Aprendizado Descritivo

Aula 13 – Mineração de modelos excepcionais

Professor Renato Vimieiro

DCC/ICEx/UFMG

Introdução

- A tarefa de descoberta de subgrupos, em certa medida, generalizou a tarefa de mineração de padrões frequentes incorporando um atributo alvo à busca pelos padrões interessantes
- Mesmo sendo útil em vários contextos, a tarefa ainda pode ser aprimorada, em particular, se considerarmos que nem sempre uma única variável alvo (ou até mesmo um conjunto) pode refletir a qualidade dos padrões descobertos
- Em algumas situações, a excepcionalidade do subgrupo é percebida somente através de um modelo ajustado sobre os dados

Introdução

- Vamos considerar um conjunto de dados em que existem duas variáveis alvo
- A figura ao lado representa essa base de dados
  - Os atributos não-alvo não são exibidos
- É perceptível que nenhuma das duas variáveis define, em princípio, um subgrupo excepcional
- Contudo, existe um subconjunto dos objetos que parece apresentar um comportamento distinto com respeito à população

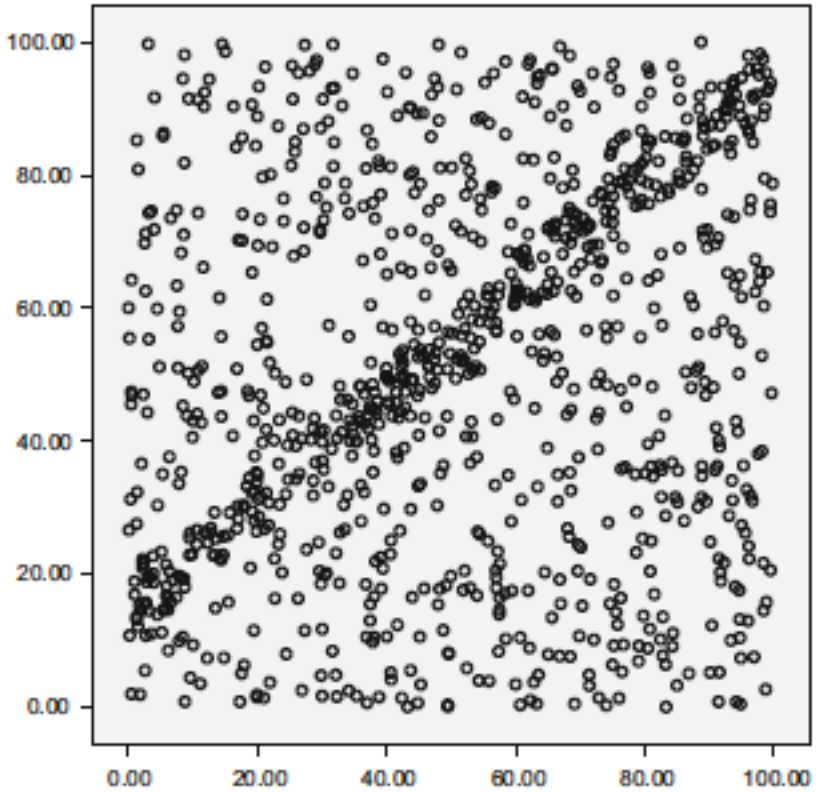
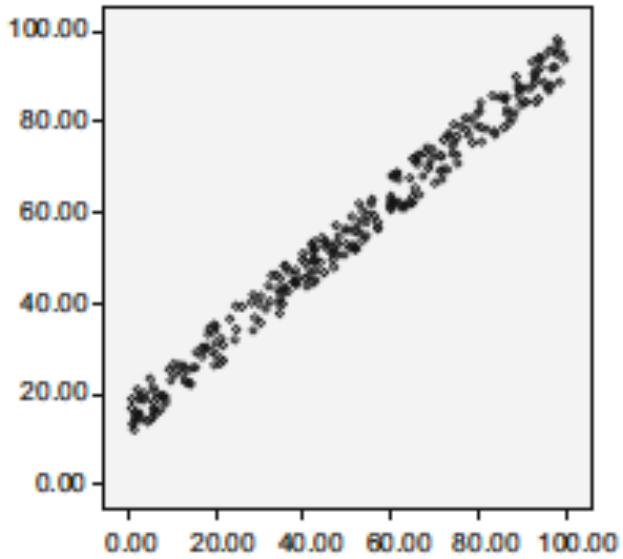
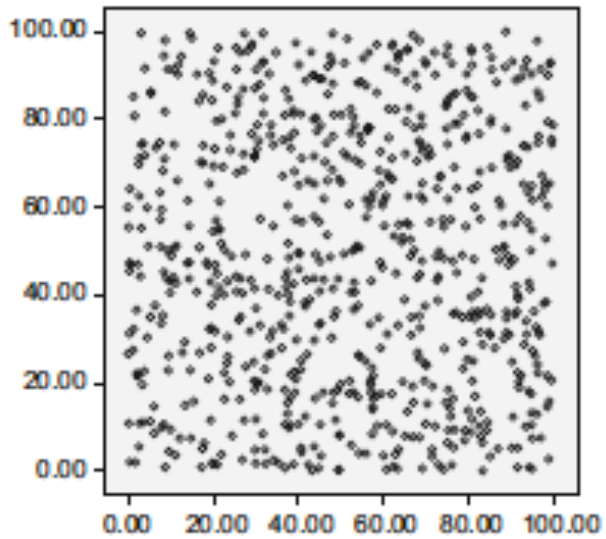


Figura retirada de Duivesteijn et al. (2016)

Introdução

- Separando os grupos, temos as duas figuras abaixo
  - Grupo à esquerda parece ter uma correlação positiva das variáveis
  - Grupo à direita parece não ter qualquer correlação





Figuras retiradas de Duivesteijn et al. (2016)

Introdução

- Nesse caso, a excepcionalidade deveria ser medida através da comparação dos modelos ajustados no grupo e em seu complemento
  - O modelo em questão seria a correlação entre as variáveis alvo em cada grupo
- Em outras palavras, em algumas aplicações, estamos interessados em encontrar subgrupos em que um modelo ajustado sobre eles é excepcional em relação à população
- Queremos generalizar a tarefa de SD de ‘encontrar subgrupos com distribuição não-usual’ para ‘encontrar subgrupos com modelos excepcionais’
  - Essa tarefa é conhecida como mineração de modelos excepcionais (exceptional model mining) (EMM)

EMM

- Formalmente, os objetos do conjunto de dados agora é descrito por dois conjuntos de atributos:
  - Um conjunto $\mathcal{A}$ de atributos descritivos (não-alvo); $|\mathcal{A}|$=q
  - Um conjunto $\mathcal{L}$ de atributos alvo (rótulos) dos objetos; $|\mathcal{L}|$=m
- O domínio tanto dos atributos descritivos quanto dos alvo é livre
- Podemos ter atributos numéricos, categóricos ou lógicos tanto para descrever os objetos quanto para rotulá-los

EMM

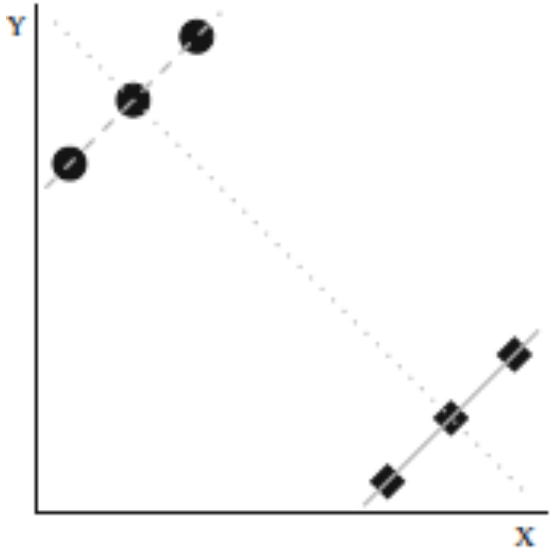
- Tal qual fazíamos em SD, devemos definir uma linguagem de descrição que estabelece o espaço de busca dos subgrupos
- Descrições nessa linguagem continuam sendo conjunções de seletores definidos sobre os atributos descritivos
- A medida de qualidade depende da classe de modelo a ser ajustado nos grupos
  - Antes tínhamos variações essencialmente entre numéricos e categóricos
  - Agora medidas de qualidade específicas para a classe de modelo que estamos considerando
- Em geral, usamos um teste de hipótese para comparar os modelos e determinar se um subgrupo é interessante ou não
  - A medida de qualidade é 1- p-value

EMM

- Usamos o p-value como heurística para a qualidade do subgrupo
  - Não queremos medir se a diferença entre os modelos é estatisticamente significativa ou não
  - Para isso, deveríamos considerar correções no nível de significância para compensar os múltiplos testes realizados durante a busca
- Outra discussão importante é com quem devemos comparar um subgrupo para determinar sua excepcionalidade?
- Fizemos essa discussão brevemente no contexto de SD, mas agora ela se torna mais relevante

EMM

- Considere a figura ao lado, as três retas (pontilhada, tracejada e sólida) representam modelos de regressão linear ajustados na população, bolinhas e quadradinhos, respectivamente
- Se compararmos a inclinação das retas dos modelos, teremos situações bem distintas se considerarmos o complemento e a população para computar a medida de qualidade
- Se o subgrupo encontrado for formado por bolinhas,
  - Ele é interessante se considerarmos a comparação com a população
  - E não interessante se compararmos com o complemento



EMM

- Um segundo aspecto não menos importante a ser considerado é o custo computacional associado
- Se avaliarmos um total de n grupos durante a busca pelos top-k subgrupos, então:
  - Ajustaremos n+1 modelos no total se considerarmos a comparação com a população; e
  - Ajustaremos 2n modelos se considerarmos o complemento
- Se o custo de ajustar o modelo for alto, a segunda opção pode ser inviável
- Além disso, a própria medida de qualidade pode ter um custo computacional alto por ser necessário computar algum valor sobre o grupo
  - Então, esse custo também deve ser levado em consideração
- Contudo, não existe uma melhor escolha entre população e complemento, a escolha vai depender da aplicação e do sentimento do especialista do domínio

Correlação

- Vamos estudar algumas medidas de qualidade propostas por Duivesteijn et al. (2016) para modelos simples e recorrentes em diversos trabalhos
- Vamos começar pela correlação de Pearson.
- Como vimos, a correlação pode ser informativa sobre a excepcionalidade de um grupo
- Assim, considerando duas variáveis alvo x e y, podemos usar o coeficiente de correlação como modelo
  - $r = \sum(x_i - \bar{x})(y_i - \bar{y}) / \sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}$

Correlação

- Uma medida de qualidade associada poderia ser a diferença absoluta das correlações
  - $\varphi_{abs}(X) = |r_x - r_C|$
- O problema dessa medida é que ela favorece grupos pequenos
  - É mais fácil observar uma correlação positiva/negativa mais alta em subgrupos pequenos, o que eventualmente destoa da população/complemento
- Podemos ajustar pelo tamanho do grupo com um fator n ou $\sqrt{n}$ como feito em SD
  - Em alguns casos, esse fator pode ser contraintuitivo, pois, buscar por grupos exageradamente grandes nos passa a noção da busca por um padrão global
- Uma forma mais sofisticada é ajustar pela entropia, que favorece partições mais equilibradas
  - $H(X) = p(X)\log p(x) - p(C)\log p(C)$, C é o grupo com o qual comparamos (população ou complemento)
- Logo, temos a medida ajustada como
  - $\varphi_{ent}(X) = H(X) \cdot |r_x - r_C|$

Correlação

- Alternativamente, a literatura de estatística apresenta um teste de hipótese para comparar o coeficiente de correlação de duas amostras independentes
- Isso envolve uma transformação nos coeficientes chamada de transformação de Fisher
  - $z = \frac{1}{2}\ln\frac{1+r}{1-r}$
- O coeficiente transformado z tem uma distribuição aproximadamente normal
  - $z^* = \frac{z_X - z_C}{\sqrt{\frac{1}{n-3}+\frac{1}{n_C-3}}}$
  - z* segue a distribuição normal padrão, e, assim, podemos usá-lo para computar o p-value

Regressão linear simples

- Outro modelo muito usado para o qual os autores adaptaram uma medida de qualidade foi a regressão linear simples
  - $y = \alpha + \beta x$
- Nesse caso, nosso interesse é, provavelmente, em avaliar se há mudança na inclinação da reta ajustada em cada grupo
- Assim, assumindo o uso do método dos mínimos quadrados para ajustar β, e uma estimativa de variância definida por
  - $s^2 = \frac{\sum(\hat{y}-y)^2}{(N-2)\sum(x_i-\bar{x})^2}$
- A literatura estatística novamente nos informa que o teste t pode ser usado com a estatística e graus de liberdade:
  - $t' = \frac{\widehat{\beta}_X - \widehat{\beta}_C}{\sqrt{s_X^2+s_C^2}}$ e $df = \frac{(s_X^2+s_C^2)^2}{\frac{s_X^4}{n-2}+\frac{s_C^4}{n_C-2}}$

Classificação

- Os autores trataram o caso de classificação através de um modelo de regressão logística

$$\text{logit}(P(y_i = 1|x_i)) = \ln\left(\frac{P(y_i = 1|x_i)}{P(y_i = 0|x_i)}\right) = \beta_0 + \beta_1 \cdot x_i$$

- Para avaliar o efeito do subgrupo, eles consideram o modelo

$$\text{logit}(P(y_i = 1|x_i)) = \beta_0 + \beta_1 \cdot D(i) + \beta_2 \cdot x_i + \beta_3 \cdot (D(i) \cdot x_i)$$

$$\text{logit}(P(y_i = 1|x_i)) = \begin{cases} (\beta_0 + \beta_1) + (\beta_2 + \beta_3) \cdot x_i & \text{if } D(i) = 1 \\ \beta_0 + \beta_2 \cdot x_i & \text{if } D(i) = 0 \end{cases}$$

- A medida de qualidade é obtida a partir do p-value associado a β_3

Leitura

- Duivesteijn, W., Feelders, A.J. & Knobbe, A. Exceptional Model Mining. *Data Min Knowl Disc* **30**, 47–98 (2016). https://doi.org/10.1007/s10618-015-0403-4

Aprendizado Descritivo

Aula 13 – Mineração de modelos excepcionais

Professor Renato Vimieiro

DCC/ICEx/UFMG