Recomendación de Códigos CIE-10 a Diagnósticos Médicos en Escenarios a Gran Escala

Por:

Blanca Bertha Velasco Sustaita

Tesis sometida como requisito parcial para obtener el grado de

Maestro en Ciencias en el área de Ciencias Computacionales

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica

Supervisada por:

Dr. Luis Villaseñor Pineda
Dr. Manuel Montes y Gómez

Luis Enrique Erro 1
Sta. Ma. Tonantzintla,
72840, Puebla, México.

Resumen

Actualmente se han aplicado métodos de procesamiento del lenguaje natural a textos clínicos con el fin de brindar un soporte para la toma de decisiones en ámbitos de administración, de ciencia o de educación. La clasificación manual de diagnósticos médicos es un proceso intensivo que consume importantes recursos humanos en los centros hospitalarios. En la asignación de códigos CIE-10-MC a diagnósticos médicos se necesita de varios expertos etiquetadores para procesar y asignar códigos de manera manual. Este trabajo presenta un sistema de procesamiento de textos clínicos para la asignación de códigos CIE-10-MC a diagnósticos médicos el cual funcionará como herramienta para facilitar la tarea de asignar códigos y automatizar la toma de decisiones.

Abstract

Currently, methods have been applied for natural language processing of clinical texts in order to provide support for decision-making in areas of administration, science or education. The manual sorting of medical diagnostics is an intensive process that consumes significant human resources in hospitals. In assigning ICD-10-CM codes to medical diagnostics several taggers experts is required for to process and assignment codes manually. This paper presents a system for processing clinical texts for assigning ICD-10-CM codes to medical diagnoses which will function as a tool to facilitate the task of assigning codes and automate decision making.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

El procesamiento automático de textos médicos es de creciente interés tanto para los profesionales de la salud como para investigadores académicos. La principal motivación es el uso de la información recaudada para toma de decisiones, temas de administración, de ciencia y de educación. Dentro de los objetivos principales en el procesamiento de textos biomédicos se encuentran la detección automática de términos, la creación de ontologías, o para la minería de datos, la clasificación y extracción de conocimiento. Este tipo de tareas consumen tiempo y muchos recursos cuando se lleva a cabo de forma manual. Esta cantidad enorme de textos médicos crea interesantes oportunidades para aplicaciones de aprendizaje, como el desarrollo de buscadores, clasificación de documentos, la minería de datos, extracción de información, y la extracción de relaciones. El sistema de cuidado de la salud emplea un gran número de sistemas de categorización y clasificación para ayudar a la gestión de datos para una variedad de tareas, incluyendo la atención al paciente, almacenamiento de registros, recuperación de información del paciente, análisis estadístico, aseguramiento y facturación, entre otros. Uno de estos sistemas es la Clasificación Internacional de Enfermedades, Décima Revisión, Modificación Clínica (CIE-10-MC) que es el sistema oficial de asignación de códigos a los diagnósticos y procedimientos médicos. Predecir automáticamente los códigos CIE-10-MC de cada historia clínica requiere el reconocimiento de la más sobresaliente enfermedad(es) o síntoma(s) en el texto del diagnóstico. Esta tarea se relaciona con identificar los conceptos clave en la historia clínica. Automatizar la asignación de códigos CIE-10 a diagnósticos médicos

no solo puede reducir costosos recursos, también puede ayudar a resolver inconsistencias en la codificación que surgen debido al error humano ya que los diagnósticos médicos suelen ser capturados a través de dictados o como texto libre, ya sea escrito a mano o en campos de formularios no estructurados. Los codificadores, a menudo formados profesionalmente, deben interpretar lo que quiso decir el médico y asignar códigos CIE-10.
Para llevar a cabo el estudio de la asignación de códigos CIE es de suma importancia tomar en cuenta todos los aspectos que encierran el proceso de generar un diagnóstico médico así como su codificación, por ejemplo:

- Persona encargada de generar el texto médico.
- Nota médica.
- Se cuenta con el diagnóstico médico definido por el doctor a cargo del paciente .
- Información completa del paciente (perfil)
- Medicamentos prescritos al paciente.

Cada uno de estos aspectos nos puede brindar información importante para la codificación CIE. En casos del mundo real donde los textos médicos son generados comúnmente al finalizar la consulta médica, los doctores a cargo no tienen el cuidado de la correcta escritura del diagnóstico médico generando un texto con errores así como la generación de textos abreviados o en claves comunes para los médicos.
Debido a la importancia de la asignación correcta de estos textos médicos es fundamental ser lo más preciso posible. Si bien es poco realista pensar que un algoritmo puede automatizar completamente el proceso de códigos de asignación estos pueden ayudar a los etiquetadores a tomar una decisión más rápida, sugiriendo un subconjunto de códigos para elegir. El apoyo a las decisiones es un enfoque que ha sido adoptado por la industria de la salud para una serie de tareas, la codificación CIE-10 es tan sólo uno de ellos. Los códigos no son independientes ya que existe una jerarquía en ellos y diferentes etiquetas pueden interactuar para aumentar o disminuir la probabilidad de la otra. Los diagnósticos médicos son documentos muy cortos y

es ahí donde se plantea el desafío de usar las técnicas actuales de procesamiento de textos biomédicos para abordar este tema. En la figura 1.1 se puede notar como se distribuye esta jerarquía en la codificación CIE-10 para la enfermedad hepatitis.

Hepatitis viral			
B15	Hepatitis aguda tipo A		
B16	HEPATITIS AGUDA TIPO B	B160	Hepatitis aguda tipo B, con agente delta (coinfección) con coma hepático
		B161	Hepatitis aguda tipo B, con agente delta (coinfección), sin coma hepático
		B162	Hepatitis aguda tipo B, sin agente delta, con coma hepático
		B169	Hepatitis aguda tipo B, sin agente delta y sin coma hepático
B17	OTRAS HEPATITIS VIRALES AGUDAS	B170	Infección (superinfección) aguda por agente delta en el portador de hepatitis B
B18	HEPATITIS VIRAL CRÓNICA	B180	Hepatitis viral tipo B crónica con agente delta
		B181	Hepatitis viral tipo B crónica sin agente delta
B19	HEPATITIS VIRAL SIN OTRA ESPECIFICACIÓN	B190	Hepatitis viral no especificada con coma
		B199	Hepatitis viral no especificada sin coma
K73	Hepatitis crónica, no clasificada en otra parte		

Figura 1.1: Los códigos correspondientes a la hepatitis de origen viral

La tarea de codificación CIE es por naturaleza compleja de evaluar, ya que consiste en una clasificación multi-etiqueta sobre una estructura de árbol, donde tanto la distancia y la ubicación en el árbol, de dos nodos dados, tiene diferentes significados. Además del mecanismo de extracción de información, varios los autores proponen el uso de clasificadores que muestran relativamente buena capacidad de predicción basado en la información contenida en archivos médicos. En particular, los árboles de decisión [7], clasificadores Naïve Bayes [13] , regresión en cresta [22] y máquinas de soporte vectorial [4, 8, 23] son empleados en estos enfoques .

1.1. Planteamiento del Problema

La asignación manual de códigos CIE-10 para diagnósticos médicos es una labor muy costosa y propensa a errores debido al gran número de códigos y a la complejidad de las reglas de asignación. Esta tarea utiliza diferentes normas de codificación que son extremadamente complejas y sólo codificadores expertos capacitados pueden realizarla de manera adecuada, haciendo que el proceso de codificación de documentos sea costoso y poco fiable. Un codificador debe seleccionar uno de miles de códigos

CIE para indexarlo a cada diagnóstico médico. El volumen de diagnósticos médicos es abrumador y la capacidad de clasificación manual resulta siempre muy costosa pues se necesitan de varios etiquetadores para procesar los diagnósticos que se generan diariamente en un hospital y esto conlleva a un significativo retraso en la tarea de asignación.

1.2. Objetivos

Objetivo general

Diseñar un método de clasificación para la asignación de códigos CIE-10 a diagnósticos médicos aprovechando la información contextual.

Objetivos específicos

- Estudiar las interrelaciones y características del conjunto de datos.
- Analizar diferentes representaciones para soportar errores de escritura.
- Proponer un sistema de asignación de códigos CIE-10 para la clasificación de diagnósticos médicos.

1.3. Solución propuesta

Se espera obtener un método de clasificación que sirva de soporte a los expertos etiquetadores en la asignación de códigos CIE-10-MC a diagnósticos médicos. De esta manera, cuando se ejecute la clasificación de un diagnóstico médico el método deberá brindar las k clases (códigos CIE) más similares para la asignación del código CIE.

1.4. Organización de la tesis

El resto del documento se organiza de la siguiente manera: el capítulo 2 describe los conceptos básicos sobre la clasificación y asignación de códigos CIE a diagnósticos médicos así como su proceso y una breve explicación de los métodos utilizados en la recuperación de información y fusión de datos; el capítulo 3 muestra los trabajos que han abordado el problema y sus diferentes características; en el capitulo 4 se describe el método propuesto y sus características principales; en el capítulo 5 se presentan los diversos experimentos desarrollados durante la investigación así como los resultados obtenidos; en el capitulo 6 se presentan las conclusiones y las posibles lineas de trabajo futuro que se desprenden de nuestros resultados.

Capítulo 2

Marco teórico

La investigación presentada en este documento se relaciona con la tarea de asignación de códigos CIE-10 a diagnósticos médicos. A continuación se detallan los conceptos básicos de la tarea así como los métodos involucrados en la investigación para la correcta interpretación y entendimiento del método desarrollado. así como d los experimentos realizados y los resultados obtenidos.

2.1. CIE-10-MC

La Decima Revisión de la Clasificación Internacional de enfermedades (CIE-10) proporciona un estándar para la codificación de historiales clínicos. El sistema de codificación se basa en directrices de la Organización Mundial de la Salud. Un código CIE-10-MC indica una clasificación de una enfermedad, síntoma, procedimiento, lesión, o información del historial personal de cada paciente. Los códigos se organizan jerárquicamente, donde las entradas de primer nivel son agrupaciones generales (Por ejemplo, enfermedades del sistema respiratorio) y los códigos de nivel inferior indican síntomas específicos o enfermedades y su ubicación (por ejemplo, "la neumonía por aspergilosis"). Cada código específico y de bajo nivel consiste de 4 o 5 dígitos, con un decimal después de la tercera. Los códigos de nivel superior incluyen típicamente sólo 3 dígitos. En general, hay miles de códigos que cubren una amplia gama de condiciones médicas. Aunque se usa principalmente para propósitos de facturación en los hospitales, los códigos CIE-10 también pueden ser útiles para detección de

epidemias y para registrar el desarrollo del problema de un paciente. Los códigos CIE-10 de un historial clínico se determinan basados en el relato de dicho registro (diagnóstico).
En la tabla 2.1 podemos observar la clasificación de códigos CIE, la cual se encuentra clasificada por categorías y cada categoría representa un grupo de enfermedades. Cada código principal comienza con una letra y es seguida por caracteres numéricos. Cuando la enfermedad es mas específica, esta se separa con un punto y le siguen tres caracteres numéricos. Con el ejemplo siguiente podemos apreciar la distribución de los códigos CIE.

- A00-B99 Certain infectious and parasitic diseases

  - A00-A09 Intestinal infectious diseases
    - A00 Cholera
      - A00.0 is a billable ICD-10-CM diagnosis code A00.0 Cholera due to Vibrio cholerae 01, biovar cholerae
      - A00.1 is a billable ICD-10-CM diagnosis code A00.1 Cholera due to Vibrio cholerae 01, biovar eltor
      - A00.9 is a billable ICD-10-CM diagnosis code A00.9 Cholera, unspecified

2.2. Clasificación de Textos

Categorización de texto (también conocido como clasificación de texto) es la tarea de asignar categorías predefinidas a documentos de texto libre. Puede proporcionar puntos de vista conceptuales de las colecciones de documentos y tiene importantes aplicaciones en el mundo real. Por ejemplo, las noticias suelen estar organizadas por categorías de temas o códigos geográficos; trabajos académicos se clasifican a menudo por los dominios técnicos y subdominios; informes de los pacientes en las organizaciones de atención de salud son a menudo indexados desde varios aspectos, el uso

Tabla 2.1: Categorías de la clasificación CIE

Capítulo	**Códigos**	**Título**
I	A00-B99	Ciertas enfermedades infecciosas y parasitarias.
II	C00-D48	Neoplasias.
III	D50-D89	Enfermedades de la sangre y de los órganos hematopoyéticos.
IV	E00-E90	Enfermedades endocrinas, nutricionales y metabólicas.
V	F00-F99	Trastornos mentales y del comportamiento.
VI	G00-G99	Enfermedades del sistema nervioso.
VII	H00-H59	Enfermedades del ojo y sus anexos.
VIII	H60-H95	Enfermedades del oído y de la apófisis mastoides.
IX	I00-I99	Enfermedades del sistema circulatorio.
X	J00-J99	Enfermedades del sistema respiratorio.
XI	K00-K93	Enfermedades del aparato digestivo.
XII	L00-L99	Enfermedades de la piel y el tejido subcutáneo.
XIII	M00-M99	Enfermedades del sistema osteomuscular y del tejido conectivo.
XIV	N00-N99	Enfermedades del aparato genitourinario.
XV	O00-O99	Embarazo, parto y puerperio.
XVI	P00-P96	Ciertas afecciones originadas en el periodo perinatal.
XVII	Q00-Q99	Malformaciones congénitas, deformidades y anomalías cromosómicas.
XVIII	R00-R99	Síntomas, signos y hallazgos anormales clínicos y de laboratorio.
XIX	S00-T98	Traumatismos, envenenamientos y algunas otras consecuencias.
XX	V01-Y98	Causas externas de morbilidad y de mortalidad.
XXI	Z00-Z99	Factores que influyen en el estado de salud.
XXII	U00-U99	Códigos para situaciones especiales.

de taxonomías de categorías de enfermedades, tipos de procedimientos quirúrgicos, códigos de reembolso de seguro, etc. Otra aplicación generalizada de la categorización de textos es el filtrado de correo no deseado, donde los mensajes de correo electrónico se clasifican en las dos categorías de *spam* y *no-spam*, respectivamente.

2.3. Representación de documentos

El primer paso en la categorización de texto es transformar los documentos, que típicamente son cadenas de caracteres, en una representación adecuada para el algoritmo de aprendizaje y la tarea de clasificación. Se sugiere que las palabras funcionan bien como unidades de representación y que su orden en un documento es de menor importancia para muchas tareas [19]. Esto conduce a una representación de texto atributo-valor. Cada palabra distinta corresponde a una función, con el número de veces que la palabra se produce en el documento como su valor.

Pre-procesamiento

La transformación de texto suele ser de los siguientes tipos:

- Eliminar etiquetas HTML.
- Eliminar palabras vacías
- Lematización de palabra.

Las **palabras vacías** son palabras frecuentes que no llevan a ninguna información como pronombres, preposiciones, conjunciones, etc.
Lematización es un método para reducir una palabra a su raíz o (en inglés) a un stem o lema, por ejemplo doctor <- doctores, doctora.

Indexado

La representación de un documento más comúnmente utilizada es el llamado modelo de espacio vectorial. En el modelo de espacio vectorial los documentos están representados por vectores de palabras. Por lo general, se tiene una colección de

documentos que está representada por una matriz A palabra-documento donde cada entrada representa las apariciones de una palabra en un documento

$$A = a_{ik} \tag{2.1}$$

donde a_{ik} es el peso de la palabra i en el documento k, dado que cada palabra no suele aparecer en cada documento la matriz A es por lo general dispersa. El número de filas M de la matriz corresponde con el número de palabras en el vocabulario M puede ser muy grande. De ahí la característica más importante, o dificultad, en los problemas de categorización de texto es la alta dimensionalidad del espacio de características.
Hay varias formas de determinar el peso a_ik de la palabra i en el documento k pero la mayoría de los enfoques se basan en dos observaciones empíricas respecto al texto:

- Cuantas más veces aparece una palabra en un documento, esta es más relevante para el tema del documento.
- Cuantas más veces una palabra aparece en todos los documentos de la colección, esta es menos discriminante entre los documentos.

Sea f_{ik} la frecuencia de la palabra i en el documento k, N el número de documentos en la colección, M el número de palabras en la colección después de la eliminación de palabras vacías y la lematización de palabras y n_i el número total de veces que la palabra i se produce en toda la colección. A continuación se describe los principales esquemas de pesado

- Booleano: Es el pesado más simple, representa la presencia (peso 1) o ausencia (peso 0) de un término en el documento.

$$a_{ik} = \begin{cases} 1 & \text{if } f_{ik} > 0 \\ 0 & \text{en otro caso} \end{cases}$$

- Frecuencia de términos: Consiste en asignar el número de veces que un término ocurre en el documento

$$a_{ik} = f_{ik} \tag{2.2}$$

Los dos esquemas anteriores no toman en cuenta la frecuencia de la palabra en todos los documentos de la colección.

- TF.IDF. Un enfoque conocido para el cálculo de los pesos de palabras es el pesado $tfidf$ que asigna el peso a la palabra i en el documento k en proporción al número de apariciones de la palabra en el documento y en proporción inversa al número de ocurrencias de la palabra en los documentos de la colección.

$$a_{ik} = f_{ik} * log(\frac{N}{n_i}) \tag{2.3}$$

2.4. Aprendizaje automático

En ciencias de la computación el aprendizaje automático o aprendizaje de máquinas es una disciplina científica que explora la construcción y estudio de algoritmos que puedan aprender de un conjunto de datos [14] . Tales algoritmos operan mediante la construcción de un modelo basado en las entradas [3], utilizando dicha información para hacer predicciones o decisiones, en lugar de seguir instrucciones programadas explícitamente.

El aprendizaje automático se puede considerar un subcampo de la ciencia y la estadística. Tiene fuertes lazos con la inteligencia artificial y la optimización, que ofrecen métodos, teorías y aplicaciones. El aprendizaje automático se emplea en una amplia gama de tareas de computación, donde el diseño y la programación de algoritmos explícitos, basados en reglas no es factible.Tiene una amplia gama de aplicaciones, incluyendo motores de búsqueda, diagnósticos médicos, detección de fraude en el uso de tarjetas de crédito, análisis del mercado de valores, clasificación de secuencias de ADN, reconocimiento del habla y del lenguaje escrito, juegos y robótica [20].

Tipos de algoritmos

Los diferentes algoritmos de Aprendizaje Automático se agrupan en una taxonomía en función de la salida de los mismos. Algunos tipos de algoritmos son:

Aprendizaje supervisado

El algoritmo produce una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema. Un ejemplo de este tipo de algoritmo es el problema de clasificación, donde el sistema de aprendizaje trata de etiquetar (clasificar) una serie de vectores utilizando una entre varias categorías (clases). La base de conocimiento del sistema está formada por ejemplos de etiquetados anteriores. Este tipo de aprendizaje puede llegar a ser muy útil en problemas de investigación biológica, biología computacional y bioinformática.

Aprendizaje no supervisado

Todo el proceso de modelado se lleva a cabo sobre un conjunto de ejemplos formado tan sólo por entradas al sistema. No se tiene información sobre las categorías de esos ejemplos. Por lo tanto, en este caso, el sistema tiene que ser capaz de reconocer patrones para poder etiquetar las nuevas entradas.

Aprendizaje semisupervisado

Este tipo de algoritmos combinan los dos algoritmos anteriores para poder clasificar de manera adecuada. Se tiene en cuenta los datos marcados y los no marcados.

Aprendizaje por refuerzo

El algoritmo aprende observando el mundo que le rodea. Su información de entrada es el feedback o retroalimentación que obtiene del mundo exterior como respuesta a sus acciones. Por lo tanto, el sistema aprende a base de ensayo-error.

Transducción

Similar al aprendizaje supervisado, pero no construye de forma explícita una función. Trata de predecir las categorías de los futuros ejemplos basándose en los ejemplos de entrada, sus respectivas categorías y los ejemplos nuevos al sistema.

Aprendizaje multi-tarea

Métodos de aprendizaje que usan conocimiento previamente aprendido por el sistema de cara a enfrentarse a problemas parecidos a los ya vistos.

2.4.1. Algoritmos de aprendizaje en la categorización de textos

Los algoritmos de aprendizaje computacional de clasificación son métodos que, dado un conjunto de ejemplos de entrenamiento, infieren un modelo de las categorías en las que se agrupan los datos, de tal forma que se pueda asignar a nuevos ejemplos una o más categorías de manera automática.
Diferentes algoritmos se han usado en la tarea de clasificación de textos, los algoritmos utilizados en este trabajo pertenecen al paradigma del aprendizaje supervisado por ejemplo, Naive Bayes un clasificador probabilista y K-NN, basado en instancias, los cuales son descritos a continuación.

Naive Bayes

El algoritmo de clasificación Naive Bayes es un clasificador probabilístico. Se basa en modelos de probabilidades que incorporan fuertes suposiciones de independencia. La idea de usar el teorema de Bayes en cualquier problema de aprendizaje automático (en especial los de clasificación) es que podemos estimar las probabilidades a posteriori de cualquier hipótesis consistente con el conjunto de datos de entrenamiento para así escoger la hipótesis más probable. Para estimar estas probabilidades se han propuesto numerosos algoritmos, entre los que cabe destacar el algoritmo Naive Bayes.
Dado un ejemplo x representado por k valores el clasificador naive Bayes se basa en encontrar la hipótesis más probable que describa a ese ejemplo. Si la descripción de ese ejemplo viene dada por los valores $<a_1$, a_2, ... ,$a_1 n >$, la hipótesis más probable será aquella que cumpla:

$$v_M AP = argmax_(v_j \epsilon V) P(v_j | a_1, \ldots a_n) \tag{2.4}$$

es decir, la probabilidad de que conocidos los valores que describen a ese ejemplo, éste pertenezcan a la clase v_j (donde v_j es el valor de la función de clasificación $f(x)$ en el conjunto finito V). Por el teorema de Bayes:

$$v_{MAP} = argmax_{v_j \epsilon V} \frac{P(a_1, ...a_n|v_j)p(v_j)}{P(a_1, ...a_n}) = argmax_{v_j \epsilon V} P(a_1, ...a_n|v_j)p(v_j) \quad (2.5)$$

Podemos estimar $P(v_j)$ contando las veces que aparece el ejemplo v_j en el conjunto de entrenamiento y dividiéndolo por el número total de ejemplos que forman este conjunto. Para estimar el término $P(a_1, \ldots a_n|v_j)p(v_j))$, es decir, las veces en que para cada categoría aparecen los valores del ejemplo x, se debe recorrer todo el conjunto de entrenamiento. Este cálculo resulta impracticable cuando no se tienen suficientes ejemplos para calcular sus probabilidades por lo que se hace necesario simplificar la expresión. Para ello se recurre a la hipótesis de independencia condicional con el objeto de poder factorizar la probabilidad [**?**]. Esta hipótesis dice lo siguiente: Los valores a_j que describen un atributo de un ejemplo cualquiera x son independientes entre sí conocido el valor de la categoría a la que pertenecen. Así la probabilidad de observar la conjunción de atributos a_j dada una categoría a la que pertenecen es justamente el producto de las probabilidades de cada valor por separado:

$$P(a_1, ...a_n|v_j) = \Pi_i P(a_i|v_j) \quad (2.6)$$

K- Vecinos más cercanos

Es un método de aprendizaje basado en instancias, también conocido como lazy learning o memory learning, debido a que los datos de entrenamiento se procesan solo hasta que se requiere una predicción y la relevancia de los datos se mide en función de una medida de distancia. Este método de clasificación ha sido usado en el análisis de texto. La idea principal de este método (knn), es almacenar el conjunto de entrenamiento (D_t), de tal modo que al clasificar un nuevo documento se busca en los documentos almacenados los k casos más cercanos (similares) usando una medida de similaridad y se le asigna al documento la clase de la mayoría de estos k documentos.

Un caso se clasifica por mayoría de votos de sus k vecinos. Si K = 1, entonces entonces simplemente se asigna a la clase de su vecino más cercano.

Los ejemplos de entrenamiento (D_t)son los vectores en un espacio de características multidimensionales, cada uno con una etiqueta de clase. La fase de entrenamiento del algoritmo consiste sólo en el almacenamiento de los vectores de características y etiquetas de clase de las muestras de entrenamiento.

En la fase de clasificación, k es una constante que representa el número de vecinos mas cercanos del vector a clasificar (una consulta o punto de prueba. DE esta manera se realiza la asignación de la etiqueta(clase) que es más frecuente entre las muestras (k muestras)de entrenamiento más próximo al punto de consulta definido por el usuario.

$$Knn(d) = argmax \sum_{\substack{d' \epsilon NN_k^L(d) \\ y(d')=c}} sim(d, d') \tag{2.7}$$

En la fórmula 2.7 se tiene a U como el conjunto de documentos sin clasificar, y L el conjunto de documentos etiquetados. Dado un documento $d \epsilon U$ se tiene $NN_k^L(d)$ como el conjunto de k documentos en L que son más similares a d con respecto a alguna medida de similitud $sim(d, d')$. d es asignada a la categoría de la mayoría de los documentos in $NN^L{}_k(d)$. Una medida de similaridad usada en la clasificación de texto es la similaridad coseno, la cual está dada por la siguiente ecuación:

$$sim(d_i, d_j) = \frac{(d_i.d_j)}{(\|d_i\| \times \|d_j\|)} = \frac{\sum_{i=1}^{n} d_i \times d_j}{\sqrt{\sum_{i=1}^{n}(d_i)^2} \times \sqrt{\sum_{i=1}^{n}(d_j)^2}} \tag{2.8}$$

donde d_i y d_j son los vectores de los documentos a comparar.

La similitud resultante varía de -1, lo cual significa que los vectores son exactamente contrarios, a 1, que significa que los vectores son exactamente lo mismo.

2.5. Fusión de datos

La eficacia de los sistemas de Recuperación de información (RI) para satisfacer las necesidades de información (peticiones), normalmente se obtiene aplicando medidas

de evaluación a nivel global. Lo anterior significa que, dado un conjunto de peticiones de prueba y una colección de búsqueda, la eficacia global del sistema se determina por el promedio de la eficacia obtenida en cada una de las peticiones. Con el creciente aumento de información y la variedad de la misma (documentos,imágenes, vídeo y audio), las investigaciones en RI han propiciado modificaciones al esquema básico. Las modificaciones más utilizadas se basan en múltiples formulaciones de la consulta procesada por un único sistema de RI, y la utilización de múltiples sistemas de RI para procesar una misma formulación de la consulta. Otra posibilidad, aunque no muy investigada actualmente, es una combinación de las dos modificaciones anteriores, es decir, tener múltiples formulaciones de una consulta y procesarlas con múltiples sistemas de RI. Algo común en los tres procesos modificados de RI es que la salida de estos procesos son diferentes listas de resultados. Dado que se debe entregar una sola lista de resultados al usuario para satisfacer su necesidad de información, estos esquemas deben decidir qué lista de resultados arrojadas por su proceso de IR debe entregarse. Usualmente, la decisión se toma con base en la observación de la eficacia de las distintas configuraciones. La lista de resultados seleccionada es aquella que obtiene los mejores resultados de recuperación en un conjunto de peticiones de prueba. Como se comentó anteriormente, una opción para aprovechar múltiples resultados de recuperación para una misma consulta, obtenidos mediante diferentes reformulaciones y/o sistemas de RI, es la Fusión de Datos.

Esta estrategia genera una nueva lista de resultados al combinar, mediante diferentes procedimientos, los resultados de recuperación disponibles. En la figura 2.1 podemos observar el proceso de fusión, donde tenemos las dos características necesarias para la fusión de datos: un conjunto de listas (l_1, l_2 y l_3) y los elementos repetidos en las listas (a,c y e).

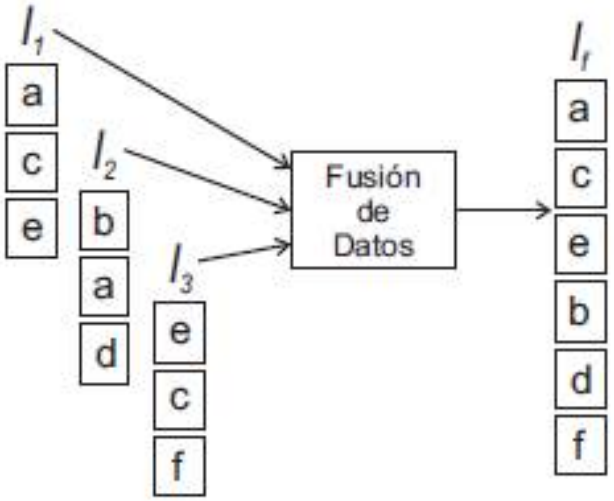

Figura 2.1: Proceso de Fusión de Datos

Definición de la tarea

Hsu y Taksa [6] definen la tarea de Fusión de Datos de la siguiente manera:

> Fusión de Datos es un proceso (adquisición, diseño e interpretación) de combinación de información obtenida por múltiples agentes (fuentes, esquemas, sensores o sistemas) en una simple representación (o resultado).

La Fusión de Datos ha sido utilizada en campos como el reconocimiento de patrones y aprendizaje automático, tomando como información a fusionar las diferentes características extraídas; en detección de señales, rastreo de blancos, procesamiento de imágenes, vigilancia y aplicaciones militares, donde los datos arrojados por sensores son la información que se fusiona.
En el área de Recuperación de Información, el concepto de Fusión de Datos ha sido utilizado para estudiar la combinación de múltiples resultados, obtenidos de diferentes formulaciones de la petición o de diferentes sistemas. Existen diferentes esquemas que pueden ofrecer diferentes resultados de una misma petición.

2.5.1. Métodos de fusion lineales

Estos méodos utilizan los valores de relevancia dados por los sistemas de RI a los elementos en la listas, para utilizarlos en una función lineal. Tres son los méetodos básicos de esta clase de métodos: MaxRSV, MinRSV y SumRSV (RSV son las siglas de $RetrievalScoreValue$ o también puede encontrarse como $RawScoreValue$). El funcionamiento de estos méetodos se basa en la redundancia de los elementos en las listas, sin embargo, el valor de similitud dado por los sistemas de recuperación de información juega un papel fundamental ya que determina la posición de los elementos en la lista final.

Métodos de fusión posicionales

Estos métodos descartan completamente los RSVs (RSV son las siglas de Retrieval Score Value o también puede encontrarse como Raw Score Value) de los elementos en las listas y calculan nuevos pesos iniciales a partir de diferentes estrategias.
El método CombMNZ ha demostrado ser el mejor de esta clase [**?**].

CombMNZ

Sea $L = \{l_1, ..., l_i, ..., l_m\}$ un conjunto de listas de resultados ordenados de la forma $l_i =< d_1, ..., d_j, ..., d_n >$ donde d_j representa a un documento recuperado. Sea $D = \{l_1 \cup ... \cup l_i \cup ... \cup l_m\}$ el conjunto unión de los elementos en las listas de L. El nuevo peso para los elementos en D se define como:

$$CombMNZ(d_j) = (\sum_{i=1}^{L} e(d_j, l_i))(\sum_{i=1}^{L} s(d_j, l_i)) \tag{2.9}$$

donde

$$e(d_j, l_i) = \begin{cases} 1 & \text{si } d_j \in l_i \\ 0 & \text{en caso contrario} \end{cases}$$

$$s(d_j, L_i) =\mid l_i \mid -r(d_j, l_i) + 1 \tag{2.10}$$

y $r(d_j, l_i)$ es la posición del elemento d_j en la lista l_i.

Por ejemplo, se tienen dos listas de resultados ordenados, $l_1 = a$,c,b, d y $l_2 = c$,b,d,a. Cada elemento en la lista tiene asignado un peso basado en la posición de la lista. En la lista l_1 el peso para el elemento a sería: $s(a, l_1) = |4| - 1 + 1$, donde $|l_1| = 4$, el total de elementos de la lista l_1, $-r(a, l_1) = 1$ ya que el elemento a esta el primera posición. Así el peso de cada elemento es:

- $s(a, l_1) = 4$
- $s(b, l_1) = 2$
- $s(c, l_1) = 3$
- $s(d, l_1) = 1$

Y de la misma forma se asignan los pesos para la lista l_2. Después se realiza la unión de todos los elementos en una lista final L en la cual sumamos los valores de todos los elementos y reordenamos de acuerdo a su peso. Nuestra lista final quedaría de la siguiente manera: $L = c(7), a(5), d(4) y b(3)$.

2.6. Evaluación de clasificadores

En el contexto de recuperación de información, la precisión, el recuerdo(recall) y la medida-f (f-measure), se definen en términos de un conjunto de documentos recuperados (por ejemplo, la lista de documentos producidos por un motor de búsqueda en la web para una consulta) y un conjunto de documentos relevantes (por ejemplo, la lista de los documentos en la Internet que son relevantes para un tema determinado).

Precisión

En el campo de la recuperación de la información, la precisión es la fracción de los documentos recuperados que son relevantes para la búsqueda:

$$precision = \frac{| \{documentos\ relevantes\} \cap \{documentos\ recuperados\} |}{| \{documentos\ recuperados\} |} \quad (2.11)$$

La Precisión toma todos los documentos recuperados en cuenta, pero también se puede evaluar a un rango de corte determinado, teniendo en cuenta sólo los resultados más altos que devuelve el sistema. Esta medida se llama precisión en n o $P@n$, donde n equivale al rango de corte de la lista de documentos recuperados.
Por ejemplo, para una búsqueda de texto en un conjunto de documentos de precisión es el número de resultados correctos, dividido por el número de todos los resultados devueltos.
El significado y el uso de precisión en el campo de la recuperación de información difiere de la definición de exactitud y precisión dentro de otras ramas de la ciencia y la tecnología.

Recuerdo

Recall en recuperación de información es la fracción de los documentos que son relevantes para la consulta que se recuperan con éxito.

$$recall = \frac{\mid \{documentos\ relevantes\} \cap \{documentos\ recuperados\} \mid}{\mid \{documentos\ relevantes\} \mid} \tag{2.12}$$

Por ejemplo para la búsqueda de un texto en un conjunto de documentos, el recall es el número de resultados correctos, dividido por el número de resultados que deberían haber sido devueltos.
Otras medidas conexas utilizadas en la clasificación incluyen la especificidad (True negative rate) y la exactitud (Accuracy).

$$True\ negative\ rate = \frac{vn}{vn + fp} \tag{2.13}$$

$$Accuracy = \frac{vp + vn}{vp + vn + fp + fn} \tag{2.14}$$

donde vn son los verdaderos negativos , $vpsonlosverdaderospositivos$, fp son los falsos positivos y fn son los falsos negativos.

Medida-F

La medida-F es la media armónica de la precisión y el recuerdo. Se considera tanto la precisión p y el recuerdo en el conjunto de prueba para calcular la puntuación: p es el número de resultados correctos dividido por el número de todos los resultados devueltos y r es el número de resultados correctos dividido por el número de resultados que debieron haber sido devueltos. La medida-F puede ser interpretado como una media ponderada de la precisión y el recuerdo, entonces la medida-F alcanza su mejor puntuación en 1 y su peor puntuación en 0.

$$F = 2 * \frac{precision * recall}{precision + recall} \tag{2.15}$$

Rango Reciproco Medio

El Rango Reciproco Medio o en en sus siglas en ingles MRR es una medida estadística para evaluar cualquier proceso que genera una lista de posibles respuestas a una muestra de consultas, ordenada por probabilidad de exactitud. El rango recíproco de una respuesta a una consulta es el inverso multiplicativo de la fila de la primera respuesta correcta. Es el promedio de los rangos recíprocos de resultados para una muestra de consultas Q.

$$MRR = \frac{1}{\mid Q \mid} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{2.16}$$

Por ejemplo, en la tabla 2.2tenemos tres consultas para un sistema que trata de trasladar las palabras en inglés en sus plurales. En cada caso, el sistema hace tres suposiciones, el sistema cree que la primera opción es la más probable:

Tabla 2.2: Ejemplo de un sistema de consulta

Query	Resultado	Respuesta correcta	Rank	Reciprocal rank
cat	catten, cati, **cats**	cats	3	1/3
torus	torri, tori, **toruses**	tori	2	1/2
virus	**viruses**, virri, viri	viruses	1	1

Dada los tres ejemplos, se puede calcular el MRR como (1/3 + 1/2 + 1)/3 = 11/18 o 0.61.

Capítulo 3

Estado del arte

3.1. Importancia de la automatización de códigos CIE.

La asignación manual de códigos CIE a diagnósticos médicos implica la revisión humana de la documentación clínica para identificar los códigos aplicables. Los códigos pueden ser asignados inmediatamente, pero en la mayoría de los casos, especialmente para los pacientes que requieren hospitalización, los códigos son asignados después de que un experto revisa la documentación médica (notas médicas, informes de laboratorio, etc.) creados durante la visita del paciente. Es decir, un codificador experto lee la documentación y, basado en el conocimiento médico, directrices, reglamentos y la experiencia, asigna uno o más códigos CIE a la visita del paciente. Cuando la aplicación de un esquema de codificación es compleja, el proceso puede ser asistido por el uso de libros de códigos, haciendo uso de las listas abreviadas, o aplicaciones que facilitan las búsquedas alfabéticas y proporcionan ediciones y consejos. La asignación de códigos puede ser llevada a cabo por los médicos, pero a menudo se lleva a cabo por otros miembros del personal, tales como profesionales de codificación. Tres de cada cuatro médicos reporta utilizar Registros Electrónicos de Salud (RES), el volumen de datos disponibles está creciendo rápidamente. Además de los beneficios de la tecnología de información de salud para la atención al paciente, la mayor importancia se encuentra en manos del análisis secundario de estos datos.
Los códigos de diagnóstico, por ejemplo, se utilizan en la RES como un mecanismo

de facturación. Pero estos códigos también han demostrado ser fundamentales en los esfuerzos de fenotipado y modelización predictiva de los estados del paciente. La codificación de diagnósticos médicos se basa en la Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados de la Salud (CIE comúnmente abreviado), creado por la Organización Mundial de la Salud (OMS) en 1977. En el escenario considerado, cuando un paciente recibe un servicio médico se le asigna un código CIE. El hecho de que los sistemas de información clínica pueden mejorar la atención médica y reducir los costos de salud ha estado en la agenda académica desde hace bastante tiempo. No obstante, los datos del paciente hoy en día siguen siendo almacenados en forma narrativa en muchos hospitales, lo que produce una gran cantidad de información que, más allá de la visita clínica, tiene una utilidad limitada debido a su alto volumen y baja accesibilidad. Sin embargo, los intentos de abordar el problema de procesamiento de texto libre han dado lugar a la demanda de aplicaciones que simulan y complementan lo que las personas son capaces de hacer. La American Health Information Management Association (AHIMA), ha convocado la exploración de la codificación asistida por computadora, informa que este flujo de trabajo de codificación manual es caro e ineficiente en una industria donde las necesidades de datos nunca han sido mayores. "La industria necesita soluciones automatizadas para permitir que el proceso de codificación se convierta en más productivo, eficiente, preciso y consistente" [**?**].

3.2. Características principales observadas en la asignación de códigos CIE.

Existen diversos aspectos observados en la revisión de trabajos relacionados a la asignación de códigos CIE que son integrados al análisis y procesamiento automático de textos médicos para mejorar la representación de las instancias de un corpus así como la detección de patrones o reglas que mejoran la discriminación entre clases.

1. Pre-procesamiento. Las notas médicas empleadas para la asignación de códigos comúnmente llevan un proceso de pre-procesamiento como la eliminación de

palabras vacías, puntuaciones, números, la conversión del texto en minúscula y la búsqueda de la raíz de cada palabra, corrección de ortografía y expansión de abreviaturas para obtener la normalización de todos los términos médicos.

2. Manejo del vocabulario. Consiste en elegir términos médicos con ayuda de tesauros para obtener diccionarios controlados, anexar sinónimos o hiperónimos a las instancias estudiadas para obtener toda la información posible.

3. Uso de palabras clave. Trata de la detección de palabras específicas que ayudan a detectar aspectos importantes en cada instancia, como negaciones, especulaciones, incertidumbre o términos clave que se relacionen directamente con una clase.

4. Similitud en textos. Se trata de buscar la ocurrencia de secuencias de palabras o términos de cada texto médico en tesauros médicos referentes a códigos CIE, términos en común con las clases establecidas o frecuencias de n-gramas en el conjunto de entrenamiento, para establecer niveles de confianza en la clasificación.

5. Interdependencia de clases. Anteriormente se explica que la codificación CIE modela un árbol jerárquico entre las enfermedades y sus diferentes variedades. Es por ello que un rasgo presente es el estudio de las relaciones entre las diversas clases y subclases para la búsqueda de umbrales que ajusten la clasificación y la discriminación de las mismas.

6. Modelo vectorial. La representación de los términos médicos puede realizarse mediante la representación por n-gramas de caracteres o n-gramas de palabras. El pesado puede ser binario, únicamente la presencia o ausencia de un término, la frecuencias de término (ft), es el número de veces que un término ocurre en un documento, la frecuencia de documentos (df) que es el número de documentos que contienen un término dado dividido por el número total de documentos en un conjunto o la implementación del tf-idf que es la frecuencia inversa de un documento y es una estadística numérica que pretende reflejar la importancia de una palabra en un documento, en una colección o conjunto de datos.

3.3. La problemática de la asignación automática de códigos CIE

Dado el gran número de códigos y su nivel de especificidad para algunos diagnósticos, éste es un proceso que lleva mucho tiempo, siendo también propenso a errores; se estima que sólo del 60 % al 80 % de los códigos asignados reflejan verdaderamente el diagnóstico de los pacientes [2].
Hay un gran número razones por las cuales una mayor precisión y eficiencia son altamente necesarias en la asignación de códigos CIE. Éstos incluyen, la clara necesidad de mantener estadísticas precisas de las enfermedades, especialmente las que son importantes para la salud pública; la necesidad de aumentar la transparencia de los costos en el sistema sanitario; y la gran importancia de tener un sistema de codificación que puede ser de apoyo en la decisión de una asignación de código CIE. Para mitigar estos problemas, los investigadores han sido motivados a desarrollar herramientas para (parcialmente) automatizar este proceso mediante el aprovechamiento de los conocimientos en la asignación manual de códigos. Una gran parte de los datos de salud se almacena en formatos narrativos y no puede ser interpretada fácilmente por los ordenadores y se utiliza únicamente para apoyar las decisiones [2]. Se han desarrollado métodos para la automatización de códigos CIE y su aplicación ha sido probada en varias áreas, incluyendo la extracción de información de textos clínicos. Sin embargo, las características intrínsecas de los textos clínicos obstaculizan la extracción de conocimiento para permitir la aplicación de la lógica y la estadística computacional para procesar grandes volúmenes de datos [12]. A pesar de los problemas en la extracción de información de textos clínicos, numerosos estudios se han dirigido a apoyar la codificación clínica utilizando sistemas basados en tratamiento de texto. En algunos casos, los estudios informan de un buen poder predictivo, pero los retos fundamentales que plantea la codificación clínica permanecen parcialmente sin respuesta. Métodos de procesamiento de texto empleados por los autores no suelen ser generalizables debido a diversos aspectos:

- uso de diferentes idiomas.
- reglas específicas generadas para el conjunto de datos estudiado.

- reglamentaciones particulares de cada institución médica [5]
- diferencias de alcance (por ejemplo, rango de condiciones clínicas).
- la complejidad de problemas y estándares usados para la asignación de un código.
- dominio del departamento de salud asociado (radiología, cardiología).
- estándar de oro. La evaluación manual contiene aspectos críticos de cada etiquetador.
- evaluación del conjunto de datos: múltiples, dos o más, revisores independientes.
- metodologías de estudio varían ampliamente dependiendo del conjunto de datos establecido.
- aspectos de la captura del diagnóstico.
- la especificidad de los diagnósticos médicos.

Así, la mayoría de las enfermedades complementarias diagnosticadas por los expertos médicos se registran en la historia personal como texto libre. La terminología usada para describir las enfermedades en los registros médicos de cada paciente por lo general suele ser diferente a la empleada en la clasificación CIE con el fin de expresar la información más específica y detallada sobre el trastorno en particular o mediante paráfrasis que hace mucho más difícil la clasificación automática. El lenguaje de los médicos es fundamental, pero carece de la estructura y la claridad necesaria para el análisis del lenguaje natural. Estas anotaciones clínicas son densas con la jerga médica y siglas que a menudo tienen múltiples significados así como la reglamentación que cada hospital estipula para sus diversas actividades.

3.4. Trabajos relacionados en la asignación de códigos CIE a textos médicos

En los estudios revisados, los autores proponen enfoques utilizando métodos de Procesamiento de Lenguaje Natural (PLN) para extraer información de textos médicos y utilizar la misma para predecir asignación de código. Dos tendencias principales se identifican: mientras que algunos estudios utilizan métodos para sugerir códigos de manera directa o basada en reglas, un enfoque más frecuente consiste en extraer conceptos de narraciones con el propósito de lograr una mejor representación del vector de características de cada instancia, empleando algoritmos de aprendizaje automático.

En [21] se presenta una división (Ver figura 3.1)entre los trabajos de acuerdo al enfoque del que parten para analizar los textos médicos. A lo largo de esta sección se explicará cada una de las divisiones así como las ventajas y desventajas y los principales trabajos relacionados a la asignación de códigos CIE a diagnósticos médicos.

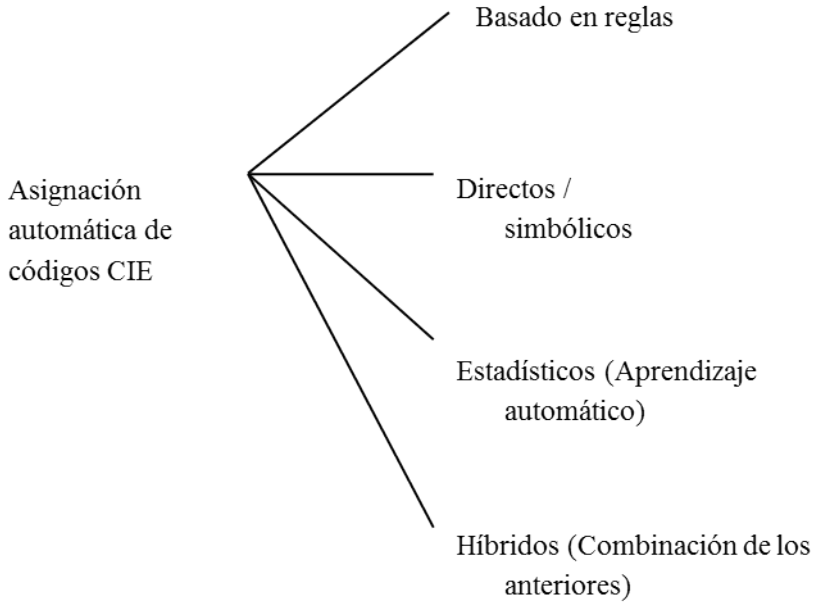

Figura 3.1: Esquema de división de métodos para asignación de códigos CIE

3.4.1. Enfoques basados en reglas.

Los enfoques basados en reglas se sustenta esencialmente en la creación de reglas deterministas para aplicar las asignaciones de códigos. En algunos casos, un conjunto de reglas es todo lo que se necesita. Las reglas son una excelente manera de encapsular ciertos tipos de conocimiento experto, tales como conocimientos lingüísticos para expresar la negación. Sin embargo, el análisis es difícil de codificar. Un ejemplo importante donde un sistema basado en reglas puede ser ventajoso es cuando no se integran nuevos ejemplos al sistema y no se tienen los suficientes documentos disponibles para formar un modelo estadístico. En este caso, una solución oportuna emplea un experto humano debidamente capacitado para intervenir y actualizar el sistema con las normas nuevas o revisadas. Quizás la ventaja más importante de este enfoque es que puede ser más rápido conseguir un sistema puesto en marcha que cubre los casos más comunes y luego mejorar la cobertura con el tiempo. Además, las normas están escritas en formas que solo los expertos entienden, así que es más fácil diagnosticar errores. La principal desventaja es que dichas normas no son generalizables y solo pueden ser utilizadas en un conjunto de datos específico y difícilmente podrá ser empleado a otro conjunto de datos. En [10], integra una parte del método con un conjunto de reglas/políticas específicas a cada clase para la asignación de códigos.

3.4.2. Enfoques Directos / simbólicos

Los sistemas directos utilizan como base de conocimiento ontologías de diversos tipos. Las principales herramientas provienen de UMLS (Unified Medical Language System), un conjunto de archivos y software que reúne diversos vocabularios, normas de salud y biomédicas para permitir la interoperabilidad entre sistemas informáticos. Uno de los usos de gran alcance de la UMLS es la vinculación de la información de salud, términos médicos, nombres de los medicamentos, y los códigos de facturación a través de diferentes sistemas informáticos. También se usan comúnmente sistemas que contienen términos CIE. Un sistema directo capta la similitud entre una entrada y los términos de estos diccionarios u ontologías y devuelve la clase con mayor grado de similitud. Para el cálculo de similitud existen diversas medidas en el campo de

procesamiento del lenguaje natural.
En [14] se emplea un sistema de indexado de diagnósticos utilizando la herramienta UMLS para el mapeo y extracción de términos médicos de vocabularios controlados y clasificaciones para ser mapeados con el MeSH (Medical Subject Headings), un vocabulario controlado integral y proveer una lista de los códigos CIE-10 con la incidencia de términos de cada diagnóstico.
El conjunto de datos empleado consta de 100 resúmenes de reportes hospitalarios escritos directamente por médicos, 50 de cardiología, y 50 de neumonía. Cada reporte contiene historial de enfermedades, procedimientos y prescripciones médicas. El sistema mapea los códigos CIE-10 con más incidencia de términos de acuerdo al diagnóstico dado y provee la lista de códigos más relevantes.
Este sistema también considera las prescripciones médicas del paciente y mediante una herramienta brindada por Le Vidal Company, provee un mapeo entre los medicamentos prescritos y los códigos CIE-10 relevantes mediante la fórmula 3.1.

$$Prescription\ Score\ (C_{(CIE-10)}) = (\frac{(x-1)}{N} + \frac{p}{N}) * 100 \tag{3.1}$$

En la fórmula (3.1) tenemos, $C_(CIE-10)$: es un código CIE-10 dado, x: sea el número de prescripciones médicas que tiene $C_(ICD-10)$ como indicio; p : prevalencia (=frecuencia) del código $C_(ICD-10)$ en el Rouen University Hospital; N : el número de drogas en el prescripción. La primera idea es que si un gran número de medicamentos con receta tiene la indicación de un mismo código de la CIE-10, este código debe probablemente ser extraído por dicho sistema. La segunda idea es que si este código es a menudo codificado por los médicos es más probable que pueda ser extraído de la información del paciente.

Este sistema no usa aprendizaje automático y únicamente se basa en clasificaciones de vocabularios y tesauros médicos para proveer los códigos CIE-10 más relevantes a cada término del diagnóstico pero no toma en cuenta las características de cada clase en el corpus empleado ni la información contextual para buscar patrones de clasificación. A diferencia de [14], en [11] Se agrega un método nombrado Extended Lexical Patterns (ELP) que realiza la extracción automática de etiquetas variantes CIE-9, consiste en dos pasos. En primer lugar, la transformación automática de una definición de clase en una extracción de términos. Detecta en cada texto una lista de

expresiones consideradas interesantes para la clase de inferencia. Después se busca coincidencias entre el diagnóstico a clasificar y las descripciones de las clases, este resultado es utilizado para la asignación. El primer paso se realiza una vez, mientras que el segundo debe repetirse para cada nuevo documento. La nomenclatura original se convierte automáticamente en transductores de estados finitos. Están hechas de: (a) los elementos léxicos de la etiqueta original de la clase; (b) otros artículos más genéricos, como códigos gramaticales o meta-etiquetas. El objetivo es aumentar la cobertura con elementos genéricos, preservando la buena precisión inducida por las unidades léxicas.

3.4.3. Enfoque Estadístico

Este tipo de enfoque se basa en que un sistema aprende las asignaciones como relaciones estadísticas de procesamiento de muchos ejemplos por ejemplo los resultados de filtros de spam de un correo electrónico. La exactitud de un modelo estadístico sube junto con el volumen de datos disponibles para el aprendizaje. Un modelo estadístico aprende en un entorno de producción los códigos más frecuentemente seleccionados. Los métodos estadísticos requieren grandes cantidades de datos anotados para entrenar y llegar a ser competentes. Pueden seguir aprendiendo con costos más bajos en comparación con los sistemas basados en reglas. Pero la exigencia de un conjunto de entrenamiento grande hace que sean inadecuados para situaciones con algunos ejemplos de datos. Un código raro que un hospital ve una vez al año, probablemente no es un buen candidato para la predicción estadística. Por otro lado, en un entorno ambulatorio con un conjunto de códigos más estrecho y un volumen más alto de codificación, el enfoque estadístico es un ajuste natural. La corrección de errores requiere encontrar nuevos rasgos lingüísticos para ayudar a discernir a estos motores estadísticos y establecer nuevas relaciones, lo cual es una tarea más compleja que mirar a casos extremos en una regla. En última instancia, sin embargo, la capacidad de un sistema para ser más preciso para los casos comunes y aprender de la producción es una gran ventaja en entornos dinámicos, como la atención de la salud. En [4, 8, 23] se clasifica usando el algoritmo SVM que devuelve la clasificación directa de diagnósticos médicos con su respectivo código CIE. En [23] se toma ventaja de las relaciones entre los diferentes códigos/clases CIE-9. Se añaden

atributos que representan las relaciones de los términos con las diferentes clases. En [8], se usa la información contextual, como: datos demográficos, diagnósticos, historia personal, alergias, recetas, medicamentos y evaluaciones; todos estos ligados a cada paciente para ser usados en la clasificación. Se comprueba que la incorporación del conocimiento del dominio ayuda al SVM a obtener un rendimiento de clasificación mejor que los enfoques usados comúnmente para clasificación multi-etiqueta. En algunos enfoques también se emplean algoritmos de aprendizaje en cascada que cubren aspectos diferentes en la clasificación para que al combinarse se obtengan mejores resultados. En [17] se describe un sistema en cascada que emplea instancias artificiales y conceptos UMLS agregados al conjunto de entrenamiento de entrenamiento usando un clasificador de mínimos cuadrados regularizados. Se emplea metatesauros para obtener hiperónimos, sinónimos y los términos médicos de cada instancia. Se crean identificadores para cada enfermedad y en el vector de característicaa se agrega el atributo de la negación e incertidumbre. Al conjunto de entrenamiento se suman los textos de las descripciones de cada clase del CIE-9.

3.4.4. Enfoques híbridos

Hoy en día, la mayoría de los sistemas de asignación de códigos son una combinación de los dos modelos, adoptando así un enfoque "híbrido". Cuando se presenta un volumen de datos en las decenas de miles de ejemplos, un enfoque estadístico tiende a superar el rendimiento de un determinado nivel de esfuerzo. Por otro lado, los pequeños tamaños de las muestras fomentan técnicas basadas en reglas. No es raro ver a un enfoque basado en normas que analiza partes de la oración como la negación o incertidumbre combinado con el análisis estadístico y la semántica. Una solución híbrida puede mezclar lo mejor de ambos mundos: la función de un enfoque basado en reglas con la escalabilidad mejorada a través de una amplia base de métodos estadísticos. Por lo menos, muchos investigadores clínicos e informáticos médicos concuerdan en que los enfoques basados en reglas y estadísticas " son complementarios". En [1, 10, 13] se implementan sistemas en cascada. En [10] se emplea tres enfoques para la clasificación de textos médicos referentes a reportes de radiología. Implementa a MIRA, un algoritmo de aprendizaje en línea, en el cual en cada documento de entrenamiento actualiza el peso del vector de acuerdo a la pérdida de

etiquetado, con respecto a la clase verdadera. La representación está conformada por n-gramas de palabras. El vector de características está formado por dos subconjuntos, los términos de cada diagnóstico médico y las características de transferencia, las cuales describen si un término es repetido en varias clases. Se crearon reglas que identifican si una instancia trata de una enfermedad o un síntoma, si existe negación, si el texto es totalmente idéntico a la descripción oficial de un código CIE y se agregan al vector.

- "instancia contiene un código de enfermedad"
- "instancia contiene un código de síntoma"
- "instancia contiene la palabra ñeumonía"'
- "instancia contiene un negación suave"

Se agregan instancias artificiales al corpus de entrenamiento obtenidas de las descripciones oficiales de la clasificación CIE-9 y se envían a un algoritmo en línea. El segundo enfoque implementa un sistema de clasificación que no requiere entrenamiento y solo mapea los términos de cada instancia con la descripción oficial de un código CIE-9 y devuelve los códigos más similares. El tercer enfoque crea una serie de políticas específicas a cada clase para la asignación de códigos CIE a cada instancia.

- Buscar por palabras clave como "cough", "fever".
- Si "pneumonia" aparece sin ninguna negación aplicar el código de pneumonia.

En [1] dada una entrada, tres sistemas de codificación seleccionan códigos de salida con sus puntuaciones de confianza, mientras que un selector basado en un árbol de decisión, C4.5 decide cuál es la mejor salida al utilizar la información tanto de la puntuación de confianza de cada sistema y las estadísticas de entrada. En [13] se desarrolla un sistema de codificación semi-automático que combina un método basado en reglas y un módulo de aprendizaje automático empleando un clasificador Naïve

Bayes. Se utiliza una base de datos del Mayo Clinic con documentación de la historia clínica de cada paciente; contiene la historia clínica de enfermedades y padecimientos, medicaciones actuales, signos vitales y propósitos de la visita. Para este proyecto se usa el HICDA (Adaptación de Clasificación de Enfermedades Hospitalarias), el cual es una adaptación del CIE-8, (Clasificación internacional de enfermedades 8a edición), un esquema propio de dicho hospital. Autocoder (nombrado por el autor) se enfoca en realizar dos tipos de procedimientos, clasificar directamente diagnósticos médicos con un alto grado de confianza basada en dos parámetros establecidos:

- MIN_EVENT_FREQ, frecuencia mínima del evento. Es el umbral relacionado a la frecuencia del evento. Si la frecuencia cae debajo de este umbral las categorías asociadas con este diagnóstico son consideradas correctas.
- MAX_NUM_CAT, máximo número de categorías principales. Controla cuantos de los eventos más frecuentes en el sistema han de ser considerados como potenciales candidatos para la asignación de una clase.

El segundo procedimiento fue realizar un clasificador de aprendizaje automático empleando Naive Bayes para ofrecer un lista de posibles códigos como sugerencia a clasificar cuando no se obtenga un alto grado de confianza lo cual deriva en una verificación humana.
En [9] se realizan tres enfoques combinados para la asignación automática de códigos CIE-9 a reportes médicos de radiología. Se utiliza Lucene, una librería de herramientas de procesamiento de texto que usa la importancia relativa de las palabras de cada reporte para obtener la similaridad entre reportes (diagnósticos médicos) y se recuperan los reportes con mayor grado de similaridad, tomando el voto mayoritario. El segundo enfoque usa BoosTexter, un algoritmo de aprendizaje automático que implementa boosting, usando secuencia de palabras únicas (unigramas), secuencia de palabras consecutivas (n-gramas) y secuencia de palabras no consecutivas (s-gramas). El tercer enfoque es un codificador basado en reglas léxicas, semánticas, de negación y basadas en sinónimos.
En [7], se explica la importancia de las interdependencias entre clases aunado al estudio de la construcción del vector de características que representará cada documento. El algoritmo de aprendizaje fue un árbol de decisión C4.5 y se enriqueció usando un

clasificador de máxima entropía. Se desarrollaron reglas mediante el conocimiento y el estudio del corpus, la presencia de algunas palabras clave en el diagnóstico radicaban en suposiciones o relaciones entre algunas clases que generaban falsos positivos. Algunos ejemplos:

- Eliminar código 786.2 (tos) cuando el código 486 (neumonía) esté presente.
- Eliminar código 780.6 (fiebre) cuando el código 486 (neumonía) esté presente.
- Eliminar código 786.2 (tos) cuando el código 493.90 (asma) esté presente.
- Eliminar código 780.6 fiebre) cuando el código 599.0 (infección de vías urinarias) esté presente.

Este enfoque se basa principalmente en el desarrollo de reglas artesanales que describen y discriminan las clases de un corpus establecido. Al emplear otra base de datos es necesario un nuevo estudio para conocer el comportamiento de los nuevos datos y generar nuevas reglas.
No hay una solución de talla única para todos los errores de cada enfoque. En lugar de ello, la mejor combinación de técnicas varía de una aplicación a otra, e incluso de un sitio a otro en base a las variaciones en el lenguaje, la estructura de documentos, etc. Lo importante es centrarse en la precisión en el contexto del flujo de trabajo específico de cada aplicación.

3.4.5. Desafío en Medicina Computacional, 2007

La tarea de codificación CIE para informes de radiología fue uno de los primeros desafíos de la comunidad informática, donde cerca de 50 participantes presentaron resultados. El objetivo principal era crear y entrenar algoritmos de inteligencia computacional para automatizar la asignación de códigos CIE- 9-MC para informes de radiología anónimos con un conjunto de entrenamiento de 978 documentos y una serie de 976 documentos para pruebas.
En el desafío del 2007 [15] los sistemas de alto desempeño lograron un F-measure: 0,8908, el mínimo fue: 0.1541 y el promedio fue de 0.7670, con una desviación estándar de 0.1340. 21 sistemas tuvieron un F- measure entre 0,81 y 0,90. Otros 14 sistemas

tuvieron F- measure de 0.70. Los sistemas más valorados utilizan diversos enfoques como: Aprendizaje automático; Métodos simbólicos; Enfoques híbridos; Estructuras UMLS en sistemas directos, etc. El sistema ganador utilizó un enfoque basado en reglas identificando negaciones, hiperónimos y sinónimos y realizando procesamiento simbólico.

3.5. Discusión

Las características de asignación de códigos planteadas en la sección 3.2 y resumidas en la tabla 3.1 son rasgos que se han intentado capturar en los diversos métodos de asignación que se explicaron anteriormente. Se muestran los aspectos primordiales que cada método cubre, características que ayudan a mejorar la clasificación o representación del vector de características.

Tabla 3.1: Consideraciones en los diferentes métodos.

limpieza	· Tokenización · Palabras vacías · Puntuaciones · Números · Minúscula · Raíz de palabras
manejo de vocabulario	· Sinónimos · Hiperónimos · Términos médicos · Diccionarios controlados
uso de palabras clave	· Palabras específicas · Negaciones · Especulaciones · Incertidumbre
similitud en textos	· Secuencias de palabras. · Términos en común · Coincidencia de términos en tesauros · Frecuencia de n-gramas
modelo vectorial	· Pesado binario · TF-IDF · Frecuencias · N-gramas de palabras

Tabla 3.2: Principales características y técnicas que se emplean en diversos métodos en el estado del arte

CARACTERÍSTICA	AUTOR					
	Pereira	**Pakhomov**	**Aramaki**	**Kevers**	**Boytcheva**	**Ferrao**
Año	**2006**	**2006**	**2007**	**2010**	**2011**	**2013**
Instancias	1000	75,000	15,551	19,692	7,500	33,670
clases	-	-	995	895	-	1869
Metadatos	X	X				X
Simbólico	X			X		
Aprendizaje automático		X	X		X	X
Basado en Ejemplos		X	X			
Basado en reglas			X			
N-gramas de palabras	X	X	X	X	X	X
Steamming		X			X	
Vocabulario controlado	X			X		
Interrelaciones de clases						
metatesauros	X			X		
Instancias artificiales						
Entrada médica	informe	nota	diagnóstico	informe de alta	registro paciente	nota
Versión	CIE-10	CIE-8 Adap.	CIE-10	CIE-9	CIE-10	CIE-9
Recall(%)	68	93.70 %	83.2	60.21	74.68	67.71
F-measure(%)		90.40 %	80.4	43.28	84.5	51.34
Precision(%)	43	86.60 %	67		97.3	44.42

En las tablas 3.2 y 3.3 se presentan las diferentes características y técnicas usadas en el estado del arte referente a la asignación de códigos CIE y que diferencian notablemente cada investigación. Por ejemplo el número de instancias empleadas en la experimentación, un punto muy importante para comparar el método propuesto En [**?**] se usaron únicamente 1,000 instancias mientras que en [13] se usaron cerca de 75,000. De la misma forma se puede apreciar que tipo de métodos emplearon, desde uno simbólico, con aprendizaje automático, basado en ejemplos o basado en reglas. También se puede apreciar los resultados obtenidos en cada investigación, dividido en el recuerdo, medida-f y precisión. A diferencia de la tabla 3.2, en la tabla 3.3 se agrupan las investigaciones obtenidas de los participantes del desafió computacional 2007, el más relevante hasta el momento en asignación de códigos CIE.

Tabla 3.3: Principales características y técnicas que se emplean en los métodos del Desafio en Medicina Computacional 2007

CARACTERÍSTICA	AUTOR					
Año	**Sotelsek 2007**	**Farkas 2007**	**Goldstein 2007**	**Crammer 2007**	**Suominen 2008**	**Yang 2010**
Instancias	1954	1954	1954	1954	1954	1954
clases	48	48	48	48	48	48
Metadatos						
Simbólico		X				
Aprendizaje automático	X	X	X	X	X	X
Basado en Ejemplos		X	X		X	
Basado en reglas		X	X		X	
N-gramas de palabras	X	X	X	X	X	X
Steamming						X
Vocabulario controlado	X		X		X	
Interrelaciones de clases		X				X
metatesauros	X				X	
Instancias artificiales					X	
Entrada médica	diagnóstico	diagnóstico	diagnóstico	diagnóstico	diagnóstico	diagnóstico
Versión	CIE-9	CIE-9	CIE-9	CIE-9	CIE-9	CIE-9
Recall (%)		90.04	89.54	85.9		
F-measure (%)	89.08	89.93	88.55	86.5	87.7	
Precision (%)		87.85	87.58	87.1		

El estado del arte nos muestra diferentes trabajos referentes a la asignación de códigos CIE a diagnósticos médicos. Dichas investigaciones motivan al desarrollo del presente trabajo a establecer un escenario mas realista para la asignación de códigos. Dentro de los conjuntos de datos solo en [13] se usaron datos reales procedentes de la Clínica Mayo, en Rochester, Minnesota. En los trabajos restantes se usaron conjuntos de datos de radiología con tan solo 1000 instancias para entrenamiento y 1000 para prueba.

Esta motivación surge también porque existen muy pocos trabajos que intenten profundizar en escenarios robustos con este tipo de problemática. De esta manera se busca el desarrollo de un sistema que procese diagnósticos reales, textos médicos escritos al momento de la consulta sin pasar por ningún tipo de procesamiento o filtro. También contar con miles instancias que pertenezcan a miles de códigos (clases) y que no solo formen parte de una clasificación médica.

Capítulo 4

Método Propuesto

En el capítulo 1 se introdujo la tarea de codificación CIE, donde podemos apreciar como la tarea manual solo puede ser realizada por expertos etiquetadores y es por lo tanto un proceso muy costoso. Un experto etiquetador debe elegir entre miles de códigos CIE para indexarlo a un diagnóstico médico. El volumen de diagnósticos es abrumador y la capacidad de clasificación siempre conlleva en un considerable atraso. Es por ello que diversos trabajos se centran en resolver este tipo de problemas.

En el capítulo 3 se muestra el trabajo relacionado, donde se explican los distintos panoramas que abarca la asignación de códigos CIE, las características de las que depende la asignación y el texto usado para la codificación. Los textos usados para esta clase de tarea pueden ser la nota médica, el historial clínico del paciente, prescripciones médicas, o el diagnóstico específico que le fue asignado al paciente al momento de una consulta médica. Cabe señalar que nuestro trabajo se centra a partir del diagnóstico especificado en la consulta médica. También se muestran los diferentes métodos aplicados a dicha tarea.

La motivación principal de esta investigación yace en las diferentes características que puede comprender un corpus de esta índole. Se ha logrado identificar que cada método elegido depende de los aspectos de cada conjunto de datos. Los corpus usados en cada estudio varían significativamente, como el texto médico usado para el procesamiento y asignación de códigos, que puede ser la nota médica o el diagnos-

tico, contener información extra del paciente o quien capturó la información, como el mismo médico u otro tipo de personal como una enfermera o secretaria, siendo la elección del método aplicado de acuerdo al contexto de cada conjunto de datos. El método propuesto se realiza bajo el contexto de nuestro conjunto de datos y su estudio.

Además de la automatización, esta investigación se centra en la obtención de un sistema de asignación de códigos CIE, un conjunto de datos formado por diagnósticos médicos derivados de consultas médicas obtenidas de un hospital general en el idioma español, empleando técnicas de recuperación de información. Cabe destacar que el corpus empleado contiene diagnósticos reales escritos directamente por médicos y por ello el reto de poder clasificarlos contando con errores de ortografía y de escritura derivados de la generación del diagnóstico.

Nuestra investigación se centra principalmente en obtener un sistema de clasificación que apoye a los expertos etiquetadores en la tarea de asignación códigos CIE, el problema en particular comprende el diagnóstico médico (sin la nota médica) como el texto a procesar. El método propuesto consiste en realizar un sistema que nos brinde los cinco códigos más similares a la consulta realizada para dar al experto etiquetador la elección final del código CIE.

En el estado del arte se utiliza principalmente el corpus empleado en el Desafió de Medicina Computacional del año 2007, con alrededor de 1000 diagnósticos médicos para entrenamiento y 1000 para pruebas, con aproximadamente 45 códigos diferentes pertenecientes al área de radiología. Un aspecto importante a resaltar es el hecho de poder abordar este tipo de problemática de nuestro trabajo al mundo real, donde se cuenta no con cientos de códigos sino con miles, procesando textos de varios departamentos escritos directamente en cada consulta médica que pueden contar con errores ortográficos y de escritura.

La siguiente sección presenta una descripción de las características de nuestro sistema de asignación de códigos CIE y los aspectos que cada etapa resuelve para

mejorar la clasificación de códigos CIE. En las secciones subsecuentes se presentan los experimentos realizados con los resultados obtenidos.

4.1. Pre-procesamiento

En primer lugar se procede a pre-procesar el conjunto de datos, donde cada documento es normalizado para optimizar la clasificación.

- Eliminar signos especiales como puntuaciones, comas, llaves, etc
- Quitar espacios extras entre palabras.
- Pasar todo el texto a minúscula.
- Eliminación de diagnósticos sin codificación.

Posteriormente se divide el conjunto de datos en subconjuntos por el metadato *servicio*. De esta manera se obtienen subconjuntos de diagnósticos que serán empleados como corpus separados y por lo tanto tendrán una evaluación individual. En la tabla 4.1 se pueden observar algunos servicios pertenecientes al corpus de la investigación. Se cuenta en total con 74 servicios diferentes.

4.2. Extracción de características e indexado

Los documentos son representados como un conjunto de características binarias, en varios trabajos relacionados el modelo vectorial es empleado para la representación de los textos médicos [13].

Modelo Vectorial

Se conoce como modelo de espacio vectorial a un modelo algebraico utilizado para filtrado, recuperación, indexado y cálculo de relevancia de información. Representa

Tabla 4.1: Ejemplo de los diferentes servicios

servicio
Epidemiología
Hematología
Nutrición
Cirugía plástica
Neumología
Neurocirugía
Inhaloterapia
Clínica del dolor
Endocrinología
Odontopediatría
Odontología
Cardiología
Nefrología
Alergología

documentos en lenguaje natural de una manera formal mediante el uso de vectores (de identificadores, por ejemplo términos de búsqueda) en un espacio lineal multidimensional.

La teoría básica es que la relevancia de un documento frente a una búsqueda puede calcularse usando la diferencia de ángulos (basada en el coseno de esos ángulos) de cada uno de los documentos respecto del vector que busca, utilizando el producto escalar entre el vector de búsqueda. Así un valor de coseno de cero significa que la búsqueda y el documento son ortogonales el uno al otro, y eso significa que no hay coincidencia.

Para determinar el coseno del ángulo entre dos vectores se usa la siguiente ecuación:

$$cos\theta = \frac{\vec{V_1}.V_2}{\parallel \vec{V_1} \parallel \parallel \vec{v_2} \parallel} \tag{4.1}$$

donde:

- θ es el ángulo entre $\vec{V_1}$ y $\vec{V_2}$.
- $\vec{V_1}$ es el primer vector.

- $\vec{V_2}$ es el segundo vector.
- . representa el producto punto.
- $\| x \|$ representa la magnitud del vector x.

Representación robusta para diagnósticos

El espacio vectorial de términos (el vocabulario para el conjunto de datos) de un conjunto de documentos y su respectiva consulta son representados por vectores. Cada vector se encuentra formado por características, en este caso el vocabulario. Asumiendo un vocabulario de tamaño N (también llamado vocabulario indexado).

$$Q_j = \{w_1t_1, w_2t_2, w_3t_3, ..., w_Nt_N\} \tag{4.2}$$

$$D_k = \{v_1t_1, v_2t_2, v_3t_3, ..., v_Nt_N\} \tag{4.3}$$

w_i's y v_i's son pesos binarios[0,1] que representan la importancia relativa de los términos. Los vectores se componen de una entrada por cada término $(t_i's)$del vocabulario. Así, la mayoría de los pesos son 0. Los t's son términos.

	T_1	T_2	....	T_t
D_1	w_{11}	w_{21}	...	w_{t1}
D_2	w_{12}	w_{22}	...	w_{t2}
:	:	:		:
:	:	:		:
D_n	w_{1n}	w_{2n}	...	w_{tn}

Figura 4.1: Representación de una matriz término-documento.

El texto es inicialmente representado por un vector de características binarias por unigramas de 3 caracteres. Por ejemplo, asma que es representado por : asm, sma. Recordemos que en la representación binaria el peso Booleano indica la presencia (peso 1) o ausencia (peso 0) de un término en un documento.

Lo que se busca obtener es una representación sensible a faltas de ortografía y de escritura debido a que los diagnósticos son escritos directamente por el médico al momento de la consulta. Se necesita una técnica que permita resolver dicho problema y esto se logra representando cada diagnóstico como un vector de unigramas de caracteres.

Por ejemplo el diagnóstico "has descontrolada estadio ii" se representaría de la siguiente forma:
=¿"has snd des esc sco con ont ntr tro rol ola lad ada dae aes est sta tad dio ioii oii ".

Segmentación de diagnósticos médicos

Un aspecto importante dentro de la clasificación de diagnósticos en un escenario real es la mezcla de distintos diagnósticos dentro del mismo texto médico, mientras que un texto es etiquetado con un código CIE, por ejemplo, diabetes mellitus II, ésta también puede tener otros diagnósticos secundarios que comúnmente ocurren simultáneamente con el primero o pueden ser totalmente diferentes, como por ejemplo "diabetes mellitus II embarazo de alto riesgo", siendo dos enfermedades con su respectivo código CIE, *E11 Diabetes mellitus II*, *Z35 Supervisión de embarazo de alto riesgo* pero etiquetados solo con uno.
Para nuestro caso en particular, el hospital que nos brindó estos diagnósticos solo puede etiquetar cada texto médico con un solo código CIE, por lo cual no existe la multi-etiqueta y como regla médica cada texto debe ser etiquetado con el código de la enfermedad principal.
Dentro del estudio realizado se comprueba que el paciente acude por una enfermedad en particular pero el médico tiene la responsabilidad de describir todas las enfermedades que el paciente sufre por lo cual es difícil poder clasificar en estos casos. Es por ello que se realiza la etapa de segmentación de diagnósticos para poder equilibrar el peso de ambos diagnósticos médicos realizando una fusión de listas (a continuación se describe esta etapa). Recordemos que la clasificación se realiza por cada subconjunto de datos (servicio). De esta manera concentramos los diagnósticos más similares. La

segmentación de cada subconjunto de datos se lleva a cabo partiendo los vectores características en 3 segmentos:

- Utilizando los vectores de características completos(máximos) por cada subconjunto de datos. Esto se refiere a utilizar todo el texto del diagnóstico. Partiendolo en n-gramas de 3 caracteres.

- Utilizando el promedio de n-gramas de caracteres por cada subconjunto de datos. Para cada servicio se obtiene el promedio de n gramas caracteres de longitud del texto médico y después cada diagnostico es segmentado hasta donde llegue el promedio obtenido y el resto del diagnóstico es eliminado.

- Utilizando el mínimo (la mitad del promedio) para cada subconjunto de datos. Para este caso, el promedio se divide el promedio en dos y el numero de n-gramas obtenido es hasta donde se partirán los diagnosticos, como en el caso anterior, se contaran los n-gramas hasta llegar a la mitad del promedio y lo demás se eliminará.

De esta manera obtenemos 3 corpus que representan a cada servicio de nuestro conjunto de datos. Uno representa el total del texto, otro el promedio y el último la mitad del promedio.

Por ejemplo: Tenemos el conjunto de datos de cardiología donde el promedio de Siguiente diagnóstico: “has descontrolada estadio ii”, supongamos que el promedio de n-gramas es de 10, entonces las tres instancias resultantes serian:

- “has snd des esc sco con ont ntr tro rol ola lad ada dae aes est sta tad dio ioii oii ”

- “has snd des esc sco con ont ntr tro rol ola lad ada dae”

- “has snd des esc sco con ont”

  De esta manera formaríamos 3 subconjuntos de datos con muestras de diferentes longitudes.

4.3. Clasificación

Para esta etapa se parte de tener un conjunto de vectores de características que describen a cada diagnóstico, cada servicio comprende 3 subconjuntos de datos segmentados que serán clasificados por el algoritmo de aprendizaje. El algoritmo empleado para nuestro sistema es un k-NN o también conocido como vecinos más cercanos.

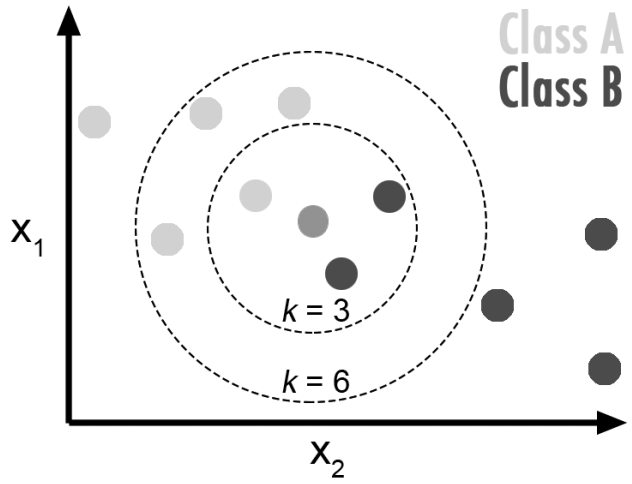

Figura 4.2: Concepto del algoritmo k-NN

Como se puede apreciar en la figura 4.2 se toman los k vecinos más cercanos usando una medida de similaridad, en este caso se emplea la medida coseno.

Por cada subconjunto de datos se emplea un k-NN con el corpus de entrenamiento de cada servicio respectivamente. El conjunto de entrenamiento es usado en el algoritmo k-NN y cuando ingresamos una nueva instancia, éste recupera los vecinos más cercanos. En lugar de recuperar los k vecinos más cercanos, buscamos obtener una lista de las 10 clases más similares obteniendo así 3 listas como resultado del algoritmo de clasificación.

Al emplear el algoritmo k-NN a cada subconjunto de datos obtenemos 3 listas con 10 clases respectivamente como lo muestra la figura 4.3 para cada instancia de prueba.

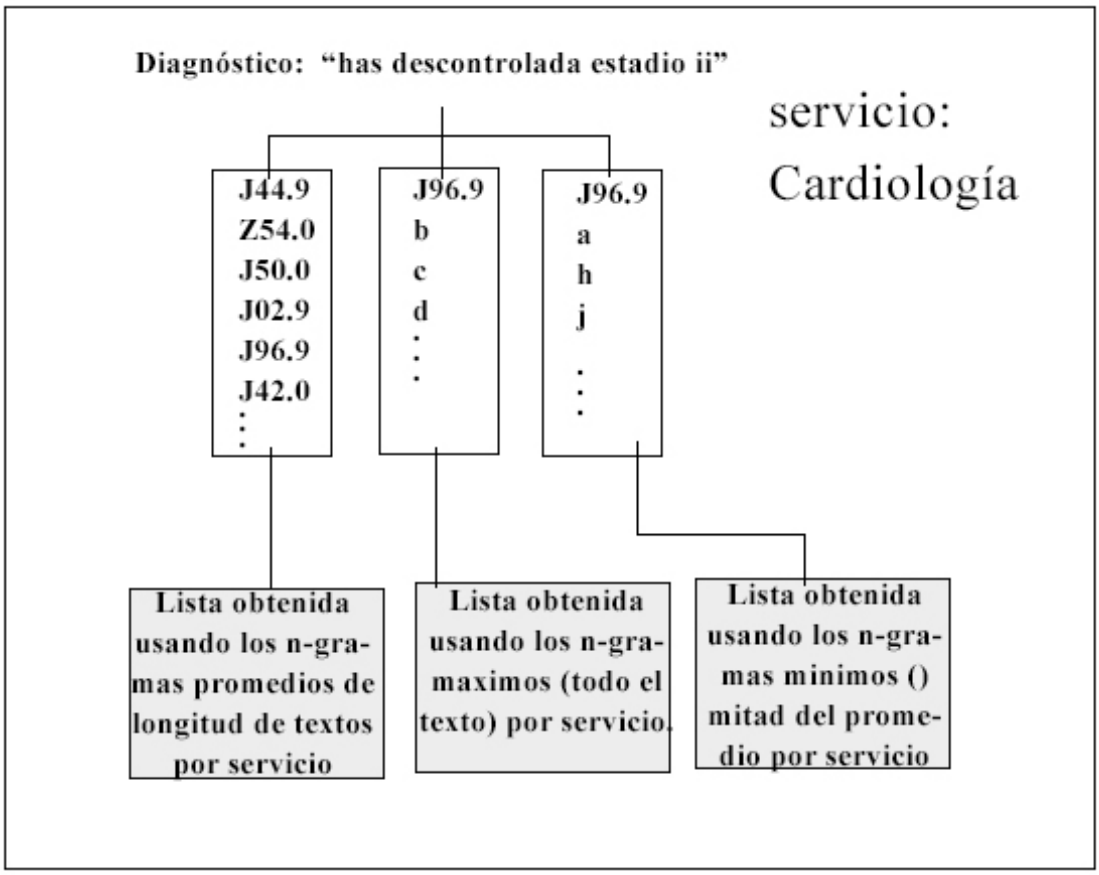

Figura 4.3: listas obtenidas para un ejemplo del servicio de Cardiología

Si observamos en la figura 4.3 en el servicio Cardiología ilustra 3 listas de clases, estas listas representan la salida de cada instancia de prueba. Recordemos que cada servicio cuenta con tres subconjuntos de datos. El primero con el subconjunto de instancias de textos completos, el siguiente subconjunto tiene textos promedio y el último cuenta con un subconjunto segmentado hasta la mitad del promedio de longitud de textos médico por servicio. Cada subconjunto es ingresado al algoritmo K-NN obteniendo una lista de las k clases mas similares por cada instancia de prueba. De esta manera podemos observar que en el servicio 1 se cuenta con tres listas de códigos. Estas listas representan la salida del algoritmo k-NN por cada instancia de prueba, obteniendo los 10 códigos mas similares.

Así por ejemplo para el diagnóstico de prueba: "has descontrolada estadio ii" del servicio Cardiología, se obtuvieron 3 listas de 10 códigos cada una.

4.4. Fusión

Con la obtención de las 3 listas de códigos por instancia para cada subconjunto de datos se procede a la fusión de listas, en este caso se emplea el método CombMNZ, el cual otorga pesos más altos a aquellos elementos presentes en varias listas.
De esta manera se procede a fusionar las 3 listas con 10 elementos cada una, al volver a realizar el ordenamiento de los elementos de acuerdo a su peso final en la lista fusionada nos quedamos con los primeros 10 elementos de la lista fusionada como se muestra en la figura 4.4

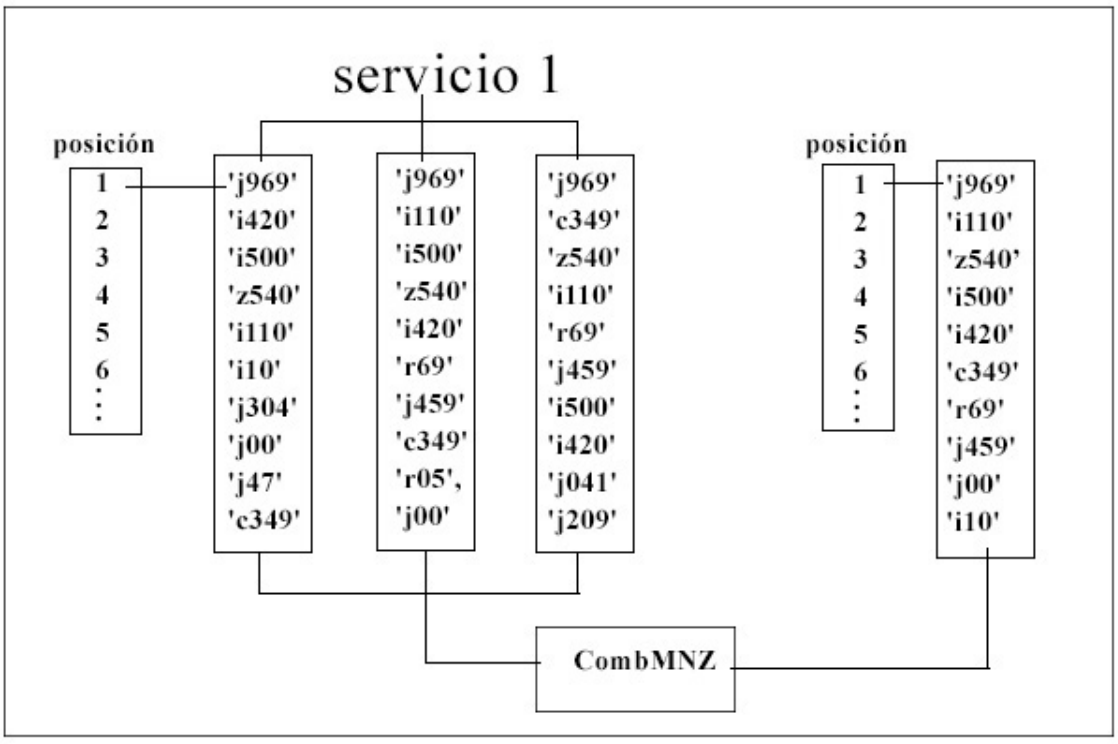

Figura 4.4: Método CombMNZ en listas

Con la fusión se pretende equilibrar las posiciones de los códigos CIE que quizás fueron castigados en la clasificación por tener diagnósticos secundarios en el texto médico.

A continuación en el algoritmo 1 se presenta un pseudocódigo del funcionamiento del método propuesto.

Algorithm 1 Método de asignación de códigos CIE a diagnósticos médicos

Require: Diagnósticos médicos de un servicio $K = \langle d_1, d_2, d_3 ... \rangle$

Ensure: lista de 10 códigos CIE por instancia, donde $d_i = instancia$.

1: donde d_i se separa en n-gramas de 3 caracteres.
2: **for** cada documento en el servicio **do**
3: Se obtiene el número de n-gramas por texto médico d.
4: Se calcula el promedio de n-gramas del servicio que será la variable PROMEDIO.
5: Se obtiene la mitad del promedio de n-gramas que sera la variable MÍNIMO.
6: **end for**
7: **for** cada documento en el servicio **do**
8: Se segmenta d_i hasta el PROMEDIO DE N-gramas obteniendo d_{i2}.
9: Se segmenta d_i hasta el MÍNIMO de n-gramas obteniendo así d_{i3}.
10: Se obtiene $K = \langle d_1, d_2, d_3 ... \rangle$ con textos completos, MÁXIMO.
11: Se obtiene $K_2 = \langle d_{12}, d_{22}, d_{32} ... \rangle$ con textos hasta el promedio, PROMEDIO.
12: Se obtiene $K_3 = \langle d_{13}, d_{23}, d_{33} ... \rangle$ con textos hasta el mínimo, MÍNIMO.
13: **end for**
14: **for** cada subconjunto de instancias, donde tenemos K, K_2, K_3 **do**
15: **for** Cada documento del subconjunto **do**
16: Se ingresa la instancia para ser evaluada con el k-NN.
17: Se obtienen 10 clases mas similares por instancia $C = \langle c_1, c_2, c_3 ... \rangle$.
18: **end for**
19: **end for**
20: **for** Cada documento del subconjunto **do**
21: Se fusionan las listas de clases $C = \langle C_{11}, C_{12}, C_{13} ... \rangle$ mediante CombMNZ.
22: Se obtiene una lista de ranking de clases para cada instancia.
23: Se limita la lista a las 10 clases con mayor ranking de la lista final fusionada.
24: **end for**

Capítulo 5

Experimentación y resultados

El presente capítulo presenta los experimentos y resultados que se obtuvieron durante la investigación. En los experimentos se evalúa a los clasificadores con: la exactitud, el recuerdo, la precisión, y la medida F (F-Measure), la precisión P@5, P@10 y el MRR. La primera sección presenta las características usadas para la selección de los subconjuntos de datos que serán empleados en la experimentación del sistema de asignación. En las secciones posteriores se muestran los subconjuntos seleccionados y los experimentos realizados así como los resultados obtenidos.

5.1. Conjunto de datos

En los experimentos se utiliza un conjunto de datos obtenido de un hospital general de la ciudad de Puebla, el ISSSTEP. de. A continuación se presenta una descripción de la colección en la tabla 5.1

Tabla 5.1: Características del conjunto de datos

			Clases (códigos)		
Año	Vocabulario	Totales	$\leq$ 20 ocurrencias	1 ocurrencia	$\leq$ 5 ocurrencias
2011	61,340 términos	5,187	3,186	1,430	2,778
2012	63,460 términos	4,953	3,507	1,333	2,811

En la tabla 5.1 podemos notar que de las clases totales para el año 2011 cerca

de 3,186 códigos (61.42 %) tiene menos de 20 ocurrencias en el conjunto de datos. Mientras que existen 2,778 (53.55 %) con menos de 5 ocurrencias y 1,430 (27.56 %) con tan solo una ocurrencia. De esta manera podemos notar que en el 2011 tan solo cerca de 2,001 códigos (38.57 %) tiene más de 20 ocurrencias mientras que en el 2012 tan solo cerca de 1450 códigos (30 %) tienen mas de 20 ocurrencias.

El corpus cuenta con 6 características: clase (código), diagnóstico, servicio, edad, sexo y clave del médico. En la tabla 5.2 se presenta un ejemplo de representación de los datos.

Tabla 5.2: Características de los diagnósticos médicos

DIAGNÓSTICO	CIE	SEXO	EDAD	CLAVEMED	SERVICIO
CARIES DE LA DENTINA	K02.1	M	9	A557	ODONTOPEDIATRÍA
RINOFARINGITIS AGUDA [RESFRIADO COMÚN]	J00	M	9	M302	PEDIATRÍA
ENTEROCOLITIS AGUDA	A09	M	8	V325	URGENCIAS ADULTOS

El conjunto de datos se divide en dos subconjuntos, año 2011 y 2012. El año 2011 con 1,048,576 instancias y el 2012 con 1,043,723 instancias.
Como se detalla en la tabla 5.1 cada subconjunto de datos cuenta con un vocabulario extenso, miles de códigos y clases desbalanceadas. Se puede observar como ejemplo en la tabla que de un total de 5187 clases del año 2011 tenemos un 27.56 % (1430 clases) de clases que solo contienen 1 ocurrencia en todo el subconjunto de datos mientras que en el año 2012 se cuenta con 38 % (1333 clases) en la misma circunstancia. Por la magnitud del problema se decidió utilizar el subconjunto de datos del 2011 como entrenamiento y el conjunto de datos del 2012 para prueba, de esa forma se logra visualizar cómo las diferentes clases (códigos) cambian o se comportan con el paso de los años.

5.2. Evaluación de subconjuntos de datos

A efectos de nuestra investigación, hemos tomado en cuenta tres diferentes características para nuestro corpus: dimensionalidad del vocabulario, desequilibrio de clases y brevedad en texto. Consideramos que estas características nos ayudan a evaluar los aspectos de una colección de documentos. El objetivo es determinar los aspectos de los diferentes subconjuntos de datos y elegir subconjuntos que cubran

características diferentes para observar el comportamiento de nuestro sistema de asignación de códigos ante diversos panoramas. Es por ello que después de evaluar cada subconjunto de datos se elegirán 12 subconjuntos con diferentes características (las mostradas a continuación)para poder estudiar como se comporta nuestro sistema de asignación. A continuación se describen las características para la evaluación de los textos médicos.

5.2.1. Dimensionalidad del vocabulario

Esta medida supone que los subconjuntos que pertenecen a un dominio estrecho compartirán el número máximo de términos de vocabulario en comparación con los subconjuntos que no lo hacen. En caso de un gran conjunto de datos de dominio, se espera que la desviación estándar de vocabularios obtenidos a partir de subconjuntos de este conjunto de datos mayor que la de un corpus de dominio estrecho.
Formalizamos la idea antes mencionada de la siguiente manera. Dado un corpus C (con vocabulario $V(C)$), que se compone de k categorías C_i, la medida de amplitud de dominio de C, (SVB) se puede escribir como se muestra en la ecuación (5.1).

$$SVB(C) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(\frac{\mid V(C_i) \mid - \mid V(C) \mid}{\mid C \mid}\right)^2} \tag{5.1}$$

5.2.2. Desbalanceo de clases

El grado de desbalanceo de clases es una característica importante que debe ser considerada cuando se clasifica cada corpus, ya que de acuerdo con el grado de desequilibrio podrían existir diferentes niveles de dificultad. Dado un corpus C (de n documentos) con un patrón oro predefinido compuesto por k clases (C_i) , el número esperado de documentos por clase se supone que es : $ENDC(C) = \frac{n}{k}$. La medida de evaluación de desbalanceo de clases (DC) se calcula como la desviación estándar de C con respecto a la cantidad esperada de documentos por clase en el estándar de oro, como se muestra en la ecuación (5.2).

$$DC(C) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(\mid C_i \mid -ENDC(C)\right)^2} \tag{5.2}$$

Se evalúa el corpus de diagnósticos médicos divididos por el servicio, en lo que se obtienen 74 servicios diferentes como se puede observar en la figura 5.1. También podemos observar en la figura el grado de desbalanceo de las clases, mientras que en pediatria tenemos cerca de 90,000 diagnósticos en alergología tenemos 10,000 e inhaloterapia con tan solo 200.

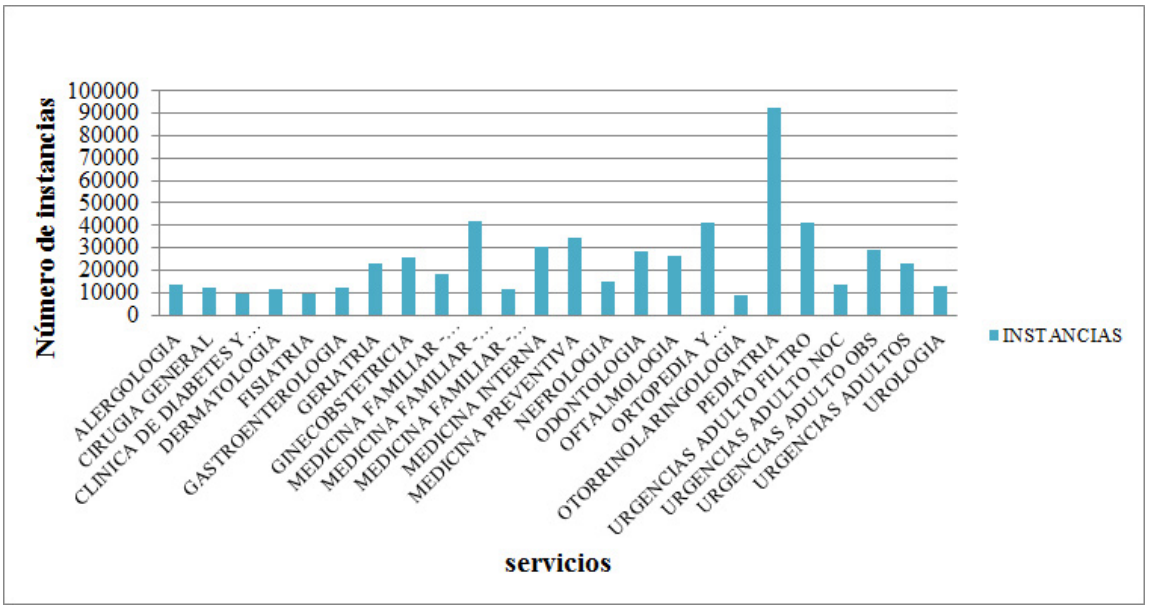

Figura 5.1: Diferentes servicios con los que cuenta el conjunto de datos

5.2.3. Brevedad (Shortness)

Cuando se trata de textos muy cortos, la frecuencia de su vocabulario es muy baja y, por lo tanto, los algoritmos de agrupación tienen problemas donde la matriz de similitud tiene valores muy bajos. Por lo tanto la longitud del texto promedio del corpus a agruparse es una característica importante que debe considerarse [16].
Esta medida evalúa características derivadas de la longitud de un texto. Dado un corpus C formado por n documentos D_i, se calcula la media aritmética de las Longitudes de los Documentos (LD) como se muestra en la ecuación (5.3).

$$LD(C) = \frac{1}{n} \sum_{i=1}^{n} \mid D_i \mid \tag{5.3}$$

Se procede a evaluar cada subconjunto de datos para obtener sus respectivas características como se observa en la figura 5.2 y se seleccionan 12 *servicios* que servirán como subconjunto de datos para evaluar al sistema de asignación de códigos CIE.

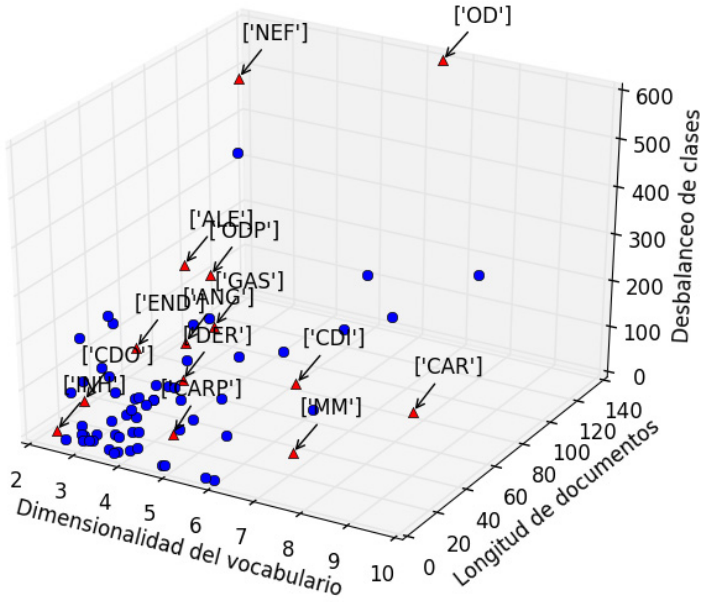

Figura 5.2: Evaluando los subconjuntos de datos

Los servicios seleccionados se muestran en la tabla 5.3 donde observan los aspectos principales como lo es el vocabulario presente en cada subconjunto, las clases totales y las instancias que los conforman. Por ejemplo, tenemos a clínica del dolor y endocrinología que cuentan con casi el mismo vocabulario a diferencia que clínica del dolor cuenta con 247 clases y endocrinología con 197 y con mas de 2,000 instancias por lo cual tenemos mas instancias de ejemplos para definir cada clase. En módulo materno tenemos un vocabulario de 1,174 con 163 clases diferentes y tan solo 2,648 instancias, lo que nos indica que los textos médicos varían notablemente y será difícil

Tabla 5.3: Subconjuntos seleccionados para la experimentación

servicio	vocabulario	clases	instancias
Inhaloterapia	95	40	167
Clínica del dolor	638	247	3,973
Endocrinología	639	197	6,155
Odontopediatría	318	83	5,212
Odontología	1,407	213	28,501
Cardiología	2,630	281	8,006
Nefrología	1,595	317	15,224
Alergología	1,133	293	13,228
Gastroenterología	1,800	364	12,001
Dermatología	1,850	421	11,674
Módulo materno	1,174	163	2,648
Angiología	808	176	4,361
Cardiología pediátrica	674	145	2,020
Clínica de displasia	698	103	3,218

definir cada clase.

En la figura (5.2) tenemos a subconjuntos como odontología[OD] que tiene un gran desbalanceo de clases, nefrología[NEF] cuenta también con desbalanceo de clases y poco vocabulario entre sus instancias. Podemos observar también a inhaloterapia[INH] con poco vocabulario y pocas instancias por clase lo cual nos dificulta la discriminación entre éstas. Cabe aclarar que al tomar los diagnósticos médicos del 2011 como entrenamiento y los diagnósticos del 2012 como prueba existen clases (códigos) inexistentes en el conjunto de entrenamiento. Dicho problema se maneja dentro del sistema como errores en la clasificación, pues no se puede predecir ni eliminar estas clases, las cuales seguirán apareciendo con el paso de los próximos años y siendo este un problema del mundo real se decide dejarlas dentro del conjunto de pruebas y manejarlas como errores.

En la tabla 5.4 se muestra el porcentaje de clases del conjunto de pruebas inexistentes en el conjunto de entrenamiento. Como anteriormente se menciona, se tomó el conjunto de datos del 2011 para entrenamiento y el conjunto del 2012 para pruebas, debido a esto existen clases(códigos CIE) que se encuentran en el 2012 pero no en el 2011 las cuales asignamos como clases inexistentes. Así en la tabla 5.4 se muestran

las clases inexistentes en el entrenamiento por servicio, las instancias totales que cuentan con una clase inexistente y el porcentaje total del servicio que cuenta con instancias con códigos que no se encuentran en el entrenamiento. Este aspecto ya se encontraba contemplado en la presente investigación debido a que queremos abordar un tema del mundo real. Sabemos de antemano que cuando se pretenda clasificar nuevas instancias en un entorno real como un hospital general, siempre tendremos códigos nuevos que nos reflejarán las nuevas enfermedades o padecimientos que surgen año con año. Es por ello que se manejan estos eventos como errores y no se sacan del conjunto de prueba.

Tabla 5.4: Porcentaje de clases e instancias inexistentes en el conjunto de entrenamiento

servicio	clases	instancias s/clases	Porcentaje total(%)
Inhaloterapia	15	51	26.70
Clínica del dolor	75	135	4.55
Endocrinología	54	77	1.92
Odontopediatría	23	41	1.07
Odontología	58	80	0.30
Cardiología	88	142	2.23
Nefrología	99	162	1.12
Alergología	80	91	0.78
Gastroenterología	71	101	0.96
Dermatología	60	173	2.50
Modulo materno	57	92	3.63
Angiología	142	536	12.86
Cardiología pediátrica	51	116	7.62
Clínica de displasia	42	72	2.64

5.3. Experimentación con los subconjuntos seleccionados

Se emplea el algoritmo Naive Bayes con unigramas de palabras como primera experimentación. Se construye el vector característica por cada diagnóstico usando valores booleanos. A continuación, en la tabla 5.5 se presentan los resultados de la experimentación.

Tabla 5.5: Resultados empleando el algoritmo Naive Bayes con unigramas de palabras

servicio	precision (%) NB	Precisión Máxima (%)	p@5 (%) NB	MRR en 5 (%
Inhaloterapia	60.21	73.3	63.35	
clínica del dolor	76.25	95.45	77.16	
Endocrinología	87.03	98.08	88.73	
Odontopediatría	87.58	98.93	93.13	
Odontología	85.01	99.7	93.73	
Cardiología	68.08	97.77	75.68	
Nefrología	87.81	98.88	90.11	
Alergología	91.33	99.22	93.54	
Gastroenterología	75.04	99.04	86.1	
Dermatología	79.45	97.5	82.3	
Módulo materno	26.9	96.37	34.44	
Angiología	54.67	87.14	56.61	
Cardiología pediátrica	64.78	92.38	72.27	
Clínica de displasia	51.78	97.36	65.67	

Algoritmo K-NN con unigramas de palabras y n-gramas de caracteres.

Se realiza una experimentación del algoritmo k-nn con unigramas y n-gramas de caracteres para ver si este último mejora la clasificación. En las gráficas 5.3, 5.4, 5.5 se presentan los experimentos realizados. En la figura 5.3 se refleja el Accuracy de cada clasificador, como se puede observar se ven reflejados los subconjuntos seleccionados para la experimentación. En la tabla 5.4 se presenta la precisión a 5, lo cual significa que se obtienen los 5 resultados más similares para saber si se encuentra la clase oro del diagnóstico en los primeros 5 resultados obtenidos. Recordemos que uno de los objetivos del proyecto es apoyar en la asignación de códigos, de esta manera se pueden presentar al experto etiquetador una lista de los 5 resultados más similares. En la figura 5.5 se observa el MRR en 5, de igual manera que la P@5 se obtienen únicamente las 5 clases más similares y se valora la posición de la clase gold en la lista, entre más cercana se encuentre a la primera posición, el MRR tendrá mayor peso.

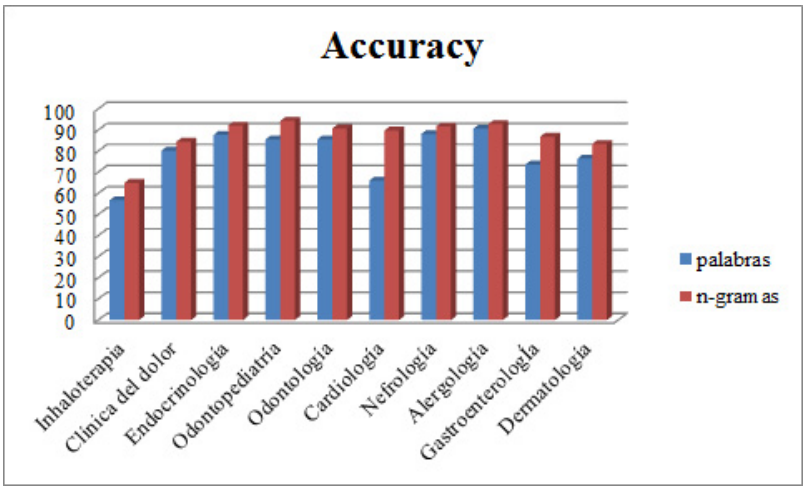

Figura 5.3: Resultados del accuracy en la experimentación

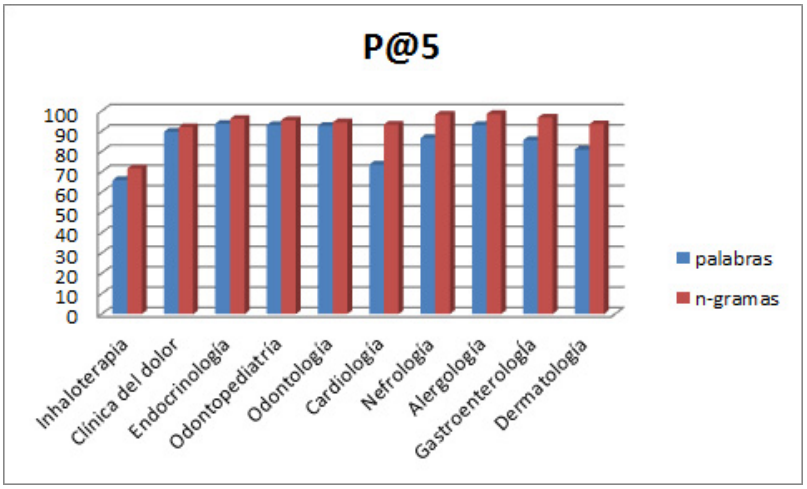

Figura 5.4: Resultados de la precisión a 5 en la experimentación

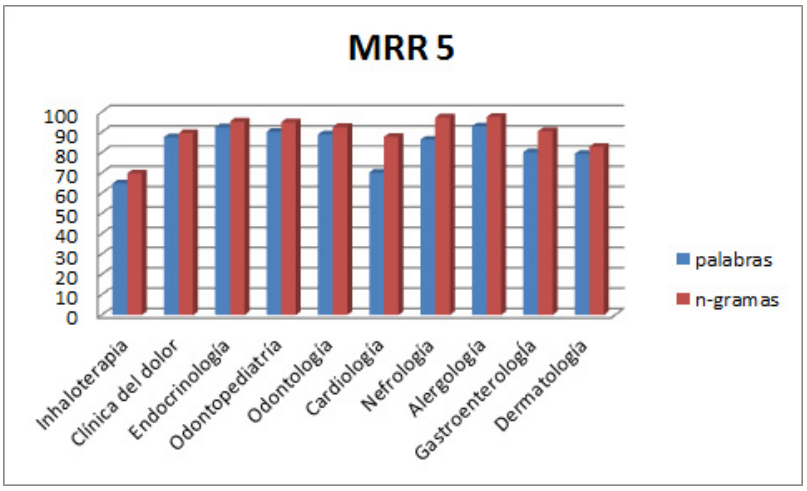

Figura 5.5: Resultados del MRR en 5 en la experimentación

El subconjunto Cardiología presenta un mejor desempeño de n-gramas de caracteres a unigramas de palabras. Este subconjunto cuenta con los diagnósticos (documentos) más largos de todo el corpus y cuenta con una dimensionalidad de vocabulario alta, lo que nos dice que cuenta con un gran vocabulario. De igual manera es el subconjunto que utiliza en menores proporciones la jerga médica y solo presenta terminología médica para cada enfermedad añadiendo siglas para diversos padecimientos, de esta manera si se escribe erróneamente las siglas o la abreviatura el significado del diagnóstico cambia completamente.

El subconjunto de Alergología obtuvo el peor desempeño, cuenta con una dimensionalidad de vocabulario muy baja, esto nos indica un subconjunto que comparte fuertemente el vocabulario entre sus instancias, estudiando más a fondo el conjunto documentos se observa que las palabras, "asma" y "rinitis" son altamente usadas, se presenta la palabra "asma" en un 6 %, y "rinitis" en un 63.65 % del total de documentos, siendo palabras cortas y muy definidas son altamente empleadas para describir una enfermedad respiratoria y la equivocación en la escritura es muy baja.

Se puede concluir que el desempeño del algoritmo K-nn con una representación basada en n-gramas mejora considerablemente la clasificación a utilizar unigramas de palabras. Debido a que los diagnósticos cuentan con diversos errores de escritura y variabilidad léxica en la definición de un problema médico el uso de n-gramas nos brinda una gran ventaja para la clasificación.

5.4. Experimentación K-NN con segmentación

Para emplear el algoritmo k-nn para la experimentación de la segmentación primero se seleccionó el rango de clases que se desean listar en cada segmentación, en nuestro caso usamos $k = 30$ para la obtención de los 30 vecinos más cercanos. Después de obtener la lista de vecinos más cercanos se obtienen la clase de cada vecino y se limpia la lista hasta obtener únicamente las 10 clases más cercanas de cada subconjunto. En caso de no obtener la lista de las 10 clases, el algoritmo busca más vecinos hasta completar su lista de clases (códigos).

A cada subconjunto de datos (servicio) se le calcula la variable PROM mediante el promedio de trigramas de caracteres y a su vez se obtienen los trigramas mínimos (MIN, la mitad del promedio) y los máximos (MAX) que serían los n-gramas totales del diagnóstico.
A continuación se muestra la variable PROM obtenida de cada subconjunto en la tabla 5.6.

Al segmentar los diagnósticos de cada servicio de acuerdo a las variables MIN, PROM y MAX(todos los n-gramas) obtenemos 3 diferentes subconjuntos por servicio que serán evaluados por el algoritmo k-NN usando la medida de similaridad coseno, obteniendo de esta manera 3 listas diferentes con las 10 clases más similares.

Recordemos que cada segmentación se realiza de acuerdo a los siguientes aspectos:

- El promedio de n-gramas de caracteres. Se calcula dicho promedio por cada servicio.

- El mínimo de n-gramas de caracteres. La mitad del promedio de n-gramas por servicio.

- El máximo de n-gramas de caracteres. El número total de n-gramas de cada diagnóstico.

Tabla 5.6: Variable PROM para segmentar los vectores de cada servicio.

servicio	promedio
inhaloterapia	7
clínica del dolor	11
endocrinología	11
odontopediatría	18
odontología	17
cardiología	19
nefrología	8
alergología	12
gastroenterología	16
dermatología	11
modulo materno	24
angiología	12
cardiología pediátrica	17
clínica de displasia	12

Los resultados de algunos subconjuntos de datos se presentan a continuación en las siguientes figuras.

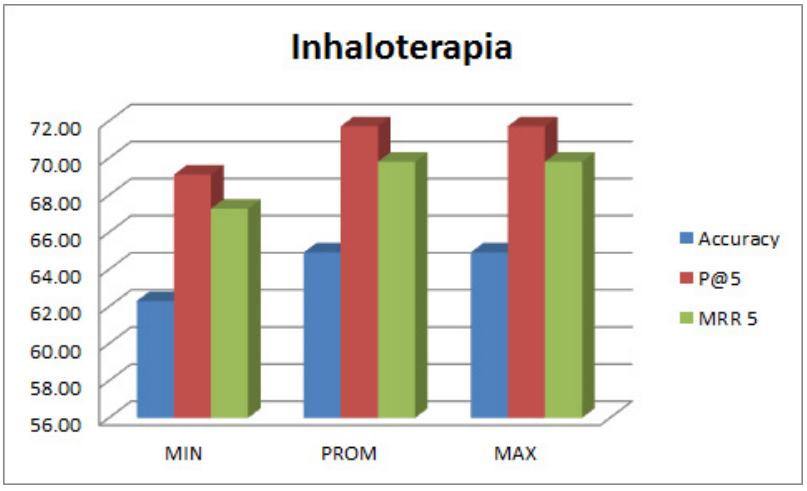

Figura 5.6: Resultados de segmentación de Inhaloterapia

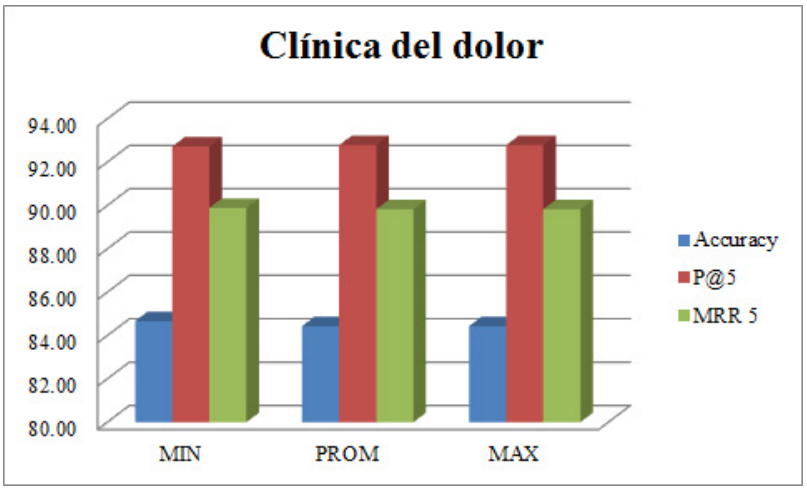

Figura 5.7: Resultados de segmentación de Clínica del dolor

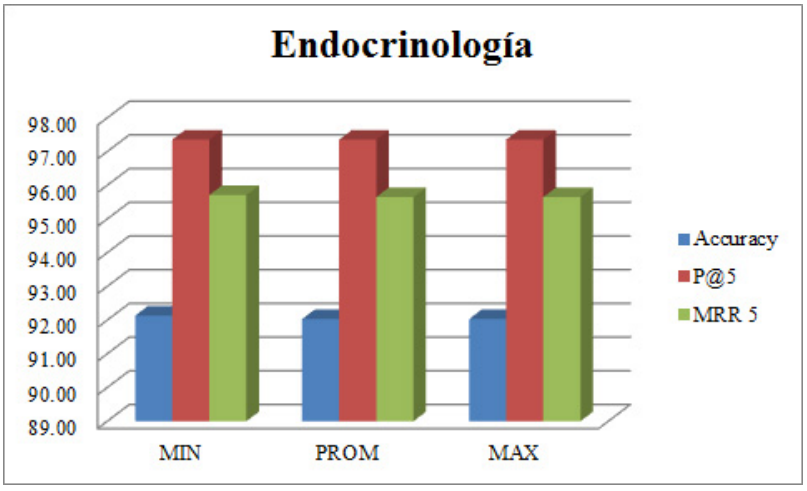

Figura 5.8: Resultados de segmentación de Endocrinología

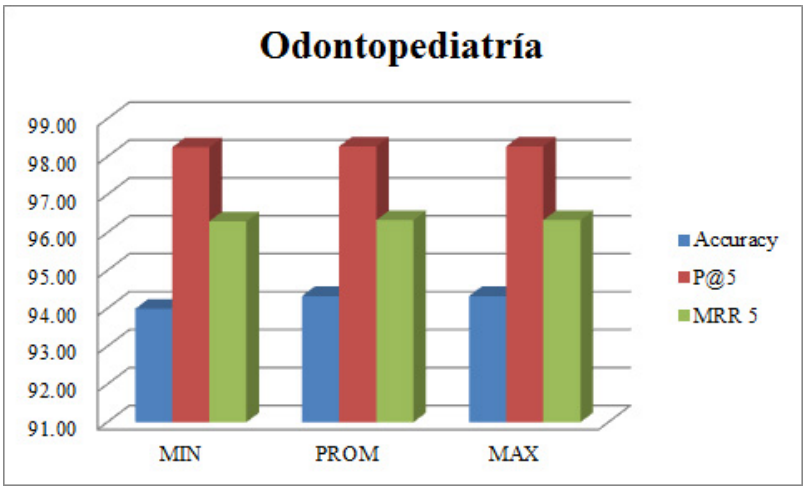

Figura 5.9: Resultados de segmentación de Odontopediatría

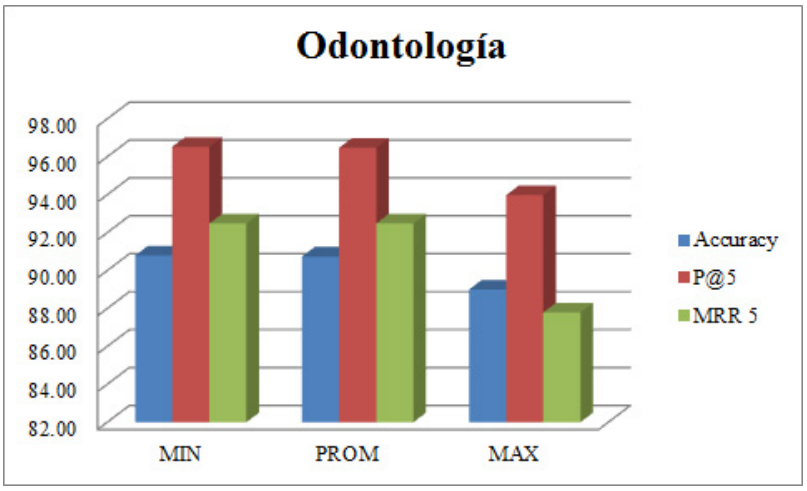

Figura 5.10: Resultados de segmentación de Odontología

5.5. Fusión de listas

Después de obtener las 3 listas por cada subconjunto de datos se procedió a la fusión de datos usando el método CombMNZ, donde solo nos quedaremos con las 10 clases más similares.

A continuación se presentan los resultados de la experimentación con segmentación y fusión de listas comparado a usar todo el diagnóstico sin segmentación y fusión.

Cabe señalar que se realizaron pruebas con otros métodos posicionales como son CombMIN y CombMAX y CombSUM y finalmente con CombMNZ obteniendo este los mejores resultados.

Tabla 5.7: Resultados obtenidos empleando el algoritmo k-NN con fusión de listas comparado con Naive Bayes.

servicio	P(%)NB	P@5(%)NB	MRR5(%)NB	P(%)kNN-F	P@5(%)kNN-F	MRR5(%)kNN-F
Inhaloterapia	57.59	60.65	61.21	65.45	71.73	68.21
Clínica del dolor	76.25	77.16	76.55	88.01	92.28	89.56
Endocrinología	87.03	88.73	87.86	94.25	97.18	95.49
Odontopediatría	87.58	93.13	90.36	95.06	98.14	96.51
Odontología	85.01	93.73	89.02	91.70	98.37	94.86
Cardiología	68.08	75.68	71.02	84.15	93.42	88.08
Nefrología	87.81	90.11	88.9	96.79	98.16	97.35
Alergología	91.33	93.54	92.39	96.59	98.56	97.49
Gastroenterología	75.04	86.1	79.79	93.10	97.15	94.89
Dermatología	79.45	82.3	80.61	88.81	95.95	92.06
Modulo materno	26.9	34.44	29.76	55.27	83.47	66.24
Angiologia	54.67	56.61	55.62	71.25	81.04	75.29
Cardiología pediátrica	64.78	72.27	68.98	74.31	87.58	80.03
Clínica de displasia	51.78	65.67	58.28	65.37	93.73	76.96

Al realizar la fusión se puede observar que favorece primordialmente a la exactitud del código y el MRR mejora en menor proporción a la exactitud.

Inhaloterapia es el servicio con mayor inexistencia de clases en el conjunto de entrenamiento (26.70 %), su conjunto de datos cuenta con textos de 1 o 2 palabras y con existencia en su mayoría de abreviaturas y logró una exactitud de 65.45 % con k-nn superando a Naive Bayes en un 7.8 %.

Aunque la fusión de listas no funcionó claramente en el servicio de Inhaloterapia fue el servicio que no tuvo diagnósticos secundarios en el texto médico obteniendo un resultado final de 65.4 %, mejorando al k-nn de unigramas por tan solo 0.62 %.

Dermatología obtuvo un mejor desempeño en fusión de listas con un 9.57 % sobre la clasificación usando el texto completo. Cuenta con una dimensionalidad baja de vocabulario lo que nos dice que comparten el vocabulario fuertemente entre todas sus instancias, bajo desbalanceo de clases pero la longitud de sus documentos es alta, aunque las clases se encuentran bien definidas, si existen diagnósticos con textos secundarios que pueden entorpecer la clasificación y al segmentarlos y fusionarlos mejoramos la asignación.

podemos notar en los resultados que la fusión de listas favoreció a aquellos servicios que contaban con mas diagnósticos secundarios como lo son Cardiología y Módulo

materno. También tenemos a Inhaloterapia que no contaba con diagnósticos secundarios y solo pudo ser beneficiado por n-gramas de caracteres.
El desempeño del sistema nos brinda buenos resultados tomando en cuenta el número de clases y el desbalanceo en diversos subconjuntos de datos. Mejora considerablemente al método ya conocido (Naive Bayes) que se empleó como base para nuestra investigación al igual que usando el texto completo con el algoritmo k-nn con unigramas de caracteres. Al considerar un escenario con instancias del mundo real podemos observar que nuestro método es aplicable a escenarios donde el conjunto de datos cuenta con diagnósticos secundarios y textos médicos con errores de escritura. Nuestro objetivo es poder emplear este sistema en este tipo de conjunto de datos, donde la generación de diagnósticos es rápida y se tiene miles de clases. En el estado del arte podemos observar diversos trabajos empleados en un conjunto de datos basados en diagnósticos de radiología que solo contaban con aproximadamente 45 códigos(clases), en esta investigación se buscó tratar un conjunto de diversas clases en un ambiente real, pudimos observar que los diagnósticos secundarios en los textos médicos podían caer en un error de clasificación, es por ello que se trató con la fusión de listas y la segmentación de diagnósticos para poder equilibrar la clasificación y dar otro peso a las clases con la segmentación. Otro punto importante surge del hecho que en el estado del arte se puede realizar la asignación multi-etiqueta de un código CIE, mientras que en nuestra investigación, llevando a cabo las reglas del hospital general que nos brindo el conjunto de datos solo se podía asignar un código CIE. Esto nos dificulto la clasificación por la existencias de mas de un diagnostico en el texto médico que también podría contar con su respectivo código CIE.

Capítulo 6

Conclusiones

En la investigación presentada se realizó una revisión del estado del arte para la asignación automática de códigos CIE-10. De igual manera se organizaron y en listaron las tareas que engloban la asignación de códigos CIE así como los diversos métodos empleados para dicha tarea. Resultado de esta investigación se presentó un sistema de recomendación de códigos CIE-10 que ayudará al experto etiquetador a asignar códigos a diagnósticos médicos. Considerando que este proceso es de suma importancia pues maneja información médica que puede ser usada para estadísticas o informes en hospitales médicos (planteando un problema del mundo real) es de suma importancia reducir los errores al asignar un código es por ello que el brindar una lista de posibles códigos al experto facilita enormemente esta tarea y le deja al etiquetador la última elección para la asignación de código.

Respecto a los métodos reportados en el estado del arte se obtuvo la conclusión que cada conjunto de datos cuenta con características diferentes y por ello el método empleado para atacar el problema dependerá de diversos factores como la existencia de diagnósticos secundarios, si se puede emplear la clasificación multi-etiqueta, si el conjunto de datos cuenta con clases de diversas áreas, la información empleada como la nota médica, el diagnóstico definitivo, las prescripciones o el historial del paciente, etcetera.

6.1. Conclusiones

Finalmente las conclusiones que se pueden extraer del trabajo realizado son:

- La utilización de n-gramas de caracteres para la representación de diagnósticos médicos ayuda a minimizar los errores de escritura o de ortografía cuando el conjunto de datos proviene de un problema real, textos médicos que son capturados en consultas médicas en tiempo real. En el estado del arte no existe evidencia de sistemas que hagan uso de este tipo de atributos para la representación de textos médicos.

- La segmentación de diagnósticos médicos y la fusión de listas integradas en el sistema logran abordar el problema de la presencia de diagnósticos secundarios dentro del mismo texto médico equilibrando nuestro conjunto de datos.

- Las investigaciones destinadas a apoyar la codificación clínica han sido basadas en la información textual y, por tanto, sufren cuestiones de aplicabilidad y generalidad. Por otra parte, los corpus utilizados en cada estudio varían de manera significativa, no sólo en el ámbito de aplicación de las condiciones clínicas, sino también en el grado de estructura del documento (por ejemplo, los textos dictados vs textos producidos al instante o durante la prestación de una consulta médica) y el lenguaje empleado en las narrativas, las cuales pueden contener abreviaturas o palabras usadas en la jerga médica. Otro rasgo en particular es la información extraída del paciente, mientras que existen algunos corpus que se basan únicamente en el texto médico del diagnóstico hay otros que cuentan con textos referentes a recetas médicas del paciente, información de sexo y edad o el historial clínico completo del paciente. En nuestra investigación se buscó generar un sistema de asignación de códigos que pueda emplearse en corpus de diagnósticos robustos reales producidos por hospitales bajo ciertas condiciones; donde cada diagnóstico se redacta al finalizar cada consulta médica, escrita por el médico en turno usando términos o jerga médica para describir cada padecimiento.

6.2. Trabajo Futuro

Tomando en cuenta la investigación y las conclusiones que se presentan en este trabajo se pretende explorar en un futuro:

- Realizar un análisis sobre diferentes atributos que aporten mayor información a la asignación como la información de recetas médicas del paciente, que puede ser utilizada como nuevos atributos en el vector de características.
- Evaluar el desempeño del sistema con otros conjuntos de datos reales.
- Integrar el sistema a una interfaz amigable para un experto etiquetador.

Bibliografía

[1] Eiji Aramaki, Takeshi Imai, Masayuki Kajino, Kengo Miyo, y Kazuhiko Ohe. Automatic matching of icd-10 codes to diagnoses in discharge letters. *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics*, págs. 645–649, 2007.

[2] C Benesch, DM Jr Witter, AL Wilder, PW Duncan, GP Samsa, y DB Matchar. Inaccuracy of the international classification of diseases icd-9-cm in identifying the diagnosis of ischemic cerebrovascular disease. *Neurology*, pág. 35, 1997.

[3] C.M. Bishop. *Pattern Recognition and Machine Learning*. McGraw-Hill Science/Engineering/Math, 2006.

[4] Svetla Boytcheva. Automatic matching of icd-10 codes to diagnoses in discharge letters. *In Proceedings of the Second Workshop on Biomedical Natural Language Processing, associated to RANLP-2011*, págs. 19–26, Septiembre,2011.

[5] W. W. Chapman, P. M. Nadkarni, L. Hirschman, L. W. D'Avolio, G. K. Savova, y O. Uzuner. "overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions". *Journal of the American Medical Informatics Association*, Volume 18, no. 5:540–543, 2011.

[6] Hsu D. F. y Taksa I. Comparing rank and score combination methods for data fusion in information retrieval. *Information Retrieval*, 8(3), Year = 2005:449–480.

[7] Richárd Farkas y Gyorgy Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC Bioinformatics*, Volume 9, Issue 3:1471–2105, Abril 2008.

[8] Jose .C Ferrao. Using structured ehr data and svm to support icd-9-cm coding. *Healthcare Informatics (ICHI), 2013 IEEE International Conference*, pág. 511 – 516, 2013.

[9] Goldstein I, Arzumtsyan A, y Uzuner O. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annu Symp Proc*, Volume 13, No. 5:279–83, 2007.

[10] Crammer K, M , Dredze, K Ganchev, y P Patrim. Automatic code assigment to medical text. department of computer and information science. *University of Pennsylvania, Philadelphia, PA*, págs. 36–40, Prague, Czech Republic; 2008.

[11] Laurent Kevers y Julia Medori. Symbolic classification methods for patient discharge summaries encoding into icd. *In Proceedings of the 7th International Conference on NLP (IceTAL)*, págs. 197–208, 2010.

[12] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, y J. F. Hurdle. "extracting information from textual documents in the electronic health record: a review of recent research". *Yearbook of Medical Informatics*, págs. 138–154, 2008.

[13] Serguei V. Pakhomov, James D. Buntrock, y Christopher G. Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association : JAMIA*, Volume 13, No. 5:516–525, 2007.

[14] Suzanne Pereira, Aurélie Névéol, Philippe Massari, Michel Joubert, y Stéfan Jacques Darmoni. Construction of a semi-automated icd-10 coding help system to optimize medical and economic coding. *MIE*, volume 124 of Studies in Health Technology and Informatics:845–850, 2006.

[15] J. Pestian, C. Brew, P. Matykiewicz, DJ Hovermale, N. Johnson, K. B. Cohen, y D. Wlodzislaw. A. shared task involving multi-label classification of clinical free text. in: Acl'07 workshop on biological, translational, and clinical language processing. *Bio=LP'07*, págs. 36–40, Prague, Czech Republic; 2007.

[16] David Pinto, Paolo Rosso, y Héctor Jiménez-Salazar. On the assessment of text corpora. *Natural Language Processing and Information Systems*, págs. 281–290, 2010.

[17] H Suominen, F Ginter, S Pyysalo, A Airola, T Pahikkala, S Salantera, y T Salakoski. Machine learning to automate the assignment of diagnosis codes to free-text radiology reports: a method description. *University of Turku, Department of Information Technology, Finland*, 2008.

[18] H. J. Tange, H. C. Schouten, A. D. M. Kester, y A. Hasman. The granularity of medical narratives and its effect on the speed and completeness of information retrieval. *Journal of the American Medical Informatics Association : JAMIA*, Volume 5:571–582, 1998.

[19] Joachims Thorsten. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, págs. 137–142, 1998.

[20] Brankov Yourganov Wernick, Yang y Strother. Machine learning in medical imaging. *IEEE Signal Processing Magazine*, Volume 27:25–38, 2010.

[21] Richard Wolniewicz. Auto-coding and natural language processing. *3M Health Information Systems*, 2001.

[22] Jian-Wu Xu, Shipeng Yu, Jinbo Bi, y L.V. Lita. Automatic medical coding of patient records via weighted ridge regression. *Sixth International Conference on Machine Learning and Applications, Cincinnati OH, USA*, págs. 260–265, 2007.

[23] Yan Yan, Glenn Fung, Jennifer G. Dy, y Rómer Rosales. Medical coding classification by leveraging inter-code relationships. *International Conference on Knowledge Discovery and Data Mining*, volume 124 of Studies in Health Technology and Informatics:193–202, 2010.