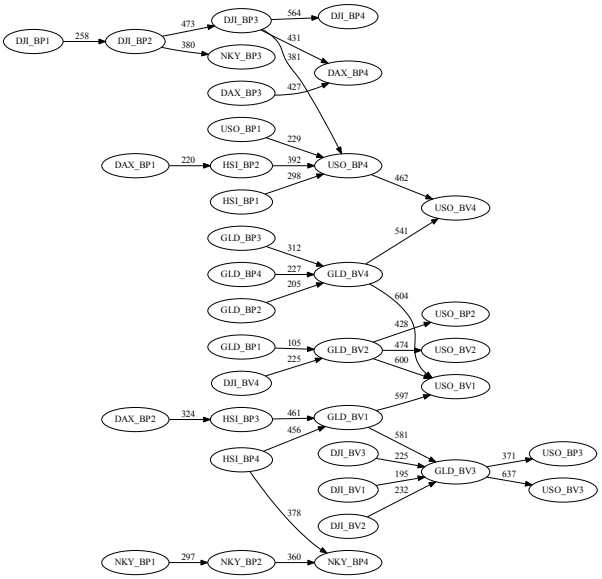
3. UMA NOVA TAXONOMIA EM DATA SCIENCE

por Luís Cavique



→

Diagrama #02 . Secção IV

Macro-padrão de uma sequência de compra/venda de produtos financeiros.

Em *Data Mining,* mais recentemente denominado *Data Science* é vulgar classificar os algoritmos em supervisionados e não supervisionados. Neste artigo iremos discutir uma taxonomia baseada nos tipos de padrões que são extraídos dos dados.

Embora as questões colocadas em base de dados sejam semelhantes às questões de *Data Science*, nas bases de dados é apresentado um padrão (consulta ou *query*) e são devolvidos dados e em *Data Science* para um conjunto de dados pretende-se extrair padrões.

Data Science é um processo computacional para descobrir "padrões" envolvendo métodos que combinam a estatística com *machine learning* e tecnologias de base de dados. Os padrões que se encontra na natureza ou manufaturados têm uma configuração específica e caracterizam-se por uma regularidade, repetição de partes e acumulação de elementos. Por exemplo, uma duna, criada pela ação do vento, é composta por várias camadas de areia e tem uma configuração reconhecível.

Neste artigo vamos diferenciar o micro-padrões dos macro-padrões. Os micro-padrões correspondem pequenas percentagens de dados; por exemplo nas regras associativas, uma medida de suporte que apresenta valores suporte ≥ 5%, sendo escolhidas as regras com maior *confidence* (ou probabilidade condicionada). Por outro lado, os macro-padrões envolvem uma grande percentagem, ou a totalidade, dos dados; por exemplo na modelação com regressão são utilizados todos os dados disponíveis. Os micro-padrões caracterizam-se por elevada confiança (*confidence*) e os Macro-padrões por elevado suporte.

Existem outros exemplos de micro-padrões: nos problemas de *sequence/episode mining* com suporte ≥ 1%; no problema de classificação, ao utilizar *decision trees*, cada ramo da árvore corresponde a uma pequena percentagem dos dados; ainda no problema de classificação ao utilizar o *k-nearest neighbor* a comparação que é efetuada é com o reduzido número de *k* elementos. Quanto aos macro-padrões em problemas como regressão, teste de hipóteses, *clustering* ou redução de atributos, todos os dados são tidos em consideração.

A origem desta dicotomia na análise de dados remonta quando do aparecimento do *Data Mining*, hoje uma área madura, mas que tinha ini-

cialmente uma conotação negativa com os nomes de "data snooping" (bisbilhotando) e "data fishing", onde o objetivo era explorar e/ou espiar subconjuntos de dados.

Drew Conway [2010], na sua definição de *Data Science*, inclui para além da matemática as tecnologias e os *hacking skills* (capacidade de decifração) que caracterizam a descoberta de micro-padrões.

Leo Breiman em 2001 [Breiman 2001] já tinha referido as duas culturas na modelação de dados. A cultura dos micro-padrões corresponde à procura de pequenas percentagem de dados com utilidade ou interesse, tendo como métricas o suporte e a confiança das regras associativas, tem origem no *machine learning* e nos *hacking skills* e tem um grande apoio dos grandes decisores dos USA em projetos de mais de 1.000.000 dólares [O'Connor 2008]. A cultura dos macro-padrões utiliza a totalidade dos dados, tem origem na matemática e na estatística e conta com projetos vinte vezes menores que os anteriores.

NOVOS PADRÕES PARA PROBLEMAS VELHOS

Nesta secção acerca de novos padrões para problemas antigos discutiremos essencialmente o problema do Market Basket Analysis (MBA) e o problema de sequence/episode mining.

As regras associativas ficaram célebres, ao encontrar um padrão nos supermercados onde jovens casais com filhos às sextas-feiras e sábado, ao comprar fraldas também compravam cerveja. A regra fraldas => cervejas, tendo um suporte de algumas centésimas, estava associada a uma confiança (probabilidade condicionada) relevante.

O algoritmo Apriori [Agrawal, Srikant 1994] foi o primeiro algoritmo para o MBA. O Apriori gera para um pequeno número de produtos um enorme conjunto de regras associativas, i.e. micro-padrões, que devem ser criteriosamente escolhidas pelo utilizador final. O trabalho de Cavique [2007a] com o algoritmo Similis, resolve o MBA para um elevado número de produtos e evita a escolha entre milhares de micro-padrões, devolvendo padrões baseados na acumulação, i.e. macro-padrões que representam a totalidade dos dados. O algoritmo Similis está dividido em duas partes, na primeira transforma o problema num grafo ponderado e na segunda encontra subgrafos completos, i.e. cliques, que correspondem aos cabazes de compras mais comprados.

Em resumo, para o MBA, o algoritmo Apriori tem como medidas o suporte e a confiança das regras associativas, que correspondem a micro-padrões, enquanto que o algoritmo Similis procura encontrar macro-padrões, i.e. padrões que resultam da acumulação das compras de vários clientes.

Para o segundo problema de *sequence mining* o algoritmo AprioriAll [Srikant, Agrawal 1996], para além de ter uma elevada complexidade temporal, encontra milhares de micro-padrões de difícil seleção e que requerem um trabalho exaustivo na atribuição de utilidade ou interesse.

No problema de *sequence mining*, tratado por Cavique [2007b], é apresentado o algoritmo Ramex que gera árvores e poli-árvores que envolvem todos os elementos numa perspetiva de macro-padrões.

Uma abordagem recente na criação de novos padrões em problemas antigos de *sequence mining* é o denominado *Process Mining* [Aalst 2011]. Segundo o autor o *Process Mining* cria pontes entre o *Data Mining* e o *Business Process Modeling,* e considera a acumulação de eventos tendo como objetivo melhorar a representação dos dados e criar equilíbrios entre a simplicidade e a exatidão dos resultados.

Na abordagem que utiliza macro-padrões existem duas fases distintas. A primeira acumula os dados em bruto numa estrutura de dados condensados, num grafo [Cavique 2007b], numa cadeia de Markov [Borges, Levene 2007] ou numa rede de Petri [Aalst 2011]. Na segunda fase é possível procurar os macro-padrões na estrutura de dados condensados.

Para os programadores e investigadores em novos algoritmos um dos grandes desafios é trazido pelo *Big Data*. É urgente a redução de complexidade temporal de quase todos os algoritmos, desde o cálculo da variância em estatística até ao mais complexo problema de sequence mining.

Os algoritmos que utilizam estruturas condensadas, e que estão na origem dos macro-padrões, são, sem dúvida, os algoritmos eleitos para lidar com *Big Data.*

CONCLUSÕES

Neste artigo apresentamos uma taxonomia de algoritmos baseada em micro e macro-padrões. Os referidos algoritmos que têm origem em escolas diferentes e complexidades temporais distintas. O analista de dados terá de escolher entre soluções únicas baseadas em macro-padrões ou descobrir padrões úteis entre milhares de micro-padrões.

Os macro-padrões podem indicar resultados óbvios e existir a necessidade que recorrer a micro-padrões. Por isso sugerimos uma metodologia combinada que procura inicialmente macro-padrões, à qual se segue uma pesquisa seletiva de micro-padrões. Os macro-padrões podem indicar resultados óbvios e existir a necessidade que recorrer a micro-padrões. Por isso sugerimos uma metodologia combinada que procura inicialmente macro-padrões, à qual se segue de uma pesquisa seletiva de micro-padrões.

Bibliografia

Agrawal, R., R. Srikant (1994), *Fast algorithms for mining association rules*, Proceedings of the 20th International Conference on Very Large Data Bases, 478-499

–

Borges J., Levene M. (2007), *Evaluating Variable-Length Markov Chain Models for Analysis of User Web Navigation Sessions*, IEEE Transaction Knowledge Data Engineering, vol. 19, issue (4), pp. 441-452.

–

Breiman L. (2001), *Statistical Modeling: The Two Cultures, Statistical Science*, vol. 16, issue 3, pp. 199-309

–

Cavique L. (2007a), *A Scalable Algorithm for the Market Basket Analysis, Journal of Retailing and Consumer Services, Special Issue on Data Mining in Retailing and Consumer Services*, vol.14, issue 6, pp. 400-407

–

Cavique L. (2007b), *Network Algorithm to Discover Sequential Patterns,* in Progress in Artificial Intelligence, J.Neves, M.Santos and J.Machado (Eds.), EPIA 2007, LNAI 4874, Springer-Verlag Berlin Heidelberg, pp. 406-414

–

Conway, D. (2010), *The Data Science Venn Diagram*, Creative Commons licensed, http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

–

Srikant R., Agrawal R. (1996), *Mining sequential patterns: Generalizations and performance improvements*, Proceedings 5th International Conference Extending Database Technology, EDBT, 1057, pp. 3–17

–

O'Connor, B. (2008), *Statistics vs. Machine Learning, fight!, AI and Social Science* blog, http://brenocon.com/blog/2008/12/statistics-vs-machine-learning-fight/