Tema 4
Actividad 1

Autor
Juan José Méndez Torrero

Uso de Weka con datos para obtener emerging patterns

Introducción

Para esta tarea se hará uso del conjunto de datos adult. Se ha elegido este conjunto de datos ya que contiene un atributo que sólo puede tomar dos valores, >50K o <=50K, además de ser la variable objetivo por la que se creó del conjunto de datos. Así, se podrá realizar un análisis de los patrones emergentes encontrados y dar una conclusión teniendo en cuenta ambos subconjuntos de datos.

Una vez dividido el conjunto de datos en las dos valores, cada uno cuenta con el siguiente número de instancias:

- Conjunto de datos >50K: 7841 instancias
- Conjunto de datos <=50K: 24719 instancias

Ambos conjuntos de datos tienen el mismo número de atributos: “age", “workclass", “education", “education-num", “marital-status”, “occupation", “relationship", “race", “sex", “capital-gain", “capital-loss", “hours-per-week", "native-country", “class”.

Al contar este conjunto con variables numéricas, se ha aplicado un filtro *NumericToNominal,* haciendo uso del programa Weka, sobre todos los atributos numéricos del conjunto de datos.

Obtención de patrones emergentes y cálculo del *Growth Ratio*

Para la extracción de patrones emergentes, primero, se ha aplicado el algoritmo *Apriori* en *Weka* con un soporte mínimo de 0.005 y una confianza del 0.3 sobre ambos subconjuntos (<=50K y >50K). El número total de patrones extraídas para ambos conjuntos ha sido limitado a 100.

Una vez se han extraído los patrones, se han volcado los resultados obtenidos en *Weka* en ficheros .txt para su procesamiento. Este procesamiento consiste en extraer el patrón y el soporte de ese patrón en porcentaje para cada subconjunto. Una vez extraído ese porcentaje, se ha comparado, para cada subconjunto de datos, qué patrones se encuentran en el conjunto de datos opuesto y se ha calculado el *Growth Ratio* para todos esos patrones utilizando la siguiente fórmula:

$$\frac{\mathrm{Sup}_{D_2}(x)}{\mathrm{Sup}_{D_1}(x)};$$

Siendo el numerador el soporte del patrón en el conjunto opuesto para el que se está calculando el *Growth Ratio*.

El *script* creado para el cálculo del *Growth Ratio* junto con los ficheros que contienen los patrones obtenidos tras ejecutar el algoritmo Apriori, serán adjuntadas en un fichero comprimido junto con la entrega.

Resultados

En la Figura 1 se puede observar los resultados obtenidos tras evaluar los patrones obtenidos en el subconjunto <=50K sobre el subconjunto >50K. Por el contrario, la Figura 2 muestra los resultados tras evaluar los patrones obtenidos en el subconjunto >50K sobre el subconjunto <=50K.

rules	support_<=50	support_>50	growth_ratio
race= White sex= Male native-country= United-States	0.481897	0.732049	1.519101
race= White sex= Male	0.529309	0.776559	1.467117
sex= Male native-country= United-States	0.541608	0.777834	1.436159
race= White sex= Male capital-loss=0 native-country= United-States	0.465796	0.658589	1.413902
sex= Male	0.611958	0.849637	1.388389
race= White sex= Male capital-loss=0	0.512318	0.698508	1.363425
sex= Male capital-loss=0 native-country= United-States	0.524131	0.700293	1.336103
sex= Male capital-loss=0	0.592621	0.763933	1.289075
race= White sex= Male capital-gain=0 native-country= United-States	0.459161	0.578115	1.259068
race= White sex= Male capital-gain=0	0.504470	0.614207	1.217529
sex= Male capital-gain=0 native-country= United-States	0.516728	0.613825	1.187907
sex= Male capital-gain=0	0.584206	0.670833	1.148280
race= White native-country= United-States	0.765241	0.854993	1.117285
race= White	0.837332	0.907665	1.083997
race= White capital-loss=0 native-country= United-States	0.740807	0.769800	1.039137
sex= Male capital-gain=0 capital-loss=0	0.564869	0.585129	1.035867
native-country= United-States	0.889923	0.914552	1.027675
race= White capital-loss=0	0.811441	0.817625	1.007622
workclass= Private race= White	0.601642	0.577988	0.960683
capital-loss=0 native-country= United-States	0.862656	0.824640	0.955931
capital-loss=0	0.969821	0.901416	0.929466
race= White capital-gain=0 native-country= United-States	0.732109	0.672363	0.918392
workclass= Private native-country= United-States	0.630851	0.579135	0.918023
race= White capital-gain=0	0.801044	0.714960	0.892535
workclass= Private	0.717383	0.632955	0.882311
capital-gain=0 native-country= United-States	0.852300	0.718148	0.842600
race= White capital-gain=0 capital-loss=0 native-country= United-States	0.707674	0.587170	0.829718
workclass= Private capital-loss=0	0.696832	0.572504	0.821580
capital-gain=0	0.958170	0.786124	0.820444
race= White capital-gain=0 capital-loss=0	0.775153	0.624920	0.806190
capital-gain=0 capital-loss=0 native-country= United-States	0.825033	0.628236	0.761468
capital-gain=0 capital-loss=0	0.927991	0.687540	0.740891

Figura 1

rules	support_>50	support_<=50	growth_ratio
capital-gain=0 capital-loss=0	0.687540	0.927991	1.349726
capital-gain=0 capital-loss=0 native-country= United-States	0.628236	0.825033	1.313253
race= White capital-gain=0 capital-loss=0	0.624920	0.775153	1.240403
capital-gain=0	0.786124	0.958170	1.218853
workclass= Private capital-loss=0	0.572504	0.696832	1.217167
race= White capital-gain=0 capital-loss=0 native-country= United-States	0.587170	0.707674	1.205229
capital-gain=0 native-country= United-States	0.718148	0.852300	1.186802
workclass= Private	0.632955	0.717383	1.133388
race= White capital-gain=0	0.714960	0.801044	1.120404
workclass= Private native-country= United-States	0.579135	0.630851	1.089298
race= White capital-gain=0 native-country= United-States	0.672363	0.732109	1.088859
capital-loss=0	0.901416	0.969821	1.075886
capital-loss=0 native-country= United-States	0.824640	0.862656	1.046101
workclass= Private race= White	0.577988	0.601642	1.040926
race= White capital-loss=0	0.817625	0.811441	0.992436
native-country= United-States	0.914552	0.889923	0.973070
sex= Male capital-gain=0 capital-loss=0	0.585129	0.564869	0.965375
race= White capital-loss=0 native-country= United-States	0.769800	0.740807	0.962337
race= White	0.907665	0.837332	0.922512
race= White native-country= United-States	0.854993	0.765241	0.895026
sex= Male capital-gain=0	0.670833	0.584206	0.870867
sex= Male capital-gain=0 native-country= United-States	0.613825	0.516728	0.841817
race= White sex= Male capital-gain=0	0.614207	0.504470	0.821335
race= White sex= Male capital-gain=0 native-country= United-States	0.578115	0.459161	0.794238
sex= Male capital-loss=0	0.763933	0.592621	0.775750
sex= Male capital-loss=0 native-country= United-States	0.700293	0.524131	0.748445
race= White sex= Male capital-loss=0	0.698508	0.512318	0.733447
sex= Male	0.849637	0.611958	0.720259
race= White sex= Male capital-loss=0 native-country= United-States	0.658589	0.465796	0.707262
sex= Male native-country= United-States	0.777834	0.541608	0.696302
race= White sex= Male	0.776559	0.529309	0.681609
race= White sex= Male native-country= United-States	0.732049	0.481897	0.658284

Figura 2

Conclusiones

Tras el cálculo del *Growth Ratio* se puede observar que no se han descubierto patrones que sean capaces de discriminar una población de otra, ya que el valor más grande de *Growth Ratio* encontrado ha sido de 1.51, no lo suficiente para confiar en la regla obtenida. Por ejemplo, el patrón con mayor valor cuenta con un soporte del 48% para el subconjunto <=50K, mientras que sólo un 73% de soporte en el subconjunto >50K, no siendo estos porcentajes lo suficientemente discriminantes.

En definitiva, el conjunto de datos no cuenta con la suficiente información como para poder encontrar patrones emergentes que cuenten con un alto *Growth Ratio*, implicando que no se ha podido encontrar un patrón que pueda identificar totalmente a un subconjunto.