REVISTA DE CIÊNCIA ELEMENTAR

Ciência dos Dados

Bases de Dados *versus* Aprendizagem Automática

Luís Cavique
Universidade Aberta

CITAÇÃO
Cavique, L. (2021)
Ciências dos Dados,
Rev. Ciência Elem., V9(02):041.
doi.org/10.24927/rce2021.041

EDITOR
José Ferreira Gomes,
Universidade do Porto

EDITOR CONVIDADO
Paulo Ribeiro-Claro
Universidade do Porto

RECEBIDO EM
29 de abril de 2020

ACEITE EM
05 de julho de 2020

PUBLICADO EM
15 de junho de 2021

rce.casadasciencias.org

A Ciências dos Dados é uma disciplina emergente que tem tido nos últimos anos muita aceitação pelo público em geral. A Ciências dos Dados aparece na interceção de três grandes áreas: a informática, a estatística e as áreas de negócio. A área de informática compreende as tecnologias e sistemas de informação bem como as ciências da computação. A área de estatística inclui a análise de dados e subáreas relacionadas da matemática. Por área de negócio designa-se o sector onde o sistema é desenvolvido, por exemplo: produção, *marketing*, finanças, turismo, educação, etc.

Na FIGURA 1 apresenta as interceções das grandes áreas, em pares e no conjunto das três áreas. Da combinação das três áreas encontramos a emergente Ciência dos Dados.

A interceção da informática com a estatística encontramos a aprendizagem automática (*machine learning*) que é uma subárea da inteligência artificial. O nome *machine learning* aparece pela primeira vez nos anos de 1980, tendo sido substituído por termos como *knowledge discovery* e *data mining*, e reaparecendo recentemente na última década.

Uma segunda interceção interessante é a que combina a informática com a área de negócio, onde encontramos o *software* aplicacional suportado pelas bases de dados. Na área das bases de dados, com o advento da *web 2.0* (a *web* das pessoas) associada aos dispositivos móveis e à *internet of things* (IoT), as clássicas aplicações empresariais foram largamente ultrapassadas em volume de dados, dando origem ao denominado *Big Data*.

Neste trabalho procuramos explorar as semelhanças e diferenças entre as bases de dados e aprendizagem automática. Estas duas subáreas lidam com o mesmo conjunto de dados, mas de formas diferentes e com resultados distintos, pelo que as convém distinguir. Nesse sentido, depois de definir base de dados e as principais técnicas de aprendizagem automática, exemplificamos as duas abordagens e refletimos sobre as semelhanças e diferenças entre elas.

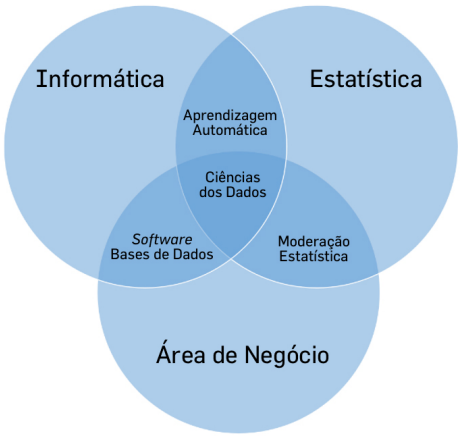

FIGURA 1. Ciência dos Dados.

A história das bases de dados[4] tem início nos anos de 1960, na sequência da utilização dos discos rígidos, que libertou os informáticos das fitas magnéticas com leitura sequencial. Com discos foi possível criar bases de dados hierárquicas e em rede que suportavam estruturas de dados mais complexas como árvores ou mesmo redes.

O modelo de base de dados relacional, que é intensamente utilizado hoje em dia, foi criado por Codd[1]. O modelo relacional para além da sua grande simplicidade é suportado por uma linguagem declarativa, o *SQL* (*Structured Query Language*). Nas linguagens declarativas ao contrário das procedimentais, basta declarar o que se pretende, cabendo ao Sistema Gestor de Base de Dados encontrar a melhor forma de realizar a consulta.

O modelo relacional é baseado num conjunto interligado de tabelas. Cada tabela tem um conjunto de linhas e colunas (ou atributos). Cada linha tem um identificador único chamado chave ou chave principal, como na TABELA 1. As tabelas relacionam-se entre si através de chaves estrangeiras, que não são mais que atributos que são chaves principais em outras tabelas.

A TABELA 1, é retirada de um estudo de caso de um hospital privado, com pagamentos associados, e que indica os doentes que tiveram complicações pós-operatórias.

O *SQL* tem dois conjuntos de comandos: um conjunto para definição da estrutura de dados e um outro para manipular os dados. Como exemplo na definição da estrutura dos dados, podemos referir a criação, alteração ou remoção de uma tabela. Como exemplo da manipulação dos dados para além da inserção, alteração e remoção de linhas, temos a possibilidade de fazer consultas (ou *queries*).

Para realizar uma consulta em *SQL* basta declarar as colunas que queremos visualizar e as respetivas tabelas, sujeitas a uma qualquer condição, com a seguinte sintaxe:

Select <colunas>, From <tabelas>, Where <condição>.

Na última década têm-se popularizado as bases de dados *NoSQL* ("não *SQL*" ou *not only SQL*) que se caracterizam por não optarem pelo modelo relacional e estar orientadas para grandes volumes de dados, associado ao movimento *Big Data*.

Na aprendizagem automática[3], tal como as bases de dados, é possível extrair informação dos dados. Podemos distinguir dois conjuntos de algoritmos, os preditivos e os descritivos, isto é, os que têm capacidade de prever algo no futuro e os que descrevem o passado.

TABELA 1. Tabela do modelo relacional.

doente (chave)	idade	#medicamentos	pagamentos	complicação
1	52	7	121€	sim
2	57	9	7,113€	sim
3	43	6	75€	sim
4	33	6	3,720€	não
5	35	8	4,489€	não
6	49	8	77€	sim
7	58	4	39€	não
8	62	3	79€	não
9	48	0	2,797€	não
10	37	6	90€	sim

De seguida, vamos exemplificar os algoritmos preditivos com a Classificação e os algoritmos descritos com a Segmentação (*clustering*) e Associação (*association rules*).

A Classificação, dado um conjunto de dados de treino com exemplos de diferentes classes, o problema consiste em criar um modelo capaz prever a classe de cada exemplo inserido num novo conjunto de dados, designado por conjunto de dados de teste. Na TABELA 1 a classe discriminante corresponde à coluna "complicação" e o problema de classificação é capaz de prever se um novo doente com 60 anos e que não toma medicação terá eventuais complicações no pós-operatório. O algoritmo de classificação na fase de teste mede da qualidade da classificação. A medida de qualidade da classificação indica também uma boa qualidade de previsão.

Por outro lado, os algoritmos descritivos, não utilizam conjuntos de dados que possuam atributos discriminantes. Eles são também conhecidos por não supervisionados, isto é, não são orientados por nenhum atributo com características especiais.

A Segmentação refere o problema de encontrar partições homogéneas nos dados. O problema de segmentação pode ser definido como dividir as observações (ou linhas) em K subconjuntos, em que a distância entre as observações de cada subconjunto seja mínima e a distância entre os diferentes subconjuntos seja máxima. Para o exemplo da TABELA 1, o problema de segmentação irá dividir os doentes em K grupos pela idade, número de medicamentos e/ou pagamento, para uma melhor caracterização. Na segmentação por pagamento identificam-se quatro doentes que são "grandes utilizadores" do hospital tendo despendido valores superiores a mil euros.

O problema de Associação procura conjuntos frequentes num mesmo atributo. Por exemplo, numa farmácia, quem compra paracetamol para a febre também compra um descongestionante nasal. O algoritmo utiliza como *input* uma tabela com várias transações de produtos, tendo como *output* um conjunto de regras do tipo $E => D$. Para o exemplo a regra é a seguinte: *paracetamol => descongestionante_nasal*. Na geração de produtos frequentes existem duas métricas fundamentais: suporte e confiança. A medida de suporte, ou frequência relativa, é obtida pela razão da frequência absoluta de *E&D* pelo número total de transações. Podemos interpretar ainda esta medida como a probabilidade de compra de E e D em conjunto. A medida de confiança é dada pelo *suporte (E&D) / SUPORTE(E)* e pode ser vista também como a probabilidade condicionada de comprar D se comprou E.

Bases de Dados versus Aprendizagem Automática

As questões colocadas a um Bases de Dados têm semelhanças com as perguntas que se fazem na Aprendizagem Automática. De seguida vamos ver exemplos das duas abordagens, para um exemplo de um Hospital Privado:

Numa Base de Dados pretende-se saber por exemplo:

B1 - quais os doentes com complicações pós-operatórias?

B2 - quais os doentes que pagaram mais de 1.000 euros?

B3 - quais os dois medicamentos mais utilizados?

Enquanto que, em Aprendizagem Automática procuram-se:

A1 - os atributos que levam os doentes a ter complicações pós-operatórias (classificação);

A2 - os grupos de doentes com base nos pagamentos (segmentação);

A3 - o medicamento *X* que é utilizado com o medicamento *Y* (associação).

As perguntas *B1* e *A1* são semelhantes, consideram os doentes com complicações no pós-operatório, mas a primeira tem uma resposta fácil em *SQL* enquanto que a segunda carece da utilização de um algoritmo de classificação.

Da mesma forma as questões *B2* e *A2*, referem os valores pagos pelos doentes no hospital privado, em que a primeira questão é respondida em *SQL*, a segunda para determinar os grupos de doentes é de utilizar um algoritmo de segmentação.

Por fim, as perguntas *B3* e *A3* também referem assuntos semelhantes, os medicamentos. Contudo, para a primeira questão basta ordenar a tabelas dos medicamentos e a segunda carece de um algoritmo de associação.

As duas abordagens tratam os mesmos dados dos sistemas de informação de formas diferentes. Embora as questões sejam semelhantes, nas Bases de Dados é apresentado um padrão (p.ex.: consulta *SQL*) e são devolvidos dados, por outro lado, e em Aprendizagem Automática são fornecidos os dados e pretende-se extrair padrões (p.ex.: regras associativas)[2].

Retomando *B1* e *A1*, em *B1* pretendemos encontrar os doentes que tiveram complicações, sendo devolvido os doentes 1, 2, 3, 6 e 10. A consulta *SQL* é a seguinte:

```
Select doente, idade, #medicamentos
From Tabela1
Where complicação = 'sim'
```

Por outro lado, na questão *A1* são fornecidos os dados e pretende-se extrair padrões. Assim, para a mesma tabela pretendemos saber os atributos que causam as complicações pós-operatórias. Com um algoritmo de classificação encontraremos a seguinte regra:

(Idade >= 35 e #Medicamentos >= 4) Complicações = 'Sim'

Isto é, quem tem 35 ou mais anos e toma 4 ou mais medicamentos tem complicações médicas.

Em resumo, com base no mesmos conjunto de dados do sistema de informação, julgamos ter identificado as semelhanças e diferenças entre Bases de Dados e Aprendizagem Automática:

- nas Bases de Dados é apresentado um padrão (*SQL*) e são devolvidos os dados, por exemplo: *Select doente, idade, #medicamentos From Tabela1 Where complicação =*

'sim', devolve os dados dos doentes com complicações pós-operatórias;

- enquanto que, na Aprendizagem Automática são fornecidos os dados e são devolvidos os padrões (regras), por exemplo: *(Idade >= 35 e #Medicamentos >= 4) Complicações = 'Sim'*, correspondendo à regra dos doentes que se prevê que tenham complicações pós-operatórias.

Desafio: em vez de um hospital ou uma farmácia, escolha uma outra área de negócio que conheça, como por exemplo uma escola; de seguida, crie perguntas que sejam suportadas por um sistema de base de dados e por um sistema de aprendizagem automática.

REFERÊNCIAS

[1] CODD, E.F., *A relational model of data for large shared data banks*, *Communications of the ACM*, vol. 13, no. 6, pp. 377–387. 1970.

[2] DHAR, V., *Data Science and Prediction*, *Communications of the ACM*, vol. 56, no.12, pp. 64-73. 2013.

[3] GAMA, J. *et al.*, *Extração de Conhecimento de Dados*, 3ª edição, Edições Sílabo, ISBN 9789726189145. 2017.

[4] SILBERSCHATZ, A. *et al.*, *Database System Concepts*, 7th edition, McGraw-Hill, ISBN 9780078022159. 2019.