RAZONADOR BASADO EN CASOS PARA LA INFERENCIA DE HIPÓTESIS DIAGNÓSTICAS, BASADO EN LA TEORÍA DE LOS CONJUNTOS APROXIMADOS

CASE-BASED REASONING FOR DIAGNOSTIC HYPOTHESES INFERENCE BASED ON ROUGH SETS THEORY

Katterine Nodarse[1], Ivett E. Fuentes[1], María M. García[1], Leticia Arco[1], Mabel M. Herrera[2], Rolando de la C. Fuentes[2]

1- Universidad Central "Marta Abreu" de Las Villas, {knmorales, ivett}@uclv.cu, {leticiaa, mmgarcia}@uclv.edu.cu

2- Universidad de Ciencias Médicas de Villa Clara Dr. "Serafín Ruíz de Zarate Ruiz", {rolandocfm, mabelmhg}@infomed.sld.cu

RESUMEN: *La Historia Clínica Electrónica (HCE) constituye el documento esencial de los sistemas de información hospitalarios (SIH). La proliferación de la HCE como consecuencia del incremento de los SIH en las instituciones hospitalarias representa enormes ventajas, debido a la riqueza de información y experiencias que en ella se recogen. Garantizar su uso productivo demanda de nuevos métodos capaces de procesar la información y extraer conocimiento en un tiempo razonable, que permita guiar a los especialistas en el proceso de admisión de nuevos pacientes. En este trabajo se presenta una propuesta novedosa que contribuye al diagnóstico clínico basado en el análisis de información textual y estructurada. La combinación del agrupamiento como base para la organización de la base de casos y la propuesta Rough Text, para enfrentar diversas tareas de la minería de textos basada en la Teoría de Conjuntos Aproximados (RST), favorece la inferencia de hipótesis diagnósticas a la llegada un nuevo paciente. El sistema CDARS implementado contribuye a asistir el proceso de toma de decisiones en la práctica del método clínico.*

Palabras Clave: Razonamiento Basado en Casos; Agrupamiento; Toma de Decisiones; HCE; Teoría de Conjuntos Aproximados.

ABSTRACT: *The Electronical Medical Record (EMC) is an essential document of the hospital information systems (HIS). The proliferation of the EMC as a result of the HIS increasing in hospitals has numerous advantages because of the abundance of information and experiences we can find in them. New methods able to process information and extract knowledge at a reasonable time are needed to help specialists in the admission process of new patients. In this paper, a new proposal that contributes to the clinical diagnosis based on the analysis of textual and structured information is presented. The combination of clustering as the foundation for the case base organization, and the proposal Rough Text to deal with several tasks of text mining based on Rough Sets Theory (RST), contributes to infer diagnosis hypothesis when a new patient arrives. The system CDARS, a result of the model implementation, contributes to assist the decision making process while practicing the clinical method.*

KeyWords: Case Based Reasoning; Clustering; Decision Making; EMR; Rough Sets Theory

1. INTRODUCCIÓN

Las enormes cantidades de datos generados cada día en las instituciones, imponen un gran reto para lograr una gestión de los datos que los transforme en información inteligente y viabilice el proceso de toma de decisiones. Por lo que, sin importar cuánto se han desarrollado los sistemas automatizados para el manejo de información, se reconoce que la complejidad mayor reside en tomar las mejores decisiones y extraer conocimiento relevante desde la información disponible; al ser la información el medio y el entorno que permite el conocimiento[1].

En el contexto de la actividad hospitalarias, los SIH

constituyen una fuente de información acerca de la historia de salud de los pacientes[2]. Estos sistemas ofrecen ventajas para los profesionales de la salud, los pacientes, y el estado; al propiciar un marco para el desarrollo de técnicas que permitan el uso productivo de la información, continuamente creciendo a partir del uso extendido de las Historias Clínicas en formato Electrónico (HCE). Por lo cual es un hecho que, gestionar el conocimiento a partir de la información almacenada es fundamental en la práctica clínica [3].

Entre las técnicas para la solución de problemas de toma de decisiones desde la Inteligencia Artificial (IA) se distinguen diversos enfoques[4]. Específicamente, el Razonamiento Basado en Casos (RBC) imita la forma de razonamiento que emplean los especialistas, basado en experiencias anteriores, a partir de la organización, generalización y reutilización para resolver problemas. Los sistemas basados en casos se han empleado en diferentes tareas de diagnóstico. En IA, esta forma de resolver problemas ha sido modelada por razonamiento por analogía [5]. Estos sistemas tienen dos componentes principales, una base de casos y el RBC. La base de casos contiene la descripción de los problemas y el solucionador tiene dos módulos, el recuperador y el adaptador. El recuperador tiene la función de buscar y recuperar el o los casos más similares de la base de casos usando una función de similitud o disimilitud. Cuando se usa RBC es importante determinar como se organiza la base de conocimientos para acceder y recuperar los casos más semejantes, lo cual puede ser un proceso complejo[6].

Debido a que, en el contexto de la gestión hospitalaria, las opiniones que refieren los expertos sobre el tratamiento y la conducta correcta de un nuevo paciente son recogidas en la HCE, ella constituyen el elemento esencial sobre el cual aprende el modelo para el razonamiento basado en casos que se presenta [6] y que se implementa en esta propuesta a partir de los elementos de la teoría RST [7]que propone *Rough Texts[8]* para asistir diversas tareas de la minería de texto.

Según la propuesta *Rough Texts,* se utiliza una relación de similitud *R*' debido a que dos documentos del universo (*U*) pueden ser similares, pero improbablemente iguales. Para determinar cuán similares son los documentos se utiliza una de las medidas de similitud propuestas en la literatura para comparar documentos.

En este, trabajo se propone una nueva alternativa para implementar la organización de la base de casos (BC) que parte de un agrupamiento, que favorece el acceso y la recuperación a los casos más similares para el completamiento eficiente del problema a resolver y brindar amplias ventajas para mejorar la calidad de las soluciones a partir del refinamiento propuesto.

2. CONTENIDO

En esta sección se realiza una descripción de los elementos utilizados para la implementación del razonador basado en casos presentado, tomando como base: (1) el modelo para la inferencia de hipótesis diagnósticas tempranas propuesto en [6] y (2) los elementos necesarios para aplicar *Rough Texts* [8].

2.1 Modelo para la toma de decisiones clínicas

En el contexto de la actividad hospitalaria, la HCE constituye la fuente esencial de datos y el documento principal de la historia del paciente, estas son almacenadas en el SIH. Luego, cada HCE está conformada por los elementos que describen las características de su(s) patología(s), las decisiones tomadas por los especialistas y la respuesta al tratamiento. De forma que, un repositorio de HCE puede ser descrito como una colección de documentos $D = \{D_1, \ldots, D_m\}$, donde cada D_i contiene una serie de Unidades Estructurales (UE) que caracterizan estas partes, tal que $UE = \{UE_1, \ldots, UE_n\}$ [10]. Las UE semánticamente identificadas en la HCE basándose en criterio de expertos son UE= (Antecedentes, Hábitos Tóxicos, Síntomas, Signos, Historia de la Enfermedad Actual, Diagnóstico Diferencial, Diagnóstico, Tratamiento, Complementarios, Evolución, Pronóstico).

Tomando como premisa lo antes enunciado, el modelo para el descubrimiento de conocimiento implícito, parte de la colección de HCE en formato semiestructurado, basado en la arquitectura de especificación para la codificación de la información clínica CDA-HL7 [9], facilitando la manipulación de las partes a partir del uso de etiquetas.

Específicamente, la etapa de agrupamiento permite delimitar y organizar la información relevante, descubrir grupos de documentos afines a partir del cálculo de la similitud entre los pares de documentos e identificar los elementos más representativos relativos a cada partición, a partir de los agrupamientos asociados a cada Unidad Estructural, identificada semánticamente según criterio de expertos.

Como salida, ante la presencia de un nuevo problema, el modelo infiere posibles hipótesis diagnósticas, a partir de los casos más similares a este almacenados en la base de casos. En la Figura 1 se muestra un esquema que ilustra la esencia del modelo. El mismo transita por diferentes etapas como se muestra en la Figura 2.

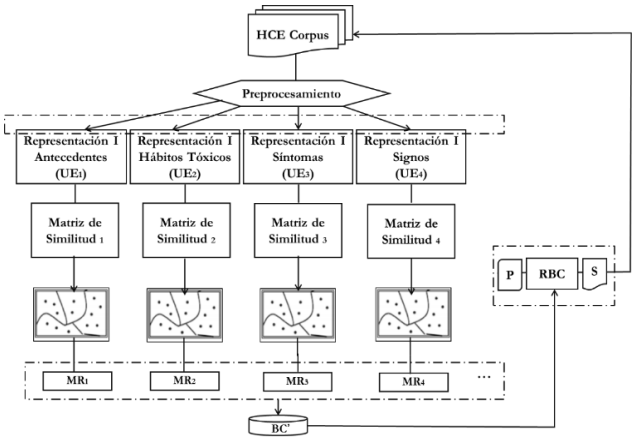

Figura. 1: Modelo para el descubrimiento de conocimiento

Módulo 1(Pre-procesamiento). Recuperación de la información, identificando en cada documento recuperado las Unidades Estructurales (UE) que constituyen los rasgos predictores para el RBC.

Módulo 2(Representación). Representación del corpus textual obtenido para cada UE, obtener una representación VSM clásica.

Módulo 3(Similitud y Agrupamiento). Cálculo de la matriz de similitud utilizando la medida de similitud *Coseno* y un agrupamiento para cada UE

- **3.1** Refinamiento del agrupamiento a partir de la calidad de los grupos.
- **3.2** Determinación de prototipos a partir del cálculo de la aproximación inferior.

Módulo 4(Razonamiento). Realizar la inferencia de posibles hipótesis diagnósticas para un nuevo problema *P*.

- **4.1** Recuperación (determinar los prototipos más similares al nuevo problema)
- **4.2** Adaptación (a partir del cálculo de la aproximación superior se revisa y se construye una solución final)

Figura. 2: Módulos principales del RBC propuesto.

2.2 Sistema de decisión necesario para aplicar RST

La teoría de los conjuntos aproximados tiene gran utilidad para evaluar la calidad del agrupamiento, a partir del cálculo de las aproximaciones inferiores, las cuales a su vez facilitan el proceso de selección de los prototipos o documentos más representativos de cada grupo, de lo cual en buena medida depende de la calidad del agrupamiento.

Para refinar los resultados obtenidos por los agrupamientos se utiliza RST.

La representación VSM del corpus y el resultado del agrupamiento como atributo de decisión, conforman el sistema de decisión necesario para aplicar RST.

Sea *Q* un conjunto de documentos de HCE y sea $s{:}Q\times Q\rightarrow\mathbb{R}$ la función que mide la similitud entre los documentos de *Q*. Pawlak define la relación R' por la expresión (1), donde $s(d_i,\ d_j)$ retorna el valor de similitud entre los documentos d_i y d_j, considerando el umbral de similitud ξ [7, 11]. Hacemos notar que, se emplea el símbolo d_z para denotar documentos que pertenecen a *Q* y D_z para denotar los documentos que además de pertenecer a *Q* pertenecen a D.

$$R'(d_i)=\{d_j\in Q : d_i R' d_j, \text{ es decir } d_j \text{ está relacionado con } d_i \text{ si y solo si } s(d_i,\ d_j)>\xi\} \quad (1)$$

Siendo *s* el coeficiente Coseno[12] adecuado para el trabajo con textos resultado de evaluación en [13] .

En las expresiones (2) y (3) se muestran las formas de cálculo a partir de la relación *R'* para cada documento en *Q,* de las aproximaciones R'-inferior $R'_*(C_j)$ y R'-superior $R'^*(C_j)$, respectivamente. Vale señalar que, en la etapa de refinamiento que se detalla más adelante en este documento, se requiere el cálculo de estas aproximaciones y sólo se consideran los documentos que en ese momento del procesamiento pertenecen a *D.* A continuación, se incluyen estas definiciones considerando solo los documentos en *D*.

Sea $C=\{C_1,\ldots,C_p\}$, los grupos obtenidos al aplicar el agrupamiento inicial de una UE que, en el momento del refinamiento constituyen las clases de decisión. A continuación, en este documento se emplea la notación C_{iz} para denotar el grupo C_i de la UE_z. No obstante, con el propósito de facilitar la formalización de las expresiones para el cálculo de las aproximaciones, solo se emplea el primer subíndice para denotar los grupos de una UE arbitraria.

$$R'_*(C_j)=\{D_i\in C_j : R'(D_i)\subseteq C_j\} \quad (2)$$

$$R'^*(C_j)=\bigcup_{D_i\in C_j} R'(D_i) \quad (3)$$

Así, los documentos de la aproximación inferior $R'_*(C_j)$ del grupo C_j, son aquellos que pueden ser clasificados con certeza como miembros del grupo C_j, ya que solamente se relacionan con documentos del mismo grupo y son los más representativos de ese grupo, mientras que los documentos en $R'^*(C_j)$ son la unión de todos los documentos que pertenecen al grupo C_j, con los documentos que ellos se relacionan según la relación R', de manera que los documentos que pertenecen a $R'^*(C_j)$ pueden ser clasificados

como posibles miembros del grupo y son aquellos que posiblemente pertenezcan a él.

2.3 Implementación del RBC

En el modelo presentado anteriormente, se propone una organización de la BC que responde a una estructura jerárquica, concebida a partir de grupos de documentos HCE. Los cuales se obtienen a partir del agrupamiento de cada UE para cada rasgo predictor. La Figura 3 muestra un esquema general de los pasos que se siguen para organizar los casos de la BC, de manera que los rasgos predictores son las UE que se obtienen del interrogatorio médico-paciente; tal que, $DUE_z \in$ Rasgos_Predictores= (Antecedentes, Hábitos Tóxicos, Síntomas, Signos, Historia de la Enfermedad Actual). Para cada grupo de casos obtenido, el caso prototipo se corresponde con el caso que más se asemeja a los restantes documentos del grupo. Para ello, se calcula la pertenencia de cada documento al grupo y se selecciona el de mayor pertenencia, el cual representa al documento, cuya similitud promedio obtenida a partir del cálculo de su similitud con todos los documentos del grupo es mayor.

Entrada: BC de HCE, UE que constituyen rasgos predictores
Salida: BC' (Particiones: Grupos homogéneos de HCE afines, HCE prototipos cada grupo y calidad del agrupamiento.)
Inicio
Para cada $DUE_z \in$ Rasgos_Predictores
1. Construir matriz de semejanza utilizando la Similitud_Coseno.
2. Estimación del umbral de similitud β_0.
3. Agrupar y Refinar.
4. Determinar *Prototipo(s)* de cada grupo.

Fin_Para
Fin

Figura. 3: Procedimiento general para el esquema de organización de la BC.

2.3.1 Procedimiento general del RBC

Este procedimiento consta de cuatro módulos, los cuales se describen a continuación.

Módulo 1: Preprocesamiento

Este módulo es el encargado de la recuperación de la información, en esta fase (etapa) se identifican en cada documento recuperado las UE que constituyen los rasgos predictores para el RBC [14,15].

Módulo 2: Representación

Debido a la necesidad que el modelo realice un análisis léxico del texto y/o permita representar la información, se ha considerado la forma de representación VSM, no costosa computacionalmente como ocurre con los modelos probabilísticos, y que puede adaptar al contexto de los criterios que emiten los especialistas en la HCE de la relatoría del paciente, en su mayoría textos cortos[6]

El resultado de esta etapa es la representación VSM para cada UE; a partir de la matriz VSM clásica [16], que contiene en sus filas el índice de términos construido utilizando *Lucene* y los documentos de la colección en sus columnas. Las celdas representan la frecuencia de aparición de cada término en la UE del documento que se procesa. A partir de la cual el sistema realiza el proceso de normalización, utilizando la frecuencia absoluta de aparición del término y la longitud del documento (se asume como longitud del documento la longitud de la UE, ya que estas son tratadas como colecciones independientes).

Módulo 3: Similitud y Agrupamiento

Este módulo incluye las etapas relativas al cálculo de la matriz de similitud y el agrupamiento de cada colección independiente. Se calcula una matriz de similitud utilizando como medida la similitud coseno que se muestra en la ecuación (4), vale señalar que el primer índice indica el documento a considerar y el segundo se refiere a las componentes del vector en la representación VSM. El agrupamiento tiene por objetivo crear grupos coherentes internamente, es decir, que los documentos de un mismo grupo sean lo más similares posibles y a la vez que sean disimilares a los documentos de otros grupos [17]. De manera que en la organización de la base de casos cada grupo, estará conformado por los pacientes cuyos rasgos coinciden con un cierto nivel de precisión. Para la implementación de este módulo se empleó la herramienta SSPACE[16] la cual facilita obtener la partición inicial (agrupamiento), que constituye la entrada del algoritmo de refinamiento. Debido al buen desempeño de los algoritmos de agrupamiento jerárquicos [8] en esta implementación se utiliza el algoritmo *Hierarchical Agglomerative Clustering* (HAC) de enlace completo [6] .

$$s(d_i, d_j) = \frac{\sum_{k=1}^{m}(d_{ik} \cdot d_{jk})}{\sqrt{\sum_{k=1}^{m} d_{ik}^2 \cdot \sum_{k=1}^{m} d_{jk}^2}} \qquad (4)$$

Refinamiento del agrupamiento

Varios algoritmos de agrupamiento tienden a crear muchos grupos unitarios o mal formados, sobre todo cuando el umbral de similitud entre los documentos no es adecuadamente seleccionado. En tales casos, es posible obtener aproximaciones inferiores vacías para esos grupos, lo cual significa que todos los documentos del grupo se relacionan con documentos de otros, indicando que el grupo no es consistente, por lo que no está adecuadamente conformado. Este elemento puede utilizarse como otro indicador de la calidad del grupo y puede ser utilizado para refinar los resultados del agrupamiento [19].

En este trabajo se realiza un refinamiento de cada uno de los agrupamientos iniciales obtenidos para cada UE, a partir de los resultados de sus aproximaciones inferiores. Para esto, cuando se analiza el agrupamiento inicial de cada UE_z se analiza la colección independiente conformada a partir de la UE_z de los documentos en *D* del grupo que se quiere eliminar y de este modo el documento *Di* es asignado al grupo C_{iz} que tenga mayor relación utilizando para esto la ecuación (5).

A continuación, se muestra el procedimiento empleado para realizar el refinamiento de los grupos, a partir de los resultados obtenidos por el agrupamiento inicial de cada colección independiente. Las entradas a este procedimiento son: el resultado del agrupamiento inicial para una UE_z, así como los resultados de las aproximaciones inferiores y superiores de cada grupo. La salida es el agrupamiento refinado.

Entrada: Aproximaciones inferiores (2) y superiores (3), Calidad del agrupamiento
Salida: Agrupamiento refinado
Mientras existan grupos con aproximación inferior vacía ($R'_* = \emptyset$) **hacer**
Seleccionar el primer grupo C_{iz} tal que $R'_*(C_{iz}) = \emptyset$
Para j=1 **hasta** Cantidad de Documentos del C_{iz}
Mientras K<=Cantidad total de grupos **hacer**
Calcular la pertenencia aproximada del documento *j* (D_j) al grupo K $\left(\varpi_{C_{Kz}}(D_j)\right)$, a través de la expresión (5)
Asignar D_j al grupo hacia al cual tuvo mayor pertenencia.
Eliminar el C_{iz}

Figura. 4: Aplicación de RST en el refinamiento del agrupamiento.

Pertenencia Aproximada:

El cálculo de la pertenencia aproximada permite obtener el grado de inclusión de un documento a un grupo. La idea es considerar de forma integrada el grupo C_i y el conjunto de documentos relacionados con *D*. Así, en el denominador del cociente se calcula la cardinalidad de la unión de estos conjuntos. La expresión (5) muestra la forma de cálculo de la pertenencia aproximada y la expresión (6) su uso para ponderar los grupos [17].

$$\varpi_{C_i}(D_j) = \frac{|C_i \cap R'(D_j)|}{|C_i \cup R'(D_j)|} \quad (5)$$

$$w_i = \frac{\sum_{D_j \in C_i} \varpi_{C_i}(D_j)}{|C_i|} \quad (6)$$

Determinación de prototipos

Después de haber realizado el agrupamiento y refinarlo, para cada grupo de casos obtenido, se determina el caso o los casos "representantes del grupo" o "prototipos", que se corresponden con los casos que pueden ser clasificados con certeza como miembros del grupo. La determinación de prototipos se basa en el cálculo de la aproximación inferior luego de obtener los grupos refinados. Para el cálculo de la aproximación inferior se utiliza la expresión (2).

Si existe más de un elemento en la aproximación inferior, los cuales serán prototipos del grupo, estos son ordenados decrecientemente atendiendo al valor de la pertenencia promedio de este, con los elementos del grupo al cual pertenece. La siguiente expresión muestra cómo obtener la pertenencia promedio de un documento D_j en el grupo C_i, siendo *m* la cantidad de documentos del grupo.

$$SG(D_j, C_i) = \frac{\sum_{k=1}^{m} \; s(D_j, D_k)}{\frac{j \neq k, D_k \in C_i}{m-1}} \quad (7)$$

De esta forma el prototipo del grupo queda en un nivel superior en la BC, lo que facilita el acceso y recuperación de los casos más semejantes al nuevo problema.

Módulo 4: Razonador Basado en Casos

- **Recuperación**

La recuperación se basa en una similitud local que determina la analogía entre valores de un mismo rasgo y una similitud global, que combina los resultados de las similitudes locales a todos los rasgos de los casos a comparar. La similitud global es el resultado de la suma ponderada de las similitudes entre las UE correspondientes de cada caso y el caso problema, según se observa en la ecuación (8). La notación $s_z(D_i, D_j)$ denota la similitud entre los documentos considerando solamente el contenido textual de la UE_z.

$$\text{SimGlobal}(D_i, D_j) = \sum_{z=0}^{m} W_i * s_z(D_i, D_j)/n \quad (8)$$

donde: $\sum W_z = 1$

Para el cálculo de los pesos de los rasgos predictores se consultó a los expertos, los cuales concluyeron que los rasgos Historia de la Enfermedad Actual, Signos y Síntomas son los más importantes, aportando en la mayoría de los casos aproximadamente un 75% de la información que se requiere para el diagnóstico, en este sentido la suma del peso de estos rasgos debe ser 0.75, y la suma total de todos 1. En [6] se realizó un estudio basado en las consideraciones anteriores. En esta propuesta, a partir de los resultados obtenidos en el estudio, se emplean los valores de pesos: HEA=0.25, SG=0.25, ST=0.25, HT=0.13, AP=0.12.

Si la similitud global supera el umbral se recupera el caso. De lo contrario se retroalimenta la búsqueda con el grupo más similar al que pertenece el caso. A partir de la aproximación superior del grupo al cual tiene mayor grado de pertenencia (Ver expresión (3)), se explora dicha aproximación superior y se recupera el documento o los documentos de la aproximación superior, atendiendo a la $SimGlobal$ mayor, es decir todos aquellos que tengan el máximo valor de similitud global con el nuevo paciente. La Figura 4, muestra el procedimiento empleado en el proceso de recuperación de los casos más similares al nuevo paciente.

Entrada: P, problema a resolver
$\mathrm{BC}' = \{\mathrm{bc}'_1, \mathrm{bc}'_2, \ldots, \mathrm{bc}'_n\}$; sub-conjunto de casos obtenidos como prototipos.
Salida: MS'
Inicio
$MS' = \emptyset$
$MS = \{p_i | i = \overline{1,k}\}$, *k* casos más similares, $MS \subset \mathrm{BC}'$.
$\forall bc_i \in MS$
Si $SimGlobal(bc_i, P) > \gamma$
$MS' = MS' \cup bc_i$
Si no
$C_g = maxSimG\{C_{iz} | \forall C_{jz},\ bc'_i \in C_{jz}, z = \overline{1,5}\}$
$LA = R'^{*}(C_g)$
$LA' = \{D_i \in LA | max\{SimGlobal(D_i, P)\}\}$
$MS' = MS' \cup LA'$
Fin

Figura. 4: Algoritmo para la recuperación de los casos más similares.

- **Adaptación**

En este paso se consideran los casos recuperados en la etapa anterior para elaborar una solución, que constituye un conjunto de diagnósticos posibles a considerar ante la admisión de un nuevo paciente. En adición a ello, de los diagnósticos sugeridos, el sistema propone un primer diagnóstico (*Hipótesis*) como el diagnóstico posible tomando en consideración los diagnósticos presentados para los casos más similares al nuevo problema *P*, y emite además un conjunto de diagnósticos diferenciales que deben descartarse. Para presentar el conjunto de diagnósticos diferenciales posibles se discurren los diagnósticos de los casos más similares y sus diagnósticos diferenciales, realizando una adaptación que elimina los diagnósticos o diagnósticos diferenciales, ya considerados los cuales se ordenan atendiendo al nivel de ocurrencia.

En la Figura 5 se describe como se realiza la adaptación al nuevo problema.

Entrada: $MS = \{c_1, c_2, \ldots, c_M\}$, *M* casos obtenidos del módulo recuperador(*M>k* casos más similares).

Salida: DF, DN; Rasgos objetivos.

Inicio

$\forall\ \mathrm{c_i} \in\ \mathrm{MS}$, $LD = \{\mathrm{ld}_z = \{\mathrm{DN}_i\ \mathrm{U}\ \mathrm{DF}_i\}\}$. $z=\overline{1,M}$

$$R(i,j) = \begin{cases} 1 & , \mathrm{ld}_i \cap \mathrm{ld}_j \neq \emptyset \\ 0 & , e.o.c. \end{cases}$$

$DM = MAJORITY\ (c_{i\,DN \cup DF}, \mathrm{R})$

Se obtiene MS′ donde $\mathrm{MS}' \subseteq \mathrm{MS}$, casos obtenidos a partir de *DM*

$\forall\ \mathrm{c_i} \in\ \mathrm{MS}'$,

construir $DC_1 = \{D_i | D_i \in DN(c_i)\}$.

$DN = \max_z\{OCCUR(D_i), \mathrm{D_i} \in DC_1\}$

construir

$DC_2 = \{D_j | D_j \in DN(c_i) \cup List(DF(c_i)) \wedge\ d_j \notin DC_1\}$.

$DF = List(max_z\{OCCUR(D_i), \mathrm{D_i} \in DC_2\})$

Fin

Figura. 5: Algoritmo para conformar la solución inicial.

- **Revisión**

La revisión se realiza cuando los casos recuperados fueron los *k* más similares iniciales, por tanto, no se explora la aproximación superior de los grupos de ninguno de los casos obtenidos como los más similares, es decir, la cantidad de casos son insuficientes para el completamiento del problema y no se pudieron discriminar atendiendo a la similitud con que estos fueron recuperados. Se realiza un proceso de selección basado en la comparación jerárquica por rasgos predictores y la importancia de estos en la recuperación con el propósito de obtener el caso más similar al nuevo problema

En la Figura 6, se describe cómo se obtiene el DN a partir de la revisión de la adaptación.

Entrada: $MS' = \{p_1, p_2, \dots, p_t\}$ t casos iniciales.
Salida: conjunto diagnósticos
Inicio
Si los casos recuperados fueron los t iniciales
1. Ordenar los **Rasgos_Predictores** de mayora menor atendiendo al peso
2. **Para** z=0 **hasta** Cantidad de UE ,**UE$_z$** ∈ **Rasgos_Predictores**
Para j=0 **hasta** Cantidad de Documentos de MS'
Calcular $s_{\mathbf{UEz}}(p_j, P)$
Fin Para
Ordenar similitudes de los p_j de mayor a menor
Si la cantidad de p_j con similitud máxima es 1
Retornar el DN del p_j que su similitud es máxima
Sino
Si z==Cantidad de UE
Para j=0 **hasta** Cantidad de Documentos de MS'
Adicionar a *listDN* el DN de p_j
Fin Para
Sino
Para i=0 **hasta** cantidad de p_i con similitud máxima
Adicionar p_i a MS''
$MS' = MS''$
Fin Para
Fin Para

Figura. 6: Algoritmo para conformar el DN a partir de la revisión de la adaptación.

Breves consideraciones sobre el umbral de similitud

Estimar adecuadamente el umbral de similitud a utilizar es fundamental, así como la selección de la similitud a emplear. En este trabajo se utiliza la similitud Coseno porque ha sido una de las más efectivas para comparar vectores documentos como antes fue enunciado. En [10] se han propuesto expresiones que permiten estimar umbrales a partir de una matriz de similitudes o distancias y que no requieren información adicional del conjunto de datos que se procesen. Una de estas variantes es la media de las similitudes entre todos los pares de documentos posibles (Ecuación (9), Variante 1). Aunque es necesario tener en cuenta que la media es sensible a observaciones extremas, y en su defecto, se pudiera utilizar la media recortada para eliminar los valores fuera de rango.

$$\bar{X} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} s(D_i, D_j) \qquad (9)$$

También se puede utilizar la media de los valores máximos de las similitudes entre cualquier par de objetos (Ecuación 10, Variante 2). Esta forma de cálculo puede provocar la obtención de un umbral muy alto, conduciendo a que exista coincidencia del grupo y sus aproximaciones. Esta situación puede arrojar valores de precisión y calidad cercanos a 1, cuando en realidad el resultado del agrupamiento no sea tan bueno (R_*(Grupo) <<Grupo<<R^*(Grupo) => Precisión = calidad = 1).

$$\overline{X}_{max} = \frac{1}{n} \sum_{i=1}^{n} max\{s(D_i, D_j)\} \qquad (10)$$

La última variante de cálculo a utilizar es la media ponderada de la media de las similitudes y la media de los máximos; (Ecuación 11, Variante 3).

$$\overline{X}_{max} = \frac{1}{n} \sum_{i=1}^{n} max\{s(D_i, D_j)\} \qquad (11)$$

La descripción de la notación utilizada es la siguiente: *n* es la cantidad de objetos de la colección, $s(D_i, D_j)$, es el valor de la similitud entre los vectores documento D_i y D_j, y α es un valor entre 0 y 1 que permite ponderar la media y la media de los máximos en la ecuación (11). En esta tesis el valor de α que se utilizó fue 0.5, según [17] arroja los mejores resultados.

3. RESULTADOS Y DISCUSIÓN

Para la evaluación del sistema implementado se utilizan como caso de estudio colecciones de HCE del Servicio de Admisión del Hospital Provincial "Celestino Hernández Robau" asociadas a diferentes enfermedades, realizándose tres experimentos. El manejo de la información se realizó en el marco de la ética profesional, la confidencialidad y la privacidad legal que se establece.

Experimento 1. Para ilustrar como el refinamiento del agrupamiento favorece obtener grupos más compactos, se diseña un experimento sobre una colección pequeña de 20 HCE, tomadas del caso de estudio antes descrito.

Las aproximaciones inferiores y superiores de cada grupo resultante del proceso de agrupamiento sugieren como fusionar grupos, de forma tal que se refinen los resultados de agrupamientos, al existir muchos grupos unitarios o con muy pocos documentos,

lo cual muestra la viabilidad del procedimiento para el refinamiento empleado. A continuación, se muestra una situación real donde se requiere la fusión y se ilustra como las aproximaciones inferiores y superiores pueden contribuir al refinamiento de los resultados.

En las Tabla 1 se pueden observar los grupos resultantes del proceso de agrupamiento inicial relativo a la UE HEA, con sus aproximaciones inferiores y superiores.

Tabla 1: Grupos de documentos y sus aproximaciones inferiores y superiores.

G_i	Documentos	$R'_*(G_i)$	$R'^*(G_i)$
1	[7]	[7]	[7]
2	[11]	[11]	[11]
3	[16]	[16]	[16]
4	[9, 13]	[]	[2, 5, 9, 13, 15, 19]
5	[5, 15]	[]	[0, 5, 9, 15, 19]
6	[1, 2]	[]	[1, 2, 3, 10, 12, 13, 17, 18]
7	[18, 3, 6]	[6]	[0, 1, 2, 3, 6, 12, 18, 19]
8	[4, 8]	[]	[4, 8, 12, 17]
9	[0, 19, 14]	[14]	[0, 3, 5, 13, 14, 18, 19]
10	[17, 10, 12]	[]	[1, 2, 3, 4, 8, 10, 12, 17, 18]

En la Tabla 2 se observan los grupos resultantes del refinamiento del agrupamiento con sus aproximaciones inferiores y superiores.

Tabla 2: Grupos refinados y sus aproximaciones inferiores y superiores.

G_i	Documentos	$R'_*(G_i)$	$R'^*(G_i)$
1	[17, 4, 7, 8, 10, 12]	[4, 7, 8]	[1, 2, 3, 4, 7, 8, 10, 12, 17, 18]
2	[11]	[11]	[11]
3	[16, 5, 9, 13, 15]	[9, 15, 16]	[0, 2, 5, 9, 13, 15, 16, 19]
4	[18, 3, 6]	[6]	[0, 1, 2, 3, 6, 12, 18, 19]
5	[0, 1, 2, 19, 14]	[14]	[0-3, 5, 10, 12, 13, 14, 17-19]

Los resultados a partir de la conformación de grupos antes y después del refinamiento evidencian que existen grupos resultantes del proceso de agrupamiento inicial con aproximación inferior vacía, lo cual no es adecuado. Este efecto se corrige al aplicar el procedimiento de refinamiento.

Experimento 2. Debido a que una de las etapas del modelo para recuperar los casos más similares al nuevo problema presentado necesita de la determinación de un umbral que permita estimar cuan similares son los objetos al nuevo problema, en este trabajo se proponen utilizar las variantes anteriormente enunciadas. Con el propósito de verificar el comportamiento del RBC propuesto a partir de la validez de las respuestas obtenidas, dependiendo de la selección de una u otra variante, se realiza el experimento que consistió en emplear el conjunto de casos correspondientes al caso de estudio y una validación cruzada *Leave One Out Cross Validation* (LOOCE) [20]. Los resultados se observan en la Tabla 3.

Tabla 3: Porcientos de validación para cada variante de cálculo de umbral.

	Variante 1	Variante 2	Variante 3
% RBC	100.0	98.5663	100.0

Experimento 3. Para evaluar la precisión del RBC se realiza una validación cruzada LOOCE [20]. En la literatura[21] se sugiere que la recuperación de los *k* casos más similares, se debe realizar sobre la base de considerar valores de *k=1*, *k=3*, *k=5* o *k=7.* Por ello, en este trabajo se consideraron diferentes valores de *k*. Los resultados alcanzados se muestran en la Tabla 4.

Tabla 4: Exactitud del RBC para los distintos valores de k.

Valor de k	k=1	k=3	k=5	k=7
Exactitud (%)	100.0	100.0	91.8919	90.0901

Los resultados obtenidos corroboran la validez del sistema propuesto que implementa el modelo, facilitando la manipulación de Historias Clínicas a partir del estándar CDA, integrando agrupamiento y clasificación automática.

3.1 Sistema CDARS

El sistema CDARS fue desarrollado a partir del RBC presentado. El mismo a partir de un repositorio de HCE disponible permite el entrenamiento asistencial que facilita a los estudiantes corroborar sus diagnósticos en contraste con las hipótesis sin necesidad de ingresar casos en el sistema. Para obtener las hipótesis diagnósticas el estudiante debe haber completado la información de las UE predictoras, como re-

sultado del interrogatorio realizado al aplicar el método clínico. Para asistir al usuario, CDARS ofrece el completamiento de los campos Diagnóstico y Diagnóstico Diferencial.

Figura. 7: Sistema para la toma de decisiones diagnóstica CDARS

4. CONCLUSIONES

Como resultado del presente trabajo se presenta un sistema basado en casos para la inferencia de hipótesis diagnósticas, basado en los elementos de la teoría de los conjuntos aproximados (CDARS), la evaluación del mismo demuestra la eficacia en la inferencia de posibles hipótesis diagnósticas. Este sistema constituye un entrenador asistencial para los estudiantes en los centros hospitalarios, auxiliándolos en la obtención de hipótesis diagnósticas tempranas a través de una interfaz amigable.

5. REFERENCIAS BIBLIOGRÁFICAS

1. Visval, A. and M. Sara, La gestión documental, de información y el conocimiento en la empresa. El caso de Cuba. Acimed, 2009. 19(5).
2. Fiks, A.G., et al., Electronic medical record use in pediatric primary care. Journal of the American Medical Informatics Association, 2010. 18(1): p. 38-44.
3. Fernández-Alemán, J.L., et al., Analysis of health professional security behaviors in a real clinical setting: An empirical study. International journal of medical informatics, 2015. 84(6): p. 454-467.
4. Martınez, I.G. and R.E.B. Pérez, Making decision in case-based systems using probabilities and rough sets. Knowledge-Based Systems, 2003. 16(4): p. 205-213.
5. Lorenzo, M.M.G. and R.E.B. Pérez, A model and its different applications to case-based reasoning. Knowledge-based systems, 1996. 9(7): p. 465-473.
6. Fuentes, I., Valdés, B., García, M., Arco, L., Herrera, M., Fuentes, R. (2017). A Case-Based Reasoning Framework for Clinical Decision Making. 16th Mexican International Conference on Artificial Intelligence MICAI 2017. Vol. 10062. 2017: Springer.
7. Pawlak, Z., Rough sets. International Journal of Computer & Information Sciences, 1982. 11(5): p. 341-356.
8. Arco, L., et al., Rough Text Assisting Text Mining: Focus on Document Clustering Validity. Granular Computing: At the Junction of Rough Sets and Fuzzy Sets, 2008: p. 229-248.
9. Dolin, R.H., et al., HL7 clinical document architecture, release 2. Journal of the American Medical Informatics Association, 2006. 13(1): p. 30-39.
10. García Lorenzo, M.M., Monografía de reconocimiento de patrones. 1999. 8(1).
11. Pawlak, Z., Rough sets. International Journal of Parallel Programming, 1982. 11(5): p. 341-356.
12. Cruañes Vilas, J., M.T. Romá Ferri, and E. Lloret, Análisis del uso de métodos de similitud léxica con conocimiento semántico superficial para mapear la información de enfermería en español. 2012.
13. Magdaleno, D., et al., Comparative Study of Clustering Algorithms using OverallSimSUX Similarity Function for XML Documents. Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial, 2015. 18(55).
14. Fuentes, I. E., Magdaleno, D., & García, M. M. (2015a). Methodology for discovery of implicit knowledge in Medical Records. Paper presented at the Fifth international workshop on Knowledge Discovery,

Knowledge Management and Decision Support, EUREKA 2015, Universidad Autónoma Metropolitana.Ciudad México. Fuentes

15. Fuentes, I., Valdés, B., García, M., Arco, L., Herrera, M., Fuentes, R. (2017). Clinical Diagnostic based-on Textual and Structured Information Analysis. Published in: Memorias del evento International Conference of Processing Information CIPI 2017, ISBN: 978-959-312-258-0.
16. Salton, G., A. Wong, and C.-S. Yang, A vector space model for automatic indexing. Communications of the ACM, 1975. 18(11): p. 613-620.
17. Guevara, D.M., Y. Miranda, and I. Fuentes, Comparative Study of Clustering Algorithms using OverallSimSUX Similarity Function for XML Documents. Inteligencia Artificial, 2015. 19(57): p. 69-80.
18. Jurgens, D. and K. Stevens. The S-Space package: an open source package for word space models. in Proceedings of the ACL 2010 System Demonstrations. 2010. Association for Computational Linguistics.
19. Magdaleno, D., Refinamiento, evaluación y etiquetamiento de grupos textuales basados en la teoría de los conjuntos aproximados, in Ciencias de la Computación. 2008, Universidad Central "Marta Abreu" de Las Villas: Santa Clara, Villa Clara.
20. Cawley, G.C. Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. in Neural Networks, 2006. IJCNN'06. International Joint Conference on. 2006. IEEE.
21. Wess, S., K.-D. Althoff, and G. Derwand. Using kd trees to improve the retrieval step in case-based reasoning. in European Workshop on Case-Based Reasoning. 1993. Springer.

6. SÍNTESIS CURRICULARES DE LOS AUTORES

Katterine Nodarse, graduada de Licenciatura en Ciencia de la Computación en la Universidad Central "Marta Abreu" de Las Villas (UCLV), Santa Clara, Cuba, en 2017. Pertenece al Laboratorio de Inteligencia Artificial (IA) del Centro de Investigaciones de la Informática, UCLV dónde trabaja actualmente para alcanzar el grado de Máster. Ha alcanzado en diferentes premios en fórums y concursos, por los resultados en sus investigaciones como estudiante investigador desde su incorporación en 2014 al grupo de investigación de IA.