UNIVERSIDADE DA CORUÑA

Departamento de Computación

Clasificación Automática de Documentación Clínica

TESIS DOCTORAL

J. David Lojo Vicente

2012

UNIVERSIDADE DA CORUÑA

Tesis Doctoral

Clasificación Automática de Documentación Clínica

J. David Lojo Vicente

Directores

Prof. Dr. Álvaro Barreiro García

Prof. Dr. David E. Losada Carril

Dr. **Álvaro Barreiro García**, Catedrático de Universidad en el área de Ciencias de la Computación e Inteligencia Artificial de la Universidade da Coruña y el Dr. **David Enrique Losada Carril**, Profesor Titular en el área de la Computación e Inteligencia Artificial de la Universidade de Santiago de Compostela

HACEN CONSTAR:

Que la memoria titulada **Clasificación Automática de Documentación Clínica** ha sido realizada bajo nuestra dirección y constituye la Tesis que presenta para optar al grado de Doctor por la Universidade da Coruña.

A Coruña, septiembre de 2012

Firmado: Dr. **Álvaro Barreiro García**
Director de la tesis

Firmado: Dr. **David Enrique Losada Carril**
Director de la tesis

Firmado: **José David Lojo Vicente**
Autor de la tesis

A Mila e Irene

El aprendizaje automático es el estudio de algoritmos computacionales que van mejorando automáticamente su desempeño a través de la experiencia.

Tom Mitchell

Hacer una tesis significa divertirse y la tesis es como el cerdo, en ella todo tiene provecho.

Humberto Eco

Agradecimientos

Quiero agradecer a mis directores de tesis Álvaro Barreiro y David Losada la labor desarrollada estos años de dedicación, apoyo y paciencia. Gracias por iniciarme y refinarme en todas la tareas relacionadas con la investigación.

Al grupo de investigación IRLAB del Departamento de Computación de la Universidade da Coruña y al Grupo de Sistemas Inteligentes (GSI) del Departamento de Electrónica y Computación de la Universidade de Santiago de Compostela por invitarme a todos los actos científicos que han organizado.

A todo el personal del servicio de Medicina Interna del Hospital de Conxo de la Xerencia de Xestión Integrada de Santiago de Compostela que han elaborado los informes de alta en el periodo 2003 – 2005, repositorio documental con la que se ha construido la colección base de los experimentos.

No me puedo olvidar de aquellos que me animaron y apoyaron a iniciar esta aventura, Ramón Pérez Otero, mi amigo. También a Jorge González por su apoyo en la realización del DEA.

Un cariñoso agradecimiento a las personas más cercanas, mi familia y mis amigos, a los que a partir de ahora podré dedicarles algo más de mi tiempo.

A Mila e Irene por su critica constructiva y la aportación de frases que han quedado en la jerga familiar "*¿qué, vai ou non vai?*", seguramente metáfora de algún adjetivo calificativo.

La alegría y emoción de mi familia y como no, la de mis directores ante su conclusión, seguramente supera la mía, gracias a todos.

Abstract

In hospitals, huge amounts of complex data are daily produced. Manually labeling every produced document is not an option because of the limited resources. One of the clinical classification tasks is the coding of diagnoses from discharge reports. Coding is a process that consists of analysing the discharge documentation and assigning the diagnostic codes associated to the clinical episode.

This doctoral dissertation aims at investigating Automatic Text Classification (ATC) in a complex area: clinical documentation. This is a supervised learning scenario, where the classes are ICD-9-CM codes and the documents are clinical discharge summaries. We use different classification strategies, such as nearest algorithm (*knn*) and Support Vector Machines (SVMs). A key contribution of this study is the construction of a new test collection from the discharge reports of a clinical service (documents written in Spanish). It is a difficult testbed because of the large number of classes, the average number of classes per document, and the lack of balance among classes. We study different representations of the documents, different retrieval models and the effect of weighting on the classification. The final objective is to build a system to assist the coders with the assignment of ICD-9-CM codes.

In addition, we also analyse Active Learning (AA) as a tool to select which documents should be coded. This helps to make good training sets and, therefore, it is a promising avenue to improve clinical classification systems.

Resumen

En los hospitales, se producen diariamente grandes cantidades de datos complejos. Puesto que los recursos humanos son limitados, la clasificación manual de los documentos producidos no es una alternativa óptima. Una de las tareas de la clasificación de la documentación clínica es la codificación de los informes de alta. La codificación es un proceso que consiste en analizar la documentación del alta, y asignar códigos de los diagnósticos de ese episodio clínico.

Esta tesis doctoral tiene como objetivo investigar la Clasificación Automática de Textos (CAT) en un área compleja: la documentación clínica. Este es un escenario de aprendizaje supervisado, donde las clases son los códigos CIE-9-MC y los documentos son los informes de alta hospitalaria. Se utilizan diferentes estrategias de clasificación, tales como los algoritmos de vecindad (*Knn*) y las Máquinas de Soporte Vectorial (SVM). Una contribución fundamental de este estudio es la construcción de una nueva colección de informes de alta de un servicio clínico (documentos escritos en español). Es un banco de pruebas difícil por la gran cantidad de clases, el número medio de clases por documento, y la falta de equilibrio entre las clases. Se estudian diferentes representaciones de los documentos, distintos modelos de recuperación y el efecto de la ponderación en la clasificación. El objetivo final es construir un sistema de ayuda a los codificadores en la asignación de códigos CIE-9-MC.

También investigamos en Aprendizaje Activo (AA) como una herramienta para seleccionar qué documentos deben ser codificados. Esto ayuda a formar buenas colecciones de entrenamiento y, por lo tanto, es una vía prometedora para mejorar los sistemas de clasificación clínicos.

Resumo

Nos hospitais, prodúcense diariamente gran cantidade de datos complexos. Como os recursos humanos son limitados, a selección manual dos documentos producidos non é unha alternativa ideal. Unha tarefa de clasificación da documentación clínica é a codificación dos informes de alta. A codificación é un proceso que consiste en analizar a documentación de alta, e asignar códigos dos diagnósticos de ese episodio clínico.

Esta tese de doutoramento ten como obxectivo investigar a Clasificación Automática de Textos (CAT), nunha área complexa: a documentación clínica. Este é un escenario de aprendizaxe supervisada, onde as clases son CIE-9-MC e os documentos son os informes de alta hospitalaria. Emprégase diferentes estratexias de clasificación, tales como os algoritmos de veciñanza (*Knn*) e as Máquinas de Soporte Vectorial (SVM). A contribución fundamental deste estudo é a construción dunha nova colección de informes de alta dun servizo clínico (documentos escritos en español). É un banco de probas difícil pola gran cantidade de clases, o número medio de clases por documento, e a falta de equilibrio entre as clases. Estudamos diferentes representacións de documentos, distintos modelos de recuperación e os efectos da ponderación na clasificación. O obxectivo final é a construción dun sistema de apoio para os codificadores na asignación de códigos CIE-9-MC.

Tamén investigamos en Aprendizaxe Activa (AA) como unha ferramenta para seleccionar que documentos deben ser codificados. Isto axuda a formar boas coleccións de adestramento e, polo tanto, é un camiño esperanzador para mellorar os sistemas de clasificación clínica .

Índice

Lista de Figuras

Lista de Tablas

Introducción

En los hospitales se genera un gran volumen de información con considerable complejidad. La capacidad de clasificación manual es limitada por lo que es imposible que todos los documentos producidos sean etiquetados. Una de las tareas de clasificación que se realizan es la codificación de los diagnósticos de los informes de alta. La codificación es un proceso que consiste en analizar la documentación del alta, y asignar los códigos de los diagnósticos de ese episodio clínico. Este proceso se realiza de forma manual por un médico codificador, con un gran coste por la complejidad del tipo de clasificación. En los hospitales, los episodios que se codifican habitualmente son los ingresos hospitalarios. Si quisiésemos codificar todos los episodios clínicos que se generan en un centro hospitalario, tendríamos que aumentar de forma considerable los recursos humanos de médicos codificadores, lo que implicaría un elevado coste económico. Debido a estas limitaciones los episodios clínicos pasan usualmente por una clasificación generalista, simplemente para generar una contabilidad básica, sin considerar la patología tratada para cada paciente. En cambio, con la codificación CIE-9-MC completa de estos episodios podríamos medir, comparar y mejorar la calidad asistencial, agrupando a los pacientes de acuerdo a requerimientos y características comunes.

Los objetivos que se pretende alcanzar con esta tesis doctoral es investigar las posibilidades que nos ofrece la clasificación automática de textos en un entorno tan complejo como la documentación clínica. Se sitúa en un entorno de aprendizaje supervisado, en donde las clases son los códigos CIE-9-MC y los documentos son los informes de alta hospitalaria. Los sistemas de clasificación que se utilizan para la asignación de códigos CIE-9-MC a un documento nuevo son algoritmos de vecindad (*Knn*) y Máquinas de Soporte Vectorial (SVM). Uno de los valores añadidos de este trabajo es la construcción de la colección, en castellano, a partir de los informes de alta de un servicio médico. Esta es una colección difícil por la gran cantidad de clases, el número de clases por documento y la descompensación entre las clases. Se estudian

diferentes representaciones de la colección, distintos modelos de recuperación y el efecto de los sistemas de pesado en la asignación de códigos CIE-9-MC. El objetivo final es construir un sistema de ayuda a la codificación de informes de alta de hospitalización u otro tipo de documentación clínica, que se pueda implementar y valorar en un centro sanitario.

En los hospitales se genera un gran volumen de información, pero sólo se codifica una pequeña parte de los informes producidos. Es por tanto un escenario donde se necesita elegir bien lo que se etiqueta para que las herramientas automatizadas de clasificación puedan surtirse de buenos conjuntos de entrenamiento, para ello utilizaremos también técnicas de Aprendizaje Activo. La evaluación de los resultados de estos procesos de selección de Aprendizaje Activo nos demuestra que esta estrategia es prometedora para mejorar este tipo de sistemas.

La estructura resultante de la Tesis doctoral es la que se detalla a continuación:

- Capítulo 1. Documentación Clínica.

  En este capítulo se define la documentación clínica, las funciones que realiza, como se gestiona, como se trasforma en actividad asistencial y producción hospitalaria a partir de la codificación de sus diagnósticos. Se describe los sistemas de terminología médica, puntualizando la Clasificación Internacional de Enfermedades (CIE-9-MC). Se detalla el proceso de codificación sobre el informe de alta hospitalaria, terminando con una valoración y análisis de los sistemas de ayuda a la codificación.

- Capítulo 2. Clasificación Automática de Textos.

  Se define la clasificación de textos, los tipos de clasificación, las posibles representaciones de los documentos, las técnicas de clasificación de textos, sus métodos y métricas de evaluación y los escenarios posibles para la comparación de métodos de clasificación.

- Capítulo 3. Clasificación de códigos CIE-9-MC con algoritmos de vecindad y Máquinas de Soporte Vectorial.

  Se especifica el proceso de creación de la colección MIR-Conxo. Análisis de la colección y comparativa con otras colecciones similares. Se describe la clasificación de textos y el procedimiento de clasificación para *knn* y SVM. Se

muestran los resultados de los experimentos para *Knn* y SVM y se realiza un análisis y una comparativa de las técnicas utilizadas, finalizando con las conclusiones y los posibles trabajos a realizar en un futuro.

- Capítulo 4. Evaluación de técnicas de Aprendizaje Activo para codificación CIE-9-MC de informes de alta hospitalaria.
  Se explica el aprendizaje activo para la clasificación de textos multietiqueta, las metodologías para evaluar aprendizaje activo, los resultados de los experimentos realizados y las conclusiones a las que hemos llegado.

- Capítulo 5. Conclusiones e investigaciones futuras.
  Se resumen las conclusiones que hemos obtenido en esta tesis y las posibles líneas de investigación a seguir.

Capítulo 1

Documentación Clínica

Gran parte de la información sobre la salud se basa en las terminologías médicas, y responde a una diversidad de criterios y disciplinas, de las distintas especialidades y ramas sanitarias, expresando la variabilidad de la actividad asistencial. Clasificar es empezar a comprender, por lo cual es necesario establecer una clasificación homologada internacionalmente que permita la normalización de toda esa información. En este capítulo se presenta un resumen de la gestión de la documentación clínica, una revisión de los sistemas de terminología médica, con especial detalle en la Clasificación Internacional de Enfermedades 9.ª revisión Modificación Clínica (CIE-9-MC), una descripción de los procesos de la codificación médica, terminando con un análisis de la situación de los sistemas de clasificación para CIE-9-MC. En el ámbito hospitalario este tipo de clasificación nos va a permitir construir un sistema de clasificación de pacientes desde el punto de vista clínico y definir un sistema de producción hospitalario.

1.1. Documentación Clínica

Se define la información clínica como todo dato, cualquiera que sea su forma, clase o tipo, que permite adquirir o ampliar conocimientos sobre el estado físico y la salud de una persona, o la forma de preservarla, cuidarla, mejorarla o recuperarla. La información relativa al estado de salud de un ciudadano está íntimamente ligada al ciclo de su vida y se va enriqueciendo desde antes de su nacimiento hasta (en algunas circunstancias) más allá del fallecimiento. Para que esto se produzca se precisan dos condiciones: que exista un contacto con un profesional sanitario y que este acto quede debidamente documentado. Estos actos médicos se documentan sobre diferentes soportes: papel, registros informáticos, estudios radiológicos, vídeo, registro de señales

analógicas, registro de señales digitales, etc. La Organización Mundial de la Salud (OMS) define el Sistema de Información Sanitaria como una estructura para la recogida, elaboración, análisis y transmisión de la información necesaria para organizar y hacer funcionar los servicios sanitarios, siendo la investigación y la docencia objetivos complementarios.

En nuestro entorno y para nuestra legislación se define la historia clínica (HC) como el conjunto de documentos que contiene datos, valoraciones e informaciones de cualquier índole sobre la situación y la evolución clínica de un paciente a lo largo del proceso asistencial. Está comúnmente aceptado que la HC está constituida por el conjunto de documentos escritos y/o gráficos que hacen referencia a los episodios de salud-enfermedad de un ciudadano y a la actividad sanitaria generada por estos, independientemente del soporte en que se encuentre. La HC se asocia comúnmente con el tradicional soporte papel. La introducción de la informática como herramienta de soporte de la información sanitaria, no varía los principios rectores de la HC, ni invalida las definiciones o propiedades de la misma. La finalidad de la HC es facilitar la asistencia del ciudadano, recogiendo toda la información clínica necesaria para asegurar, bajo un criterio médico, el conocimiento veraz, exacto y actualizado de su estado de salud por los sanitarios que lo atienden. A su vez, la HC se considera el documento clínico por excelencia, al ser el soporte de la información generada por el equipo sanitario y actuar como vehículo de transmisión entre los diferentes miembros que intervienen en la atención, o para otros equipos que puedan prestar atención sanitaria al ciudadano en otro lugar o tiempo.

Las funciones clásicas de la HC son la asistencial, la docencia y la investigación. Desde estas, se desarrollan otras, que estando íntimamente ligadas a ellas, tienen la suficiente trascendencia como para ser destacadas:

- **Asistencial.** Es un documento básicamente asistencial, siendo su misión principal recoger toda la información patográfica relevante, con objeto de poder prestarle al ciudadano la atención más adecuada a su caso.
- **Docente.** Cuando en cada HC se refleja exactamente cuál es el modo correcto de tratar cada caso clínico, explicando razonadamente las decisiones exploratorias y terapéuticas que se toman.
- **Investigación clínica.** Estableciendo los mecanismos precisos para localizar las historias clínicas que pertenecen a una determinada patología, o a un

determinado tratamiento y como fuente de conocimiento de la propia actividad clínica.

- **Investigación epidemiológica.** Cuando además de conocer lo anterior, se conocen los denominadores poblacionales adecuados.
- **Gestión clínica y planificación de recursos asistenciales.** Sirve para la gestión clínica, la evaluación de la utilización de los recursos sanitarios disponibles y la planificación de futuras inversiones.
- **Jurídico-legal.** Al ser el testimonio documental de la asistencia prestada.
- **Controles de calidad asistencial.** Las HC sirven para la evaluación de los objetivos científico-técnicos.

La HC será acumulativa cuando toda la información clínico-sanitaria que genera la asistencia de un paciente, independientemente del soporte en que se presente, pase a formar parte de la misma. Esta será integrada cuando contenga las distintas relaciones y/o episodios del paciente, definiéndose éstos como los distintos actos asistenciales relacionados con un proceso de atención sanitaria.

Las fuentes de información son múltiples y los usos de la HC tan variados que los datos que potencialmente se necesitan deben ser asimismo múltiples y variados. No por ello debemos considerar como relevante cualquier información o dato que no reúna la condición de importante o significativo si no queremos desvirtuar su significado y sobre todo, si no queremos llenar la HC de informaciones inútiles que nos entorpezcan la búsqueda de las importantes. La HC es la herramienta de trabajo de los profesionales sanitarios y, como tal herramienta, ha de contener todo aquello que facilite su tarea.

Los actos sanitarios se caracterizan fundamentalmente por diagnosticar y tratar a los pacientes que demandan asistencia. Un médico u otro profesional sanitario atiende a diversos pacientes y no siempre los actos asistenciales se repiten en períodos cortos de tiempo (caso de las consultas externas), por lo que no se puede pretender (sin degradar la calidad asistencial) que utilice su capacidad memorística para recordar esas anotaciones sobre cada uno de sus pacientes. Por otro lado no siempre es el mismo interlocutor sanitario quien atiende a un paciente, y habitualmente, no solo un determinado profesional de una especialidad concreta es el único encargado de prestar

la asistencia. La atención sanitaria es multidisciplinar y la HC es el medio de comunicación entre los distintos profesionales que intervienen en dicha atención

Los principios de la Documentación Clínica se establecieron a finales de los 80 y comienzos de los 90, con la creación de los Servicios de Admisión y Documentación Clínica (SADC) en los hospitales. En este periodo empezaron a funcionar los Archivos de Historias Clínicas centralizados, dentro del contexto de la introducción de nuevos modelos de gestión sanitaria.

En los hospitales siempre ha existido personal, generalmente ligado a la administración del centro, entre cuyos cometidos estaba el registro de pacientes, y el archivado de los documentos generados en la asistencia a los pacientes. Esto surge como respuesta al incremento de la demanda asistencial, con el fin garantizar la equidad en el acceso y de optimizar los recursos disponibles. Por otro lado, el aumento de la complejidad de los procesos asistenciales obliga a la creación de un sistema de información que facilite la gestión y el control de la calidad. En este contexto se constituyen los SADC, como una estructura central de apoyo al funcionamiento de los hospitales, orientados a la integración, ordenación y coordinación de la actividad hospitalaria y quedando adscritas a la Gerencia del hospital las siguientes áreas de actividad: Admisión, recepción e información y a la Dirección médica del hospital los servicios y unidades que incluyan el área de actividad de Documentación y archivo de Historias Clínicas. Este modelo organizativo parece adecuado para grandes hospitales, pioneros en la creación de estos servicios, en donde coexisten como servicios diferentes la Admisión y la Documentación Clínica. Por el contrario se muestra poco adecuado para hospitales pequeños y medianos, dado que es el mismo servicio para la gestión de pacientes como para la documentación clínica.

1.1.1. Gestión de la documentación clínica

La gestión de la documentación clínica tiene como objetivo organizar y gestionar toda la información clínica generada a lo largo de los sucesivos procesos asistenciales del paciente. Aspectos fundamentales sobre esta gestión están regulados por la Ley 41/2002, de 14 de noviembre, básica reguladora de la autonomía del paciente y de derechos y obligaciones en materia de información y documentación clínica; así como por desarrollos normativos autonómicos de dicha ley.

La gestión de la documentación clínica se concreta en tres aspectos fundamentales: gestionar la HC, los archivos de documentación e historias clínicas y la codificación clínica.

Gestionar la HC: En general todos los aspectos relacionados con la gestión de la HC implica:

- Identificar la HC: creación, actualización y mantenimiento del fichero de pacientes en el centro, garantizando su coherencia, integridad y fiabilidad, así como la confidencialidad de la información.
- Diseñar y mantener actualizado el formato de la HC: normalización de la documentación clínica del centro para su correcta homogeneización, en colaboración con la comisión de historias clínicas.
- Clasificar, integrar y coordinar toda la información clínico-asistencial generada independientemente de su soporte físico (impresos, películas).
- Controlar la calidad de la HC: evaluación sistemática de la calidad formal y de contenido de los documentos empleados en la asistencia, así como elaboración de informes sobre los resultados de las evaluaciones y difusión de los mismos.
- Garantizar la accesibilidad de la historia, elaborando –en colaboración con las instancias determinadas por cada centro– la normativa acerca de la localización, el préstamo y la devolución de las historias clínicas, estableciendo mecanismos que aseguren su disponibilidad y velen por su confidencialidad.
- Gestionar y organizar los archivos de documentación e historias clínicas, asegurando que su configuración y utilización se ajustan a las previsiones contenidas en la Ley Orgánica 15/1999, de 13 de diciembre, de protección de datos de carácter personal. Supone:
- Custodiar, prestar y recepcionar las historias clínicas: supervisión y ejecución de las normas del centro que regulan el acceso y disponibilidad de la HC y de la información en ella contenida, preparación y préstamo (registro del tipo de documentación solicitada, solicitante, motivo, fines y fecha en que se necesita, desarchivado, registro de documentación prestada y envío de la misma) y recepción (registro de documentación devuelta al Archivo y archivado).
- Realizar el seguimiento de la documentación prestada: mantenimiento del registro de préstamo-recepción de la HC en el Archivo y reclamación activa de la documentación no devuelta en los plazos establecidos.

- Identificar, mantener y tratar la documentación clínica de menor probabilidad de uso asistencial posterior: definición y mantenimiento de pasivo, así como la relación activo-pasivo (reactivación).
- Evaluar la actividad y control de calidad del Archivo de historias clínicas.

Organizar y gestionar operativamente la codificación clínica:

- Definir las fuentes de datos del sistema de información clínico.
- Tratar la información clínica extraída de las historias clínicas:
- Indización: identificación de diagnósticos y procedimientos, selección de diagnóstico principal y codificación a través del sistema de clasificación vigente actualmente, la Clasificación Internacional de Enfermedades. 9ª Revisión Modificación Clínica (CIE-9-MC).
- Elaboración y validación de la información recogida en el episodio asistencial para configurar el Conjunto Mínimo Básico de Datos (CMBD) de ingresos, cirugía ambulatoria, hospital de día u otras modalidades asistenciales de las que se defina el CMBD correspondiente: captura de datos administrativos del episodio, registro informatizado de datos clínicos resultantes del proceso de codificación y validación.
- Recuperación, análisis y difusión de la información tratada: envío del CMBD a los organismos oficiales correspondientes, elaboración y difusión del cuadro de mandos del CMBD, realización de búsquedas o informes "ad hoc" para satisfacer las necesidades de información de los usuarios internos y elaboración y difusión del análisis de la casuística, utilizando sistemas de clasificación de pacientes basados en el CMBD.
- Controlar la calidad del sistema de información clínico, desarrollando medidas para garantizar su fiabilidad: evaluación de los documentos fuente y circuitos de información establecidos, auditorías internas y externas del proceso de codificación, revisión sistemática de indicadores de calidad del CMBD (registros agrupados en inespecíficos y otros) y establecimiento de mecanismos de retroalimentación continua de los usuarios internos.

1.1.2. Volumen de la documentación clínica

Uno de los lugares en donde se genera un mayor volumen de información y con considerable complejidad son los centros sanitarios, y en concreto los centros

hospitalarios. Los factores que determinan este volumen en los hospitales del Sistema Nacional de Salud (SNS) de España (Sistema Nacional de Salud de España 2010. Madrid. Ministerio de Sanidad y Política Social, Instituto de Información Sanitaria. Disponible en: http://www.msps.es/organizacion/sns/librosSNS.htm), son sus recursos humanos con 202.355 profesionales sanitarios y algunos de los datos principales de su actividad en el 2008, con 5,2 millones de ingresos, 77,1 millones de consultas, 26,3 millones de urgencias y 4,5 millones de actos quirúrgicos. De la propia actividad de estos profesionales nace esta información, que se representa casi siempre en un documento de texto. Así podemos afirmar que cada acto médico tiene asociado como mínimo un documento, que puede estar representado por una sola línea o por múltiples de líneas de texto. Es de fácil deducción, con los datos anteriores, que en los hospitales del SNS de España se escriben como mínimo más de 100 millones de documentos por año.

1.1.3. Actividad Asistencial, producción hospitalaria y codificación

El Sistema Nacional de Salud se organiza en dos niveles diferenciados de Atención Sanitaria, uno caracterizado por actos sanitarios personales, más en contacto con el entorno que rodea al paciente, que es la Atención Primaria, y otro con actos sanitarios más complejos, que requieren generalmente la utilización de más recursos, caracterizado por el trabajo en equipo, donde la atención sanitaria se efectúa por diferentes profesionales, pertenecientes a su vez a diferentes estamentos (facultativos, sanitarios no facultativos, etc.) y que se presta fundamentalmente en los Hospitales.

En el caso de la Atención Especializada, considerándose ésta como el escalón más complejo de la asistencia sanitaria, tenemos tres grandes grupos de actos asistenciales que dan lugar a diferentes maneras de organizarlos y por consiguiente, deben dar lugar a diferentes modelos de entender los soportes documentales. Estos actos asistenciales son los relacionados con la hospitalización, los relacionados con las formas ambulatorias de asistencia especializada, en gran medida las consultas externas, y los relacionados con las asistencias en los Servicios de Urgencias. Los datos de actividad reflejados en el apartado anterior nos muestran que los ingresos son un 4,5% de los actos médicos, las urgencias un 23,6% y las consultas externas un 71,9%. Estos actos sanitarios se caracterizan fundamentalmente por diagnosticar y tratar a los pacientes que demandan asistencia. Pero el tratamiento de estos procesos asistenciales de cara al diagnóstico es diferente. En la actualidad solo los diagnósticos de los ingresos

hospitalarios son codificados, es decir un 4,5%, el resto de episodios no se codifica, en términos generales. Al no disponer de diagnósticos codificados no podemos aplicar sistemas para medir producción hospitalaria de forma más exhaustiva. Disponer de una codificación de diagnósticos nos proporciona el Conjunto Mínimo Básico de Datos (CMBD) y a su vez los Grupos Relacionados por el Diagnóstico (GRD), como observamos en la Figura 1.1. Este proceso solo se realiza para los episodios de hospitalización, un 4,5% del total, en el resto de episodios (consultas externas, urgencias) en la mayoría de los casos no se codifican, por ello no podemos obtener ni el CMBD ni los GRD de estos episodios clínicos. El motivo de este limitado alcance en la codificación es el alto coste que conlleva a nivel de recursos humanos de médicos codificadores. Si estos episodios fuesen codificados tendríamos un mayor conocimiento clínico de los pacientes, un conocimiento normalizado y estructurado, que permitiría una mejora sustancial en los indicadores de producción hospitalarios.

La gestión está muy desarrollada en el ámbito de la hospitalización por ser un recurso muy costoso, y por lo mismo, es más precisa en la patología quirúrgica. Mientras que en la gestión de consultas y urgencias las herramientas e indicadores de gestión son muy rudimentarios, en la gestión de la hospitalización existen sofisticados métodos para medir qué se hace y cómo se hace. Los episodios ambulatorios no necesitan ingreso para resolver problemas de salud, estos se han incrementado en los últimos años. En el pasado muchos de estos episodios requerían de hospitalización. Estos episodios ambulatorios empiezan a ser cada día más complejos con un incremento de los costes. A medida que los costes hospitalarios aumentan, las administraciones quieren mejorar la eficiencia en la sanidad. Pero tenemos grandes áreas de consumo de recursos en donde los indicadores de gestión son muy básicos, y el principal motivo para no poder aplicar una gestión más precisa y eficaz, es la falta de codificación de estos episodios. La ayuda a la codificación mediante técnicas de clasificación automática nos permiten, sin un coste desorbitado, codificar todos los actos médicos que se realizan en un hospital.

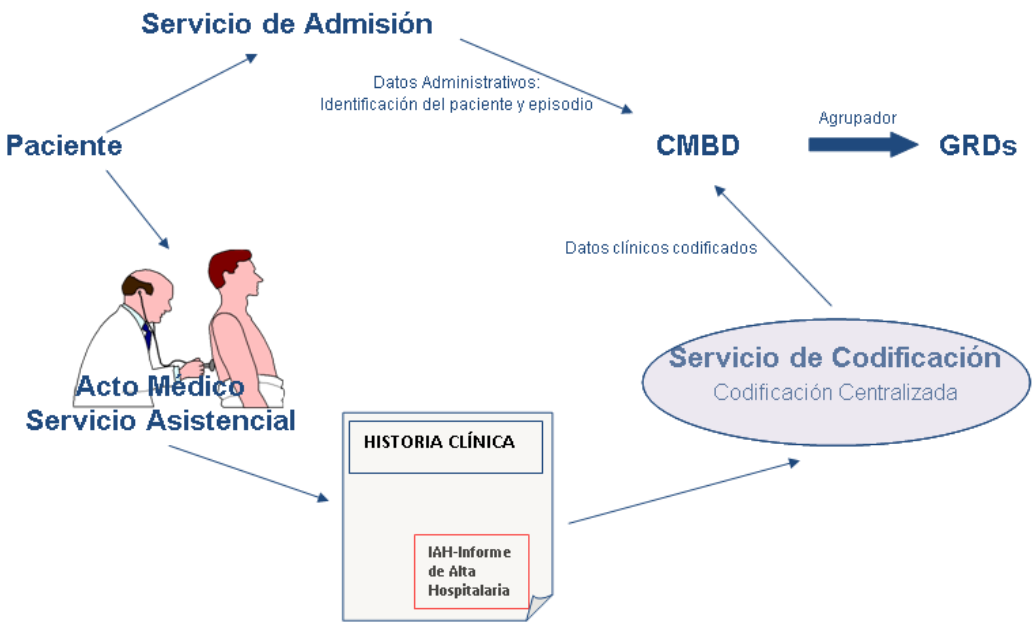

Figura 1.1: Circuito de la documentación clínica

1.1.4. Conjunto Mínimo Básico de Datos (CMBD) y Grupos Relacionados por el Diagnóstico (GRD)

En 1975 el Comité de Información y Documentación Científica y Tecnológica de la CEE creó un grupo de trabajo sobre información biomédica y de la salud con el objeto de normalizar la información clínica en los hospitales de la Comunidad. De este grupo surge un subgrupo denominado BM3 con el encargo concreto de seleccionar una serie de datos que pasarían a formar parte de estos resúmenes clínicos.

Como resultado surge el llamado *European Minimun Basic Data Set* sobre el que el 14 de diciembre de 1987 el Consejo Interterritorial incluye la propuesta de un Conjunto Mínimo Básico de Datos para todo el territorio nacional.

Podemos definir el CMBD del paciente como un conjunto de variables obtenidas en el momento del alta que proporcionan datos sobre el paciente, su entorno, la institución que lo atiende y su proceso asistencial. Representa aquella información básica cuya necesidad es común a diferentes usuarios (clínicos, gestores, planificadores, epidemiólogos, investigadores,...) sin que esto signifique que sea suficiente a cada uno de ellos.

Desde hace años en todos los países de la Unión Europea, la realización del CMBD al alta de un paciente es una práctica obligatoria para todos los hospitales. Se trata de extraer la información del paciente en su proceso de hospitalización recogiendo datos administrativos, clínicos y demográficos. El CMBD constituye una importante

herramienta para los distintos actores de la empresa sanitaria entre cuyas utilidades cabe destacar:

- Informa de la casuística hospitalaria.
- Proporciona conocimiento de las características de la morbilidad ingresada en hospitales, su frecuencia, distribución geográfica y por grupos de edad y sexo. Se convierte en instrumento de epidemiología (analítica y experimental).
- Produce información útil para la financiación, ordenación y distribución de recursos sanitarios.
- Sirve de punto de entrada para la realización de estudios clínicos específicos.
- Aproxima al conocimiento del consumo de recursos por patologías.
- Recoge información de calidad de los procesos asistidos.
- Permite introducir técnicas de agrupación de pacientes (GRDs, PMCs, ...) sirviendo como base para la identificación de las líneas de producción en los hospitales.
- Identifica los movimientos geográficos y utilización del hospital por parte de la población. Sustituye y mejora la información obtenida a través de la Encuesta de Morbilidad Hospitalaria al crear una base censal del 100% de las altas y al desagregar la información a nivel del ámbito hospitalario en lugar del provincial. Mejora la información recogida de los Diagnósticos y procedimientos utilizados en el hospital.
- Permite disponer de información uniforme y comparable entre diferentes hospitales, administraciones y países.

Los componentes del CMBD son 14:

1. Identificación del hospital.
2. Identificación del paciente.
3. Fecha de nacimiento.
4. Sexo.
5. Residencia.
6. Financiación.
7. Fecha de ingreso.
8. Circunstancias del ingreso.
9. Diagnóstico principal y secundario.

10. Procedimientos quirúrgicos y obstétricos,
11. Otros procedimientos.
12. Fecha de alta.
13. Circunstancias del alta.
14. Identificación del médico responsable del alta.

Una de utilidades con mayor importancia del CMBD es para obtener los Grupos Relacionados por el Diagnóstico (GRD). Los GRD constituyen un sistema de clasificación de pacientes que permite relacionar los distintos tipos de pacientes tratados en un hospital (es decir, su casuística), con el coste que representa su asistencia. La finalidad de los GRD es relacionar la casuística de un hospital con el consumo de recursos, esto implica disponer de un sistema que determine el tipo de pacientes tratados y que relacione cada tipo de pacientes con los recursos que consumen. Cada paciente es único, pero los grupos de pacientes tienen atributos comunes demográficos, diagnósticos y terapéuticos que determinan su intensidad de consumo de recursos.

El diseño y desarrollo de los GRD comenzó a finales de los años sesenta en la Universidad de Yale. El motivo inicial por el cual se desarrollaron los GRD era la creación de una estructura adecuada para analizar la calidad de la asistencia médica y la utilización de los servicios en el entorno hospitalario.

La primera aplicación a gran escala de los GRD fue a finales de los años setenta en el Estado de Nueva Jersey. El Departamento de Sanidad del Estado de Nueva Jersey utilizó los GRD como base para un sistema de pago prospectivo en el cual se reembolsaba a los hospitales una cantidad fija específica para cada GRD y por cada paciente tratado.

En 1982 la *Tax Equity and Fiscal Responsability Act* modificó la sección 223 sobre los límites de reembolso de gastos hospitalarios de *Medicare* (programa de seguro de salud del gobierno de los Estados Unidos para personas mayores de 65 años y algunas personas menores de 65 años con ciertas patologías) para incluir una corrección según la casuística de los hospitales y basada en los GRD. En 1983 el Congreso americano modificó la *Social Security Act* (Ley de Seguridad Social) para dar cabida a un sistema nacional de pago prospectivo a los hospitales, basado en los GRD y para todos los pacientes de *Medicare*. La evolución de los GRD y su uso como unidad básica de pago en el sistema de financiación hospitalaria de *Medicare* es el reconocimiento del papel

fundamental que juega el *case mix* o la casuística de un hospital a la hora de determinar sus costes. En nuestro entorno los GRD se utilizan en cualquier valoración organizativa, asistencial o económica del área de hospitalización.

1.2. El diagnóstico y su normalización

Los conocimientos acerca de la salud y la enfermedad y su plasmación en una forma de lenguaje médico son el resultado de la interacción de una serie de elementos que esquematizamos en la figura 1.2 del siguiente modo:

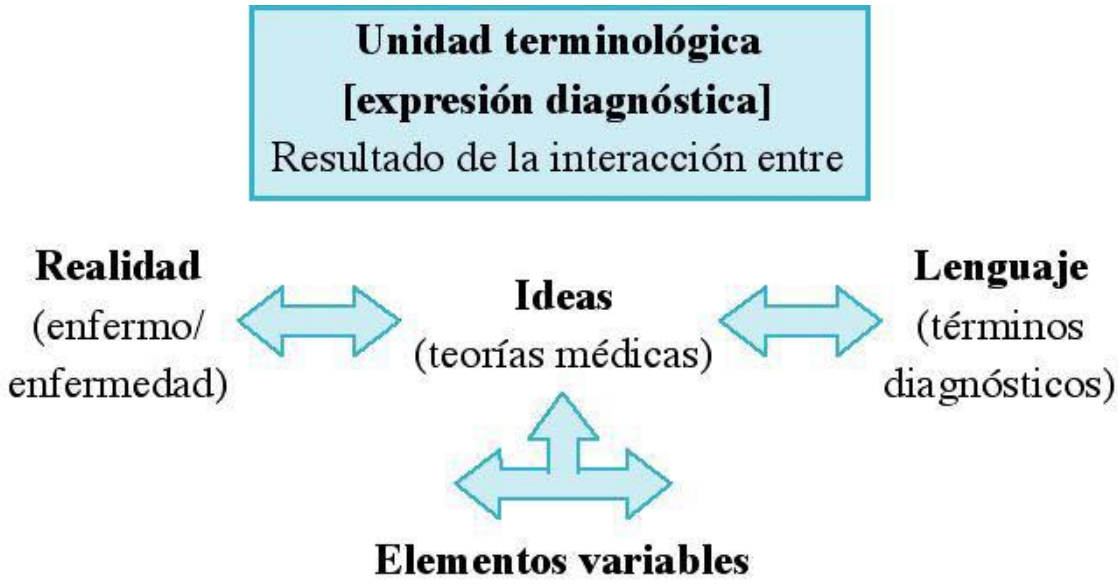

Figura 1.2: Esquema del diagnóstico y su normalización

Estos procesos generan la documentación clínica, un conjunto de documentos o cualquier otra cosa que pruebe y acredite los acontecimientos o datos relacionados con la salud del paciente y la asistencia prestada. La documentación clínica es testimonio y reflejo de la relación entre el médico y el paciente. Estamos hablando de un conjunto de documentos que contienen los datos, valoraciones e informaciones de cualquier índole sobre la situación y la evolución clínica de un paciente a lo largo de un proceso asistencial. Dentro de los distintos tipos de clasificación, en los procesos asistenciales con ingresos se realiza la asignación de códigos CIE-9-MC. Este trabajo lo realizan unas unidades específicas de codificación médica.

La utilidad de una clasificación es conseguir convertir los términos diagnósticos, terapéuticos y otros términos relacionados con la salud en códigos normalizados que permitan la explotación de la información que contienen. La calidad de la información

depende de las fuentes seleccionadas, del lenguaje documental, así como de los criterios utilizados para realizar el análisis documental. Por lo tanto es necesario disponer de fuentes de datos correctas, herramientas adecuadas para pasar de lenguaje natural a lenguaje documental y disponer de criterios claros para realizar el análisis, siendo el denominador común la existencia de profesionales formados que dominen estas herramientas.

1.3. Sistemas de terminología médica

Los sistemas de información sanitaria tienen como finalidad proporcionar datos que faciliten el conocimiento sobre la salud de las poblaciones y el funcionamiento de los sistemas sanitarios y constituyen un elemento clave para la comunicación de todos sus integrantes, ya sean éstos ciudadanos, pacientes, profesionales sanitarios, investigadores o políticos. Contar con datos comparables y normalizados redunda en una mayor exactitud, eficiencia, fiabilidad de la información sanitaria a nivel local, regional, nacional e internacional [Fenton, 2000].

Se entiende por normalización la formulación de especificaciones aceptadas (definiciones, normas, unidades y reglas) que establezcan un lenguaje común como base para la interpretación y el intercambio de información entre distintas partes. Con el volumen de documentación médica que producimos necesitamos representar esta información de manera estructurada y controlada. El modo más adecuado de plantearse una solución es hacer cumplir el uso de términos estándar. Las terminologías médicas proporcionan una manera estándar de nombrar los conceptos del dominio clínico. En la actualidad existen múltiples sistemas de terminología médica de los cuales destacamos los siguientes:

- CIE-9-MC. Clasificación Internacional de Enfermedades 9.ª revisión Modificación Clínica, de la Organización Mundial de la Salud (OMS). La CIE-9 está diseñada para clasificar los datos sobre morbilidad y mortalidad recogidos con fines estadísticos y para la clasificación de las historias clínicas por enfermedad y operaciones, con objeto de facilitar el almacenamiento y recuperación de dichos datos.
- CIE-10-MC. Clasificación Internacional de Enfermedades 10ª revisión Modificación Clínica. Es la evolución natural de la CIE-9-MC, en unos años esta sustituirá a la CIE-9-MC.

- SNOMED-CT.Son las siglas de *Systematized Nomeclature of Medicine Clinical Terms* una extensa terminología médica desarrollada por el *College of American Pathologists* (CAP) y mantenida por *The Internacional Health Terminology Standard Development Organisation* (IHTSDO). Facilita un lenguaje común para la indexación, el almacenamiento, la recuperación y la agregación de datos médicos. Está disponible en inglés, francés y español, la versión actual en español tiene más de 357.000 conceptos, 800.000 descriptores y 1.500.00 relaciones semánticas.
- CIAP-2. Clasificación Internacional de la Atención Primaria, realizada por la WONCA (Organización Mundial de los Médicos Generales/de Familia). La CIAP permite codificar tanto la razón de consulta (lo que dice el paciente) como el problema de salud (lo que dice el profesional sanitario) y el proceso de la atención (lo que se hace en el curso de la consulta); también permite la codificación de la gravedad y del estado funcional del paciente.
- CPT. Codificación de Procedimientos y Tratamientos médicos utilizada por *American Medical Association* (AMA) para procesos de facturación.
- ICNARC *Coding Method* (ICM). Método de codificación para unidades de cuidados Intensivos desarrollada en 1995 [ICNARC, 1995].
- *The NHS Clinical Terms*, llamada también *Read Classification* o *Read Codes*, tiene la finalidad práctica de llevar un seguimiento sistemático de grupos de pacientes con enfermedades crónicas o problemas de salud específicos. Cuenta con 100.000 términos y 150.000 sinónimos codificados y su implantación se desarrolla en el *Nacional Health Service* (NHS) de Inglaterra.
- GALEN, *General Architecture for Languages Encyclopoedias and Nomenclatures in Medicine*, es un proyecto financiado por la Unión Europea que tiene como objetivo el desarrollo de herramientas y métodos para una construcción y mantenimiento de clasificaciones de procedimientos quirúrgicos. Está elaborado con un modelo semántico para la gestión de terminología clínica denominado CORE (*Coding Reference*) que está estructurado en torno a tres módulos e incluye relaciones que reflejan posibles combinaciones de términos, y conceptos complejos que son combinaciones de conceptos más simples. La manipulación de estos conceptos y relaciones se

hace a través del lenguaje GRAIL (*GALEN Representation and Integration Language*) y el CM (*Concept Model*), una herramienta de modulación conceptual a partir de la cual los terminógrafos crean modelos con conceptos y relaciones que sirven, a su vez, para derivar otros nuevos siempre que GRAIL determine que es una composición válida. Se consiguen por tanto modelos conceptuales robustos, consistentes y predecibles que están separados del denominado Módulo Multilingüe que contiene las frases y términos utilizados para referirse al primero. Además en la estructuración conceptual, GALEN también tiene en cuenta los sistemas de clasificación vigentes en Salud y, entre otras cosas, relacionan los conceptos de estos sistemas con los conceptos estructurados en el modelo CORE que actúan a modo de interlingua. Este proceso tiene lugar en un tercer módulo, el Módulo de Conversión (*Code Conversion Module*).

- MeSH son las siglas de *Medical Subject Headings*, un tesauros desarrollado por la *National Library of Medicine* (NML) de los Estados Unidos, es un vocabulario controlado de términos biomédicos formado por un conjunto de términos denominados descriptores, que constituyen una estructura jerárquica que permite la búsqueda en distintos niveles de especificidad. Los descriptores se organizan de dos formas distintas: alfabéticamente y en una estructura jerárquica. MeSH, cuenta en el 2010 con 25.588 descriptores, 172.000 conceptos suplementarios (*Supplementary Concept Records*) y más de 97.000 términos de ayuda para localizar el descriptor más conveniente. MeSH es utilizado por la NLM para la indización de los artículos de las 5400 principales revistas biomédicas del mundo para la base de datos MEDLINE, y para la catalogación de libros, documentos y audiovisuales.
- UMLS (*Unified Medical Language System*) desarrollado por la *National Library of Medicine* (NLM) de los Estados Unidos. El sistema posee referencias cruzadas entre más de treinta vocabularios y clasificaciones, incluyendo CIE-9, SNOMED-CT y MESH. Este presenta tres fuentes de conocimiento:
    - ✓ El Meta Tesauro es una base de datos multilingüe que contiene información sobre conceptos médicos, incluyendo sus nombres y relaciones. Está construido a partir de otros tesauros, de clasificaciones

y listas de términos controlados utilizados en el cuidado de los pacientes, y en el indexado y catalogación de la literatura médica y la investigación clínica.

- ✓ El Léxico Especializado contiene en la versión actual 108.000 informes léxicos y más de 186.000 cadenas de términos. Está en inglés y cada entrada presenta información sintáctica, morfológica y ortográfica, incluyendo la categoría sintáctica (verbo, sustantivo, adjetivo, pronombre,…), las inflexiones de género y número, las conjugaciones de verbos, y los comparativos y superlativos de los adjetivos y adverbios.
- ✓ La Red semántica tiene 132 tipos semánticos y garantiza una categorización estable de todos los conceptos representados en el meta tesauro.

- Codificación de procedimientos radiológicos: E-ACR, Radlex y el Catálogo de Exploraciones Radiológicas de la Sociedad Española de Radiología Médica (SERAM).

Los sistemas de terminología médica más importantes dentro del ámbito hospitalario del SNS de España (y en la mayoría de los países) que se utilizan en el trabajo diario y requieren de un proceso de clasificación son: CIE-9-MC, SNOMED-CT y CIAP-2. Con el CIE-9-MC se codifican los episodios hospitalarios con ingreso, con SNOMED-CT se recopilan los diagnósticos anatomopatológicos de los informes del Servicio de Anatomía Patológica y con el CIAP-2, o en algunos casos CIE-9-MC, se realiza la clasificación de los procesos de Atención Primaria.

En la actualidad el CIE-9-MC es el que mayor importancia tiene en los sistemas de información sanitaria. Pero en un futuro próximo puede ser que SNOMED-CT alcance un mayor protagonismo. Las razones de esta afirmación están apoyadas en la adopción por parte de la Unión Europea de elegir SNOMED-CT como sistema de terminología médica e interoperabilidad semántica para la Historia Clínica Electrónica (HCE). SNOMED-CT es la terminología clínica de referencia seleccionada para la Historia Clínica Digital del Sistema Nacional de Salud (HCDSNS), lo que supone un primer

paso hacia la interoperabilidad semántica de la HCDSNS. SNOMED-CT tiene además mapeo con CIE-9-MC y CIE-10-MC.

1.4. CIE-9-MC Clasificación Internacional de Enfermedades

La Clasificación Internacional de Enfermedades es un inventario donde se recogen todos los posibles diagnósticos médicos y su uso es universal. Es el sistema más importante de codificación y clasificación que permite las comparaciones internacionales y la monitorización de los problemas de salud en todos los ámbitos de actuación asistenciales.

En el ámbito concreto de los sistemas de información hospitalaria, y en especial el entorno relacionado con la actividad clínico-asistencial, como es el caso del Conjunto Mínimo Básico de Datos (CMBD), la clasificación que se adapta mejor a la práctica clínica es la denominada modificación clínica de la novena versión de esta clasificación (CIE-9-MC). Publicada y distribuida por el *Council on Clinical Classifications* de Estados Unidos, incorpora volúmenes específicos de códigos para procedimientos, morfología de las neoplasias, así como de causas externas de lesiones y envenenamientos, por lo que permite reflejar de una manera más completa y fidedigna lo acontecido en el proceso de atención de los episodios clínicos.

La CIE-9-MC es la clasificación que se lleva utilizando en España desde hace ya dos décadas para la codificación clínica de los procesos de ingresos atendidos en los hospitales. Esto ha permitido la normalización del registro de altas CMBD del Sistema Nacional de Salud, de forma que en el momento actual disponemos de unas bases de datos del CMBD consistentes y estables, siendo la principal fuente de datos sobre morbilidad atendida en España. Además, esta información es la base para la mayoría de sistemas de clasificación de pacientes (Grupos Relacionados por el Diagnóstico - GRD), lo que permite a su vez acercarnos a la medición de la producción hospitalaria. Por último, el uso del CMBD se ha acuñado como fuente para la obtención de importantes indicadores relacionados con la calidad asistencial y seguridad del paciente.

1.4.1. Perspectiva histórica de la CIE-9-MC

La Clasificación Internacional de Enfermedades, Revisión 9.ª, Modificación Clínica (CIE-9-CM), se basa en la versión oficial de la 9.ª Revisión de Clasificación Internacional de Enfermedades (CIE-9) de la Organización Mundial de la Salud. CIE-9

está diseñada para clasificar los datos sobre morbilidad y mortalidad recogidos con fines estadísticos y para la clasificación de las historias clínicas por enfermedad y operaciones, con objeto de facilitar el almacenamiento y recuperación de dichos datos.

La idea de ampliar la Clasificación Internacional de Enfermedades para su uso en los archivos de hospitales se desarrolló originariamente como respuesta a la necesidad de una base más eficaz para el almacenamiento y recuperación de datos diagnósticos. En 1950, el Servicio de Salud Pública de Estados Unidos y la Administración de Veteranos de Guerra iniciaron pruebas independientes con la Clasificación Internacional de Enfermedades dirigidas a la clasificación de los archivos hospitalarios. El año siguiente, el *Columbia Presyterian Medical Center* en Nueva York adoptó la Clasificación Internacional de Enfermedades, 6.ª Revisión, con algunas modificaciones para uso en su departamento de archivos médicos. Unos años más tarde, la Comisión sobre Actividades Profesionales y Hospitalarias adoptó la Clasificación Internacional de Enfermedades con modificaciones similares para su uso en aquellos hospitales que participaban en el Estudio de Actividad Profesional.

El problema de la adaptación de la CIE para clasificar las historias clínicas hospitalarias, fue abordado por el Comité Nacional de Estados Unidos sobre Estadísticas Vitales y de Salud, a través del subcomité sobre Estadísticas Hospitalarias. Este subcomité revisó las modificaciones realizadas por los distintos usuarios de la CIE y propuso la realización de cambios normalizados. Dicha tarea fue llevada a cabo por un pequeño grupo de trabajo.

A la vista del creciente interés en el uso de la Clasificación Internacional de Enfermedades para la ordenación de datos hospitalarios, se emprendió en 1956 un estudio por la Asociación de Hospitales Americanos y la Asociación de Archivos Médicos de Estados Unidos (en aquel entonces llamada Asociación Americana de Bibliotecarios de Historias Clínicas), sobre la eficiencia relativa a los distintos sistemas de codificación para la clasificación diagnóstica. Este estudio indicó que la Clasificación Internacional de Enfermedades proporcionaba un marco adecuado y eficiente para la clasificación de historias clínicas. Los principales usuarios de la Clasificación Internacional de Enfermedades en hospitales consolidaron entonces sus experiencias y en diciembre de 1959 se publicó la primera adaptación. En 1962 se publica una revisión, incluyendo en esta ocasión la primera "Clasificación de Operaciones y Tratamientos".

En 1966, la Conferencia Internacional para la revisión de la Clasificación Internacional de Enfermedades resaltó que la 8.ª Revisión de la CIE había sido realizada pensando en las clasificaciones para hospitales y consideró que la clasificación revisada sería idónea para su uso en los hospitales de ciertos países. Sin embargo, se reconoció que la clasificación básica posiblemente proporcionaría un detalle inadecuado para la clasificación diagnóstica en otros países. Se pidió a un grupo de asesores que estudiara la revisión 8.ª de la CIE (CIE-8) para su aplicabilidad a los distintos usuarios en Estados Unidos. Dicho grupo recomendó que se proporcionaran más detalles para la codificación de datos hospitalarios y de morbilidad. Se pidió a la Asociación de Hospitales Americanos que desarrollara las propuestas de adaptación que fuesen necesarias. Esa tarea fue llevada a cabo por un comité asesor (el Comité Asesor ante la Oficina Central sobre la ICDA). En 1968, el Servicio de Salud Pública de Estados Unidos publicó la Octava Revisión de la Clasificación Internacional de Enfermedades, adaptada para su uso en Estados Unidos (publicación PHS, 1963). Este documento se llegó a conocer comúnmente como CDA-8 y a partir de 1968 sirvió como base para la codificación de los datos diagnósticos, tanto de morbilidad como de mortalidad en Estados Unidos.

1.4.2. Otras Adaptaciones

En 1968, la Comisión sobre Actividades Profesionales y Hospitalarias (CPHA) de Ann Arbor, Michigan, publicó la Adaptación Hospitalaria de la ICDA (H-ICDA) basada tanto en el documento original de la CIE-8 como en la ICDA-8. En 1973, la CPHA publicó una revisión de la H-ICDA-2. Los hospitales en las distintas partes de Estados Unidos se han mostrado divididos en la utilización de dichas clasificaciones. Con su entrada en vigor, en enero de 1979, la CIE-9-MC proporcionaba una única clasificación para su utilización en Estados Unidos, sustituyendo las clasificaciones anteriores, interrelacionadas, pero algo diferentes.

1.4.3. Antecedentes de CIE-9-MC

En febrero de 1977, un Comité Directivo fue convocado por el Centro Nacional de Estadística Sanitaria para proporcionar asesoramiento y consejo para el desarrollo de una modificación clínica de la CIE-9. Las organizaciones representadas en dicho Comité Directivo fueron:

- *American Association of Health Data Systems*

- *American Hospital Association*
- *American Medical Record Association*
- *Association for Health Records*
- *Council on Clinical Classifications*
- WHO *Center for Classification of Diseases for North America, sponsored by the National Center for Health Statistics,* DHEW

El Consejo sobre las Clasificaciones Clínicas está patrocinado por:

- *American Academy of Pediatrics*
- *American College of Obstetricians and Gynecologists*
- *American College of Physicians*
- *American College of Surgeons*
- *American Psychiatric Association*
- *Commission on Professional and Hospital Activities*

El Comité Directivo se reunió a intervalos periódicos a lo largo de 1977. La orientación clínica y las aportaciones técnicas fueron proporcionadas por los Grupos de Trabajo sobre la Clasificación establecidos por las organizaciones patrocinadoras del Consejo de la Clasificación Clínica.

La CIE-9-MC es una modificación clínica de la Clasificación Internacional de Enfermedades, 9ª Revisión (CIE-9) de la Organización Mundial de la Salud. El término "clínico" se utiliza para subrayar el propósito de la modificación: el de servir tanto como herramienta útil en el campo de las clasificaciones de los datos de morbilidad para la ordenación de las historias clínicas, las revisiones de los cuidados médicos y los programas de cuidados ambulatorios y otros cuidados médicos, como para las estadísticas de salud básicas. Para describir el cuadro clínico del paciente, los códigos deben ser más precisos que aquellos que se necesitan exclusivamente para agrupaciones estadísticas y análisis de tendencias.

La actual edición en castellano de la CIE-9-MC (7ª Edición. Enero 2010) se presentan en un solo libro agrupando:

TOMO I o Índice Alfabético de Enfermedades, que contiene:

- Índice de enfermedades, en el que a su vez se incluyen:
  - Tabla de hipertensión
  - Tabla de neoplasias

- Tabla de fármacos y químicos
- Índice Alfabético de causas externas de lesiones y envenenamientos (Códigos E)

TOMO II o Lista Tabular de Enfermedades incluyendo:

- Lista Tabular de Enfermedades. Dividida en 17 capítulos que comprenden las categorías desde 001 hasta 999
- Clasificación suplementaria de factores que influyen en el estado de salud y contacto con los servicios sanitarios que incluyen las categorías desde V01 hasta V89
- Clasificación suplementaria de causas externas de lesiones y envenenamientos que incluyen las categorías desde E800 hasta E999

TOMO III o Índice Alfabético de Procedimientos

TOMO IV o Lista Tabular de Procedimientos, dividiéndose en 16 capítulos que comprenden las categorías desde 00 hasta 99.

TOMO V de Apéndices, que se encuentra conformado por los siguientes anexos:

- Apéndice A. Morfología de las neoplasias que incluyen los códigos del M8000/X al M9970/X
- Apéndice B. Subdivisiones de cuarto dígito para el código de Causas Externas (Código E)

1.4.4. Estructura de los códigos CIE-9-MC

CIE-9-MC es un sistema de categorías numéricas que han sido asignadas a las enfermedades de acuerdo con unos criterios que han sido establecidos previamente. Una clasificación de enfermedades debe reunir una serie de condiciones como son:

- Debe tener un número restringido de categorías y para ello todas las enfermedades se agrupan en categorías, lo que implica un beneficio para la clasificación.
- Cualquier enfermedad solo se puede clasificar dentro de una categoría ya que las categorías entre si son excluyentes y una categoría excluye todas las demás.

La CIE-9-MC cumple todos estos requisitos y aunque no es perfecta tiene la ventaja de que está siendo utilizada en los hospitales de todo el mundo para codificar las altas y se actualiza anualmente.

La CIE-9-MC se estructura siguiendo un criterio principalmente anatómico en capítulos, 17 para enfermedades y 16 para procedimientos. Cada uno de los capítulos a su vez se divide en secciones (sólo en la Lista Tabular de Enfermedades), cada sección se divide siempre en categorías, cada categoría puede dividirse en subcategorías y cada subcategoría puede hacerlo en subclasificaciones. En la tabla 1.1 representamos esta clasificación para enfermedades y procedimientos.

Enfermedades:

Capítulo
Sección
- Categoría
- Subcategoría
- Subclasificación

Procedimientos:

Capítulo
- Categoría
- Subcategoría
- Subclasificación

Tabla 1.1: Estructura de enfermedades y procedimientos CIE-9-MC

Existen códigos válidos desde categoría en adelante y a cada subdivisión corresponde un nuevo dígito en el código. Un punto decimal separa las categorías de subcategorías y subclasificaciones. En la tabla 1.2 definimos los tipos de códigos posibles con su posible nomenclatura.

TIPO DE CÓDIGOS	CATEGORÍA	SUBCATEGORÍA	SUBCLASIFICACIÓN
Enfermedades (000-999)	XXX	XXX.X	XXX.XX
Clasificación Suplementaria (V01-V86)	VXX	VXX.X	VXX.XX
Causas Externas de Lesiones y Envenenamientos (E800-E999)	EXXX	EXXX.X	—
Procedimientos (00-99)	XX	XX.X	XX.XX
Morfología de las Neoplasias (M8000-M9970)	M8000/X-M9970/X		

Tabla 1.2: Tipos de códigos CIE-9-MC

Para asignar un código a un diagnóstico o procedimiento se elegirá siempre el código de mayor nivel de especificidad posible (mayor número de dígitos).

1.4.5. Evolución del CIE-9-MC

En el momento actual, la CIE-9-MC tiene capítulos que no permiten la adicción de más códigos, agotamiento de códigos, y cada vez se hace más difícil su actualización con la tecnología. Presenta además dificultad para la codificación en algunas áreas y de los procedimientos de enfermería. Partiendo de esta necesidad de evolucionar surge la CIE-10-MC, cuyas diferencia más importante con respecto a CIE-9-MC figuran en la tabla 1.3.

CIE-9-MC	CIE-10-MC
Numérica	Alfanumérica – Incluye todas la letras excepto la "U"
17 capítulos	21 capítulos
V y E códigos adicionales	Códigos V y E en clasificación general
Códigos entre 3 y 6 dígitos	Códigos entre 3 y 7 dígitos

Tabla 1.3: Diferencias entre CIE-9-MC y CIE-10-MC

La implantación del CIE-10-MC no es una tarea fácil, ni puede acometerse a corto plazo. Esta tiene que estar sustentada en una decisión política de todo el SNS, acompañada de una estrategia de implantación, una estrategia de formación, con las herramientas necesarias para su desarrollo. Esto implica un periodo de estudio para realizar la estrategia del cambio que requiere la implantación de la CIE-10-MC que incluya el desarrollo de herramientas adecuadas para la formación en el uso de la clasificación, de ayuda a la codificación, sin olvidarse del cronograma para formación del personal de codificación. Como ejemplo, Estados Unidos tiene previsto adoptar CIE-10-MC en el año 2014.

1.5. Descripción de la codificación

Podemos definir la codificación con la Clasificación Internacional de Enfermedades 9ª Revisión – Modificación Clínica (CIE-9-MC) como un proceso de análisis documental (indización y codificación) mediante el que, tras analizar la información contenida en uno o varios documentos en lenguaje natural (en nuestro caso el informe de alta, no la historia clínica), se selecciona la información relevante

para traducirla a un lenguaje normalizado (CIE-9-MC). Esta traducción debe reflejar de la manera más fiel posible lo acontecido en el episodio a codificar.

Al tratarse de un proceso de selección y traducción, la codificación será tanto más completa cuanto:

- Mayor sea el conocimiento de ambos lenguajes:
  - Lenguaje documental y normalizado: CIE-9-MC
  - Lenguaje natural: Terminología médica
- Mejor sea la capacidad del codificador para sintetizar la totalidad de la información relevante de la fuente documental.

Para codificar correctamente un episodio es necesario seguir una secuencia de procesos que, salvo algunas excepciones ya falta de normas o instrucciones particulares que indiquen otro proceder, consisten en los pasos que se explican a continuación.

1.5.1. Indización del episodio asistencial

1.5.1.1. Identificar los diagnósticos y procedimientos que deben ser codificados

Para ello, hay que localizar y leer toda la información correspondiente al episodio. Una buena aproximación la dará el informe de alta. Sin embargo, éste no será suficiente en muchos casos, ya que a veces se hace imprescindible conocer la circunstancia del ingreso, la circunstancia del alta, una descripción de las técnicas quirúrgicas, etc.

Por ello, es conveniente revisar además otros documentos:

- Informe e asistencia en urgencias
- Hoja de anamnesis y exploración clínica
- Hoja operatoria
- Informes: anatomía patológica, radiología, interconsulta, y otros
- Hoja de curso clínico
- Hojas de evolución de enfermería

En cuanto a los procedimientos habrá que confirmar que se han realizado durante el episodio, e incluir también los procedimientos que hayan podido realizarse fuera del centro durante el ingreso siempre que cumplan los requisitos establecidos.

1.5.1.2. Identificar los diagnósticos y procedimientos principales y secundarios

El paso fundamental para una correcta interpretación del episodio es distinguir los conceptos de diagnóstico y procedimiento principal y los secundarios, que se definen como:

1.5.1.2.1. Diagnóstico Principal [DP]

- En el ámbito de la hospitalización, se define diagnóstico principal como "el proceso que, tras el estudio pertinente y al alta hospitalaria, se considera el responsable del ingreso del usuario en el hospital" (Orden 6 septiembre 1984, del Ministerio de Sanidad y Consumo). Este es el criterio que vamos a utilizar en nuestro trabajo, ya que los episodios que pretendemos codificar automáticamente son de pacientes hospitalizados.

- En el ámbito del hospital de día quirúrgico (HDQ), se define diagnóstico principal como "el proceso, que al alta hospitalaria, se considera responsable del procedimiento o del grupo de procedimientos relacionados que se han realizado al paciente en este ámbito".

- En el ámbito del hospital de día quirúrgico, en el caso de que al paciente se le realicen procedimientos no relacionados entre si (ej.: herniorrafia inguinal y reparación de fimosis) quedará al arbitrio del documentalista identificar el diagnóstico principal, e introducirá el resto en campos de diagnósticos secundarios.

- En el ámbito del hospital de día médico (HDM), se define diagnóstico principal como "el proceso, que al alta hospitalaria, se considera responsable del procedimiento o grupo de procedimientos relacionados que se han realizado al paciente en este ámbito".

1.5.1.2.2. Diagnósticos Secundarios [DS]

"Se consideran diagnósticos secundarios a los procesos patológicos que no son el principal y que coexisten con él en el momento del ingreso o contacto,

que se desarrollan a lo largo de éste, o que influyen en su duración o en el tratamiento administrado. Deben excluirse los diagnósticos relacionados con un episodio anterior y que no tengan que ver con el que ha ocasionado el actual ingreso o contacto "(Orden 6 septiembre 1984, del Ministerio de Sanidad y Consumo).

En general se tratará de los diagnósticos y comorbilidades presentes al ingresar, así como las complicaciones que hayan podido aparecer en el transcurso del mismo.

Para propósitos de información la definición de "diagnósticos secundarios" es interpretada como las afecciones adicionales que influyan en el cuidado del paciente cuando impliquen algunas de las siguientes condiciones:

- Evaluación clínica
- Tratamiento terapéutico
- Procedimientos diagnósticos
- Alargamiento de la estancia hospitalaria
- Cuidados o monitorización de enfermería

Los diagnósticos descritos en otros episodios de ingresos anteriores y resueltos no se codificarán

1.5.1.2.3. Procedimiento Principal [PP]

En el CMBD de hospitalización esta variable registrará el primer procedimiento quirúrgico programado (incluidos los diferidos) que acontezca en el tiempo, que esté relacionado con el diagnóstico principal y que haya ocupado un quirófano. Se incluirán las cesáreas programadas. En el CMBD de hospitalización contendrá un código comprendido entre el 00.01 y el 86.99.

Si un procedimiento no cumple estas condiciones ocupará una de las restantes posiciones de procedimientos quirúrgicos/diagnósticos/ terapéuticos (P2 a P15), pero nunca se registrará en este campo (P1); tal es el caso de los procedimientos quirúrgicos realizados de forma urgente.

En los CMBD de hospital de día quirúrgico y hospital de día médico esta variable recogerá tanto los procedimientos quirúrgicos como los obstétricos,

terapéuticos o diagnósticos.

1.5.1.2.4. Procedimientos Secundarios [PS]

Se incluirán tanto los procedimientos quirúrgicos como los obstétricos, terapéuticos o diagnósticos.

Cualquier de los conceptos definidos anteriormente se asignará el código de mayor nivel de especificidad.

Después de detallar los conceptos y procedimientos del proceso de codificación manual que realizan los codificadores médicos, vamos a definir el entorno que utilizaremos en la codificación automática.

1.6. El Informe de Alta Hospitalaria en la Codificación CIE-9-MC

El Informe de Alta Hospitalaria (IAH) es el documento final emitido por un médico responsable acerca de la atención a un paciente, que hace referencia a un episodio de hospitalización. El IAH refleja un resumen del historial clínico, de la actividad asistencial prestada, el diagnóstico principal y los secundarios, así como el tratamiento recomendado. Es un documento fundamental de la asistencia sanitaria porque facilita la continuidad asistencial, reduce el tiempo de búsqueda de información, evita la repetición de pruebas y disminuye los errores. Pero para nosotros su principal característica es su estructura, lo que nos va a permitir realizar la codificación de CIE-9-MC. El IAH debe tener unos contenidos mínimos que han sido definidos en un real decreto, en el que han participado el Ministerio de Sanidad, un grupo de expertos de diferentes sociedades médicas y de enfermería, así como el Consejo Interterritorial del Sistema Nacional de Salud. Los contenidos básicos del IAH incluyen:

- ✓ Los datos administrativos del paciente
- ✓ Los referidos a sus antecedentes personales
- ✓ La enfermedad actual y la situación previa
- ✓ Las pruebas y procedimientos
- ✓ El juicio clínico
- ✓ El tratamiento y otras recomendaciones

El IAH debe ser un resumen sintético, preciso y conciso y ha de redactarse en términos médicos. El médico especialista hospitalario tiene que expresar con rigor y exhaustividad los fundamentos del diagnóstico y tratamiento, y una información

amplia acerca de los datos complementarios y las pruebas realizadas. Para el codificador médico su interés primordial son los diagnósticos y procedimientos, principales y secundarios, así como las complicaciones durante la estancia hospitalaria. Todo episodio de hospitalización debe concluir con un informe de alta independientemente del destino del enfermo. El informe de alta tiene especial importancia por dos aspectos, uno clínico y otro de gestión. El obvio para los médicos es el interés clínico, pues informa al individuo de lo acontecido durante la hospitalización y le garantiza la continuidad asistencial al mantener informados a los sucesivos médicos que lean ese informe, pero también tiene un interés en la evaluación de la gestión de la estancia, para el CMBD y para obtener los GRD.

La calidad del informe en relación con los dos aspectos no está reñida, sino todo lo contrario. Cuanto mejor sea el informe desde el punto de vista clínico también lo será desde el punto de vista de la gestión.

La forma más habitual de codificación de altas es a partir de los IAH, dada la complejidad y la carga de trabajo que implica la codificación desde la historia clínica completa. En la mayoría de los centros hospitalarios del SNS se codifica a través del IAH. El uso exclusivo del informe de alta para codificar un episodio tiene una gran ventaja para el codificador médico ya que agiliza el proceso, si bien tiene un inconveniente, si no está bien cumplimentado perdemos información y por lo tanto peso en la codificación. En nuestro caso, este condicionante es evidente, ya que es el documento disponible electrónicamente y relacionado inequívocamente con el ingreso hospitalario.

Nuestra única fuente de conocimiento documental disponible para realizar la codificación automática de códigos (CAC) CIE-9-CM es el IAH. Por esto creemos necesario analizar y estudiar la influencia del IAH en la codificación, y en concreto la sección del IAH en donde el médico especialista realiza el juicio clínico, enumerando los diagnósticos del episodio de hospitalización.

La primera cuestión que se plantea es si hay alguna reciprocidad entre la sección de diagnósticos que escribe el médico especialista en el IAH y la codificación que realizada de este informe el codificador médico. Lo lógico sería que los diagnósticos que redacta el médico especialista tenga una trascripción a leguaje normalizado como el CIE-9-MC. Para ello realizamos un estudio orientado a analizar la diferencia entre el número de diagnósticos que tiene un IAH y el número de códigos que asigna el codificador. Los datos que mostramos a continuación corresponden a la colección de

informes de alta del servicio de Medicina Interna del Hospital de Conxo dentro del Complejo Hospitalario Universitario de Santiago de los años 2003, 2004 y 2005 (hasta mayo).

La figura 3.1 resultante tiene las siguientes coordenadas: en el eje de las abscisas**,** la diferencia entre el número de códigos asignados a un informe por el codificador y las líneas de diagnósticos que escribe el médico especialista en el IAH, en el eje de las ordenadas, el número de documentos con esa diferencia. La figura 3.1 nos muestra que el área bajo la curva que está en lado positivo del eje x es mayor que la situada en el lado negativo. Esto indica que los codificadores incluyen más códigos de diagnóstico que líneas de diagnóstico redacta el médico especialista. Una posible explicación podría ser a que los codificadores se apoyan en otros documentos diferentes al documento de alta. Pero la experiencia práctica que pudimos contrastar en una unidad de codificación no lo certifica. Esta tendencia es mínima ya que el punto máximo de esta función está en un valor de +1.Si nos fijamos los gráficos tiene cierta simetría, lo único que su centro está ligeramente desplazado hacia la derecha (más códigos que líneas de diagnósticos).

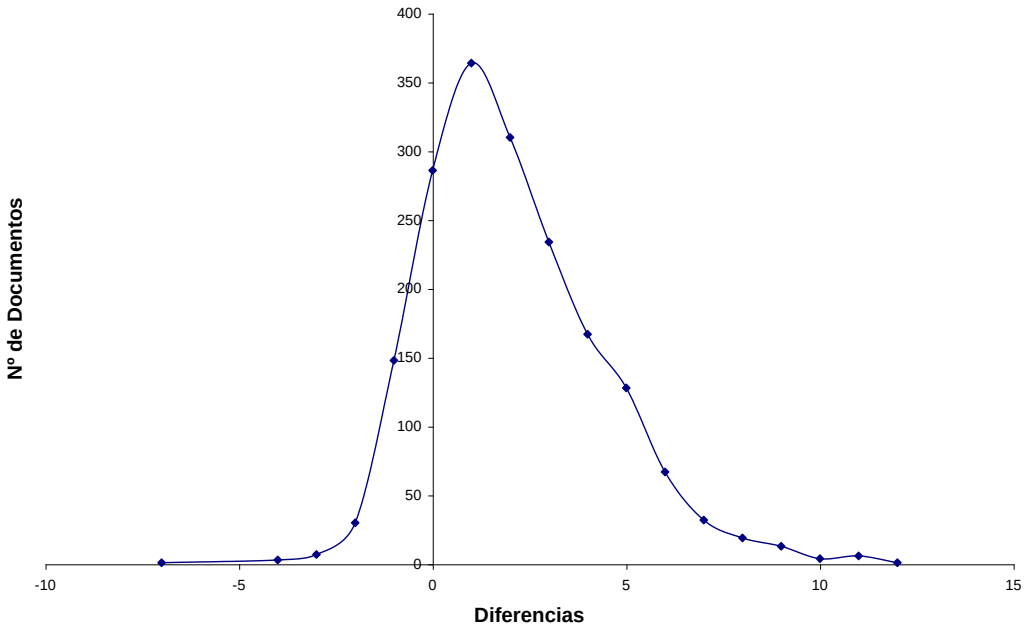

Figura 1.3: Diferencias entre el nº de diagnósticos codificados y el nº líneas de diagnósticos en los documentos

Las razones de este comportamiento creemos que están asentadas en varias circunstancias. En primer lugar, tendríamos la propia condición humana del médico codificador que realiza un trabajo personalizado y no simplemente mecánico. Estamos

ante los IAH con mayor complejidad clínica del ámbito hospitalario, esto puede influir en este incremento, patologías más complejas, en donde el clínico no valora en sus líneas de diagnósticos algún diagnóstico de menor importancia. No hemos realizado un estudio exhaustivo de los episodios con diferencias significativas, pero podemos aventurar que los motivos pueden ser varios entre los que destacamos situaciones como la que surge cuando el clínico resume en la frase "Los previos" los diagnósticos previos del paciente y no los redacta en la sección de diagnósticos. O cuando el paciente ha estado ingresado en otros servicios y el codificador asigna todos los códigos al servicio donde se realiza el alta. También cuando existe una agrupación de diagnósticos en una misma línea, etc. En esta línea podría ser útil realizar un estudio de la eficacia de clasificador automático en función del especialista médico que redacta el informe de alta.

En cualquier caso lo importante de este análisis es que el informe de alta tiene la mayoría de la información que necesitamos para poder utilizarlos como elemento único y principal en la creación de un codificador automático.

1.7. Sistemas de ayuda a la codificación

Desde los inicios de la codificación se está intentando mejorar su productividad [Zieserl and Dowell 1989] en los sistemas de codificación. Este tipo de mejoras siempre surgieron de la mano de las nuevas tecnologías. Al principio, se utilizaban programas de ayuda a la codificación, denominados encoders. Estos básicamente se dividían en dos tipos. El primero de ellos manejaba un sistema de subdivisiones lógicas. Se introduce el término principal del diagnóstico o procedimiento y el sistema mediante una serie de preguntas y a partir de las respuestas obtenidas acaba sugiriendo la asignación de un código [Surjan and Heja 2001]. El segundo tipo de programas se considera más una codificación asistida por ordenador, una especie de libro de ayuda a la codificación, en donde mediante el ordenador se consulta el índice alfabético y la lista tabular del CIE-9-MC. El personal con mayor experiencia encargado de la codificación, prefiere normalmente el segundo tipo de programas debido a que tienen mayor pericia en el manejo de las entradas alfabéticas y en la utilización de la lista tabular. Por tanto, codifican más rápidamente con este sistema. Al contrario, los codificadores con poca experiencia prefieren el primer sistema que les va guiando hasta realizar la selección de un código.

En la última década hemos visto un torrente de adelantos en tecnologías de la información para automatizar y agilizar diferentes procesos relacionados con la salud. Unos de estos procesos pendientes de una solución óptima es la CAC CIE-9-MC. Obtener procesos automáticos para una correcta codificación de diagnósticos es importante ya que nos proporciona información muy útil en muchas decisiones de los sistemas de salud, incluyendo la práctica clínica, la investigación, la gestión, decisiones políticas sobre sanidad, etc.

En los últimos años están surgiendo nuevas técnicas para abordar este problema [Surjan, 1999], y estos entornos empiezan a ser considerados atractivos y un difícil reto para los investigadores.

Las investigaciones y los desarrollos realizados para la codificación CIE-9-MC mediante tecnologías de la información las podemos clasificar en dos grupos. La codificación asistida, un sistema de software que asiste al usuario en la asignación del código. Y la clasificación automática, en donde obtenemos los códigos CIE-9-MC propuestos o asignados automáticamente sin una intervención humana directa en el proceso.

Dentro de estos dos tipos de codificación podemos diferenciar varias técnicas, destacamos las más utilizadas:

- Codificación asistida
  - Software construido para la navegación en una taxonomía jerárquica. Permiten en cada nivel visualizar los niveles más bajos.
  - Herramientas basadas en una búsqueda léxica. Permiten buscar un término en el texto del diagnóstico. Son fáciles de desarrollar y poner en práctica, pero sus resultados siguen siendo limitados.
- Codificación Automática
  - Codificación a través de las Técnicas de Procesamiento de Lenguaje Natural (PLN)
  - Codificación automática o semiautomática mediante Técnicas de Aprendizaje Automático.
  - Soluciones hibridas que combinan Técnicas de PLN y de Aprendizaje Automático.

En la literatura no existen revisiones sistemáticas sobre los sistemas clínicos automatizados de codificación CIE-9-MC. Existe un estudio reciente [Stanfill et al., 2010] que realiza una revisión de la literatura científica para identificar todos los trabajos sobre la codificación automática y sistemas de clasificación clínicos. Dentro de los propósitos de los sistemas automáticos de codificación clínica para resolver problemas prácticos del mundo real, nos encontramos que en los entornos de CAC no son de los más estudiados.

La codificación automática y los sistemas de clasificación son una tecnología emergente, donde los investigadores construyen y evalúan nuevos proyectos. Es importante explorar el funcionamiento de los sistemas de clasificación automática para determinar su aplicabilidad en los procesos de codificación. Disponer de una codificación correcta se ha convertido en una de las tareas más críticas de la asistencia sanitaria, ya que las necesidades de la atención médica han evolucionado. Prueba de ello es el desafío internacional para la CAC CIE-9-MC propuesto a la comunidad de investigadores en Procesamiento de Lenguaje Natural (PLN) por *Computational Medicine Center*. En esta competición internacional del año 2007 (www.computationalmedicine.org/challenge) participaron 44 grupos de investigación, y se hizo una clasificación de los mejores equipos para las métricas definidas en el desafío. La mayoría de los trabajos de investigación en clasificación de documentación médica surgen de colaboraciones entre hospitales e investigadores universitarios, como es nuestro caso. Estas colaboraciones ofrecen un entorno realista y práctico en donde aplicar estas técnicas. Una de las aportaciones importantes del desafío fue proporcionar un corpus público para que los científicos puedan experimentar sus técnicas en CAC. Este corpus está formado por informes de radiología para realizar una codificación CIE-9-MC. El primer corpus construido para realizar experimentos en codificación CIE-9-MC es de 1972 [Dinwoodie 1972]. Se han construido otros corpus, como en nuestro caso, gracias a la colaboración con los centros hospitalarios, para poder experimentar. Los grupos de investigación trabajan con corpus diferentes entre sí lo que dificulta la valoración de los resultados. En este misma línea están las conclusiones aportadas por [Stanfill et al., 2010] donde los Sistemas de CAC en sí no son generalizables, y tampoco lo son los resultados de su evaluación, por las diferencias en cuanto a sistemas de codificación clínica y en cuanto a las características de los entornos. Dada la complejidad del proceso es necesario utilizar

unas métricas específicas para este tipo de clasificadores, esta tesis contribuye a este fin y ayuda al desarrollo de un entorno en donde los resultados tengan un significado real para los codificadores manuales. Esto tendría que venir acompañado de un corpus público (en varios idiomas) para que la comunidad de investigadores en clasificación automática de documentación clínica pudiese evaluar y comparar sus sistemas. Mientras tanto, nos situamos en un contexto en donde los distintos Servicios de Salud que forman el SNS de España están implantando de forma progresiva la HCE. Esto permite a los investigadores en clasificación automática, disponer en formato electrónico de la información clínica, base fundamental para crear colecciones propias en donde experimentar.

American Health Information Management Association (AHIMA) dispone de un grupo de trabajo, para explorar los sistemas de codificación. Una de sus conclusiones es que la codificación manual es un sistema caro e ineficiente y el sector necesita soluciones automatizadas para permitir que el proceso de codificación sea más productivo, eficiente, preciso y consistente. Estos atributos son aún más importantes en entornos como el sistema de salud americano en donde la codificación se utiliza como proceso de facturación de servicios sanitarios. Por ello, estos sistemas de codificación automático son un buen mecanismo de lucha contra el fraude [Garvin, Watzlaf, and Moeini 2006]. Sin olvidar que el software de estos sistemas tienen que cumplir una guía de buenas prácticas. En la actualidad, una de las premisas importantes de los sistemas automáticos de codificación es la supervisión y control por codificadores humanos.

En esta etapa de experimentación e investigación en la que nos encontramos, los codificadores humanos, como evaluadores cualificados, son un elemento imprescindible en la certificación del funcionamiento de los sistemas automáticos de codificación. El codificador humano no puede ver a los codificadores automáticos como un peligro para su situación laboral. Tampoco los investigadores en CAC pueden pretender encontrar un sistema que supere al codificador humano, ya que estos son procesos muy complicados. Tenemos que encontrar sistemas de permitan que la codificación sea más productiva, eficiente, precisa y consistente.

En concreto, la CAC debe aportar una situación ventajosa con relación a la codificación manual en alguno de los siguientes términos:

- Aumentar la productividad en la codificación

- Obtener una codificación más coherente
- Asignación de códigos más completos
- Aumentar la precisión en la codificación
- Facilitar las auditorías
- Disminuir los coste de codificación
- Facilitar la codificación y clasificación de la documentación clínica
- Posibilitar la asignación de códigos a personal que no sea codificador médico
- Gestión de recursos más precisa

Capítulo 2

Clasificación Automática de Textos

La clasificación es una motivación natural en el ser humano y de especial interés en la comunidad científica. La clasificación de textos tiene una cantidad importante de aplicaciones prácticas, como la desarrollada en esta tesis. En este capítulo presentamos una revisión de las técnicas que se han utilizado hasta la actualidad en el área de la clasificación automática de textos. Se analizarán los diferentes tipos de clasificación existentes, sus características y propiedades. Por otra parte, se expondrán las técnicas de clasificación más conocidas que han sido tratadas en la literatura científica. También se presentarán algunos modelos para la representación de documentos y los métodos de evaluación específicos para este tipo de clasificadores.

2.1. Definición de Clasificación de textos

En la actualidad, la mayoría de la información que se genera está disponible en formato electrónico, y últimamente su volumen se está incrementado. Por esta razón, cada vez es más necesario poder clasificarla y disponer de herramientas que nos ayuden a realizar este proceso.

La Clasificación Automática de Textos (CAT) está altamente relacionada con la Recuperación de Información. Hay autores que sitúan la clasificación de textos en la frontera entre el Aprendizaje Automático y la Recuperación de Información [Sebastiani, 2002], y hay quienes se refieren a este área de estudio como una parte de la Minería de Textos [Knight 1999]. En la comunidad científica el enfoque dominante para abordar este problema se basa en técnicas de aprendizaje automático. En concreto, a través de un proceso estadístico o inductivo que crea automáticamente un clasificador por el aprendizaje adquirido a partir de un conjunto de documentos. La

categorización de textos es un campo muy tradicional. Sebastiani [Sebastiani, 2002] ubica los primeros trabajos en el año 1961 con las investigaciones de Maron [Maron, 1961]. En la década de los 80, los clasificadores que existían eran construidos por expertos de forma manual mediante el uso de reglas. En los años 90 este enfoque pierde popularidad en favor del aprendizaje automático. Con la Recuperación de Información tiene amplia relación, de hecho, hay varias técnicas de RI que también son utilizadas en CAT. Estas técnicas son usadas en algunas de las fases del ciclo de vida del clasificador:

- Indexación, del mismo modo que se utiliza en Recuperación de Información, se trabaja con esta técnica en la fase de representación de los documentos de texto.
- Distintas técnicas de Recuperación de Información se usan en la construcción estadística o inductiva del clasificador.
- Evaluación al estilo Recuperación de Información, para medir la efectividad alcanzada por el clasificador.

En el aprendizaje supervisado, se sabe a qué clases pertenecen algunos documentos. Esto es, se dispone de una colección de documentos etiquetados, generando un clasificador. Una vez terminada esta fase, el clasificador que hemos construido se utilizará para la clasificación de documentos de los que no se conoce su clase.

En el aprendizaje no supervisado se extraen los patrones de clasificación sin disponer de una colección de documentos etiquetados. La clasificación se realiza en grupos no predefinidos, lo que se denomina *clustering*.

En el presente trabajo cuando utilizamos el término clasificación o categorización automática de textos estamos describiendo un modelo de aprendizaje supervisado.

El objetivo de este trabajo en clasificación automática de texto (CAT) es categorizar o clasificar documentos dentro de un número de clases predefinidas en función de su contenido. El clasificador va aprendiendo de manera estadística o inductiva a partir de ejemplos preclasificados. Para ello tenemos que decidir qué características seleccionamos de los textos y como las vamos a utilizar. Una ventaja, sin duda muy importante, es que es más fácil clasificar documentos mediante técnicas de Aprendizaje Automático que construir y afinar reglas de clasificación. Esto último tiene un alto coste en términos de construcción y mantenimiento.

El proceso de clasificación lo podemos definir de la siguiente forma. Dado un conjunto de documentos $d_j \epsilon\ D$ que pertenecen a un determinado dominio, y un conjunto fijo de clases $C = \{c_1, c_2, \dots, c_{|c|}\}$, CAT persigue aproximar una función objetivo desconocida $\emptyset : D\ x\ C \rightarrow \{T,\ F\}$ (que describe como los documentos deberían ser clasificados). Esta función ideal es aproximada por medio de una función $\widehat{\emptyset} : D\ x\ C \rightarrow \{T,\ F\}$ que llamamos clasificador. Si $\emptyset(d_j, c_i) = T$, entonces d_j es un ejemplo positivo o pertenece a la clase c_i, mientras que si $\emptyset(d_j, c_i) = F$ es un ejemplo negativo, es decir no pertenece a la clase c_i. La clasificación en sí misma es una tarea subjetiva, muchas veces, en un determinado dominio, los expertos no se ponen de acuerdo en la clasificación de un documento d_j en una clase c_i. Este factor lo debemos tener en cuenta a la hora de comparar los resultados de la clasificación automática y la realizada por un experto.

2.2. Tipos de Clasificación Automática de Textos

Dependiendo de los aspectos propios de la clasificación esta se puede catalogar en diferentes tipos, que se mencionan a continuación.

2.2.1. Única etiqueta *vs* multi-etiqueta

Dependiendo el tipo de dominio en el que estemos trabajando, un documento puede estar asociado a una o varias clases. Denominamos clasificación con etiqueta única cuando cada uno de los documentos de la colección tiene una y sólo una clase asignada, y clasificación multi-etiqueta cuando el número de clases asociadas a un documento puede variar desde 0 hasta el número total de clases. Cuando un documento pertenece a más de una clase, esto genera un mayor grado de complejidad en la clasificación.

Un caso especial de única etiqueta es la clasificación binaria, que se produce cuando únicamente tenemos dos posibles etiquetas a asignar a cada documento. Por ejemplo, la clasificación de emails como spam o no spam es un típico ejemplo de clasificación binaria donde únicamente existen dos etiquetas en el sistema y estas se corresponden con la pertenencia o no del documento a la clase de documentos basura. Un clasificador binario es más general que uno multi-etiqueta [Sebastiani, 2002] puesto que podemos transformar el problema multi-etiqueta en $|C|$ problemas de clasificación binarios independientes, cada uno de ellos asociado a la pertenencia o no del documento a cada una de las $|C|$ clases de referencia.

2.2.2. Clasificación pivotada por categorías *vs* clasificación pivotada por documentos

Otro aspecto relevante a la hora de estudiar una solución en CAT es el pivote. Se define categorización con pivote en documentos (DPC – *Document Pivoted Categorization*) como aquella CAT que pretende encontrar todas las clases a las que pertenece un documento.

Por otro lado, existe la categorización con pivote en categorías o clases (CPC – *Category Pivoted Categorization*) como aquella CAT que pretende encontrar todos los documentos que pertenecen a una determinada clase.

La diferencia es importante si el conjunto de clases o el de documentos, no están completamente disponibles desde el principio. También es importante para escoger el método de construcción del clasificador.

La clasificación con pivote en documentos (DPC) es la clasificación más común y se suele recomendar cuando los documentos están disponibles en distintos momentos, como los correos electrónicos.

La clasificación con pivote en clases (CPC), en cambio, suele ser apropiada cuando una nueva clase puede ser agregada después de que existan documentos ya clasificados, y los documentos necesitan ser reclasificados con $|C| + 1$ clases.

2.2.3. Clasificación 'hard' *vs* clasificación en ranking

Cuando se relaciona una categoría con un documento, lo normal es pensar en que el documento pertenece o no a una determinada clase en términos categóricos. Es decir, se toma una decisión booleana respecto a la pertenencia en esa clase. Pero no siempre es fácil tomar una decisión taxativa puesto que es un proceso de clasificación en el que existe incertidumbre. Por ello, otra alternativa consiste en generar un ranking, en el que ordenamos las propuestas del sistema en base a algún tipo de estimación de lo apropiado que es la asignación de la clase al documento en cuestión. En el caso de DPC, construiríamos un ranking de clases para cada documento (se supone que la primera sería aquella que el sistema estima más claramente asociada al documento) y, en el caso de CPC, construiríamos un ranking de documentos para cada clase.

En clasificación hard se obtiene el listado no ordenado de documentos que pertenece a cada clase (CPC) o el listado no ordenado de clases asociadas a un documento (DPC).

La categorización gradual es especialmente útil cuando se trata de aplicaciones críticas, donde los documentos o las clases se ordenan de acuerdo a criterios

probabilísticos, y posteriormente se deja la decisión final de la asignación normalmente un humano. También se recurre a este tipo de clasificación cuando el clasificador obtenido no es suficientemente bueno.

Para un experto humano encargado de tomar la decisión final de clasificación estos rankings serían de gran ayuda, ya que así puede restringir su análisis a los elementos de la parte superior del ranking, en lugar de tener que examinar todo el ranking. Estos clasificadores semiautomatizados son útiles especialmente en aplicaciones críticas donde la eficacia de un sistema automatizado puede ser significativamente inferior a la de un experto humano.

En este trabajo que presentamos, por la complejidad y lo crítico que es la asignación de clases, vamos a utilizar una clasificación con ranking, orientándonos desde un principio a una clasificación semiautomática.

2.3. Representación de los documentos

Representar un documento del modo adecuado es una tarea fundamental que repercute de forma importante en la clasificación. Los documentos, que son típicamente secuencias de cadenas de caracteres, tienen que transformarse en una representación adecuada para los algoritmos de aprendizaje utilizados en clasificación.

Para poder aplicar técnicas de clasificación automática es necesario realizar una serie de pasos previos. Una cuestión básica consiste en definir cómo vamos a representar los documentos, de manera que la representación pueda generarse automáticamente a partir del texto. El sistema más utilizado para la representación de documentos es el modelo vectorial, bien conocido y ampliamente utilizado en RI. El modelo del espacio vectorial ha constituido la base de gran parte de experimentos y sistemas desarrollados. En RI y CAT, el modelo de espacio vectorial [Salton y McGill, 1983; Salton, 1989] se encuentra entre los métodos de representación más utilizados, y es un modelo de altas prestaciones cuando se utiliza con esquemas de pesado y normalización de longitud de documentos.

Este sistema fue desarrollado por Salton [Salton, 1971] en los años 70, y consiste básicamente en representar cada documento con un vector de términos. Cada término lleva asociado un peso o puntuación que trata de reflejar el grado de representatividad o importancia de ese término en ese documento. El cálculo de ese peso se puede definir de varias formas posibles. Un mismo término puede ser más o menos significativo en un contexto que en otro, de manera que tendrá diferente peso en un

documento que en otro. Dependiendo del ámbito de conocimiento en que se inscriba la colección documental, unos términos cobran más importancia que en otros; así, términos que aparecen en casi todos los documentos parecen poco aprovechables para clasificar documentos a partir de ellos. El tamaño o número de términos de cada documento también juega un papel importante. No es lo mismo que un mismo término aparezca dos veces en un documento largo de muchas páginas, a que aparezca dos veces en un documento corto. Hay muchas fórmulas definidas para estimar los pesos. Todas se basan, de una u otra forma, en los términos que aparecen en los documentos y, como es obvio, pueden ser calculadas de manera automática. Los términos definen el espacio vectorial y los documentos se modelan como vectores de términos que pueden ser individualmente tratados y pesados.

Dada una colección de documentos $D = \{d_1, d_2, \cdots, d_m\}$ y sea $T = \{t_1, t_2, \cdots, t_n\}$ conjunto de términos extraídos de *D*. Denominamos función de pesado del término t_j en el documento d_i, a una función *w* que asocia un peso al par $\left(t_j, d_i(t_j)\right)$. Este peso puede ser un valor binario o real y que representa la importancia del término t_j en el documento d_i.

En un documento, estos términos se pueden estructurar en función del nivel al que analicemos el texto [Joachims, 2002]:

- Nivel sub-palabra: descomposición de palabras y su morfología. *n-Grams* son la representación más popular dentro de este nivel. En lugar de utilizar palabras como términos de indexado, construiremos bloques de cadenas de *n* caracteres.
- Nivel palabra: palabras y su información léxica. Esta técnica es muy común y consiste en representar los documentos en un modelo basado en vectores de palabras. Es el que se utiliza en la aproximación denominada bolsa de palabras (*bag of words).*
- Nivel multi-palabra: frases e información sintáctica. La representación con nivel multi-palabra generalmente se usa para indexar términos que incorporan información sintáctica. La estructura sintáctica más comúnmente usada son los sintagmas nominales. Otras formas utilizadas en la generación de términos indexados multi-palabra están basadas en métodos estadísticos.
- Nivel semántico: el significado del texto. Los clasificadores de textos pueden trabajar óptimamente si son capaces de encontrar la semántica de los documentos de una forma eficiente. Desafortunadamente no es posible todavía

extraer automáticamente la semántica de un texto y representarla de forma operativa.

- Nivel pragmático: el significado del texto en función del contexto y la situación.

Dentro de estos niveles, el que se utiliza mayoritariamente en el entorno de investigación de CAT es el nivel de palabra, que es el que manejamos en los experimentos de este trabajo. El modelo vectorial para el nivel de palabra se caracteriza, fundamentalmente, porque asume el principio de independencia, por el que se considera que los términos de un mismo texto no tienen relación entre sí y, por tanto se tratan de manera independiente. Además, no tiene en cuenta el orden en el que aparecen las palabras en el texto. Estas suposiciones dan lugar a un modelado simplista de los textos pero reducen drásticamente la complejidad computacional del problema, ya que permiten representar el documento simplemente como un vector. Además, en muchos casos, esta aproximación tan simple funciona de manera eficaz en distintos problemas de RI y CAT [Sebastiani, 2002].

2.3.1. Funciones de pesado de términos

En la literatura pueden encontrarse multitud de funciones de pesado de términos para calcular la importancia, o relevancia, de un término en el contenido de un texto. Las funciones de pesado se basan fundamentalmente en un cómputo de frecuencias, ya sea dentro del documento a representar, o en el conjunto de documentos de la colección.
Del total de funciones que pueden encontrarse en la literatura, se presentan aquí algunas de las más utilizadas, en donde las diferenciamos en funciones de carácter local y global. En los siguientes apartados se presentan las características principales de cada tipo.

2.3.2. Funciones Locales

Se consideran funciones de pesado local aquellas que toman únicamente información del propio documento para obtener una representación, sin necesidad de ninguna información externa.

- Función binaria (*Binary*, *Bin*). El método de representación más sencillo, dentro de los modelos de representación vectorial, es el conocido como

conjunto de palabras o espacio vectorial binario. La función de pesado es una función binaria, que considera únicamente la presencia o ausencia de un término en un documento para calcular su relevancia dentro del mismo.

$$bin(t_i, d_j) = \begin{cases} 1, & si\ el\ término\ t_i\ aparece\ en\ d_j \\ 0, & si\ no\ aparece \end{cases} \quad (2.1)$$

- Frecuencia de aparición (*Term Frequency*, *TF*). La representación más sencilla dentro de los modelos no binarios es la generada con la función TF. La relevancia se representa por la frecuencia de aparición del término en el documento y puede representarse como:

$$TF(t_i, d_j) = f_{ij} \quad (2.3)$$

donde f_{ij} es la frecuencia del término t_i en d_j.

- Frecuencia normalizada (*Weighted Term Frequency*, *WTF*). Con esta función se genera una representación conocida como frecuencia normalizada, donde la relevancia se calcula como la frecuencia de aparición normalizada del término en el documento:

$$WTF(t_i, d_j) = \frac{f_{ij}}{\sum_{t_p \in d_j} f_{pj}} \quad (2.4)$$

Esta función supone una normalización de la frecuencia de un término en un documento por la suma total de frecuencias del conjunto de términos presentes en el mismo.

- Frecuencia aumentada y normalizada (*Augmented Normalized Term Frequency*, *ANTF*); esta función representa una frecuencia normalizada de un término en un documento y la normalización se realiza con la mayor de las frecuencias presentes en el documento:

$$ANTF(t_i, d_j) = 0.5 + 0.5\ \frac{f_{ij}}{max(\{f_{pj} | t_p\ \in\ d_j\})} \quad (2.5)$$

2.3.3. Funciones Globales

Las funciones de pesado global son aquellas que toman información de la colección para generar los pesos. En [Spärck Jones, 1972] se analiza el uso de esquemas de pesado global para mejorar los sistemas de RI. Este estudio dedujo que los términos que aparecían frecuentemente en una colección podían considerarse importantes en tareas de recuperación de información; sin embargo, si lo que se pretendía era encontrar las diferencias entre los documentos, entonces los términos poco frecuentes en la colección deberían ser tenidos muy en cuenta, y pesados en mayor grado que los términos más frecuentes. Las funciones de pesado globales más conocidas son las siguientes:

- Frecuencia Inversa en Documentos (*Inverse Document Frequency*, *BinIDF*). Esta función trata de enriquecer la representación binaria suponiendo que los términos que aparecen en muchos documentos de la colección no son tan descriptivos como aquellos que aparecen en unos pocos, y se puede expresar como:

$$binIDF(t_i, d_j) = \begin{cases} 1 + \log\left(\frac{N}{df(t_i)}\right), & si\ f_{ij} \neq 0 \\ 0, & si\ f_{ij} = 0 \end{cases} \quad (2.6)$$

  donde $df(t_i)$ es el número de documentos de la colección en los que aparece el término t_i , f_{ij} la frecuencia de t_i en d_j y *N* el número de documentos en la colección.

- Frecuencia del Término × Frecuencia Inversa en Documentos (*Term Frequency - Inverse Document Frequency*, *TFIDF*). Para evitar que el peso de un término sea constante $\forall d_j \in D$, [Salton, 1989] propuso combinar la función $TF(t_i, d_j)$ con el factor $IDF(t_i, d_j)$:

$$TFIDF(t_i, d_j) = f_{ij} \times \log\left(\frac{N}{df(t_i)}\right) \quad (2.7)$$

  La frecuencia f_{ij} del término t_i en d_j, afecta al peso de forma que el valor que toma un mismo término en dos documentos es diferente siempre que la frecuencia de dicho término en cada documento sea también diferente.

- Frecuencia inversa ponderada (*Weighted Inverse Document Frequency*, *WIDF*). Esta función normaliza las frecuencias f_{ij} de un término t_i en un documento d_j con la frecuencia de dicho término en la colección. Esta función tiene la forma:

$$WIDF(t_i, d_j) = \frac{f_{ij}}{\sum_{d_k \in D} f_{ik}} \tag{2.8}$$

  Esta función supone una corrección a la sobreponderación que realiza la función *TF* con los términos frecuentes.

En algunas colecciones, manejando representaciones creadas con funciones de carácter global se obtienen mejores resultados que cuando se usan funciones de pesado local. En general, la función *TF* suele mejorar la representación binaria en problemas CAT; a su vez, las representaciones con factor *IDF* (la función *TFIDF* es la más utilizada) suelen ofrecer mejores resultados que la representación *TF*. La función *TFIDF* es la que vamos a utilizar en esta tesis para representar los documentos.

2.3.4. Funciones de selección de términos (*Feature Selection*)

Teóricamente, cuanto más términos tiene un documento mayor facilidad para discriminar a la hora de hacer clasificación. Sin embargo, la experiencia con algoritmos de aprendizaje ha demostrado que no es siempre así, detectándose algunos inconvenientes: aparición de muchos atributos redundantes o irrelevantes, una degradación en la eficacia de la clasificación y, además, tiempos de ejecución aumentan.

[Luhn, 1958] establece una relación entre el grado de discriminación o poder de resolución de un término y su frecuencia de aparición en la colección. Así las palabras con mayor poder de resolución tienen una frecuencia de aparición media. La justificación para la eliminación de términos infrecuentes se basa en una observación, realizada por [Zipf, 1949] y conocida como Ley de Zipf. Esta establece que, ordenadas las palabras de una colección por su frecuencia total de uso, el producto de su frecuencia total de uso por su posición en el ordenamiento es constante. Esta relación se muestra gráficamente en la figura 2.1. Lo que hizo Luhn fue establecer dos umbrales, corte superior y corte inferior (indicados en la figura 2.1), tratando de

excluir así las palabras no significativas. Los términos que excedían en frecuencia el corte superior eran considerados palabras de uso común, mientras que las que no llegaban al corte inferior se consideraban muy poco comunes, y tienen un uso

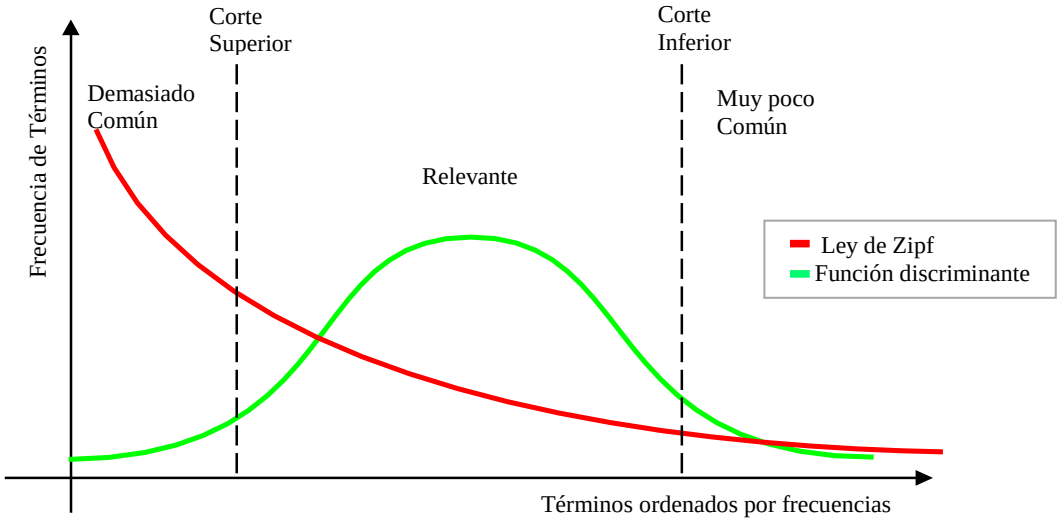

Figura 2.1: Relación entre la frecuencia de aparición de los términos y su relevancia

muy marginal. Las funciones de reducción de términos permiten realizar una ponderación en base a la cual se ordenan todos los términos y se seleccionan un subconjunto de ellos. La selección se puede realizar con un umbral de ponderación mínima o estableciendo una dimensión reducida, generando así un subconjunto del conjunto inicial de términos. Las técnicas de reducción de términos las agrupamos en dos grupos [Joachmis, 2002]:

- Selección de un subconjunto de términos: La nueva representación consiste en un subconjunto de los términos originales.
- Construir términos nuevos. Nuevos términos son introducidos por combinación de los términos originales.

2.3.4.1. Selección de un subconjunto de términos

La selección de términos significa escoger aquellos que son más significativos. De $|T|$ términos escogemos $|T'| < |T|$, el resto se descarta por irrelevante o redundante.

2.3.4.1.1. Eliminación de palabras vacías (*stop-words*)

En este proceso se eliminan aquellas palabras que son muy utilizadas de forma repetitiva en la redacción de los documentos, tales como artículos, preposiciones, conjunciones, etc…, pero que no tienen por sí solas una semántica relevante en el

contenido de un texto. Se considera que este tipo de palabras no tienen capacidad discriminante, ya que aparecen con alta frecuencia en los documentos. La comunidad científica dispone de listas de *stop-words* para numerosos idiomas, entre las que se incluyen también algunos verbos, adverbios o adjetivos de uso frecuente.

La idea de eliminar estas palabras surgió de un trabajo de [Salton et al., 1975], en el que se constató que se obtenían mejores resultados en tareas de RI cuando los documentos eran menos similares entre sí. Es decir, si se les quitaban muchas de las palabras que compartían, logrando reducir así la densidad del espacio vectorial de los documentos.

2.3.4.1.2. Ganancia de información (*Information Gain, IG*).

Esta medida se utiliza para establecer la calidad de un determinado término en una tarea de CAT. Calcula la aportación de información que tiene un término valorando la predicción de una clase en ausencia y presencia de este término.

Así IG puede definirse de la siguiente forma:

$$IG(t_i) = -\sum_{j=1}^{|C|} P(c_j)\, logP(c_j) + P(t_i) \sum_{j=1}^{|C|} P\left(c_j|t_i\right) logP\left(c_j|t_i\right) + P(\bar{t}_i) \sum_{j=1}^{|C|} P\left(c_j|\bar{t}_i\right) logP\left(c_j|\bar{t}_i\right) \quad (2.9)$$

En donde $P(c_j)$ representa la probabilidad a priori de una clase c_j, $P(t_i)$ es la probabilidad de seleccionar un documento que contiene el término t_i, $P(c_j|t_i)$ es la probabilidad condicional de que un documento con el término t_i pertenezca a la clase c_j, $P(\bar{t}_i)$ es la probabilidad de seleccionar documentos que no contiene el término t_i y $P(c_j|\bar{t}_i)$ es la probabilidad condicional de que un documento no contiene el término t_i y pertenezca a la clase c_j.

A partir del cálculo de la ganancia de información de cada término es posible identificar aquellos términos con mayor poder discriminativo. Usualmente se seleccionan aquellos términos que sobrepasan un cierto umbral.

2.3.4.1.3. Información mutua (*Mutual Information, MI*)

Esta función se utiliza fundamentalmente para encontrar relaciones entre términos, muy recurrida en el modelado estadístico del lenguaje. Toma un valor particular para cada clase, y el $MI(t_i)$ se calcula como el valor medio ponderado del conjunto de

todas las clases $MI_{avg}(t_i)$, o como el valor máximo sobre el total de todas las clases $MI_{max}(t_i)$.

$$MI_{avg} = \sum_{j=1...|C|} P(c_j) I_{avg}(t_i, c_j) \approx \sum_{j=1...|C|} P(c_j) \log \frac{|D| \cdot P(t_i, c_j)}{P(t_i) \cdot P(c_j)} \qquad (2.10)$$

$$MI_{max} = max_{j=1...|C|}\{I_{avg}(t_{i_i}, c_j)\} \approx max_{j=1...|C|}\left\{\log \frac{|D| \cdot P(t_i, c_j)}{P(t_i) \cdot P(c_j)}\right\} \qquad (2.11)$$

En donde se considera $|D|$ el número de documentos de la colección, $P(t_i, c_j)$ es la probabilidad de seleccionar documentos de una clase c_j que tiene un término t_i, $P(t_i)$ es la probabilidad de seleccionar documentos que contiene el término t_i, $P(c_j)$ es la probabilidad de seleccionar documentos que pertenece a la clase c_j.

2.3.4.1.4. Chi-square (χ^2)

La función $\chi^2(t_i, c_j)$ mide la falta de independencia entre un término t_i y un documento d_j.

$$\chi^2(t_i, c_j) = \frac{|D| \cdot \left[P(t_i, c_j) \cdot P(\bar{t}_i, \bar{c}_j) - P(t_i, \bar{c}_j) \cdot P(\bar{t}_i, c_j)\right]^2}{P(t_i) \cdot P(\bar{t}_i) \cdot P(c_j) \cdot P(\bar{c}_j)} \qquad (2.12)$$

En donde $P(\bar{t}_i, \bar{c}_j)$ es la probabilidad de un documento no tenga el término t_i ni pertenezca a la clase c_j, $P(t_i, \bar{c}_j)$ es la probabilidad de seleccionar documentos que contienen el término t_i y no pertenecen a la clase c_j, $P(\bar{c}_j)$ es la probabilidad de seleccionar documentos que no pertenezca a la clase c_j.

$\chi^2(t_i, c_j)$ igual que MI, tiene un valor para cada clase, así se puede estimar de dos formas: el valor medio ponderado sobre el conjunto total de clases o el valor máximo sobre el total de clases.

2.3.4.1.5. Odds Ratio

Es la probabilidad $P(t_i|c_j)$ de que un término t_i sea característico de una determinada clase c_j en relación a la probabilidad $P(t_i|\overline{c_j})$ del resto de clases, y la definimos con la siguiente función:

$$OR(t_i, c_j) = \frac{P(t_i|c_j) \times \left(1 - P(t_i|\overline{c_j})\right)}{\left(1 - P(t_i|c_j)\right) \times P(t_i|\overline{c_j})} \quad (2.13)$$

2.3.4.2. Construir términos nuevos

La idea es aplicar una transformación de un espacio a otro, de modo que el espacio destino sea de dimensionalidad menor y puede contener términos no existentes explícitamente en el espacio original de términos. Se pretende que la nueva representación conserve las diferencias relevantes entre los documentos.

2.3.4.2.1. Lematización y truncado (*stemming*)

El proceso de lematización es aquel en el que a cada forma flexiva se le asigna su lema. Este proceso requiere recursos lingüísticos adecuados como pueden ser un software específico o diccionarios electrónicos. Esto consiste en crear un lematizador, programa basado en diferentes algoritmos, que trabaja sobre una colección de textos en lenguaje natural y realiza una extracción automática de términos simplificados a su lema.

El truncamiento (*stemming*), tiene como objetivo reducir el número de términos del vocabulario. En este caso, a cada palabra encontrada en un documento se le eliminan caracteres de los prefijos o sufijos para lograr así agrupar diferentes palabras con una misma forma. Con esto no solo conseguimos reducir el número de términos del vocabulario, sino también agrupar palabras de significado potencialmente similar.

2.3.4.2.2. Indexado Semántico Latente (*Latent Semantic Index, LSI*)

LSI [Deerwester et al., 1990] es una técnica desarrollada para abordar los problemas derivados de la utilización de palabras sinónimas, homónimas y polisémicas en la representación de los documentos. En esencia trata de permitir comparaciones de similitudes semánticas entre textos. Este método consiste en determinar la estructura semántica latente de la relación entre documentos y términos con el fin de superar las deficiencias de los sistemas basados en la mera similitud por coincidencia de términos.

Esta técnica comprime los vectores de documentos en un espacio de dimensiones inferior a partir de los documentos originales, analizando la estructura de los términos en la totalidad de los documentos, de manera que los documentos similares que no comparten los mismos términos se colocan en la misma categoría. Existen varias

técnicas para determinar dicha estructura semántica, pero la más común está basada en la descomposición *SVD* (*Singular Value Decomposition*) de la matriz de términos y documentos de la colección. Con ella se pretende medir la similitud entre diferentes términos de la colección en base a coocurrencia en documentos. De este modo, se pretende incrementar la semejanza en la representación entre documentos cercanos semánticamente. Un problema de aplicar LSI para CAT es que si algún término original es particularmente bueno en sí mismo para clasificar, este poder de discriminación se puede perder en el espacio semántico reducido.

2.3.4.2.3. Agrupamiento de términos (*Term clustering*)

Esta tarea pretende descubrir una estructura latente y oculta en el conjunto de términos. En concreto, el objetivo es la agrupación de palabras con gran parecido semántico. Se pretende encontrar grupos de palabras que presenten relaciones semánticas basadas en su coocurrencia o coausencia en los documentos, de modo que estos grupos puedan utilizarse en lugar de los términos como las dimensiones del espacio vectorial. Se trata de un agrupamiento no supervisado.

Con todas estas técnicas pretendemos elegir atributos que sean relevantes para CAT y lograr el máximo rendimiento con el mínimo esfuerzo. Con menos términos los algoritmos pueden aprender más rápido, con mayor exactitud el clasificador generaliza mejor, y si a la vez conseguimos resultados más simples, estos serán más fáciles de entender.

2.4. Técnicas de clasificación

Hay una buena cantidad de algoritmos propuestos para clasificación. La mayor parte de ellos no son, en realidad, específicos para clasificar documentos, sino que se han propuesto para clasificar todo tipo de objetos. Entre los más utilizados, tenemos:

- Algoritmos probabilísticos
- Algoritmo de Rocchio
- Algoritmos por vecindad
- Árboles de decisión
- Reglas de decisión
- Máquinas de Soporte Vectorial (*Support Vector Machines, SVM*)
- Combinación de clasificadores (Multiclasificadores)

2.4.1. Algoritmos probabilísticos

Se basan en la teoría probabilística, en especial en el teorema de Bayes. Éste permite estimar la probabilidad de un suceso a partir de la probabilidad de que ocurra otro suceso, del cual depende el primero, y aplicar la teoría de las probabilidades condicionadas. Dada una clase c_k y un documento d_j a clasificar calculamos:

$$P(c_k|d_j) = \frac{P(d_j|c_k)\,P(c_k)}{P(d_j)} \tag{2.14}$$

Una vez estimados estos valores de probabilidad $P(c_k|d_j), \forall c_k$, la tarea de un clasificador es simplemente elegir la clase con mayor probabilidad. Para ello, deben estimarse primero las probabilidades a priori de cada clase, $P(c_k)$, y la del documento $P(d_j)$, así como la probabilidad condicionada del documento dada una clase, $P(d_j|c_k)$.

Para calcular $P(d_j|c_k)$ se recurre típicamente a las características o rasgos que definen d_j. A nuestros efectos, las características o rasgos son los términos que conforman los documentos, y tanto su probabilidad de aparición en general, como la probabilidad de que aparezcan en los documentos de una determinada clase, pueden obtenerse a partir de estadísticas sobre la colección.

Sin embargo, la dificultad de calcular el valor $P(c_k|d_j)$ hace que en muchos casos se tenga que recurrir a la suposición de independencia que toma el modelo espacio vectorial. Esto implica que dos términos cualquiera del documento son estadísticamente independientes. Aunque muchos términos no son realmente independientes, esta suposición en la práctica reduce en gran medida la complejidad del clasificador, y la reducción del rendimiento es muy pequeña o inexistente. A este algoritmo se le denomina el clasificador de Naive Bayes, calculando $P(d_j|c_k)$ como:

$$P(d_j|c_k) = \prod_{i=1}^{|T|} P(t_i|c_k) \tag{2.15}$$

Con dichas probabilidades obtenidas de una colección de entrenamiento, podemos estimar la probabilidad de que un nuevo documento pertenezca a cada una de las clases. La implementación del Naive Bayes es sencilla y rápida, y sus resultados son aceptables, como prueban numerosos trabajos experimentales [Yang and Liu, 1999], [Dumais et al., 1998], [Li and Jain, 1998] [Joachmis, 1997] y [Lewis, 1992].

Cuando las colecciones de aprendizaje son pequeñas, pueden producirse errores al estimar las probabilidades. Por ejemplo, cuando un determinado término no aparece nunca en la colección de aprendizaje, pero aparece en los documentos a categorizar. Esto implica la necesidad de aplicar técnicas de suavizado, a fin de evitar distorsiones en la obtención de las probabilidades.

2.4.2. Algoritmo de Rocchio

El algoritmo de Rocchio [Rocchio, 1971] es bien conocido y aplicado en la realimentación de relevancia de RI. En este ámbito, la idea es simple: formulada y ejecutada una primera consulta, el usuario examina los documentos recuperados y determina cuáles le resultan relevantes y cuáles no. Con estos datos, el sistema genera automáticamente una nueva consulta, basándose en los documentos que el usuario señaló como relevantes o no relevantes. En este contexto, el algoritmo de Rocchio proporciona un método para construir el vector de la nueva consulta, recalculando los pesos de los términos de ésta y aplicando un coeficiente a los pesos de la consulta inicial, otro a los de los documentos relevantes y otro distinto a los de los no relevantes.

En el ámbito de la categorización, el mismo algoritmo de Rocchio proporciona un sistema para construir los patrones de cada una de las clases o categorías de documentos. Así, partiendo de una colección de entrenamiento, clasificada manualmente de antemano, y aplicando el modelo vectorial, podemos construir vectores patrón para cada una de las clases, considerando como ejemplos positivos los documentos de entrenamiento de esa clase, y como ejemplos negativos los que no pertenecen a esa clase.

Una vez que se tienen los patrones de cada una de las clases, el proceso de entrenamiento o aprendizaje está concluido. Para clasificar nuevos documentos, simplemente se estima la similitud entre el nuevo documento y cada uno de los patrones. El que proporciona un valor mayor de similitud nos indica la clase a la que se debe asignar ese documento. El algoritmo de Rocchio ha sido utilizado en tareas de categorización con buenos resultados. Algunos trabajos donde se aplica este algoritmo son [Lewis et al., 1996], [Joachims, 1997] y [Figuerola et al., 2001].

2.4.3. Algoritmos por vecindad

El algoritmo del vecino más próximo (*Nearest Neighbour, NN*) es uno de los más sencillos de implementar. Se basa en la aplicación de una métrica que establezca la similitud entre el documento a clasificar y cada uno de los documentos de entrenamiento. La clase o categoría que se asigna al documento será la clase de documento más cercano según la métrica establecida. Una vez localizado el documento de entrenamiento más similar, dado que éstos han sido previamente categorizados manualmente, sabemos a qué categoría pertenece y, por lo tanto, a qué categoría debemos asignar el documento que estamos clasificando.

Una de las variantes más conocidas de este algoritmo es la del *k-nearest neighbour* o *Knn* que consiste en tomar los *k* documentos más parecidos, en lugar de sólo el primero. Como esos *k* documentos tendrán varias clases asociadas, se asignará aquella clase que más veces haya aparecido. El *Knn* une a su sencillez una eficacia notable como lo demuestran los experimentos realizados por [Joachims, 1998], [Yang, 1999], [Yang and Liu, 1999] y que logramos confirmar con los resultados de esta tesis.

2.4.4. Árboles de decisión

Se basan en un particionado recursivo del dominio de definición de los rasgos predictores (términos en nuestro caso). El conocimiento sobre el problema es representado por medio de una estructura de árbol, que se denomina árbol de decisión. La construcción de los árboles de decisión se hace recursivamente de forma descendente (parte de los conceptos generales que se van especificando conforme se desciende en el árbol), por lo que se emplea el acrónimo TDIDT (*Top Down Induction of Decision Trees*) para referirse a la familia completa de algoritmos de este tipo.

Uno de los algoritmos de inducción más populares con árboles de clasificación es el denominado ID3 introducido por [Quinlan, 1986]. El criterio escogido para seleccionar la variable más informativa está basado en el concepto de cantidad de información mutua entre dicha variable y la variable clase. La terminología usada en este contexto para denominar a la cantidad de información mutua es la de Ganancia en Información (*Information Gain, IG*).

[Quinlan, 1993] propone una mejora del algoritmo ID3, al que denomina C4.5. El algoritmo C4.5 se basa en la utilización del criterio ratio de ganancia (*Gain Ratio*). De esta manera se consigue evitar que las variables con mayor número de posibles valores

salgan beneficiadas en la selección. Además, el algoritmo C4.5 incorpora una poda del árbol de clasificación una vez que este ha sido inducido. La poda está basada en la aplicación de un test de hipótesis que trata de responder a la pregunta de si merece la pena expandir o no una determinada rama.

2.4.5. Reglas de decisión

Son clasificadores construidos a partir de métodos inductivos de reglas tipo condicional, donde los literales en la premisa denotan presencia o ausencia de un término o palabra clave. En este sentido, tienden a ser similares a los árboles de decisión, pero además tienden a generar clasificadores más compactos. Inicialmente los documentos se expresan como un vector de términos.

Un clasificador puede expresarse como un conjunto de reglas de tipo Si-Entonces, en las que el antecedente de cada regla está formado por una serie de condiciones que debe cumplir un objeto para que se considere que pertenece a la clase indicada.

2.4.6. Máquinas de Soporte Vectorial (*Support Vector Machines, SVM*)

Los fundamentos de las Máquinas de Vectores de Soporte o *Support Vector Machines* (SVM) se encuentran en los trabajos de Vapnik [Vapnik, 1995] y otros autores sobre la teoría del aprendizaje estadístico basada en el principio de Minimización del Riesgo Estructural desarrollados a finales de los años setenta y durante los ochenta [Vapnik, 1982]. SVM pertenecen a la familia de los clasificadores lineales, que calculan separadores lineales (hiperplanos) en espacios que pueden ser de muy alta dimensionalidad. Presentan un sesgo inductivo muy particular, a través de maximizar el margen de separación entre dos clases. Estos sistemas dan lugar a clasificadores binarios que toman como entrada dos conjuntos de muestras que denominaremos ejemplos positivos y ejemplos negativos. Se trabaja en un modelo de espacio vectorial de d dimensiones, y se asume que esos dos conjuntos son separables en el espacio de representación; en base a ello, se trata de buscar un hiperplano que separe ambos conjuntos de muestras. Las SVM para clasificación binaria intentan encontrar un hiperplano que maximice el margen entre los ejemplos positivos y negativos, mientras que simultáneamente minimice el error de clasificación, como podemos ver en la figura 2.2. Los ejemplos más cercanos a la frontera, los más difíciles de clasificar, se denominan vectores de soporte o *support vectors*. En general, cuanto mayor sea el margen de separación menor será el error de generalización del clasificador. Cada

ejemplo se representa como un vector de dimensión *d*, siendo el objetivo separar dichos ejemplos con un hiperplano de dimensión *d - 1*. Este modelo que representamos en la figura 2.2 es el más sencillo e intuitivo de SVM, aunque también el que tiene condiciones de aplicabilidad más restringidas, ya que parte de la hipótesis de que el conjunto de datos es linealmente separable en el espacio de entrada. Aun así, es explicativo de muchas de las ideas subyacentes en la teoría de las SVM y es la base de todas las demás extensiones.

2.4.6.1. SVM lineal

Una SVM lineal se queda entre todos los hiperplanos posibles que separan las clases, con aquel que maximiza la distancia entre los documentos de cada clase y el propio hiperplano, lo que se denomina margen. Supongamos el conjunto de entrada representado por $\{(x_1, y_1),\ (x_2, y_2), \cdots, (x_n, y_n)\}$ donde $x_i \in R^d$ e $y_i \in \{-1, +1\}\ \forall\ i = 1, \cdots, n$ donde y_i indica la pertenencia o no de cada ejemplo a la clase de referencia. Este conjunto se dice separable si existe algún hiperplano en R^d: $\langle \vec{w}, \vec{x} \rangle + b = 0$ que separa los vectores $\vec{X} = \{x_1, \cdots, x_n\}$ con etiqueta $y_i = +1$ de aquellos con etiqueta $y_i = -1$. El clasificador en este caso sería de la siguiente forma:

$$f(x) = sgn(\langle \vec{w}, \vec{x} \rangle + b) \tag{2.16}$$

El clasificador tiene que dividir el espacio de entrada en dos zonas, como se representa en la figura 2.2. En términos matemáticos, es equivalente a decir que existe un hiperplano de manera que en cada lado del mismo sólo hay ejemplos de una clase.

Geométricamente esa frontera de decisión se representa mediante un hiperplano tal que $f(x) = 0$ y por definición se simbolizara por la ecuación $\langle \vec{w}, \vec{x} \rangle + b = 0$ como muestra la figura 2.2 en un espacio R^2. El vector $\vec{w}$ define la pendiente del hiperplano ya que tiene que ser un vector ortogonal (perpendicular). Solo así su producto escalar es igual a cero. El término *b*, permite determinar cuál es el hiperplano entre los infinitos hiperplanos paralelos que existen. Así, este par de valores $(\vec{w}, b)$ definen el hiperplano que necesitamos encontrar.

Dado un conjunto linealmente separable, existen muchos hiperplanos capaces de separar las clases, sin embargo uno de ellos está más distanciado de ambas clases.

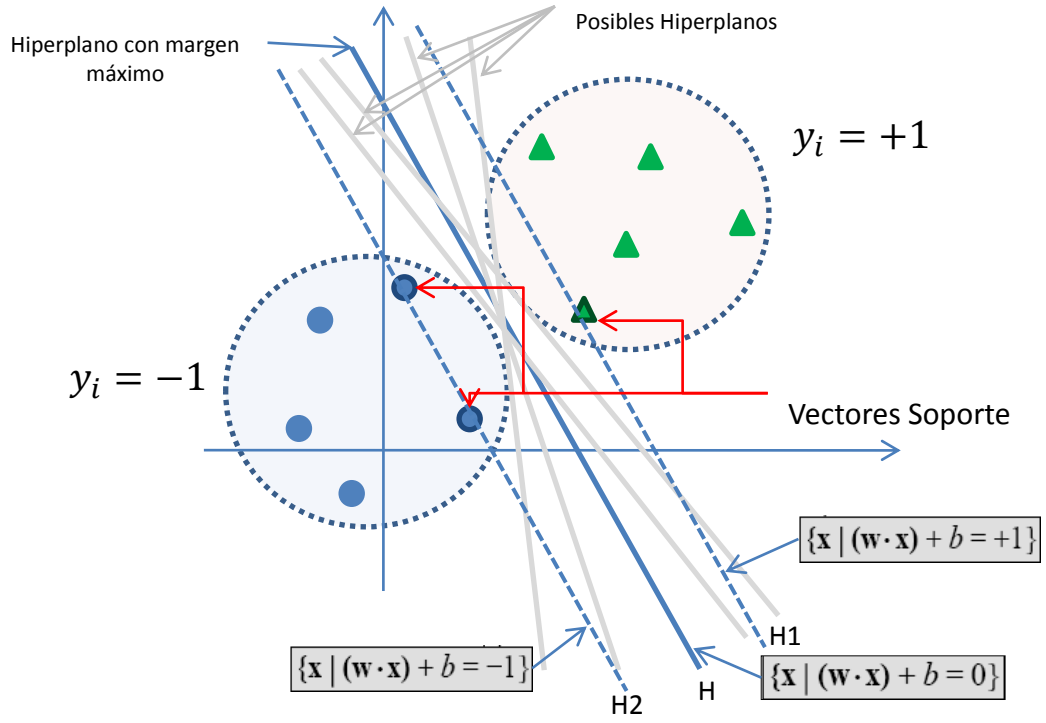

Figura 2.2: Representación de SVM lineal en R^2. El símbolo ▲ los ejemplos positivos y el símbolo ● representa los ejemplos negativos.

Teniendo en cuenta que en el espacio bidimensional la distancia de un punto $p = (x_0, y_0)$ a una recta $H = Ax + By - C = 0$ es:

$$d(p, H) = \frac{Ax_0 + By_0 + C}{\sqrt{A^2 + B^2}} \tag{2.17}$$

De forma similar la distancia entre un punto de H_1 y el hiperplano H es

$$d(H1, H) = \frac{\langle \vec{w} \cdot \vec{x} \rangle + b}{\|\vec{w}\|} = \frac{1}{\|\vec{w}\|} \tag{2.18}$$

Y, por lo tanto, la distancia entre H_1 y H_2 es

$$d(H1, H2) = \frac{2}{\|\vec{w}\|} \tag{2.19}$$

En consecuencia, el hiperplano óptimo es aquel que maximiza la distancia entre H_1 y H_2, por ello deben cumplirse las siguientes condiciones:

$$\begin{aligned} &\text{mín} && \frac{1}{2}\|\vec{w}\| \\ &\text{sujeto a} && y_i[\langle \vec{w} \cdot \vec{x}_i \rangle + b] \geq 1 \quad \forall i = 1, \cdots, n \end{aligned} \tag{2.20}$$

A nivel algorítmico, el aprendizaje de las SVM representa un problema de optimización que se puede resolver usando técnicas de programación cuadrática. El problema consiste en un problema de programación cuadrática donde la función objetivo es convexa, y los vectores que satisfacen las restricciones forman un conjunto convexo. Para resolver el problema de optimización con restricciones se utiliza los multiplicadores de Lagrange:

$$L_p(\vec{w}, b, \alpha_i) = \frac{1}{2}\|\vec{w}\| - \sum_{i=1}^{n} \alpha_i (y_i(\vec{w} \cdot \vec{x}_i + b) - 1) \tag{2.21}$$

Esto significa que se puede resolver el siguiente problema dual asociado al problema primal: maximizar la función $L_p(\vec{w}, b, \alpha_i)$ respecto a las variables duales α_i sujeta a las restricciones impuestas para que los gradientes de L_p con respecto a *w* y *b* sean nulos, y sujeta también al conjunto de restricciones.

$$\frac{dL_p}{dw_j} = w_j - \sum_{i=1}^{n} \alpha_i y_i x_{ij} = 0 \quad \Rightarrow w_j = \sum_{i=1}^{n} \alpha_i y_i x_{ij} \tag{2.22}$$

$$\frac{dL_p}{db} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{2.23}$$

Las condiciones de Karush-Kuhn-Tucker (KKT) son necesarias para satisfacer que los problemas de optimización no lineal con restricciones de desigualdad tengan una solución óptima global. Se trata de una generalización del método de los multiplicadores de Lagrange para restricciones de desigualdad. Así añadimos al Lagrangiano las condiciones suficientes de optimización o condiciones KKT:

$$y_i[\langle \vec{w} \cdot \vec{x}_i \rangle + b] \geq 1 \quad \forall i = 1, \cdots, n \tag{2.24}$$

$$\alpha_i \geq 0 \tag{2.25}$$

$$\alpha_i (y_i[\langle \vec{w} \cdot \vec{x}_i \rangle + b] - 1) = 0 \tag{2.26}$$

Sustituyendo la expresión 2.22 en la expresión 2.21 obtenemos que el Lagrangiano se puede formular de la siguiente forma:

$$L_p(\vec{w}, b, \alpha_i) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,s=1}^{n} \alpha_i \alpha_s y_i y_s \vec{x}_i \vec{x}_s \tag{2.27}$$

A partir de esta expresión, aplicando dualidad Lagrangiana se puede obtener un problema dual conocido como el Dual de Wolfe:

$$\begin{aligned} máx \;\; L_p(\vec{w}, b, \alpha_i) &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,s=1}^{n} \alpha_i \alpha_s y_i y_s \vec{x}_i \vec{x}_s \\ \text{s.a} \quad & \textstyle\sum_{i=1}^{n} \alpha_i y_i = 0 \\ & \alpha_i \geq 0 \end{aligned} \tag{2.28}$$

La resolución se reduce, entonces, a obtener los valores óptimos para los multiplicadores α_i. Una vez conocidos éstos, los valores óptimos de las variables primales $\vec{w}$ se obtienen de la ecuación 2.22. Se puede definir la función discriminante o de clasificación como:

$$f(x) = sgn(\langle \vec{w}, \vec{x} \rangle + b) = \sum_{i=1}^{n} \alpha_i \, y_i (\vec{x}_i \cdot \vec{x}) + b \tag{2.29}$$

El objetivo era encontrar los valores de $(\vec{w}, b)$ que definen el hiperplano de separación óptimo, se ha formulado este problema de optimización con restricciones y se ha explicado cómo resolverlo utilizando una formulación dual que consiste en calcular los multiplicadores óptimos α_i. Las restricciones de la fórmula 2.25 se pueden descomponer en dos tipos:

- Las restricciones en donde $\alpha_i > 0$, son las activas.
- Las restricciones en donde $\alpha_i = 0$, son las inactivas.

De la ecuación $\alpha_i(y_i[\langle \vec{w} \cdot \vec{x}_i \rangle + b] - 1) = 0$, una restricción es activa ($\alpha_i > 0$) solo si la distancia es igual al margen y es una restricción inactiva ($\alpha_i = 0$) sí la distancia es superior al margen. Los elementos que resaltamos en la figura 2.2 son los que denominamos Vectores de Soporte (VS) que corresponden a una restricción activa ($\alpha_i > 0$), el resto de elementos son restricciones inactivas ($\alpha_i = 0$).

El vector óptimo $\vec{w}$ lo podemos representar como $w_j = \sum_{i=1}^{n} \alpha_i y_i x_{ij}$, en donde $\alpha_i = 0$ en una restricción inactiva y $\alpha_i > 0$ en una restricción activa. Si descartamos los términos nulos, se puede reescribir el vector óptimo $\vec{w}$ de la siguiente forma:

$$w_j = \sum_{x_{ij}\ es\ un\ VS} \alpha_i y_i x_{ij} \tag{2.30}$$

Poder calcular el vector óptimo $\vec{w}$ como una combinación de los vectores de soporte tiene consecuencias importantes:

- El número de vectores de soporte puede ser muy pequeño en comparación con el tamaño de la colección de entrenamiento (en la figura 2.2 hay 3 vectores de soporte de los 12 elementos de la colección de entrenamiento). Por lo tanto $\vec{w}$ puede ser una combinación de un pequeño número de vectores.
- La definición del hiperplano óptimo está condicionada exclusivamente por los vectores soporte. Es decir los elementos que no son vectores soporte se pueden quitar de la colección de entrenamiento sin que esto influya sobre el hiperplano óptimo.
- Desde el punto de visa computacional es más eficiente desarrollar algoritmos especiales para conocer α_i partiendo de la base que la mayoría de ellos son iguales a cero. Como los vectores de soporte son normalmente un número muy pequeño en comparación al tamaño de la colección de entrenamiento, la clasificación de un nuevo elemento suele ser muy rápida.

2.4.6.2. SVM lineal con margen blando (*soft margin*)

En las aplicaciones reales, el problema planteado hasta ahora tiene pocas posibilidades de que se pueda implementar, ya que en gran medida estos no son linealmente separables. Por ello es interesante la posibilidad de permitir que las condiciones impuestas en (2.20) no siempre se cumplan para todos los ejemplos. Esto da lugar a lo se denomina SVM con margen blando (*soft margin*). Las razones por las que se producen estas situaciones pueden ser diversas: ejemplos con ruido, errores humanos de clasificación, o la más interesante en algunos casos, ampliar el margen de separación entre clases, incrementando la facilidad de clasificación, como podemos observar en la figura 2.3. Con este planteamiento estamos asumiendo un riesgo

adicional, y es posible que nos sobreajustemos a los datos de entrenamiento al emplear un espacio más enriquecido y favorecedor.

Permitir errores de clasificación desde la perspectiva de nuestro problema de optimización equivale a consentir que las restricciones impuestas en (2.20) para estos ejemplos no se cumplan. Para representar estas situaciones en nuestro problema de optimización introducimos en cada restricción original una variable de holgura (*slack*), ξ_i, que mide el coste de violar esta restricción, respecto a la clasificación correcta de este ejemplo. La expresión de estas nuevas restricciones es:

$$f(x_i) = y_i[\langle \vec{w} \cdot \vec{x}_i \rangle + b] \geq 1 - \xi_i \quad \forall i = 1, \cdots, n \tag{2.31}$$

Las variables ξ_i tendrán el valor cero cuando el ejemplo este situado en el margen definido para su clase y un valor mayor que cero cuando no lo esté. Es decir, $\xi_i \neq 0$ en (2.31) en aquellos ejemplos que no cumplan (2.20).

Después de definir las nuevas restricciones que nos permiten la presencia de errores, falta plantear la función a optimizar. Esta función en el caso de no permitir errores consistía en maximizar el margen. Es evidente que en esta nueva situación esto no basta, ya que podríamos maximizar el margen simplemente con el aumento de ejemplos mal clasificados. La función debe incluir de alguna forma los errores que está permitiendo el hiperplano. Este aspecto se representa añadiendo un término que indique un coste o una penalización de la solución:

$$\frac{1}{2}\|\vec{w}\| + C\sum_{i=1}^{n} \xi_i \tag{2.32}$$

La constante *C*, que multiplica al término relativo al coste, nos permite controlar en qué grado influye dicho término en la minimización. Esta constante nos permitirá controlar el grado de sobreajuste que permitimos. El valor de *C* se determinará empíricamente observando los resultados obtenidos para distintas configuraciones. Si el valor de *C* es grande, pocas variables de holgura ξ_i podrán tener un valor distinto de 0. Es decir, pocos elementos x_i podrían violar la restricción. Si por el contrario, *C* es pequeño permitimos que más ejemplos violen la restricción.

El problema de optimización en el caso general no separable linealmente o para el caso separable linealmente que queremos ampliar el margen queda definido de la siguiente forma:

$$\begin{aligned} & \frac{1}{2}\|\vec{w}\| + C\sum_{i=1}^{n}\xi_i \\ \text{sujeto a} \quad & y_i[\langle \vec{w}\cdot\vec{x}_i\rangle + b] \geq 1-\xi_i \quad \forall i = 1,\cdots,n \\ & \xi_i \geq 0 \end{aligned} \tag{2.33}$$

Si todos los ejemplos están en el margen correspondiente a su clase, entonces ξ_i siempre es cero y las condiciones de (2.33) se transforman en las de (2.20) (problema separable linealmente).

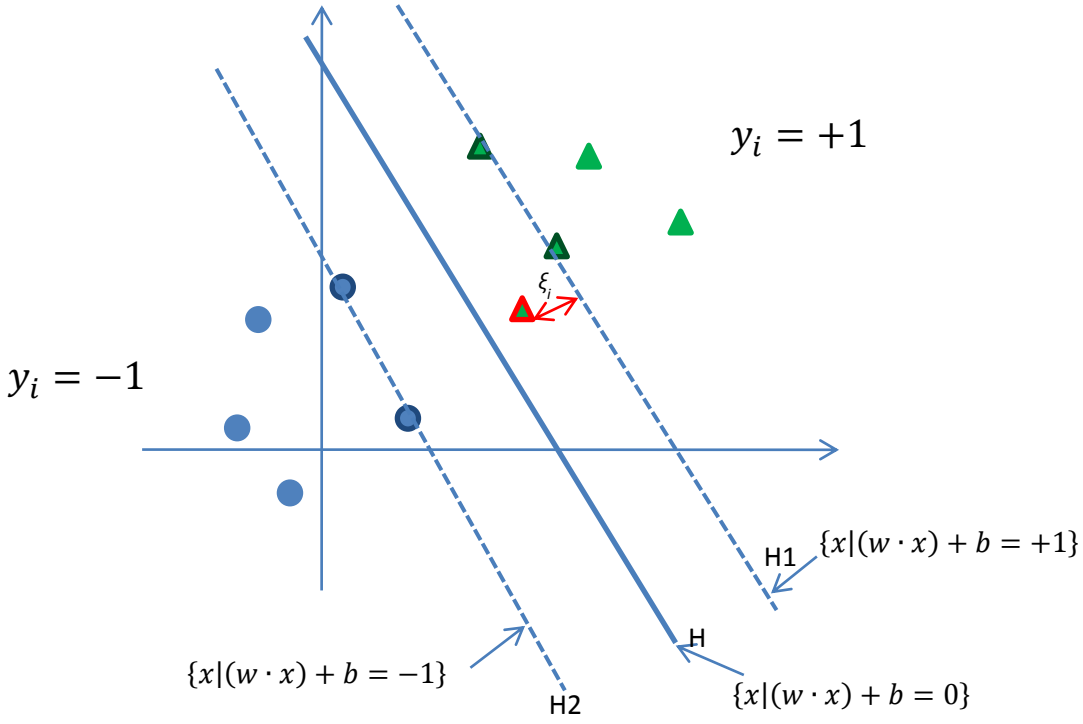

Figura 2.3: SVM lineal con margen blando

En el formato dual obtenemos las siguientes expresiones:

$$\begin{aligned} & máx L_p(\vec{w}, b, \alpha_i) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,s=1}^{n}\alpha_i\alpha_s y_i y_s \vec{x}_i\vec{x}_s \\ \text{s.a.} \quad & \textstyle\sum_{i=1}^{n}\alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned} \tag{2.34}$$

Por las condiciones de KKT si:

- $\alpha_i = 0 \Longrightarrow f(\vec{x}_i) \geq 1 \; y \; \xi_i = 0$ entonces $\vec{x}_i$ no es un vector de soporte.
- $0 < \alpha_i < C \implies f(\vec{x}_i) = 1 \; y \; \xi_i = 0$ ahora $\vec{x}_i$ es un vector de soporte.
- $\alpha_i = C \implies f(\vec{x}_i) \leq 1 \; y \; \xi_i > 0$, $\vec{x}_i$ es un vector de soporte.

2.4.6.3. SVM no lineal

En muchas ocasiones las soluciones anteriores para ciertos problemas no aportan una respuesta adecuada. Hay otras formas de conseguir que un problema no linealmente separable se transforme en un problema linealmente separable y así poder construir un clasificador lineal en el espacio transformado. El teorema de Cover [Cover, 1965] establece que un problema de clasificación de patrones en un espacio de dimensionalidad alta es más probable que sea linealmente separable que en un espacio de dimensionalidad baja. La función que realiza esta transformación la denominaremos función núcleo o *kernel*. Dicho de otro modo, supongamos que existe una transformación no lineal del espacio de entrada en un espacio $\Im$ de mayor dimensión (espacio de características), en el que si pueden ser separados por un hiperplano (Figura 2.4). Para ello, se utiliza una función Φ, tal que:

$$\begin{aligned} \Phi: \Re^D &\rightarrow \Im \\ x &\rightarrow \Phi(x) \end{aligned} \tag{2.35}$$

Φ está dotada de un producto escalar $<\Phi(x),\Phi(y)>$ ($\Im$ es un espacio de Hilbert). Para ciertos espacios de características y ciertas transformaciones existe una forma de calcular el producto escalar usando las funciones núcleo o *kernels*. Una función núcleo es una función $K: X \times X \rightarrow \Re$ tal que $K(x,y) = <\Phi(x),\Phi(y)>$, donde Φ es una transformación de X en un cierto espacio de Hilbert $\Im$. Es decir, el producto escalar se puede calcular usando la función núcleo, quedando implícita la transformación del espacio de entrada en el espacio de características.

La modificación que tenemos que realizar para formular y resolver el problema de la clasificación con SVM consiste en reemplazar x_i por $\Phi(x_i)$ quedando la definición del hiperplano óptimo de la siguiente manera:

$$\begin{aligned} &\text{mín} && \frac{1}{2}\|\vec{w}\| \\ &\text{sujeto a} && y_i[\langle \vec{w} \cdot \Phi(x_i)\rangle + b] \geq 1 \quad \forall i = 1, \cdots, n \end{aligned} \tag{2.36}$$

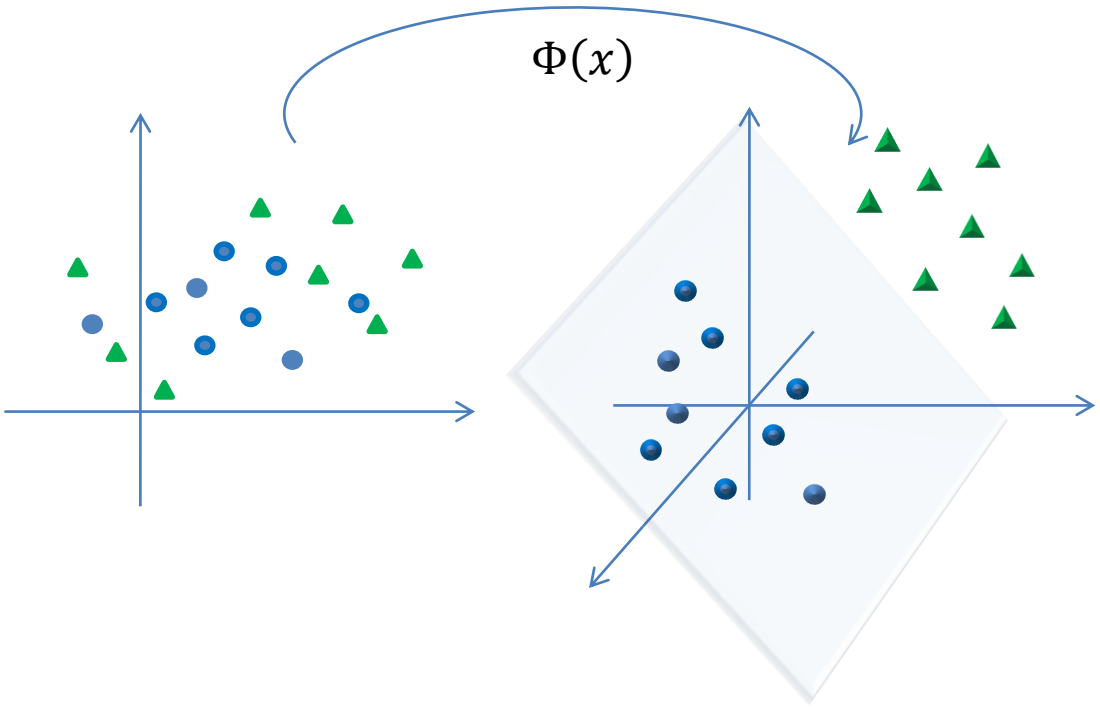

Figura 2.4: Transformación de los datos de entrada a un espacio de mayor dimensión

Recordemos que no se necesita una representación explícita de los vectores en el espacio de características, sino sólo un modo de calcular el producto escalar. Así que no debemos preocuparnos por la dimensión del espacio en cuanto a coste computacional. Lo único que necesitamos conocer es la función kernel. Existe un gran número de funciones posibles, las más comunes las representamos en la tabla 2.1:

Lineal	$k(x_i, x_j) = x_i^T x_j$
Polinómico	$k(x_i, x_j) = (x_i^T x_j + 1)^n$
Gaussiano de base Radial (RBF)	$k(x_i, x_j) = exp\left(-\frac{\|x_i^T - x_j\|^2}{2\sigma^2}\right)$

Tabla 2.1: Kernels más comunes en SVM

En la representación dual la enunciación quedaría de la siguiente forma:

$$\begin{aligned} máx L_p(\vec{w}, b, \alpha_i) &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,s=1}^{n} \alpha_i \alpha_s y_i y_s K(\vec{x}_i \vec{x}_s) \\ \text{s.a} \quad & \textstyle\sum_{i=1}^{n} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned} \tag{2.37}$$

Y la función de clasificación en la formulación dual sería:

$$f(\vec{x}) = \sum_{i \; es \; SV} \alpha_i \, y_i K(\vec{x}_i \cdot \vec{x}) + b \tag{2.38}$$

SVM es uno de los clasificadores que obtiene mejores resultados en clasificación de textos, como lo demuestran los trabajos realizados por [Dumais et al., 1998], [Joachims, 1998], [Li and Yamanishi, 1999] y [Yang and Liu, 1999].

2.4.7. Combinación de clasificadores (Multiclasificadores)

En distintas aplicaciones de clasificación automática se ha puesto de manifiesto que muchas veces no hay un solo algoritmo de clasificación que siempre funcione de forma adecuada para todos los casos. Por esto, han surgido estrategias de clasificación compuestas de múltiples clasificadores. Los clasificadores individuales se complementan entre si y se consiguen resultados más fiables combinando las predicciones individuales. Esta combinación de varios modelos de clasificación se denomina comités de clasificadores (*committees ensembles*) o multiclasificadores. Esto implica la selección de un determinado número *k* de clasificadores, y la elección de un método para combinar los resultados.

Dentro de las formas de caracterizar los distintos tipos de comités de clasificadores, distinguimos si se trabaja siempre o no con el mismo clasificador, si se trabaja siempre con el mismo espacio de características, etc. Si se trabaja con el mismo espacio de características puede ser utilizando distintos clasificadores o el mismo clasificador pero entrenando con distintas muestras.

Los multiclasificadores que trabajan con distintas muestras se pueden clasificar en: muestras aleatorias (*bagging*), muestras en serie dando más énfasis en el siguiente clasificador a las muestras mal clasificadas por el clasificador anterior (*boosting* y

cascada). También se puede usar siempre el mismo clasificador pero con parámetros distintos.

Según su estructura los clasificadores se pueden clasificar en: en paralelo, los resultados de cada uno se pasan al que los combina (votación, *stacking*); en serie, cada clasificador es llamado secuencialmente, y usa los resultados del anterior clasificador, (cada vez van siendo más complejos); y jerárquico, se combinan en jerarquía, con las salidas de uno alimentando a un nodo padre.

En cuanto a la decisión del método de combinación de resultados, son varias las opciones que se han utilizado en la literatura:

- Por mayoría, se obtienen los valores binarios devueltos por los clasificadores, y la opción con más votos será escogida. Existen distintas alternativas:
  - Voto por mayoría
  - Mayoría absoluta
  - Voto por mayoría con umbral
  - Voto por mayoría con peso

- Combinación de los pesos**:** la combinación de los resultados de todos los clasificadores da como resultado el valor global, que se utiliza para la clasificación final. Entre las alternativas de combinación distinguimos:
  - Media aritmética
  - Mínimo
  - Máximo
  - Mediana
  - Producto
  - Media generalizada

- Selección dinámica del clasificador**:** de todos los clasificadores utilizados, se comprueba cuál de ellos es el más eficiente en la fase de validación, y su decisión es la que se toma por válida.

- Combinación adaptativa de clasificadores: se suman las decisiones de todos los clasificadores, pero su valor es ponderado en función del resultado obtenido en la fase de validación por cada clasificador.

Los buenos resultados que se pueden conseguir con este tipo de clasificadores lo demuestran los trabajos de [Weiss et al., 1999] y [Schapire and Singer, 2000].

2.5. Métodos de evaluación

Validar un clasificador nos sirve para medir su capacidad de predicción sobre nuevas peticiones que le lleguen en el futuro para que las clasifique. El objetivo de la clasificación supervisada es la inducción de modelos de clasificación que tengan una buena capacidad generalizadora. Los clasificadores ante un nuevo caso del que se conocen los valores de las variables predictoras tienen que ser capaces de clasificarlo correctamente con una alta probabilidad. Precisamente el objetivo de la evaluación es estudiar métodos que estimen dicha probabilidad con objeto de que tengamos una idea de la habilidad del modelo clasificación.

Como ya se ha dicho, la categorización de texto basada en aprendizaje necesita ejemplos para poder construir un clasificador. Normalmente, se recolectan ejemplos de textos correctamente clasificados. La tarea de etiquetación suele ser realizada por expertos humanos. A estos ejemplos recolectados se les denomina generalmente colección o corpus; y está definido como el conjunto $D = \{d_1, \cdots, d_{|D|}\}$ preclasificados en $C = \{c_1, \cdots, c_{|C|}\}$, donde *C* es el conjunto de categorías existentes y *D* el conjunto de documentos.

2.5.1. Métodos para estimar la probabilidad de clasificación correcta de un clasificador

La colección o corpus se puede dividir de diferentes formas. Lo más normal es dividirla en dos conjuntos disjuntos: entrenamiento y prueba o test. En donde el conjunto de entrenamiento sirve para educar al clasificador, y el de prueba para medir la efectividad conseguida. En algunas ocasiones se ajusta alguno de los parámetros del clasificador para mejorar la efectividad del clasificador. Para ello se puede reservar una parte del conjunto de entrenamiento no utilizado antes, que permite observar el resultado del clasificador con el ajuste de estos parámetros. A este subconjunto de la colección de entrenamiento se le suele denominar conjunto de validación.

Al dividir la colección en entrenamiento y test, no parece justo el estimar la probabilidad de correcta clasificación a partir del porcentaje de casos que el modelo clasificatorio inducido es capaz de clasificar correctamente en el conjunto de casos a partir del cual se ha inducido el clasificador. Es intuitivo que el proceder de esta manera proporcionaría unas estimaciones demasiado buenas. Por ello debemos utilizar formas que consigan estimaciones honestas de la probabilidad de éxito del clasificador. Esta se realiza a partir del conjunto de casos que usamos para construir el clasificador final, y el conjunto de peticiones de las cuáles sabemos su verdadera clasificación, bajo el supuesto de que estas provienen de una muestra aleatoria. Así podemos definir bajo estos criterios los siguientes métodos para el diseño de los experimentos: método H y métodos basados en el remuestreo.

2.5.1.1. Método H

El método H (*holdout*) también conocido como método de entrenamiento-test, se basa en particionar la colección de tamaño M en dos colecciones de tamaño M_1 y M_2 en donde $M_1 + M_2 = M$. La primera colección (M_1) se denomina de entrenamiento, ya que a partir del mismo se forma el modelo de clasificación. La evaluación de dicho modelo, es decir la estimación de la probabilidad de éxito de dicho modelo frente a casos nuevos, se obtiene por medio del porcentaje de casos bien clasificados obtenidos para la segunda colección (M_2). A esta segunda colección se denomina colección de test, ya que es la que testea la bondad del modelo de clasificación.

Conviene tener presente que con este método el modelo clasificatorio obtenido, y que posteriormente se va a aplicar, se ha instruido a partir de M_1 casos. Suele ser habitual el utilizar las proporciones $\frac{2}{3}M$ y $\frac{1}{3}M$ respectivamente para las colecciones de entrenamiento y test. Se suele utilizar el método H en el caso de que M sea del orden de millares o superior.

2.5.1.2. Métodos basados en remuestreo

En este punto se explicarán tres métodos basados en el remuestreo que sirven para estimar la probabilidad de éxito de un sistema clasificación. La gran diferencia con relación al método H descrito anteriormente radica en que los métodos basados en el remuestreo se estiman la probabilidad de éxito en toda la colección.

Los tres métodos que se exponen en este punto son: submuestreo aleatorio (*random subsampling*), validaciones cruzadas de *k* partes (*k-fold cross validation*) y dejando uno fuera (*Leave one out*).

- Submuestreo aleatorio (*random subsampling*)

  El método Submuestreo aleatorio viene a ser una generalización del método H, realizándose éste múltiples veces sobre diferentes particiones independientes de la colección de entrenamiento y la colección de test. Así, la estimación de la tasa de error se efectúa a partir de la media de las tasas de error obtenidas en los diferentes experimentos.

- Validaciones cruzadas de *k* partes (*k-fold cross validation*)

  En el método de validaciones cruzadas de *k* partes también es una generalización del método H. La colección se particiona en *k* subconjuntos disjuntos de aproximadamente el mismo tamaño, donde *k - 1* subconjuntos constituyen la colección de entrenamiento y el subconjunto restante la colección de test. Tenemos que repetir el proceso *k* veces sobre las distintas combinaciones de *k - 1* subconjuntos de entrenamiento. Los *k* porcentajes de casos bien clasificados se promedian para estimar el rendimiento del clasificador.

- Dejando uno fuera (*Leave one out*)

  La validación dejando uno fuera es un caso particular de la validación cruzada, en la cual el parámetro *k* viene a ser igual al número de elementos de la colección. De esta forma, la colección de test está formada por un único elemento y la colección de entrenamiento por la colección total menos ese único elemento que pertenece a la colección de test.

2.5.2. Métricas de evaluación en CAT

La evaluación experimental de un clasificador utiliza básicamente dos tipos de medidas: aquellas que estiman la efectividad (capacidad de satisfacer las necesidades de los usuarios en cuanto a toma de decisiones de clasificación correctas) y las que estiman la eficiencia. Estas últimas tratan de medir los tiempos de respuesta, la complejidad teórica o los recursos computacionales. En este trabajo nos vamos a

centrar en aquellas métricas que determinan la efectividad. La utilización de unas métricas u otras va a depender del tipo de problema de clasificación y de otros factores que examinaremos. En la subsección 2.2.3, presentamos dos formas de clasificar texto: una clasificación *hard*, donde se toma una decisión booleana respecto a la pertenencia o no del documento a una determinada clase; o una gradual, donde se estima una probabilidad de pertenencia. De manera similar, hay dos formas de crear los clasificadores: una manera dura o automatizada, donde se deja la responsabilidad completa de la clasificación del texto al clasificador; y una parcialmente automatizada, donde el clasificador pondera la pertenencia de asignar la clase al documento en cuestión. Dependiendo del tipo de clasificación utilizada es más recomendable utilizar uno u otro método de validación del clasificador.

Vamos a introducir algunos conceptos relacionados con la evaluación experimental del clasificador. Empleamos para ello una colección de 2 clases que denominamos 0 o (+) y 1 o (-). Definimos la tabla de contingencia o matriz de confusión de la siguiente forma:

	Clase Verdadera	
Clase Predecida	Clase 0(+)	Clase 1(-)
Clase 0(+)	TP	FP
Clase 1(-)	FN	TN

Tabla 2.2: Tabla de contigencia para dos clases

Categorías $C = \{c_1, \cdots, c_{\lvert C\rvert}\}$		Clase Verdadera	
		SI	NO
Clase Predecida	SI	$\sum_{i=1}^{\lvert c\rvert} TP_i$	$\sum_{i=1}^{\lvert c\rvert} FP_i$
	NO	$\sum_{i=1}^{\lvert c\rvert} FN_i$	$\sum_{i=1}^{\lvert c\rvert} TN_i$

Tabla 2.3: Tabla de contigencia global

Las tablas de contingencia 2.2 y 2.3 nos permiten ver la distribución de los aciertos y errores cometidos por un clasificador para dos o más clases. En estas tablas de contingencia se cruza la variable derivada de la clasificación predecida por el clasificador con la variable que determina la verdadera clasificación.

En donde,

- ***TP$_i$*** representa el número de casos que el clasificador predijo que eran de la clase c_i (en la tabla 2.2 es la clase 0), y los ejemplos efectivamente pertenecían a c_i. *TP* significa verdaderos positivos (*True Positive*).
- ***TN$_i$*** representa el número de casos que el clasificador predijo que no eran de la clase c_i, y los ejemplos efectivamente no pertenecían a c_i. *TN* significa verdaderos negativos (*True Negative*).
- ***FP$_i$*** representa el número de casos que el clasificador clasifico como pertenecientes la clase c_i, pero no pertenecían a tal clase. *FP* significa falsos positivos (*False Positive*).
- ***FN$_i$*** representa el número de casos que el clasificador no clasificó en la clase c_i en la cual tendrían que haber sido clasificados. *FN* significa falsos negativos (*False Negative*).

2.5.2.1. Precisión y recall

La mayoría de las métricas que se utilizan en CAT proceden de las definiciones realizadas en el entorno clásico de RI, como son la precisión (π) y el índice de recuperación – recall (ρ). Estas probabilidades se pueden estimar en términos de la tabla de contingencia para una colección de test y una clase c_i de la siguiente forma:

$$\hat{\pi}_i = \frac{TP_i}{TP_i + FP_i} \qquad \hat{\rho}_i = \frac{TP_i}{TP_i + FN_i} \tag{2.39}$$

Para obtener valores estimativos de π y ρ para la colección completa tenemos dos métodos diferentes:

– **Microaveraging**: los cálculos se basan en la suma total de todas de las decisiones individuales de clasificación

$$\hat{\pi}^{\mu} = \frac{TP}{TP + FP} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i + FP_i)} \tag{2.40}$$

$$\hat{\rho}^{\mu} = \frac{TP}{TP + FN} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i + FN_i)} \tag{2.41}$$

en donde μ indica microaveraging. La tabla de contingencia global (tabla 2.3) se obtiene sumando las tablas de contingencia específicos para cada clase. Con

microaveraging todos los documentos tiene igual peso, pero las clases con más documentos influyen en mayor medida en la evaluación.

– **Macroaveraging**: los cálculos se basan en la obtención local de los valores de precisión y recall para cada clase, obteniendo después la media total.

$$\hat{\pi}^M = \frac{\sum_{i=1}^{|C|} \hat{\pi}_i}{|C|} \qquad \hat{\rho}^M = \frac{\sum_{i=1}^{|C|} \hat{\rho}_i}{|C|} \tag{2.42}$$

en donde *M* indica macroaveraging. En este método todas las clases tienen el mismo peso en el resultado estimado para toda la colección. Esta medida nos capta mejor la capacidad del funcionamiento del clasificador en todas las clases.

Por tanto, microaveraging trata de dar a las clases una importancia proporcional al número de ejemplos positivos que le corresponden, mientras que con macroaveraging todas las clases importan lo mismo.

Estos dos métodos pueden dar resultados muy diferentes, sobre todo si las diferentes clases son muy desiguales. La elección del método de evaluación va depender del tipo de clasificación que realicemos y de la colección con la que se realizan los experimentos, siendo microaveraging la más utilizada dentro de la literatura en CAT [Sebastiani, 2002]. En aquellas colecciones que tenga una distribución de clases asimétrica es recomendable utilizar los dos métodos de medida.

Existen otras métricas de efectividad alternativa a precisión y recall, en la literatura sobre aprendizaje automático. Por ejemplo *Accuracy*, que se calcula como $\hat{A} = \frac{TP+TN}{TP+TN+FP+FN}$ y el error $\hat{E} = 1 - \hat{A} = \frac{FP+FN}{TP+TN+FP+FN}$. Yang [Yang ,1999] explica que habitualmente no son apropiados estos valores en CAT, debido al gran valor que adquiere el denominador de la división, haciendo la evaluación relativamente insensible a las variaciones en el número de decisiones correctas. En algunos entornos se utilizan también medidas como $fallout = \frac{FP}{FP+FN}$.

2.5.2.2. Medidas de combinación de la efectividad

En una clasificación *hard* si queremos obtener una única medida de evaluación del clasificador podemos utilizar la medida F_β. La medida F_β , en base a las medidas clásicas de precisión y recall, la función F_β permite estimar, a través de la media armónica de ambas métricas, la bondad del clasificador mediante un único valor.

$$F_\beta = \frac{(\beta^2 + 1)\pi\rho}{\beta^2\pi + \rho} \tag{2.43}$$

β es un parámetro que regula la influencia de π o ρ en la formula. Usualmente se utiliza $\beta=1$, obteniendo la métrica F_1 :

$$F_1 = \frac{2\pi\rho}{\pi + \rho} \tag{2.44}$$

En otro tipo de clasificaciones (como puede ser una colección multiclase) donde el clasificador produce un ranking de clases por cada documento, la métrica que se suelen utilizar es la siguiente:

- Media de los 11 puntos de precisión (*11-point average precisión*) que se calcula como la precisión (π) medida sobre cada valor 0.0, 0.1, 0,2, … ,1.0 de recall (ρ).

2.5.2.3. Medidas para clasificadores específicos

En determinados entornos se definen métricas específicas en donde el usuario analiza y valora mejor la función del CAT. En este caso las métricas se adaptan al modelo de clasificación, y así el usuario juzga mejor el funcionamiento del clasificador. Para entornos singulares como la clasificación de documentos de alta mediante códigos CIE-9-MC tenemos las siguientes métricas especiales [Larkey and Croft, 1996]:

- **Top candidato**. Proporción de casos donde la clase principal del documento de test (esto es la clase del diagnóstico principal) es el primer candidato en el ranking de clases de CAT para ese documento.
- **Top 10.** Proporción de casos donde la clase principal del documento de test está en los 10 primeros candidatos del ranking de clases de CAT.
- **Recall 15.** Nivel de *recall* en los 15 primeros candidatos, esto es la proporción de todas las clases correctas para un documento de test que aparecen en los 15 primeros candidatos.
- **Recall 20.** Nivel de *recall* en los 20 primeros candidatos. Proporción de todas

las clases correctas para un documento de test que aparecen en los 20 primeros candidatos.

***Mean Average Precision* (MAP)**

Mean Average Precision (MAP), calcula la precisión cada vez que se agrega un TP, resultando ser el promedio de las precisiones medias calculadas para cada una de los documentos de la colección de test. Si Q es la colección de test, y si las clases relevantes de un documento $q_j \in Q$ es $\{c_1, c_2, \cdots, c_{m_j}\}$ y R_{jk} es el conjunto de ranking de resultados desde el mejor puesto hasta llegar a la clase c_k, entonces

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\pi(R_{jk}) \tag{2.45}$$

2.6. Comparación de métodos de clasificación.

Cuando precisamos comparar dos o más métodos de clasificación, es necesario definir un entorno de trabajo apropiado para obtener unas conclusiones honestas, ciertas y tangibles. Desafortunadamente, esto no siempre se traduce en unos resultados fiables y comparables, en el sentido de que muchos de estos experimentos se han llevado a cabo en condiciones ligeramente diferentes. En general, los diferentes conjuntos de experimentos de clasificación se pueden analizar en función de si cumplen las siguientes condiciones [Sebastiani, 2002]:

- Trabajan exactamente con la misma colección (es decir, los mismos documentos y las mismas categorías).
- Trabajan con la misma división, en la colección de entrenamiento y la test.
- Trabajan con la misma medida de la evaluación y, siempre que esta medida depende de algunos parámetros, con los mismos valores de esos parámetros.

Las comparaciones son fiables cuando se basan en experimentos llevados a cabo por el mismo autor y bajo unas condiciones cuidadosamente controladas. En cambio las comparaciones son más problemáticas cuando se refieren a diferentes experimentos realizados por diferentes autores. En este caso, las diversas condiciones de fondo, a menudo ajenas al algoritmo de clasificación en sí, pueden influir en los resultados. Nos referimos, entre otros, a las diferentes opciones de pre-procesamiento (*stemming*,

stopwords, *etc*), la indexación, la reducción de dimensionalidad, los valores de los parámetros del clasificador, y a las diferentes normas de honestidad en el cumplimiento de la práctica científica (tales como los parámetros de ajuste en la colección de test cuando son separados para realizar la validación), que a menudo no se describen en los artículos publicados.

Como consecuencia de lo anterior, hay dos métodos diferentes que se pueden aplicar para comparar clasificadores [Yang, 1999]:

1. Comparación directa: los clasificadores $\emptyset_1$ y $\emptyset_2$ se puede comparar cuando se han probado en la misma colección, por lo general por los mismos investigadores y con las mismas condiciones de fondo. Este es el método más fiable.
2. Comparación indirecta: los clasificadores $\emptyset_1$ y $\emptyset_2$ se puede comparar cuando
   a) han sido probados en las colecciones Ω_1 y Ω_2, respectivamente, por lo general por diversos investigadores y posiblemente con las condiciones de fondo diferentes;
   b) uno o más clasificadores de referencia, $\overline{\emptyset}_1, \overline{\emptyset}_2, \cdots, \overline{\emptyset}_m$, han sido probados en ambas colecciones Ω_1 y Ω_2 por el método de comparación directa.

Este método de comparación indirecta es menos fidedigno, siendo imprescindible para ambos métodos disponer de colecciones estándar con las cuales podamos realizar los experimentos y evaluar los resultados.

2.6.1. Colecciones

Para realizar balance de los resultados con los diferentes métodos necesitamos que existan colecciones a disposición de los investigadores para que estos puedan comparar su efectividad. Existe una gran cantidad de colecciones que ejercen de estándares. De las cuales destacamos la colección Reuters-21578 que es un referente en CAT, Oshumed que es un corpus específico y estándar para la biomedicina y una colección reciente CCHMC que está clasificada por códigos CIE-9-MC.

2.6.1.1. La colección Reuters

La colección documental Reuters está compuesta por noticias reales que aparecieron en noticias de la agencia Reuters durante 1987. Los documentos fueron recopilados y categorizados manualmente por personal de la agencia y de la compañía Carnegie

Group, Inc, en 1987. En 1990, la agencia entregó los documentos al Laboratorio de Recuperación de Información de la Universidad de Massachussets.

La colección se distribuyó bajo la denominación Reuters 22173 desde 1991 hasta 1996. En ese año durante la conferencia SIGIR (Special Interest Group on Information Retrival) un grupo de investigadores realizó un trabajo sobre esta colección con el objetivo de poder comparar los resultados de las investigaciones que utilicen la colección. De aquí surge una distribución de 21578 documentos, que es la que actualmente se utiliza en muchos trabajos de CAT con el fin de asegurar una colección de prueba uniforme.

La colección se compone de 21578 documentos (cantidad que le da nombre a la misma), con aproximadamente 27.000 términos, distribuidos en 22 archivos y está disponible en [Lewis, 1997]. Esta colección tiene un total de 135 clases asociadas.

Esta colección de documentos Reuters-21578 [Lewis, 1997] se ha convertido en un estándar de facto dentro del dominio de la CAT, y ha sido la más utilizada en la evaluación de sistemas de clasificación de textos.

En esta colección hay varias particiones, pero la más utilizada en entornos de investigación es la que se denomina “ModApte”. Está formada por 9.603 documentos en la colección de entrenamiento y 3.299 documentos en la colección de test. De las 135 clases de toda la colección, sólo 90 clases están representadas tanto en los documentos de entrenamiento, como en los de test.

2.6.1.2. Colección Oshumed

En las conferencias TREC (Text REtrieval Conference) se utilizaban colecciones de prueba del dominio médico para la evaluación de los sistemas de RI, pero hasta el año 2000 no se creó un corpus específico para biomedicina. En concreto, se midió la capacidad de distintos sistemas para clasificar los documentos de OhsuMed (Oregon Health Sciences University).

Se trata de un subconjunto de MEDLINE con orientación clínica, que consta de 348.566 referencias (de un total de más de 7 millones), y que cubre todas las referencias de 270 revistas médicas en un período de cinco años (1987-1991). La colección se define y está accesible en [Hersh, 1994], con aproximadamente 38.000 términos y tiene unos 400 megabytes de tamaño. La colección tiene una partición que es la que se utiliza habitualmente en los experimentos de 10.000 documentos de entrenamiento y 10.000 documentos de test, con 23 clases. Cada documento tiene una

serie de campos asociados. Los campos que la forman son: título, resumen, términos MeSH de indexación, autor, fuente y tipo de publicación.

2.6.1.3. Colección CCHMC

Esta colección de 978 documentos ha sido preparada por The Computational Medicine Center [CMC, 2007] para el desafío internacional: Classifying Clinical Free Text Using Natural Language Processing. El corpus incluye registros médicos anónimos recopilados en el Departamento de Radiología del Hospital Infantil de Cincinnati (the Cincinnati Children's Hospital Medical Center's Department of Radiology – CCHMC).

Estos documentos son informes radiológicos que están etiquetados con códigos CIE-9-MC. Cada documento contiene dos campos de texto a partir del cual se ha construido la colección: CLINICAL_HISTORY e IMPRESSION. Ambos campos son, por lo general, muy breves.

La colección se encuentra clasificada manualmente por tres expertos. En cada documento existen tres conjuntos de codificaciones, una por cada uno de los expertos. A la colección se añade una nueva codificación que representa los códigos que han sido elegidos mayoritariamente por los expertos, y que se han etiquetado con <code origin="CMC_MAJORITY" type="ICD-9-CM">. Esto indica la enorme dificultad de este tipo de clasificación en donde diferentes codificadores expertos manuales no coinciden en bastantes ocasiones en sus criterios de clasificación. El número de códigos distintos que utiliza la colección es de 142.

Esta colección no nos sirve para realizar pruebas en clasificación automática de informes de alta, ya que los informes de alta de hospitalización y los informes radiológicos son bastantes diferentes en su finalidad, su contenido y en la estructura del mismo. Por esta razón fue necesario crear una colección específica de informes de alta hospitalaria para poder realizar los experimentos.

Los trabajos de [Joachmis, 1998], [Aas K. and Eikvil L., 1999], [Yang, 1999] y [Sebastiani, 2002] aplican los métodos de comparación definidos en este punto, con algunas de las colecciones descritas anteriormente. En la mayoría de los casos se obtienen los mejores resultados con las técnicas de clasificación SVM y *Knn*. Teniendo en cuenta esto, consideramos que estos dos sistemas de clasificación sean los referentes de los experimentos de esta tesis.

Capítulo 3

Clasificación de códigos CIE-9-MC con algoritmos de vecindad y Máquinas de Soporte Vectorial

Este capítulo plantea la resolución de un problema de clasificación automática de textos en un dominio médico. El proceso consiste en una clasificación de textos en donde las clases son los códigos CIE-9-MC y los documentos son los informes de alta hospitalaria. Los sistemas de clasificación que se utilizan para la asignación de códigos CIE-9-MC a un documento nuevo son algoritmos de vecindad y Máquinas de Soporte Vectorial. Uno de los valores añadidos de este trabajo es la construcción de la colección, a partir de los informes de alta de un servicio médico. Esta es una colección difícil por la gran cantidad de clases, el número de clase por documento y la descompensación entre las clases. Se estudian diferentes representaciones de la colección, distintos modelos de recuperación y el efecto de los sistemas de pesado en la asignación de códigos CIE-9-MC. La expansión de documentos es especialmente original, al ampliar la colección de entrenamiento con las descripciones de los códigos CIE-9-MC asignados.

3.1. Creación y análisis de la colección

El primer problema que se encuentra al investigar en clasificación de informes de alta hospitalaria con códigos CIE-9-MC es que no existen colecciones públicas. Por ello, una tarea prioritaria es la construcción de la colección. Se efectúa un estudio previo en los servicios que elaboran informes de alta mediante documentos informatizados. De este análisis se selecciona los informes de alta del servicio de Medicina Interna del Hospital de Conxo, que es uno de los hospitales del Complejo Hospitalario

Universitario de Santiago de Compostela (España). Los motivos de esta selección son el elevado número de documentos disponibles, el tamaño de los documentos, el modelo uniforme de documentos, y la complejidad de los diagnósticos utilizados por este servicio.

Estos informes de alta se elaboran mediante una plantilla, que determina el esquema del documento, como mostramos en el siguiente ejemplo:

MOTIVO DE INGRESO: dolor en hemitorax izquierdo, disnea y síndrome confusional.

ANTECEDENTES PERSONALES:

Hepatopatía crónica de probable origen etílico. Diabetes Mellitus insulino-dependiente. Intervenido de hernia inguinal. Ulcus gástrico. Hernia de hiato. Celulitis en MMII en marzo/03. No alergias medicamentosas conocidas.

A tratamiento domiciliario con: Duphalac, Hidroxil B12,B6, B1, Aldactone 100, Seguril, Parizac, Insulina Humaplus, Cozaar e Idalpren y Besitran.

HISTORIA ACTUAL:

Paciente que refiere un cuadro de un mes y medio de evolución de dolor en región costal izquierda irradiado hacia región escapular y bazo izquierdo. El dolor en mas intenso al inspirar y con los movimientos respiratorios. Presenta además tos no productiva. El paciente refiere que acudió en 3 ocasiones al Servicio de Urgencias siendo tratado con Adolonta gotas y Actira.

EXPLORACION FISICA: P: 135. TA: 126/76. Tº: 36. Consciente, orientado y colaborador. Sobrepeso. Cabeza y cuello: no IVY a 45.
AC: arrítmica, taquicardica. AP: roncus dispersos mas intensos en campos izquierdo. Abdomen: sin alteraciones. EEII: edema de estasis, ulcera diabética en pie izquierdo, pulsos pedios positivos.

EXPLORACIONES COMPLEMENTARIAS:

GASOMETRIA AL INGRESO: pH 7,44, pCO2 35,9, pO2 81,1, CO3H 24,2, SAT: 96,4%.

ANALITICA AL INGRESO: hemograma: leucos 9,13 (84%N, 9,8%L), Hb 9,9, Hto 27,8, plaquetas 125.000. Coagulación: TP: 75%. Bioquímica: glucosa 504, creatinina 1,2, amilsa 15, sodio 128, bilirrubina total 1,2, GGT 81, troponina I 0,46.

ANALITICA DE CONTROL: hemograma: leucos 4,49, (57%N), Hb 8,7, Hto 23,8, plaquetas 107.000. Coagulación: TP: 70%. APTT:29,2.
Bioquímica: glucosa 125, urea 86, potasio 5,2.

PROTEINOGRAMA: proteínas totales 5,7, albumina 41%, alfa1 9%, alfa2 14%, beta 12%, gamma 24%.

INMUNOGLOBULINAS: IgG 1100, IgA 255,IgM 163.

H. TIROIDEAS: TSH: 2,73,T4 LIBRE: 0,78, T3 LIBRE: 2,34.

Hb A1c: 11,3.

ESTUDIO DEL HIERRO: hierro 45, transferrina 127, ferritina 232.

MARCADORES T: ALFAFETOPROTEINA: 1,6, CEA-II: 1,4, PSA TOTAL: 0,186.

FROTIS DE SANGRE PERIFERICA: morfología plaquetar normal. Moderada desviación izquierda. 1% mielocitos, 2% cayados, normocitica, normocromica.

LIQUIDO PLEURAL: leucos 1020 (13%N,30%L, 25%Mac. 32%Mes), pH 7,4, proteínas liquido: 2,3, glucosa 112, triglicéridos 14, colesterol 41, amilasa 17, albumina 1,1, LDH 135.

ADA LIQUIDO:14, CEA LIQUIDO: 0,6.

A. PATOLOGICA DE LIQUIDO PLEURAL: citología negativa para malignidad. Reacción mesotelial.

ORINA: hematíes indicios.Sedimento: 1-5 hematíes/campo.

UROCULTIVOS: negativos.

SEROLOGIA DE LEGIONELLA: negativo.

ECG (al ingreso): FA a 131 lpm. ECG: de control, ritmo sinusal a 74 lpm.

RX TORAX: condensación hiliar izquierda espiculada con una pequeña banda de condensación parenquimatosa a nivel del LII que podría corresponder a una pequeña zona de neumonitis obstructiva.
Pinzamiento del seno costodiafragmatico izquierdo.

RX ABDOMEN: calcificación de los conductos deferentes.

ECOGRAFIA ABDOMINAL: derrame pleural izquierdo. Imagen hiperecogenica sobre el LDH con sombra sonica posterior de aproximadamente 1,5 cm en probable relación con granuloma calcificado. Porta de 1,6 cm. Moderada esplenomegalia de 18x 11,5 cm. Ambos riñones de tamaño y morfología normal. Páncreas sin alteraciones.

TAC TORACO-ABDOMINAL: moderado derrame pleural izquierdo, lesión con aspecto de masa en LSI con un amplio contacto con la pleura parietal, la lesión presenta un amplio componente necrótico con nivel hidro-aereo, dicha lesión presenta un aspecto de masa mas que de absceso pulmonar. No hay evidencia de crecimientos adenopaticos mediastinicos. El estudio abdominal, suprarrenales normales. Agrandamiento del lobulo caudado y moderada esplenomegalia compatible con patología alcohólica.

BIOPSIA GUIADA POR TAC:(informe verbal);cambios inflamatorios. Áreas de bono. Negativa para malignidad.

ESPUTOS: positivo para streptococcus pneumoniae.

BAAR: negativo

INTERCONSULTA AL S DE OFTALMOLOGIA: retinopatia diabetica no proliferante leve en ambos ojos. En ojo izquierdo nevus temporal a papila.

INTERCONSULTA AL S DE DERMATOLOGIA: ezcema de extasis.

EVOLUCION Y COMENTARIOS:
Paciente que ingresa en M. Interna por cuadro de dolor torácico en hemitorax izquierdo con hallazgos en la radiografía al ingreso de una masa en hemitorax zquierdo. El cultivo de esputo resulto positivo para neumococo sensible a betalactamicos. Se realizo TAC torácico demostrando una masa pulmonar cavitada en contacto con la plerual parietal y con nivel hidroaereo sin evidencia de adenopatías y derrame pleural izquierdo. Se realizo toracocentesis con citología negativa para malignidad. Con las hipótesis de neoplasia pulmonar sobreinfectada / absceso pulmonar se inicio tratamiento con cefotaxina y clindamicina con buena respuesta clínica y desaparición de la expectoración purulenta . Se realizo TAC torácico de control demostrando gran disminución del tamaño de la lesión pulmonar y desaparición del nivel hidroaereo. Se realizo PAAF de dicha lesión guiada por TAC con hallazgos compatibles con cambios inflamatorios y descartándose malignidad, por lo que consideramos que se trata de un absceso pulmonar. En el momento actual el paciente se encuentra estable y asintomático por lo que se procede al alta. Se continuara seguimiento en consultas hasta resolución de masa pulmonar mediante TAC torácico.

DIAGNOSTICOS:

1. MASA PULMONAR CAVITADA (PROBABLE ABSCESO PULMONAR).

2. DERRAME PLEURAL IZQUIERDO SECUNDARIO A 1).

3. HEPATOPATIA CRONICA ETILICA.

4. DIABETES MELLITUS TIPO 1.

5. HEPATOPATIA CRONICA DE ORIGEN ETILICO.

6. EZCEMA DE EXTASIS.

7. RETINOPATIA DIABETICA NO PROLIFERATIVA.

TRATAMIENTO:

1. Dieta de Diabetes Mellitus de 2500 calorías sin sal.

2. ZINNAT 500: 1 cp con desayuno y cena durante 15 días y suspender.

3. DALACIN 300: 1 cp con desayuno, comida y cena.

4. INSULINA HUMAPLUS: 34 unidades antes del desayuno y 24 unidades antes de la cena.

5. PEITEL CREMA: 1 aplicación en ambas piernas diaria.

6. DACNOLUX COLIRIO: 1 aplicación cada 6 horas en ambos ojos.

7. PARIZAC: 1 cp diario.

8. ALDACTONE 100, SEGURIL, BESITRAN, COZAAR, HIDROXIL, IDALPREN: como venia realizando.
Glucemias capilares antes de cada comida y ACTRAPID: según pauta adjunta.

Se construye la colección a partir de los informes de alta en un formato estándar de la *Text REtrieval Conference* (TREC), con la siguiente estructura:

<DOC>
<DOCNO>*document_number*</DOCNO>
<TEXT>
.
. Informe de Alta
.
</TEXT>
</DOC>

La colección final se forma con los informes de alta de enero 2003 a mayo 2005, un total de 1823 documentos. Aleatoriamente la colección se divide en dos partes: 1501 documentos de entrenamiento y 322 documentos de test. Hay 1238 clases diferentes en la colección de entrenamiento y 544 clases diferentes en la colección de test. De todas las clases, 71 están presentes en el conjunto de test pero no aparecen en el conjunto de entrenamiento. Estas clases están en 74 documentos, que no se eliminan de la colección de test. En una clasificación multiclase, estos documentos pertenecen a otras clases que si están representadas en la colección de entrenamiento. Este es un problema real, y en este tipo de entornos, es muy difícil obtener un conjunto de

entrenamiento con los más de 21.000 códigos distintos que posee CIE-9-MC. Los datos más importantes de la colección, que se denomina MIR–Conxo, figuran en la tabla 3.1.

	Entrenamiento	Test
Número de documentos	1501	322
Tamaño	5963 Kb	1255 Kb
Media número de códigos por documento	7.06	7.05
Máximo número de códigos por documento	23	19
Media número de términos por documento	519.5	508.1
Máximo número de términos en documento	1386	1419
Mínimo número de términos en documento	64	109

Tabla 3.1: Propiedades de la colección MIR-Conxo

Es una colección desbalanceada ya que un 20% de los códigos representan aproximadamente un 80% de los diagnósticos codificados.

Para valorar los resultados obtenidos con otros trabajos de CAT con códigos CIE-9-MC, es necesario conocer las características de las colecciones con la que se han realizado los experimentos. Al intentar comparar las características de la colección MIR–Conxo con otras colecciones de informes de alta de hospitalización, descubrimos que no existen prácticamente colecciones públicas. Los pocos experimentos realizados de CAT con códigos CIE-9-MC son con colecciones propias creadas para desarrollar los mismos.

La única colección pública disponible en la actualidad, es la colección CCHMC de informes radiológicos [Pestian et al., 2007]. La colección CCHMC se forma a partir de una colección de 20.275 documentos, en donde se incluyeron solo aquellas clases que están representadas en 100 o más documentos. Se selecciona un subconjunto por un muestreo, de forma que contenga el 20% de los documentos de cada clase. Con una selección manual se eliminan alrededor de un 50% de los documentos, para mantener el anonimato de los textos. En la colección de test se elimina los documentos que pertenezcan a una clase que no esté representada en la colección de entrenamiento. Al final obtenemos una colección de 1.954 documentos repartidos en 978 en la colección

de entrenamiento y 976 en la colección de test, con 45 clases/códigos CIE-9-MC. Aunque los documentos que la forman surgen de una preselección que favorece la clasificación, son de un área hospitalaria distinta, y contienen diferente información (no son informes de alta hospitalaria), al final se utiliza para clasificar códigos CIE-9-MC.

En la literatura hay muy pocos trabajos de CAT para codificar informes de alta en donde que se pueda extraer las características de la colección que utilizan. Uno de los pocos disponibles es el realizado por [Larkey y Croft, 1995]. Se contrasta las características de las colecciones MIR–Conxo, Larkey–Croft y CCHMC y se muestran en la tabla 3.2.

	MIR-Conxo	Larkey-Croft	CCHMC
Media códigos por doc	7.05	4.43	1.9
% docs con menos de 9 códigos	63,7	90	93
Número de documentos	1823	11293	1954

Tabla 3.2: Características de la colecciones MIR-Conxo, Larkey-Croft y CCHMC

Las colecciones tienen características muy diferentes, siendo la colección MIR – Conxo la de mayor complejidad. Por ejemplo, el número de clases de la colección, MIR–Conxo es de 1238, en cambio CCHMC solo son 45. La figura 3.1 representa el número de documentos por código para la colecciones MIR–Conxo y CCHMC, y refuerza el criterio de una mayor dificultad para realizar CAT con la colección MIR–Conxo. Las razones que hacen que nuestra colección sea tan desigual es que está formada por episodios clínicos con múltiples patologías. Las conclusiones que se deducen de estos datos son que los resultados obtenidos no pueden ser comparables por las diferentes características que la forman. Esta situación se acrecienta con la colección CCHMC, dada el origen de su información y el criterio de selección de los documentos.

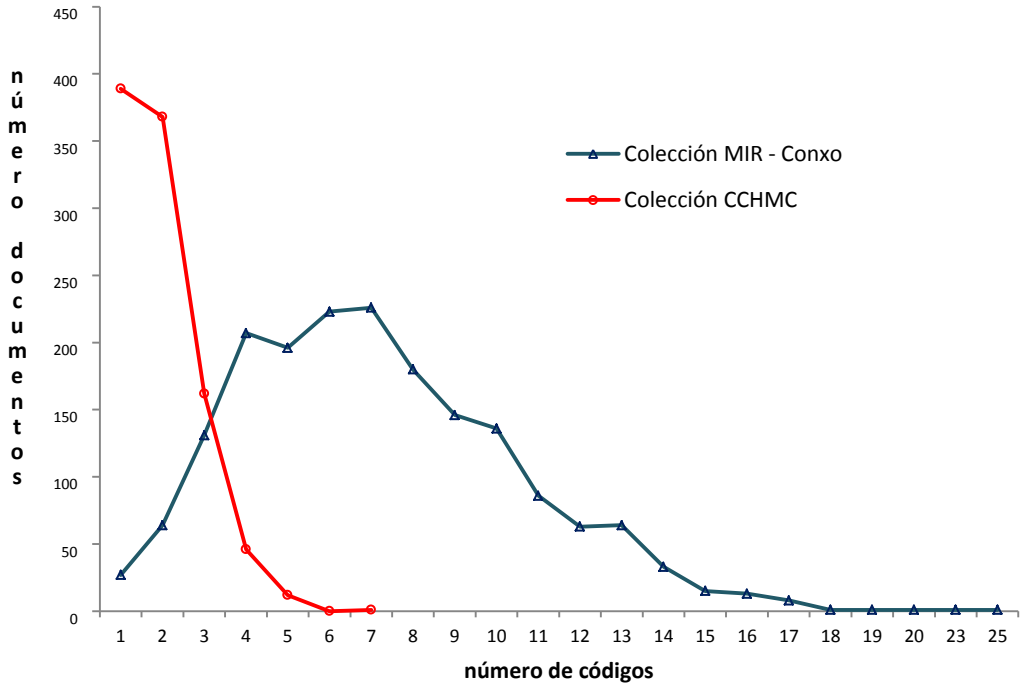

Figura 3.1: Distribución de número de documentos por número de códigos asignados para la colección MIR– Conxo y la colección CCHMC

En la actualidad con la llegada de la historia clínica electrónica a nuestro entorno es más fácil que aparezcan colecciones públicas anónimas. En el momento de desarrollar esta tesis no existía una colección pública CIE-9-MC de informes de alta como la utilizada en los experimentos.

3.2. Clasificación de textos basada en *Knn*

Las técnicas de clasificación y aprendizaje basadas en criterios de vecindad constituyen uno de los ejemplos más representativos en CAT. Dada su simplicidad conceptual, intuitiva, su fácil implementación y aplicación, lo convierte en un instrumento de gran popularidad. El clasificador únicamente requiere la definición de una métrica de similitud (o distancia) entre los distintos espacios de representación de los documentos.

El método de los *k* vecinos más cercanos (*knn*) [Dasarathy, 1991] se puede resumir en que la clase asignada a un documento de test, d_{new}, será la clase más votada entre los *k* vecinos más próximos (según la métrica de similitud) del conjunto de entrenamiento. Esta situación la representamos de forma gráfica en la figura 3.2. En ella, tenemos 16 documentos que pertenecen a dos clases distintas: la clase 1 está formada por 7

círculos y la clase 2 por 9 triángulos. Se simboliza un documento nuevo a clasificar, d_{new}, con una estrella. En este ejemplo, se selecciona los cinco vecinos más cercanos, se delimita su área de influencia por el círculo representado en la figura. De los 5 vecinos más cercanos a d_{new}, dos de ellos pertenecen a la clase 1 y tres a la clase 2. Por lo tanto, 5nn asignará el documento d_{new}, a la clase 2.

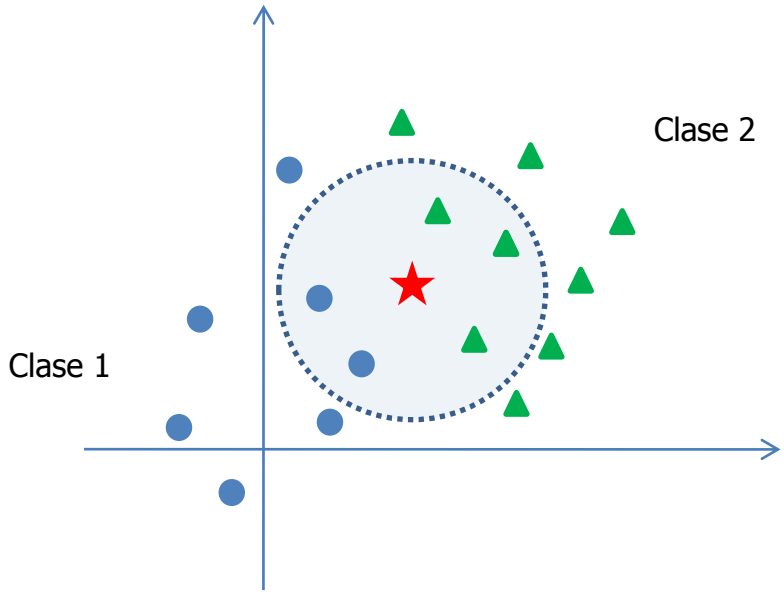

Figura 3.2: Ejemplo de un clasificador *knn*

Está situación se complica cuando se trabaja en un clasificación de textos multi-etiqueta, como es nuestro caso. En donde un mismo informe de alta tiene asignado varios códigos CIE-9-MC. Se asigna al documento d_{new} los códigos CIE-9-MC a partir de los documentos más cercanos recuperados en base a algún tipo de combinación de las puntuaciones de las múltiples clases involucradas.

3.3. Clasificación de textos con SVM

Las máquinas de soporte vectorial han demostrado en los últimos años una gran efectividad en la clasificación automática de textos [Yang and Liu, 1999] y en otros muchos problemas de aprendizaje. SVM parte de un problema de clasificación binaria (es decir, dos clases). El modelo de SVM permite definir un clasificador lineal basado en un hiperplano que actúa como frontera entre las dos clases. Los documentos, en nuestro caso particular, se representan con un modelo de espacio vectorial. Supongamos que los documentos de cada clase se pueden separar en este espacio de representación. SVM busca un hiperplano que separe a las clases y, entre las alternativas, el hiperplano de margen máximo de separación entre clases.

3.3.1. Aplicación al dominio clínico

Nuestro problema de asignación de códigos CIE-9-MC es inherentemente multi-etiqueta, pero SVM originalmente se ha diseñado para clasificación binaria. Por la naturaleza binaria de SVM, surge la necesidad de implementar métodos que puedan resolver los problemas de clasificación multi-etiqueta. Con este objetivo, se han propuesto en la literatura diferentes aproximaciones. Algunas intentan resolver el problema de forma directa [Weston and Watkins, 1998], proponen una modificación de la función de optimización que define el hiperplano óptimo y que tenga en cuenta todas las clases. Por otro lado, se han desarrollado diversas técnicas para la solución de SVM multiclase, a partir de la combinación de clasificadores binarios [Hsu and Lin, 2002]. Cuando el conjunto de entrenamiento tiene más de dos clases existen fundamentalmente dos alternativas parar resolver SVM.

- (*1-vs-todos*): se construyen *c* clasificadores *1-vs-todos* que separan los documentos de cada clase de los restantes. Se opta por la clase que consigue el hiperplano con mayor margen al clasificar un documento de test.

- (*1-vs-1*): se construyen $\frac{c \cdot (c-1)}{2}$ clasificadores *1-vs-1*, uno por cada par de clases posibles. A un documento de test se le aplica todos estos clasificadores, y se computa un voto a la clase ganadora para cada caso. Finalmente, aquella clase que obtenga más votos será la clase propuesta por el clasificador.

Analizando cada uno de los modelos encontramos los siguientes inconvenientes. Recordemos que un documento pertenece en nuestro problema a varias clases. En un clasificador *1-vs-todos*, surge el problema de que se clasifica una clase contra el resto de clases, en donde algunas de ellas también pertenecen a ese documento. Además, el conjunto de entrenamiento va a estar muy desbalanceado. El clasificador *1-vs-1*, se elabora con los datos extraídos de dos clases del conjunto de entrenamiento y no proporciona información del resto de clases, se entrena con un subconjunto de la colección, lo que nos puede suponer una preocupante pérdida de información. Además, con este modelo el número de clasificadores que tenemos que realizar es alto, y más aún si el número de clases es elevado como en nuestro caso.

Estas técnicas pueden crear una zona ambigua para el clasificador. Para apreciar esta zona representamos varios ejemplos en una dimensión $\mathbb{R}^2$. En la figura 3.3 mostramos una colección de cuatro clases con los hiperplanos de separación para un clasificador *1-vs-todos*. Con este tipo de clasificador surge una zona ambigua de ejemplos que no pertenecen a ninguna de las clases.

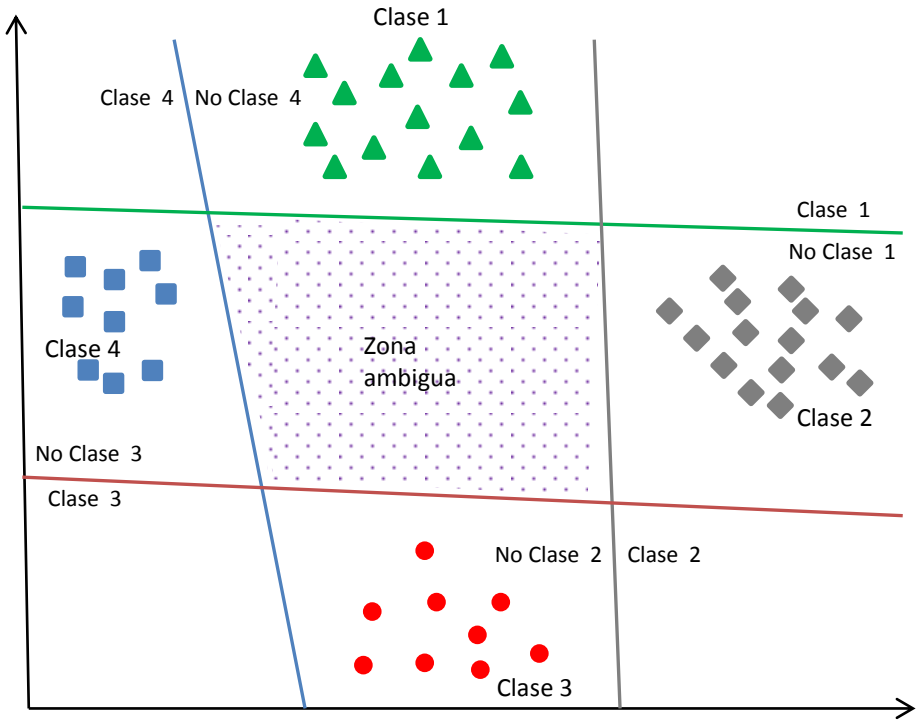

Figura 3.3: Zona ambigua en un clasificador *1-vs-todos*

En la figura 3.4 representamos la zona ambigua para un clasificador *1-vs-1* en una colección de tres clases. Obsérvese que en este ejemplo la ambigüedad se produce porque esa zona está incluida en todas las clases.

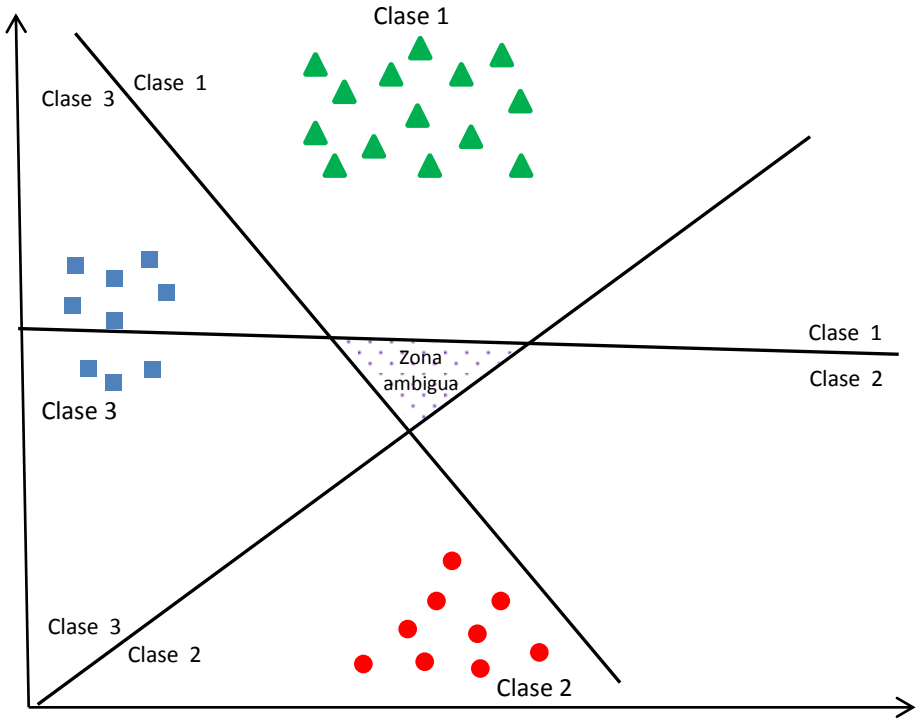

Figura 3.4: Zona ambigua en un clasificador *1-vs-1*

La clasificación de un documento que esté situado en estas zonas no se puede definir de forma clara la pertenencia o no a una clase.

La técnica *1-vs-todos* es la seleccionada para realizar los experimentos por las siguientes razones: la colección tiene muchas clases y con *1-vs-todos* construimos un número razonable de clasificadores. Además, en CAT es la técnica preferida y es recomendada por Vapnik en sus trabajos.

Estas técnicas están diseñadas para asignar una clase única a los documentos de test y nuestro problema de CAT es encontrar múltiples clases para un documento de test. Al utilizar *1-vs-todos,* los documentos pertenecen a una clase y a su vez esos mismos documentos tienen otras clases que están dentro de la clase enfrentada. Por estos motivos y los descritos anteriormente, lo más apropiado es utilizar una clasificación en ranking. Donde se construye un ranking de posibles códigos para cada documento de test. Para conseguir este ranking, se utiliza el margen del hiperplano de separación como medida de certeza de que el documento de test es apropiado para esa clase. Con estas medidas se elabora un ranking de clases, en donde las clases se clasifican por orden decreciente del valor del margen entre el documento de test y los hiperplanos de las clases.

3.4. Método de clasificación

El proceso de asignación de códigos CIE-9-MC se realiza en los hospitales de forma manual por médicos codificadores después de la lectura de los informes de alta. CIE-9-MC es un sistema de categorías alfanuméricas que han sido asignadas a las enfermedades de acuerdo con unos criterios internacionales establecidos previamente. Los códigos CIE-9-MC están divididos en dos partes separadas por un punto, a excepción de los códigos M (Morfología de las neoplastias). La parte izquierda del punto se denomina categoría o sección. El símbolo de separación para las categorías de los códigos M está representado por el símbolo "/". Denominamos código CIE-9-MC al código de mayor nivel de especificidad y son códigos válidos aquellos que tienen como mínimo los mismos dígitos que la categoría.

La asignación de códigos CIE-9-MC a un episodio clínico tiene los siguientes elementos importantes. El diagnóstico principal [DxP] es la enfermedad que tras su estudio y en el momento del alta, el médico que atendió al paciente establece como causa del ingreso. Los diagnósticos secundarios [DxS] se consideran aquellas enfermedades que coexisten con el [DxP] en el momento del ingreso o que se han desarrollado durante la estancia hospitalaria y que han influido en la duración del ingreso.

3.4.1. Procedimiento de clasificación *Knn*

El método de clasificación consiste en recuperar primero aquellos *k* documentos ya codificados que son muy similares al documento nuevo a codificar. Asignando a continuación al documento a clasificar los códigos CIE-9-MC de los documentos recuperados. Utilizaremos *Lemur*, un conjunto de herramientas *(toolkit)* de RI en código abierto desarrollado por la Universidad de Massachussets y la Universidad de Carnegie Mellon. Esta herramienta nos permite realizar todas las etapas de un sistema de RI.

Se construye un índice con la colección de entrenamiento y los documentos de test van a actuar como consultas como se muestra en la figura 3.5. En esta figura se representa globalmente el proceso de clasificación.

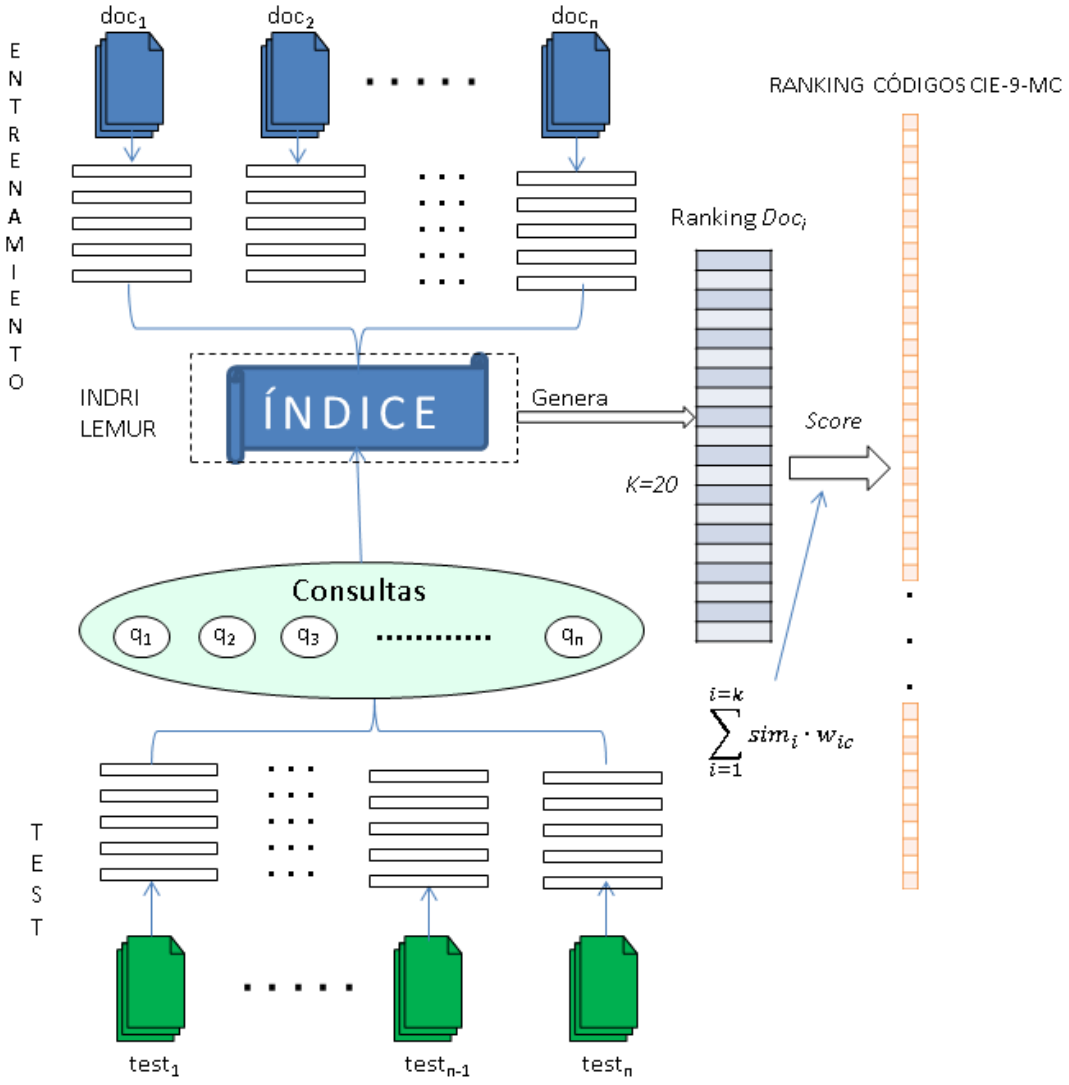

Figura 3.5: Esquema global del clasificador *knn*

Como se observa en la figura 3.5, cada documento de test genera una consulta que interactúa con el sistema RI y nos devuelve un conjunto de documentos con una puntuación de similitud (*score*) con el documento de test. El conjunto de documentos recuperados se ordena en orden decreciente de la puntuación de similitud.

En la tabla 3.3 se presenta el ranking de documentos recuperados para un determinado documento de test, en donde se incluyen los códigos asociados. Se diferencian en este ranking el código del diagnóstico principal de los códigos de los diagnósticos secundarios.

Doc	Rank	Puntuación	[DxP]	[DxS]
51007762	1	-5.60631	787.91	787.01 553.3 438.82 438.11 438.20 272.9 401.9
41000982	2	-5.63082	507.0	491.21 518.84 427.31 438.11 438.20 V55.1
51007699	3	-5.65442	787.2	335.20
31021009	4	-5.67955	290.3	
41010468	5	-5.68983	009.0	285.9 287.5 434.91 584.9 715.35 591.
41034118	6	-5.69316	507.0	290.0 250.00 414.8 402.90 V45.01 553.3
51010532	7	-5.69714	507.0	402.90 427.31 V58.61 332.0 294.8 V12.59 790.93
……	*k*	…….	…….	……

Tabla 3.3: Ranking de documentos para un documento a clasificar

Una cuestión básica es cuantos documentos debemos recuperar (valor de *k*). Aunque algunos estudios [Larkey y Croft, 1995], [Lojo, et al., 2009] sugieren el uso de *k*=20, realizamos experimentos con distintos valores de *k*. Cada código asociado a un documento recuperado se convierte en un candidato para ser asignado al documento de test. Con los códigos de los documentos recuperados, se genera el ranking de códigos CIE-9-MC para el documento de test, como muestra la tabla 3.4.

Se usa la siguiente expresión para calcular la puntuación total de un código:

$Score_c = \sum_{i=1}^{i=k} sim_i \cdot w_{ic}$, en donde w_{ic} es el peso asociado al código c en el documento i y sim_i es la similitud del documento i con el documento a clasificar.

Código CIE-9-MC	Num. Docs. recuperados	Puntuación final $Score_{nc}$	Descripción del código
438.20	7	0.023865	HEMIPLEJIA AFECTANDO UN LADO INESPECIFICADO
507.0	5	0.016901	NEUMONITIS POR INHALACION DE COMIDA/VOMITOS
414.8	5	0.016639	OTRAS FORMAS ESPECIFICADAS ENF.CARDIACA ISQUEMICA CRONICA
402.90	5	0.016639	CARDIOPATIA HIPERTENSIVA SIN ESPECIFICAR. CON FALLO C.CONG.
…….	….	…….	…….

Tabla 3.4: Ranking de códigos para un documento a clasificar

Para el cálculo de la puntuación final de un código se suma la puntuación obtenida para cada documento que tiene asignado a cada código y la multiplicamos por w_{ic}. Dentro de los diferentes sistemas de pesado utilizamos inicialmente el más básico (*baseline*): el peso w_{ic} es 1 si el código está entre los códigos del documento recuperado y 0 en caso contrario. En este caso el resultado final del $Score_c$ es la suma de los sim_i de los documentos recuperados que contengan ese código *c*.

Lemur tiene varios modelos de RI y algunos de ellos, como el mostrado en la tabla 3.3 nos devuelven valores de similitud negativos. Dado que la definición del $Score_c$ está pensada para valores positivos, realizamos la siguiente operación para obtener un valor normalizado y positivo del $Score_c = \sum_{i=1}^{i=k} e^{sim_i} \cdot w_{ic}$.

3.4.2. Procedimiento de clasificación SVM

El proceso de clasificación mediante una Máquina de Soporte Vectorial consta de dos fases: entrenamiento y clasificación. En el primero se reconocen los patrones de la colección de entrenamiento con el fin de crear un modelo que será utilizado en la clasificación de nuevos documentos. La fase de entrenamiento trata de encontrar los vectores soporte que definen el hiperplano óptimo de separación. Estos vectores son los puntos de entrenamiento que no están clasificados con confianza. Por lo tanto, los vectores de soporte son puntos esenciales de la fase de entrenamiento, y su objetivo es descubrirlos.

La implementación de SVM que se utiliza es SVM^{light} [Joachims, 1999], se fundamenta en una propuesta de mejora al algoritmo planteado por Osuna [Osuna et al., 1997]. Consta de dos módulos, uno de aprendizaje (*svm-learn*) y otro de clasificación (*svm-classify*). Al aplicar *1-vs-todos*, implica generar un fichero por clase, esto implica construir tantos clasificadores como clases tiene la colección. El fichero que necesita s*vm-learn* para obtener los vectores de soporte para cada clase, se construye con la representación vectorial de la colección de entrenamiento. De forma similar se aplica el mismo procedimiento para el modulo de clasificación (*svm-classify*). El clasificador nos devuelve su predicción para el documento de test con cada clase/código CIE-9-MC. Con las predicciones de cada clasificador para el documento de test se construye un ranking de códigos.

Dentro de un dominio clínico en una clasificación de textos multi-etiqueta con SVM se puede utilizar varios modelos. Por los motivos expuestos en el punto 3.3.1 de este capítulo se selecciona el modelo *1-vs-todos* para realizar los experimentos.

3.5. Métricas de evaluación

Los parámetros de evaluación que están descritos en esta sección requieren la existencia de estándares de referencia. En una clasificación automática supervisada como es este caso, se dispone de los códigos asignados por los codificadores médicos para cada documento de la colección de entrenamiento y test. Por supuesto, la lista de códigos asociados a los documentos de test solo se maneja para fines de evaluación. El sistema de CAT produce un ranking de códigos (lista de códigos candidatos) para cada documento de test. En estos casos una de las métricas más utilizada es la media de los 11 puntos de precisión (*11-point average precisión*).

Además, en un entorno específico y singular como el que se desarrolla este trabajo, se definen métricas concretas en donde el usuario analiza y valora mejor la función CAT. Para un entorno de clasificación de documentos de alta mediante códigos CIE-9-MC tenemos las siguientes métricas especiales [Larkey and Croft, 1996]: *Top candidato, Top 10, Recall 15* y *Recall 20.*

Estas son las definiciones de las métricas que se calculan en este capítulo:

- **Media de los 11 puntos de precisión (*11-point average precisión*).** Precisión (π) y recall (ρ) son dos medidas estándar en la evaluación de RI. En nuestro caso, precisión es la proporción de los códigos propuestos por el clasificador que son correctos. Recall es la proporción de todos los códigos correctos que han sido propuestos por el clasificador. La media de los 11 puntos de precisión se calcula como la precisión (π) promedio obtenida para cierto valores espaciados unifórmente de recall.
- **Top candidato**. Proporción de casos donde la clase principal del documento de test (esto es la clase del diagnóstico principal) es el primer candidato en el ranking de clases de CAT para ese documento.
- **Top 10.** Proporción de casos donde la clase principal del documento de test está en los 10 primeros candidatos del ranking de clases de CAT.
- **Recall 15.** Nivel de recall en los 15 primeros candidatos, esto es la proporción de todas las clases correctas para un documento de test que aparecen en los 15 primeros candidatos.

- **Recall 20.** Nivel de recall en los 20 primeros candidatos. Proporción de todas las clases correctas para un documento de test que aparecen en los 20 primeros candidatos.

3.6. Representación de los documentos

Una de las primeras fases en los procesos de RI es la representación de los documentos. El preprocesado léxico inicial consiste en convertir el texto de cada documento en un conjunto de palabras o lemas que puedan servir como términos a las otras fases del clasificador. Es importante determinar qué palabras o lemas van a incluirse en ese conjunto. Parece claro que cualquier palabra formada por caracteres alfabéticos puede incluirse como candidato a término de indexación. Un aspecto que debe tenerse en cuenta es el idioma de los documentos: la colección en español contiene caracteres como la eñe, vocales acentuadas, símbolos especiales, etc, que hay que tener en cuenta. Un aspecto previo en esta fase consiste en identificar cada palabra del texto. Parece claro que el separador por excelencia es el espacio y los caracteres de puntuación. La herramienta *Lemur* tiene utilidades para realizar algunas de estas tareas, como identificación de palabras o el tratamiento de los signos de puntuación. *Lemur* soporta una codificación UTF-8.

En general, las vocales acentuadas incluyen una carga semántica importante a la palabra, pero en RI no suelen considerarse; el motivo no es otro que el alto grado de errores ortográficos que se cometen con los acentos. En nuestros experimentos se ha optado por convertir a vocales no acentuadas aquellas que lo están en el texto.

En los sistemas de RI lo normal es que los términos de indexación se conviertan a minúsculas. En nuestro caso, se han transformado los textos originales de la colección a minúsculas.

Como ya comentamos en la sección 2.3.4.1.1 del capítulo 2, para reducir el número de términos del índice no se incluyen palabras que, por su poca capacidad semántica o por su alta frecuencia, son poco significativas en el proceso de clasificación. Este conjunto de palabras, que se denomina conjunto de palabras vacías (*stop words*), se compone de preposiciones, artículos, adverbios, conjunciones, posesivos, demostrativos, pronombres y algunos verbos muy comunes (Anexo C).

El lenguaje médico, al igual que todo lenguaje científico, tiene como objetivo referirse con precisión a los conceptos propios de su área de conocimiento. Aunque los lematizadores de textos han demostrado su eficacia para reducir el vocabulario y para

acortar su espacio de almacenamiento, en entornos con muchos términos técnicos aplicar un lematizador es problemático. El algoritmo para lematizar términos médicos es una tarea compleja [Peinado, 2003] y en la actualidad no se dispone de algoritmos adecuados.

En la colección estándar Ohsumed (con términos médicos), se ha demostrado que la funcionalidad de aplicar un lematizador perjudica los resultados obtenidos por el clasificador [Joachims, 2002]. Por estas razones no se utiliza la extracción de raíces (stemming) en los experimentos que se van a desarrollar en esta tesis.

Se puede resumir que el preprocesado inicial de la colección de entrenamiento y test implica modificar las vocales acentuadas por vocales sin acentuar, cambiar las letras mayúsculas por minúsculas, eliminar las palabras vacías y no utilizar stemming.

Los documentos de la colección se construyen con diferentes representaciones, que a su vez dan lugar a tres representaciones de la colección que se denominan:

- *Diagnósticos:* representación de los documentos en donde solo va a estar presente la sección del informe de alta en donde el especialista de Medicina Interna redacta los diagnósticos del paciente (esto es, el texto explicativo del informe se descarta y sólo se considera el listado de diagnósticos formulados por el médico). Ver informe de alta que se incluye en este capítulo.
- *Total:* se considera todo el documento de alta.
- *Total + CIE-9-MC:* está formada por todo el documento de alta, y se añaden las descripciones de los diagnósticos de los códigos *CIE-9-MC* codificados por el médico codificador. Esta tabla 3.5, a modo de ejemplo, nos muestra la estructura de un código CIE-9-MC real, con sus descripciones, que son las que se agregan en una representación *Total + CIE-9-MC*. En definitiva, en esta última representación se realiza una expansión de los documentos con las descripciones de los códigos *CIE-9-MC*.

Hay que destacar que la estructura de la información de los documentos de entrenamiento y test en las representaciones *Diagnósticos* y *Total* es la misma. En cambio, en la representación *Total + CIE-9-MC*, los documentos de entrenamiento incorporan las descripciones de los códigos de diagnósticos al texto del documento. En cambio, los documentos de test no pueden usar ninguna información de los códigos (o descripciones) asignados por los médicos codificadores.

- 534 **Ulcera gastroyeyunal**

  Incluye:
  - úlcera (péptica) o erosión:
  - - anastomótica
  - - estomacal
  - - gastrocólica
  - - gastrointestinal
  - - gastroyeyunal
  - - marginal
  - - yeyunal

  Excluye:
  úlcera primaria del intestino delgado (569.82)
  La siguiente subclasificación de quinto dígito debe emplearse con la categoría 534:
  0 Sin mención de obstrucción
  1 Con obstrucción

  - 534.0 Aguda con hemorragia [0,1]
    - 534.00 Sin mención de obstrucción
    - 534.01 Con obstrucción
  - 534.1 Aguda con perforación [0,1]
    - 534.10 Sin mención de obstrucción
    - 534.11 Con obstrucción
  - 534.2 Aguda con hemorragia y perforación [0,1]
    - 534.20 Sin mención de obstrucción
    - 534.21 Con obstrucción
  - 534.3 Aguda sin mención de hemorragia ni perforación [0,1]
    - 534.30 Sin mención de obstrucción
    - 534.31 Con obstrucción
  - 534.4 Crónica o no especificada con hemorragia [0,1]
    - 534.40 Sin mención de obstrucción
    - 534.41 Con obstrucción
  - 534.5 Crónica o no especificada con perforación [0,1]
    - 534.50 Sin mención de obstrucción
    - 534.51 Con obstrucción
  - 534.6 Crónica o no especificada con hemorragia y perforación [0,1]
    - 534.60 Sin mención de obstrucción
    - 534.61 Con obstrucción
  - 534.7 Crónica sin mención de hemorragia ni perforación [0,1]
    - 534.70 Sin mención de obstrucción
    - 534.71 Con obstrucción
  - 534.9 No especificada como aguda ni como crónica, sin mención de hemorragia ni de perforación [0,1]
    - 534.90 Sin mención de obstrucción
    - 534.91 Con obstrucción

Tabla 3.5: Descripciones de la categoría 534 CIE-9-MC

Para valorar el mejor rendimiento se experimentará con diferentes modelos de RI. En SVM se utilizará el modelo vectorial *TFIDF* por los buenos resultados que nos ofrece [Joachims, 2002] y los experimentos se desarrollarán con distintos parámetros y núcleos de SVM.

3.7. Experimentos con *Knn*

En esta sección se muestran los experimentos con *Knn* en los sistemas de CAT para codificación CIE-9-MC. Para la fase de recuperación con *Knn* se ha selecciona el modelo de recuperación usado por defecto por *Indri*[1] y se contrasta con dos variantes del modelo vectorial para RI (*Lemur TFIDF* y *coseno*).

Los resultados están estructurados en función del formato de los códigos CIE-9-MC. Cada código CIE-9-MC se divide en varias partes. Los códigos tienen categorías y cada categoría puede dividirse en subcategorías y cada subcategoría puede hacerlo en subclasificaciones, como mostramos en la tabla 1.1 del capítulo 1. En los experimentos de esta tesis (y en general) se denomina código aquel de mayor nivel de especificidad posible (mayor número de dígitos).

En la tabla 1.2 se definen todos los tipos de códigos, con su nomenclatura y el formato de división entre categorías y códigos. Un código de enfermedad (la mayoría de nuestra colección) tiene el formato *CCC.S[X]*, en donde *CCC* es la categoría, *S* la subcategoría y *X* la subclasificación, como se puede observar en la tabla 3.5. Los experimentos van encaminados a encontrar las posibles categorías y los posibles códigos a asignar a un nuevo documento de alta. Determinar correctamente un código es una clasificación de grano fino, por lo tanto, más difícil.

La validación de los experimentos se realiza con el método *Holdout*. La colección de entrenamiento se crea de una selección aleatoria de la colección de tamaño 2/3. El resto de la colección (1/3) es la colección de test.

La tabla 3.6 muestra los resultados obtenidos con *K=20* para la clasificación de códigos y categorías. Se detalla los resultados para las diferentes representaciones de la colección y en las métricas definidas para estos experimentos. Se utiliza un sistema de pesado básico, en donde w_{ic} va a ser 1, si el si el código de diagnóstico aparece en el documento recuperado y 0 si no aparece. Los resultados los calculamos con *microaveraging*, estos se basan en la suma total de todas de las decisiones individuales de clasificación y *macroaveraging*, donde se obtienen los valores locales para cada clase, obteniendo después la media total.

[1] www.lemurproject.org

Representación	11-pt Avg	Top Candidato	Top 10	Recall 15	Recall 20
Clasificación Códigos					
Diagnósticos	44.0	14.9	58.7	52.6	57.0
Total	43.1	16.1	64.9	52.5	57.7
Total + CIE-9-MC	43.8	17.4	64.3	53.1	58.2
Clasificación Categorías					
Diagnósticos	52.0	21.1	67.0	60.8	67.9
Total	51.2	22.7	74.2	62.4	67.7
Total + CIE-9-MC	51.8	24.5	73.9	62.9	68.2

Tabla 3.6: Rendimiento de los resultados con *microaveraging* (*K=20*, pesado básico, y modelo *Indri*)

Representación	Top Candidato	Top 10	Recall 15	Recall 20
Clasificación Códigos				
Diagnósticos	10,4	28,0	18,1	21,0
Total	9,3	39,3	23,2	28,8
Total + CIE-9-MC	9,7	40,9	24,8	30,6
Clasificación Categorías				
Diagnósticos	11,8	39,7	25,2	28,5
Total	11,5	53,7	29,2	35,7
Total + CIE-9-MC	14,6	52,9	31,5	37,8

Tabla 3.7: Rendimiento de los resultados con *macroaveraging* (*K=20*, pesado básico, y modelo de IR *Indri*)

Dentro de los diversos experimentos realizados en *Knn* destacamos aquellos con *K=10* en la tabla 3.8 y *K*=30 en la tabla 3.9 para la representación *Total*. Los resultados para *K=10* y *K=30* no mejoran los resultados obtenidos para *K= 20*.

Representación	Top Candidato	Top 10	Recall 15	Recall 20
Clasificación Códigos				
Total	15.8	63,3	50.6	54.5
Clasificación Categorías				
Total	21.4	72.4	60.5	64.8

Tabla 3.8: Rendimiento de los resultados con *microaveraging* (*K=10*, pesado básico, y modelo *Indri*)

Representación	Top Can	Top 10	Recall 15	Recall 20
Clasificación Códigos				
Total	14	64,9	53.1	58.5

Tabla 3.9: Rendimiento de los resultados con *microaveraging* (*K=30*, pesado básico, y modelo *Indri*)

Los datos de 11-pt Avg representan una combinación de las medidas de precisión y recall. En algunos casos es importante observar como se va modificando la precisión en función de recall. La curva precisión-recall permite visualizar estos cambios. En las figuras 3.6 y 3.7 se dibujan estos cambios para códigos y categorías para las distintas representaciones de la colección, con *K=20*, pesado básico y modelo *Indri*.

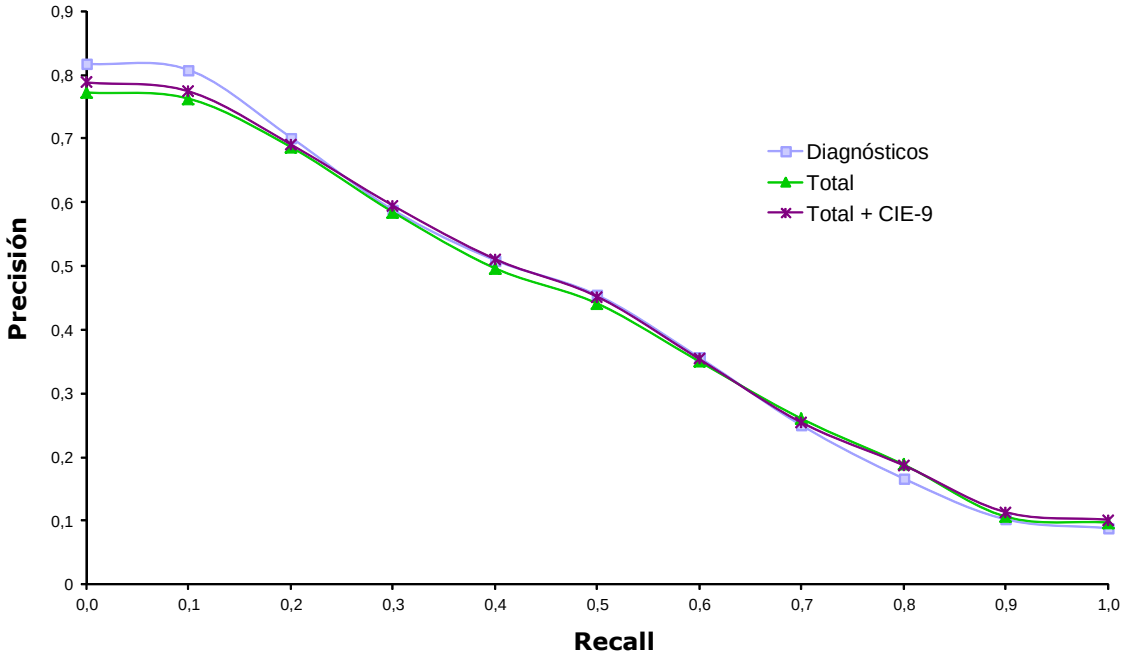

Figura 3.6: Curva Precisión-Recall códigos con *K=20*, pesado básico y modelo *Indri*

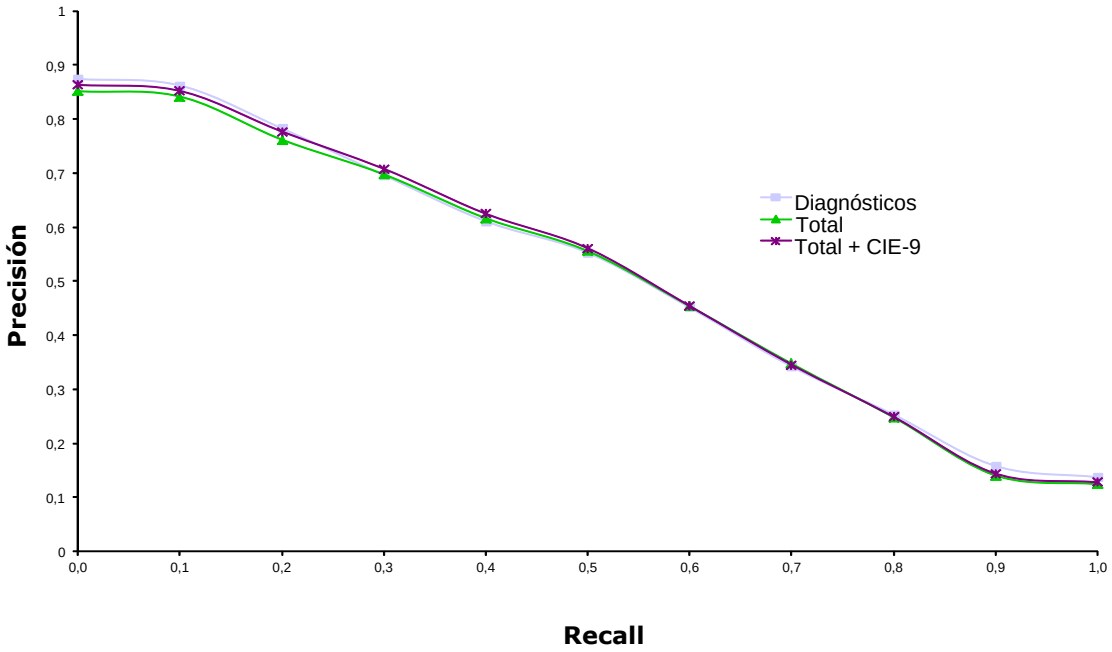

Figura 3.7: Curva Precisión-Recall categorías con *K=20*, pesado básico y modelo *Indri*

En un primer análisis de los resultados nos sorprende las diferencias entre los datos *microaveraging* y *macroaveraging*. Como ya se suponía, suele ser mejor los datos de *microaveraging*, pero no con estas diferencias. Esto nos hace pensar que la clasificación funciona mejor para unos códigos CIE-9-MC que para otros. Con *microaveraging* cada clasificación tiene un voto, en cambio con *macroaveraging* cada clase tiene un valor (calculado con la media de las clasificaciones individuales para esa clase) en el cálculo final de la métrica. Con los valores de las métricas calculadas en los experimentos, se generan histogramas en donde figuran las frecuencias para los diferentes valores de medición. En el eje de las abscisas está el rango de clases (Bin range) para esa métrica, que se ha definido de 0 a 10, de 10 a 20, …, de 90 a 100. El eje de las ordenadas muestra el porcentaje de frecuencia de esa métrica para cada rango de clases definido. Todos los histogramas se realizan para las métricas de los códigos CIE-9-MC que representan las clases. En los histogramas de *Recall 15* y *Recall 20* se representa los valores de las frecuencias de cada rango de clases (CIE-9-MC) y se añaden los porcentajes de frecuencia para cada rango de documentos. Para la representación *Total*, con *K=20*, pesado básico y modelo *Indri* se exponen en el eje x las métricas *Top Candidato*, *Top 10*, *Recall 15* y *Recall 20* en las figuras 3.8, 3.9, 3.10 y 3.11.

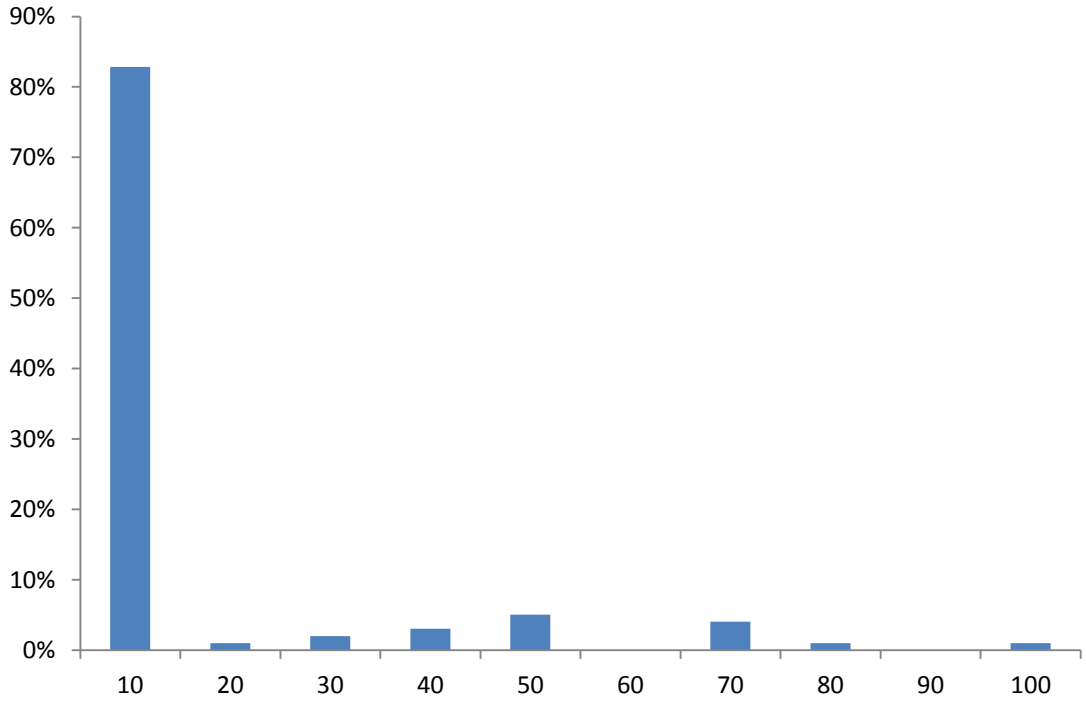

Figura 3.8: Histograma *Top Candidato* para códigos CIE-9-MC en *Knn*

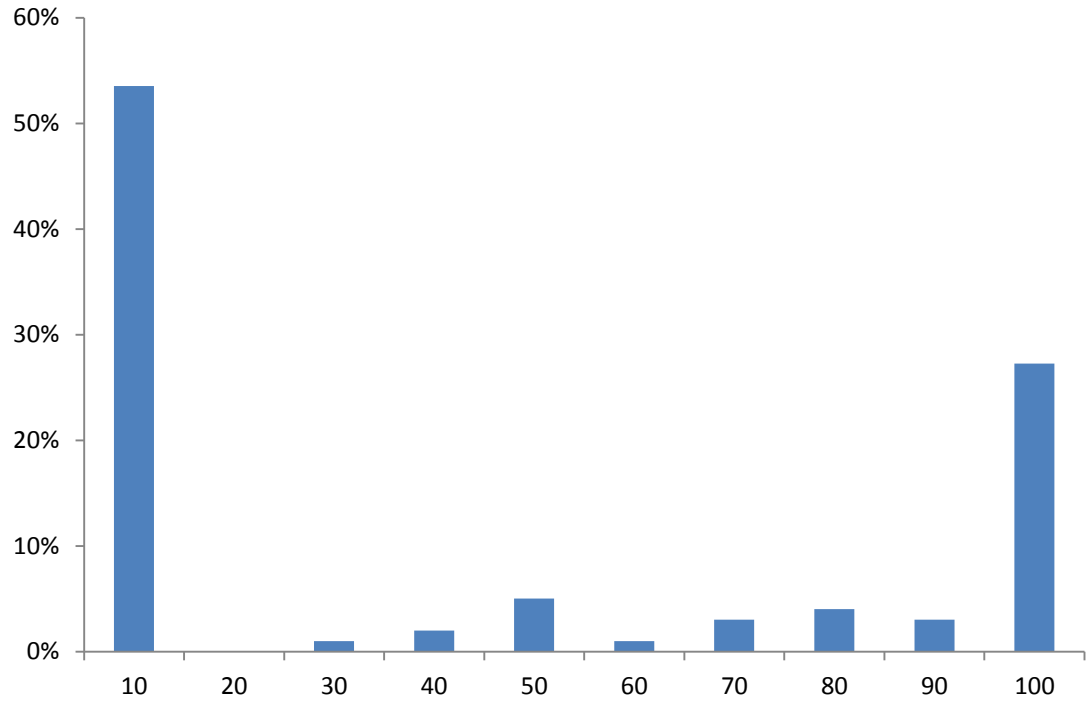

Figura 3.9: Histograma *Top 10* para los códigos CIE-9-MC en *Knn*

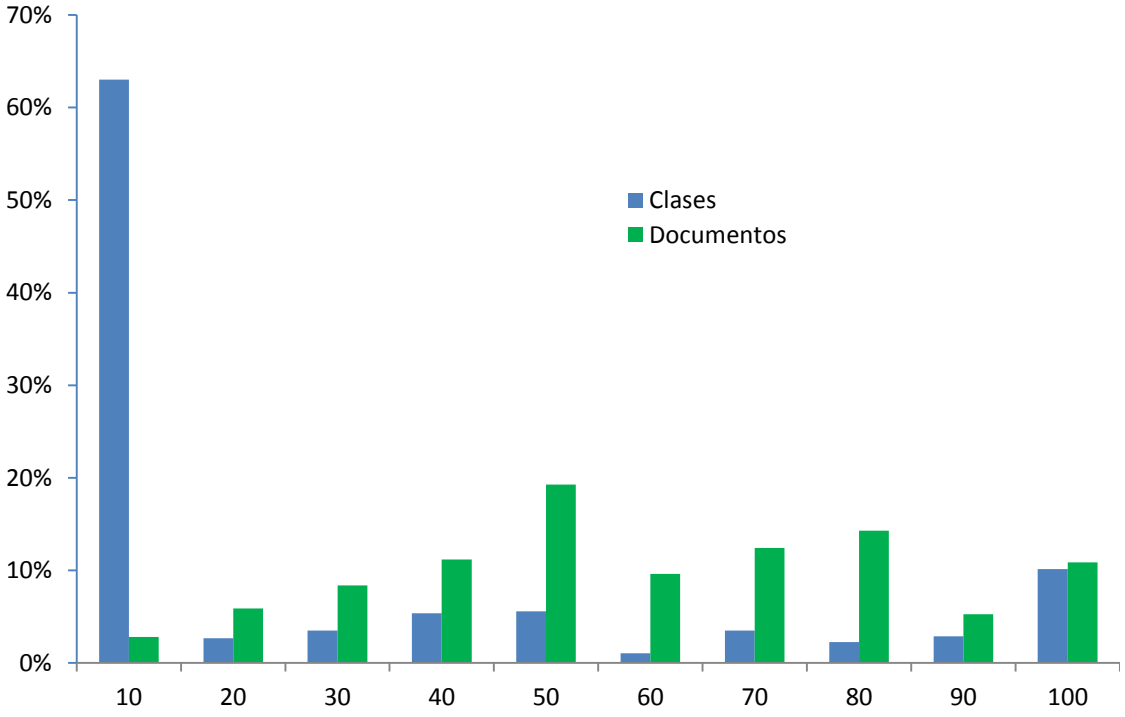

Figura 3.10: Histograma *Recall 15* para los códigos CIE-9-MC y documentos en *Knn*

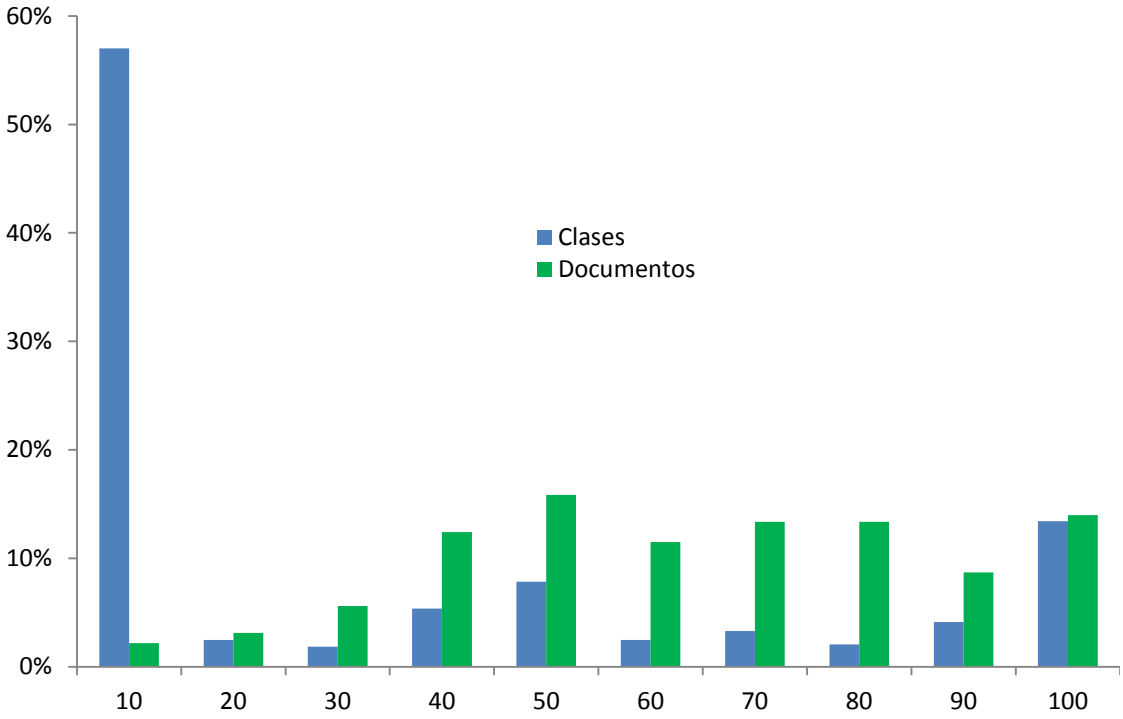

Figura 3.11: Histograma *Recall 20* para los códigos CIE-9-MC y documentos en *Knn*

En todas las figuras anteriores se visualiza que hay un conjunto importante de clases en donde el clasificador no obtiene unos buenos resultados. Esta tendencia no se produce en los documentos. La colección, como ya se ha mencionado, está desbalanceada, el 20% de las clases representa prácticamente al 80% de los documentos. Esto nos sugiere a pensar que las clases más representadas en la colección, el clasificador funciona mejor. Parece que el desbalanceo de la colección afecta a un clasificador *Knn*, como el propuesto en los experimentos.

3.7.1. Resultados con diferentes modelos de recuperación

Los resultados anteriores se obtienen con el sistema de recuperación de *Indri* [Strohman et al., 2005], que es un sistema avanzado de búsqueda construido por *Lemur*. Este modelo de RI se fundamenta en una combinación entre el lenguaje de modelado [Ponte y Croft. 1998] y una red de inferencia [Turtle y Croft, 1991].

Los resultados con *Indri* se van a comparar con otros modelos vectoriales de *Lemur*, *TFIDF* y *coseno.* Los resultados en la representación *Total* con *k=20* y pesado básico para la asignación de códigos con *microaveraging* y *macroaveraging* se muestran en las tablas 3.10 y 3.11.

Clasificación Códigos					
Modelo	11-pt Avg	Top Candidato	Top 10	Recall 15	Recall 20
Indri	43.1	16.1	64.9	52.5	57.7
TFIDF	40.5	10.5	55.6	50.7	55.6
Coseno	43.8	17.1	65.5	54.3	60.0

Tabla 3.10: Rendimiento de distintos modelos de RI con *microaveraging* (*K=20*, pesado básico)

Clasificación Códigos				
Modelo	Top Candidato	Top 10	Recall 15	Recall 20
Indri	9,3	39,3	23,2	28,8
TFIDF	5,5	29,0	22,4	27,5
Coseno	14,1	47,0	26,8	32,5

Tabla 3.11: Rendimiento de distintos modelos de RI con *macroaveraging* (*K=20*, pesado básico)

El modelo *Lemur Coseno* obtiene mejores resultados que *Indri* y *Lemur TFIDF*. Los resultados *Lemur Coseno* superan ligeramente resultados de *Indri*. Las mayores diferencias entre *Coseno* e sitúan en*: Top 10* con un 6.8% de incremento y *Recall 20* con un 3%.

3.7.2. Sistema de pesado en la asignación de códigos

En todos los experimentos anteriores se ha utilizado el sistema de pesado básico ($w_{ic} = 1$). En esta sección se analiza como influyen distintos sistemas de pesado en los resultados de la clasificación. Para ello se asigna un peso igual a 1 ($w_{ic} = 1$) para los códigos secundarios [DxS] y un peso mayor que 1 ($w_{ic} > 1$) para el código principal [DxP]. Los resultados alcanzados en la representación *Total* para distintos pesos del diagnóstico principal están reflejados en la tabla 3.12.

Los resultados anteriores nos revelan que *Top candidato* y *Top 10* mejoran con un mayor peso al código principal. En cambio, *Recall 15* y *Recall 20* van empeorando ligeramente a medida que aumentamos el peso. Lo más destacable es la mejora del *Top candidato*, al utilizar un peso de 1.8 para los códigos principales: se produce un incremento de un 99.7% con respecto al obtenido con un peso igual a 1. Este incremento es menos sustancial en las categorías, en donde un peso de 1.8 implica una

mejora de un 59%. Esto demuestra que la estrategia de ponderación descrita anteriormente funciona bien en estos problemas de clasificación.

Peso (DP)	11-pt Avg	Top Candidato	Top 10	Recall 15	Recall 20
Clasificación Códigos					
1	43.1	16.1	64.9	52.5	57.7
1.5	42.5	28.9	68.9	52.4	57.5
1.8	41.7	31.9	69.9	52.3	57.5
2.3	40.8	34.5	73.3	51.0	54.3
2.5	40.5	34.5	73.3	50.9	54.3
2.7	40.3	35.4	73.6	50.9	54.3
4.3	37.2	37.3	76.7	45.5	52.7
Clasificación Categorías					
1	51.2	22.7	74.2	62.4	67.7
1.5	50.6	33.8	77.0	62.4	67.5
1.8	50.3	36.0	78.6	62.4	67.4
2.3	49.3	38.8	80.1	61.2	65.7
2.5	48.9	39.1	80.1	61.2	65.7
2.7	48.5	40.3	80.4	61.2	65.7
4.3	46.0	41.3	82.9	57.0	64.5

Tabla 3.12: Rendimiento de distintos pesos con *Indri* para *K=20* en la colección *Total*

3.8. Experimentos con SVM

En los experimentos con SVM se utiliza la representación *Total*, ya que es una de las representaciones con mejores resultados en *Knn*. El preprocesamiendo aplicado a la colección para SVM es el definido en la sección 3.6. La colección de entrenamiento y test con la que se realizan los experimentos es la misma que la de *Knn*. Para desarrollar los experimentos se usa el paquete SVM^{light} de *Joachims* [Joachims, 1999], una implementación en lenguaje C para SVM.

Para la representación de los documentos se recurre al modelo vectorial TFIDF, que genera un vocabulario de 19.924 términos. El proceso de clasificación emplea las típicas fases de aprendizaje y clasificación. SVM^{light} está desarrollado para trabajar con dos clases, problema binario de SVM. Los experimentos están en un entorno

multiclase, para solucionar está dificultad y poder trabajar con *SVM*^{light} se recurre a la alternativa de *1-vs-todos*. La colección tiene 1238 clases/códigos y la colección de test está compuesta por 322 documentos. Si la colección tiene 1238 clases/códigos en una clasificación *1-vs-todos* tenemos que construir 1238 clasificadores.

En los experimentos se ha trabajado con distintos parámetros configurables y núcleos. Con núcleos polinómicos, gaussianos y con variaciones de sus parámetros, los resultados han sido bastante pobres con respeto a los obtenidos con una clasificación lineal. Los experimentos con clasificación lineal se han realizados para distintos valores de C, que mostramos en la tabla 3.13[2] para *microaveraging* y en la tabla 3.14[2] para *macroaveraging*.

C	11-pt Avg	Top Candidato	Top 10	Recall 15	Recall 20
Clasificación Códigos					
Default[2]	58.1	16.1	74.8	67.3	72.8
0.5	59.4	16.7	73.2	67.3	72.8
1000	59.4	16.7	73.2	67.3	72.8
Clasificación Categorías					
Default	66	22.0	84.1	77.6	82.2
0.5	67.3	22.9	83.2	77.8	82.3
1000	67.3	22.9	83.2	77.8	82.3

Tabla 3.13: Resultados *microaveraging* de SVM lineal para la representación *Total*

C	Top Candidato	Top 10	Recall 15	Recall 20
Clasificación Códigos				
Default[2]	11,6	51,5	39,9	48,4
0.5	9,8	51,4	41,6	48,1
1000	10,8	52,8	42,5	49,0
Clasificación Categorías				
Default	15,9	67,4	51,9	59,5
0.5	16,8	71,3	54,4	62,2
1000	16,8	71,3	54,4	62,2

Tabla 3.14: Resultados *macroaveraging* de SVM lineal para la representación *Total*

[2] Por defecto *SVM*^{light} establece $C = {}^{n}/_{\sum_{i=1}^{n} x_i \cdot x_i}$, donde n es el número de documentos de la colección de entrenamiento

Se visualiza mediante los siguientes histogramas el comportamiento de la colección para las métricas definidas.

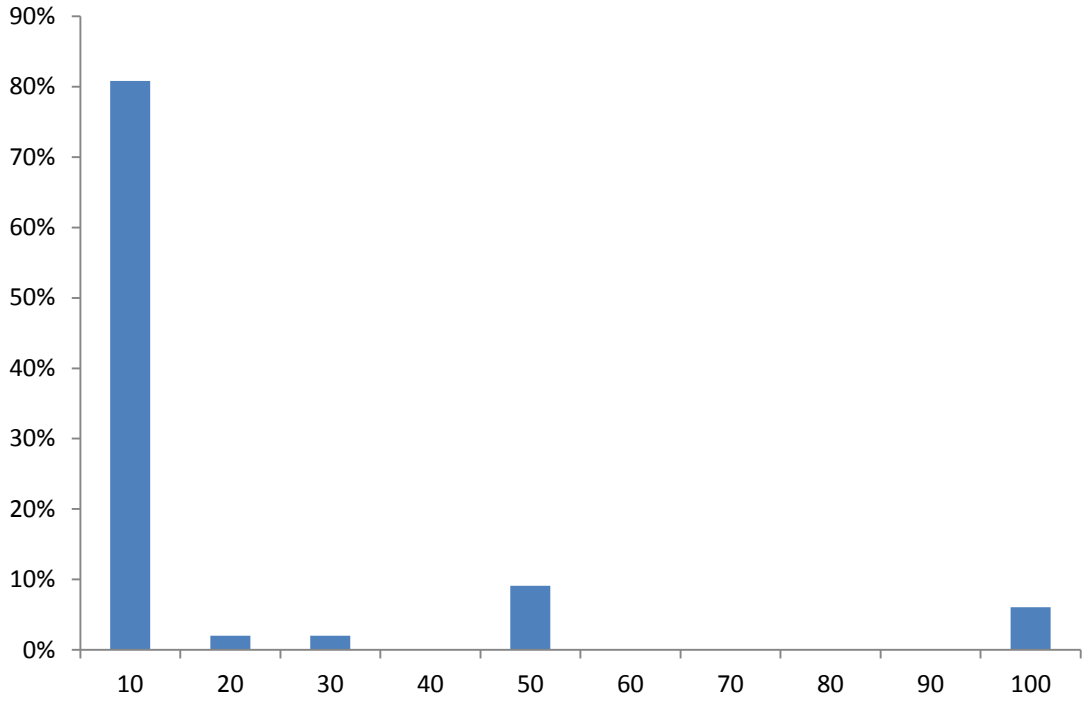

Figura 3.12: Histograma *Top Candidato* para los códigos CIE-9-MC en SVM

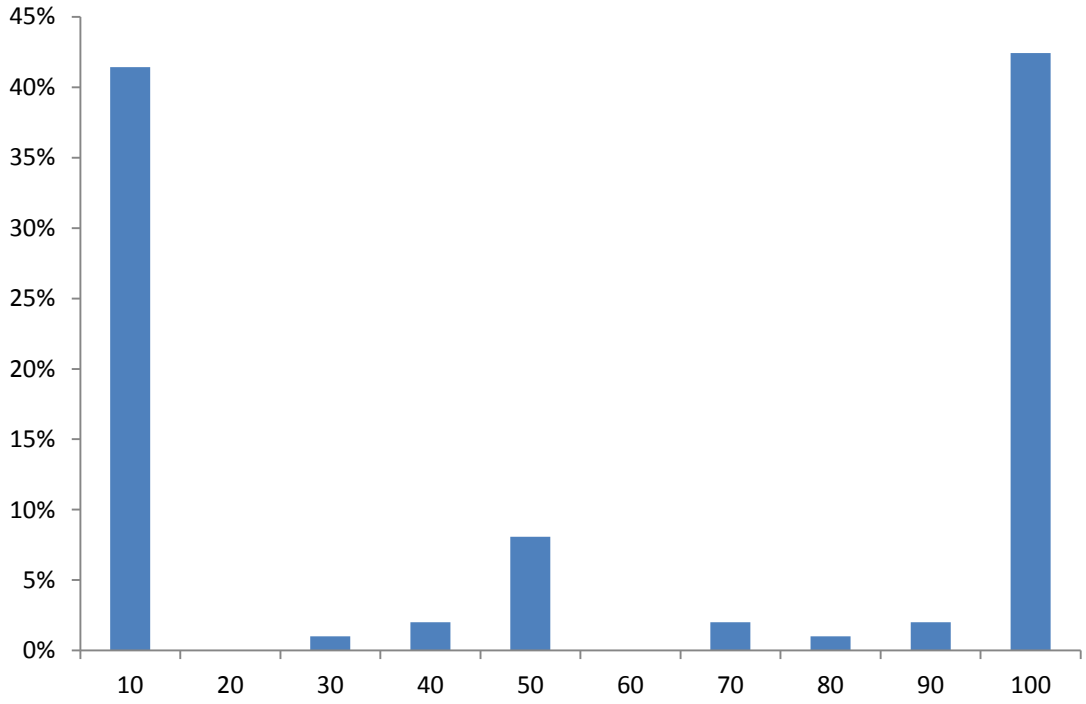

Figura 3.13: Histograma *Top 10* para los códigos CIE-9-MC en SVM

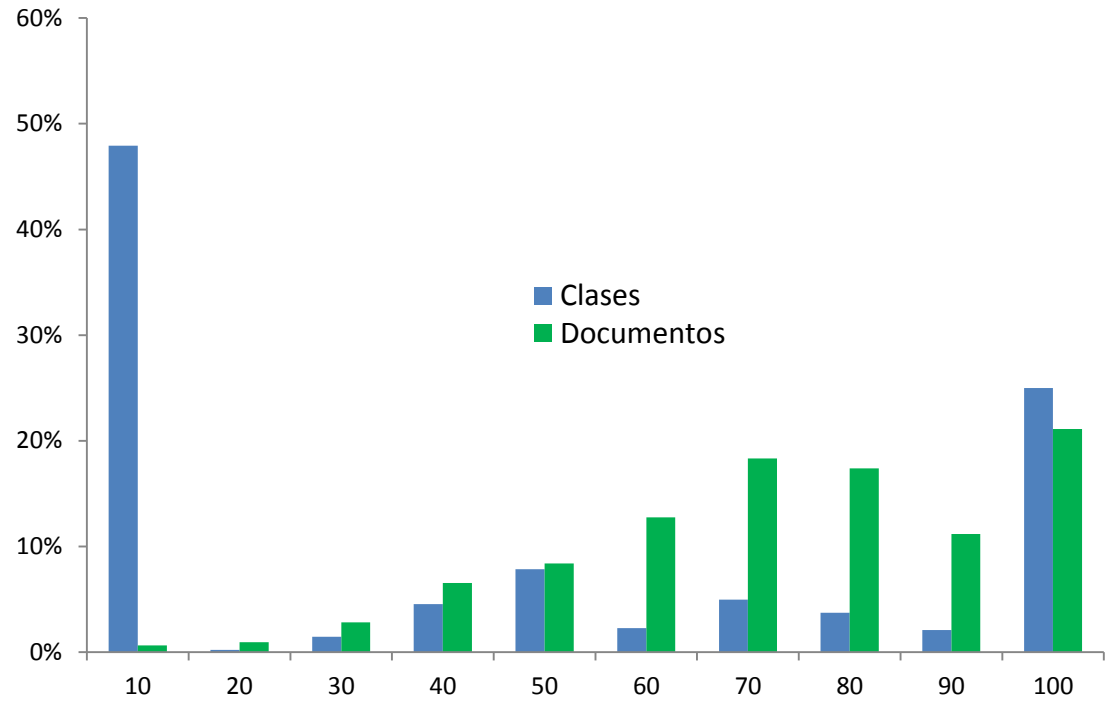

Figura 3.14: Histograma *Recall 15* para los códigos CIE-9-MC y documentos en SVM

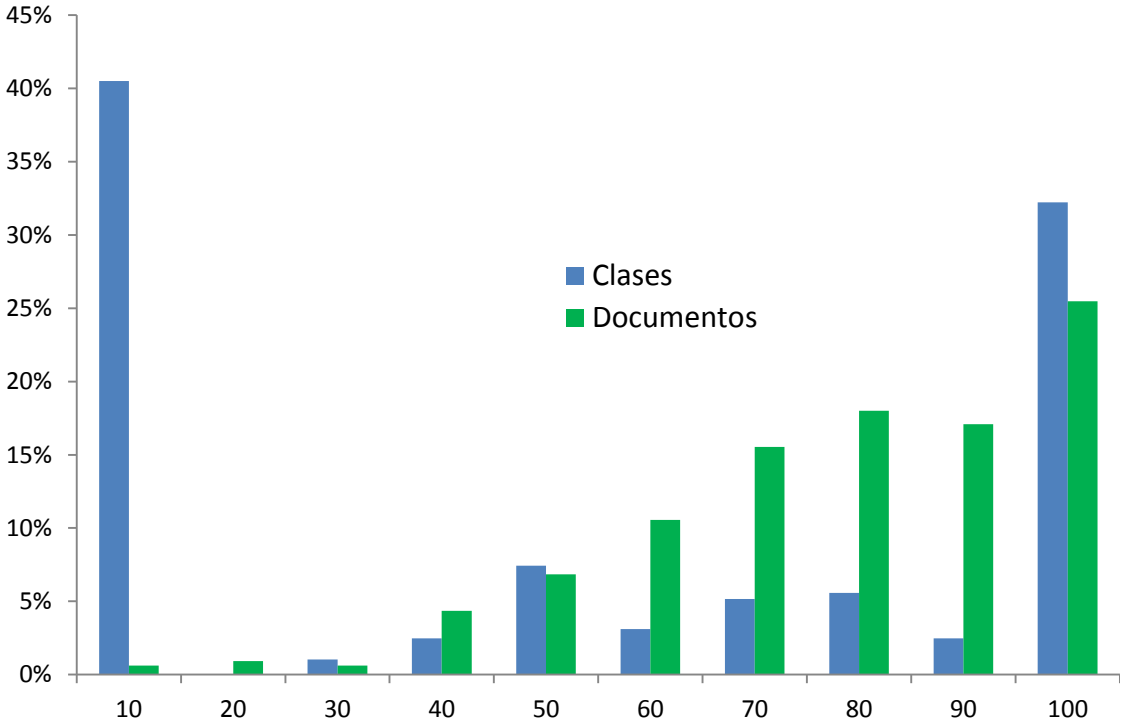

Figura 3.15: Histograma *Recall 20* para los códigos CIE-9-MC y documentos en SVM

Las figuras anteriores nos muestran que para SVM hay más clases en donde el clasificador funciona bien, en comparación con *Knn*. Los histogramas nos develan que los datos son mejores para SVM que para *Knn*. Una posible explicación puede estar en que para SVM el desbalanceo de la colección influye menos que en *Knn*. Estos resultados nos pueden dar una idea interesante sobre la interpretación de los vectores de soporte en SVM. A menudo se argumenta que los vectores de soporte proporcionan una representación suficiente para los documentos en una clasificación. Ya que los vectores de soporte son una descripción suficiente de la frontera de decisión pero no de los ejemplos mismos. Los vectores de soporte nos proporcionan una estimación para la probabilidad de clasificar un ejemplo, pero no representan la probabilidad de los elementos de la colección de entrenamiento en el clasificador. Por esta razón el clasificador SVM es más independiente de la composición de la colección de entrenamiento, y no le afecta de forma tan directa el desbalanceo de la colección. Sin embargo, al igual que otros modelos de clasificación, SVM tiene como objetivo minimizar el error en el conjunto en toda la colección, por lo que es intrínsecamente proclive a la clase mayoritaria. Las colecciones reales de CIE-9-MC, por sus propia naturaleza intrínseca, es un conjunto de datos no balanceados, ya que contienen muchos ejemplos de una clase, pero muy pocos para otras. Esto se acentúa al aplicar en SVM *1-vs-todos*, ya que la colección se desbalancea aún más. Esto supone en la colección MIR–Conxo una relación de 1238:1 para algunas clases, lo que se considera un desequilibrio severo. En estos casos para el clasificador SVM, la clase positiva (minoritaria) es el 0,08% de los documentos y la clase negativa (mayoritaria) el 99,92%. La mayoría de los algoritmos de clasificación tienen como objetivo obtener un modelo con un acierto alto y una buena capacidad de generalización, que beneficia la cobertura de los ejemplos mayoritarios. Esta tendencia inductiva supone un serio problema para la clasificación de datos muy desbalanceados. El clasificador no puede diferenciar entre ejemplos ruidosos y ejemplos de la clase minoritaria y de esta forma pueden ser ignorados por el clasificador. Aun así, la mejoría de los resultados de SVM viene acompañada con un incremento de las clases en donde el clasificador funciona de forma correcta y obtiene un acierto pleno.

3.9. Comparativa *Knn* – SVM

En la tabla 3.15 se presenta una comparación de los mejores resultados de *Knn* y SVM, para *microaveraging*. Los resultados más robustos en *Knn* son con un peso = 2.7 para códigos y un peso = 2.5 para categorías (ver tabla 3.12). Estos datos demuestran que *Knn* con una ponderación adecuada es muy eficaz para lograr un buen rendimiento en Top candidato. En todo el resto de métricas SVM es superior a *Knn*.

	11-pt Avg	Top Candidato	Top 10	Recall 15	Recall 20
Clasificación Códigos					
Knn	40.3	35.4	73.6	50.9	54.3
SVM	59.4	16.7	73.2	67.3	72.8
Clasificación Categorías					
Knn	48.9	39.1	80.1	61.2	65.7
SVM	67.3	22.9	83.2	77.8	82.3

Tabla 3.15: *Knn vs* SVM. *Microaveraging*

3.10. Conclusiones y trabajo futuro

Nuestro trabajo se diferencia de otros estudios en que la colección en la que se realizan los experimentos está en castellano. La gran mayoría de los estudios en CAT para CIE-9-MC, que se mencionan en esta tesis son desarrollados sobre colecciones en lengua inglesa. Se ha creado una colección específica compuesta con los informes de alta de un Servicio de Medicina Interna. El tipo de episodios clínicos que la forman genera una colección muy difícil, como ya se ha detallado en este capítulo.

Los experimentos se han desarrollado con diferentes métodos de clasificación y con diferentes representaciones de la colección. Comparando varias técnicas de CAT se ha encontrado que SVM funciona mejor que *Knn* en casi todos los ámbitos. Excepto en la métrica Top candidato en donde *Knn* con un peso mayor que 1 en el diagnóstico principal consigue mejores resultados que SVM. Esto demuestra la potencia de aprendizaje que se consigue con SVM.

Los resultados de rendimiento obtenidos son suficientemente buenos para construir una herramienta evaluable en el trabajo real dentro de un centro hospitalario. Estos datos son mejores que los obtenidos en una colección con una estructura similar, pero en inglés y con una dificultad inferior. La gran mayoría de trabajos realizados en el

ámbito de la investigación de clasificación CIE-9-MC son sobre la colección pública CCHMC. Esta colección de informes radiológicos está en otro entorno diferente al propuesto en nuestro trabajo y que se ha adaptado en su composición para favorecer unos buenos resultados.

Las posibles mejoras se pueden enfocar en varias líneas. La primera evitar que la colección esté desbalanceada, procurando que el número de documentos de una clase no sea muy superior al de las otras. Y si esto es inevitable se puede intentar solucionar este problema en SVM mediante la separación imperfecta, en la que se utilizan factores de penalización distintos para cada clase, lo que permite ajustar el coste de los falsos positivos y de los falsos negativos de manera independiente. Dentro de este ámbito se podrían utilizar técnicas para seleccionar solo aquellos documentos más informativos para la colección de entrenamiento, intentado conseguir una colección lo menos desbalanceada posible que mejore el clasificador.

Otra de las posibles líneas de trabajo podría ser utilizar expansión de las consultas (*query expansion)*, ya que con la expansión de documentos los resultados mejoraron. Dentro de SVM, también podemos considerar la utilización de otras técnicas de clasificación multietiqueta, o la búsqueda de un núcleo que facilite el sesgo en este tipo de clasificación.

En la actualidad la clasificación automática de textos es un problema real que afecta a todos los centros sanitarios del mundo. Con una solución automática fiable y estable se podrían codificar otras áreas de los centros sanitarios que en estos momentos no se codifican por la cantidad de recursos humanos necesarios. Sin olvidarnos que esto repercutiría en una mejor gestión económica, un perfeccionamiento en la gestión asistencial y en la atención sanitaria del paciente.

Capítulo 4

Evaluación de técnicas de Aprendizaje Activo para codificación CIE-9-MC de informes de alta hospitalaria

El Aprendizaje Activo es una técnica según la cual, a partir de un conjunto de documentos sin etiquetar, se ordenan y seleccionan los documentos para ser etiquetados de modo que el nuevo conjunto de entrenamiento mejore el clasificador construido. En los hospitales se genera un gran volumen de información, pero sólo se codifica una pequeña parte de los informes producidos. Es por tanto un escenario donde se necesita elegir bien lo que se etiqueta para que las herramientas automatizadas de clasificación puedan surtirse de buenos conjuntos de entrenamiento. En nuestro trabajo, vamos a utilizar técnicas de Aprendizaje Activo para elegir los informes de alta hospitalaria que se deben etiquetar con códigos CIE-9-MC y, a continuación, evaluaremos la calidad de ese proceso de selección. Los documentos se representan utilizando técnicas populares en Recuperación de Información y la calidad de los conjuntos de entrenamiento se evalúa utilizando clasificación con Máquinas de Soporte Vectorial. El dominio clínico donde trabajamos es muy complejo, con un gran número de clases, y con un desbalanceo significativo entre las clases. Los resultados de experimentación demuestran que nuestra estrategia es prometedora para mejorar este tipo de sistemas.

4.1. Introducción

En los hospitales se genera un gran volumen de información con considerable complejidad. La capacidad de clasificación manual es limitada por lo que es imposible que todos los documentos producidos sean etiquetados. Una de las tareas de clasificación que se realizan es la codificación de los diagnósticos de los informes de alta. La codificación es un proceso que consiste en analizar la documentación del alta, y asignar los códigos de los diagnósticos de ese episodio clínico. Este proceso se realiza de forma manual por un médico codificador, con un gran coste por la complejidad del tipo de clasificación. En los hospitales, los episodios que se codifican habitualmente son los ingresos hospitalarios. Otros episodios clínicos no son codificados (por ejemplo los que corresponden a episodios de consultas externas, urgencias, pruebas funcionales, etc.). Actualmente el porcentaje de episodios clínicos que se codifican con respecto al total es mínimo. Si quisiésemos codificar todos los episodios clínicos que se generan en un centro hospitalario, tendríamos que aumentar de forma considerable los recursos humanos de médicos codificadores, lo que implicaría un elevado coste económico.

Debido a estas limitaciones los episodios clínicos pasan usualmente por una clasificación generalista, simplemente para generar una contabilidad básica, sin considerar la patología tratada para cada paciente. En cambio, con la codificación CIE-9-MC completa de estos episodios podríamos medir, comparar y mejorar la calidad asistencial, agrupando a los pacientes de acuerdo a requerimientos y características comunes.

En la literatura existen propuestas de clasificación automática para dar soporte a la codificación de informes [Larkey y Croft, 1995], [Larkey y Croft, 1996]. Para poder aplicar técnicas de clasificación automática a los episodios clínicos, debemos elegir bien los episodios que actúan como entrenamiento para el clasificador. Para ello proponemos aplicar técnicas de Aprendizaje Activo (AA), para seleccionar aquellos documentos con los cuales poder entrenar un clasificador con resultados eficaces. El aprendizaje activo [Cohn et al., 1994], [Tong y Koller, 2001] es una técnica empleada en el entrenamiento de clasificadores que a partir de un conjunto de documentos sin etiquetar, escoge los documentos más informativos para la construcción del clasificador, obteniendo un conjunto etiquetado con una alta capacidad discriminante.

Con el fin de minimizar el impacto que la limitación de la información pueda tener sobre la eficacia del clasificador es fundamental seleccionar adecuadamente los documentos que forman la colección. Las colecciones con un alto grado de desequilibrio en las clases, puede provocar un efecto negativo en la clasificación [Weiss y Provost, 2001]. AA ha sido estudiado en la comunidad científica en problemas de clasificación con colecciones desbalanceadas [He y García, 2009].

En este capítulo evaluamos AA sobre una colección real del dominio clínico que presenta alto desbalanceo entre clases y en la que el problema de clasificación es difícil. Nuestros experimentos demuestran que AA se puede utilizar en este escenario con resultados razonables.

4.2. Aprendizaje activo para la clasificación de textos multietiqueta

En muchas ocasiones en un esquema de aprendizaje supervisado la colección de entrenamiento es pequeña, y es muy costoso obtener nuevos documentos etiquetados. Sin embargo, se suele disponer de una gran cantidad de documentos sin etiquetar. Esta es la situación que nos encontramos en los hospitales en donde para ciertas áreas clínicas no disponemos de una colección etiquetada de sus episodios. Además, la etiquetación manual es costosa y ha de ser realizada por expertos.

Para conseguir la colección de entrenamiento aplicamos AA en un entorno multietiqueta. Dado un conjunto de clases predefinidas $C = \{C_1, \ldots, C_m\}$, y una colección de documentos a clasificar *(D)*, el objetivo es encontrar una función $\varphi: D \times C \rightarrow \{-1, +1\}$, que denominamos clasificador (-1 y +1 representan la pertenencia o no a una clase). Un documento puede no pertenecer a ninguna clase, a una clase o a varias clases. En una clasificación de textos multietiqueta usualmente se generan m clasificadores binarios, uno por cada clase c_j. En nuestro trabajo nos centramos en clasificadores que tienen la forma $\hat{\varphi}: D \times C \rightarrow [-1, +1]$. Esta función permite estimar la clase a la que el clasificador cree que el documento pertenece (signo de $\left(\hat{\varphi}(d_i, c_j)\right)$) y, además proporciona una estimación de la confianza del clasificador en la decisión tomada $\left|\left(\hat{\varphi}(d_i, c_j)\right)\right|$.

Las técnicas de AA para entornos con una única clase consisten básicamente en escoger primero aquellos ejemplos no etiquetados sobre los que el clasificador automático tiene menos confianza. El codificador humano de esta forma etiquetará manualmente solo aquellos documentos más informativos para el aprendizaje. En

situaciones como la nuestra, con múltiples clases, cada documento tiene un valor de confianza para cada clase por lo que es necesario combinar esas puntuaciones para estimar qué documento no etiquetado es globalmente más informativo para el proceso de clasificación multietiqueta.

En la literatura [Esuli y Sebastiani, 2009], nos encontramos dos opciones para seleccionar los documentos que vamos a incorporar al conjunto de entrenamiento. Por un lado, generar *m* rankings independientes de documentos (cada uno de ellos asociado a una clase). El médico tendría que codificar los documentos que figuran más arriba en el ranking para cada clase. Esta opción se denomina *etiquetado local*, ya que se realiza localmente en cada clase. La otra opción, consiste en generar un único ranking de documentos en base a la combinación de los *m* valores de confianza asociados a un mismo documento, y se denomina *etiquetado global*. En colecciones con un número elevado de clases, como ocurre con la colección utilizada en esta investigación, el etiquetado local obliga al médico codificador a trabajar con muchos rankings diferentes. Esto supone un gran esfuerzo para la etiquetación puesto que un documento puede aparecer en múltiples rankings y el codificador debe revisarlo cada vez para asignar, o no, la etiqueta de la correspondiente clase. En cambio, la opción de etiquetado global se adapta mejor a los requisitos del trabajo a desarrollar, pues el médico codificador revisa exclusivamente un único ranking y está garantizado que un documento se lee como máximo una vez.

En este trabajo vamos a comparar algunas estrategias propuestas en [Esuli y Sebastiani, 2009] y aplicadas en [Lojo, et al., 2010] , para la obtención de un ranking de documentos en orden decreciente en cuanto a su capacidad informativa, dentro de una estrategia de etiquetado global. Las distintas variantes se definen a través de tres dimensiones: dimensión "evidencia", dimensión "clase" y dimensión "peso". Cada estrategia que apliquemos va a ser una combinación de decisiones en estas tres dimensiones, y las representaremos con una secuencia de tres letras, en donde cada letra representa una elección en cada una de las dimensiones.

Antes de entrar en detalles sobre las dimensiones, conviene aclarar terminología: dado el clasificador $\hat{\varphi}: D \times C \rightarrow [-1, +1]$, el valor $\hat{\varphi}(d_i, c_j)$ lo denominaremos *puntuación* de la clase c_j para el documento d_i y el valor $|\hat{\varphi}(d_i, c_j)|$ la *confianza* de la clase c_j para el documento d_i y $sgn\left(\hat{\varphi}(d_i, c_j)\right)$ será el signo de la categoría c_j para el documento d_i.

4.2.1. Dimensión "evidencia"

Esta dimensión hace referencia al tipo de evidencia que utilizamos como base para la estimación de lo informativo que es un documento para una clase. Una posibilidad es utilizar el valor $\left|\hat{\varphi}\left(d_i, c_j\right)\right|$, que representa la ***Confianza*** (**C**) del clasificador de la clase c_j con respecto al documento d_i. La intuición es que cuanto menor sea el valor de la confianza con el clasificador actual, el documento aportará más información al clasificador tras ser etiquetado (esto es, tenderemos a etiquetar los documentos sobre los que haya más duda).

La otra alternativa es usar directamente la puntuación $\hat{\varphi}\left(d_i, c_j\right)$ como evidencia. Aquí la intuición es otra, se trata de promover documentos que son claramente ejemplos positivos de la clase porque en muchas situaciones los casos positivos son los que ayudan más en escenarios supervisados. A esta opción le denominaremos ***Puntuación*** (en inglés *Score*, ***S***).

Tenemos pues dos estrategias, *confianza* y *puntuación*, a la hora de estimar lo informativo que es un documento para una clase. La siguiente dimensión, de clase, hace referencia a cómo se combinan los *m* valores de evidencia que tiene un documento en un único valor.

4.2.2. Dimensión "clase"

En la dimensión "clase" lo que se pretende es generar una valoración global para cada documento independiente de la clase. Una opción consiste en maximizar la capacidad informativa esperada calculada sobre todas las clases. Esto significa que si la dimensión de evidencia es Confianza (C), para la dimensión de "clase" se tomaría $min_{c_j \in C}\left|\hat{\varphi}\left(d_i, c_j\right)\right|$. Si la dimensión de evidencia es Puntuación (S), para la dimensión de clase escogeríamos $max_{c_j \in C}\, \hat{\varphi}\left(d_i, c_j\right)$. Intuitivamente pretendemos que el médico codificador se concentre en documentos que se consideran de gran valor por lo menos para una clase. Esta elección se denomina ***Min / Max (M)***.

Otra opción consiste en utilizar el promedio de todos los valores obtenidos para un documento en todas las clases. De este modo seleccionaremos los documentos útiles globalmente para el conjunto de clases. Esta aproximación se llama ***Promedio*** (***Avg, A***).

Una última opción, ***Round Robin (R)***, consiste en seleccionar los documentos mejores de cada clase de la siguiente forma: se toma el mejor documento para cada clase

(según dimensión de evidencia) y se crea un ranking de documentos (como mucho de tamaño *m*, pues se eliminan repetidos) ordenados decrecientemente según evidencia; a continuación, se toman los segundos documentos de cada clase, se incorporan al final del ranking (ordenados entre sí por evidencia), y así sucesivamente.

4.2.3. Dimensión "Peso"

La dimensión "Peso", ***Weight (W)*** tiene la función de no tratar a todas las clases por igual. Uno de los objetivos es dar más peso a las clases en donde el clasificador obtiene peores resultados. Para ello utilizamos una función de evaluación $f(\emptyset_j)$ que tenga un valor entre [0,1] y que nos indica qué rendimiento tiene el clasificador automático para la clase c_j. Cuando estemos trabajando con Confianza (C), multiplicaremos el valor de la confianza $|\hat{\varphi}(d_i, c_j)|$ por la función de evaluación $f(\emptyset_j)$, que indica la efectividad del clasificador en la clase. Para Puntuación (S) calcularíamos el producto de $\hat{\varphi}(d_i, c_j)$ por $\left(1 - f(\emptyset_j)\right)$. Estos ajustes sobre los valores de evidencia consiguen promover aquellas clases que no dan buen rendimiento ($f(\emptyset_j)$ bajo).[3] En nuestros experimentos, al igual que se hizo en [Esuli y Sebastiani, 2009], utilizaremos la conocida medida F_1 para definir $f(\emptyset_j)$, aplicando además un suavizado de *Laplace* ($\varepsilon = 0.05$) para evitar multiplicaciones por 0. La alternativa de no utilizar pesos para las clases la denominaremos ***No Weight (N)***.

4.3. Metodología para evaluar Aprendizaje Activo

Para evaluar las estrategias de AA mantendremos una colección separada de documentos de test (*TestSet*) y crearemos incrementalmente una colección de entrenamiento cuya calidad iremos contrastando mediante la construcción de un clasificador automático a partir del conjunto de entrenamiento y su evaluación contra el *TestSet*.

En el proceso de clasificación vamos a utilizar una representación vectorial de los documentos en un espacio de características. En este trabajo hemos seleccionado una representación vectorial basada en el popular esquema de pesado *TFIDF*. Para clasificar utilizaremos la metodología aplicada en el capítulo 3 sobre Máquinas de

[3] Para confianza se multiplica por $f(\emptyset_j)$ porque la elección de documentos se hace en orden creciente de evidencia. Para una clase *j* con pobre rendimiento ($f(\emptyset_j)$ bajo), $f(\emptyset_j)$ multiplicado por su confianza será bajo con lo que se favorecerá la elección de los documentos. Análogamente sucede con Puntuación al multiplicarle por $\left(1 - f(\emptyset_j)\right)$ y escoger los documentos en orden decreciente.

Soporte Vectorial (SVM). La implementación de SVM utilizada fue SVM^{Light} [Joachims, 1999]. Para cada clase el clasificador SVM utilizado nos devuelve un valor que representa $\hat{\varphi}(d_i, c_j)$. La distancia al hiperplano lo interpretamos como la puntuación y su valor absoluto como la confianza.

Sea D_{Todo} el conjunto de documentos etiquetados con el que vamos a surtir a la colección de entrenamiento. El algoritmo es como sigue:

1. Seleccionamos aleatoriamente 100 documentos de D_{Todo} que denominamos D_{Train}, en donde $|D_{Train}| = 100$ y $D_{Todo} = D_{Todo} - D_{Train}$.
2. Construimos *m* clasificadores SVM, un clasificador para cada clase, utilizando D_{Train} y los evaluamos con *TestSet.*
3. Aplicamos *AA* para seleccionar 50 nuevos documentos para incorporar a D_{Train}. Para ello, según la variante utilizada:

   a. *Evidencia de Confianza* y *Clase Min (CM)*

      $\forall d_i \in D_{todo}$ calculamos $d_{score} = min|\hat{\varphi}(d_i, c_j)| \quad \forall c_j \in (C_1, \dots, C_m)$

      Realizamos un ranking creciente por d_{score} $\left(d^i_{score} \leq d^{i+1}_{score}\right)$ y seleccionamos el *Top 50*.

   b. *Evidencia de Puntuación* y *Clase Max (SM)*

      $\forall d_i \in D_{todo}$ calculamos $d_{score} = max\left(\hat{\varphi}(\mathrm{d_i}, \mathrm{c_j})\right) \quad \forall c_j \in (C_1, \dots, C_m)$

      Realizamos un ranking decreciente por d_{score} $\left(d^i_{score} \geq d^{i+1}_{score}\right)$ y seleccionamos el *Top 50*.

   c. *Evidencia de Confianza* y *Clase Promedio (CA)*

      $\forall d_i \in D_{todo}$ calculamos $d_{score} = Avg|\hat{\varphi}(d_i, c_j)| \quad \forall c_j \in (C_1, \dots, C_m)$

      Realizamos un ranking creciente por d_{score} y seleccionamos el *Top 50*.

d. Evidencia de Puntuación y *Clase Promedio (SA)*

$\forall d_i \in D_{todo}$ calculamos $d_{score} = Avg\left(\hat{\varphi}(d_i, c_j)\right) \quad \forall c_j \in (C_1, \dots, C_m)$

Realizamos un ranking decreciente por d_{score} y seleccionamos el *Top 50*.

e. Evidencia de Confianza y *Clase Round Robin (CR)*

$\forall c_j \in (C_1, \dots, C_m)$ calculamos $d_{C_j} = min\left|\hat{\varphi}(d_i, c_j)\right| \quad \forall d_i \in D_{todo}$

Con el documento d_{C_j} seleccionado para cada clase realizamos un ranking de documentos por orden creciente de confianza, eliminamos los duplicados y seleccionamos el *Top 50*. Si no hay 50 documentos en el ranking se tomarían los siguientes documentos seleccionados por cada clase y así sucesivamente.

f. Evidencia de Puntuación y *Clase Round Robin (SR)*

$\forall c_j \in (C_1, \dots, C_m)$ calculamos $d_{C_j} = max\left(\hat{\varphi}(d_i, c_j)\right) \quad \forall d_i \in D_{todo}$

Con el documento d_{C_j} seleccionado para cada clase realizamos un ranking de documentos por orden decreciente de confianza, eliminamos los duplicados y seleccionamos el *Top 50*. Si no hay 50 documentos en el ranking se tomarían los siguientes documentos seleccionados por cada clase y así sucesivamente.

4. Los 50 documentos obtenidos los incorporamos a D_{Train}: $D_{train} = D_{train} \cup Top\ 50$ y $D_{Todo} = D_{Todo} - Top\ 50$.
5. Volver al paso 2 si $|D_{train}| < 1000$.

Si además se quiere aplicar la dimensión de Peso, lo primero es calcular F_1 para cada clase, F_1^j. Estos valores se calculan con los clasificadores construidos en el punto 2 y evaluados con *TestSet.* Y este valor lo obtenemos de la siguiente forma:

$$F_1^j = \frac{2TP_j}{2TP_j + FP_j + FN_j} \quad \forall j \in (C_1, \dots, C_m) \qquad (4.1)$$

Si $TP_j = FP_j = FN_j = 0$ entonces $F_1^j = 1$

Donde TP, FP y FN son el número de positivos verdaderos, positivos falsos y negativos falsos, respectivamente, obtenidos por el clasificador.

Al cálculo de F_1^j le aplicamos el suavizado de *Laplace* con un valor $\varepsilon = 0.05$. Como se explicó antes estos valores se utilizan para ajustar en el proceso anterior las evidencias de los documentos en las clases.

Con este proceso incremental de creación de una colección de entrenamiento podemos elegir iterativamente los 50 mejores documentos para etiquetar. En nuestra evaluación comparamos estas estrategias de AA entre sí y, además, incluimos una alternativa aleatoria en la que se toman siempre 50 documentos al azar.

4.4. Experimentos

Los experimentos se han realizado con la misma colección definida en el punto 3.2 del capítulo 3. La asignación de códigos CIE-9-MC a un episodio clínico tiene los siguientes elementos importantes ya definidos:

- El diagnóstico principal (DxP) es la enfermedad que tras su estudio y en el momento del alta, el médico que atendió al paciente establece como causa del ingreso.
- Los diagnósticos secundarios (DxS) se consideran aquellas enfermedades que coexisten con el DxP en el momento del ingreso o que se han desarrollado durante la estancia hospitalaria y que han influido en la duración del ingreso.

La asignación de códigos CIE-9-MC es un problema multietiqueta, pero SVM se diseñó para realizar clasificación binaria. Es posible resolver los problemas de entornos SVM multiclase, basándose en la combinación de clasificadores binarios. Estas técnicas descomponen el problema multiclase en múltiples problemas binarios. Las dos principales alternativas que se aplican en la literatura para utilizar SVM cuando el número de clases, c, es superior a dos, son *1-vs-todos* y *1-vs-1*. Si añadimos el factor de trabajar con una colección con un elevado número de clases, lo más recomendable es utilizar *1-vs-todos*.

Las métricas de evaluación requieren la existencia de un *gold standard*, como ya definimos en el capitulo 3. En nuestro caso, conocemos los códigos correctos por cada documento de test, porque cada documento de entrenamiento o test tiene una lista de clases asignada por el médico codificador. Por supuesto, la lista de los códigos

asociados a los documentos de test sólo se usa para fines de evaluación. Adoptamos las mismas métricas del capítulo 3, que se han utilizado en la evaluación de los clasificadores de documentación clínica [Larkey y Croft, 1995], [Larkey y Croft, 1996]:

- *Top candidato:* proporción de casos en los que el código principal es el principal candidato (*top 1*) propuesto por el sistema automático.
- *Top 10:* proporción de casos en los que el código principal está en los primeros 10 candidatos producidos por el sistema.
- *Recall 15 y recall 20:* Nivel de recall en los primeros 15 o 20 candidatos.

Se recurre a una cadena de tres letras para definir cada una de las variantes de AA. Por ejemplo, CAN es la combinación de elegir *Confianza (C)* para la dimensión "evidencia", *Avg(A)* para la dimensión "clase" y (*N)* indica que no se utiliza la dimensión "peso". Con los distintos tipos de dimensiones posibles obtenemos 12 combinaciones posibles, que generan otros tantos experimentos.

A los 12 experimentos de AA, tenemos que añadir un nuevo experimento, por el cual obtendremos de forma aleatoria los documentos que vamos incorporando a la colección de entrenamiento. De este modo podemos comparar el rendimiento entre los modelos de *AA* y el modelo aleatorio (que supone una técnica base realista ya que es lo que ocurre en los hospitales pues no tienen criterios específicos para codificar episodios).

Los resultados de *Top candidato* para todos los experimentos se muestran en la tabla 4.1. Los resultados para *Top 10*, *Recall 15* y *Recall 20* se pueden consultar en el Anexo A. Se representa en negrita el mejor resultado de *Top candidato* para cada tamaño del conjunto de entrenamiento.

Los experimentos nos demuestran que con AA mejoramos *Top candidato*, en relación a una selección aleatoria a lo largo de todo el proceso. Tras incorporar 1000 documentos obtenemos mediante CAN un incremento del 34,7%, en comparación con el método aleatorio.

No existen mayores diferencias entre los distintos métodos de AA pero, en todo caso, CAN, SAN y SMW resultan ligeramente más robustos. Los resultados nos manifiestan que la dimensión de "peso", con pocos documentos en la colección de entrenamiento, obtiene los mejores resultados (por ejemplo, SMW y SRW). En cambio, a medida que

el número de documentos de la colección de entrenamiento se incrementa, los resultados más favorables están en los experimentos que utilizan la dimensión de clase *Avg(A)* sin necesidad de realizar pesado por clases, como lo demuestran los resultados de SAN y CAN.

#Docs.	CMN	SMN	CAN	SAN	CRN	SRN	CMW	SMW	CAW	SAW	CRW	SRW	Aleatorio
100	6.52	6.52	6.52	6.52	6.52	6.52	6.52	6.52	6.52	6.52	6.52	6.52	6.52
150	8.07	8.07	6.52	7.76	8.07	8.07	8.07	**8.39**	6.83	7.14	8.07	**8.39**	7.76
200	9.94	9.32	7.45	8.70	9.94	9.32	8.07	**10.25**	8.07	8.39	8.07	**10.25**	6.83
250	**11.49**	9.94	9.32	9.94	**11.49**	9.94	9.94	9.32	7.45	10.25	9.94	9.32	10.87
300	11.18	**11.80**	11.49	12.42	11.18	**11.80**	10.87	**11.80**	11.49	**11.80**	10.87	**11.80**	10.87
350	12.11	14.29	12.42	13.04	12.11	14.29	11.18	13.98	**14.60**	13.66	11.18	13.98	12.11
400	13.04	**15.22**	13.04	13.04	13.04	**15.22**	13.04	14.29	13.98	13.98	12.73	14.29	11.80
450	13.66	13.98	13.66	14.91	13.66	13.98	12.42	**15.53**	14.60	13.98	12.11	**15.53**	13.04
500	14.60	13.35	14.91	15.22	14.60	13.35	**15.53**	14.91	**15.53**	15.22	15.22	14.91	12.42
550	14.91	15.84	15.53	16.15	14.91	15.84	16.46	**16.77**	16.46	16.15	16.15	**16.77**	11.80
600	14.60	16.77	17.08	**17.70**	14.60	16.77	15.84	17.39	16.77	16.46	15.53	17.39	12.73
650	16.46	16.46	17.70	**20.50**	16.46	16.46	17.70	16.46	17.39	16.77	17.39	16.46	12.42
700	16.46	16.46	18.32	**18.94**	16.46	16.46	17.70	17.08	17.70	16.77	17.39	17.08	12.73
750	15.84	17.08	19.25	**19.57**	15.84	17.08	15.84	17.39	18.63	16.77	15.53	17.39	12.42
800	16.15	16.46	19.25	**20.50**	16.15	16.46	17.39	17.39	18.32	17.70	17.08	17.39	13.66
850	16.15	16.46	18.63	**19.57**	16.15	16.46	17.08	18.01	19.25	16.77	17.08	18.01	14.60
900	17.70	17.08	**19.57**	18.32	17.70	17.08	18.01	18.63	18.94	15.84	18.01	18.63	15.53
950	18.32	17.39	**19.57**	17.70	18.32	17.39	18.94	18.32	18.63	16.77	18.94	18.32	14.60
1000	18.94	17.08	**20.50**	18.32	18.94	17.08	18.63	17.39	18.94	17.70	18.63	17.39	15.22

Tabla 4.1: Resultados *Top candidato*

Con pocos documentos en la colección de entrenamiento, SMW (Puntuación – Máximo – Peso) funciona mejor que las otras combinaciones, en cambio al aumentar el número de documentos la dimensión *Avg(A)* nos facilita los mejores valores para cualquiera de las dimensiones de "evidencia", *Puntuación (S*) o *Confianza (C).* Podemos deducir, que con pocos documentos, aquellos que pertenecen claramente a una clase, son los más representativos para el clasificador. En cambio, a medida que el número de documentos se incrementa, la media de los valores obtenidos para todas las clases, es más informativo para el clasificador. No podemos contrastar directamente estos resultados con otros experimentos [Esuli y Sebastiani, 2009] de AA con colecciones multiclase, ya que las métricas utilizadas no son las mismas y las colecciones son diferentes. Para nuestra colección la dimensión de peso no funciona bien cuando se combina con la dimensión de clase *Avg(A),* en cambio consigue mejores resultados con la combinación de las otras dimensiones. Las figuras 4.1, 4.2,

4.3 y 4.4 nos muestra la evolución de las métricas en función del número de documentos que vamos incorporando a la colección de entrenamiento para los experimentos SAN, CAN, SMW y aleatorio.

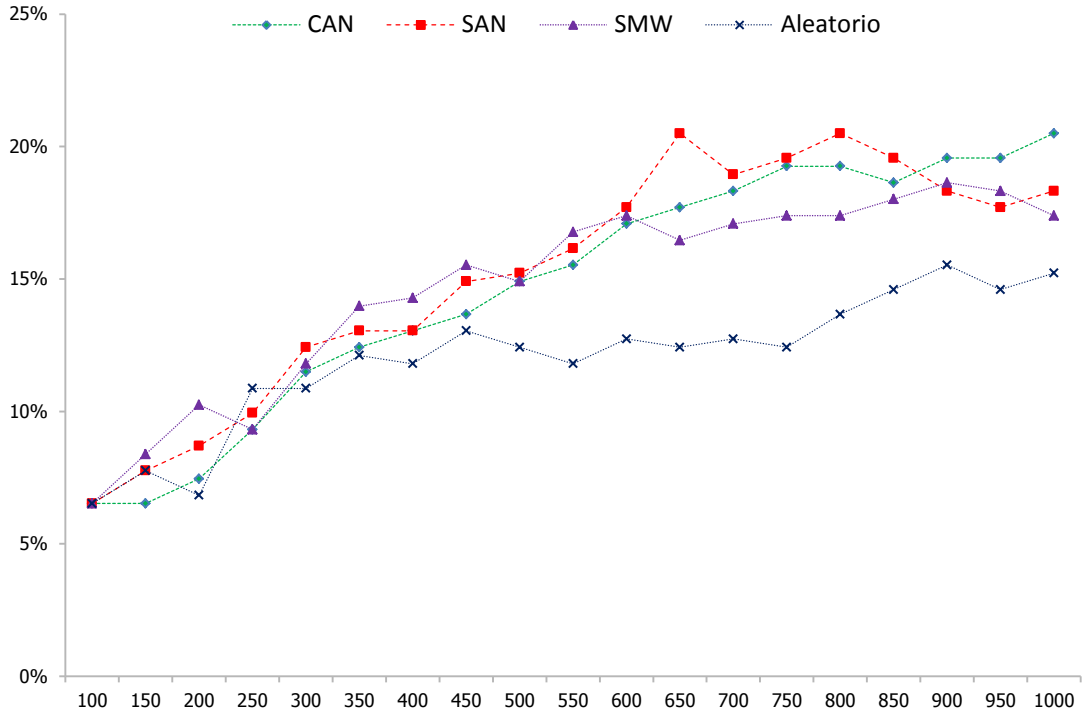

Figura 4.1**:** *Top candidato*

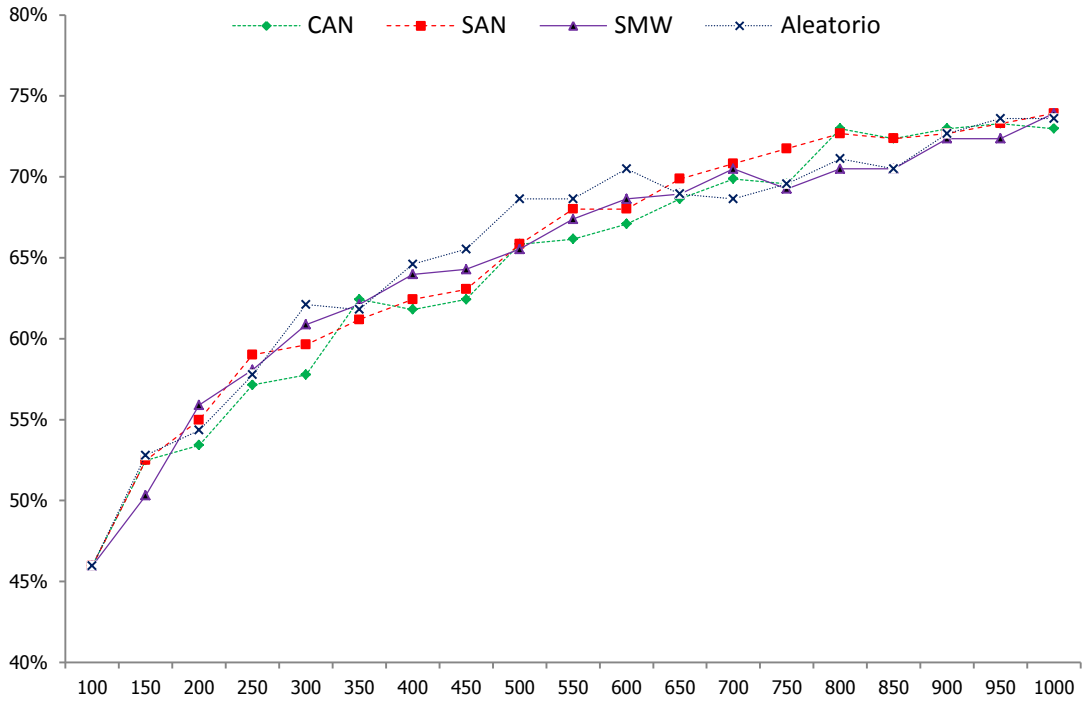

Figura 4.2: *Top 10*

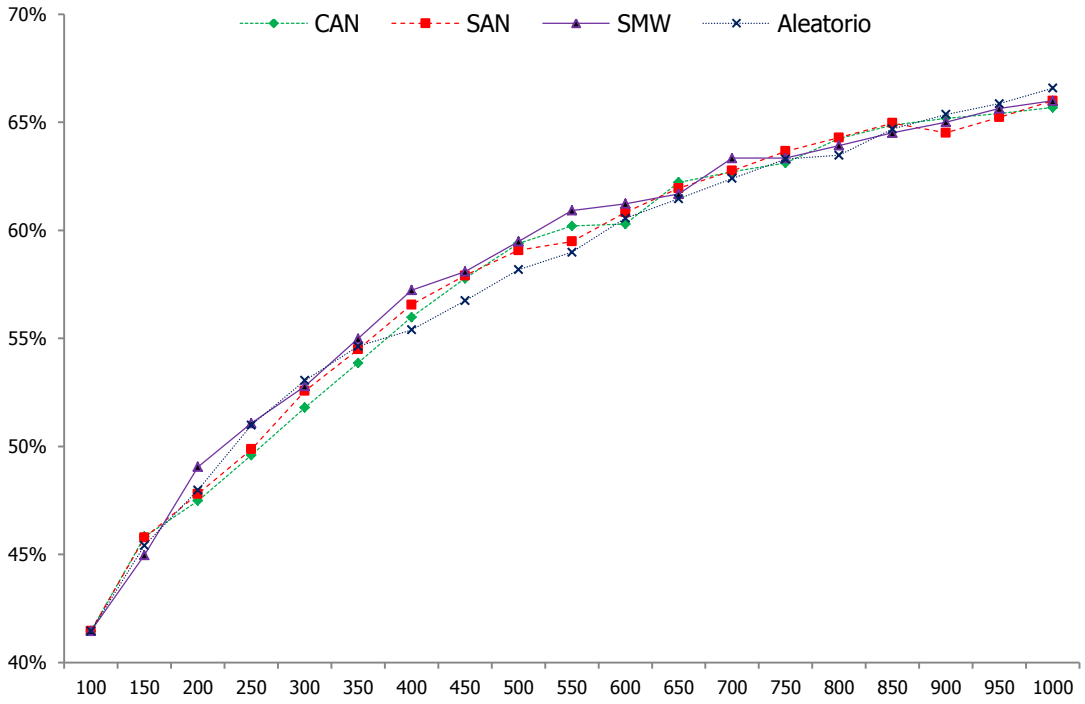

Figura 4.3: *Recall 15*

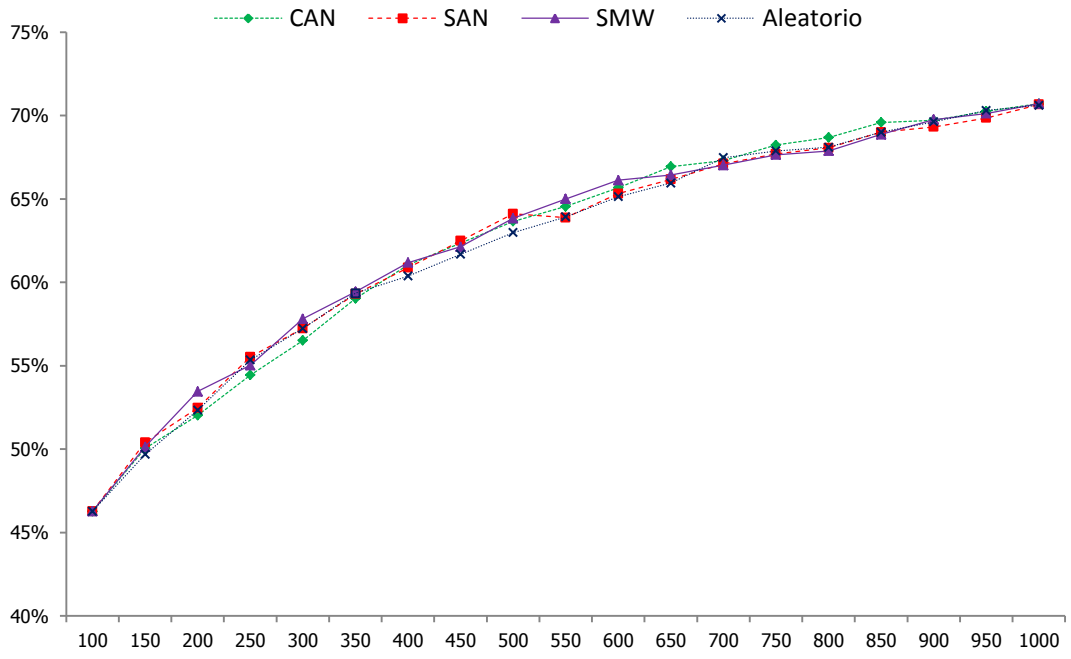

Figura 4.4: *Recall 20*

En una clasificación multiclase donde el clasificador produce un ranking de clases por cada documento, una de las métricas que se suelen utilizar es la siguiente:

Mean Average Precision (MAP), calcula la precisión cada vez que se agrega un TP, resultando ser el promedio de las precisiones medias calculadas para cada uno de los documentos de la colección de test. Si Q es la colección de test, y si las clases relevantes de un documento $q_j \in Q$ es $\{c_1, c_2, \cdots, c_{m_j}\}$ y si R_{jk} es el conjunto de ranking de resultados desde el mejor puesto hasta llegar a la clase c_k, entonces

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\pi(R_{jk}) \tag{4.2}$$

Con esta métrica se pretende valorar si el ranking que produce una colección obtenida con técnicas de AA mejora en relación a otra seleccionada aleatoriamente. Se calcula MAP para un ranking con diferentes números de códigos CIE-9-MC recuperados. Los valores seleccionados para las distintas cantidades de códigos recuperados que forman el ranking son 5, 10, 15 y 20. Los resultados de MAP con las técnicas de AA SAN, CAN y SMW tomando como referencia (*baseline*) el modelo aleatorio se muestran en las figuras 4.5, 4.6, 4.7 y 4.8 para los diferentes números de códigos recuperados. Los resultados de MAP para todos los métodos AA se pueden consultar en el Anexo B.

Los resultados nos muestran que aplicar AA con CAN, SMW y SRW mejora los resultados de MAP en relación a una selección aleatoria. Los métodos de AA SMW y SRW son superiores cuando la colección tiene pocos documentos. A medida que la colección crece, lo deseable es utilizar CAN donde muestra un funcionamiento sobresaliente.

El problema de clases desbalanceadas en una clasificación de textos ocurre cuando el número de documentos que pertenece a cada clase es muy diferente. Esto provoca que los clasificadores tengan mayor exactitud para clasificar la clase mayoritaria pero una menor exactitud predictiva sobre las clases minoritarias. El clasificador intenta reducir el error global, de forma que el error de clasificación no tiene en cuenta la distribución de los documentos.

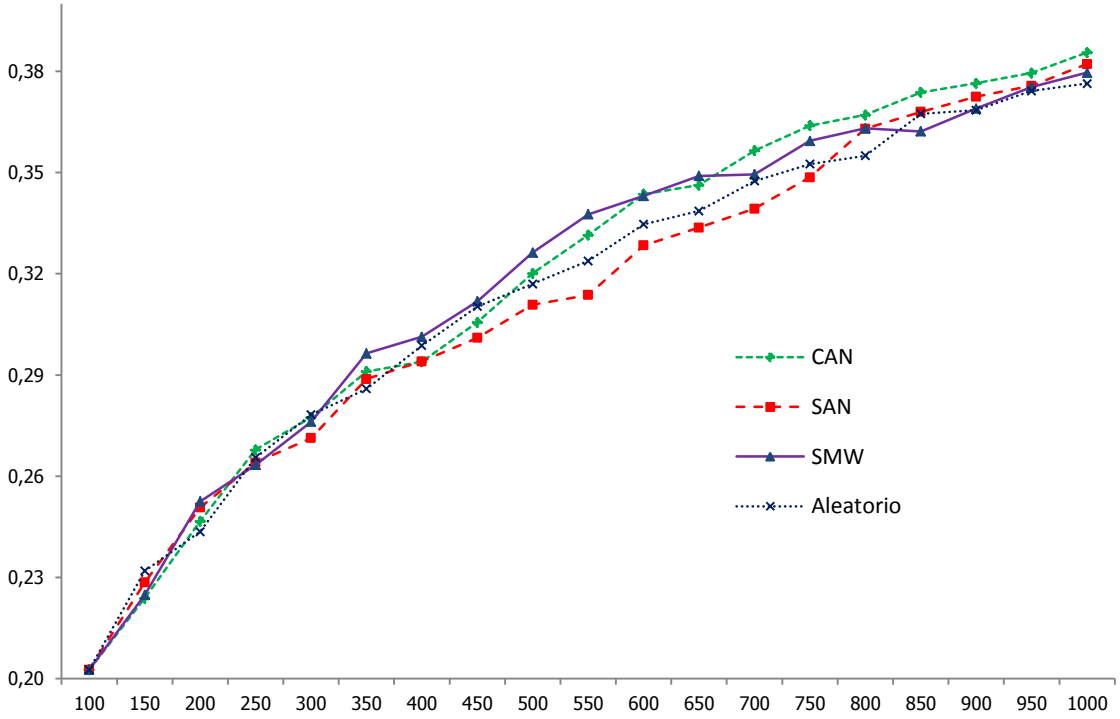

Figura 4.5: Resultados de MAP 5

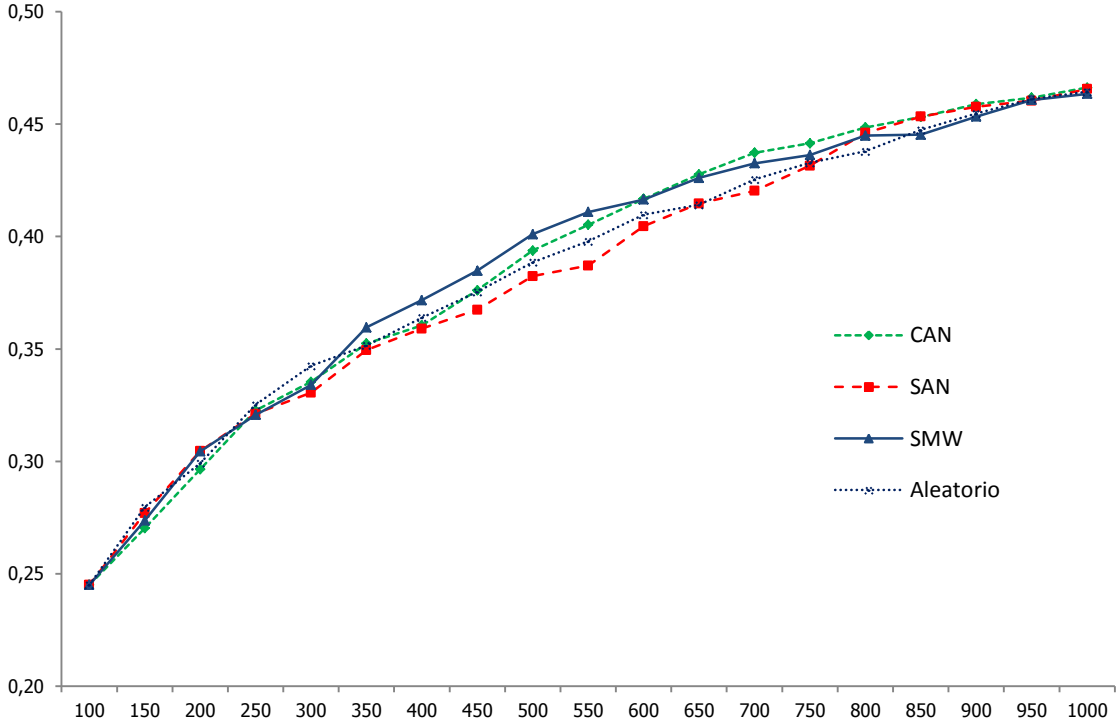

Figura 4.6: Resultados de MAP 10

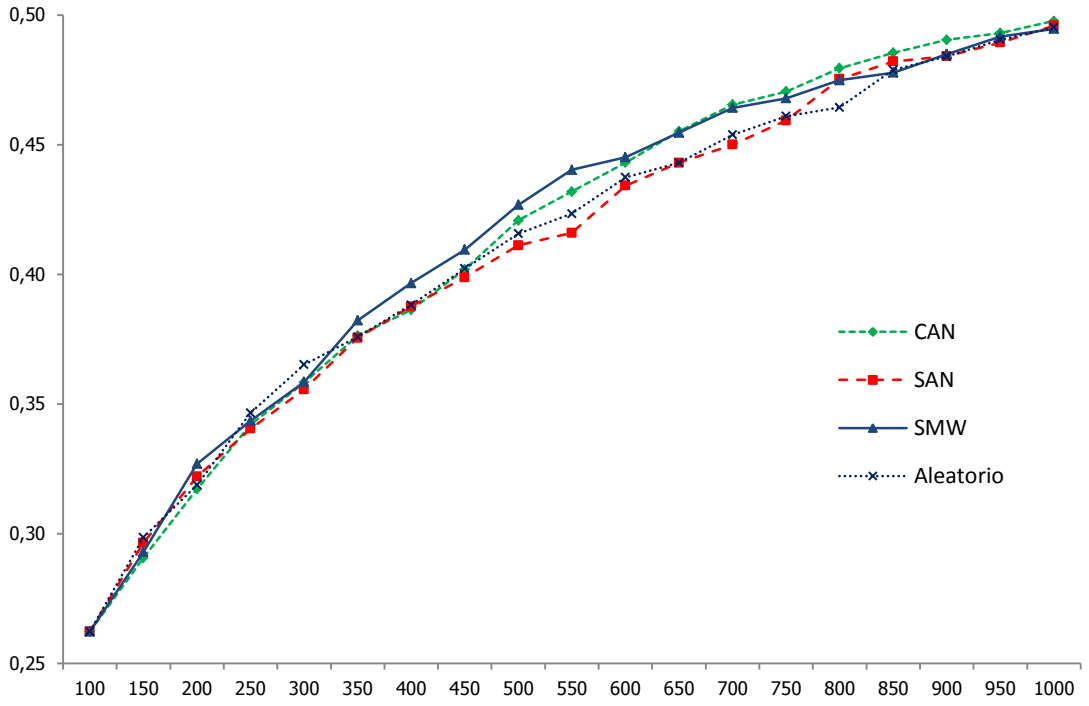

Figura 4.7: Resultados de MAP 15

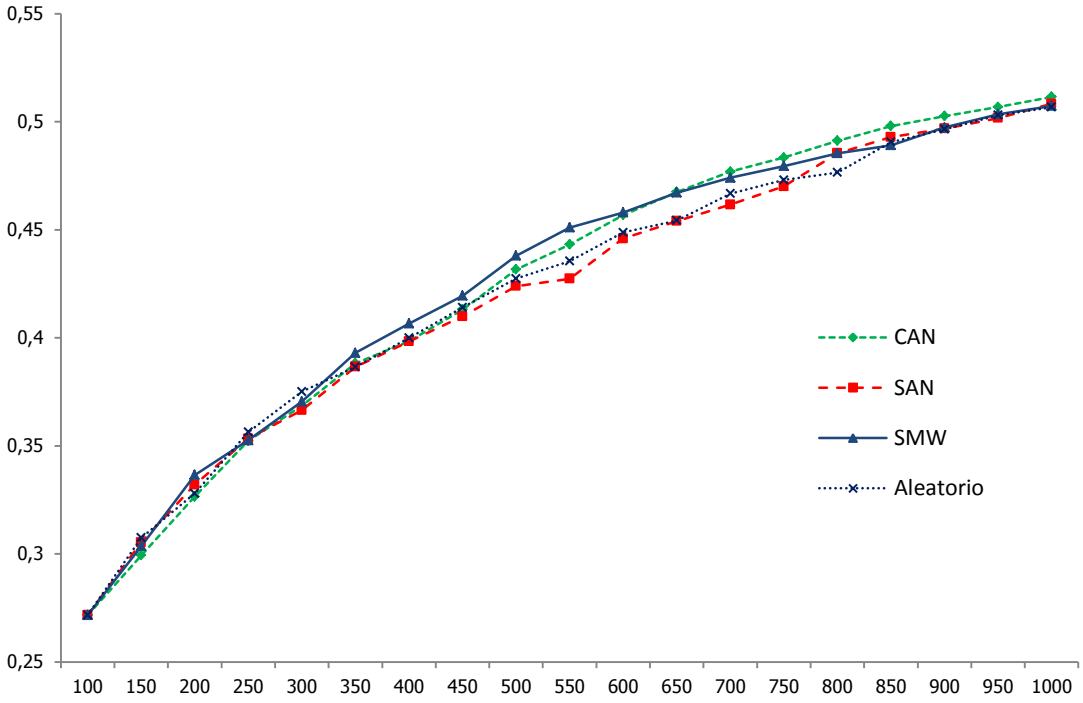

Figura 4.8: Resultados de MAP 20

Hay varias líneas de investigación en el campo de las colecciones desbalanceadas. Dentro de las distintas propuestas para resolver la clasificación con colecciones desbalanceadas está AA. En la literatura se ha discutido sobre la utilización de AA en colecciones desbalanceadas [Provost, 2001], [Abe, 2003], [Ertekin et al., 2007] y [Ertekin et al., 2007].

La colección que se utiliza en esta tesis es desbalanceada. Esto se acentúa con la utilización de clasificadores binarios, como SVM. Se intenta resolver problemas de clasificación multiclase, donde generalmente se adopta la estrategia de *1-vs-todos*, como en nuestro caso. El clasificador se entrena para cada clase (datos positivos); el resto de los datos formarían el conjunto de ejemplos de la otra clase (datos negativos). El modelo de clasificación con SVM está diseñado para minimizar el error que se comete sobre la colección de entrenamiento. El clasificador intentará clasificar los ejemplos negativos correctamente, ya que estos son una gran mayoría en relación a los ejemplos positivos.

# docs	CMN	SMN	CAN	SAN	CRN	SRN	CMW	SMW	CAW	SAW	CRW	SRW	Aleatorio
100	302	302	302	302	302	302	302	302	302	302	302	302	302
150	**393**	382	380	381	393	382	390	381	385	387	390	381	389
200	**486**	464	457	457	**486**	464	474	458	444	460	474	458	458
250	**537**	527	519	528	**537**	527	552	529	513	530	552	529	502
300	587	575	573	581	587	575	**598**	586	571	587	**598**	586	547
350	624	633	623	**658**	624	633	646	640	618	655	646	640	604
400	683	673	670	**705**	683	673	685	678	662	696	685	678	645
450	727	716	709	**744**	727	716	723	721	706	742	723	721	692
500	766	754	758	**781**	766	754	756	756	738	780	756	756	725
550	799	797	800	816	799	797	791	798	774	**828**	791	798	760
600	826	816	838	845	826	816	840	822	815	**857**	840	822	794
650	860	854	873	867	860	854	861	860	853	**875**	861	860	839
700	895	884	895	898	895	884	887	891	890	**906**	887	891	872
750	928	917	932	922	928	917	918	911	929	**932**	918	911	915
800	950	934	956	946	950	934	940	935	**964**	960	940	935	951
850	961	958	969	965	961	958	962	962	**991**	984	962	962	982
900	989	996	1001	996	989	996	992	986	**1024**	1002	992	986	1015
950	1006	1021	1025	1019	1006	1021	1023	1009	**1038**	1036	1023	1009	1032

Tabla 4.2: Número de códigos CIE-9-MC en cada colección y para cada modelo

Se analiza y se valora ahora la captación de códigos CIE-9-MC (clases) para AA y su correlación con el desbalanceo. En la tabla 4.2 se expone el número de códigos CIE-9-

MC para cada método de AA y el método aleatorio en cada colección. En la figura 4.9 se representan los datos de la tabla 4.2. En los resultados se observa que con AA se capta más códigos CIE-9-MC, esta tendencia empieza a ser favorable para el método aleatorio cuando se aproxima al número total de códigos CIE-9-MC de la colección y solo el método CAW mantiene su superioridad. Esto significa que las aproximaciones AA funcionan razonablemente bien en cuanto a ir trayendo más códigos nuevos a la colección de entrenamiento.

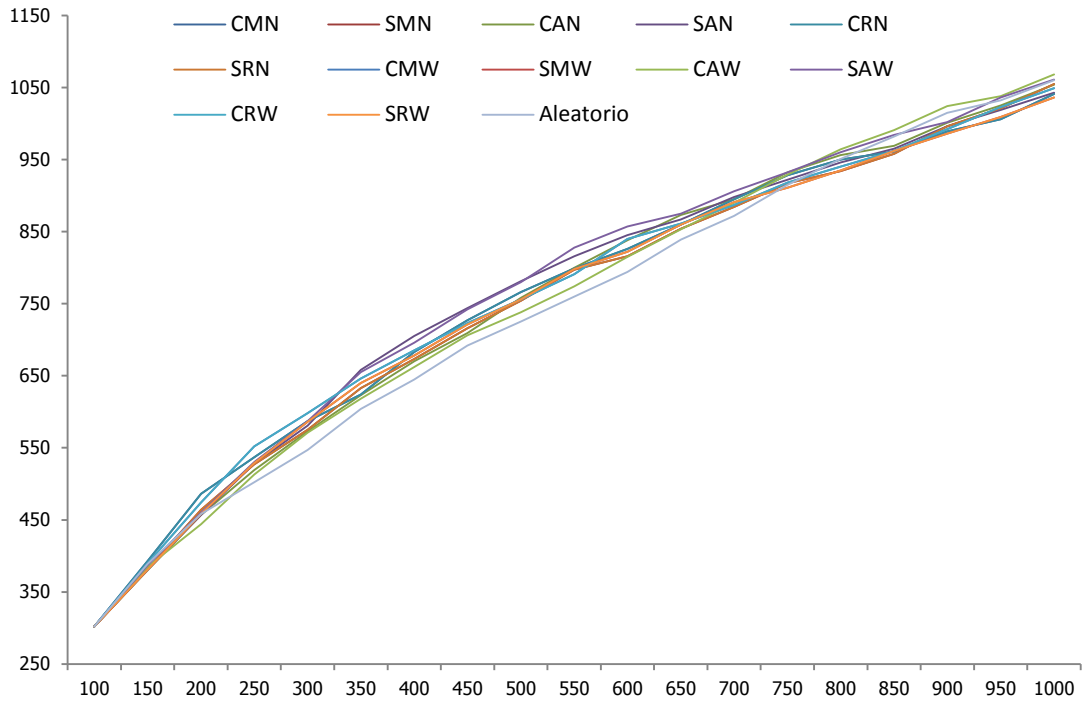

Figura 4.9: Captación de códigos CIE-9-MC

4.5. Conclusiones

En los experimentos de aprendizaje activo realizados con todas las estrategias, observamos que solo *Top candidato* mejora claramente con respecto a una selección aleatoria. Para las otras métricas *Top 10*, *Recall 15* y *Recall 20* no se aprecia una mejora clara. Una de las posibles explicaciones es que las métricas *Top 10*, *Recall 15* y *Recall 20* están caracterizadas por definir un rango de posibles códigos válidos sin importarle el lugar que estos ocupan dentro del ranking. Esto se demuestra con los

resultados de MAP para diferentes valores, en donde son siempre superiores a los resultados de MAP para una colección aleatoria. El aprendizaje activo mejora la posición en el rango de posibles códigos correctos, pero esta mejora no se transmite en el valor absoluto de estas métricas. Esto lo demuestra también en los resultados de *Top candidato* y *Top 10, y*a que, mejorando los códigos correctos que ocupan la primera posición (*Top candidato),* este ascenso de posición del código correcto en el ranking, no se refleja en *Top 10*.

Una de las dimensiones que favorece las técnicas de AA es la mejora en el desbalanceo. Como se ha podido demostrar AA descubre más códigos CIE-9-MC que el método aleatorio, lo cual permite que las colecciones que construye sean menos desbalanceadas.

Destacar que las técnicas AA con la dimensión "Peso" se comportan mejor en colecciones con pocos documentos, en cambio en colecciones con un mayor número de documentos la dimensión "Clase" con la aproximación Promedio (*Avg*) y sin peso realiza un progreso considerable.

En un trabajo futuro consideraremos también la posibilidad de utilizar otros clasificadores que sean más interpretables (en la línea de que el médico codificador pueda comprender mejor las sugerencias del sistema).

Capítulo 5

Conclusiones e investigaciones futuras

Este capítulo recopila las conclusiones de esta tesis y algunas de las sugerencias para futuras investigaciones.

5.1. Conclusiones

Esta tesis investiga sobre la clasificación automática de la documentación clínica, centrándonos en los informes de alta de hospitalización y su codificación con códigos CIE-9-MC. Lo que se pretende es investigar las posibilidades que nos ofrece CAT en un entorno tan complejo como la documentación clínica, con la finalidad de construir un sistema de ayuda a la codificación de informes de alta de hospitalización u otro tipo de documentación clínica, que se pueda implementar y valorar en un centro sanitario. Este es un reto dentro de la investigación en CAT, por su dificultad y por su necesidad, al aumentar sustancialmente cada día la documentación clínica en formato electrónico que se genera en un centro sanitario sin clasificar y por su universalidad, al ser utilizado en la gran mayoría de centros sanitarios.

Las propuestas que han sido analizadas y estudiadas en esta tesis van desde la creación de la colección, las diferentes representaciones de la colección, técnicas de clasificación, métricas de evaluación de la eficacia, técnicas de aprendizaje activo para crear un conjunto de entrenamiento que mejore el clasificador y a su vez valorar el nivel de desbalanceo.

- Las innovaciones de esta tesis empiezan por la creación de una colección en lengua castellana con la cual realizar los experimentos. La gran mayoría de los estudios en CAT para CIE-9-MC, se han desarrollado sobre colecciones en lengua inglesa.

- Otro desafío es la dificultad que tiene la codificación CIE-9-MC, y esta se incrementa al aspirar automatizarla o semiautomatizarla. La complejidad de este proceso se manifiesta en que todas las administraciones sanitarias publican boletines en donde se aclaran o marcan criterios de codificación CIE-9-MC y a su vez de forma continua instruir a su personal con cursos de formación especializados. Como ejemplo de esta problemática, en la colección CCHMC, que está codificada por tres expertos, usualmente hay disparidad de criterio entre los mismos y es necesario recurrir a seleccionar el código que al menos dos expertos hayan señalado (esto es, por mayoría de votos).

- Hemos estudiado diferentes representaciones de la colección, distintos modelos de recuperación y el efecto de los sistemas de pesado en la asignación de códigos CIE-9-MC de los experimentos. Dentro de las representaciones de los documentos consideramos original aquella en donde realizamos una expansión de los documentos de entrenamiento con las descripciones de los códigos CIE-9-MC.

- Los experimentos nos indican que SVM funciona mejor que *Knn* en casi todos los ámbitos. Excepto en la métrica Top candidato en donde *Knn* con un peso mayor que 1 en el diagnóstico principal consigue mejores resultados que SVM. Esto demuestra la potencia de aprendizaje que se consigue con SVM.

- Los resultados nos manifiestan que las métricas calculadas con *microaveraging* son muy superiores a las obtenidas con *macroaveraging*. Lo que indica que hay clases en donde el clasificador no funciona bien, como lo demuestran los gráficos de la sección 3.7 y 3.8. Para SVM hay más clases en que el clasificador funciona bien, en comparación con *Knn*. Lo que nos indica que *Knn* es más susceptible al desbalanceo de clases que SVM.

- El tipo de episodios clínicos que genera los informes de alta la convierten en una colección de documentos muy difícil, como ya se ha detallado en la sección 3.1. Con técnicas de Aprendizaje Activo (AA) se eligen los informes de alta que se deben etiquetar con códigos CIE-9-MC, con el fin de crear un nuevo conjunto de entrenamiento que mejore el clasificador. Los resultados de

los experimentos en AA demuestran que nuestra estrategia es satisfactoria para este tipo de sistemas. En los experimentos de aprendizaje activo solo *Top candidato* mejora claramente con respecto a una selección aleatoria. Para las otras métricas *Top 10, Recall 15* y *Recall 20* no se aprecia una mejora clara. Una de las posibles explicaciones es que las métricas *Top 10, Recall 15* y *Recall 20* están caracterizadas por definir un rango de posibles códigos válidos sin importarle el lugar que estos ocupan dentro del ranking. Con los cálculos de MAP para diferentes valores, en donde son siempre superiores las colecciones de AA que una colección aleatoria, demuestran que AA mejora la posición en el rango de posibles códigos correctos. Este ascenso no se transmite en el valor absoluto de *Top 10, Recall 15* y *Recall 20*. Esto lo demuestra también en los resultados de *Top candidato* y *Top 10*, ya que, mejorando los códigos correctos que ocupan la primera posición (*Top candidato)*, este ascenso de posición del código correcto en el ranking, no se refleja en *Top 10*.

- Con las técnicas de AA se obtienen colecciones de entrenamiento menos desbalanceadas. AA descubre más códigos CIE-9-MC que el método aleatorio, esta característica en si misma es un avance en el desbalanceo.

- Las técnicas AA con la dimensión "Peso" se comportan mejor en colecciones con pocos documentos. Cuando en las colecciones aumenta el número de documentos, la combinación de la dimensión "Clase" con la aproximación Promedio (*Avg*) y sin la dimensión "Peso" los resultados son superiores.

- El rendimiento obtenido en los experimentos es suficientemente bueno para construir una herramienta evaluable en el trabajo real dentro de un centro sanitario. Esto lo certifica resultados como el de *Recall 20* con *microaveraging* en categorías para los clasificadores SVM en donde los aciertos son superiores al 80%. Hay estudios que señalan un 20% de errores en la codificación manual realizada por médicos codificadores. A esto hay que añadir el estudio de la sección 1.6, en donde aparecen siempre más diagnósticos codificados que diagnósticos descritos por el médico especialista que firma el informe de alta, sin olvidarnos que está colección es una de las de mayor dificultad que nos podemos encontrar en un hospital. Todo parece indicar que los sistemas de

clasificación automática de códigos CIE-9-MC para informes de alta hospitalaria pueden ser una alternativa fiable, cómoda y estable a la codificación manual.

5.2. Investigaciones futuras

Las posibles investigaciones se pueden enfocar en varias líneas. Una de ellas es intentar evitar que la colección esté desbalanceada, procurando que el número de documentos de una clase no sea muy superior al de las otras. En esta tesis ya se han utilizado AA con técnica para mejorar el desbalanceo con resultados positivos. Dentro de este ámbito se podrían utilizar técnicas a nivel de datos o nivel de algoritmos para seleccionar solo aquellos documentos más informativos para la colección de entrenamiento, intentado conseguir una colección lo menos desbalanceada posible que mejore el clasificador. También se puede evaluar la creación de documentos artificialmente para aquellas clases minoritarias o que no existen en la colección de entrenamiento.

Dentro de SVM, también podemos considerar la utilización de otras técnicas de clasificación multietiqueta, o la búsqueda de un núcleo que facilite el sesgo en este tipo de clasificación. Y si esto es inevitable se puede intentar solucionar este problema mediante factores de penalización distintos para cada clase, lo que permite ajustar el coste de los falsos positivos y de los falsos negativos de manera independiente.

En un trabajo futuro consideraremos también la posibilidad de utilizar otros clasificadores que sean más interpretables (en la línea de que el médico codificador pueda comprender mejor las sugerencias del sistema).

La codificación automática de códigos CIE-9-MC está básicamente enfocada al ámbito clínico asistencial. Pero hay otros entornos menos críticos en donde puede ser aplicada. Nos estamos refiriendo a la gestión económica, a nuevas formas de medir la producción asistencial que conlleve evaluar adecuadamente la actividad y poder desenvolver mecanismos más avanzados. En estos casos se podrían utilizar clasificadores totalmente automáticos, como contrapunto a los clasificadores para la codificación de enfermedades asistenciales en donde nuestra propuesta se desarrolla con la visión de una herramienta de ayuda.

ANEXO A

Resultados y figuras de la métricas *Top 10*, *Recall 15* y *Recall 20* en cada colección creada para los diferentes métodos AA y aleatorio. Se representa en negrita el mejor resultado para cada tamaño del conjunto de entrenamiento.

#Docs	CMN	SMN	CAN	SAN	CRN	SRN	CMW	SMW	CAW	SAW	CRW	SRW	Aleatorio
100	45,96%	45,96%	45,96%	45,96%	45,96%	45,96%	45,96%	45,96%	45,96%	45,96%	45,96%	45,96%	45,96%
150	52,17%	52,48%	52,48%	52,48%	52,17%	52,48%	52,17%	50,31%	51,86%	**53,42%**	51,86%	50,31%	52,80%
200	53,11%	56,21%	53,42%	54,97%	53,11%	56,21%	56,52%	55,90%	56,21%	54,66%	**56,52%**	55,90%	54,35%
250	54,04%	57,45%	57,14%	59,01%	54,04%	57,45%	**60,25%**	58,07%	58,70%	57,45%	**60,25%**	58,07%	57,76%
300	58,70%	61,80%	57,76%	59,63%	58,70%	61,80%	60,87%	60,87%	59,94%	55,59%	60,87%	60,87%	**62,11%**
350	60,87%	62,42%	62,42%	61,18%	60,87%	62,42%	60,87%	62,11%	**63,04%**	61,18%	60,87%	62,11%	61,80%
400	61,18%	63,35%	61,80%	62,42%	61,18%	63,35%	61,49%	63,98%	62,73%	62,73%	61,49%	63,98%	**64,60%**
450	59,94%	**65,84%**	62,42%	63,04%	59,94%	**65,84%**	63,35%	64,29%	64,60%	64,91%	63,35%	64,29%	65,53%
500	61,49%	66,15%	65,84%	65,84%	61,49%	66,15%	64,91%	65,53%	66,15%	64,91%	64,91%	65,53%	**68,63%**
550	63,98%	67,70%	66,15%	68,01%	63,98%	67,70%	66,77%	67,39%	66,15%	66,77%	66,46%	67,39%	**68,63%**
600	66,46%	68,63%	67,08%	68,01%	66,46%	68,63%	67,70%	68,63%	65,84%	67,70%	67,70%	68,63%	**70,50%**
650	66,77%	70,50%	68,63%	69,88%	66,77%	**70,50%**	67,70%	68,94%	67,70%	68,63%	67,70%	68,94%	68,94%
700	67,70%	69,88%	69,88%	**70,81%**	67,70%	69,88%	69,25%	70,50%	68,63%	68,94%	69,25%	70,50%	68,63%
750	69,25%	69,57%	69,57%	**71,74%**	69,25%	69,57%	68,32%	69,25%	70,19%	69,88%	68,01%	69,25%	69,57%
800	69,88%	71,12%	**72,98%**	72,67%	69,88%	71,12%	70,19%	70,50%	71,12%	71,12%	69,88%	70,50%	71,12%
850	70,81%	71,12%	**72,36%**	**72,36%**	70,81%	71,12%	69,88%	70,50%	70,81%	72,36%	69,57%	70,50%	70,50%
900	72,05%	71,74%	72,98%	72,67%	72,05%	71,74%	71,43%	72,36%	72,36%	**73,29%**	71,43%	72,36%	72,67%
950	72,36%	72,67%	73,29%	73,29%	72,36%	72,67%	71,74%	72,36%	72,36%	72,67%	71,74%	72,36%	**73,60%**
1000	72,67%	72,98%	72,98%	**73,91%**	72,67%	72,98%	72,05%	**73,91%**	74,22%	73,60%	72,05%	**73,91%**	73,60%

Tabla A.1: Resultados *Top 10*

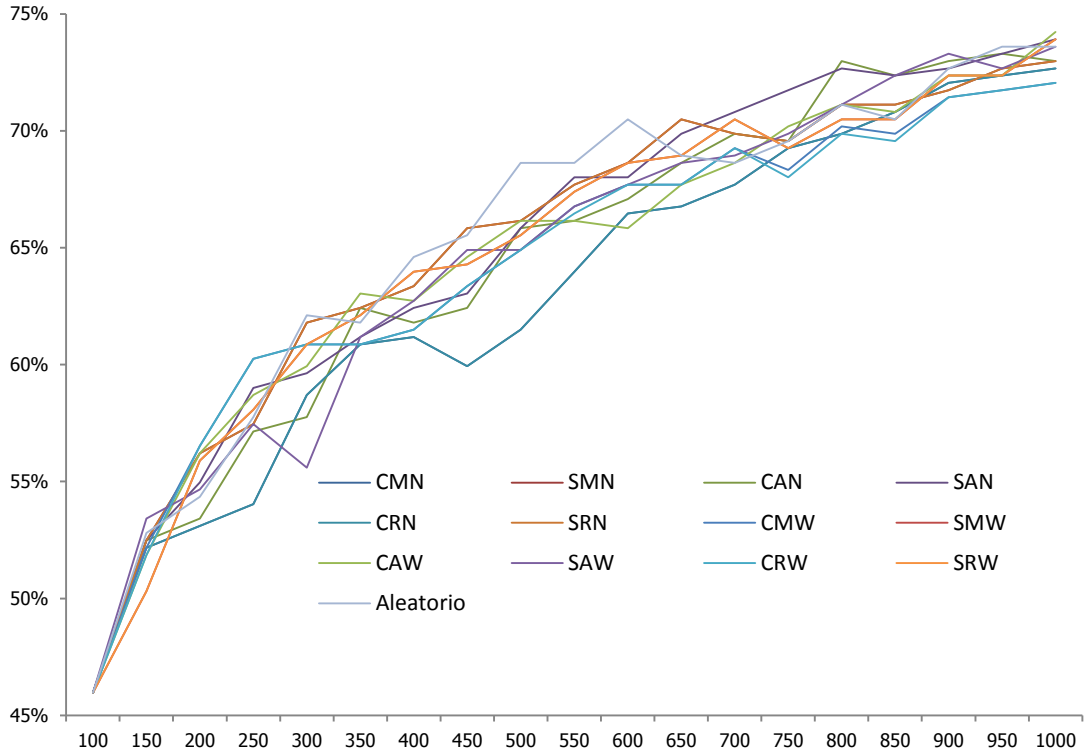

Figura A.1: Resultados *Top 10*

#Docs	CMN	SMN	CAN	SAN	CRN	SRN	CMW	SMW	CAW	SAW	CRW	SRW	Aleatorio
100	41,46%	41,46%	41,46%	41,46%	41,46%	41,46%	41,51%	41,46%	41,46%	41,46%	41,46%	41,46%	41,46%
150	45,06%	44,97%	**45,82%**	45,78%	45,06%	44,97%	45,51%	44,97%	45,60%	45,78%	45,51%	44,97%	45,42%
200	47,62%	47,89%	47,48%	47,80%	47,62%	47,89%	47,21%	49,06%	**49,28%**	48,07%	47,21%	49,06%	47,98%
250	49,78%	49,60%	49,60%	49,87%	49,78%	49,60%	50,72%	**51,08%**	49,87%	49,87%	50,67%	51,08%	50,99%
300	51,17%	51,48%	51,80%	52,56%	51,17%	51,48%	52,11%	52,79%	51,75%	51,35%	52,11%	52,79%	**53,05%**
350	52,34%	52,96%	53,86%	54,49%	52,34%	52,96%	54,49%	**54,99%**	53,68%	54,04%	54,49%	**54,99%**	54,63%
400	54,49%	55,35%	55,97%	56,56%	54,49%	55,35%	55,44%	**57,23%**	55,12%	54,85%	55,44%	**57,23%**	55,39%
450	56,15%	57,01%	57,77%	57,91%	56,15%	57,01%	58,00%	**58,09%**	56,92%	57,46%	58,00%	**58,09%**	56,74%
500	56,87%	58,09%	59,39%	59,07%	56,87%	58,09%	**59,61%**	59,48%	58,72%	58,89%	**59,61%**	59,48%	58,18%
550	57,68%	59,39%	60,20%	59,48%	57,68%	59,39%	60,06%	**60,92%**	59,52%	59,39%	60,06%	**60,92%**	58,98%
600	59,70%	60,02%	60,29%	60,83%	59,70%	60,02%	**61,23%**	**61,23%**	60,06%	60,83%	**61,23%**	**61,23%**	60,56%
650	60,69%	61,05%	**62,22%**	61,95%	60,69%	61,05%	61,90%	61,68%	61,37%	61,46%	61,90%	61,68%	61,46%
700	62,58%	62,08%	62,71%	62,76%	62,58%	62,08%	62,94%	63,34%	62,08%	62,35%	**62,98%**	63,34%	62,40%
750	62,98%	63,43%	63,12%	**63,66%**	62,98%	63,43%	63,39%	63,34%	62,31%	62,89%	63,43%	63,34%	63,30%
800	63,79%	63,88%	64,24%	**64,29%**	63,79%	63,88%	63,79%	63,93%	63,12%	63,97%	63,84%	63,93%	63,48%
850	64,56%	64,42%	64,87%	**64,96%**	64,56%	64,42%	63,97%	64,51%	63,70%	64,65%	63,97%	64,51%	64,69%
900	64,65%	**65,41%**	65,18%	64,51%	64,65%	65,41%	65,18%	65,00%	64,24%	64,87%	65,18%	65,00%	65,36%
950	65,36%	65,32%	65,41%	65,23%	65,36%	65,32%	65,36%	65,63%	65,05%	65,23%	65,36%	65,63%	**65,86%**
1000	65,90%	65,77%	65,68%	65,99%	65,90%	65,77%	65,59%	65,99%	65,72%	65,86%	65,59%	65,99%	**66,58%**

Tabla A.2: Resultados *Recall 15*

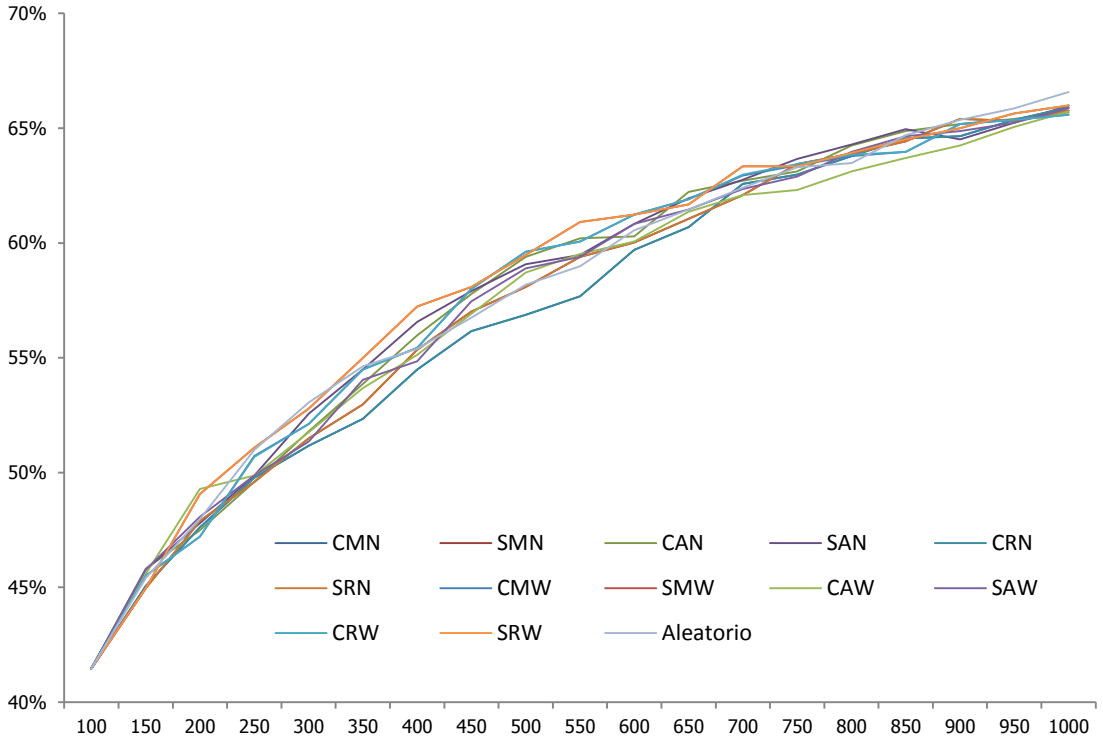

Figura A.2: Resultados *Recall 15*

#Docs	CMN	SMN	CAN	SAN	CRN	SRN	CMW	SMW	CAW	SAW	CRW	SRW	Aleatorio
100	46,27%	46,27%	46,27%	46,27%	46,27%	46,27%	46,23%	46,27%	46,27%	46,27%	46,27%	46,27%	46,27%
150	49,55%	49,42%	50,04%	50,40%	49,55%	49,42%	49,82%	50,13%	49,91%	**50,49%**	49,78%	50,13%	49,69%
200	52,38%	52,88%	52,02%	52,47%	52,38%	52,88%	52,29%	53,46%	**53,95%**	52,79%	52,16%	53,46%	52,34%
250	54,72%	54,49%	54,45%	55,53%	**54,72%**	54,49%	55,30%	55,03%	55,08%	54,58%	55,35%	55,03%	55,35%
300	56,06%	56,65%	56,51%	57,23%	56,06%	56,65%	56,87%	**57,82%**	56,24%	55,75%	56,96%	57,82%	57,23%
350	57,41%	58,67%	59,03%	59,30%	57,41%	58,67%	**59,52%**	59,43%	58,09%	59,07%	59,48%	59,43%	59,34%
400	58,76%	60,02%	61,05%	60,87%	58,76%	60,02%	60,11%	**61,19%**	59,84%	60,38%	60,11%	61,19%	60,38%
450	61,28%	61,90%	62,35%	62,49%	61,28%	61,90%	**62,76%**	62,13%	61,59%	62,17%	**62,76%**	62,13%	61,68%
500	62,31%	62,67%	63,66%	**64,11%**	62,31%	62,67%	64,02%	63,84%	63,21%	63,79%	63,97%	63,84%	62,98%
550	63,25%	64,33%	64,56%	63,88%	63,25%	64,33%	65,59%	65,00%	64,02%	64,65%	**65,59%**	65,00%	63,93%
600	64,69%	65,09%	65,68%	65,32%	64,69%	65,09%	**66,17%**	66,13%	64,65%	65,36%	**66,17%**	66,13%	65,14%
650	66,04%	65,59%	66,94%	66,17%	66,04%	65,59%	**67,03%**	66,44%	65,86%	65,81%	66,98%	66,44%	65,95%
700	67,34%	66,80%	67,30%	67,12%	67,34%	66,80%	**67,57%**	67,03%	66,71%	67,12%	67,52%	67,03%	67,48%
750	67,48%	**68,28%**	68,24%	67,70%	67,48%	68,28%	67,61%	67,65%	67,48%	67,34%	67,57%	67,65%	67,88%
800	**68,73%**	68,33%	68,69%	68,06%	**68,73%**	68,33%	68,06%	67,88%	68,19%	67,88%	68,06%	67,88%	68,10%
850	68,96%	68,87%	**69,59%**	69,00%	68,96%	68,87%	68,46%	68,87%	68,73%	68,87%	68,46%	68,87%	69,00%
900	69,45%	69,68%	**69,72%**	69,32%	69,45%	69,68%	69,18%	69,77%	69,14%	69,05%	69,23%	69,77%	69,63%
950	69,81%	70,04%	70,26%	69,86%	69,81%	70,04%	70,04%	70,13%	70,08%	69,68%	70,04%	70,13%	**70,31%**
1000	70,35%	70,53%	70,71%	**70,66%**	70,35%	70,53%	70,22%	70,71%	70,49%	70,31%	70,22%	70,71%	70,62%

Tabla A.3: Resultados *Recall 20*

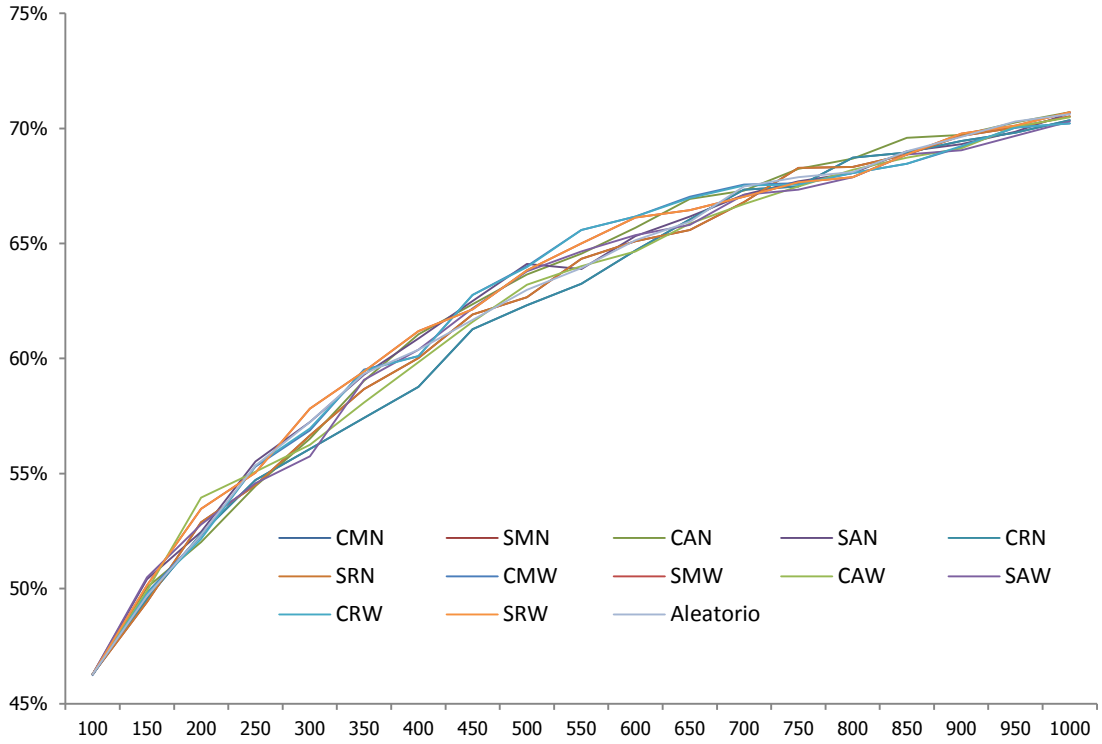

Figura A.3: Resultados *Recall 20*

ANEXO B

Resultados para MAP en el modelo aleatorio y cada uno de los modelos AA en cada colección y para distinto número de códigos recuperados. Las figuras muestran los valores de las tablas.

Se representa en negrita el mejor resultado para cada tamaño del conjunto de entrenamiento.

#Docs	CMN	SMN	CAN	SAN	CRN	SRN	CMW	SMW	CAW	SAW	CRW	SRW	Aleatorio
100	0,2026	0,2026	0,2026	0,2026	0,2026	0,2026	0,2026	0,2026	0,2026	0,2026	0,2026	0,2026	0,2026
150	0,2294	0,2287	0,2238	0,2285	0,2294	0,2287	**0,2357**	0,2248	0,2242	0,2266	0,2357	0,2248	0,2319
200	0,2490	0,2469	0,2466	0,2507	0,2490	0,2469	0,2504	**0,2525**	0,2444	0,2490	0,2504	**0,2525**	0,2435
250	0,2567	0,2618	0,2678	0,2640	0,2567	0,2618	0,2681	0,2633	0,2608	0,2645	**0,2681**	0,2633	0,2655
300	0,2741	0,2796	0,2774	0,2713	0,2741	0,2796	0,2800	0,2761	**0,2828**	0,2722	0,2800	0,2761	0,2782
350	0,2882	**0,2988**	0,2910	0,2888	0,2882	**0,2988**	0,2852	0,2964	0,2988	**0,2899**	0,2852	0,2964	0,2859
400	0,3079	0,3049	0,2939	0,2940	**0,3079**	0,3049	0,3014	0,3013	0,3020	0,2972	0,3014	0,3013	0,2987
450	**0,3185**	0,3117	0,3056	0,3010	**0,3185**	0,3117	0,3080	0,3118	0,3076	0,3086	0,3080	0,3118	0,3103
500	0,3221	0,3139	0,3201	0,3108	0,3221	0,3139	0,3176	**0,3263**	0,3194	0,3152	0,3176	**0,3263**	0,3170
550	0,3273	0,3252	0,3315	0,3137	0,3273	0,3252	0,3240	**0,3376**	0,3256	0,3200	0,3240	**0,3376**	0,3238
600	0,3339	0,3336	**0,3436**	0,3284	0,3339	0,3336	0,3277	0,3431	0,3329	0,3320	0,3277	0,3431	0,3347
650	0,3404	0,3423	0,3463	0,3337	0,3404	0,3423	0,3404	**0,3490**	0,3375	0,3337	0,3404	**0,3490**	0,3386
700	0,3450	0,3511	**0,3565**	0,3393	0,3450	0,3511	0,3478	0,3495	0,3432	0,3424	0,3478	0,3495	0,3475
750	0,3537	0,3521	**0,3639**	0,3486	0,3537	0,3521	0,3543	0,3594	0,3542	0,3500	0,3543	0,3594	0,3525
800	0,3559	0,3562	**0,3671**	0,3630	0,3559	0,3562	0,3571	0,3631	0,3594	0,3585	0,3571	0,3631	0,3550
850	0,3632	0,3585	**0,3738**	0,3680	0,3632	0,3585	0,3645	0,3622	0,3617	0,3620	0,3645	0,3622	0,3674
900	0,3675	0,3650	**0,3765**	0,3725	0,3675	0,3650	0,3692	0,3690	0,3754	0,3699	0,3692	0,3690	0,3685
950	0,3734	0,3659	**0,3795**	0,3757	0,3734	0,3659	0,3738	0,3754	0,3783	0,3743	0,3738	0,3754	0,3742
1000	0,3789	0,3735	**0,3856**	0,3822	0,3789	0,3735	0,3790	0,3796	0,3802	0,3799	0,3790	0,3796	0,3764

Tabla B.1: Resultados MAP 5

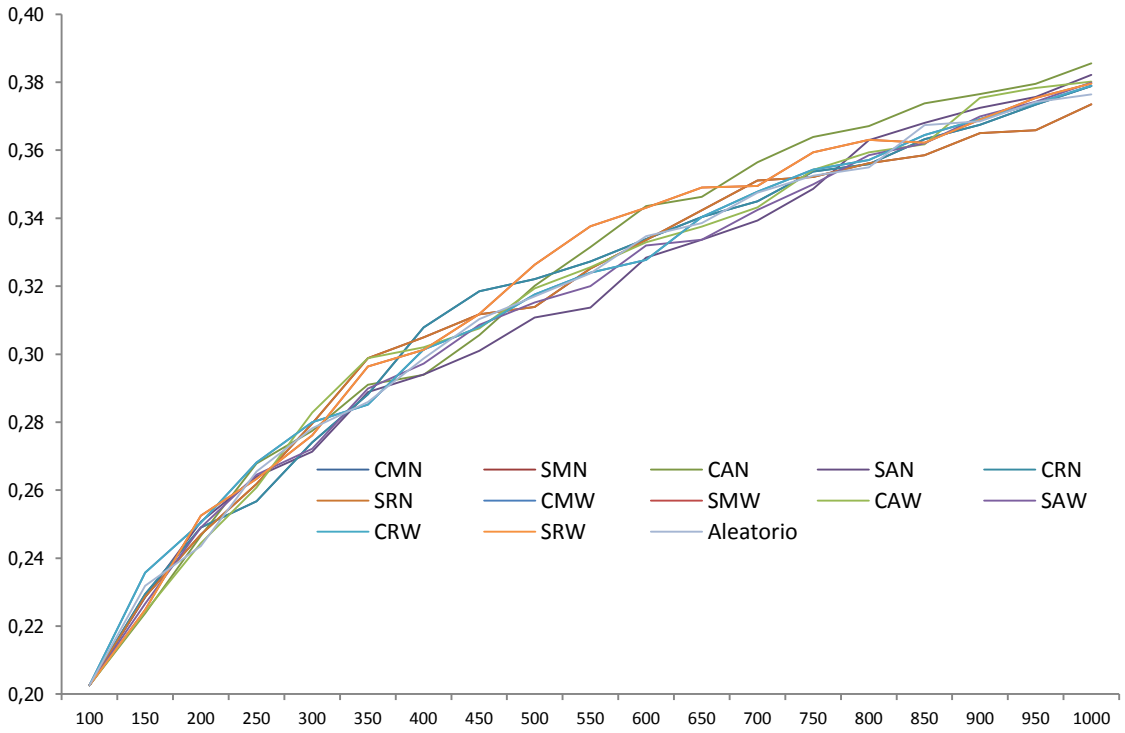

Figura B.1: Resultados MAP 5

# docs	CMN	SMN	CAN	SAN	CRN	SRN	CMW	SMW	CAW	SAW	CRW	SRW	Aleatorio
100	0,2450	0,2450	0,2450	0,2450	0,2450	0,2450	0,2450	0,2450	0,2450	0,2450	0,2450	0,2450	0,2450
150	0,2763	0,2757	0,2702	0,2769	0,2763	0,2757	**0,2828**	0,2736	0,2714	0,2762	**0,2828**	0,2736	0,2795
200	0,3007	0,2977	0,2964	**0,3045**	0,3007	0,2977	0,2997	0,3044	0,2993	0,3009	0,2997	0,3044	0,2991
250	0,3111	0,3151	0,3226	0,3211	0,3111	0,3151	0,3248	0,3206	0,3164	0,3189	0,3248	0,3206	**0,3251**
300	0,3288	0,3364	0,3354	0,3305	0,3288	0,3364	0,3363	0,3340	**0,3380**	0,3302	0,3363	0,3340	0,3425
350	0,3455	0,3557	0,3524	0,3495	0,3455	0,3557	0,3495	**0,3596**	0,3559	0,3526	0,3495	**0,3596**	0,3515
400	0,3654	0,3666	0,3605	0,3590	0,3654	0,3666	0,3652	**0,3717**	0,3639	0,3638	0,3652	**0,3717**	0,3639
450	0,3821	0,3800	0,3761	0,3674	0,3821	0,3800	0,3748	**0,3848**	0,3770	0,3745	0,3748	**0,3848**	0,3754
500	0,3875	0,3826	0,3938	0,3824	0,3875	0,3826	0,3918	**0,4010**	0,3907	0,3836	0,3918	**0,4010**	0,3886
550	0,3954	0,3952	0,4052	0,3871	0,3954	0,3952	0,3977	**0,4109**	0,3972	0,3934	0,3977	**0,4109**	0,3979
600	0,4043	0,4076	**0,4168**	0,4046	0,4043	0,4076	0,4020	0,4165	0,4037	0,4060	0,4020	0,4165	0,4097
650	0,4152	0,4163	**0,4277**	0,4147	0,4152	0,4163	0,4140	0,4261	0,4140	0,4146	0,4140	0,4261	0,4141
700	0,4217	0,4247	**0,4373**	0,4203	0,4217	0,4247	0,4224	0,4326	0,4189	0,4230	0,4224	0,4326	0,4254
750	0,4325	0,4279	**0,4415**	0,4315	0,4325	0,4279	0,4338	0,4363	0,4323	0,4324	0,4338	0,4363	0,4328
800	0,4383	0,4355	**0,4485**	0,4461	0,4383	0,4355	0,4377	0,4449	0,4395	0,4402	0,4377	0,4449	0,4379
850	0,4464	0,4384	0,4532	**0,4534**	0,4464	0,4384	0,4445	0,4453	0,4473	0,4472	0,4445	0,4453	0,4475
900	0,4505	0,4467	**0,4589**	0,4577	0,4505	0,4467	0,4536	0,4534	0,4572	0,4526	0,4536	0,4534	0,4547
950	0,4602	0,4511	**0,4618**	0,4604	0,4602	0,4511	0,4613	0,4608	0,4615	0,4588	0,4613	0,4608	0,4612
1000	0,4657	0,4587	**0,4663**	0,4656	0,4657	0,4587	0,4660	0,4635	0,4656	0,4609	0,4660	0,4635	0,4644

Tabla B.2: Resultados MAP 10

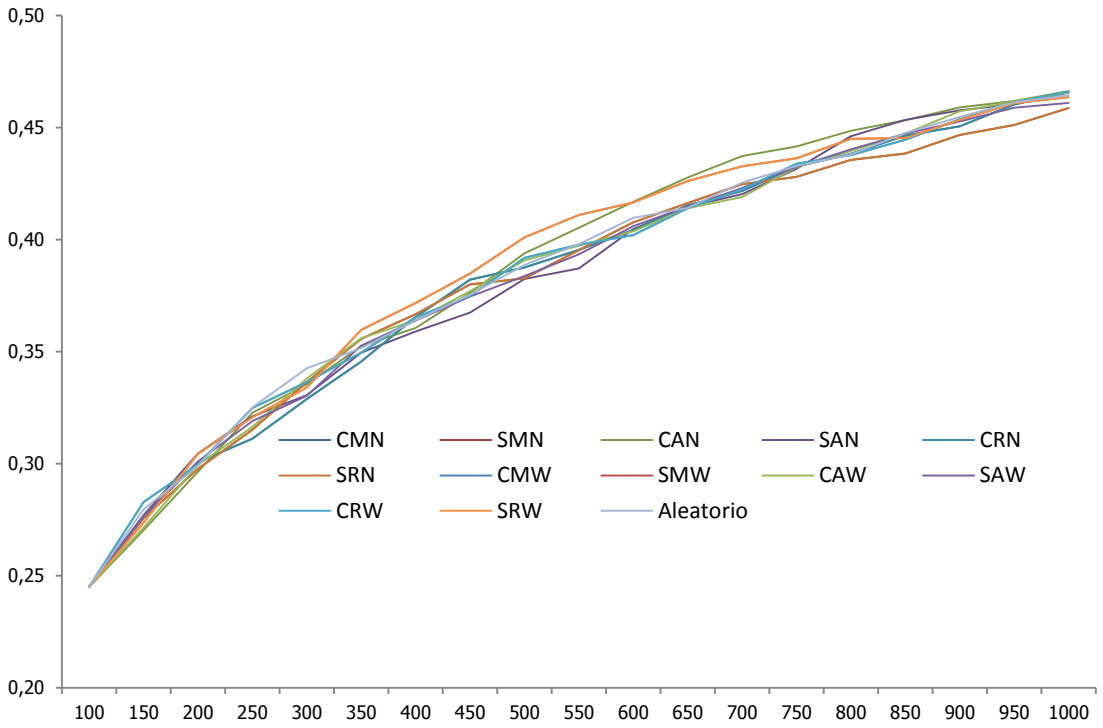

Figura B.2: Resultados MAP 10

# docs	CMN	SMN	CAN	SAN	CRN	SRN	CMW	SMW	CAW	SAW	CRW	SRW	Aleatorio
100	0,2623	0,2623	0,2623	0,2623	0,2623	0,2623	0,2623	0,2623	0,2623	0,2623	0,2623	0,2623	0,2623
150	0,2930	0,2944	0,2907	0,2965	0,2930	0,2944	**0,3014**	0,2930	0,2905	0,2960	**0,3014**	0,2930	0,2986
200	0,3187	0,3173	0,3172	0,3220	0,3187	0,3173	0,3181	**0,3269**	0,3212	0,3212	0,3181	**0,3269**	0,3188
250	0,3325	0,3353	0,3421	0,3405	0,3325	0,3353	0,3447	0,3435	0,3361	0,3404	0,3447	0,3435	**0,3466**
300	0,3507	0,3581	0,3583	0,3556	0,3507	0,3581	0,3581	0,3585	0,3614	0,3528	0,3581	0,3585	**0,3652**
350	0,3666	0,3770	0,3763	0,3755	0,3666	0,3770	0,3714	**0,3822**	**0,3802**	0,3758	0,3714	**0,3822**	0,3759
400	0,3894	0,3907	0,3863	0,3876	0,3894	0,3907	0,3886	**0,3966**	0,3889	0,3860	0,3886	**0,3966**	0,3882
450	0,4076	0,4049	0,4019	0,3988	0,4076	0,4049	0,4032	**0,4095**	0,4019	0,4012	0,4032	**0,4095**	0,4022
500	0,4136	0,4100	0,4208	0,4112	0,4136	0,4100	0,4202	**0,4268**	0,4193	0,4130	0,4202	**0,4268**	0,4157
550	0,4205	0,4241	0,4319	0,4160	0,4205	0,4241	0,4254	**0,4403**	0,4256	0,4207	0,4254	**0,4403**	0,4234
600	0,4311	0,4342	0,4431	0,4342	0,4311	0,4342	0,4336	**0,4452**	0,4333	0,4355	0,4336	**0,4452**	0,4374
650	0,4431	0,4447	**0,4552**	0,4430	0,4431	0,4447	0,4443	0,4546	0,4444	0,4418	0,4443	0,4546	0,4430
700	0,4533	0,4550	**0,4654**	0,4500	0,4533	0,4550	0,4539	0,4642	0,4505	0,4519	0,4539	0,4642	0,4538
750	0,4627	0,4609	**0,4705**	0,4593	0,4627	0,4609	0,4637	0,4679	0,4608	0,4601	0,4637	0,4679	0,4610
800	0,4695	0,4671	**0,4795**	0,4753	0,4695	0,4671	0,4692	0,4749	0,4684	0,4704	0,4692	0,4749	0,4644
850	0,4765	0,4701	**0,4854**	0,4821	0,4765	0,4701	0,4747	0,4778	0,4759	0,4770	0,4747	0,4778	0,4788
900	0,4795	0,4806	**0,4904**	0,4841	0,4795	0,4806	0,4838	0,4849	0,4861	0,4824	0,4838	0,4849	0,4840
950	0,4886	0,4819	**0,4931**	0,4894	0,4886	0,4819	0,4896	0,4917	0,4914	0,4890	0,4896	0,4917	0,4905
1000	0,4969	0,4898	**0,4977**	0,4959	0,4969	0,4898	0,4933	0,4947	0,4965	0,4930	0,4933	0,4947	0,4952

Tabla B.3: Resultados MAP 15

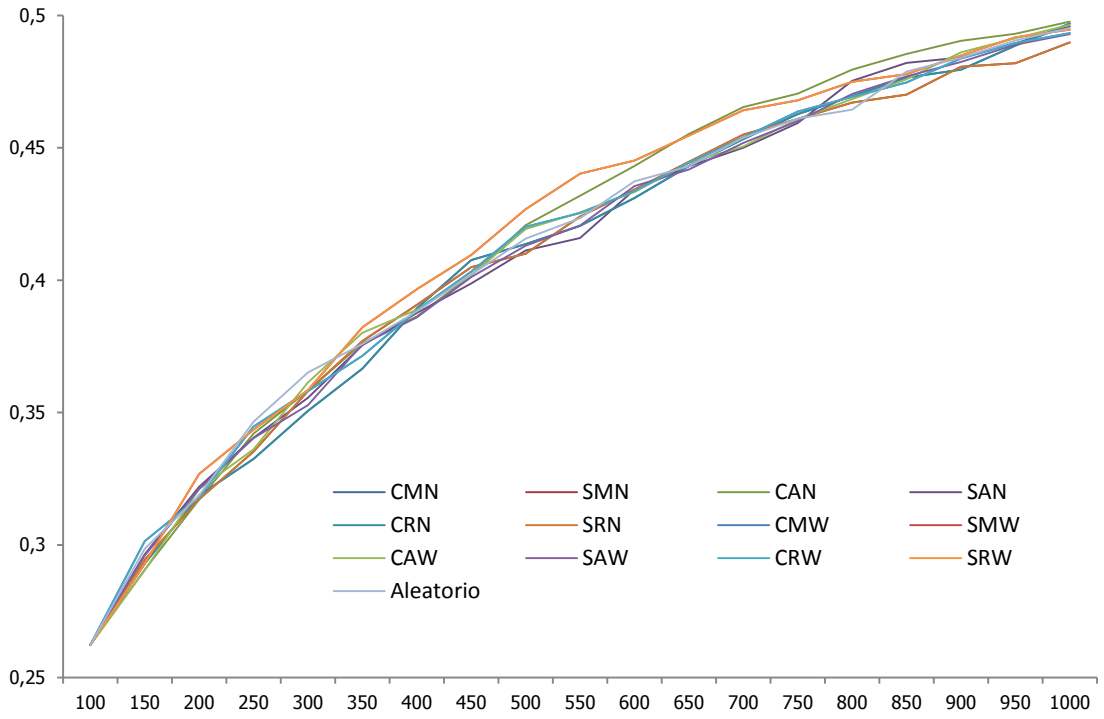

Figura B.3: Resultados MAP 15

# docs	CMN	SMN	CAN	SAN	CRN	SRN	CMW	SMW	CAW	SAW	CRW	SRW	Aleatorio
100	0,2716	0,2716	0,2716	0,2716	0,2716	0,2716	0,2716	0,2716	0,2716	0,2716	0,2716	0,2716	0,2716
150	0,3027	0,304	0,2993	0,3053	0,3027	0,304	**0,3103**	0,3036	0,2997	0,305	**0,3103**	0,3036	0,3075
200	0,3288	0,3283	0,3265	0,332	0,3288	0,3283	0,3288	**0,3365**	0,3313	0,3314	0,3288	**0,3365**	0,328
250	0,3437	0,3461	0,3523	0,3533	0,3437	0,3461	0,3551	0,3526	0,3477	0,3504	0,3551	0,3526	**0,3564**
300	0,3622	0,3696	0,3686	0,3665	0,3622	0,3696	0,3695	0,3705	0,3713	0,3621	0,3695	0,3705	**0,3751**
350	0,378	0,3902	0,3882	0,3866	0,378	0,3902	0,3842	**0,393**	0,3904	0,3875	0,3842	**0,393**	0,3865
400	0,3998	0,4022	0,3984	0,3983	0,3998	0,4022	0,3997	**0,4066**	0,4001	0,3993	0,3997	**0,4066**	0,3999
450	0,4202	0,4171	0,4131	0,4099	0,4202	0,4171	0,4147	**0,4195**	0,4134	0,4134	0,4147	**0,4195**	0,4141
500	0,4267	0,4216	0,4316	0,4239	0,4267	0,4216	0,4311	**0,438**	0,4306	0,4255	0,4311	**0,438**	0,4274
550	0,4342	0,4363	0,4432	0,4273	0,4342	0,4363	0,4394	**0,451**	0,437	0,4342	0,4394	**0,451**	0,4355
600	0,4434	0,4471	0,4568	0,4459	0,4434	0,4471	0,4462	**0,458**	0,4446	0,4475	0,4462	**0,458**	0,4487
650	0,4566	0,4567	**0,4675**	0,454	0,4566	0,4567	0,4576	0,4672	0,4558	0,4532	0,4576	0,4672	0,4543
700	0,4653	0,4676	**0,4769**	0,4616	0,4653	0,4676	0,4656	0,4741	0,4623	0,4639	0,4656	0,4741	0,4667
750	0,4743	0,4738	**0,4834**	0,4701	0,4743	0,4738	0,4742	0,4795	0,4742	0,4714	0,4742	0,4795	0,473
800	0,4824	0,4787	**0,4912**	0,4854	0,4824	0,4787	0,4802	0,4854	0,482	0,4801	0,4802	0,4854	0,4765
850	0,4882	0,4823	**0,498**	0,4929	0,4882	0,4823	0,4862	0,4891	0,4897	0,4878	0,4862	0,4891	0,4906
900	0,4922	0,4918	**0,5026**	0,4968	0,4922	0,4918	0,4942	0,4973	0,499	0,4936	0,4942	0,4973	0,4965
950	0,5001	0,4949	**0,5068**	0,5017	0,5001	0,4949	0,5017	0,5035	0,5052	0,5002	0,5017	0,5035	0,503
1000	0,5082	0,5025	**0,5115**	0,5083	0,5082	0,5025	0,5056	0,5073	0,5098	0,5049	0,5056	0,5073	0,5067

Tabla B.4: Resultados MAP 20

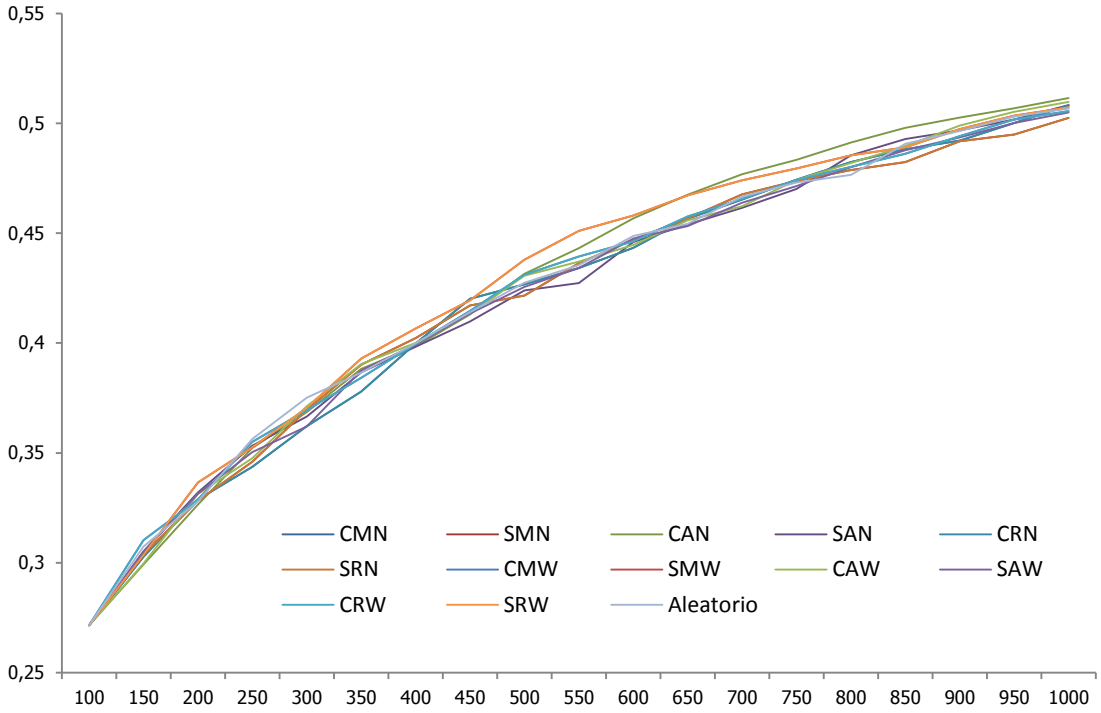

Figura B.4: Resultados MAP 20

ANEXO C. Palabras vacías

a
acá
ahí
ajena
ajenas
ajeno
ajenos
al
algo
alguna
algunas
alguno
algunos
algún
allá
allí
aquel
aquella
aquellas
aquello
aquellos
aquí
b
c
cada
cierta
ciertas
cierto
ciertos
como
cómo
con
conmigo
consigo
contigo
cualquier
cualquiera
cualquieras
cuan
cuanta
cuantas
cuánta
cuántas
cuanto
cuantos
cuán
cuánto
cuántos
d
de
dejar
del
demasiada
demasiadas
demasiado
demasiados
demás
e
el
ella
ellas
ellos
él
esa
esas
ese
esos
esta
estar
estas
este
estos
f
g
h
hacer
hasta
i
j
jamás
junto
juntos
k
l
la
las
lo
los
m
mas
más
me
menos
mía
mientras
mío
misma
mismas
mismo
mismos
mucha
muchas
muchísima
muchísimas
muchísimo
muchísimos
mucho
muchos
muy
nada
n
ni
ninguna
ningunas
ninguno
ningunos
no

nos
nosotras
nosotros
nuestra
nuestras
nuestro
nuestros
nunca
o
os
otra
otras
otro
otros
p
para
parecer
poca
pocas
poco
pocos
por
porque
q
que
querer
qué
quien
quienes
quienesquiera
quienquiera
quién
r
ser
s
si
siempre
sí
sín
Sr
Sra
Sres
Sta
suya
suyas
suyo
suyos
t
tal
tales
tan
tanta
tantas
tanto
tantos
te
tener
ti
toda
todas
todo
todos
tomar
tuya
tuyo
tú
u
un
una
unas
unos
usted
ustedes
v
varias
varios
vosotras
vosotros
vuestra
vuestras
vuestro
vuestros
w
y
yo
z

Referencias

Aas, K., Eikvil, L. Text categorisation: A survey. Technical report, Norwegian Computing Center, 1999

Abe N., Invited Talk: Sampling Approaches to Learning from Imbalanced Data Sets: Active Learning, Cost Sensitive Learning and Deyond, Proc. Int'l Conf. Machine Learning, Workshop Learning from Imbalanced Data Sets II, 2003.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, pages 144-152. ACM Press, 1992.

[CMC, 2007]. Computational Medicine Center. 2007. The computational medicine center's 2007 medical natural language processing challenge. http://computationalmedicine.org/challenge/index.php.

Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning.Machine Learning 15(2), 201–221 (1994)

Cover, T. M. Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern recognition. IEEE Transactions on Electronic Computers, 14, pp. 326–334, 1965.

Dasarathy, B. V. Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. Los Alamitos, CA: IEEE Computer Society Press. 1991

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. Indexing by latent semantic indexing. Journal of the American Society for Information Science 41, 6, 391-407, 1998

Dinwoodie, H. P. Morbidity recording. J.R.Coll.Gen.Pract. 22.119 (1972): 417-20.

Dumais, Susan, Platt, John, Heckerman, David, & Sahami, Mehran. Inductive learning algorithms and representations for text categorization. Pages 148-155 of: Cikm '98: Proceedings of the seventh international conference on information and knowledge management. New York, NY, USA: ACM, 1998

Ertekin S., Huang J., Bottou L., and Giles L., Learning on the Border: Active Learning in Imbalanced Data Classification, Proc. ACM Conf. Information and Knowledge Management, pp. 127-136, 2007.

Ertekin S., Huang J., and Giles C.L., Active Learning for Class Imbalance Problem, Proc. Int'l SIGIR Conf. Research and Development in Information Retrieval, pp. 823-824, 2007.

Esuli A. and Sebastiani F. Active Learning Strategies for Multi-Label Text Classification**.** Proceedings of the 31st European Conference on Information Retrieval (ECIR'09), Toulouse, FR, 2009

Figuerola, C., Rodríguez, A., y Berrocal, J. Automatic vs. Manual Categorization of Documents in Spanish. Journal of Documentation, 57(6):763–773, 2001

Fenton, S. H. Clinical vocabularies and terminologies: impact on the future of health information management. Top.Health Inf.Manage. 21.2 (2000): 74-80.

Garvin, J. H., V. Watzlaf, and S. Moeini. Automated coding software: development and use to enhance anti-fraud activities. AMIA.Annu.Symp.Proc. (2006): 927

He H. and Garcia E. A., Learning from Imbalanced Data, IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009.

Hersh WR, Buckley C, Leone TJ, Hickam DH, OHSUMED: An interactive retrieval evaluation and new large test collection for research, Proceedings of the 17th Annual ACM SIGIR Conference, 1994, 192-201. http://ir.ohsu.edu/ohsumed/

Hsu, C.W. and Lin, C.J. A comparison of methods for multiclass support vector machines. Neural Networks, IEEE Transactions on 13 (2), 415-425. 2002.

Joachims, T. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In In Proceedings of ICML-97, 14th International Conference on Machine Learning. Douglas H. Fisher, editor., pages 143–151, Nashville, US. Morgan Kaufmann Publishers, San Francisco, US, 1997

Joachims, T. Text categorization with support vector machine: Learning with many relevant features. In European Conference on Machine Learning (ECML), pages 137–142, Berlin. Springer. 1998

Joachims, T.: Making large-scale SVM learning practical. In: Advances in Kernel Methods - Support Vector Learning. MIT press, Cambridge (1999)

Joachims, Learning to Classify Text using Support Vector Machines, Kluwer/Springer, 2002

Knight, K. Mining online text. Communications of the ACM 42, 11, 58-61. 1999

Larkey, L. and Croft, W. B., Automatic Assignment of ICD9 Codes to Discharge Summaries, Center for Intelligent Information Retrieval Technical Report (1995).

Larkey, L. and Croft, W. B., Combining Classifiers in Text Categorization, Proceedings of the 19th International Conference on Research and Development Information Retrieval (SIGIR96), Zurich, Switzerland, pp. 289-297

Lewis, D. D. Text representation for intelligent text retrieval: a classification-oriented view. pages 179–197, 1992.

Lewis, D., Schapire, R., Callan, J., y Papka, R. Training algorithms for linear text classifiers. In In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Special Issue of the SIGIR Forum),pages 298–306, ACM, 1996

Lewis, David. Reuters-21578 Text Categorization Collection Distribution 1.0, 1997. http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

Li, H. and Yamanishi, K. Text classification using ESC-based stochastic decision lists. In Proceedings of CIKM-99, 8th ACM International Conference on Information and Knowledge Management (Kansas City, US, 1999), pp. 122-130. 1999.

Li Y. and Jain A.K, Classification of Text Documents, Proc. 14th Int'l. Conf. Pattern Recognition, Brisbane, pp. 1295-1297, August 1998.

Lojo D., Losada D. y Barreiro A. CIE-9-MC code Classification with knn and SVM. 3rd International Work-conference on the Interplay between Natural and Artificial Computation, IWINAC 2009, Santiago de Compostela (Spain), Jun 2009.

Lojo D., Losada D. y Barreiro A. Evaluación de técnicas de Aprendizaje Activo para codificación CIE-9-MC de informes de alta hospitalaria. 1st Spanish Conference on Information Retrieval, CERI 2010, Madrid (Spain), Jun 2010.

Luhn, H. P. The automatic creation of literature abstracts. IBM Journal of Research Development, 2(2):159-165. [2, 3, 6, 8], 1958.

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

Osuna, E., R. Freund, and F. Girosi. An improved training algorithm for support vector machines. in Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop. 1997.

Peinado Rodríguez, Jesús. Lematización para palabras médicas complejas: implementación de un algoritmo en LISP. *Rev Med Hered*, oct. 2003, vol.14, no.4, p.223-228. ISSN 1018-130X.

Pestian J. P., Brew C., Matykiewicz P.M., Hovermale D.J., Johnson N., Cohen K.B., Duch W.: A shared task involving multi-label classification of clinical free text. Proceedings of ACL BioNLP; 2007 Jun; Prague. (2007)

Platt, J., Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.

Platt, J., Fast training of support vector machines using sequential minimal optimization. 1999: p.185-208

Ponte, J. and Croft, W. B. , A Language Modeling Approach to Information Retrieval. In Proc. SIGIR, pp. 275-281. ACM Press. 1998.

Provost F., Machine Learning from Imbalanced Data Sets 101, Proc. Learning from Imbalanced Data Sets: Papers from the Am. Assoc. for Artificial Intelligence Workshop, 2000 (Technical Report WS-00-05).

Quinlan J.R. Induction of Decision Trees, Machine Learning, (1), 81-106, 1986

Quinlan J.R. C4.5 Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993

Rocchio, J. Relevance feedback in information retrieval. The SMART Retrieval System. Experiments in Automatic Document Processing. Prentice Hall, Englewoods Cliffs, N. J., 1971

Salton, G. The SMART Retrieval System. Experiments in Automatic Document Processing, Prentice Hall, Englewood Cliffs, NJ, 1971

Salton G., Wong A., y Yang C.. A vector space model for automatic indexing. Communications of the ACM, 18(11), 1975.

Salton, G. y M. McGill. Introduction to Modern information Retrieval. McGraw Hill, New York, 1983.

Salton G.. Automatic Text Processing: the transformation, analysis and retrieval of information by computer. Addison Wesley, 1989.

Schapire, R. E. and Singer, Y. BoosTexter: a boosting-based system for text categorization. Machine Learning 39, 2/3, 135-168. 2000.

Sebastiani, F. Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1):1–47, 2002.

Sparck, J. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28(1):11–21, 1972.

Stanfill, M. H., et al. A systematic literature review of automated clinical coding and classification systems. J.Am.Med.Inform.Assoc. 17.6 (2010): 646-51.

Strohman, T., Metzler, D., Turtle, H., Croft, W. B.: Indri: A language model-based search engine for complex queries. Proceedings of the International Conference on Intelligence Analysis, May 2-6, 2005, McLean, VA.

Surjan, G. Questions on validity of International Classification of Diseases-coded diagnoses. Int.J.Med.Inform. 54.2 (1999): 77-95.

Surjan, G. and G. Heja. Indexing of medical diagnoses by word affinity method. Stud.Health Technol.Inform. 84.Pt 1 (2001): 276-79.

Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. Journal of Machine Learning Research 2, 45–66 (2001)

Turtle, H. and Croft, W. B. , Efficient Probabilistic Inference for TextRetrieval, Proceedings of RIAO 3, 644-661. 1991

Vapnik, V. Estimation of Dependences Based on Empirical Data. Springer Verlag, 1982.

Vapnik, V. The nature of statistical learning theory. Springer–Verlag, New York, 1995.

Wang, G., A Survey on Training Algorithms for Support Vector Machine Classifiers. Networked Computing and Advanced Information Management, International Conference on, 2008. 1: p. 123-128.

Weiss, S. M., Apte, C., Damerau, F. J., Johnson, D. E., Oles, F. J., Goetz, T., and Hampp, T. Maximizing text-mining performance. IEEE Intelligent Systems 14, 4, 63-69, 1999.

Weiss G.M. and Provost F., The Effect of Class Distribution on Classifier Learning: An Empirical Study, Technical Report MLTR-43, Dept. of Computer Science, Rutgers Univ., 2001.

Weston, J. and Watkins, C. Multi-class support vector machines. Proceedings ESANN, Brussels. 1999.

Yang, Y. An evaluation of statistical approaches to text categorization. Inform. Retr. 1, 1–2, 69–90. 1999.

Yang, Y. y Liu, X. A re-examination of text categorization methods. In 22nd Annual International SIGIR, pages 42–49, Berkley, 1999

Zieserl, R. M. and S. P. Dowell. Using microcomputers to improve the timeliness, accuracy, and accessibility of clinical data. J.Soc.Health Syst. 1.2 (1989): 12-24.

Zipf, G. K. Human behavior and the principle of least effort. Addison-Wesley, Reading, MA, 1949