Enunciat del projecte de PROP
Quadrimestre de primavera, curs 19/20

Regles d'associació

El procés de mineria de dades tracta de descobrir patrons o relacions en un conjunt de dades donat. Una de les possibles tècniques de *data mining* són les regles d'associació. Una regla d'associació és una implicació entre condicions sobre atributs de l'estil

> SI (practica_esport = fals) I (fruita_en_dieta <10%) I (colesterol_en_sang > 3g/L)
> LLAVORS malaltia_cardiaca = cert

Aquí, "practica_esport" i "malaltia_cardiaca" són atributs booleans mentre que "fruita_en_dieta" i "colesterol_en_sang" són atributs numèrics.

Es tracta de fer una eina que trobi les regles d'associació que en les dades superen certs llindars de suport (% de registres on l'antecedent i el conseqüent de la regla són certs) i de confiança (% de registres on, sent l'antecedent cert, a més el conseqüent és cert).

Les dades vindran en fitxers de text. Cada línia del fitxer s'anomenarà registre, i contindrà els valors de diversos atributs. Com a mínim han de poder haver-hi atributs numèrics, booleans o categòrics (ex: blau, verd, vermell).

El programa ha d'oferir un entorn tant còmode com sigui possible per al necessari preprocés de les dades. Obligatòriament, el programa ha de:
- Permetre la definició d'atributs
- Incloure la discretització o binarització d'atributs numèrics.
- Donar l'opció de que l'usuari pugui guardar aquestes dades preprocessades per no haver de repetir el preprocés en treballar de nou amb el mateix joc de dades.

I opcionalment:
- Es podria incloure la inspecció visual de les dades, eliminació de registres escollits (p.ex., *outliers* o redundants), eliminació d'atributs (que podrien ser, p.ex., poc rellevants), afegir atributs nous (p.ex., calculats o derivats dels altres), etc.
- Es podria tractar el cas dels *missing values* (emplenar valors d'atributs que puguin faltar).

L'entorn haurà de permetre les següents funcionalitats:
1. Inducció de regles d'associació a partir d'un conjunt de dades
2. Validació d'un conjunt de regles obtingudes en un conjunt de dades diferent (és a dir, calcular la rellevància i fiabilitat de les regles en el nou conjunt
3. Preprocés de les dades d'un conjunt
4. Guardar i recuperar el resultat de l'algorisme principal (regles d'associació trobades)

Seria bo que l'usuari pogués triar entre diversos nivells d'interactivitat: des del cas que el programa ho fa gairebé tot automàticament fins aquell en què l'usuari pot establir manualment tots els paràmetres de l'algorisme, seguir el procés pas a pas, modificar el resultat final, etc.

Funcionalitats principals a entregar al primer lliurament:
1. Inducció de regles d'associació utilitzant l'algorisme *Apriori*
2. Funcionalitats automàtiques definides pel preprocés de les dades (aquelles en què no intervé interactivament l'usuari)
3. Validació d'un conjunt de regles

Dates dels lliuraments:

- Primer: divendres 24 d'abril de 2020
- Segon: divendres 22 de maig de 2020
- Tercer: divendres 29 de maig de 2020 (lliuraments interactius: a partir del 2 de juny de 2020)