O que pode um teste de diagnóstico dizer sobre detecção de anomalias?

M. Rosário de Oliveira[1], L. Gonçalves[2], A. Pacheco[1], R. Valadas[3]

[1]Departamento de Matemática, IST e CEMAT
[2]Unidade de Epidemiologia e Bioestatística, IHMT e CEAUL
[3]Departamento de Engenharia Electrotécnica e de Computadores, IST e IT

Seminário no Âmbito Mestrado Bioestatística, Lisboa, 5 Novembro, 2010

Sumário:

1. Introdução

1. Introdução

2. O Problema do Médico...

- **Testes de diagnóstico** são de grande importância em Medicina visando **classificar** um indivíduo em Doente ou Não Doente

- **Problema:** Avaliar o **desempenho** de novos testes de diagnóstico, na ausência de um teste de referência perfeito (***Gold Standard***)

3. Indicadores do desempenho de um teste de diagnóstico

- Seja X_i ($i = 1, \ldots, p$) o resultado do i-ésimo teste de diagnostico

$$X_i = \begin{cases} 1, & \text{se o teste dá } \textbf{indicação} \text{ que o indivíduo está doente i.e. } (+) \\ 0, & \text{caso contrário } (-) \end{cases}$$

- Seja Y o verdadeiro estado do indivíduo - Variável Latente

$$Y = \begin{cases} 1, & \text{se o indivíduo está doente} \\ 0, & \text{caso contrário} \end{cases}$$

3. Indicadores do desempenho de um teste de diagnóstico

As medidas usuais de avaliação do desempenho de um teste de diagnóstico são:

- **Sensibilidade** (Se): probabilidade de um indivíduo doente (D) ser correctamente identificado como doente, pelo teste de diagnóstico (**recall**) ($+$), i.e.

$$Se = P(+|D) = P(X = 1|Y = 1)$$

- Taxa de Falsos Negativos: $TFN = 1 - Se = P(X = 0|Y = 1)$
- **Especificidade** (Sp): probabilidade de um indivíduo não doente ($\bar{D}$) ser correctamente identificado como não doente, pelo teste de diagnóstico ($-$), i.e.

$$Sp = P(-|\bar{D}) = P(X = 0|Y = 0)$$

- Taxa de Falsos Positivos: $TFP = 1 - Sp = P(X = 1|Y = 0)$

3. Indicadores do desempenho de um teste de diagnóstico

- **Valor Predictivo Positivo** (VPP): probabilidade de um indivíduo diagnósticado como doente ($+$) estar de facto doente (**precision**) (D), i.e.

$$VPP = P(D|+) = P(Y = 1|X = 1)$$

- **Valor Predictivo Negativo** (VPN): probabilidade de um indivíduo diagnóstico como não doente ($-$) não estar doente ($P(\bar{D}|-)$, i.e.

$$VPN = P(\bar{D}|-) = P(Y = 0|X = 0)$$

4. Abordagem clássica

Avaliação do desempenho de testes de diagnóstico:

- **Abordagem clássica – Médicos**: comparação do novo teste com um teste de referência idealmente perfeito (***Gold Standard:*** $Se = Sp = 1$)

- Prova-se que o uso de um *Gold Standard* **Imperfeito** como referência conduz a **estimativas enviesadas!**

4. Abordagem clássica - o que os Médicos sabem...

Como ***Gold Standard imperfeito*** escolhe-se frequentemente o melhor teste disponível

E.g. em parasitologia ainda se escolhe, com frequência, um teste parasitológico, Se<1 e $Sp(X_1) = 1$, como referência

- "*If culture for pertussis is assumed to be <100% sensitive and 100% specific and culture* $(\tilde{Y})$ *is used as the gold standard for assessing the index test* (X)*, then the index test's* ***sensitivity*** *estimate will be* ***unbiased*** *but the* ***specificity*** *estimate will be* ***biased*** *. . .*"

(Baughman et al., 2008)

Verdade se $(\tilde{Y} \amalg X | Y = j)$, onde Y é o *Gold Standard*

4. Abordagem clássica - o que os Médicos sabem...

Em geral, se $See(X) = P(X = 1|\tilde{Y} = 1)$, $(\tilde{Y} \amalg X|Y = 1)$ e $\eta = P(Y = 1)$ então

$$See(X) = \frac{Se(X)Se(\tilde{Y})\eta + (1 - Sp(X))(1 - Sp(\tilde{Y}))(1 - \eta)}{Se(\tilde{Y})\eta + (1 - Sp(\tilde{Y}))(1 - \eta)}$$

Logo, $Sp(\tilde{Y}) = 1$ implica $See(X) = Se(X)$

De um modo geral,

$$See(X) - Se(X) = \frac{(1 - Sp(\tilde{Y}))(1 - \eta)}{P(\tilde{Y} = 1)}(1 - Se(X) - Sp(X))$$

Logo, $See(X) - Se(X) < 0$ sse $Se(X) + Sp(X) > 1$

4. Abordagem clássica - o que os Médicos sabem...

"*. . .the* ***specificity*** *estimate will be* ***biased in the direction of lower estimates***" (Baughman et al., 2008)

Verdade se $Se(X) + Sp(X) \geq 1$, $(\tilde{Y} \amalg X | Y = 0)$ e $\eta = P(Y = 0)$

De um modo geral,

$$Spp(X) - Sp(X) = \frac{(1 - Se(\tilde{Y}))\eta}{P(\tilde{Y} = 0)} (1 - Se(X) - Sp(X))$$

Logo, $Spp(X) - Sp(X) \leq 0$ sse $Se(X) + Sp(X) > 1$

4. Abordagem clássica - o que os Médicos sabem...

- "*Under the assumption that the index test and culture are conditionally independent, the negative bias of the specificity estimate* ***increases*** *as the sensitivity of culture* ***decreases*** *and as the prevalence of pertussis increases*"
  (Baughman et al., 2008)

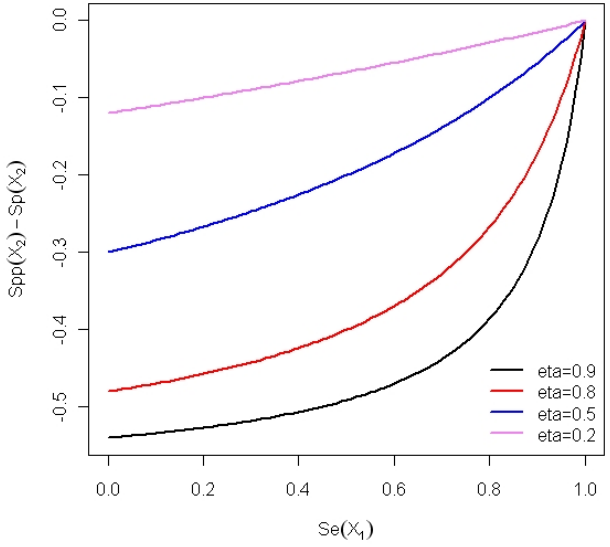
  Verdade se $Se(X) + Sp(X) \geq 1$
- Reformulando, se $(\tilde{Y} \amalg X | Y = 0)$ então:

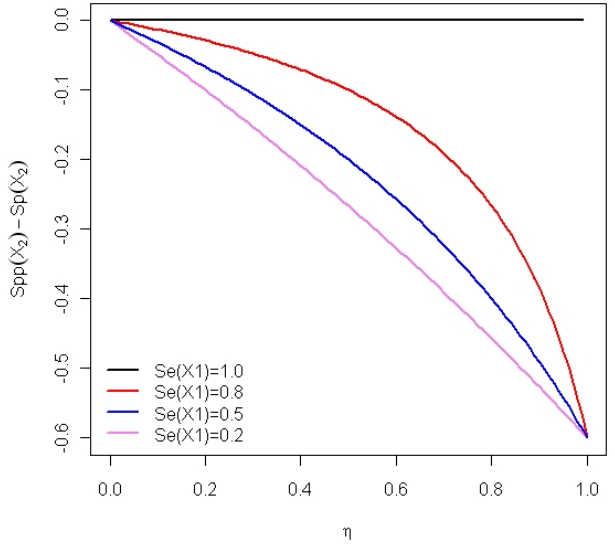
$$Bias = Spp(X) - Sp(X) \leq 0 \text{ sse } Se(\tilde{Y}) + Sp(\tilde{Y}) \geq 1,$$
$$\text{onde } Spp(X) = P(X = 0 | \tilde{Y} = 0)$$

4. Abordagem clássica - o que os Médicos sabem...

Em particular,

$$Bias = Spp(X) - Sp(X) = \frac{\eta\left[1-Se(\tilde{Y})\right][1-Se(X)-Sp(X)]}{\eta\left[1-Se(\tilde{Y})\right]+(1-\eta)Sp(\tilde{Y})}$$





5. Análise de discrepâncias: O que alguns Médicos fazem...

- **Objectivo:** Se o teste de referência é tal que $Sp(\tilde{Y}) = 1$ (e.g. teste parasitológico), sabe-se que se $\{\tilde{Y} = 1\}$ então o indivíduo está parasitado. No entanto, se $\{\tilde{Y} = 0\}$ este pode ser ou não um resultado correcto
- A ideia é "**confirmar**" se os casos $\{\tilde{Y} = 0\}$ e $\{X = 1\}$ são de facto **falsos positivos**
- Aplica-se um novo teste (idealmente, 100% fiável) a este indivíduos. Se o novo teste der positivo considera-se que o resultado $\{\tilde{Y} = 0\}$ estava errado e esta observação passa a ser considerada um **positivo**: $\{\tilde{Y} = 1\}$ e $\{X = 1\}$
- **Actualizam-se** as estimativas de $Se(X)$ e $Sp(X)$

5. Análise de discrepâncias: o que os Médicos sabem...

- "*. . .discrepant analysis (DA) merely* ***substitutes*** *incorporation bias for imperfect gold standard bias. . .* "
  (Baughman et al., 2008)

- "*DA-based estimate of specificity is typically less biased than that based on culture and that the DA-based estimate of specificity shows little appreciable bias...*" *(Green et al., 1998)*

  "I show that those conclusions are incorrect.... I demonstrate that the concept of discrepant analysis is ***profoundly flawed*** *and* ***unscientific***" (Hadgu, 1999)

5. Análise de discrepâncias: o que os Médicos sabem...

- De facto, pode provar-se que se $Sp(\tilde{Y}) = 1$ e $(\tilde{Y} \amalg X | Y = j)$ então:

$$Se^{DA}(X) = \frac{Se(X)}{Se(X) + Se(\tilde{Y})(1 - Se(X))}$$

$$Bias = Se^{DA}(X) - Se(X) \geq 0$$

$$Sp^{DA}(X) = \frac{Sp(X)Sp(\tilde{Y})(1-\eta) + (1 - Se(X))(1 - Se(\tilde{Y}))\eta}{(1 - Se(X))(1 - Se(\tilde{Y}))\eta + Sp(\tilde{Y})(1-\eta)}$$

$$Bias = Sp^{DA}(X) - Sp(X) \geq 0$$

5. Análise de discrepâncias: o que os Médicos sabem...

Alternativas:

- Observar todas os testes de diagnóstico e aplicar MCL (Baughman et al., 2008)
- Admitir mecanismos de censura apropriados e estimar $Se(X)$ e $Sp(X)$ baseado no algoritmo SEM ou usar metodologias bayesianas (Achar et al., 2005; CEB, 2009)

6. Modelo de classes latentes

- O modelo de classes latentes é um dos mais **populares** modelos de variáveis latentes. Foi proposto (e tem sido largamente utilizado) no contexto de problemas das **ciências sociais e humanas**

  (Lazarsfeld e Henry, **1968**)

- A partir de meados dos anos 80, têm sido aplicado em áreas ligadas à saúde, sobretudo na estimação da **sensibilidades** e **especificidades** de testes de diagnóstico (Young, 1985)

6. Modelo de classes latentes

- **Filosofia**: Admite-se que as **associações** entre as variáveis observáveis podem ser explicadas por uma **variável latente**
- **Pressuposto**: Fixo um valor (classe latente) para a variável latente as variáveis observáveis são independentes — **Hipótese de Independência Condicional (HIC)**
- Por vezes esta é uma hipótese pouco realista. Existem generalizações deste modelo que contemplam dependências locais (Qu et al., 1996 e Dendukuri et al, 2008)

6. Modelo de classes latentes

Parametros do modelo:

- $\eta_j = P(Y = j),\ j = 0, 1$ (Prevalencia $= \eta$)
- $\pi_{ij} = P(X_i = 1 | Y = j)$, $i = 1, \ldots, p$; $j = 0, 1$
  - $\pi_{i1} = P(X_i = 1 | Y = 1) = Se(X_i)$, Sensibilidade do i-ésimo teste de diagnóstico
  - $\pi_{i0} = P(X_i = 1 | Y = 0) = 1 - Sp(X_i)$, 1-Especificidade do i-ésimo teste de diagnóstico

- **Hipótese**:
  - Dado o verdadeiro estado da doença ($Y = j$), os resultados dos testes de diagnóstico, (X_i) são independentes – **Hipótese de Independência Condicional (HIC)**.

7. Um exemplo...

Leishmaniasis conjunto de doenças provocadas por um parasita

- Veículo de transmissão: mosquito;
- Potenciais infectados: Homens e cães;
- Os cães são uma fonte de infecção para os humanos;
- Infecção nos humanos detectada em **16** países europeus, que incluem a França, Italia, Grécia, Malta, Espanha e Portugal (***OMS***);

7. Um exemplo...

Leishmaniasis: conjunto de doenças provocadas por um parasita

- Ocorre em diversas formas, a doença é geralmente reconhecida por sua forma cutânea que provoca lesões não-fatais mas desfigurantes, embora epidemias da variante visceral causam milhares de mortes por ano potencialmente fatal milhares forma visceral causa de mortes; (***OMS***)
- Uma das 10 doenças que *Organização Mundial de Saúde* (***OMS***) financia projectos de investigação.

- Um indivíduo pode estar infectado ***sem*** mostrar sinais clínicos da doença
- **Os testes de diagnóstico** são usados para diagnosticar a doença

7. Exemplo...

Seja X_k o resultado de um **teste parasitologico.**Sabe-se que se um parasita é **detectado** então o indivíduo está **infectado**. No entanto, se o parasita **não é detectado**, **não se pode concluir** que o indivíduo **não está infectado**

- Se o parasita é detectado ($X_k = 1$), $\Longrightarrow$ o indivíduo está infectado ($Y = 1$)

  ENTÂO

- $P(Y = 1|X_k = 1) = 1$

  LOGO

- $\pi_{k0} = 0 \quad \Leftrightarrow \quad Sp(X_k) = 1$

7. Um exemplo...

Testes Parasitologicos:

- Medula (Bone Marrow) (*ParMar*),
- Fígado (Liver) (*ParL*),
- Baço (Spleen) (*ParS*)

- Uma vez que o **Fígado** e o **Baço** não são os principais alvos deste estudo, constroi-se a variável (*ParLS*):

$$ParLS = \begin{cases} 1, & \text{se } ParL = 1 \text{ ou } ParS = 1 \\ 0, & \text{c.c.} \end{cases}$$

7. Um exemplo...

- **Teste de DNA**: Bone Marrow PCR (*PCRMar*),
- **Testes Serologicos**: *CIE* e *IFI*

- Combinados por razões médicas e para superar violações da HIC (*ParLS*):

$$CIE\_IFI = \begin{cases} 1, & \text{se } CIE = 1 \text{ e } IFI = 1 \\ 0, & \text{c.c.} \end{cases}$$

7. Um exemplo...

$\hat{\eta}_1$		*ParLS*	*PCRMar*	*ParMar*	*CIE_IFI*
0.231	Especificidade	1.000	1.000	1.000	0.960
	Sensibilidade	0.966	0.966	0.833	0.867

- Melhor teste diagnóstico: *ParLS* e *PCRMar*
- A nossa recomedação: *PCRMAR*, já que *ParLS* não é interessante (só pode ser aplicável em animais mortos)

8. O Problema do Engenheiro...

1. **Identificação de aplicações**:
   - Existem ferramentas de inspecção de pacotes (***Deep Package Inspection*** - **DPI**)
   - Não conseguem identificar todas as aplicações, e.g. **tráfego cifrado**
   - Existe sempre uma percentagem de tráfego que **não** é **identificado** (**Orange DPI**: entre 14% a 39% tráfego não identificado, Pietrzyk et al., 2009)
   - Sabe-se que não são 100% fiáveis. Segundo Pietrzyk et al. (2009), Orange DPI é melhor que Tstat, mas acreditam que **cometem erros**...

9. O Problema do Engenheiro...

2. **Detecção** automática **de anomalias** em redes de computadores:
   - Invasões a computadores, volumes anormais de tráfego
   - Também tem aplicações em tráfego **rodoviário**
   - **Problema:** Avaliar o **desempenho** de novos métodos de identificação de aplicações ou detecção de anomalias

9. Abordagem clássica - o que os Engenheiros sabem...

- ***Ground truth*** perfeito muito difícil de estabelecer
  - Exige o conhecimento da **lista** completa de **anomalias** em conjuntos volumosos de dados. Por sua vez, estes podem ter pouca qualidade

    P. ex. apenas se observam sub-conjuntos de dados (amostragem)

- Muitos dos estudos efectuados em detecção de anomalias e identificação de aplicações carecem de estudos de análise de sensibilidade e de **determinação efectiva** das suas **propriedades** (Ringberg et al., 2008)

8. Análise de discrepâncias: O que alguns Engenheiros fazem...

- Se o procedimento de referência identifica o tráfego como sendo anómalo então é verificado **manualmente** se a anomalia se confirma ou não
- Assim, **actualizam** o *ground truth* imperfeito, que servirá de referência para avaliar o classificador em estudo

8. Os Engenheiros e o modelo de classes latentes

- **Parâmetros do modelo:** $P(Y = j)$ e $P(X_i = x_i | Y = j)$, $i = 1, \ldots, p$, $j = 0, \ldots, k-1$
- Importando as ideias bem estabelecidas em Medicina, pretendemos aplicar este método para **avaliar métodos** de detecções de anomalias ou classificadores
- **Variáveis observáveis**, X_i, são o resultado da classificação de cada objecto em anomalia ou não anomalia
- **Classes latentes** são a verdadeira classe de tráfego: anomalia ou não anomalia

8. Os Engenheiros e o modelo de classes latentes

- **Possíveis problemas:** **HIC** e necessidade de comparar pelo menos **3** métodos. As matrizes **esparsas** provavelmente deixarão de ser problema, dado o elevado volume de dados envolvidos nos problemas de telecomunicações
- As **estimativas** (de máxima verosimilhança) dos parâmetros do modelo permitem-nos facilmente estimar **indicadores** como *recall* (sensibilidade) ou *precision* (valor predictivo poitivo)...
- Está largamente explorado na literatura médica as **vantagens** e desvantagens desta abordagem sobre as anteriormente descritas

9. Comentário Final