Tema 4
Actividad 2

Autor
Juan José Méndez Torrero

Uso de Weka (o cualquier otra librería python o R) con datos para contrast sets

Introducción

Para la realización de esta actividad, se ha hecho uso del lenguaje de programación Python para el preprocesamiento del conjunto de datos y para el cálculo de los contrast set. Además, se ha usado el programa Weka para el cálculo de las reglas de asociación de cada subconjunto de datos.

El conjunto de datos proporcionado para esta actividad cuenta con un total de 2201 instancias y 4 atributos, entre los cuales se encuentran: *age*, *sex, survived* y *class,* siendo este último el atributo que utilizaremos para dividir el conjunto de datos.

Los script creados, junto con los conjuntos de datos utilizados serán adjuntados, junto con este PDF, en la entrega de la actividad.

Preprocesamiento

Antes de realizar el cálculo de las reglas de asociación, hemos divido el conjunto de datos dependiendo de la clase a la que pertenece cada instancias. Esta división da como resultado un total de 4 subconjuntos de datos:

- 1st class: 325 instancias
- 2nd class: 285 instancias
- 3rd class: 706 instancias
- Crew: 885 instancias

Para esta división se ha creado un script sencillo el cual lee el conjunto de datos inicial según la clase y guarda cada subconjunto en un fichero CSV para su posterior importación en el programa Weka.

Algoritmo Apriori

Una vez hemos dividido el conjunto de datos, cada subconjunto ha sido exportado al programa Weka. Una vez importados, para cada uno, se ha ejecutado el algoritmo Apriori con un soporte mínimo del 0.05 y una confianza del 0.001.

Además, para el cálculo de las reglas hemos activado la opción CAR para que el consecuente de la regla siempre contenga el atributo class, para así encontrar reglas que puedan diferenciar los subconjuntos.

Tras ejecutar el algoritmo, para cada subconjunto se han conseguido extraer la siguiente cantidad de reglas de asociación:

- 1st class: 21
- 2nd class: 23
- 3rd class: 26
- Crew class: 15

Cálculo de contrast sets

Para el cálculo de los contrast sets, se ha creado un script en el lenguaje de programación Python, el cual realiza un procesamiento sobre las reglas de asociación extraídas de cada subconjunto de datos, para extraer los antecedentes de las reglas para poder, después, evaluarlo con respecto a los demás subconjuntos.

Primero, este script calcula el soporte de cada regla en modo de porcentaje y lo guarda en un pandas dataframe junto con el antecedente. Una vez se han calculado los soportes de cada subconjunto, se ha calculado el soporte del antecedente para el resto de subconjuntos. Por último, se ha calculado la diferencia máxima calculada para cada uno de los subconjuntos.

Este dataframe ha sido ordenado según la diferencia máxima extraída. La Figura 1 muestra los soportes y diferencias calculadas para las reglas de asociación extraídas para el subconjunto de datos 1st class.

patterns	supp_1st	supp_2nd	diff_2nd	supp_3rd	diff_3rd	supp_crew	diff_crew	max_diff
age=adult sex=male	0.538462	0.589474	0.051012	0.654391	0.115929	0.974011	0.435550	0.435550
sex=male	0.553846	0.628070	0.074224	0.722380	0.168533	0.974011	0.420165	0.420165
sex=female	0.446154	0.371930	0.074224	0.277620	0.168533	0.025989	0.420165	0.420165
age=adult sex=female	0.443077	0.326316	0.116761	0.233711	0.209366	0.025989	0.417088	0.417088
sex=female survived=yes	0.433846	0.326316	0.107530	0.127479	0.306367	0.022599	0.411247	0.411247
age=adult sex=female survived=yes	0.430769	0.280702	0.150067	0.107649	0.323121	0.022599	0.408170	0.408170
sex=male survived=no	0.363077	0.540351	0.177274	0.597734	0.234657	0.757062	0.393985	0.393985
age=adult sex=male survived=no	0.363077	0.540351	0.177274	0.548159	0.185082	0.757062	0.393985	0.393985
age=adult survived=yes	0.606154	0.329825	0.276329	0.213881	0.392273	0.239548	0.366606	0.392273
age=adult survived=no	0.375385	0.585965	0.210580	0.674221	0.298836	0.760452	0.385067	0.385067
survived=no	0.375385	0.585965	0.210580	0.747875	0.372491	0.760452	0.385067	0.385067
survived=yes	0.624615	0.414035	0.210580	0.252125	0.372491	0.239548	0.385067	0.385067
sex=female survived=no	0.012308	0.045614	0.033306	0.150142	0.137834	0.000000	0.012308	0.137834
age=adult sex=male survived=yes	0.175385	0.049123	0.126262	0.106232	0.069152	0.216949	0.041565	0.126262
age=adult sex=female survived=no	0.012308	0.045614	0.033306	0.126062	0.113755	0.000000	0.012308	0.113755
sex=male survived=yes	0.190769	0.087719	0.103050	0.124646	0.066123	0.216949	0.026180	0.103050
age=adult	0.981538	0.915789	0.065749	0.888102	0.093436	1.000000	0.018462	0.093436
age=child	0.018462	0.084211	0.065749	0.111898	0.093436	0.000000	0.018462	0.093436
age=child survived=yes	0.018462	0.084211	0.065749	0.038244	0.019782	0.000000	0.018462	0.065749
age=child sex=male	0.015385	0.038596	0.023212	0.067989	0.052604	0.000000	0.015385	0.052604
age=child sex=male survived=yes	0.015385	0.038596	0.023212	0.018414	0.003029	0.000000	0.015385	0.023212

Figura 1

De igual manera, las figuras 2, 3 y 4, muestran los soportes y diferencias calculadas para los subconjuntos 2nd class, 3rd class y Crew class respectivamente.

patterns	supp_2nd	supp_1st	diff_1st	supp_3rd	diff_3rd	supp_crew	diff_crew	max_diff
age=adult sex=male	0.589474	0.538462	0.051012	0.654391	0.064917	0.974011	0.384538	0.384538
sex=male	0.628070	0.553846	0.074224	0.722380	0.094309	0.974011	0.345941	0.345941
sex=female	0.371930	0.446154	0.074224	0.277620	0.094309	0.025989	0.345941	0.345941
sex=female survived=yes	0.326316	0.433846	0.107530	0.127479	0.198837	0.022599	0.303717	0.303717
age=adult sex=female	0.326316	0.443077	0.116761	0.233711	0.092605	0.025989	0.300327	0.300327
age=adult survived=yes	0.329825	0.606154	0.276329	0.213881	0.115944	0.239548	0.090277	0.276329
age=adult sex=female survived=yes	0.280702	0.430769	0.150067	0.107649	0.173053	0.022599	0.258103	0.258103
sex=male survived=no	0.540351	0.363077	0.177274	0.597734	0.057383	0.757062	0.216711	0.216711
age=adult sex=male survived=no	0.540351	0.363077	0.177274	0.548159	0.007808	0.757062	0.216711	0.216711
survived=no	0.585965	0.375385	0.210580	0.747875	0.161910	0.760452	0.174487	0.210580
age=adult survived=no	0.585965	0.375385	0.210580	0.674221	0.088256	0.760452	0.174487	0.210580
survived=yes	0.414035	0.624615	0.210580	0.252125	0.161910	0.239548	0.174487	0.210580
age=adult sex=male survived=yes	0.049123	0.175385	0.126262	0.106232	0.057109	0.216949	0.167826	0.167826
sex=male survived=yes	0.087719	0.190769	0.103050	0.124646	0.036927	0.216949	0.129230	0.129230
sex=female survived=no	0.045614	0.012308	0.033306	0.150142	0.104528	0.000000	0.045614	0.104528
age=adult	0.915789	0.981538	0.065749	0.888102	0.027687	1.000000	0.084211	0.084211
age=child survived=yes	0.084211	0.018462	0.065749	0.038244	0.045967	0.000000	0.084211	0.084211
age=child	0.084211	0.018462	0.065749	0.111898	0.027687	0.000000	0.084211	0.084211
age=adult sex=female survived=no	0.045614	0.012308	0.033306	0.126062	0.080448	0.000000	0.045614	0.080448
age=child sex=female	0.045614	0.000000	0.045614	0.043909	0.001705	0.000000	0.045614	0.045614
age=child sex=female survived=yes	0.045614	0.000000	0.045614	0.019830	0.025784	0.000000	0.045614	0.045614
age=child sex=male	0.038596	0.015385	0.023212	0.067989	0.029392	0.000000	0.038596	0.038596
age=child sex=male survived=yes	0.038596	0.015385	0.023212	0.018414	0.020183	0.000000	0.038596	0.038596

Figura 2

patterns	supp_3rd	supp_1st	diff_1st	supp_2nd	diff_2nd	supp_crew	diff_crew	max_diff
age=adult survived=yes	0.213881	0.606154	0.392273	0.329825	0.115944	0.239548	0.025667	0.392273
survived=no	0.747875	0.375385	0.372491	0.585965	0.161910	0.760452	0.012577	0.372491
survived=yes	0.252125	0.624615	0.372491	0.414035	0.161910	0.239548	0.012577	0.372491
age=adult sex=female survived=yes	0.107649	0.430769	0.323121	0.280702	0.173053	0.022599	0.085050	0.323121
age=adult sex=male	0.654391	0.538462	0.115929	0.589474	0.064917	0.974011	0.319620	0.319620
sex=female survived=yes	0.127479	0.433846	0.306367	0.326316	0.198837	0.022599	0.104880	0.306367
age=adult survived=no	0.674221	0.375385	0.298836	0.585965	0.088256	0.760452	0.086231	0.298836
sex=male	0.722380	0.553846	0.168533	0.628070	0.094309	0.974011	0.251632	0.251632
sex=female	0.277620	0.446154	0.168533	0.371930	0.094309	0.025989	0.251632	0.251632
sex=male survived=no	0.597734	0.363077	0.234657	0.540351	0.057383	0.757062	0.159328	0.234657
age=adult sex=female	0.233711	0.443077	0.209366	0.326316	0.092605	0.025989	0.207722	0.209366
age=adult sex=male survived=no	0.548159	0.363077	0.185082	0.540351	0.007808	0.757062	0.208904	0.208904
sex=female survived=no	0.150142	0.012308	0.137834	0.045614	0.104528	0.000000	0.150142	0.150142
age=adult sex=female survived=no	0.126062	0.012308	0.113755	0.045614	0.080448	0.000000	0.126062	0.126062
age=adult	0.888102	0.981538	0.093436	0.915789	0.027687	1.000000	0.111898	0.111898
age=child	0.111898	0.018462	0.093436	0.084211	0.027687	0.000000	0.111898	0.111898
age=adult sex=male survived=yes	0.106232	0.175385	0.069152	0.049123	0.057109	0.216949	0.110717	0.110717
sex=male survived=yes	0.124646	0.190769	0.066123	0.087719	0.036927	0.216949	0.092303	0.092303
age=child survived=no	0.073654	0.000000	0.073654	0.000000	0.073654	0.000000	0.073654	0.073654
age=child sex=male	0.067989	0.015385	0.052604	0.038596	0.029392	0.000000	0.067989	0.067989
age=child sex=male survived=no	0.049575	0.000000	0.049575	0.000000	0.049575	0.000000	0.049575	0.049575
age=child survived=yes	0.038244	0.018462	0.019782	0.084211	0.045967	0.000000	0.038244	0.045967
age=child sex=female	0.043909	0.000000	0.043909	0.045614	0.001705	0.000000	0.043909	0.043909
age=child sex=female survived=yes	0.019830	0.000000	0.019830	0.045614	0.025784	0.000000	0.019830	0.025784
age=child sex=female survived=no	0.024079	0.000000	0.024079	0.000000	0.024079	0.000000	0.024079	0.024079
age=child sex=male survived=yes	0.018414	0.015385	0.003029	0.038596	0.020183	0.000000	0.018414	0.020183

Figura 3

patterns	supp_crew	supp_1st	diff_1st	supp_2nd	diff_2nd	supp_3rd	diff_3rd	max_diff
age=adult sex=male	0.974011	0.538462	0.435550	0.589474	0.384538	0.654391	0.319620	0.435550
sex=male	0.974011	0.553846	0.420165	0.628070	0.345941	0.722380	0.251632	0.420165
sex=female	0.025989	0.446154	0.420165	0.371930	0.345941	0.277620	0.251632	0.420165
age=adult sex=female	0.025989	0.443077	0.417088	0.326316	0.300327	0.233711	0.207722	0.417088
sex=female survived=yes	0.022599	0.433846	0.411247	0.326316	0.303717	0.127479	0.104880	0.411247
age=adult sex=female survived=yes	0.022599	0.430769	0.408170	0.280702	0.258103	0.107649	0.085050	0.408170
sex=male survived=no	0.757062	0.363077	0.393985	0.540351	0.216711	0.597734	0.159328	0.393985
age=adult sex=male survived=no	0.757062	0.363077	0.393985	0.540351	0.216711	0.548159	0.208904	0.393985
survived=no	0.760452	0.375385	0.385067	0.585965	0.174487	0.747875	0.012577	0.385067
age=adult survived=no	0.760452	0.375385	0.385067	0.585965	0.174487	0.674221	0.086231	0.385067
survived=yes	0.239548	0.624615	0.385067	0.414035	0.174487	0.252125	0.012577	0.385067
age=adult survived=yes	0.239548	0.606154	0.366606	0.329825	0.090277	0.213881	0.025667	0.366606
age=adult sex=male survived=yes	0.216949	0.175385	0.041565	0.049123	0.167826	0.106232	0.110717	0.167826
sex=male survived=yes	0.216949	0.190769	0.026180	0.087719	0.129230	0.124646	0.092303	0.129230
age=adult	1.000000	0.981538	0.018462	0.915789	0.084211	0.888102	0.111898	0.111898

Figura 4

Conclusiones

Como se puede observar en las figuras anteriormente mostradas, los resultados obtenidos nos indican que no se ha encontrado ninguna regla que pueda diferenciar un conjunto de datos en su totalidad. Sólo algunas reglas han podido superar la diferencia de 0.4, y estas reglas están relacionadas con los subconjuntos 1st class y Crew class.

Un ejemplo sería la primera regla observada en las figuras 1 y 4. Como se puede observar, hemos obtenido una diferencia del 0.43, indicando que las personas que son adultas y son hombres, tiene más probabilidad de pertenecer a la clase Crew que a la clase 1st.

Otro ejemplo podría ser la regla de asociación obtenida “age=adult, sex=female”, la cual ha obtenido una diferencia del 0.417. Esto nos indica que, si una nueva persona, mujer adulta, entrara dentro del conjunto de datos, sería muy poco probable que esta instancia se clasificara como clase Crew.

Por último, se puede observar que las reglas de asociación encontradas no son lo suficientemente restrictivas como para poder diferencias las clases 2nd y 3rd.