Importanta suportului, confidența si corelația in identificarea regulilor de asociere

Matei-Bledea Alexandru

15 noiembrie 2022

1 Introducere

Odată cu avansul in domeniul tehnologiei, au apărut și magazinele online. [1] Pe lângă faptul că acestea au permis clienților o experiența foarte comodă și rapidă de cumpărături, ele au permis de asemenea și colectarea imensă de date obținută pe urma tranzacțiilor făcute de clienți. De când resursele calculatoarelor au devenit tot mai puternice, a devenit posibilă procesarea unor seturi imense de date. Această analiză duce la identificarea anumitor strategii pe care deținătorii magazinelor le pot aplica pentru a-și maximiza profiturile, iar una dintre aceste strategii este identificarea anumitor reguli de asociere între anumite produse, care permite prezentarea produselor asociate într-un mod atractiv pentru clienți, crescând șansele de achiziție.

2 Reguli de asociere

În această secțiune, vom prezenta o definiție a regulilor de asociere, concentrându-ne mai mult pe Market Basket Analysis. Mai mult, vom defini niște noțiuni care ajută în identificarea acestor reguli, și formulele care calculează valorile acestor noțiuni. Vom încheia capitolul ilustrând intuiția din spatele noțiunilor, precum și analizând cum sunt legate valorile formulelor.

2.1 Market Basket Analysis si reguli de asociere

Semnificația expresiei "Market Basket Analysis" se poate intui din traducerea ei în limba română: "analiza coșului de cumpărături." [2] Aceasta este o tehnică in Data Mining a cărui scop este identificarea anumitor pattern-uri din totalul achizițiilor efectuate de către clienții unui magazin. Pattern-urile descoperite pot fi foarte utile în implementarea anumitor strategii pentru a maximiza profitul obținut de către magazin.

Figura 1: Exemplu de produse cumpărate într-o singura tranzactie [3]

Printre pattern-urile posibile, regăsim determinarea unor reguli de asociere între anumite seturi de produse. Nu vom da o definiție formală unei reguli de asociere deoarece ele pot fi oarecum subiective, însă cel mai ușor mod de a explica ce reprezintă aceste reguli este prin intermediul unui exemplu. Să presupunem ca într-un magazin se vinde lapte, respectiv cereale, iar din analiza tranzacțiilor efectuate se observă ca aceste două produse sunt cumpărate frecvent împreună. O posibilă regula de asociere ar fi sa spunem că daca un client cumpăra cereale, este foarte probabil ca acel client sa cumpere și lapte, probabil pentru a-l folosi împreuna cu cerealele. Notația pe care o vom folosi ca sa ilustrăm asocierea dintre cereale si lapte (și anume ca dacă un client cumpară cereale, sunt șanse mari să cumpere și lapte) ar fi {cereale} $\Rightarrow$ {$lapte$}.

Aceste reguli de asociere se pot folosi pentru implementarea unor strategii eficiente a căror scop este maximizarea profitului magazinului. Pornind de la exemplul precedent și presupunând că s-a identificat laptele ca fiind adesea cumpărat de clienții care își cumpăra cereale, o posibilă strategie ar fi amplasarea celor doua obiecte în locații îndepărtate din magazin. Aceasta strategie expune clienții la o mulțime de alte produse pe drumul efectuat în colectarea celor doua obiecte, crescând posibilitatea clientului de a mai adaugă și alte articole in coșul de cumpărături.

O altă posibilă strategie din punct de vedere al prețului acestor alimente ar fi stabilirea unei reduceri la articolul principal, în cazul nostru cerealele. Aceasta creste posibilitatea ca un client sa îl cumpere, iar el fiind achiziționat deseori împreuna cu articolul secundar, va crește de asemenea numărul de vânzări ale articolului secundar, in cazul acesta fiind laptele.

În domeniul online, una din posibilele strategii ar fi implementarea unui algoritm care in momentul checkout-ului îi arată clientului o posibilă listă de obiecte pe care le-ar mai putea cumpăra, determinată în funcție de ce prezintă clientul în coșul de cumpărături virtual. De asemenea, acea listă de articole ar putea conține și articole care au fost identificate ca fiind achiziționate frecvent de către acel client, după mai multe tranzacții. Una din aplicațiile care a adoptat recent acești algoritmi este Glovo.[4]

Bineînțeles, acestea sunt doar niște reguli intuitive oferite cu scop de exemplu, dar actuala implementare a lor poate să difere sau să varieze. In continuarea capitolului, vom defini niște noțiuni care se pot folosi in identificarea acestor reguli, și ilustra care este relația dintre ele. Noțiunile despre care vom discuta sunt suportul, confidența si corelația. [5] Pentru secțiunile de mai jos, vom presupune că încercam să identificam o regula de asociere între seturile de obiecte A si B, si anume $A \Rightarrow B$. Precizăm ca obiectele din cele două seturi sunt distincte.

2.2 Suportul

Suportul unei mulțimi de obiecte D este definit ca raportul dintre numărul de tranzacții care conțin toate elementele din mulțimea respectiva și numărul total de tranzacții efective, și se notează cu $P(D)$, unde $P(E)$ reprezintă de asemenea probabilitatea ca evenimentul E să se întâmple. Aceasta ne arată cât de des obiectele din mulțimea D apar în coșul de cumpărături. În particular, ne interesează suportul tranzacțiilor care conțin și A și B:

$$\text{suport}(A, B) = P(A \cap B) = \frac{\text{number of transactions containing both } A \text{ and } B}{\text{total number of transactions}}$$

De preferință, se caută reguli de asociere în care acest procent este cât mai mare, deoarece sunt aplicabile unui număr mai mare de tranzacții.

2.3 Confidența

Confidența este o măsura definită asupra ambelor mulțimi din regula de asociere, și anume reprezintă raportul dintre numărul de tranzacții conținând obiectele din ambele mulțimi și numărul de tranzacții conținând doar partea stânga a regulii de asociere. In cazul unei asocieri de tipul $A \Rightarrow B$, ea s-ar traduce ca fiind probabilitatea de a cumpăra obiectele din mulțimea B, dat fiind că s-au cumpărat și toate obiectele din mulțimea A:

$$\text{confidența}(A, B) = P(B|A) = \frac{\text{number of transactions containing both } A \text{ and B}}{\text{number of transactions containing } A}$$

Cu cât este mai mare confidența, cu atâta sunt mai mari șansele să se cumpere obiectele din mulțimea B, daca s-au cumpărat toate obiectele din mulțimea A, așadar se prefera in general reguli de asociere cu confidența mai mare.

Un lucru care poate fi remarcat este relația dintre confidența si suport. Se poate observa ca daca împărțim fiecare parte a raportului din definiția confidentei cu numărul total de tranzacții, obținem chiar raportul dintre suportul mulțimilor $A \cap B$ si A. Cu alte cuvinte, putem scrie:

$$\text{confidența}(A, B) = \frac{\text{suport}(A, B)}{\text{suport}(A)}$$

O alta remarcă care poate fi adusă ar fi faptul că expresia confidenței reprezintă de fapt o probabilitate condiționala, și anume probabilitatea lui B condiționata de A.

2.4 Corelatia

O ultimă măsura de asociere pe care o vom folosi este corelația. Măsura de corelație este definită prin formula următoare:

$$\text{corelație}(A, B) = \frac{P(A \cap B)}{P(A)P(B)}$$

Înlocuind cu definiția suportului, regula poate fi exprimată doar funcție de suportul mulțimilor A, B și $A \cap B$:

$$\text{corelație}(A, B) = \frac{\text{suport}(A \cap B)}{\text{suport}(A) \cdot \text{suport}(B)}$$

Aceasta măsura joaca și ea un rol important în identificarea regulilor de asociere. [6] Pornind de la regula de asociere dintre A și B, putem identifica următoarele 3 situații:

- Daca corelația$(A, B) < 1$, atunci șansele ca elementele din setul B sa fie cumpărate daca elementele din setul A au fost cumpărate sunt mici.
- Daca corelația$(A, B) > 1$, atunci șansele ca elementele din setul B sa fie cumpărate daca elementele din setul A au fost cumpărate sunt mari.
- Daca corelația$(A, B) = 1$, atunci nu se poate identifica o asociere clara intre seturile A si B.

2.5 Identificarea unei reguli de asociere

In secțiunile de mai sus, am definit diverse noțiuni, unde acum vom explica ce rol joacă îm identificarea anumitor reguli de asociere. În general, următoarele condiții trebuiesc îndeplinite pentru a putea fi considerată existența unei reguli de asociere:

- O regulă este interesanta daca satisface valori minime pentru suport si confidență, unde aceste praguri sunt specificate de către un expert in domeniul aplicației;
- Pe lângă acestea, corelația trebuie sa fie cât mai pozitivă, adică cât mai mare decât 1.

Urmează să discutăm in următoarea subsecțiune in ce mod aceste valori decid puterea unei astfel de reguli de asociere, mai ales corelația.

2.6 Importanta corelației in identificarea regulilor de asociere

Ultimul punct pe care îl vom adresa este importanța valorii corelației (și celorlalte două noțiuni) în identificarea unei reguli de asociere. Intuitiv, ar părea suficient să folosim doar valoarea suportului și confidenței pentru a identifica o regulă de asociere: până la urmă, dacă există un număr semnificativ de tranzacții în care obiectele din seturile A și B sunt cumpărate des împreună, iar dacă un procent semnificativ din tranzacțiile care conțin obiectele din setul A conțin de asemenea și obiectele din setul B, ar părea natural că exista un fel de asociere între A și B, și anume $A \Rightarrow B$, sau că dacă un client cumpăra elementele din setul A, va cumpăra de asemenea elementele din setul B. Însa, acest lucru nu este mereu adevărat, iar aici vom detalia in ce măsura valoarea corelației influențează relația dintre cele doua seturi de obiecte.

Din formula corelației, distingem următoarele două situații speciale:

- A și B sunt independente, caz în care $P(A \cap B) = P(A) \cdot P(B)$;
- A și B sunt mutual exclusive, caz în care $P(A \cap B) = 0$.

Echivalentul primei situații ar fi că probabilitatea de a cumpăra elementele din mulțimea A nu este deloc influențata de probabilitatea de a cumpăra elementele din mulțimea B, însemnând că nu există nicio asociere între cele doua seturi de elemente. În cazul acesta, se observă că valoarea corelației este chiar 1, direct din definiție. În cea de-a doua situație, valoarea corelației este 0, iar A și B fiind mutual exclusive, înseamna că dacă elementele din A sunt cumpărate, cele din B nu vor fi cumpărate.

Aceste două situații speciale ilustrează intuiția din spatele corelației. În momentul în care valoarea ei este 1, nu putem identifica o anumita relație. Pe măsura ce crește mai mult decât 1, relația $A \Rightarrow B$ este tot mai puternică, iar pe măsura ce se apropie de 0, relația $A \Rightarrow \neg B$ este tot mai puternică.

Urmează sa prezentăm o abordare matematică care demonstrează în general semnificația corelației în stabilirea unei reguli de asociere. Pentru început, vom presupune că suportul lui A si $A \cap B$ sunt fixate; cu alte cuvinte, $P(A)$ si $P(A \cap B)$ au valori fixe, iar $P(B)$ este variabil. Putem exprima următoarele:

$$P(B \setminus A) = P(B) - P(A \cap B) = P(B) - c_1,$$
$$\text{corelație}(A, B) = \frac{P(A \cap B)}{P(A) \cdot P(B)} = c_2 \cdot \frac{1}{P(B)},$$

unde c_1 și c_2 sunt constante. Probabilitatea ca o tranzacție să conțină doar elemente din setul B și nu A variază liniar cu probabilitatea unei tranzacții sa conțină toate elementele din B. În schimb, valoarea corelației este invers proporțională cu $P(B)$. Mai precis, putem exprima corelația în felul următor:

$$\text{corelație}(A, B) = c_2 \cdot \frac{1}{P(B \setminus A) + c_1}$$

Deci, în cazul in care știm si suportul si confidența, putem observa cum corelația este invers proporțională cu suportul mulțimii $B \cap A$. Cu cât valoarea corelației este mai mică, cu atâta elementele care fac parte din setul B sunt mai probabil a fi cumpărate de către clienți independent de existența elementelor din setul A in tranzacție, sau chiar în lipsa elementelor din setul A.

Un alt unghi din care putem privi situația ar fi să considerăm că $P(A)$ și $P(B)$ sunt fixe, variabila fiind aici $P(A \cap B)$. În situația asta, corelația ar avea o expresie de tipul

$$\text{corelație}(A, B) = \frac{P(A \cap B)}{c},$$

unde c este o constanta. Deci, valoarea ei este direct proporțională cu suportul mulțimii $A \cap B$, ca și confidența. Știind exact procentul de tranzacții care conțin obiectele din A si obiectele din B, șansele sa existe o asociere pozitivă $A \Rightarrow B$ sunt tot mai mari cu cat exista mai multe tranzacții care conțin atât obiectele din A, cat si obiectele din B, iar asocierea inversă $A \Rightarrow \neg B$ este mai probabila cu cât obiectele din cele două seturi sunt cumpărate tot mai rar împreună.

În final, mai considerăm și $A \subset B$. In acest caz, suportul este $P(B)$, confidența este $\frac{P(B)}{P(A)}$ iar corelația este $\frac{1}{P(A)}$. Ar putea părea surprinzător faptul că corelația este strict determinata de $P(A)$ și nu $P(B)$, dar interpretând ipoteza $B \subset A$, ar însemna că elementele cumpărate din setul B sunt mereu cumpărate doar dacă și elementele din setul A au fost cumpărate. Fixând $P(B)$, cu cât elementele din A sunt cumpărate mai des, cu atâta confidența scade și corelația se apropie de 1, sugerând posibilitatea mai mica a existentei unei relații. Asta pare intuitiv corect: daca niște obiecte sunt aproape tot timpul cumpărate, este dificil sa stabilim dacă datorita existenței lor în coșul de cumpărături se află și alte produse sau nu.

Iar in același caz, daca fixăm $P(A)$, corelația rămâne fixă peste 1, însă confidența scade pe măsura ce $P(B)$ scade. Acest caz ilustrează perfect motivul pentru care valorile tuturor celor 3 formule sunt importante in identificarea regulilor de asociere, nefiind suficient să avem o corelație pozitivă. Din nou, semnificația se poate observa intuitiv: daca produsele din B au fost mereu cumpărate când cele din A sunt cumpărate, dar produsele din B au fost totuși cumpărate foarte rar, nu putem stabili foarte clar o regula și ar putea fi o pură coincidență.

3 Exemplu practic

În continuare, vom prezenta un exemplu simplu în care vom aplica noțiunile definite mai sus pentru a încerca identificarea unei reguli de asociere. Pentru simplitate, se presupune existența unui magazin fictiv care include printre produsele puse la vânzare un telefon mobil și o husă de telefon, și vom încerca să identificăm o regulă de asociere $A \Rightarrow B$ între setul A conținând telefonul mobil, și setul B conținând husa de telefon.

Să zicem că următoarele date au fost colectate:

- Numărul total de tranzacții efectuate a fost 10,000;
- Aproximativ 1,000 de tranzacții includ telefonul mobil în setul de obiecte cumpărate;
- Aproximativ 750 de tranzacții includ husa de telefon în setul de obiecte cumpărate;
- Aproximativ 600 de tranzacții includ și telefonul mobil, și husa în setul de obiecte cumpărate.

Vom continua prin a calcula suportul, confidența și corelația pentru seturile A și B:

$$\text{suport}(A, B) = \frac{600}{10,000} = 0.06;$$
$$\text{confidența}(A, B) = \frac{600}{1,000} = 0.6;$$
$$\text{corelația}(A, B) = \frac{0.06}{0.1 \cdot 0.075} = 8;$$

Suportul calculat pentru setul format din telefonul mobil și husă nu are o valoare mare, ceea ce înseamnă ca produsele nu sunt des achiziționate împreună. În schimb, valoarea confidenței ne spune că husa este deseori cumpărata împreună cu telefonul, iar valoarea corelației confirmă că într-adevăr avem o corelație pozitivă între elementele din setul A și B. Astfel, putem deduce că există o regulă de asociere între telefonul mobil și husa pentru telefon. Un motiv care ar putea explica suportul mic ar fi posibila diversitate de articole găsite în magazinul respectiv.

4 Concluzie

În acest scurt articol, am analizat identificarea anumitor reguli de asociere intre diferite seturi de obiecte, pornind de la totalul tranzacțiilor de care dispunem. In vreme ce definiția unei reguli de asociere este oarecum vaga si nu foarte riguroasa, valorile pe care le obținem din calcularea suportului, confidentei si corelației pot totuși să ne ghideze către obținerea anumitor relații intre produse. In particular, ele ne oferă o direcție spre care ne putem îndrepta pentru a putea răspunde întrebării de tipul "daca elementele din setul A sunt achiziționate, cât de siguri putem fi că elementele din setul B vor fi de asemenea achiziționate".

Bibliografie

[1] Frans Coenen. Data mining: Past, present and future. *Knowledge Eng. Review*, 26:25–29, 03 2011.

[2] David Olson. *Market Basket Analysis*, pages 29–41. 12 2017.

[3] Lynsey McColl. Market Basket Analysis: Understanding Customer Behaviour. https://select-statistics.co.uk/blog/market-basket-analysis-understanding-customer-behaviour/. Online; accessed Nov 12 2022.

[4] Glovo. Algorithms. https://about.glovoapp.com/algorithms/. Online; accessed Nov 14 2022.

[5] Selvam Jesiah and Uma Kalakada. Employability recommender system for learning environments in management education using association rules. 12 2013.

[6] Tutorial and Example. Association Rule in Data Mining. https://www.tutorialandexample.com/association-rule-in-data-mining. Online; accessed Nov 13 2022.