07/21 DM 下午場

洪子軒

Sent: Thursday, July 21, 2016 4:06 PM
To: 洪子軒

GMC（Gaussian Mixture Clustering 貝氏理論之應用） EM clustering
計算每群的中心和共變數（n*n 距陣）
→屬於每一群的機率（前置機率 prior ，多變量常態分配計算可能性）
→By MLE（expectation-maximization algorithm）
→所有資料點都會對每個中心產生貢獻
→Mahalanbis distance

【關連規則】因果性不成立，只有關連

support 支持度：左邊發生的機率（越大越好）
confidence 信賴度：在左邊發生的條件下，右邊也會發生的機率（越大越好）
定臨界值
lift 提升度：統計上的（同時發生> 個別發生機率相乘 → 正相關；相等 → 獨立） P(Y|X) / P(Y)
PS： P(X,Y) - P(X)P(Y)

＞apiori principle：If an itemset is frequent, then all of its subsets must also be frequent
子集合的 support 永遠>= 母集合
所以子集合 support 不超過臨界值的情況下，往下和其他商品結合的 support 也不會超過（反單調 anti-monotone property）→ pruning
apriori algorithm：join + prune 直到找不到 large itemset

縮減 pruning：商品項目、交易次數…
找出來的規則還是太多，例如太單調 AB → D 後不需要再看 ABC → D（早知道，不感興趣），如何砍？
還有其他指標

其他演算法：FP-growth

反單調：confidence of rules generated from the same itemset
c(ABC→D) >= c(AB → CD) >= c(A → BCD)

應用：cross selling

--
洪子軒 Tzu-Hsuan Hung
中華電信研究院 巨量資料所
TEL: (03)-4245128
Email: Lucas@cht.com.tw
32661桃園市楊梅區電研路99號