuda de un sistema inmunológico adaptativo y un mecanismo de señalización mostró tener un potencial grande para la clasificación de historias clínicas.

Declaración de conflictos

Declaramos no tener ningún conflicto de interés en la realización de este estudio de investigación.

Referencias

1. Sandria JC. Sistema experto para el diagnóstico diferencial de enfermedades dermatológicas usando una red neuronal. In Memorias del 1er Simposio Científico Internacional. "Aplicaciones de la Matemática y la Cibernética a la Medicina". Centro de Cibernética Aplicada a la Medicina, 1999.
2. Sahan S, Polat K, Kodaz H, and Güne S. A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. Computers in Biology and Medicine 2007.
3. Imhoff M, Bauer M, Gather U, and Lohlein D. Time series analysis in intensive care medicine. Applied Cardiopulmonary Pathophysiology 1997; 6: 203-281.
4. Morik K, Brockhausen P, and Joachims T. Combining statistical learning with a knowledge based approach - a case study in intensive care monitoring. In Proceedings of 16th International Conference on Machine Learning, Morgan Kauffman, 1999; 268-277.
5. Ichise R and Numao M. Learning first-order rules to handle medical data. NII Journal 2001; 3(2): 9-14.
6. McCowan I, Moore D, and Fry MJ. Automated cancer stage classification from free-text histology reports. In Proceedings of the Australian Health Informatics Conference (HIC), 2006.
7. Taira RK and Soderland SG. A statistical natural language processor for medical reports. In AMIA'99 Annual Symposyium, Nov. 1999.
8. Pinckney RE and Pinckney C. The Sciences 1989.
9. Shawe-Taylor J and Cristianini N. An Introduction to Support Vector Machines and other Kernel-Based Learning Methods. Cambridge University Press, 2000.
10. Witten IH and Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kauffman, 2nd edition, 2005.
11. Cohen WW. Fast, effective rule induction. In Proceedings of the 12th International Conference on Machine Learning, Morgan Kauffman, 1995; 115-123.
12. Kowal C, DeGiorgio LA, Nakaoka T, Hetherington H, Huerta PT, Diamond B, and Volpe BT. Cognition and immunity: Antibody impairs memory. Immunity 2004; 21(2): 179-188.
13. Wallace R and Wallace RG. Immune cognition and vaccine strategy: beyond genomics. Microbes and Infection 2002; 4(4): 521-527.
14. Cayzer S and Aickelin U. A recommender system based on the immune network. Technical Report 1, Hewlett Packard, Mar. 13 2002.
15. Secker A, Freitas AA and Timmis J. Aisec: an artificial immune system for e-mail classification. In Proceedings of the Congress on Evolutionary Computation, Canberra, Australia, 2003; 131-139.
16. Toma N, Endo S, and Yamada K. An immune co-evolutionary algorithm for n-th agent's traveling salesman problem. In Proceedings 2003 IEEE International Symposium on Computational Intelligence in Robotics and Automation, 2003; 3: 1503-1508.
17. Matzinger P. The danger model in its historical context. Scand J Immunol 2001; 54: 4-9.
18. Matzinger P. The danger model: A renewed sense of self. Science 2002; 296: 301-305.
19. Matzinger P. The real function of the immune system. http://cmmg.biosci.wayne.edu /asg/polly.html, Abr. 6 2004.
20. Bentley PJ, Greensmith J, and Ujjin S. Two ways to grow tissue for artificial immune systems. In Lecture Notes in Computer Science, volume 3627 of Artificial Immune Systems: 4th International Conference, ICARIS 2005, pages 139-152, Banff, Alberta, Canada, Ago. 14-17 2005. Springer.
21. Watkins A, Timmis J and Boggess L. Artificial immune recognition system (AIRS): An immune-inspired supervised machine learning algorithm. Genetic Programming and Evolvable Machines 2004; 5(3): 291-317.