Dados socioeconómicos são bons preditores de resultados eleitorais para a Assembleia da República?

Diamantino Azevedo, Graça Gaspar, Luís Correia

Laboratório de Modelação de Agentes, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Portugal

dsazevedo@gmail.com {gg, luis.correia}@di.fc.ul.pt

Abstract. Pretende-se analisar a possibilidade de previsão de resultados eleitorais utilizando dados socioeconómicos públicos sobre Portugal e as percentagens de votos expressos dos quatro tradicionais partidos concorrentes às treze eleições para as legislativas, Assembleia da Republica, entre 1974 e 2009.

O trabalho permitiu determinar automaticamente variáveis socioeconómicas relevantes e produzir modelos prevendo resultados com estimativas de erro absoluto da ordem dos 3,5% ou inferior, dependendo dos partidos.

***Keywords*:** Previsão eleitoral, *Data Mining*, Correlação, Dados socioeconómicos, Assembleia da Republica.

1 Introdução

A previsão eleitoral comum utiliza sondagens ou inquéritos geralmente aplicando amostragens. Outros estudos na área incluem previsões de reeleições presidenciais [1,2,3], de governos ou partidos governantes [4], frequentemente utilizando apenas o Produto Interno Bruto (PIB), ou o Emprego (Desemprego) como variáveis económicas [5,6]. A democracia portuguesa, sendo recente, tem poucos trabalhos relacionados com este tema. O único identificado [4] apresenta um modelo linear, com uma estrutura geral fixa, para previsão da percentagem eleitoral nas eleições legislativas do partido no poder, no qual a principal variável é uma transformação não linear do PIB.

Este trabalho pretende analisar, utilizando técnicas de Data Mining, a possibilidade de previsão de resultados eleitorais para quatro partidos, a partir de 654 variáveis socioeconómicas (SE). Não foram referenciados trabalhos equivalentes em que fossem previstos resultados eleitorais para mais do que dois partidos em simultâneo.

2 Variáveis e processos utilizados

Foram considerados os quatro partidos tradicionais do período democrático português iniciado em 1974, o Partido Popular, (CDS, CDS/PP e PP), identificado como PP, o

Partido Social Democrata, (PPD, PPD/PSD e PSD), como PSD, o Partido Comunista Português (PCP) e o Partido Socialista (PS).

2.1 Dados utilizados

Os dados utilizados foram recolhidos em finais de 2010 e são as percentagens de votação para a Assembleia da República, com origem em www.cne.pt, e 654 variáveis SE, obtidas em www.pordata.pt, abrangendo o período entre 1974 e 2009.

2.2 Processamento

Agregação e normalização de dados SE: Os dados SE são anuais com 35 ocorrências e os eleitorais apenas treze, pelo que se optou por converter as variáveis SE para a frequência dos dados eleitorais.

Essa conversão foi efectuada por intermédio de dezanove formas de agregação diferentes que no total produziram 19*654 = 12426 novos atributos. Estas agregações resultaram de aplicar diferentes funções, nomeadamente médias, medianas, diferenças e declives, a diferentes conjuntos de valores: os relativos aos anos entre uma eleição e a anterior, os dos quatro anos anteriores a uma eleição ou os de todos os anos desde 1974 até à eleição em causa.

Os atributos obtidos foram em seguida normalizados, para média zero e desvio padrão um.

Selecção de conjuntos de atributos partidários: O elevado número de atributos resultantes das agregações levou à necessidade de efectuar uma selecção antes da aplicação de qualquer método de *Data Mining*. Para cada um dos quatro partidos seleccionou-se um conjunto de 10 atributos em cada uma das 19 formas de agregação. Escolheram-se os dez atributos com maior correlação, em valor absoluto, com o resultado eleitoral do partido, resultando em valores no intervalo [0,7928; 0,9804]**.**

Para cada partido, os 190 atributos assim seleccionados foram finalmente reduzidos a seis conjuntos diferentes constituídos, respectivamente, por:

- Os dez, cinco e um atributos com maior correlação com os resultados eleitorais, chamados respectivamente **CMax(10)**, **CMax(5)** e **CMax(1)**;
- Os atributos resultantes da Análise de Componentes Principais que explicam 95% da variância contida nos 190, chamados **ACP(95)**;
- Os cinco atributos com maior correlação com os resultados eleitorais, mas que não possuíssem entre si correlação superior a 0,8, chamados **CS**;
- Os cinco atributos cujas correlações com os resultados se situam a igual distância entre si ocupando toda a amplitude dos valores das correlações existentes, **CL**.

2.3 Aplicação dos métodos de Data Mining

Só houve treze eleições para a Assembleia da Republica entre 1974 e 2009 pelo que se optou por utilizar a validação cruzada (VC) como método de avaliação aplicando-se a variante *Leave-one-out* (VC-LOO). Utilizou-se, para medida de ajuste dos modelos obtidos, o *root mean squared error* (RMSE).

Para cada partido político e cada um dos seis conjuntos de atributos partidários seleccionados, aplicaram-se os métodos de máquina de vectores de suporte (SVM), regressão linear (RLM) e perceptrão multicamada (MLP). Foram utilizadas as implementações disponíveis na ferramenta Weka. Ao conjunto CMax(5) foi ainda aplicado um outro método (Grad), que consiste na optimização [7] de uma soma ponderada dos atributos. Os pesos óptimos foram obtidos por um método de descida de gradiente.

Em todos os métodos os valores escolhidos para os parâmetros foram os que produziram os menores valores de RMSE na VC-LOO. Nas SVM foram também testadas diversas funções *kernel*, e na maioria dos casos os melhores resultados foram obtidos com *kernel* linear.

3 Resultados

	Metodo.Atributos	RMSE	p	v	|v-p|	s
PCP	Grad.CMax(5)	0,53%	7,72%	7,86%	0,14%	6,50%
	MLP.ACP.95	0,73%	6,50%		1,36%	a
	SVM.ACP.95	0,80%	6,90%		0,96%	8,70%
PP	SVM.CS	1,07%	8,00%	10,43%	2,43%	7,70%
	MLP.CS	1,10%	7,23%		3,20%	a
	Grad.CMax(5)	1,34%	8,66%		1,77%	9,90%
PSD	Grad.CMax(5)	2,34%	27,58%	29,11%	1,53%	26,90%
	SVM.CMax(5)	3,14%	26,47%		2,64%	a
	SVM.CS	3,15%	29,95%		0,84%	30,70%
PS	SVM.CMax(5)	2,64%	41,09%	36,56%	4,53%	36,20%
	Grad.CMax(5)	2,74%	40,02%		3,46%	a
	MLP.ACP.95	3,04%	45,15%		8,59%	40,40%

Quadro 1. Melhores "Método.Atributos" obtidos com treino usando as primeiras 12 eleições. Previsões e diferenças para a 13ª eleição (2009). Todos os valores são em pontos percentuais.

Para cada partido apresenta-se, Quadro 1, os pares "Método.Atributos" que obtiveram os três menores valores de RMSE, relativos à VC-LOO a partir dos dados das primeiras 12 eleições. A coluna "p" apresenta as previsões de votação obtidas para as eleições de 2009 e a "v" os valores reais. Em "v-p" apresentam-se as diferenças absolutas entre os valores reais e previstos. A coluna "s" representa os intervalos de previsão obtidos pela Eurosondagem por sondagem à boca das urnas no dia das eleições.

3.1 Análise de resultados

O SVM e o Grad apresentaram sempre resultados que os colocaram num dos primeiros três lugares em todos os partidos. No outro extremo, a RLM nunca apresentou resultados que a classificassem nos três primeiros lugares em qualquer partido.

Os conjuntos CMax(1), CMax(10), e CL não aparecem nos três primeiros lugares. Dos restantes, CMax(5) produz o melhor resultado em três partidos. Note-se que estes atributos variam com o partido.

Analisando os atributos utilizados [8], nota-se que alguns habitualmente salientes nunca foram seleccionados, como é o caso do PIB. Constata-se igualmente que os mais vezes seleccionados são os correspondentes a dados populacionais e à educação.

O partido com erros mais reduzidos é o PCP e com maiores erros são o PSD e o PS. O facto de estes valores serem de erro absoluto sugere que os erros relativos à votação real tenham graus de importância similares para os diferentes partidos.

4 Conclusão

Determinaram-se variáveis socioeconómicas relevantes e produziram-se modelos prevendo resultados das eleições para a Assembleia da República, com estimativas de erro absoluto da ordem dos 3,5% ou inferior, dependendo dos partidos.

Não houve um método, um conjunto de atributos ou um par método/conjunto de atributos que se destacasse para todos os partidos. Não obstante, os métodos Gradientes e Máquina de Vectores de Suporte e os conjuntos de atributos CMax(5) apresentaram geralmente bons resultados. Os resultados indicam também que não são as variáveis de escolha comum (ex: PIB) as mais indicadas para previsão eleitoral.

Conclui-se que é possível utilizar dados SE para previsão eleitoral nas eleições legislativas. No entanto há que aprofundar o estudo para conclusões mais sólidas acerca das variáveis e dos métodos mais úteis.

5 Bibliografia

1. Wlezien, C., & Erikson, R. S. (1996). Temporal Horizons and Presidential Election Forecasts. American Politics Research, pp. 492-505.
2. Brown, L. B., & Chappell Jr., H. W. (1999). Forecasting presidential elections using history and polls. International Journal of Forecasting, 15, pp. 127-135.
3. Rennó, L., & Spanakos, A. P. (2006). Fundamentos da Economia, Mercado Financeiro e Intenção de Voto: As Eleições Presidenciais Brasileiras de 1994, 1998 e 2002. Revista de Ciências Sociais, 49, pp. 11-40.
4. Magalhães, P. C., & Aguiar-Conraria, L. (2009). Growth, centrism and semi-presidentialism: Forecasting the Portuguese general elections. Electoral Studies, 28, pp. 314-321.
5. Nordhaus, W. D. (1974). The Political Bussiness Cycle. Yale: Yale University.
6. Pennings, P., & Keman, H. (2002). Towards a New Methodology of Estimating Party Policy Positions. Quality & Quantity, pp. 55-79.
7. Wolfe, P., & Frank, M. (1956). An algorithm for quadratic programming. Naval Research Logistics Quarterly, 3, pp. 95-110.
8. Azevedo, D. (2012). Eleições para a Assembleia da Républica e as variações socieconomicas em Portugal. Tese Mestrado em Gestão da Informação, FCUL. Submetida.