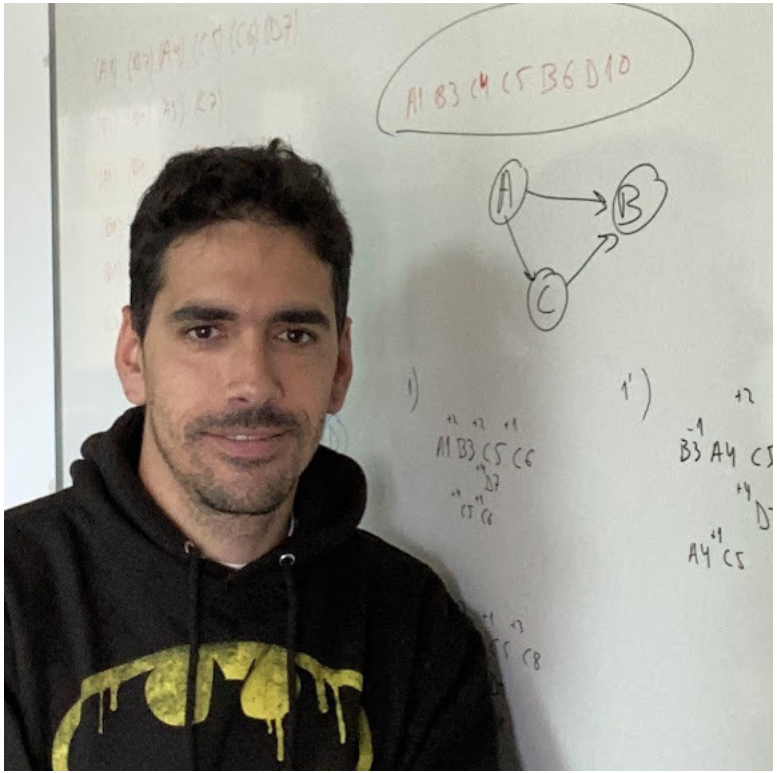
Métodos descriptivos supervisados

Máster en Ciencia de Datos

UCO ONLINE

José María Luna

recibió el título de Doctor en Ciencias de la Computación en 2014, por la Universidad de Granada. Su carrera investigadora fue inicialmente subvencionada por el Ministerio de Educación de España bajo el programa FPU (predoctoral) y el programa Juan de la Cierva (postdoctoral). Actualmente es profesor de la Universidad de Córdoba en el departamento de Informática y Análisis Numérico. Dr. Luna ha sido autor de los libros monográficos"Pattern Mining with Evolutionary Algorithms" y "Supervised Descriptive Pattern Mining", ambos publicados por la editorial Springer. Además, ha publicado más de 30 artículos en revistas científicas de alto impacto. Actualmente tiene un total de 2266 citas en Google Scholar y un índice H de 25. Su investigación es llevada a cabo en el grupo de investigación *Knowledge Discovery and Intelligent Systems* , donde investiga temas relativos a computación evolutiva, minería de patrones, reglas de asociación y sus aplicaciones.



Aplicaciones de métodos descriptivos (Parte 1)

Postprocesado de datos

UCO ONLINE

Postprocesado de datos

- Introducción
  - Tras aplicar algoritmos de extracción de patrones frecuentes, trabajamos sobre un espacio de búsqueda de 2^k-1 soluciones posibles
  - Los algoritmos de reglas de asociación trabajan sobre un espacio de búsqueda de 3^k-2^{k+1}+1
  - Todas estas soluciones pueden ser proporcionadas al usuario final:
    - Profesores en el análisis de datos educativos
    - Médicos en el análisis de datos médicos
    - Cliente en el análisis de la cesta de la compra (tipo Amazon)

Postprocesado de datos

- Introducción
  - Para obtener los resultados, se suele trabajar con dos métricas generales:
    - Soporte (frecuencia)
    - Confianza (exactitud de la regla)
  - Sin embargo, estas métricas no son las únicas existentes, y en ocasiones se necesita de otras métricas que permitan reducir el conjunto total de resultados
    - Métricas subjetivas
      - Dependientes del usuario y de sus conocimientos
    - Métricas Objetivas
      - Dependientes de la distribución de los datos. Son métricas estadísticas en su mayoría

Postprocesado de datos

- Ejemplo

Sexo	Edad	Pais	Estudios	Estado civil	Economía
Hombre	25	España	Superiores	Soltero	Normal
Mujer	45	Alemania	Superiores	Casado	Normal
Mujer	62	Italia	Primaria	Casado	Normal
Mujer	56	España	Secundaria	Casado	Normal
Mujer	37	Suiza	Doctorado	Soltero	Rico
Mujer	55	Alemania	Doctorado	Casado	Rico
Hombre	71	Suiza	Primaria	Soltero	Rico
Hombre	18	Suiza	Secundaria	Soltero	Rico

Postprocesado de datos

- Ejemplo

Hombre -> Soltero	soporte = 0.375	confianza = 1.0
Hombre, Soltero -> Rico	soporte = 0.25	confianza = 0.66
Doctorado -> Mujer	soporte = 0.25	confianza = 1.0
Suiza -> Rico	soporte = 0.375	confianza = 1.0
Edad > 45 -> Casado	soporte = 0.375	confianza = 0.6
Mujer, Edad > 45 -> Casado	soporte = 0.375	confianza = 0.75
….		

Postprocesado de datos

- Ejemplo: queremos un conjunto de reglas, exactas, y que representen muy bien el dataset
  - Ordenamos las reglas por confianza y soporte en caso de empate

Hombre -> Soltero	soporte = 0.375	confianza = 1.0
Suiza -> Rico	soporte = 0.375	confianza = 1.0
Doctorado -> Mujer	soporte = 0.25	confianza = 1.0
Mujer, Edad > 45 -> Casado	soporte = 0.375	confianza = 0.75
Hombre, Soltero -> Rico	soporte = 0.25	confianza = 0.66
Edad > 45 -> Casado	soporte = 0.375	confianza = 0.6

Postprocesado de datos

- Ejemplo: queremos un conjunto de reglas, exactas, y que representen muy bien el dataset
  - Analizamos regla por regla, qué registros cumplen

    Hombre -> Soltero soporte = 0.375 confianza = 1.0

    Cubre los registros: 1, 7 y 8

Postprocesado de datos

- Ejemplo: queremos un conjunto de reglas, exactas, y que representen muy bien el dataset
  - Analizamos regla por regla, qué registros cumplen

    Suiza -> Rico soporte = 0.375 confianza = 1.0

    Cubre los registros: 5, 7 y 8

    Los registros que cubre son muy similares a los de la regla anterior, por lo que descartamos esta regla

Postprocesado de datos

- Ejemplo: queremos un conjunto de reglas, exactas, y que representen muy bien el dataset
  - Analizamos regla por regla, qué registros cumplen

    Doctorado -> Mujer soporte = 0.25 confianza = 1.0

    Cubre los registros: 5 y 6

    Los registros que cubre son diferentes a la primera regla que consideramos, por lo que nos quedamos con esta regla y ya tenemos cubiertos los registros 1, 5, 6, 7 y 8

Postprocesado de datos

- Ejemplo: queremos un conjunto de reglas, exactas, y que representen muy bien el dataset
  - Analizamos regla por regla, qué registros cumplen

    Mujer, Edad > 45 -> Casado soporte = 0.375 confianza = 0.75

    Cubre los registros: 3, 4 y 6

    Los registros que cubre son diferentes a los que ya tenemos por reglas previas, excepto el registro 6 que ya estaba cubierto. Nos quedamos con esta regla y ya tenemos cubiertos los registros 1, 3, 4, 5, 6, 7 y 8

Postprocesado de datos

- Ejemplo: queremos un conjunto de reglas, exactas, y que representen muy bien el dataset
  - Analizamos regla por regla, qué registros cumplen

    Hombre, Soltero -> Rico soporte = 0.25 confianza = 0.66

    Cubre los registros: 7 y 8

    Estos registros ya los tenemos cubiertos, así que descartamos la regla

Postprocesado de datos

- Ejemplo: queremos un conjunto de reglas, exactas, y que representen muy bien el dataset
  - Analizamos regla por regla, qué registros cumplen

    Edad > 45 -> Casado soporte = 0.375 confianza = 0.6

    Cubre los registros: 3, 4 y 6

    Estos registros ya los tenemos cubiertos, así que descartamos la regla

Postprocesado de datos

- Ejemplo: queremos un conjunto de reglas, exactas, y que representen muy bien el dataset
  - Nos quedamos con el conjunto de reglas:

Hombre -> Soltero	soporte = 0.375	confianza = 1.0
Doctorado -> Mujer	soporte = 0.25	confianza = 1.0
Mujer, Edad > 45 -> Casado	soporte = 0.375	confianza = 0.75

  - Estas reglas cubre 7/8 registros de los datos

Postprocesado de datos

- Algunas métricas adicionales
  - p_x es la frecuencia del antecedente
  - p_y es la frecuencia del consecuente
  - p_{xy} es la frecuencia de la regla

Support	p_{xy}	[0, 1]
Coverage	p_x	[0, 1]
Prevalence	p_y	[0, 1]
Confidence	$\frac{p_{xy}}{p_x}$	[0, 1]
Lift	$\frac{p_{xy}}{p_x \times p_y}$	[0, n]
Cosine	$\frac{p_{xy}}{\sqrt{p_x \times p_y}}$	[0, 1]
Leverage	$p_{xy} - (p_x \times p_y)$	[-0.25, 0.25]
Conviction	$\frac{p_x \times p_{\bar{y}}}{p_{x\bar{y}}}$	$[\frac{1}{n}, \frac{n}{4}]$
CC	$\frac{p_{xy}}{p_x} - p_y$	$[-1, 1-\frac{1}{n}]$
CF	$\begin{cases} \frac{\frac{p_{xy}}{p_x} - p_y}{p_{\bar{y}}} & if\ (\frac{p_{xy}}{p_x} - p_y) \geqslant 0 \\ \frac{\frac{p_{xy}}{p_x} - p_y}{p_y} & Otherwise \end{cases}$	[-1, 1]

¡Gracias!

UCO ONLINE