Pattern mining for label ranking

Pinho Rebelo de Sá, C.F.

Resumo

É comum lidarmos com preferências no nosso dia-a-dia. Quando compramos um carro, procuramos uma casa ou mesmo quando decidimos o que comer, estamos a tomar decisões que revelam informação sobre as nossas preferências. Nos dias que correm, cada vez mais dados são recolhidos, onde se incluem também dados sobre preferências.

A extracção e a criação de modelos de preferências, podem fornecer informações valiosas sobre determinados grupos ou indivíduos. Em áreas de negócio como o comércio electrónico, que lidam com informações de milhares de utilizadores, a modelação de preferências pode constituir um desafio. Por isso, métodos de Inteligência Artificial (em particular, *machine learning*), têm sido cada vez mais usados para a descoberta e aprendizagem automática de modelos sobre preferências.

A área de *machine learning* que lida a modelação e estudo de preferências é chamada de *Preference Learning* (PL). O tema deste doutoramento, foca em uma sub-área de PL denominada de Label Ranking (LR). Em LR, os dados consistem em observações constituídas por *atributos* (variáveis independentes) e *rankings* de um conjunto finito de objetos (*target* ou variáveis dependentes). O objectivo é prever esses rankings para novas observações, baseando-se nos valores fornecidos das variáveis independentes. Neste trabalho, foram propostas várias abordagens ao problema de LR.

Exploramos as Label Ranking Association Rules (LRAR), que são equivalentes às Class Association Rules no contexto de LR. Por definição, uma LRAR é uma *regra de associação* onde o *antecedente* é um conjunto de itens baseados nos valores das variáveis independentes, e o *consequente* é um ranking. Com uma estrutura semelhante, também propusemos as Pairwise Association Rules (PAR), definidas como regras de associação onde o consequente é um conjunto de *pairwise comparisons*. Tal como as LRAR, as PAR podem ser usadas como abordagens descritivas e como modelos de previsão. No entanto, a nossa análise foca-se nas propriedades descritivas das PAR,

enquanto que as LRAR foram usadas como modelos preditivos.

Métodos de pré-processamento são uma parte essencial nos processos de *machine learning*. As LRAR, tal como regras de associação comuns, não conseguem lidar directamente com variáveis numéricas, que, por sua vez, têm que ser discretizadas à priori. Dado que não existiam métodos de discretização especificamente para dados de LR, foram propostas duas abordagens baseadas em medidas de *entropia de rankings*.

Apesar de a maior parte deste trabalho focar em métodos de *pattern mining*, tendo em conta a popularidade de métodos como *árvores de decisão* e pela forma clara como expressam informação, propusemos as *Entropy Ranking Trees*. Mesmo já existindo árvores de decisão para LR, uma vez que tinha sido proposta a medida de entropia de rankings, achamos natural estudar a sua integração neste modelos. Outra abordagem também muito popular em *machine learning* é *ensemble learning*. Nomeadamente, um algoritmo denominado *Random Forests* (RF), tem sido bem sucedido, mas nunca tinha sido adaptado para LR. O método de RF, combina vários modelos de árvores de decisão que são geradas usando algumas técnicas de randomização. Por isso, propusemos *ensembles* de árvores de decisão, baseados em RF, que chamamos de Label Ranking Forests.

Continuamos a nossa jornada na área de PL, combinando-a com técnicas de *local pattern mining*. O método, a que chamamos de Exceptional Preferences Mining (EPM), pode ser visto como uma técnica de *local pattern mining* que encontra sub-conjuntos de observações onde as preferências se desviam do normal. Por outras palavras, é uma variante de *Subgroup Discovery*, em que os rankings são o *target*. Par isso, foram propostas três medidas (*quality measures*) que procuram sub-conjuntos que apresentem preferências consideradas excepcionais. Os resultados obtidos realçam também uma forma proposta de representar preferências, a *Preference Matrix*.

Por último, apresentamos formas de testar a relação entre variáveis independentes e rankings, em dados de LR. Uma técnica denominada *target swap randomization*, também aplicada em problemas de classificação, foi implementada para este tipo de testes. Além disso, também foram propostas duas variantes, baseadas em *target swap randomization*, para se adequarem melhor ao problema.

Os resultados experimentais apresentados demonstram o potencial dos métodos aqui propostos.