Capítulo 8

Conclusión y perspectivas

Ya sabes
lo que hay que hacer en este mundo:
andar,
como un arado, andar entre la tierra.
Blas de Otero, Que trata de España.

El objetivo del trabajo presentado en esta memoria ha sido el estudio de la inferencia gramatical, esto es, el aprendizaje de las reglas sintácticas que rigen la construcción de las frases de un lenguaje, realizado a partir de fuentes aleatorias de ejemplos. El interés de este problema está motivado porque numerosos problemas prácticos de clasificación de patrones, tales como el reconocimiento de voz o el procesamiento de lenguaje natural, sugieren que la descripción más adecuada ha de realizarse en términos de ejemplos ruidosos o aleatorios.

Aquí nos hemos limitado a estudiar modelos sencillos en los que la generación de los ejemplos se puede realizar mediante un sistema de memoria acotada, lo que normalmente se denomina un autómata finito. En cuanto a los lenguajes estudiados, estos se componían de estructuras lineales (cadenas) o arbóreas.

En el capítulo 3 se han desarrollado algunos métodos para medir la calidad de una hipótesis con respecto al modelo probabilístico real basándose en la teoría de la información. También se ha discutido la importancia de identificar en el límite la estructura del modelo.

En el capítulo 5 se ha propuesto un algoritmo que identifica cualquier lenguaje regular determinista de cadenas construido sobre un alfabeto arbitrario $\mathcal{A}$. La identificación se consigue a partir de ejemplos aleatorios de las cadenas del lenguaje, sin que sean necesarios los contraejemplos. Además las probabilidades del modelo se aproximan rápidamente a las correctas, consiguiendo una rápida reducción de la distancia entre ambas distribuciones expresada por la entropía relativa. Este comportamiento se consigue gracias a la identificación en el límite de la estructura del autómata finito mínimo M que genera el lenguaje. Esta identificación reduce el número de probabilidades que se deben determinar, que pasa de infinito a finito, en concreto a un número $|M||\mathcal{A}|$. Experimentalmente, el algoritmo requiere muy poco tiempo y muestras comparativamente pequeñas para identificar el lenguaje. Para muestras muy grandes, consume un tiempo lineal en función del número de ejemplos (aproximadamente 4 minutos por millón de ejemplos en un computador Hewlett-Packard 715 con 40 MIPS). El algoritmo es, por tanto, adecuado para tareas de reconocimiento en las que aparezcan ejemplos con ruido o fuentes aleatorias.

Recientemente (Ron, Singer & Tishby 1995) han propuesto algoritmos que identifican una subclase de autómatas estocásticos en el sentido PAC. Dichos algoritmos admiten sólo fusiones de estados entre un mismo nivel del árbol de prefijos y conducen a grafos acíclicos. Los autores reconocen que no poden demostrar la corrección del algoritmo cuando se admiten fusiones entre distintos niveles, lo que indica la dificultad de demostrar la convergencia PAC de algoritmos del tipo de los aquí propuestos. Sin embargo, ésta sería una línea de investigación futura de cierto intere's, así como la comparación de los rendimientos de ambos tipos de algoritmos en problemas prácticos.

En el capítulo 6 se ha presentado un algoritmo que es capaz de

aprender gramáticas independientes del contexto a partir de muestras aleatorias que contienen esqueletos de árboles de derivación de las cadenas del lenguaje. El resultado es en el límite, idéntico estructuralmente a la gramática objetivo, es decir, ambas generan el mismo lenguaje estocástico de esqueletos. El algoritmo encuentra su hipótesis en un tiempo que crece sólo linealmente con el tamaño de la muestra. Experimentalmente, la identificación se alcanza con número de ejemplos moderado y la velocidad del algoritmo resulta adecuada para futuras aplicaciones a tareas de reconocimiento.

Una línea de trabajo para el futuro que podría dar lugar a resultados interesantes es la aplicación de estos algoritmo a la inferencia de gramáticas de grafos para su utilización en problemas de reconocimiento de formas. Una extensión más inmediata podría ser la generalización del algoritmo de identificación de lenguajes de árboles para la identificación de lenguajes de grafos dirigidos acíclicos, cuyo interés se justifica por la naturalidad de la descripción de figuras planas (tales como caracteres manuscritos) en términos de este tipo de grafos.

En el capítulo 7 se ha estudiado el poder de generalización de las redes neurales de segundo orden entrenadas con muestras estocásticas de lenguajes regulares. Se ha visto que en algunos casos aparecen cúmulos en el espacio interno de representación asociados a los estados del autómata finito que genera el lenguaje. Si se calculan las probabilidades con el autómata extraído de la red, la distancia entre la distribución verdadera y el modelo se reduce de forma sustancial. Si bien, en general, el comportamiento de los algoritmos simbólicos es muy superior al de los algoritmos basados en redes neurales, las redes neurales de segundo orden se presentan como candidatas para el modelado de procesos estocásticos donde no se dispone todavía de algoritmos gramaticales. Una línea de investigación por desarrollar es la comparación de ambos tipos de métodos cuando la información de las muestras es parcialmente incorrecta o está afectada por ruido. En esa situación, las redes neurales son más robustas frente a defectos en la información y las desventajas implícitas en su entre-

namiento pueden ser total o parcialmente compensadas. Una de las dificultades de la extracción de estados en las redes entrenadas con muestras estocásticas es la necesidad de definir un método de minimización compatible con las incertidumbres estocásticas. Un intento en este sentido se encuentra en la tesis de S.C. Kremer (1996), lo que resulta en una acercamiento a los algoritmos de fusión de estados del tipo `rlips`. Otro problema por resolver es la aparición frecuente de inestabilidades durante el entrenamiento.

Finalmente, una línea de interés en la actualidad es la predicción de series temporales. Esta tarea involucra muestras probabilísticas (por lo que puede ser interesante extender los métodos simbólicos aquí discutidos a este tipo de aplicaciones), pero también fuertes componentes de ruido, no estacionariedad etc, lo que ha originado una exhaustiva utilización en el pasado de modelos neurales. El diseño de un sistema que integre las ventajas de ambos métodos puede resultar una tarea de interés en el futuro próximo.

Alicante, 13 de diciembre de 1999