Capítulo 3

Processamento e Modelagem de Dados

A tomada de decisão é um processo cognitivo que envolve tanto o nosso racional quanto o emocional. É um processo que fazemos o tempo inteiro, seja de forma quase automática como atos reflexos, ou de forma consciente e bem criteriosa. No caso de uma tomada de decisão consciente, quanto mais dados tiver em mãos, menor será a incerteza e a margem de erro nas decisões tomadas. Pois, combinando adequadamente as informações úteis contidas nestes dados com os conhecimentos prévios, aumenta a chance de sucesso numa escolha. Porém, em decorrência do aumento vertiginoso dos dados produzidos nesta nossa Era Digital, o grande desafio constitui a descoberta de informações efetivamente relevantes e escondidas nestes dados.

Para que as informações úteis sejam reveladas a partir de um amontoado de dados, é necessário gerenciar estes dados brutos, remover os dados irrelevantes, integrar e transformar os dados relevantes em dados semanticamente significativos (informações) e estruturá-los de forma adequada para a nossa estrutura cognitiva. Nas últimas décadas, diversas pesquisas tem se empenhado para solucionar os problemas relacionados com a transformação de um volume gigantesco de dados brutos em informações essenciais, fáceis de serem entendidas, interpretadas e analisadas.

Pesquisas na área de Banco de Dados tem se ocupado com o desenvolvimento de técnicas eficientes para gerenciamento de um grande volume de dados, incluindo os problemas de consistência, integridade, integração e padronização. Estudos na área de Processamento de Dados tem contribuído com técnicas de validação, ordenação, síntese, agregação, filtragem e classificação de dados. Ferramentas da Estatística Descritiva tem sido

amplamente aplicadas. A comunidade de Mineração de Dados, ou KDD (*Knowledge Discovery in Databases*), por sua vez, tem se focado na descoberta de padrões intrínsecos aos dados que sejam relevantes para embasar uma tomada de decisão. Além dos métodos já mencionados, as técnicas de Estatística Inferencial se mostraram valiosas na análise preditiva.

Neste capítulo daremos uma visão geral das técnicas relacionadas com o gerenciamento e processamento de dados brutos. Veremos que há técnicas de visualização, a maioria inventada pelos estatísticos, que suportam esta etapa de processamento.

3.1 Tipos de Dados

Sendo dados o domínio de todas as operações desenvolvidas, é essencial distinguir os tipos de dados a serem processados em cada contexto.

Sob o ponto de vista de tecnologia de informação, distinguem-se duas formas de organização de dados[1]: **dados estruturados**, armazenáveis em estruturas tabulares e descritíveis pelos modelos relacionais, e **dados não-estruturados**.

Para gerenciar os dados estruturados, sistemas de gerenciamento de banco de dados (SGBD) baseados no modelo relacional e providos da linguagem de interface SQL (*Structured Query Language*) são utilizados. O mais popular é o MySQL [2], lançado em 1995 e mantido pela *Oracle Corporation*. Este *software* é livre para uso não-comercial. No entanto, bancos de dados tem se evoluído para bancos multidomensionais em que cada entrada (fato) é associada a mais de uma tabela, como o modelo estrela (*Star Schema*). Este modelo multidimensional permite uma análise histórica dos dados e desempenho nas consultas, tornando possível distinguir os dados em termos da frequência de acessos: dados transacionais e dados históricos. Os **dados transacionais** constituem os dados referentes às transações rotineiras que ocorrem com a base de dados, ou seja dados para OLTP (*OnLine Transaction Processing*), enquanto os **dados históricos** englobam os dados que suportam gerenciamento, recuperação e análise dos dados, isto é, dados para operações analíticas OLAP (*OnLine Analysis Processing*).

Para os dados não-estruturados, como documentos e modelos geométricos, ainda não existe um sistema de gerenciamento de dados consolidado. Para se beneficiar do conjunto de programas responsávies pelo gerenciamento de

[1] Leandro Augusto da Silva e colegas. Introdução à Mineração de Dados: com aplicações em R

[2] https://pt.wikipedia.org/wiki/MySQL

um banco de dados relacional, disponíveis em MySQL e sistemas similares, é necessário remodelar os dados não-estruturados em dados estruturados. Da Silva e seus colegas apresentaram um exemplo de conversão dos dados não-estruturados de mensagens textuais para os dados estruturados usando R[3]. Vale comentar aqui que a maioria dos problemas envolvem dados não estruturados.

Em termos de visualização, é comum classificar os dados em dados físicos e dados abstratos[4]. Os **dados físicos** são usualmente dados espaciais acerca o mundo físico que nos rodeia, como formas geométricas dos objetos, estruturas moleculares dos genes, amostras dos fluidos (líquidos ou gases) em movimento. Enquanto os **dados abstratos** são aqueles que não tem uma posição espacial associada. Para exibí-los numa tela do computador, é necessário mapeá-los aos pontos de um espaço métrico.

Ainda em relação aos dados físicos, é comum sub-classificá-los em termos dos seus valores[5]. Isso facilita a escolha das representações gráficas para os dados de interesse. Existem dados escalares, vetoriais e tensoriais. Os **dados escalares** são representados por um valor numérico. Os **dados vetoriais** de um espaço de dimensão n são descritos por n valores numéricos. E os **dados tensoriais** de ordem m num espaço de dimensão n possuem n^m valores.

3.2 Estatística Descritiva

É comum perguntarmo-nos por onde começar quado estamos diante de um volume gigantesco dos dados armazenados numa base de dados. A forma mais intuitiva é ter antes de tudo uma visão sumária destes dados. E esta visão pode ser proporcionada pelas ferramentas da estatística descritiva.

A estatística é um conjunto de técnicas que permite de forma sistemática, organizar, descrever, analisar e interpretar um conjunto de dados. A estatística descritiva é um ramo da estatística que visa sintetizar em medidas simples os dados que tem pelo menos uma características em comum (**população**). Quando a população é muito grande, as medidas podem ser inferidas a partir de um subconjunto de elementos representativos da população (**amostra**).

[3]Seções 2.5 a 2.7 do livro "Leandro A. da Silva e colegas. Introdução à Mineração de Dados: Com aplicações em R"

[4]Daniel Keim e colegas. Mastering the information age: solving problems with visual analytics

[5]Alexandru C. Telea. Data Visualization: Principles and Practice

Chamamos de **variável** uma característica associada a uma população. Ela pode ser **qualitativa** (valores não numéricos) ou **quantitativa** (valores numéricos). Uma variável qualitativa pode ser, por sua vez, classificada em nominal (os valores não tem uma ordem natural de ocorrência) e ordinal (os valores tem uma ordem natural de ocorrência). E uma variável quantitativa pode ser diferenciada em contínua (valores contínuos) e discreta (valores discretos).

Reorganizando os dados conforme o número de ocorrências de faixas de valores de uma variável e representando esta **distribuição de frequências**/ocorrências de forma compacta num gráfico conhecido por **histograma**, podemos ganhar uma percepção melhor da resposta de um sistema em relação a uma variável de interesse. O número de ocorrências de uma faixa de valores em relação ao número total de ocorrência de todos os valores é denominado **frequência relativa**. A frequência relativa acumulada até uma certa faixa de valores é chamada **frequência acumulada relativa**.

Existem ainda algumas medidas estatísticas descritivas que sumarizam os valores que uma variável quantitativa pode assumir. Elas são as medidas de tendência central, ou de posição, e as medidas de dispersão em torno de uma medida central denominada a média.

As **medidas de tendência central** mais conhecidas são:

Moda: é o valor que ocorre com maior frequência.

Média (μ): é a média aritmética dos valores x_i de toda população de tamanho N.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{3.1}$$

Quando a população é muito grande, é usual estimar a média da população a partir da média dos N valores x_i representativos

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{3.2}$$

Percentis: de ordem $p \times 100$ em um conjunto de dados de tamanho N é o valor que ocupa a posição $p \times (N+1)$ do conjunto de dados ordenados. Um percentil de ordem 25 (p=0.25) é denominado o primeiro quartil, e de ordem 75, o terceiro quartil.

Mediana: é o valor da variável que corresponde a um percentil de ordem 50. Ele ocupa a posição central de um conjunto de dados ordenados.

E das **medidas de dispersão**, destacamos:

Amplitude: é a diferença entre o maior e o menor valor do conjunto.

Variância (σ^2): é a média dos quadrados das diferenças entre cada valor x_i da população de tamanho N e a média desta população.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2. \tag{3.3}$$

A variância, considerada não enviesada, para os valores x_i das N observações de uma amostra é dada pela expressão

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2. \tag{3.4}$$

Desvio-padrão (σ): é a raiz quadrada da variância. O desvio-padrão amostral $\sigma_{\overline{x}}$ de uma amostra de tamanho N é relacionado com o desvio-padrão da população, da qual ela pertence, através da seguinte relação:

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{N}} \tag{3.5}$$

Coeficiente de dispersão (cv): é o desvio-padrão em relação à média da população. Também conhecido por **dispersão relativa**.

$$cv = \frac{\sigma}{\mu}. \tag{3.6}$$

Há uma outra medida que expressa a variabilidade de um valor x dentro de um conjunto de dados em termos dos múltiplos de desvio-padrão σ. Ela é o **escore-z**, ou a variável normal normalizada:

$$z = \frac{x - \mu}{\sigma}. \tag{3.7}$$

A estatística descritiva oferece também ferramentas apropriadas para estudar a relação entre duas variáveis quantitativas. O coeficiente de correlação de Pearson é o mais conhecido. Dados os N valores x_i e y_i de duas variáveis aleatórias x e y. O **coeficiente de correlação de Pearson**, $\rho \in [-1, 1]$, é obtido através de

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{[\sum_{i=1}^{N}(x_i - \overline{x})^2][\sum_{i=1}^{N}(y_i - \overline{y})^2]}} \tag{3.8}$$

3.3 Filtragem

Em se tratando de analítica visual (uma área emergente que envolve tanto a tecnologia de informação quanto a de visualização), vale a pena darmos um destaque ao conceito de filtragem de dados.

Para tecnologia de IT, filtragem de dados consiste essencialmente em remover os dados indesejados sem alterar os valores do restante dos dados, a fim de aprimorar a qualidade dos dados, reduzir o volume de dados a ser processado, remover os dados sintaticamente incorretos [6], ou floquear dados de ameaça à segurança [7]. Funções lógicas de acesso seletivo aos dados estruturados são suportadas pela maioria dos SGBD.

Na visualização, o conceito de filtragem é usualmente definido no espaço de frequências para o qual as imagens são convertidas. Neste novo espaço, as frequências de variação dos sinais correspondentes às variações das intensidades das cores nas imagens podem ser filtradas/removidas. Quando as altas frequências são removidas, por exemplo através de um filtro Gaussiano, as imagens ficam mais borradas. E se as baixas frequências são removidas, as bordas dos objetos na imagem são realçadas. A operação correspondente no domínio espacial é conhecida por **convolução**[8]. Vale chamar atenção aqui que esta filtragem no domínio espectral pode alterar os valores dos dados no domínio espacial.

3.4 Pré-processamento

É comum que os dados brutos apresentam erros ou falhas que os tornam impróprios para os algoritmos disponíveis. Dispondo das ferramentas da estatística descritiva podemos sumarizar as principais características da massa dos dados, como a média e o desvio-padrão, e aplicá-las no aprimoramento dos dados.

Dois tipos de erros mais comuns na aquisição de dados brutos são:

Falta de dados (*data missing*): algum campo associado à amostra fica vazio. Medidas mais comuns são: descarte da amostra, inserção manual do valor faltante, inserção automático de um valor padrão.

Ruídos (*noisy data*): dados originais são alterados indesejadamente. São também conhecidos em alguns casos por *outliers*. Correções mais apli-

[6] https://www.displayr.com/what-is-data-filtering/
[7] https://www.paloaltonetworks.com/features/data-filtering
[8] https://lodev.org/cgtutor/filtering.html

cadas são: remoção manual e filtragem/suavização automática. Ferramentas estatísticas podem ser aplicadas para detectar automaticamente os *outliers*. Por exemplo, todos os valores com escore-z fora do intervalo $[-3, 3]$.

Essas ações de correção são usualmente conhecidas por **limpeza/polimento de dados**.

Outro problema comum entre os dados brutos é a redundância dos dados, replicando um mesmo valor em diferentes entradas no banco de dados. Esta replicação de dados pode levar o sistema a um estado inconsistente quando alguma entrada deixa de ser atualizada numa transação. Neste caso, é interessante que seja aplicada uma **redução de dados**, integrando um mesmo valor disperso em vários pontos de uma base de dados num único registro.

Pode acontecer que os dados gerados por uma fonte sejam inconsistentes com os dados previamente armazenados. Para evitar que estes dados levem o sistema a um estado inconsistente, é comum removê-los ou modificá-los para um mesmo valor, aseegurando a **consistência dos dados** mesmo que tais ações venham a implicar em valores errados.

Para evitar uma grande quantidade de algoritmos, só pela variabilidade na forma de representação dos dados oriundos de fontes distintas, uma prática comum é uniformizar a representação dos dados num mesmo espaço através de procedimentos conhecidos por **transformação de dados** ou **normalização dos dados**. Em comparações numéricas de dados de diferentes fontes, o uso de dados normalizados pode simplificar bastante um algoritmo de comparação.

É interessante comentar que, quando os dados numéricos estão espacialmente estruturados numa grade regular, é comum aplicar a técnica de filtragem do domínio espectral para suavizar/atenuar os valores dos dados. Outra observação pertinente é, quando possível, preserve uma cópia dos dados brutos. Pois, as técnicas de pré-processamento estão em constante aprimoramento. Dados filtrados hoje em dia podem não corresponder aos dados filtrados daqui a algum tempo.

3.5 Modelos de Dados

A fim de assegurar um ambiente confortável de exploração, conforme o *Visual Information Seeking Mantra*[9] proposto pelo Schneiderman em 1996, um sistema de analítica visual deve ser projetado levando em conta não só a renderização dos dados em imagens inteligíveis como também a responsividade

[9] "*Overview first, zoom and filter, then details-on-demand*".

às ações dos usuários. Como muitas respostas às ações dos usuários implicam em diversos acessos aos dados armazenados, uma organização/modelagem adequada aos possíveis padrões de acessos é fundamental.

Embora o modelo relacional seja adotado na maioria dos SGBD, ele não é apropriado para representar os dados físicos. A diversidade na forma como estes dados são adquiridos, a variedade nos padrões de acesso adotados pelos algoritmos e as relações espaciais e temporais inerentes a eles levaram a inúmeros modelos bem específicos de dados. Modelos tridimensionais CSG (*Constructive Solid Geometry*), Brep (*Boundary Representation*), *Halfedge datastructure*, e enumeração da ocupação espacial (*spatial occupancy enumeration*) são exemplos de modelos propostos para atender demandas particulares de certas aplicações.

Vale comentar que uma estratégia que temos adotado nos nossos projetos de analítica visual é derivar a partir dos dados brutos modelos específicos de suporte às interações preditivas, ao invés de gerá-los sob demanda. Isso tem melhorado a responsividade dos nossos projetos. Um exemplo é o projeto de interações com os exames imagiológicos médicos. Embora os dados destes volumes sejam discretos, a percepção que se tem sobre os volumes renderizados numa tela é que eles sejam contínuos e, para dar melhor suporte às interações com estes dados discretos, porém visualmente contínuos, construímos modelos temporários "contínuos".

É ainda importante lembrar aqui que a tecnologia das unidades (dedicadas) ao processamento gráfico (GPU) tem contribuído muito para melhorar tanto a qualidade quanto à velocidade na renderização das imagens. É quase mandatório implementar os algoritmos de renderização em cima desta tecnologia. Porém, para se tirar máximo proveito dos recursos otimizados oferecidos por ela, devemos procurar modelar os nossos dados compatíveis com as primitivas disponíveis nas GPUs.

Sintetizando, para conciliar o desempenho dos modelos dedicados e as facilidades de gerenciamento da integridade, consistência e segurança dos dados oferecidas pelos SGBD, acreditamos que um sistema de analítica visual deve suportar a co-existência de vários modelos de trabalho apropriados para interações (renderização), porém coordenados com um modelo relacional persistente gerenciado por um SGBD. O grande desafio seria elaborar uma estratégia de coordenação entre estes modelos num cenário de multi-usuários e multi-dados.

3.6 Visualização

Visualização de uma massa de dados com o intuito de ganhar "insights" sobre as relações escondidas nela foi um dos focos da área de pesquisa da Estatística. Os estatísticos inventaram vários gráficos e diagramas para aprimorar a apresentação dos dados estruturados. Muitos deles revelaram padrões e ajudaram no esclarecimento de alguns fenômenos aparentemente inexplicáveis[10].

No ambiente R[11], que provê interfaces a diferentes SGBD, são implementados os seguintes recursos básicos de visualização[12]:

Histograma (gráfico de densidade) (*density plot*): apropriado para visualizar a distribuição de frequências de uma massa de dados.

Gráfico de barras (*bar plot*): apropriado para visualizar comparativamente os valores de uma variável qualitativa.

Gráfico de pizza (*pie char*): apropriado para visualizar comparativamente os valores de uma variável qualitative em relação ao todo.

Gráfico de pontos (*dot plot*): apropriado para visualizar possível clusterização de variáveis quantitativas (escalares).

Gráfico de linhas (*line chart*): apropriado para visualizar a relação linear por parte entre duas variáveis quantitativas.

Gráfico de caixas (*boxplot*): apropriado para visualizar a dispersão dos valores de uma variável quantitativa.

Gráfico de dispersão (*scatterplot*): apropriado para visualizar a correlação entre duas variáveis

e alguns formatos mais avançados[13]

Gráfico de mosaico (*mosaic plot*) apropriado para visualizar a relação de duas ou mais variáveis qualitativas.

Gráfico de reticulado (*lattice plot*) apropriado para visualizar alternativas representações de um mesmo conjunto de dados organizadas numa grade.

[10] William S. Cleveland. Visualizing Data
[11] https://www.statmethods.net/r-tutorial/index.html
[12] https://www.statmethods.net/graphs/creating.html
[13] https://www.statmethods.net/advgraphs/index.html

Gráfico de coordenadas paralelas (*parallel coordinates plot*[14]) apropriado para comparar valores de multivariáveis.

3.7 Exercícios

1. Como os objetos, ou entidades, e seus atributos são representados e identificados de forma unívoca no modelo relacional?

2. Leia as seções 2.5 a 2.7 do livro "Leandro A. da Silva, Sarajane Marques Peres e Clodis Boscarioli. Introdução à Mineração de Dados: Com aplicações em R". Reproduza o exemplo no ambiente R.

3. Discuta com suas palavras as possíveis razões para classificação distinta de tipos de dados na área de Banco de Dados e na área de Visualização.

4. *Fisher's Iris dataset* é um conjunto clássico de dados de três variedades de íris[15]. Compute os sumários estatísticos com uso de R para as quatro variáveis associadas às amostras (largura e comprimento de sépalas, largura e comprimento de pétalas): medidas de tendência central (média, mediana, moda e primeiro e terceiro quatis) e medidas de dispersão (variância, desvio-padrão e coeficiente de dispersão). É possível distinguir as 3 variedades com base nos sumários estatísticos e os coeficientes de correlação de Pearson?

5. *Ascombe'se quartet*[16] consiste de quatro conjuntos de dados que demonstram a insuficiência da estatística descritiva para diferenciar conjuntos de dados nitidamente distintos. Compute os sumários estatísticos com uso de R para as variáveis y_1, y_2, y_3 e y_4 associadas às amostras: medidas de tendência central (média, mediana, moda e primeiro e terceiro quatis) e medidas de dispersão (variância, desvio-padrão e coeficiente de dispersão). É possível distinguiros os quatro conjuntos de dados através dos coeficientes de correlação de Pearson?

6. Por quê, em comparações numéricas, o uso de valores normalizados pode simplificar um algoritmo de comparação?

7. O que você entende por "*visual data reduction techniques*" e "*new interaction paradigms for large sets of users*" no trecho abaixo extraído

[14] https://plot.ly/r/parallel-coordinates-plot/
[15] https://en.wikipedia.org/wiki/Iris_flower_data_set#Use_of_the_data_set
[16] https://en.wikipedia.org/wiki/Anscombe%27s_quartet

do livro de D. Keim e colegas? (Dica: Lembre-se da evolução do Mantra de Schneidermann proposto em 1996,"*Overview first, zoom/filter, details on demand*", para a demanda atual,"*Analyze first, show the important, zoom/filter, analyse further, details on demand*".)

"*In particular, we need better logic based integration systems, a tighter integration between visualisation and data, more precise quality indicators, visual oriented data reduction techniques, and new interaction paradigms for large sets of users.*"

8. O formato OBJ é um formato tabular padrão de representação de objetos geométricos 3D desenvolvido pela empresa *Wavefront Technologies*. Veja a descrição de um cubo no formato OBJ em http://paulbourke.net/dataformats/obj/. Identifique os dados redundantes. Pesquise uma forma de redução de dados que minimize esta redundância (Dica: *Halfedge Data Structure*[17])

9. Elabore um algoritmo de conversão dos modelos geométricos CSG, Brep e enumeração espacial para o modelo relacional.

10. Na seção 3.5 foi apresentado um caso de interações com dados discretos, porém visualmente contínuos. Para evitar que um usuário receba *missing data* como resposta às suas interações, qual estratégia você adotaria para contornar esta falta de dados? (Dica: Veja uma solução em https://www.ncbi.nlm.nih.gov/pubmed/21301030)

11. Represente graficamente as relações entre as quatro variáveis da questão 4 através de *scatterplots*[18]. É possível distinguir as três variedades com esta visualização? Explique.

12. Represente graficamente as relações entre as quatro variáveis da questão 5 através de *scatterplots*[19]. É possível distinguir os quatro conjuntos com esta visualização? Explique.

[17] https://doc.cgal.org/latest/HalfedgeDS/index.htmlhttps://doc.cgal.org/latest/HalfedgeDS/index.html
[18] https://www.statmethods.net/graphs/scatterplot.htmlhttps://www.statmethods.net/graphs/scatterplot.html
[19] https://www.statmethods.net/graphs/scatterplot.htmlhttps://www.statmethods.net/graphs/scatterplot.html