# Pattern mining for label ranking

Pinho Rebelo de Sá, C.F.

# Nederlandse Samenvatting

Voorkeuren hebben altijd een rol gespeeld in ons dagelijks leven. Het kopen van de juiste auto, het kiezen van een geschikt huis, en zelfs beslissen wat te eten, zijn enkele triviale voorbeelden van keuzes die expliciet of impliciet iets vertellen over onze voorkeuren. De recente trend om alsmaar groeiende hoeveelheden data te verzamelen geldt ook voor data over voorkeuren.

Het extraheren en modelleren van voorkeuren levert ons waardevolle informatie over de keuzes van groepen en individuen. In gebieden als e-commerce, die typisch gaan over de keuzes van duizenden mensen, kan het vastleggen van voorkeuren een moeilijke taak zijn. Om deze redenen zijn methoden uit de kunstmatige intelligentie (specifieker, het machinaal leren) van toenemend belang voor het ontdekken en automatisch leren van modellen over voorkeuren.

Het deelgebied van machinaal leren dat zich richt op de studie en het modelleren van voorkeuren staat bekend als *Preference Learning* (PL). We kijken specifiek naar een deeltaak binnen PL, namelijk Label Ranking (LR). Kort gezegd, een LR dataset bestaat uit een verzameling observaties, beschreven door attributen (de onafhankelijk variabelen), en een ordening over een (eindige) verzameling labels (de afhankelijke variabele). In LR zijn we geïnteresseerd in het voorspellen van de ordening van de labels, gebaseerd op de waarden van de onafhankelijk variabelen.

In dit promotieproject zijn meerdere aanpakken voor het LR-probleem voorgesteld en geanalyseerd. We onderzochten Label Ranking Association Rules (LRAR), wat het LR-equivalent is van de Class Association Rules. Een LRAR is een associatieregel waarvan de items gebaseerd zijn op waarden voor de onafhankelijke variabelen, en de conclusie van de regel een ordening over de labels voorstelt. Verder stelden we de zogenaamde Pairwise Association Rules (PAR) voor, die gedefinieerd zijn als associatieregels waarvan de conclusie een verzameling van paarsgewijze voorkeuren weergeeft. PAR kunnen, net als LRAR, zowel beschrijvend als voorspellend gebruikt worden.

Onze analyse is echter gericht op de beschrijvende eigenschappen, terwijl de LRAR vooral gemodelleerd zijn als voorspellende modellen.

Het is algemeen bekend dat preprocessing (het voorbewerken van data) een belangrijke component is van machinaal leren. Dit is net zo zeer het geval voor LR als voor elke andere machinaal leren taak. Om een voorbeeld te noemen: LRAR, net als associatieregels, kunnen niet direct met numerieke data omgaan, zodat de data vooraf gediscretiseerd moet worden. Er bestond echter nog niet zo'n methode voor het geval van LR. Om die reden stelden we twee LR-specifieke discretisatiemethoden voor. Beide methoden zijn gebaseerd op nieuwe maten voor de entropie van ordeningen (ranking entropy).

Het grootste deel van dit project is gericht op methoden voor het ontdekken van patronen. Echter, gezien de de populariteit van beslisboommethoden en hoe deze methoden inzichtelijk informatie over de taak kunnen weergeven, introduceerden we Entropy Ranking Trees. Hoewel eerdere methoden bestonden voor het aanpassen van beslisboomalgoritmen aan LR, was het natuurlijk om vanuit een maat voor de entropie van ordeningen te onderzoeken hoe deze in deze algoritmen geïntegreerd konden worden. Een andere erg populaire modelleeraanpak is die van de ensemble learning. Specifiek het Random Forest (willekeurige woud) algoritme is zeer succesvol gebleken, maar was nog nooit aangepast naar LR. Het betreft hier een ensemble-methode die verschillende bomen combineert die op willekeurige manier verkregen zijn. We stelden dus een ensemble voor Label Ranking voor die gebaseerd is op Random Forest, genaamd Label Ranking Forest.

Onze zoektocht door het veld van de preference learning werd voortgezet met het ontdekken van lokale patronen. De taak heet Exceptional Pattern Mining (het ontdekken van uitzonderlijke patronen) en kan gezien worden als het ontdekken van lokale patronen over deelverzamelingen van de observaties waarvan de voorkeursverhoudingen tussen een deel van de labels afwijken van de norm. In andere woorden, het is een variant op de zogenaamde Subgroup Discovery taak, waarbij de doelvariabele een ordening betreft. We gebruiken drie kwaliteitsmaten die subgroepen benadrukken die uitzonderlijke voorkeuren vertonen, waarbij de specifieke nadruk wat betreft 'uitzonderlijk' verschilt per maat. De resultaten tonen ook aan hoe de visualisatie van voorkeuren in een voorkeursmatrix (Preference Matrix) kan helpen bij het interpreteren van subgroepen van uitzonderlijke voorkeuren.

Als laatste suggereerden we een aanpak voor het testen van de relatie tussen ordeningen en de onafhankelijke variabelen in een LR dataset. Zoals in andere leertaken met toezicht (supervised learning) is randomiseren van de

doelvariabele door middel van omwisselen (swap randomisation) gebruikt om deze test te doen. We stelden twee manieren voor om dit te doen voor LR, en hebben ze toegepast op LR datasets.

De experimentele resultaten tonen het potentieel van de genoemde aanpakken.