Probabilidades e Estatística

Editor Convidado: Nelson Antunes

Funcionais aditivos de processos de exclusão

Patrícia Gonçalves
Centro de Matemática da Universidade do Minho
Campus de Gualtar
4710-057 Braga, Portugal
e-mail: patg@math.uminho.pt

Resumo: Neste trabalho consideram-se processos de exclusão partindo da medida de Bernoulli produto. Obtém-se o limite em escala de funcionais aditivos, como por exemplo do tempo de ocupação, a partir das flutuações da densidade.

Abstract In this work we consider exclusion processes starting from the Bernoulli product measure. We obtain the scaling limits of additive functionals, as for example the occupation time, from the density fluctuations.

palavras-chave: Processos de exclusão; funcionais aditivos.

keywords: Exclusion processes; additive functionals.

Neste trabalho pretende-se explorar a seguinte questão: dado um processo de Markov a tempo contínuo $(X_t)_{t\geq 0}$ com espaço de estados compacto $\mathcal{E}$ e uma função $f : \mathcal{E} \to \mathbb{R}$ qual o limite em escada do funcional aditivo

$$\int_0^t f(X_s)ds \ ?$$

O processo de Markov que se considera é o *processo de exclusão*. Vai-se definir a evolução do processo em $\mathbb{Z}$ e o seu espaço de estados é $\mathcal{E} = \{0,1\}^{\mathbb{Z}}$. A dinâmica deste processo pode ser descrita da seguinte forma. Em cada sítio $x \in \mathbb{Z}$, existe um relógio aleatório com distribuição exponencial de parâmetro 1 e independente dos relógios de outros sítios. Cada estado deste processo é uma configuração $X \in \{0,1\}^{\mathbb{Z}}$, i.e. $X = (\cdots, x_{-1}, x_0, x_1, \cdots)$ é um vetor com infinitas coordenadas onde $x_i \in \{0,1\}$ para todo $i \in \mathbb{Z}$. A interpretação física é a seguinte, se para $x \in \mathbb{Z}$, $X(x) = 1$ então o sítio x está ocupado com uma partícula, se $X(x) = 0$ então o sítio x está vazio. Aqui $X(x)$ representa a entrada do vetor correspondente ao sítio x. Cada partícula movimenta-se como um passeio aleatório, mas sujeito a restrições locais devido ao movimento das restantes partículas. Quando o relógio em x toca, pode acontecer o seguinte. Se não existe partícula em x, então nada acontece e os relógios iniciam uma nova contagem; se existe uma partícula em x ela decide saltar para $x+1$ ou $x-1$ de acordo com $r(\cdot)$ definida abaixo;

mas o salto só ocorre se o sítio está vazio, caso contrário nada acontece - daqui provém o nome processo de exclusão.

Seja $r : \{0,1\}^{\mathbb{Z}} \to \mathbb{R}$ uma função que satisfaz as seguintes condições:

i) Existe $\varepsilon_0 > 0$ tal que $\varepsilon_0 < r(X) < \varepsilon_0^{-1}$ para qualquer $X \in \{0,1\}^{\mathbb{Z}}$. (**Elipticidade**)

ii) Para qualquer $X, Y \in \{0,1\}^{\mathbb{Z}}$, tal que $X(x) = Y(x)$ para $x \neq 0, 1$, então $r(X) = r(Y)$. (**Reversibilidade**)

iii) A função $r(\cdot)$ depende do vetor X apenas num número finito de coordenadas, ou seja, $r(\cdot)$ é uma função *local*.

A dinâmica descrita acima pode ser definida formalmente através de um generador, cujo cerne é o conjunto das funções locais e tem a expressão:

$$\mathcal{L}f(X) = \sum_{x\in\mathbb{Z}} r(\tau_x X)(f(X^{x,x+1}) - f(X)),$$

onde: $X^{x,x+1}$ é o vetor obtido de X trocando o valor de $X(x)$ pelo valor de $X(x+1)$; τ_x representa a translação espacial por x, ou seja $y \in \mathbb{Z}$ $\tau_x X(y) := X(x+y)$ e $f : \{0,1\}^{\mathbb{Z}} \to \mathbb{R}$ é uma função local.

As suas medidas invariantes denotam-se por $\{\nu_\rho : \rho \in [0,1]\}$, onde ν_ρ é a medida de Bernoulli produto com parâmetro $\rho \in [0,1]$ e definida em $\{0,1\}^{\mathbb{Z}}$. Como consequência, sob estas medidas, as variáveis de ocupação $\{X(x) : x \in \mathbb{Z}\}$ são independentes e $\nu_\rho(X : X(x) = 1) = \rho$.

Um dos problemas centrais no estudo deste processo consiste em determinar a sua *equação hidrodinâmica*. A equação hidrodinâmica é uma equação diferencial parcial que descreve a evolução espaço-temporal da densidade de partículas. Este problema é conhecido na literatura por *limite hidrodinâmico*. Outro problema central, consiste em determinar a equação diferencial parcial estocástica que rege as flutuações da densidade. Para enunciar este último resultado, define-se o campo de flutuações por $\mathcal{Y}_t^n(g) := \frac{1}{\sqrt{n}} \sum_{x\in\mathbb{Z}} g\Big(\frac{x}{n}\Big)\Big(X_{tn^2}(x) - \rho\Big)$, onde $g \in \mathbb{S}(\mathbb{R})$ e $\mathbb{S}(\mathbb{R})$ denota o espaço de Schwarz. O dual topológico de $\mathbb{S}(\mathbb{R})$ com respeito ao produto interno de $L^2(\mathbb{R})$ denota-se por $\mathbb{S}'(\mathbb{R})$.

Foi provado em [1], que $\{\mathcal{Y}_t^n : t \in [0,T]\}$ converge em distribuição, com respeito à topologia de Skorohod de $\mathbb{D}([0,T], \mathbb{S}'(\mathbb{R}))$ (o espaço de trajetórias contínuas à direita e com limite à esquerda que tomam valores em $\mathbb{S}'(\mathbb{R})$), quando $n \to +\infty$, para a solução estacionária da equação de *Ornstein-Uhlenbeck* dada por

$$d\mathcal{Y}_t = D(\rho)\partial_x^2 \mathcal{Y}_t dt + \sqrt{2D(\rho)\rho(1-\rho)}\partial_x d\mathcal{B}_t, \tag{1}$$

onde $\mathcal{B}_t$ é o movimento Browniano com valores em $\mathbb{S}'(\mathbb{R})$ e $D(\rho)$ é o chamado *coeficiente de difusão*. Em particular, as trajetórias de $\mathcal{Y}_t$ são contínuas e $\mathcal{Y}_0$ é um campo Gaussiano com variância $\rho(1-\rho)$.

O primeiro resultado que se segue está relacionado com a convergência do funcional aditivo de $\mathcal{Y}_t$. Fixe uma solução estacionária $\{\mathcal{Y}_t : t \in [0,T]\}$ de (1). Para $x \in \mathbb{R}$, seja $i_\varepsilon(x) : y \mapsto \varepsilon^{-1} 1_{(0,1]}((y-x)\varepsilon^{-1})$. Para cada $\varepsilon \in (0,1)$, seja $\{\mathcal{Z}_t^\varepsilon : t \in [0,T]\}$ definido por

$$\mathcal{Z}_t^\varepsilon = \int_0^t \mathcal{Y}_s(i_\varepsilon) ds.$$

Então, $\{\mathcal{Z}_t^\varepsilon : t \in [0,T]\}$ converge em distribuição com respeito à topologia uniforme de $\mathbb{C}([0,T],\mathbb{R})$, quando $\varepsilon \to 0$, para o *movimento Browniano fracionário* $\{\mathcal{Z}_t : t \in [0,T]\}$ com expoente de Hurst $H = 3/4$. Note que $\mathbb{C}([0,T],\mathbb{R})$ denota o espaço de trajetórias contínuas que tomam valores em $\mathbb{R}$.

No segundo resultado que se segue, obtém-se o limite em escala de observáveis do processo de Markov $(X_t)_{t\geq 0}$. Para $f : \{0,1\}^{\mathbb{Z}} \to \mathbb{R}$ local, o processo $\{\Gamma_{tn^2}(f) : t \in [0,T]\}$ definido por

$$\Gamma_{tn^2}(f) = \frac{1}{n^{3/2}} \int_0^{tn^2} (f(X_s) - \varphi_f(\rho)) ds$$

converge em distribuição com respeito à topologia uniforme de $\mathbb{C}([0,T],\mathbb{R})$, quando $n \to +\infty$, para o processo $\{\varphi_f'(\rho)\mathcal{Z}_t : t \in [0,T]\}$, onde $\{\mathcal{Z}_t : t \in [0,T]\}$ é o processo obtido no limite de $\{\mathcal{Z}_t^\varepsilon : t \in [0,T]\}$ quando $\varepsilon \to 0$. Aqui $\varphi_f(\rho) := E_{\nu_\rho}[f(\eta)]$.

Note-se que as funções locais definidas em $\{0,1\}^{\mathbb{Z}}$ podem ser classificadas pelo seu grau. Diz-se que uma função local tem grau n se $\varphi_f^j(\tilde{\rho})\Big|_{\tilde{\rho}=\rho} = 0$, para $j = 0, \cdots, n-1$ e $\varphi_f^n(\tilde{\rho})\Big|_{\tilde{\rho}=\rho} \neq 0$. Esta condição é equivalente a dizer que $f(\cdot)$ pode ser escrita como $f(X) = c\prod_{x\in A}(X(x)-\rho)$, onde c é uma constante e $A \subseteqq \mathbb{Z}$ com $|A| = n$.

Sendo assim, o resultado previamente enunciado, refere que o funcional aditivo para funções de grau 1 converge e o seu limite é identificado. No entanto, para funções com grau $n \geq 2$, o resultado acima refere que o limite é zero. Considerando $f(X) = X(0)$, o funcional $\Gamma_t(f)$ corresponde ao *tempo de ocupação da origem* durante o intervalo de tempo $[0,t]$.

É oportuno referir que o primeiro resultado enunciado acima não se restringe a (1). De facto, considerando a equação de *Burgers estocástica* dada por:

$$d\mathcal{Y}_t = \frac{1}{2}\partial_x^2 \mathcal{Y}_t dt + \left(\partial_x \mathcal{Y}_t\right)^2 dt + \sqrt{\rho(1-\rho)}\partial_x d\mathcal{B}_t \tag{2}$$

pode-se provar o seguinte resultado. Seja $\{\mathcal{Y}_t : t \in [0,T]\}$ uma solução estacionária de (2). Para $\varepsilon > 0$, seja $\tilde{\mathcal{Z}}_t^\varepsilon = \int_0^t \mathcal{Y}_s(i_\varepsilon)ds$. Então, existe $\{\tilde{\mathcal{Z}}_t : t \in [0,T]\}$ tal que, $\{\tilde{\mathcal{Z}}_t^\varepsilon : t \in [0,T]\}$ converge em distribuição com respeito à topologia uniforme de $\mathbb{C}([0,T],\mathbb{R})$, quando $\varepsilon \to 0$, para $\{\tilde{\mathcal{Z}}_t : t \in [0,T]\}$.

Em [4] apresenta-se a prova dos resultados mencionados acima e em [3] considera-se uma classe geral de modelos partindo da medida de Bernoulli produto, cujas flutuações são regidas por (2). A prova do resultado sobre os funcionais aditivos, assenta no chamado *Príncipio de Boltzmann-Gibbs*. Em [4] prova-se esse Princípio através de um argumento multi-escala introduzido em [2]. Em [5] estende-se esse resultado para dinâmicas e medidas iniciais muito mais gerais.

O argumento multi-escala assenta na seguinte ideia. Primeiro, substitui-se o funcional aditivo de $f(\cdot)$ pelo funcional aditivo da esperança condicional de $f(\cdot)$, com respeito à quantidade de partículas num intervalo com tamanho ℓ. De seguida, substitui-se esse funcional pelo funcional da esperança condicional num intervalo com tamanho 2ℓ e repete-se o argumento. Finalmente, ao fim de m passos, quando o intervalo tem tamanho $2^m\ell$, substitui-se esse funcional pelo funcional da densidade de partículas. Este Princípio é um dos grandes desafios no estudo do comportamento deste tipo de processos.

Referências

[1] C. Chang, "Equilibrium fluctuations of gradient reversible particle systems", Probab. Theory Relat. Fields, Vol. 100, No. 3 (1994), pp. 269–283.

[2] P. Gonçalves, "Central limit theorem for a tagged particle in asymmetric simple exclusion", Stoch. Proc. Appl., Vol. 118, (2008), pp. 474–502.

[3] P. Gonçalves e M. Jara, "Universality of KPZ equation", disponível em arXiv.org.

[4] P. Gonçalves e M. Jara, "Scaling limits of additive functionals of interacting particle systems", accepted for publication in Comm. Pure Appl. Math.

[5] P. Gonçalves, M. Jara e S. Sethuraman, "A stochastic Burgers equation from a class of microscopic interactions", disponível em arXiv.org.

O perfil dos campeões de jogos matemáticos

Paulo Infante, *Anabela Afonso*

Centro de Investigação em Matemática e Aplicações e
Departamento de Matemática, ECT - Universidade de Évora
Rua Romão Ramalho, 59
7000-671 Évora, Portugal
e-mail: pinfante@uevora.pt
aafonso@uevora.pt

Jorge Nuno Silva

Centro de História das Ciências da Universidade de Lisboa e
Secção Autónoma de História e Filosofia das Ciências (SAHFC), FCUL
Campo Grande, C4, Gab. 4.3.18
1749-016 Lisboa, Portugal
e-mail: jnsilva@cal.berkeley.edu

Resumo: Neste trabalho apresentamos os primeiros resultados de um inquérito realizado no último Campeonato Nacional de Jogos Matemáticos. Começamos por caracterizar todos os alunos que disputaram o campeonato e posteriormente os vencedores dos vários torneios. Avaliamos a significância estatística de associações entre algumas variáveis, por exemplo, até que ponto o desempenho no campeonato está associado com o desempenho a Matemática e o Português. Finalmente, traçamos o perfil mais provável dos vencedores dos torneios em função das disciplinas preferidas, da frequência de prática de jogos matemáticos e outros, do interesse por jogos matemáticos, do desempenho escolar, do sexo e das habilitações literárias dos pais.

Abstract In this paper we present the first results obtained with a questionnaire applied in the last National Championship of Mathematical Games. We begin with a brief characterization of all players of this championship and after the the winners in the several turnings. We evaluate the statistical significance of associations between some variables, for example, is the performance in the championship related with Mathematic and Portuguese. Finally, we draw the most probable profile of the turning winners according to their preferred subjects, how many times they play mathematical and other games, interest in mathematical games, scholarship performance, gender and parents qualifications.

palavras-chave: Jogos matemáticos; regressão logística; testes não paramétricos.

keywords: Logistic regression; mathematical games; non parametric tests.

1 Introdução

O Campeonato Nacional de Jogos Matemáticos (CNJM) tem sido promovido, desde a sua primeira edição em 2004, pela Associação LUDUS, pela Sociedade Portuguesa de Matemática (SPM), pela Associação de Professores de Matemática (APM) e pela Ciência Viva. Em Março de 2012 disputou-se em Coimbra a 8ª edição, onde participaram quase 2100 alunos dos 3 ciclos do ensino básico e ensino secundário, pertencentes a 526 escolas de todo o país. Um número tão elevado de participantes é ainda mais significativo porque cada escola apenas pode inscrever um aluno por jogo e por nível de ensino: Semáforo (1º ciclo), Cães e Gatos (1º e 2º ciclos), Ouri (1º, 2º e 3º ciclos), Hex (2º e 3º ciclos e secundário), Rastros (3º ciclo e secundário) e Avanço (secundário). A adesão de alunos ao CNJM ultrapassou todas as espectativas, estimando-se que nas escolas o número de alunos jogadores que disputam as eliminatórias de apuramento para a final esteja próximo dos 100000, número muito superior ao de praticantes de qualquer outra atividade lúdica e desportiva em Portugal.

Nesta edição foi realizado um inquérito por questionário em papel distribuído aos alunos jogadores, com o objectivo de identificar quem joga e o quê, há quanto tempo, como e quando joga, qual o desempenho escolar, quais as variáveis que influenciam o desempenho nos jogos e a influência dos jogos na sua formação escolar e pessoal. Obtiveram-se 1148 questionários válidos (cerca de 56% dos jogadores presentes no CNJM). Das 12 questões, a que apresentou taxa de resposta mais baixa (84%) foi a relativa à profissão pretendida para o futuro, e em todas as outras a taxa de resposta foi de pelo menos 90%. Confrontando a distribuição das respostas obtidas por ciclo e jogo obtivemos uma distribuição semelhante à dos alunos presentes no campeonato, o que indicia uma boa representatividade da amostra obtida. De seguida apresentam-se e discutem-se os primeiros resultados alcançados.

2 Breve caracterização dos jogadores presentes

De um modo geral os alunos tiveram nota positiva nas disciplinas de Matemática (96%), Português (96%) e Educação Física (99%). A nota dominante a Matemática é o 5 (42%) e a Português o 4 (39%). As raparigas apresentam notas a Português significativamente superiores às dos rapazes (*Wilcoxon-Mann-Whitney*, $z = 4,219$, $p < 0,001$). A nota a Matemática difere entre ciclos (*Kruskal-Wallis*, $\chi^2_3 = 54,7$, $p < 0,001$) tal como a nota a Português (*Kruskal-Wallis*, $\chi^2_3 = 109,7$, $p < 0,001$) ambas decrescendo com o aumento do nível de escolaridade.

Há diferença significativa entre sexos tanto na disciplina que mais gostam

($\chi^2_8 = 37,7$, $p < 0,001$) como na segunda disciplina preferida ($\chi^2_8 = 35,9$, $p < 0,001$). A Matemática foi a disciplina preferida mais mencionada por ambos os sexos. Nos rapazes a segunda referência recai sem qualquer dúvida sobre a Educação Física, ao passo que nas raparigas divide-se entre a Educação Física e as Ciências da Natureza/Estudo do Meio. A preferência pela Matemática diminui gradualmente com o aumento do nível de ensino, sendo a Educação Física a mais preferida pelos alunos do secundário.

Não foi detetada relação significativa entre a frequência de jogo de jogos de tabuleiro por ciclo ($\chi^2_6 = 10,1$, $p = 0,112$), mas sim para todos os outros jogos (puzzle: $\chi^2_6 = 21,2$, $p = 0,002$; consola: $\chi^2_6 = 13,0$, $p = 0,044$; cartas: $\chi^2_6 = 30,4$, $p < 0,001$; *online*: $\chi^2_6 = 15,2$, $p = 0,019$). Os alunos do 1º ciclo são os que jogam mais com puzzles e menos *online*, os alunos do secundário jogam menos consola que os restantes, sendo que a frequência de jogo de cartas aumenta com o ciclo. Não foi detetada relação significativa entre a frequência de jogo de jogos de tabuleiros e o sexo dos jogadores ($\chi^2_2 = 0,2$, $p = 0,887$). Nos restantes jogos a associação é bastante significativa (todos $p < 0,001$). Nos jogos puzzle, *online* e consola a percentagem de jogadores frequentes do sexo masculino é superior ao sexo feminino, verificando-se o inverso nos jogos de cartas. Há diferenças significativas na distribuição do número de vezes que o aluno pratica jogos matemáticos semanalmente entre ciclos ($\chi^2_3 = 151,3$, $p < 0,001$). Os alunos do 1º ciclo são os que treinam mais e os alunos do ensino secundário os que treinam menos.

3 Perfil dos finalistas: campeões

Para traçar o perfil dos alunos vencedores dos torneios recorreu-se a um modelo linear generalizado com componente aleatória com distribuição binomial e componente estrutural com a função de ligação canónica (*logit*), conhecido por modelo de regressão logística [1, 2]. A variável dicotómica a explicar foi codificada como 1 se venceu o torneio, e 0 c.c. e o critério de seleção das variáveis foi o apresentado em [2].

Após uma análise de resíduos para avaliação de *outliers*, observações influentes e observações com repercussão elevada, constatou-se o bom ajuste do modelo obtido (tabela 1) aos dados através do teste de Hosmer e Lemeshow ($\chi^2_8 = 4,41$; $p = 0,82$), podendo também concluir-se que o modelo final tem uma capacidade discriminativa aceitável ($AUC = 0,73$; $IC_{95\%} =]0,67; 0,78[$), tendo sido também feita uma validação por *bootstrap*.

Admitindo fixas as restantes covariáveis do modelo, pelo cálculo dos *odds ratio* do modelo ([2]) podemos retirar as seguintes conclusões: um jogador cujo pai ou mãe tem habilitações ao nível do ensino superior tem quase o dobro de possibilidades de vencer o torneio relativamente a um

Tabela 1: Coeficientes estimados ($\hat{\beta}$), desvios-padrão estimados ($\hat{\sigma}_{\hat{\beta}}$) e *valores p (teste de Wald)* para o modelo de regressão logística estimado para o vencedor dos torneios (ref = classe de referência).

Variáveis	$\hat{\beta}$	$\hat{\sigma}_{\hat{\beta}}$	p
Sexo (ref = masculino)	-1,006	0,458	0,028
Hab. pais (ref = sem ensino superior)	0,647	0,252	0,010
Disc. pref. (ref = todas / Mat. e Ed. Física)	0,566	0,279	0,043
Cartas raram. (ref = nunca)	-0,683	0,441	0,122
Cartas freq. (ref = nunca)	-0,906	0,457	0,047
Anos int. (ref = no máximo 1)	1,639	0,481	<0,001
Jogos mat. 1 ou 2× (ref = nenhuma)	-0,681	0,335	0,042
Jogos mat. >2× (ref = nenhuma)	-0,236	0,360	0,512
Consola (ref = nunca ou raramente)	-0,508	0,282	0,072
Consola*Sexo	0,839	0,680	0,217
Constante	-2,548	0,633	<0,001

cujos pais tenham habilitações inferiores ($IC_{90\%} =]1,3; 2,9[$); um que tenha como disciplinas preferidas Matemática e Educação Física tem mais de 75% de possibilidades relativamente a um que tenha outras preferências ($IC_{90\%} =]1,1; 2,8[$); um que nunca jogue cartas tem 2,5 vezes mais possibilidades relativamente a um que jogue frequentemente ($IC_{90\%} =]1,2; 5,2[$); um com mais de 1 ano de interesse por jogos matemáticos tem 5 vezes mais possibilidades relativamente a um que tenha um interesse muito recente ($IC_{90\%} =]2,3; 11,4[$); para os que não jogam consola, um rapaz tem quase 3 vezes mais possibilidades relativamente a uma rapariga ($IC_{90\%} =]1,3; 5,8[$); um rapaz que não jogue ou jogue raramente consola tem 2/3 mais possibilidades do que um que jogue frequentemente ($IC_{90\%} =]1,1; 2,6[$).

4 Conclusão

O perfil mais provável de um aluno que venceu um torneio é ser rapaz cujo o pai ou a mãe tem ensino superior, que se interessa por jogos matemáticos há mais de um ano, mas que não os pratica uma ou duas vezes por semana, que nunca joga cartas e que gosta de Matemática e Educação Física.

Referências

[1] A. Agresti, *An Introduction to Categorical Data Analysis*, Second Edition, Wiley-Interscience, New Jersey, 2007.

[2] D. Hosmer e S. Lemeshow, *Applied Logistic Regression*, Second Edition, John Wiley & Sons, New York, 2000.

Aplicação do coeficiente de afinidade a dados de natureza complexa

Áurea Sousa, Osvaldo Silva
Universidade dos Açores, Dep. de Matemática, CEEAplA, e CMATI
Rua da Mãe de Deus - Apartado 1422
9501-801 Ponta Delgada, Portugal
e-mail: aurea@uac.pt; osilva@uac.pt

Helena Bacelar-Nicolau
Universidade de Lisboa, Faculdade de Psicologia, LEAD, e DataScience
Alameda da Universidade
1649-013 Lisboa, Portugal
e-mail: hbacelar@fp.ul.pt

Fernando C. Nicolau
Universidade Nova de Lisboa, FCT, Dep. de Matemática, e DataScience
Quinta da Torre
2829-516 Caparica, Portugal
e-mail: geral@datascience.org

Resumo: É ilustrada a aplicação da Análise Classificatória Hierárquica Ascendente a dados de natureza complexa, com base no coeficiente de afinidade generalizado ponderado e em critérios de agregação clássicos e probabilísticos, estes últimos no âmbito da metodologia *VL*.

Abstract: We illustrate an application of Ascendant Hierarchical Cluster Analysis to complex data, based on the weighted generalized affinity coefficient. Classical and probabilistic aggregation criteria are used, the probabilistic ones in the scope of the *VL* methodology

palavras-chave: Análise classificatória hierárquica; dados simbólicos; coeficiente de afinidade generalizado ponderado; metodologia *VL*.

keywords: Cluster analysis; symbolic data; weighted generalised affinity coefficient; *VL* methodology.

1 Introdução

A existência de bases de dados de elevada dimensão levou à necessidade de sumariar esses dados em termos dos seus conceitos mais relevantes, os quais podem ser descritos por tipos de dados complexos, também designados por

dados simbólicos. Na tabela de dados, as linhas representam unidades de dados ou objetos simbólicos e as colunas variáveis simbólicas, respetivamente, podendo as células da tabela conter, por exemplo, subconjuntos de números reais, distribuições de frequências, ou intervalos da reta real, em vez de um único valor, como usualmente (Bock e Diday, 2000).

Na Secção 2 é indicada a fórmula do coeficiente de afinidade generalizado ponderado e referido um coeficiente probabilístico a este associado, para o caso em que as unidades de dados são descritas por variáveis cujos valores são intervalos da reta real. Na Secção 3, apresentam-se os principais resultados obtidos com a aplicação da Análise Classificatária Hierárquica Ascendente (ACHA) a dados de natureza complexa, com base naqueles coeficientes e em critérios de agregação clássicos e probabilísticos. Finalmente, a Secção 4 contém algumas considerações sobre o trabalho desenvolvido.

2 Coeficiente de afinidade generalizado ponderado: Caso de variáveis de intervalo

A partir do coeficiente de afinidade entre duas distribuições de probabilidade discretas, proposto por Matusita, Bacelar-Nicolau (1980, 1988) introduziu o coeficiente de afinidade na análise classificatória, para avaliar a semelhança básica entre pares de colunas ou pares de linhas de uma matriz de dados, ou seja, entre variáveis ou entre indivíduos. Posteriormente, estendeu o coeficiente a diferentes tipos de dados, incluindo os heterogéneos e de natureza complexa (Bacelar-Nicolau, 2000; Bacelar-Nicolau et al., 2009, 2010).

Seja $E = \{1, ..., N\}$ um conjunto de unidades de dados descritas por p variáveis, $Y_1, ..., Y_p$, cujos valores são intervalos da reta real. Seja Y_j, com $j \in \{1, \ldots, p\}$ a variável associada à coluna j, na qual cada célula (k, j) contém um intervalo I_{kj}, com $k = 1, ..., N$. Neste caso, Bacelar-Nicolau definiu o coeficiente de afinidade generalizado ponderado entre k e k' como:

$$a(k,k') = \sum_{j=1}^{p} \pi_j \, aff\,(I_{kj}, I_{k'j}) = \sum_{j=1}^{p} \pi_j \, aff\,(k,k';j) = \sum_{j=1}^{p} \pi_j \frac{|I_{kj} \bigcap I_{k'j}|}{\sqrt{|I_{kj}|.|I_{k'j}|}},$$

onde $aff\,(k, k'; j)$ designa a afinidade "parcial" relativa à j-ésima variável e os pesos π_j verificam $0 \leq \pi_j \leq 1$, $\sum \pi_j = 1$. Este coeficiente toma valores no intervalo [0,1] e satisfaz um conjunto de propriedades que o tornam um coeficiente de semelhança robusto. O coeficiente assintoticamente centrado e reduzido, sob uma hipótese de referência permutacional baseada no teorema limite de Wald e Wolfowitz, e denotado por $a_{WW}(k, k')$, foi analisado em Bacelar-Nicolau et al. (2010). O coeficiente $a_{WW}(k, k')$ permite, por sua

vez, definir um coeficiente probabilístico no contexto da metodologia *VL*, na linha iniciada por Lerman (1972) e desenvolvida por Bacelar-Nicolau (e.g. 1980, 1988) e Nicolau (e.g. 1983, 1998). Aplicações desta metodologia foram apresentadas, por exemplo, em Bacelar-Nicolau et al. (2009, 2010).

3 Aplicação a dados: temperaturas das cidades

Considera-se as temperaturas mínimas e máximas, em graus centígrados, registadas em 37 cidades (Guru et al., 2004), durante um ano. A partição intuitiva, efetuada por um grupo de observadores humanos, resultou em quatro classes de cidades: $\{2, 3, 4, 5, 6, 8, 11, 12, 15, 17, 19, 22, 23, 29, 31\}$; $\{0, 1, 7, 9, 10, 13, 14, 16, 20, 21, 24, 25, 26, 27, 28, 30, 33, 34, 35, 36\}$; $\{18\}$; $\{32\}$.

Foi efetuada a ACHA das 37 cidades, utilizando-se primeiro o coeficiente $a(k, k')$, com a opção de pesos $\pi_j = \frac{1}{p}$, e depois o coeficiente probabilístico associado a $a_{WW}(k, k')$, no âmbito da metodologia *VL*, com quatro critérios de agregação, um dos quais clássico, Single Linkage (*SL*), e três probabilísticos, *AVL*, *AV1* e *AVB* (Nicolau, 1983; Bacelar-Nicolau, 1988; Nicolau e Bacelar-Nicolau, 1998; Lerman, 1972).

Segundo a estatística global de níveis *STAT* (Bacelar-Nicolau, 1980), a partição mais significativa obtida com a aplicação do coeficiente $a(k, k')$ e o critério *SL* é: $\{0, 21, 7, 35, 10, 14, 25, 1, 26, 33, 16, 28, 13, 9, 36, 24, 27, 34, 30, 20\}$; $\{2, 4, 23, 8, 3, 6, 29, 17, 12, 5, 15, 19, 22, 31\}$; $\{11\}$; $\{18\}$; $\{32\}$. A melhor partição obtida com a aplicação do coeficiente $a(k, k')$ e os critérios *AVL*, *AV1* e *AVB* é: $\{0, 21, 34, 14, 25, 7, 35, 10, 20, 30, 1, 26, 33, 13, 27, 9, 36, 16, 28, 24\}$; $\{32\}$; $\{2, 8, 5, 15, 4, 23, 19, 11, 3, 12, 6, 29, 17, 18, 22, 31\}$. Estas classes estão próximas das fornecidas pelo painel de observadores humanos, tendo os valores de *STAT* sido, respetivamente, 17.7227 e 18.6694. No caso do coeficiente probabilístico, a partição em quatro classes (nível 32) do dendrograma obtido pelo *SL* foi idêntica à fornecida pelo painel de observadores.

4 Considerações finais

O exemplo apresentado permitiu ilustrar a aplicação do coeficiente $a(k, k')$ e do coeficiente probabilístico associado a $a_{WW}(k, k')$, à ACHA de dados de natureza intervalar (complexa). A utilização do coeficiente probabilístico VL associado a $a_{WW}(k, k')$, em vez do coeficiente $a(k, k')$, permite-nos trabalhar com valores de semelhança comparáveis numa escala probabilística. No exemplo tratado, foi com este coeficiente que se obteve o resultado que melhor se ajustou aos dados.

Referências

[1] Bacelar-Nicolau, H., "Contribuições ao estudo dos coeficientes de comparação em análise classificatória", Tese de doutoramento, Faculdade de Ciências da Universidade de Lisboa, Lisboa, 1980.

[2] Bacelar-Nicolau, H.: Two probabilistic models for classification of variables in frequency tables. In: Bock, H.-H. (eds.) Classification and Related Methods of Data Analysis, pp. 181–186. North Holland (1988).

[3] Bacelar-Nicolau, H.: The affinity coefficient. In: Bock, H.-H. (eds.) Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data, pp. 160–165. Springer (2000).

[4] Bacelar-Nicolau, H., Nicolau, F.C., Sousa, A., "Measuring similarity of complex and heterogeneous data in clustering of large data sets", *Biocybernetics and Biomedical Engineering*, Vol. 29, No. 2 (2009), pp. 9–18.

[5] Bacelar-Nicolau, H., Nicolau, F.C., Sousa, A. and Bacelar-Nicolau, H., "Clustering complex heterogeneous data using a probabilistic approach", *Stochastic Modeling Techniques and Data Analysis International Conference (SMTDA2010)*, 2010, (electronic publication).

[6] Bock, H.-H. and Diday, E. (Eds.), *Analysis of symbolic data: Exploratory methods for extracting statistical information from complex data*, Springer, Heidelberg, 2000.

[7] Guru, D.S, Kiranagi, Bapu B. and Nagabhushan, P., "Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns", *Pattern Recognition Letters*, Vol. 25, No. 10 (2004), pp. 1203–1213.

[8] Lerman, I.C. *Étude distributionelle de statistiques de proximité entre structures algébriques finies du même type: Apllication à la classification automatique*, Cahiers du B.U.R.O., 19, Paris 1972.

[9] Nicolau, F.C., "Cluster analysis and distribution function", *Methods of Operations Research*, 45 (1983), pp. 431–433.

[10] Nicolau, F.C.; Bacelar-Nicolau, H., Some trends in the classification of variables. In: C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock, Y. Baba (eds.) Data Science, Classification, and Related Methods, pp. 89-98. Springer-Verlag (1998).

Comparação de Pares de Partições em Análise Classificatória

Osvaldo Silva, Áurea Sousa

Universidade dos Açores, Dep. de Matemática, CMATI, e CEEAplA
Rua da Mãe de Deus - Apartado 1422
9501-801 Ponta Delgada, Portugal
e-mail: osilva@uac.pt; aurea@uac.pt

Helena Bacelar-Nicolau

Universidade de Lisboa, Faculdade de Psicologia, LEAD, e DataScience
Alameda da Universidade
1649-013 Lisboa, Portugal
e-mail: hbacelar@fp.ul.pt

Fernando C. Nicolau

Universidade Nova de Lisboa, FCT, Dep. de Matemática, e DataScience
Quinta da Torre
2829-516 Caparica, Portugal
e-mail: geral@datascience.org

Resumo: Existem diversos índices para a comparação de partições, o que dificulta a tomada de decisão, dado que cada um desses índices põe em evidência uma determinada peculiaridade das partições a comparar. Com o intuito de auxiliar nessa avaliação, é apresentada e exemplificada uma abordagem para a comparação de partiçoes, com base na semelhança *VL* (*Validade da Ligação*), a qual tem, entre outras, a vantagem de uniformizar a escala de medida. Por último, serão tecidas algumas considerações sobre os resultados obtidos usando as abordagens clássicas e a abordagem *VL*.

Abstract: There are several indexes to compare partitions, which complicates decision-making, given that each of these indexes highlights a particular peculiarity of these partitions. In order to assist in this assessment, is presented and exemplified an approach to the comparison of partitions based on the similarity *VL* (*Validity of the Link*), which among others, has the advantage of standardizing the measurement scale. Finally, we will present some considerations about the obtained results using the classic approaches and the *VL* approach.

palavras-chave: Análise classificatória hierárquica; comparação de partições; coeficiente de afinidade; metodologia *VL*.

keywords: Hierarchical cluster analysis; comparing partitions; affinity coefficient; *VL* methodology.

1 Coeficientes clássicos

A comparação de duas partições, no âmbito da Análise Classificatória, pode ser efetuada usando diversos índices ou coeficientes clássicos no contexto de três abordagens (com base, respetivamente, na contagem de pares, no emparelhamento das classes e na variação da informação). No entanto, cada um desses coeficientes assume um determinado valor e alguns apresentam intervalos de variação diferentes e não variam no intervalo previsto mas somente num subintervalo desse intervalo. Para que esses coeficientes sejam mais facilmente comparáveis, deve-se ter em atenção as suas características intrínsecas, categorizando-os em grupos com características similares. Em Silva (2011), por exemplo, foi considerada uma classificação dos coeficientes de comparação de partições, com base na contagem de pares.

2 Coeficientes probabilísticos

Lerman (1970) propôs a utilização de um coeficiente de semelhança de natureza probabilística entre variáveis binárias, que depois generalizou a coeficientes de proximidade entre estruturas do mesmo tipo (e.g. Lerman, 1981). Bacelar-Nicolau (e.g. 1980, 1987) desenvolveu um estudo distribucional dos coeficientes de comparação para dados binários, tendo verificado e comprovado a equivalência distribucional de uma vasta classe de coeficientes, sob a hipótese de margens fixas da tabela de contingência 2×2 associada a cada par de elementos do conjunto a classificar. Para outros coeficientes, bem como na hipótese de margens livres, embora não se verifique a equivalência distribucional exata, podemos encontrar classes de coeficientes assintoticamente distribucionalmente equivalentes e tomar sempre, como informação associada a um coeficiente, a sua função de distribuição limite (Bacelar-Nicolau, 1980; Lerman, 1981), que é um coeficiente de semelhança probabilístico γ ou da Validade da Ligação, *VL*. Tem-se então, para um coeficiente de semelhança S: $\gamma = F_S(s) = Prob_{H_0}(S \leq s) \cong Prob_{H_0}(S^* \leq s^*) \cong \phi(s^*)$, onde H_0 é uma hipótese de referência adequada, $F_S(s)$ é a função de distribuição de S, $S^* = (S - E(S))/\sigma_S$, s^* é uma realização de S^*, ϕ é a função de distribuição da lei normal reduzida e $E(S)$ e σ_S são, respetivamente, o valor médio e o desvio padrão de S, geralmente assintóticos. O coeficiente probabilístico assume valores em [0,1] (segue a distribuição Uniforme $(0,1)$) e, em geral, é calculado assintoticamente, porque a função de distribuição exata de S pode não ser conhecida. O coeficiente *VL* foi posteriormente

estendido a outros tipos de dados (e.g. Bacelar-Nicolau, 1980, 1985, 1987, 1988) e a misturas de dados de diferentes tipos.

A abordagem à comparação de partições, com recurso a coeficientes probabilísticos do tipo *VL* (Silva, 2011), apoia-se nos estudos relativos aos coeficientes de comparação para dados binários de Bacelar-Nicolau e processa-se do seguinte modo: Parte-se de um índice de semelhança s para a comparação de duas partições, tendo por base a contagem de pares de elementos que existem nas duas partições. Em seguida, calcula-se o valor de $\gamma_{P,P'}$ da função de distribuição do índice de semelhança utilizado S no ponto s, sob a hipótese de referência considerada:

$$\gamma_{P,P'} = F_S(s) = Prob_{H_0}(S \leq s) \cong Prob_{H_0}(S^* \leq s^*) \cong \phi(s^*).$$

Duas partições, P e P', serão consideradas tanto mais concordantes quanto maior for o valor de $F_S(s)$, ou seja, quanto mais improvável for ultrapassar a realização s de S na hipótese de referência. A aplicação da metodologia *VL* permite, portanto, obter índices de comparação de partições cujos valores podem ser interpretados numa escala probabilística. Assim, utilizando um coeficiente probabilístico, podemos escolher, para a comparação de partições, unicamente um coeficiente s em cada uma das classes de coeficientes distribucionalmente equivalentes.

3 Resultados e considerações finais

Foi efetuada uma Análise Classificatória Hierárquica Ascendente utilizando o coeficiente de afinidade (e.g. Bacelar-Nicolau, 1980, 1985) entre variáveis e os critérios de agregação probabilísticos *AVL*, *AV1* e *AVB* (e.g. Nicolau, 1983; Bacelar-Nicolau, 1988; Nicolau e Bacelar-Nicolau, 1998; Lerman, 1981). A metodologia utilizada na avaliação e comparação das partições, com recurso à reamostragem e a diversos coeficientes clássicos e probabilísticos, pode ser encontradas em Silva (2011). Constatou-se que os valores médios dos índices clássicos, considerando os valores obtidos em r reamostragens, variam muito de índice para índice. Já no contexto da abordagem *VL* verificou-se que todos os índices utilizados apresentam valores muito similares entre si, pelo que basta calcular o coeficiente probabilístico *VL* associado a um só desses índices para se proceder à comparação de partições (Silva, 2011). Este resultado está de acordo com a propriedade da equivalência distribucional dos coeficientes para dados binários (Bacelar-Nicolau, 1980, 1987). A comparação de partições com recurso a coeficientes proba-

bilísticos do tipo *VL* é, portanto, uma abordagem mais robusta do que a clássica: em vez de se determinarem vários destes índices, usamos o coeficiente probabilístico *VL* associado a um desses índices, para calcularmos as semelhanças entre a melhor partição, ou "*partição mais significativa*", na matriz inicial de dados e as partições (com o mesmo número de classes) obtidas nas r reamostragens, o que nos permitirá posteriormente avaliar o respetivo consenso.

Referências

[1] Bacelar-Nicolau, H., "Contribuições ao estudo dos coeficientes de comparação em análise classificatória", Tese de doutoramento, Faculdade de Ciências da Universidade de Lisboa, Lisboa, 1980.

[2] Bacelar-Nicolau, H., "The affinity coefficient in cluster analysis", *Methods of Operations Research*, Vol.53 (1985), pp. 507–512.

[3] Bacelar-Nicolau, H., "On the distribution equivalence in cluster analysis", *Pattern Recognition Theory and Applications*, Vol. 30 (1987), pp. 73–79.

[4] Bacelar-Nicolau, H.: Two probabilistic models for classification of variables in frequency tables. In: Bock, H.-H. (eds.) Classification and Related Methods of Data Analysis, pp. 181–186. North Holland (1988).

[5] Lerman, I.C., *Les Bases de la classification automatique*, Gauthier-Villars, Paris, 1970.

[6] Lerman, I.C., *Classification et analyse ordinale des données*, Dunod, Paris, 1981.

[7] Nicolau, F.C. , "Cluster analysis and distribution function", *Methods of Operations Research*, Vol. 45 (1983), pp. 431–433.

[8] Nicolau, F.C.; Bacelar-Nicolau, H., Some trends in the classification of variables. In: C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock, Y. Baba (eds.) Data Science, Classification, and Related Methods, pp. 89–98. Springer-Verlag (1998).

[9] Silva, O., "Contributos para a avaliação e comparação de partições em análise classificatória", Tese de Doutoramento, Universidade dos Açores, Portugal, 2011.