FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

**Ano Letivo de 2015/2016**

Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão

*Unidade Curricular de Extração de Conhecimento de Dados I*

# RELATÓRIO DO TRABALHO PRÁTICO II

## *"Regras de Associação"*

**Docentes:** Professor Doutor João Gama
Professor Doutor Pavel Bradzil
**Discentes:** Rui Pedro Machado, 201300292
Hélder Filipe Russa, 201508409

# 1. INTRODUÇÃO

No âmbito da disciplina de Extração de Conhecimento de Dados I, do Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão da Faculdade de Economia da Universidade do Porto 2015/2016, foi desenvolvido este trabalho, denominado "Regras de Associação", pretendendo-se desenvolver um modelo baseado na extração de conjuntos de itens frequentes.

Este tópico de descoberta de conhecimento em base de dados é comummente utilizado no mercado para diversos fins, nomeadamente, análise de carrinho de compras, análise de visitas a um determinado sítio web, ou até, compreensão do padrão de consumo de um determinado consumidor (ex. "*Caso um cliente compre bananas, qual a probabilidade de comprar também iogurtes?*"). Este tipo de análises são extremamente valiosas para, por exemplo, estruturar campanhas de marketing, bem como, organizar a distribuição de produtos dentro de uma loja física, entre outros. (Gama, Carvalho, Facelli, Lorena, & Oliveira, 2012)

Assim, o problema especifico a abordar neste trabalho, diz respeito à análise de carrinho de compras (*market basket problem*) que através de regras de associação visa fornecer uma abordagem para análise de dados respeitantes a itens comprados por consumidores em diferentes dias por fim a permitir de uma forma cientifica, baseada em probabilidades, obter uma compreensão sobre os itens que são mais frequentemente adquiridos em conjunto e sobre a variabilidade das sus frequências.

O objetivo principal é de explorar e trabalhar um conjunto de dados previamente disponibilizados de 5000 carrinhos de compras, dispersos por 28 dias de atividade de um determinado supermercado, com vista a (1) estudar a variação do número de grupos de itens frequentes, variando o seu suporte; (2) estudar a variação do número de regras de associação, fixando o seu suporte; (3) selecionar a regra com maior lift; (4) e por último analisar a variação de *itemsets* do início do mês para o fim do mês.

# 2. REGRAS DE ASSOCIAÇÃO

## 2.1 Suporte e Confiança

As regras de associação, tal como anteriormente descrito, mostram fortes associações existentes entre itens em estudo (Como por exemplo produtos) e podem ser obtidas através da análise de frequência de ocorrência dos mesmos e consequentemente na probabilidade de serem observados num dado conjunto de ocorrências. A descoberta de regras de associação é então baseada na retirada de informação de *item sets* frequentes (Sendo a frequência dada pelo número de transações que cada item tem). Estas regras podem ser interpretadas como uma instrução "Se A Então B" e podem ser escritas no formato A ⇨B.

Para uma correta construção de regras de decisão existe um conjunto de conceitos que se deve compreender, começando pelas duas medidas de interesse utilizadas para a compreensão quer da relevância (ou utilidade) quer da confiança (ou eficácia), das regras, sendo respetivamente o valor do **suporte** e o valor da **confiança**. Estas medidas são importantes dado que uma dada regra de associação só tem interesse se respeitar um valor mínimo de suporte e de confiança.

O suporte é a medida mais simples e pode ser explicada como sendo a probabilidade de encontrar o conjunto {A,B} num determinado carrinho de compras. Matematicamente pode ser formulado como sendo: **SUPORTE DA REGRA (A ⇨ B) = P({A,B})**

Para o mesmo conjunto de itens {A,B}, a confiança pode ser explicada como sendo a probabilidade condicionada de B, sabendo A. Matematicamente pode ser formulado como sendo: **CONFIANÇA DA REGRA (A ⇨ B) = P(B | A),** podendo ainda ser formulada com base no suporte: **CONFIANÇA DA REGRA (A ⇨ B) = SUPORTE({A,B}/ SUPORTE(A)**

## 2.2 Algoritmo Apriori

A utilização do **algoritmo Apriori**, é uma técnica comummente utilizada e trata-se de um algoritmo que permite não só extrair conjuntos de itens frequentes dado um determinado suporte mas também gerar regras de associação baseadas nos resultados obtidos no primeiro passo e para uma determinada confiança definida. Tornou-se interessante por recorrer ao princípio de que qualquer subconjunto de *itemsets* frequentes deve ser também um *itemset* frequente. Este algoritmo apresenta algumas limitações, nomeadamente o facto de percorrer diversas vezes as bases de dados para o cálculo do suporte de cada *itemset* frequente candidato, desta forma existe hoje em dia outros algoritmos mais eficientes, nomeadamente o ***FP-growth*** que, no entanto, não foi utilizado neste trabalho, ficando simplesmente essa ressalva.

Após terminar a sua execução, o algoritmo retorna o conjunto de itens frequentes com um suporte superior ao estipulado pelo analista e inicia-se a segunda fase do mesmo - extrair regras de associação. Esta fase é crítica e depende bastante da qualidade do passo anterior dado que caso o conjunto de *itemsets* frequentes seja muito grande, o número de regras de associação geradas pode ser igualmente grande, acabando por gerar regras sem interesse, mesmo respeitando o valor de suporte e confiança definidos.

## 2.3 Coeficiente de lift

Quando o suporte e a confiança não se revelam suficientes para filtrar regras de associação sem interesse existem algumas heurísticas que fornecem uma resposta mais assertiva tal como o **coeficiente de *LIFT*** de uma regra de associação. Esta métrica reflete a noção estatística de independência entre duas variáveis aleatórias. Um valor de *lift* superior a 1 indica que A tem um efeito positivo sobre a ocorrência de B. Se o valor for menor do que 1 indica que a ocorrência de A tem um efeito negativo sobre a ocorrência de B. Um valor próximo de 1 indica que A e B aparecem frequentemente juntos. Matematicamente o coeficiente de *lift* pode ser formulado como sendo: **LIFT(A ⇨ B) = CONFIANÇA DA REGRA (A ⇨ B) / SUPORTE(B)**, podendo ainda ser formulado com base nos valores de suporte: **LIFT(A ⇨ B) = SUPORTE(A∪B) / (SUPORTE(A) * SUPORTE(B))**.

### 2.4 *Itemsets* frequentes *closed* e *maximal*

Por último existe um caraterística de *itemsets* frequentes que permite otimizar as regras de associação extraídas através de uma técnica de sumarização de *itemsets*. O conjunto de todos os *itemsets* frequentes é através desta técnica, representado num conjunto com menos elementos por via da eliminação de todos os *itemsets* que são subconjuntos de outros *itemsets* frequentes. Esta técnica baseada na propriedade da monotonia do valor de suporte permite-nos filtrar quais os itemsets frequentes que queremos utilizar para a geração de regras de associação: ***closed***, utilizando aqueles que não tenham nenhum superconjunto frequente com a mesma frequência dele próprio; ***maximal*** utilizando aqueles em que todos os superconjuntos próprios não são frequentes (A é um superconjunto próprio de B se e só se, a cardinalidade de A é maior qua a de B, e todos os elementos de B pertencem a A).

## 3. EXPERIÊNCIAS REALIZADAS

Para a elaboração das experiências ilustradas nos seguintes pontos, foi utilizada a plataforma de desenvolvimento analítico KNIME. O fluxo base utilizado para a obtenção dos resultados demonstrados pode ser visualizado na figura 1.

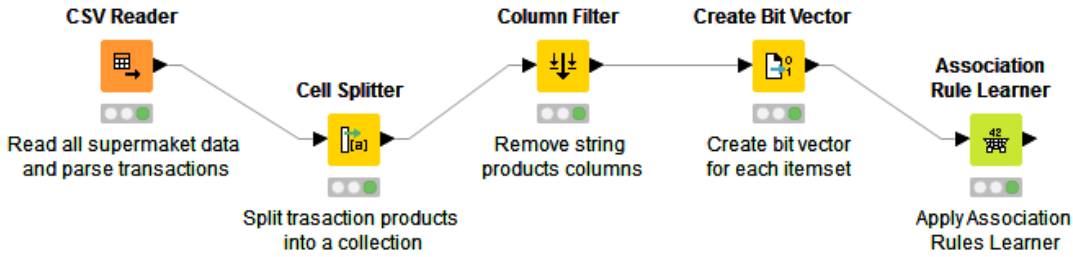


Figura 1 - Fluxo do KNIME para aplicação de algoritmo de regras de associação

Este objeto de KNIME *"Association Rule Learner"* utiliza o algoritmo APRIORI, previamente descrito e foi implementado por *Christian Borgelt*.

## 3.1 Estudo da variação de *itemsets* frequentes com variação de suporte

Na tabela 1, é demonstrada a variação do número de *itemsets* frequentes *closed* e *maximal*, face à variação do suporte mínimo, recorrendo ao *software* KNIME. O intervalo de valores de suporte mínimo foi definido através da análise do valor máximo para o suporte mínimo, com obtenção de resultados, o qual foi identificado como sendo de 0,11. Assim sendo especificamos o nosso intervalo de análise de 0,007 a 0,12, por forma a totalizar 15 observações, esperando 0 *itemsets* para o valor 0,12.

| Iteração | Suporte Mínimo | #*Itemsets* Frequentes closed | #*Itemsets* Frequentes maximal |
|---|---|---|---|
| **1** | 0,007 | 116 | 32 |
| **2** | 0,008 | 110 | 26 |
| **3** | 0,009 | 110 | 26 |
| **4** | 0,01 | 110 | 26 |
| **5** | 0,02 | 110 | 26 |
| **6** | 0,03 | 82 | 35 |
| **7** | 0,04 | 57 | 38 |
| **8** | 0,05 | 39 | 37 |
| **9** | 0,06 | 38 | 38 |
| **10** | 0,07 | 31 | 31 |
| **11** | 0,08 | 17 | 17 |
| **12** | 0,09 | 7 | 7 |
| **13** | 0,10 | 2 | 2 |
| **14** | 0,11 | 1 | 1 |
| **15** | 0,12 | 0 | 0 |

Tabela 1 Variação do número de itemsets frequentes closed e maximal, face à variação do suporte mínimo

É possível verificar pela análise da tabela 1, que os valores quando o suporte mínimo é mais reduzido, são mais díspares, sendo que quando o valor do suporte aumenta tendem a estabilizar e a não haver diferença entre ambos. Um outro ponto curioso é o facto de o número de *itemsets closed* ser sempre superior ou igual ao número de *itemsets maximal*, o que se deve ao facto de que o conjunto dos *itemsets* do tipo *closed* contém ou é igual ao conjunto dos *itemsets* do tipo *maximal.*

## 3.2 Estudo da variação do número de regras de associação com variação de confiança e suporte mínimo fixo

Para o estudo em causa e tal como solicitado, o número de regras de decisão geradas é ilustrado, variando a confiança mínima entre 0,1 e 1 enquanto que o suporte mínimo é fixado em 0,03, o que permite ter valores diferentes para *itemsets* frequentes *closed* e *maximal.*

Na tabela 2 é possível visualizar o resultado de tal análise.

| ***Valor de suporte mínimo: 0,03*** | | | |
|---|---|---|---|
| **Iteração** | Confiança Mínima | *#Regras de Decisão (Closed)* | *#Regras de Decisão (Maximal)* |
| **1** | 0,1 | 70 | 70 |
| **2** | 0,2 | 70 | 70 |
| **3** | 0,3 | 70 | 70 |
| **4** | 0,4 | 60 | 60 |
| **5** | 0,5 | 29 | 29 |
| **6** | 0,6 | 16 | 16 |
| **7** | 0,7 | 16 | 16 |
| **8** | 0,8 | 13 | 13 |
| **9** | 0,9 | 12 | 12 |
| **10** | 1,0 | 4 | 4 |

**Tabela 2 Variação do número de regras de associação com variação de confiança e suporte mínimo fixo**

Tal como é possível verificar pela tabela acima, o número de regras de decisão não varia entre *itemsets* frequentes *closed* e *maximal* para este conjunto de dados, pelo que nenhuma análise de interesse pode ser retirada daqui. A análise interessante é verificar que consoante a confiança aumenta o número de regras diminui e isto está relacionado com a própria definição de confiança. Se esta significa o número de consumidores que adquiriu o produto B sabendo que adquiriu A, aumentar a confiança significa aumentar o número esperado de ocorrências deste tipo e sendo o *dataset* estático, vai fazer reduzir o interesse da regra de decisão.

## 3.3 Estudo da regra com maior *lift*

Para a procura da regra com maior Lift, foi definido um suporte mínimo de 0,005 e uma confiança mínima de igualmente 0,005, o que por ter valores mínimos baixos vai provocar a geração de diversas regras de decisão, permitindo assim escolher aquela com maior *Lift*. Desta forma foi identificado que o seguinte "23 ⇦ [24, 40, 41, 43]", é o que obtém o valor observado mais elevado com 15.625 de *lift* para um suporte de 0.0212 e uma confiança de 1.0.

Frequent itemsets/Association rules - 0:5 - Association Rule Learner (Apply Association)

File

Table "default" - Rows: 443 | Spec - Columns: 6 | Properties | Flow Variables

| Row ID | D Support | D Confide... | D Lift | S Conseq... | S implies | (...) Items |
|---|---|---|---|---|---|---|
| rule303 | 0.021 | 1 | 15.625 | 23 | <--- | [24,40,41,...] |
| rule302 | 0.021 | 1 | 15.576 | 24 | <--- | [23,40,41,...] |
| rule336 | 0.026 | 1 | 15.576 | 24 | <--- | [23,40,41] |
| rule337 | 0.026 | 0.992 | 15.507 | 23 | <--- | [24,40,41] |
| rule304 | 0.021 | 1 | 15.432 | 40 | <--- | [23,24,41,...] |
| rule338 | 0.026 | 1 | 15.432 | 40 | <--- | [23,24,41] |
| rule340 | 0.026 | 0.95 | 14.792 | 24 | <--- | [23,40] |
| rule305 | 0.021 | 1 | 14.749 | 41 | <--- | [23,24,40,...] |
| rule341 | 0.026 | 0.943 | 14.732 | 23 | <--- | [24,40] |

**Figura 2 Lifts observados**

Pela análise do valor de *lift* obtido, temos que, dado que este é superior a 1 os itens que constituem os *itemsets* da regra aparecem em conjunto mais vezes do que o esperado, ou seja a compra do item 23 tem um efeito positivo (provoca) na compra do *itemset* [40,43,24,41]. Caso quiséssemos aumentar as vendas do *itemset* [40,43,24,41], poderíamos por exemplo definir com esta informação, uma campanha de marketing para aumentar as vendas do item 23, o que poderia provocar o aumento de compras do *itemset* em causa.

## 3.4 Estudo da variação de *itemsets* frequentes em diferentes períodos: início do mês versus fim do mês

Para o estudo em causa, foram selecionados os primeiros 5 (Dia 1 a Dia 5) e os últimos 5 dias (Dia 24 a 28) do mês para refletir respetivamente o início e o fim do mês, dado que nada foi especificado a este nível e posteriormente para cada um destes subconjuntos do conjunto de dados iniciais, aplicada uma análise de *itemsets* frequentes gerados com o algoritmo Apriori, para um suporte mínimo de referência de 0,03.

Pelos valores observados no KNIME foi construída a tabela 3 onde constata o top 5 dos *itemsets* frequentes, comparando o início com o fim do mês e igualmente os valores observados para os *itemsets* frequentes *closed* e *maximal*.

| ***Valor de suporte mínimo: 0,03*** | | | | |
|---|---|---|---|---|
| **Rank** | ***Itemsets frequentes maximal*** | | ***Itemsets frequentes closed*** | |
| | Início do Mês | Fim do Mês | Início do Mês | Fim do Mês |
| **1** | [28] Sup:0,106 | [7] Sup:0,107 | [28] Sup:0,106 | [35] Sup:0,107 |
| **2** | [7] Sup:0,102 | [45] Sup:0,107 | [42] Sup:0,103 | [7] Sup:0,107 |
| **3** | [45] Sup:0,101 | [17] Sup:0,093 | [7] Sup:0,102 | [45] Sup:0,107 |
| **4** | [1] Sup:0,097 | [37] Sup:0,089 | [45] Sup:0,101 | [18] Sup:0,104 |
| **5** | [22] Sup:0,094 | [27] Sup:0,088 | [1] Sup:0,097 | [17] Sup:0,093 |

**Tabela 3 Análise de variação de itemsets início do mês versus fim do mês**

Pela análise da tabela três podemos observar que existe uma variabilidade no conjunto de produtos mais adquiridos no início do mês quando comparados com o fim do mês. Por exemplo olhando para os *itemsets* frequentes *maximal*, constatamos que o produto 28 é o mais vendido no início do mês enquanto que no final nem sequer se encontra no top 5. Não temos no entanto dados suficientes para compreender o porquê desta variação (Serão produtos de uma categoria de luxo? De uma categoria afetada pela altura do mês? etc.).

Podemos por último constatar que existe variabilidade dependendo de se analisamos os produtos recorrendo ao filtro *closed* ou *maximal*, o que pode ser explicado pela tabela 1 onde se demostra o número de *itemsets* gerados. Para um suporte Mínimo de 0,3, o número de *itemsets closed* gerados é superior ao número de *itemsets* gerados com *maximal*, o que aumenta o espectro de itemsets em análise e consequentemente faz variar este *ranking* de *top* 5.

# 4. CONCLUSÃO

A utilização de regras de associação é bastante útil para muitos processos de tomada de decisão tais como a organização de produtos em loja, venda e *marketing* cruzado de produtos, assim como descobrir grupos de produtos que são frequentemente comprados em conjunto e com base nos mesmos inferir os que serão comprados sabendo que foram comprados outros produtos.

Neste exemplo foi-nos possível constatar que apesar da sua simples implementação em *softwares* especializados, como o KNIME, o valor obtido com a sua aplicação é imensurável e pode ajudar as organizações a muito rapidamente redesenharem as suas campanhas de marketing ou até mesmo reorganizar o seu catálogo de produtos e potenciar as suas vendas.

Constatamos no entanto que o algoritmo Apriori utilizado é de elevada complexidade computacional e num momento em que as organizações aderem a conceitos como *Internet of Things* (IoT), ficamos curiosos quanto à performance do mesmo quando inserido numa arquitetura de sistemas de informação baseada em *Big Data*, com um volume e variedade de dados, abrindo-nos assim a oportunidade de explorar outros tais como FP-growth para estes propósitos.

# 5. BIBLIOGRAFIA

Gama, J., Carvalho, A., Facelli, K., Lorena, A., & Oliveira, M. (2012). *Extração de Conhecimento de Dados*. (E. Silabo, Ed.) (1st ed.). Lisboa: Silabo, Edições.

Regras de Associação, João Gama, MADSAD (FEP), 2015-2016;