Estudo Comparativo sobre Seleção de Variáveis em Classificação Supervisionada

Ana Sousa Ferreira[1]

Anabela Marques[2]

[1]Faculdade de Psicologia, Universidade de Lisboa e Bussiness Research Unit – IUL

[2] Escola de Tecnologia do Barreiro, Instituto Politécnico de Setúbal

Plano da apresentação

1. Introdução
2. Seleção de Variáveis
    - 2.1. Métodos Descritivos
    - 2.2. Métodos Inferenciais
3. Análise Discriminante Discreta (ADD)
4. Combinação de Modelos em ADD
5. Análise dos Resultados
6. Conclusões

Introdu

d

Os dados

Introdução

O Questionário SOC

Com a ad
versão curt
se conhecer
compensação
com as perdas e m que
ocorrem na terceira e idade.

Exemplo para três classes e 2 variáveis explicativas binárias

Estado	Classe 1	Classe 2	Classe 3
00	5	1	0
01	2	0	1
10	1	5	2
11	0	2	5

Neste estudo iremos considerar 10 variáveis binárias do Questionário SOC (que geram 1024 estados possíveis).

Introdução

O Questionário SOC

Exemplo de ítem:

Escolha, em cada descrição que se segue, qual o sujeito mais parecido consigo (A ou B):

A - **(Cotação 1)**

Quando as coisas não correm tão bem como antes, continuo a tentar outras maneiras de as fazer até conseguir o mesmo resultado de antes.

B - **(Cotação 0)**

Quando as coisas não correm tão bem como antes, aceito esse facto.

Seleção de variáveis

Motivação

Nas áreas das Ciências Sociais e Humanas encontramos frequentemente problemas de dimensionalidade, pois, mesmo num fenómeno descrito por um número moderado de variáveis, o número de resultados possíveis é muito grande quando comparado com o número de participantes.

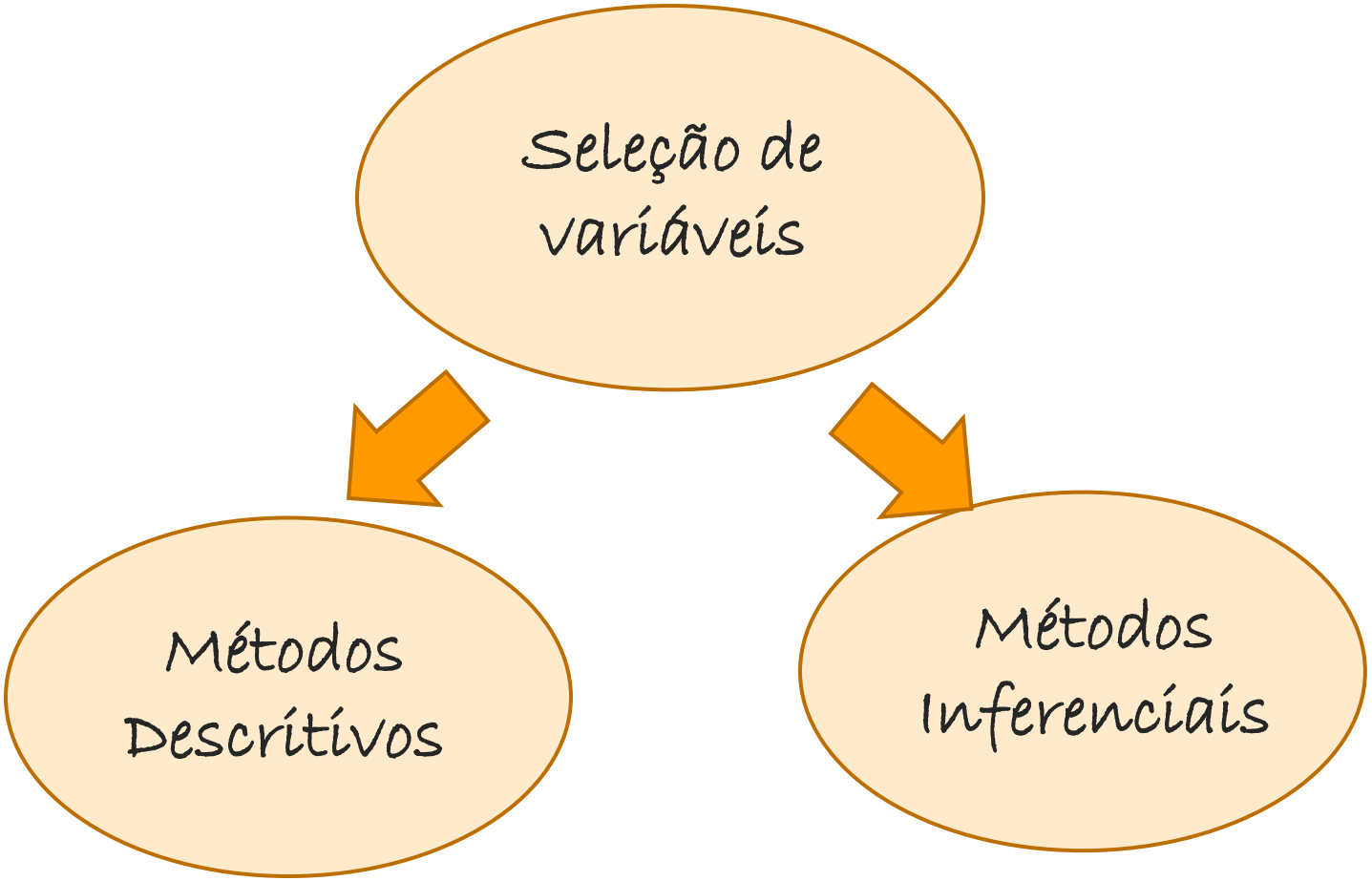
Neste contexto, **a seleção das variáveis** mais discriminativas num problema de classificação discreto tem, pois, um **importante papel** no desempenho de qualquer modelo.

Seleção de variáveis

Em Análise Discriminante Discreta (ADD), o problema da dimensionalidade conduz, frequentemente, a desempenhos fracos dos modelos.

Objetivo: Selecionar de entre as P variáveis qualitativas que caracterizam o fenómeno em estudo, as P* variáveis, com P*<<P, que conduzam a uma "boa" regra de decisão.

Seleção de variáveis



Estatística Qui-Quadrado (QQ)

Para estudar a associação entre cada variável explicativa qualitativa e as classes definidas *a priori, determina-se o valor da* Estatística Qui-quadrado *associada a cada uma das* tabelas de contingência que cruzam cada variável explicativa com as classes definidas *a priori* e selecionam-se as P* variáveis que conduzem aos maiores valores observado da Estatística Qui-quadrado.

Informação Mútua (IM)

Como é sabido, a informação mútua entre duas variáveis aleatórias *X* e *Y* *indica, de certa forma, a redução no nível de* incerteza associado a uma variável graças à informação trazida pela outra variável. Determinando a IM entre cada variável explicativa e a variável que define as classes a *priori selecionam-se as* P* variáveis que conduzem aos maiores valores observado da Informação Mútua.

Taxa de Erro da Família de Testes (TEFT)

Considerando a família de m Testes de Independência do Qui-quadrado entre cada variável explicativa e as classes definidas *a priori* e os respetivos valores–p $p_{1:m}$, $p_{2:m}$, ... $p_{K:m}$, a correção de Bonferroni (Benjamini e Hochberg,1995) controla a TEFT. São, então selecionadas as K variáveis tais que $p_{K:m} < \alpha/m$.

Taxa de Falsas Descobertas (TFD)

Considere-se a mesma família de *m* Testes de Independência do Qui-quadrado e os respetivos valores–p $p_{1:m}$, $p_{2:m}$, ... $p_{K:m}$. A TFD (Benjamini e Hochberg,1995; Duarte Silva, 2010) consiste no valor médio da proporção de hipóteses nulas verdadeiras que serão consideradas falsas, entre as hipóteses nulas rejeitadas e são, então, selecionadas as K variáveis tais que $p_{K:m} < (K/m) * \alpha$

Análise Discriminante Discreta

O objetivo principal da ADD é discriminar as classes de uma partição definida *a priori*, construindo uma regra de decisão que permita, no futuro, classificar novos indivíduos (*indivíduos anónimos*), minimizando os erros de classificação.

Modelos de referência

Modelo Multinomial Completo (MMC)

$$\hat{P}_M(x \mid c_k) = \frac{N(x \mid k)}{n_k}$$

Desvantagem: exige, para cada classe a priori, a estimação de um grande número de parâmetros (se as p variáveis forem binárias - 2^p-1)

Modelo de Independência Condicional de ordem um (MIC)

$$\hat{P}_I(x \mid c_k) = \prod_{j=1}^{p} \frac{N(x_j \mid k)}{n_k}$$

Desvantagem: propõe uma grande redução dos parâmetros a estimar mas em certos casos é demasiado simplificativo

Modelo Gráfico Decomponível (MGD)

Considera as interações mais importantes entre pares de variáveis para estimar a função de probabilidade por grupo, utilizando uma estrutura de árvore (grafo), que se baseia na informação mútua. O algoritmo considerado foi o proposto por Chow e Liu (1968).

Isto é, se tivermos P=5 variáveis explicativas e se as interações mais importantes, no grupo k, forem (x_2, x_1), (x_3, x_2), (x_4, x_2) e (x_5, x_2) teremos

$$\hat{P}\left(X \middle| C_k\right) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)P(X_4 \mid X_2)P(X_5 \mid X_2)$$

Combinação de Modelos em ADD

Como já referido em ADD existem frequentemente problemas de dimensionalidade e tendo observado que quando se comparam os erros de classificação obtidos por diversos modelos, se verifica, frequentemente, que eles não ocorrem sobre os mesmos objetos propusemos duas abordagens de combinação de modelos:

Combinação linear entre os modelos MMC e MIC (Sousa Ferreira, 2000):

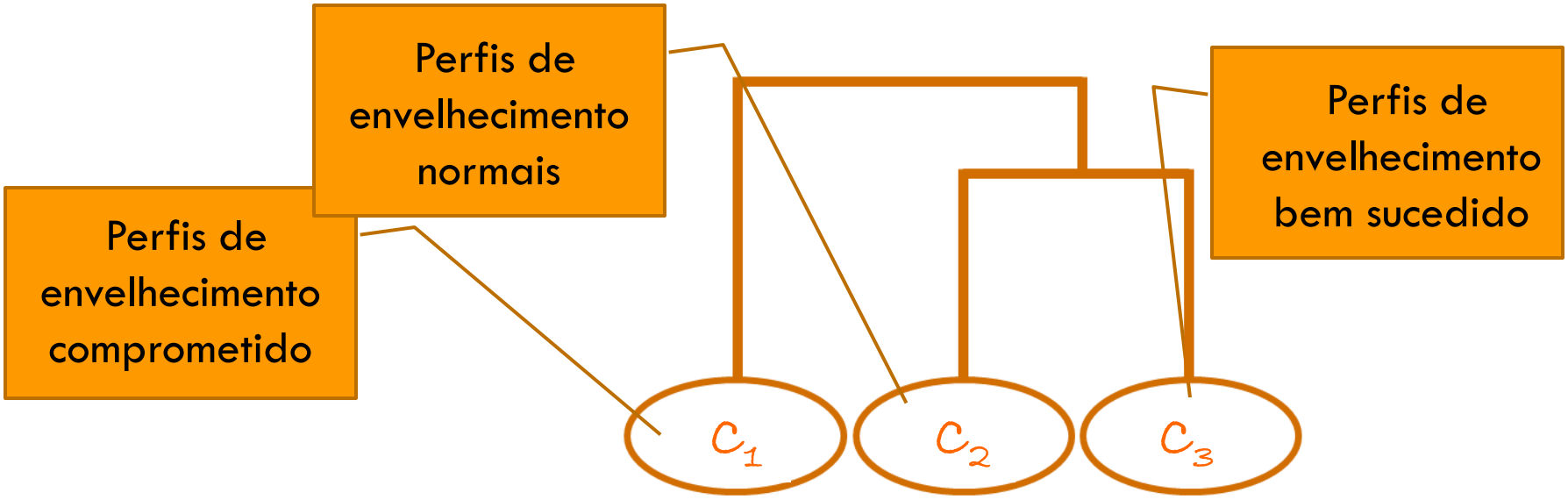
$$\hat{P}_k(x \mid \beta) = (1-\beta)\,\hat{P}_{MMC}(x \mid c_k) + \beta\,\hat{P}_{MIC}(x \mid c_k), \quad 0 \le \beta \le 1$$

Combinação linear entre os modelos MGD e MIC (Marques, 2008):

$$\hat{P}(\underline{x}|\beta) = \beta\,\hat{P}_{MIC}(\underline{x}|\beta) + (1-\beta)\,\hat{P}_{MGD}(\underline{x}|\beta), \quad 0 \le \beta \le 1$$

Combinação de Modelos em ADD

No caso de mais duas classes *a priori* propusemos o Modelo de Emparelhamento Hierárquico (MHIERM) que decompõe um problema com múltiplas classes *a priori* em diversos problemas de dois grupos *a priori*, utilizando uma estrutura de árvore binária.



Análise dos resultados

Seleção das variáveis

Método	Itens selecionados
Estatística Qui-quadrado	SOC_9, SOC_11, SOC_2, SOC_10 e SOC_4
Informação Mútua	SOC_9, SOC_11, SOC_2, SOC_4, SOC_10, SOC6, SOC_3, SOC_8 , SOC_6 e SOC_5
Taxa de Erro da Família de Testes	SOC_9, SOC_11e SOC_2 (α=.15)
Taxa de Falsas Descobertas	SOC_9, SOC_11, SOC_2, SOC_10 e SOC_4 (α=.15)

Análise dos resultados

Comparação dos modelos

Percentagens de bem classificados por método e nº de variáveis (estimadas por 2-fold)

Nº de classes	Nº de variáveis	Métodos				
		MMC	MGD	MIC	MMC-MIC	MGD-MIC
3	10	30.6	39.0	46.8	59.4	**60.3**
3	5	42.3	39.0	43.6	**65.6**	61.0
2	10	52.6	60.3	61.0	58.4	**64.3**
2	5	55.8	61.0	57.8	**66.8**	64.9

Note-se que, no caso dos modelos combinados se consideraram diversos valores do coeficiente β da combinação, estimados por estratégias de regressão ou numa grelha com valores no intervalo (0,1).

Conclusões

Os estudos de seleção de variáveis em ADD que já realizámos apontam para que, em pequenas amostras os métodos inferenciais poderão ter dificuldades em selecionar variáveis. No entanto, poderão ter um desempenho interessante na seleção de variáveis em amostras de dimensão moderada.

Neste estudo é relevante notar que em situações muito desvantajosas (por exemplo 1024 estados e n=154), a seleção de variáveis permite reduzir drasticamente os tempos de execução e melhorar o desempenho dos modelos.

A combinação de modelos mostra a sua capacidade de incrementar a capacidade preditiva dos modelos, particularmente quando usa o Modelo de Emparelhamento Hierárquico no caso de mais de dois grupos.

Referências

[1] Benjamini, Y., Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological), *57*, 289-300.

[2] Brito, I., Celeux, G., Sousa Ferreira, A (2006). Combining Methods in Supervised Classication: A Comparative Study on Discrete and Continuous Problems. RevSTAT-Statistical Journal, *4*(3), 201-225.

[3] Gaudêncio, J., Duarte Silva, M.E., Doria, I. (2014). Perfis de envelhecimento em idosos com idade avançada: Resultados de um estudo no sotavento algarvio. Em Anica, A. et al. Editores, Envelhecimento Ativo e Educação, Edição da Universidade do Algarve, 41-53.

[4] Marques, A. (2014). Análise discriminante sobre variáveis qualitativas. Tese de Doutoramento, ISCTE-Instituto Universitário de Lisboa.

[5] Marques, A., Sousa Ferreira, A., Cardoso, M. (2008). Uma proposta de combinação de modelos em Análise Discriminante. Actas do XVI Congresso Anual da Sociedade Portuguesa de Estatística, 393-403.

[6] Marques, A., Sousa Ferreira, A., Cardoso, M. (2013). Selection of variables in Discrete Discriminant Analysis. Biometrical Letters, 50(1), 1-14.

[7] Sousa Ferreira, A. (2001). Combinação de modelos em Análise Discriminante sobre Variáveis Qualitativas. Tese de Doutoramento, FCT – Universidade Nova de Lisboa.