6. Óra

Adatelemzési platformok, BME, 2018. Február 28., IV. Gyakorlati óra.

Telecom adatelemzési probléma

Olyan ügyfélklasztereket akarunk alkotni, ahol az egyes embereknek nagyon hasonló a telefonálási szokása.

Előkészítés

- Not-supported operator: Custom model visualizer
- CSV helyett txt-t keresünk
    - Breakpointot teszünk a csv-re, hogy megnézzük, hogy mit csinálunk

Lépések

Nem készítjük elő az adatot, sőt kikapcsoljuk a következő operátorokat:

1. Data preparation
2. Outlier detections
3. Calculating ssb és sst
4. Flatten, és két multiply

Illetve kiszedjük a második kimeneteket a clustering operatorból

Klaszterező operátorok

1. Clustering 1
    1. Complete link: a legtávolabbiakat nézi
2. Clustering 2
    1. K-means
    2. Numeric measure
    3. Euclidean distance

Első futtatás

Run: missing values

1. Multiply - K-means clustering közé: Filter examples
2. Hiba: nem-numerikus attribútumokkal nem tud mit csinálni (customer id-ra)
3. Kapcsoljuk ki a Hierarchikus klaszterezőt

Data preparation operátor

1. Kijelöli a customer id-t
2. Kizárja a nominális változókat
    - Csinálhatná, hogy kijelöli csak a numerikusokat, de akkor nem venné ki a binomiálisokat (ami nem nominális)
3. Normalizál, range tranformációval 0 és 1 közé: ne dominájanak azok a változók, amelyek nagyságrendekkel nagyobbak mint más változók (pl darab és perc változók)
4. Kiszűrjük a hiányzó értékeket
5. Kiszűrjük az egymással minimum +/- 0.8-cal korreláló változókat
    - Ezzel próbálja megelőzni, hogy túldominálják az elemzést (mivel a hatásuk össze fog adódni)

Data preparation bekapcsolása

1. Visszatesszük a data preparation
2. Kitöröljük a filter-t
3. Visszakapcsoljuk a Hierarchikus klaszterezést
4. Breakpoint a data preparation-ra
5. Lefuttatjuk breakpointig
    - Sok helyen nem 1 a maximum, mivel előbb normalizáltunk, és aztán filtereztünk
6. Végig Lefuttatjuk

Klaszterelemzés

1. 5 db k-means klastzer
    1. Nagy különbség a klaszterek között
    2. Plot: mutatja a klaszter középpontokat
    3. Az alső kettő klaszter nagyon közel van:
        - Öreg, keveset használók
        - Fiatal keveset használók

4. Nem kell minden attribútum szerint különbözőnek lenniük
5. A k-means kezdőpontjai véletlenszerűek, de maga az elemzés megismételhető

2. Hierarchikus klaszterezés
    - A klaszterek száma, az elemek felezéséből jön ki (talán?)
    - Dendogram-nézet: jobb oldalt a magas pontok talán outlierek

SSB/SST számolás

1. Visszatesszük a ssb/sst számolókat
2. Flatten clustering: a Hierarchikus clustering és a ssb/sst kalkulátor közé tesszük
3. Futtassuk:
    - SSB/SST kicsik
    - De probléma lehet, hogy a két klaszterezés egymáshoz nincs normálva, mivel az egyiknél csak 5 klaszter van, míg a másiknál 10
4. Klaszter vizualizáló (lincensz kell hozzá)
    - Klaszterek vizuális szétválasztása alapból nem nagyon sikerül, erre jó a dimenziók csökkentése
5. Rapidminer: metaoperatorok
    - Loop operátorral ki lehet számolni a 'könyökpontot' az összes esetre

Outlier detection

Outlier detection operátor:

1. Távolság alapú
2. Lehetőségek
    - Dimemziók mentén
    - Hierarchikus klazsterezés keztedi értékei
    - Statisztikai nézetben: a nagy szórású változókat megvizsgáljuk
3. DE
    1. Sorokra akarjuk megtudni
    2. A fentiek leginább a külső kiugrókat találja meg, a 'belsőket' nem

Minden egyes pontra kiszámoljuk, hogy mely pontok vannak hozzájuk a legközelebb (LOF, local outlier factor)