Uma metodologia para avaliar modelos de previsão de eventos a partir de redes sociais*

Denise E. F. Brito[1], Wagner Meira Jr.[1], Roberto C. S. N. P. Souza[1], Bruna O. Neuenschwander[1], Walter dos Santos F.[1], Mauro M. Teixeira[2]

[1]Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte – MG – Brasil

{denise.brit, meira, nalon, bruna, walter}@dcc.ufmg.br, mmtex@icb.ufmg.br

Abstract. *Social networks have been used as data source to predict the occurrence of real events. However, the success of such predictions depends on the model's ability to capture the characteristics of the focused event, and determine the best models is a challenge. In this work we present a methodology to evaluate comparatively different models towards selecting the best one. We also apply our methodology to the context of predicting dengue surges, being able to identify the best model to be employed per city and the characteristics that justify the choice.*

Resumo. *Redes sociais têm sido utilizadas como fontes de dados para prever a ocorrência de eventos reais. Entretanto, o sucesso de tais previsões depende da capacidade do modelo de capturar as características do evento em foco, e determinar os modelos mais adequados se mostra como um desafio. Neste trabalho apresentamos uma metodologia para avaliar comparativamente diferentes modelos com vistas a selecionar o mais adequado. Também aplicamos a nossa metodologia no contexto de previsão da epidemia de dengue, sendo capaz de identificar o melhor modelo para cada cidade a ser empregado e as características que justificam a sua escolha.*

1. Introdução

Recentemente, informações extraídas da web têm sido amplamente utilizadas por pesquisadores para detecção dos mais variados tipos de eventos ocorridos no mundo real. Dentre as informações mais utilizadas, destacam-se os termos pesquisados em máquinas de busca e mensagens postadas publicamente em redes sociais.

As redes sociais crescem diariamente, com milhões de usuários de diferentes países, que publicam mensagens de conteúdos diversos, inclusive aspectos de suas vidas pessoais, tais como suas preferências musicais, seu humor no momento presente e até mesmo seu estado de saúde.

Os eventos do mundo real possuem localização espacial e temporal bem definidas. Logo, para detectar eventos, se faz necessário escolher uma rede social cuja estrutura permita que o usuário poste mensagens públicas com um certo contexto, e que essas mensagens possam ser coletadas juntamente com a localização do usuário, quando ele declarar, e a data e hora da mensagem. O Twitter cumpre essas premissas, permitindo que

*Este trabalho é parcialmente financiado pelo CNPq, Capes, Fapemig e InWeb.

o usuário poste mensagens de até 140 caracteres, e fornece uma API pública[1], além de ser muito popular no Brasil.

Uma tarefa chave no processo de utilização de dados de redes sociais para prever a ocorrência de eventos é o modelo de previsão utilizado, que deve ser capaz de se adequar às características dos dados, como ruído e variações motivadas por aspectos externos. Neste trabalho, propomos uma metodologia que permite comparar experimentalmente diferentes modelos e escolher, para cada caso, o modelo mais adequado para fins de previsão. Utilizamos mensagens coletadas do Twitter para estimar o número de casos reais de dengue. A dengue é uma doença de regiões de clima tropical e subtropical[2], transmitida pela picada do mosquito Aedes aegypti infectado, que pode evoluir para sérias complicações e causar até mesmo a morte.

Nos últimos anos, vários pesquisadores vêm utilizando dados de redes sociais e de pesquisas feitas em máquinas de busca para prever eventos reais. Em [Chunara et al. 2012], dados do HealthMap e do Twitter foram usados para modelar o surto de cólera ocorrido no Haiti em 2010. Em [Culotta 2010], foi demonstrado haver alta correlação entre tweets sobre doenças com sintomas de gripe e dados oficiais sobre as mesmas nos Estados Unidos. Em [Signorini et al. 2011], [Lampos and Cristianini 2010], mensagens do Twitter foram utilizadas para monitorar a epidemia de gripe. Em [Chan et al. 2011] e [Althouse et al. 2011], epidemias de dengue foram monitoradas através de termos de pesquisas na web.

Em trabalhos anteriores [Gomide 2012] [Brito et al. 2012], foi demonstrado haver alta correlação entre o número de tweets relatando experiência pessoal com dengue e o número de casos reais da doença registrados pelo Ministério da Saúde. Foi utilizada uma função de previsão linear, para encontrar o número de casos previstos através do número de tweets sobre dengue, e calculado o Z-score, para inferir, respectivamente, o nível de incidência e a tendência da doença. De acordo com a classificação do Ministério da Saúde, que considera a incidência de dengue dividida em três faixas, baixa, média e alta, em relação ao número de casos por cem mil habitantes, foi atingida uma acurácia maior que 99%.

O objetivo deste trabalho é identificar um modelo mais adequado ao contexto da dengue, não apenas para inferir a faixa de incidência da doença, mas para aproximar o número de casos previstos do número de casos reais, investigando várias funções de previsão geradas por modelos de regressão diferentes, comparando-os ao método de mínimos quadrados.

2. Metodologia

Nesta seção apresentamos a nossa metodologia para avaliar experimentalmente e selecionar o modelo mais adequado para previsão. O primeiro passo consiste em selecionar os modelos que podem descrever os dados reais. Em uma primeira análise, há uma variável aleatória Y, que corresponde à intensidade do evento real a ser detectado, em função de outra variável aleatória X, que é o número de tweets. O objetivo é calcular o valor estimado de Y, tendo amostras de X.

[1]http://apiwiki.twitter.com/

[2]http://www.who.int/topics/dengue/en/

Os dados são discretos e não negativos, o que sugere um problema de contagem, onde a regressão de Poisson pode ser adequada [Coelho and Codeço 2012]. A regressão de Poisson assume que Y possua distribuição de probabilidades de Poisson. A distribuição de Poisson é caracterizada pelo fato de a esperança ser igual à variância, entretanto, isso raramente ocorre com dados do mundo real. Quando a variância é maior que a esperança, dá-se o nome de sobredispersão. O modelo de Poisson não é adequado neste caso, sendo normalmente substituído por um modelo de quasi-Poisson ou binomial negativa.

O método de mínimos quadrados utilizado nos trabalhos anteriores assume que a variável Y esteja perturbada. Como analisamos dados informais coletados de redes sociais e classificados automaticamente, a variável X também pode conter ruído. Sendo assim, foi considerada também a regressão ortogonal linear. Essa regressão assume que haja erro em ambas as variáveis [Markovsky and Van Huffel 2007], ou seja, que tanto X quanto Y possuam distorção.

Definidos os modelos a serem testados, calculamos os parâmetros das funções de previsão. Para as regressões de Poisson, quasi-Poisson e binomial negativa, o valor esperado de Y condicionado a X é [Zeileis et al. 2008] [Ver Hoef and Boveng 2007]: $E[y_i|x_i] = \mu_i$, $g(\mu_i) = ax_i + b$, onde x_i é a amostra de tweets na semana i, $g()$ é a função de ligação. Para essas regressões geralmente é utilizada a função logarítmica, onde a e b são os parâmetros. Para a regressão ortogonal linear, assim como para o método de mínimos quadrados: $E[y_i|x_i] = ax_i + b$. Os parâmetros da regressão ortogonal linear foram calculados conforme [Petras and Bednarova 2010].

Para avaliar as regressões, verificamos se os dados cumprem as premissas dos modelos, e em seguida, para cada um deles, calculamos o erro médio ponderado pelo número de habitantes dos municípios analisados (aqueles que possuem dados suficientes). Escolhemos o de menor erro para ser contrastado com a regressão linear, por meio de um teste estatístico.

3. Aplicação: modelo de previsão de epidemia de dengue

Os dados informais utilizados foram as contagens de tweets coletados contendo os termos "dengue" ou "Aedes aegypti", com a localização do usuário em nível de cidade, classificados pelo algoritmo LAC [Veloso et al. 2006] como sendo experiência pessoal. Os dados oficiais foram as contagens de casos reais de dengue registrados pelo Ministério da Saúde. Ambos foram agrupados por semana, para todas as cidades brasileiras com mais de cem mil habitantes, totalizando 285 cidades. O período de tempo considerado para o cálculo dos parâmetros foi de 50 semanas de 2011 (todo o ano, exceto as semanas 1 e 22 por falta de dados) e para o teste dos modelos, as semanas de 1 a 30 de 2012.

Ao calcularmos o parâmetro de sobredispersão de acordo com o modelo quasi-Poisson: $\sum_{i=1}^{n} \frac{model\$weights \times model\$residuals^2}{model\$df.residual}$ verificamos que em mais de 10% das cidades há sobredispersão (o parâmetro é muito maior do que 1) e, portanto, não é recomendado utilizar o modelo de Poisson.

Avaliamos então os demais métodos de forma global utilizando o erro médio ponderado: a regressão binomial negativa obteve 25220.54414; a quasi-Poisson, 9739.89995; a ortogonal, 6532.08542 e a linear, 5883.06794. Embora o modelo linear tenha obtido o

melhor resultado na média, percebemos que ele não é capaz de tolerar ruídos como os verificados em alguns casos. Neste caso, a regressão ortogonal linear pode ser mais efetiva, em particular quando analisamos as correlações entre tweets e casos de dengue para várias cidades e verificamos que existem muitas semanas nas quais o número de casos é significativamente diferente e o número de tweets permanece o mesmo.

4. Resultados

A Figura 1 mostra os dados de treino e o número de casos previstos por número de tweets, calculados com a regressão quasi-Poisson e binomial negativa, para Fortaleza. A função de ligação logarítmica faz com que a curva de tweets por casos previstos tome a forma de uma exponencial, o que não corresponde à realidade, pelo menos para as cidades avaliadas.

Algumas vezes, ocorrem picos de tweets por razões que extrapolam o contexto das mensagens e, nesses casos, tanto a regressão quasi-Poisson quanto a binomial negativa levariam a alarmes falsos, visto que não se assemelham nem mesmo aos dados de teste. No entanto, entre essas, a quasi-Poisson se aproxima mais da curva real.

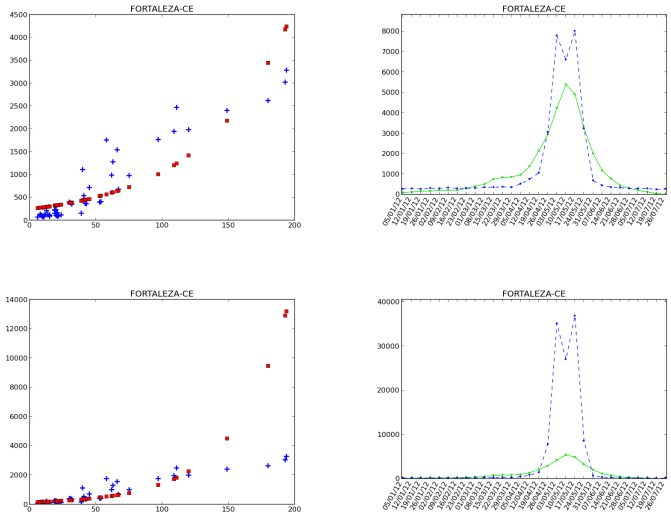

Figura 1. Resultados para Fortaleza. Acima, regressão quasi-Poisson. Abaixo, regressão binomial negativa. Do lado esquerdo, dados de treino. As cruzes azuis são os casos reais e os quadrados vermelhos, casos previstos. À direita, dados de teste. A linha em verde representa os casos reais e a azul, pontilhada, os casos previstos.

A regressão ortogonal linear é mais sensível do que a linear simples, por isso, também gera alarmes falsos para algumas cidades. Porém, para aquelas com tweets suficientes (com 7 tweets ou mais por semana, durante pelo menos 20 das 30 semanas de 2012 analisadas), o número de casos previstos com esse modelo ficou bem semelhante ao gerado pelo método de mínimos quadrados e para Recife, a ortogonal aproximou melhor a curva de casos reais do que a linear simples (Figura 2).

Para testar a hipótese de que o modelo ortogonal linear é estatisticamente diferente do linear simples para o contexto deste trabalho, calculamos o teste F com confiança de 95% e graus de liberdade 48, com os dados de treino. Assim, para todas as cidades

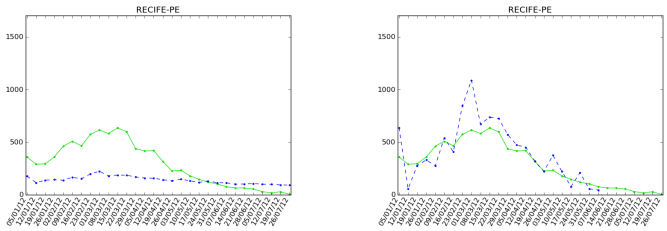

Figura 2. Resultados para Recife. À esquerda, regressão linear simples e à direita, regressão ortogonal linear. A linha em verde representa os casos reais e a azul, pontilhada, os casos previstos.

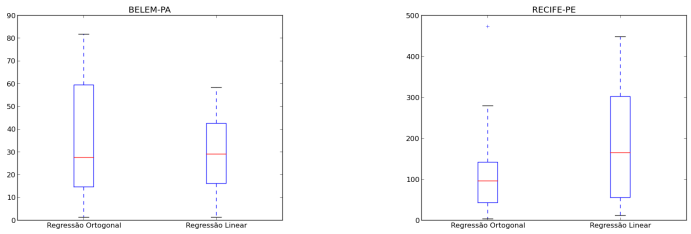

Figura 3. Diferenças absolutas entre casos previstos e reais das 30 semanas de teste usando as funções geradas por regressão ortogonal linear e por mínimos quadrados.

que obtiveram o valor do teste acima de 1.61537, descartamos a hipótese nula de que os modelos são equivalentes e, então, o modelo ortogonal é melhor. A Tabela 1 mostra os valores do teste para todas as cidades com tweets suficientes.

Tabela 1. Valores do Teste F

Cidade	UF	Teste F	Cidade	UF	Teste F
RIO DE JANEIRO	RJ	1.007859	FORTALEZA	CE	1.003830
CUIABÁ	MT	1.204502	BELÉM	PA	1.698209
MANAUS	AM	1.004843	RECIFE	PE	3.443550
CURITIBA	PR	1.000772	BRASÍLIA	DF	1.172616
PORTO ALEGRE	RS	1.001448	SALVADOR	BA	1.104516
SÃO PAULO	SP	1.080026	JOÃO PESSOA	PB	1.102533
GOIÂNIA	GO	1.090568	SANTOS	SP	1.052270
ARACAJU	SE	2.083052	NATAL	RN	1.034813
BELO HORIZONTE	MG	1.615080	MOSSORÓ	RN	1.097748

A Figura 3 mostra o erro absoluto durante as 30 semanas de dados de teste para as cidades com teste F acima de 1.61537. Para Belém, o valor mediano do erro para a ortogonal foi praticamente igual, porém com dispersão maior. Para Recife, a regressão ortogonal aproximou melhor os casos previstos dos reais.

5. Conclusões

Como uma direção futura, pretendemos validar o modelo em outros contextos. Para o cenário da dengue, os modelos de Poisson, quasi-Poisson e binomial negativa com função

de ligação logarítmica não se adequam, pois, em relação à primeira, os dados apresentarem sobredispersão, e as demais por produzirem uma curva de tweets por casos previstos exponencial, gerando um erro médio ponderado muito grande, devido aos alarmes falsos gerados por picos inexplicados de tweets.

De acordo com o teste F utilizando os dados de treino, a regressão ortogonal linear foi significativamente diferente do que o método de mínimos quadrados para as cidades de Belém, Recife e Aracaju. Para as demais, os dois métodos se mostraram equivalentes, e sendo assim, ambos podem ser utilizados. Para os dados de teste de 2012, a regressão ortogonal obteve erro mediano e dispersão menores do que a linear para Recife.

Referências

Althouse, B., Ng, Y., and Cummings, D. (2011). Prediction of dengue incidence using search query surveillance. *PLoS Negl Trop Dis*, 5(8):e1258.

Brito, D., Gomide, J., Santos, W., Jr., W. M., Veloso, A., and Almeida, V. (2012). Um sistema de alarme para vigilância epidemiológica de rumores utilizando redes sociais. In *Proceedings of the 27th Brazilian Symposium on Databases*, São Paulo, BR.

Chan, E. H., Sahai, V., Conrad, C., and Brownstein, J. S. (2011). Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. *PLoS Negl Trop Dis*, 5(5):e1206.

Chunara, R., Andrews, J. R., and Brownstein, J. S. (2012). Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. *The American Journal of Tropical Medicine and Hygiene*, 86(1):39–45.

Coelho, F. and Codeço, C. (2012). Análise preliminar: Twitter x dengue.

Culotta, A. (2010). Detecting influenza outbreaks by analyzing Twitter messages. *ArXiv e-prints*.

Gomide, J. S. (2012). Mineração de Redes Sociais para Detecção e Previsão de Eventos Reais. Master's thesis, Universidade Federal de Minas Gerais, BR.

Lampos, V. and Cristianini, N. (2010). Tracking the flu pandemic by monitoring the social web. In *IAPR CIP 2010*, pages 411–416.

Markovsky, I. and Van Huffel, S. (2007). Overview of total least-squares methods. *Signal Process.*, 87(10):2283–2302.

Petras, I. and Bednarova, D. (2010). Total least squares method. http://www.mathworks.com/matlabcentral/fileexchange/31109.

Signorini, A., Segre, A. M., and Polgreen, P. M. (2011). The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic. *PLoS ONE*, 6(5):e19467.

Veloso, A., Meira Jr., W., and Zaki, M. J. (2006). Lazy associative classification. In *International Conference on Data Mining*, pages 645–654. IEEE Computer Society.

Ver Hoef, J. and Boveng, P. (2007). Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–72.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in r. *Journal of Statistical Software*, 27(8):1–25.