ANÁLISE ASSOCIATIVA EM MINERAÇÃO DE DADOS

Bolsista: Alexandre Esteves Almeida – *almeida.xan@gmail.com*
Orientador: Prof. Dr. Emanuel Pimentel Barbosa
Unidade: Departamento de Estatística – IMECC, UNICAMP
Bolsa/Agência: PIBIC/CNPq
Palavras-chave: Mineração de Dados – Algoritmo Apriori – Análise Associativa

Introdução

Neste projeto, estudamos os conceitos, fundamentos matemáticos, algoritmos e realizamos implementações práticas do método de *análise associativa*, tópico muito útil no campo de mineração de dados, com objetivo de descobrir *relações interessantes* (não aparentes e que tragam informação nova e útil) em (grandes) conjuntos de dados. Tais relações interessantes são apresentadas, fundamentalmente, no formato de *regras de associação*.

Um dos desafios que este projeto procurou enfrentar, foi o fato da literatura da área frequentemente ocultar formalizações matemáticas mais adequadas (particularmente em relação à teoria dos conjuntos e funções) assim como o uso de um software apropriado que pudesse lidar, de forma eficiente, com os conjuntos de dados utilizados nas aplicações.

Metodologia

O desenvolvimento do projeto foi feito, essencialmente, em torno de um dos algoritmos da análise associativa: o *algoritmo Apriori*, o qual baseia-se em inúmeras *podas* do enorme conjunto de itens e regras possíveis, de modo que não seja necessário o cálculo de *medidas de interesse* utilizadas para definir se uma associação extraída do conjunto de dados é, ou não, interessante. Tais medidas são denominadas *medidas objetivas de interesse*. Como exemplo dessas medidas, temos o *suporte* e a *confiança*, que são, basicamente, frequências relativas.

As regras de associação em questão são apresentadas no formato "X implica Y", onde X e Y são dois conjuntos disjuntos de itens/atributos dos dados. Tais regras associativas são frequentemente representadas pela notação

$$X \Longrightarrow Y$$

e, ao final, são exibidas juntamente com as suas respectivas medidas objetivas de interesse. Outras medidas complementares de interesse podem ser calculadas para o auxílio na detecção de regras que supostamente sejam falsas ou que não apresentem informação útil. Como é o caso da medida *lift*.

No software R, utilizamos o pacote *arules* para a geração das regras, assim como o pacote complementar *arulesViz*, que dispões de alternativas gráficas para visualização e seleção de regras interessantes.

Resultados

Dois conjuntos de dados foram analisados: *Titanic*, com quatro atributos de 2201 passageiros da embarcação e *Groceries*, com 9835 transações de até 169 itens, de uma mercearia.

Abaixo estão dois gráficos do tipo grafos (um para cada conjunto de dados), que sumarizam as principais regras de associação obtidas, apresentadas juntamente com a informação de duas medidas objetivas de interesse:

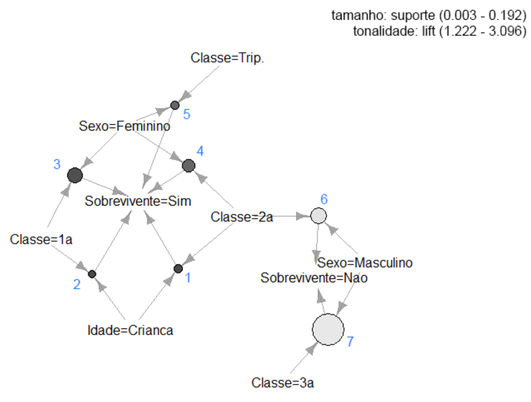

Gráfico do tipo de grafos para *todas* as 7 regras geradas pelo conjunto de dados *Titanic.*

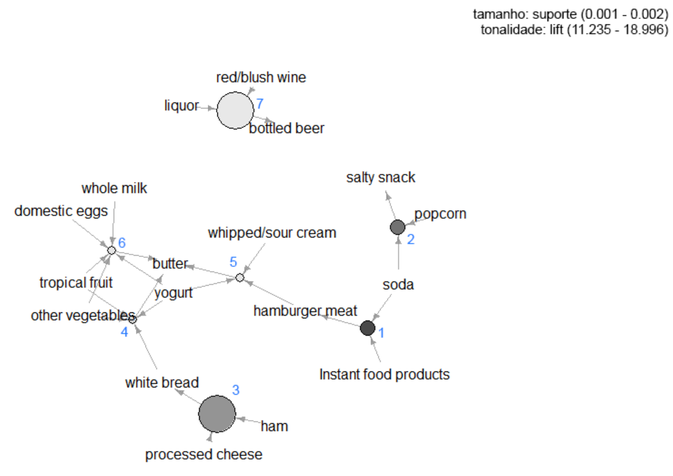

Gráfico do tipo de grafos para as 7 regras com maiores valores de *lift* do conjunto de dados *Groceries*.

Conclusões

- A formalização matemática adequada é de grande utilidade para o entendimento da maneira que as medidas de interesse são calculadas, assim como o funcionamento do algoritmo Apriori e seus procedimentos.
- O software R apresentou um bom desempenho com os pacotes arules e arulesViz para aplicação do método de análise associativa e para lidar com conjuntos de dados médios, possivelmente até grandes.
- Algoritmo Apriori tem grande potencial e aplicabilidade até em problemas que fogem de áreas comumente aplicadas, como foi perceptível com o conjunto de dados Titanic.
- Produção de um material didático que une as facilidades de leitura e entendimento de uma das referências, com a formalização matemática da outra, e aplicações em um software apropriado, não presente em nenhuma das referências bibliográficas.

Referências Bibliográficas

- Djeraba, C. & Simovici, D. A., 2008. *Mathematical Tools for Data Mining.* Verlag: Springer.
- Tan, P.-N., Steinbach, M. & Kumar, V., 2005. *Introduction to Data Mining.* Addison-Wesley.