A survey on pre-processing techniques: relevant issues in the context of environmental data mining

Noa Yu Ventura Vila

MEI Q1 2023-2024

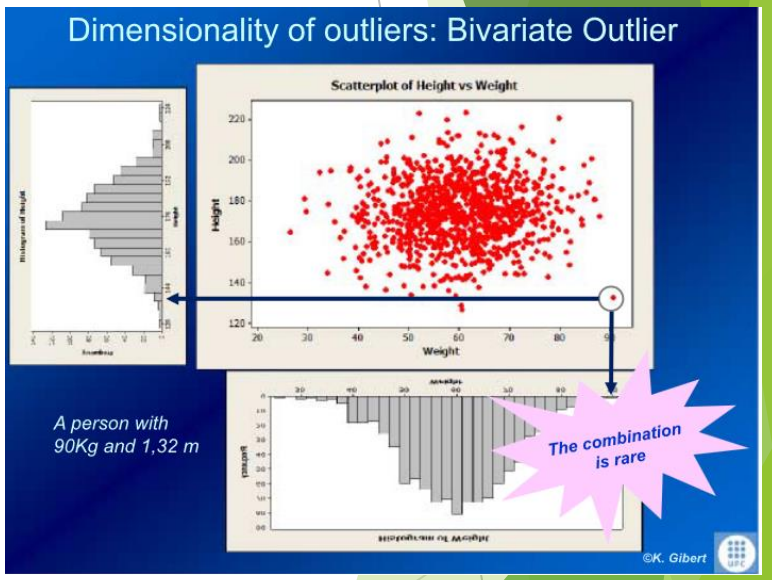
Gibert, K., Sànchez-Marrè, M., and Izquierdo, J. (2016). A survey on pre-processing techniques: Relevant issues in the context of environmental data mining. AI Communications, 29(5), 627-663

Capítol 5: visualització de dades

- Visualitzar les dades amb grafs és una estratègia molt bona per a tenir una idea de com estan distribuïdes.
- Grafs clàssics no són suficients per a representar sets de data tan grans i sofisticats. D'altres són:
  - Grafs de distribució
  - Grafs de 2-5 dimensions
  - Etc.
- Per a fer el pre-processament grafs d'aquests són molt útils.
- Mètodes de ML inductiu necessiten dades balancejades per a tenir resultats confiables. Es poden fer tests en processos repetitius però a canvi aquests mateixos tenen les seves tècniques i assumpcions.

Capítol 6: valors atípics i observacions influents: detecció i tractament

- Cada valor atípic té una natura, una raó per la qual és atípic. Segons el seu tipus li apliquem un tractament o un altre. El tractament no només depèn del tipus sinó també d'altres factors:
  - Tipus de mètode:
    - **Mètodes robusts**: fins a cert grau poden resistir distorsions produïdes per un valor atípic.
    - **Mètodes no robusts**: és molt important ser capaços de "netejar" el dataset abans de fer servir les dades.
  - **Observacions influents**: observació que determina molt fortament els resultats d'un anàlisi d'un dataset. Només és un problema quan està lluny de les altres observacions i no segueix el model.
- **Dimensions** dels valors atípics: tenir valors per a dues variables relacionades pot donar que no hi ha cap valor atípic, però al relacionar-los i fer el graf 2D podem veure que ara sí que és un valor atípic.



Capítol 6: valors atípics i observacions influents: detecció i tractament

- **Natura** dels valors atípics:
  - Error: que no sigui físicament possible i hagi estat un error.
  - Punt informatiu: que sigui un cas concret i especial.
  - Individu d'una altra població.
  - Valor extrem intrínsec: que realment sigui un valor extrem.
  - Codi faltant: falta informació d'aquell punt.
- **Detecció** de valors atípics: de manera visual només ens serveix fins a 3D, per tant necessitem altres mètodes que tinguin en compte el data set sencer.
  - Quan el punt és una equivocació lo millor és tornar a observar-ho i corregir-lo. Si no es pot observar un altre cop doncs s'ha de substituir per un altre.
  - Si és perquè no forma part de la població (per exemple, n'és d'una altra) i no és representatiu d'aquesta llavors s'ha de fer encara més gran aquesta. O investigar una subpoblació.
  - Quan és un valor atípic natural llavors no sol ser una observació influent i segueix el model, de manera que no ens hem de preocupar per aquest tipus.
  - Quan és un codi numèric per representar que falten dades llavors s'ha de trobar el valor real o treure-ho del data set.
- És molt important després explicar quins mètodes s'han fet servir perquè l'usuari final ho pugui tenir present.

Aplicacions i resultats

- El pre-processament de dades és molt important fer-lo.
- No tots els valors atípics són perjudicials però s'han de tractar de totes maneres, i els que ho són és imprescindible tractar-los correctament.
- A vegades no es pot fer servir una manera gràfica de representar les dades per a trobar els valors atípics.
- Documentar el tractament que s'ha fet amb els valors atípics és mandatori.