Segunda Práctica Aprendizaje Automático

Inducción de Reglas Gramaticales con ILP

4º Ingeniería en Informática
Curso 2004-05 (Junio)

1. Introducción

Hoy en día existe una cantidad considerable de trabajos de investigación que relacionan Aprendizaje Automático con Procesamiento de Lenguaje Natural. Las principales tareas que se plantean en el Procesamiento de Lenguaje Natural requieren una cantidad sustancial de conocimiento, que normalmente se debe propocionar a mano. La aplicación de Aprendizaje Automático a estas tareas está orientada a reducir el trabajo a mano, ya que aporta algoritmos que permiten adquirir este conocimiento a partir de un conjunto de datos disponible. La idea básica es representar los problemas que se plantean en Procesamiento de Lenguaje Natural como problemas de clasificación.

Una de las partes en las que tradicionalmente se divide el estudio de la lengua es la sintaxis o gramática. En esta práctica nos plantearemos el problema de determinar si una frase dada pertenece o no a una gramática. Es relativamente sencillo resolver este problema si se cuenta con las reglas gramaticales. Si no es así, nos tendríamos que plantear como primera tarea generar estas reglas, aunque también podríamos pensar en utilizar alguna técnica de Aprendizaje Automático que a partir de frases correctas nos permita aprender las reglas gramaticales, de esta forma nuestra tarea consistiría en generar una batería de frases para alimentar al correspondiente algoritmo (en vez de generar a mano todas las reglas gramaticales). Esto es precisamente lo que trataremos de hacer en esta práctica con frases expresadas en castellano. Para ello utilizaremos Programación Lógica Inductiva (ILP) con el sistema Aleph.

2. Descripción de la Tarea

Para llevar a cabo la práctica se deberá:

- Generar los ficheros de ejemplos positivos (y negativos si procede). Las frases deberán estar formadas al menos por un sujeto y un predicado. Como el lenguaje de representación es Prolog, tendremos que elegir cómo representar en Prolog los ejemplos. Una posible forma de representarlos es mediante un predicado `frase_valida` de aridad 1, cuyo único argumento sea una lista con las palabras de la frase. Por ejemplo:

```
frase_valida([el,niño,lee]).
```

- Generar el fichero con el Conocimiento de Base. En éste habrá que añadir los parámetros, modos y determinaciones que sean pertinentes, además del Conocimiento de Base en sí, que podría incluir conocimiento sobre las palabras del lenguaje (que constituirían los símbolos terminales de la gramática) como los tipos de palabras y características morfológicas de las mismas. Por ejemplo:

```
articulo([el],masculino,singular).
sustantivo([niño],masculino,singular).
```

```
verbo([lee],tercera_persona,singular).
```

Además, en este fichero habrá que añadir algún predicado que permita extraer cada una de las palabras de una frase en el mismo orden en que en ésta aparecen. (Se puede utilizar cualquier predicado que permita extraer las palabras y refleje la posición de las mismas en la frase). Por ejemplo:

```
partes([X|R],[X],R).
```

que divide la lista que se pasa como primer argumento en dos partes, la primera es una lista con el primer elemento (en las frases será por lo tanto una lista con la primera palabra), y la segunda el resto.

Algunos tipos de palabras que podéis considerar son:

- artículos: la, el, los, ...
- demostrativos: éste, ese, aquel, ...
- posesivos: mi, mío, tuyo, ...
- indefinidos: uno, varios, demasiados ...
- preposiciones: a, ante, bajo, ...
- adverbios: luego, antes, entonces, ...
- pronombres: yo, tú, ellos, ...
- sustantivos: niño, libros ...
- adjetivos: azul, grande ...
- verbos: lee, come ...

- Utilizar Aleph para tratar de inducir reglas generales que nos permitan determinar si una nueva frase (es decir una frase que no se encuentra entre los ejemplos de entrenamiento) es válida o no. Las reglas que se espera obtener serán reglas gramaticales (aunque con el modelo de los ejemplos anteriores únicamente contendrán símbolos terminales). Es decir, expresarán información del tipo: *Una frase es válida si contiene un artículo con género G y número N, un sustantivo con género G y número N y un verbo en tercera persona y número N en este orden*. Suponiendo que únicamente conjugamos el presente indicativo de los verbos. En Prolog:

```
frase_valida(A):- partes(A,B,C),
                  artículo(B,G,N),
                  partes(C,E,F),
                  sustantivo(E,G,N),
                  verbo(F,tercera_persona,N).
```

- Evaluación de la calidad de las reglas que induce Aleph utilizando un conjunto de test que contenga frases nuevas.

Algunos comentarios:

- Se recomienda que el proceso de realización de la práctica sea incremental, generando inicialmente un conjunto de ejemplos y un conocimiento de base pequeños para hacer pruebas.
- La forma de representar el conocimiento que aquí se expone en los ejemplos no es la única posible. Podéis utilizarla o no, siempre que se consiga que la práctica cubra su objetivo: aprender reglas gramaticales a partir de frases. Podéis consultar con el profesor cualquier idea sobre otras formas de llevarlo a cabo.
- La forma de conseguir las frases de entrada se deja a criterio del alumno.

3. Entrega de la Práctica

La práctica se debe entregar como muy tarde el día del examen de Junio de la asignatura.

Para realizar la entrega se deberá:

- Enviar un e-mail a rfuentet@inf.uc3m.es con `Subject: Práctica 2 de AA` en el que aparezcan los nombres de los integrantes del grupo y en el que se manden adjuntos los ficheros utilizados para alimentar a Aleph y un fichero con las reglas inducidas.
- Entregar una memoria breve en papel con la siguiente estructura:

  1. **Introducción** (1 hoja de descripción de la práctica desde vuestro punto de vista)
  2. **Descripción del trabajo realizado**, que debe incluir:
     - Una descripción del conocimiento expresado en los ficheros utilizados con las explicaciones que se consideren necesarias (principalmente del fichero que contiene el Conocimiento de Base).
     - Descripción de cualquier prueba realizada que se considere de interés.
     - Análisis del resultado y evaluación de la teoría aprendida.
  3. **Conclusiones**: En las que además de cualquier tipo de conclusión a la que se llegue tras la realización de la práctica, se debe incluir:
     - Diferencias que encontráis entre la Programación Lógica Inductiva y los métodos de aprendizaje inductivo simbólicos utilizados en la primera práctica.
     - Opinión acerca de la aplicación de este tipo de aprendizaje automático al tipo de problemas tratado en la práctica. ¿Se te ocurren otras aplicaciones?.
     - Opinión sobre la conveniencia de la práctica, de lo que se ha aprendido, . . . etc.

     Las conclusiones son una parte importante de la práctica, no las descuidéis.