Tema-3-Tecnicas-automatizadas-pa...

ParmigianoReg

Ingenieria del Conocimiento

3º Grado en Ingeniería Informática

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación
Universidad de Granada

Tema 3 - Técnicas (automatizadas) para la Adquisición de Conocimientos

Profesor Juan Luis Castro Peña, Curso 20-21

Tareas a Automatizar

- La conceptualización de dominio por medio de la obtención de conceptos y factores relevantes.
  - Se utiliza la Repertory Grid o Rejilla de Repertorio.
  - Un concepto es una reunión de varios objetos básicos que comparten una característica.
- La obtención de reglas, por medio de la obtención reglas candidatas a partir de ejemplos.
  - Aprendizaje con árboles de decisión
  - Aprendizaje de reglas

Rejilla de Repertorio:

- También conocida como emparrillado o Repertory Grid.
- Fue desarrollada por Kelly en 1955 para ayudar enfermos mentales.
- Se aplica en muchos ámbitos
  - Asesoramiento.
  - Estudios demográficos.
  - Dinámica de grupos.
  - Adquisición de conocimiento.

Idea de la Rejilla

- Cada persona tiene su propia visión del mundo: asocia y distingue elementos de acuerdo a sus propios criterios. Su propia forma de agrupar diferentes conceptos. Esto se le denomina como **Constructor**, también conocido como **Construcción**.
  - Es una característica bipolar que se aplica de forma gradual a los elementos.
- Cuando a una persona se le pide que dé razones para ir distinguiendo o agrupando elementos, va a indicar los criterios y su forma de usarlos.

Proceso de la rejilla de repertorio

- Dialogo inicial con el experto
- Sesión de valoración
- Análisis de los resultados

Constructores

- Una construcción es una característica bipolar en la cual cada elemento tiene un cierto grado o escala.
  - Por ejemplo: Pesado - Ligero, Valiente - Cobarde, Largo - Corto.
- Estas características se pueden representar en una escala lineal con valores ponderados relativos a la característica. No son características binarias de Sí o No solamente, aunque pueden incluir valores binarios de ser necesario.

Valoración de objetos según el constructor

- Como es el experto quien crea las construcciones, es trabajo suyo comprender qué hace que una construcción sea válida y cómo se usa. Los ratios pueden expresarse también con nombres en vez de números.
- La escala no debe variar en una misma construcción pero puede variar de una construcción a otra.
- Un elemento se pondera según cada construcción usando un ratio subjetivo dado por el experto, esto permite clasificar los elementos en vez de compararlos.

Utilidad de la rejilla de repertorio en la IC

- Es un sistema que da una correspondencia entre elementos y construcciones, esta técnica es útil para Ingenieros de Conocimiento porque:
    - Hace que el experto piense sobre el problema y por tanto ayuda a clarificar consecuencias en su mente.
    - Los grids se pueden analizar para encontrar modelos, clases y conceptos o asociaciones a investigar con mayor profundidad.

Obtención de la Rejilla
Preparación

- Se elige el problema:
    - Clasificación de objetos, taxonomía.
    - Evaluación de candidatos o alternativas.
    - Obtención de las características que determinan una o varias clases de objetos.
- Recoger una lista de objetos iniciales.

Rellenado

- Implica repetidas comparaciones de elementos, se suelen elegir un grupo de tres elementos para empezar a describir similitudes o diferencias.
- Los grupos de tres pueden elegirse al azar o sistemáticamente.
- Luego de tener ese constructo, se aplica al resto de elementos y se comienza otra comparación.

	C1. Responsable/Irresponsable	C2. Clasificaciones Pobres/Buenas	C3. Inseguro/Seguro ...	C4. Preguntan/No preguntan	C5. Atentos/Distraídos	C6. Trabajan/No trabajan	C7. Integrados/No Integrados	C8. Organizados/No organizados
E1. Francisco	3	1	3	1	1	1	1	3
E2. Elena	1	2	1	3	3	3	3	1
E3. Manolo	3	1	1	2	3	2	3	1
E4. José	1	2	2	2	1	1	2	2
E5. Isabel	1	3	3	1	2	1	2	3
E6. Yolanda	1	2	1	3	2	3	3	2
E7. David	1	2	3	2	3	1	1	1

Análisis de Entidades

- Es necesario para ejecutar la técnica del clustering un criterio que mida la distancia entre pares de elementos, en este caso se puede utilizar la suma de diferencias absolutas entre cada par de elementos.
    - Es decir, resta absoluta entre los valores de una fila y los de otra fila.
- Luego, se realiza una matriz de Entidades con Entidades con los valores obtenidos.

NEW

WUOLAH Print

Lo que faltaba en Wuolah

Imprimir

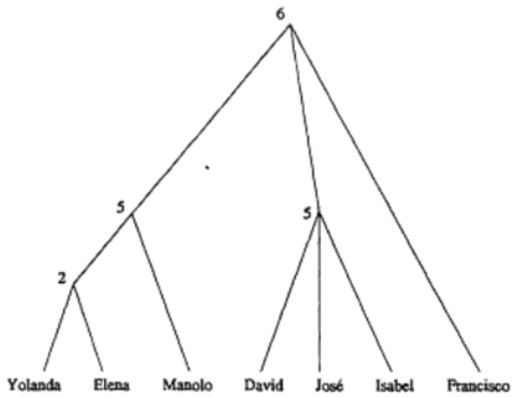
- [ ] Todos los apuntes que necesitas están aquí
- [ ] Al mejor precio del mercado, desde **2 cent.**
- [ ] Recoge los apuntes en tu copistería más cercana o recíbelos en tu casa
- [x] Todas las anteriores son correctas

	E1	E2	E3	E4	E5	E6	E7
E1		15	10	7	6	13	8
E2			5	10	11	**2**	7
E3				9	12	7	8
E4					5	6	5
E5						9	6
E6							9
E7							

- A partir de esta tabla se construye un árbol siguiendo el siguiente procedimiento:
  - Se señala en la rejilla los elementos con distancia mínima.
  - Esos elementos se reemplazan en la rejilla por el conjunto de ambos, por ejemplo si se elige $E2$ y $E6$, se reemplaza por $[E2, E6]$
  - La distancia del resto de los elementos al conjunto es el mínimo de las distancias de los elementos del conjunto.
  - Se repite el proceso con la matriz resultante hasta que solo quedan dos categorías en la tabla.

	E1	E3	E4	E5	E7	[E2-E6]
E1		10	7	6	8	13
E3			9	12	8	**5**
E4				**5**	**5**	6
E5					6	9
E7						7
[E2-E6]						

Luego de iterar…



Análisis de Construcciones

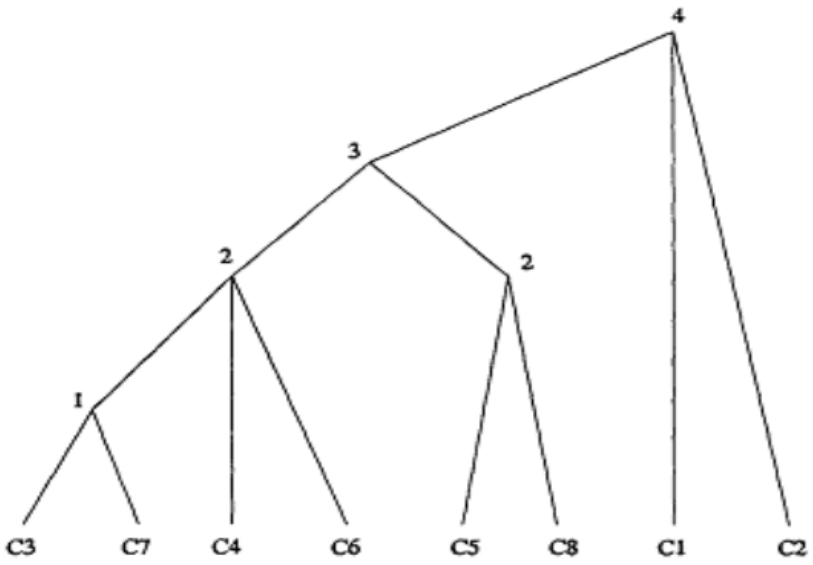
- El procedimiento es el mismo que para las entidades/elementos, aunque la distancia entre las construcciones requiere de observarla en ambas direcciones, es decir obtener la resta absoluta de las dos columnas de arriba hacia abajo y otra resta absoluta con una de las columnas en dirección invertida. Luego se reduce al valor más pequeño que se obtiene.

	C1	C2	C3	C4	C5	C6	C7	C8
C1		10	7	9	8	7	9	6
C2	4		5	5	6	7	6	4
C3	7	7		10	9	12	11	3
C4	5	5	2		5	2	3	9
C5	6	6	5	9		5	4	10
C6	7	7	2	10	9		3	9
C7	6	6	1	9	8	11		7
C8	8	6	9	3	2	5	4	

Superior derecha en orden directo, inferior izquierda en orden inverso

	C1	C2	C3	C4	C5	C6	C7	C8
C1		4	7	5	6	7	6	6
C2			5	5	6	7	6	4
C3				2	5	2	1	3
C4					5	2	3	3
C5						5	4	2
C6							3	5
C7								4
C8								

Quedándose con la distancia mínima entre el orden directo e inverso



Análisis de Resultados

- Cuando se analizan las categorizaciones realizadas, el Experto puede corroborar, refutar o matizar los agrupamientos realizados, pudiendo producirse los casos siguientes:
    - Dos elementos aparecen ligados cuando no deberían estarlo, en este caso se pueden reconsiderar valores atribuidos a los objetos, si los valores parecen correctos, el IC debe solicitar un motivo para diferenciar esos elementos, dicha característica debe añadirse al estudio y generar dos nuevos árboles de agrupamiento.
    - Dos elementos aparecen como disjuntos cuando deberían estar ligados, el proceso es similar al anterior. Se añade una característica que los ligue.
    - Dos características aparecen ligadas cuando no deberían estarlo, si los valores atribuidos a los objetos para esas dos características parecen correctos se le solicita al Experto un ejemplo de un elemento que contradiga la relación. Si existe, se añade el elemento a la rejilla inicial y se repite el proceso.

Árboles de Decisión:

- Un árbol de decisión toma de entrada un objeto o una situación descrita a través de un conjunto de atributos y devuelve una "decisión", el valor previsto de la salida dada la entrada.
- Los atributos pueden ser discretos o continuos.
- Salida:
    - Discreta: Clasificación

    - Continua: Regresión
- Los árboles de decisión se pueden convertir en reglas "*if-then*" siguiendo cada rama del árbol.
    - Los árboles pueden tener en las hojas diferentes casos e inclusive casos en lo que la regla no se cumple, aquí se introduce la noción de certeza de una regla.
    - Es posible que un nodo hoja posea un solo caso, esto no es muy representativo, esa regla no está bien respaldada. Se conoce como el "soporte" y en el árbol se descartan aquellas reglas que tengan un soporte mínimo.
- Es necesario que los árboles tengan un grado de profundidad bajo para que sean entendibles por una persona.
- Lo que se obtiene en un árbol de decisión es un "pre-conocimiento" el cual luego debe ser validado por el experto, el experto luego añade excepciones a la reglas o bien existan reglas que no tengan sentido, lo cual luego debe modificarse.
- Existe la posibilidad en que dado un grupo de entrenamiento no sea posible generarse un árbol de decisión, pueden aproximarlo con el grado de certeza.

Ejemplos positivos y negativos

- Los ejemplos positivos son aquellos en los que la meta a esperar es verdadera.
- Los negativos son aquellos en la que es falsa.
- El conjunto de ejemplos completo se denomina conjunto de entrenamiento.

Expresividad de los árboles de decisión

- Los árboles de decisión pueden expresar cualquier función a partir de los atributos de entrada.

Inducción en los árboles de decisión

- Existen múltiples maneras de inferir el árbol
    - Trivial: Se crea una ruta del árbol por cada instancia de entrenamiento
        - Producen árboles excesivamente grandes.
        - No funcionan bien con instancias nuevas.
    - Óptimo: El árbol más pequeño posible compatible con todas las instancias (Navaja de Ockham)
        - Inviable computacionalmente
    - Pseudo-óptimo (Heurístico): Selección del atributo en cada nivel en función de la calidad de la división que lo produce
        - Los principales programas de generación de árboles utilizan procedimientos similares.

Elección de Atributos

- Un buen atributo debería dividir el conjunto de ejemplos en subconjuntos que sean o "todos positivos" o "todos negativos". Un atributo perfecto dividiría los ejemplos en conjuntos que contienen solo ejemplos positivos y solo ejemplos negativos.
- Se tiene que definir una medida de atributo "bastante adecuado" o "Inadecuado".
- Para un conjunto de entrenamiento que contenga p ejemplos positivos y n ejemplos negativos, se utiliza la función del grado de entropía.

$$I\left(\frac{p}{p+n},\frac{n}{p+n}\right)=-\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right)-\frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

- Permite medir la ausencia de "homogeneidad" de la clasificación.
- La entropía esperada luego de utilizar un atributo A en el árbol es:

$$resto(A)=\sum_{i=1}^{v}\frac{p_i+n_i}{p+n}I\left(\frac{p_i}{p_i+n_i},\frac{n_i}{p_i+n_i}\right)$$

- La ganancia de información esperada después de usar el atributo es:

$$Ganancia(A)=I\left(\frac{p}{p+n},\frac{n}{p+n}\right)-resto(A)$$

- Se elige el atributo con mayor valor de G.
- Este criterio tiene un fuerte sesgo a favor de tests con muchas salidas, por lo que se puede mitigar con un ratio de ganancia.

Ratio de Ganancia:

$$RGanancia\ \ (A) = \frac{Ganancia\ \ (A)}{dINFO\ \ (A)}$$

con

$$dINFO\ \ (A) = -\sum_{i=1}^{v} \frac{p_i + n_i}{p + n} \log_2 (\frac{p_i + n_i}{p + n})$$

- Se puede reemplazar la entropía con la heurística Gini, que es similar a la ganancia de información.

Atributos de Entrada Continuos

- Se ordenan los valores del atributo, y se especifica la clase a la que pertenecen; se crean nodos de decisión discretos tomando los puntos de corte entre los valores.
- Puede dejar de ser efectivo con muchos valores.

Atributos de Salida Continuos

- Se utiliza un árbol de regresión: En cada hoja se dispone de una función lineal de algún subconjunto de atributos numéricos.
- El algoritmo de aprendizaje debe decidir cuándo dejar de dividir para comenzar a aplicar la regresión lineal utilizando los atributos restantes.